Machine Learning for Computational Biology

Ameet Soni

Associate Professor

Computer Science

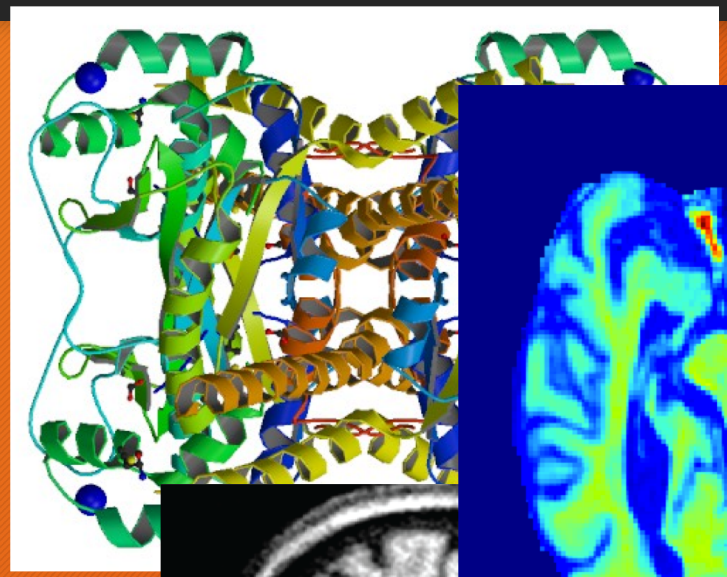
Swarthmore College

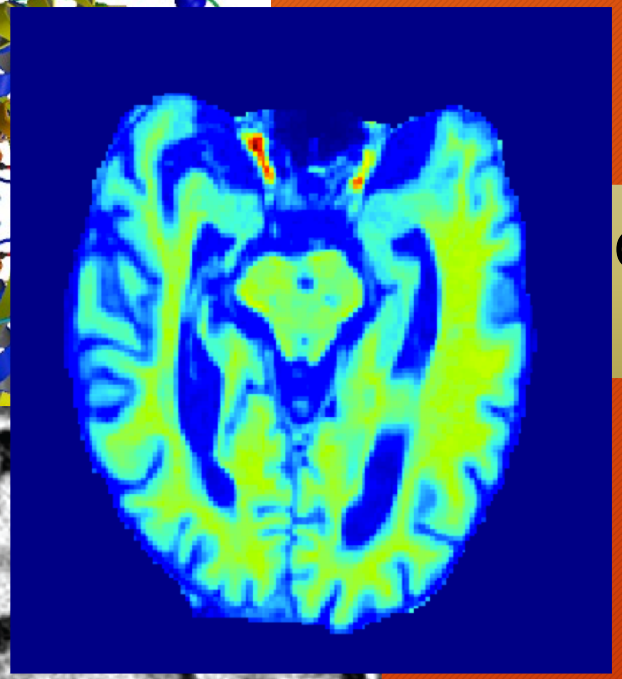
September 29, 2018 NSF Combio Workshop, Carleton College

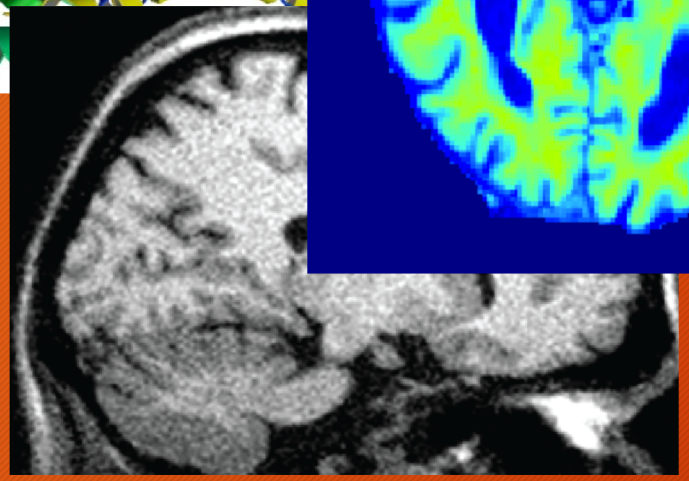
What is Machine learning?

- “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”

Research Projects







Brain Imaging

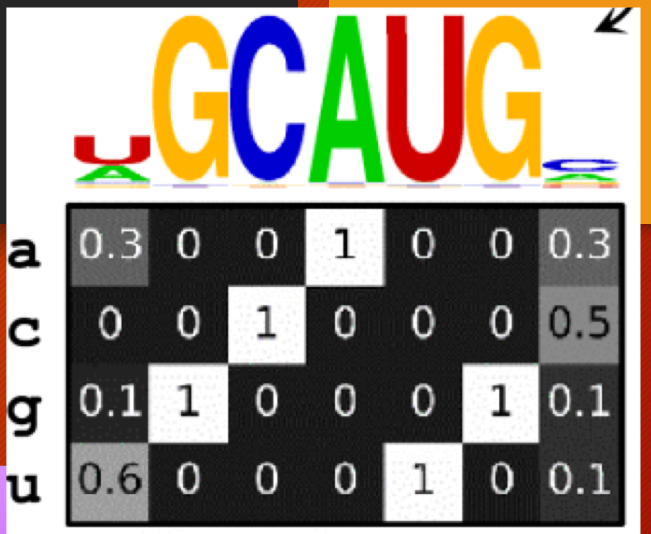
Doctor's notes

Genetic Data

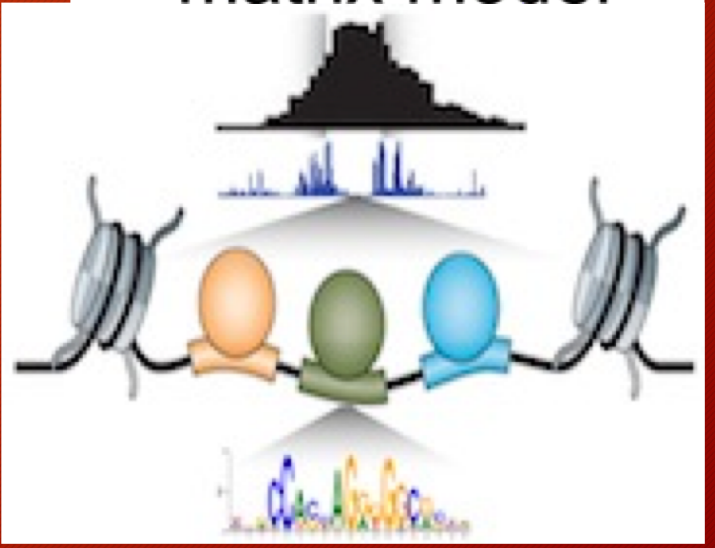
Blood work

Train Model

Predict Parkinson's Disease



position-frequency matrix model



Why Machine Learning?

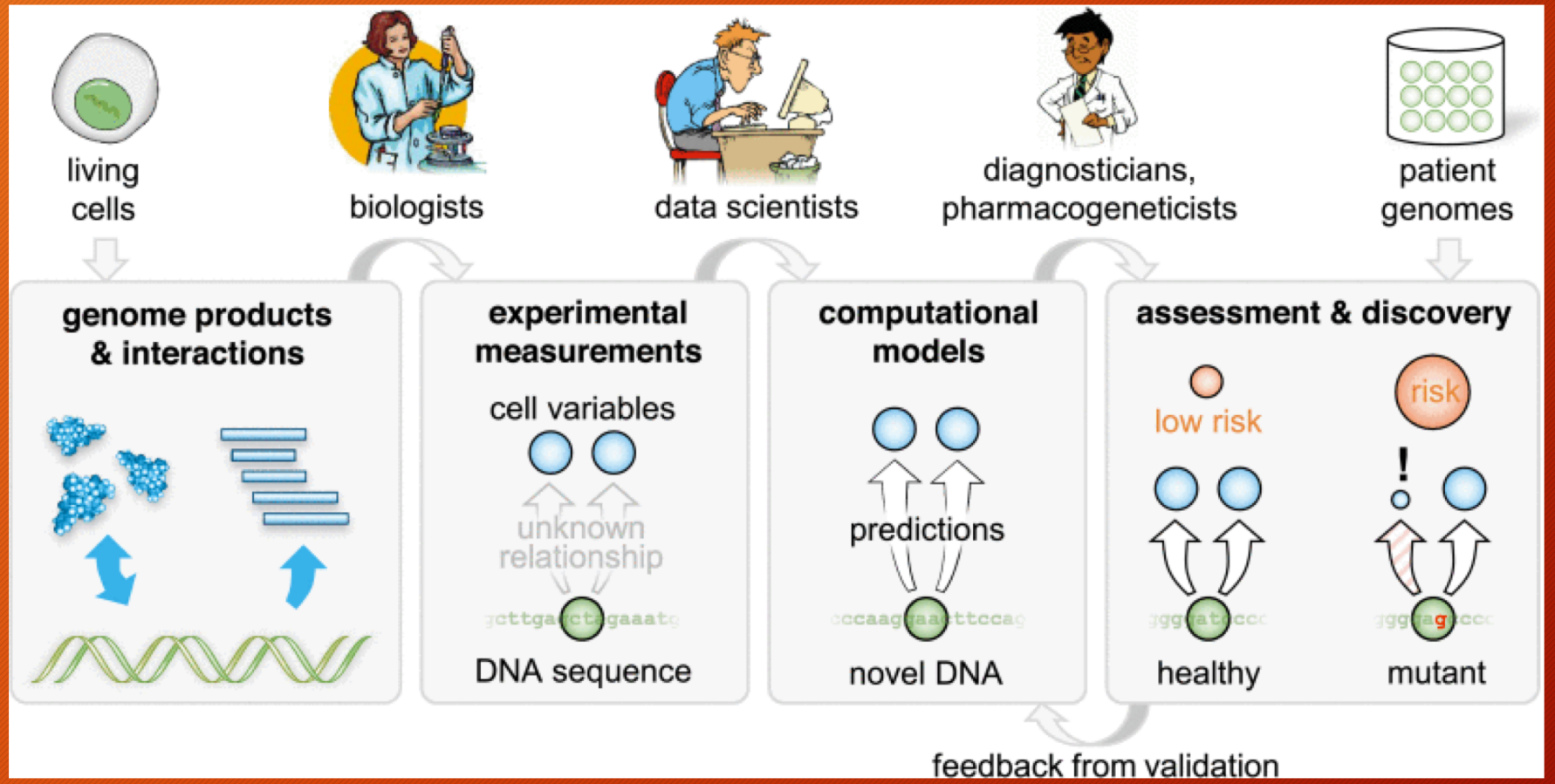
Many systems are too difficult/expensive because of:

- 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



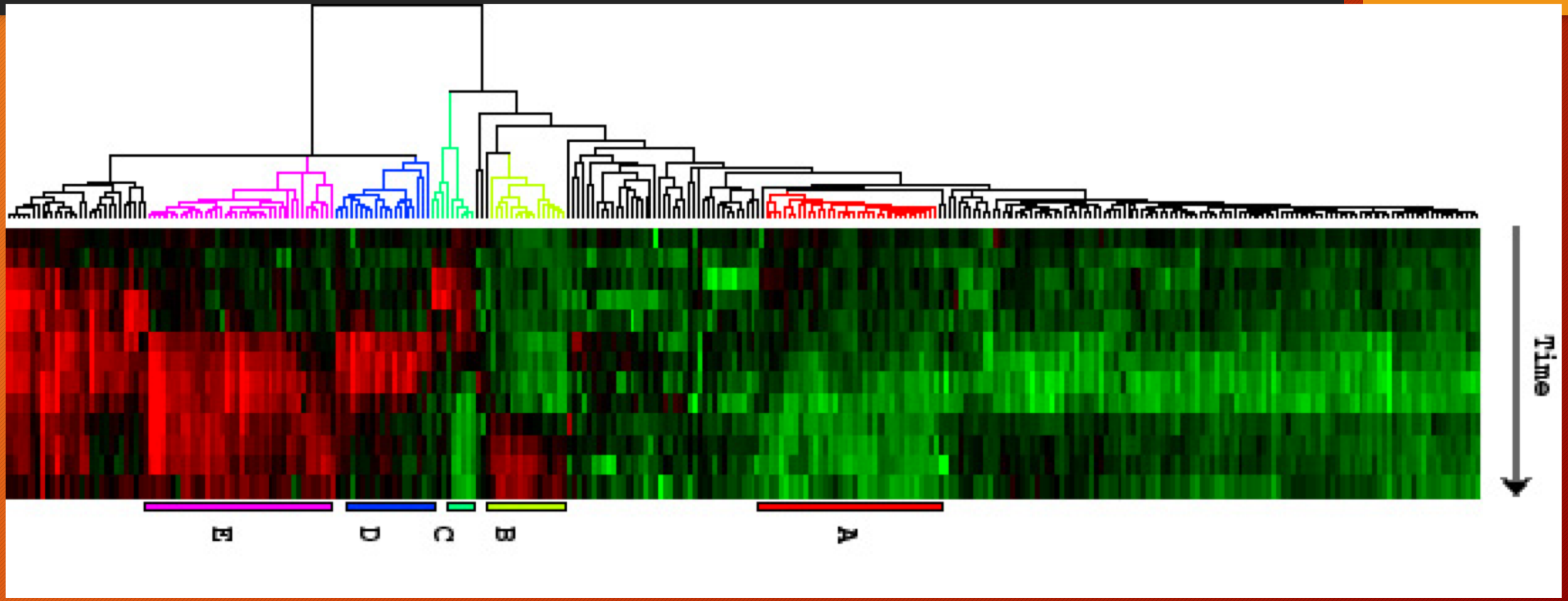
Leung et al., *IEEE*, 2016

Machine Learning Frameworks

- Supervised learning: learn from answers (prediction)
- Unsupervised learning: learn without answers (clustering)

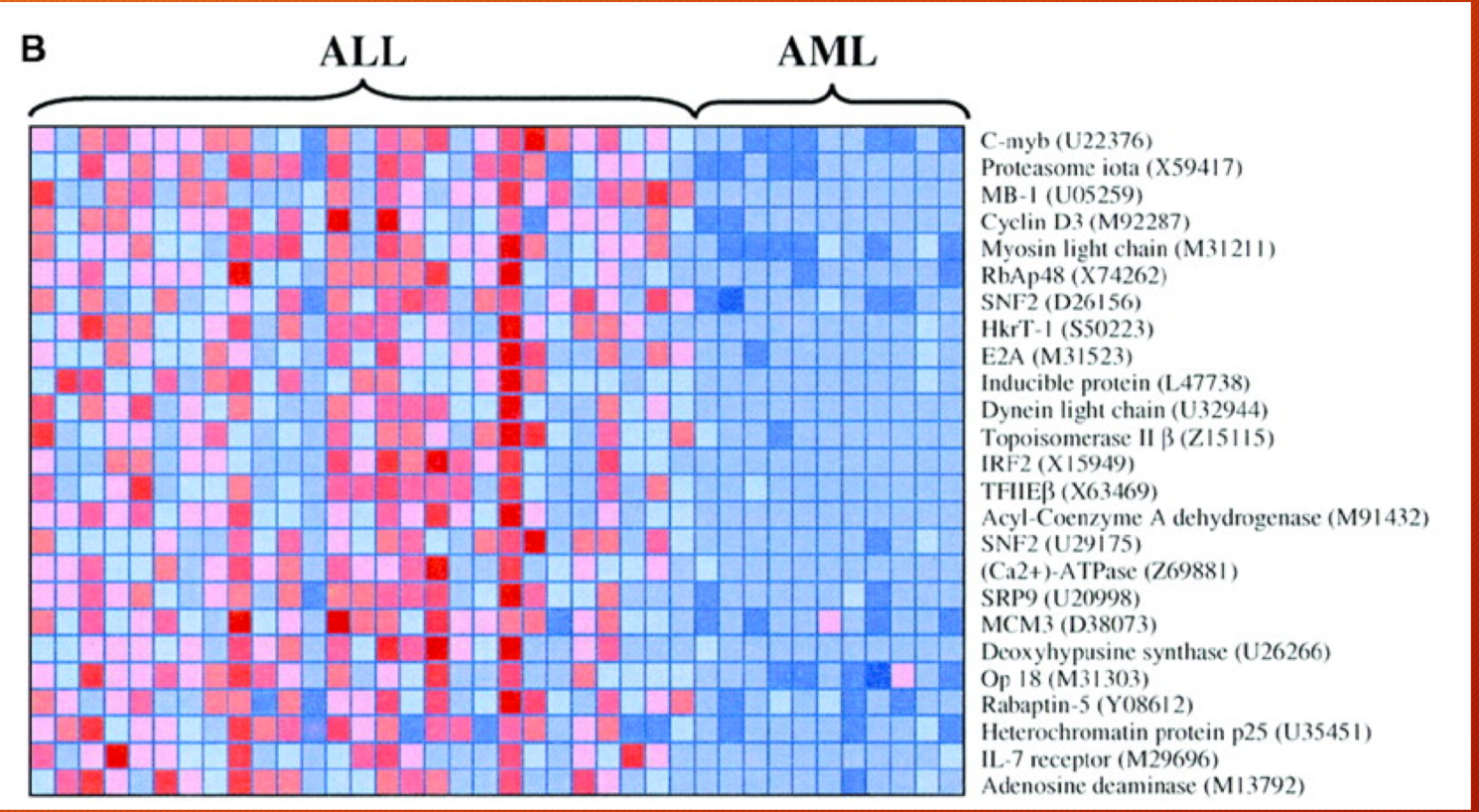
- Other:
  - Semi-supervised
  - Time-series
  - Structured prediction
  - Active learning

Unsupervised: Gene Function



Gillet et al., *PNAS* 2011

Supervised: Disease Diagnosis



Golub et al., *Science*, 1999

Sequence: Gene Finding

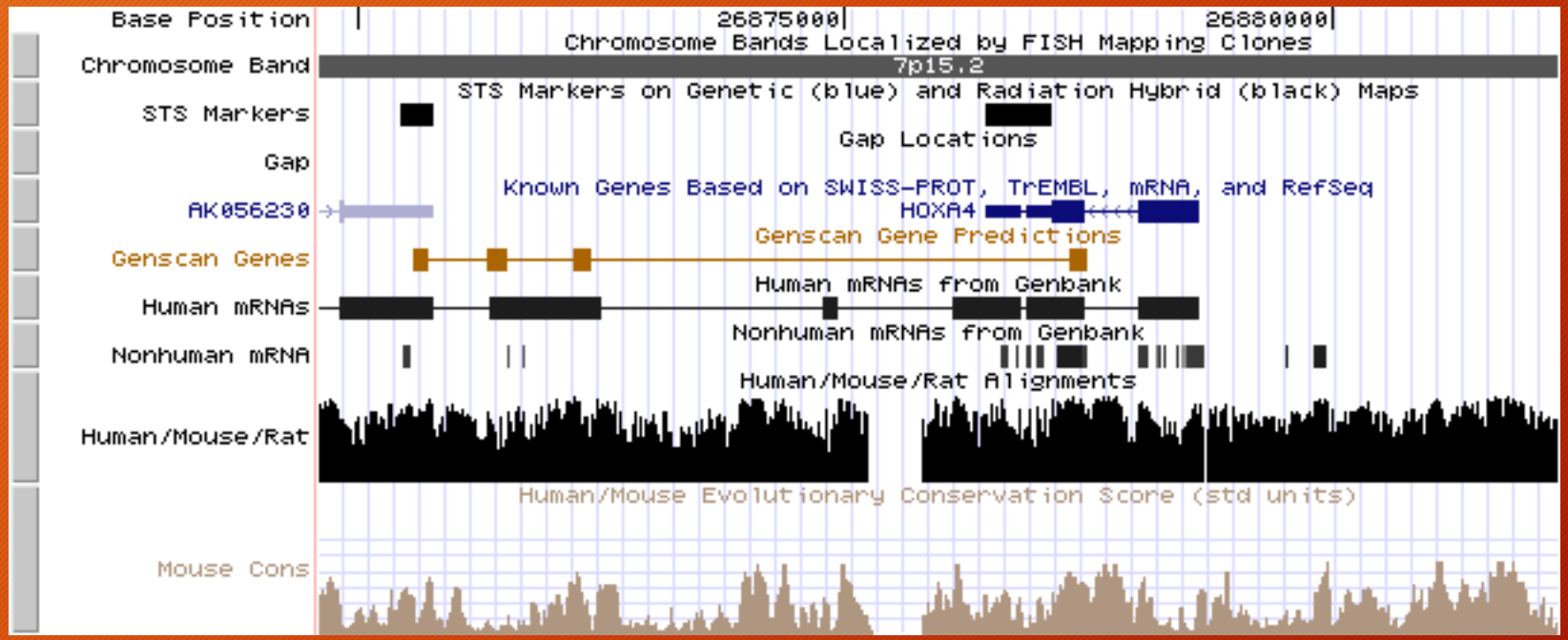
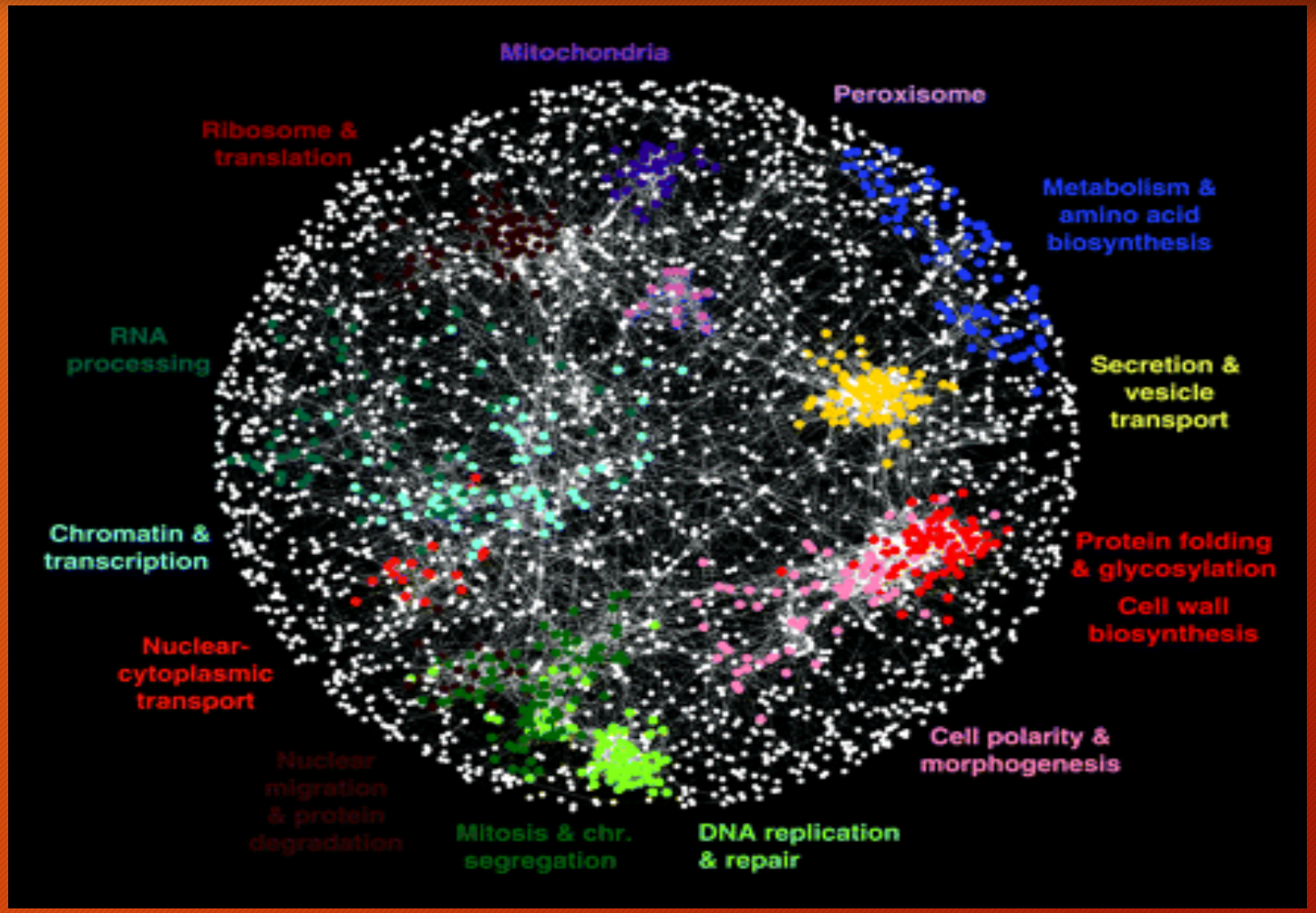


image from the UCSC Genome Browserhttp://genome.ucsc.edu/

Structure: Gene Interaction Network



Costanzo et al., *Biotechnology* 2011

Outline

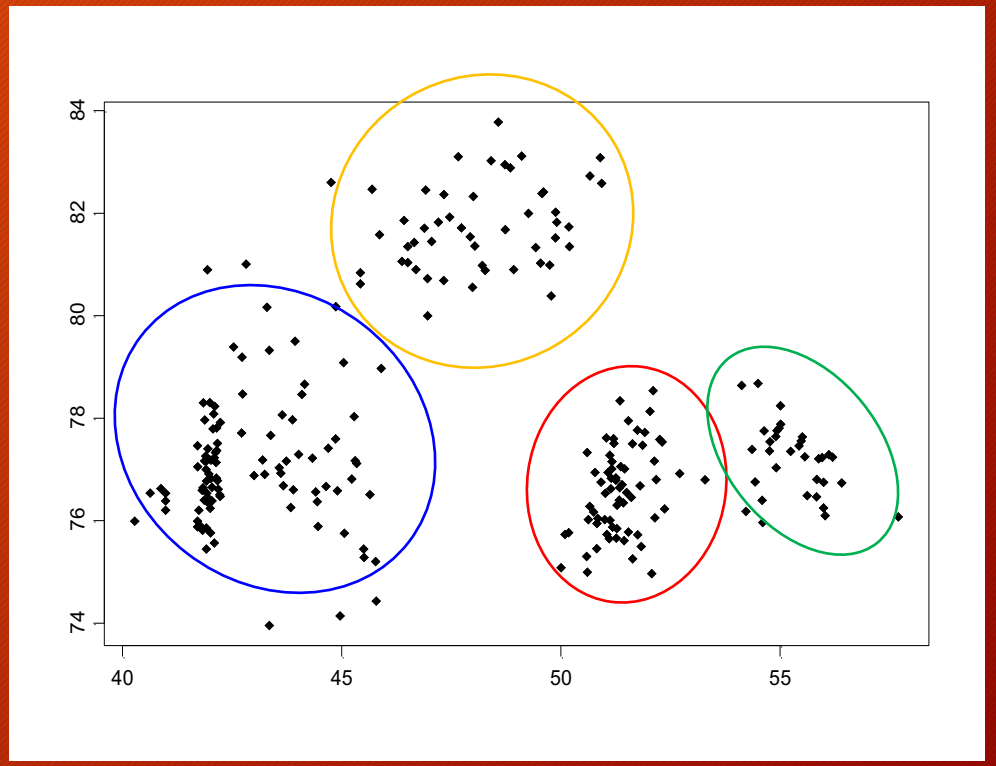
- Clustering Algorithms
  - K-Means algorithm
  - Application: gene expression
- Classification Algorithms
  - Application: disease prediction
  - Application: sequence annotation

What is Clustering?

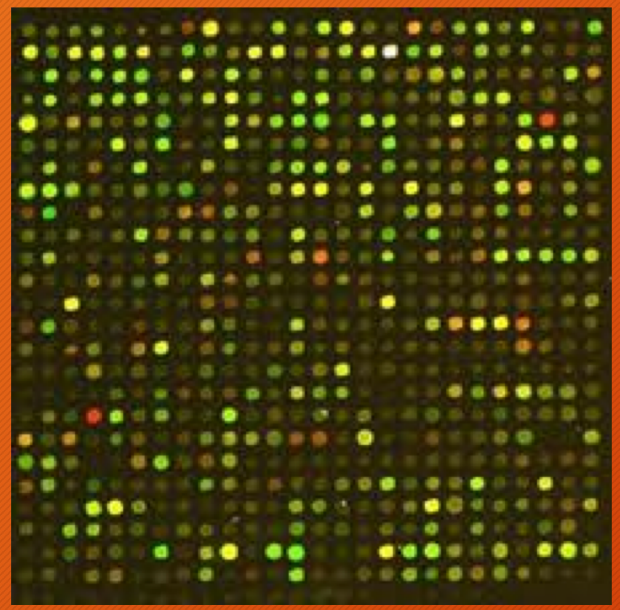
- *Clustering* - procedure that detects the presence of distinct groups

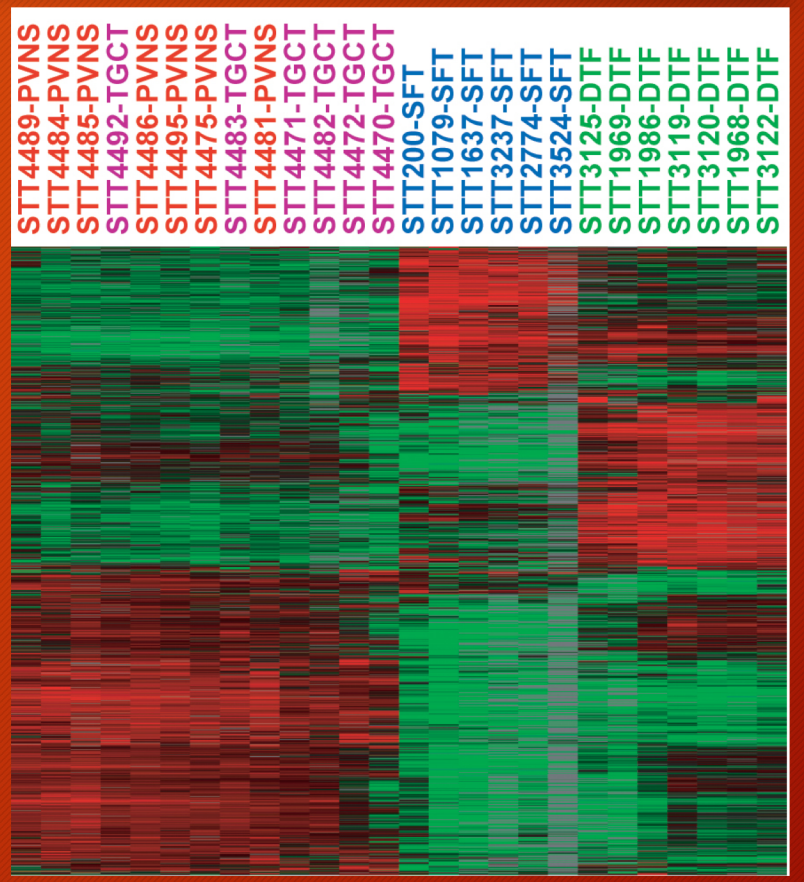
  A form of unsupervised learning

- Applications:
  - Visualization
  - Data exploration

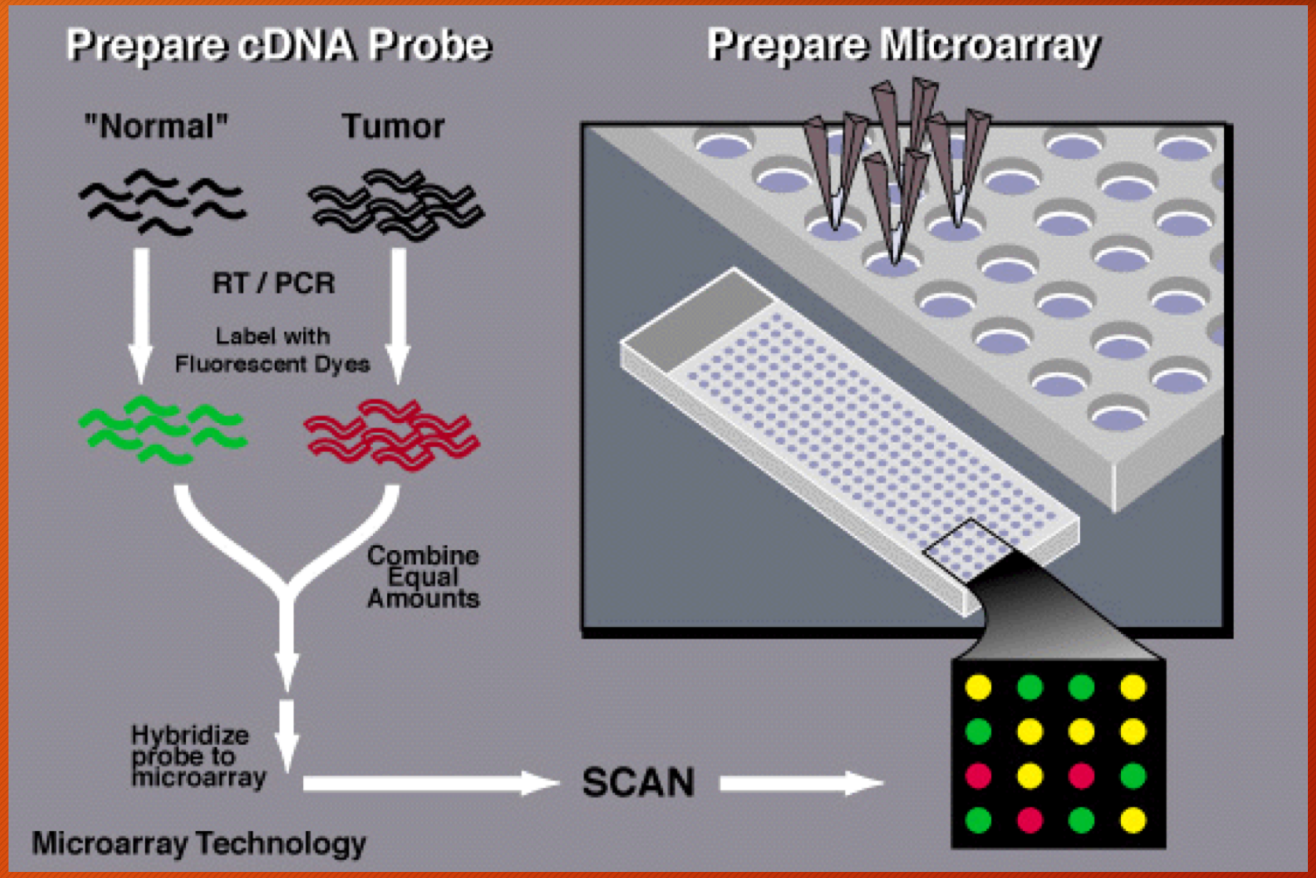


Application: Gene Expression





Spotted cDNA Microarrays



Also look at this animation: Microarray Animation

Microarray Data – Intensity matrix

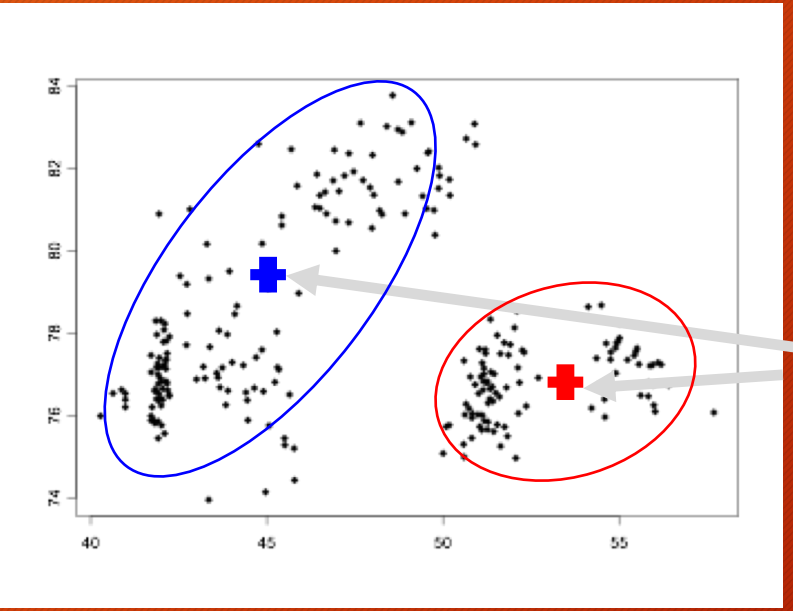
Time:	Time X	Time Y	Time Z
Gene 1	10	8	10
Gene 2	10	0	9
Gene 3	4	8.6	3
Gene 4	7	8	3
Gene 5	1	2	3

Other Biological Applications

- Population genetics (SNP profiles)
- Tree of life/phylogenetics (evolutionary markers/DNA)
- Gene function analysis (expression values)
- Protein function (structure)
- Image analysis (segmentation)

K-Means Clustering

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



Cluster Centers

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

Lloyd's Algorithm

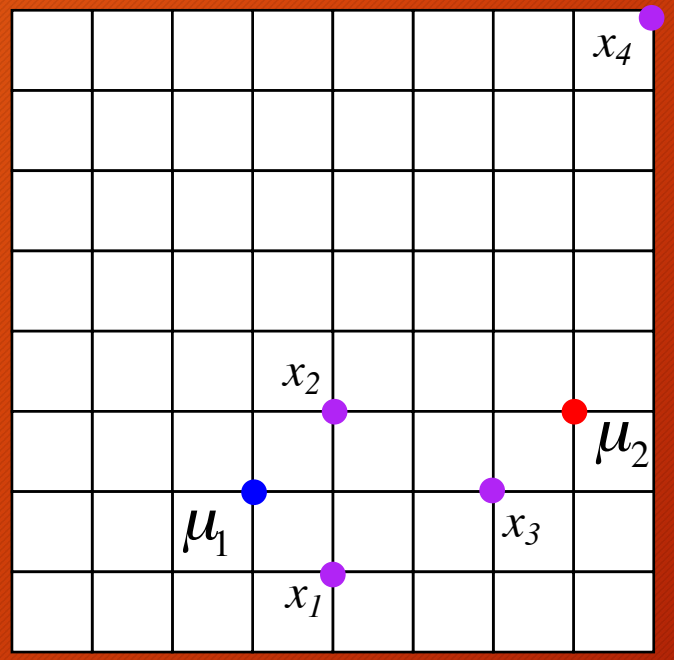
- Pick *k* initial cluster centers
- Put each data point in nearest cluster center
- Re-estimate cluster center
- Repeat

Efficient, but approximate solution

K-means Clustering Example

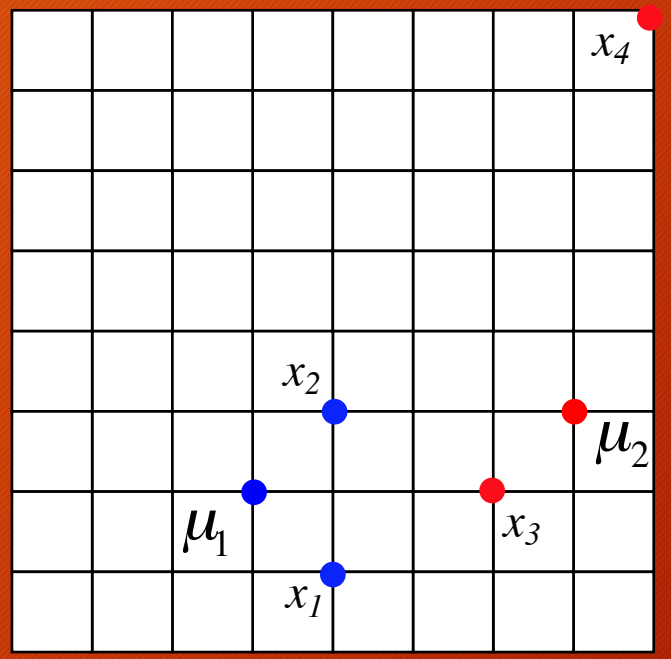
Step 1: assign points to clusters

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



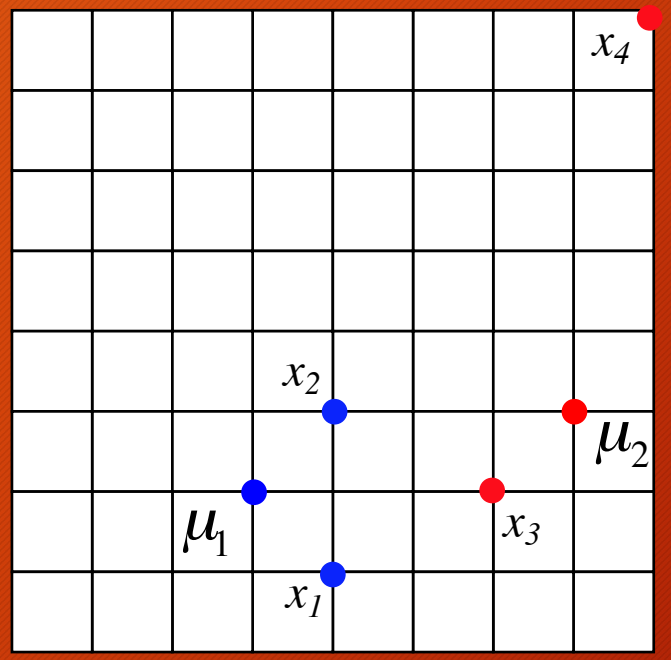
K-means: Assignment

Step 2: Calculate the updated means



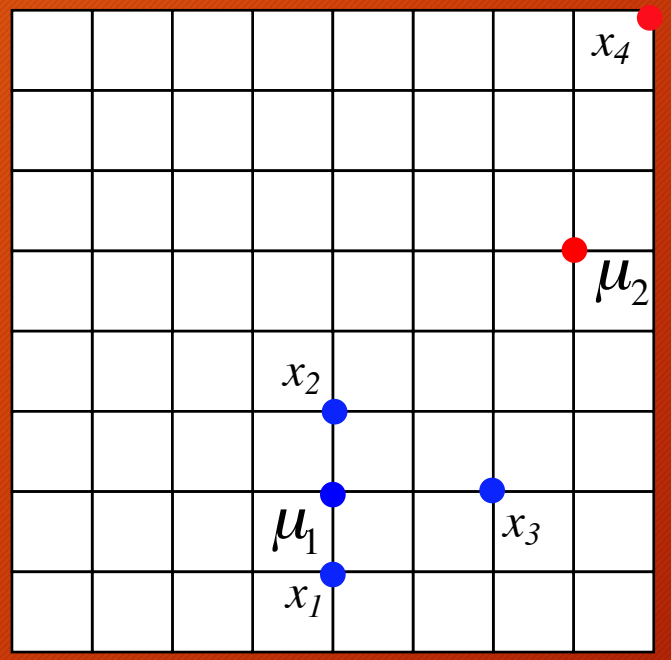
K-means: Update Means

Step 1: Reassign points to clusters

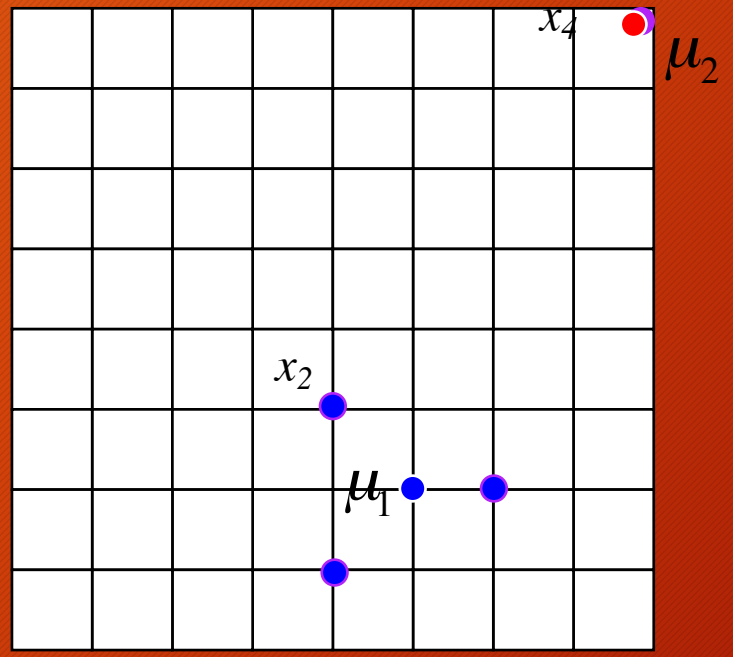


K-means: Reassign

Step 2: Update means



K-means: FINAL



Algorithmic Design Choices

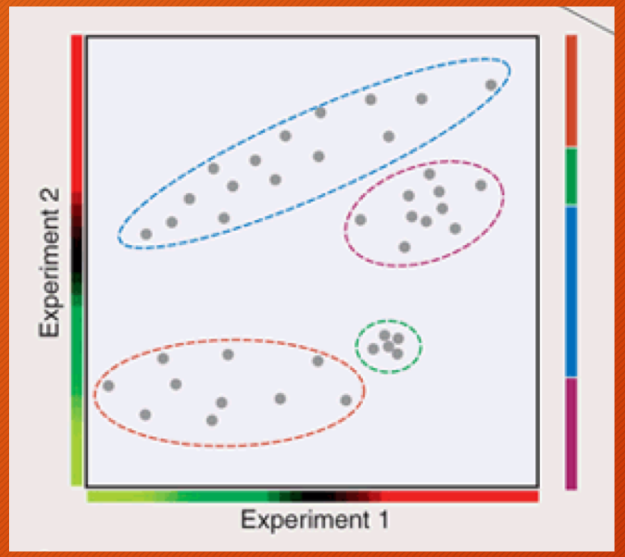
- Goal: cluster related (or similar) items together in a group

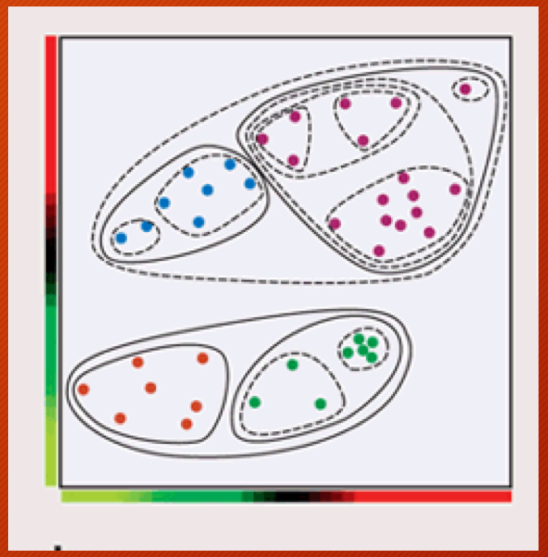
Questions for practitioner

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

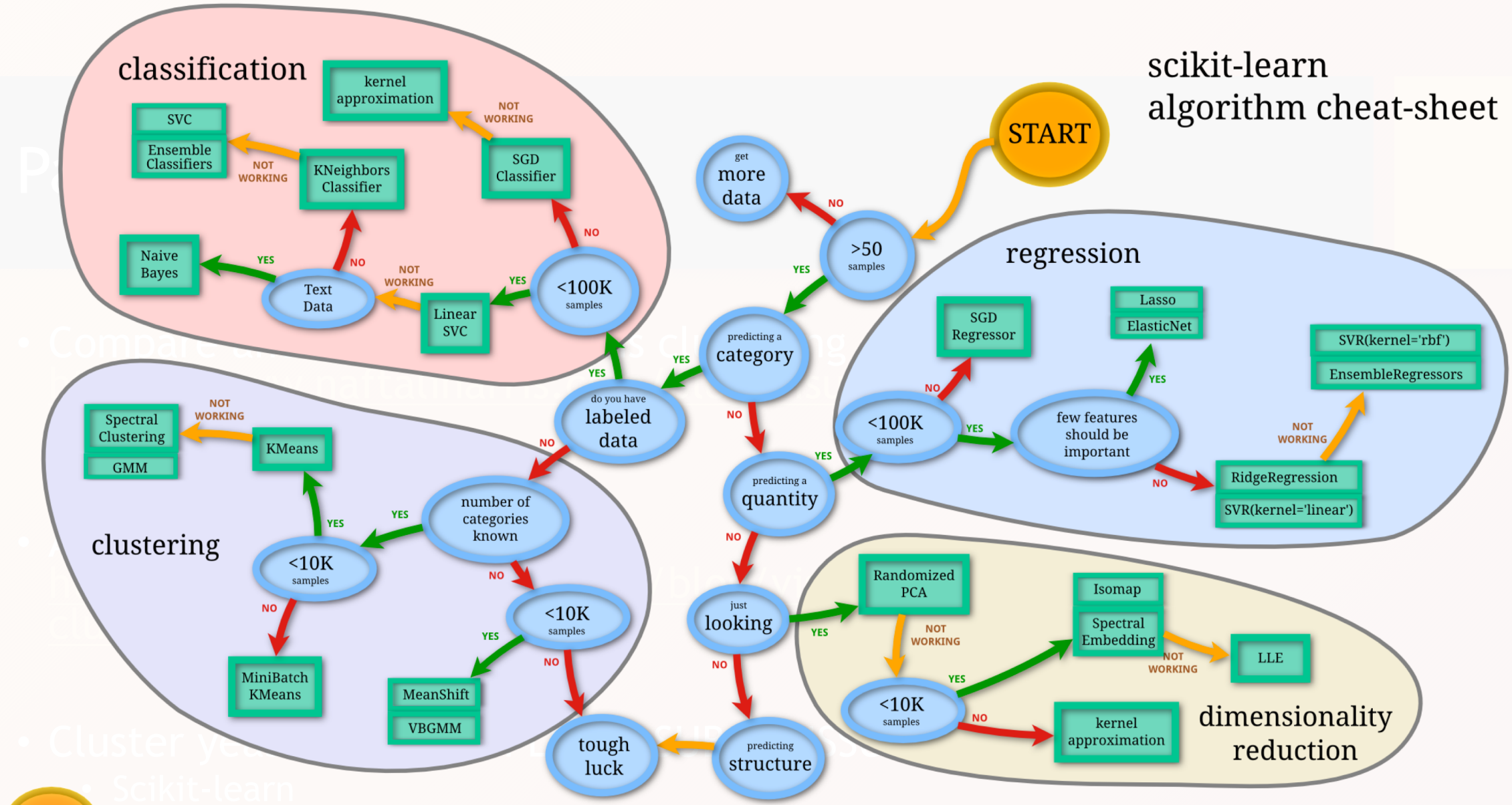
Clustering Algorithms

- Hard vs Soft – can group assignments have uncertainty?
- Flat vs. Hierarchical
- Graph-based



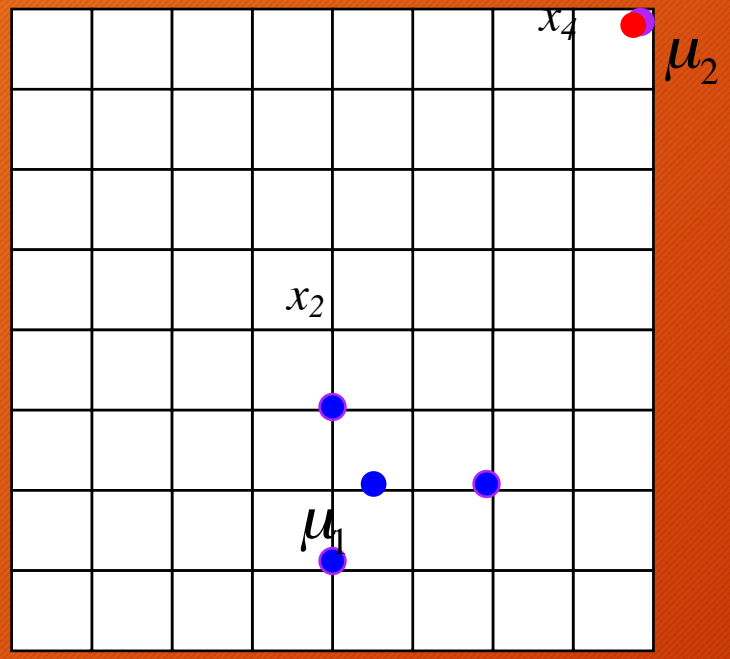


D'haeseleer, Nature Biotechnology, 2005

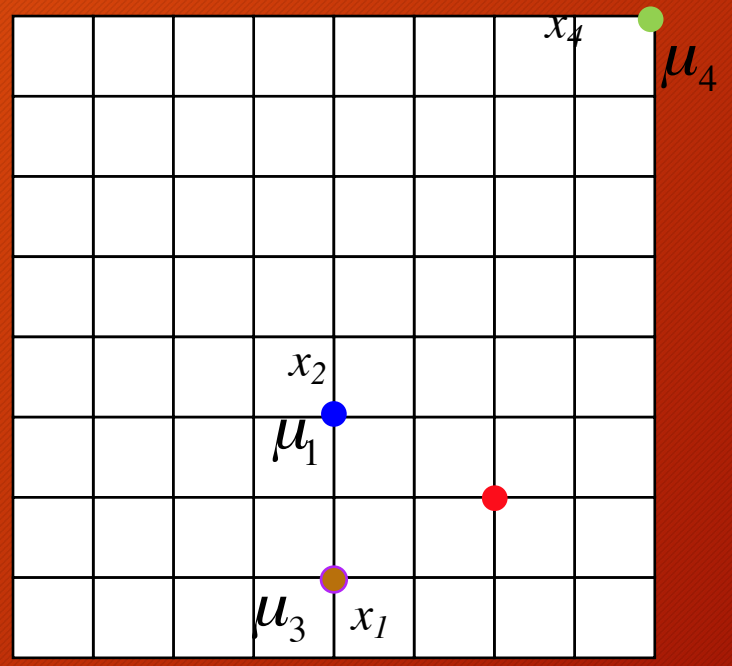

scikit-learn
algorithm cheat-sheet
START
classification
kernel approximation
NOT WORKING
SVC
Ensemble Classifiers
NOT WORKING
KNeighbors Classifier
SGD Classifier
NO
Naive Bayes
YES
NO
Text Data
NOT WORKING
YES
<100K samples
Linear SVC
get more data
NO
>50 samples
YES
regression
SGD Regressor
Lasso
ElasticNet
SVR(kernel='rbf')
EnsembleRegressors
predicting a category
YES
NO
<100K samples
YES
few features should be important
YES
NOT WORKING
RidgeRegression
SVR(kernel='linear')
NO
do you have labeled data
YES
NO
clustering
Spectral Clustering
GMM
NOT WORKING
KMeans
YES
number of categories known
YES
<10K samples
NO
MiniBatch KMeans
NO
<10K samples
YES
MeanShift
VBGMM
NO
predicting a quantity
YES
NO
just looking
YES
Randomized PCA
NOT WORKING
Isomap
Spectral Embedding
NOT WORKING
LLE
YES
<10K samples
NO
kernel approximation
dimensionality reduction
NO
predicting structure
tough luck
Back
scikit learn

Selecting number of clusters

K=2



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

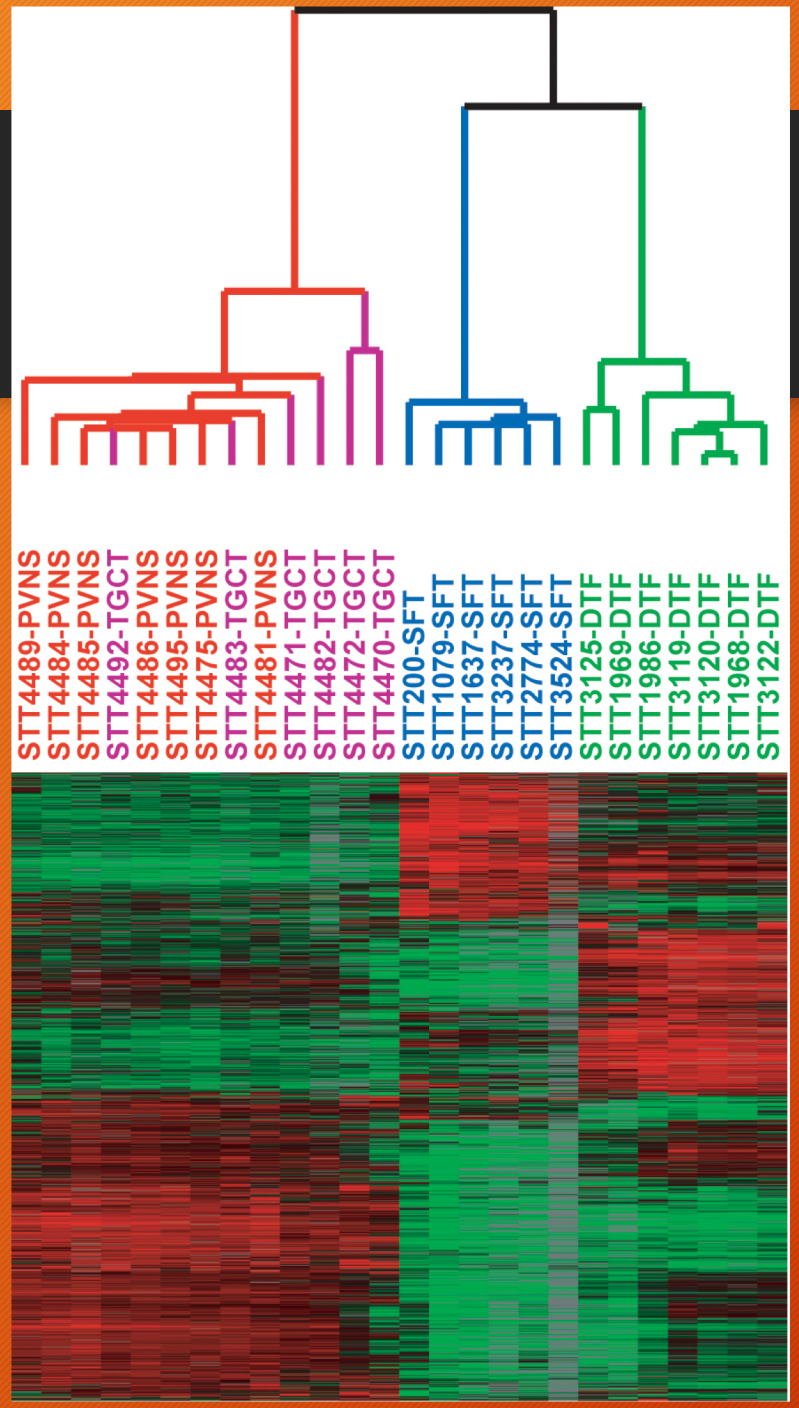
C. Provide a model that accurately represents given examples of a task

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 (bias, imbalance)
- 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 – penalize higher values of k
- Hold aside data, see which models best “fit” the unseen data



Clustering Example

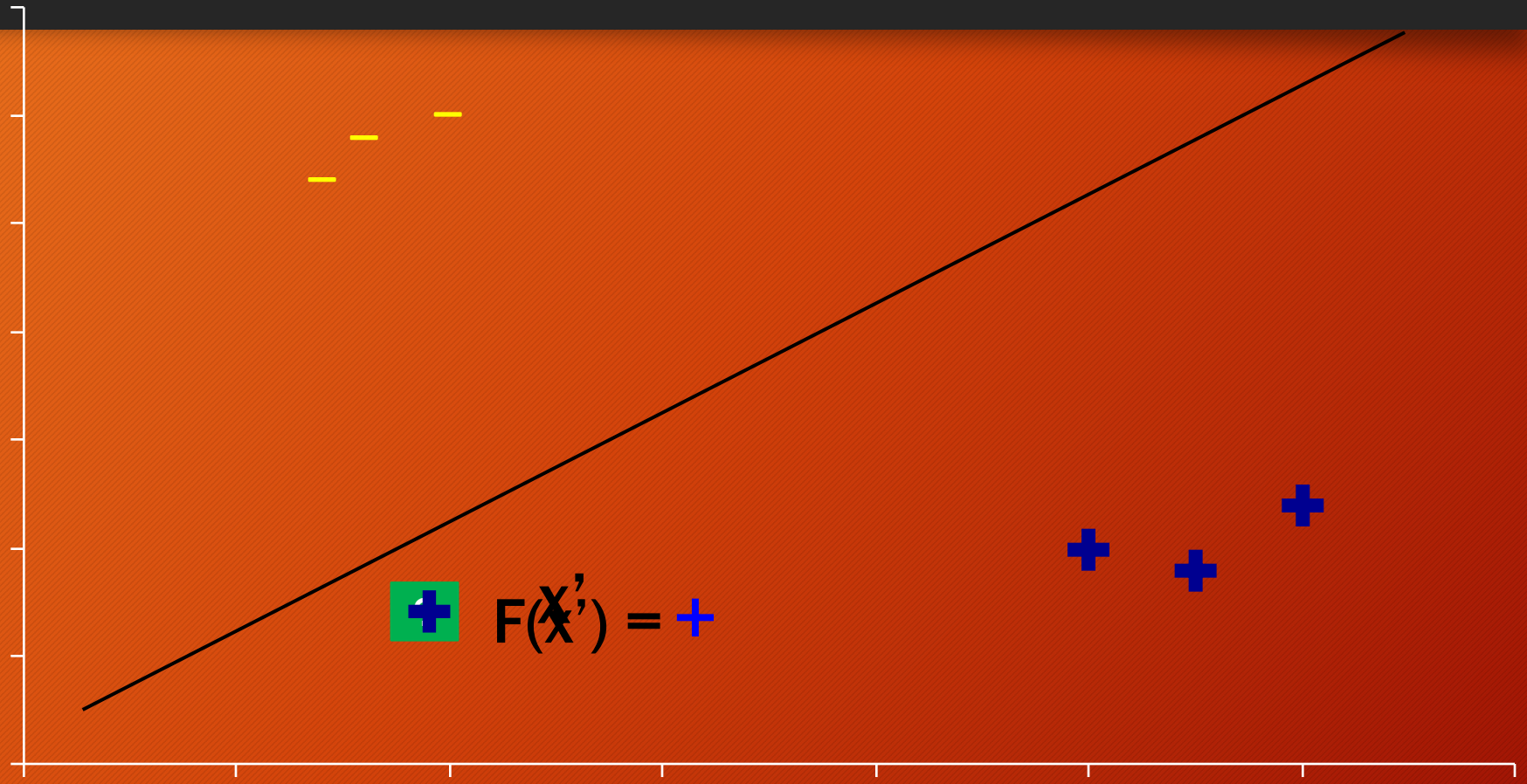
- clustering of related cancers and an inflammatory disorder
  - *TGCT*: Tenosynovial giant-cell tumor (purple)
  - *PVNS*: pigmented villonodular synovitis (red)
  - *SFT*: solitary fibrous tumor (blue)
  - *DTF*: desmoid-type fibromatosis (green)

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

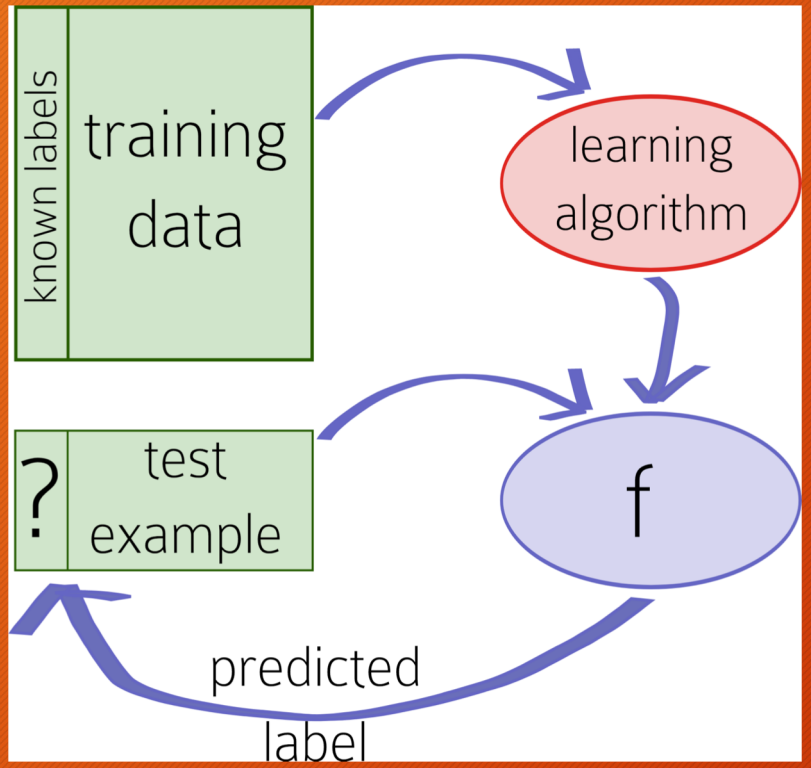
Outline

- Clustering Algorithms
  - K-Means algorithm
  - Application: gene expression
- Classification Algorithms
  - Application: disease prediction
  - Application: sequence annotation

Supervised Learning Example



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?
+
+
-
-
+

Task 2 Exercise

- Predict Colon Cancer samples using machine learning

- Default: logistic regression
- Alternative: pick any classifier (I recommend random forests)

- Extensions:
  - Tuning model parameters: the bias-variance dilemma
  - Proper evaluation methodology
  - Interpreting models

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!

"Deep Learning" is only part of this picture

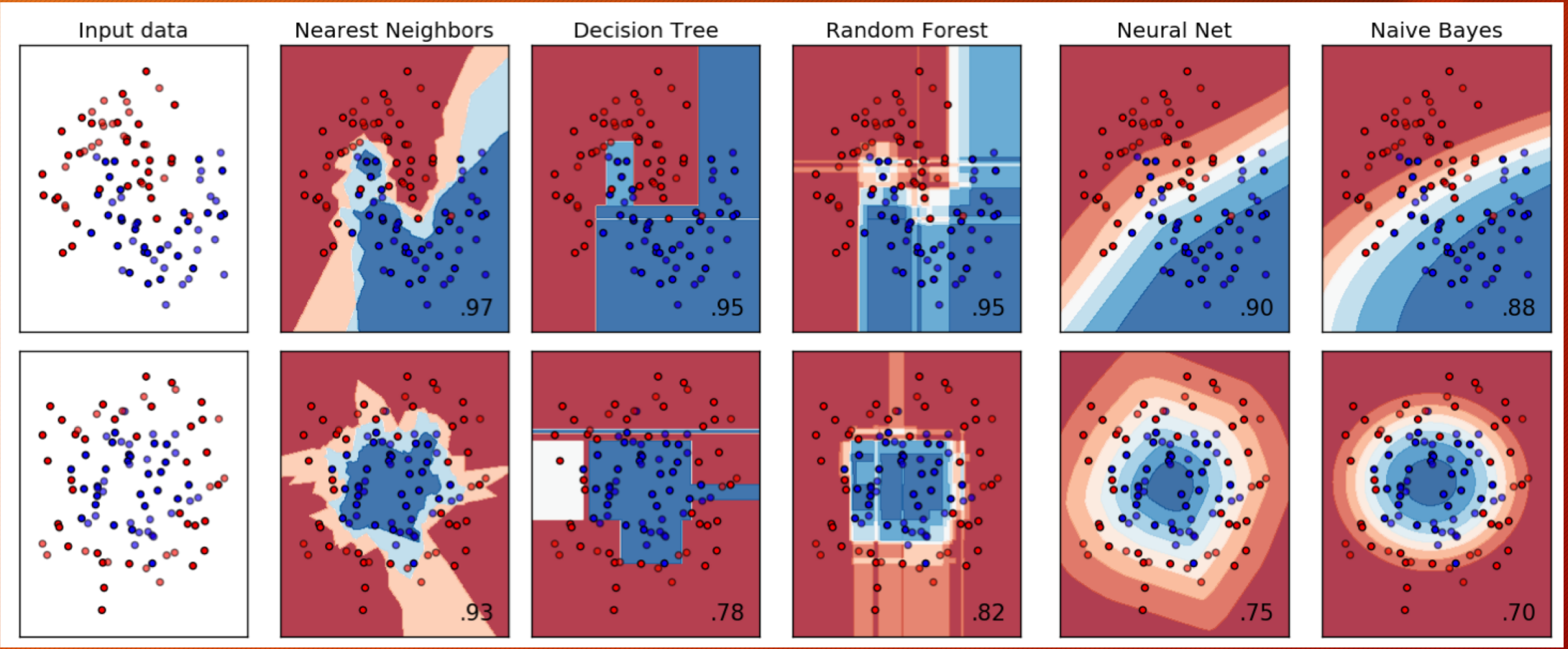
Algorithmic Design Choices

No Free Lunch

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

Decision Surface

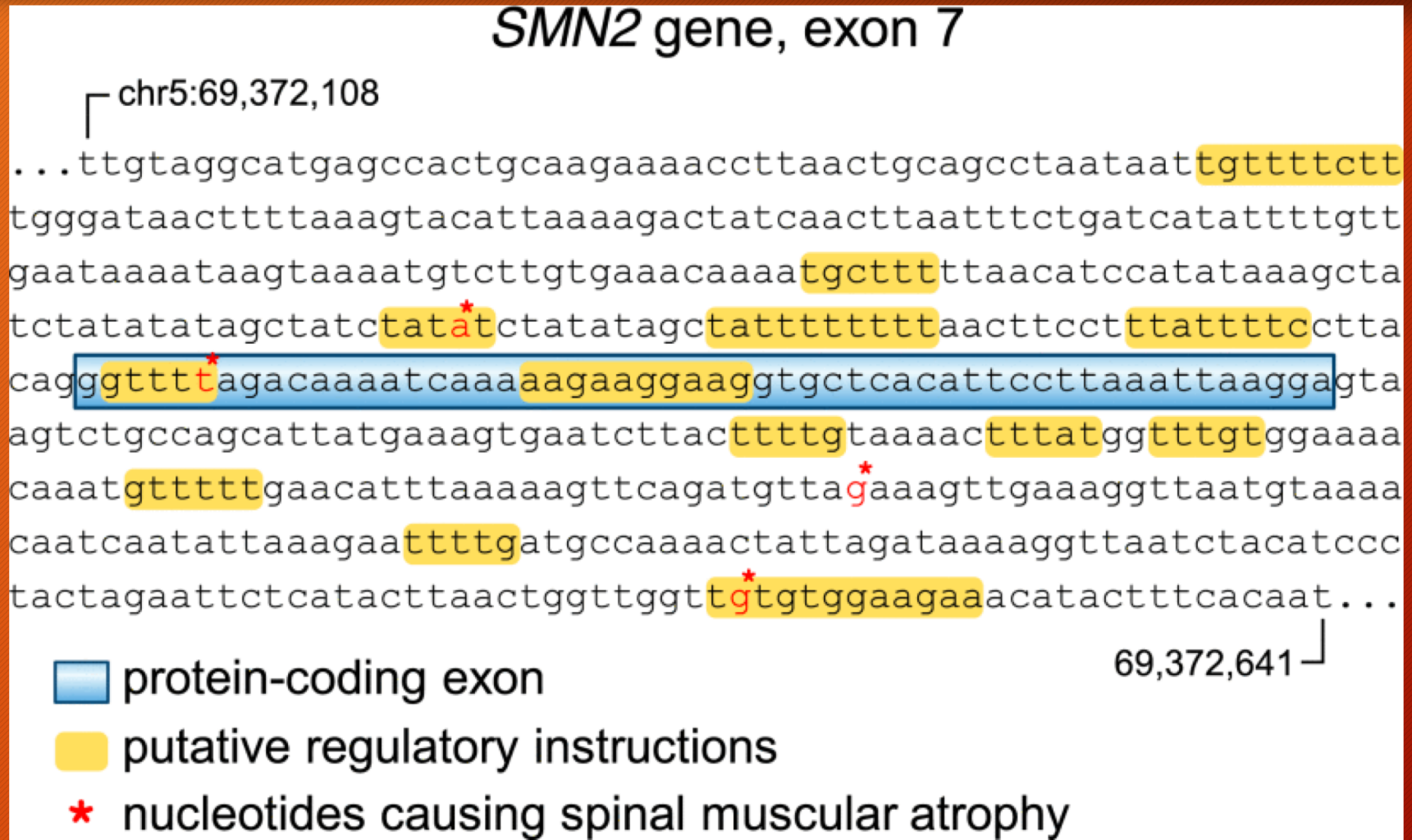


scikit-learn.org

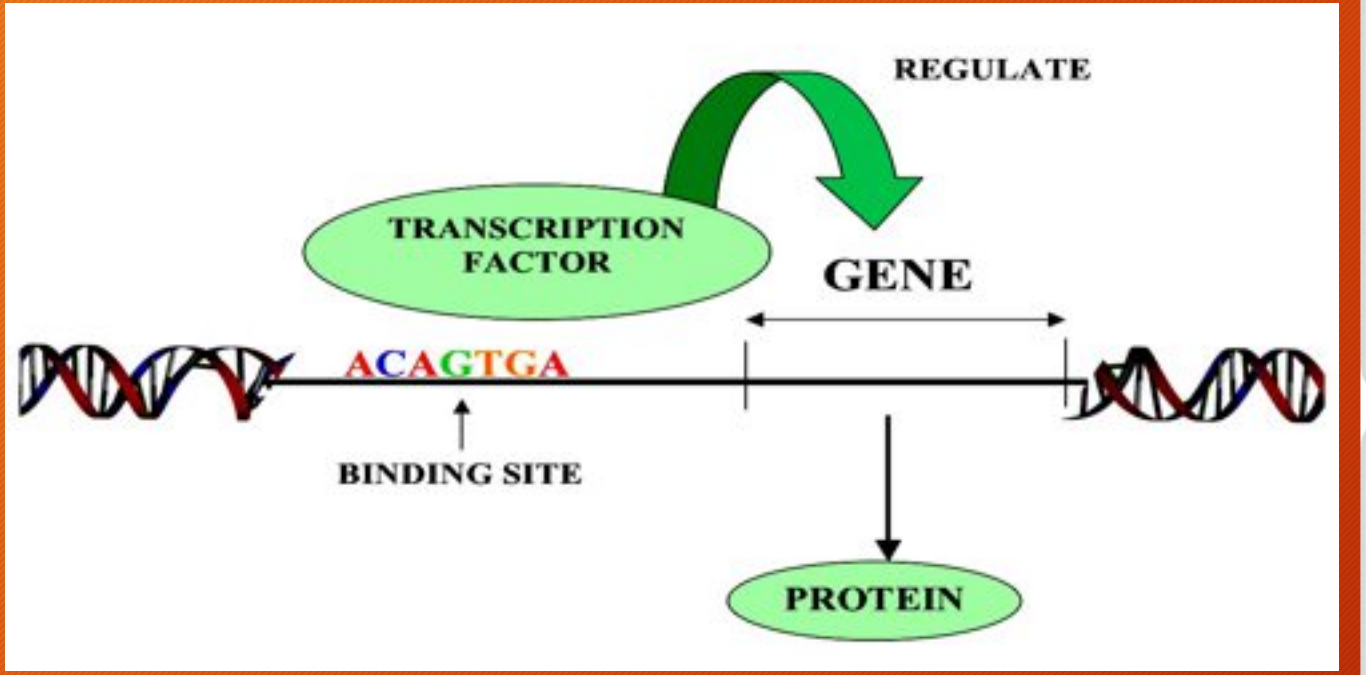
Feature Representation

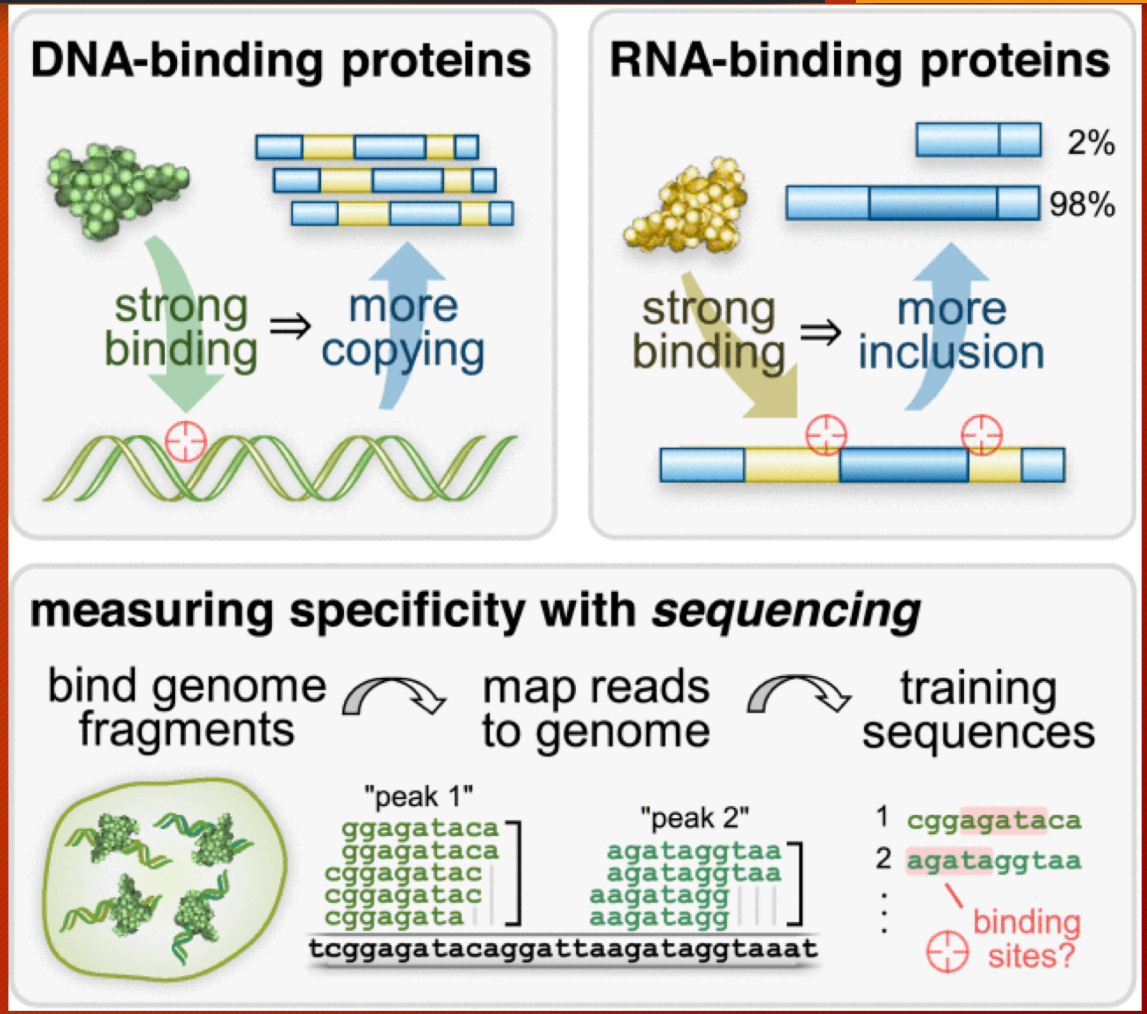
- To machine learning, all data is just a matrix
- Codification of data (e.g., DNA sequence) is crucial design choice

Application: Sequence Analysis



Protein Binding Motifs





Feature Representation

Seq1	**AGGCTTTAACCAGATT…**
Seq2	GAGACTTAACCAGCCT…
Seq3	CCCGGTTAACCAGTGG…
…	…

Rule: TTAACCAG starting at sequence positions 5 → predict this is a binding site

Feature Representation

Seq1	**AGGCTTTAACCAGATT…**	**Rule predicts TRUE**
Seq2	GCCAAATCCAGGAGACTTAACCAGCCT…	Rule predicts FALSE
Seq3	TTAACCAGTGGCCGTAATC…	Rule predicts FALSE
…	…	

Rule: TTAACCAG starting at sequence positions 5 → predict this is a binding site

Problem: We don't know the binding site before hand, can't align binding region

Our representation needs to **translation invariant**

Solution 1: K-mer Matrix

AAAAAATCGTTATAGGCGCTAGATCGT…

CAAGCGAGTCTAATCGAATAAAAAAG…

AAAAAA	AAAAAT	AAAAAC	AAAAAG	AAAATA	
[1	1	0	0	0	….
[1	0	0	1	0	….

Simple Neural Network

Input Sequence

AAAAAA

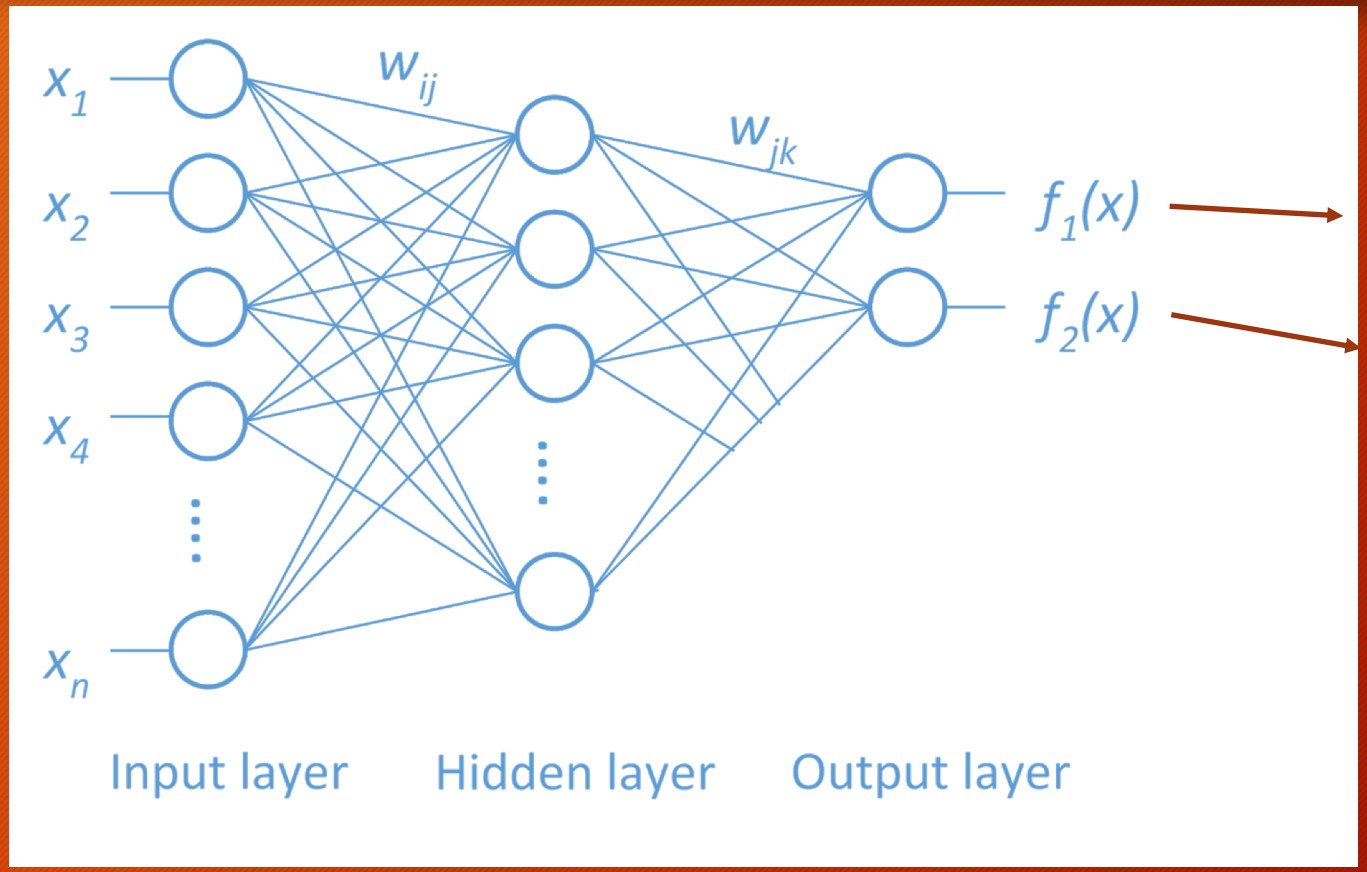
AAAAAT

AAAAAC

AAAAAG

⋮

GGGGGG



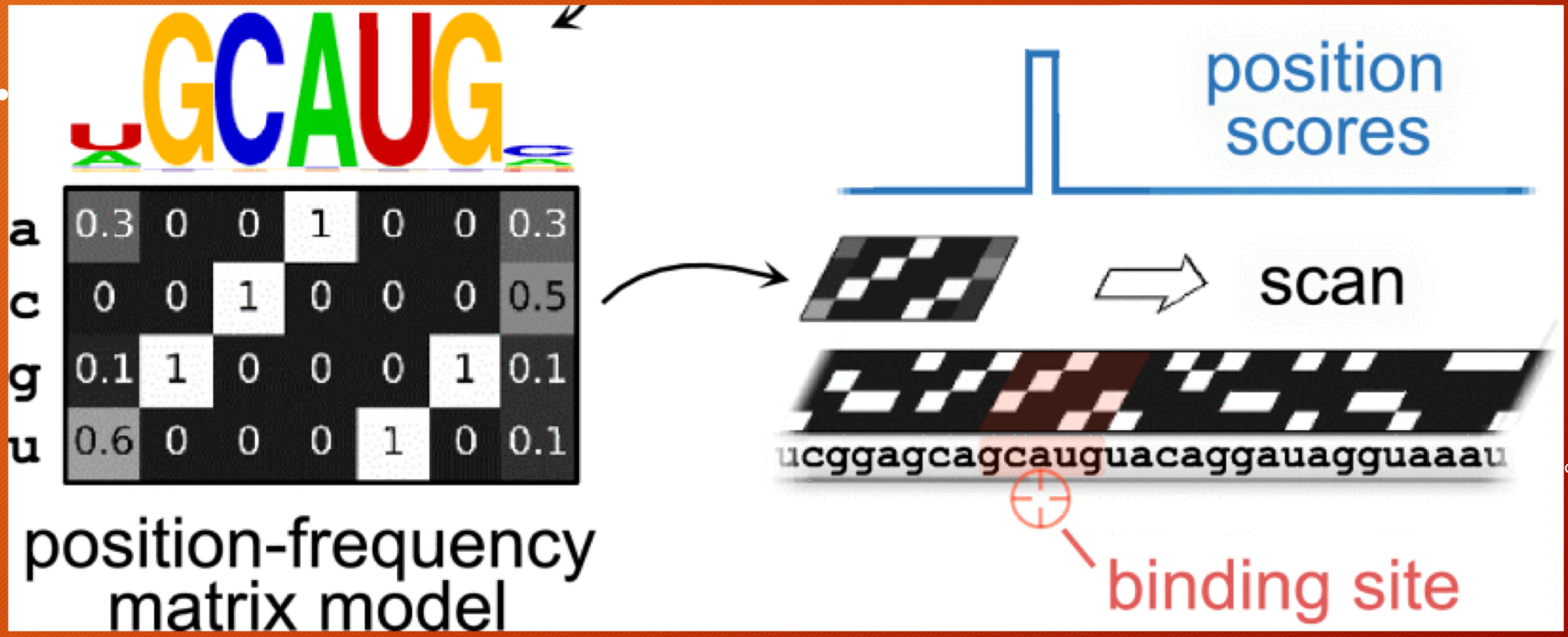
Prediction

$f_1(x)$ → TF bound score

$f_2(x)$ → TF **not** bound score

Alternative: Convolution Neural Network

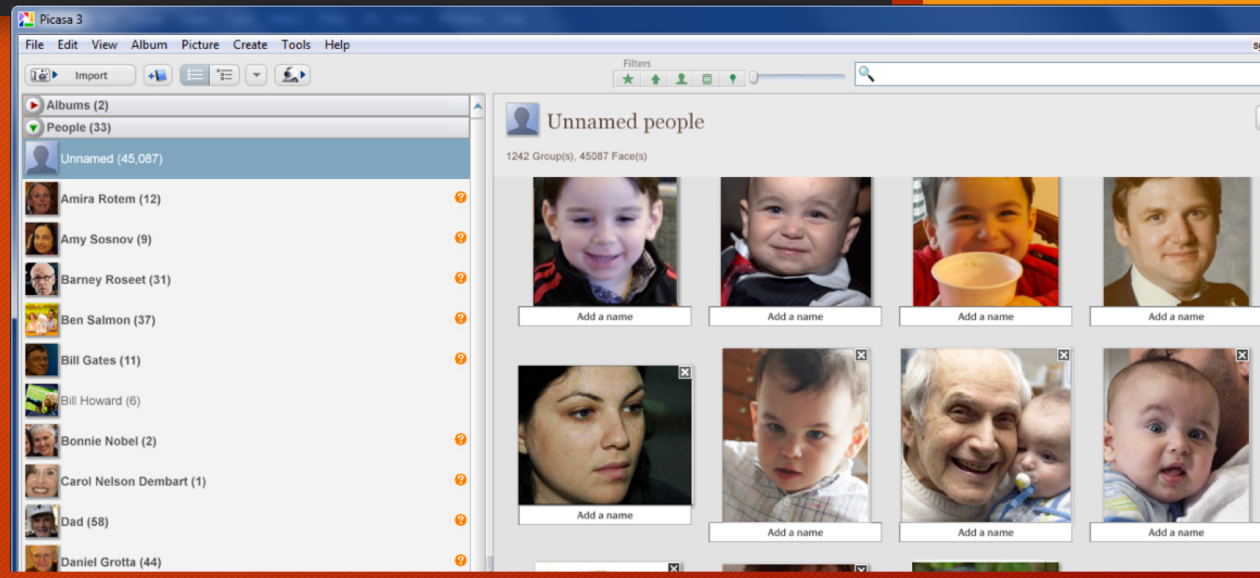
- Convolution: learn a small network and slide it over the data
-



ding_no_strides.gif

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)

Summary

- Machine learning is an important tool for computational biology
- Utilizing ML requires is more just picking the “hot” algorithm
  - Framework
  - Representation
  - Evaluation methodology
  - Bias

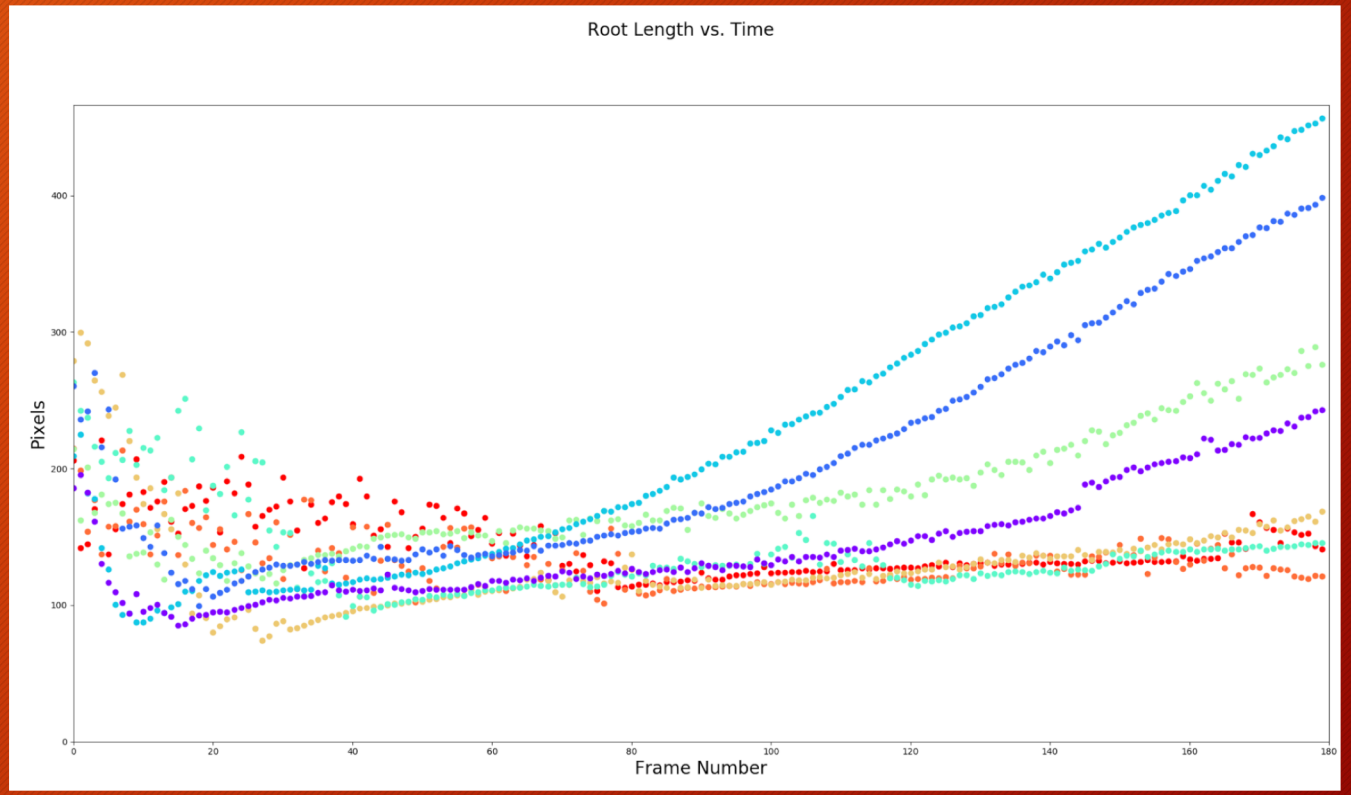
My Research

- Deep Learning Algorithms
- Probabilistic Graphical Models

- Image analysis (Alzheimer's disease, Chest X-Ray)
- Sequence analysis (TF binding prediction)
- Machine learning methodology (weak supervision, feature elicitation)

Plant Reaction to Heat

https://www.dropbox.com/s/r9tnpi5tz0v1twi/rootTracerExample.mpg?dl=0



Conclusions

- “Big data” necessitates improved machine learning approaches
- Biomedical applications have important implications for society
- Biomedical data is a particular good test bed for analyzing new approaches
- Thank you for attending!