Curso 2 – Olhando para os dados: CD, AM e DM

Profa. **Roseli Ap. Francelin Romero**

MBA em Inteligência Artificial e BigData

Depto. de Ciências de Computação
ICMC - USP

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

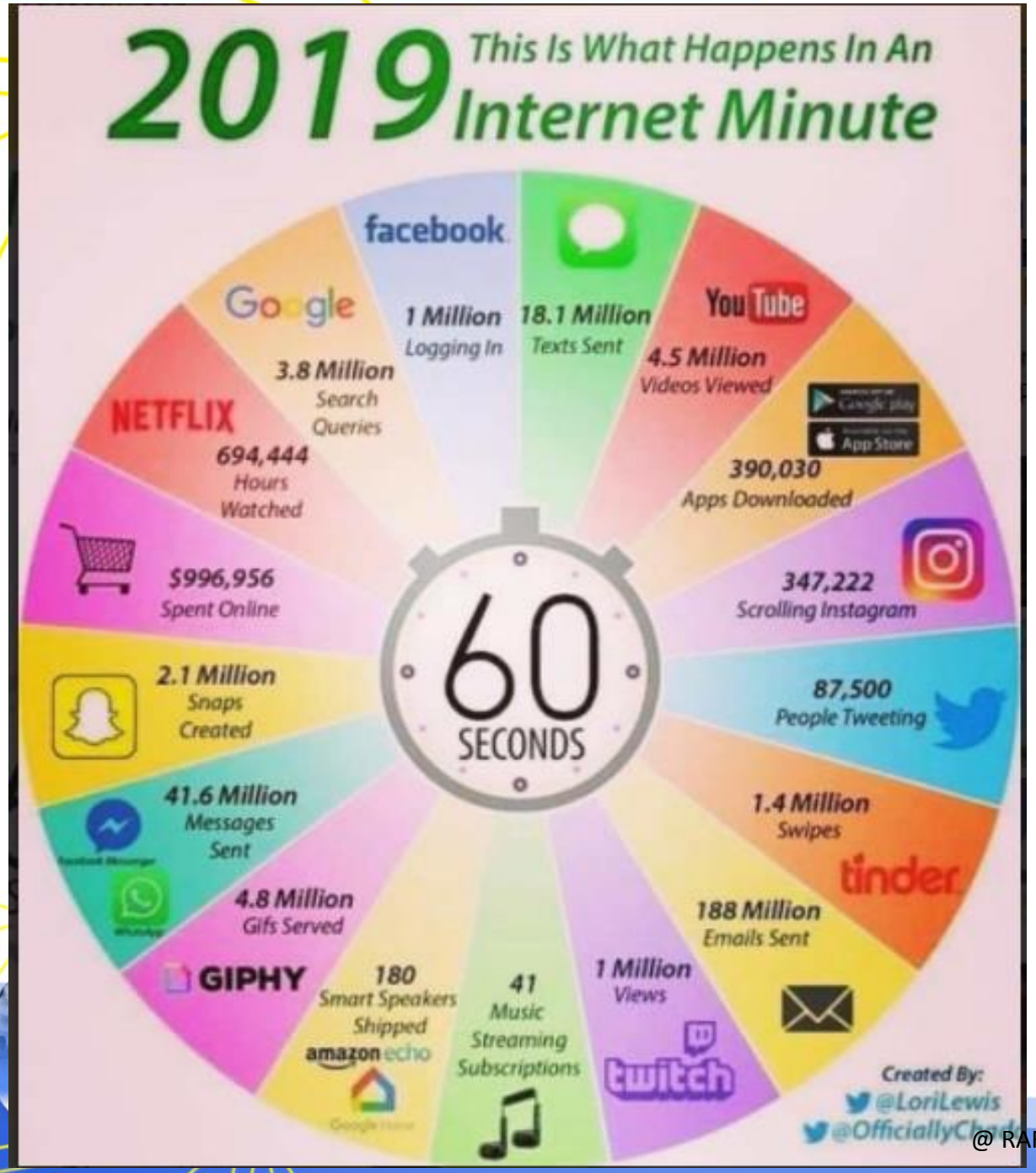


Image Source: http://www.marketwatch.com/story/one-chartshows-everything-that-happens-on-the-internet-in-just-oneminute-2016-04-26

Por Minuto

204 Million emails

facebook

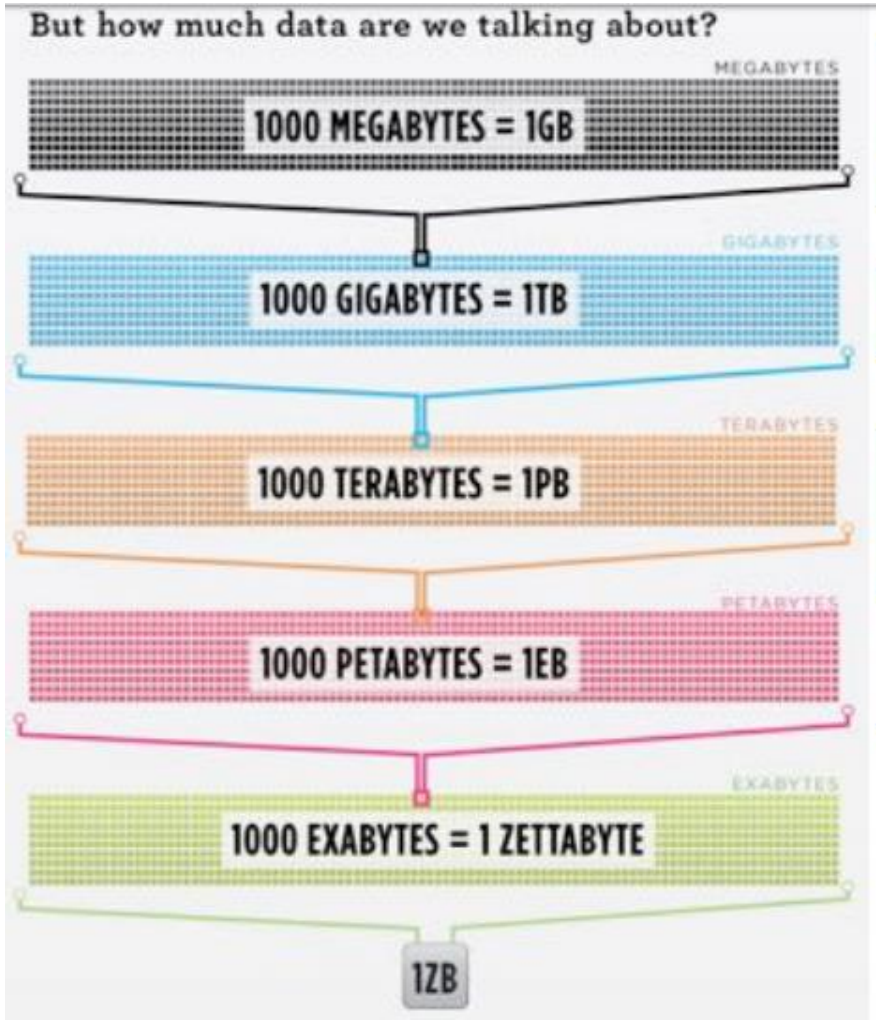
200,000 photos

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

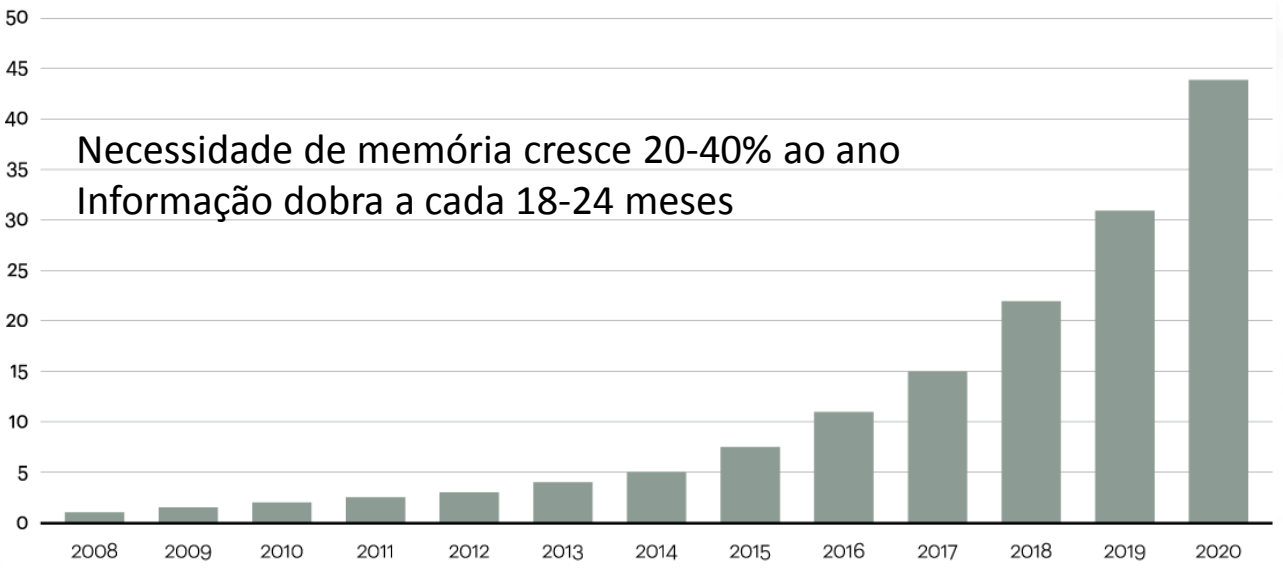
1 TB ~= 300 hours of good quality video

LHC ~= 15 PBs a year

Como cresce a quantidade de dados?

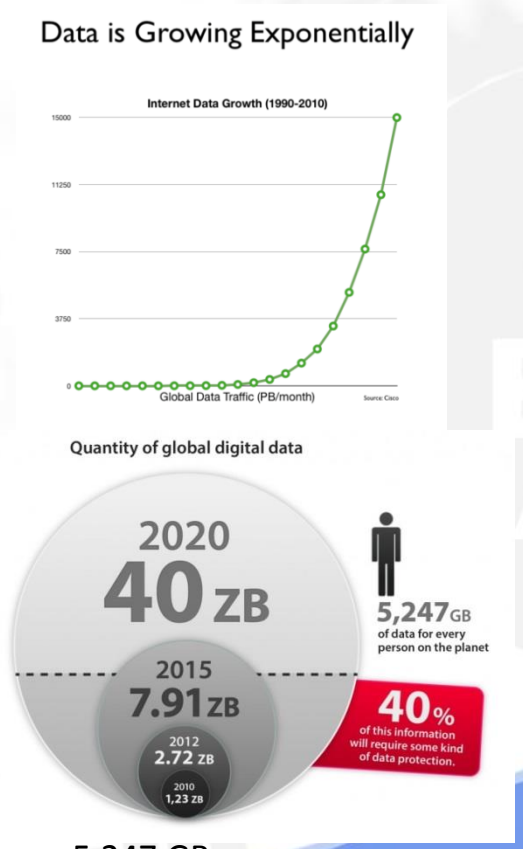
Figure 1

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?

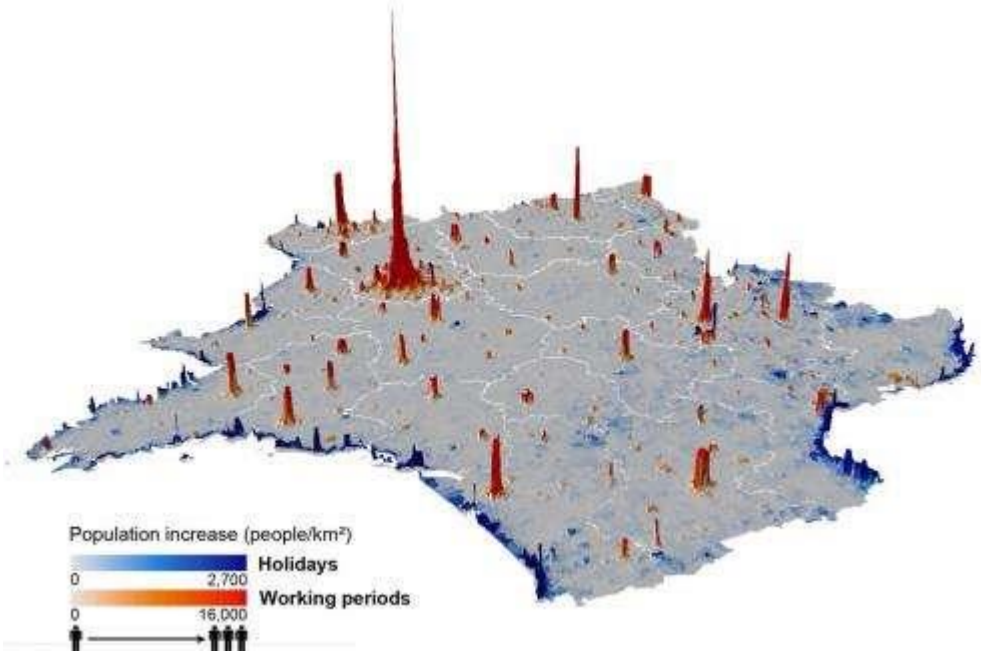
- Dispositivos eletrônicos
  - Sinais de localização de smartphones
  - Logs de servidores de aplicações
    - Jogos e web sites
  - Sensores de dados
    - Climáticos, reservatórios de água, corpo humano
  - Imagens e vídeos
    - Câmeras de monitoramento de trânsito, de segurança

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

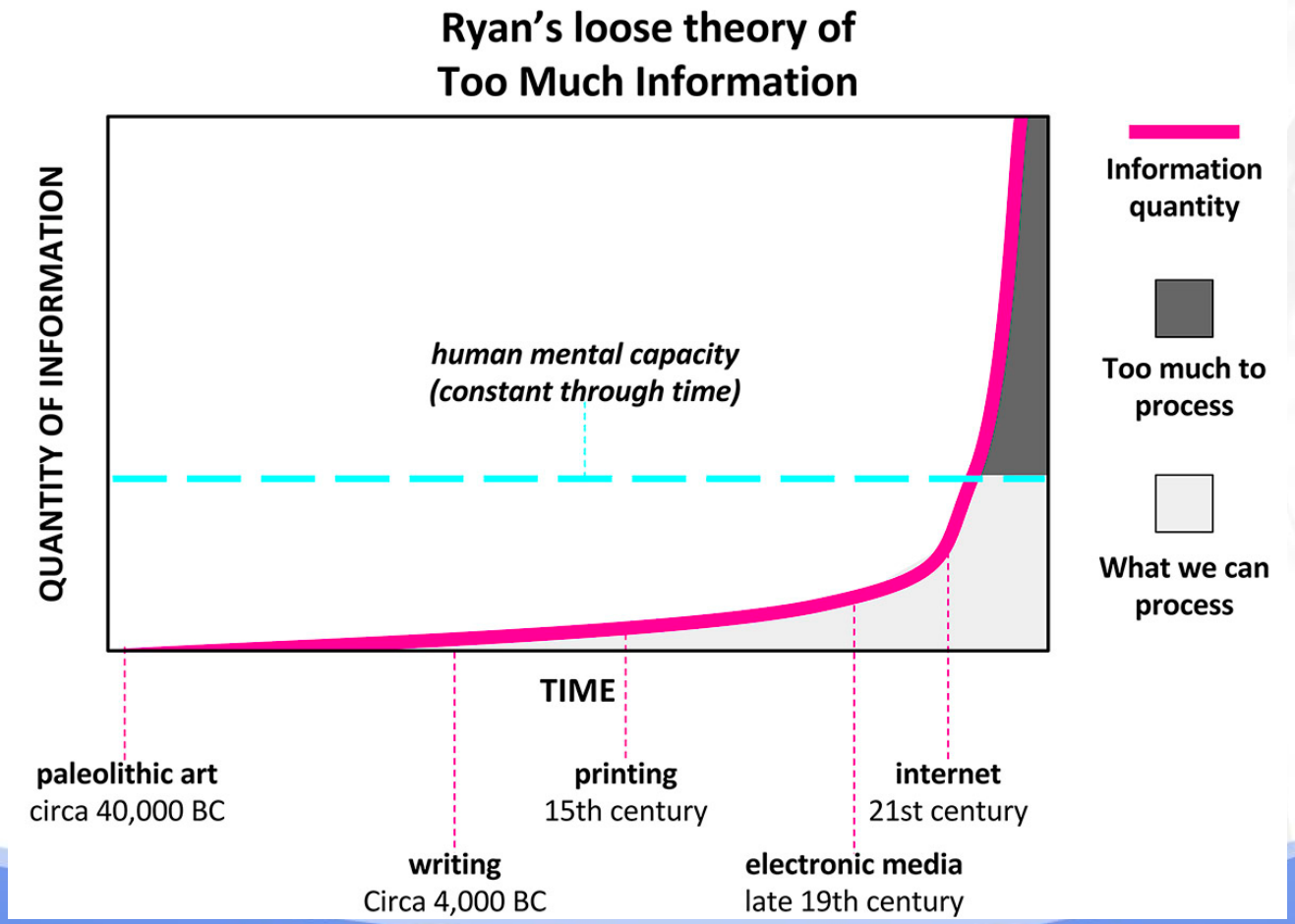
http://phys.org/news/2014-10-cellphone-population-density.html#jCp

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

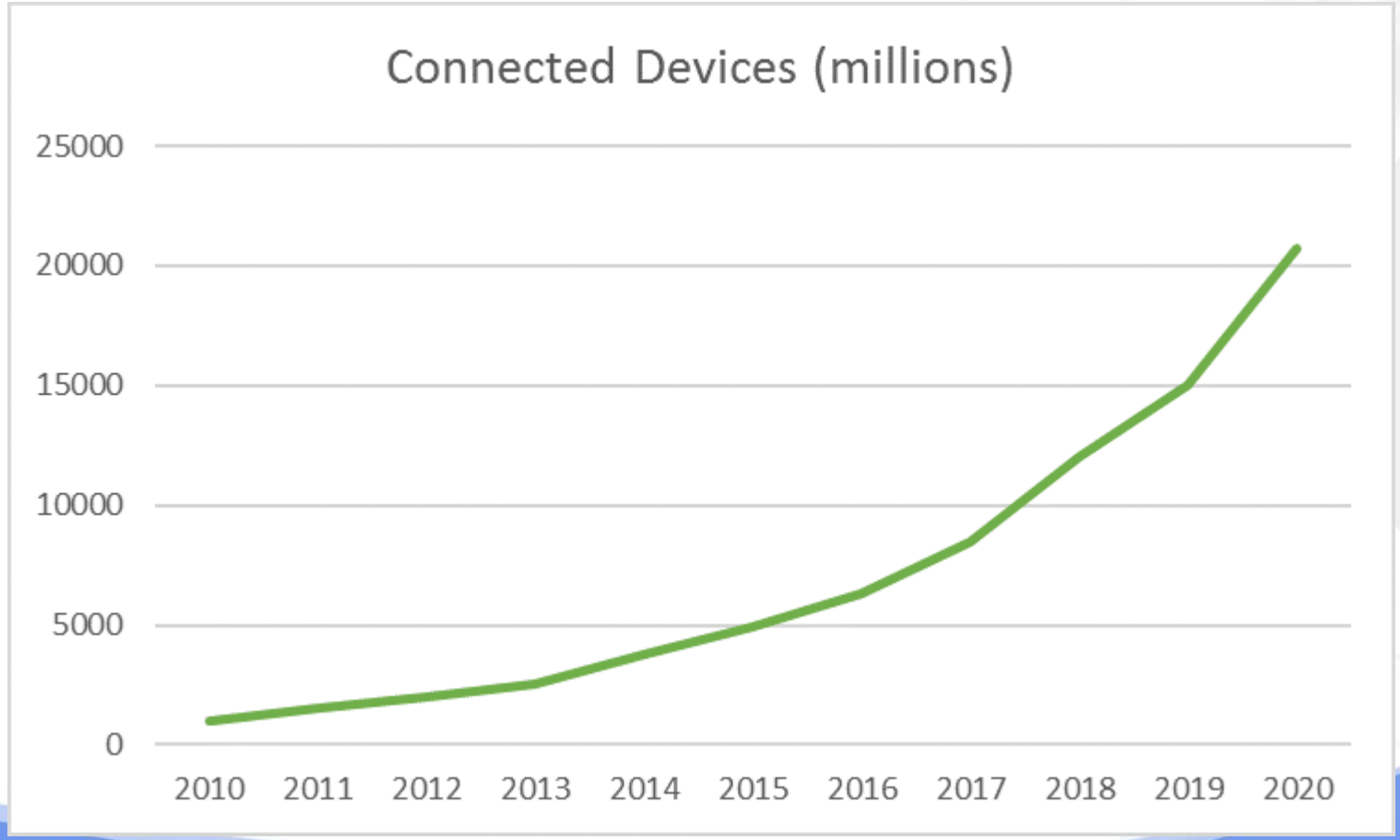
Avanços recentes nas tecnologias para **aquisição, transmissão**, **armazenamento** e **processamento** de dados

Big Data

Sobrecarga de informação

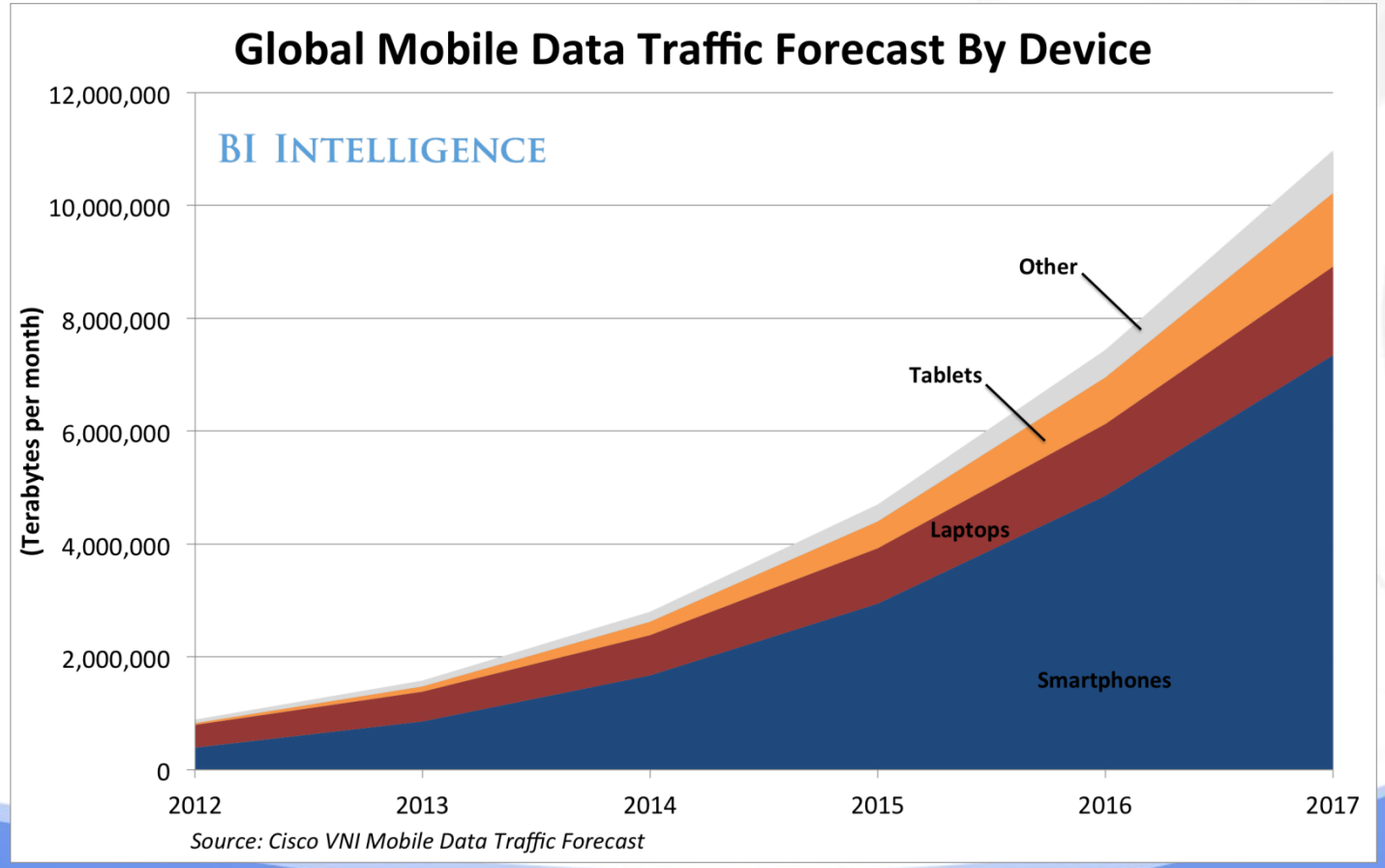


Como avança a transmissão de dados? (Velocidade)

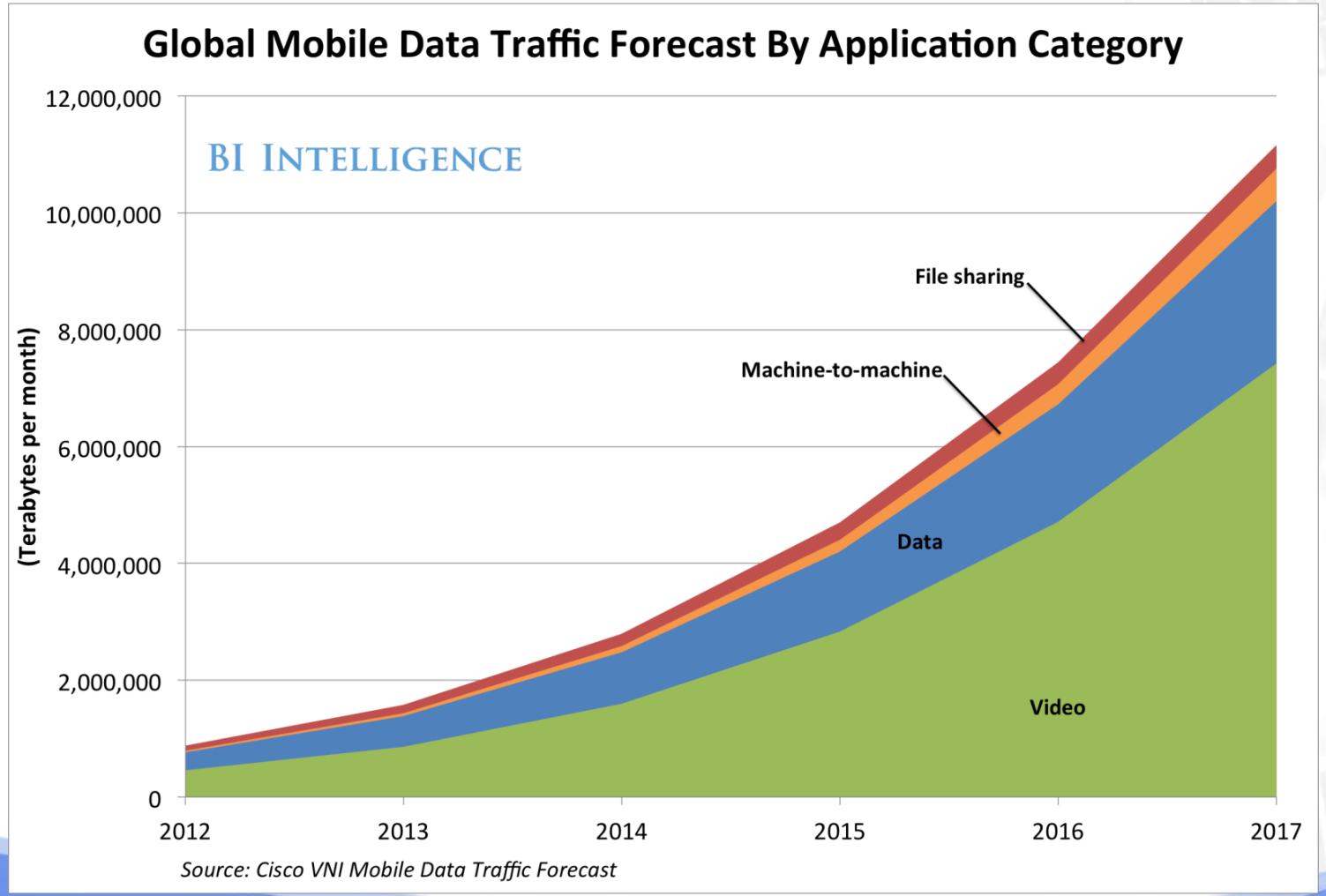


http://motherboard.vice.com/blog/the-next-five-years-of-explosive-internet-growth-in-seven-graphs

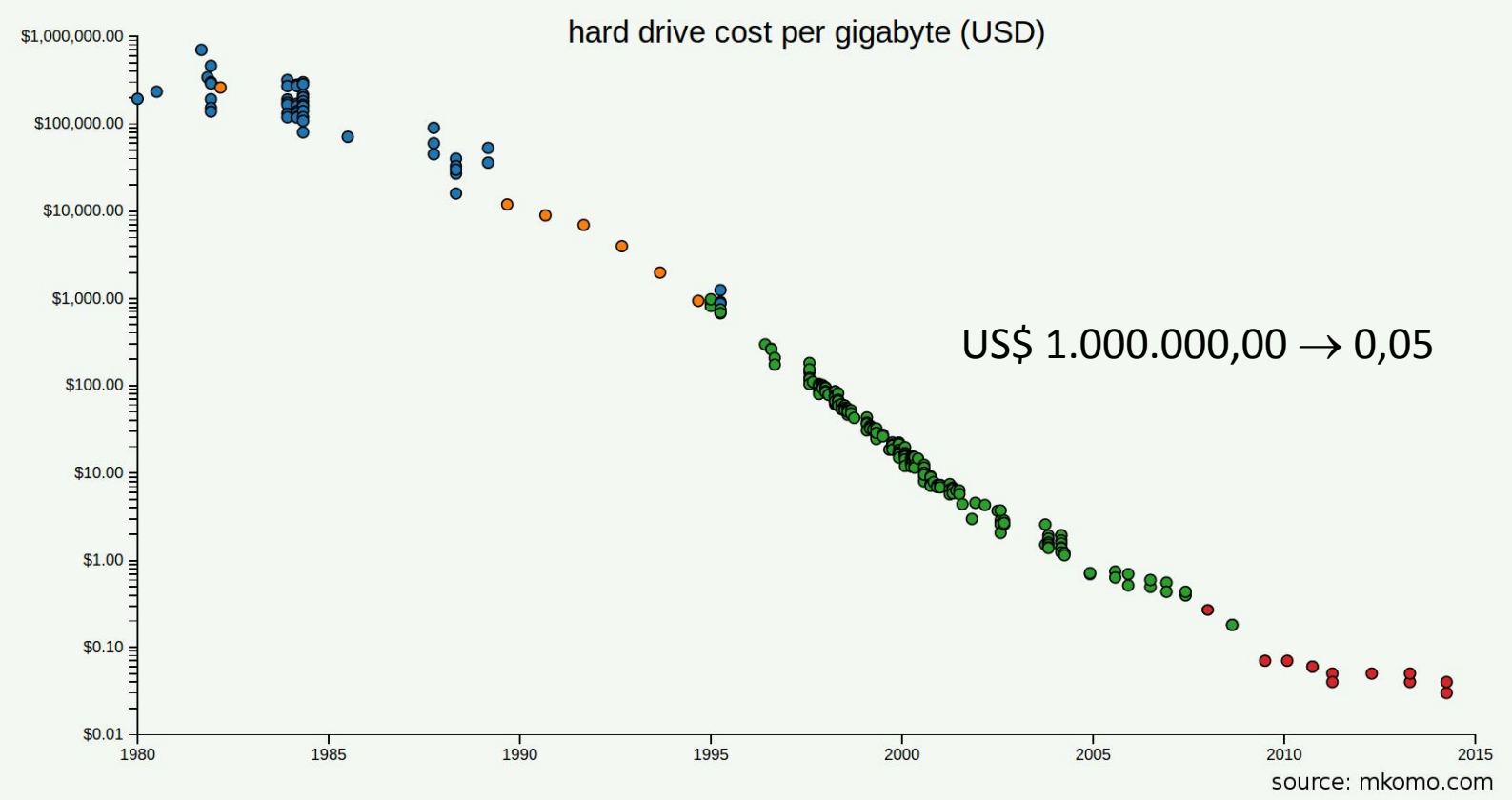
De que tipo de dispositivos os dados chegam?



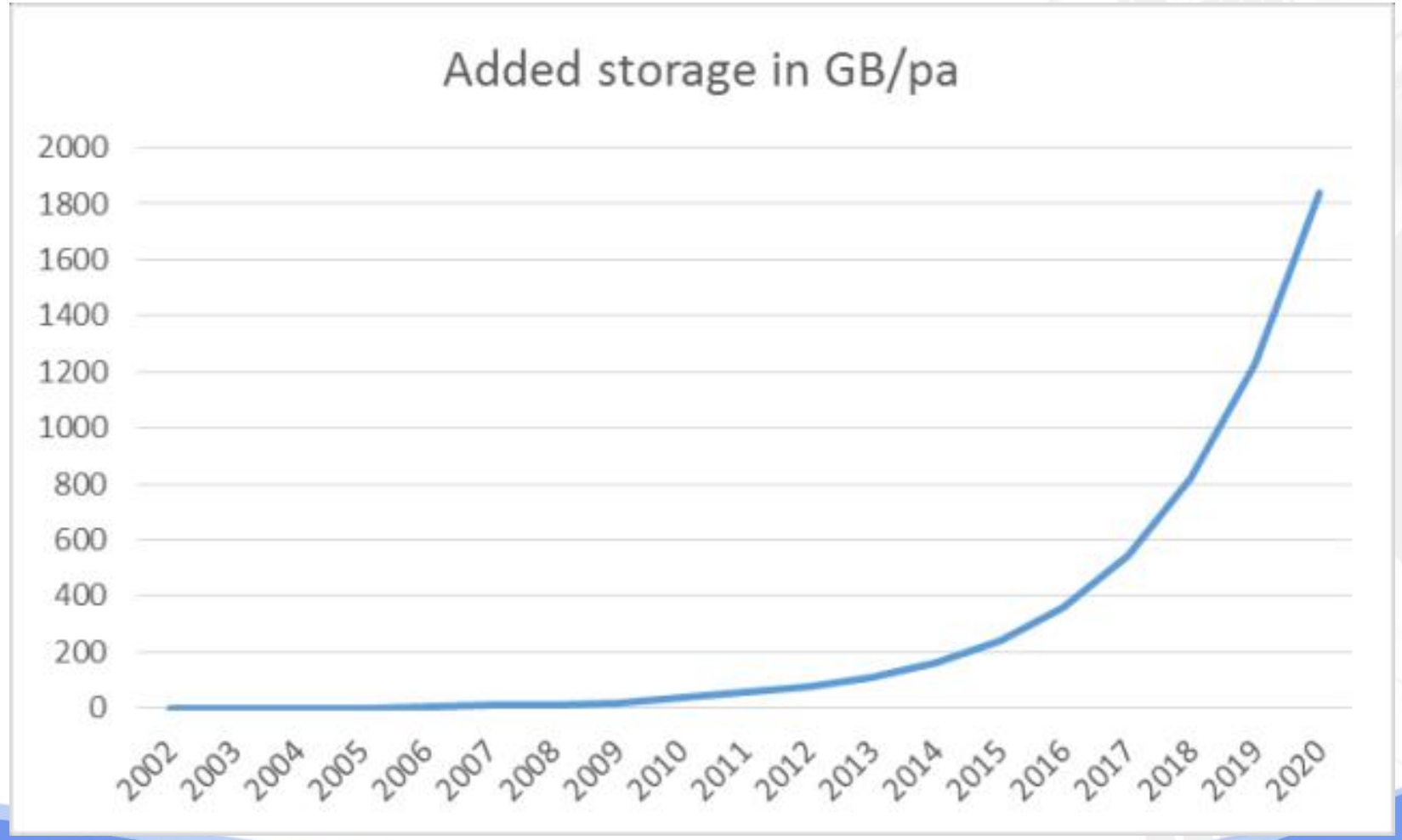
De que tipo são esses dados ?



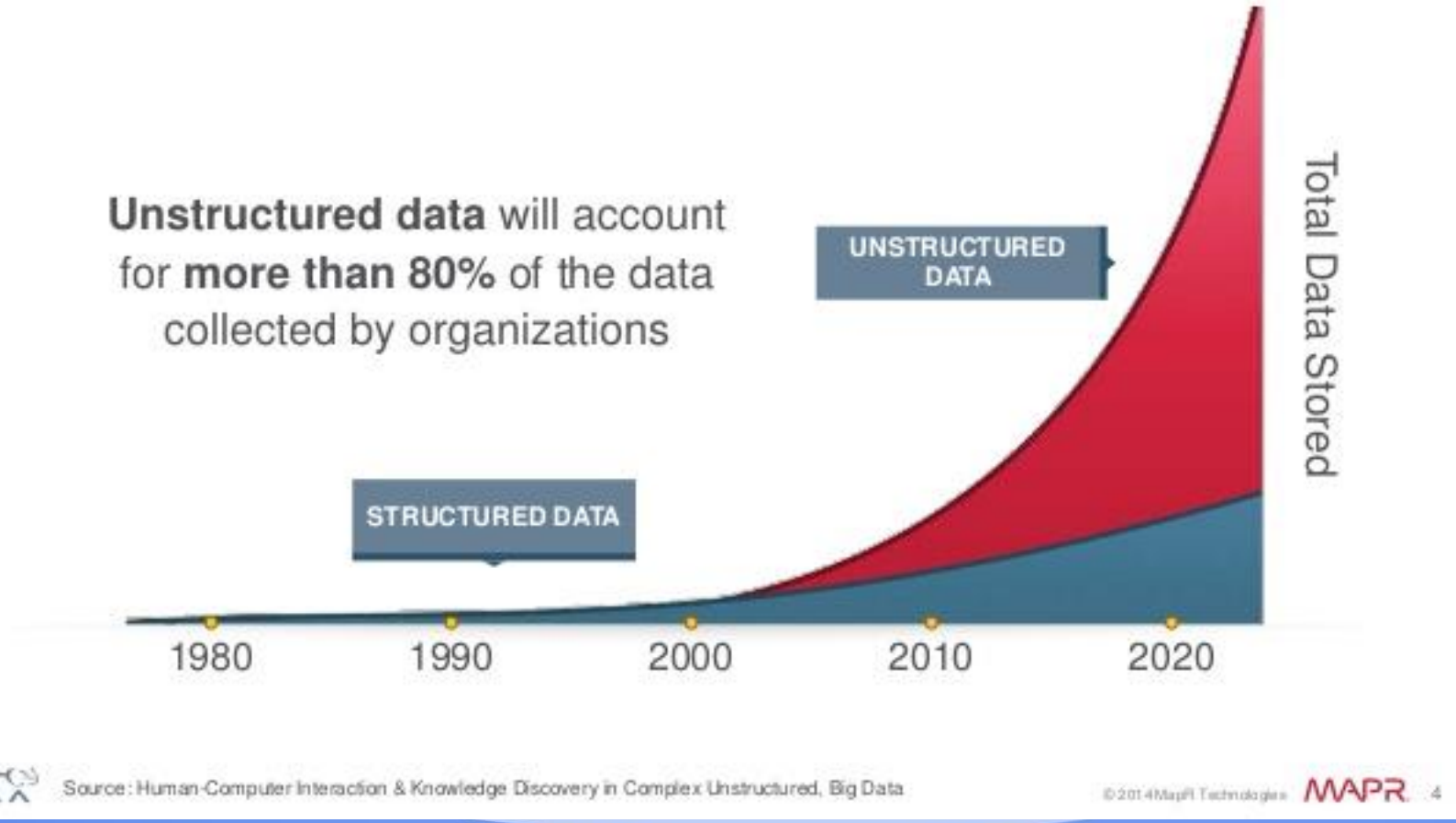
E o custo de armazenamento?



E a capacidade de armazenamento?



Como são esses dados?



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

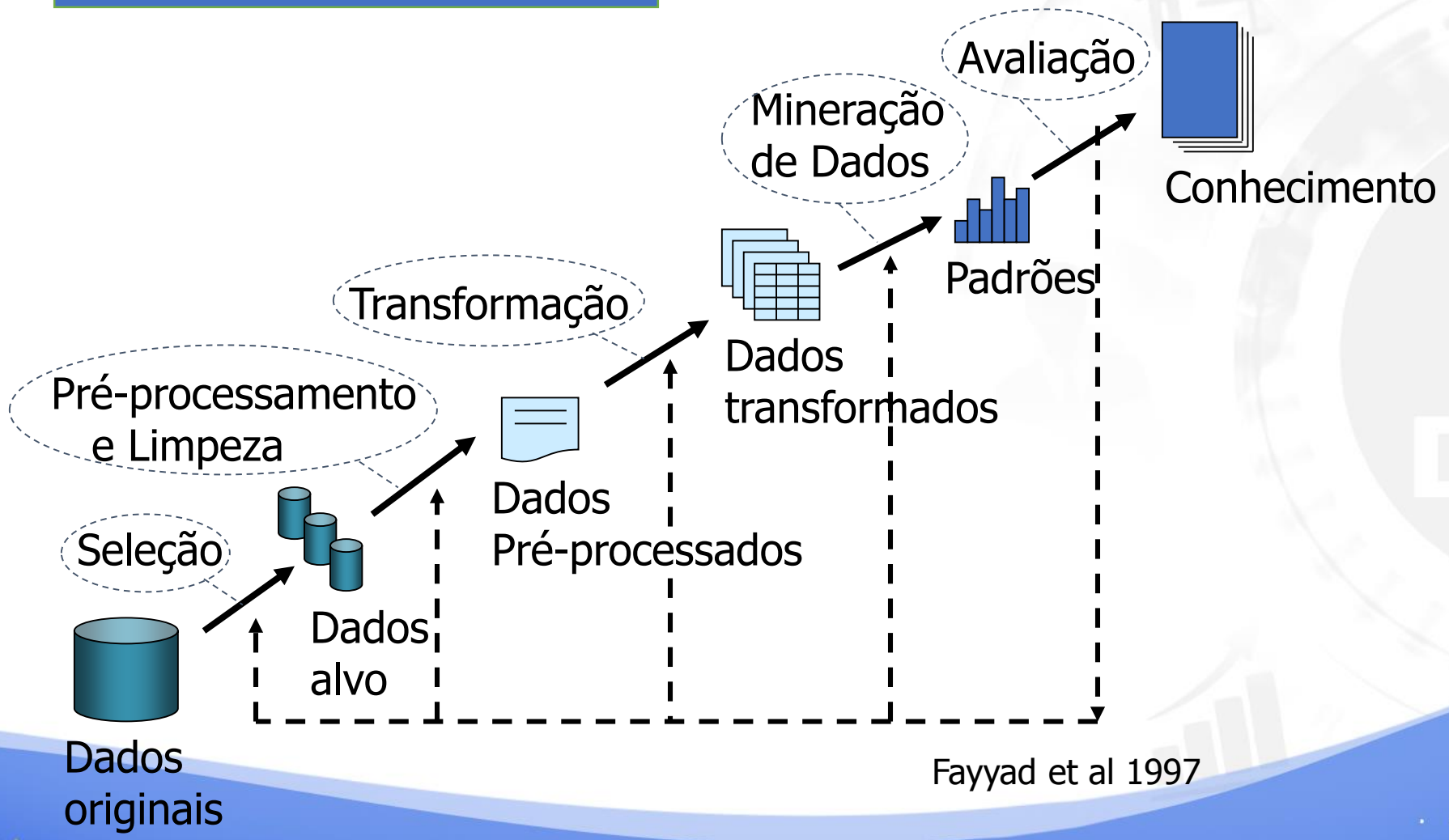
KDD - *Knowledge Discovery in Databases*

- 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

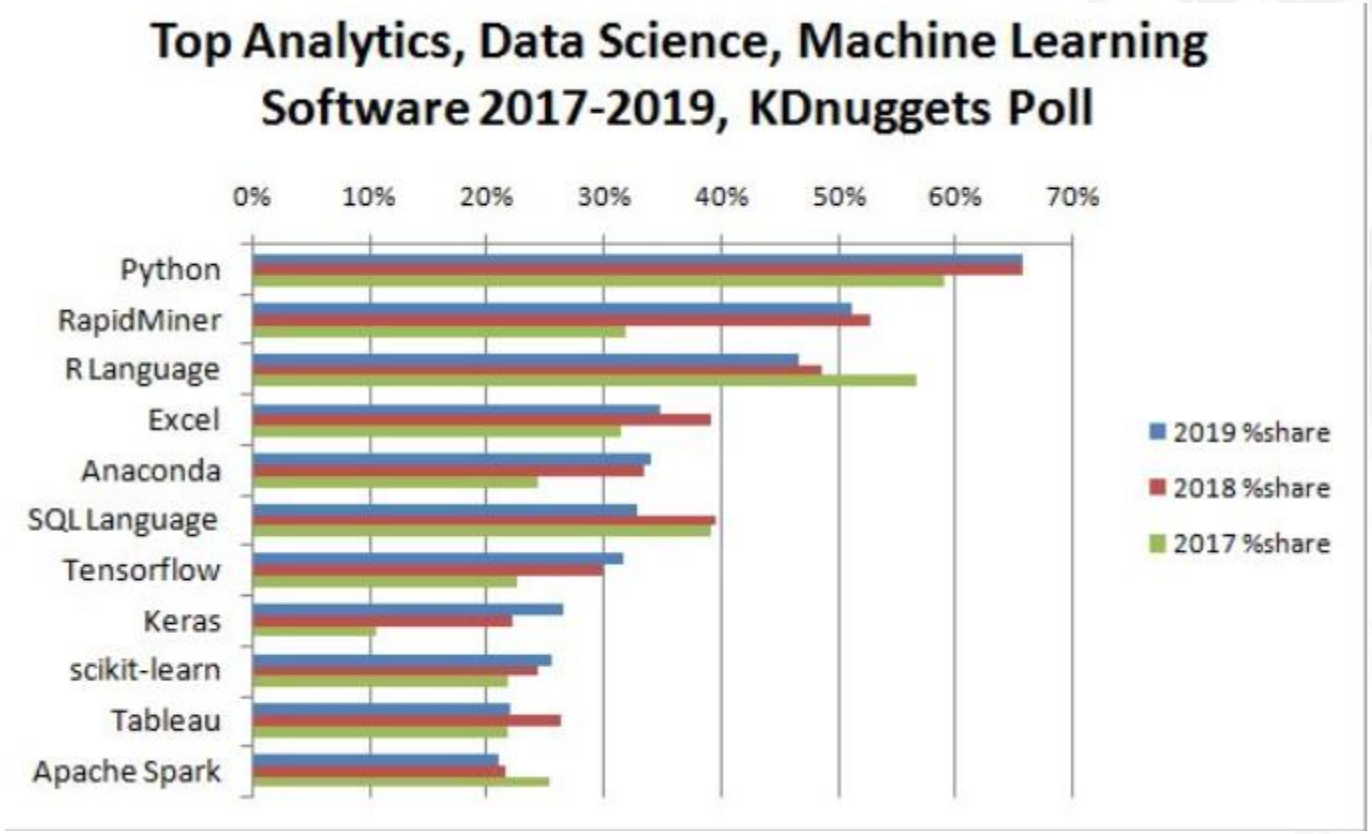


Fayyad et al 1997

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