BI na Era do Big Data para Cientistas de Dados

indo além de cubos e dashboards
na busca pelos porquês, explicações e padrões

Autor: Stanley Loh

Stanley Loh

BI na era do big data para cientistas de dados:
indo além de cubos e dashboards na busca pelos porquês, explicações e padrões

1a edição

Porto Alegre

Stanley Loh

2014

Prefixo Editorial: 916683
Número ISBN: 978-85-916683-1-1

Formas de citação:

LOH, Stanley. **BI na era do big data para cientistas de dados** - indo além de cubos e dashboards na busca pelos porquês, explicações e padrões. Porto Alegre, 2014.

Loh, S. (2014). *BI na era do big data para cientistas de dados: indo além de cubos e dashboards na busca pelos porquês, explicações e padrões.* Porto Alegre, 158 p.

Conteúdo

Lista de Figuras

1 Introdução

O melhor exemplo para explicar o que é Business Intelligence (BI) para um leigo é o caso da GM e o sorvete de baunilha. Conta a lenda que um consumidor comprou um carro da GM e depois mandou uma carta se queixando. A queixa era a seguinte: quando ele ia na sorveteira e pegava o sorvete de baunilha, ele voltava para o carro e este demorava a dar partida; se ele pegasse qualquer outro sabor de sorvete, ele voltava para o carro e este "pegava" de primeira.

Conta ainda a lenda que isto virou piada na GM, uma vez que ninguém imaginava o que o sabor de um sorvete teria a ver com o problema no carro. Acredita-se que um engenheiro foi investigar o caso. Apresentou-se ao cliente e juntos foram testar a teoria que o cliente alegava. Foram até a sorveteria e compraram o sorvete de baunilha. Voltaram para o carro e realmente o carro não deu partida na primeira tentativa nem nas seguintes. Esperaram um pouco, e tentaram de novo. Aí sim o carro ligou. Voltaram para a casa e depois de comerem o sorvete fizeram o mesmo teste só que pegando um sorvete de sabor diferente. Quando voltaram para o carro, a surpresa: o carro "pegou" de primeira. Bom, mas poderia ser acaso ou coincidência. Então testaram diversas vezes, usando métodos estatísticos e o resultado ... sempre o mesmo.

O engenheiro sabia que o sabor do sorvete não poderia influenciar o problema, mas certamente ali havia algum fator que estaria associado ao problema. E este fator tinha a ver com o sabor. Então ele descobriu que o sorvete de baunilha ficava na entrada da sorveteria, enquanto que os demais ficavam nos fundos. Ao entrar e comprar o sorvete de baunilha, o dono do carro demorava menos que se pegasse outro sabor. Havia uma peça no carro que precisava resfriar para o carro poder ligar. Menos tempo na sorveteria, menos tempo para a peça resfriar e o carro não ligava. Desta forma, o engenheiro descobriu a causa para o problema.

Eu sempre cito isto como um exemplo de BI, mesmo tendo sido feito manualmente, isto é, sem ajuda de bancos de dados e software (tecnologias da informação). Mas este caso ilustra bem o objetivo de um processo de BI e como ele pode ser feito, não só para leigos mas também para analistas de BI experientes.

Hoje em dia há diversas definições para BI e muitas vezes profissionais dizem estar fazendo BI quando na verdade estão gerando informações com sistemas de informações gerenciais, ou seja, através de ferramentas para geração de dashboards, gráficos, relatórios e análises visuais (visualização de informações).

A seguir, explicarei um pouco melhor o que entendo de BI e qual sua diferença para sistemas gerenciais. Também falaremos da buzzword Big Data, o que significa e o que implica para processos de BI.

O livro tem o objetivo primeiro de explicar técnicas e métodos que ajudem processos de BI. Mas vamos procurar dar ênfase ao que ainda não foi dito em outros livros do gênero. Por isto, vamos enfatizar que o objetivo principal de um processo de BI é encontrar causas, explicações e padrões.

Estaremos trazendo conhecimentos de outras áreas. Em muitas partes do livro, o leitor talvez imagine estar lendo um livro sobre investigações e descobertas científicas. Isto

não está errado. Não é o único enfoque, mas é uma das formas de se ver o BI. Temos muito a aprender com a história dos grandes cientistas da Humanidade. A diferença talvez não esteja nos métodos, apesar de que eles também evoluem. Mas hoje temos muito mais dados e mais complexos (Big Data) e ferramentas mais avançadas, principalmente ferramentas de software. Por isto, o termo Cientista de Dados é tão atual.

Por isto, vamos enfatizar que os dados são muito importantes para o processo, incluindo a forma e as condições como são coletados e armazenados. Não basta discutirmos as formas de análise se os dados analisados não tiverem qualidade (*garbage in, garbage out*).

O leitor se quiser poderá pular algumas seções, conforme seu interesse. Os capítulos não estão numa sequência de aprendizado. Dentro dos capítulos sim, a ideia é manter uma certa ordem de leitura.

1.1 A Evolução dos Sistemas de Informação

A Tecnologia da Informação, que inclui computadores, redes de comunicação e software, iniciou nas organizações para armazenar dados em grande volume e auxiliar pessoas em cálculos. Por isto, as primeiras aplicações a serem automatizadas eram controle de estoque, folha de pagamento e contabilidade. Os sistemas deste tipo chamam-se rotineiros ou transacionais.

Com o passar do tempo, viu-se que era possível extrair novas informações daquelas armazenadas e apresentar isto na forma de relatórios. Então, de um sistema de controle de estoque, era possível saber quais os produtos mais vendidos, os que menos saíam e desenhar um gráfico médio das saídas dos produtos ao longo do tempo. Da mesma forma, de um sistema de folha de pagamento era possível saber qual o cargo ou setor que mais custo dava para a empresa. E de sistemas de contabilidade, era possível medir o que já tinha sido gasto ao longo o tempo e o que se esperava recebe no tempo futuro. Os relatórios evoluíram para se tornarem sofisticados sistemas de informações gerenciais (SIGs), incluindo a geração de diferentes tipos de gráficos e painéis com diferentes informações (dashboards). O livro de Bertin (1983) apresenta e explica as aplicações de diferentes tipos de gráficos.

Apesar da utilidade incontável dos sistemas de informações gerenciais, o que faz deles úteis até hoje em qualquer empresa, profissionais tais como administradores, tomadores de decisão, gestores de informações e executivos ainda precisavam de um tipo de apoio mais sofisticado, algo que pudesse facilitar a tomada de decisão.

Primeiro, era necessário descrever dados para encontrar características para ajudar a entender o que estava acontecendo ou o que havia acontecido. Esta é a função dos modelos descritivos, que buscam identificar padrões. Os sistemas de BI entram aqui, auxiliando a entender por que as coisas acontecem, quais são as causas ou explicações para certos eventos ou fenômenos.

Após os sistemas de BI, vêm os sistemas de Business Analytics, que utilizam modelos preditivos para tentar prever eventos futuros ou predizer valores para atributos. Incluem-

se neste tipo de apoio, os sistemas conhecidos como sistemas de apoio à decisão (SADs).

Então podemos dividir o processo todo da seguinte forma, sistematizando o que se quer saber em relação a como encontrar tais respostas:

- O que aconteceu? Exemplo: quais os totais de venda no mês anterior.

Para isto, existem os SIGs, que buscam informações em sistemas transacionais e geram relatórios (novas informações ou novas formas de apresentação).

- O que está acontecendo ? Exemplo: nossas vendas estão crescendo ou diminuindo ?

Para isto, podemos usar também SIGs ou sistemas de Data Mining, que encontram padrões estatísticos nos dados.

- Por quê ? Exemplo: por que as vendas estão caindo ?

Aqui é que entra o BI, procurando descobrir as causas para os eventos observados.

- O que acontecerá no futuro ? Exemplo: se mantivermos os níveis de venda mas diminuirmos o preço de venda, o que acontecerá com nosso lucro ?

As previsões e análises *what-if* são feitas com sistemas de Business Analytics e Sistemas de Apoio à Decisão.

- O que gostaríamos que acontecesse ? Exemplo: queremos aumentar a receita total em 10%.

Aqui são essenciais técnicas de planejamento e definição de metas. Mas elas só funcionam quando entendermos as causas e inter-relações entre variáveis.

1.2 BI X Sistemas Gerenciais

Hoje em dia, BI é confundido com as aplicações que geram relatórios, chamadas há muito tempo de Sistemas de Informações Gerenciais - SIGs (em inglês, Management Information Systems - MIS). SIGs e EIS (Executive Information Systems) geram relatórios, geralmente gráficos, sintetizando informações ou permitindo compará-las. Eles geram informações novas, que não estavam explícitas na base de dados, ou permitem visualizar as informações de tal forma que o usuário do sistema descubra rápida e facilmente algo novo. Como exemplos, temos relatórios que apontam os produtos mais vendidos ou mais lucrativos, melhores vendedores ou lojas com melhores resultados, época em que cada produto sai mais ou menos (vendas ao longo do tempo) e etc.

Tais sistemas são há muito tempo importantes para as empresas. Entretanto, o BI deve ir mais fundo que os SIGs, seu papel é mais nobre. O processo de BI deve ajudar as pessoas a descobrirem as causas para tais acontecimentos ou descobertas. Assim, o SIG aponta qual o produto mais vendido, mas o BI deve procurar descobrir porque este produto é mais vendido que os outros ou porque os outros não vendem tão bem. O SIG aponta a época em que um produto vende mais, já o BI busca saber por que o produto vende mais nesta época e menos nas outras.

Em resumo, SIGs ajudam a entender o que aconteceu ou o que está acontecendo (ex.: totais de venda no mês anterior, qual a taxa de crescimento de nossas vendas); BI procura por causas ou explicações (ex.: por que as vendas estão caindo).

Ambos os tipos de sistemas de informação (SIGs e BI) procuram auxiliar na tomada de decisão, uma vez que este é o objetivo geral de qualquer sistema de informação. Entretanto, a forma de apoio é que é diferente em cada tipo.

1.3 Dados X Informação X Conhecimento X Inteligência

É importante distinguir dados, informação, conhecimento e acrescentar o conceito de inteligência. A Figura 1 apresenta uma tabela. O valor 35 na 2a linha com a 2a coluna é um dado. Dados são representações de informações. Sozinhos não dizem nada. Quando entendemos que o 35 significa a idade do cliente José, em anos, estamos transformando o dado em informação. As pessoas trabalham com informações mas a tecnologia armazena dados.

Já conhecimento seria: "Todos os clientes da cidade de SP têm saldo médio maior que 9 mil reais". Notem, isto não é uma informação explícita na tabela. Só conseguimos chegar a este conhecimento se cruzarmos informações diferentes. Conhecimento, portanto, vem das informações, mas está acima. As pessoas recebem muitas informações no seu dia a dia, mas nem tudo fica retido, nem tudo é útil, nem tudo será utilizado mais adiante. O que resta, o que é útil, o que é utilizado forma o conhecimento desta pessoa.

Cliente	Idade	Saldo Médio	Cidade
José	35	9000	SP
João	30	4000	Santos
Ana	25	8600	Rio
Maria	23	3000	Ribeirão Preto
Carlos	34	9700	SP

Figura 1: Dados X Informação X Conhecimento

Já o conceito de Inteligência (alguns chamam Sabedoria) está acima de conhecimento. Imagine um grupo de pessoas numa sala fechada (nada entra ou sai) recebendo uma tarefa: quebrar a cadeira onde estão sentados. Admitamos que todos possuem a mesma força física e foram criados e educados em famílias e escolas semelhantes. Ou seja, possuem o mesmo nível de conhecimento, obtido por estudos nas escolas, leituras em casa, viagens, experiências, etc. Algumas destas pessoas conseguirão resolver o problema e outras não. Mas por que, se todas possuem a mesma força física e os mesmos conhecimentos ? A diferença está na forma como cada um utiliza o

conhecimento que tem e as conexões que faz em seu cérebro. Isto é inteligência, ou seja, saber resolver problemas utilizando o conhecimento que possui. E isto se aplica também a poder resolver problemas novos, usando adaptações, analogias, etc.

BI então, como o nome "inteligência" indica, deve ajudar pessoas e organizações a resolverem seus problemas e alcançarem seus objetivos.

1.4 O que é BI então ?

Primeiro de tudo, cabe salientar que BI é um processo. Existem técnicas, tecnologias e software para BI, mas BI é um processo que envolve métodos, técnicas, tecnologias, pessoas, informações, fontes de informações, métricas, ferramentas, etc.

Em resumo, o processo de BI tem por objetivo encontrar causas ou explicações para eventos ou resultados. E estes resultados podem ser bons ou ruins, ou seja, o BI deve procurar causas dos problemas e as melhores práticas do sucesso. Não basta saber qual o problema mais comum em máquinas de uma indústria; a empresa precisa saber o porquê disto, para poder atacar as causas e diminuir os prejuízos. Não basta saber qual o melhor vendedor, a empresa precisa saber por que ele é o melhor, para que as tais boas práticas deste vendedor possam ser replicadas para todos os outros vendedores.

O processo de BI pode fazer uso de sistemas gerenciais, ferramentas de Data Mining e tudo isto com dados vindos de sistemas rotineiros ou transacionais. Podemos dizer que BI está na ponta do fluxo de informação, muito próximo de quem toma decisões.

O grande objetivo do BI é acabar com o "achismo" ou "empirismo". Ouve um caso em que os ouvidores de uma concessionária de rodovias achavam que o trecho mais problemático era um. Quando foram feitas análises estatísticas sobre as ocorrências registradas, descobriu-se que o trecho com mais problemas era outro.

O conhecimento nos faz mais inteligentes; pessoas e empresas que aprendem. Como Kuhn relata, até a metade do século 19 não se usava conhecimento na indústria, somente nas Ciências. Assim como o conhecimento científico mudou o paradigma da Ciência na idade média, o uso intensivo de conhecimento acelerou inovações e permitiu à indústria aproveitar os que as metodologias científicas ensinavam nas ciências, fazendo a prática da indústria menos empírica.

BI também procura encontrar explicações para eventos mas fundamentadas em dados. Não basta saber o que está acontecendo, é preciso analisar as causas para poder repetir o sucesso ou evitar fracassos.

A busca por padrões também é objetivo do BI. No oceano de dados, é preciso tentar encontrar uma ordem para que os dados possam fazer sentido e serem úteis. Uma base de clientes onde não conseguimos identificar quem é nosso cliente, o que ele quer, quais suas características, não serve para nada além de confundir.

E isto tudo fez surgir a Era do Conhecimento, apoiada pelas chamadas tecnologias da informação.

1.5 Big Data e Information Explosion

Estamos vivendo numa era de grandes volumes de informações. O volume de informações é medido em exabytes. A escala é assim: bit, byte, kylobyte, megaybte, gigabyte, terabyte, petabyte, exabyte, zettabyte, yottabyte.

Chamam isto de Big Data (Tole, 2013), mas anos atrás Korth e Silberschatz já falavam sobre isto e chamavam esta nova revolução de "explosão de informações". Sim eles comparavam estes novos acontecimentos a revoluções como a invenção da imprensa por Gutenberg (distribuição de informações a todo canto do mundo) e invenção do telefone por Graham Bell (informação distribuída imediatamente, em tempo real).
O volume aumenta a cada ano pelas seguintes razões:

- o armazenamento de dados hoje é barato (discos rígidos e DVDs) ou mesmo de graça (serviços de hospedagem free na Web);
- as pessoas estão mais familiarizadas com a tecnologia e consequentemente geram e armazenam mais informações (crianças de 2 anos já sabem usar celulares e computadores e a 3a idade está menos tecnofóbica);
- a tendência atual de "não jogar nada fora", que começou com o Gmail dizendo que ninguém precisava "deletar' seus e-mails;
- mais possibilidades de serviços para publicar e difundir informações (blogs, twitter, e-mail, redes globais, conexões sem fio, etc.).

A Revista Veja, edição de maio de 2013 (ed.2321, n.20, ano 46) tratou deste assunto na sua reportagem de capa. Eles falam que o Big Data se deve a 3 Vs: volume, velocidade e variedade. Além do grande volume de dados gerados, coletados, armazenados, etc, a velocidade de transmissão (banda larga por cabo ou 3G ou wifi etc.) e a diversidade de tipos de informações (planilhas, textos, imagens, sons) ajudam a sobrecarregar o ser humano e as organizações.

Segundo a reportagem da revista Veja, a cada dia:
- 2,5 exabytes de informação são produzidos pela humanidade;
- 375 megabytes de dados são acumulados por cada família;
- 24 petabytes são processados pelo site do Google;
- 10 petabytes correspondem aos e-mails enviados;

E ainda, 385 terabytes guardam todo o catálogo da Biblioteca do Congresso americano, a maior do mundo, enquanto que 1,8 zettabyte armazena todos os dados acumulados pela civilização em um ano. Comparando com os 3 exabytes que a humanidade conseguia guardar em 1986 (hoje produzimos quase o dobro disto em 2 dias), estamos vivendo em tempos exponenciais.

Além disto, a complexidade do ser humano foi passada para a Tecnologia da Informação. Hoje podemos armazenar dados não estruturados, ou seja, imagens, vídeos, sons e textos.

E some-se a isto tudo a possibilidade de análises mais complexas com o desenvolvimento de softwares com funções de Inteligência Artificial. Se antes, os gestores apenas queriam encontrar endereço de clientes num banco de dados, hoje querem saber qual a faixa de idade que mais compra os produtos de uma certa faixa de preço e isto tudo apresentado por loja, cidade e país.

Alguém vai dizer que o volume de informações é bom, porque as pessoas e organizações possuem mais informação para tomar decisões. Por outro lado, vivemos no stress por termos mais opções para escolher, mais informações para ler, mais conhecimento para aprender e por não conseguirmos lidar com tanta informação disponível e nem mesmo conseguir encontrar as informações que precisamos (*information overload*). É como uma mesa cheia de papéis e a gente sabendo que a informação que a gente procura está em algum destes papéis nesta mesa.

BI passa então a ser primordial para as organizações poderem funcionar de forma "organizada" e não se afogarem com tanta informação.

2 Busca e Identificação de Padrões

Quero ratificar mais uma vez que o objetivo do processo de BI é ajudar pessoas e organizações a encontrarem causas e não só apresentar informações, como fazem sistemas gerenciais. A busca por causas passa por analisar dados, talvez grandes quantidades, à procura de padrões, modelos ou repetições. Se não encontrarmos padrões, não temos como afirmar quais eventos geram quais consequências. Será uma confusão de dados, sem ordem, sem explicações.

A identificação de padrões é parte da nossa vida. A descoberta de padrões iniciou há milhares de anos atrás. Nossos antepassados conseguiam prever as variações do tempo, as estações, os ciclos das plantações, as fases lunar e eclipses, e até mesmo o surgimento de reis. E hoje em dia não é diferente. Quem não dá palpites sobre como será o tempo, se vai chover, fazer sol, calor, observando as nuvens ? Ou se o próximo inverno será mais frio ou menos frio do que o ano anterior, pelo que viu no outono ? Se um local público vai lotar ou não para um evento, observando o movimento das pessoas chegando ? Ou quantas pessoas há num concerto ao ar livre num parque público, lembrando o último evento que ocorreu ali ? Mesmo algumas superstições são exemplos de padrões, que acreditamos que irão se repetir. Numa entrevista de negócios, usar a mesma roupa de um acontecimento bom. Sentar no mesmo lugar do último título para torcer por seu time. Não quebrar espelho, pois quando isto ocorreu, um evento de má sorte também ocorreu junto.

Vemos padrões no ambiente, no que vemos e sentimos e daí criamos modelos para o clima, trânsito, estereótipos de pessoas, etc. Alguns modelos mais completos que outros, alguns mais precisos, outros com mais exceções. Vemos até mesmo padrões na nossa própria vida. Wolf (2010) relata uma série de casos de pessoas analisando seus próprios dados. Como o cara que descobriu estatisticamente que café não ajudava na concentração dele (ele acreditava no contrário, mas fez experimentos e descobriu um novo padrão, mais exato).

2.1 Modelos e Padrões

A classificação é um instinto do ser humano. Tentamos colocar tudo em grupos (pessoas, produtos, eventos, animais, plantas, etc.). Mesmo num texto como este, as informações estão agrupadas. Acreditamos que podemos reduzir tudo a um modelo único ou a poucas regras. Esta é a busca eterna dos físicos, para entender a Natureza e o Universo. Einstein acreditava que há uma ordem na desordem, mas que os padrões ainda devem ser descobertos.

A classificação facilita nosso entendimento do mundo e agiliza nossa tomada de decisão. Os padrões servem para minimizar a incerteza. Se encontramos uma situação nova e verificamos que ela se encaixa num padrão já entendido, já sabemos que atitudes tomar naquela situação. Este é um dos conceitos de inteligência: saber adaptar-se a novas situações e conseguir resolver problemas novos. Isto não significa que vamos

usar exatamente as mesmas ações. A inteligência humana pressupõe a adaptação dos padrões para novas realidades.

A melhor forma de entender um conjunto de dados é estabelecer um modelo para ele. O modelo explicaria as características comuns aos dados, as relações entre os dados, as relações de causalidade e influência ao longo do tempo. O ser humano busca padrões no seu contexto porque se sentirá mais parte do contexto e menos um alienígena. É como uma necessidade humana, para não ficarmos loucos. O que não se encaixa nos nossos padrões, como por exemplo eventos paranormais, acabamos considerando como bruxarias.

Mas o que é um modelo ? Vejamos algumas definições de modelo: aquilo que serve de objeto de imitação; aparelho ou conjunto de aparelhos que permitem a reprodução de determinada peça por processos usados em fundição para o preparo de objetos de metal; molde; protótipo ou exemplo que se pretende reproduzir ou imitar; um exemplar que se deve seguir e imitar pela sua perfeição; imagem ou desenho que representa o objeto que se pretende reproduzir esculpindo, pintando ou desenhando; pessoa exemplar, perfeita, digna de ser imitada; esquema teórico em matéria científica representativo de um comportamento, de um fenômeno ou conjunto de fenômenos. No contexto deste livro, a melhor definição é a última: um esquema ou estrutura que representa um comportamento (de um evento ou conjunto de eventos). Ao longo deste livro, usaremos alguns sinônimos para modelo, tais como: padrão, regras, leis, teoria, regularidade, código, paradigma. Não há uma explicação científica para tais escolhas. Isto demandaria muito espaço num livro que pretende ser prático.

A finalidade dos modelos é permitir o entendimento de um conjunto de eventos, poder comunicar a outros, poder reproduzir este comportamento. Os modelos são construídos a partir de experiências passadas, de registros de casos que já aconteceram, com suas características descritas (o que, quando, onde, por que, com quem e como aconteceram - os 5W e 1H). Sem registros históricos não há como identificar padrões e daí montar modelos.

Modelo não inclui tudo, é uma representação da realidade, de parte dela, para um fim especifico. O processo de BI então procura por modelos que possam explicar os acontecimentos passados ou atuais. Estamos interessados nas características deste modelo e em como ele pode relacionar os eventos entre si. Isto tudo para evitar ou eliminar as causas de problemas ou para que possamos repetir as causas de boas práticas.

O modelo permite completar um cenário. A partir de dados que temos como fatos, encaixados no modelo, podemos saber de outros dados que não temos (a chamada inferência). Se um evento aconteceu dentro de um modelo, podemos completar os dados que nos faltam sobre este evento. Por exemplo, usando modelos matemáticos e físicos aplicados a dados observados no ambiente, peritos podem saber a velocidade a que um carro estava no momento de um acidente. Os modelos também nos ajudarão a encontrar causas para os eventos, como será discutido adiante neste livro.

Mas os modelos estão também associados ao futuro. Eles nos servem para direcionar nossas decisões e ações. Por isto, usamos muitas vezes o termo "modelo de predição", porque usando modelos podemos "prever" o futuro (ou tentar, pelo menos). Modelos

são utilizados para previsão do tempo, para previsão de colheitas, de níveis de vendas, de quebra de máquinas, possibilidade de voto numa eleição (como discutido em Moraes, 2012), possibilidade de um cliente fechar uma venda, possibilidade de ocorrer um sinistro (em empresas de seguro), de um cliente pagar ou não um empréstimo, e para outros tantos fins como veremos neste livro.

2.2 Contextualização dos Modelos e Comparações

Os modelos então servem para entendimento de alguns aspectos da realidade (talvez a maioria, mas não todos), para predizer algumas situações (não todas, nem algumas poucas com total acurácia), para que possamos diferenciar contextos (gerais ou específicos), para que nossa vida não seja uma total escuridão e nossos caminhos possam ser trilhados com um mínimo de planejamento.

Todo modelo é uma especialização ou generalização da realidade, e toda especialização é uma abstração, ou seja, só absorve parte da realidade. Um protótipo de automóvel de tamanho reduzido terá apenas o design exterior do produto final, mas servirá para avaliar a aerodinâmica do projeto. Já um protótipo em tamanho real deste carro servirá para *crash testes* mas não terá os acessórios interiores, nem a pintura final. Por outro lado, uma classificação étnica é uma generalização, pois tenta encaixar todas as pessoas em algum grupo existente ou pré-definido. Portanto, o modelo deve ser estudado dentro do seu contexto específico.

Entretanto, entender os limites (escopo) do contexto não é uma tarefa fácil. Se temos um modelo que prediz o quanto um cliente com perfil Y irá gastar no Dia das Mães, temos que entender que este modelo de predição só serve para os parâmetros definidos no modelo (sexo, renda, idade, estado civil, etc. do cliente). Se algum outro atributo fora deste contexto (por exemplo, altura do cliente) puder influenciar os resultados, a predição dada pelo modelo conterá uma margem de erro. E se houver vários destes atributos, a margem de erro aumenta.

O conhecimento das informações ou dos dados isolados é insuficiente (Morin, 2000, p.36). Precisamos sempre estar fazendo comparações. É assim que o ser humano pensa. O preço das coisas é determinado pela relatividade, em relação ao preço de outras coisas e em relação ao que outras pessoas estão pagando.

Como discutiremos neste livro, encontrar as causas que levam um produto a ser mais vendido que outros exige também comparar tais causas com problemas que impedem a venda de outros produtos. Descobrir que um modelo explica por que uma máquina estraga mais frequentemente que outra, inevitavelmente nos leva a pensar em que boas práticas são utilizadas nas máquinas que não estragam tão facilmente.

O ritmo com que produtos são vendidos, os seus períodos de baixa, de alta e de normalidade é naturalmente uma comparação. Da mesma forma, encontrar o que é comum no comportamento dos melhores alunos é uma comparação, assim como identificar um aluno raro, com alto desempenho, só é possível por comparação.

A granularidade da comparação é relativa a cada objetivo. O BI pode preocupar-se em comparar vendas dentro de uma cidade ou no país todo. E mesmo uma empresa que não

faça vendas no exterior, pode querer comparar seu desempenho com empresa similares em outros países.

Portanto, um modelo deve obrigatoriamente permitir comparações. Eventos ou valores absolutos não dizem nada. E um modelo só funciona no contexto onde foi identificado. Se conseguirmos extrair de um caso real um modelo matemático que simule e explique como duas populações de espécies diferentes (por exemplo, lobos e ovelhas) irão se comportar, tal modelo só irá funcionar no contexto em que foi identificado. Se foi numa ilha, que tipo de ilha e com que recursos. Qual o número inicial de cada população e quais as características de cada componente dos grupos. O modelo não irá funcionar se colocarmos os mesmos grupos junto com outros.

2.3 Padrões X Exceções: imprecisão dos modelos

Os modelos podem não ser precisos. É preciso avaliar quando, onde, como e por que os modelos acertam ou erram. Para tanto, devem ser feitos experimentos controlados. Fazendo novas observações, poderemos verificar se elas se encaixam no modelo. Se sim, confirmam o modelo. Se não, exigem algum refinamento do modelo ou mesmo a desistência dele. Entretanto, é impossível fazer todos os testes necessários ou coletar ou observar todos os eventos necessários. Hans Reichenbach comenta o caso de avaliar remédios utilizando placebo; isto pode durar muito tempo ou não ser possível de ser realizado pela dificuldade em encontrar cobaias.

Karl Popper propôs o falseamento para comprovação de modelos e teorias. A ideia consiste em procurar um caso (exemplo) que não se encaixa no modelo ou padrão. Se não for possível encontrar tal caso, a teoria poderia ser dita verdadeira. Se não for possível procurar por um tal caso, a teoria não poderá ser provada. Por isto é que Popper (1980) diz que Astrologia e Numerologia explicam tudo.

É claro que o falseamento depende do modo como esta busca foi feita. Nunca será possível dizer com total certeza que todos os casos foram testados ou que não existe um caso tido como exceção. A verdade sempre será que não foi encontrado nenhum caso pelo modo como a busca foi feita.

A predição de eventos futuros pode ajudar a validar um modelo. Se um modelo puder ser utilizado para prever o que vai acontecer, e tais acontecimentos realmente se realizarem, então a teoria pode estar certa. As dificuldades incluem determinar que casos selecionar para testes e em que situações, quantas predições fazer, que margem de erro considerar aceitável. Além disto, há o problema de saber exatamente quais fatores influenciam. Em muitos casos, os eventos previstos podem ocorrer mas por coincidência, influenciados por outros fatores (ruídos). Nate Silver (2013) comenta diversos casos assim, muitos deles ligados ao baseball nos Estados Unidos. Por exemplo, ele recomenda não usar vitórias ou derrotas para avaliar um jogador, porque tais índices são afetados por outros desempenhos. Jogadores são responsáveis por suas estatísticas, mas também depende de quem está do outro lado jogando contra.

Apesar de invalidar um modelo, a descoberta de exceções pode ser benéfica porque gera mais conhecimento sobre o contexto, podendo vir a aprimorar modelos existentes ou

gerar um novo modelo mais moderno e preciso. Kuhn (2011) discute os paradigmas científicos e conclui que a existência de exceções pressupõe o surgimento de um novo paradigma (leia-se, modelo).

Uma vez que só capturam parte da realidade, os modelos nunca são completos. Esta falta de dados, entretanto, não impede o uso e os benefícios dos modelos. Se não puder ajudar a prever eventos futuros com exatidão, pelo menos nos darão tendências (em cima de probabilidades). E é assim que as pessoas tomam decisões. Nenhuma empresa deixa de produzir um produto só porque as vendas caíram durante os últimos dias. Se houver uma tendência de retomada das vendas, ou seja, se houver uma certa probabilidade (aceita por alguma razão) de que as vendas irão subir novamente, a empresa segue no mesmo caminho.

Por outro lado, reduzir sistemas complexos a sistemas simples ou a um conjunto de regras simples pode ser perigoso. Podemos estar fazendo suposições erradas, julgando com critérios errados em cima de fatos não observados ou mal interpretados. E o resultado pode ser desastroso, como na caça às bruxas e na ascensão de ditadores. Daqui a pouco vamos estar categorizando e estereotipando todas as pessoas, como fazemos com grupos musicais e criações artísticas. O perigo do rótulo é não conseguir sair dele. Esta é uma tendência perigosa do ser humano: criar um modelo ou teoria e sair procurando casos que confirmem a sua validade, tentando encaixar tudo no modelo. E se algo não se encaixar ? Forçaremos o encaixe ou mudaremos nosso modelo ?

Tem sido assim ao longo da História da Humanidade e da Ciência. Teorias surgem mas após anos elas podem ser refinadas ou mesmo refutadas, devido a novas descobertas, novos fatos ou novas formas de interpretar os velhos fatos. No início, o ser humano achava que todos os planetas e o sol giravam ao redor da Terra. Depois, descobriu-se a teoria heliocêntrica. Agora já há um pesquisador dizendo que a Terra é o centro do Universo. E está sendo tachado de louco, como já foram classificados Kepler e Galileu. Então o que existe é um modelo mais aceito pela maioria das pessoas (estudiosos, cientistas, ou mesmo pessoas comuns). A Teoria de Newton, que fazia isto (apesar de tantos acertos) caiu. A de Einstein, sucessora da de Newton, também possui lacunas. As leis da macro Física não se ajustam aos comportamentos no mundo micro, os quais são regidos por outras leis. Mas ambas são aceitas até que uma nova teoria consiga reunir e acomodar casos de ambos os grupos, provavelmente a partir de novas regras.

2.4 Analisar passado para criar modelos

Analisar o histórico de dados é geralmente a alternativa utilizada para montar um modelo ou teoria. Se pudermos encontrar eventos que aconteceram repetidamente no passado, é possível que se repitam no futuro. E assim teremos um padrão ou modelo.

A primeira alternativa para encontrar causas, de problemas ou boas práticas, é procurar por algo que também aconteceu quando estes eventos ocorreram. Se várias vezes uma máquina quebrou, procuramos observar o que ocorreu junto com estas quebras. Se tivermos todas as quebras registradas ou descritas, procuraremos por eventos comuns. Se tivermos um grupo de campeões num esporte, procuraremos saber o que eles fazem ou fizeram de comum. Se sabemos quais são os produtos que mais vendem, queremos

saber que características são comuns a todos. Se quisermos diminuir gastos com peças defeituosas, vamos procurar pelas causas mais frequentes. Se estamos precisando diminuir desperdícios de energia, vamos olhar para os casos mais frequentes. A nossa busca então é por repetições.

E aí é que entra a estatística, nos permitindo separar repetições interessantes das que não são significativas. Mas como os modelos não são perfeitos, precisam ser aperfeiçoados ou corrigidos. Isto pode ser feito por retroalimentação, aprendendo por experiência, por exemplo, com os erros cometidos e pela medição da incerteza (Stewart, 2000).

2.5 Modelos para prever futuro

Por que um computador ganha de humanos no jogo de xadrez ? Porque consegue reunir conhecimento de vários enxadristas (através da análise de jogos anteriores) e porque consegue realizar simulações e previsões de consequências de jogadas possíveis em situações atuais, ou seja, consegue avaliar o que vai acontecer caso uma determinada peça seja movida para uma determinada casa. Mesmo que o computador não consiga fazer todas as combinações possíveis, conforme teoria de Simon (1972), ainda sim poderá avaliar melhores jogadas do que um ser humano. Para tanto, os computadores são programados com modelos heurísticos e não algorítmicos. E usando probabilidades, conseguem avaliar qual a melhor alternativa. Pode ser então que um modelo não consiga prever o futuro com exatidão, mas permitirá avaliar quais eventos mais provavelmente poderão ocorrer.

Nate Silver (2013) comenta que o verbo "prever" em português possui duas versões em inglês: *predict* e *forecast*. Ele comenta que hoje elas são usadas como sinônimas, mas na época de Shakespeare tinham significados diferentes: *predict* era aquilo que faziam os adivinhos; *forecast*, por outro lado, implicava em planejar em condições de incerteza. Os modelos discutidos neste livro pretendem fazer previsões do segundo tipo, baseados em dados. Seria como tentar predizer valores para atributos ou acontecimentos de eventos a partir da análise de causas (valores de outros atributos).

A previsão com modelos já é uma realidade nas mais diversas áreas de conhecimento humano. Gorr (1999) discute a perspectiva de analisar dados históricos para entender estratégias e tentar prever concentrações de futuros crimes. Maltz e Klosak-Mullany (2000) utilizaram a técnica de sequência de tempo (um tipo de Data Mining) para encontrar padrões estatísticos no comportamento de jovens delinquentes nos EUA e antever eventos ruins em suas vidas, para intervir antes que aconteçam. Bill Gates, numa palestra recente, sugeriu utilizar tais tecnologias de predição na educação (http://www.technologyreview.com.br/read_article.aspx?id=43501). A ideia seria analisar dados sobre desempenho e comportamento de alunos, para entender por que um aluno pode estar faltando às aulas, e com isto tomar ações para diminuir taxas de abandono. Além disto, podemos pensar em modelos que permitam entender causas de desempenho de alunos, para evitar problemas de baixo rendimento ou replicar as boas práticas dos melhores alunos.

Tendências futuras também podem ser inferidas de ações ou comportamentos coletivos. Estudos sobre Sabedoria das Massas ou Multidões (Wisdom of Crowds) analisam o que

a maioria das pessoas está fazendo, e assim poder prever resultados ou entender o que está acontecendo. Por exemplo, o Google Trends é usado para monitorar epidemias nos EUA. Quando há muitas pesquisas no Google, vindas de uma mesma região, por palavras-chave relacionadas a uma determinada doença, isto significa que o número de casos desta doença está aumentando nesta região. Há um experimento do Google (http://www.google.org/flutrends/br/#BR) para monitorar casos de gripe. O artigo de Dugas et al. também trata do mesmo assunto.

A análise de redes sociais virou uma maneira fácil de observar as multidões. Um artigo de 2011 (Bollen et al.), conseguiu provar a correlação entre o tipo de humor nas postagens do twitter e o índice Dow Jones da bolsa de valores americana. Outros artigos provaram ser possível prever receitas de filmes, aumento no número de turismo e mesmo prever eventos futuros analisando postagens ou buscas (Asur et al. 2010; Mishne, 2006; Radinsky & Horvitz, 2013; Choi & Varian, 2012). Spink e colegas (2001) analisam o comportamento de multidões em mecanismos de busca para realizar diversas inferências.

Sargut e McGrath (2011) sugerem a gestores estabelecer um modelo que agregue três tipos de informação preditiva:

- informações passadas: dados sobre o que já aconteceu, incluindo indicadores financeiros e de desempenho;
- informações presentes: alternativas de caminhos, ações, estratégias, oportunidades ou decisões que podem ser tomados;
- informações futuras: o que pode acontecer como consequência das alternativas, incluindo respostas do meio-ambiente ou mudanças internas.

O modelo deve integrar estes 3 tipos de informações. Geralmente, são usados modelos matemáticos, ou seja, é preciso reduzir as informações para valores quantitativos (nominais, categóricos ou numéricos) e a forma de interligação entre as variáveis são fórmulas matemáticas.

O fato é que as novas técnicas estão permitindo predizer com maior precisão alguns valores e ainda verificar a interligação entre eventos ou variáveis. Desta forma, é possível saber se uma determinada ação vai impactar positivamente ou negativamente em algum contexto futuro. E quanto irá impactar. Por exemplo, se aumentarmos a exposição do produto em X dias na mídia convencional, quanto teremos de aumento de vendas e, com base nos custos desta estratégia, o quanto teremos de retorno financeiro (ou lucro).

Se tivéssemos como prever o futuro, poderíamos evitar problemas futuros (como no filme *Minority Report*, dirigido por Steven Spielberg e estrelado por Tom Cruise). Ou poderíamos indicar melhores alternativas ou mesmo saber se uma certa alternativa daria certo ou não. Mas isto não existe. Nenhuma decisão é certa. Ninguém tem como saber se uma escolha vai funcionar ou não.

Mas nem por isto (porque vivemos na incerteza) vamos tomar decisões sem critérios. Justamente, as técnicas, os padrões, os dados, nos ajudam a diminuir a incerteza e com isto melhorar nossas decisões e consequentemente seus resultados. Há alguns autores que são contra as técnicas de planejamento, porque acreditam que não vale a pena planejar, pois o futuro nunca acontece como planejado. Entretanto, se não planejamos,

se não tomamos decisões e ações, temos grandes chances de chegar onde estamos agora ou pior, chegar em algum lugar que não queremos.

É claro que os planos e caminhos, e digamos os modelos e padrões, no contexto deste livro, devem ser ajustados com retroalimentação durante a jornada. Mas uma viagem sem planos tem mais probabilidade de dar errado ou chegar num destino não desejado. O modelo utilizado por Maltz e Klosak-Mullany (2000) para prever comportamento de jovens delinquentes justamente permite que ações sejam tomadas para modificar um futuro muito provável e ruim na vida daqueles jovens. Resumindo os modelos permitem entender o passado e o presente, para que tomemos melhores decisões para um futuro melhor.

2.6 Análise de Correlação e Causa-Efeito

Como já dissemos antes, e vamos estressar muito neste livro, BI é um processo que busca encontrar causas (para problemas ou para bons resultados). Portanto, BI é um processo de investigação e descoberta, com algumas semelhanças com o processo criativo, como discutiremos mais adiante.

Pessoas e empresas querem tomar melhores decisões, para alcançar melhores resultados ou poder direcionar seu futuro. Entender quais condições geram quais resultados é uma das formas de fazer este tipo de planejamento. Entretanto, como discutiremos neste livro, encontrar causas não é tão simples quanto parece. Traremos estudos de áreas tais como investigação criminal, diagnóstico médico, previsão do tempo, ecologia, biologia, mecânica, física, engenharia, ciências sociais, economia, política, etc.

BI é análise de dados. E isto ocorre em diversas disciplinas, não sendo restrito ao meio computacional ou empresarial. O problema é comum a diversas áreas e talvez analogias possam ser utilizadas, para aplicarmos soluções que já deram bons resultados, mesmo que em outras áreas. Várias ciências ou áreas estão sempre à procura de modelos que possam explicar fenômenos e que possam ajudar as pessoas a preverem acontecimentos.

Então, BI também inclui como objetivo descobrir as relações causais, mesmo que estas envolvam diversas variáveis e diversos tipos de relações, inclusive indiretas em vários níveis. Para uma empresa é importante avaliar a correlação entre suas ações e os resultados. Por exemplo, uma empresa descobriu que um aumento de 5 pontos na atitude comportamental dos empregados implicava em 1,3 ponto de incremento na satisfação dos clientes, e isto fazia aumentar em 0,5% o faturamento da empresa. Tal descoberta permite à empresa avaliar onde investir e o quanto. Neste exemplo, se ela quiser aumentar 1% das vendas talvez tenha que aumentar 10 pontos na atitude dos colaboradores.

2.7 Dificuldades para identificar padrões - pessoas e sistemas complexos

O problema de prever eventos futuros é que o futuro é feito *COM* pessoas. A maioria dos modelos incluem pessoas. Se precisamos saber a causa por que máquinas quebram, temos que lembrar que elas são operadas por pessoas, pessoas fazem sua manutenção, pessoas as programam. Se quisermos prever índices de vendas, temos que lembrar que são as pessoas que compram produtos e serviços, e há também vendedores, promotores, publicitários, especialistas em moda e por fim administradores determinando preços.

Apesar de todos os autômatos embutidos em sistemas computacionais, das regras e procedimentos de qualidade, das interfaces planejadas para guiar o usuário, ainda sim as pessoas agem de forma não planejada, não prevista. E o "*ser humano é um ser racional e irracional, capaz de medida e desmedida; sujeito de afetividade intensa e instável*" (Morin, 2000, p.60)

Não temos como prever o que as pessoas farão em qualquer situação. Os genes condicionam vários comportamentos dos seres humanos (Winston, 2006; Dawkins, 2007) e muitas vezes agimos por instintos bastante primitivos, enraizados em nós nos tempos das savanas (Winston, 2006). Mas as pessoas também são condicionadas ou influenciadas pelo meio que as cerca, podem receber treinamento para fazerem algo dentro de certos procedimentos e há ainda as várias possibilidades do erro humano. E por fim, ainda há o livre arbítrio: os genes nos moldam como roteiristas de filmes mas o resultado final é nós que decidimos, porque os genes nos dão modelos de decisão e não a decisão final (Winston, 2006; Dawkins, 2007).

Nate Silver (2013) conta o caso dos modelos utilizados pelos cientistas políticos prevendo a vitória esmagadora de Al Gore na eleição presidencial de 2000 nos Estados Unidos. Mas quem ganhou as eleições foi George W. Bush, e um dos motivos foi a cédula de votação, com marcadores mal associados aos nomes, confundindo eleitores que iriam votar em Al Gore.

A raiz do problema está em que estamos tratando com sistemas complexos. Sistemas complicados são aqueles compostos por muitas partes, mas para os quais podemos prever o resultado final, se cada parte funcionar de forma planejada. Se conhecermos os dados de entrada, as condições ambientes e o sistema funcionar segundo o padrão conhecido, ou seja, um contexto estável, é certo que saberemos o resultado final (Sargu and McGrath, 2011). Um exemplo de sistema complicado é um carro: um mecânico conhece as partes, suas interações e consegue prever o funcionamento. Se algum problema ocorrer, ele poderá determinar a causa usando seus conhecimentos e coletando alguns dados diagnósticos.

Por outro lado, sistemas complexos podem até ter poucas partes mas as interações entre as partes podem causar funções ou resultados imprevisíveis. As partes interagem de forma inesperada e por isto seu comportamento passado não pode ser usado para antecipar eventos futuros com acurácia (Sargu and McGrath, 2011). Sistemas complexos contêm interações dinâmicas e portanto as mesmas condições de entrada podem levar a resultados diferentes em tempos diferentes. Há 3 características que

determinam um sistema complexo: multiplicidade (relativa ao número de elementos ou partes do sistema), interdependência (o nível de conexões entre as partes) e a diversidade (heterogeneidade dos elementos). Conforme Sagu e McGrath, quanto maior o nível de cada característica, mais complexo será o sistema. Um exemplo de sistema complexo foi a campanha (ou guerra) contra pardais na China em 1958. Os pardais estavam atacando as plantações de arroz e então o Governo chinês fez uma campanha para dizimar os pardais. O problema é que, após a eliminação dos pardais, os gafanhotos é que começaram a comer grãos de arroz, porque os pardais eram predadores naturais dos gafanhotos.

As loucuras que acontecem nos mercados econômicos e nas bolsas de valores também são resultados dos comportamentos complexos das multidões. Muitas vezes não há uma explicação lógica para a correria de venda ou compra nos mercados. Simples boatos podem se difundir rapidamente e levantar medo na população, gerando comportamentos ilógicos de indivíduos e levando as massas para direções inesperadas.

Entender o comportamento de multidões é um desafio. Conforme a teoria de Herbert Simon (1972), o ser humano toma decisões sob uma Racionalidade Limitada à informação disponível, à capacidade cognitiva das mentes e ao tempo disponível. Na maioria das vezes não vale a pena (pelo custo ou tempo) coletar todas as informações necessárias para tomar uma decisão. Por exemplo, se uma pessoa quiser comprar um sapato, pensará em verificar na cidade qual a loja com o preço mais barato. Entretanto, se for avaliar o preço de cada loja, ao terminar o processo, terá levado tanto tempo que os primeiros preços consultados já poderão ter sido alterados e o custo total de deslocamentos e perda de tempo não valerá o desconto que conseguir. É impossível que o indivíduo conheça todas as alternativas para uma decisão e que possa avaliar todas as suas consequências. A tendência do ser humano é simplificar as escolhas. Isto quer dizer que não temos como saber se a decisão tomada foi a mais acertada antes de tomá-la; somente após saberemos se deu certo ou não. E mesmo tendo alcançado êxito, talvez não tenhamos certeza se foi a melhor alternativa.

Em geral então, as pessoas procuram diminuir a incerteza das decisões mas assumem certos riscos pela racionalidade limitada. Por exemplo, se alguém quiser traçar uma rota de fuga em caso de incêndio num prédio, talvez não consiga avaliar todas as alternativas possíveis (local de início do fogo, quantidade de pessoas, etc.). E no momento da situação de incêndio, o ser humano tem que simplificar ao máximo seu processo de decisão para acelerar as ações. Isto quer dizer que os planos iniciais podem ter sido esquecidos ou terão que ser simplificados. E assim, as atitudes planejadas mudam pela racionalidade limitada. E o ser humano se torna imprevisível. Tversky e Kahneman (1974, 1983) discutem o problema de avaliações probabilísticas erradas em decisões humanas. Eles apresentam diversos experimentos que comprovam que o ser humano avalia de forma errada muitas situações, usando modelos probabilísticos errados ou incompletos.

Além disto, a ação de uma pessoa acaba por influenciar a decisão dos que estão próximos. Isto pode modificar o comportamento dos outros, que podem imitar ou fazer algo bem diferente. Por vezes, algumas decisões de pessoas pensando no benefício próprio e único podem prejudicar ainda mais o sistema. Há o famoso caso do paradoxo de Braess, que diz que criar atalhos em redes rodoviárias pode não diminuir o tempo médio, porque a maioria das pessoas irá escolher o atalho, gerando novos

engarrafamentos. Tomar decisões de forma independente, talvez não seja a melhor alternativa, conforme a teoria do Equilíbrio de John Nash. Talvez a melhor alternativa para todos seja cada um "perder" um pouco de algo para todos "ganharem".

As técnicas relativas à Teoria dos Jogos ajudam a entender os resultados nestes tipos de sistemas complexos. A Teoria dos Jogos é uma tentativa de tentar prever resultados em sistemas complexos. Através da análise da combinação de diferentes estratégias dos jogadores (componentes do sistema que possuem poder de decisão), pode-se prever os resultados possíveis. A dificuldade está em prever as decisões que serão tomadas.

Apesar das dificuldades, das incertezas, mesmo assim precisamos procurar padrões para entender a complexidade dos sistemas. Isto nos ajudará em situações futuras, mesmo que não nos permitindo chegar a previsões exatas. Ghani e Simmons (2004), por exemplo, conseguiram prever com 96% de acerto, o preço final em leilões no eBay, um tipo de situação bastante complexa, envolvendo diversas variáveis e além disto intuições, sentimentos, percepções e escolhas humanas.

3 Processo Geral de BI

BI tem a ver com descobrir conhecimento, para poder gerar inteligência e resolver problemas, como discutido no capítulo anterior. O objetivo final então é poder gerar conhecimento novo e útil.

Vários autores descrevem o processo geral de descoberta de conhecimento como o descrito na Figura 2. A entrada do processo é um banco de dados e a saída um conjunto de conhecimentos. A etapa principal é a de Mineração ou Análise dos Dados (Data Mining). A análise nunca é feita sobre todos os dados e sim sobre amostras. Para tanto, é necessário ter antes uma etapa de preparação dos dados, a partir da base de entrada. Nesta etapa, os dados serão tratados (limpeza, integração, deduplicidade) e amostras diferentes serão geradas, como será discutido adiante.

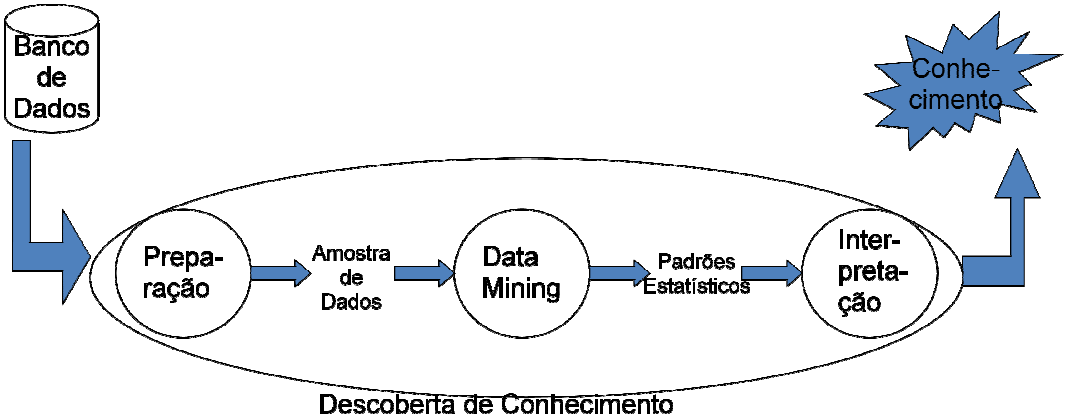


Figura 2: Processo Geral de Descoberta de Conhecimento

A etapa de análise tem como entrada uma amostra dos dados e gera como saída padrões estatísticos. Mas estes padrões não são conhecimento. Eles precisam ser interpretados dentro do contexto da organização ou do domínio, para aí sim se tornarem conhecimento. Por exemplo, uma análise de pacientes com diabetes descobriu que 95% dos pacientes com diabetes do tipo 1 recebiam o medicamento X. Isto, para um médico especialista da área, não é novidade nenhuma, pois é o tratamento usual dado a pacientes deste tipo. O conhecimento interessante e novo está nas exceções, nos 5% que tem o mesmo tipo de diabetes mas não recebe o mesmo medicamento. Pode ser que tenham alguma outra característica que os impede de tomar tal medicação.

O processo de descoberta de conhecimento é iterativo e interativo. **Iterativo** (ou cíclico) porque precisa ser feito várias vezes, com diferentes amostras ou até mesmo com diferentes técnicas e ferramentas. Os padrões estatísticos são, na maioria das vezes, hipóteses de causas, devendo ser melhor avaliados. Isto muito em razão da Teoria do Mundo Fechado, que será discutida mais adiante. O processo também é **interativo**, porque precisa intervenção humana. Para realizar a preparação dos dados e depois a interpretação dos resultados, pessoas com conhecimento sobre o domínio precisam

colocar seu intelecto a serviço da descoberta de conhecimento. Ainda não conseguimos colocar este tipo de conhecimento ou inteligência em máquinas.

3.1 Premissas do Processo de BI

Para que o processo de BI tenha um resultado satisfatório e de qualidade, algumas premissas devem ser observadas.

- Objetivo do BI

Como trataremos mais adiante neste livro, o processo de BI pode ser feito de forma reativa ou proativa. Mas em ambos os casos há um objetivo. No primeiro tipo (BI reativo), o objetivo é bem definido e busca identificar ou monitorar indicadores quantitativos. Já no segundo caso, o objetivo é mais vago e tem mais a ver com uma exploração (estamos procurando algo mas não sabemos bem o que é, nem se vamos encontrar). Este "algo" que se procura no modo proativo pode ser simplesmente "algo novo", sem definição, forma ou qualidades.

- Coletar as informações certas

Coletar os dados que realmente influenciam os objetivos é crucial para que o processo de BI atinjas os objetivos. Quanto mais informações, menos incertezas. Entretanto, só quantidade não é suficiente. Precisamos também de dados com qualidade.
Falaremos da etapa de coleta num capítulo só sobre isto e sobre qualificação de dados quando tratarmos de ETL.

- Formato certo das informações

Depois de coletados os dados, é importante colocá-los no formato adequado para análise. Dados numéricos são mais fáceis de serem analisados estatisticamente. Mas também podemos tratar informações não-estruturadas com técnicas como text mining. Se vamos predizer o total de espectadores de um filme e só temos informações qualitativas como diretor, estúdio, atores, produtores, gênero, resumo da história, local de gravação, etc., seria interessante primeiro transformar tais informações para um formato que permita aplicar técnicas de análise quantitativa para podermos relacionar tais informações com um dado estruturado e numérico como o total de espectadores ou valores monetários arrecadados.

- Qualidade das informações

Como os americanos falam "*garbage in, garbage out*". Ou seja, se o processo for feito com dados sem qualidade, o resultado será compatível, isto é, também sem qualidade. Em alguns pontos deste livro discutiremos técnicas para tratamento de dados e para enriquecimento. Mas há tantas outras técnicas para avaliação da qualidade de dados que fogem ao escopo deste livro.

- Organizar as informações

Como discutiremos neste livro, a separação dos dados em amostras é um passo importante para o processo de BI. Isto permite analisar os resultados e interpretá-los à luz da amostra. Se estamos analisando dados históricos dos 2 últimos anos, os resultados se referem a esta amostra. Se formos utilizar os padrões encontrados neste

histórico para nos preparar para o futuro (ou tentar prever o futuro), poderemos ter surpresas bem desagradáveis.
Além disto, a separação em amostrar permite comparar os padrões encontrados nas amostras. Separando dados por dias da semana, talvez possamos descobrir um padrão diferente para cada dia da semana.

- Técnicas e métodos de análise

Utilizar a técnica correta é fundamental. Por isto, discutiremos neste livro várias técnicas de análise e alguns cuidados na interpretação dos resultados.

- Recuperação e disseminação do conhecimento

O processo de BI só se completa quando o conhecimento descoberto chega até as pessoas que precisam dele, no formato correto e no tempo exato. Se o processo demorar demais, se o resultado chegar num formato não adequado, o processo de decisão (razão da existência das informações) será comprometido.

3.2 Quem deve participar do Processo de BI

Hoje há um cargo conhecido como Analista de BI. Este conhece principalmente as ferramentas de software utilizadas para a análise dos dados e apresentação dos resultados em dashboards.

Entretanto, deve haver um Analista de Negócios, que possa interpretar os resultados no contexto da organização. Este também deverá propor objetivos para o BI, como por exemplo a análise de certos indicadores de desempenho (KPIs), pois fará a ponte entre os problemas e objetivos da organização e as técnicas e ferramentas de BI e TI. O Analista de Negócios também deverá auxiliar na preparação dos dados, indicando que tipo de amostrar poderá ser utilizada e que atributos ou valores são mais importantes para serem analisados. Na falta de um profissional deste cargo, o Analista de BI deverá assumir tal responsabilidade, e portanto deverá procurar conhecer a organização, seus problemas e objetivos. E a participação de gestores, administradores, executivos ou tomadores de decisão também é importante, pois são os clientes das informações a serem geradas pelo BI.

Por fim, seria interessante contar com um cientista social ou estatístico, que pudesse ajudar na geração de amostras e na seleção das técnicas estatísticas a serem utilizadas.

3.3 Processo de BI Pró-ativo X Reativo: começar com ou sem hipóteses

De acordo com Choudhury e Sampler (1997), existem dois modos para aquisição de informação: o modo reativo e o modo proativo. No primeiro caso, a informação é adquirida para resolver um problema específico do usuário (uma necessidade resultante de um estado anômalo de conhecimento). Nestes casos, o usuário sabe o que quer e poderá identificar a solução para o problema quando há encontrar.

Por outro lado, no modo proativo, o propósito de adquirir informação é exploratório, para detectar problemas potenciais ou oportunidades. Neste segundo caso, o usuário não tem um objetivo específico.

Oard e Marchionini (1996) classificam as necessidades de informação em estáveis ou dinâmicas e em específicas ou abrangentes (gerais). Taylor, citado por Oard e Marchionini (1996), define 4 tipos de necessidades, os quais formam uma escala crescente para a solução do problema:

- necessidades viscerais: quando existe uma necessidade ou interesse, mas esta não é percebida de forma consciente;
- necessidades conscientes: quando o usuário percebe sua necessidade e sabe do que precisa;
- necessidades formalizadas: quando o usuário expressa sua necessidade de alguma forma;
- necessidades comprometidas: quando a necessidade é representada no sistema.

As necessidades tratadas pela abordagem de descoberta reativa poderiam ser classificadas como estáveis e específicas, segundo a classificação de Oard e Marchioninni, e como conscientes (no mínimo), segundo Taylor. Isto porque o usuário sabe o que quer, mesmo que não consiga formalizar.

Exemplos de objetivos que caracterizam um processo reativo são:

- encontrar características comuns nos produtos mais vendidos;
- encontrar motivos que levam à evasão ou a reclamações de clientes;
- achar perfis de grupos de clientes;
- encontrar clientes potenciais para propaganda seletiva;
- encontrar concorrentes no mercado.

No modo reativo, o usuário tem uma ideia, mesmo que vaga, do que pode ser a solução ou, pelo menos, de onde se pode encontrá-la. Pode-se dizer então que o usuário possui algumas hipóteses iniciais, que ajudarão a direcionar o processo de descoberta. Neste caso, é necessário algum tipo de pré-processamento, por exemplo para selecionar atributos (colunas em uma tabela) ou valores de atributos (células). Isto exige entender o interesse ou objetivo do usuário para limitar o espaço de busca (na entrada) ou filtrar os resultados (na saída). É o caso típico de quando se deseja encontrar uma informação específica, por exemplo, um valor para um atributo ou um processo (conjunto de passos) para resolver um problema.

Já as necessidades da abordagem proativa poderiam ser classificadas como dinâmicas e abrangentes, segundo a classificação de Oard e Marchioninni. São dinâmicas porque podem mudar durante o processo, já que o objetivo não está bem claro, e são abrangentes porque o usuário não sabe exatamente o que está procurando. Pela taxonomia de Taylor, as necessidades do modo proativo são viscerais. Isto quer dizer que há uma necessidade ou objetivo, mas o usuário não consegue definir o que precisa para resolver o problema. A necessidade típica do modo proativo poderia ser representada pela expressão: "*diga-me o que há de interessante nesta coleção de dados*". Neste caso, o usuário não tem de forma definida o que lhe seja de interesse (o que precisa), podendo tal interesse mudar durante o processo. Pode-se dizer que é um processo exploratório, sendo, em geral, iterativo (com retroalimentação) e interativo (com ativa participação e intervenção do usuário).

Na abordagem proativa, não há hipóteses iniciais ou elas são muito vagas. O usuário deverá descobrir hipóteses para a solução do seu problema e explorá-las, investigá-las e testá-las durante o processo. Em geral, acontece porque o usuário não sabe exatamente o que está procurando. É o caso típico de quando se quer monitorar alguma situação ou encontrar algo de interessante que possa levar a investigações posteriores. Depois que hipóteses são levantadas, o processo pode seguir como no paradigma reativo, talvez sendo necessário avaliar as hipóteses, para verificar se são verdadeiras ou não.

Pode-se dizer que a abordagem proativa é mais difícil de ser conduzida e até mesmo pode não levar a descobertas interessantes. A princípio, deve-se sempre procurar iniciar com indicadores bem definidos, ou seja, usando uma abordagem reativa. A pró-atividade é útil quando os indicadores já foram esgotados ou quando se quer descobrir algo realmente novo e inesperado. Muitas empresas utilizam abordagens para Gestão da Inovação baseadas em descobertas por acidente ou acaso (o que os americanos chamam de *serendipity*), e este "pulo do gato" pode fazer a grande diferença em mercados competitivos. Mas isto é papo para outro capítulo.

4 Pré-processamento e Preparação de dados

Esta etapa também é conhecida pelo termo ETL (extração, transformação e carga/load) ou *cleansing* (limpeza).
O objetivo é melhorar a qualidade dos dados e gerar uma base separada para análise (um data warehouse) para não sobrecarregar as bases usadas pelos sistemas transacionais.
A limpeza serve para eliminar inconsistências da base, completar dados, tratar valores nulos, eliminar registros duplicados, etc. (por exemplo, uma pessoa com dois telefones diferentes ou com um endereço incompleto ou faltando).

O Data Mining na verdade veio de processos de correção de integridade em bases de dados. Por exemplo, num hospital, os procedimentos de cesariana só podem ser feitos em pacientes do sexo feminino. Então, eram criadas regras de integridade e um software automaticamente verificava a probabilidade da regra. Neste caso, 100% dos procedimentos de cesariana deveriam ter sido feitos em mulheres. Se o resultado não fosse 100%, algum registro estava inconsistente.

A grande ideia foi construir um software que identificasse regras automaticamente (sem que operadores humanos precisam definir as regras) e avaliasse a probabilidade. Quando os criadores viram que regras novas e interessantes eram descobertas, aí nasceu a Mineração de Dados como é conhecida hoje.

A seguir serão discutidas algumas técnicas desta etapa.

4.1 Tratamento de valores nulos

O que fazer se acontecer de pegarmos para analisar uma base de dados onde 50% dos registros não possuem dados para um determinado atributo (por exemplo, campo sexo). Isto pode gerar resultados não confiáveis. Por exemplo, se uma análise estatística gerar um padrão dizendo que 80% dos registros possuem valor "masculino" para este campo. Como não sabemos o que acontece com os outros 50% dos registros, é possível que todos eles sejam do mesmo sexo e com isto a regra descoberta estaria completamente distante da realidade.

Uma possibilidade é desconsiderar os valores nulos e interpretar os padrões descobertos dentro deste contexto, como uma tendência. Se os registros com valores nulos são apenas 10% do total, a margem de erro nas regras descobertas será também de 10%.

Outra possibilidade é gerar dados por aproximação. Por exemplo, na mineração de uma base com dados climáticos da região sul do estado do Rio Grande do Sul, havia muitos dados faltantes. O que se fez foi completar os dados faltantes com os dados de estações próximas, uma vez que a variação dos valores de uma estação de coleta para outra não é muito grande.

A média e a interpolação também podem ser utilizadas, mas isto pode gerar distorções drásticas nos resultados se os valores faltantes justamente destoavam da maioria (eram *outliers*). Se o conjunto de registros compunha uma minoria, os resultados finais terão um desvio muito pequeno.

Outra possibilidade é utilizar regras de classificação coletadas fora da empresa. Por exemplo, se não tivermos a renda de um cliente, podemos usar dados estatísticos sobre a renda da cidade onde ele mora. Se não tivermos o estado civil, podemos supor se ele é casado ou solteiro analisando outros dados referentes a esta pessoa.

4.2 Deduplicidade de registros

A eliminação de registros duplicados evita contar duas vezes uma entidade. Além disto, pode resolver problemas com dados conflitantes (ex.: cliente com dois endereços). Há técnicas que avaliam probabilidades para saber qual o valor mais correto.

A identificação de registros duplicados pode ficar mais fácil se houver uma identificação única. Hoje em dia, não só CPF e RG são usados como identificadores, mas também endereços, números de celular, e-mail e *logins* em redes sociais.

Entretanto, há muitos casos em que isto não é feito por alguma razão histórica (mal planejamento, por exemplo) ou quando duas bases são unidas por aquisição de empresas diferentes. Imagine o caso em que o identificador utilizado é o nome de uma pessoa. É muito provável que o nome de uma mesma pessoa seja escrito de formas diferentes em oportunidades diferentes. Um operador humano pode registrar o nome completo, outro pode abreviar algum nome intermediário ou mesmo o dono do nome pode não querer dizer todos os seus sobrenomes. O uso de atributos complementares pode ajudar a encontrar registros duplicados desta pessoa. Também pode-se utilizar técnicas de avaliação de similaridade entre vetores, como a medida de similaridade de Pearson usada em sistemas de Raciocínio Baseado em Casos (RBC ou CBR).

4.3 Integração de bases (*merge*)

O melhor seria ter padronização de todos os campos. Se isto não for possível, devemos usar técnicas como as discutidas anteriormente para deduplicar registros.

Em muitos casos, é imprescindível integrar duas ou mais bases, como no caso de uma empresa que adquire outra e quer unificar as duas bases. Em outros casos, a integração pode ser feita para gerar enriquecimento dos dados. Por exemplo, integrar a base de dados de uma loja física com a base de uma loja na Internet.

A integração de bases pode ser feita também para aumentar a possibilidade de identificação de padrões estatísticos. Por exemplo, minerar vendas juntando o cadastro de produtos e o cadastro de clientes pode permitir identificar associações entre bairro do cliente e tipo de embalagem do produto.

Se temos uma base de pedidos de clientes residentes em cidades diferentes, podemos adicionar dados referentes às cidades. Por exemplo, o tamanho da cidade, o tipo de atividade econômica principal, se é de montanha ou é praia, a idade da cidade, o partido

do prefeito, etc. Isto pode ajudar a encontrar padrões como por exemplo o tipo de produto mais comprado para cada perfil de cidade. Podemos hipoteticamente descobrir que clientes de cidades grandes compram em maior quantidade ou que cidades litorâneas não fazem pedidos nas sextas-feiras. Até mesmo a renda média da cidade pode ser usada para completar a renda dos clientes, em caso de valores nulos.

Se formos analisar pacientes de um hospital, talvez seja interessante acrescentar informações sobre o ambiente familiar e profissional de cada paciente, seus hábitos alimentares e cotidianos, e até mesmo a história pregressa de doenças suas e de seus familiares.

Para aumentar as chances de haver padrões estatísticos, pode-se gerar novos campos a partir dos existentes. Por exemplo, um hospital possui dados de baixa e alta de pacientes que foram internados. Mas o dado mais importante para este hospital é o número de dias que o paciente ficou internado (tempo de permanência). Uma simples subtração entre datas.

É claro que isto aumenta o volume de dados, mas certamente também aumenta a probabilidade de encontramos padrões. Em geral, é utilizada uma tabela não normalizada para agilizar as análises, uma vez que não é preciso passar de uma tabela para outra através de chaves estrangeiras (códigos que relacionam registros).

4.4 Enriquecimento de dados

O enriquecimento de dados compreende acrescentar dados à base existente. Por exemplo, se tivermos dados cadastrais de clientes, podemos incorporar dados externos da empresa, por exemplo vindos de outras empresas parceiras ou mesmo de comportamentos capturados fora da empresa.

A vantagem do enriquecimento é ter mais dados para análise estatística, aumentando as chances de encontramos padrões. Por exemplo: uma base de vendas contém dados como data da venda, número de nota fiscal, os itens adquiridos, valor total pago. Se incorporarmos dados dos clientes (cidade, idade, sexo) e dados dos produtos (preço, categoria, tamanho), há mais chances de haver repetições. Além disto, pela técnica de associação, poderemos cruzar dados de produto com dados de clientes e, por exemplo, encontrar padrões entre faixa etária e faixa de preço (ex.: jovens tendem a adquirir produtos de menor valor).

Outro exemplo de enriquecimento: cada click de uma pessoa num site é monitorado. Aí estes dados são cruzados com o que a pessoa comprou pela internet. E depois estes dados são cruzados com dados dos cadastros de lojas físicas. E então a estes dados são somados dados sobre as compras que esta pessoas fez na loja física, fora da Internet. E tudo isto é complementado com dados vindos dos perfis da pessoa nas redes sociais e com o que a pessoa diz em fóruns e blogs (é o Social CRM).
E é possível pegar dados públicos, disponíveis livremente na Internet. Estes dados não identificam pessoas individualmente, mas dão estatísticas sobre grupos de pessoas. Uma empresa pode comprar uma lista telefônica com nome, endereço e telefone de clientes. Mas não sabe classificar os clientes por dados sócio-demográficos. Então, a empresa pode consultar bases públicas sobre setores censitários. Um setor censitário é diferente

de um bairro ou quadra; é uma região, geralmente menor que um bairro mas podendo abranger partes de 2 bairros, que foi pesquisada pelo censo do IBGE. Então, há informações estatísticas sobre cada setor específico. Imagine que a empresa então possui os seguintes dados sobre uma pessoa: o nome é José da Silva e mora na Rua X, n.41. Bom, usando um sistema de GIS simples, pode-se saber o setor censitário onde ela mora. Depois, procuram-se dados estatísticos sobre este setor e, digamos, temos que neste setor:

- 100% das residências possuem 3 TVs;
- 98% possuem 2 banheiros;
- 90% possuem aparelhos de DVD;
- 90% possuem TVs LCD;
- etc.

Agora, de posse destas informações estatísticas, podemos estimar alguns dados sobre José da Silva. Que ele tem 3 TVs, com 100% de chances, que há 98% de chances de ele ter 2 banheiros em casa, e assim por diante.
Então, desta forma, uma empresa combina a lista telefônica com dados censitários e poderá obter um banco de dados de clientes potenciais.
Empresas parceiras também costumam compartilhar dados sobre clientes, por exemplo, administradoras de cartões de crédito, instituições financeiras, redes de varejo, escolas, postos de gasolina, editoras, etc. E há empresas que vendem este tipo de informação (cadastros).

E a cada pesquisa que participamos, com o objetivo de concorrer a prêmios, estamos fornecendo mais dados sobre nós.

Mas não precisa ser só enriquecimento de dados sobre pessoas. Se tivermos o campo cidade em alguma base de dados, podemos incorporar dados sobre as cidades, tais como número de habitantes, geografia, economia principal, nível de escolaridade, renda per capita, índices sócio-culturais como IDH e outros.
Neste caso, pode-se cruzar a cidade do cliente com dados dos produtos adquiridos. Isto nos permitirá, por exemplo, descobrir que tipo de cidade compra mais um certo tipo de produto. Num caso real, uma empresa de comércio eletrônico descobriu que somente clientes de cidades pequenas (com menos de 50 mil habitantes) compravam produtos eletrônicos mais caros (depois descobriu-se que a razão era porque naquelas cidades não havia lojas físicas vendendo tais produtos; enquanto que em cidades maiores, o preço do produto estava muito alto em relação a um concorrente com loja física).

4.5 Seleção de Amostras

É muito difícil minerar ou analisar todos os dados de uma base. Em geral, é preciso fazer uma seleção inicial. Isto porque alguns dados, mesmo presentes por direito na base, talvez não sirvam os propósitos. Por exemplo, produtos que não são mais vendidos e não interferem mais no processo, e por consequente não são interessantes para serem analisados, devem ser excluídos.

O primeiro passo então num processo de BI é selecionar um conjunto de dados (uma amostra) sobre os quais serão aplicadas as técnicas de análise ou mineração.

A seguir, são discutidas algumas técnicas para geração de amostras.

4.5.1 Tipos de amostras

Existem 4 tipos de técnicas de seleção de amostras. Discutiremos elas através de um exemplo: uma loja querendo analisar a satisfação de seus clientes. Também discutiremos duas situações possíveis: a loja já ter um cadastro de clientes e o caso de a loja não conhecer seus clientes (porque entram e saem da loja sem mesmo a loja saber se são homens ou mulheres).

- Amostras aleatórias

Neste caso, são selecionados aleatoriamente elementos do universo (conjunto todo). Por exemplo, a loja determina o tamanho da amostra (valor N) e a seleção é feita pegando-se os N primeiros clientes da base de dados (do cadastro) ou são selecionados N elementos dentro do cadastro, pulando de forma aleatória. Se a loja não tiver um cadastro, ela irá selecionar clientes que saiam da loja com sacolas (produtos comprados), "atacando" N clientes pulando alguns (a critério da pessoa que fará a abordagem).
Este tipo de amostra pode trazer problemas, pois imagine que os N selecionados são todos do mesmo tipo (homens X mulheres, classe A ou classe C, etc). E pior ainda se forem selecionadas justamente as exceções.
Alguns pesquisadores julgam a técnica eficiente pois acreditam na distribuição aleatória (aquela velha história da moedinha, se jogarmos uma moeda 1000 vezes e só analisarmos os 100 primeiros resultados, a distribuição será a mesma).
Entretanto, para que a técnica seja utilizada adequadamente, a aleatoriedade deve ser total. No caso de clientes saindo da loja, não se pode selecionar clientes apenas num dia. Deve-se levar em conta as variedades (dia da semana, dia do mês, mês, turno, etc).
Esta técnica só deve ser usada quando não se pode utilizar uma técnica melhor.

- Amostras por conveniência

Neste caso, a seleção é feita pelo que for mais fácil. Por exemplo, a loja seleciona os N primeiros clientes que saírem da loja num determinado dia ou liga para N clientes cadastrados que tiverem telefone e só utiliza dados dos N primeiros que atenderem o telefone.
É a pior técnica pois não há critério algum, nem mesmo a aleatoriedade, o que pode levar a tendências (selecionar somente elementos de um tipo).
Esta técnica só deve ser usada quando não se pode utilizar uma técnica melhor.

- Amostras por julgamento

As amostras por julgamento são formadas por elementos que satisfaçam regras previamente determinadas. Por exemplo, analisar somente a satisfação de clientes mulheres que compraram mais de um produto até uma semana após o Dia das Mães.
Neste caso, o critério de seleção está bem definido e é justificado (por exemplo, só querer analisar certos tipos de elementos do conjunto todo). E portanto os resultados da análise serão condizentes somente com as regras definidas (não valem para o universo todo).
Podem ser utilizada regras de seleção ou de exclusão. O segundo caso pode ser melhor para se ter uma visão melhor do todo. Por exemplo, a loja pode querer analisar todos os

tipos de clientes, mas vai excluir quem só veio uma vez por ano ou quem comprou num valor muito abaixo da média de gasto.

- Amostras estratificadas

Esta é a forma correta de gerar amostras. Para tanto, precisa-se identificar que variáveis podem interferir na análise. Por exemplo, no caso da loja, atributos como sexo, idade, classe sócio-econômica, bairro e cidade, valor gasto e forma de pagamento podem fazer diferença para entender os tipos de clientes. E talvez altura, peso e escolaridade não sejam diferenciais para campanhas de marketing ou para entender comportamentos de compra.
Depois de identificadas as variáveis, precisa-se saber a proporção de elementos no universo todo para cada variável. Por exemplo, digamos que há 60% de mulheres e 40% de homens entre todos os clientes da loja, e que 25% são da classe A, 50% da classe B e 25% da classe C, e assim por diante nas demais variáveis.
Então, a amostra será definida com a mesma proporção que a divisão no universo. Ou seja, a amostra deve conter 60% de mulheres, 40% de homens, 25% de pessoas da classe A, 50% de pessoas da classe B, 25% da classe C e assim por diante.

4.5.2 Como separar amostras (subcoleções ou subconjuntos)

Para agilizar o processo de análise, pode-se separar subconjuntos dos dados. Além de tornar o processo mais rápido, evita também a descoberta de padrões com suporte muito baixo.

A formação de subconjuntos pode ser feita por corte vertical ou horizontal. O corte vertical significa selecionar alguns atributos para análise, eliminando outros. O corte horizontal trata de selecionar alguns registros, eliminando outros.

O corte vertical (*feature selection*) será discutido mais adiante. Para o corte horizontal, as amostras podem ser definidas por tempo (ano a ano, mês a mês, etc) ou por algum outro atributo que permita separar os dados com significado e não aleatoriamente. Pode-se pegar um atributo específico e fazer a separação (ex: sexo, tipo de cliente, produto ou tipo de produto). Por exemplo, separar uma base de clientes em homens X mulheres, separar para análise somente produtos de um certo setor ou faixa de preço, classificar empresas por porte e analisar em separado cada grupo.

Ou então separar um conjunto de dados por outros dados relacionados. Por exemplo, pode-se comparar as vendas feitas por homens X vendas feitas por mulheres, compras de adultos X jovens X 3a idade, vendas separadas por tipo de produto ou por loja ou por região, etc.

Mas qual o melhor campo para separar em subcoleções ? Utilizar apenas um campo ou uma combinação de vários campos (amostra estratificada) para separar a coleção toda em subconjuntos ? A escolha deve ser feita por humanos ou automaticamente, como na técnica de *clustering* ?

Bom, não vi ainda uma regra que dê estas respostas. Normalmente, é um processo de tentativa e erro, utilizando feeling do analista, pela sua experiência.

Uma constatação, entretanto, é que campos com predomínio de valores não são bons para separação. Por exemplo, num hospital é possível que mais de 90% dos pacientes sejam atendidos pelo SUS. Então não adiante separar os pacientes em "particulares" e "SUS". Até porque o subgrupo do "SUS" deve ser muito pequeno e não irá gerar um número mínimo de elementos para se ter significância estatística (discutida adiante).

O que este tipo de campo nos diz é que podemos sim separar um subgrupo para análise, mas seria o da maioria. Isto é, eliminar registros com valores minoritários. Por exemplo, se estamos analisando uma base de clientes, e há apenas 0,1% dos clientes que moram numa determinada cidade, não vale a pena minerar estes registros quando queremos analisar padrões pela cidade.

A lição é que devemos analisar diversas amostras e comparar os padrões encontrados em cada uma. Assim, pode-se descobrir que um padrão aparece numa amostra e não aparece noutra (ex.: o caso acima citado do produto X), ou que um padrão aparece com uma probabilidade numa amostra (ex.: 80% dos clientes do bairro K utilizam serviço Z) e com outra probabilidade em outra amostra (ex.: apenas 40% dos clientes do bairro L utilizam o serviço Z).

Exemplos de como separar amostras:

- Numa base de vendas ou pedidos, pode-se separar por período de tempo, por exemplo, uma amostra para cada ano ou mês. Isto permitirá (como será discutido adiante), comparar os padrões encontrados em cada amostra.
- Também é possível separar por dia da semana, mas neste caso é preciso juntar dados do mesmo dia, ou seja, se tivermos dados de vários meses, agrupar os dados por dia da semana. Assim, o grupo da 2a-feira terá dados de todos os meses mas somente da 2a-feira.
- Outra forma de fazer a separação é por tipo de cliente. Se a empresa já trabalha com clientes segmentados, por exemplo, por plano de serviços ou pessoa jurídica X pessoa física, pode-se criar uma amostra para cada tipo de cliente. A amostra deve conter não somente os dados demográficos dos clientes (nome, endereço, sexo, etc.), mas também dados comportamentais (compras feitas, hábitos, preferências, ações). Isto permitirá comparar os clientes entre si, para realizar ações focadas.
- Pode-se também segmentar a base de dados por características geográficas, por exemplo, país, região, estado, cidade, bairro ou setor censitário. E isto vale para vendas, clientes, pedidos ou até mesmo para origem de produtos.
- Se quisermos separar amostras por produto, podemos utilizar categorias de produtos, faixas de preços, composição (ex. plástico X metal), tamanho, tipo de pacote ou embalagem.

4.5.3 Generalizações e Especializações

Em muitos casos, podemos encontrar atributos que são hierarquias de tipos. Por exemplo, o caso de cidade e estado cai nesta situação. Temos uma hierarquia entre os seguintes atributos: país → estado → cidade → bairro.

Se analisarmos juntos todos os atributos que formam uma hierarquia, muitos padrões descobertos irão mostrar estas relações. E isto não é interessante porque já sabemos

destas relações. A solução é utilizar um dos atributos de cada vez, em cada ciclo de análise.

Agora note que, se usarmos o atributo mais genérico (neste exemplo, o país), a probabilidade de encontramos padrões é maior, pois há menos valores possíveis para este atributo. Entretanto, pode haver predomínio de um ou dois valores, e como já comentamos antes isto também não é bom.

Se usarmos o atributo mais específico (no exemplo, bairro), pode ser que não haja repetições e o suporte das regras encontradas seja muito baixo (ou mesmo não encontremos padrões).

A navegação por uma hierarquia dá nome às operações de drill-down e drill-up (ou roll-down e roll-up), seja para analisar os dados com mais detalhes ou para se ter uma visão mais superficial dos dados.

Bom, o que foi dito acima vale também para outros tipos de hierarquias como datas (ano → mês → dia), classificações de produtos (tipo do produto "brick" → marca → embalagem), pedidos e vendas (carrinho → item do carrinho), etc.

4.5.4 Amostras por período de tempo - analisar ritmo

Normalmente não se costuma analisar todos os dados disponíveis, por causa do enorme volume ou por limitações das ferramentas. Mas também porque é perigoso trabalhar com o conjunto todo de dados. Por exemplo, uma loja analisou 10 anos de vendas e descobriu um padrão: 90% das mulheres com perfil A compravam o produto X. Ao analisarem amostras ano a ano, descobriram que a probabilidade do padrão era de 100% nos 9 primeiros anos (ou seja, todas as mulheres do perfil A compraram o produto X nos 9 primeiros anos). Mas no último ano, nenhuma das mulheres do perfil A comprou o produto X.

Geralmente, dividimos as amostras por tempo utilizando alguma unidade como ano, semestre, mês, dia da semana, hora, turno, etc. Aqui a dica é a tentativa e erro e a geração de diversas amostras para comparação. Pode-se começar com uma granularidade intermediária (por exemplo, mês) e depois aumentar ou diminuir a granularidade, utilizando uma unidade menor ou maior. O feeling de um especialista do domínio pode ajudar a determinar as melhores unidades para análise, mas também pode influenciar o processo e acabar deixando fora amostras interessantes (o tal de "achômetro").

É claro que a seleção da unidade de tempo também passa pelo conhecimento do domínio. Se não interessa saber qual o turno em que os eventos ocorrem, se manhã, tarde, noite ou madrugada) ou se já se sabe de antemão que não há diferença de comportamento no início, meio ou fim de mês, então podemos eliminar a separação de amostras por estas unidades de tempo. As unidades menos utilizadas são o dia do mês e a quinzena. Então, o melhor é trabalhar com dia da semana ou semanas (1a semana do mês é diferente da última, e ambas são diferentes das duas semanas intermediárias).

Muitas vezes, a granularidade alta (unidade menor, como por exemplo a hora) pode dificultar a interpretação dos resultados. O que significa um padrão de vendas que ocorre às 9 horas todos os dias, mas não ocorre às 8 horas nem às 10 horas ? Que estratégias devem ser usadas para aquela hora específica e que não valem a pena ser usadas uma hora antes ou depois ? E também trabalhar com hora e minuto pode gerar padrões muito específicos, que até podem ser interessantes mas como traçar estratégias de ações para um minuto específico ?

Outra dificuldade é a seleção de dados por estações climáticas. Não há como analisar as vendas feitas no inverno. Porque não sabemos exatamente quando o inverno começa e termina. Não podemos usar as datas tradicionais, porque muitas vezes o frio começa antes, ou só chega bem depois, ou a estação é mais curta ou mais extensa. Neste caso, o melhor seria associar a temperatura como uma atributo. Entretanto, deve-se cuidar que alguns eventos só são desencadeados um certo tempo após seu estímulo. Por exemplo, propagandas na TV não geram vendas no mesmo dia, nem talvez no dia seguinte. Isto quer dizer que se uma onde de calor acontecer no meio do inverno, não significa que as pessoas vão correr para as lojas para comprar roupas de verão. Em alguns casos, a reação é quase imediata: se a temperatura sobe, as vendas de sorvete sobem quase que instantaneamente.

Mais adiante discutiremos a correlação entre variáveis com comportamentos semelhantes mas em períodos de tempo diferentes (correlação assíncrona).

O importante na análise temporal é entender o comportamento do gráfico correspondente, incluindo subidas, descidas, platôs, e as características destes tipos de acidentes (altura ou profundidade, a frequência com que ocorrem, o comprimento do período). Também é interessante analisar padrões que podem ser encontrados nas sequências: por exemplo, sempre depois de um platô e uma pequena queda, ocorre uma subida ao dobro do platô.

Não devemos também negligenciar padrões que ocorrem com frequências maiores que meses. Por exemplo, para uma revenda de carros pode ser interessante descobrir que um cliente troca de carro a cada 3 ou 4 anos. As lojas de varejo já descobriram que nos anos de Copa do Mundo (a cada 4 anos então), as vendas de TVs aumentam muito.

Os registros feitos ao longo do tempo formam uma série temporal. Como discutiremos adiante, a técnica de mineração mais apropriada é a de análise de séries temporais.

4.5.5 Tamanho da amostra - quantidade de elementos na amostra

Como determinar o número ideal de elementos numa amostra ? Se olharmos para as pesquisas para presidente do Brasil, a amostra normalmente é composta por aproximadamente 2 mil pessoas. Isto quer dizer que cada pessoa representa em torno de 50 mil outras.
O cálculo estatístico do tamanho da amostra depende do erro amostral (a diferença entre o valor estimado pela pesquisa e o verdadeiro valor e isto pode ser um valor estabelecido como meta); do nível de confiança (a probabilidade de que o erro amostral efetivo seja menor do que o erro amostral admitido pela pesquisa); da população (número de elementos existentes no universo da pesquisa, valor que pode não ser

conhecido); entre outros (percentuais máximo e mínimo). Há uma calculadora online para fazer tais cálculos: http://www.calculoamostral.vai.la/

Tversky e Kahneman (1971) discutem os problemas com amostras muito pequenas. Por exemplo, se você jogar uma moeda não viciada três vezes e der duas vezes cara e uma vez coroa, você estará inclinado a acreditar que a probabilidade é 66,66% contra 33,33%. Mas se jogar mil vezes a mesma moeda, certamente haverá uma proporção próxima de 50/50. Pior seria se nas três primeiras jogadas, desse somente um lado. Como sabemos que, no caso da moeda, a probabilidade é 50/50, isto pode gerar a chamada "falácia do jogador": acreditar que o jogo vai mudar para reverter uma tendência e voltar ao padrão estatístico. Por exemplo, jogando 5 vezes a mesma moeda e dando sempre o mesmo lado (digamos, cara), vamos acreditar que na 6a vez irá dar o outro lado (coroa). E na 7a também vamos estar inclinados que dará coroa para equilibrar o jogo e voltar à proporção 50/50. Entretanto, a proporção só acontece com amostrar maiores. Então, as próximas jogadas só minimizam os desvios e não os corrigem logo em seguida.

4.6 Seleção de atributos ou campos para análise - *feature selection*

Como dito antes, nem sempre é interessante analisar todos os atributos disponíveis. Para tanto, precisa-se selecionar alguns e eliminar outros. Esta separação pode ser feita por *benchmarking* ou analogia, ou seja, utilizando o que normalmente se analisa tais como vendas, perfil de clientes, etc.

Mesmo assim, a base ainda pode conter muitos atributos e isto pode gerar milhares de padrões estatísticos. O ruim é que não se consegue interpretar todos estes padrões, por serem muitos e isto ser uma tarefa intelectual. Então é necessário ainda eliminar alguns atributos.

Uma dica é evitar atributos com valores que não se repetem, como por exemplo identificadores e campos chave como CPF, RG, CNPJ e códigos criados para relacionar tabelas. Se estes atributos forem utilizados como chaves estrangeiras, aí talvez se consiga algum padrão. Por exemplo, o código de clientes pode ser utilizado em vendas para se descobrir algo específico sobre um determinado cliente. Mas para casos em que se queira um padrão mais genérico, estes atributos não servem.

As datas devem ser "quebradas" em dia da semana, dia do mês, mês e ano, senão dificilmente se repetirão. E se isto acontecer, de que adianta conhecer um padrão que aconteceu numa data específica ?

Quando há campos calculados (ex: total), isto também pode gerar muitos padrões. No caso de associações, certamente aparecerão diversas regras relacionando o campo calculado com seus parâmetros. Um exemplo: imagine uma base de vendas com um campo sendo o total da venda e outro sendo o imposto. Sabe-se que o imposto é calculado pelo valor total da venda. Assim, é possível que sejam identificados diverso padrões do tipo: SE imposto = X e outro_atributo = Y ENTÃO total_da_venda = Z.
Note que neste exemplo, várias regras aparecerão alterando somente os atributos relacionados na parte do SE (outro_atributo). Para resolver tal problema, basta utilizar somente um dos campos (origem ou calculado) de cada vez.

4.6.1 Valores que predominam

Outra dica é evitar campos com valores dominantes. Por exemplo, se numa base de dados sobre clientes, 98% dos registros são de homens (ou seja, 98% dos clientes são homens), não vale a pena minerar o campo "sexo", pois ele estará presente em diversas regras de associação do tipo SE atributo = X ENTÃO sexo = "M".

Também é possível que apareçam regras do tipo SE atributo_1 = X e sexo = "M" ENTÃO atributo_2 = Y.

Neste exemplo, mesmo que apareçam regras com o sexo = "F", provavelmente o suporte será muito baixo, já que somente 2% dos registros têm este valor.

Outro caso é de entidades ou registros que predominam. Por exemplo, suponha uma base de pedidos onde 90% pedidos sejam de uma empresa X, e que a cidade desta empresa seja Y. Então é possível que sejam encontrados diversos padrões com o código desta empresa. Se juntarmos os pedidos e os dados das empresas que fazem pedidos (clientes), vão aparecer muito mais padrões com a cidade Y.

4.6.2 Dependências funcionais

Uma dependência funcional acontece quando um atributo tem seu valor determinado pela presença de outro (seria uma probabilidade condicional de 100%). Por exemplo, se numa base aparecer a cidade = "Porto Alegre", o estado será "RS" sempre (em 100% dos casos). Então o atributo "cidade" determina o valor do atributo "estado" (e "estado" depende de "cidade").

De maneira formal, temos a seguinte definição:
Um atributo Y é dependente funcionalmente do atributo X, se, para cada valor do atributo X, existe exatamente um único valor do atributo Y. A dependência funcional é representada por $X \rightarrow Y$ (ou seja, X determina o valor de Y). Por exemplo, o atributo "nome" é dependente funcionalmente do atributo "CPF", pois o valor do CPF determina o nome da pessoa. O inverso não ocorre porque pode haver duas pessoas com o mesmo nome (CPF → NOME).

Em termos de análise de dados, as dependências funcionais tendem a gerar diversos padrões associativos que não são novos. O problema é a quantidade. E isto pode gerar sobrecarga na hora da interpretação dos resultados.

Para evitar tal problema, basta utilizar um dos atributos por vez, em cada ciclo de análise. Ou seja, utilizar somente um dos campos da dependência. Eliminar os campos-chave, como códigos e identificadores, também minimiza o problema.

4.7 Discretização - faixas ou grupos de valores

Quando há valores numéricos (contínuos ou discretos), pode ocorrer de não encontramos repetições. Por exemplo, uma base de clientes com o atributo "idade", uma base de vendas com o atributo "total da venda", uma base sobre produtos com o atributo "preço" e assim por diante.

Uma possibilidade é agrupar os valores criando faixas ou intervalos de valores. Há técnicas automáticas e mesmo software que realizam este processo sem intervenção humana, utilizando técnicas estatísticas. Este processo é chamado de "discretização".

Se uma pessoa for realizar a separação dos valores, pode incorrer em erros. Por exemplo, como separar por idade. De que idade até que idade seria o grupo dos jovens, dos adultos, das crianças e da chamada 3a idade ?

Além disto, da dificuldade em fazer tal separação, ainda há o problema de onde classificar os valores próximos dos limites dos grupos. Por exemplo, se definirmos que crianças vão até 14 anos e adolescentes começam com 15 anos, como tratar justamente quem está nestes limites (têm 14 ou 15 anos).

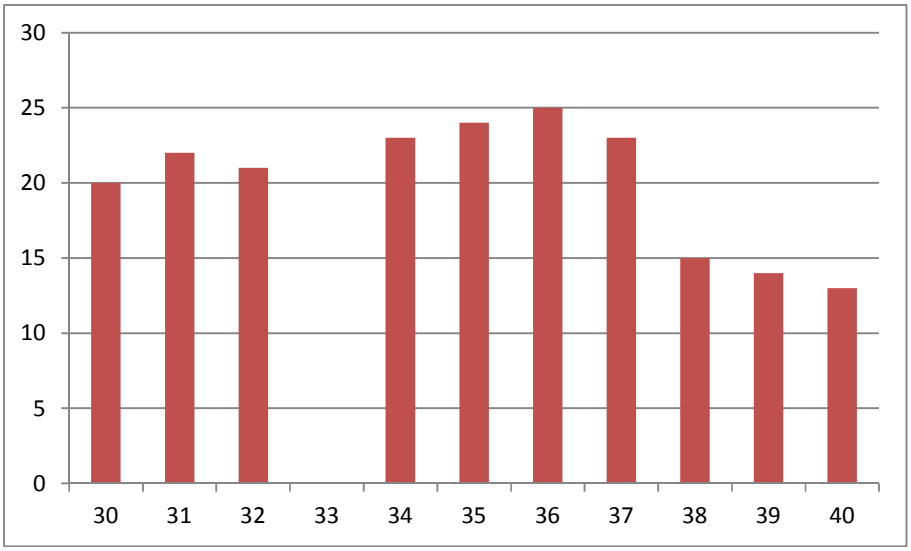


Figura 3: Gráfico para mostrar discretização de forma intuitiva

Uma saída para isto é utilizar a lógica difusa (*fuzzy*). Ela permite que um elemento seja classificado em diferentes grupos mas com graus de pertinência diferentes. Isto significa que alguém pode dizer que hoje está frio e quente ao mesmo tempo. Mas a pessoa dirá que está frio com grau 0,7 (por exemplo) e está quente com grau (0,3). Então, no caso do jovem com 14 anos, ele poderá ser classificado tanto como criança como adolescente (o mesmo com o jovem de 15 anos). Para efeitos de marketing, isto é bom, porque as campanhas não irão perder a oportunidade de atingir certas pessoas por dúvida na classificação.

Para realizar a discretização, há técnicas estatísticas e softwares que fazem isto automaticamente. Para entender intuitivamente como ocorre a discretização, vejamos a

Figura 3. Ela apresenta o número de pessoas (eixo vertical "y") que possuem uma determinada idade (eixo horizontal "x").
Alguém poderia dividir este grupo de pessoas em 2 ou 3 grupos. No caso de 2 grupos, teríamos pessoas com idade entre 30 e 32 (inclusive) num grupo e pessoas de 34 a 40 anos num segundo grupo. Se quisermos podemos dividir ainda o segundo grupo em 2, um com idade entre 34 e 37 e outro com idades entre 38 e 40.

4.8 Data Warehouse

Em geral, os processos de BI devem ser feitos sobre bases de dados separadas, e não sobre a base transacional, onde ocorrem as operações do dia a dia, para não onerar servidores ou atrapalhar operações de colaboradores. Imagine uma empresa multinacional com revendas espalhadas por todo o mundo e de repente seus vendedores não conseguem realizar nenhuma venda, porque os servidores de aplicação e banco de dados estão ocupados com algum executivo realizando análises complexas de dados.

Então a solução é gerar uma base só para análise, chamada de base OLAP (*on line analytical processing*). Este é o conceito de Data Warehouse: uma base centralizada formada por dados copiados de outras bases, as chamadas bases OLTP (*on line transactional processing*). Então separamos as bases de dados e os servidores: um esquema para aplicações transacionais a nível operacional da empresa (com tarefas de inclusão, exclusão, alteração e consulta simples de registros e valores) e outro esquema com dados só para análise (dados não voláteis, onde somente haja inclusão), apoiando decisões táticas e estratégicas.

5 Técnicas de Análise de Dados

Este capítulo pretende apresentar diversas técnicas para análise de dados, incluindo um conjunto de técnicas conhecidas como Data Mining, mas também técnicas de análise de dados cúbicos (montagem de cubos) ou análise OLAP, as quais são popularmente chamadas de BI.

Análise qualitativa X quantitativa

Começamos explicando que a maioria das técnicas de análise de dados é baseada em técnicas estatísticas. E estas por sua vez, devem ser aplicadas sobre dados quantitativos ou estruturados. Dados quantitativos incluem variáveis que podem ser expressas com valores numéricos (ex.: idade, quantidade de produtos em estoque, quantidade vendida, tempo de permanência de pacientes em hospitais), valores temporais (data e hora, por exemplo) ou valores conhecidos com nominais, categóricos ou qualitativos (ex.: bairro, cidade, sexo, classe social). Estes últimos são valores selecionados de um conjunto limitado, e não incluem atributos textuais que podem ser preenchidos com texto livre (ex.: descrição de um problema ou uma solução). Os dados nominais são semelhantes a dados numéricos porque poderiam ter um correspondente numérico. Por exemplo, sexo pode ser armazenado com um número representado os diferentes tipos (e há bases de dados que trabalham com mais de 2 sexos). Satisfação de clientes poderia ser representada por escalas numéricas; bairros, cidades e países podem ser representados por códigos numéricos.

Por outro lado, há também análises qualitativas. Estas têm por objetivo encontrar as variáveis envolvidas, para depois então serem aplicadas técnicas quantitativas. Por exemplo, uma pesquisa sobre refrigerantes preferidos por uma população pode começar por uma pesquisa qualitativa, para que fossem identificados as diferentes preferências. Também pode-se fazer uma análise qualitativa para identificar possíveis motivos para cada preferência. Após então, pode-se conduzir pesquisas quantitativas para determinar quantidades (quantas pessoas preferem cada tipo e quantas vezes cada motivo foi citado). A análise qualitativa pode ser feita de forma manual ou intelectual por humanos, mas já há ferramentas de software que auxiliam tal processo. Neste caso, normalmente a análise qualitativa é feita sobre dados chamados não-estruturados, os quais incluem textos, sons e imagens (figuras, desenhos, diagramas, fotos, vídeos, etc.).

Um processo de BI normalmente aplica técnicas quantitativas sobre variáveis. Mas não deve excluir análises qualitativas. Isto envolve, por exemplo, a descoberta de quais variáveis devem ser incluídas em um modelo para análise, que eventos podem interferir nos resultados e como representar quantitativamente cada atributo. Por exemplo, a idade de pessoas pode ser representada utilizando um valor numérico absoluto (ex.: 35 anos), um valor relativo (ex.: mais jovens, mais velhos), uma faixa ou intervalo de valor (ex.: pessoas entre 15 e 20 anos) ou uma categoria ou valor nominal (ex.: crianças, jovens, adultos, terceira idade).

Qualitativo para quantitativo

"Todas as coisas são números", já dizia Pitágoras há mais de 2 mil anos atrás. Se as coisas não nascem números, nós as transformamos em números. A representação por

números começou para facilitar as comparações e depois o comércio. E isto permitiu identificar padrões, para entendermos o passado e podermos nos preparar para o futuro. Assim foi contando dias, estações e anos. Só assim entendemos os ciclos da agricultura e de morte-vida de animais, inclusive nós mesmos. Usamos números e funções matemáticas para encontrar padrões, para fazer raciocínio probabilístico e previsões, para tomar decisões com base em probabilidades. É mais fácil assim entender a natureza, os sistemas, os comportamentos e relações, e até mesmo a complexidade.

A Teoria do Caos diz que há padrão em tudo, até mesmo no nosso livre arbítrio. E tudo se reduz a funções matemáticas. A dificuldade não é nem encontrar a função que rege cada sistema, mas sim saber quais variáveis influenciam cada resultado, e depois conseguir coletar em tempo hábil cada medida. O Prêmio Nobel de Economia geralmente é dado a matemáticos, porque estes descobrem funções matemáticas para explicar comportamentos econômicos. Um destes casos é o de John Nash. No filme "Uma Mente Brilhante", a vida deste gênio é bem retratada. Em algumas passagens podemos ver como sua mente funcionava, tentando encontrar padrões matemáticos em tudo, por exemplo, pássaros voando, pessoas caminhando no campus da universidade.

A famosa série de Fibonacci foi encontrada em vários casos na natureza (sementes de girassóis, caracóis, alinhamento de planetas). A regra desta série é de que o próximo número é soma dos 2 anteriores (0, 1, 1, 2, 3, 5, 8, 13, 21, ...). Ela era utilizada na arquitetura antiga como uma forma de estética e beleza. Outra série famosa é a que virou a lei de Titius-Bode. Ela foi formulada inicialmente por Johann Titius em 1776 e depois formulada como uma expressão matemática por J. E. Bode em 1778. Esta série descreveria a distância dos planetas ao sol (com algum ajuste). Com esta série William Herschel descobriu um novo planeta além de Saturno: Urano (Losee, 2001). Também procuraram o planeta que faltava entre Marte e Júpiter e descobriram os asteroides Ceres e Pallas. É claro que os céticos falam em coincidência.

A Geometria, um ramo da matemática, também está presente na arquitetura e nas relações do corpo humano. O desenho do Homem Vitruviano de Leonardo Da Vinci apresenta as proporções do corpo humano. E Da Vinci se inspirou em Vitrivius, que acreditava que a arquitetura deveria imitar as proporções da natureza (Christianson, 2012).

Tornar o subjetivo em objetivo, o abstrato em mensurável, o incompreendido e intocável em algo simples: este é o desafio de transformar o qualitativo em quantitativo..

Os programas que "escutam" um trecho de música e nos dizem que música é, transformam música (sons) em números para poder fazer a comparação rápida. Os softwares de biometria (identificação por características físicas da pessoa) também transformam um ser humano em números. Nossas características (traços do rosto ou das impressões digitais, atributos de nossa voz ou pupilas) são transformadas para equações matemáticas para uma rápida análise.

No nosso dia a dia, também usamos simplificações deste tipo. Por exemplo, para representar a qualidade de um filme ou uma música, usamos estrelas. Quanto mais estrelas, melhor a qualidade. Mas como traduzir um conjunto de diretores, atores, temas, cenários, etc. a um único número. Já há estudos para análise automatizada de textos e

imagens. Isto certamente passa por números e fórmulas matemáticas. Só assim poderemos num futuro breve pesquisar no Youtube por um vídeo onde apareça um casal numa praia com coqueiros, ao entardecer.

O humor da humanidade já pode ser representado por uma série temporal, a partir da análise de postagens no Twitter. E já se pode entender o aumento de vendas pela análise dos gráficos do Google Trends. E isto inclui prever a bilheteria de um filme analisando blogs. Os artigos de Bollen et al. (2011), Choi e Varian (2012) e Mishne (2006) explicam o que eu estou dizendo.

Até a vida das pessoas está sendo representada em números. O método do biorritmo pretende mostrar as características físicas, emocionais, intelectuais e intuitivas em momentos no tempo. Isto permitiria considerar o momento de cada dimensão para tomar decisões. O pressuposto é que a vida de uma pessoa, representada por estas 4 dimensões, segue ciclos de períodos regulares (contados em dias). A Figura 4 abaixo mostra o meu Biorritmo no dia 18 de agosto de 2013, calculado pelo site http://www.profcardy.com/numerologia/biorritmo.php

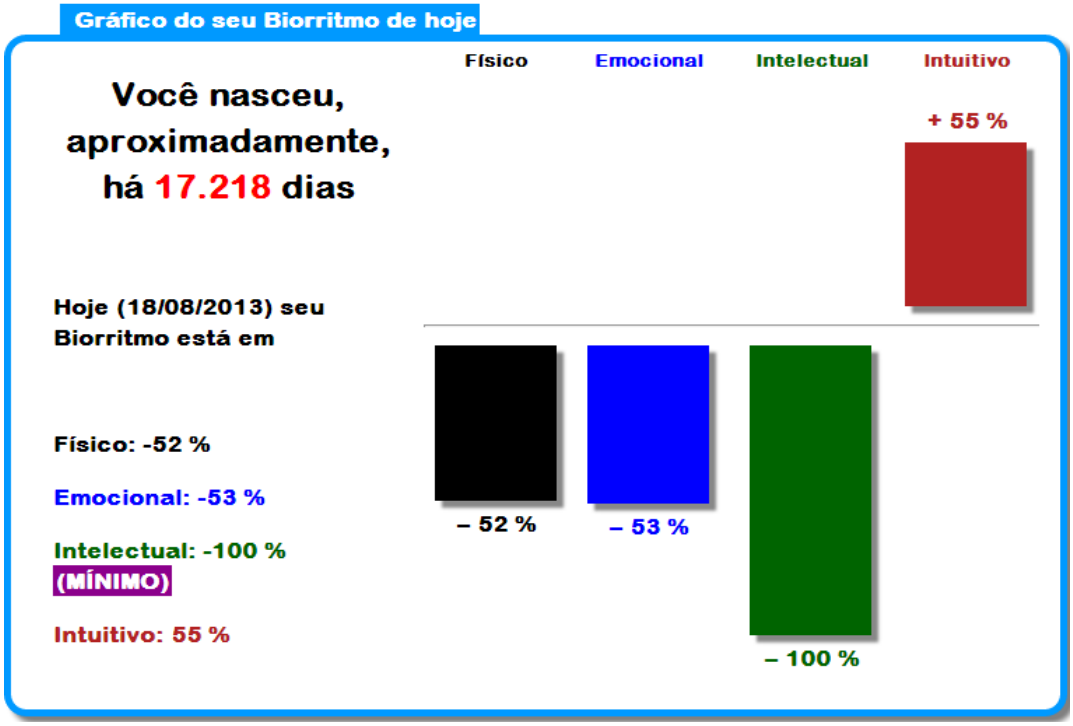


Figura 4: biorritmo num determinado dia

Já a Figura 5 mostra a previsão para os próximos 2 meses. Quando as 4 linhas estiverem lá embaixo, não vou marcar nenhum compromisso.

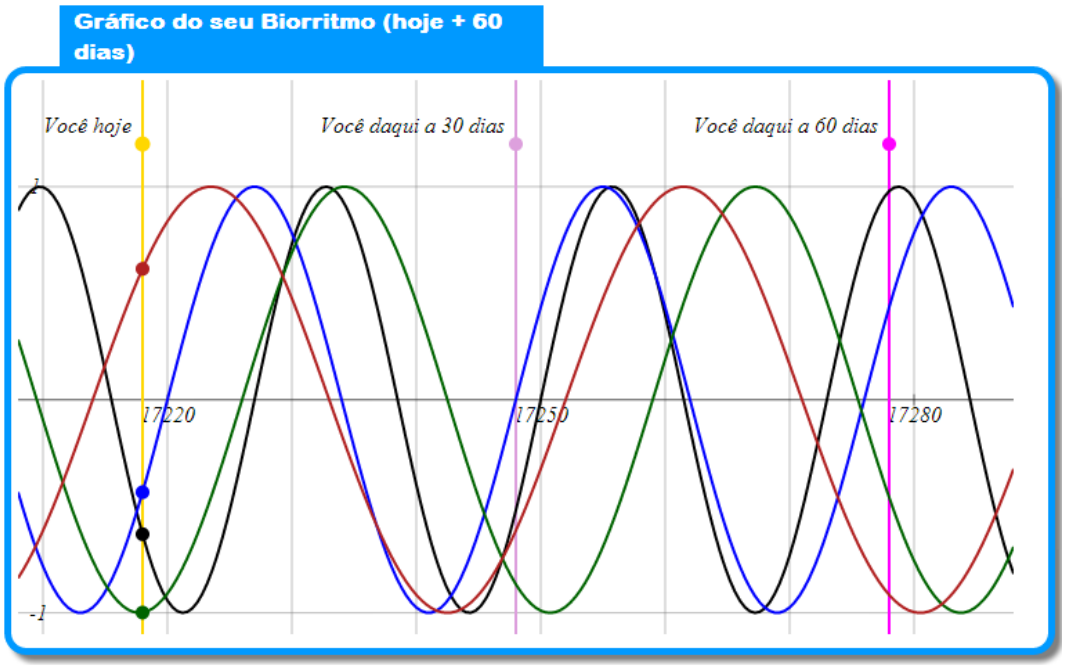


Figura 5: biorritmo para vários dias

Nem sempre os números são melhores que nossas intuições e sentimentos. Nate Silver relata que olheiros humanos tiveram melhores desempenhos que as estatísticas do sistema Pecota em vários casos no baseball (Silver, 2013). Apesar de Lewis (2004) relatar alguns casos contrários, (em Moneyball, as estatísticas foram melhores que olheiros), a conclusão é que ambos devem se ajudar. A prova disto é que a equipe de Obama mesclou dados e sentimentos das pessoas para fazer uma campanha vitoriosa (Moraes, 2012).

5.1 Data Mining - técnicas tradicionais sobre dados estruturados

Nesta seção, apresentamos as principais técnicas para Data Mining, seu funcionamento e suas aplicações.

Associação

Esta técnica é a mais famosa e ficou conhecida depois que uma rede de supermercados, ao utilizar uma ferramenta de Data Mining com esta técnica, descobriu que, nas 6as-feiras, quem comprava fraldas também comprava cerveja.

O objetivo da técnica é avaliar que valores aparecem muito juntos nas mesmas transações ou eventos (por exemplo, carrinhos de compras), mas também pode ser utilizada para identificar relações entre atributos dentro de uma mesma entidade (ex.: clientes do sexo feminino costumam morar mais no bairro X).

Para isto, a técnica é baseada na probabilidade condicional. A Figura 6 apresenta uma amostra exemplo de uma tabela num banco de dados. Nela podemos ver que há 2 campos, C1 e C2, e os valores que aparecem nas linhas (transações). Pode-se notar que os valores X e Y aparecem em comum em muitas linhas.

A probabilidade condicional resulta em implicações do tipo X ➔ Y, que são chamadas regras condicionais e podem ser lidas como "se X aparecer, então Y tem grandes chances de aparecer também". A implicação tem um grau de probabilidade ou confiança (confidence), que é calculado pela razão entre o número de registros onde X e Y aparecem juntos, dividido pelo número de registros em que X aparece (independente da presença de Y).

No exemplo da Figura 6, temos que a regra X ➔ Y possui confiança de 80%. Isto quer dizer que há 80% de chances de Y aparecer no campo C2 na mesma linha em que X estiver no campo C1. Ou olhando para o passado, Y aparece em 80% das linhas onde X aparece.

Note que a relação inversa pode possuir outro grau de confiança. No exemplo, a regra Y ➔ X tem confiança de 100%, calculada pela divisão do número de registros onde Y e X aparecem juntos pelo número de vezes em que Y aparece.

É importante também observar o suporte da regra, ou seja, o número de casos. Imagine que um supermercado descubra que 100% dos clientes que compraram o sapato de número 48 também compraram o Xampu de Abacate. Seria interessante fazer uma campanha de marketing para isto ? Se o número de casos (suporte) for muito baixo, não vale a pena.

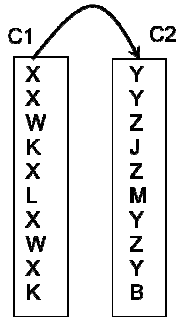


Figura 6: Associações de valores entre 2 campos para Data Mining

Os algoritmos para este tipo de técnica não são muito complicados. O que complica é que todas as combinações deverão ser avaliadas, ou seja, todos os tipos de regras. Isto quer dizer que o campo C1 será avaliado implicando em C2, C3, C4, etc. Depois C2 será avaliado contra C3, C4, etc, e assim por diante. Depois faz-se o caminho inverso.
Além disto, regras complexas, com mais de um campo na parte anterior (no "se") também serão avaliadas e aí poderemos ter regras complexas tais como "Se cliente é mulher, mora no bairro X, tem idade entre 20 e 30 anos, é solteira, tem curso superior, Então compra o produto X". A Figura 7 apresenta uma ideia de como será feita a combinação de campos. Note que a técnica avalia um campo contra outro, 2 campos contra um 3o, 3 campos contra um 4o e assim por diante, fazendo todas as combinações possíveis.

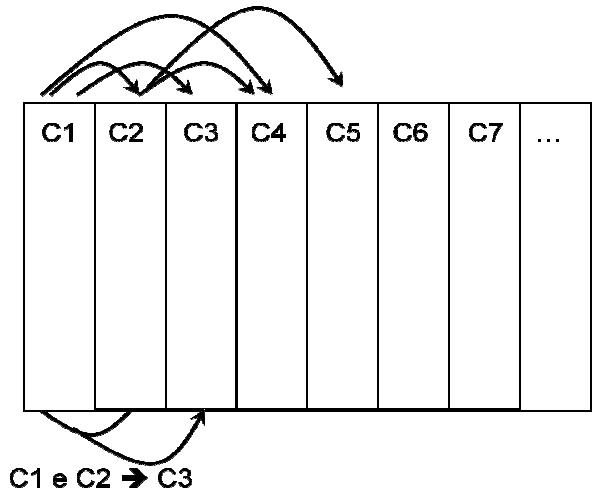


Figura 7: Comparação de valores entre campos para Data Mining

Correlação

A técnica de correlação procura avaliar a similaridade entre duas variáveis numéricas. Este tipo de análise não indica causalidade, ou seja, não diz se um atributo leva a outro, se é causa de outro (ou se um é consequência de outro). Apenas indica que há uma forte relação entre os atributos, pelos seus valores quantitativos. A análise de correção pode inclusive indicar a força da correlação.
Esta força de correção pode ser medida numa escala de 1 a -1. O valor 1 indica que as variáveis estão fortemente correlacionadas. O valor zero indica que não há nenhum relação entre elas, e o valor -1 indica uma relação inversa (quanto o valor de uma variável é alto, o valor da outra é baixo).

A Figura 8 apresenta um exemplo com diversos vetores com valores numéricos. Os vetores de V2 a V6 serão comparados com o vetor base V1, tendo as seguintes características em relação ao vetor base:

- V2: metade dos valores são iguais e outros bem diferentes;
- V3: valores muito próximos (para mais ou a menos);
- V4: valores exatamente iguais;
- V5: valores bem diferentes;
- V6: valores pela metade.

Pode-se notar que o vetor V4 tem um grau de correlação igual a 1 em relação ao vetor V1, pois todos os valores são idênticos. Já o vetor V3, com valores muito próximos, tem um a correlação em mais de 99%. O vetor V2 tem correlação de 97,4% porque metade dos valores são iguais ao vetor V1. O vetor V6 com valores pela metade tem correlação de 88,7% e por fim o vetor V5 com valores bem diferentes tem só 14,2% de correlação com o vetor V1.

Esta técnica é útil para verificar se há uma relação entre atributos quantitativos, por exemplo, temperatura e vendas. Como os valores de temperatura oscilam entre 0 e 50 e as vendas possuem valores bem diferentes, é preciso fazer uma normalização, ou seja, levando ambas as faixas de valores para o mesmo intervalo (por exemplo, entre 0 e 1).

Uma maneira de fazer isto é dividir o intervalo original por um valor base (por exemplo, temperatura dividida por 100) ou fazer a transposição proporcional de valores mínimos e valores máximos, mantendo a proporcionalidade entre os valores.

V1	V2	V3	V4	V5	V6
40	70	39	40	20	20
120	120	123	120	300	120
60	80	62	60	120	30
300	300	301	300	150	150
150	120	148	150	80	75
200	200	202	200	90	100
80	60	79	80	140	40
180	180	179	180	100	90
correlação=	0,974583	0,999773	1	0,142469	0,887595

Figura 8: Planilha de vetores e grau de correlação

Outras aplicações incluem a análise de correlação entre indicadores dentro da empresa. Eis alguns exemplos:

- número de horas de treinamento X número de falhas: note que na normalização, será preciso inverter algum vetor, pois quanto mais horas, menos falhas são esperadas;
- número de vendedores X tamanho da receita;
- aumento nas vendas X aumento no salário;
- número de promoções X aumento de clientes.

Isto é útil para se saber quais ações estão realmente impactando em objetivos. Mais adiante discutiremos a questão da causalidade, ou seja, se uma forte relação entre duas variáveis pode indicar que uma é causa da outra.

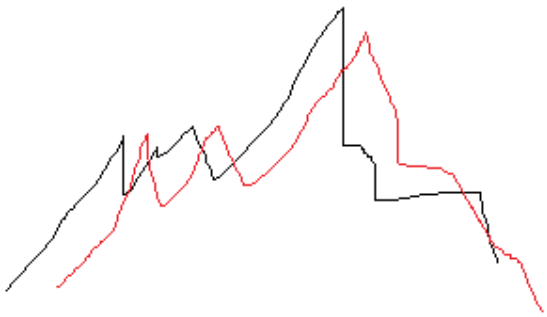


Figura 9: Gráficos semelhantes indicando correlação entre variáveis

Correlação assíncrona

Podemos ver na Figura 9, os gráficos em cor preta e vermelha são muito parecidos. Provavelmente, se usarmos a técnica de correlação iremos verificar um alto grau de similaridade entre estas duas variáveis.
Agora veja a Figura 10. Há correlação entre estes 2 gráficos ? Talvez sim, se posicionarmos eles de forma diferente, fazendo coincidir os picos.
Pode haver correlação entre duas variáveis mas utilizando como marco zero momentos diferentes no tempo. Steven D. Levitt (Freakonomics) sugere haver uma relação entre a redução de crimes verificada no Natal de 1989 nos EUA e a legalização do aborto naquele país 20 anos antes. Quando há uma relação de causa-efeito, nem sempre o efeito é imediato.

Figura 10: Correlação assíncrona entre duas variáveis

Análise de Regressão e Modelos de Predição

A Análise de Regressão é uma técnica estatística que estuda a relação entre duas ou mais variáveis, procurando elaborar um modelo para explicar o comportamento relativo destas variáveis. É útil para inferir a relação de uma variável dependente (variável de resposta) com variáveis independentes específicas (variáveis causais ou explicativas do resultado).

O modelo em questão, normalmente, é uma função matemática que relaciona as variáveis, ou seja, que permite calcular o valor da variável dependente com base nos valores das outras variáveis (causais ou explicativas).

Por exemplo, imagine que a Coca Cola tivesse uma função relacionando o dia do ano com a quantidade vendida de seu principal produto. A função iria dizer o quanto a Coca Cola iria vender num determinado dia futuro e assim ela poderia produzir somente o que espera vender (ver Figura 11). A variável dependente é a quantidade vendida do produto e a variável independente é o dia do ano (pois não depende das vendas e sim o contrário). Infelizmente a coisa não é tão simples assim, pois outros fatores influenciam a quantidade de vendas, incluindo temperatura, promoções, ações da concorrência, etc.

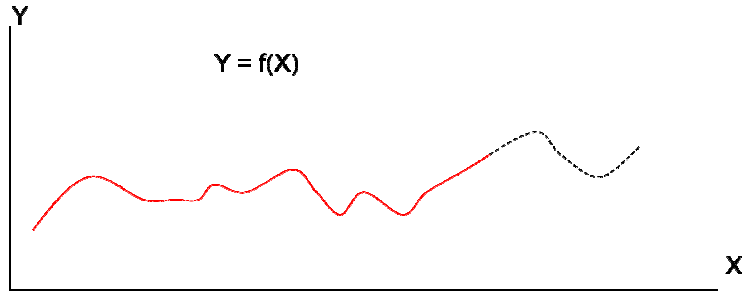


Figura 11: Técnica de Modelo de Predição

A principal vantagem de poder determinar a relação entre duas variáveis é poder realizar previsões sobre o comportamento futuro das variáveis, calculando um valor quantitativo futuro ou até mesmo podendo prever acontecimentos (eventos) que ainda não ocorreram.

Por exemplo, Thomas Morus equacionou o crescimento da população como uma função exponencial enquanto que previu o crescimento linear da produção de alimentos, chegando então à conclusão que iria faltar comida no futuro. Os serviços de meteorologia utilizam modelos matemáticos desta forma, juntando diversas variáveis para poder prever o tempo (temperatura, se vai chover ou não, o quanto vai chover, qual será a velocidade do vento, etc.).

Outra forma de aplicação dos modelos construídos desta forma é poder fazer simulações, fornecendo como entrada dados ainda não observados. Imagine que há um modelo que representa a relação entre número de vagas nas escolas, número de empregos e que tenhamos informações sobre a idade e nível de escolaridade de cada pessoa num pequeno país. E que ainda seja possível determinar a taxa de crescimento da população, vagas nas escolas e empregos. Então, usando a análise de regressão seria possível ter uma função matemática relacionando estas variáveis. Isto seria útil para prever as quantidades futuras destas variáveis, assumindo uma linearidade. E também é claro assumindo que outras variáveis não interferissem (não haverá evasão, migrações, repetições de ano, etc.). Outro benefício do modelo seria poder avaliar eventos futuros caso alguma variável tivesse alteração de comportamento. Por exemplo, e se o número de nascimentos aumentasse muito (acima do esperado), e se o número de vagas de emprego não crescesse tanto quanto esperado (acima do linear), e assim por diante.

A relação entre as variáveis pode ser funcional (por exemplo, a área de um círculo em relação à medida do raio deste círculo) ou estatística. A relação pode existir mas não necessariamente ser exata. Por exemplo, a idade das pessoas em relação à altura; são funções lineares que progridem juntas com uma certa relação até certo ponto. Mas talvez não seja possível identificar uma função matemática que, a partir de uma, seja possível calcular o valor de outra.

Outro exemplo é a relação (hipotética) inversa entre o aumento das vendas de TVs num determinado país e o índice de mortalidade infantil neste mesmo país. Até podemos encontrar uma função matemática que relacione os índices, ou seja, pode haver uma forte correlação estatística (como discutido na técnica anterior) mas certamente uma variável não é causa de outra. E este tipo de correlação é que pode desviar a análise de causa-efeito, como discutiremos mais adiante.

Os modelos de regressão podem ser:
- simples: quando uma variável depende somente de outra variável; ou
- múltiplos ou multivariados: quando uma variável depende de um conjunto de outras variáveis (o caso das vendas).

E os modelos também podem ser:
- lineares: quando a função de relação entre as variáveis é linear; ou
- não lineares: quando a função tem outra forma, como por exemplo, exponencial, logarítmica, geométrica, etc.

A Teoria do Caos diz que temos funções para descrever tudo. O filme "Uma Mente Brilhante" mostra a vida do matemático John Nash, que descobria funções matemáticas para tudo. A dificuldade da técnica de análise de regressão não é descobrir a função que relaciona as variáveis, pois isto os softwares de Data Mining podem fazer. O problema está em conseguir dados de todas as variáveis envolvidas e numa quantidade suficiente para tornar a previsão significativa em termos estatísticos. E isto inclui também em conhecer ou determinar quais variáveis influenciam o resultado (discutiremos isto mais adiante quando tratamos de descobrir hipóteses para causas).

Média

Na falta de uma função, podemos utilizar valores médios. Imagine, como na Figura 12, termos histórico de vendas em 3 anos seguidos. Podemos fazer uma função média com os valores médios dos 3 anos ou mesmo utilizar intervalos, e isto ajudaria a prever o comportamento para anos futuros.

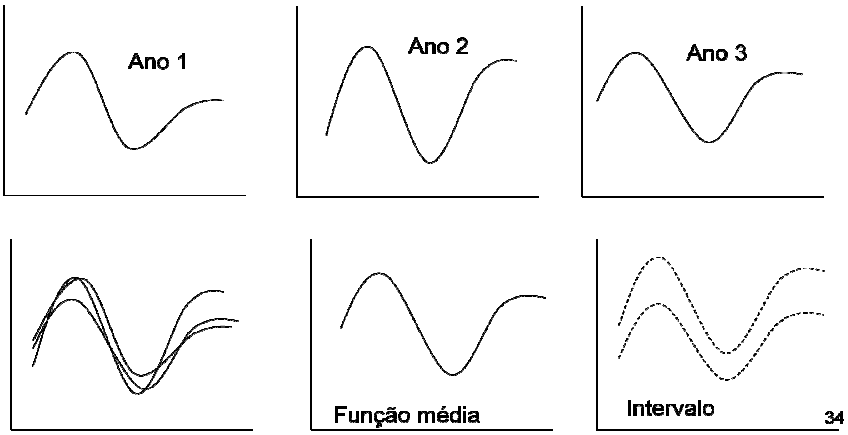


Figura 12: Técnica da Média

Detecção de desvios (outliers)

Normalmente, o ser humano tem a tendência de procurar por padrões que se repetem, ou seja, que sejam comuns ou mais frequentes. Por exemplo, quais os produtos mais vendidos, qual o tipo de cliente mais comum, qual o comportamento típico dos consumidores. Mas algumas vezes o incomum também é interessante. Por exemplo,

investigar por que somente uma pessoa comprou o produto Y no último mês, por que um vendedor não atingiu a meta (o normal seria premiar o melhor vendedor e descobrir o que os melhores fizeram de bom e em comum para que tais melhores práticas sejam repetidas).

Estas peças fora do padrão são chamadas de Outliers. Em alguns casos, eles são mais importantes que os casos normais. Por exemplo, analisando saídas de um determinado material do almoxarifado de uma empresa, tem-se uma padrão de saída (uma quantidade média ou intervalo normal), como na Figura 13. Entretanto, num determinado mês, houve muito mais saídas que o normal. Isto deveria gerar um alerta na empresa. Isto pode estar acontecendo por roubo ou pode estar indicando uma tendência que a empresa não soube prever.

A técnica de detecção de desvios utiliza funções ou intervalos médios (padrões), mas seu objetivo é estar atento ao que se desvio dos valores médios, os *outliers*. Em alguns casos, eles são mais importantes que os casos normais.

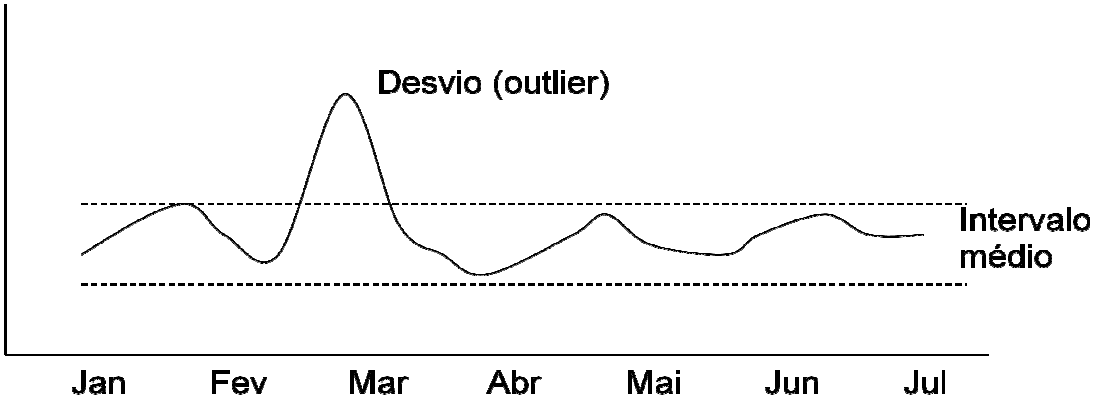


Figura 13: Detecção de desvios (outliers)

Esta técnica também é utilizada por instituições financeiras e administradoras de cartões de crédito. Se você tem um limite de mil reais num cartão, mas nunca fez compras acima de 500 reais, quando fizer uma compra de 700 reais, a operação será autorizada mas imediatamente irão lhe telefonar para confirmar a operação, pois ela "fugiu" do seu padrão.

Sequência de tempo

Esta técnica analisa sequências de eventos. Por exemplo, a técnica de associação pode identificar que fraldas são compradas em conjunto com cerveja, mas na mesma transação. Agora, se muitas pessoas compra um TV fina hoje e voltam depois de 3 meses para comprar um *home theater*, isto é função da técnica de sequência de tempo.
A Figura 14 apresenta um exemplo. Imaginem que são pacientes com suas linhas de tempo, e cada forma colorida indica um determinado evento importante na saúde desta pessoa. Podemos prever que há grande probabilidade de ocorre um evento do tipo "bolinha vermelha" na linha de tempo da paciente Ana, logo no início do ano de 2006,

já que todos os pacientes que tiveram eventos do tipo "triângulo amarelo" no início de um ano, tiveram "bolinha vermelha" no início do ano seguinte. É claro que isto é só um exemplo e a probabilidade deve ser levada em conta e não somente um número pequeno de casos.

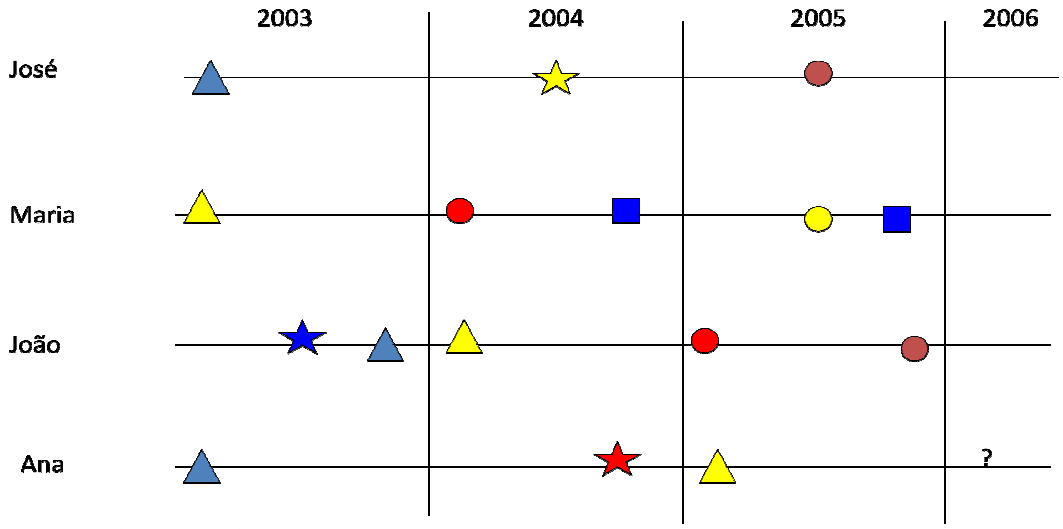


Figura 14: Técnica de análise de sequência temporal

Maltz e Klosak-Mullany (200) utilizaram a técnica de sequência de tempo (um tipo de Data Mining) para encontrar padrões estatísticos no comportamento de jovens delinquentes nos EUA e antever eventos ruins em suas vidas, para intervir antes que aconteçam.

Séries Temporais

Quando não é possível encontrar uma função que descreve o comportamento de uma variável (por exemplo, valor das ações de uma empresa ao longo do tempo), pode-se tentar prever pelo menos valores futuros num pequeno espaço de tempo. No caso das ações, por exemplo, pode-se querer saber se vão descer ou subir no dia seguinte.
Uma das formas de se fazer isto é analisando repetições de séries ao longo do tempo. Para isto, utilizam-se valores numéricos registros em sequência por vários períodos de tempo (a unidade de tempo não é fixa).

A Figura 15 apresenta o comportamento de uma variável ao longo do tempo, com seus altos e baixos. Imagine que se deseje saber o que vai acontecer após a linha contínua (mais à direita). Pode-se notar que um segmento deste gráfico repete-se. Então, é possível que o segmente que se repete seja maior e com isto saberíamos que a linha irá subir (como no trecho pontilhado).

É claro que as séries temporais são baseadas na premissa de que os comportamentos se repetem, pelo menos em parte (trechos ou momentos ao longo do tempo). Se isto não acontecer, não há por que usar séries temporais. Entretanto não se sabe qual o tamanho

de cada repetição. Além disto, há a premissa que outros fatores não irão influenciar o comportamento. Por exemplo, no caso das ações, uma notícia ou evento relevante pode influenciar o comportamento de compra e venda das ações, e o que era esperado (subir ou descer) pode não acontecer devido a isto.

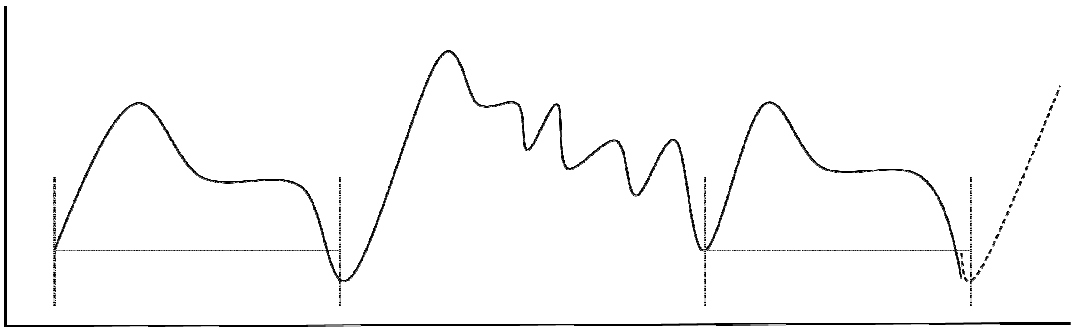


Figura 15: Exemplo de análise de séries temporais - dentro da mesma série

Outra possibilidade de utilizar séries temporais é comparar comportamentos de entidades diferentes. No caso anterior, usamos como exemplo a série de uma mesma entidade e as repetições eram procuradas dentro da mesma série. Na Figura 16, temos uma série principal acima e 3 relacionadas abaixo. Podemos supor que são gráficos referentes a totais de vendas ou receitas na matriz (acima) e filiais (abaixo). Pode-se notar que a série da matriz é semelhante à série da filial mais à esquerda, se analisarmos subidas e descidas em sequência e em momentos próximos no tempo. No caso deste exemplo, pode significar que a matriz e esta filial possuem práticas semelhantes. Se quisermos que as demais tenham comportamento semelhante ao da matriz, as filiais devem utilizar práticas semelhantes à da filial mais à esquerda.

Nesta mesma figura, pode-se notar que a filial mais à direita tem um gráfico quase que exatamente inverso ao da matriz. Isto pode significar comportamentos competidores: quando um gráfico está em cima, o outro está em baixo e vice-versa. Então, a comparação entre séries também pode ser feita para encontrar séries inversas ou contrárias.

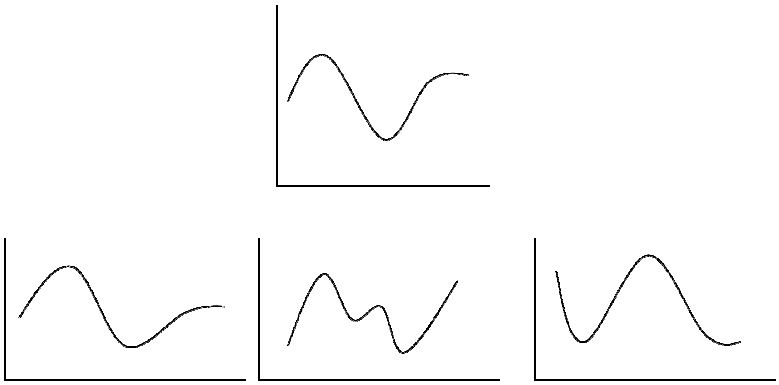


Figura 16: Exemplo de análise de séries temporais - comparação entre séries

A comparação de séries se dá não por proximidade de valores no tempo mas por semelhança no gráfico (subidas e descidas). Isto quer dizer que duas séries são semelhantes não importando o momento no tempo. Na Figura 17, podemos ver que as séries A e B são semelhantes e iniciam ao mesmo tempo. Por outro lado, a série C é também semelhante à série A, mas se inicia um pouco depois. Isto pode ser útil para avaliar retorno de campanhas de marketing. Por exemplo, ao se colocar propaganda na TV, talvez as vendas não cresçam logo no dia seguinte. E se tirarmos a campanha do ar, talvez as vendas ainda sigam aquecidas por um certo tempo.

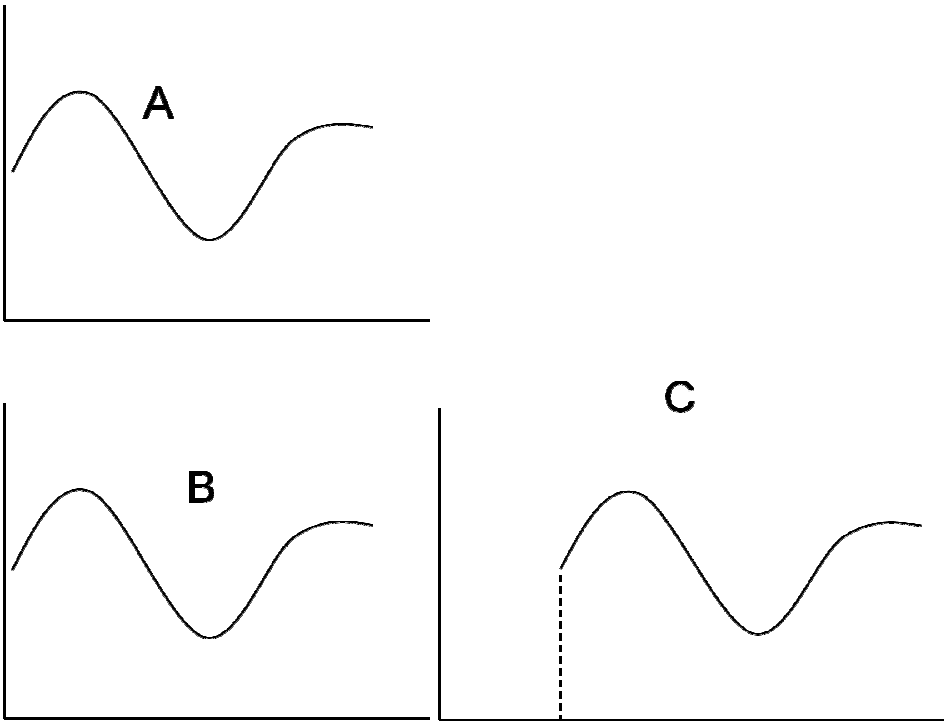


Figura 17: Séries temporais com diferença no momento de início da série

Classificação (categorização)

A técnica de classificação tem por objetivo encontrar a classe de um elemento. Note que por "classe", pode-se até mesmo entender uma ação (por exemplo, aprovar ou rejeitar um pedido de empréstimo). Para que a técnica funcione, as classes deverão já existir previamente.

O processo de avaliar a qual classe pertence um elemento novo pode fazer uso de regras determinísticas, probabilísticas, heurísticas, árvores de decisão, tabelas de decisão ou RBC (baseado em exemplos), conforme discutido no capítulo sobre Sistemas Especialistas.

Indução

O objetivo desta técnica é a identificação de um modelo para classificação, ou seja, a descoberta das regras de classificação. Isto é feito através do chamado "aprendizado supervisionado", onde exemplos de treino são avaliados para identificar padrões. Os algoritmos clássicos para indução incluem ID3 e C4.5.

Também é possível identificar, ao invés de regras, apenas as características de cada classe. Para isto, pode-se calcular o "centróide" da classe, que é um elemento hipotético que representa a classe, tendo a média das características dos elementos da classe ou um elemento hipotético que tenha todas as características de todos os elementos da classe.

Clusterização ou Agrupamento (clustering)

A técnica de Clustering recebe um grupo de elementos e daí identifica as classes. Ou seja, diferentemente da técnica de classificação, as classes não existem ainda ou não são conhecidas.

O princípio básico da técnica é colocar no mesmo grupo os elementos mais similares e em grupos diferentes os elementos pouco similares. Este agrupamento é feito por algoritmos automáticos como o k-Means e algoritmos baseados em grafos como Stars, Single-link, Strings e Cliques.

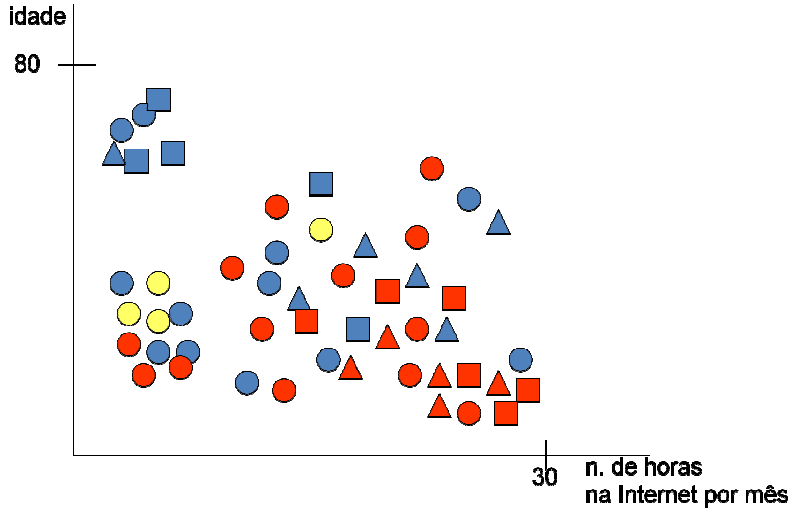


Figura 18: Exemplo de clustering

Mas para entender o processo, observe a Figura 18. Este gráfico posiciona clientes num plano que relaciona a idade da pessoa e o número de horas que passa na Internet por mês. Os símbolos no meio do gráfico representam o tipo de produto adquirido pelo cliente (quadrado, círculo ou triângulo) e a forma de pagamento (vermelho = cartão; azul = boleto; amarelo = depósito bancário).

Visualmente pode-se verificar que clientes de mais idade e que passam menos horas na Internet (quadrante mais à esquerda e em cima), é dominado por clientes que pagam por boleto bancário (cor azul). Clientes que compram por depósito bancário (cor amarela), só compram produtos do tipo círculo. Os clientes de menos idade tendem a passar mais horas na Internet e pagar com cartão (cor vermelha).

Uma empresa de telefonia segmentou seu portfólio de 70 aparelhos em quatro grupos, correspondendo a quatro categorias de clientes. A análise de perfis foi feita com base em atitudes dos clientes e resultou em 4 grupos de clientes: o "descomplicado", "multifuncional", "fashion" e "high tech".

O cliente "descomplicado" é o que pretende apenas falar ao telefone, é sensível a preço e não se importa com marcas, quer um aparelho de boa qualidade, durável e fácil de usar. O "multifuncional" faz questão de aproveitar todas as funcionalidades, como agenda, emails, vídeo, foto e tudo o mais que o aparelho oferecer para facilitar seu trabalho. O "fashion" é aquele que busca personalização, quer que o seu celular se identifique com ele, unindo as funções do anterior ao aspecto de estética. Por fim, o "hight tech" é aquele que faz questão de ter o aparelho mais sofisticado, com bluetooth, câmera com alta resolução, e tudo o que a tecnologia oferecer. Em geral não se importa com preços.

Esta segmentação atitudinal não tem nada a ver com o poder aquisitivo do cliente.
A separação foi feita manualmente mas poderia ter utilizado ferramentas automáticas de clustering.

5.2 Análise de cubos e análise multidimensional OLAP

Geralmente, os dados que compõem um Data Warehouse são organizados numa estrutura chamada Multidimensional. Isto porque há uma estrutura principal de dados (fatos) e estruturas auxiliares (dimensões). Por exemplo, um banco de dados sobre vendas de uma empresa teria como fatos os dados sobre as vendas (nota fiscal, códigos de produtos, código de clientes, data, valor pago, forma de pagamento, código da loja, código do vendedor), enquanto haveria outros dados relacionados a vendas (dimensões). As dimensões normalmente possuem uma estrutura particular e separada. Neste nosso exemplo, as dimensões e seus dados seriam: produtos (descrição, preço, setor), clientes (nome, endereço, idade), lojas (endereço, tamanho, gerente) e vendedores (nome, endereço, salário, data de admissão). Então o modelo deste exemplo possui 4 dimensões e uma base de fatos.

A vantagem de utilizar dimensões é que os fatos podem ser vistos sob diferentes perspectivas. Neste exemplo das vendas, o total de vendas pode ser apresentado por produto, por cliente, por loja ou por vendedor. O interessante também dos dados multidimensionais é que as dimensões podem ser cruzadas: por exemplo, comparar a idade do cliente com o preço do produto. Tal tipo de cruzamento nos dará informações que não poderiam ser vistas antes (como discutiremos nos próximos parágrafos).

A Figura 19 e a Figura 20 apresentam o mesmo conjunto de dados (vendas: produto X loja X quantidade). Na primeira representação, foi utilizado um modelo relacional não-

normalizado, enquanto que na segunda temos uma representação multidimensional (com matrizes). O formato multidimensional é mais compacto e também ajuda nas operações de análise. Neste exemplo, há somente duas dimensões: lojas e produtos.

Loja	Produto	Quantidade
1	X	10
1	Y	15
2	X	25
2	Y	20
2	Z	30
3	X	10
3	Z	20

Figura 19: Comparação de esquemas relacional X multidimensional para DWH

PRODUTOS

LOJAS

	X	Y	Z
1	10	15	-
2	25	20	30
3	10	-	20

Figura 20: Comparação de esquemas relacional X multidimensional para DWH

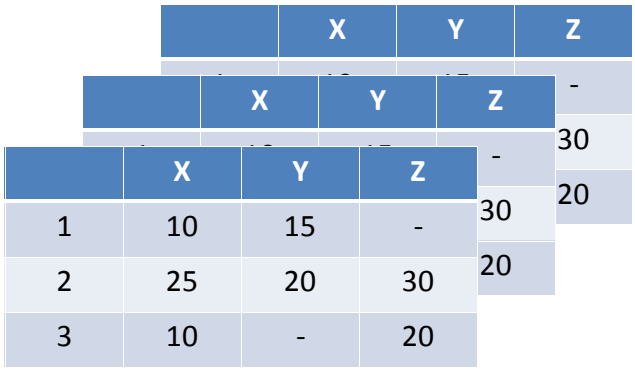


Figura 21: Dados multidimensionais - exemplo para 3 dimensões

Imagine agora que se queira acrescentar uma 3a dimensão, por exemplo, o cliente. As vendas de cada cliente formariam uma matriz e assim teríamos tantas matrizes quanto forem os clientes. Assim, teríamos o esquema da Figura 21. Isto dá a ideia de 3a

dimensão como visto na imagem. Se for necessário acrescentar mais dimensões (por exemplo, vendedor), isto será feito nas estruturas internas de armazenamento, pois não será possível ao ser humano imaginar visualmente tal estrutura (4 dimensões).

Se olharmos melhor, esta imagem lembra a de um cubo, por isto, muitas vezes os dados multidimensionais são também conhecidos como dados cúbicos (ou sua representação é conhecida como cubo de dados).

A vantagem dos dados cúbicos é acelerar as análises e dar respostas mais rapidamente para usuários que tomam decisões. Além disto, a visualização de dados em duas ou mais dimensões ajuda a ver padrões que são difíceis de identificar em tabelas normalizadas (*flat*).

Por exemplo, se tivermos uma base de dados sobre falhas que ocorreram em máquinas numa empresa, provavelmente a estrutura será similar à que pode ser vista na Figura 22, onde todos os atributos das falhas estão como colunas: identificação da máquina, setor onde ocorreu a falha, quem era o operado no momento da falha, data e hora da ocorrência, tipo de problema que ocorreu, quantas horas a máquina ficou parada devido à falha, custo por hora da máquina parada e prejuízo total que a falha gerou, multiplicando-se as horas paradas pelo custo-hora.

Neste tipo de estrutura, fica difícil verificar quais os problemas que mais ocorrem com cada máquina, qual o total de falhas por operador, etc., especialmente se são muitas falhas (muitas linhas ou registros).

3	Máquina	Setor	Operador	Data Ocorrência	Hora	Tipo problema	Horas Paradas	Custo-hora	Custo total da falha
4	empilhadeira	produção	João Maria	30/07/00	9	falta peças	2	50	100
5	empilhadeira	produção	João Maria	30/07/00	9	falta peças	1	50	50
6	empilhadeira	produção	João Maria	30/07/00	9	falta peças	2	50	100
7	computador	almoxarife	Beltrão	12/08/03	9	falta software	2	10	20
8	computador	almoxarife	Rudinei	16/09/03	7	falta software	3	10	30
9	computador	almoxarife	Rudinei	16/09/03	7	falta software	4	10	40
10	computador	almoxarife	Rudinei	16/09/03	7	não liga	5	10	50
11	computador	almoxarife	Rudinei	16/09/03	7	não liga	1	10	10
12	computador	almoxarife	Rudinei	16/09/03	7	não liga	2	10	20
13	computador	almoxarife	Rudinei	08/11/03	7	não liga	1	10	10
14	computador	almoxarife	Rudinei	08/11/03	7	não liga	1	10	10
15	computador	almoxarife	Rudinei	08/11/03	7	não liga	1	10	10
16	computador	almoxarife	Rudinei	08/11/03	7	não liga	12	10	120
17	computador	almoxarife	Rudinei	08/11/03	7	falha HD	2	10	20
18	computador	almoxarife	Rudinei	08/11/03	7	falha HD	3	10	30
19	computador	almoxarife	Beltrão	11/12/03	9	não liga	2	10	20
20	computador	almoxarife	Beltrão	11/12/03	9	não liga	2	10	20
21	computador	almoxarife	Beltrão	11/12/03	9	falta software	2	10	20
22	computador	almoxarife	Beltrão	11/12/03	9	falha HD	2	10	20
23	computador	almoxarife	Beltrão	11/12/03	9	não liga	1	10	10
24	computador	almoxarife	Beltrão	11/12/03	9	não liga	1	10	10
25	computador	almoxarife	Rudinei	13/12/03	8	não liga	1	10	10
26	empilhadeira	produção	João Maria	13/01/04	10	falta peças	1	50	50
27	empilhadeira	produção	João Maria	13/01/04	10	falta peças	2	50	100
28	empilhadeira	produção	João Maria	13/01/04	10	falta peças	3	50	150
29	empilhadeira	produção	João Maria	19/01/04	15	travou entrada	4	50	200

Figura 22: Estrutura de dados flat - todos atributos como colunas

Por isto, uma estrutura multidimensional, como a apresentada na Figura 23, permite mais rapidamente verificar padrões. Na estrutura multidimensional, os atributos podem aparecer como linhas ou colunas. Isto permite relacionar atributos entre si e encontrar padrões que não podem ser verificados nas estruturas unidimensionais (tipo "*flat*").

No exemplo da Figura 23, estamos relacionando duas dimensões: identificação da máquina (nas linhas) X tipo de problema (nas colunas). Na figura, podemos ver o total de falhas para cada máquina (última coluna à direita), o total de falhas por tipo de problema (última linha) e a quantidade de falhas para cada par máquina X tipo de problema. Por exemplo, pode-se notar que ocorreram 51 registros no entroncamento da linha da "empilhadeira" com a coluna de "falta peças", indicando que a máquina Empilhadeira teve 51 falhas por falta de peças. Rapidamente também podemos notar qual o tipo de problema mais comum relacionado a cada máquina.

3	Contar de Data Ocorrência	Tipo problema						
4	Máquina	falha HD	falta peças	falta software	não imprime	não liga	travou entrada	Total geral
5	computador	6		7	3	32		48
6	empilhadeira		51				33	84
7	limpadora		14					14
8	Total geral	6	65	7	3	32	33	146

Figura 23: Estrutura multidimensional - máquina X tipo de problema

Na Figura 24, estamos relacionando o operador com a hora em que a falha ocorreu. Aqui a estrutura multidimensional permite visualizar que as falhas com o operador Beltrão só ocorrem às 9h da manhã e que o operador Rudinei só teve falhas no início do dia (entre 7 e 8h da manhã). Também pode-se notar que as falhas com o operador João Maria ocorrem mais frequentemente de manhã, enquanto que para Menezes e Otto as falhas são mais frequentes à tarde. A estrutura multidimensional também dá uma visão diferenciada das falhas que ocorreram com o operador José Carlos: elas ocorrem em ambos os turnos, mas acontecem mais no início dos turnos. Este tipo de análise não poderia ser feita com dados na estrutura *flat*.

3	Contar de Data Ocorrência	Hora										
4	Operador	7	8	9	10	11	13	14	15	16	19	Total geral
5	Beltrão			12								12
6	João Maria		12	5	6	3		3	4			33
7	José Carlos		9	3			3	15	6	4	1	41
8	Menezes			1				5	8	2	2	18
9	Otto							2	2	2		6
10	Rudinei	18	18									36
11	Total geral	18	39	21	6	3	3	25	20	8	3	146

Figura 24: Estrutura multidimensional - operador X hora em que ocorreu a falha

Para o caso de ser necessário analisar mais de 2 dimensões, já que as telas de computadores ainda não permitem visualizar dados em 3D, deve-se utilizar uma visualização 2D adaptada, como mostra a Figura 25, onde se pode ver que há 3 dimensões relacionadas: operador, tipo de problema e hora. Note que as dimensões (ou atributos) operador e tipo de problema foram colocados nas linhas, formando uma hierarquia.

3	Contar de Data Ocorrência		Hora										
4	Operador	Tipo problema	7	8	9	10	11	13	14	15	16	19	Total geral
5	⊟Beltrão	falha HD			2								2
6		falta software			2								2
7		não imprime			1								1
8		não liga			7								7
9	Beltrão Total				12								12
10	⊟João Maria	falta peças		12	5	6	3		2	1			29
11		travou entrada							1	3			4
12	João Maria Total			12	5	6	3		3	4			33
13	⊟José Carlos	falta peças		9	3			3				1	16
14		travou entrada							15	6	4		25
15	José Carlos Total			9	3			3	15	6	4	1	41
16	⊟Menezes	falta peças			1				4	8	1	2	16
17		travou entrada							1		1		2
18	Menezes Total				1				5	8	2	2	18
19	⊟Otto	falta peças							1	2	1		4
20		travou entrada							1		1		2
21	Otto Total								2	2	2		6
22	⊟Rudinei	falha HD	2	2									4
23		falta software	4	1									5
24		não imprime	1	1									2
25		não liga	11	14									25
26	Rudinei Total		18	18									36
27	Total geral		18	39	21	6	3	3	25	20	8	3	146

Figura 25: Estrutura multidimensional - máquina + tipo de problema X hora

3	Soma de Horas Paradas	
4	Tipo problema	Total
5	falha HD	15
6	falta peças	202
7	falta software	22
8	não imprime	6
9	não liga	71
10	travou entrada	122
11	Total geral	438

Figura 26: Análise OLAP com somente uma dimensão

A análise OLAP também pode ser feita com uma dimensão somente, como no caso da Figura 26, onde há somente o atributo "tipo de problema" e a análise é feita pela soma de horas paradas.

No link abaixo, há uma animação mostrando como fazer análises multidimensionais com tabelas dinâmicas no software MS Excel:
http://www.youtube.com/watch?v=4hZN2YWKuy8

6 Interpretação dos resultados da análise

Como discutido anteriormente, o processo de descoberta de conhecimento tem por objetivo identificar conhecimentos novos e úteis. Por outro lado, as técnicas de Data Mining e análise OLAP apenas apresentam padrões estatísticos, e isto não é conhecimento. Portanto, é necessário interpretar cada padrão para poder extrair conhecimento.

Por exemplo, no caso da análise de pacientes com diabetes, onde se descobriu que 95% dos pacientes com diabetes com tipo 1 tinham um determinado tratamento, não é novidade para quem já é familiarizado com a área. Entretanto, este padrão evoca a dúvida sobre o que estaria acontecendo com os 5% dos pacientes que têm o mesmo diagnóstico e não estão recebendo o mesmo tratamento. Neste caso, o fato interessante estava nas exceções, e portanto será necessário investigar as exceções e não a normalidade para poder extrair conhecimento novo.

Em outros casos, talvez o conhecimento mais interessante esteja na conjunção entre dois padrões. Por exemplo, ao se descobrir (a) que 80% das máquinas da marca XYZ quebravam com 3 anos de uso e (b) que 77% das máquinas desta marca eram operadas por pessoas altamente experientes (mais de 10 anos no ramo), levanta-se a curiosidade de saber qual o percentual para a conjunção dos 2 casos, ou seja, o que estaria acontecendo com máquinas da marca XYZ com 3 anos de uso e operadas por profissionais com mais de 10 anos de experiência. Ou então, o interessante pode estar em combinar um padrão com o negativo de outro: o que acontece com as máquinas XYZ com menos de 3 anos e operadas por pessoas com mais de 10 anos de experiência, e o que acontece com máquinas XYZ com 3 anos e operadas por pessoas com menos de 10 anos de experiência.

Também pode ser necessário realizar comparações entre padrões, como discutiremos adiante. Um padrão que apareça com 80% de probabilidade numa amostra e com 60% de probabilidade em outra merece ser investigado (investigar o porquê da diferença).

Para entendimento dos padrões, é necessário conhecimento sobre o domínio, o qual pode não estar presente nos dados analisados. Por exemplo, o famoso caso da relação entre fraldas e cervejas num supermercado, exigiu, para a interpretação das causas, conhecimento sobre padrões já conhecidos no supermercado mas que não estava formalizado em algum meio físico (estava só na cabeça de algumas pessoas com experiência no ramo, na forma de conhecimento tácito).

Desta forma, para a interpretação dos resultados é importante ter alguém com conhecimento sobre o domínio, ramo, mercado ou específico da empresa.

Além disto, a interpretação dos padrões identificados depende do contexto, ou seja, de como tais padrões foram identificados. Isto quer dizer que os padrões se referem somente à amostra de dados analisada. Uma amostra pretende ser representativa de um universo, mas ela nunca o é de forma completa.

Outro cuidado que devemos ter é que os dados são influenciados por eventos externos e assim a interpretação dos resultados deve entender que eventos aconteceram ou estão acontecendo. Por exemplo, no famoso caso da associação entre fraldas e cervejas, o tal supermercado tomou atitudes após esta descoberta. Ou eles colocaram os produtos próximos ou colocaram bem longe. E isto deve ter influenciado o padrão, aumentado seu percentual ou talvez até acabando com ele. Então o tal supermercado precisa refazer o processo de análise e comparar os novos resultados com os anteriores. A interpretação não pode estar dissociada do tempo em que os fenômenos ocorrem e de seu contexto.

Um certo manual de investigação criminal aponta algumas falhas na interpretação de dados, descritas a seguir:

- Excesso de simplificação ou excesso de complicação: é o perigo de assumir a interpretação mais simples; nós seremos humanos temos tendência, por preguiça mental ou falta de tempo, acolher como melhor alternativa aquela que é mais simples; por outro lado, quando as pessoas determinam que um problema é complicado, normalmente procuram soluções complicadas; a dica é comparar as interpretações possíveis à luz do contexto, sem se deixar influenciar por simplificação ou complicação;
- Erros de causa: como discutiremos adiante neste livro, encontrar causas é uma tarefa difícil em qualquer situação; muitas vezes falhamos ao estabelecer relações de causa-efeito; a presença de certos elementos com alta frequência a certos tipos de eventos conduz erroneamente as interpretações para alternativas que não são verdadeiras; adiante, teremos um capítulo só para discutir relações de causa-efeito;
- Falsos dilemas ou dicotomias: quando nos concentramos em duas explicações para um fenômeno que são opostas; é normal no ser humano considerar apenas frio-quente, bom-ruim, perto-longe; mas existe o meio termo;
- Amostras inadequadas: Nate Silver também descreve problemas com amostras inadequadas; é muito difícil conseguir coletar todos os dados úteis então acabamos sempre ficando com uma amostra dos dados; este tipo de simplificação acaba nos levando a resultados irreais; alguns exemplos serão dados ao longo deste livro.

6.1 Resultados condizem com a técnica usada

Uso de técnicas erradas ou dados pobres pode levar a conclusões ou interpretações erradas. Por exemplo, a Figura 27 apresenta em vermelho o gráfico de vendas (eixo y) em um site de comércio eletrônico ao longo do tempo (eixo x). Em azul, temos a média de vendas neste período. O gerente deste site considerou baixa a média de vendas e descontinuou as vendas pelo site. Entretanto, o que ele não notou é que no momento em que foi descontinuado o site, as vendas estavam no seu auge. Ou seja, se ele tivesse usado a técnica de tendência, teria visto que as vendas estavam subindo e talvez fosse bom esperar um pouco mais para ver o resultado final.

Este é o mesmo tipo de análise que é feita quando se fala em aquecimento global. Independente da discussão se a causa é humana ou não, estatisticamente, está comprovado que a média de temperatura anual no mundo todo está crescendo. Algumas pessoas não acreditam nisto porque olham o inverno de um ano e verificam que ele foi mais frio que o inverno do ano anterior. Sim, isto pode acontecer. Mas o que está sendo

medido são médias por ano e levando em conta todas as medições pelo mundo todo. Realmente, pode acontecer que, em algumas regiões, a média pode ter baixado de um ano para outro. Mas nem isto mesmo é argumento contrário. O filme de Al Gore, "Uma verdade inconveniente", mostra claramente este gráfico. E a tendência é de subida. Ou seja, temos que usar a técnica correta.

Com relação ao aquecimento global, muitos acreditam que estamos nos aproximando de uma era de temperaturas altas. Entre os anos 1000 e 1200 d. C., tivemos uma época com média de temperatura 6 graus acima da média atual. Foi assim que os Vikings fizeram fazendas na Groenlândia e chegaram até a América. E isto pode estar novamente acontecendo. Por outro lado, segundo alguns estudiosos, há também ciclos de eras glaciais e é possível que estejamos a algumas dezenas de anos de uma pequena era do gelo. Então, talvez até este tipo de acontecimentos pode ser regido por padrões. Mas é bom deixar claro que ainda não li nenhum estudo que comprovasse que há um padrão. Nate Silver conta que já tentaram encontrar padrões temporais ou sazonais em terremotos e os resultados não foram bons, pois deixaram de prever os grandes que aconteceram na Itália em 2006 (L´Aquila) e no Japão em 2011 (Fukushima).

Bom, para completar um pouco a discussão e ver também o outro lado, a ONU divulgou recentemente (em setembro de 2013) um relatório apontando que uma das causas para as mudanças climáticas é a ação do Homem sobre a natureza.

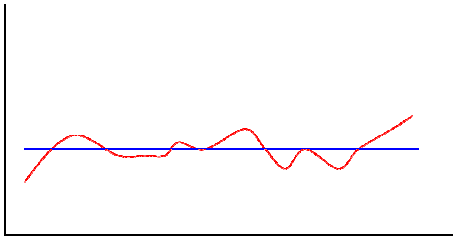


Figura 27: Média X Tendência

Média de gasto	
Média de Valor-gasto	
Acompanhado	Total
amigos	80,64516
casal	121,3768
familia	137,1711
sozinho	104,661
Total geral	110,2214

Figura 28: Média de gastos de clientes num supermercado, por perfil

Voltando à discussão sobre interpretação de resultados, eu queria discutir ainda um outro exemplo de má interpretação. A Figura 28 apresenta a média de gasto para cada tipo de cliente. Estes valores foram calculados assim: as vendas num supermercado

foram registradas em separado, tendo associado a cada uma delas um atributo informando o tipo do cliente, ou seja, como ele veio ao estabelecimento (se sozinho, acompanhado de amigos, se era um casal ou se era uma família com crianças). Depois, o valor total de cada venda foi somado para cada tipo de cliente em separado e então feita a média (total por tipo dividido por número de carrinhos/vendas para cada tipo). Nesta figura, podemos ver que a média de gasto do cliente tipo "sozinho" (ou seja, pessoas que estavam desacompanhadas no momento da compra) era de 104 reais.

Por outro lado, temos a Figura 29, que apresenta o % de carrinhos em cada faixa ou categoria de gastos (valores arredondados para múltiplos de 50). Nesta tabela, podemos notar que 44,6% dos clientes do tipo "sozinho" gastam em torno de 50 reais e apenas 26,5% gastam perto de 100 reais. O que contradiz o valor resultante da figura anterior.

A causa para esta discrepância é que a média não leva em conta o desvio padrão. Assim, se uma pessoa sozinha fizer uma compra de 5 mil reais neste supermercado, vai aumentar a média de gasto dos clientes deste tipo. Ou seja, os chamados "outliers", valores que se distanciam muito da média, também acabam sendo contados. Então, a segunda tabela é mais precisa em nos dizer a expectativa de gasto de cada tipo de cliente.

Contar de CÓD	Acompanhado				
Valor-gasto	amigos	casal	familia	sozinho	Total geral
50	57,26%	31,88%	30,92%	44,63%	41,67%
100	33,06%	21,01%	13,16%	26,55%	23,96%
150	3,23%	25,36%	15,13%	9,89%	12,63%
200	4,84%	18,12%	32,89%	14,69%	17,32%
250	0,81%	1,45%	7,24%	2,26%	2,86%
300	0,81%	2,17%	0,66%	1,98%	1,56%
Total geral	100,00%	100,00%	100,00%	100,00%	100,00%

Figura 29: Gastos de clientes num supermercado, por perfil, e classificados por faixa de gasto

6.2 Indicadores escolhidos para BI - certos ou errados

As ferramentas, técnicas e softwares utilizados nos processos de BI apenas apresentam os dados solicitados pelos usuários. A interpretação é sempre humana. Como os dashboards são criados por pessoas, muitas vezes eles podem estar apresentando indicadores equivocados para uma determinada análise ou tomada de decisão. Por exemplo, muitas empresas criam um ranking de vendedores utilizando somente o indicador de "soma de valores monetários referentes às vendas feitas por cada vendedor". Entretanto, muitas vezes, este indicador pode estar premiando quem não é o melhor vendedor. Há outros indicadores que talvez tenham que ser levados em conta, como por exemplo:

- custos para realizar a venda: um vendedor X pode ter vendido 100 mil reais no mês mas ter gerado um custo de 70 mil para a empresa (lucro de 30 mil), enquanto que o vendedor Y faturou apenas 50 mil mas teve um custo de apenas 10 mil (lucro de 40 mil); então a lucratividade talvez seja um melhor indicador;

- tempo despendido: um vendedor talvez tenha faturado menos que outros porque teve mais tempo de deslocamento ou porque teve que realizar mais tarefas burocráticas; se ele tivesse o mesmo tempo para dedicar aos clientes em contato direto, talvez pudesse ter o mesmo índice de vendas;
- número de clientes a visitar: muitas empresas determinam os clientes que os vendedores devem visitar; o mais correto neste caso, seria avaliar a média de vendas por cliente;
- número de clientes novos: alguns vendedores acumulam tarefas de prospecção, ou seja, precisam, além de concretizar vendas, encontrar novos clientes; alguns realmente conseguem conquistar novos clientes, mas que talvez não gastem tão alto, justamente por serem novos; mas estes novos clientes talvez sejam repassados para outros vendedores no próximo mês e aí as vendas futuras subsequentes serão contabilizadas para outro vendedor;
- desistências de clientes: avaliar vendedores somente por pedidos feitos pode ser perigoso se os pedidos não se concretizarem; da mesma foram, avaliar somente pelas vendas concretizadas pode deixar de fora desistências, principalmente quando os pagamentos dos clientes são realizados a prazo; a inadimplência dos clientes também deveria ser somada (ou subtraída) aos respectivos vendedores.

Discussão similar ocorre na hora de determinar os melhores produtos para a empresa. Só levar em conta quantidade vendida não é suficiente. O custo e o preço final também interferem, ou seja, talvez seja melhor utilizar a lucratividade de cada produto.

O mesmo ocorre na hora de "rankear" clientes. Qual é o melhor cliente: aquele que compra todo mês e só gasta 100 reais por mês ou aquele que só vem uma vez por ano mas gasta 3 mil reais ? Pela lucratividade, o segundo cliente é melhor (cliente de maior valor) mas o primeiro é pode ser um "cliente de maior potencial", já que vem mais seguido.

E o caso de quem compra 1000 pequenos produtos num supermercado (como sabonete, pasta de dente, desodorante, etc.) totalizando 3 mil reais, é melhor cliente que alguém que compra um eletrônico no mesmo valor total ? Para levar todos os 1000 produtos talvez seja necessário um caminhão e várias pessoas, mas para transportar o eletrônico talvez um carro e uma pessoa sejam suficientes.

A conclusão é que os indicadores devem ser bem definidos, por quem realmente conhece o negócio. Analistas de BI só devem gerar as análises ou apresentações. O BI não é culpado por apresentar dados equivocados; ele só apresenta o que é solicitado.

6.3 Teoria do Mundo Fechado

É comum entre analistas de dados e mesmo entre cientistas de diversas áreas incutir no erro conhecido como a Teoria do Mundo Fechado. Vejamos um exemplo. A Figura 30 apresenta o gráfico de vendas de laranjas num supermercado ao longo do tempo (apenas 6 meses são mostrados). Nos 4 primeiros meses, o supermercado comprava do fornecedor "vermelho". No 5o mês, trocaram para o fornecedor "azul", mas voltaram a comprar do fornecedor "vermelho" no 6o mês. Pode-se notar que o nível das vendas nos

primeiros 5 meses é semelhante mas no 6o mês as vendas caíram muito. De quem é a culpa ? Do fornecedor vermelho ou do azul ?

Uma resposta possível é culpar o fornecedor "vermelho". A explicação seria assim: os clientes compraram as laranjas "azuis" e acharam de melhor qualidade que as laranjas "vermelhas". Quando vieram ao supermercado no 6o mês, viram que havia voltado o fornecedor "azul" e aí não compraram no mesmo nível.

Outra possibilidade é culpar o fornecedor "azul". A explicação seria esta: os clientes estavam acostumados às laranjas "vermelhas", e acabaram comprando as "azuis" no 5o mês por acomodação, mesmo sabendo que era de outro fornecedor. Mas quando provaram as laranjas azuis, não gostaram. Então não voltaram a comprar laranjas neste supermercado no mês seguinte (e o nível das vendas no 6o mês caiu).

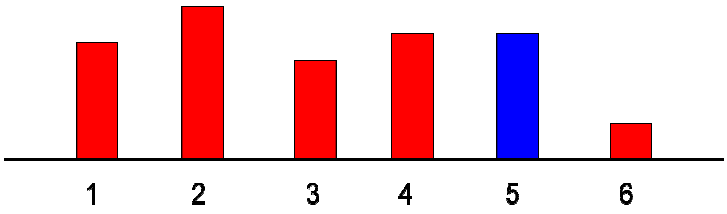


Figura 30: Venda de laranjas num supermercado

Estas explicações funcionam se só tivermos estes dados disponíveis. E é assim que as pessoas costumam tomar decisões. Entretanto, há outras possibilidades de causas. Uma delas é a sazonabilidade, ou seja, sempre no 6o mês do ano as vendas de laranjas diminuem. Se não tivermos dados dos anos anteriores, não vamos entender esta padrão e acabar culpando o fornecedor.

Outra explicação é que no 6o mês o supermercado concorrente fez uma promoção de laranjas e por isto as vendas diminuíram no primeiro supermercado. Mas de novo sem esta informação, acabaríamos culpando fornecedores.

Isto acontece porque as pessoas só fazem análises com os dados armazenados na tecnologia (por exemplo, nos bancos de dados das empresas). Funciona como esquematizado na Figura 31. Coletamos dados do mundo real através de diferentes formas e os armazenamos em bancos de dados (tecnologia). As análises são feitas sobre estes dados armazenados. Os padrões encontrados nos dados são interpretados gerando conhecimento novo. E aí acreditamos que este conhecimento explica o mundo real. O problema é que o conhecimento só explica os dados armazenados. Como no exemplo citado antes (vendas de laranjas no supermercado), se não temos todos os dados que podem influenciar as análises, acabamos chegando a conclusões que não valem no mundo real (condizem apenas aos dados armazenados).

Por isto, é importante a etapa de preparação dos dados, para que todos os dados relevantes para entendimento dos padrões estejam disponíveis para análise.

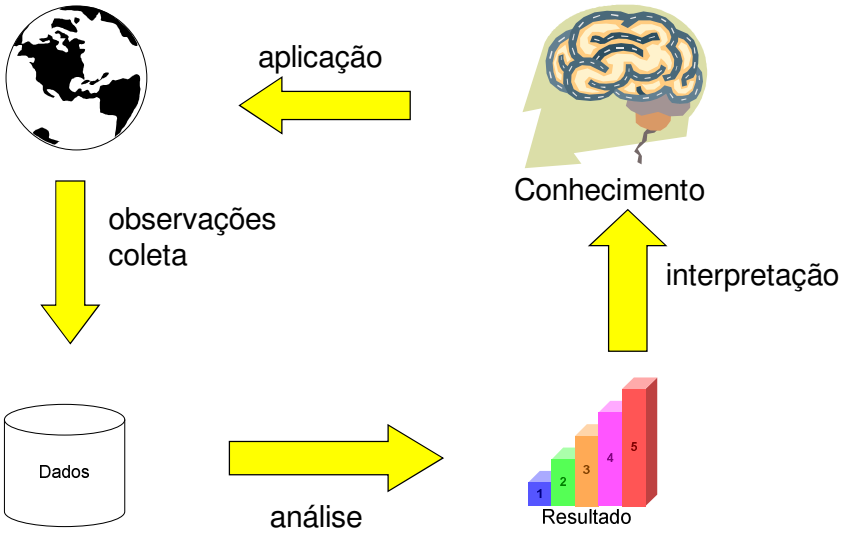


Figura 31: Teoria do Mundo Fechado

Entretanto, não há como coletar todos os dados; por isto, nosso mundo não é fechado. Até o planeta Terra troca energia e matéria com resto do Universo. E o acelerador de partículas do CERN na Suíça precisa de alguém para ligar (e há sensores também). Desta forma, temos que ter ciência de que os resultados das análises dizem respeito tão somente aos dados analisados, isto é, às amostras analisadas. O conhecimento descoberto é então uma hipótese ou tendência, que deverá ser confirmada analisando o mundo real ou através de tentativa e erro.

6.4 Correlações erradas

O perigo da análise de correção é supor causas erradas para eventos. Por exemplo, anos atrás os americanos achavam que o sorvete era causador da pólio, porque os gráficos eram muito semelhantes; as vendas de sorvete e os casos de pólio cresciam no verão. As duas variáveis tinham uma correlação estatística, mas uma não era causa ou efeito de outra (Levitt e Dubner, 2009).

Nate Silver, no livro "O sinal e o ruído" comenta diversos casos de correlações erradas. Um deles fala de uma pesquisa inglesa que concluiu que vacas com nome produziam mais leite que vacas anônimas. Na verdade, o fator que influenciava a produção era a personalização no cuidado com o animal. As vacas mais bem cuidadas recebiam nomes. Desta forma, a produção era maior não pelo nome em si para pelo maior cuidado que recebiam dos tratadores.

Há também um caso famoso (citado em http://epocanegocios.globo.com/Informacao/Acao/noticia/2013/08/jornalista-americana-vira-suspeita-de-terrorismo-por-buscar-panela-de-pressao-na-internet.html), em que uma jornalista americana se tornou suspeita de terrorismo. Aconteceu que ela fez várias buscas na Internet por panelas de pressão, enquanto seu marido buscou mochilas no mesmo período. E ainda por cima, seu filho leu diversas notícias sobre o atentado em

Boston, onde uma bomba foi feita com uma panela de pressão e colocada numa mochila.

Vários casos de correlações estranhas são listadas em http://www.tylervigen.com/ e você pode fazer a sua própria escolhendo variáveis.

Max Gunther (no livro "O Fator Sorte") conta o caso de um sujeito que costumava tocar clarim e abanar uma bandeira verde numa esquina, dizendo que servia para espantar girafas. Quando perguntado se dava certo, ele respondia dizendo que nenhuma girafa havia passado por ali.

Muitas vezes os padrões podem dar certo talvez pelo efeito placebo: achamos que vamos ter melhor rendimento usando certos padrões ou superstições. Aí, repetimos o padrão e o resultado acontece como esperado. Neste caso, há relação entre duas variáveis mas uma não implica na outra. É pura coincidência ou sorte. É como regular sua alimentação e ver efeitos positivos, e então acreditar que descobriu um novo método. E isto aí vira sabedoria popular e vai passando de boca em boca. Como os sacrifícios humanos para os deuses ou para ajudar na agricultura e clima (a civilização Maia fazia isto).

Outro engano típico é supor relações de causa e efeito em variáveis que possuem comportamentos similares. A correlação existe porque os valores são similares ao longo do tempo, mas não necessariamente pode haver uma relação entre elas. Por exemplo, vendas de sorvete e vendas de maiôs aumentam no verão e diminuem no inverno, mas uma variável não implica na outra. Neste caso, há uma causa comum (a temperatura ou estações) que determina estes comportamentos mas não há relação direta entre os dois tipos de vendas.

Muitas vezes, ocorrem coincidências. Gunther também fala da Sincronicidade. Segundo a Wikipedia, Sincronicidade "*é um conceito desenvolvido por Carl Gustav Jung para definir acontecimentos que se relacionam não por relação causal e sim por relação de significado. Desta forma, é necessário que consideremos os eventos sincronísticos não a relacionado com o princípio da causalidade, mas por terem um significado igual ou semelhante. A sincronicidade é também referida por Jung de 'coincidência significativa'* ". Um exemplo é o caso de um americano que lutou na Guerra da Coréia e teve um filho por lá. Mas nem sabia disto. O filho foi trabalhar nos EUA e não sabia nada do pai, a não ser seu nome. Um dia, aquele americano estava andando dirigindo pela estrada e resolveu parar num restaurante que não costumava. Quando foi pagar em cartão, o atendente viu o nome e adivinhem: era seu pai. Uma grande coincidência, uma sincronicidade: tais eventos são comuns de ocorrer; o que determina sua relevância é que aconteceu com pai e filho que não se conheciam.

Max Gunther, no seu livro "O Fator Sorte" diz que há duas leis estatísticas: (a) tudo pode acontecer e (b) se algo pode acontecer, vai acontecer algum dia, pelo grande volume de casos (por exemplo, cair 5 vezes o mesmo número na roleta em algum cassino do mundo, algum dia).

Descobrir correlações entre variáveis é fácil; há métodos matemáticos/estatísticos para isto, inclusive nas planilhas eletrônicas. O problema é saber se um fator determina outro (implica em outro), ou seja, se há uma relação de causa-efeito entre duas variáveis. Para

isto, precisamos separar as relações que são estatísticas das que são coincidência ou acaso ou sorte. Este problema será discutido mais adiante.

Além disto, a correlação entre duas variáveis pode perdurar por apenas um certo período de tempo. Vejamos dois exemplos. Imagine uma fruta que só apareça no verão e que seja quase impossível guardá-la numa câmera fria para ser comercializada no inverno. É certo que as vendas desta fruta serão maiores no verão. Se analisarmos a correlação entre as vendas da fruta e a variável "dias de verão", encontraremos uma forte relação. Mas só neste período. O perigo é generalizar para outras estações.

Em outro exemplo, imagine que durante 10 anos as vendas de um certo produto infantil estiveram fortemente correlacionadas com aparições de uma certa atriz em novelas, programas de TV, noticiários, etc. Entretanto, esta atriz envelheceu, a moda mudou, as crianças cresceram, e a correlação não enfraqueceu ou desapareceu. Então, temos que admitir que a correlação entre duas variáveis pode ser forte mas talvez não dure para sempre.

6.5 Sobrecarga e Ruídos

Michael Lewis no seu livro “Moneyball” fala do perigo de estatísticas baseadas em variáveis que não interferem no resultado. No baseball, algumas estatísticas utilizam o número de vitórias e derrotas para avaliar o desempenho de um jogador. Entretanto, vitória ou derrota não dependem unicamente de um jogador. Há diversos fatores alheios à competência do jogador. Por isto, Lewis sugere que o melhor, neste caso, é considerar um grupo menor de características. No caso, foram escolhidos cinco requisitos: potência ao rebater, média de rebatidas, velocidade, força do braço e alcance defensivo.

Além de determinar corretamente quais fatores devem ser analisados, deve-se também determinar pesos, ou seja, o quanto um fator é mais importante que outro. No exemplo do baseball, a potência ao rebater é muito mais importante do que a força do braço e o alcance defensivo na maioria das posições, exceto para as posições de shortstop e receptor.

E por que utilizar um número reduzido de características ? Nate Silver comenta que muitos analistas econômicos utilizam 4 mil variáveis para fazer previsões econômicas. Muitas destas variáveis são irrelevantes e confundem os resultados. O correto é avaliar quais fatores são os únicos ou os mais determinantes de um resultado.

Alguns autores falam da chamada “causa-raiz”, ou seja, separar as causas que realmente levam a um determinado resultado. Por exemplo, uma empresa pode detectar que, para aumentar a satisfação de seus clientes, precise diversos fatores tais como: tornar o ambiente mais confortável, baixar preços, fazer mais promoções e mais diversificadas, melhorar o relacionamento do funcionário com o cliente, etc. Entretanto, cada um destes fatores pode ser conseguido através de subfatores. Por exemplo, para tornar o ambiente mais agradável, talvez seja necessário ter uma melhor disposição dos produtos e uma decoração mais atraente; para baixar preços e fazer promoções, talvez seja necessário que gerentes financeiros aprendam novas técnicas; para melhorar o relacionamento na loja, talvez seja melhorar a cordialidade do funcionário e seu modo de abordagem ao

cliente. Bom, estes fatores parecem exigir uma qualificação melhor dos funcionários. E isto tudo exige orientações e cursos para funcionários. Desta forma, o fator-chave, a causa-raiz talvez seja o treinamento dos funcionários.

Sobre a análise de causa-raiz, discutiremos mais adiante neste livro.

7 Processo de BI reativo

Normalmente o processo de BI (Business Intelligence) recebe como entrada solicitações para gerar como resultado indicadores quantitativos tais como níveis de venda, custos e lucratividade (por produto, loja, vendedor, departamento, etc). Neste caso, o objetivo do BI é apresentar graficamente os indicadores e monitorá-los, atualizando-os em tempo real. Estes indicadores são também chamados KPI (Key Performance Indicators), um termo que vem da metodologia de planejamento e gestão chamada BSC (Balanced Scorecard).

Para apresentar tais indicadores, então são utilizados os famosos DASHBOARDS, que são painéis visuais (como na Figura 32). Nestes painéis, os indicadores são apresentados de diferentes formas gráficas (linhas, barras, mostradores, mapas, etc). O interessante é que os dados podem ser apresentados em diferentes granularidades de tempo, ou seja, por semana, mês, semestre, ano, etc, e os painéis podem usar mostradores diferentes para cada período (por exemplo, ano a ano). o que permite ao usuário comparar indicadores temporais (ex.: comparar as vendas nos últimos 5 anos, apresentando indicadores ano a ano).

Figura 32: exemplos de dashboards

Também é possível comparar indicadores entre si. Por exemplo, analisar as vendas na semana anterior ao Dia das Mães em comparação às vendas na semana anterior ao Dia dos Namorados. Ou então comparar a lucratividade de cada produto com o grau de satisfação dos clientes em relação a cada produto.

Os indicadores podem ser apresentados como números (ex. total de vendas), escalas numéricas ou nominais (ex.: bom, médio, ruim, inclusive com cores tais como verde, amarelo e vermelho), direcionais (ex.: setas indicando tendência de subida ou descida no número de clientes), mapas (ex: cores indicando níveis de venda por região). Menos comuns mas também úteis podem ser representações de variáveis qualitativas, como por exemplos as *tag clouds* (ex.: palavras mais frequentes nas reclamações dos clientes).

Este tipo de abordagem pode ser considerada **reativa**, pois há uma entrada ou objetivo bem definido e o analista de BI sabe exatamente o que procurar e o que apresentar para o cliente.

A minha crítica a este tipo de processo de BI é que ele é apenas uma evolução dos antigos SIGs (Sistemas de Informações Gerenciais) e dos EISs (Executive Information Systems). A meu ver, o verdadeiro processo de BI deve procurar causas para o que está acontecendo.

Deixemos claro que os SIGs têm seu valor pois ajudam a apontar qual o produto mais vendido, em que épocas saem mais ou menos, qual o melhor vendedor, qual o setor que mais gasta, etc. Mas o verdadeiro BI deve procurar encontrar o porquê de um produto vender mais que outro, de sair mais numa época que noutra, o porquê de um vendedor ser melhor que outro.

Aí então é que entram as técnicas de análise multidimensional ou cúbica (OLAP) e as técnicas de Data Mining. Mas o processo passa a ser um processo de descoberta, como uma investigação ou pesquisa científica. Em outro capítulo, metodologias para tal processo serão abordadas.

Outra forma de fazer BI reativo é analisando a organização, conversando com clientes e usuários e daí então definindo os indicadores. Isto acontece porque muitas vezes o cliente não sabe exatamente o que deve monitorar. Ele tem objetivos ou preocupações (aumentar vendas, diminuir custos, reduzir reclamações de clientes, etc) mas não sabe bem por onde começar. Aí o trabalho do analista de BI é procurar entender que tipo de informações seriam úteis para o gestor atingir seus objetivos. Neste caso, conhecimentos prévios do analista sobre a empresa podem ajudar mas também informações do ramo (por exemplo, coletadas por *benchmarking*).

8 Metodologia para BI proativo

Agora vamos falar de BI **proativo**, uma abordagem não muito comum. Neste caso, a entrada é puramente uma base de dados. O cliente não diz o que está querendo, quais seus objetivos ou problemas, mas apenas informa que deseja encontrar algo interessante nos dados. Este paradigma seria bem representado pela seguinte questão: "o que há de interessante nos meus dados ?".

Neste tipo de abordagem, o objetivo não está bem definido. Ele existe (encontrar algo útil e novo), mas não está claro ou bem detalhado. Isto funciona como uma busca exploratória, onde o analista está procurando encontrar coisas interessantes, sem bem saber por onde ir ou como fazer isto. E não há hipóteses iniciais; o objetivo é justamente tentar descobrir hipóteses para poder depois testar.

Em geral, a falta de hipóteses iniciais se dá porque o usuário ou cliente não consegue definir exatamente o que está procurando. Ele sabe que tem um problema, mas não tem uma ideia exata do que pode ser a solução. É o caso típico de monitorar alguma situação ou encontrar algo de interessante que possa levar a investigações posteriores. Depois que hipóteses são levantadas, o processo pode seguir como no paradigma reativo. Por exemplo, o cliente sabe que há funcionários desmotivados mas não sabe a causa. Ou então um gerente que sabe que as vendas caíram mas não sabe onde procurar as explicações. Ou um diretor que descobre que uma de suas filiais está muito abaixo da média de vendas e não sabe por onde começar sua investigação. Para estes casos, a abordagem proativa deve ser utilizada.

Um dos problemas do paradigma proativo é definir um plano de uso das técnicas ou de como a coleção de dados deverá ser analisada, a fim de serem descobertas hipóteses. Kuhlthau (1991) determinou seis fases em processos de descoberta de informação: iniciação, seleção, exploração, formulação, coleção e apresentação. Cada fase é caracterizada por atitudes diferentes do usuário (por exemplo, em relação a sentimentos, pensamento, ações e tarefas). Uma das descobertas mais interessantes desta pesquisadora é que o usuário inicia procurando algum tipo de conhecimento mais geral, depois ele procura informação relevante em grupos mais restritos e termina procurando informações mais focadas ou específicas. Durante este processo, o usuário reconhece, identifica, investiga, formula, reúne e complementa o conhecimento.

Infelizmente não existe uma máquina de indução, como discutido por Popper, senão seria fácil para analistas de BI, gerentes, etc. A ideia da tal máquina seria que ela aprendesse automaticamente as leis vigentes no universo observando os fenômenos da natureza e daí generalizando comportamentos. Mas como ela não existe (pelo menos ainda), então cabe aos seres humanos fazerem tal processo de investigação e descoberta.

Sugere-se a seguir uma estratégia para análise proativa de dados. Não se pode considerar esta estratégia uma metodologia, mas sim um esboço (*framework*), que poderá conduzir os analistas no processo, indicando os passos principais (técnicas ou ferramentas a serem usadas). Os passos são resumidamente descritos a seguir.

8.1 Seleção de dados e amostras

Como já foi discutido anteriormente neste livro, no capítulo sobre preparação dos dados, o primeiro passo é gerar amostras (mais de uma). Pode-se considerar a base toda como uma amostra, mas certamente devemos também criar subgrupos.

8.2 Seleção da técnica de análise

Uma forma de fazer um processo proativo é utilizar técnicas de Data Mining próprias para tal. As técnicas já foram discutidas anteriormente. O problema agora é saber qual técnica utilizar. Quando apresentamos as técnicas, discutimos algumas formas de aplicação. Se tivermos algumas hipóteses iniciais ou se tivermos um problema bem definido, fica fácil saber que técnica usar. Mas num processo proativo os parâmetros iniciais para se definir que técnica usar é justamente o que está faltando.

Neste caso então, podemos seguir pelo processo de tentativa e erro, usando uma técnica de cada vez e analisando seus resultados para gerar hipóteses iniciais. A escolha do tipo de técnica depende do tipo de dados que temos.

Há valores nominais ou categóricos (ex.: bairro, cidade, profissão, sexo) e numéricos discretos (idade, renda, totais). Os valores numéricos discretos podem ainda ser categorizados por faixas de valor (pelo processo de discretização). Neste caso, podemos usar a técnica de associação pode ser usada para procurar relações entre variáveis. Foi com este tipo de técnica e uma abordagem proativa que o Walmart descobriu que quem comprava cerveja na 6a-feira também comprava fraldas (a famosa lenda do Data Mining).

Outra forma de encontrar relações entre variáveis é utilizar a técnica de correlação ou a técnica de modelos de predição. A primeira indica se há uma relação entre duas variáveis e qual a força desta relação. A segunda gera uma função matemática que possa relacionar os valores das variáveis sendo analisados. Note que a primeira técnica exige duas variáveis, enquanto que a segunda pode ser aplicada a muitas variáveis ao mesmo tempo. É claro que, para utilizar estas técnicas, ou as aplicamos a todas as variáveis e combinações existentes ou possíveis, ou fazemos uma seleção, como discutido em capítulo anterior.

As técnicas de média e detecção de outliers também são simples de serem utilizadas e podem ser aplicadas sobre cada variável em separado (uma por vez).

Já a técnica de análise de séries temporais exige trabalhar sobre uma variável com valores contínuos ao longo do tempo (um valor para cada unidade de tempo, não podendo haver falta de valores num certo período).

A técnica de sequência de tempo exige trabalharmos sobre eventos discretos (acontecimentos), que estejam ordenados cronologicamente.

A técnica de classificação não pode partir do nada, pois exige algum esquema de classificação prévio. Mas as técnicas de clustering e indução podem ser usadas para gerar as regras de classificação.

Já dados temporais (ex.: ano, mês, dia da semana, turno, hora) podem ser utilizados com valores discretos ou contínuos.

8.3 Análise da coleção toda

Neste ponto, o analista deve decidir se irá aplicar as técnicas de descoberta sobre todos os dados ou sobre partes da base; a sugestão é que se comece analisando toda a base e depois sejam examinados subconjuntos. Em alguns casos, nada de interessante é encontrado na coleção toda, o que leva o usuário, necessariamente, a investigar pequenas subcoleções.

8.3.1 Analisar percentual ou valores absolutos

Admitindo que vamos analisar cada atributo em separado, vamos ter informações estatísticas sobre os valores que aparecem associados a este atributo. A frequência de cada valor pode ser apresentada como um valor percentual ou valor absoluto. Por exemplo, analisando o atributo "cidade" numa base de clientes, podemos ter cada cidade apresentada com sua frequência absoluta na base (número de registros em que aparece cada nome de cidade) ou apresentada por percentuais (ex.: a cidade de Bagé aparece em 23% dos registros).

O valor percentual é bom para saber quem predomina num conjunto (os famosos gráficos em pizza). Já o valor absoluto serve para comparar um valor com ele mesmo, em períodos de tempo diferentes. Por exemplo, quantos registros eram de Bagé na medição anterior em comparação à frequência atual.

Ambos os valores são interessantes para saber quem está subindo, quem caiu, quem está surgindo, etc. Entretanto, se o conjunto (número de elementos) aumenta, o valor absoluto não permite saber a relação com outros valores (

Por exemplo, uma empresa notou que reclamações sobre um produto XYZ haviam diminuído em número absoluto, mas em valores percentuais em relação ao conjunto todo, o valor aumentou. Isto significa que as reclamações realmente perderam força, mas que agora este produto era um dos principais em termos de reclamações. A empresa então mudou o foco para este produto, tentando diminuir as reclamações sobre ele (e com isto, tendo como consequência a diminuição do total geral de reclamações).

Lembrando que podemos estar falando de atributos de produtos, empresas, clientes ou outros tipos de atributos como forma de pagamento, mês ou dia, sexo, etc.

8.3.2 Soma X Contagem X Média

Nate Silver conta de uma piada onde um estatístico afogou-se num rio que tinha, em média, 1 metro de profundidade. Ou seja, havia partes mais rasas e outras bem mais fundas.

Já comentamos sobre enganos com a média (Figura 27 e Figura 28). Continuemos com este exemplo, das vendas em um supermercado. A Figura 33 apresenta o total de carrinhos (ou vendas ou notas fiscais) para cada perfil de cliente (dentro de uma determinada amostra). Note que os carrinhos de "famílias" são menos da metade dos carrinhos de pessoas "sozinhas".

Número de carrinhos/compras/clientes	
Contar de CÓD	
Acompanhado	Total
amigos	124
casal	138
familia	152
sozinho	354
Total geral	768

Figura 33: Análise de vendas, utilizando contagem de registros

Soma de valor gasto por cada tipo de cliente (perfil)	
Soma de Valor-gasto	
Acompanhado	Total
amigos	10000
casal	16750
familia	20850
sozinho	37050
Total geral	84650

Figura 34: Análise de vendas, utilizando soma de valores

Para a mesma amostra, a Figura 34 apresenta a soma de gastos de cada perfil. Agora podemos ver que a diferença diminui. Isto porque famílias gastam mais (o que pode ser visto na Figura 28).

A conclusão é que devemos utilizar diferentes técnicas e comparar os resultados. Não há uma técnica melhor que outra. As técnicas existem para apresentar pontos de vista diferentes. O melhor é saber escolher a melhor técnica para cada objetivo ou problema. Se não souber qual a melhor, utilize várias e compare os resultados.

No exemplo dado, a contagem de carrinhos permite descobrir que a maioria dos clientes vêm sozinhos ao supermercado. Já a média de gastos permite ver que famílias gastam mais que os demais perfis. E a soma de gastos pode nos dizer qual o tipo de cliente que mais impacta na receita.

8.3.3 Percentual por linha X por coluna

A Figura 35 abaixo apresenta pedidos de produtos por cidade e por dia da semana. Os valores foram definidos pelo percentual da linha, ou seja, mostra a proporção com que os pedidos foram feitos em cada dia da semana, mas dentro de cada cidade (por isto os 100% estão no total da linha). Este tipo de análise permite descobrir qual o dia da semana com mais incidência de pedidos dentro de cada cidade.

Por exemplo, podemos notar que, na cidade de Uruguaiana, a maioria dos pedidos é feita na 3a-feira, enquanto que na cidade de Itaqui os pedidos predominam na 5a-feira e já na cidade de Bagé há um empate entre 4a e 5a-feira. Também podemos notar que a única cidade que tem predominância na 2a-feira é a cidade de Dom Pedrito.

Já a Figura 36 apresenta os valores percentuais mas por coluna. Isto significa separar os pedidos de cada dia da semana entre as cidades, para ver a proporção dos pedidos entre as cidades (100% está no total da coluna). Isto permite verificar, por exemplo, que na 6a-feira a cidade onde mais são feitos pedidos é a cidade de Uruguaiana (apesar de este não ser o dia de mais pedidos nesta cidade).

Cidade	2a-feira	3a-feira	4a-feira	5a-feira	6a-feira	Total
Bagé	18,2%	21,2%	22,7%	22,7%	15,2%	100,0%
Alegrete	19,4%	22,2%	19,4%	16,7%	22,2%	100,0%
Uruguaiana	16,9%	26,8%	16,9%	18,3%	21,1%	100,0%
Itaqui	16,0%	18,0%	20,0%	24,0%	22,0%	100,0%
Marau	20,0%	20,0%	24,0%	20,0%	16,0%	100,0%
Dom Pedrito	24,4%	17,1%	19,5%	22,0%	17,1%	100,0%

Figura 35: Valores percentuais por linha

Cidade	2a-feira	3a-feira	4a-feira	5a-feira	6a-feira
Bagé	22,2%	22,6%	25,9%	25,0%	18,2%
Alegrete	13,0%	12,9%	12,1%	10,0%	14,5%
Uruguaiana	22,2%	30,6%	20,7%	21,7%	27,3%
Itaqui	14,8%	14,5%	17,2%	20,0%	20,0%
Marau	9,3%	8,1%	10,3%	8,3%	7,3%
Dom Pedrito	18,5%	11,3%	13,8%	15,0%	12,7%
Total	100,0%	100,0%	100,0%	100,0%	100,0%

Figura 36: Valores percentuais por coluna

Os 2 tipos de análise de percentuais, tanto por linha quanto por coluna, são importantes, pois cada um mostra um padrão diferente.

Aqui mostramos o exemplo de vendas por cidade e dia da semana. Mas imagine ter uma base de clientes e cruzar dados como faixa etária (linhas) X bairro (colunas). Podemos fazer o percentual por linha e analisar em que bairro predomina cada faixa etária (por exemplo, jovens estão mais localizados no bairro Praia enquanto que 3a idade está mais no bairro Centro). Ou então fazer o percentual por coluna e assim saber qual a faixa etária que predomina em cada bairro (por exemplo, no bairro XYZ predominam jovens, enquanto que no bairro KLM predominam adultos).

Na amostra do supermercado, extraímos o total de carrinhos que têm algum tipo de brinquedo e classificamos por perfil. O resultado está na Figura 37. Podemos notar que pessoas sozinhas compram mais brinquedos (inclusive que as famílias).

Contar de brinquedos	
Acompanhado	Total
amigos	3
casal	2
familia	19
sozinho	25
Total geral	49

Figura 37: total de carrinhos com brinquedos - por perfil

Entretanto, devemos lembrar que há mais clientes com perfil "sozinho" e isto gera uma tendência. Por isto, fizemos outra tabela, apresentada na Figura 38, onde podemos ver duas colunas referentes a brinquedos: uma que indica o número de carrinhos que tinha algum brinquedo (valor 1) e os que não tinham brinquedos (vazio).

Para facilitar a comparação, a mesma tabela foi reformatada para apresentar valores percentuais (por linha), como está na Figura 39. Agora pode-se ver mais claramente que 12,5% das famílias compra brinquedos enquanto que apenas 7,06% das pessoas sozinhas compram brinquedos.

Contar de CÓD	brinquedos		
Acompanhado	1	(vazio)	Total geral
amigos	3	121	124
casal	2	136	138
familia	19	133	152
sozinho	25	329	354
Total geral	49	719	768

Figura 38: carrinhos com ou sem brinquedos - valor absoluto

Contar de CÓD	brinquedos		
Acompanhado	1	(vazio)	Total geral
amigos	2,42%	97,58%	100,00%
casal	1,45%	98,55%	100,00%
familia	12,50%	87,50%	100,00%
sozinho	7,06%	92,94%	100,00%
Total geral	6,38%	93,62%	100,00%

Figura 39: carrinhos com e sem brinquedos - % por linha

8.3.4 O que predomina

Uma tendência nas análises estatísticas é procurar por valores que predominam. Por exemplo, numa base de vendas, encontrar o vendedor que mais vende, o produto que mais vende, a época em que um produto mais sai, etc. Então a técnica é procurar por valores predominantes em cada atributo.

Outra possibilidade é separar um subgrupo de registros com o valor que predomina (por exemplo, cidade com maior frequência entre os clientes) e aí analisar somente estes registros (clientes de uma determinada cidade). Isto nos permitiria descobrir predominâncias dentro de cada atributo. E isto pode ser feito em vários níveis consecutivos.

Exemplo de uma estratégia de análise de predominância:
a) selecionar clientes da cidade que mais predomina;
b) analisar valores de um atributo específico (ex.: forma de pagamento), dentro deste subgrupo;
c) separar os registros do valor que mais predomina (ex.: pagamentos por cartão);
d) voltar ao passo (b) e selecionar outro atributo, mas utilizando o subgrupo do item (c).

8.3.5 O que é mais importante: o que é raro ou o que é comum ?

Em Business Intelligence (BI), ambos são importantes. Encontrar um padrão que seja muito frequente é ótimo. Por exemplo, um supermercado descobrir que a maioria das pessoas compra feijão na 3a-feira (hipotético). Ou um engenheiro descobrir que 90% das causas de quebra nas máquinas é devidos a mau uso delas. Isto permitirá a estas organizações melhorarem suas estratégias de marketing, investimentos, produção, logística, estoque, vendas, compras, etc. É por isto que uma pessoa que queira compreender um assunto novo irá procurar os livros ou artigos mais citados dentro desta área.

Por outro lado, imagine se o supermercado descobrisse que tem gente comprando feijão no domingo e são uma minoria, talvez duas ou três pessoas. O que isto tem de interessante ? E se o engenheiro descobrir que 1% das quebras são devido a uma única peça ? E se uma pessoa descobrir um livro raro, nunca antes lido ? Ou algum livro publicado, mas pouco vendido ou citado ?

Primeiro, o valor da descoberta pode estar associado ao retorno do investimento (ROI), ou seja, o quanto a informação pode gerar resultados financeiros para a empresa. Por exemplo, aquele 1% de quebras que pode ser evitado ao se descobrir a peça defeituosa pode poupar muito dinheiro para a empresa.

Segundo, algumas raridades de padrões podem suscitar hipóteses para novas teorias. No caso do supermercado, talvez seja interessante fazer campanhas para as pessoas comprarem feijão no domingo e fazerem feijoada em casa na 2a-feira com os restos do churrasco do domingo. Pode ser um novo padrão, ainda adormecido (que precisa ser despertado). Talvez o padrão não seja muito frequente por falta de estímulos. As fábricas de cerveja já descobriram que muitas mulheres bebem cerveja, apesar de serem

a minoria. Mas as propagandas são todas machistas. Então pode estar aí uma nova oportunidade de promoção. São os chamados Nichos de mercado, a estratégia do Oceano Azul. Steve Jobs não perguntou se as pessoas queriam um iPad. Ele fez e foi o maior sucesso.

Terceiro, mas não esgotando as possibilidades, o que é raro pode fazer uma enorme diferença no mundo competitivo. Saber o que ninguém mais sabe, pode ser uma vantagem econômica (veja os investidores nas Bolsas de Valores). Há uma lenda de um inglês que ficou sabendo, durante a guerra entre Inglaterra e França, que a Inglaterra iria vencer. Então ele voltou às pressas para seu país e começou a vender tudo o que tinha. As pessoas, sabendo que ele voltava do campo de batalha, também começaram a vender tudo, achando que a Inglaterra tinha perdido. Aí ele então passou a comprar tudo por baixíssimos preços.

Agir de forma diferente pode chamar atenção (produtos personalizados, novos estilos de moda). O novo gênio do xadrez, o norueguês Magnus Carlsen (o "Mozart do Xadrez") não usa técnicas usuais. Todos grandes jogadores conhecem todas as estratégias. Então o norueguês costuma fazer algo inesperado, fora dos padrões, e isto desconcerta os adversários, que não entendem o padrão, não conseguem prever as próximas jogadas e ficam nervosos. Foi assim que ele deixou nervoso o grande campeão Gary Kasparov.

Na batalha por segurança de informação, para impedir invasões de sistemas computacionais, analistas de segurança com softwares de Data Mining procuram padrões. Mas uma ação nova pode ser uma nova estratégia de ataque.

Por isto, processos de BI devem procurar padrões com alta frequência ou probabilidade estatística, mas os analistas de BI devem também estar atentos a momentos raros, eventos pouco frequentes.

8.3.6 Investigar padrão normal e exceções ou minorias

Uma variação da estratégia descrita no item anterior, seria analisar valores minoritários ou separar um subgrupo de registros com valores que menos aparecem. No caso de valores numéricos, os valores minoritários (*outliers*) podem ser os valores acima ou abaixo da média ou intervalo médio. Por exemplo, se temos uma base de clientes com média de idade num intervalo entre 20 e 60 anos, poderíamos analisar a minoria que tem idade abaixo de 20 ou acima de 60.

Como discutido anteriormente, a análise de exceções ou minorias pode ajudar a encontrar hipóteses de novos conhecimentos. Exceções podem alertar para novos padrões ou especializações dos padrões existentes. Por exemplo, num caso de análise de pacientes com diabetes, foi descoberto um padrão: 95% dos pacientes que tinham o tipo 1 de diabetes estavam recebendo o mesmo tratamento. Um especialista não viu nada de interessante neste padrão, pois é o procedimento normal. O interessante estava justamente com os 5% que eram exceção, ou seja, que tinham o mesmo tipo de diabetes mas não tinham o mesmo padrão de tratamento.

Outro caso interessante de análise de minorias ou exceções (*outliers*) aconteceu numa revenda de carros. A revenda, analisando dados de seus clientes, relacionou

estatisticamente o perfil do cliente com o tipo de carro adquirido. O perfil incluía tipos como “mulheres jovens”, “casais”, “jovens homens solteiros”, etc.

Quando uma exceção ocorre, por exemplo um jovem homem solteiro comprando um carro tipicamente de casais, isto chama atenção, mas ninguém costuma investigar pois é uma exceção. Entretanto, este caso isolado pode ser uma hipótese para novo tipo de comportamento, quem sabe levantando a possibilidade de novas propagandas para atrair novos públicos.

Outro caso interessante aconteceu num site de comércio eletrônico que descobriu que havia muitos homens comprando “chapinha” (para alisar cabelos). Apesar de ser uma minoria que faz isto (a grande maioria dos clientes que compra chapinha é de mulheres), o site resolveu investigar o caso. Constatou-se que eles estavam comprando para presente, mas isto não ficava explícito na hora da compra. Este tipo de informação pode até influenciar de forma errada as campanhas de marketing e os sistemas de recomendação que traçam perfis de clientes. O site então inclui uma opção para o cliente poder dizer que “estava comprando para dar de presente” (e não era para uso do próprio cliente). O mais interessante entretanto é que o site passou a gerar campanhas no dias dos namorados para homens comprarem o tal produto para darem de presente para suas namoradas (e a campanha trouxe bons resultados).

Em várias situações, as exceções são até mais importantes que a regra. Numa investigação criminal, o fato de haver somente uma ligação entre um suspeito e outra pessoa (um possível cúmplice) pode ser mais útil que o caso de o mesmo suspeito ter feito diversas ligações para uma mesma pessoa (por exemplo, um familiar).

Um modo de observar com mais detalhe os chamados *outliers* é tentar relacioná-los com eventos do mundo real. Os picos (subida ou descida) em valores numéricos, como por exemplo os valores extremos em gráficos de vendas, podem ser indicativos importantes para se entender por que as vendas subiram ou cairam tanto. Neste caso, notícias publicadas ou eventos ocorridos no mesmo período (mesmo dia ou dia anterior) podem ajudar a explicar o ocorrido. O ideal seria analisar se tais correlações ocorrem mais vezes, para evitar analisar coincidências ou sincronicidades.

8.3.7 Qual probabilidade mínima é interessante

Se encontramos um padrão estatístico, como vamos saber se ele é interessante ou não ? Um padrão com probabilidade acima de 90% certamente é interessante. Mas pode não ser novo (como o caso do diabetes, relatado antes).

E uma probabilidade de 80% ? E de 70% ? No caso de um valor aparecer em 50% dos registros, isto pode ser interessante, se forem vários valores (por exemplo, cidade do cliente num site de comércio eletrônico que vende para todo o Brasil). Mas se estivermos falando do atributo sexo, 50% não é interessante porque se espera justamente esta divisão num conjunto normal de pessoas.

A sugestão é começar procurando por padrões com alta probabilidade (para não gerar muitos resultados) e depois ir diminuindo. Um valor mínimo ideal não existe. Se houver

um atributo que não tenha um valor com alta frequência (por exemplo, que não apareça em 40% ou mais dos registros), então a probabilidade de 30% pode ser interessante.

Além da probabilidade, é importante ficar atento ao chamado suporte (número de registros onde o padrão ocorre). Por exemplo, uma empresa descobriu um padrão que dizia que 100% (probabilidade) dos distribuidores de uma mesma cidade estavam atrasando 10 dias o pagamento. O problema é que só havia um distribuidor nesta cidade, ou seja, 100% se referia a uma única empresa.

Eu costumo usar o seguinte caso como piada e exemplo: um supermercado descobriu que 100% dos clientes que compravam sapatos de tamanho 48 também compravam o xampu de abacate. Ao saber disto, o pessoal de marketing já começou a pensar em campanhas para aumentar este tipo de venda cruzada. Entretanto, a regra aparecia somente num caso (suporte = 1), ou seja, era somente um cliente que tiha este comportamento.

8.3.8 Medidas de *Interestingness*

O interessante, em geral, é o evento inesperado, que contradiz as expectativas. Pode ser um padrão (ordem) para a maioria dos casos ou simplesmente algo que sai do padrão, como uma exceção.

Descobrir que a maioria dos clientes de um supermercado compra em média 2 kg de feijão é interessante. Mas também é interessante observar quem está comprando abaixo ou acima disto. O que sai da média, o que está fora do previsto, também pode ser interessante.

Para tanto, precisamos de um sistema de crenças, com conceitos básicos ou primitivos ou atômicos que formem um senso comum (ou conhecimento comum ou ordinário). Alguma coisa que, quando solta no ar, sobe contradiz nossos conhecimentos sobre gravidade. Isto é algo interessante que merece ser investigado.

Os povos são cheios de crenças populares e superstições (sabedoria popular). Isto poderia ser incorporado num sistema de crenças, para ajudar a descobrir contradições ou exceções. Ou então a empresa poderia gerar um conjunto de regras de negócio e comparar com padrões encontrados em seus dados. Por exemplo, houve o caso de uma empresa de BPM (Business Process Management) que aplicou Data Mining em nas medições de processos. Ela descobriu uma sequência muito frequente de tarefas que ia contra suas regras de negócio. Ela admitia exceções em alguns processos, mas a exceção ser mais frequente que a regra, isto sim era interessante.

Geng e Hamilton (2006) propõe 9 critérios para determinar se um padrão é interessante ou não. Aí vão eles:

- concisão: um padrão que trata de poucos atributos é mais interessante porque é mais fácil de ser entendido; por exemplo, o que se entende uma regra que diz que “89% dos clientes que compram refrigerante, carne, salada e leite num supermercado, também compram queijo” ? Agora, se a regra for “89% dos clientes que compram presunto também compram queijo”, aí fica mais fácil de se entender o padrão e tomar algumas atitudes;

- cobertura ou generalidade: um padrão é geral se cobre um conjunto grande de dados; é o caso contrário da exceção como já discutido antes;
- confiabilidade: um padrão é confiável se tem suporte maior, ou seja, se ocorre com alta frequência ou percentual (em grande parte dos casos);
- raridade: um padrão é interessante se se distancia muito dos demais padrões (é caso das exceções);
- diversidade: um padrão que trata de atributos bem diferentes dos que são tratados em outros padrões é considerado diverso e por isto tem um certo grau de importância;
- novidade: se um padrão não puder ser inferido de outros padrões, então ele é interessante;
- surpresa: é o caso já comentado de contradizer as crenças ou expectativas;
- utilidade: é útil se contribui para alcançar um objetivo;
- aplicabilidade: se ajuda em alguma tomada de decisão ou em ações futuras.

É claro que o primeiro objetivo do processo de BI é encontrar padrões, não importando de que tipo. O problema é que geralmente um grande número de padrões surgem como resultado, dificultando separar os mais interessantes e consequentemente dificultando a análise e descoberta de conhecimento útil. Por isto, as tais medidas de *interestingness* podem ajudar a filtrar resultados. E isto pode ser feito com auxílio automático de ferramentas de software.

8.4 Comparação de subcoleções entre si ou em relação à coleção toda

Como discutido antes, a separação da base de dados pode ser feita em subconjuntos associados a aspectos temporais (por ano, mês, bimestre, semana, dia da semana) ou , separando os fatos (vendas, clientes, produtos, pedidos, etc) por alguma característica ou atributo.

A granularidade ou unidade temporal a ser utilizada para extrair amostras ou separar a base de dados é importante pois irá influenciar nos resultados. Por exemplo, pode-se separar uma base por ano, mês ou dia. Os padrões encontrados serão condizentes com a unidade de tempo definida. Se dividirmos uma base de fatos por ano e encontrarmos um padrão em um determinado ano, não se sabe se este padrão irá acontecer durante todos os meses deste ano. Sendo assim, talvez tenhamos que analisar mês a mês e aí poderemos saber se o padrão acontece em todos os meses, ou na maioria ou em somente alguns. Por isto, pode ser útil separar a base por turno e não por dia semana, para se observar padrões que acontecem somente de manhã ou somente de noite.

A comparação entre grupos fica mais fácil de ser feita quando as amostras dizem respeito a períodos de tempo. Assim, pode-se comparar vendas entre os meses do ano, ou reclamações a cada ano e entre eles. Isto permite acompanhar as mudanças nos padrões ao longo do tempo e identificar tendências (de queda ou subida), ou mesmo encontrar um padrão que aconteça a cada X anos (por exemplo, vendas de TV têm seu pico a cada 4 anos, coincidindo com os anos em que há Copa do Mundo de Futebol).

Os critérios para separar a coleção em grupos pode ser qualquer um e não somente o

tempo. Por exemplo, podemos trocar tempo por espaço e assim comparar padrões em regiões geográficas diferentes. Ou até mesmo combinar vários atributos. Por exemplo, comparar doenças entre países de hemisfério Sul e Norte a cada ano.

Cada grupo ou amostra pode ser analisado separadamente, mas o interessante é poder comparar os padrões encontrados para cada grupo (internamente) com os padrões de outros grupos ou mesmo com o padrão da coleção toda.

Por exemplo, um processo de análise de reclamações de clientes de uma empresa de TV por assinatura dividiu os clientes (e seus reclamações) por perfil (plano adquirido) e por tipo de programação preferida (pelo canal mais assistido). Esta separação, e a posterior comparação dos padrões entre os grupos, permitiu descobrir que os clientes que mais reclamavam do custo do serviço eram os que tinham o plano mais barato. Da mesma forma, os clientes que menos reclamavam da programação de filmes eram os que tinham como canal preferido algum de filme (os que mais reclamavam dos filmes preferiam notícias ou esportes).

A comparação de padrões entre subcoleções pode ser feita avaliando o que predomina em cada subgrupo ou então buscando saber a probabilidade (ou frequência) de cada padrão em cada subgrupo. Podemos descobrir que um padrão aparece com probabilidade de 90% num subgrupo e com apenas 50% noutro. Ou então podemos verificar o tipo de valor para um determinado atributo que predomina em cada subgrupo. Por exemplo, podemos descobrir que num subgrupo há mais homens e noutro mais mulheres, ou então ficar sabendo que a faixa etária predominante num subgrupo é de jovens enquanto que em outro subgrupo predomina a faixa etária mais velha.

Isto significa tomar cada atributo e avaliar os padrões encontrados para cada um deles em cada grupo e aí comparar os resultados entre os grupos.

Outra possibilidade é descobrir regras de associação (ex.: Se cliente é do sexo X, Então valor gasto está na faixa Y) e aí comparar a probabilidade da regra em cada subgrupo.

Mas também podemos comparar os padrões encontrados em cada grupo com o padrão da coleção toda. Por exemplo, pelo Google Trends, comparamos as buscas pelos termos "dengue" e "gripe A", feitas no Brasil todo, com buscas originadas no Rio Grande do Sul, sobre os mesmos termos e no mesmo período. O resultado está nas Figura 40 e Figura 41.

Os gráficos têm certas semelhanças em alguns períodos, mas são bem diferentes em outros. Pode-se notar que a preocupação com Dengue não é tão grande no Rio Grande do Sul, em nenhuma época, enquanto que no Brasil teve um pico em abril de 2013. Por outro lado, não há no Brasil, como um todo, grandes variações nas quantidades de buscas pelo termo "gripe A", enquanto que no Rio Grande do Sul pode-se ver um período de maior preocupação anterior a outubro de 2012.

Figura 40: Google Trends sobre Gripe A e Dengue no Brasil

Figura 41: Google Trends sobre Gripe A e Dengue no Rio Grande do Sul

A comparação de gráficos correspondentes a grupos permite descobrir grupos com comportamento similar (com correlação) ou inverso. Apesar de ser mais difícil descobrir comportamentos inversos, tais descobertas são importantes e muitas vezes menosprezadas. Se compararmos as vendas entre duas filiais, e os gráficos forem inversos (isto é, quando um está em cima, o outro está em baixo e vice-versa), é possível que elas estejam concorrendo entre si (uma tirando mercado de outro). E isto pode ser feito para comparar vendedores, produtos, lojas, etc.

A descoberta de gráficos inversos pode até mesmo ser interessante para levantar hipóteses que ainda não tinham sido consideradas. Por exemplo, uma indústria gerou gráficos de desempenhos de máquinas (com índices numéricos representando quantidade e velocidade de produção e incluíam as quebras). Cada gráfico correspondia a uma máquina específica. Ao comparar os gráficos entre si, notou-se que dois deles eram exatamente inversos, ou seja, quando a produção de uma máquina estava no máximo, a produção da outra estava em queda. Este tipo de padrão nunca havia sido notado ou mesmo pensado pelos especialistas em manutenção das máquinas. Uma investigação mais profunda descobriu que o cronograma de uso das máquinas estava sendo programado pelo gerente da produção de forma a poupar máquinas aos pares, ou seja, alternando períodos de uso excessivo com períodos mais amenos.

8.5 Combinação e Integração de padrões

Já vimos antes como comparar padrões. Agora vamos discutir como combiná-los, para gerar um padrão único ou um novo padrão. Imagine que haja duas regras com atributos comuns, por exemplo:

a) Se operador = Rudinei Então máquina = computador
significando que as falhas ocorridas numa empresa, quando tinham Rudinei como operador, ocorriam no computador

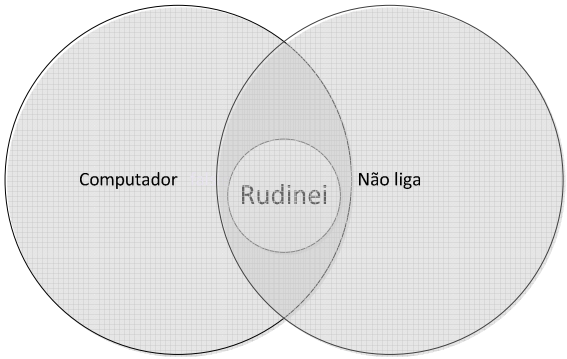
b) Se operador = Rudinei Então problema = não liga
significando que as falhas ocorridas numa empresa, quando tinham Rudinei como operador, eram problemas do tipo "a máquina não liga".

Se ambas as regras possuírem probabilidade de 100%, podemos juntar as duas regras (a) e (b), e teremos que:

Se operador = Rudinei Então máquina = computador E problema = não liga.

ou seja, toda as falhas ocorridas com Rudinei aconteceram no computador e eram do tipo "não liga".

O uso de diagramas de Venn ajuda a visualizar melhor a situação.



Se as probabilidades não forem 100%, não podemos tirar a mesma conclusão, mas pelo menos isto levanta uma hipótese nova que não havia sido considerada ainda, pois os padrões estavam sendo analisados em separado.

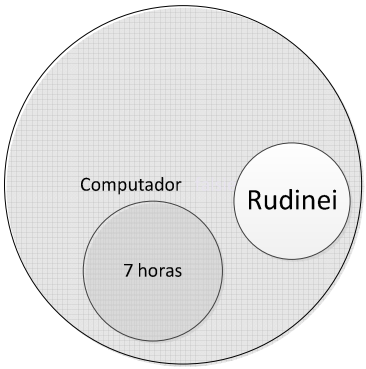
Outro caso:

c) Se operador = Rudinei Então máquina = computador (probabilidade de 100%)
significando que as falhas do Rudinei como operador ocorriam no computador.

d) Se horário = 7 horas Então máquina = computador (probabilidade de 100%)
significando que as falhas que ocorriam às 7 horas da manhã ocorriam no computador.

Aqui, não podemos juntar as regras (c) e (d) porque não sabemos se os casos do Rudinei são comuns aos casos ocorridos às 7 horas ou não.

Como mostra a figura abaixo, pode mesmo acontecer de não haver casos ocorridos com Rudinei e às 7 horas, ou seja, não haver intersecção entre as duas condições (nenhum caso do Rudinei aconteceu às 7 horas e todos os casos que aconteceram às 7 horas não eram com o Rudinei).



8.5.1 Hierarquia de padrões e regras

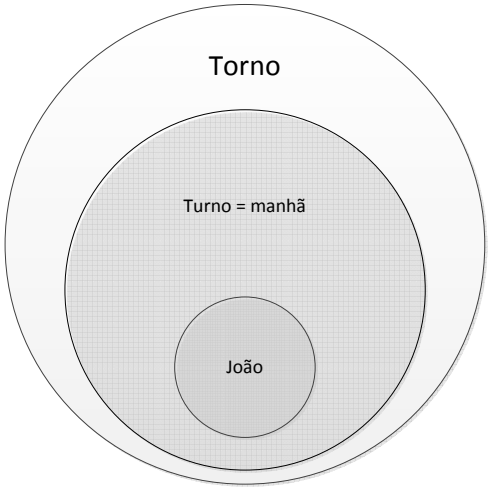
Um caso interessante para comparar ou combinar padrões é quando temos hierarquias. Por exemplo, a regra

"*todas as falhas do João ocorreram no turno da manhã*"

é mais genérica (ou geral) que a regra

"*todas as falhas do João que ocorreram no turno da manhã foram no torno*"

(que é mais específica que a anterior).



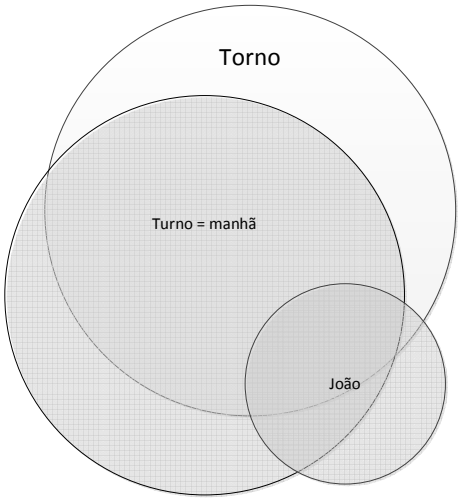
Aqui pode ocorrer de as regras não terem probabilidade 100%, mas o tipo de análise segue o mesmo.

Por exemplo, a regra

"*70% das falhas do João ocorreram no turno da manhã*"

é mais genérica que a regra

"*80% das falhas do João que ocorreram no turno da manhã foram no torno*"



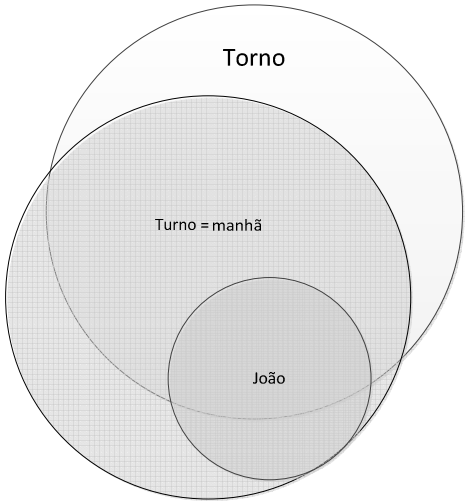
Mas pode acontecer, por exemplo, que a regra mais genérica seja

"*100% das falhas do João ocorreram no turno da manhã*"

é mais genérica que a regra

"*70% das falhas do João que ocorreram no turno da manhã foram no torno*"

Note que a probabilidade das regras (genérica X específica) não necessariamente deva ser maior numa ou noutra.



Agora note que neste último caso, a regra
(e) "*70% das falhas do João que ocorreram no turno da manhã foram no torno*"

pode ser resultado das duas regras a seguir:
(f) "*70% das falhas do João ocorreram no torno*"
(g) "*100% das falhas do João ocorreram no turno da manhã*"

Note o seguinte na regra (e): "*70% das falhas do João que ocorreram no turno da manhã ...* "; mas todas as falhas do João foram de manhã, então pode-se dizer simplesmente que "*70% das falhas do João foram no torno*" (regra "f").

O interessante é que, em alguns casos, podemos suprimir algumas regras e trabalhar somente com um grupo reduzido. O mais indicado é ficar com a regra mais específica (neste exemplo, a regra combinada "e"). Entretanto, a regra mais geral (g) pode ser útil em alguma situação, se não quisermos considerar a máquina onde a falha ocorreu, e a regra (f) pode ser útil se não quisermos considerar o turno ou se já sabemos que o turno da manhã é mais predominante.

Outro caso é seguinte: se descobrirmos que a maior parte das nossas vendas são feitas para clientes do estado de SP, e se soubermos que a maior parte das vendas (todas) são feitas para clientes da cidade de SP, então é melhor ficar com a regra mais específica (a 2a).

A sugestão então é procurar agrupar os padrões por semelhança, ou seja, pelos atributos em comum, e tentar verificar se é possível juntar os padrões ou eliminar alguns, ficando com os mais sucintos. Se for necessário escolher, os autores sugerem ficar com padrões mais genéricos, pois muita especificidade pode gerar sobrecarga.

8.5.2 Regras inversas

Outro caso interessante para avaliar são as regras inversas ou complementares, como por exemplo:
"Se operador = Rudinei Então máquina = computador"
"Se máquina = computador Então operador = Rudinei"

Não necessariamente que elas tenham que ter a mesma probabilidade. Se ambas tiverem probabilidade 100%, temos o chamado "se e somente se".

O interessante é, quando encontrarmos uma regra, verificar a regra inversa, e comparar as probabilidades.

8.6 Avaliação e Teste de Hipóteses

Após terem sido levantadas hipóteses de causas, é necessário avaliá-las ou testá-las, com o intuito de verificar sua veracidade ou a extensão de sua validade.

Uma das maneiras de testar hipóteses é fazer novas observações no mundo real. Por exemplo, se descobrimos um padrão que "*a maioria dos clientes homens com mais de 30 anos adquire o produto X*", basta observar se este padrão aparece em novos casos. Seria a mesma situação que fazer uma previsão ("*o próximo cliente homem com mais de 30 anos irá comprar o produto X*") e verificar se ela ocorre ou não.

Este tipo de validação era muito feita por cientistas no início do método científico e com o surgimento de teorias científicas, segundo Losee). Por exemplo, se temos 3 observações tais que:

A1 é P
A2 é P
A3 é P

podemos inferir uma regra tal que "Todo A é P".

Para validar a regra, temos que saber se todos A realmente são P. Isto significa procurar por As que não são P. Se houver um A que não seja P, então a regra é inválida. Mas como testar com todos os casos ? Isto pode ser muito custoso. Além disto, como vamos saber se conseguimos testar todos os casos ?

Outro problema com este tipo de abordagem, segundo Losee, é que podemos encontrar conclusões verdadeiras mas as premissas serem falsas. Estaremos validando premissas ou regras inválidas. Então teríamos que validar todas as premissas antes.

Mas ter que validar todas as premissas anteriores pode ser muito trabalhoso (validar a premissa da premissa e assim por diante). Para evitar tais problemas, a Humanidade utiliza conceitos e princípios básicos. São definições estabelecidas e aceitas pela comunidade científica. O que Thomas Kuhn chamou de "paradigma". Isto evita ter que fazer regressões infinitas e explanações de todos os princípios. Mas aí surge outro problema (que será discutido adiante), que é justamente não haver mais discussão sobre conceitos básicos. Mas e se eles estiverem errados ? E se até hoje não tivéssemos questionado o heliocentrismo ?

Para testar as novas observações, devemos manter as mesmas condições de quando a regra foi descoberta, ou seja, o mesmo contexto. Por exemplo, no caso anterior, se a inflação subir muito, então é possível que a regra não valha mais. Assim, iríamos considerar a regra inválida, quando na verdade ela vale mas somente numa determinada situação (por exemplo, com a condição de que a inflação esteja baixa). Lembra do famoso padrão em supermercado que dizia que "*clientes que compravam fraldas também compravam cerveja*" ? E se o supermercado mudar a disposição dos produtos e colocá-los em locais próximos, o padrão irá se manter ?

É possível que uma hipótese só valha em certas situações (por exemplo, para alguns tipos de clientes ou produtos, ou somente em alguns períodos de tempo, ou até mesmo só tenham sido verdadeiras no passado, não valendo mais no presente). Então devem ser determinadas as condições ou premissas para validade de uma hipótese.

Outra maneira de validar hipóteses é criar 2 grupos controlados, um que confirme a hipótese e outro que contradiga. Por exemplo, se acharmos uma hipótese de que clientes que gostam de esportes gastam 2 vezes mais que suas esposas, vamos preparar 2 grupos para testar com novas observações. Um com os clientes que gostam de esportes e outro grupo com clientes que não gostam de esportes ou preferem outro tipo de hobby. Se o padrão (gastar 2 vezes mais que suas esposas) só acontecer num grupo, a hipótese é válida. Mas se o padrão aparecer nos dois grupos, então a hipótese não vale. A curva ROC permite comparar resultados (experimentais X observacionais).

A forma então de validar uma hipótese é fazer uma predição e avaliar o resultado com novas observações. O problema pode ser uma questão de tempo entre a predição e seus resultados. Imagine ter que esperar anos para saber se algo que acontece na infância é causa de câncer (Maathuis et al., 2010).

Outro problema da validação pode ser seu custo (tendo que refazer experimentos ou situações). Imagine que foi descoberto um padrão que diz que "*máquinas da marca XYZ quebram mais no verão e quando utilizadas por operadores novatos*". Então, para avaliar esta regra, devemos esperar o próximo verão e colocar um operador novato para operá-la. Mas e se fizermos manutenção preventiva neste meio tempo ? Então o contexto foi alterado, como já discutido no parágrafo anterior. O custo também pode advir de ter que realizar novos experimentos. Por exemplo, ao realizar uma campanha de marketing na TV em horário nobre (ou seja, com custo alto), uma empresa de publicidade descobriu um determinado padrão. Para avaliar o tal padrão, seria necessário repetir a campanha, que já foi custosa. É claro que, se a empresa acredita que a campanha foi boa, ela irá repeti-la e comparar os resultados financeiros. A isto chamamos "taxa de retorno" de campanhas de marketing. E se os resultados da segunda campanha não forem bons ? Então o padrão estava errado. E o custo foi desperdiçado. Mas não havia como testar a hipótese sem refazer a campanha.

Outro exemplo de custo para avaliar uma teoria: uma empresa descobriu que, se os vendedores usassem de um determinado artifício na negociação, a venda seria perdida. Como testar esta hipótese ? Refazendo o modo considerado errado ? Neste caso, o normal para qualquer pessoa é evitar repetir o erro. Mas se a teoria estiver errada e não for mesmo um erro tal procedimento ?

Uma alternativa seria utilizar software para simulação e testar as hipóteses. Para isto, precisamos do modelo real e parâmetros. Por exemplo, equipes de Fórmula-1 utilizam simuladores chamados túneis de vento para testar o design do carro. Utilizando leis da Física e modelos computacionais do carro é possível avaliar sua performance dentro do computador, sem precisar de um túnel de vento real ou um carro em tamanho real.

Outra questão a ser pensada quando se quer avaliar uma hipótese, é definir o período pelo qual a hipótese será avaliada. No mesmo exemplo anterior, devemos considerar quantos clientes do mesmo tipo ? Ou devemos considerar todos os clientes que fizerem compras nos próximos "n" dias, mas qual o valor de "n" ?.

Outra técnica de validação de hipóteses é a "redução ao absurdo" (de Euclides e Arquimedes, segundo John Losee). Para confirmar que um padrão ou regra é inválido, bastaria inferir ou derivar um fato que fosse uma contradição da regra ou um resultado absurdo. Por exemplo, imaginemos que um sistema de Data Mining automaticamente

descobre que todos os pacientes que são tratados na ala sul do hospital recebem, como procedimento cirúrgico, uma cesariana. Se conseguirmos encontrar um paciente do sexo masculino que foi tratado na ala sul, a regra então não é mais válida.

Assim, uma forma de invalidar uma regra é encontrar um caso que seja exceção, ou seja, onde a regra não se aplica. Entretanto, exceções existem aos montes e aí estaríamos simplesmente desconsiderando a regra. O que pode acontecer é diminuir a probabilidade da regra. Por exemplo, se encontrarmos uma regra que diz que "*clientes homens entre 20 e 30 anos praticam algum tipo de esporte*", talvez ela não valha para 100% dos casos. Se houver exceções neste caso, elas não invalidam a regra mas somente diminuem sua força.

Conforme Popper, é fácil obter confirmações ou casos positivos; basta procurá-los. Entretanto, as confirmações (casos positivos) só devem ser consideradas como prova se resultarem de predições arriscadas ou pouco prováveis. Popper também argumenta que toda teoria ou modelo é de certa forma uma proibição (ela proíbe certas coisas de acontecerem). E assim, quanto mais a teoria ou modelo proíbe, melhor ela é. Popper encerra dizendo que a Astrologia e algumas teorias psicológicas aceitam e explicam tudo, e portanto não devem ser consideradas teorias científicas.

Uma anomalia na refutação de padrões ou teorias é eliminar também alguns efeitos positivos. Houve um caso numa empresa que descobriu que certas reuniões eram desnecessárias para atingir alguns objetivos. Então, para diminuir custos, aquele tipo de reunião foi cancelada. Os objetivos continuaram a ser alcançados e os custos diminuíram. Entretanto, as tais reuniões traziam benefícios paralelos e ajudam em outros objetivos, os quais tiveram perdas com o fim destas reuniões. Apenas após alguns meses, a tal anomalia foi detectada. A solução não foi trazer de volta aquelas reuniões mas utilizar outros tipos de procedimentos para substituir as reuniões e obter os mesmos resultados paralelos.

O problema todo é que vivemos num mundo cada vez mais complexo. Há muitos padrões, mas também muitas exceções. Há muitos efeitos colaterais, positivos e negativos. Conseguir mapear todas estas influências é uma tarefa muito difícil.

Um perigo na avaliação de hipóteses é querer acomodar fenômenos (dados observados) no modelo proposto. John Losee e muitos cientistas chamam a isto de "*salvar as aparências*". Isto pode ser feito distorcendo dados e observações para confirmarem hipóteses, ou mesmo escondendo ou minimizando exceções. Há o famoso caso de um vidente que previu tantas catástrofes, que precisou forjar algumas para não passar ridículo.

8.7 Retroalimentação

Todo processo de descoberta e investigação é cíclico, ou seja, alguns passos ou mesmo o processo todo devem ser refeitos. Quando hipóteses são descobertas, é necessário validá-las. Confirmadas ou não, é necessário voltar e refazer o processo para descoberta de novas hipóteses e continuar o ciclo.

Muitas vezes, até mesmo para validar uma hipótese é necessário refazer o processo, mas aí utilizando uma abordagem reativa e não proativa.

Em outros casos, é necessário refazer o processo várias vezes para gerar um conjunto grande de hipóteses ou mesmo de conhecimentos já validados, para que possam ser combinados (já discutimos anteriormente como integrar e combinar padrões ou regras).

9 Processo de BI como Descoberta e Investigação

Encontrar porquês é uma característica típica de processos de investigação e descoberta. Isto inclui a investigação científica, a descoberta de fontes de recursos naturais, o diagnóstico médico, a busca por causas de efeitos ou problemas, o planejamento de recursos para atingir objetivos.

A maioria dos pesquisadores concorda que o processo de descoberta é cíclico, tendo como passos principais: (Agrawal e Imielinski, 1993; Parsaye et al., 1989; Ingwersen, 1996):
a) a formulação de hipóteses;
b) o teste das hipóteses;
c) a observação dos resultados (para refutar ou confirmá-las);
d) a revisão das hipóteses e a sua modificação (reiniciando o processo), até que o usuário se dê por satisfeito.

Portanto, a estratégia inicia com a geração de hipóteses iniciais. Hipóteses são roteiros para direcionar a investigação ou o processo de descoberta e análise. Elas sugerem que dados coletar e analisar. Se estivermos investigando causas de acidentes de trânsito, podemos começar pesquisando quantos acidentes foram causados por condutores embriagados. Isto não significa que a causa principal é esta ou que iremos somente nos concentrar neste tipo de causa. A avaliação desta hipótese inicial pode até mesmo nos desviar para a causa real, caso se descubra que esta causa inicial não é muito frequente.

Entretanto, levantamento de hipóteses ou mesmo sua investigação não é uma tarefa simples. Não é um algoritmo ou programa com passos bem definidos. O que há é uma estratégia (algo como um *framework*). Edgar Morin (2000, p.90) nos avisa: "*a estratégia deve prevalecer sobre o programa. O programa estabelece uma sequência de ações que devem ser executadas sem variação em um ambiente estável, mas, se houver modificação das condições externas, bloqueia-se o programa. A estratégia, ao contrário, elabora um cenário de ação que examina as certezas e as incertezas da situação, as probabilidades, as improbabilidades. O cenário pode e deve ser modificado de acordo com as informações recolhidas, os acasos, contratempos ou boas oportunidades encontradas ao longo do caminho. Podemos, no âmago de nossas estratégias, utilizar curtas sequências programadas, mas, para tudo que se efetua em ambiente instável e incerto, impõe-se a estratégia.*"

Este levantamento exige conhecimento do domínio, criatividade e certas habilidades que talvez não possam ser aprendidas em cursos ou manuais. Muitas vezes, inventores, investigadores, criadores, etc, conseguem chegar a soluções de problemas por *insights*, que não podem ser explicados, de onde vieram, ou como foram gerados (tópico que será discutido mais adiante neste capítulo).

O conhecimento prévio sobre o assunto ou domínio é importante, bem como estar ciente do contexto, parâmetros, limitações e condições em que a investigação ocorre. Entretanto, o conhecimento é subjetivo, flexível, mutável e depende das pessoas. Por isto, Moscarola e Bolden (1998) sugerem o modelo construtivista ao invés do positivista

para os processos de descoberta. Isto é, o processo deve ser de construção e guiado por um especialista humano. A construção forma-se a partir de fundamentos, que podem ser dados novos ou conhecimentos e teorias prévias. E vai se desenvolvendo com aprendizados, erros e correções. Os caminhos podem ser refeitos, os objetivos redirecionados, hipóteses novas podem surgir, anteriores podem ser refinadas, refeitas ou mesmo descartadas. As conclusões iniciais devem ser validadas. As primeiras nunca devem ser tomadas como verdadeiras de imediato.

Este capítulo se concentra no problema de descoberta de hipóteses iniciais.

9.1 Descobrindo hipóteses de causas

Não devemos esquecer que primeiro vêm as observações e depois as hipóteses, senão é ficção científica ou invenção. BI é procurar por porquês, explicações, causas, padrões. Mas eles só surgem se houver pistas anteriores. E estas hipóteses iniciais nem sempre são as conclusões ou respostas definitivas. Podem ser simplesmente um passo inicial para algo bem diferente. Já houve casos de teorias erradas que levaram a objetivos certos. Lembre do caso da navegação de Cristóvão Colombo.

Segundo Popper, existem boas e más teorias. Só precisamos saber distingui-las. A ideia é começar com qualquer teoria, ir testando e melhorando-a. Não é ruim fazer tentativas. Pode ser custoso, mas o processo é de construção e aprendizado.

Pior seria não ter hipóteses para começar. Conforme Clarke e Eck, a falta de hipóteses pode gerar "paralisia de análise", conduzindo a investigação a lugar nenhum.

Podemos começar pelas teorias já conhecidas em outros ramos e verificar se se aplicam no nosso contexto (as chamadas analogias). Ou então procurar por hipóteses bem diferentes. Popper sugere que teorias mais prováveis são pouco interessantes, porque possuem pouco poder de explicação. Entretanto, o objetivo não é se preocupar com a probabilidade da teoria, mas sim com seu poder para explicar fenômenos.

A coleta inicial de dados

Na investigação criminal, como nos seriados CSI, os investigadores primeiro procuram por evidências, pistas ou sinais. O mesmo ocorre no diagnóstico médico: o médico analisa primeiro sinais, sintomas, queixas. Existem evidências primárias e evidências secundárias. Por exemplo, num crime, os elementos que estão disponíveis na cena do crime são elementos primários. Mas há informações importantes como endereços visitados por vítimas e suspeitos, amigos e relações profissionais, ligações telefônicas, etc., que compõem as chamadas evidências secundárias.

Se não tivermos dados iniciais, ou não soubermos por onde começar coletando dados, a famosa estratégia dos 5W e 2H funciona bem para este início de processo:

– What (o que aconteceu),
– Who (quem fez ou participou ou foi vítima ou prejudicado),
– When (quando aconteceu o fato),
– Where (onde ocorreu o evento),

– How (como ocorreu o evento),
– How Much (quanto: quantificar algumas variáveis já conhecidas).

Um dos W (Why - por que ocorreu) será deixado de fora neste início, já que é justamente o alvo da investigação.

A ordem dos eventos também pode influenciar o resultado. Por exemplo, vendedores com as mesmas ações podem ter tido resultados diferentes; e isto pode ser devido à ordem das ações. A aplicação da técnica de Data Mining para análise de Sequências de Tempo, apresentada na seção 5.1, pode ser útil para descobrir padrões em relações sequenciais entre eventos (uma ordem significativa de acontecimentos).

Estes primeiros passos geram volumes grandes de dados. O famoso Big Data. Deve-se ter ferramentas próprias para registrar os dados, seus relacionamentos, de forma a facilitar a análise e filtragem posteriores.

Não se deve fazer filtragens no início. Tudo é importante; nenhum dado deve ser descartada ou menosprezado. Lembre do efeito Borboleta no clima. O mesmo se diz para relações entre variáveis. Tudo deve ser anotado para análise futura. Lembre que um supermercado descobriu uma relação entre as vendas de fraldas e cervejas, o que a princípio pode parecer um absurdo.

Quantidade de informação X sobrecarga X ruídos

A quantidade de informação é importante. Quanto mais informação melhor. Mas em muitos casos a quantidade pode gerar sobrecarga, tirando ou desviando o foco de causas importantes. Já comentamos antes os problemas do excesso de variáveis em análises estatísticas no baseball, muito discutido no livro Moneyball de Lewis. Houve também um caso de uma empresa que usava 50 variáveis para diferenciar perfis de clientes. No final, descobriram que apenas 5 atributos seriam suficientes para distinguir as principais classes.

Aqui não podemos deixar de lembrar os estudos de George Miller, na década de 50, sobre o número mágico 7 mais ou menos 2. Resumidamente, para quem não conhece esta teoria, ela diz que o ser humano normal tem a capacidade para gerenciar de 5 a 9 subsistemas (7-2=5 e 7+2=9). Então, se tivermos que dividir um sistema em partes, o melhor é que ele tenha de 5 a 9 partes. Na prática, podemos pegar o exemplo de montar uma equipe de trabalho. Se ela tiver mais de 9 pessoas, o líder do grupo terá dificuldades para gerenciar todos. Se o grupo tiver menos de 5 pessoas, haverá capacidade ociosa. E estas conclusões perduram até hoje. Ninguém ainda fez um estudo capaz de contradizer as conclusões de Miller. Concluindo, o ideal é trabalhar com um número reduzido de variáveis e este número poderia ficar entre 5 e 9.

Nate Silver escreveu um livro todo (e não é pequeno) para tratar do problema de informações a mais que acabam deturpando as análises e desviando pessoas do objetivo. Ele caracterizou tais dados como ruídos e chamou de "*overfitting*" o engano de interpretar ruído como sinal. Por isto se deve a importância de saber filtrar, armazenar, buscar corretamente, fazer resumos, interpretar e saber distinguir o que é importante.

Silver declara que um dos maiores riscos na era da informação é que a massa de conhecimento no mundo está aumentando (e exponencialmente). Então a diferença entre o que sabemos e o que pensamos saber pode estar aumentando. E como consequência temos um crescente aumento de stress, porque as pessoas querem e precisam saber e armazenar mais informações e conhecimento.

A observação é direcionada, seletiva

Segundo Darwin, "*ninguém pode ser bom observador se não tiver uma teoria antes*". É preciso direcionar o foco da observação, porque pode haver muita informação. Isto não significa apaixonar-se pela teoria e não enxergar outros caminhos. Darwin mesmo tinha algumas teorias iniciais (vindas de Lamarck) que acabou refutando com suas descobertas.

Se estivermos numa aula e pedirmos aos alunos para "observarem", eles perguntarão "observar o quê ?". Se estivermos numa cidade nova com fome, a observação será para encontrar algum lugar para comer. Se estivermos sem compromisso, talvez nos interessemos pela arquitetura e pelo ambiente. Se estamos procurando uma pessoa, só vamos olhar para pessoas. O ser humano recebe muitas informações pelos 5 sentidos, externas e internas, mas não dá relevância a tudo. E nem pode. Para evitar a sobrecarga, é preciso fazer filtros e selecionar dados.

Koestler diz que o "bom observador" é aquele que direciona suas observações. Popper diz que usamos quadros de referência. Somos condicionados pelas necessidades e vontades, primeiro momentâneas, depois relativas a nossa expectativa de futuro, mas isto tudo moldado pelo nosso passado. O passado pode ajudar, acelerando buscas, eliminando lixos. Mas pode nos condicionar por um vício de interesse e fechar nossos olhos a novas observações. A filtragem pode ser boa para evitar o acúmulo de grande volume de dados. Mas pode ser ruim, por deixar coisas importantes de fora da análise.

Por isto, devemos usar técnicas e nossa experiência para saber selecionar e filtrar dados. Um dos auxílios pode ser o uso de ferramentas de software, para ajudar no armazenamento, recuperação e seleção de dados. O ser humano possui limitações para estas tarefas. O computador não é tão inteligente. Mas a parceria de ambos pode ser uma solução ótima.

Bancos de dados e planilhas são úteis para armazenar dados estruturados. Há formas diferentes de recuperação, começando pelas mais técnicas como a linguagem SQL ou XQUERY ou XPATH (para XML), mas também podemos usar classificações (taxonomias) e consultas por palavras-chave (como o Google). A grande dificuldade está em lidar com dados não estruturados (discutida mais adiante), tais como textos, imagens e sons.

As ferramentas para visualização de dados ajudam a gerar resumos e filtros visuais. Há diversas formas diferentes de ver os mesmos dados. Isto nos dá pontos de vista diferentes, como já discutido antes, quando falamos de dados multidimensionais. No link abaixo, a Universidade de Maryland apresenta diversas técnicas que estão sendo pesquisadas para visualização de informações.

http://www.cs.umd.edu/hcil/research/

E há também o livro de Jacques Bertin, sobre o assunto.

A intuição para seleção de dados

A intuição é um palpite, mas não uma adivinhação. Ela deve ser precedida por dados. A questão é que a intuição acontece numa decisão sem muita explicação de onde veio, se ela está certa ou não ou por que devemos utilizá-la. É saber algo sem saber explicar como. Max Gunther acredita que usamos dados do inconsciente, que foram colhidos e armazenados antes, mas que não temos consciência de quando os estamos usando. É como reconhecer um amigo na rua ou a voz de alguém no telefone. Não tem explicação, mas a gente faz e na maioria das vezes não erra.

Simon (1972) apresenta a teoria da racionalidade limitada nas decisões. A premissa é que as pessoas procuram tomar decisões de forma racional, analisando dados, usando a lógica, etc., mas nem sempre isto acontece na prática. Em parte, o processo de decisão é limitado por não termos todos os dados disponíveis, ou por eles estarem incompletos, ou por não sabermos se são verdade ou não. E na maioria das vezes, não vale a pena coletar todos os dados necessários e verificá-los. Por exemplo, se uma pessoa quiser comprar um sapato, pensará em verificar na cidade qual a loja com o preço mais barato. Entretanto, se for avaliar o preço de cada loja, ao terminar o processo, terá levado tanto tempo que os primeiros preços consultados já poderão ter sido alterados e o custo total de deslocamentos e perda de tempo não valerá o desconto que conseguir. É impossível que o indivíduo conheça todas as alternativas para uma decisão e que possa avaliar todas as suas consequências. A tendência do ser humano é simplificar as escolhas. Isto quer dizer que não temos como saber se a decisão tomada foi a mais acertada antes de tomá-la; somente após saberemos se deu certo ou não. E mesmo tendo alcançado êxito, talvez não tenhamos certeza se foi a melhor alternativa.

Malcolm Gladwell, no livro Blink (2005), fala de experimentos de psicólogos analisando vídeos de casais conversando e tentando prever se o casal iria continuar junto ou não depois de 15 anos. Ao analisar 1 hora de vídeo, eles conseguiram uma acurácia de 95%, enquanto que analisando apenas 15 minutos de vídeos, atingiram 90% de precisão nas predições. Ou seja, não são necessários muitos dados nesta situação. Padrões podem ser identificados em resumos. Gladwell também comenta sobre técnicas utilizadas por americanos para reconhecer operadores alemães de código Morse. Como saber distinguir operados numa tarefa tão rápida como transmitir código Morse ? É algo que não pode ser explicado conscientemente.

Uchida, Kepecs e Mainen (2006) concluem que as pessoas vão acumulando dados, a partir de experiências e sentidos, os quais vão sendo agregados até o momento em que uma decisão é tomada. Mas isto acontece em frações de segundo.

Wilson (2004) discute o inconsciente adaptativo, um sistema de percepção não consciente, que utiliza funções de menor ordem (percepção, compreensão da linguagem), ao contrário de funções de alta ordem, envolvendo raciocínio. Segundo Wilson, nossos sentidos recebem 11 milhões de pedaços de informação num dado momento, nossos olhos recebem e enviam para cérebro 10 milhões de sinais a cada segundo, mas só conseguimos processar 40 partes de informação por segundo, de forma

consciente. Por exemplo, se você pedir para um pianista explicar que sequência de teclas ele usa numa música que saiba tocar sem partitura (de memória ou de cabeça), dificilmente ele conseguirá explicar, ou pelo menos, levará um bom tempo tentando relembrar. Mas no momento de tocar a música, a sequência vem sem ele precisar pensar sobre isto.

A intuição também é utilizada, segundo Gunther, sem a necessidade de pressa. Ela não deve ser confundida com caminho mais fácil (preguiça). Gunther não recomenda confiar na primeira impressão, mas sugere que coletemos muitos dados.

O hábito e a experiência para seleção de dados

A experiência, o hábito pode ajudar a aprimorar o uso de intuições, tanto para filtragem do que coletar quanto para seleção de hipóteses ou causas prováveis.

O hábito é uma vantagem quando não há tempo para raciocinar. Ele nos ajuda a tomar as decisões certas. Mas deve ser treinado, para não ser usado como sorte ou preguiça. Por exemplo, o jogador que assume a função de líbero num time de vôlei, quando ele faz uma defesa, ele não pensa conscientemente. A reação é em milésimos de segundos. Mas seu cérebro precisa tomar decisões quanto a posicionamento do corpo (pernas, braços, mãos, etc.), para rebater a bola para frente, em direção ao meio da quadra, sem passar a rede e sem ficar muito perpendicular a ele mesmo. Para tanto, ele vai dispor braços, mãos e restante do corpo, mas a decisão não é consciente. As decisões rápidas (e acertadas) neste caso vêm devido a treino (a força do hábito). É como digitar um texto num computador sem olhar para o teclado. Se perguntarmos a uma pessoa que digite textos rapidamente onde fica uma determinada letra, ele terá que parar para pensar. Mas se pedirmos para ele digitar uma palavra, esta sairá rapidamente. O mesmo com um piloto de corrida. As decisões são tomadas rapidamente, parecendo ser instinto, mas na verdade é um hábito que foi muito treinado.

Outro exemplo de hábito ou habilidade muito treinado é o caso de Ayrton Senna correndo na chuva. No início de carreira, no kart, ele não sabia andar na chuva. Então começou a treinar exaustivamente até que pudesse fazer disto um hábito, ou seja, uma habilidade que ele desempenhava sem precisar pensar (eram decisões rápidas).

Kahneman fala da importância de praticar o hábito. Ele afirma que os grandes jogadores de xadrez não veem o mesmo tabuleiro como um novato. Eles conseguem visualizar jogadas possíveis pela força do hábito. O treino gera uma habilidade para acessar mais rapidamente certas informações no cérebro e organizá-las melhor. Gladwell, no seu livro sobre Outliers (2011), chega a um número mágico de 10 mil horas de treino, que distingue os grandes campeões dos demais. Ele vê isto em grandes músicos e jogadores. Se uma pessoa treinar 8 horas por dia, todos os dias, sem folgas, precisará de 3,4 anos para chegar a este número. É por isto que podemos notar que grandes campeões de esportes ou músicos virtuoses começaram com pouca idade.

Duhigg (2012) diz que podemos instalar hábitos em nossos cérebros. Eles ficam armazenados em áreas específicas do cérebro e podem ser recuperados de forma inconsciente. Começa com um estímulo que manda o cérebro entrar em modo automático, e indica qual hábito deve ser usado. As recompensas (dor, prazer, etc.)

ajudam o cérebro a saber se vale a pena memorizar este hábito para o futuro ou não. A sugestão é definir um plano para uma rotina que traga a mesma recompensa. Para o líbero do vôlei, seria treinar exaustivamente defesas e recompensar com felicidade ou tristeza cada resultado. Recompensas diferentes ajudam o cérebro a diferenciar ações boas de ruins.

Heurísticas para seleção de dados

Regras heurísticas orientam decisões mesmo sem garantir resultados. Não são algoritmos ou procedimentos. Devem ser usadas conforme a situação do momento.

Em muita decisões, não há informações suficientes para uma boa escolha. Por exemplo, ao chegarmos a uma encruzilhada, no caminho em direção a um destino, e se não tivermos um mapa, vamos usar heurísticas para escolher o caminho a seguir. Talvez alguém olhe para o céu, e mesmo sem saber orientar-se por ele, tenha um lampejo de informação, lembrando de uma situação semelhante em que ficou perdido. Outros olharão para o chão, lembrando situações que viram num filme (nunca experimentaram a mesma situação mas reusarão soluções que foram úteis para outras pessoas).

Algumas heurísticas estão enraizadas no ser humano como hábito ou instinto. Por exemplo, segurar coisas que caem, fugir do fogo ou de animais que rosnam. Mas as heurísticas também são usadas para acelerar a solução de problemas. Conforme, Gigerenzer e Gaissmaier (2011), uma heurística é uma estratégia que ignora parte da informação com o objetivo de fazer decisões mais rápidas do que métodos complexos. Em casos onde não há tempo para pensar, as heurísticas podem funcionar.

Por outro lado, como as heurísticas são usadas sem consciência, podem gerar resultados catastróficos, quando seria melhor raciocinar sobre alternativas. Gladwell no livro Blink (2011) descreve o caso de um bombeiro que sobreviveu a um incêndio na floresta parando para pensar numa solução, enquanto que seus companheiros não tiveram a mesma sorte porque seguiram seus instintos (ou heurísticas).

Lenat (1982) diz que as heurísticas podem ser construídas por especialização ou por generalização. Por exemplo, se uma decisão foi útil numa caminhada por uma floresta, é possível que também seja útil em qualquer tipo de caminhada (generalização) ou em caminhadas menores (especialização). A força das heurísticas está na analogia que proporcionam. Se uma heurística H foi útil numa situação S, então heurísticas similares a H serão úteis em situações similares a S (analogia). Entretanto, se o ambiente muda rapidamente, as heurísticas possuem pouco tempo de vida.

Em resumo, as heurísticas funcionam bem para ajudar o raciocínio, eliminando alternativas quando há muitas e não há tempo ou recursos suficientes para avaliar todas.

A observação influencia o ambiente

Nate Silver comenta sobre o princípio da incerteza de Heisenberg: “assim que começamos a medir algo, seu comportamento começa a mudar”. Em muitos casos que envolvem atividades humanas, o próprio ato de observar pode alterar o comportamento

das pessoas. Se as pessoas souberem que estão sendo avaliadas ou observadas, mudam seu comportamento (para melhor ou pior).

Hoje em dia se discute muito no Brasil se as pesquisas de opinião para eleições influenciam ou não os que votam. Conforme vamos coletando dados e formando hipóteses, nosso conhecimento vai mudando, vai-se moldando. Não há como impedir tal modificação. Isto pode nos direcionar na coleta de mais dados, fazendo-nos eliminar certas hipóteses ou circunstâncias, ou fazendo com que nos atenhamos mais a certos detalhes.

Como já dito antes, não é errado formular hipóteses iniciais. O problema é só ficar com estas e descartar outras possibilidades.

Fazer as perguntas certas

Conforme Koestler, o que diferenciou Darwin de outros pesquisadores que acreditavam e estudavam a teoria da evolução foi que conseguiu provar a teoria com o seu porquê e como. Mas para isto, ele precisou fazer as perguntas certas. Neste caso, por que as espécies evoluíam e como (origem das modificações e como passavam entre as gerações). Além disto, ele foi atrás de fatos para explicar sua teoria.

Fazer as perguntas certas significa coletar e armazenar os dados certos, ou seja, já ter algumas hipóteses do que pode ser a causa ou o que pode influenciar. Se a causa para quebras de máquinas é a temperatura ambiente, então temos que coletar estes dados e inseri-los na base de dados para depois poder utilizar as técnicas de análise com ajuda de software. Se esta for a causa e tais dados não estiverem na base, ou não descobriremos nunca a causa ou então estaremos calcados em descobertas enganosas.

Detalhes podem fazer a diferença. O ser humano tem a tendência de analisar o que é comum, mais frequente, o que aparece mais. É assim com a moda. Ninguém dá atenção para um tipo de acessório que só uma pessoa usa. Se vários estiverem usando o mesmo estilo, isto chama a atenção das pessoas comuns. Entretanto, num processo de descoberta ou investigação, os pequenos sinais podem ser muito úteis. Pergunte a um investigador policial. Então, num primeiro momento nada deve ser descartado. Todos os dados possíveis devem ser coletados e analisados. Todos os caminhos devem ser considerados. E várias hipóteses iniciais devem ser construídas.

Descobrir as hipóteses iniciais é um processo de tentativa e erro. Podemos acelerar com analogias e benchmarking, como será discutido adiante. Mas muito provavelmente será necessário refazer o processo de descoberta, analisar novos padrões ou outras causas possíveis, gerar novas hipóteses, testá-las com casos reais e aí refazer tudo de novo.

Visão Holística - Análise do Contexto

A visão holística significa a "Visão do Todo", ver todos os elementos e suas relações. Isto ajuda a entender como o todo (problema) está composto e pode ajudar a direcionar o foco ou mesmo ver detalhes pouco percebidos.

Procure observar as interações, não só estabelecendo as conexões entre os elementos mas entendendo que tipo de conexão existe. X pode estar conectado a Y por ser sua causa, mas pode estar conectado a Z por que são ideias contrárias e pode estar conectado a W por outra razão diferente. Não estabeleça regras de tipos de conexões, não fique preso a paradigmas, tenha mente aberta.

Os gregos só conseguiram entrar em Troia porque estudaram o povo troiano. Se tivessem visto o todo (problema) somente como uma cidade-fortaleza com muros altos, poço de fogo, portão forte e guerreiros, estariam até hoje tentando entrar. A ideia do Cavalo de Troia veio porque eles entenderam que o problema incluía o povo troiano, e este detalhe fez a diferença. Eles descobriram que o povo troiano era supersticioso, muito religioso e acreditava em presentes dos deuses. Daí veio o *insight* da solução.

Visão holística também tem a ver com Sinergia (o todo é maior que a mera soma das partes). Se ao analisar a molécula de água (H2O), observássemos os elementos hidrogênio e oxigênio em separado, não saberíamos que o estado natural da água é líquido. Quando os elementos de um todo interagem entre si, formam um sistema complexo que pode levar a resultados imprevisíveis. Só listar os elementos não é suficiente; temos que entender as relações entre eles.

Segundo Morin (2000, p.42), até meados do século XX, a maioria das ciências obedecia ao princípio de redução, que limitava o conhecimento do todo ao conhecimento de suas partes, como se a organização do todo não produzisse qualidades ou propriedades novas em relação às partes consideradas isoladamente. A sinergia se resume em dizer que 1 + 1 = 3. Quando as partes se juntam, podem formar algo novo e bem diferente. Foi assim que a primeira forma de vida deve ter começado segundo a teoria evolucionista.

É preciso entender o contexto e coletar dados externos. Lembre que o mundo não é fechado, como já discutido na seção 6.3 Lembre do exemplo das vendas de laranja (Figura 30).

Não entender isto é como procurar a causa para defeitos num carro observando apenas o comportamento do motorista e as peças, sem olhar para a estrada, o clima, o que outros motoristas fizeram, etc. Houve um caso interessante numa cidade do interior do Rio Grande do Sul. Estavam acontecendo suicídios entre agricultores em número fora dos padrões normais. As primeiras hipóteses levavam para investigação de aspectos sociais, como família, ambiente social onde trabalhavam, perspectivas econômicas, etc. Depois descobriu-se que o uso excessivo de agrotóxicos estava influenciando o organismo e o lado psicológico dos agricultores.

Em biologia, diz-se que os ecossistemas são formados pela união de dois fatores:
a) fatores abióticos: conjunto de todos os fatores físicos que podem incidir sobre as comunidades de uma certa região (ex.: luz, temperatura, chuva, tipo de solo);
b) fatores bióticos: conjunto de todos seres vivos que interagem numa certa região

Se estivermos lidando com sistemas biológicos, temos que considerar estes aspectos. E aí a sinergia é bem maior e complexa. Há o famoso caso da guerra dos pardais na China em 1958. O governo identificou que os pardais estavam comendo arroz nas plantações e diminuindo a produção. Então fez uma ampla campanha para que os cidadãos ajudassem a matar pardais. Com a população de pardais quase extinta, os gafanhotos

começaram a atacar as plantações de arroz, porque os pardais eram seus predadores naturais.

Verificar o que é comum a um conjunto de casos

Uma das maneiras de levantar hipóteses de causas é identificar características (atributos ou valores ou intervalos de valores) que se repetem entre todos os elementos ou casos de um grupo de eventos. Este é o Método da Concordância de Duns Scotus (segundo Losee): analisar instâncias de ocorrências de um evento procurando o que é comum nas instâncias.

Por exemplo, nos EUA, a polícia conseguiu capturar um franco atirador porque ele se escondia dentro de um carro, atirando por um buraco no porta-malas; as imagens dos locais sempre tinham este carro. Analisando os dados capturados e armazenados no banco de dados, não se tinha nenhum atributo comum a todos os casos. O bairro era diferente, horário e tipo de vítima também. Mas as imagens revelaram o sinal comum que permitiu identificar o criminoso.

Um cuidado que se tem que ter é que em alguns conjuntos pode não haver uma característica comum a todos elementos. Por exemplo, na maioria das empresas, se formos procurar o que há de comum a todos clientes, a resposta será "nada" (conjunto vazio). O que acontece é que os clientes formam grupos, porque justamente o mercado é segmentado. Então, neste caso, seria melhor utilizar a técnica de agrupamento (*clustering*), que separa automaticamente os elementos por similaridade. Depois, a técnica de indução permite descobrir as características de cada grupo.

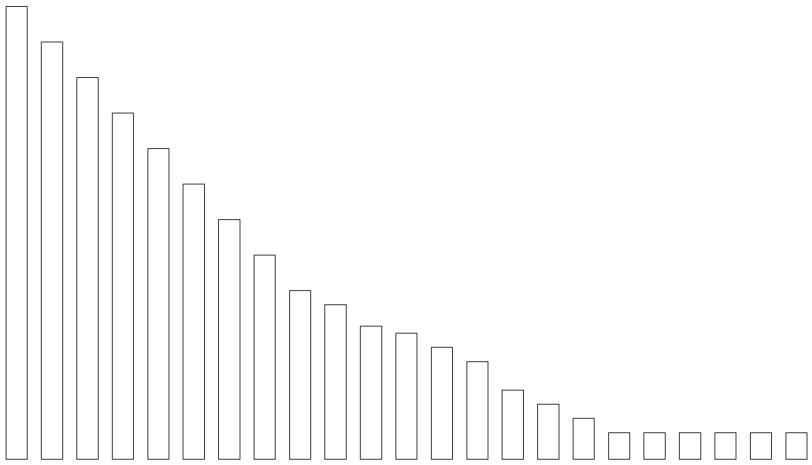


Figura 42: Gráfico de Pareto

Outra foram de trabalhar com elementos diferentes, é montando um ranking dos elementos, uma ordem segundo algum critério. Pode-se por exemplo utilizar o gráfico de Pareto, colocando mais à esquerda os elementos com maior valor para um determinado atributo. Por exemplo, na Figura 42, imagine que cada barra corresponde a um cliente e o tamanho da barra indica o número de itens adquiridos pelo cliente. Neste caso, o atributo considerado para a geração do ranking é o número de itens adquiridos, mas poderia ser a soma total de valores monetários gastos por cada cliente, a idade de

cada cliente. E os elementos podem ser produtos, lojas, vendedores, etc. Em geral, o Gráfico de Pareto então se assemelha a uma curva descendente.

Para identificar características comuns, é preciso formar um grupo para análise ou separar por grupos usando técnicas de discretização. Por exemplo, pode-se trabalhar somente com os mais bem posicionados no ranking ou os intermediários ou aqueles posicionados mais embaixo (a curva ABC funciona assim para classificar clientes). Pode-se definir um limiar numérico para os valores do atributo sendo considerado para corte dos elementos ou considerar os N primeiros elementos ou separar o grupo todo em N partes iguais.

Uma variação do método seria trabalhar com valores fuzzy para classificar elementos. Por exemplo, se estamos classificando pessoas por faixa etária, temos a tendência de definir limites. Poderia ser assim: jovens são pessoas menores de 24 anos, adultos têm entre 25 e 40, sêniors têm entre 41 e 60 anos e a 3ª idade é formada por pessoas com mais de 60 anos. Mas o que acontece com pessoas próximas das fronteiras (por exemplo, alguém com 24 anos e alguns meses) ? A lógica fuzzy permite classificar um mesmo elemento em duas classes distintas mas com graus de pertinência diferentes. Então, se uma empresa for fazer campanhas de marketing para clientes segmentados por idade, usando o raciocínio fuzzy, a empresa não perde oportunidades deixando clientes das fronteiras somente numa campanha.

Depois, tendo um grupo selecionado e separado, pode-se:
a) procurar médias de valores para um determinado atributo (pode não ser o utilizado para montar o ranking);
b) um limiar mínimo para um atributo;
c) um atributo comum ao grupo selecionado;
d) uma combinação de atributos comuns.

Por exemplo, se foi utilizado o total de itens para montar o ranking de clientes, pode-se descobrir:
a) que a média de itens comprados entre eles era X;
b) que a idade mínima era 23 anos;
c) que todos os clientes do grupo selecionado tinham residência na cidade;
d) que todos estes clientes selecionados tinham renda acima de Y salários mínimos e moravam em residência própria.

Verificar o que é incomum ou diferenças entre grupos

Uma anomalia do método descrito antes, é que as características comuns num grupo podem também estar presentes em outros grupos. Um exemplo: uma empresa utilizou o gráfico de Pareto para tentar encontrar as boas práticas de seus vendedores. Ao identificar os atributos dos melhores vendedores, descobriu que estavam também presentes nos piores.

Pode-se dividir o grupo todo em partes para comparar características, a fim de encontrar o que diferencia um grupo de outro. Por exemplo, uma técnica muito utilizada para entender bons e maus pagadores em sistemas financeiros é dividir os clientes nestes 2

tipos de grupos e então analisar o que há de comum internamente a cada grupo. Depois então, os atributos que caracterizam cada grupo são comparados em busca das diferenças.

Um problema que pode acontecer é se não forem encontradas características comuns em cada grupo. Então, pode-se analisar algumas amostras de cada grupo. Por exemplo, ao se analisar um grupo de máquinas que falhavam antes do tempo previsto, não se encontrou nenhuma característica comum. Entretanto, cada uma delas tinha uma característica única, que a distinguia das demais deste grupo. Comparando 1 caso onde o defeito ocorreu com 1 caso onde não houve o defeito, chegou-se a uma característica que havia no primeiro e não ocorria no segundo. Esta foi uma tomada então como uma hipótese de diferença entre os grupos. Então fez-se uma análise estatística para saber a frequência da presença ou ausência da característica em cada grupo. A probabilidade não era 100% em cada grupo mas bastante significativa.

Este é o Método da Diferença de William of Ockham (segundo Losee).

O grande defeito é seguir por caminhos errados, levando a perda de tempo e esforços. Imagine pegar justamente as exceções. Mas muitas vezes, são caminhos que devem ser trilhados na falta de hipóteses.

Benchmarking e Analogias

Uma maneira de levantar hipóteses iniciais é utilizando *benchmarking*, ou seja, vendo o que já havia acontecido antes com outras empresas ou em situações semelhantes. Este é o princípio das heurísticas, já comentado antes. Muitas vezes podemos reutilizar soluções que deram certo em outra área. A técnica de *benchmarking* significa olhar e aprender com outras empresas. A solução de um programa de computador que não "roda" pode vir de uma ideia de um brinquedo que não funciona. A causa para defeitos num processo de produção pode vir da análise de defeitos em carros. É claro que pessoas, empresas e mercados são diferentes, são organismos vivos. E por isto, talvez seja necessário alguma adaptação na solução, pois ela provou funcionar em outro contexto ou área, mas pode não funcionar neste exatamente igual numa nova situação.

Por isto, um esquema visual é importante, pois podemos visualizar problemas e soluções. Se compararmos dois casos com informações diferentes, talvez o padrão visual seja o mesmo. Mapas mentais, anagramas, grafos podem ajudar (adiante veremos um caso com mapas mentais).

E também é preciso ter informações e conhecimentos diversos. Por isto é tão importante conhecer vários assuntos e não ser um "especialista burro".

O perigo das analogias, segundo Popper, é generalizar demais ou de forma errada. Ele conta o caso de cachorros que foram aterrorizados com cigarros. Após, cada vez que um destes via um papel branco enrolado, ele fugia. No caso, o fator real (causa raiz) era o fumo e não o papel que enrolava o fumo. Mas até mesmo o ser humano confunde as causas.

"Reframe", repensar o problema

Eu gosto do termo "*reframe*" associado a criatividade e solução de problemas. *Reframe* é repensar o problema com outros esquemas, pontos de vista, elementos, dados, contextos, regras, etc. Talvez o momento Eureka dependa de vermos o problema com outros olhos, sem mesmo precisar mudar as informações ou o contexto. Basta "pensar diferente".

Não se pode simplesmente ficar em cima de um problema usando os mesmos paradigmas; o resultado será sempre o mesmo. Repensar tem que ser "reformular". Por isto que quando temos um problema devemos sair do ambiente, fazer outra coisa (ex. Arquimedes). Muitas vezes fazemos isto e quando voltamos "enxergamos" a solução de primeira e pensamos: "por que não vi isto antes ?"

Para reformular, temos que nos libertar das regras que estamos usando. Einstein, Galileu e Darwin quebraram paradigmas. Mas para isto, precisaram se libertar das teorias aceitas em suas épocas. Se pensarmos que um problema só tem uma solução possível (ou caminho para a solução), a tendência é tentar colocar os dados num esquema que leve por este caminho. É por isto que muitas soluções aparecem em sonhos, porque quando dormimos a parte do cérebro que dita regras e conexões lógicas está dormindo também. Por isto é que sonhamos coisas estranhas, sem lógica. Mas é também o que permite conectar diferentes matrizes e fazer associações novas (que acordados não fazemos).

Uma sugestão é utilizar esquemas diferentes para representação ou descrição do problema. Podemos usar diagramas (esquemas visuais), textos, imagens em sequência (*storytelling*), planilhas e até mesmo gravações de áudio (segundo a Neurolinguística, algumas pessoas retém melhor as informações ouvindo, outras vendo, outras tocando, etc.).

Precisamos voltar, tomar direções diferentes, usar dados diferentes, observar detalhes que talvez não fossem considerados tão importantes, refazer as perguntas. Sair das regras normais e hábitos, ver o que está escondido (hidden analogies). É justamente o contrário de usar analogias e *benchmarking*.

Recentemente, surgiu uma explicação possível para as pedras (algumas com mais de 300 quilos) que se movem sozinhas no lago seco de Racetrack Playa, no deserto de Mojave nos EUA. Elas deslizam pelo solo deixando marcas bem visíveis atrás delas. O geólogo da NASA Ralph Lorenz acredita que as rochas são movidas pela ação dos ventos e da água. Ele acredita que elas ficam envoltas em gelo durante o inverno, então quando o leito do lago derrete e fica lamacento, o gelo permite às pedras deslizar sobre o barro, impulsionadas pelos ventos fortes do deserto.

Quebra de Paradigmas

Já comentamos que os hábitos são bons para filtrar opções e economizar tempo. Mas há o perigo de ficar preso a soluções pré-determinadas ou tradicionais. Há heurísticas (não comprovadas cientificamente) que acabam guiando as nossas decisões. Por exemplo, muitos executivos demitem funcionários para reduzir custos. É a solução mais comum, mais tradicional e muita vezes mais fácil para quem faz (não para quem é demitido). Outro exemplo: muitas empresas pensam que não se investe na crise; mas o livro de Carlos Domingos ("Oportunidades Disfarçadas") conta justamente casos de sucesso que contrariaram esta regra.

Muitas vezes, a solução passa por quebrar paradigmas. Segundo Thomas Kuhn, no seu famoso livro "A estrutura das revoluções científicas", paradigmas são realizações científicas universalmente reconhecidas que, durante algum tempo, fornecem problemas e soluções modulares para uma comunidade de praticantes de uma ciência. O paradigma orienta pesquisas de um grupo; é um modelo ou padrão aceito.

Um paradigma é uma maneira de ver o mundo. E isto pode mudar. Kuhn comenta o experimento de utilizar um óculos que inverte a imagem (descrito por Harvey Carr). As pessoas se acostumam e conseguem viver normalmente.

A quebra de paradigma é uma nova forma de ver as mesmas coisas talvez até com os mesmos instrumentos. Foi o que aconteceu em várias quebras de paradigma na Astronomia. Em muitos casos, o mesmo instrumento (luneta) era utilizado focando no mesmo lugar no espaço. Mas as hipóteses eram diferentes. E aí novos detalhes aparecem, fazendo então a teoria se modificar.

Entretanto, uma teoria pode ser aceita mesmo sem explicar todos os fenômenos. Quando surgem contra-exemplos, a teoria não deve ser rejeitada mas adaptada. Para rejeitar uma teoria, é preciso ter outra para substituí-la.

As revoluções científicas são justamente episódios de desenvolvimento não cumulativo, nos quais um paradigma mais antigo é total ou parcialmente substituído por um novo, incompatível com o anterior. O pré-requisito para a substituição é o funcionamento defeituoso do modelo. Uma nova teoria não precisa estar em conflito com a antiga; pode tratar de assunto novo (como a física quântica) ou ser de maior grau (englobar outras menores).

É claro que há propriedades inatas e irredutíveis, as quais não são nunca questionadas e não precisam ser constantemente avaliadas, o que tornaria o raciocínio muito mais lento. Em cada empresa, há princípios básicos irredutíveis. Por exemplo, algumas empresas de varejo definem posições de estoque mínimo e não voltam a questioná-las. Setores de RH definem critérios de avaliação de pessoal e nunca os rediscutem. Departamentos de venda definem índices para premiar vendedores e são sempre os mesmos que ganham.

Descoberta por acaso (serendipity)

Serendipity é um neologismo inglês que significa fazer descobertas por acidente, sorte ou acaso. A origem da palavra é relatada no artigo de Pek Van Andel e é creditada ao escritor britânico Horace Walpole em 1754.

Entretanto, a sorte favorece a mente preparada (frase associada a Pasteur por vários autores, entre eles Koestler e Johnson). Isto quer dizer que, para descobrir algo por acaso, é preciso ter informações, hipóteses, testes, ideias, etc.

É falso acreditar que Arquimedes resolveu o problema do Rei sem nada saber. Antes, ele estudou muito o problema e possíveis soluções. Mesmo aqueles que sonharam com soluções é porque estavam, durante o dia, colhendo informações. Talvez o momento Eurekha tenha sido a junção das peças do quebra-cabeça (como Koestler e Johnson dizem ser um dos passos essenciais para a criatividade). Mas então antes era preciso colher e analisar as peças.

9.2 Sinais fracos, fatos X opiniões, rumores e boatos

Estamos acostumados a pensar que toda decisão deve ser baseada em fatos e raciocínio lógico. Em geral, as pessoas relutam em usar dados não confirmados ou mesmo que não sejam quantitativos (números). Entretanto, como já discutimos antes, pela racionalidade limitada, nem sempre é possível coletar e analisar todos os dados e alternativas necessários, ou mesmo verificar a veracidade de tudo o que ouvimos e lemos. Em muitos casos, utilizar uma informação não confirmada, pode ser o pulo do gato na frente dos demais concorrentes.

Por exemplo, a maioria dos investidores das bolsas de valores utilizam softwares que analisam dados históricos e fazem previsões através de técnicas de Data Mining (mineração de dados). Mas todos os investidores tomarem decisões da mesma forma (com os mesmos dados e técnicas), ninguém vai ganhar. Para vender, é preciso que alguém compre e vice-versa. Então, para ganhar na Bolsa é preciso ter uma visão diferente dos outros, sobre algo que pode dar certo ou errado, enquanto os outros estão pensando o contrário. Isto, é claro, aumenta o risco e a probabilidade de erro, mas também aumenta as chances de sair ou estar à frente. Se formos esperar para confirmar todas as informações, nunca vamos tomar uma decisão.

Gunther diz que precisamos também utilizar dados subjetivos, como os sentimentos. Precisamos ouvir os nossos próprios sentimentos. Isto não significa confundir intuição com desejo. Um forte desejo pode parecer uma forte intuição. Decisões também podem ser tomadas com base em informações ainda não confirmadas, como opiniões e rumores.

Existem informações que sozinhas não significam muito, mas quanto integradas podem ajudar a predizer eventos. Estes são os chamados sinais fracos (weak signals) segundo Ansoff (1980). Sinais fracos são aqueles pedaços de informação, ambíguos, vagos, incompletos, imprecisos e controversos. Não são claros; são quase mudos. Estão normalmente escondidos no ruído e não recebem muita atenção no processo de decisão.

São informações mal estruturadas, esparsas e desconexas. Não são certezas, mas pistas. Podem surgir na forma de frases, fotos, cheiros, imagens, desenhos, pedaços de artigos ou qualquer observação pronunciada por alguém. Nesses fragmentos esparsos pode residir um potencial informativo importante para a investigação.

Os sinais fracos analisados separadamente, não significam nada. Gradualmente se integram para formar um padrão de inteligência, que dão alertas de necessidades de mudanças. Tornam-se fortes quando combinados com outros sinais. "Uma andorinha sozinha não faz verão", mas vários bandos voando na mesma direção em dias diferentes e por um certo tempo devem significar algo importante. É por isto que as postagens em twitter, blogs, Facebook e outras redes sociais estão sendo monitorados por departamentos de inteligência, seja em empresas, governo, partidos políticos e até pais e familiares (a importância da análise de textos na Web será discutida mais adiante).

Sinais fracos podem gerar grandes influências nos resultados. A Teoria do Caos (Gleick, 1989) explica que pequenas alterações em algumas variáveis podem modificar completamente o resultado final. Daí é que surge o tal efeito borboleta (uma borboleta voando no Brasil pode gerar uma tempestade no Texas). Gladwell, no livro Ponto da Virada (2013) também comenta sobre pequenos eventos que desencadeiam grandes revoluções. Há muitos exemplos na moda e no marketing. Nate Silver fala de um sinal que foi desconsiderado para um terremoto na Itália: sapos deixaram de desovar 5 dias antes.

Outro caso interessante é com relação à análise de atrasos em voos. Muitas companhias descobriram que a etapa de limpeza da aeronave nas escalas era um determinante para os atrasos. Antes relegada a um fator de pouca importância no tempo da viagem, a etapa de limpeza recebeu foco de equipes de planejamento. A TAM então passou a usar um tapete vermelho para clientes limpares os pés na entrada. A GOL projetou um esquema em que os clientes ajudam na limpeza interna. E tudo isto deu certo.

Já Pentland estuda outros tipos de sinais fracos, o que ele chama de “sinais honestos”. São sinais que aparecem nos rostos das pessoas, impercebíveis no cotidiano pelo olhar humano, muito porque acontecem num tempo menor que um piscar de olhos. Quando assistimos vídeos em câmera lenta, tais sinais aparecem claramente. Pentland e sua equipe utilizam tecnologias para detectar estes sinais honestos. Os sinais podem ser demonstrações de empatia para facilitar e encorajar comunicação (ex.: acenos com cabeça) ou podem indicar estresse. E não aparecem somente no rosto, mas são demonstrados por todo o corpo humano. A linguagem dissimula emoções, mas o corpo não as consegue esconder. Já há até taxonomias para análise de expressões faciais (Ekman e parceiros; Kring e Sloan, 2007)

O ser humano intuitivamente consegue identificar tais sinais. É o que muitos dizem de uma conversa olho no olho para conhecer melhor uma pessoa. As decisões referentes a escolha ou avaliação de pessoas são feitas assim. Mas também servem para avaliar veracidade de argumentos e informações que os outros nos passam. Saber reconhecer tais sinais pode melhorar nossa tomada de decisão. Os estudos de Pentland concluíram que empregados que se valem de interações cara a cara acabam sendo 30% mais produtivos.

Gladwell comenta sobre o poder dos boatos, que podem ajudar ou atrapalhar. Ele comenta o caso de um boato espalhado entre americanos para resistência aos ingleses, e que acabou tendo uma forte influência na independência americana.

O importante é saber juntar os sinais fracos, entender suas relações, seu poder de conjunto e para onde apontam. Para Nate Silver, havia dados suficientes para prever o ataque terrorista de 11 de setembro. O problema não era a carência de informações, mas sim que as peças não foram corretamente juntadas (exatamente como aconteceu nos ataques a Pearl Harbor). O que faltava era exatamente uma teoria que pudesse explicar os dados em conjunto, um padrão que indicasse um evento significativo ou mesmo uma hipótese por menos provável que fosse.

Lesca (2003) apresenta uma metodologia para análise de dados sobre mercado competitivo, onde os chamados "sinais fracos" são também considerados. Isto inclui opiniões e até mesmo boatos. A ideia é não descartar nada. A metodologia de Lesca é interessante porque demonstra como conectar dados e sinais fracos, para gerar hipóteses. Talvez o conjunto final de dados possa mostrar uma tendência que os números não apresentavam. Parte desta metodologia será discutida adiante.

Outro exemplo: uma empresa não sabia mais como lidar com quebras em suas máquinas. Já havia investigado tudo: fornecedores, tempo de uso, qualidade dos operadores, qualidade das peças que substituíam outras, temperatura durante o uso, as variações de temperatura (uso X descanso) e até mesmo a temperatura ambiente. E nada de encontrar um padrão. Aí alguém suspeitou que a trepidação das máquinas era diferente. Colocaram sensores para medir o quanto cada máquina trepidava. Descobriram que as medidas eram diferentes mas não havia um padrão. Não encontram um motivo para haver diferenças nas trepidações, analisando as variáveis já descritas antes. Aí, outro alguém suspeitou que a diferença nas trepidações poderia estar no tipo de piso usado na empresa. Nada. Eram todos iguais. Aí outro alguém, analisando onde ficavam as máquinas que mais davam problemas, descobriu que o andar onde estava é que fazia a diferença. Máquinas em andares mais altos tinham histórico maior de falhas e quebras. Concluindo: as diferenças na estrutura do prédio eram a causa dos problemas.

9.3 Análise de causa-efeito

Um dos grandes objetivos de um processo de BI é encontrar causas para eventos ocorridos ou padrões descobertos. Por exemplo, sistemas gerenciais podem ajudar a descobrir que tipo de cliente é mais lucrativo para a empresa. Mas BI tem que explicar por que este tipo é mais lucrativo e outros não. Sistemas gerenciais apontam os produtos mais lucrativos. BI deve dizer por que estes produtos e não outros são mais lucrativos. Sistemas gerenciais sobre dados de estoque e logística identificam quais produtos dão mais custo porque precisam ficar mais tempo armazenados. BI tem que permitir descobrir por que tais produtos não podem girar ou viajar mais rapidamente.

Então uma das tarefas do processo de BI engloba buscar causas para efeitos observados. Por exemplo, no McDonald´s, o sanduíche Big Mac fica pronto mais rápido que os demais porque vende mais ou vende mais porque fica pronto mais rápido ? Se outro

sanduíche ficasse pronto primeiro, ele seria o mais vendido ? É o velho problema de o que vem primeiro: o ovo ou a galinha ? Gladwell, no livro "O ponto da virada", comenta a relação entre pessoas confiantes e o ato de fumar. O que gera o quê ? É a confiança que faz a pessoa fumar ou é o ato de fumar que deixa a pessoa mais confiante.

A primeira tarefa identificar causas de efeitos é avaliar a correlação entre as variáveis ou eventos. A correlação é uma técnica estatística que avalia a similaridade entre 2 vetores de números, 2 gráficos ou 2 séries. O coeficiente de Pearson é um dos métodos mais utilizados. Quanto mais próximos os números na ordem, maior o grau de correlação entre os vetores. Para uma empresa é importante avaliar a correlação entre suas ações e os resultados. Por exemplo, uma empresa descobriu que um aumento de 5 pontos na atitude comportamental dos empregados implicava em 1,3 ponto de incremento na satisfação dos clientes, e isto fazia aumentar em 0,5% o faturamento da empresa.

Tal descoberta permite à empresa avaliar onde investir e o quanto. Neste exemplo, se ela quiser aumentar 1% das vendas talvez tenha que aumentar 10 pontos na atitude dos colaboradores.

Como já discutimos antes, correlação entre eventos ou variáveis não necessariamente implica em que um seja causa de outro. Conforme Hans Reichenbach, citado por Tsamardinos e Sofia Triantafillou (2011), se A e B estão correlacionados, ou A causa B, ou B causa A, ou eles compartilham uma causa comum. Eu ainda acrescentaria que pode ser uma sincronicidade, como discutido antes, caso não haja uma frequência mínima. O famoso teste de Granger pode ajudar a identificar se há uma relação causal numa correlação.

Além disto, uma causa pode ser direta ou indireta. Em muitas empresas, costuma-se relacionar os índices de venda ao desempenho dos vendedores. Mas muitas vezes são esquecidas causas indiretas. Por exemplo, as propagandas feitas pela empresa podem ajudar um vendedor e prejudicar outro. Os tipos de clientes ou regiões pelas quais cada vendedor ficou responsável pode ser o determinante, isentando o vendedor e suas atitudes do resultado final. Outro exemplo: a causa para o custo elevado de um produto pode estar na raiz da cadeia de suprimentos.

Causas indiretas podem gerar o evento mas com muitos laços intermediários. Imagine o caso de uma virose que deixa várias pessoas com problemas estomacais. Se todos comeram no mesmo restaurante, isto pode ser uma causa comum e direta. Entretanto, pode ter ocorrido de uma pessoa ter comido algo e depois passado o vírus para outro que passou para outro e assim por diante.

Na área de saúde, é muito comum confundir sintoma com causa. A causa vem primeiro e os sintomas ou sinais aparecem depois. Mas há casos complexos onde fica difícil determinar o que é causa e o que é efeito. Por exemplo, água no pulmão é consequência ou causa de problemas cardíacos ? e diabetes, é causa ou consequência de problemas de má circulação ?

Outra questão a cuidar é que a causa pode ter ocorrido logo antes do evento efeito ou muito tempo antes. Uma promoção publicada num jornal talvez gere resultados no mesmo dia. Mas uma campanha nas redes sociais talvez demore mais tempo para gerar resultados positivos. Levitt e Dubner (no livro Freakonomics) levantam a possibilidade de a liberação de abortos ser uma das causas para diminuição de crimes nos EUA no final de 1989. Mas os 2 eventos estariam relacionados numa diferença de tempo de 20 anos. Esta é a chamada correlação assíncrona que já foi discutida antes neste livro.

Aqui devemos distinguir causas determinísticas de causas prováveis. O determinismo ocorre quando a causa leva aos efeitos em 100% dos casos e sem nenhuma dúvida. Quando não há certeza, devemos tratar a relação causa-efeito de forma probabilística. Isto acontece em modelos ou padrões onde há exceções.

Para poder avaliar o determinismo da causa sobre o efeito, é necessário avaliar também outros eventos no contexto. Como já discutimos antes, BI não acontece num mundo fechado. Se as vendas caem ou sobem inesperadamente num determinado mês, não significa que as ações da empresa foram a causa. As ações dos competidores, os eventos que acontecem na cidade ou sociedade, as questões econômicas, etc., podem ser causas mascaradas.

Uma maneira de avaliar qual realmente é a causa para um efeito é colocar num banco de dados todos os eventos que podem estar relacionados e aí utilizar técnicas estatísticas (como análise de correlação e teste de Granger) para filtrar candidatos a causas.

Em muitas vezes teremos que refazer as situações ou eventos e então fazer novas observações. A cada novo experimento realizado, precisamos monitorar as causas candidatas e registrar tudo num banco de dados para análise estatística.

Outra técnica útil é gerar um grafo relacionando possíveis causas a efeitos. Cada relação de causa-efeito recebe uma probabilidade. Depois podemos analisar o grafo com os seguintes Axiomas Causais de Markov e de Rei:
a) causas imediatas geram efeitos independente de causas remotas; por exemplo, infecção causa doença, independente de como se foi infectado (axioma de Markov);
b) uma causa comum pode gerar dois ou mais efeitos independentes; por exemplo, fumar pode causar câncer e dedos amarelados, mas um efeito não tem a ver com outro (axioma de Reichenbach).

Um último cuidado é com causas escondidas. Forster discute a alta frequência de doenças de coração entre os que bebem café. Há estatisticamente uma correlação entre as duas variáveis: a doença aparece mais no grupo dos que bebem, do que na população em geral. Entretanto, pode haver uma causa escondida e neste caso há: quem bebe também fuma. Então precisamos comparar o número de casos de câncer de pulmão em relação ao número de fumantes na população e em relação ao número de pessoas que bebem café. E por fim verificar a proporção em que as características aparecem na mesma pessoa.

Análise de causa-raiz

Gladwell, no livro Outliers, comenta que acidentes com aviões acontecem por acúmulo de erros triviais e pequenos. Ele comenta o caso de uma companhia aérea que precisou treinar sua tripulação para se comunicarem melhor em inglês com as torres de controle em outros países. Isto porque a má comunicação gerava outros pequenos erros e daí poderia até mesmo causar um grave acidente.

O importante então é tentar descobrir a chamada causa-raiz, aquela que gera outras causas em sequência ou cascata. Se conseguirmos eliminar a causa-raiz, as demais causas não acontecerão e assim o efeito também não acontecerá.
Veja o caso das empresas de Eike Batista que perderam muito valor em 2013: uma empresa estava escorada em outra e dependia dos investimentos se confirmarem nas outras. Quando a base ruiu, todas caíram junto.

Uma das maneiras de analisar causas de efeitos é usando o Diagrama de Ishikawa (1990), também conhecido como diagrama de causa-efeito ou espinha-de-peixe (Figura 43). A ideia é ir dividindo um efeito em suas causas. Cada causa pode ser subdividida também, formando novos diagramas de níveis mais detalhados.

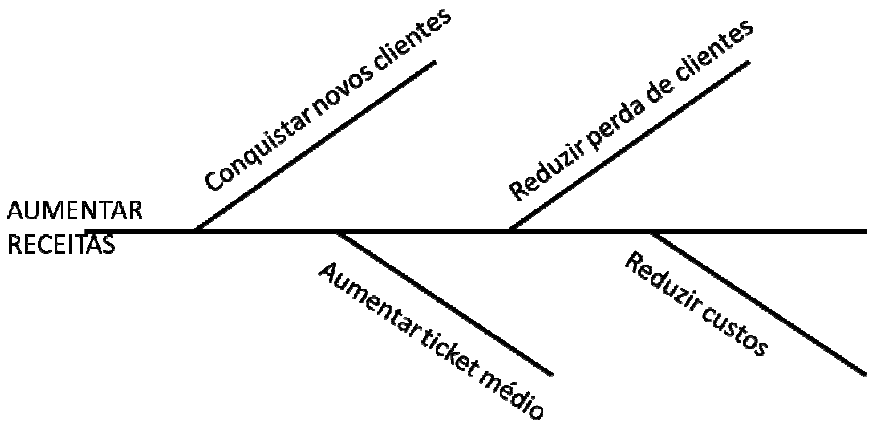


Figura 43: Diagrama de Ishikawa (causa-efeito ou espinha-de-peixe)

Já comentamos antes o caso da empresa que descobriu que um aumento de 5 pontos na atitude comportamental dos empregados implicava em 1,3 ponto de incremento na satisfação dos clientes, e isto fazia aumentar em 0,5% o faturamento da empresa. A causa raiz era a atitude comportamental.

Em outro caso, uma empresa gerou diversos diagramas de Ishikawa para entender causas das reclamações de clientes. A divisão foi feita em vários níveis, isto é, cada causa era estudada e suas causas analisadas, gerando diagramas interrelacionados, formando uma árvore de causas interconectadas. O interessante é que um fator se repetia em diversas subcausas e diversos ramos da árvore: a competência dos funcionários. A empresa então focou em treinamentos e conseguir diminuir muito as reclamações.

Avaliação sistêmica dos dados

Nate Silver discute em seu livre as previsões feitas para a eleição de presidente nos EUA. Ele acredita que, para uma previsão ser precisa, não basta saber qual candidato tem mais preferências nos estados; é preciso avaliar os estados mais importantes e o conjunto (relações entre estados).

Além disto, na maioria dos casos, não há uma causas única e simples; pode haver causas múltiplas ou multivariadas (como na regressão). Para tanto, é necessário analisar conjuntos de dados e não dados isolados.

Muitos fenômenos têm como causa um conjunto de eventos anteriores, ou seja, uma combinação de causas menores. Um usuário acessando um sistema computacional e errando a senha é um evento corriqueiro e normal. Agora, este mesmo usuário errando diversas vezes a senha, pode ser indício de tentativa de fraude.

Em outros casos, a causa pode ser um volume grande de eventos do mesmo tipo. Veja a moda por exemplo. Uma pessoa sozinha usando uma marca ou um tipo específico de acessório não gera efeito, mas várias fazendo isto gera um efeito exponencial. Este tipo de fenômeno é chamado de ponto da virada, muito bem descrito no livro de Malcolm Gladwell.

Um dos desafios é extrair significado (a chamada abstração semântica) a partir de um conjunto de dados aparentemente desconexos. Uma pessoa comprando plástico que pode ser usado para fazer bombas não significa nada, é um evento isolado. Mas se ela também comprar uma mochila, uma passagem de avião e estiver indo para um lugar onde não tem nenhum conhecido, pode ser algo significativo.

Uma empresa descobriu que suas máquinas só tinham problemas quando a temperatura no ambiente passava dos 30 graus "e" um operador inexperiente (menos de 1 ano de trabalho) estava manipulando a máquina. Notem: eu frisei o "E". Ambos os fatores deveriam estar juntos para gerar o problema.

A abstração pode ser feito de duas formas: por Generalização ou por Agregação, conforme Smith e Smith (1977). Generalizar é formar conceitos de mais alto nível a partir de fatores menores. Um exemplo de generalização seria notar que todos os problemas com uma determinada máquina industrial ocorreram com operadores que tinham menos de 20 anos. A agregação seria compor eventos mais complexos a partir de fatores menores. Por exemplo, o mesmo caso (operadores jovens) mas somente em máquinas adquiridas há menos de um ano (máquinas novas, de modelos novos).

Em alguns casos, os fatores talvez não apareçam simultaneamente mas em sequência. Então a causa é uma sequência específica de eventos. E a ordem pode ser importante. Se os mesmos eventos ocorrerem em uma sequência diferente talvez não gerem o efeito.

Então, resumindo, uma causa pode ser identificada:

- pela presença de algum evento específico; exemplo: um vendedor é melhor que outro porque visita seus clientes enquanto que os outros não o fazem;
- pela frequência de eventos; exemplo: o melhor vendedor visita cada cliente toda semana (os outros só uma vez por mês);
- pela ordem dos eventos; exemplo: o melhor vendedor liga após visitar seus clientes, enquanto que os demais ligam e depois visitam.

É importante lembrar que, quando estamos falando de causa-efeito, nem sempre estamos só preocupados com efeitos ruins. Um objetivo é um efeito desejado. E procurar por suas causas também é importante.

Parcimônia – conjunto mínimo de causas

Em Ciência, parcimônia é a preferência pela explicação mais simples para uma observação. Parcimônia é um conceito utilizado para focar em relações ou eventos mais importantes, aqueles que realmente determinam o efeito. No exemplo anterior dos problemas em máquinas industriais, podemos ter vários fatores que são possíveis causadores. Por exemplo: idade do operador, tempo de experiência, marca ou modelo da máquina, tempo de uso, tempo de vida da máquina, condições ambientais. É possível que apenas algumas poucas características sejam realmente causa dos problemas. A análise de correlação pode filtrar fatores que não estão associados estatisticamente. Mas é difícil ter um modelo probabilístico que se encaixe em 100% dos casos. O mais comum é ter várias exceções.

Então digamos, para exemplificar, que:
a) 30% dos casos de quebras das máquinas ocorra com operadores jovens;
b) 20% ocorram com operadores inexperientes;
c) 15% ocorram com máquinas modelo A;
d) 10% ocorram com máquinas modelo B (75% ocorrem com outros modelos em menor proporção);
e) 50% ocorram em dias de muita umidade;
f) apenas 5% dos casos ocorrem quando os fatores (a), (b), (c) e (e) estão presentes;
g) apenas 3% dos casos ocorrem quando os fatores (a), (b), (d) e (e) estão presentes;
h) 40% dos problemas ocorrem com operadores jovens e inexperientes.

Se a empresa precisa otimizar os investimentos para reduzir os problemas (não há como gastar para atacar todas as causas), o que ela deve fazer ?

Uma possibilidade seria atacar somente a causa (e) que é a que tem maior probabilidade (50%). Outra seria atacar as causas (a) e (b), pois juntas (conforme item "h") dão 40% de probabilidade e podem ser dirimidas com ações semelhantes (e de menor custo).

9.4 Métodos e Teorias para Investigação

As investigações científicas, a perícia criminal e o diagnóstico médico são facetas de um mesmo problema: encontrar causas ou explicações para eventos.

Por isto, nesta seção vamos falar de algumas metodologias (se é que se pode chamar assim) para investigação. Elas nos orientam como coletar, filtrar e analisar dados, como desenvolver teorias, como criar e validar modelos, como identificar e definir regras e leis científicas ou não.

Método Cartesiano

O método de René Descartes, que ficou conhecido como método Cartesiano, possui os seguintes passos ou preceitos:

1. Busca pela verdade: nunca aceitar algo como verdadeiro sem conhecer; receber as informações com ceticismo, examinando sua racionalidade e sua justificação;
2. Análise, ou divisão do assunto em tantas partes quanto possível e necessário: dividir cada uma das dificuldades em tantas partes quanto for possível e necessário para melhor entendê-las e resolvê-las;
3. Síntese, ou elaboração progressiva de conclusões abrangentes e ordenadas a partir de objetos mais simples e fáceis até os mais complexos e difíceis.
4. Enumerar e revisar minuciosamente as conclusões, garantindo que nada seja omitido e que a coerência geral exista.

Método Científico

Os passos do método científico, de forma geral:

1. Fazer observações, sistemáticas e controladas
2. Levantar hipóteses
3. Montar um modelo ou teoria científicas
4. Realizar novos experimentos e fazer novas observações
5. Avaliar se as novas observações corroboram a teoria
6. Caso não corroborem, reciclar as hipóteses ou refazer a teoria.

Método indutivo-dedutivo de Aristóteles

Segundo Losee, a observação leva a princípios explanatórios (pela indução) e os princípios geram novas observações pela dedução (para confirmar).

Por exemplo, ao identificar que um cliente jovem comprou o produto X e depois que outro cliente jovem também comprou o mesmo produto, começamos a pensar na hipótese de haver uma regra (princípio) que diga que "todo cliente jovem compraria o produto X (se soubesse que ele existisse). Isto é indução. Ela olha para o passado e procura explicações.

Se esta regra for verdadeira, então um novo cliente jovem deverá comprar o produto X, e esta é a dedução de uma possibilidade. Ela olha para o futuro.

Podemos também pensar na dedução como uma maneira de produzir fatos (mesmo que históricos) que devem ser verdades. Por exemplo, se descobrirmos que várias máquinas da marca XYZ quebraram 2 anos após o início de utilização, podemos pensar que esta é uma regra. E portanto, podemos dizer que as demais máquinas desta marca, mesmo que ainda não avaliadas, também quebraram no mesmo período (casos passados mas aidna não confirmados). Estas deduções (novos fatos) devem ser verificados para confirmar a regra.

Um exemplo mais formal:
A dedução funciona assim:
Tendo a regra “A ==> B” (A implica em B), se A é verdadeiro, então deduzimos B.

A indução por sua vez é assim:
Tendo várias instâncias de A e B e notando a relação de implicação de um A em um B, induzimos a regra A ==> B (se A, então B)

Método de Análise e Síntese de Newton

Análise significa dividir um problema em problemas menores ou identificar as partes de um elemento que está sendo estudado. A síntese é o caminho inverso, ou seja, a partir de elementos menores (partes), construir um elemento maior (agregado das partes).

Por exemplo, se estivermos fazendo um caminho e encontrarmos um rio, o qual devemos transpor. Sabemos que uma ponte pode resolver o problema. Então, segundo o método de análise, pensaríamos nas partes que podem compor a solução (a ponte) e procuraríamos elementos que pudéssemos usar para formar a solução (talvez árvores e galhos próximos ao lugar).

Pelo lado da síntese, a ideia seria procurar elementos que estivessem disponíveis no momento (próximos ao lugar) e daí tentar construir uma solução com eles. Talvez a solução não fosse uma ponte, mas uma canoa ou tirolesa. A solução dependeria dos elementos encontrados.

Método de Galileu

Galileu revolucionou o modo como a Astronomia era feita. De seus aprendizados, surge um método para construção de teorias. Os passos são:
1. Fazer a observação do fenômeno;
2. Resolver a complexidade do fenômeno, identificando elementos, relações, quantidades, medidas, etc;
3. Elaborar uma hipótese explicativa;
4. Verificar a hipótese através de experimentações ou novas observações.

Raciocínio Abdutivo

Segundo Charles Sanders Peirce: "*a abdução é o processo para formar hipóteses explicativas. A dedução prova algo que deve ser; a indução mostra algo que atualmente é operatório; já a abdução faz uma mera sugestão de algo que pode ser. Para apreender ou compreender os fenômenos, só a abdução pode funcionar como método. O raciocínio abdutivo são as hipóteses que formulamos antes da confirmação (ou negação) do caso.*"

A abdução funciona assim:

Tendo a regra “A ==> B” (A implicando em B), se B é um fato comprovado, podemos abduzir (como hipótese) que A é verdadeiro e também é causa de B.
Somente testes posteriores podem comprovar se isto é verdade. Mas a hipótese está aí.

A abdução se contrapõe ao método cartesiano. Ela não identifica verdades, nem prova nada. Mas é uma boa maneira de levantar hipóteses.

Visão Sistêmica e Pensamento Sistêmico

Visão sistêmica consiste na habilidade em compreender os sistemas de acordo com a abordagem da Teoria Geral dos Sistemas. Para entender a visão sistêmica, primeiro é preciso entender as principais características de um sistema, dentre as quais:

a) Um sistema é um conjunto de elementos inter-relacionados.

Ou seja, um sistema é composto por elementos ou partes e assim infinitamente. Os elementos de um sistema são também sistemas (neste caso, subsistemas). Por exemplo, o motor de um carro também é um sistema. E desta forma, cada subsistema também possui as 4 características básicas. E se os elementos são sistemas, então eles também são formados por subsistemas (e isto se repete infinitamente).

As partes possuem conexões entre si, segundo alguma ordem ou objetivo comum. Nem todos elementos estão conectados a todos outros. Podem haver subgrupos, mas sempre haverá alguma ligação entre os grupos.

b) Todo sistema é parte de um sistema maior (e isto ocorre infinitamente).

Por exemplo, o sistema “carro” é parte de um sistema maior de tráfego, que por sua vez pode ser considerado subsistema de uma cidade e assim infinitamente.

O que está fora do sistema é seu meio-ambiente. O meio-ambiente não pode ser controlado pelo sistema, mas pode trocar “coisas” com o sistema (energia, produtos, materiais, informações) e por isto, dizemos que o sistema pode influenciar o meio-ambiente e vice-versa.

Por exemplo: o meio-ambiente de um carro inclui a pista ou estrada, postes e árvores, edificações, placas e sinaleiras, outros carros, o clima e a natureza (ex: chuva), etc. Um exemplo de troca é a de combustível (meio para sistema) e gases poluentes (sistema para meio).

Às vezes, é difícil determinar o que está fora ou dentro do sistema. Por exemplo, os alunos de uma universidade são elementos do sistema “universidade” ou são meio-ambiente. Para tirar esta dúvida (e outras), verifique se o sistema pode controlar este elemento. Se sim, ele será um elemento do sistema. Se não, ele será um elemento do meio-ambiente. Neste exemplo, a universidade não pode controlar que o aluno venha à aula, portanto os alunos são parte do meio-ambiente. Um cuidado: a universidade pode influenciar (persuadir) o aluno a vir às aulas mas não tem controle sobre esta decisão do aluno.

c) “Quanto maior a fragmentação do sistema (ou seja, o número de subsistemas), maior será a necessidade para coordenar as partes”.

Por exemplo, é mais fácil coordenar um time de futebol de campo (com 11 jogadores em campo) do que um time de futebol de salão (com 5 jogadores em campo). Por isto, ninguém vê peças pequenas (como parafusos) quando pensa em elementos de um carro. A razão disto é que é mais fácil visualizar menos sistemas e entender sua integração; por esta razão, as pessoas procuram agrupar os elementos em subsistemas.

O número de subsistemas é arbitrário e depende do ponto de vista de cada pessoa ou de seu objetivo. Por exemplo, um carro pode ser visto formado por 2 subsistemas somente (motor e estrutura); já outras pessoas poderão subdividir um carro em parte elétrica, motor, rodas, chassis, carroceria e estofamentos.

d) O número mágico 7 ± 2.

Na década de 50, George Miller conclui de suas pesquisa que as pessoas normais possuem uma certa capacidade de processamento de informações. Uma das descobertas é que podemos gerenciar de 5 a 9 subsistemas (por isto, o número 7 + 2 e 7 – 2). Isto quer dizer que uma pessoa consegue gerenciar melhor uma equipe com 5 a 9 membros. Ou que devemos subdividir os sistemas de 5 a 9 partes para poder entender melhor o todo.

Se tivermos mais de 9 elementos, teremos dificuldade para gerenciar os subsistemas ou entender o sistema como um todo. Abaixo disto, estamos com capacidade ociosa.

Esta regra é seguida na área de dividir um sistema baseado em tecnologia em subsistemas. Ou exemplo na área, é que devemos colocar de 5 a 9 opções no menu (interface) de um sistema automatizado.

e) Homeostase.

Este princípio diz que os sistemas sempre procuram o equilíbrio. Isto quer dizer que, se uma parte não está funcionando bem, outras terão que trabalhar mais para manter o equilíbrio e para que o sistema consiga atingir seu objetivo.

Por exemplo, se uma pessoa está mancando de uma parte, a outra perna será sobrecarregada. Uma infecção no pé pode gerar febre e isto afeta todo o corpo; da mesma forma, outras partes poderão ficar infeccionadas. Numa empresa, se o setor de vendas não está bem, outros setores devem trabalhar mais ou melhor (por exemplo, marketing).

f) Sinergia
A sinergia pode ser exemplificada pela fórmula 1 + 1 = 3. Isto significa que as partes de um sistema podem interagir para gerar algo maior, o que as partes não conseguiriam fazer ou atingir se trabalhando isoladamente.

Tal princípio também pode ser entendido através da frase "O todo não é a mera soma das partes". Um bom exemplo é a água (cuja fórmula é H2O). Se estudarmos cada parte isoladamente, teremos que as moléculas de hidrogênio se encontram na natureza em estado gasoso, e o mesmo acontecendo com o oxigênio. Mas quando esta partes se juntam formam uma substância cujo estado natural é líquido.

A sinergia também explica por que, muitas vezes, uma equipe de futebol com um jogador a menos consegue ganhar de outra com maior número de jogadores. A resposta está na integração entre as partes, que conseguem gerar algo novo.

O pensamento sistêmico é considerado a Quinta Disciplina, segundo Peter Senge. As demais são: Domínio pessoal, Modelos mentais, Objetivo comum (visão compartilhada) e Aprendizado em grupo.

Abordagem Sistêmica

A abordagem sistêmica é uma maneira de resolver problemas sob o ponto de vista da Teoria Geral de Sistemas. Muitas soluções surgem quando observamos um problema como um sistema e, desta foram, sendo formado por elementos, com relações, objetivos e um meio-ambiente.

Aí vão algumas dicas da abordagem sistêmica:

a) dividir para conquistar
Procure dividir o problema em problemas menores. Alguém que quer ir de uma cidade a outra, divide o caminho em partes por onde deve passar (estradas a tomar, saídas, entradas, conexões).

b) identificar todas as partes do sistema
Procure identificar tudo o que faz parte do sistema. Algumas partes podem fazer a diferença. Um exemplo clássico é o cavalo de tróia na guerra entre gregos e troianos. Se os gregos vissem o problema apenas como uma cidade (Tróia) com muros altos e fortes portões, não teriam conseguido entrar. A diferença aconteceu porque eles entenderam que o sistema ainda era composto de pessoas e, neste caso, supersticiosos e religiosos (que não poderiam rejeitar um presente dos deuses).

c) atentar para detalhes
A falta de uma caneta pode gerar o insucesso de um sistema automatizado. Os analistas se preocupam geralmente com as coisas grandes como computadores, redes e software de banco de dados. Mas num supermercado, se não houver uma caneta para o cliente assinar o cheque, de nada terá adiantada gastar milhares de dólares com hardware, software e treinamento de pessoal.

d) olhar para o todo (visão holística)
Se alguém está perdido numa floresta, sobe numa árvore para poder enxergar onde está a saída. O mesmo acontece com labirintos. A visão do todo permite entender como as partes se relacionam.

e) analogias
A analogia consiste em utilizar uma solução S' num problema P', similar a uma solução S que já teve sucesso num problema P similar a P'. Ou seja, é o reuso de soluções em problemas similares, com alguma adaptação da solução. Não é a toa que o Homem criou o avião observando os pássaros voarem.

O 4o Paradigma de Jim Gray - a eScience

Milhares de anos atrás, a ciência era empírica, descrevendo apenas fenômenos naturais. E isto durou até a Renascença e o Iluminismo. Há poucas centenas de anos, ramos teóricos surgiram usando modelos e generalizações. Com o surgimento do computador e do software, foi possível elaborar teorias complexas e testá-las com simulações computacionais (Hey et al., 2009).

Hoje fala-se na eScience: um processo de exploração massiva de dados, combinado com unificação de teorias, experimentos e simulação.

Cientistas realizam análises de Big Data, armazenados em bancos de dados não estruturados, capturados por instrumentos e sensores de última geração, usando computadores de alto desempenho para simulações, técnicas de gestão de informações e estatísticas, armazenados nas nuvens e construídos de forma colaborativa.

A capacidade de processamento paralelo, em clusters e *grids* de computadores (e ainda processos de comunicação machine to machine) somada à inteligência artificial tem proporcionado análises mais complexas, levantamento pela força bruta de relações escondidas, validação de fatos e modelos, e a captura de dados reais em ambientes reais.

Método de Investigação Criminal

Existem diversos manuais e artigos com dicas para investigação criminal e perícia. Basta procurar na Web por "crime scene analysis/investigation/evidence". Separei algumas dica de um manual que encontrei na Internet. (Clarke e Eck)

- analisar o ambiente do crime
- usar o triângulo de análise de problema: criminoso + vítima + local; para o criminoso, sempre há pessoas que o conhecem; para a vítima, há pessoas também que o conhecem; para o local, deve haver um gerente ou dono
- saiba que a oportunidade faz o ladrão
- coloque-se no lugar do criminoso
- eventos podem ser recorrentes ou ter outros similares (método CHEERS)
- estude a jornada ou sequência temporal do crime
- fique atento aos ritmos temporais (dia, semana, mês)
- utilize o método 5W+2H
- procure os facilitadores do crime
- considere as características geográficas
- monte uma história que faça sentido

O famoso "Unabomber", que enviava cartas bomba para cientistas com o intuito de parar a evolução tecnológica, foi identificado por suas próprias cartas: seu estilo de escrita denunciou sua formação, detalhes do papel e da impressão indicaram o tipo de máquina que usava e ainda os locais de postagem. Mas a dica final veio de um familiar.

Método do Sherlock Holmes

Investigar causas é como investigar um crime. Sherlock Holmes tinha seu método, utilizado em vários livros deste personagem mas descrito primeiramente no livro "Um estudo em vermelho" de Doyle.

Holmes usava deduções baseado em princípios universais. Por exemplo, no seu primeiro livro, Holmes infere a altura da pessoa que escreveu uma mensagem na parede, usando como fundamento o princípio (a regra) de que as pessoas costumam escrever na altura dos olhos.

Em outros casos, ele mesmo gerava suas regras, segundo o método indutivo.

Mas muitas vezes, o método de Sherlock Holmes era o método abdutivo e não o dedutivo. Em alguns casos, ele tinha um fato confirmado (um evento já ocorrido) e utilizava uma regra universal de causalidade. A partir de relações de causa-efeito, ele supunha causas para os eventos ocorridos.

Holmes também usava os métodos de análise e síntese, o método cartesiano, e outros. Mas talvez seu grande diferencial estivesse na sua forma única de coletar informações e fazer observações que nenhum outro conseguia repetir. Como já discutimos antes em outra seção, o método de coleta e observação é importante para a análise de causas.

Holmes criticava as pessoas que atulhavam o cérebro com detalhes inúteis, soterrando hipóteses promissoras. O personagem ressalta a importância também do estudo meticuloso e sistemático, aconselhando evitar formar teorias antes de possuir todos os indícios, pois isto poderia distorcer o raciocínio.

Holmes também aconselha utilizar o raciocínio retrospectivo, reconstruindo passo a passo os acontecimentos e sua ordem. Ele complementa dizendo que é mais fácil raciocinar para frente, na direção do tempo, mas isto pode fazer esquecer o processo inverso.

Quanto às circunstâncias fora do comum, ele diz que constituem mais uma orientação do que um obstáculo.

Diagnóstico Médico

O processo de diagnóstico médico tem por objetivo primeiro identificar a doença (causa) para as queixas de pacientes (e depois então prescrever tratamentos). Para tanto, é preciso analisar sinais (visíveis ao médico), sintomas (informações prestadas pelo paciente sobre o que está sentindo) e também exames técnicos (imagens, radiografias, etc.).

As primeiras informações são coletadas na chamada anamnese. Além das informações atuais (sinais, sintomas e exames recentes), é necessário perguntar sobre a história pregressa do paciente, o que inclui sabre sobre doenças anteriores. Complementa a anamnese a coleta do histórico familiar (informações sobre doenças de familiares), dos hábitos (alimentares, diários, etc.) do paciente e de suas condições e ambientes sociais e profissionais (fonte: Porto, 2005).

O objetivo é compor um quadro que possa classificar o paciente segundo casos semelhantes já estudados e aí poder determinar a causa (doença). É claro que as doenças são as mesmas, mas os pacientes são diferentes. E portanto a forma como uma doença se manifesta ou sua origem em cada paciente pode ser bem diferente. Não estamos nem falando de doenças raras ou desconhecidas, o que seria um trabalho ainda mais complexo.

Um sintoma deve ser analisado de forma contextual. Ele possui um início no tempo, uma duração e pode evoluir para características diferentes. É importante entender as características no momento em que o sintoma surgiu e também as mudanças ao longo do tempo.

Neste momento, talvez alguém que trabalhe com máquinas industriais esteja se perguntando o que pode aprender com o diagnóstico médico. Mas temos que lembrar que máquinas também apresentam sintomas e sinais, só que não nos dizem isto. Mas podemos observar e até mesmo coletar tais dados com sensores.

9.5 BI como um ato de criação

Um objetivo dá a direção, o foco, ilumina o caminho; mas a criatividade faz sair das regras e encontrar novos caminhos (hipóteses). O processo de BI é, de certa forma, semelhante a um músico procurando uma nota que faça a conexão entre 2 partes de uma música, um investigador policial procurando o autor de um crime, um mecânico investigando a causa de um defeito em uma máquina, um pintor procurando um meio de expressar suas ideias mentais e surpreender aqueles que olham sua obra.

Mas para que o momento Eureka ocorra, algumas coisas devem acontecer antes. O insight da solução não vem por acaso, como Koestler e Johnson descrevem em tantos exemplos nos seus livros. Arquimedes só viu a solução porque tinha estudado ardentemente o problema que lhe havia sido imposto, porque estava estudando outros temas e conseguiu conectá-los.

Segundo Koestler e Johnson, 2 elementos principais são necessários (entre outros):

a) Maturação de ideias
Koestler fala em *ripeness*. Steven Johnson fala em palpite lento (*slow hunch*).
Isto significa muito estudo. Coletar muitas informações, propor teorias (hipóteses), testar a teoria com exemplos reais e refazer o processo muitas vezes. Tim Berners-Lee maturou a ideia da WWW por mais de 10 anos. E perseverou. Christianson (2012)

inclusive apresenta uma cópia do manuscrito original, onde o orientador de Tim escreve a mão: "vago mas excitante ...".

b) Junção de contextos diferentes
Koestler fala em bissociação de matrizes (*bisociation of matrices*); Johnson, em colisão de ideias (*collision of hunches*). Koestler descreve como passar repentinamente de um plano (assunto) para outro (como Arquimedes), conectando as partes e gerando uma solução nova. Johnson diz que é preciso completar nossas teorias com as ideias de outros.

É preciso também ter conhecimentos generalizados, além dos especializados. Darwin foi influenciado pelo trabalho do economista Thomas Malthus sobre o crescimento da população, a falta de alimento e a possível morte de pessoas por causa desta disparidade. E Darwin iniciou sua jornada de estudos investigando pedras (na área de geologia). Steve Jobs revolucionou as interfaces homem-computador, criando telas encantadoras. Boa parte deste sucesso se deve a seus estudos de caligrafia, que o ajudaram a criar as fontes de textos.

9.6 Associações Visuais - Análise de Grafos, Redes e Mapas Mentais

Processos de BI utilizam muito representações visuais que permitam análises rápidas e descobertas através de pontos de vista. A representação multidimensional ou através de cubos de dados permite relacionar atributos e verificar associações entre valores. É claro que a técnica de Data Mining baseada em associações pode nos revelar associações estatisticamente significativas, mas um gráfico permite que a experiência ou o insight de especialistas humanos possa identificar padrões interessantes. Como se diz por aí, uma imagem vale por mil palavras.

Os grafos e mapas mentais ou conceituais podem ser úteis para representar conexões entre conceitos ou ideias. Os grafos podem ser direcionados, como um DAG (*directed acyclic graph*), representando por exemplo relações de causa-efeito ou se um conceito influencia ou implica em outro. Mas as relações (representadas graficamente por arestas entre nodos do grafo) também podem representar outros tipos de significados quaisquer. Por exemplo, podem representar ideias conflitantes, podem representar generalizações ou agregações entre conceitos ou objetos, podem indicar sequências ou caminhos e por aí vai. Se as relações não tiverem direção, as arestas podem simplesmente significar que há uma relação entre 2 conceitos ou ideias.

O interessante é que um grafo permite ciclos. Uma hierarquia é um grafo onde um conceito só pode ter um "pai", ou seja, um conceito de mais alto nível. Um grafo de generalizações segue esta regra. Mas uma representação em rede permite ciclos, as relações podem representar voltas. Os trabalhos independentes de Albert Laszlo Barabasi e Paul Baran discutem os princípios básicos e tipos de redes.

Os mapas mentais são muito utilizados pela Gestão do Conhecimento para representar conhecimento (e não informações). Já as estruturas e análises multidimensionais são a base para o BI. Como juntar estes dois paradigmas ?

A Figura 44apresenta um mapa mental que representa também a visão multidimensional dos dados envolvidos na venda de um produto. Se alguém quiser ver pelo ponto de vista do BI tradicional, conseguirá ver uma tabela fato sobre vendas, tabelas de dimensões (vendedores, loja, propaganda, dados de clima, marca, data e hora) e tabelas secundários formando um esquema tipo floco de neve (snowflake).

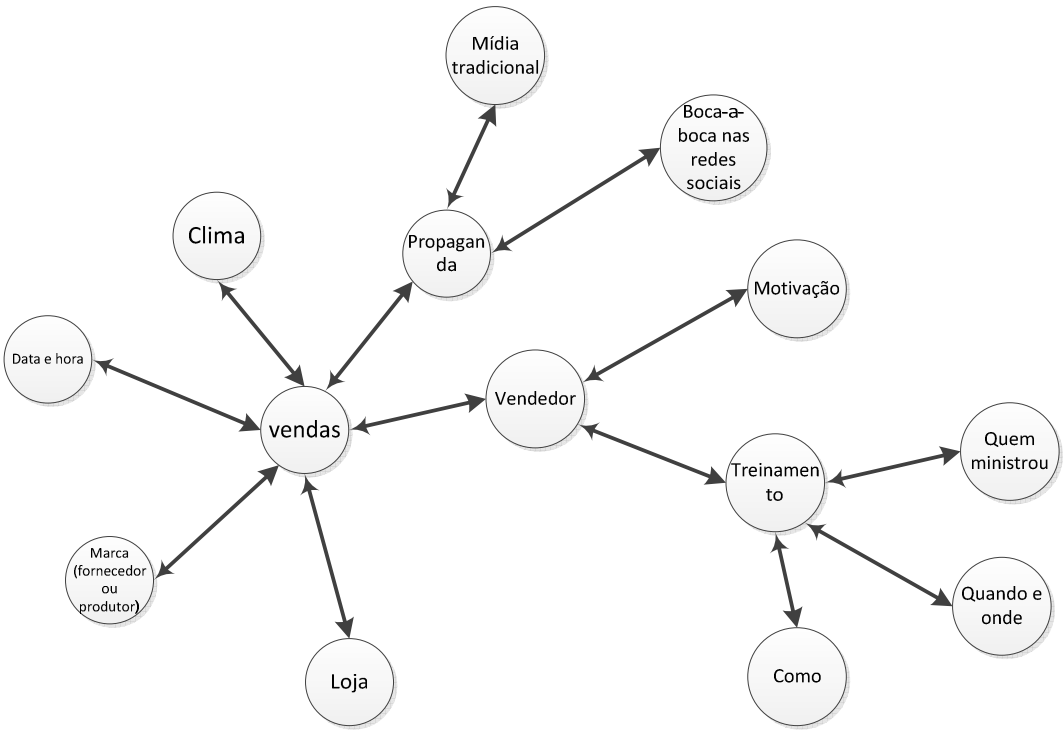


Figura 44: Mapa Conceitual sobre Fatos e Dimensões

Como um mapa mental, podemos ver os fatores que influenciam a venda. Diretamente, temos clima, loja, marca, propaganda, data hora e vendedor. Entretanto, o esquema mostra que o vendedor é influenciado pela sua motivação e pelo treinamento que recebeu. E o treinamento possui 3 fatores que influenciam.

Desta forma, podemos pensar nas causas para índices de vendas bons ou ruins analisando as causas diretas ou indiretas. O diferencial deste tipo de visualização é poder descobrir uma causa distante. Por exemplo, um baixo índice de vendas pode estar associados a quem ministrou o treinamento (que influencia a qualidade do treinamento, que por sua vez influencia o desempenho do vendedor, que finalmente influencia as vendas). Ou quem sabe o aumento das vendas pode ser devido à atitude dos vendedores, que por sua vez receberam um bom treinamento, e este foi de qualidade porque o ambiente do treinamento foi especial (quando e onde).

Uma rede de varejo estava tendo muitos problemas com mercadorias defeituosas, e queria diminuir tal prejuízo. Estes problemas foram detectados em todas as lojas. Então

o problema não era na loja. As mercadorias defeituosas vinham de diferentes fornecedores. Então o problema não estava também no fornecedor (ou na fabricação). Notou-se também que as mercadorias defeituosas vinham apenas dos Centros de Distribuição (CDs) número 1 e 2. Mas todos os CDs utilizam o mesmo processo padrão. Fez-se uma análise por observação (invisível) para saber se os funcionários estavam realizando o processo de forma diferente do planejado. Nada foi encontrado. Então o problema não era no processo específico de um ou alguns CDs.

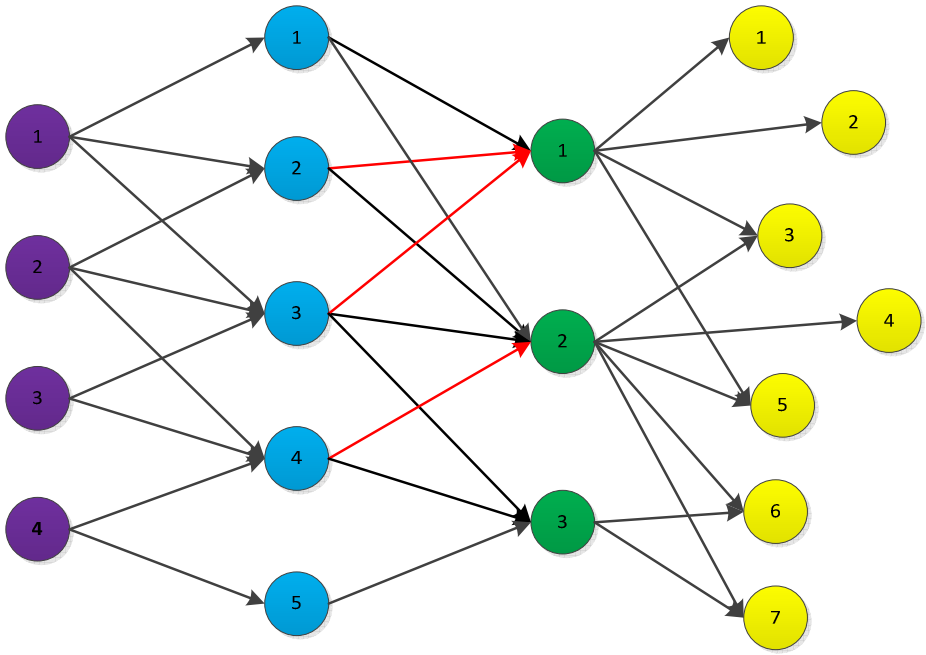


Figura 45: grafo para análise de causas

Uma constatação importante foi que as mercadorias defeituosas tinham sido entregues por apenas 3 transportadores: a 2, a 3 e a 4. Então procurou-se saber o que havia de comum entre estes transportadores. Nada foi encontrado. Pois estes 3 transportadores utilizam diferentes tipos de caminhões. Utilizando informações de rastreamento, ou seja, caminho percorrido pelas mercadorias defeituosas, procurou-se saber se algum tipo específico de caminhão havia sido utilizado para as mercadorias defeituosas. De novo, nada foi encontrado, pois as mercadorias defeituosas chegavam com diferentes tipos de caminhões.

Então, alguém teve a ideia de fazer um grafo, representando os caminhos percorridos e, incluir no grafo os diferentes tipos de caminhões utilizados. A Figura 45 representa o grafo gerado. Os círculos em roxo representam os fornecedores, os azuis representam as transportadoras, os verdes os CDs e os círculos amarelos são as lojas. As flechas representam o fluxo de mercadorias (todos os tipos), desde os fornecedores até as lojas.

Nesta figura, estão marcados em vermelho os caminhos que geraram mercadorias defeituosas. Então notou-se um padrão: 2 tipos de caminhões (X e Z) levaram as tais mercadorias. Mas estes caminhões levaram também mercadorias do mesmo tipo das

defeituosas e que não apresentavam problemas. E também levaram o mesmo tipo de mercadoria para o CD número 3, e ali não foram constatados defeitos neste tipo de mercadoria. Então o tipo de caminhão não era determinante do problema.

Mas uma constatação importante foi feita: quando um caminhão do tipo X ou Z fazia entregas no CD 1 ou 2, a entrega era feita de forma um pouco diferente. Como nestes CDs, a movimentação era maior, o processo de descarregar as mercadorias era feito com algumas alterações, feitas pelas pessoas sem conhecimento de quem planejou o processo todo. O mesmo tipo de caminhão, ao fazer entregas no CD 3, que tem menos movimento, não alterava o processo.

Em resumo, pode-se descobrir que a causa dos problemas era uma combinação de elementos do sistema de logística desta empresa. A representação visual permitiu identificar a combinação que gerava os problemas, algo que as planilhas e bancos de dados não mostravam.

A Figura 46 apresenta uma taxonomia (classificação hierárquica) de assuntos da área de Computação. Sobre ela, foram desenhadas conexões direcionadas, que significam a sequência com que os temas foram discutidos num fórum.

Pode-se verificar as mudanças de tema, se foram drásticas ou pequenas. Pode-se verificar os temas mais discutidos. Pode-se verificar se houve ciclos, ou seja, se a discussão voltou a temas já discutidos antes. E o número de conexões num ciclo permite saber se esta volta demorou a acontecer ou não.

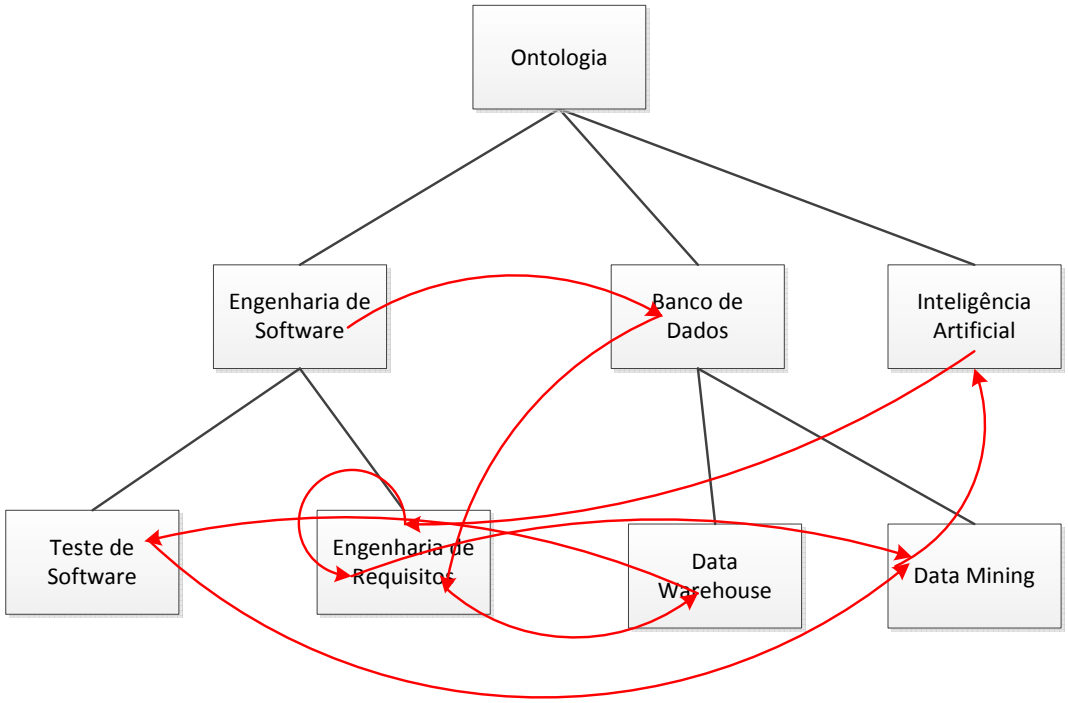


Figura 46: grafos combinados com hierarquias

Outro tipo de análise interessante sobre mapas mentais é fazê-los representando fluxos de informações, ou seja, quem fornece informação para quem (ver Formanski et al.). Nodos representam pessoas e arestas (setas) representam o fluxo de informação de uma pessoa para outra. As cores indicam o departamento ou setor de cada pessoa. A largura da seta representa o quanto de informação que passou naquela via. A Figura 47 mostra um exemplo.

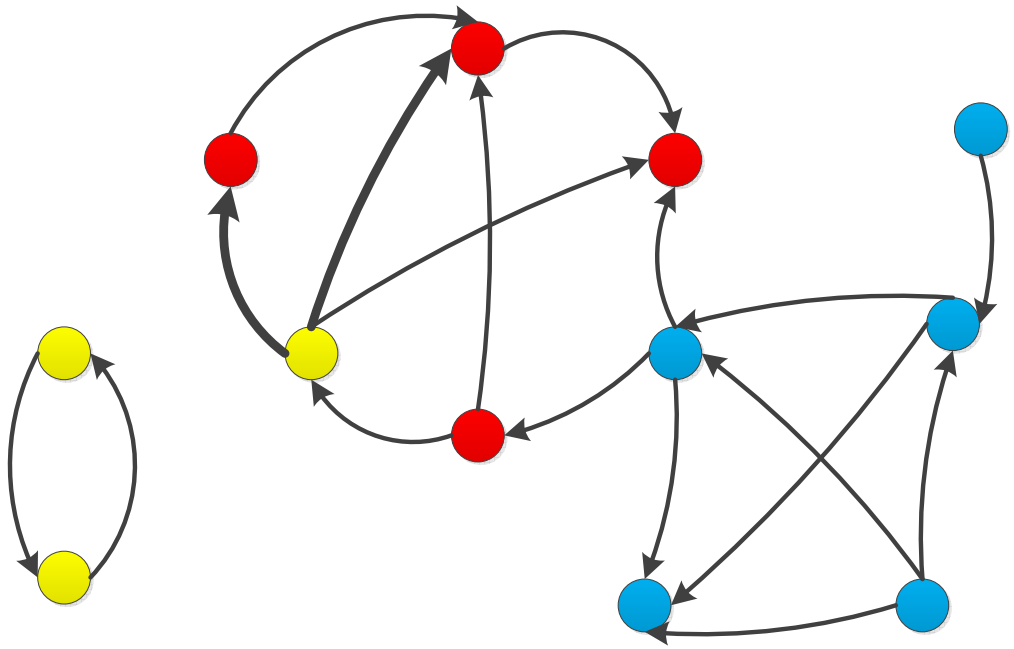


Figura 47: Grafo de comunicação entre membros de equipes

Várias análises podemos fazer a partir deste mapa:

a) podemos notar uma sub-rede isolada à esquerda com duas pessoas do departamento amarelo (identificadas como 1 e 2). Elas não trocam informações com pessoas de outro departamento. Além disto, há uma pessoa do departamento "amarelo" (3) que não interage com estas duas para trocar informações, mas que está bem "enturmada" com pessoas de outros departamentos. Provavelmente isto indica um problema a ser contornado. É necessário que esta 3a pessoa (identificado por 3) interaja com seus pares. E também seria possível pensar em como fazer com que os 2 funcionários "amarelos" (1 e 2) pudessem interagir com pessoas de outros departamentos.

b) podemos notar que há uma pessoa (11) que só recebe informações. Pode ser um novato, ainda aprendendo. E há alguém (12) que só fornece; pode ser alguém experiente, mas será que ele ou ela não deve receber algum tipo de informação de alguma outra pessoa ?

c) a pessoa identificada como 8 está interligando duas sub-redes, a azul e a amarela, provavelmente um elo ligação importante para juntar duas áreas de conhecimento.

d) a pessoa identificada por 9 está isolada, tendo somente contato com a pessoa identificada por 10. Pode ser que 9 seja um aprendiz, que deve ser "sombra" de 10.

e) fora a pessoa 9, a rede azul é a mais conectada, pois todos as pessoas deste setor interagem entre si. Já na sub-rede vermelha, o nodo 4 não interagem com 6 e 7. Há que se investigar o porquê disto, se é planejado assim ou se é um problema.

Determinismo X probabilismo

As arestas num grafo podem representar relações determinísticas de, por exemplo, causa-efeito. Mas também podemos usar grafos de probabilidades. Neste caso, as relações são prováveis e não há certeza absoluta. As Redes de Markov e as Redes Bayesianas utilizam o conceito de probabilidade para marcar relações entre nodos num grafo. As Redes Neurais Artificiais também utilizam pesos probabilísticos para as conexões entre os neurônios artificiais.

Num grafo de relações causais, as relações entre conceitos (causas e efeitos) recebem pesos numéricos indicando a probabilidade da relação. Isto permite raciocínio lógico (crisp ou fuzzy) sobre qual a causa mais provável, independente se a causa está direta ou indiretamente conectada ao efeito.

Os grafos ponderados (com pesos nas relações) também são úteis para que se possa identificar quais relações são de maior interesse para análise. Pesos muito altos podem sugerir relações mais importantes num contexto e relações com pesos muito baixos podem ser eliminadas por insignificância (principalmente para limpar um grafo com muitas conexões).

Descobrir novas ligações

Um dos casos mais interessantes de descoberta por mineração foi feita por Swanson e Smalheiser (1997). Eles conseguiram encontrar uma possível relação entre 2 textos de assuntos distintos. O texto 1 falava que "...o óleo de peixe é bom para a circulação do sangue...". O texto 2 dizia que "... a síndrome de Raynaud está associada com a vaso-constrição nas pessoas ...". A partir da leitura destes 2 textos, eles chegaram à hipótese de que "o óleo de peixe poderia ajudar no tratamento da síndrome de Raynaud". Entretanto, não havia na literatura médica científica nenhum texto que falasse de tal hipótese. Então eles partiram para experimentos práticos e os resultados comprovaram a hipótese.

Este problema pode ser esquematizado utilizando-se um mapa mental (ou grafo). Considerando os seguintes conceitos e suas relações:

- Síndrome de Raynaud ➔ vaso-constrição (relação de causa-efeito);
- Óleo de peixe ➔ boa circulação (relação de causa-efeito);
- vaso-constrição ⇔ boa circulação (relação de associação).

O mapa pode levantar a hipótese que há uma relação entre a Síndrome de Raynaud e o óleo de peixe. Generalizando, poderíamos construir um autômato que sugere novas ligações (a serem investigadas) a partir de grafos.

A partir da Figura 48, que relaciona conceitos, pode-se:
a) sugerir ligações para procurar: por exemplo, verificar se há ligação entre A e D (e de que tipo);
b) procurar evidências que liguem os conceitos: por exemplo, com a hipótese de relação entre A e D, procurar se há algum texto falando da relação entre A e D ou então podemos fazer algum experimento que concretize ou comprove esta relação;
c) procurar um conceito que ligue outros: por exemplo, será que existe um conceito X, tal que A ⇔ X ⇔ D ?

O método regressivo, proposto por Descartes, assume que uma solução existe como hipótese, bastando procurar por ela para comprovar a hipótese.

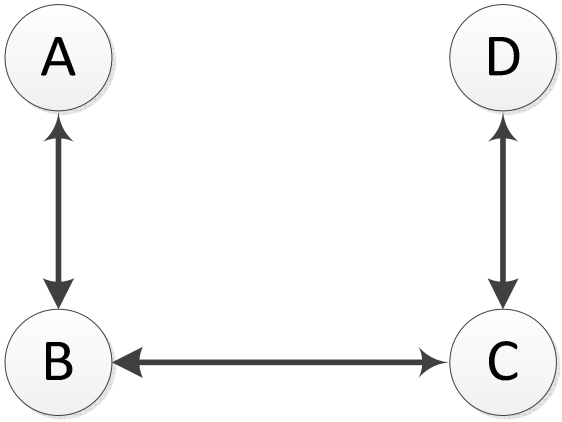


Figura 48: Grafo com relações entre conceitos

Este é um ótimo exemplo de como fazer as perguntas certas ajuda a encontrar as respostas certas. No caso, os referidos autores perguntaram qual a relação entre dois conceitos, que eles ainda não haviam notado em nenhum texto. O uso de mapas mentais pode ajudar neste tipo de investigação, para encontrar hipóteses iniciais, necessitando pouco ou talvez nenhum entendimento do domínio (para iniciar; depois sim será necessário um *background* especializado).

Mapas e informações geográficas

Mapas geográficos podem ser muito úteis para levantamento e validação de hipóteses. O diferencial do mapa geográfico é trazer informações que não aparecem em bancos de dados tradicionais, tais como distância, proximidade, tipo de terreno, etc.

Um dos casos mais famosos de análise de mapas e que permitiu descobrir causas está relatado no livro "The Ghost Map" de Steven Johnson. O livro conta a história do médico John Snow que descobriu a causa de mortes e a origem da cólera em Londres, em 1854. Naquela época, todos diziam (sabedoria popular) que a doença se alastrava pelo ar. Dr. Snow, a partir de seus conhecimentos, não acreditava nesta hipótese, mas

não sabia a real causa. Após posicionar num mapa da cidade todos os casos, Dr. Snow percebeu que havia mais mortes próximas de uma fonte de água. Sua hipótese então era de que a água seria o meio de transmissão. A análise temporal da disseminação de casos fortaleceu ainda mais a hipótese, pois os casos aumentavam com o tempo a partir da fonte de água. Por fim, as hipóteses do doutor foram confirmadas e muitas vidas salvas.

O trabalho de conclusão de Robson Jardim resultou num sistema automatizado para o registro colaborativo de casos de doenças e a geração posterior de relatórios de evolução da doença. Usuários cadastrados podem registrar o local onde o caso ocorreu, o tipo de doença e a data. O sistema permite aos usuários comparar a evolução e o deslocamento de casos de doenças em mapas ao longo do tempo. Na Figura 49, há um exemplo de como podemos ver o surgimento de novos casos em dois momentos diferentes, permitindo inferir uma direção de deslocamento da doença.

Hoje em dia, com a constante preocupação com novos vírus e a disseminação cada vez mais rápida de epidemias, uma ferramenta visual pode apoiar análises e dar subsídios para decisões de entidades de saúde e governos.

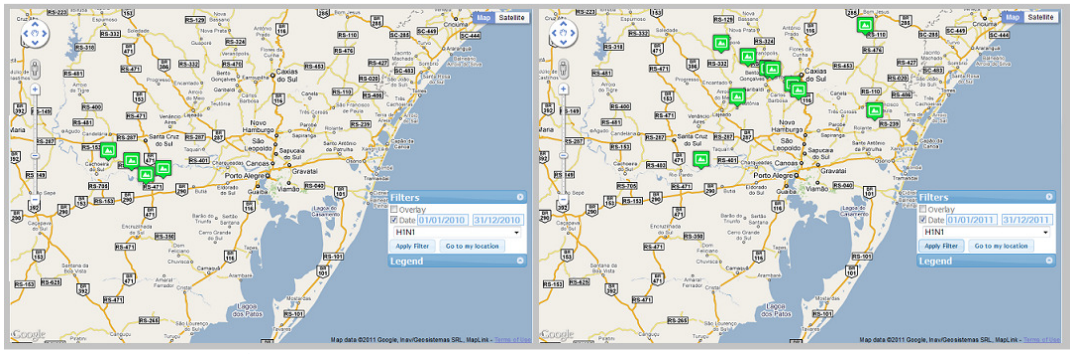


Figura 49: mapa para análise de evolução e disseminação de doenças

Uma Metodologia Associativa

Nesta seção, apresento um framework (esboço de metodologia) para utilização de mapas mentais/conceituais para análise de informações.

As informações são representadas por conceitos (nodos do grafo) e relações entre os conceitos (arestas, direcionadas ou não). Estas relações podem ser de diversos tipos (causalidade, conflito, exemplo, corroboração, instância, etc.). Associações entre conceitos são a forma como o cérebro humano funciona. É de onde tiramos a inteligência, conectando conhecimentos e ideias.

O mapa serve para representar causas (diretas ou indiretas) de eventos, relações entre eventos, relações entre causas, instâncias, generalizações, etc. É uma metodologia genérica que pode ser aplicada a diferentes contextos com o objetivo de facilitar o entendimento de um problema.

O framework aqui apresentado é baseado na metodologia L.E.SCAnning de Humbert Lesca (também discutida nos trabalhos de Caron-Fasan, Janissek-Muniz e Blanco.

Os passos da proto-metodologia são:

1. Levantar fatos ou evidências ou sinais fracos
2. Agrupar informações relacionadas (ex.: assuntos ou temas)
3. Identificar relações entre as partes de informações
4. Finalizar o Mapa Mental (informações e relações entre elas)
5. Validar Mapa (reavaliar conceitos e conexões)
6. Descoberta de conhecimento.

A coleta de informações (passo 1) pode inclusive considerar informações ainda não verificadas ou confirmadas. O objetivo é trabalhar com unidades de informação, que podem ser representadas por expressões ou frases curtas. Estas informações podem vir de fontes tais como notícias, livros, artigos, palestras, postagens em fóruns e redes sociais, boatos, relatórios internos da empresa, publicidade e reportagens públicas, etc. Também podem ser utilizados dados numéricos, vindos de relatórios ou estatísticas. Pode-se fazer um filtro inicial para focar num objetivo (ou então coletar tudo como no modo proativo discutido antes).

O passo 2 serve para agrupar informações por similaridade ou por assunto, usando marcadores (labels) para os grupos. O mapa mental já pode começar a ser feito. Pode-se usar um símbolo diferente (ex. quadrado) para dizer que há várias unidades de informação dentro de um conceito. A Figura 50 apresenta um exemplo deste passo da metodologia. Nela podemos ver grupos de informações já categorizados por assunto. Imagine que estas informações foram coletadas a partir de notícias, reportagens, propagandas, blogs, comentários de especialistas no assunto e etc.

Figura 50: Metodologia Associativa - passo 2

O passo 3 deverá identificar relações entre as unidades ou grupos de informações. Estas relações podem ser de causa, contradição, explicação, efeito, consequência, conflito, etc

(não há limites). O importante é deixar explícito no mapa o tipo da relação. Podem ser usados símbolos diferentes para os diferentes tipos de relações. Neste momento, se há uma relação entre 2 conceitos mas não se sabe o tipo, deve-se manter a conexão, mesmo sem a determinação do tipo (que depois será avaliado). É possível manter conceitos contraditórios, marcando este tipo de relação entre eles. A verificação da veracidade será feita mais tarde. A Figura 51 apresenta o grafo da figura anterior (mesmo exemplo) já com as relações entre conceitos.

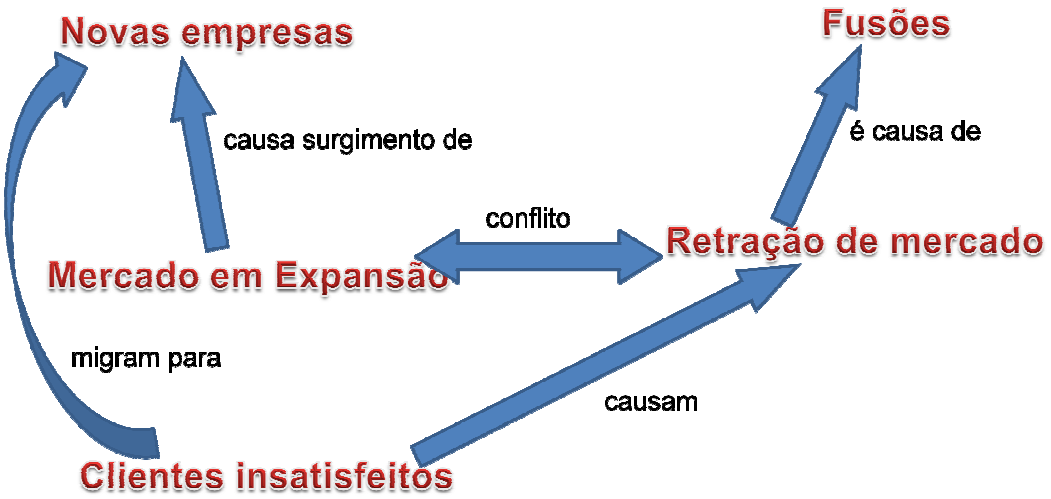


Figura 51: Metodologia Associativa - passo 3

O passo 4 consiste em analisar a consistência do mapa, revisando conceitos ou grupos e suas relações, eliminando conexões sem significado ou sem início ou fim.

Já no passo 5, devemos validar o mapa e suas informações. Neste ponto, deve-se revisar as conexões e os tipos e até mesmo a importância e veracidade dos conceitos. Pode-se inclusive colocar um grau de certeza nas informações e relações.

Por fim, o passo 6 refere-se à descoberta e análise. O objetivo é identificar hipóteses no mapa, identificar conhecimento novo e também identificar o que está faltando. Aqui também é possível incluir novas relações e mesmo verificar a falta de conexões (e incluir se for necessário). Deve-se interpretar o conjunto de informações e tirar as primeiras conclusões. Também as contradições devem ser resolvidas aqui (podendo-se eliminar informações não verificadas). A Figura 52 mostra o surgimento de um novo conceito ("novo serviço"), como uma nova hipótese e um conceito ("retração de mercado") que foi eliminado pois se verificou não ser verdade.

O mapa é como um quebra-cabeça (puzzle). Em alguns casos, pegamos uma unidade de informação separada e procuramos encaixá-la no mapa. Em outros momentos, verificamos a falta de alguma peça (conceito ou conexão) e vamos em busca de novas informações.

Figura 52: Novas hipóteses e revisão do mapa - metodologia associativa

O mapa serve também para raciocinarmos sobre as informações. Podemos até mesmo pensar em relações que ainda não foram descobertas ou não são existentes. Por exemplo, pensar o que os conceitos A e B tem a ver entre si (qual a relação) e aí buscar informações

Neste último passo também ocorre a análise de causalidade e influência: quem influência o que, ou é causa de.

Hipóteses podem ser acrescentadas ao mapa, sendo marcadas desta forma (para não confundir com informações já confirmadas).

O tal mapa mental poderá ser, num futuro breve, um autômato que sugira análises a serem feitas, novos conceitos ou relações, informações a verificar e até mesmo possíveis causas para efeitos.

Será como o conceito de biblioteca do futuro, proposto por Feigenbaum (1989). Ele compara as bibliotecas de hoje com as do futuro: as primeiras são como um armazém de objetos passivos, enquanto que as bibliotecas do futuro serão uma coleção de documentos ativos que ajudarão às pessoas fornecendo conexões desconhecidas, fazendo associações e analogias, sugerindo conceitos novos, descoberta de novos métodos e teorias.

10 Business Analytics

A evolução da área de BI gerou a chamada Business Analytics. O objetivo é poder prever acontecimentos ou predizer valores para variáveis. Por exemplo, "neste ritmo de vendas, alcançaremos a meta no dia ...". A ideia não é nova, apenas teve uma nova roupagem. Os sistemas de apoio à decisão (SAD ou DSS, em inglês) já há muitos anos vêm ajudando os tomadores de decisão. O funcionamento é simples: a partir de dados de entrada (parâmetros) e utilizando um modelo de decisão, pode-se prever valores futuros. Os modelos de decisão geralmente são do tipo what-if ("e se eu fizer isto, o que vai acontecer"), e utilizam técnicas como projeção, regressão e simulação.

O processo de BI está mais inclinado para explicações ou explanações e não tanto para previsão ou predição. Há uma diferença entre tentar explicar o que aconteceu e prever o que vai acontecer. As explicações, principalmente de causas, podem ser usadas para as previsões. Em geral, as previsões são baseadas em dados históricos e na construção de modelos de previsão.

Mas não pode haver confusão. O barômetro permite prever chuva mas não é causa do tempo. O pluviômetro mede índices de chuva mas também não são causas (e não servem para fazer previsões, mas seus registros podem ser utilizados para tal). Não é o ato de fumar que causa câncer mas sim as substâncias que estão no cigarro.

Business Analytics (BA) complementa BI uma vez que os padrões encontrados no passado podem ser testados no futuro. Por exemplo, uma rede de varejo identificou por BI que um aumento de 1% no preço final de produtos de um setor sempre reduzia as vendas totais deste setor em 0,5%. Uma etapa posterior de BA poderia avaliar as mudanças no lucro final da empresa para um período futuro, considerando que as demais condições não mudem. É claro que são utilizados simuladores (software) para fazer as previsões futuras e estes simuladores são baseados em modelos matemáticos (fórmulas sobre dados quantitativos).

Previsões

Como já dito antes, as previsões ajudam as empresas no seu planejamento e no seu dia a dia. Uma empresa que trabalhe com estoques que consiga prever quanto vai vender nos próximos dias, pode produzir ou comprar somente a quantidade que irá vender. Estoque parado é prejuízo porque a empresa precisa pagar infraestrutura para armazenar (local, pessoas, climatização, etc.) e se não vender o produto pode deteriorar (perder prazo de validade, estragar por condições climáticas adversas, etc.). Dizem que a Amazon será capaz de prever vendas e com isto antecipar sua logística. Ou seja, se ela predizer que um determinado cliente vai comprar um certo livro dentro de um mês, ela já vai enviar este livro para um local próximo ao cliente.

E as previsões também servem para validar hipóteses. Faça uma previsão a partir de um modelo e verifique se os eventos previstos acontecem. Isto permite refinar um modelo ou descartá-lo.

Mas o que é uma previsão boa ? Ela precisa acertar tudo, sempre e nos mínimos detalhes ? A qualidade de uma previsão é dada pela precisão. Mas nem sempre os valores ou eventos acontecem realmente como previstos, pode haver um certo desvio, que chamamos de margem de erro. A tendência é que os modelos e suas previsões errem mais no início e com o passar do tempo vão melhorando. Para isto é preciso fazer mais previsões e refinar o modelo a partir da avaliação das causas dos erros.

Também podemos avaliar os modelos e suas previsões pelo seu valor. Talvez a previsão erre, mas a margem de erro pode ser aceitável e a previsão ajude a tomar decisões. Imagine também uma indústria de refrigerantes. Deixar produto estocado é perda na certa. Ela precisa produzir quase como just-in-time. Então talvez uma previsão boa não precise de um valor exato para quanto ela vai vender (quanto as pessoas vão consumir ou comprar), mas um intervalo de valores já ajude.

A previsão tem que ser honesta, como nos aconselha Nate Silver. Ela não deve suscitar a fama pela sua grandiosidade. Ela precisa ser a melhor previsão que poderia ter sido feita. É claro que a previsão do clima para uma semana é inútil. Ela precisa ser boa para o dia corrente e não interessa se errar para mais dias, pois ela poderá ser refeita.

Nate Silver distingue previsão de projeção. Uma previsão é uma declaração definitiva e específica sobre quando e como acontecerá um evento (por exemplo, um terremoto de grandes proporções atingirá tal cidade no dia tal). Já uma projeção é uma declaração probabilística (por exemplo, há 60% de chance de ocorrer um terremoto em tal cidade nos próximos trinta anos).

Os grandes desafios dos modelos de previsão são:
1) construir o modelo e refiná-lo;
2) determinar os dados ou parâmetros que influenciam as previsões;
3) coletar estes dados a tempo de poder predizer e não só explicar os ocorridos.

E como já discutido no início deste livro, os modelos de comportamento que se aplicam a um determinado contexto talvez não funcionem em outros contextos ou épocas. Uma pequena mudança nas condições pode inviabilizar um modelo. As analogias, como discutido antes, precisam ser adaptadas.

Estas pequenas variações podem ser ruídos, como discutido por Nate Silver, ou podem ser variações do ambiente real. A Teoria do Caos (discutida no livro de James Gleick) diz que uma borboleta batendo asas no Brasil pode influenciar o clima no Japão. Esta ideia veio de um artigo apresentado em 1972, por Edward Lorenz. Lorenz descobriu que truncar um dado na terceira casa decimal fazia uma enorme diferença. A conclusão é que uma pequena mudança nas condições iniciais (o bater de asas de uma borboleta no Brasil) pode produzir uma divergência grande e inesperada nos resultados (um tornado no Japão). Não significa que o comportamento do sistema seja aleatório, como o termo "caos" talvez possa sugerir. Significa apenas que é muito difícil prever a atuação de certos tipos de sistemas, pois seria necessários coletar todas as variáveis que implicam no resultado e saber seu valor com muita precisão em tempo hábil.

As previsões mudam com o passar do tempo

Segundo Nate Silver, os sistemas complexos são influenciados pelo:
1. Dinamismo: significa que o comportamento do sistema em um dado momento influencia seu comportamento no futuro; e pela
2. Não linearidade: significa que o comportamento segue padrões exponenciais.

A extrapolação tende a causar dificuldades de previsão porque valores de alguns parâmetros crescem de forma exponencial. Por exemplo, o crescimento populacional e a disseminação de doenças precisam ser previstos de forma exponencial e não linear.

Ray Kurzweil fala da teoria do retorno acelerado. No caso, ele usa esta teoria para discutir previsões tecnológicas. Uma previsão não é linear. Imagine fazer previsões para 10 anos. Entretanto, esta previsão foi feita no tempo Zero. Após algum tempo após o marco Zero, digamos 2 ou 3 anos, as condições iniciais já mudaram. Ou seja, a previsão inicial não vale mais, precisaria ser refeita com as novas condições. E como as informações surgem de forma exponencial (por isto também o Big Data), elas podem ajudar a melhorar as previsões. E isto vai acelerando de forma exponencial.

Raposas X Porcos-espinhos

Nate Silver diz que há 2 tipos de pessoas que fazem previsões: as raposas e os porcos-espinhos.

Porcos-espinhos são personalidades que acreditam em grandes ideias, em princípios básicos ou leis que regeriam o mundo (como as leis da física) e que sustentam praticamente todas as interações que ocorrem na sociedade.

Raposas, por outro lado, são criaturas que vivem de fragmentos, que acreditam numa infinidade de pequenas ideias que juntas produzem algo maior. Tendem a ser mais tolerantes em relação à incerteza e às opiniões discordantes. Se os porcos-espinhos são caçadores e estão sempre em busca de uma grande presa, as raposas são animais coletores.

Previsões grandiosas e ousadas podem levar os porcos-espinhos à TV. Mas informações em excesso se tornam um mau negócio pois há mais variações. Porcos-espinhos constroem histórias que são mais nítidas e mais coerentes do que o mundo real, com protagonistas e vilões, vencedores e perdedores, clímax e desfechos, e, geralmente, um final feliz para o time pelo qual torcem.

Raposas usam mais dados. Porcos-espinhos usam poucos índices (reduzir algo complexo a poucas variáveis).

Estatísticas X Percepções humanas

As previsões podem não ser projeções, mas ainda assim são feitas com dados. Se não tivermos dados, é adivinhação, como tendo uma bola de cristal. Por isto, as estatísticas

são muito importantes. Não há como fazer previsões sem olhar para o passado e aprender com ele.

Por exemplo, no Brasil, o técnico de vôlei Bernardinho e sua equipe têm conseguido grandes resultados para o time nacional de vôlei usando estatísticas. Eles monitoram tudo o que é feito por cada jogador do time do Brasil e também dos adversários. Registram todos os tipos de jogadas, se resultaram em fracasso ou sucesso, como estava a posição dos jogadores, e com isto extraem relatórios de que jogadores estão melhor e quais estão com pior desempenho. Então, quando um brasileiro for "sacar", eles analisam em tempo real as estatísticas e verificam para que adversário deve ser direcionado o saque e de que forma (tipo de saque). E isto é feito para outras estratégias além do saque.

Michael Lewis, no livro Moneyball (que virou filme com Brad Pitt), faz uma grande discussão sobre esta dicotomia entre usar ou não estatísticas. Ele discorre sobre o caso real do Oakland Athletics, time de baseball americano, para expor seus argumentos. A questão toda se desenrola na diferença entre olheiros humanos e sistemas estatísticos para fazer previsões sobre jovens jogadores. Cada time escolhe os jogadores mais promissores no início da temporada. A grande maioria dos clubes utiliza, até hoje, os olheiros (scouts).

Os olheiros muitas vezes erram porque se preocupam mais com aparências. Então os sistemas baseados em estatísticas podem ser melhores pois não são influenciados por ruídos e variáveis que não implicam em resultados e conseguem se adaptar melhor a pequenas variações nos parâmetros. Por outro lado, os olheiros vão melhor em alguns casos porque usam uma abordagem híbrida, com uma quantidade maior de informações do que a oferecida apenas pelas estatísticas. E ainda acumulam informações com o passar do tempo (não são sistemas estáticos). Um bom olheiro também consegue informações privilegiadas, que a maioria não pode obter (por exemplo, no baseball, dados sobre a situação social e familiar do jogador).

É também o mesmo caso dos investidores das bolsas de valores. Se um investidor utilizar somente as informações públicas, a que todos têm acesso, não terá nenhuma vantagem. Os investidores precisam encontrar detalhes de informações que os outros não possuem.

O Oakland de Billy Beane teve um grande sucesso com estatísticas. Em outros casos porém, olheiros venceram o sistema Pecota de estatísticas. Já os Red Sox uniram olheiros (scouts) e estatísticas (nerds) e foram campeões em 2004. Lewis concluiu que as estatísticas funcionam melhor para jogadores de divisões inferiores do que para os da primeira liga. Mas nos níveis ainda mais inferiores, elas não funcionam.

Segundo Silver, meteorologistas melhoram em 25% as previsões de precipitações feitas por computador e em 10% as da temperatura. Neste caso, as informações visuais são melhores interpretadas por seres humanos do que pelo computador. É por isto que muitos sistemas na Web utilizam figuras tipo *captcha* para distinguir usuários humanos de robôs.

O uso de intuições para previsões

Uma boa ideia então é combinar dados estatísticos com intuição, e não somente usar um ou outro. Onde a intuição não é detalhista, os dados podem nos ajudar a lembrar detalhes. Onde a estatística não é completa, a observação humana pode completar uma análise.

Em geral, as pessoas procuram diminuir a incerteza das decisões mas assumem certos riscos pela racionalidade limitada. Por exemplo, se alguém quiser traçar uma rota de fuga em caso de incêndio num prédio, talvez não consiga avaliar todas as alternativas possíveis (local de início do fogo, quantidade de pessoas, etc.). E no momento da situação de incêndio, o ser humano tem que simplificar ao máximo seu processo de decisão para acelerar as ações. Isto quer dizer que os planos iniciais podem ter sido esquecidos ou terão que ser simplificados. E assim, as atitudes planejadas mudam pela racionalidade limitada. E o ser humano então utiliza intuições para acelerar a decisão.

Já falamos antes que a intuição é um palpite, mas não uma adivinhação. Ela deve ser precedida por dados. O ser humano possui uma certa capacidade para tomar decisões rápidas com pouca informação. Isto não significa que devemos tomar decisões por pressa. A intuição não deve ser confundida com caminho mais fácil (preguiça). Gunther recomenda não confiar na primeira impressão, e sugere que coletemos muitos dados. Kahneman também concorda: é um grande risco tomar decisões usando a área preguiçosa e irracional do cérebro.

Daniel Kahneman (2012), ganhador do Prêmio Nobel de Economia em 2002, diz que temos dois sistemas de tomada de decisão: um rápido e outro devagar. O sistema rápido é utilizado por exemplo para reconhecer rostos. Até bebês o usam. E a gente não precisa raciocinar, é automático, sem esforço. Utiliza associações e reconhecimento de padrões, sendo difícil de controlar ou modificar. Já o sistema devagar é usado para, por exemplo calcular quantas horas tem em 4 dias. Ele é serial, controlável, flexível, governado por regras e exige muito esforço.

Ambos os sistemas são importantes. O segundo sistema é o que acreditamos ser mais comum e mais correto. Seria como um processo racional de decisão. Entretanto, nossas vidas estão cheias de exemplos de decisões certas que foram tomadas pelo sistema rápido.

Por exemplo, grandes negócios são fechados somente após o encontro presencial entre as partes. Os homens de negócios dizem que é importante "olhar nos olhos". Isto também serve para contratações para empregos. Koestler sugere que as pessoas devam ter conhecimentos generalizados, sobre outras áreas, além da sua especialização. Isto pode ajudar inconscientemente, com dados novos e analogias. Gunther cita Alfred P. Sloan, ex-executivo da GM: "o ato final da decisão é intuitivo". Isto porque é uma escolha entre alternativas. Ninguém sabe qual a melhor alternativa ou se uma delas vai dar certo ou não. Se soubéssemos, não seria decisão e sim "bola de cristal".

Não há nada que garanta o resultado, seja utilizando dados estatísticos ou intuições. Mas é melhor para uma decisão ter mais dados (sejam confirmados ou não).

11 Novos tipos de dados, técnicas de coleta e análise

Este capítulo aborda questões periféricas ao tema de BI, mas que podem ajudar cientistas de dados e analistas de BI.

11.1 Coleta explícita X implícita X por inferência

A coleta de dados explícita acontece quando perguntamos algo a alguém (num entrevista ou questionário) e a pessoa nos dá os dados em forma de resposta. Ou então quando alguém preenche um formulário na Web ou nos diz algo, mesmo sem a gente pedir.

Já a coleta implícita é aquela que utiliza a observação. Não conheço estabelecimento que faça isto, mas é um futuro provável: quando você paga em dinheiro num supermercado, este só registra o que você comprou e como; não ficam registrados dados como seu sexo, idade, etc. Mas imagine que o operador do caixa (check-out) possa observar o cliente e utilizar códigos para dar entrada no sistema de dados que ele está vendo (sexo, faixa etária, estilo de se vestir, se está acompanhado ou não).

Num futuro um pouco mais distante isto já poderá ser feito através da análise de imagens gravadas com câmeras. Já foi feito um experimento que, pelo contorno da pessoa diante de um banner, era possível identificar o sexo e a faixa etária. Paco Underhill e parceiros fazem consultoria para empresas de varejo analisando estatisticamente o comportamento de clientes em lojas. As informações são coletadas por observação direta no ambiente ou em gravações de imagens.

Com esta onda de Big Data por aí, está todo mundo coletando dados sobre todos. A operadora de celular sabe por onde a gente anda e quando. Qual o caminho que costumamos fazer, por onde costumamos andar em cada dia da semana e horário. E se instalarmos aplicativos tipo o Waze no nosso celular, a Google (que comprou o Waze) vai saber até a que velocidade estamos andando. E daí inferir se estamos a pé ou de carro, ou num engarrafamento. Aí alguém inventou a tecnologia de RFID, e ela está em cartões com chips, carros, produtos novos e vai estar em sacolas, carrinhos de supermercados, etc. Então não é só por celular. Os aplicativos e softwares que usamos em celulares, tablets, notebooks e etc também estão avisando onde estamos, se estivermos conectados via Wifi, 3G ou 4G.

Este tipo de coleta também é considerada implícita, apesar de não usar a observação humana. Neste caso, a observação é feita sobre dados eletrônicos, capturados por dispositivos eletrônicos.

Inferir é gerar uma informação a partir de outra. Se você compra muito produto congelado no supermercado, a análise destes dados pode ajudar a inferir que:
a) você tem um bom freezer em casa;
b) você não sabe cozinhar ou não gosta;

c) você é uma pessoa muita atarefada e não tem tempo nem para cozinhar.

A coleta por inferência então é quando o sistema gera informações novas a partir de outras. O nível de inferência é subjetivo de cada organização e certamente aumenta a incerteza sobre a veracidade da informação. Mas muitas empresas assumem o risco desta incerteza, porque mais incerto ainda é não saber nada sobre o cliente.

Tempos atrás surgiram alguns artigos falando sobre Phenomenal Data Mining. Que significa tentar inferir eventos ou atributos de entidades a partir de coleções de dados. É na prática e com seriedade fazer aquela brincadeira de analisar os restos no lixo de alguém. Aí você saberá que tipo de pessoa é, pelo que compre e consome (marcas, tipos de produtos, faixas de preços, etc). Assim, se você compra Xampu feminino e desodorante feminino juntos na mesma compra, você é uma mulher. Se comprar Xampu para carro, esponja para lavar carro e creme para polimento de carro, você certamente tem um carro. É claro que há margem para erros.

E utilizando a sabedoria das massas, se numa loja de supermercado a venda de água mineral foi muito acima do normal, é porque faltou água neste bairro. E se na mesma cidade, várias farmácias estão vendendo antigripal, é porque há um surto de gripe. E provavelmente a temperatura também esfriou ou a umidade aumentou.

E isto já chegou à Internet. O Facebook já consegue inferir nossa orientação sexual e tendência política só analisando nossas "curtidas" (ler a reportagem "Estudo mostra que botão 'Curtir' do Facebook revela muito mais do que se imagina sobre o usuário http://oglobo.globo.com/tecnologia/estudo-mostra-que-botao-curtir-do-facebook-revela-muito-mais-do-que-se-imagina-sobre-usuario-7812419).
Um exemplo caso aconteceu em algumas sinaleiras de grandes cidades. Uma pessoa passava pelos carros perguntando ao motorista se queria ganhar um brinde. A grande maioria das pessoas dizia que sim, mesmo que desconfiadas. Então o "entrevistador de sinaleiras" pedia o nome e o telefone do motorista, alegando que depois entraria em contato.

A princípio, parece que só foi utilizada a coleta explícita (perguntas e respostas). Mas se pararmos para pensar, a pessoa só se dirigia a certos tipos de carros. Além disto, anotava mais que o nome e o telefone. Ela anotava o tipo de carro e outros dados que conseguisse coletar (adesivos informando que há bebês no carro, sobre estacionamentos hospitalares, associações e clubes, etc). Então este é um tipo de coleta implícita, por observação.

Além disto, os dados iam para centrais onde eram então analisados. A partir dos dados coletados explícita ou implicitamente, alguém iria fazer uma inferência. Por exemplo, a partir do selo de estacionamento de médicos num hospital, pode-se inferir a profissão de médico; daí tem-se o perfil de pessoas com boa renda e alto senso crítico. Se o carro tinha cadeira de bebês, infere-se que há uma família por trás.

11.2 Novas tecnologias para coletar e monitorar dados

Novas tecnologias estão surgindo para coletar dados. Chips e antenas de RFID permitem rastrear produtos e até mesmo pessoas (bem como GPS e celulares). A análise de vídeos (imagens) permite capturar movimentos e gestos. Capturas de sons permitem a posterior análise e o reconhecimento de fala. Já há diversos dispositivos para identificação de pessoas por biometria (até mesmo tatuagens já servem para isto).

Diversos sensores estão sendo fabricados e utilizados nas mais diversas situações. Sensores de movimento alertam para intrusos. Sensores de umidade e luminosidade são utilizados na agricultura de precisão. Sensores de rotação são comuns em jogos em aparelhos móveis, mas também servem para estabilizar veículos. Computadores de bordo também usam sensores de proximidade para estacionar de forma autônoma um carro. A medicina no futuro irá utilizar sensores para medir sinais de saúde nas pessoas.

O professor Petland faz pesquisas com sensores para coletar expressões faciais e utilizar isto para melhorar a comunicação. O pesquisador Kevin Warwick implantou sensores em seu corpo. O futurista Michio Kaku fala que haverá em breve diagnóstico médico por imagens capturadas pelo espelho do banheiro ou pela câmera do celular.

11.3 Web Mining

Técnicas de Web Mining procuram encontrar padrões no comportamento de usuários na Web. Tais técnicas estatísticas são aplicadas sobre dados de usuários web e sobre históricos de suas ações em sites Web. Como as ações de usuários ficam registradas em arquivos de log, nos servidores Web onde os sites ficam instalados, é possível ter relatórios estatísticos sobre diferentes tipos de informações, tais como a origem dos visitantes (analisando o número IP de suas máquinas), qual seu sistema operacional e navegador, qual a última página vista antes de chegar ao site, etc.

As técnicas de Web Mining mais básicas calculam médias ou somas de variáveis numéricas tais como tempo que usuários passam num site ou lendo uma página, número de visitantes por hora, mês ou dia da semana, número de hits (ações de um usuário no site), e com isto geram relatórios de páginas mais acessadas, assuntos mais procurados ou lidos, e métricas como a taxa de conversão (quantos usuários compraram um produto em relação à quantidade que viu o produto no site).

Uma técnica mais avançada é a que analisa a sequência de clicks ou páginas vistas por um usuário numa sessão em um site. Esta sequência é chamada de *clickstream*, e indica o caminho percorrido pelo usuário desde que entrou no site até sua saída (última página vista). A análise de *clickstreams* é importante para conhecer a estratégia dos usuários até seu objetivo, ou para saber se alguém estava perdido no site sem saber como chegar ao objetivo, ou para diferenciar as estratégias de usuários com perfis diferentes. Por exemplo, pode-se comparar os *clickstreams* mais comuns entre usuários que compram e comparar com o padrão de usuários que não compra. Talvez o projeto do site não esteja ajudando estes últimos a chegarem a seus objetivos. Ou a empresa pode descobrir que o diferencial está na página que apresenta o preço dos produtos.

Se o usuário puder ser identificado, seja por login, cookies ou outra forma, é possível saber quantas revisitas são feitas ao site, inferir o interesse do usuário e também enriquecer tais dados com informações vindas de outras bases, tais como cadastros em lojas físicas.

Hal Varian, economista-chefe do Google, na sede da empresa em Mountain View, Califórnia diz que eles podem prever o número de pedidos iniciais de seguro-desemprego com mais antecedência porque, se correrem boatos de que haverá demissões em alguma empresa, as pessoas vão começar a pesquisar 'onde e como dar entrada no seguro-desemprego' e termos afins (citado no livro de Nate Silver).

Para mais detalhes sobre esta tecnologia ver o meu livro sobre 31 tipos de sistemas de informação.

11.4 Text Mining

Estima-se que 80% das informações de uma companhia estão contidas em documentos textuais. Os textos podem ser e-mails, postagens em blogs, microblogs e redes sociais, arquivos eletrônicos (txt, doc, pdf, ppt, documentos digitalizados), comentários em páginas web e até mesmo textos resultantes de pesquisas e questões abertas. Este volume grande de informações textuais impossibilita a análise das informações de forma manual. Isto não só pela quantidade, mas pela complexidade das informações neste formato, o que exige trabalho intelectual para interpretação dos textos. Outro problema com análise manual é que se perde a noção estatística do conteúdo destes textos.

Text Mining ou Mineração de Textos ou Descoberta de Conhecimento em Textos (KDT – Knowledge Discovery in Texts) é uma evolução das áreas de Recuperação de Informações (*Information Retrieval*) e Extração de Informações (*Information Extraction*). As técnicas de Text Mining tem por objetivo aplicar técnicas estatísticas diretamente sobre os textos. No caso de Data Mining, que é aplicado sobre dados estruturados, as técnicas estatísticas são aplicadas sobre campos e valores de tabelas ou planilhas. Entretanto, no caso de textos, não temos campos, valores ou mesmo tabelas. E precisamos aplicar as técnicas sobre o conteúdo dos textos. Pois bem, o conteúdo dos textos é formados por palavras (unidade de informação). Então, Text Mininig iniciou-se desta forma, aplicando técnicas estatísticas sobre palavras de textos.

Entretanto, a análise de palavras isoladas traz problemas de interpretação, conhecidos como o "problema do vocabulário" (*vocabulary problem*). O mesmo assunto ou evento pode ser abordado ou relatado com diferentes palavras (sinônimos, variações linguísticas, etc). Além disto, há palavras polissêmicas (com mais de um significado).

Uma das soluções é utilizar um vocabulário controlado, como fazem os médicos através do CID (Classificação Internacional de Doenças), para evitar mal entendidos. Mas quando se trata de web e textos populares, não há como garantir uniformidade.

Da mesma forma, poderemos ter problemas analisando reclamações de clientes se encontramos a expressão "gostei" e não analisarmos as palavras ao seu redor. Pode ser que exista um "não" antes e isto muda completamente o significado.

Então, a estratégia mais apropriada para Text Mining é identificar conceitos (contextos ou temas ou assuntos) nos textos e aplicar as técnicas estatísticas sobre os conceitos. Para identificar os conceitos, deve-se usar uma base ou ontologia de conceitos, na qual estão definidas as diferentes formas de um conceito aparecer num texto (sinônimos, expressões, etc).

Por exemplo, a presença de sintomas de alcoolismo em prontuários médicos pode ser verificado pela presença de uma das seguintes expressões: álcool, hálito etílico, faz uso de bebidas, bebe imoderadamente.

Então o conceito "alcoolismo" será definido de forma a serem analisadas estas expressões. Se uma delas aparecer, o texto estará tratando deste conceito.

Uma vez que as palavras formam a unidade básica de informação dos textos e sobre elas será feito o text mining, é necessário algum tratamento prévio antes de aplicar estatística. Por exemplo, corretores ortográficos ajudam a eliminar variações incorretas de palavras.

Text Mining utiliza as mesmas técnicas de Data Mining que podem ser aplicadas a variáveis nominais ou qualitativas, tais como classificação, clustering, associação, sequência temporal e análise de distribuição. Além disto, há nova técnicas como análise de diferenças e similaridade entre textos e a técnica de geração automática de resumos de textos.

Para mais detalhes sobre esta tecnologia ver o meu livro sobre 31 tipos de sistemas de informação.

11.5 Análise de Sentimentos

As empresas estão preocupadas com sua imagem. É importante saber o que estão falando dela ou de seus produtos e serviços. Para obter tal conhecimento, a empresa pode usar pesquisas de campo com clientes potenciais ou fazer pesquisas tipo "survey" com uma amostra de seus clientes. Entretanto, nem sempre as pessoas se sentem confortáveis para reclamar ou falar mal.
Para estes casos existe a Internet, zona livre de censura e restrições. Mas não estamos falando de analisar notícias, nem sites específicos para reclamações como o *Reclameaqui*. No primeiro caso, depende-se da parcialidade da fonte e, no segundo caso, pode ficar em aspectos muitos específicos de alguns poucos clientes (há uma estatística que diz que apenas 95% dos clientes insatisfeitos fazem reclamações formais).

A ideia é vasculhar a Web atrás de oceanos de opiniões, procurando saber o que a grande massa tem por dizer (Wisdom of Crowds - Sabedoria das Massas). Hoje cada cliente é um "prosumidor" (consumidor + produtor), que deseja expressar suas opiniões, dar ideias, ajudar a empresa ou outras pessoas. E para isto utiliza as redes sociais (Twitter, Facebook, Google+) ou cria blogs e fóruns para reunir grupos de pessoas interessadas na mesma discussão.

O sucesso depende da capacidade de coletar tais dados informais e da velocidade em analisar seu conteúdo, para gerar decisões sábias em tempo hábil. A área de Análise de

Sentimentos (Sentiment Analysis) ou Mineração de Opiniões (Opinion Mining) nasce como uma das alternativas. Seu objetivo é encontrar opiniões e analisar seu conteúdo. Na prática, o que deve ser feito é encontrar na Web textos que possam conter opiniões de pessoas e analisar o tipo de sentimento presente nos textos: se positivos ou negativos (se falam bem ou falam mal).

O processo depende da existência de uma ontologia de tarefa ou de domínio, que permita entender como as pessoas escrevem sobre um determinado assunto e como elas expressam seus sentimentos positivos e negativos. Após, um processo de inferência probabilístico ou determinístico é utilizado para identificar o tipo de sentimento.

A ontologia de tarefa ou de domínio é um conjunto organizado de palavras e expressões linguísticas (multipalavras), separadas por tipo de sentimento. Pode-se utilizar um método determinístico (quando a presença de certas palavras diz com certeza que um sentimento está presente num texto) ou um método probabilístico. Neste último caso, as palavras da ontologia devem ter pesos associados, indicando a probabilidade de a palavra ou expressão indicar um certo tipo de sentimento. A inferência então é feita com métodos probabilísticos (por exemplo, métodos bayesianos). Assim, o resultado é um grau de certeza de que um sentimento esteja presente no texto sendo analisado.

Esta ontologia de aplicação pode ser incrementada para que a análise seja feita sobre sentimentos mais detalhados (e não somente positivos ou negativos). Alguns autores utilizam o modelo POMS (Profile of Mood States), utilizado por psicólogos, para identificar o estados de humor. Este modelo utiliza 6 tipos de humor:

1. Tensão-Ansiedade:
    - tenso, tranquilo, nervoso, impaciente, inquieto e ansioso.
2. Depressão-Melancolia:
    - triste, desencorajado, só, abatido (deprimido), desanimado e infeliz
3. Hostilidade-Ira:
    - irritado, mal humorado, (rabujento), aborrecido, furioso, com mau feitio, e enervado.
4. Vigor-Actividade:
    - animado, activo, enérgico, alegre e cheio de boa disposição
5. Fadiga-Inércia:
    - esgotado, fatigado, exausto, sem energia, cansado e estourado.
6. Confusão-Desorientação:
    - confuso, baralhado, desnorteado, inseguro, competente e eficaz.

Tal modelo já foi utilizado para comprovar a correlação entre postagens do twitter e acontecimentos do mundo real. Por exemplo, pode-se analisar o sentimento predominante nas postagens antes, durante ou depois de um evento, sejam as eleições presidenciais ou o Dia de Ação de Graças. Também é possível saber o ritmo das postagens para cada tipo de humor, analisando-se subidas e descidas num gráfico que represente o total de postagens de cada tipo.

Outro modelo que pode ajudar a detalhar sentimentos, é o Modelo OCC de Ortony, Clore e Colins. Este modelo trabalha com 22 tipos de emoções, agrupando adjetivos que exprimem tais emoções em textos.

Resumindo, técnicas de análise de sentimentos são úteis para avaliar opiniões de clientes efeitos ou potenciais, mas também para refinar ideias (pois a empresa pode analisar o sentimento das pessoas sobre determinados assuntos antes que produtos e serviços sejam lançados).

12 Conclusão

Ao fazer BI, o cientista ou analista deve ter em mente que é preciso ter um objetivo. Como já discutimos durante o livro, talvez o objetivo não esteja muito claro no início (esta é a abordagem proativa), mas irá se delinear durante o processo. Portanto, não há como terminar um processo de BI sem se ter avaliado se algum objetivo foi alcançado. Muitas empresas coletam todos os tipos de dados possíveis, sem mesmo saber se vão usar ou não. Outras fazem todo tipo de análise sem bem saber qual o objetivo por trás disto. Empresas analisam perfis de clientes, coletam dados pessoais e privativos, invadem privacidade, mas para quê ?

O Big Data pode ser analisado com técnicas e ferramentas. Mas será que precisamos de tantos dados ? Isto muitas vezes causa a sobrecarga e depois o estresse de quem faz. E também pode causar problemas para clientes. Muitas empresas são coletando dados demais sobre as pessoas, como invasão de privacidade. O que temos que nos perguntar é se o que estamos fazendo irá trazer mais resultados positivos ou negativos. Ou seja, vai fazer mais mal ou bem ? E para quem.

Outro cuidado para o cientista de dados é querer encontrar padrão em tudo. Isto pode virar um TOC (transtorno obsessivo-compulsivo). Popper (1980, p.17) nos diz: "... fenômeno psicológico do pensamento dogmático ou, de modo geral, do comportamento dogmático: esperamos encontrar regularidades em toda parte e tentamos descobri-las mesmo onde elas não existem; os eventos que resistem a essas tentativas são considerados como 'ruídos de fundo'; somos féis a nossas expectativas mesmo quando elas são inadequadas - e deveríamos reconhecer a derrota". O mundo é caótico por natureza. Em alguns casos a gente vê padrões, mas na maioria parece uma bagunça mesmo. E daí ? O importante é conseguir viver neste contexto. Foi isto que causou a evolução dos seres vivos, justamente a capacidade de adaptar-se a ambientes diferentes. Isto implicou no desenvolvimento de habilidades melhores e a consequente sobrevivência por mais tempo.

Um conselho final é aproveitar o que os números podem nos dar mas não acreditar que os números sempre serão melhores que nossas intuições e sentimentos.

O Futuro do BI

O futuro do BI provavelmente está no 4o paradigma: o uso intensivo de dados (data-intensive science) com novos métodos científicos, com sistemas de software mais poderosos, com mais semântica a partir dos dados, mas acima de tudo com o intelecto e a sensibilidade de humanos.

Sistemas inteligentes poderão sugerir novas conexões, descobrir novas regras, padrões, hipóteses e conhecimentos. Mas somente humanos poderão incorporar técnicas de criatividade e conhecimentos para a integração de diferentes disciplinas, para análise de novos cenários, para solução de problemas, para identificação de causas.

Bibliografia

AGRAWAL, Rakesh; IMIELINSKI, Tomasz. Database mining: a performance perspective. IEEE Transactions on Knowledge and Data Engineering, v.5, n.6, Dezembro de 1993.

ANDEL, Pek Van. Anatomy of the Unsought Finding. Serendipity: Origin, History, Domains, Traditions, Appearances, Patterns and Programmability. The British Journal for the Philosophy of Science, v.45, n.2, Junho, 1994, p.631-648.

ANSOFF, H. Igor. Strategic issue management. Strategic Management Journal, v.1, n.2, April/June 1980, p.131–148.

ASUR, Sitaram; HUBERMAN, Bernardo A. Predicting the Future with Social Media. Proceedings WI-IAT '10 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - v.1, 2010, p. 492-499.

BARABASI, Albert-Laszlo; BONABEAU, Eric. Scale-free networks. Scientific American, n.288, Maio de 2003, p.50-59.

BARABASI, Albert-Laszlo. Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life. Plume, 2003.

BARAN, Paul. On Distributed Communications Networks. the Rand Corporation, Setembro de 1962.

BERTIN, Jacques. Semiology of Graphics: Diagrams, Networks, Maps. University of Wisconsin Press, 1983.

BLANCO, S.; CARON-FASAN, M. L.; ,LESCA, H. Developing capabilities to create collective intelligence within organizations. Journal of Competitive Intelligence and Management, v.1, n.1, Spring 2003.

BOLLEN, Johan; MAO, Huina; ZENG, Xiao-Jun. Twitter mood predicts the stock market. Journal of Computational Science, 2(1), March 2011, 1-8.

CARON-FASAN, Marie Laurence; JANISSEK-MUNIZ, Raquel. Análise de informações de inteligência estratégica antecipativa coletiva: proposição de um método, caso aplicado e experiências. Revista de Administração, São Paulo, v.39, n.3, jul/ago/set 2004, p.205-219.

CHOI, H.; VARIAN, H. Predicting the Present with Google Trends. Economic Record, special issue selected Papers from the 40th Australian Conference of Economists, v. 88, n.1, p.2–9, June 2012.

CHOUDHURY, Vivek; SAMPLER, Jeffrey L. Information specificity and environmental scanning: an economic perspective. MIS Quarterly, Março de 1997.

CLARKE, Ronald V.; ECK, John E. Crime analysis for problem solvers in 60 small steps. Center for Problem-Oriented Policing, U.S. Department of Justice.

DAWKINS, Richard. O Gene Egoísta. Companhia das Letras, 2007.

DESCARTES, René. O discurso do método. São Paulo: Martins Fontes, 2001. (original: Discours de la methode, 1637)

DOMINGOS, Carlos. Oportunidades disfarçadas: histórias reais de empresas que transformaram problemas em grandes oportunidades. Sextante, 2009.

DOYLE, Arthur Conan. Um Estudo em Vermelho. Tradução de Hamílcar de Garcia. Publicado em "As Aventuras de Sherlock Holmes, Volume I". Círculo do Livro. (Original: A Study in Scarlet. Almanaque Beeton's Christmas Annual, novembro, 1887).

DUGAS, A. F. et al. Influenza Forecasting with Google Flu Trends. Online Journal of Public Health Informatics, v.8, n.2, Fevereiro de 2013.

DUHIGG, Charles. O Poder do Hábito - Por que fazemos o que fazemos na vida e nos Negócios. Objetiva, 2012.

EKMAN, Paul; ROSENBERG, Erika L. (ed.) What the Face reveals - basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). New York: Oxford University Press Inc., 1997.

EKMAN, P.; FRIESEN, W.V.; HAGER, J.C. FACS - the Facial Action Coding System. 2a. ed. Salt Lake City: Research Nexus eBook. London: Weidenfeld & Nicolson, 2002.

FEIGENBAUM, E. A.. Toward the Library of the Future. Long Range Planning, v. 22, n. 1, 1989, p.118-123.

FORMANSKI, José Gilberto; FORMANSKI, Filipi Naspolini; RODRIGUEZ y RODRIGUEZ, Martius Vicente. A contribuição da análise de redes sociais na identificação dos conhecimentos críticos em uma organização: um estudo de caso. Anais do KM Brasil 2012. São Paulo: SBGC, agosto 2012.

FORSTER, Malcolm R. Probabilistic Causality and the Foundations of Modern Science. Ph.D. Thesis, University of Western Ontario. 1984.

GENG, Liqiang; HAMILTON, Howard J. Interestingness Measures for Data Mining: A Survey. ACM Computing Surveys, v.38, n.3, 2006.

GHANI, Rayid; SIMMONS, Hillery. Predicting the End-Price of Online Auctions. International Workshop on Data Mining and Adaptive Modelling Methods for Economics and Management held in conjunction with the 15th European Conference on Machine Learning (ECML/PKDDD), Pisa, Itália, 2004.

GIGERENZER, Gerd; GAISSMAIER, Wolfgang. Heuristic Decision Making. Annual Review of Psychology, v.62, 2011, p.451–482.

GLADWELL, Malcolm. Blink - a decisão num piscar de olhos. Rocco, 2005.

GLADWELL, Malcolm. Outliers - the story of success. Back Bay Books, 2011.

GLADWELL, Malcolm. O ponto da virada - como pequenas coisas podem fazer uma grande diferença (original: the tipping point). Rio de Janeiro: Sextante, 2013.

GLEICK, James. Caos - a criação de uma nova ciência. Rio de Janeiro: Campus, 1989.

GUNTHER, Max. O Fator Sorte. Rio de Janeiro: Best Business, 2013 (original: The luck factor, 1977).

HARVEY A. Carr. An introduction to space perception. 1935.

HEY, Tony; TANSLEY, Stewart; TOLLE, Kristin. The Fourth Paradigm: data-intensive scientific discovery. Redmond: Microsoft Research, 2009.

INGWERSEN, Peter. Congnitive perspectives of information retrieval interaction: elements of a cognitive IR theory**.** Journal of Documentation, v.52, n.1, Março de 1996.

ISHIKAWA, K. Introduction to quality control. Productivity Press, 1990.

JARDIM, Robson Bresolin; LOH, S (orientador). Portal colaborativo para construção de mapas sobre evolução de doenças epidemiológicas. 2011. (Trabalho de Conclusão do Curso de Sistemas de Informação, Universidade Luterana do Brasil)

JOHNSON, Steven Berlin. The Ghost Map: The Story of London's Most Terrifying Epidemic and How It Changed Science, Cities, and the Modern World. Riverhead Hardcover, 2006.

KAHNEMAN, Daniel. Rápido e Devagar - Duas Formas de Pensar. Objetiva, 2012.
GORR, Wilpen L. et al. Forecasting Crime. 1999.

KOESTLER, Arthur. The Act of Creation - a study of the conscious and unconscious processes in humor, scientific discovery and art. New York: Arkana (The Penguin Group), 1964.

KORTH, Henry; SILBERSCHATZ, Abraham. Database Research Faces the Information Explosion. Communications of the ACM, v. 40, n.2, Fevereiro de 1997, p.139-142.

KRING, Ann M.; SLOAN, Denise M. The Facial Expression Coding System (FACES): development, validation, and utility. Psychological Assessment, v.19, n.2, Junho de 2007, p.210-24.

KUHLTHAU, Carol C. Inside the search process: information seeking from the user's perspective. Journal of the American Society for Information Science, v.42, n.5, June 1991.

KUHN, Thomas S. A Estrutura das Revoluções Científicas. 10.ed. São Paulo: Perspectiva, 2011 (original: 1962).

LENAT, Douglas B. The nature of Heuristics. Artificial Intelligence, v.19, n.2, Outubro de 1982, p.189-249.

LESCA, Humbert. Veille stratégique: la méthode L.E.SCAnning. Colombelles: Editions SEM, 2003.

LEWIS, Michael. Moneyball: The Art of Winning an Unfair Game. W. W. Norton & Company, 2004.

LEVITT, Steve. D.; DUBNER, S. J. Freakonomics: A Rogue Economist Explores the Hidden Side of Everything. William Morrow Paperbacks, 2009.

LOH, Stanley. 31 tipos de sistemas de informação - 31 maneiras de a tecnologia da informação ajudar as organizações. Porto Alegre, 2014.

LOSEE, John. A Historical Introduction to the Philosophy of Science. 4a.ed. New York: Oxford University Press, 2001. (original 1972)

MALTZ, Michael D.; KLOSAK-MULLANY, Jacqueline. Visualizing Lives: New Pathways for Analyzing Life Course Trajectories. Journal of Quantitative Criminology, v.16, n.2, June 2000, p.255-281.

MAATHUIS, Marloes H.; COLOMBO, Diego; KALISCH, Markus; BÜHLMANN, Peter. Predicting causal effects in large-scale systems from observational data. Nature Methods 7, April 2010, p.247–248.

MILLER, George A. The Magical Number Seven, Plus or Minus Two: Some Limits on OurCapacity for Processing Information. The Psychological Review, v. 63, 1956, p. 81-97.

MISHNE, Gilad. Predicting movie sales from blogger sentiment. In AAAI Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW) 2006.

MORAES, Maurício. Big Brother Obama. Revista Info, Editora Abril, n.324, dezembro de 2012.

MORIN, Edgar. Os Sete Saberes Necessários à Educação do Futuro. 2.ed. São Paulo: Cortez; Brasília: UNESCO, 2000.

MOSCAROLA, Jean; BOLDEN, Richard. From the data mine to the knowledge mill: applying the principles of lexical analysis to the data mining and knowledge discovery process. Note de Recherche nº 98-15, Université de Savoie. Setembro de 1998.

OARD, Douglas W.; MARCHIONINI, Gary**.** A conceptual framework for text filtering. Technical Report, University of Maryland. Maio de 1996.

ORTONY, A.; CLORE, G. L.; COLINS, A. The Cognitive Structure of Emotions. Cambridge University Press. 1988.

PARSAYE, Kamran et alli. Intelligent databases: object-oriented, deductive hypermedia technologies. New York: John Wiley & Sons, 1989.

POPPER, Karl. The logic of scientific discovery. Londres: Hutchinson & Co., 1959.

POPPER, Karl. Conjecturas e Refutações. Brasília: Editora da UnB. 1980.

PORTO, Celmo Seleno. Semiologia Médica. 5.ed. Guanabara Koogan, 2005.

RADINSKY, Kira; HORVITZ, Eric. Mining the web to predict future events. Proceedings WSDM '13 Proceedings of the sixth ACM international conference on Web search and data mining, 2013, p. 255-264.

SARGUT, Gökçe; McGRATH, Rita Gunther. Learning to Live with Complexity. Harvard Business Review, special issue on Complexity, September 2011.

SENGE, P. The Fifth Discipline: The art & practice of the learning organization. New York: Doubleday, 1990.

SENGE, P. et al. A Quinta Disciplina: Caderno de Campo. Rio de Janeiro: Qualitymark, 1995.

SILVA, Ricardo. Causality. Encyclopedia of Machine Learning, Springer, 2010, p.159-166.

SILVER, Nate. O sinal e o ruído: por que tantas previsões falham e outras não. Rio de Janeiro: Intrínseca, 2013.

SIMON, Herbert A. "Theories of Bounded Rationality". In McGUIRE, C.B. & RADNER, R. (ed.). Decision and Organization. Amsterdam: North-Holland Publishing Company, 1972.

SMITH, John Miles; SMITH, Diane C. P. Database abstractions: aggregation and generalization. ACM Trans. on Database systems, v.2, n.2, junho, 1977, p.105-133.

SPINK, Amanda; WOLFRAM, Dietmar; JANSEN, Major B. J.; SARACEVIC, Tefko. Searching the web: The public and their queries. Journal of the American Society for Information Science and Technology, v. 52, n.3, 2001, p. 226–234.

STEWART, Thomas R. Uncertainty, judgment and error in prediction. In: SAREWITZ, D.; PIELKE, R. A.; BYERLEY, R. Prediction: Science, Decision Making and the Future of Nature. Washington: Island Press, 2000, p. 41-57.

SWANSON, Don R.; SMALHEISER, N. R. An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artificial Intelligence, Amsterdam, v.91, n.2, p.183-203, Apr. 1997.

TOLE, A. A. Big Data Challenges. Database Systems Journal, v. IV, n. 3, 2013, p.31-40.

TSAMARDINOS, Ioannis; TRIANTAFILLOU, Sofia. Introduction to causal discovery: A Bayesian Networks approach. ECML-PKDD, Causal Discovery Tutorial, 2011.

TVERSKY, Amos; KAHNEMAN, Daniel. Belief in the law of small numbers. Psychological Bulletin, v.76, n.2, 1971, p.105-110.

TVERSKY, Amos; KAHNEMAN, Daniel. Judgment under uncertainty: heuristics and biases. Science, n.185, 1974, p.1124-1131.

TVERSKY, Amos; KAHNEMAN, Daniel. Extensional versus intuitive reasoning: the conjunction fallacy in probabilistic reasoning. Psychological Review, n.90, 1983, p.293-315.

UCHIDA, Naoshige; KEPECS, Adam; MAINEN, Zachary F. Seeing at a glance, smelling in a whiff: rapid forms of perceptual decision making. Neuroscience, v.7, Junho de 2006, p.485-491.

UNDERHILL, Paco. Why we buy: the science of shopping. Simon & Schuster, 1999.

WILSON, Timothy D. Strangers to Ourselves: Discovering the Adaptive Unconscious. Belknap Press of Harvard University Press, Maio de 2004.

WINSTON, Robert. Instinto humano. São Paulo: Globo, 2006.

WOLF, Gary. The Data-Driven Life - What happens when technology can analyze every quotidian thing that happened to you today ? The New York Times Magazine Maio de 2010.