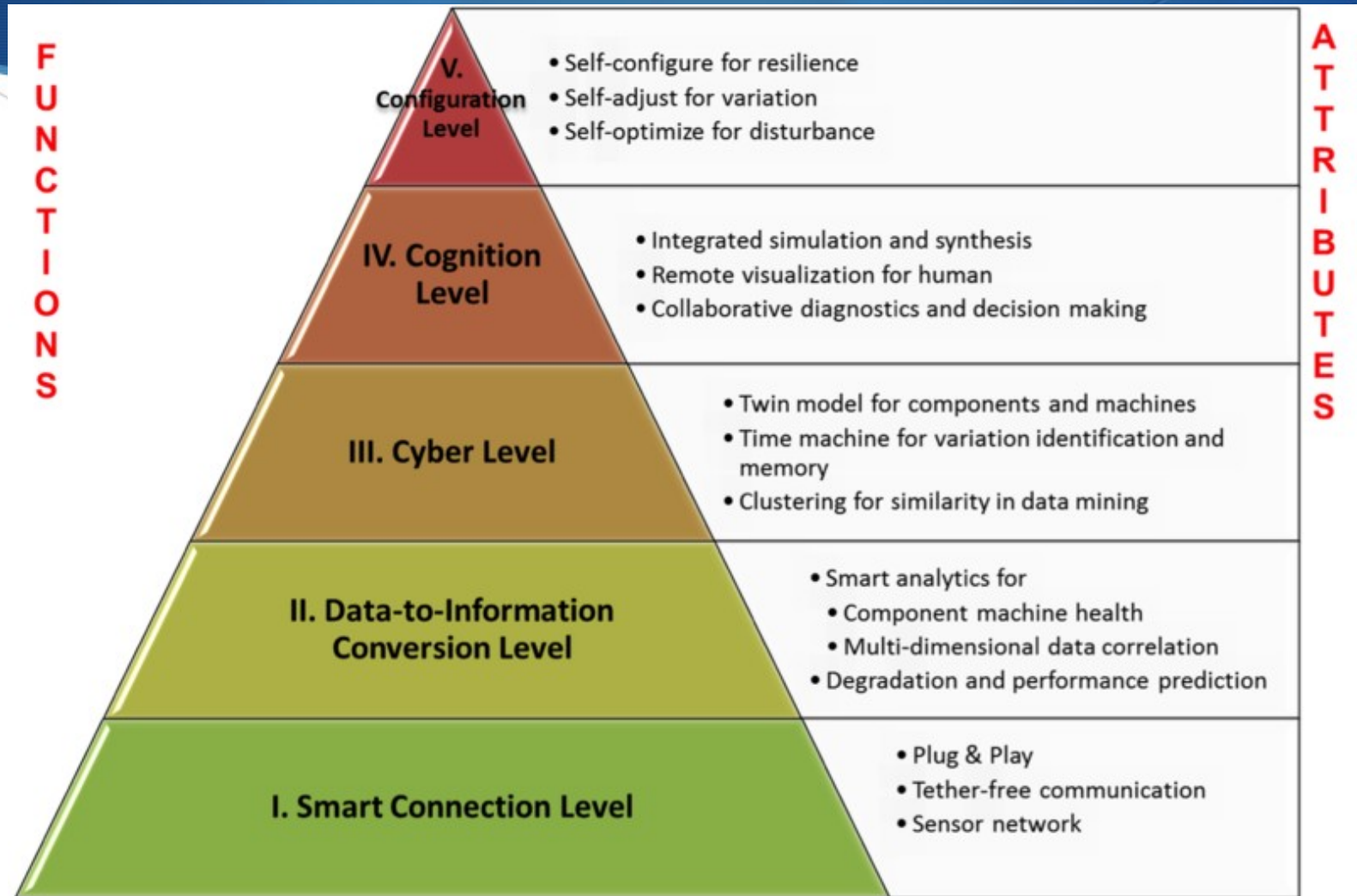


Big Data

Daniela Maria Uez
dani.uez@gmail.com

11 JULHO 2019

Arquitetura 5C

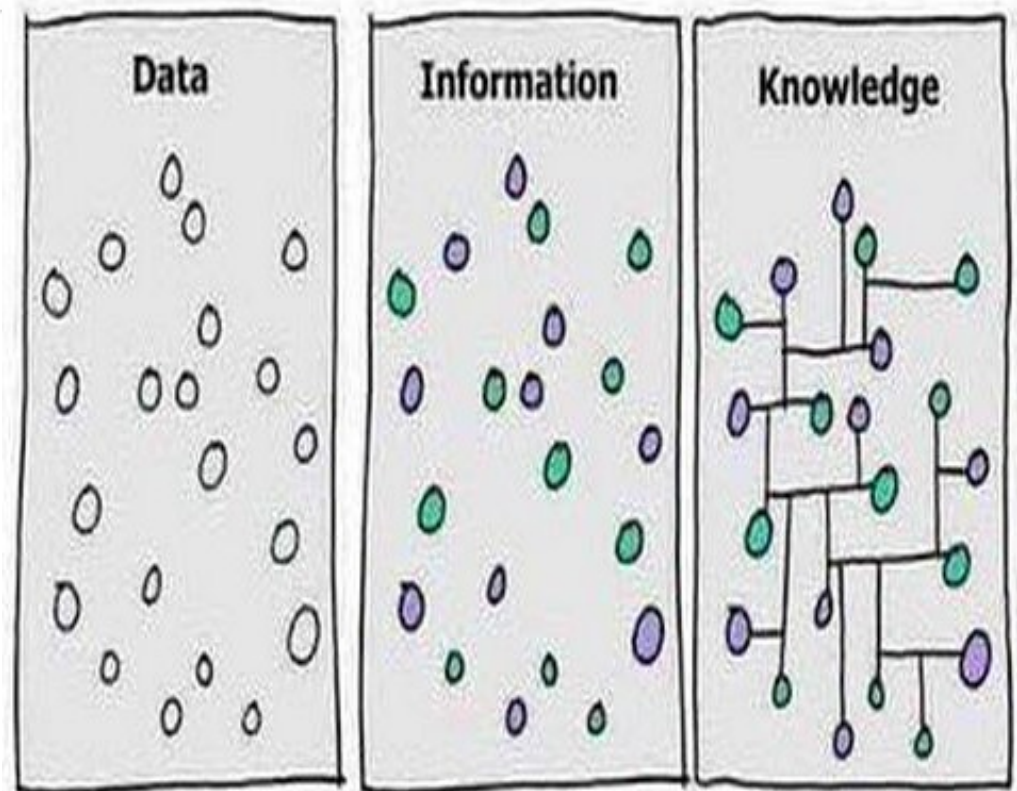


Dados x Informação x Conhecimento

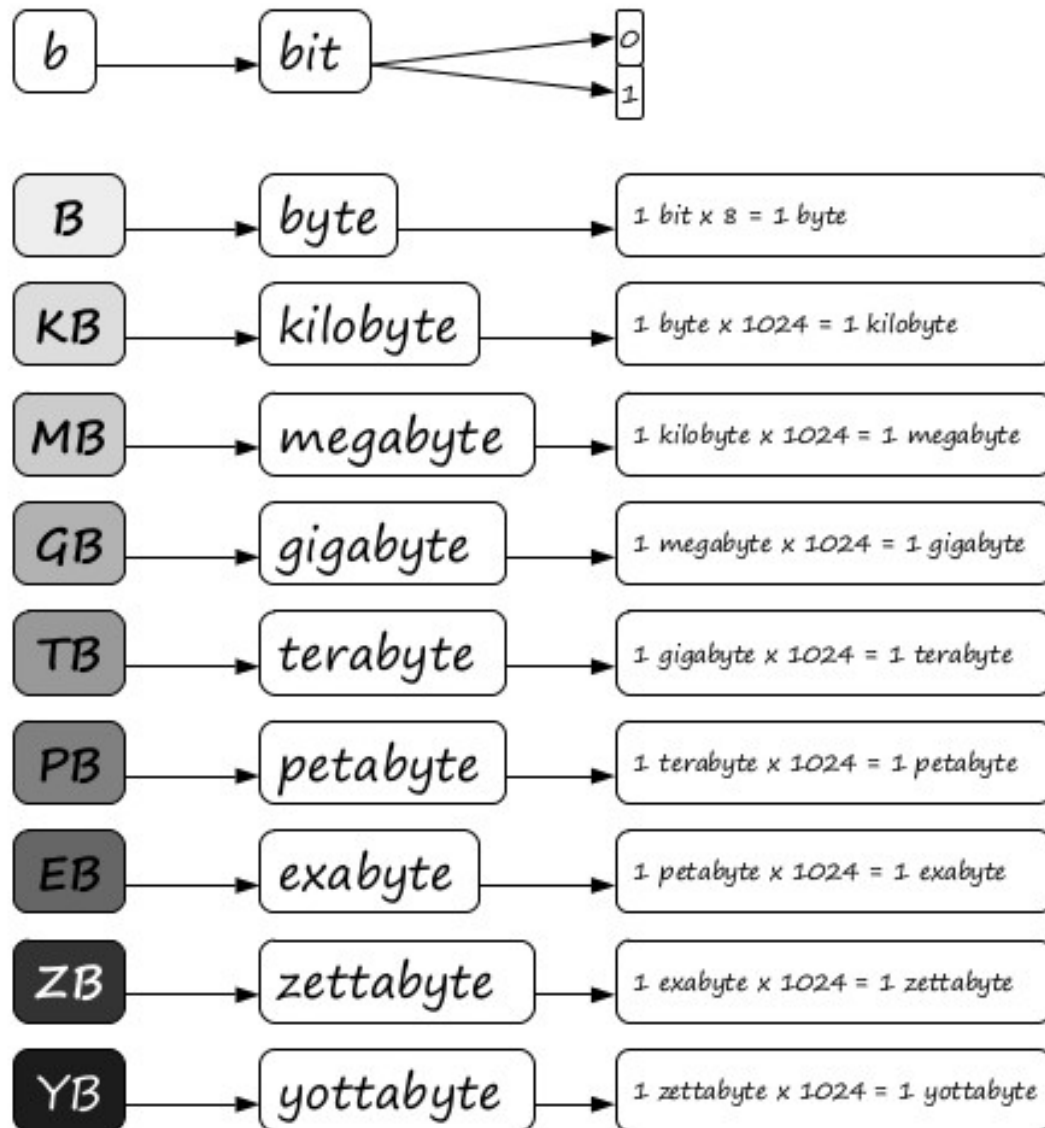
Dado: observação sobre o estado do mundo

Informação: dado dotado de relevância e propósito

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



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”

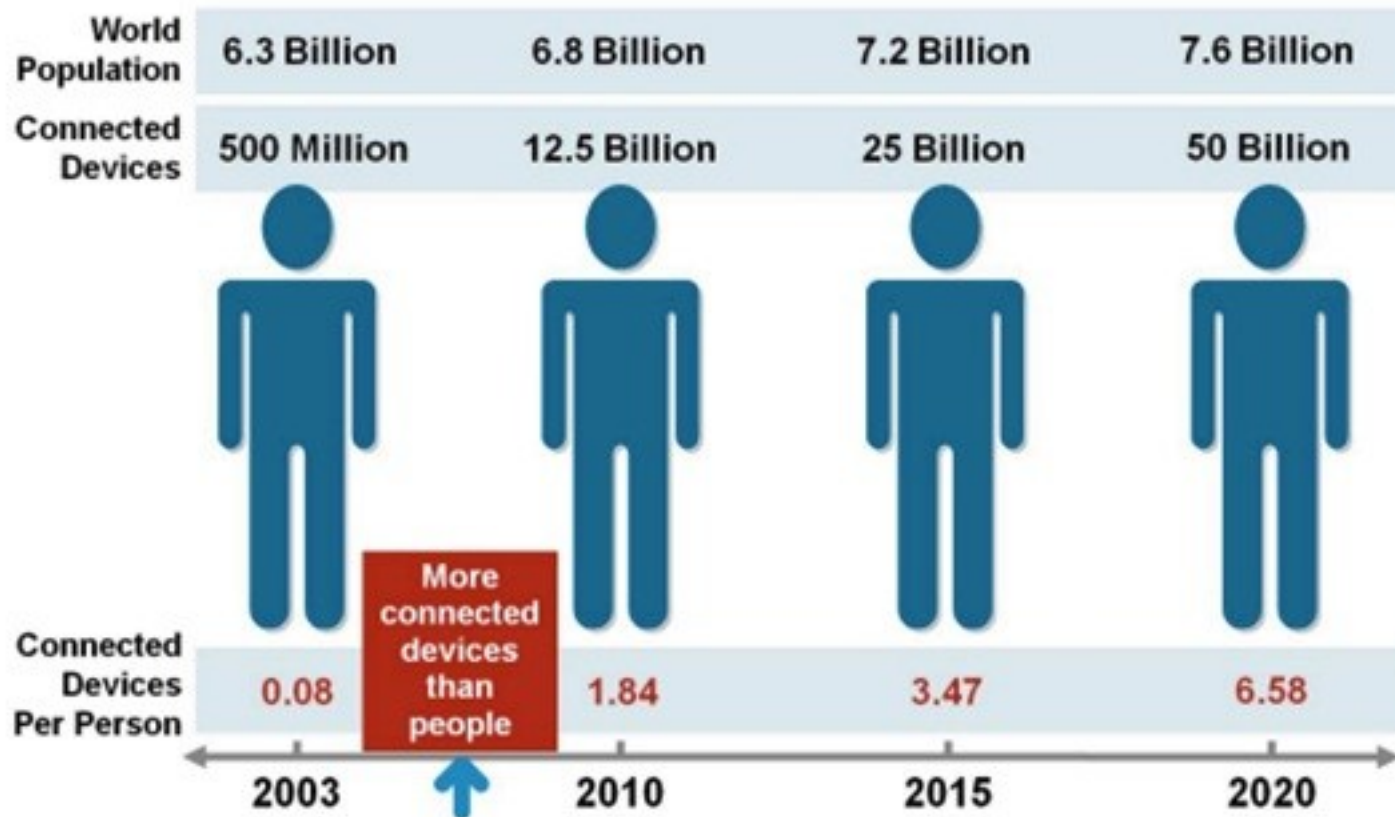
Quão Grande?

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)



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

... velocidade com a qual os dados são gerados e transmitidos

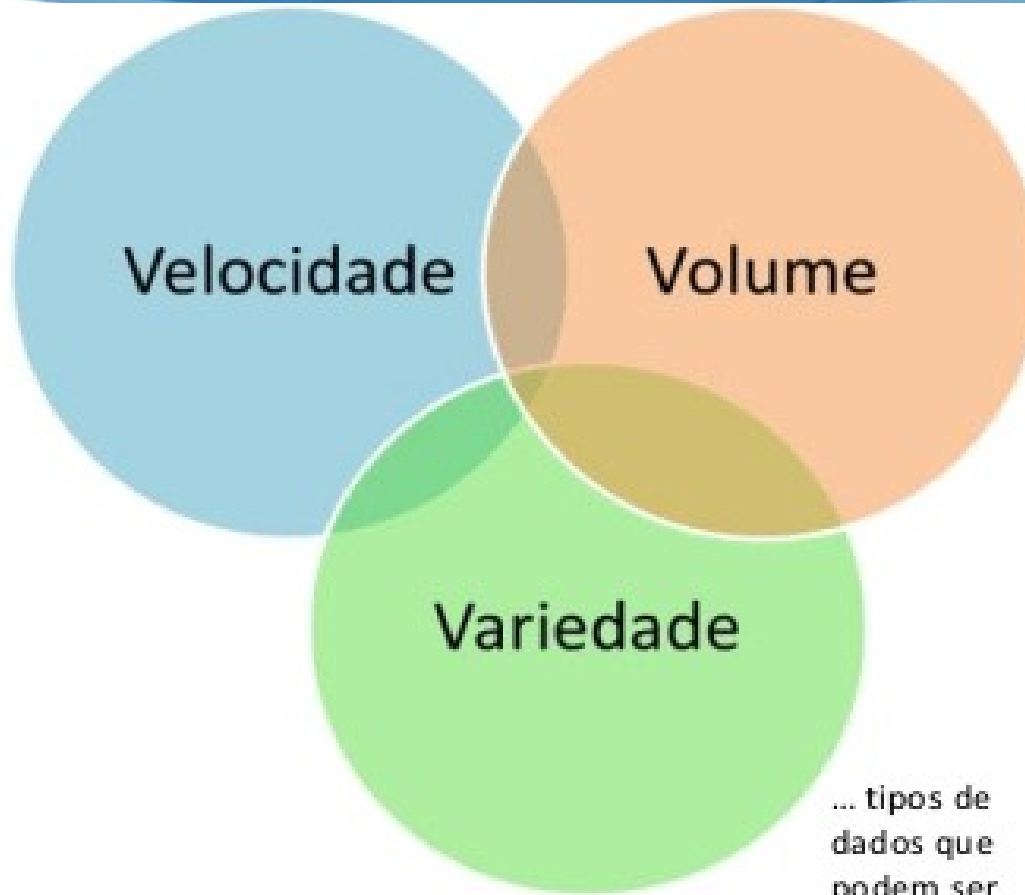
Velocidade

... vasta quantidade de dados que são gerados a cada segundo

Volume

Variedade

... tipos de dados que podem ser utilizados

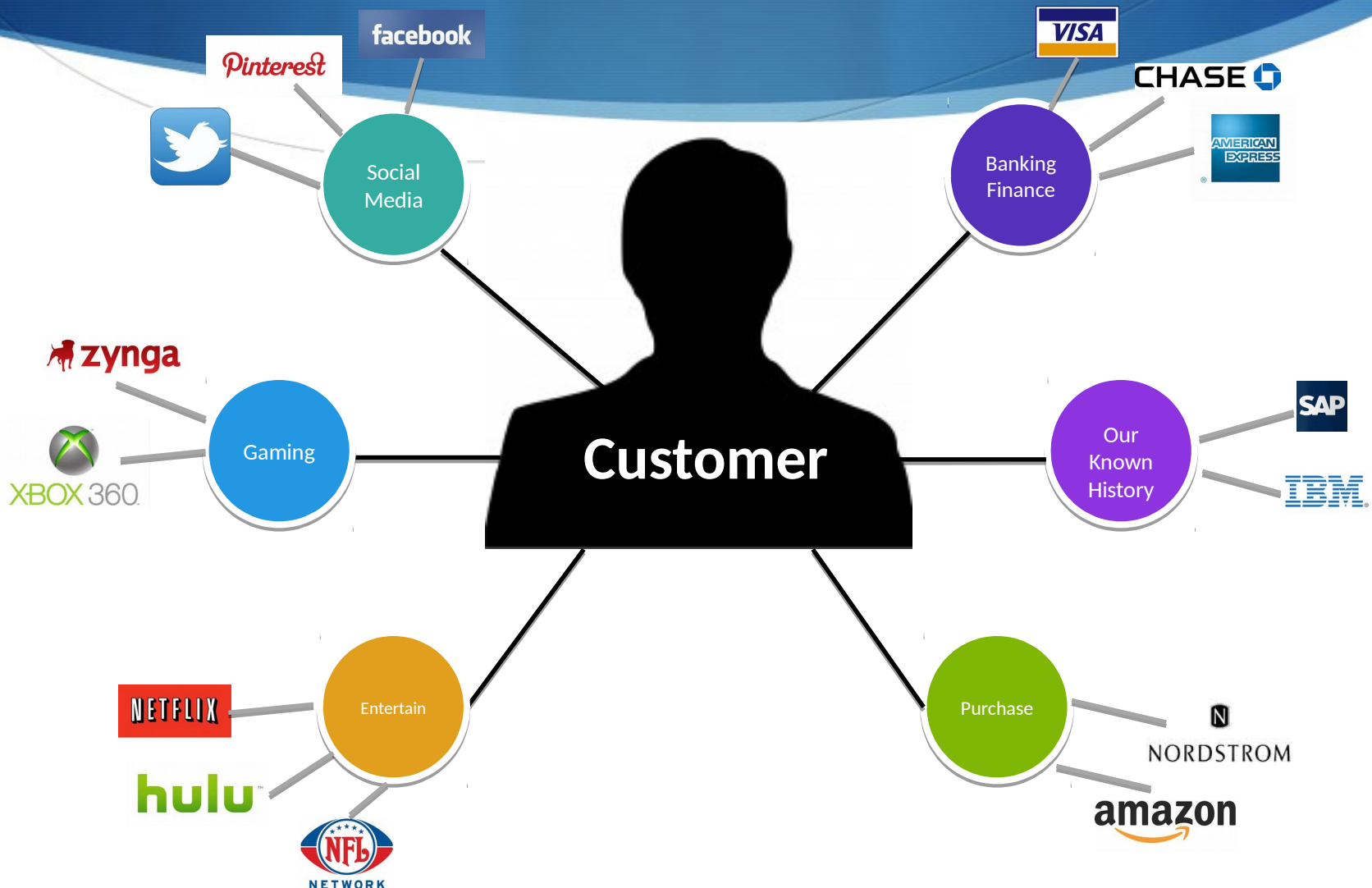


Variedade

- 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

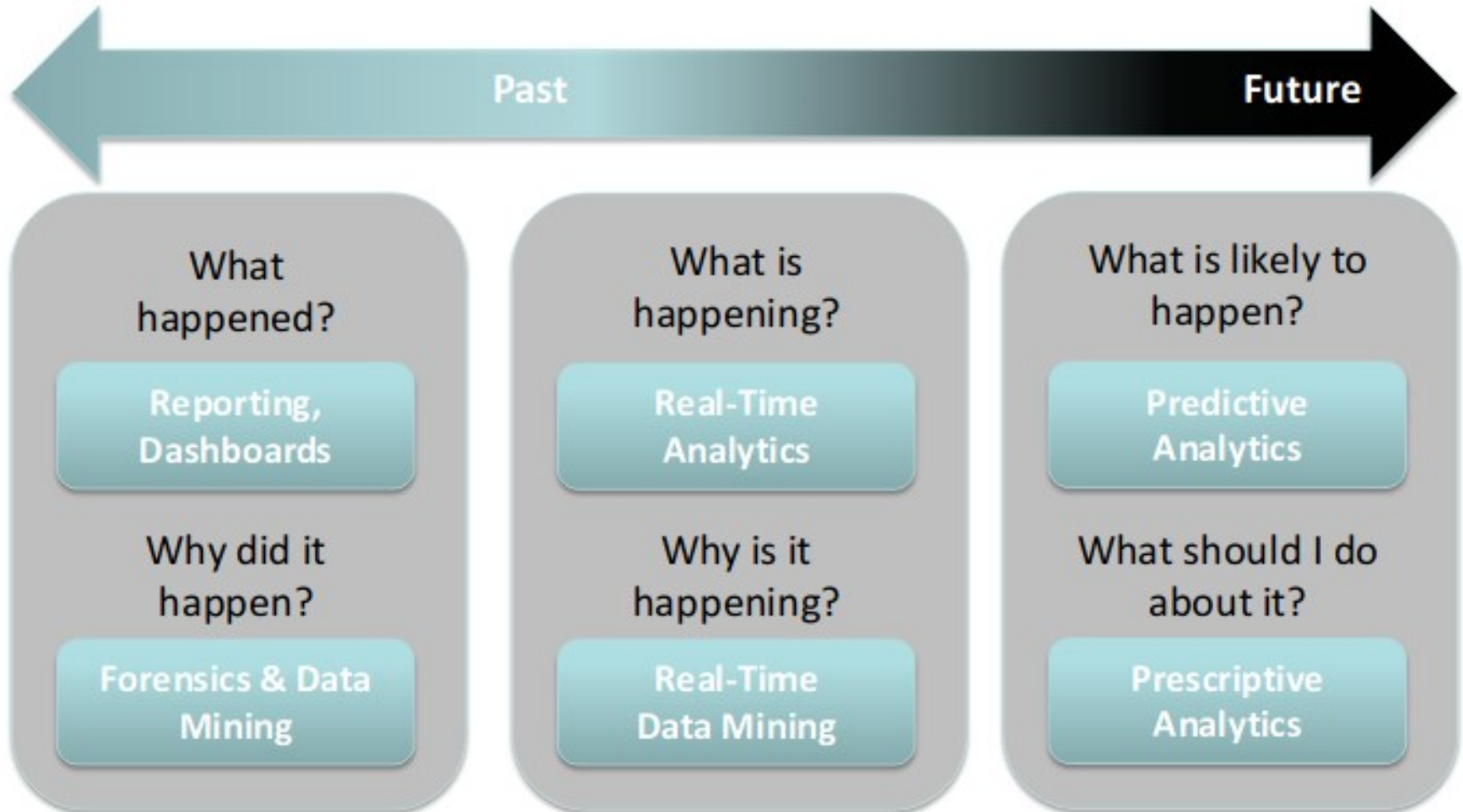


Pra que serve?

MATERIAIS + MÉTODOS = RESULTADOS



Pra que serve?



BI x Big Data

Business Intelligence (BI)

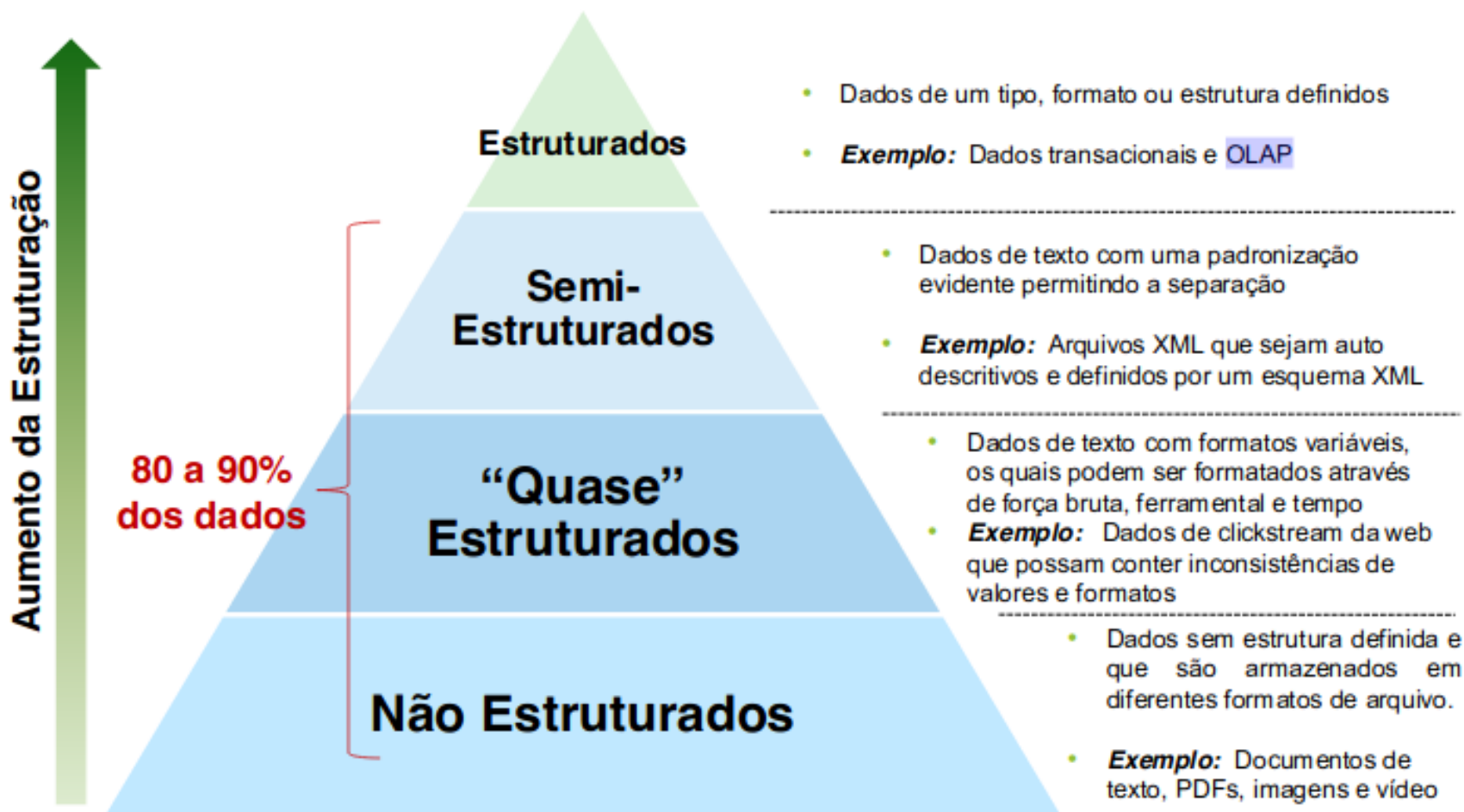
- Focado na coleta, transformação e disponibilização de dados estruturados para a tomada de decisões
- Analisa o que já existe, definindo as melhores hipóteses
- Ideal para quando já se conhece as perguntas
- Mais específico, voltado apenas para negócios

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

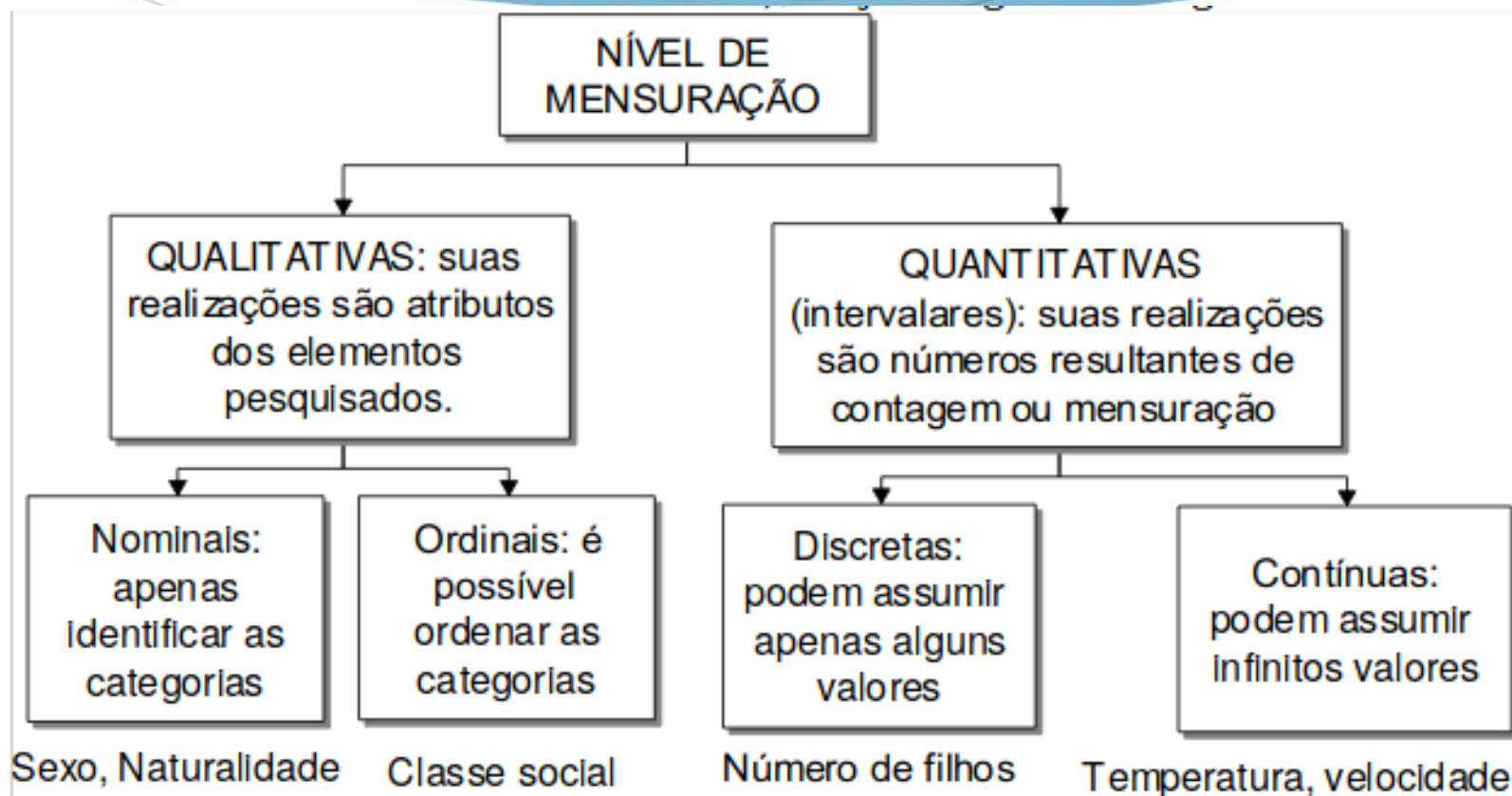
Análise Exploratória

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

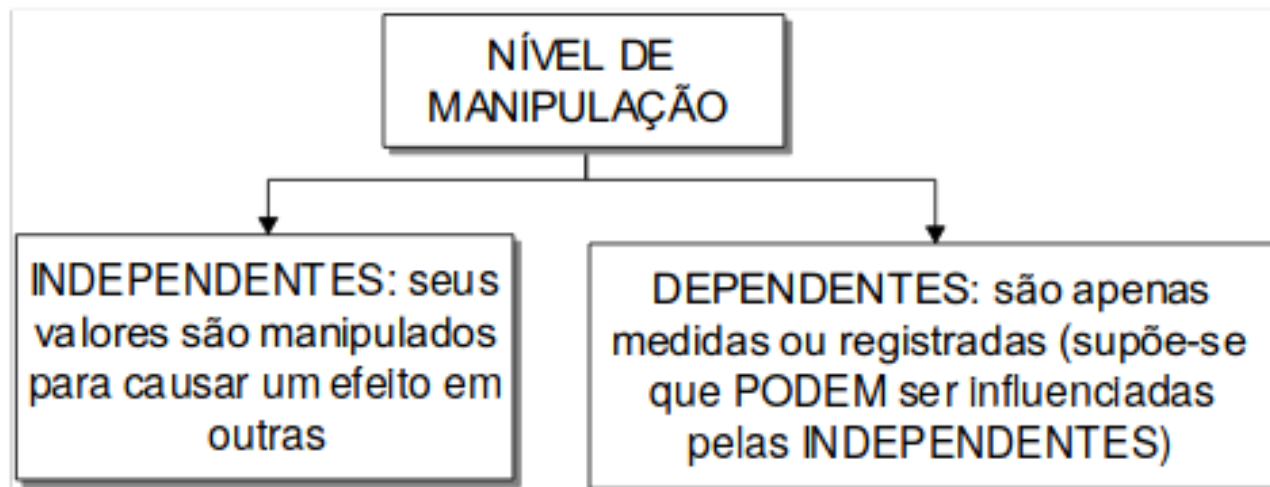
A partir dos dados resumidos, procurar regularidades (INTERPRETAR os dados)

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



Variáveis



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 técnicas 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 técnicas 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)

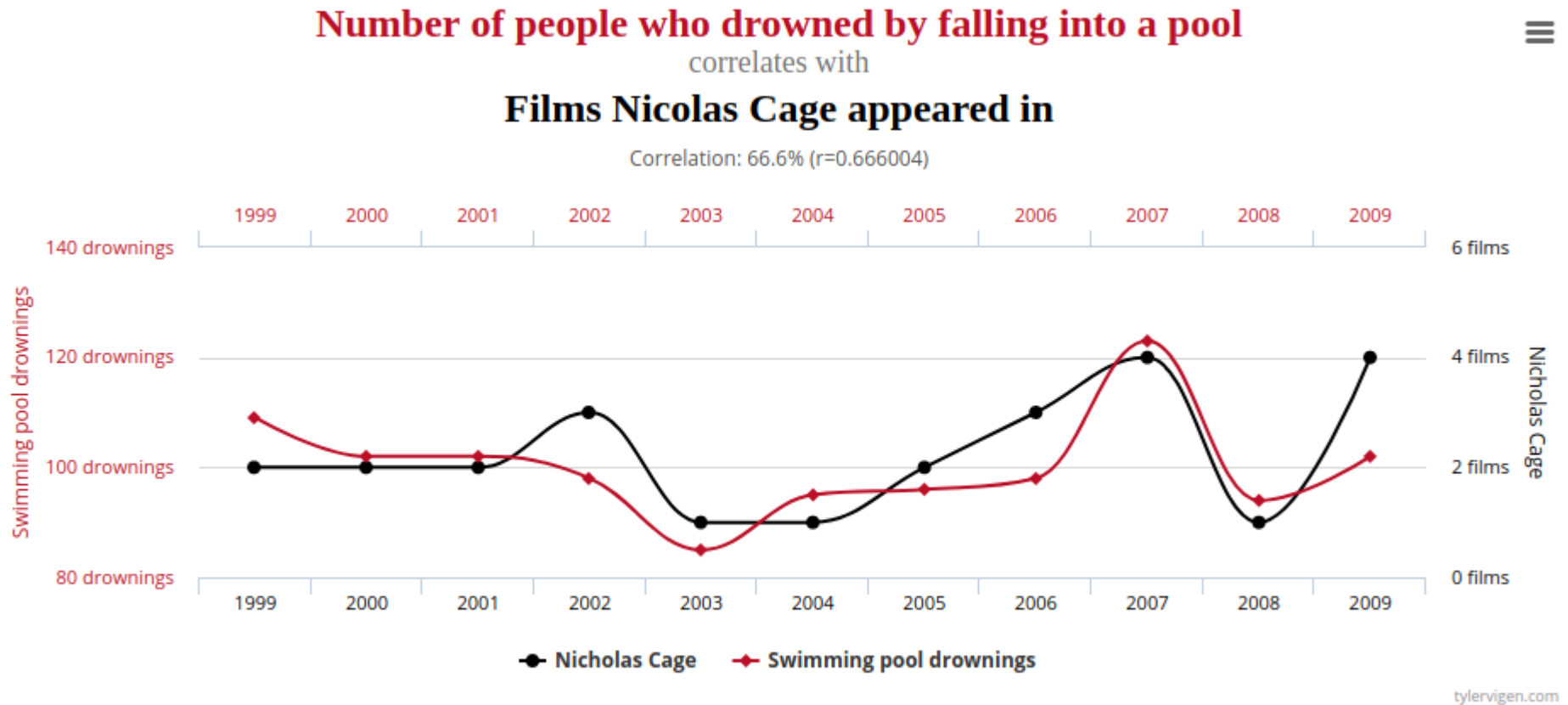
Análise Exploratória (II)

- 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



Data sources: Centers for Disease Control & Prevention and Internet Movie Database

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

Cuidado com os gráficos!

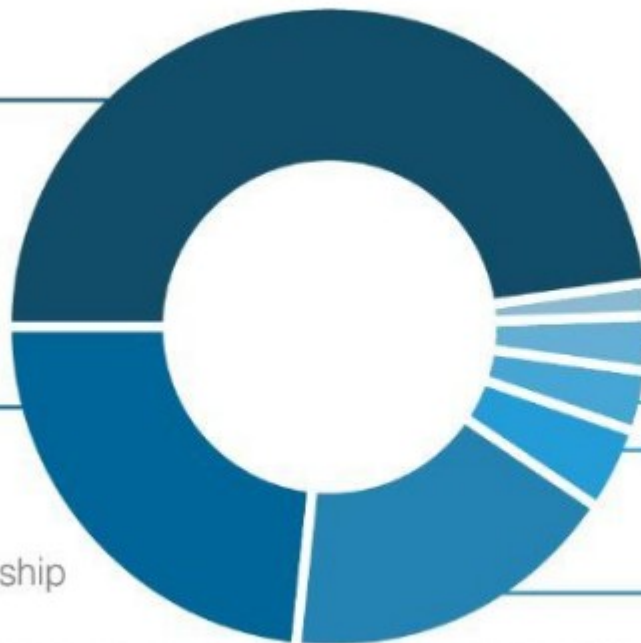
WHY PEOPLE KEEP DOGS IN CHINA



93.6%
To guard



45.1%
For companionship



3.4%
Others



5.3%
To catch mice



6.1%
To help
stray animals



8.2%
To eat



33.8%
Just for fun



SOURCE: Animal Asia survey of 1,432 people in rural areas. Respondents could give multiple answers.

195.5% of dog lovers will find this concerning

<https://viz.wtf/>

Ferramentas

Muitas e muitas ferramentas pra trabalhar com Big Data!!

Big Data Landscape 2016 (Version 3.0)

Infrastructure

Hadoop On-Premise
cloudera, Hortonworks, MAPR, Pivotal, IBM InfoSphere, bluedata, jethro

Hadoop in the Cloud
amazon, Microsoft Azure, Google Cloud Platform, IBM InfoSphere, CAZENA, altiscale, quable

Spark
databricks, GridGain, TACHYON NEXUS

Cluster Services
amazon, kubernetes, HPCC SYSTEMS, MESOSPHERE, CoreOS, pepperdata, StackIQ

Analytics

Analyst Platforms
Palantir, AYASDI, Quid, enigma, Digital Reasoning, ORBITAL INSIGHT

Analytics Platforms
Microsoft, guavus, Datarameer, Bottlenose, interana

Data Science Platforms
context relevant, DataRobot, CONTINUUM, Alpine, MODE, dataiku, DOMINO, yhat, ARIMO, ntonian, sense, ALGORITHMIA

Visualization
tableau, Google Cloud Platform, Qlik, looker, Roambi, BISSENSE, QOMDATA, datarama, CHARTIO

Applications

Sales & Marketing
RADIUS, Gainsight, bloomreach, Zeta, EVERSTRING, livefyre, blueyonder, Lattice, kahuna, infer, SAILTHRU, persado, AVISO, sense, QUANTIFIND, ACTIONIQ, fuse/machines, ENGAGIO

Customer Service
MEDALLIA, ATTENTIVITY, CLARABRIDGE, CLICKFOX, STELLAService, NGDATA, Preact, DigitalGenius, appuri, Wiseio

Human Capital
gild, Connectifier, textic, entelo, hiQ, RAVEL, JUDICATA, Everlaw, Brevia, PREMIATION

Legal

NoSQL Databases
amazon, DynamoDB, Google Cloud Platform, Microsoft Azure, ORACLE, mongoDB, MarkLogic, DATASIX, Couchbase, KEROPIKE, SequoiaDB, redislabs, influxdata

NewsSQL Databases
SAP, Clustrix, Pivotal, paradigm4, memsql, nuODB, splice, MariaDB, VOLTD, citusdata, deepdb, Trafalgar, Cockroach LABS

BI Platforms
Power BI, amazon, Wave Analytics, DOMO, GoodData, birst, kyvos insights, platforma, atscale, ACADA, BISSENSE

Statistical Computing
sas, SPSS, MATLAB

Log Analytics
splunk, sumologic, kibana, CLOUD PHYSICS, loggly

Social Analytics
Hootsuite, NETBASE, DATASIFT, truck, bitly, synthetio, simplereach

Ad Optimization
AppNexus, MediaMath, critico, rocketfuel, OpenX, theTradeDesk, Integral, Ad Science, Algorithms, dstillery, LiveIntent, TAPAD, DataXu, Appier, MOAT

Security
CYCLANCE, CounterTack, cyberreason, AREA 1 SECURITY, SentinelOne, Recorded Future, Guardian Analytics, FORTSCALE, sift science, Keybase, feedzai, SIGNIFYD

Vertical AI Applications
facebook, Clara, KASIST, lumia

Graph Databases
neo4j, OrientDB, InfiniteGraph

MPP Databases
TERADATA, VERTICA, Netezza, Acton, Kognitio, SASOL, dremio

Cloud EDW
amazon, Google Cloud Platform, Microsoft Azure, Pivotal, snowflake, WATERLINE DATA, Infoworks

Data Transformation
alteryx, talend, TRIFACTA, tamr, StreamSets, Alation

Data Integration
informatica, MuleSoft, snaplogic, BedrockData, xplenty

Real-Time
amazon, METAMARKETS, striim, confluent, DATATOPIC, dataArtisans

Machine Learning
Azure Machine Learning, H2O, amazon, SKYTREE, rapidminer, DATAFORM, deepsense, VIZEN, PredictionIO, glowfish

Speech & NLP
NarrativeScience, NUANCE, WolframAlpha, semantic machines, ARRIA, apiai, corticalio, maluba, MindMeld, IDIBON, VESQ

Horizontal AI
IBM Watson, Cortana, sentient, viv, nermana, nora, Numenta, HyperScience, SI, Decartes Labs, clarifai, MetaMind

Publisher Tools
Outbrain, Taboola, quantcast, Chartbeat, yieldbot, Yieldmo

Govt / Regulation
Socrata, OPENGOV, FN, FiscalNote, enigma, PREPOLL, mark43, OpenDataSoft

Finance
affirm, LendingClub, OnDeck, Kreditech, res finance, LendUp, Kabbage, tidemark, Fyfi, INSIGHT, ZUORA, Dataminr, Lenddo, KENSHO, AIDYA, ISENTIENT, Quantopian, sentient

Management / Monitoring
New Relic, APPDYNAMICS, amazon, acinfo, splunk, DATADOG, DRIVEN, Yrroana, Anodot

Security
TANUM, Illumio, CODE42, DataGravity, CipherCloud, VECTRA, sqrl, BlueTalon

Storage
amazon, Google Cloud Platform, Microsoft Azure, panasas, nimblestorage, COHO, Qumulo

App Dev
apigee, CASK, Typesafe, DRIVEN

Crowd-sourcing
amazon, mechanicalturk, CrowdFlower, WorkFusion

Search
hp, Oracle, ENCEA, EXALEAD, Lucidworks, elastic, ThoughtSpot, MAANA, swifttype, Algolia, SINEQUA

Data Services
UO OPERA, MU Sigma, EXL, DATA SCIENCE, kaggle, dataSCOPE, DataKind

For Business Analysts
OrigamiLogic, ClearStory, CIRRO, import io

Web / Mobile / Commerce
Google Analytics, mixpanel, RJMetrics, BLUECORE, AMPLITUDE, granify, sumall, Airtale, retention, custora

Education / Learning
KNEWTON, Clever, Declara, PANORAMA, knowtre

Life Sciences
23andMe, Counsyl, RECOMBINE, KYRUS, FLATIRON, zymogen, HealthTap, METABIOTA, ZEPHYR, HEALTH, ovia, Ginger.io, transcriptic, Glow, @elnic, AiCure, Atomwise

Industries
OPOWER, eHarmony, RetailNext, STITCH FIX, WorkFusion, BLUE RIVER, TACHYUS, SwiftKey, Seeg, FarmLogs, HowGood, select, SIGN MACHINE, statmuse, BOXEVER

Cross-Infrastructure/Analytics

amazon, Google, Microsoft, IBM, SAP, sas, data, hp, Autonomy, VERTICA, vmware, TIBCO, TERADATA, ORACLE, NetApp

Open Source

Framework
hadoop, HDFS, YARN, Spark, MESOS, TEZ, Flink, CDAP

Query / Data Flow
SLAMDATA, HIVE, DRILL, Google Cloud Dataflow

Data Access
cassandra, mongoDB, CouchDB, riak, SCIO, OPENSTDB, nifi

Coordination
talend, Apache Zookeeper, Apache Ambari

Real-Time
STORM, Spark, APEX, Flink, TACHYON, druid

Stat Tools
ScalaLab, NumPy, SciPy

Machine Learning
mlilb, Aerosolve, Apache SINGA, MADlib, CNTK, TensorFlow, jupyter, DL4J

Search
elasticsearch, Solr, Lucene

Security
Apache Ranger, Zeppelin

Data Sources & APIs

Health
JAWBONE, GARMIN, practicefusion, fitbit, Withings, VALIDIC, netatmo, kinsq, Human API

IOT
UPTAKE, ThingWorx, helium, samsara, AUGURY, estimate

Financial & Economic Data
Bloomberg, DOW JONES, THOMSON REUTERS, YODLEE, PREMISE, S&P CAPITAL IQ, quandl, xignite, CB INSIGHTS, mattermark, Stocktwits, estimate, PLAID

Air / Space / Sea
PLANET LABS, spire, WINDWARD, CRUISE, SKYCATCH, Airware, DroneDeploy

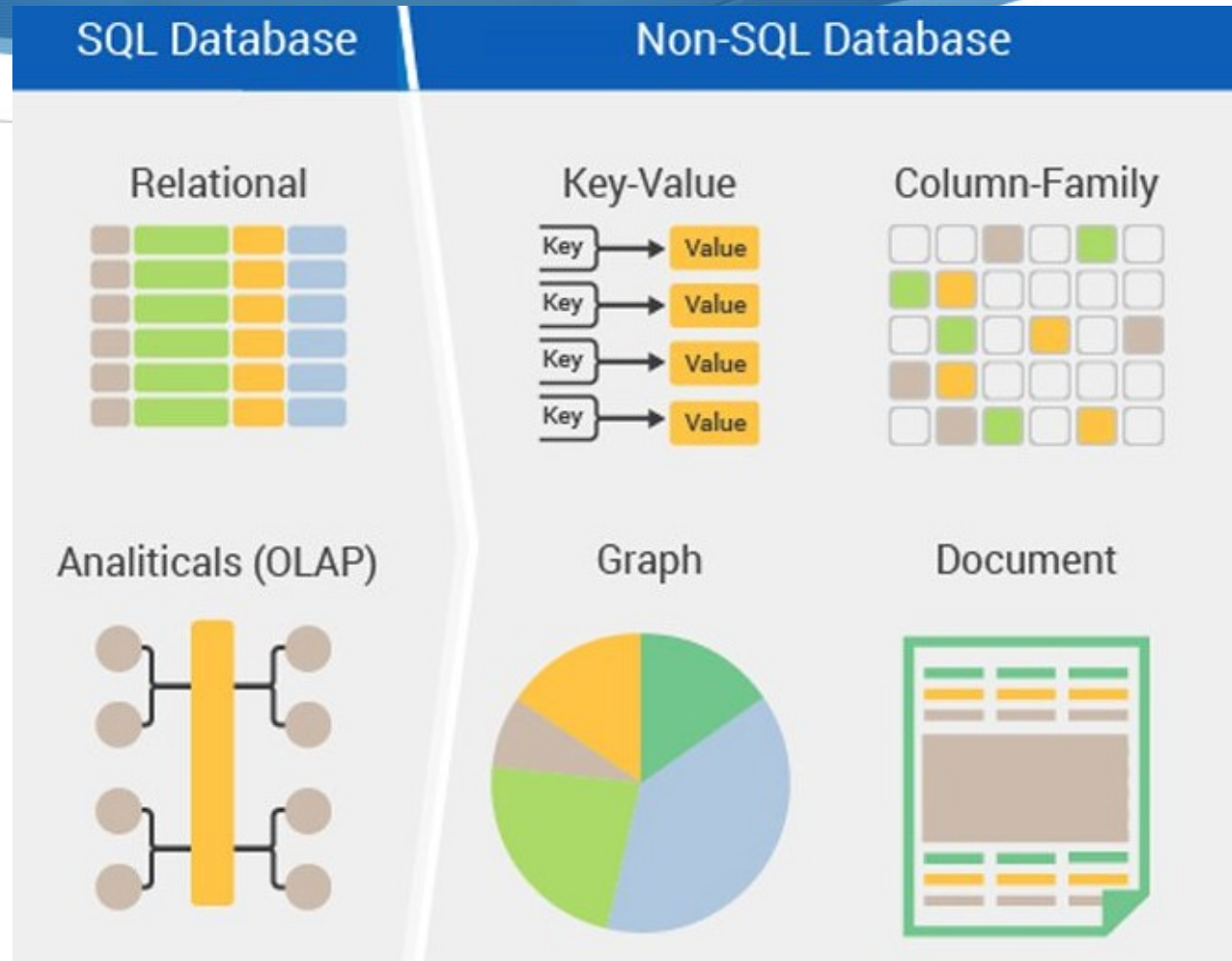
Location / People / Entities
axiom, Experian, EPSILON, InsideView, GARMIN, foursquare, STREETLINE, esri, Crism Hexagon, CARTODB, factual, PlaceIQ, CIRCULATE, placemeter, BASIS, Sense

Other
qualtrics, panjiva, DATA.GOV

Incubators & Schools
GA, PLURALSIGHT, DataCamp, INSIGHT, DataElite, The Data Incubator, METIS

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/>
- Olhometro 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/>