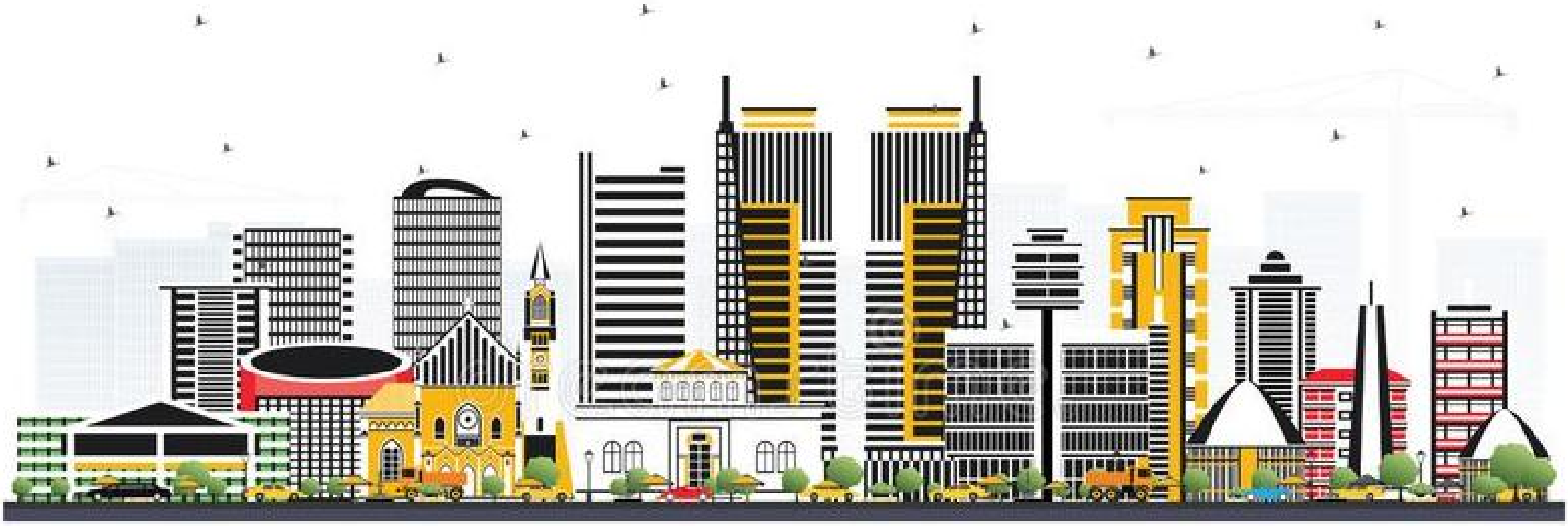
pycon
tanzania



DATA SCIENCE AND MACHINE LEARNING LIBRARIES

Bio

Zephania Reuben

The University of Dodoma

College of Informatics and Virtual Education

10th December, 2019

pycon tanzania

Introdution

Problem.

You need to predict how much user “A” will like a movie that she hasn't seen based on her ratings of movies that she has seen.

Traditional Methods

Machine Learning

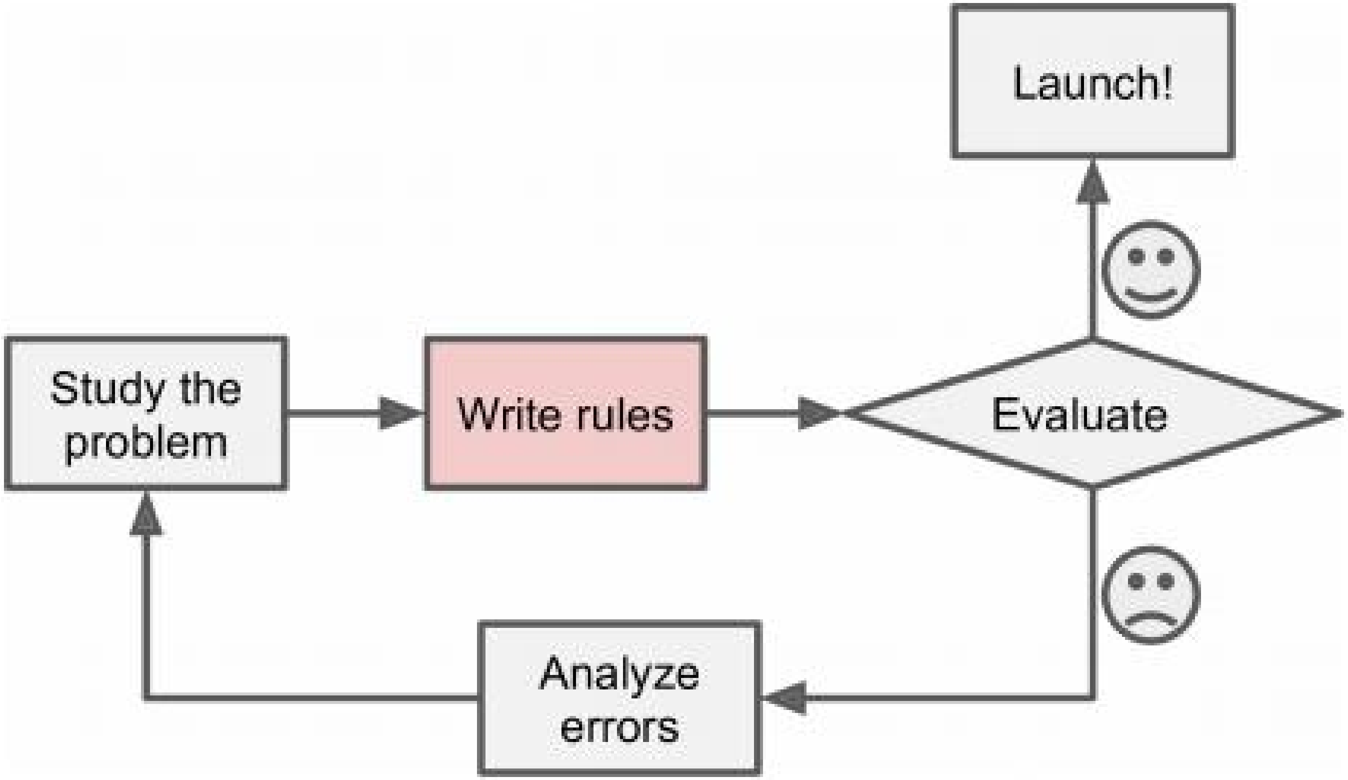
Traditional Method

Complex rules.

Hard to maintain

pycon tanzania

Traditional Method



Machine Learning

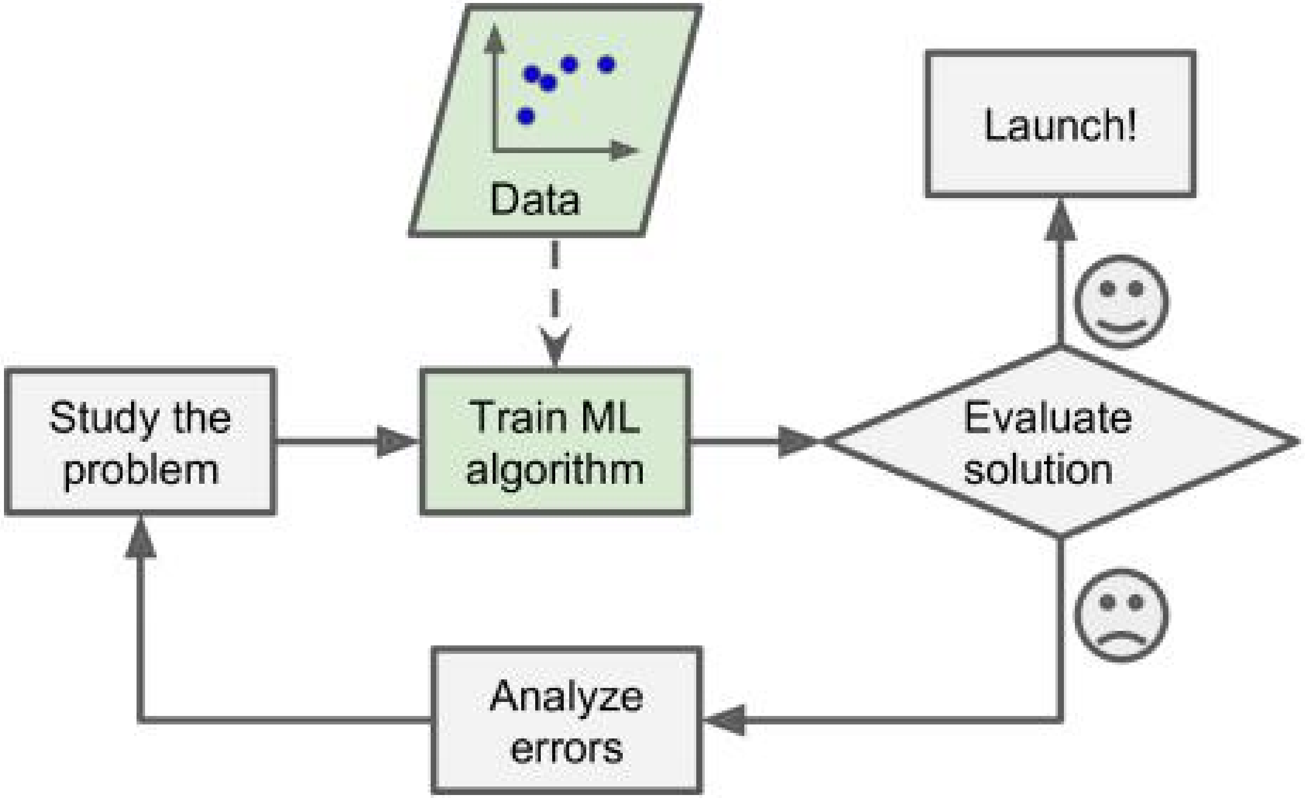
Automatic pattern learning

Easy to maintain

Adopt to changes

More accurate

Machine Learning



Machine Learning

What Does it Mean to Learn?

In Machine Learning an important concept is the ability to generalize.

Machine Learning

A computer program is said to learn from experience E with respect to some task T and some performance P, if its performance on T, as measured by P, improves with experience E.
- Tom Mitchell, 1997.

Task

Performance

Experience

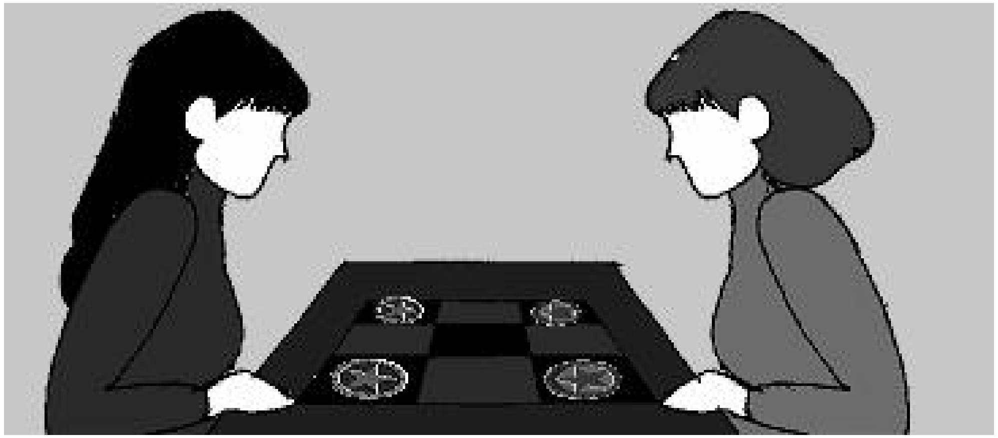
Machine Learning

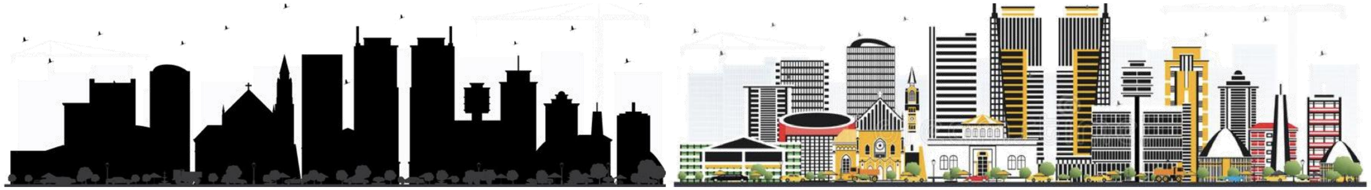
Checker Learning Problem

Task **T** : Playing Checker.

Experience **E**: Playing practice game against itself.

Performance Measure **P**: % of games won against opponents.





Machine Learning

Types of Machine Learning

Supervised

Reinforcement

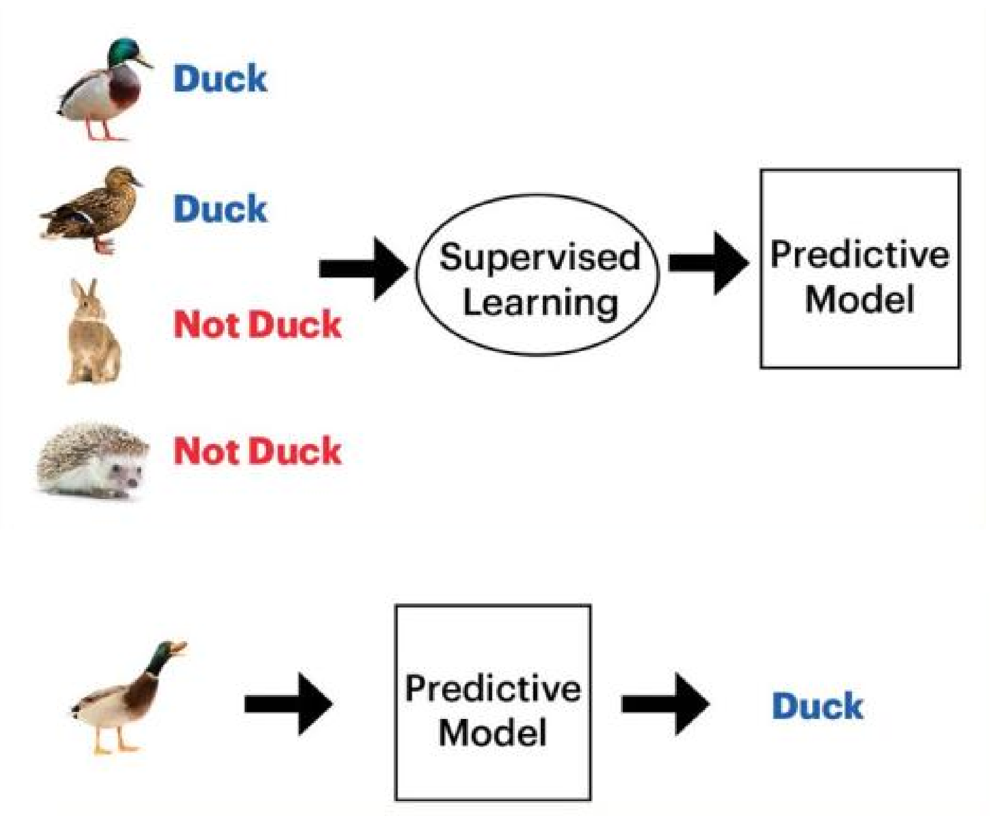
Semi-supervised

Unsupervised

ML Algorithms

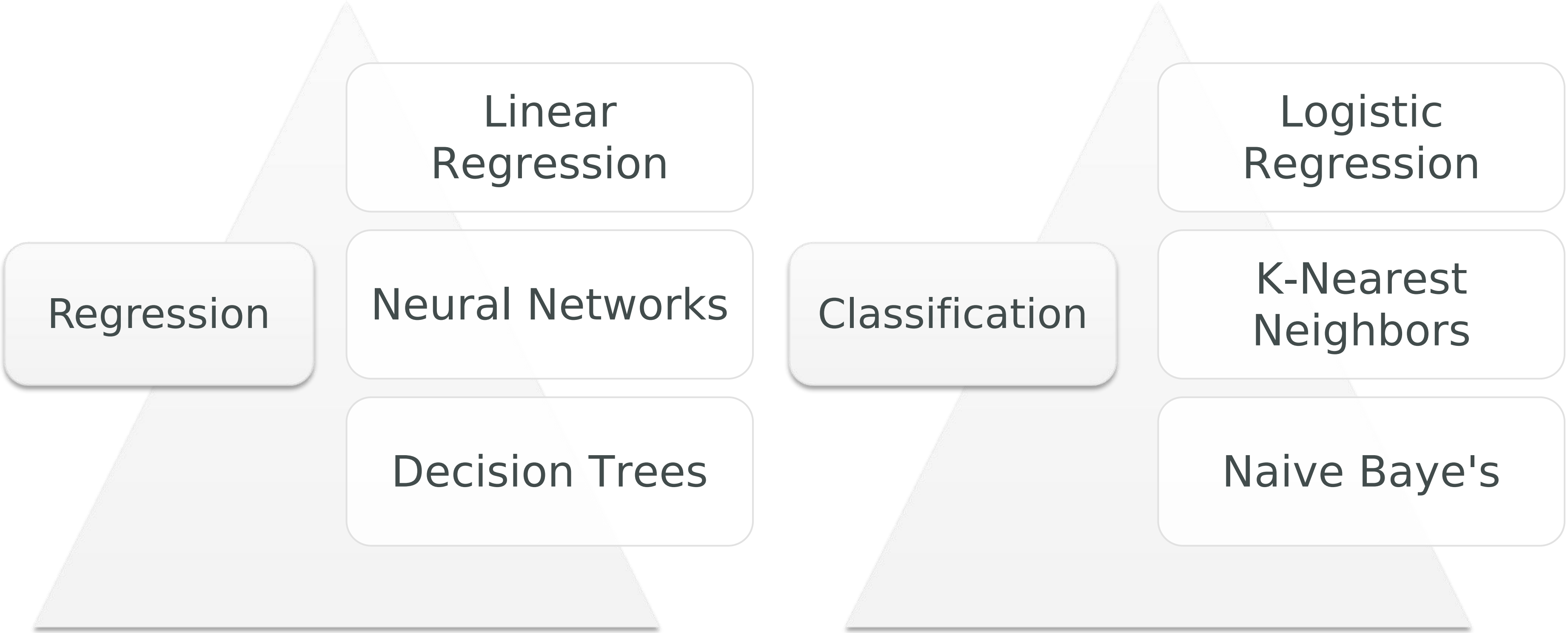
Supervised Machine Learning Algorithms

Training data includes the desired solutions called labels.



ML Algorithms

Supervised Machine Learning Algorithms.



Regression

Linear Regression

Neural Networks

Decision Trees

Classification

Logistic Regression

K-Nearest Neighbors

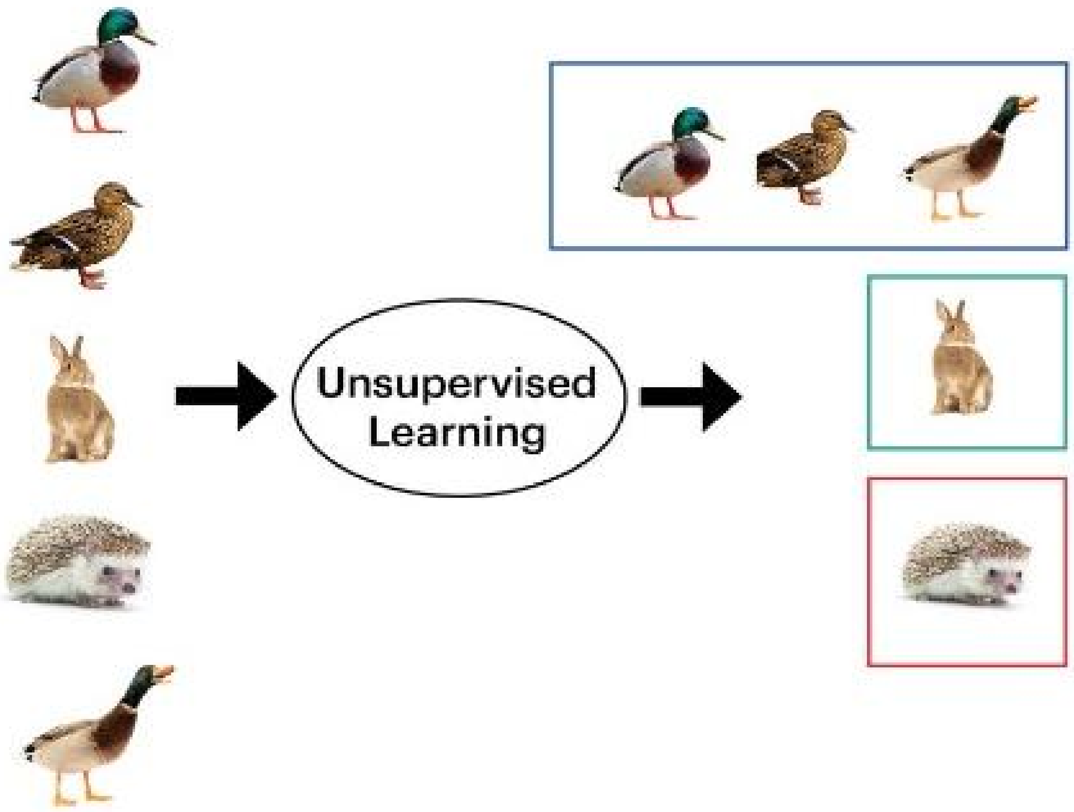
Naive Baye's

pycon tanzania

ML Algorithms

Unsupervised Machine Learning Algorithms

They only extracts pattern from the provided data during learning.



ML Algorithms

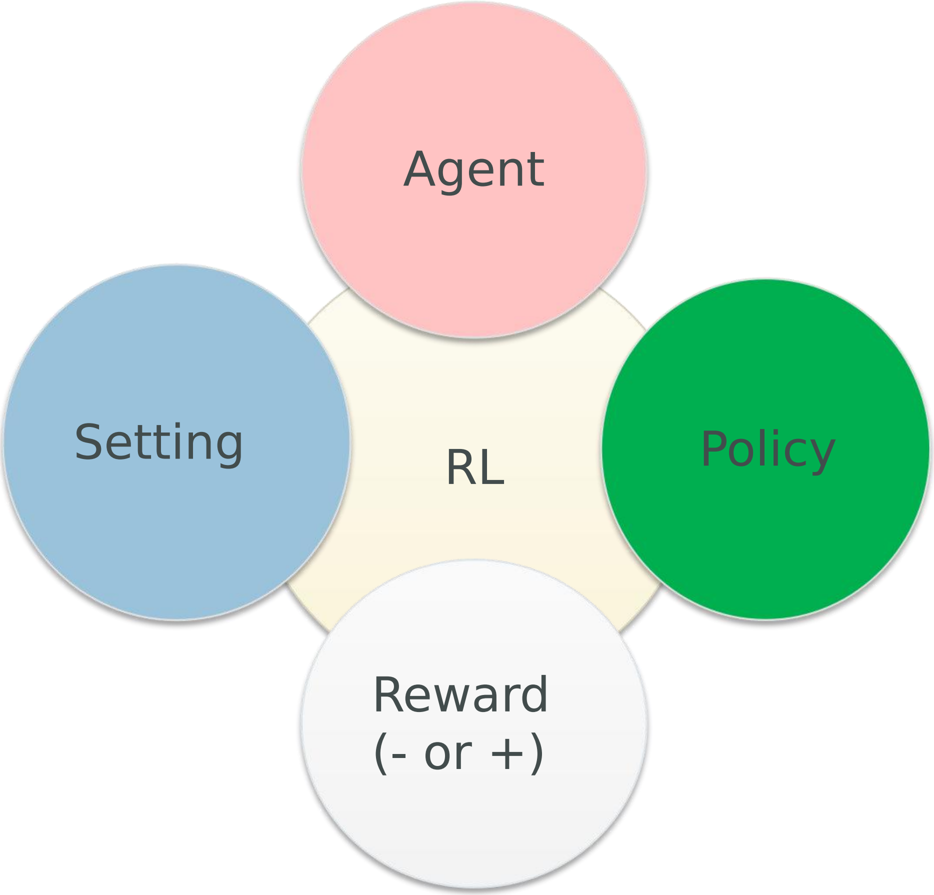
Unsupervised Machine Learning Algorithms

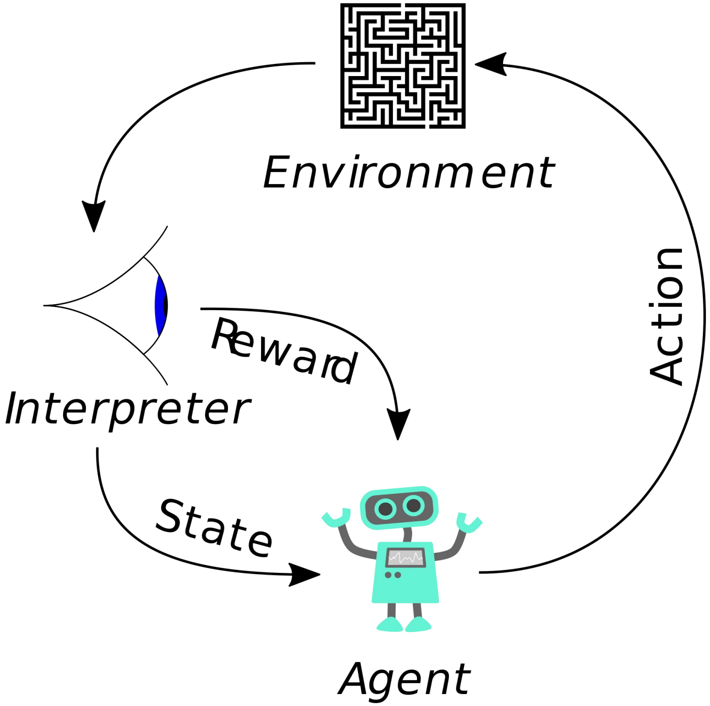
They only extracts pattern from the provided data during learning.

- Clustering
- Anomaly Detection
- Dimensionality Reduction

ML Algorithms

Reinforcement Learning Algorithm





ML Algorithms

Python Libraries for DS and ML.

matplotlib

NumPy

Seaborn

Pandas

scikit learn

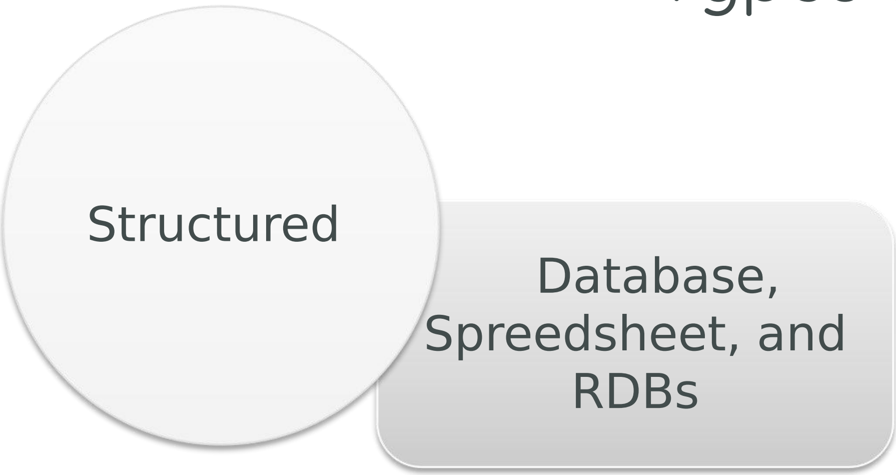
statsmodels

SciPy

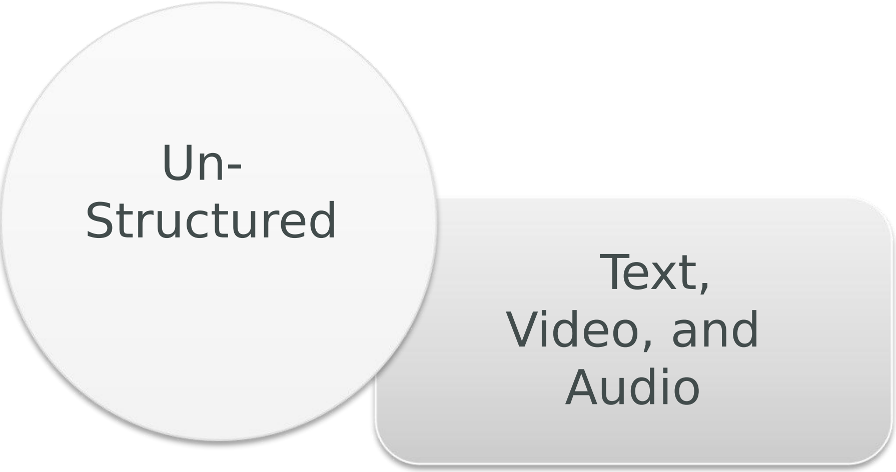
pycon tanzania

Processing Data

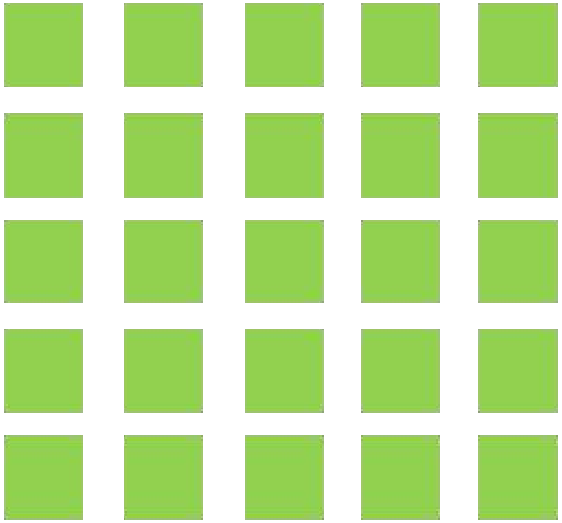
Types of Data

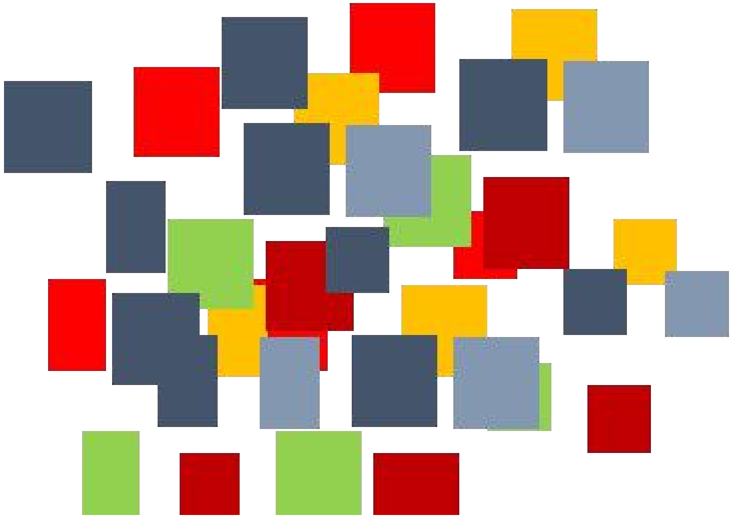

Structured

Database, Spreedsheet, and RDBs


Un-Structured

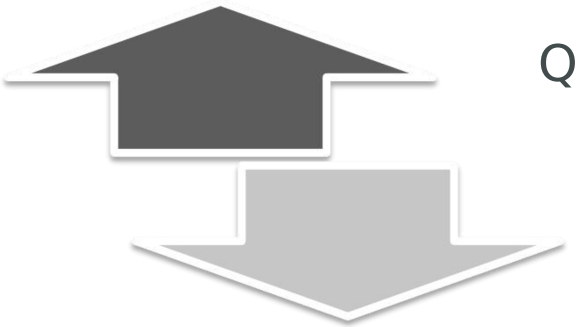
Text, Video, and Audio





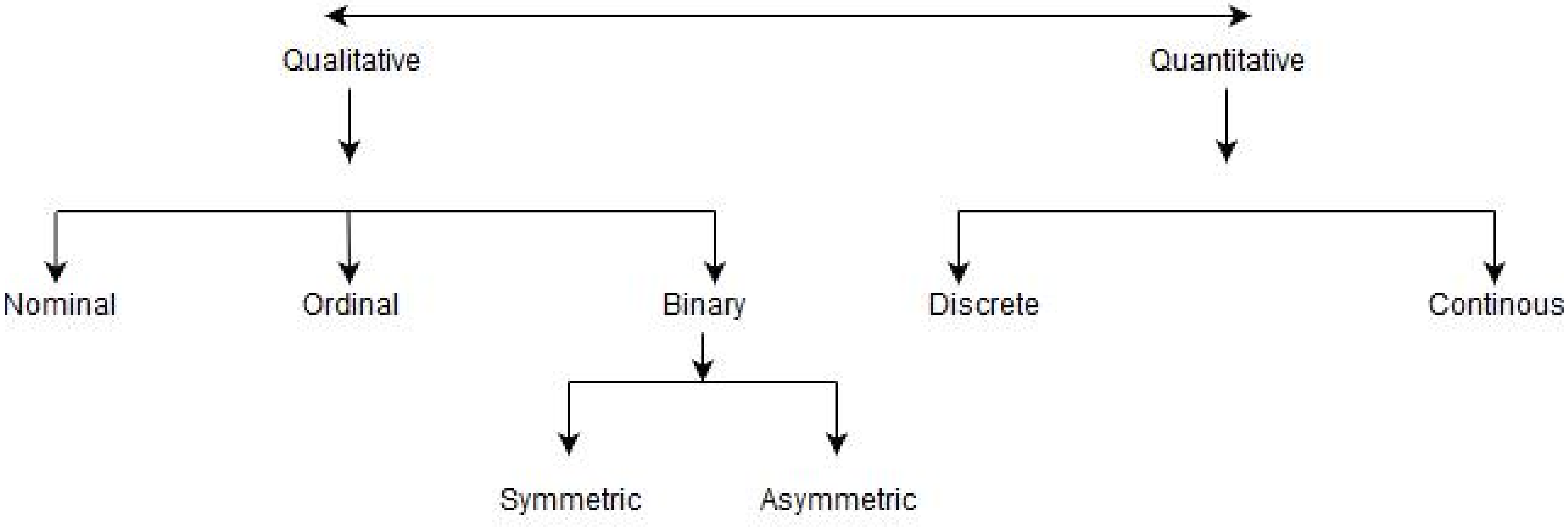
Processing Data

Types of Data Attributes



Quantitative/Numerical

Qualitative/Categorical



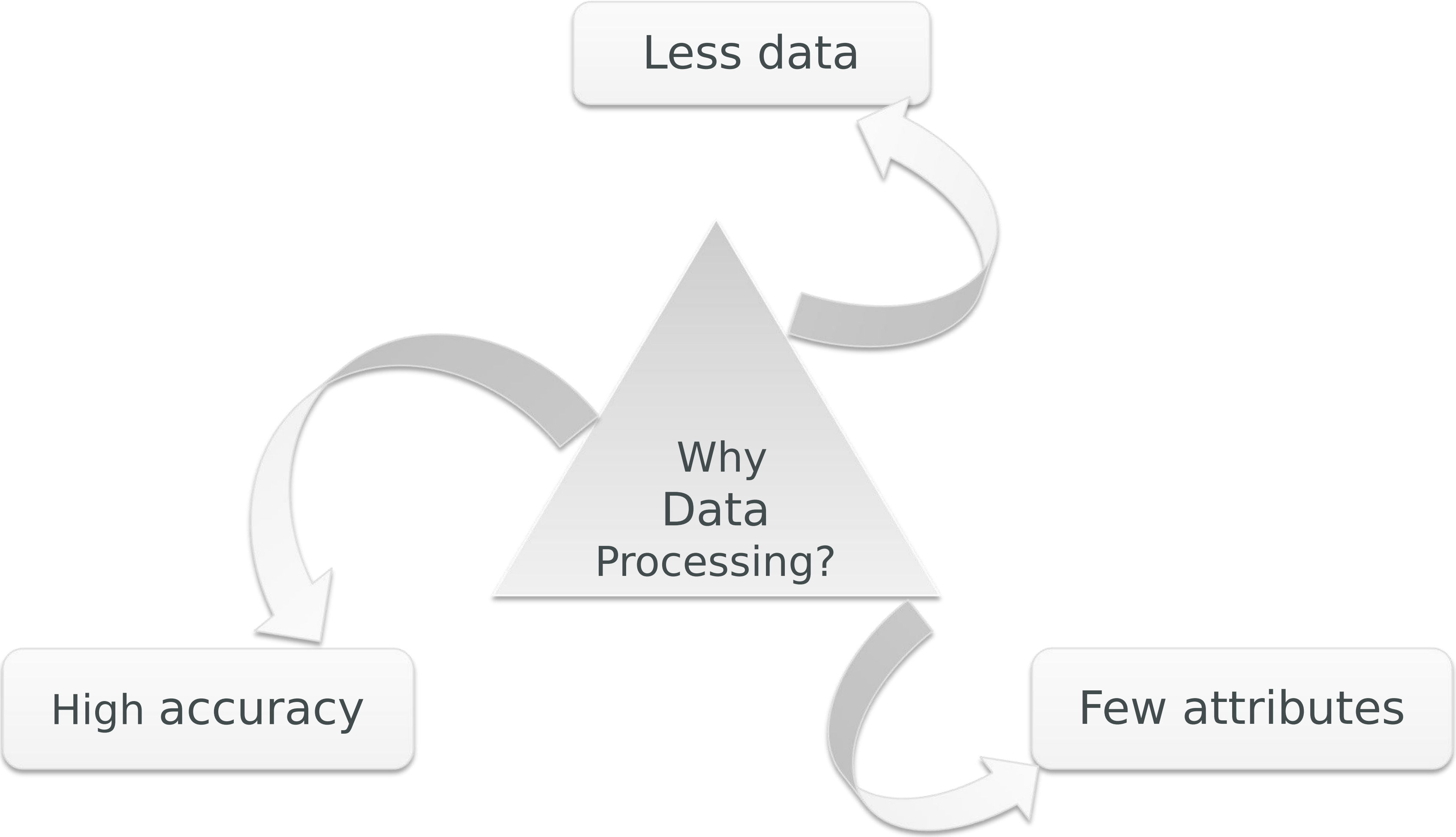
Processing Data

Data Wrangling

Transforming raw data to a clean and organized format.

Common data structure used to “wrangle” data is “data frame” .

Processing Data



Less data

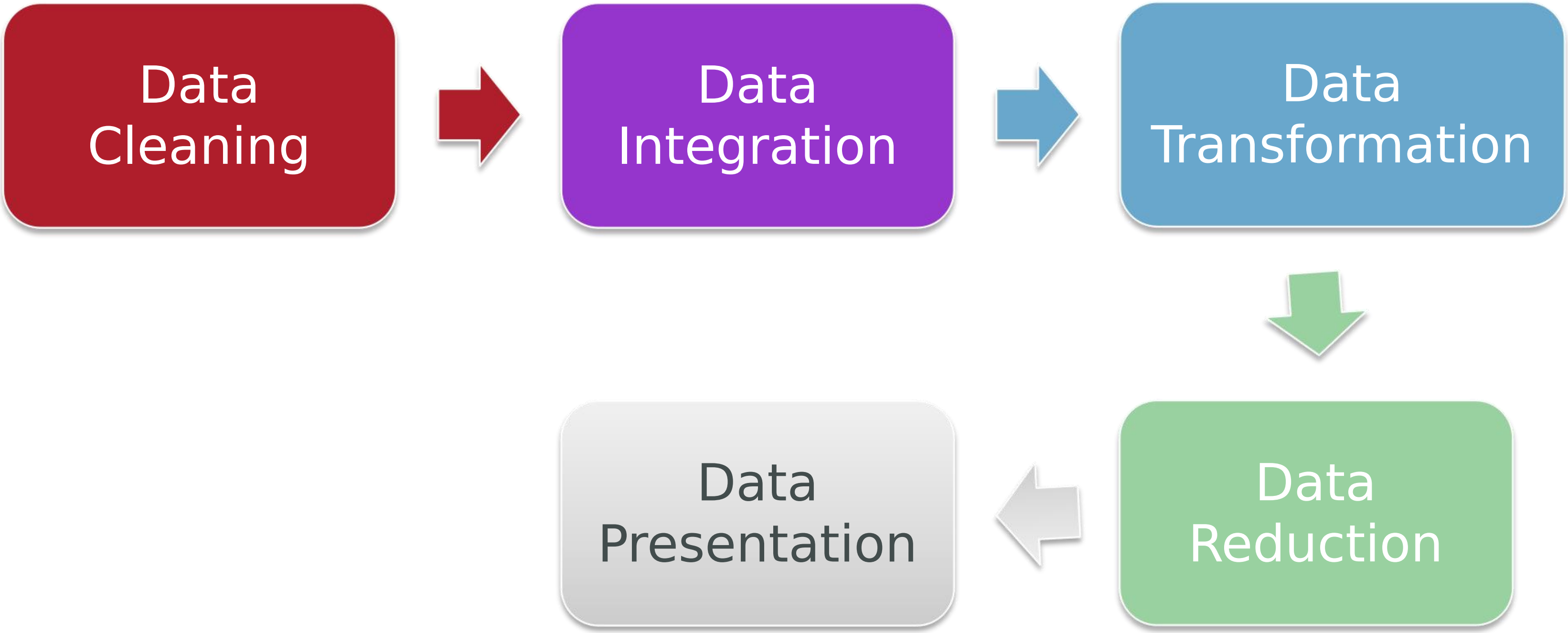
Why Data Processing?

High accuracy

Few attributes

Processing Data

Major Tasks in Data Processing



Data Cleaning

Data Integration

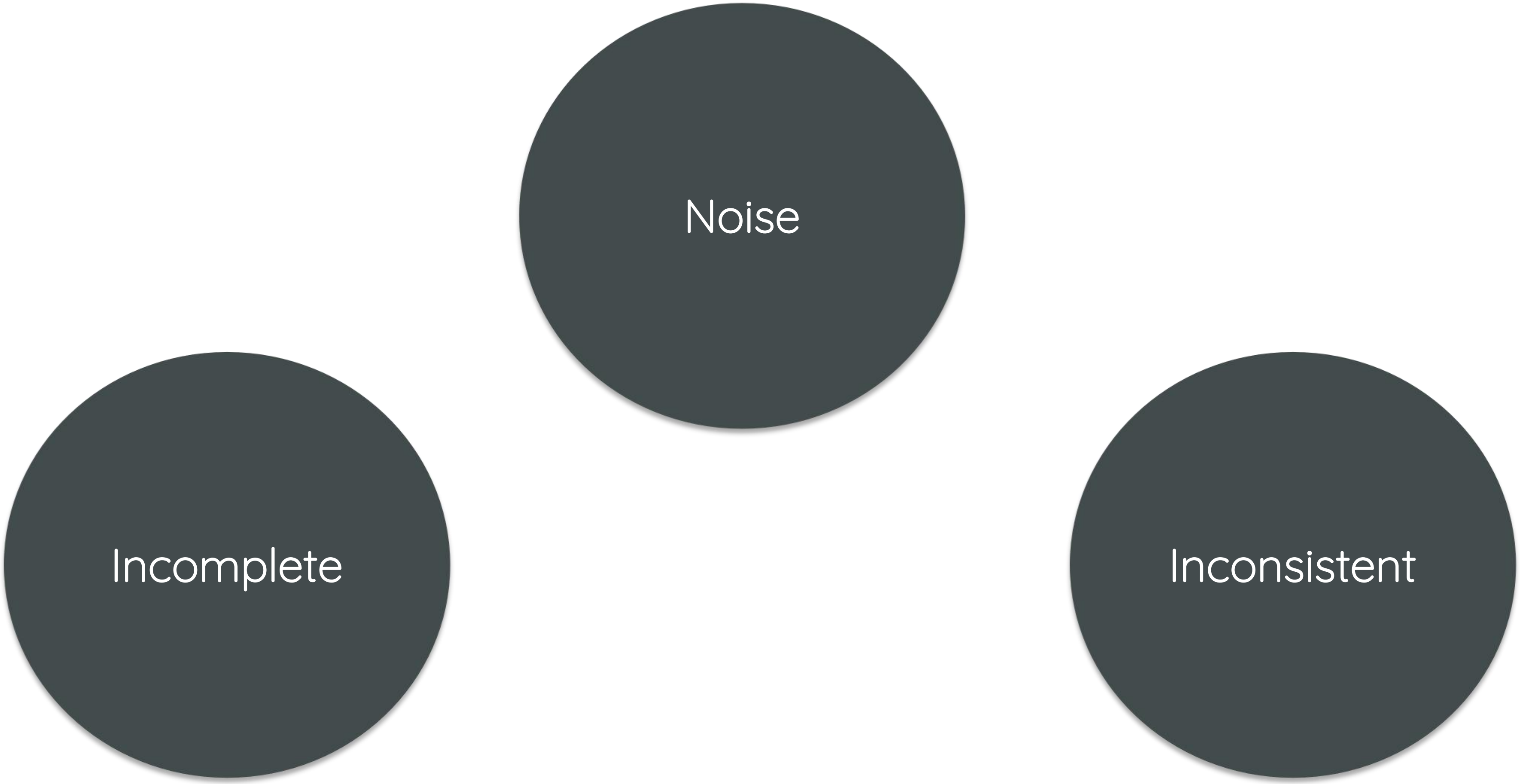
Data Transformation

Data Reduction

Data Presentation

Data Cleaning

Why Data is “Dirty”?



Noise

Incomplete

Inconsistent

Data cleaning

Types of Data Cleaning Methods

Missing Value

Fill in missing values.

-Mean,

-Median or Zero

Drop missing values.

- Ignore

Noisy Data

- Identify outliers.

Smooth out noisy data.

- Binning

pycon tanzania

Data Integration

Why Integarate Data

Schema Conflict

custom_id and cust_number , Use: Metadata)

Value conflict

“H” and “S” , and 1 and 2 for pay_type, (Metadata)

Redundant data

Use : Correlation and Chi-Square Test

pycon tanzania

Data Transformation

Ways of Transforming Data

Normalization

- Min-Max Normalization
- Z-score Normalization

Attribute Engineering

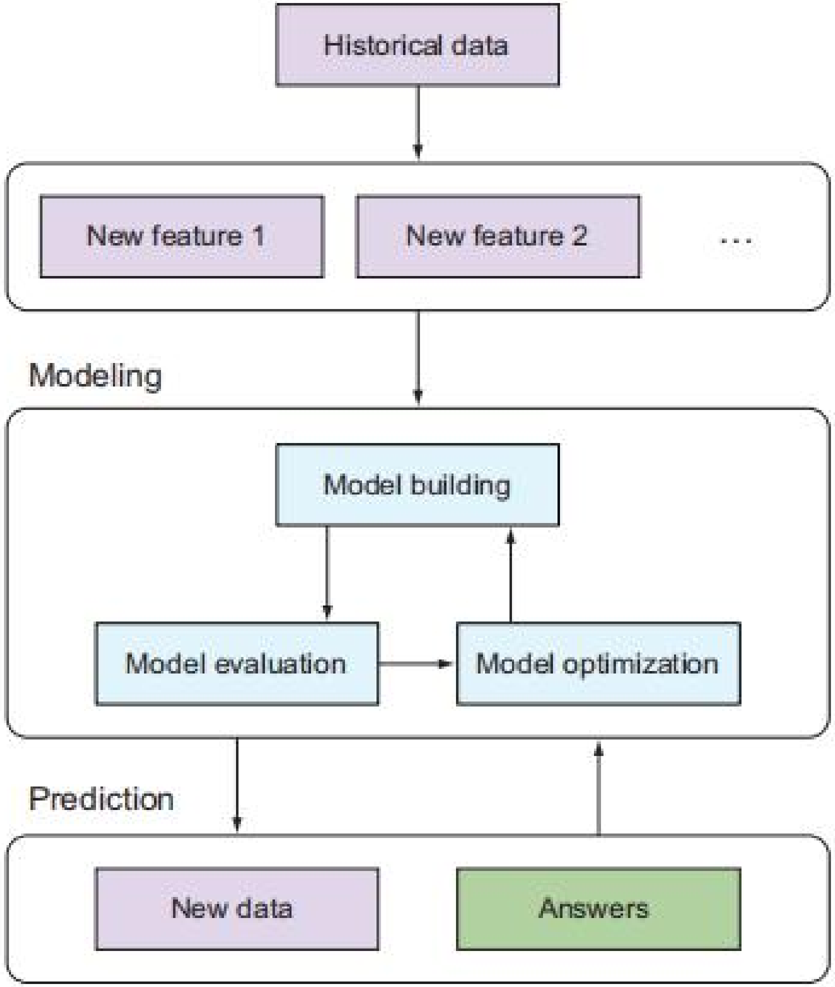
- Attribute Extraction
- Attribute Selection

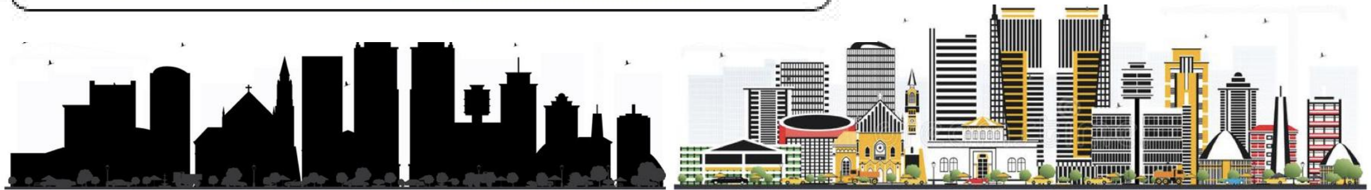
Data Reduction

- Data Discretization
- Dimensionality Reduction

Training ML Algo's

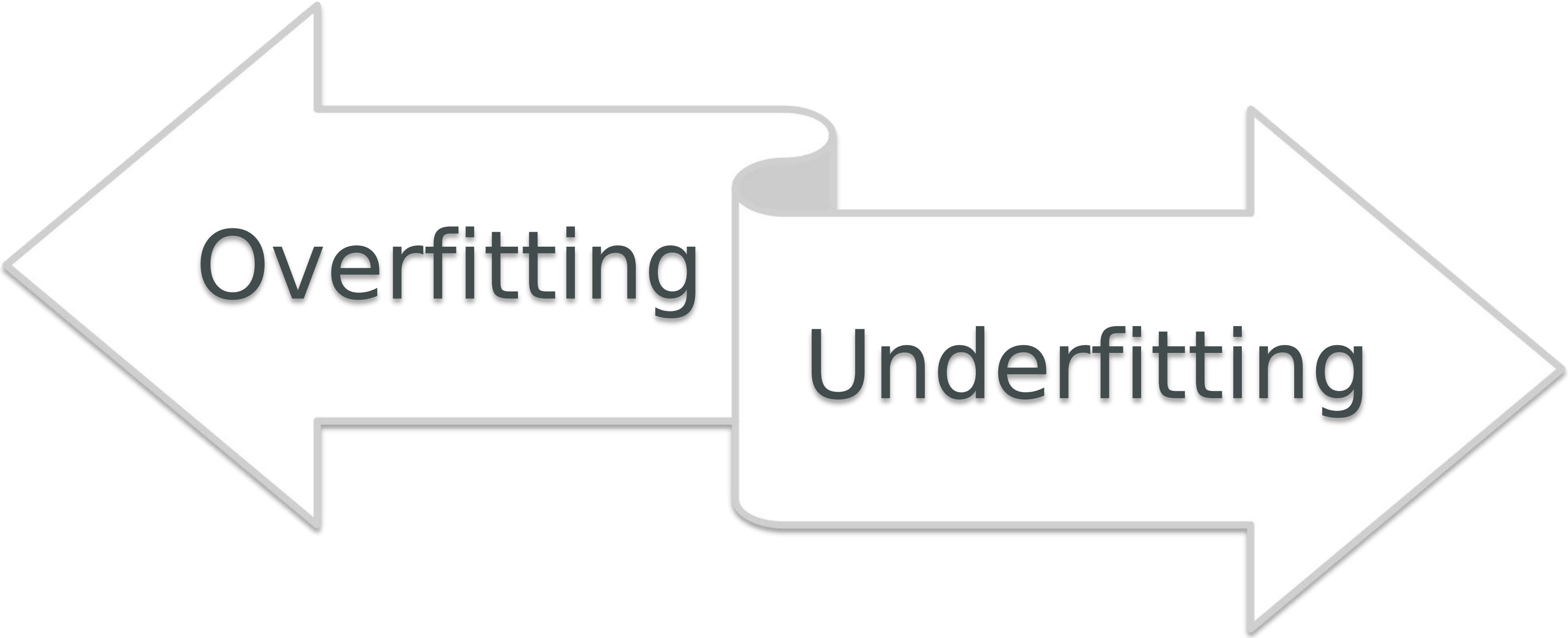
Machine Learning Workflow





Evaluating a Model

Generalization Problems



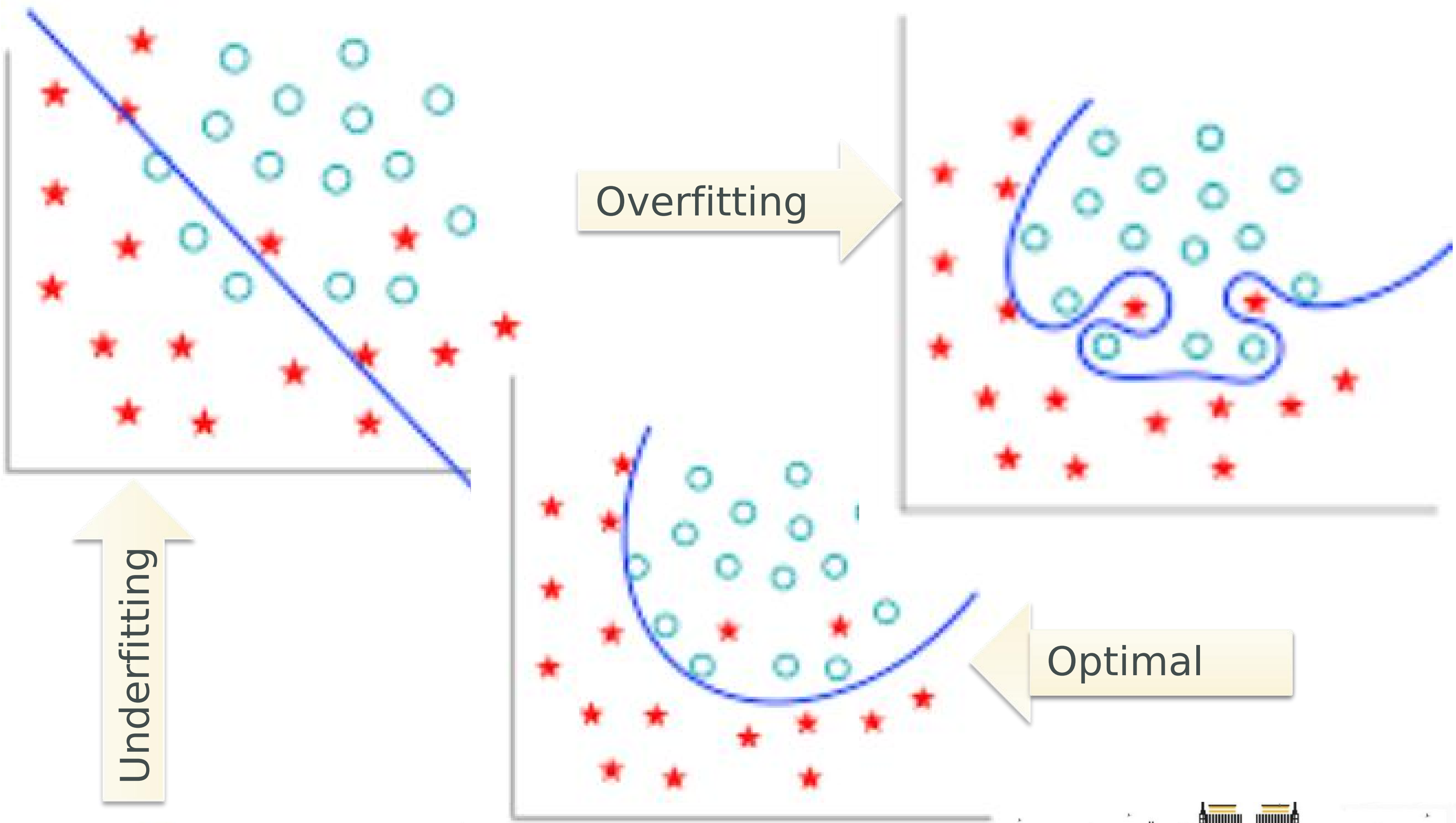
Overfitting

Underfitting

pycon tanzania

Evaluating a Model

Generalization Problems



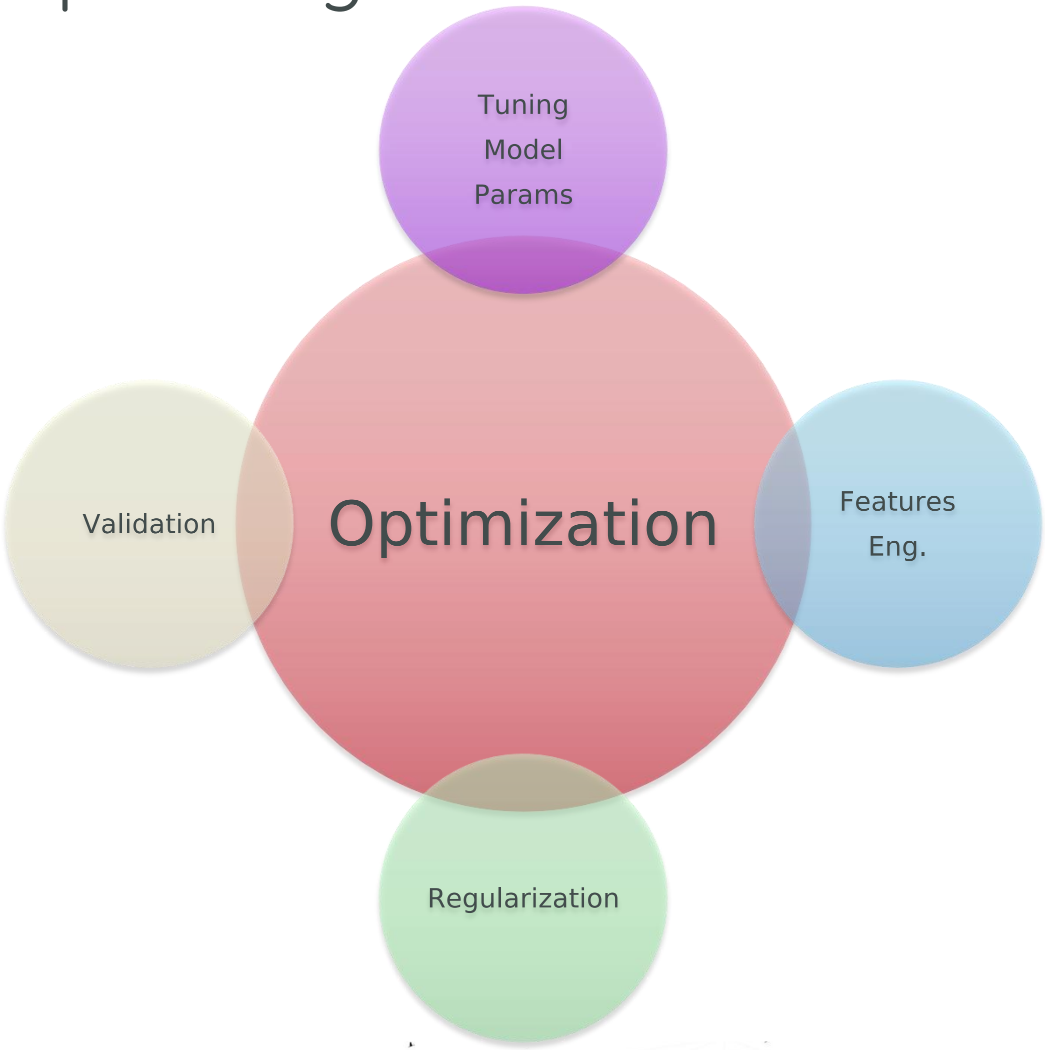
Overfitting

Underfitting

Optimal

Model Selection

Optimizing Model Performance



Tuning Model Params

Validation

Optimization

Features Eng.

Regularization

THANK YOU

pycon tanzania