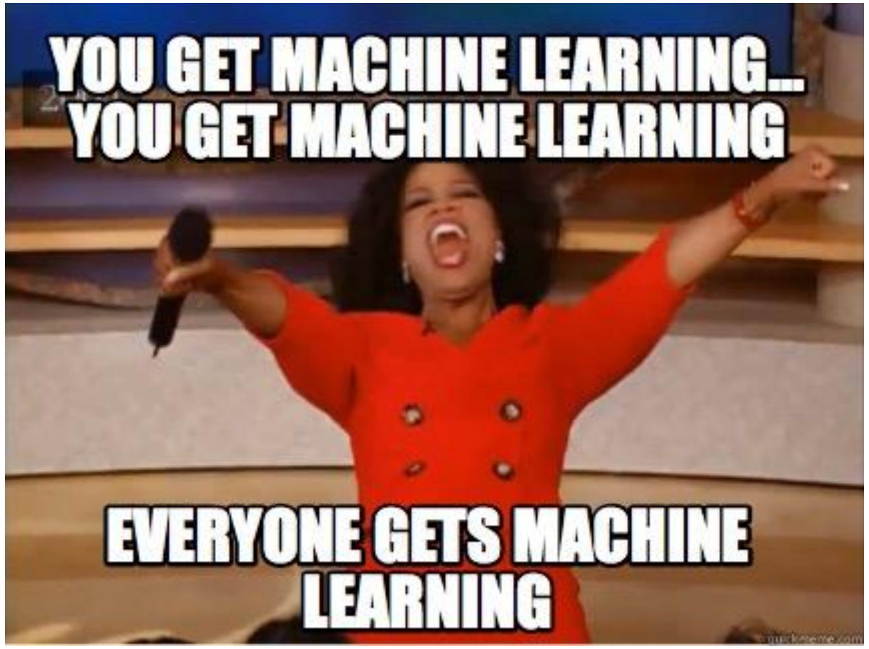
USC University of Southern California

TRGN 515:
Advanced Human Genomic Analysis Methods

Week 6 – Lecture 12:

Machine learning models for genomic data analysis



Bilgenur Baloglu, Ph.D.
- Clinical Instructor of Translational Genomics,
USC Keck School of Medicine
- Bioinformatics scientist III,
Thermo Fisher Scientific
Email: baloglu@usc.edu

Some of the slides adapted from:
Usman Roshan (NJIT)
Saleh Alkhalifa (Amgen)
Andrew Ng (Coursera)
Bing Liu (UIC)

On tap today!

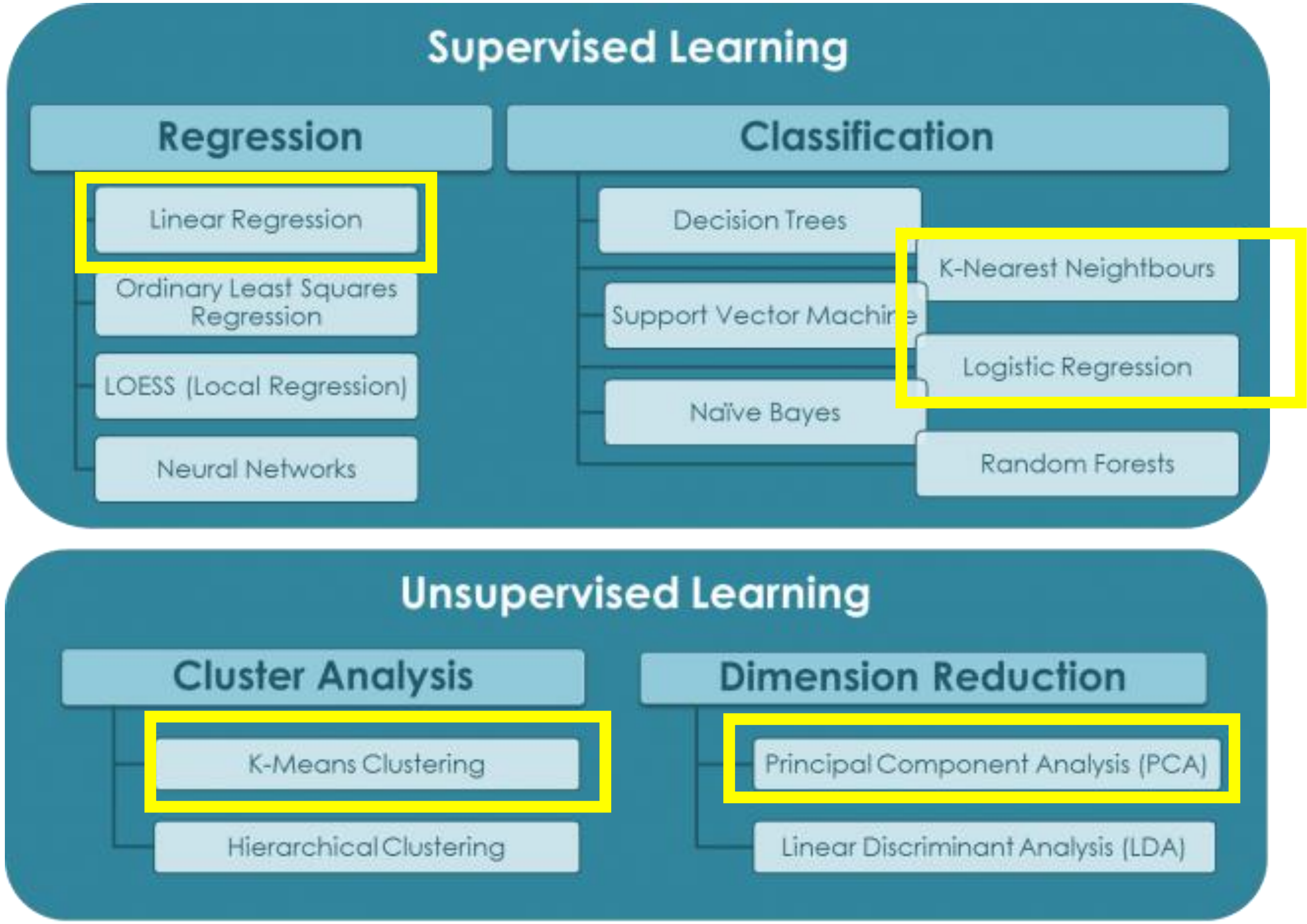
- Machine learning models
  - Supervised learning
    - Regression
    - Classification
  - Unsupervised learning
    - Cluster Analysis
    - Dimensionality reduction

What we focus on in this class: Machine Learning, not Deep Learning

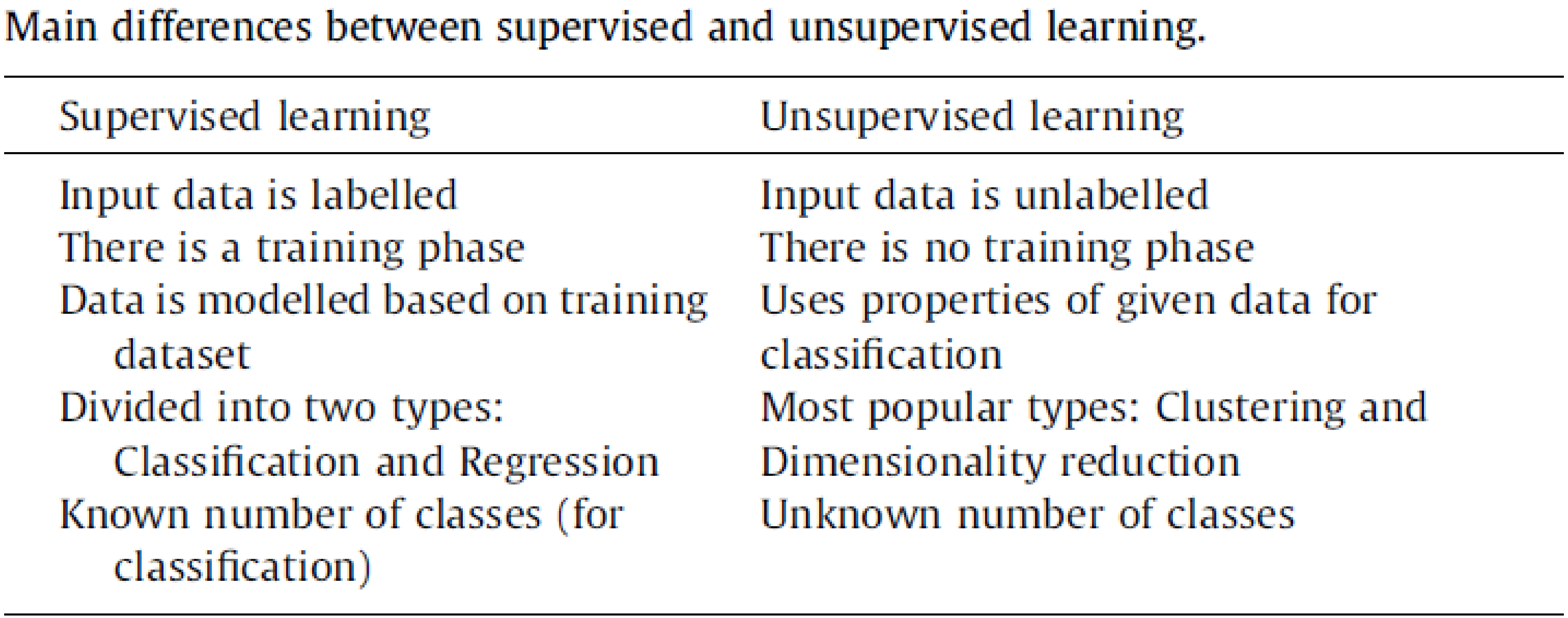
- Find articles in the resources/articles directory
- This week’s reading material:
  - Yang et al 2020, Review on the Application of Machine Learning Algorithms in the Sequence Data Mining of DNA
  - Monaco et al 2021, A primer on machine learning techniques for genomic applications
  - Wan et al 2021, Beyond sequencing: machine learning algorithms extract biology hidden in Nanopore signal data

Machine learning models

What models we'll cover



Types of Learning



Main differences between supervised and unsupervised learning.

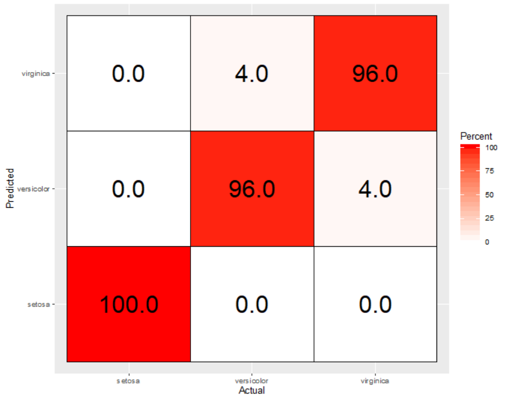
Supervised learning	Unsupervised learning
Input data is labelled	Input data is unlabelled
There is a training phase	There is no training phase
Data is modelled based on training dataset	Uses properties of given data for classification
Divided into two types: Classification and Regression	Most popular types: Clustering and Dimensionality reduction
Known number of classes (for classification)	Unknown number of classes

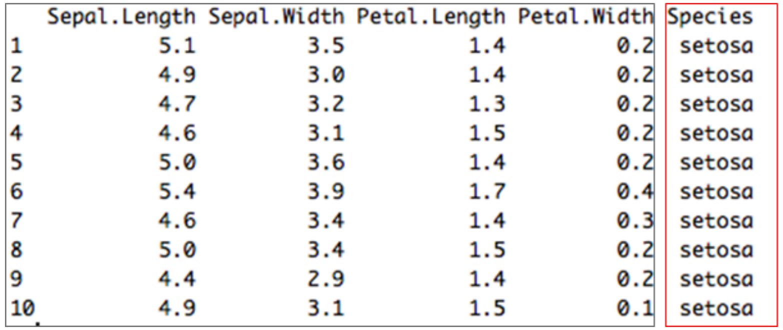
Monaco et al. 2021

Supervised learning

Supervised learning

- For every example in the data there is always a predefined outcome
- Models the relations between a set of descriptive features and a target
- 2 groups of problems:
  - Regression: Predicts continuous values
    - prediction of a quantitative phenotype such as age
  - Classification: Predicts which class a given sample of data (sample of descriptive features) is part of (**discrete value**)
    - cancer type or disease trait classification



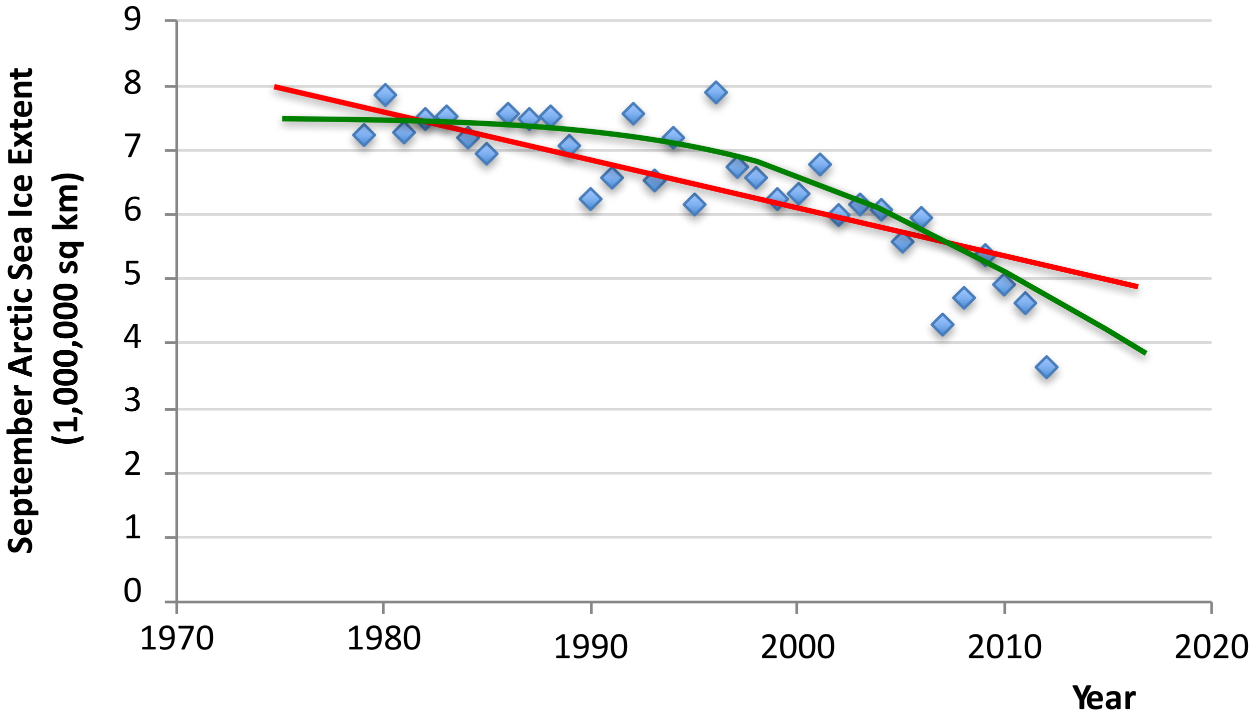


```
   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1           5.1         3.5          1.4         0.2  setosa
2           4.9         3.0          1.4         0.2  setosa
3           4.7         3.2          1.3         0.2  setosa
4           4.6         3.1          1.5         0.2  setosa
5           5.0         3.6          1.4         0.2  setosa
6           5.4         3.9          1.7         0.4  setosa
7           4.6         3.4          1.4         0.3  setosa
8           5.0         3.4          1.5         0.2  setosa
9           4.4         2.9          1.4         0.2  setosa
10          4.9         3.1          1.5         0.1  setosa
```

Supervised learning: Regression

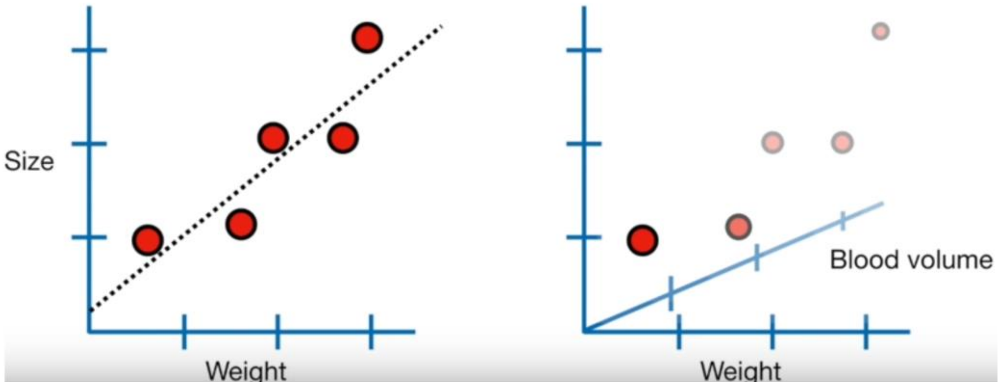
Supervised Learning: Regression

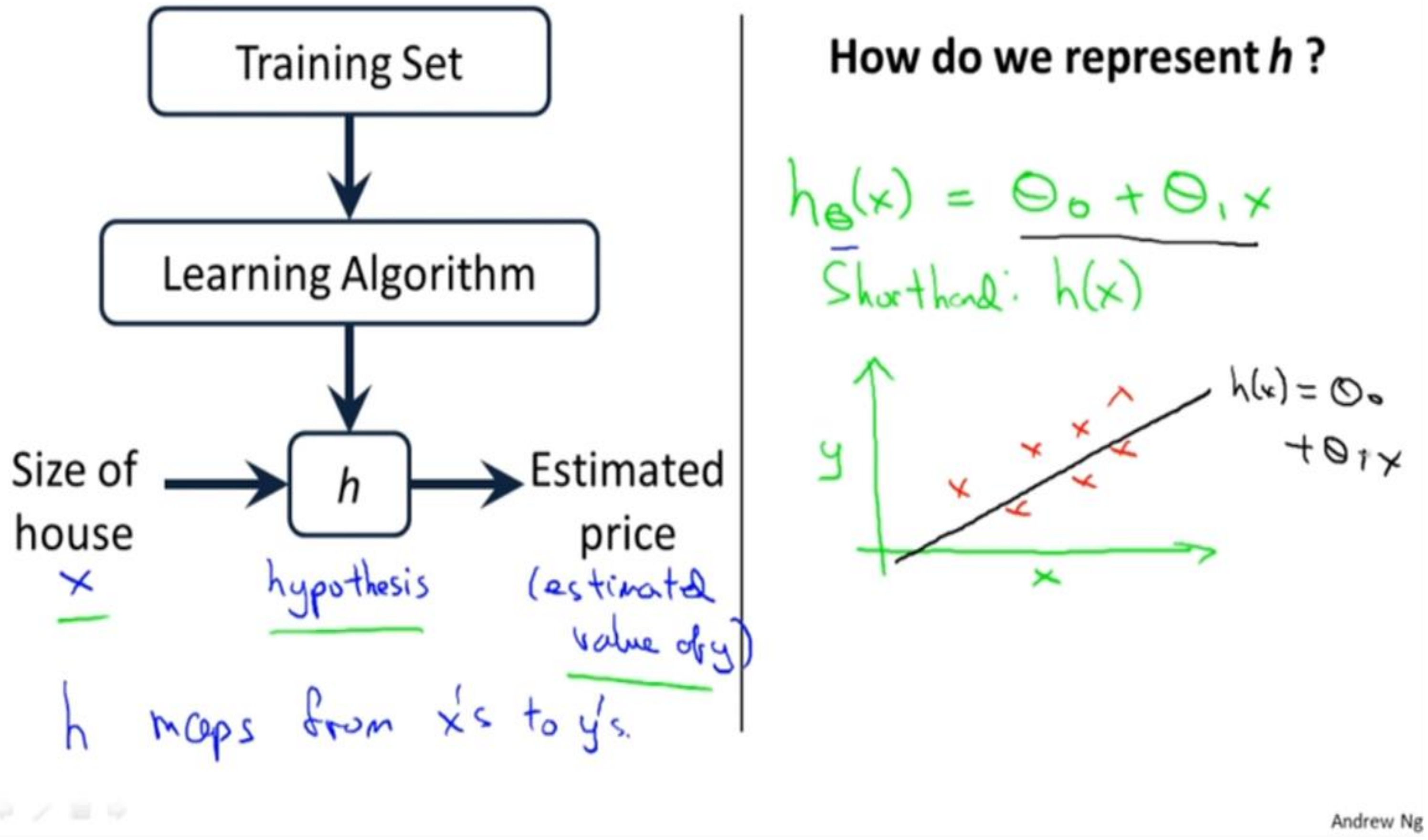
- Given $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Learn a function $f(x)$ to predict y given x
  - y is real-valued == regression



Data from G. Witt. Journal of Statistics Education, Volume 21, Number 1 (2013)

Linear regression in simple terms





Training Set

Learning Algorithm

Size of house → h → Estimated price

x

hypothesis

(estimated value of y)

h maps from x's to y's

How do we represent *h* ?

$h_\theta(x) = \theta_0 + \theta_1 x$

Shorthand: $h(x)$

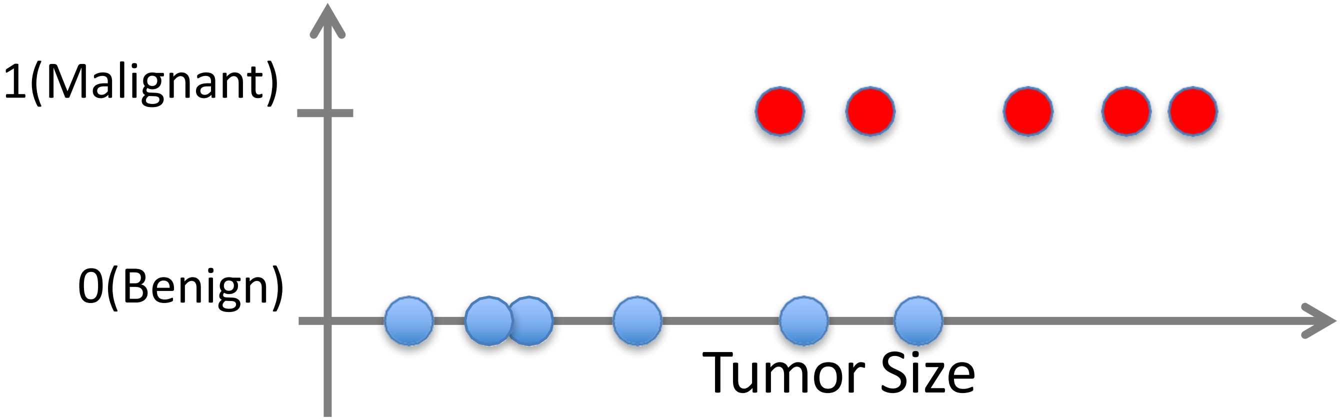
Andrew Ng

Supervised learning: Classification

Supervised Learning: Classification

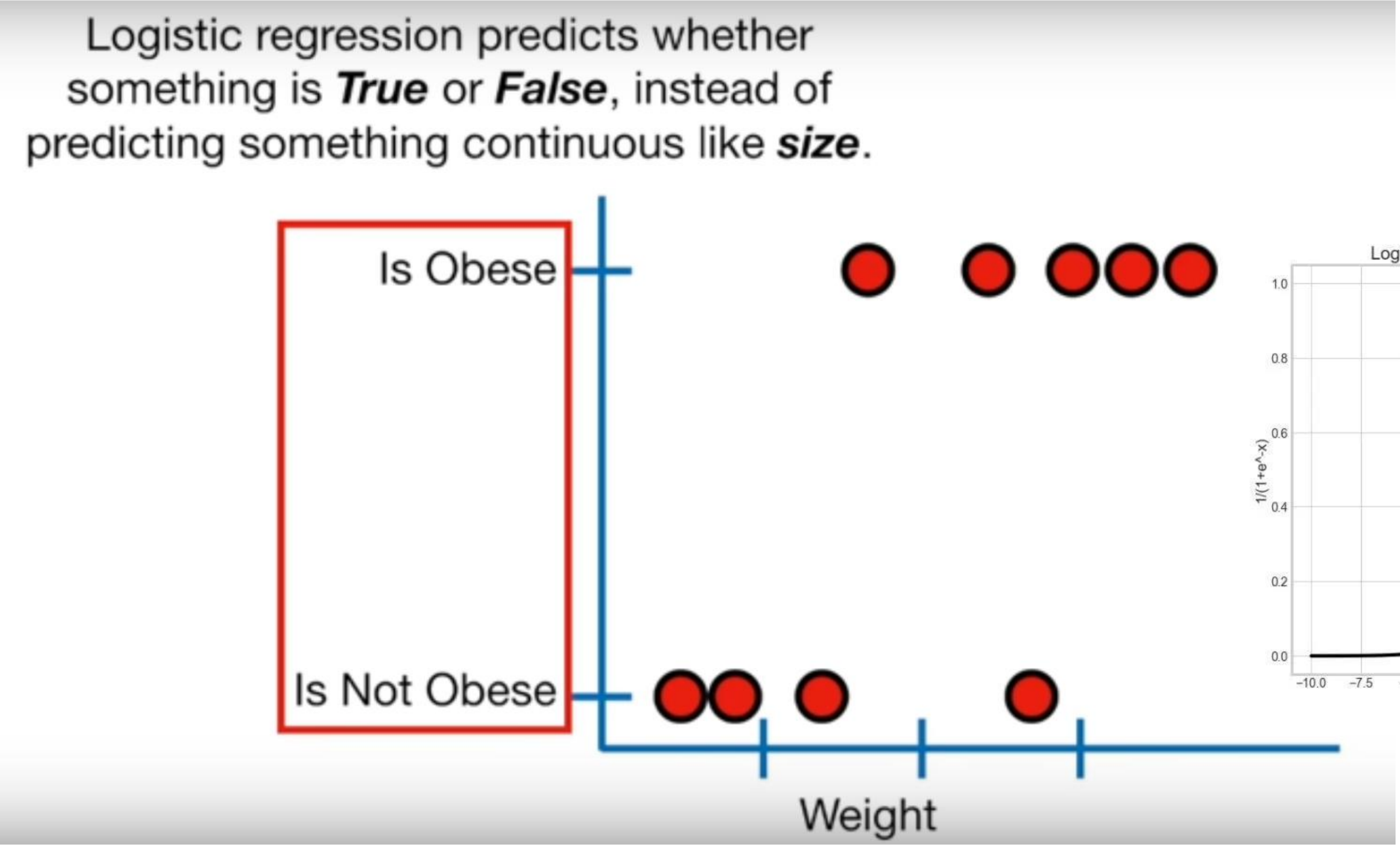
- Given $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Learn a function $f(x)$ to predict y given x
  - y is categorical == classification

Breast Cancer (Malignant / Benign)

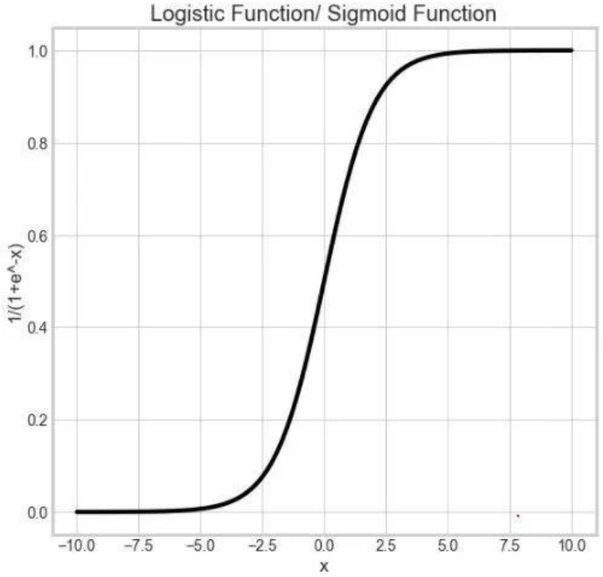


Based on example by Andrew Ng

Logistic regression



Logistic regression predicts whether something is ***True*** or ***False***, instead of predicting something continuous like ***size***.



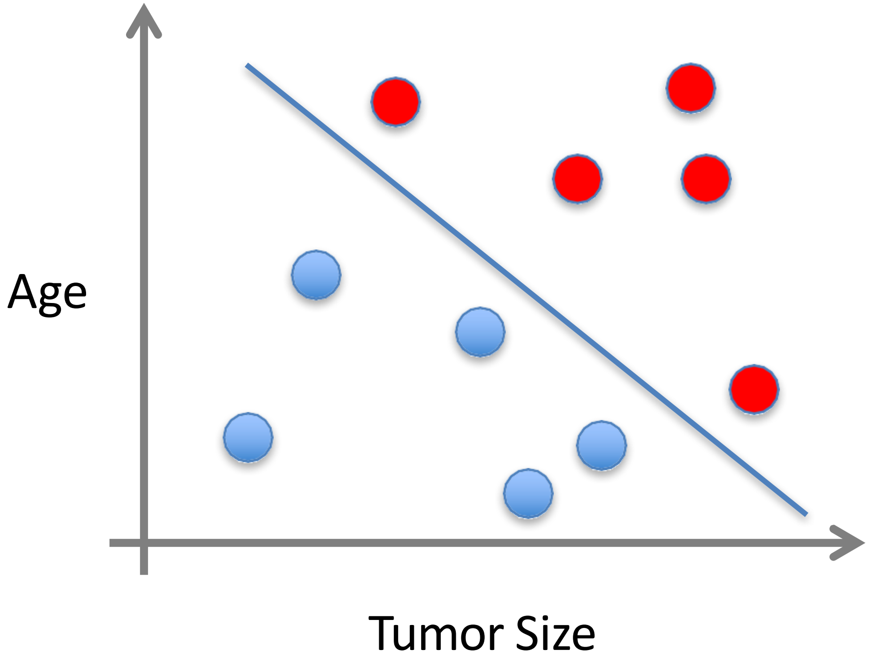
$$S(x) = \frac{1}{1 + e^{-x}}$$

$x \to \infty$ ➡ $S(x) \to 1$

$x \to -\infty$ ➡ $S(x) \to 0$

Supervised Learning: Classification

- *x* can be multi-dimensional
  - – Each dimension corresponds to a feature



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

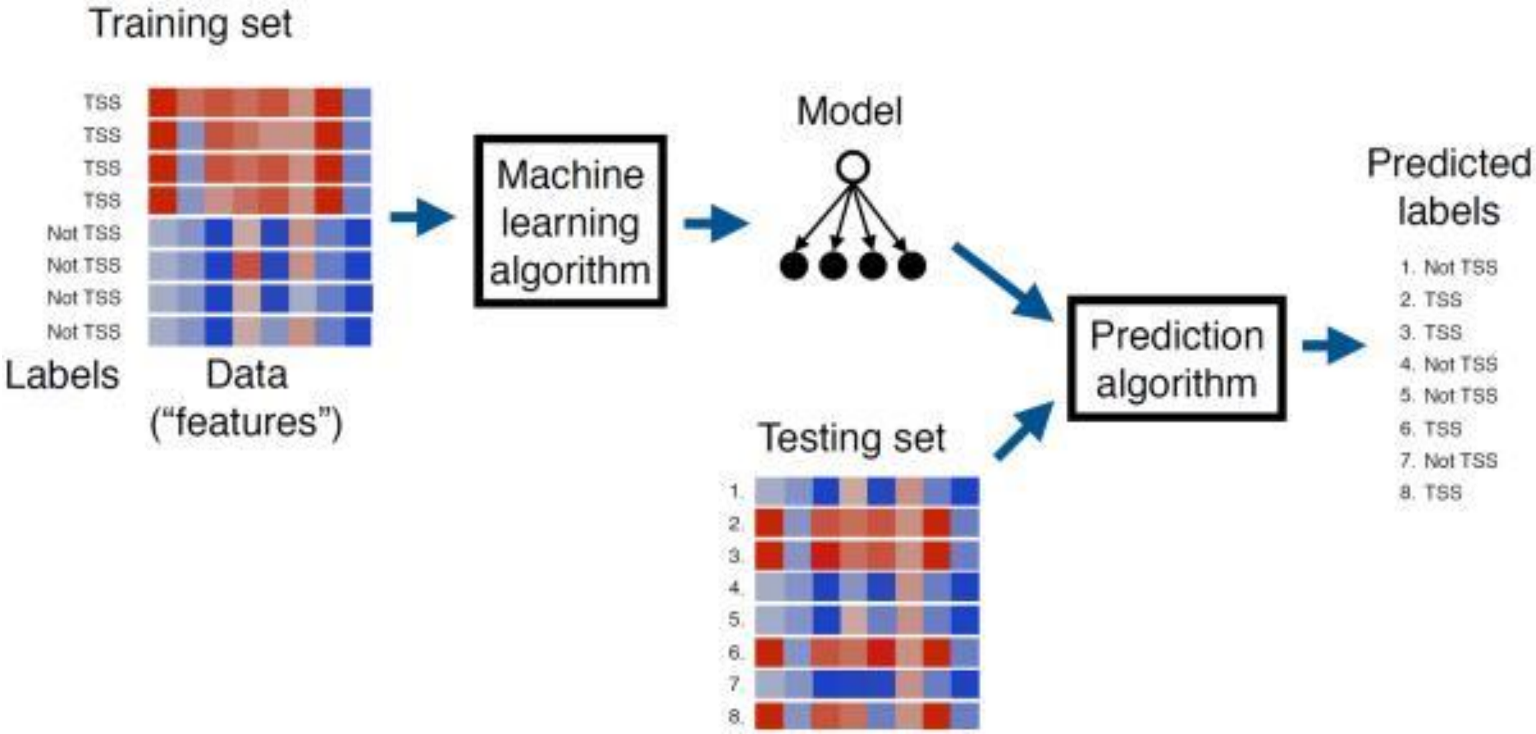
…

Based on example by Andrew Ng

Classification in Bioinformatics

- Computational diagnostic: early cancer detection
- Tumor biomarker discovery
- Protein structure prediction (threading)
- Protein-protein binding sites prediction
- Gene function prediction

Supervised learning for genomics data



Libbrecht, 2015

Example: Breast tumor classification

van 't Veer et al (2002) Nature 415, 530

Dutch Cancer Institute (NKI)

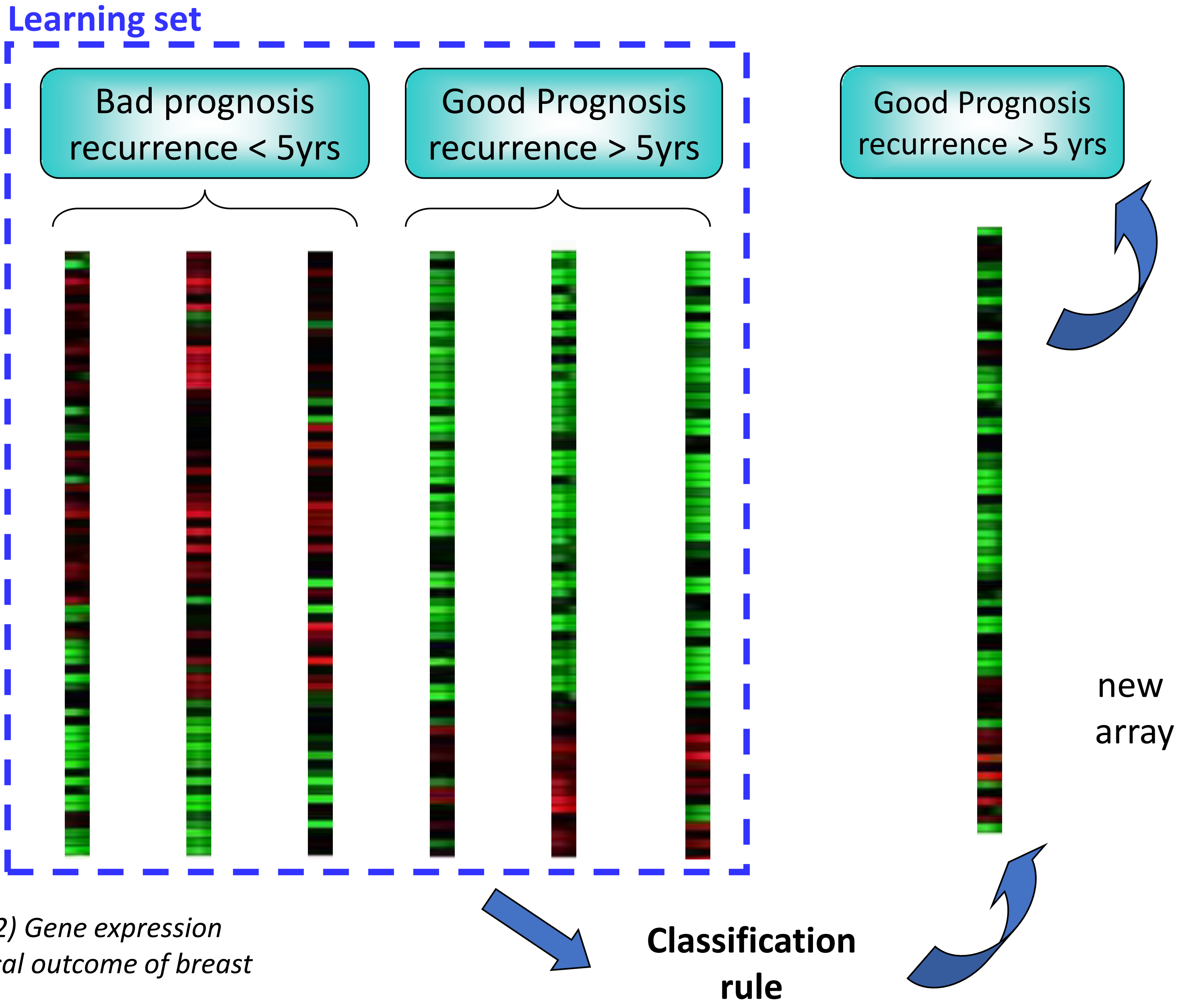
Prediction of clinical outcome of breast cancer

DNA microarray experiment
117 patients
25000 genes

Predefine classes
Clinical outcome

Objects
Array

Feature vectors
Gene expression



Learning set

Bad prognosis
recurrence < 5yrs

Good Prognosis
recurrence > 5yrs

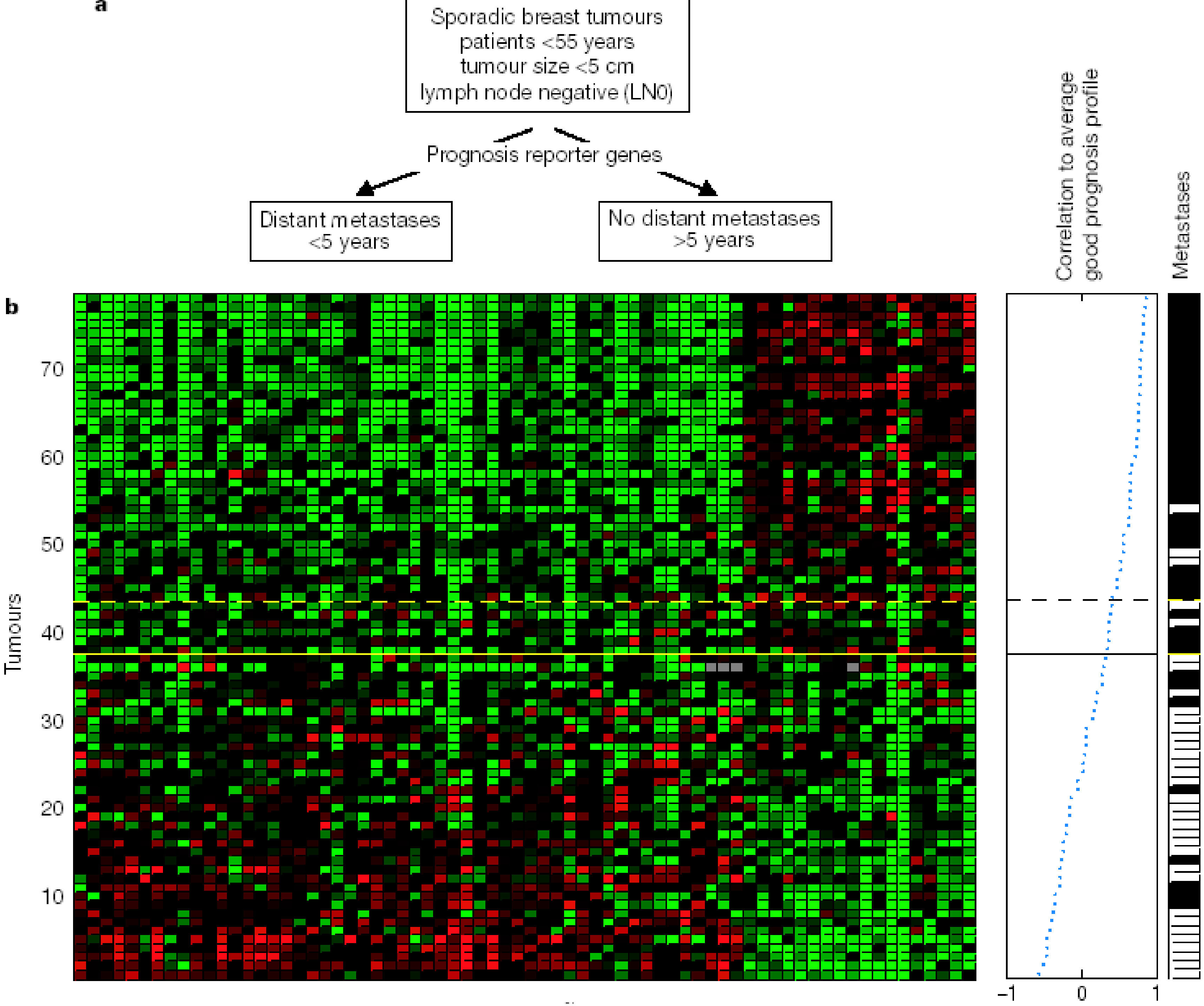
Good Prognosis
recurrence > 5 yrs

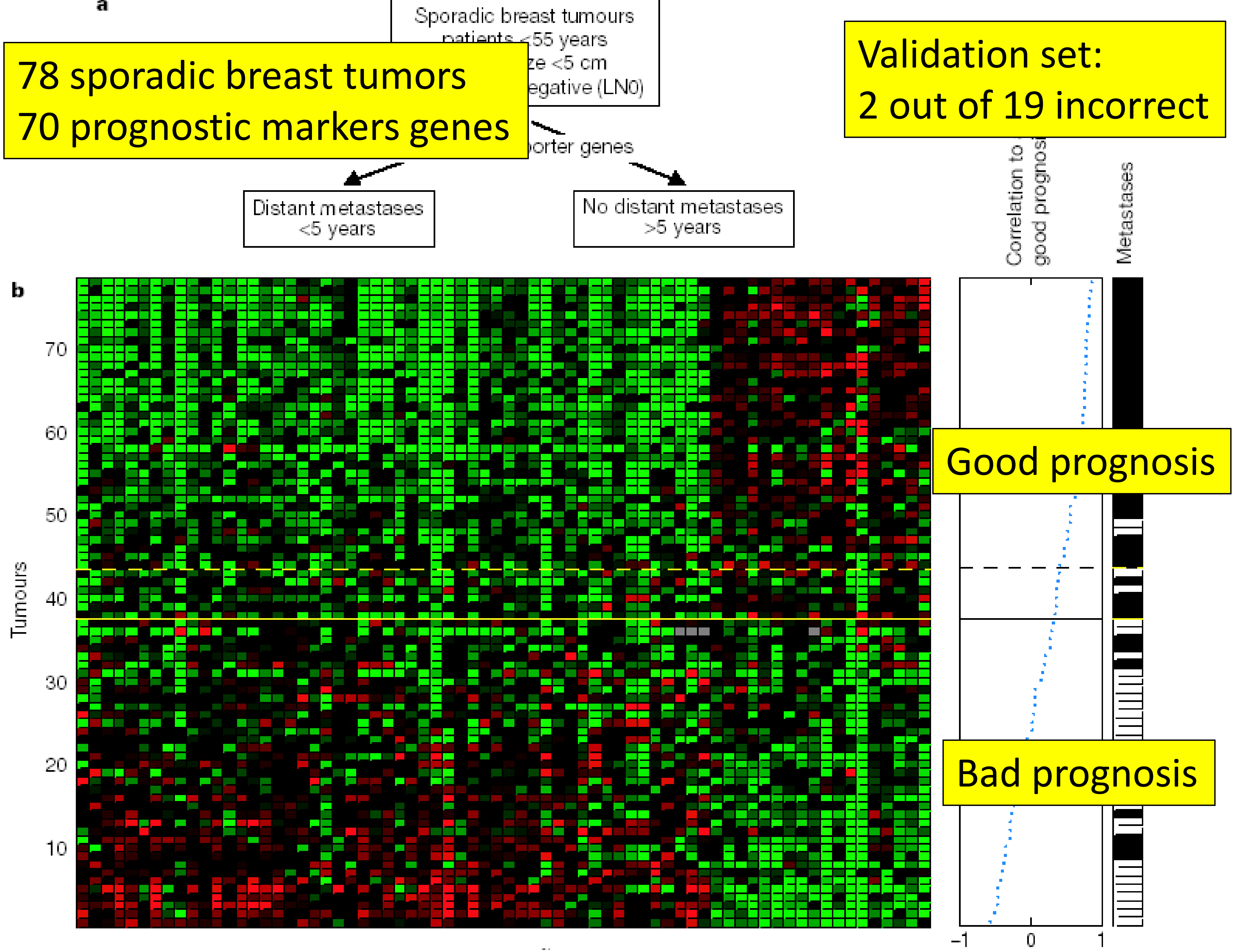
new
array

Classification rule

Reference
L van't Veer *et al (2002) Gene expression profiling predicts clinical outcome of breast cancer.* Nature, Jan.

.


a
Sporadic breast tumours
patients <55 years
tumour size <5 cm
lymph node negative (LN0)
Prognosis reporter genes
Distant metastases
<5 years
No distant metastases
>5 years
b
Tumours
70
60
50
40
30
20
10
Correlation to average
good prognosis profile
Metastases
–1
0
1



78 sporadic breast tumors
70 prognostic markers genes

Validation set:
2 out of 19 incorrect

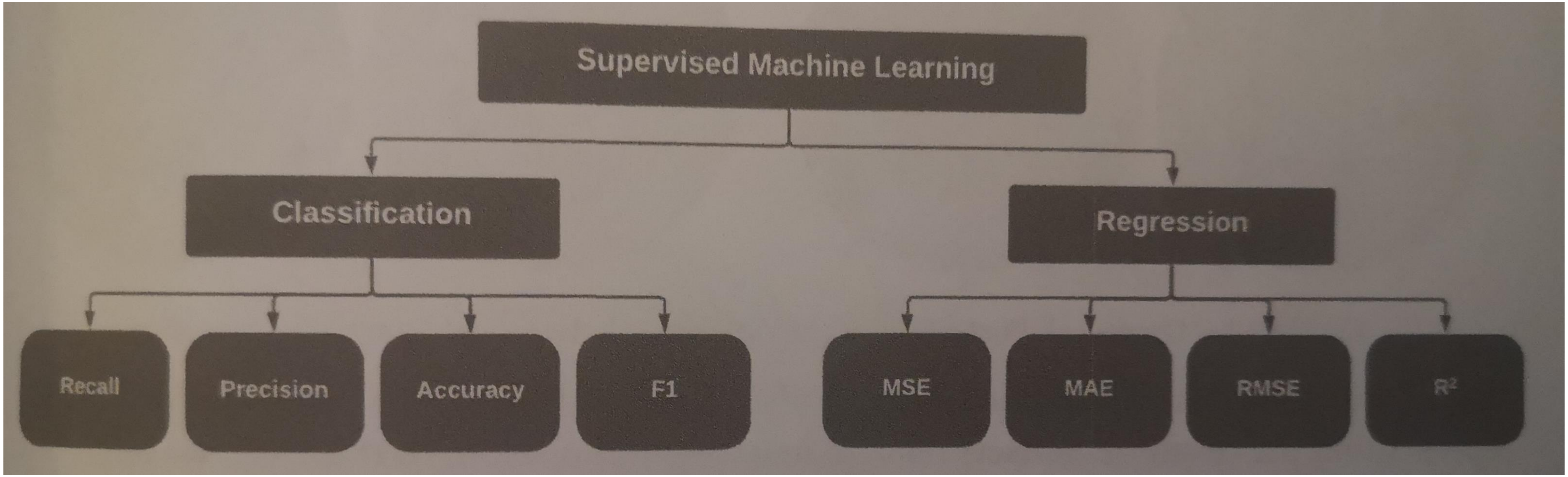
Good prognosis

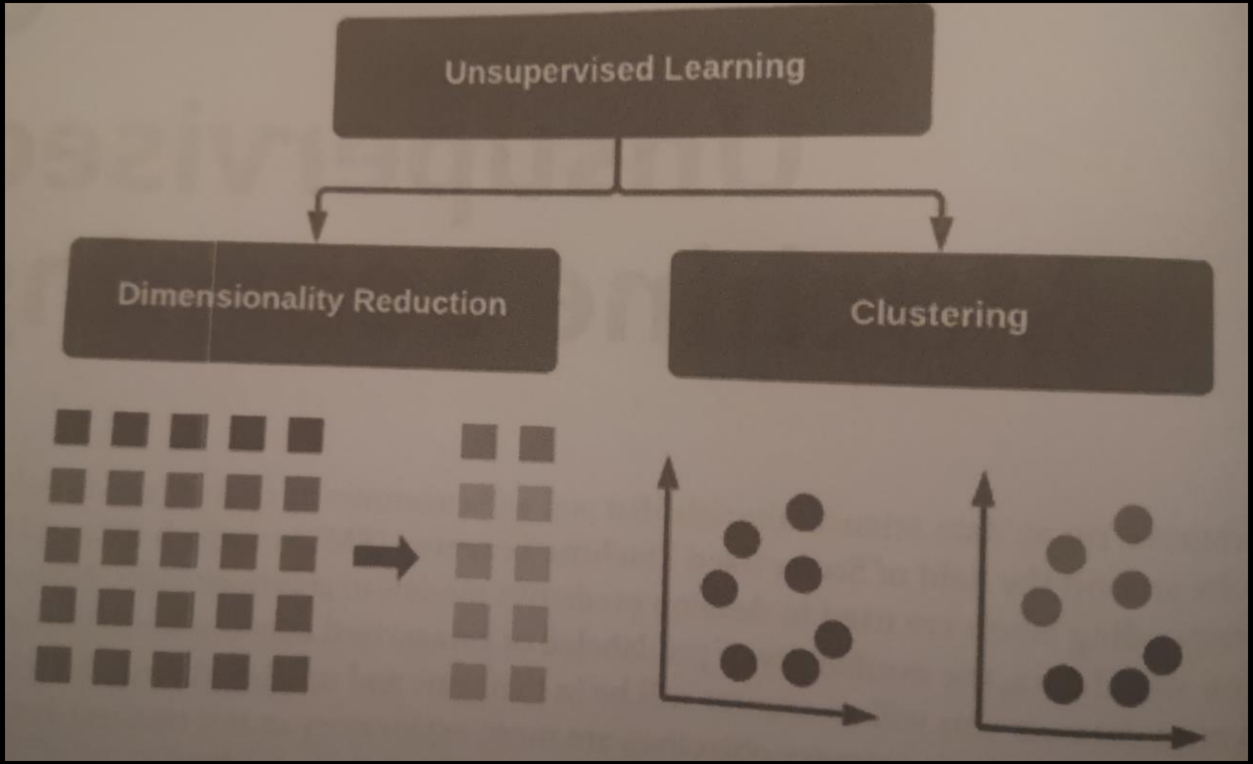
Bad prognosis

Is there work to do on van 't Veer et al. data ?

- What is the minimum number of genes required in these classification models (to avoid chance classification)
- What is the maximum number of genes (avoid overfitting)

Evaluation metrics for supervised learning

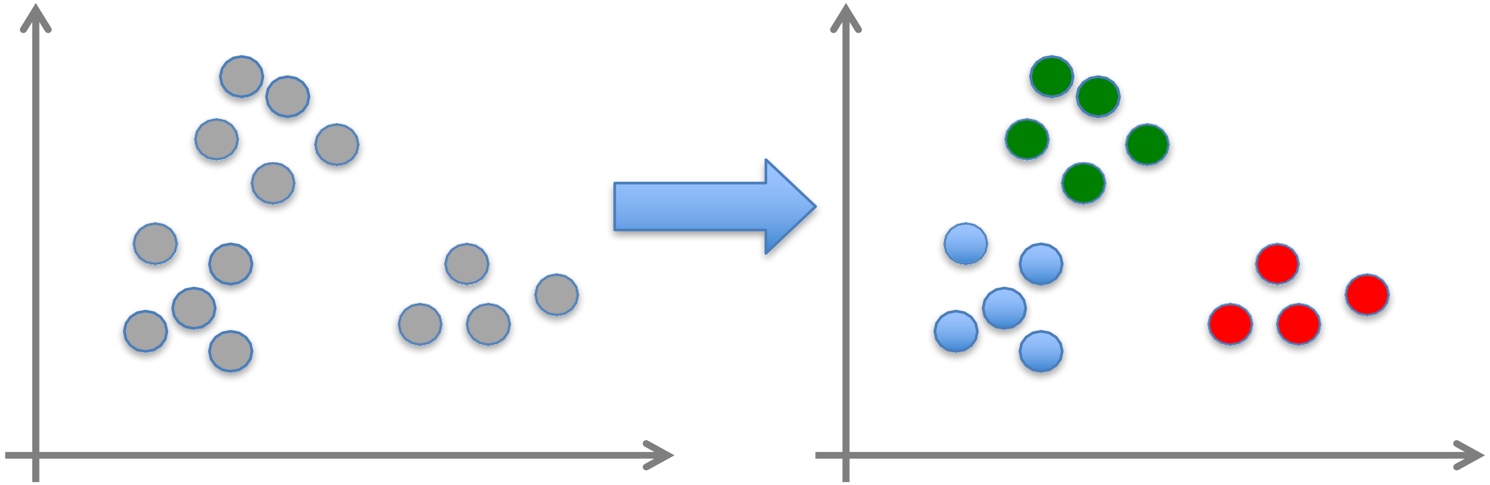




Unsupervised learning

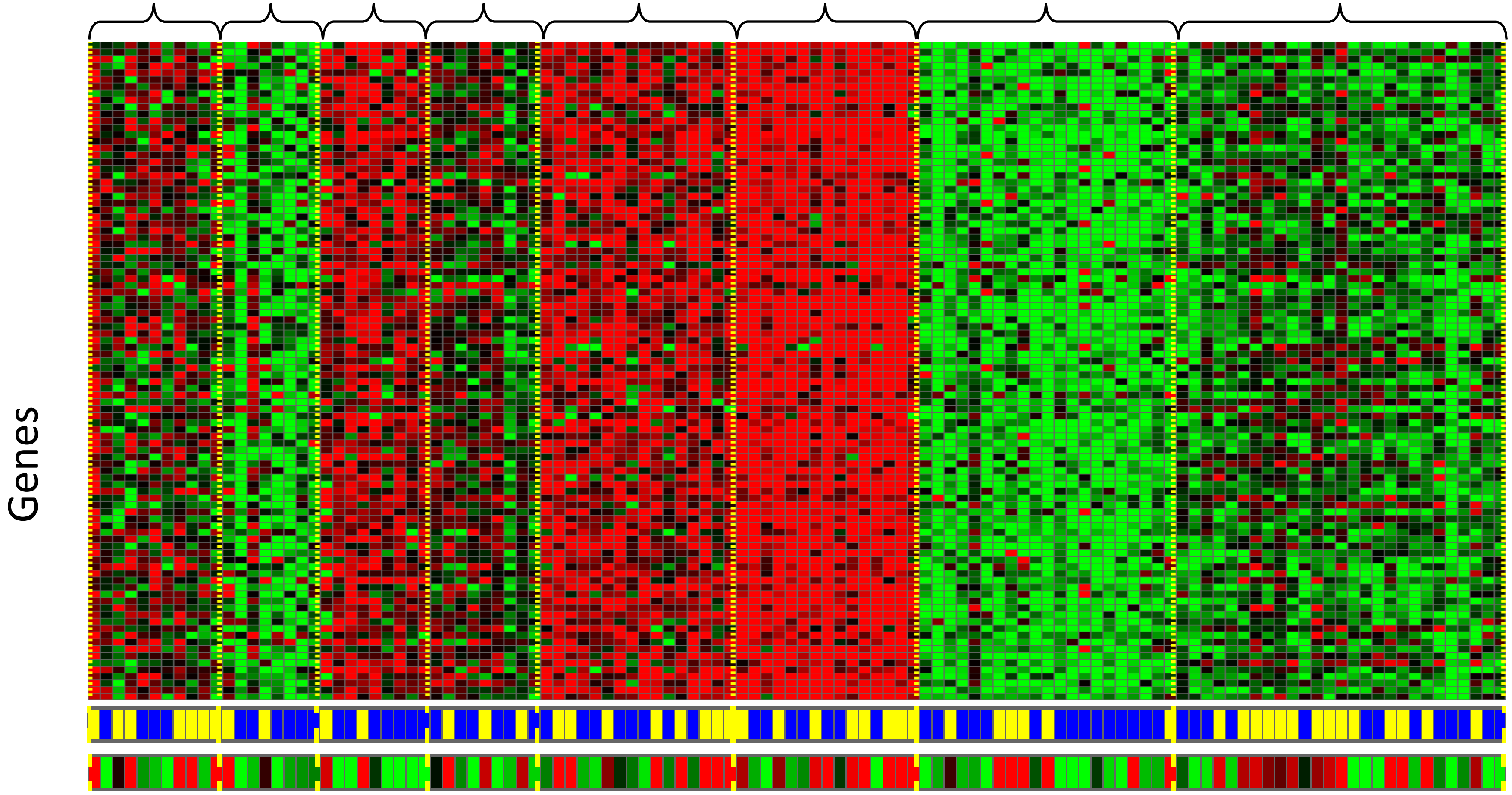
Unsupervised Learning

- The data have no target label
- We want to explore the data to capture some pattern
- Given x_1, x_2, ..., x_n (without labels)
- Output hidden structure behind the *x*'s
  - E.g., clustering



Unsupervised Learning

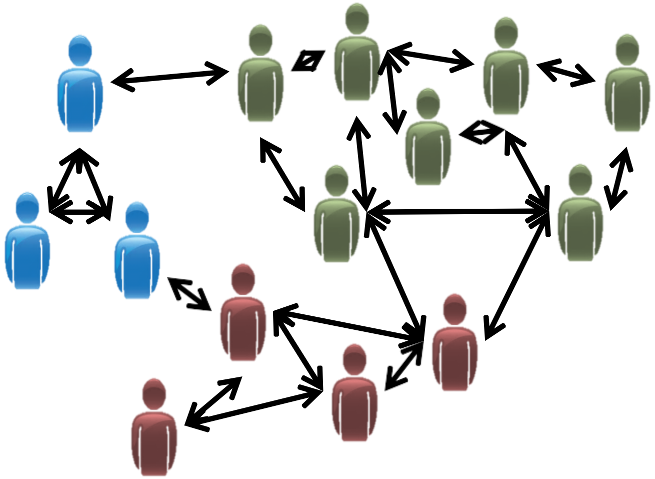
Genomics application: group individuals by genetic similarity



Individuals

Daphne Koller

Unsupervised Learning



Social network analysis

Market segmentation

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

Astronomical data analysis

Andrew Ng

Some applications of unsupervised learning in genetics and genomics

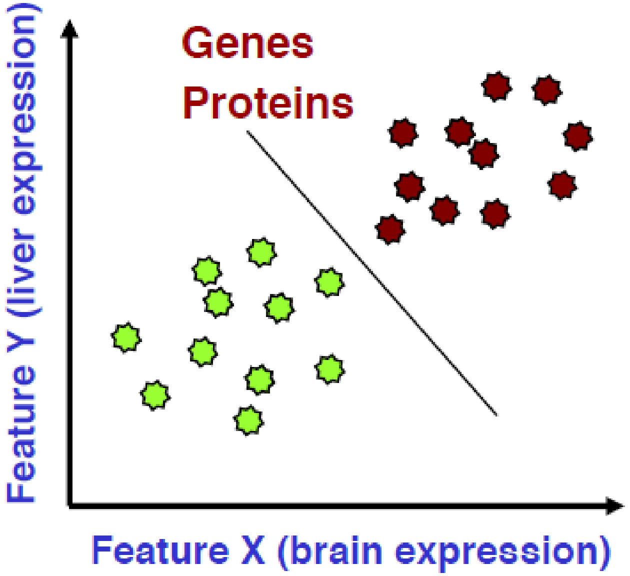
- binning of metagenomics contigs
- identification of plasmids and chromosomes
- clustering reads into chromosomes for better assembly
- clustering of reads as a preprocessor for assembly of reads

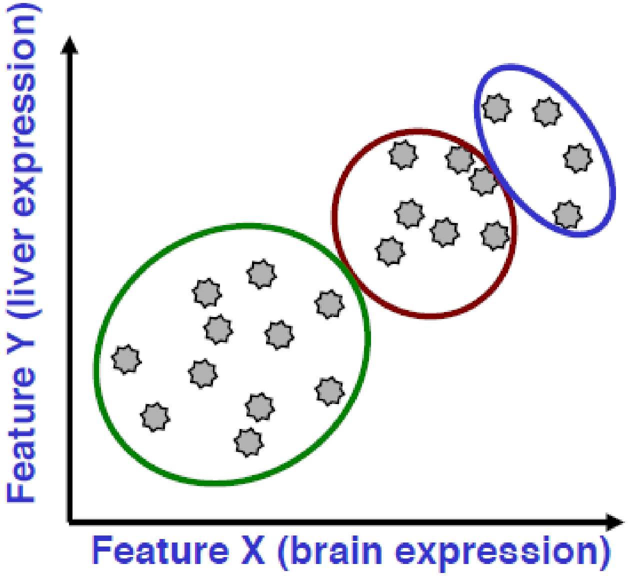
Unsupervised learning: Cluster Analysis

Recap: Classification vs. Clustering

Objects characterized by one or more features

- **Classification (supervised learning)**
  - Have labels for some points
  - Want a "rule" that will accurately assign labels to new points
  - Metric: Classification accuracy
- **Clustering (unsupervised learning)**
  - No labels
  - Group points into clusters based on how "near" they are to one another
  - Metric: independent validation features

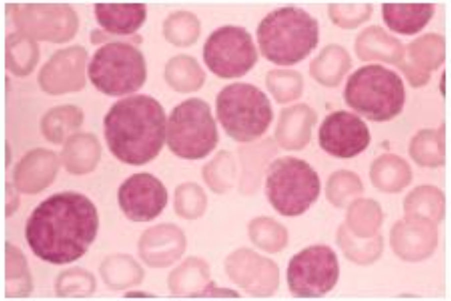




Why use unsupervised approach?

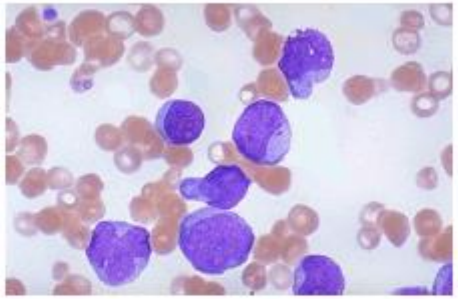
- Class discovery: Unsupervised approach
  - Clinical heterogeneity: Only 40% of patients respond to chemotherapy
    - Hypothesis: Reflects molecular heterogeneity in tumors
    - Approach: Use clustering to discover new classes (There may be classes we are unaware of)
- Class prediction: Supervised approach
  - No single biomarker is enough

Acute Lymphoblastic Leukemia (ALL)



Golub et al., 1999

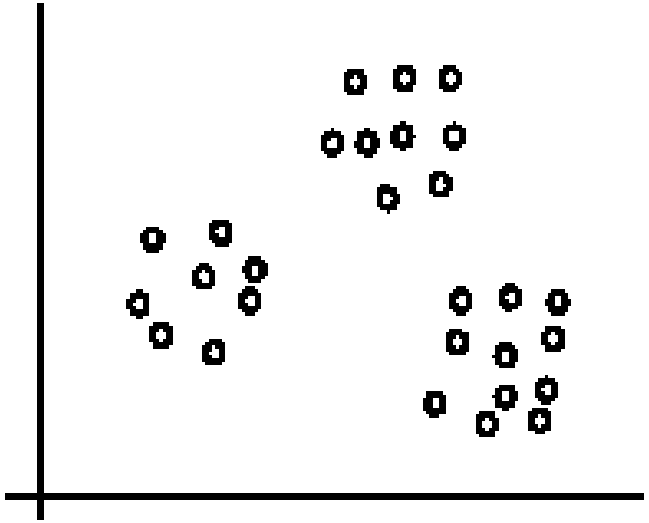
Acute Myelogenous Leukemia (AML)



Clustering process

- Clustering is a technique for finding similarity groups in data, called **clusters**
  - It groups data instances that are similar to (near) each other in one cluster and data instances that are very different (far away) from each other into different clusters

3 clusters



Aspects of clustering

- A clustering algorithm
  - Partitional clustering
    - Distance based (i.e., K-means)
    - Probabilistic
  - Hierarchical clustering
    - Agglomerative
    - Divisive
- A distance (similarity, or dissimilarity) function
- Clustering quality
- The quality of a clustering result depends on the algorithm, the distance function, and the application.

Examples to distance measures

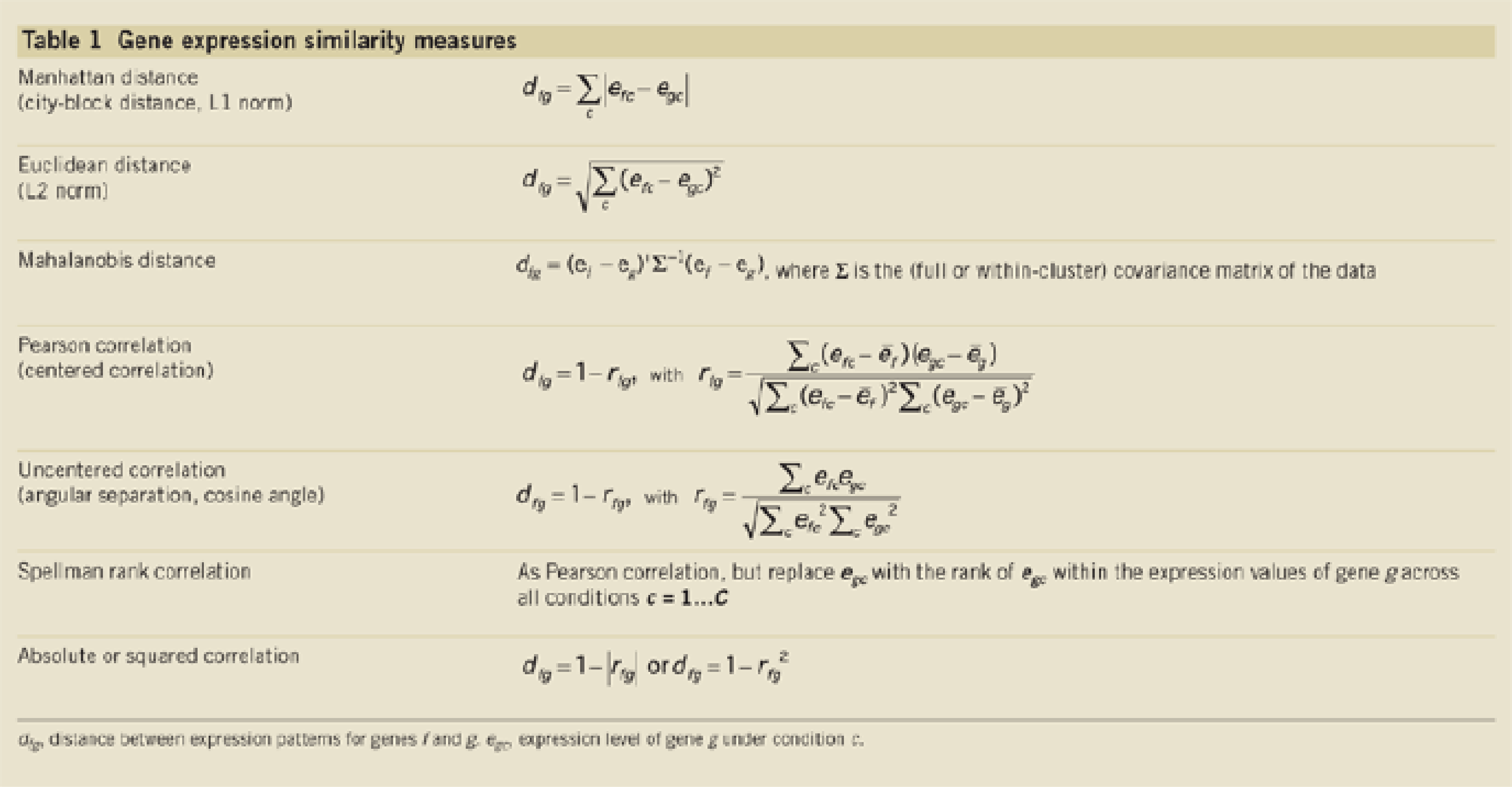


Table 1 Gene expression similarity measures

Manhattan distance (city-block distance, L1 norm)	$d_{fg} = \sum_c \lvert e_{fc} - e_{gc} \rvert$
Euclidean distance (L2 norm)	$d_{fg} = \sqrt{\sum_c (e_{fc} - e_{gc})^2}$
Mahalanobis distance	$d_{fg} = (\mathbf{e}_f - \mathbf{e}_g)' \Sigma^{-1} (\mathbf{e}_f - \mathbf{e}_g)$, where Σ is the (full or within-cluster) covariance matrix of the data
Pearson correlation (centered correlation)	$d_{fg} = 1 - r_{fg}$, with $r_{fg} = \frac{\sum_c (e_{fc} - \bar{e}_f)(e_{gc} - \bar{e}_g)}{\sqrt{\sum_c (e_{fc} - \bar{e}_f)^2 \sum_c (e_{gc} - \bar{e}_g)^2}}$
Uncentered correlation (angular separation, cosine angle)	$d_{fg} = 1 - r_{fg}$, with $r_{fg} = \frac{\sum_c e_{fc} e_{gc}}{\sqrt{\sum_c e_{fc}^2 \sum_c e_{gc}^2}}$
Spellman rank correlation	As Pearson correlation, but replace e_{gc} with the rank of e_{gc} within the expression values of gene g across all conditions $c = 1 \ldots C$
Absolute or squared correlation	$d_{fg} = 1 - \lvert r_{fg} \rvert$ or $d_{fg} = 1 - r_{fg}^2$

d_{fg}, distance between expression patterns for genes f and g. e_{gc}, expression level of gene g under condition c.

http://www.nature.com/nbt/journal/v23/n12/pdf/nbt1205-1499.pdf

Unsupervised learning: K-means clustering

- K-means is a partitional clustering algorithm
- The *k*-means algorithm partitions the given data into *k* clusters
  - Each cluster has a cluster **center**, called **centroid**.
  - *k* is specified by the user

K-means algorithm

- Given *k*, the *k-means* algorithm works as follows:
  1. Randomly choose *k* data points (seeds) to be the initial centroids, cluster centers
  2. Assign each data point to the closest centroid
  3. Re-compute the centroids using the current cluster memberships.
  4. If a convergence criterion is not met, go to 2).

How do we select k?

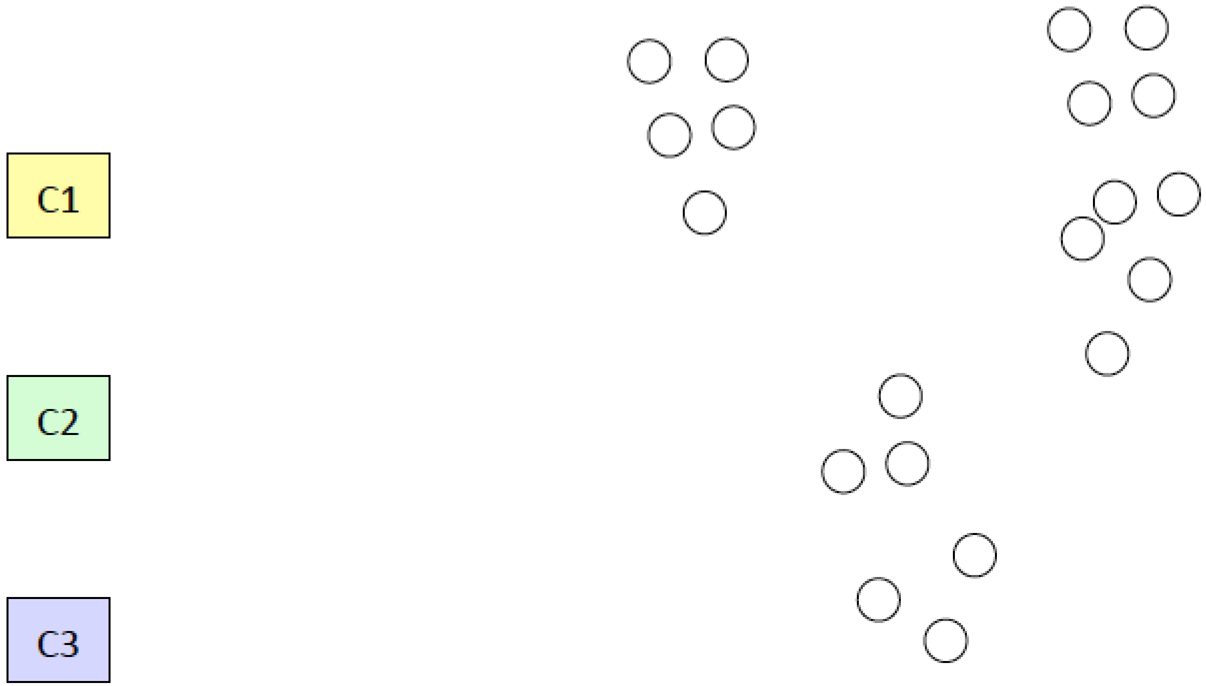
Problem: we can always make clusters more compact if we increase their number (in the extreme case, number of clusters = number of samples

Guess: Educated guess or try out and see what we like

Robustness: e.g. remove samples or add noise to measurements at random and see how resilient to change

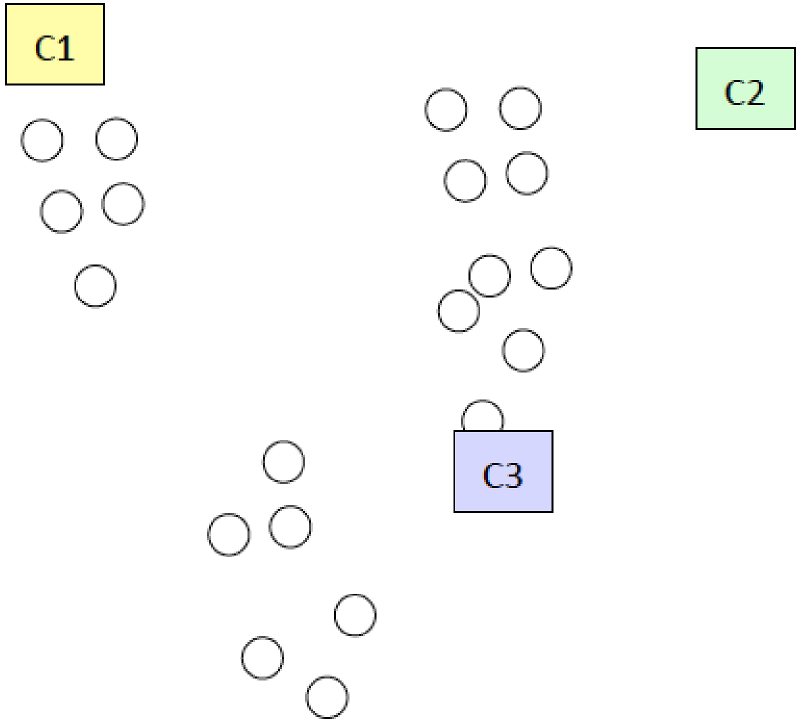
K-means clustering

Define *K (how?)*



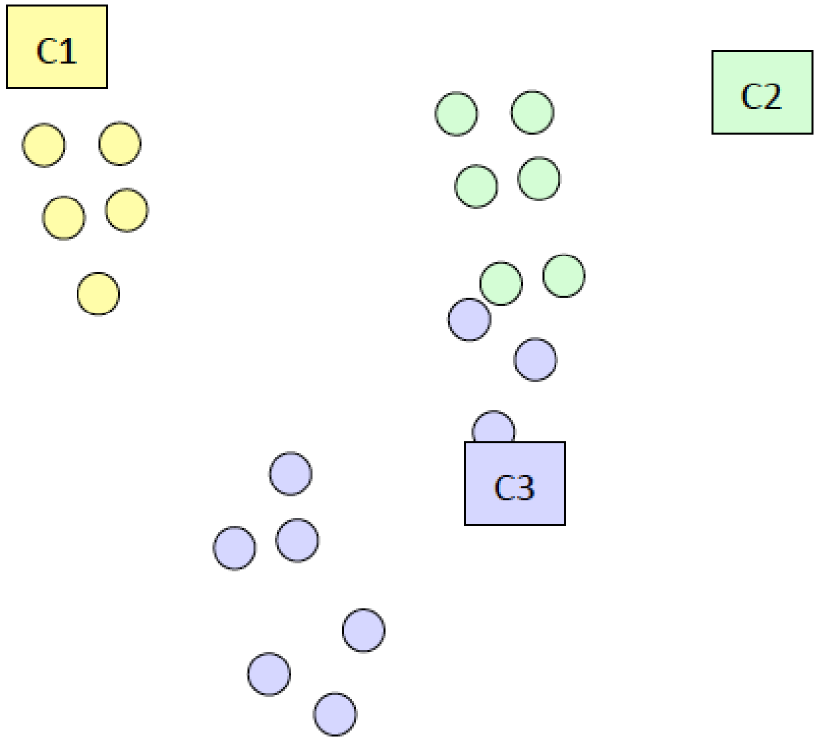
K-means clustering

Randomly initialize clusters
(*K*=3)



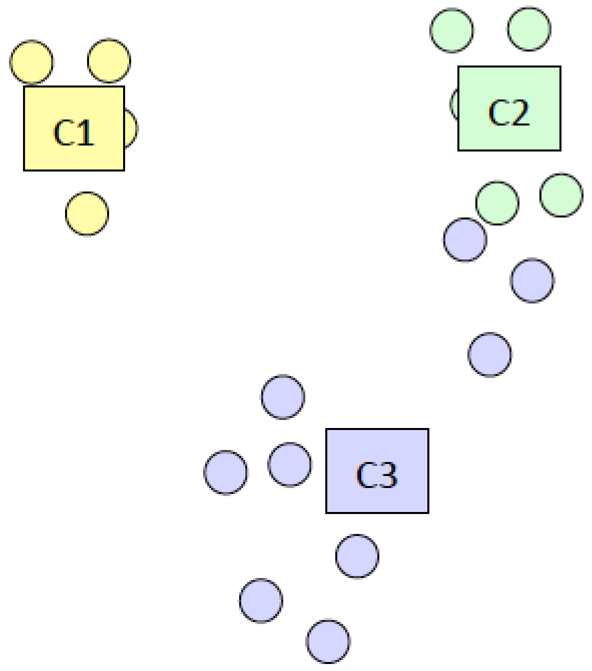
K-means clustering

Assign data points to nearest clusters



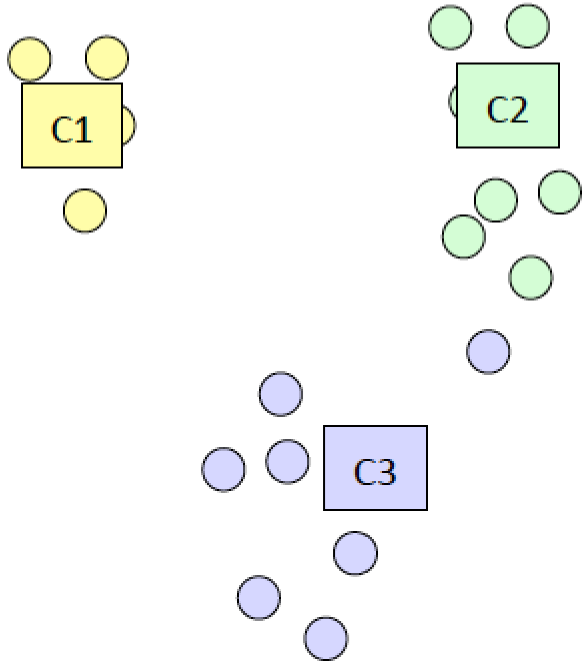
K-means clustering

Re-calculate clusters, as the centroids of current samples



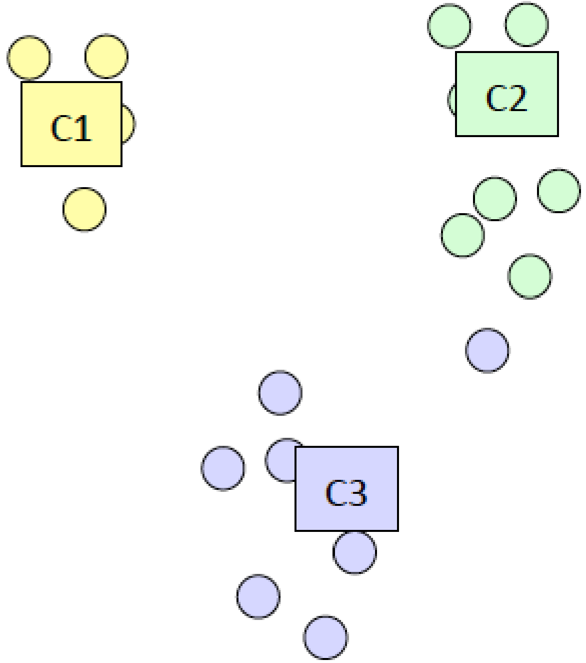
K-means clustering

Repeat (1): assign to clusters



K-means clustering

Repeat (1): re-calculate clusters, as the centroids of current samples

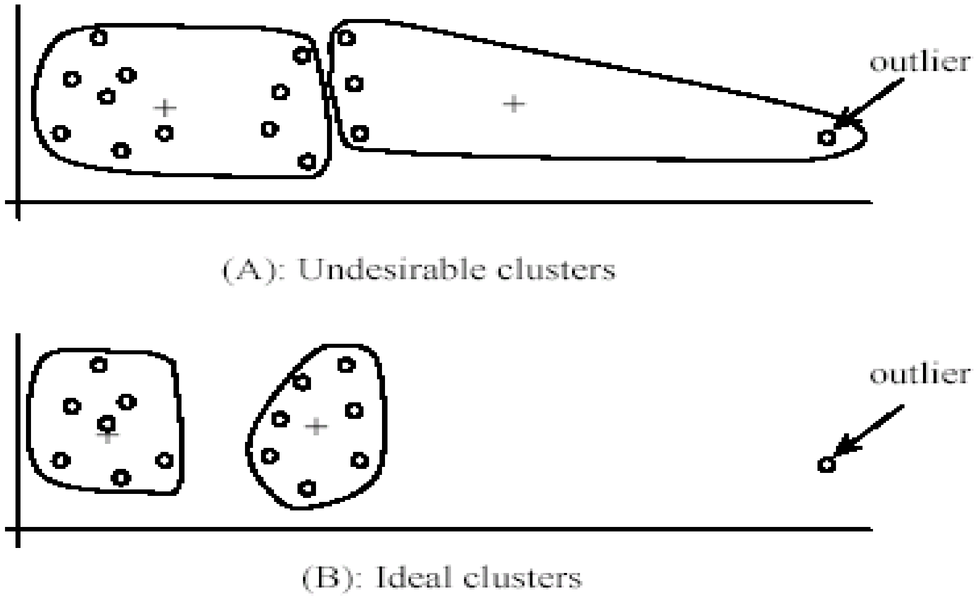


Strengths of k-means

- Strengths:
  - Simple: easy to understand and to implement
  - Works well with high dimensional data
  - Efficient: Time complexity: *O*(*tkn*),
    where *n* is the number of data points,
    k is the number of clusters, and
    t is the number of iterations
  - Since both *k* and *t* are small. *k*-means is considered a linear algorithm

Weaknesses of k-means

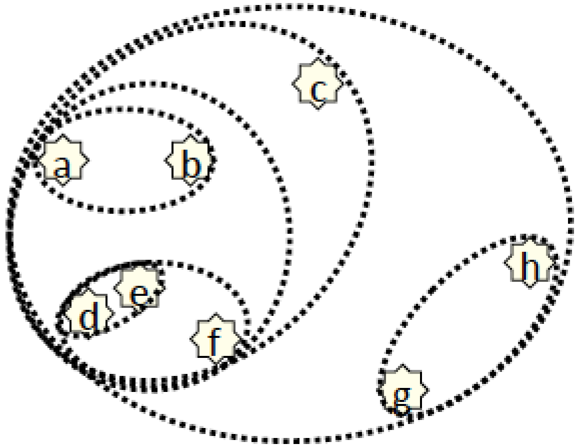
- The algorithm is only applicable if the mean is defined.
  - For categorical data, *k*-mode - the centroid is represented by most frequent values.
- The user needs to specify *k*.
- The algorithm is sensitive to **outliers**
  - Outliers are data points that are very far away from other data points.
  - Outliers could be errors in the data recording or some special data points with very different values.

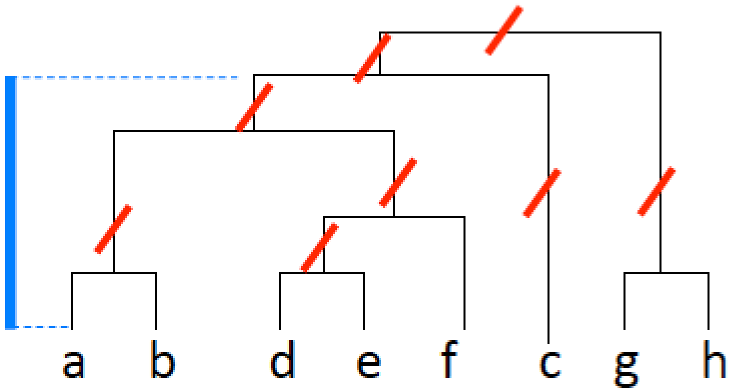


(A): Undesirable clusters

(B): Ideal clusters

Unsupervised learning: Hierarchical clustering





In the dendrogram the order of the leaves within a cluster is random; this can have a big visual impact

Hierarchical clustering

Start with each point in a separate cluster (“leaves”)

At each step:

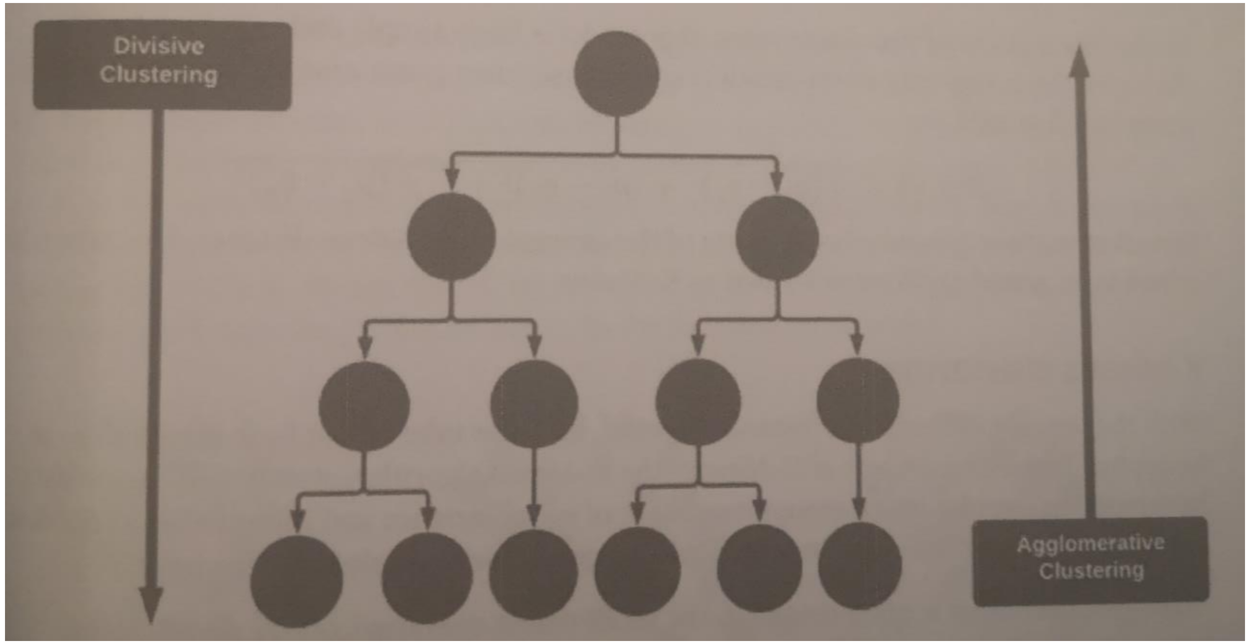
- Choose a pair of closest clusters
- Merge

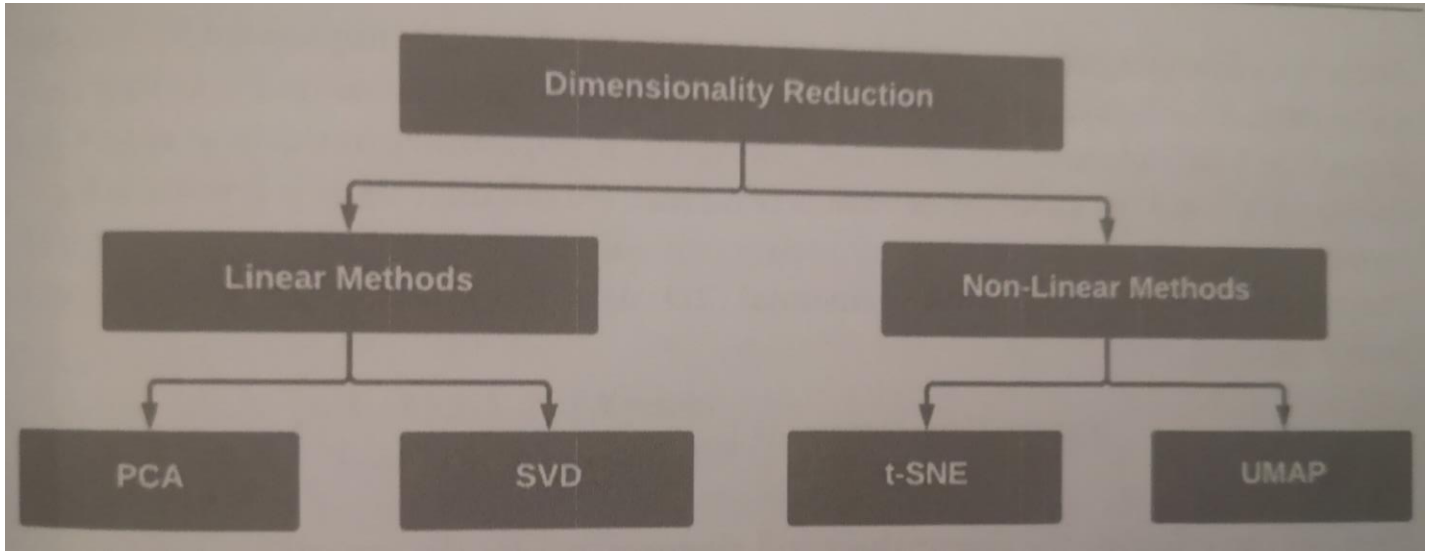
Repeat until only one cluster remains, with all samples (“root”)

Hierarchical clustering

Choose a pair of closest clusters: based on some distance measure (e.g. Euclidean, correlation)

Merge: re-calculate distance between new cluster and other clusters



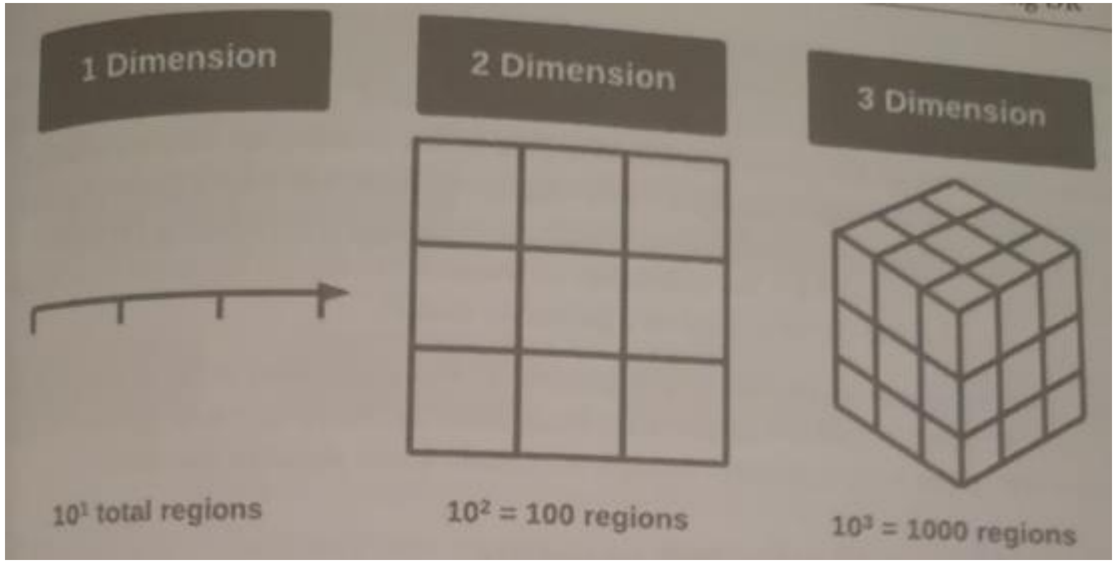


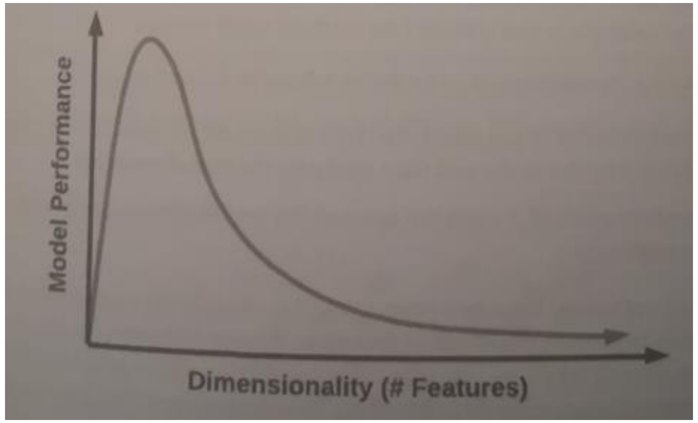
Unsupervised learning: Dimensionality Reduction

Dimensionality reduction

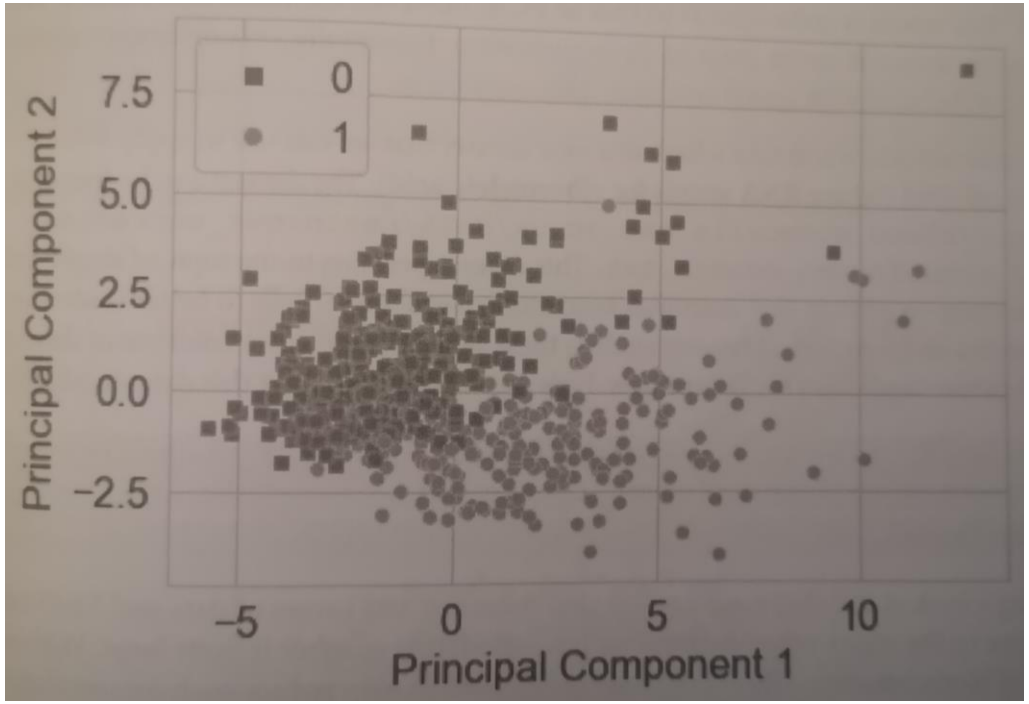
Uses:

- Data Visualization
- Data Reduction
- Data Classification
- Trend Analysis
- Making the process faster and less computationally intensive
- Noise Reduction

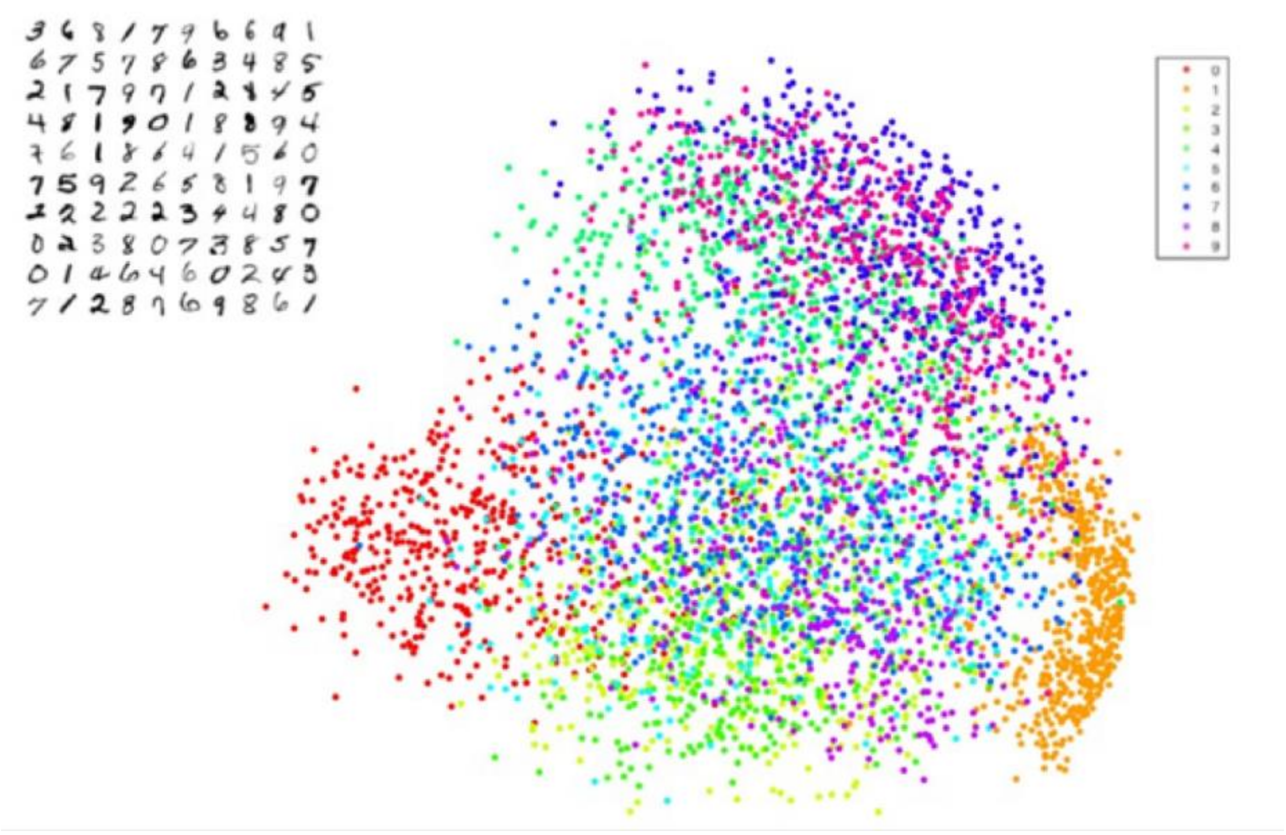




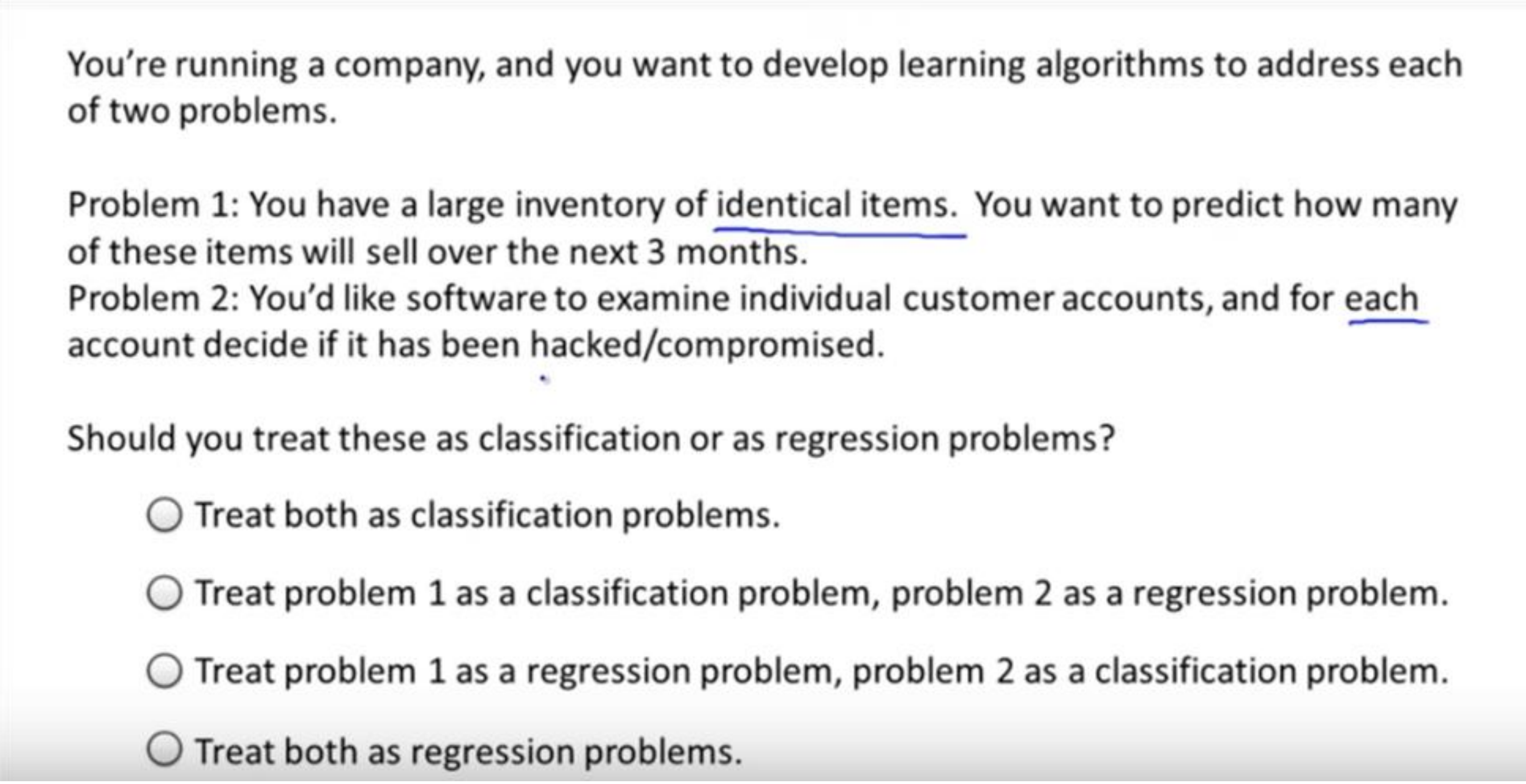
Unsupervised learning: principal component analysis



PCA of MNIST digits



What kind of problem is it?



You're running a company, and you want to develop learning algorithms to address each of two problems.

Problem 1: You have a large inventory of identical items. You want to predict how many of these items will sell over the next 3 months.
Problem 2: You'd like software to examine individual customer accounts, and for each account decide if it has been hacked/compromised.

Should you treat these as classification or as regression problems?

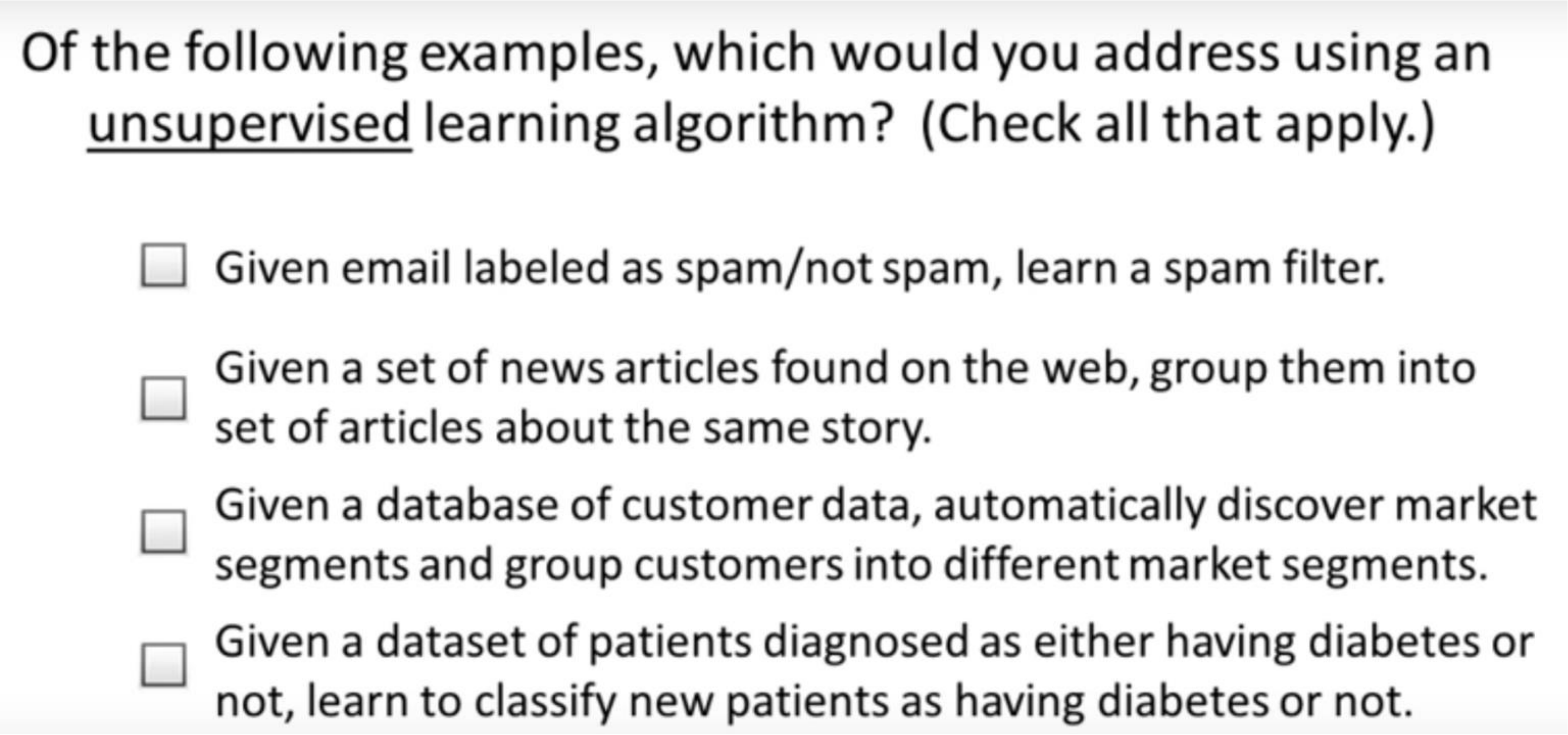
○ Treat both as classification problems.

○ Treat problem 1 as a classification problem, problem 2 as a regression problem.

○ Treat problem 1 as a regression problem, problem 2 as a classification problem.

○ Treat both as regression problems.

What kind of problem is it?



Of the following examples, which would you address using an unsupervised learning algorithm? (Check all that apply.)

□ Given email labeled as spam/not spam, learn a spam filter.

□ Given a set of news articles found on the web, group them into set of articles about the same story.

□ Given a database of customer data, automatically discover market segments and group customers into different market segments.

□ Given a dataset of patients diagnosed as either having diabetes or not, learn to classify new patients as having diabetes or not.