

# AULA 16 – DATA WAREHOUSE

PROFA. DRA. LEILA BERGAMASCO

CC5232 – Banco de Dados

# AGENDA

- Hoje – Big Data
- 16/11 – Dicionário de Dados
- 17/11 – Desenvolvimento do projeto ou dúvidas
- 23/11 – Desenvolvimento do projeto ou dúvidas
- Prova
  - Individual
  - Sem consulta
  - Moodle
  - Questões majoritariamente de múltipla escolha
    - Conceitos gerais
    - Comandos SQL
    - Normalização

# BIG DATA

- O termo big data refere-se legitimamente a conjuntos de dados cujo tamanho está além da capacidade típica das ferramentas de software de banco de dados: capturar, armazenar, gerenciar e analisar.
- Espera-se que a internet das coisas (IOT — Internet of Things) promova uma revolução que melhore a eficiência operacional das empresas e abra novas fronteiras para o aproveitamento de tecnologias inteligentes.
- Big data inclui dados estruturados, semiestruturados e não estruturados em diferentes proporções com base no contexto.
- Dois desafios: a credibilidade da fonte e a adequação dos dados ao público-alvo.

## DEFINIÇÃO E ETIMOLOGIA

- Quando uma massa de dados se torna Big data?
  - IGB? IM? ITB?

Quando é necessário, sumarizar, analisar usando estatística os dados.

- BYGGJA (Vikings ~1000 DC) → BIGGEN (Inglês ~1400) → BIG (América~1600)
- DATUM (Inglês ~1600DC) → Plural =DATA

1600 - John Graunt - surto de peste bubônica

Ele usou estatísticas de mortalidade para alertar sobre o aparecimento e propagação da peste bubônica em Londres.

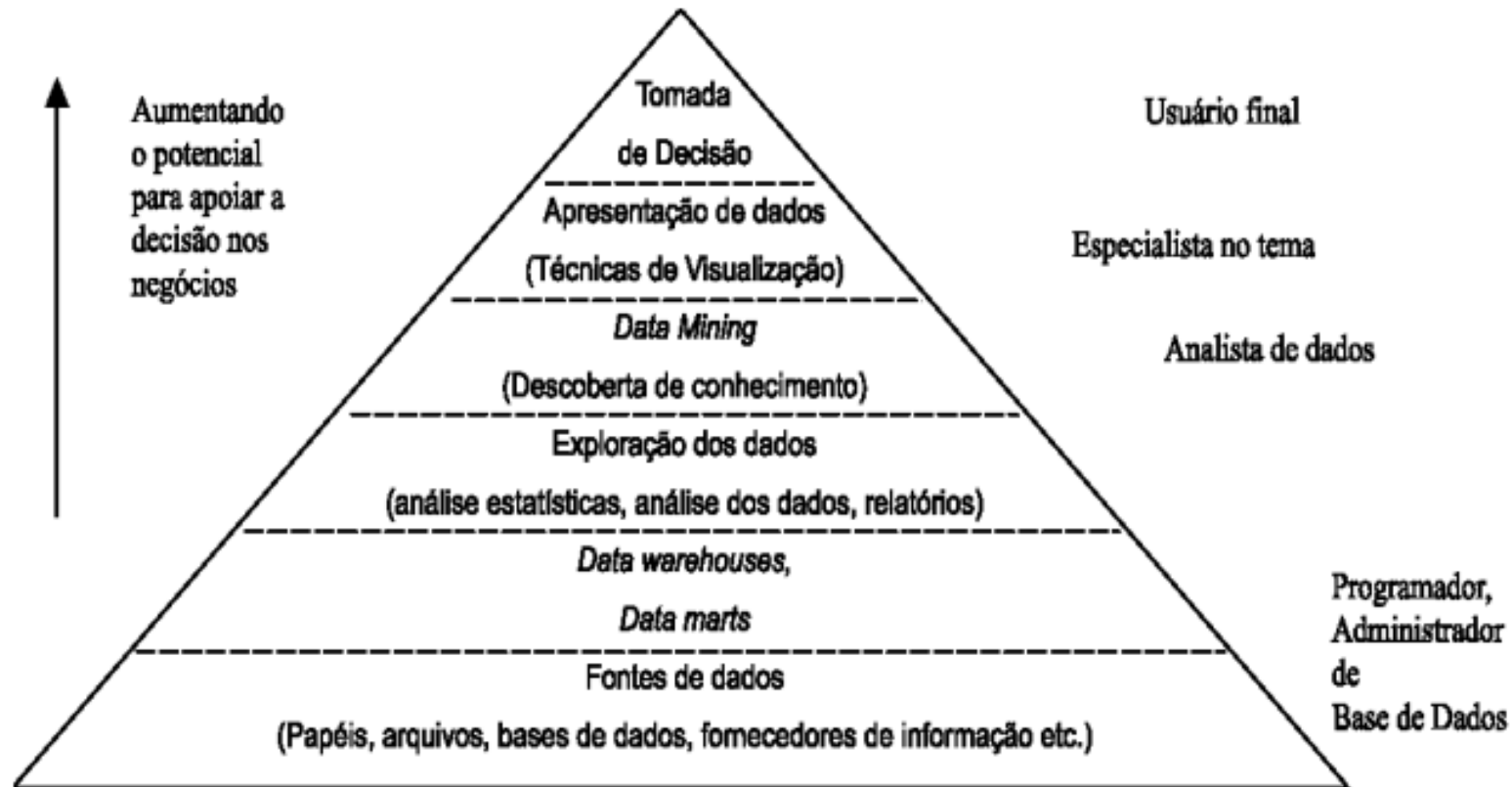
aquilo que nos foi “dado” e usamos como base de nossos cálculos

Dados são representações simbólicas de observações ou pensamentos sobre o mundo

# BIG DATA E 3VS

- Encontram-se até 5 Vs na literatura
- Volume
  - Exige estatística para gerar análises
- Velocidade
  - Existe a necessidade de ter a informação de forma rápida
- Variedade
  - Estruturados e não estruturados
- Valor
  - Potencial para gerar insights
- Veracidade
  - Dados confiáveis

# VALOR





## INFRASTRUCTURE

**HADOOP ON-PREMISE**  
cloudera Hortonworks  
MAPR Pivotal  
IBM InfoSphere  
jethro

**HADOOP IN THE CLOUD**  
aws Microsoft Azure  
Google Cloud  
Cloud Platform  
IBM InfoSphere  
arm  
Dagster CAZENA

**STREAMING / IN-MEMORY**  
Amazon Kinesis  
databricks  
Cloud Platform  
CONFLUENT  
confluent  
stream  
hazelcast  
GridGain  
GIGASPACE  
WOLKEM  
FASTONIX  
KX

**NoSQL DATABASES**  
Google Cloud AWS  
ORACLE Microsoft Azure  
MongoDB MarkLogic  
Couchbase DRYSTRA  
redis ELASTICSEARCH  
AristaDB SCYLLA

**NewSQL DATABASES**  
SAP Clustrix  
Pivotal  
Microsoft SQL Server  
CockroachDB  
VOLTDB  
spilix  
Google Cloud AWS  
ORACLE Microsoft Azure  
Couchbase DRYSTRA  
redis ELASTICSEARCH  
AristaDB SCYLLA

**GRAPH DBs**  
Neo4j  
Amazon Neptune  
ORACLE  
Couchbase DRYSTRA  
redis ELASTICSEARCH  
AristaDB SCYLLA

**MPP DBs**  
Teradata  
VERTICA  
IBM Data Warehouse System  
Celonis  
Exasol  
dremio  
Infoworks

**CLOUD EDW**  
aws  
Google Cloud  
Microsoft Azure  
Pivotal  
Snowflake  
nuclio  
Infoworks

**SERVICES**  
Google Cloud AWS  
ORACLE Microsoft Azure  
MongoDB MarkLogic  
Couchbase DRYSTRA  
redis ELASTICSEARCH  
AristaDB SCYLLA

**DATA TRANSFORMATION**  
talend pentaho  
alteryx TIBCO  
Informatica  
StreamSets UNIFI

**DATA INTEGRATION**  
SAP Data Services  
Informatica  
Talend  
Informatica  
StreamSets UNIFI

**DATA GOVERNANCE**  
Informatica  
IBM  
Celonis  
Exasol  
dremio  
Infoworks

**MGMT / MONITORING**  
aws New Relic  
APPOYNAMICS  
Signalix  
Celonis  
Exasol  
dremio  
Infoworks

**STORAGE**  
aws Google Cloud  
Microsoft Azure  
ORACLE  
MongoDB MarkLogic  
Couchbase DRYSTRA  
redis ELASTICSEARCH  
AristaDB SCYLLA

**CLUSTER SVCS**  
aws Google Cloud  
Microsoft Azure  
ORACLE  
MongoDB MarkLogic  
Couchbase DRYSTRA  
redis ELASTICSEARCH  
AristaDB SCYLLA

**DATA GENERATION & LABELLING**  
aws Google Cloud  
Microsoft Azure  
ORACLE  
MongoDB MarkLogic  
Couchbase DRYSTRA  
redis ELASTICSEARCH  
AristaDB SCYLLA

**AI OPS**  
aws Google Cloud  
Microsoft Azure  
ORACLE  
MongoDB MarkLogic  
Couchbase DRYSTRA  
redis ELASTICSEARCH  
AristaDB SCYLLA

**GPU DBs & CLOUD**  
aws Google Cloud  
Microsoft Azure  
ORACLE  
MongoDB MarkLogic  
Couchbase DRYSTRA  
redis ELASTICSEARCH  
AristaDB SCYLLA

**HARDWARE**  
aws Google Cloud  
Microsoft Azure  
ORACLE  
MongoDB MarkLogic  
Couchbase DRYSTRA  
redis ELASTICSEARCH  
AristaDB SCYLLA

## CROSS-INFRASTRUCTURE/ANALYTICS

aws Google Cloud Microsoft IBM SAP Oracle NetApp Synacor MAPR cloudera

## ANALYTICS &amp; MACHINE INTELLIGENCE

**DATA ANALYST PLATFORMS**  
Microsoft pentaho alteryx  
guavus AYASDI  
ATTIVO Detameer incoffa  
interiana MODE ENDOR  
sisu switchboard Starburst

**DATA SCIENCE PLATFORMS**  
IBM databricks dataiku  
DOMINGO rapidminer TIBCO  
sas  
ANACONDA  
KNIME  
MathWorks

**BI PLATFORMS**  
looker  
aws  
ATSCALE  
Qlik  
birst  
MindStrategy  
Klein ID

**VISUALIZATION**  
tableau  
Google Cloud  
CRONIS  
Chartio

**MACHINE LEARNING**  
aws  
Google Cloud  
CRONIS  
Chartio

**COMPUTER VISION**  
Microsoft Azure  
Amazon Rekognition  
Clarifai  
Everii  
deepomatic  
Liquidity  
Wibitix  
synthesis

**HORIZONTAL AI**  
Microsoft Azure  
Amazon Rekognition  
Clarifai  
Everii  
deepomatic  
Liquidity  
Wibitix  
synthesis

**SPEECH & NLP**  
Google Cloud  
Amazon Rekognition  
Clarifai  
Everii  
deepomatic  
Liquidity  
Wibitix  
synthesis

**SEARCH**  
ORACLE  
Clarifai  
Everii  
deepomatic  
Liquidity  
Wibitix  
synthesis

**LOG ANALYTICS**  
splunk  
sumologic  
elasticsearch  
logz.io

**SOCIAL ANALYTICS**  
Hootsuite  
sprinklr  
NETBASE  
synthesis  
billy  
SimilarWeb

**WEB / MOBILE / COMMERCE ANALYTICS**  
Google Analytics  
mixpanel  
Airtable  
SIGOPT  
custora

## APPLICATIONS - ENTERPRISE

**SALES**  
Salesforce CHURUS  
INSIDE SALES.COM people  
conversa  
clari  
fusemachines

**MARKETING - B2B**  
RADIUS  
EVERESTING  
HINTIGO  
JENGAGIO  
KNATCH mype

**MARKETING - B2C**  
Zeta  
Brite  
Amplero  
Simon

**CUSTOMER EXPERIENCE / SERVICE**  
Qualtrics  
CLARABRIDGE  
HEAD  
DigitGenix  
KAROLIN

**ENTERPRISE PRODUCTIVITY**  
slack  
ORACLE  
GURU lumicha  
SHRIM  
talla

**HUMAN CAPITAL**  
Workday  
ADP  
Paycom  
BambooHR  
Gigamonks  
Deel  
Oyster  
Payscale  
Gigamonks  
Deel  
Oyster  
Payscale

**LEGAL**  
Ravel  
Lexipol  
JUDICATA  
REVIEW  
ROSS  
Casepoint

**REGTECH & COMPLIANCE**  
RegTech  
Compliance  
RegTech  
Compliance

**FINANCE**  
Anaplan  
ZUORA  
Veeva  
SAP  
Oracle  
SAP  
Oracle

**BACK OFFICE AUTOMATION & RPA**  
UiPath  
Automation Anywhere  
Blue Prism  
Workday  
SAP  
Oracle

**SECURITY**  
Tanium  
CyberArk  
SentinelOne  
Vectra  
Palo Alto  
Cisco  
Fortinet

## APPLICATIONS - INDUSTRY

**ADVERTISING**  
Adaptive  
Criteo  
Oracle  
Mucata  
Tribuna  
TAPAB

**EDUCATION**  
Blackboard  
Canvas LMS  
FutureLearn  
FutureLearn

**REAL ESTATE**  
Redfin  
OpenDoor  
VTS  
Credentia  
Zillow

**GOVT**  
OpenDoor  
VTS  
Credentia  
Zillow

**INTELLIGENCE**  
Palantir  
Dataminr  
Quint  
PRIMER  
FRODO

**FINANCE - INVESTING**  
Kenshuc  
Quantopian  
Aladdin  
Gentium  
Aladdin

**FINANCE - LENDING**  
OnDeck  
Affirm  
Kabbage  
Upstart  
LendingClub  
Acorns

**INSURANCE**  
Thomson  
Aviva  
Siemens  
Predix  
Mettler  
Siemens

**HEALTHCARE**  
Flatiron  
Cerner  
Allscripts  
Epic  
Cerner  
Allscripts  
Epic

**LIFE SCIENCES**  
Illumina  
Roche  
Novartis  
Pfizer  
Roche  
Novartis  
Pfizer

**TRANSPORTATION**  
Uber  
Tesla  
Waymo  
Cruise  
Waymo  
Cruise

**AGRICULTURE**  
John Deere  
Case IH  
New Holland  
Fendt  
John Deere  
Case IH  
New Holland  
Fendt

**COMMERCE**  
Amazon  
eBay  
Walmart  
Target  
Amazon  
eBay  
Walmart  
Target

**INDUSTRIAL**  
Siemens  
ABB  
Schneider Electric  
ABB  
Schneider Electric

## OPEN SOURCE

**FRAMEWORKS**  
TensorFlow  
PyTorch  
Keras  
Caffe  
MXNet  
Theano  
OpenAI  
PyTorch

**QUERY / DATA FLOW**  
Spark SQL  
Dremio  
SLAMDATA  
GraphLab

**DATA ACCESS & DATABASES**  
Couchbase  
Redis  
CockroachDB  
Dgraph  
Couchbase  
Redis  
CockroachDB  
Dgraph

**ORCHESTRATION & MGMT**  
Talend  
Apache Airflow  
Rundeck  
Kubernetes

**STREAMING & MESSAGING**  
Spark  
Kafka  
Storm  
Apache Spark

**STAT TOOLS & LANGUAGES**  
R  
Python  
Julia  
R  
Python  
Julia

**AI OPS & INFRA**  
Kubernetes  
Docker  
Prometheus  
Grafana  
Kubernetes  
Docker  
Prometheus  
Grafana

**AI / MACHINE LEARNING / DEEP LEARNING**  
TensorFlow  
PyTorch  
Keras  
Caffe  
MXNet  
Theano  
OpenAI  
PyTorch

**SEARCH**  
Elasticsearch  
Solr  
Elasticsearch  
Solr

**LOGGING & MONITORING**  
Elasticsearch  
Logstash  
Fluentd  
Grafana  
Elasticsearch  
Logstash  
Fluentd  
Grafana

**VISUALIZATION**  
matplotlib  
Tableau  
PowerBI  
Tableau  
PowerBI

**COLLABORATION**  
Slack  
Microsoft Teams  
Slack  
Microsoft Teams

**SECURITY**  
Knox  
Sentry  
Knox  
Sentry

## DATA SOURCES &amp; APIs

**HEALTH**  
Apple  
Fitbit  
Garmin  
Kinsa

**IOT**  
GE Digital  
Uptake  
Helium  
Samsara

**FINANCIAL & ECONOMIC DATA**  
Bloomberg  
Thomson Reuters  
Dow Jones  
Capital IQ  
CB Insights  
PwC  
Deloitte  
EY  
KPMG  
PwC  
Deloitte  
EY  
KPMG

**AIR / SPACE / SEA**  
Airbus  
Boeing  
SpaceX  
Airbus  
Boeing  
SpaceX

**PEOPLE / ENTITIES**  
Axiom  
Experian  
InsideView  
Quantcast  
SafeGraph

**LOCATION INTELLIGENCE**  
Foursquare  
Proton  
Esri  
Factial  
Cuebio  
A Radar

**OTHER**  
DATA.GOV  
IMAGENET  
CRUX  
SignalFire

## DATA RESOURCES

**DATA SERVICES**  
QIQA  
DataCamp  
DataCamp  
DataCamp

**INCUBATORS & SCHOOLS**  
Pluralist  
DataCamp  
DataCamp  
DataCamp

**RESEARCH**  
OpenAI  
MIRI  
Vector Institute  
A2



## DATA & AI LANDSCAPE 2019

### INFRASTRUCTURE

**HADOOP ON-PREMISE**  
cloudera Hortonworks  
MAPR Pivotal  
IBM InfoSphere  
jethro

**HADOOP IN THE CLOUD**  
aws Microsoft Azure  
Google Cloud  
SAP Cloud Platform  
IBM InfoSphere BigInsights arm  
du bole CAZENA

**STREAMING / IN-MEMORY**  
Amazon Kinesis Google Cloud Dataflow databricks  
SAP Cloud Platform ORACLE confluent  
stream hazelcast GridGain  
GIGASPACE Wallaroo FASTDATA kx

### ANALYTICS & MACHINE INTELLIGENCE

**DATA ANALYST PLATFORMS**  
Microsoft pentaho alteryx  
Digital Reasoning GUAVUS AYASDI  
ATTIV/O Datameer Incorta  
inter|ana MODE ENDOR  
sisu switchboard Starburst

**DATA SCIENCE PLATFORMS**  
IBM databricks dataiku  
DOMINO rapidminer TIBCO  
ANACONDA SAS Allaire  
KNIME MathWorks

### APPLICATIONS - ENTERPRISE

**SALES**  
Verstein CHORUS  
INSIDESALES.COM people.ai  
conversica  
clari aviso tact.ai TROOPS  
fuse machines Clearbit

**MARKETING - B2B**  
RADIUS App Annie  
EVERSTRING Lattice  
MINTIGO sense  
tubular  
ENGAGIO Refuel  
KNOTCH mpro

**MARKETING - B2C**  
zeta bloomreach SendGrid  
braze ACTIONIQ BLUECORE  
CONTENT SQUARE TEALUM mparticle  
Amplero amperity QUANTIFIND  
Simon Lytica PERSADO  
remesh

**CUSTOMER EXPERIENCE / SUPPORT**  
qualtrics MEDALLIA  
CLARABRIDGE zendesk  
INTERCOM Drift LIVEPERSON  
HEAP Amplitude Watson AI  
DigitalGenius ASAPP ada  
Ca#Desk metomi

**NoSQL DATABASES**  
Google Cloud AWS  
ORACLE Microsoft Azure  
mongoDB MarkLogic  
Couchbase DATASIX  
redislabs REDOSPIKE  
ArangoDB SCYLLA

**NewSQL DATABASES**  
SAP Clustrix  
Pivotal  
MEMSQL influxdata  
Cockroach LABS  
VOLTDB splice  
paradigm

**GRAPH DBs**  
neo4j  
Amazon Neptune  
ORACLE  
OrientDB  
InfiniteGraph  
Objectivity

**MPP DBs**  
TERADATA  
IBM Data Warehouse Systems  
Cobion  
Kognitio  
Exasol  
dremio  
Yellowbrick

**CLOUD EDW**  
aws  
Google Cloud  
Microsoft Azure  
Pivotal  
snowflake  
Infoworks

**SERVERLESS**  
AWS Lambda  
PULSAR  
nuclio  
PaaS Function Service

**BI PLATFORMS**  
looker  
aws  
ATSCALE  
Qlik  
GoodData  
MicroStrategy  
Keen IO

**VISUALIZATION**  
+tableau  
Power BI  
SAP  
Google Cloud  
celonis  
zepl  
CHARTIO

**MACHINE LEARNING**  
Amazon SageMaker  
H2O  
DataRobot gamalor  
VISENZE ELEMENT  
deepense.ai

**DATA TRANSFORMATION**  
talend pentaho  
alteryx TRIFACTA  
tamr Paxata  
StreamSets UNIFI

**DATA INTEGRATION**  
SAP Data Services Informatica  
MuleSoft TEALUM  
Inlogix enigma  
Segment ATTUNITY  
xplenty ZALONI  
Infoworks Fivetran  
SNOWFLOW MATILLION

**DATA GOVERNANCE**  
Informatica  
IBM  
collibra  
Alation  
OKERA  
MANTA data.world

**MGMT / MONITORING**  
aws New Relic octrio  
rubrik APPDYNAMICS  
dynatrace  
SignalFx dnuvo  
splunk Moogsoft pagerduty  
unravel Numentary  
zenoss OpRamp MAGNITUDE

**COMPUTER VISION**  
Microsoft Azure  
Amazon Rekognition  
clarifai  
EVER AI deepomatic  
neurologic twentybn  
UBIQUITY ADEE  
YITU trax  
synthesis

**HORIZONTAL AI**  
IBM Watson Cortana Face++  
sentient  
Voyager  
Affectiva  
Numenta  
PETUUM  
nalogics  
BLUE VISION

**SPEECH & NLP**  
Google Cloud  
Amazon Alexa Amazon Translate  
twilio  
narrative science semantic machines  
SoundHound Inc.  
Mindfield  
cogito snips  
SMARTING UN Unbabel PolyAI

### APPLICATIONS - INDUSTRY

**ADVERTISING**  
AppNexus Rubicon  
criteo xAd Integral  
ORACLE MOAT  
theTradeDesk  
dstillery LiveIntent  
TAPAD dataxu gumgum  
Appier

**EDUCATION**  
Lullis huo  
KNEWTON  
Clever  
edureka  
kidaptive  
PANORAMA  
knowre  
gradescope

**REAL ESTATE**  
REDFIN  
Opendoor  
VTS  
CREDIFI  
GEOPHY  
reonomy  
COMPSTAK  
SPACE MAKER

**GOV'T**  
OPENGOV  
mark43  
FiscalNote  
LiveStories  
Passport  
SmartProcure  
STREETLIGHT DATA  
OpenDataSoft

**INTELLIGENCE**  
Palantir  
Dataminr  
Quid  
PRIMER  
FORGE

**FINANCE - INVESTING**  
KENSHC  
Quantopian  
ADDEPAR  
NUMERAI  
iSENTIUM  
ALGORZ  
PAGAYA

**FINANCE - LE**  
ondeck  
JIANPU.AI  
TALA  
aura  
100Credit  
TrueAccord  
cignifi

**STORAGE**  
aws Google Cloud  
Microsoft Azure  
PURE STORAGE  
ALLUXIO wasabi  
nimblestorage  
Dumulo panasas  
COHERITY

**CLUSTER SVCS**  
Amazon ECS  
IBM Amazon EKS  
MESOSPHERE  
packet  
Bright Computing  
CYCLOCLOUD

**DATA GENERATION & LABELLING**  
amazon mechanical turk  
Upwork  
appen scale  
HIVE Labelbox  
Mighty AI  
LIONBRIDGE

**AI OPS**  
ALGORITHMIA  
comet  
Verta.ai datmo  
datastrato  
Determined AI  
fiddler

**GPU DBs & CLOUD**  
kinetica  
SOREM  
bryllyt  
BLAZINGDB  
PG-Strom  
FLOYDHUB

**HARDWARE**  
Google TPU arm  
intel intel Xeon  
NVIDIA GRAPHCORE MYTHIC  
GARDIAN  
Movidius habana  
VIAVE  
CERNAMI  
LEONIX

**SEARCH**  
elasticsearch ORACLE  
algolia covéo  
Lucidworks ATTIV/O  
swiftype EXALFO  
alphasense MAANA  
omni:us SINEQUA

**LOG ANALYTICS**  
splunk  
sumologic  
solarwinds loggly  
TIMBER  
kibana  
logz.io

**SOCIAL ANALYTICS**  
Hootsuite sprinkr  
NETBASE  
synthesio tracx  
smile reach  
bitly SimilarWeb

**WEB / MOBILE / COMMERCE ANALYTICS**  
Google Analytics  
mixpanel AMPITUDE  
Airtable RESCI  
SIGOPT granify  
custora

### CROSS-INFRASTRUCTURE/ANALYTICS

aws Google Cloud Microsoft IBM SAP Hewlett Packard Enterprise SAS 1010DATA vmware TIBCO TERADATA ORACLE NetApp syncsort MAPR cloudera



# VISÃO GERAL DE MINERAÇÃO DE DADOS

- A mineração de dados pode ser usada em conjunto com um *data warehouse* para ajudar com certos tipos de decisões.
- Pode ser aplicada a bancos de dados operacionais com transações individuais.
- Para tornar a mineração de dados mais eficiente, o data warehouse deve ter uma coleção de dados agregada ou resumida.
- Ajuda na extração de novos padrões significativos que não necessariamente podem ser encontrados apenas ao consultar ou processar dados ou metadados no DE
- As aplicações de mineração de dados devem ser fortemente consideradas desde cedo, durante o projeto de um DW

# KDD

- A descoberta de conhecimento nos bancos de dados, (Knowledge Discovery Database - KDD), normalmente abrange mais que a mineração de dados.

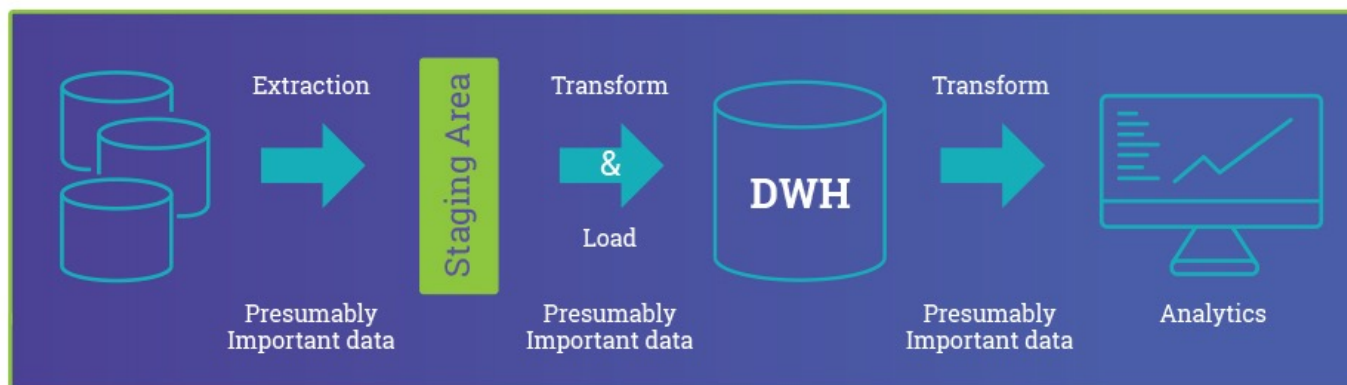
- O processo de descoberta de conhecimento compreende seis fases:

- seleção de dados,
  - limpeza de dados,
  - enriquecimento,
  - transformação ou codificação de dados,
  - análise de dados e o relatório e
  - exibição da informação descoberta.
- Pré-processamento
- Mineração de Dados
- KDD

# PRÉ-PROCESSAMENTO

- Em teoria, o ETL é feito antes do dado ser inserido no DW
- Entretanto mesmo estando em um DW o dado pode ainda precisar de enriquecimento

## ETL



# ETAPAS DO PRÉ-PROCESSAMENTO

- Seleção: Coletar, agrupar dados mais relevantes
  - Consultas à base de dados (SQL, noSQL)
- Limpeza: Inputar valores, reduzir ruídos, eliminar duplicadas
  - Causas: Mau funcionamento do equipamento de coleta, Usuário (Não considerou importante, Engano)
  - Tratamento: Ignorar, Eliminar registro, Preencher os valores ausentes manualmente, usar uma constante global para representar os valores ausentes (não recomendado, pois o sistema pode identificar esse valor como um conceito); Ex. categoria “Vazio”, usar a média (ou a moda), Usar o valor mais provável segundo um modelo (regressão, regra de Bayes, árvores de decisão)
  - Ferramentas: Python (pandas, numpy), R, Excel
- Enriquecimento/Transformação: Normalizar, discretizar, criar atributos

Paciente	Compatibilidade sanguínea
1	{AB, O+, O-}
2	{AB}



Paciente	Compatibilidade sanguínea		
	AB	O+	O-
1	1	1	1
2	1	0	0

- Redução: de dimensão, de dados, balanceamento



# OBJETIVOS

- A mineração de dados costuma ser executada com alguns objetivos finais ou aplicações.
- De modo geral, esses objetivos se encontram nas seguintes classes:
  - previsão, identificação, classificação e otimização.
- Onde se tem aplicado Mineração de dados com resultados satisfatórios?
  - Bancos: auxiliar no gerenciamento de relacionamento com o cliente;
  - Cartão de Crédito: identificar segmentos de mercado e rotatividade;
  - Cobrança: detecção de fraudes;
  - Eleitoral: identificação de um perfil para possíveis votantes;
  - Medicina: indicações de diagnósticos mais precisos;
  - Tomada de decisão: filtro de informações relevantes fornecendo indicadores de probabilidade.

# MINERAÇÃO DE DADOS

- Constituído de 3 etapas
  - Escolha da tarefa: pode ser feita uma combinação de tarefas para uma melhor extração de padrões. Entre as tarefas estão: descrição, classificação, estimação ou regressão, predição, agrupamento e associação
  - Escolha da Técnica: de acordo com a tarefa ou conjunto de tarefas selecionadas é escolhida a técnica que será utilizada.
  - Aplicação da técnica

# TIPOS DE APRENDIZADO

- Para extrair o conhecimento a mineração de dados pode aplicar dois tipos de abordagens:
  - Aprendizado supervisionado: tenta explicar ou categorizar dados em particular;
  - Aprendizado não supervisionado: tenta encontrar padrões ou similaridades entre grupos de registros sem o uso de um campo em particular como alvo ou de conjuntos de classes pré-definidos.

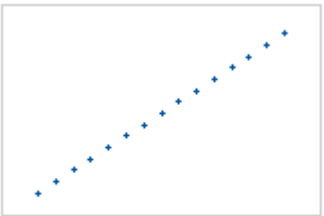
Tarefas & Métodos de Mineração de Dados	Algoritmos de Mineração de Dados	Tipo de Aprendizado
<div>Previsão</div> <div>Classificação</div> <div>Regressão</div> <div>Série temporal</div>	<div>Árvores de Decisão, Redes Neurais, Máquinas de Vetores de Suporte, kNN, Naïve Bayes, GA</div> <div>Regressão Linear/Não linear, ANN, Árvores de Regressão, SVM, kNN, GA</div> <div>Métodos Autorregressivos, Métodos de Extração de Médias, Suavização Exponencial, ARIMA</div>	<div>Supervisionado</div> <div>Supervisionado</div> <div>Supervisionado</div>
<div>Associação</div> <div>Cesta de mercado</div> <div>Análise de elos</div> <div>Análise sequencial</div>	<div>Apriori, OneR, ZeroR, Eclat, GA</div> <div>Maximização de Expectativa, Algoritmo Apriori, Correspondência Baseada em Gráficos</div> <div>Algoritmo Apriori, FP-Growth, Correspondência Baseada em Gráficos</div>	<div>Não supervisionado</div> <div>Não supervisionado</div> <div>Não supervisionado</div>
<div>Segmentação</div> <div>Agrupamento</div> <div>Análise de discrepâncias</div>	<div><i>k-means</i>, Maximização de Expectativa (ME)</div> <div><i>k-means</i>, Maximização de Expectativa (ME)</div>	<div>Não supervisionado</div> <div>Não supervisionado</div>



# SUPERVISIONADO

## ■ Previsão

- Regressão: A partir de dados históricos descreve a relação entre variáveis.
  - Ela pode ser positiva ou negativa



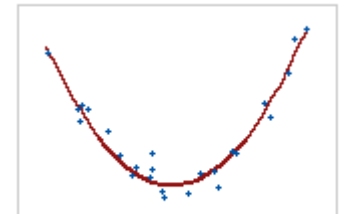
Pearson = +1, Spearman = +1



Pearson = -0,093, Spearman = -0,093



Pearson = -1, Spearman = -1



Coefficiente 0



iris setosa



iris versicolor

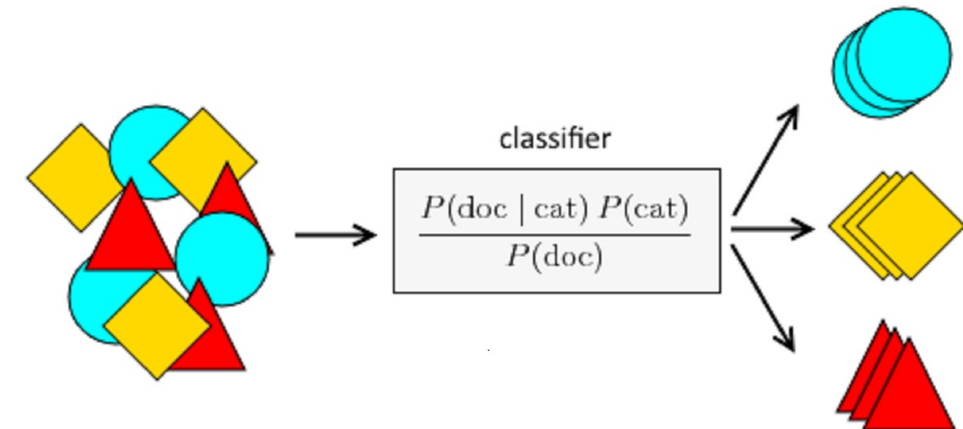
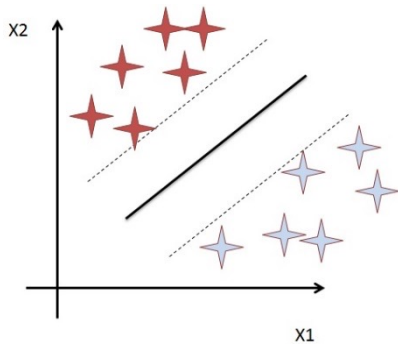


iris virginica

# SUPERVISIONADO

## ■ Previsão

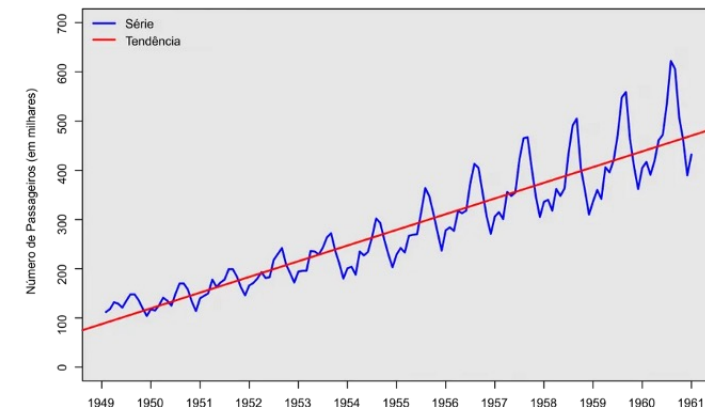
- Classificação: A partir de dados históricos classifica determinado item
  - Diferença da regressão: Regressão retorna uma probabilidade e classificação um rótulo
    - Árvores de Decisão, SVM, Ensemble, Naive Bayes



# SUPERVISIONADO

## ■ Previsão























- Série temporal: Sequência de realizações (observações) de uma variável ao longo do tempo. É possível extrair tendências a partir desses dados
  - Economia: Preços diários de ações, taxa mensal de desemprego;
  - Saúde: Número mensal de novos casos de alguma doença, registro de um eletrocardiograma de uma pessoa;
  - Climatologia: Temperatura diária;
  - Marketing: Previsão de aquisição de novos clientes, previsão de volume de viagens;
  - Administrativo: Previsão de vendas de produtos;



# NÃO SUPERVISIONADO

## ■ Associação

- Cestas de mercado - esta tarefa consiste em identificar quais atributos estão relacionados; é uma das tarefas mais conhecidas. Por exemplo, cestas de compras em que produtos são levados juntos pelos consumidores.
- Detecção de sequências - utilização de algum tipo de padrão nos dados para determinar que tipos de sequências possam ser determinados. Por exemplo, um cliente compra um determinado produto e meses depois compra um produto associado ao primeiro;
- Análise de elos - descreve e estuda a regularidade de modelos ou tendências para objetos cujo comportamento muda ao longo do tempo. Por exemplo, ampliação de estoque;
- Algoritmos: Apriori

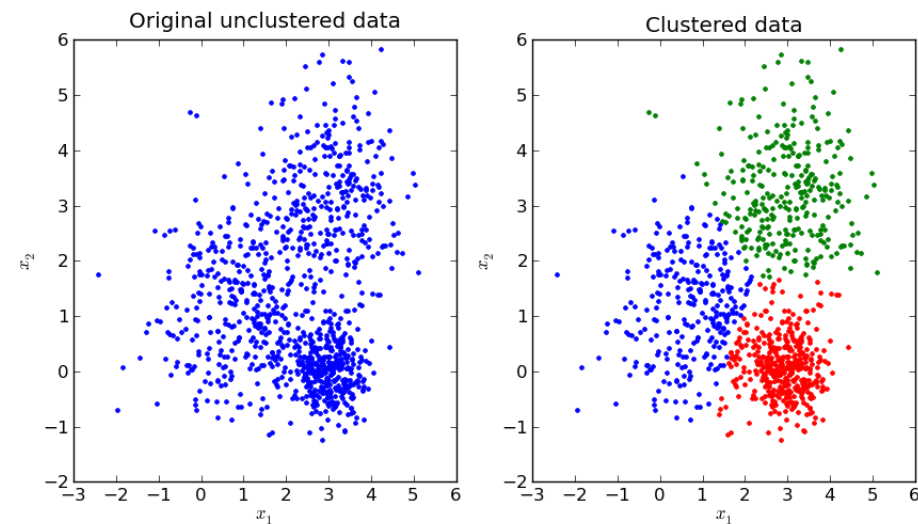
Transaction 1	   
Transaction 2	  
Transaction 3	 
Transaction 4	 
Transaction 5	   
Transaction 6	  
Transaction 7	 
Transaction 8	 



# NÃO SUPERVISIONADO

## ■ Segmentação

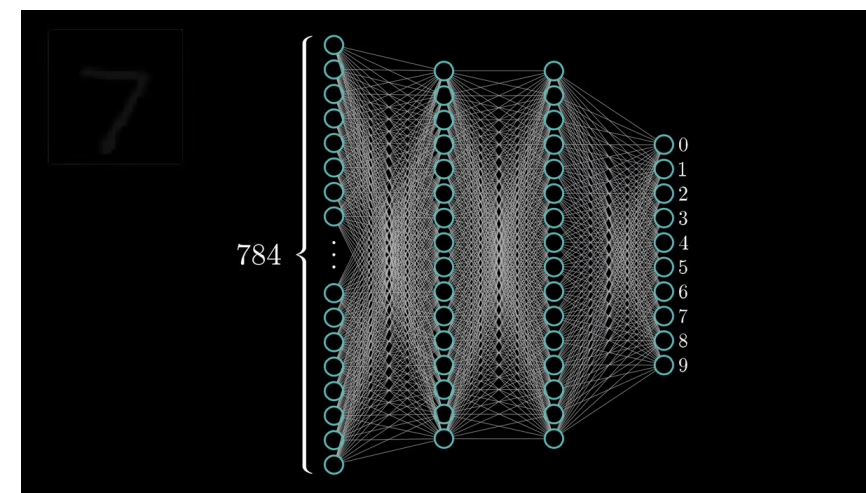
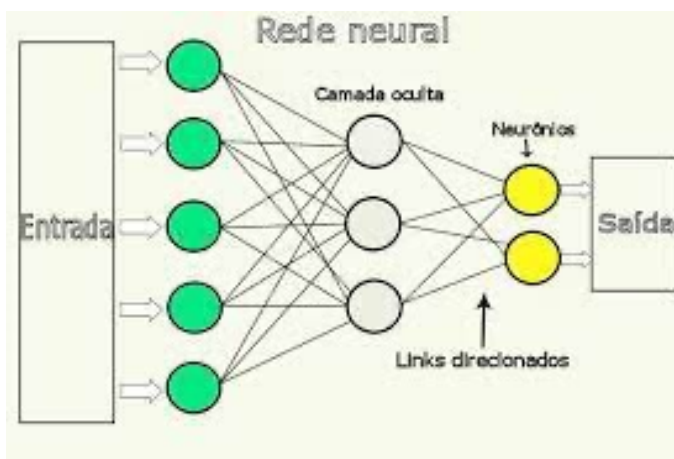
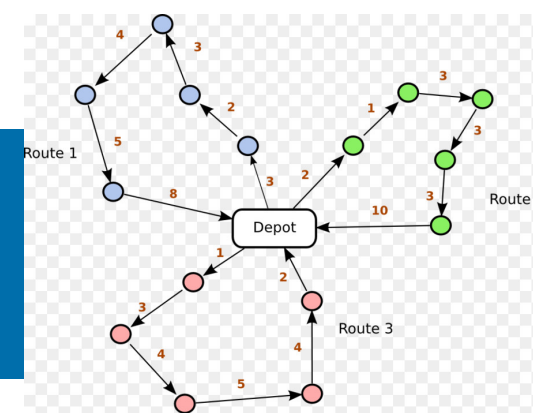
- Agrupamento ou Clustering - tem o objetivo de identificar e aproximar os registros similares. Consiste de uma coleção de registros similares entre si, porém diferentes de outros tipos de registros em demais agrupamentos.
- Identificação - utilizar padrões de dados para identificar a existência de um item, um evento ou uma atividade. Por exemplo, aplicações biológicas para autenticação de usuário específico ou de classe autorizada;



# NÃO SUPERVISIONADO

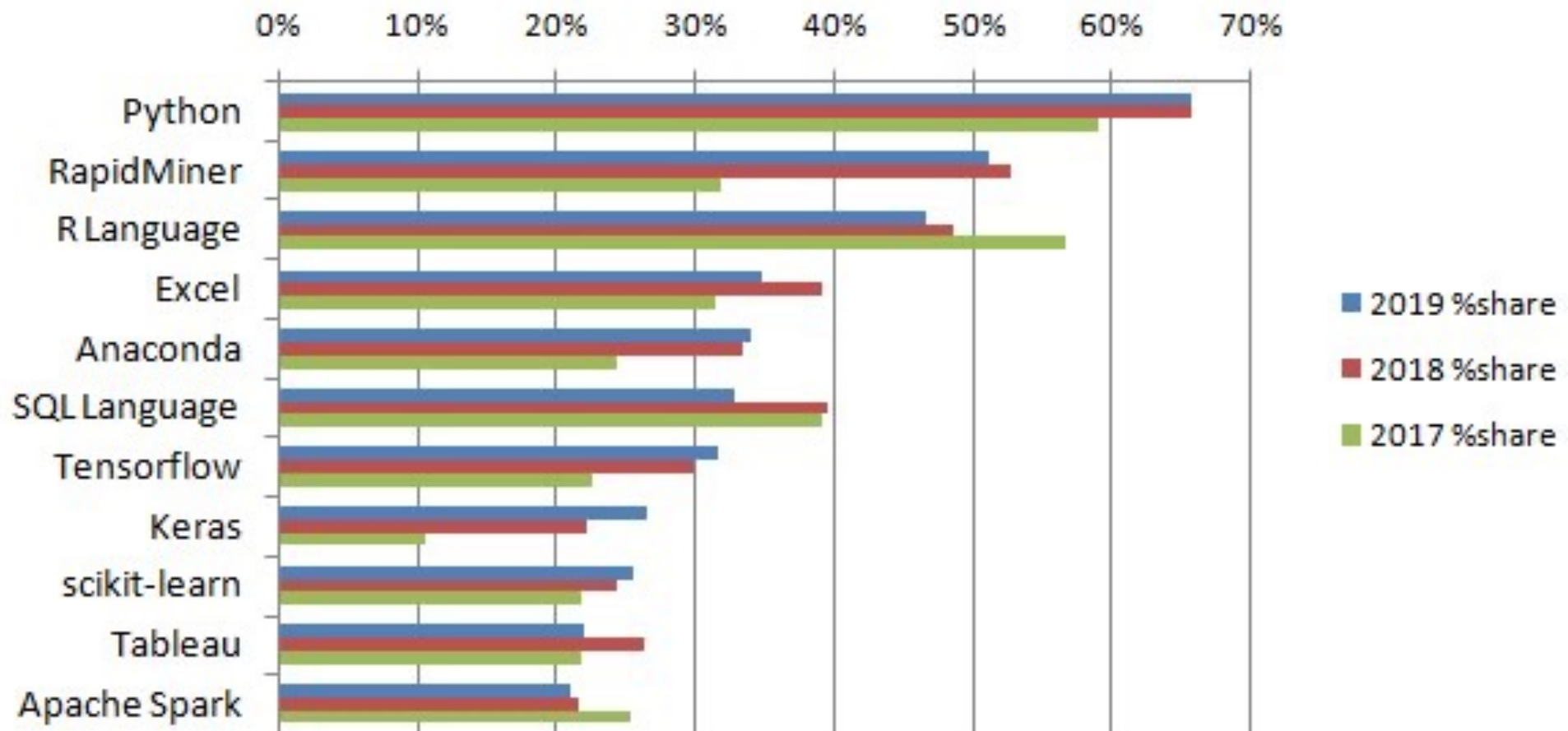
## ■ Outros

- Análise em dados no formato texto, vídeo, voz - trabalhar dados em formatos complexos visando transformar em formato de uso e extrair seus resultados baseados em técnicas de tratamento e exploração de dados complexos. Por exemplo: Alexa, Automoveis inteligentes
- Otimização - visa otimizar recursos limitados como tempo, espaço, dinheiro, matéria-prima, dentre outros, buscando maximizar resultados com venda, lucros, distribuição, economia de espaço e etc. Por exemplo, estudo das vendas de um supermercado;



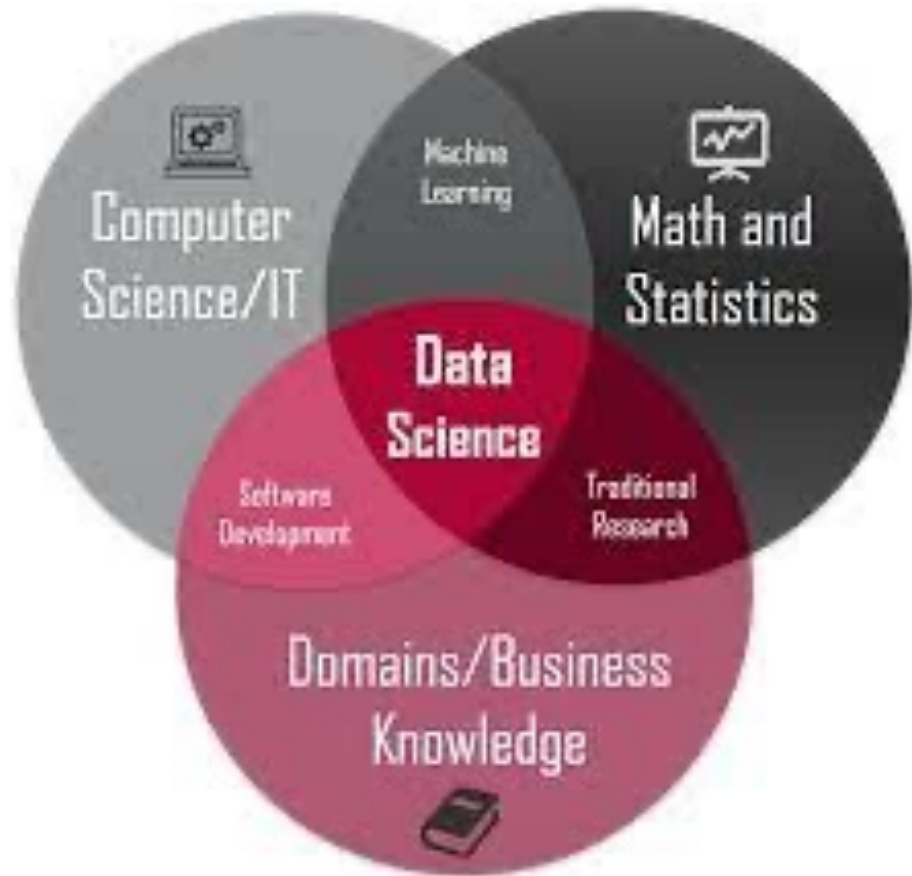
Ferramenta	Fornecedor	Tarefas
WEKA	University of Waikato	Classificação, Regressão e Regras de Associação.
Intelligent Miner	IBM Corp.	Classificação, Regras de Associação, Clusterização e Sumarização.
Oracle Data Miner	Oracle	Classificação, Regressão, Associação, Clusterização e Mineração de Textos.
SAS Enterprise Miner Suite	SAS Inc.	Classificação, Regras de Associação, Regressão e Sumarização.
Clementine	SPSS Inc.	Classificação, Regras de Associação, Clusterização, Sequência e Detecção de Desvios.
Darwin	Thinking Machines	Classificação.
Business Objects	Business Objects	Classificação, Regras de Associação, Clusterização e Sumarização.
Microsoft Data Analyser	Microsoft Corp.	Classificação e Clusterização.
MineSet	Silicon Graphics Inc.	Classificação, Regras de Associação, Análise Estatística.
DBMiner	DBMiner Technology Inc.	Classificação, Regras de Associação e Clusterização.
Gemanics Expression	Gemanics Developer	Análise de Sequências.
SAS Text Miner	SAS Inc.	Mineração de Textos.

## Top Analytics, Data Science, Machine Learning Software 2017-2019, KDnuggets Poll





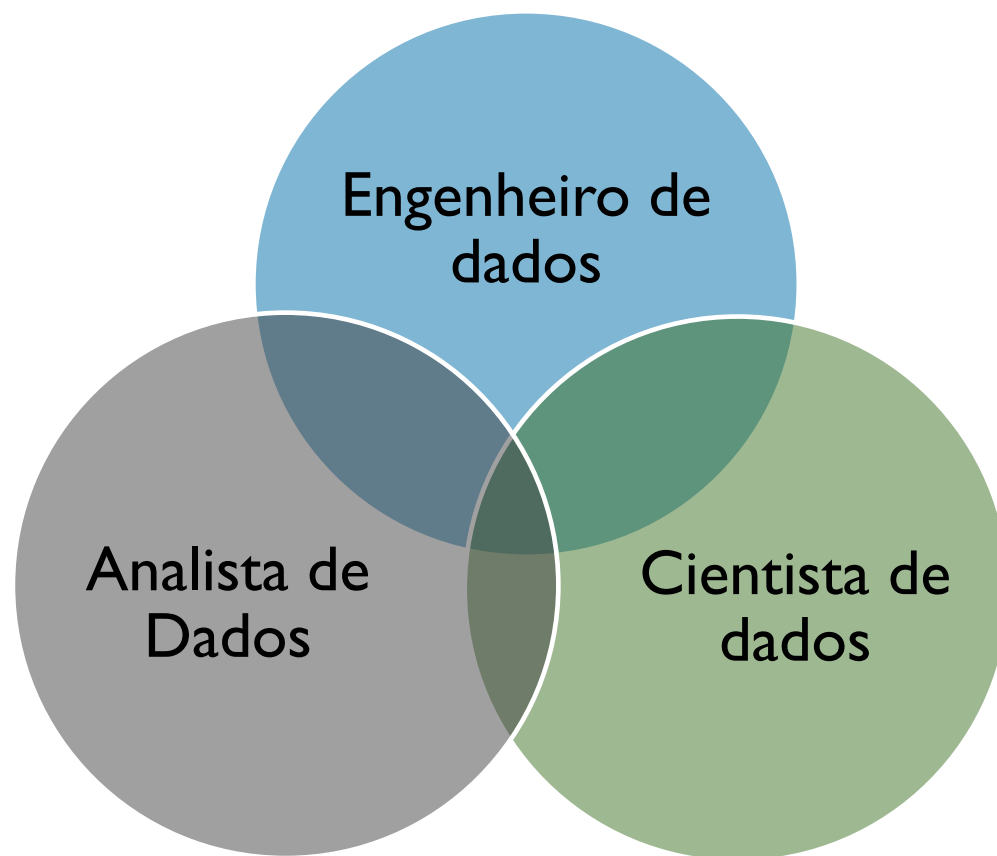
# CARREIRA EM DADOS



- Interdisciplinar
- Orientada a problemas
- Dados como fonte de trabalho

# TRILHAS DE CARREIRAS

- Ciência de dados é um grande guarda chuva
  - **Engenheiro de dados:** data pipeline (e2e), qualidade do dado, escalabilidade
    - MLOps: delivery da solução em nuvem, normalmente
  - **Cientista de dados:** modelos preditivos, prescritivos, hipóteses, avaliação
  - **Analista de dados (BI):** interface com cliente, *storytelling*, criação de visuais relevantes, KPIs



empregos na área.

## As 25 profissões de tecnologia mais buscadas por empregadores:

1. Desenvolvedor back-end – 21.802 vagas
2. Desenvolvedor front-end – 18.680 vagas
3. Engenheiro de software – 12.870 vagas
4. Product manager – 12.005 vagas
5. Desenvolvedor full-stack – 8.366 vagas
6. Gestor de mídias sociais – 5.086 vagas
- 7. Analista de Business Intelligence – 3.781 vagas
8. Especialista em machine learning – 3.414 vagas
9. Engenheiro de dados – 2.928 vagas
10. Cientista de dados – 2.213 vagas
11. Scrum master – 2.093 vagas
12. Gestor de projetos – 2.085 vagas
13. Desenvolvedor web – 1.551 vagas
14. Desenvolvedor de banco de dados – 1.476 vagas
15. DPO (Data Protection Officer) – 1.366 vagas

A pesquisa entrevistou seis mil profissionais de todo o Brasil no ano passado, em cargos de suporte à gestão até alta e média gerência. Depois, a empresa analisou a remuneração mensal de 719 cargos em 15 setores.

Na comparação com o ano passado, 41% das posições tiveram reposição ou manutenção salarial, quanto 4% dos cargos registraram queda.

Segundo o estudo, o maior acréscimo de remuneração foi na área de tecnologia da informação. O cargo de analista de projetos teve aumento de 75%. Em segundo lugar, fica o analista de business intelligence, com 70%.

  
<https://exame.com/carreira/cargos-salario-em-alta-2022-pagegroup/>

OBRIGADO E ATÉ A PRÓXIMA AULA!