Machine Learning for Computational Biology

Ameet Soni

September 29, 2018

NSF Workshop, Carleton College

What is Machine learning?

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

- The study of algorithms that improve performance with experience
  →predict the future based on the past

- Replace “human writing code” with “human supplying data”

Why machine learning?

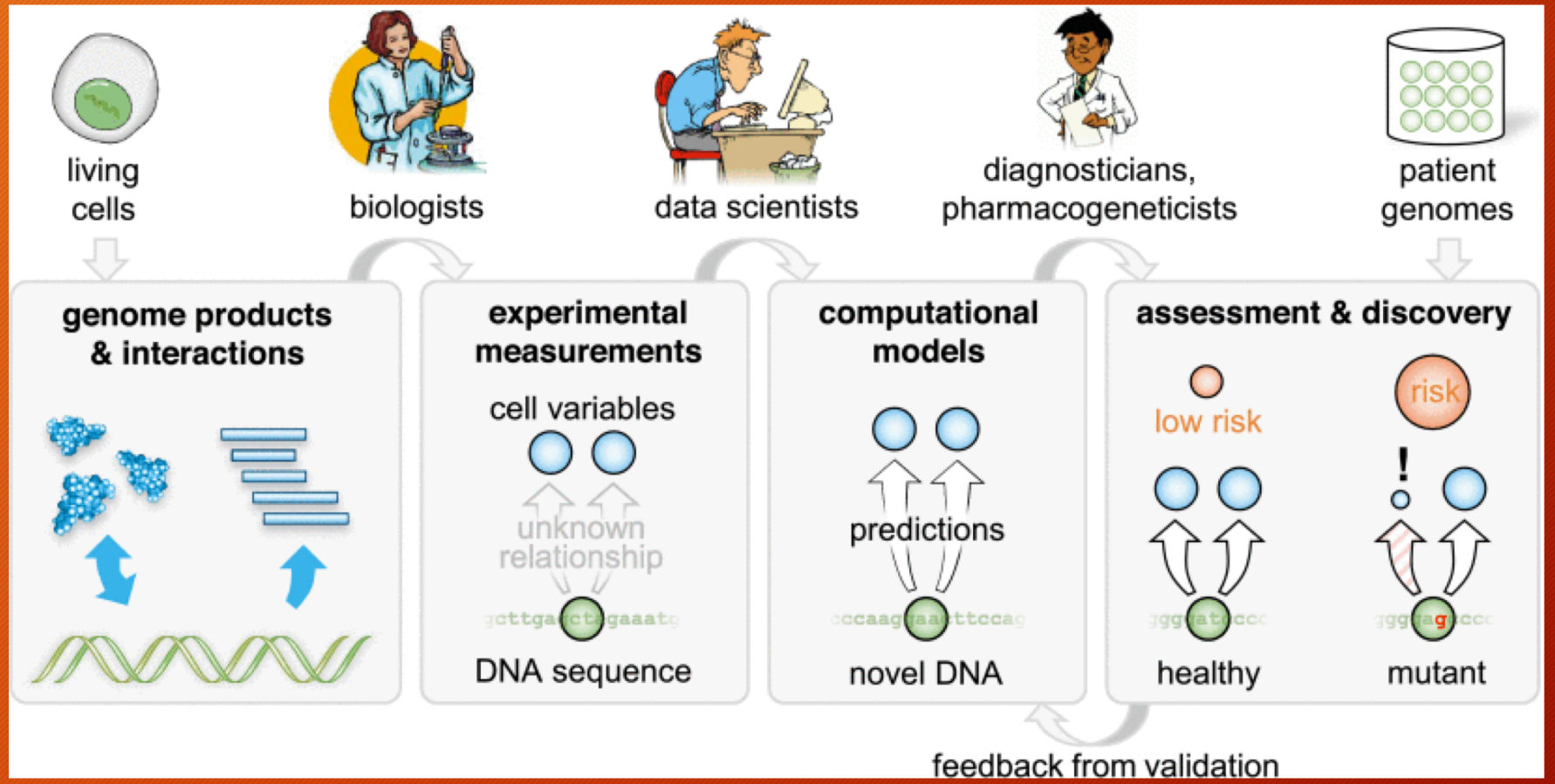
Many systems are too difficult/expensive because of:

- Lack of human expertise (e.g., drug binding prediction)
- Complexity of expertise (e.g., natural language)
- Need for individually-tailored solution (e.g., email filter)
- Dynamic nature (e.g., network intrusion, stock market, product stocking)

Why Computational Biology?

- Genomic era:
  - Data characteristics: large data set of sequences; very little noise
  - Solutions: efficient algorithms (e.g., dynamic programming) and mathematical models
- Post-genomic era data:
  - Measure system dynamics
  - Data characteristics: noisy, indirect, complex interactions
  - Solution: statistical/probabilistic methods

Comp Bio Pipeline

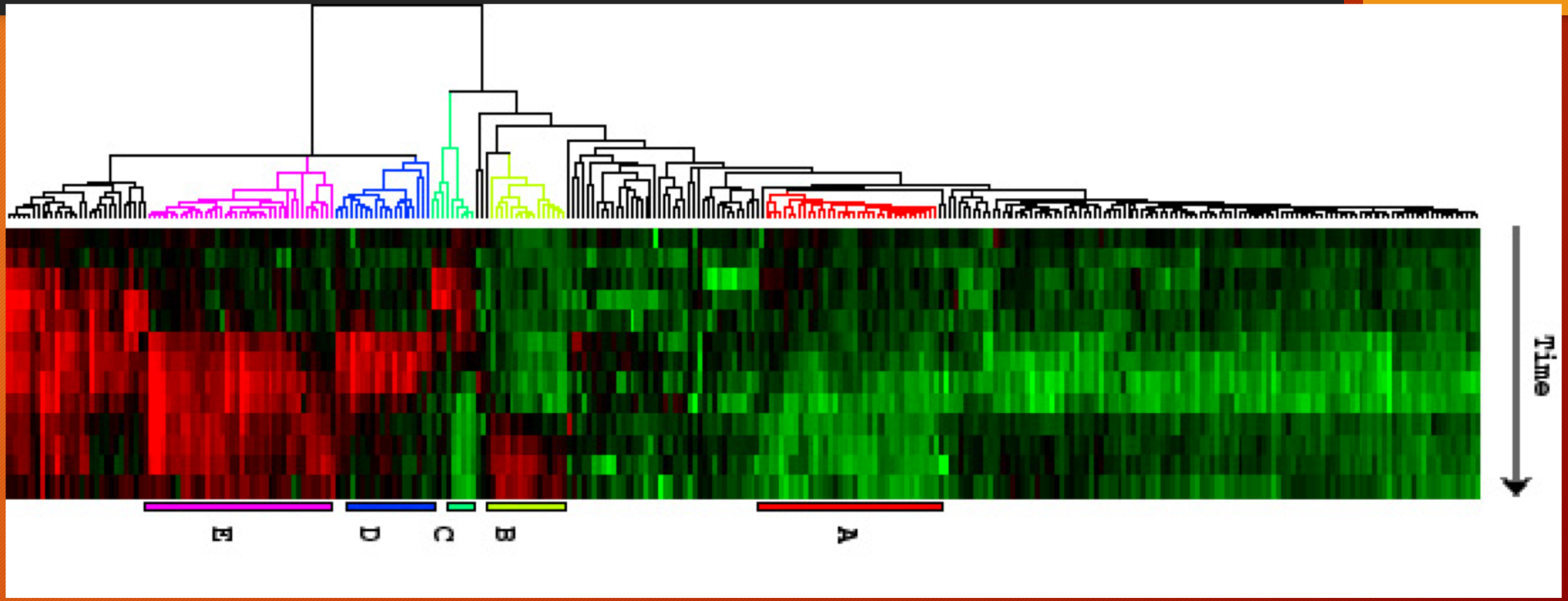


Leung et al., *IEEE*, 2016

Machine Learning Frameworks

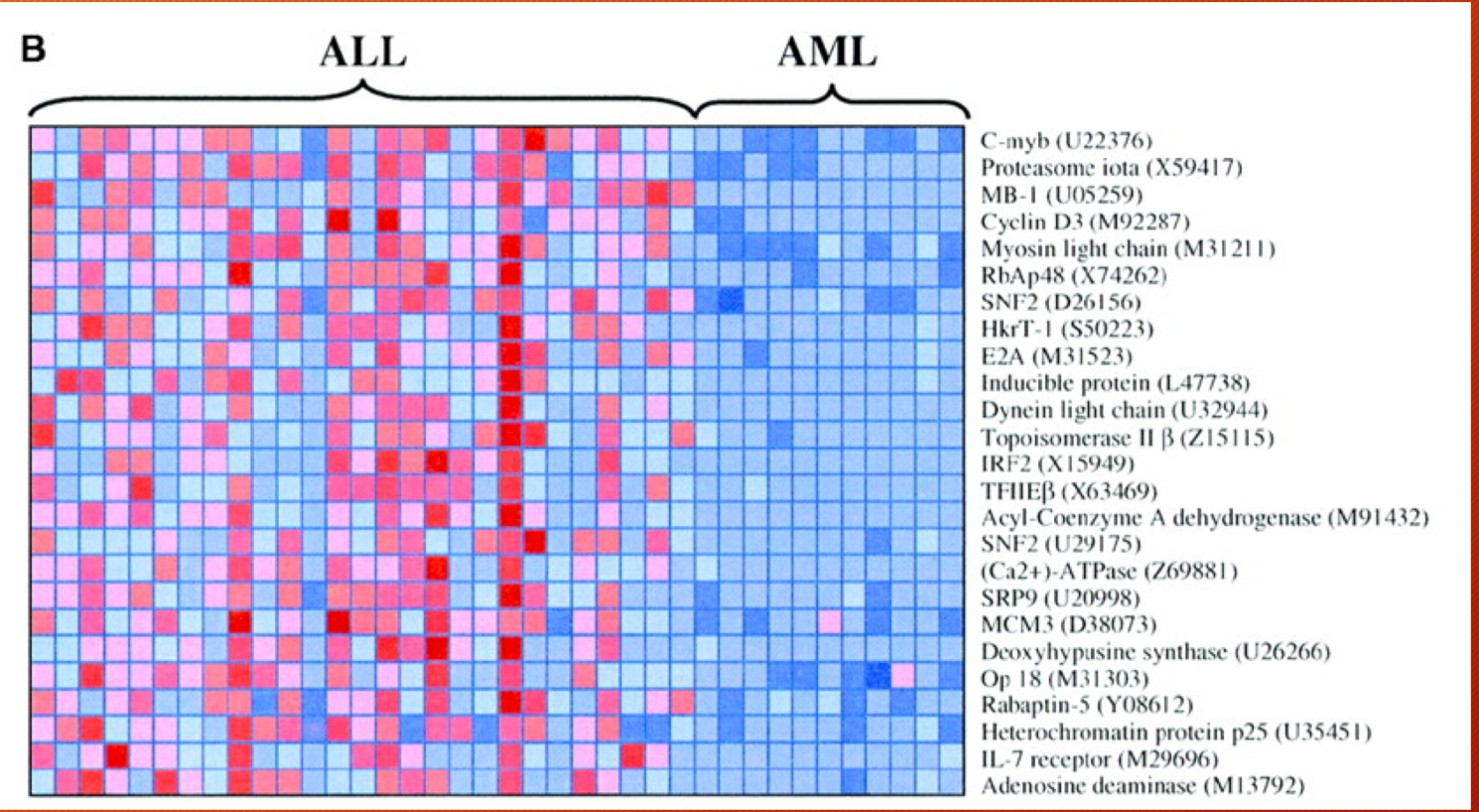
- Supervised learning: learn from answers (prediction)
- Unsupervised learning: learn without answers (clustering)
- Other:
  - Semi-supervised
  - Time-series
  - Structured prediction
  - Active learning

Unsupervised: gene function



Gillet et al., *PNAS* 2011

Supervised: Disease Diagnosis



Golub et al., *Science*, 1999

Sequence: Gene Finding

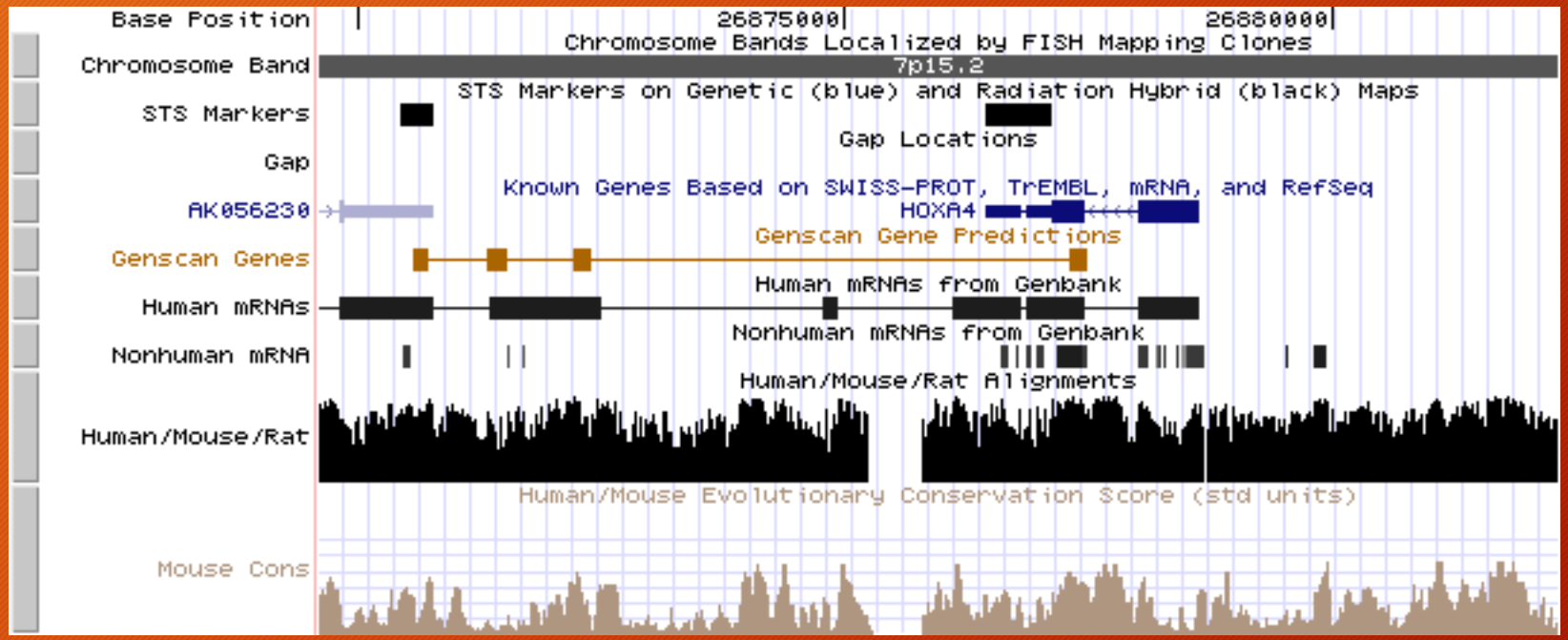
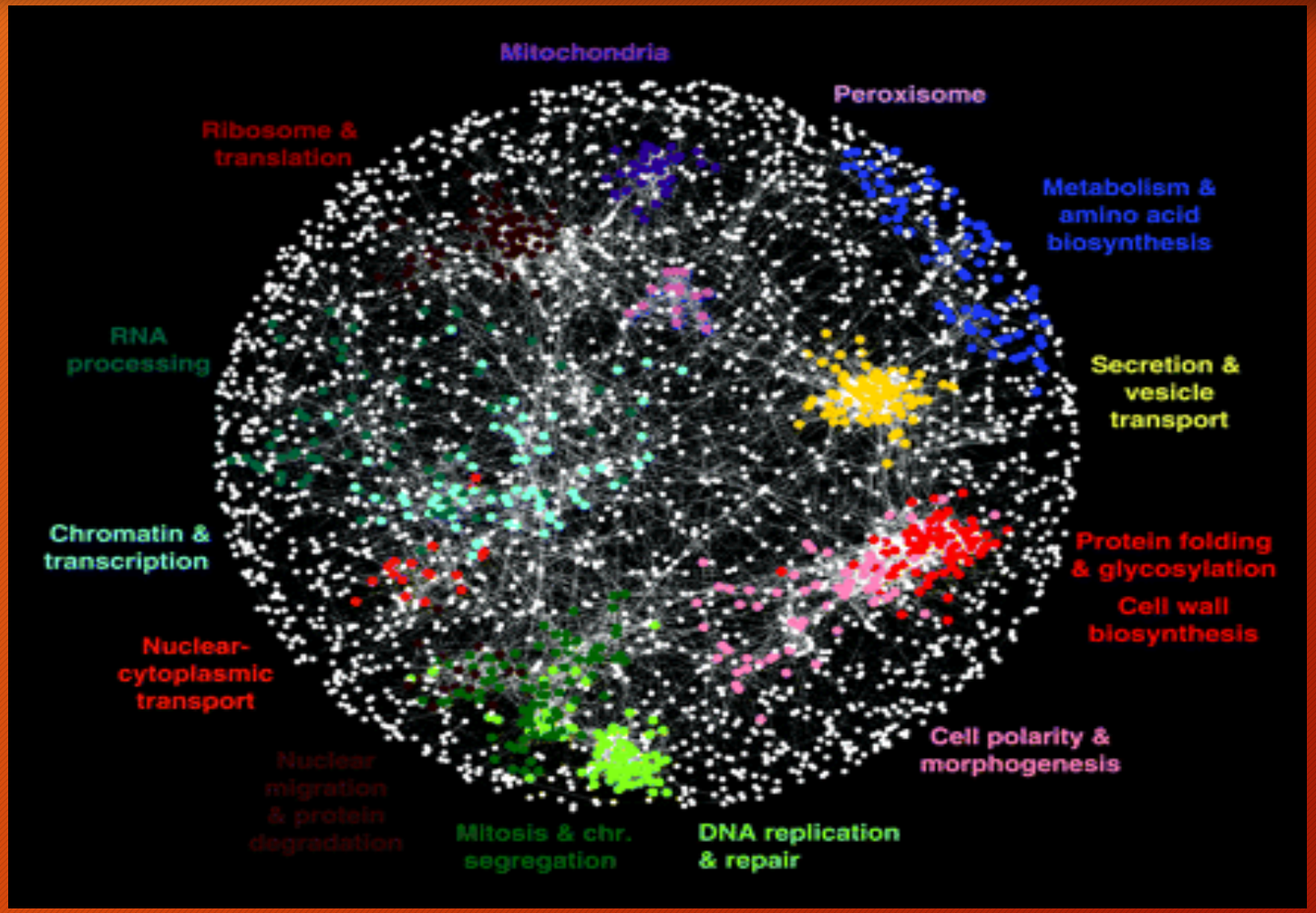


image from the UCSC Genome Browserhttp://genome.ucsc.edu/

Structure: Gene Interaction Network



Costanzo et al., *Biotechnology* 2011

Learning framework

- **Given**: observed properties of an object
- **Do:** make informed guess about unobserved property of object

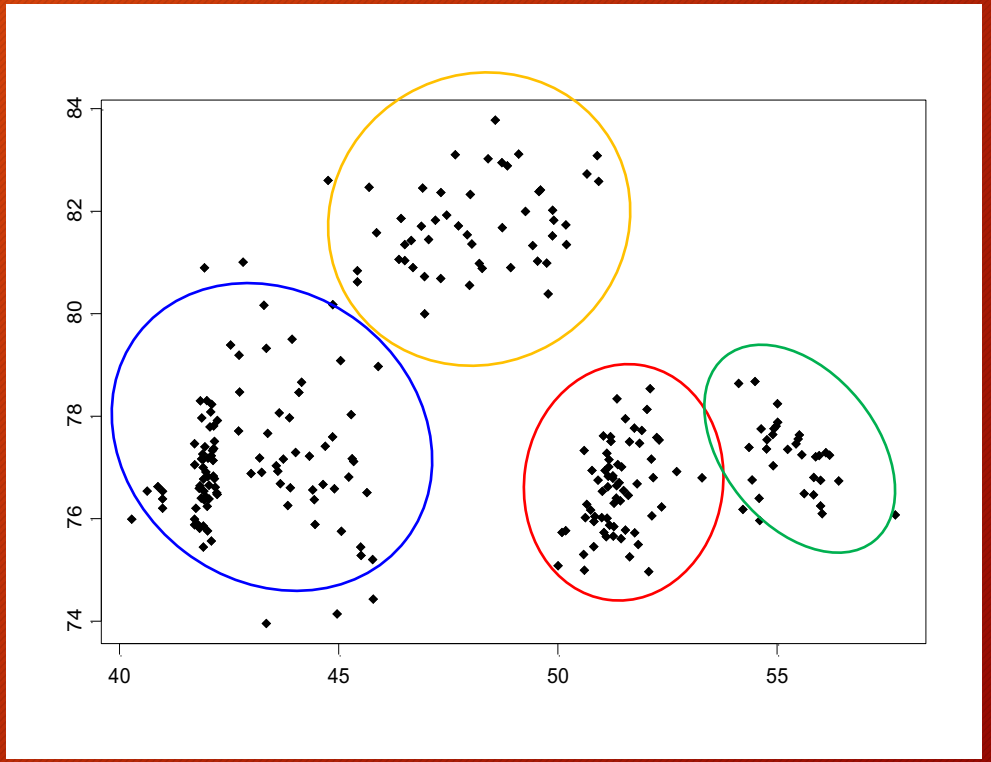
- Assumption: all objects have the same number of descriptions (features)

What is clustering?

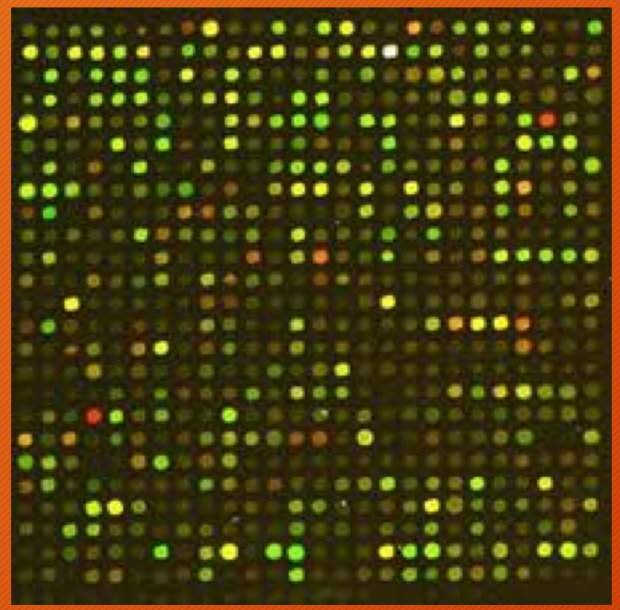
- *Clustering* - procedure that detects the presence of distinct groups

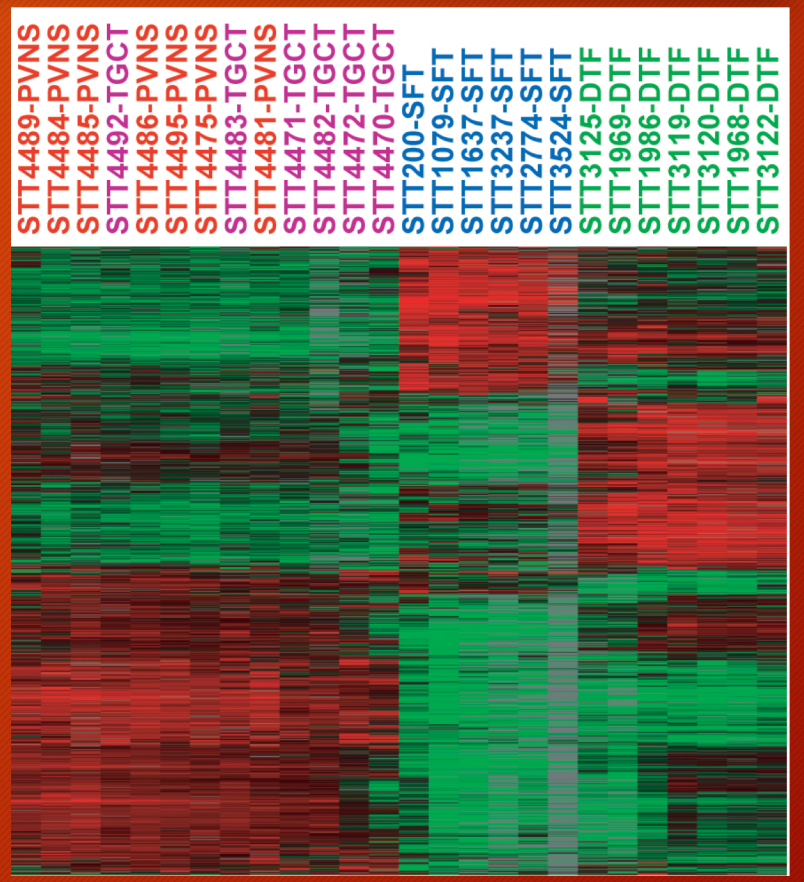
  A form of unsupervised learning

- Uses:
  - Visualization
  - Data exploration

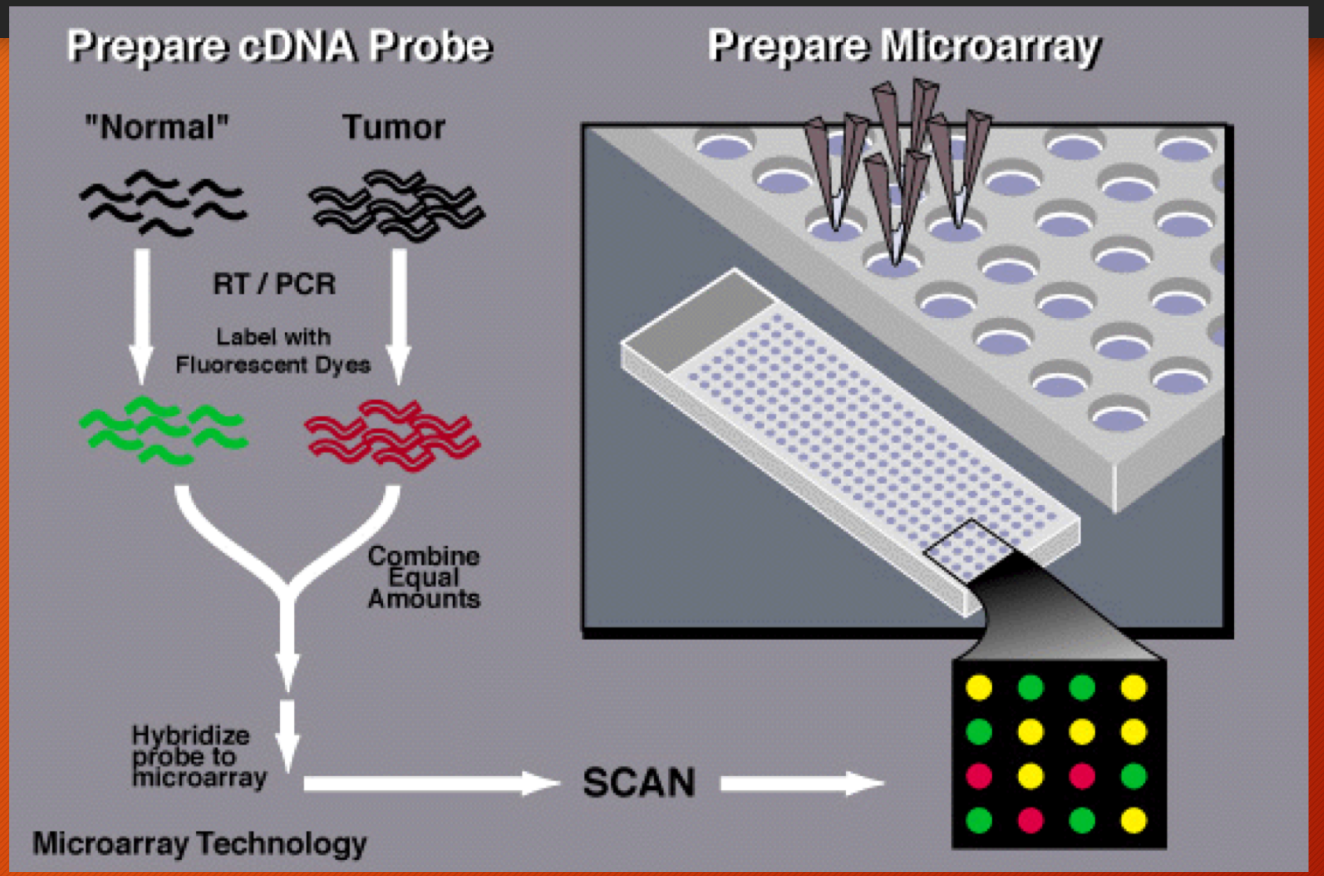


Application: gene expression





Spotted cDNA Microarrays

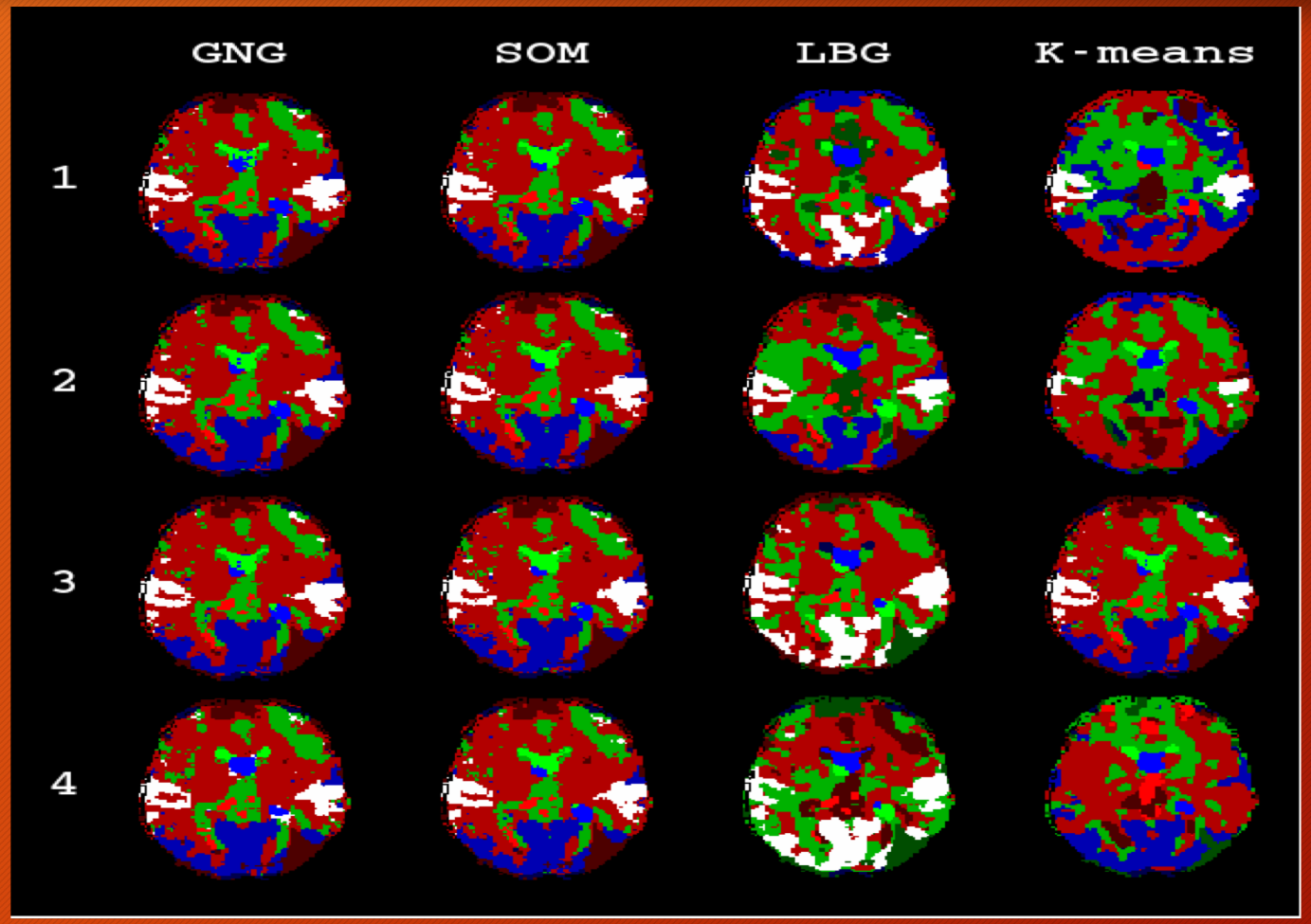


Also look at this animation: Microarray Animation

Microarray Data – Intensity matrix

Time:	Time X	Time Y	Time Z
Gene 1	10	8	10
Gene 2	10	0	9
Gene 3	4	8.6	3
Gene 4	7	8	3
Gene 5	1	2	3

Application: Medical Imaging Segmentation

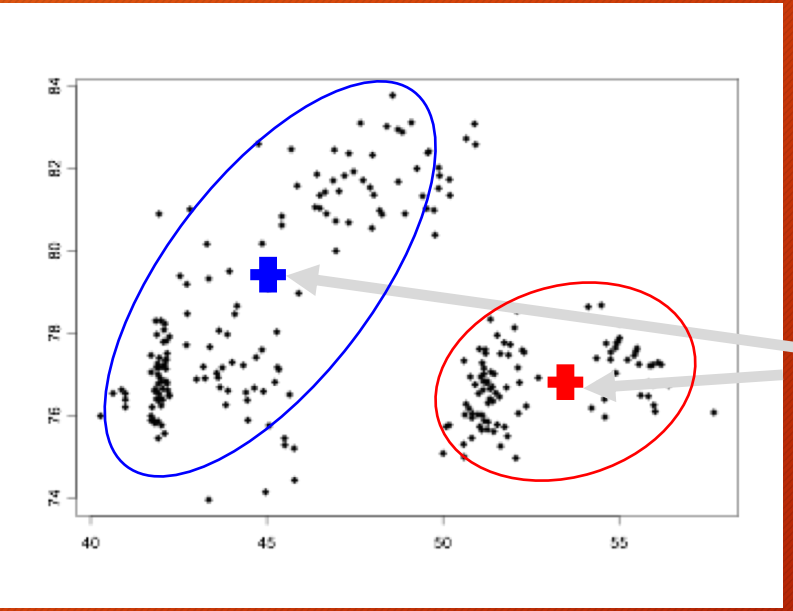


Other Biological Applications

- Population genetics
- Tree of life (phylogenetics)
- Gene function analysis
- Protein structure/function

K-Means Clustering

- K-means clustering: partition examples into k groups such that each example joins the group with the closest mean value



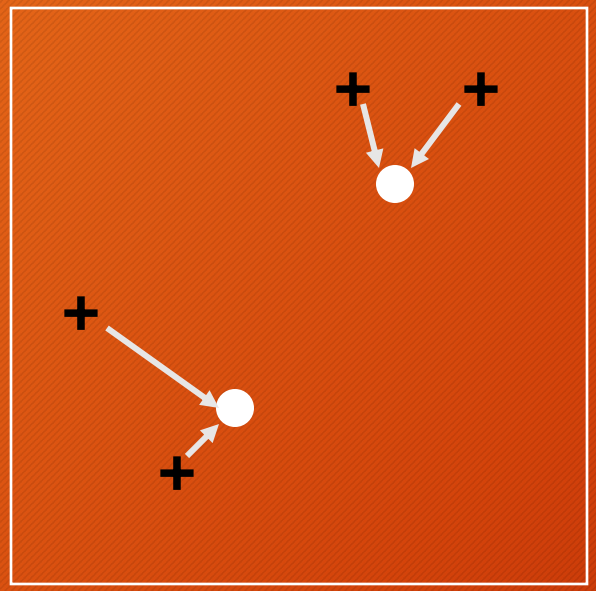
Problem: there are 2^N possible clusters with just $k=2$!

Lloyd's Algorithm

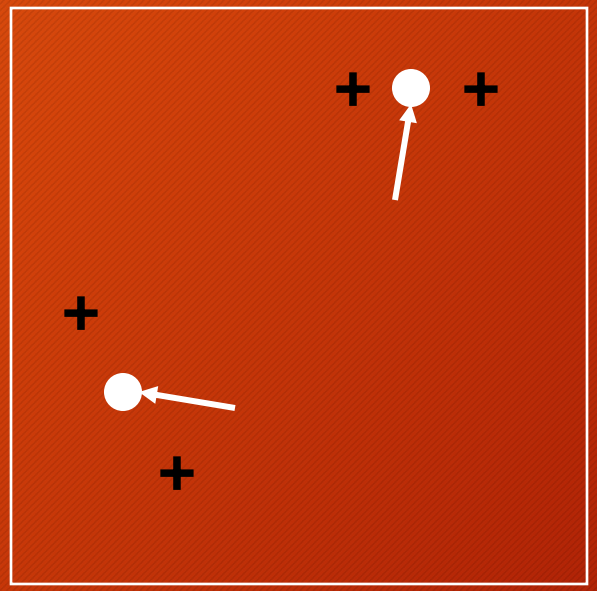
- Pick *k* initial cluster centers
- Put each data point in nearest cluster center
- Re-estimate cluster center
- Repeat

Efficient, but approximate solution

K-means clustering



assignment



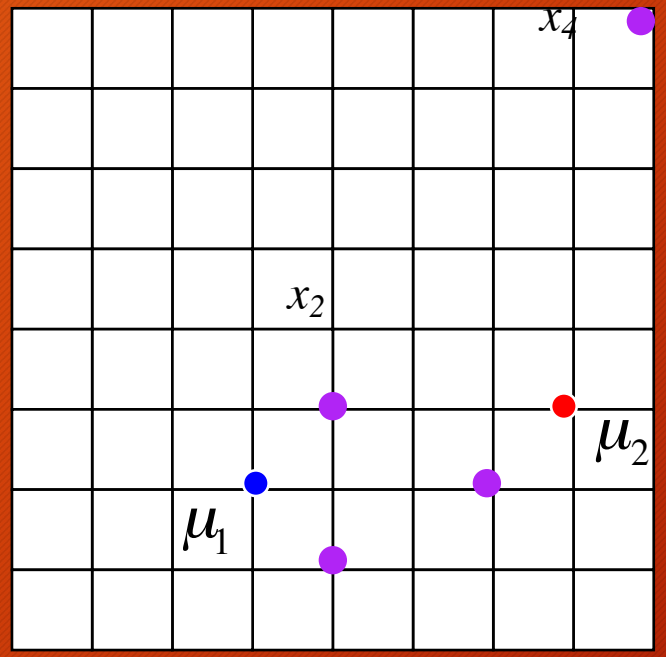
re-computation of means

K-means clustering example

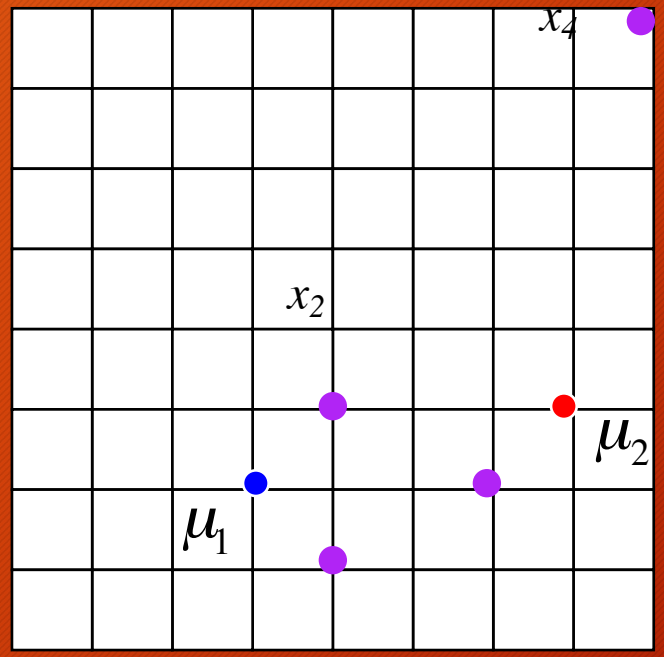
Given the following 4 profiles and 2 clusters initialized as shown.
Assume the distance function is

$$\text{dist}(x_i, x_j) = \sum_e \left| x_{i,e} - x_{j,e} \right|$$

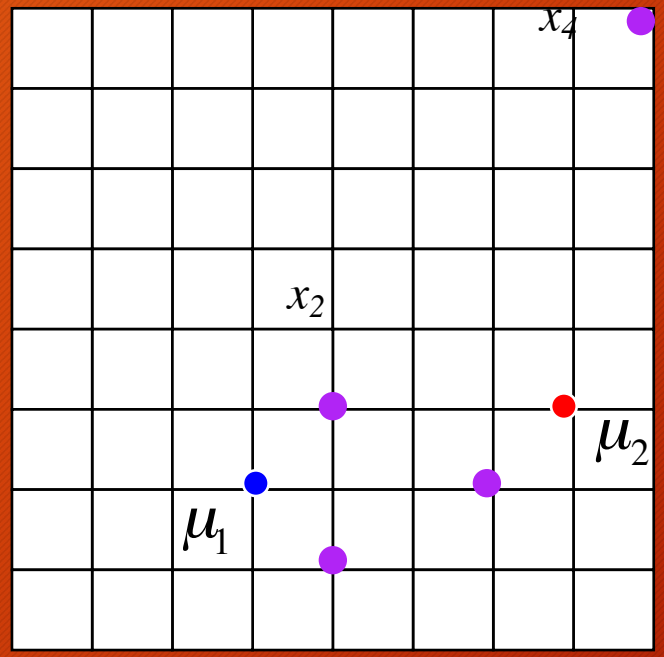
Exercise: find the new cluster means



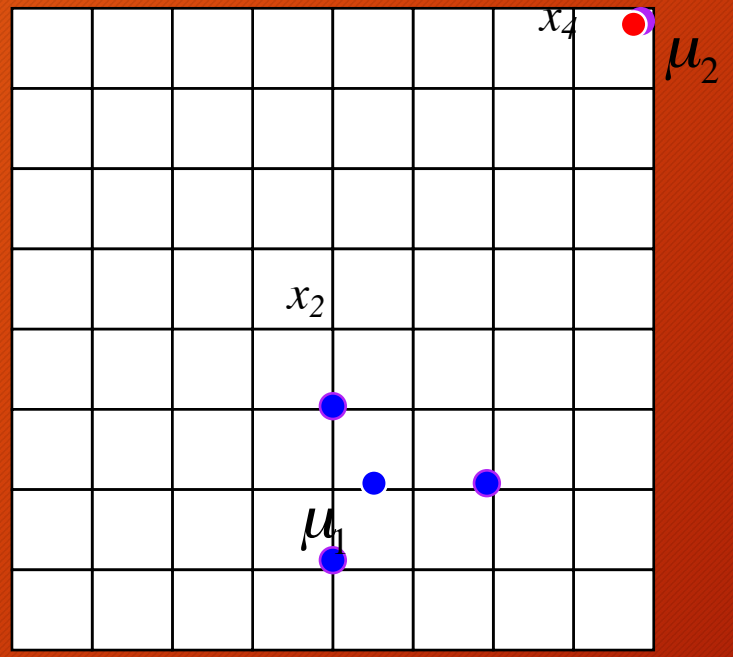
K-means: Assignment



K-means: UPDATE



K-means: FINAL



Algorithmic Design Choices

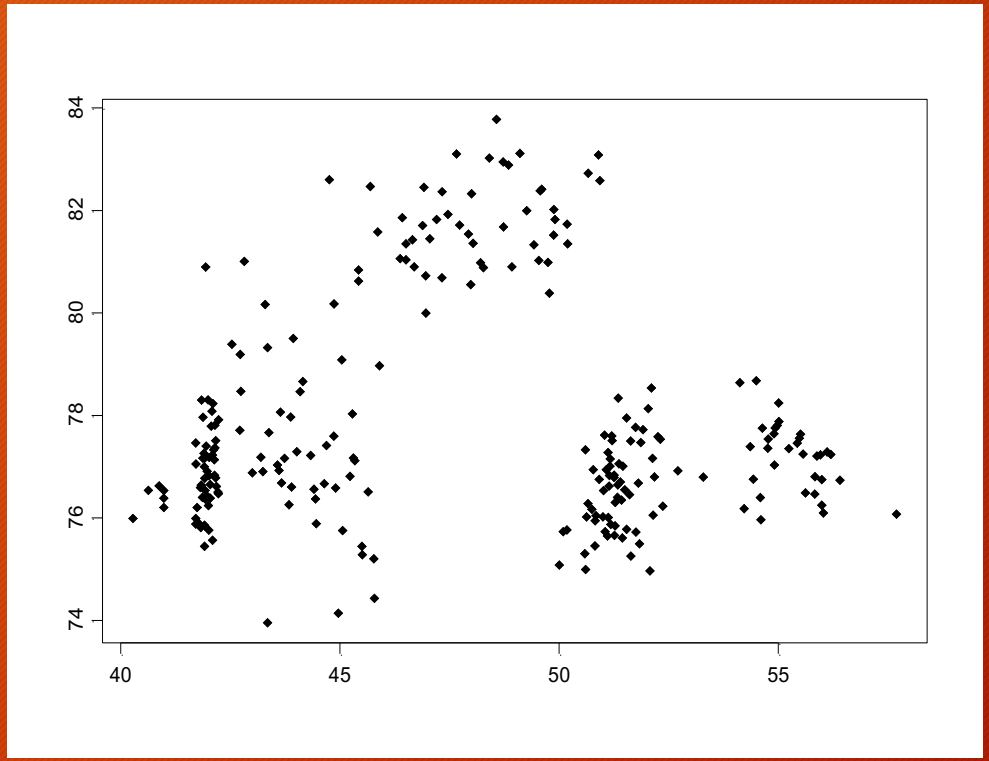
- Goal: cluster related (or similar) items together in a group

Questions for practitioner

- What makes a good group? Is the clustering result “good”?
- What type of algorithm should we choose?
- How many groups are there?

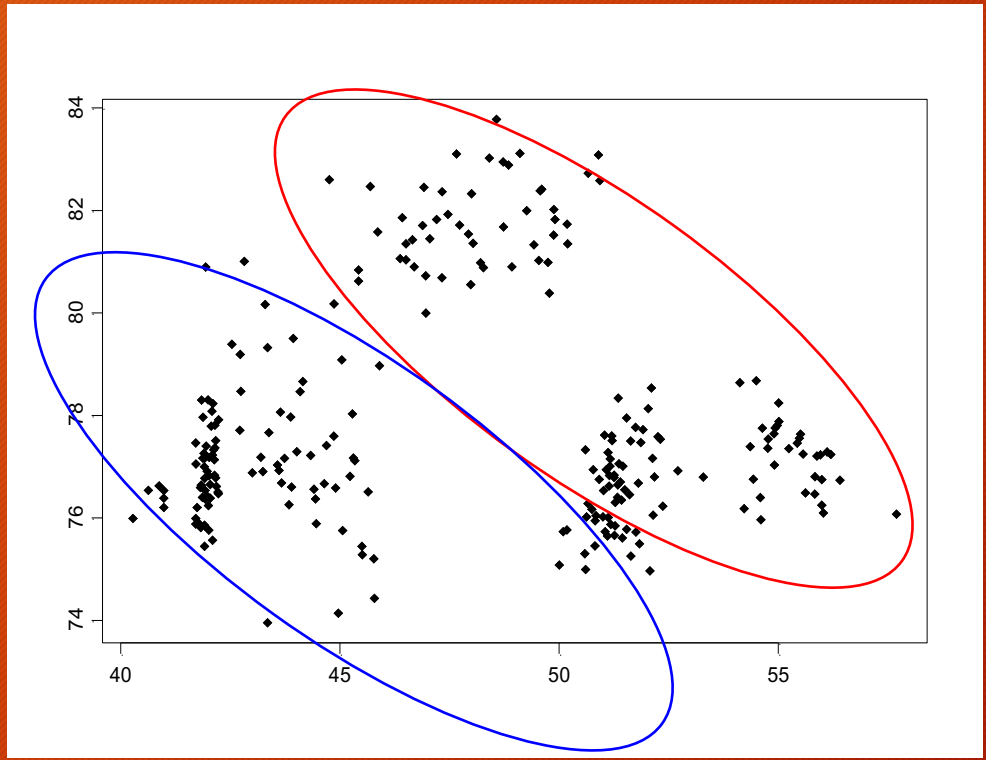
Cluster quality

- How should this data be clustered?



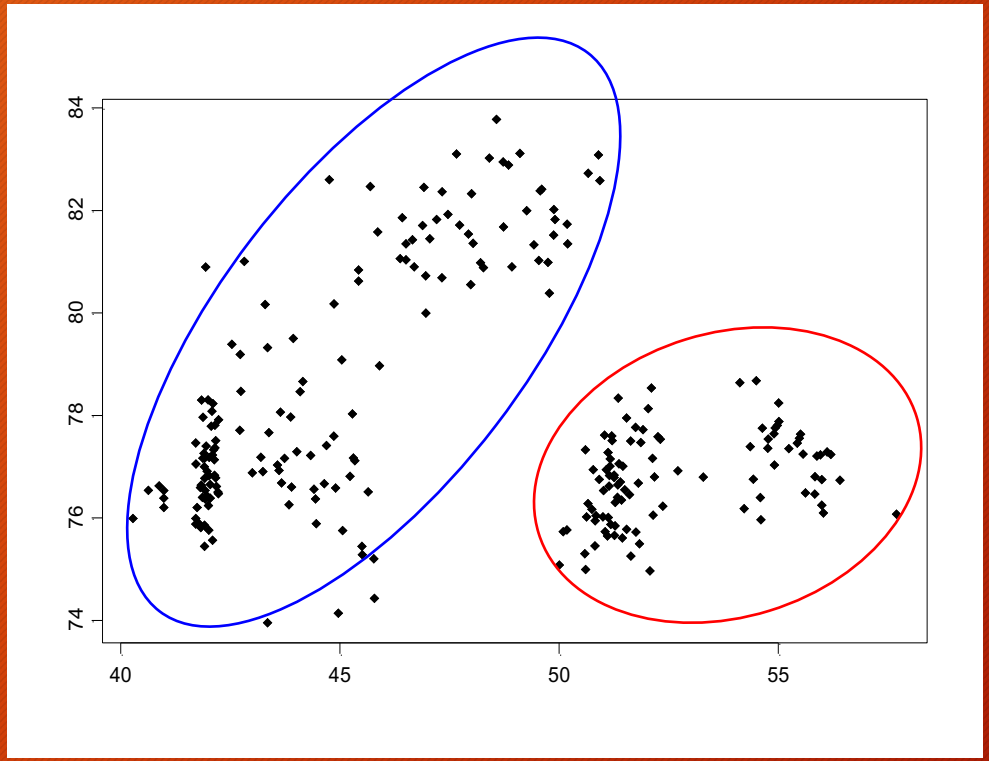
Cluster properties

- Is this a good cluster?



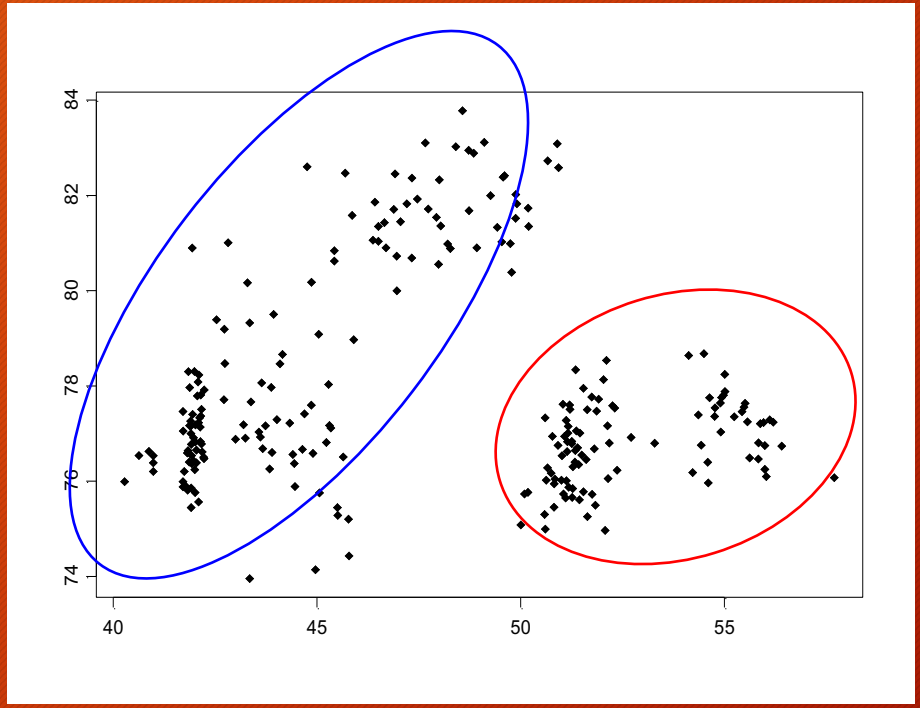
Cluster properties

- Is this better?



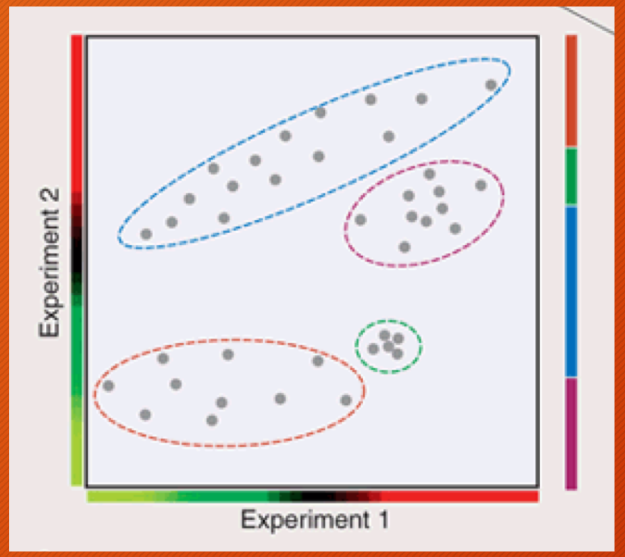
Cluster properties

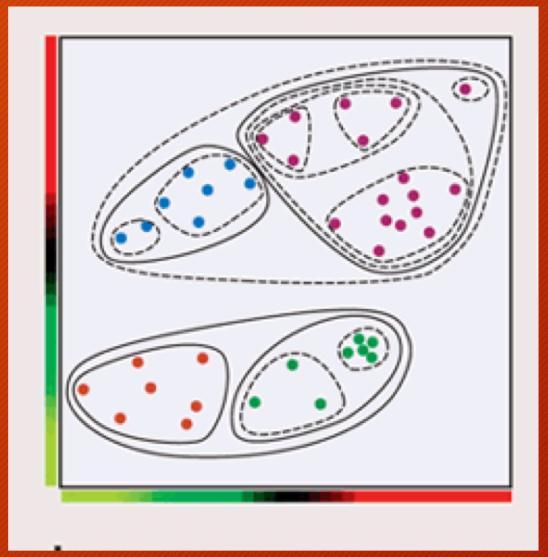
- Items within a group have high similarity (**homogeneity**)
- Items across groups have low similarity (**separation**)



Clustering Algorithms

- Dozens of approaches. Two main distinctions:
  - Hard vs Soft – can group assignments have uncertainty?
  - Flat vs. Hierarchical



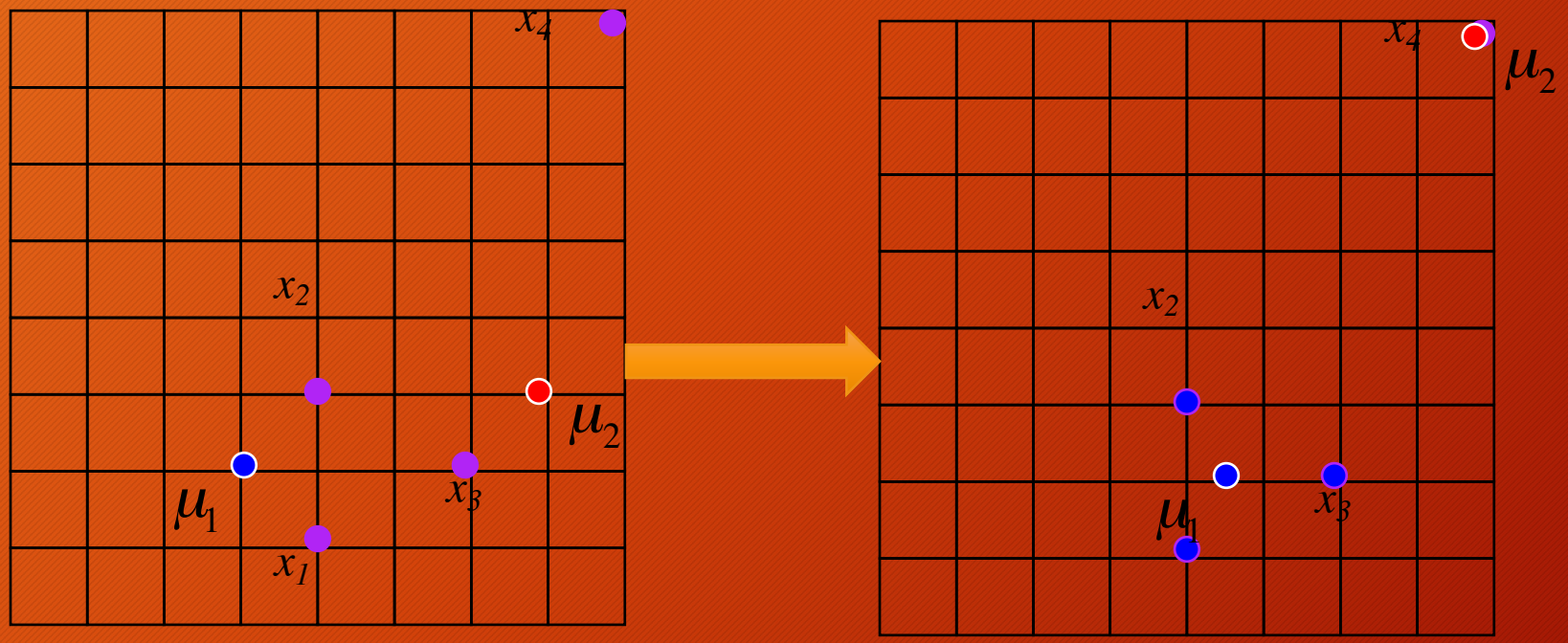


D'haeseleer, Nature Biotechnology, 2005

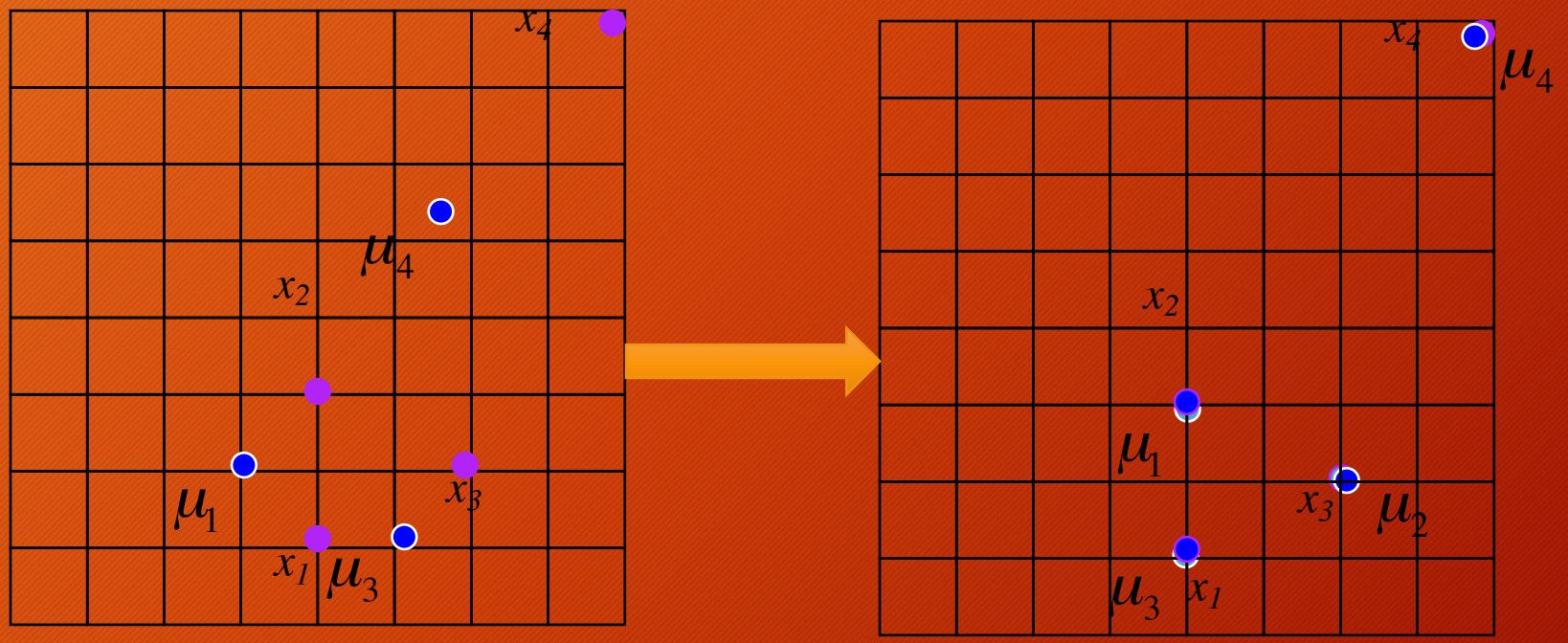
Pause for Exercises

- Compare and contrast k-means clustering https://www.naftaliharris.com/blog/visualizing-k-means-clustering/
- And DBScan https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/
- Cluster yeast genome

K-means, k=2



K-means, k=4



What is the primary goal of “learning”?

A. Provide an explanatory model of the task

B. Provide a model that generalizes to unseen examples of a task

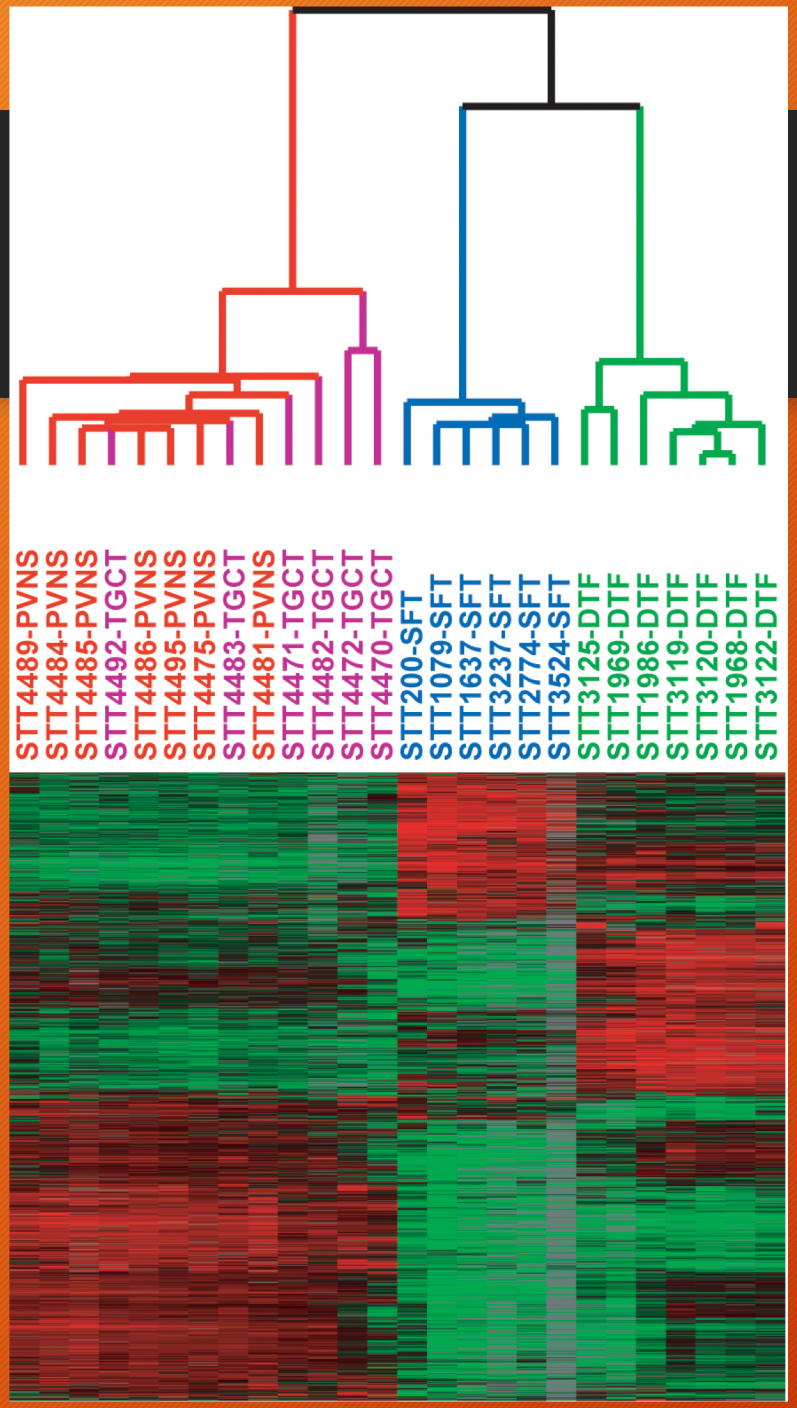
C. Provide a model that accurately represents given examples of a task

Generalization

- Primary goal of ML: develop models that *generalize* from specific examples to apply to future examples
  - Not the same as memorization, but related
  - Explanatory power can help
- Many causes for poor generalization
  - Algorithm assumptions
  - Learning procedures
  - Data
- Must analyze generalization at every point in ML pipeline

Selecting number of clusters

- Common solution: use domain knowledge to inform *k*
  - *e.g.*, "Doctors have characterized four subtypes of the cancer"
- *Regularize* the model by giving a penalty for too much complexity
- Hold aside data, see which models best "fit" the unseen data

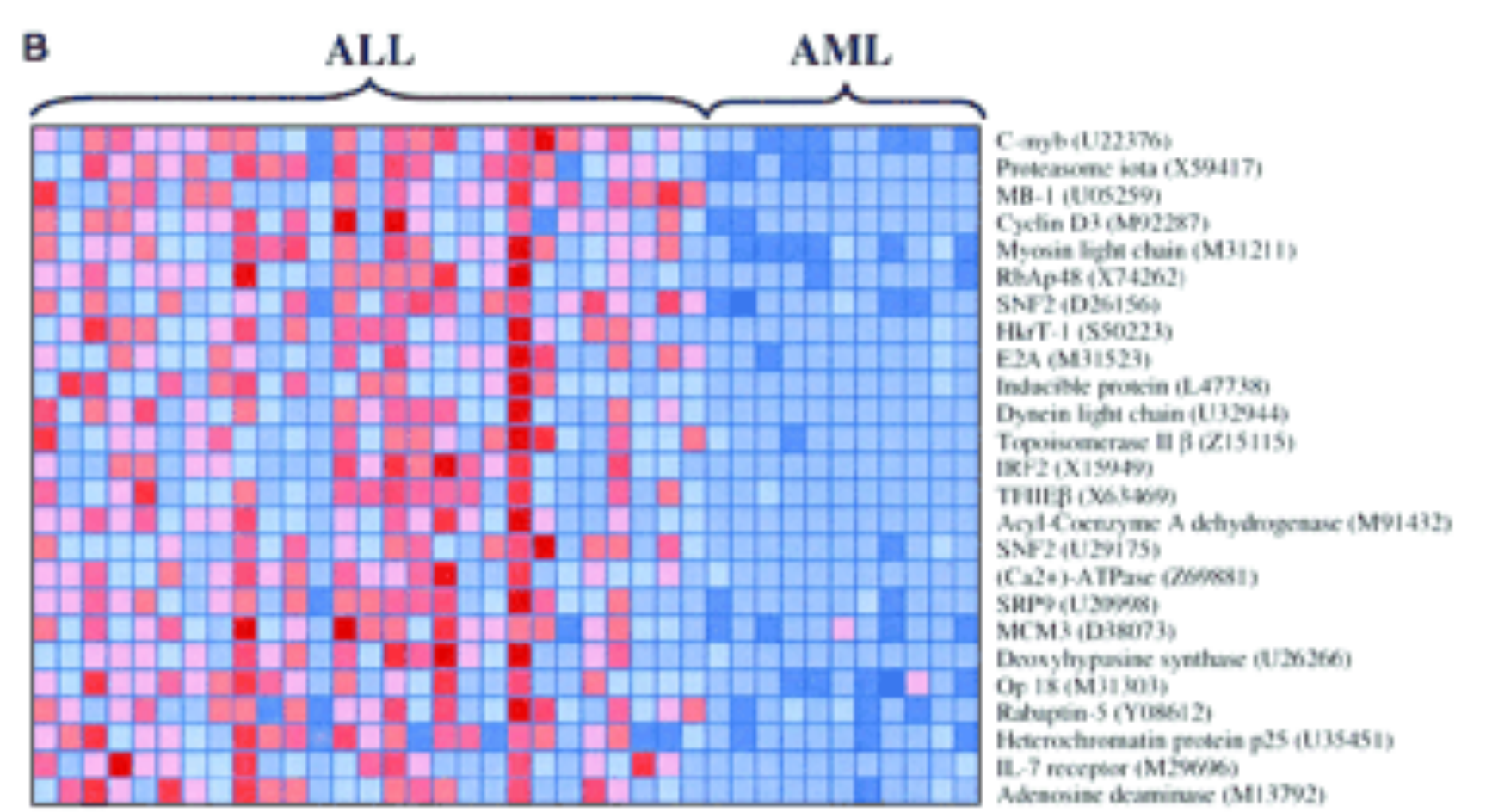


Hierarchical clustering example

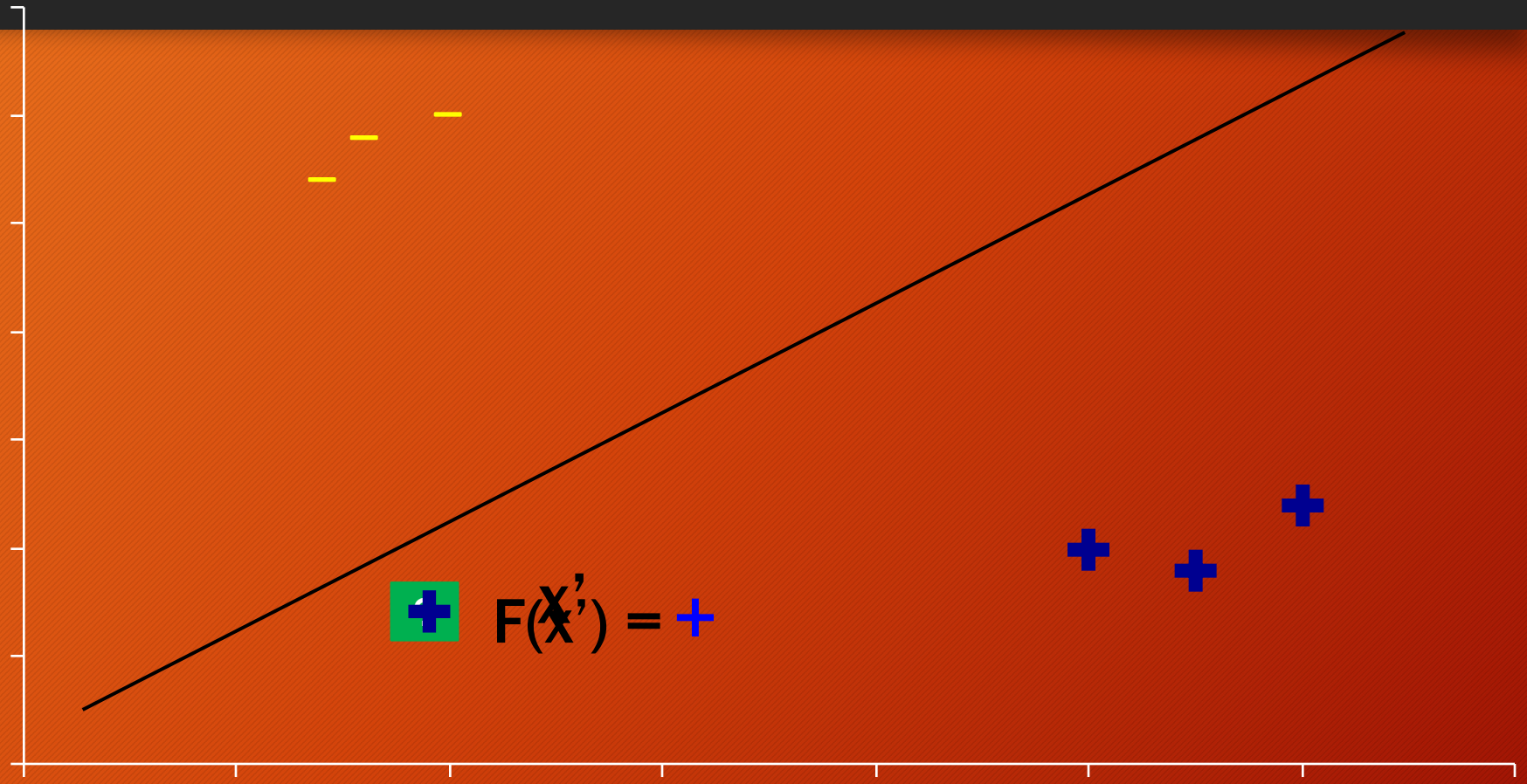
- clustering of related cancers and an inflammatory disorder
  - *TGCT*: Tenosynovial giant-cell tumor (purple)
  - *PVNS*: pigmented villonodular synovitis (red)
  - *SFT*: solitary fibrous tumor (blue)
  - *DTF*: desmoid-type fibromatosis (green)

figure from: West et al. *PNAS* 103, 2006

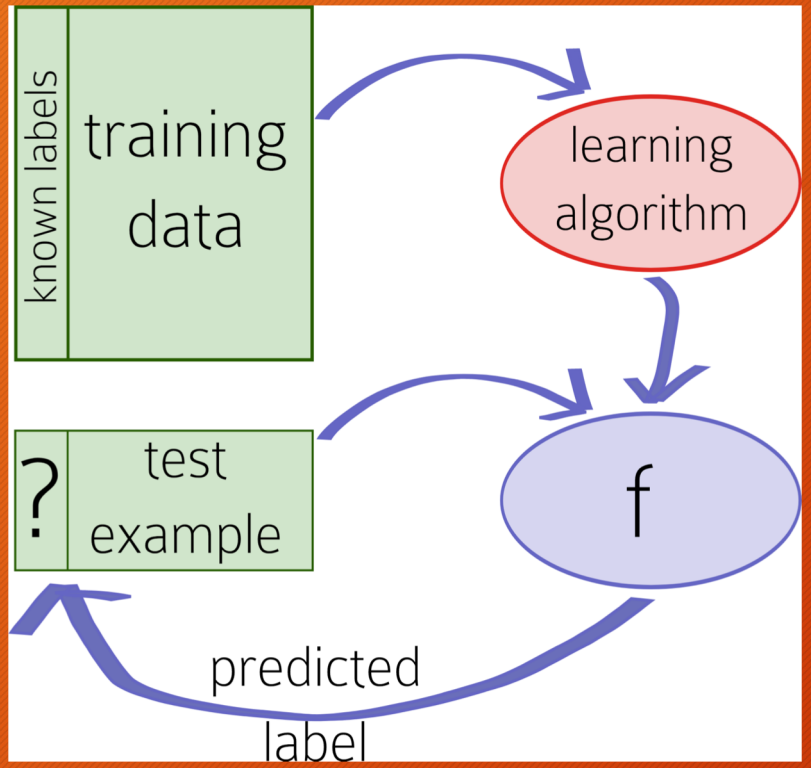
Supervised Learning



Supervised learning



Supervised Learning (induction)



- Labels: what we want to predict about an example
- **training data** – labels are known
  use for learning model *f*
- *f* is used to **predict label** for **test examples**

Typical Data set

X

Color	Shape	Size
red	square	big
blue	square	big
red	circle	small
yellow	square	small
red	circle	big

Y

Likes toy?
+
+
-
-
+

Machine Learning In Practice

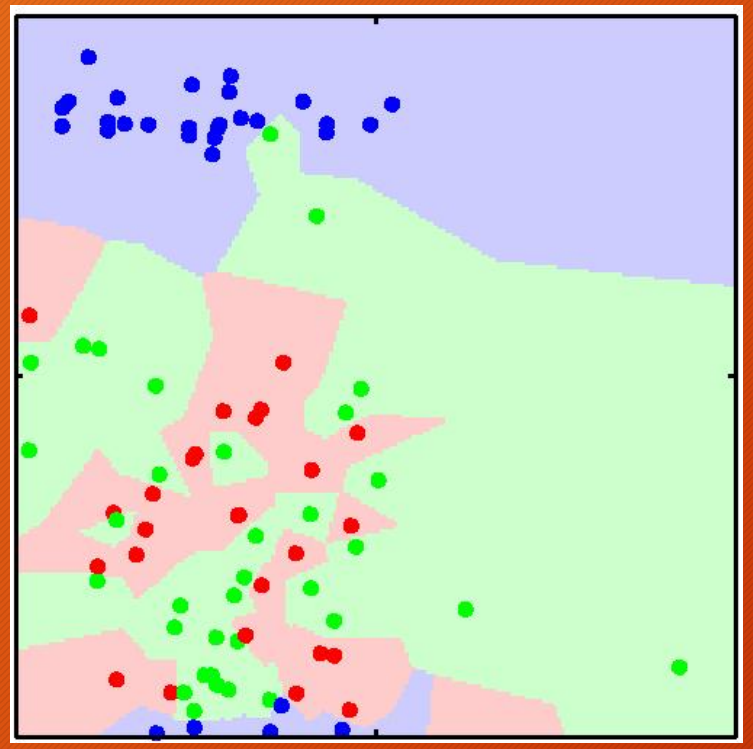
- Define the learning goal of your system?
- Collect and preprocess your data
- Pick a learning **framework**
- Pick a **data representation**
  - Inductive bias
  - Learning algorithm
- Train model, picking hyperparameters
- Evaluate model on test data
- Deploy!

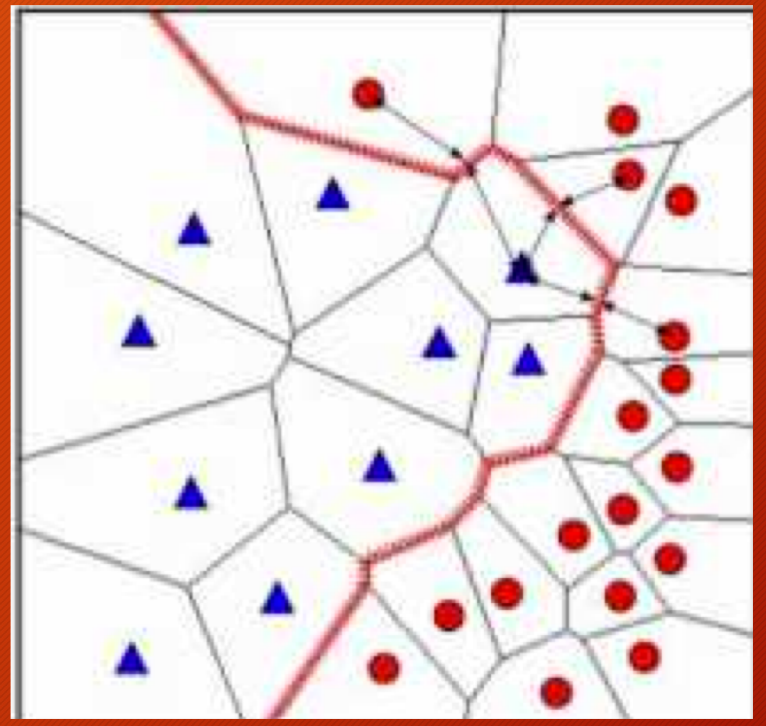
The world of classifiers

Representation	Evaluation	Optimization
Instances	Accuracy/Error rate	Combinatorial optimization
K-nearest neighbor	Precision and recall	Greedy search
Support vector machines	Squared error	Beam search
Hyperplanes	Likelihood	Branch-and-bound
Naive Bayes	Posterior probability	Continuous optimization
Logistic regression	Information gain	Unconstrained
Decision trees	K-L divergence	Gradient descent
Sets of rules	Cost/Utility	Conjugate gradient
Propositional rules	Margin	Quasi-Newton methods
Logic programs		Constrained
Neural networks		Linear programming
Graphical models		Quadratic programming
Bayesian networks		
Conditional random fields		

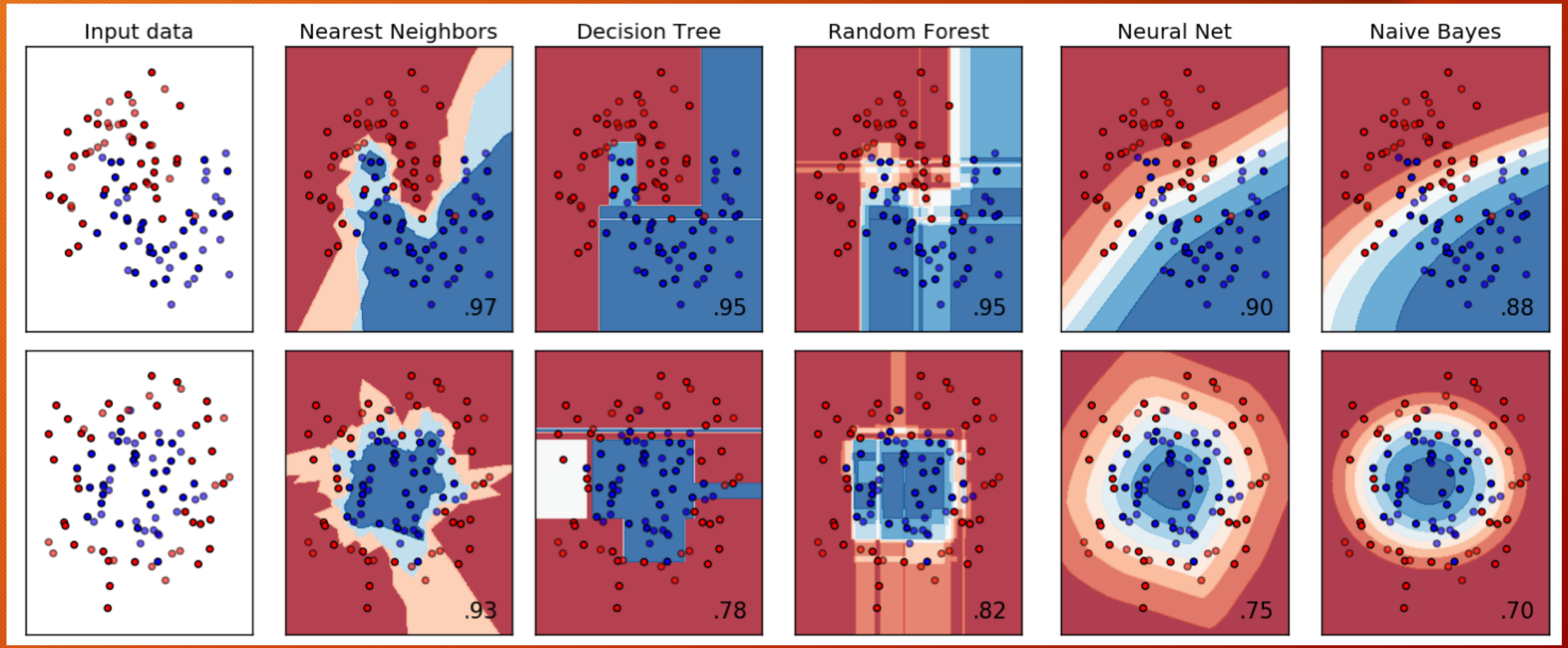
Pedro Domingos, *Communications of the ACM*, 2012

K-Nn Decision surface

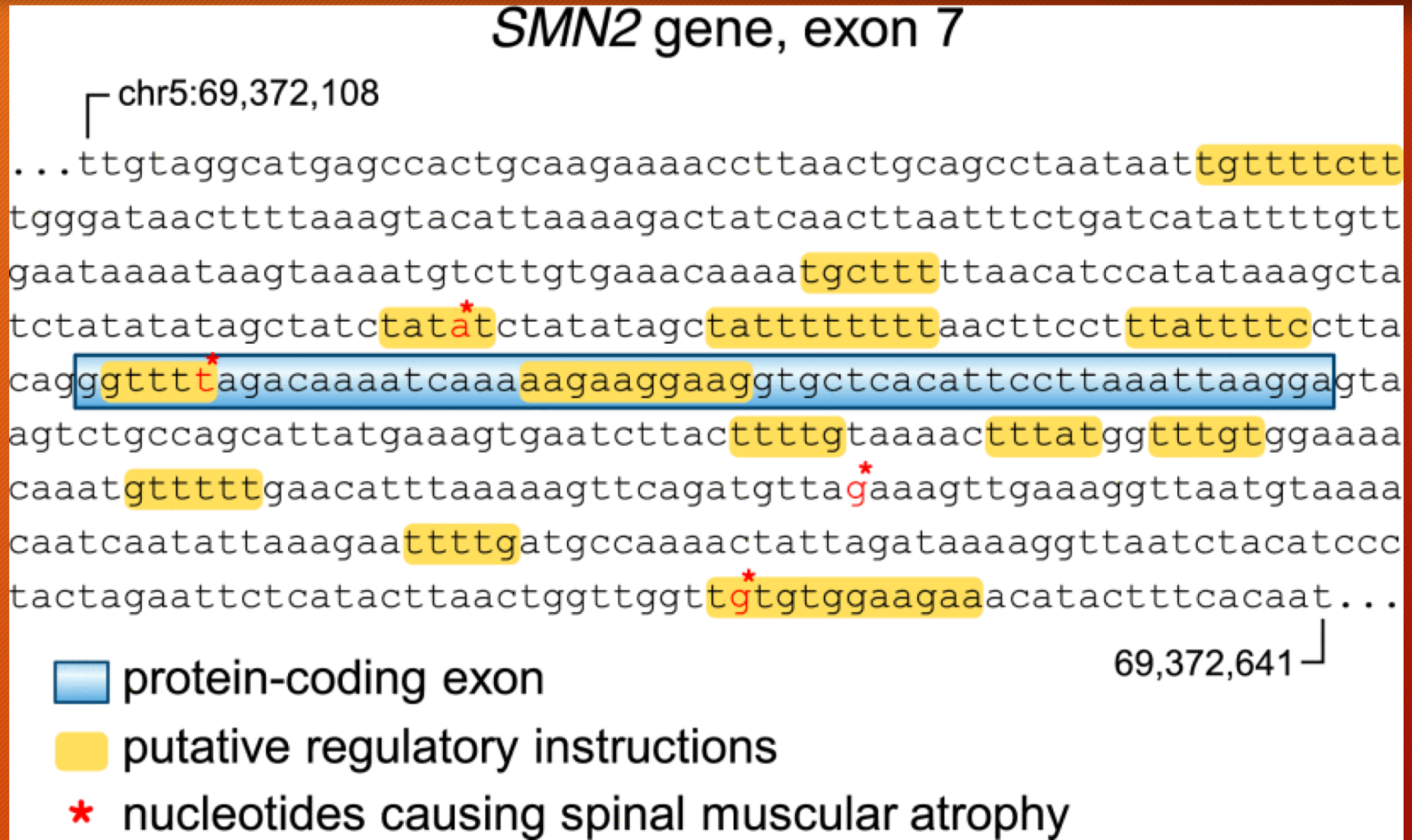




Decision surface



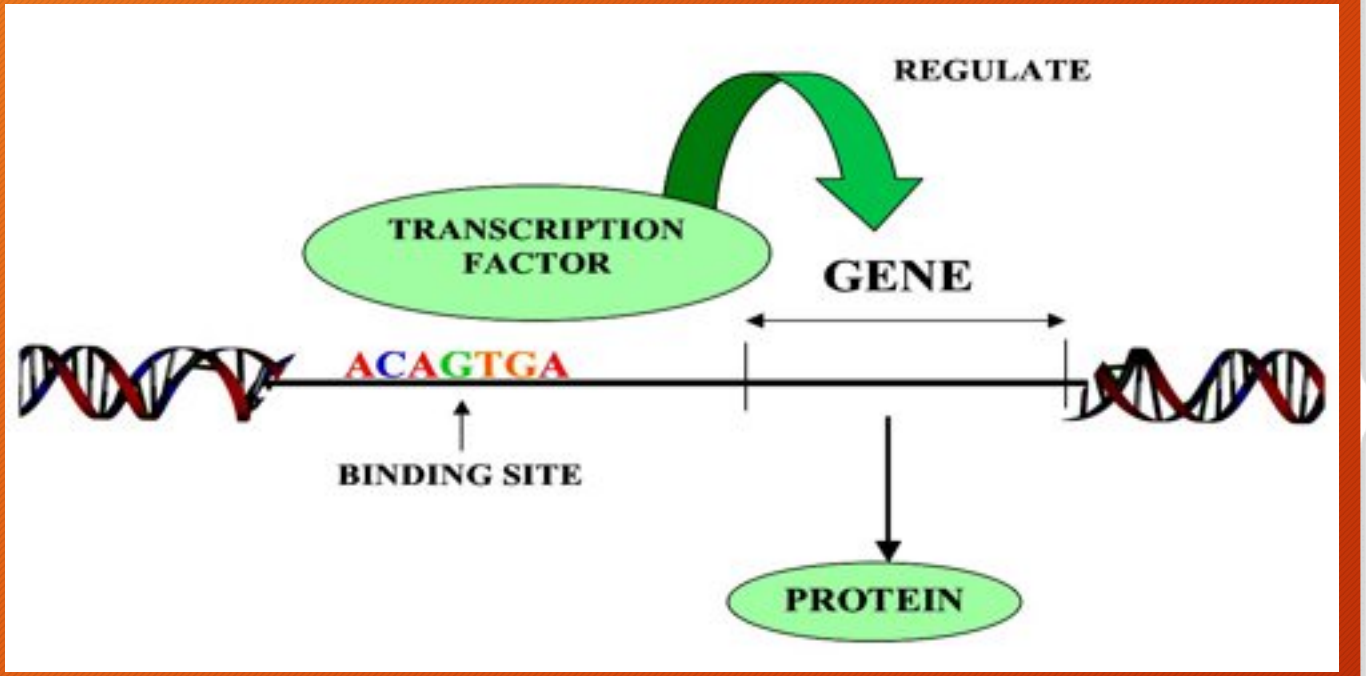
Application: Sequence Analysis

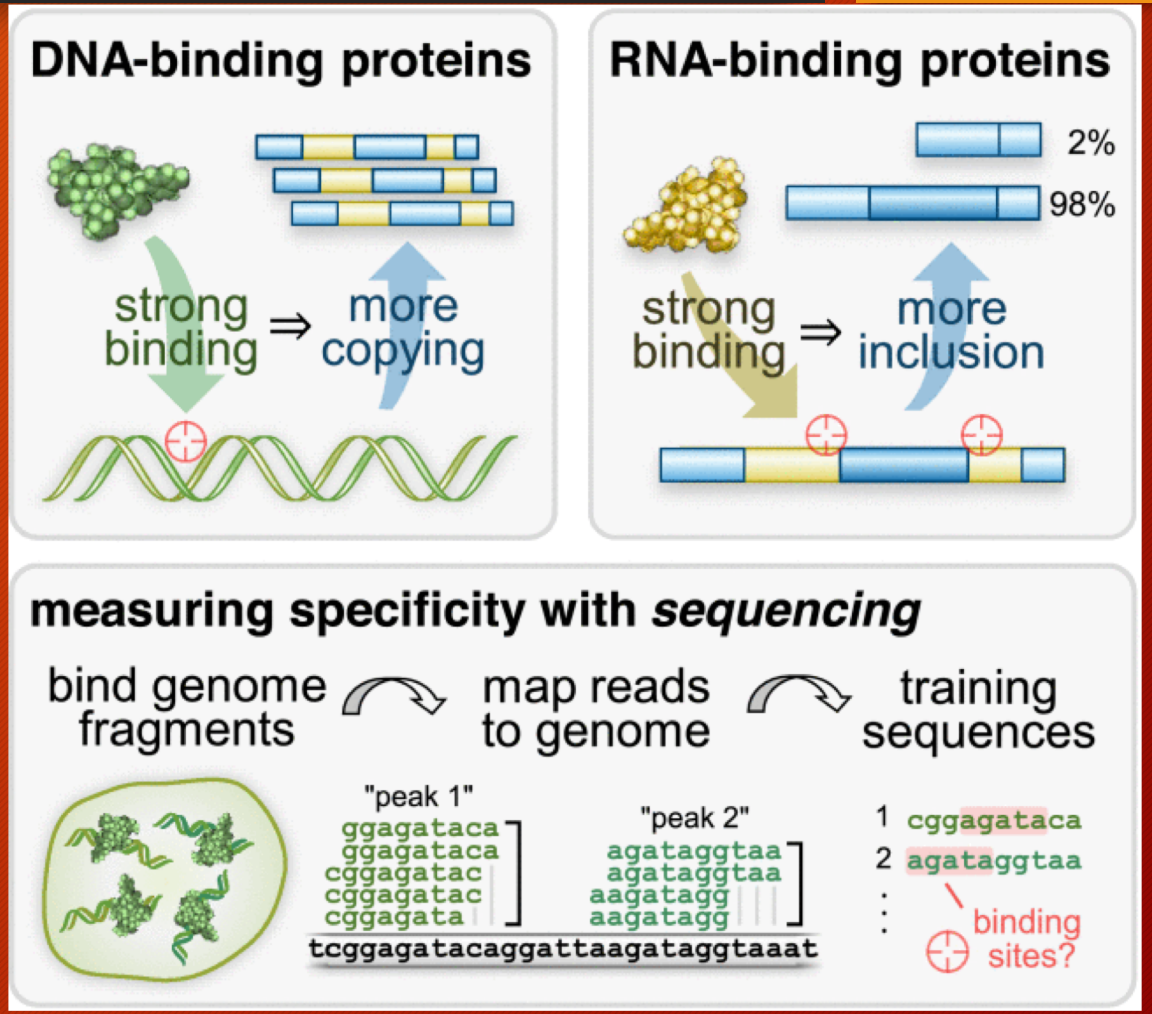


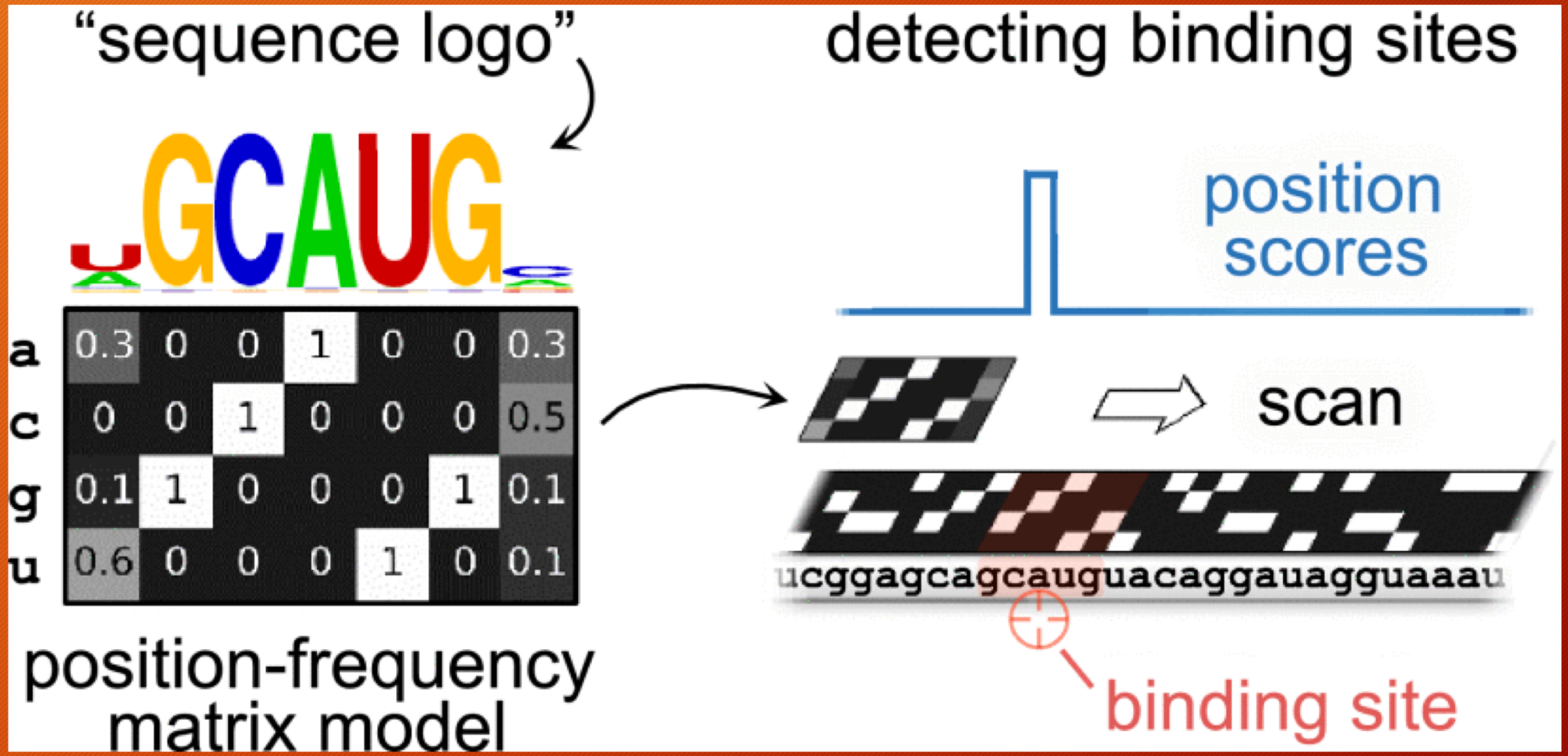
Feature Representation

- To machine learning, all data is just a matrix
- Codification of data (e.g., DNA sequence) is crucial design choice

Protein binding motifs

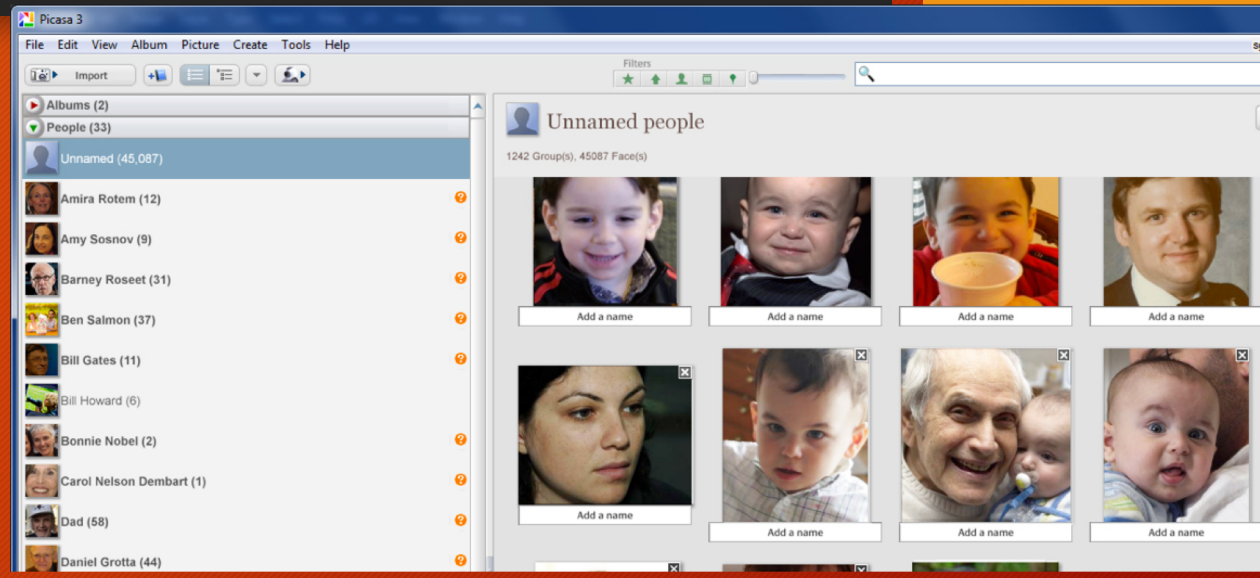





"sequence logo"
detecting binding sites
position scores
scan
a 0.3 0 0 1 0 0 0.3
c 0 0 1 0 0 0 0.5
g 0.1 1 0 0 0 1 0.1
u 0.6 0 0 0 1 0 0.1
position-frequency matrix model
ucggagcagcauguacaggauagguaaau
binding site

Why not always use supervised learning?

- Why not supervised?
  - Labels are expensive/time consuming
  - Big Data: Tons of data, very few labels
  - More data —> mitigate
    curse of dimensionality



- Advantages of unsupervised learning:
  - less burden on user (biologist)
  - better generalization - reduce selection bias/blind spots
  - algorithms can identify drift (aging) and novelty (new person)