Aprendendo o ABC: Aprendizado de Máquina, Big Data e Ciência de Dados

Prof. Dr. André C. P. L. F. de Carvalho
ICMC-USP

Tópicos

- Introdução
- Big Data
- Ciência de dados
- Aprendizado de Máquina
- Crescimento da área
- Oportunidades na área
- Ciência de dados para o bem

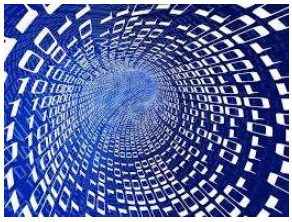
Presente

Inundação de dados

https://inside.igneous.io/storing-files-beyond-big-data

Causas da inundação

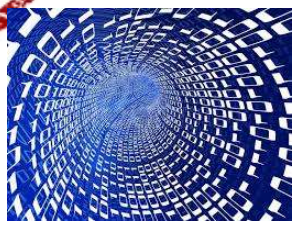
- Avanços recentes nas tecnologias para
  - **Aquisição, armazenamento transmissão** e **processamento** de dados
  - Maior **quantidade**, mais **rápido,** melhor **qualidade** e menor **custo**



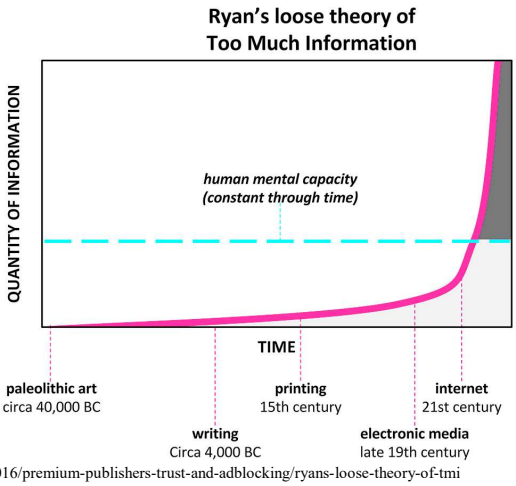
Causas da inundação

- Avanços recentes nas tecnologias para
  - **Aquisição, armazenamento transmissão** e **processamento** de dados
  - Maior **quantidade**, mais **rápido,** melhor **qualidade** e menor **custo**

Big Data



Sobrecarga de informação



https://pagefair.com/blog/2016/premium-publishers-trust-and-adblocking/ryans-loose-theory-of-tmi

Como cresce a aquisição?

Figure 1
Data is growing at a 40 percent compound annual rate, reaching nearly 45 ZB by 2020

Source: Oracle, 2012

44 ZB

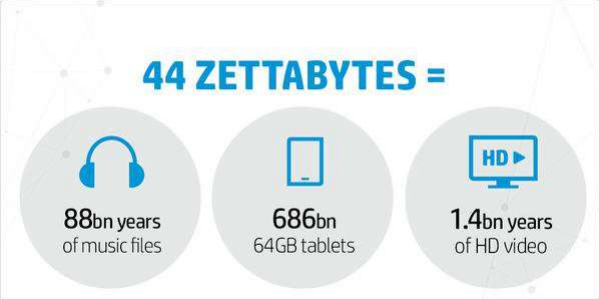
Como cresce a aquisição?

Figure 1
Data is growing at a 40 percent compound annual rate, reaching nearly 45 ZB by 2020

Source: Oracle, 2012

44 ZB

O que cabe em 44 ZB?

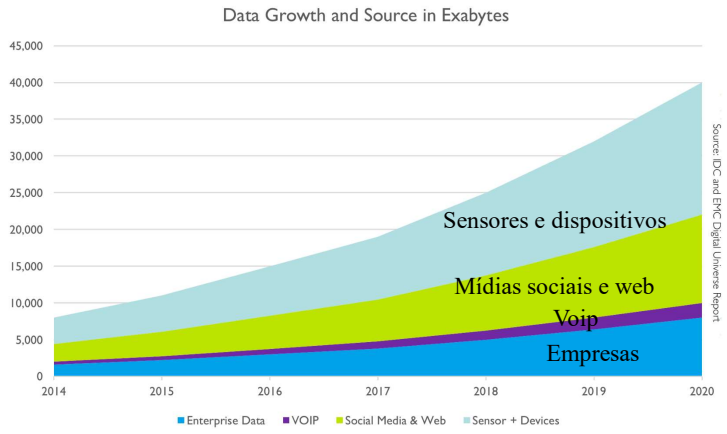


5,3 GB para cada pessoa na terra

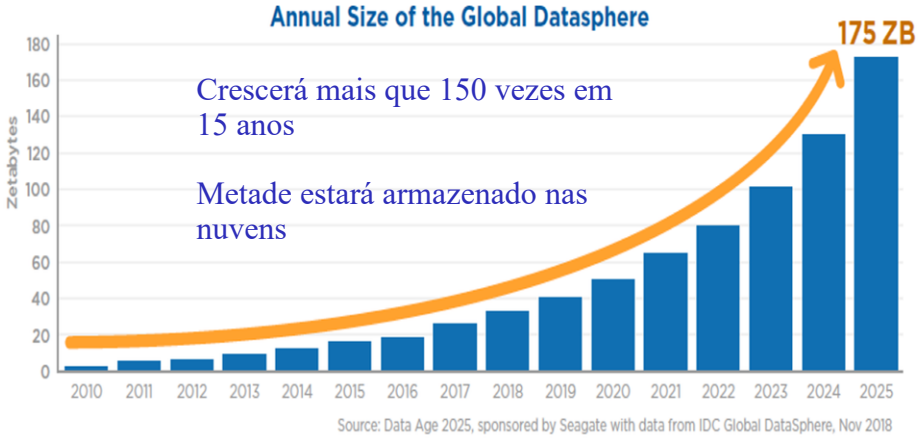
40% desses dados precisam de proteção

https://premioinc.com/blog/storage-overview-glimpse-digital-data-storage-technology

De onde vêm os dados?



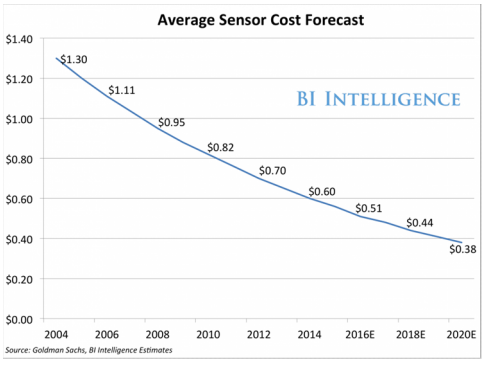
Como segue o crescimento?

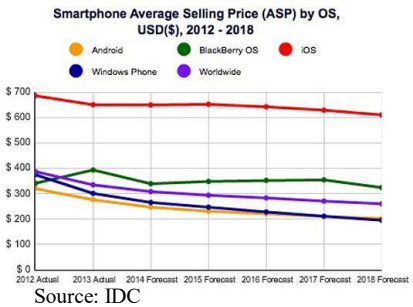


Crescerá mais que 150 vezes em 15 anos

Metade estará armazenado nas nuvens

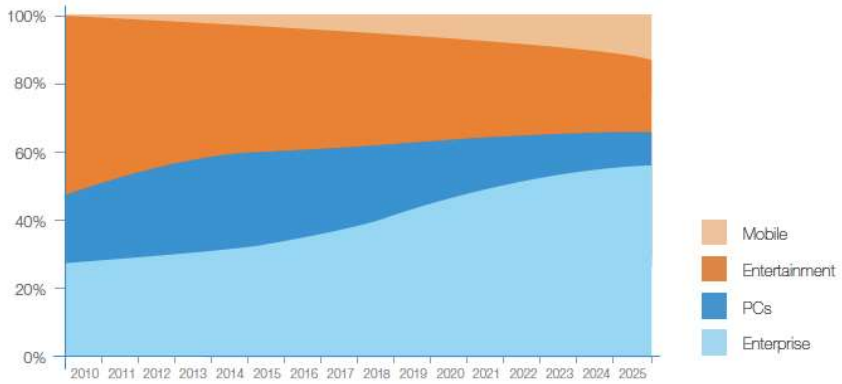
Custo de coleta de dados





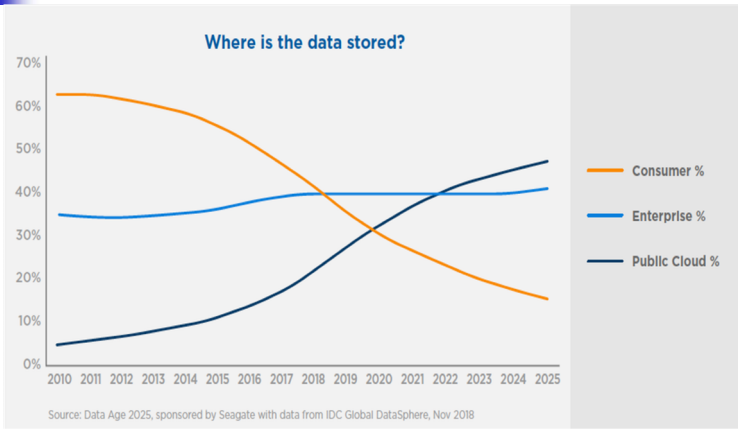
Source: IDC

Onde serão armazenados?



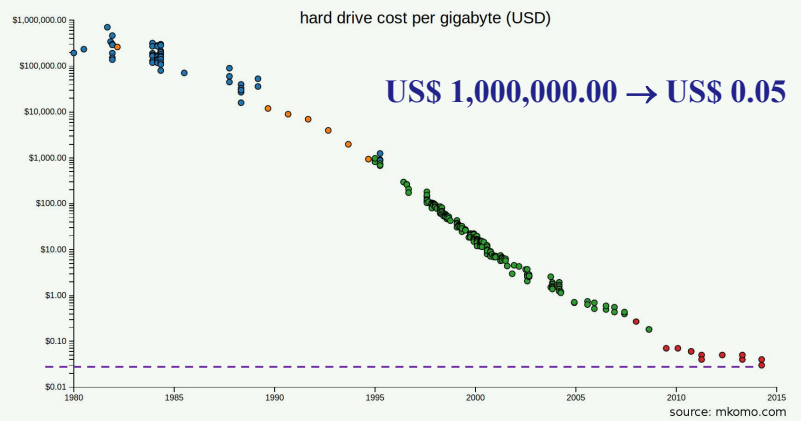
Source: IDC's Data Age 2025 study, sponsored by Seagate, April 2017

Onde serão armazenados?

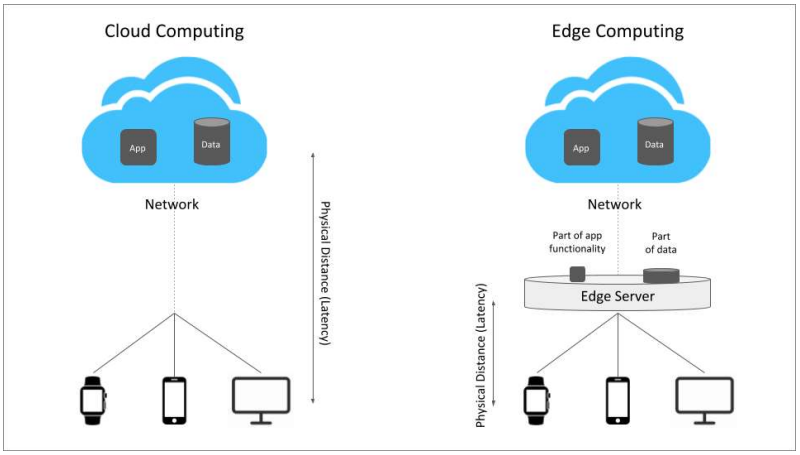


https://www.forbes.com/sites/tomcoughlin/2018/11/27/175-zettabytes-by-2025/#736993dc5459

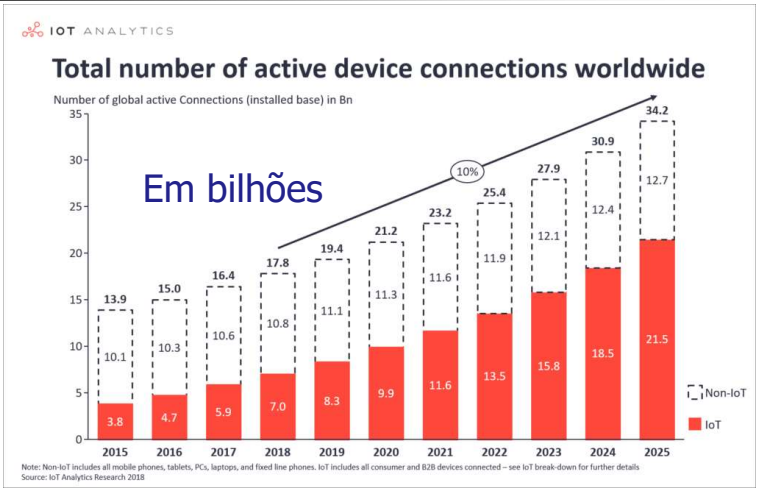
Custo de armazenamento



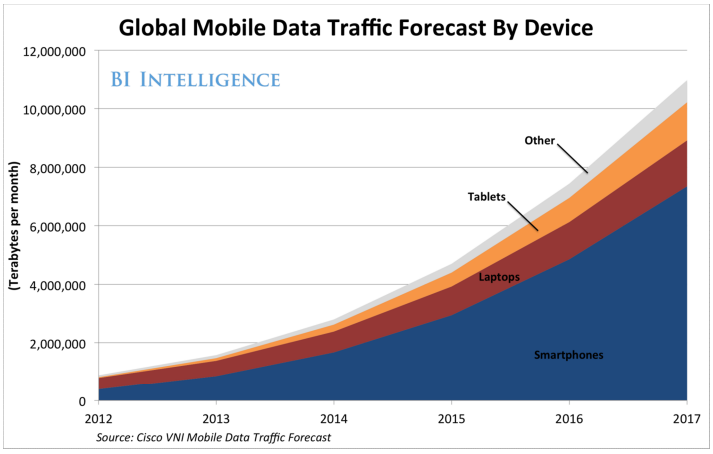
Computação de borda



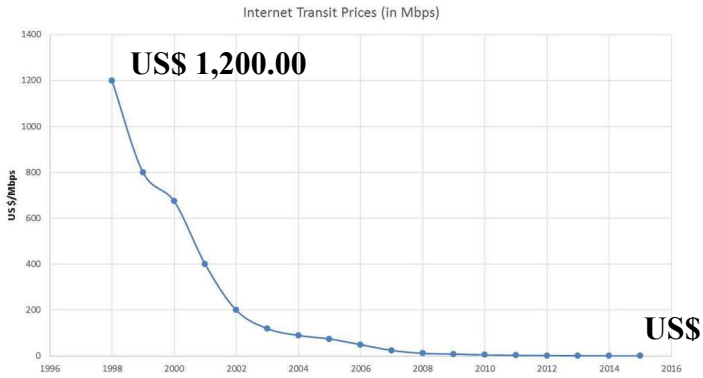
Dispositivos conectados



Como são transmitidos?

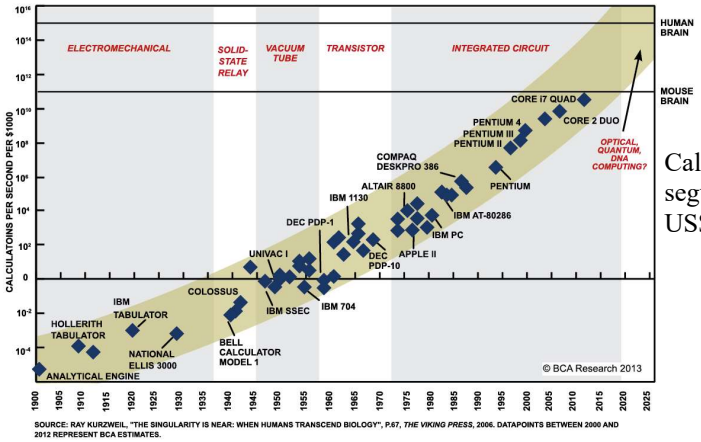


Custo de transmissao



Source: DrPeering.net

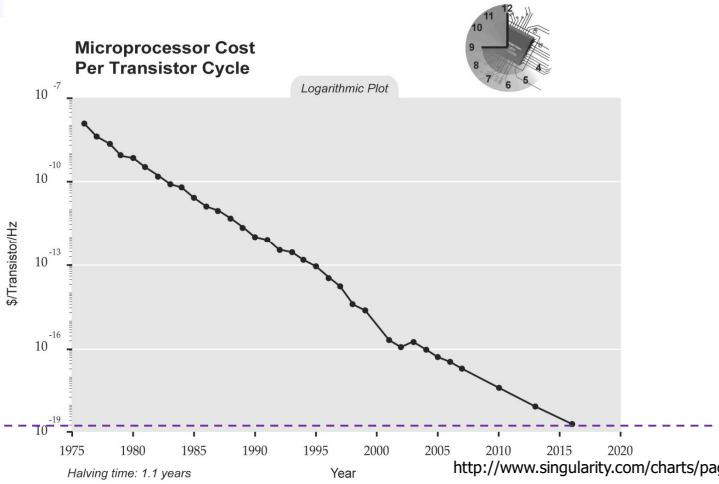
Capacidade de processamento



Calculos por segundo por US$ 1000

SOURCE: RAY KURZWEIL, "THE SINGULARITY IS NEAR: WHEN HUMANS TRANSCEND BIOLOGY", P.67, *THE VIKING PRESS*, 2006. DATAPOINTS BETWEEN 2000 AND 2012 REPRESENT BCA ESTIMATES.

Custo de processamento



http://www.singularity.com/charts/page61.html

O que é Big Data?

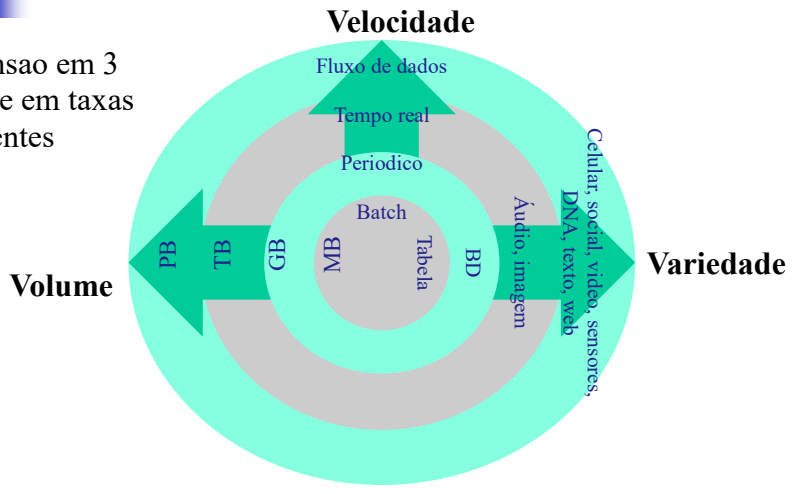
Características de Big Data

- Grande **volume** de dados, gerados a uma grande **velocidade** e com uma grande **variedade** (3 Vs)
  - Volume: tanto de dados estruturados quanto de não estruturados
  - Variedade: vindos de fontes diferentes e com diferentes formatos
  - Velocidade: gerados em fluxos cada vez mais rápidos

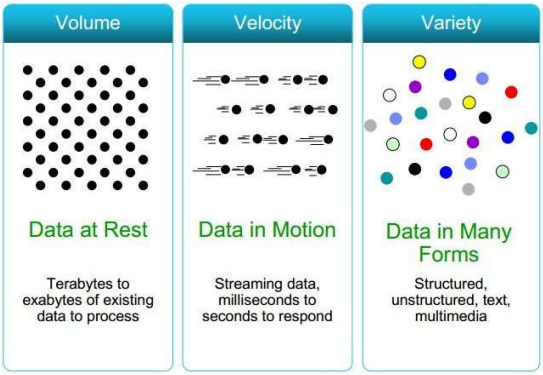
Propostas por Doug Laney, Consultoria Gartner, 2001

Características de Big Data

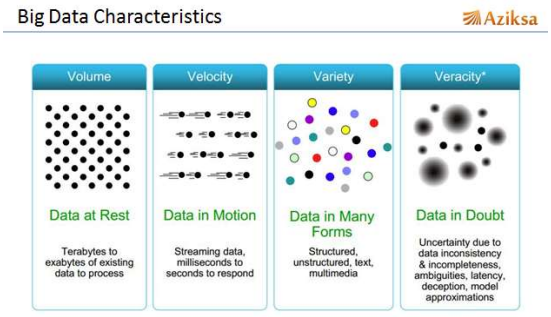
Expansao em 3 eixos e em taxas crescentes



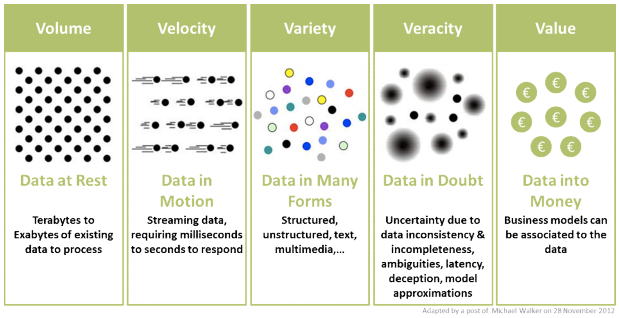
Três Vs



Quatro Vs



Cinco Vs



Vale o que é ouro

Tio Patinhas

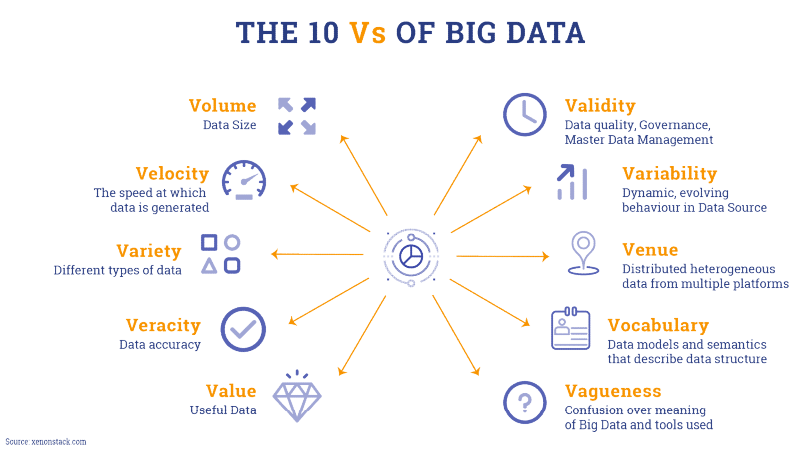
Vale o que é dado

- 1 tonelada de ouro vale US$ 39,289,360.00
- Até hoje, 185.000 toneladas de ouro foram extraídos
  - Tudo somado: US$ 7.4 trilhões (7400 000 000 000)
  - FAAMG + BAT: US$ 5 trilhões (Junho de 2018)
    - Facebook, Amazon, Apple, Microsoft and Google
    - Baidu, Alibaba and Tencent

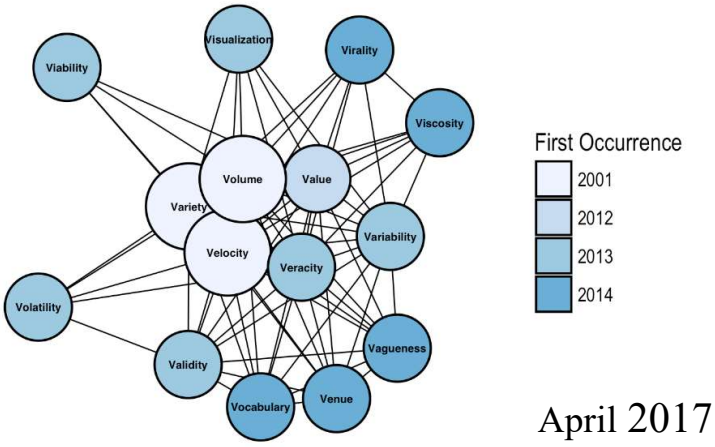
Vale o que é dado

https://www.nytimes.com/es/2017/05/12/amazon-apple-facebook-microsoft-y-google-nos-tienen-atrapados

10 Big Vs (2014)

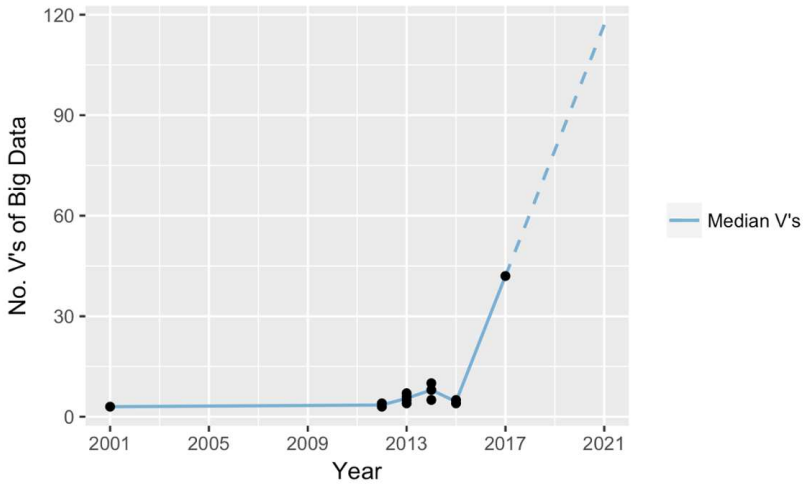


42 Big Vs

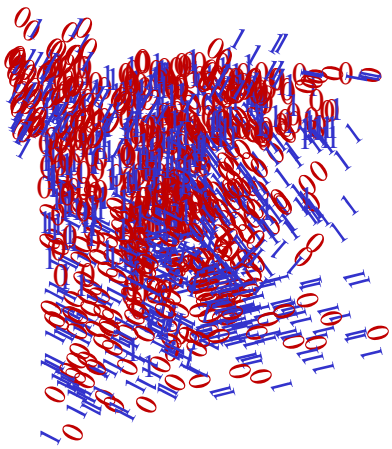


April 2017

Inflação de Vs



Garbage Data



Exploração de Big data

- Dados gerados geralmente contêm informações relevantes
  - Uma vez analisados, podem trazer vários benefícios
    - Sociais e econômicos
  - Crescente interesse na análise de dados

Analise de dados

- Analise de dados por analistas
  - Faltam especialistas
  - Quando tem, o custo é elevado
  - Dificuldade de lidar de forma eficiente com grandes volumes de dados
- Necessário melhores ferramentas de analise
- Vários avanços nas últimas décadas
  - Ciência de Dados

Ciência de Dados

- Estuda princípios e técnicas para extrair conhecimento de um conjunto de dados
  - Novo, relevante e útil
- Pergunta chave da área:
  - Como extrair (de forma eficiente) conhecimentos em (grandes) conjuntos (fluxos) de dados

Big Data x Ciência de Dados

- Frequentemente usados como sinônimos
  - Big Data lida com tecnologias para coletar, gerenciar e processar (Big) dados
  - Ciência de Dados lida com criação de soluções para modelagem de dados
    - Capazes de extrair conhecimento de dados reais

Processar x Descobrir

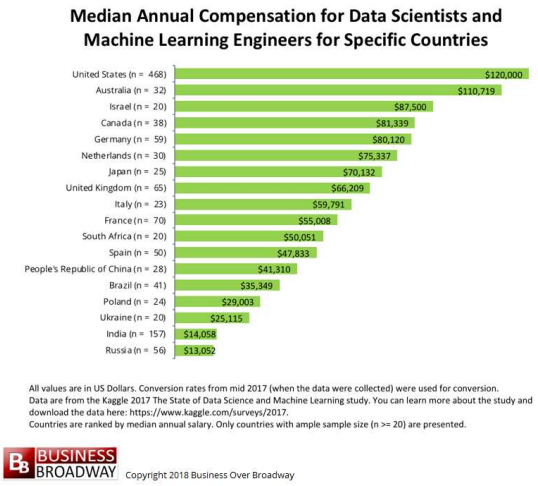
Empregos em Ciência de dados

https://medium.com/indeed-data-science/transitioning-from-academia-to-industry-perspectives-from-indeeds-data-scientists-de890acd1bfc

Quem esta contratando CD

- Amazon
- Apple
- Electronic Arts
- Expedia
- Facebook
- Ford
- Google
- IBM
- Microsoft
- Booking.com
- Disney
- Ford
- Greepeace
- Hasbro
- Mattel
- Mercedes-Benz
- Nespresso
- Red Bull F1

Salários em Ciência de Dados



Ciência de Dados

- Mineração de dados é o conceito mais semelhante a ciência de dados
  - Mas CD é muito mais que isso
- Inclui
  - Compreensão do problema
  - Planejamento de experimentos
  - Pré-processamento
  - Modelagem
  - Avaliação e Validação

Pré-processamento, Modelagem: Mineração de dados

Aprendizado de Máquina (AM)

- Computadores podem aprender a realizar uma tarefa
  - Ao invés de serem explicitamente programados para isso
  - Aprendem a partir de um conjunto de dados a criar modelos (funçõs, hipóteses)
- Dezenas de milhares de algoritmos
  - Centenas de algoritmos novos a cada ano

AM *vs* computação tradicional

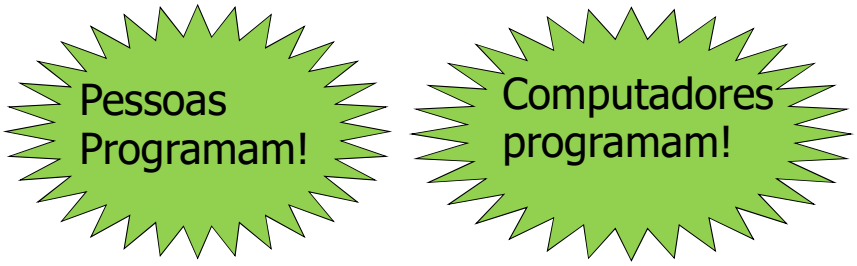
Escreve algoritmo detalhando como resolver um problema

Computação Tradicional

Aprende a resolver um problema observando dados coletados dele

Aprendizado de Máquina

AM *vs* computação tradicional



Computação tradicional

Aprendizado de Máquina

AM *vs* computação tradicional

- Programação
  - Programa que funciona 90% das vezes é ruim
- Aprendizado de máquina
  - Modelo que acerta 90% das vezes pode ser o possível
    - E muito bom
    - Muitas vezes, é mais que suficiente para ser útil

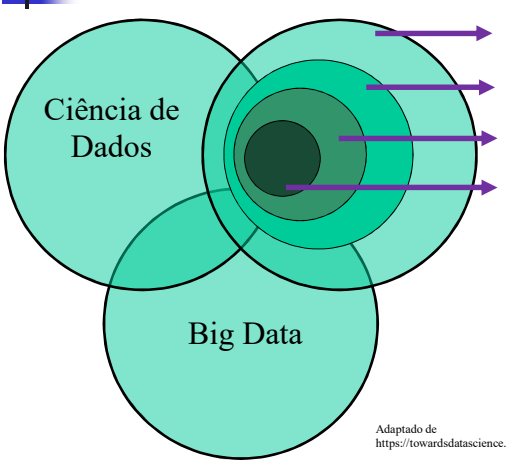
Aprendizado de Máquina

- Investiga algoritmos que possam aprender modelos a partir de dados
  - De forma automática, reduzindo (eliminando) intervenção humana
- Bem sucedido em vários problemas reais de modelagem
  - Descritivos
  - Preditivos

Aplicações de AM

- AM está presente em várias atividades do nosso dia-a-dia
  - Recomendar que mensagens mostrar em aplicativos de redes sociais
  - Filtrar spams de seus *emails*
  - Decidir que resultados (e anúncios), e em que ordem, mostrar após uma busca na internet
  - Diagnóstico médico
  - Predizer aproveitamento de estudantes

Como A, B e C se relacionam?



Inteligência Artificial

Aprendizado de Máquina

Redes Neurais

Redes Neurais profundas/ Aprendizado Profundo

Adaptado de https://towardsdatascience.com/role-of-data-science-in-artificial-intelligence-950efedd2579

AM de ponta-a-ponta

Utilizando Aprendizado de Máquina

Escolher/Modificar algoritmo de AM

Lidar com dados desbalanceados

Engenharia de atributos

Selecionar atributos

Ajustar hiper-parâmetros

Lidar com valores ausentes

Verificar overfitting

Descobrir bugs

Adaptado de Rick Caruana, Research opportunities in AutoML Microsoft Research

AM de ponta-a-ponta

Aprendizado de Máquina na Prática

Escolher/Modificar algoritmo de AM

Lidar com dados desbalanceados

Engenharia de atributos

Selecionar atributos

Ajustar hiper-parâmetros

Lidar com valores ausentes

Verificar overfitting

Descobrir bugs

Adaptado de Rick Caruana, Research opportunities in AutoML Microsoft Research

Forbes Billionaires Innovation Leadership Money Consumer Industry

19,061 views | Apr 15, 2018, 02:05am

Why AutoML Is Set To Become The Future Of Artificial Intelligence

3) Say "Hello" To AutoML

One of the biggest trend that will dominate the AI industry in 2019 v automated machine learning (**AutoML**). With automated learning capabilities, developers will be able to tinker with machine learning and create new machine learning models that are ready to handle future AI

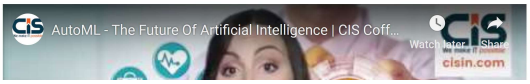
ZDNet

MUST READ: These hackers broke into 10 telecoms companies to steal customers' phone records

AutoML is democratizing and improving AI

Once a niche technology, Automated Machine learning (AutoML) is now a thing. Helping non-data scientists do simple AI, and helping trained data scientists do complex work ever-faster, AutoML technology is catching on, and may well put AI in the Enterprise fast lane.

AutoML - A Short Overview: Why AutoML Is Ready To Be The Future Of Artificial Intelligence



HEARTBEAT MOBILE MACHINE LEARNING NEWSLETTER COMMUNI

The Rise of Automated Machine Learning

No matter what industry you're in, autoML can help you use machine learning successfully and and leverage business insights hidden in places where only machine learning can reach.

By Abhi Yadav

January 15, 2019

Freedom. Wisdom. Excellence.

PRODUCTS ARTICLES SERIES CATEGORIES ABOUT

AutoML: The Next Wave of Machine Learning

Parul Pandey Follow

Apr 18 · 9 min read

AutoML and the Rise of Advanced Machine Learning Platforms

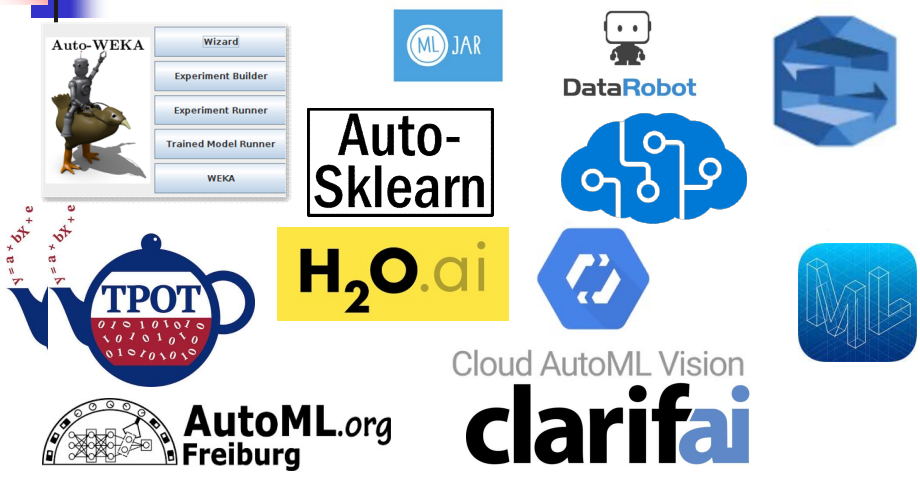
FEBRUARY 7, 2019 — 1 COMMENT

Analytics

Love it or Hate It: Auto ML is Here to Stay

By David Sweenor - March 28, 2019

Ferramentas de AutoML



CD Responsável

- Reprodutibilidade
  - Disponibilização e curadoria de dados e códigos
- Privacidade
  - Aplicando AM a 10 (300) likes, é possível conhece melhor sua personalidade que colega de trabalho (cônjuge)
  - *Fair Information Practices* (FIPs)
- Responsabilidade (*accountability*)
  - Alguém responde pela ferramenta de CD

CD Responsável

- Justiça
  - Tomada de decisão não deve embutir preconceito
- Transparência
  - General Data Protection Regulation (GDPR-UE)
    - Direito a informação
  - Brasil tem legislação semelhante

Ciência de Dados para o Bem

- Movimento sem fins lucrativos
  - Levar benefícios sociais para as pessoas e comunidades
  - Alguns programas são adotados por empresas
- Como ele ocorre?
  - Reuniões
  - Eventos
  - Estágios acadêmicos
  - Redes sociais

Ciência de Dados no ICMC USP

- Encaixa perfeito com perfil do ICMC
- Pesquisa
  - Um dos 3 pilares do CEPID CeMEAI, sediado no ICMC
  - Centro de Excelência da Intel em Inteligência Artificial
- Extensão
  - Escolas de Ciência de Dados e de Big Data
  - MBA em Ciência de Dados

Ciência de Dados no ICMC USP

- Graduação
  - Ênfase de graduação em Ciência de Dados
  - Curso de graduação em Ciência de Dados (trâmites legais iniciados)
- Pós-graduação
  - Mestrado Profissional

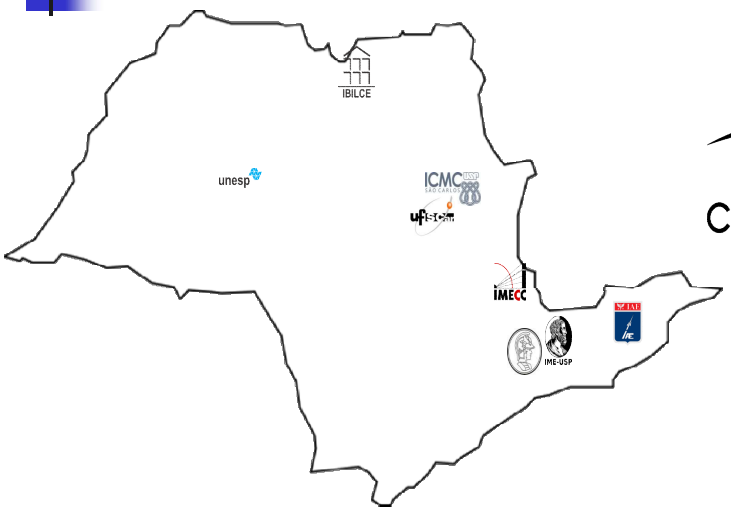
Ciência de Dados no ICMC USP

- Mais de 20 docentes trabalhando em
  - Big Data
  - Ciência de Dados
  - Aprendizado de Máquina
- Laboratórios
  - Analytics, BioCom, LABIC, Estatística
- Formação de recursos humanos
  - Graduação e Pós-graduação

CEPID CeMEAI

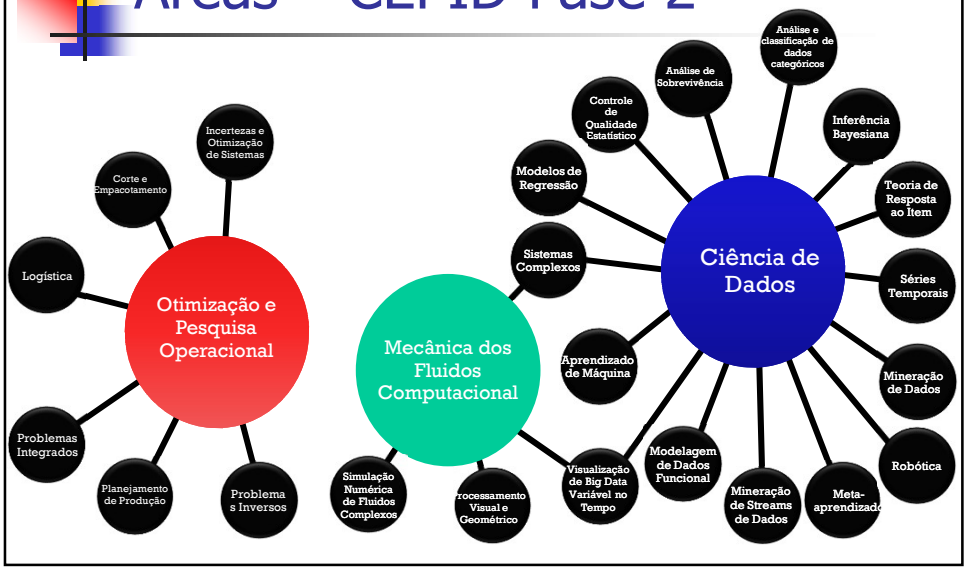
- Centro de Ciências Matemáticas Aplicadas à Indústria
  - ICMC-USP
- Inicio em 2013, duração de até 11 anos
  - Orçamento para primeiros 5 anos de 15 milhões de Reais da FAPESP
    - Mesma quantidade obtida de empresas
- Objetivo principal
  - Transferir conhecimento em computação, matemática e estatística para empresas e órgãos públicos

Instituições Parceiras



CeMEAI

Áreas – CEPID Fase 2



Workshops com indústrias

2015
7 problemas de 6 empresas
139 participantes
Resultados
- 1 proposta de PIPE
- 1 proposta proposta de cooperação
- 2/3 problemas completamente resolvidos

2016
6 problemas de 5 empresas
137 participantes
Resultados
- 1 problema completamente resolvido
- 1 problema com 4 soluções possíveis de patenteamento
- 4 problemas com grande avanço para a solução

2017
6 problemas de 6 empresas
132 participantes
Resultados
- 2 problemas completamente resolvidos
- 2 com grandes avanços

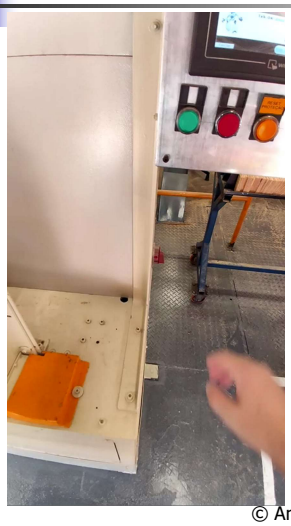
Mestrado Profissional

6ª turma em Ciência de Dados
Alunos de empresas:

MECAI

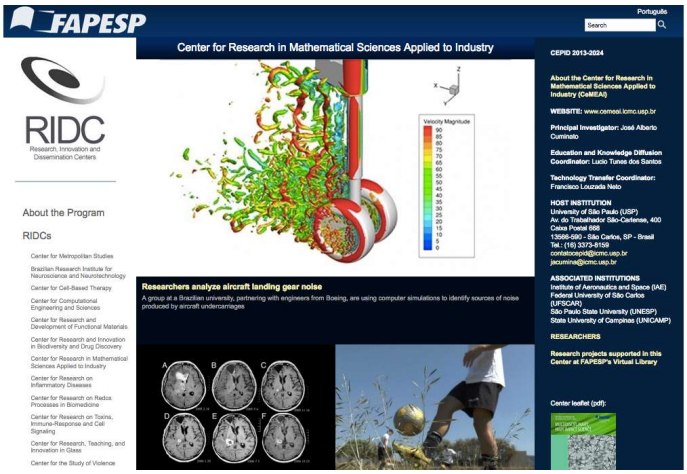
amdocs
SIEMENS
Virgos
BSI tecnologia
SAP
TAM
nutresa
Disfrutar la vida te alimenta
QW3
CREDIT SUISSE
hp
CUTRALE
DELTA
sistemas de informação
Embrapa
central açucareira santo antonio s.a.
Animall TAG
Bradesco
Itaú

Classificação de madeiras

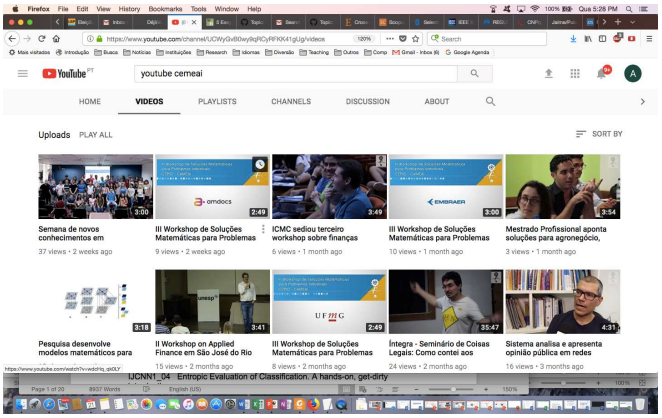


- Protótipo para Faber Castell
- 100 tábuas
  - 50% boas and 50% ruins
    - 14 tipos de defeitos
  - 150 placas por minute (27,7 metros por minute)
  - Acurácia preditiva: 97%
    - Operador humano: 93%

CEPID CeMEAI



Canal YouTube CeMEAI



Conclusão

- Big data
- Ciência de dados
- Aprendizado de Máquina
- A, B e C no ICMC

Perguntas