CIÊNCIA DE DADOS

Profa. Roseli A. F. Romero

SCC-ICMC-USP

http://www.digitalzenway.com/2011/12/data-diet-a-resolution-you-can-stick-to/

CONTEÚDO

- INTRODUÇÃO
- **PARTE I** – EXPLORAÇÃO
  EXEMPLO
- **PARTE II** – PRE-PROCESSAMENTO
  EXEMPLOS
- **PARTE III** – MODELAMENTO E ANÁLISE DE EXPERIMENTOS
  EXEMPLO

Cientistas de Dados:
O que fazem?

- Cientistas de dados são os grandes mineradores de dados. Eles recebem uma enorme massa de dados desorganizados (estruturados e não estruturados) e usam suas habilidades em matemática, estatística e programação para limpar, tratar e organizá-los.
- Em seguida, eles aplicam suas capacidades analíticas – conhecimento da indústria, compreensão contextual, ceticismo de suposições existentes – para descobrir soluções para os desafios de negócios ocultos.

Cientistas de Dados:
O que fazem?

- Entre suas principais responsabilidades estão:

1 - Realizar pesquisas sem direção e formular perguntas abertas aos dados

2 - Extrair grandes volumes de dados de múltiplas fontes internas e externas

3 - Empregar os programas de análise sofisticadas, aprendizado de máquina e métodos estatísticos para preparar os dados para uso em modelagem preditiva e descritiva.

DADOS

Estruturados

Mais facilmente analisados por técnicas de MD
Ex.: Planilhas e tabelas atributo-valor

Não estruturados

Ex.: Conteúdo de página na web, emails, vídeos, sequencia de DNA, ...

Tipos de atributos

- Simbólicos ou qualitativos
  - Nominal ou categórico
    - Ex.: cor, código de identificação, profissão
  - Ordinal
    - Ex.: gosto (ruim, médio, bom), dias da semana
- Numéricos, contínuos ou quantitativos
  - Intervalar
    - Ex.: data, temperatura em Celsius
  - Racional
    - Ex.: peso, tamanho, idade

1) PONTO INICIAL: Escolha um Problema

- Escolha algo que o entusiasme, como um projeto de análise musical do Spotify
- Projeto de análise de aluguel em uma cidade

2) Pense nos diferentes passos

Coleta de Dados

Análise de Dados

Visualização:

Implantação

Kaggle
Web Scraping
API

Os dados podem estar bagunçados

Existem diferentes ferramentas

Desenvolvimento WEB
Interface

Análise de Dados

- EXTRAIR CONHECIMENTO DOS DADOS
- REALIZAR A INTERPRETAÇÃO
- TOMAR AÇÕES

PARTE I

Exploração dos Dados

- Gerar Hipóteses
- Entendimento por meio de técnicas
- Reavaliar as Hipóteses
- Vantagens e desvantagens de técnicas
- Sumarizar as informações

1º.ETAPA

EXPLORAÇÃO

- Medidas de Localização
  - Média
  - Mediana
  - Percentil
- Momentos
  - Variancia
  - skewness
  - Kurtosis
- Visualização
  - Histograma
  - Box Plot

GRANDES VOLUMES DE DADOS

- BIG DATA
- KDD – Knowledge Discovery in Databases

Os dados nunca dormem

- Nossas vidas são cercadas e preenchidas

  por dados de todos os tipos
- Nós vivemos num mundo repleto de dados e o montante

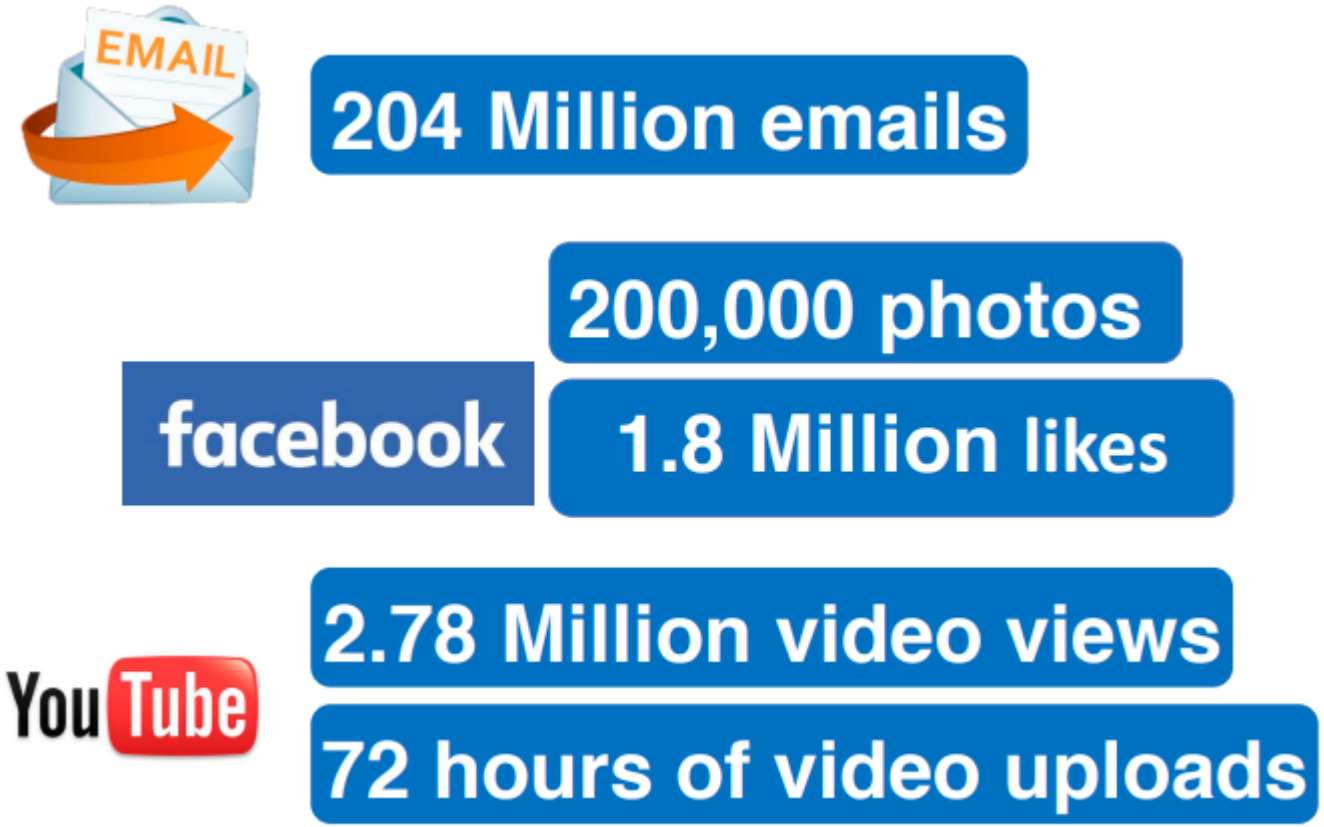
armazenado diariamente é assustador

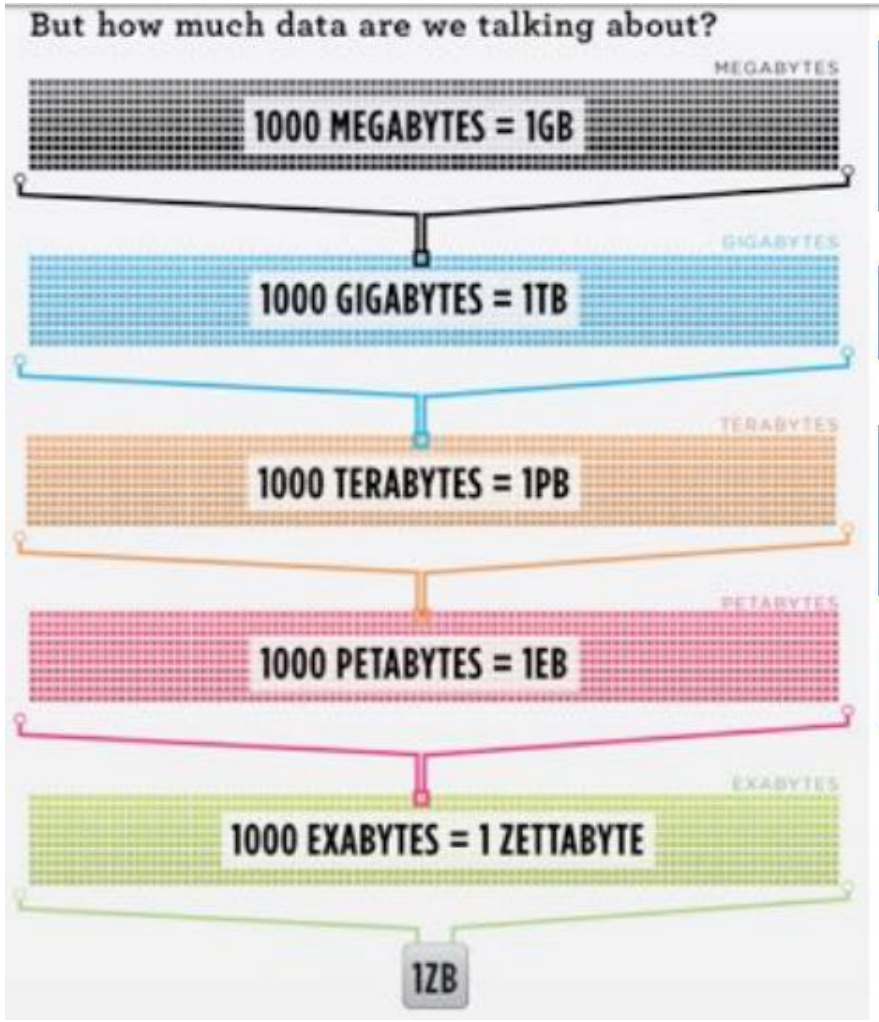
- Nós podemos desligar nossos dispositivos para descansar ou desligar do mundo dos dados, mas os DADOS NUNCA DORMEM.

Internet por minuto

Image Source: http://www.marketwatch.com/story/one-chartshows-everything-that-happens-on-the-internet-in-just-oneminute-2016-04-26

Por Minuto


EMAIL
204 Million emails
200,000 photos
facebook
1.8 Million likes
YouTube
2.78 Million video views
72 hours of video uploads



100 MBs ~= couple of volumes of Encyclopedias

A DVD ~= 5 GBs

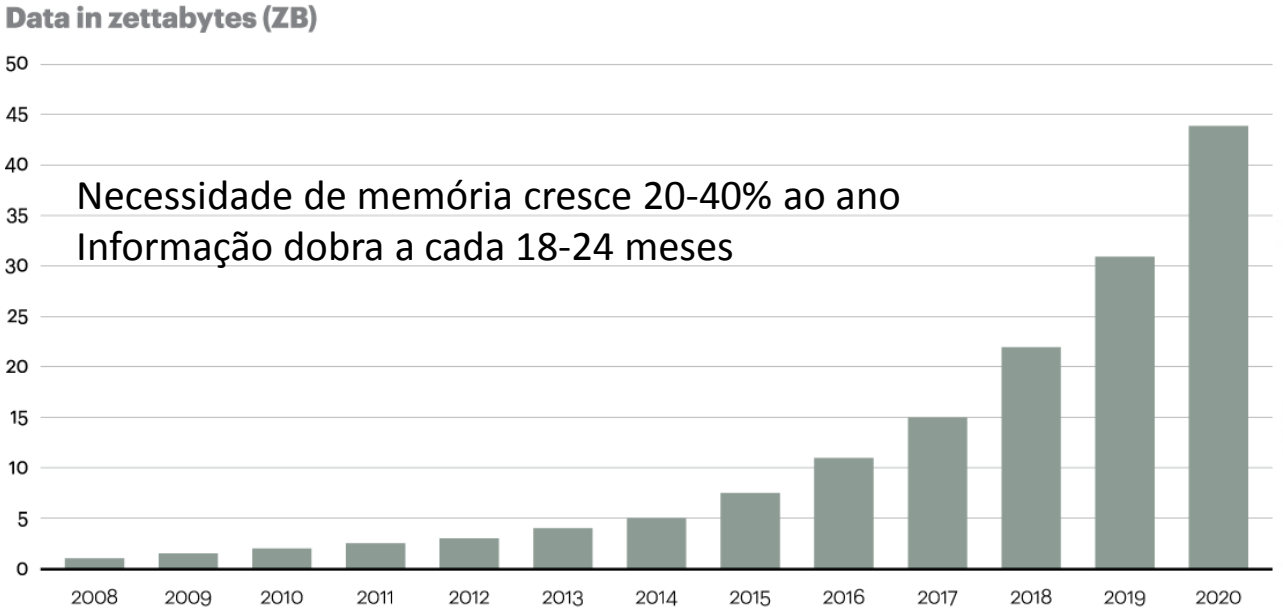
1 TB ~= 300 hours of good quality video

LHC ~= 15 PBs a year

Como cresce a quantidade de dados?

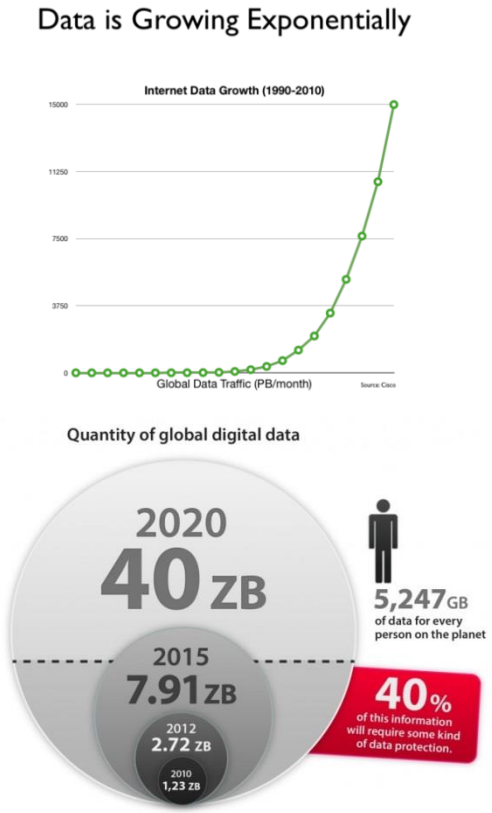
Figure 1

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



Source: Oracle, 2012

Necessidade de memória cresce 20-40% ao ano
Informação dobra a cada 18-24 meses



5,247 GB por pessoa
40% requer proteção

De onde vêm os dados?

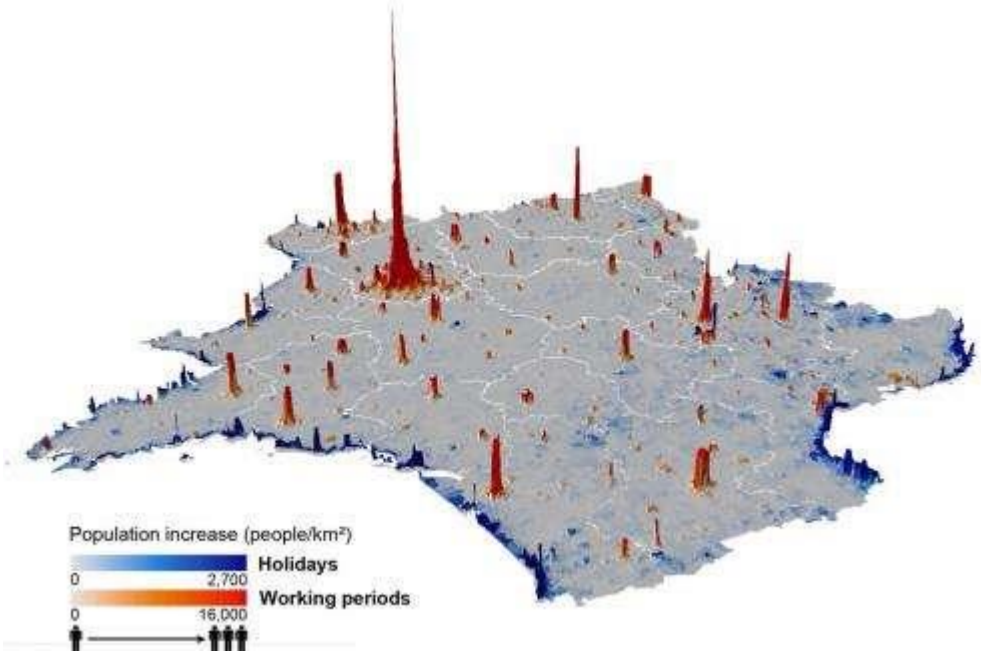
- Dispositivos eletrônicos
  - Sinais de localização de smartphones
  - Logs de servidores de aplicações
    - Jogos e web sites
  - Sensores de dados
    - Climáticos, reservatórios de água, corpo humano
  - Imagens e vídeos
    - Câmeras de monitoramento de trânsito, de segurança

De onde vêm os dados?

- Atividades realizadas por seres humanos
  - *Blogs*
  - *Emails*
  - Formulários
  - Navegações e buscas
  - Redes sociais
    - Compartilhamento de músicas, fotos, vídeos, envio de informações, troca de mensagens curtas...

Dados de smartphones

França



Population dynamics between the main holiday period (July and August) and working periods in France.
Credit: Catherine Linard

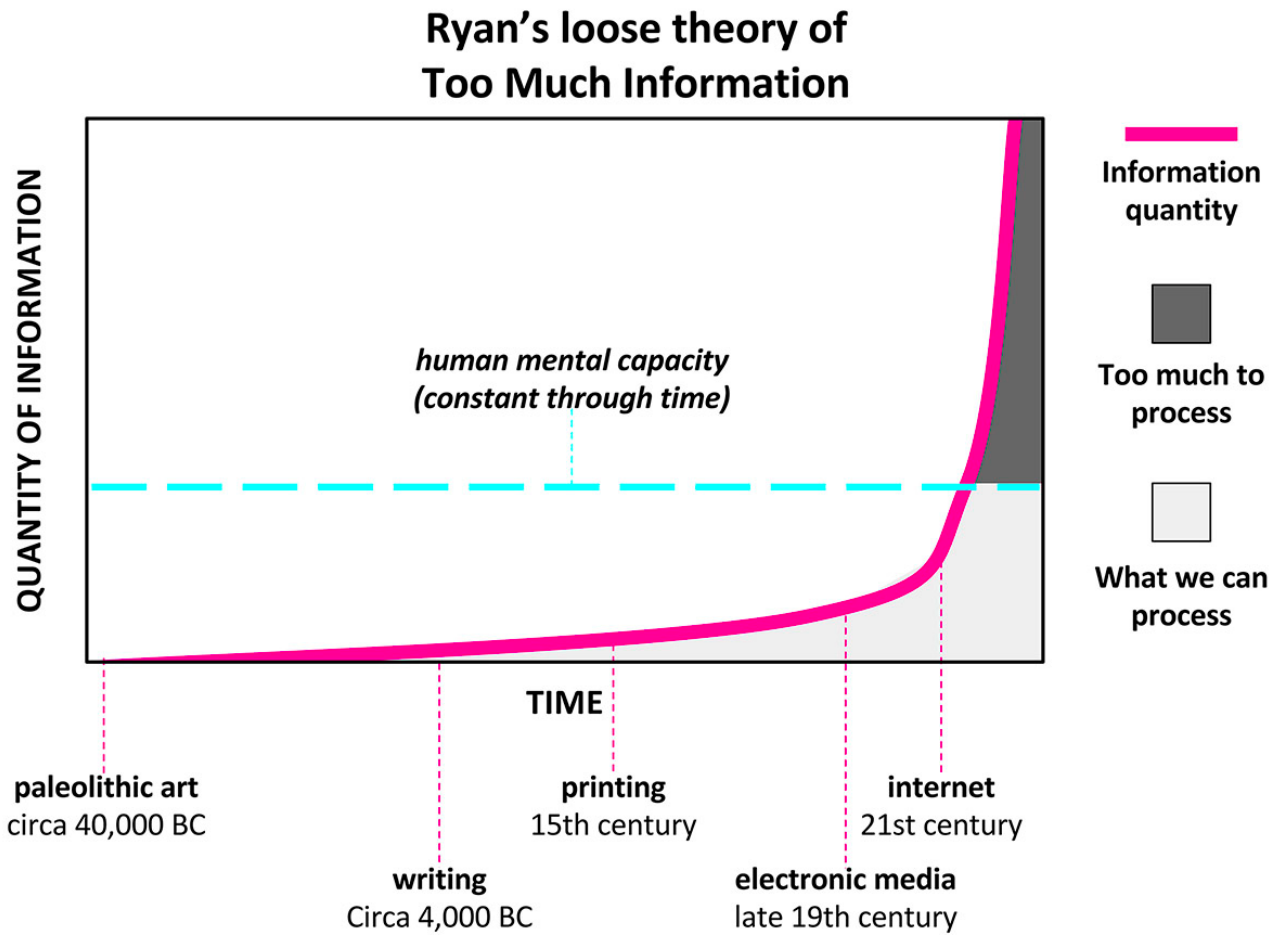
http://phys.org/news/2014-10-cellphone-population-density.html#jCp

Cada vez mais dados (Volume)
e cada vez mais complexos (Variedade)

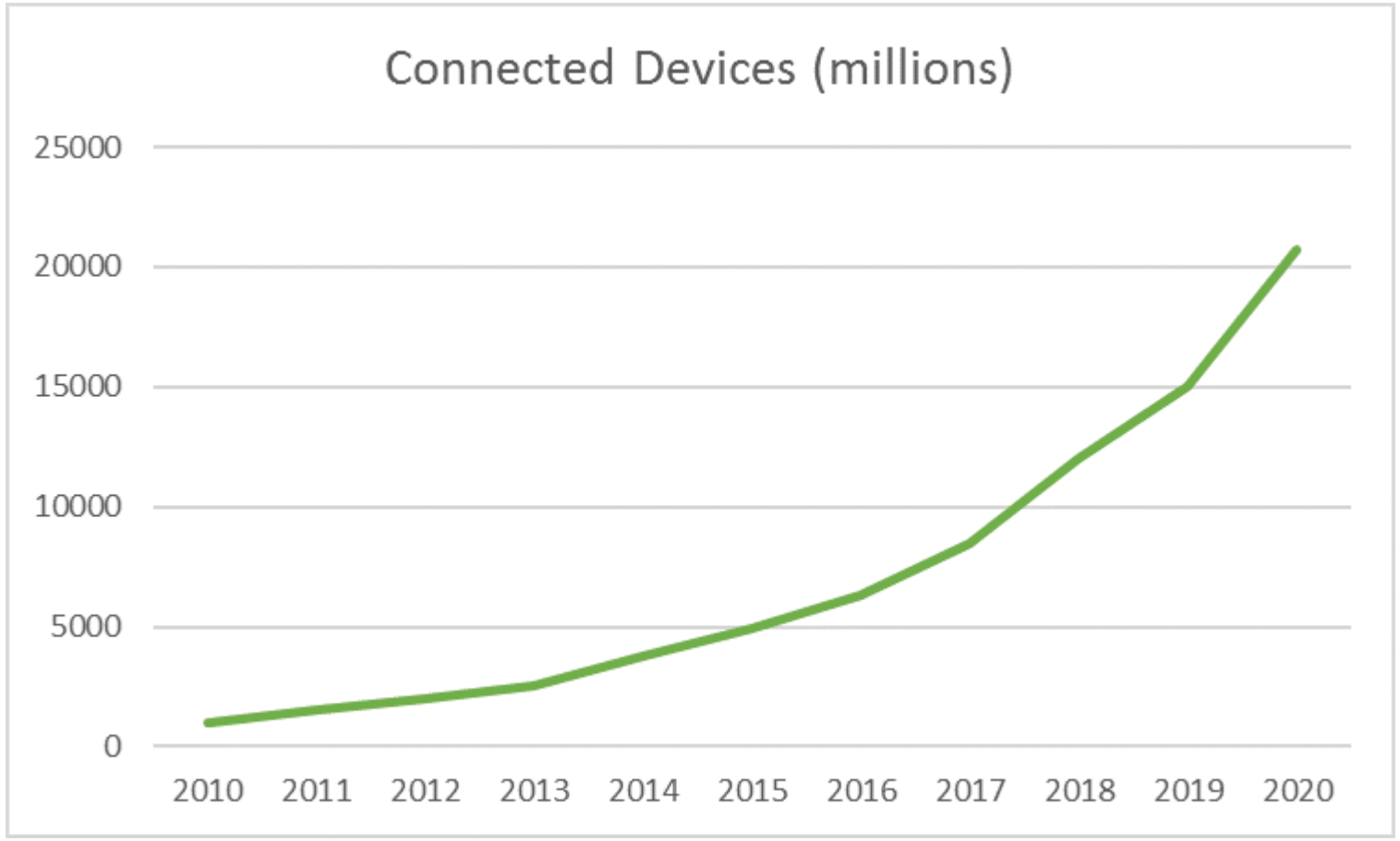
Avanços recentes nas tecnologias para **aquisição, transmissão**, **armazenamento** e **processamento** de dados

Big Data

Sobrecarga de informação

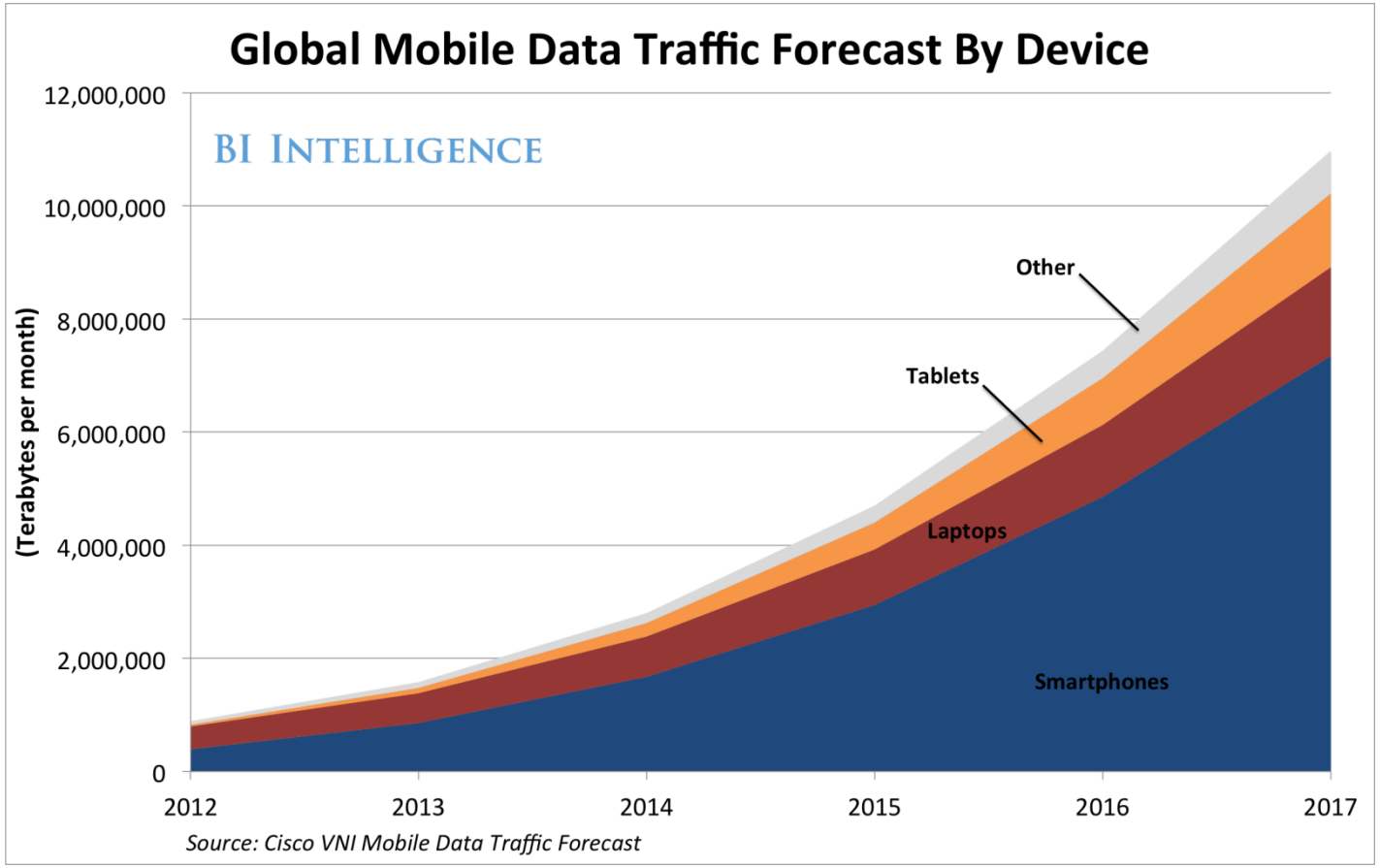


Como avança a transmissão de dados? (Velocidade)

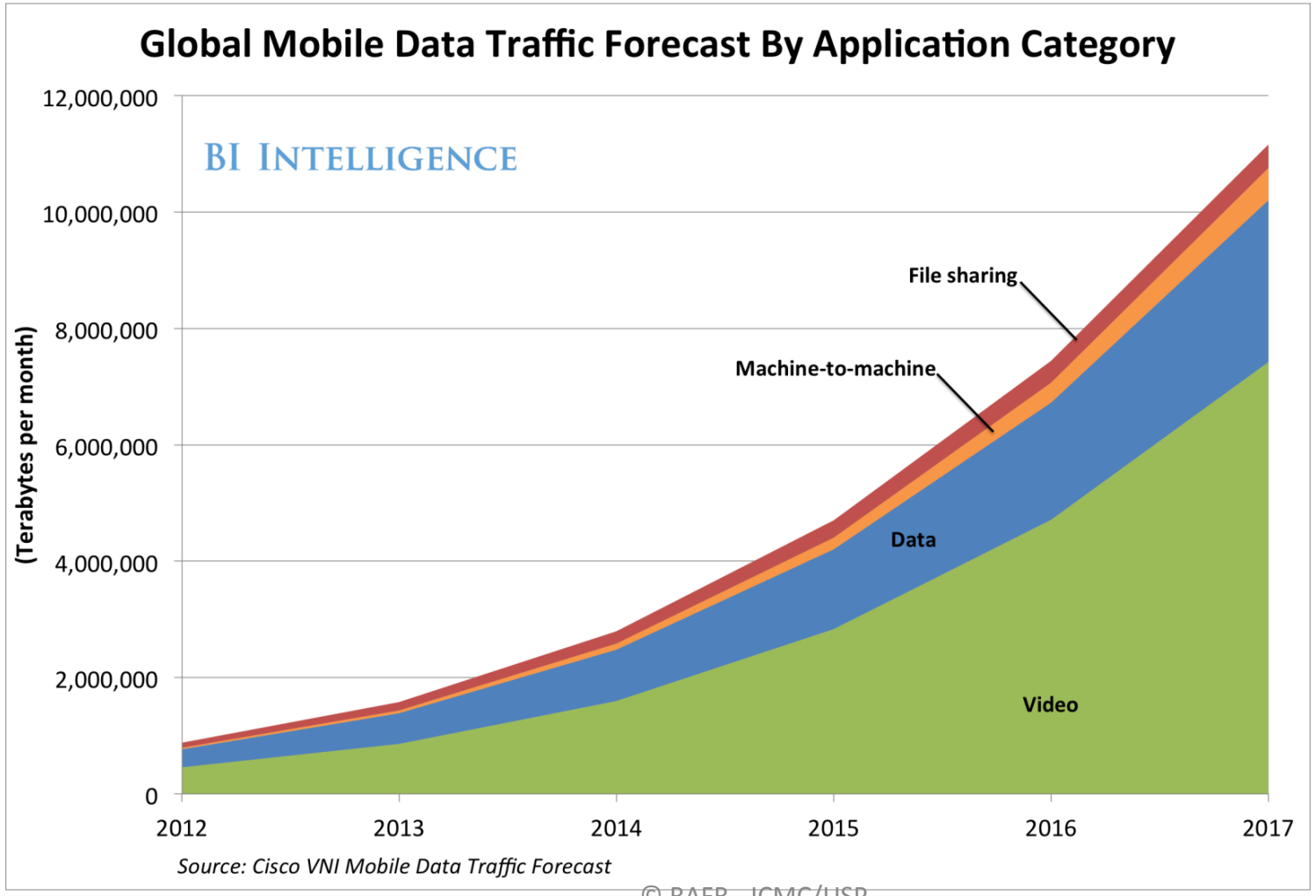


http://motherboard.vice.com/blog/the-next-five-years-of-explosive-internet-growth-in-seven-graphs

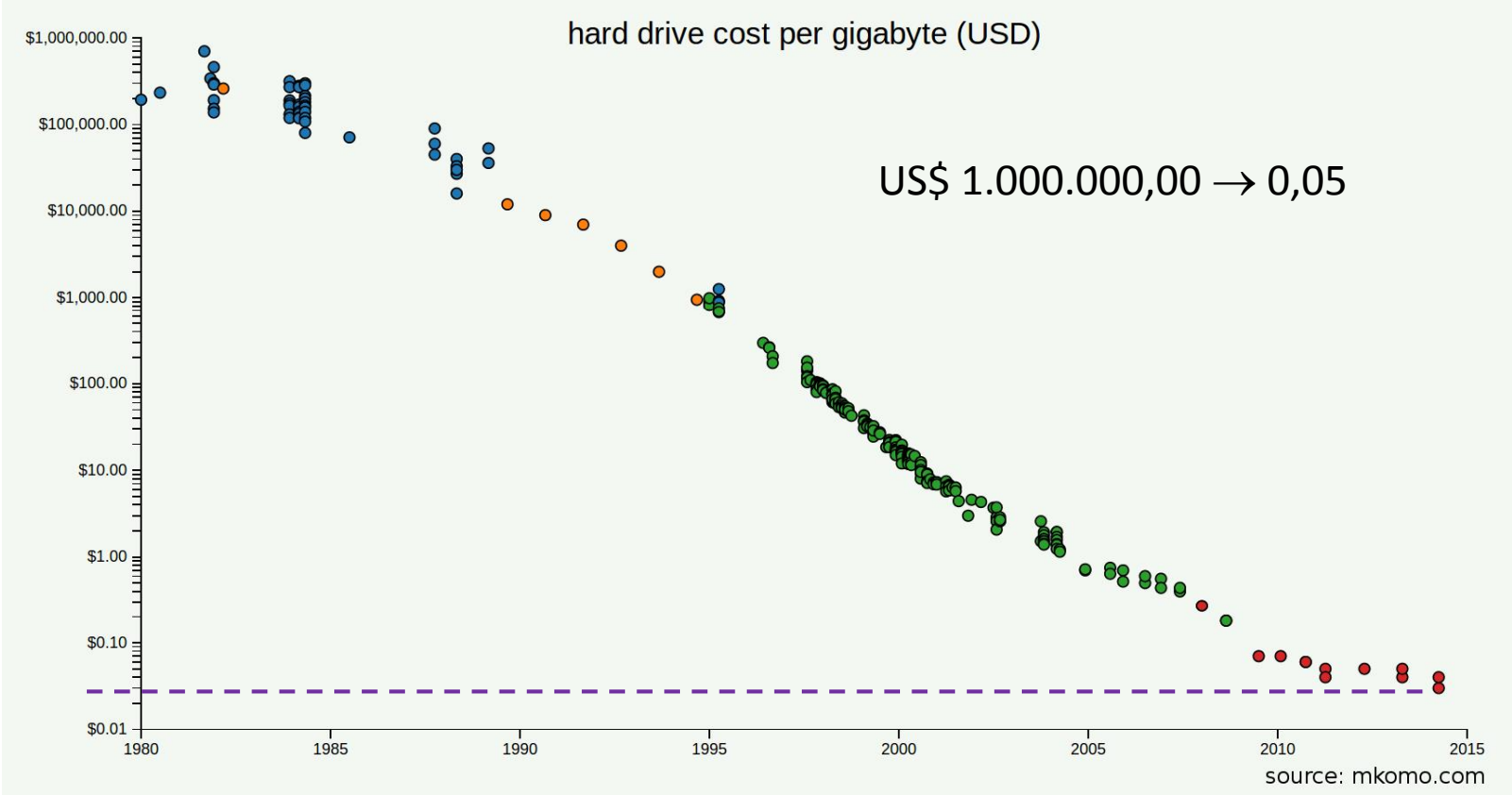
De que tipo de dispositivos os dados chegam?



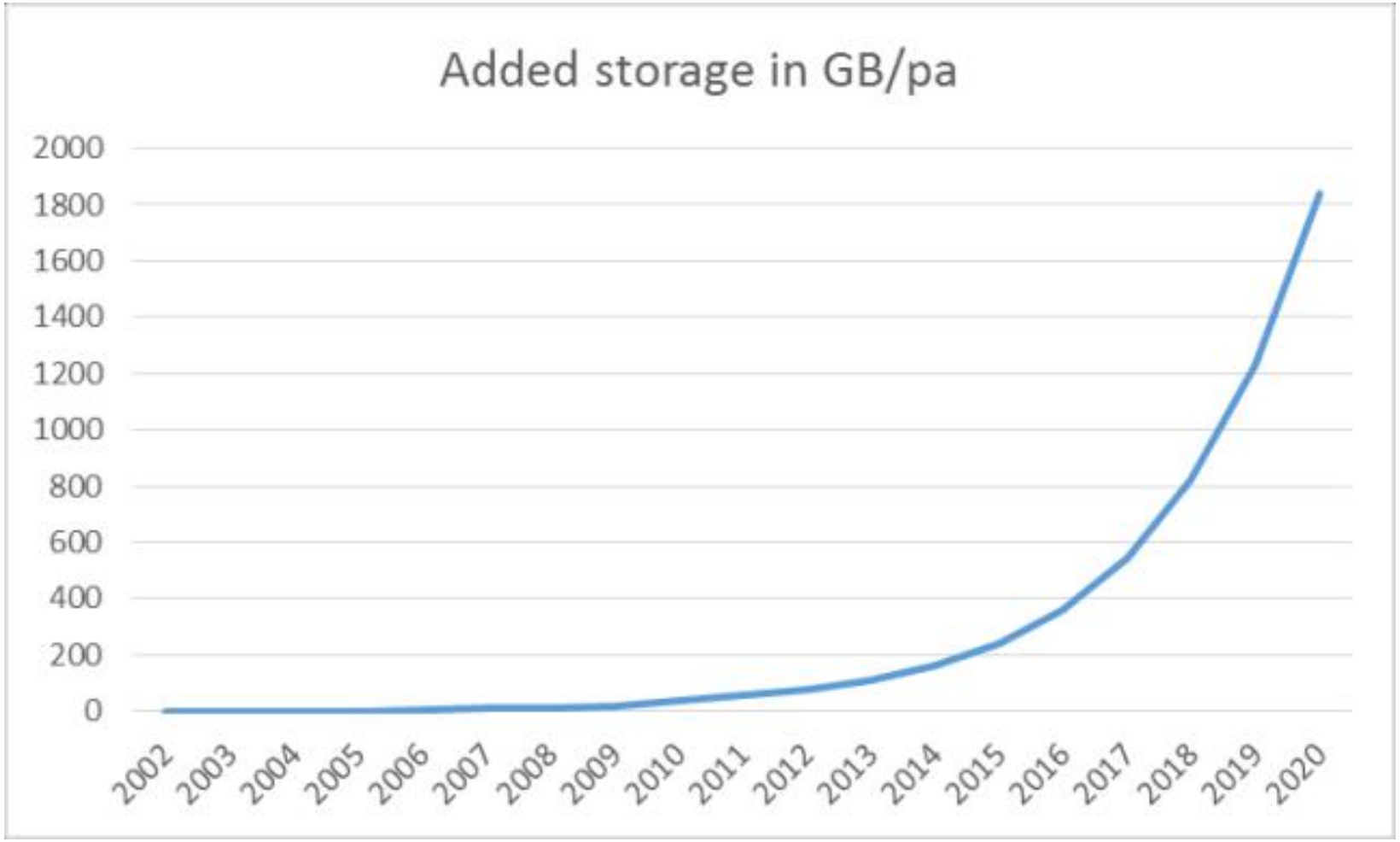
De que tipo são esses dados ?



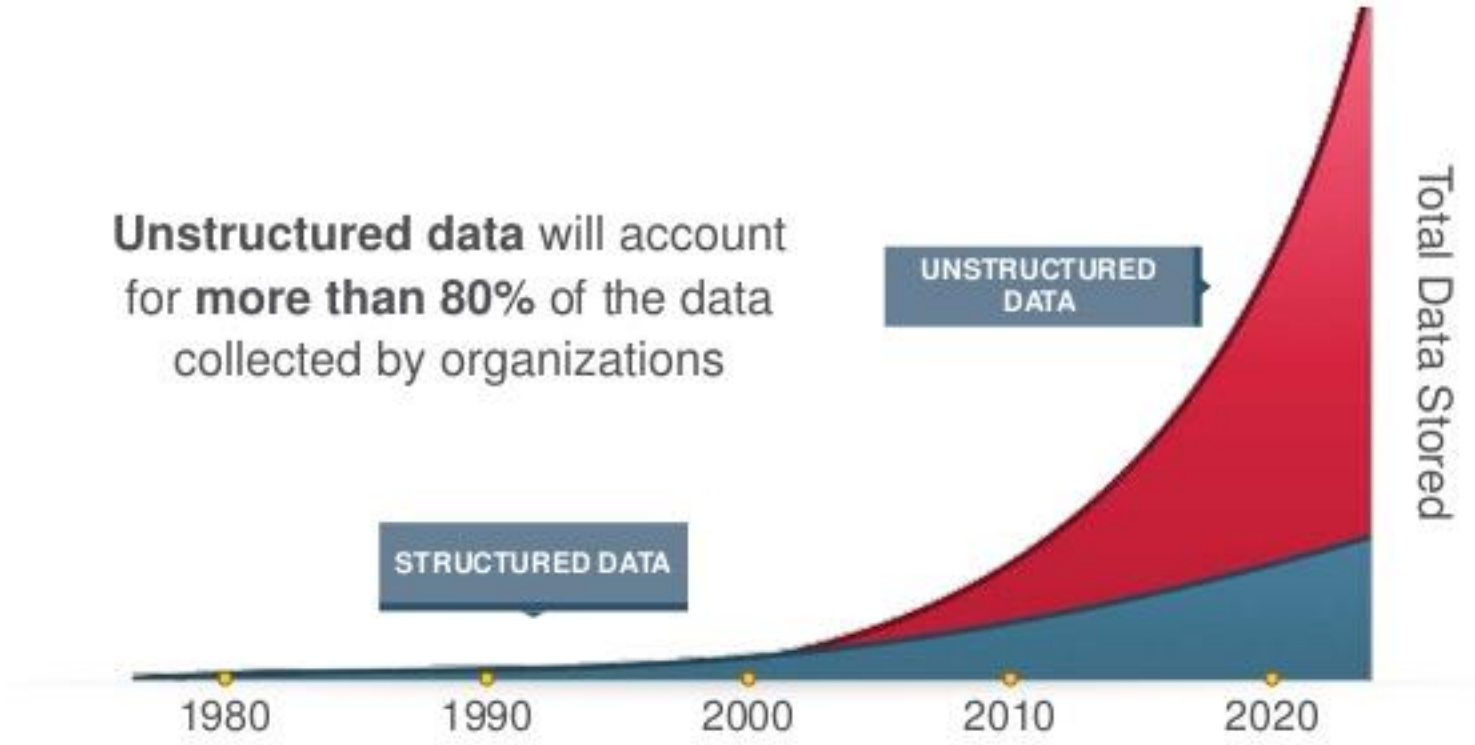
E o custo de armazenamento?



E a capacidade de armazenamento?



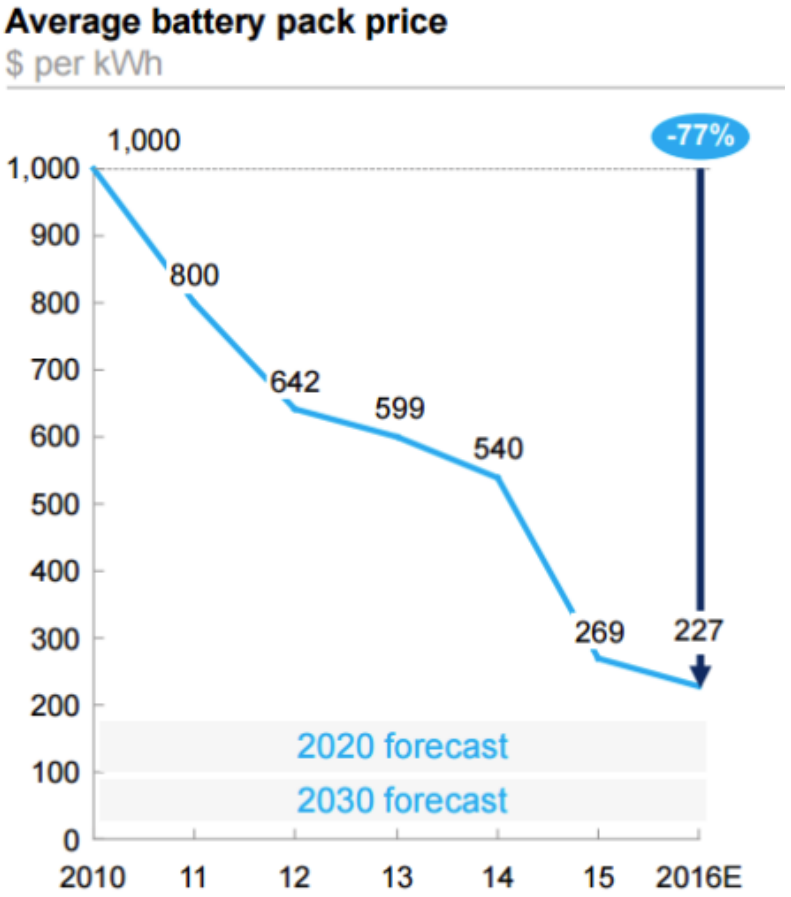
Como são esses dados?



Source: Human-Computer Interaction & Knowledge Discovery in Complex Unstructured, Big Data

 MAPR 4

E o custo das baterias?



O que é Big Data?

- Várias definições
  - Dados que são grandes demais para sistemas tradicionais de processamento de dados
  - Dados que precisam de novas técnicas para serem processados
  - Dados que são muito complexos
  - Dados que são importantes
  - Coletar dados agora para entendê-los depois

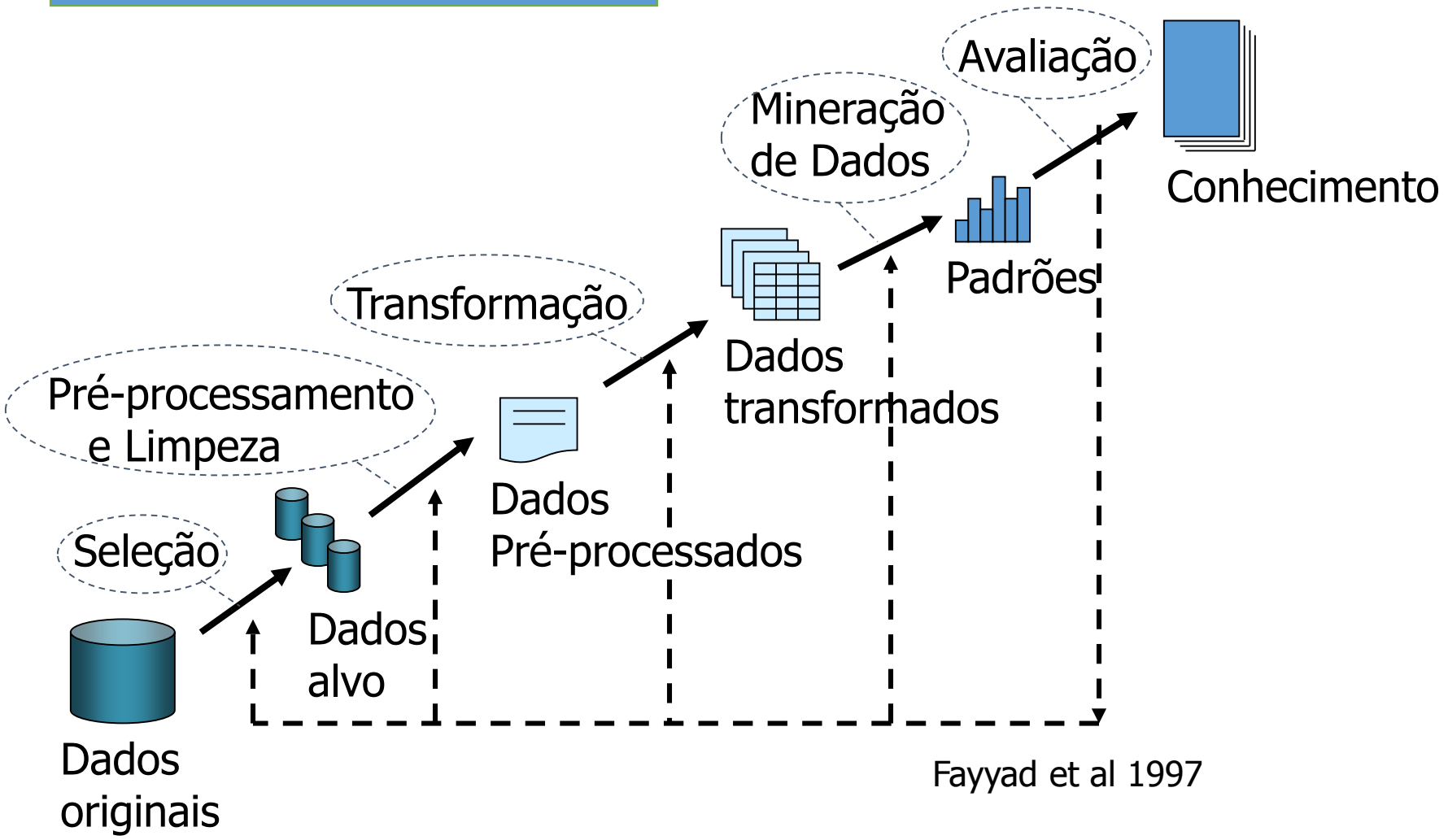
KDD - *Knowledge Discovery in Databases*

- Bases de Dados podem conter (esconder) dados preciosos
- Existe um interesse crescente em explorar esses dados armazenados
  - Descobrir conhecimento novo
  - Apoio à tomada de decisão

KDD - *Knowledge Discovery in Databases*

- Processo de extrair conhecimento de dados
  - Útil
  - Novo
  - Válido
  - Potencialmente compreensível
- Processo interativo e iterativo
  - Várias etapas

KDD


Avaliação
Mineração
de Dados
Conhecimento
Padrões
Transformação
Dados
transformados
Pré-processamento
e Limpeza
Dados
Pré-processados
Seleção
Dados
alvo
Dados
originais
Fayyad et al 1997

SELEÇÃO

- Extrai uma amostra do conjunto de dados para extração de conhecimento
  - Seleciona “manualmente” entre os dados disponíveis
    - Subconjunto de registros (instâncias ou exemplos)
    - Subconjunto de atributos considerados relevantes para o problema
      - Elimina atributos que sejam claramente irrelevantes

Pré-processamento e Limpeza

- Melhora a qualidade dos dados e facilita sua posterior utilização
- Engloba várias operações
  - Seleção “automática” de atributos
  - Conversão de valores
  - Tratamento de atributos com valores ausentes
  - Eliminação de dados duplicados
  - Detecção (e remoção) de ruído

Transformação

- Inclui operações que modificam valores para um dado atributo
  - Cada operação deve ser aplicada a todos os valores do atributo
    - Em todos os objetos
  - Ex.: normalização, valor absoluto, padronização, log, codificação 1-de-m,...

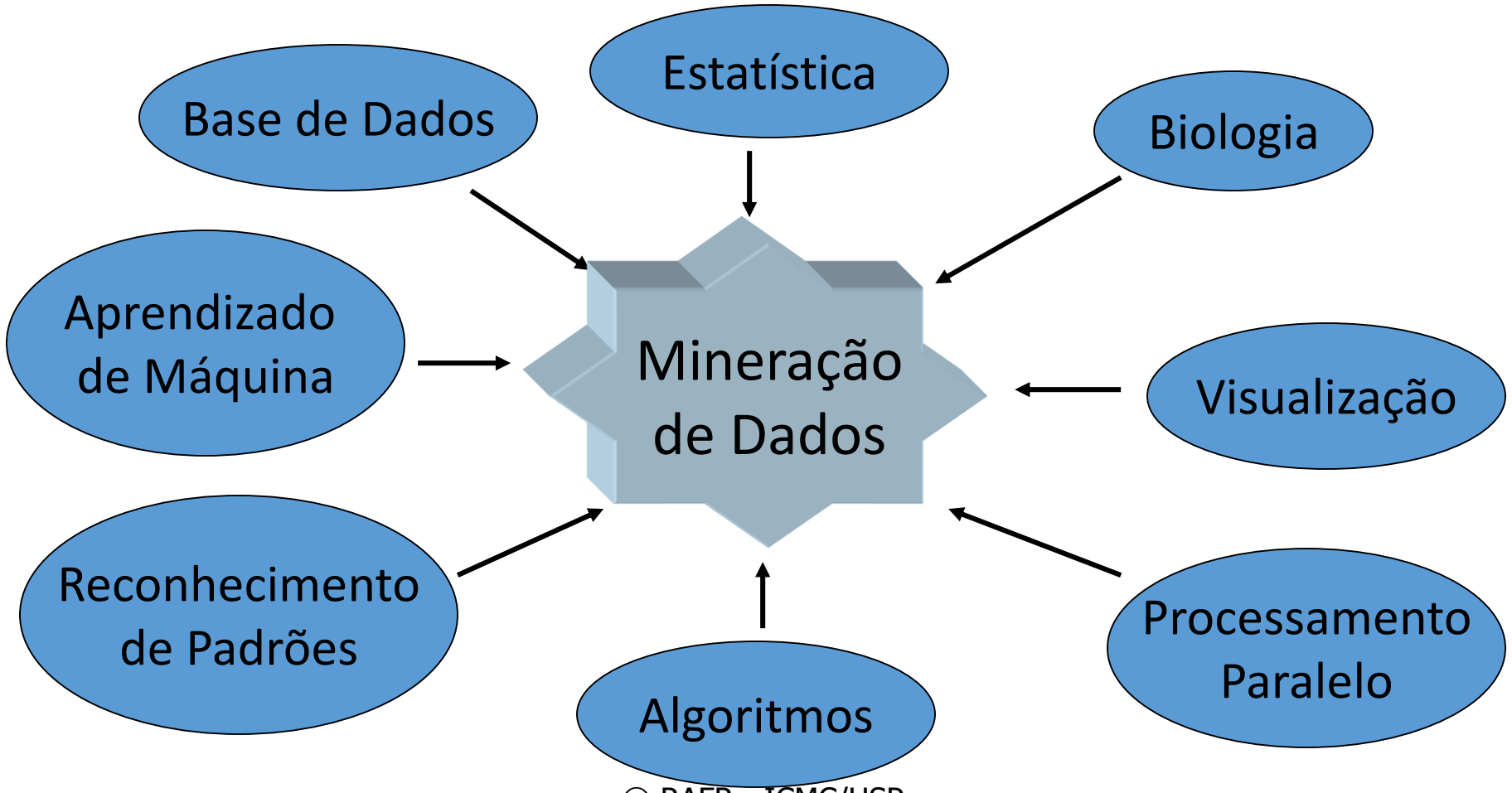
Mineração de Dados

- Principal passo no processo de KDD
  - Mineração de Dados (DM) e KDD são frequentemente utilizados como sinônimos
- Difícil identificar fronteiras da etapa de MD no processo de KDD
  - Pré-processamento e transformação de dados são geralmente vistos como uma parte da MD

MD x KDD

- MD: ferramentas básicas utilizadas para extrair padrões de dados
- KDD: processo que engloba o uso dessas ferramentas, além de:
  - Seleção, pré-processamento, transformação dos dados
  - Interpretação e validação do conhecimento
    - Geração de conhecimento
    - Suporte à tomada de decisão

Mineração de Dados



Mineração de Dados

- Outros termos utilizados para MD, KDD e CD
  - Extração de conhecimento
  - Descoberta de informação
  - Extração de padrões
  - Análise exploratória de dados
  - Analítica (*Data analytics* ou *analytics*)

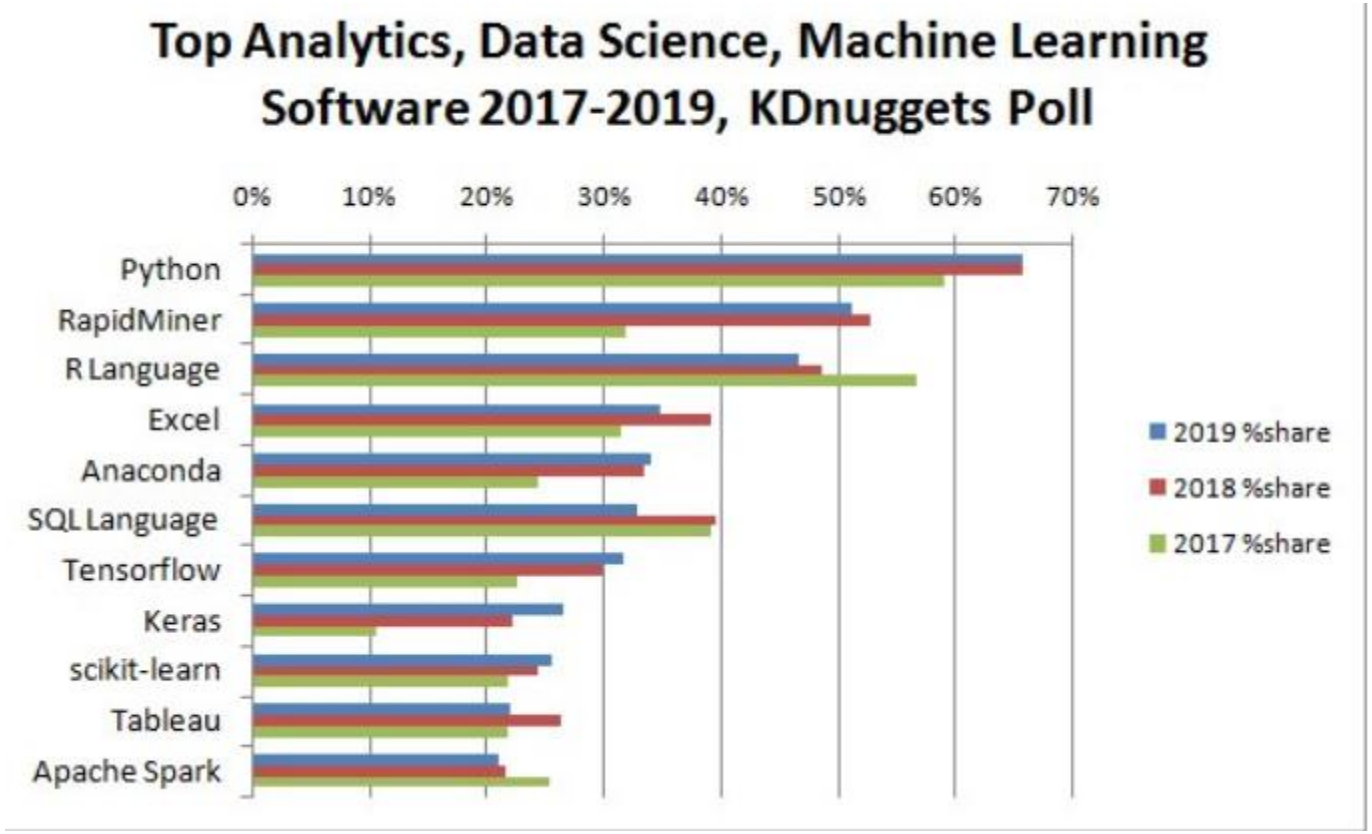
Analítica

- Ciência que analisa dados crus para extrair padrões desses dados
  - Pode englobar coleta e organização dos dados
- Analítica preditiva (*predictive analytics*)
  - Extrai modelos (conhecimento) a partir de dados para realizar predições futuras
- Analítica descritiva (*descriptive analytics*)
  - Sumariza ou condensa dados para extrair conhecimento

Interpretação e Avaliação

- Interpretação dos padrões encontrados na etapa de MD
  - Possível retorno a qualquer uma das etapas anteriores para iteração adicional
- Validar padrões encontrados
  - Importante consulta a um especialista
- Inclui análise estatística
- Ferramentas de visualização fornece um suporte importante

Ferramentas



https://www.kdnuggets.com/2019/05/poll-top-data-science-machine-learning-platforms.html

O Mercado

- A demanda está crescente por profissionais, que tenham conhecimento tanto de **análise de grande volume de dados** quanto das **tecnologias de Inteligência Artificial** para aplicação prática em Inteligência de Negócios.
- É eminente a necessidade de **um novo tipo de profissional** capaz de atuar na **nova era da conhecimento e da conectividade.**

The top 15 emerging jobs in the U.S.

#1 74% annual growth

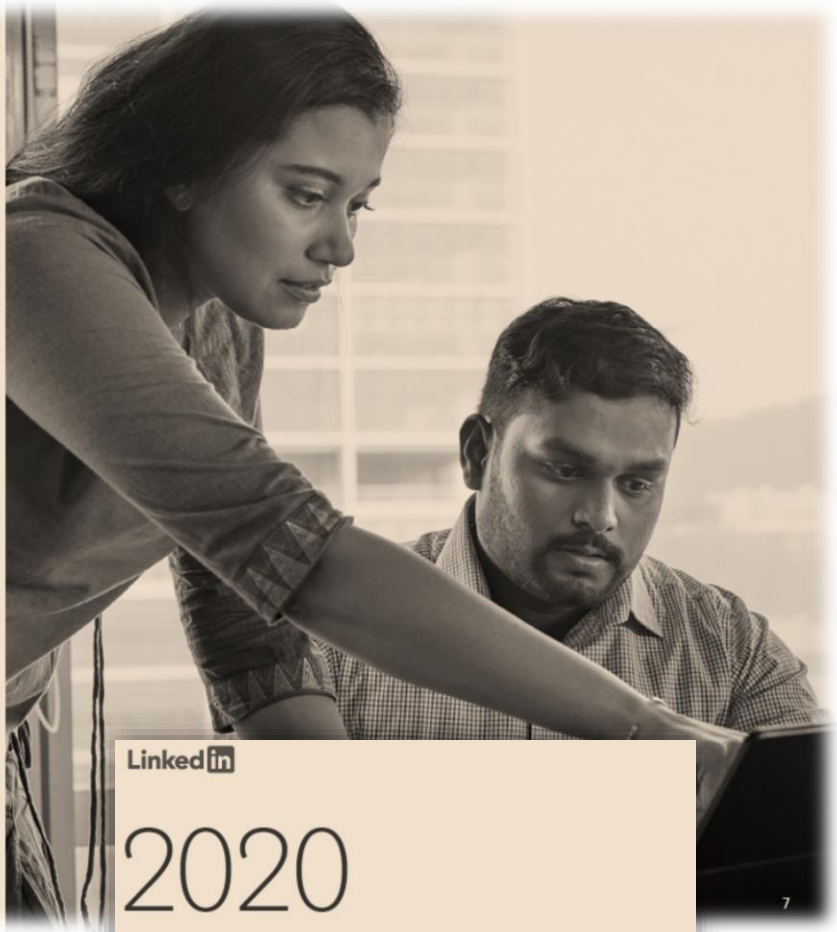
Artificial Intelligence Specialist

What you should know:
Artificial Intelligence and Machine Learning have both become synonymous with innovation, and our data shows that's more than just buzz. Hiring growth for this role has grown 74% annually in the past 4 years and encompasses a few different titles within the space that all have a very specific set of skills despite being spread across industries, including artificial intelligence and machine learning engineer.

Skills unique to the job:
Machine Learning, Deep Learning, TensorFlow, Python, Natural Language Processing

Where the jobs are:
San Francisco Bay Area, New York, Boston, Seattle, Los Angeles

Top industries hiring this talent:
Computer Software, Internet, Information Technology & Services, Higher Education, Consumer Electronics



LinkedIn

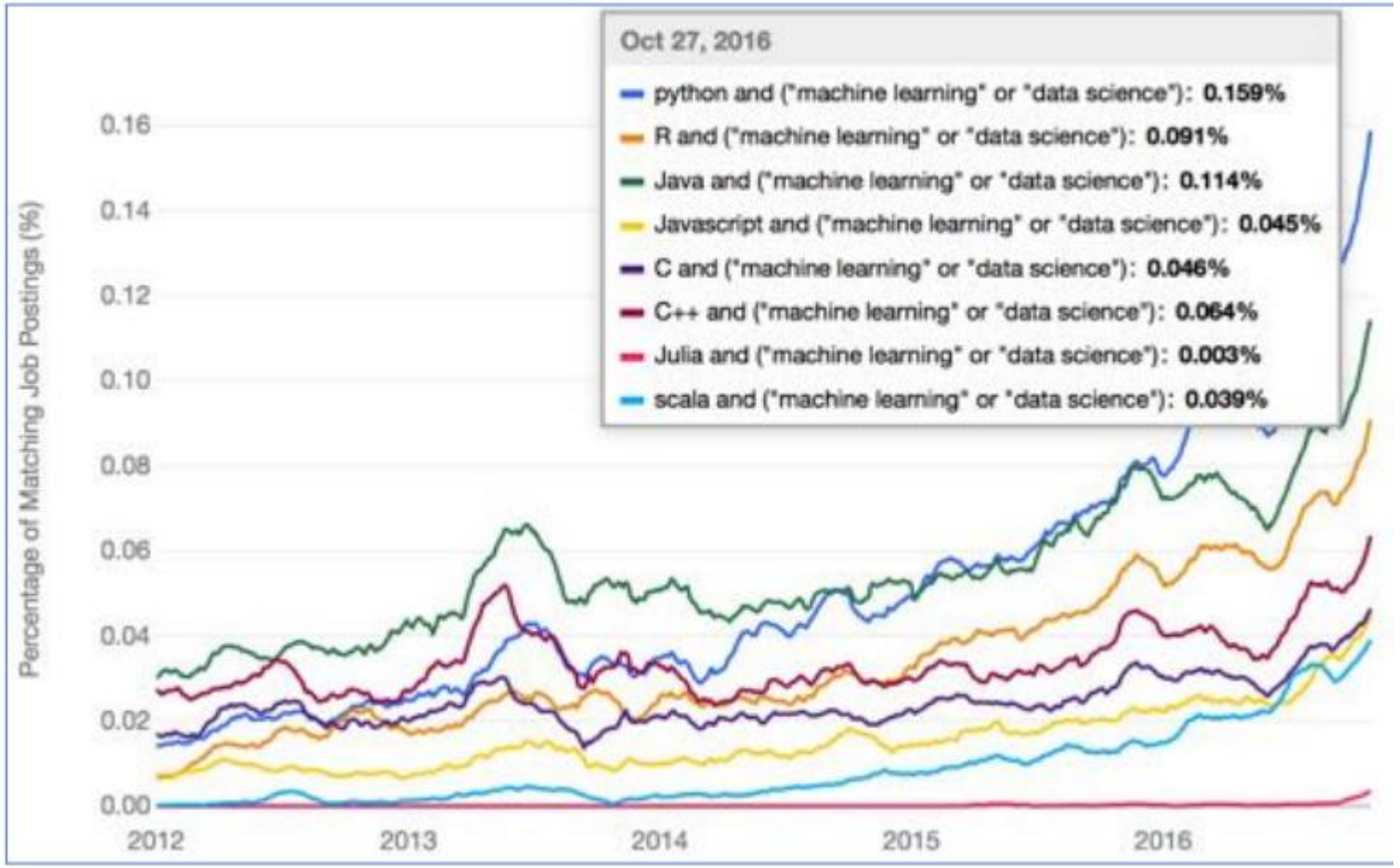
2020 Emerging Jobs Report

7

O **crescimento de contratações** para esta função cresceu 74% ao ano nos últimos 4 anos.

Inteligência Artificial e Aprendizado de Máquina tornaram-se sinônimos de inovação.

Por que usar Python?



http://www.kdnuggets.com/2017/01/most-popular-language-machine-learning-data-science.html

Por que usar Python?

- Fácil de ler e aprender
- Comunidade vibrante
- Conjunto crescente e em evolução de bibliotecas
- Gestão de dados
- Processamento analítico
- Visualização
- Aplicável a cada etapa do processo de ciência de dados
- Notebooks

Aprofundar em Python

- **Python for Data Science and Machine Learning Bootcamp** – UDEMY
- **Coursera:** Udacity's Intro to Data Analysis