Descoberta do Conhecimento

Descoberta do Conhecimento

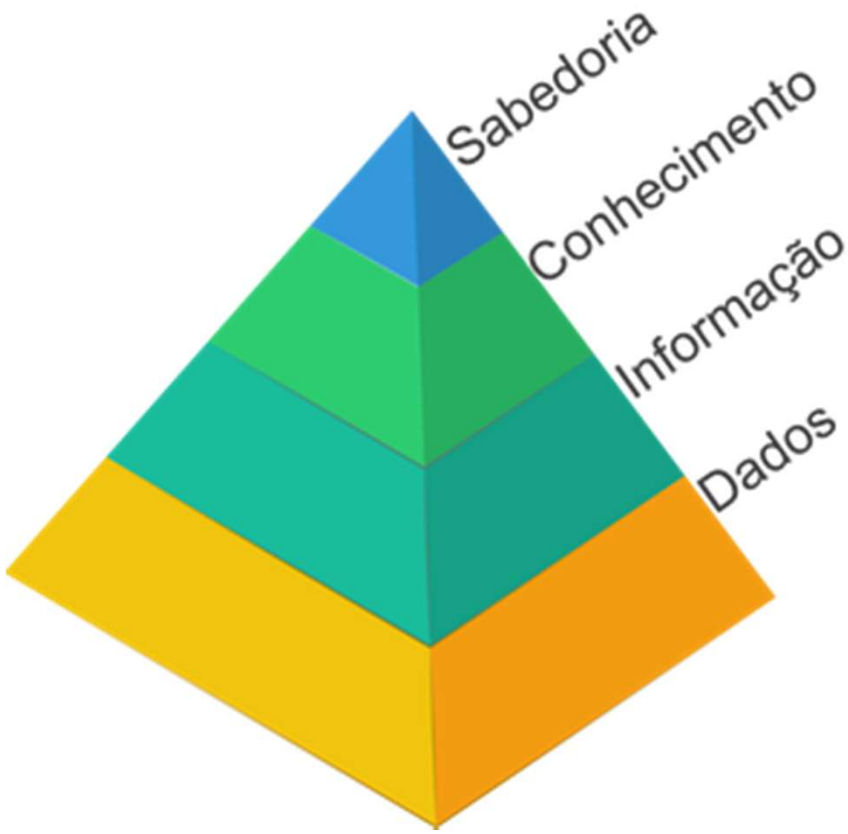
Cleilton Lima Rocha

Universidade 7 de Setembro
Fortaleza – CE, Brasil

Agenda

- Introdução ao Processo de Descoberta de Conhecimento e Data Science
- Exploração de Dados
- Feature engineering:
  - Pré-processamento de dados
    - Modelagem dos dados
    - Seleção de Features ...
- Modelos de aprendizagem supervisionada e não supervisionada
- Análise do *bias variance threshold*
- Introdução a Aprendizagem por reforço e Deep Learning
- Projeto prático aplicado à Data Science.

Processo de Descoberta de Conhecimento



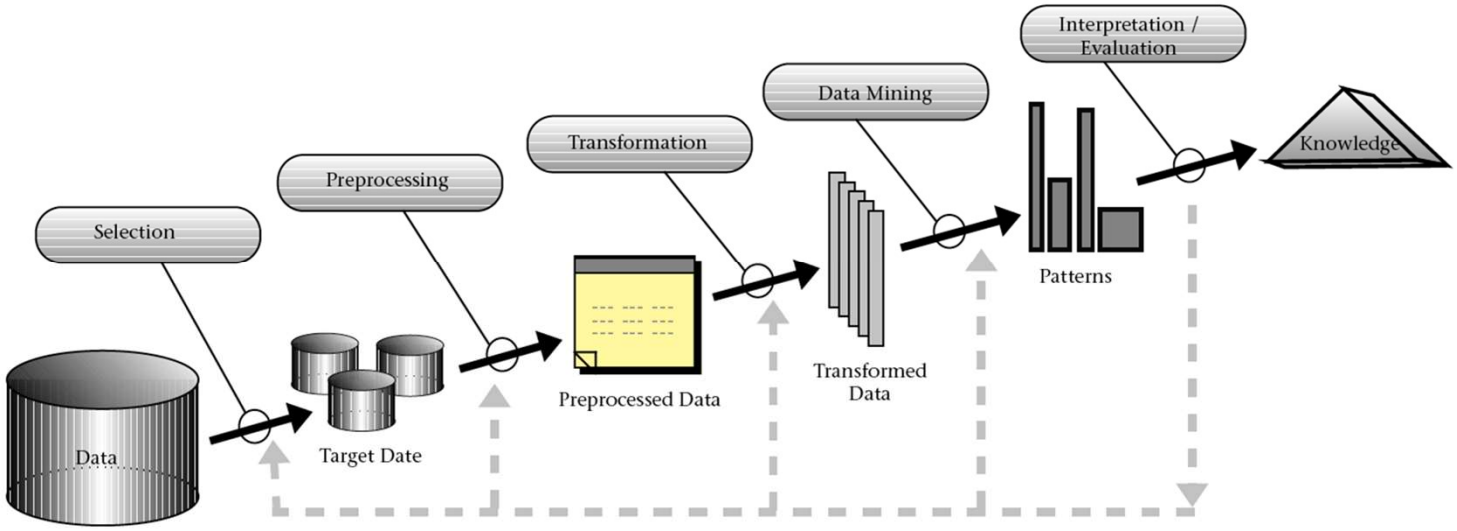
“O KDD pode ser visto como o processo de descoberta de padrões e tendências por análise de grandes conjuntos de dados, tendo como principal etapa o processo de mineração, consistindo na execução prática de análise e de algoritmos específicos que, sob limitações de eficiência computacionais aceitáveis, produz uma relação particular de padrões a partir de dados FAYYAD et al (1996).”

“Informação é o resultado do processamento de dados num formato que tem significado para o usuário respectivo e que tem valor real ou potencial nas decisões presentes ou prospectivas DAVIS (1974).”

“Segundo DAVENPORT e PRUSAK (1998), a GC pode ser vista como uma série de ações gerenciais constantes e sistemáticas que facilitam os processos de criação, registro e compartilhamento do conhecimento nas organizações.”

“O conhecimento necessário para se decidir e/ou avaliar torna-se disponível por meio de informações SANCHES (1997).”

Fases do KDD


Selection
Preprocessing
Transformation
Data Mining
Interpretation / Evaluation
Knowledge
Data
Target Date
Preprocessed Data
Transformed Data
Patterns

Data Mining e seus métodos

- Aprendizagem supervisionada
- Aprendizagem não supervisionada
- Modelos de regras de associação
- Modelos de relacionamento entre variáveis

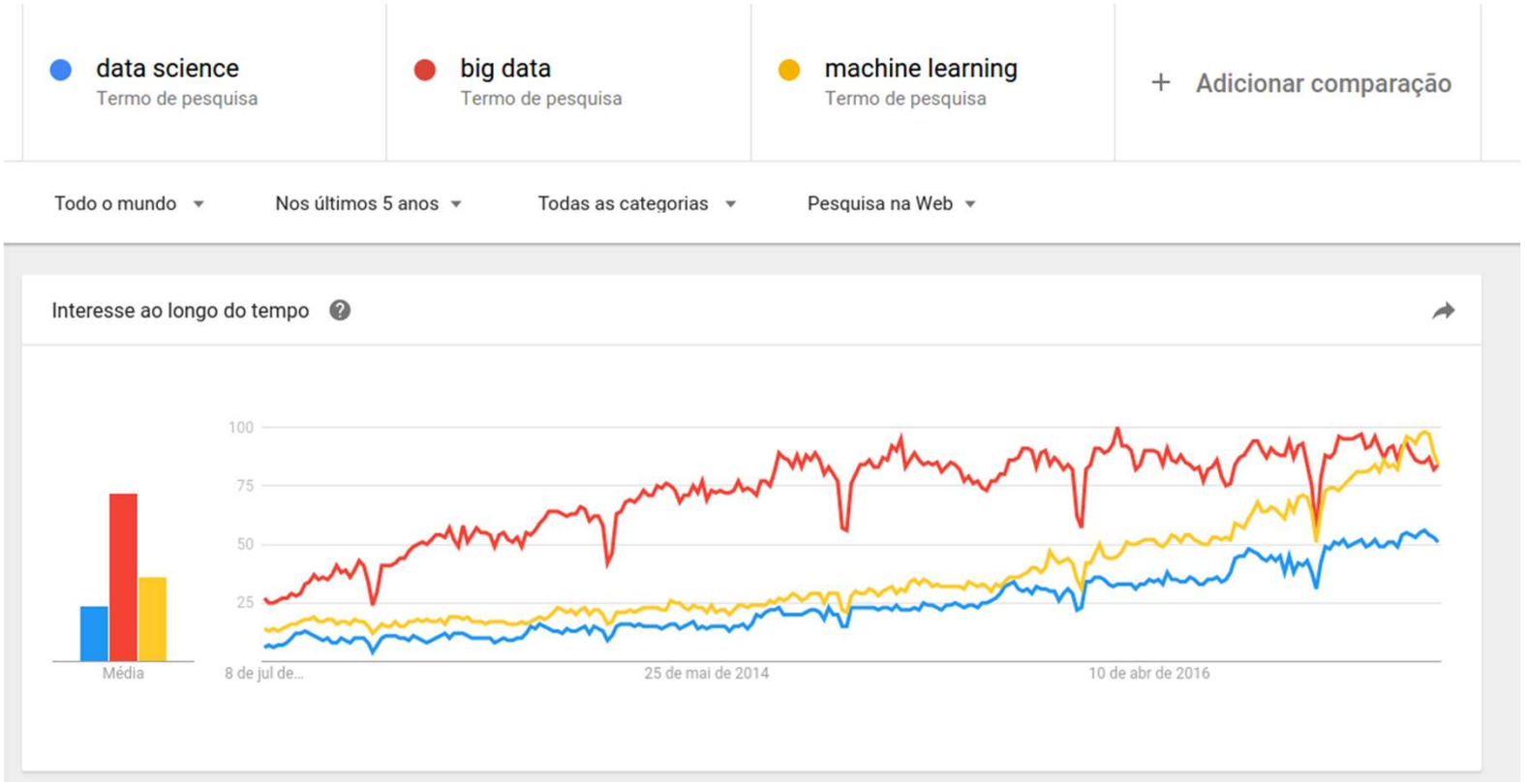
ATD

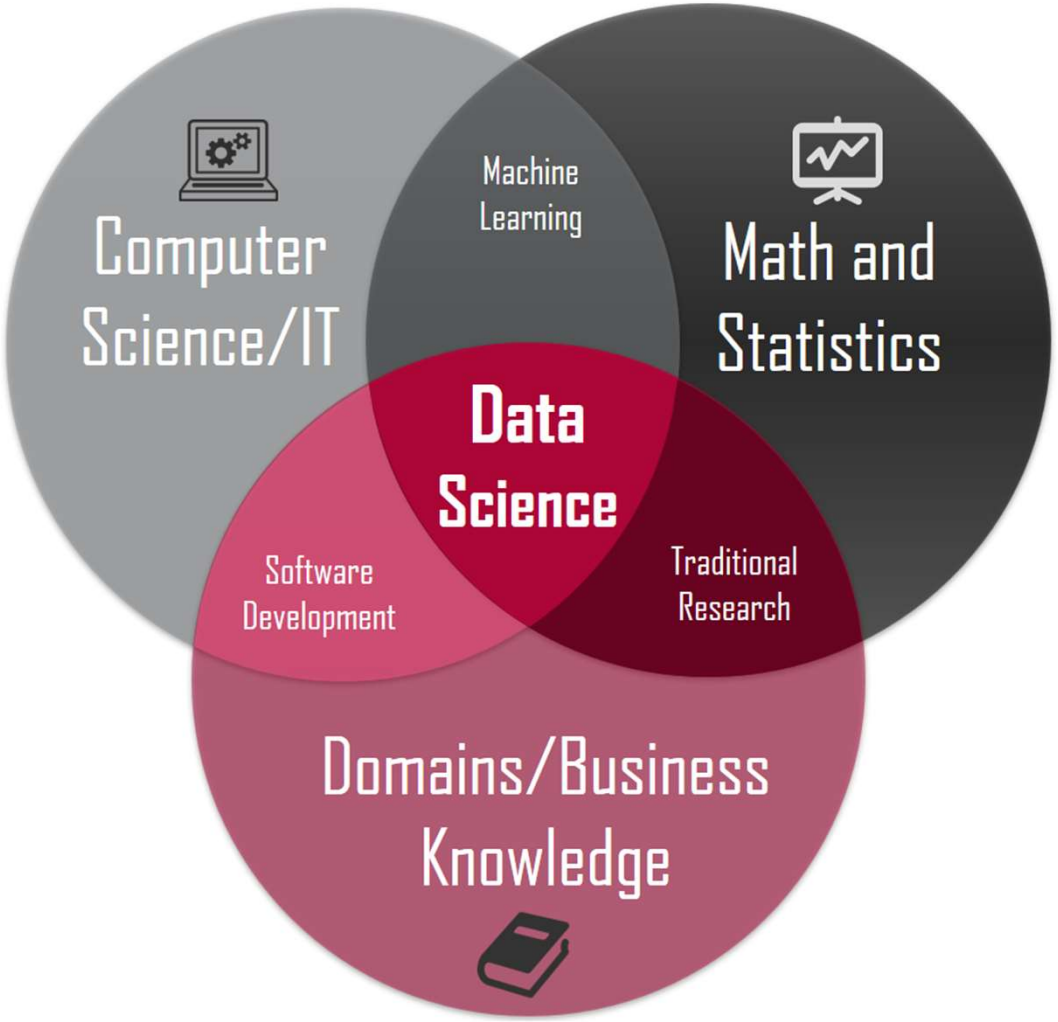
Apoio à **tomada de decisão**

Volume de dados gerados em um minuto, nos anos de 2017 e 2018

Riqueza dos Dados

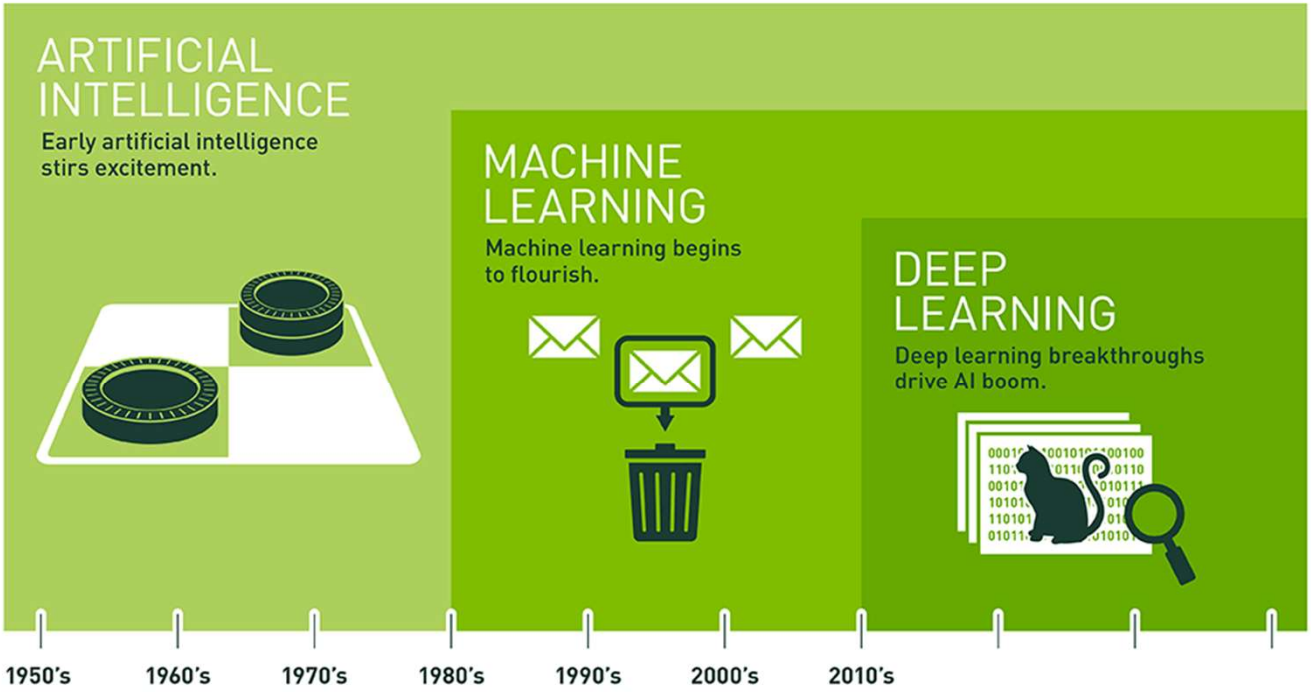
Interesse em Data Science





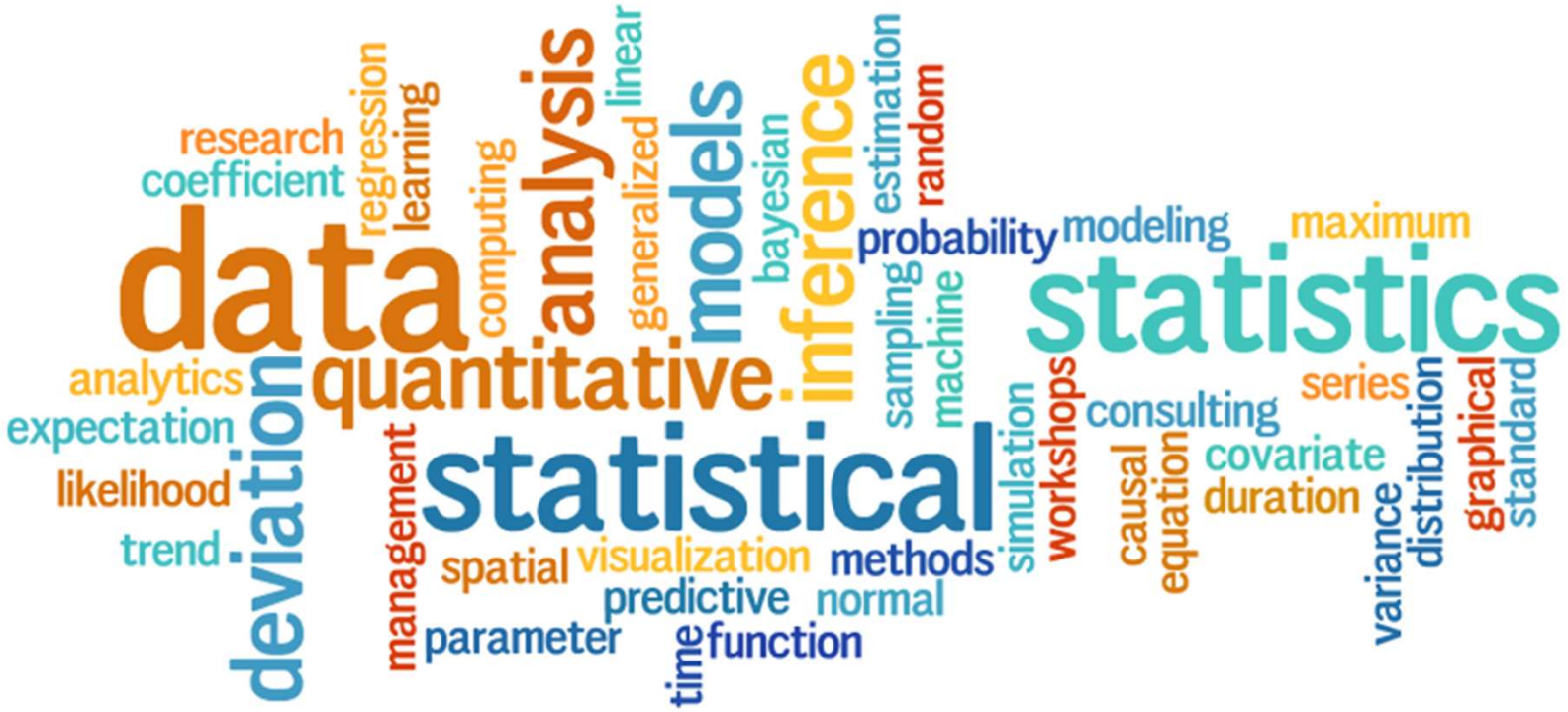
Data Science – uma ciência interdisciplinar

Machine Learning Overview



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

Data Science Overview



Exemplos

GISSA
Inteligência em Saúde

OCARIoT

NETFLIX

amazon

Recommendation Systems

TARGET

Walmart

Inventory planning

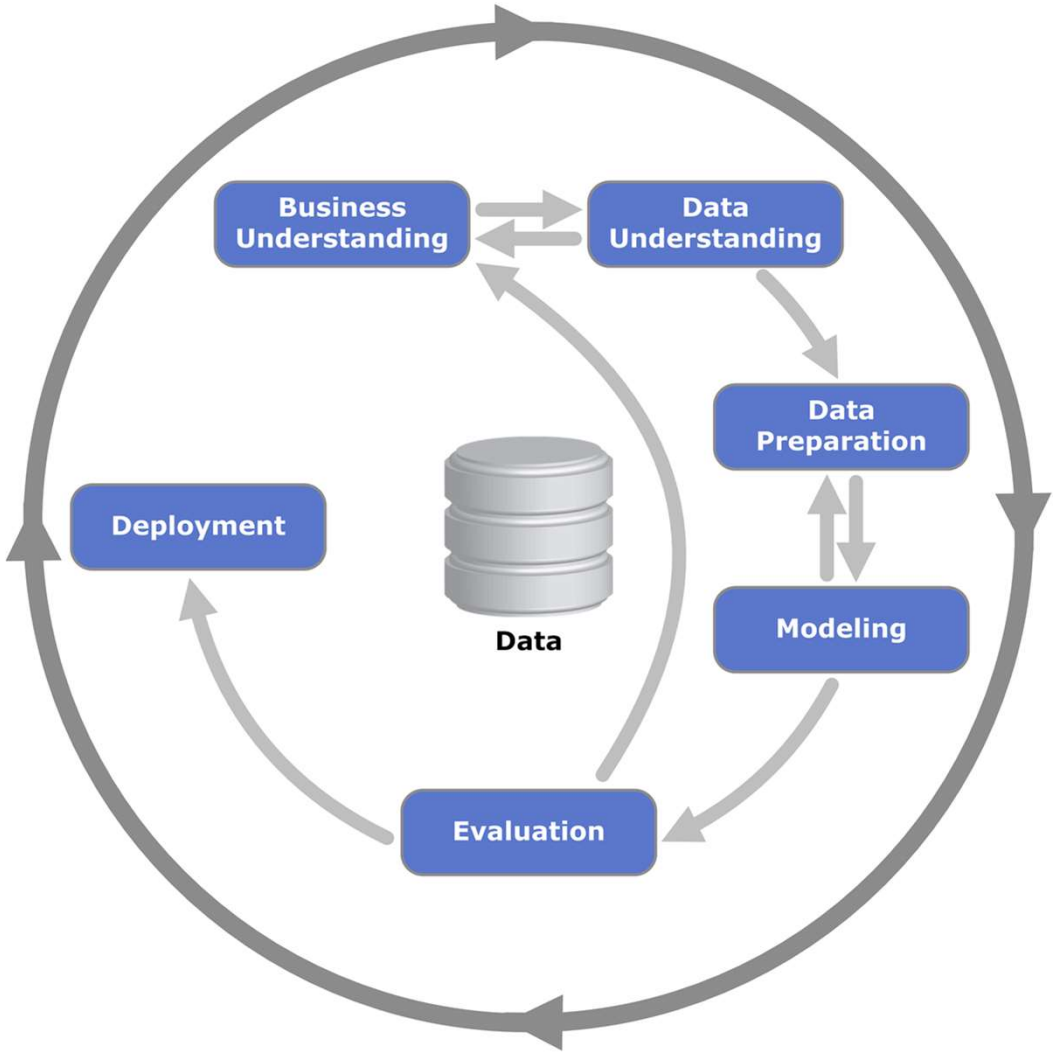
UBER

Dynamic pricing

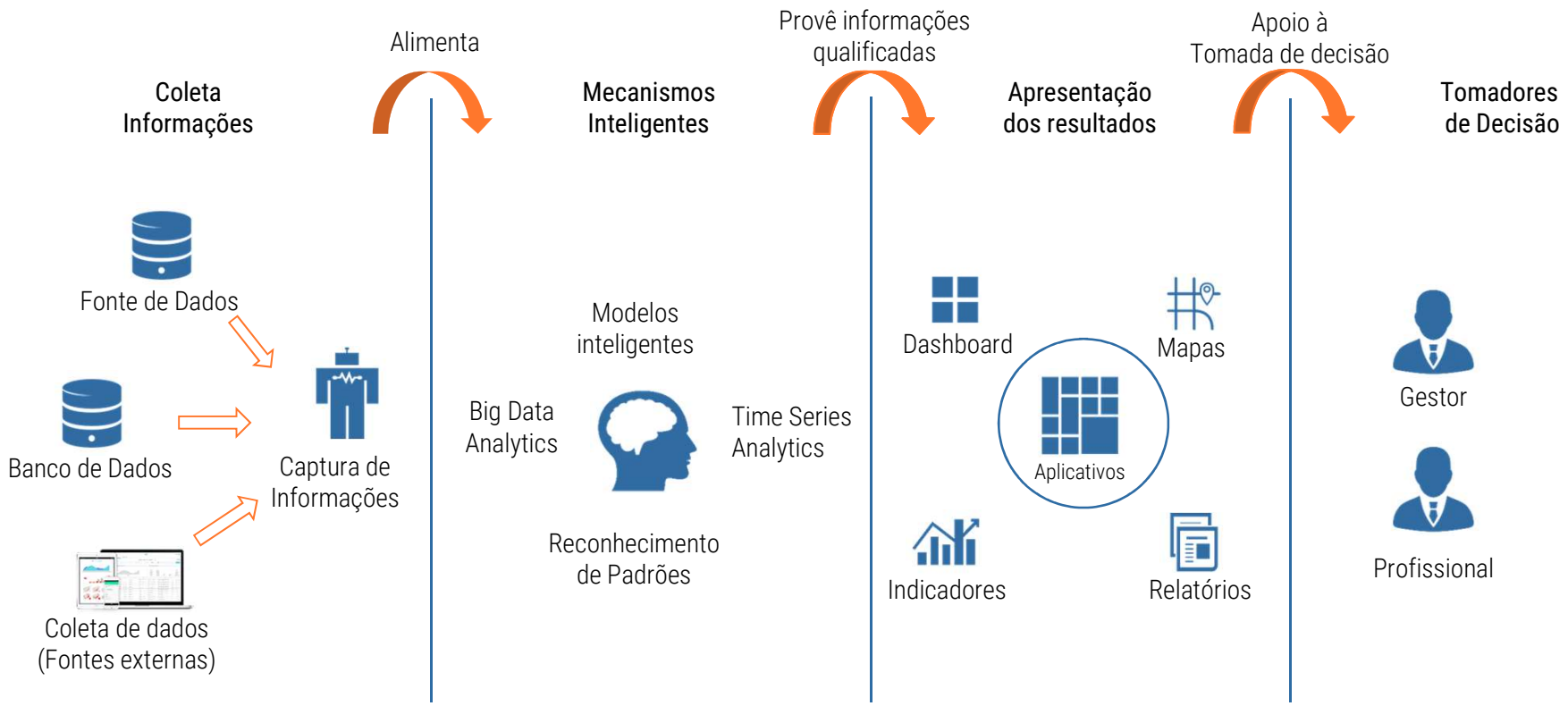
© Manish Kurse, 2016

5

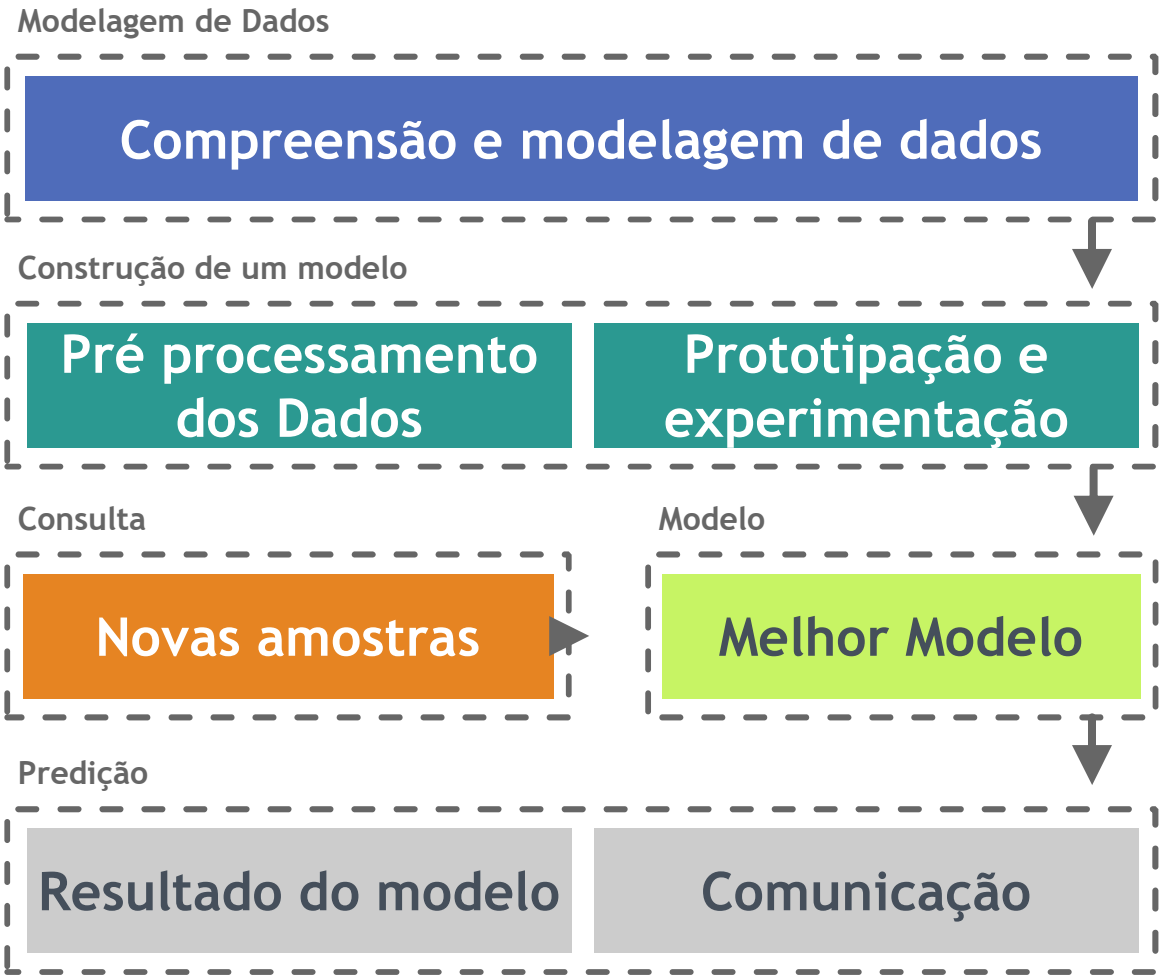
CRISP-DM


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

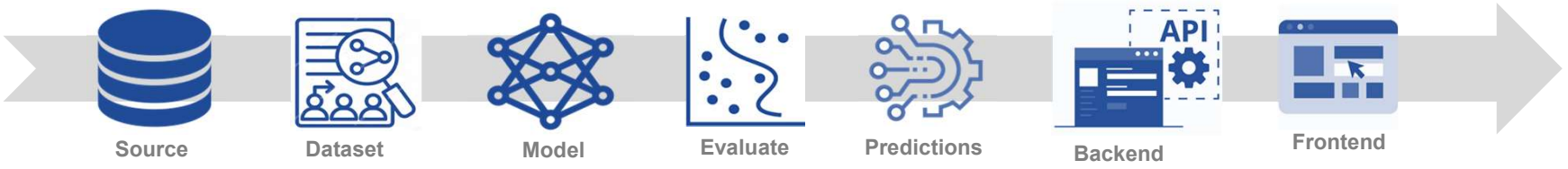
Metodologia geral adotada



Metodologia

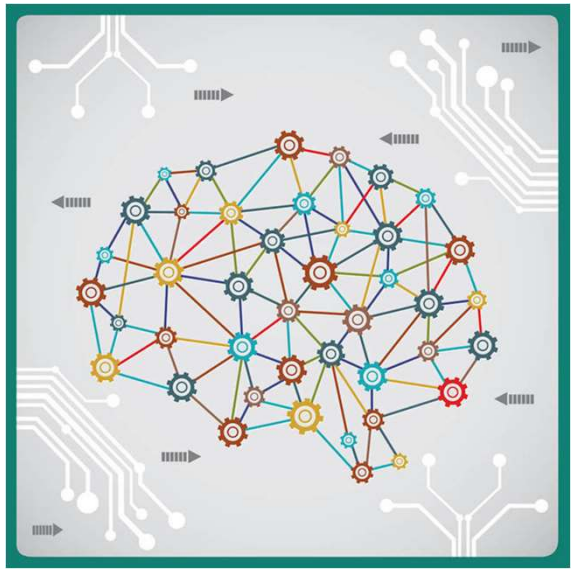


Exemplo de uma solução end-to-end

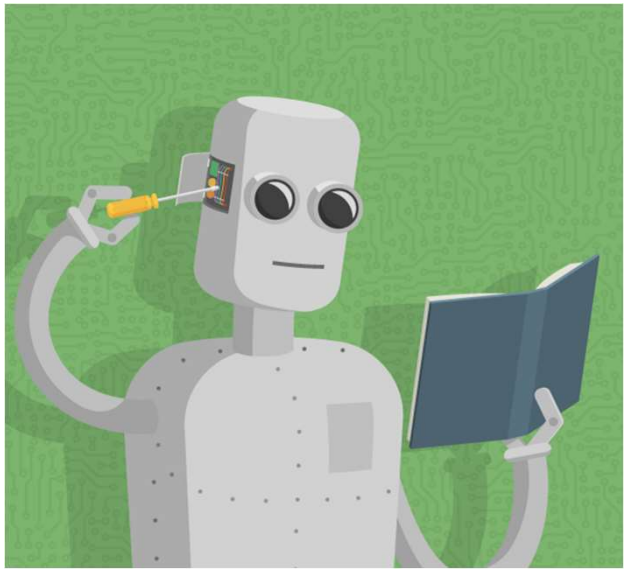


Conhecimentos desejáveis

Business Intelligence



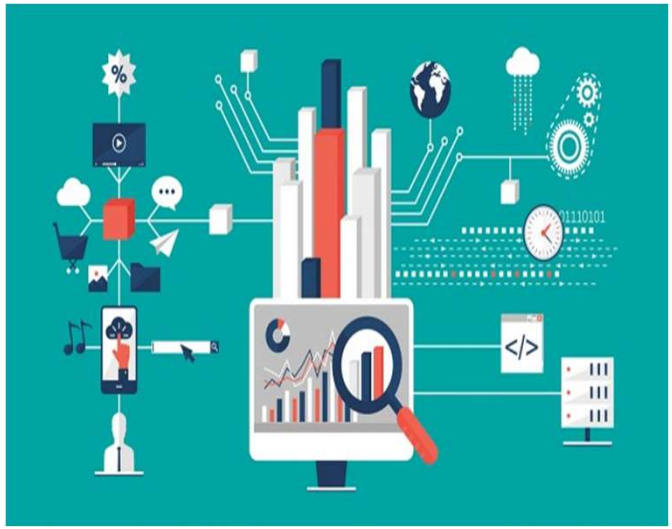
Gestão do conhecimento



Grafos e Machine Learning

Conhecimentos desejáveis

Big Data



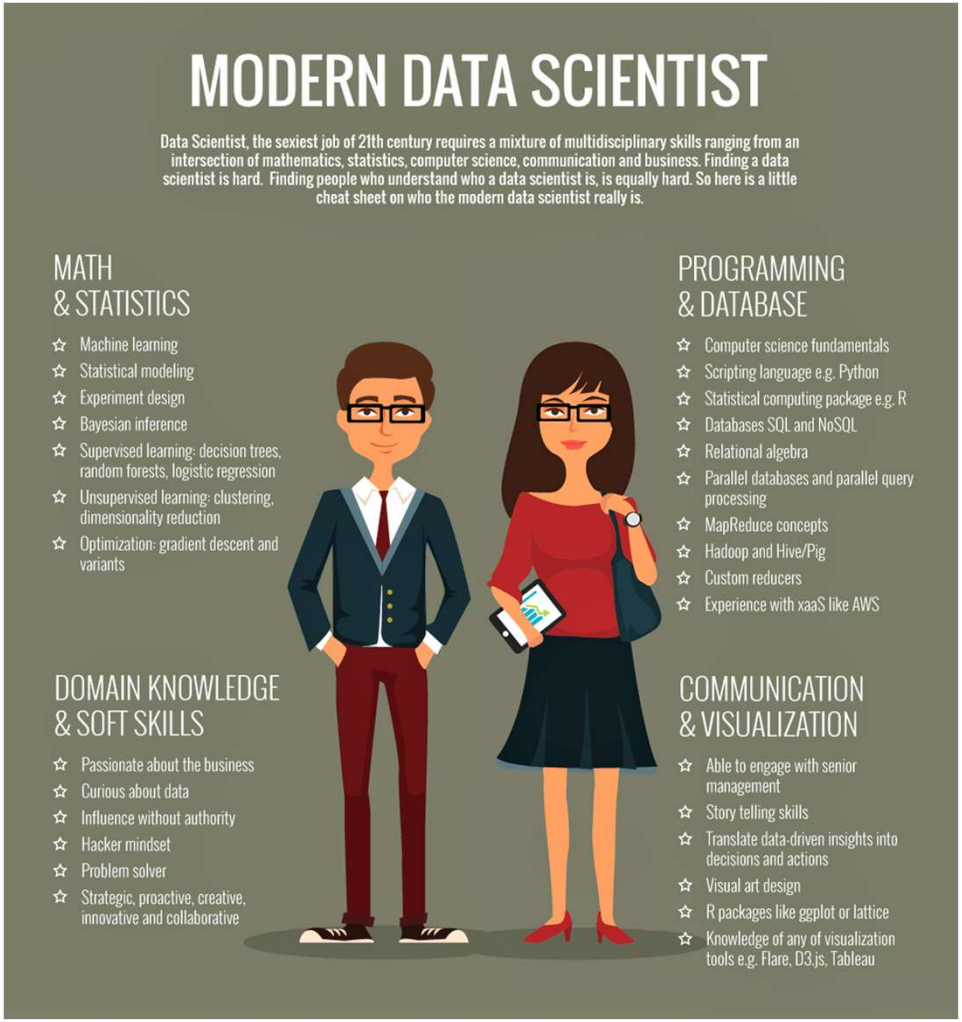
Processamento de stream e séries temporais

Processamento de Linguagem Natural

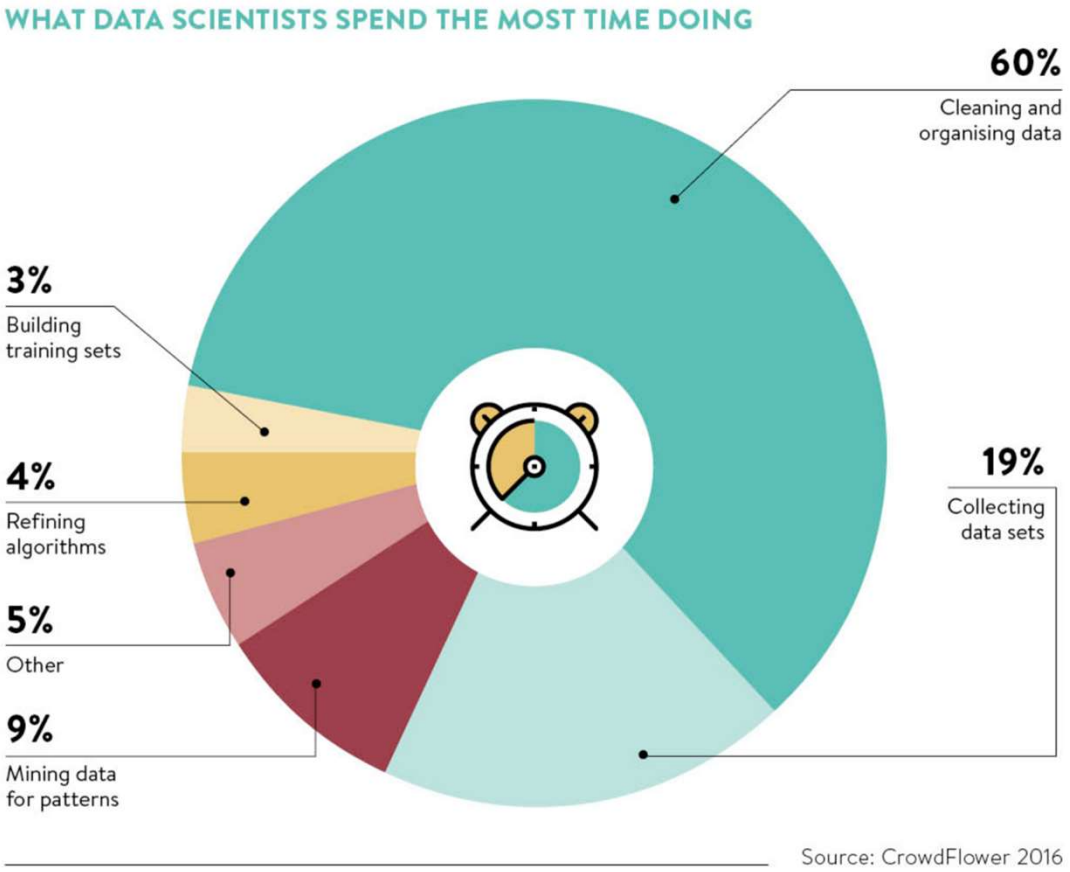
O que um cientista de dados faz?

Conhecimentos desejáveis

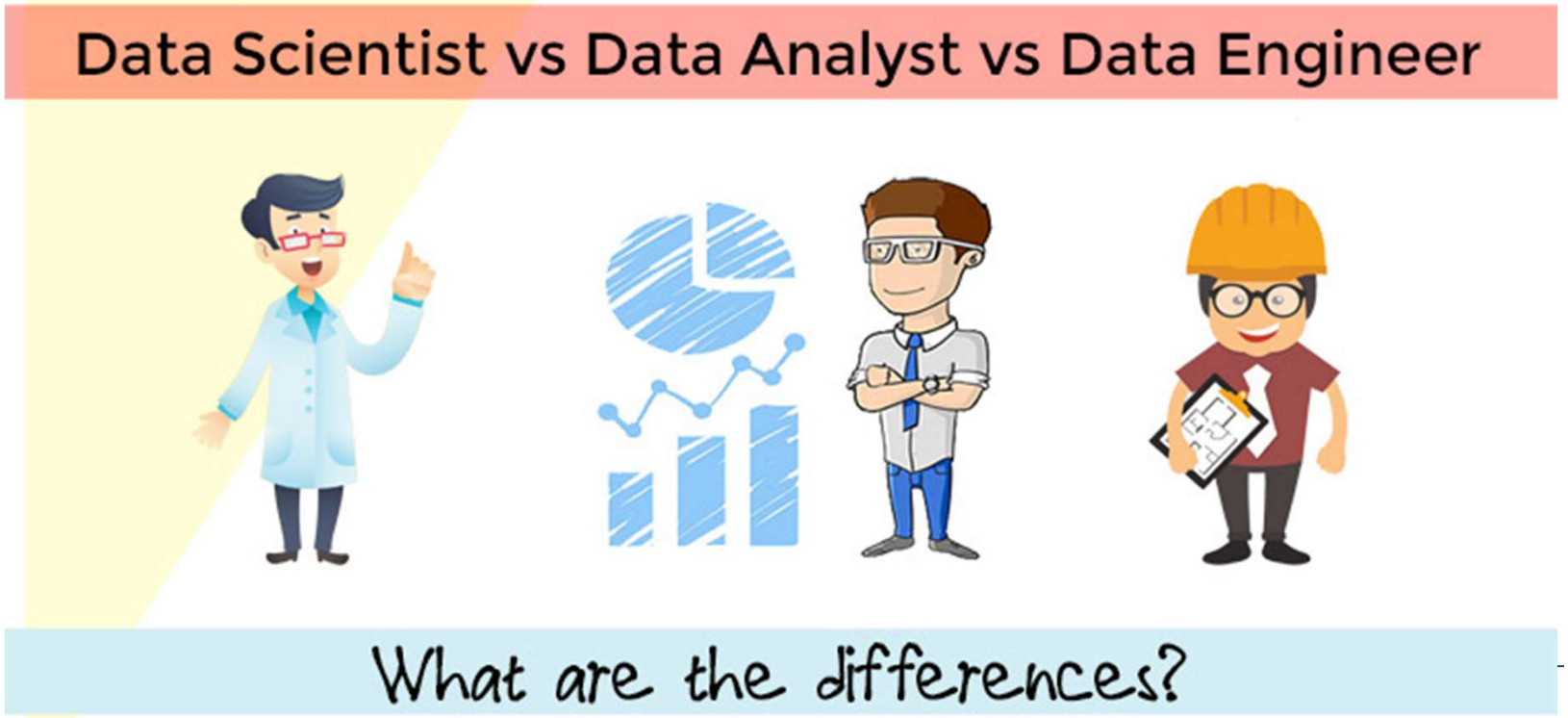
Cientista de Dados



Cientista de Dados

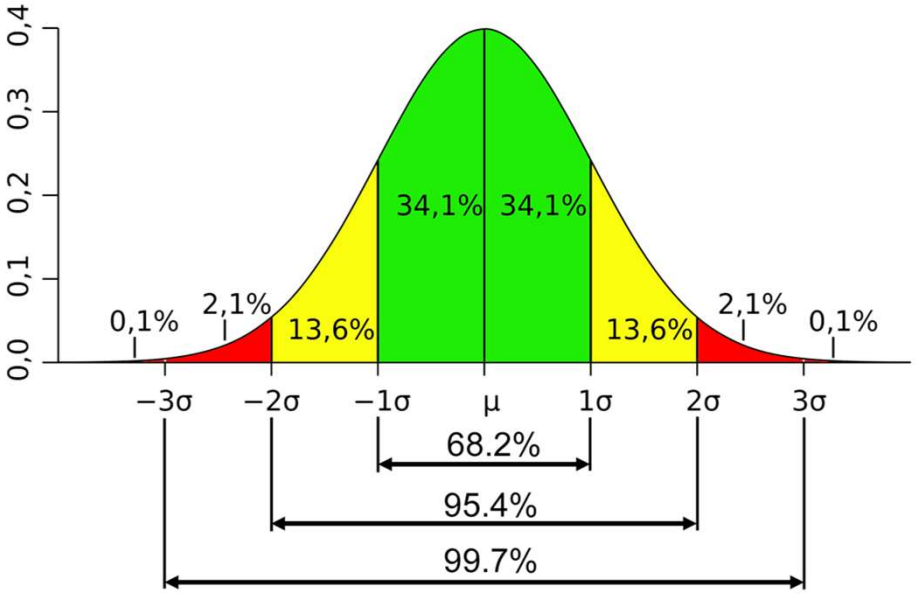


Definição de Especialistas



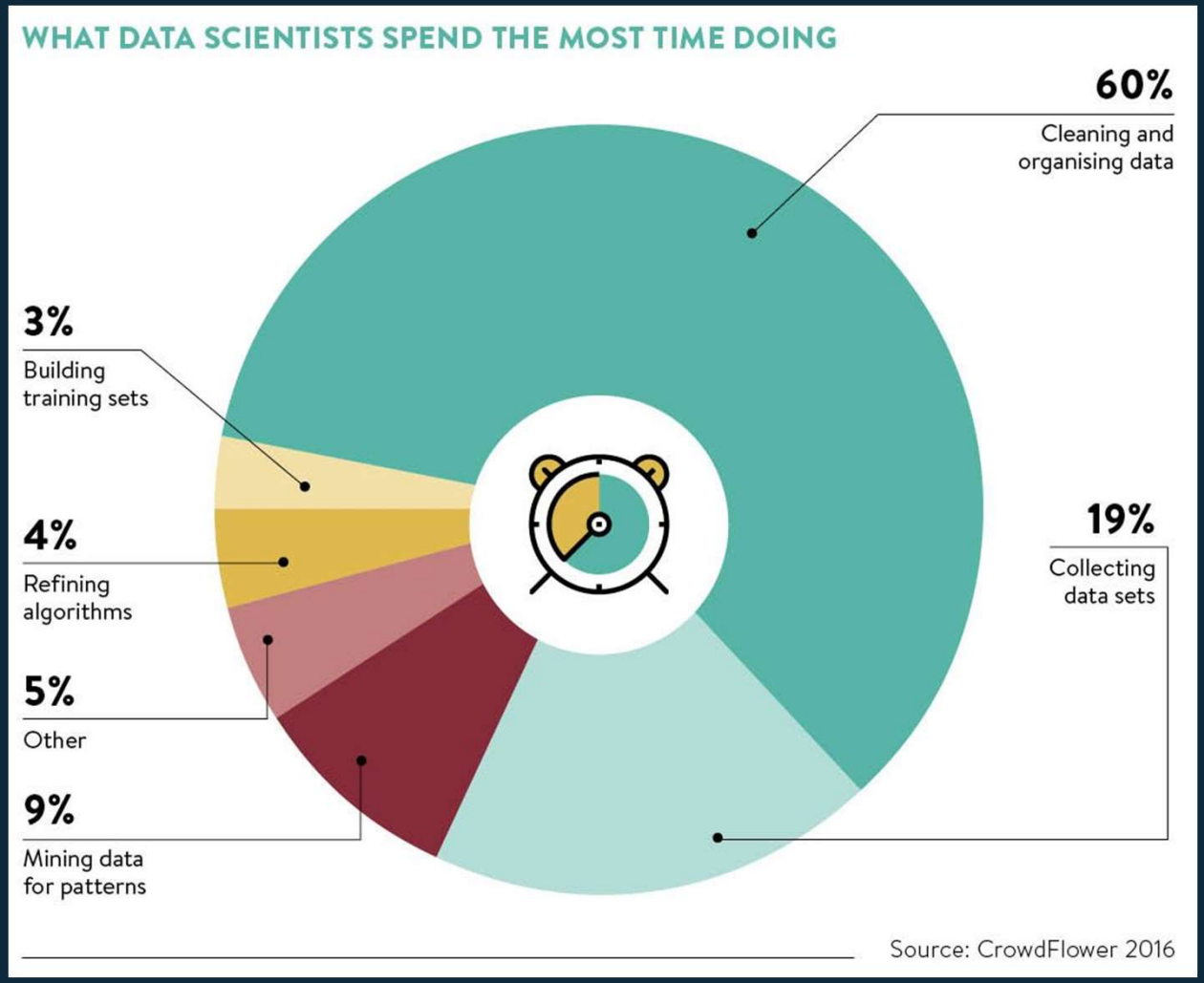
Distribuição Normal

Com a curva normal definida, temos informações importantes sobre a distribuição dos nossos dados:

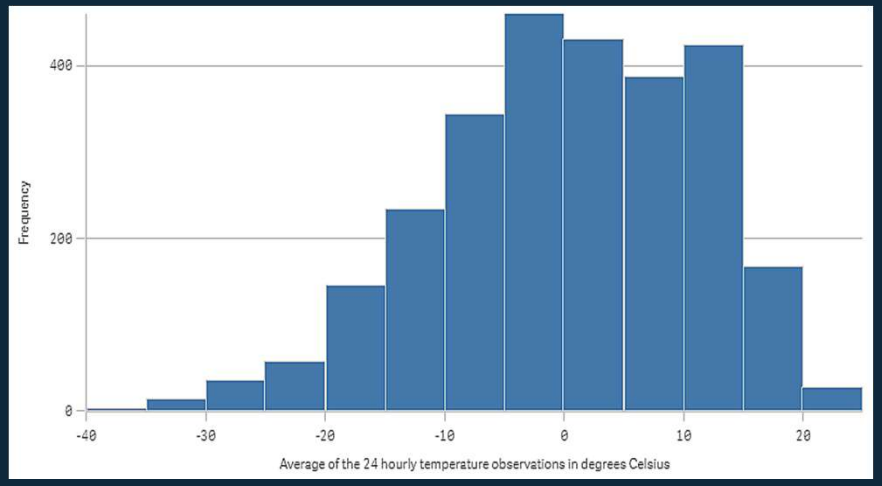


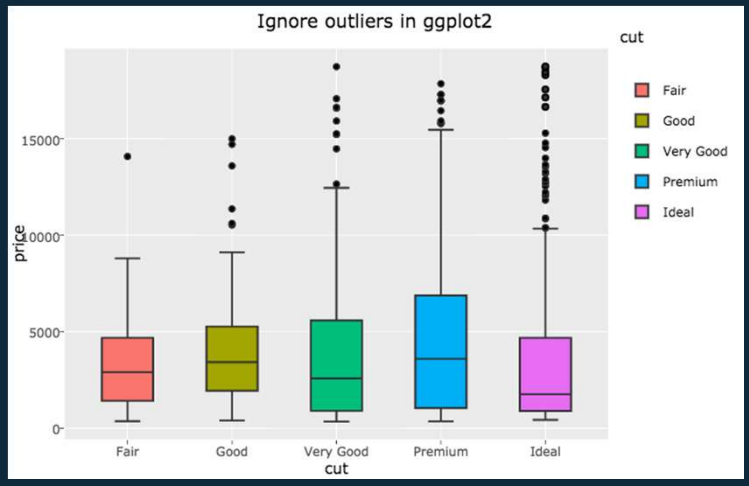
Intervalo	Proporção
μ +- 1σ	68.2%
μ +- 2σ	95.4%
μ +- 3σ	99.7%

Exploração de Dados (EDA)

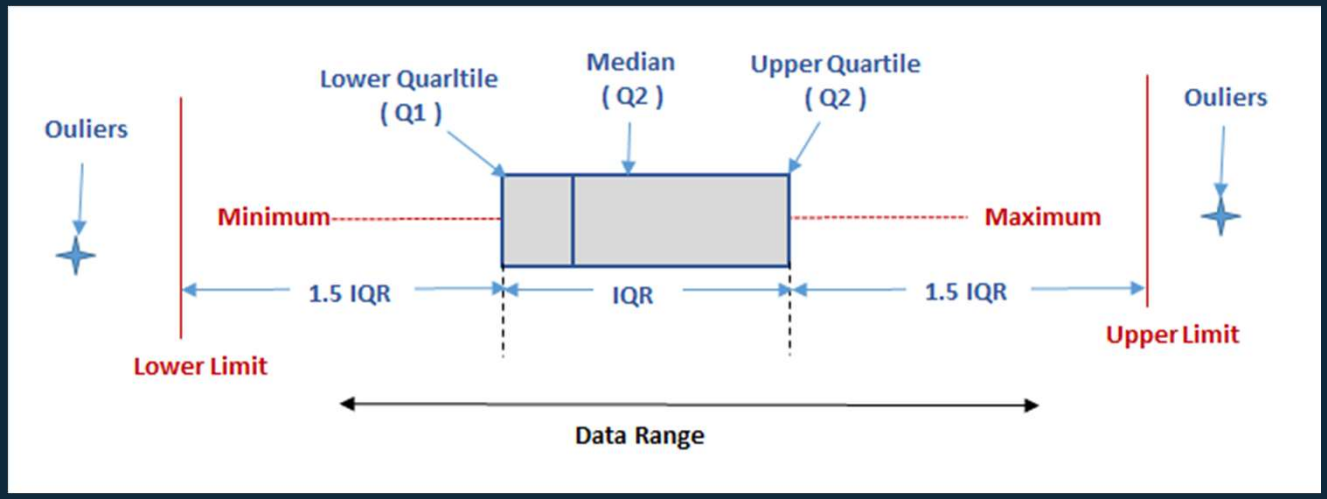

WHAT DATA SCIENTISTS SPEND THE MOST TIME DOING
60%
Cleaning and organising data
3%
Building training sets
4%
Refining algorithms
5%
Other
9%
Mining data for patterns
19%
Collecting data sets
Source: CrowdFlower 2016

Exploração do dado





Box Plot


Ouliers
Lower Quarltile
(Q1)
Median
(Q2)
Upper Quartile
(Q2)
Ouliers
Minimum
Maximum
1.5 IQR
IQR
1.5 IQR
Lower Limit
Upper Limit
Data Range

Hands-On