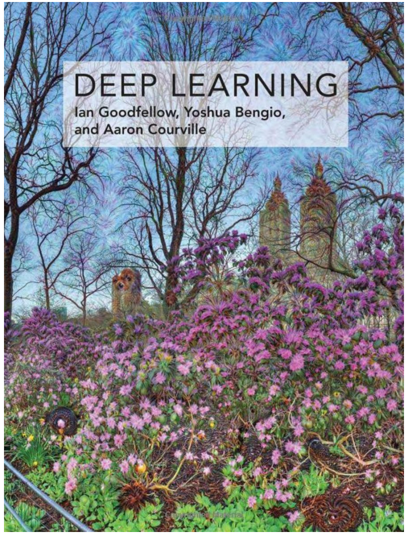
Machine Learning is Fun!

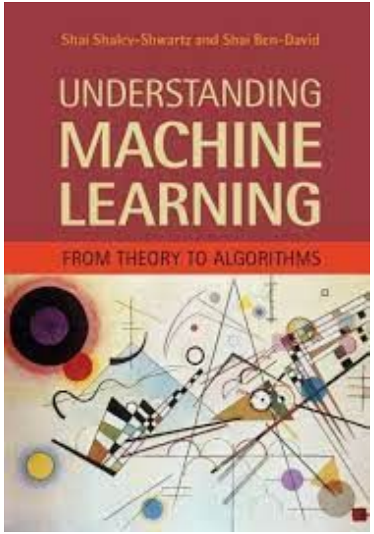
RAPIDMINER

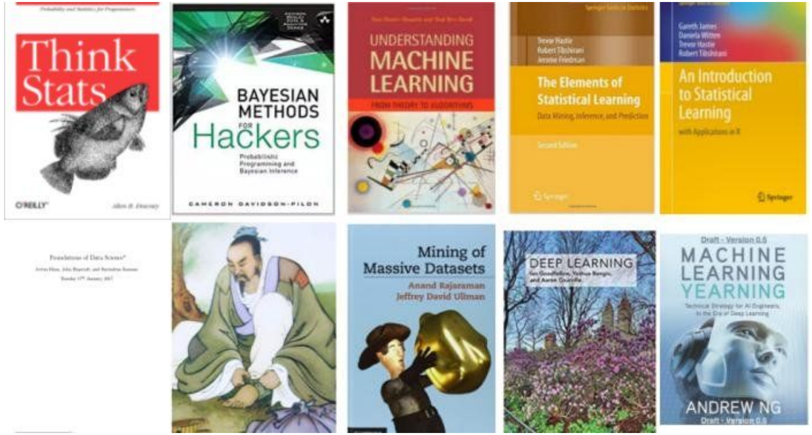
https://github.com/kaopanboonyuen/CS101

Reference

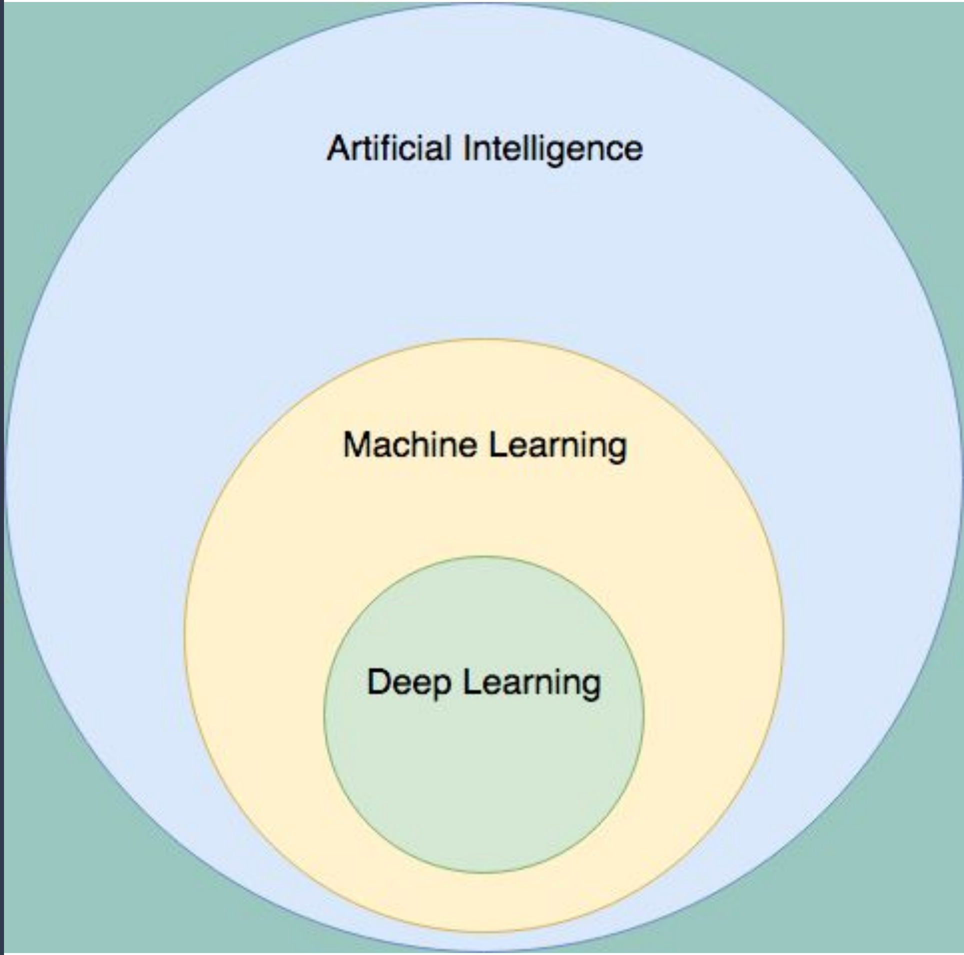
1. https://towardsdatascience.com/types-of-machine-learning-algorithms-you-should-know-953a08248861
2. https://omdena.com/blog/supervised-and-unsupervised-machine-learning/
3. https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/understanding-machine-learning-theory-algorithms.pdf
4.



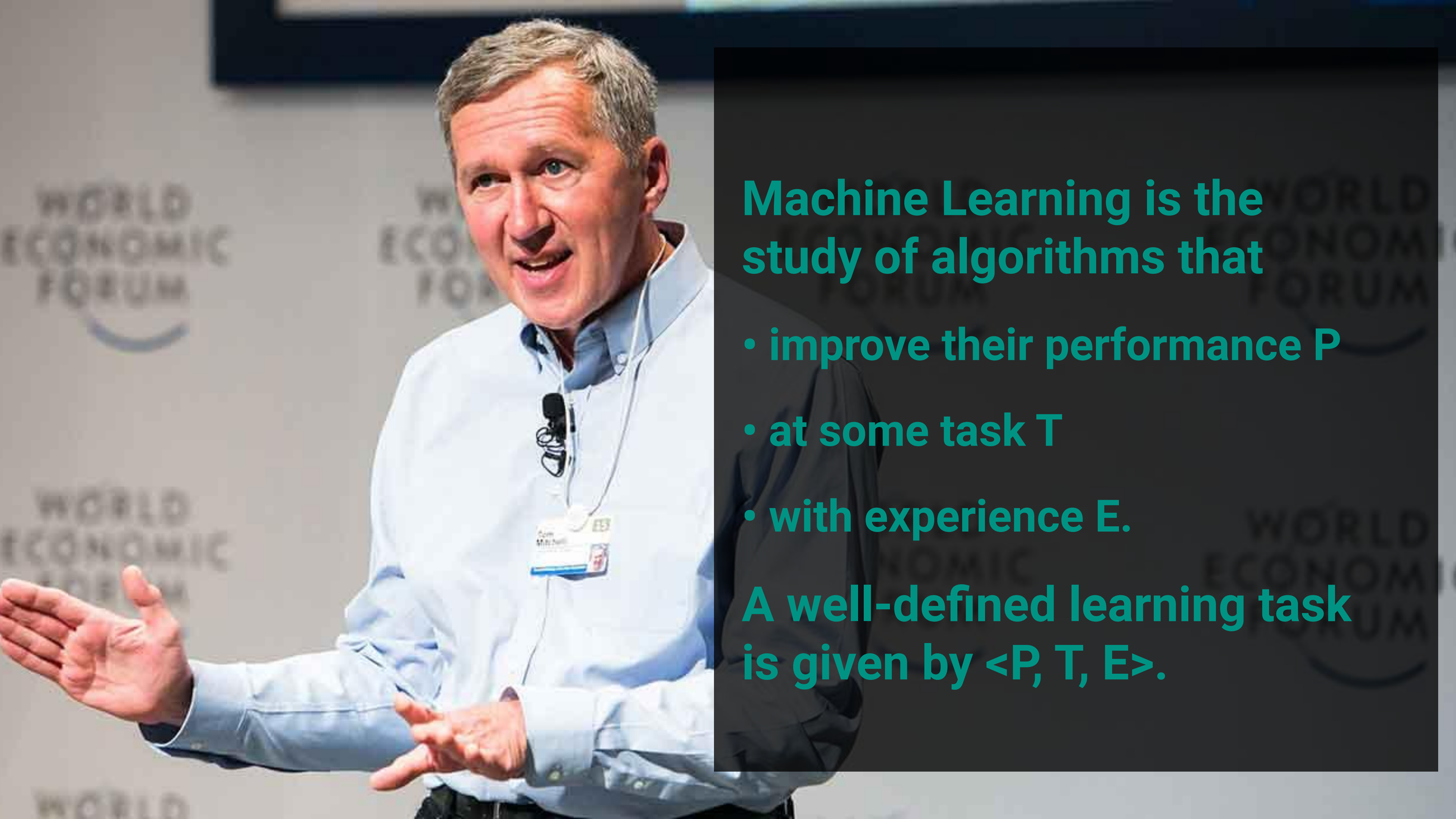

Shai Shalev-Shwartz and Shai Ben-David
UNDERSTANDING MACHINE LEARNING
FROM THEORY TO ALGORITHMS


Think Stats
O'REILLY
BAYESIAN METHODS FOR Hackers
CAMERON DAVIDSON-PILON
UNDERSTANDING MACHINE LEARNING
The Elements of Statistical Learning
An Introduction to Statistical Learning
Mining of Massive Datasets
DEEP LEARNING
MACHINE LEARNING YEARNING
ANDREW NG

What is Machine Learning?



> “Learning is any process by which a system improves performance from experience.”
>
> - Herbert Simon

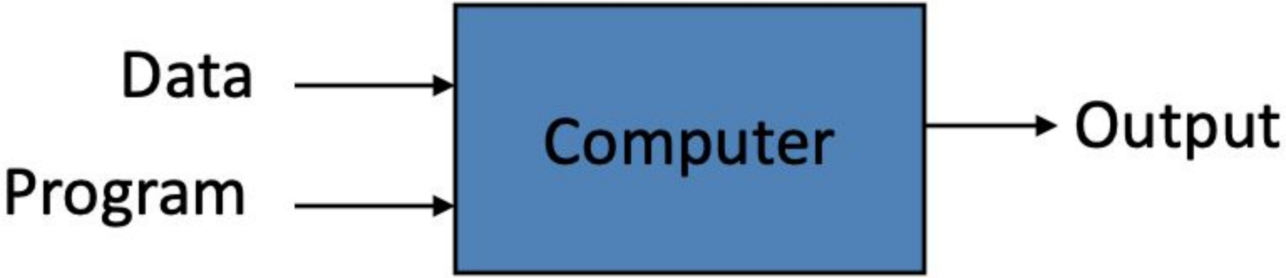


Machine Learning is the study of algorithms that

- improve their performance P
- at some task T
- with experience E.

A well-defined learning task is given by <P, T, E>.

Traditional Programming

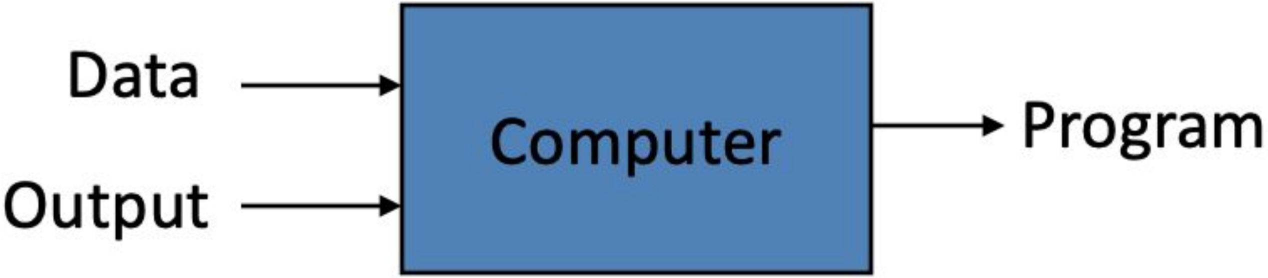

Data

Program

Computer

Output

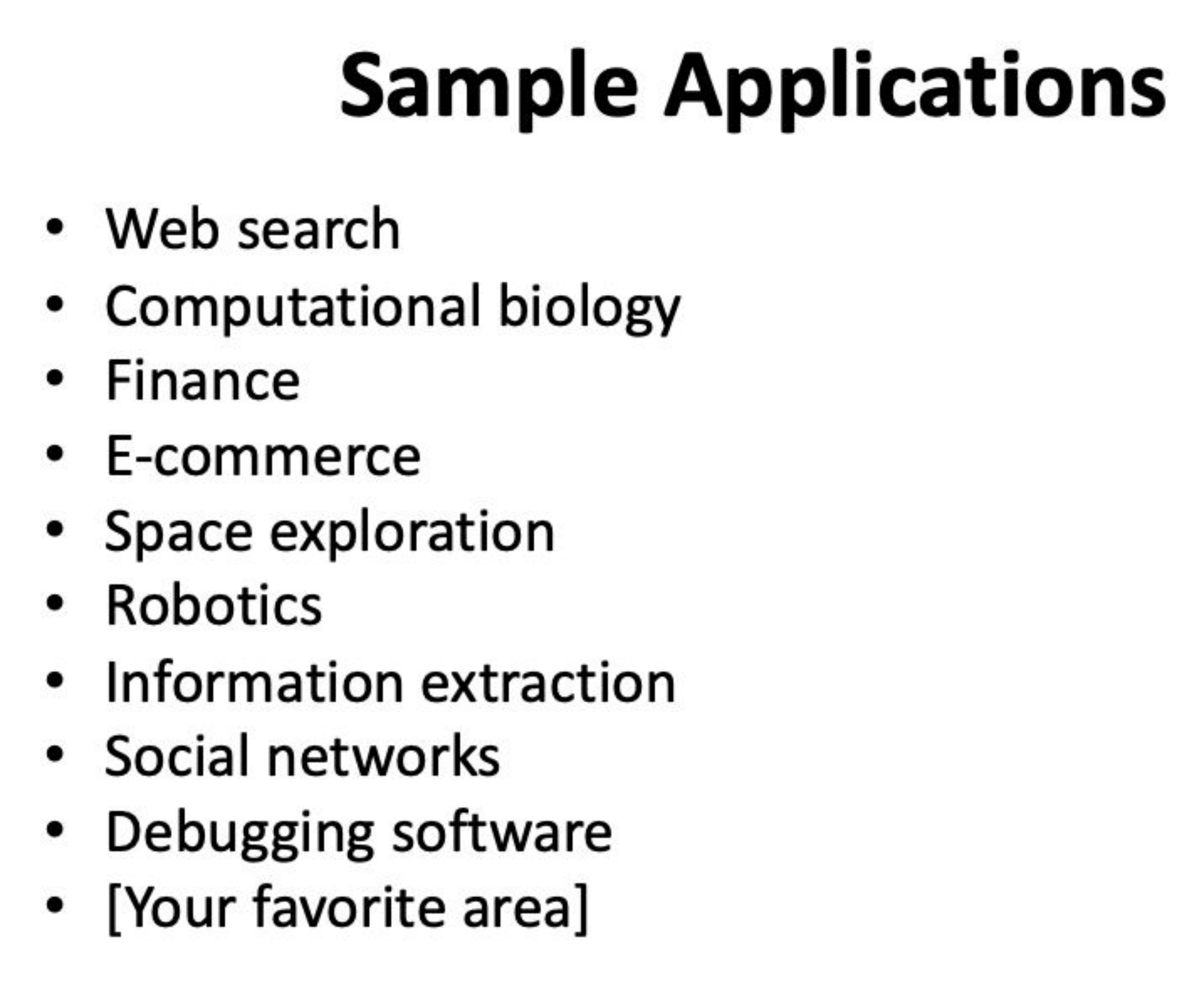
Machine Learning


Data

Output

Computer

Program



Sample Applications

- Web search
- Computational biology
- Finance
- E-commerce
- Space exploration
- Robotics
- Information extraction
- Social networks
- Debugging software
- [Your favorite area]

Types of Learning

- Supervised (inductive) learning

  Given: training data + desired outputs (labels)

- Unsupervised learning

  Given: training data (without desired outputs)

- Semi-supervised learning

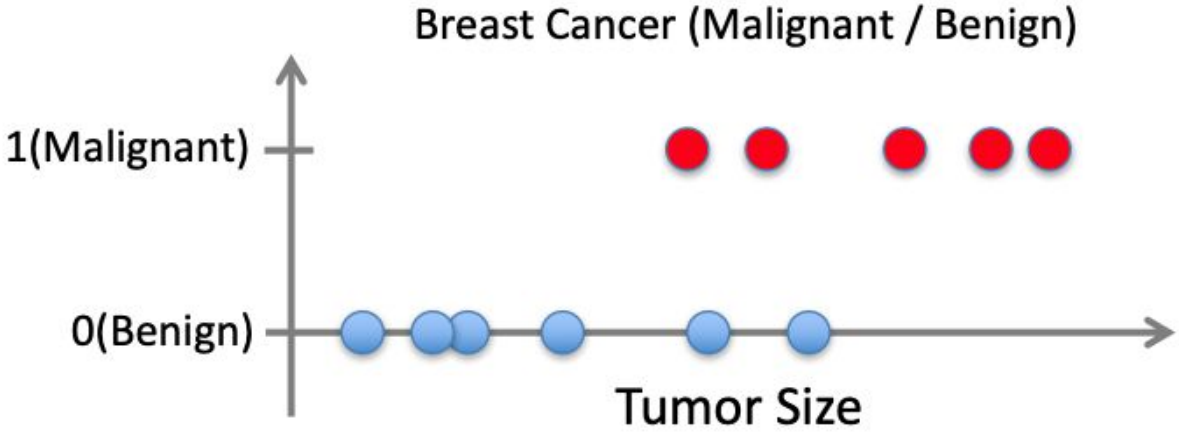
  Given: training data + a few desired outputs

- Reinforcement learning

  Rewards from sequence of actions

Supervised Learning: Classification

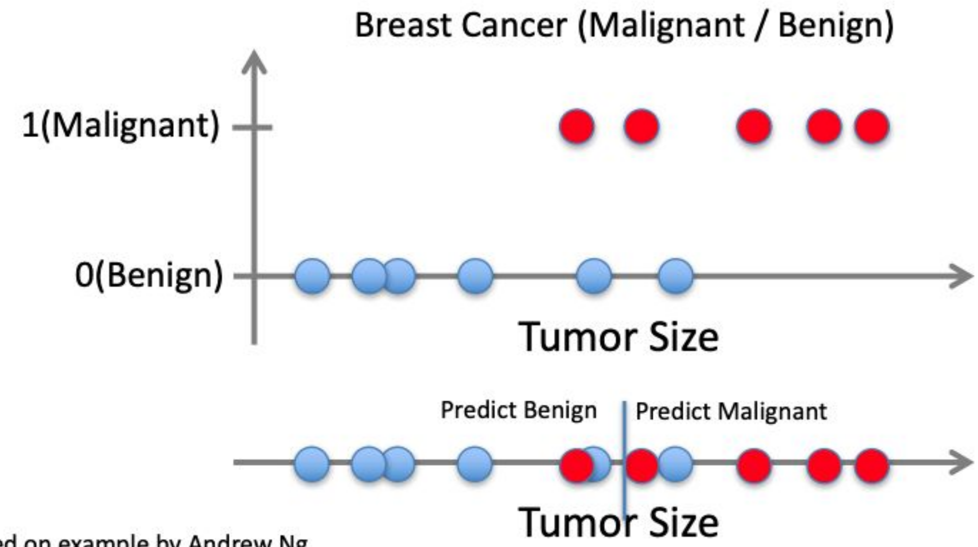
- Given $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Learn a function $f(x)$ to predict y given x
  - y is categorical == classification



Breast Cancer (Malignant / Benign)

Supervised Learning: Classification

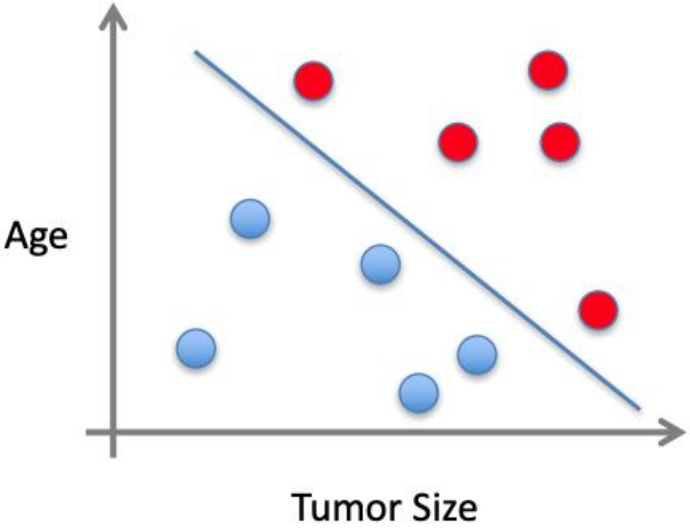
- Given $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Learn a function $f(x)$ to predict y given x
  - y is categorical == classification



Based on example by Andrew Ng

Supervised Learning

- x can be multi-dimensional
  - Each dimension corresponds to an attribute

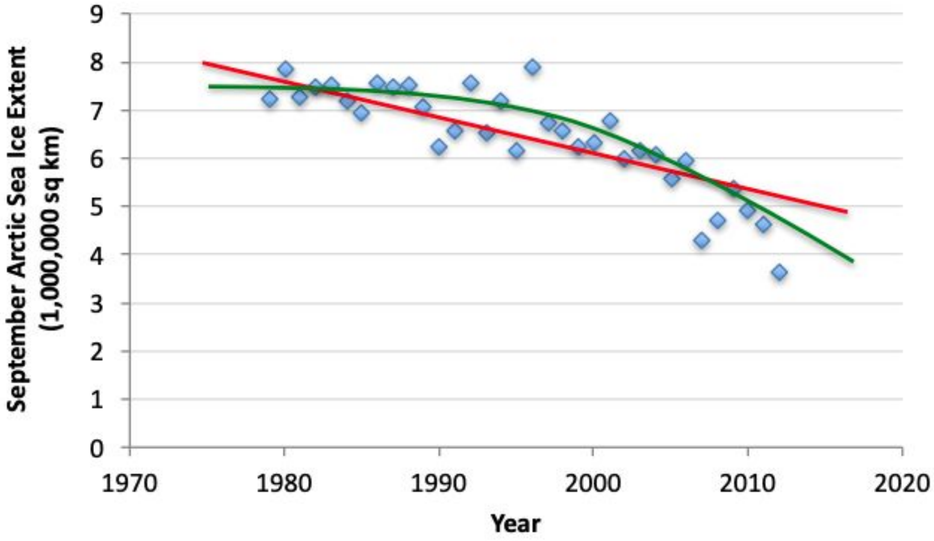


- Clump Thickness
- Uniformity of Cell Size
- Uniformity of Cell Shape

…

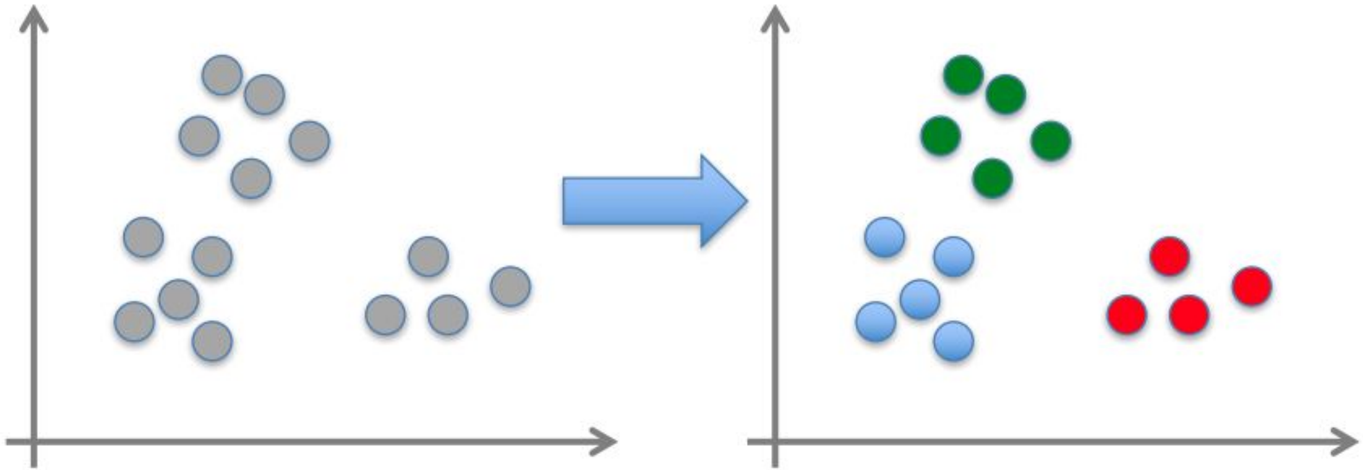
Supervised Learning: Regression

- Given (x_1, y_1), (x_2, y_2), ..., (x_n, y_n)
- Learn a function $f(x)$ to predict y given x
  - y is real-valued == regression



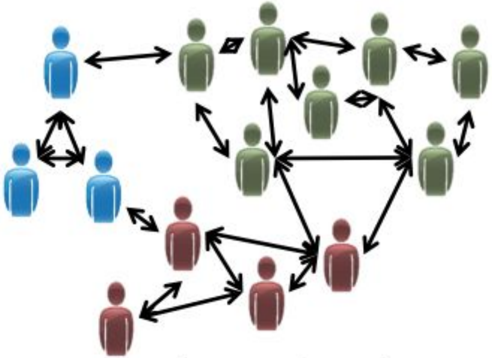
Unsupervised Learning

- Given x_1, x_2, …, x_n (without labels)
- Output hidden structure behind the x’s
  - E.g., clustering



Unsupervised Learning

Organize computing clusters



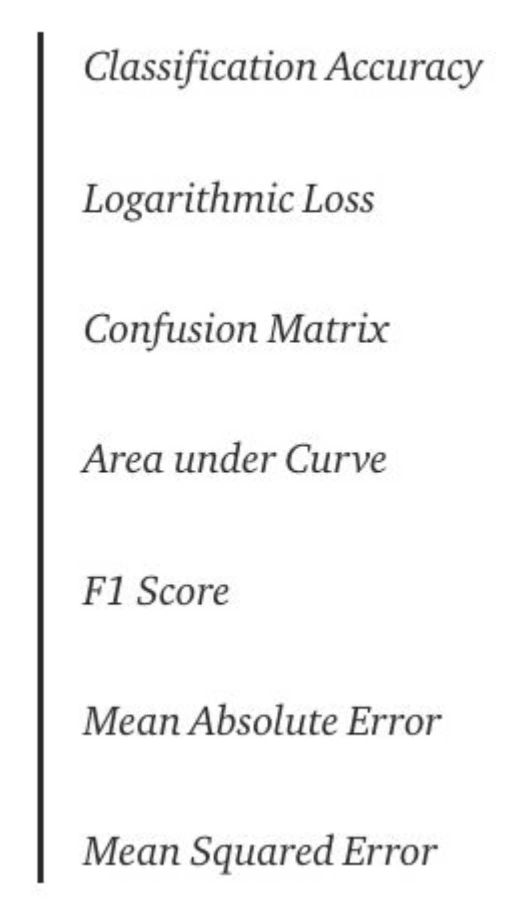
Social network analysis

Market segmentation

Image credit: NASA/JPL-Caltech/E. Churchwell (Univ. of Wisconsin, Ma

Astronomical data analysis

Metrics to Evaluate your Machine Learning Algorithm



Classification Accuracy

Logarithmic Loss

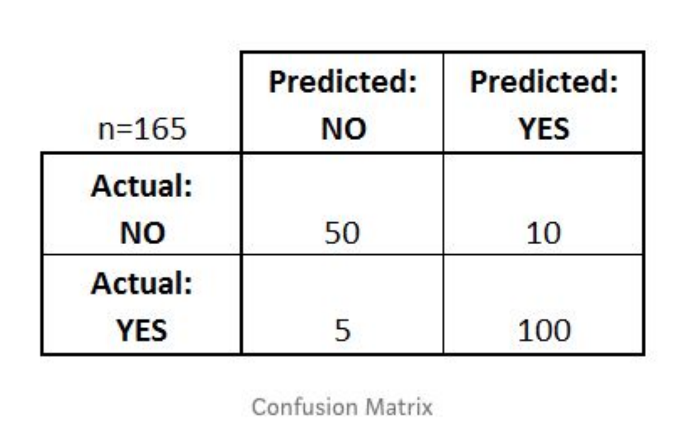
Confusion Matrix

Area under Curve

F1 Score

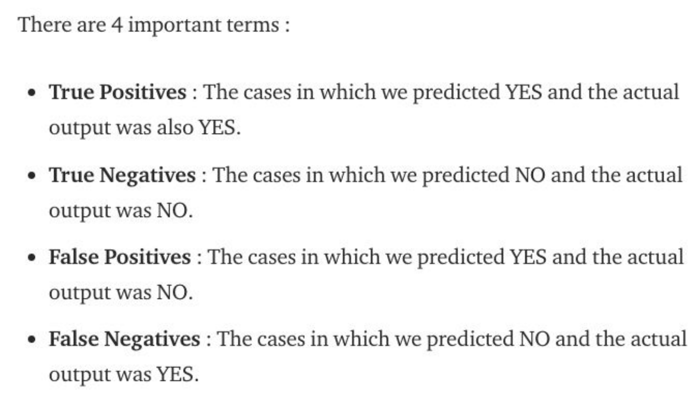
Mean Absolute Error

Mean Squared Error



n=165	**Predicted: NO**	**Predicted: YES**
Actual: NO	50	10
Actual: YES	5	100

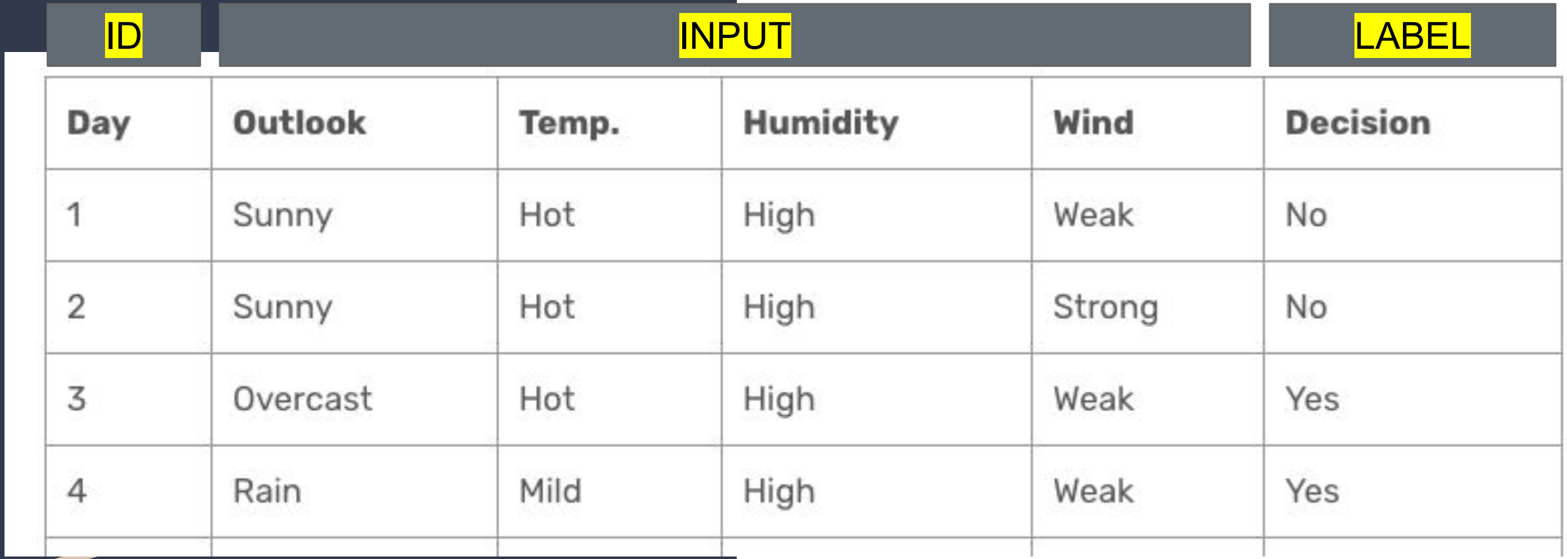
Confusion Matrix



There are 4 important terms :

- **True Positives** : The cases in which we predicted YES and the actual output was also YES.
- **True Negatives** : The cases in which we predicted NO and the actual output was NO.
- **False Positives** : The cases in which we predicted YES and the actual output was NO.
- **False Negatives** : The cases in which we predicted NO and the actual output was YES.

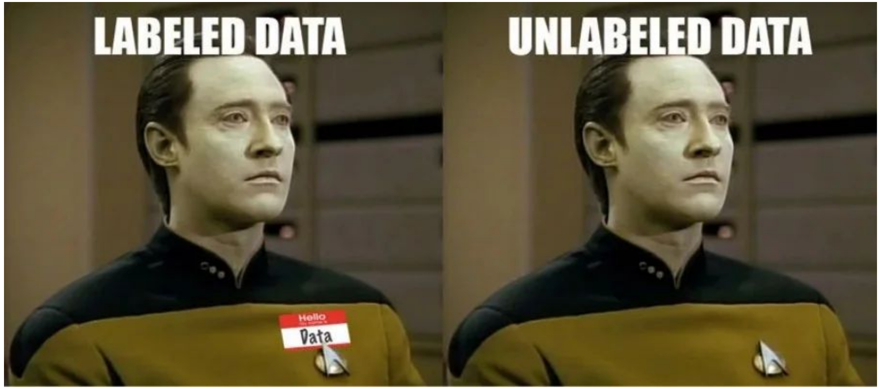
Terminology: Data table

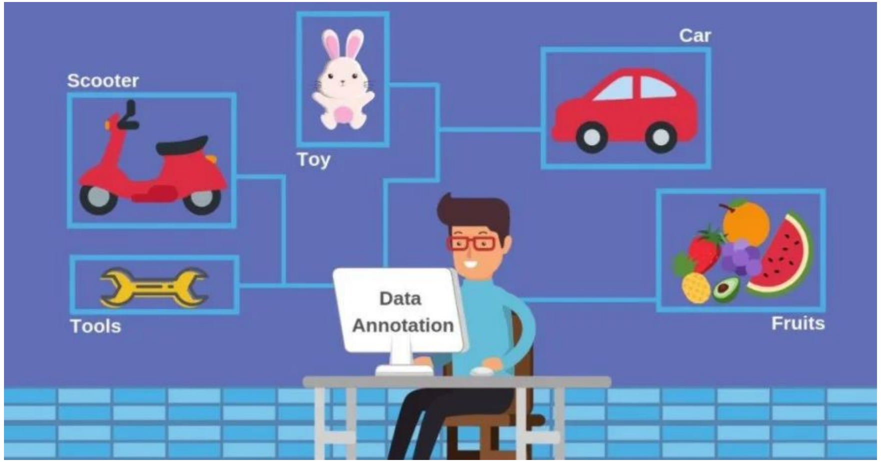


ID	INPUT				LABEL
Day	**Outlook**	**Temp.**	**Humidity**	**Wind**	**Decision**
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes

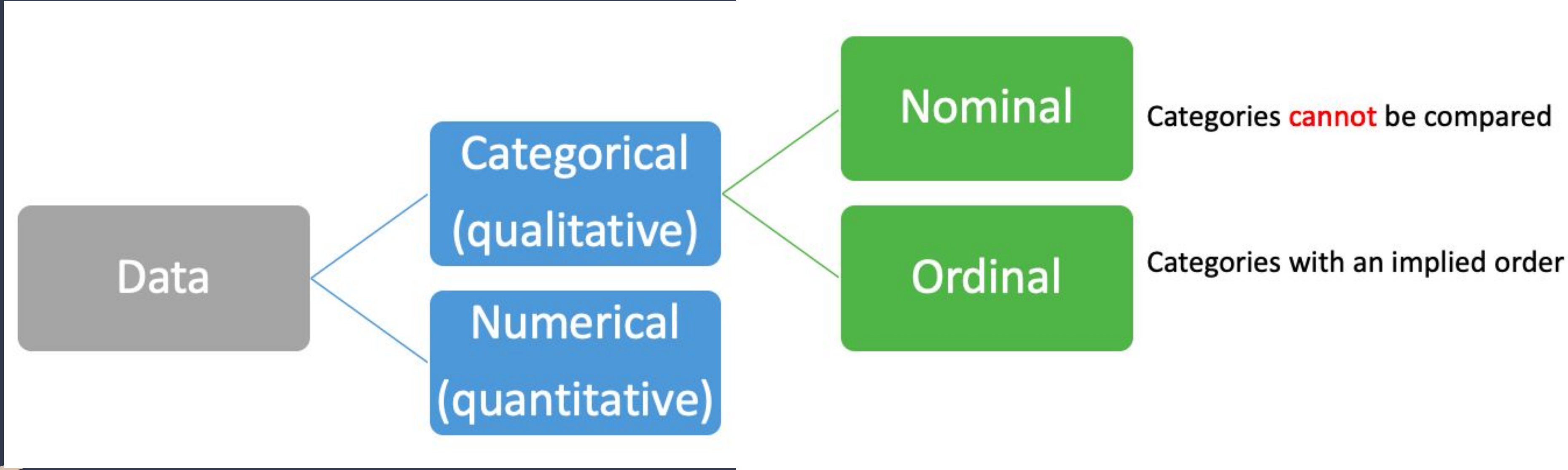
Credit to https://medium.com/@george.drakos62/decision-trees-decoded-part-1-23b45f69111c

What is the difference between supervised and unsupervised learning techniques?





Terminology: Data table



Data

Categorical (qualitative)

Numerical (quantitative)

Nominal

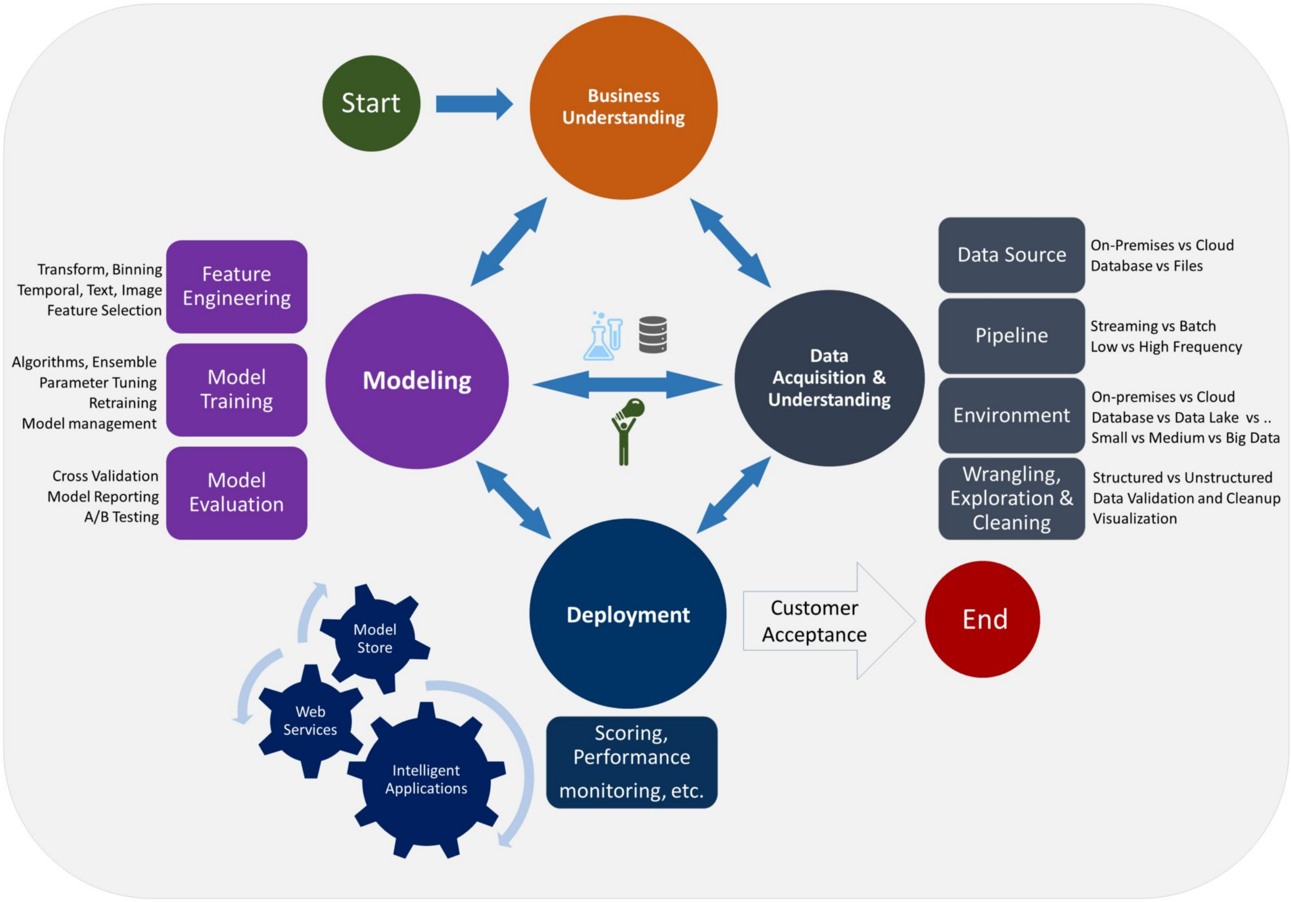
Categories cannot be compared

Ordinal

Categories with an implied order

Credit to Aj.Peerapon Vateekul

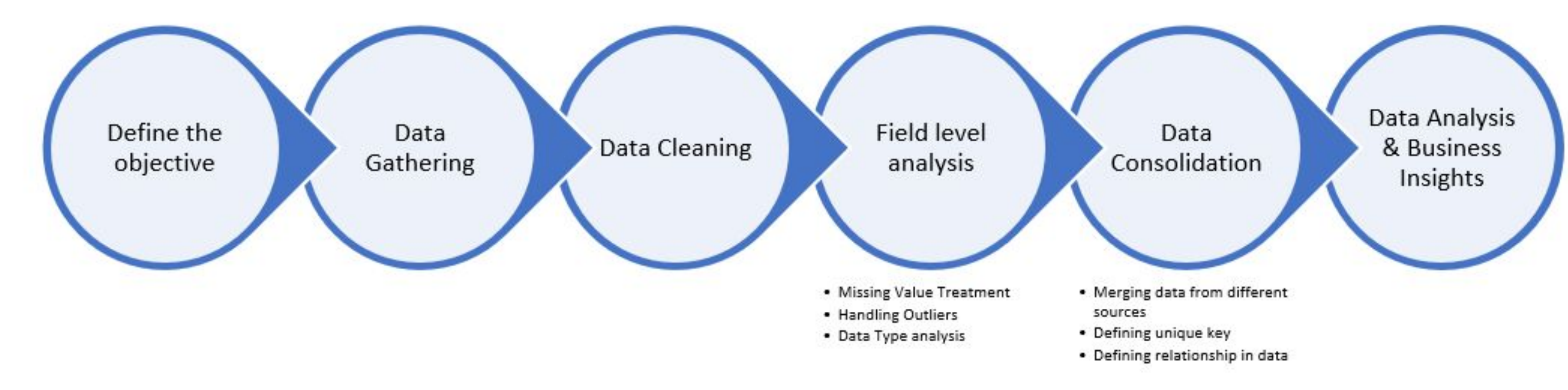
Data Science Lifecycle



Ref: https://medium.com/@rathi.ankit/data-science-introduction-e03773919c6

Data Analytics: Step by Step Approach

The Most Important Thing: Define Your Questions !!!!



- Missing Value Treatment
- Handling Outliers
- Data Type analysis

- Merging data from different sources
- Defining unique key
- Defining relationship in data

https://medium.com/datadriveninvestor/data-analytics-step-by-step-approach-757c6a0bd8a2

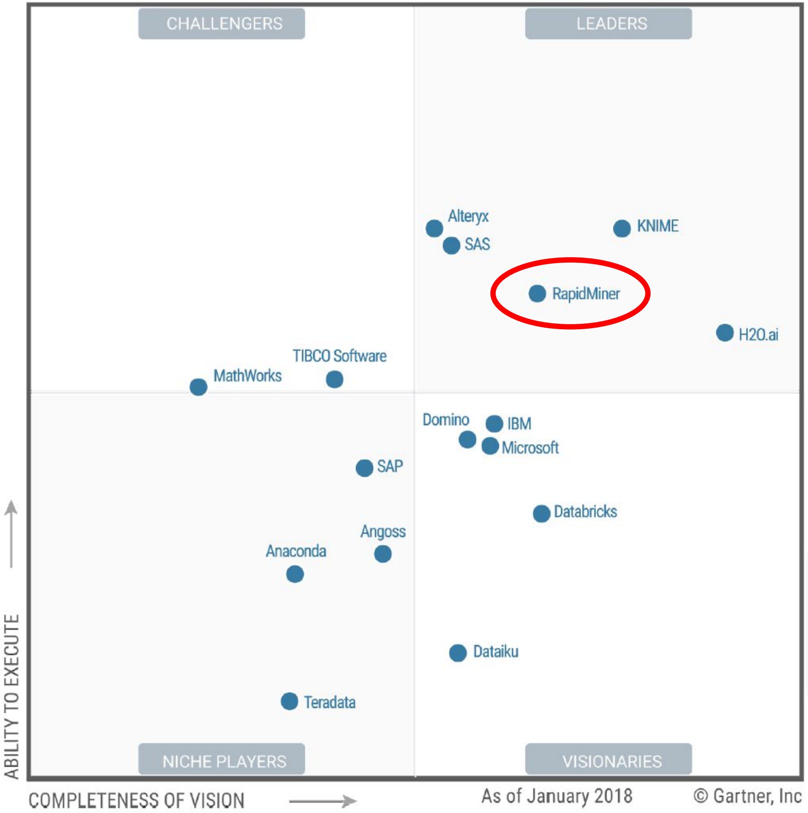
Let's explore the exercises!

rapidminer

kaggle

UCI

Data Mining Tools



As of January 2018 © Gartner, Inc

Code

sas®

R

python™

GUI

SAS® Enterprise Miner™ 12.3

WEKA The University of Waikato

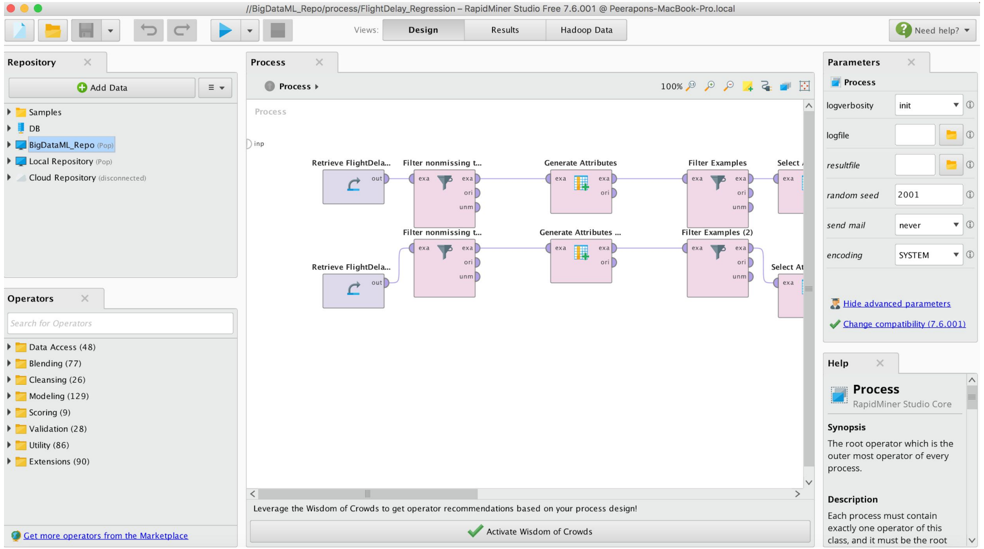
rapidminer

SPSS Modeler

https://i0.wp.com/r4stats.com/wp-content/uploads/2018/02/Gartner-2018.p

Rapid Miner

rapidminer



//BigDataML_Repo/process/FlightDelay_Regression – RapidMiner Studio Free 7.6.001 @ Peerapons-MacBook-Pro.local

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

logverbosity: init

logfile:

resultfile:

random seed: 2001

send mail: never

encoding: SYSTEM

Process

RapidMiner Studio Core

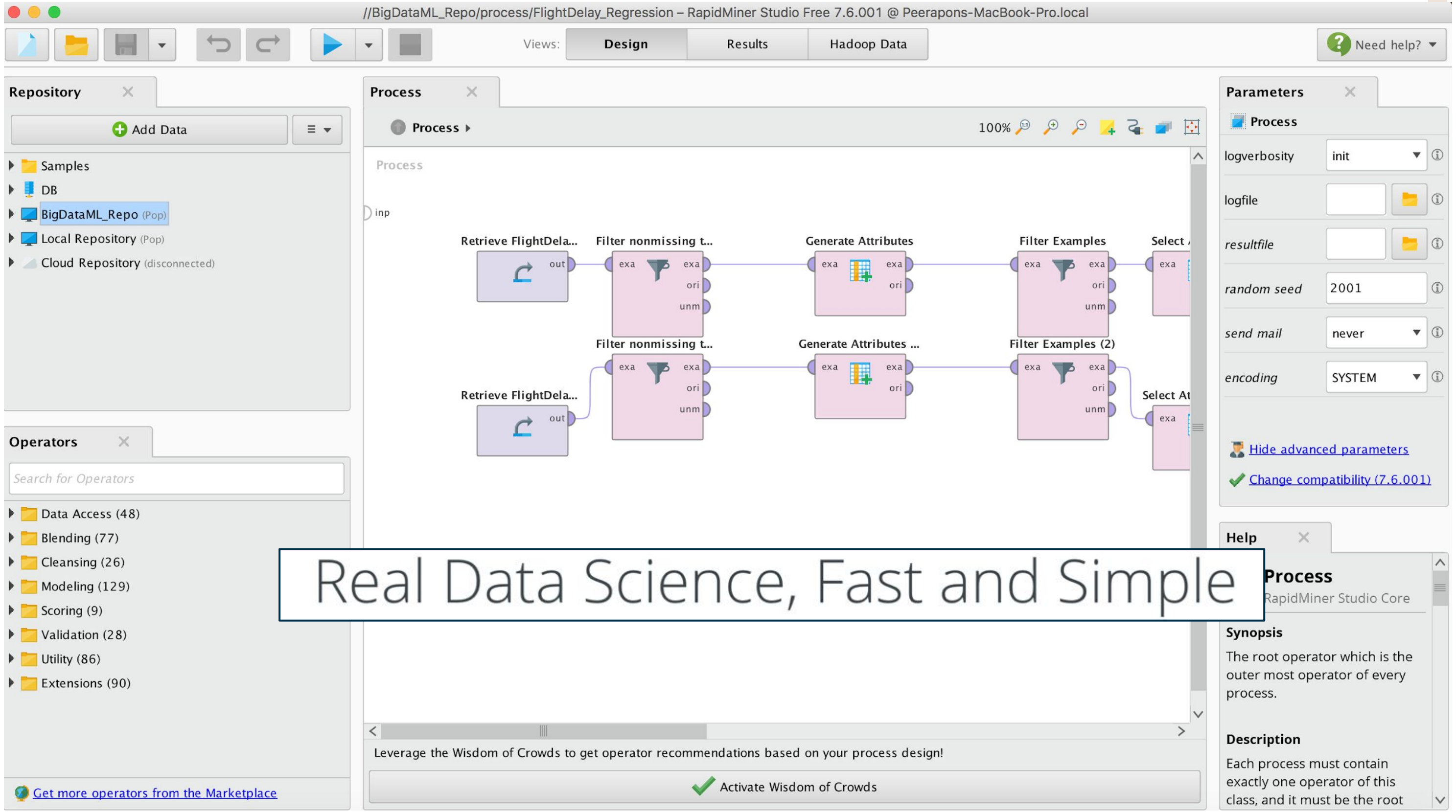
Synopsis

The root operator which is the outer most operator of every process.

Description

Each process must contain exactly one operator of this class, and it must be the root

What is Rapid Miner?



Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

logverbosity	init
logfile	
resultfile	
random seed	2001
send mail	never
encoding	SYSTEM

Process

RapidMiner Studio Core

Synopsis

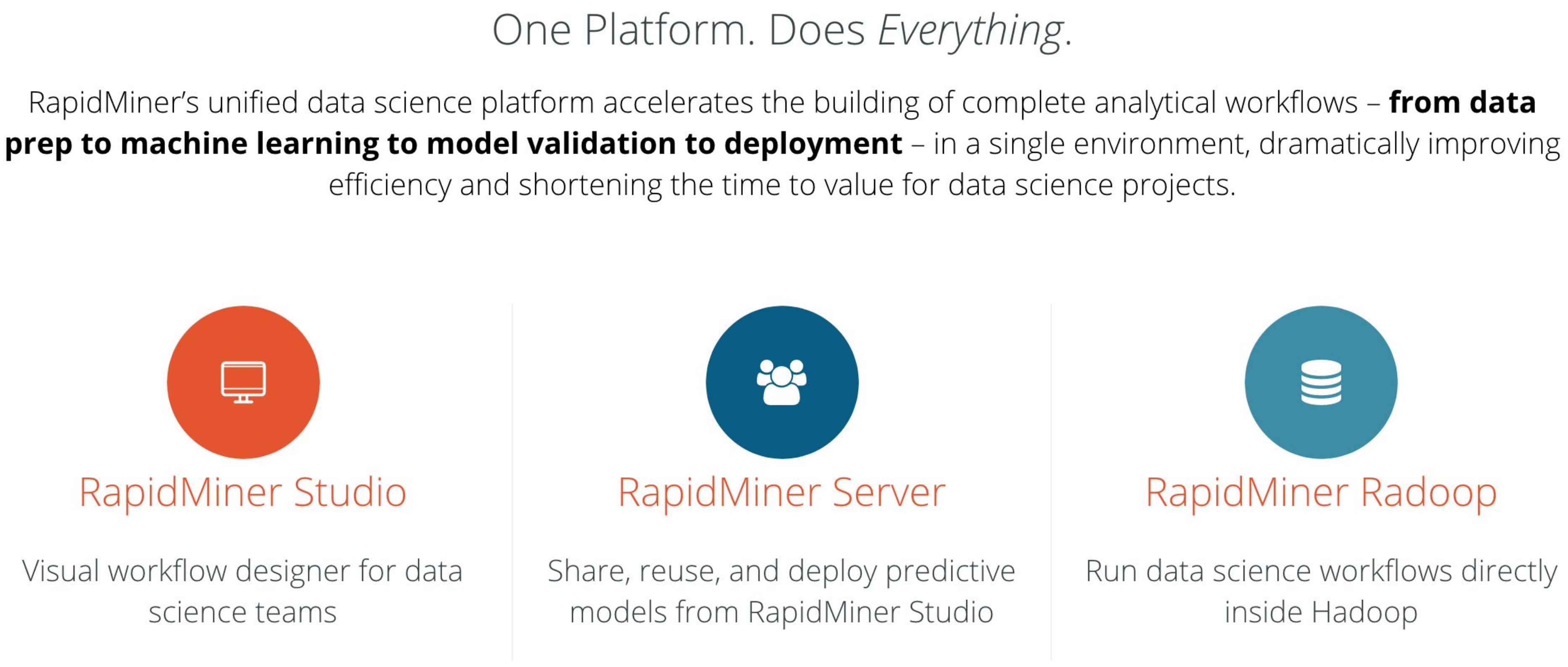
The root operator which is the outer most operator of every process.

Description

Each process must contain exactly one operator of this class, and it must be the root

Real Data Science, Fast and Simple

Platform



One Platform. Does *Everything*.

RapidMiner's unified data science platform accelerates the building of complete analytical workflows – **from data prep to machine learning to model validation to deployment** – in a single environment, dramatically improving efficiency and shortening the time to value for data science projects.

RapidMiner Studio

Visual workflow designer for data science teams

RapidMiner Server

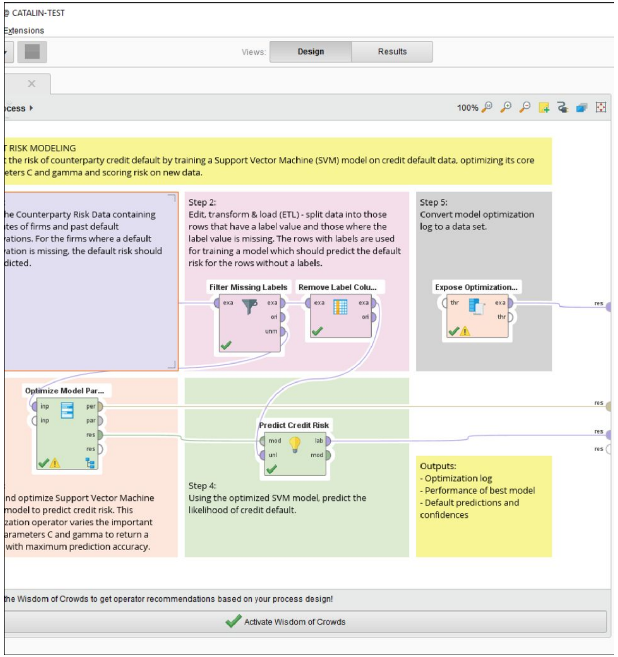
Share, reuse, and deploy predictive models from RapidMiner Studio

RapidMiner Radoop

Run data science workflows directly inside Hadoop

+ Process and Operators

- An analytical workflow is called “Process”
- Each process consists of one or more “Operators”
- Connect output of an operator to input of the next operator



T RISK MODELING
t the risk of counterparty credit default by training a Support Vector Machine (SVM) model on credit default data, optimizing its core eters C and gamma and scoring risk on new data.

he Counterparty Risk Data containing tes of firms and past default ations. For the firms where a default ation is missing, the default risk should dicted.

Step 2:
Edit, transform & load (ETL) - split data into those rows that have a label value and those where the label value is missing. The rows with labels are used for training a model which should predict the default risk for the rows without a labels.

Step 5:
Convert model optimization log to a data set.

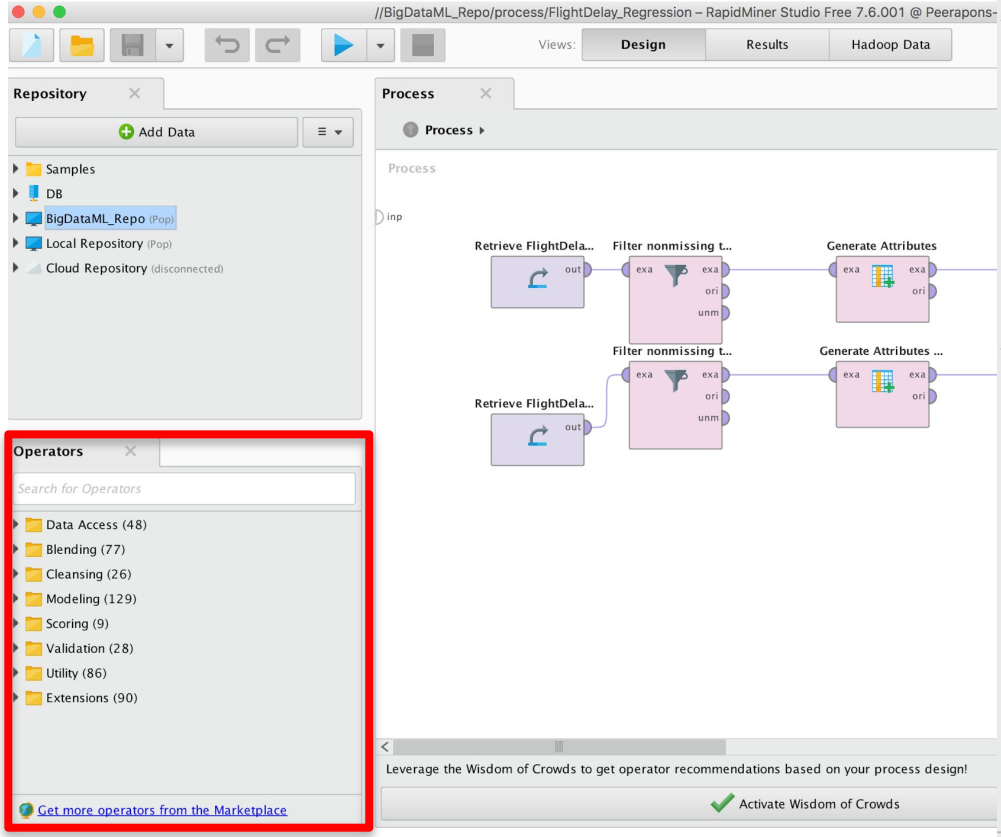
nd optimize Support Vector Machine model to predict credit risk. This zation operator varies the important arameters C and gamma to return a with maximum prediction accuracy.

Step 4:
Using the optimized SVM model, predict the likelihood of credit default.

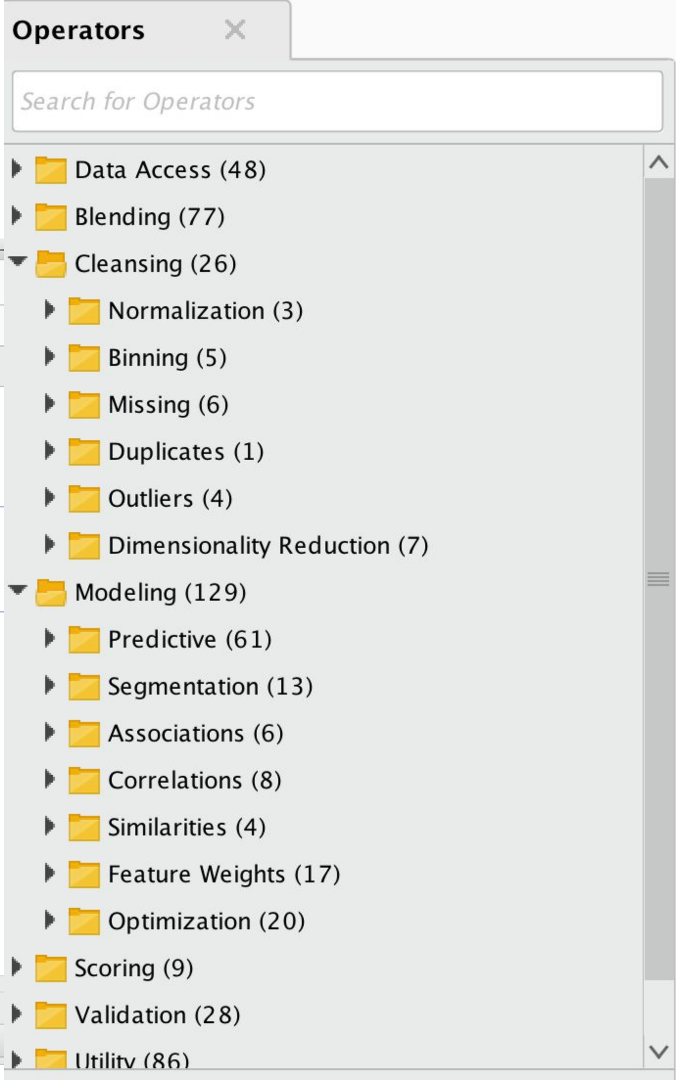
Outputs:
- Optimization log
- Performance of best model
- Default predictions and confidences

he Wisdom of Crowds to get operator recommendations based on your process design!

+ Tools in Rapid Miner



Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

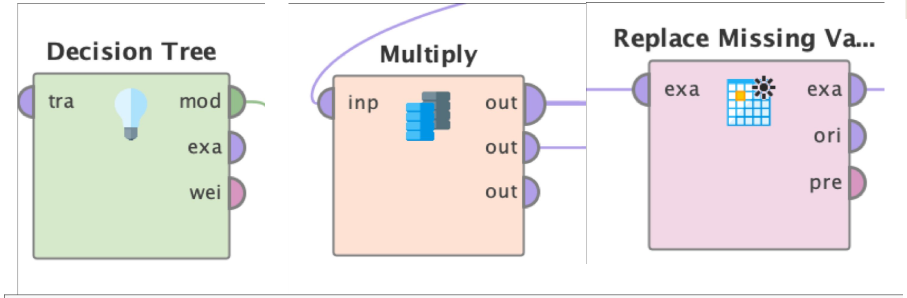


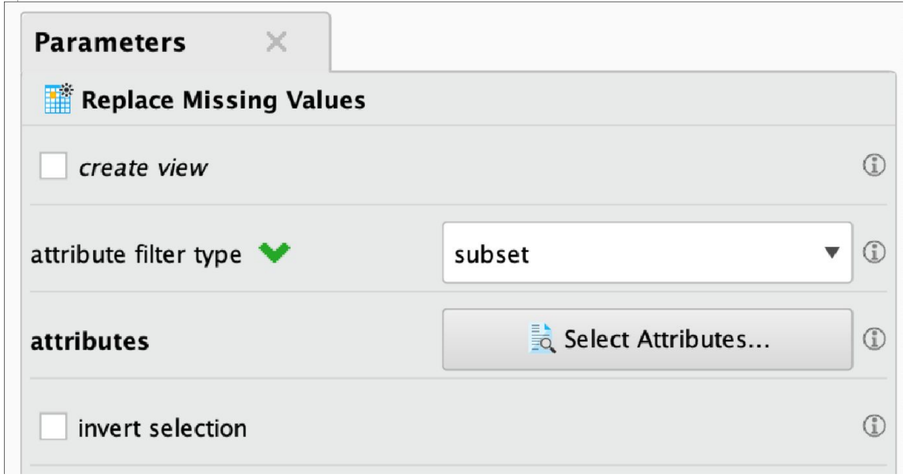
Examples of Operators

- Data accessing
  - e.g. file, cassandra, mongoDB, Amazon S3
- Data blending
  - e.g. mapping, filter, select, aggregate, split
- Data cleansing
  - e.g. normalize, deduplication, outlier
- Data modeling
  - e.g. Bayesian, decision tree, neural net

Anatomy of Operators in RapidMiner

- “inp” – Input
- “out” – Output
- “thr” – Through
- “fil” – File
- “exp” – Example
- “ori” - Original
- “lab” – Label
- “tra” – Training
- “mod” – Model
- “wei” -Weight




Parameters ✕

Replace Missing Values

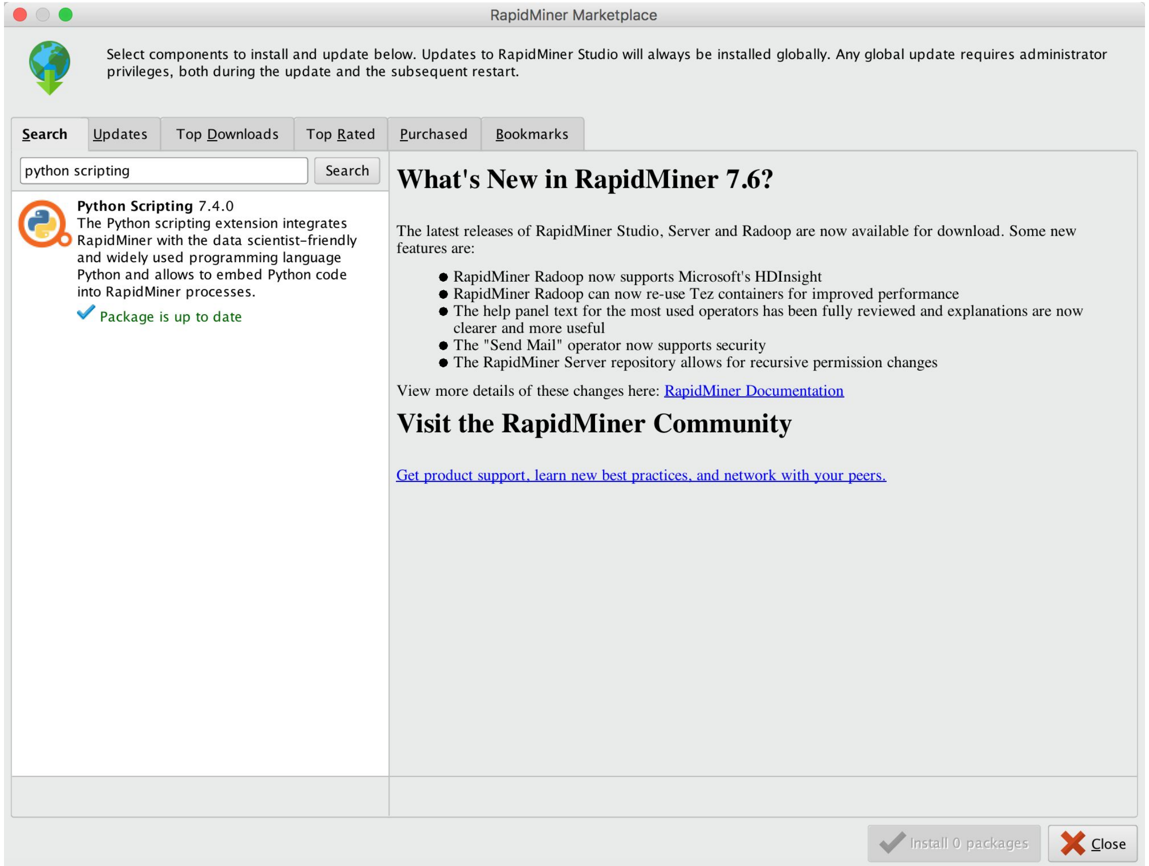
□ *create view*

attribute filter type ✓ subset ▼

attributes Select Attributes...

□ invert selection

+ Extensions in Rapid Miner: Marketplace



Select components to install and update below. Updates to RapidMiner Studio will always be installed globally. Any global update requires administrator privileges, both during the update and the subsequent restart.

Python Scripting 7.4.0
The Python scripting extension integrates RapidMiner with the data scientist-friendly and widely used programming language Python and allows to embed Python code into RapidMiner processes.
✓ Package is up to date

What's New in RapidMiner 7.6?

The latest releases of RapidMiner Studio, Server and Radoop are now available for download. Some new features are:

- RapidMiner Radoop now supports Microsoft's HDInsight
- RapidMiner Radoop can now re-use Tez containers for improved performance
- The help panel text for the most used operators has been fully reviewed and explanations are now clearer and more useful
- The "Send Mail" operator now supports security
- The RapidMiner Server repository allows for recursive permission changes

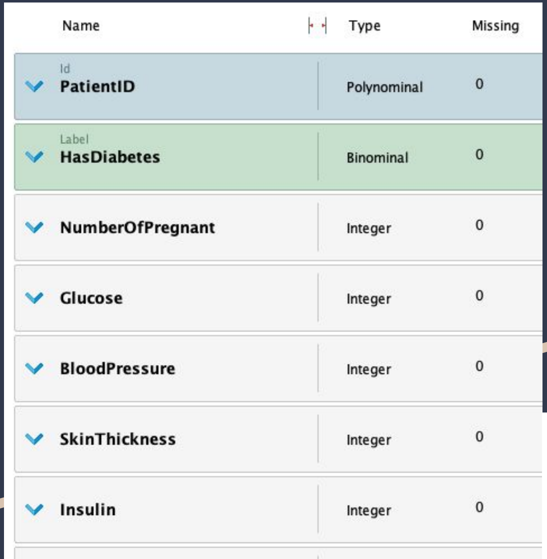
View more details of these changes here: RapidMiner Documentation

Visit the RapidMiner Community

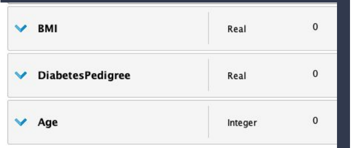
Get product support, learn new best practices, and network with your peers.

Machine Learning Exercises using RapidMiner

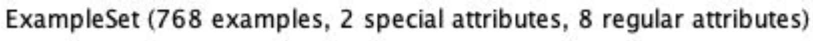
Lab1: Classification Lab | Pima Indians Diabetes



Name	Type	Missing
Id PatientID	Polynominal	0
Label HasDiabetes	Binominal	0
NumberOfPregnant	Integer	0
Glucose	Integer	0
BloodPressure	Integer	0
SkinThickness	Integer	0
Insulin	Integer	0



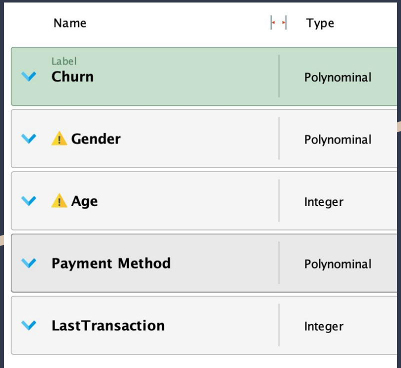
BMI	Real	0
DiabetesPedigree	Real	0
Age	Integer	0



ExampleSet (768 examples, 2 special attributes, 8 regular attributes)

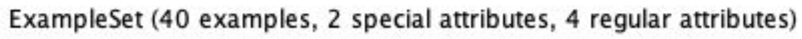
https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

Lab2: Customer Churn



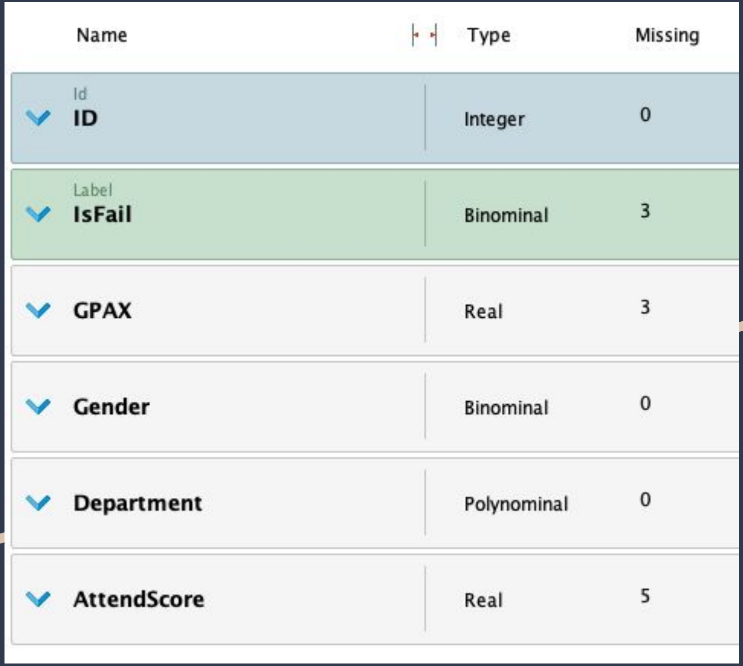
Name	Type
Label Churn	Polynominal
⚠ Gender	Polynominal
⚠ Age	Integer
Payment Method	Polynominal
LastTransaction	Integer

Loyal Customers
High Value High Risk
Low Value High Risk
High Value Low Risk



ExampleSet (40 examples, 2 special attributes, 4 regular attributes)

Exercise: Student Grade



Name	Type	Missing
Id ID	Integer	0
Label IsFail	Binominal	3
GPAX	Real	3
Gender	Binominal	0
Department	Polynominal	0
AttendScore	Real	5