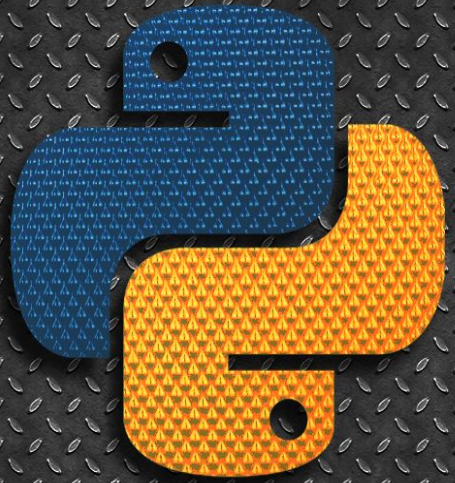
Machine Learning in Python

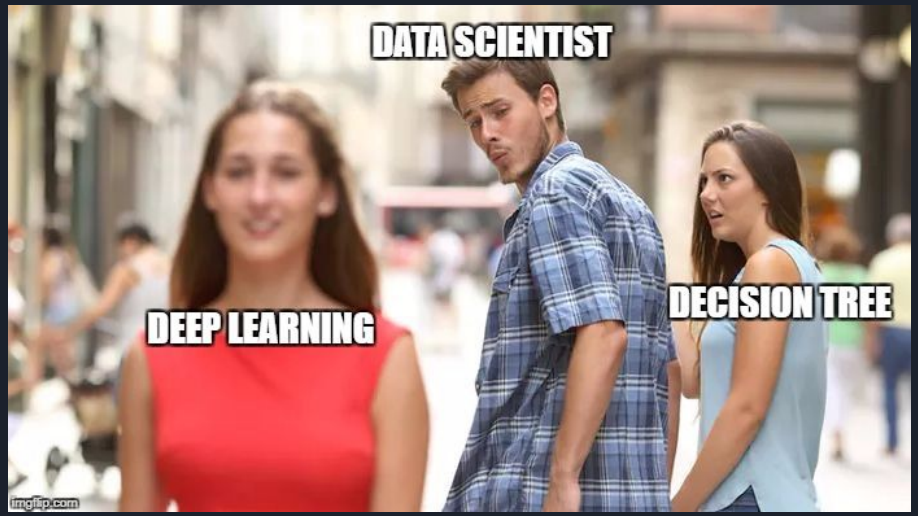


BDAA

BIG DATA & ANALYTICS ASSOCIATION

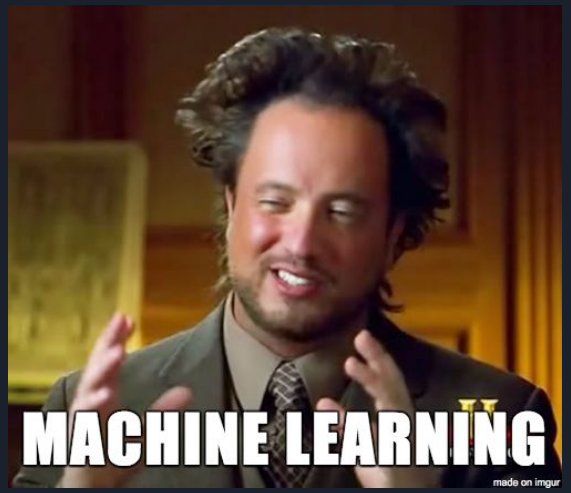
Outline

- Foundational Machine Learning Concepts
- Machine Learning!
  - Decision Trees
  - Decision Tree - Exercise
  - Random Forests
- Questions



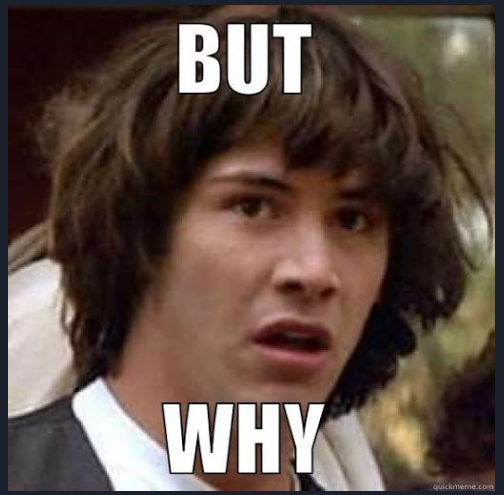
What is Machine Learning (ML)?

- Computers "learning" from data
- "Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed."
    - https://expertsystem.com/machine-learning-definition/

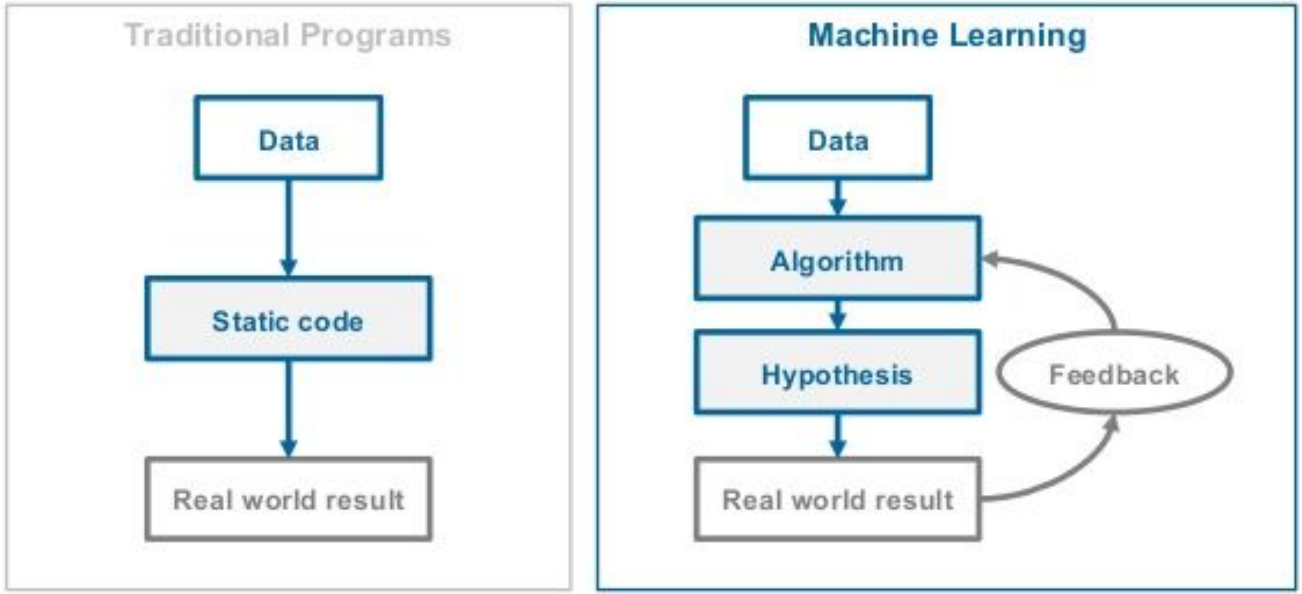


Why use Machine Learning?

- You don't have to explicitly program something
  - ML is powered by data - not instructions
- When problems are much too complex to be explicitly programmed
- When the environment changes over time



Traditional Programs vs. Machine Learning



IBM

Applications of Machine Learning

Sarah Michelle Gellar
74 mutual friends

Stephen King
13 mutual friends

AI vs. ML vs. DL

Artificial Intelligence

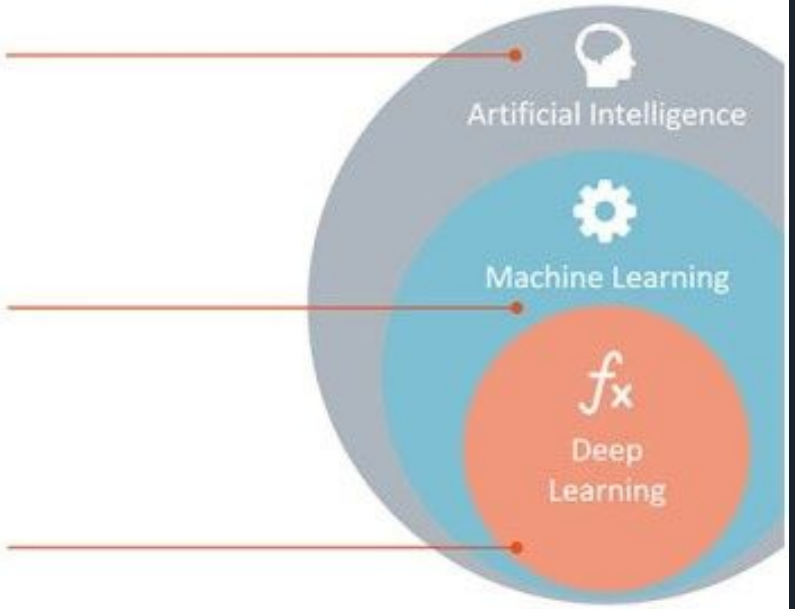
Any technique which enables computers to mimic human behavior.

Machine Learning

Subset of AI techniques which use statistical methods to enable machines to improve with experiences.

Deep Learning

Subset of ML which make the computation of multi-layer neural networks feasible.



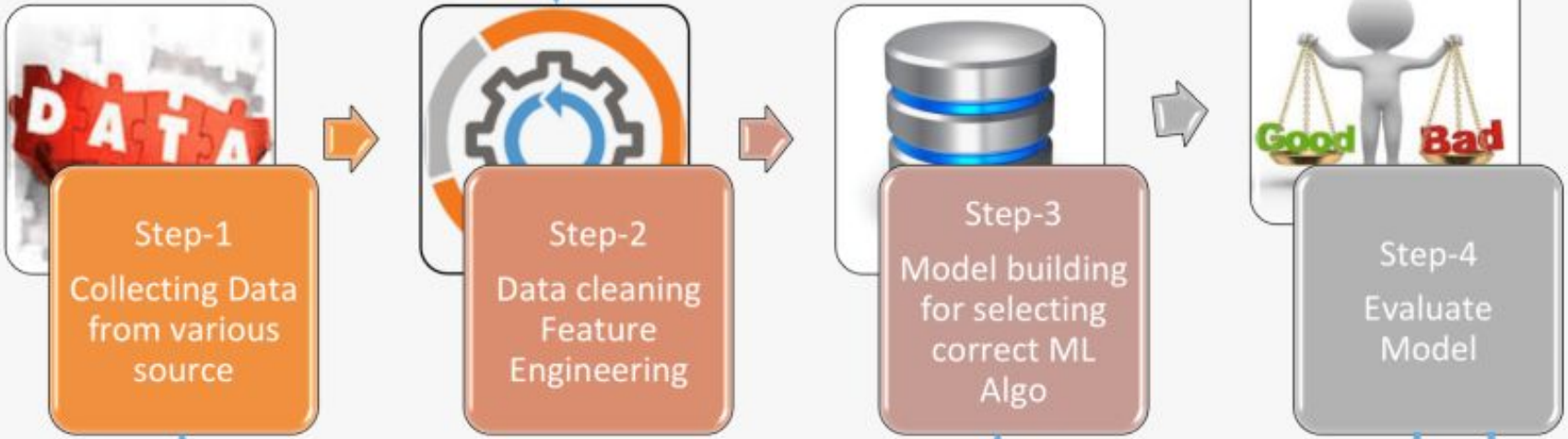
https://rapidminer.com/blog/artificial-intelligence-machine-learning-deep-learning/

ML Terminology

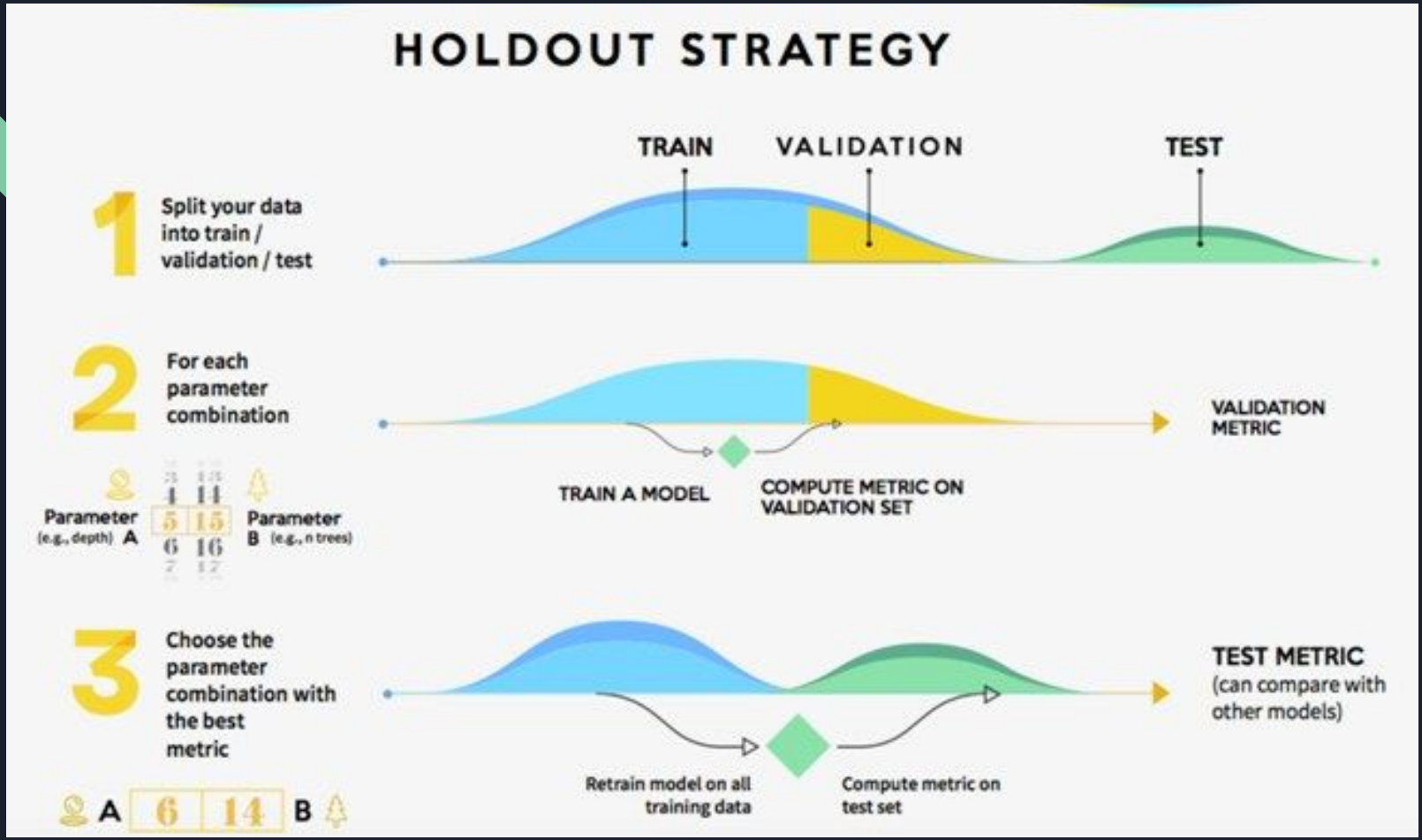
- Training Data: the examples that the system uses to learn
- Features: The input "variables"
- Labels: The value(s) you're trying to predict given the features
- Algorithm: The model that is trained and used to predict the label(s)

ML Process

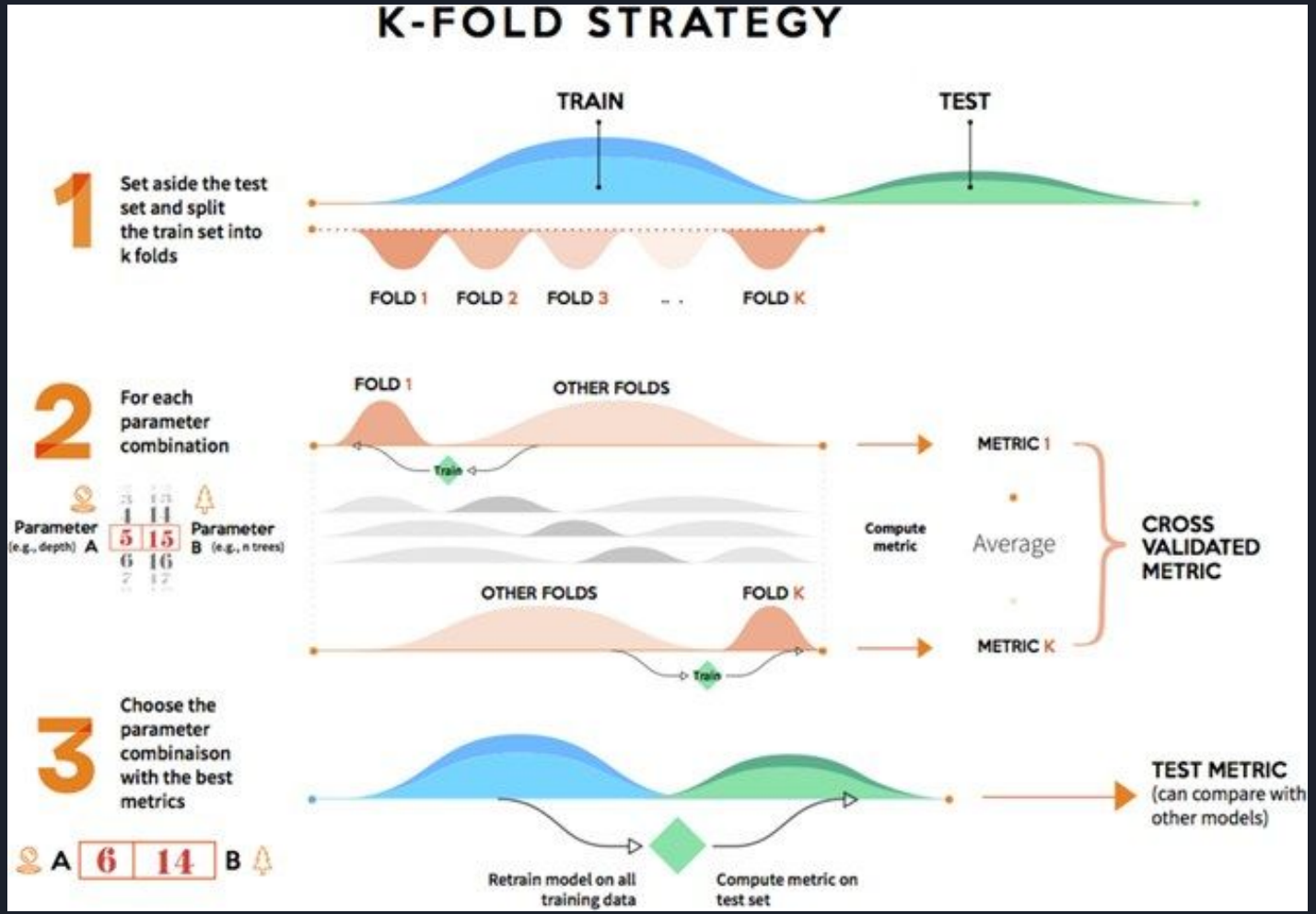
Machine Learning Process at High Level



AILabPage Source - Open Internet various sources Image Source - https://vinodsblog.com via @vinod1975

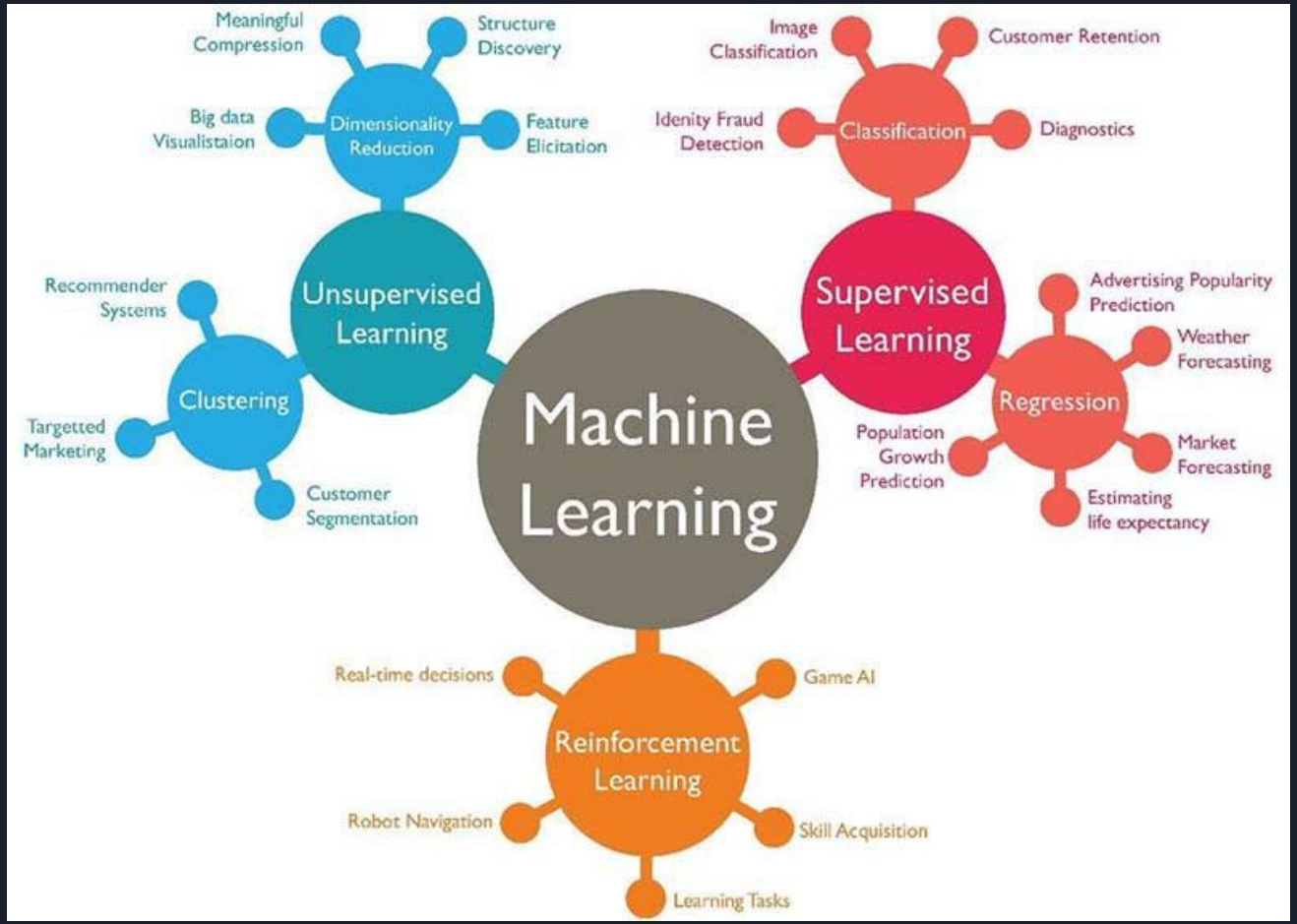


https://www.kdnuggets.com/2017/08/dataiku-predictive-model-holdout-cross-validation.html



https://www.kdnuggets.com/2017/08/dataiku-predictive-model-holdout-cross-validation.html

Types of ML


Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

Types of Machine Learning

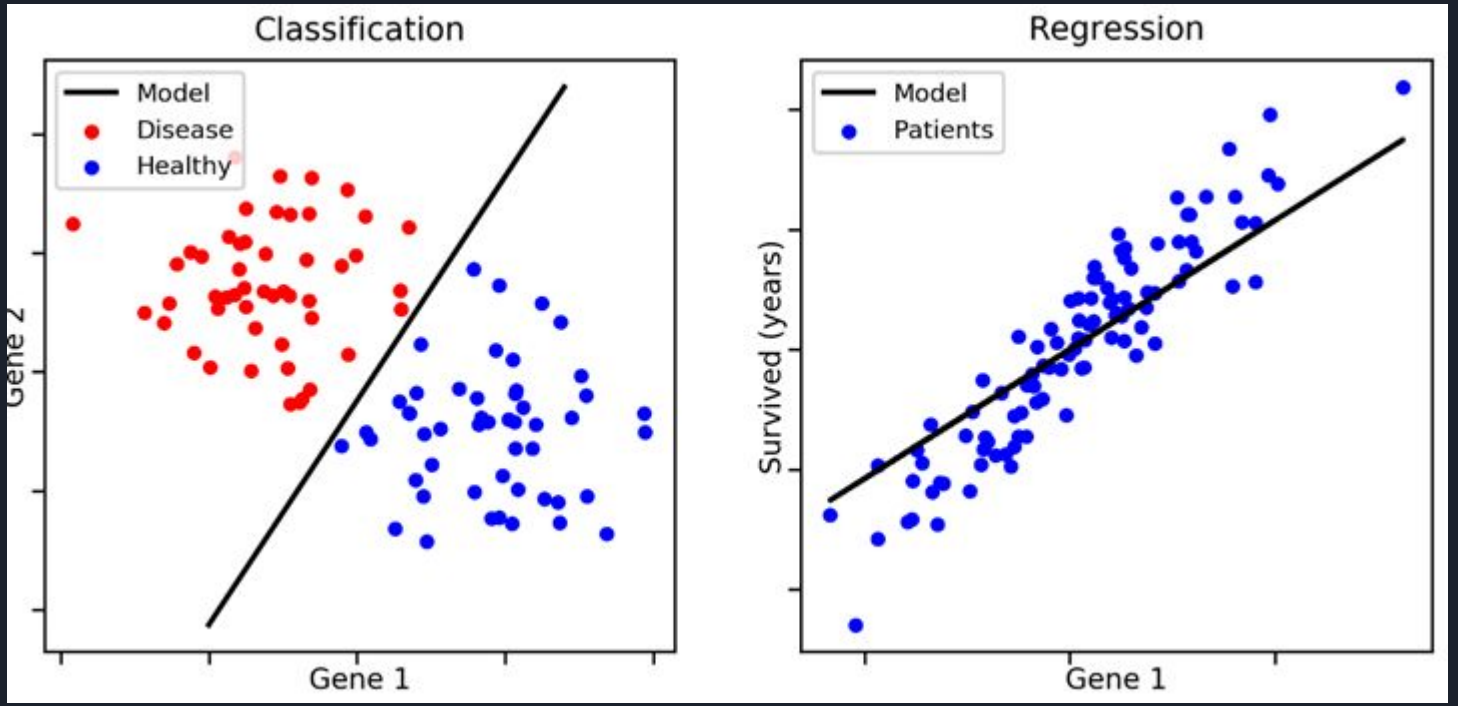
- Supervised Learning
  - Labeled Training Data ("right" answer)
  - Predict outcome
  - Ex: Classification, Regression
- Unsupervised Learning
  - Unlabeled Training Data (no "right" answer, or "right" answer is unknown)
  - Find structure and patterns in data
  - Ex: Clustering, Anomaly Detection
- Reinforcement Learning
  - Machine Agent explores environment
  - Learning is based on rewards and punishments

Common ML Tasks

- Supervised:
  - Classification
  - Regression

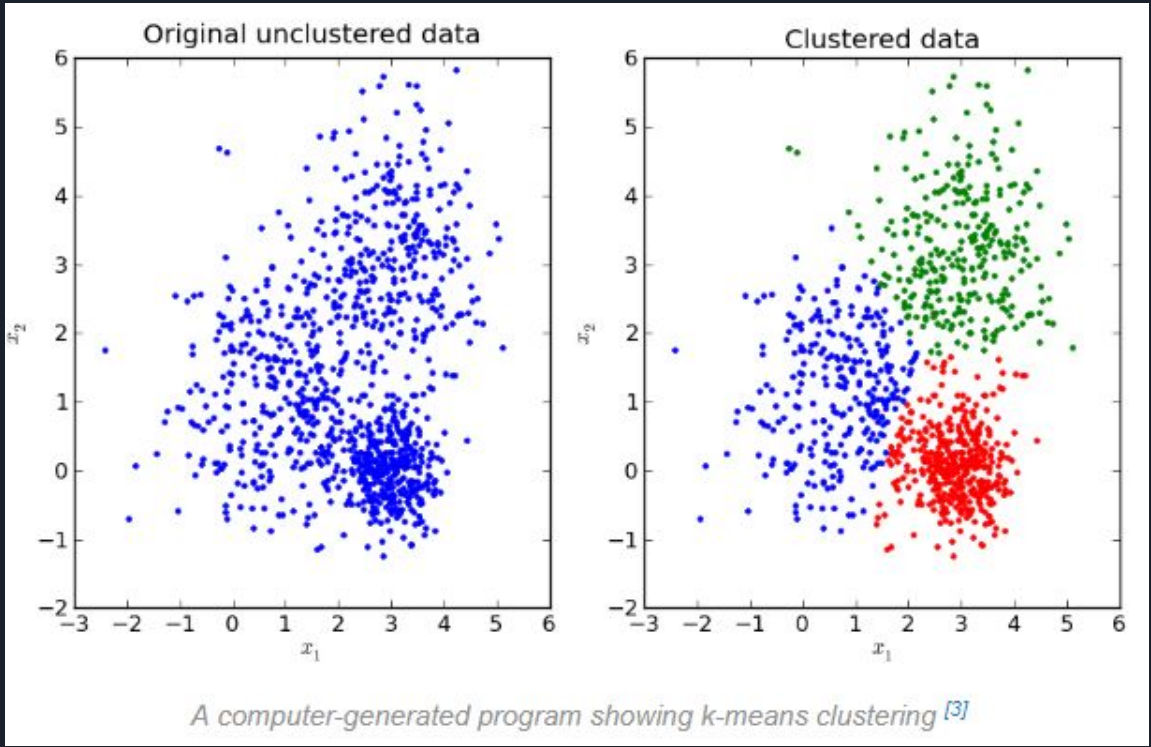
- Unsupervised:
  - Clustering
  - Anomaly Detection
  - Association Rule Learning

Common Supervised ML Tasks: Classification vs. Regression



https://medium.freecodecamp.org/using-machine-learning-to-predict-the-quality-of-wines-9e2e13d7480d

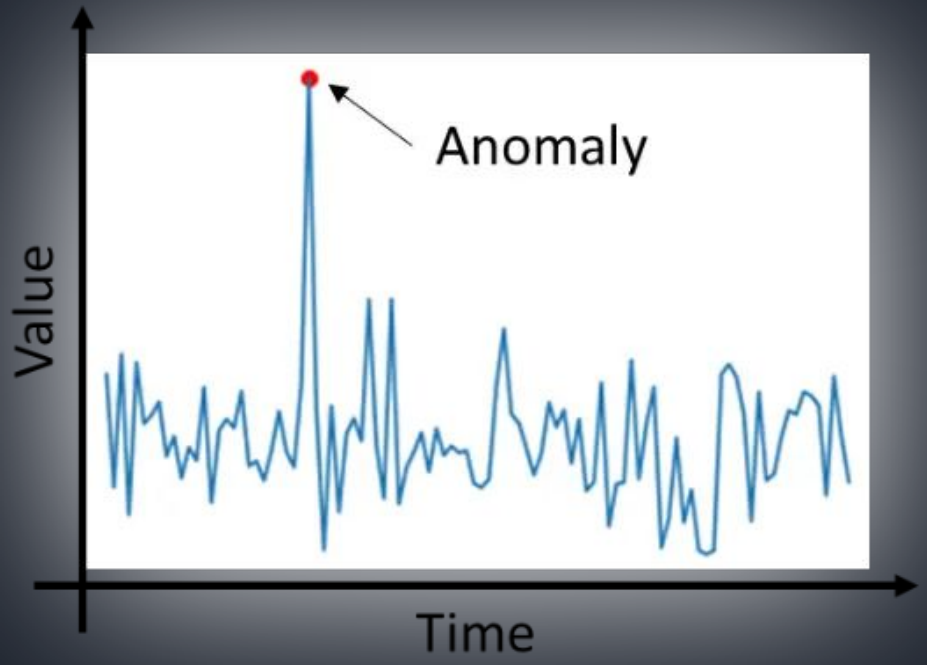
Common Unsupervised ML Tasks: Clustering



A computer-generated program showing k-means clustering [3]

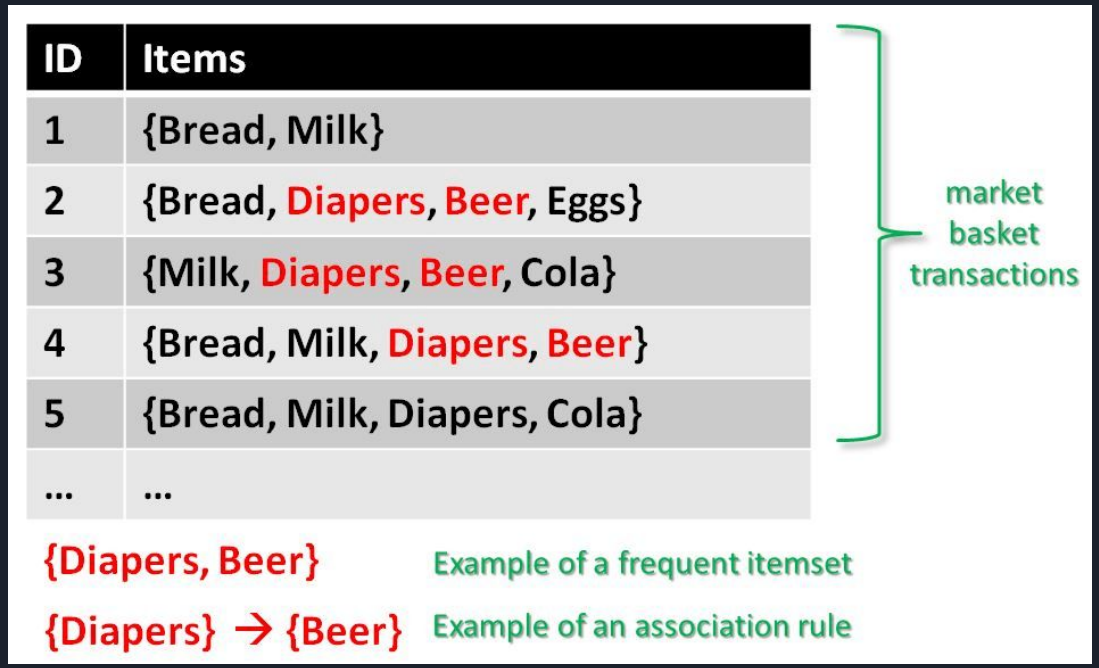
https://brilliant.org/wiki/k-means-clustering/

Common Unsupervised ML Tasks: Anomaly Detection



https://www.engineering.com/AdvancedManufacturing/ArticleID/19058/Anomaly-Detection-Industrial-Asset-Insights-Without-Historical-Data.aspx

Common Unsupervised ML Tasks: Association Rule Mining

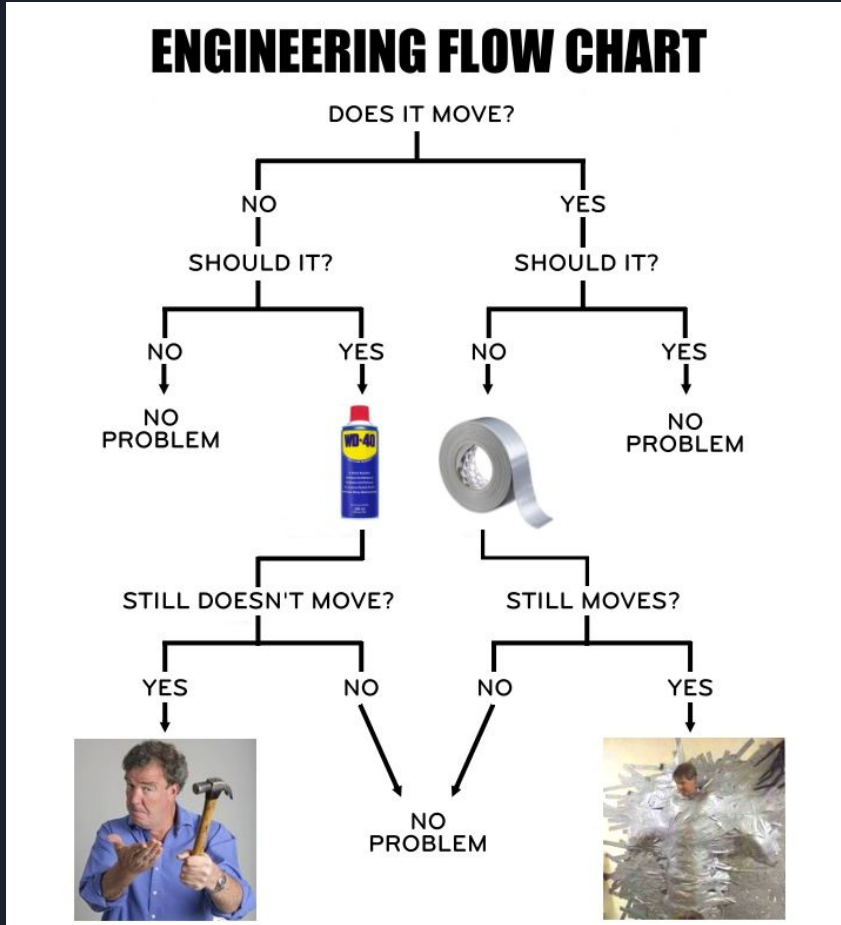


https://www.datacamp.com/community/tutorials/market-basket-analysis-r

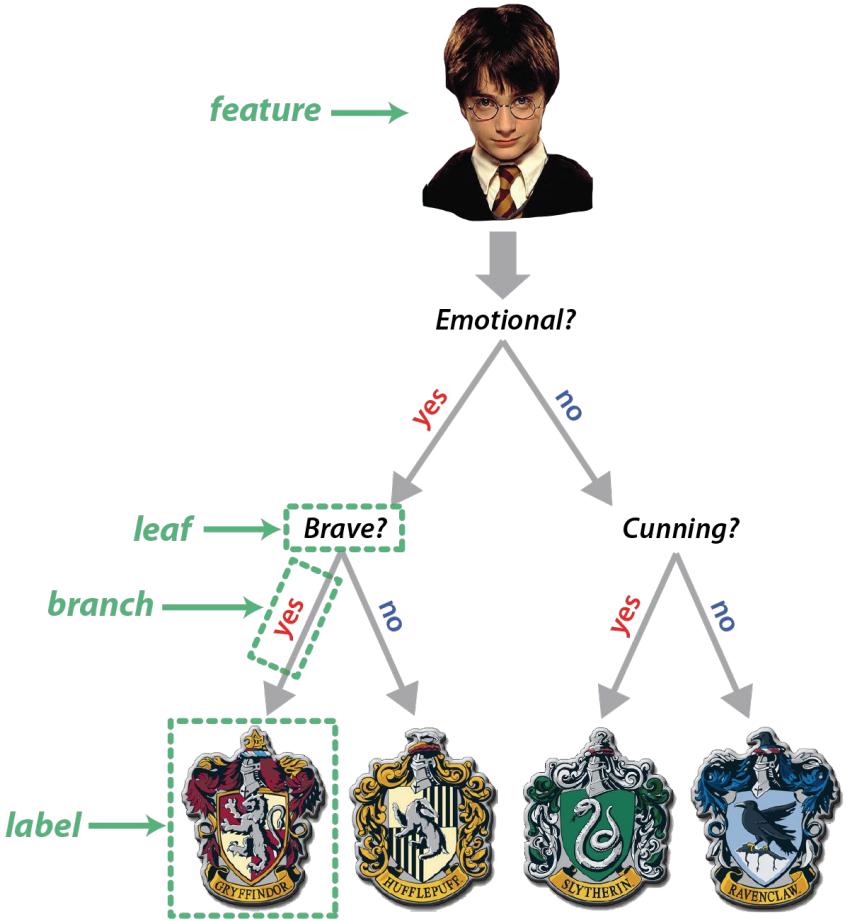
Decision Trees

Decision Trees

- Series of Q's
- Versatile: Classification or Regression
- Powerful: Can fix complex datasets
- "White-box"
- Few assumptions about training data
- Backbone of Random Forests



reddit.com/r/thegrandtour/comments/5yl6el/engineers_flow_chart/



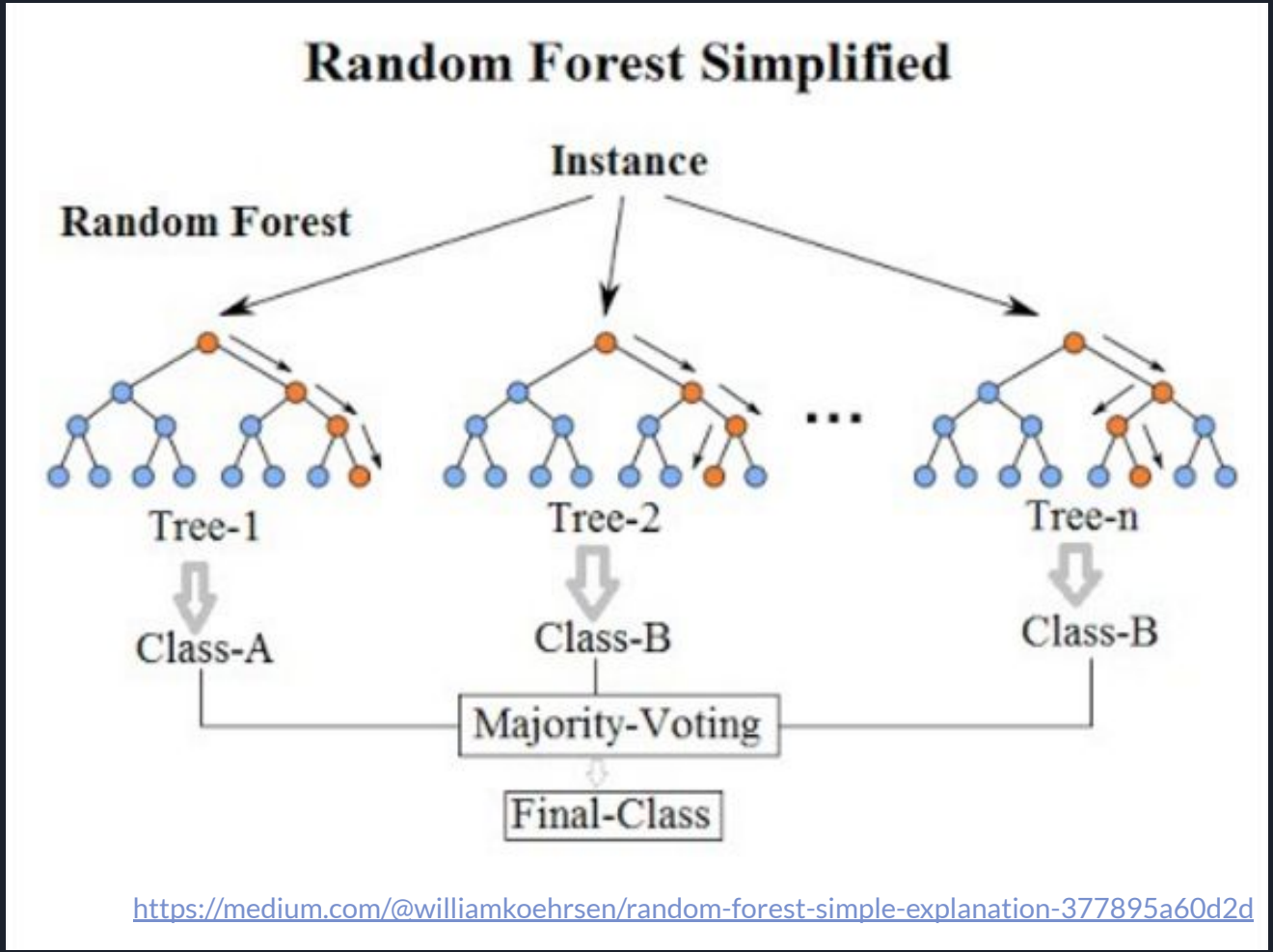
https://medium.com/@tmckenzie.nz/learning-machine-learning-at-hogwarts-55e6ee85eef7

Ensemble Learning

- A group of predictors
  - AKA training more than one model and combing the predictions of them

Random Forests (RF)

- Each decision tree in the RF:
  - has random subset of features
  - uses random set of the training data points
- Final prediction is a majority vote for the predicted class



https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d

Further Resources

- Online Classes:
  - Coursera
- Books:
  - *Data Science for Business* - Foster Provost and Tom Fawcett
  - *Hands-On Machine Learning with Scikit-Learn and TensorFlow* - Aurélien Géro
- Internet:
  - Youtube: Siraj Raval
- Classes at OSU (have prereqs):
  - CSE 5243: Introduction to Data Mining (3 cr hrs)
  - CSE 3521: Survey of Artificial Intelligence I: Basic Techniques (3 cr hrs)
  - CSE 5523: Machine Learning and Statistical Pattern Recognition (3 cr hrs)
  - CSE 5524: Computer Vision for Human-Computer Interaction (3 cr hrs)
  - CSE 5526: Introduction to Neural Networks (3 cr hrs)
- Projects:
  - Kaggle

Questions?

- Any Questions?

- My Contact Info:
  - Leo Glowacki
  - Message me on Slack!
  - www.LeoGlowacki.com