



# DATA DISCOVERY E ANALYTICS

Rodrigo Moravia



PUC Minas  
Virtual

# Data-Driven

# Panorama geral

A movimentação de dados está a todo vapor.

Usar dados de forma eficaz não é apenas sobre qual banco de dados você usa ou quantos cientistas de dados você tem na equipe, mas sim uma interação complexa entre os dados que você possui, onde eles estão armazenados e como as pessoas trabalham com eles, e quais problemas são considerados valiosos.

Enquanto a maioria das pessoas se concentra na tecnologia, as melhores organizações reconhecem que as pessoas estão no centro dessa complexidade. Em qualquer organização, as respostas a perguntas como quem controla os dados, a quem se reportam e como escolhem no que trabalhar são sempre mais importantes do que usar um banco de dados.

# continuação

Deseja-se ver mais organizações bem-sucedidas com dados.

Os dados mudarão a maneira como as empresas interagem com o mundo e queremos que mais pessoas tenham acesso.

Para ter sucesso com os dados, as empresas devem desenvolver uma cultura de dados.

# continuação

Um cientista de dados não faz nada fundamentalmente novo. Há muito tempo temos estatísticos, analistas e programadores.

O que há de novo é a maneira como os cientistas de dados combinam várias habilidades diferentes em uma única profissão:

- Matemática, principalmente estatística e álgebra linear.
- Habilidades de computação, incluindo programação e design de infraestrutura.
- Habilidade de Comunicação

# continuação

Os cientistas de dados são valorizados por sua capacidade de criar narrativas em torno de trabalho deles. Eles não vivem em um mundo abstrato e matemático; eles entendem como integrar os resultados em uma história maior e reconhecem que, se seus resultados não levarem à ação, esses resultados não terão sentido.

Além dessas habilidades, um cientista de dados deve ser capaz de fazer as perguntas certas. Essa habilidade é mais difícil de avaliar do que qualquer habilidade específica, mas é essencial. Fazer as perguntas certas envolve conhecimento e experiência de domínio, juntamente com uma capacidade aguçada de ver o problema, ver os dados disponíveis e combinar os dois.



PUC Minas  
Virtual

# Organizações Data Driven

# O QUE É UM ORGANIZAÇÃO DATA-DRIVEN?

As organizações orientadas a dados mais conhecidas são as empresas de Internet de consumo: Google, Amazon, Facebook e LinkedIn. No entanto, ser orientado a dados não se limita à Internet.

O Walmart é pioneiro no uso de dados desde a década de 1970. Foi uma das primeiras organizações a construir grandes DWs para gerenciar inventário em seus negócios. Isso permitiu que ela se tornasse a primeira empresa a ter mais de US\$ 1 bilhão em vendas durante seus primeiros 17 anos.



# continuação

Na década de 1980, o Walmart percebeu que a qualidade de seus dados era insuficiente e, para adquirir dados melhores, tornou-se a primeira empresa a usar leitores de código de barras nas caixas registradoras. A empresa queria saber quais produtos estavam sendo vendidos e como a colocação desses produtos na loja afetava as vendas. Também precisava entender as tendências sazonais e como as diferenças regionais impactavam seus clientes. À medida que o número de lojas e o volume de mercadorias aumentavam, aumentava a complexidade de sua gestão de estoque. Graças aos seus dados históricos, aliados a um modelo preditivo rápido, a empresa conseguiu gerenciar sua curva de crescimento.

# continuação

- GRANDE INVESTIMENTO EM TECNOLOGIA DE INFORMAÇÃO
- MAIOR DB DO PLANETA QUE GERA O MAIOR DW DO PLANETA
- 30 X MAIOR QUE TODO O CONTEÚDO DA BIBLIOTECA DO CONGRESSO DOS EUA
  - N TRILHÕES DE DADOS ( n TERABYTES)
  - DW DO WAL-MART= 600 Tb - 1 Petabyte - TERADATA
  - INFORMAÇÕES GERENCIAIS
    - PERFIL DE CONSUMIDOR - SAMS CLUBE
    - DIAS MAIS PROPÍCIOS PARA COMPRA DE PRODUTO X POR CONSUMIDOR TIPO Y
- INVESTIMENTOS TOTAIS DE MAIS DE 3 BI DE DÓLARES DESDE 88
- SÓ EM 2002 APLICOU 400 MI DE DÓLARES



Realizaram um movimento gigantesco de logística, com deslocamento de frotas de caminhões gigantes, contendo esses dois produtos, para as áreas ameaçadas. Ganharam muito dinheiro!



Em setembro de 2004, o furacão Frances se aproximava das costas da Flórida



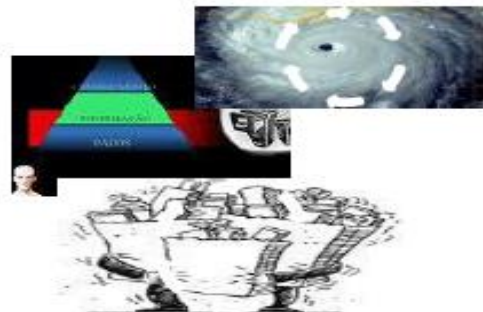
Poder da Informação

A equipe de BI do Wal Mart especialista em análises preditivas atípicas: Quais seriam os produtos de maior consumo e que mereceriam cuidado da área de logística da empresa, nas regiões atingidas?



Verificaram que o perfil de consumo durante o furacão anterior(Charley) apontava para dados não óbvios. Não era lanterna, pregos, reforço de portas e janelas, etc

Analisaram os dados de consumos, registrados no DW, relativo à passagem do outro Furacão-Charley, em agosto de 2004





Realizaram um movimento gigantesco de logística, com deslocamento de frotas de caminhões gigantes, contendo esses dois produtos, para as áreas ameaçadas. Ganharam muito dinheiro!



Em setembro de 2004, o furacão Frances se aproximava das costas da Flórida



Verificaram que havia tido um consumo extraordinário de cerveja (fase pré-tufão) e de torta de morango (strawberry tarts, na fase pós-tufão). Essa tinha sido consumida 7 x mais que a média



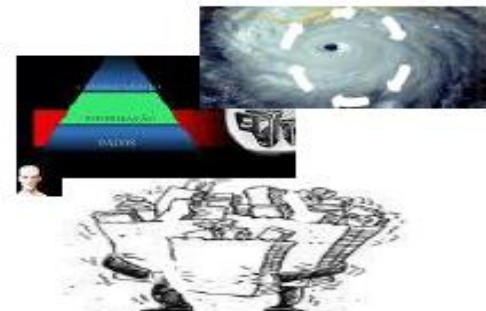
Poder da Informação

A equipe de BI do Wal Mart especialista em análises preditivas atípicas: Quais seriam os produtos de maior consumo e que mereceriam cuidado da área de logística da empresa, nas regiões atingidas?

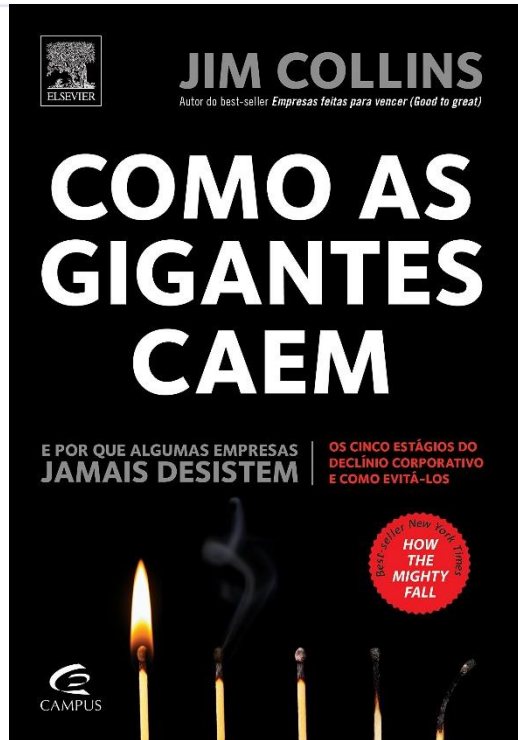


Verificaram que o perfil de consumo durante o furacão anterior(Charley) apontava para dados não óbvios. Não era lanterna, pregos, reforço de portas e janelas, etc

Analisaram os dados de consumos, registrados no DW, relativo à passagem do outro Furacão-Charley, em agosto de 2004



# O que NÃO é?





# O que é?



Let data drive decisions, not the Highest Paid Person's Opinion.

confiança  
objetividade  
transparência  
**cultura**  
iteratividade  
aprendizado  
questionamento

“

Uma empresa orientada aos dados  
adquire, processa e utiliza dados em  
tempo hábil para criar eficiências, iterar  
e desenvolver novos produtos, e  
navegar no cenário competitivo.

”

# Como criar essa cultura?

As melhores organizações orientadas a dados se concentram incansavelmente em manter seus dados limpos.

Os dados devem ser:

- organizados
- bem documentados
- livres de erros.
- A limpeza dos dados costuma ser a parte mais desgastante. Organizações bem-sucedidas investem pesadamente em ferramentas, processos e auditorias regulares.
- Em resumo... Investir pesado na **Qualidade de Dados**



# Como criar essa cultura?

- Uma das distinções mais importantes entre organizações que são orientadas por dados e aquelas que não são é como elas abordam a formulação de hipóteses e a resolução de problemas.
- Todas as organizações orientadas por dados seguem alguma variante do método científico, que chamamos de método científico de dados:
  - Comece com dados.
  - Desenvolva intuições sobre os dados e as perguntas que eles podem responder.
  - Aproveite seus dados atuais para entender melhor se é a pergunta certa a ser feita. Se não, itere até que você tenha uma hipótese testável.

# Dicas - Tempo

- Ocorrerá algum tipo de alerta, por exemplo, por e-mail, se houver uma mudança nos dados?
- Se o painel contiver um gráfico de pizza, você saberá se há uma mudança?
- Em vez de um único painel, crie diferentes painéis que reflitam diferentes escalas de tempo. Por exemplo, alguns painéis podem estar em escala horária, enquanto outros podem estar em escala trimestral.
- Mas se houver muitos alarmes, você cria fadiga de alarmes: a equipe fica insensível aos alertas, porque eles ocorrem com muita frequência e geralmente não têm sentido.

# Dicas - Valor

- Gerencie seus painéis em vez de permitir que eles gerenciem você. Revise-os e pergunte se eles ainda estão dando valor a você. Se não, altere-os.
- É surpreendente a frequência com que as pessoas consideram seu painel “fixo” e imutável. É exatamente o oposto: o painel é uma entidade viva que permite gerenciar sua organização.
- À medida que a sofisticação dos dados da organização aumenta, é provável que as formas antigas de medir o sistema se tornem muito simplistas. Portanto, essas medidas mais antigas devem ser substituídas por outras mais recentes.

# Dicas – Visual e História

- Dashboards precisam ser atraentes para prender atenção do usuário
- Montar de acordo com a história que ele precisa transmitir ao usuário



PUC Minas  
Virtual

# Estudo de caso Twitter

# Estudo de Caso

- Em 2009, o **Twitter** enfrentou um desafio. Houve uma grande empolgação com o serviço, mas as pessoas simplesmente não o usavam regularmente: três em cada quatro pessoas parariam de usá-lo em dois meses. Para resolver o problema de engajamento, o Twitter começou fazendo perguntas e analisando seus dados atuais. Ele encontrou uma série de resultados surpreendentes. Primeiro, os usuários que usaram o serviço pelo menos sete vezes no primeiro mês tiveram mais de 90% de probabilidade de retornar nos meses subsequentes. Para muitas organizações, identificar esse número mágico seria mais do que suficiente. Mas o Twitter continuou a estudar os dados e foi bem recompensado. Entre os usuários com alta retenção, descobriu-se que uma vez que um usuário seguisse 30 ou mais pessoas, essa pessoa quase certamente se tornaria um usuário de longo prazo.

# continuação

- A empresa continuou a cavar e descobriu que a natureza das pessoas seguidas também era essencial. Dois terços dos novos usuários que seguira eram puramente por conteúdo, mas um terço teve que seguir os novos usuários de volta.
- De posse desses fatos, a equipe do Twitter conseguiu descobrir uma solução que contrariava o pensamento convencional sobre a integração de novos usuários. Sites como Facebook e LinkedIn apresentavam aos novos usuários um “importador de catálogo de endereços” que rastreava os endereços de e-mail do usuário. Em seguida, um novo usuário veria uma página cheia de sugestões para “pessoas que você pode conhecer”. Quaisquer outras páginas que possam adicionar “atrito” à experiência do usuário fariam com que um número significativo (20%+) de usuários abandonasse a integração.

# continuação

- A análise mostrou que o Twitter precisava:
  - (a) ensinar novos usuários o que era um tweet,
  - (b) sugerir contas com conteúdo de alta qualidade segmentado por categorias (por exemplo, NFL, NBA, sites de notícias) e, em seguida...
  - (c) sugerir outros usuários com alta probabilidade de seguir alguém depois de conhecer essa pessoa estava no Twitter.
- A implementação dessas ideias adiciona atrito ao processo de integração ao ensinar os usuários sobre o tweet; também coloca as pessoas com as quais um usuário provavelmente irá interagir por último.



# continuação

- No entanto, o resultado não foi uma diminuição de novos usuários, mas sim um aumento de 30% no número de pessoas completando a experiência e um aumento de 20% no engajamento de longo prazo!
- Em retrospectiva, o processo e os resultados parecem quase mágicos. Eles estão longe disso; eles representam adesão dedicada ao método científico de dados. **Demorou cerca de 2,5 anos para chegar e testar esses resultados** – e o processo está longe de estar completo. Testes regulares estão em andamento para melhorar ainda mais o que acontece quando novos usuários chegam. O método científico de dados nunca para.



PUC Minas  
Virtual

# Perguntas de Organizações

## Data Driven Vs Não Data Driven

# Alguns exemplos empresas NÃO Data Driven

- Quantos itens do produto X foram vendidos no mês passado?
- Como foram as vendas por região no natal do ano passado?
- Quantas unidades do produto X foram devolvidas no mês passado?
- Qual foi a receita e o lucro da empresa no último trimestre?
- Quantos empregados foram contratados no último ano?

Em contraste, considere *análises preditivas*, como modelos que:

- Otimizam o gasto com publicidade,
- reposição da cadeia de abastecimento, ou
- minimizam a rotatividade de clientes.

Empresas *data-driven* respondem mais geralmente *quem, o quê, quando, por que e onde* fazendo recomendações e previsões, contando uma história em torno das descobertas.

# Alguns exemplos empresas Data Driven

- Quantos itens do produto X serão vendidos no próximo mês?
- Como serão as vendas por região no natal desse ano?
- Quantas unidades do produto X serão devolvidas no próximo mês?
- Qual será a receita e o lucro da empresa no próximo trimestre?
- Quantos empregados devem ser contratados para o próximo ano?

No entanto, tais recomendações e previsões (*insights*) exigem:



Coletar os dados corretos.  
Os dados devem ser dignos de  
confiança.



Uma boa análise, onde as  
perspectivas são considerados na  
decisão.



Conduzem a decisões concretas.



Comece com os dados



Desenvolva intuições sobre os dados e as perguntas que podem ser respondidas.



Formule perguntas.



Aproveite os dados atuais para melhor entender se é a pergunta certa a fazer. Se não, repita até que você tenha uma hipótese testável.



Crie um framework onde você pode executar testes/experimentos.



Analise os resultados para desenhar insights sobre a pergunta.

Ser orientado por dados não significa ignorar seu instinto.

# Alguns exemplos empresas Data Driven



*Coca-Cola*





A ideia de *Stranger Things* foi gerada a partir da coleta de dados sobre as preferências de milhares de usuários da Netflix, de acordo com aquilo que eles costumam assistir. A partir da análise do chamado *big data*, chegou-se à conclusão de que uma série com determinadas características teria tudo para dar certo. E foi exatamente isso o que aconteceu.

# REFERÊNCIAS

ANDREATTO, R. Construindo um Data Warehouse e Analisando suas Informações com Data Mining e OLAP. Monografia Final de Curso. Faculdade de Ciências Administrativas, Faculdade de Valinhos. 1999.

BARBIERI, Carlos. BI - Business Intelligence: Modelagem e tecnologia. Rio de Janeiro, Axcel Books, 2001.

CARVALHO, B.F. Arquiteturas de Ferramentas OLAP. SQL Magazine, Rio de Janeiro, ano 1, ed. 9, p.12-16, 2004.

DWBrasil. OLAP. 2007. Disponível em Acesso em 24 outubro 2007.

INMON, W. H. Como Construir o Data Warehouse. Rio de Janeiro: Campus, 1997.

# REFERÊNCIAS

Imagens: 270948935 Nome do autor: Rose Carson URL de Portfólio:

<https://www.shutterstock.com/g/Mosesstudio>

Imagens: 1911700096 Nome do autor: JLStock Sobre: Multidisciplinary acumen in all creative designs. URL de Portfólio: <https://www.shutterstock.com/g/JLStock>

Imagens: 2156187219 Nome do autor: kavi designs Sobre: Hai peoples, I am a graphic designer, vector illustrator. and photographer. URL de Portfólio: [https://www.shutterstock.com/g/kavi designs](https://www.shutterstock.com/g/kavi%20designs)



**PUC Minas**  
**Virtual**