Introdução a Ciência de Dados

Wellington Franco
Universidade Federal do Ceará – UFC
Campus da UFC em Crateús
wellington@crateus.ufc.br

Autor

- Wellington Franco
  - Email: wellington@crateus.ufc.br
  - Minicurrículo:
    - Graduado em Ciência da Computação pela UECE
    - Mestrado em Lógica e IA pela UFC
    - Doutor em Banco de dados e IA pela UFC

Big Data

- Estrutura do Curso:
  - Módulo 1: Análise Estatística de Dados
  - Módulo 2: Machine Learning
  - Módulo 3: Infraestrutura de Big Data

Big Data

- Estrutura do Curso:
  - Módulo 1: Análise Estatística de Dados
    - Objetivo:
      - Introduzir conceitos estatísticos básicos a fim de fornecer uma visão geral do conjunto de dados.

Big Data

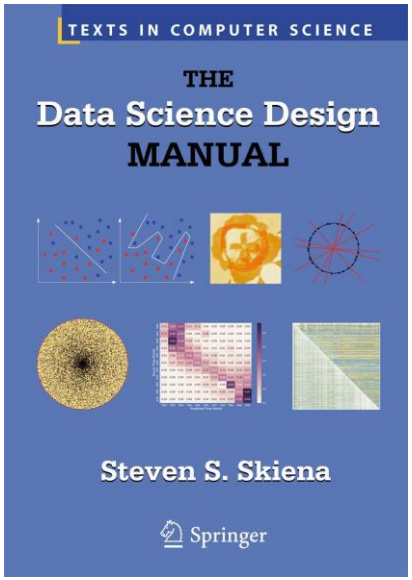
- Estrutura do Curso:
  - Módulo 1: Análise Estatística de Dados
    - Introdução à Data Science
    - Engenharia de Dados
    - Distribuição dos Dados
    - Conceitos Básicos de Visualização de Dados

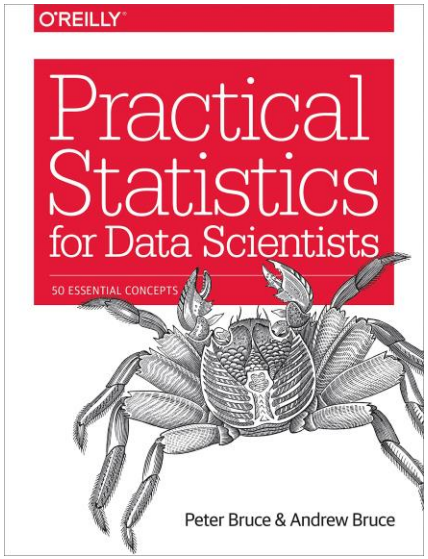
Big Data

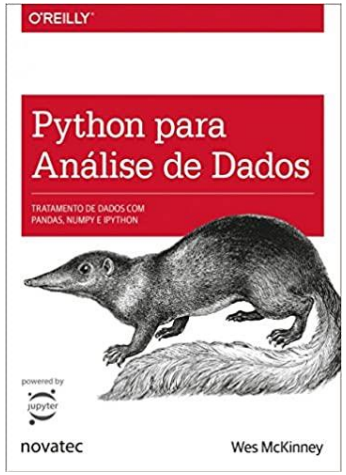
- Estrutura do Curso:
  - Módulo 1: Análise Estatística de Dados
    - Bibliografia:
      - The Data Science Design Manual, 1st ed. 2017. Steve Skiena.
      - Python para Análise de Dados: Tratamento de dados com Pandas, NumPy e IPython, Novatec Editora; 1ª Edição, 2019. Wes McKinney.
      - Practical 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.

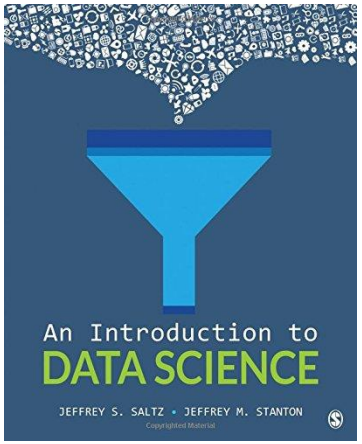
Big Data

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









Big Data

- Estrutura do Curso:
  - Módulo 2: 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

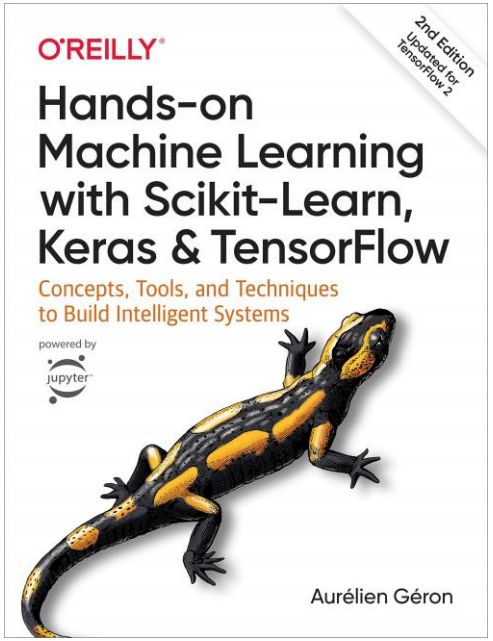
- Estrutura do Curso:
  - Módulo 2: Machine Learning
    - Fundamentos de Machine Learning
    - Problemas de Regressão
    - Problemas de Classificação
    - Problemas de Clustering
    - Redes Neurais Artificiais
    - Introdução às Redes Neurais Profundas
    - Guideline para Aplicar Machine Learning com Segurança
    - Aplicando Machine Learning em Grafos de Conhecimento

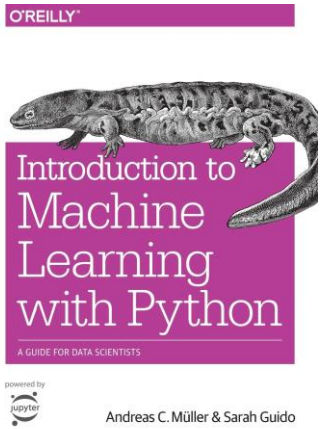
Big Data

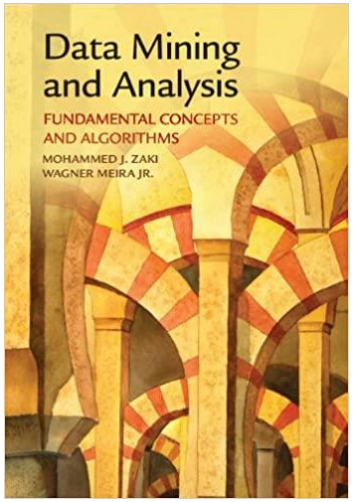
- Estrutura do Curso:
  - Módulo 2: Machine Learning
    - Bibliografia:
      - Hands–On Machine Learning with Scikit–Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, 2o. Edition, 2019. Aurelien Geron.
      - Introduction to Machine Learning with Python: A Guide for 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.

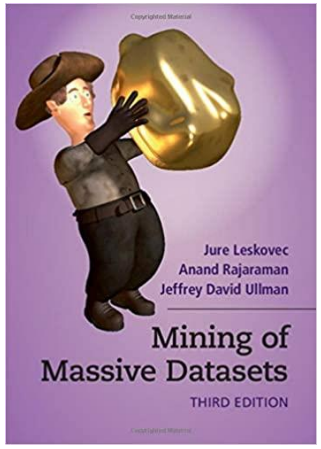
Big Data

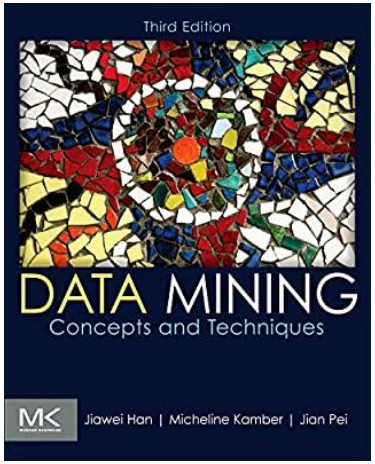
- Estrutura do Curso:
  - Módulo 2: Machine Learning
    - Bibliografia:

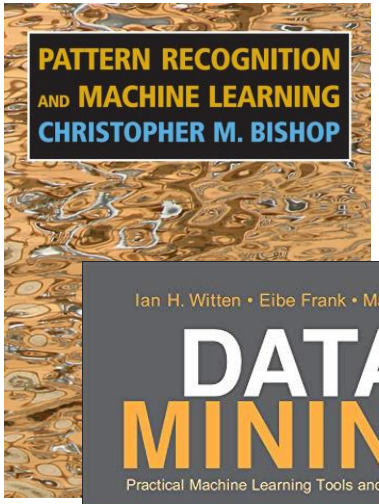


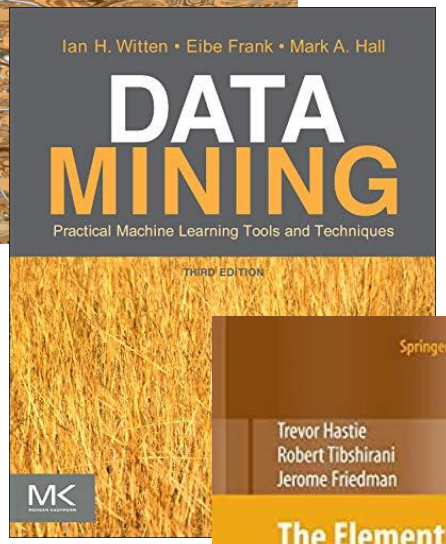


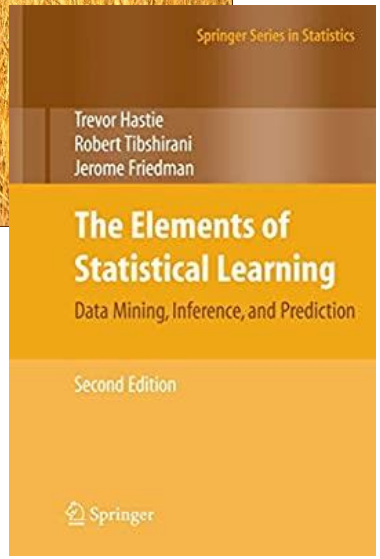












Big Data

- Estrutura do Curso:
  - Módulo 3: 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).
      - Conhecer os princípios teóricos que norteiam as arquiteturas técnicas de Big Data (ACID, CAP, multi-tenancy, replication factor).
      - Discutir os principais tópicos técnicos que afetam provisionamento, desempenho e manutenção das soluções de Big Data.

Big Data

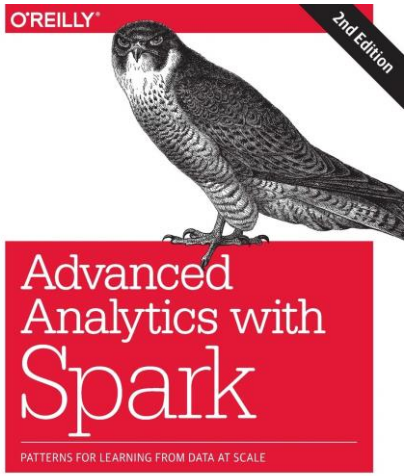
- Estrutura do Curso:
  - Módulo 3: Infraestrutura de Big Data
    - Introdução a Big Data
    - Montando um Ambiente de Big Data
    - Desenvolvendo com o Spark
    - Utilizando o Apache Spark
    - Bancos de Dados NoSQL

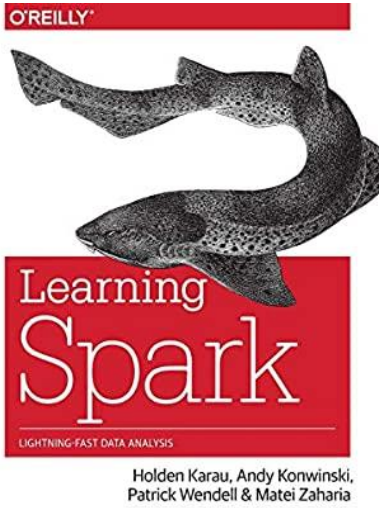
Big Data

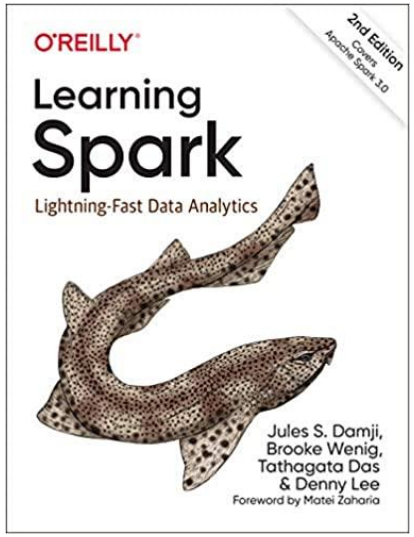
- Estrutura do Curso:
  - Módulo 3: Infraestrutura de Big Data
    - Bibliografia:
      - Learning Scala Programming: Object-oriented Programming Meets Functional Reactive to Create Scalable and Concurrent Programs, First Edition, 2018. Vikas Sharma.
      - Advanced Analytics with Spark: Patterns for 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.

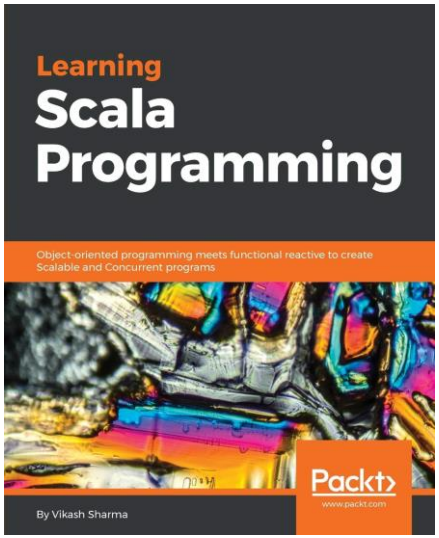
Big Data

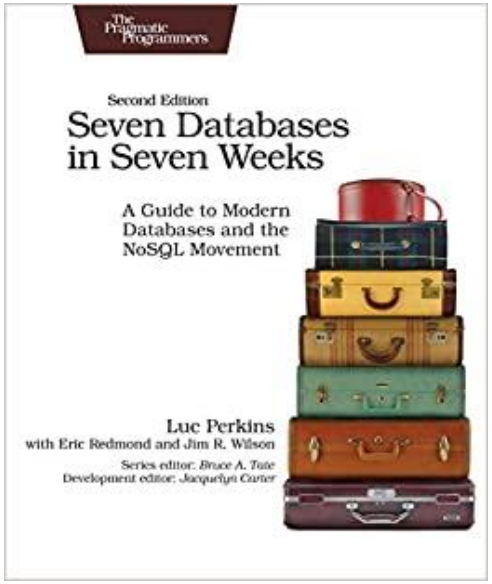
- Estrutura do Curso:
  - Módulo 3: Infraestrutura de Big Data
    - Bibliografia:











Big Data

- Metodologia
  - Aulas expositivas:
    - Discursão teórica;
  - Exercícios práticos:
    - Individuais ou em grupo;
  - Mentoria:
    - Discursões de problemas específicos;

Apresentação da Turma

Nome

Área de Formação

Experiência em Programação

Experiência em Big Data / Data Science

Qual a primeira coisa que vem na sua cabeça ao ouvir o termo Big Data?

Big Data

Módulo 1: Análise Estatística de Dados

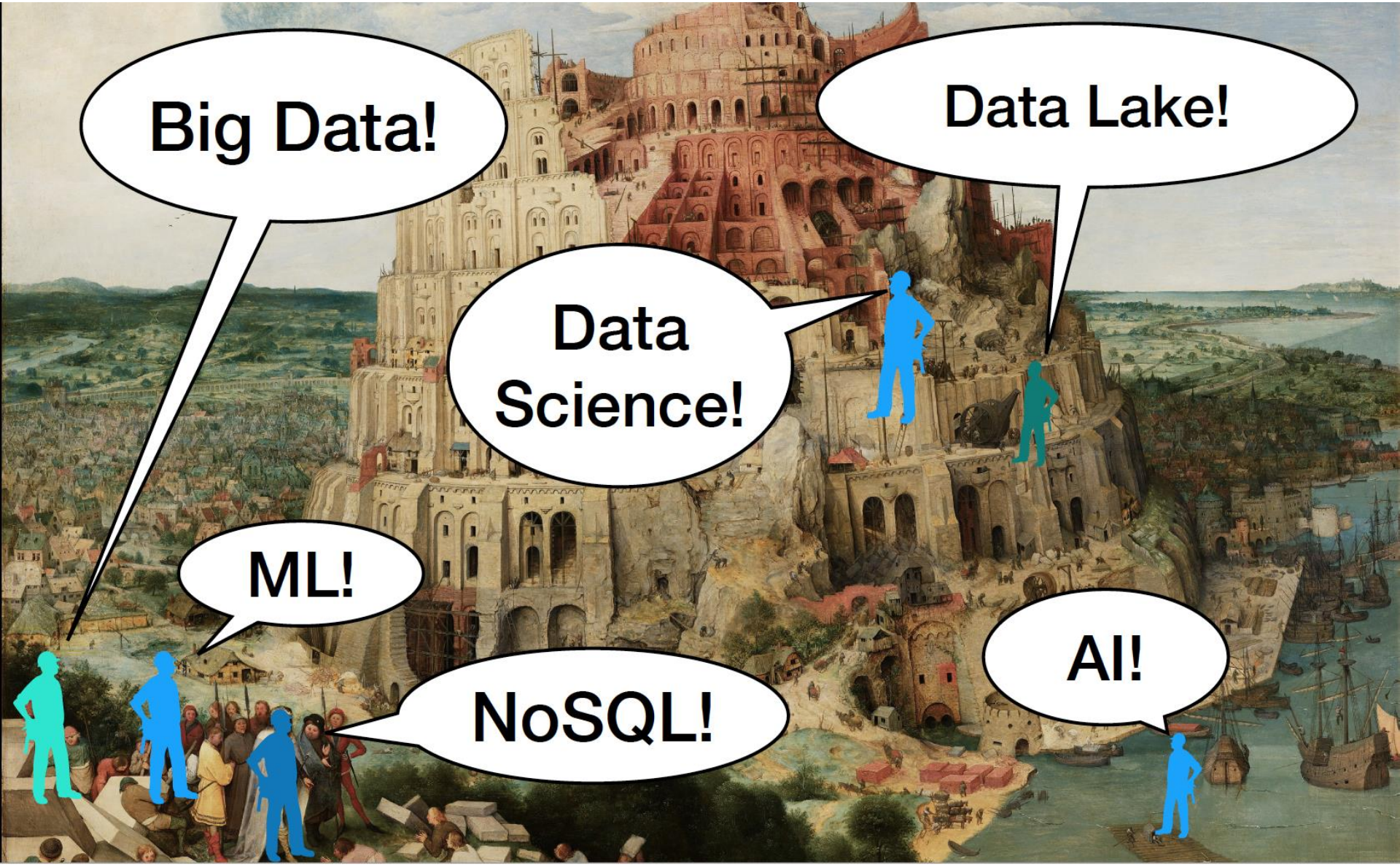
Aula 1: Introdução à Data Science

Introdução à Data Science

- Considerações Iniciais
  - Big Data e Data Science
    - São áreas relativamente recentes;
    - Ainda em pleno desenvolvimento;
    - São áreas multidisciplinares
      - Computação
        - » Banco de Dados, Inteligência Artificial, Visualização de Dados, ...
      - Estatística
      - Matemática
      - Negócio
    - Termos utilizados com finalidade comercial;

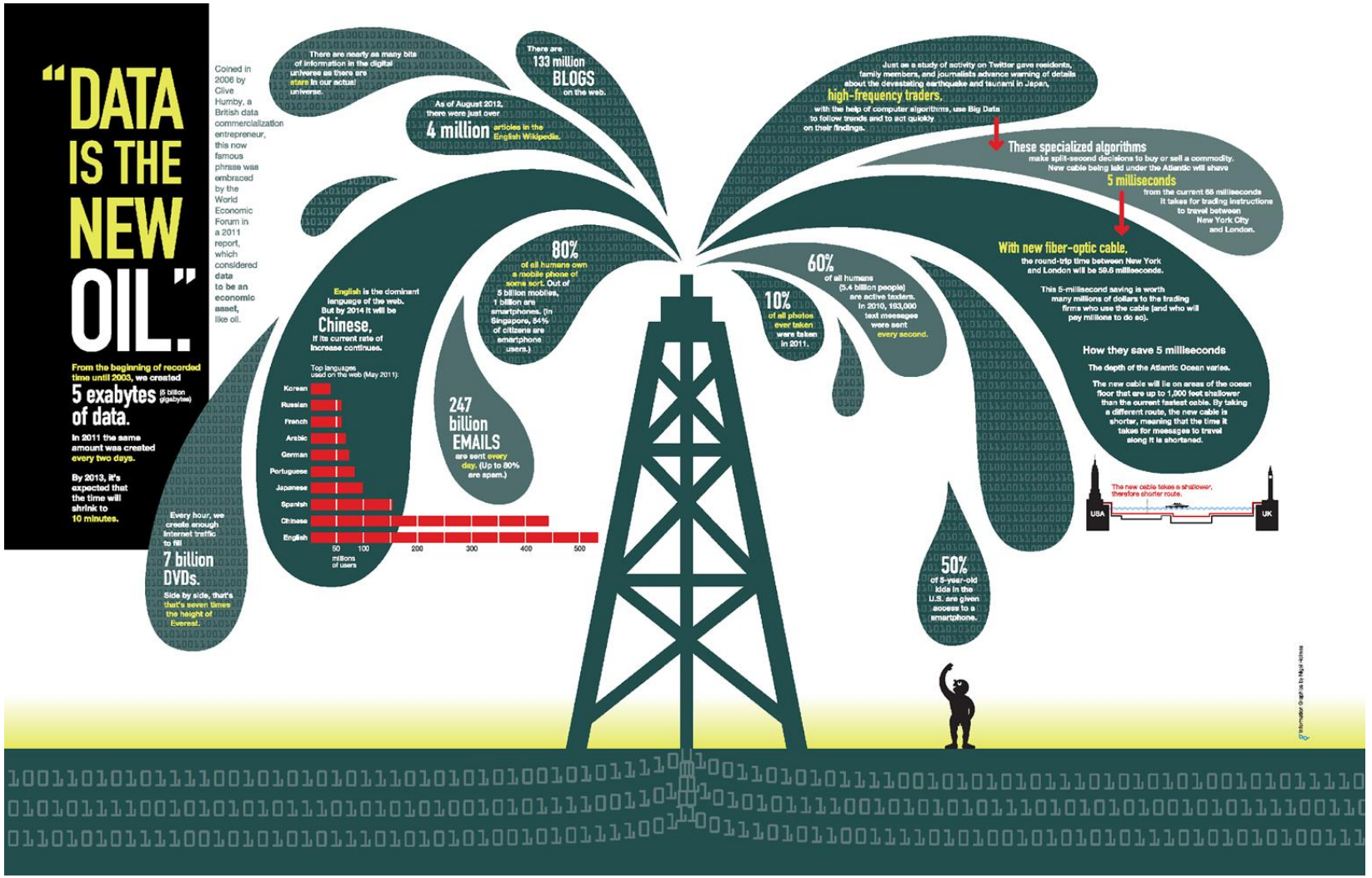
Introdução à Data Science

- Considerações Iniciais
  - Big Data e Data Science
    - Não existe uma definição única (padrão);
    - Não existe uma taxonomia ou classificação padrão;
    - Profissionais de diferentes áreas possuem visões diferentes;
  - Existem muitos termos inter-relacionados:
    - Big Data, Data Science, Data Mining, Machine Learning, Pattern Recognition, Analytics, Predictive Analytics, Big Data Analytics, ...


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,...

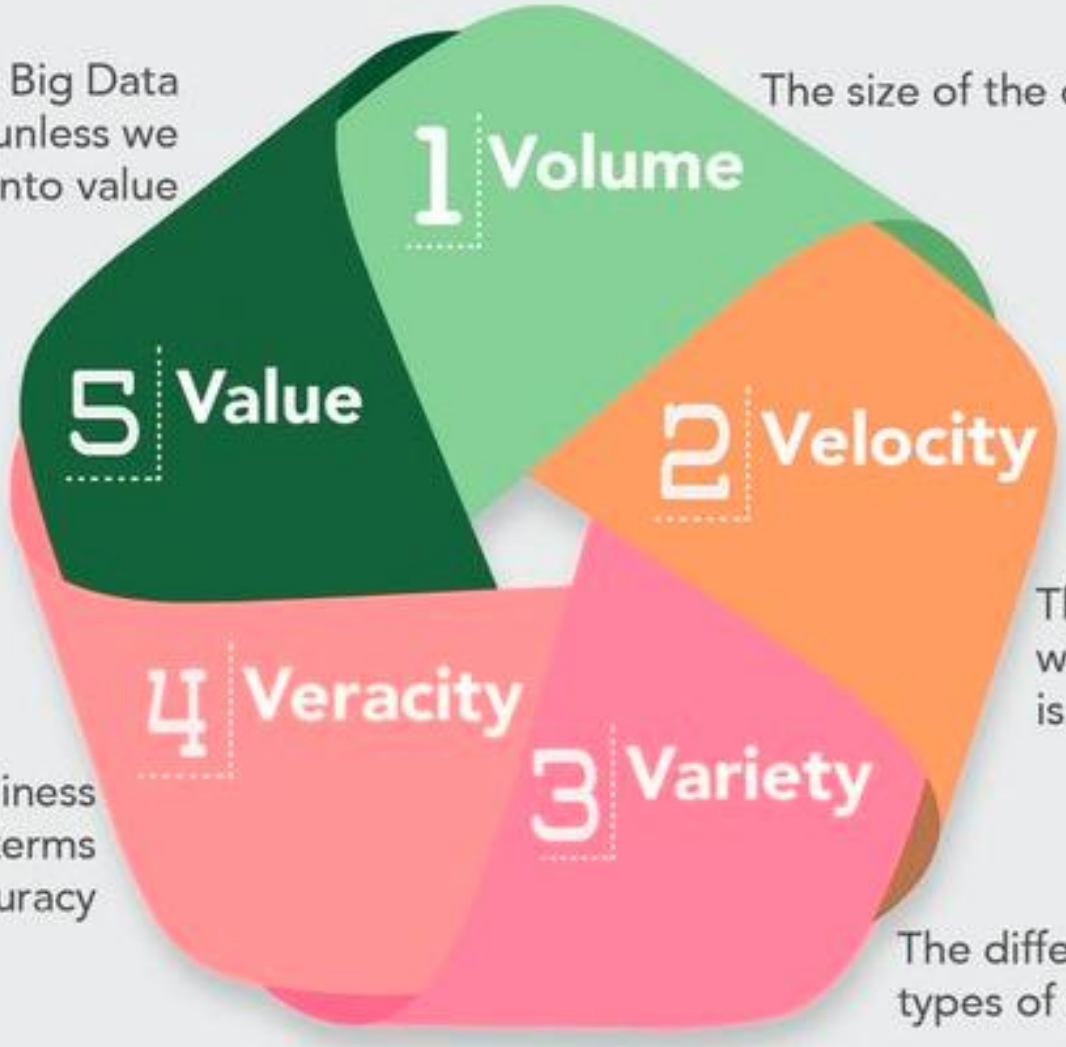
Vamos tentar organizar essa bagunça...


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

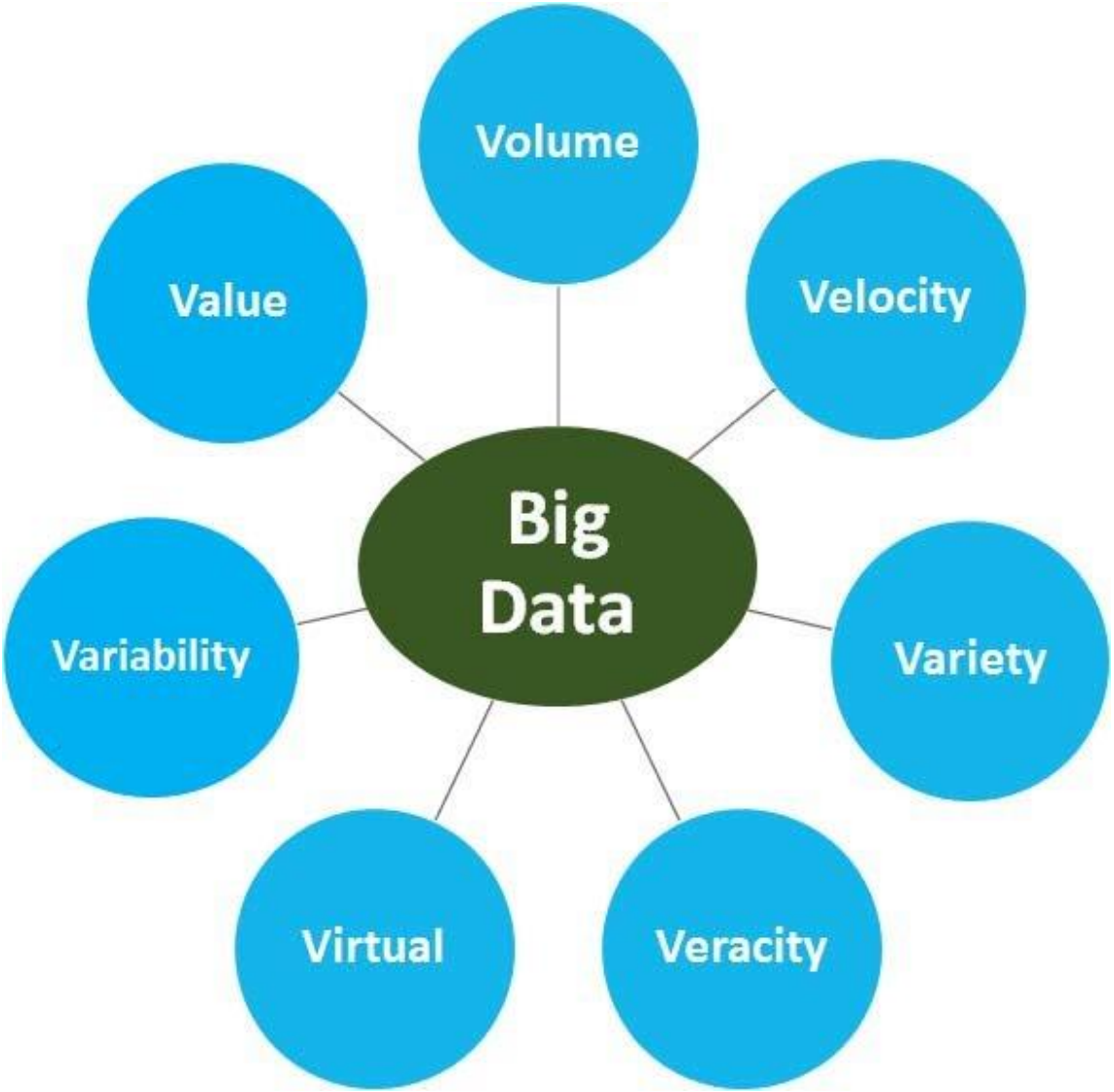
Big Data

- 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

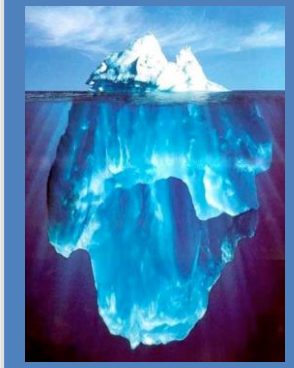


THE 7 Vs OF BIG DATA



“Big Data”: De onde vêm os dados

Tudo que acontece On-line



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

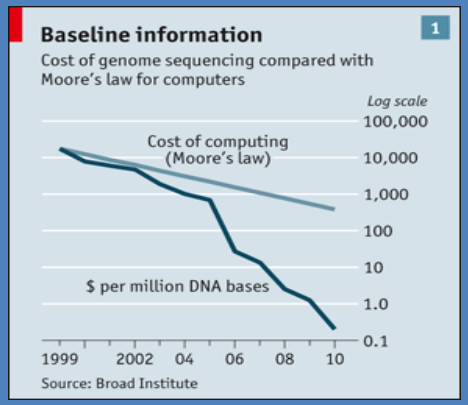
Gerados pelos Usuários (Web & Mobile)




....
.

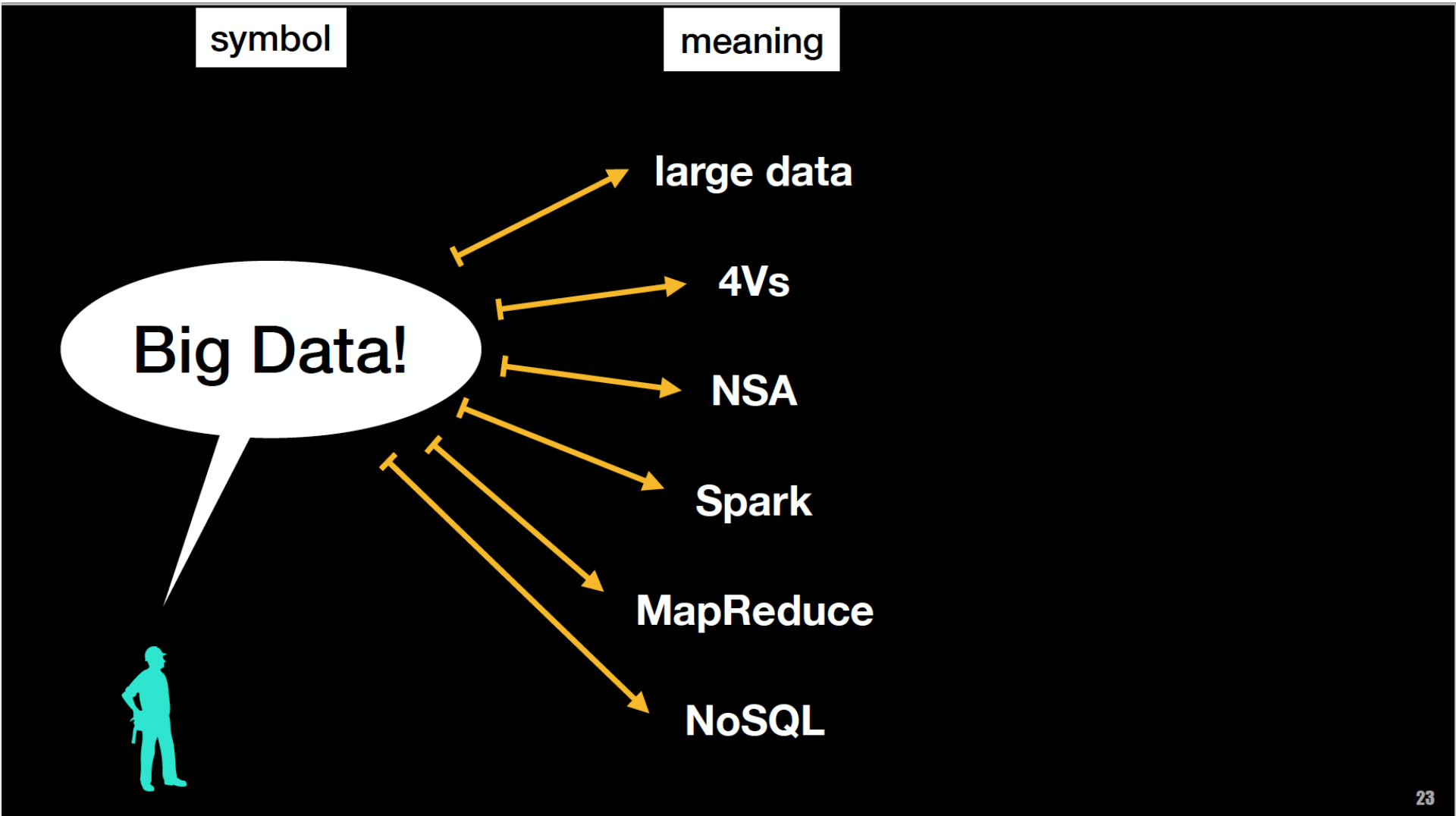

Internet das Coisas (IoT)

Computação Científica/Saúde

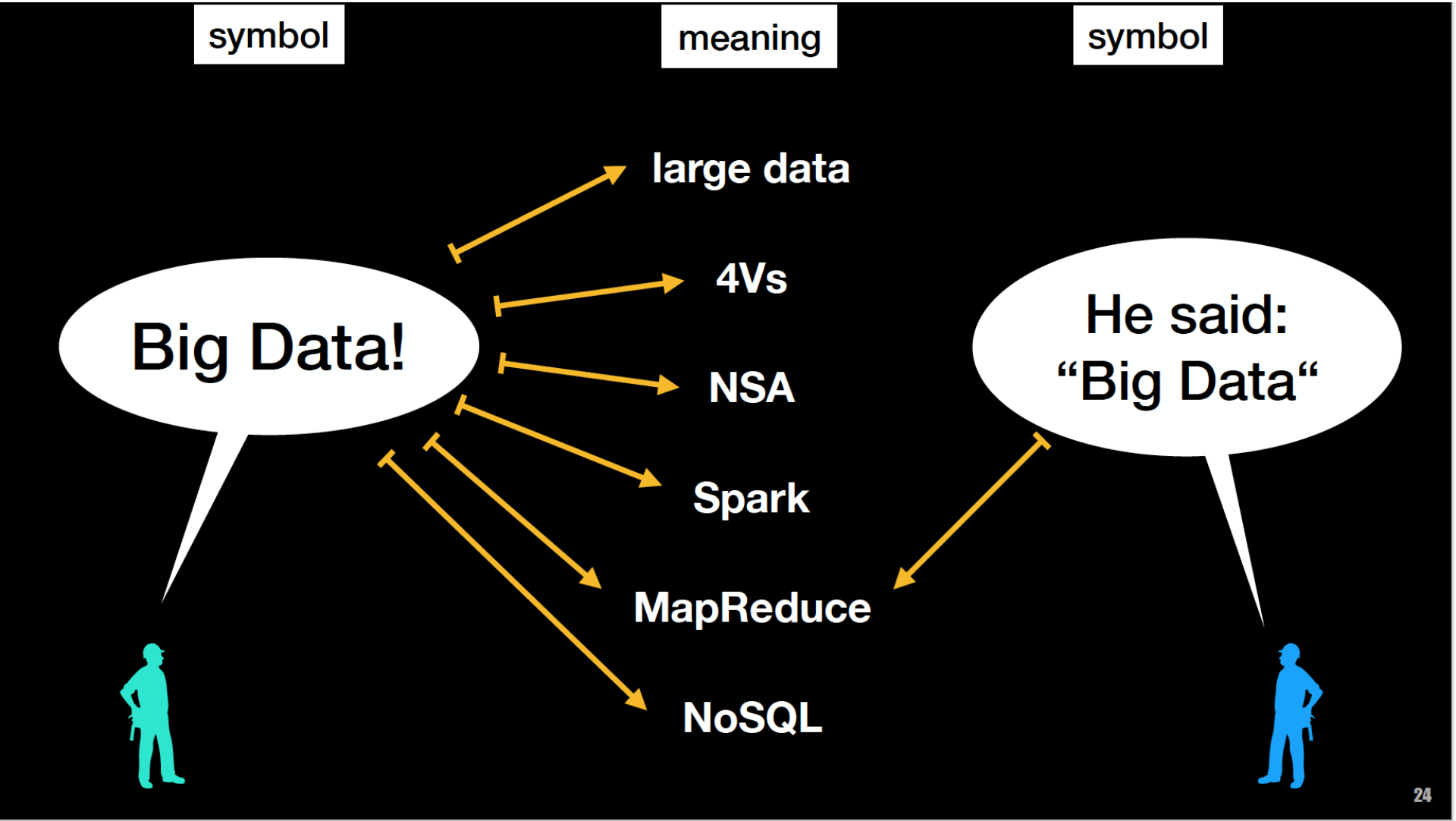


Problem:
ambiguous communication

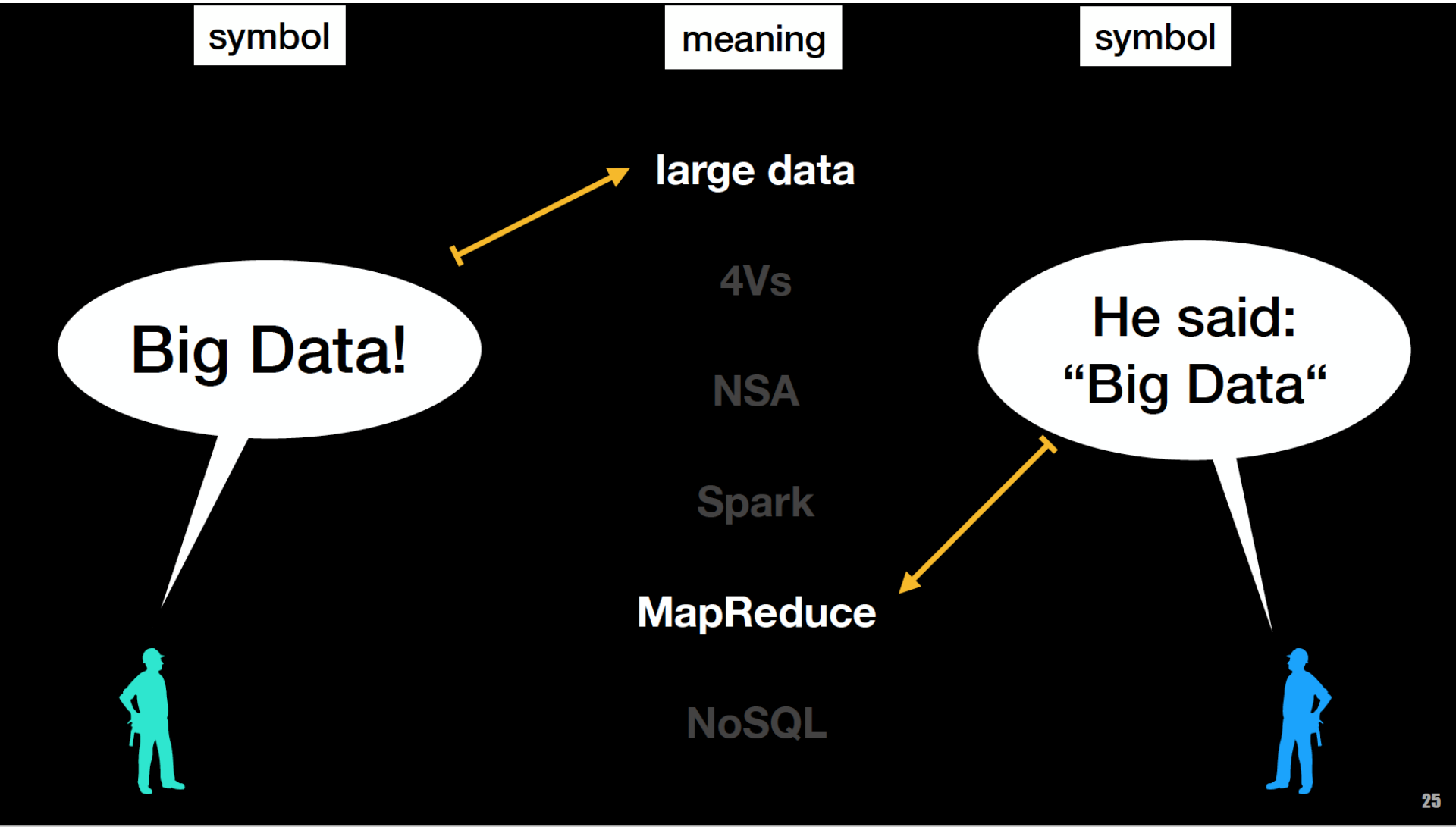
“Big Data”: Cuidado com a Comunicação



“Big Data”: Cuidado com a Comunicação

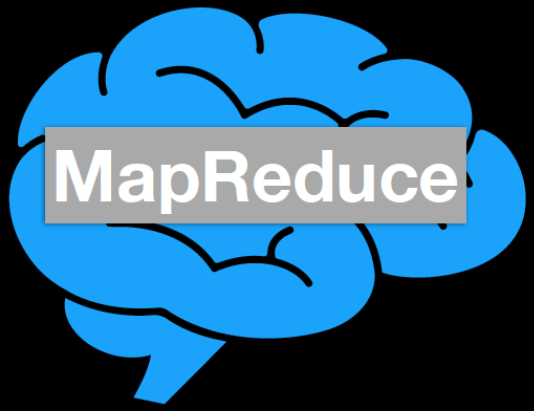


“Big Data”: Cuidado com a Comunicação



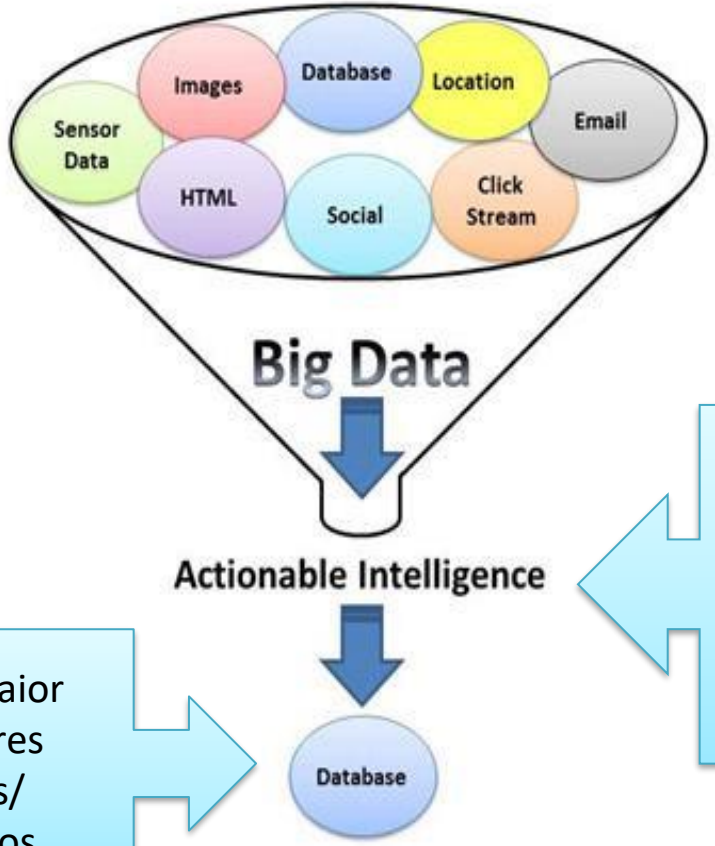
“Big Data”: Cuidado com a Comunicação

translated to:



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