



Descoberta do Conhecimento





Descoberta do Conhecimento

Cleilton Lima Rocha

Universidade 7 de Setembro
Fortaleza – CE, Brasil



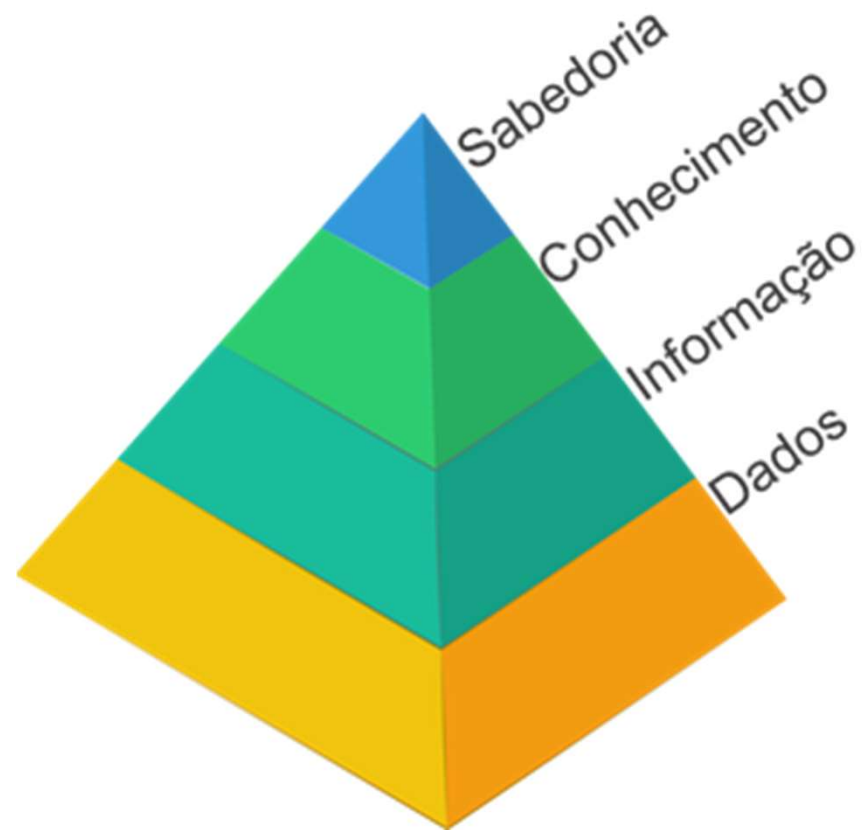


Agenda

- ◇ 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).”

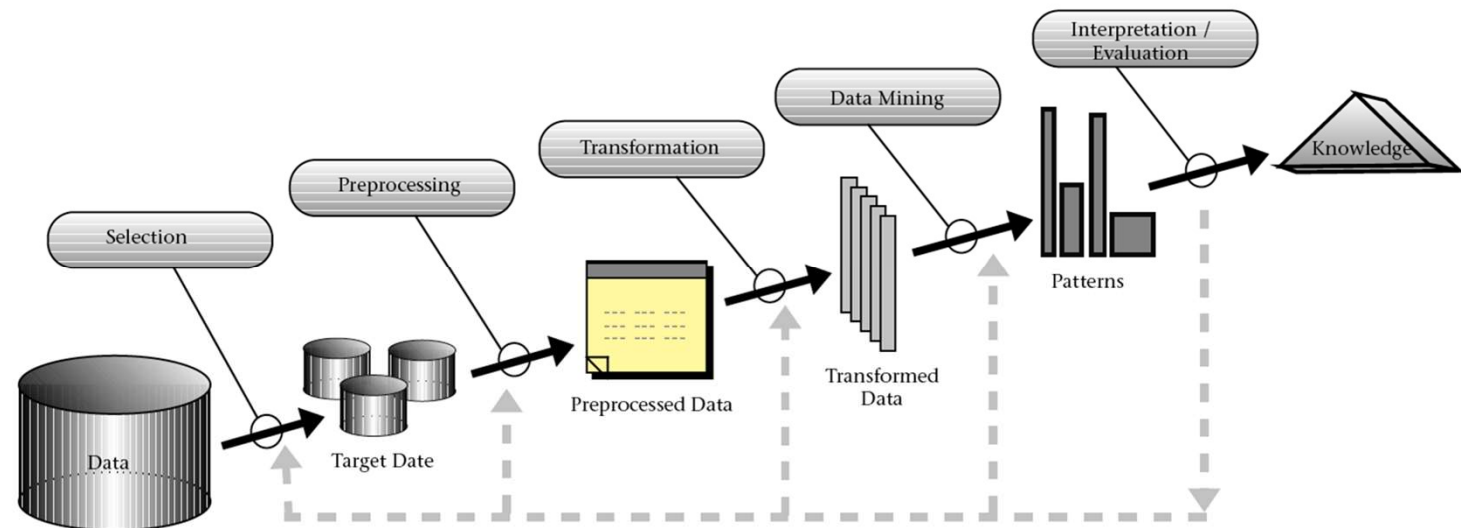


“Segundo DAVENPORT e PRUSAK (1998), a GC pode ser vista como uma série de ações gerenciais constantes e sistemáticas que facilitam os processos de criação, registro e compartilhamento do conhecimento nas organizações.”



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

Fases do KDD



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





2017 *This Is What Happens In An Internet Minute*



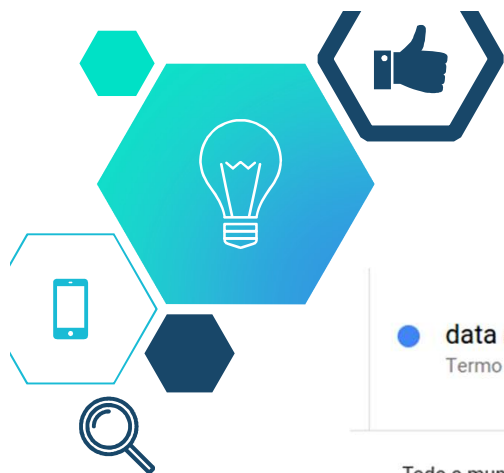
2018 *This Is What Happens In An Internet Minute*





Riqueza dos Dados





Interesse em Data Science

● **data science**
Termo de pesquisa

● **big data**
Termo de pesquisa

● **machine learning**
Termo de pesquisa

+ Adicionar comparação

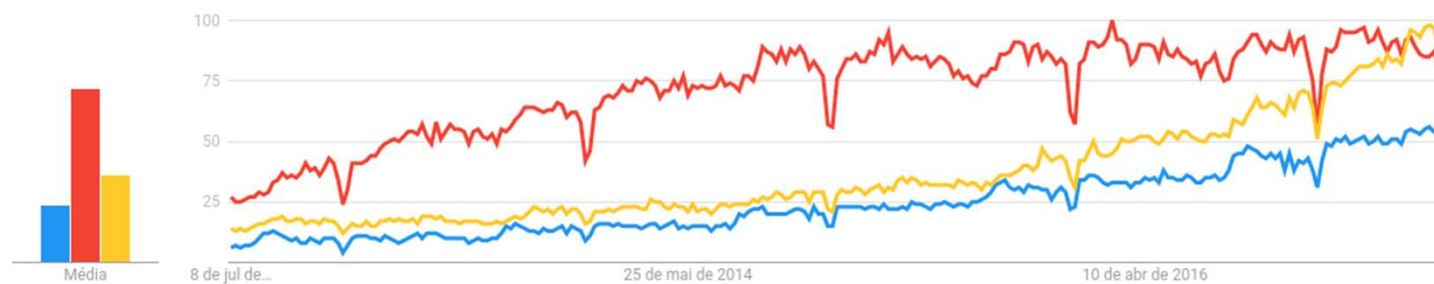
Todo o mundo ▼

Nos últimos 5 anos ▼

Todas as categorias ▼

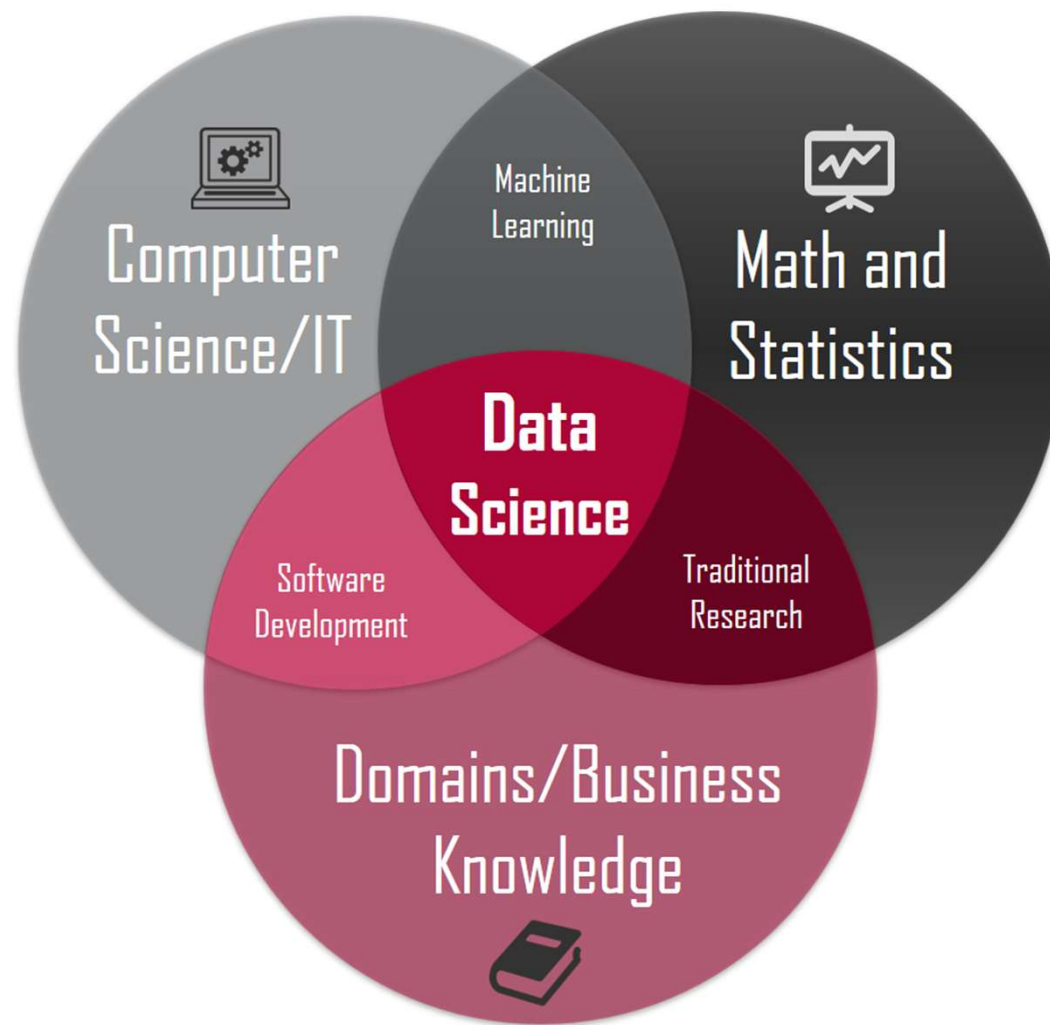
Pesquisa na Web ▼

Interesse ao longo do tempo ?



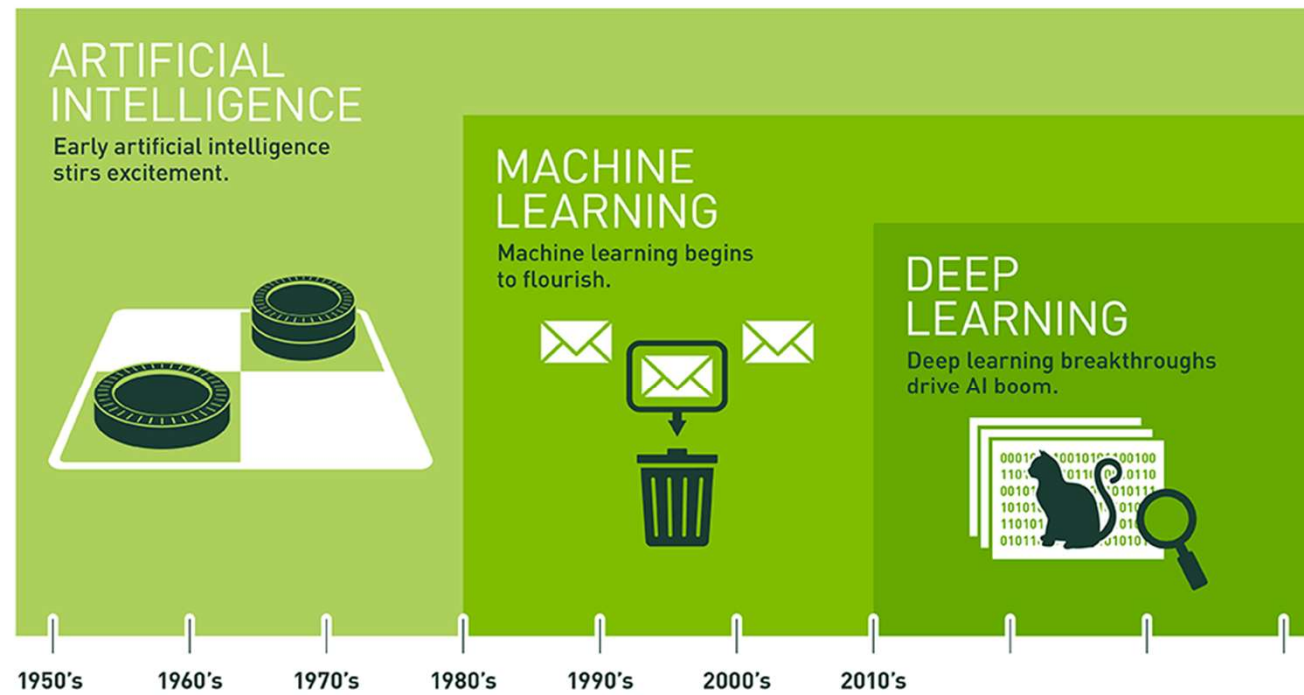


Data Science – uma ciência interdisciplinar





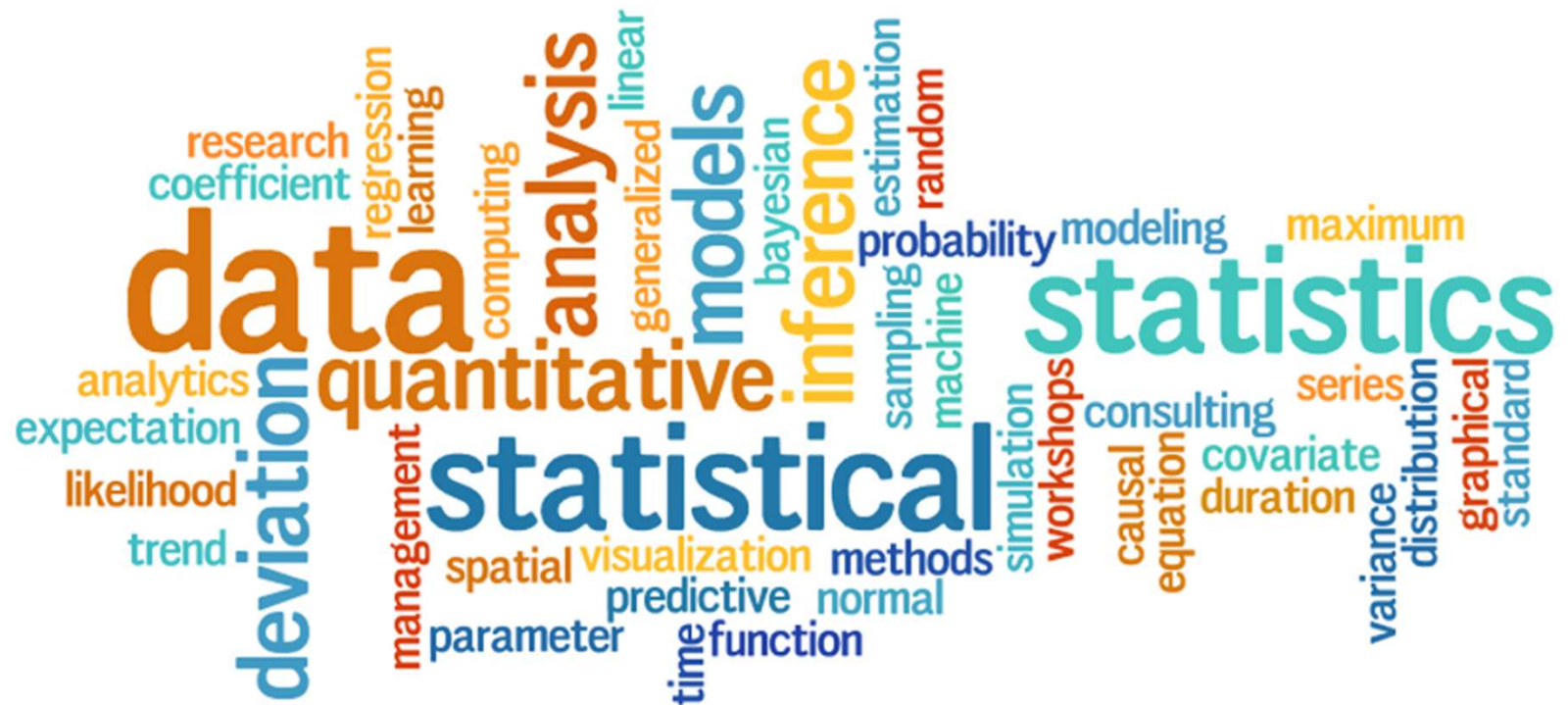
Machine Learning Overview



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



Recommendation Systems



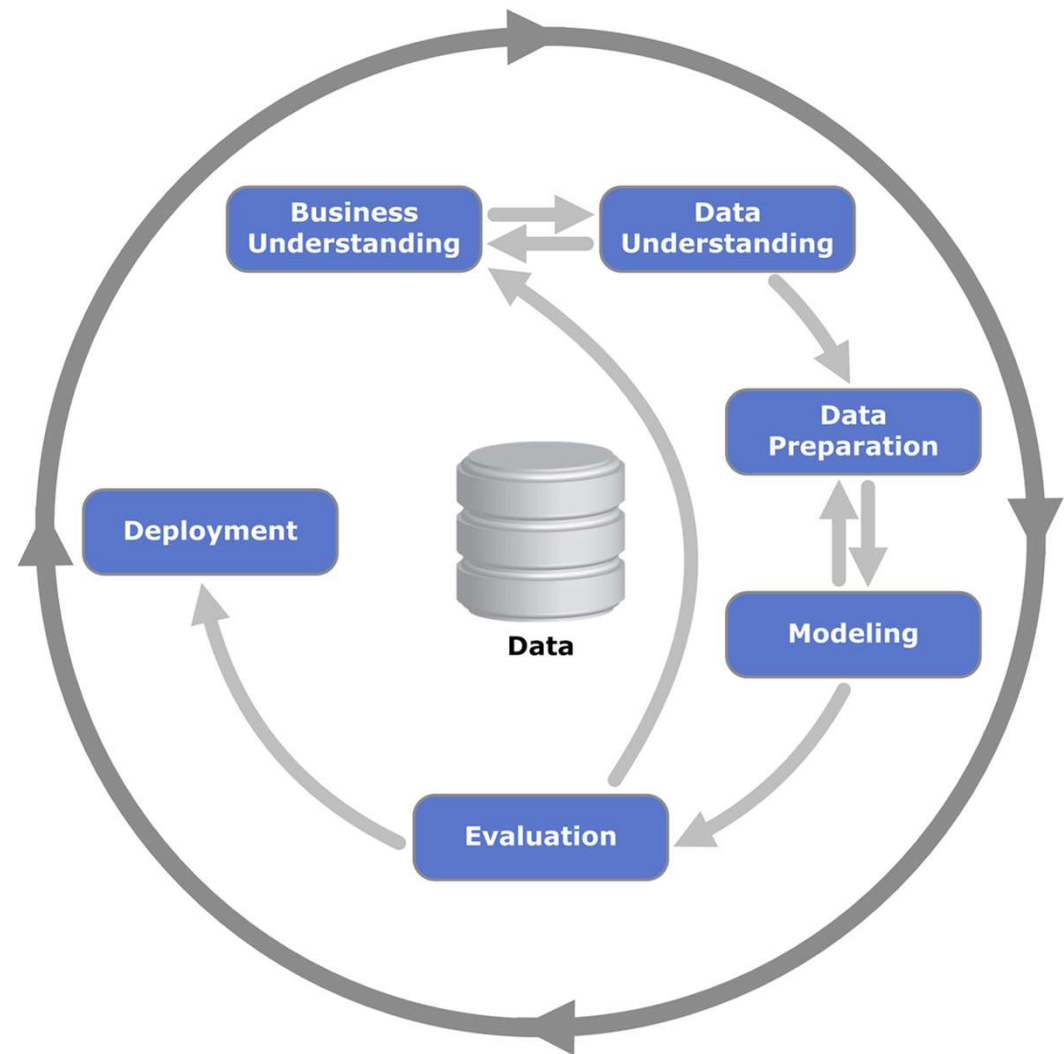
Inventory planning



Dynamic pricing

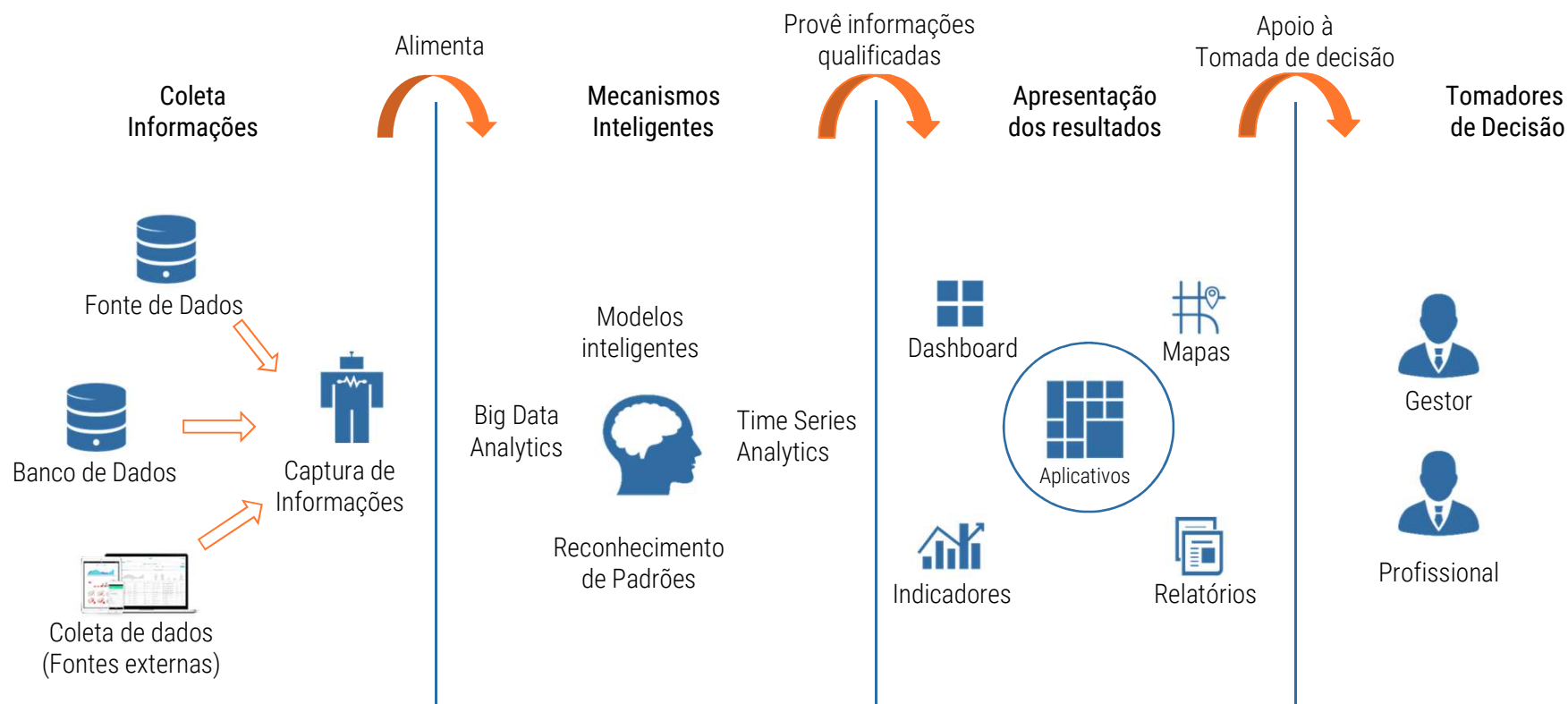


CRISP-DM





Metodologia geral adotada





Metodologia

Modelagem de Dados

Compreensão e modelagem de dados

Construção de um modelo

Pré processamento dos Dados

Prototipação e experimentação

Consulta

Novas amostras

Modelo

Melhor Modelo

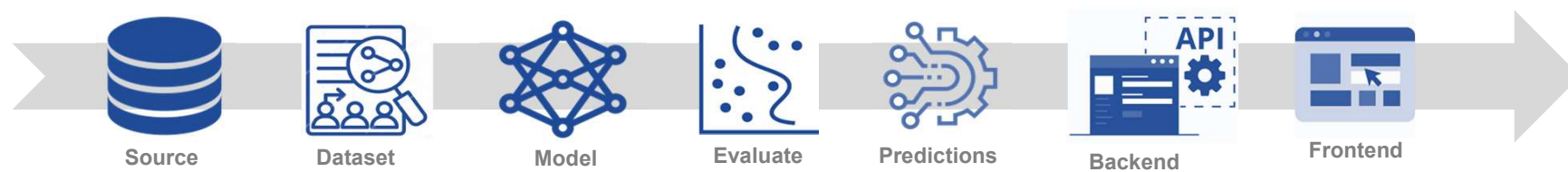
Predição

Resultado do modelo

Comunicação



Exemplo de uma solução end-to-end

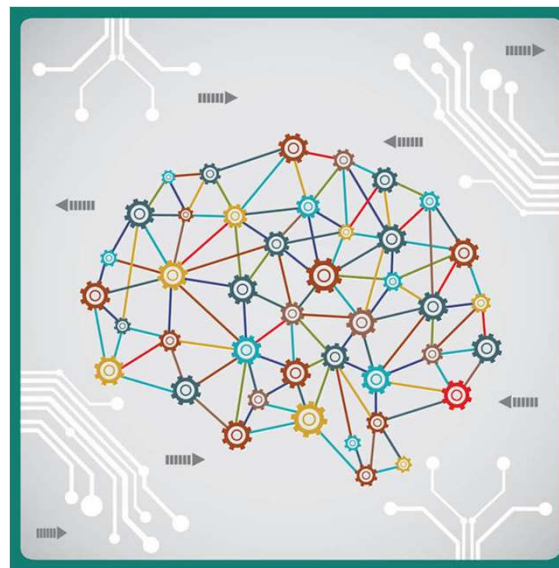




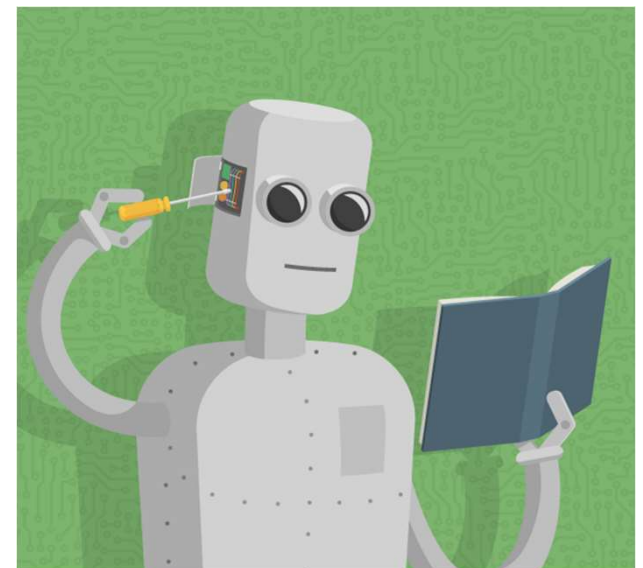
Conhecimentos desejáveis



Business Intelligence



Gestão do conhecimento



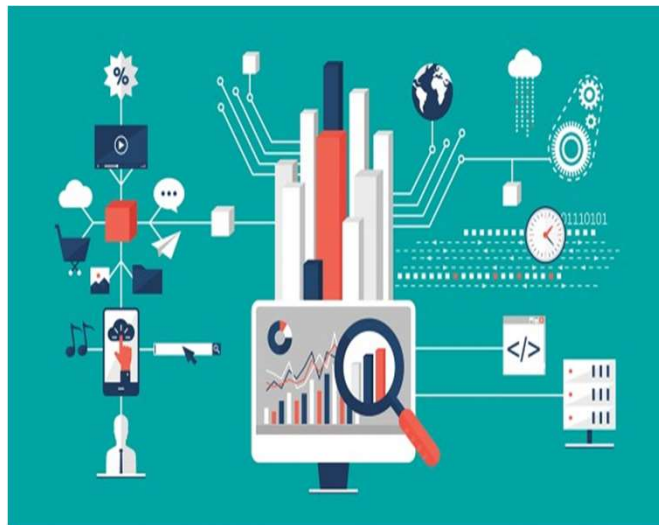
Grafos e Machine Learning



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

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21st century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing package e.g. R
- ☆ Databases SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

COMMUNICATION & VISUALIZATION

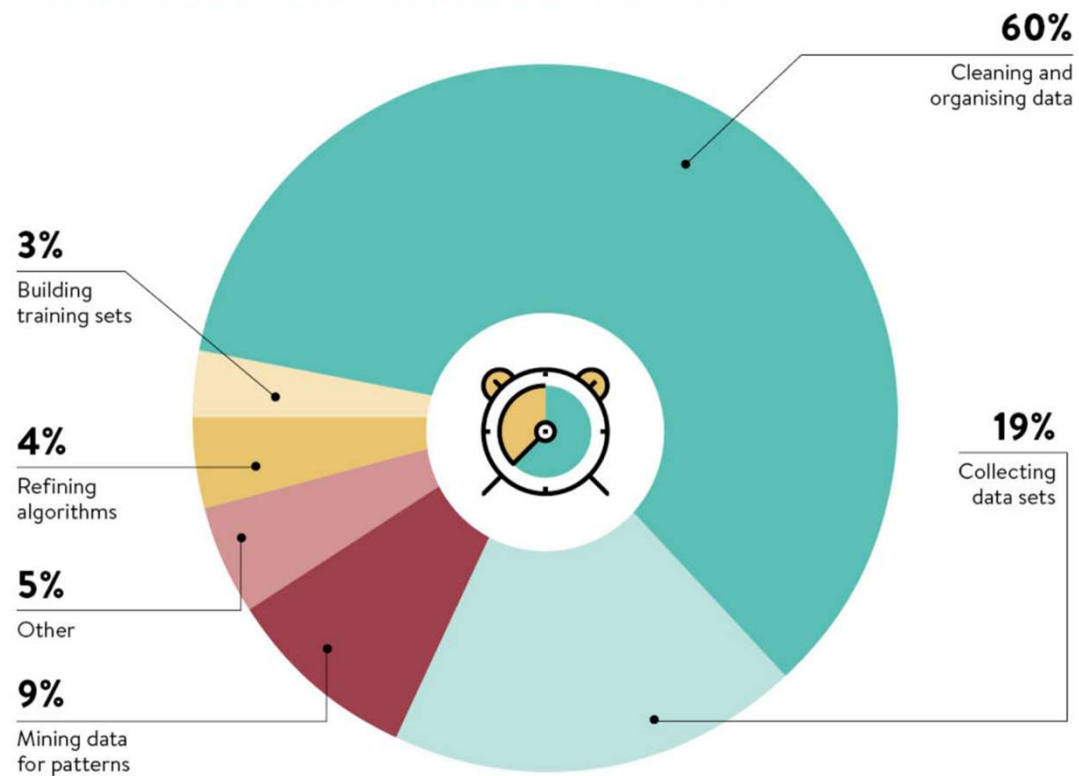
- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau





Cientista de Dados

WHAT DATA SCIENTISTS SPEND THE MOST TIME DOING



Source: CrowdFlower 2016



Definição de Especialistas

Data Scientist vs Data Analyst vs Data Engineer

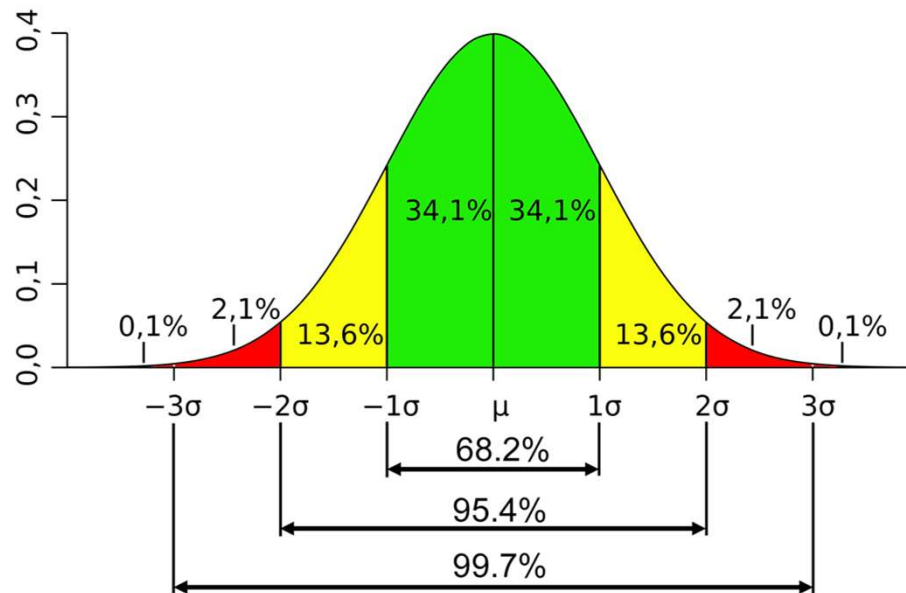


What are the differences?



Distribuição Normal

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



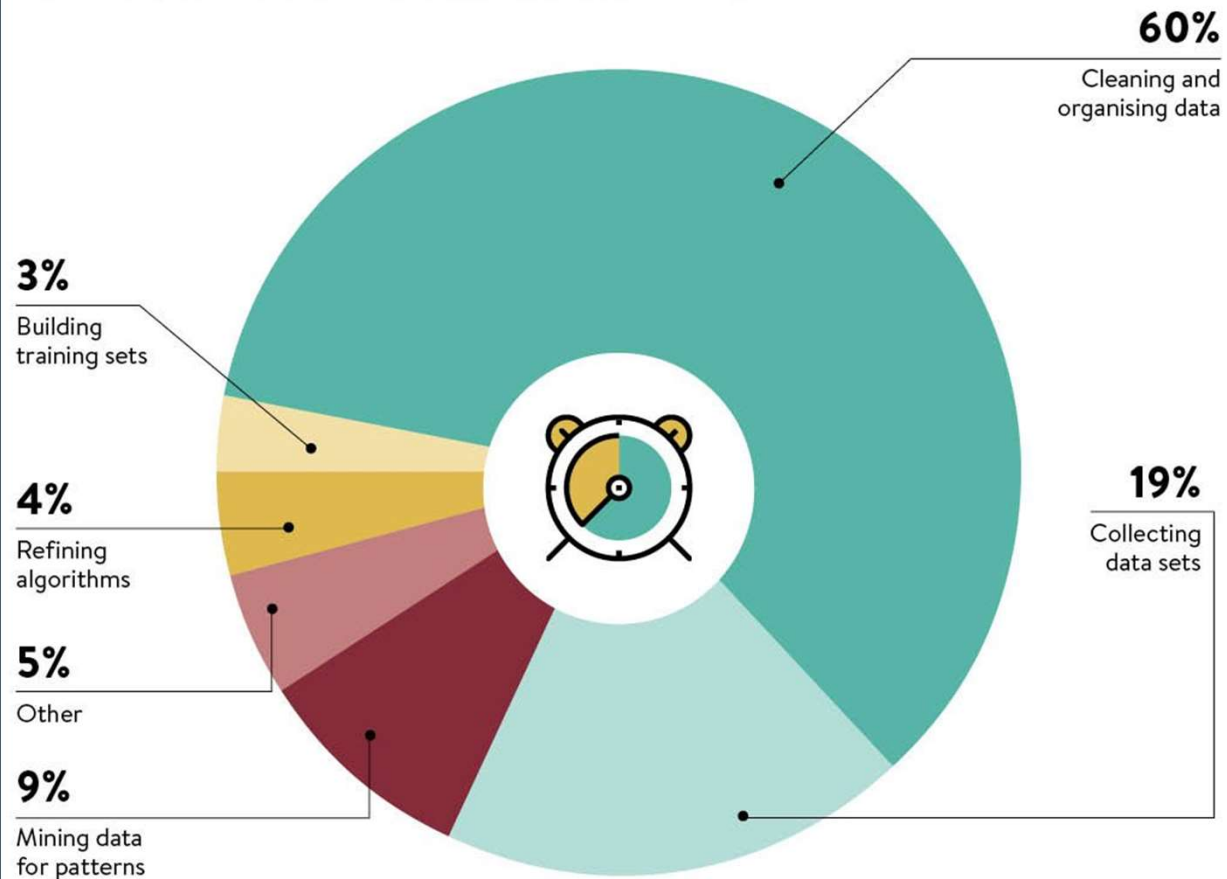
Intervalo	Proporção
$\mu \pm 1\sigma$	68.2%
$\mu \pm 2\sigma$	95.4%
$\mu \pm 3\sigma$	99.7%



Exploração de Dados (EDA)

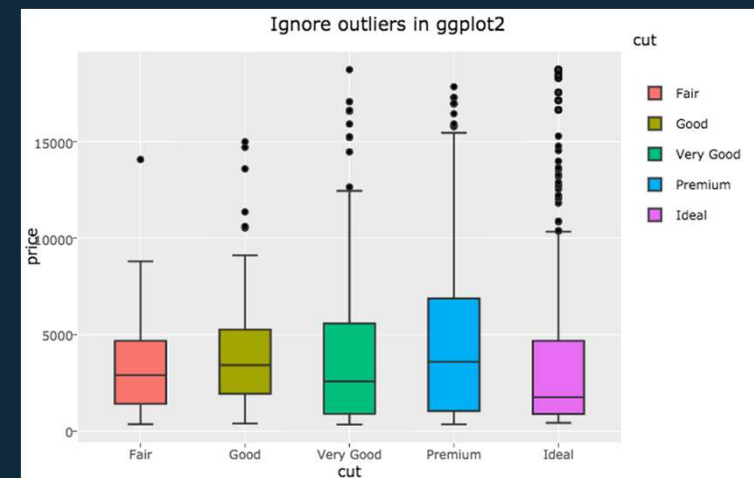
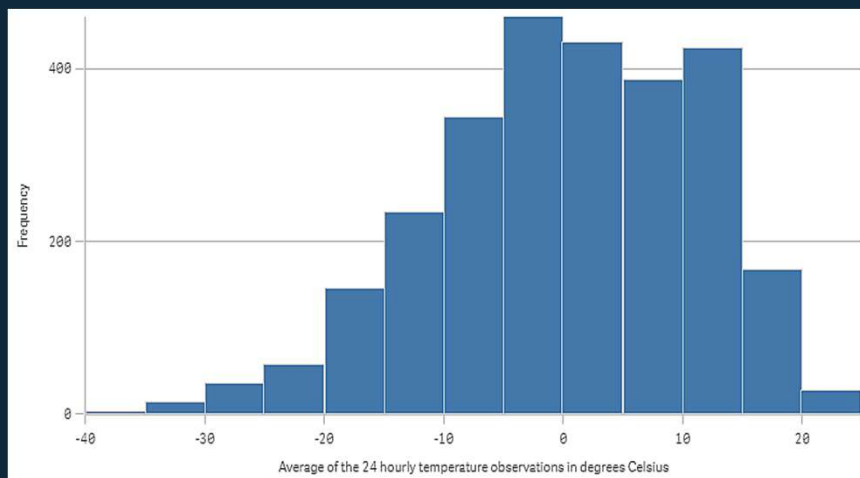


WHAT DATA SCIENTISTS SPEND THE MOST TIME DOING

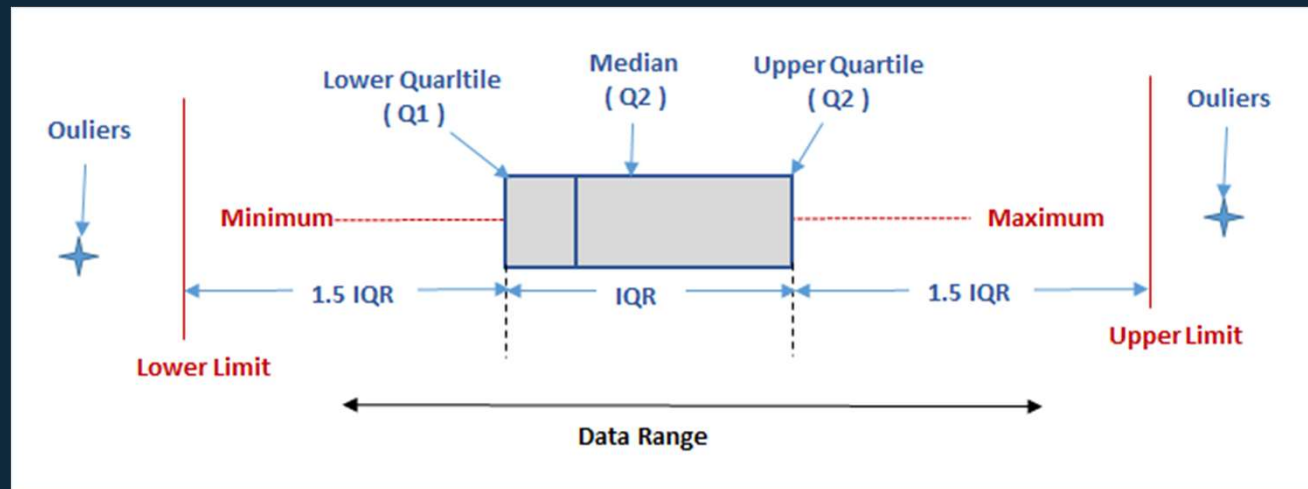


Source: CrowdFlower 2016

Exploração do dado



Box Plot



Exploração de Dados (EDA)

Hands-On

