Machine Learning Overview

After this video you will be able to..

- Explain what machine learning is
- List three applications of machine learning encountered in everyday life

What is Machine Learning?

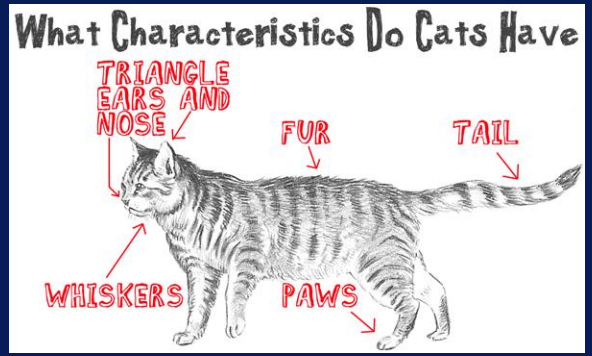
Machine Learning is…

…learning from data

Machine Learning is…

… learning from data

… no explicit programming



Machine Learning is…

… learning from data

… no explicit programming

… discovering hidden patterns

Machine Learning is…

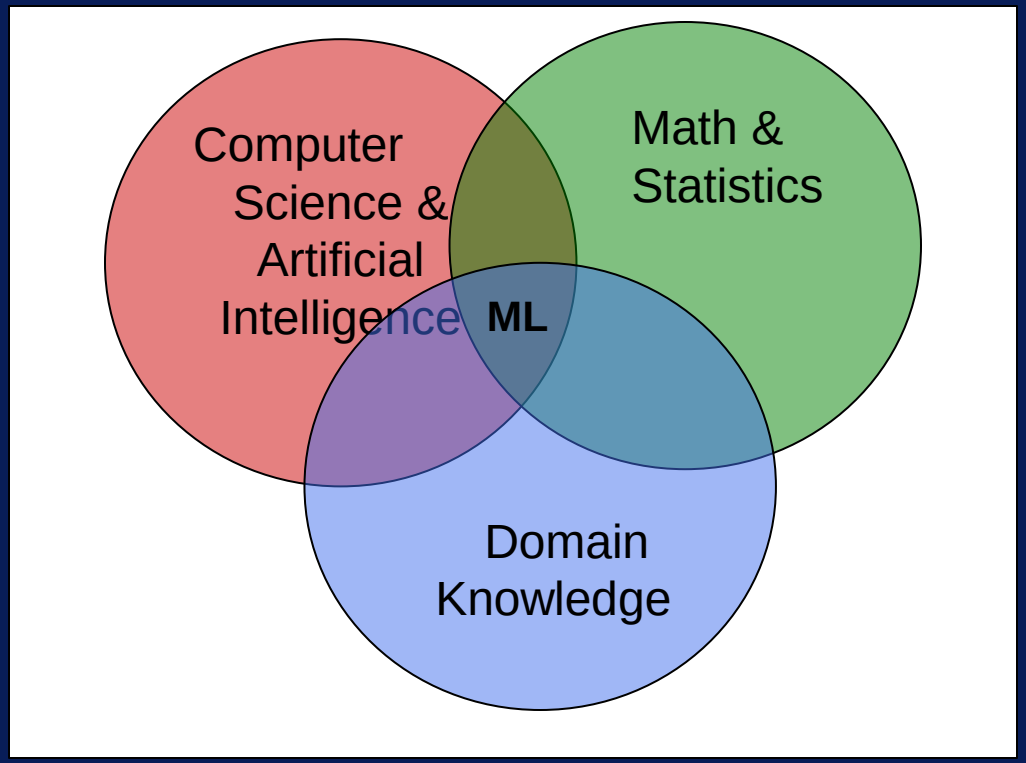
… learning from data

… no explicit programming

… discovering hidden patterns

… data-driven decisions

Machine Learning (ML) is an Interdisciplinary Field

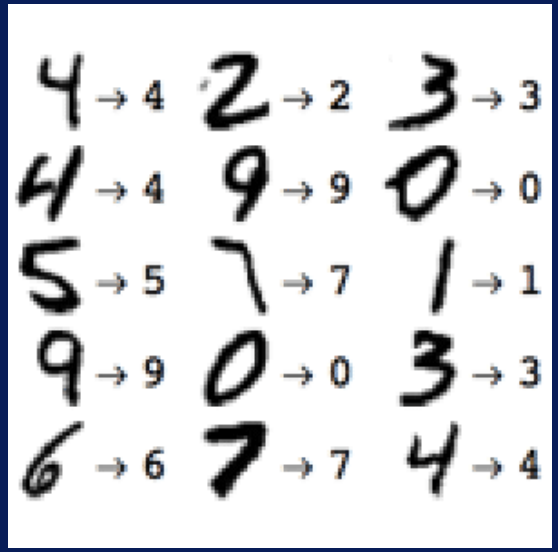


Example Application of Machine Learning

- **Credit card fraud detection**

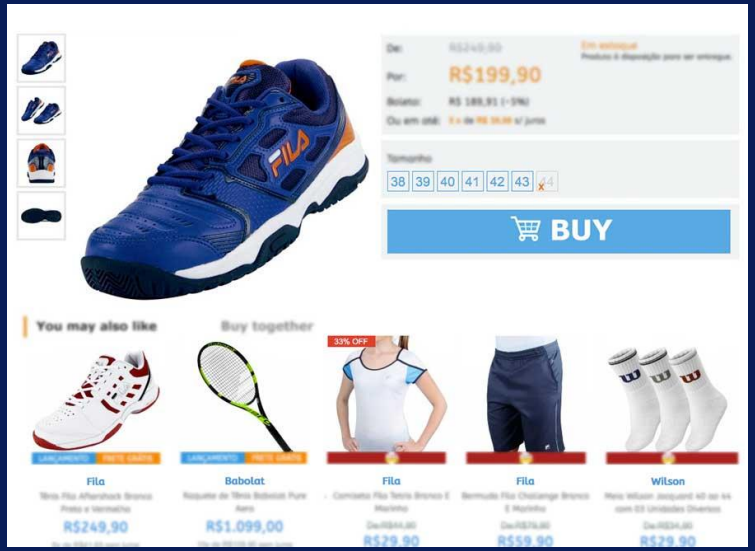
Example Application of Machine Learning

- **Handwritten digit recognition**



Example Application of Machine Learning

- Recommendations on websites



More Applications of Machine Learning

- Targeted ads on mobile apps
- Sentiment analysis
- Climate monitoring
- Crime pattern detection
- Drug effectiveness analysis

What's in a Name?

Machine learning

Data mining

Predictive analytics

Data science

Machine Learning Models

- Learn from data
- Discover patterns and trends
- Allow for data-driven decisions
- Used in many different applications

Categories of Machine Learning Techniques

After this video you will be able to..

- Describe the main categories of machine learning techniques
- Summarize how supervised learning differs from unsupervised learning

Categories of Machine Learning Techniques

- Classification
- Regression
- Cluster Analysis
- Association Analysis

Classification

Goal: Predict category

Sunny

Windy

Rainy

Cloudy

Image source: http://www.davidson.k12.nc.us/parents___students/inclement_weather

Classification Examples

- Classify tumor as benign or malignant
- Predict if it will rain tomorrow
- Determine if loan application is high-, medium-, or low-risk
- Identify sentiment as positive, negative, or neutral

Regression

Goal: Predict numeric value

Regression Examples

- Estimate demand for a product based on time of year
- Predict score on a test
- Determine likelihood of drug effectiveness for patient
- Predict amount of rain

Cluster Analysis

Goal: Organize similar items into groups.

Seniors

Adults

Teenagers

Presenter

Image source: http://www.monetate.com/blog/the-intrinsic-value-of-customer-segmentation

Cluster Analysis Examples

- Identify areas of similar topography (desert, grass, etc.)
- Categorize different types of tissues from medical images
- Determine different groups of weather patterns
- Discover crime hot spots

Presenter

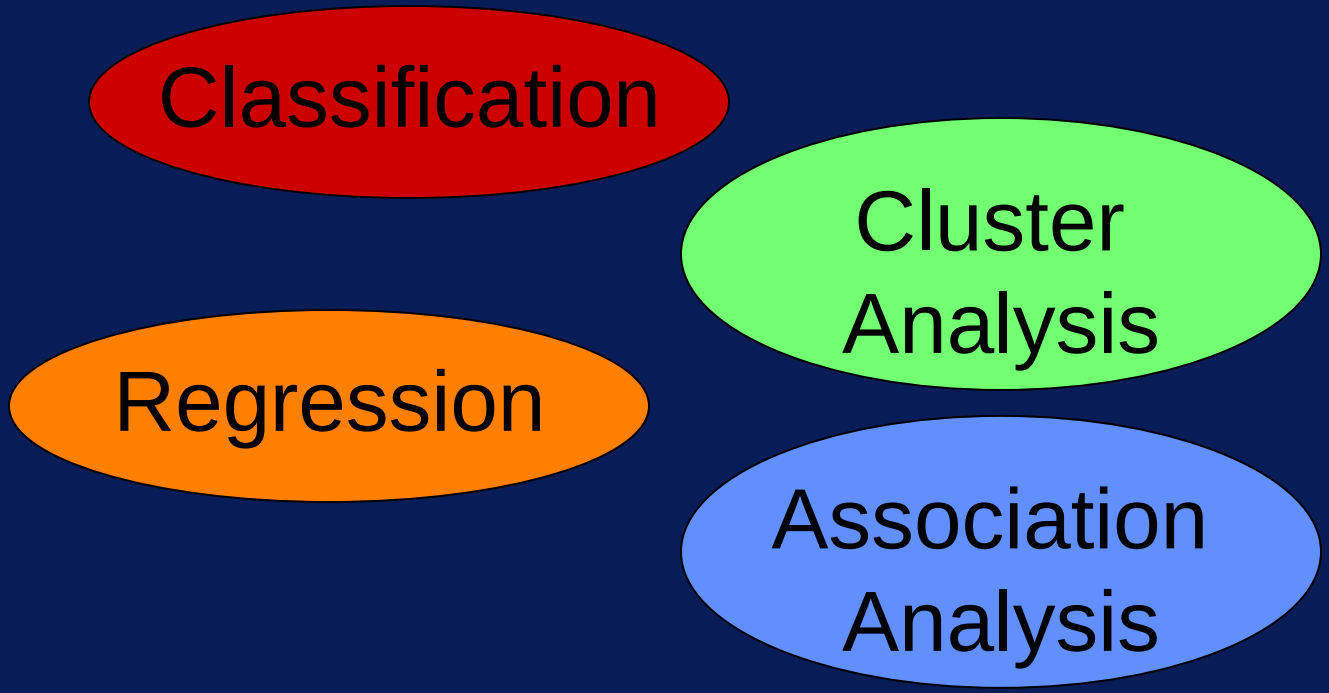
Association Analysis

Goal: Find rules to capture associations between items.

Association Analysis Examples

- Recommend items based on purchase/browsing history
- Have sales on related items often purchased together
- Identify web pages accessed together

Categories of Machine Learning Techniques



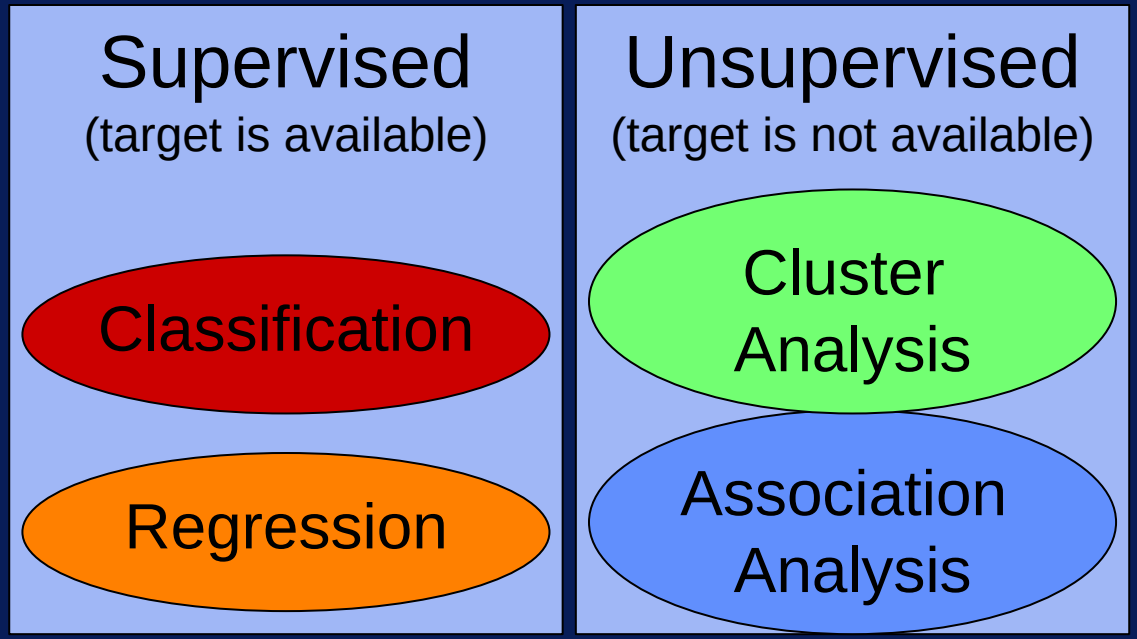
Supervised vs. Unsupervised

- **Supervised Approaches**
  - Target (what model is predicting) is provided
  - ‘Labeled’ data
  - Classification & regression are supervised.

Supervised vs. Unsupervised

- **Unsupervised Approaches**
  - Target is unknown or unavailable
  - ‘unlabeled’ data
  - Cluster analysis & association analysis are unsupervised.

Categories of Machine Learning Techniques



Machine Learning Process

After this video you will be able to..

- Identify the steps in the machine learning process

- Discuss why the machine learning process is iterative

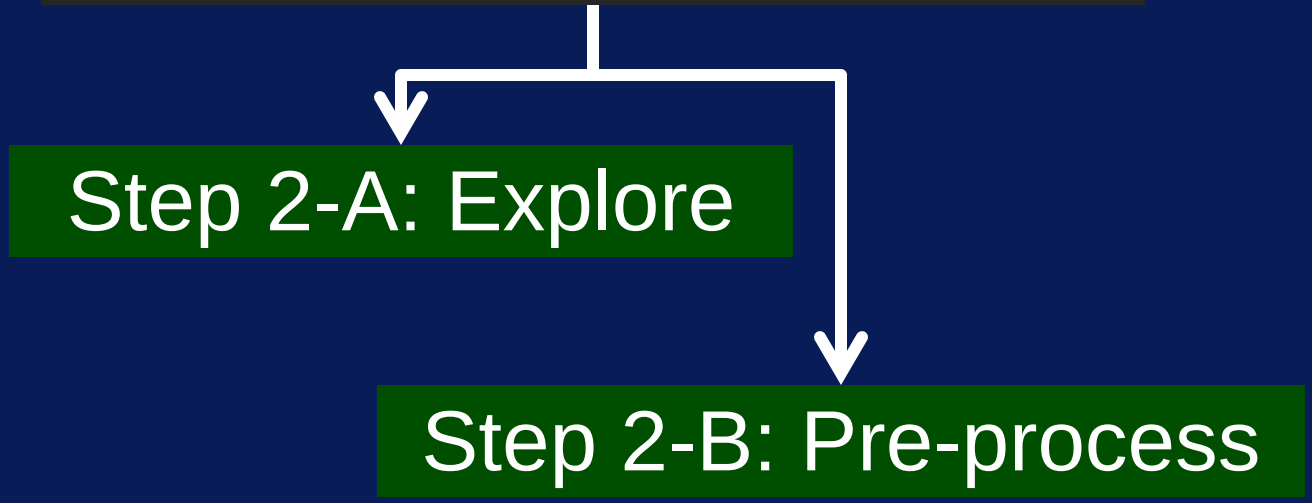
ACQUIRE PREPARE ANALYZE REPORT ACT

ACQUIRE PREPARE ANALYZE REPORT ACT

Step 1: Acquire Data

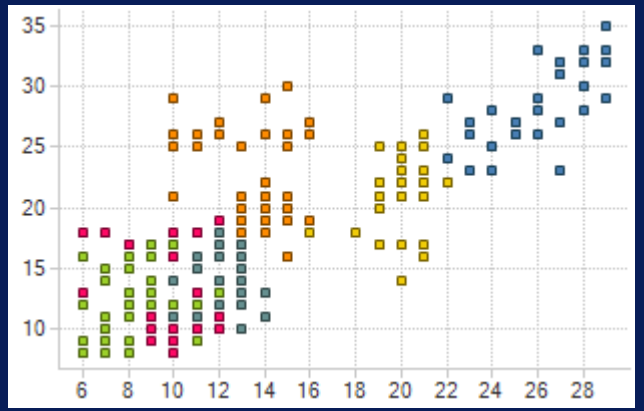
- Identify data sources
- Collect data
- Integrate data

Step 2: Prepare Data

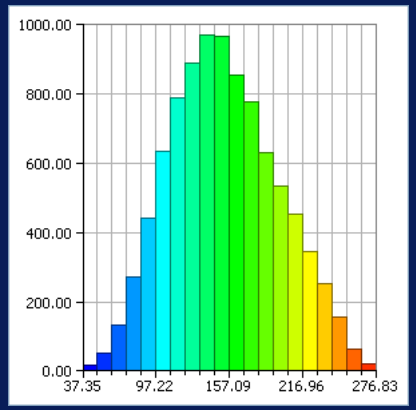


ACQUIRE | **PREPARE** | ANALYZE | REPORT | ACT

Step 2-A: Explore Data



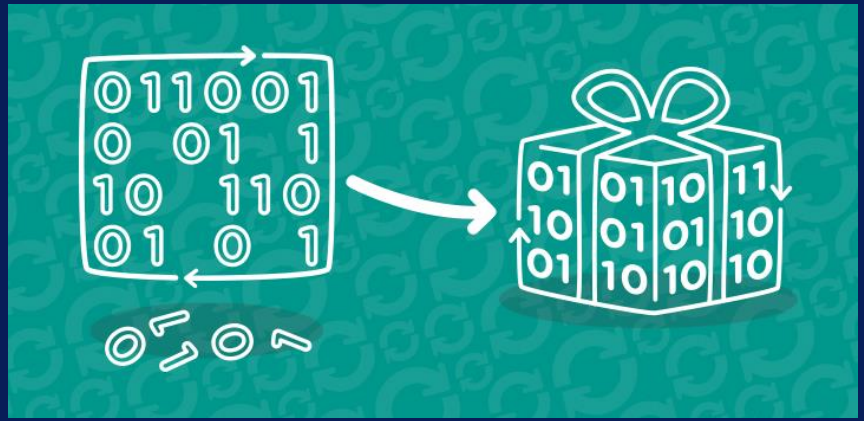
Understand nature of data



Preliminary analysis

ACQUIRE | **PREPARE** | ANALYZE | REPORT | ACT

Step 2-B: Pre-process Data



Clean | Select | Transform

ACQUIRE | PREPARE | **ANALYZE** | REPORT | ACT

Step 3: Analyze Data

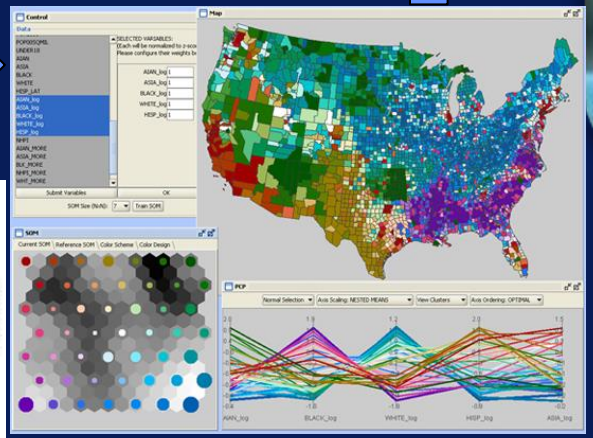
Select analytical techniques

Build models

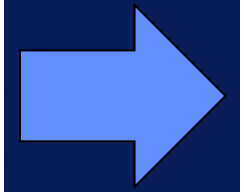
Assess results

ACQUIRE PREPARE ANALYZE **REPORT** ACT

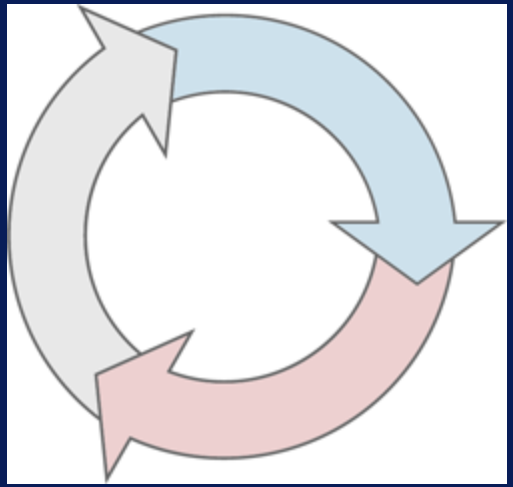
Step 4: Communicate Results



Step 5: Apply Results



ACQUIRE PREPARE ANALYZE REPORT ACT



Iterative process

Goals and Activities in the Machine Learning Process

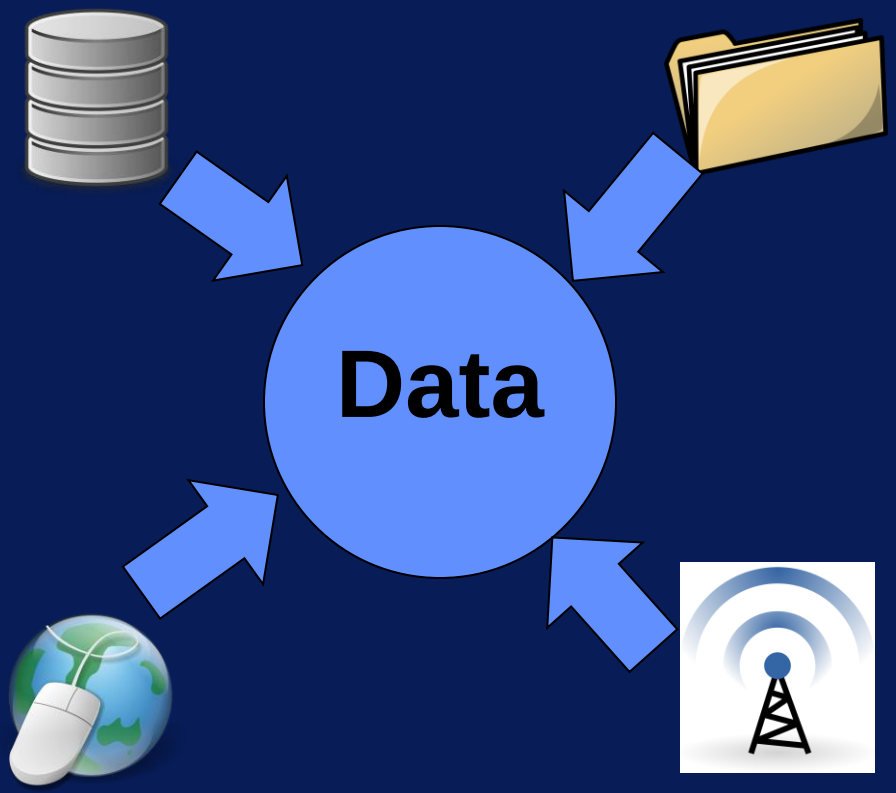
After this video you will be able to...

- Explain the goals of each step in the machine learning process
- List key activities in each step in the process


ACQUIRE
PREPARE
ANALYZE
REPORT
ACT

Goal: Identify and obtain all data related to problem

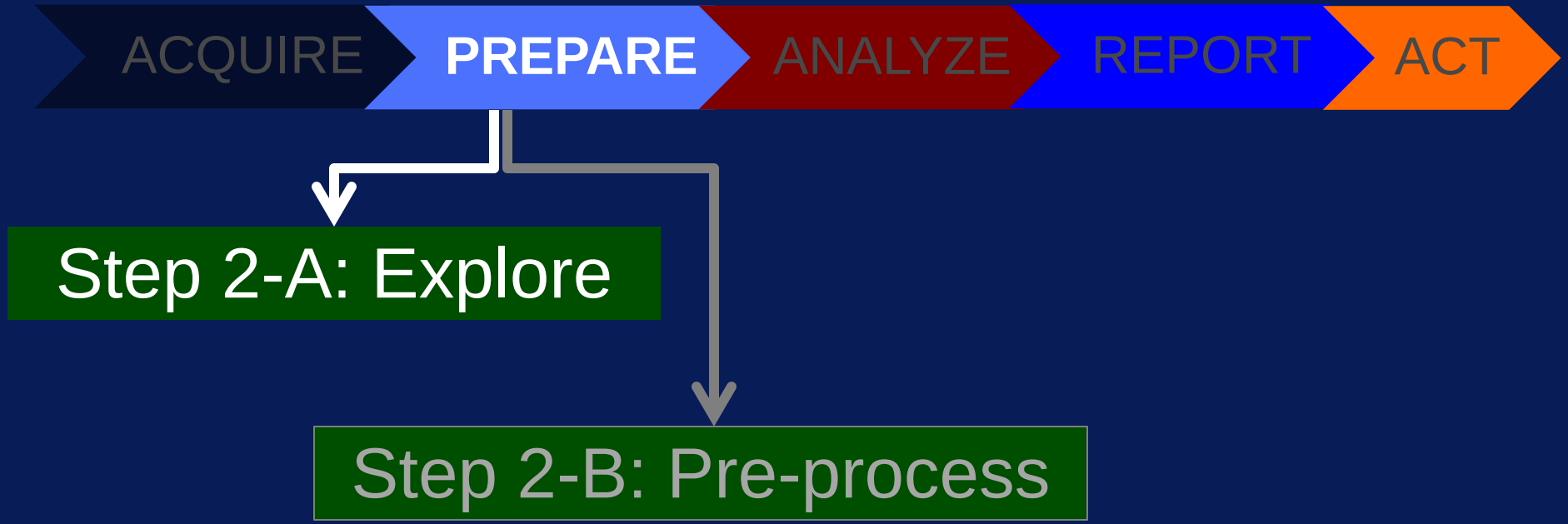
Acquire Data



Identify data sources

Collect data

Integrate data

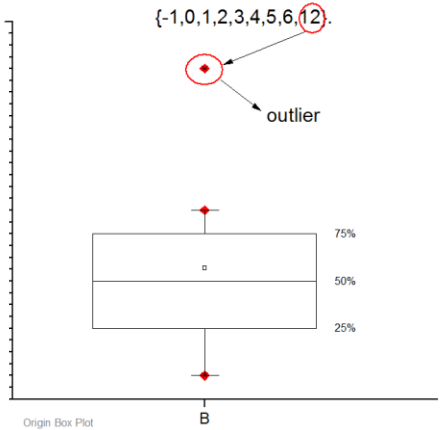

ACQUIRE
PREPARE
ANALYZE
REPORT
ACT
Step 2-A: Explore
Step 2-B: Pre-process

Why Explore?

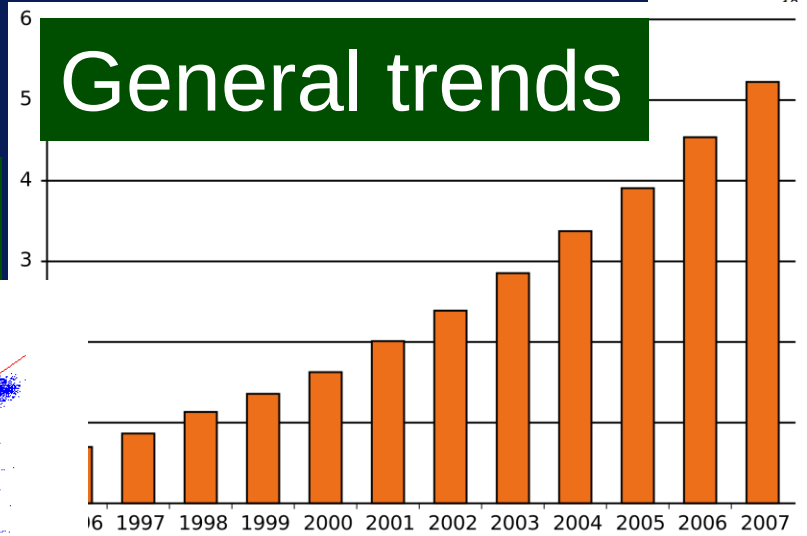
Goal: Understand your data

Why Explore?

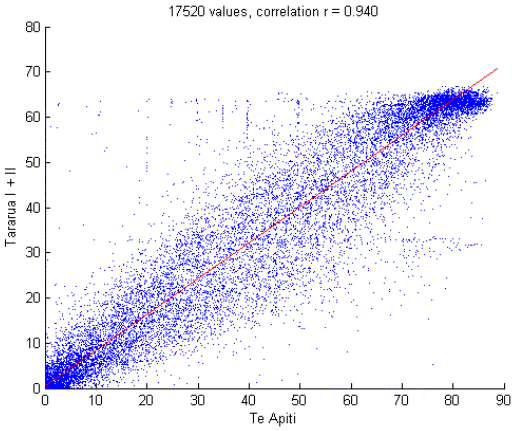
Outliers



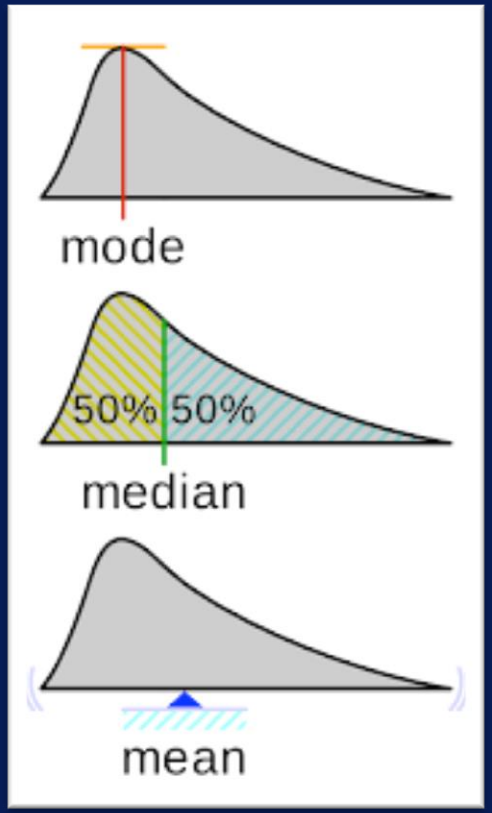
General trends

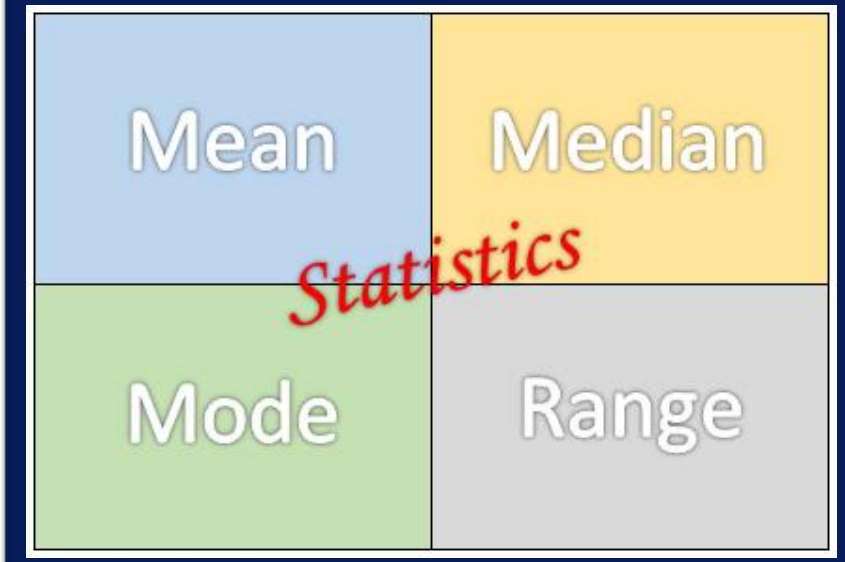


Correlations



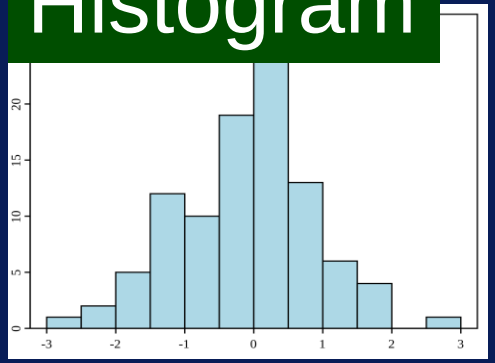
Describe Your Data



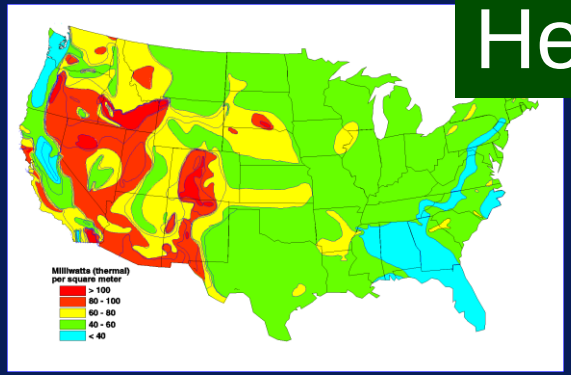


Visualize Your Data

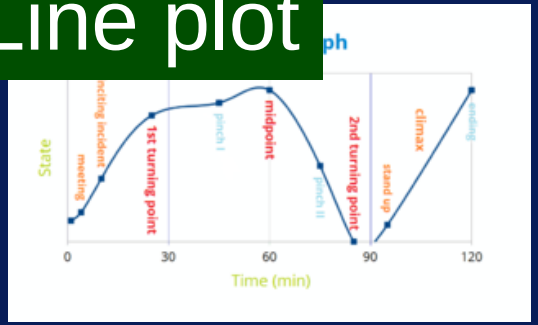
Histogram



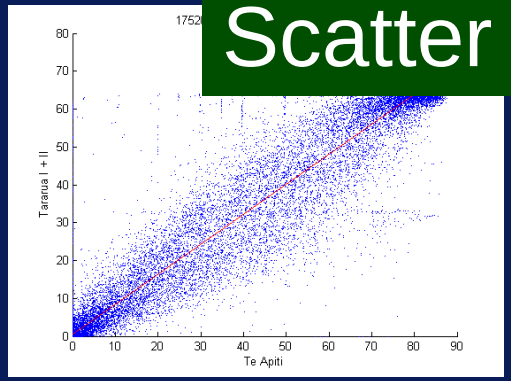
Heat map

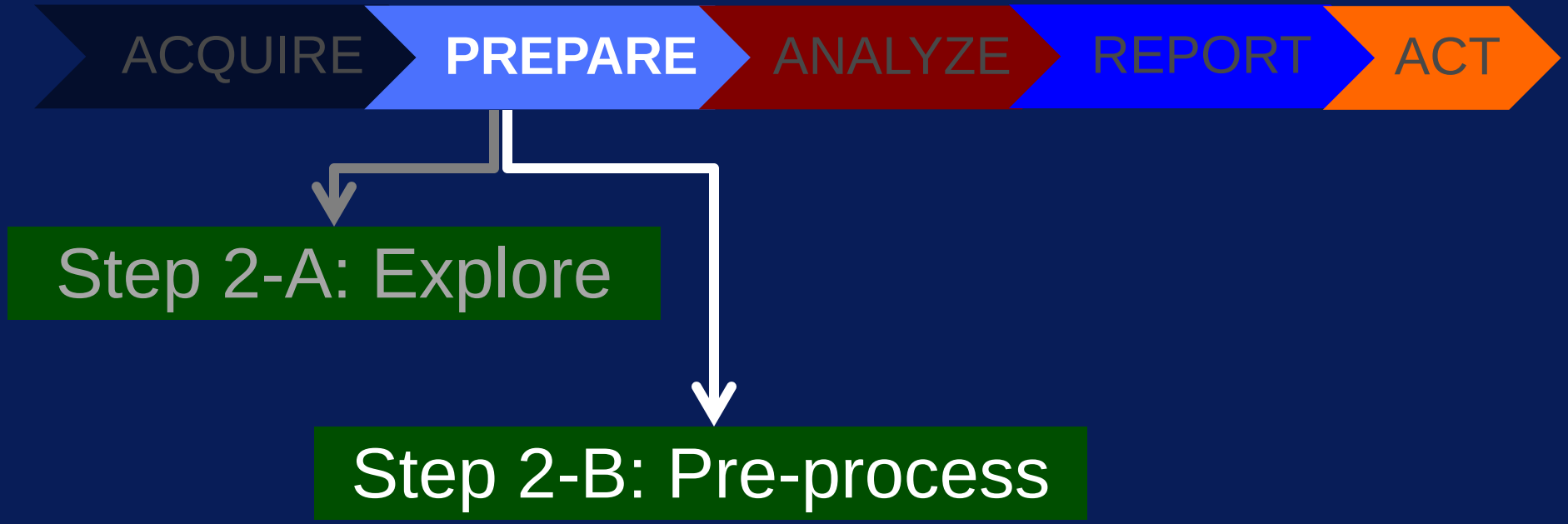


Line plot



Scatter plot




ACQUIRE
PREPARE
ANALYZE
REPORT
ACT
Step 2-A: Explore
Step 2-B: Pre-process

Step 2-A: Explore

Goal: Create data for analysis

Step 2-B: Pre-process

Clean

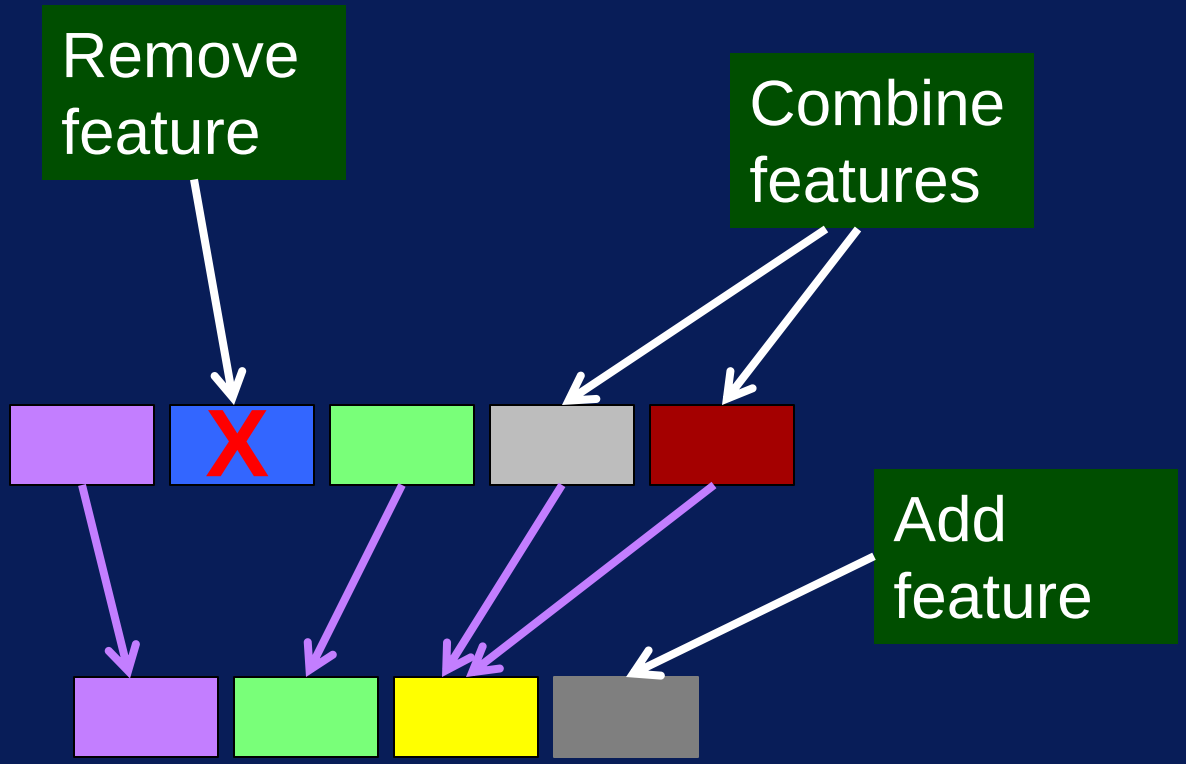
Select

Transform

Data Cleaning

- Missing values
- Duplicate data
- Inconsistent data
- Noise
- Outliers

Feature Selection



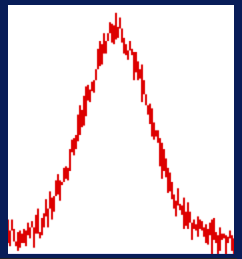
Feature Transformation

Weight

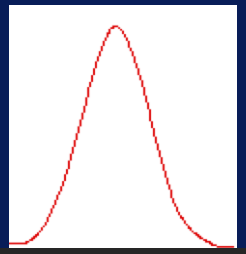
Scaled Values

Height

Scaling



Original



Filtered

Filtering

3D

2D

Dimensionality Reduction

Goals:

- Build model
- Evaluate results

Analyze

Select technique

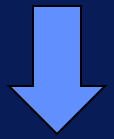
Build model

Evaluate

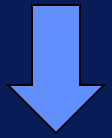
Classification
Regression
Cluster Analysis
Association
Analysis

Analyze

Select technique



Build model



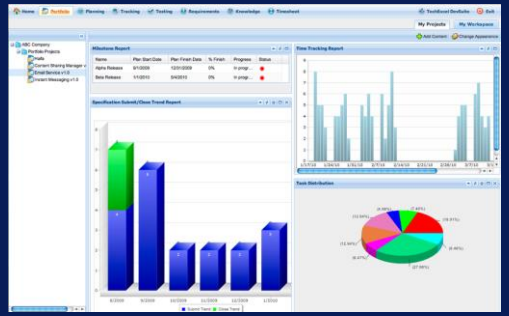
Evaluate results

Goal: Communicate results and recommend actions

Present

Report

with



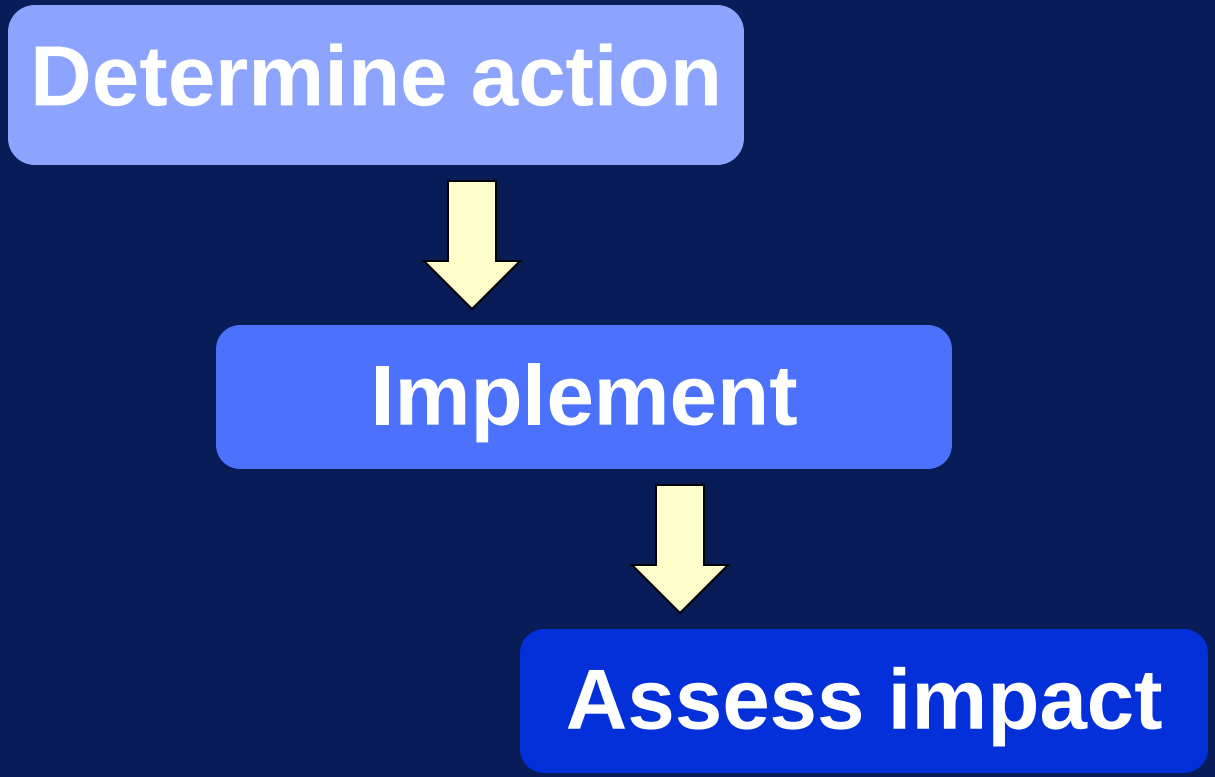
using

tableau

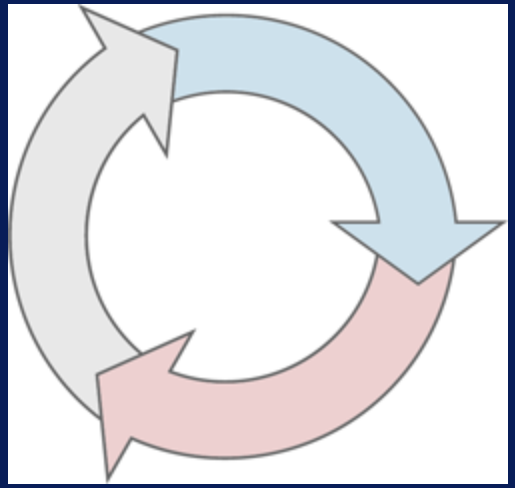
Google Developers CHARTS

Goal: Determine actions based on insights

Act



ACQUIRE PREPARE ANALYZE REPORT ACT



Iterative process

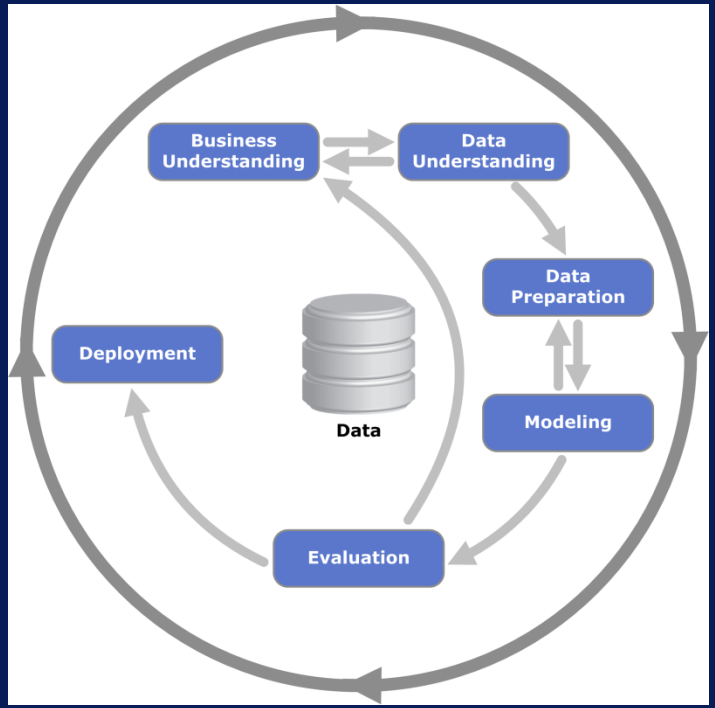
Cross Industry Standard Process for Data Mining (CRISP-DM)

After this video you will be able to..

- Summarize what CRISP-DM is
- List the phases of CRISP-DM
- Describe the goals of each phase

CRISP-DM

CRoss Industry Standard Process for Data Mining



CRISP-DM Phases

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

Phase 1 – Business Understanding

- Define problem or opportunity
- Assess situation
- Formulate goals

Phase 2 – Data Understanding

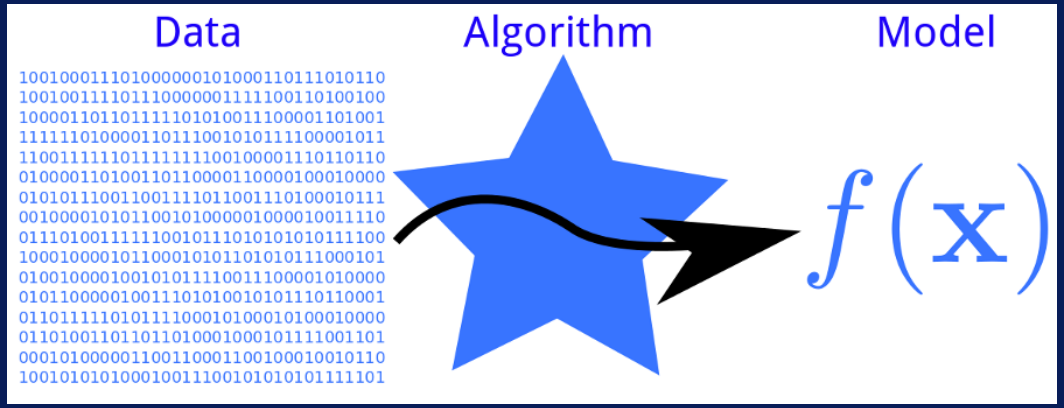
- Data acquisition
- Data exploration

Phase 3 – Data Preparation

- Prepare data for modeling
- Address quality issues, select features to use, process data for modeling

Phase 4 – Modeling

- Determine type of problem
- Select modeling technique(s) to use
- Build model

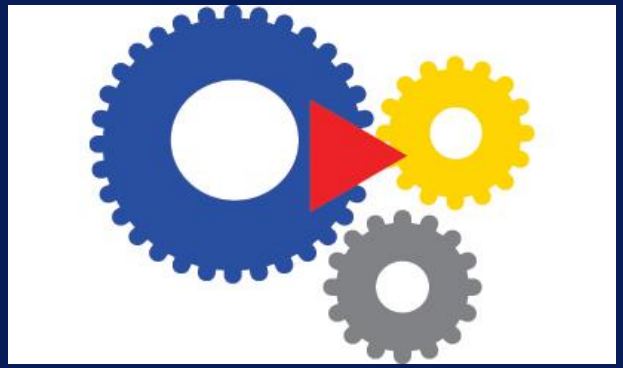


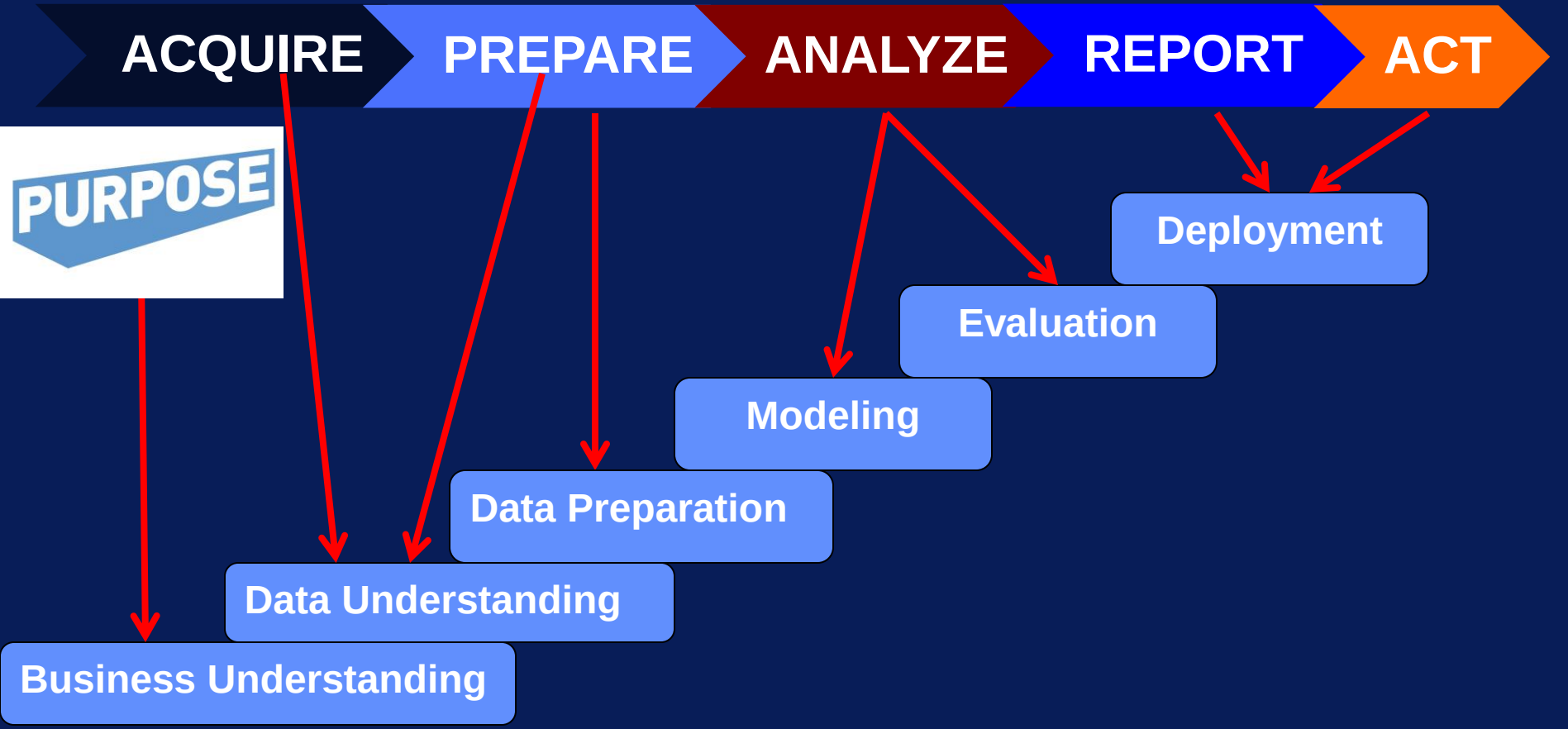
Phase 5 – Evaluation

- Assess model performance
- Evaluate model results with respect to success criteria

Phase 6 – Deployment

- Produce final report
- Deploy model
- Monitor model




ACQUIRE
PREPARE
ANALYZE
REPORT
ACT
PURPOSE
Deployment
Evaluation
Modeling
Data Preparation
Data Understanding
Business Understanding

Tools Used in This Course

After this video you will be able to..

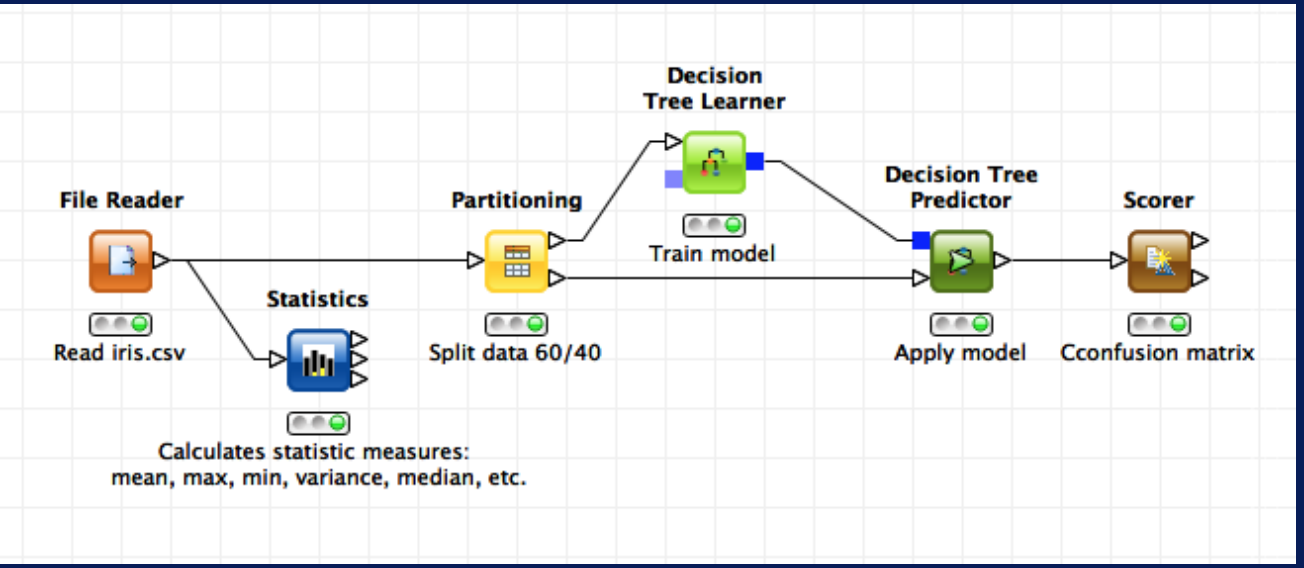
- Describe what KNIME is
- Describe what Spark MLlib is
- Contrast KNIME and ML as machine learning tools

Tools for This Course

KNIME Analytics Platform

- Platform for data analytics, reporting, and visualization
- GUI-based approach with drag-and-drop interface
- Nodes provide functionality
- Nodes are assembled to create workflows

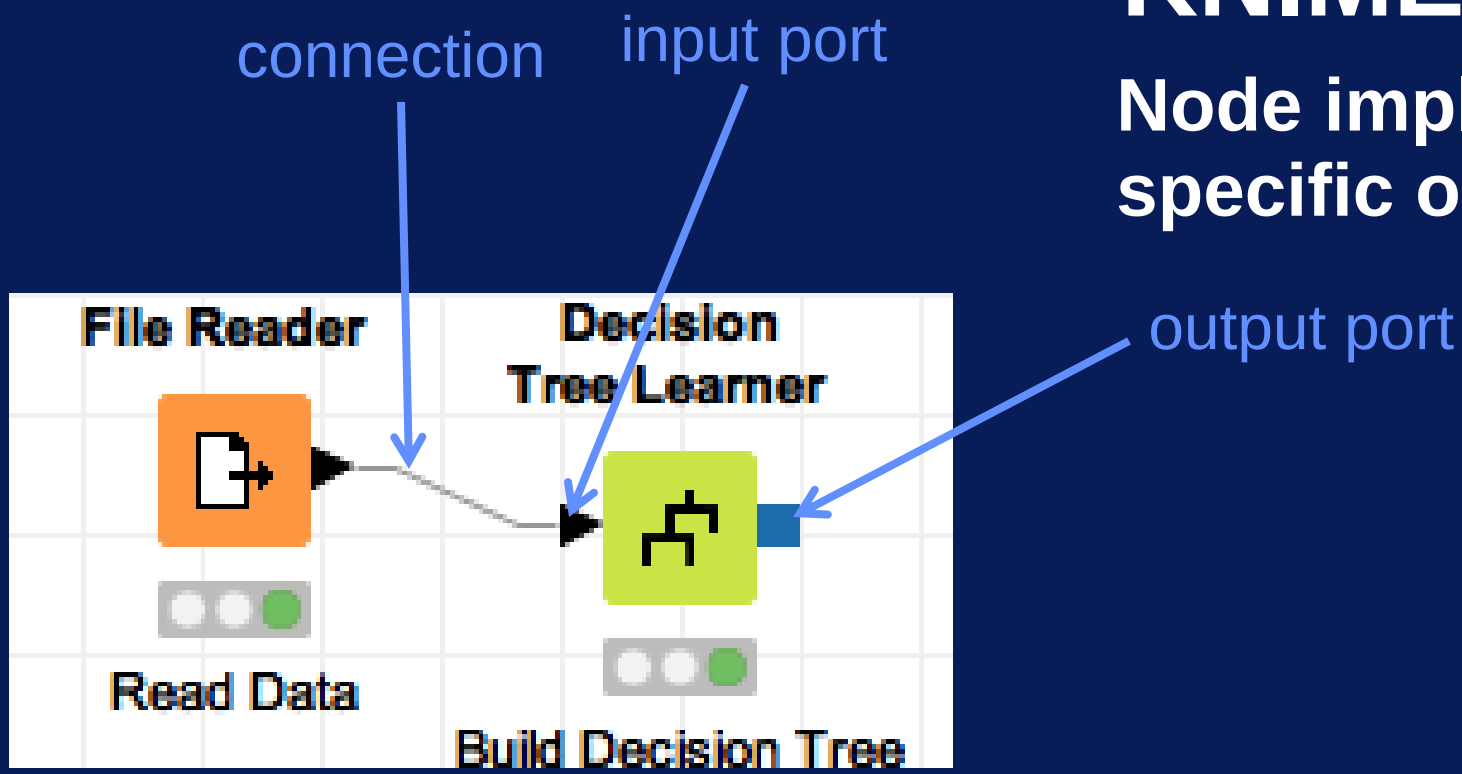
KNIME Workflow

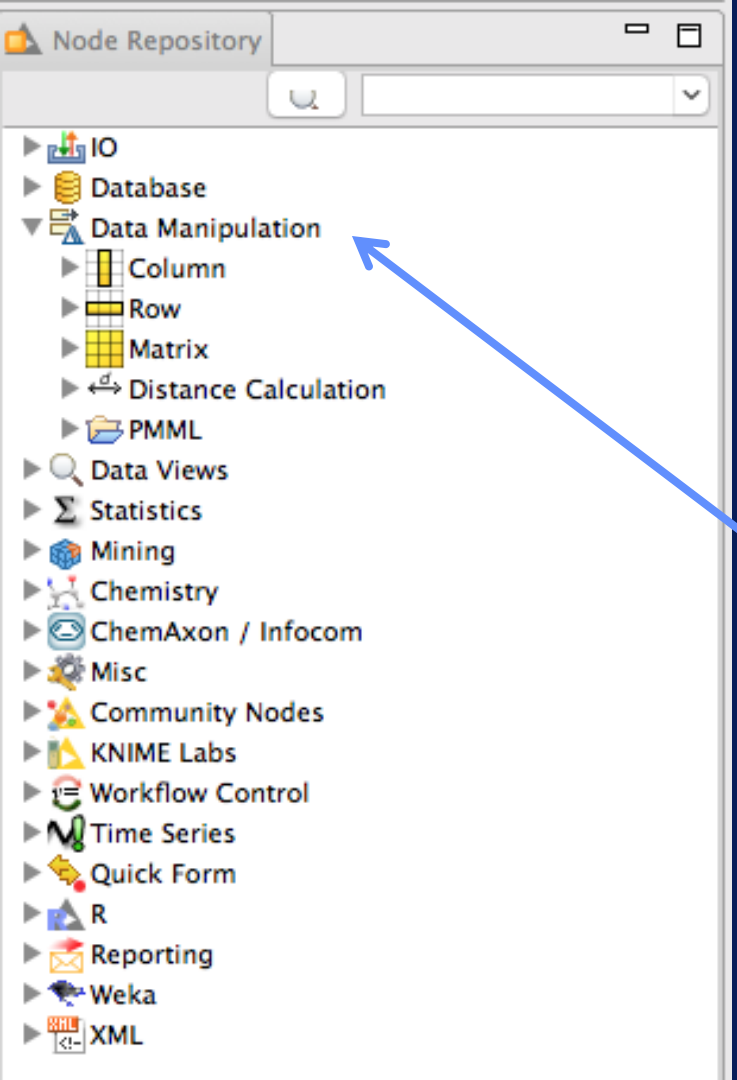


- Visual representation of steps in analysis process
- Workflow is composed of nodes

KNIME Node

Node implements specific operation





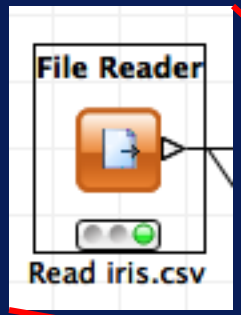
Node Repository

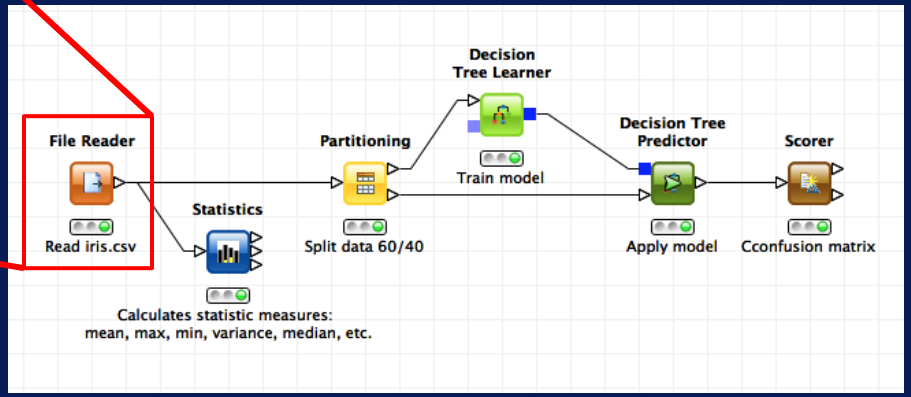
Contains nodes organized by category

Expand category to see subcategories and nodes

KNIME

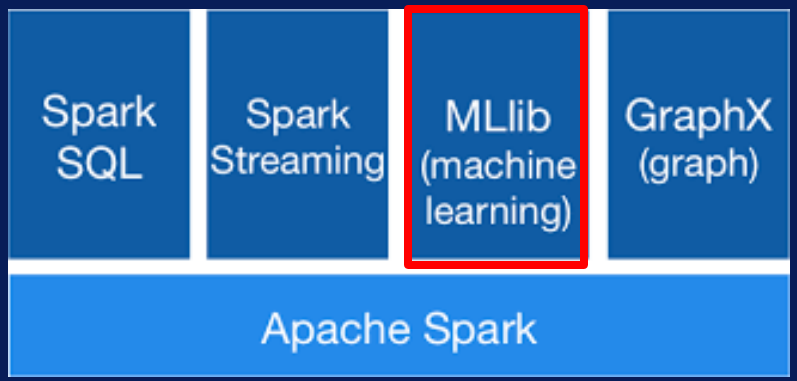
- GUI-based
- Drag-and-drop
- Interactive
- For small datasets





Spark MLlib

- **Scalable machine learning library**
- **Runs on Spark**
  - Distributed computing platform



Spark MLlib

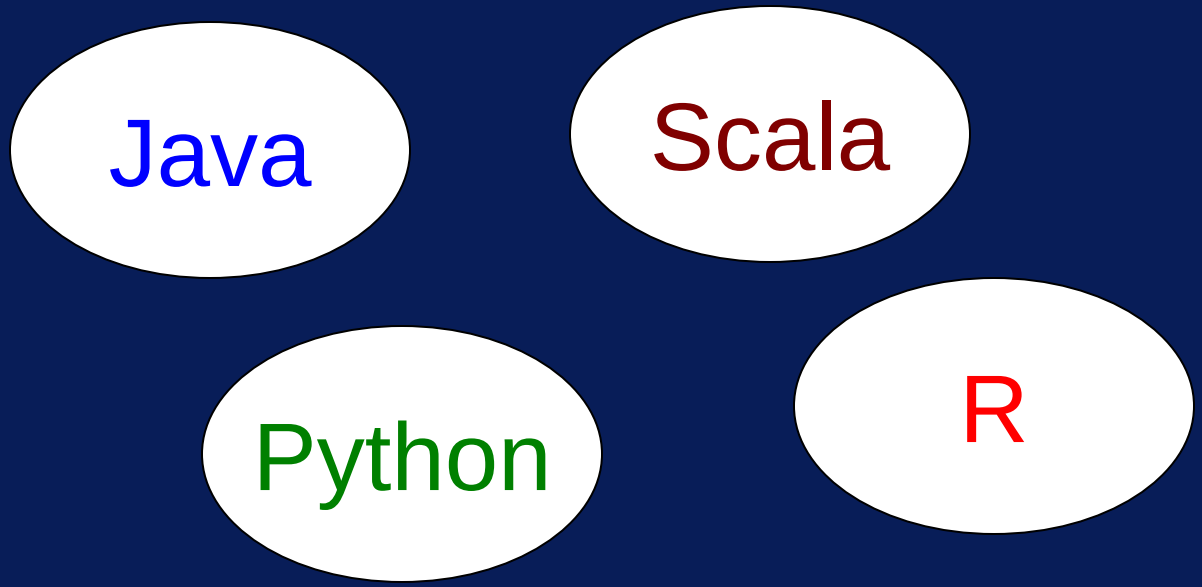
- **Write code to implement machine learning operations**

```
# Read and parse data
data = sc.textFile("data.txt")

# Decision tree for classification
model = DecisionTree.trainClassifier
        (parsedData, numClasses=2)
print(model.toDebugString())
model.save(sc, "decisionTreeModel")
```

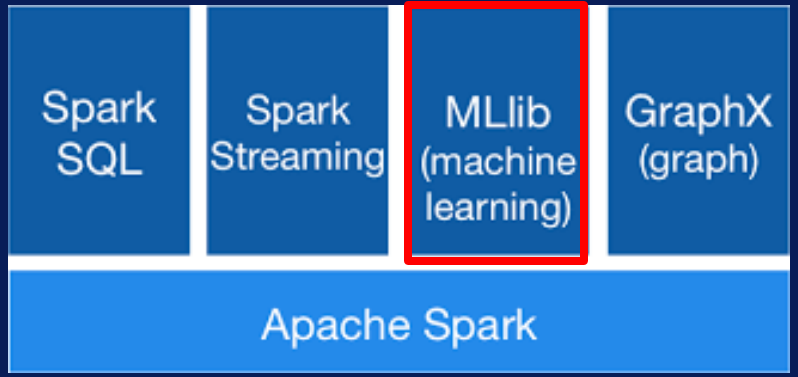
Spark MLlib

- **Provides APIs for**



Spark MLlib

- Distributed platform
- Scalable algorithms & techniques
- For large datasets
- Requires coding



KNIME & Spark MLlib