COSC 3337 : Data Science I

N. Rizk

College of Natural and Applied Sciences

Department of Computer Science

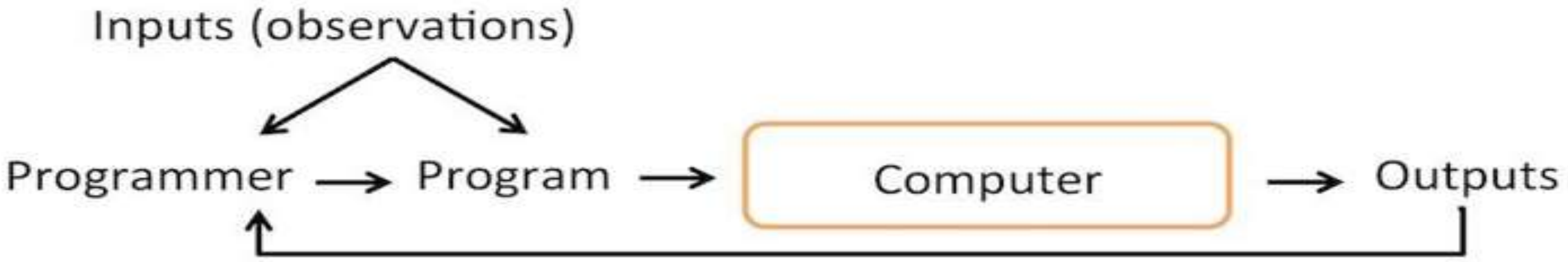
University of Houston

Machine  Learning

"Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed"

— Arthur L. Samuel, AI pioneer, 1959

The Traditional Programming Paradigm



Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed
– Arthur Samuel (1959)

Machine Learning

Learning

“A computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P*, if its performance at tasks in *T*, as measured by *P*, improves with experience *E*.”

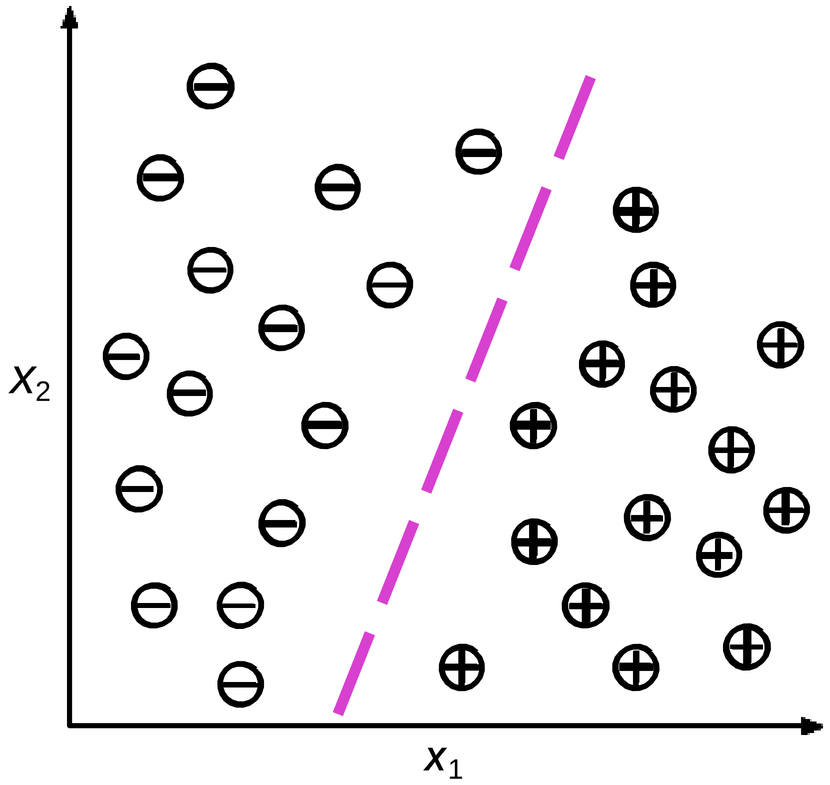
— Tom Mitchell, Professor at Carnegie Mellon University

Learning in this context is the process of gaining understanding by constructing models of observed data with the intention to use them for prediction.

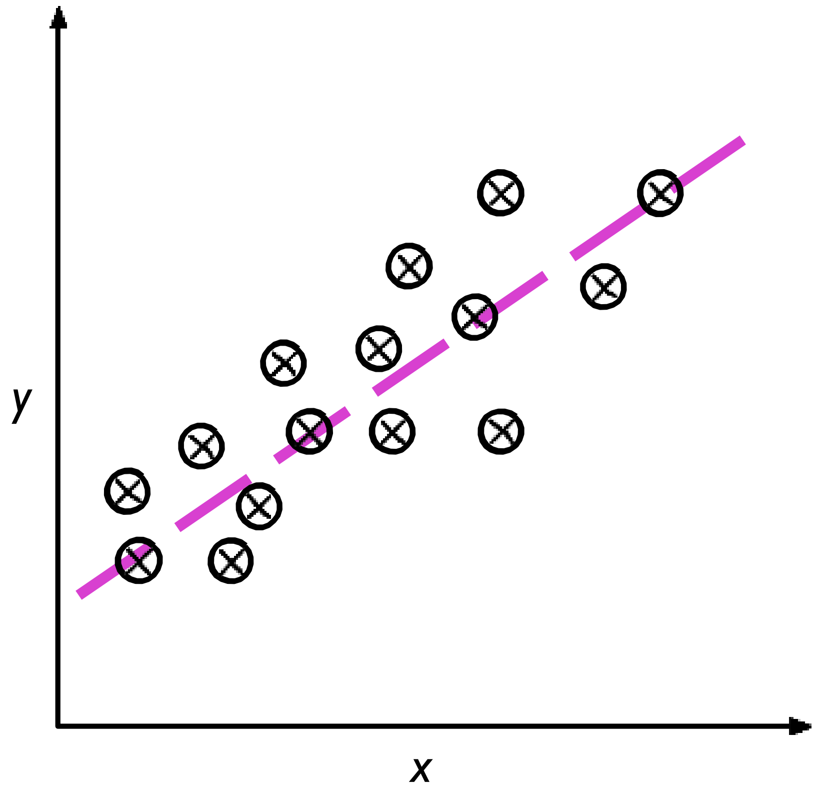
Categories of Machine Learning

Supervised Learning	- Labeled data - Direct feedback - Predict outcome/future

Supervised Learning: Classification



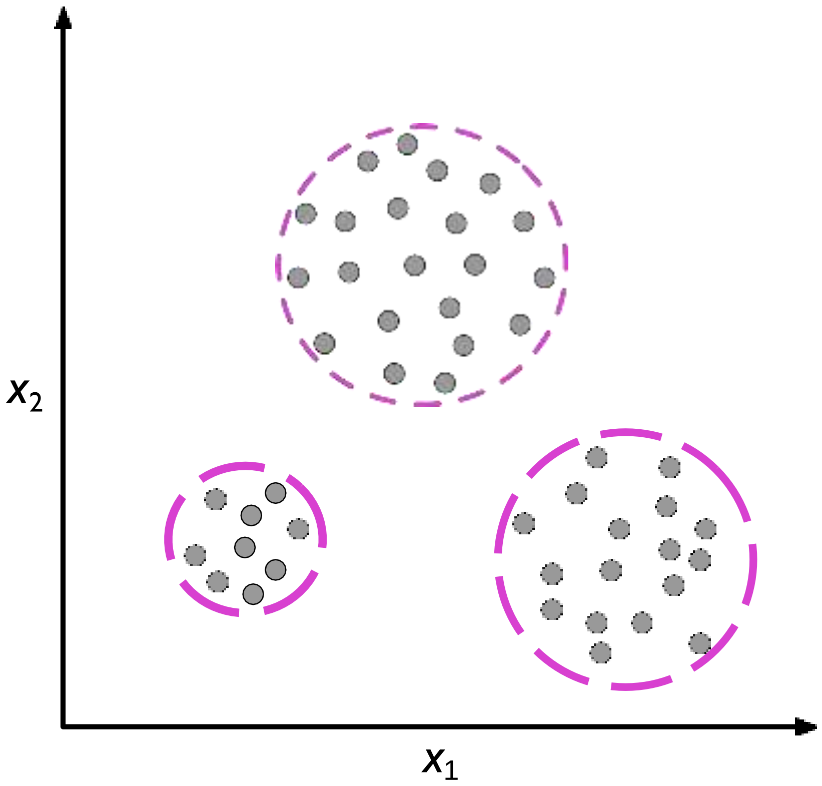
Supervised Learning: Regression



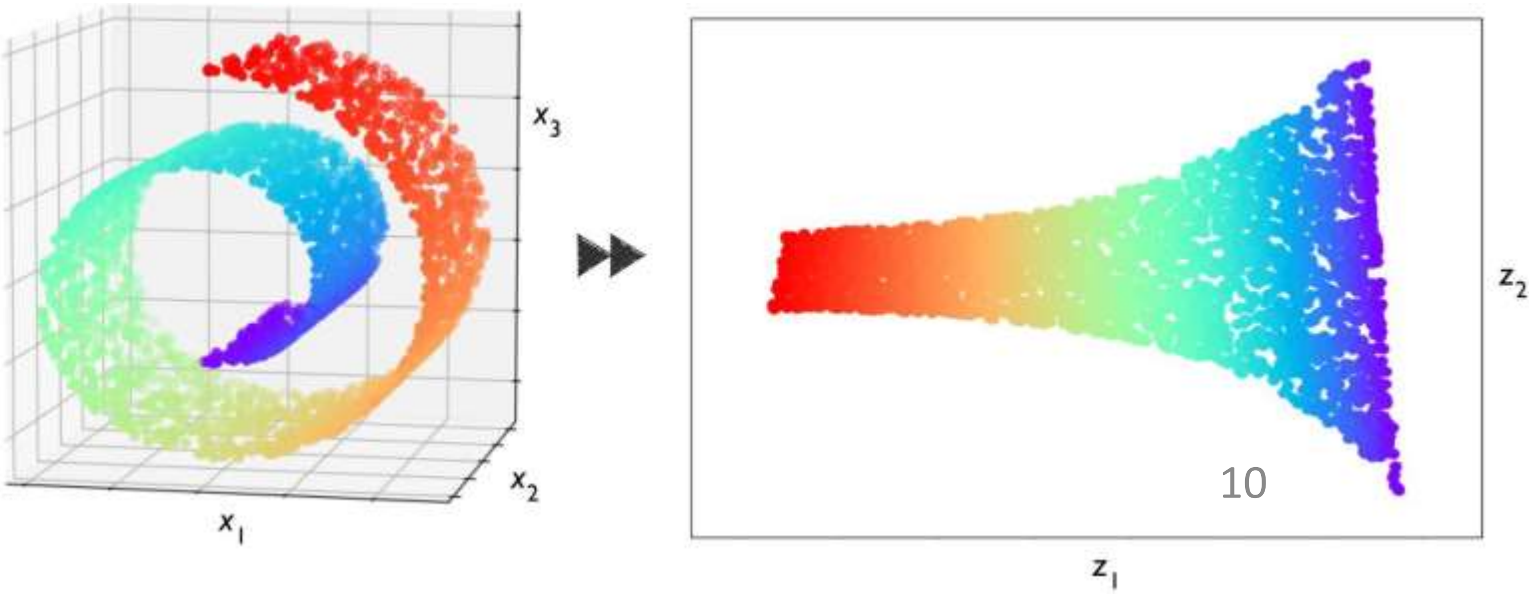
Categories of Machine Learning

Supervised Learning	Labeled data Direct feedback Predict outcome/future
Unsupervised Learning	No labels/targets No feedback Find hidden structure in data

Unsupervised Learning - Clustering



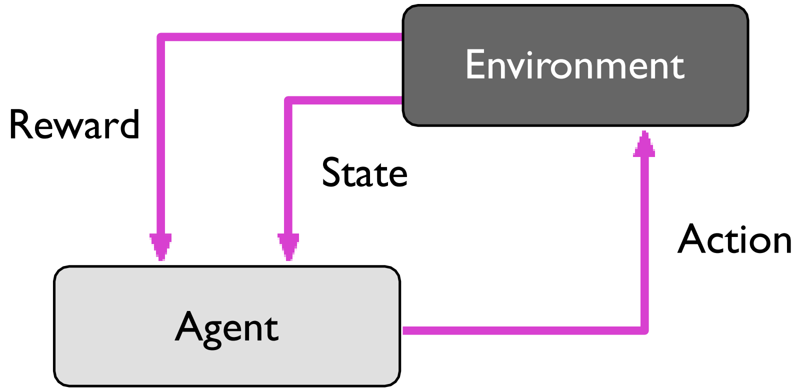
Unsupervised Learning: Dimensionality Reduction



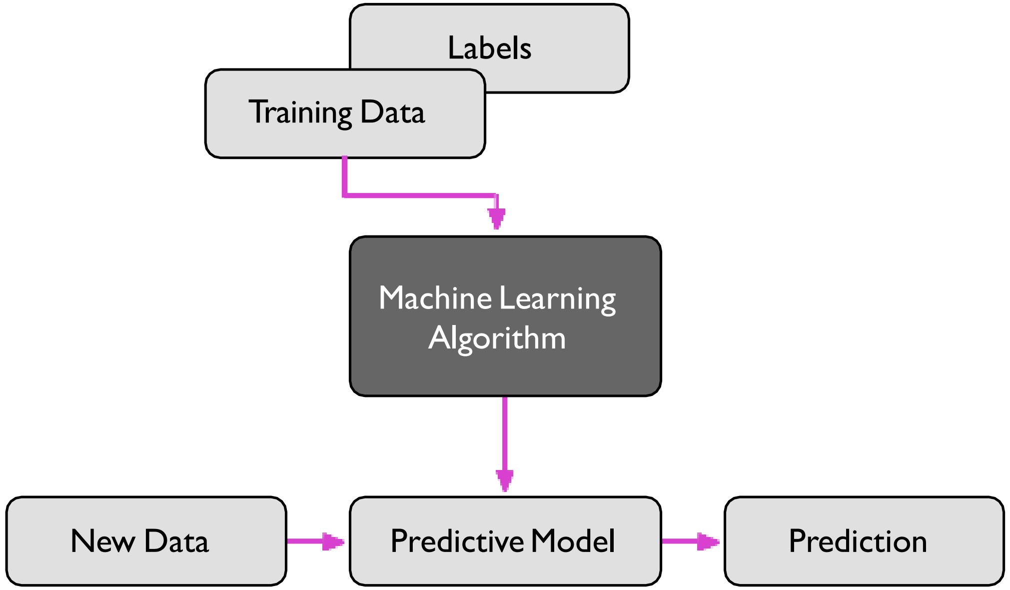
Categories of Machine Learning

Supervised Learning	Labeled data Direct feedback Predict outcome/future
Unsupervised Learning	No labels/targets No feedback Find hidden structure in data
Reinforcement Learning	Decision process Reward system Learn series of actions

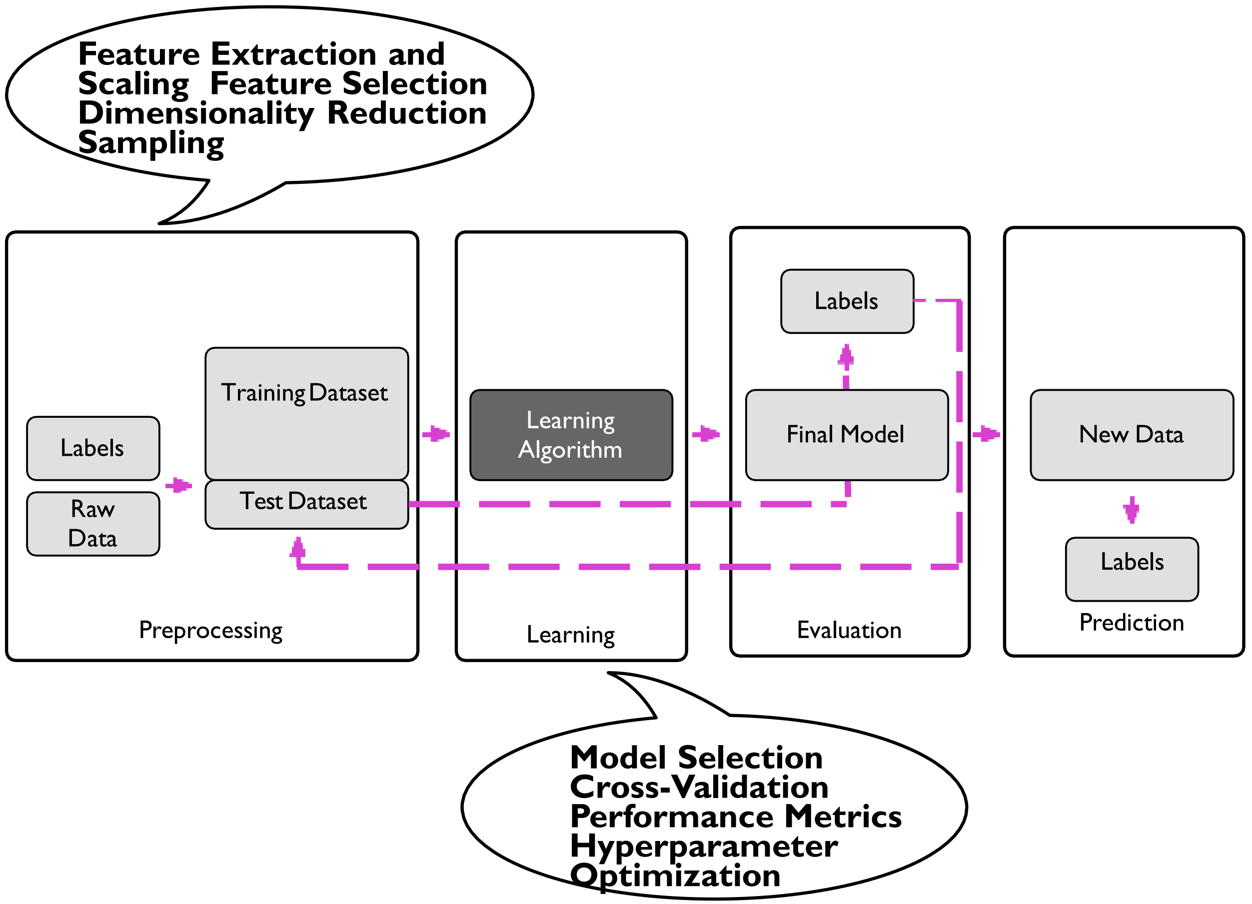
Reinforcement Learning



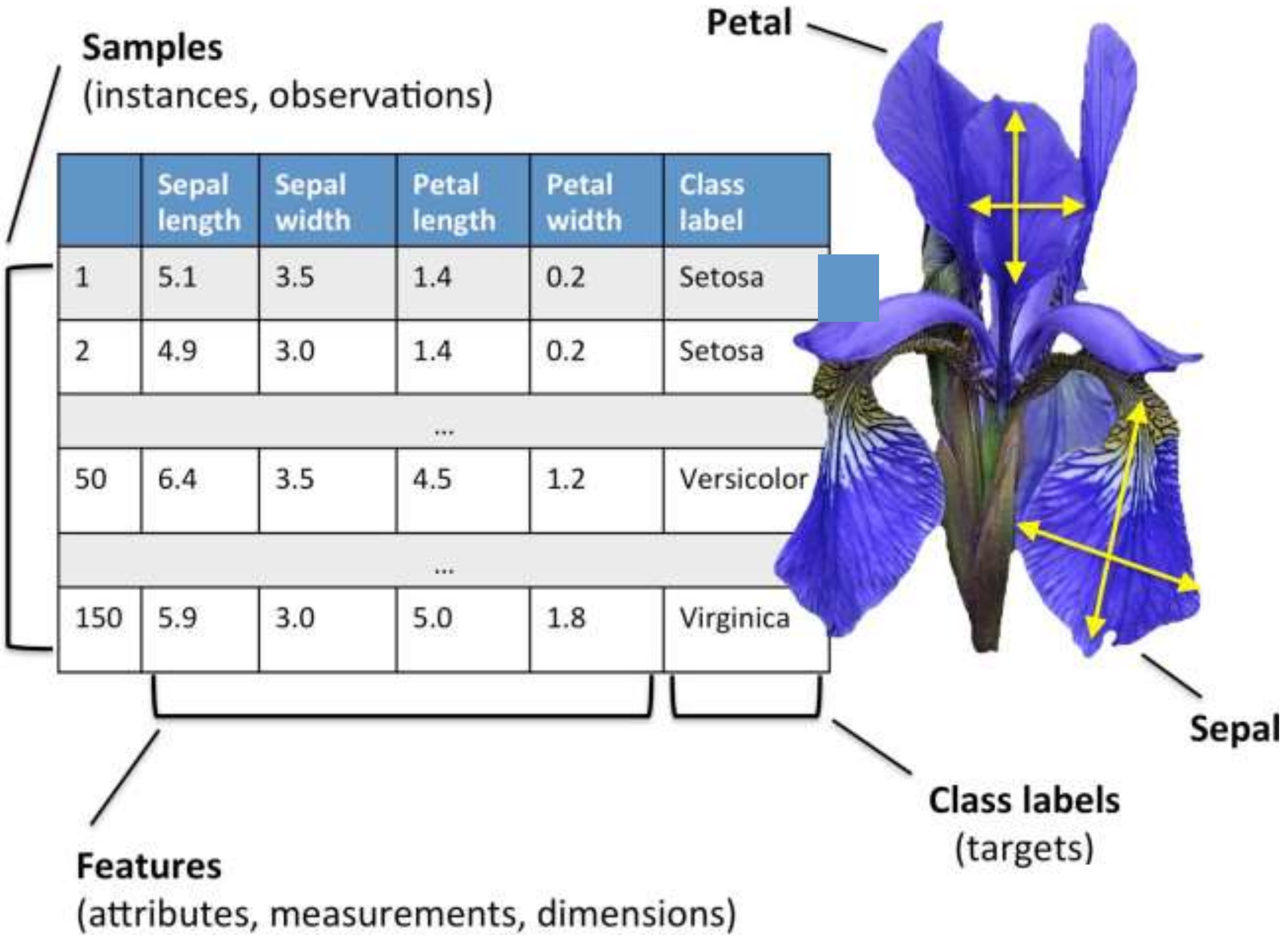
Supervised Learning Workflow



Supervised Learning Workflow Detailed



IRIS data set



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
			...		
50	6.4	3.5	4.5	1.2	Versicolor
			...		
150	5.9	3.0	5.0	1.8	Virginica

5 Steps for Approaching an Application

1. Define the problem to be solved.
2. Collect (labeled) data.
3. Choose an algorithm class.
4. Choose an optimization metric for learning the model.
5. Choose a metric for evaluating the model.

Objective Functions

Maximize the posterior probabilities (e.g., naive Bayes)

- Maximize a fitness function (genetic programming)
- Maximize the total reward/value function (reinforcement learning)
- Maximize information gain/minimize child node impurities (CART decision tree classification)
- Minimize a mean squared error cost (or loss) function (CART, decision tree regression, linear regression, adaptive linear neurons
- Maximize log-likelihood or minimize cross-entropy loss (or cost) function
- Minimize hinge loss (support vector machine)

Metrics

- Accuracy (1-Error)
- ROC AUC
- Precision
- Recall
- (Cross) Entropy
- Likelihood
- Squared Error/MSE
- L-norms
- Utility
- Fitness

...

Categorizing Machine Learning Algorithms

Lazy : K - Nearest Neighbor, Case - Based Reasoning
Eager : Decision Tree, Naive Bayes, Artificial Neural Networks

- Eager vs Lazy;

Lazy learner:

Just store Data set without learning from it

Start classifying data when it receive Test data

So it takes less time learning and more time classifying data

Eager learner:

When it receive data set it starts classifying (learning)

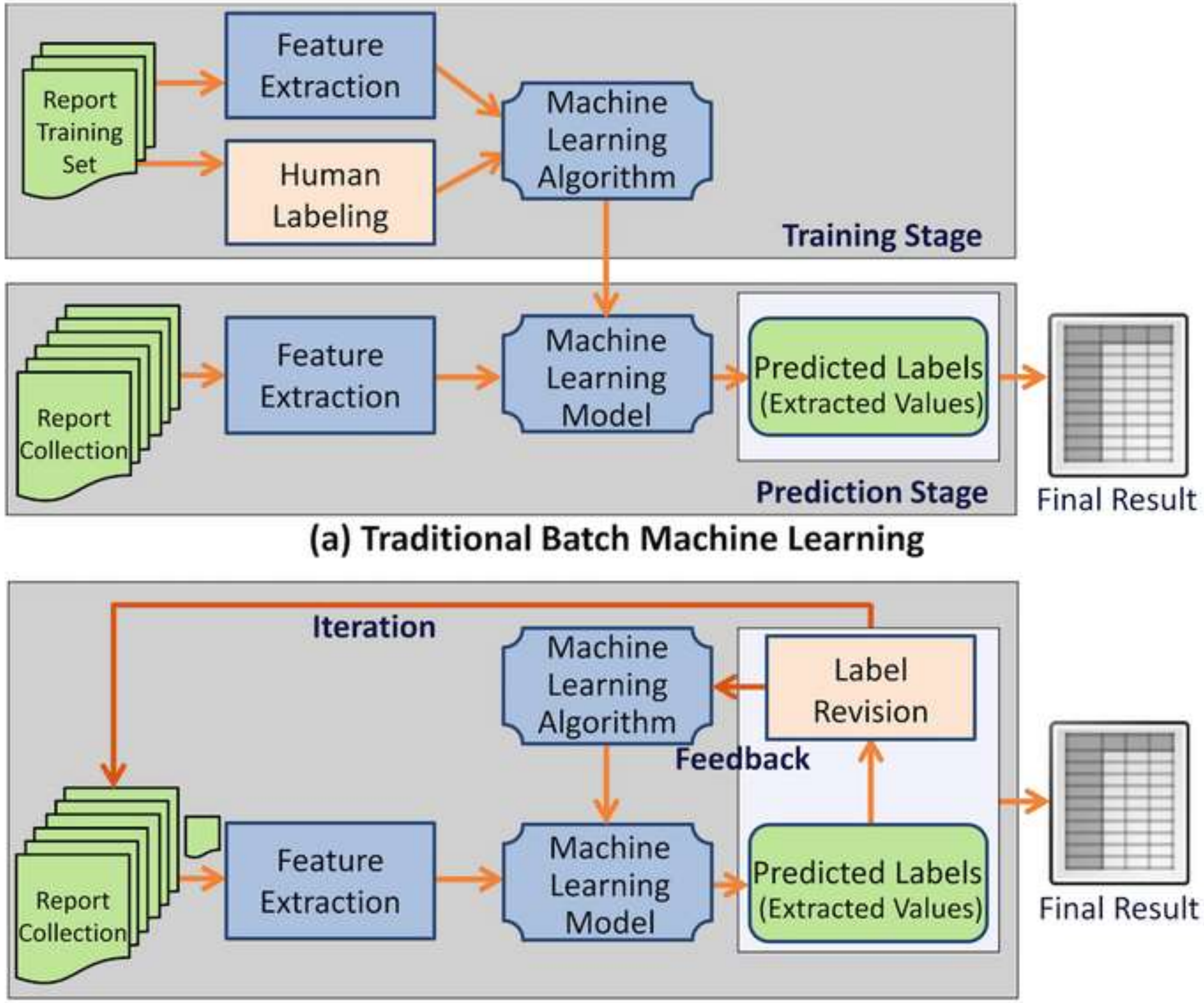
Then it does not wait for test data to learn

So it takes long time learning and less time classifying data

Categorizing Machine Learning Algorithms

Online: Learning
Batch: Learning

- Eager \
- Batch \



(a) Traditional Batch Machine Learning

(b) Online Machine Learning

Categorizing Machine Learning Algorithms

The trade-offs between parametric and non-parametric algorithms are in computational cost and accuracy.

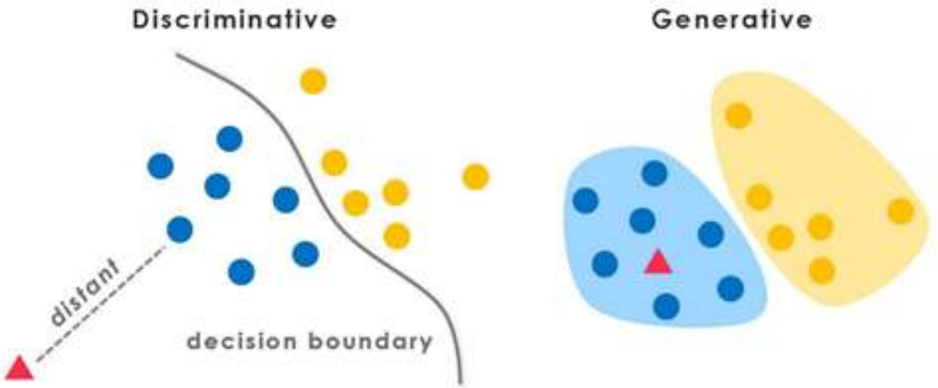
- Eager vs Lazy;
- Batch vs Online;
- Parametric vs Nonparametric;

A non-parametric algorithm uses a flexible number of parameters, and the number of parameters often grows as it learns from more data. A non-parametric algorithm is computationally slower:KNN

A parametric algorithm has a fixed number of parameters. A parametric algorithm is computationally faster, but makes stronger assumptions about the data: Linear Regression

Categorizing Machine Learning Algorithms

- Eager vs Lazy;
- Batch vs Online;
- Parametric vs Nonparametric;
- Discriminative vs Generative.



Goals in Analyzing data

Case 1:

Prediction. To be able to predict what the responses are going to be to future input variables

Information. To extract some information about how the algorithm is associating the response variables to the input variables.

Goals in Analyzing data

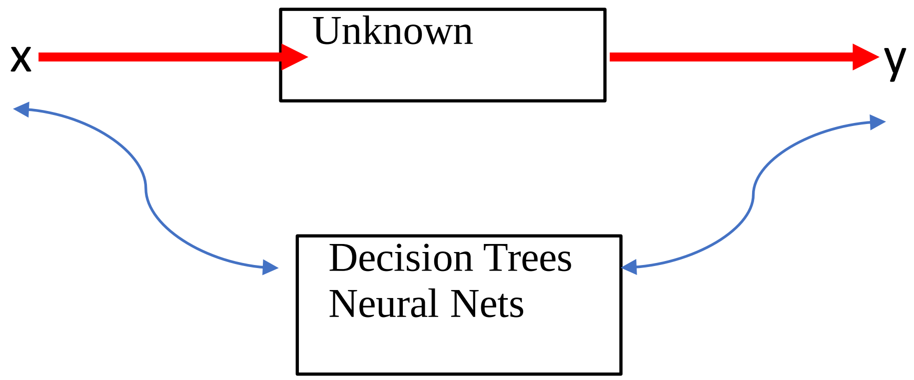
Case 2

The values of the parameters are estimated from the data and the model then used for information and/or prediction. Thus the black box is filled in like this:

Model validation. Yes–no using goodness-of-fit tests and residual examination.

Goals in Analyzing data

Case 3

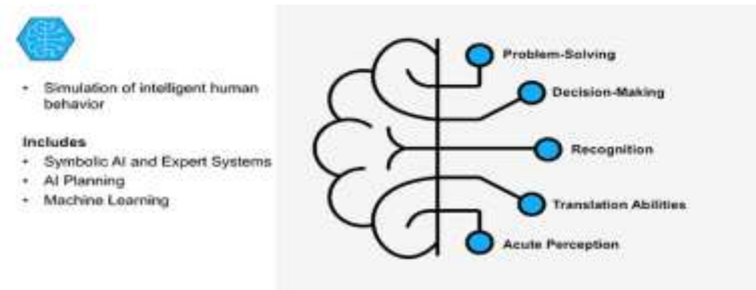


The analysis in this culture considers the inside of the box complex and unknown. Their approach is to

- find a function f→**x**)—an algorithm that operates on **x** to predict the responses **y**

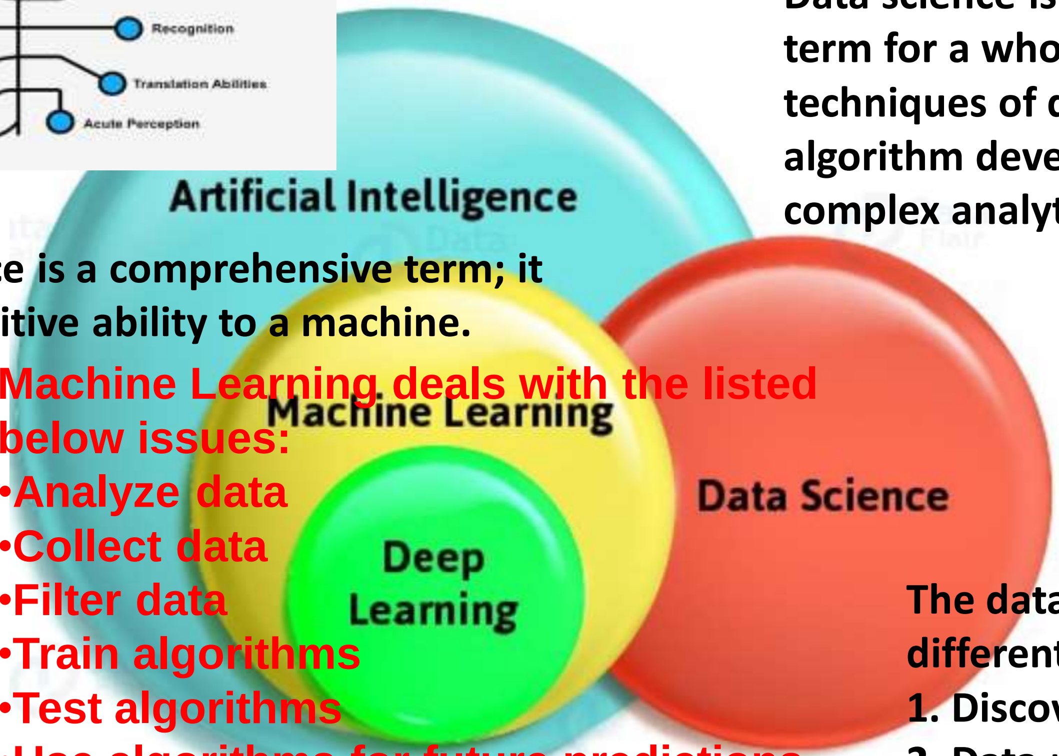
Model validation. Measured by predictive accuracy.

Machine Learning, AI, Deep Learning, and DATA SCIENCE



Data science is a multidisciplinary term for a whole set of tools and techniques of data inference and algorithm development to solve complex analytical problems.

Artificial Intelligence is a comprehensive term; it is conveying a cognitive ability to a machine.



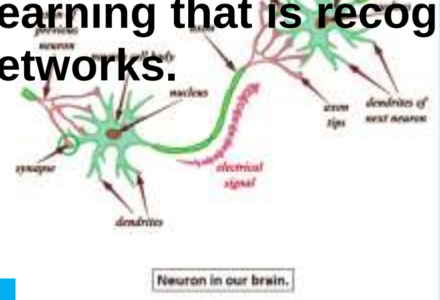
Machine Learning deals with the listed below issues:

- **Analyze data**
- **Collect data**
- **Filter data**
- **Train algorithms**
- **Test algorithms**
- **Use algorithms for future predictions**

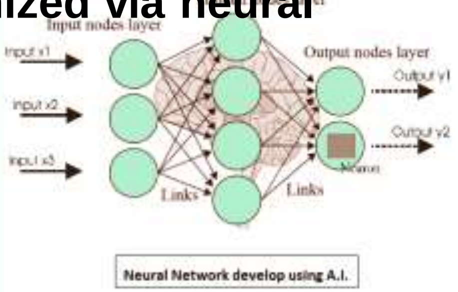
The data science life cycle has six different phases:

1. Discovery
2. Data preparation
3. Model planning
4. Model building
5. Communicating results
6. Operationalizing

Deep Learning is an approach to Machine Learning that is recognized via neural networks.



Neuron in our brain.



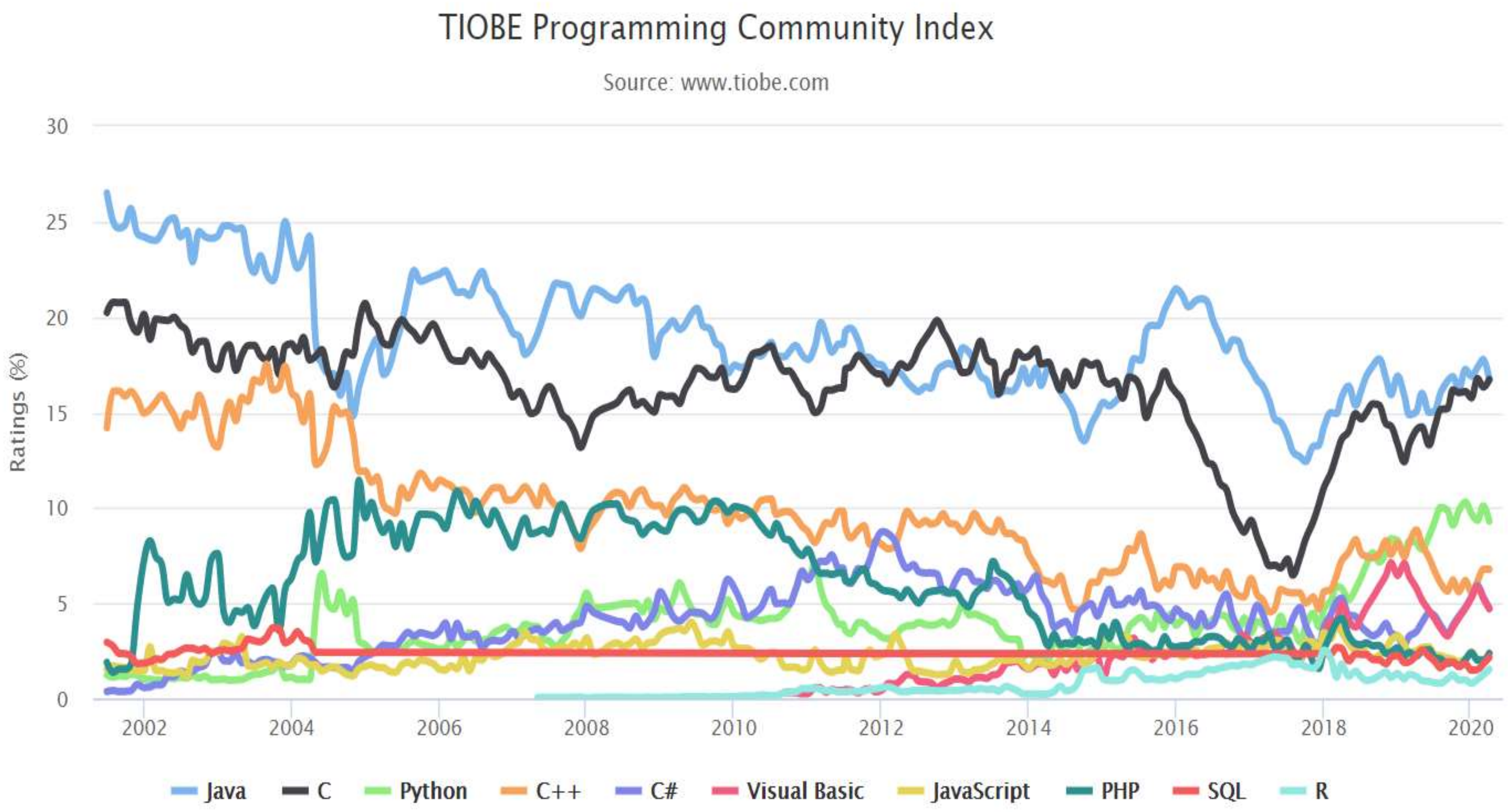
Neural Network develop using A.I.

	Machine Learning	Deep Learning
Data Dependencies	Superior performance on a small and medium dataset	Performs excellent on a big dataset
Hardware dependencies	Performs on a low-end machine	Preferable requires a machine with GPU. Deep Learning performs on a noteworthy matrix multiplication
Feature engineering	Carefully understand the features of how it represents the data	Required to understand the specific best functionality that represents the data
Execution time	From a few minutes to hours	It requires a time of up to 2-3 weeks.
Interpretability	Some algorithms are easy to interpret like, logistic and decision tree. Whereas some are almost impossible like, SVM and XGBoost	Difficult to impossible

Machine Learning vs Deep Learning

	Machine Learning	Deep Learning
Training dataset	Small	Large
Choose features	Yes	No
Number of algorithms	Many	Few
Training time	Short	Long

Why Python?



TIOBE Programming Community Index
Source: www.tiobe.com
Ratings (%)
30
25
20
15
10
5
0
2002
2004
2006
2008
2010
2012
2014
2016
2018
2020
Java
C
Python
C++
C#
Visual Basic
JavaScript
PHP
SQL
R

Python Libraries

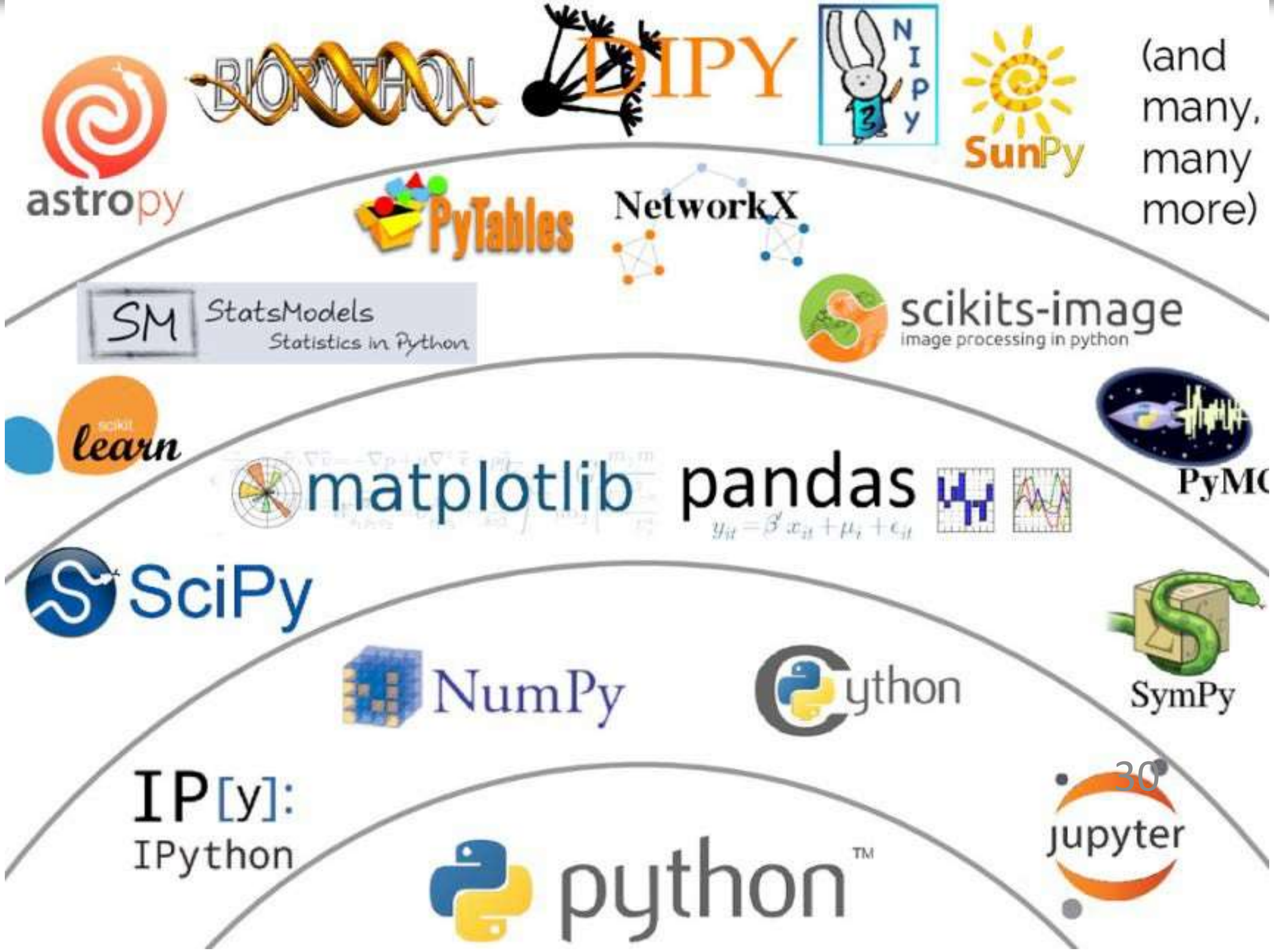


Image by Jake VanderPlas; Source: https://speakerdeck.com/jakevdp/the-state-of-the-stack-scipy-2015-keynote?slide=8)