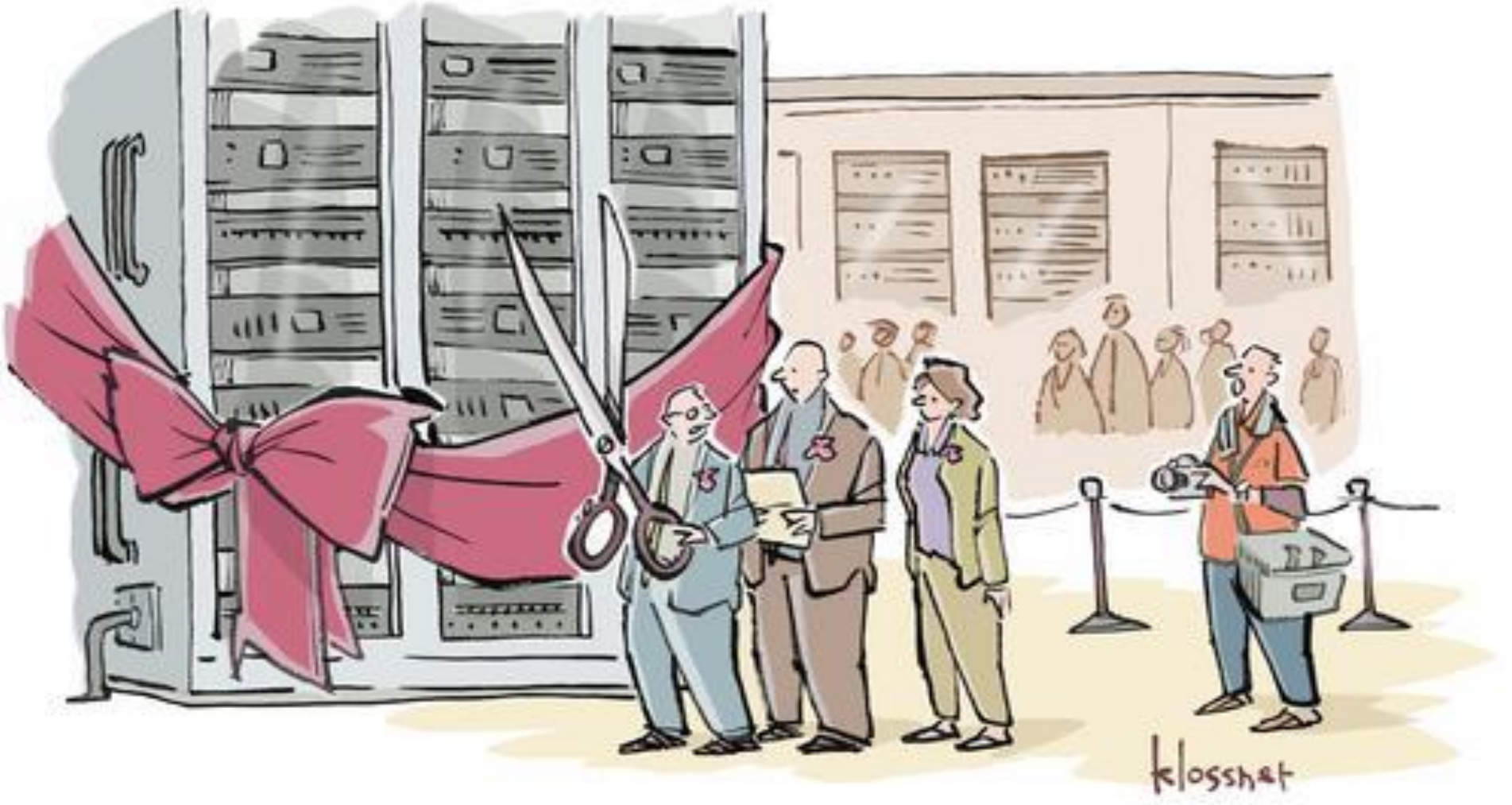
FGV EMAp
ESCOLA DE MATEMÁTICA APLICADA

Introdução à Data Science & Data Mining


klosshat
"IS THIS A GOOD TIME TO TELL YOU I DON'T KNOW WHAT 'BIG DATA' MEANS?"

O que é Big Data?

UNIT	SIZE	WHAT IT MEANS
Bit (b)	1 or 0	Short for "binary digit," after the binary code (1 or 0) computers use to store and process data.
Byte (B)	8 bits	Enough information to create an English letter or number in computer code. It is the basic unit of computing.
Kilobyte (KB)	1,000 or 2^{10} bytes	From "thousand" in Greek. One page of typed text is 2KB.
Megabyte (MB)	1,000KB; 2^{20} bytes	From "large" in Greek. The complete works of Shakespeare total 5MB. A typical pop song is about 4MB.
Gigabyte (GB)	1,000MB; 2^{30} bytes	From "giant" in Greek. A two-hour film can be compressed into 1-2GB.
Terabyte (TB)	1,000GB; 2^{40} bytes	From "monster" in Greek. All the catalogued books in America's Library of Congress total 15TB.
Petabyte (PB)	1,000TB; 2^{50} bytes	All letters delivered by America's postal service this year will amount to around 5PB. Google processes around 1PB every hour.
Exabyte (EB)	1,000PB; 2^{60} bytes	Equivalent to 10 billion copies of *The Economist*.
Zettabyte (ZB)	1,000EB; 2^{70} bytes	Approximately 1,000 Exabytes. Data in 2010 cracked the zettabyte barrier.
Yottabyte (YB)	1,000ZB; 2^{80} bytes	Currently too big to imagine.

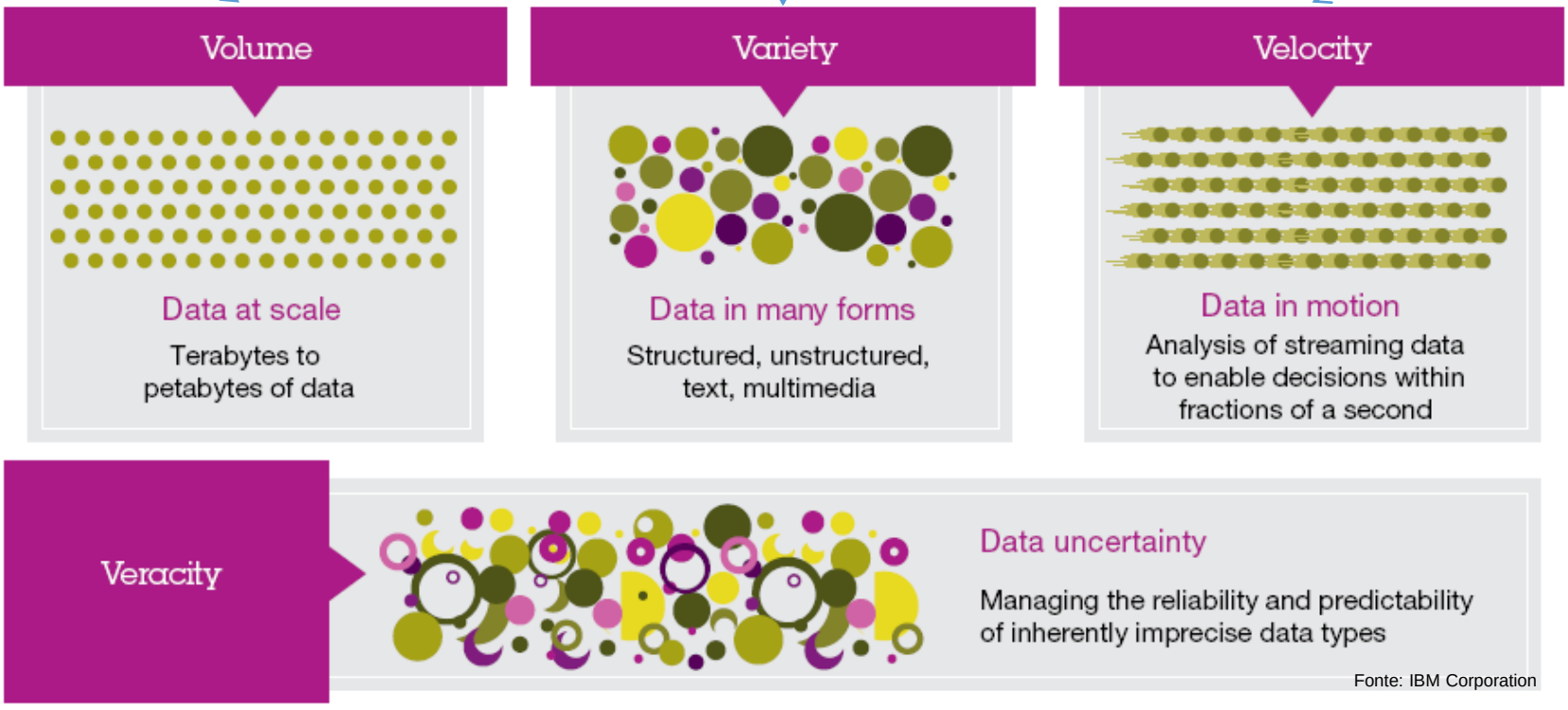
- Big Data se refere usualmente a datasets com tamanho além da capacidade de processamento por ferramentas usuais de software, em tempo hábil
- O "tamanho" do que é considerado Big Data é fugidio. Em 2012 eram poucos terabytes; hoje já são muitos petabytes.
- Big Data é também um conjunto de técnicas e tecnologias que requerem novas abordagens e métodos para se buscar valor em conjuntos de dados diversificados, complexos e em escala massiva.

Tamanho importa. Mas não é só tamanho...

- Como transformar 12TB de *tweets* diários em um produto de análise de sentimento e marketing digital?
- Como converter dezenas de gigabytes de notícias em insight sobre oportunidade de negócios?

- Dados estruturados e não estruturados;
- Bancos de dados, texto, áudio, vídeo, logs de servidores, streams de redes sociais, câmeras de vigilância, shapefiles, dados de sensores;
- 70% do tempo de análise gasto na limpeza de dados.

- Milhares de operações no Bovespa a cada segundo;
- Milhares de acessos ao site do ENEM por minuto;
- Centenas de tentativas de ataque aos servidores FGV por minuto durante eventos como concursos.



Fonte: IBM Corporation

Quem é "Big"?

Google

Estimativas obtidas através da energia consumida indicam que a Google possua 2.4 M servidores e 15 exabytes de dados (2013)

facebook

500 TB por dia (2013)

Chevron

Walmart

WayBack Machine

ebay

Dropbox

Microsoft

E no Brasil?

Metodista

P3NS4 SALA DE IDEIAS
PREFEITURA DO RIO DE JANEIRO - CASA CIVIL

Minas em números

buscapé

Votorantim Cimentos

MINISTÉRIO DA JUSTIÇA
SECRETARIA NACIONAL DE JUSTIÇA

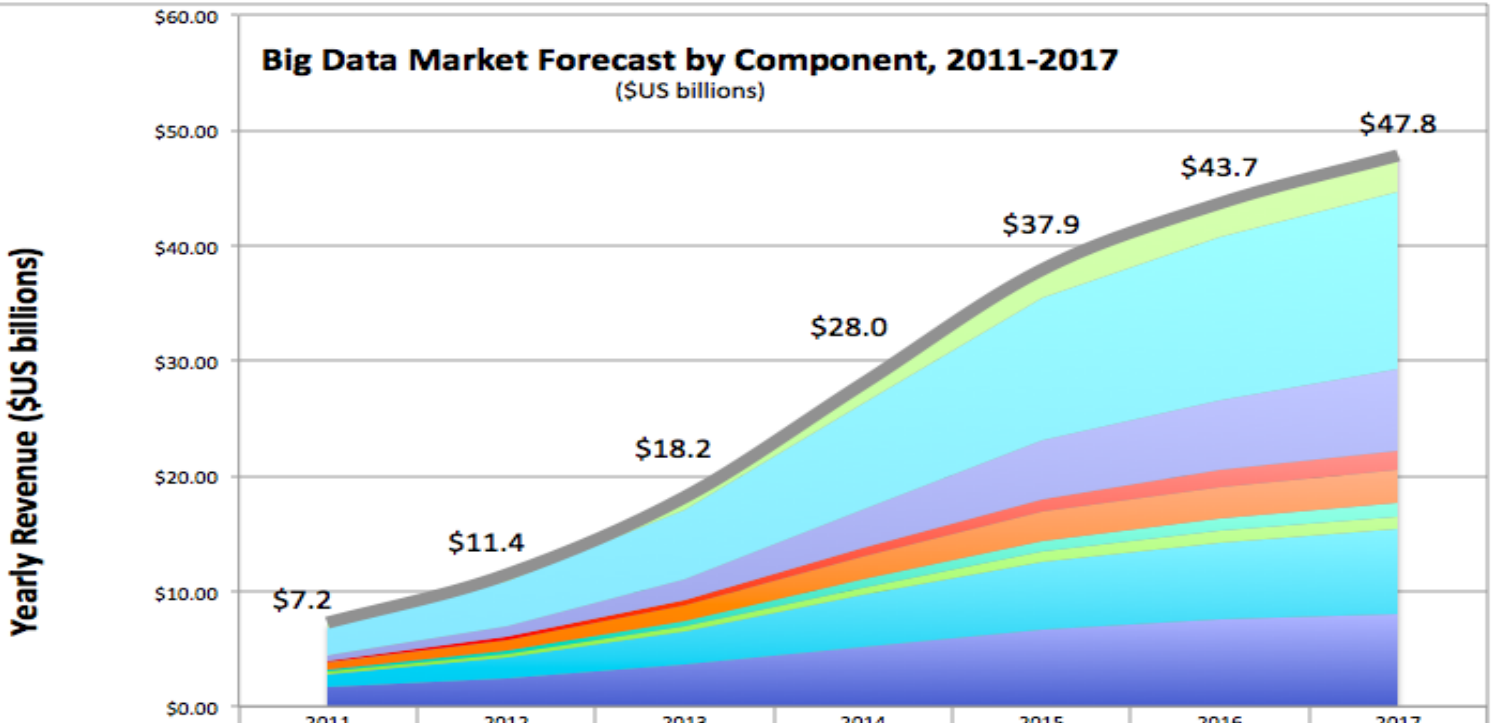
SUS Sistema Único de Saúde

Quanto vale este mercado?

TECH 1/12/2014 @ 7:25PM | 51.636 views

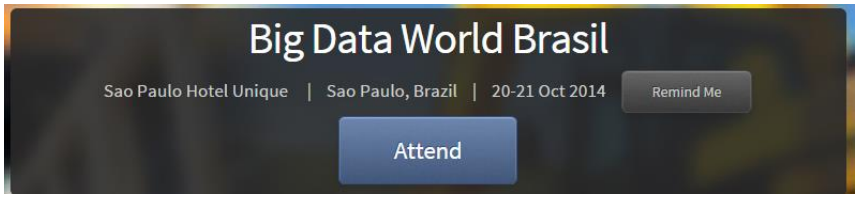
2014: The Year Big Data Adoption Goes Mainstream In The Enterprise - Forbes

Wikibon



	2011	2012	2013	2014	2015	2016	2017
Big Data XaaS Revenue	$0.35	$0.61	$1.05	$1.74	$2.47	$2.91	$3.24
Big Data Professional Services Revenue	$2.45	$3.87	$6.10	$9.29	$12.37	$14.14	$15.38
Big Data Application (Analytic and Transactional) Software	$0.49	$0.94	$1.80	$3.29	$5.02	$6.15	$7.00
Big Data NoSQL Database Software	$0.10	$0.19	$0.39	$0.73	$1.14	$1.41	$1.62
Big Data SQL Database Software	$0.72	$1.02	$1.45	$1.99	$2.47	$2.73	$2.90
Big Data Infrastructure Software	$0.16	$0.26	$0.43	$0.70	$0.96	$1.12	$1.24
Big Data Networking Revenue	$0.18	$0.28	$0.44	$0.67	$0.89	$1.02	$1.11
Big Data Storage Revenue	$1.16	$1.83	$2.89	$4.40	$5.86	$6.70	$7.28
Big Data Compute Revenue	$1.64	$2.45	$3.64	$5.23	$6.70	$7.50	$8.06
Total Big Data Revenue	$7.2	$11.4	$18.2	$28.0	$37.9	$43.7	$47.8

Eventos (Brasil)



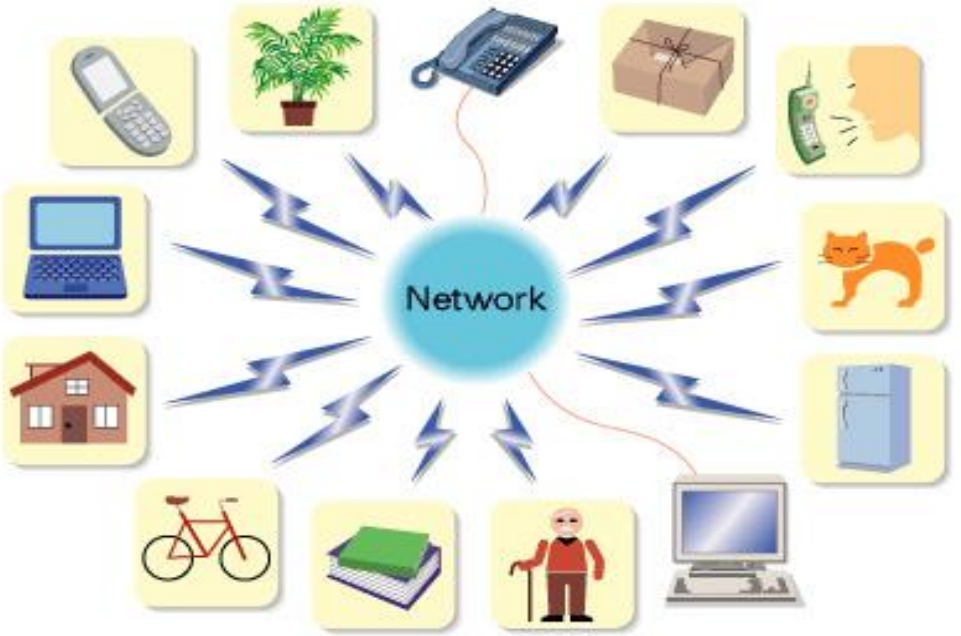
O que esperar do futuro?

"640K ought to be enough for anybody" (Bill Gates, 1981).

THE INTERNET OF THINGS

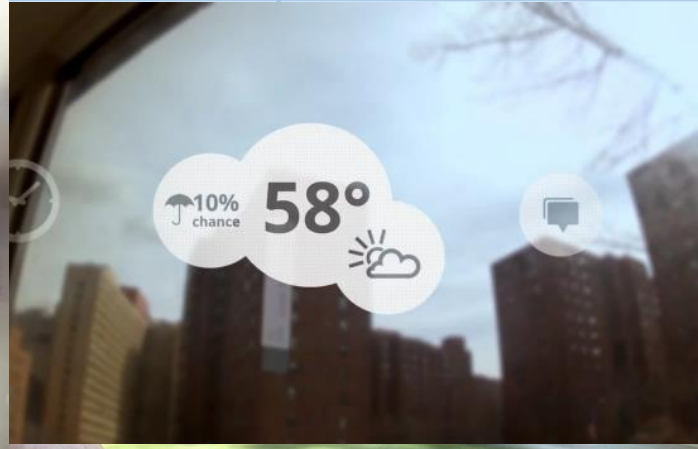
From RFID to the Next-Generation Pervasive Networked Systems

Auerbach Publications



Ubiquitous computing will enable diverse wireless applications, including monitoring of pets and houseplants, operation of appliances, keeping track of books and bicycles, and much more.


12:30 PM
Lunch
JFK
SFO
Virgin Airlines
Flight 23
10:31 AM
On Time
80°
60°
30%
tomorrow
New York
72°


10% chance
58°


How's the view up there?
just now


Google


The M90 Bus Service Will Arrive in 23 Min.
Are you Going the right Way?
You have been to Blue Fin Before. Whant to Check in?
Status: Work :-(
I <3 Lady GaGa!
This Lady Smells
Hey Bro, I see you're Close to Blue Fin, wanna eat?
You are Really Hot But your hat is Really Stupid
What is that?
You have 72 New Emails
RE: Your Lunch Date
FWD: Stupid Dancing Dog
Sponsorship Opportunities
RE: Google Glasses??
Jessica Wants to Chat with you
TARGETING SYSTEM:
OFFLINE
Citytv
DELICATESSEN
BLUE
ATM INSIDE
GRAND SLAM

Casos de uso de BIG Data em negócios

Big Data Exploration

Find, visualize, understand all big data to improve decision making

Security/Intelligence Extension

Lower risk, detect fraud and monitor cyber security in real-time

Enhanced 360º View of the Customer

Extend existing customer views (MDM, CRM, etc) by incorporating additional internal and external information sources

Operations Analysis

Analyze a variety of machine data for improved business results

Data Warehouse Augmentation

Integrate big data and data warehouse capabilities to increase operational efficiency

Fonte: IBM Corporation

Competências para Big Data

Math and Operations Research Expertise

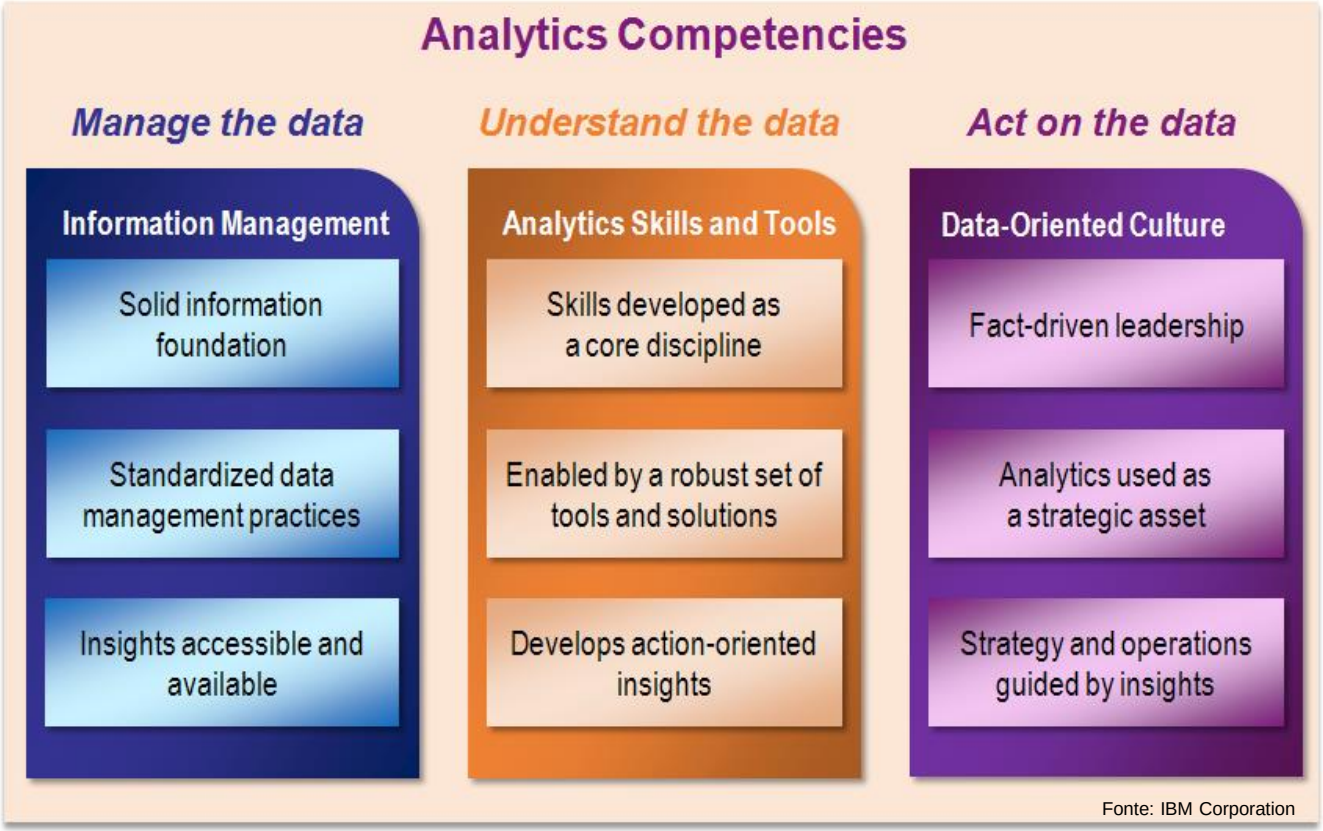
- Develop analytic algorithms

Visualization Expertise

- Interpret data sets
- determine correlations
- present in meaningful ways

Data Experts

- Data architecture
- Management
- Governance
- Policy



Fonte: IBM Corporation

Industry Vertical Domain Expertise

- Develop hypothesis,
- Identify relevant business issues,
- ask the right questions

Tool Developers

Mask complexity and analytics to lower skills boundaries

Decision Making Executive and Management

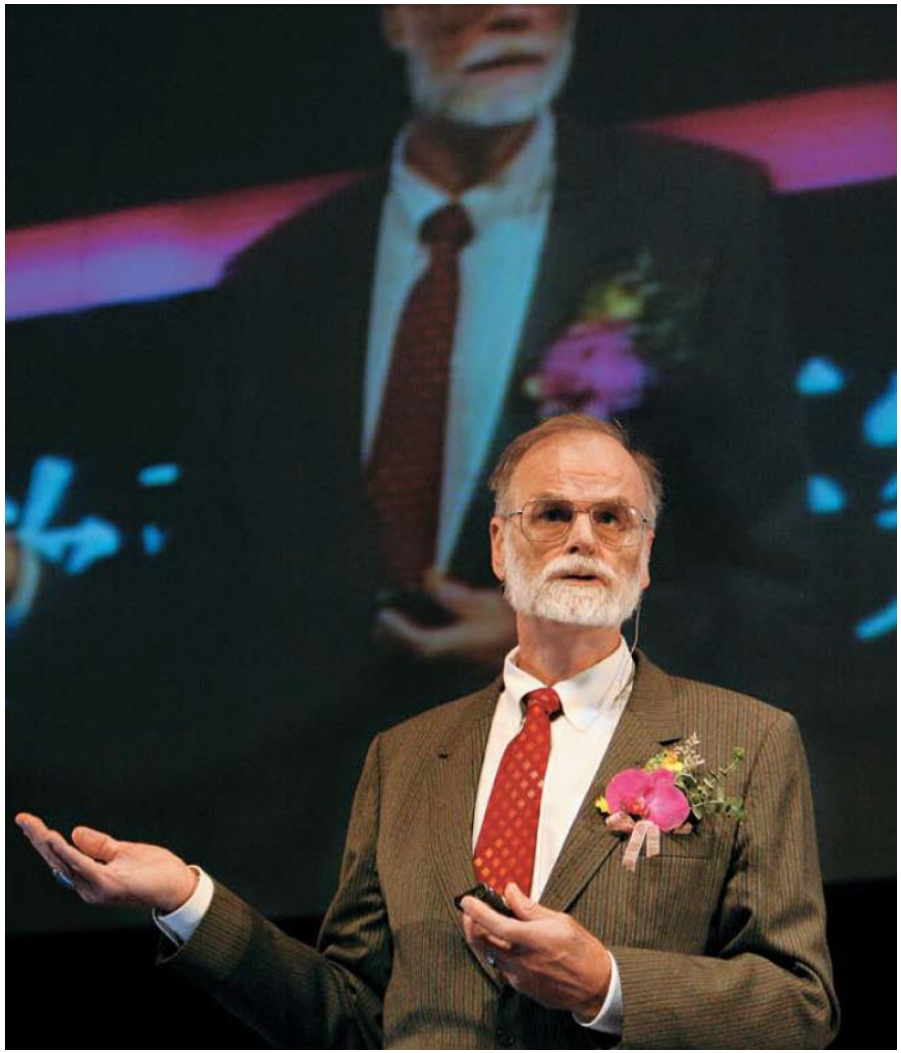
Apply information to solve business issues

"By 2015, big data demand will reach 4.4 million jobs globally, but only one-third of those jobs will be filled." *Source: "Gartner's Top Predictions for IT Organizations and Users, 2013 and Beyond: Balancing Economics, Risk, Opportunity and Innovation" 19 Oct 2012*

Data Science

'Fourth paradigm' –science led from Big Data

Data Science ou e-Science



Jim Gray: e-Science is where "IT meets scientists."

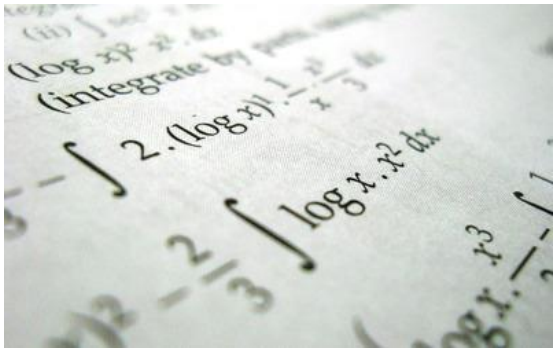
Jim Gray on eScience: A Transformed Scientific Method

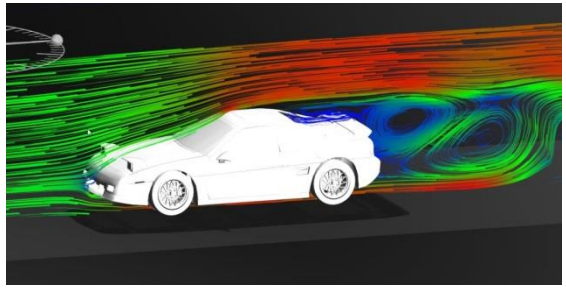
Based on the transcript of a talk given by Jim Gray to the NRC-CSTB[1] in Mountain View, CA, on January 11, 2007[2]

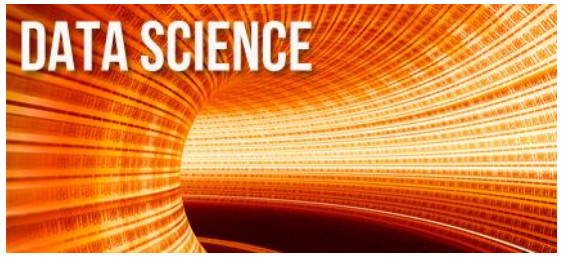
http://research.microsoft.com/en-us/collaboration/fourthparadigm

Paradigmas científicos

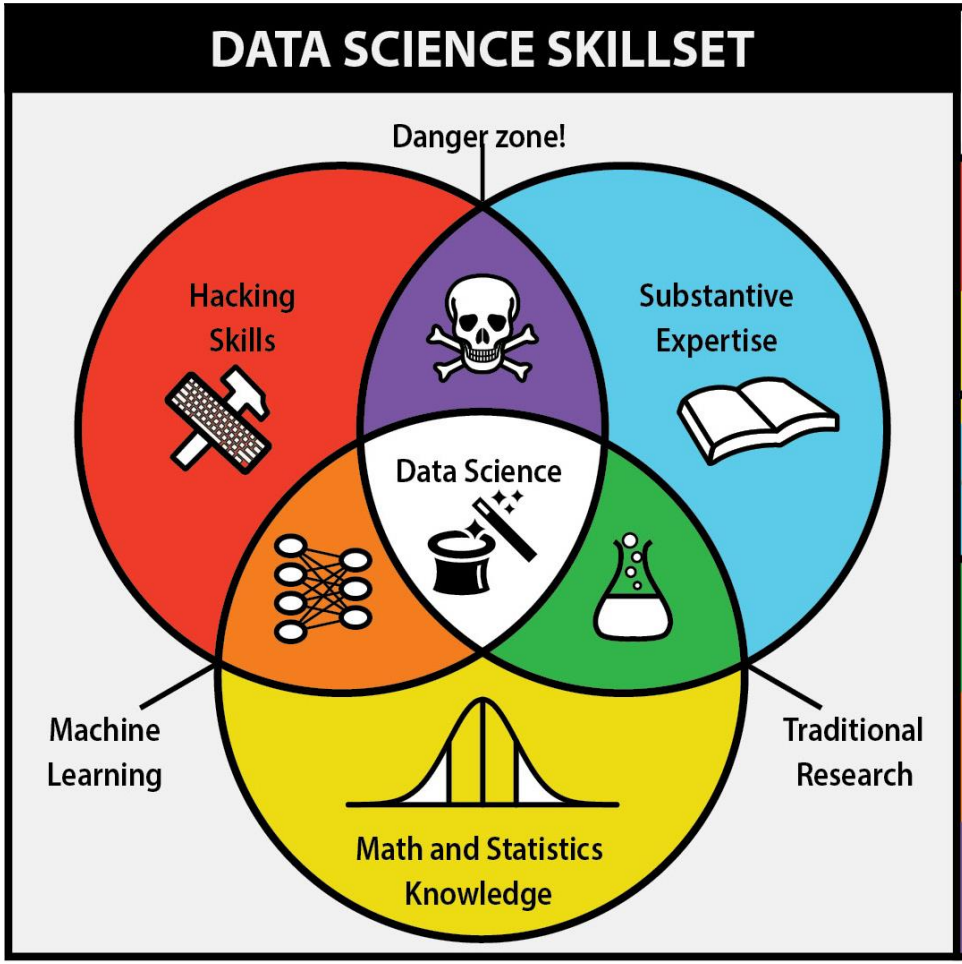
- Há mais de mil anos: a ciência era empírica
  - Baseada na descrição de fenômenos naturais
- Há algumas centenas de anos: surgem ramos teóricos
  - Uso de modelos e generalizações
- Há algumas décadas: surgem ramos computacionais
  - simulação de fenômenos complexos
- Hoje: exploração de dados (eScience)
  - Unificação de teoria, experimentos e simulação
  - Dados capturados por instrumentos ou gerados por simuladores
  - Processamento intensivo via software científico
  - Informação e Conhecimento armazenados no computador
  - Cientistas analisam bancos de dados e arquivos usando ferramentas analíticas e estatísticas







Data Science ou e-Science



Data science, due to its interdisciplinary nature, requires an intersection of abilities: **hacking skills**, **math and statistics knowledge**, and **substantive expertise** in a field of science.

Hacking skills are necessary for working with massive amounts of electronic data that must be acquired, cleaned, and manipulated.

Math and statistics knowledge allows a data scientist to choose appropriate methods and tools in order to extract insight from data.

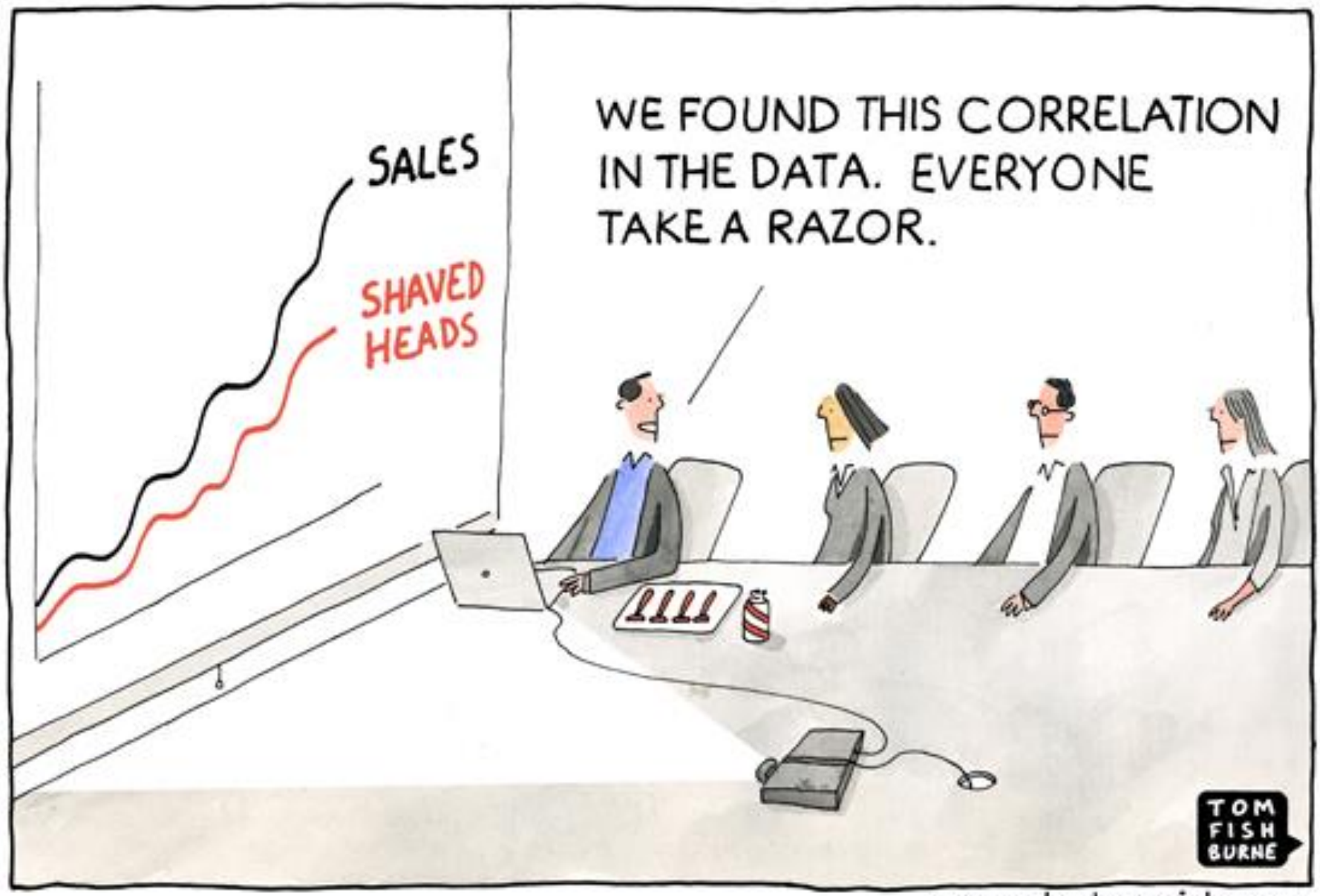
Substantive expertise in a scientific field is crucial for generating motivating questions and hypotheses and interpreting results.

Traditional research lies at the intersection of knowledge of math and statistics with substantive expertise in a scientific field.

Machine learning stems from combining hacking skills with math and statistics knowledge, but does not require scientific motivation.

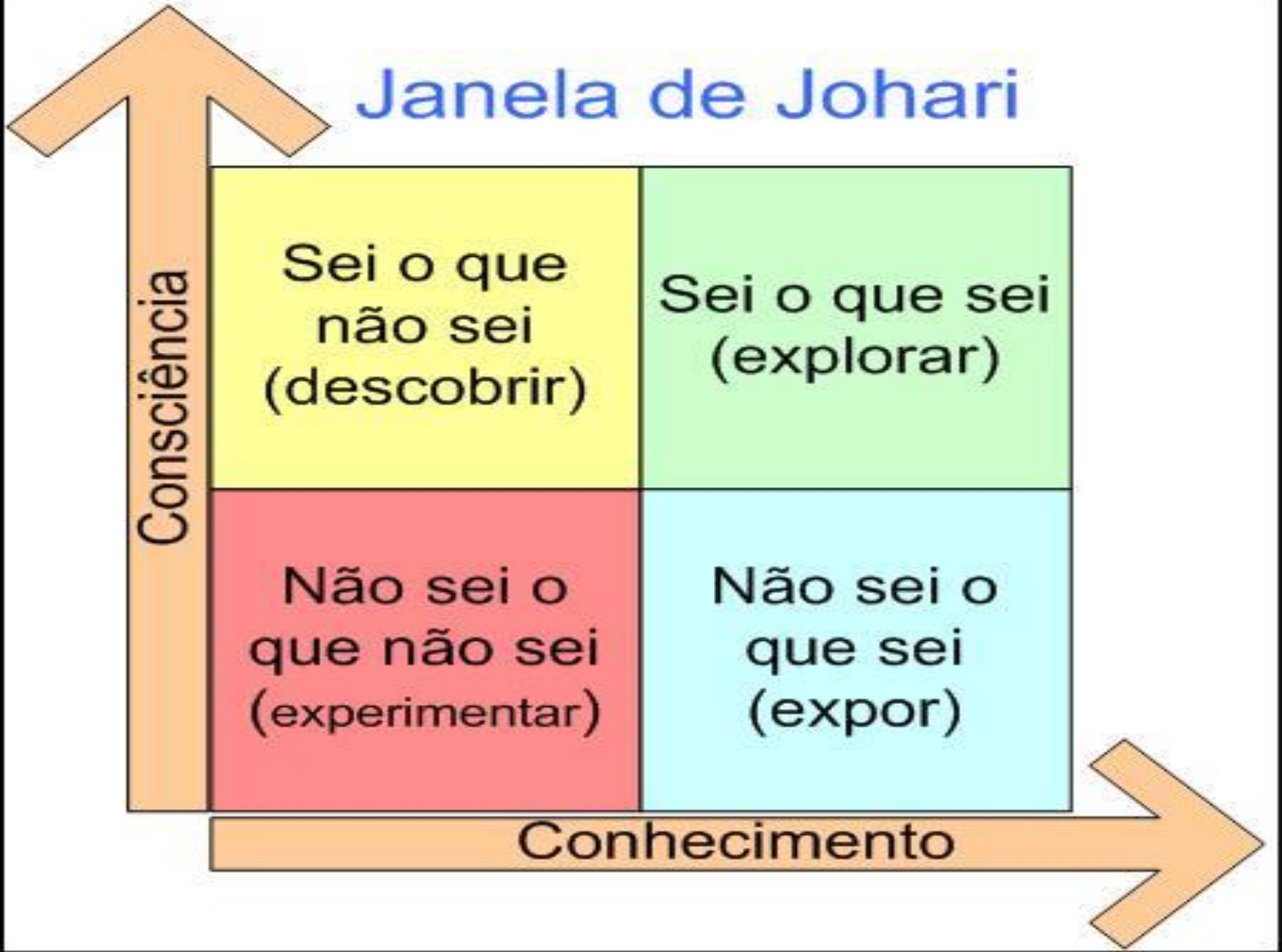
Danger zone! Hacking skills combined with substantive scientific expertise without rigorous methods can beget incorrect analyses.

Design: Natalia Bilenko, modified from Drew Conway: https://s3.amazonaws.com/aws.drewconway.com/viz/venn_diagram/data_science.html


SALES
SHAVED HEADS
WE FOUND THIS CORRELATION IN THE DATA. EVERYONE TAKE A RAZOR.
TOM FISH BURNE
©marketoonist.com

O que é Data Mining?

- Campo interdisciplinar e relativamente novo que tem como objetivo procurar por padrões ocultos em grandes massas de dados, combinando, para isso, técnicas de inteligência artificial, estatística, visualização, e gestão de bancos de dados;
- Embora seja um campo novo, algumas das técnicas de análise de dados são bastante antigas:
    - 1700s - Estatística bayesiana
    - 1800s - Análise de regressão
    - 1950s - Redes neurais, clustering, algoritmos genéticos
    - 1960s - Árvores de decisão
    - 1990s – Classificação baseada em support vector machines

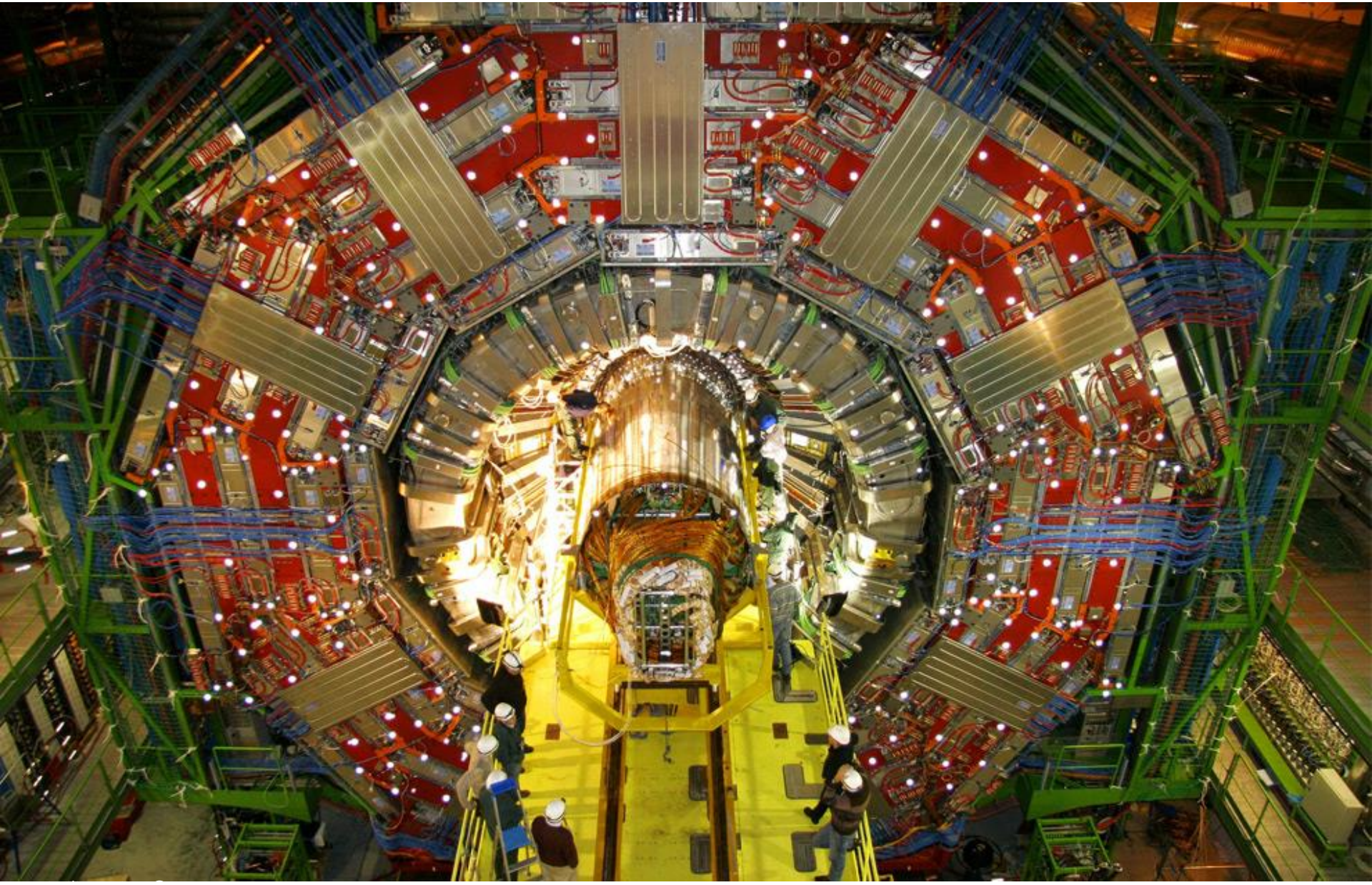

Janela de Johari
Consciência
Sei o que não sei (descobrir)
Sei o que sei (explorar)
Não sei o que não sei (experimentar)
Não sei o que sei (expor)
Conhecimento

Data Mining em Ambientes de Pesquisa

- Objetivo: melhoria contínua do negócio vs. pesquisa teórica e aplicada
  - **Busca por insights e desenvolvimento científico**
  - **"Data-driven research" – o fim do método científico tradicional?**

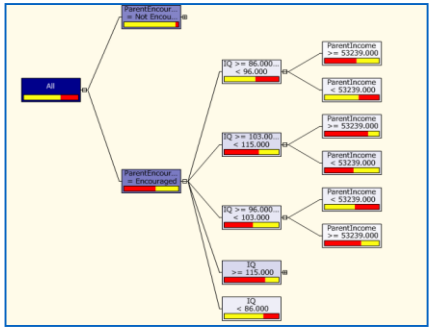
- Características dos dados:
  - **Ao invés de dados transacionais: imagens, sensores, simulações, observações**
  - **Dimensões espaciais e temporais, heterogeneidade, falta de estrutura**

- Tendência: As diferenças tecnológicas estão diminuindo!
  - **Bancos de dados, ferramentas de integração, serviços na web**
  - **Dados semi-estruturados ou não estruturados predominam em casos reais**
  - **Novos campos: text mining, web mining, image mining, etc.**

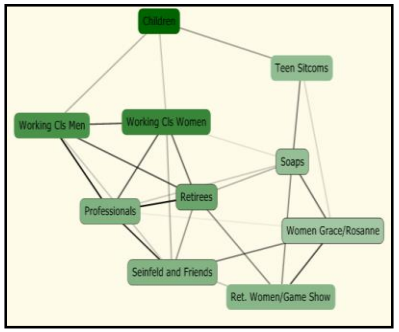


Data Mining em Ambientes de Pesquisa

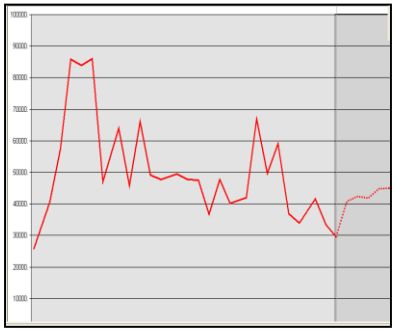
Algoritmos:



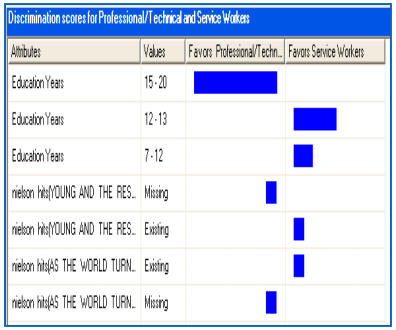
Árvores de Decisão



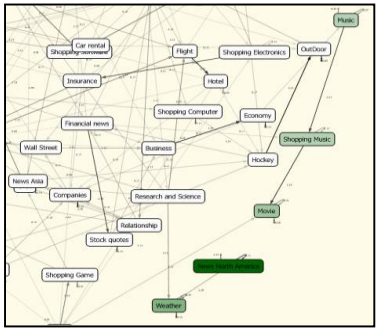
Clustering



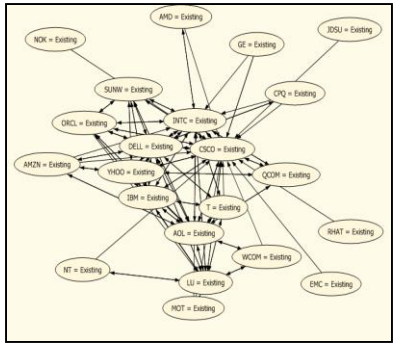
Séries Temporais



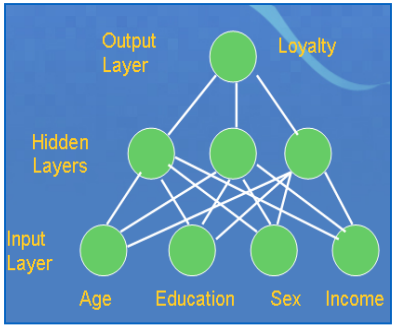
Naive Bayes



Sequence Clustering



Associação



Redes Neurais

Text Mining

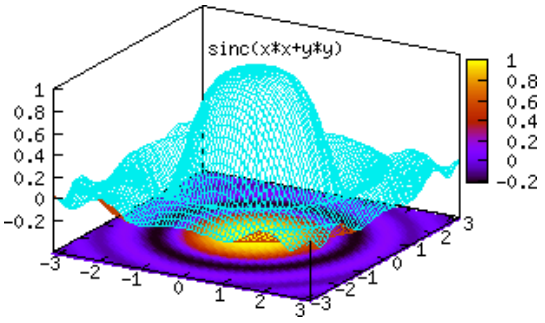
Ferramentas *open source* para Mineração de Dados

WEKA
The University of Waikato

scikit learn

Data Mining Tool orange

Gnuplot



sage

python

R

GATE
general architecture for text engineering

DATA SCIENCE WORKFLOW

1 Data acquisition and cleanup

Many **Python** libraries and specialized tools like **OpenRefine** and **Wrangler** aim to lower costs of data cleanup, which can claim up to 80% of development time.

2 Storage and management

Novel tools such as **NoSQL** and **MapReduce** are bolstered by growth of global data, expected to reach 40 zettabytes by 2020.

Analysis often involves revisiting raw data

3 Analysis

Data scientists who use open-source tools such as statistical packages in **R** and **Python** report higher salaries than those who use commercial software.

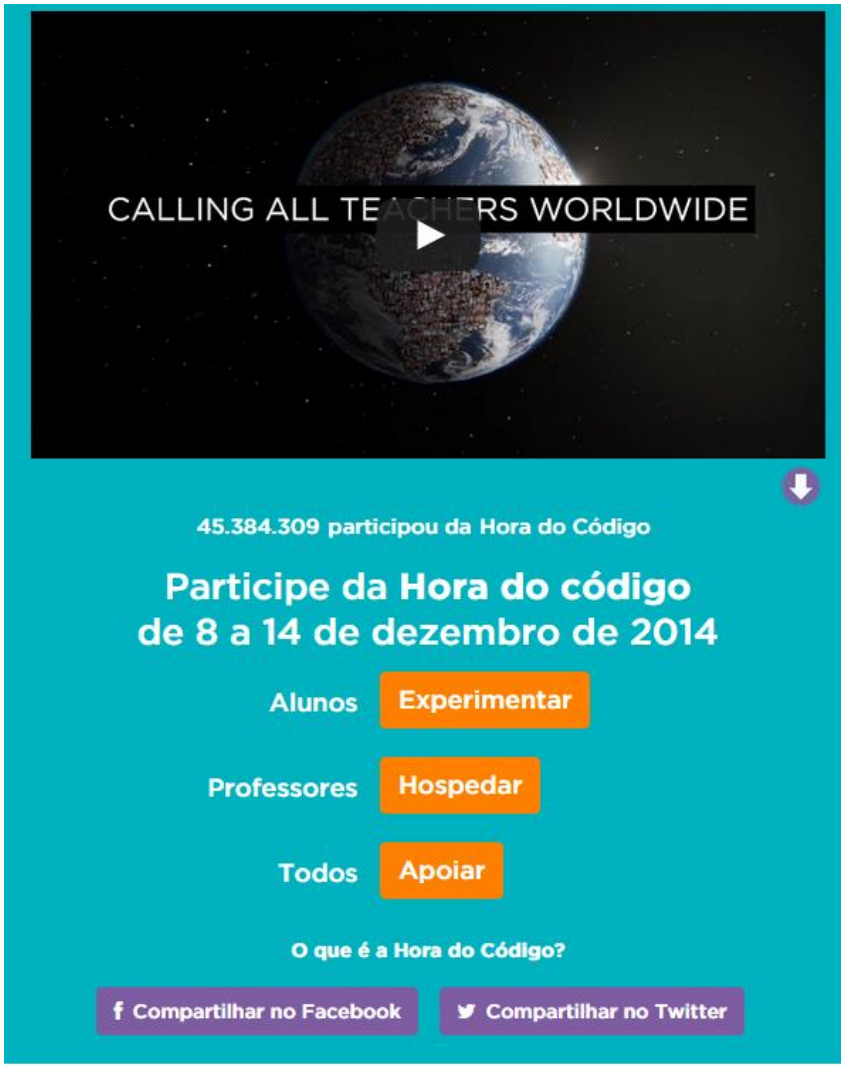
4 Visualization

Flexible visualization tools such as **D3.js** and **Processing** extract insight from data and easily integrate with existing frameworks.

5 Communication

Collaborative services such as **GitHub** and **Bitbucket** simplify sharing code and distributing results, which in turn increases reproducibility.

Codecademy



code

code school

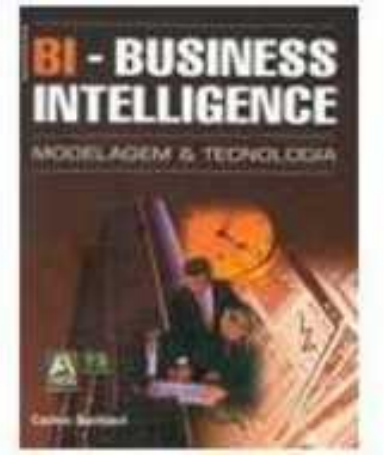
learn by doing

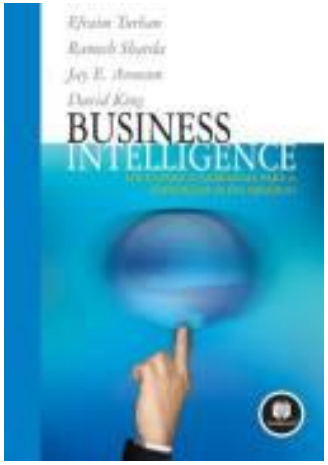
LEARN programming by visualizing code execution

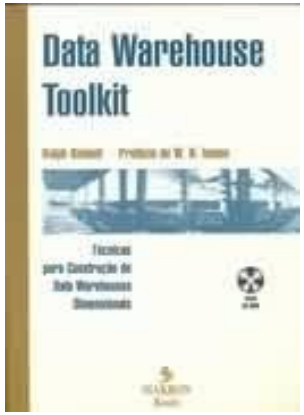
Online Python Tutor is a free educational tool created by Philip Guo that helps students overcome a fundamental barrier to learning programming: understanding what happens as the computer executes each line of a program's source code. Using this tool, a teacher or student can write a Python program in the Web browser and visualize what the computer is doing step-by-step as it executes the program.

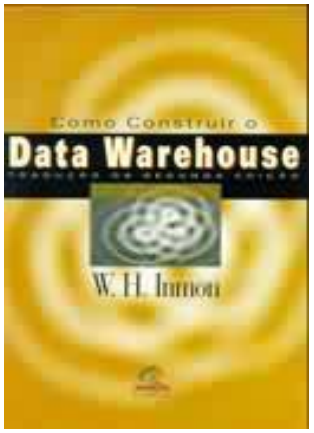
FGV EMAp ESCOLA DE MATEMÁTICA APLICADA

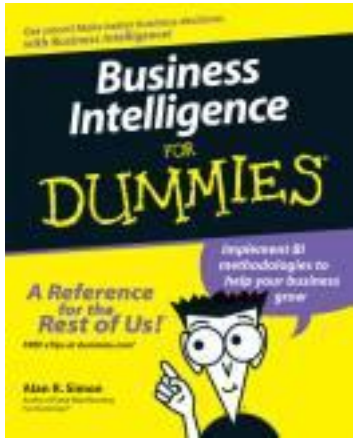
Data Mining em Ambiente Corporativo: Data Warehouse e Business Intelligence



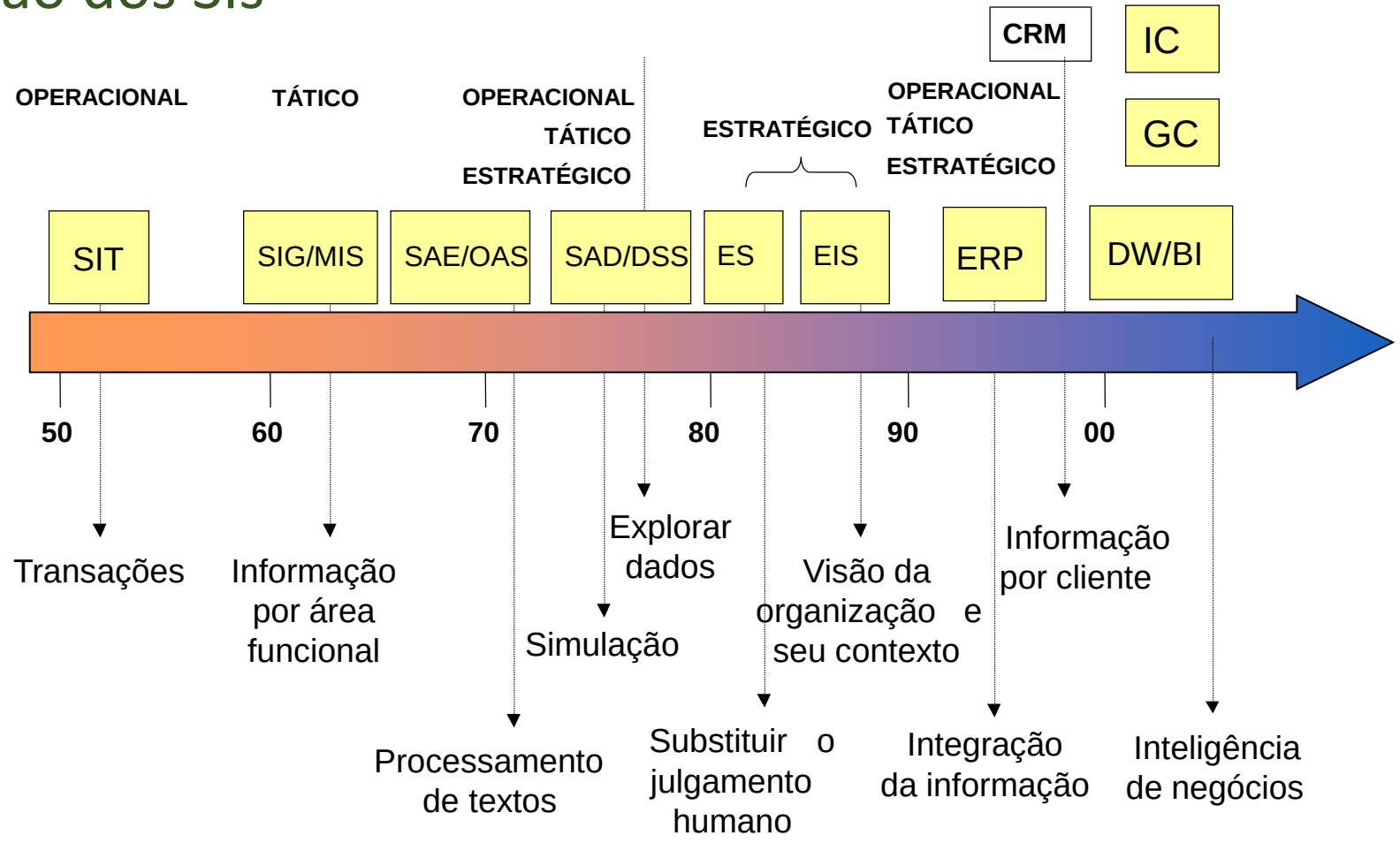








Evolução dos SIs



Data Mining em Ambiente Corporativo: Data Warehouse e Business Intelligence: projeto

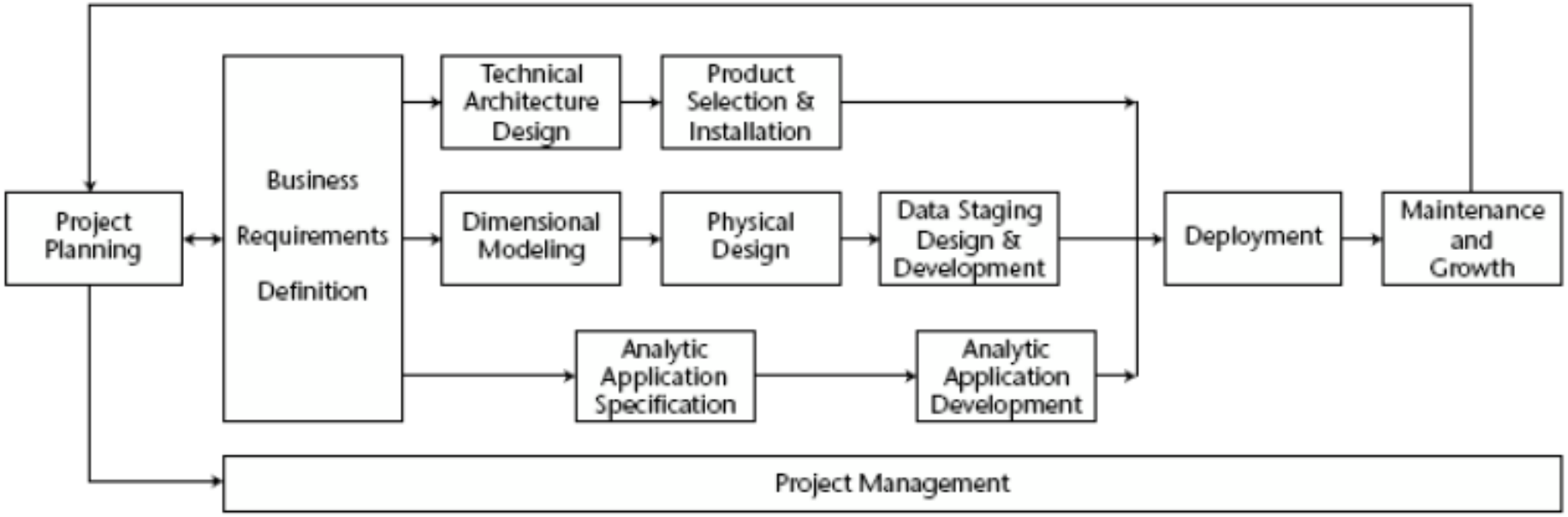
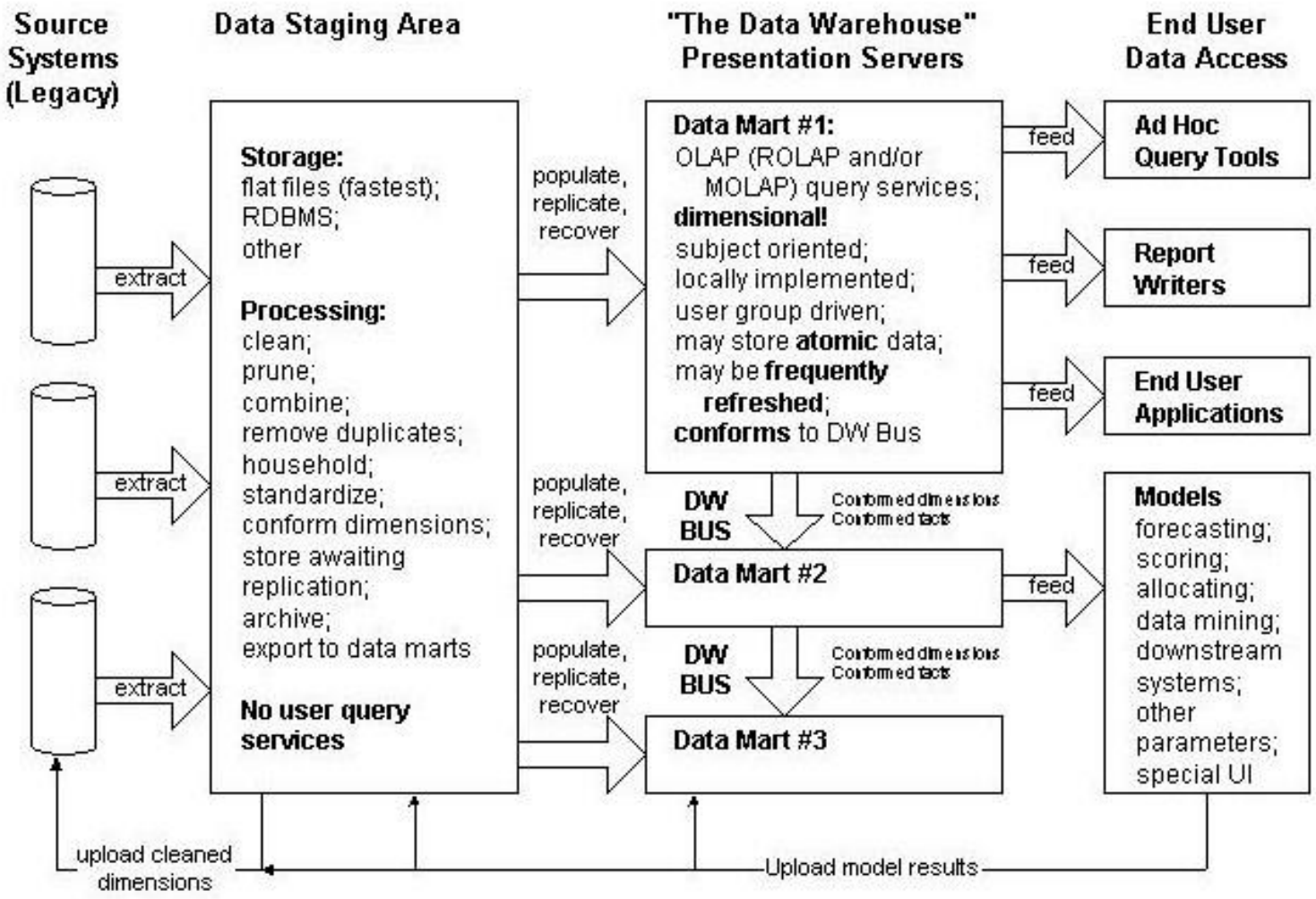


Figure 16.1 Business dimensional lifecycle diagram.

Building the Data Warehouse de Kimball, Ralph; Ross, Margy. The Data Warehouse Toolkit. John Wiley, 2002

FGV EMAp ESCOLA DE MATEMÁTICA APLICADA

Data Mining em Ambiente Corporativo: Data Warehouse e Business Intelligence: processo



Building the Data Warehouse de Kimball, Ralph; Ross, Margy. The Data Warehouse Toolkit. John Wiley, 2002

FGV EMAp ESCOLA DE MATEMÁTICA APLICADA

Data Mining em Ambiente Corporativo: Data Warehouse e Business Intelligence: dimensões

BUSINESS PROCESSES	COMMON DIMENSIONS: Date	Product	Store	Promotion	Warehouse	Vendor	Contract	Shipper
Retail Sales	X	X	X	X				
Retail Inventory	X	X	X					
Retail Deliveries	X	X	X					
Warehouse Inventory	X	X			X	X		
Warehouse Deliveries	X	X			X	X		
Purchase Orders	X	X			X	X	X	X

Building the Data Warehouse de Kimball, Ralph; Ross, Margy. The Data Warehouse Toolkit. John Wiley, 2002

Data Mining em Ambiente Corporativo: Data Warehouse e Business Intelligence: exemplos

Marketing
- Identificar o padrão de compra dos clientes
- Identificar associações entre clientes por características demográficas
- Prever resposta a campanhas de Marketing
- Análise de “carrinho de compras”

Mercado Financeiro
- Identificar padrões de uso fraudulento de cartões de crédito
- Identificar clientes leais
- Identificar perfil de inadimplência

Telecomunicações
- Identificar perfis de uso de celulares clonados
- Analisar perfis de utilização para sugerir planos mais vantajosos

Seguros
- Análise de solicitações
- Prever a aceitação de novos tipos de seguros por perfil de consumidor

Saúde
- Análise de pacientes visando ações preventivas
- Identificar o sucesso de terapias para diferentes doenças

Igrejas...

Data Driven Research

Data-Driven Design

THE DATA-DRIVEN ECONOMY

Data driven storytelling
The impact of visualisation on teaching and research

Data Driven Security

Data-Driven Marketing

FGV EMAp ESCOLA DE MATEMÁTICA APLICADA

Ética em Mineração de Dados

- Legislação x prática
- O “Big Brother” é aqui!
- Transações financeiras e novos comportamentos sociais