



**Universidade Federal do Ceará**  
**Departamento de Engenharia de Teleinformática**

# Big Data

**Flávio R. C. Sousa**

**[flaviosousa@ufc.br](mailto:flaviosousa@ufc.br)**

 **[@flaviosousa](https://twitter.com/flaviosousa)**

**[www.lia.ufc.br/~flavio](http://www.lia.ufc.br/~flavio)**

Pig DATA is  
GOING TO  
IMPACT US  
SOON!



CALL  
BRUCE  
"WILIS"



# Introdução

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

**Os dados armazenados vão crescer 50 vezes mais até 2020**



**64 Bilhões** de mensagens em **24 horas**



**100 GB** para análise **3 seg/decisão**

# Introdução

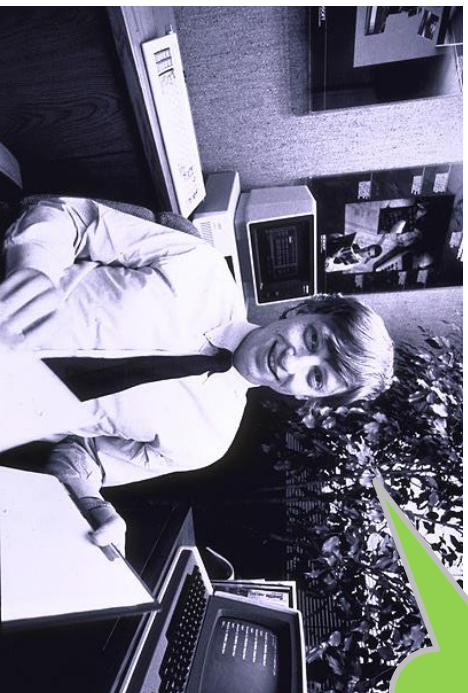
- **Facebook**
  - **1.2B** de usuários
  - **1,13 Trilhões** de "likes"
  - **240B** de fotos e **140.3B** de relacionamentos
  - Crescimento de **7PB** por mês
- **Youtube**
  - **100 horas** de vídeos adicionado a **cada minuto**
- **Bolsa de valores de Nova Iorque**
  - + **1 TB** de dados a cada sessão do pregão
- **Boeing**
  - **640 TB** gerados em um voo transatlântico
- **Wal-Mart**
  - **2,5 PB** e **1 milhão** de transações/hora

# Introdução

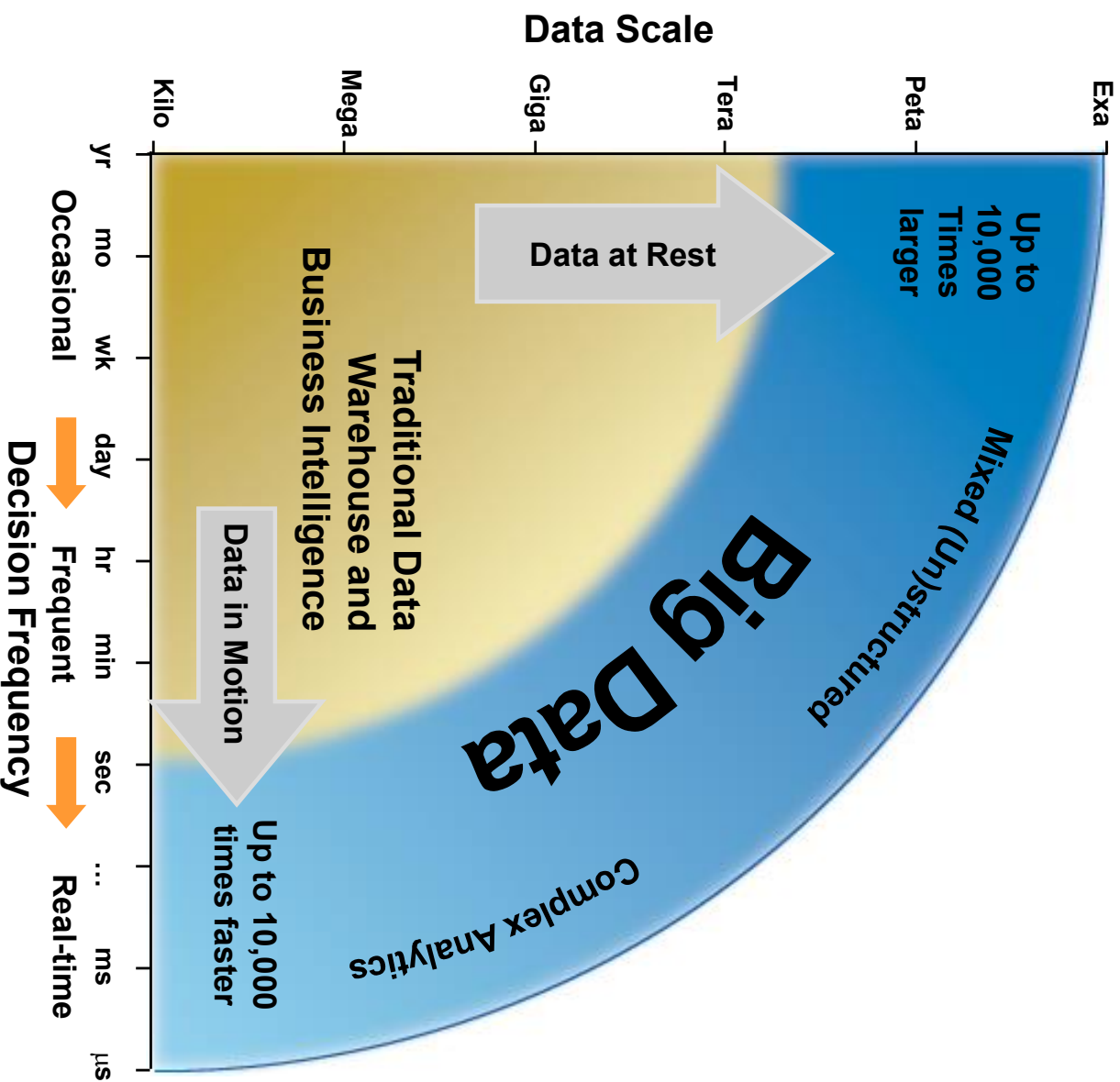
- **LHC CERN**
  - **15 Petabytes** por ano
- **Sloan Digital Sky Survey**
  - **10 Petabytes** gerados a cada varredura
- **Google**
  - **24 Petabytes** processados por dia



**640K** ought to  
be enough for  
anybody.



# Introdução



## Homeland Security

600,000 records/sec, 50B/day

1-2 ms/decision

320TB for Deep Analytics

## Telco Promotions

100,000 records/sec, 6B/day

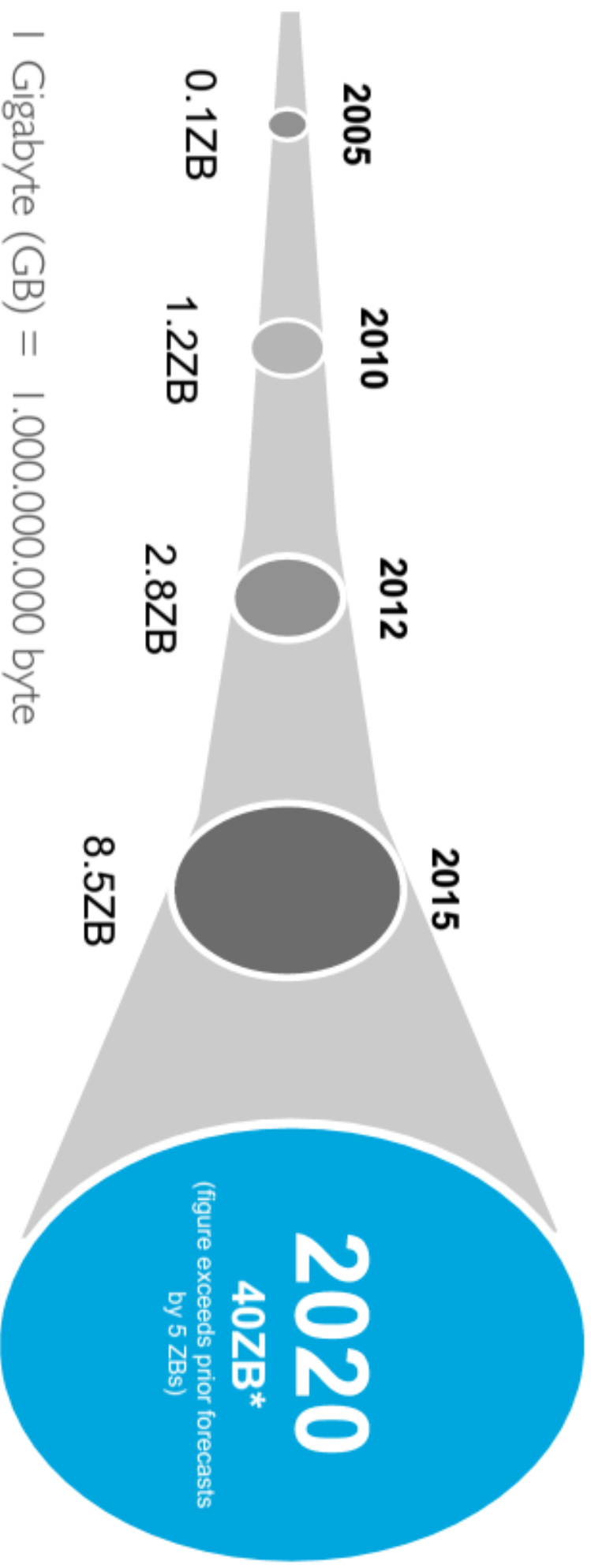
10 ms/decision

270TB for Deep Analytics

Fonte: IBM



# Introdução



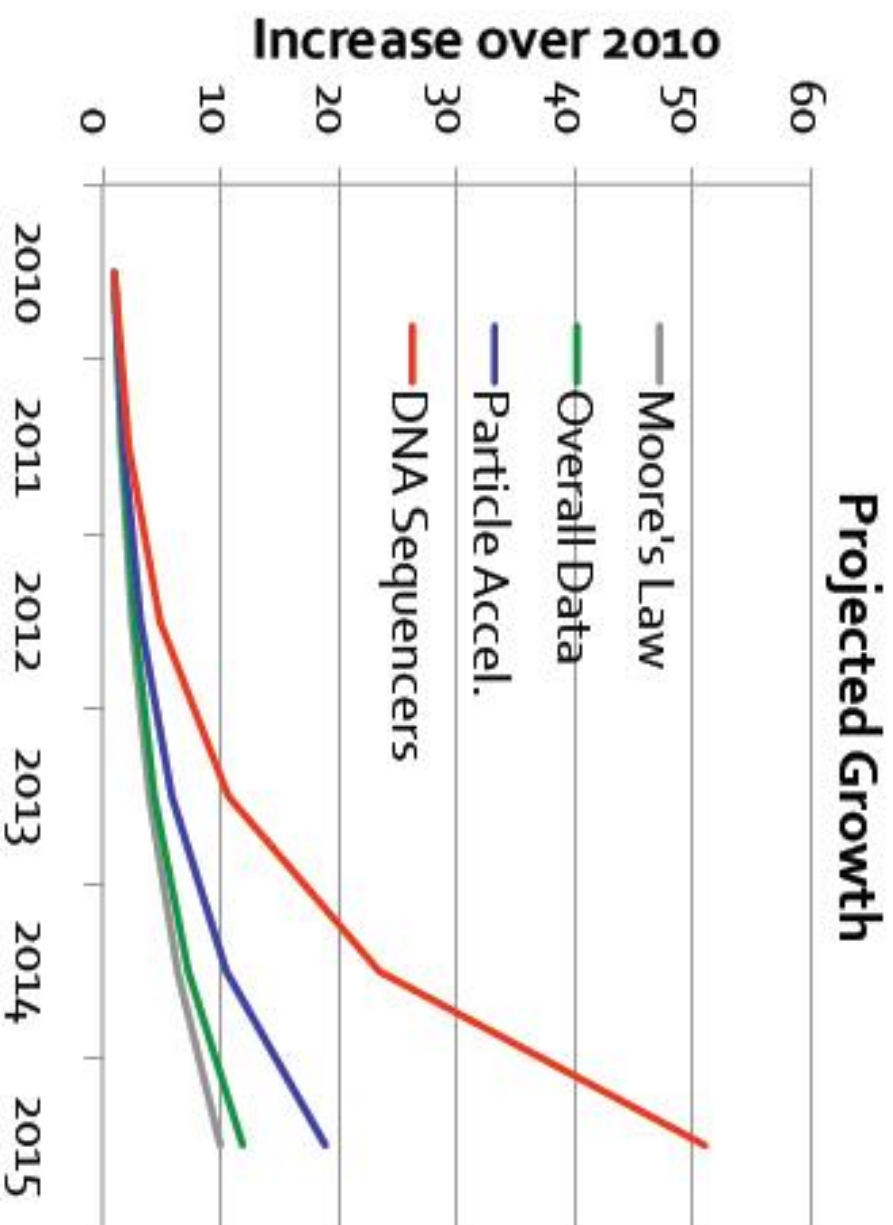
- | Gigabyte (GB) = 1.000.000.000 byte
- | Terabyte (TB) = 1.000 (GB)
- | Petabyte (PB) = 1.000.000 (GB)
- | Exabyte (EB) = 1.000.000.000 (GB)
- | Zettabyte (ZB) = 1.000.000.000.000 (GB)

$$2.7 \text{ ZB} = 85 \text{ B} \times$$



32 GB

# Os dados são “Grandes”



**Data Grows faster than Moore's Law**

[IDC report, Kathy Yelick, LBNL]



Fonte: [Amplab UC Berkeley](#)



# Dados gerados por IoT

## IoT

### Market Size

(by 2025)

McKinsey&Company

**\$6.1T**



**\$7.1T**



**\$14.4T**

## Connected

### Devices

(by 2020)

Gartner

**26B**



**32B**



**50B**

## Data

### Growth

(2013 vs 2020)



Total Data

**4.4ZB → 44.4ZB**

**10x**

IoT Data

**.09ZB → 4.4ZB**

**49x**

# Os dados são “Sujos”

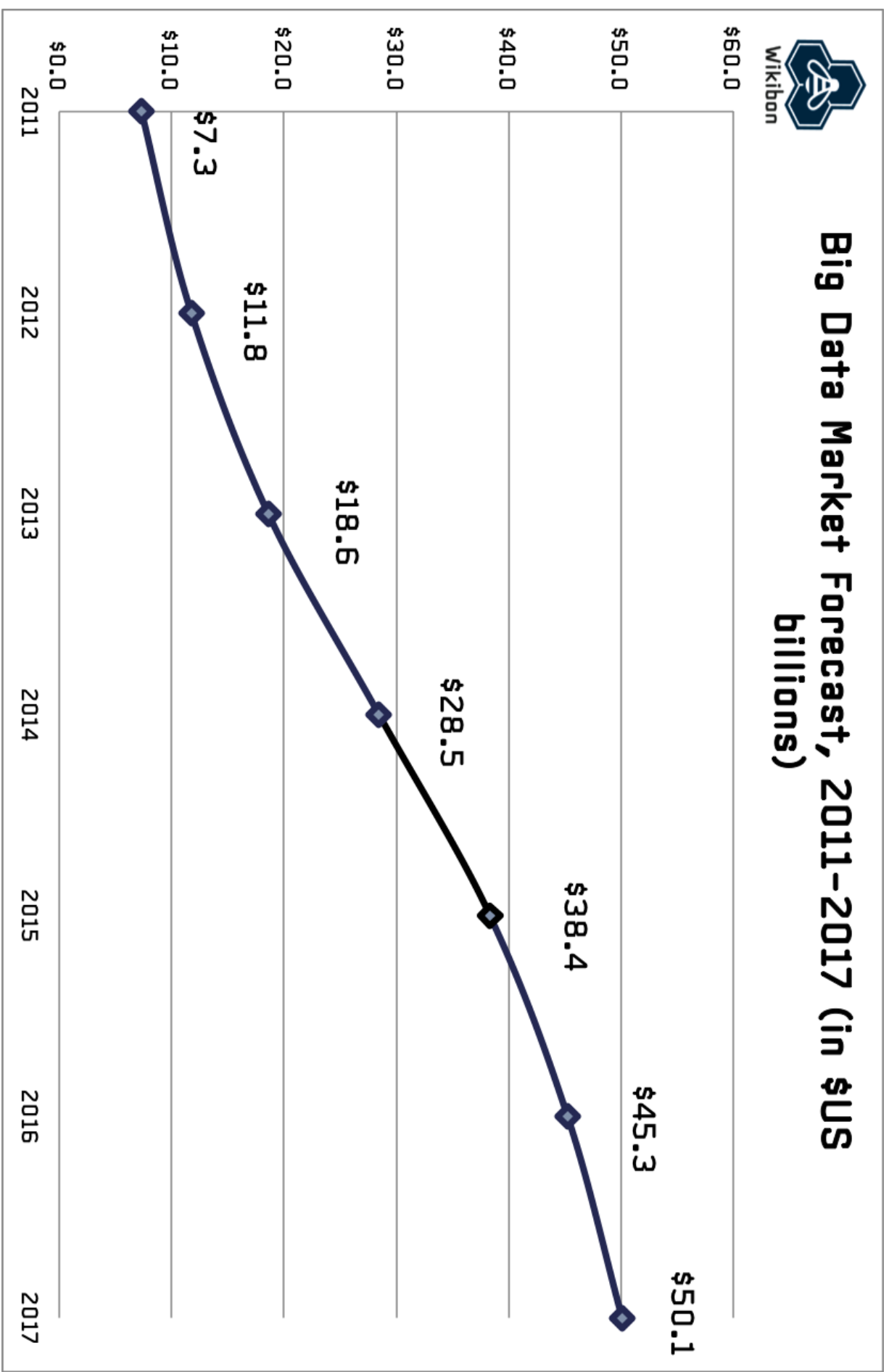
- Diversas fontes de dados
- Sem esquema
- Sintaxe e semântica inconsistente



# 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

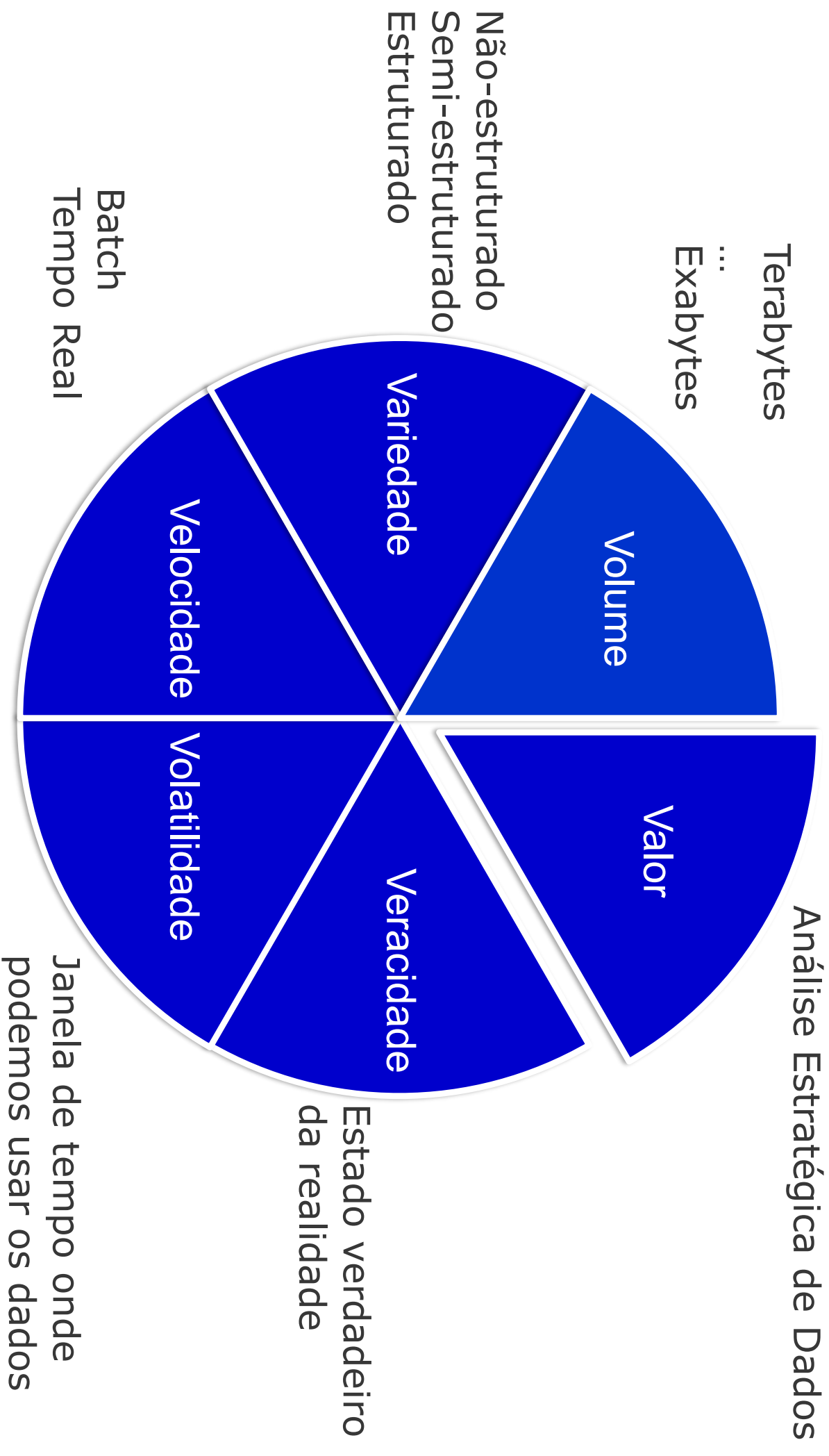


# Big Data

- 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**



# 6 V's do Big Data



# Tecnologias para Big Data



Fonte: [Jordi Torres](#)



**Key:**

- 
- General purpose
- Specialist analytic
- as-a-Service
- NoSQL extension
- BigTables
- Graph
- Document
- Key value stores
- Key value direct access
- Hadoop
- NewSQL extension
- MySQL storage engines
- Advanced clustering/sharding
- New SQL databases
- Data caching extension
- Data caching
- Data grid
- Search
- Appliances
- Off-heap memory
- \* Xeround closed May 2013
- until further notice



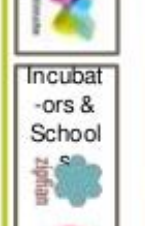
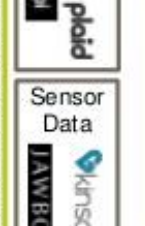
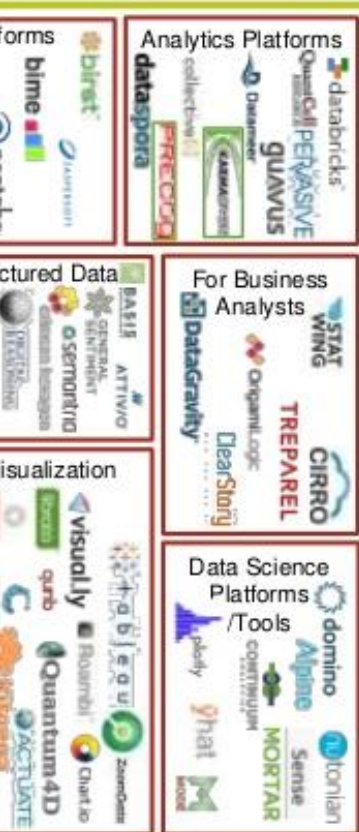
## Infrastructure



## Analytics



## Applications



© Matt Turk (@matturk), Sutan Dong (@sutandong) & FirstMark Capital (@firstmarkcap)



# Tecnologias para Big Data





# Análise para Big Data: Gera Valor

Smarter Healthcare



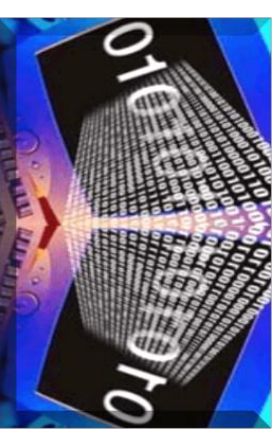
Multi-channel



Finance



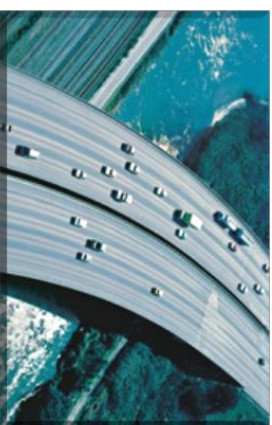
Log Analysis



Homeland Security



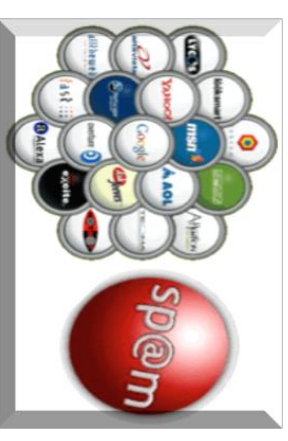
Traffic Control



Telecom



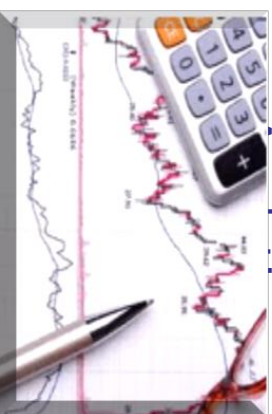
Search Quality



Manufacturing



Trading



Fraud and Risk



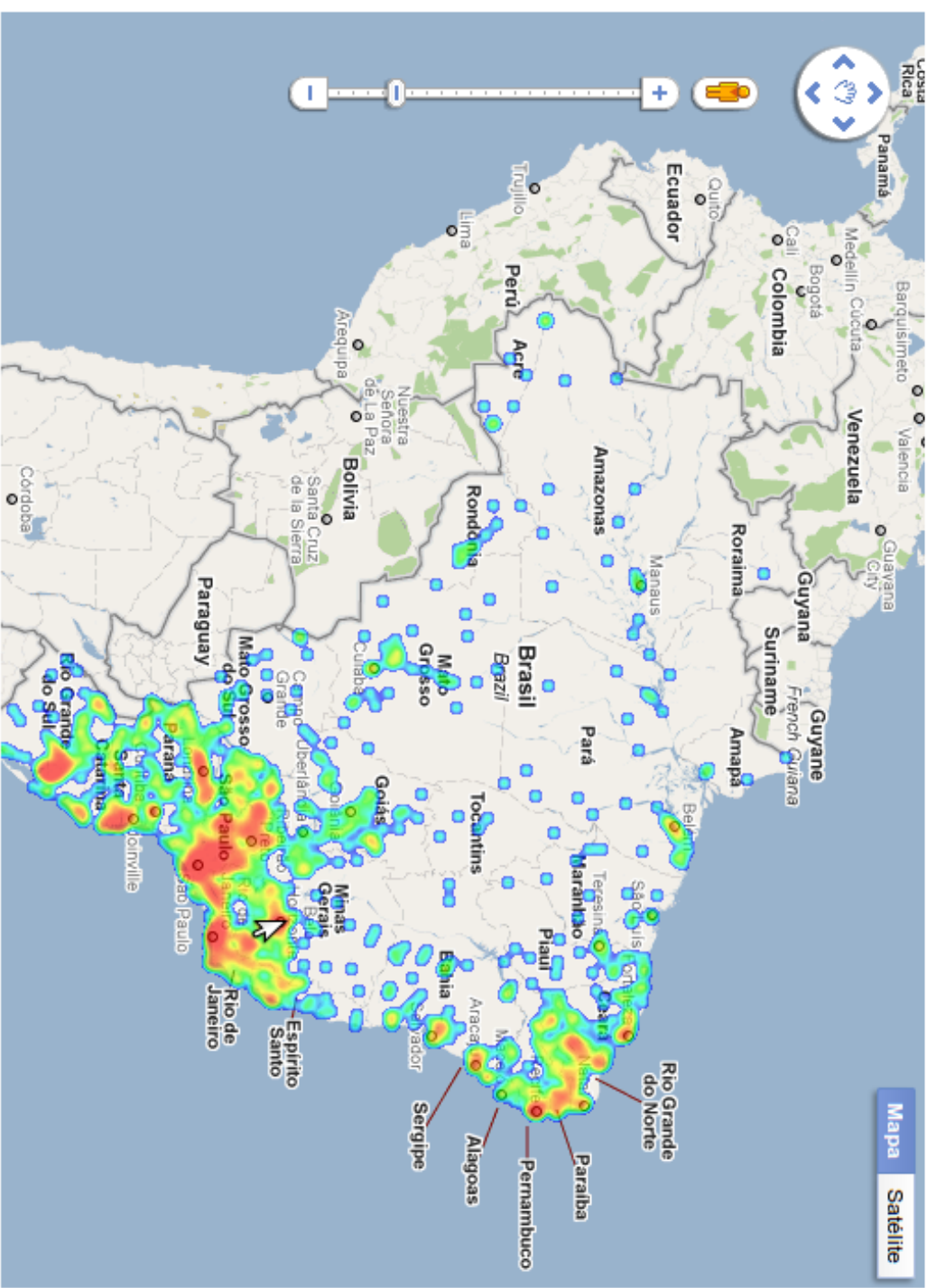
Retail: Churn, NBO



# Dengue Watch: Heat Map

## observatório da dengue

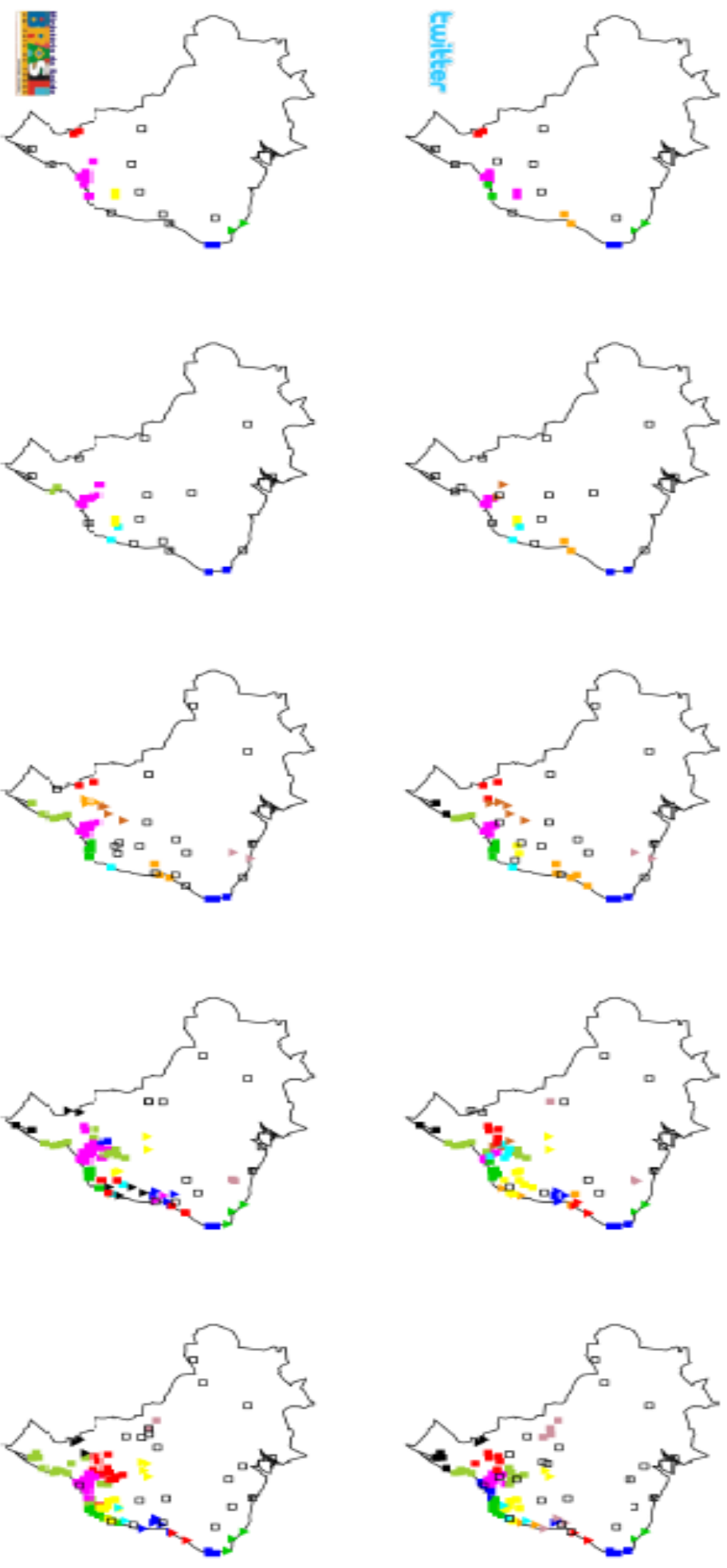
Menções à dengue no Twitter no mês de fev/2011



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
Ibirité	159026	4	1.2110e-04
Itabira	109551	6	2.7305e-04
Itauna	85396	1	5.2124e-05
João Monlevade	73451	4	2.7146e-04
Matosinhos	32973	1	1.4765e-04
Ribeirão 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.](#)

**Como armazenar e processar este  
grande volume de dados?**





# Computação em Nuvem

- Serviços básicos e essenciais são todos entregues de uma forma transparente
- A mesma ideia tem sido aplicada no contexto da informática
  - *Cloud Computing* ou Computação em Nuvem
- Computação em Nuvem
  - Ideia antiga: Software como um Serviço (SaaS)
    - Entrega de aplicações através da Internet
  - Recentemente: “[Hardware, Infraestrutura, Plataforma] como um serviço”
    - “X como um serviço”



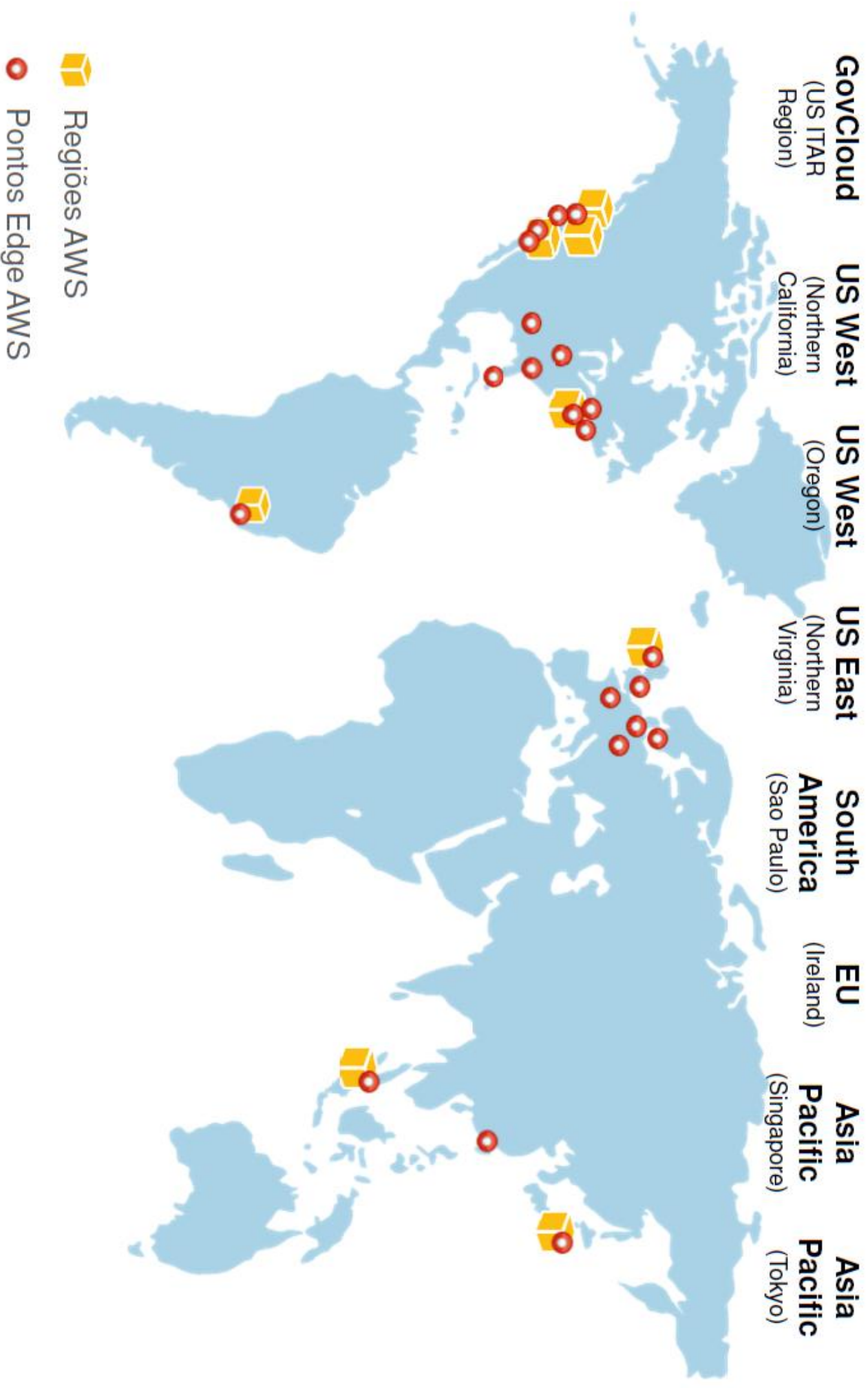


# Computação em Nuvem





# Computação em Nuvem: Amazon AWS

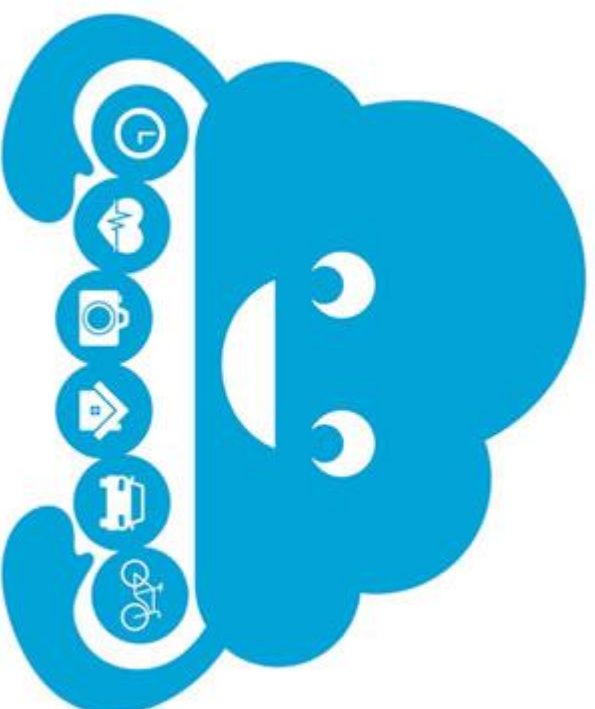


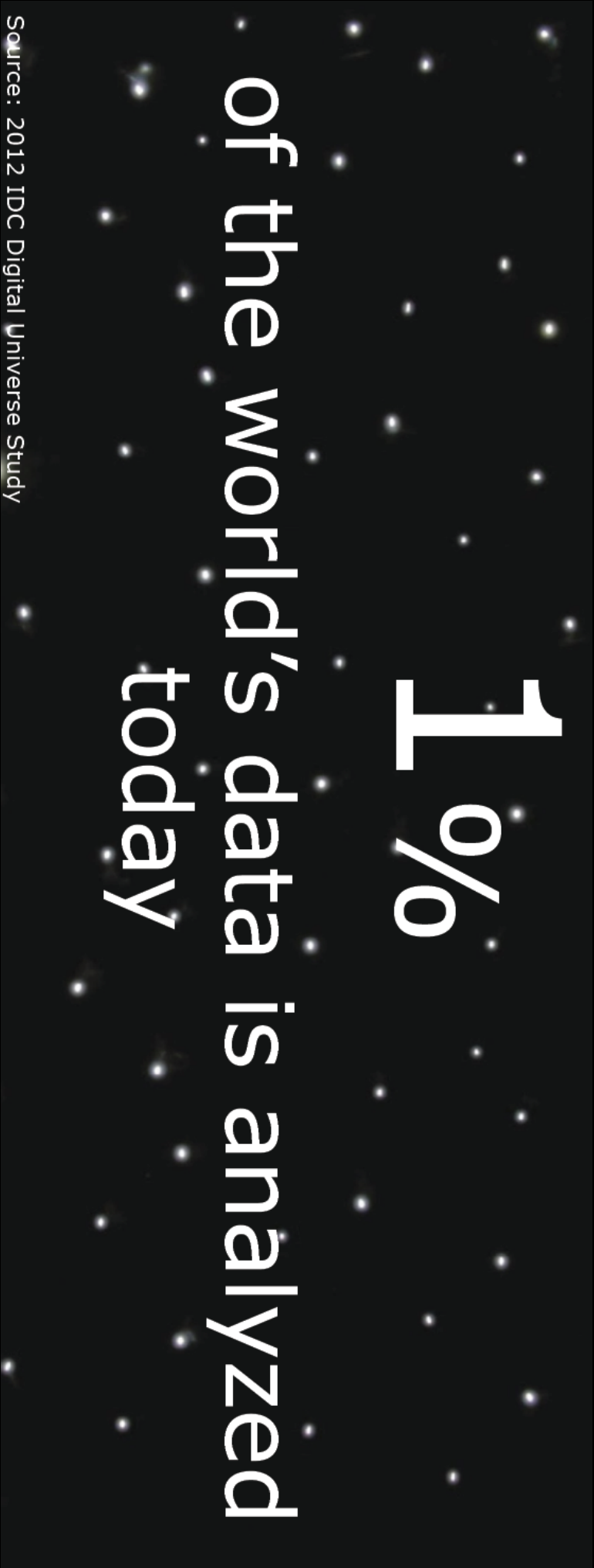
# Computação em Nuvem

- Lista **Top 500**
  - Elenca os 500 supercomputadores mais rápidos do mundo
- **1.064 instâncias** do EC2 foram usadas para criar um supercomputador com **17.024 cores**
- **240 teraflops** de velocidade
  - 240 trilhões de operações por segundo
- Esse supercomputador é o **72º computador** mais rápido do mundo
  - Lista do Top 500 (jun/2012)

**Você** pode alugá-lo por **menos de US\$ 1.000/h**

# Big Data: Desafios





1%  
of the world's data is analyzed  
today

Source: 2012 IDC Digital Universe Study



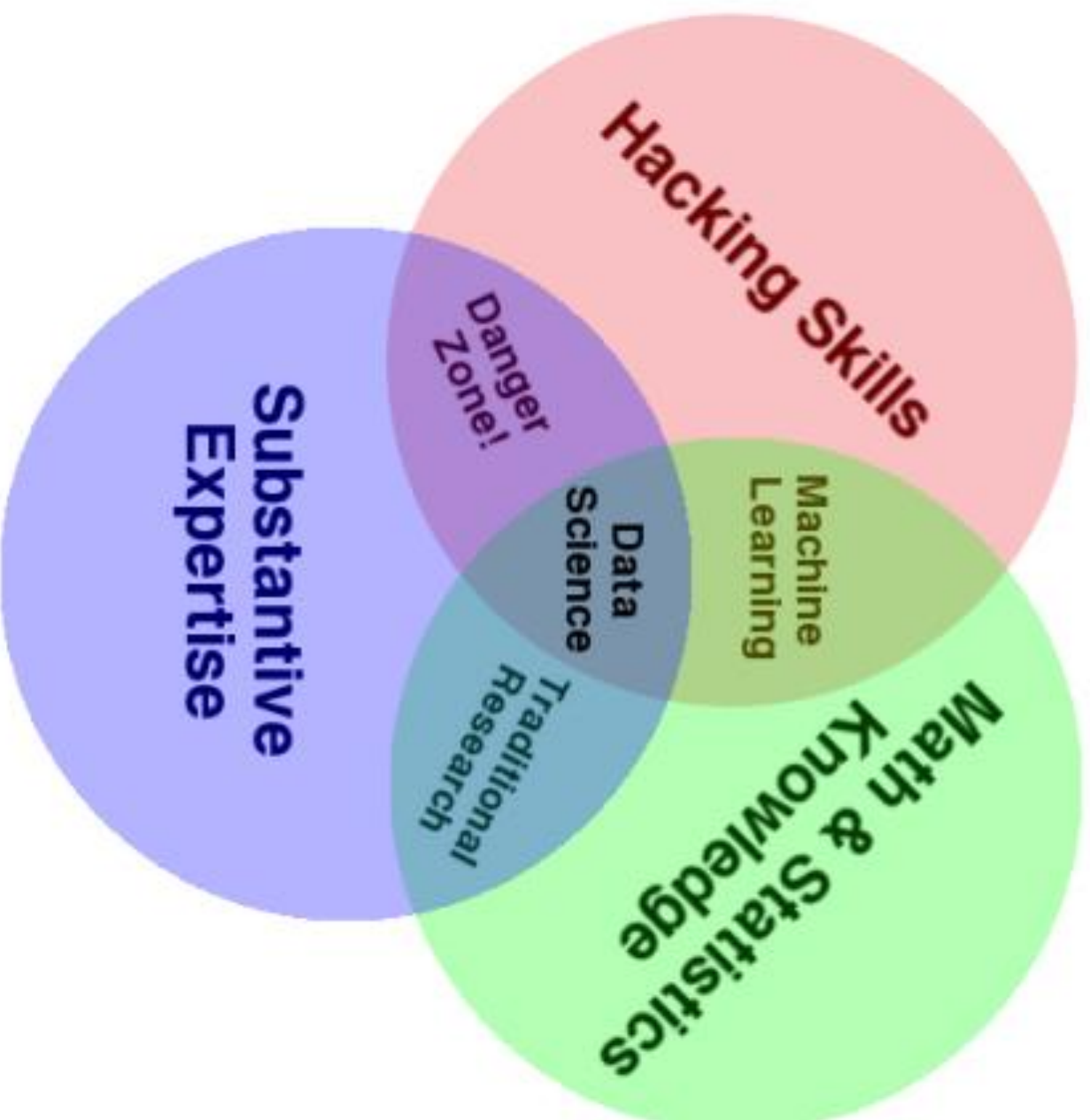
Cientista de Dados

# Data Scientist: *The Sexiest Job of the 21st Century*

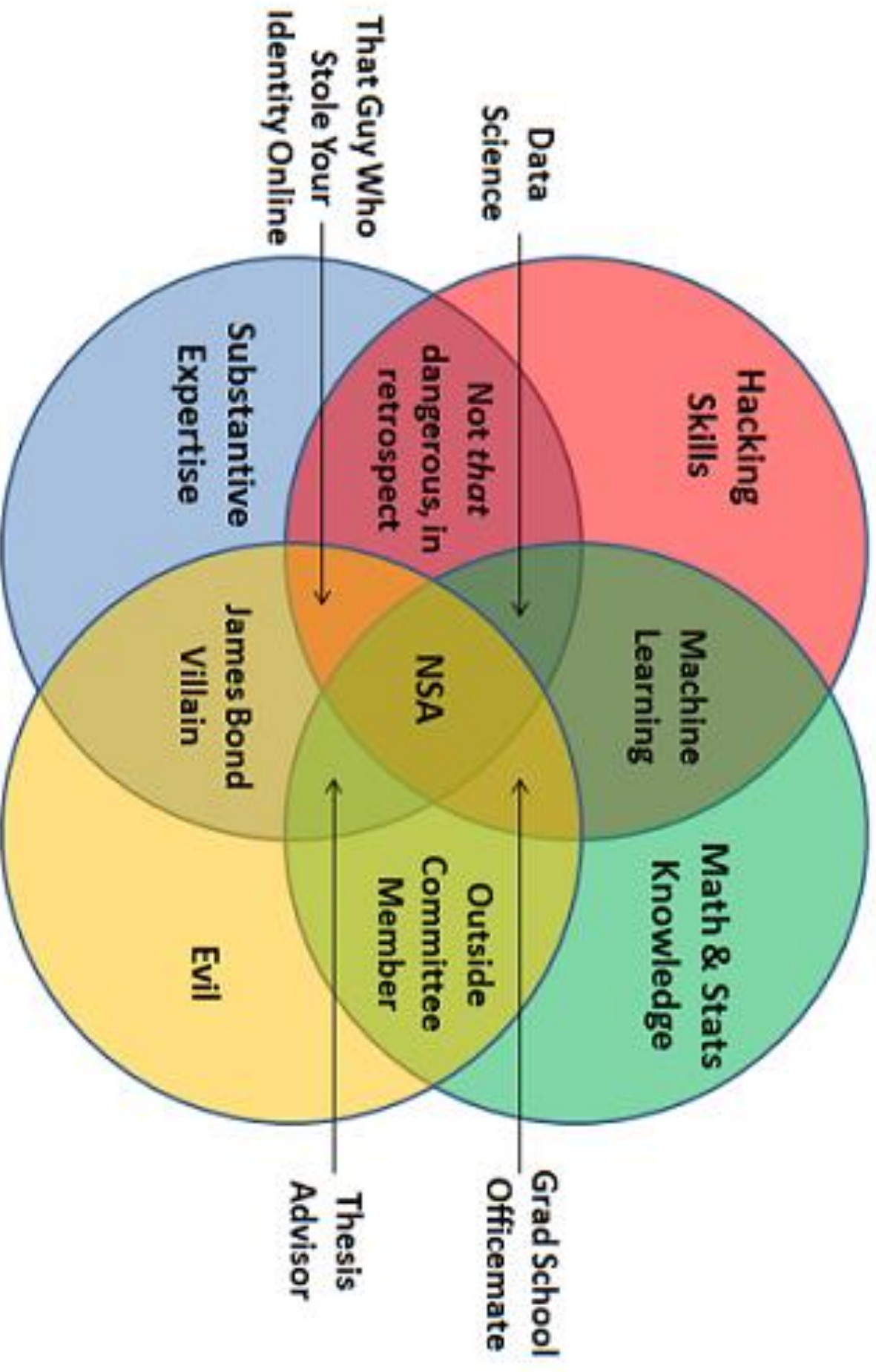
***Meet the people who  
can coax treasure out of  
messy, unstructured data.***  
by Thomas H. Davenport  
and D.J. Patil

**W**hen 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."

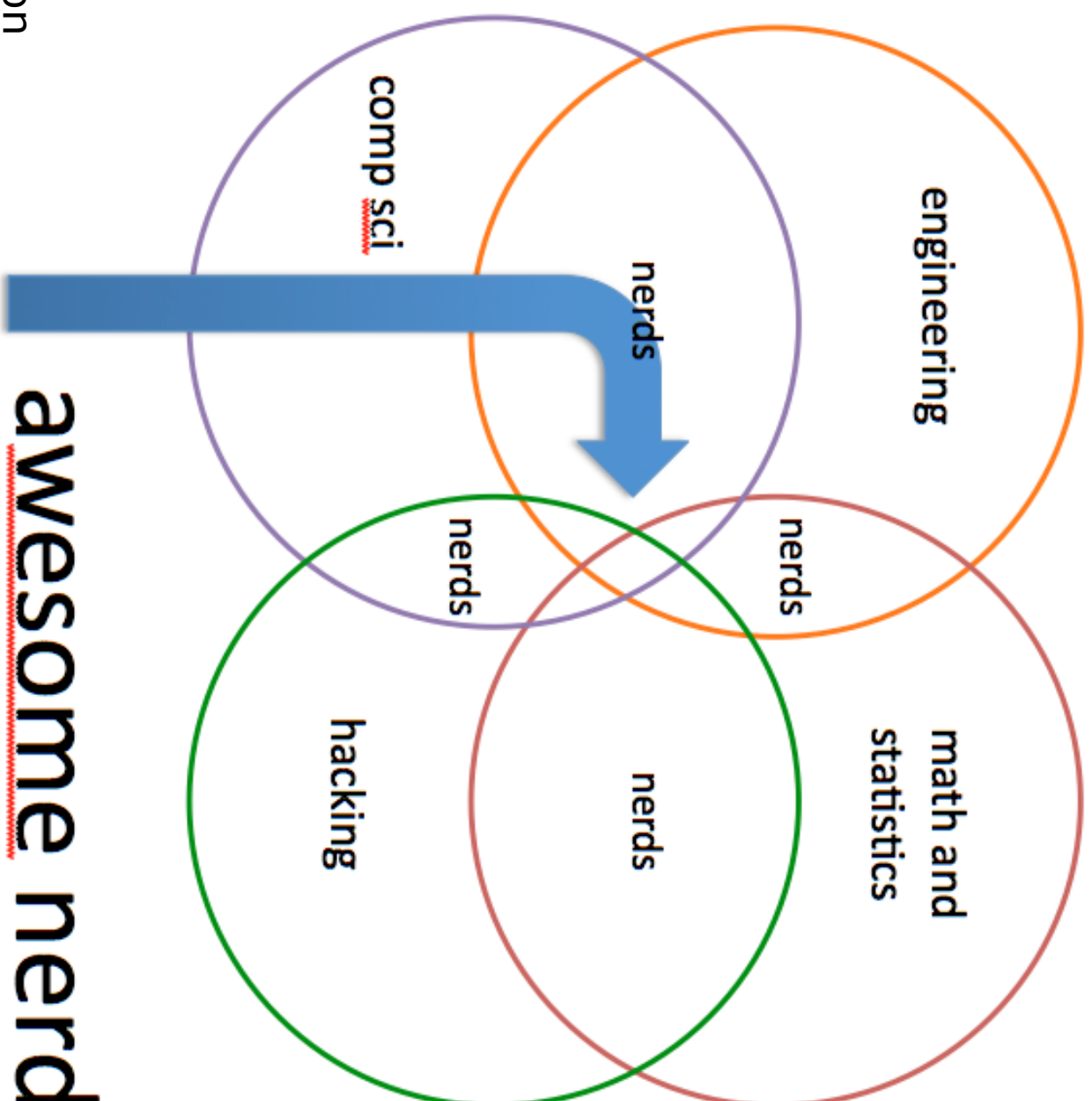
# Cientista de Dados: Visão Clássica



# Cientista de Dados: Visão NSA



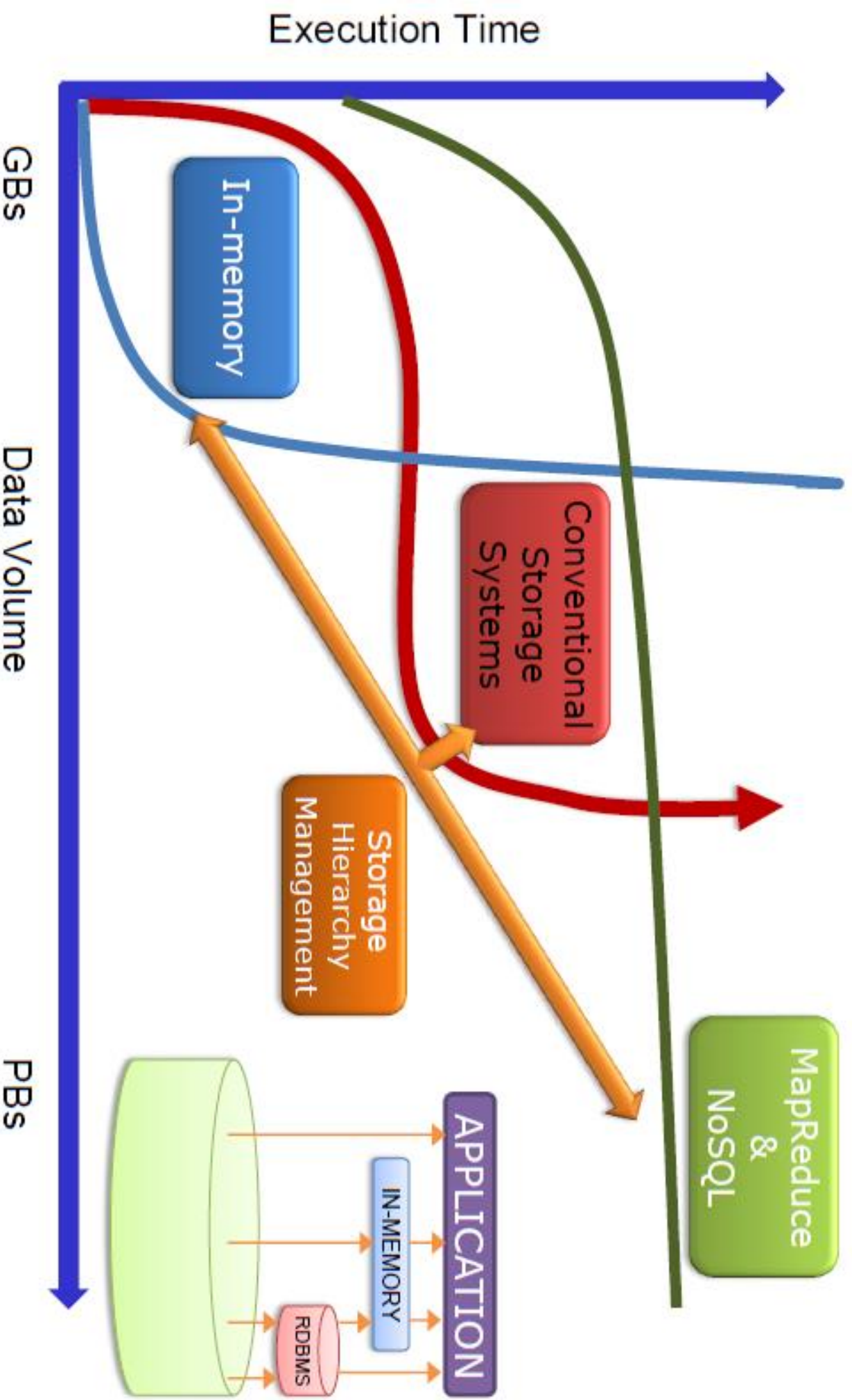
# Data scientists?



**awesome nerds**



# Novos Sistemas para Big Data



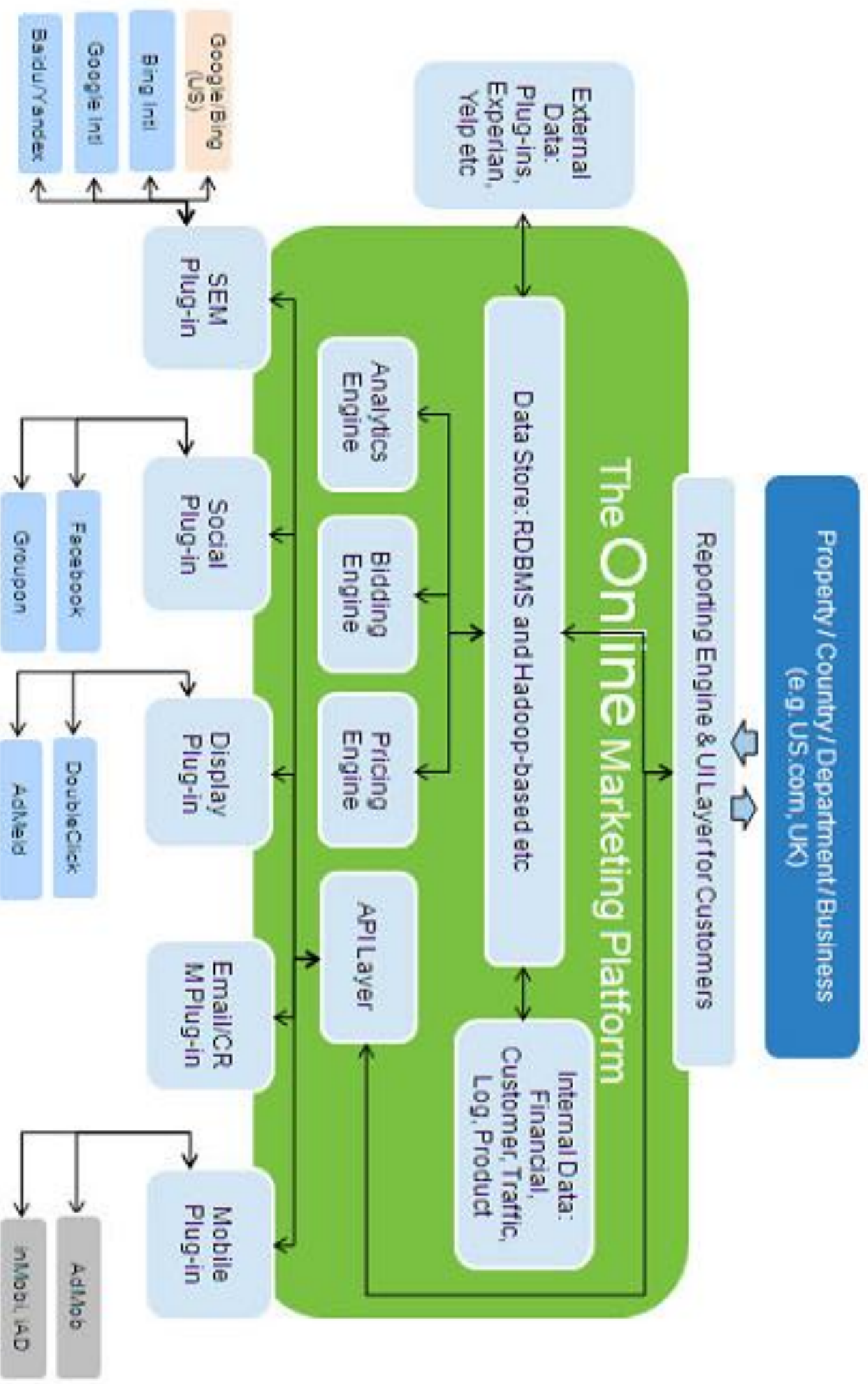
# Novos Sistemas para Big Data

- Armazenamento
  - SSD
- Processamento
  - MapReduce
- Gerenciamento
  - NoSQL, NewSQL
- Análise
  - Aprendizagem de máquina
  - Computação autônoma



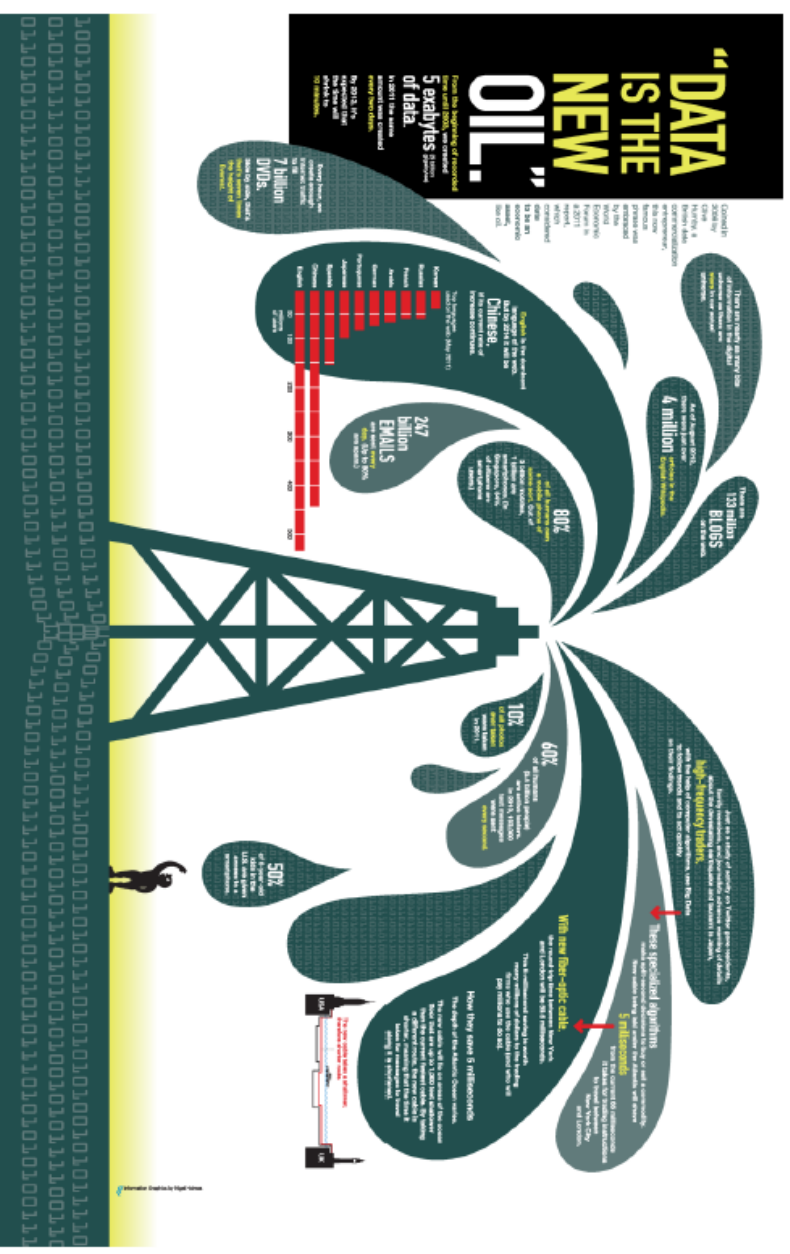


# Tech Architecture and Online Marketing Ecosystem



**e melhorando nossas vidas.**

# "Data is the new gold"



Fonte: ODI European Commission



Work Hard. Have Fun. Make History



amazon.com

Obrigado!

**Flávio R. C. Sousa**

**flaviosousa@ufc.br**

 **@flaviosousa**

**[www.lia.ufc.br/~flavio](http://www.lia.ufc.br/~flavio)**

