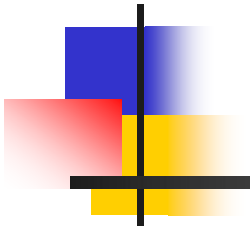
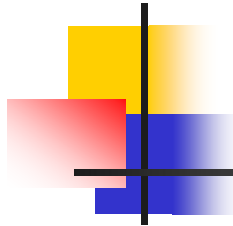


# Big Data e Ciência de Dados



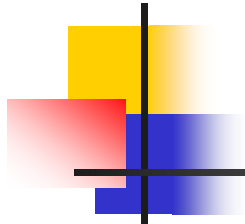
André C. P. L. F. de Carvalho  
Centro de Aprendizado de Máquina em Análise de Dados  
Universidade de São Paulo



# Tópicos

---

- Explosão de dados
- Big Data
- Ciência de Dados
- Crescimento da área
- Oportunidades na área
- Ciência de Dados para o bem
- Áreas de interesse



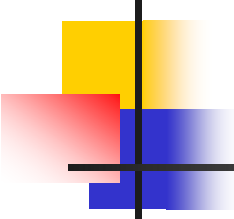
# Inteligência Artificial USP-SC

---

- 10 Docentes
- 9 Pós-doutorandos
- Cerca de 80 alunos de doutorado e mestrado
- 5 laboratórios
  - Analytics, BioCom, LABIC, LAR e NILC
- 2 Núcleos de Apoio à Pesquisa
  - Centro de Pesquisa AMDA



- Centro de Pesquisa de Aprendizado de Máquina em Análise de Dados
  - Interdisciplinar
  - Mais de 60 pesquisadores
    - Universidade de São Paulo
    - Centros de Pesquisa e Universidades brasileiras
    - Centros de Pesquisa e Universidades internacionais

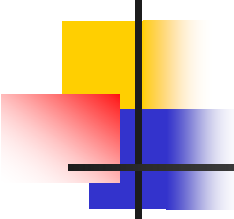


## Colaboradores nacionais do NAP-AMDA

---

- CATI
- Embrapa
- IBM
- INPE
- Instituto Jardim Botânico
- PUC-Rio
- UFABC
- UFC
- UFF
- UFMG
- UFPE
- UFPR
- UFRJ
- UFRN
- UFSCar
- UFTPR
- UFU
- UNB
- UNESP
- UNIFESP
- UNICAMP



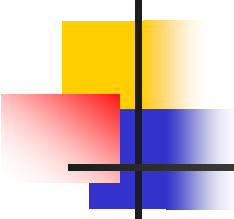


# Colaboradores internacionais do NAP-AMDA

---

- Auckland University of Technology
- Arizona State University
- Central Queensland University
- East China Normal University
- Humboldt University
- Hong Kong Baptist University
- Kyushu Institute of Technology
- National Research Council of Canada
- Nanyang Technological University
- Norwegian University of Science and Technology (NTNU)
- Rutgers University
- The Ohio State University
- Technical University of Ostrava (VSB)
- Universidade do Porto
- Universidad de Salamanca
- Universidad Nacional de Rosario
- University of Alberta
- University of California, Riverside
- University of Kent
- University of Leipzig
- University of Lyon 2
- University of Pittsburgh
- University of Texas, Austin
- University of Regensburg
- University of Surrey
- Uppsala University



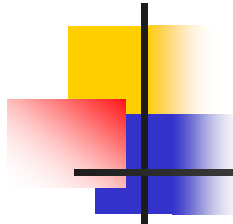


# Áreas de interesse do NAP AMDA

- Aprendizado de máquina
- Pré-processamento de dados
- Agrupamento de dados
- Classificação de dados
- *Data warehouse*
- Otimização bioinspirada
- Sistemas dinâmicos

Aplicado a

- Agricultura
- Bioinformática
- Diagnóstico de falhas
- Ecologia
- Engenharia
- Finanças
- Logística
- Medicina
- Petróleo e gás
- Redes sociais
- Robótica



# Introdução

---

- Sem perceber, as pessoas geram dados a todo momento
  - Aplica para um cartão de fidelidade
    - Empresa aérea, supermercado, ...
  - Faz uma compra com cartão de débito ou crédito
  - Navega na internet
  - Vai ao médico
- Esses dados são armazenados em computadores (pessoais ou nuvens)

# Introdução

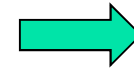
- Esses dados geralmente contém informações relevantes
  - Uma vez analisados, podem trazer vários benefícios
  - Análise de dados não é uma tarefa recente
  - Começou no Egito antigo
    - Recenseamentos periódicos eram realizados para a construção de pirâmides
    - Período dos faraós, em torno de 3200 AC



# Explosão de dados

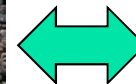
- Prática anterior

- Poucas empresas geravam dados
- Todo o resto (empresas e pessoas) consumia dados



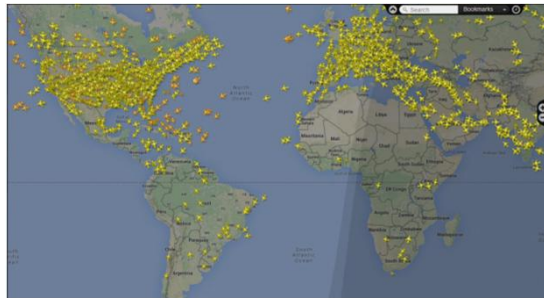
- Prática atual

- Todo mundo produz dados
- Todo mundo consome dados



# Explosão de dados

- Máquinas e pessoas continuamente geram, coletam e processam dados

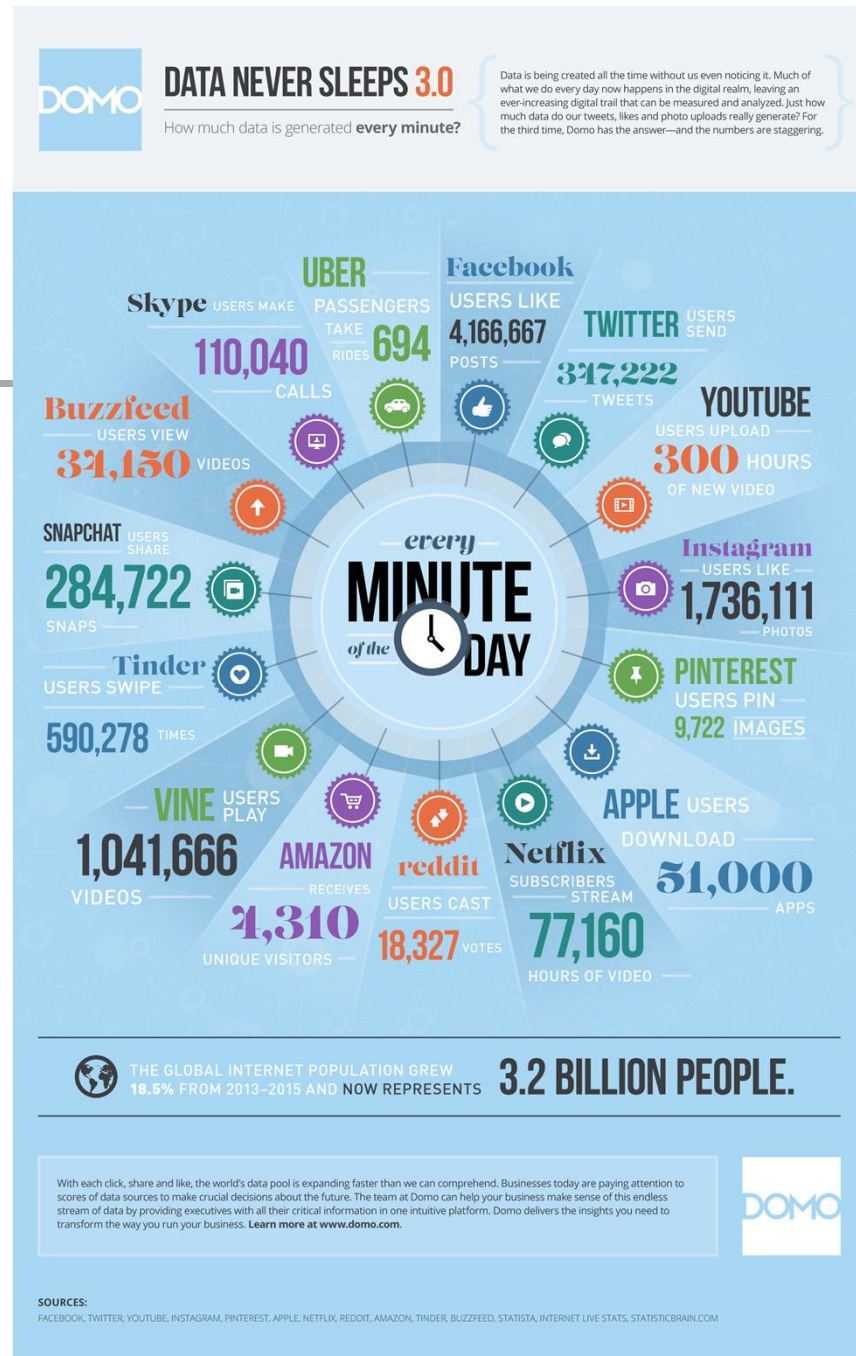


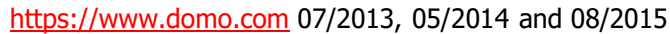
# Dados nunca dormem

Quantos dados são gerados a cada minuto

Origem: *Domo business management platform*

<https://www.domo.com>

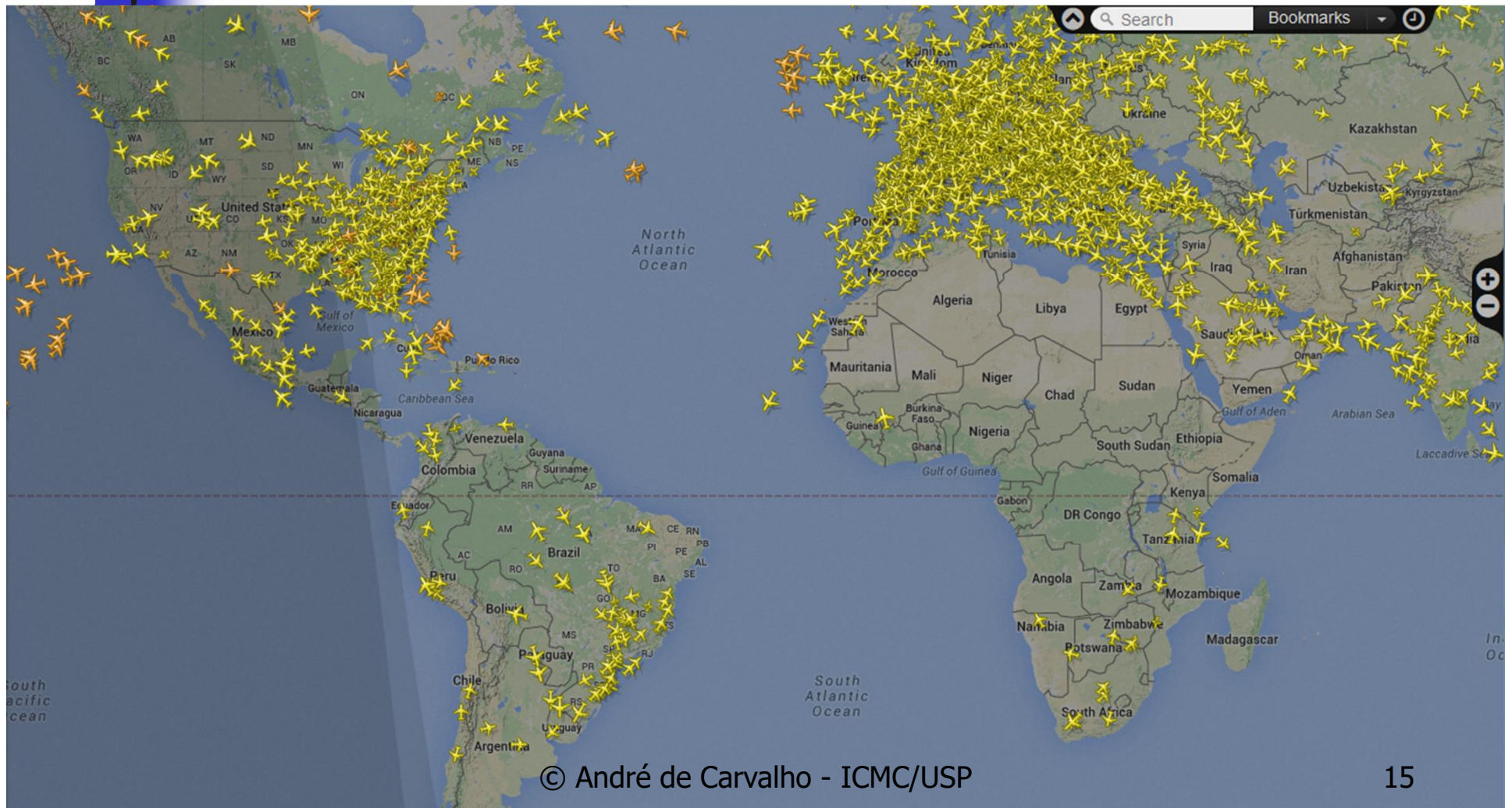




# Dados nunca dormem (manhã)

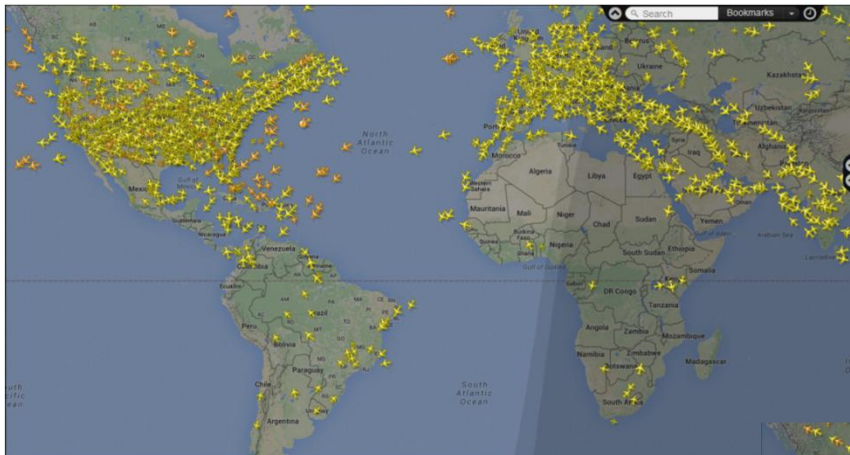


# Os dados nunca dormem (tarde)

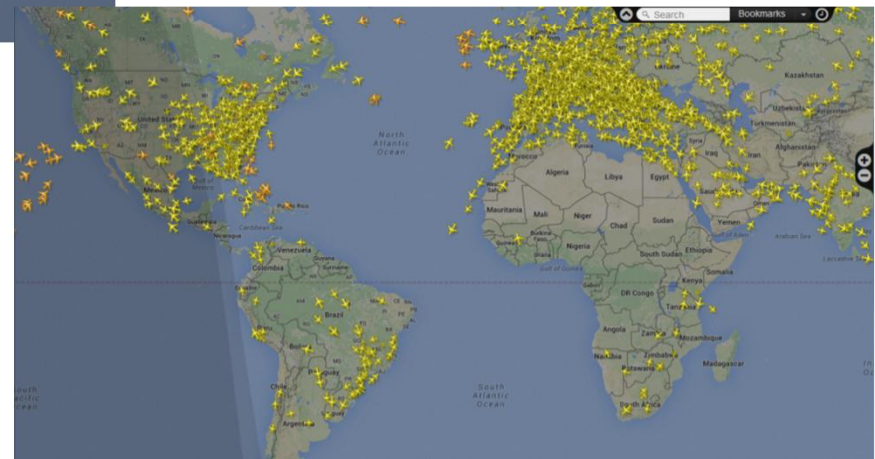


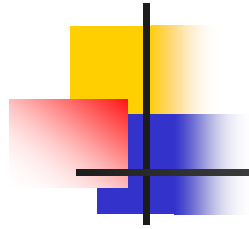
# Os dados nunca dormem

Dia 1 - Tarde



Dia 2 - Manhã



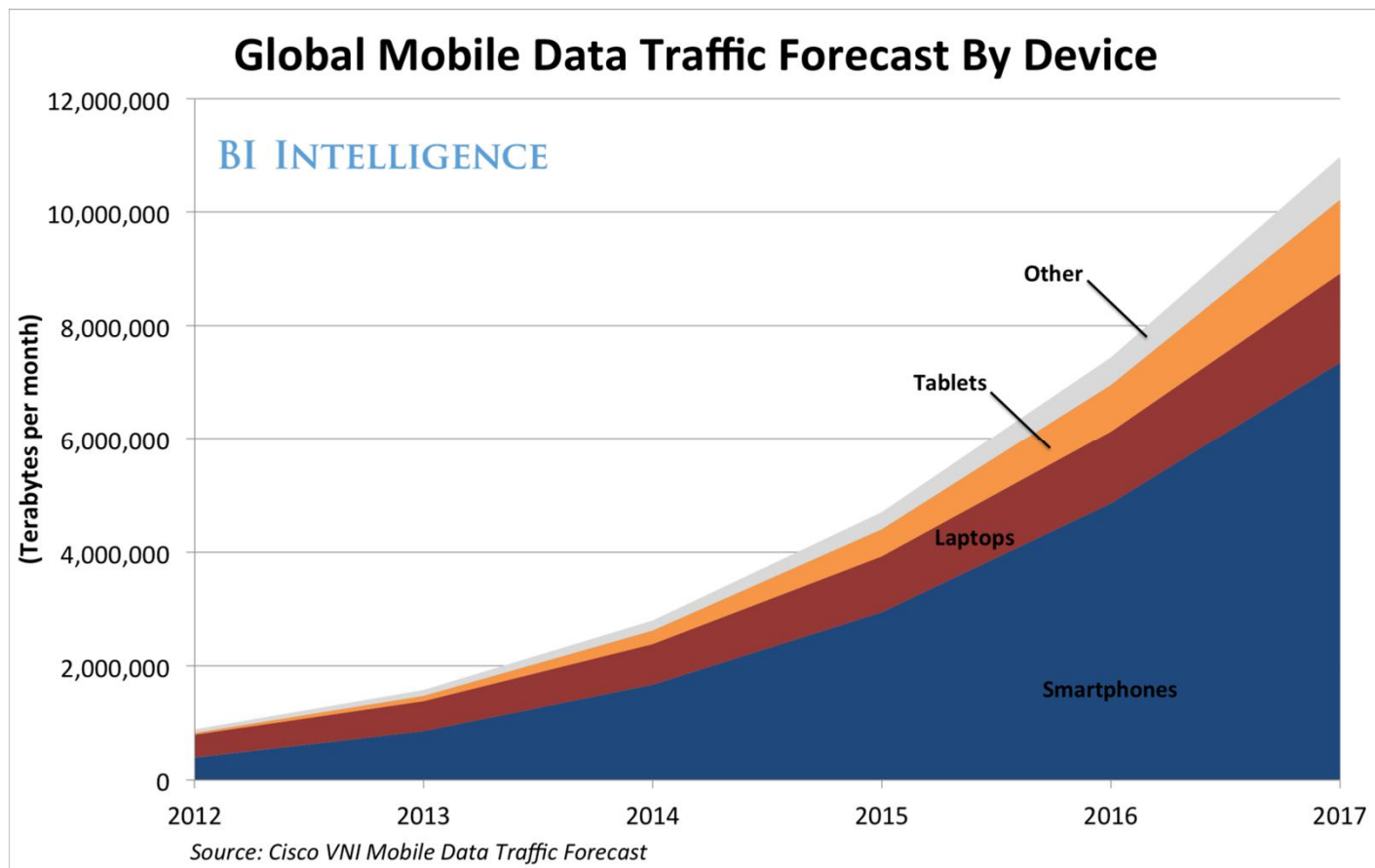


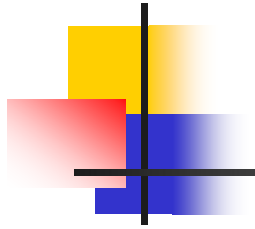
# Dados nunca dormem

---

<http://www.flightradar24.com>

# Tráfego de dados





# Dados de smartphones

Portugal

Onde estão os portugueses quando trabalham e vão de férias?

Distribuição da população  
(pessoas por km<sup>2</sup>)

Férias em Julho e Agosto

0 1200

Período de trabalho

0 7600



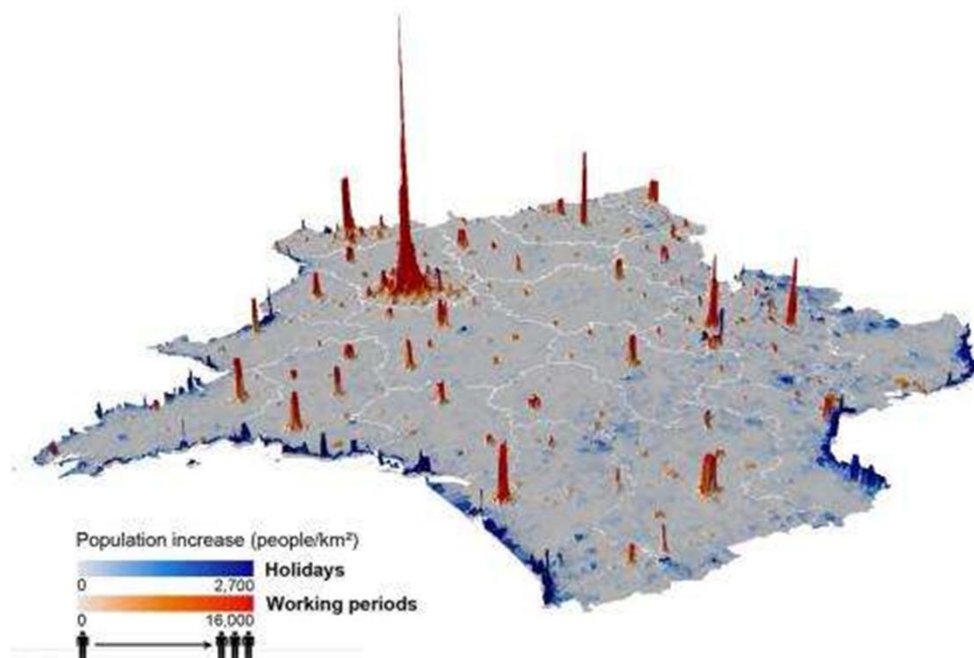
Fonte: Catherine Linard e PÚBLICO

PÚBLICO

<http://www.publico.pt/ciencia/noticia/telemoveis-fornecem-quase-em-tempo-real-mapas-da-densidade-populacional-portuguesa-1677020>

# Dados de smartphones

França

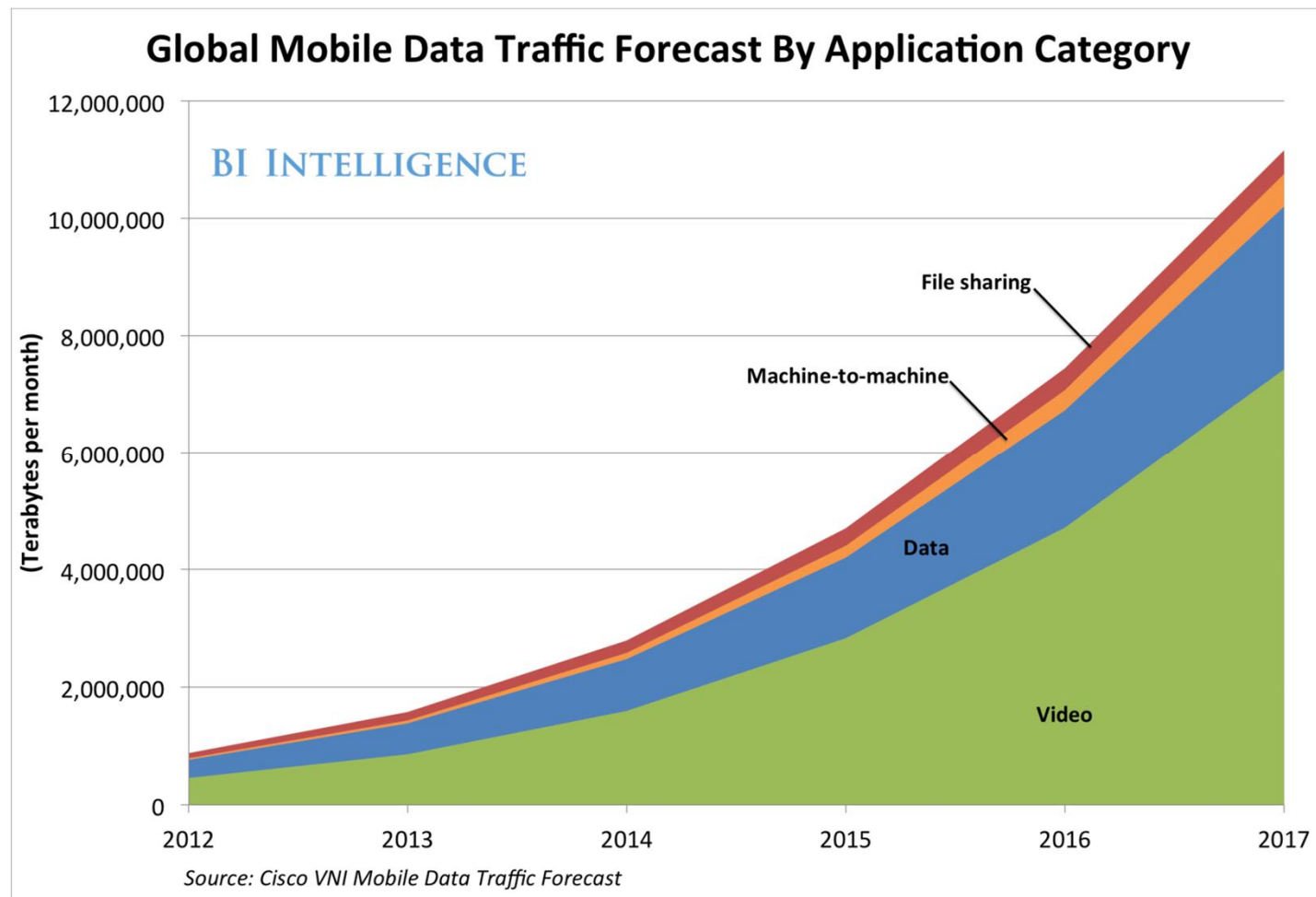


Population dynamics between the main holiday period (July and August) and working periods in France.

Credit: Catherine Linard

<http://phys.org/news/2014-10-cellphone-population-density.html#jCp>

# Dados nunca dormem

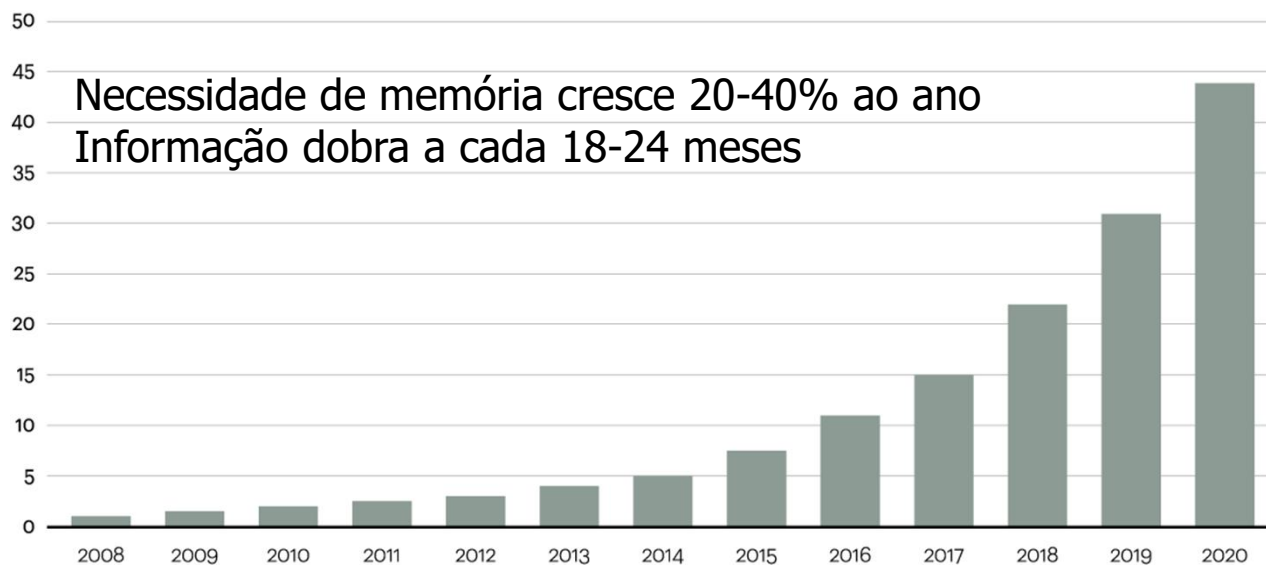


# Geração de dados

Figure 1

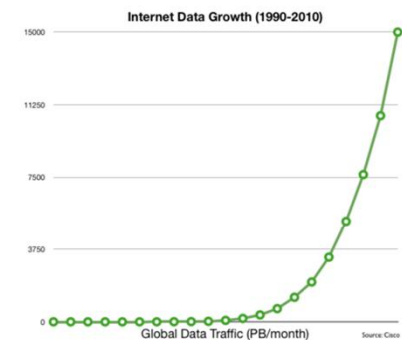
**Data is growing at a 40 percent compound annual rate, reaching nearly 45 ZB by 2020**

## Data in zettabytes (ZB)

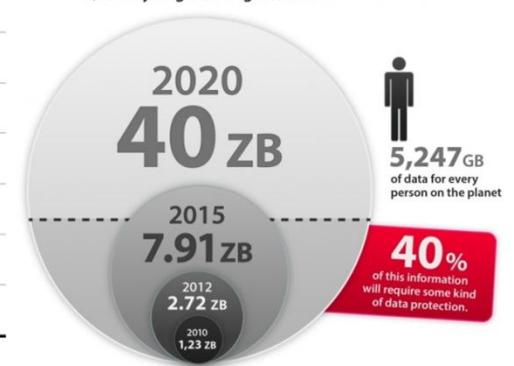


Source: Oracle, 2012

## Data is Growing Exponentially

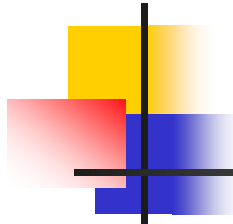


## Quantity of global digital data



# Inundação de dados

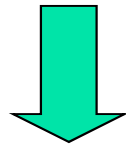




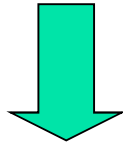
# Big Data

---

- Avanços recentes nas tecnologias para **aquisição, armazenamento e transmissão** de dados



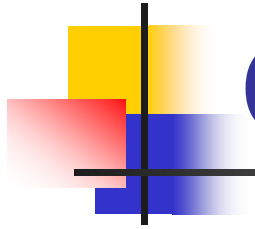
Bases de dados cada vez maiores



# Big Data





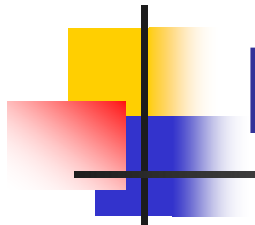


# O que é Big Data?

---

- Para alguns existe uma confusão entre os termos Big Data e Ciência de Dados
  - Confusão ocorre principalmente por interesses de mercado
  - Ciência de Dados procura criar modelos capazes de extrair padrões de sistemas complexos
    - E usar esses modelos em aplicações reais
  - Big Data procura dar suporte à coleta e ao gerenciamento de grandes quantidades de dados

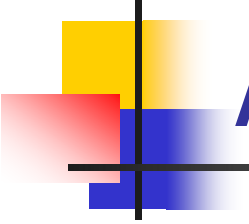
**Colecionar x Descobrir**



# Do que trata Big Data?

---

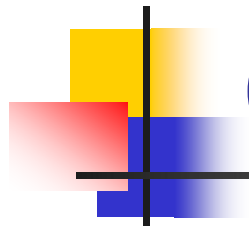
- Conjuntos de dados que são grandes demais para sistemas tradicionais de processamento de dados
- Requer novas tecnologias para:
  - Armazenamento
  - Processamento
  - Transmissão



# Armazenamento de dados

---

- Computadores atuais já vêm com 1 ou 2 terabyte (TB) de memória
- Cabem em 1 petabyte (1000 TB):
  - 20 milhões de arquivos de 4 gavetas cheios
  - 500 bilhões de páginas de texto
    - Metade do conteúdo de todas as bibliotecas acadêmicas americanas combinadas
  - 7 bilhões de fotos no *facebook*
  - 200 milhões de músicas



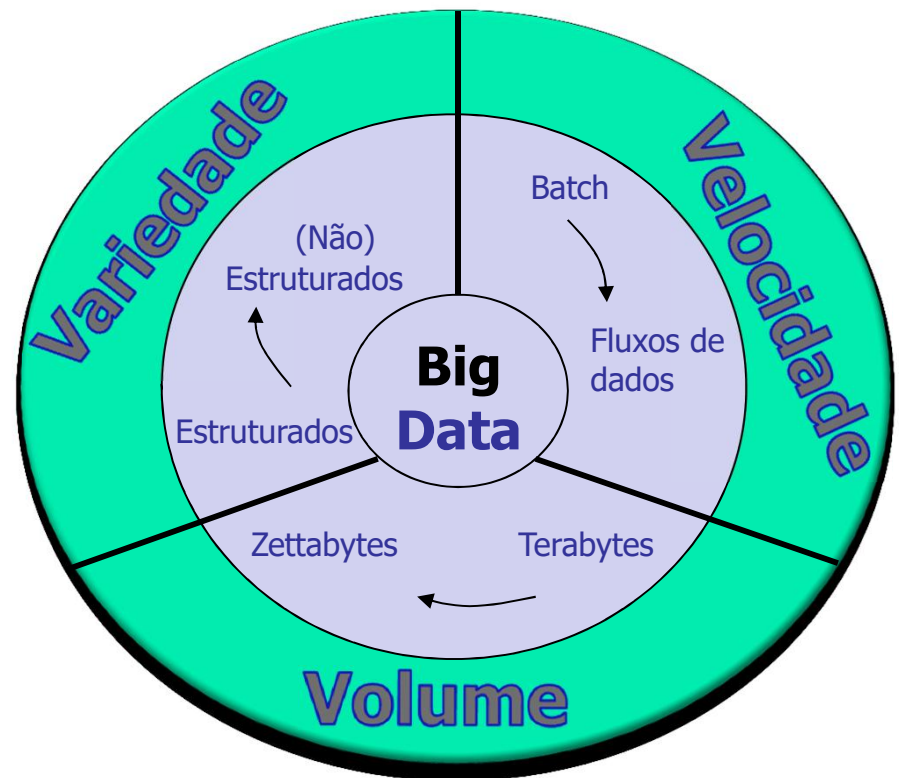
# Características de Big Data

---

- Grande **volume** de dados, gerados a uma grande **velocidade** e com uma grande **variedade** (3 Vs)
  - Volume: tanto de dados estruturados quanto de não estruturados
  - Variedade: vindos de fontes diversas e que precisam ser integrados
  - Velocidade: gerados em fluxos cada vez mais intensos

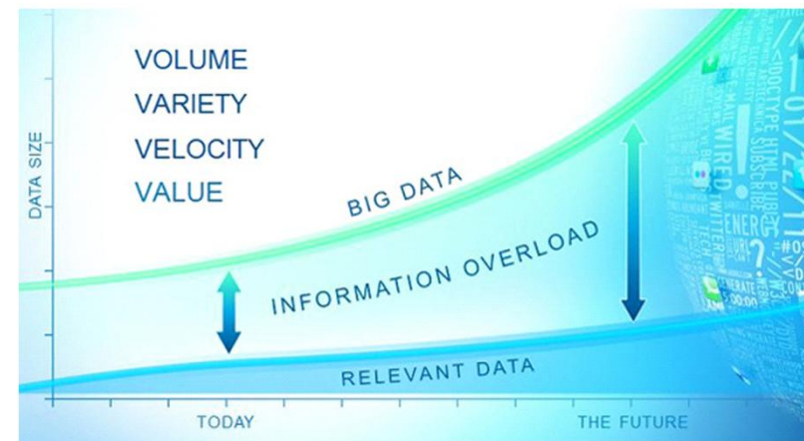
# Características de Big Data

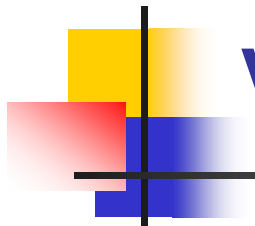
- Variedade:
  - Complexidade de dados
  - Dados com diferentes estruturas
    - Relacionais, Logs, textos
- Velocidade
  - Fluxos de dados em grande velocidade
- Volume
  - Escalas de Terabytes a Petabytes (1K TBs) a Zetabytes (1000K TBs)



# Quarto V

- Valor
  - Valor das informações contidas nos dados cresce rapidamente
    - Porém menos rapidamente que dados irrelevantes



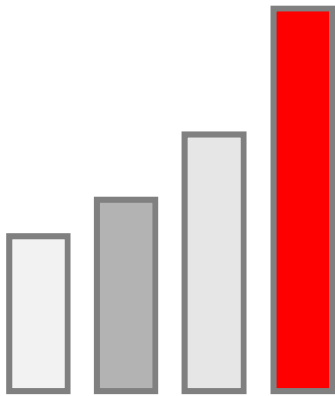


# Valor de Big Data

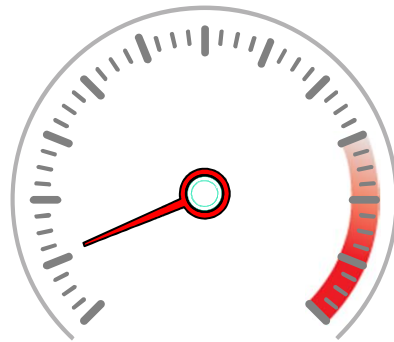
---

- Valor dos dados de 1 bilhão de perfis de usuários do facebook
  - Estimado em US\$ 32 bilhões (Nov 2012), US\$ 141 bilhões (maio 2014) e US\$ 300 (Julho 2015)
- Valor global de vendas relacionadas a aplicações de Big Data
  - Estimado em mais de US\$ 7 bilhões em 2012
  - Espera-se que cresça para mais de US\$ 100 bilhões em 2016

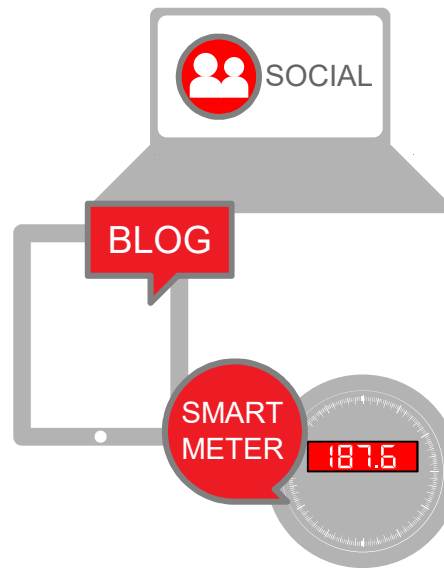
# Características de Big Data



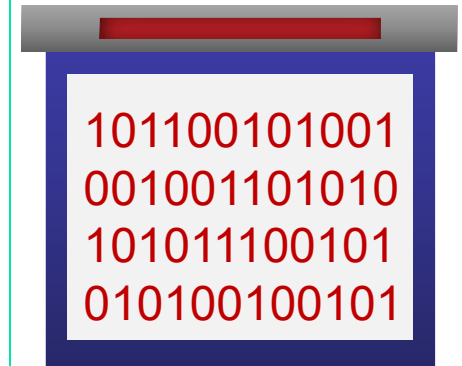
VOLUME



VELOCIDADE

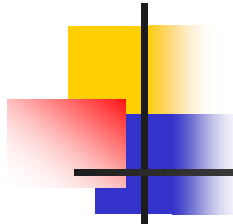


VARIEDADE



VALOR

Fonte: Oracle

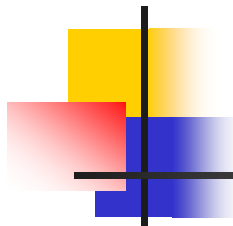


# Quinto V

---

- Veracidade

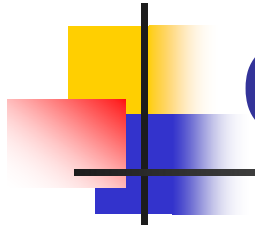
- 1 Em 3 tomadores de decisão não confia nas informações que usa para decidir
  - Como usar uma informação em que não confia?
- Um dos principais desafios de Big Data é mostrar que extrai informação confiável
  - Desafio aumenta com o crescimento na variedade e no número de fontes



# Análise de dados

---

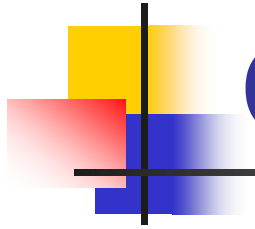
- Quantidade crescente de dados esta sendo gerada
  - Respeitando os 5 Vs de Big Data
  - Tecnologias de Big Data fornece meios para armazenar, processar e transmitir esses dados
  - Dados contêm conhecimento precioso, que precisa ser extraído
  - Ciência de Dados



# Ciência de Dados

---

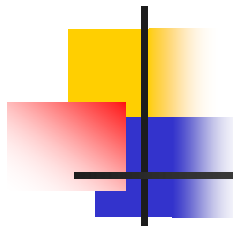
- Várias definições
  - Estuda princípios, métodos e sistemas computacionais para extrair conhecimento de dados
- Pergunta chave da área:
  - Como encontrar de forma eficiente conhecimento (padrões) em (grandes) conjuntos (fluxos) de dados



# Ciência de Dados

---

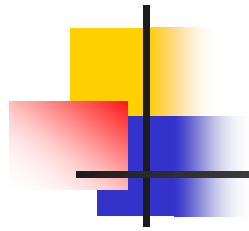
- Teorias e princípios gerais ainda estão sendo formulados
- Também chamada de *Analytics*
- Área basicamente experimental
- Mas a mudança esta sendo rápida
  - Inclusive com nova forma de abordar teoria da computação



# Ciência de Dados

---

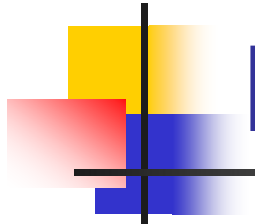




# Etapas de Ciência de Dados

---

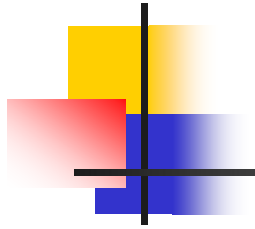
- Planejamento de experimentos
- Pré-processamento
- Modelagem
- Avaliação



# Planejamento de experimento

---

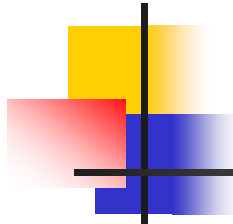
- Entender o problema a ser resolvido
- Definir:
  - Técnicas de pré-processamento
  - Técnicas de modelagem
  - Medidas de avaliação
  - Meta para cada etapa
  - Tempo a ser alocado a cada etapa



# Pré-processamento

---

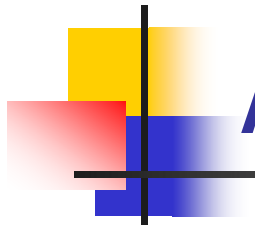
- Em geral, dados não foram gerados para uso em Ciência de Dados
  - Produzidos para outros propósitos
  - Frequentemente apresentam problemas
- Etapa de modelagem precisam geralmente de dados “limpos”
  - Entra lixo, sai lixo
  - Problemas nos dados precisam ser detectados e corrigidos



# Modelagem

---

- Extrai modelos capazes de extrair conhecimento dos dados
  - Mineração de Dados (MD)
  - Analítica
- Várias técnicas foram criadas para extrair modelos durante a MD
  - Maioria dessas técnicas é baseada em Aprendizado de Máquina (AM)

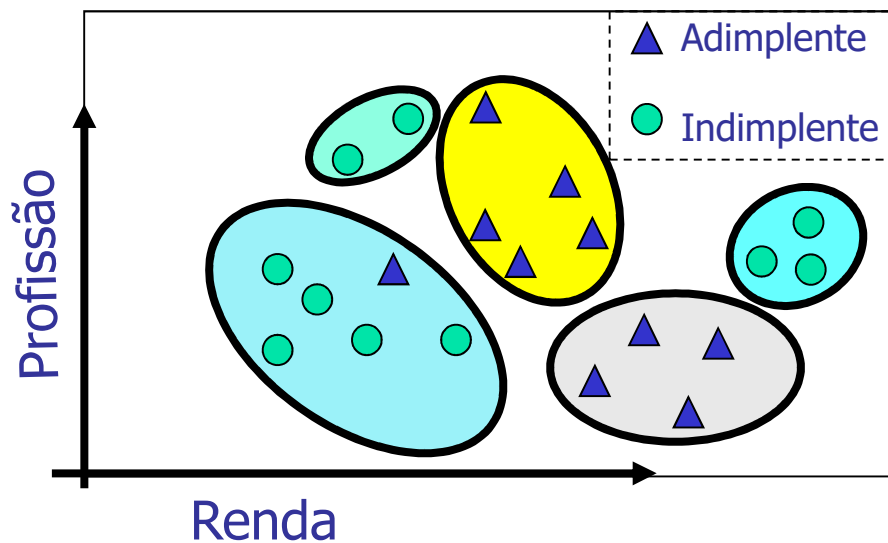


# Aprendizado de Máquina

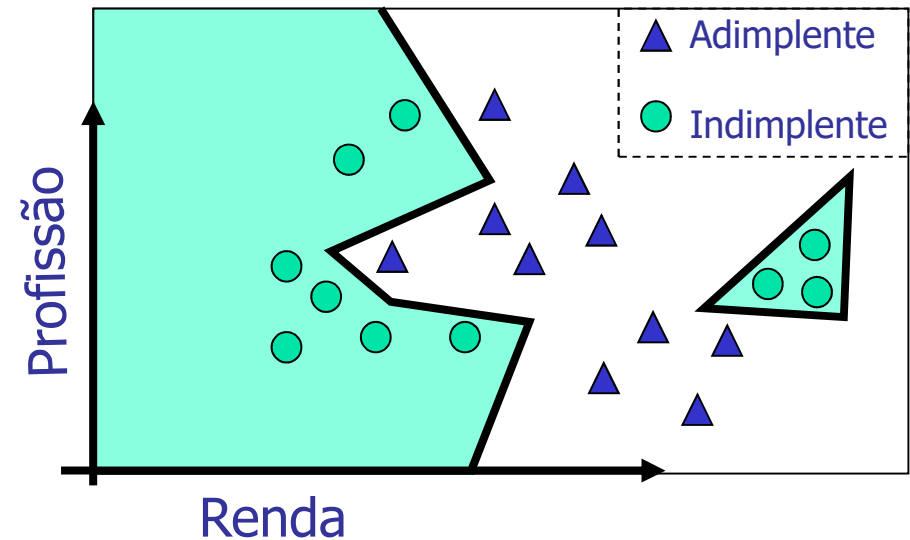
---

- Investiga técnicas capazes de aprender a resolver problemas
  - De forma automática, sem intervenção humana
- Bem sucedido em vários problemas reais de modelagem
  - Descritivos
  - Preditivos

# Modelagem por AM



Descritivo  
Agrupamento



Preditivo  
Classificação

# Algoritmos de classificação

## Indução

conjunto de dados

|          |          |     |          |          |
|----------|----------|-----|----------|----------|
| $x_{11}$ | $x_{12}$ | ... | $x_{1m}$ | $y_1$    |
| $x_{21}$ | $x_{22}$ | ... | $x_{2m}$ | $y_2$    |
| $\vdots$ | $\vdots$ |     | $\vdots$ | $\vdots$ |
| $x_{n1}$ | $x_{n2}$ | ... | $x_{nm}$ | $y_n$    |

Algoritmo



Treinamento

$f(\mathbf{x})$

Modelo para  
Classificação

## Dedução

$\mathbf{x}_i$

|          |          |     |          |
|----------|----------|-----|----------|
| $x_{i1}$ | $x_{i2}$ | ... | $x_{im}$ |
|----------|----------|-----|----------|

novo dado

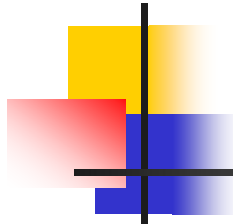


modelo  $f(\mathbf{x})$

Teste

*Classe ou probabilidade*

previsão



# Avaliação

---

- Interpretação do conhecimento extraído
  - Possível retorno a qualquer uma das etapas anteriores
- Validação de conhecimento extraído
  - Importante consulta a um especialista
- Análise estatística
- Ferramentas de visualização fornece um suporte importante

# Oportunidades

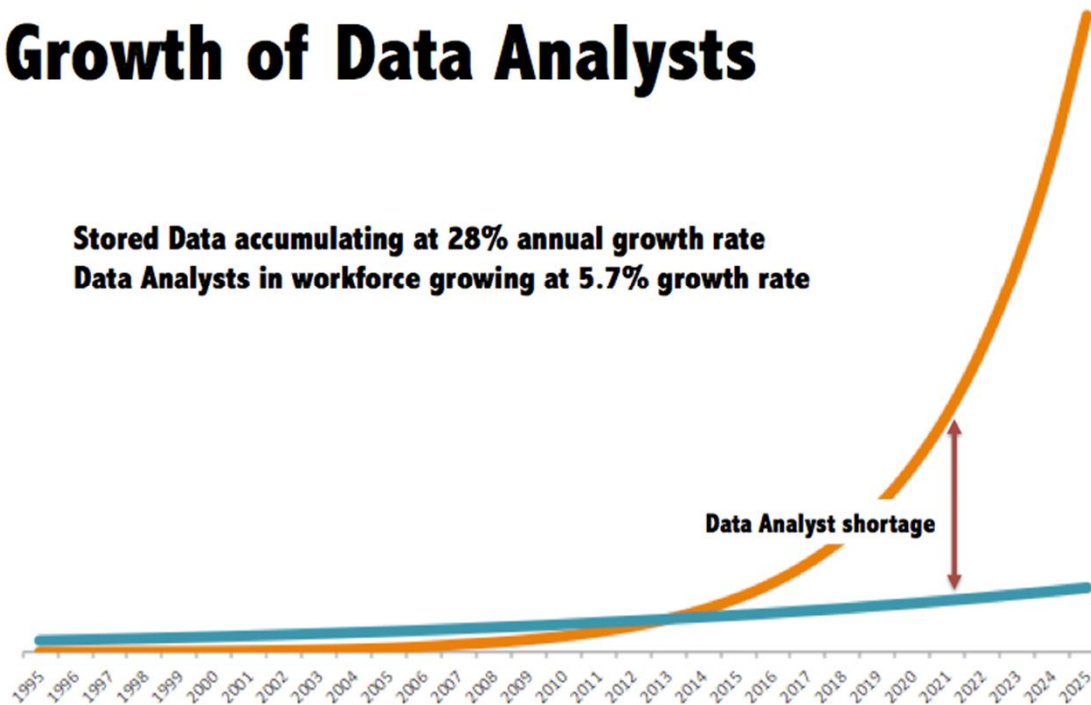
- “Data Scientist: The Sexiest Job of the 21st Century”
  - *Harvard Business Review*, Outubro de 2012
- Ajuda tomadores de decisão a mudar análise subjetiva para análise baseada em dados



# Falta de Cientistas de Dados

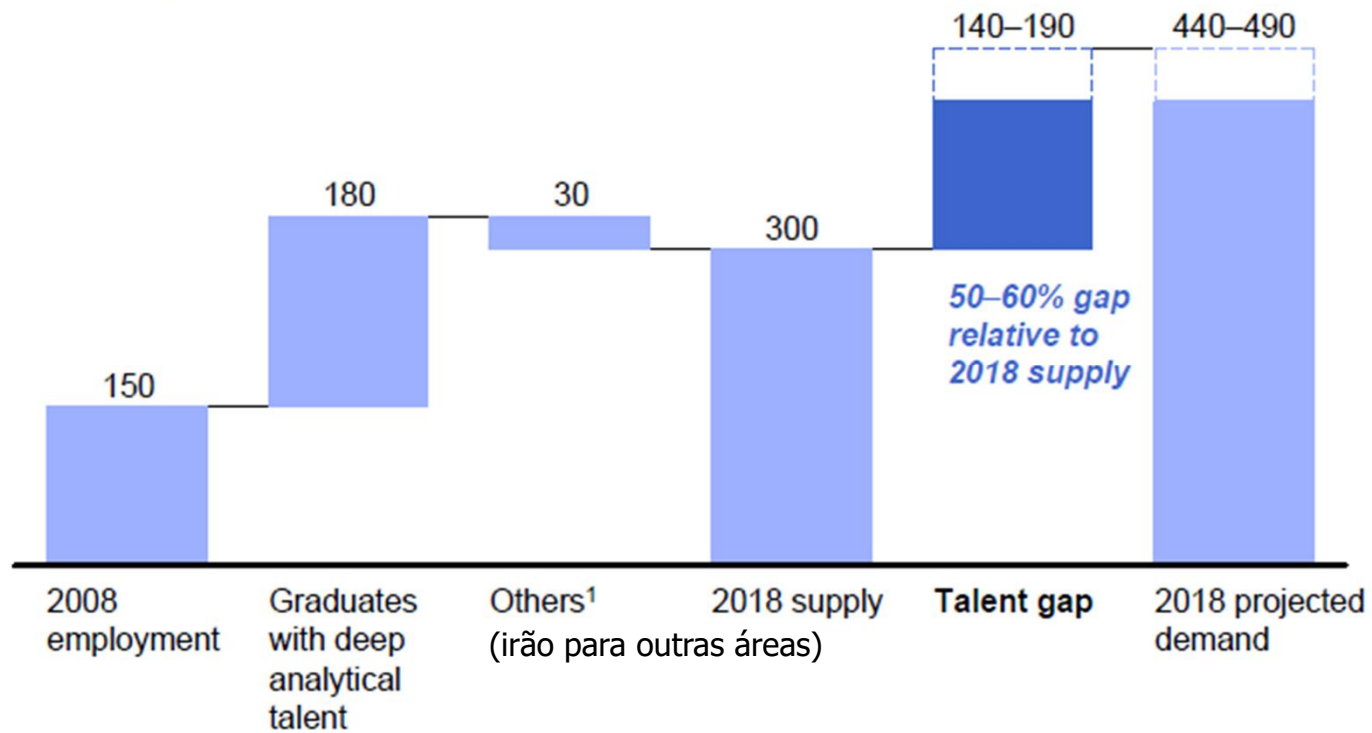
## Growth of Data vs. Growth of Data Analysts

**Stored Data accumulating at 28% annual growth rate**  
**Data Analysts in workforce growing at 5.7% growth rate**



Fonte: [www.delphianalytics.net](http://www.delphianalytics.net)

# Falta de cientistas de dados



## **Haverá falta de especialistas em data science.**

Em 2018, faltarão nos EUA 140.000 a 190.000 analistas com capacidade para análises detalhadas de dados.

# Falta de Cientistas de Dados

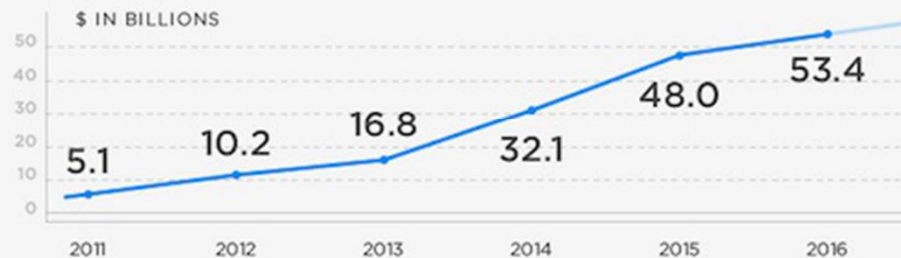


# Por que essa necessidade?

## EXPLAINING THE SUDDEN NEED FOR DATA SCIENTISTS

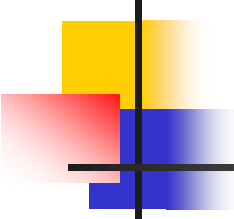
These scientists don't just happen to be getting far more job offers without reason. Today's modern business needs to manage far more data than ever before, and few have the talent on staff for the job.

**Projections indicate that the market will experience meteoric growth in the next several years.**



The Big Data  
Market Forecast

**Conclusion:** With so much activity going on in the big data space and new data touch points being measured every day, there will be an increasing need for data-driven individuals within organizations to make sense of it all. Is that data-savvy person you?



# Quem esta contratando CD

---

- Apple
- Booking.com
- Disney
- Google
- Greepeace
- Mercedes-Benz
- Red Bull F1

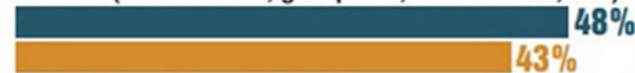
# Que aplicações se interessam

## Factors Driving Interest in Big Data Analysis

What data sources or challenges are driving, or would drive, your organization's interest in doing big data analysis?

■ 2015 ■ 2014

Finding correlations across multiple, disparate data sources (clickstreams, geospatial, transactions, etc.)



Predicting customer behavior



Predicting product or service sales



Predicting fraud or financial risk



Analyzing social network comments for consumer sentiment



Analyzing high-scale machine data from sensors, web logs, etc.



Identifying computer security risks



Analyzing web clickstreams



Other



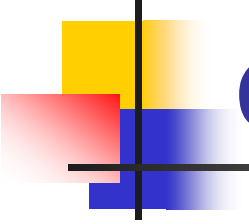
Big data analytics is not of interest to my organization



[blog.raremile.com](http://blog.raremile.com)

# Mercado profissional

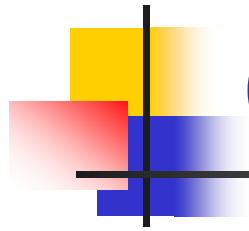




# Centros e Institutos de Ciência de Dados

---

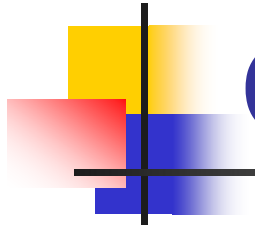
- Columbia University, EUA
- Eindhoven University of Technology, Holanda
- Imperial College, Reino Unido
- Leiden University, Holanda
- New York University, EUA
- Tilburg University, Alemanha
- University of Edinburgh, Reino Unido
- University of Massachusetts at Amherst, EUA



# Cursos em Universidades

---

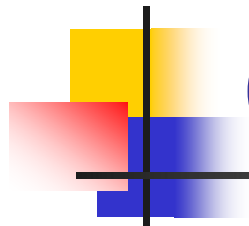
- Graduações, mestrados e doutorados
- Graduações
  - Eindhoven University of Technology, Holanda
  - Tilburg University, Alemanha
  - University of Nottingham, Reino Unido
  - University of Warwick, Reino Unido
  - University of Essex, Reino Unido



# Ciência de Dados para o Bem

---

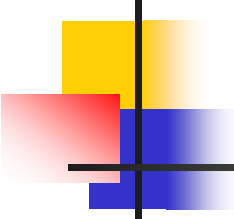
- Movimento sem fins lucrativos
  - Trazer benefícios sociais para as pessoas e comunidades
  - Alguns programas são adotados por empresas
- Como isso ocorre?
  - Reuniões
  - Eventos
  - Estágios acadêmicos
  - Redes sociais



# Ciência de Dados para o Bem

---

- Abordagens existentes:
  - Uso de dados (abertos) para resolver problemas de defesa civil
    - Normalmente, desenvolvimento de aplicativos móveis / web
  - Uso de Ciência de Dados para resolver problemas sociais
    - Principalmente buscando suporte de cientistas de dados



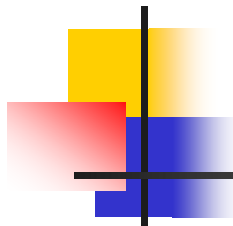
# Ciência de Dados para o Bem

---

- Abordagens existentes:
  - Democratização de dados
    - Permitir que qualquer pessoa tenha acesso a dados públicos
    - Primeiro Cientista Chefe de Dados foi nomeado em 2015 pelo presidente dos EUA
      - *First U.S. Chief Data Scientist*
      - Estimular pesquisas e desenvolvimento tecnológico em medicina de precisão, dados abertos, decisão apoiada por dados

- [illegible]

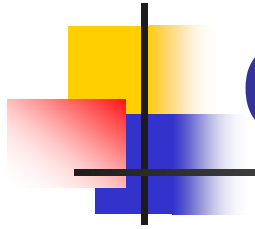
<http://www.kdnuggets.com/2014/07/data-for-good-data-driven-projects-social-good.html>



# Ciência de Dados para o Bem

---

- Traz benefícios sociais para pessoas e comunidades
  - Bons serviços de saúde para todos
  - Desenvolvimento econômico de países pobres
  - Educação pública de qualidade
  - Energia limpa e barata
  - Melhor exercício da cidadania
  - Proteção ambiental
  - Meios de transportes mais seguros, rápidos e limpos



# Conclusão

---

- Explosão de dados
- Big Data
- Ciência de Dados
- Crescimento da área
- Oportunidades na área
- Ciência de Dados para o bem
- Áreas de interesse