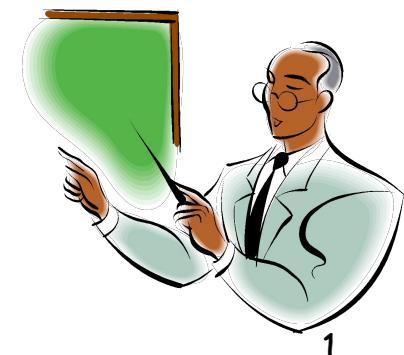


DCA- 0131: Ciência de Dados

Luiz Affonso Guedes - affonso@dca.ufrn.br



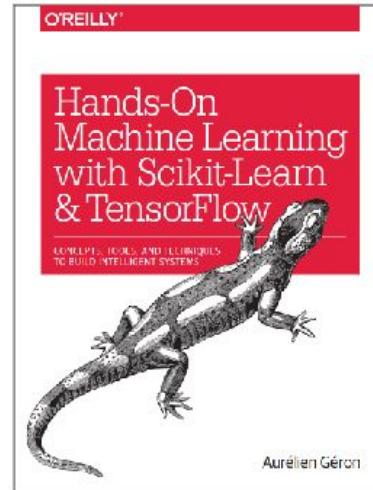
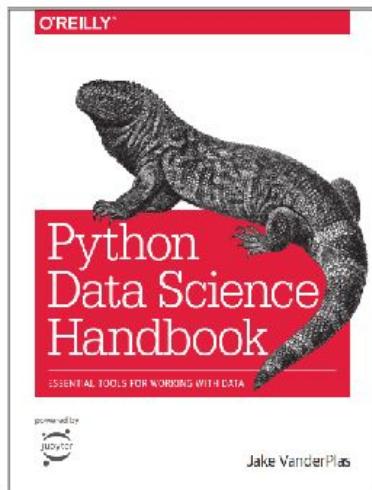
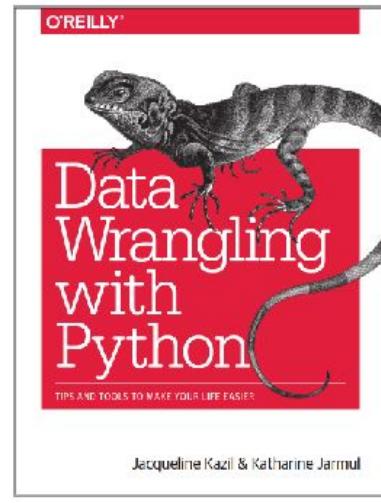
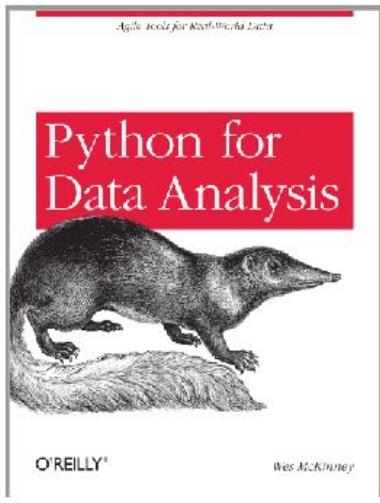
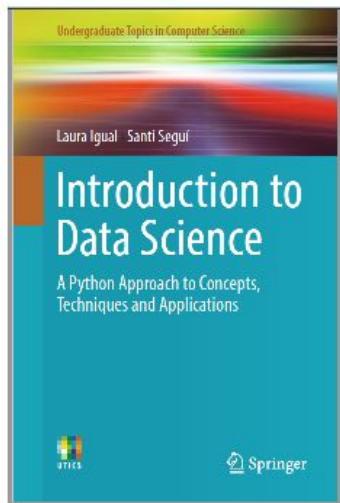
Ementa

- Introdução à Ciência de Dados
 - Conceitos fundamentais, importância, aplicações e ciclo de vida de dados.
- Tratamentos de dados
 - Captura, limpeza, filtragens, imputação e armazenamento de dados.
 - Pacote Pandas do Python
- Análise Exploratória de Dados
 - Conceitos de estatísticas e visualização de dados.
 - Pacotes Scipy e Seaborn do Python
- Processamento de Dados
 - Técnicas de agrupamento e classificação de dados.
 - Pacote Scikit-learn do Python
- Interpretação de Dados
 - Explainable AI (**XAI**)
 - Pacotes Shap e XAI do Python

Metodologia Adotada

- Aulas Presenciais
 - Apresentação do conteúdo e desenvolvimento de programas no ambiente Colab-Google.
 - Turma 1 - Terças e quintas-feiras: 35T34
 - Turma 2 - Sextas-feiras de manhã - 6M1234.
- Plantão para tirar dúvidas: na sala do professor ou via Google Meets
 - Apoio na resolução de exercícios no Colab-Google.
- Material de Apoio
 - Notebooks em Python
 - Livros
 - Cursos da Internet

Bibliografia Sugerida



Avaliação

- Primeira Avaliação:
 - Aplicação de Ciência de Dados – Visualização de dados em notebook
- Segunda Avaliação
 - Aplicação de Ciência de Dados com Visualização de dados na Web
- Terceira Avaliação
 - Aplicação com técnicas de Machine Learning com Visualização de dados na Web

Pré-requisitos

- Conceitos de Probabilidade
- Programação básica.
- Não ter medo de programar
- Não ter medo de aprender coisas novas
- Motivação

Abordagem do Curso

- Metodologia baseada em resolução de problemas
 - "Mão-na-massa"
- Uso do Ambiente Anaconda – Jupyter ou Colab-Google
- Uso da Linguagem Python

Abordagem do Curso

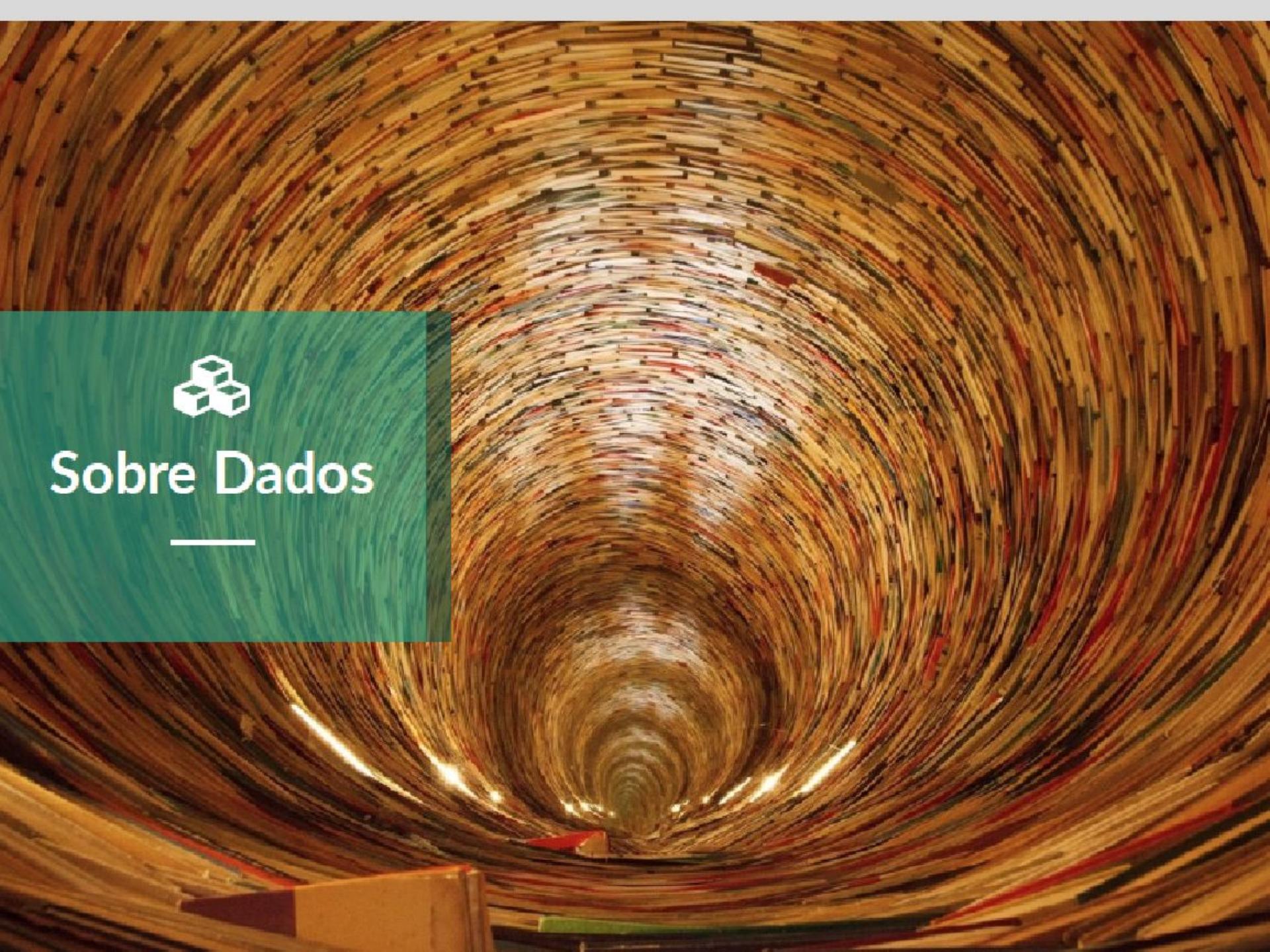


Motivação

- Ampla gama de aplicações.
- Conciliação entre teoria e aplicação.
- Dados estão em todos os lugares. Só falta a gente analisá-los.
- Modelagem científica de dados é a base da ciência quantitativa atual.
 - Aprendizagem de máquinas.
 - Bioinformática
 - Epidemiologia
 - Reconhecimento de imagens
 - Etc.

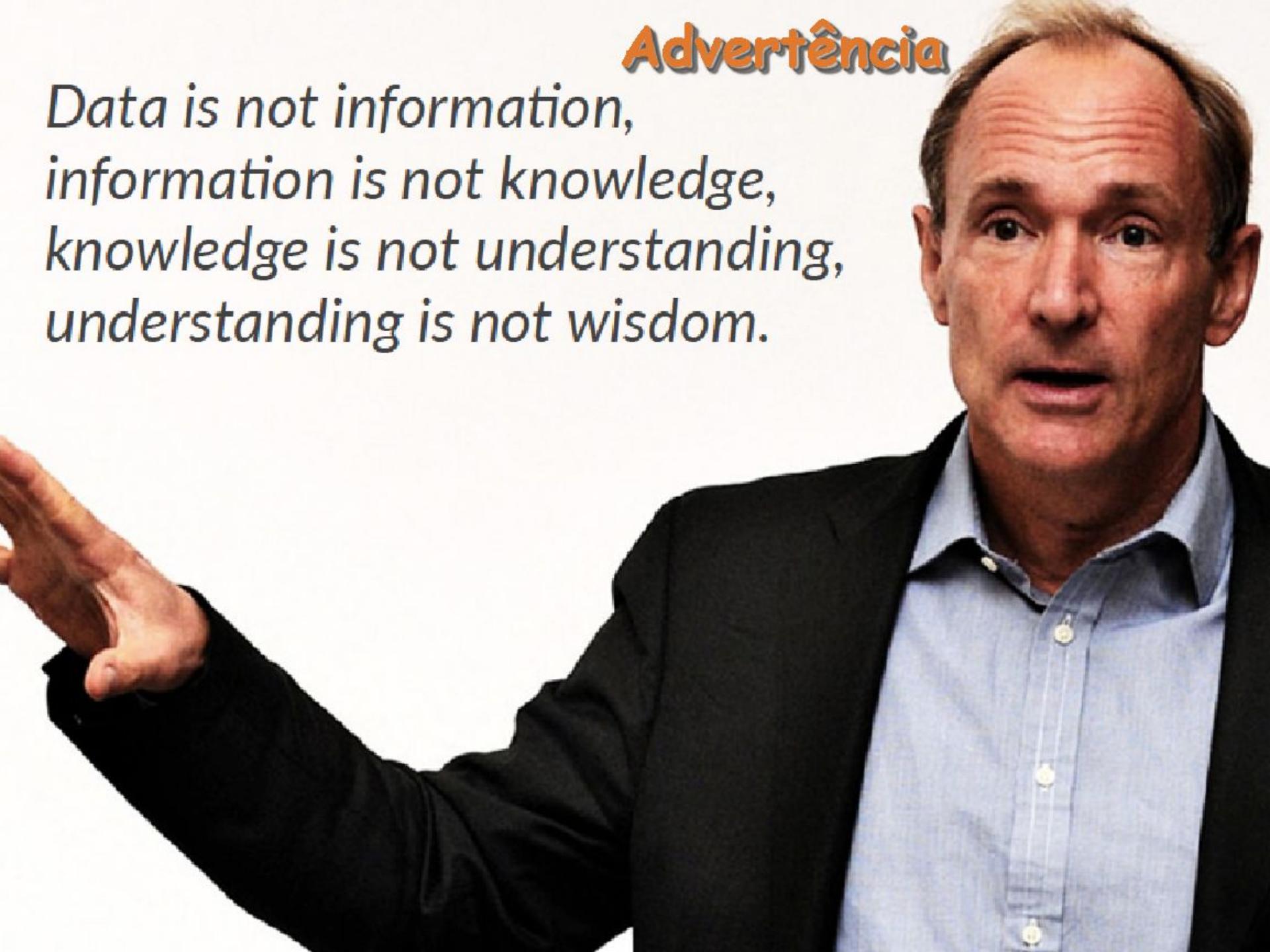


Sobre Dados



Advertência

*Data is not information,
information is not knowledge,
knowledge is not understanding,
understanding is not wisdom.*



Motivação

Muitas aplicações importantes são baseadas em conhecimento.



predisposing factors
symptoms
test results
diseases
treatment outcomes

Thanks to: Eric Horvitz, Microsoft Research

Medical Diagnosis

Appllet started

OnParenting May 14 - May 20, 1997

Fidelity Investments Our home on [is where] click here

SEARCH ON STAGE ESSENTIALS COMMUNICATE FIND

COVER CONTENTS news EXPERTS fun handbook TALK FIND help feedback

There are two ways to search for specific information in OnParenting. In **Find by Word**, type the word(s) you want to find and get a list of titles relevant to that word. **Find by Symptom** will help you get information about children's symptoms. Help has tips to forget your search.

Find by Word **Find by Symptom**

Describe the child in the drop-down boxes at the right. Relevant information will appear below.

Age: Toddler Sex: Female

Complaint: Abdominal pain

Localized pain: Can the child localize, or point to, the site of the pain?

- No, unable to localize
- Below the navel to the child's left
- Above the child's navel
- Either of the child's sides
- Below the navel to the child's right
- Above the navel in the child's right
- Above the navel to the child's left
- Don't Know

Results so far

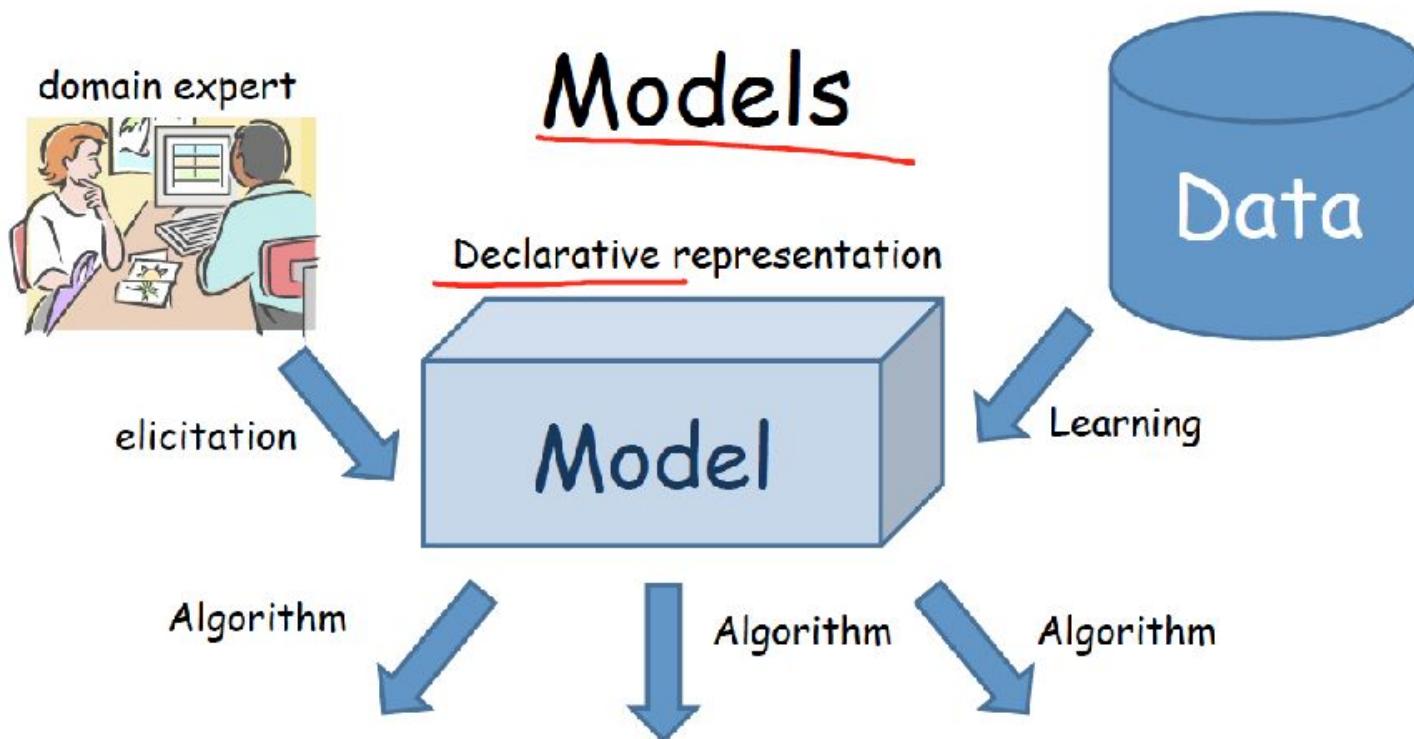
Disorder	Relevance
Viral gastroenteritis	High
Psychomotoric pain	Medium
Urinary tract infection	Low
Other	Low

Start Over Review

Next > Finish

Daphne Koller

União de conhecimento com dados melhoram os resultados



Modelagem a partir de dados

1	-0.185294206	0.027055773	-0.171819975	0.093276222	-0.027994229	0.098787646	-0.020583535	0.032817213	-0.132101895	0.007443381
-0.185294206	1	0.046731844	0.192725409	0.06221338	-0.121835155	0.007001147	0.072787318	0.256501727	-0.332697412	0.301121428
0.027055773	0.046731844	1	0.192725409	0.06221338	-0.121835155	0.007001147	0.072787318	0.256501727	-0.332697412	0.301121428
-0.171819975	0.192725409	0.06221338	1	0.093276222	-0.027994229	0.098787646	-0.020583535	0.032817213	-0.132101895	0.007443381
0.093276222	-0.027994229	0.098787646	-0.020583535	1	0.032817213	-0.132101895	0.007443381			
-0.020583535	0.032817213	-0.132101895	0.007443381		1					
0.032817213	-0.132101895	0.007443381				1				
-0.132101895	0.007443381						1			
0.007443381								1		
									1	
										1
1	0.046731844	0.192725409	0.06221338	-0.121835155	0.007001147	0.072787318	0.256501727	-0.332697412	0.301121428	0.007443381
0.046731844	1	0.192725409	0.06221338	-0.121835155	0.007001147	0.072787318	0.256501727	-0.332697412	0.301121428	0.007443381
0.192725409	0.06221338	1	-0.121835155	0.007001147	0.072787318	0.256501727	-0.332697412	0.301121428	0.007443381	0.007443381
0.06221338	-0.121835155	0.007001147	1	0.072787318	0.256501727	-0.332697412	0.301121428	0.007443381	0.007443381	0.007443381
-0.121835155	0.007001147	0.072787318	0.256501727	1	-0.332697412	0.301121428	0.007443381	0.007443381	0.007443381	0.007443381
0.007001147	0.072787318	0.256501727	-0.332697412	0.301121428	1	0.007443381	0.007443381	0.007443381	0.007443381	0.007443381
0.072787318	0.256501727	-0.332697412	0.301121428	0.007443381	0.007443381	1				
0.256501727	-0.332697412	0.301121428	0.007443381	0.007443381		1				
-0.332697412	0.301121428	0.007443381	0.007443381	0.007443381			1			
0.301121428	0.007443381	0.007443381	0.007443381	0.007443381				1		
0.007443381	0.007443381	0.007443381	0.007443381	0.007443381					1	
										1
1	0.046731844	0.192725409	0.06221338	-0.121835155	0.007001147	0.072787318	0.256501727	-0.332697412	0.301121428	0.007443381
0.046731844	1	0.192725409	0.06221338	-0.121835155	0.007001147	0.072787318	0.256501727	-0.332697412	0.301121428	0.007443381
0.192725409	0.06221338	1	-0.121835155	0.007001147	0.072787318	0.256501727	-0.332697412	0.301121428	0.007443381	0.007443381
0.06221338	-0.121835155	0.007001147	1	0.072787318	0.256501727	-0.332697412	0.301121428	0.007443381	0.007443381	0.007443381
-0.121835155	0.007001147	0.072787318	0.256501727	1	-0.332697412	0.301121428	0.007443381	0.007443381	0.007443381	0.007443381
0.007001147	0.072787318	0.256501727	-0.332697412	0.301121428	1	0.007443381	0.007443381	0.007443381	0.007443381	0.007443381
0.072787318	0.256501727	-0.332697412	0.301121428	0.007443381	0.007443381	1				
0.256501727	-0.332697412	0.301121428	0.007443381	0.007443381		1				
-0.332697412	0.301121428	0.007443381	0.007443381	0.007443381			1			
0.301121428	0.007443381	0.007443381	0.007443381	0.007443381				1		
0.007443381	0.007443381	0.007443381	0.007443381	0.007443381					1	
										1

Imagine looking at thousands of numerical values of data!

Looking for a needle in a haystack?

Without a proper data analysis tool, such an exercise will generate more **heat** than light!

Data \neq Information



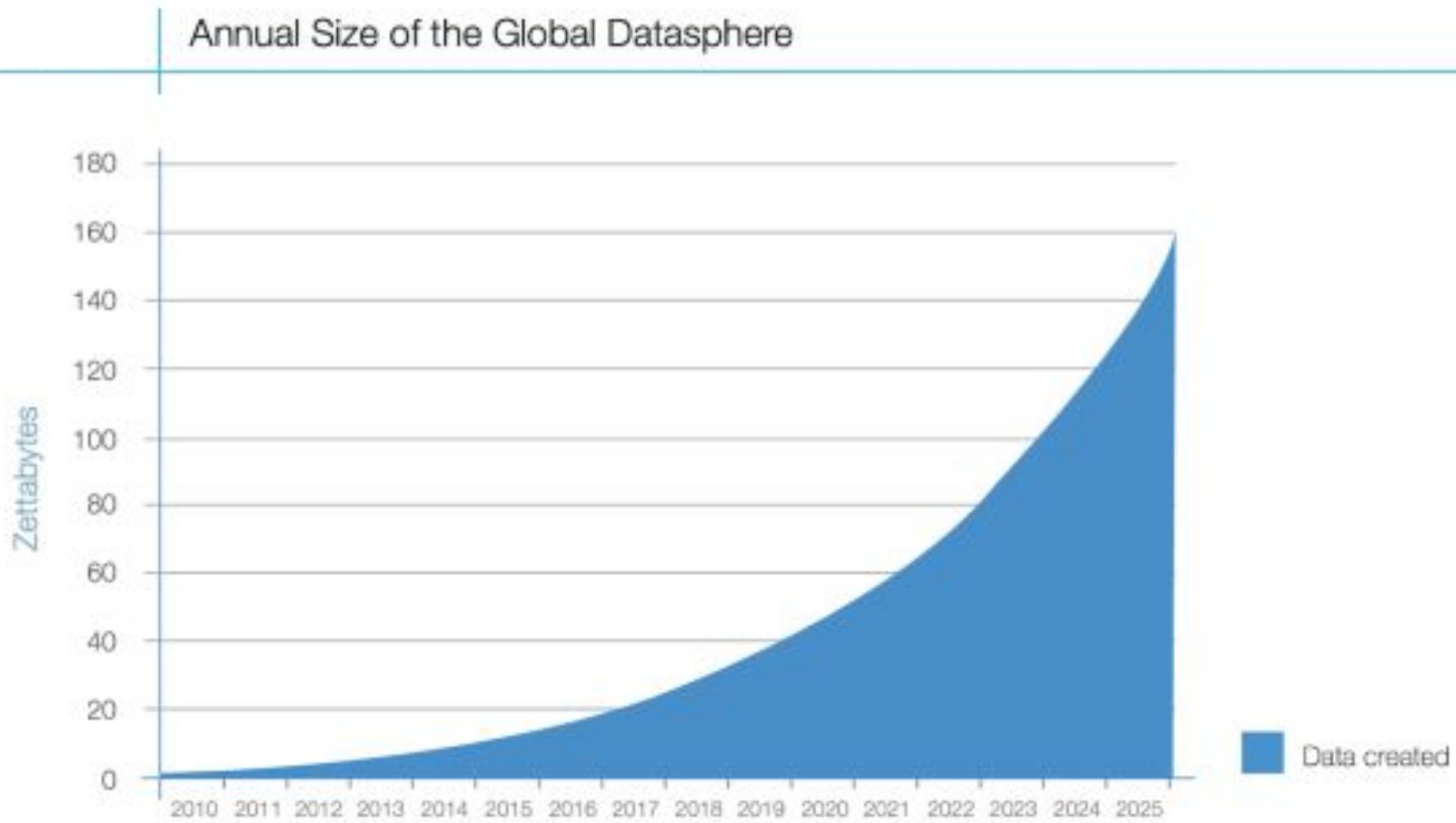
- Motivação - A quantidade de dados está crescendo como uma torrente



Motivação: dados em todos os lugares

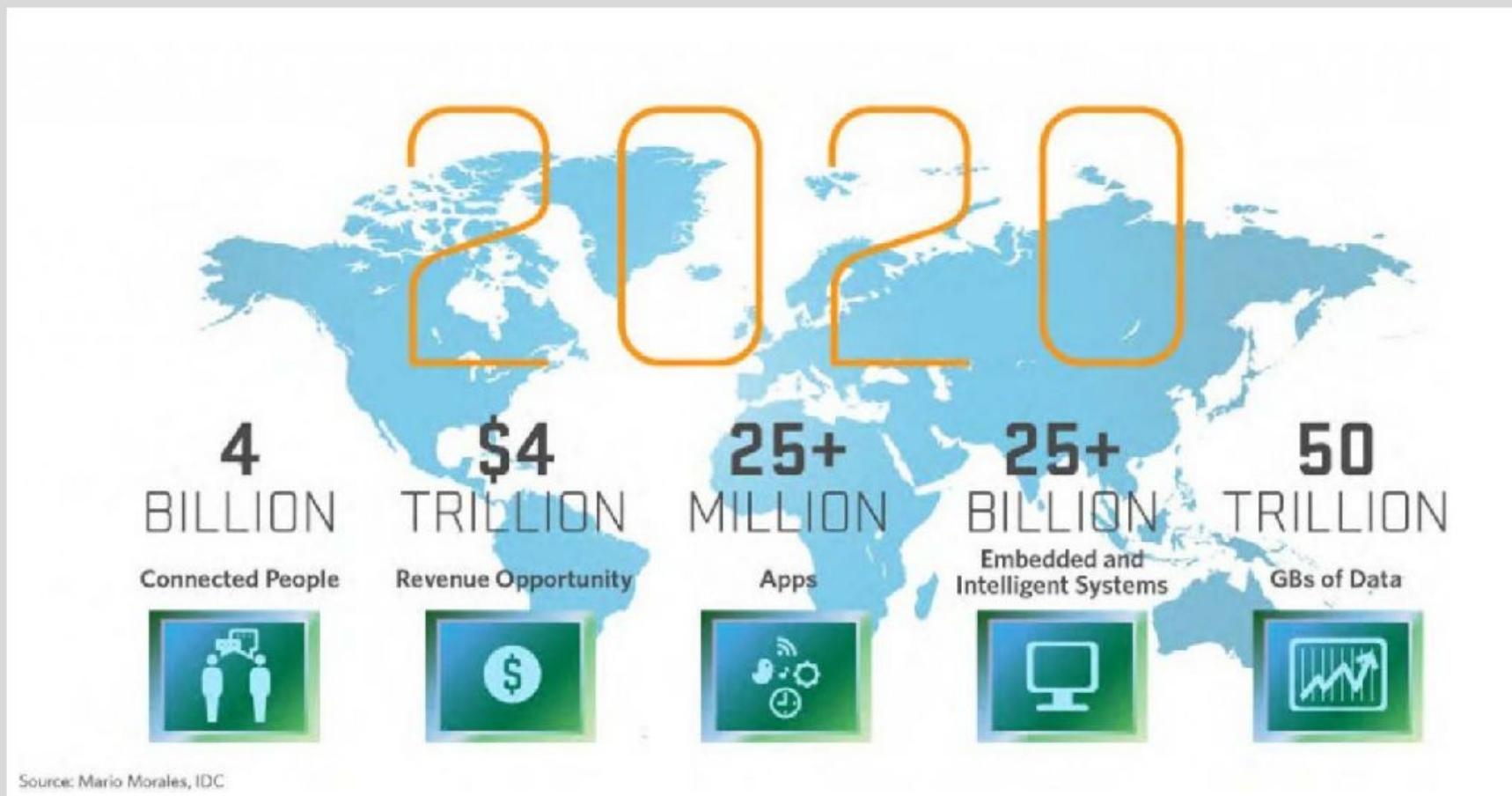
- 2.5 quintilhões de bytes gerados por dia.
- 30 bilhões de compartilhamento de conteúdo no Facebook por mês.
- 13 bilhões de dispositivos conectados à Internet em 2013 e 50 bilhões em 2020.
- Crescimento projetado de 40% em dados globais por ano.
- Demanda de 160.000 posições de trabalho para analistas de dados só nos EUA.

Por que estudar Ciência de Dados?

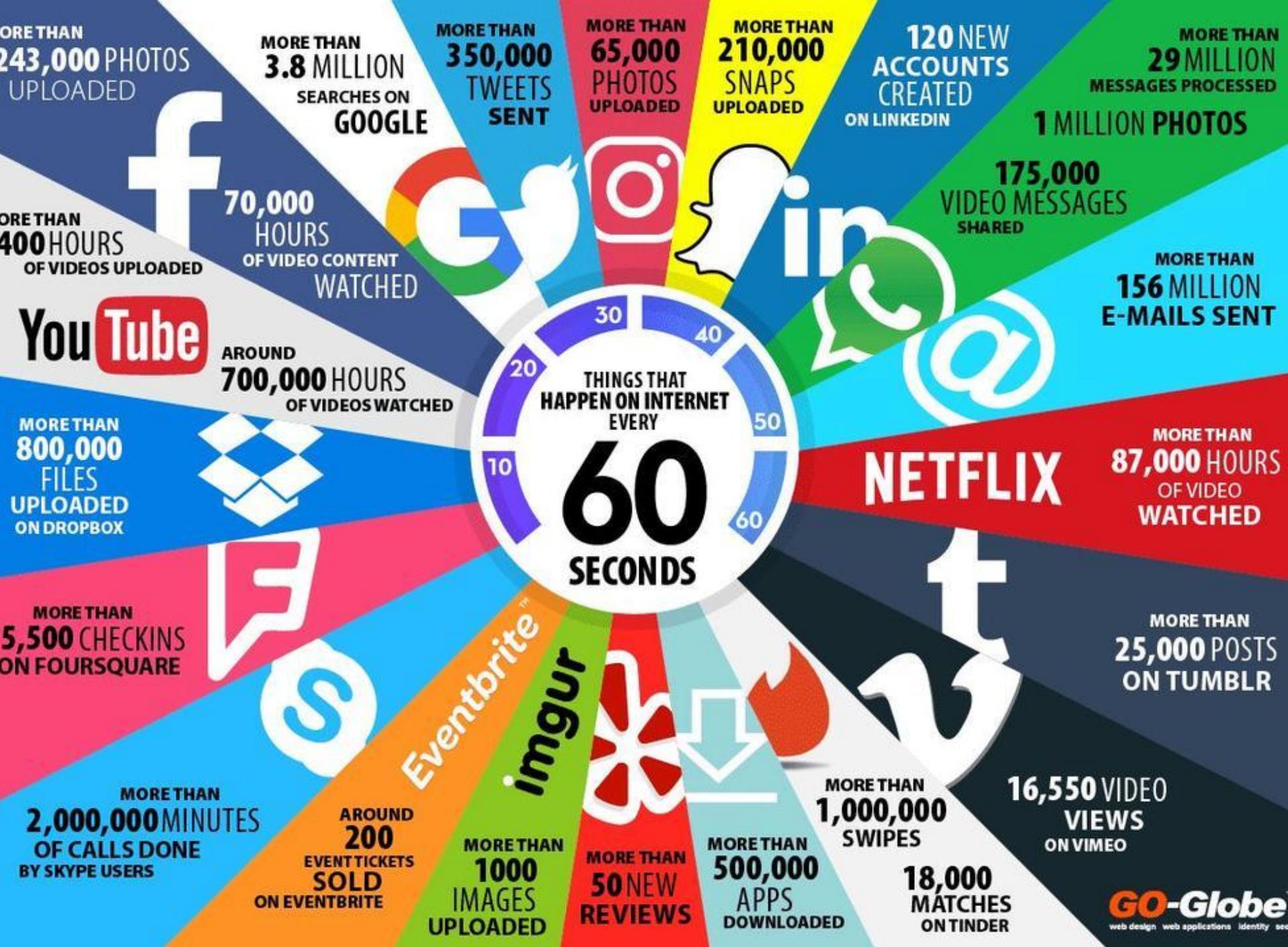


Source: IDC's Data Age 2025 study, sponsored by Seagate, April 2017

Motivação- A quantidade de dados está crescendo como uma torrente







A DAY IN DATA

The exponential growth of data is undisputed, but the numbers behind this explosion – fuelled by internet of things and the use of connected devices – are hard to comprehend, particularly when looked at in the context of one day

500m

tweets are sent every day

Twitter

294bn

illion emails are sent



3.9bn

people use emails

320bn
emails to be sent each day by 2021

306bn
emails to be sent each day by 2020

dell Group

4PB

of data created by Facebook, including

350m photos

100m hours of video watch time

Facebook Research



4TB

of data produced by a connected car

Intel

4.4ZB

ACCUMULATED DIGITAL UNIVERSE OF DATA

44ZB

PwC

2013

2020

DEMYSTIFYING DATA UNITS

From the more familiar "bit" or "megabyte", larger units of measurement are more frequently being used to explain the masses of data

Unit	Value	Size
b bit	0 or 1	1/8 of a byte
B byte	8 bits	1 byte
KB kilobyte	1,000 bytes	1,000 bytes
MB megabyte	1,000 ² bytes	1,000,000 bytes
GB gigabyte	1,000 ³ bytes	1,000,000,000 bytes
TB terabyte	1,000 ⁴ bytes	1,000,000,000,000 bytes
PB petabyte	1,000 ⁵ bytes	1,000,000,000,000,000 bytes
EB exabyte	1,000 ⁶ bytes	1,000,000,000,000,000,000 bytes
ZB zettabyte	1,000 ⁷ bytes	1,000,000,000,000,000,000,000 bytes
YB yottabyte	1,000 ⁸ bytes	1,000,000,000,000,000,000,000,000 bytes

*A lowercase "b" is used as an abbreviation for bits, while an uppercase "B" represents bytes.

65bn

messages sent over WhatsApp and two billion minutes of voice and video calls made

Facebook



463EB

of data will be created every day by 2025

IDC

95m

photos and videos are shared on Instagram

Instagram Business



28PB

to be generated from wearable devices by 2020

Statista



RACONTEUR

Motivação: depoimentos importantes

“ Data constitutes a new natural resource, which promises to be for the 21st century what steam power was for the 18th, electricity for the 19th and hydrocarbon for the 20th. ”

V. Rometty, IBM CEO

Motivação: depoimentos importantes - Hal Varian - Economista Chefe do Google

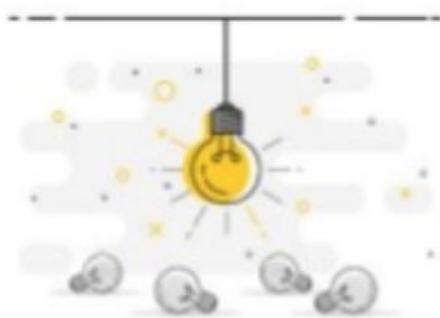
" If you are looking for a career where your services will be in high demand, you should find something where you provide a scarce, complementary service to something that is getting ubiquitous and cheap.

- ❖ So what's getting ubiquitous and cheap? Data!
- ❖ And what is complementary to data? Analysis

So my recommendation is to take lots of courses about how to manipulate and analyze data: databases, machine learning, econometrics, statistics, visualization, and so on."

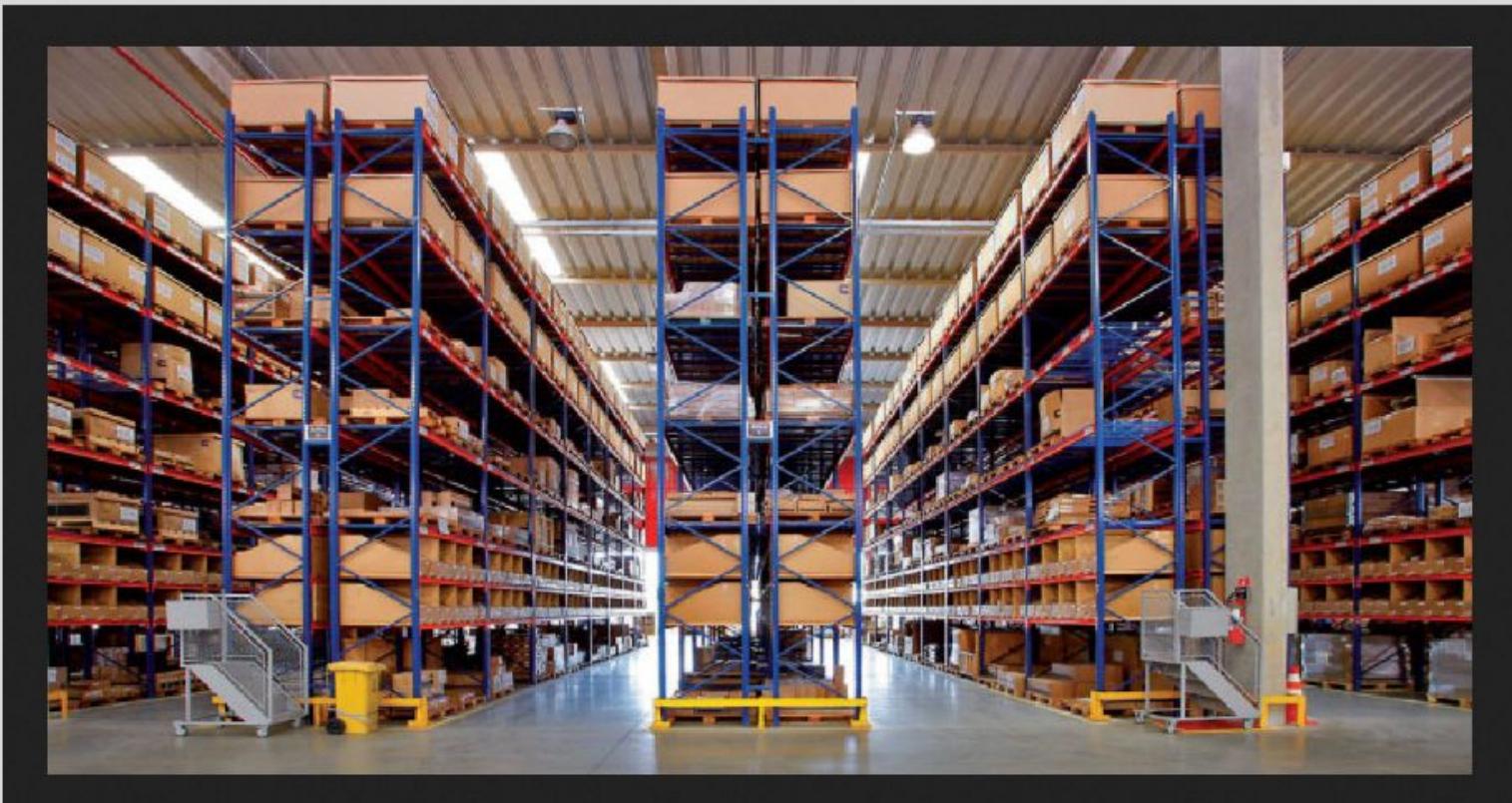
Data Science

“Ciência de Dados é a prática de transformar **dados** brutos em insights de **negócio** utilizando **métodos científicos**.”



Desafios:

Como armazenar grande volumes de dados?



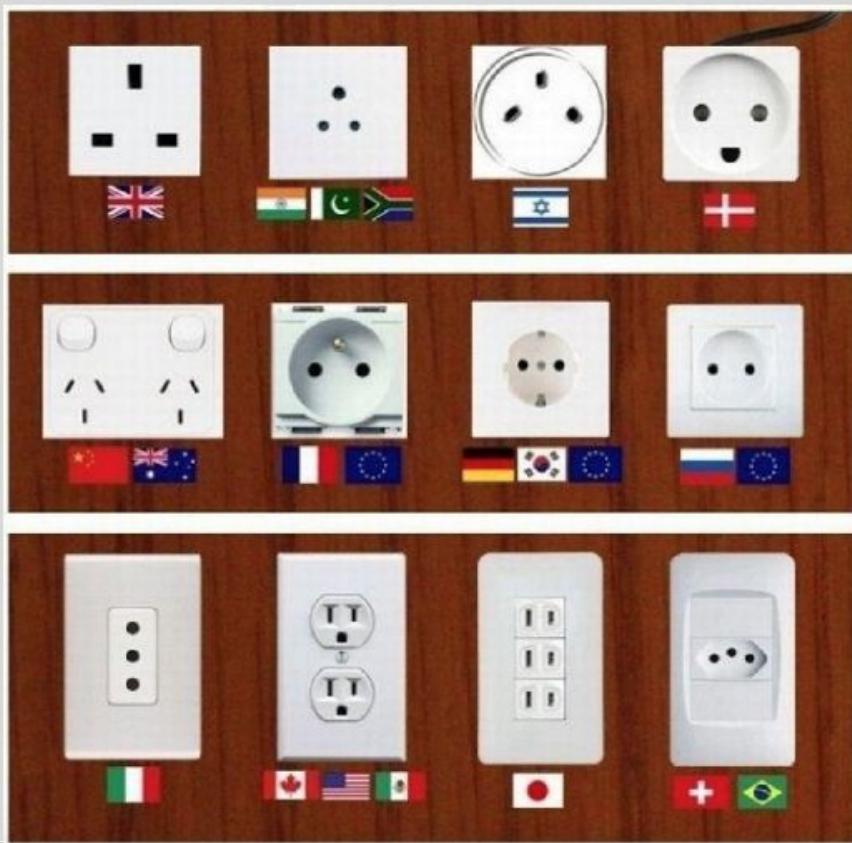
Desafios:

Como processar grande volumes de dados?



Desafios: Variedade

Como acessar fontes de dados diversas?



Desafios: Onde buscar os dados?



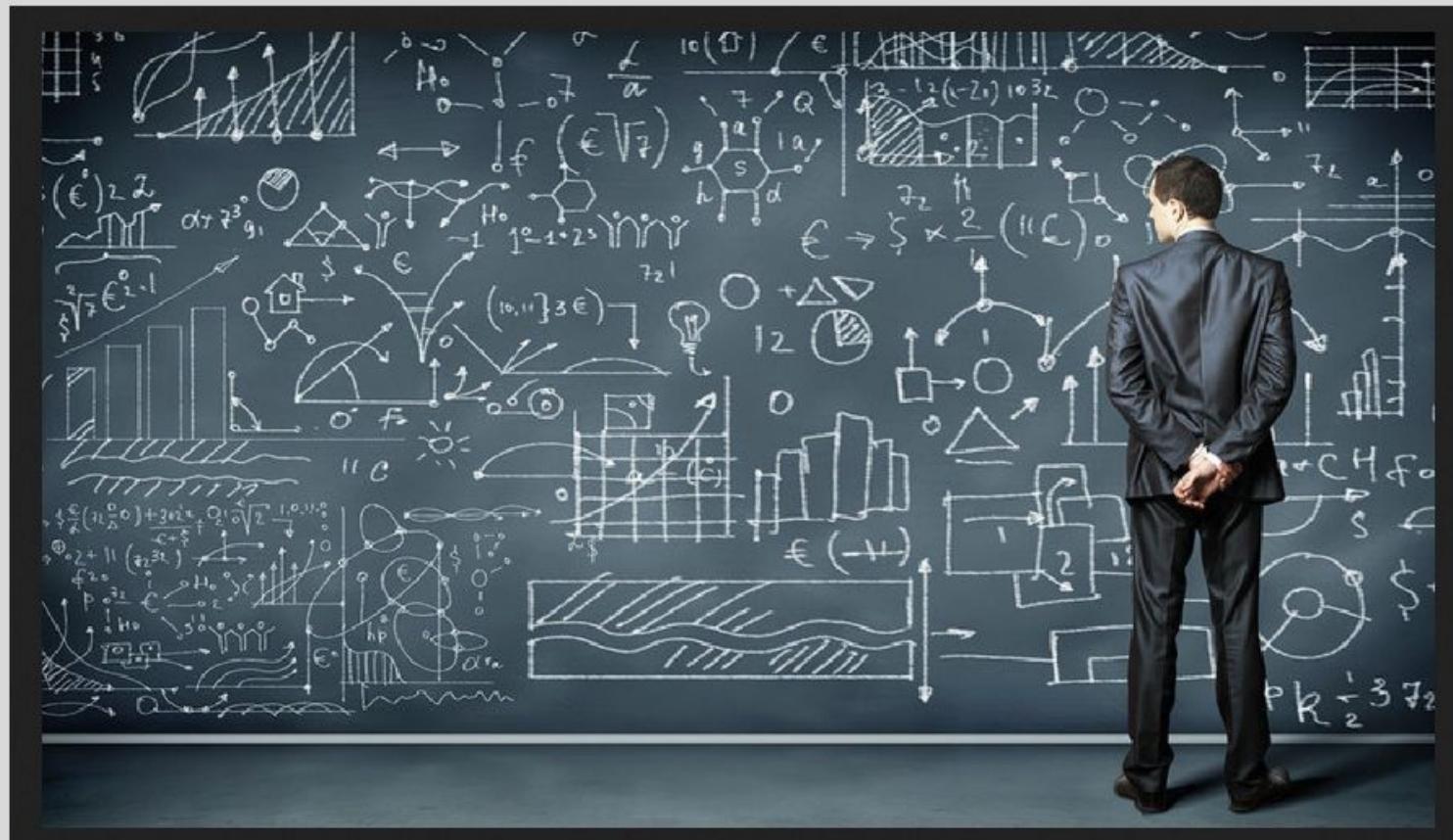
Desafios: Organização

Como organizar os dados/informação?



Desafios: Análise de Dados

Como extrair informação útil dos dados?



Desafios: Visualização

Como apresentar os dados de forma a obter informação útil?



Visualização de Dados

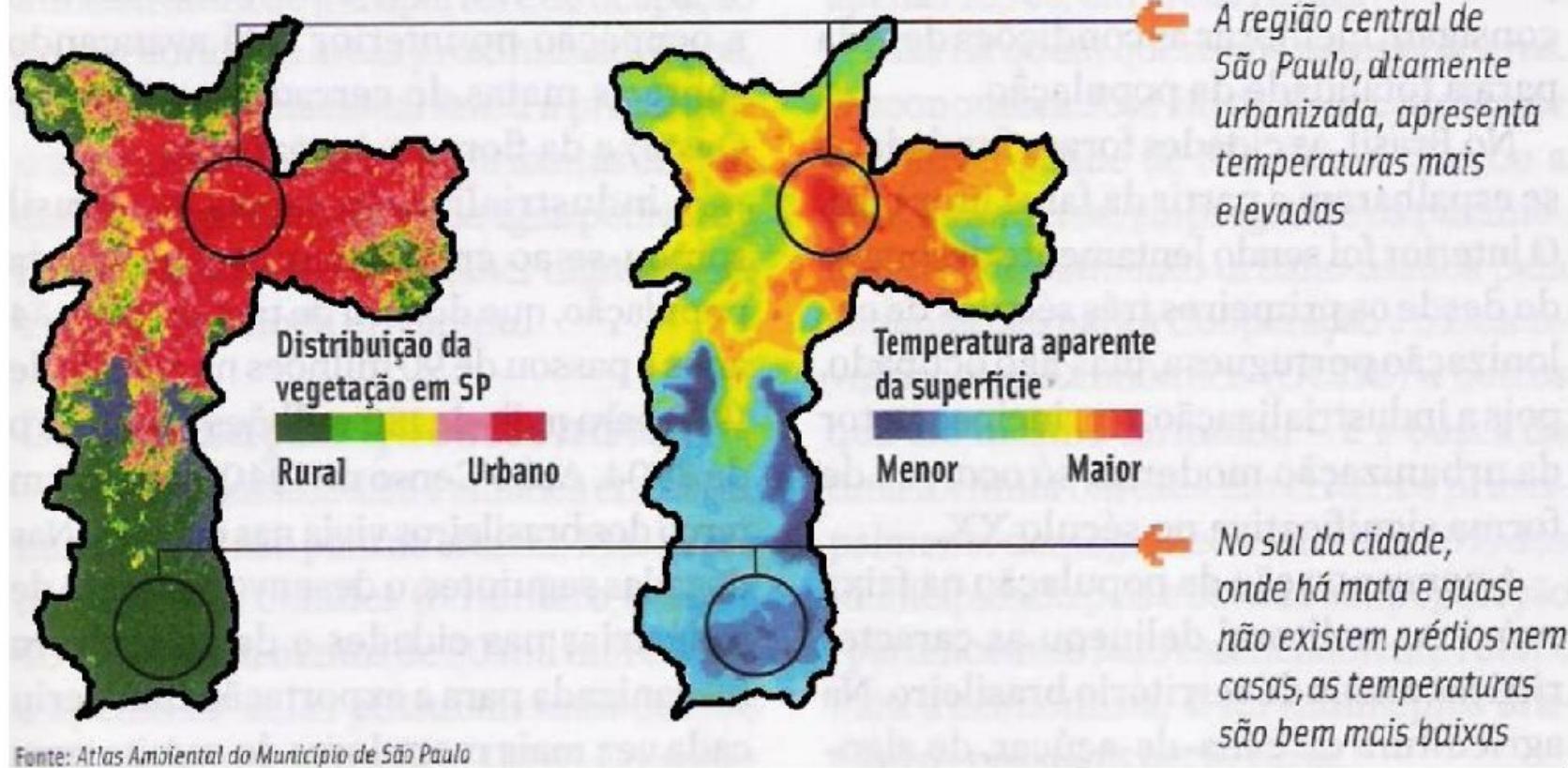


John Snow, médico higienista inglês, considerado um dos pais da epidemiologia¹, ficou conhecido pela solução dada a um surto de cólera no bairro do Soho, na Inglaterra, baseando-se nos princípios básicos na análise espacial.

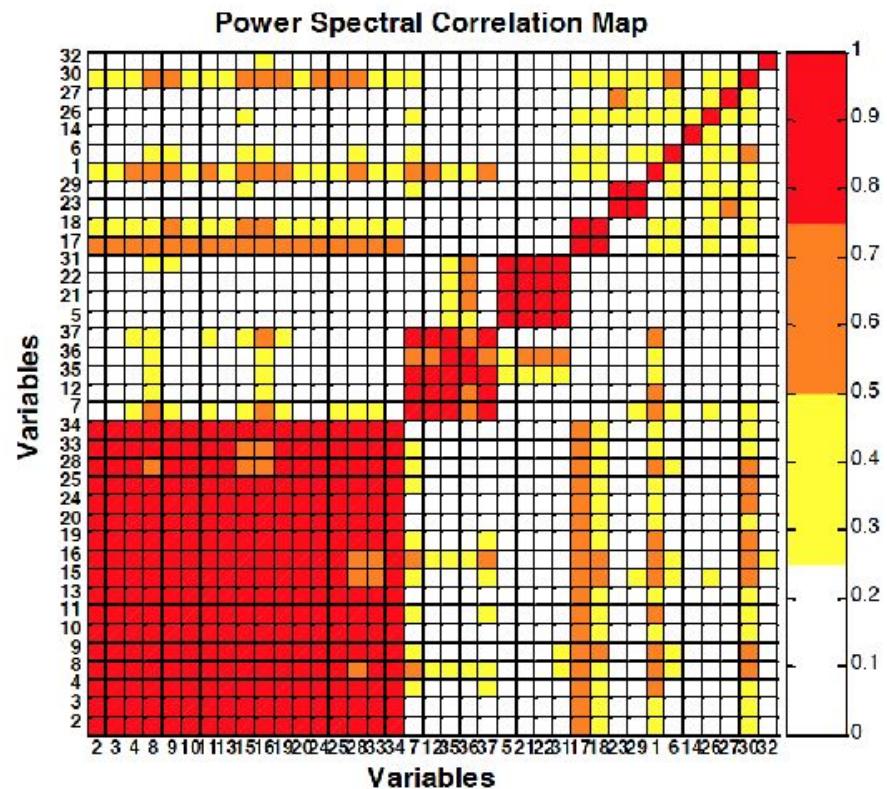
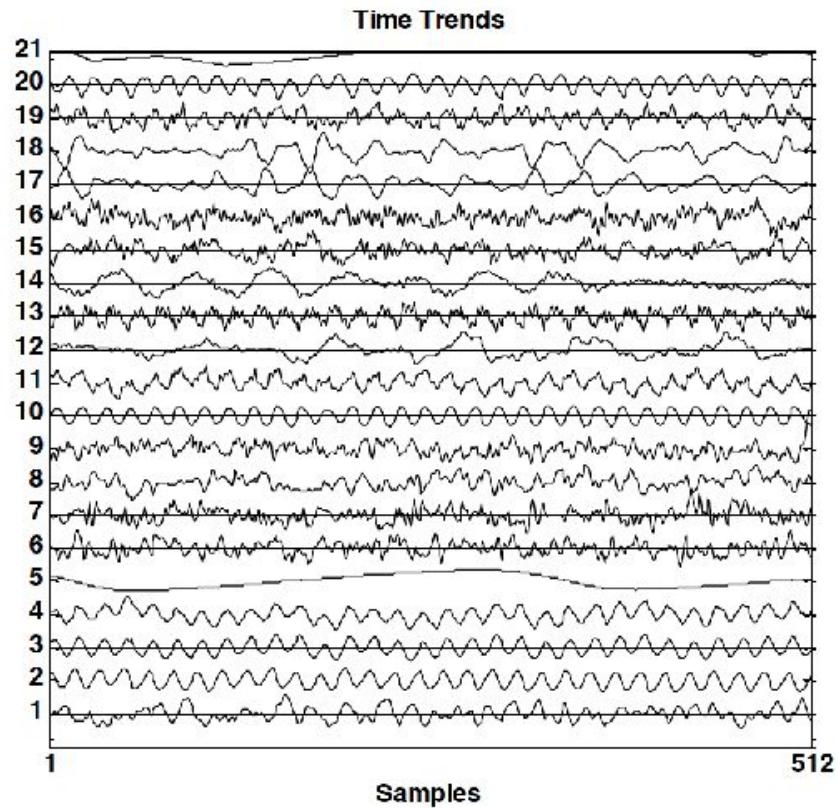
Visualização de Dados

DENSIDADE DEMOGRÁFICA E ILHAS DE CALOR

Município de São Paulo, com variação de temperatura de 24 °C a 32 °C, em 3/9/1999



Visualização de Dados



Desafios: Segurança

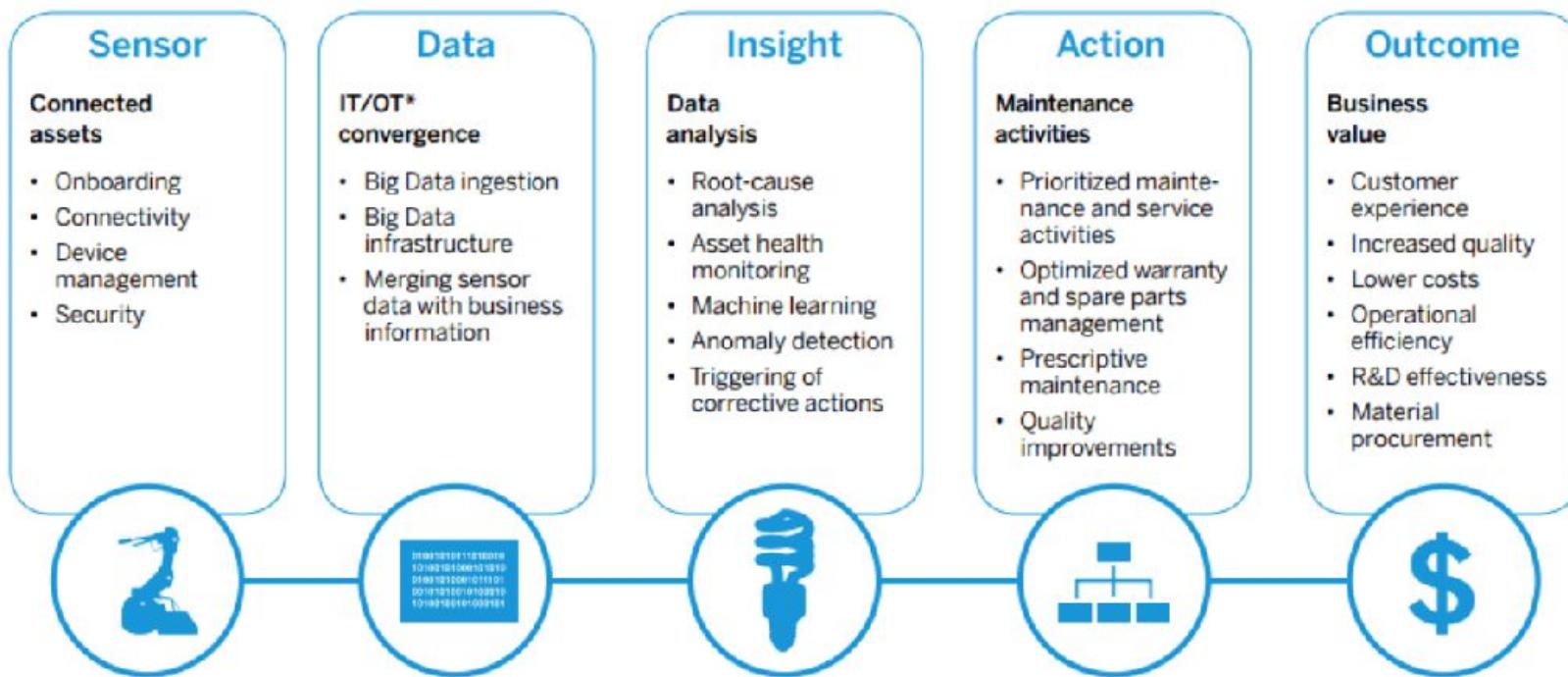
- Como proteger os dados, se eles estão na nuvem?



Desafios: Uso ético da informação

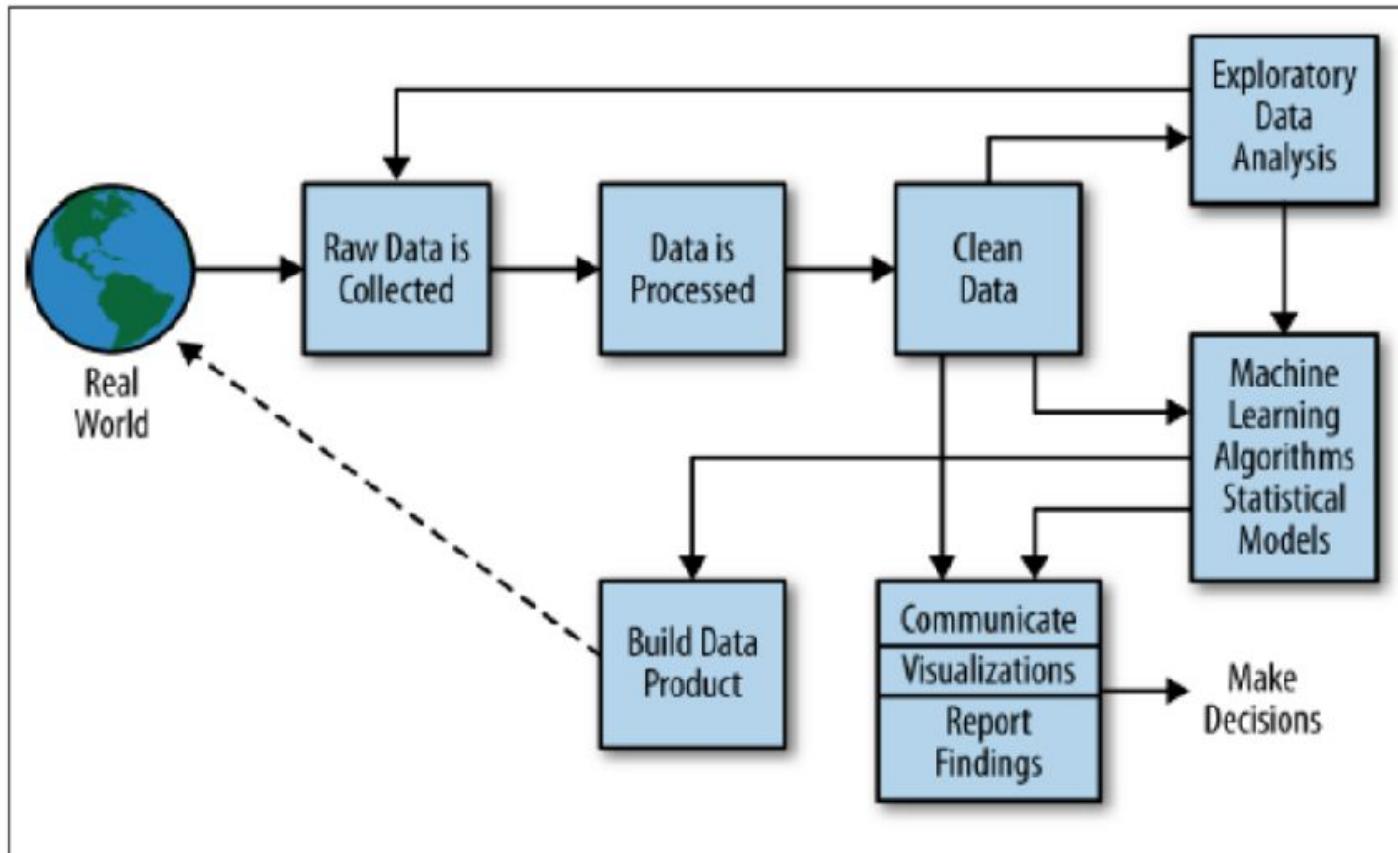


Caminho dos Dados



* OT = operational technology

Ciclo de Atividades de Ciência de Dados



Pipeline de Dados

Parte 1 Planejamento

Parte 2 Preparação dos dados

Parte 3 Modelagem

Parte 4 Acompanhamento

Pipeline de Ciência de Dados

O sucesso e aplicação de projetos de Ciência de Dados depende muito do nível de maturidade em dados das empresas:

1	Empírico	Ambiente caótico. Sem coleta de dados e decisões empíricas individualizadas
2	Adhoc	A maioria das empresas brasileiras. Dados coletados sem uma arquitetura de informação orientada a dados.
3	Definido	Dados coletados , com indicadores validados e orientados à cultura de dados. Decisões pautadas em sistemas de monitoramento e BI.
4	Otimizado	Dados coletados e enriquecidos . Geração automática de análises preditivas e prescritivas . Decisões baseadas em métricas.

Pipeline de Ciência de Dados



Toda análise precisa de um **objetivo bem definido e uma métrica**. A análise deve ser realizada em **conjunto com a área de negócio** para evitar a perda de foco do analista e facilitar a geração de insights.

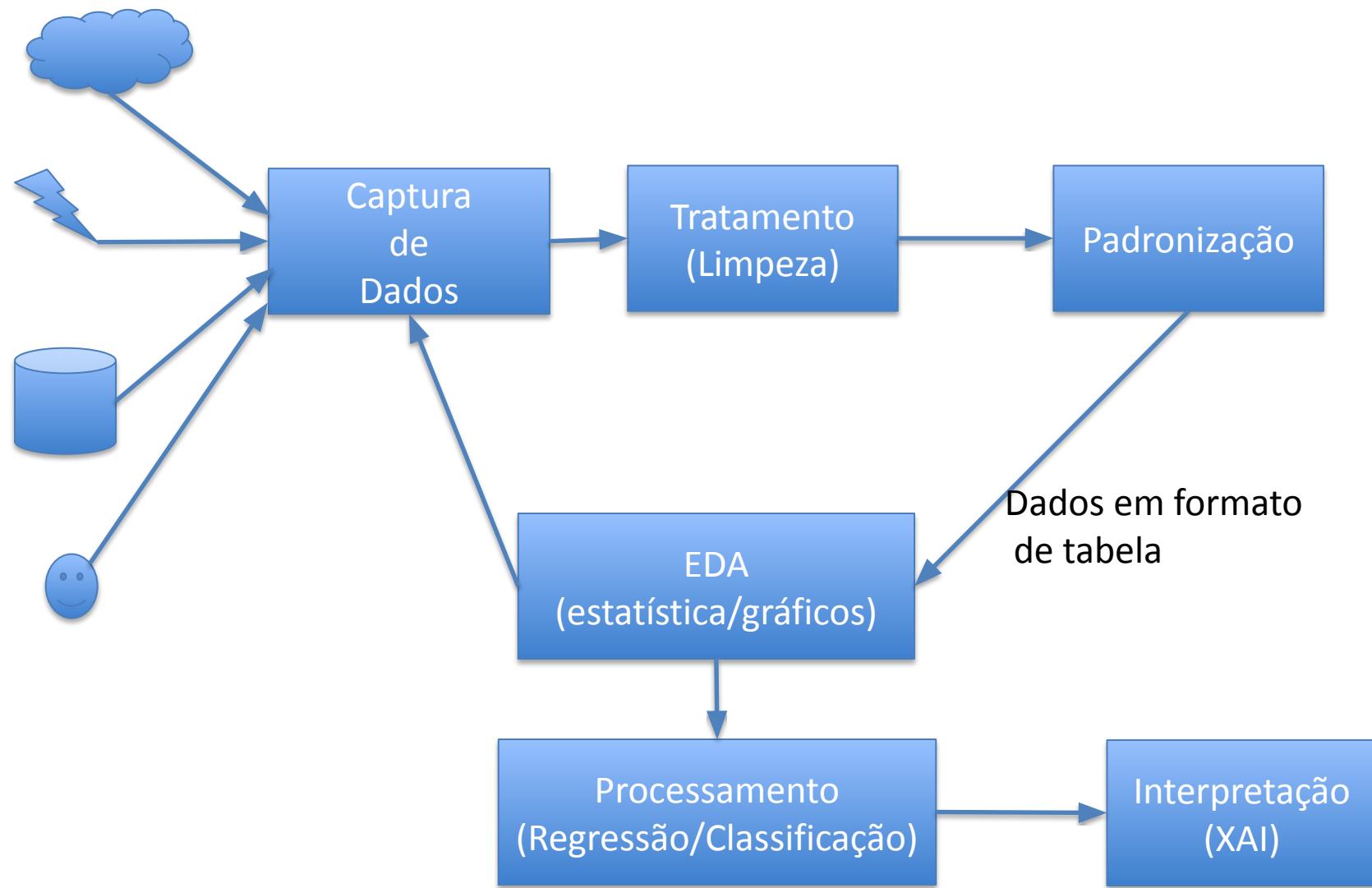
A **fonte de dados pode ser diversas**, desde redes sociais, bancos relacionais, csv.. etc. É importante que todos os dados necessários para a análise estejam disponíveis.

Métodos estatísticos e de Inteligência artificial são utilizados para extração de padrões. Geralmente é necessário combinar múltiplas fontes de dados e transformar variáveis para extração de padrões.

Avaliação e Implantação do Insight

Uma análise deve servir de base para a mudança de processos que geralmente visam retorno financeiro. É necessário validar se as decisões tomadas surtem efeitos reais, abrindo portas para outras análises.

Nossa Abordagem



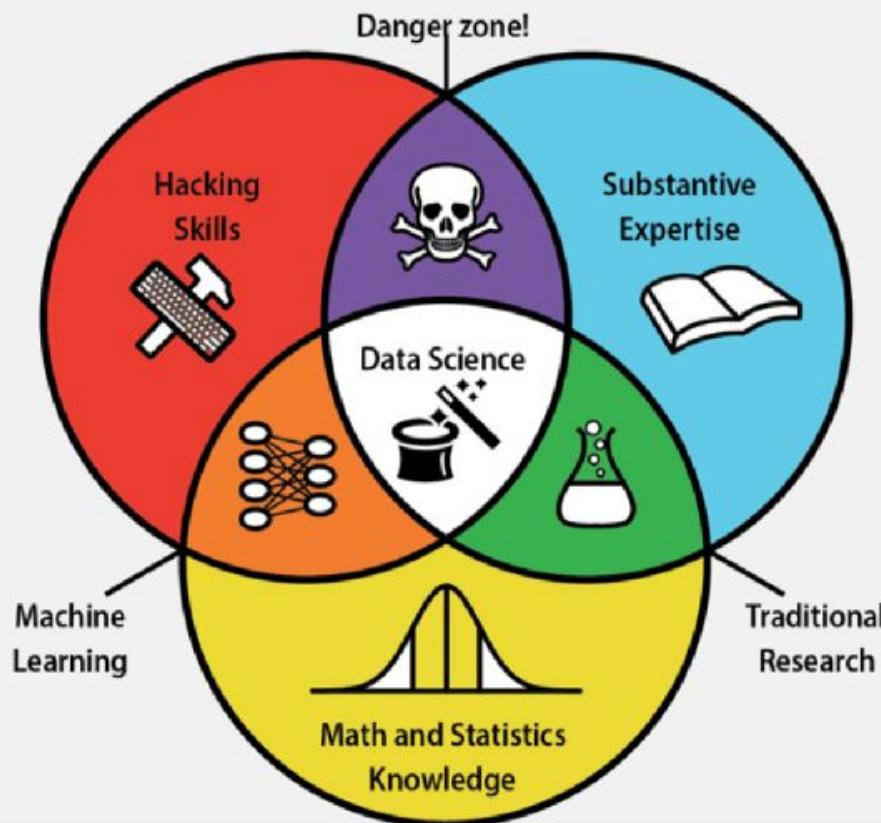
O que é um Cientista de dados?

Data
Scientist



[https://www.youtube.com/watch?
v=i2jwZcWicSY](https://www.youtube.com/watch?v=i2jwZcWicSY)

Cientista de Dados



O que faz um(a) Cientista de Dados?



Profissionais de Dados

Engenheiro de Dados



- Coleta de dados
- Infraestrutura de armazenamento
- Qualidade dos dados

#sql #nosql #ETL #bigdata
#python #cloud

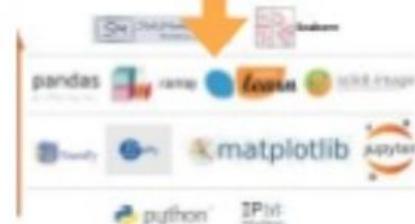


Cientista de Dados



- Análise estatística e ML
- Criação de Modelos
- Visualização dos resultados
- Geração de insights

#estatística #python #r
#machinelearning #dataviz



Engenheiro de Machine Learning



- Operacionalizar modelos
- Implementar modelos escaláveis
- Integração

#matemática #machinelearning
#cloud #MLops



Profissionais de Dados

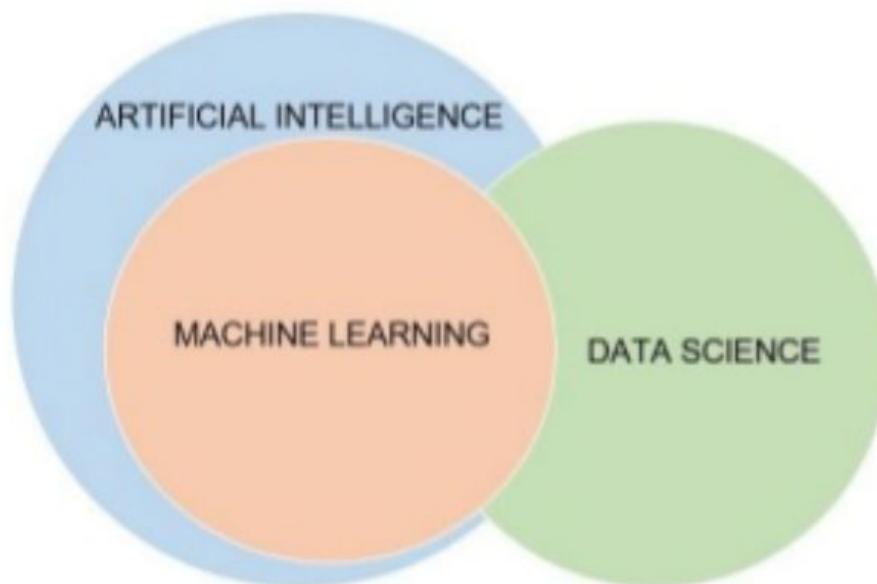
O cientista de dados não precisa saber tudo da área de Inteligência Artificial, apenas os métodos e algoritmos de extração de padrões nos dados (ML).

Cientista de Dados



- Análise estatística e ML
- Criação de Modelos
- Visualização dos resultados
- Geração de insights

#estatística #python #
#machinelearning #dataviz



Profissionais de Dados

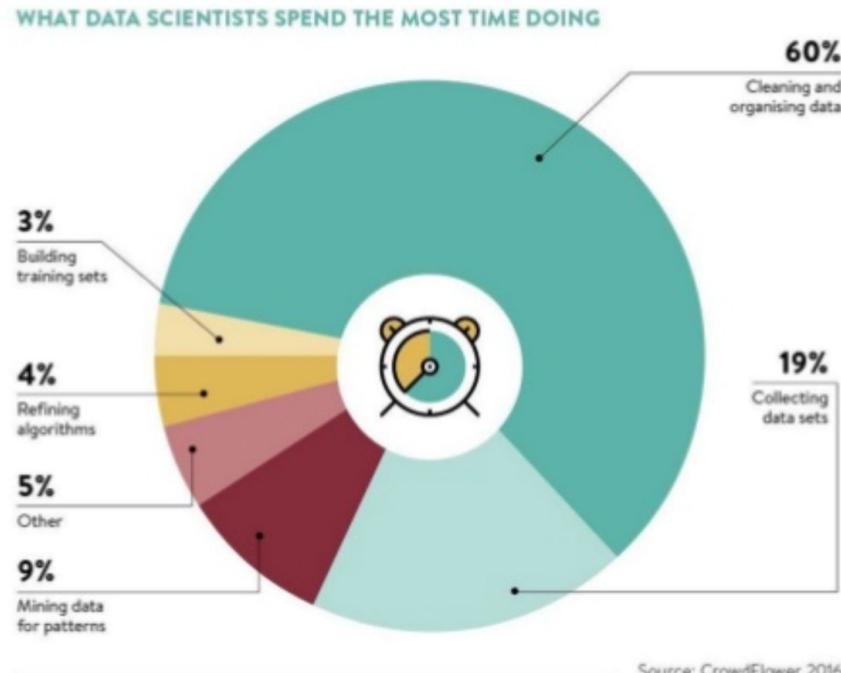
Um cientista de dados gasta aproximadamente **~80% do tempo** de uma análise coletando, limpando e organizando dados.

Cientista de Dados



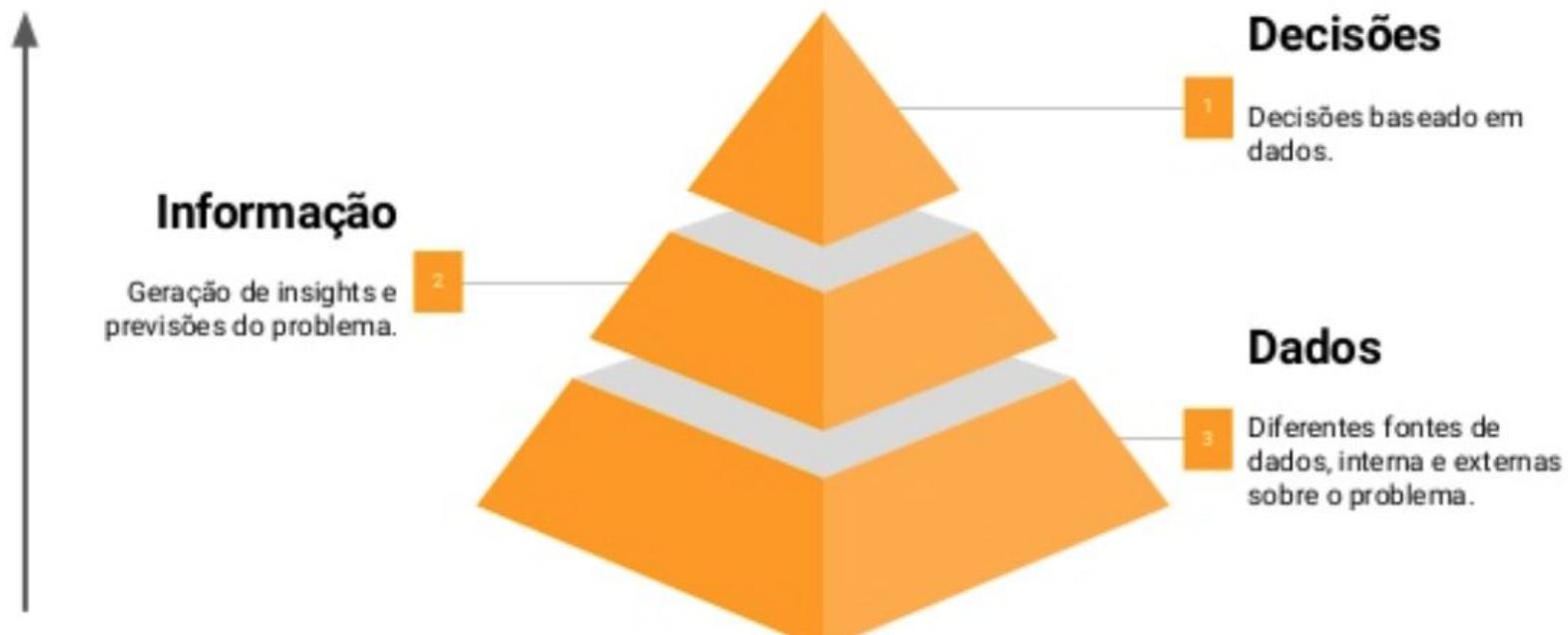
- Análise estatística e ML
- Criação de Modelos
- Visualização dos resultados
- Geração de insights

#estatística #python #
#machinelearning #dataviz



Aplicações de Data Science

A área de Data Science é **transversal** dentro de qualquer empresa e atende todas os setores (financeiro, logístico, comunicação..).





Cientista De Dados: mensal
Itaú Unibanco (Itaú BBA e Rede)
35 salários

R\$ 7.287/mês

R\$ 6 mil

R\$ 12 mil



Cientista De Dados: mensal
IBM
8 salários

R\$ 5.626/mês

R\$ 3 mil

R\$ 16 mil



Cientista De Dados: mensal
Propz
7 salários

R\$ 6.450/mês

R\$ 4 mil

R\$ 11 mil



Cientista De Dados: mensal
Hospital Israelita Albert Einstein
7 salários

R\$ 12.187/mês

R\$ 2 mil

R\$ 19 mil



Cientista De Dados: mensal
Semantix
6 salários

R\$ 7.533/mês

R\$ 7 mil

R\$ 16 mil



Cientista De Dados: mensal
TOTVS

R\$ 8.271/mês

R\$ 7 mil

R\$ 10 mil

Data Science

Profissionais de dados são altamente requisitados e estão em falta no mercado.



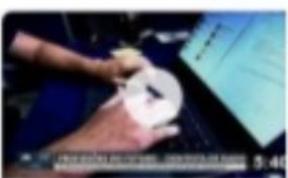
Mercado procura por cientistas de dados e promete salários de mais de R\$ 20 mil

Record TV - R7.com - 9 de ago de



Cientistas de dados: muito prazer, vocês são os profissionais do futuro

CDTV
YouTube - 27 de set de 2017



Profissões do Futuro:
cientista de dados

G1 - Globo - 20 de nov de 2018

Empresas como **Nubank**, **Itaú**, **Ifood** e **Globo.com** estão com **vagas permanentes** em Ciência de Dados.

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

DATA 100

Você gostaria de se tornar um(a) Cientista de Dados?



Pense bem. Depois que você fizer sua escolha, não tem mais volta.



