Ciência de Dados

Aula 1- Apresentação da Disciplina

Prof. Wellington Franco

UNIVERSIDADE FEDERAL DO CEARÁ
CAMPUS DE CRATEÚS

enginelab
Laboratório de Engenharia de Software e Sistemas

Apresentação

- Wellington Franco
- Graduação
  - Graduado em Ciência da Computação pela UECE
- Mestrado
  - Mestrado em Lógica e Inteligência Artificial pela UFC
- Doutorado
  - Doutor em Banco de Dados e Inteligência Artificial pela UFC

Apresentação

- Experiência
  - Líder técnico de projetos de P&D
  - Mais de 10 anos de experiência em pesquisa
  - Professor da Universidade Federal do Ceará - Campus Crateús
- Contatos
  - Linkedin: https://www.linkedin.com/in/wellington-franco-565a60a7/
  - E-mail: wellington@crateus.ufc.br

Agenda

1. Introdução
2. Apresentação geral sobre o curso
    - Análise Estatística de Dados
    - Processamento de Linguagem Natural
    - Machine Learning
    - Infraestrutura de Big Data
3. Avaliação e próximos passos
4. Introdução a Big Data

Big Data

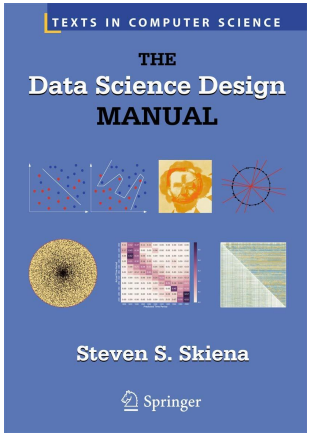
- Estrutura do Curso:
  - Módulo 1: Análise Estatística de Dados
  - Introdução à Data Science
  - Engenharia de Dados
  - Distribuição dos Dados
  - Conceitos Básicos de Visualização de Dados

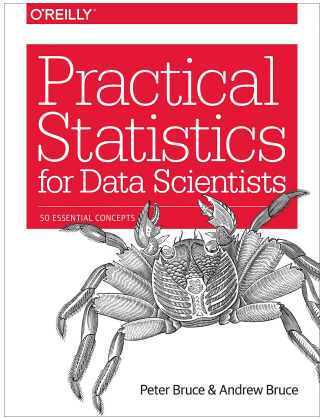
Big Data

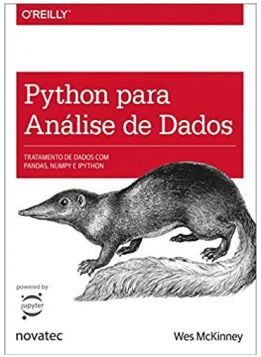
- Estrutura do Curso:
- Módulo 1: Análise Estatística de Dados
    - Bibliografia:
        - The Data Science Design Manual, 1st ed. 2017. Steve Skiena.
        - Python para Análise de Dados: Tratamento de dados com Pandas, NumPy e IPython, Novatec Editora; 1ª Edição, 2019. Wes McKinney.
        - Practical Statistics for Data Scientists, O'REILLY, 2017. Peter Bruce and Andrew Bruce.
        - An Introduction to Data Science, 2017. Jeffrey Saltz and Jeffrey Stanton.
        - Data Science do Zero: Primeiras Regras com o Python, Alta Books, 2016. Joel Grus.

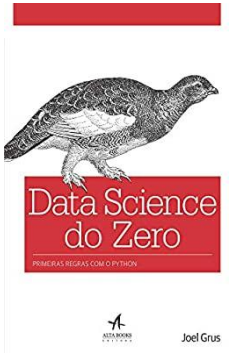
Big Data

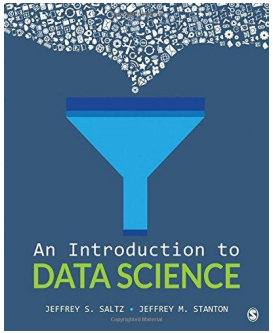
1. Estrutura do Curso:
    - Módulo 1: Análise Estatística de Dados
        - Bibliografia:











Big Data

- Estrutura do Curso:
  - Módulo 2: Processamento de Linguagem Natural
    - Objetivos:
      - Aplicar técnicas para extração de conhecimento sobre dados não estruturados.
      - Identificar as abordagens mais apropriadas para uso no contexto textual.

Big Data

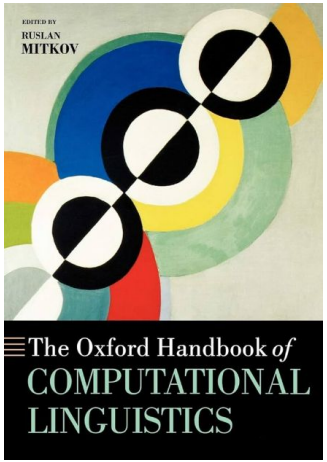
- Estrutura do Curso:
  - Módulo 2: Processamento de Linguagem Natural
    - Fundamentos de Processamento de Linguagem Natural
    - Correlação de termos; Concorrência de termos;
    - Ferramentas para processamento de texto;
    - Problemas da análise sintática e semântica de linguagem natural;
    - Extração de Informação em Textos;
    - Reconhecimento de Entidades Nomeadas;

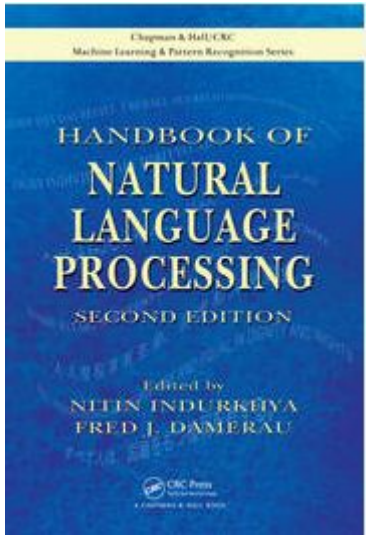
Big Data

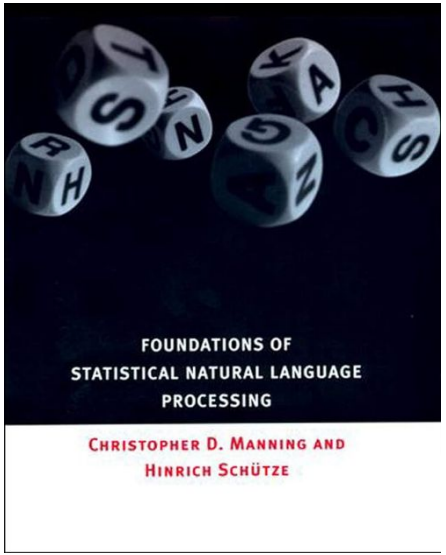
- Estrutura do Curso:
- Módulo 2: Processamento de Linguagem Natural
    - Bibliografia:
        - Ruslan Mitkov, The Oxford Handbook Of Computational Linguistics, Oxford Universitty Press, 2003.
        - Robert Dale, Hermani Moisi, Harold Somers, Handbook Of Natural Language Processing, Markcel Dekker Inc.
        - James Allen, Natural Language Processing, Pearson Education, 2003.
        - Christopher D.Manning & Henrich Schutze, Foundations Of Statistical Natural Language Processing, The MIT Press, 2001
        - Douglas Biber, Susan Conrad, Randi Reppen, Corpus Linguistics – Investigating Language Structure And Use, Cambridge University Press, 2000.

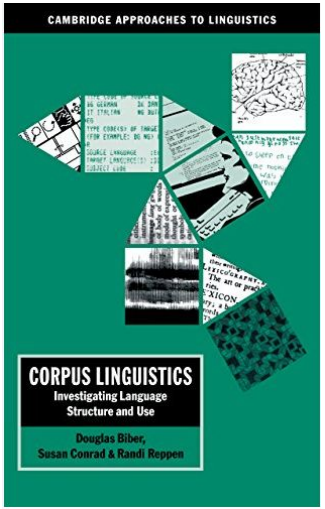
Big Data

1. Estrutura do Curso:
    - Módulo 2: Processamento de Linguagem Natural
        - Bibliografia:









Big Data

- Estrutura do Curso:
  - Módulo 3: Machine Learning
    - Objetivos:
      - Identificar problemas de regressão, classificação, clusterização, detecção de outliers e reamostragem de dados.
      - Identificar as abordagens mais apropriadas para gerar modelos de aprendizagem automática em cada um desses cenários.

Big Data

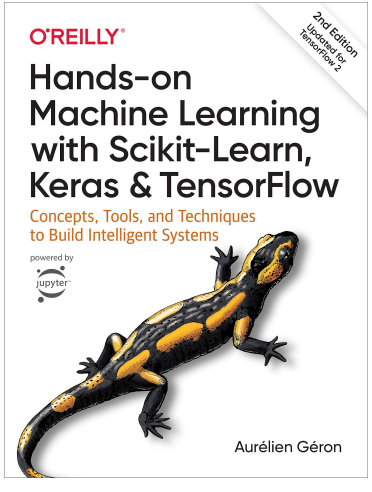
- Estrutura do Curso:
  - Módulo 3: Machine Learning
    - Fundamentos de Machine Learning
    - Problemas de Regressão
    - Problemas de Classificação
    - Problemas de Clustering
    - Redes Neurais Artificiais

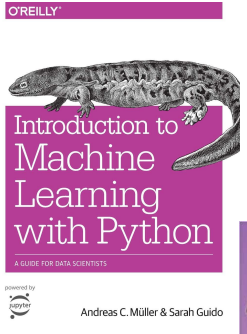
Big Data

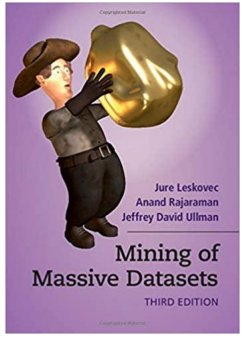
- Estrutura do Curso:
  - Módulo 2: Machine Learning
    - Bibliografia:
      - Hands–On Machine Learning with Scikit–Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, 2o. Edition, 2019. Aurelien Geron.
      - Introduction to Machine Learning with Python: A Guide for Data Scientists, O'REILLY, 2016. Andreas C. Mueller and Sarah Guido.
      - Pattern Recognition and Machine Learning, 2011. Christopher M. Bishop.
      - Data Mining: Concepts and Techniques, 2011. Jiawei Han, Jian Pei and Micheline Kamber.
      - Data Mining and Analysis: Fundamental Concepts and Algorithms, First Edition, 2020. Mohammed J. Zaki e Wagner Meira Jr.
      - Mining of Massive Datasets, 3rd Edition, 2020. Jure Leskovec, Anand Rajaraman and Jeff Ullman.
      - The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, 2009. Trevor Hastie, Robert Tibshirani and Jerome Friedman.
      - Data Mining: Practical Machine Learning Tools and Techniques (The Morgan Kaufmann Series in Data Management Systems), 3rd Edition, 2011. Ian H. Witten, Eibe Frank and Mark A. Hall.

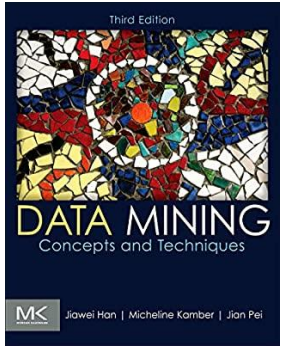
Big Data

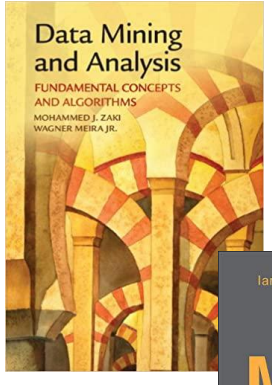
- Estrutura do Curso:
  - Módulo 3: Machine Learning
    - Bibliografia:

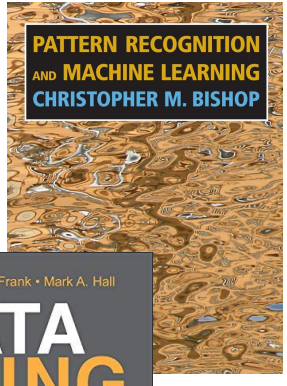


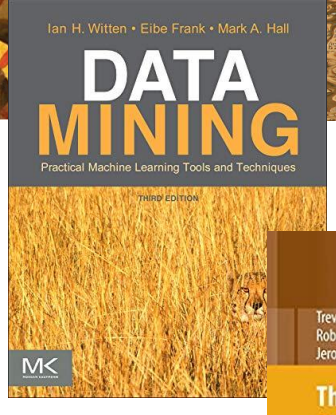


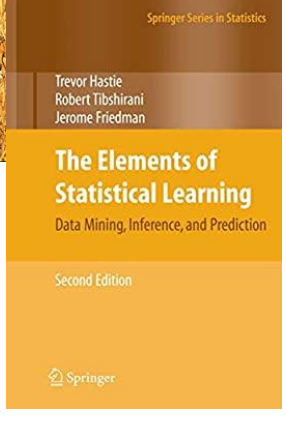












Big Data

- Estrutura do Curso:
  - Módulo 4: Infraestrutura de Big Data
    - Objetivos:
      - Comparar as arquiteturas de armazenamento e processamento de dados tradicionais com as novas arquiteturas para big data.
      - Apresentar os principais componentes e conceitos do ecossistema Hadoop (HDFS, Sqoop, Hive, Arquitetura Lambda).

Big Data

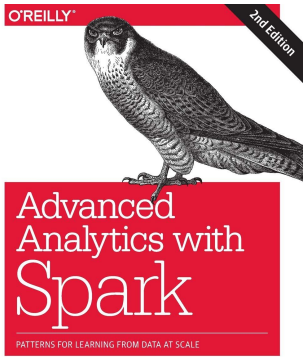
- Estrutura do Curso:
  - Módulo 4: Infraestrutura de Big Data
    - Introdução a Big Data
    - Montando um Ambiente de Big Data
    - Desenvolvendo com o Spark

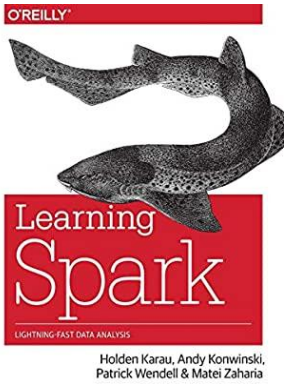
Big Data

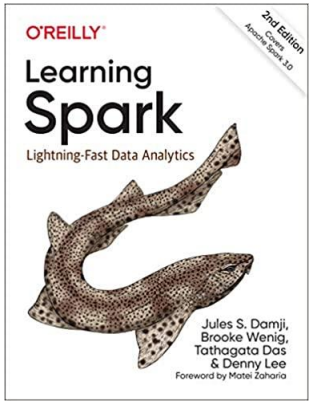
- Estrutura do Curso:
  - Módulo 4: Infraestrutura de Big Data
    - Bibliografia:
      - Learning Scala Programming: Object-oriented Programming Meets Functional Reactive to Create Scalable and Concurrent Programs, First Edition, 2018. Vikas Sharma.
      - Advanced Analytics with Spark: Patterns for Learning from Data at Scale, O'Reilly; 2nd Edition, 2017. Uri Laserson, Sean Owens, Sandy Ryza and Josh Wills.
      - Learning Spark: Lightning-Fast Big Data Analysis, O'Reilly Media; 1ª Edição, 2015. Mark Hamstra, Matei Zaharia and Holden Karau.
      - Learning Spark: Lightning-Fast Data Analytics, O'Reilly Media, 2020. Jules S. Damji, Brooke Wenig, Tathagata Das and Denny Lee.
      - Seven Databases in Seven Weeks: A Guide to Modern Databases and the Nosql Movement, O'Reilly, 2nd Edition, 2018. Luc Perkins, Eric Redmond and Jim Wilson.

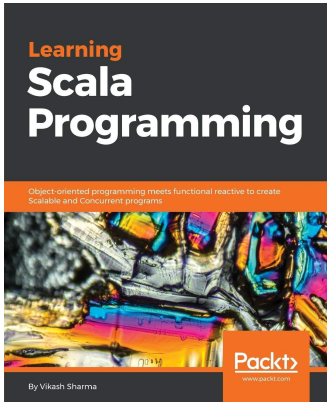
Big Data

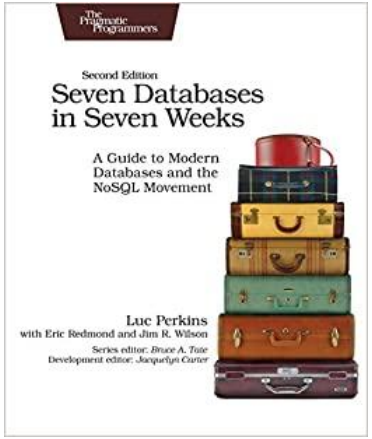
- Estrutura do Curso:
  - Módulo 3: Infraestrutura de Big Data
    - Bibliografia:

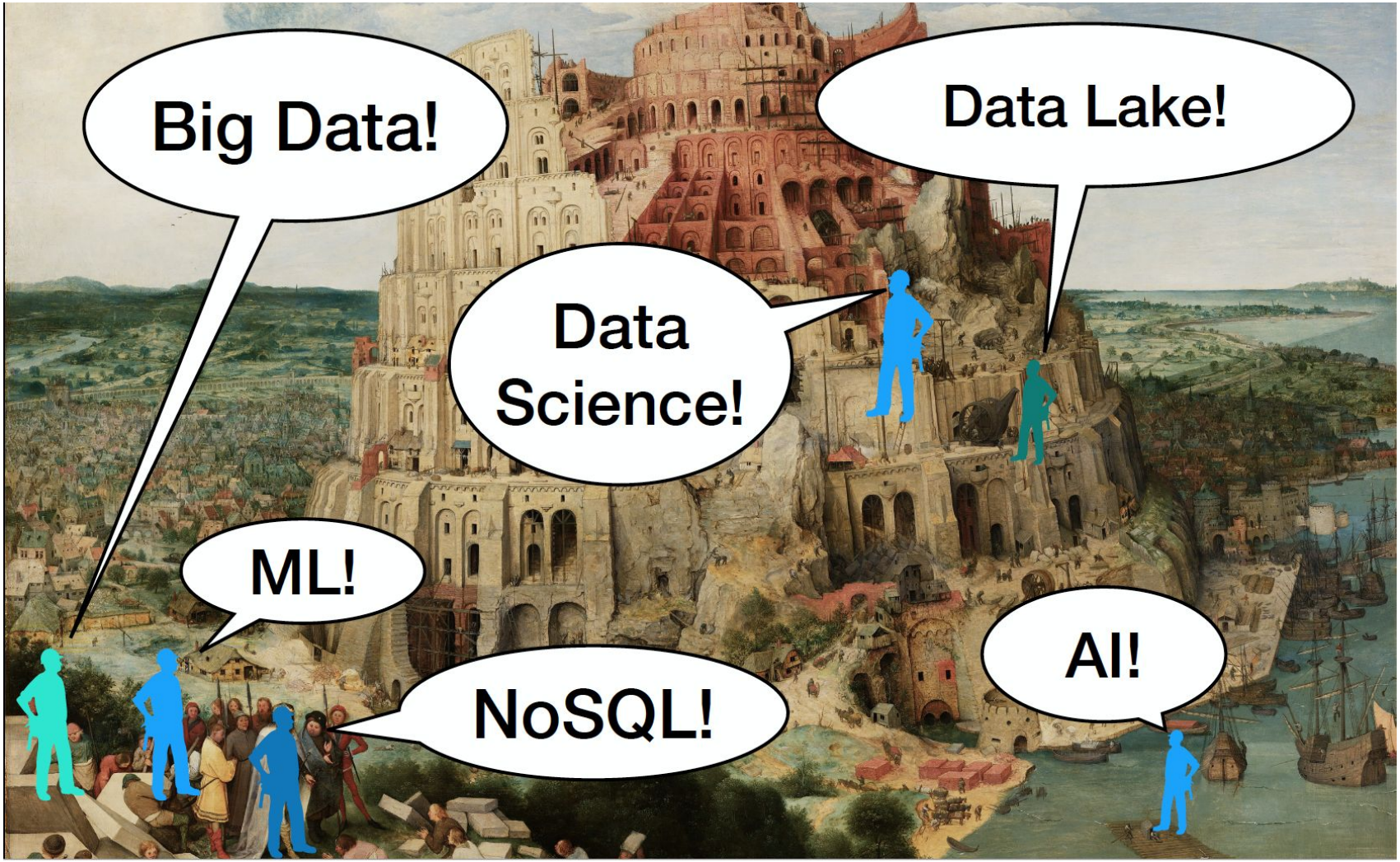






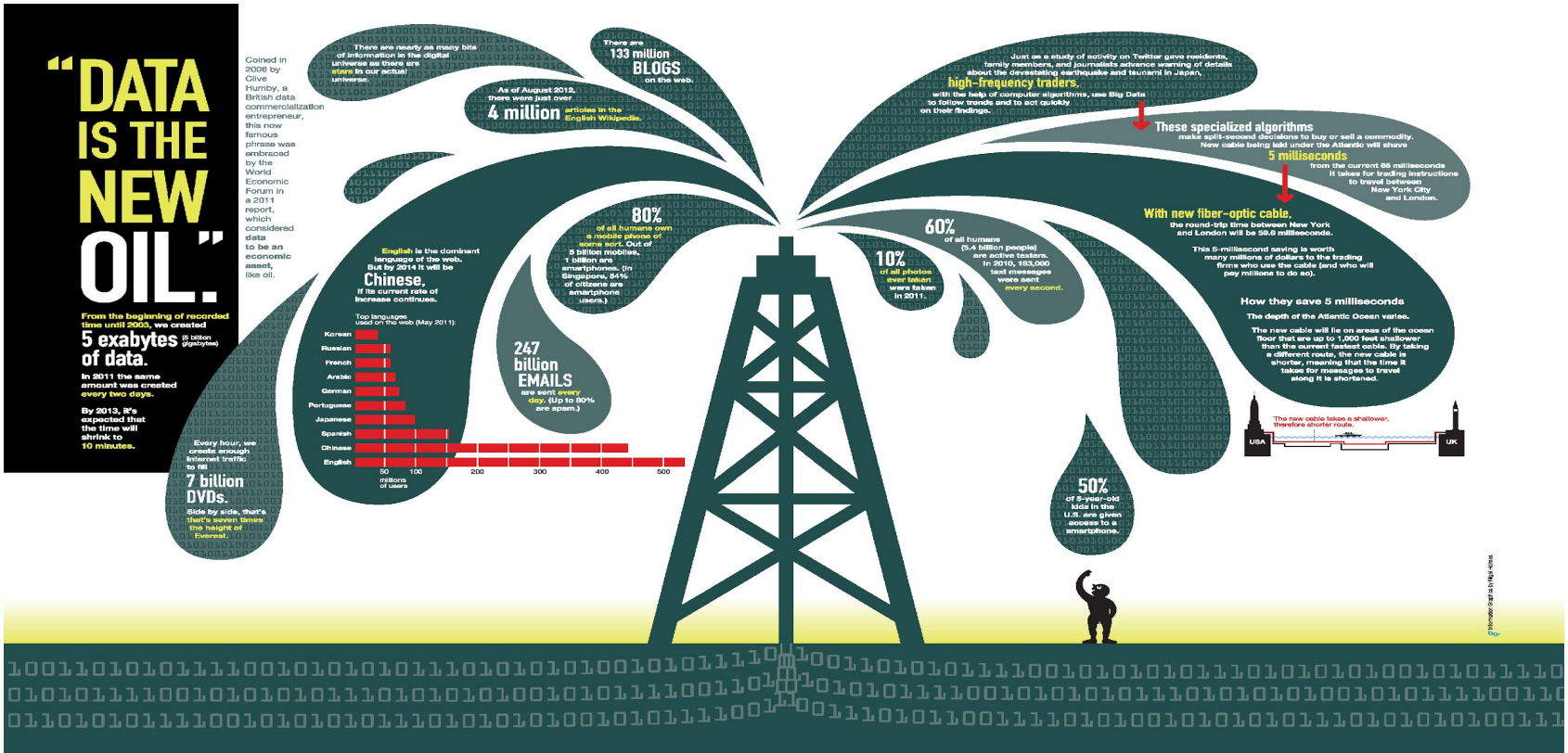





Big Data!
Data Lake!
Data Science!
ML!
AI!
NoSQL!

O Importante é que...

“Data is the New Oil” – World Economic Forum 2011



Data contains value and knowledge

Mas,...

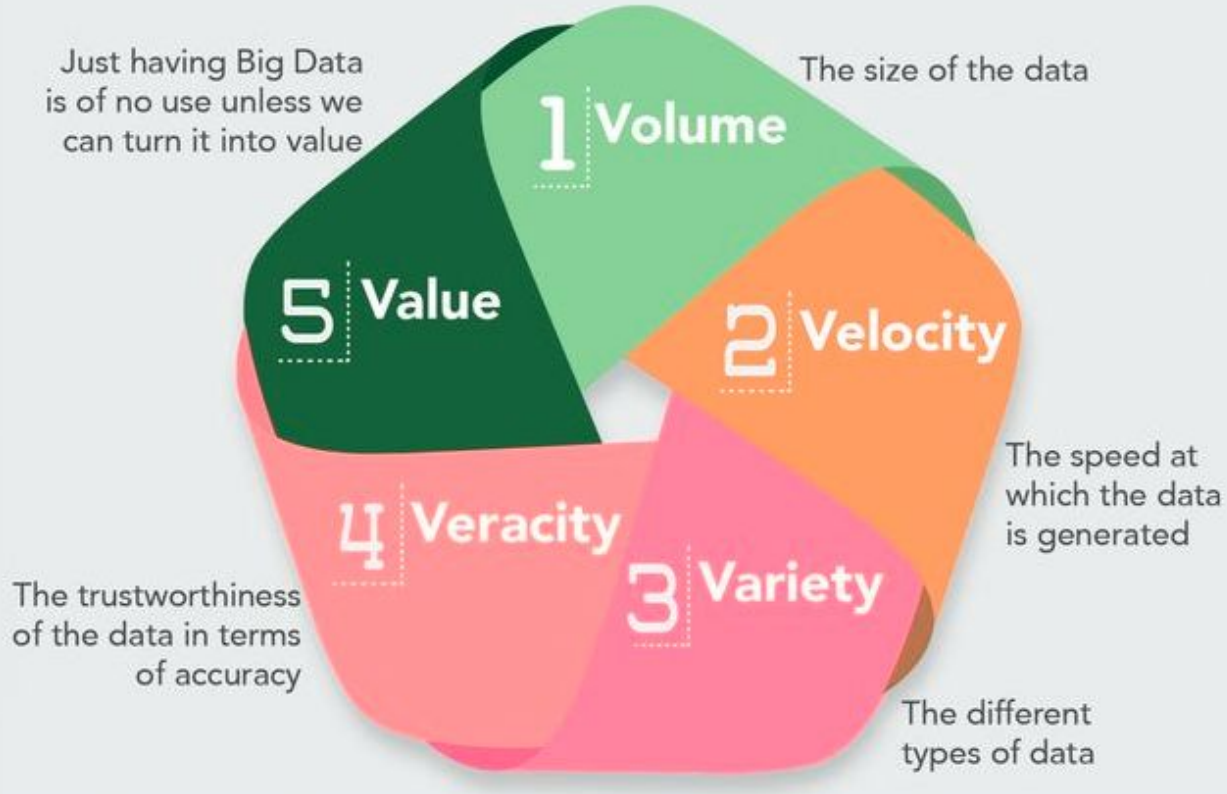
Vamos tentar organizar essa bagunça...


BIG
DATA
ALSO
TOOLS
DISK
TARGET
APPLIED
SENSOR
MAY
EXAMPLES
SHARED
DEFINITION
CURRENT
MOVING
EVERY
RECONSIDER
WITHIN
ZETTABYTES
PRACTITIONERS
CAPTURE
THOUGHT
INTERNET
BUSINESS
SETS
GENOMICS
DISTRIBUTED
DESCRIBING
RADIO-FREQUENCY
COMPLEXITY
MANAGEMENT
TERABYTES
CASE
ABILITY
PETABYTES
INCLUDE
ELAPSED
SYSTEMS
TOLERABLE
STORAGE
SIZE
MPP
PARALLEL
SAN
QUALITIES
SINCE
MASSIVELY
GROW
DIFFICULTY
SEARCH
RECORDS
DATABASES
NETWORKS
HUNDREDS
DESKTOP
CAPACITY
BURIED
WORLD'S
NOW
BIOLOGICAL
PROCESSING
UBIQUITOUS
TYPES
SOLID
CURRENTLY
GARTNER
WORKING
AMOUNT
OPPORTUNITIES
ANALYTICS
TENS
COMPLEX
CITATION
INDEXING
LARGER
CONTINUES
SOCIAL
USE
SET
CONNECTOMICS
ORGANIZATIONS
RELATIONAL

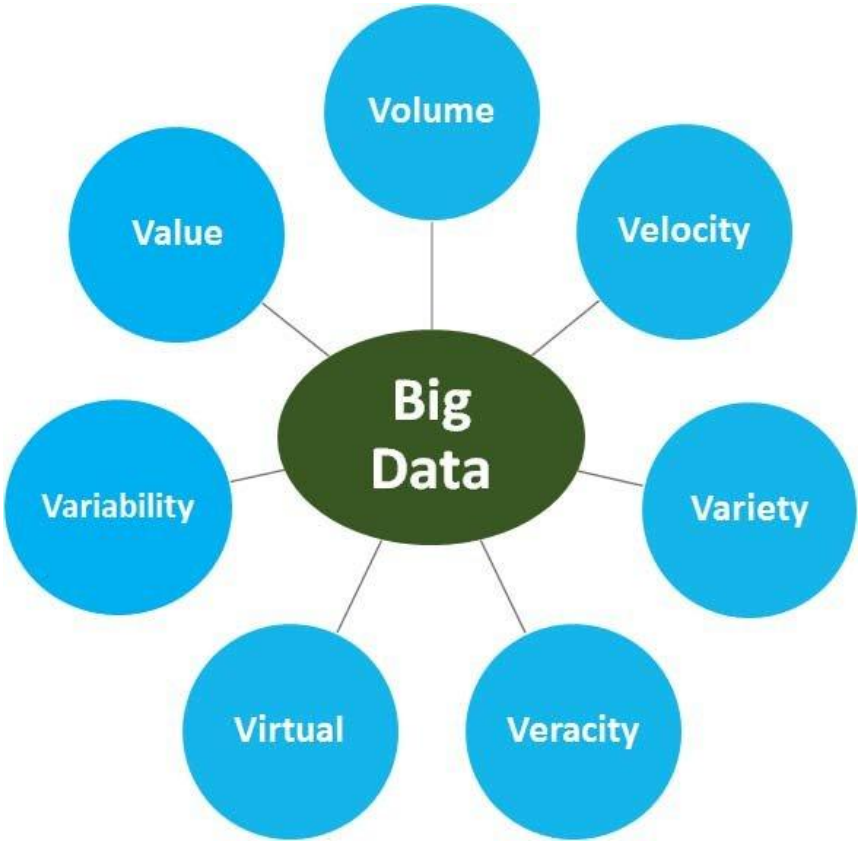
Big Data

- Uma Tentativa de Definição:
  - Área do conhecimento que estuda como tratar, analisar e obter informações a partir de conjuntos de dados “grandes” demais para serem analisados por sistemas tradicionais;
    - Onde o termo “grande” tem diversos sentidos:
      - 3Vs, 5Vs, 7Vs
    - Diferentes técnicas/métodos podem ser utilizados para tratar, analisar e obter informações:
      - Ex: Contar quantas vezes cada palavra (item/produto) aparece em um conjunto de documentos gerados continuamente (NFEs);

THE 5 Vs OF BIG DATA

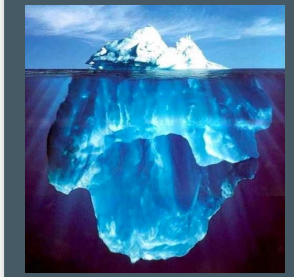


THE 7 Vs OF BIG DATA



“Big Data”: De onde vêm os dados

Tudo que acontece On-line



Every:
Click
Ad impression
Billing event
Fast Forward, pause
Network message
Fault
…

Gerados pelos Usuários (Web & Mobile)

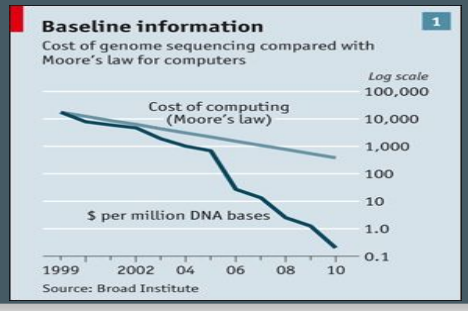





…
..

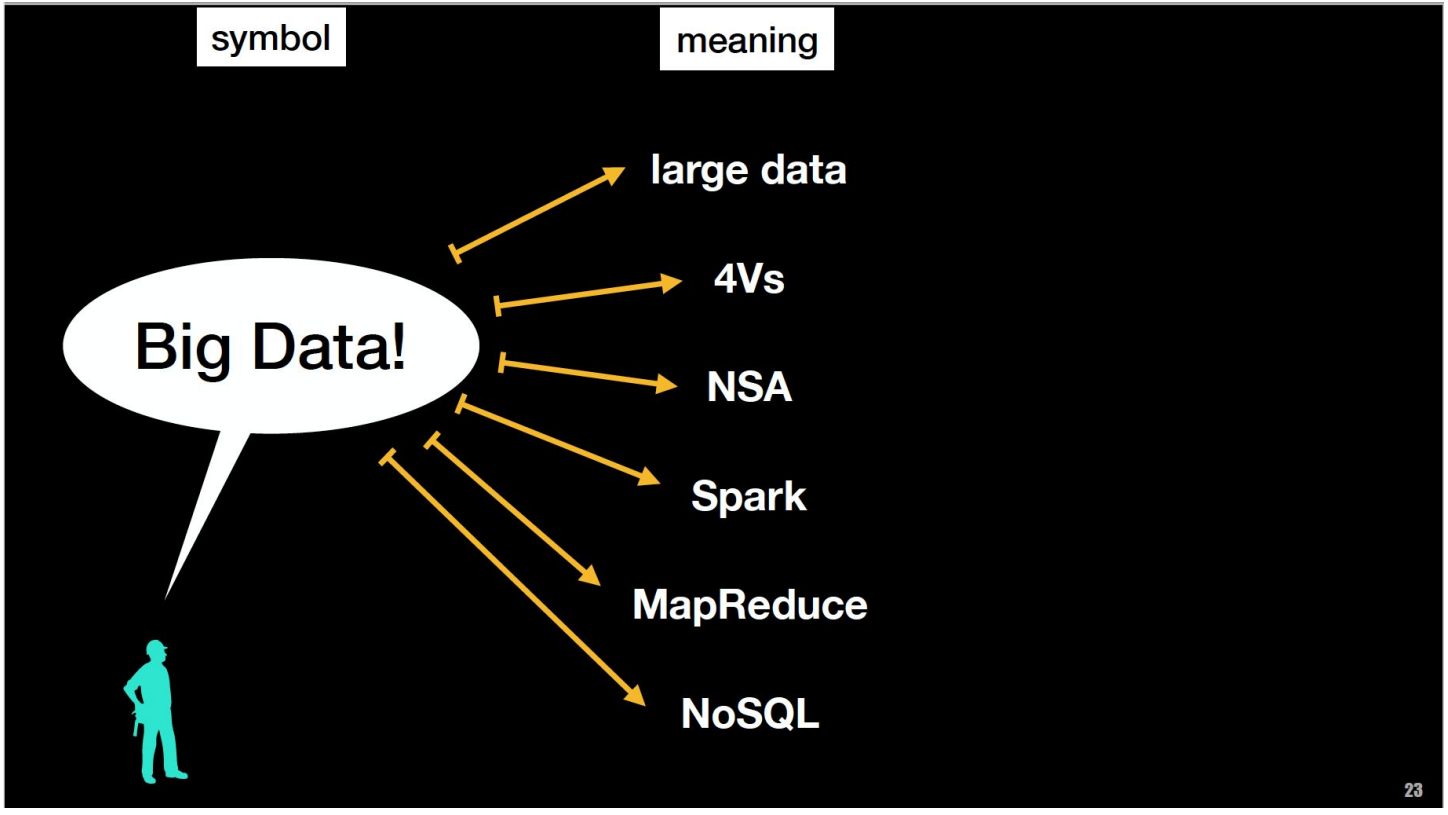
Internet das Coisas (IoT)

Computação Científica/Saúde

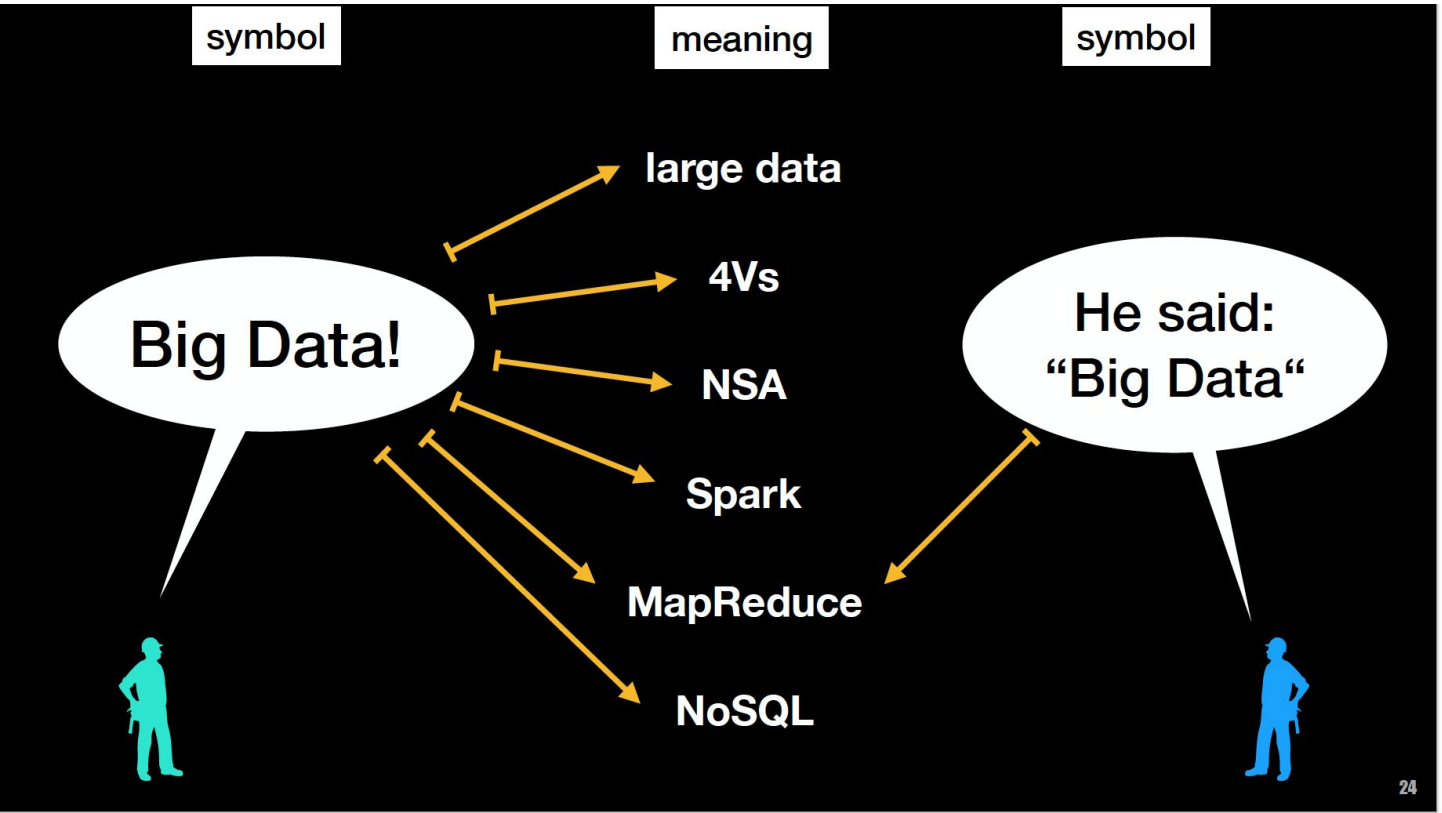


Problem:
ambiguous communication

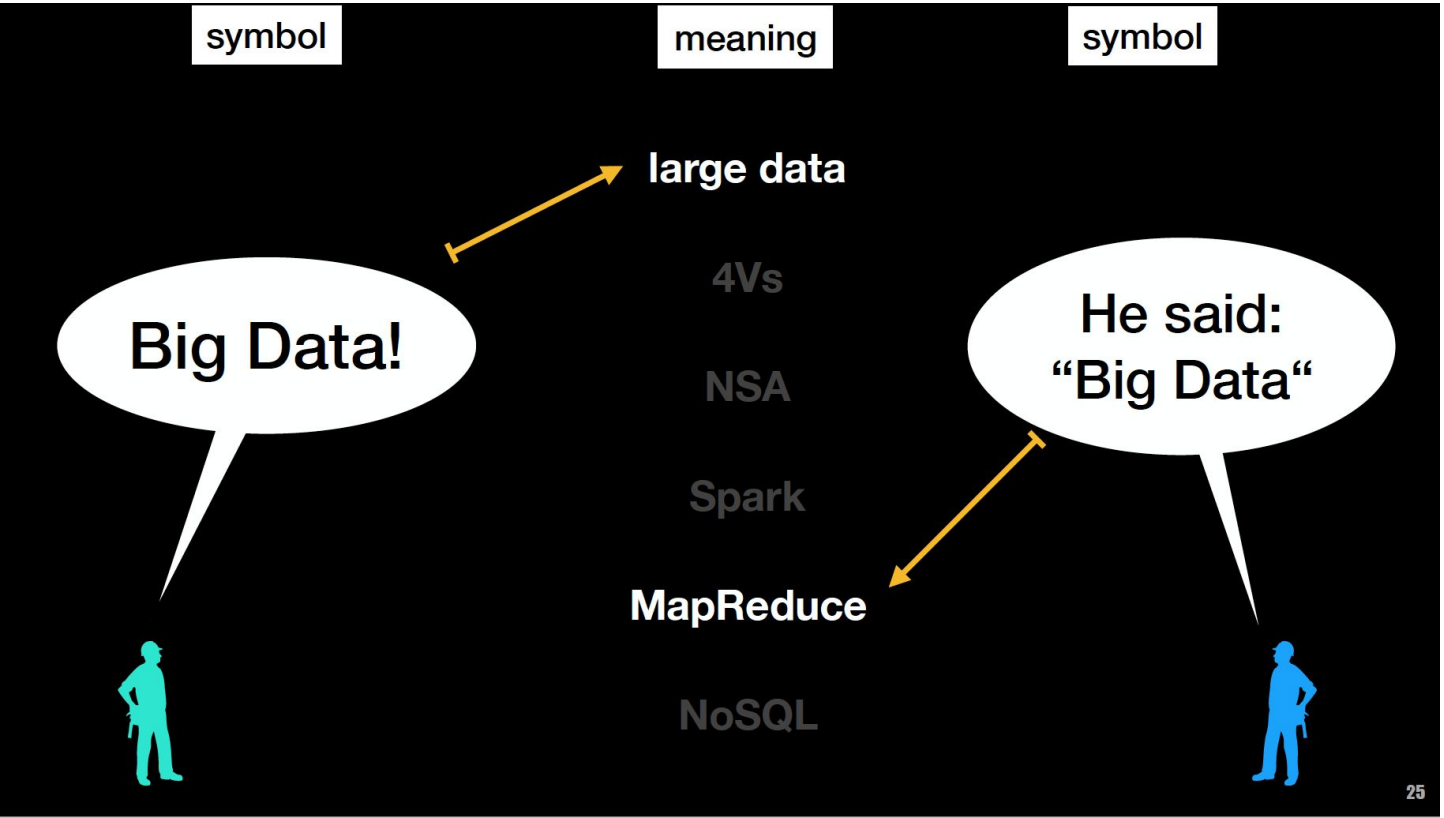
“Big Data”: Cuidado com a Comunicação



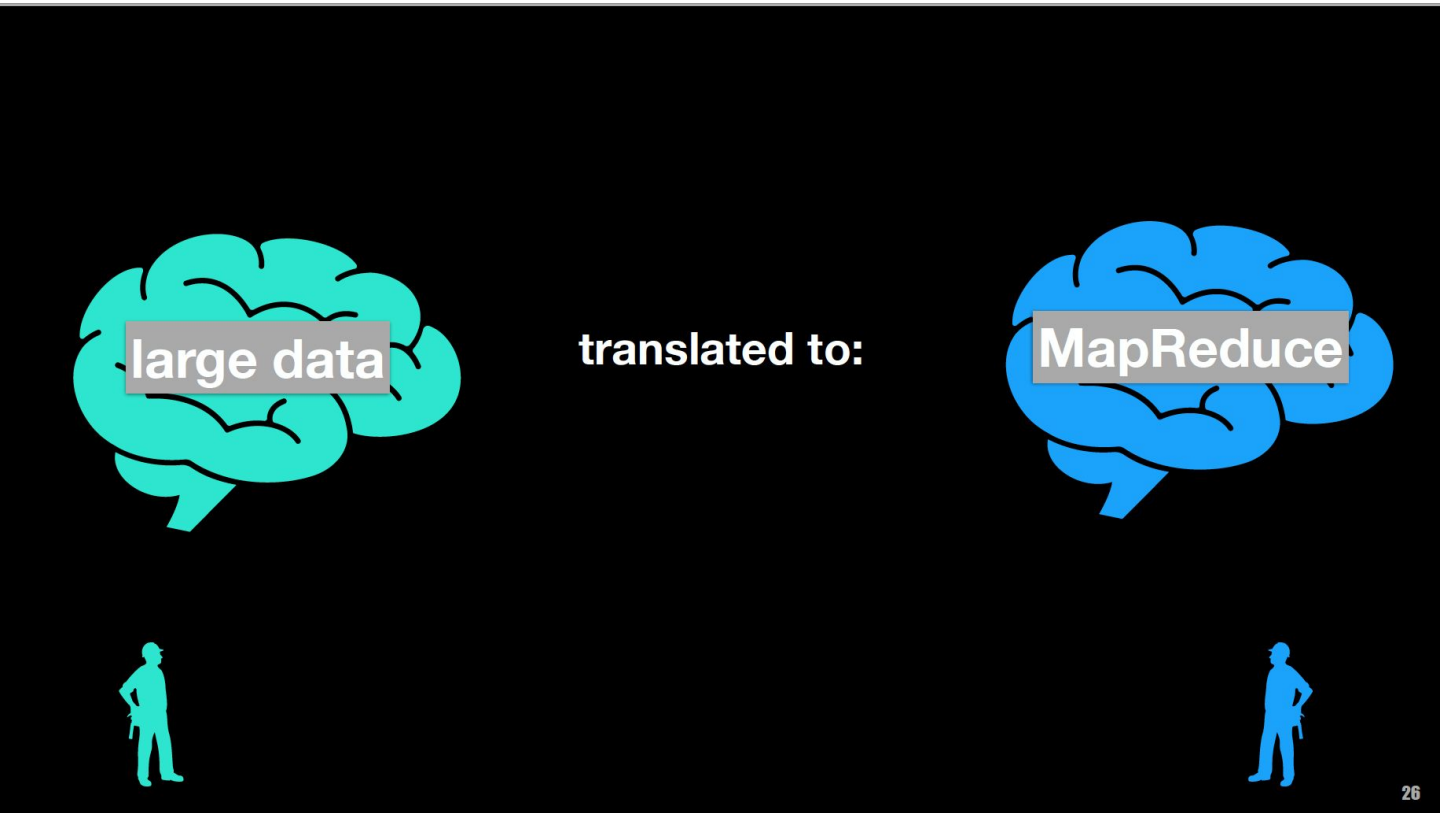
“Big Data”: Cuidado com a Comunicação



“Big Data”: Cuidado com a Comunicação

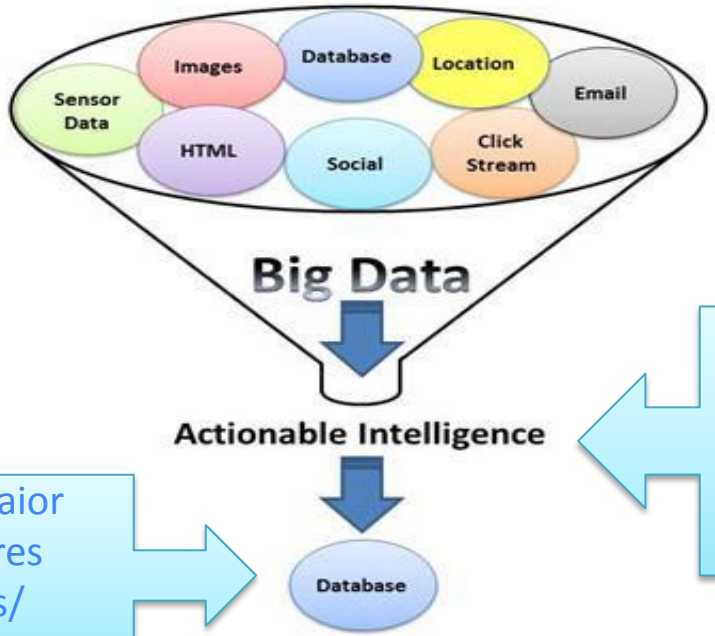


“Big Data”: Cuidado com a Comunicação



Big Data Analytics

Informação original "bruta": geralmente *Terabytes* em dados de baixo valor.



Processo de limpeza, amostragem e relacionamento dos dados

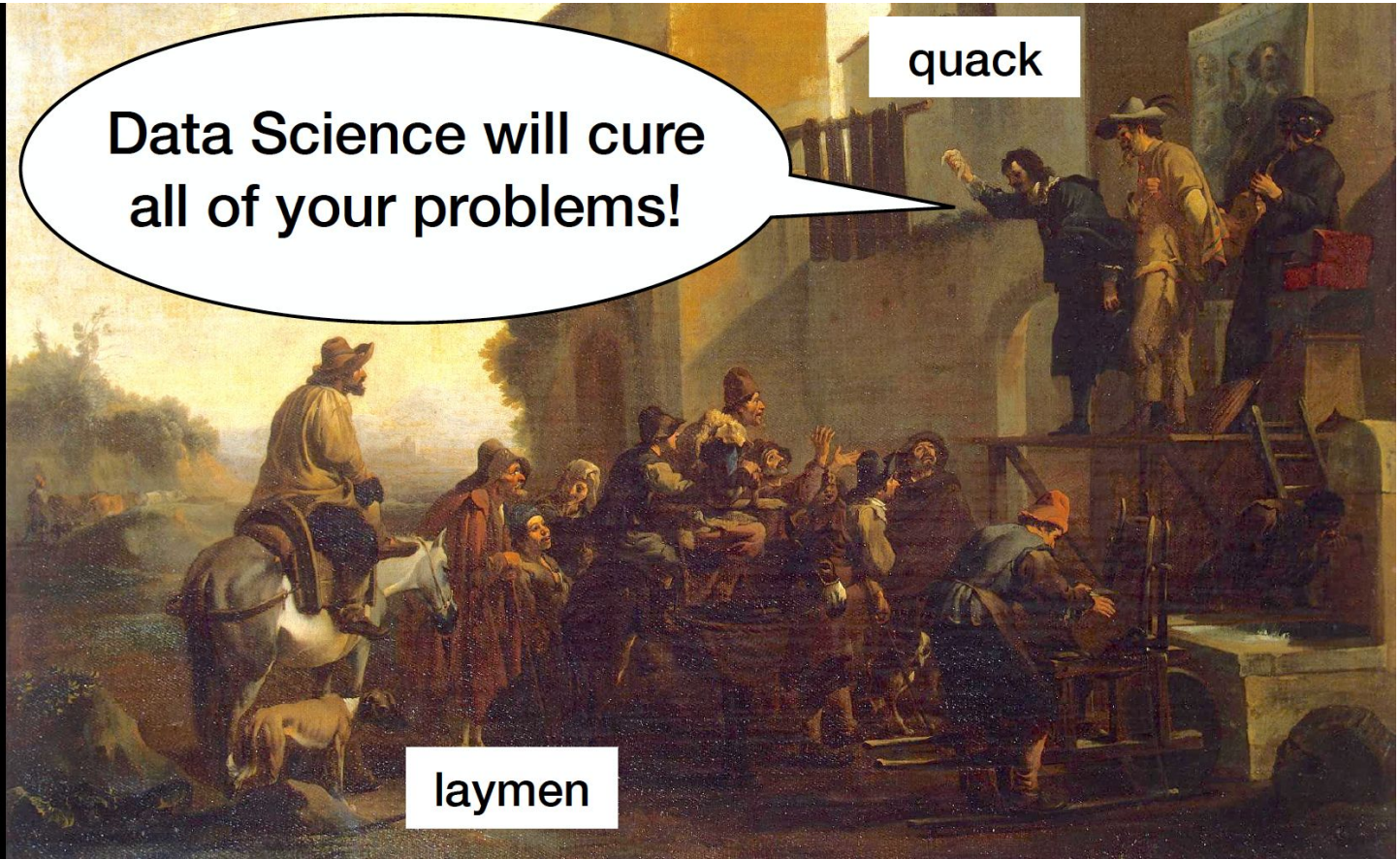
Dados de maior valor: valores agregados/ sumarizados

Ex: Valor total mensal que uma determinada empresa comercializou: NFEs X Cartões de Crédito

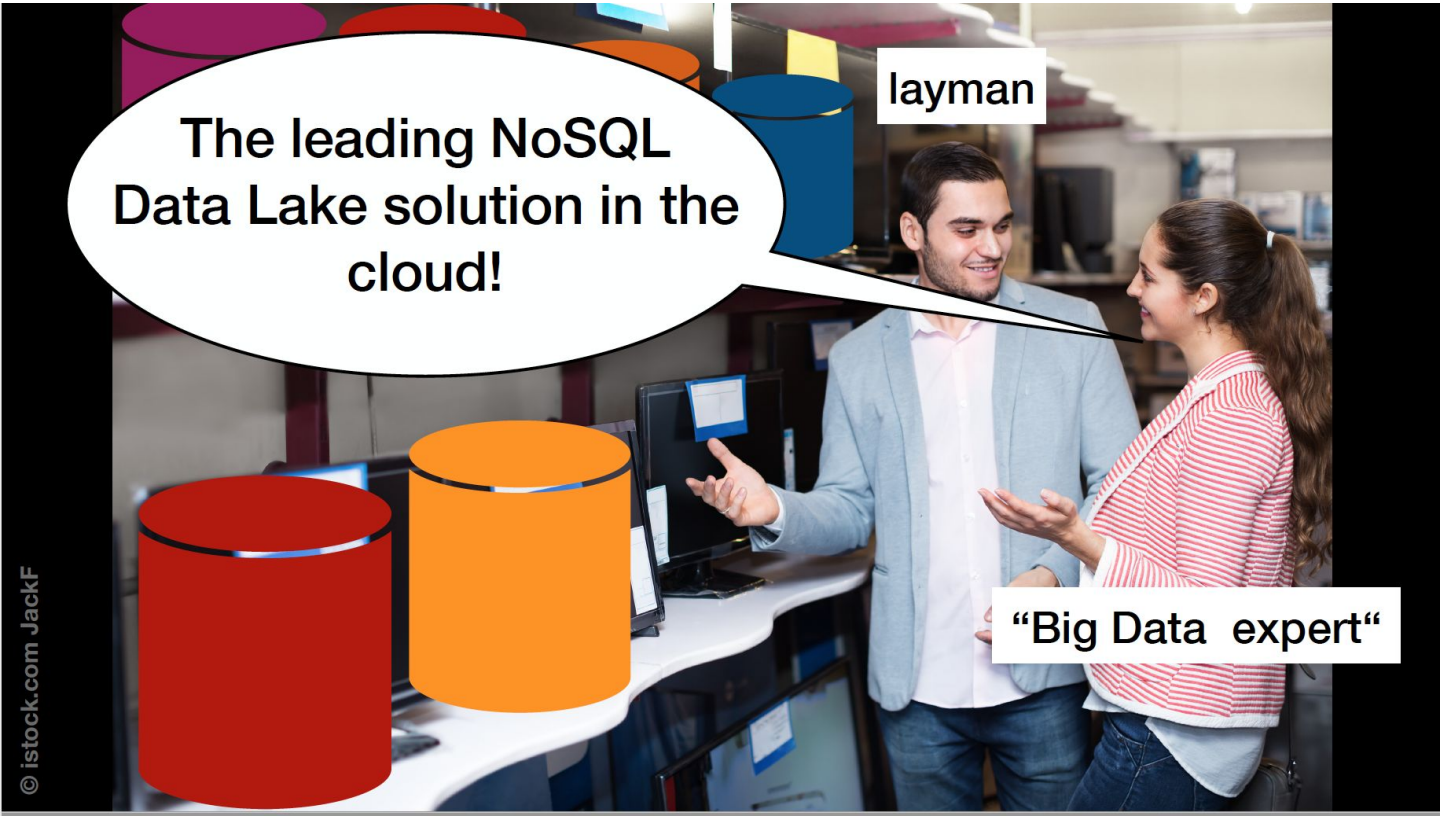
Data Science

Data Science

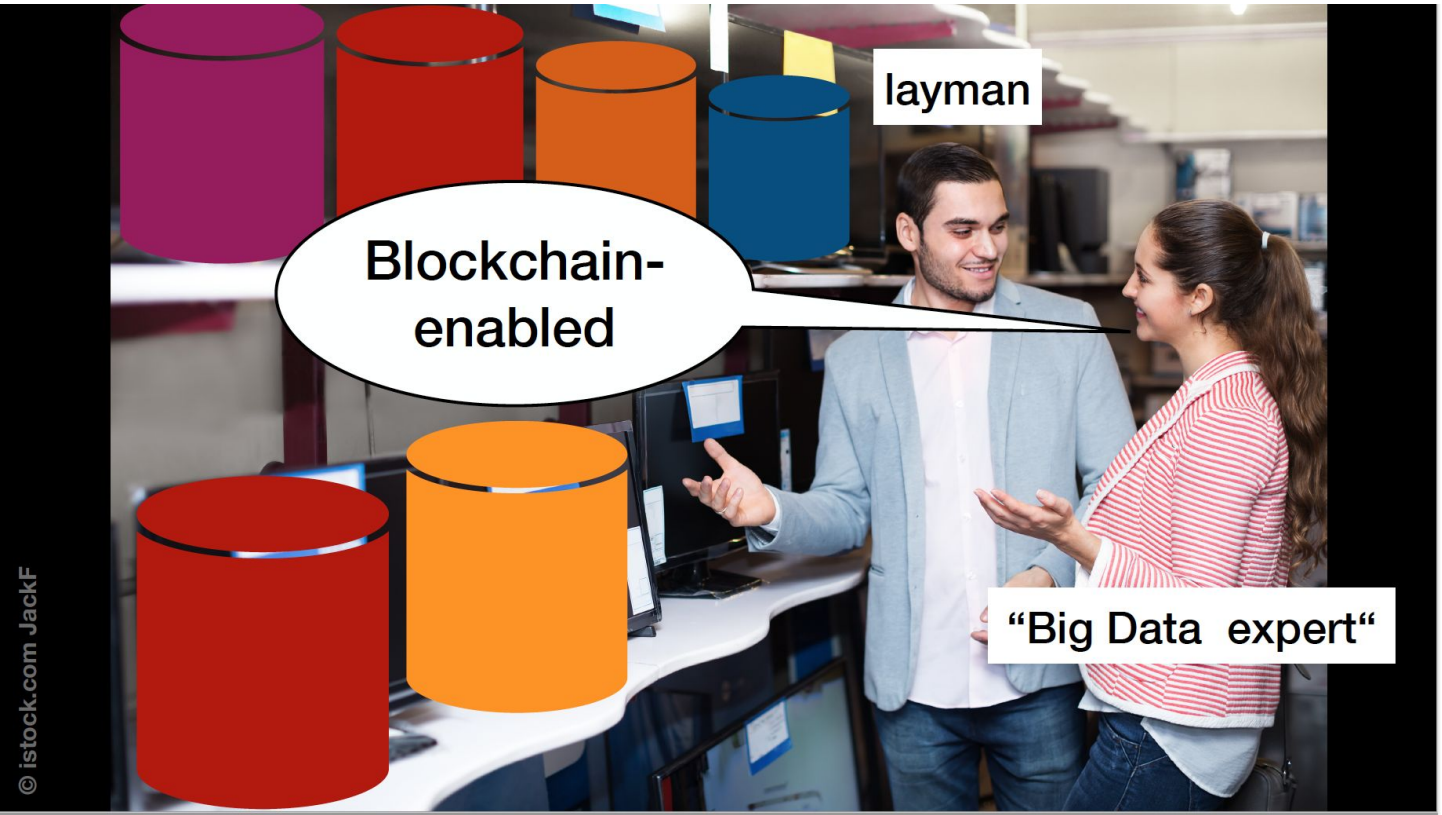
Data Science



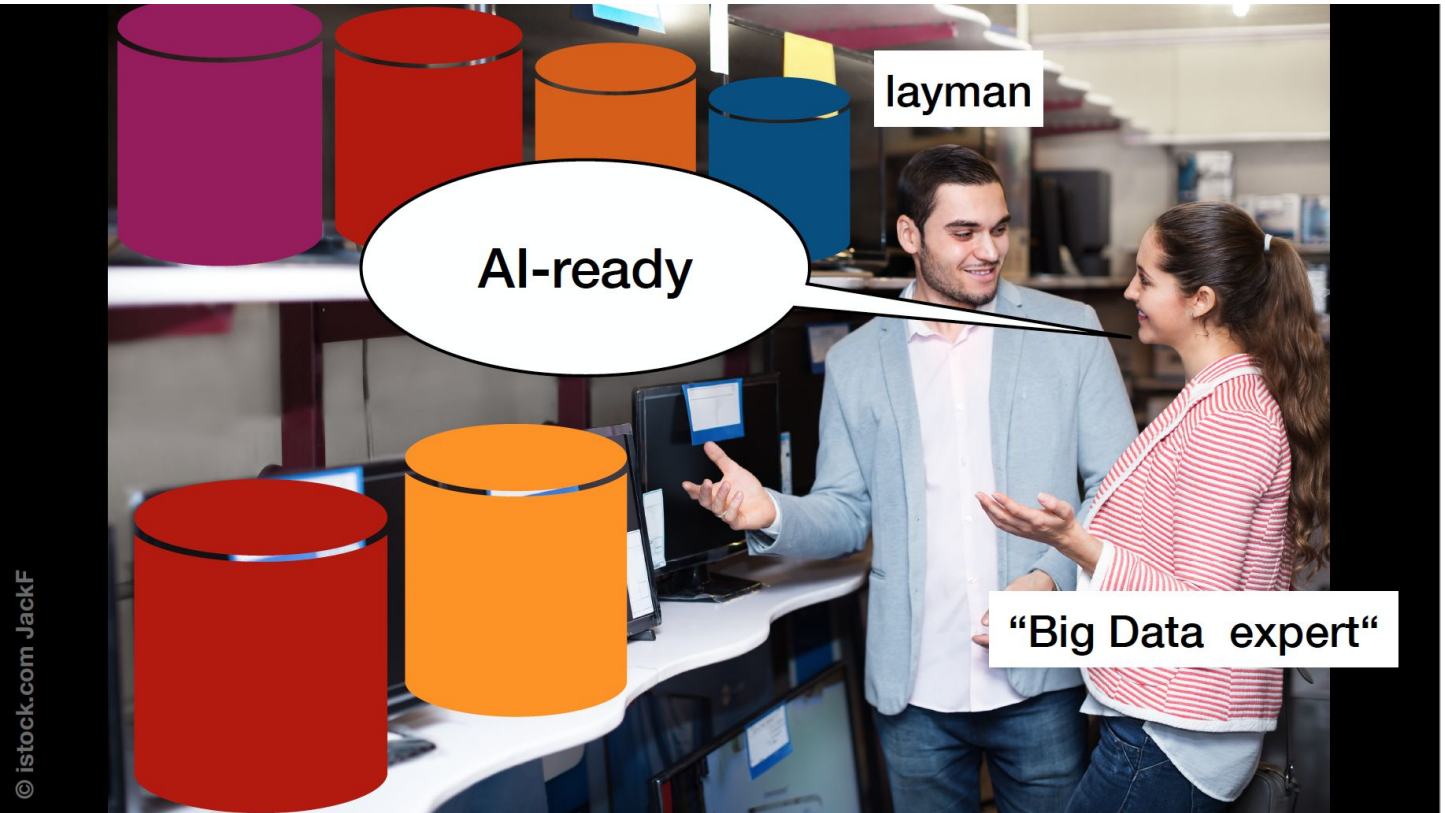
Cuidado com Promessas Milagrosas



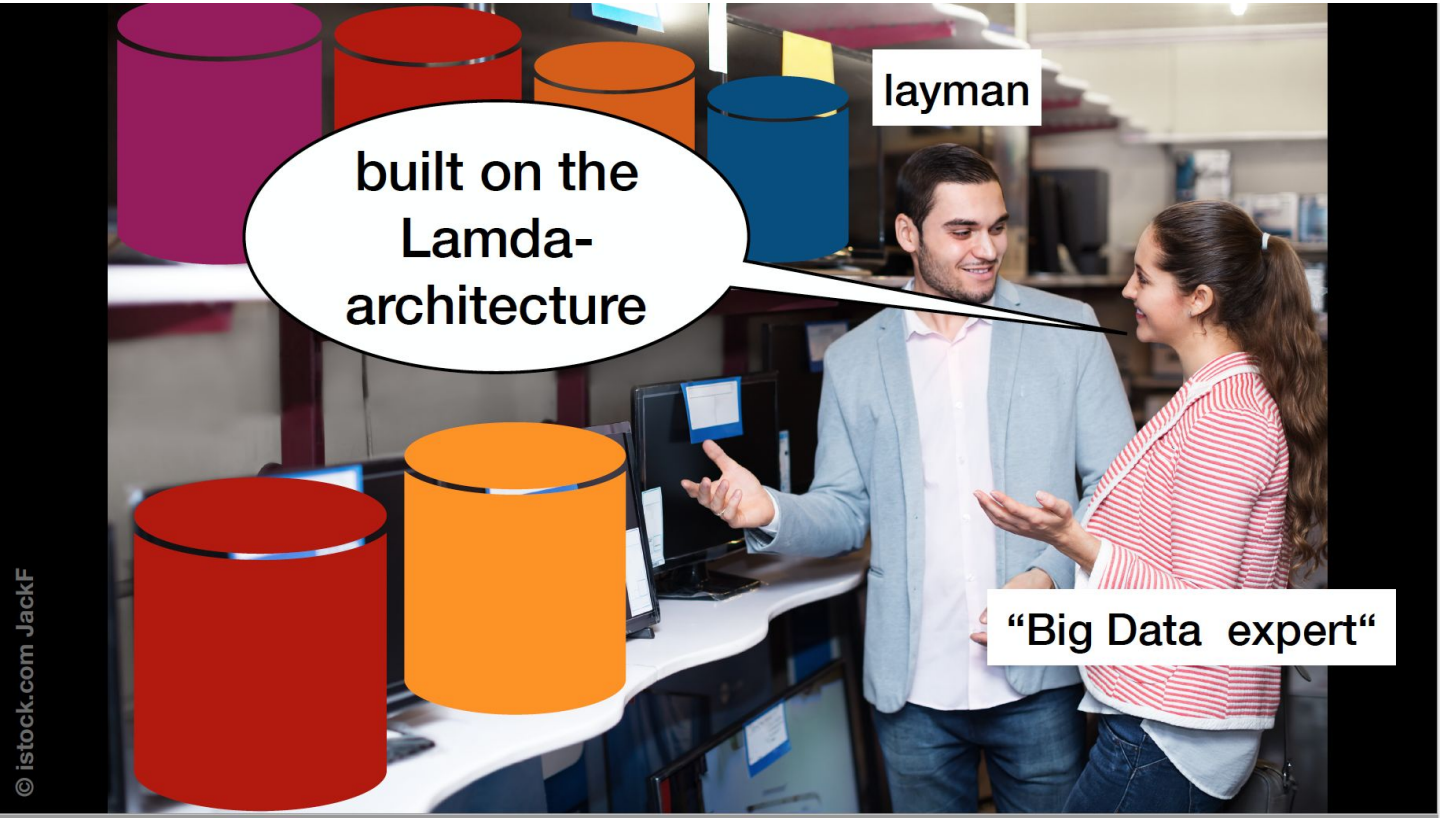
Cuidado com Promessas Milagrosas



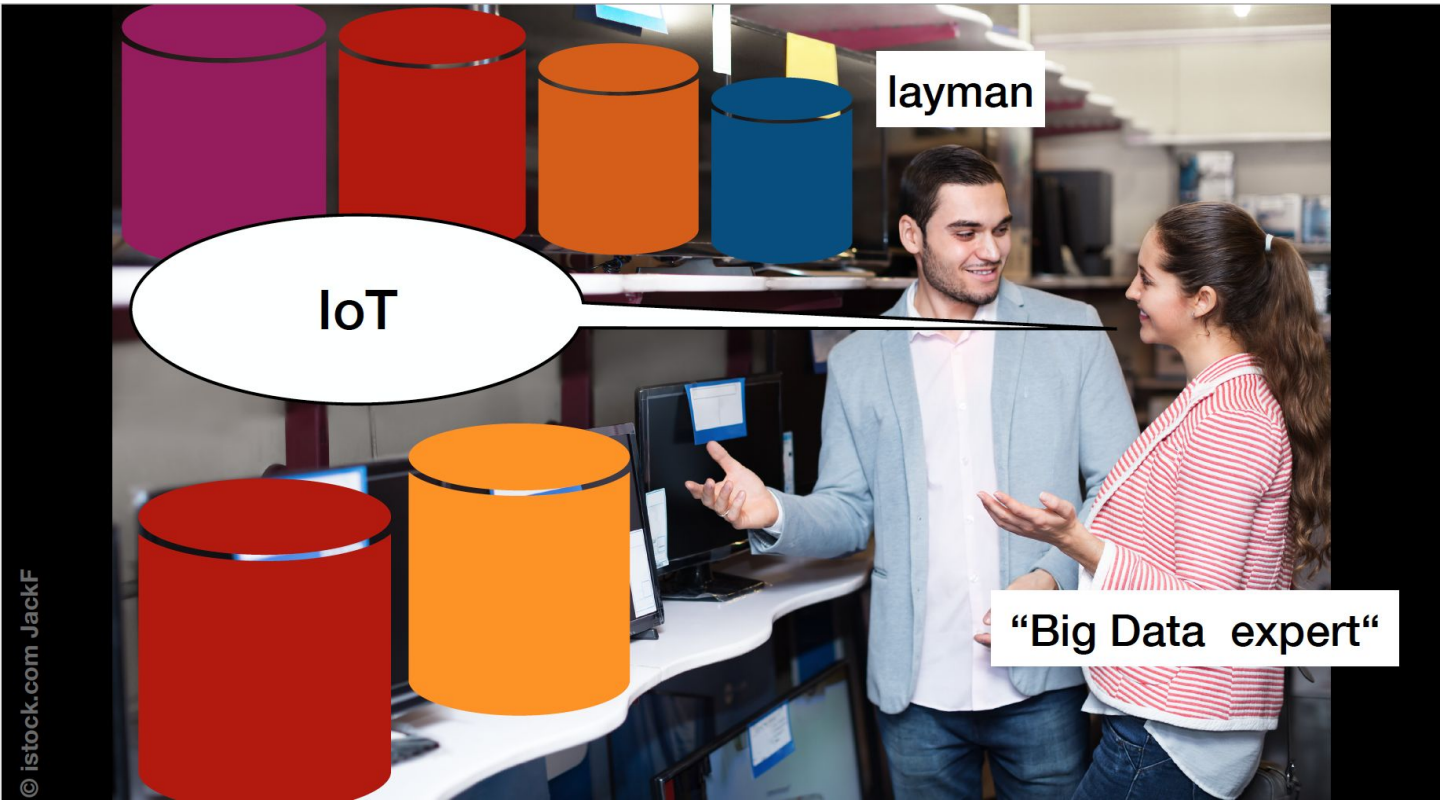
Cuidado com Promessas Milagrosas



Cuidado com Promessas Milagrosas



Cuidado com Promessas Milagrosas

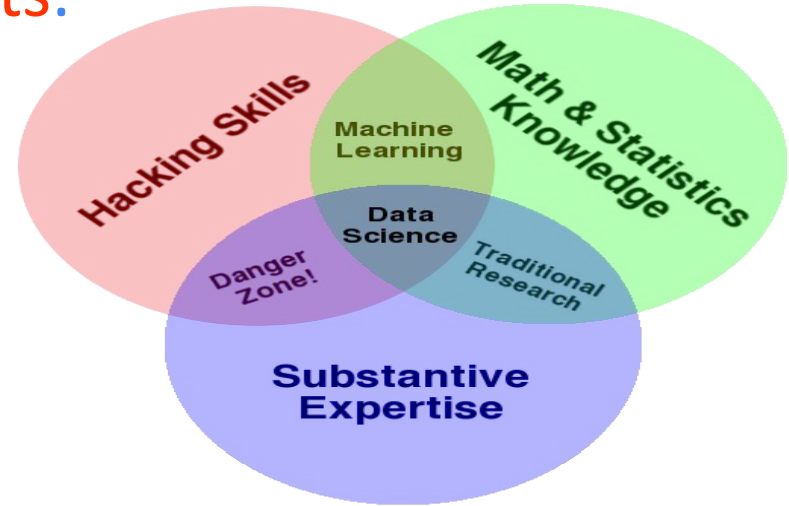


Buzzwords

O que é Data Science???

Data Science is the science which uses computer science, statistics and machine learning, visualization and human-computer interactions to collect, clean, integrate, analyze, visualize, interact with data to create data products.

Turn data into data products.



O que é Data Science???

A Ciência de Dados é uma área multidisciplinar que orienta a extração de informação e conhecimento a partir de grandes volumes de dados [Provost and Fawcett 2013].

Trata da coleta, integração, gerenciamento, exploração dos conjuntos de dados e da utilização do conhecimento adquirido com a finalidade tomar decisões, entender o passado/presente, prever o futuro e criar novos serviços/produtos [Ozdemir 2016]

A Ciência de Dados busca obter novas ideias ("insights") que estejam escondidas nesses grandes repositórios de dados.

Ciclo de Vida de DS

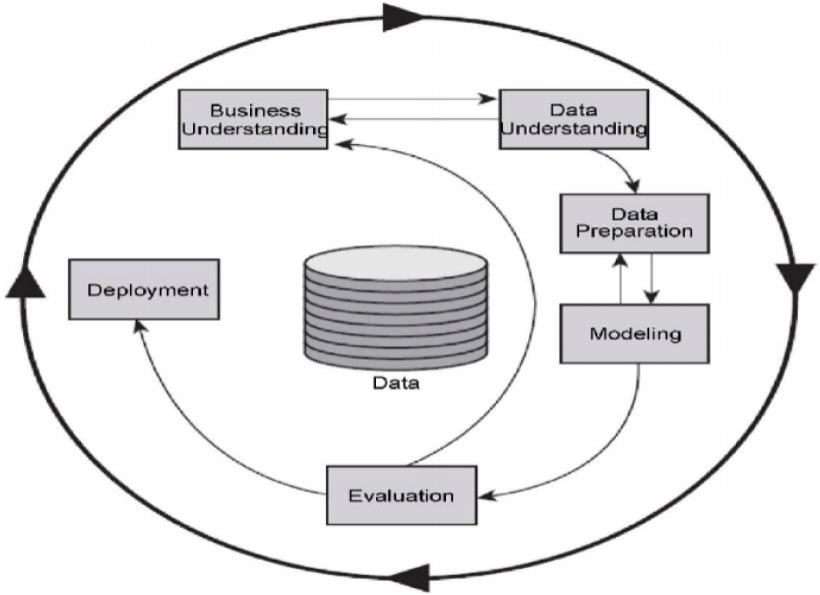
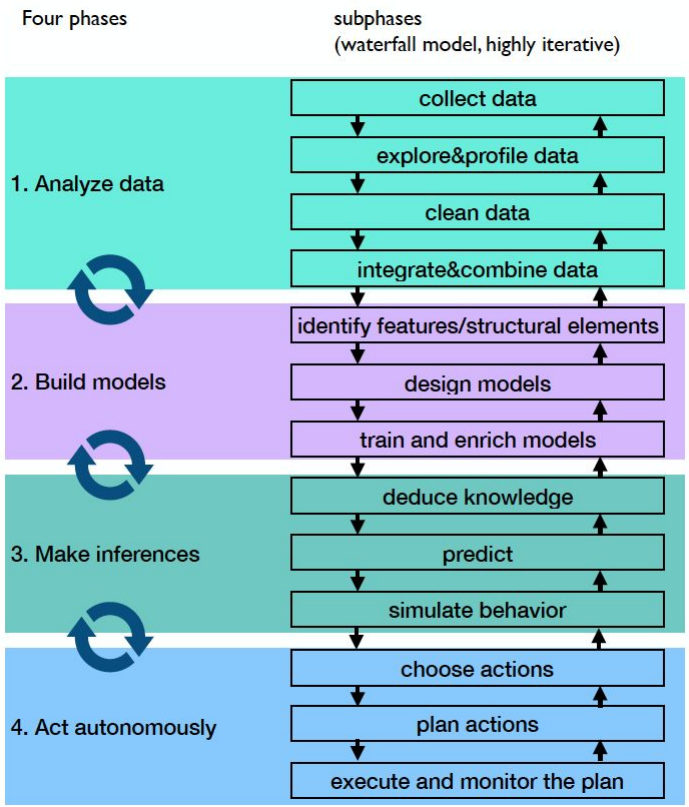


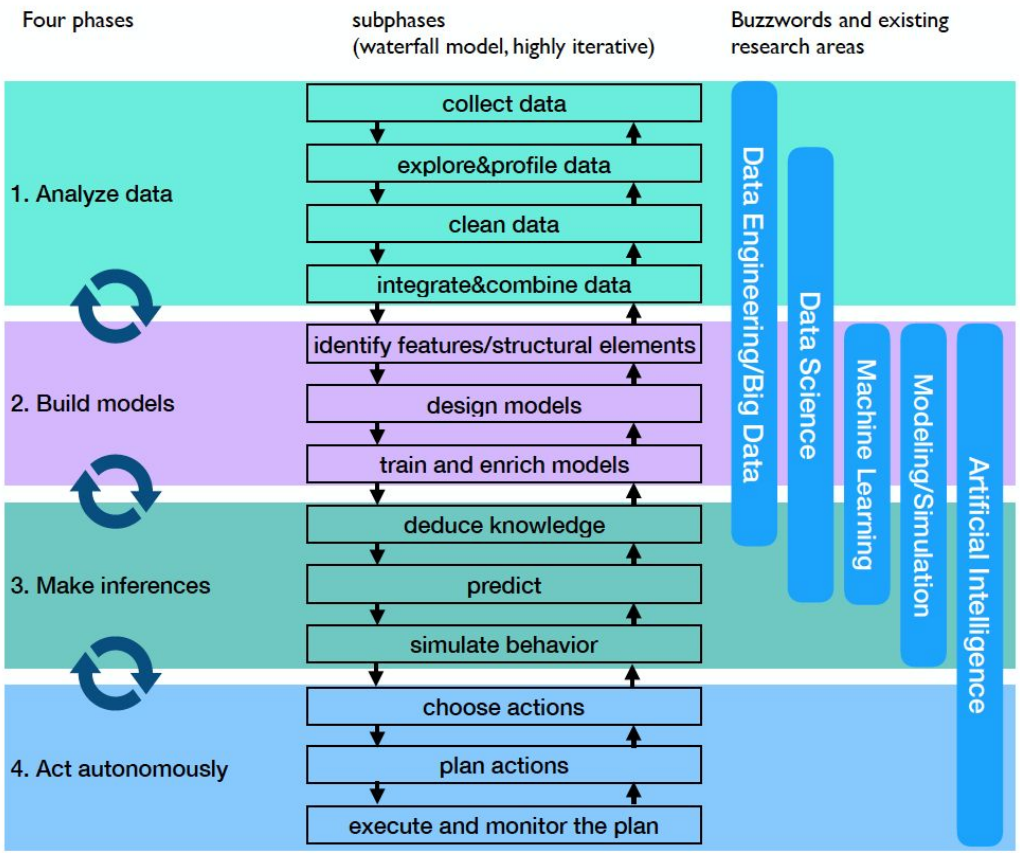
Figura 1. O ciclo de vida da ciência de dados — Fonte: [Chapman et al. 2019]

Ciclo de Vida de DS

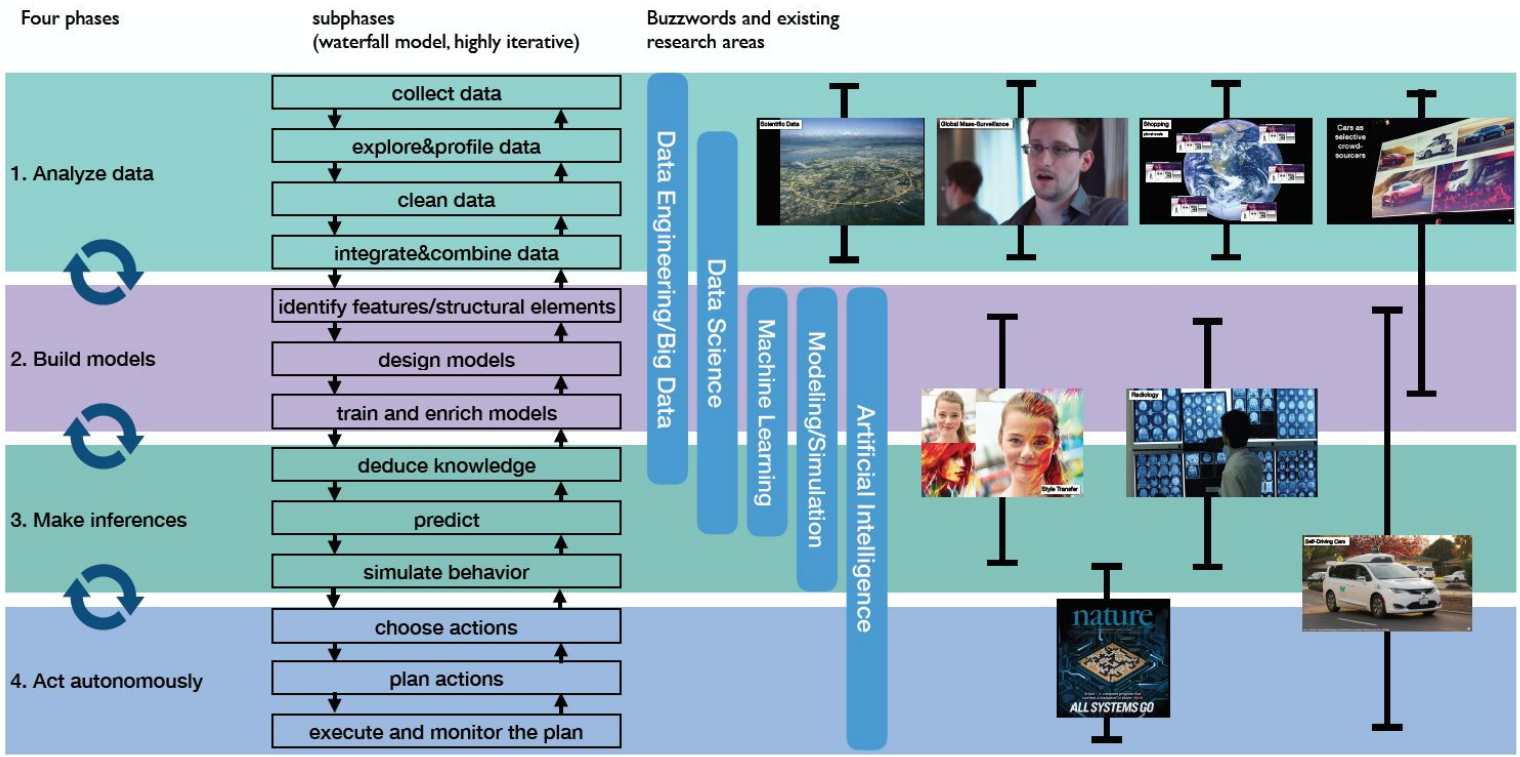
Um Modelo de Processamento para DS



Um Modelo de Processamento para DS



Um Modelo de Processamento para DS



Data Science

- Por que DS se tornou viável?
  - Muitos dados sendo coletados;
  - Avanços em áreas relacionadas:
    - ML, Visualização de Dados, etc;
  - Avanços em hardware:
    - GPUs, TPUs, etc;
  - Muitos dados sendo coletados;
    - Muitas fontes de dados disponíveis:
    - Muitos deles abertos, públicos;
  - Casos de sucesso na pesquisa e indústria;

Desafios

1º. Desafio

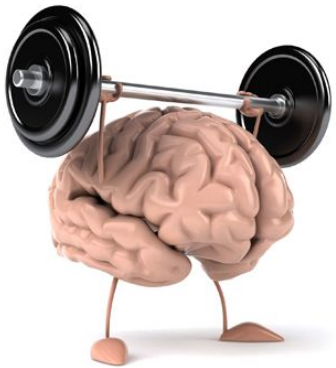
- Capturar e armazenar esse grande volume de dados;

Novos Modelos de Dados

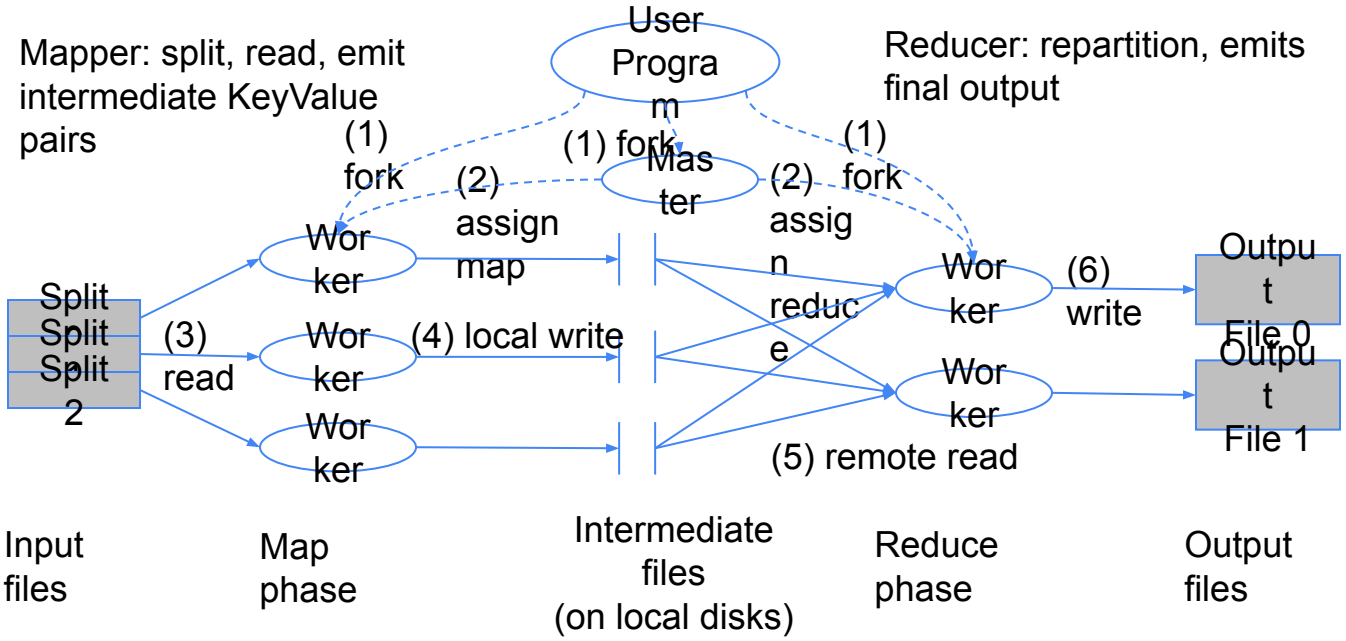
Type	Example
Key-Value Store	 
Wide Column Store	 
Document Store	 
Graph Store	 

2º. Desafio

- Processar esse grande volume de dados;



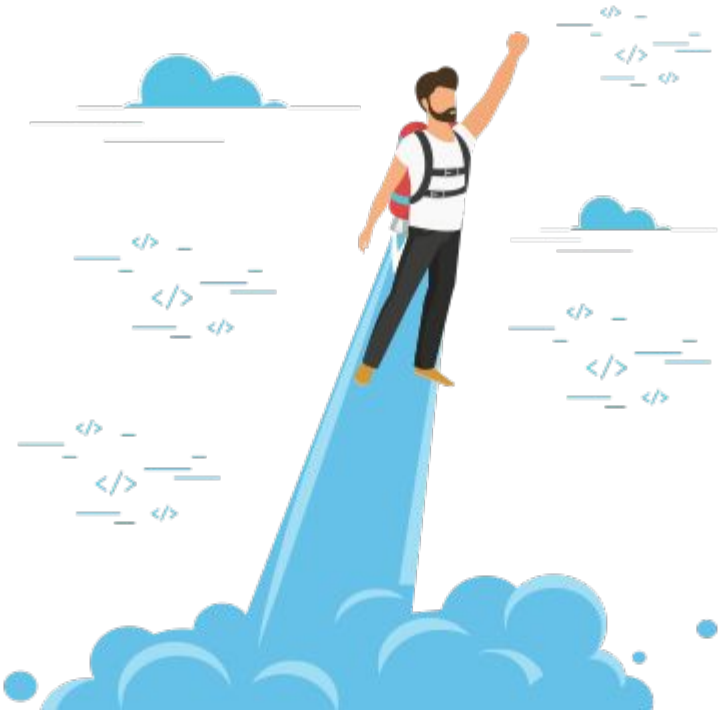
Google MapReduce




You
Cloud
Computing

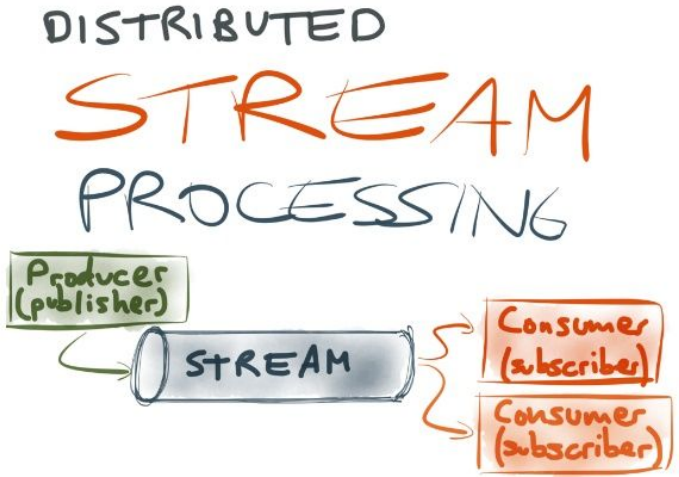
3º Desafio

- Processar um grande volume de dados transmitidos via Streaming;



Processamento de Dados em Streaming

- Ao criar aplicações para coletar, processar e analisar dados em streaming, é necessário levar em consideração uma abordagem diferente daquela utilizada para tratar dados que normalmente são processados em lote;



Processamento de Dados em Streaming

- Frameworks interessantes:

Apache **Kafka**

Apache **Samza**

samza

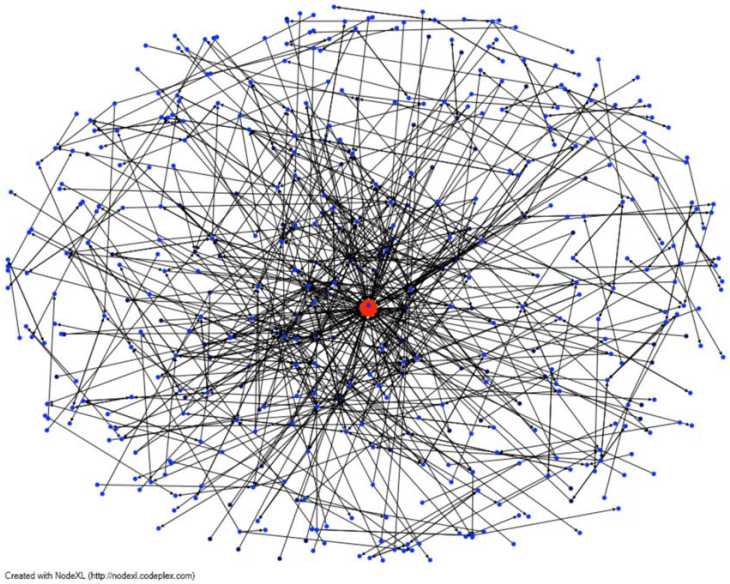
Spark Streaming

APACHE STORM™

Distributed • Resilient • Real-time

4º. Desafio

- Processar dados em grafos;



GraphX

neo4j

Gephi

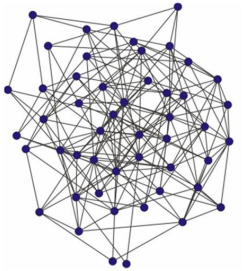
powergraph

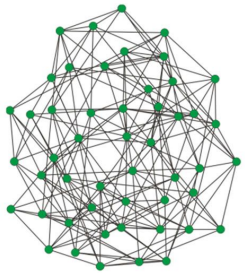
NetworKit

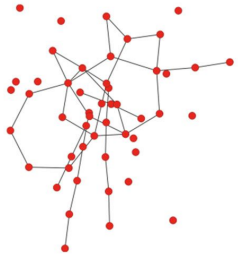
Large-Scale Network Analysis — Interactive and Fast!

5º. Desafio

- Processar redes complexas;







6º Desafio

- Garantir a privacidade dos usuários;

7º Desafio

- Assegurar que os algoritmos (soluções) não promovam qualquer tipo de discriminação;

8º Desafio

- Simplificar a pilha de Software;

Data Store

mongoDB.

MySQL

Joins are what RDBMS's do for a living

Transform

Spark

SQL

If you change the way you look at things, the things you look at change.

Model

scikit learn

pandas

Spark

AI usually beats natural stupidity

Visualize

ggplot2

Dont trust everything you see, even salt looks like sugar

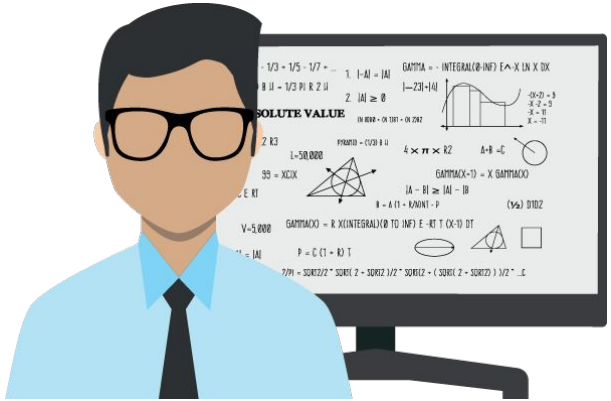
Other Tools

kafka

LUIGI

I have enough tools, said no data scientist ever

Full Stack DC
Data Science



MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th 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

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

DOMAIN KNOWLEDGE & SOFT SKILLS

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

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

Hands-On

Como seremos avaliados?

Dúvidas?

Email: wellington@crateus.ufc.br

UNIVERSIDADE FEDERAL DO CEARÁ
CAMPUS DE CRATEÚS

enginelab
Laboratório de Engenharia de Software e Sistemas