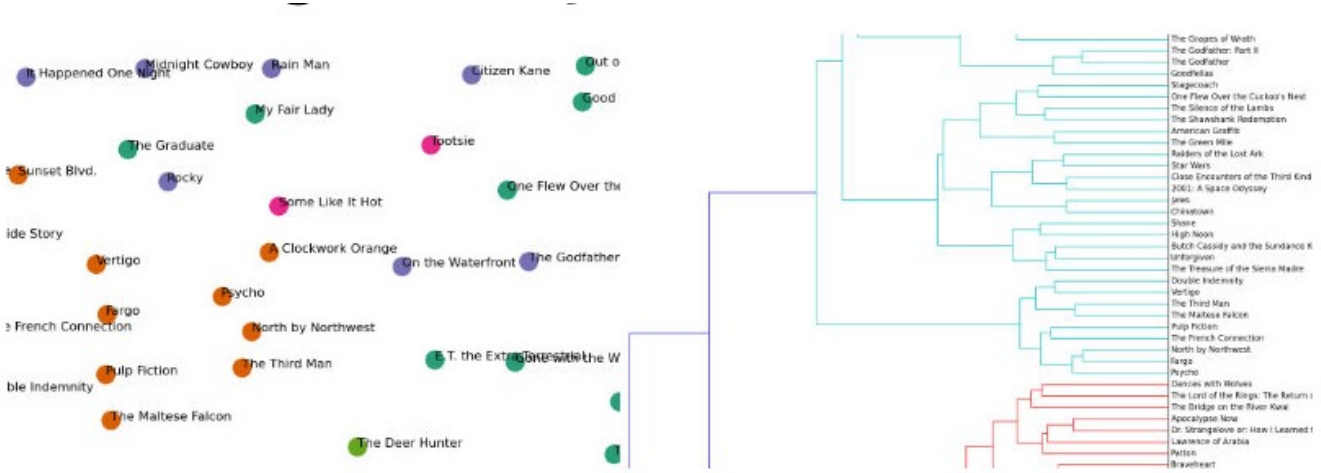
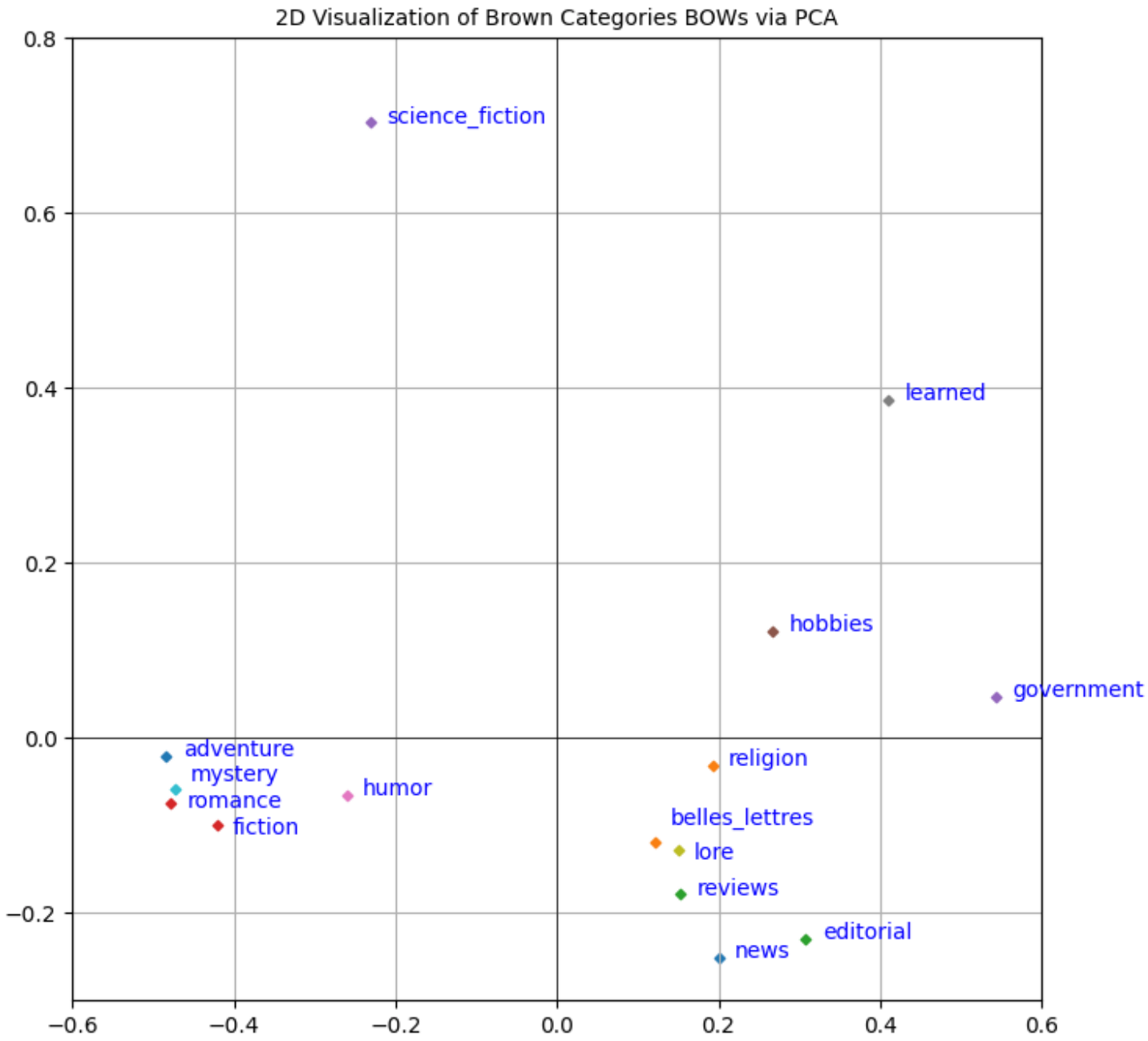
CS 505: Introduction to Natural Language Processing

Wayne Snyder
Boston University

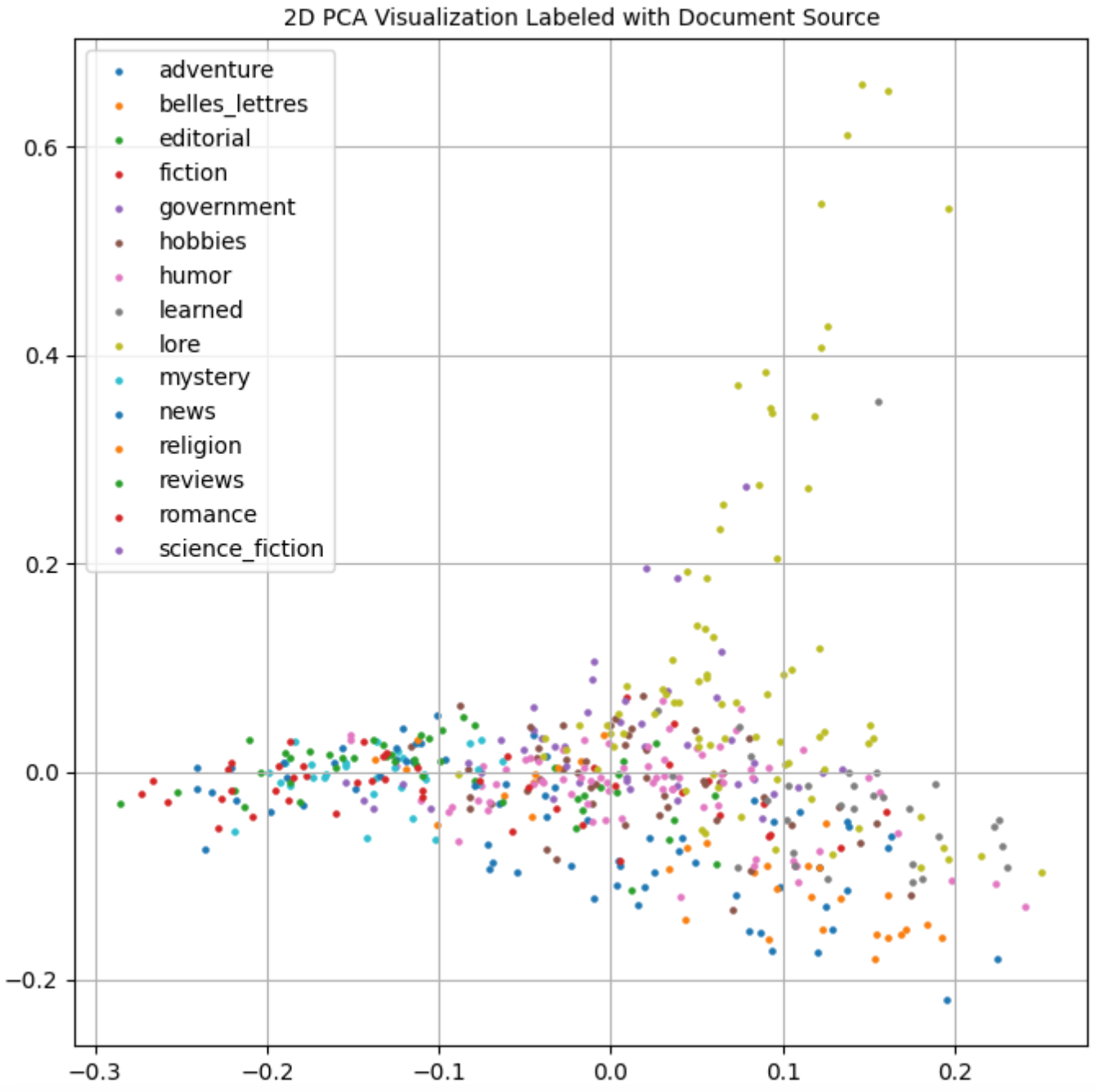
Lecture 7: Introduction to Machine Learning; Unsupervised ML; Clustering with K-Means, Hierarchical Clustering



Addendum to PCA from last time



Addendum to PCA from last time



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)

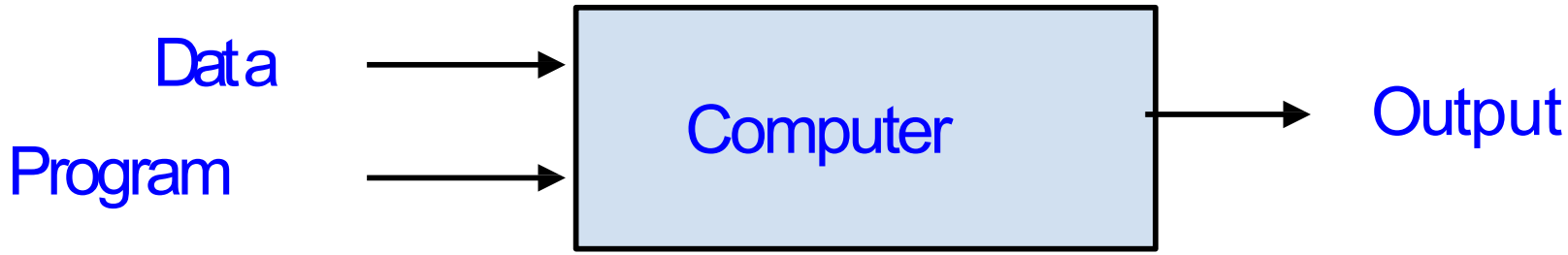
A well-defined learning task is given by <P, T, E>.

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

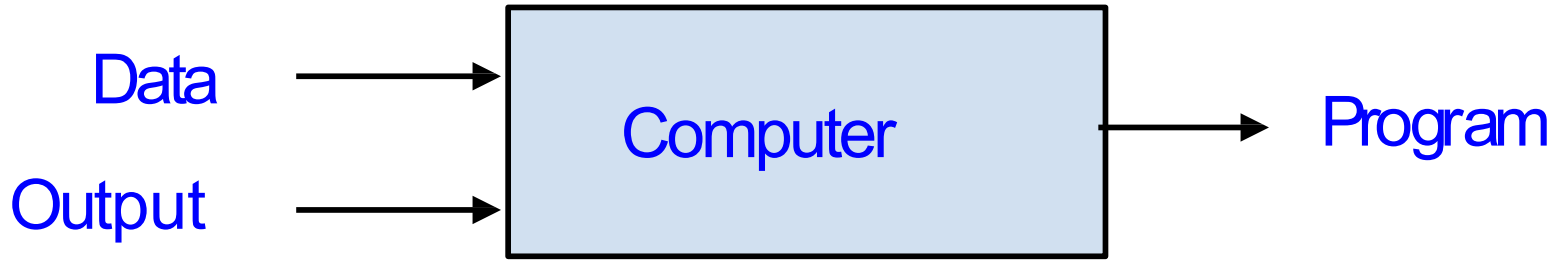
--Tom Mitchell (1998)

What is Machine Learning?

Traditional Programming



Machine Learning

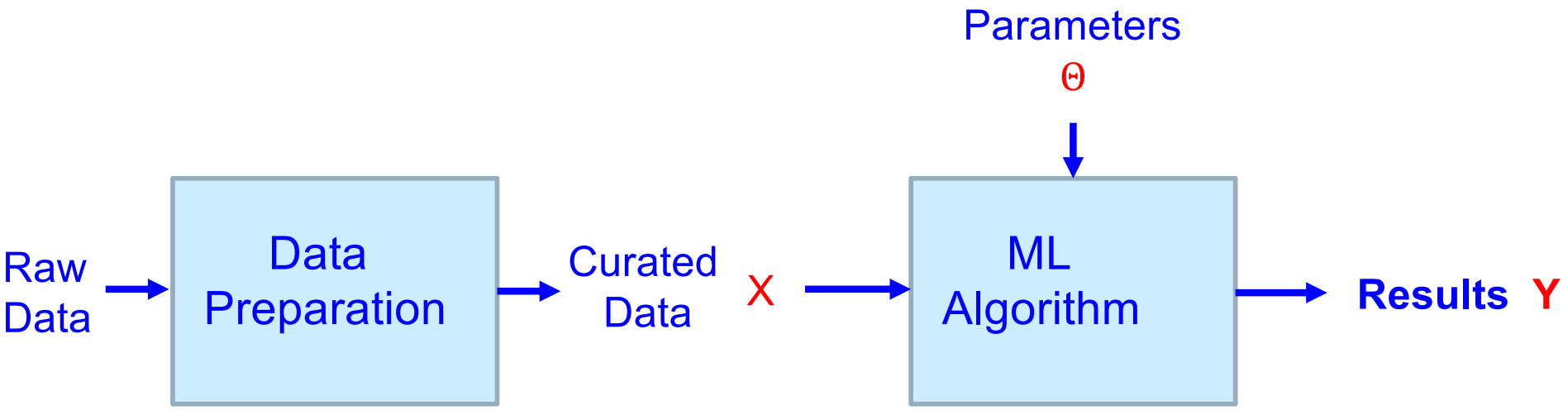


Slide credit: Pedro Domingos

Introduction to Machine Learning

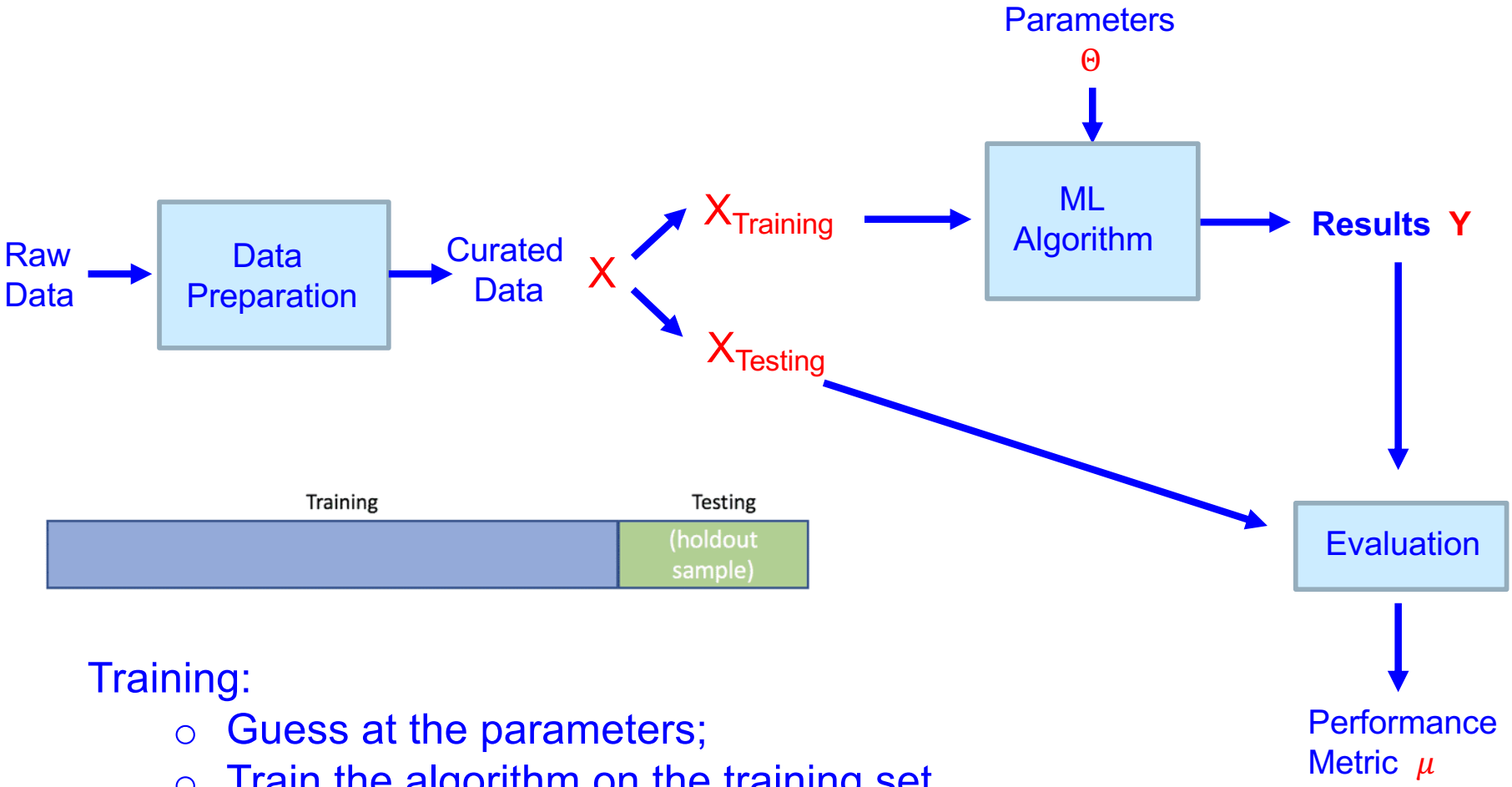
There are several flavors of Machine Learning….

Unsupervised Machine Learning Workflow:



Introduction to Machine Learning

Supervised Machine Learning Workflow (Naive Version):

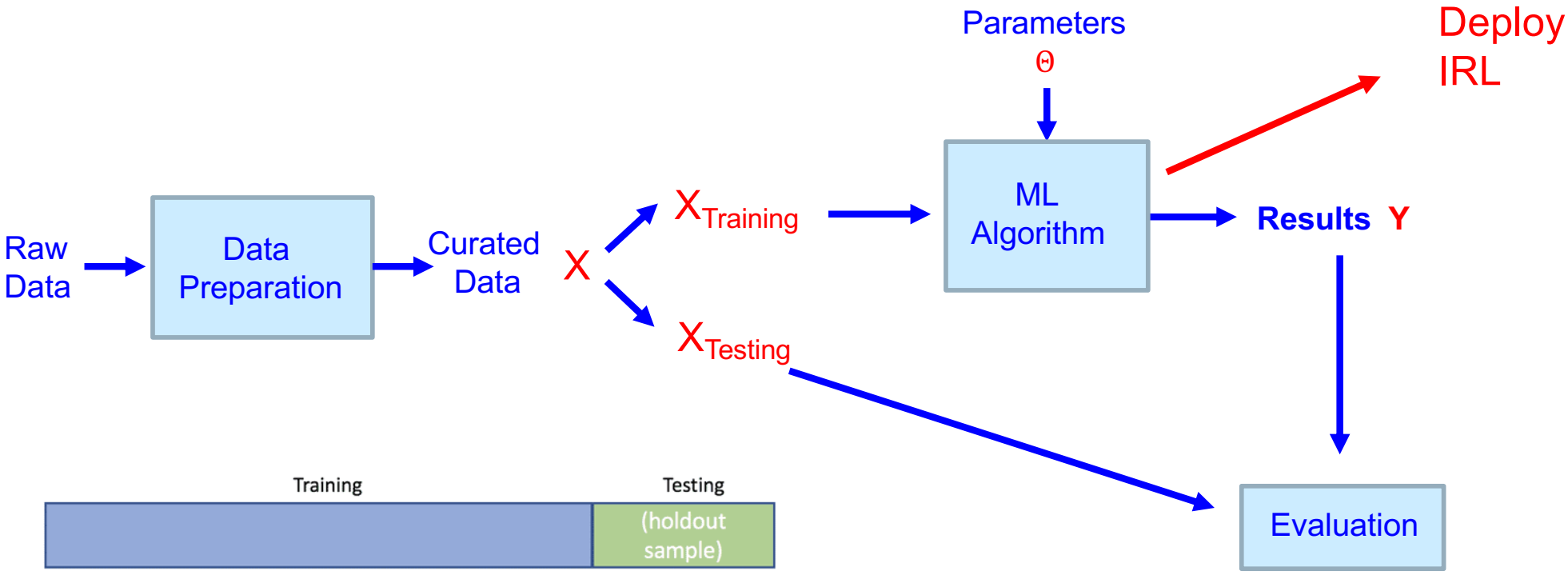


Training:

- Guess at the parameters;
- Train the algorithm on the training set
- Evaluate using the testing set
- (Adjust parameters to get better evaluation)

Introduction to Machine Learning

Supervised Machine Learning Workflow (Naive Version):

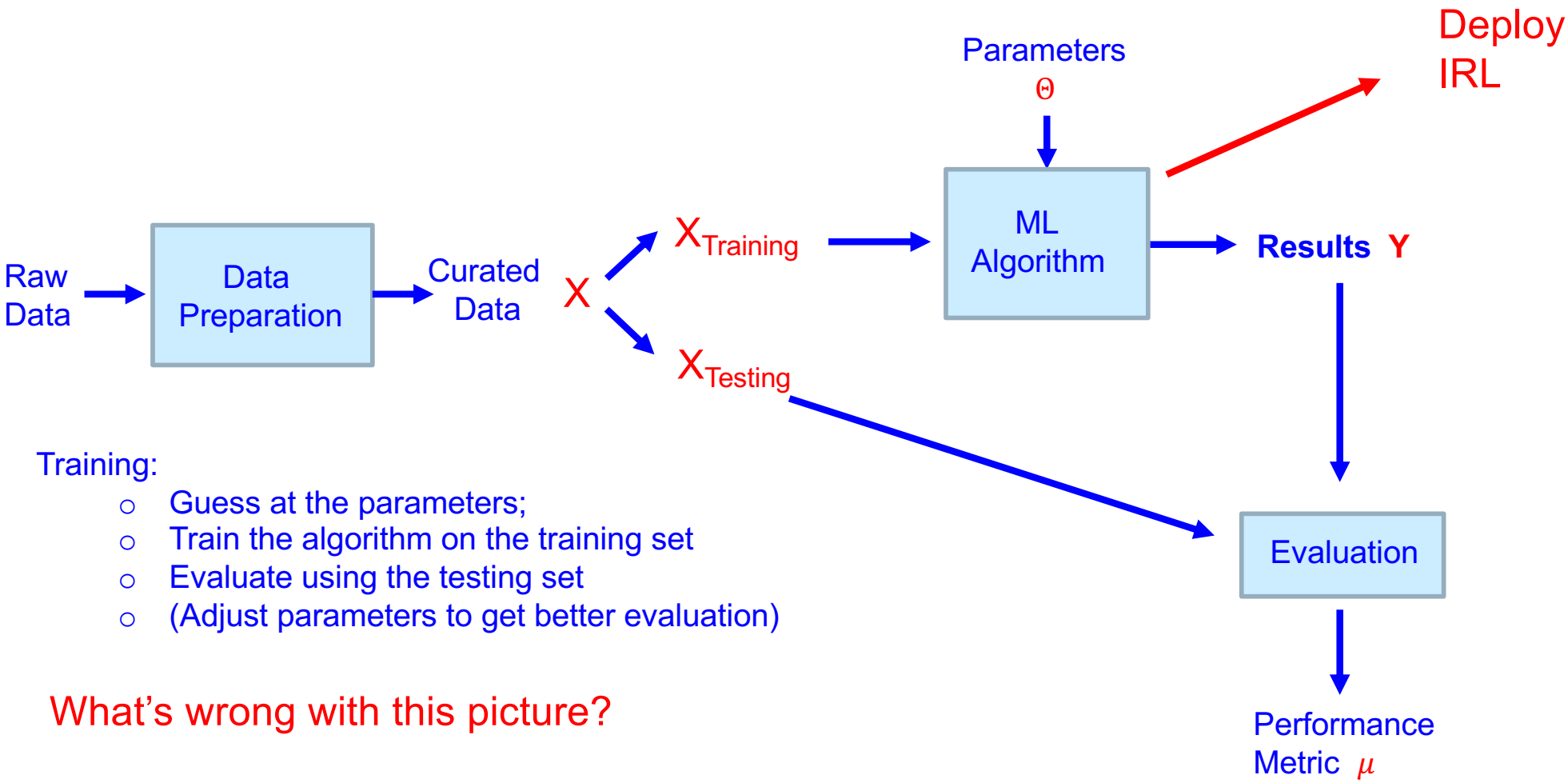


Training:

- Guess at the parameters;
- Train the algorithm on the training set
- Evaluate using the testing set
- (Adjust parameters to get better evaluation)

Introduction to Machine Learning

Supervised Machine Learning Workflow (Naive Version):



Training:

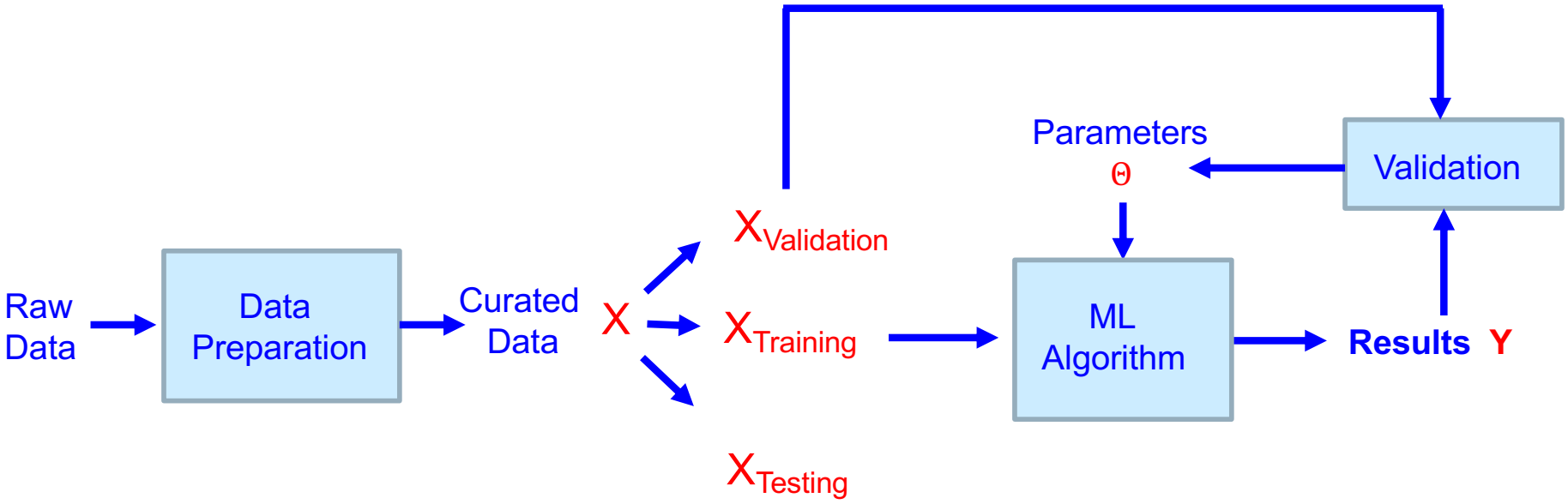
- Guess at the parameters;
- Train the algorithm on the training set
- Evaluate using the testing set
- (Adjust parameters to get better evaluation)

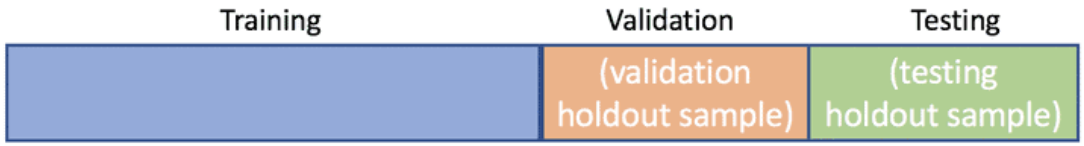
What's wrong with this picture?

Either you guess at the parameters, or adjust them to fit the testing set. You have no objective measure of quality!

Introduction to Machine Learning

Supervised Machine Learning Workflow (Actual):

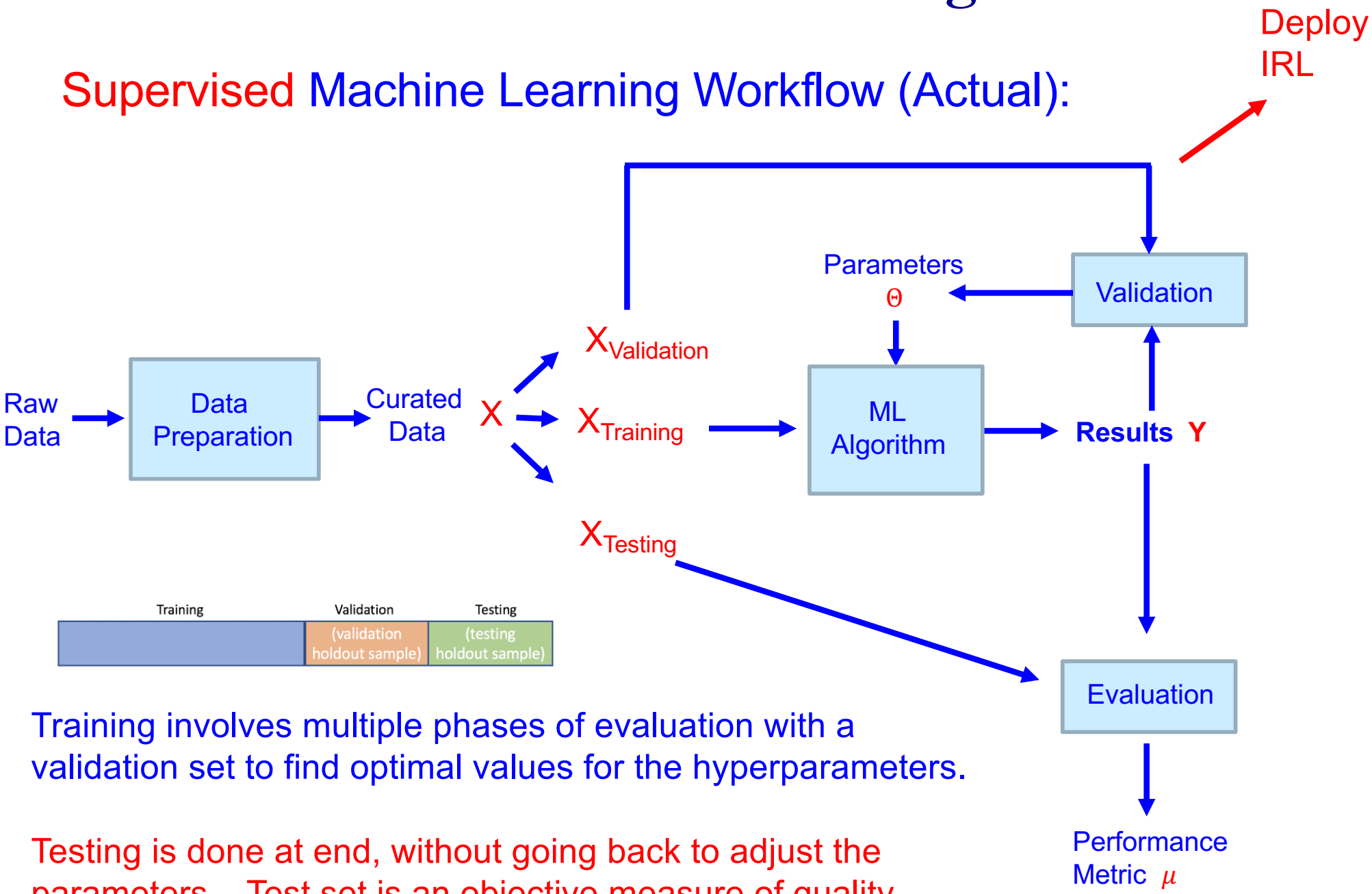




Training involves multiple phases of evaluation with a validation set to find optimal values for the hyperparameters.

Introduction to Machine Learning

Supervised Machine Learning Workflow (Actual):

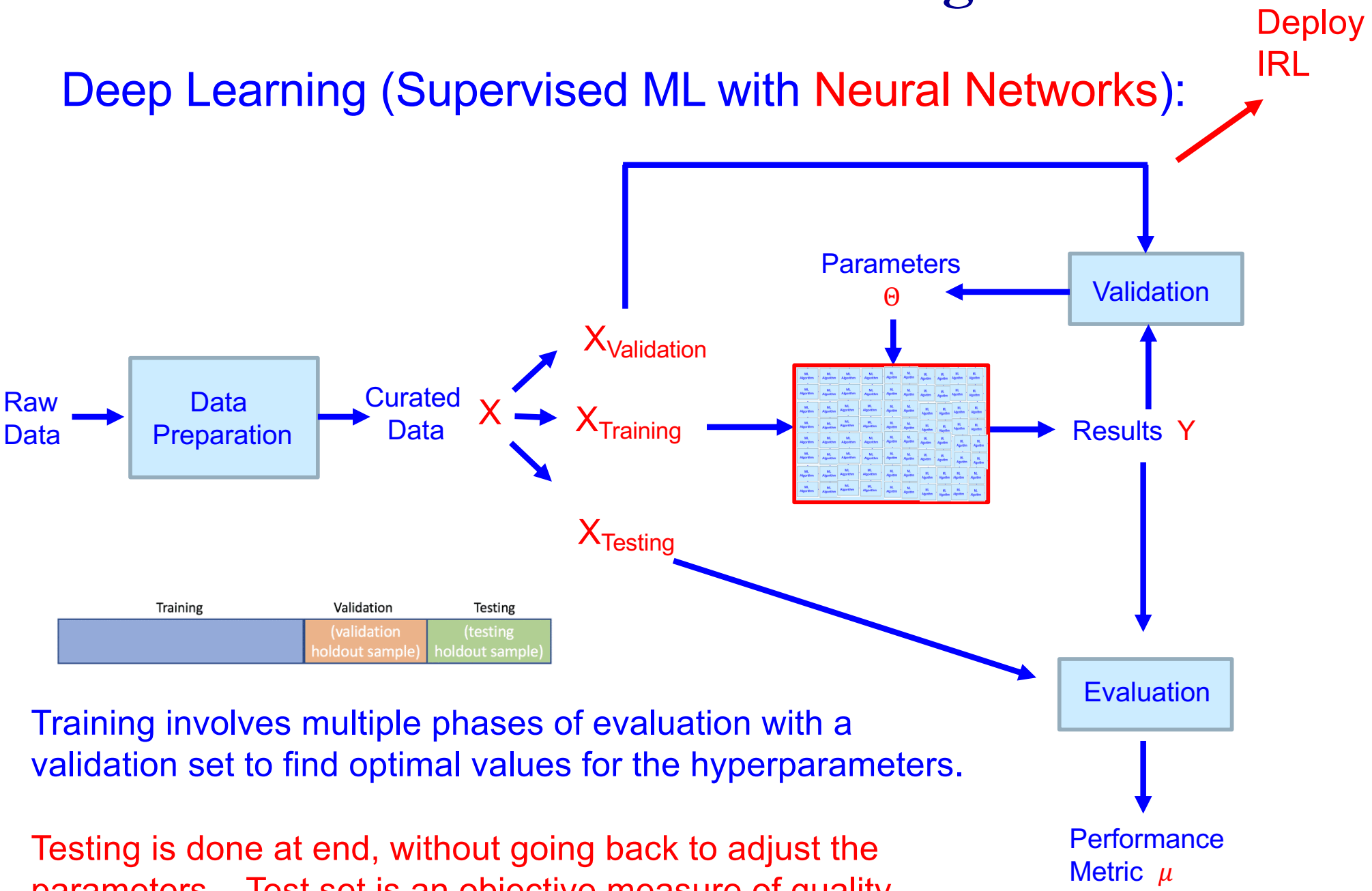


Training involves multiple phases of evaluation with a validation set to find optimal values for the hyperparameters.

Testing is done at end, without going back to adjust the parameters. Test set is an objective measure of quality.

Introduction to Machine Learning

Deep Learning (Supervised ML with Neural Networks):

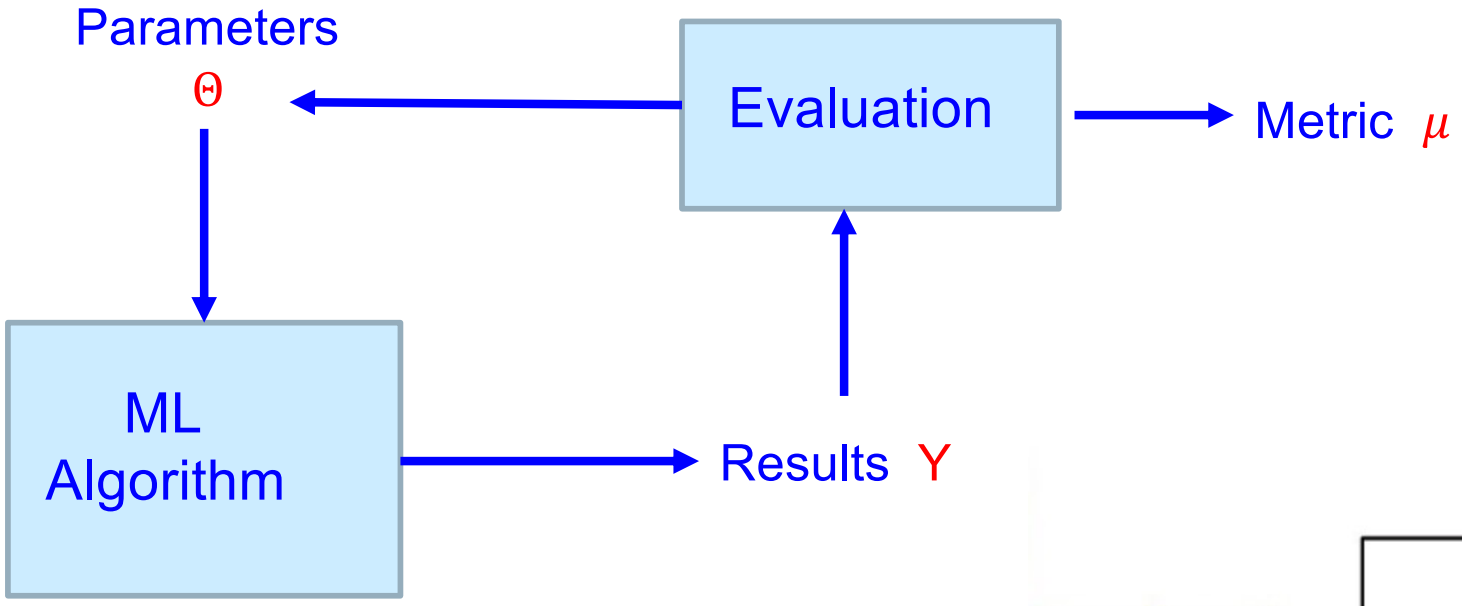


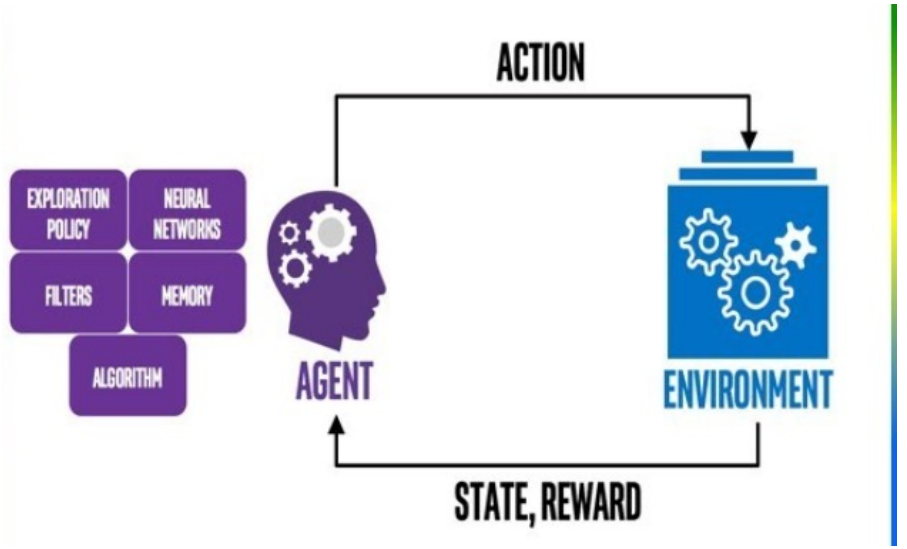
Training involves multiple phases of evaluation with a validation set to find optimal values for the hyperparameters.

Testing is done at end, without going back to adjust the parameters. Test set is an objective measure of quality.

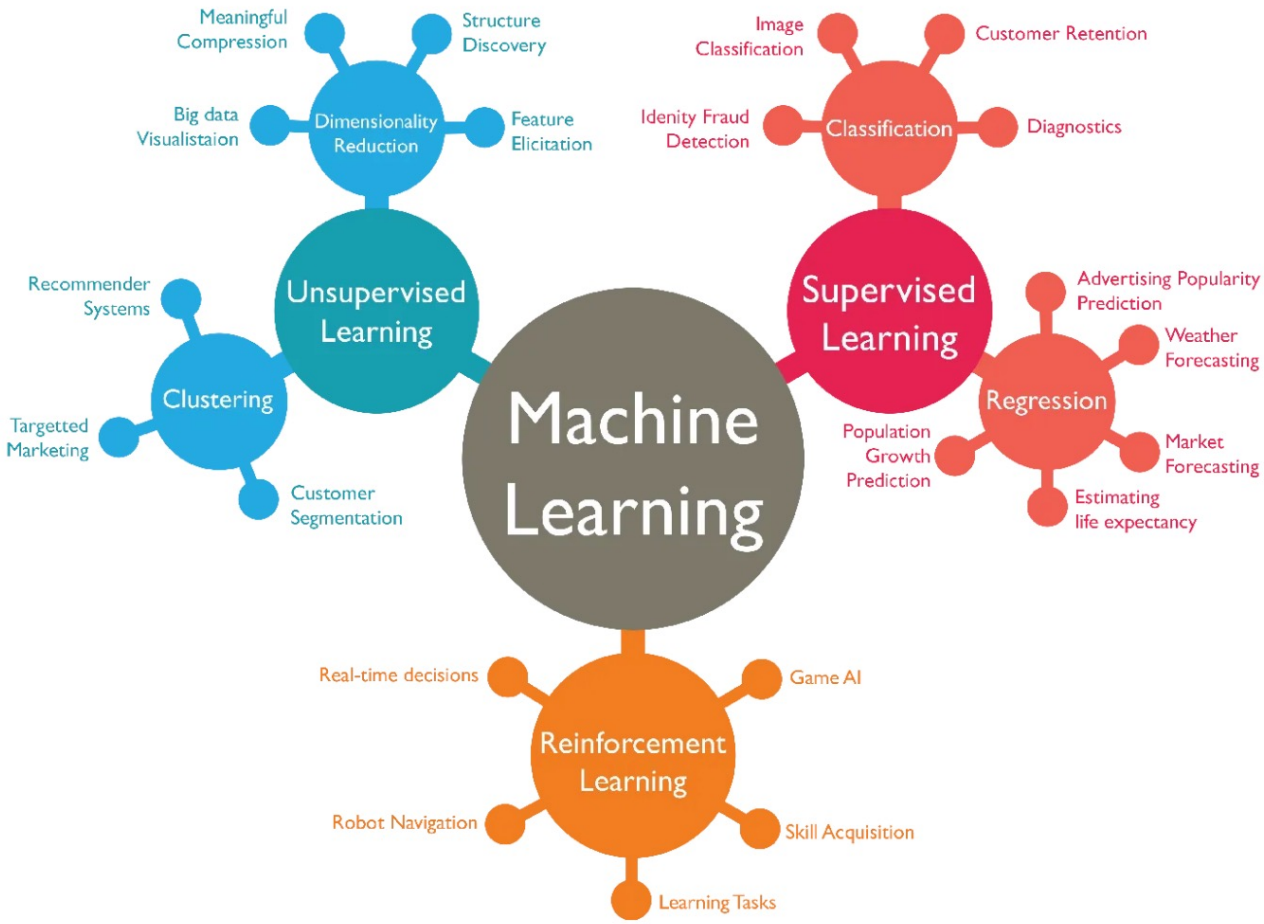
Introduction to Machine Learning

Reinforcement Machine Learning Workflow:





Introduction to Machine Learning

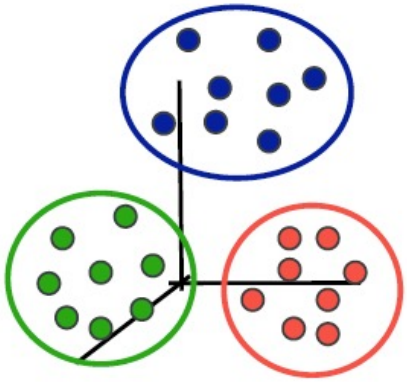


Unsupervised Learning: Clustering

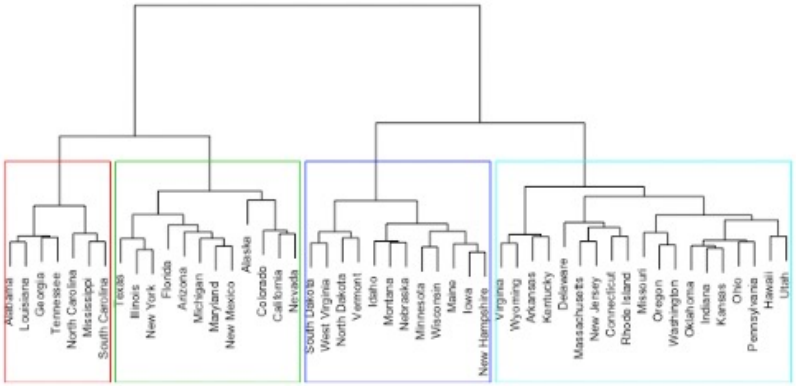
What is Clustering?

There are two basic types of clustering:

Partitioning



Hierarchical



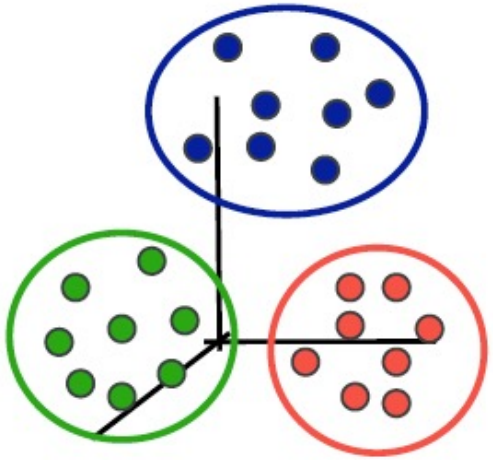
For now we will only consider partitioning algorithms: each object belongs to exactly one cluster.

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:

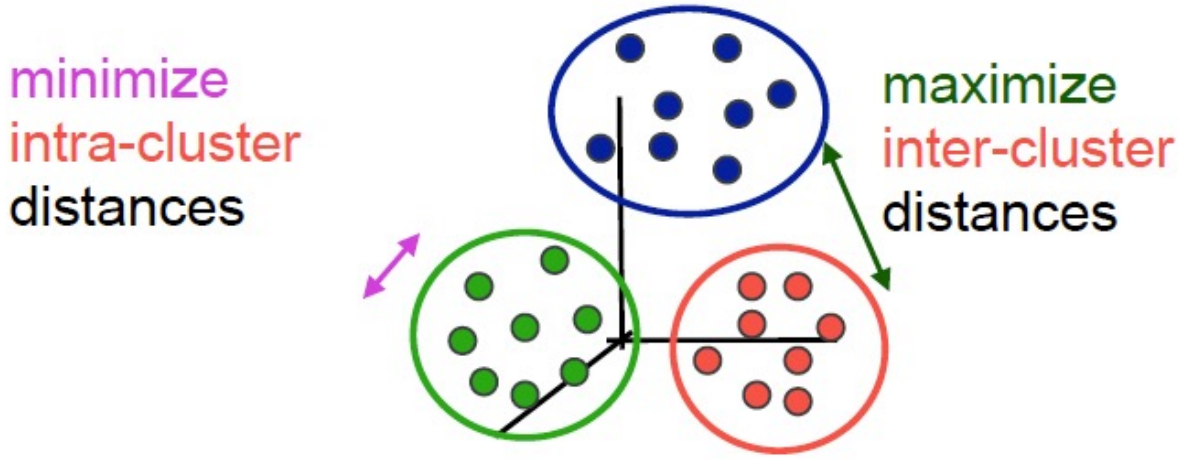
Cluster = group



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 a 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
- Classification where the only data available is unlabeled

Unsupervised Learning: Clustering

The Clustering Problem

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

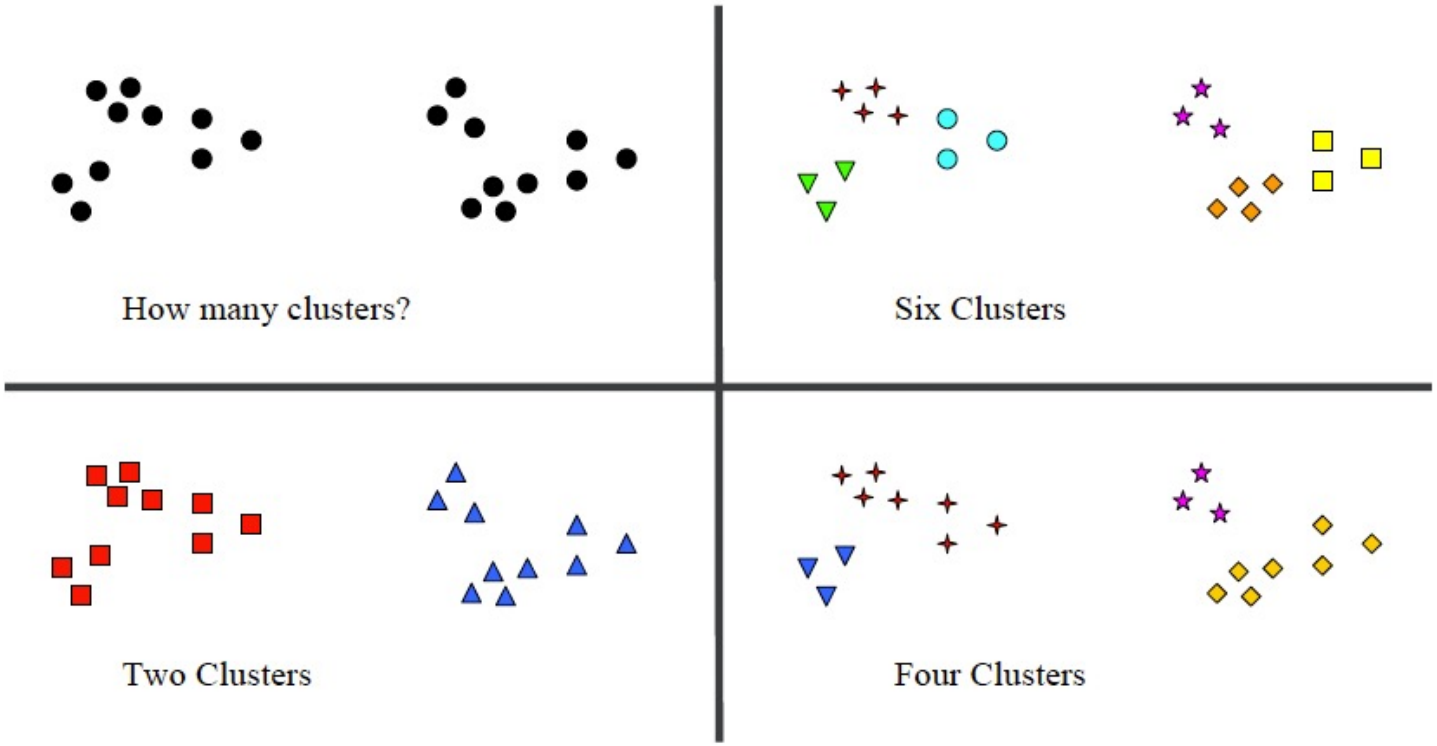
- Similar objects are in the same cluster
- Dissimilar objects are in different cluster

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.

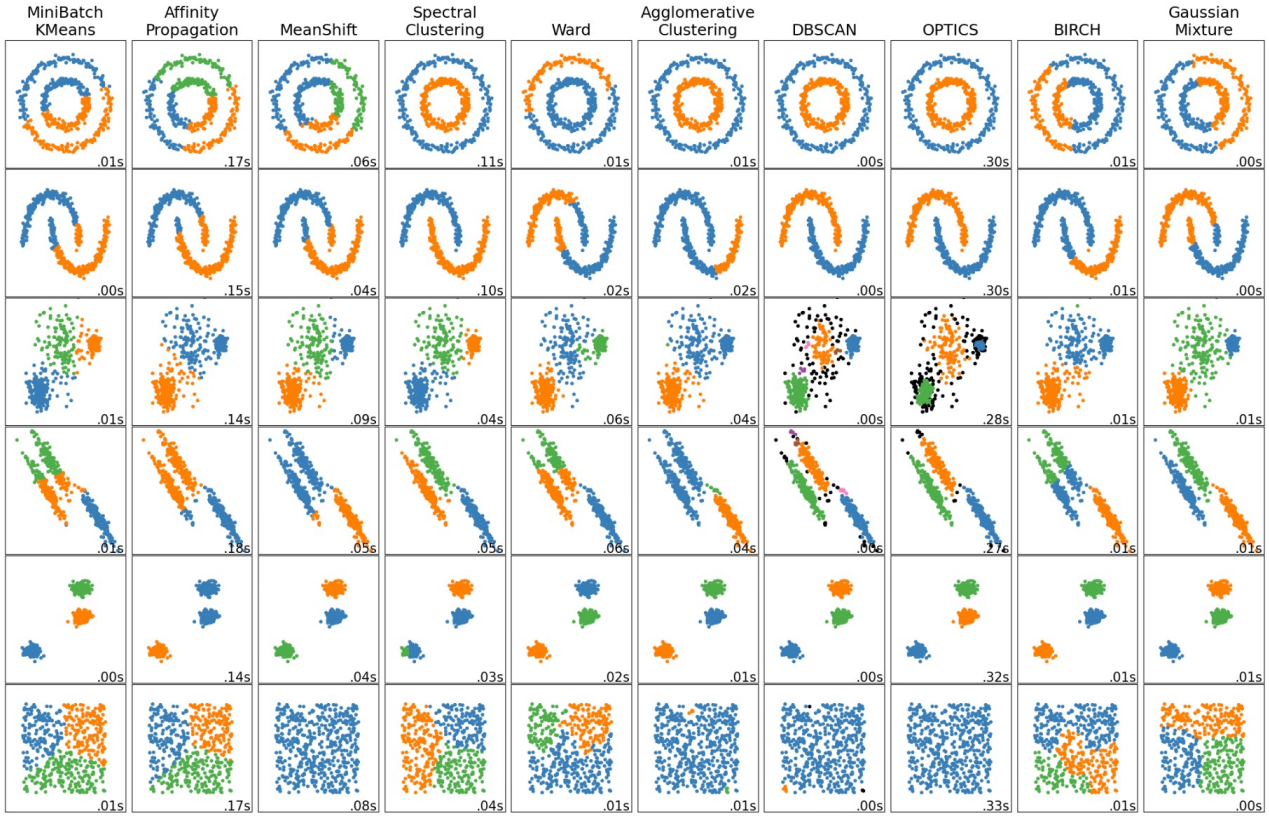
But in practice, iterative greedy algorithms work quite well!

Unsupervised Learning: Clustering

The K-Means Problem

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

2.3.1. Overview of clustering methods



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.

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

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

Unsupervised Learning: Clustering

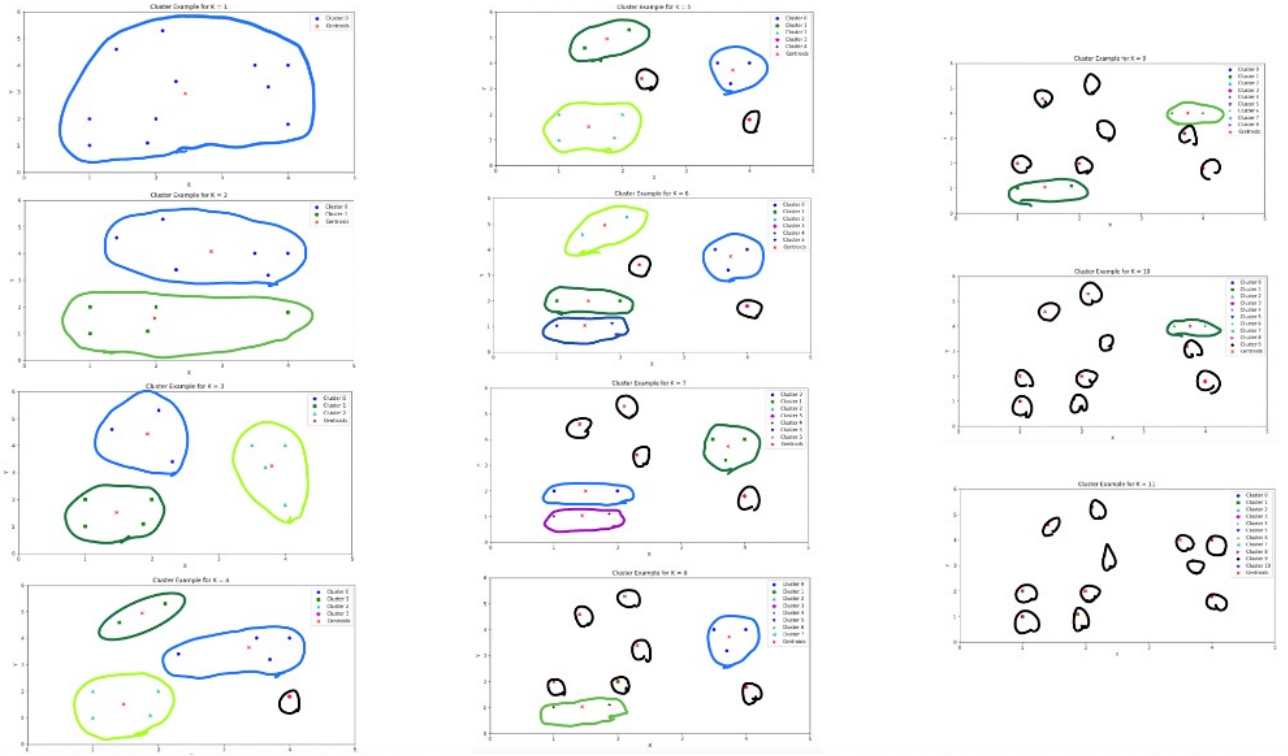
Lloyd's Algorithm for K-Means

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

Unsupervised Learning: Clustering

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

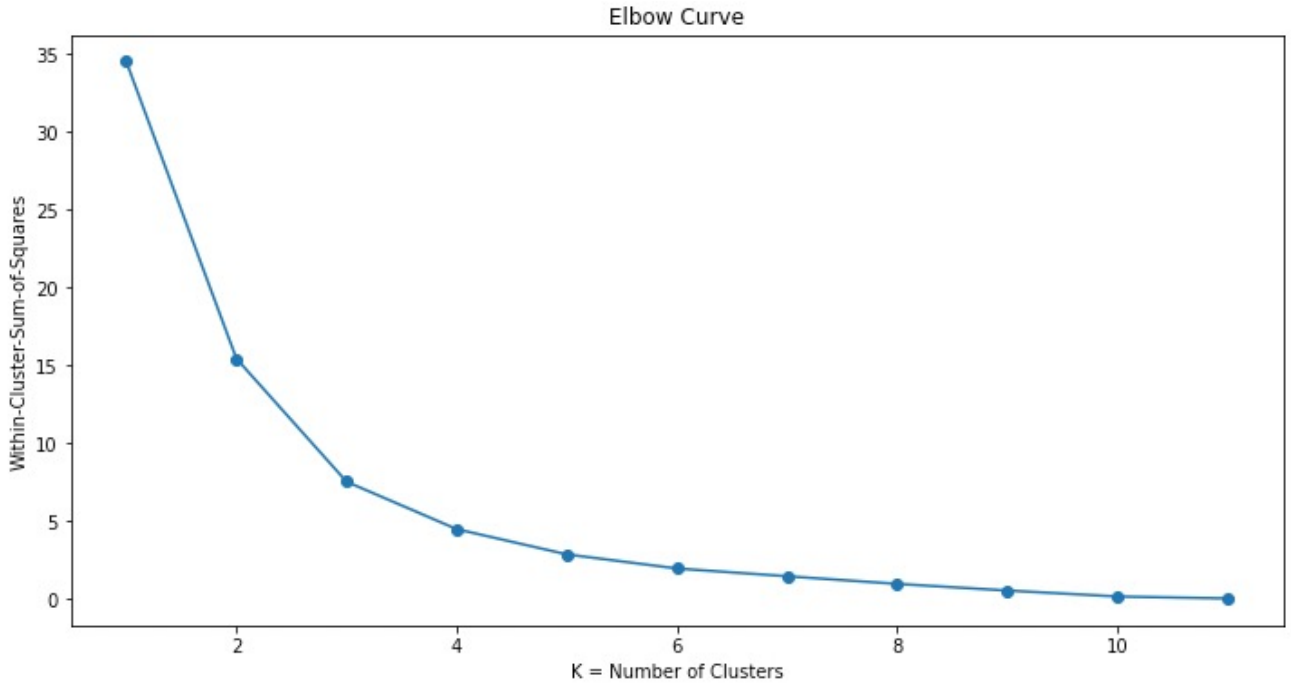
The main problem is that the cost (sum of squared intra-cluster distances) decreases as number of clusters gets smaller! Which is the best one?



Unsupervised Learning: Clustering

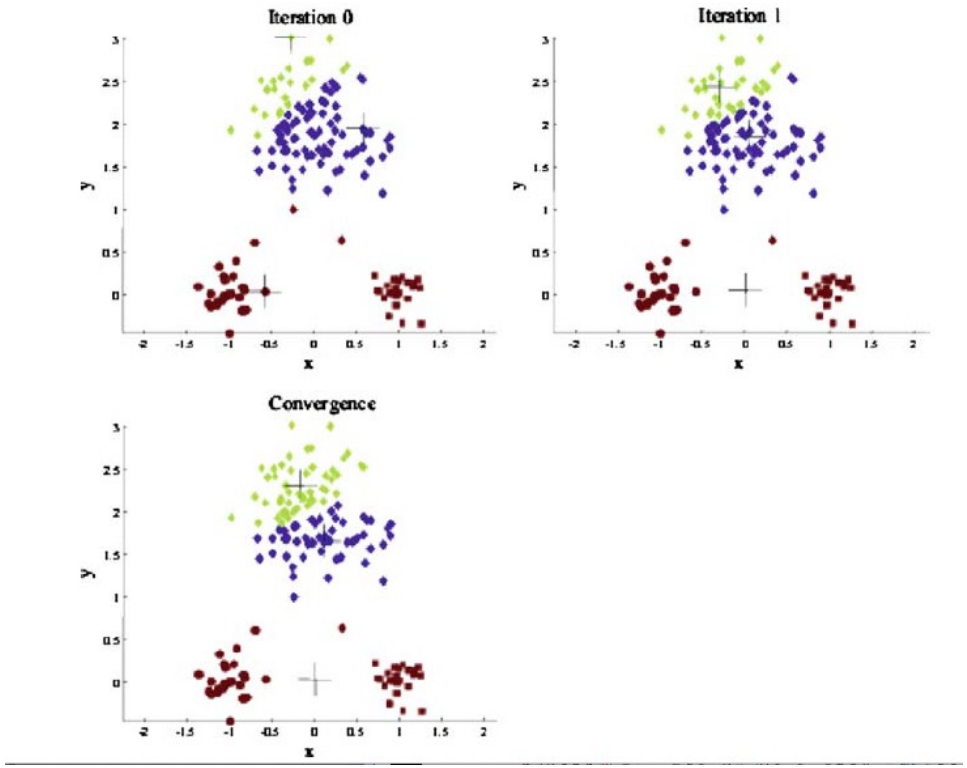
How to choose the right K??

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.



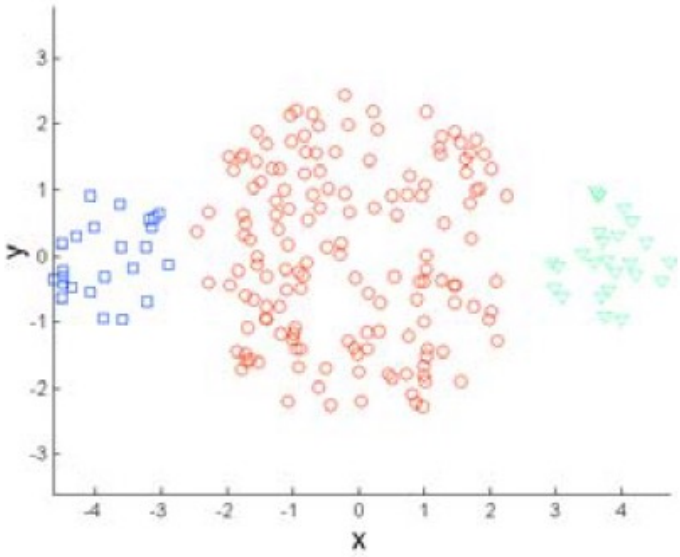
Unsupervised Learning: Clustering

Effect of a Bad Initialization

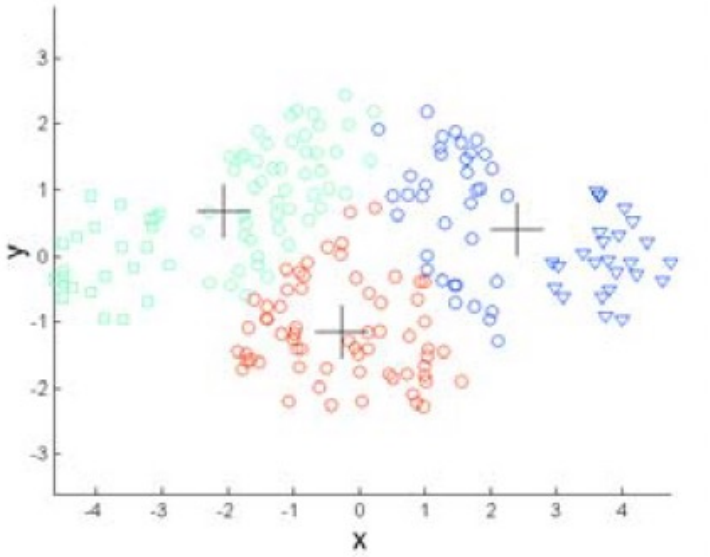


Unsupervised Learning: Clustering

Limitations of K-means: Clusters of different sizes



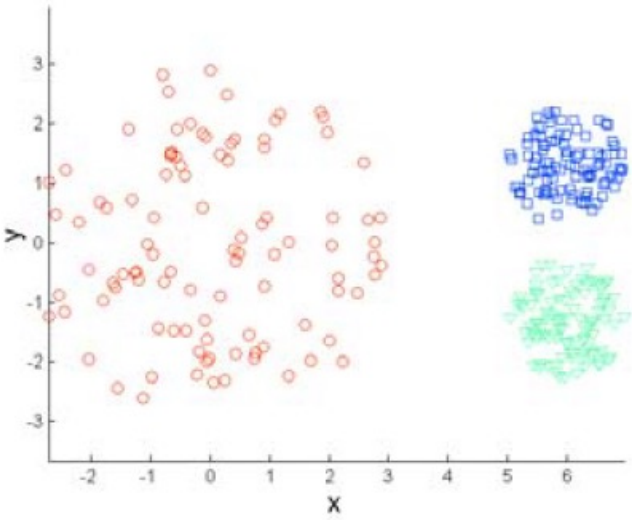
Original Points



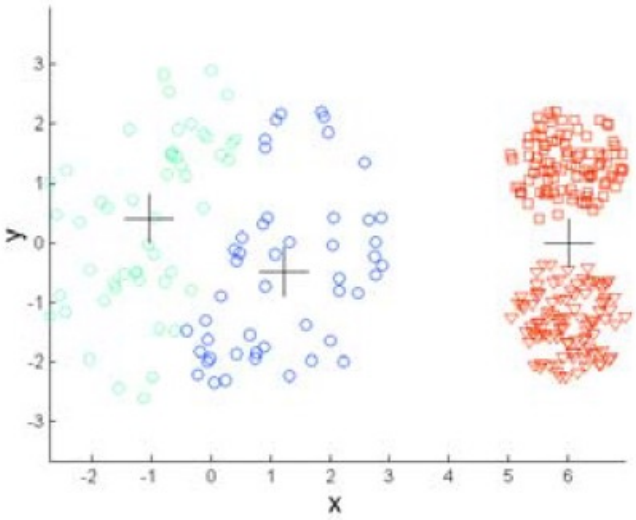
K-means (3 Clusters)

Unsupervised Learning: Clustering

Limitations of K-Means: Different Cluster Densities



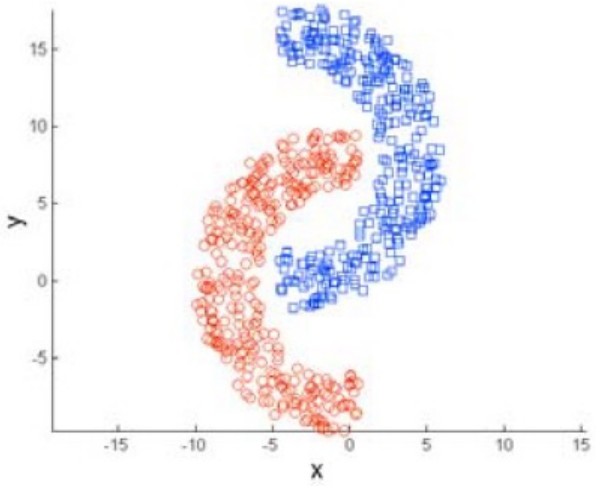
Original Points



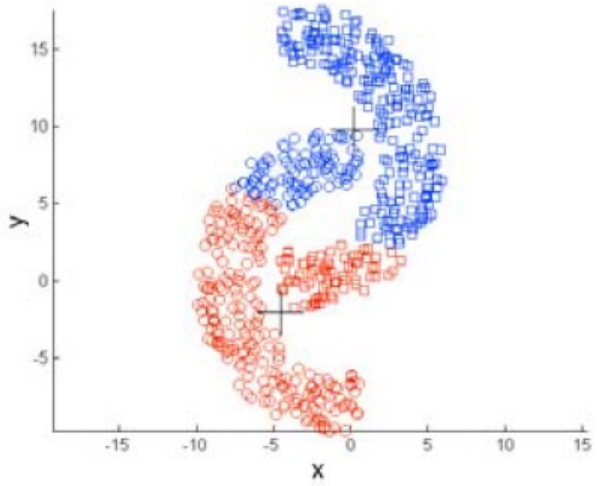
K-means (3 Clusters)

Unsupervised Learning: Clustering

Limitations of K-Means: non-spherical clusters



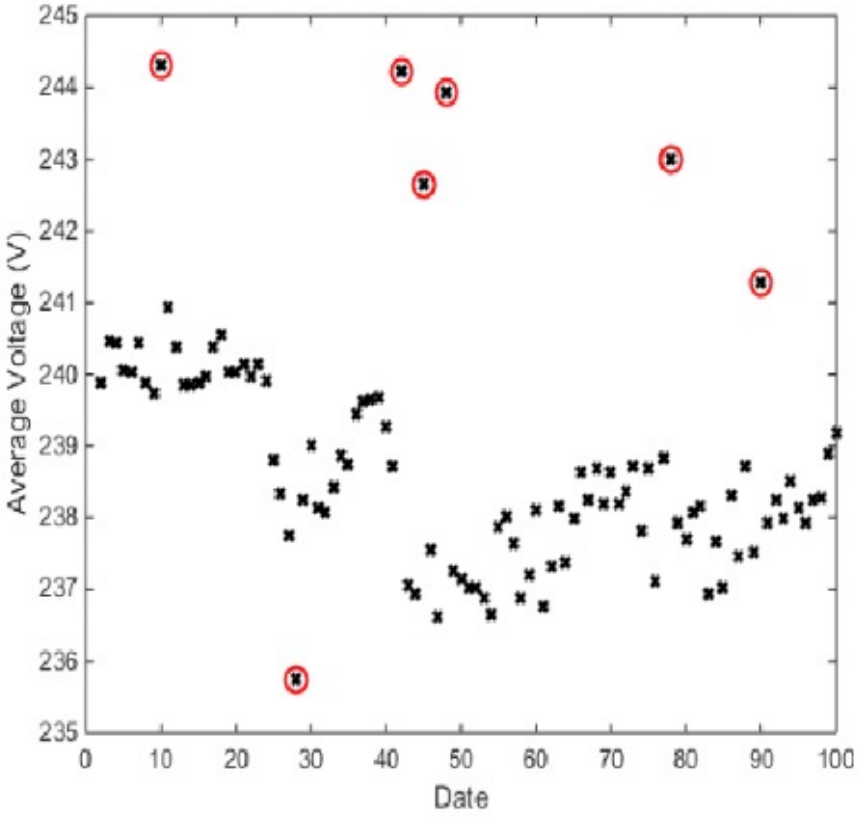
Original Points



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.

Hierarchical Clustering

Another framework for clustering, essentially different from K-Means and its variants, is Hierarchical Clustering, which has two flavors:

Divisive:

1. Start with all points in one cluster;
2. At each step, split until every point is in its own cluster

Agglomerative Clustering Algorithm:

1. Start with each point in its own cluster;
2. Compute the distance between all pairs of clusters;
3. Merge the two closest clusters;
4. Repeat 3 & 4 until only one cluster remains.

How do we define distance between clusters?

Each answer will give us a slightly different Hierarchical Clustering algorithm....

Hierarchical Clustering

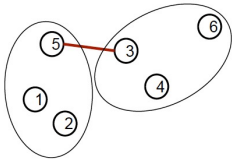
Agglomerative Clustering Algorithm:

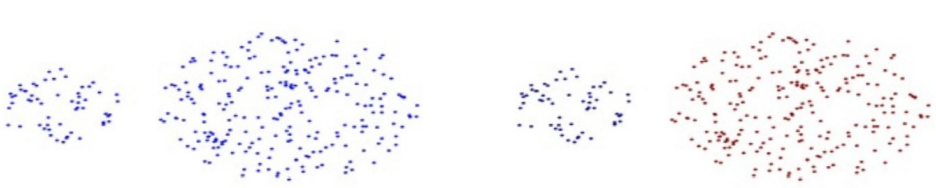
1. Start with each point in its own cluster;
2. Compute the distance between all pairs of clusters;
3. Merge the two closest clusters;
4. Repeat 3 & 4 until only one cluster remains.

Distance calculation:

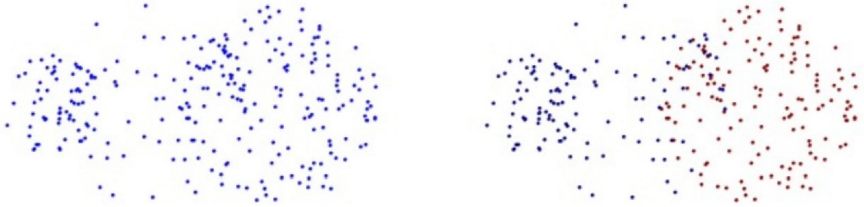
1. Single-Link Distance: Minimum distance between a point in one and a point in the other cluster:

$$D_{SL}(C_1, C_2) = \min \{d(p_1, p_2) \mid p_1 \in C_1, p_2 \in C_2\}$$



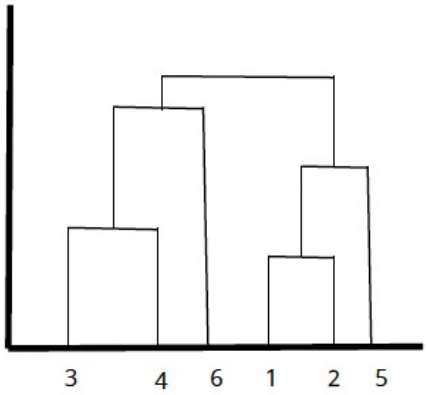


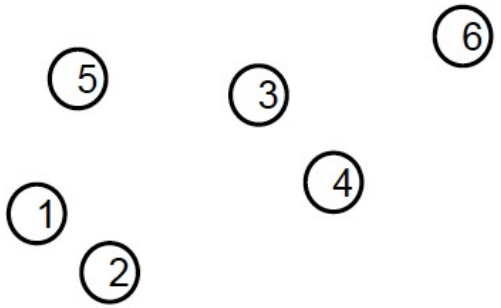
Can handle clusters of different sizes



But... Sensitive to noise points
Tends to create elongated clusters

Hierarchical Clustering





Hierarchical Clustering

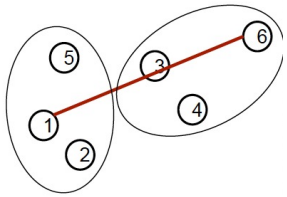
Agglomerative Clustering Algorithm:

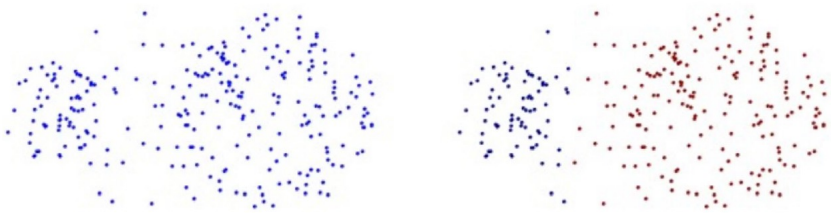
1. Start with each point in its own cluster;
2. Compute the distance between all pairs of clusters;
3. Merge the two closest clusters;
4. Repeat 3 & 4 until only one cluster remains.

Distance calculation:

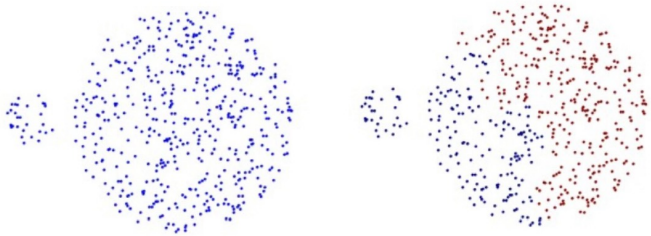
2. Complete-Link Distance: Maximum distance between a point in one and a point in the other cluster:

$$D_{CL}(C_1, C_2) = \max \{d(p_1, p_2) \mid p_1 \in C_1, p_2 \in C_2\}$$





Less susceptible to noise
Creates more balanced (equal diameter) clusters



But... Tends to split up large clusters.
All clusters tend to have the same diameter

Hierarchical Clustering

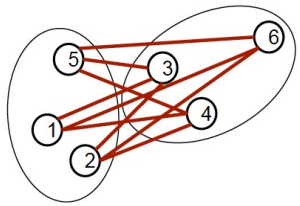
Agglomerative Clustering Algorithm:

1. Start with each point in its own cluster;
2. Compute the distance between all pairs of clusters;
3. Merge the two closest clusters;
4. Repeat 3 & 4 until only one cluster remains.

Distance calculation:

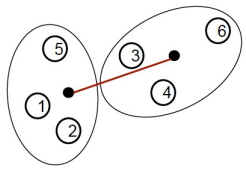
3. Average-Link Distance: Average distance between a point in one and a point in the other cluster:

$$D_{AL}(C_1, C_2) = \frac{1}{|C_1| \cdot |C_2|} \sum_{p_1 \in C_1, p_2 \in C_2} d(p_1, p_2)$$



Less susceptible to noise and outliers, but tends to be biased toward globular clusters....

Also: Centroid Distance, Ward's Distances, etc., etc.



Hierarchical Clustering

Agglomerative Clustering Algorithm:

1. Start with each point in its own cluster;
2. Compute the distance between all pairs of clusters;
3. Merge the two closest clusters;
4. Repeat 3 & 4 until only one cluster remains.

Finding the threshold with which to cut the dendrogram requires exploration and tuning. But in general hierarchical clustering is used to expose a hierarchy in the data (ex: finding/defining species via DNA similarity).