Universidade Federal do Ceará

Mestrado e Doutorado em Ciência da Computação

Análise em Big Data via Mineração de Dados

Ticiana L. Coelho da Silva

Campus Quixadá

ENUCOMP 2014

Encontro Unificado de Computação em Parnaíba

arida

Advanced Research in Database Group

Quem somos nós ?

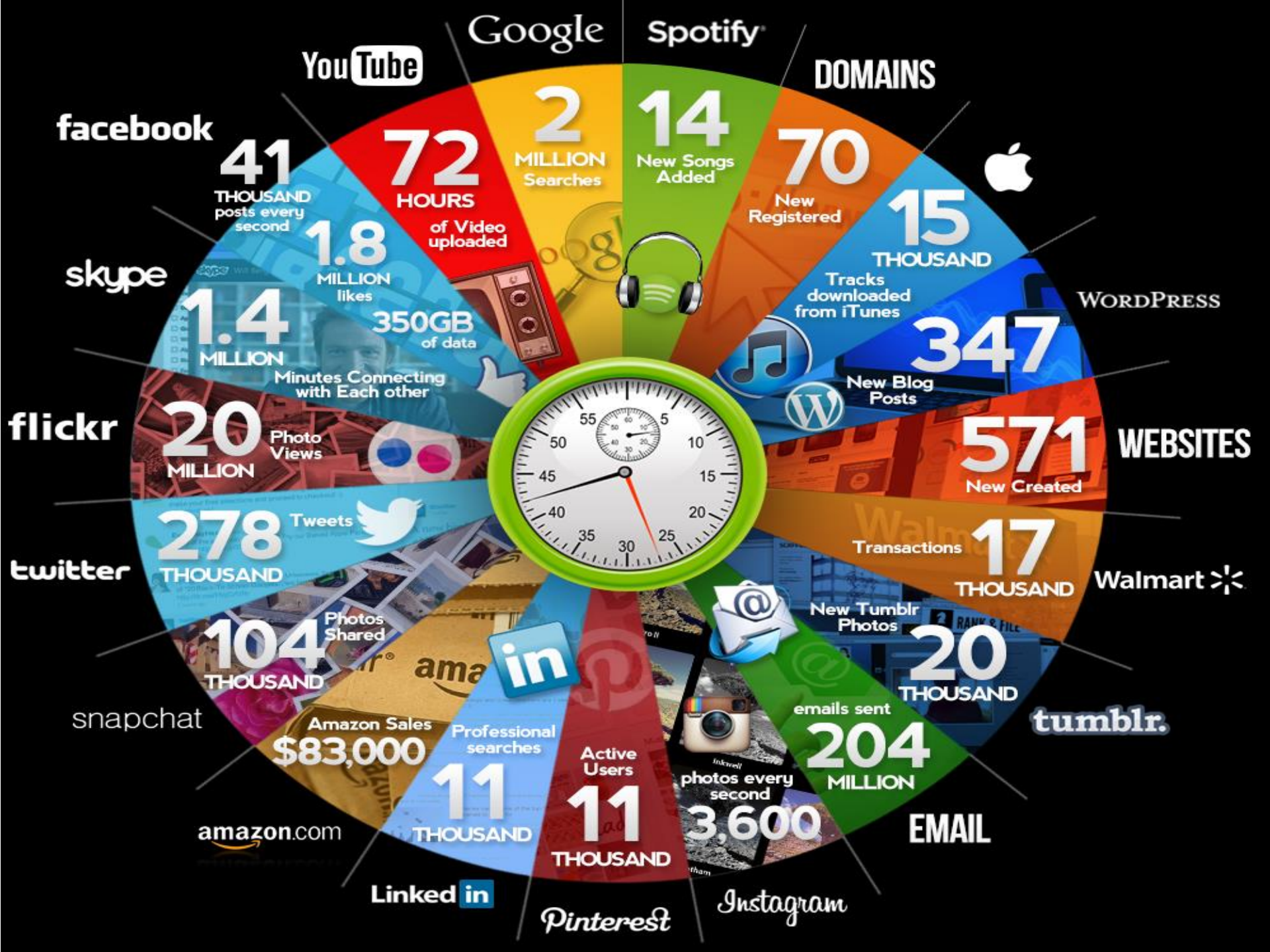
Advanced Research in Database Group


UNIVERSIDADE
FEDERAL DO CEARÁ
VIRTVS VNITA FORTIOR
Campus Quixadá

If you are looking for a career where your services will be in high demand, you should find something where you provide a scarce, complementary service to something that is getting ubiquitous and cheap. So what's getting ubiquitous and cheap? Data. And what is complementary to data? Analysis.

– Prof. Hal Varian, UC Berkeley, Chief Economist at Google

Introdução


Google
Spotify
YouTube
DOMAINS
facebook
2
MILLION
Searches
14
New Songs
Added
70
New
Registered
41
THOUSAND
posts every
second
72
HOURS
of Video
uploaded
15
THOUSAND
Tracks
downloaded
from iTunes
skype
1.8
MILLION
likes
WordPress
1.4
MILLION
350GB
of data
347
New Blog
Posts
Minutes Connecting
with Each other
flickr
20
MILLION
Photo
Views
571
New Created
WEBSITES
278
THOUSAND
Tweets
twitter
Transactions
17
THOUSAND
Walmart
New Tumblr
Photos
104
THOUSAND
Photos
Shared
20
THOUSAND
snapchat
tumblr.
emails sent
204
MILLION
Amazon Sales
$83,000
Professional
searches
11
THOUSAND
Active
Users
11
THOUSAND
photos every
second
3,600
amazon.com
EMAIL
LinkedIn
Pinterest
Instagram

Introdução

2,5 quintilhões de bytes de dados **por dia**

90% dos dados no mundo hoje foram produzidos nos **últimos dois anos**

2,7 bilhões no de usuários na internet

5 bilhões de celulares

64 Bilhões de mensagens em **24 horas**

Fonte: IBM/Whatsapp

Introdução

1990s

2010s

Os dados armazenados vão crescer 50 vezes mais até 2020

Fonte: KPCB/ SAS

Introdução

- **Facebook**
  - **1B** de usuários, **1,13 Trilhões** de "likes", **219B** de fotos e **140.3B** de relacionamentos
- **Youtube**
  - **100 horas** de vídeos adicionado a **cada minuto**
- **Bolsa de valores de Nova Iorque**
  - + **1 TB** de dados a cada sessão do pregão
- **Flick**
  - + de **5B de fotos**
- **Twitter**
  - **80 TB** e **1B** de tweets por dia

Introdução

- **Boeing**
  - **640 TB** gerados em um voo transatlântico
- **Wal-Mart**
  - **2,5 PB** e **1 milhão** de transações/hora

- **LHC CERN**
  - **15 Petabytes** por ano
- **Sloan Digital Sky Survey**
  - **14 milhões** de estrelas e galáxias
  - **80 atributos** por objeto
  - **10 Petabytes** gerados a cada varredura
- **Google**
  - **24 Petabytes** processados por dia

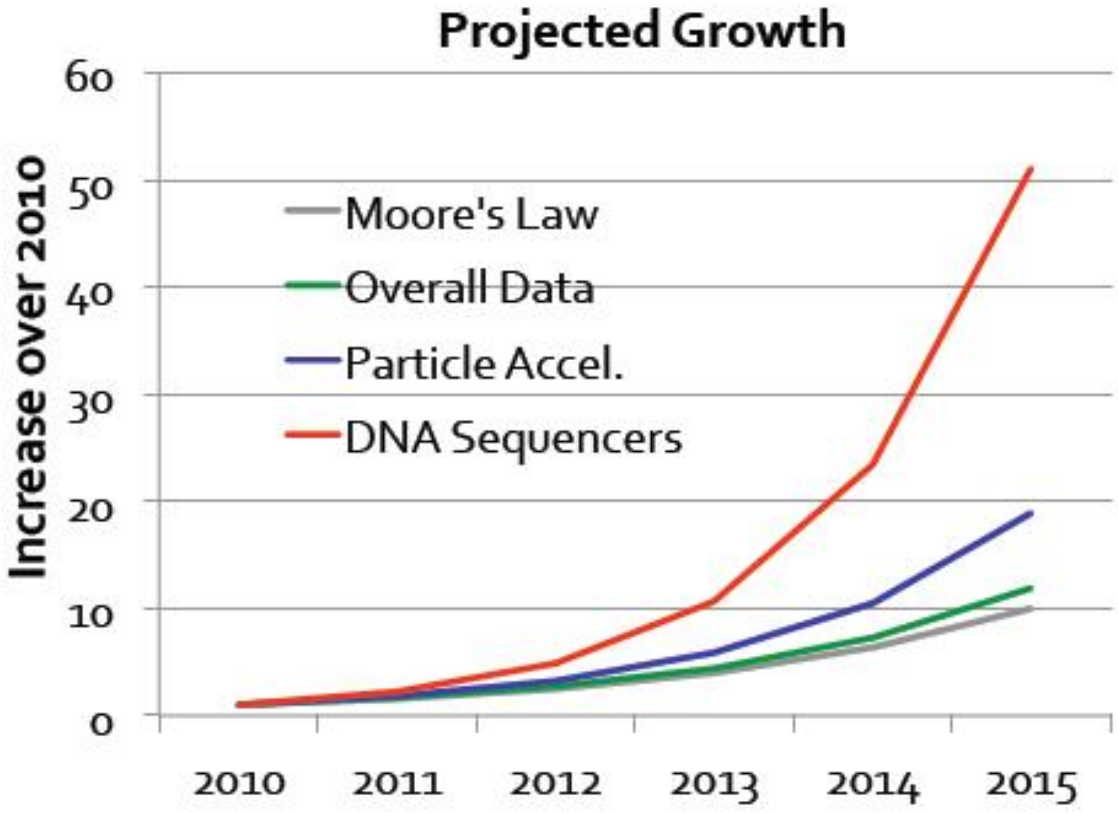
Introdução

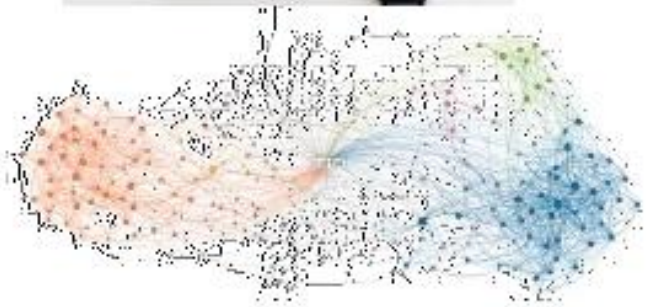
- 2000, 800 Terabytes
- 2006, 160 Exabytes
- 2009, 500 Exabytes(Internet)
- 2012, 2.7 Zettabytes
- 2020, 35 Zettabytes, 2020

Múltiplos do byte					
Prefixo binário (IEC)			Prefixo do SI		
Nome	Símbolo	Múltiplo	Nome	Símbolo	Múltiplo
byte	B	2^{0}	byte	B	10^{0}
kibibyte	KiB	2^{10}	Kilobyte	kB	10^{3}
mebibyte	MiB	2^{20}	megabyte	MB	10^{6}
gibibyte	GiB	2^{30}	gigabyte	GB	10^{9}
tebibyte	TiB	2^{40}	terabyte	TB	10^{12}
pebibyte	PiB	2^{50}	petabyte	PB	10^{15}
exbibyte	EiB	2^{60}	**exabyte**	EB	10^{18}
zebibyte	ZiB	2^{70}	zettabyte	ZB	10^{21}
yobibyte	YiB	2^{80}	yottabyte	YB	10^{24}

2.7 ZB = **85 Bilhões** x  **32 GB**

Os dados são "Grandes"





Data Grows faster than Moore's Law

[IDC report, Kathy Yelick, LBNL]

Fonte: Amplab UC Berkeley

Os dados são "Sujos"

- Diversas fontes de dados
- Sem esquema
- Sintaxe e semântica inconsistente

Dirty Data worse than Big Data

Fonte: Amplab UC Berkeley

Questões "Complexas"

- Perguntas difíceis
  - Qual é o impacto no trânsito e no preços das casas com construção de uma nova ponte?
- Perguntas em tempo real
  - Existe um ataque cibernético acontecendo?
- Perguntas em abertas
  - Quantos supernovas aconteceram no ano passado?

Big Data Must Enable Decisions

Fonte: Amplab UC Berkeley

Internet of Things (IoT)

Mobile Sensors

FACEBOOK GROWS BY 250 MILLION PHOTOS / DAY

Social Media

Video Surveillance

Video Rendering

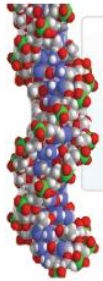
READING METERS EVERY 15 MINS. IS 3,000X MORE DATA INTENSIVE

Smart Grids

Geophysical Exploration

Medical Imaging

Gene Sequencing



COST TO SEQUENCE ONE GENOME HAS FALLEN FROM $100M IN 2001 TO $10K IN 2011

Fonte: EMC

Inovação

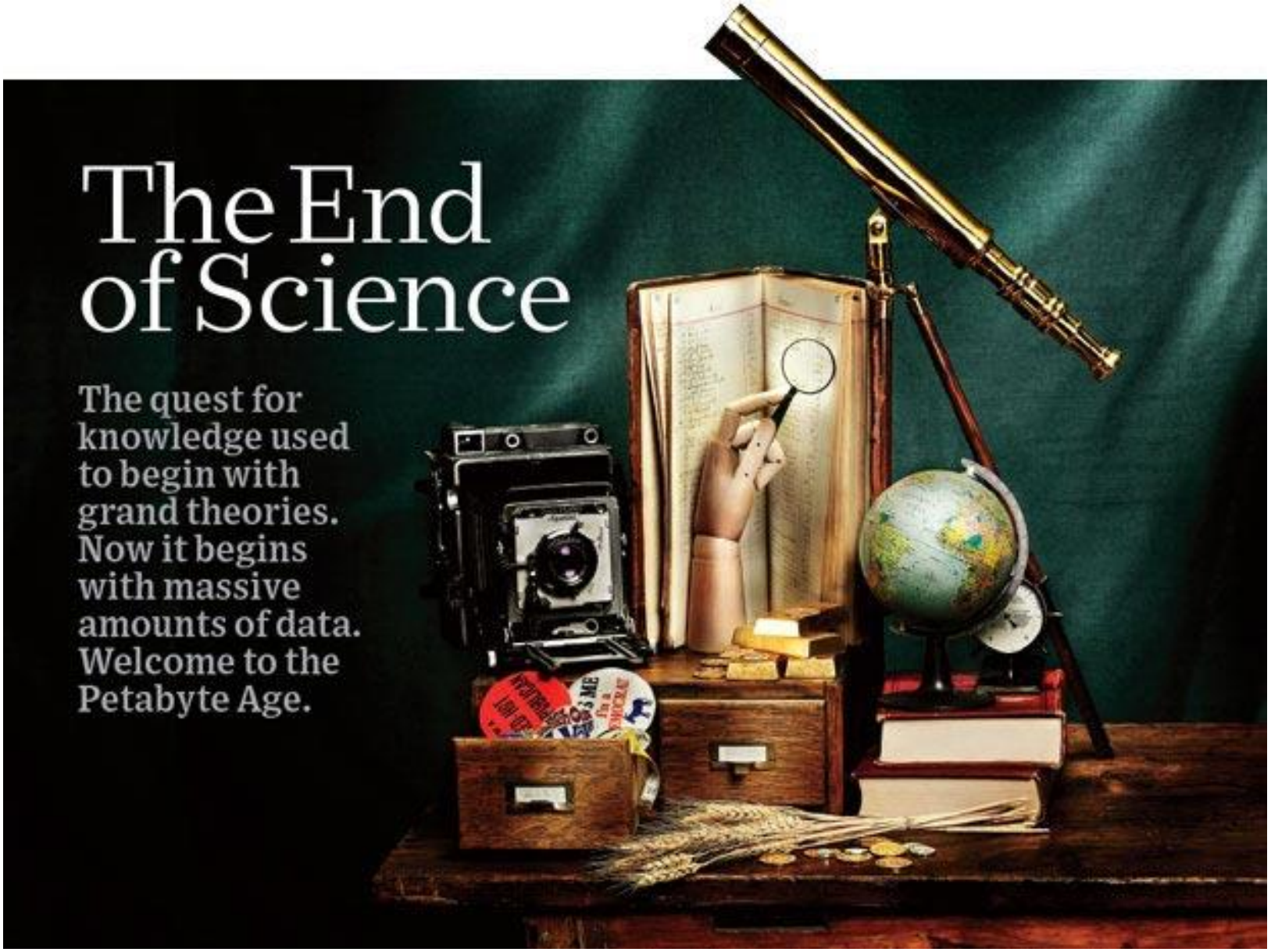
- Inovação é...
  - Criar algo "melhor" ou "mais eficaz"
  - Uma nova ideia ou método que é economicamente valioso
  - Fazer algo diferente

- Inovação melhora...
  - Experiência do cliente
  - Desenvolvimento do produto
  - Processo operacional

Como utilizar dados para a inovação?

Inovação orientada a dados

Teoria + Experimentação + Simulação ?



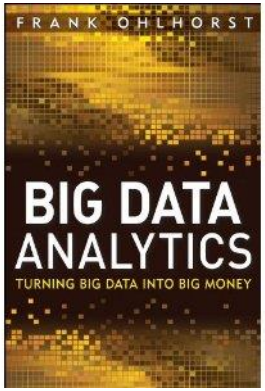

BIG
DATA


computing
in SCIENCE & ENGINEERING
p. 76 "Your Homework" Returns
BIG DATA

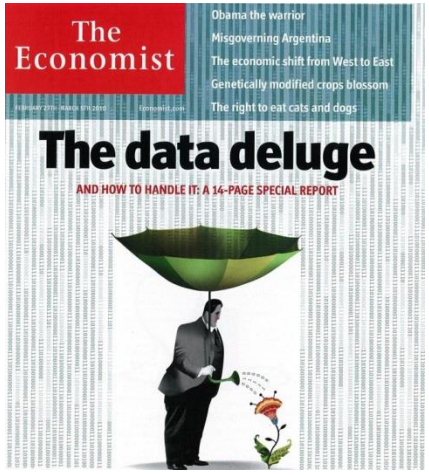

4 September 2008 | www.nature.com/nature | £10
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
nature
THE BITER BIT
Viral infections for viruses
TROPICAL CYCLONES
The strong get stronger
BLACK HOLE PHYSICS
A new window on the Galactic Centre
BIG DATA
NATUREJOBS
Minnesota musings
SCIENCE IN THE PETABYTE ERA


scientific
11 February 2011 | $10
Science
data
research
information
climate
analysis
visualization
researchers
access
knowledge


Harvard Business Review
GETTING CONTROL OF BIG DATA
How vast new streams of information are changing the art of management

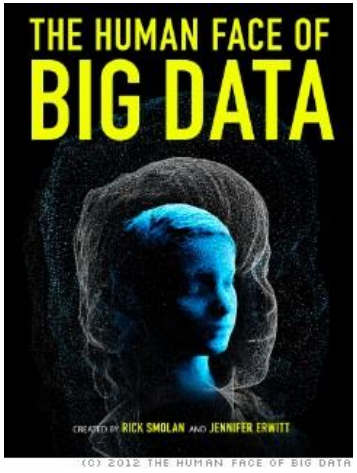

FRANK OHLHORST
BIG DATA ANALYTICS
TURNING BIG DATA INTO BIG MONEY

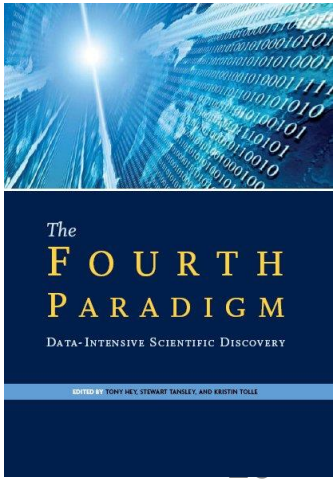

CIÊNCIA HOJE
LUIZ WERNECK VIANNA
Câncer de pele
PEDRO NAVA
DESAFIOS DO BIG DATA
Como escolher e guardar tanta informação?


The Economist
Obama the warrior
Misgoverning Argentina
The economic shift from West to East
Genetically modified crops blossom
The right to eat cats and dogs
The data deluge
AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT


POPULAR SCIENCE
THE FUTURE NOW
THE CONTROL CENTERS
OFFICER ALGORITHM
NEW WAYS OF SEEING
PLUS
SPECIAL ISSUE
DATA IS POWER
HOW INFORMATION IS DRIVING THE FUTURE


cnrs
international magazine
The BIG DATA REVOLUTION


THE HUMAN FACE OF BIG DATA
CREATED BY RICK SMOLAN AND JENNIFER ERWITT
(C) 2012 THE HUMAN FACE OF BIG DATA


The FOURTH PARADIGM
DATA-INTENSIVE SCIENTIFIC DISCOVERY
EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

Big Data

"Big Data é como sexo no colegial: "Ninguém faz, mas todo mundo diz que faz. Então todos pensam que alguém está fazendo e dizem que fazem também"

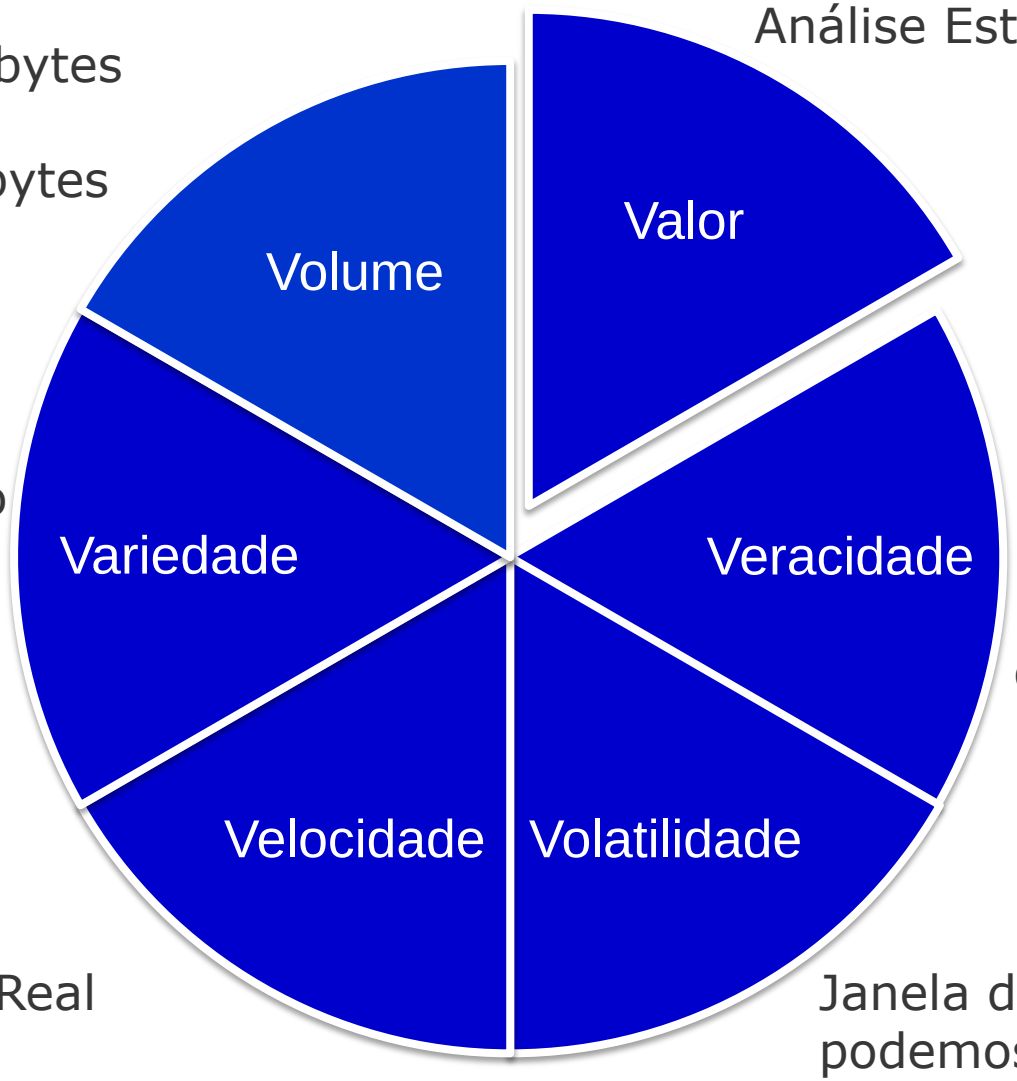
Fonte: Jay Kidd, CTO da NetApp

Big Data

- Big Data são dados que **excedem** o armazenamento, o processamento e a capacidade dos sistemas convencionais
  - Volume de dados muito grande
  - Dados são gerados rapidamente
  - Dados não se encaixam nas estruturas de arquiteturas de sistemas atuais
- Além disso, para obter **valor** a partir desses dados, é **preciso mudar a forma de analisá-los**

Fonte: Jordi Torres

6 V's do Big Data



Terabytes
...
Exabytes

Análise Estratégica de Dados

Não-estruturado
Semi-estruturado
Estruturado

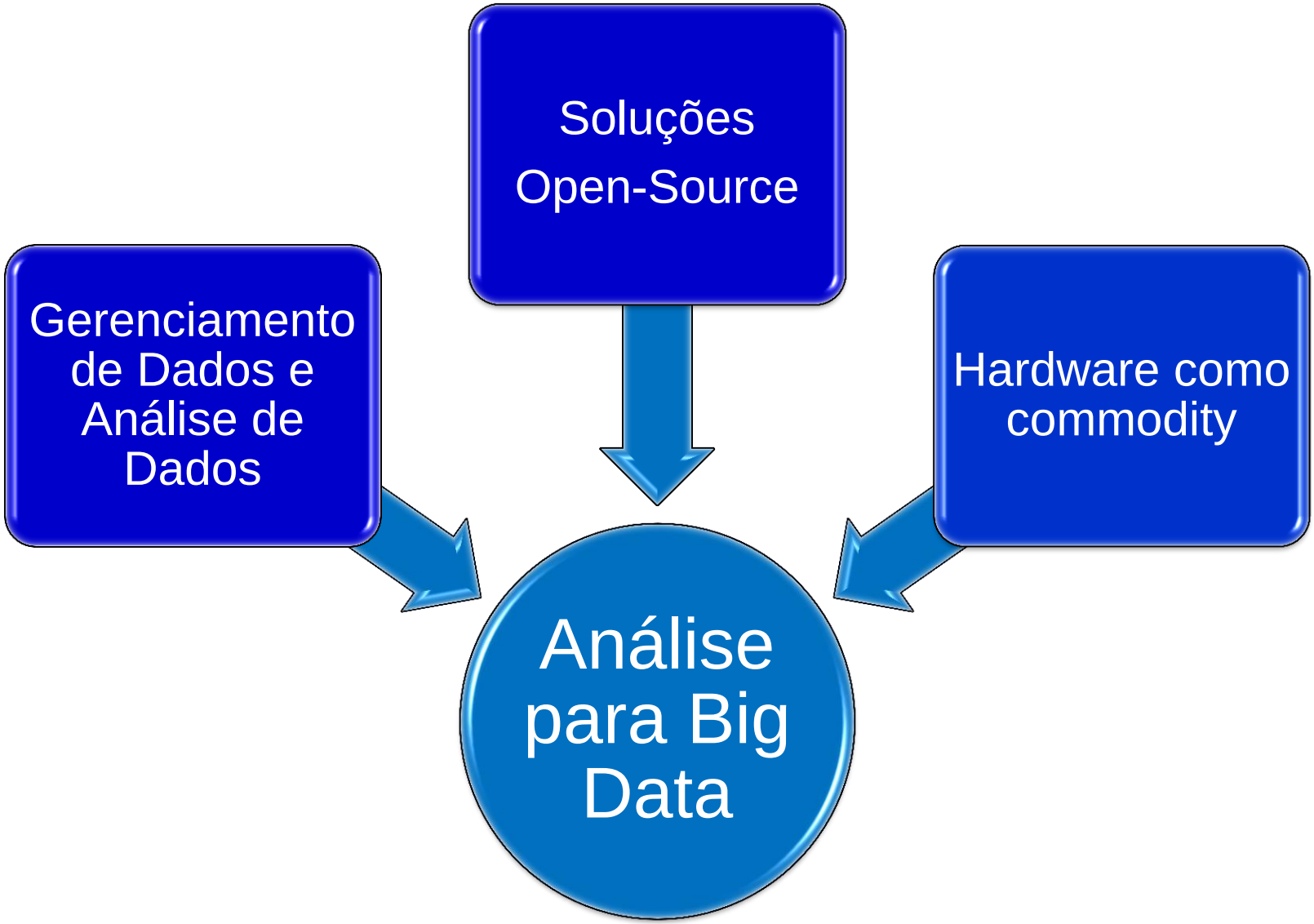
Estado verdadeiro
da realidade

Batch
Tempo Real

Janela de tempo onde
podemos usar os dados

Big Data: Aplicações

Convergência Hardware e Software




1%
of the world's data is analyzed today
Source: 2012 IDC Digital Universe Study

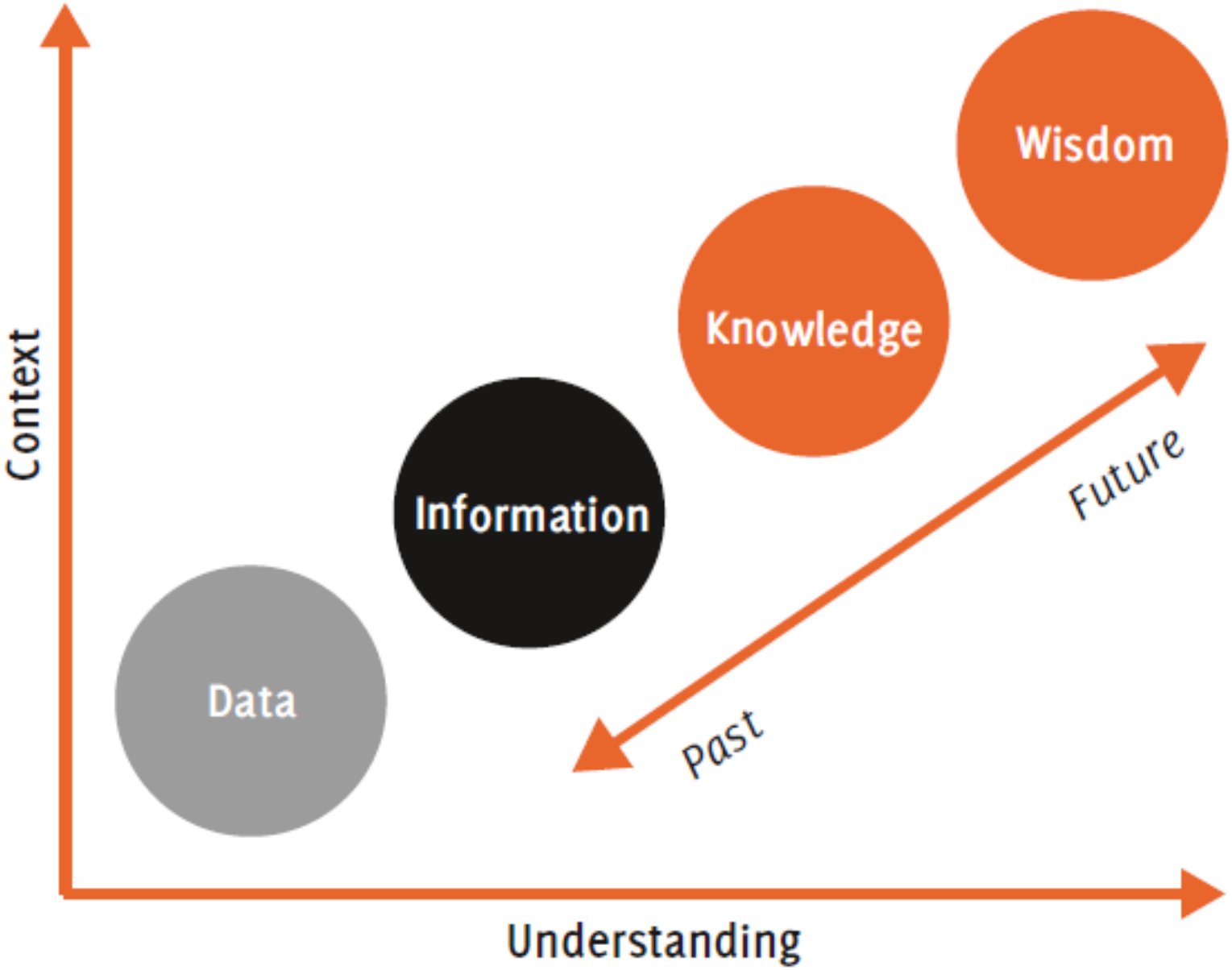
Análise para Big Data

"O desafio fundamental para as aplicações de Big Data é explorar os grandes volumes de dados e extrair informações úteis ou conhecimento para futuras ações"

Fonte: Rajaraman and Ullman 2012

Análise para Big Data

Fonte: VLDB 2010



Fonte: Sogeti

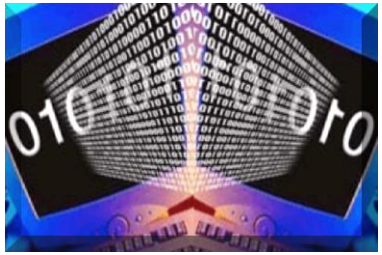
Análise para Big Data: Gera Valor

Smarter Healthcare

Multi-channel

Finance

Log Analysis



Homeland Security

Traffic Control

Telecom

Search Quality

Manufacturing

Trading

Fraud and Risk

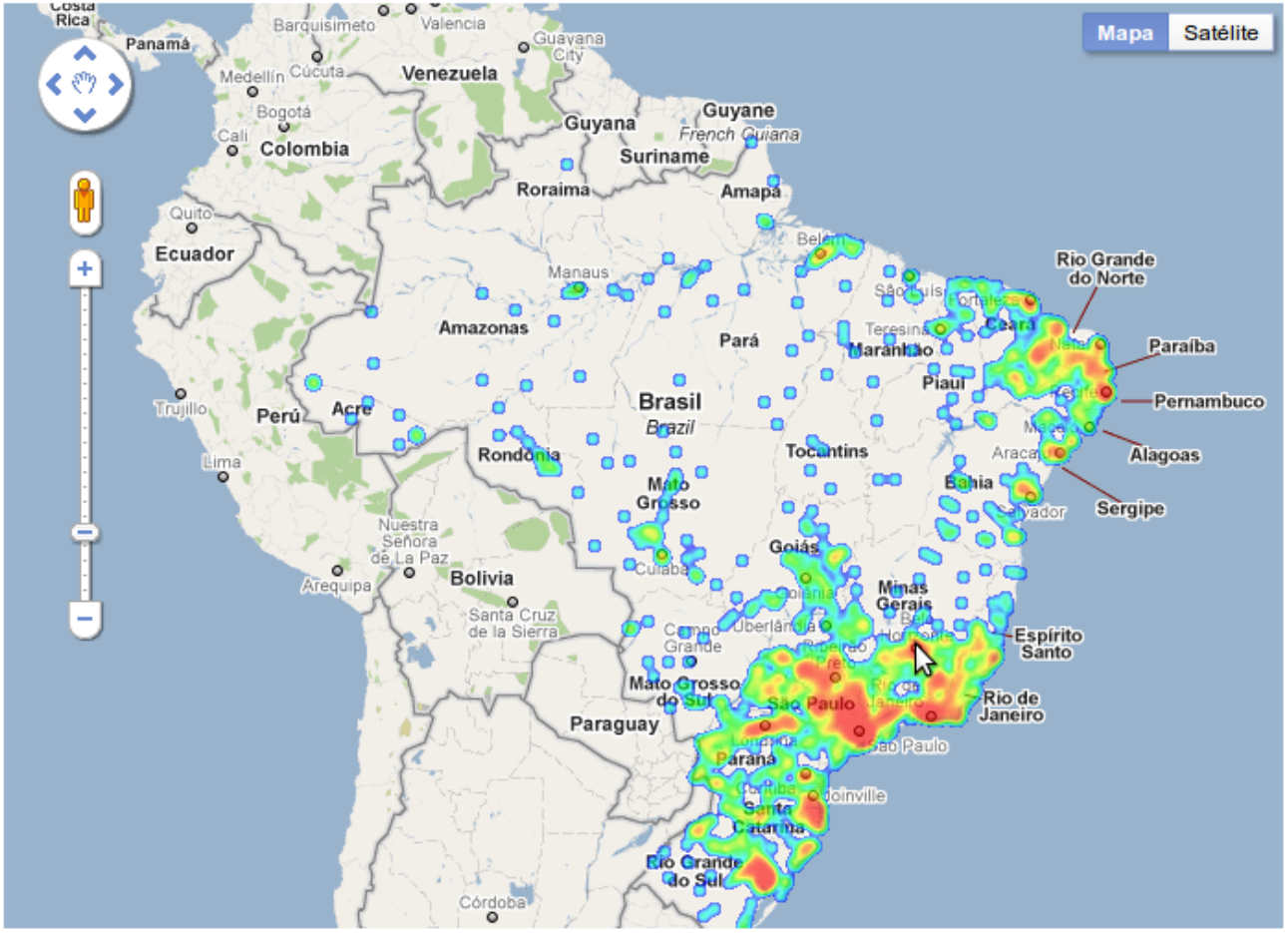
Retail: Churn, NBO

Fonte: Alberto Laender

Dengue Watch
Heatmap

observatóriodadengue

Menções à dengue no Twitter no mês de fev/2011

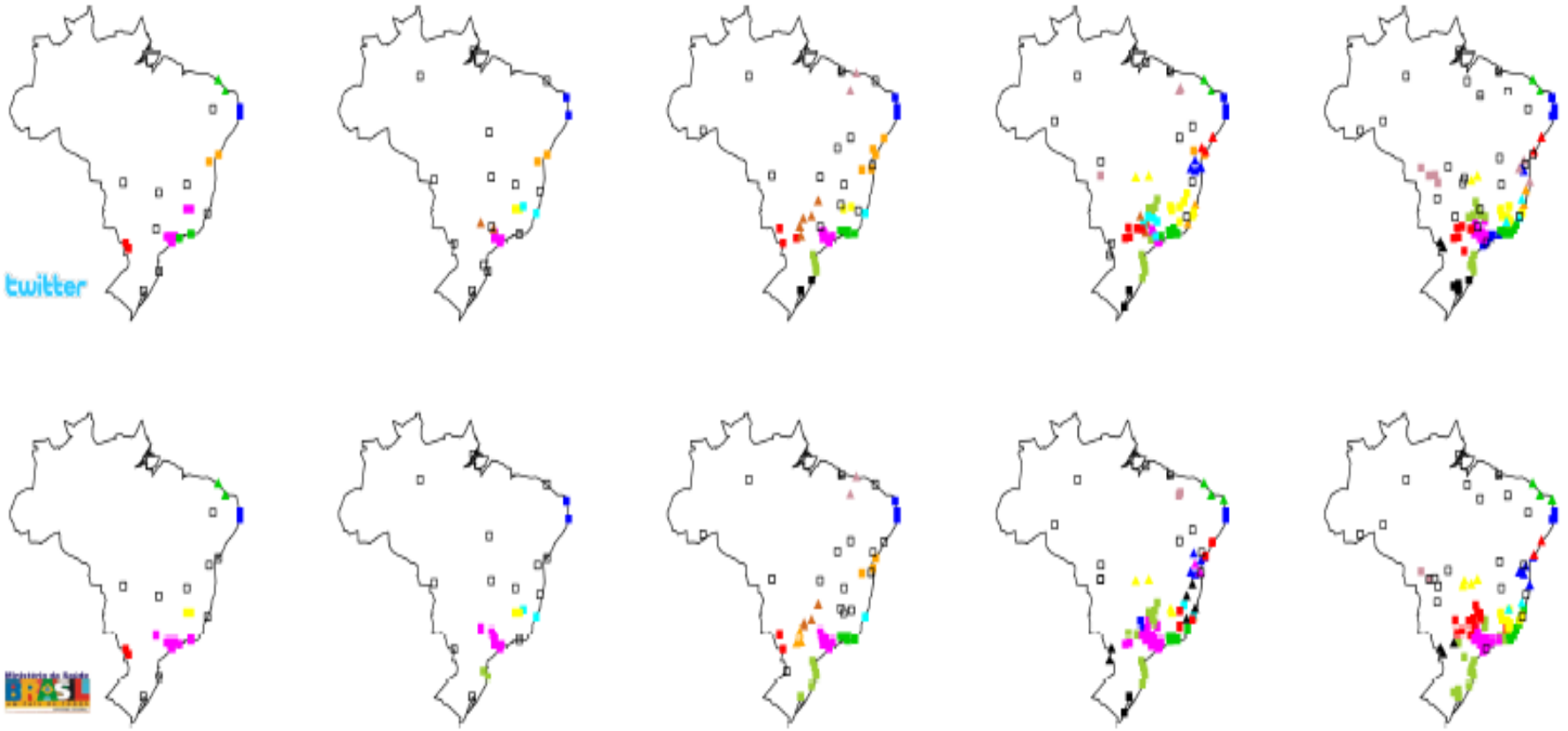


Clique nos pontos do mapa para informações

Cidades: 11 Tweets: 59 População: 1925450 Tx. Inc. Méd.: 1.5334e-04

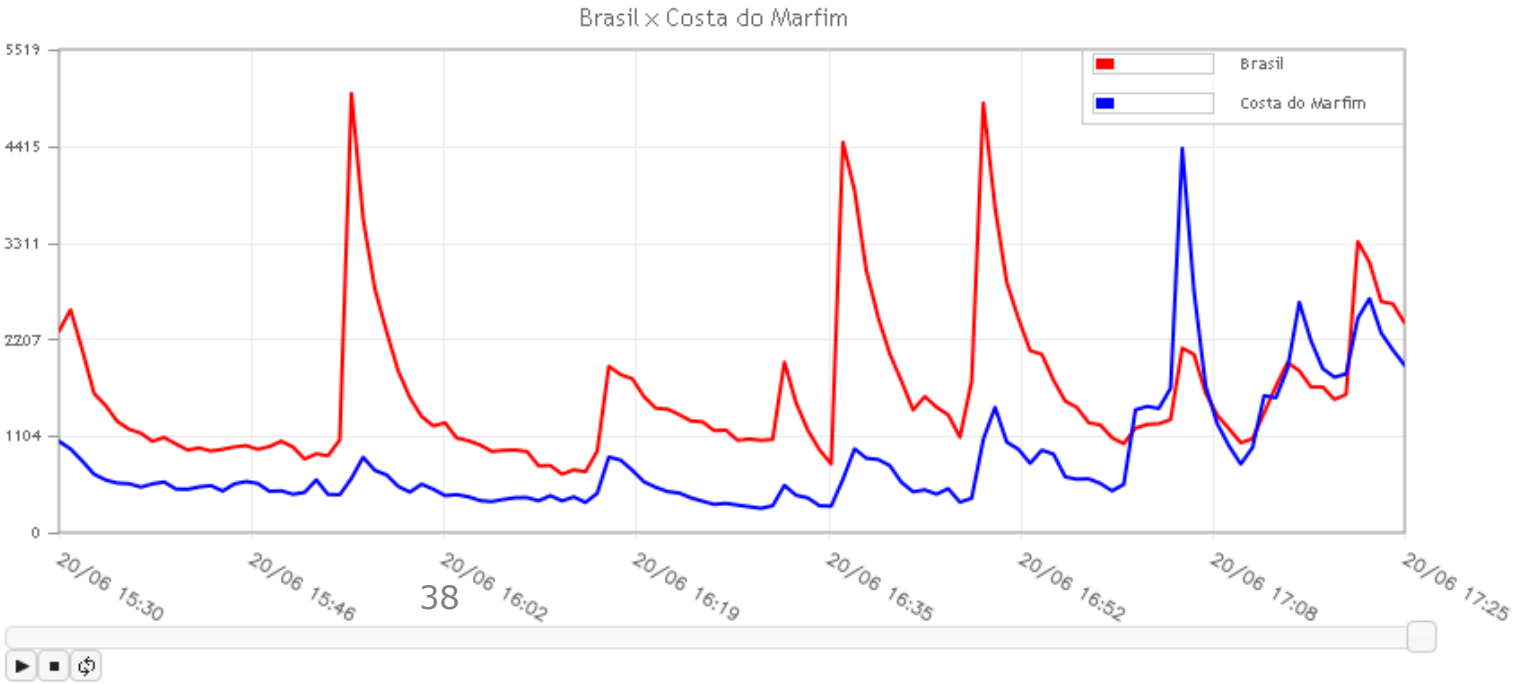
Cidade	Pop.	Tweets	Tx.Inc
Betim	377547	14	1.8230e-04
Brumadinho	34013	1	1.4289e-04
Contagem	603048	22	1.7922e-04
Ibirite	159026	4	1.2110e-04
Itabira	109551	6	2.7305e-04
Itauna	85396	1	5.2124e-05
Joao Monlevade	73451	4	2.7146e-04
Matozinhos	32973	1	1.4765e-04
Ribeirao das Neves	296376	2	2.6665e-05
Sabara	126219	3	1.1399e-04
Santa Barbara	27850	1	1.7627e-04

Dengue Surveillance
Twitter X Official Data



Fonte: Janaína Gomide, Adriano Veloso, Wagner Meira Jr., Virgílio Almeida, Fabrício Benevenuto, Fernanda Ferraz, Mauro Teixeira: Dengue Surveillance Based on a Computational Model of Spatio-temporal Locality of Twitter. WebSci 2011: 1-8.

World Cup Watch
Twitter View of a Match



Mineração de Dados Abertos: Enem 2010

- **0** – Baixa Renda, Zona Urbana, Possui Internet, Escolaridade do pai: ensino médio
- **1** – Baixa Renda, Zona Urbana, Não possui internet, Escolaridade do pai: ensino médio
- **2** – Baixa Renda, Escolaridade do pai: ensino fund., Possui internet
- **3** – Média Renda, Escolaridade do pai: ensino fund., Possui internet
- **4** – Baixa Renda, Não possui internet, Zona Rural, Escolaridade do Pai: ensino fund.
- **5** – Baixa Renda, Escolaridade do pai: ensino fund., Zona Urbana, Possui Internet
- **6** – Média/Alta Renda, Escolaridade do pai: ensino superior, Escola Particular
- **7** – Média Renda, Escolaridade do pai: ensino fund., Possui internet

Questão	Resposta	Qtd	%
Você tem em sua casa? Acesso à Internet	**Não**	**105.849**	**26.46%**
	Sim	**294.151**	**73.53%**
Até quando seu pai estudou?	**Até o ensino fundamental**	**218.811**	**54.70%**
	Até o ensino médio	**115.290**	**28.82%**
	Mais que o ensino médio	**65.899**	**16.47%**
Somando a sua renda com a renda das pessoas que moram com você, quanto é, aproximadamente, a renda familiar mensal?	**Até 2 salários mínimos**	**282.033**	**70.50%**
	Entre 2 e 12 salários mínimos	**101.427**	**25.35%**
	Acima de 12 salários mínimos	**16.540**	**4.13%**
Sua casa está localizada em?	**Rural ou indígena**	**21.478**	**5.36%**
	Urbana	**378.522**	**94.63%**
Tipo de escola?	**Escola pública**	**365.036**	**91.25%**
	Escola Particular	**34.964**	**8.74%**
Nota	**<500 pontos**	**146466**	**36.61%**
	>= 500 e <700 pontos	**244131**	**61.03%**
	>= 700 pontos	**9.403**	**2.35%**

Fonte: R. de Acácio Leonel Júnior, João H. F. Júnior, Tércio J. da Silva, Ticiana L. Coelho da Silva, Régis P. Magalhães: Mineração de Dados Abertos. JORNASCI 2014.

Eleições 2014 - Análise de Sentimento

Eleições 2014 - Análise de Sentimento

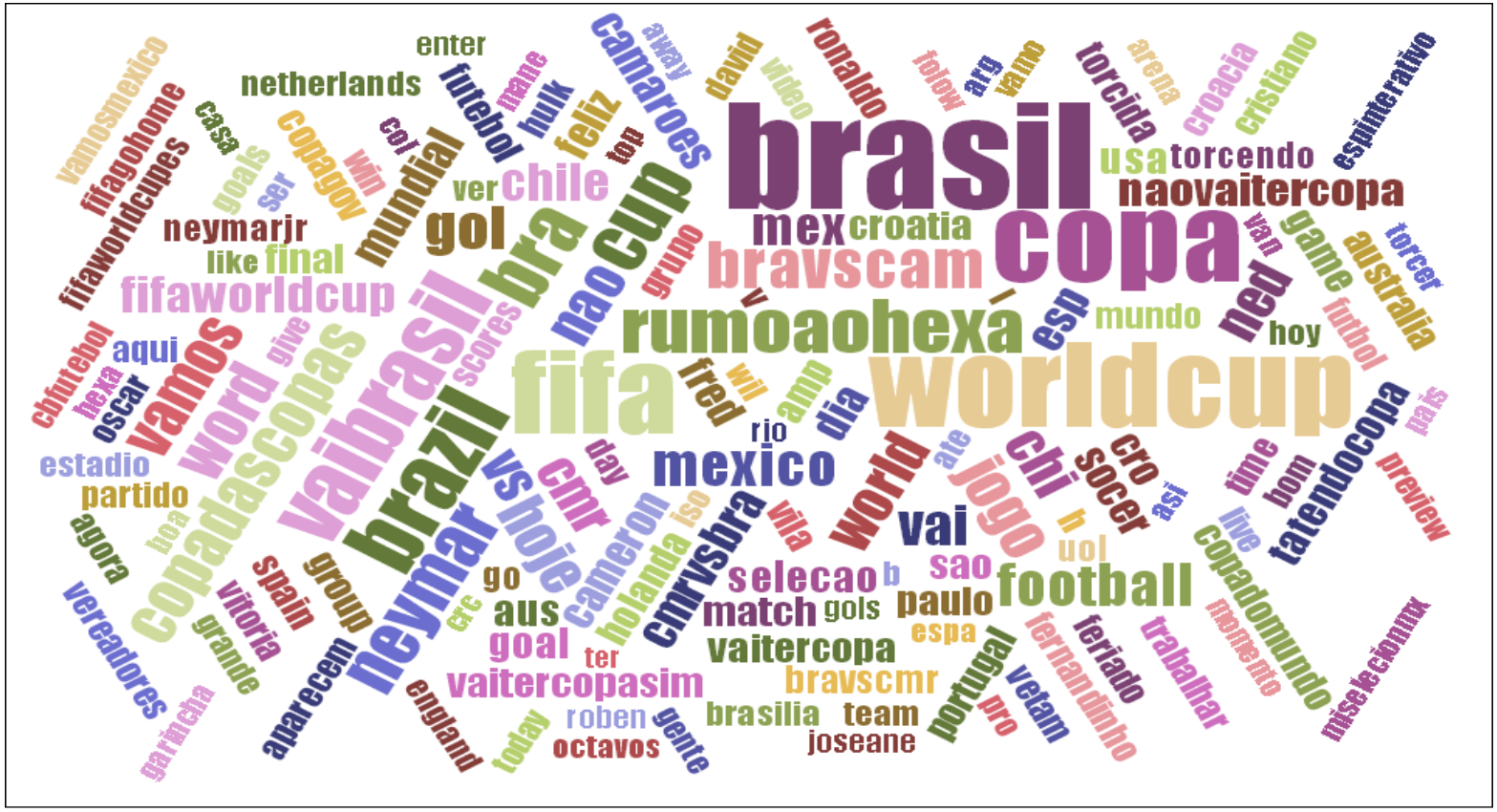
Copa FIFA 2014 – Análise de Sentimentos – Brasil X Croácia (1º jogo)

Fonte: José Adail Carvalho Filho, João Lucas Leite, Ticiana L. Coelho da Silva: Desenvolvimento de um modelo de classificação de tweets para analisar as opiniões de usuários do Twitter sobre osjogos da Seleção Brasileira de Futebol na Copa do Mundo da FIFA Brasil 2014. ENUCOMP 2014.

Copa FIFA 2014 – Análise de Sentimentos – Brasil X México (2º jogo)

Fonte: José Adail Carvalho Filho, João Lucas Leite, Ticiana L. Coelho da Silva: Desenvolvimento de um modelo de classificação de tweets para analisar as opiniões de usuários do Twitter sobre osjogos da Seleção Brasileira de Futebol na Copa do Mundo da FIFA Brasil 2014. ENUCOMP 2014.

Copa FIFA 2014 – Análise de Sentimentos – Brasil X Camarões (3º jogo)



Fonte: José Adail Carvalho Filho, João Lucas Leite, Ticiana L. Coelho da Silva: Desenvolvimento de um modelo de classificação de tweets para analisar as opiniões de usuários do Twitter sobre osjogos da Seleção Brasileira de Futebol na Copa do Mundo da FIFA Brasil 2014. ENUCOMP 2014.

Copa FIFA 2014 – Análise de Sentimentos – Brasil X Chile (4º jogo)

Fonte: José Adail Carvalho Filho, João Lucas Leite, Ticiana L. Coelho da Silva: Desenvolvimento de um modelo de classificação de tweets para analisar as opiniões de usuários do Twitter sobre osjogos da Seleção Brasileira de Futebol na Copa do Mundo da FIFA Brasil 2014. ENUCOMP 2014.

Copa FIFA 2014 – Análise de Sentimentos – Brasil X Colômbia (5º jogo)

Fonte: José Adail Carvalho Filho, João Lucas Leite, Ticiana L. Coelho da Silva: Desenvolvimento de um modelo de classificação de tweets para analisar as opiniões de usuários do Twitter sobre osjogos da Seleção Brasileira de Futebol na Copa do Mundo da FIFA Brasil 2014. ENUCOMP 2014.

Copa FIFA 2014 – Análise de Sentimentos – Brasil X Alemanha (6º jogo)

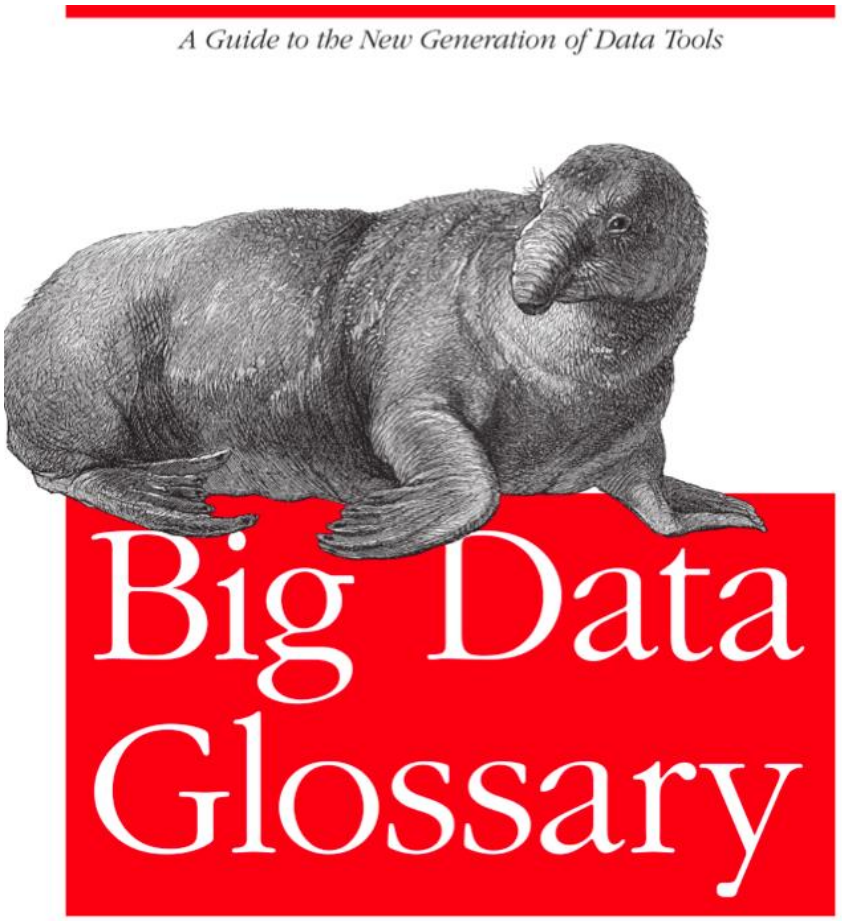
Fonte: José Adail Carvalho Filho, João Lucas Leite, Ticiana L. Coelho da Silva: Desenvolvimento de um modelo de classificação de tweets para analisar as opiniões de usuários do Twitter sobre osjogos da Seleção Brasileira de Futebol na Copa do Mundo da FIFA Brasil 2014. ENUCOMP 2014.

Copa FIFA 2014 – Análise de Sentimentos – Brasil X Holanda (7º jogo)

Fonte: José Adail Carvalho Filho, João Lucas Leite, Ticiana L. Coelho da Silva: Desenvolvimento de um modelo de classificação de tweets para analisar as opiniões de usuários do Twitter sobre osjogos da Seleção Brasileira de Futebol na Copa do Mundo da FIFA Brasil 2014. ENUCOMP 2014.

O que podemos usar agora?

Tecnologias para Big Data



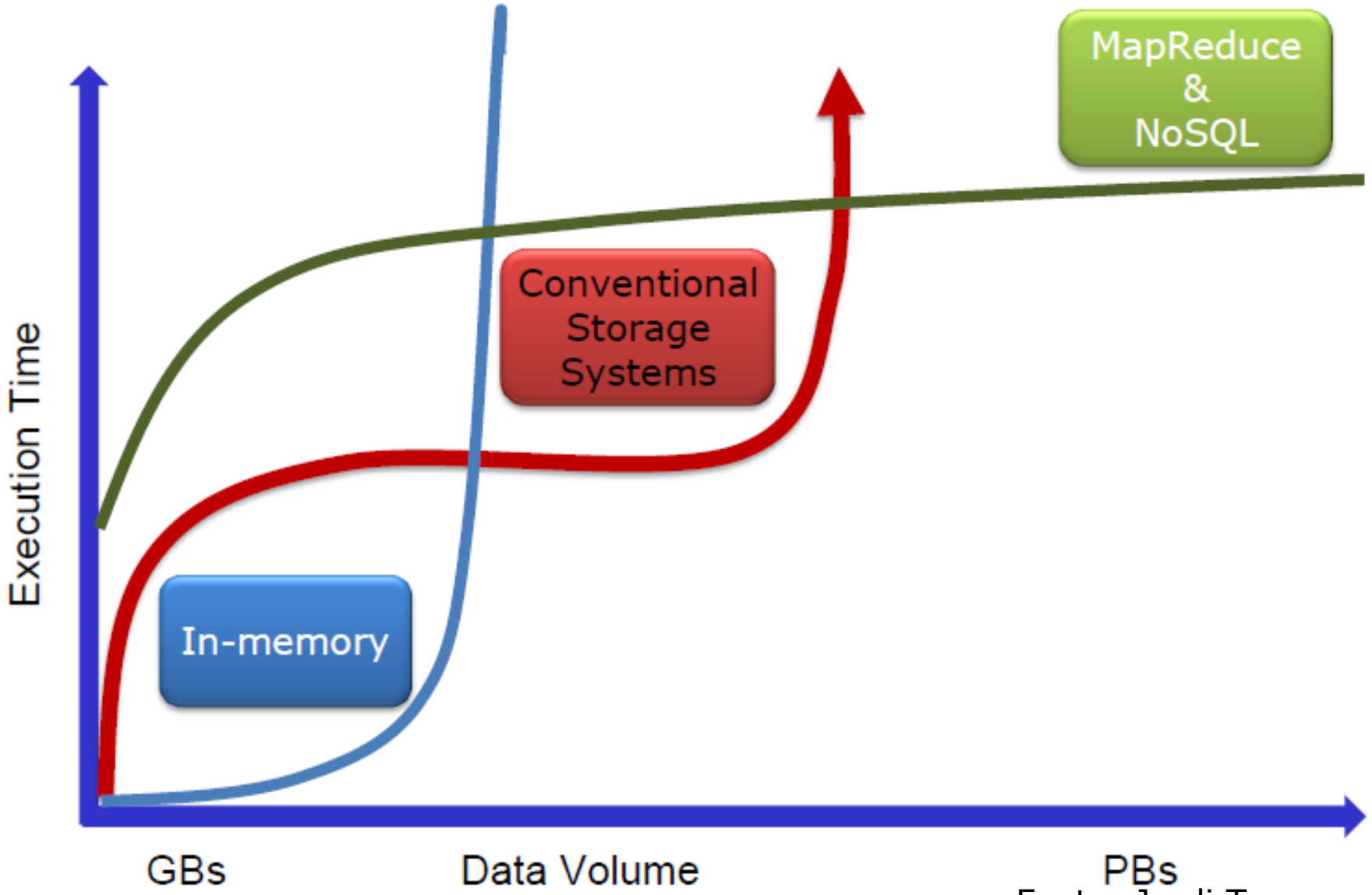
Tecnologias para Big Data

- NoSQL Databases
  - MongoDB, CouchDB, Cassandra
- Map Reduce
  - Hadoop, Hive, Pig
- Storage
  - S3, Hadoop Distributed File System
- Servers
  - EC2, Google App Engine
- Processing
  - R, Yahoo! Pipes
- NLP
  - NL Toolkit, OpenNLP
- Machine Learning
  - WEKA, Mahout
- Visualization
  - Gephi, GraphViz
- Serialization
  - JSON, BSON

Novos Sistemas para Big Data

- SGBDs Relacionais não podem suportar tudo
  - NoSQL
  - NewSQL

Novos Sistemas para Big Data

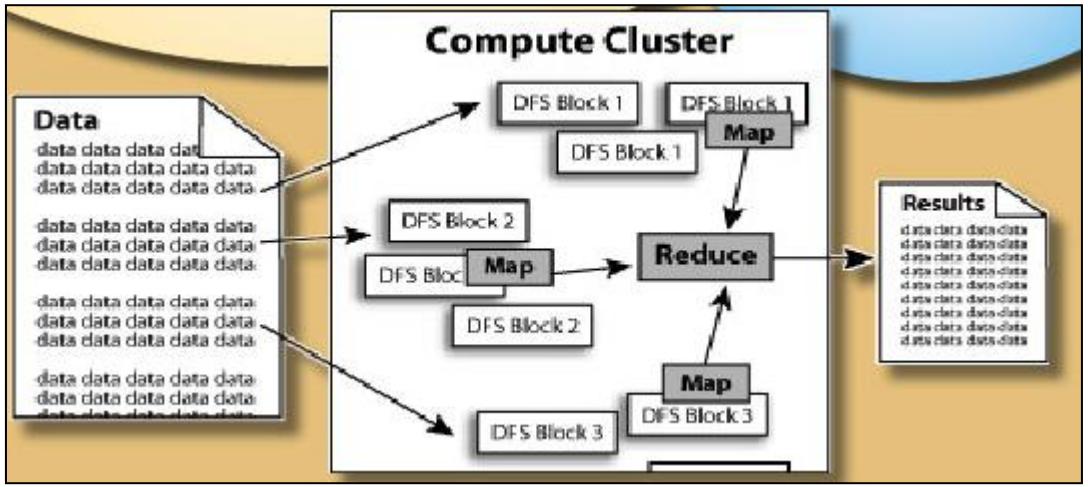


Fonte: Jordi Torres

- *“Let’s start with the obvious observation: data intensive processing is beyond the capability of any individual machine and requires clusters—which means that large-data problems are fundamentally about organizing computations on dozens, hundreds, or even thousands of machines. This is exactly what MapReduce does…”*

Data-Intensive Text Processing with MapReduce
Jimmy Lin and Chris Dyer
University of Maryland, College Park

MapReduce



- Os dados são distribuídos entre os computadores de um cluster.
- Programas **Map** em cada computador analisam e processam seu subconjunto dos dados e retornam resultados intermediários como pares chave-valor.
- O passo **Reduce** ordena e combina os resultados intermediários para retornar um resultado final.

Bancos de dados NoSQL

- Capazes de tratar imensos volumes de dados estruturados e não estruturados.
- Tipos / Orientados a:
  - **Colunas / Tabular**
    - **Google Big Table** – usado internamente pelo Google e Google App Engine.
    - Apache **HBase** (baseado no Big Table)**;** Apache **Cassandra** (baseado no DynamoBD da Amazon)**.**
      - **Facebook (criador do Cassandra) substituiu Cassandra por HBase (Mensagens).**
      - **Netflix usa Cassandra como BD de seus serviços de streaming.**
      - **Twitter usa Cassandra (Analytics, TopTweets, ...) e MySQL (para Tweets)**
  - **Documento**
    - **MongoDB**, Apache **CouchDB** (documentos JSON).
  - **Chave/Valor**
    - **DynamoDB** da Amazon; Riak; **Redis**; Cache; Voldemort.
  - **Grafo**
    - **Neo4j**, Allegro, Virtuoso.

http://nosql-database.org/

Infraestrutura para Análise em Big Data

Computação em Nuvem

The Big Data Landscape

Apps

Vertical

PREDICTIVE POLICING · Atigeo · MYRRIX · ellucian. · Placed.

Operational Intelligence

splunk> · loggly · sumologic

Ad/Media

METAMARKETS · collective[i] · rocketfuel · Recorded Future · bluefin · bloomreach · Quantivo · LuckySort · Media Science · TURN · DataXu

Data As A Service

factual. · FICO · kaggle · knoema beta · GNIP · DATASIFT · Windows Azure Marketplace · INRIX · LexisNexis · SPACE CURVE · LOCATE Everything Location

Business Intelligence

ORACLE | Hyperion · SAP Business Objects · RJMetrics · Microsoft Business Intelligence · IBM COGNOS · birst · JASPERSOFT · pentaho · MicroStrategy · Autonomy · bime · DOMO · Chart.io · GoodData · ATTIVIO

Analytics and Visualization

+ableau · Palantir · OPERA · metaLayer · dataspora · TERADATA ASTER · centrifuge · SAS · TIBCO · KARMASPHERE · panopticon Real-Time Visual Data Analysis · Datameer · CIRRO · platfora · ClearStory · saffron · alteryx · visual.ly · AYATA · QlikView

Infrastructure

Analytics

Hortonworks · VERTICA An HP Company · MAPR · cloudera · INFOBRIGHT · ParAccel · EMC² · GREENPLUM. · NETEZZA · kognitio · DATASTAX · EXASOL · calpont

Operational

COUCHBASE · 10gen · HADAPT · TERRACOTTA · VoltDB · MarkLogic · INFORMATICA

As A Service

Qubole · amazon web services · Windows Azure · MORTAR · infochimps · Google BigQuery

Structured DB

ORACLE · MySQL · Microsoft SQL Server · PostgreSQL · IBM DB2 · SYBASE · memsql · TERADATA

Technologies

hadoop · hadoop MapReduce · mahout · APACHE HBASE · Cassandra

dave@vcdave.com

bigdatalandcsape.com

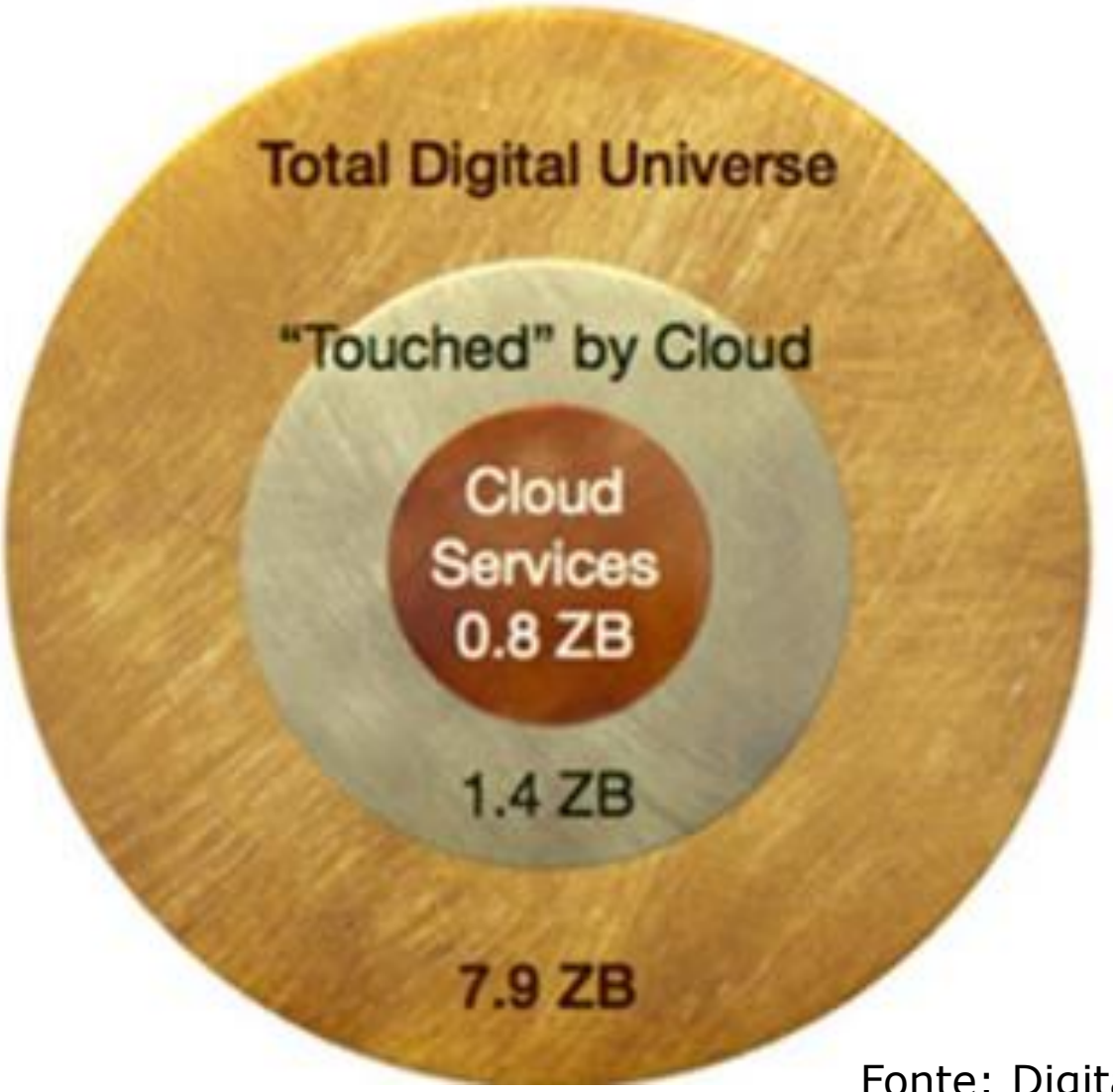
Desafios em Big Data

O que precisamos ?

"Um projeto Big Data requer uma transformação sincronizada entre pessoas, processos e tecnologias. Todas as três devem marchar em sincronia, caso contrário o projeto falhará."

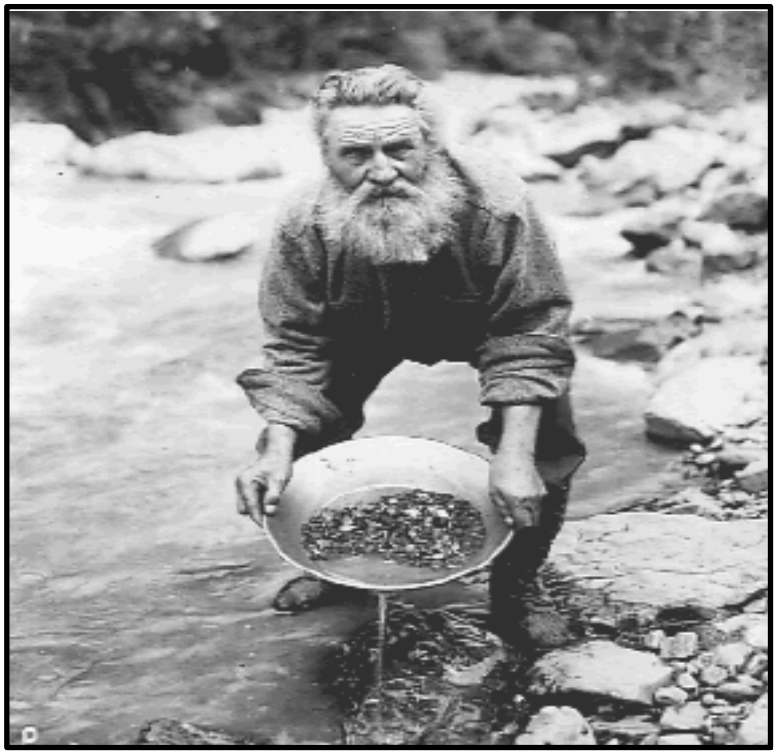
Fonte: Big Data, Big Analytics;
Minelli, Chamber, Dhira;
Wiley CIO Series, 2013

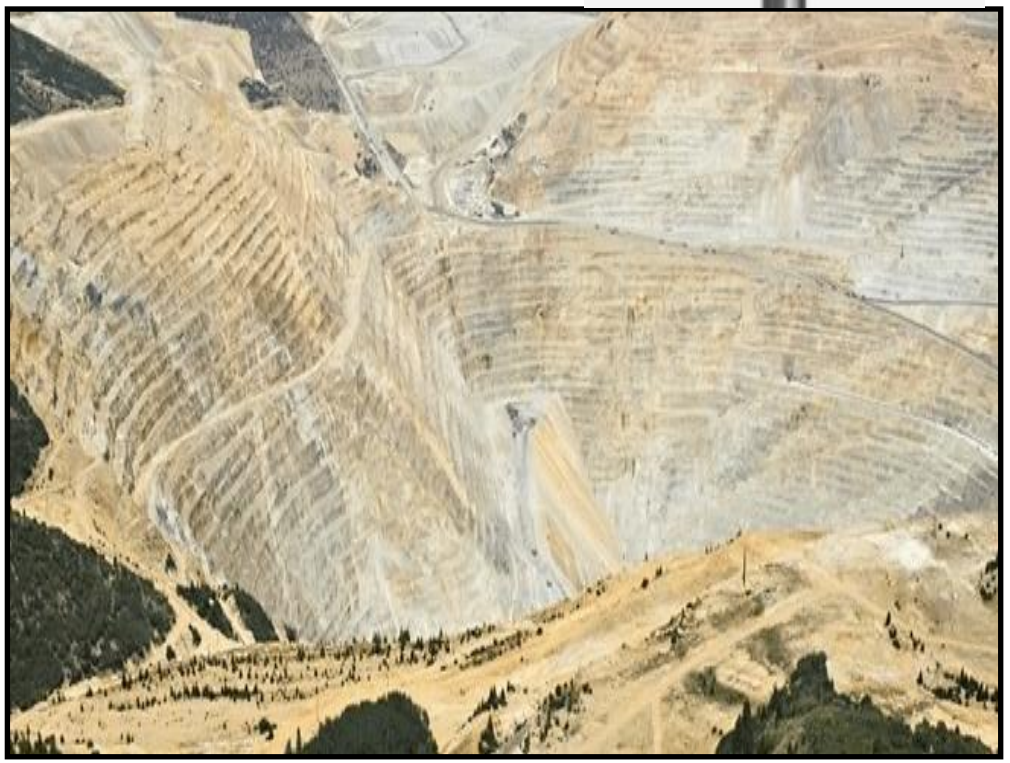
Existe muito dado intocado ...



Fonte: Digital Universe Study

As ferramentas de análise são pobres …





Precisamos de ferramentas de análise sofisticadas ...

- Lidem com 6 V’s do Big Data;
- Não mais limitada à sumarização e agregação de dados;
- Precisamos realizar análises complexas baseadas em mineração de dados e técnicas estatísticas;
- Precisamos urgentemente de técnicas de gerenciamento de Big Data;

Usar tecnologias maduras ...

- 40 anos de pesquisa em BD;
- Tecnologias existentes lidam bem com:
  - Volume de Dados,
  - Variedade de Modelos;
  - Stream de Dados;
  - Fonte de dados heterogeneas e distribuídas;
- Porém, são necessárias novas técnicas:
  - Análise de Padrões Temporais;
  - Processamento em tempo real;
  - Incerteza, subjetividade e ambiguidade;

e adaptá-las ao novo contexto ...

- Otimização baseada em I/O não deve ser uma premissa;
- Muitas análises precisam iterar sobre o dado diversas vezes;
- Devemos mover os programas e não os dados;
- Falha é uma regra e não uma exceção !

E ainda tem mais …

- Precisamos de algoritmos eficientes: O(n), O(log n)

Cientista de Dados

"Até 2015 serão necessários **4,4 milhões** de experts em interpretação de dados em larga escala, sendo que **500 mil deles serão para o Brasil**".

Data Scientist:

The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data.

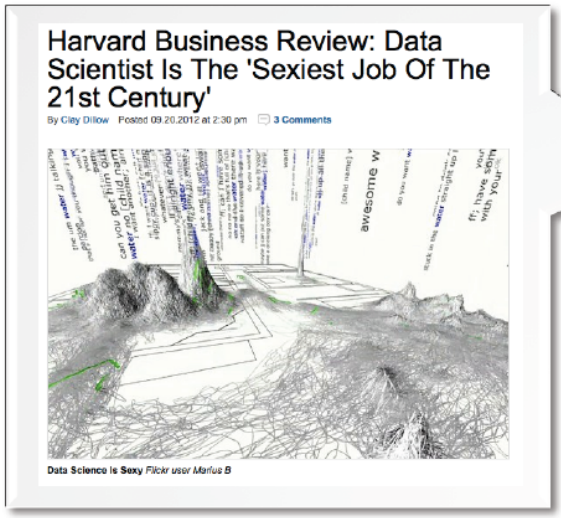
by Thomas H. Davenport and D.J. Patil

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

70 Harvard Business Review October 2012

Fonte: Instituto Gartner

Cientista de Dados

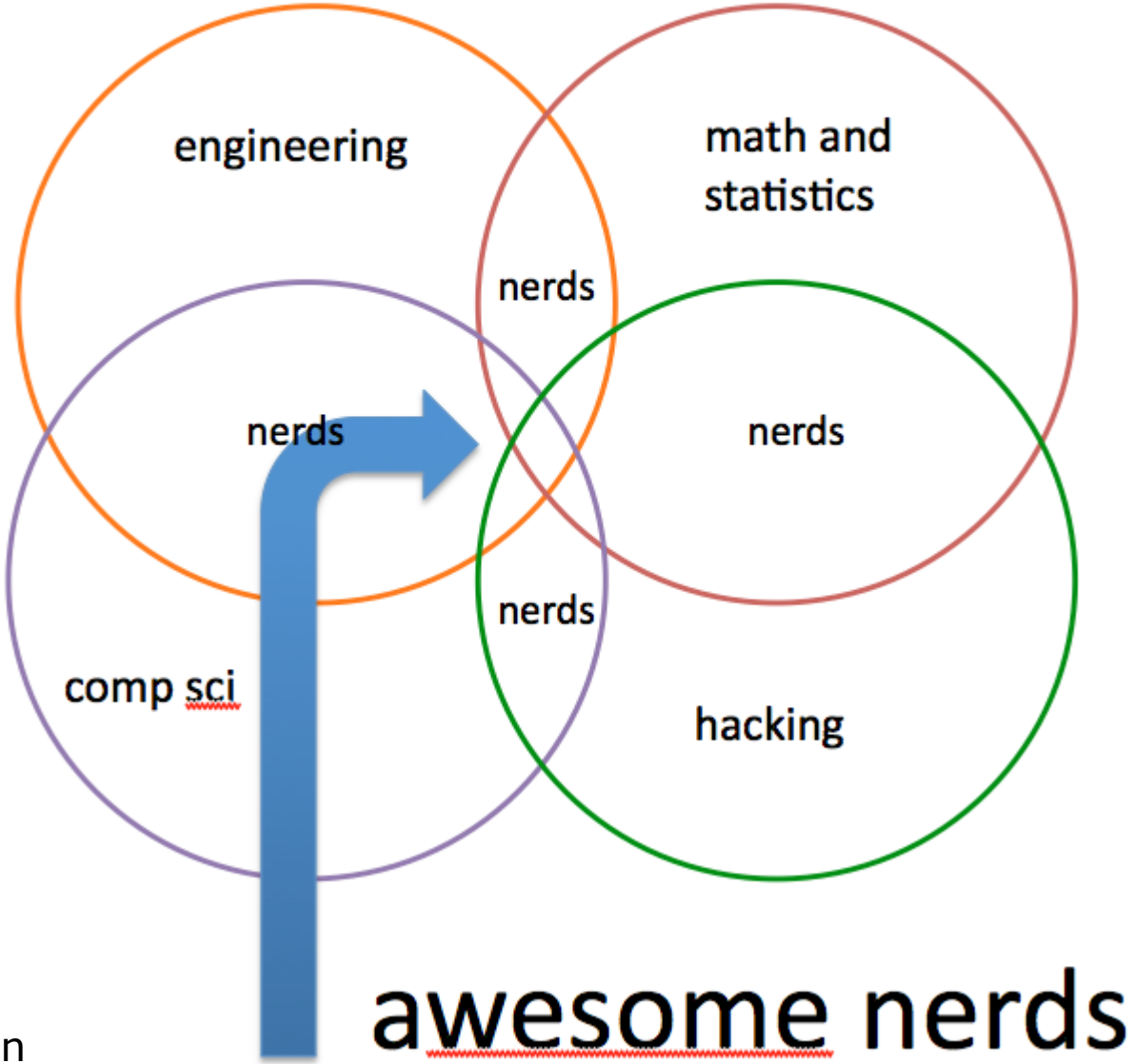


Data Science applies advanced **analytical** tools and algorithms to generate **predictive insights** and **new** product **innovations** that are a direct result of the data

Who Is The Data Scientist?

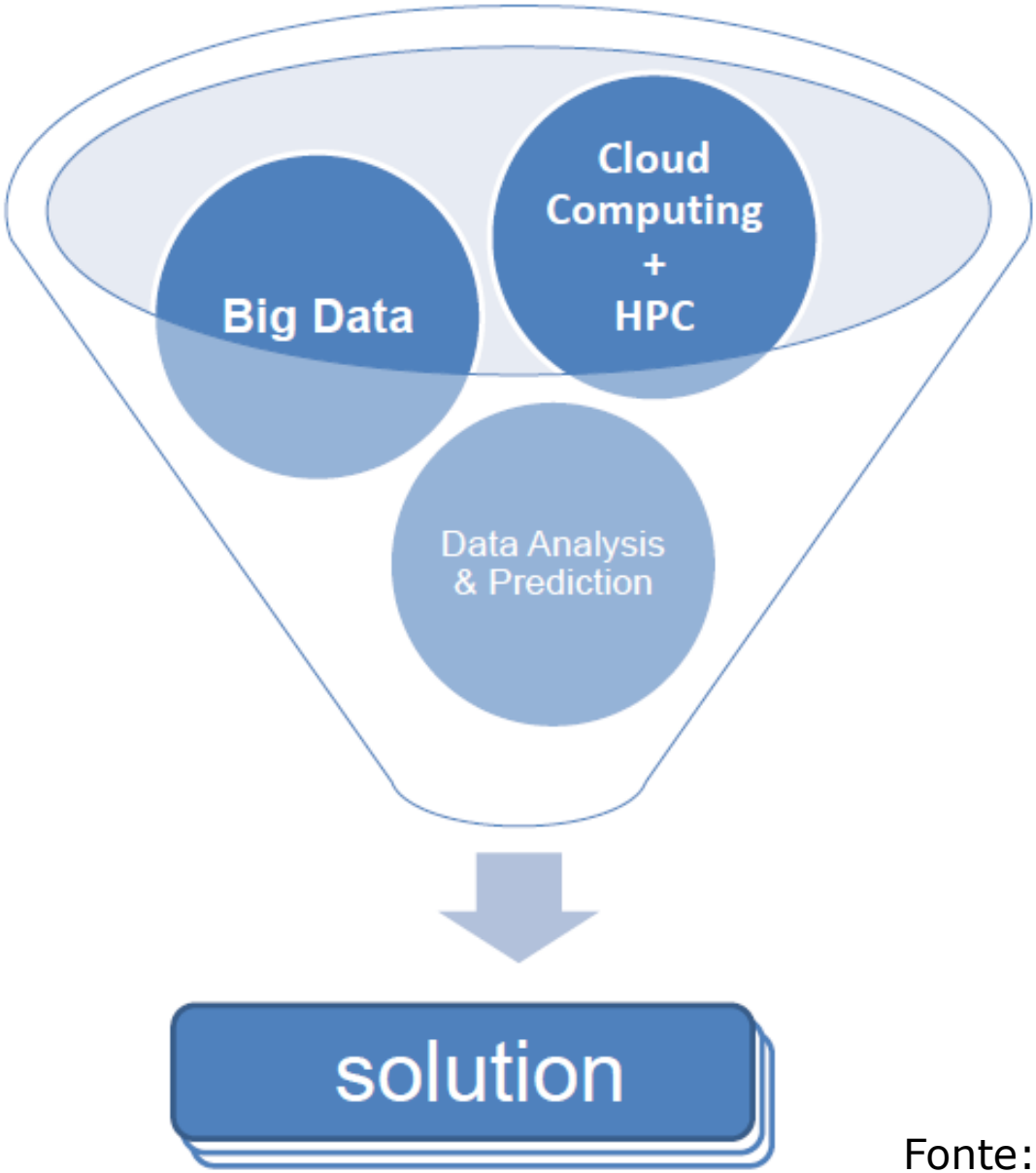
Fonte: EMC

Data scientists?



Fonte: Hilary Mason

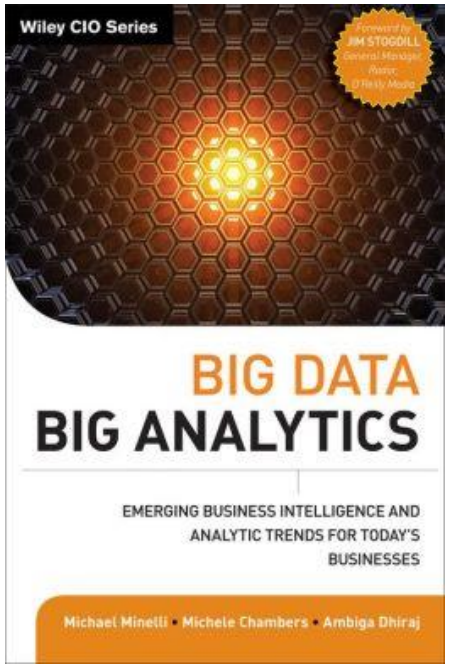
Novo modelo de Inovação?

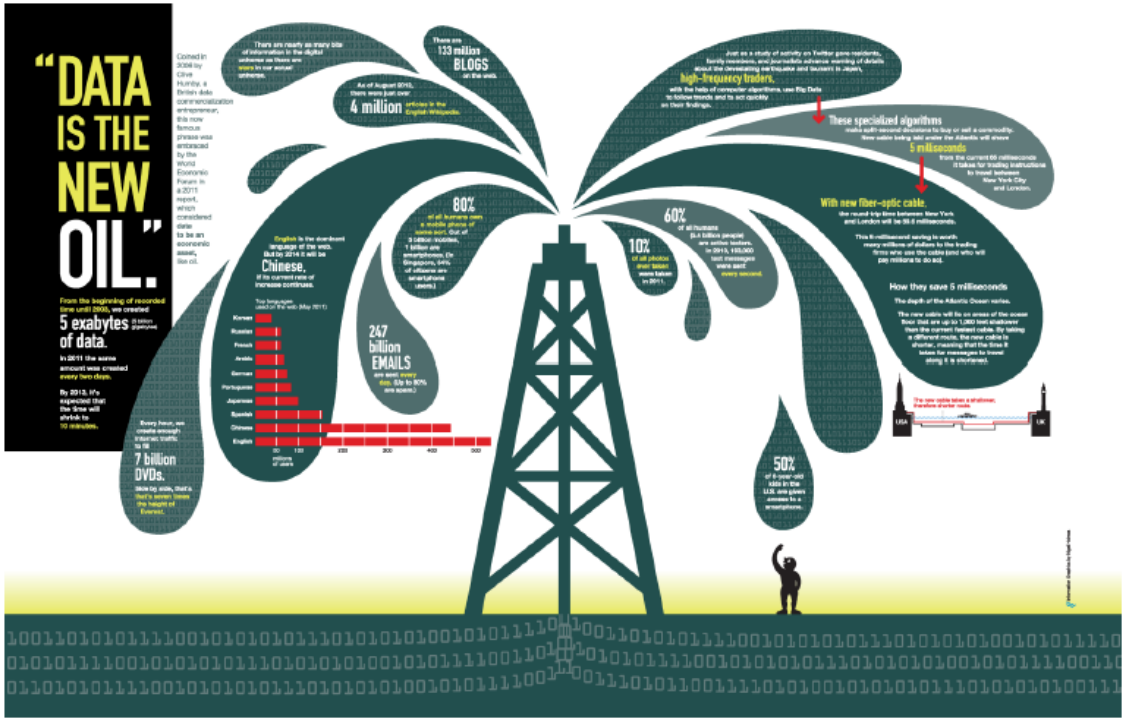


Fonte: Jordi Torres

Big Data está acelerando a **inovação** e melhorando nossas vidas.

“Data is the new gold”





Fonte: ODI European Commission