Ciência de Dados
Visão geral

Dojo do time Coca-Cola.
02/04/2020

Erik Yuri Dutzig
CWI SOFTWARE

SUMÁRIO

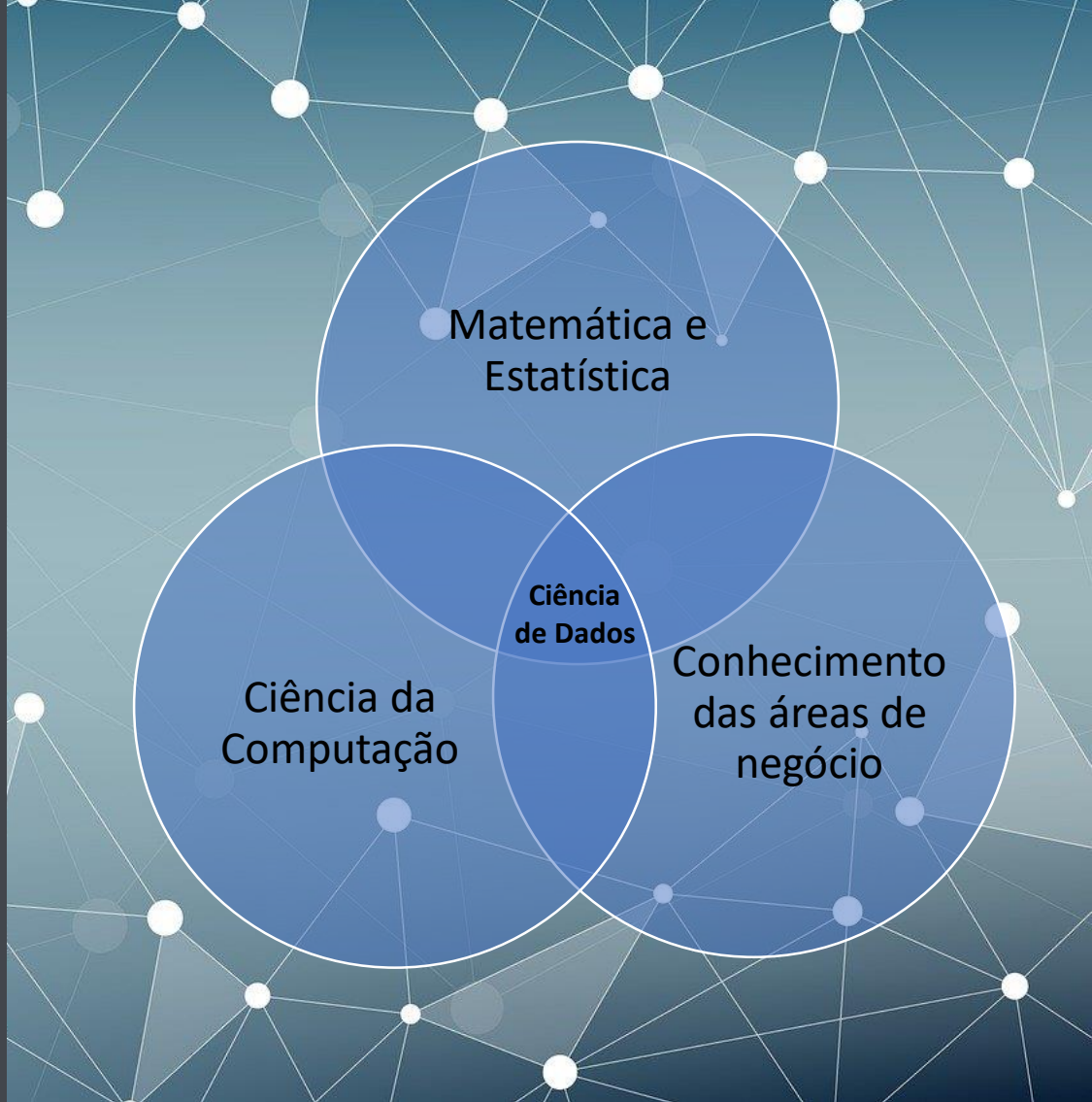
CWI SOFTWARE

01.

O que é Ciência de Dados.

O que é Ciência de Dados

Na sua essência, a **Ciência de Dados** envolve o uso de métodos automatizados (**ciência da computação**) para analisar dados (**estatística**) e extrair conhecimento (**áreas de negócio**) a partir deles.

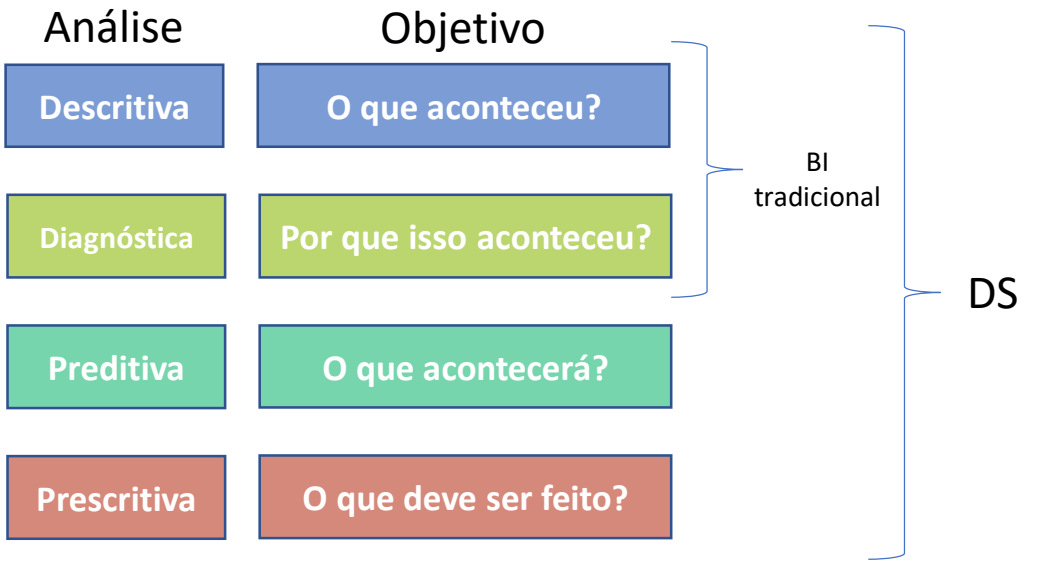


Ciência

O que é Ciência de Dados

Business Intelligence **não** é a mesma coisa que Data Science

Ambos usam dados, mas a sua abordagem, tecnologia e função diferem de diversas maneiras.



O que é Ciência de Dados

Data Science **não** é estatística

Estatística desempenha um papel fundamental dentro da Ciência de Dados. Porém, a Ciência de Dados compreende outras áreas de conhecimento, como já vimos anteriormente.

"Through this statement, the ASA and its membership acknowledge that data science encompasses more than statistics, but at the same time also recognize that statistical science plays a critical role in the fast-growing field. It is our hope the statement will reinforce the relationship of statistics to data science and further foster mutually collaborative relationships among all key contributors in data science."

ASA issues statement on role of statistics in data science
American Statistical Association (ASA), Washington, 2015.
https://www.amstat.org/asa/files/pdfs/pressreleases/2015-ASA-StatisticsFoundationaltoDataScience.pdf

02.
Áreas do Conhecimento.

Estatística.

Probabilidade

Estudo da aleatoriedade e da incerteza

> “
>
> **Estatística** é a **ciência**, parte da **Matemática Aplicada**, que fornece métodos para **coletar**, **descrever**, **analisar**, **apresentar** e **interpretar** dados, para a utilização dos mesmos na **tomada de decisões**.

Se lançarmos um dado perfeito, qual a probabilidade de sair um número menor que 3?

Sendo o dado perfeito, todas as 6 faces têm a mesma chance de caírem voltadas para cima. Temos 6 casos possíveis (1, 2, 3, 4, 5, 6) e que o evento "sair um número menor que 3" tem 2 possibilidades, ou seja, sair o número 1 ou o número 2. Assim, temos:

p(A): probabilidade da ocorrência de um evento A
n(A): número de casos que nos interessam (evento A)
n(Ω): número total de casos possíveis

$$P(A) = \frac{n(A)}{n(\Omega)}$$

$$P(A) = \frac{2}{6}$$

$$P(A) \cong 0{,}33 \cong \mathbf{33\%}$$

Estatística.

Estatística Descritiva

Utiliza métodos para coleta, organização, apresentação, análise e síntese de dados obtidos em uma população ou amostra.

> “
>
> **Estatística** é a **ciência**, parte da **Matemática Aplicada**, que fornece métodos para **coletar**, **descrever**, **analisar**, **apresentar** e **interpretar** dados, para a utilização dos mesmos na **tomada de decisões**.

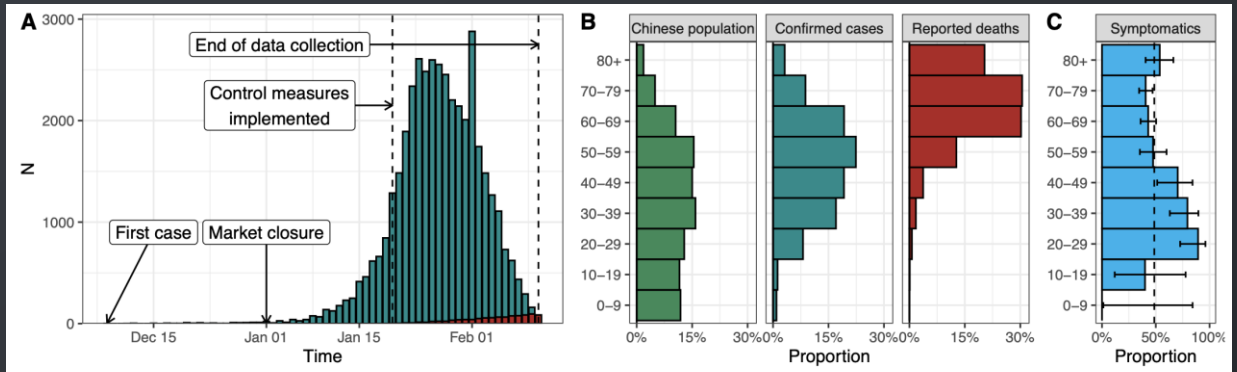


Figure 1: (A) Reported confirmed cases of COVID-19 in Hubei by date of disease onset (blue) and reported deaths (red) from 8 December, 2019 until 11 February, 2020. (B) Age distribution of the Chinese population compared to that of confirmed cases of and deaths due to COVID-19. (C) Proportion of individuals infected by COVID-19 showing symptoms among passengers of the Diamond Princess ship (with 95% credible interval).

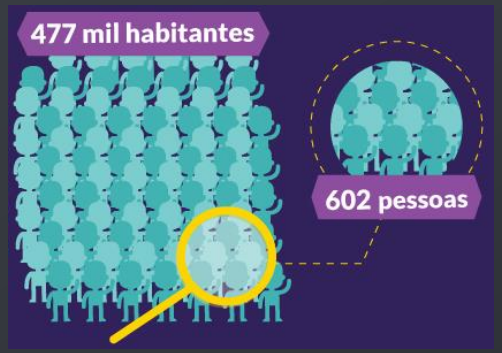
Estatística.

> “
>
> **Estatística** é a **ciência**, parte da **Matemática Aplicada**, que fornece métodos para **coletar**, **descrever**, **analisar**, **apresentar** e **interpretar** dados, para a utilização dos mesmos na **tomada de decisões**.

Estatística Inferencial

É o processo de estimar informações sobre uma população a partir dos resultados observados em uma amostra.

Inferência sobre a intenção de voto da população



Imagens: Isabela Souza, Politize!
https://www.politize.com.br/pesquisas-eleitorais-como-sao-feitas/, acesso em 30/03/2020

Conhecimento da área de negócio.

> "
>
> É necessário entender o problema que deu origem à investigação e coleta de dados (por isso o conhecimento de áreas de negócios é tão importante). O contexto é o que faz cada investigação estatística diferente.

Descoberta de tratamento para o covid-19

www.evqlv.com

Columbia University

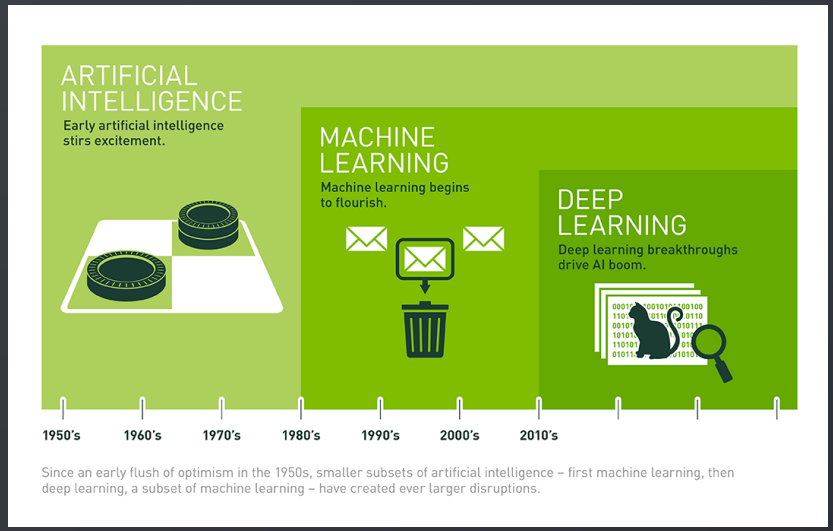
Com algoritmos de Machine Learning, Satz e Averso esperam acelerar a velocidade com que as terapias com Coronavirus são descobertas, desenvolvidas e fornecidas.

"O que nossos algoritmos fazem é reduzir a probabilidade de falha na descoberta de drogas no laboratório" Andrew Satz

"Esse esforço requer conhecimento, colaboração e a capacidade de processar quantidades incríveis de dados (...)" Rensselaer Polytechnic Institute

Ciência da Computação.

> “
>
> Na Ciência de Dados a Ciência da Computação oferece as soluções necessárias para o processamento, coleta e armazenamento de dados, em grande quantidade, velocidade e variedade, assim como a automatização dos processos daquela ciência.

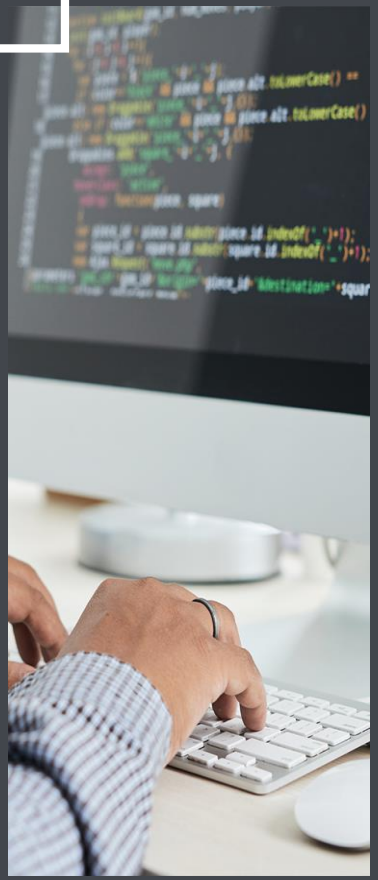


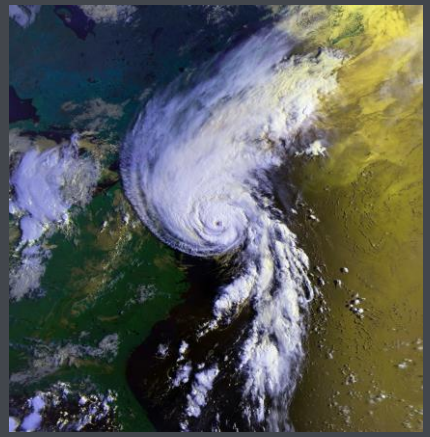
Nvidia

www.dataopsmanifesto.org

03.

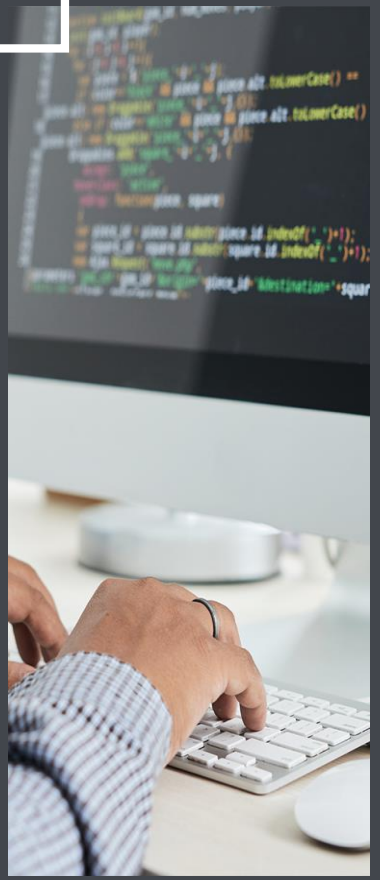
Exemplos de aplicações.





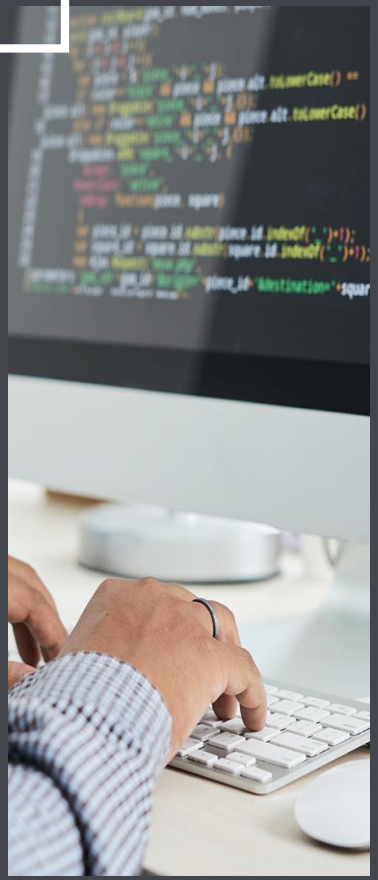
Aplicações

A Ciência de Dados pode ser utilizada em praticamente todas as atividades humanas, desde que dados sejam gerados e possam ser coletados.



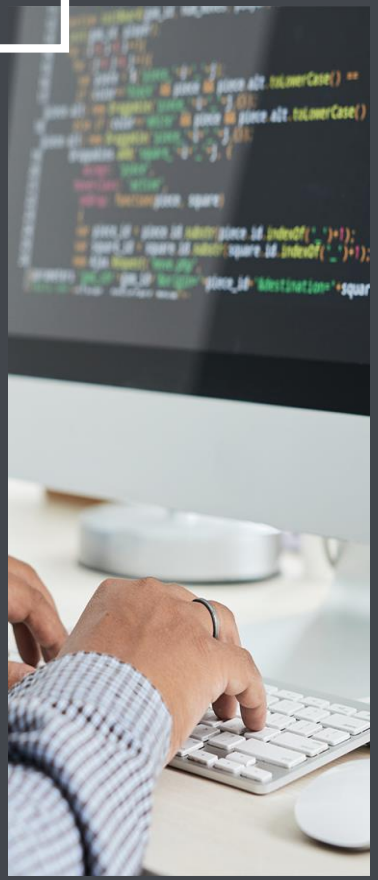
Walmart

- Servidores de dados com mais de 4 Petabytes de dados
- Cada venda é registrada
- Aproximadamente 267 milhões de transações por dia, nas 6000 lojas em todo o mundo
- Análise de Dados focada na avaliação da efetividade de estratégias de preço e campanhas de marketing
- Busca de melhoria na sua gestão logística e de inventário



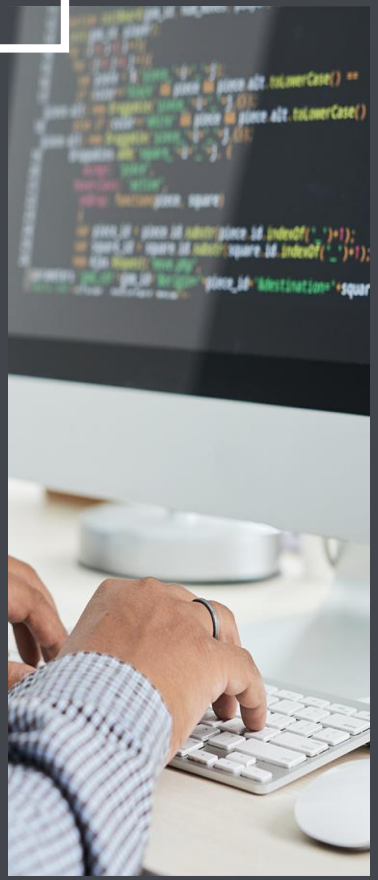
Amazon

- Personalização da experiência de compra online
- Cada cliente possui sua própria loja, baseada nas suas preferências
- Influência das avaliações de outros usuários nas decisões de compra



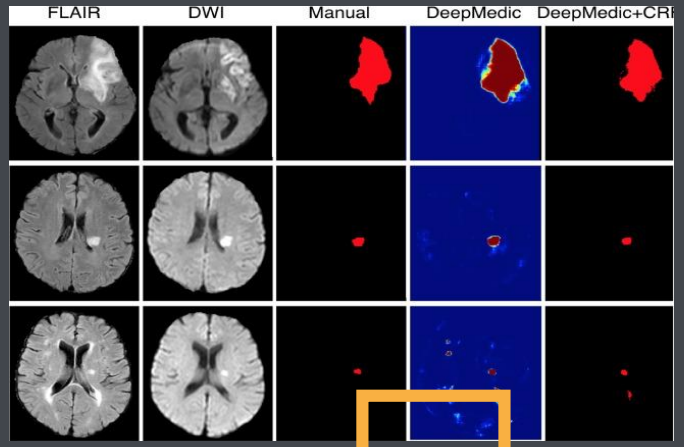
Citibank

A análise de cada uma das transações realizadas pelo banco, nos mais de 100 países em que opera, permite a geração de insights relacionados a investimentos, mudanças de mercado, padrões de operação e condições econômicas

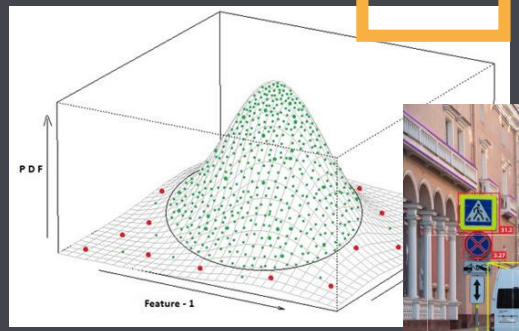


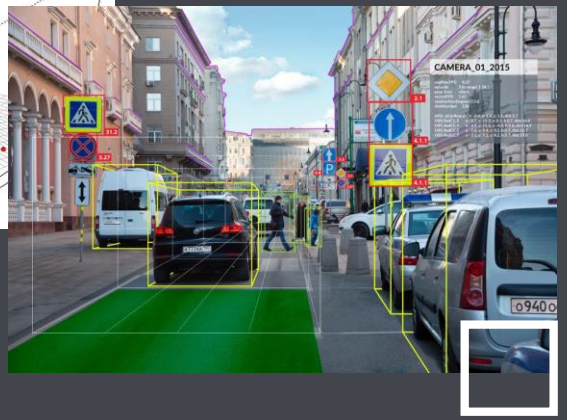
Outros exemplos

- Netflix
- Mídias Sociais (Facebook, LinkedIn, Twitter)
- Web Apps (Uber, AirBnB)
- Planejamento urbano (Cidades Europeias)
- Astrofísica (Nasa)
- Saúde Pública (Hospitais Americanos)
- Esportes (NFL)
- Chatbots e assistentes virtuais (Siri)
- Detecção de fraudes (UNIMED – CWI)



Estado da arte





Segmentação de imagens medicas

Visão computacional

Robôs investidores

Robôs advogados

Tratamentos para o Covid-19 *(Columbia University)*

04.

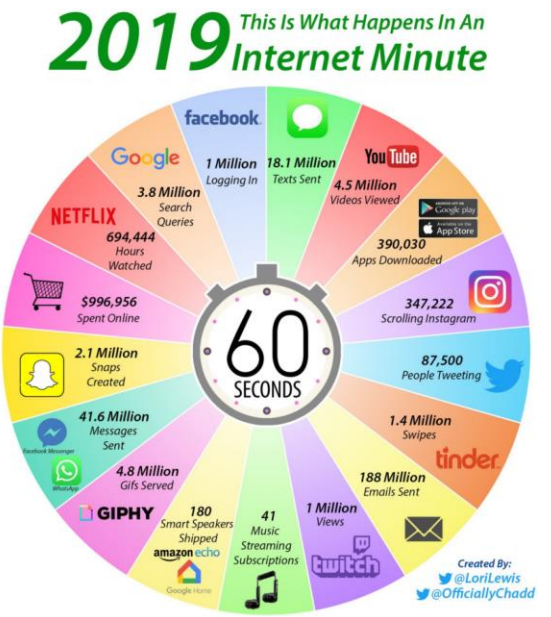
Expectativas para o futuro.

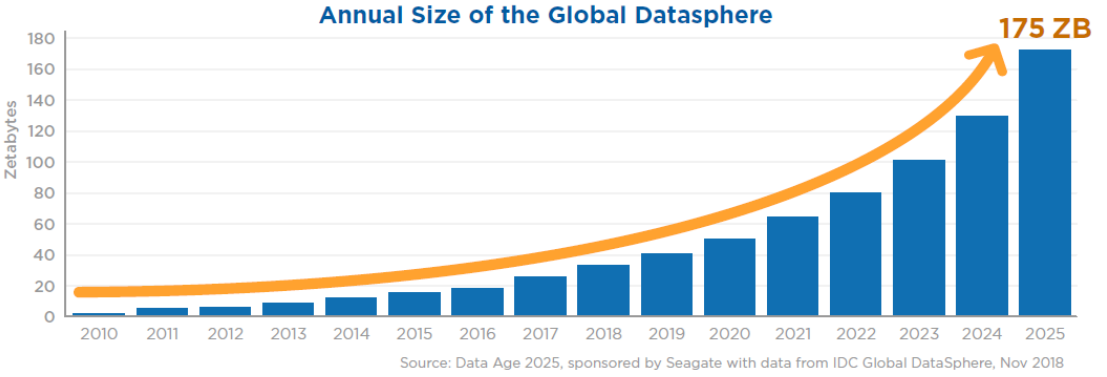
Futuro

DADOS

Estima-se que mais de 80% dos dados gerados atualmente são não estruturados

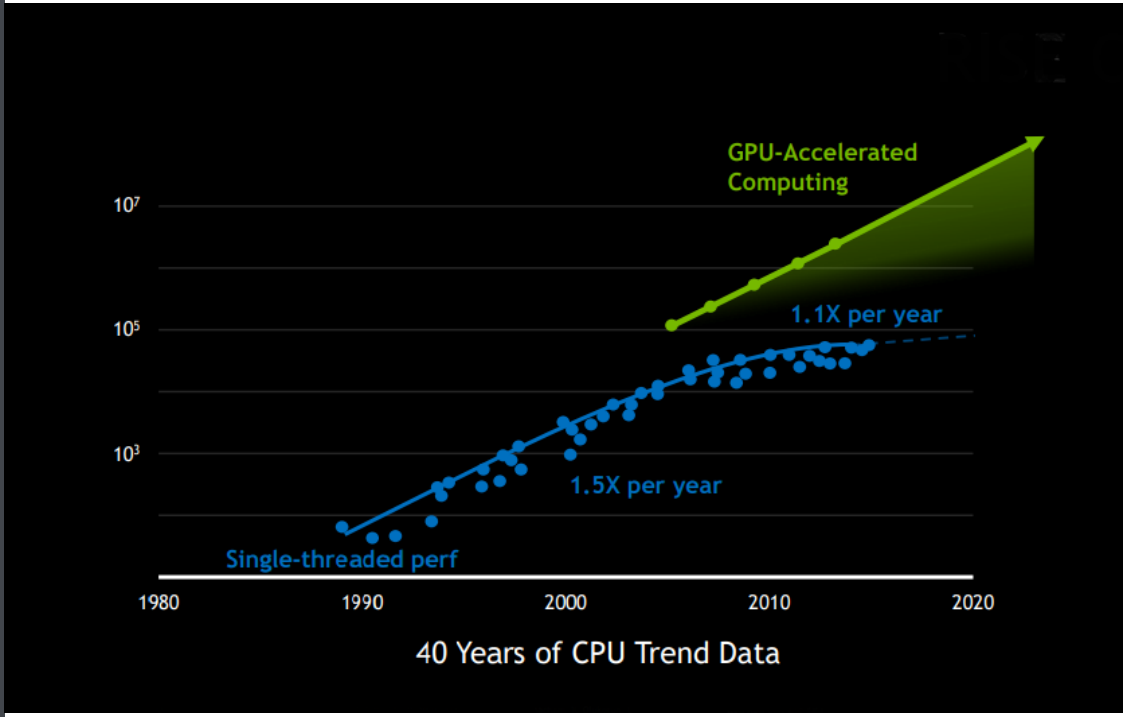
International Data Corporation (IDC)





Futuro

Capacidade computacional

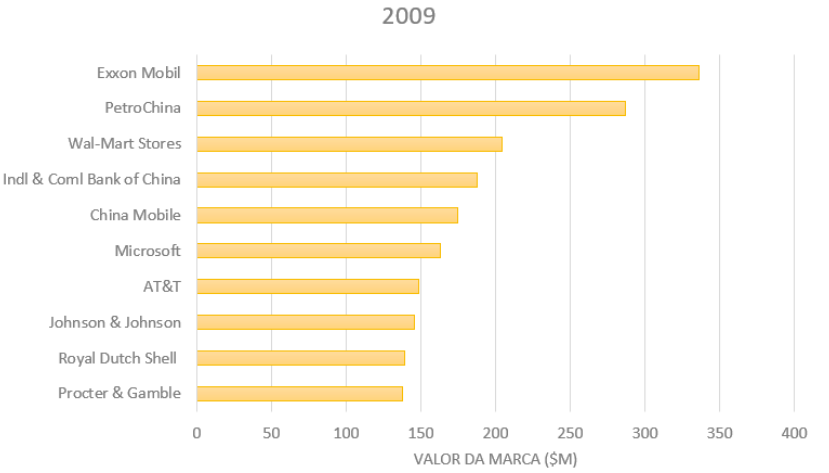


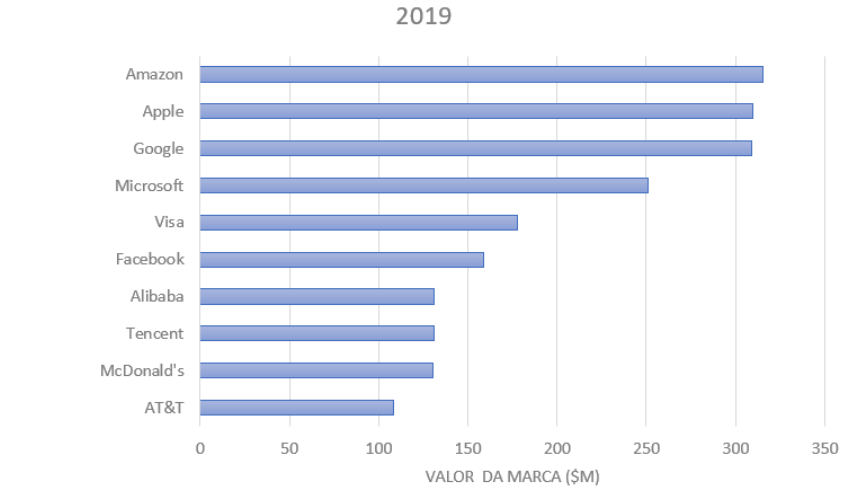
Fonte: Nvidia

Futuro

“The world’s most valuable resource is no longer oil, but data”

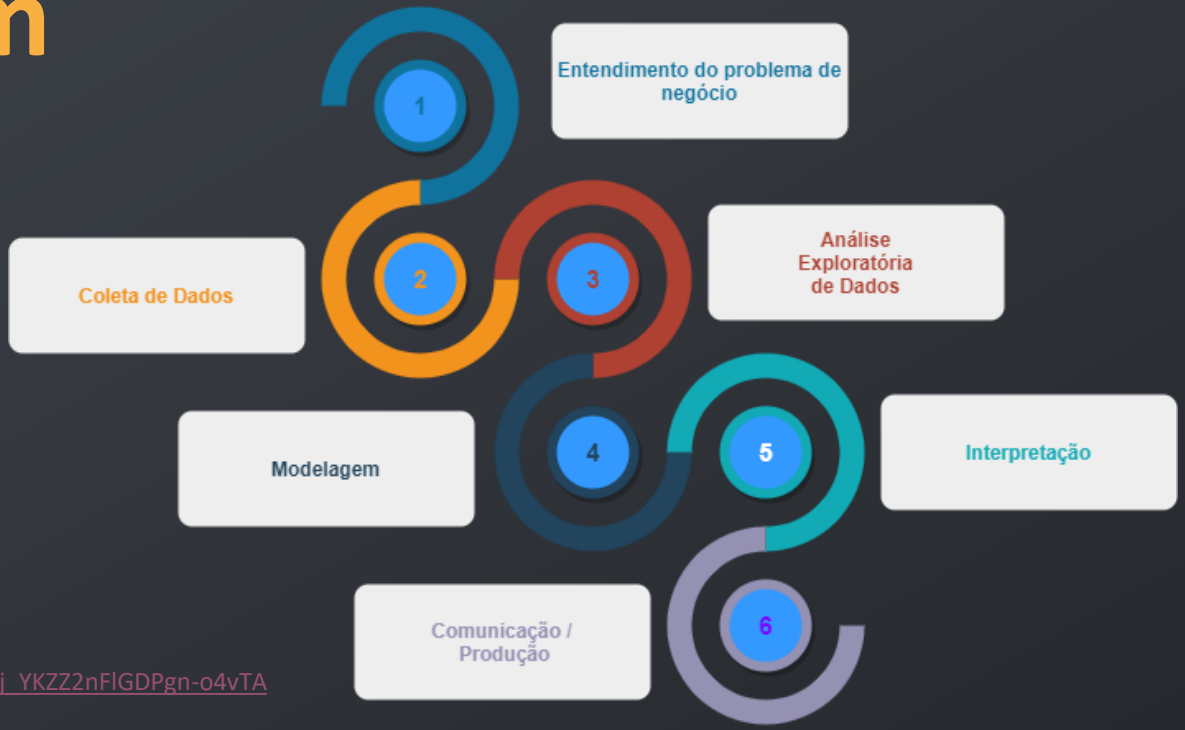
The Economist, 2017





Fonte: Global 500 2009 e Global 500 2019, Financial Times

Fases de um projeto.



Exemplo

https://colab.research.google.com/drive/17aX8PAw6S83j_YKZZ2nFlGDPgn-o4vTA

Material complementar

Blog com conteúdo sobre Ciência de Dados e Big data
http://datascienceacademy.com.br/blog/

Conteúdos sobre Deep Learning
http://deeplearningbook.com.br/

Coleção de notebooks
https://github.com/jupyter/jupyter/wiki/A-gallery-of-interesting-Jupyter-Notebooks

Canal do Mario Filho com tutoriais e exemplos
https://www.youtube.com/channel/UCIFd_i2iwYox1PPm9rD8wFA

Podcast brasileiro sobre Ciência de Dados
https://open.spotify.com/show/5k0Ei0MSg5BuiHshr43aSg

Instituto de pesquisa em IA
https://www.asimovinstitute.org/

Módulos Python para computação científica e Data Science
https://www.anaconda.com/enterprise/

Linguagem R, voltada para visualização e análise de dados
https://www.r-project.org/

Serviço de computação em nuvem da IBM
https://cloud.ibm.com/login

Sugestões de livros sobre Ciência de Dados
http://datascienceacademy.com.br/blog/10-sugestoes-de-livros-sobre-data-science-para-voce-ler-em-2020/

Erik Yuri Dutzig

erik.dutzig@cwi.com.br

CWI SOFTWARE