Fundamentals of Machine Learning Workshop

Alexander Ioannidis
ioannidis@stanford.edu

Gabriel Maher
gdmaher@stanford.edu

Institute for Computational and Mathematical Engineering,
Stanford University

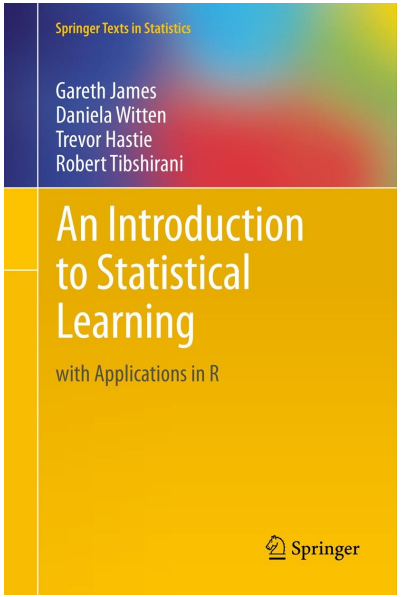
Workshop Goals

- Overview of well-known machine learning techniques
- Practical tips, good practices, and pitfalls
- Intuition about what works when, and why

Useful References

- *"An Introduction to Statistical Learning with Applications in R" (ISL)* by James, Witten, Hastie and Tibshirani*
  - available online (pdf): www-bcf.usc.edu/~gareth/ISL/

- *"The Elements of Statistical Learning" (ESL)* by Hastie, Tibshirani and Friedman
  - available online (pdf): statweb.stanford.edu/~tibs/ElemStatLearn/



*Some of the figures in this presentation are taken from *"An Introduction to Statistical Learning, with applications in R"* (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani

Introduction to Machine Learning

- What is machine learning?
  - A set of tools for understanding data by building models *from data*
  - Methods for *automatically* learning and recognizing complex patterns from data

Applications of Machine Learning

Automating Complex Tasks:

Self-driving Cars

http://driving.stanford.edu/

Automated decision-making:

Credit Card Fraud Detection

Self-customizing algorithms:

Movie Recommendations

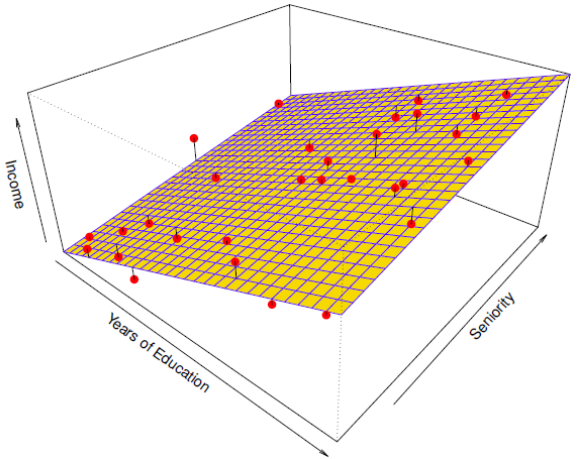
Questions?

Machine Learning

- Two general categories of learning problems:
  - Supervised Learning
  - Unsupervised Learning

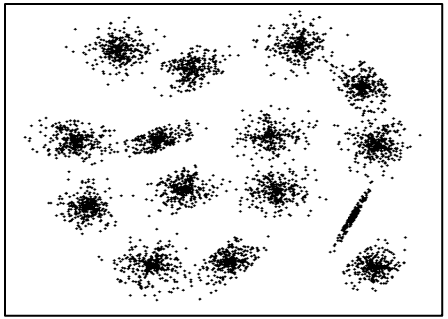
Supervised Learning

- For training data, both the input variable and the associated response are available
  - $X^{(i)}$ and associated $Y^{(i)}$ are available to learning algorithm for training
- Goal: *generalize* to new data

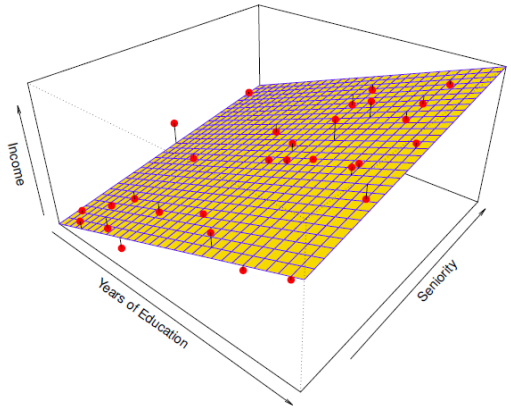


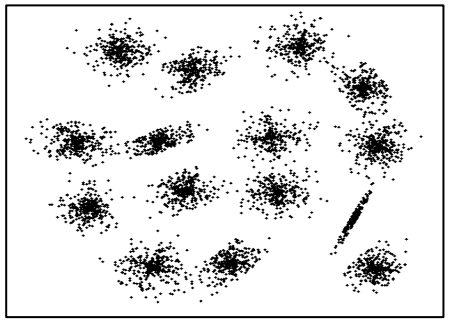
Unsupervised Learning

- Measurements available for each observation, but no associated response
  - $X^{(i)}$ available but not $Y^{(i)}$
- Goal: *understand relationships* between variables or among observations



Types of Algorithms





Supervised Learning: Classification and Regression

- Supervised learning problems can be divided into *classification* and *regression*

Supervised Learning: Classification and Regression

- *Regression*: the output is quantitative (continuous / numerical / ordered)

  e.g. predicting

  - the value of stock Z one year from now
  - a person's income based on demographic factors

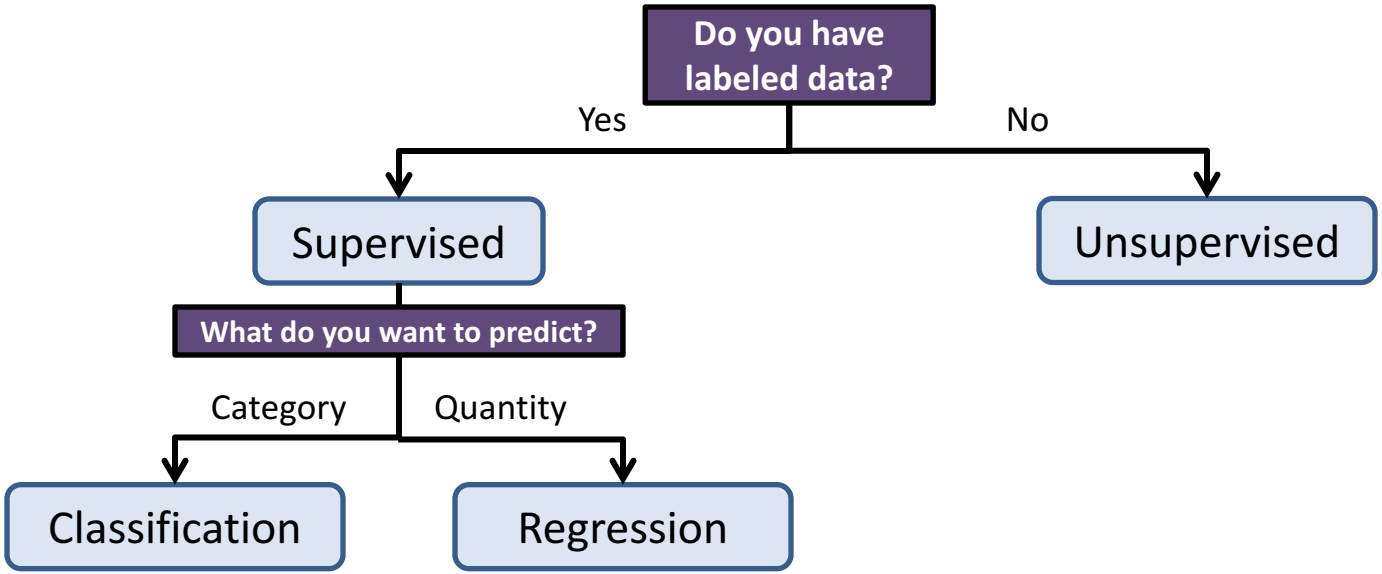
Supervised Learning: Classification and Regression

- *Classification*: the output is qualitative (categorical)

  e.g. predicting

  - whether the value of stock Z will have increased or decreased one year from now
  - whether a credit card transaction is fraudulent or legitimate

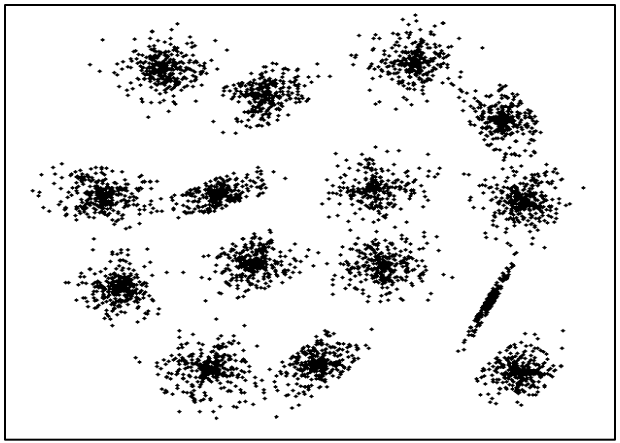
Types of Algorithms



Unsupervised Learning:
Clustering & Dimensionality Reduction

- *Cluster analysis*

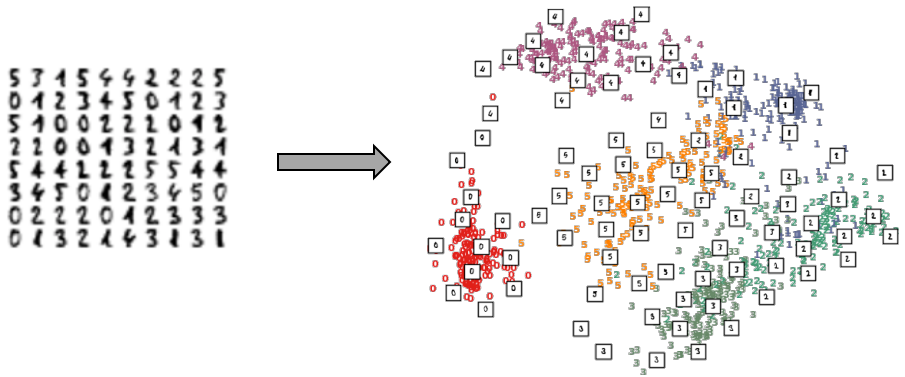
  partition data into subsets that share common characteristics



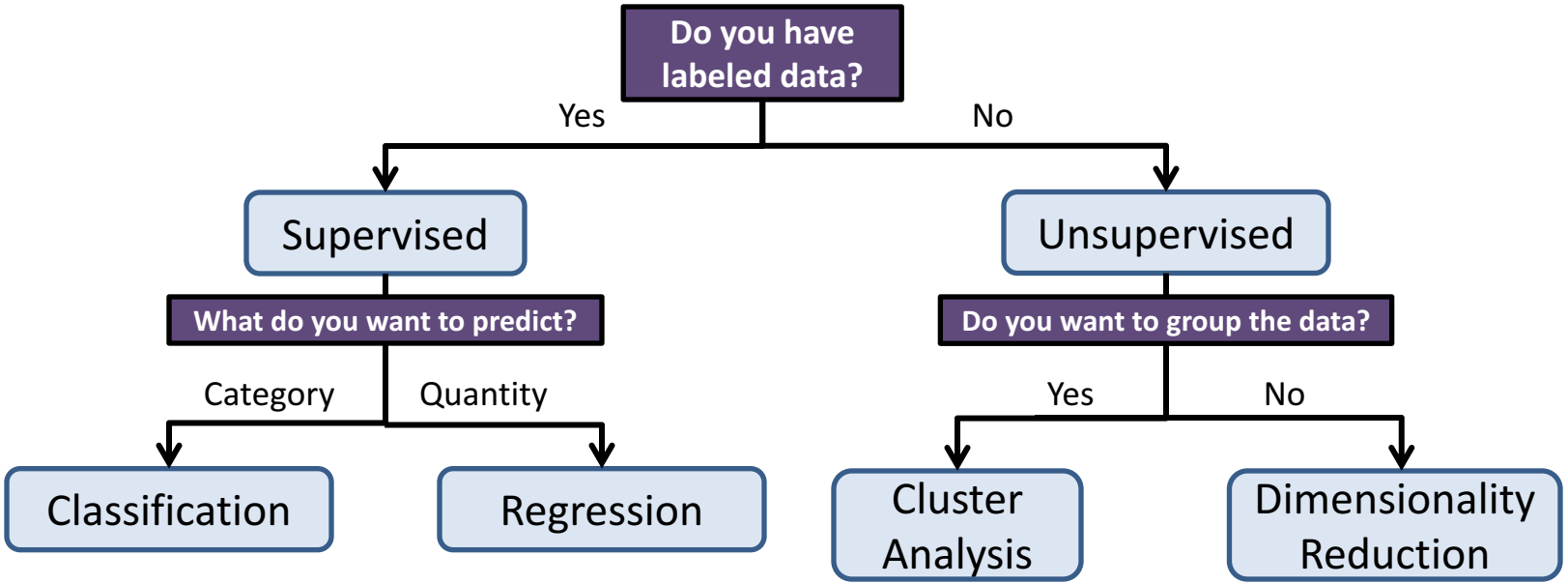
Unsupervised Learning:
Clustering & Dimensionality Reduction

- *Dimensionality reduction*

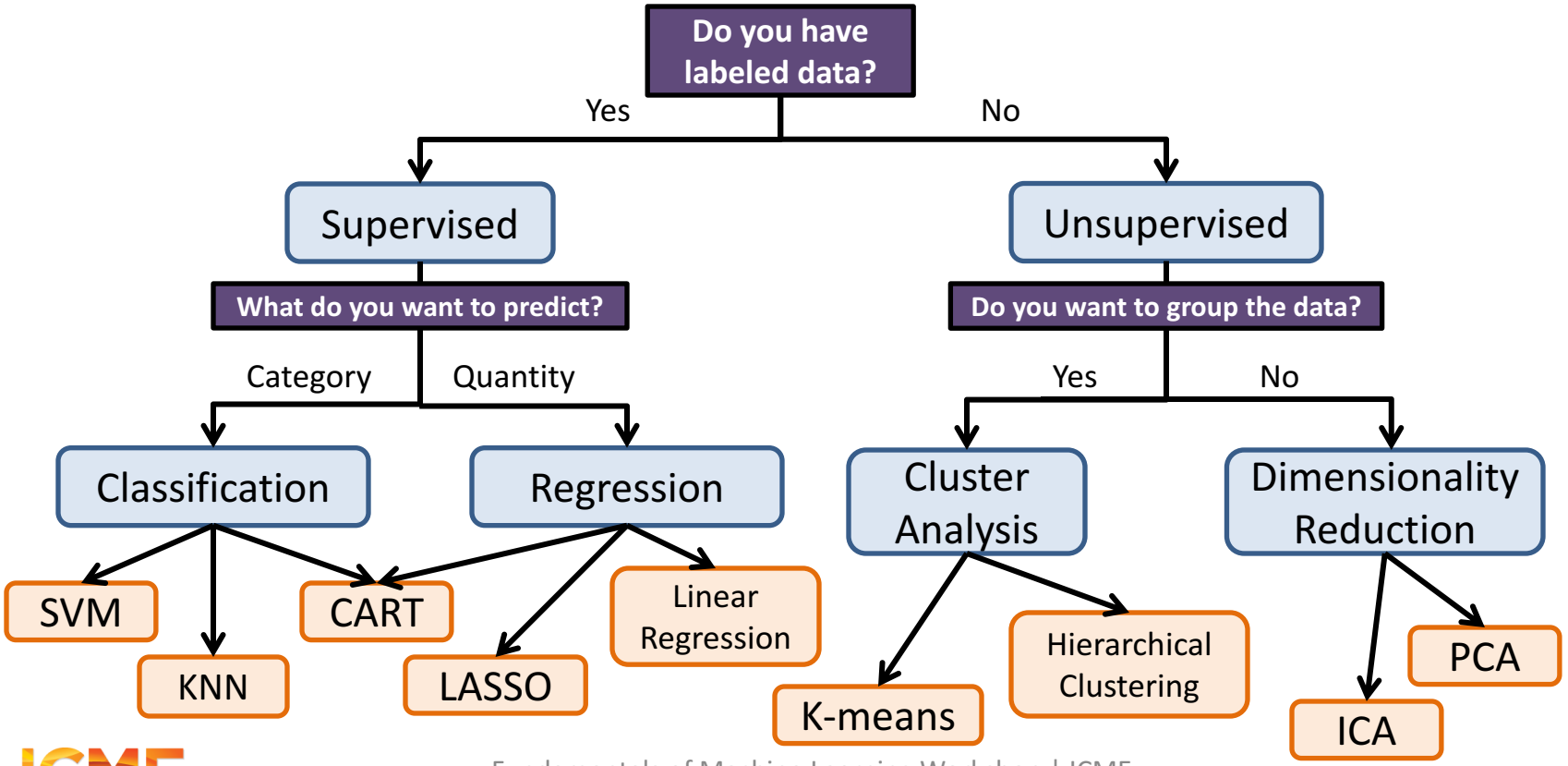
  create new features from original inputs that retain important information



Types of Algorithms



Machine Learning Algorithms



“Best” Machine Learning Algorithm

- Bad news: no algorithm is the best
  - No machine learning algorithm will perform well on every task
- Good news: all of them are the best
  - Each machine learning algorithm will perform well on some task
- “No free lunch” theorem
  - Wolpert (1996): all algorithms perform equally when averaged over all possible problems

Trade-offs in Machine Learning

- Bias vs. variance
- Accuracy vs. interpretability
- Accuracy vs. scalability
  - Some models / algorithms for computing them may not scale to large data sets
- Domain-knowledge vs. data-driven
- More data vs. better algorithm

Trade-off: Accuracy vs. Interpretability

- Methods vary in terms of model flexibility
  - Models that are too restrictive may have poor accuracy
  - e.g. Linear regression is restrictive – can't capture non-linear effects
- Why choose a more restrictive model?
  - More interpretable – advantage for inference
  - Simpler models can be more accurate (less risk of over-fitting)
- For prediction, interpretability is not always required
  - Prediction model can be a black box

Trade-off: Accuracy vs. Interpretability

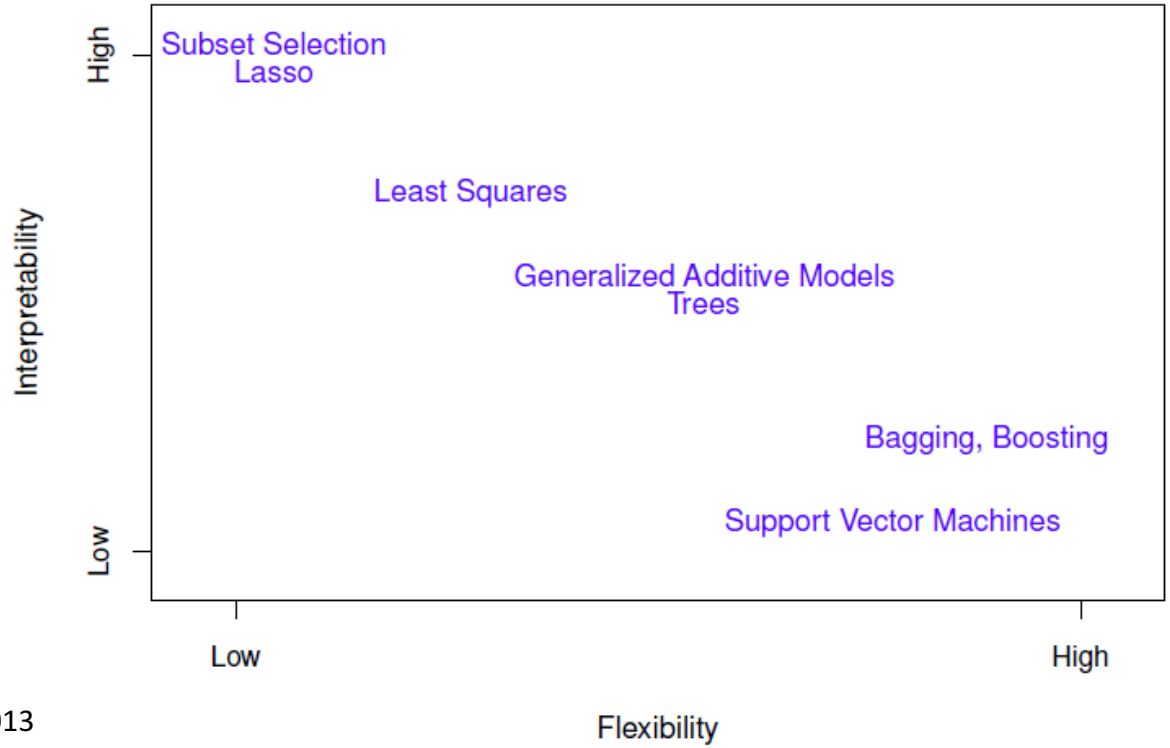


Figure 2.7 , ISL 2013

Trade-off: Bias vs. Variance

Knn Example

Figure 2.7 , ISL 2013

Choice of K (KNN classifier)

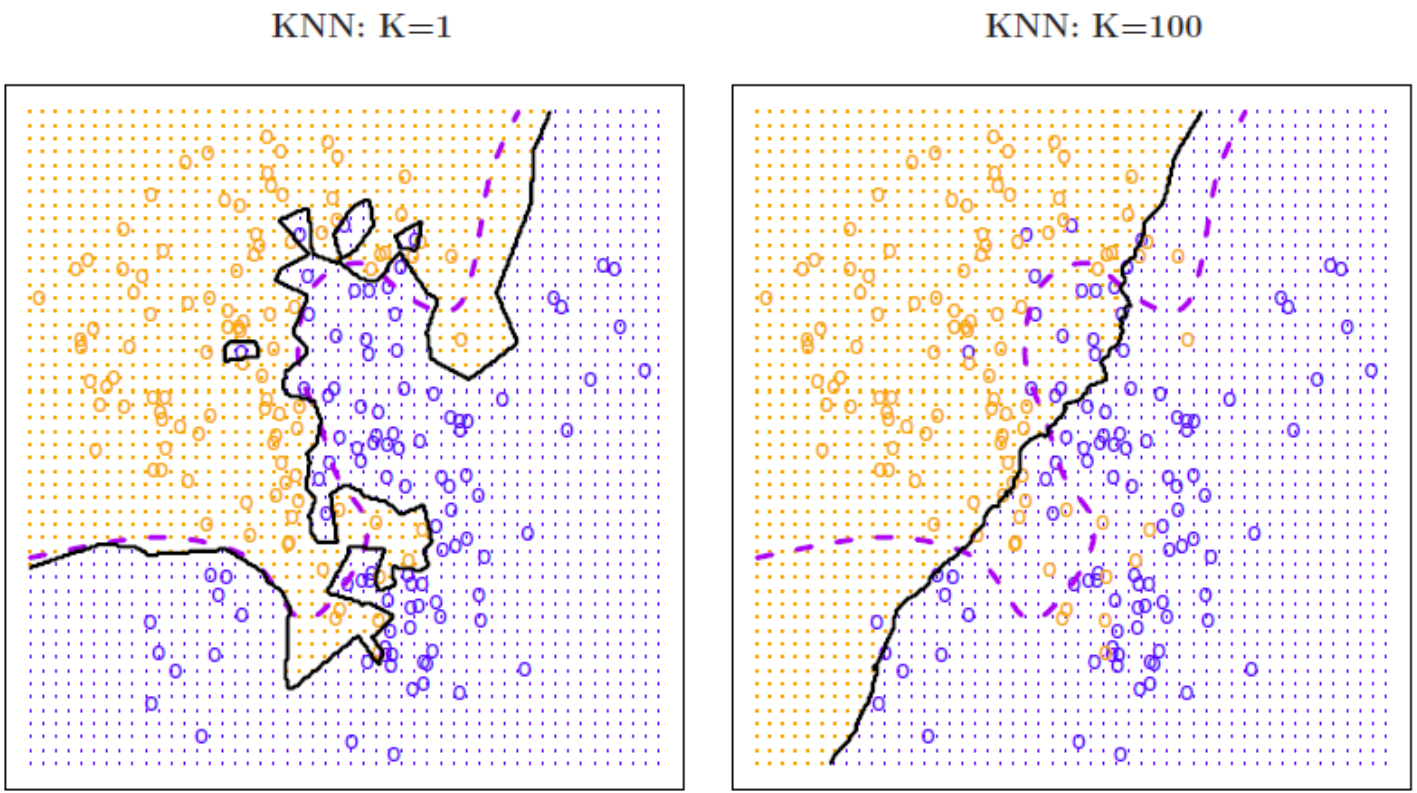


Figure 2.16, ISL 2013

Choice of K

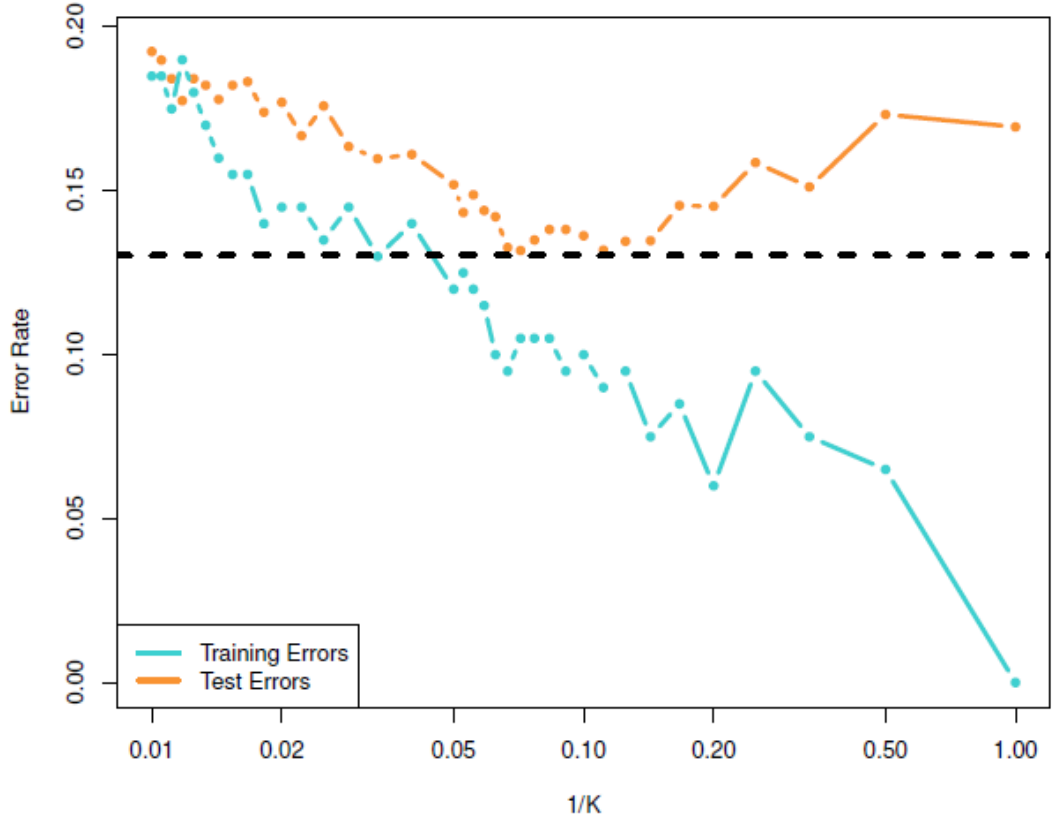


Figure 2.17, ISL 2013

Choice of K (KNN classifier)

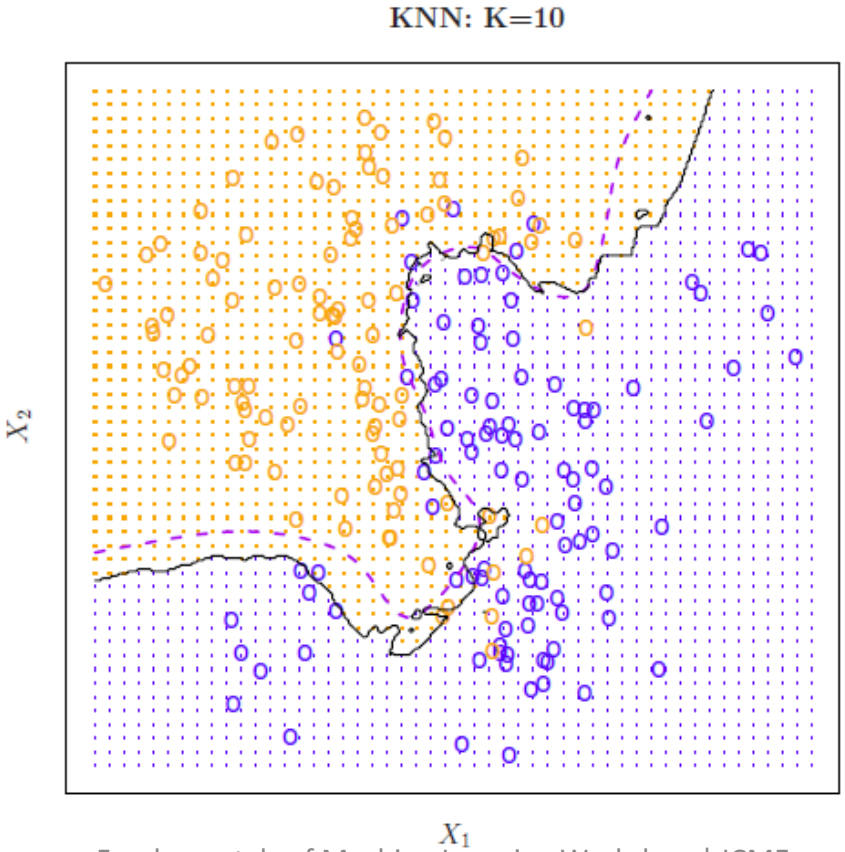


Figure 2.15, ISL 2013

K-Nearest Neighbor classifier (KNN)

- Advantages:
  - Simple to implement
  - Few tuning parameters (K, distance metric)
  - Flexible, classes do not have to be linearly separable

- Disadvantages:
  - Computationally expensive
  - Sensitive to imbalanced datasets
  - Sensitive to irrelevant inputs

Programming

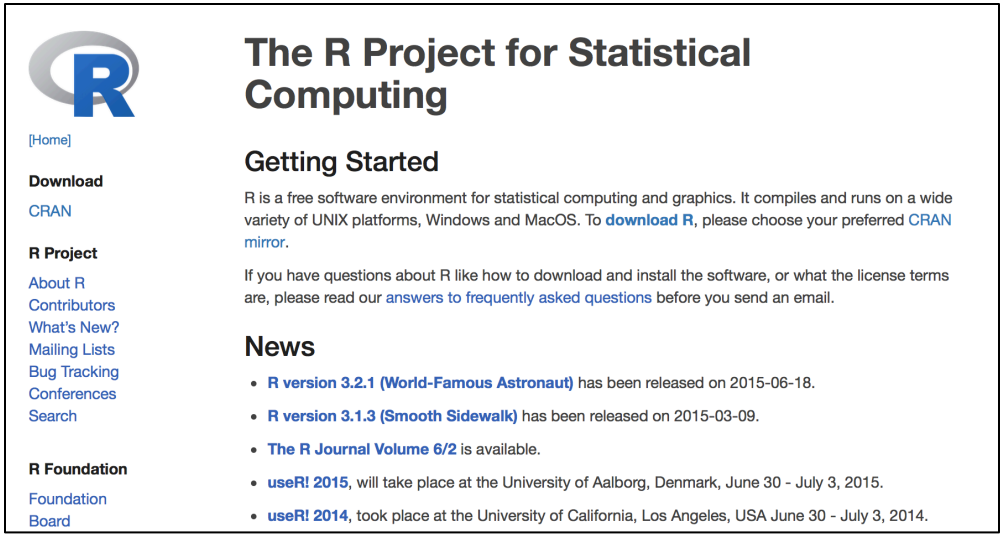
Figure 2.7 , ISL 2013

R/Python Programming

- Recommended programming languages with existing libraries for machine learning:
  - ***R***: tailored for statistical analysis
  - ***Python***: general-purpose programming language

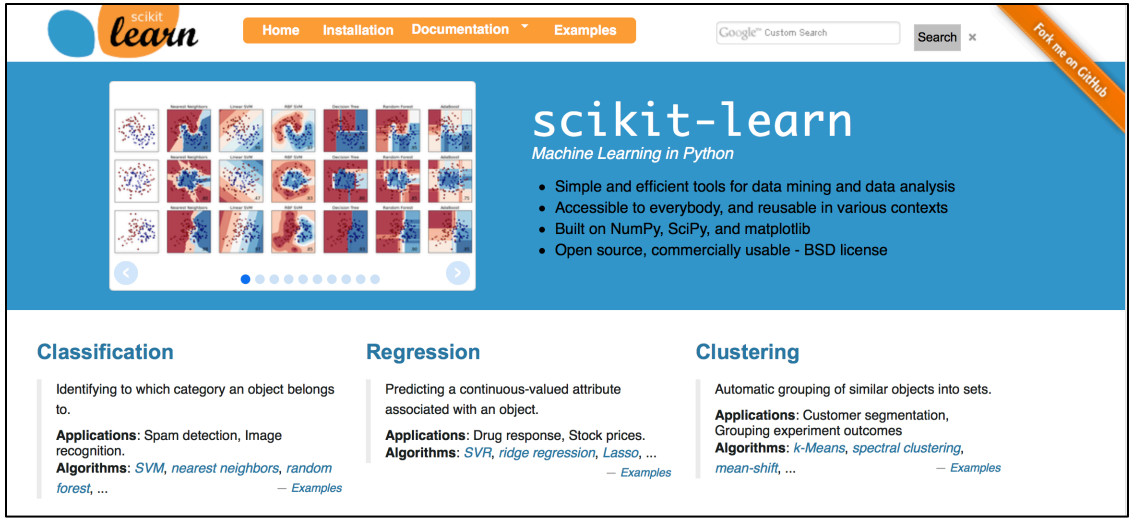
R/Python Programming

- Download and documentation
  - R: www.r-project.org



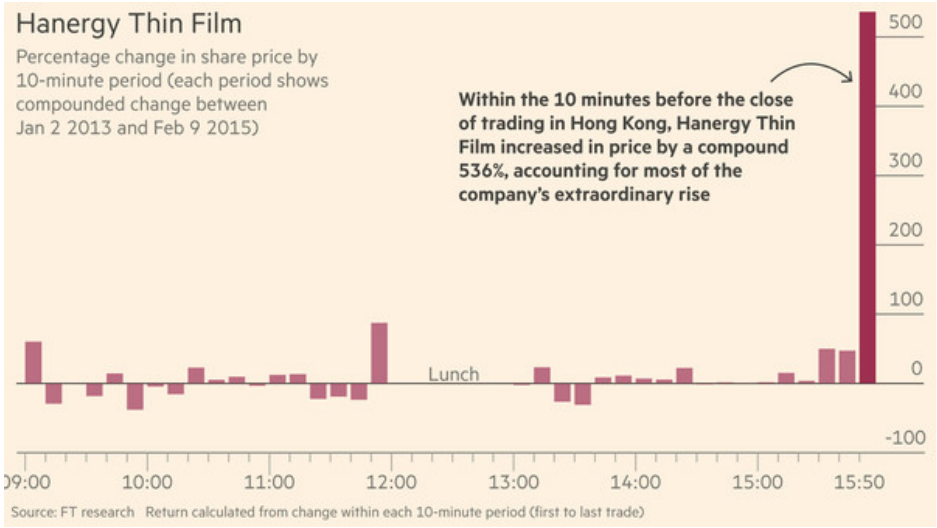
R/Python Programming

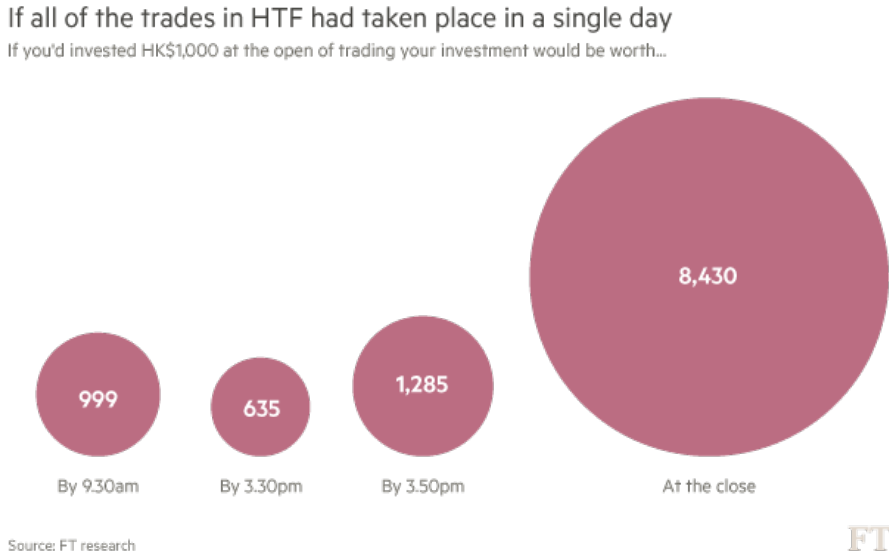
- Python: www.python.org
  - scikit-learn: http://scikit-learn.org/



Questions?

R



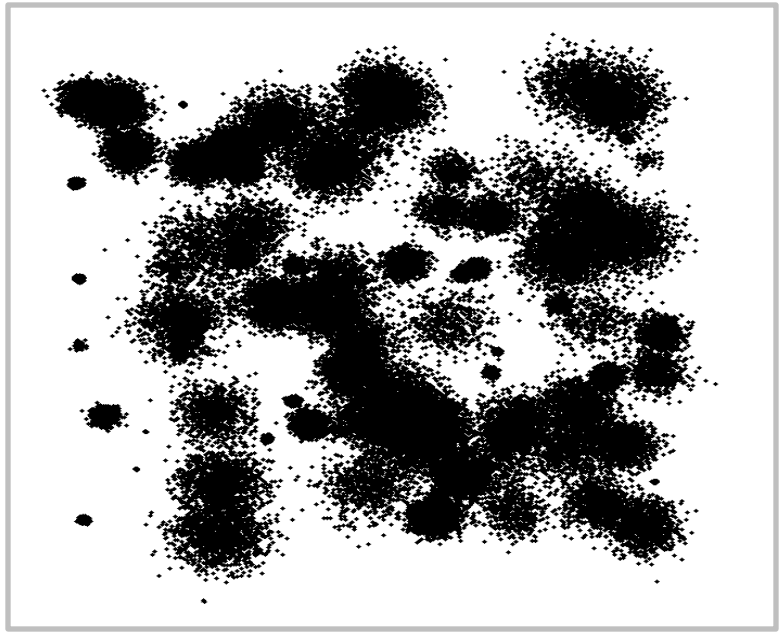


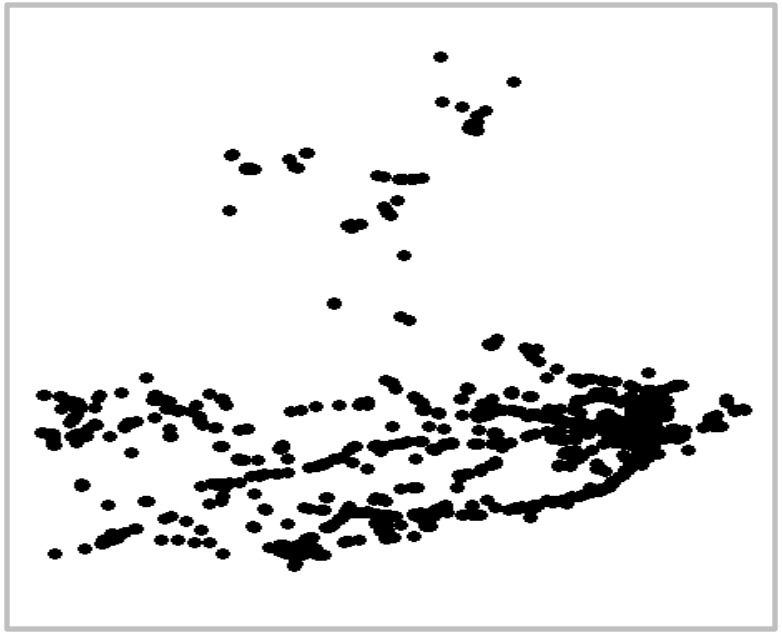
The FT used the statistical programming language R as well as SQL in its analysis.

Unsupervised Learning

Clustering & Dimensionality Reduction

Clustering





Data sets from: http://cs.joensuu.fi/sipu/datasets/

Hierarchical Clustering

- Hierarchical clustering
  - Clusters based on distances between observations
  - Represented as a hierarchy rather than a partition of data

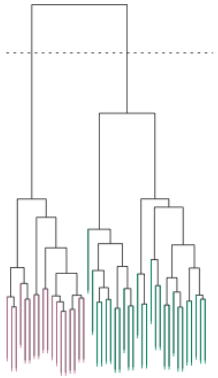
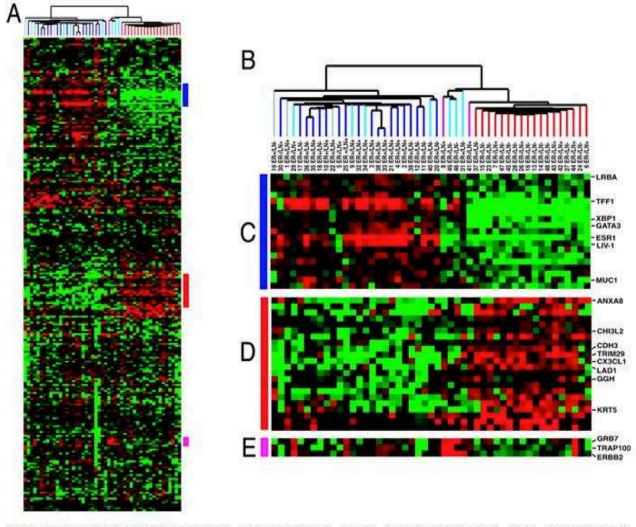
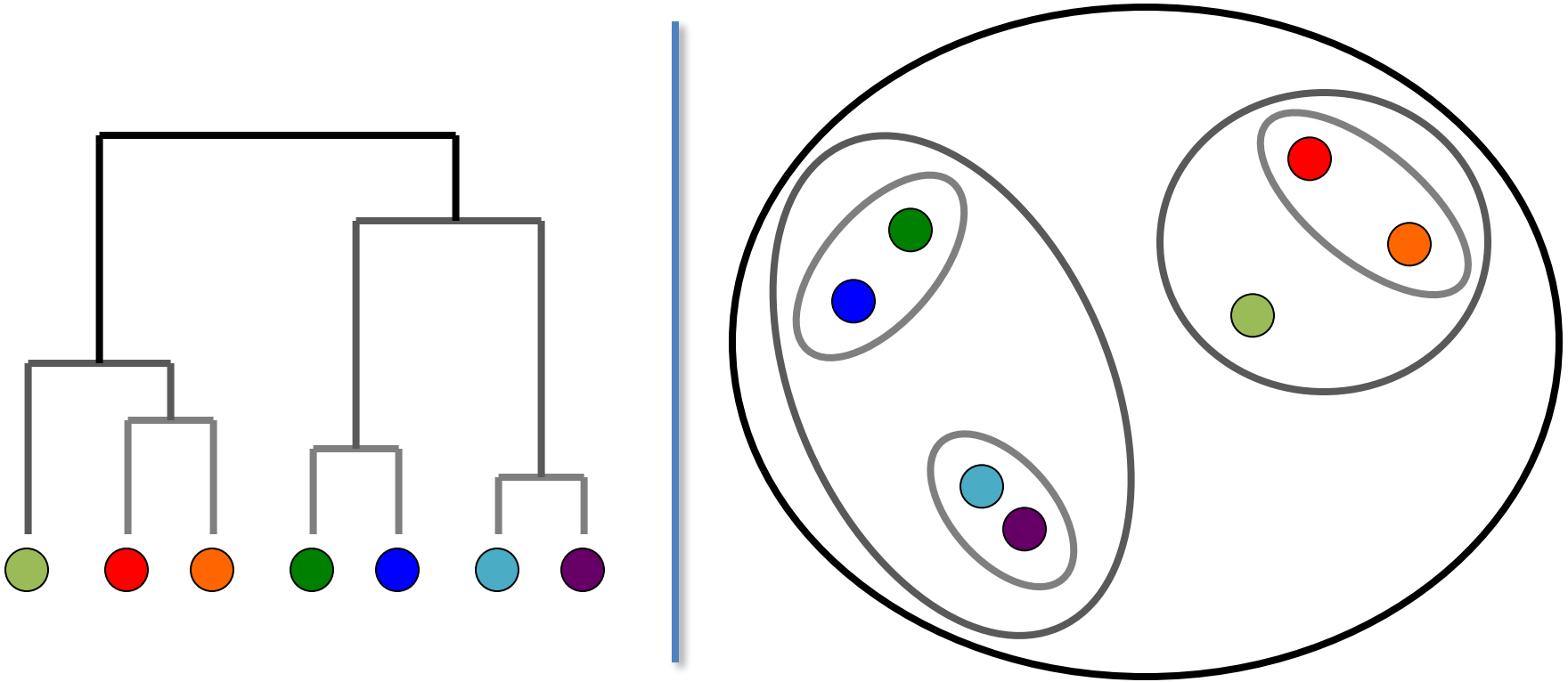


Figure 10.9 , ISL 2013

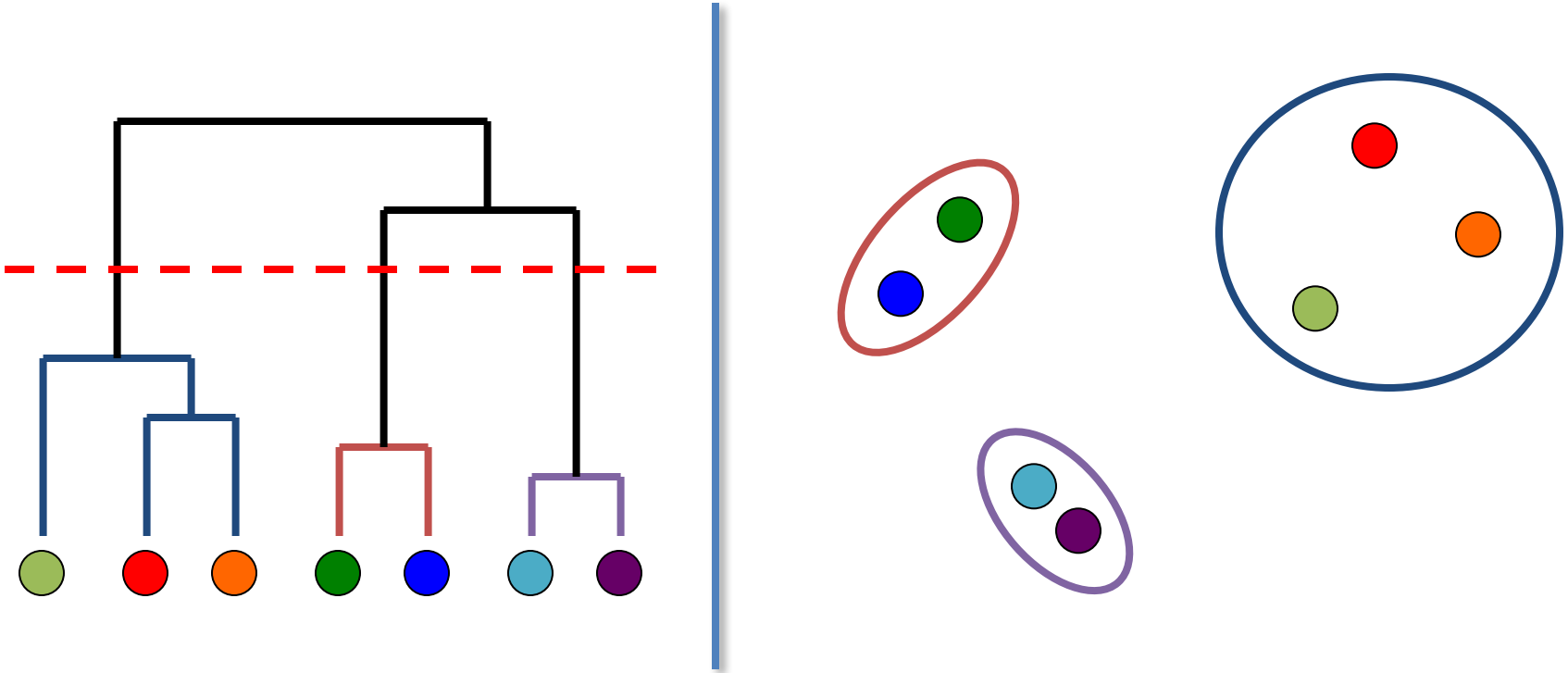


Sørlie, Therese, et al. (2003) "Repeated observation of breast tumor subtypes in independent gene expression data sets," *PNAS*.

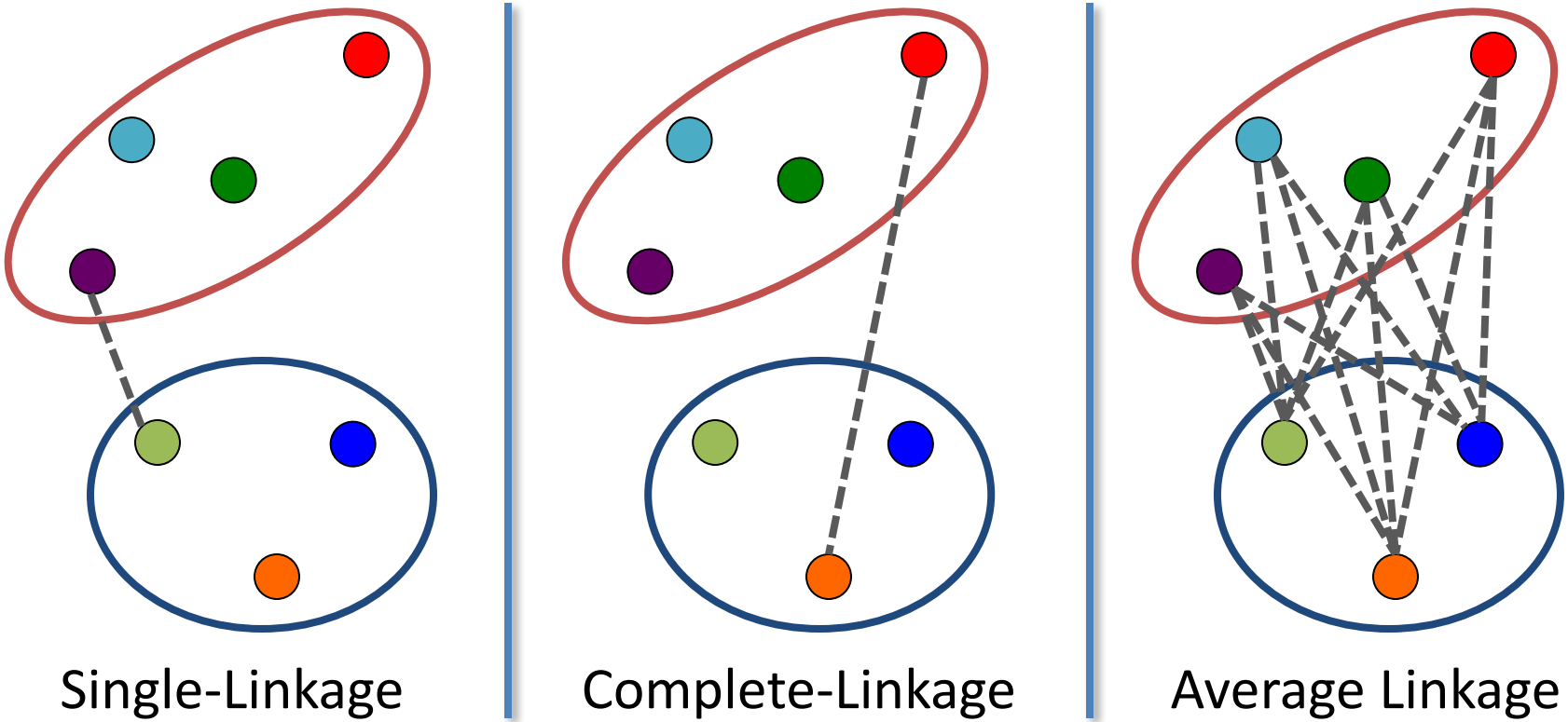
Hierarchical Clustering



Hierarchical Clustering



Hierarchical Clustering



Single-Linkage

Complete-Linkage

Average Linkage

K-means Clustering

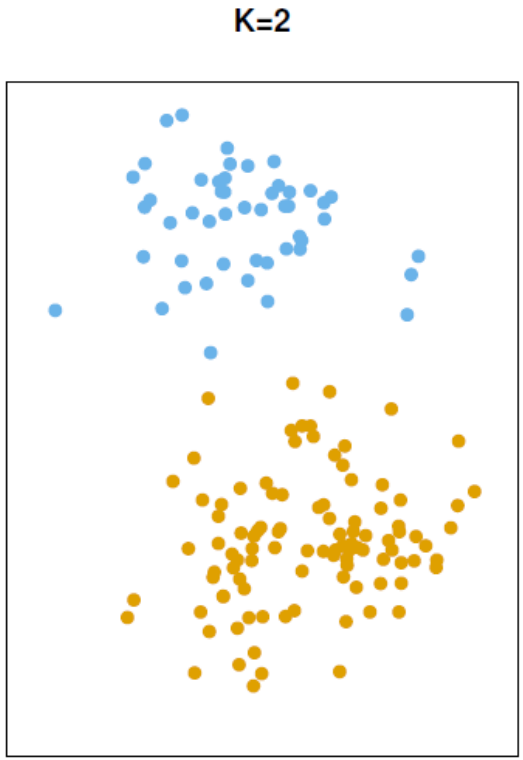


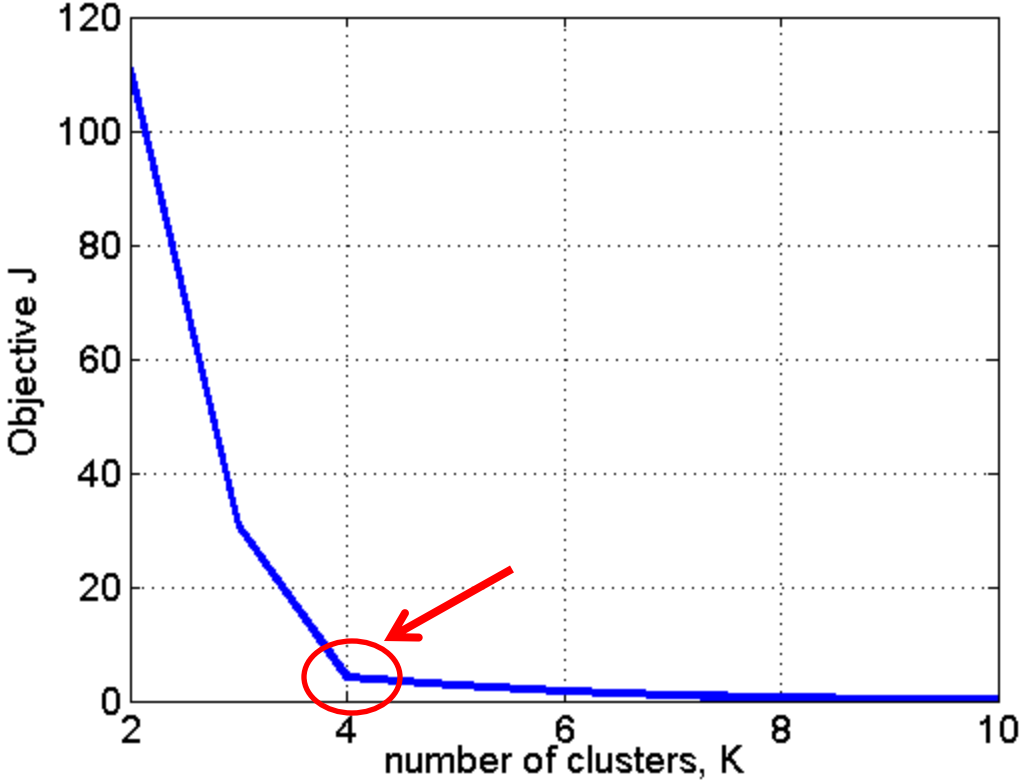
Figure 10.5 , ISL 2013

K-means Algorithm*

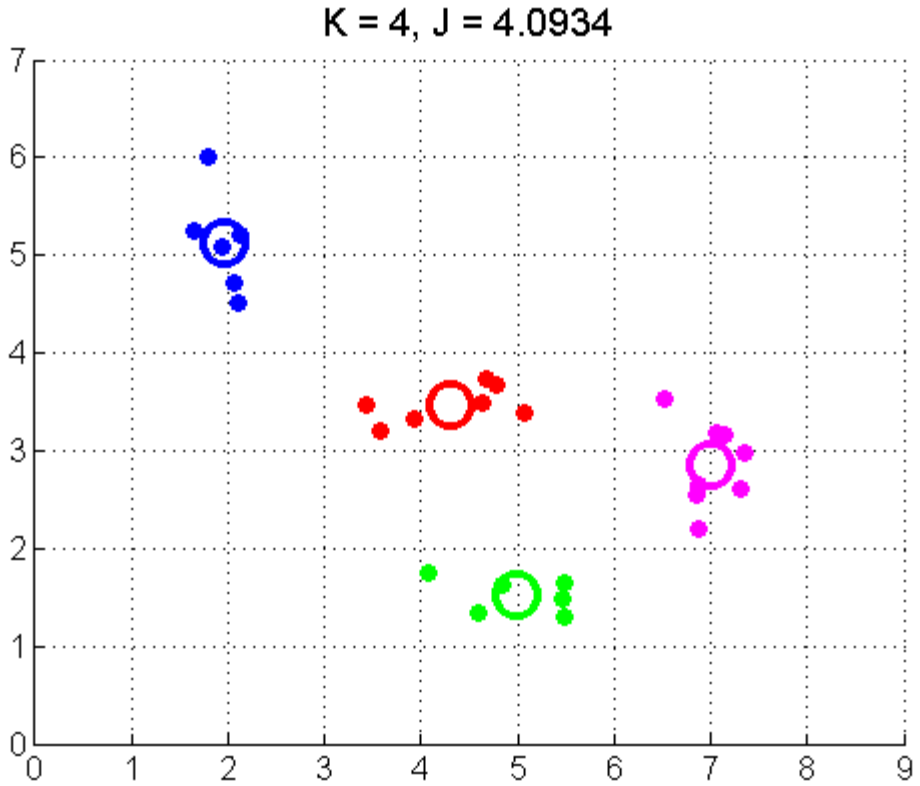
1) Select initial set of centroids
2) Partition data by assigning each sample to cluster associated with its nearest centroid
3) Compute new centroids within each cluster
4) Repeat 2 and 3 until convergence
   - "converged" when centroids stabilize and samples do not move between clusters

* also known as "Lloyd's algorithm"

How many clusters?



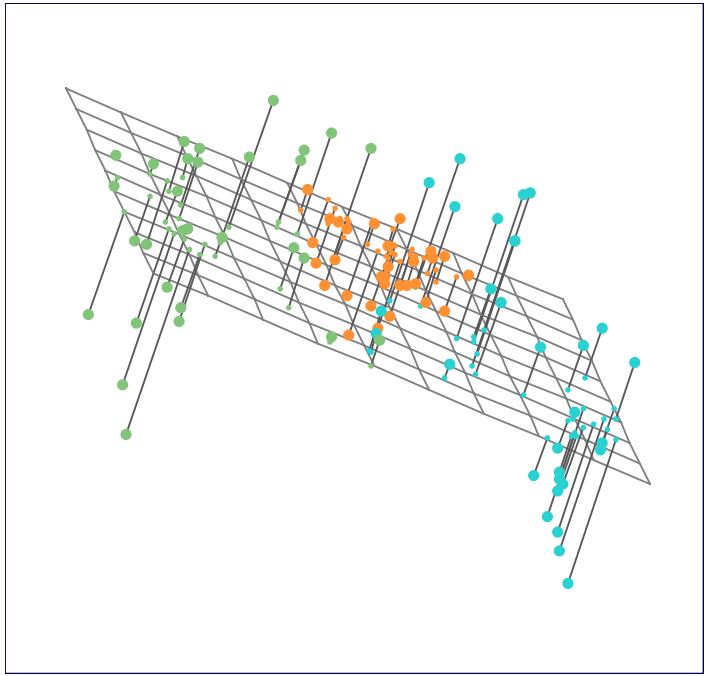
How many clusters?

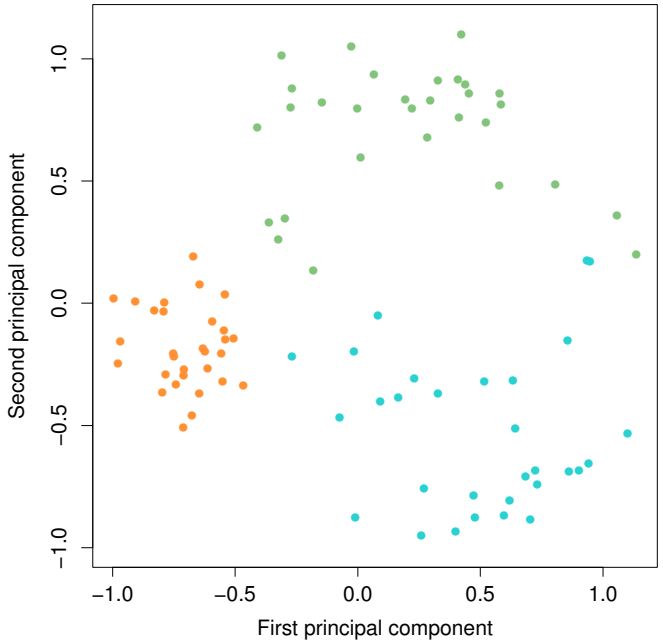


K-means algorithm

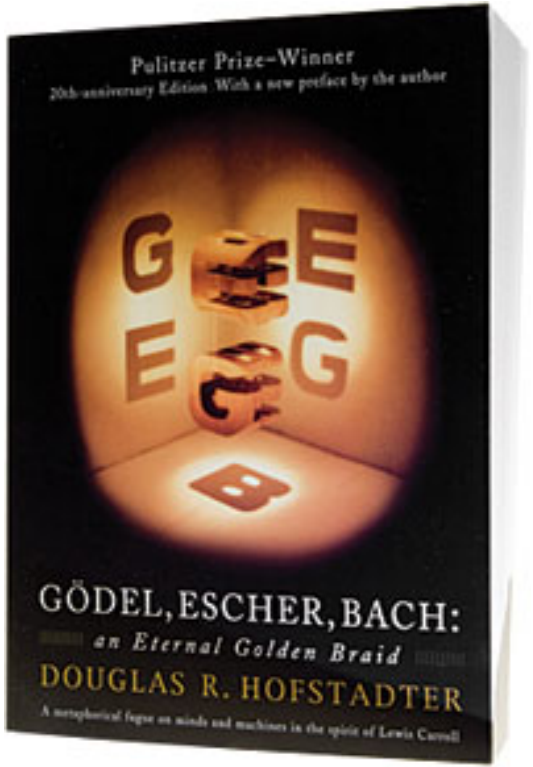
- Advantages
  - Easy to implement
  - Often converges in a small number of iterations
  - Can be applied on data with a large number of features
- Disadvantages
  - K is input parameter
  - Iterative algorithm returns local minimum*

Dimensionality Reduction



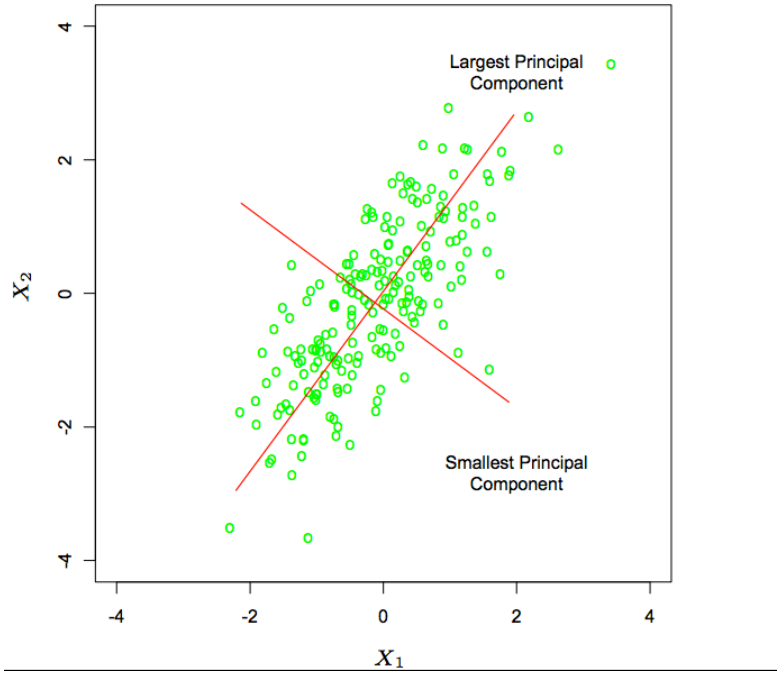


Projections



I. Principal Component Analysis (PCA)

Maximal Variance Projection (PCA)



Maximal Variance Projection (PCA)

Find eigenvectors with largest eigenvalues of the *sample covariance matrix*:

$$X^T X = V D^2 V^T$$

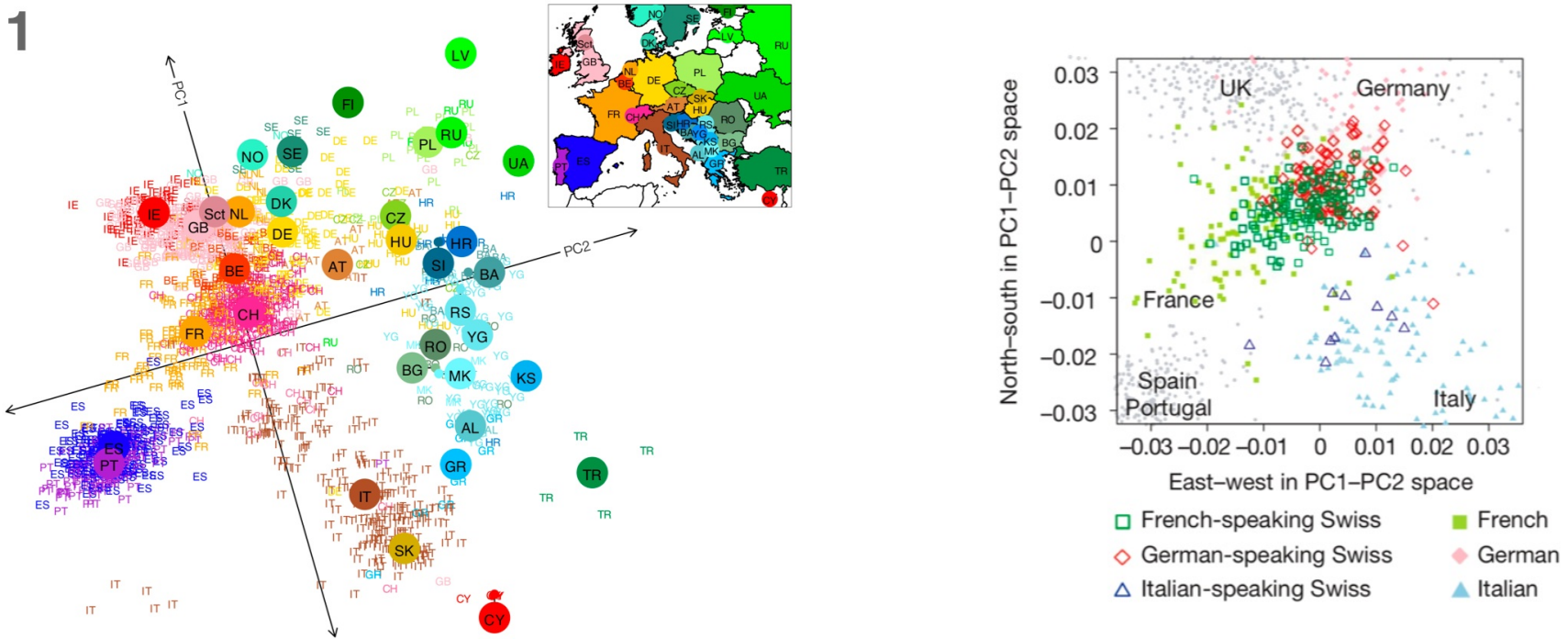
Or, equivalently by the Singular Value Decomposition of the N by p data matrix X:

$$X = U D V^T$$

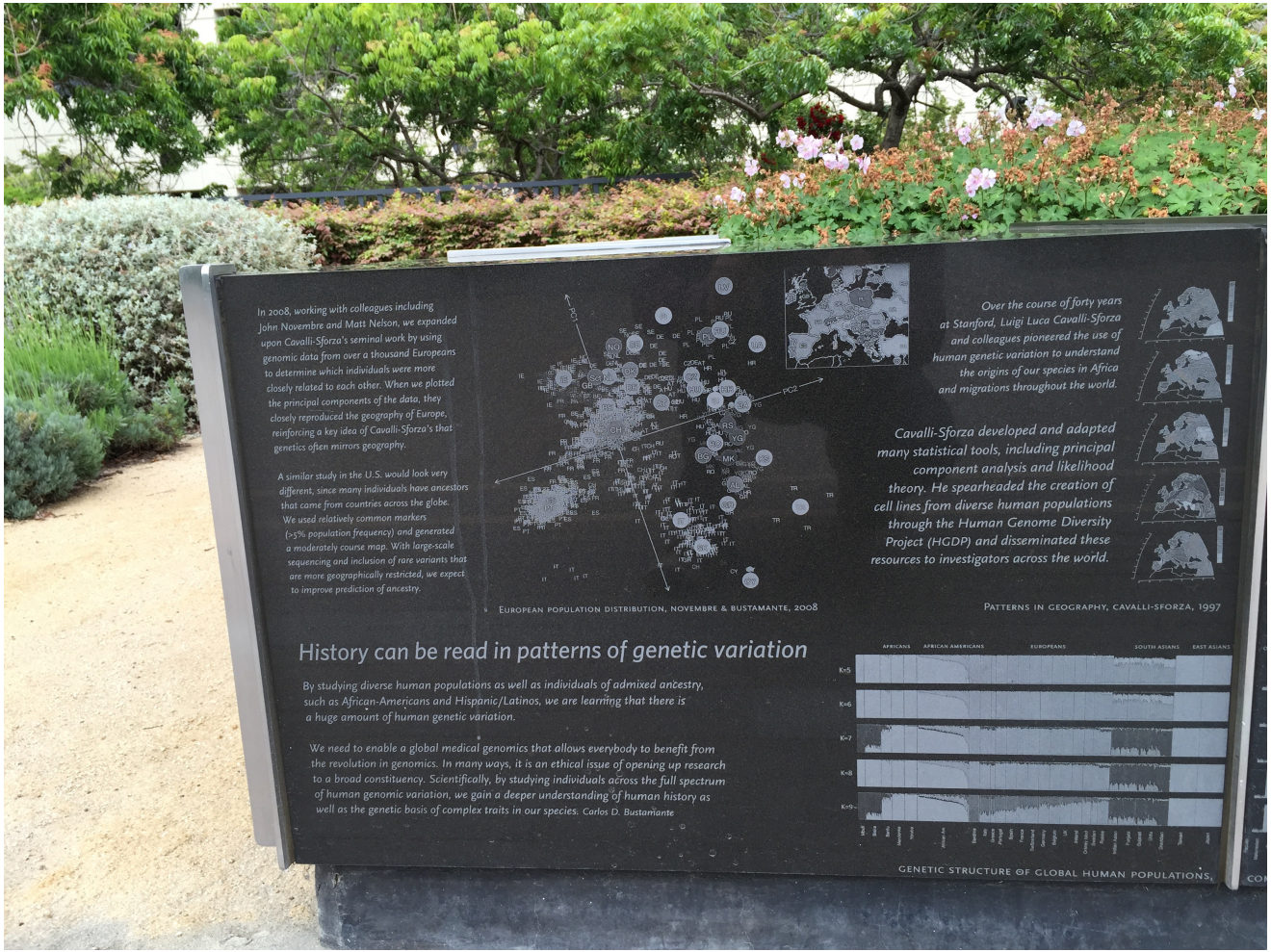
Maximal Variance Projection (PCA)

- Find a set of q uncorrelated variables that are linear combinations of the original p variables and explain most of the variation.
- Approximate the N x p data matrix by the best (in a squared error sense) rank-q matrix.
- Given N points in a *p* dimensional space find the least-squares optimal manifold of co-dimension q, (which is less than p).

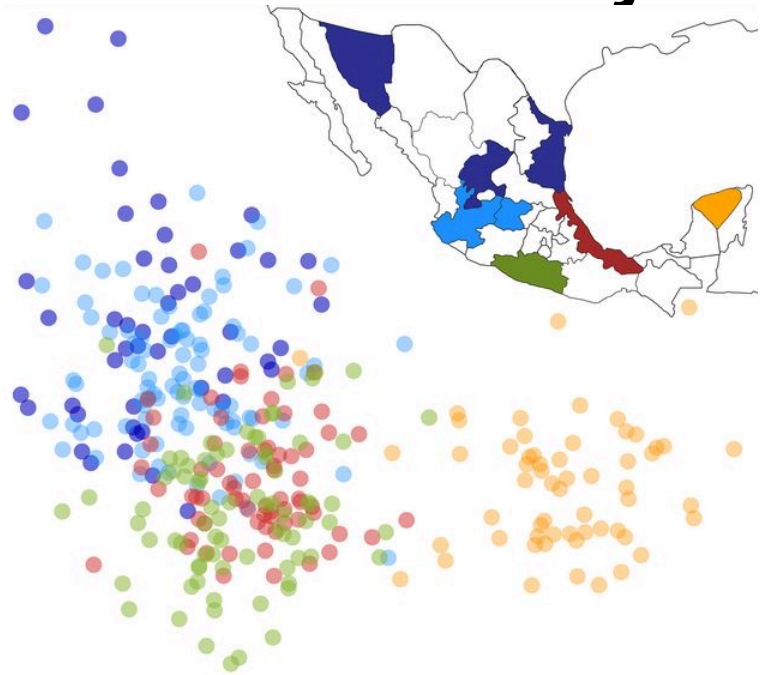
Maximal Variance Projection (PCA)



Novembre, John, et al. "Genes mirror geography within Europe." *Nature* 456.7218 (2008): 98-101.

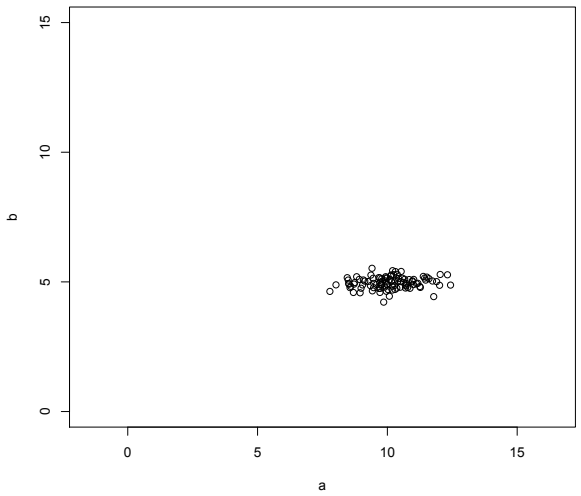

In 2008, working with colleagues including John Novembre and Matt Nelson, we expanded upon Cavalli-Sforza's seminal work by using genomic data from over a thousand Europeans to determine which individuals were more closely related to each other. When we plotted the principal components of the data, they closely reproduced the geography of Europe, reinforcing a key idea of Cavalli-Sforza's that genetics often mirrors geography.
A similar study in the U.S. would look very different, since many individuals have ancestors that came from countries across the globe. We used relatively common markers (>5% population frequency) and generated a moderately course map. With large-scale sequencing and inclusion of rare variants that are more geographically restricted, we expect to improve prediction of ancestry.
PC1
PC2
EUROPEAN POPULATION DISTRIBUTION, NOVEMBRE & BUSTAMANTE, 2008
Over the course of forty years at Stanford, Luigi Luca Cavalli-Sforza and colleagues pioneered the use of human genetic variation to understand the origins of our species in Africa and migrations throughout the world.
Cavalli-Sforza developed and adapted many statistical tools, including principal component analysis and likelihood theory. He spearheaded the creation of cell lines from diverse human populations through the Human Genome Diversity Project (HGDP) and disseminated these resources to investigators across the world.
PATTERNS IN GEOGRAPHY, CAVALLI-SFORZA, 1997
History can be read in patterns of genetic variation
By studying diverse human populations as well as individuals of admixed ancestry, such as African-Americans and Hispanic/Latinos, we are learning that there is a huge amount of human genetic variation.
We need to enable a global medical genomics that allows everybody to benefit from the revolution in genomics. In many ways, it is an ethical issue of opening up research to a broad constituency. Scientifically, by studying individuals across the full spectrum of human genomic variation, we gain a deeper understanding of human history as well as the genetic basis of complex traits in our species. Carlos D. Bustamante
AFRICANS
AFRICAN AMERICANS
EUROPEANS
SOUTH ASIANS
EAST ASIANS
K=5
K=6
K=7
K=8
K=9
GENETIC STRUCTURE OF GLOBAL HUMAN POPULATIONS

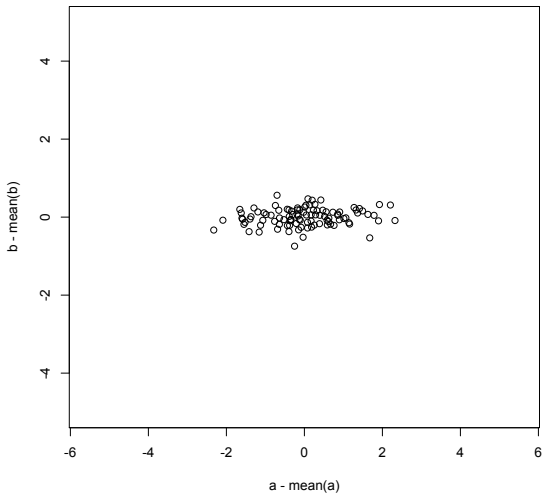
Maximal Variance Projection (PCA)

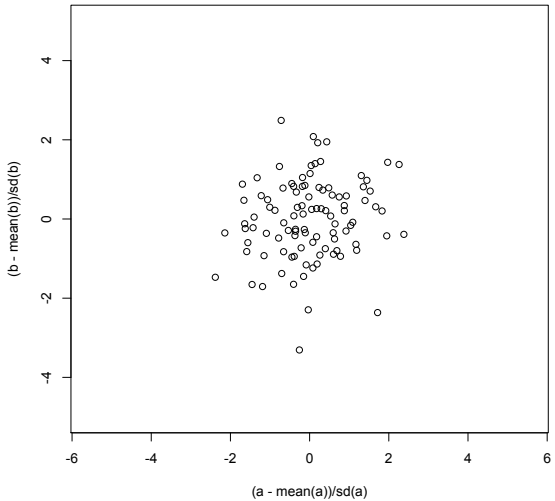


Moreno-Estrada, Andrés, et al. "The genetics of Mexico recapitulates Native American substructure and affects biomedical traits." *Science* 344.6189 (2014): 1280-1285.

Centering (always), normalizing (correlation PCA), whitening (never)







How many principal components?

$$W_k = \frac{1}{N} \sum_{i=1}^{k} d_i^2$$

Variance Explained

How many principal components?

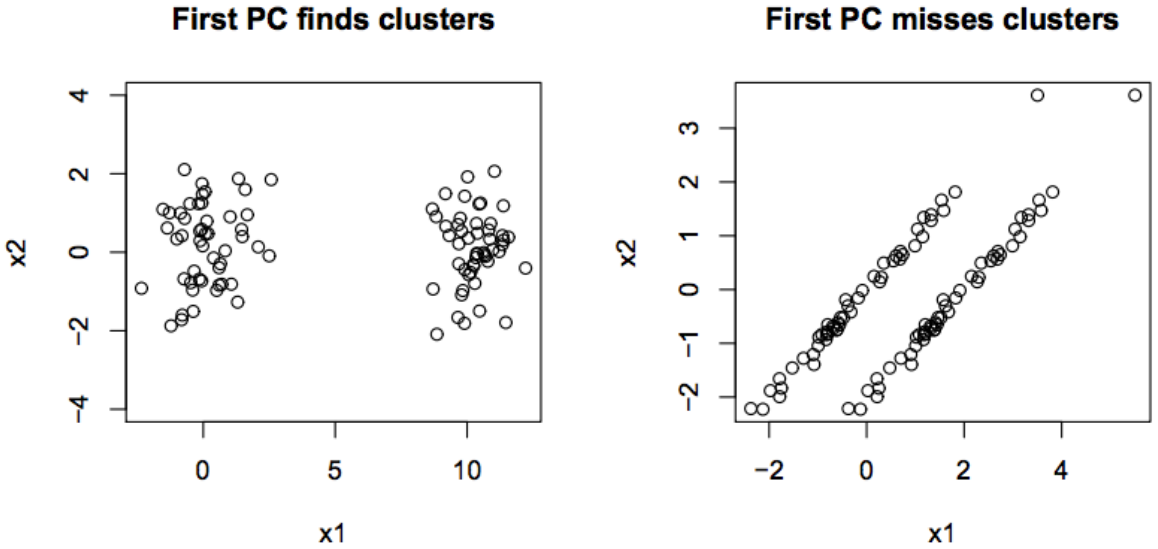
$$W_k = \frac{1}{N} \sum_{i=1}^{k} d_i^2$$

Variance Explained

$$\log W_k - \log W_k^*$$

Gap Statistic

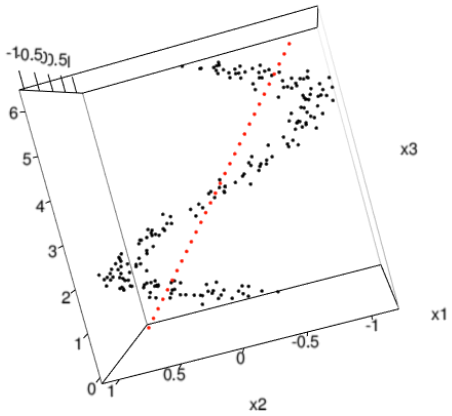
Does PCA give the most interesting projection?



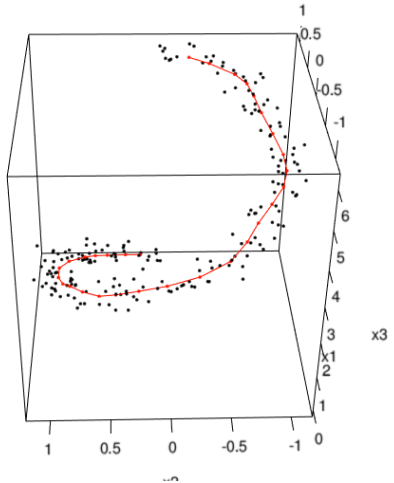
II. Self Organizing Maps (SOM)

Self Organizing Maps (SOM)

$$m_k \leftarrow m_k + \alpha \bullet (x_i - m_k)$$



Prototypes initialized along 1st principal component axis

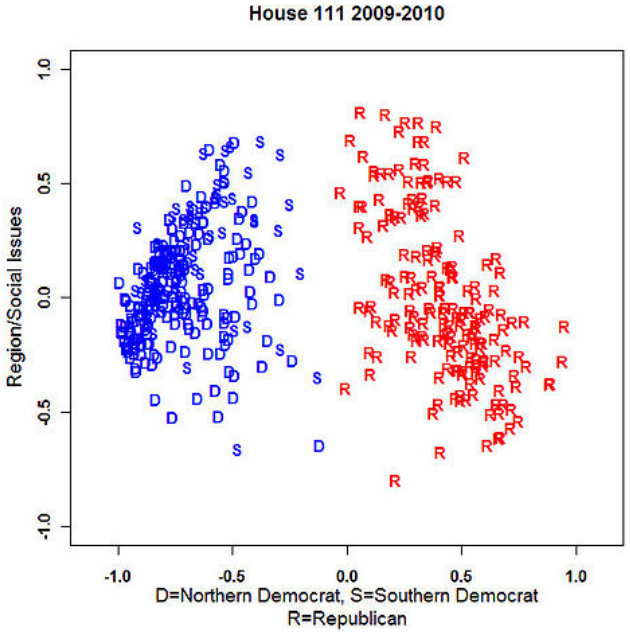


Final result of SOM iteration

III. Multi-Dimensional Scaling (MDS)

Multi-dimensional Scaling (MDS)

$$\min_{x1,\ldots,x_I} \sum_{i<j} (\|x_i - x_j\| - \delta_{i,j})^2.$$

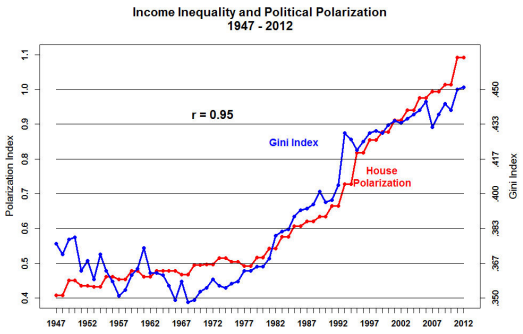


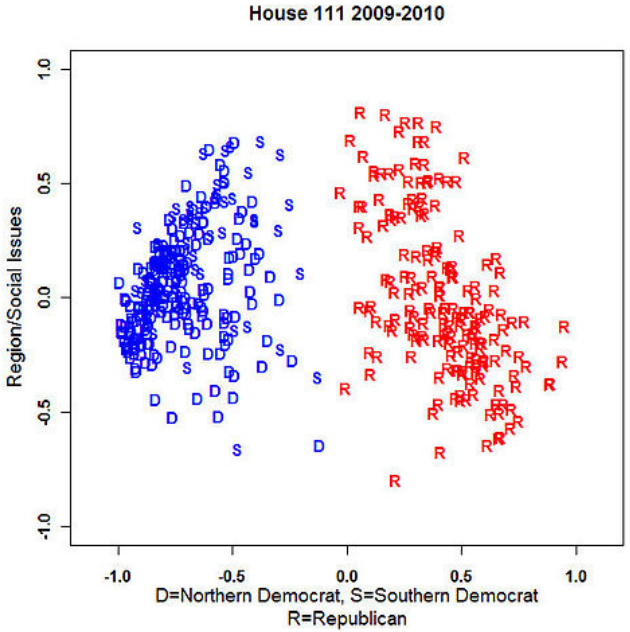
http://voteview.com/polarized_america.htm

Nicholas Kristof, "America's Political Dysfunction." *The New York Times*, Nov. 6, 2014.

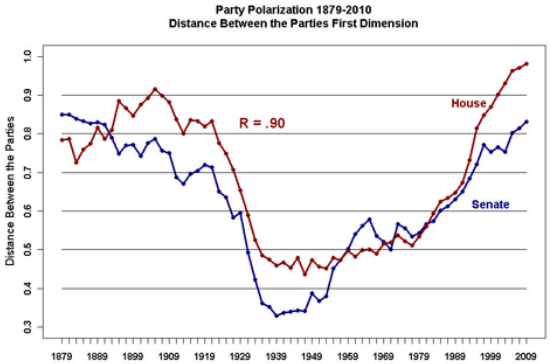
Multi-dimensional Scaling (MDS)

$$\min_{x_1,\ldots,x_I} \sum_{i<j} (\|x_i - x_j\| - \delta_{i,j})^2.$$





http://voteview.com/polarized_america.htm



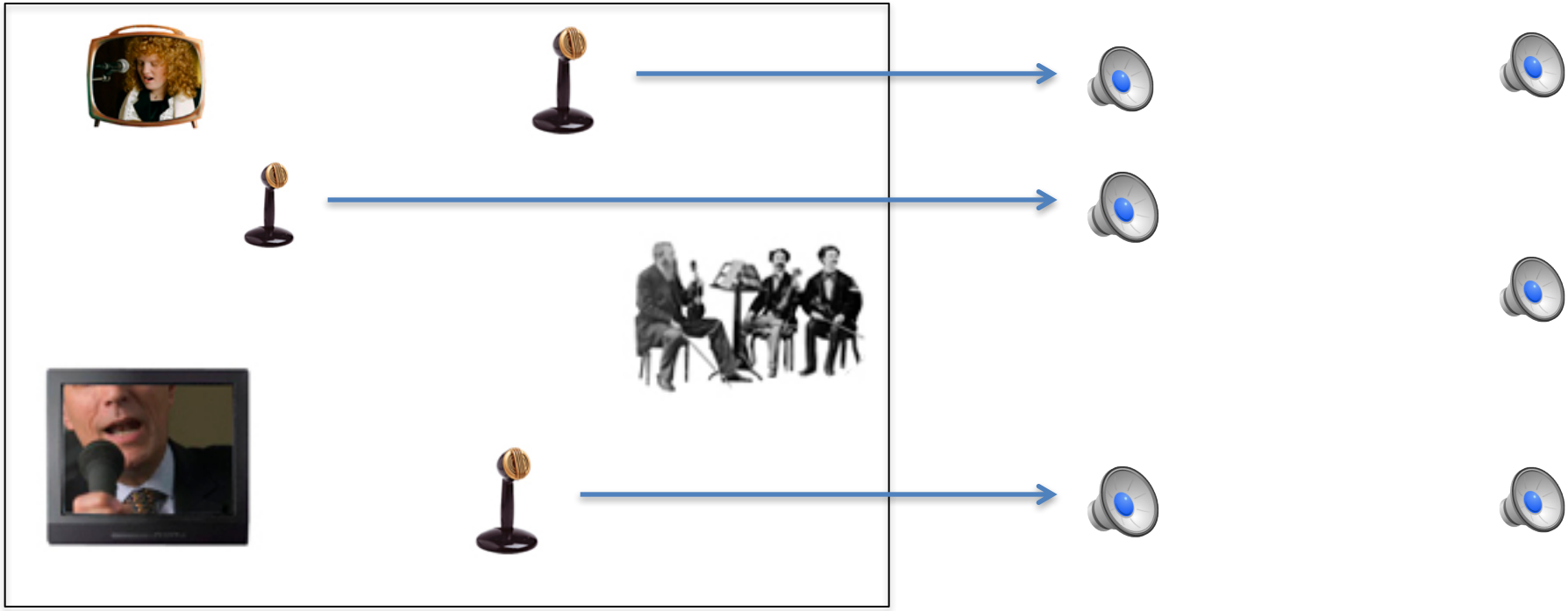
IV. Independent Component Analysis (ICA)

Independent Component Analysis

How to unmix linearly superimposed signals?

$$x_1(t) = a_{11}s_1 + a_{12}s_2$$
$$x_2(t) = a_{21}s_1 + a_{22}s_2$$

Independent Component Analysis



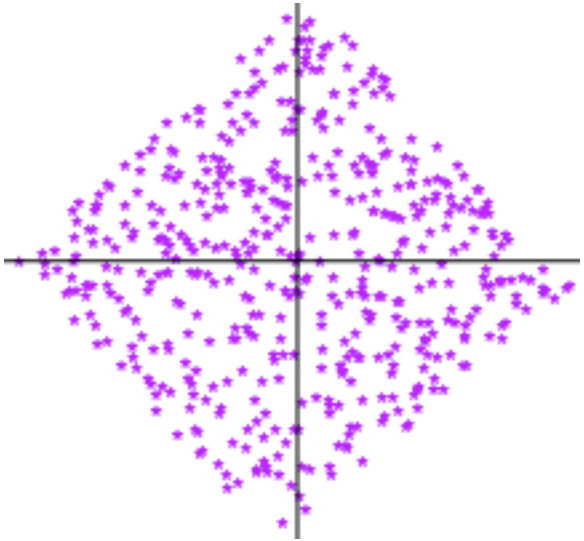
Independent Component Analysis

1) Whiten (Sphere) Data $\Sigma^{-\frac{1}{2}} X$

Directions become uncorrelated, but not independent

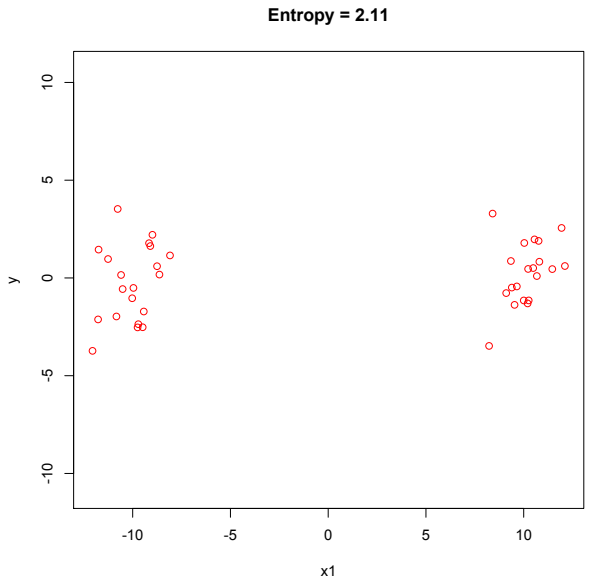
2) Find low entropy (high information) or non-Gaussian projections.

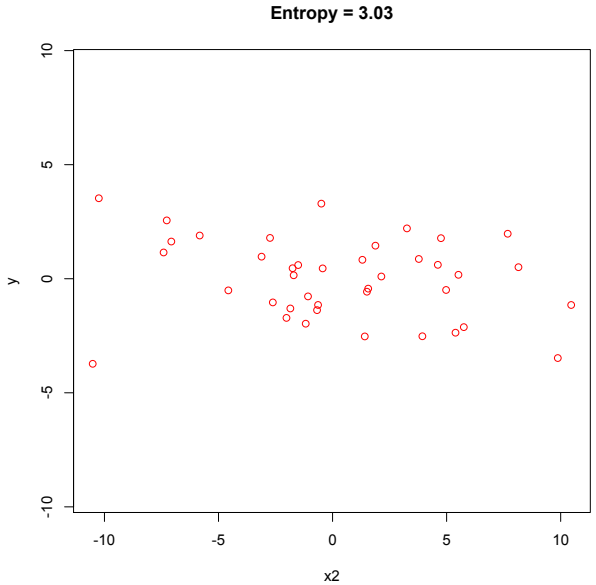
Uncorrelated but not independent



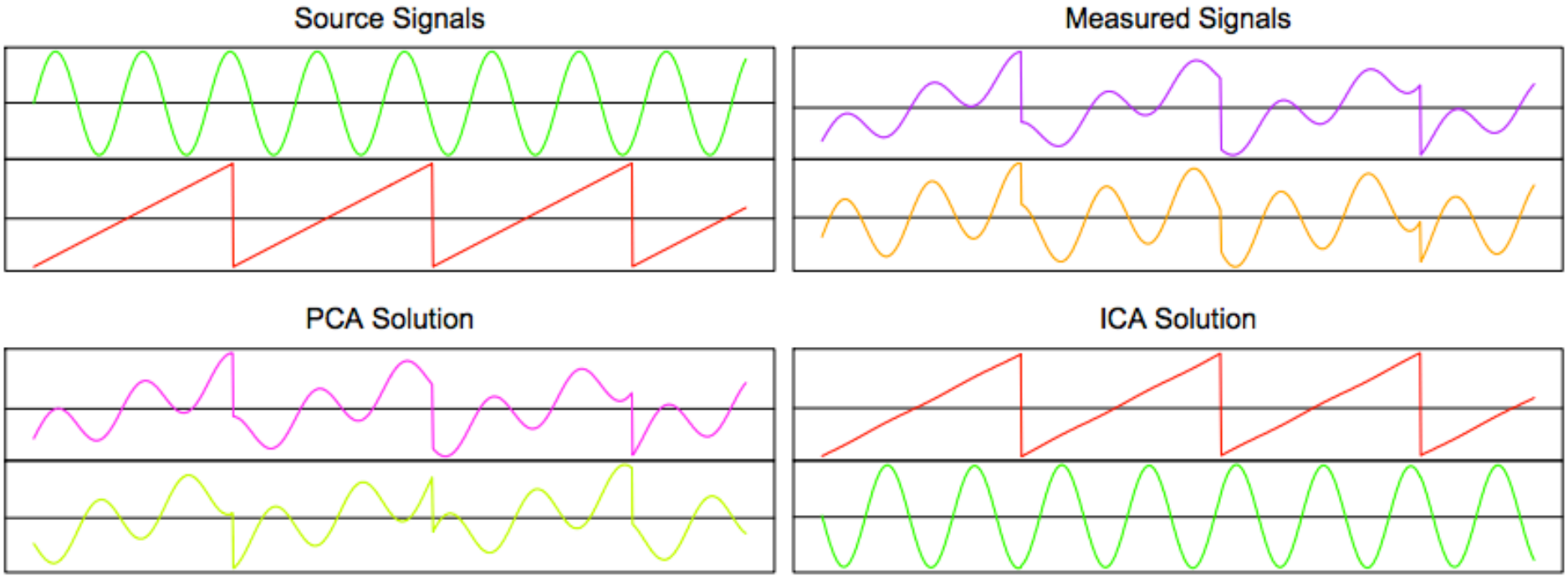
Entropy

$$h(X) = -\int_{\mathbb{X}} f(x) \log f(x)\, dx$$





Independent Component Analysis



Other Unsupervised Techniques

- Hierarchical Clustering
- Nonnegative Matrix Factorization (NMF)
- Multidimensional Scaling (MDS)

Questions?