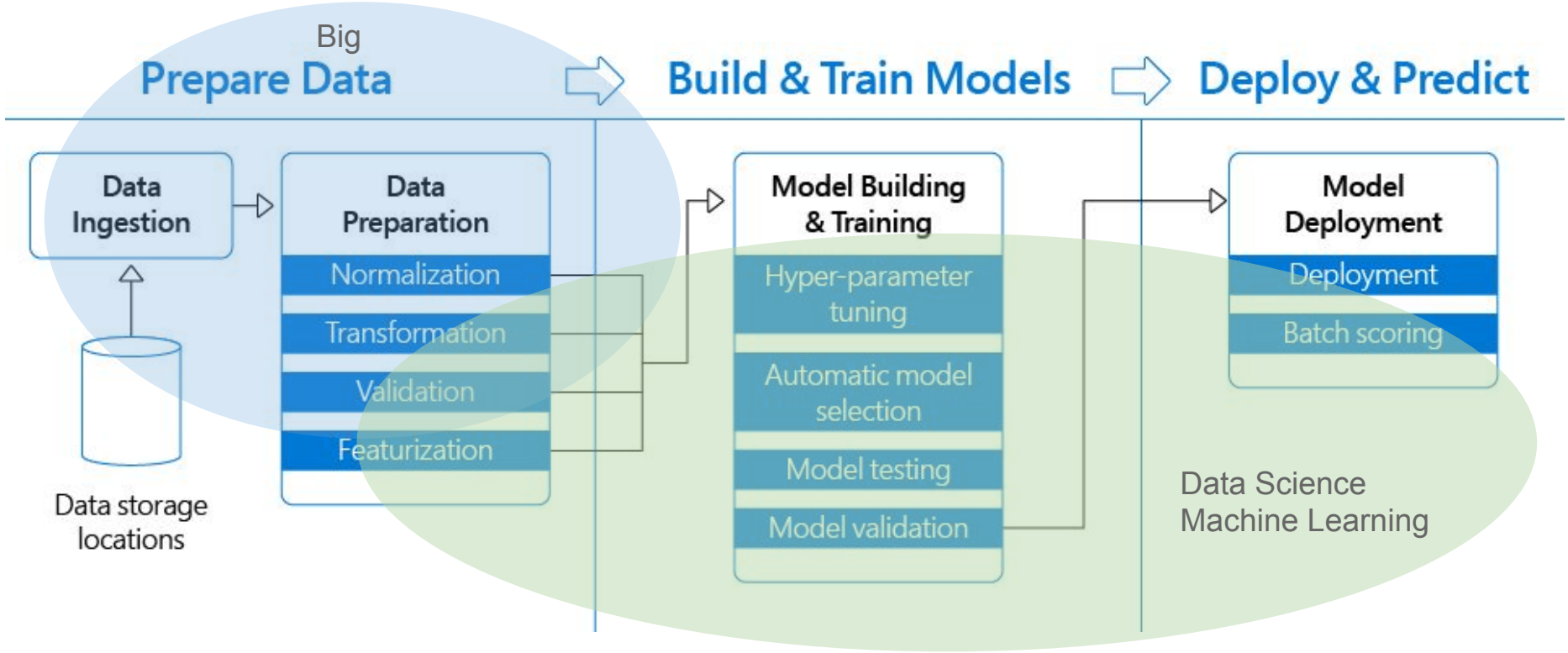
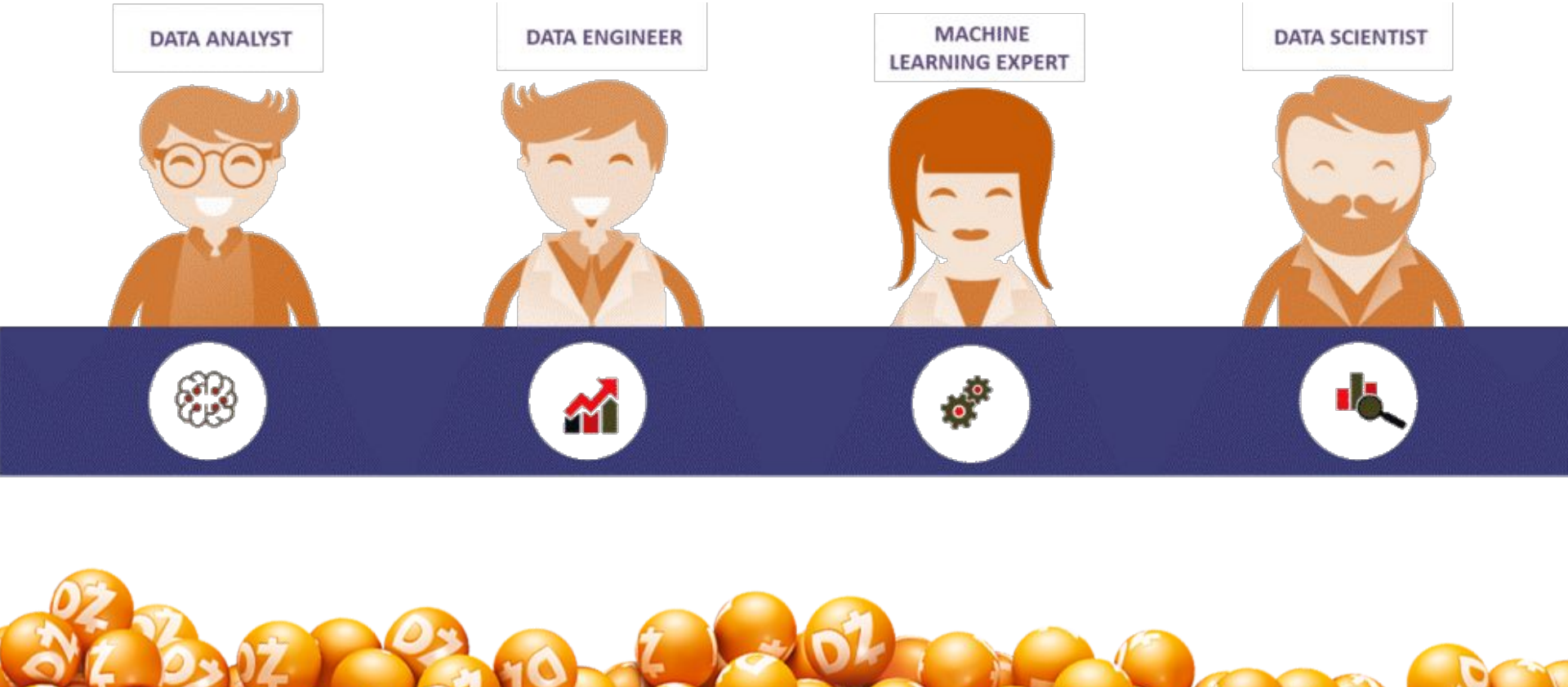
Da ingestão do dado ao machine learning

Mini-curso 1: 29/04/2020

Diógenes Justo

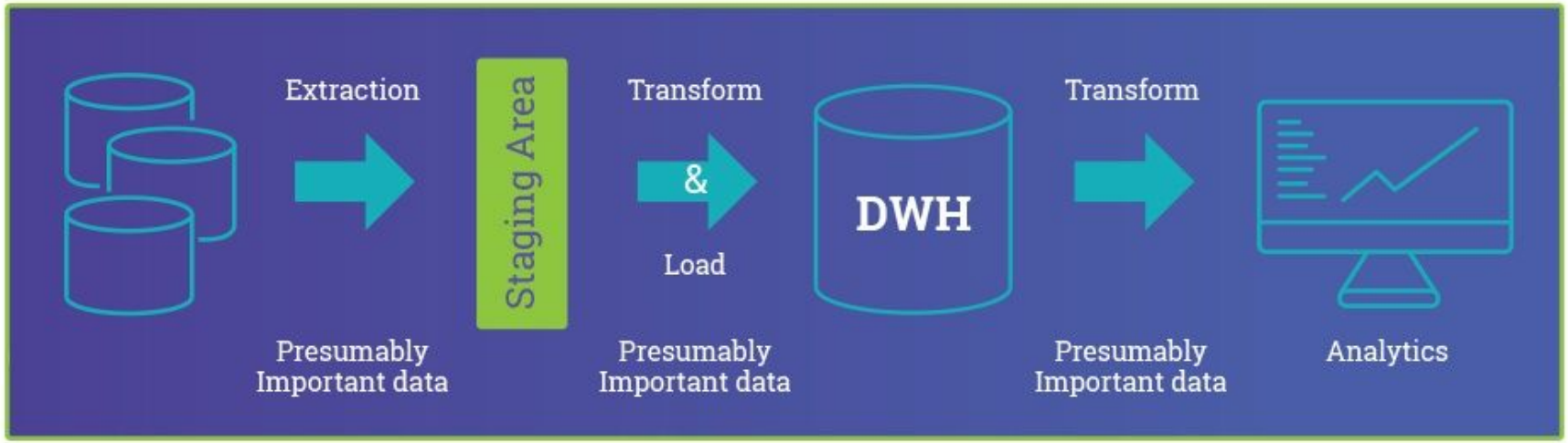

Big
Prepare Data
Build & Train Models
Deploy & Predict
Data Ingestion
Data storage locations
Data Preparation
Normalization
Transformation
Validation
Featurization
Model Building & Training
Hyper-parameter tuning
Automatic model selection
Model testing
Model validation
Model Deployment
Deployment
Batch scoring
Data Science
Machine Learning


DATA ANALYST
DATA ENGINEER
MACHINE LEARNING EXPERT
DATA SCIENTIST

Ingestão de dados

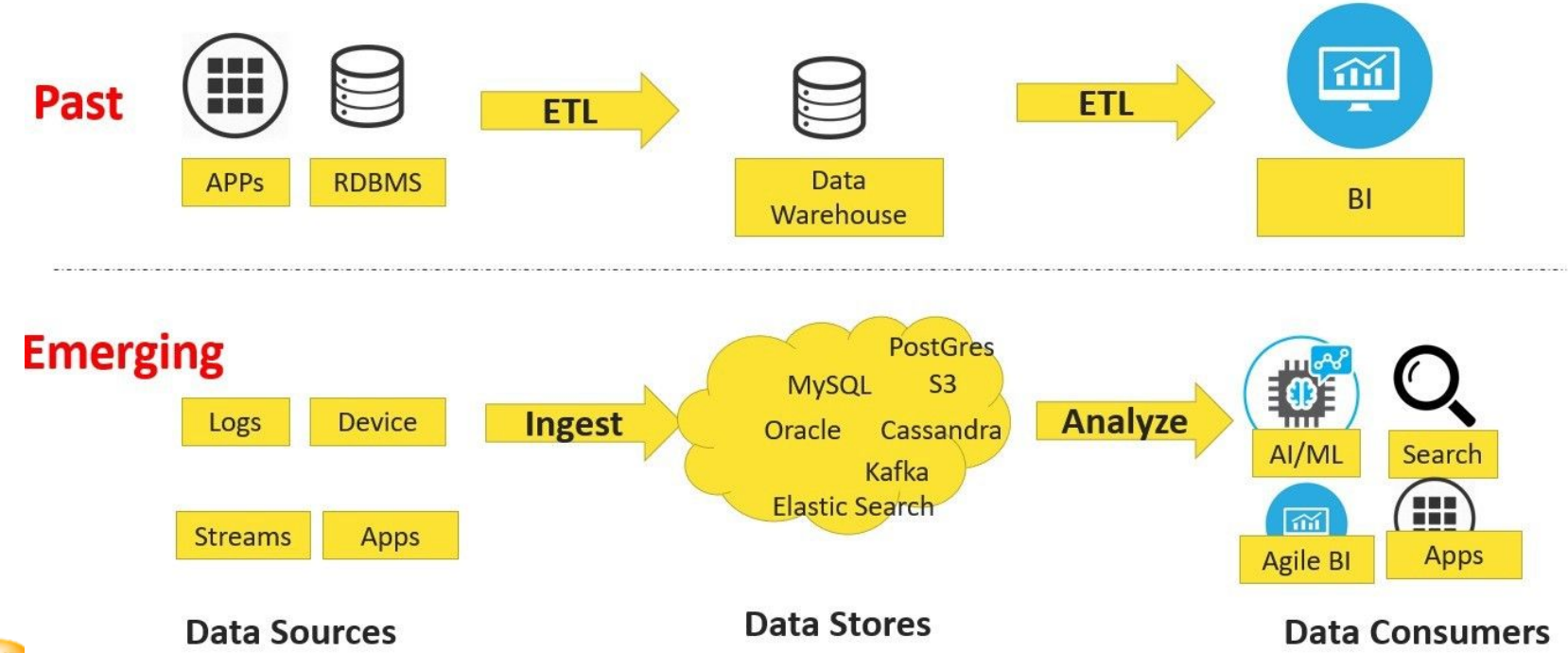
80 - 2000's

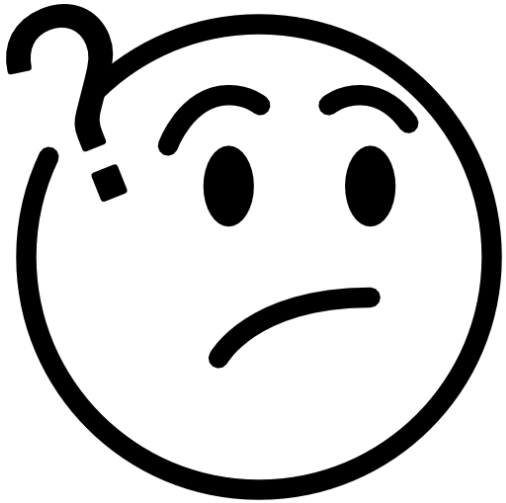
ETL


Extraction
Staging Area
Transform
&
Load
DWH
Transform
Analytics
Presumably
Important data
Presumably
Important data
Presumably
Important data

2010' -

Evolution of Data in Motion





Por que raios copiar dados de um sistema transacional e criar um ambiente analítico?

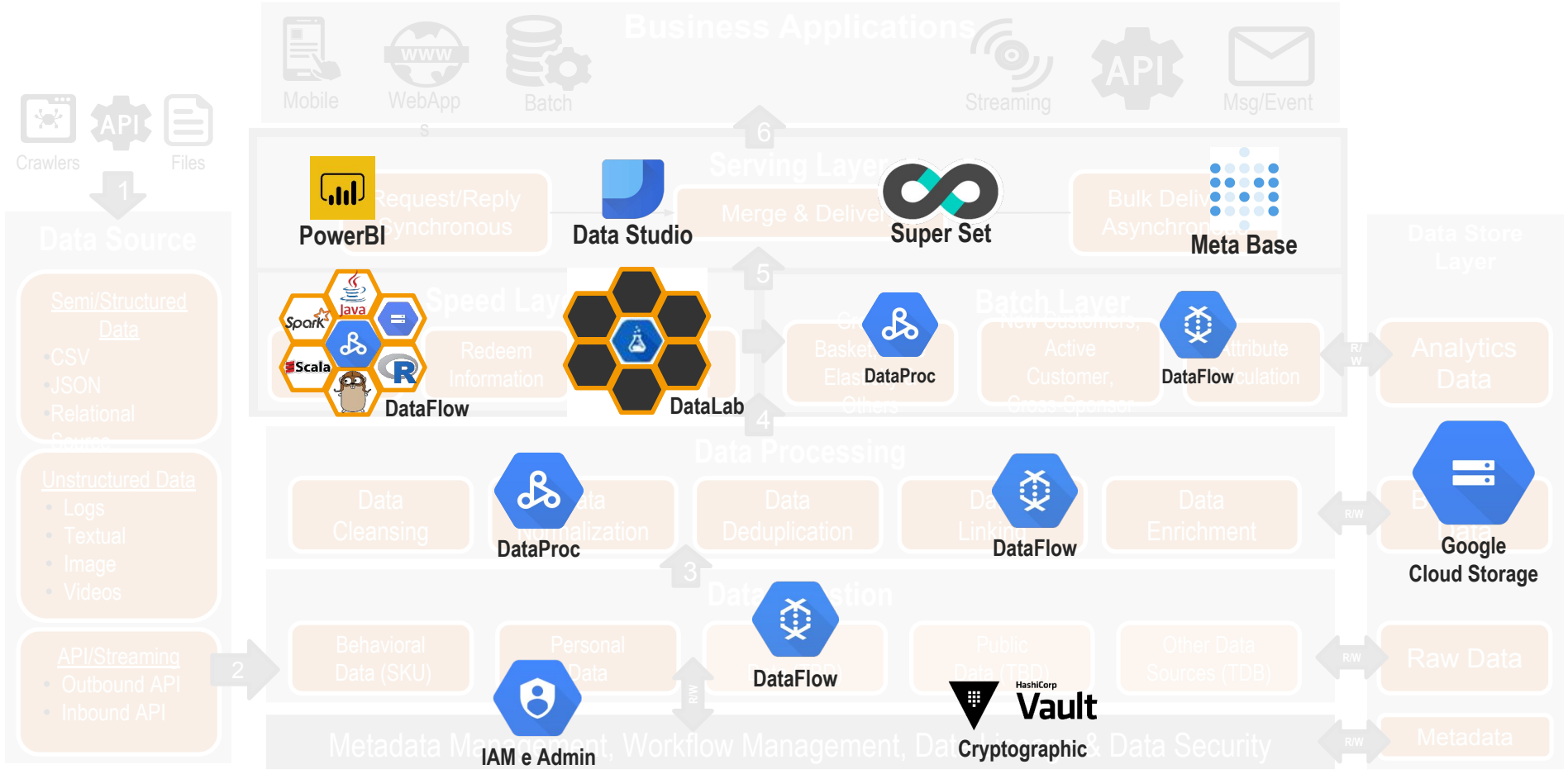
OLTP vs. OLAP

ONLINE TRANSACTION PROCESSING	ONLINE ANALYTICAL PROCESSING
Handles recent operational data	Handles all historical data
Size is smaller, typically ranging from 100 Mb to 10 Gb	Size is larger, typically ranging from 1 Tb to 100 Pb
Goal is to perform day-to-day operations	Goal is to make decisions from large data sources
Uses simple queries	Uses complex queries
Faster processing speeds	Slower processing speeds
Requires read/write operations	Requires only read operations

E na Dotz?

by Jonathas Mendes

Ecossistema Tecnológico **Data Lake**

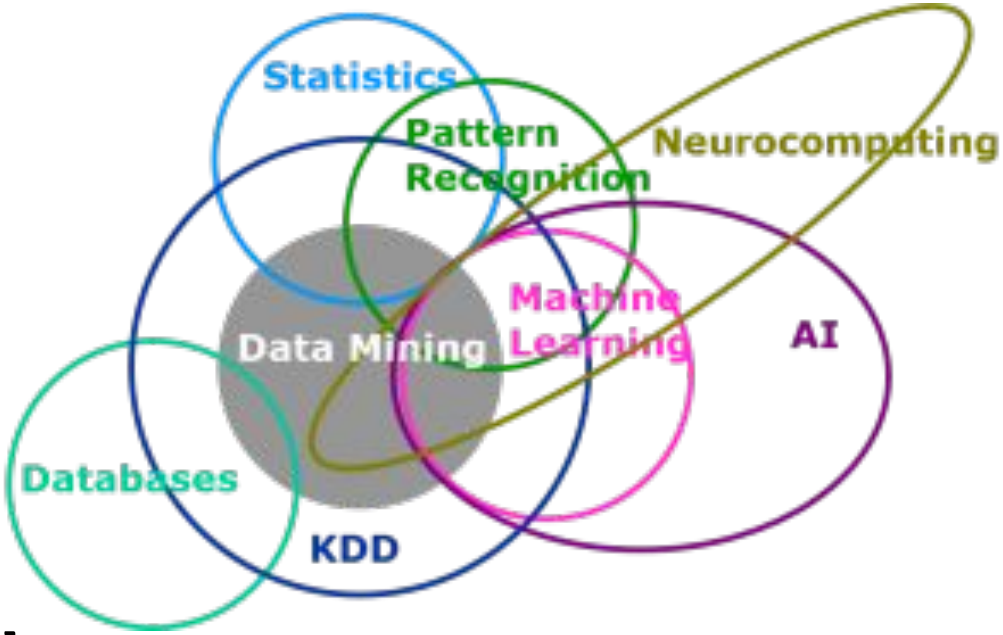


DATA MINING, BIG DATA, DATA SCIENCE...

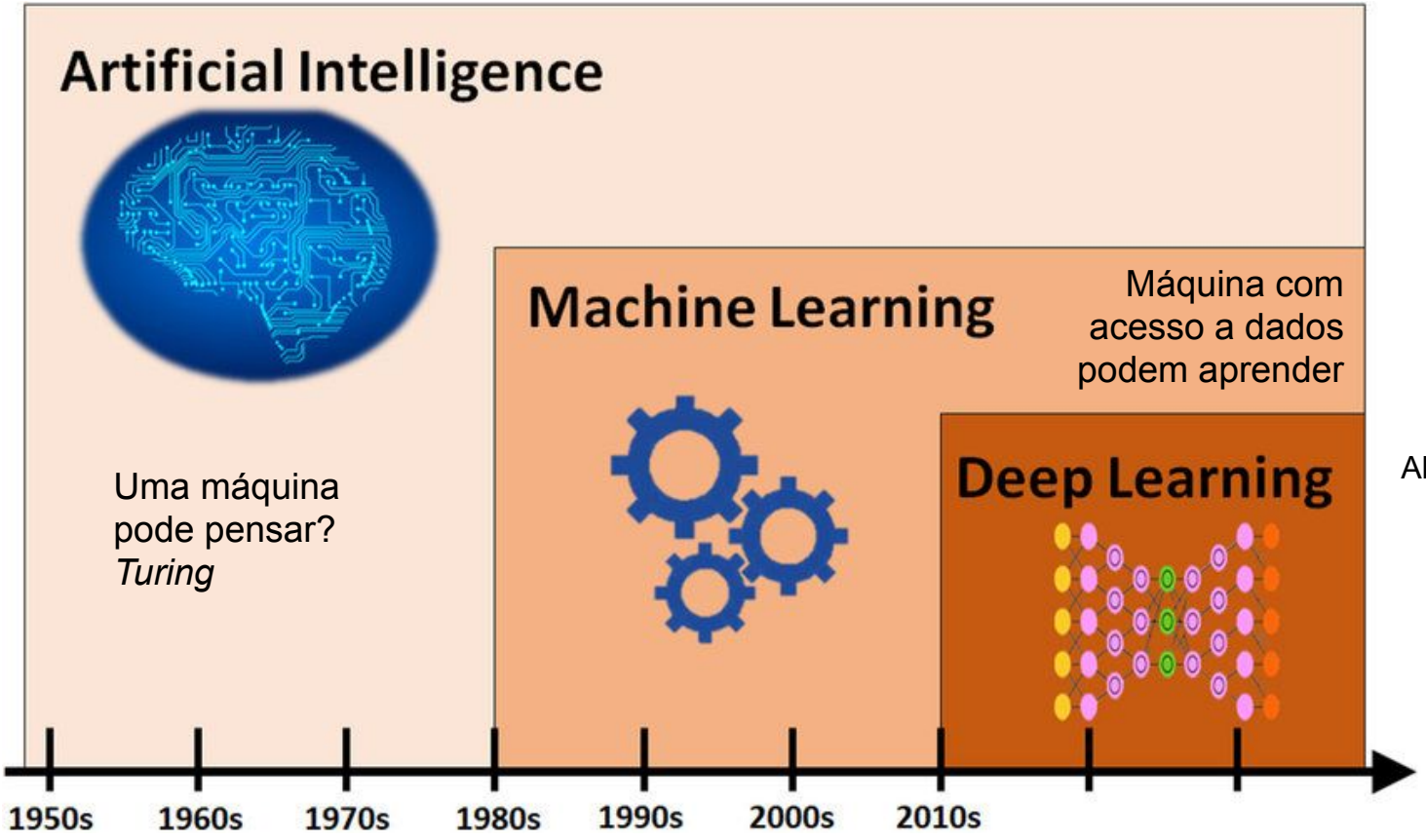
Data Mining ≈ Big Data ≈
Predictive Analytics ≈
Data Science

J. Leskovec, A.Rajaraman, J.Ullman: Mining of Massive Datasets
http://www.mmds.org

DATA MINING, MACHINE LEARNING...



[DEAN, 14]


Artificial Intelligence
Uma máquina pode pensar?
Turing
Machine Learning
Máquina com acesso a dados podem aprender
Deep Learning
Algoritmos que aprendem "como humanos" a resolver problemas complexos
1950s
1960s
1970s
1980s
1990s
2000s
2010s

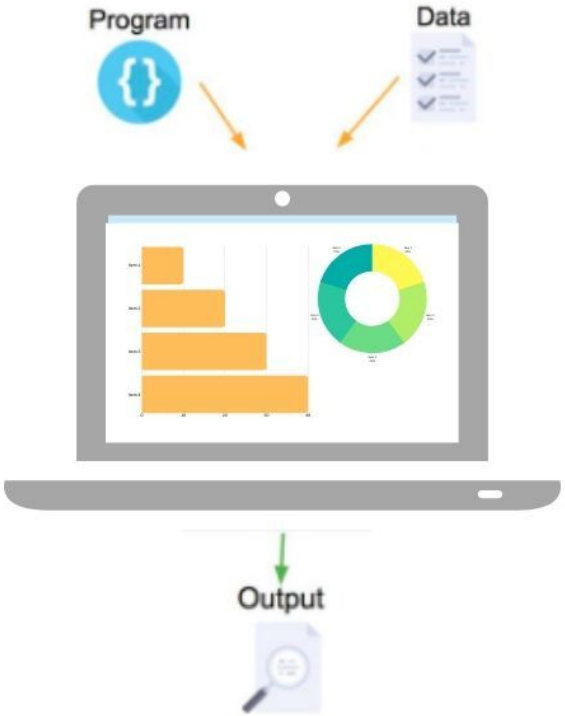
(Procure no Netflix)
Máquinas que aprendem "per si" e "ensinam" humanos

https://thispersondoesnotexist.com/

Geração de dados "sintéticos" - imagens (sons, textos, etc)

GAN, Ian Goodfellow, 2014 paper

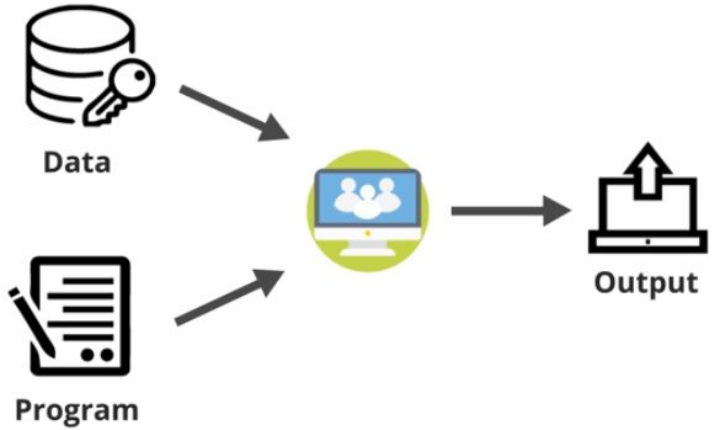
Traditional Programming



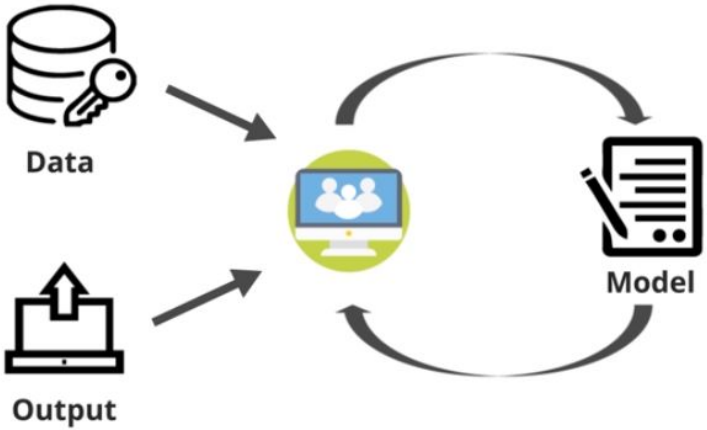
Machine Learning

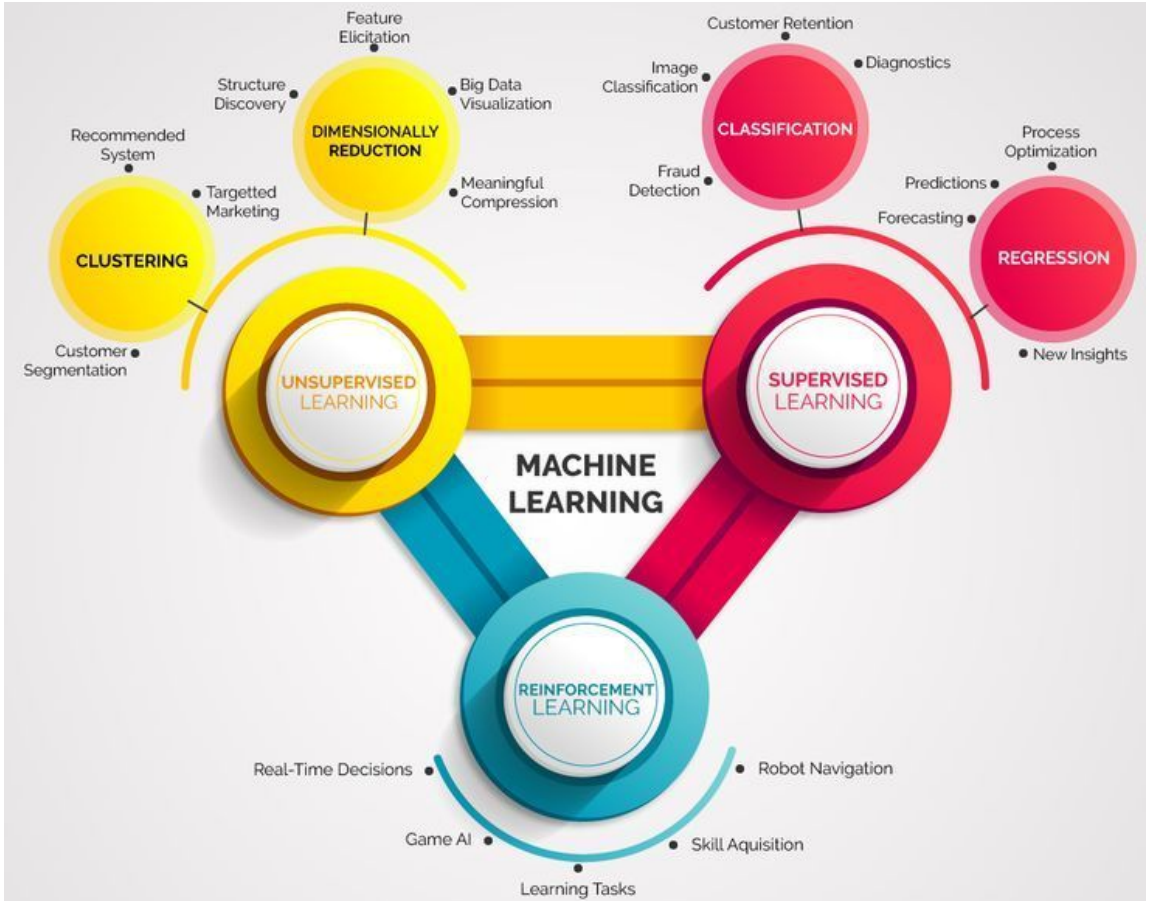
Traditional Approach vs. Machine Learning Approach

Traditional Programming: you code the behavior of the program

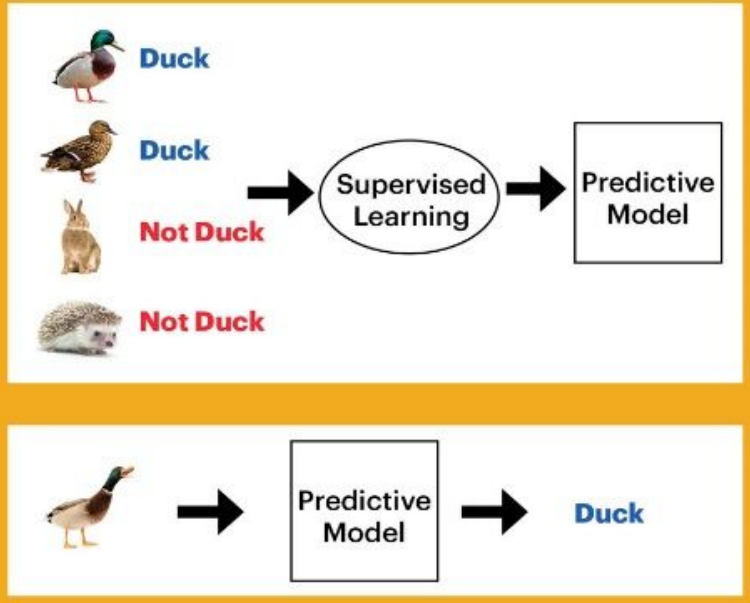


Machine Learning: you leave a lot of that to the machine to learn from data

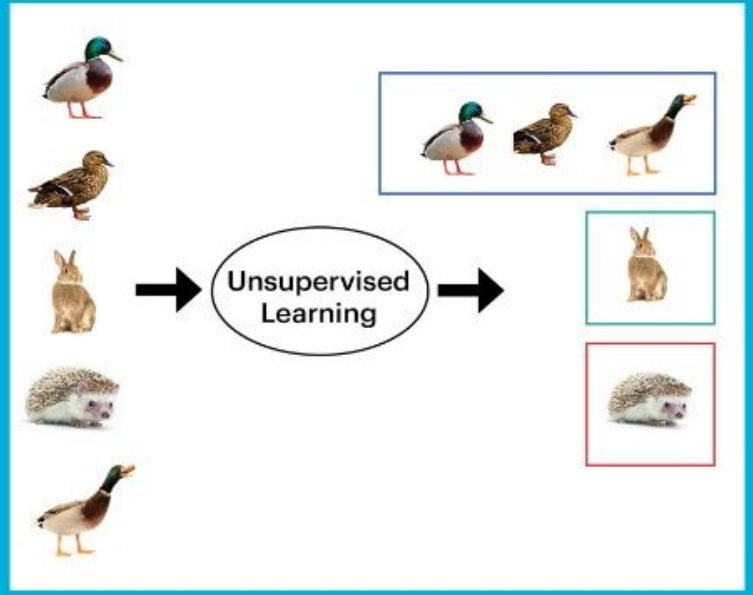


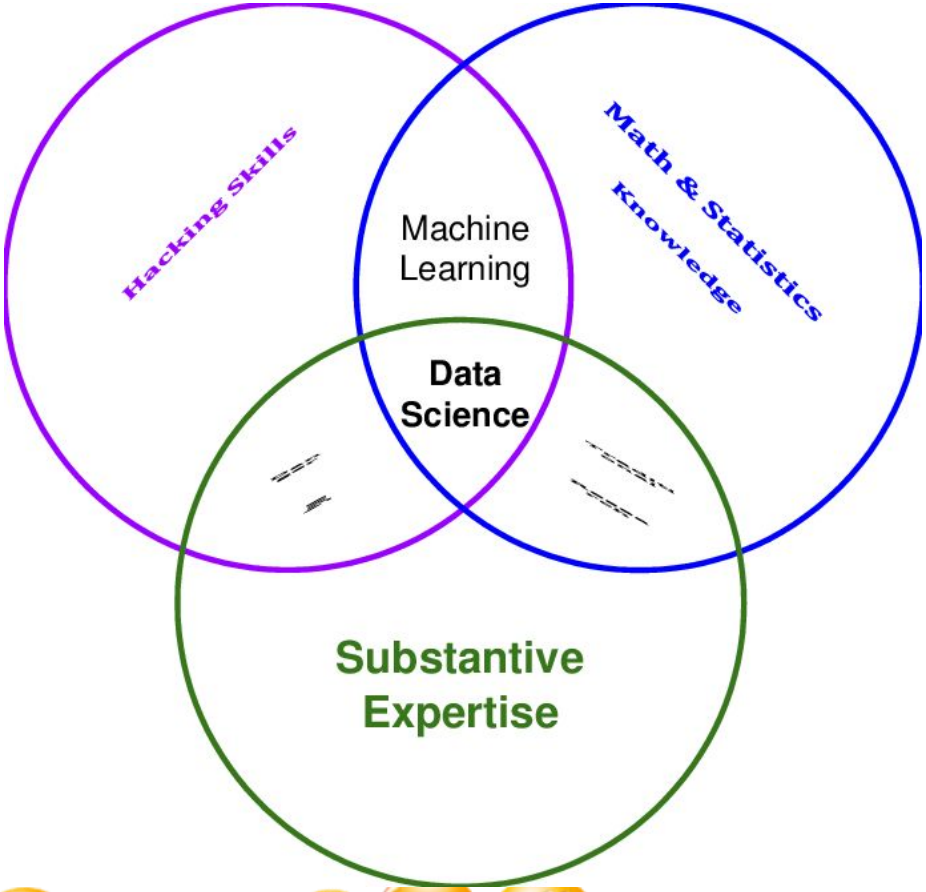

Feature Elicitation
Structure Discovery
Big Data Visualization
DIMENSIONALLY REDUCTION
Meaningful Compression
Recommended System
Targetted Marketing
CLUSTERING
Customer Segmentation
UNSUPERVISED LEARNING
Customer Retention
Image Classification
Diagnostics
CLASSIFICATION
Fraud Detection
Process Optimization
Predictions
Forecasting
REGRESSION
New Insights
SUPERVISED LEARNING
MACHINE LEARNING
REINFORCEMENT LEARNING
Real-Time Decisions
Robot Navigation
Game AI
Skill Aquisition
Learning Tasks

Supervised Learning (Classification Algorithm)

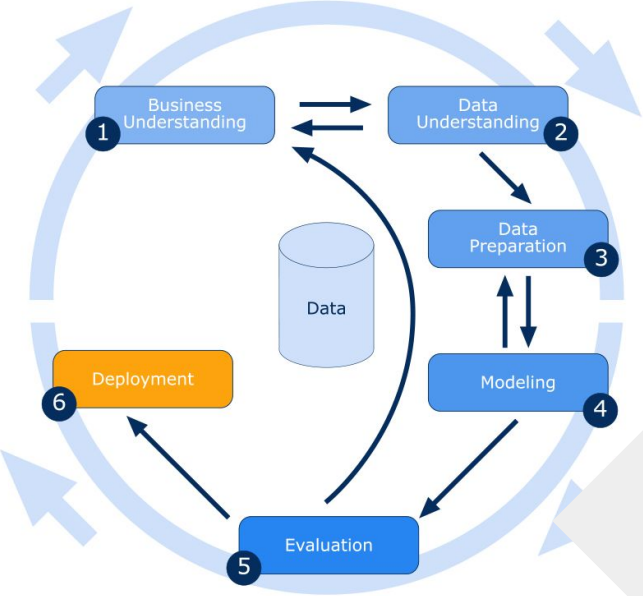


Unsupervised Learning (Clustering Algorithm)




Hacking Skills
Machine Learning
Math & Statistics Knowledge
Data Science
Substantive Expertise

Metodologia Referencial

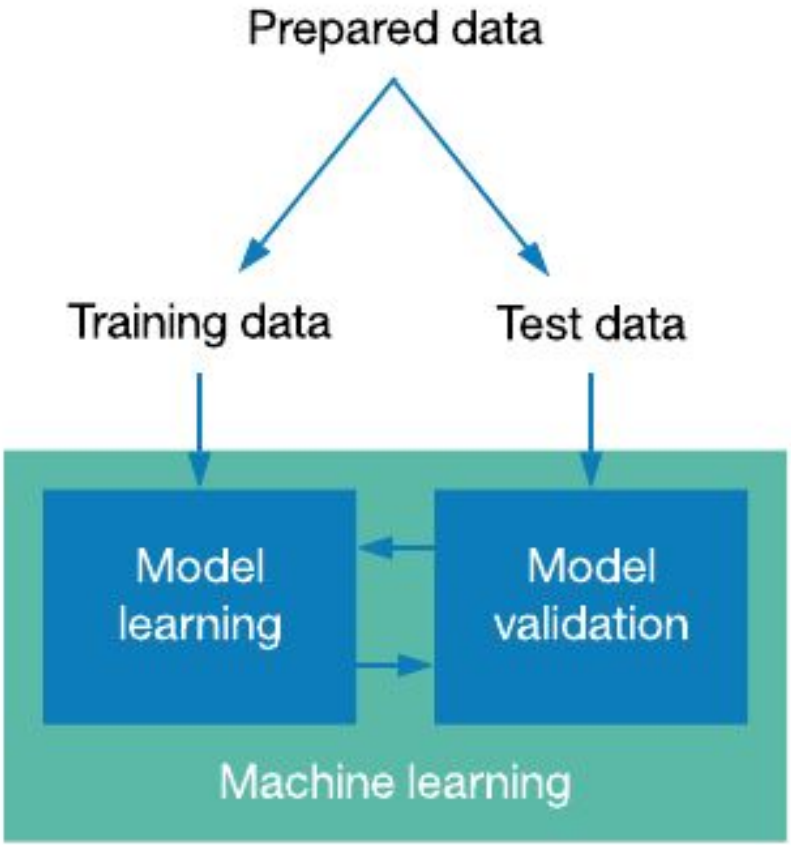


Dados já estão ingeridos?
Features precisam ser preparadas?

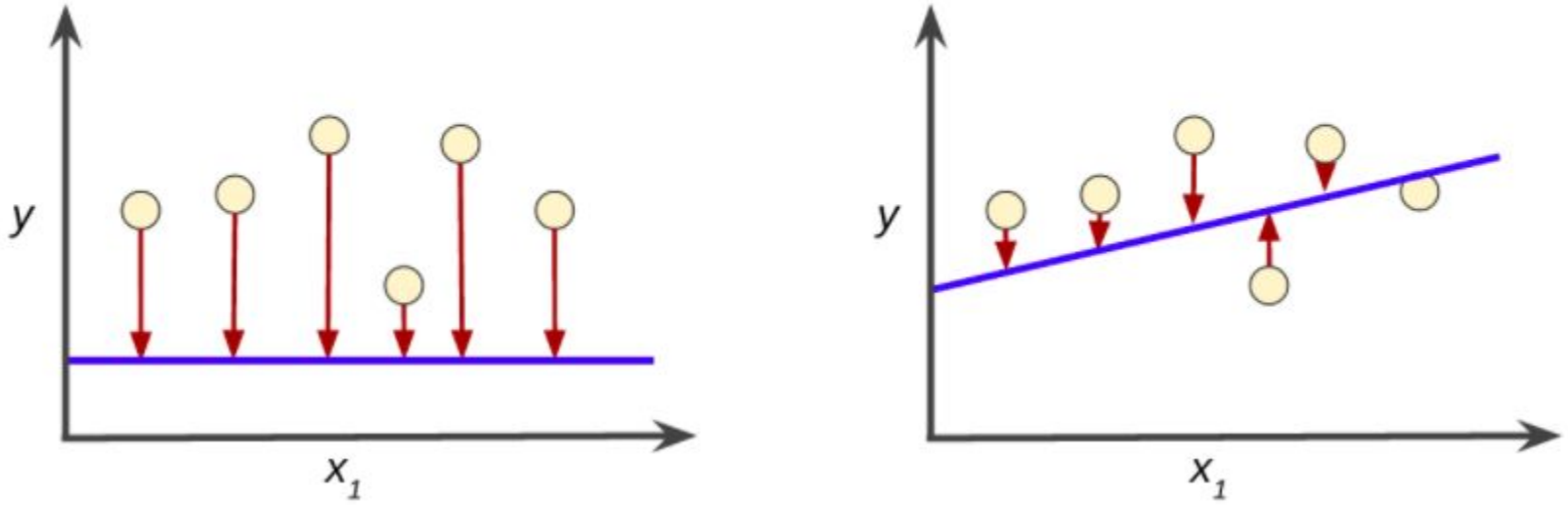
Modelagem e avaliação de modelos

Fonte: CRISP-DM
Outras fontes que versam para frameworks muito parecidos

- KDD (escola Data Mining)
- Jeff Hammerbacher's (ex-facebook, fund. da Cloudera)
- Steps on Data Analysis (John's Hopkin's University)


Prepared data
Training data
Test data
Model learning
Model validation
Machine learning

- The arrows represent loss.
- The blue lines represent predictions.



Fonte: https://developers.google.com/machine-learning/crash-course/descending-into-ml/training-and-loss

Mean square error (**MSE**) is the average squared loss per example over the whole dataset. To calculate MSE, sum up all the squared losses for individual examples and then divide by the number of examples:

$$MSE = \frac{1}{N} \sum_{(x,y) \in D} (y - prediction(x))^2$$

where:

- (x, y) is an example in which
  - x is the set of features (for example, chirps/minute, age, gender) that the model uses to make predictions.
  - y is the example's label (for example, temperature).
- $prediction(x)$ is a function of the weights and bias in combination with the set of features x.
- D is a data set containing many labeled examples, which are (x, y) pairs.
- N is the number of examples in D.

Although MSE is commonly-used in machine learning, it is neither the only practical loss function nor the best loss function for all circumstances.

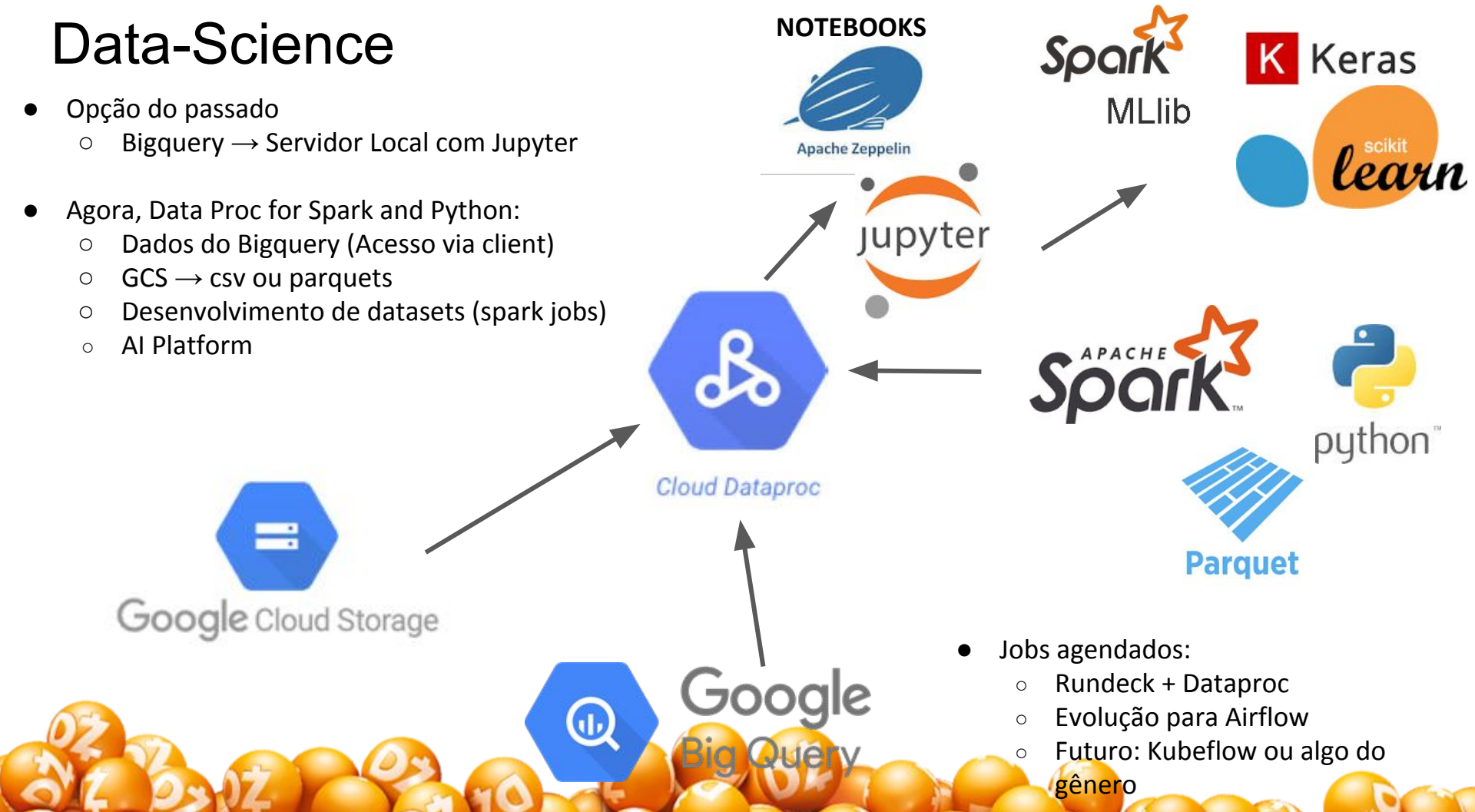
Fonte: https://developers.google.com/machine-learning/crash-course/descending-into-ml/training-and-loss

e daqui pra frente?

Data-Science

- Opção do passado
  - Bigquery → Servidor Local com Jupyter

- Agora, Data Proc for Spark and Python:
  - Dados do Bigquery (Acesso via client)
  - GCS → csv ou parquets
  - Desenvolvimento de datasets (spark jobs)
  - AI Platform



- Jobs agendados:
  - Rundeck + Dataproc
  - Evolução para Airflow
  - Futuro: Kubeflow ou algo do gênero

Muito obrigado!

https://github.com/diogenesjusto/dotz-mini-curso-ingestao-ml