



Introdução à Ciência de Dados

Luciano Barbosa



UNIVERSIDADE
FEDERAL
DE PERNAMBUCO

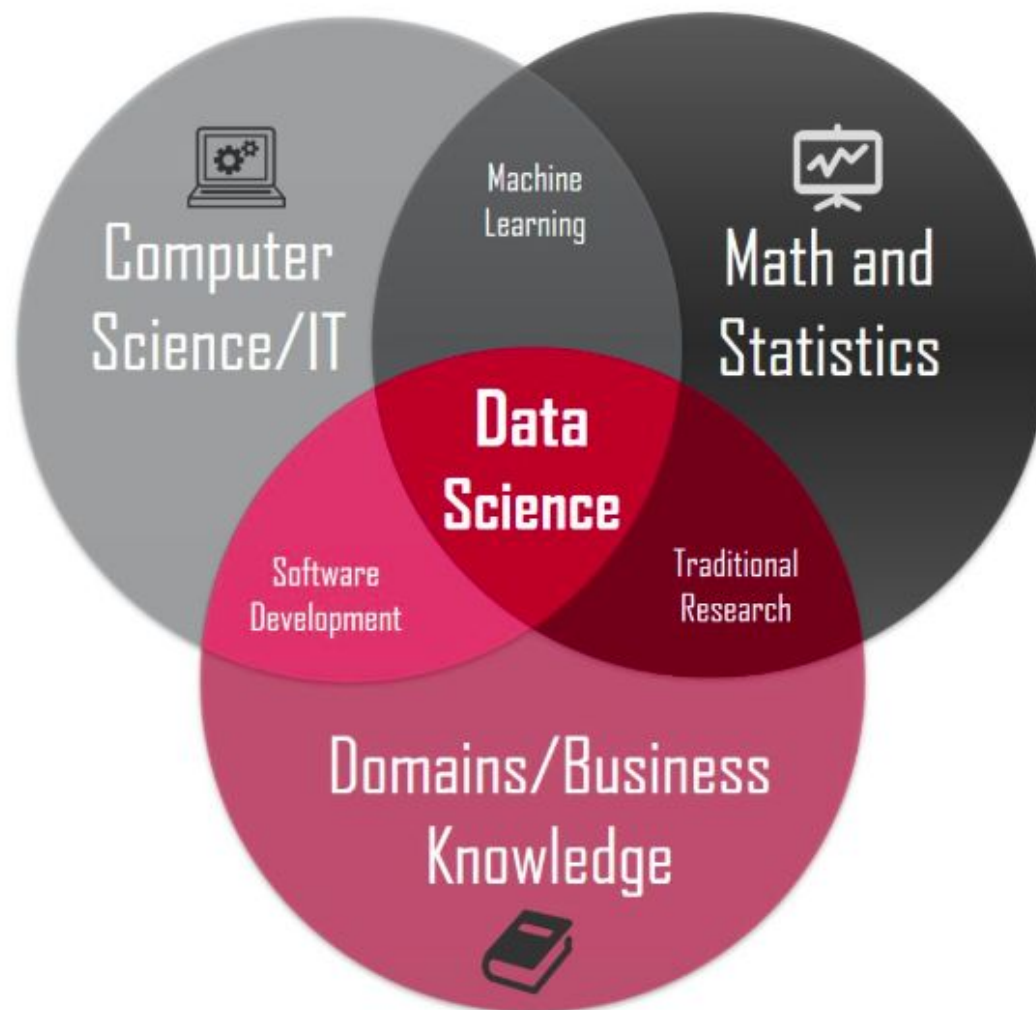


Data Science vs. Computer Science

- CS foca em algoritmos
 - Especifica entrada e saída
 - Algoritmo tem que ser correto e eficiente
 - Dados de entrada podem ser qualquer coisa que vai de acordo com a especificação da entrada
- Data science foca em dados
 - Objetivo de compreender, modelar os dados
 - Algoritmos usados para identificar padrões



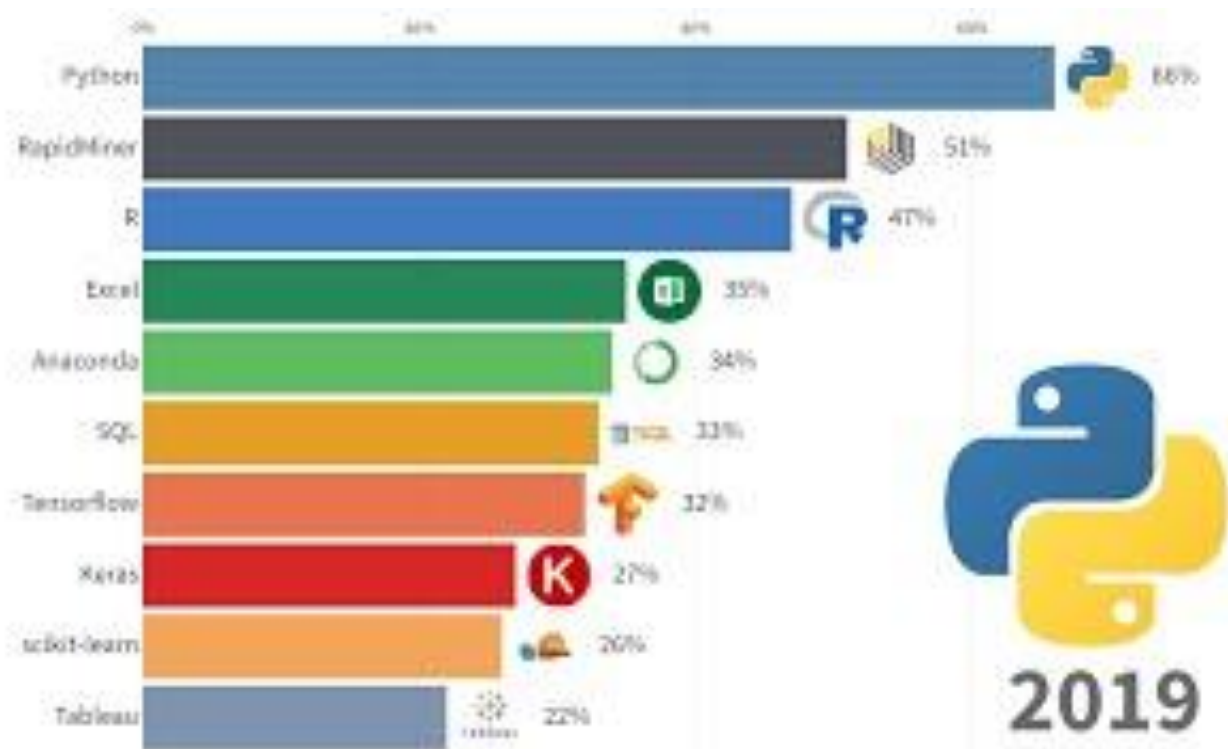
O que é Ciência de Dados





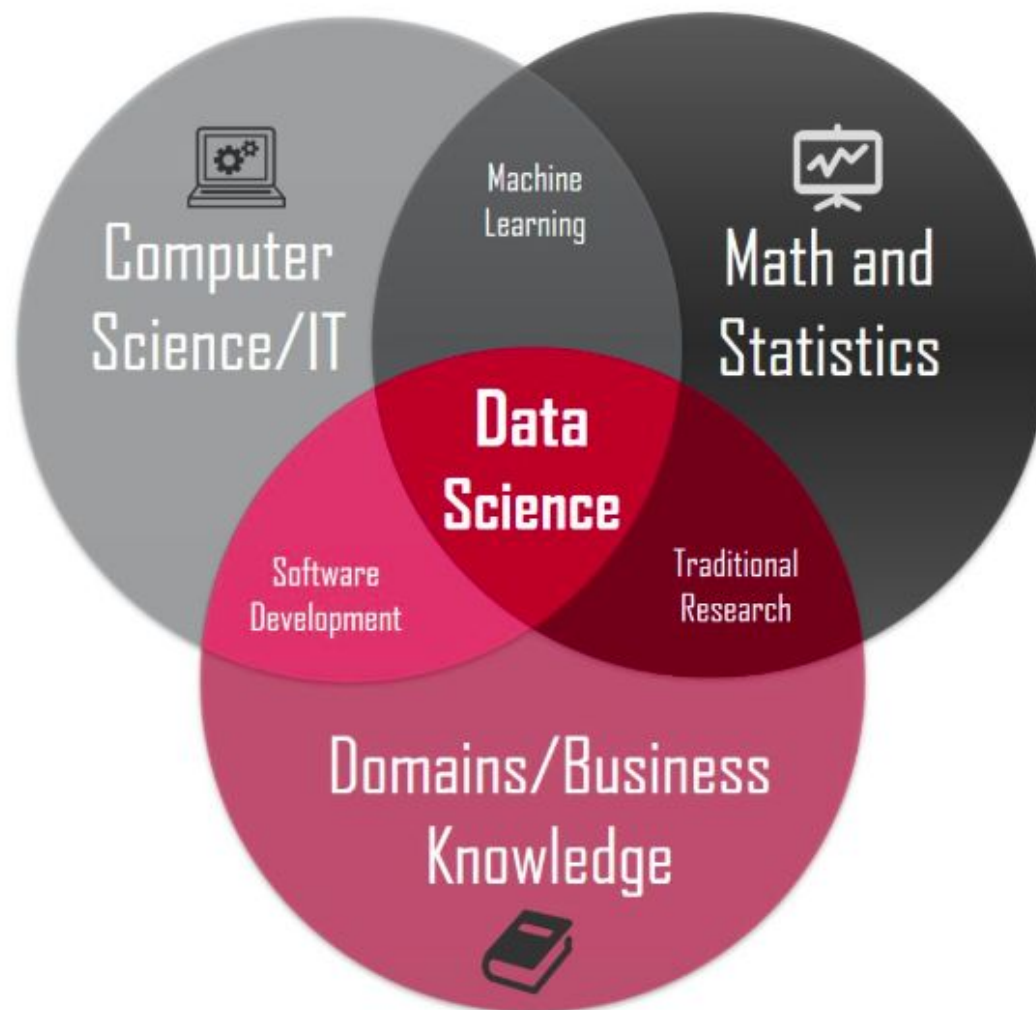
Computação

- Habilidade de construir sistemas
 - Programação: python, R etc
 - Banco de dados: MySQL, MongoDB etc
 - Visualização: D3, Tableau etc
 - Processamento de dados: MapReduce, Spark etc





O que é Ciência de Dados



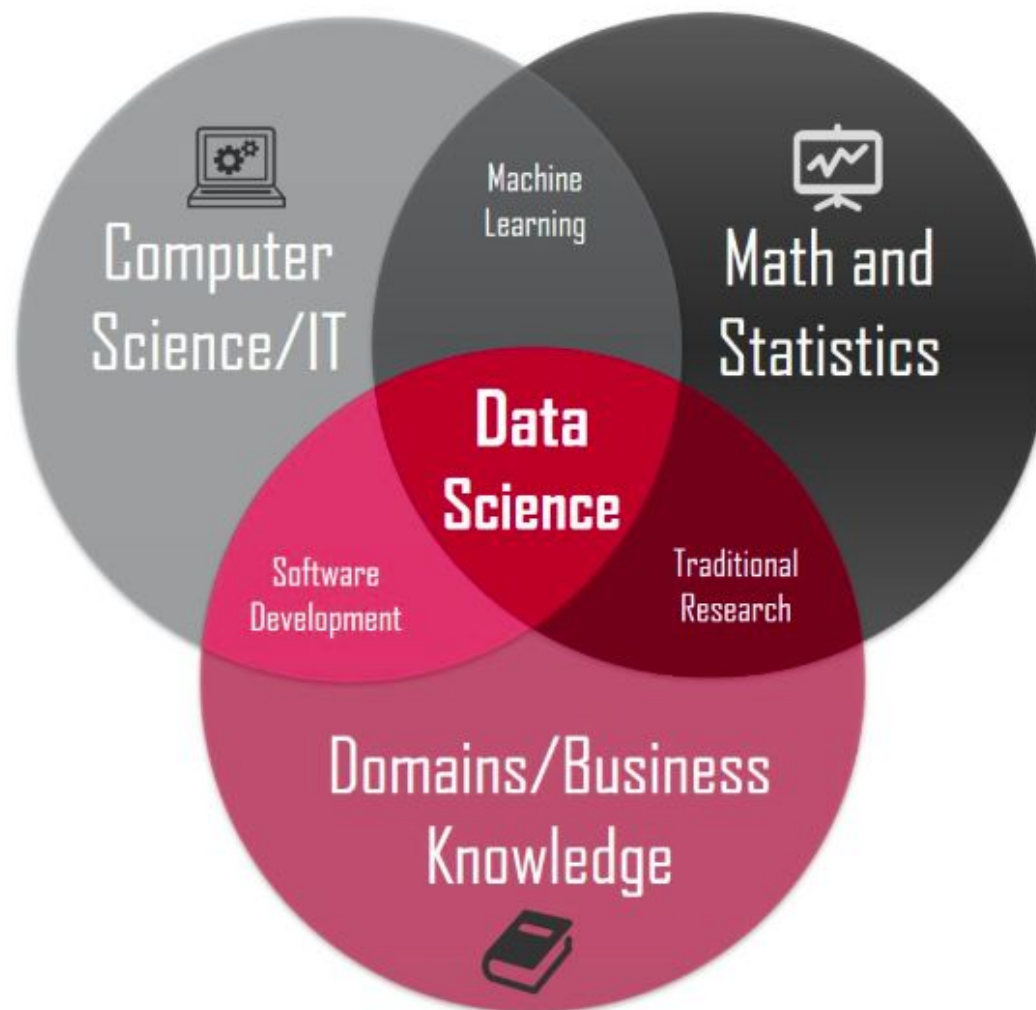


Matemática e Estatística

- Identificar a solução correta para o problema
 - Aprendizado de máquina
 - Estatística descritiva



O que é Ciência de Dados





Conhecimento do Domínio

- Habilidade de fazer perguntas relevantes
- Requer conhecimento do domínio
- Qual o tipo de problema que estamos tentando atacar?



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



Grande Demanda

- Data scientist: um dos top-10 empregos de acordo com a Forbes e Glassdoor
- Inúmeras vagas no Glassdoor e LinkedIn



Bons Salários

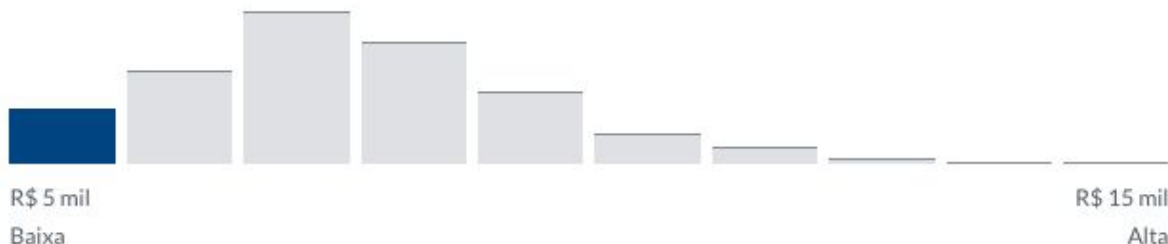


Confiança muito alta

R\$ 9.500 /mês

Média salarial

1.070 salários



Empresa

Faixa salarial base em (BRL)



Itaú Unibanco (Itaú BBA e Rede)

Cientista De Dados: mensal

R\$ 10.453 /mês

4,5 ★

52 salários [Ver 123 salários de todas as localizações](#)



Hospital Israelita Albert Einstein

Cientista De Dados: mensal

R\$ 13.167 /mês

4,5 ★

19 salários [Ver 27 salários de todas as localizações](#)



IBM

Cientista De Dados: mensal

R\$ 10.000 /mês

4,0 ★

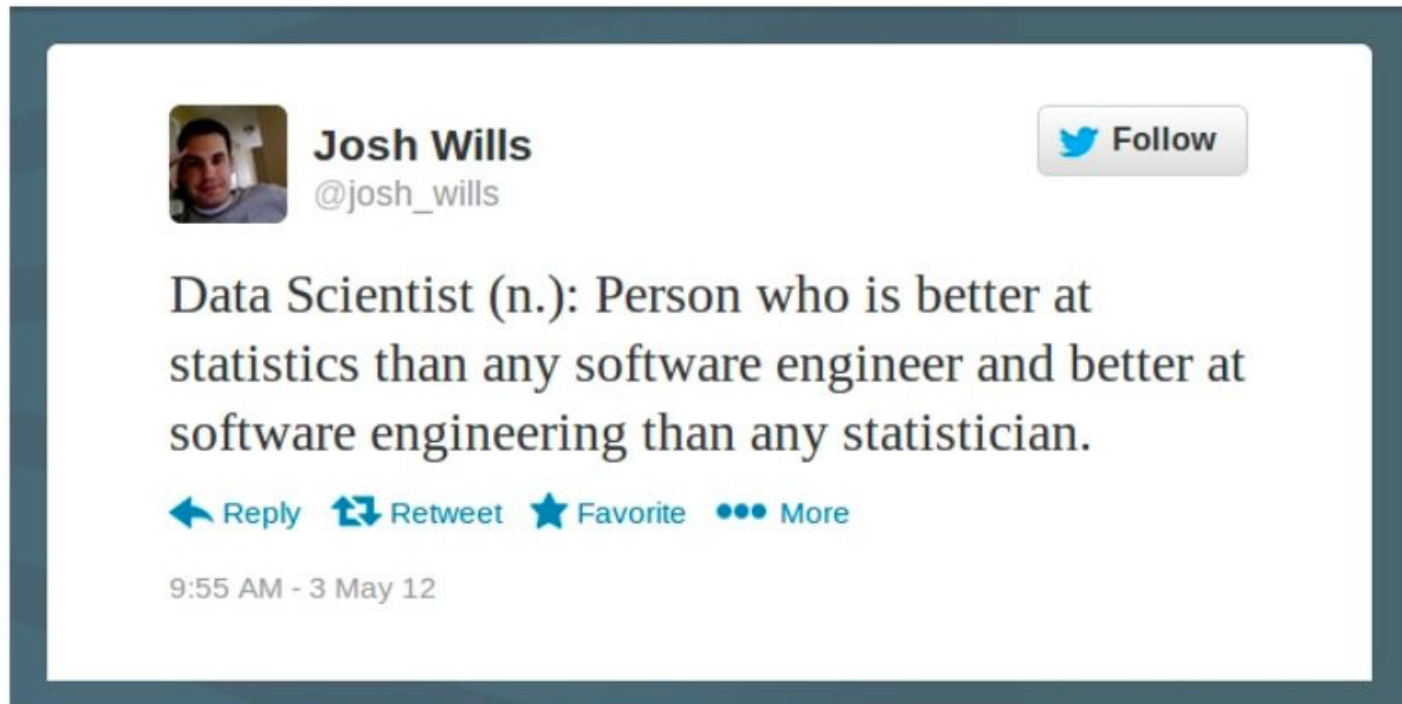
15 salários [Ver 58 salários de todas as localizações](#)

Fonte: Glassdoor 2023

CIn.ufpe.br



O que é um Cientista de Dados?



Fonte: <https://www.slideshare.net/ryanorban/how-to-become-a-data-scientist>



10 Principais Tarefas que Cientistas de Dados Desempenham

- 1 Ask Good Questions. What is What...
...we don't know?
...we'd like to know?
- 2 Define and Test an Hypothesis. Run experiments
- 3 Scoop, Scrap, Sink, & Sample Business Relevant Data
- 4 Munge and Wrestle Data. Tame Data
- 5 Explore Data, Discover Data Playfully. Discover unknowns.
- 6 Model Data. Model Algorithms.
- 7 Understand Data Relationships
- 8 Tell the Machine How to Learn from Data
- 9 Create Data Products that Deliver Actionable Insight
- 10 Tell Relevant Business Stories from Data



Perfil de Cientista de Dados




Cientista de dados

AMcom Sistemas de Informação · Blumenau e Região, Brasil

Posted 2 weeks ago · 717 views

Save

 Easy Apply

Job description

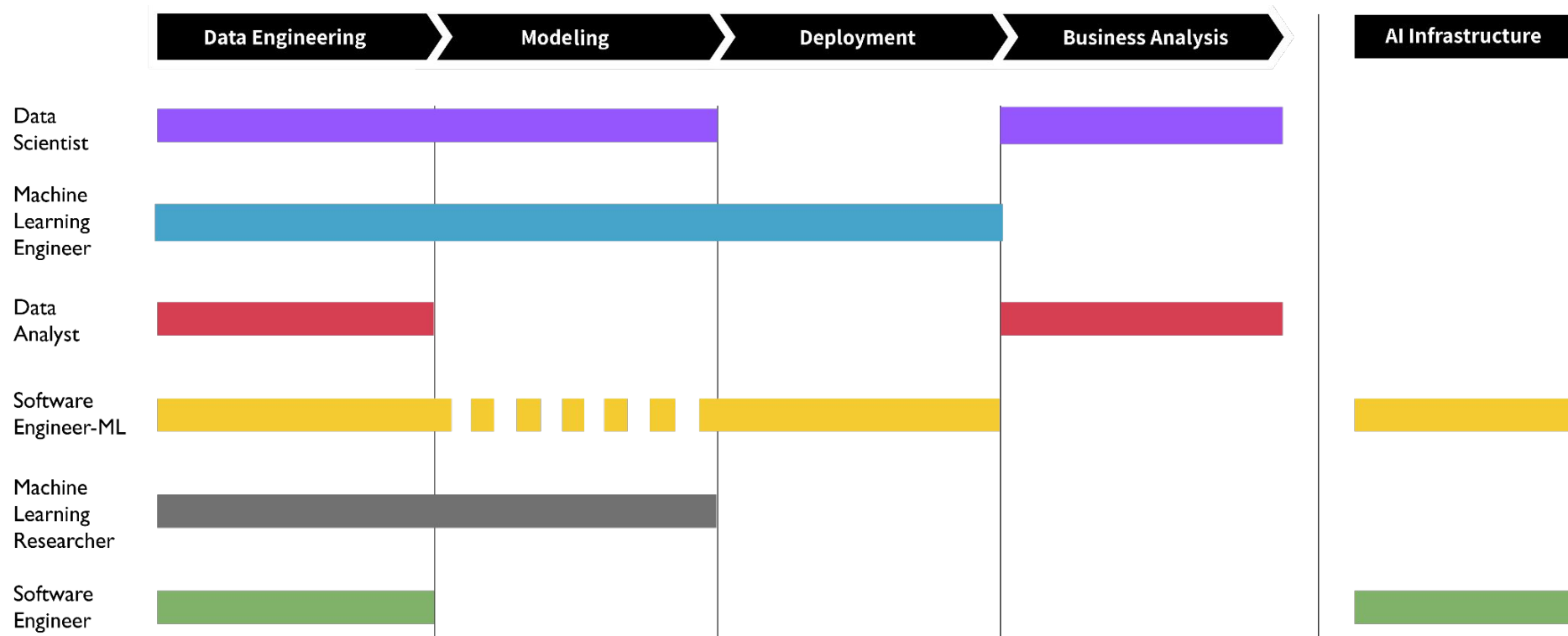
Definir modelos de dados e análise estatística, utilizando inteligência artificial.

Experiência:

- Ensino Superior completo em Sistemas de Informação, Ciência da Computação ou áreas afins.
- Experiência na criação de modelos preditivos.
- Implantação e utilização de ferramentas de aprendizado de máquina.
- Participação em projetos de Machine Learning.
- Deep Learning.
- Data Mining.
- Árvore de decisão/regressão.
- Desejável conhecimento em .NET, Python/R/Ruby.
- Desejável conhecimento em Spark, Hadoop, Azure ML, Kafka, Hive.



Diferentes Perfis





Ferramentas

- Java, R, Python... (bonus: Clojure, Haskell, Scala)
- Hadoop, HDFS & MapReduce... (bonus: Spark, Storm)
- HBase, Pig & Hive... (bonus: Shark, Impala, Cascalog)
- ETL, Webscrapers, Flume, Sqoop... (bonus: Hume)
- SQL, RDBMS, DW, OLAP...
- Knime, Weka, RapidMiner... (bonus: SciPy, NumPy, scikit-learn, pandas)
- D3.js, Gephi, ggplot2, Tableau, Flare, Shiny...
- SPSS, Matlab, SAS... (the enterprise man)
- NoSQL, Mongo DB, Couchbase, Cassandra...
- And Yes! ... MS-Excel: *the most used, most underrated DS tool*

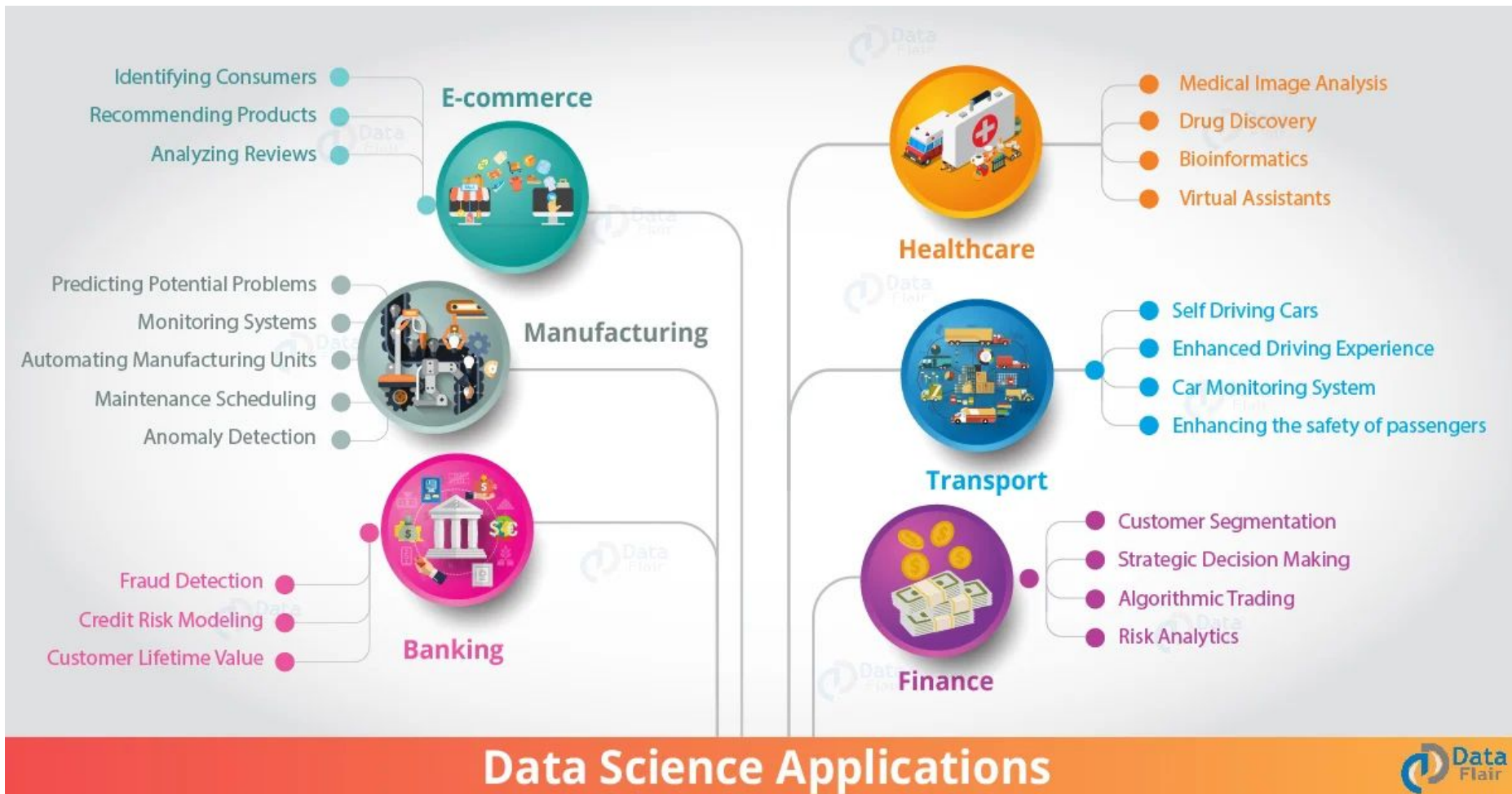


Alguns Princípios

- Dados evoluem
- Dados são sujos
- Preparação e limpeza dos dados tomam muito tempo
- “Torture os dados até que eles falem a verdade”

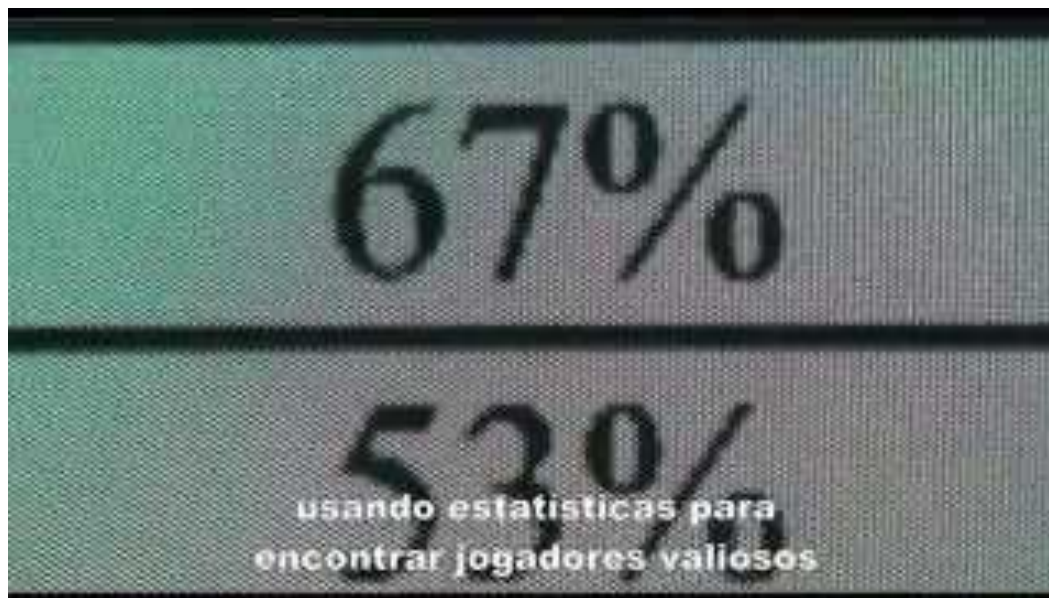


Exemplos de Aplicação de Ciência de Dados



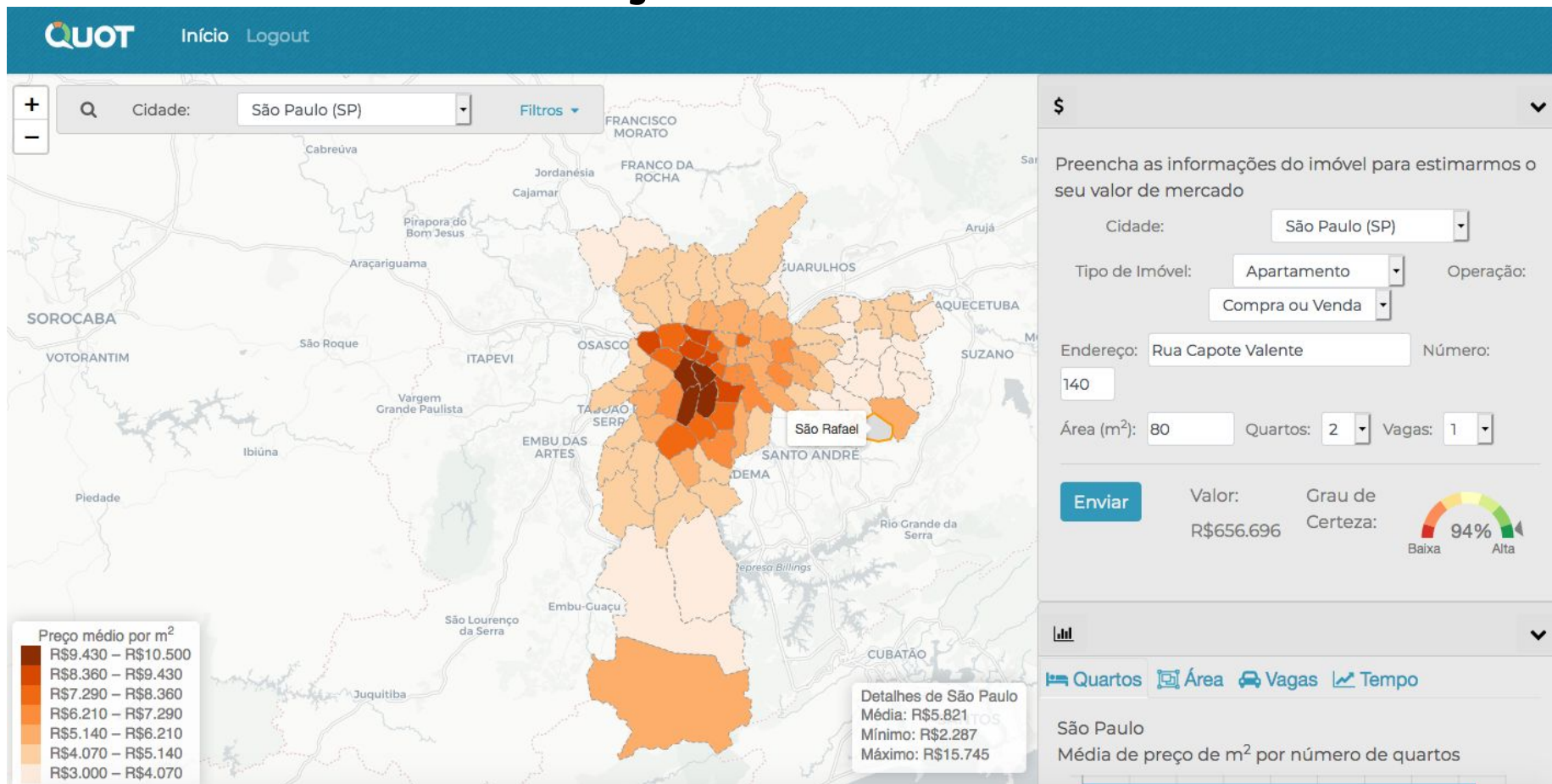


Exemplo de Aplicação: Análise de Jogadores de Beisebol





Exemplo de Aplicação: Predição de Preço de Imóvel





Exemplo de Aplicação: Análise de Dados de Passagens Aéreas

Find the Fair Price and When to Buy Airline Tickets Based on Airfare Sales Data

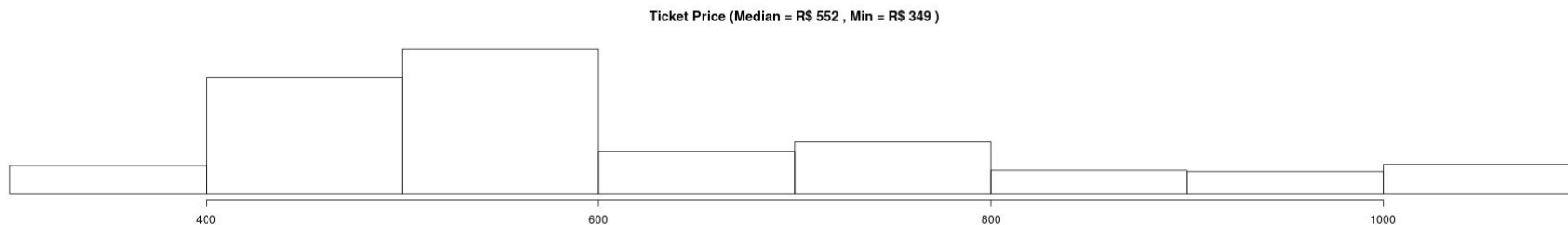
Origin

Recife

Destination

Aracaju

Search

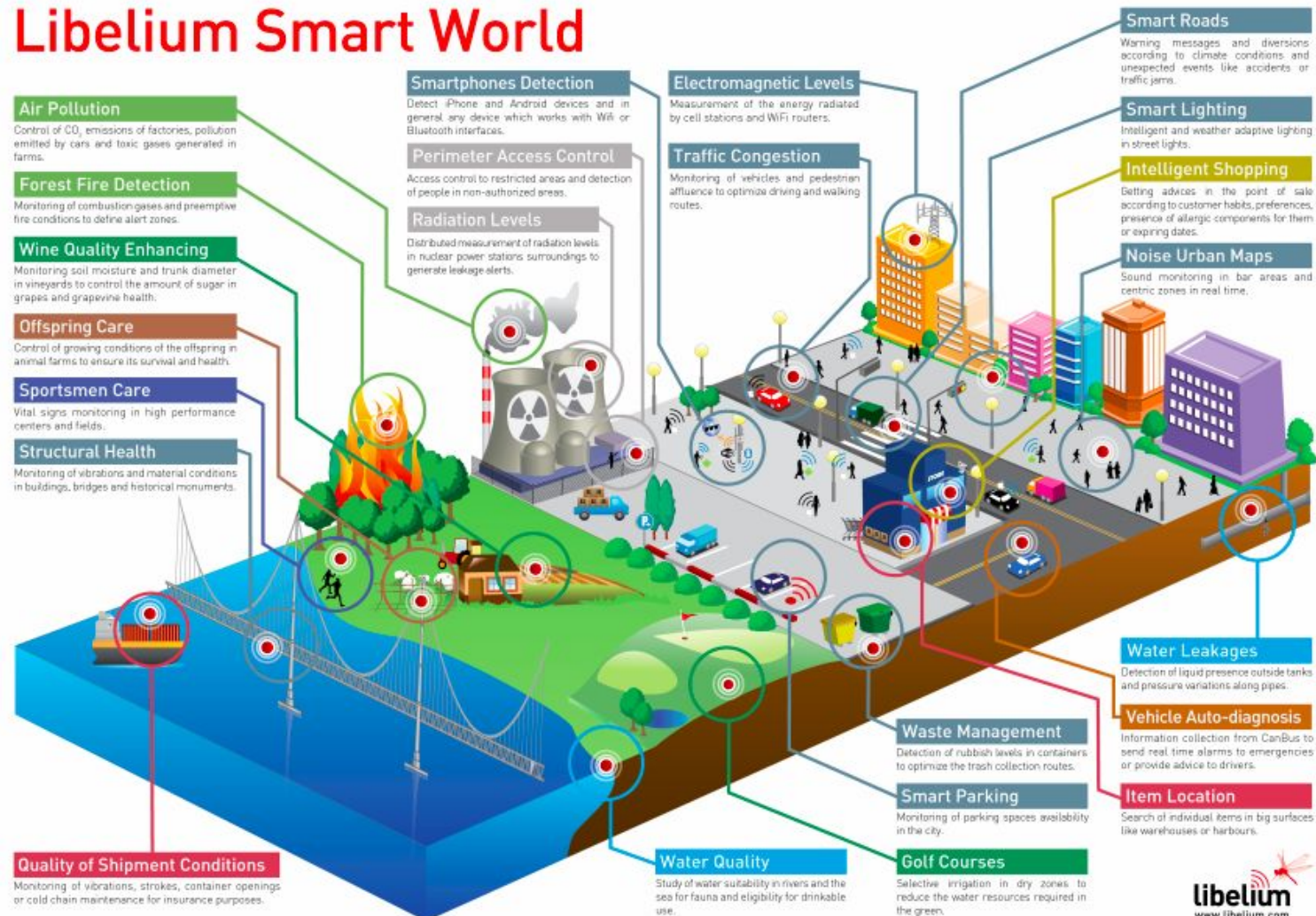


Days prior to the trip (Median = 41 days)



Data Science & IOT

Libelium Smart World



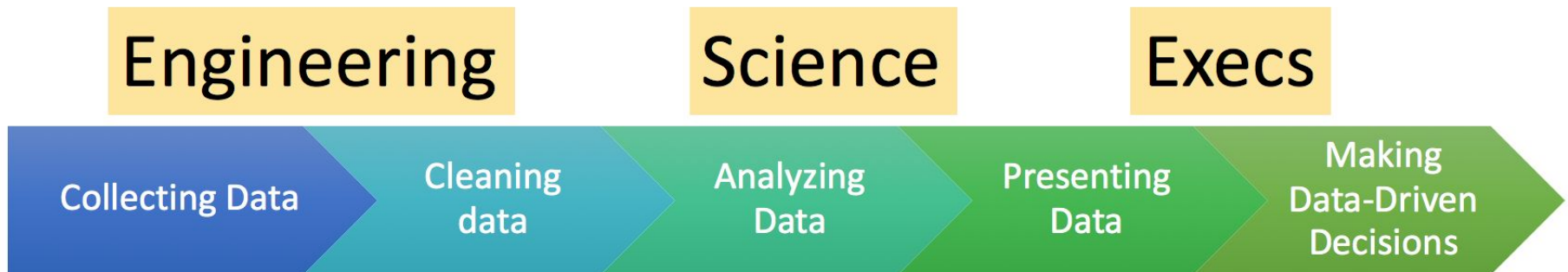


A melhor forma de aprender
ciência de dados é praticar!

<https://www.kaggle.com/competitions>

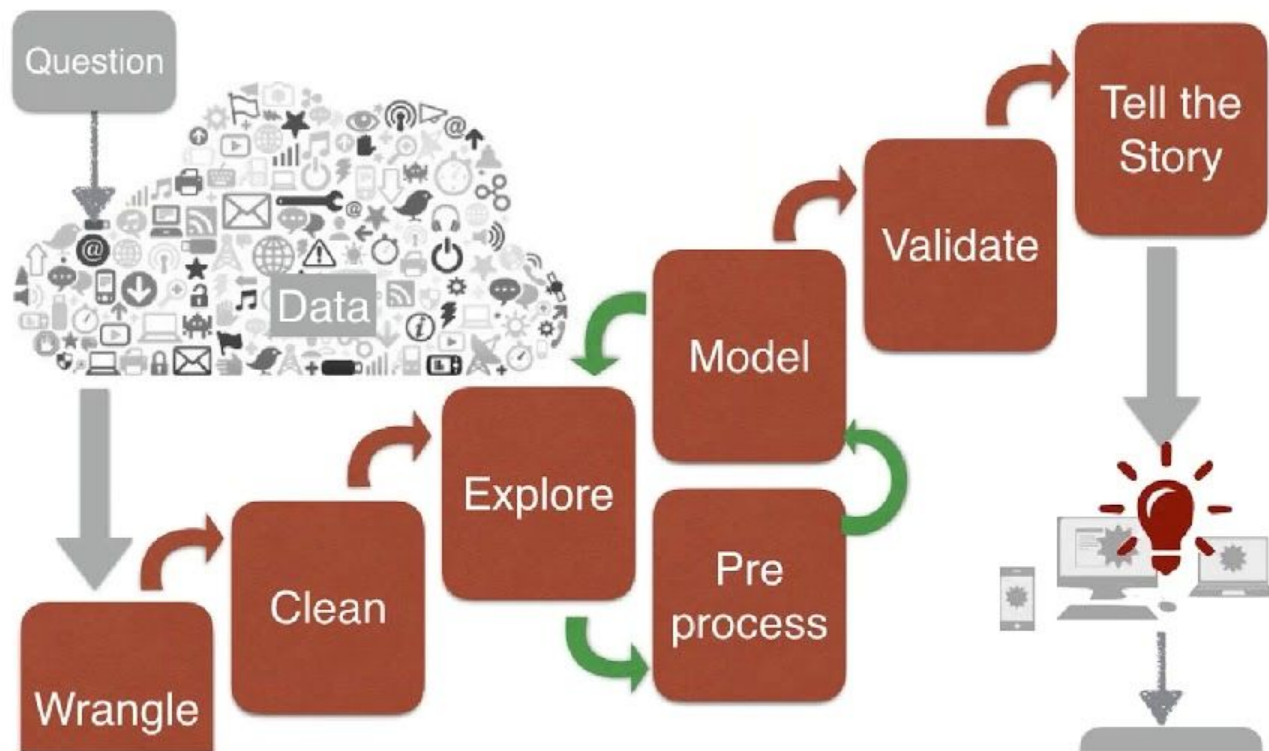


Exemplos de Pipelines





The Data Science pipeline





Pipeline

- Definir o problema
 - Ex.: Posso prever infecção antes que ela ocorra?
- Identificar e coletar dados
 - Ex.: Dados de batimento cardíaco, pressão sanguínea etc
- Entender e preparar os dados
 - Ex.: Limpar e agregar os dados
- Construir e avaliar modelos
 - Ex.: Comparar diferentes modelos de predição
- Apresentar resultados
 - Ex.: Criar dashboard para profissionais de saúde
- Colocar modelo em produção
 - Ex.: Verificar escalabilidade



**STATISTICIANS, LIKE
ARTISTS, HAVE THE BAD
HABIT OF FALLING IN LOVE
WITH THEIR MODELS
– GEORGE BOX**



Visão Geral do Curso

- Foco em aplicação
- Apresentação de conceitos-chave
- Importante: conhecimento básico de programação
- Aulas focadas em exemplos
- Avaliação: projetos



Conteúdo

- Introdução à Ciência de Dados
- Processamento de dados colunares (Pandas)
- Estatísticas descritivas
- Visualização de dados
- Testes de hipótese
- Pré-processamento:
 - Limpeza (detecção de outliers)
 - Normalização e imputação etc



Conteúdo

- Processamento de séries temporais
- Processamento em larga escala
- Modelos preditivos:
 - Classificação
 - Regressão
 - Avaliação e diagnóstico de modelos
 - Interpretabilidade da predição
- Agrupamento
- Rastreamento e reproducibilidade