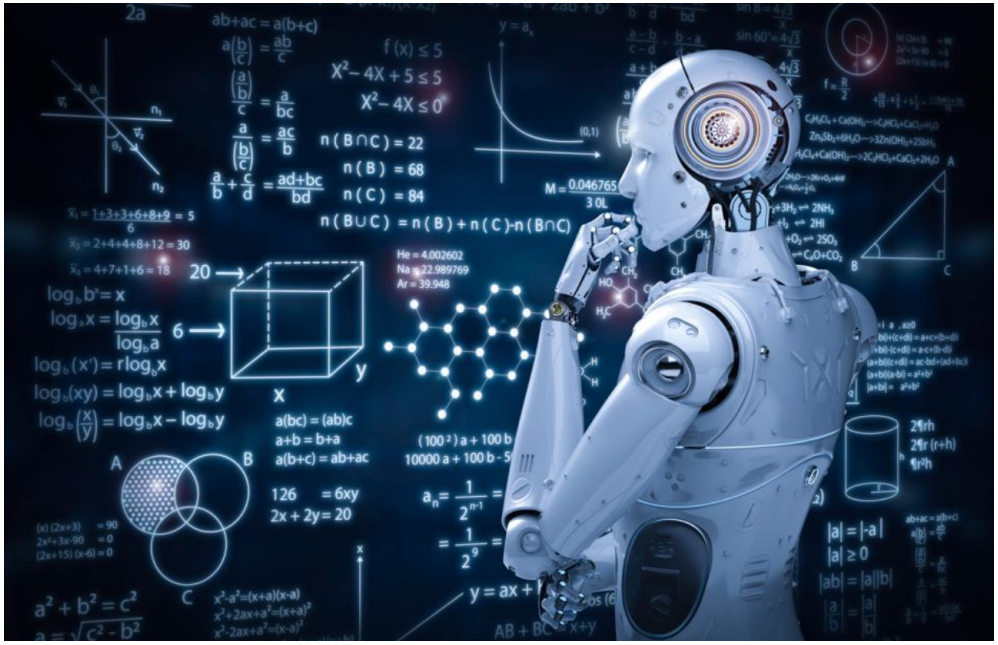
CS 4100: Introduction to AI

Wayne Snyder
Northeastern University

Lecture 14: Introduction to Machine Learning



What is Machine Learning?

"Learning is any process by which a system improves its performance from experience."

- Herbert Simon (A founder of AI)

"Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed."

-Arthur Samuel (Creator of first checker-playing program, 1959)

Machine Learning is the study of algorithms that perform

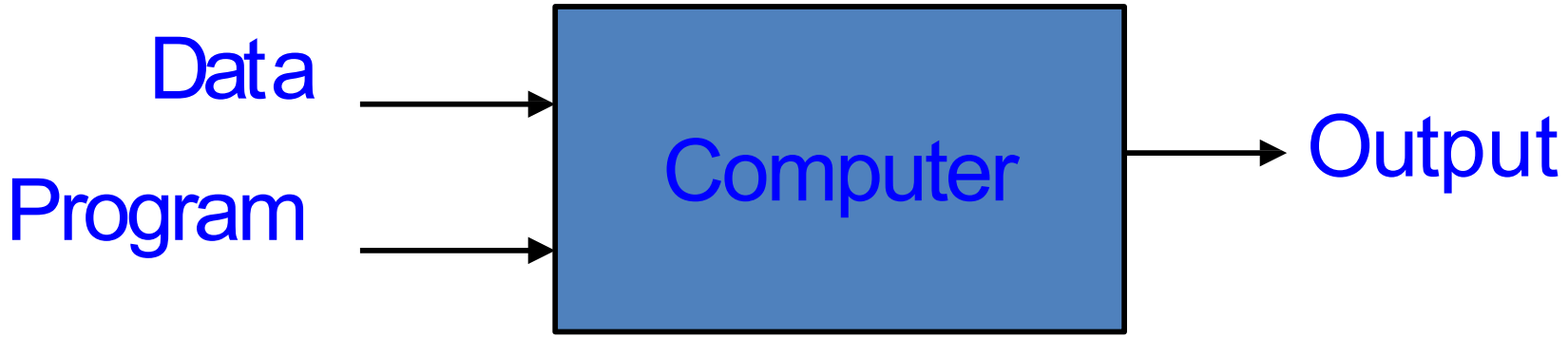
- Some task T (The problem/task to solve)
- After some experience E (Training)
- And improve in some performance metric P (Testing)

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

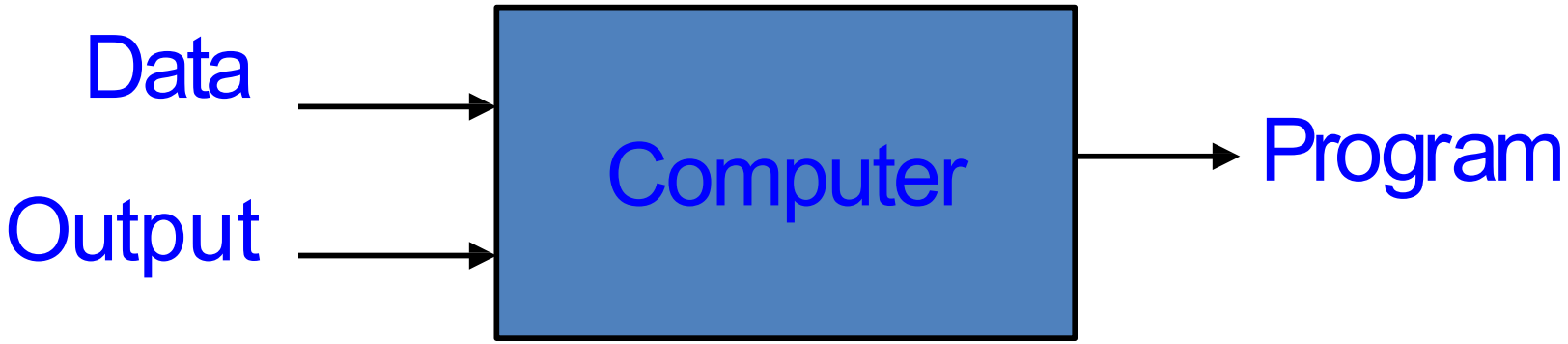
The ways that these three parameters are defined gives rise to the variety of different approaches to Machine Learning.

--Tom Mitchell (1998)

Traditional Programming



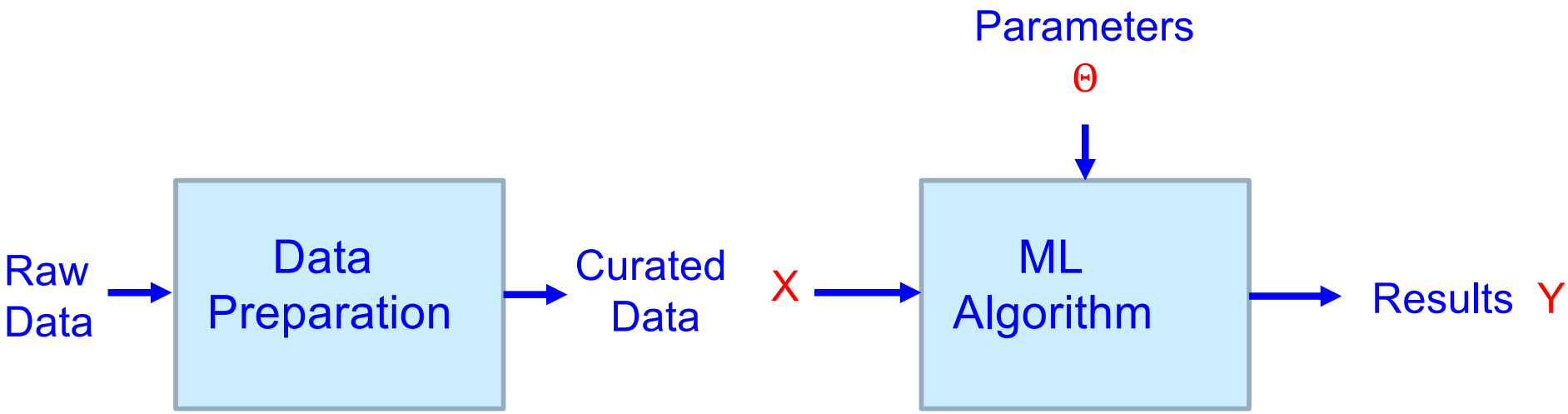
Machine Learning



Slide credit: Pedro Domingos

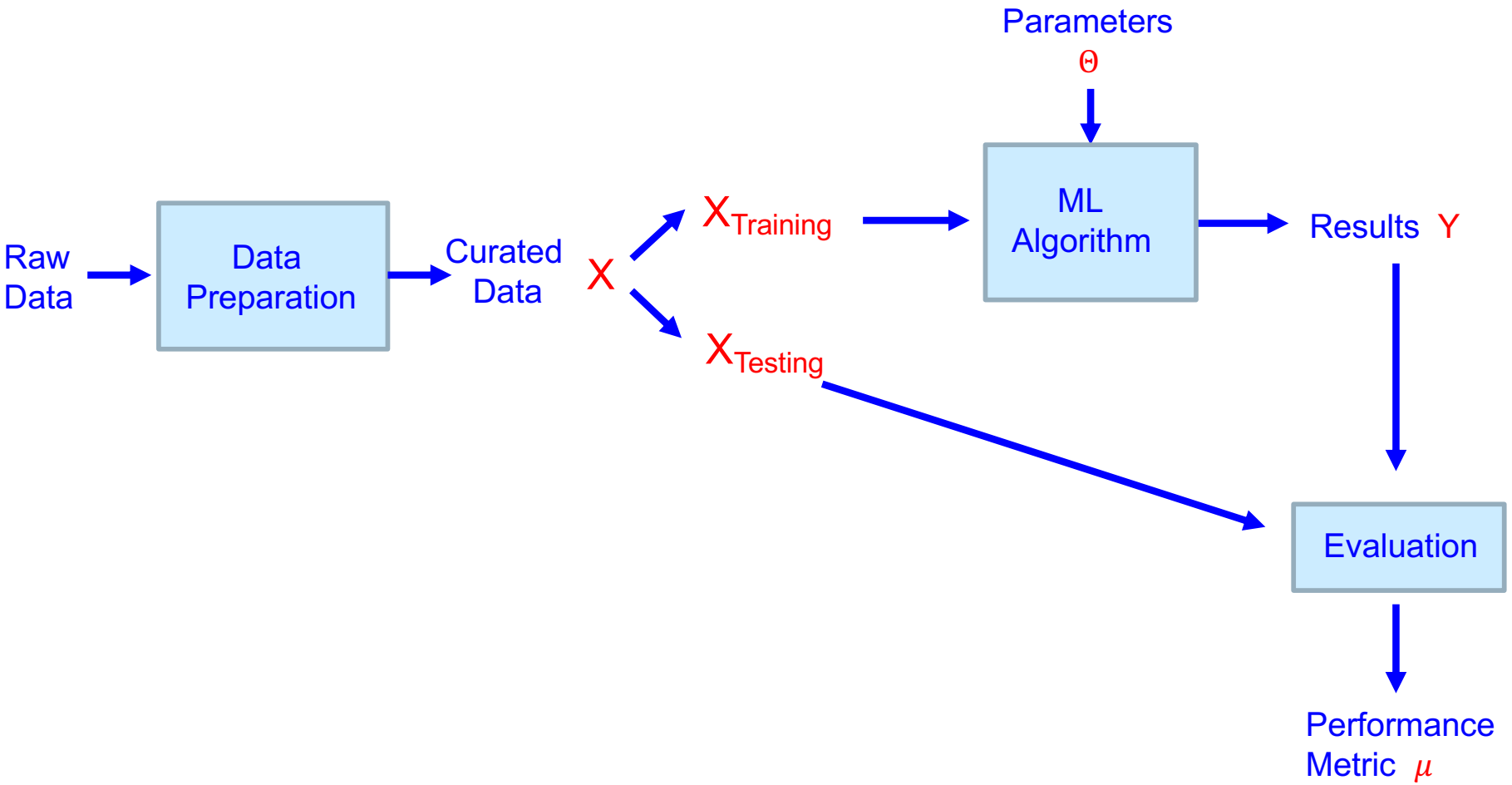
Introduction to Machine Learning

Unsupervised Machine Learning Workflow:



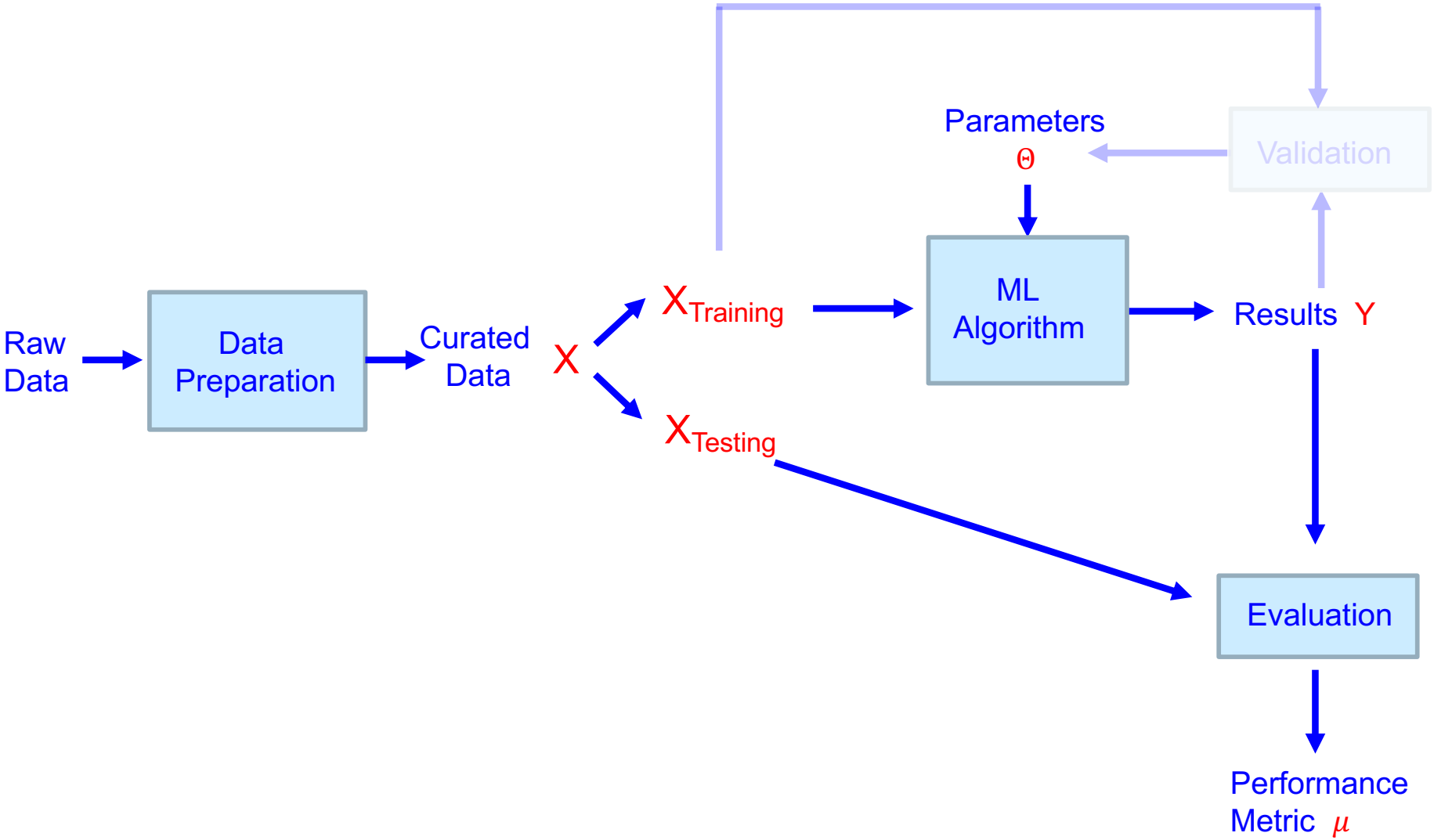
Introduction to Machine Learning

Supervised Machine Learning Workflow (Naive Version):



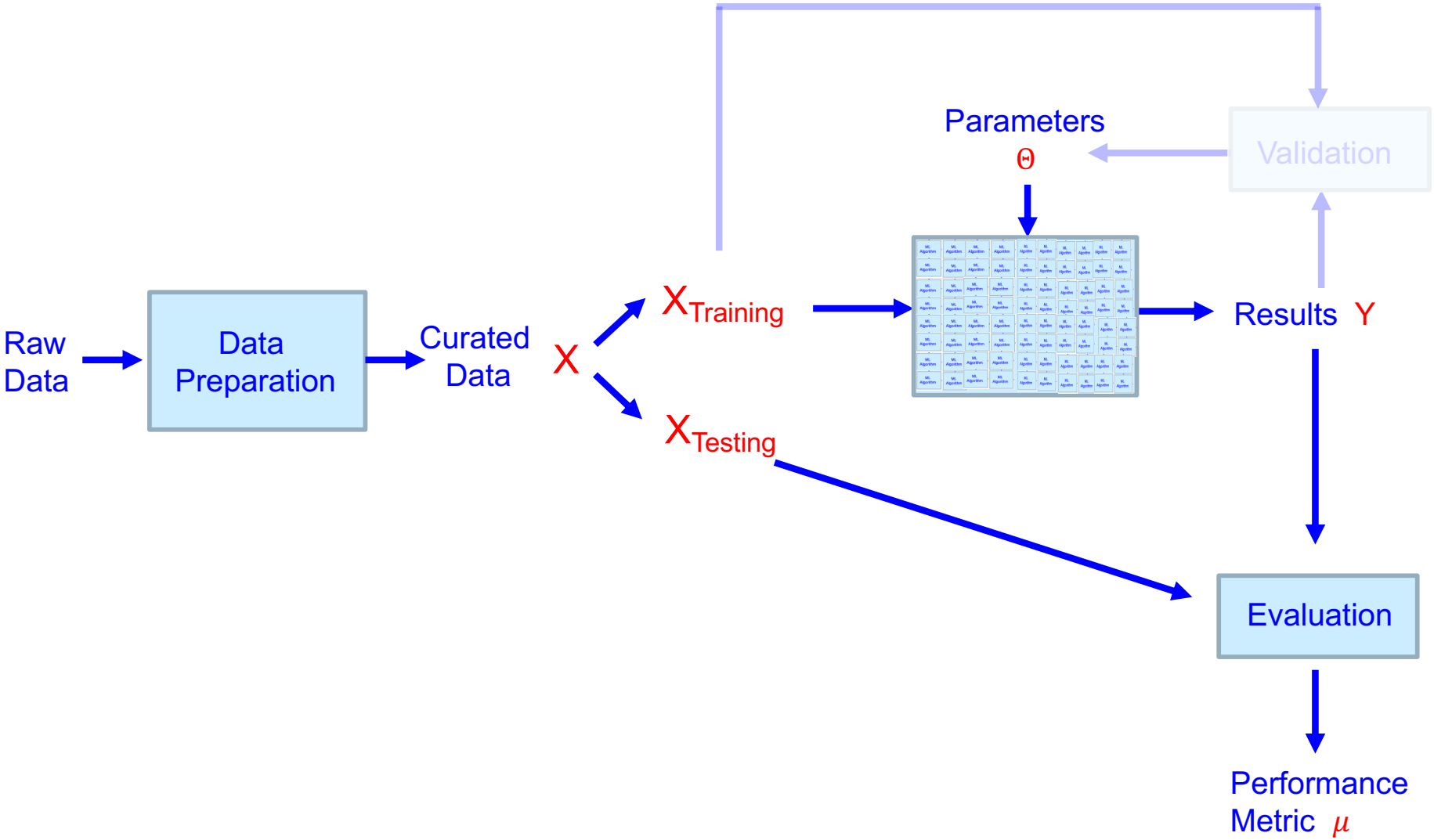
Introduction to Machine Learning

Supervised Machine Learning Workflow (Actual):



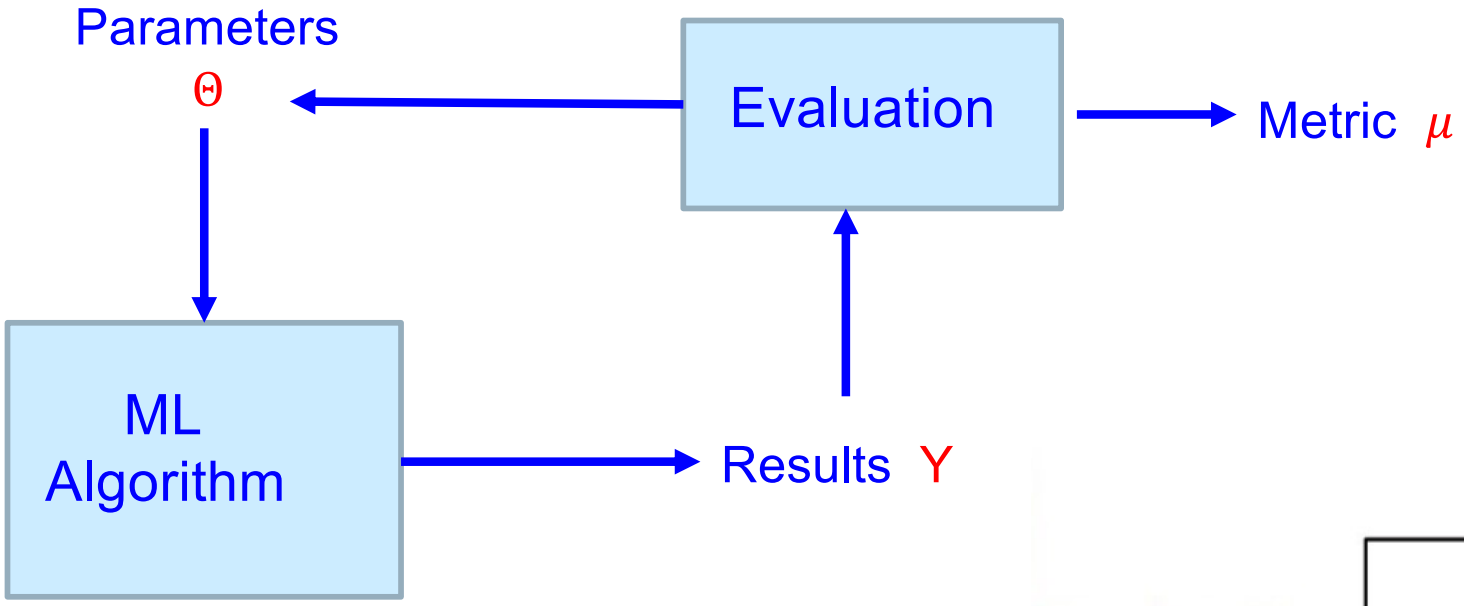
Introduction to Machine Learning

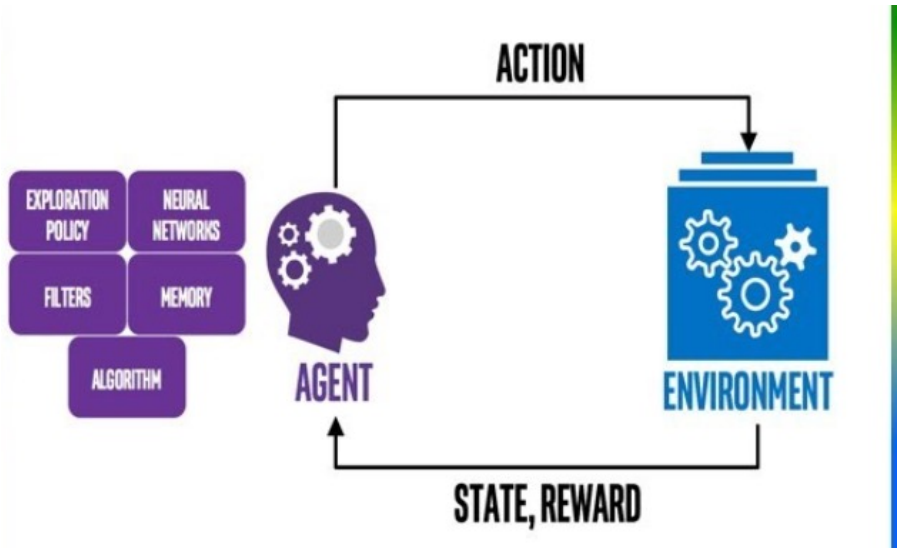
Deep Learning (Supervised ML with Neural Networks):



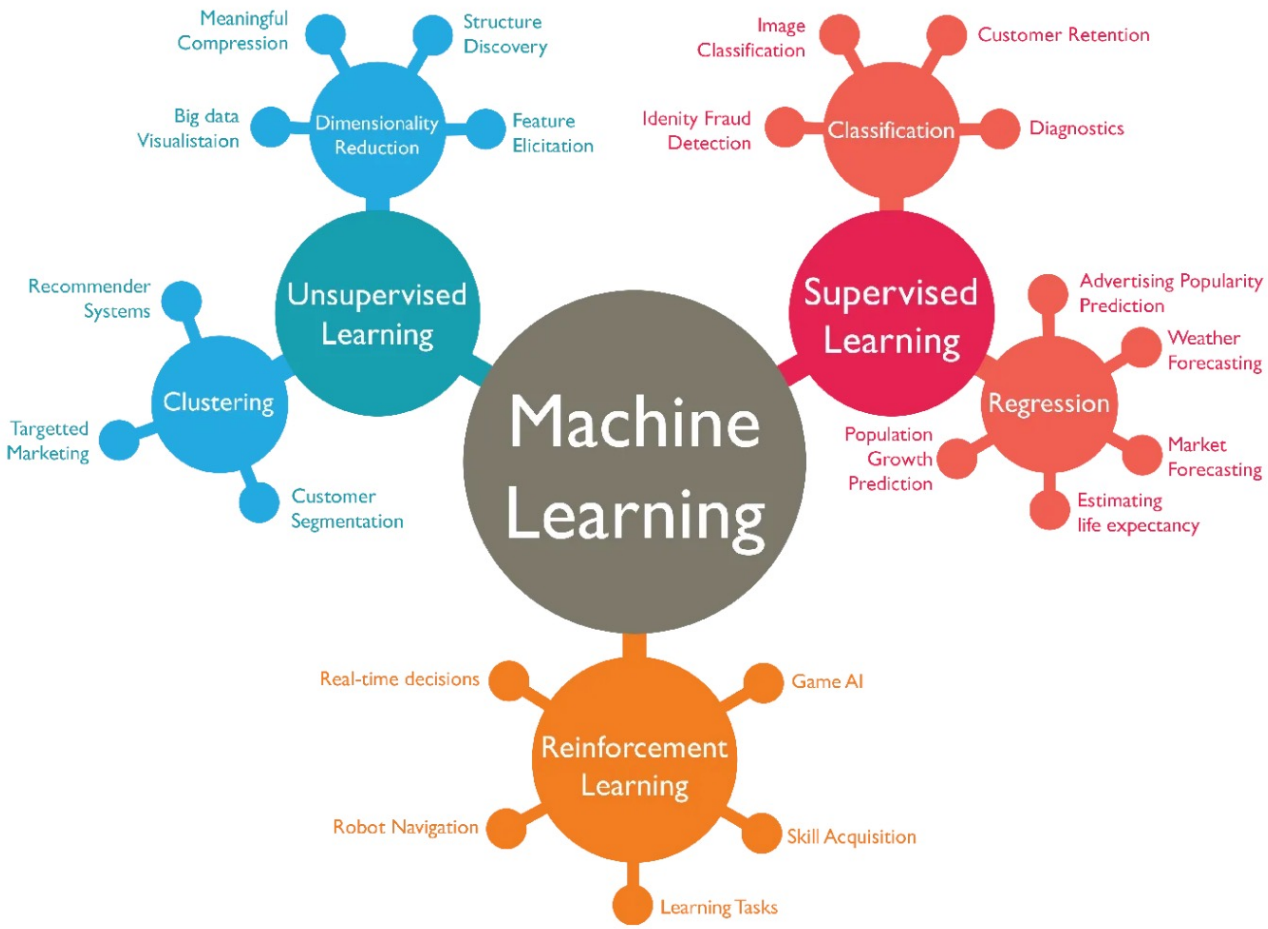
Introduction to Machine Learning

Reinforcement Machine Learning Workflow:





Introduction to Machine Learning



Introduction to Machine Learning

The plan for the rest of the semester....

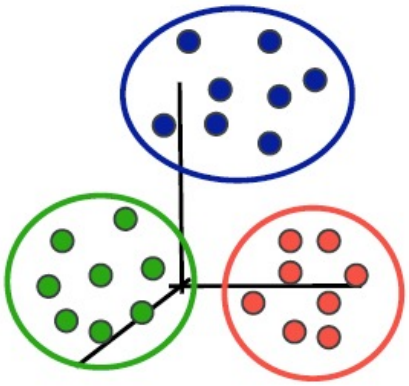
14	M 2/27	Introduction to Machine Learning; Unsupervised Learning: Clustering with K-Means
15	W 3/1	K-Means concluded; Hierarchical Clustering; Dimensionality Reduction with PCA
	Week 3/6 - 3/10	SPRING BREAK
16	M 3/13	Supervised Learning: Regression and Logistic Regression
17	W 3/15	Supervised Learning: Decision Trees and other "non-deep" ML
18	M 3/20	Introduction to Artificial Neural Networks; Perceptrons; Feedforward Networks
19	W 3/22	Feedforward networks; Backpropagation; training and evaluation workflow
20	M 3/27	Classification with FF Networks
21	W 3/29	Convolutional networks for image classification
22	M 4/3	Recurrent Networks: RNN, GRU, LSTM
23	W 4/5	Sequence to sequence models: Attention, BERT, GPT, ...
24	M 4/10	Reinforcement Learning
25	W 4/12	Reinforcement Learning
26	M 4/17	Patriots Day (Marathon Monday): No class!
27	W 4/19	Last Class: Conscious AI?
	4/21 - 28	Final Exam Period

Unsupervised Learning: Clustering

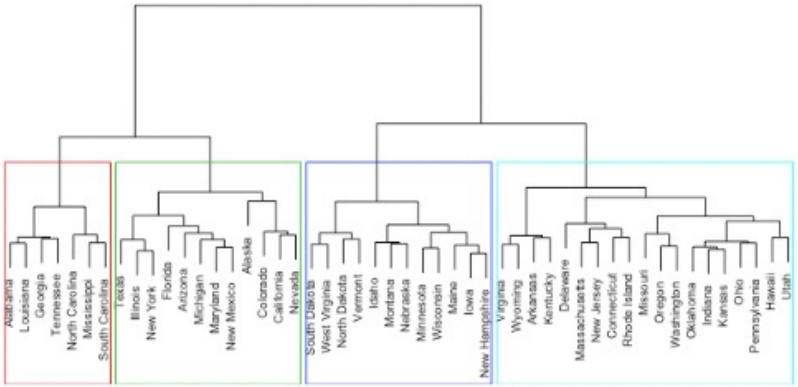
What is Clustering?

There are two basic types of clustering:

Partitioning



Hierarchical



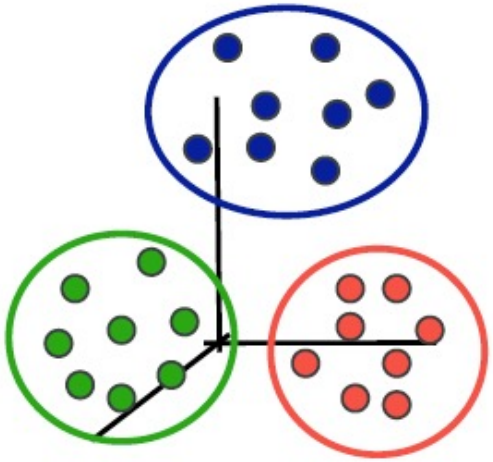
For now we will only consider partitioning algorithms: each objects belongs to exactly one cluster.

Unsupervised Learning: Clustering

What is Clustering?

A grouping of data objects such that the objects within a group are similar (or near) to one another and dissimilar (or far) from the objects in other groups:

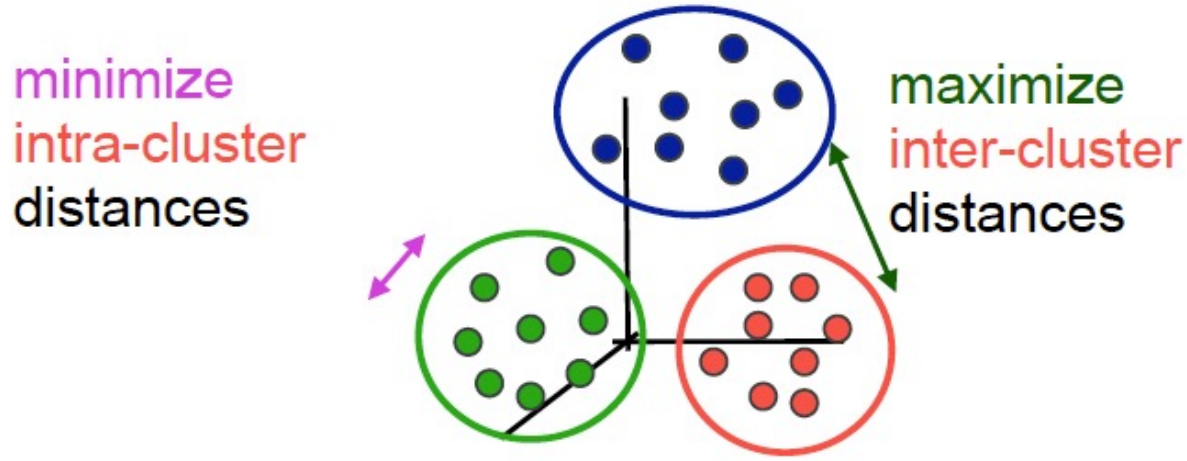
Cluster = group



Unsupervised Learning: Clustering

A grouping of data objects such that the objects within a group are similar (or near)to one another and dissimilar (or far) from the objects in other groups:

Cluster = group



Unsupervised Learning: Clustering

The Clustering Problem

Given a collection of data objects, find a grouping such that

- Similar objects are in the same cluster
- Dissimilar objects are in different cluster

Why is this important?

- A stand-alone tool for visualizing and understanding the data
- A preprocessing step for other algorithms
  - Creating group labels for supervised learning
  - Indexing or compression often relies on clustering

Unsupervised Learning: Clustering

The Clustering Problem

Given a collection of data objects, find a grouping such that

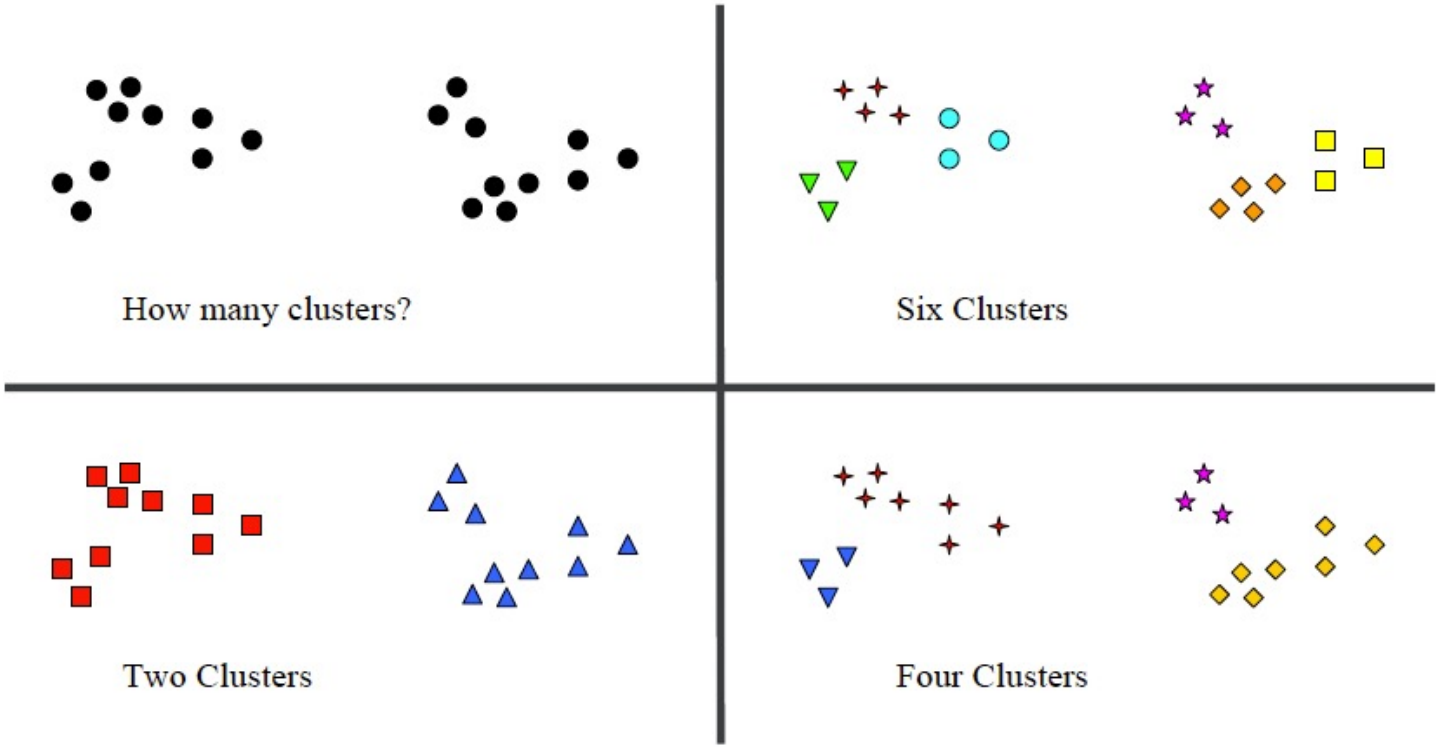
- Similar objects are in the same cluster
- Dissimilar objects are in different cluster

Basic Questions:

- What does similar mean?
- What are the most efficient algorithms?
- How do we evaluate the quality of the resulting partition?

Unsupervised Learning: Clustering

The Big Problem: The notion of a cluster is ambiguous.



Unsupervised Learning: Clustering

K-Means Clustering

BIG Assumption: Assume in advance you want exactly k clusters.

The K-Means Algorithm:

Given k and a set X = { x_1, x_2, ..., x_n } of points in R^d (d = number of dimensions), find

- k points {c_1, ..., c_k} (called centers, means, or centroids) and
- a partition of X into k clusters {X_1, ..., X_k} by assigning each point x_i to its nearest cluster center,

such that the cost

$$\sum_{j=1}^{k} \sum_{x \in X_j} ||x - c_j||_2^2$$

is minimized.

L2 norm: square of distance between points x and c_j.

Unsupervised Learning: Clustering

The K-Means Problem

For K = 1 and K = n the solution is trivial (why?)

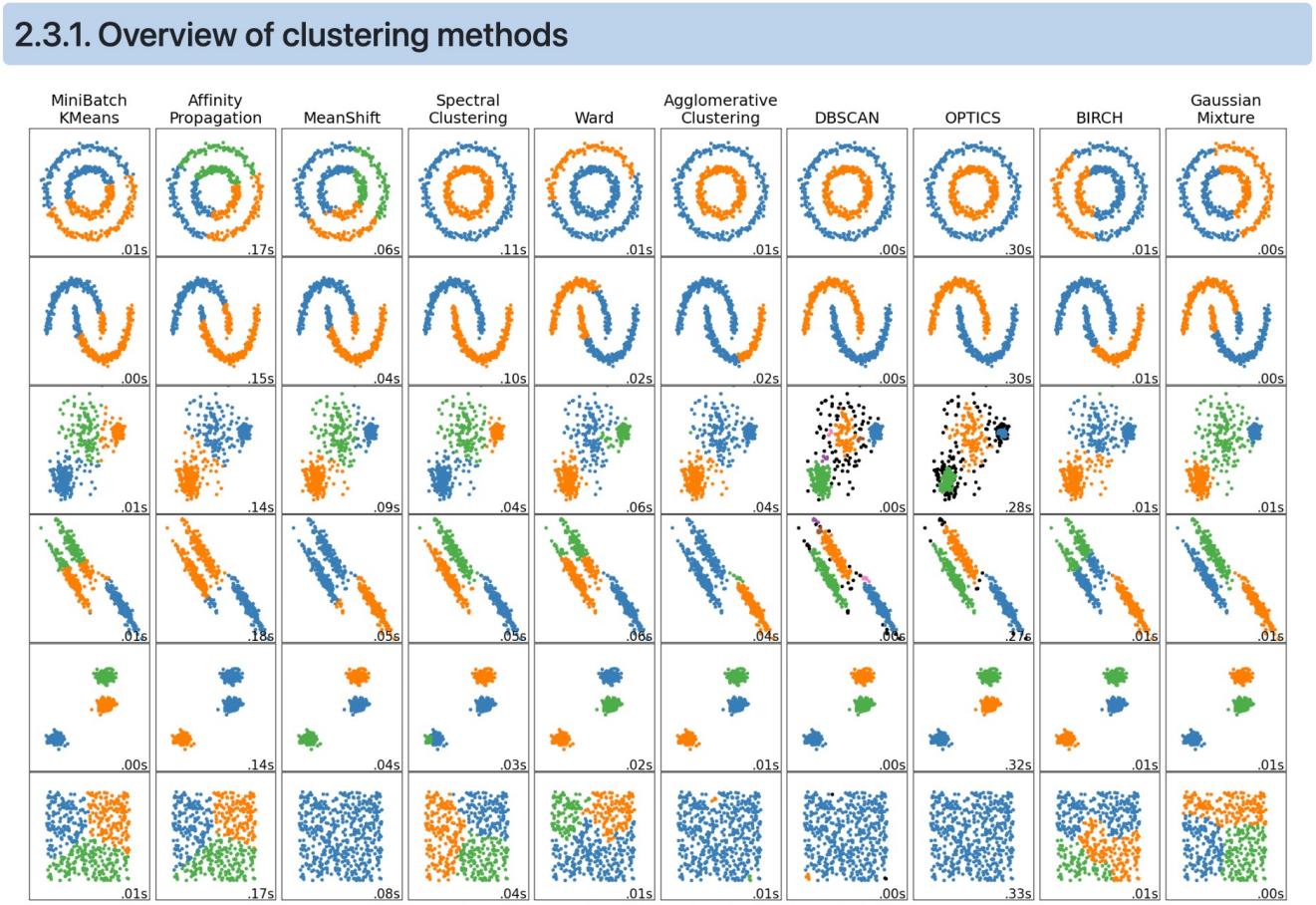
For other cases, it is NP-hard (probably exponential) for d > 2.

But in practice, iterative greedy algorithms work quite well!

Unsupervised Learning: Clustering

The K-Means Problem

There are many flavors of K-Means! We will only look at the most basic.



A comparison of the clustering algorithms in scikit-learn

Unsupervised Learning: Clustering

Lloyd's Algorithm for K-Means

Repeat until some termination criterion is met:

1. Randomly* choose the k centroids { c_1, ..., c_k };
2. For each $1 \leq j \leq k$ set the cluster X_j to be the set of points in X which are closest to the center c_j;
3. For each j, update the value of c_j to be the mean of the vectors in X_j.

* NOTE: This is a Hill-Climbing (search) algorithm, where the cost is the squared intra-cluster distances; it often converges quickly, but the choice of the initial set of centroids is critical, and essentially all of the refinements to this algorithm have to do with how this initialization step is done.

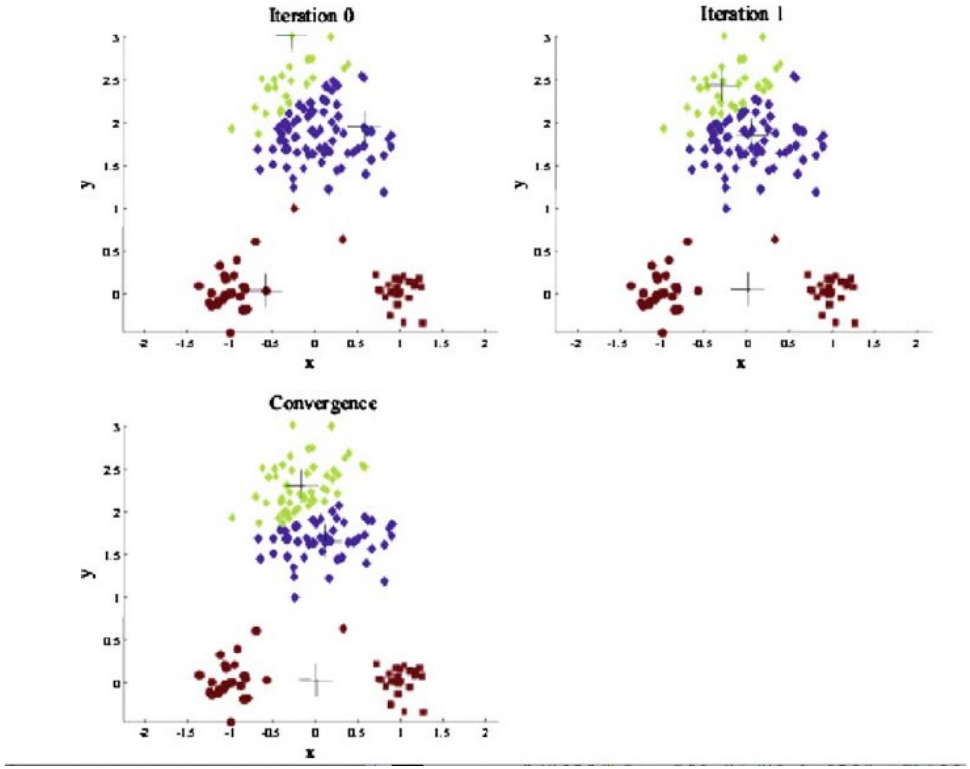
Unsupervised Learning: Clustering

Lloyd's Algorithm for K-Means

Let's look at some code to see what happens.....

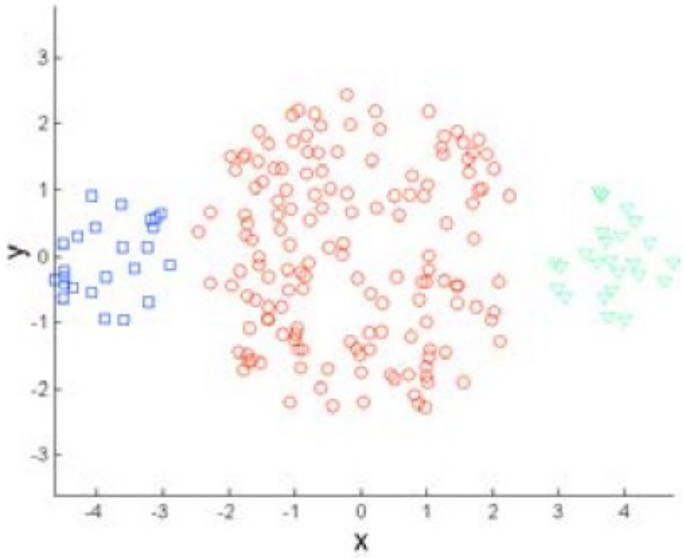
Unsupervised Learning: Clustering

Effect of a Bad Initialization

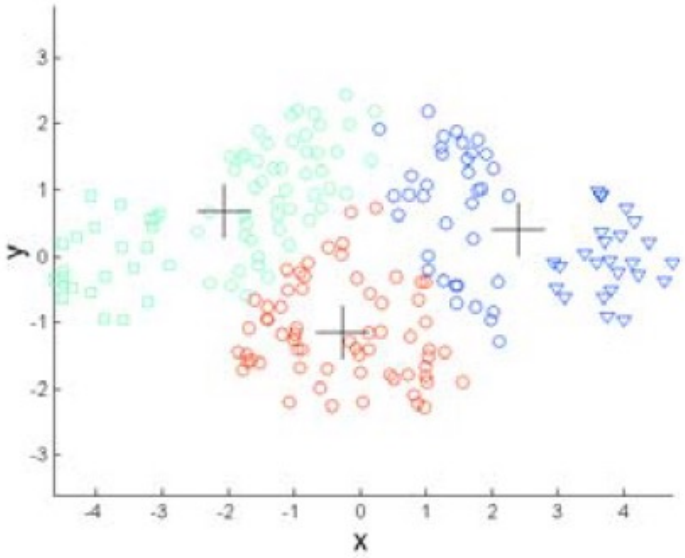


Unsupervised Learning: Clustering

Limitations of K-means: Clusters of different sizes



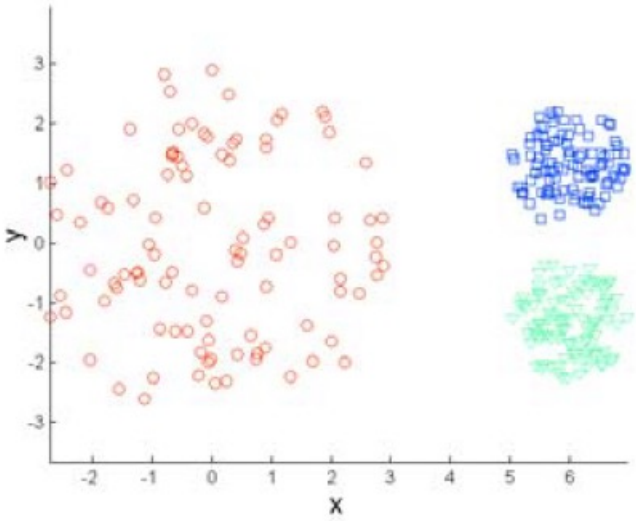
Original Points



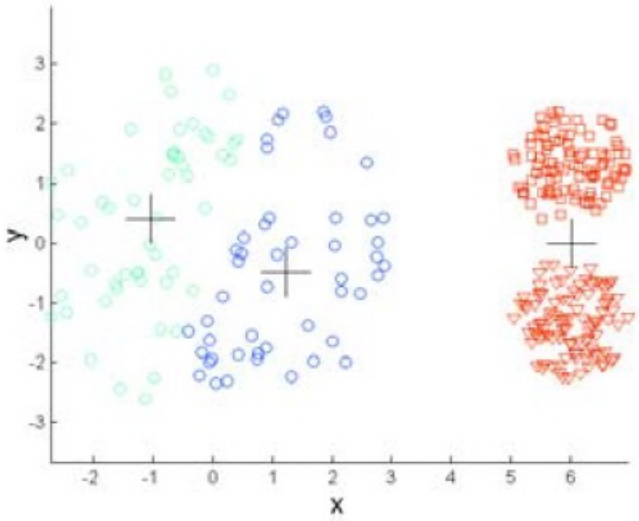
K-means (3 Clusters)

Unsupervised Learning: Clustering

Limitations of K-Means: Different Cluster Densities



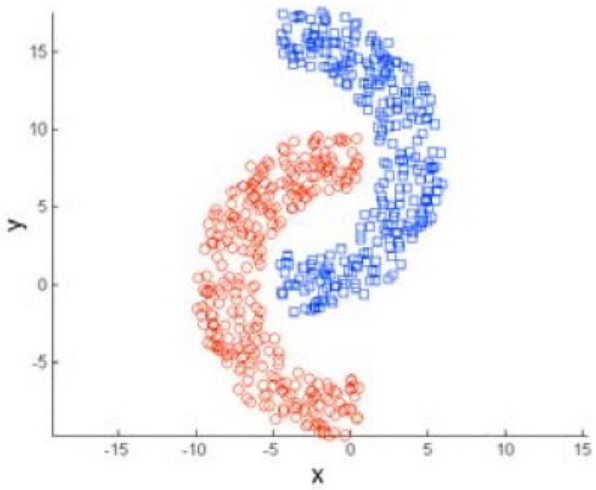
Original Points



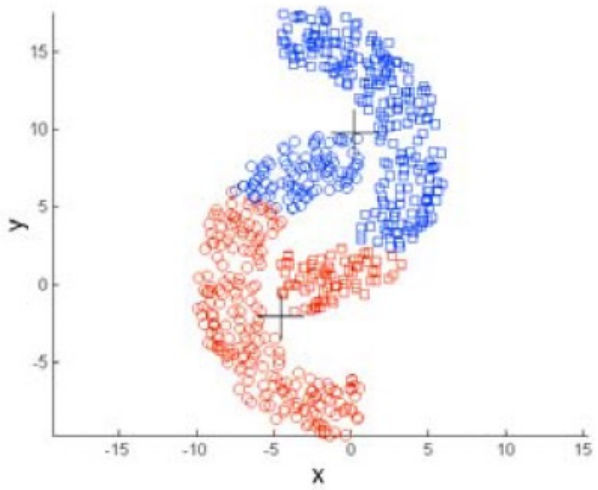
K-means (3 Clusters)

Unsupervised Learning: Clustering

Limitations of K-Means: non-spherical clusters



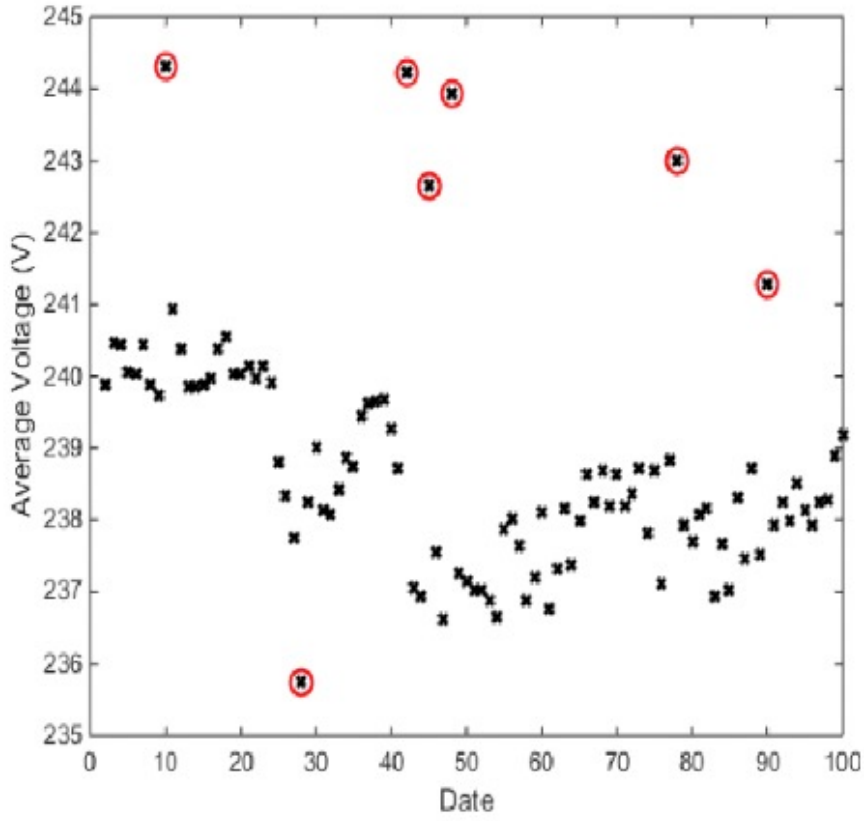
Original Points



K-means (2 Clusters)

Unsupervised Learning: Clustering

Limitations of K-Means: Outliers are a problem!



Unsupervised Learning: Clustering

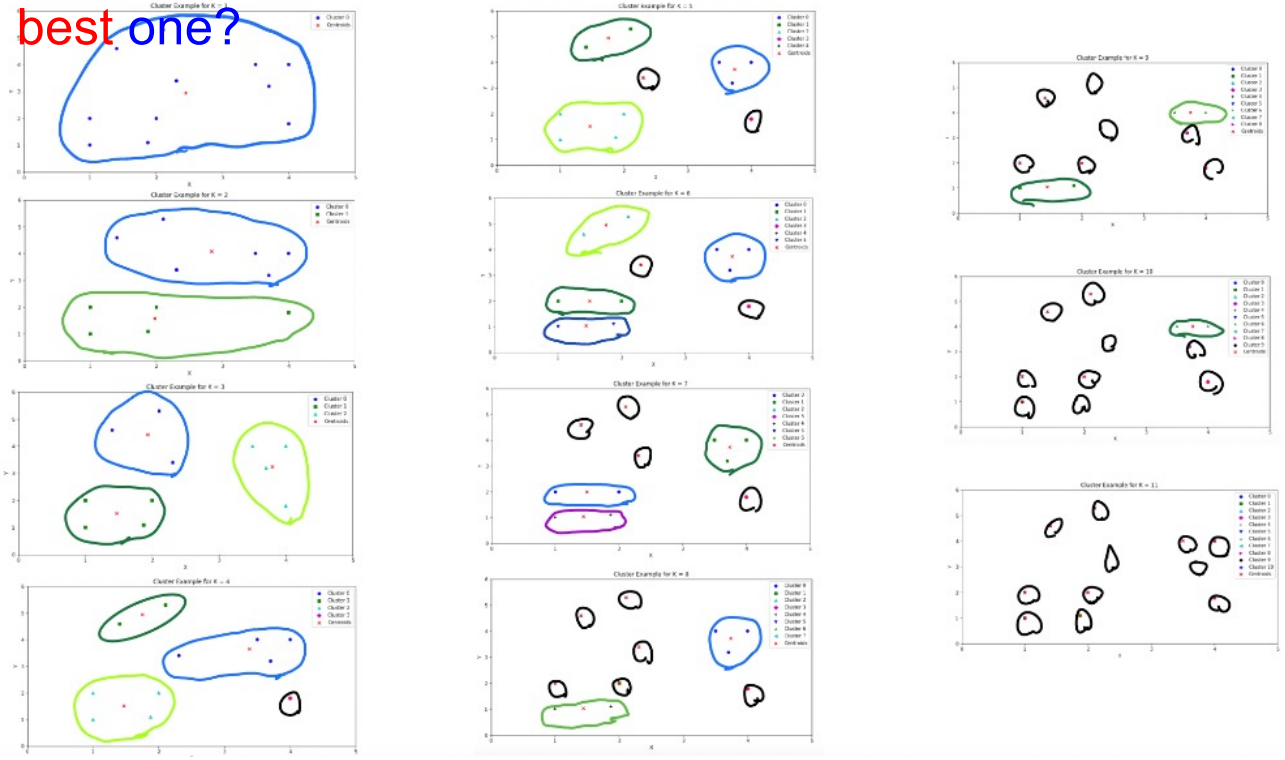
Improvements based on Initialization

- Random Initialization
- Repeat random initialization multiple times and take the best solution
- Pick random points which are distant from each other
  - Basis of K-Means++ algorithm
  - There are provable guarantees about quality of solutions.

Unsupervised Learning: Clustering

Evaluating K-Means: What should K be????

The main problem is that the cost (sum of squared intra-cluster distances) decreases as number of clusters gets smaller! Which is the best one?



Unsupervised Learning: Clustering

How to choose the right K??

Well... it depends.... but a naive method is to look at the graph of K vs cost and pick an appropriate midpoint between extremes, the so-called “elbow” of the curve.