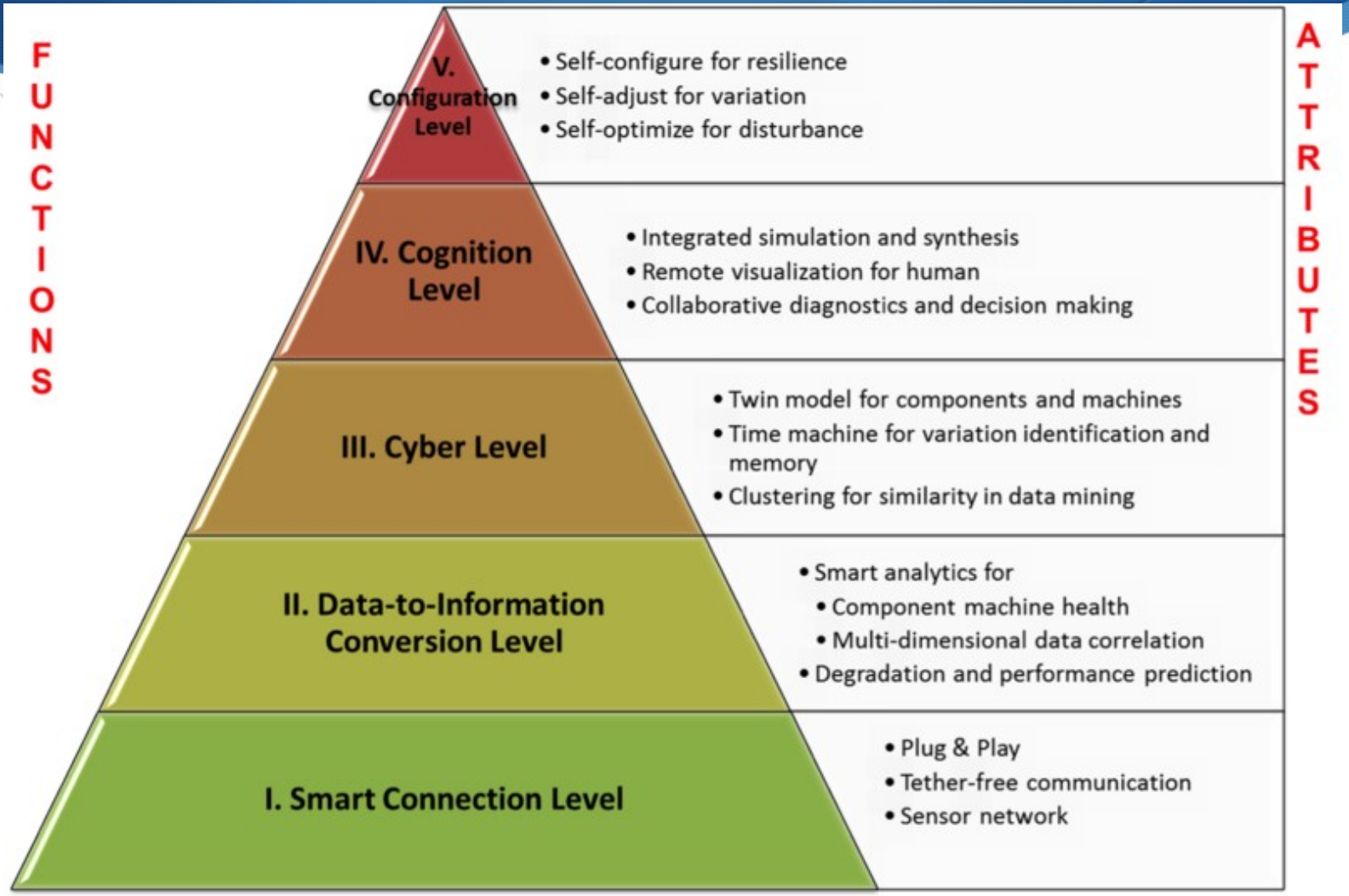
Big Data

Daniela Maria Uez
dani.uez@gmail.com

11 JULHO 2019

UCS
UNIVERSIDADE
DE CAXIAS DO SUL

Arquitetura 5C

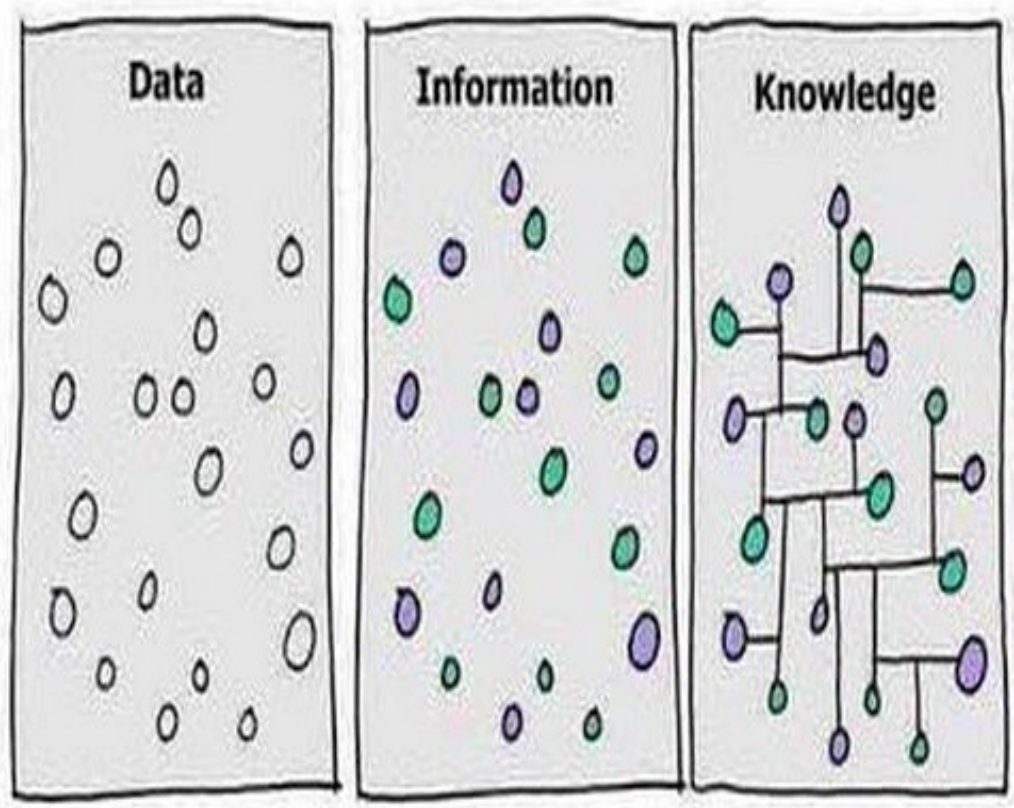


Dados x Informação x Conhecimento

Dado: observação sobre o estado do mundo

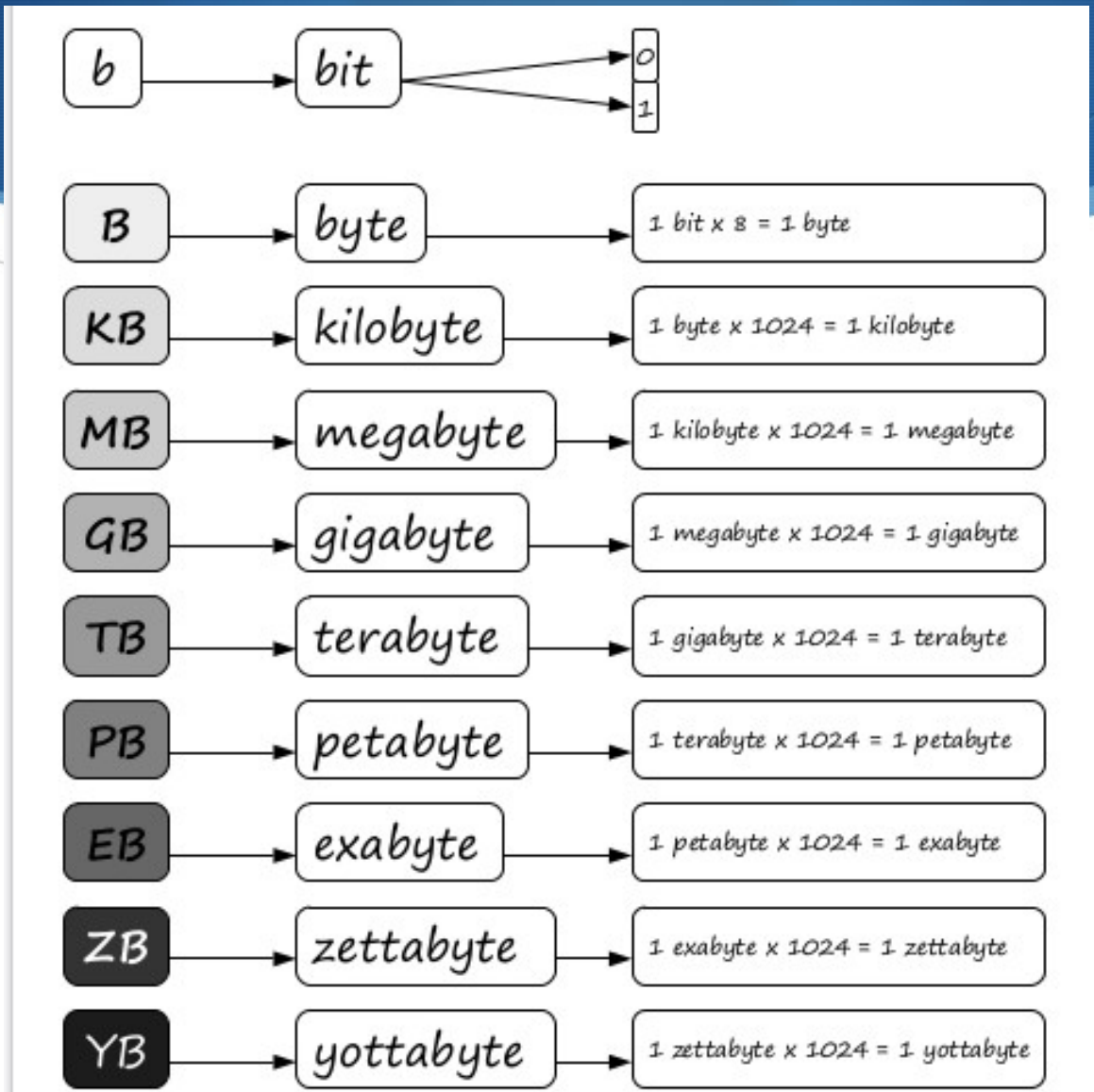
Informação: dado dotado de relevância e propósito

Conhecimento: reflexão que inclui reflexão, síntese, contexto sobre uma informação



(Fonte: www.worldwidewebsize.com)

Unidades de medida dados



O que é Big data?

Não existe uma definição consensual

Dados com maior variedade que chegam em volumes crescentes e com velocidade cada vez maior.

“**Grandes** dados”

O que é Big data?

Não existe uma definição consensual

Dados com maior variedade que chegam em volumes crescentes e com velocidade cada vez maior.

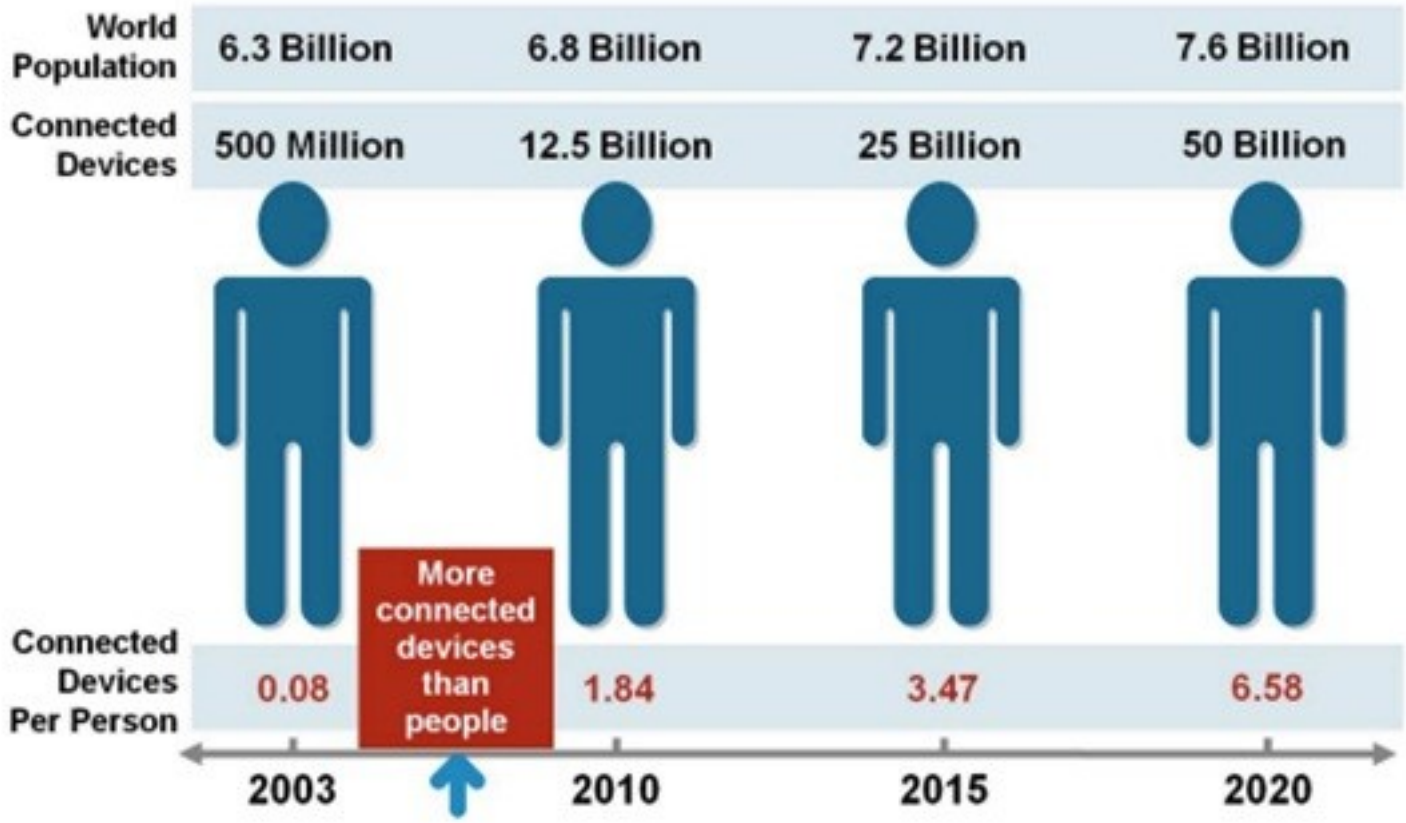
“**Grandes** dados”

Quão Grande?

Quantidade dados gerados

A quantidade de informação gerada no primeiro dia de vida de um bebê é equivalente a 70 vezes a quantidade de informação contida na biblioteca do congresso americano. (Fonte: PBS 2016)

Dispositivos conectados



O que acontece na Internet em um minuto?(2019)

Fonte: http://portalmakingof.com.br/o-que-acontece-em-1-minuto-de-internet

Quantos dados?

Google - 20 PB de dados processados por dia (2008)

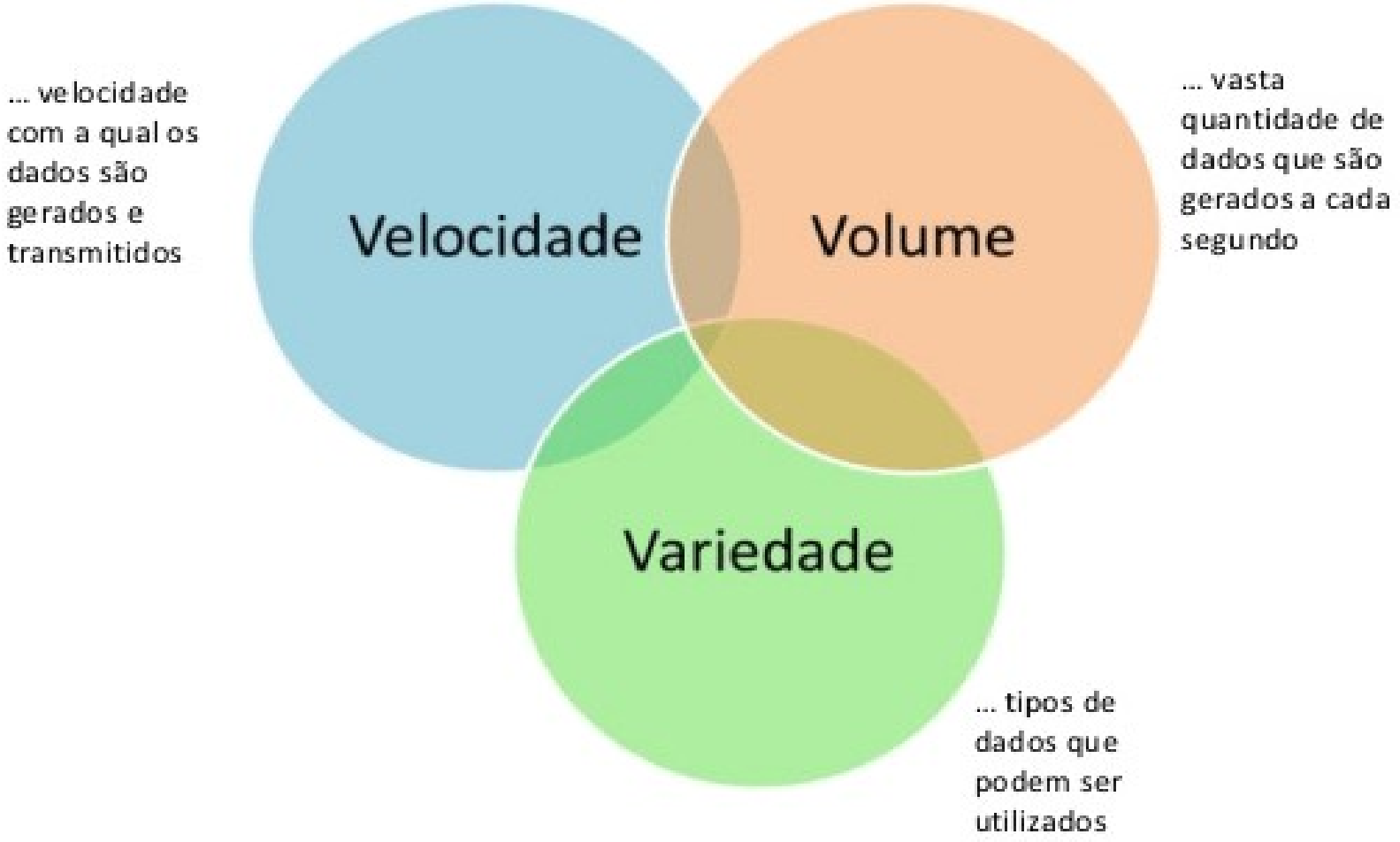
Facebook - 2.5 PB dados + 15 TB/dia (4/2009)

eBay - 6.5 PB dados + 50 TB/dia (5/2009)

Netflix 1.3 petabyte dados/dia (1.000.000 GB)

CERN’s Large Hydron Collider (LHC) - gera 15 PB/ano

3 V’s do Big Data

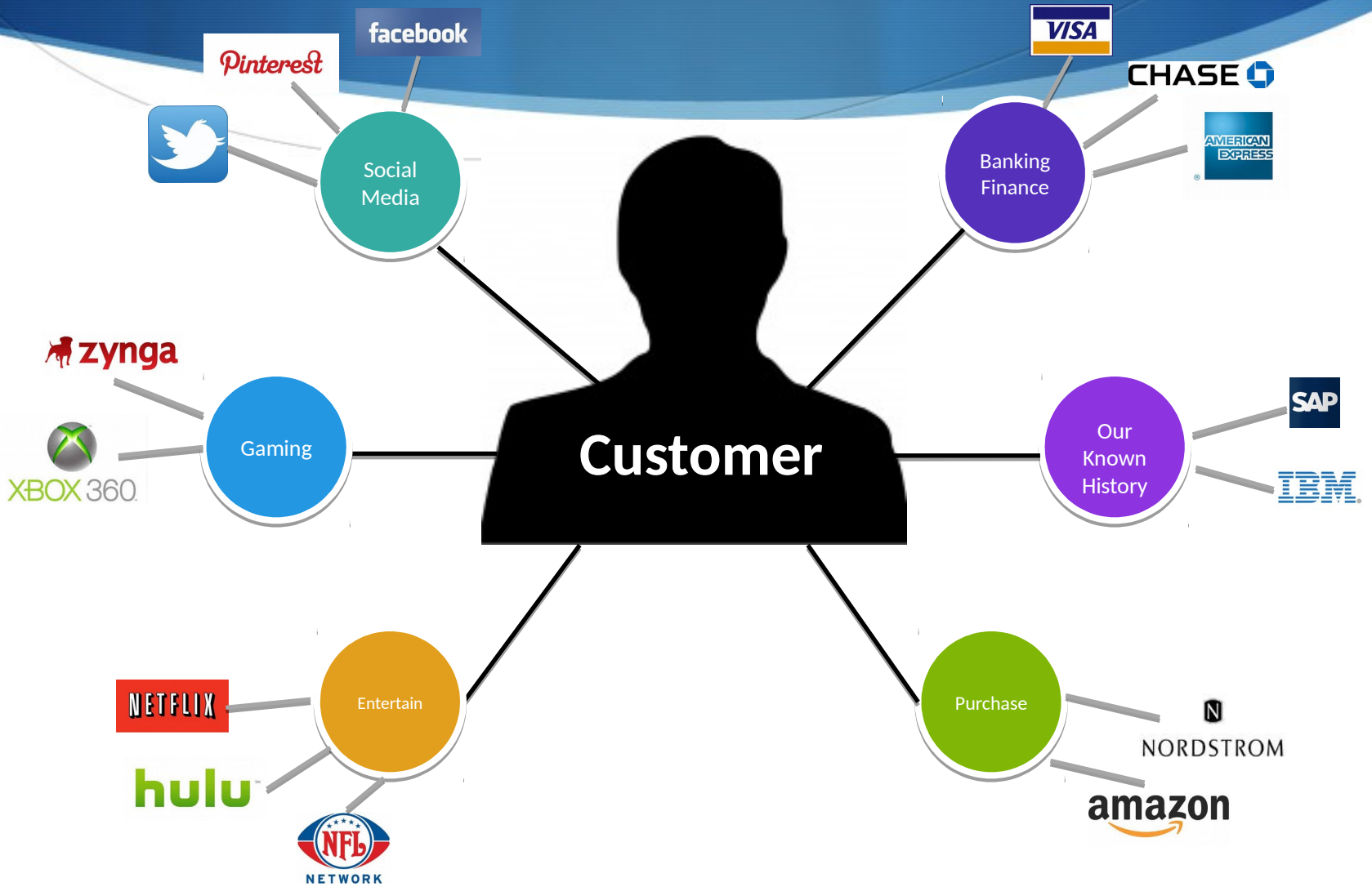


Variedade

- Sensores
- Banco de dados
- Arquivos de texto (word, txt, pdf…)
- Arquivos de imagens (fotos, gráficos…)
- Emails, blogs, páginas da web
- Videos e músicas - Streaming
- Mensagens texto (SMS, Whatsapp…)
- Mensagens de voz

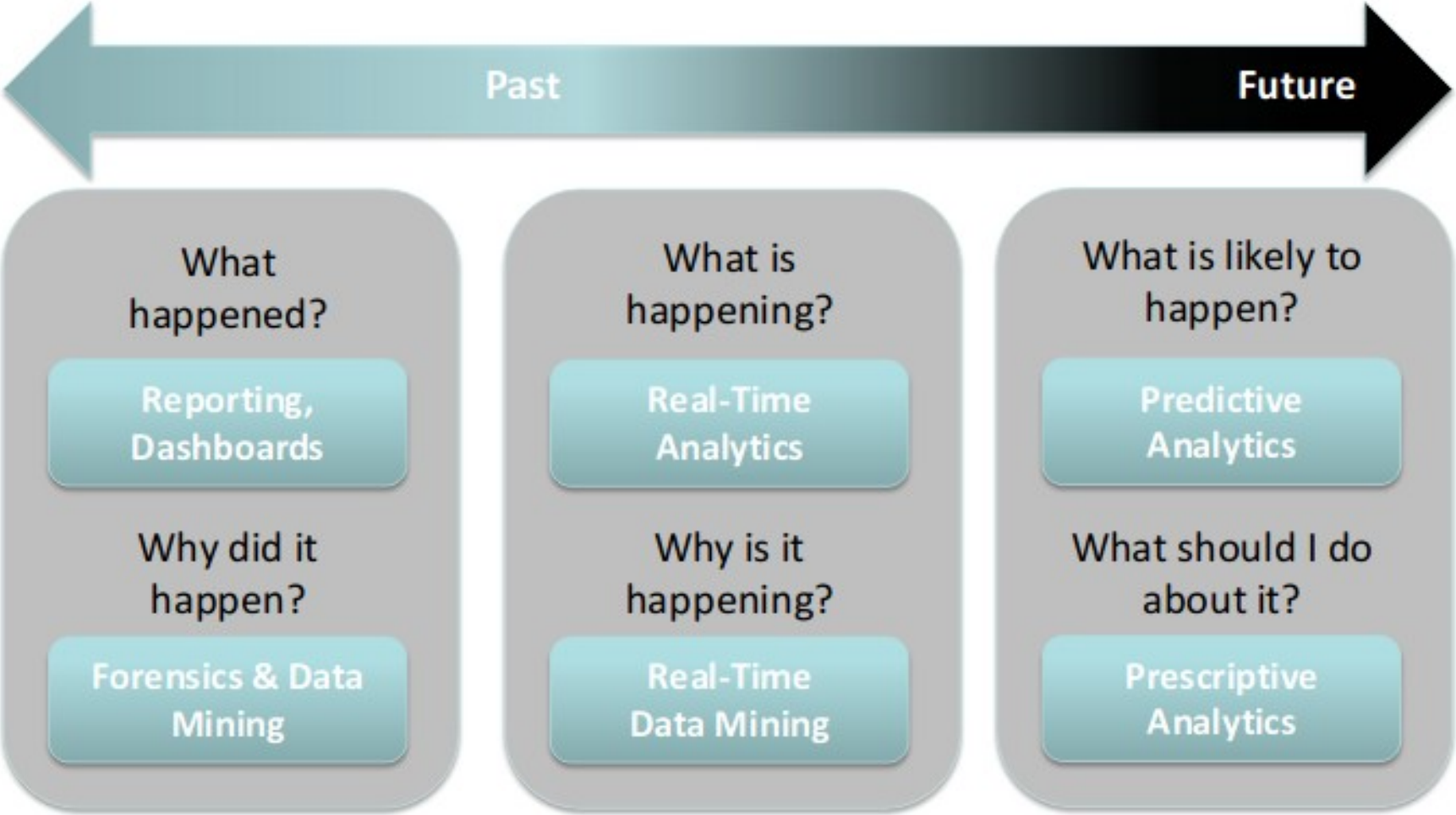
Todos os tipos de dados são importantes para gerar conhecimento

Variedade: muitas visões da mesma pessoa



Pra que serve?

Pra que serve?



BI x Big Data

Business Ingelligence (BI)

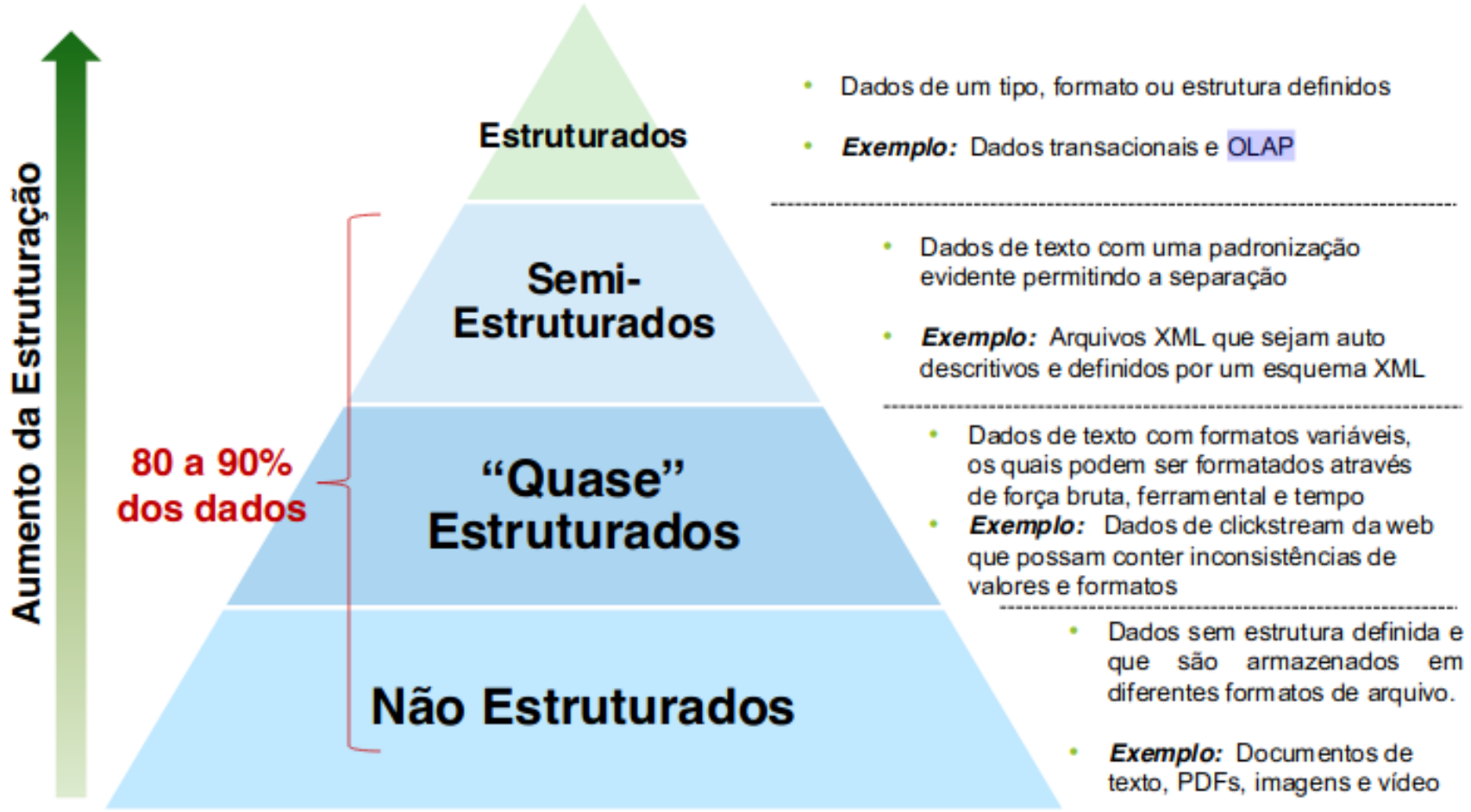
- Focado na coleta, transformação e disponibilização de dados estruturados para a tomada de decisões
- Analisa o que já existe, definindo as melhores hipóteses
- Ideal para quando já se conhece as perguntas
- Mais específico, voltado apenas para negócios

BI x Big Data

Big Data Analytics:

- Focado no processamento de dados estruturados e não estruturados, bem como nas correlações e descobertas que desse processamento podem advir
- Analisa o que já existe e o que está por vir, apontando novos caminhos
- Ideal para quando se quer explorar novas possibilidades, descobrir novos padrões e explorar perguntas que ainda não haviam sido feitas
- Mais amplo, voltado não apenas para negócios, mas para qualquer área/segmento, como saúde, entretenimento, educação

Como trabalhar com os dados



Trabalhando com dados

- Análises Estatísticas

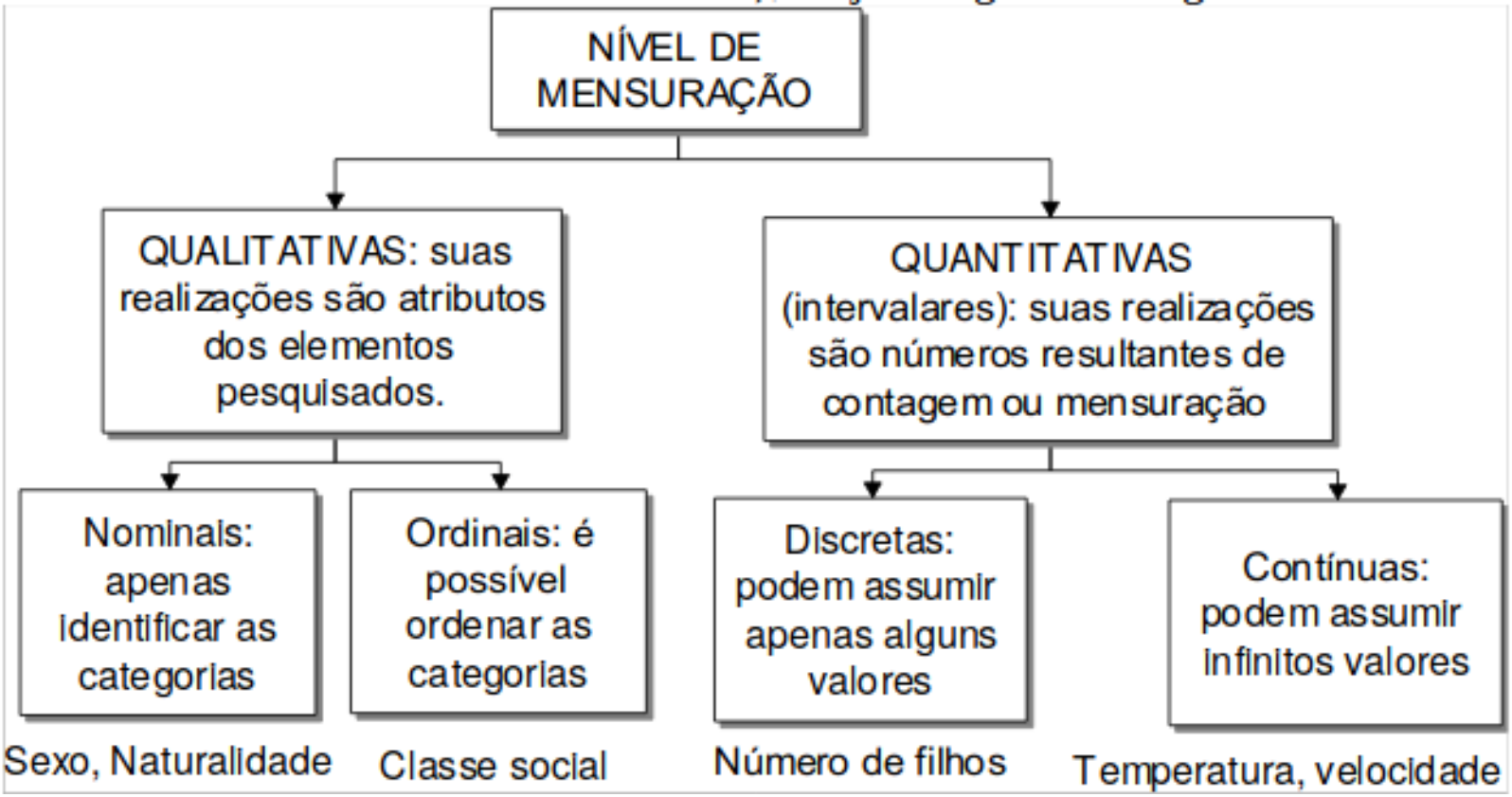
Análise Exploratória

RESUMIR E ORGANIZAR os dados coletados através de tabelas, gráficos ou medidas numéricas

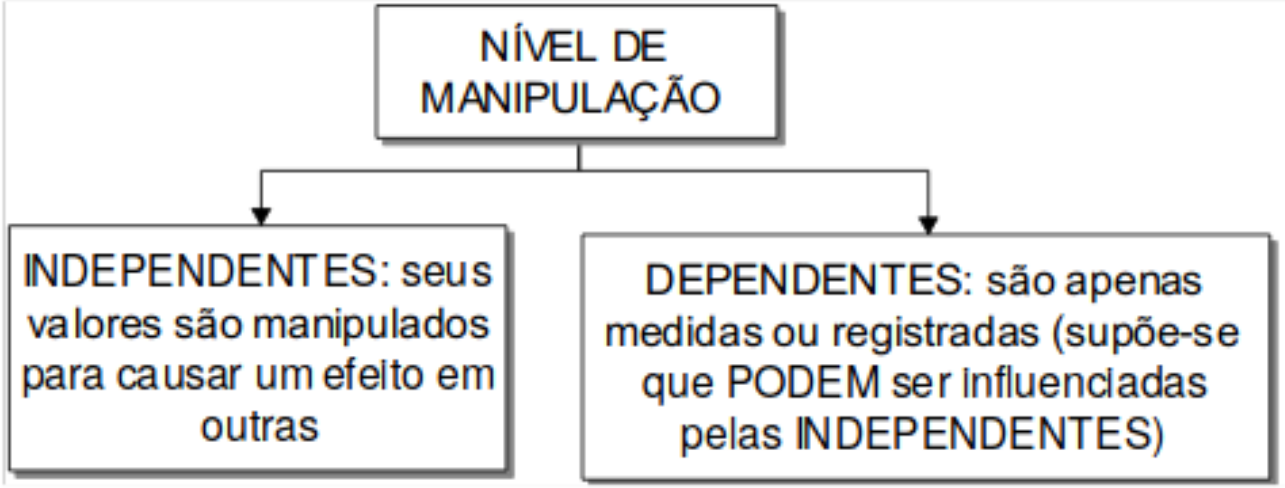
A partir dos dados resumidos, procurar regularidades (INTERPRETAR os dados)

É possível identificar se os dados seguem algum modelo conhecido, que permita estudar o fenômeno sob análise, ou se é necessário sugerir um novo modelo.

Variáveis


NÍVEL DE MENSURAÇÃO
QUALITATIVAS: suas realizações são atributos dos elementos pesquisados.
QUANTITATIVAS (intervalares): suas realizações são números resultantes de contagem ou mensuração
Nominais: apenas identificar as categorias
Ordinais: é possível ordenar as categorias
Discretas: podem assumir apenas alguns valores
Contínuas: podem assumir infinitos valores
Sexo, Naturalidade
Classe social
Número de filhos
Temperatura, velocidade

Variáveis


NÍVEL DE MANIPULAÇÃO
INDEPENDENTES: seus valores são manipulados para causar um efeito em outras
DEPENDENTES: são apenas medidas ou registradas (supõe-se que PODEM ser influenciadas pelas INDEPENDENTES)

Análise Exploratória (II)

1) Normalização dos dados

- **Dados perdidos:** dados que não foram coletados.
  - Eliminação do registro,
  - Preenchimento com base na média da variável ou por interpolação
  - Criação de uma categoria “Não respondeu”
- **Erros no registro:** valores que foram armazenados incorretamente. Pode ser por falta de uniformidade no armazenamento dos valores (qualitativos) ou valores impossíveis para a variável

Análise Exploratória (II)

1) Normalização dos dados
   - **Valores discrepantes :** estão muito acima, ou muito abaixo da maioria dos valores do conjunto de dados. Pode ser identificado usando distribuição de frequências

   - **Inconsistências:** dados deturpados (deliberadamente ou não). Precisa usar tecnicas avançadas para descobrir. Normalmente cruza com outras variáveis para identificar a discrepância

   - **Recodificação:** criar novas variáveis a partir das existentes para facilitar a sua análise individual ou o cruzamento com outra para atingir os objetivos da análise. Usada para:
     - Agrupar valores de variáveis qualitativa com muitos valores possíveis
     - Transformar variável quantitativa em qualitativa (categorizada)

Análise Exploratória (II)

1) Normalização dos dados
   - **Valores discrepantes :** estão muito acima, ou muito abaixo da maioria dos valores do conjunto de dados. Pode ser identificado usando distribuição de frequências

   - **Inconsistências:** dados deturpados (deliberadamente ou não). Precisa usar tecnicas avançadas para descobrir. Normalmente cruza com outras variáveis para identificar a discrepância

   - **Recodificação:** criar novas variáveis a partir das existentes para facilitar a sua análise individual ou o cruzamento com outra para atingir os objetivos da análise. Usada para:
     - Agrupar valores de variáveis qualitativa com muitos valores possíveis
     - Transformar variável quantitativa em qualitativa (categorizada)

Análise Exploratória (II)

2) Medidas de Posição: valor numérico que represente a tendência do conjunto: Média, Mediana, e Moda.

3) Medidas de Dispersão: Intervalo, Variância, Desvio Padrão e Coeficiente de Variação.

4) Separatrizes dividem o conjunto em um certo número de partes iguais: Quartis (4 partes), Decis (10 partes), Centis (100 partes).)

Análise Exploratória (II)

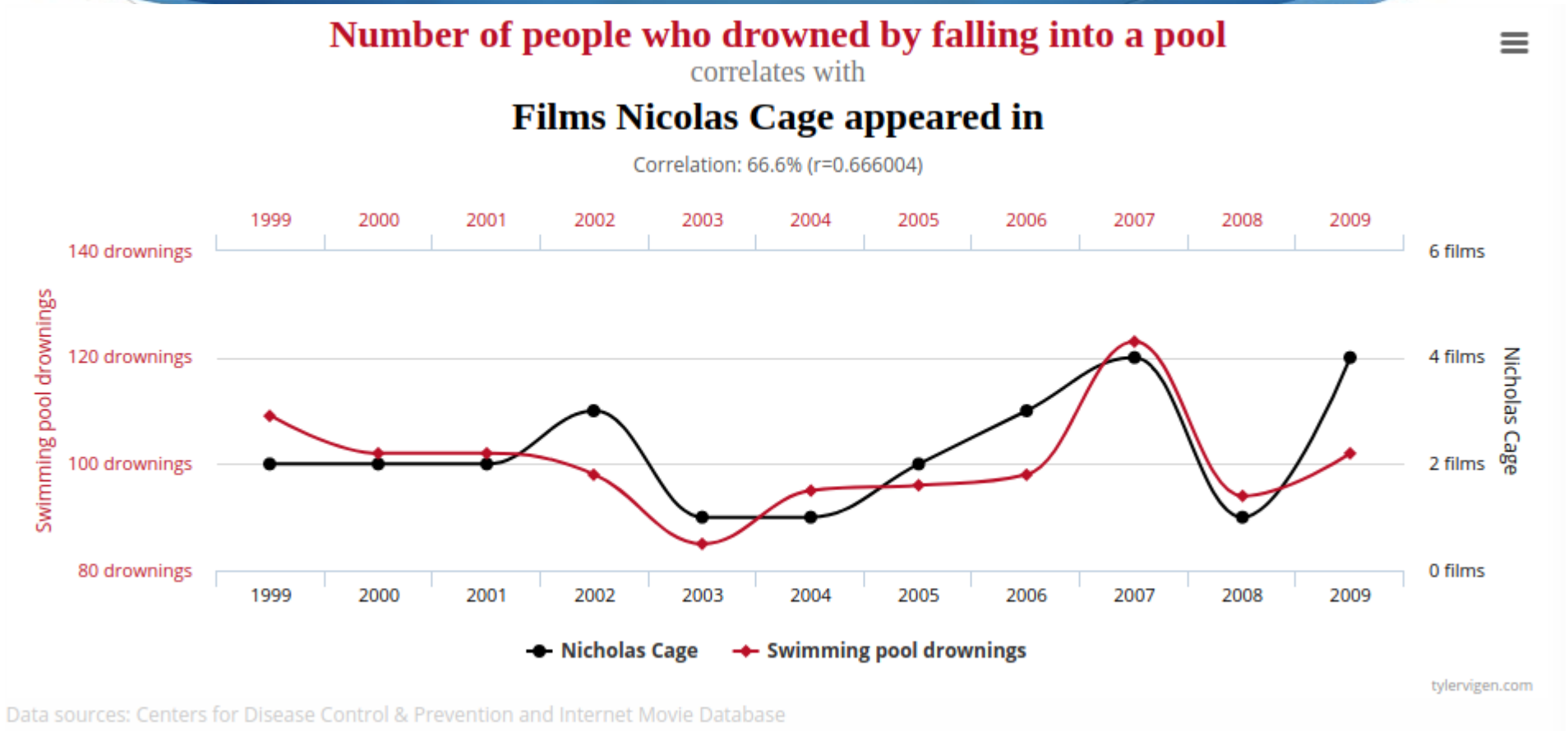
2) Distribuição das frequências: contar os valores e mostrar quantas vezes cada valor aparece em uma tabela

3) Medidas de Posição: valor numérico que represente a tendência do conjunto: Média, Mediana, e Moda.

4) Medidas de Dispersão: Intervalo, Variância, Desvio Padrão e Coeficiente de Variação.

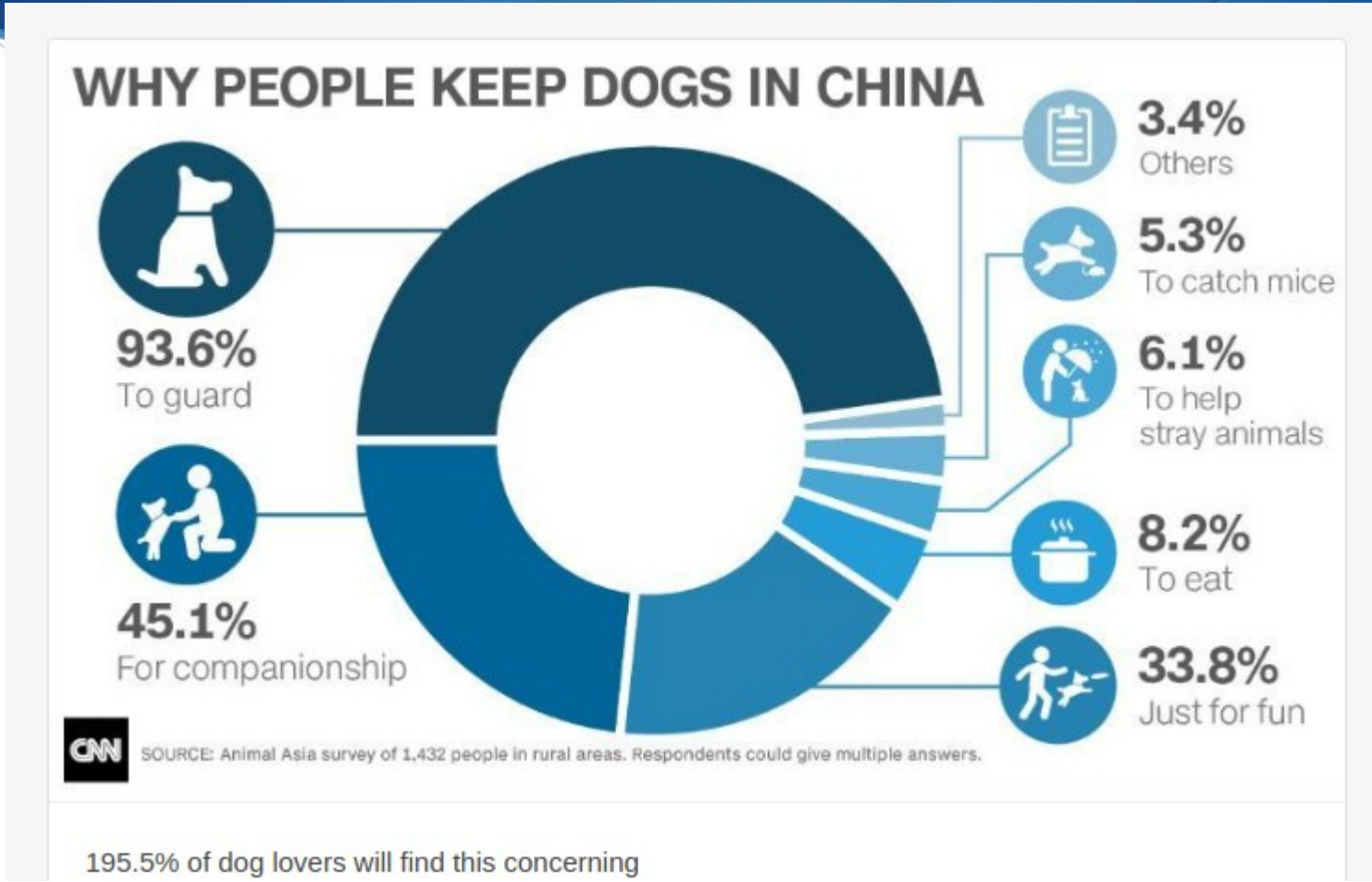
5) Separatrizes dividem o conjunto em um certo número de partes iguais: Quartis (4 partes), Decis (10 partes), Centis (100 partes).)

Correlação != Causalidade



https://www.tylervigen.com/spurious-correlations

Cuidado com os gráficos!

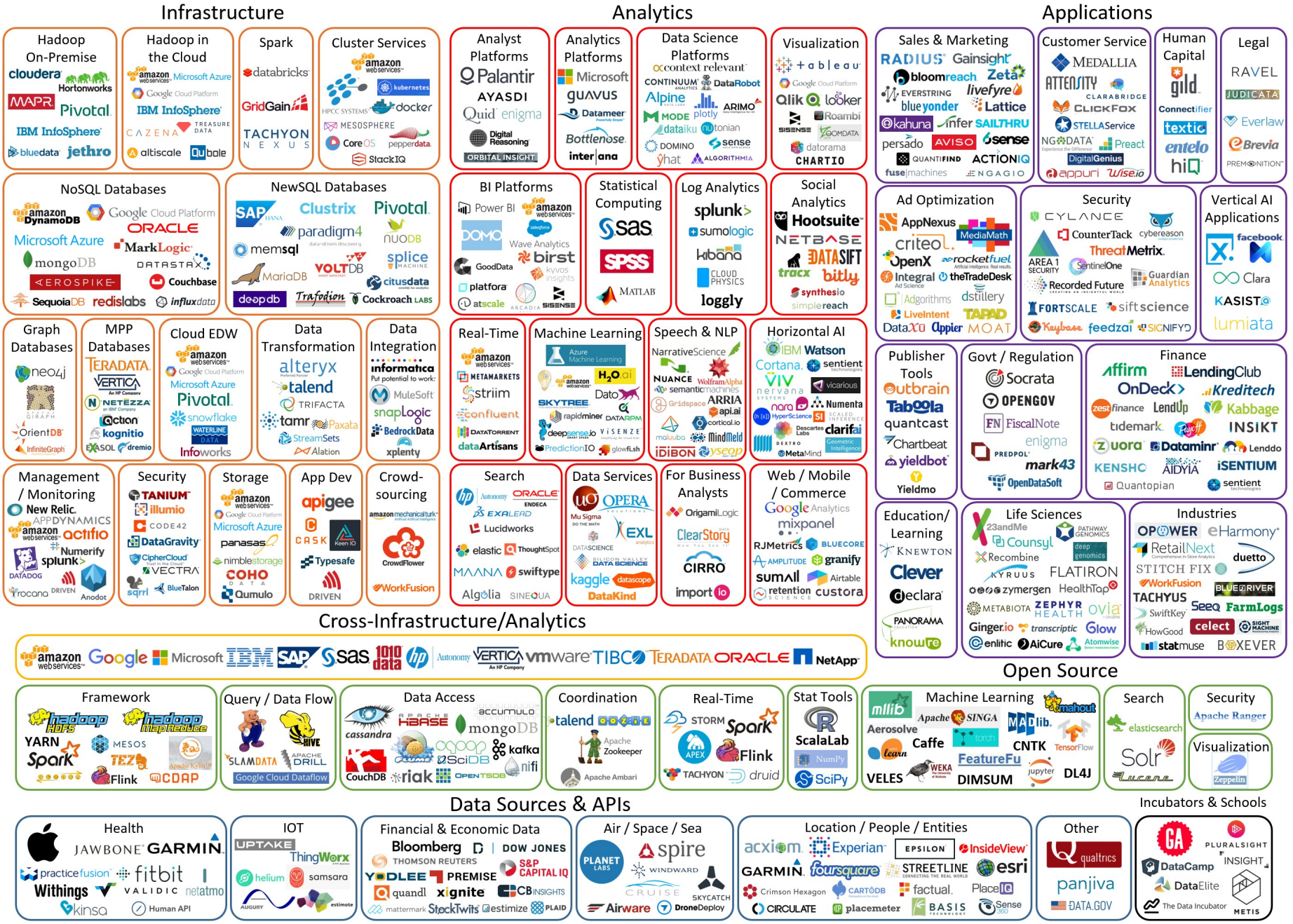


https://viz.wtf/

Ferramentas

Muitas e muitas ferramentas pra trabalhar com Big Data!!

Big Data Landscape 2016 (Version 3.0)

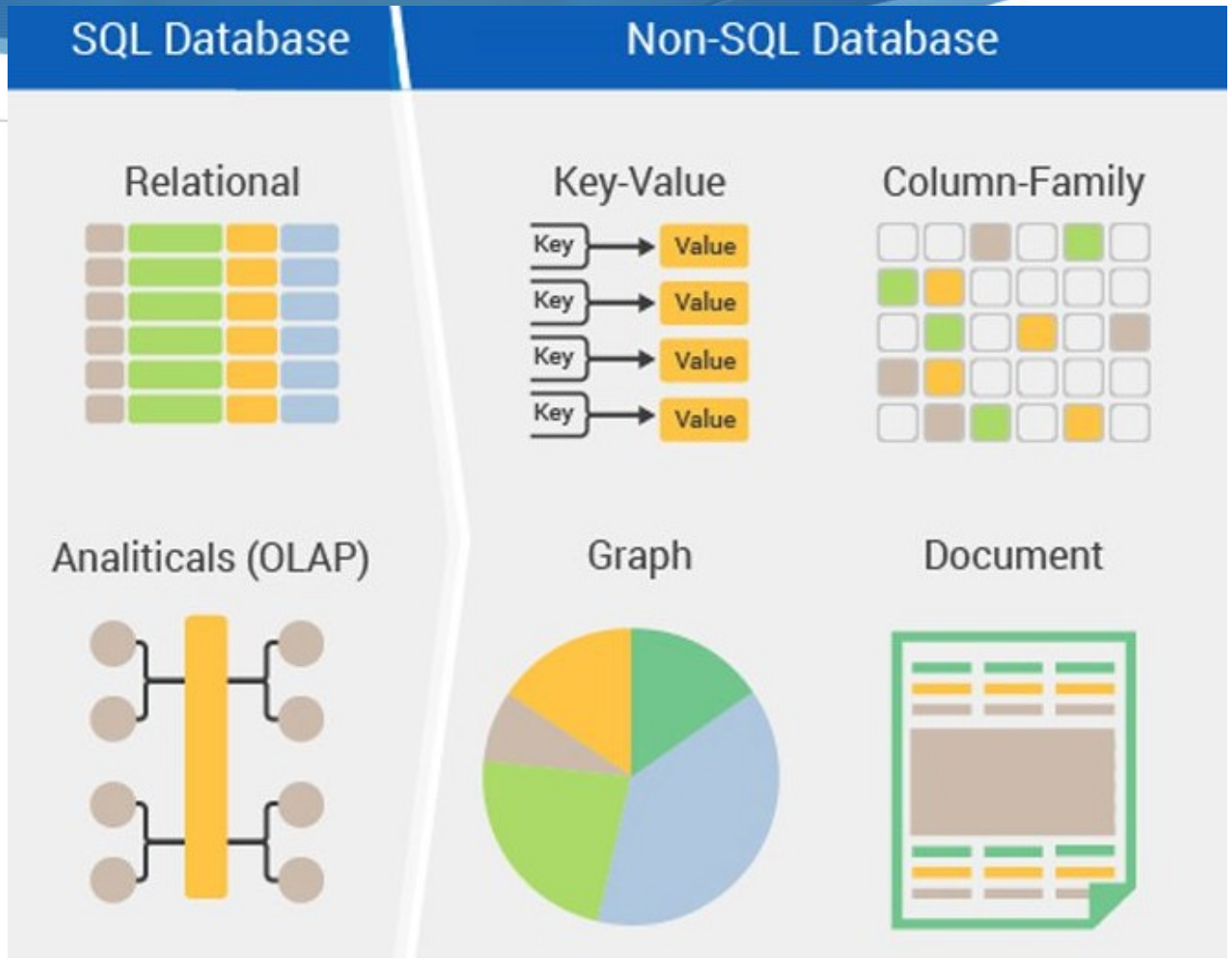


Last Updated 3/23/2016

FIRSTMARK

Armazenamento -Banco de dados -

- SQL
  - Oracle
  - PostgreSQL
  - MySQL
  - DB2
  - SQL Server
- NoSQL
  “not only SQL”
  - Cassandra
  - MongoDB
  - Aerospike



Dúvidas?

Referências
- Bases de Dados -

- Censo do Legislativo - http://dados.gov.br/dataset/censo-do-legislativo
- IBGE - https://downloads.ibge.gov.br/downloads_estatisticas.htm
- 33 Brilliant And Free Data Sources Anyone Can Use - https://www.forbes.com/sites/bernardmarr/2016/02/12/big-data-35-brilliant-and-free-data-sources-for-2016/#39b6b626b54d
- Portal da transparência – dados do governo federal

Referências

- Pizza de dados - https://podcast.pizzadedados.com/ - Spotify
- Serenata de amor - https://serenata.ai/
- Olhômetro Pioneiro - http://especiais-pio.clicrbs.com.br/olhometro/index.html
- CCs governo municipal - http://especiais-pio.clicrbs.com.br/ccsdeguerra/index.html
- Mapa da transparência – RS - http://www.mapa.rs.gov.br/
- Data with story - http://datawithstory.com/o-guia-do-cientista-de-dados/