STATISTICS 454/563, Lecture 1

Sept 05, 2018

Xuekui Zhang

Announcements

- Class meets: Every Tue, Wed, Fri at 11:30 – 12:20 in room ELL 162 (Sep 5- Dec 5, 2018)

- (Tentative) office hours: Tuesday 12:30-1:30 in room DTB A523

- (Tentative) grading:
  - 2-4 in class quizzes (10%)
  - 2 Course projects (60%)
  - 2-3 assignments (30%)

Acknowledgement

- The slides are revised from CS540 (machine learning) that I took at UBC
- I removed some material to add new topics based on my research

Pre-requisites

- Maths: multivariate calculus, linear algebra, probability, statistics.
- CS: programming skills, knowledge of data structures and algorithms helpful but not crucial.
- I will discuss Bayesian statistics at length.
- Programming language preferred R and/or C

Textbook

- C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.
-
- Check the webpage http://research.microsoft.com/~cmbishop/PRML/

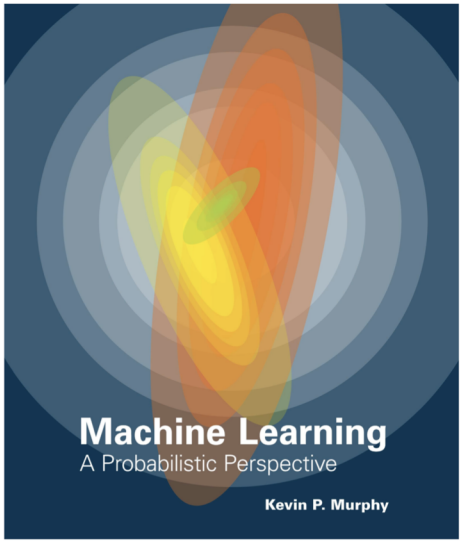
- Corrections of exercises available on this webpage.

- Reference book: Hastie, Tibshirani and Friedman, Elements of Statistical Learning - Data Mining, Inference and Prediction, Springer-Verlag, 2009.

Best reference book for Matlab users

Machine Learning: a Probabilistic Perspective

by Kevin Patrick Murphy



- Hardcopy available from Amazon.com. There is only one edition of the book. However, there are multiple print runs of the hardcopy, which have fixed various errors (mostly typos). The latest printing is the fourth printing (Sep. 2013). This is what Amazon (at least in the USA) is shipping. Note: page numbering can be different between printings, although the section numbers, figure numbers, and equation numbers are the same.
- Ecopy available from MIT Press. (The Kindle version is still (as of 4 March 2014) from the first printing, which has many errors, so do not buy the ecopy from Amazon! The MIT Press version is up to date.)
- As of 10/19/15, a Korean version of the book is available.
- Table of contents
- Chapter 1 (Introduction)
- Chapter 19 (Undirected graphical models/ Markov random fields). Note: this is from the third printing. This corrects some errors that were found (by Sebastien Bratieres) in sec 19.7.
- Bibliography
- Errata
- Matlab software
- All the figures, together with matlab code to generate them
- **My book has won the 2013 De Groot Prize for best textbook on Statistical Science.**
- Best selling machine learning book on amazon.com (22 October 2012).
- Best selling book at MIT Press (24 November 2012).
- Resources for instructors from MIT Press. If you are an official instructor, you can request an e-copy, which can help you decide if the book is suitable for your class. You can also request the solutions manual. Slides are not available.

What is Machine Learning

- Machine learning is a field of computer science that uses statistical techniques to give computer systems the ability to "learn" (e.g., progressively improve performance on a specific task) with data, without being explicitly programmed" --- Wikipedia

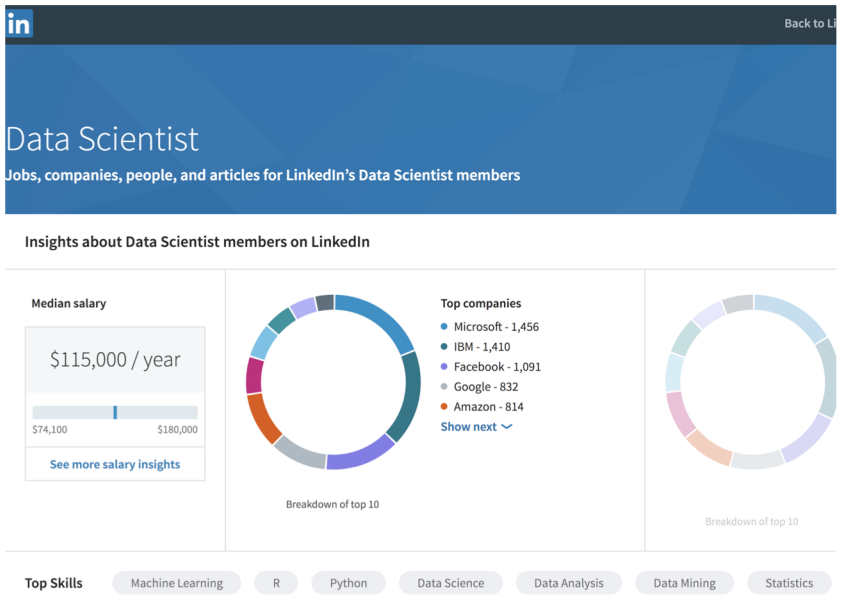
- "Learning denotes changes in the system that are adaptive in the sense that they enable the system to do the task or tasks drawn from the same population more efficiently and more effectively the next time" ---Herb Simon.

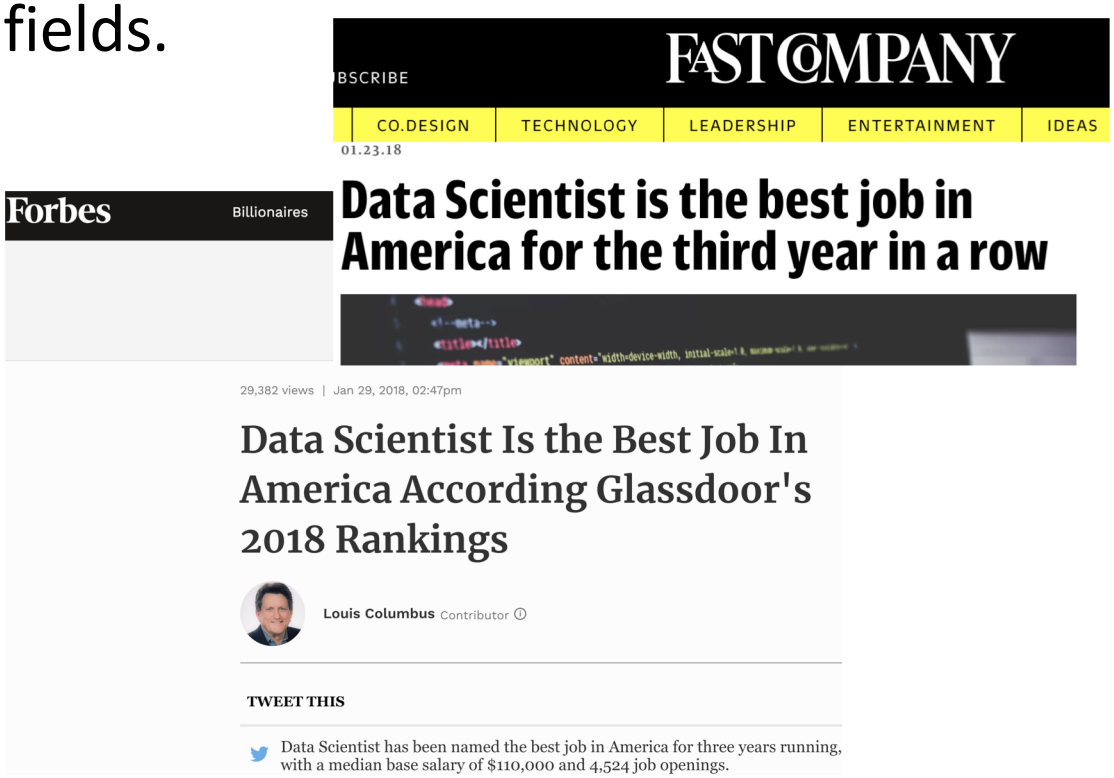
- Machine Learning is an interdisciplinary field at the intersection of Statistics, CS and EE etc.

- At the beginning, Machine Learning was fairly heuristic but it has now evolved and become -in my opinion- Applied Statistics with a CS flavor.

Motivations for learning Machine Learning

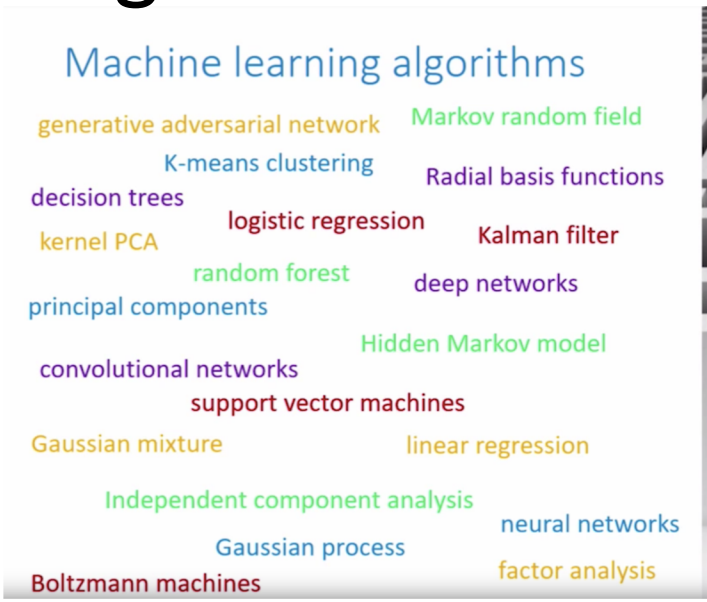
- Applications in many other fields.
- Job market





Why not just one best algorithm for all?

Deep learning



No Free Lunch Theorem

Wolpert, D.H., Macready, W.G. (1997)

If an algorithm performs well on a certain class of problems then it necessarily pays for that with degraded performance on the set of all remaining problems

Validation

- So many models, which is the best for “my” task? -- Validation

- Types of validation
  - Internal validation
  - Independent validation
  - Cross validation

Validation

- Applying a model to a new sample shrinks goodness of fit
  - Wherry, R. J. (1931). A New Formula for Predicting the Shrinkage of the Coefficient of Multiple Correlation. *Annals of Mathematical Statistics, 2*, 440—457.
- Psychologists became aware of this problem in the 1960s
  - Robyn Dawes, Lou Goldberg, Paul Slovic, Lee Cronbach, Amos Tversky, They shrank their *R*2 values with Wherry's formula when presenting models
  - Soon thereafter they did cross validation for other models
    They used cross validation to guard against over-fiting

Cross Validation

- Types
  - Split-half
  - Leave one out (impractical)
  - K-fold CV (popular)
    - Split file into k pieces
    - For each k, train on other k-1 pieces, test on the kth
    - Average k goodness-of-fit statistics
- Problems
  - What is the population?
    - This is a major problem for Big Data
    - Other researchers testing (replicating) your method can't use your data
      - They can't find another dataset from the same population
      - because yours was a convenience batch (you had the whole population)
  - K-fold CV is not replication (in the same sense that scientists use the word)
    - Yu, B. (2013). Stability. *Bernoulli, 19*, 1484-1500.
  - Researchers often use CV to select best model or optimize parameter values

(Tentative) Topics to Be Covered

- *Introduction*
- *Linear Models for Regression and Classification*
- *Nonlinear Models for Regression and Classification*
- *Model comparison*
- *Staking*
- *Kernel Methods*
- *Graphical Models*
- *Mixture Models*
- *Approximate Inference*
- *Sampling Methods*
- *Continuous Latent Variables*
- *Hidden Markov Models*

Types of machine learning tasks

- Supervised learning
- Un-supervised learning
- Semi-supervised learning
- Pseudo-supervised learning
- Active learning
- …

Supervised learning

- *Supervised learning*: given a training set of N input-output pairs $\{x_n, t_n\} \in \mathcal{X} \times \mathcal{T}$, construct a function $f : \mathcal{X} \to \mathcal{T}$ to predict the output $\widehat{t} = f(x)$ associated to a new input x.
  - Each input x_n is a p-dimensional feature vector (covariates, explanatory variables).
  - Each output t_n is a target variables (response).
- Regression corresponds to $\mathcal{T} = \mathbb{R}^d$.
- Classification corresponds to $\mathcal{T} = \{1, ..., K\}$.
- Aim is to produce the correct output given a new input.

Polynomial regression

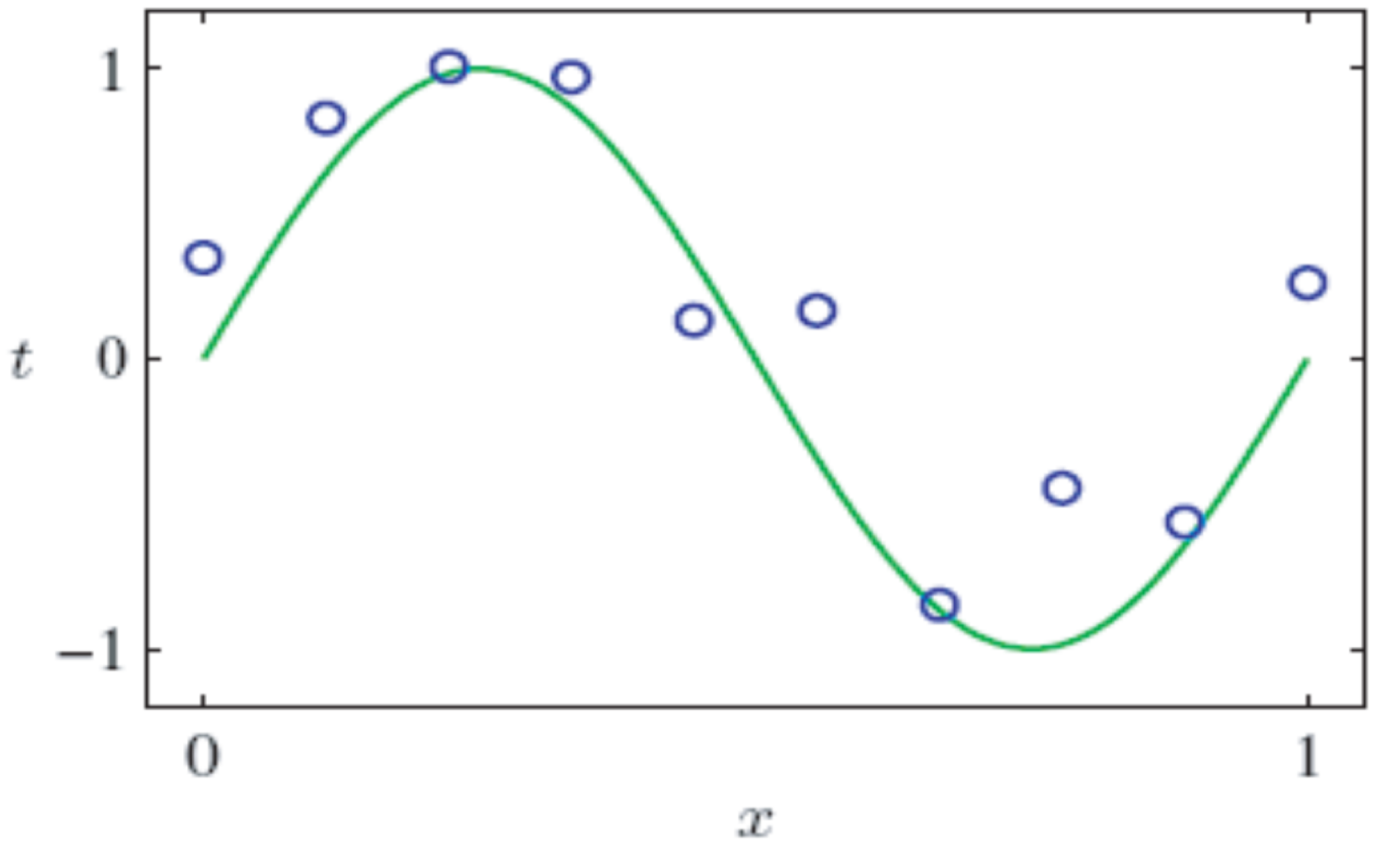


Figure: Polynomial regression where $x \in \mathbb{R}$ and $t \in \mathbb{R}$

Linear regression

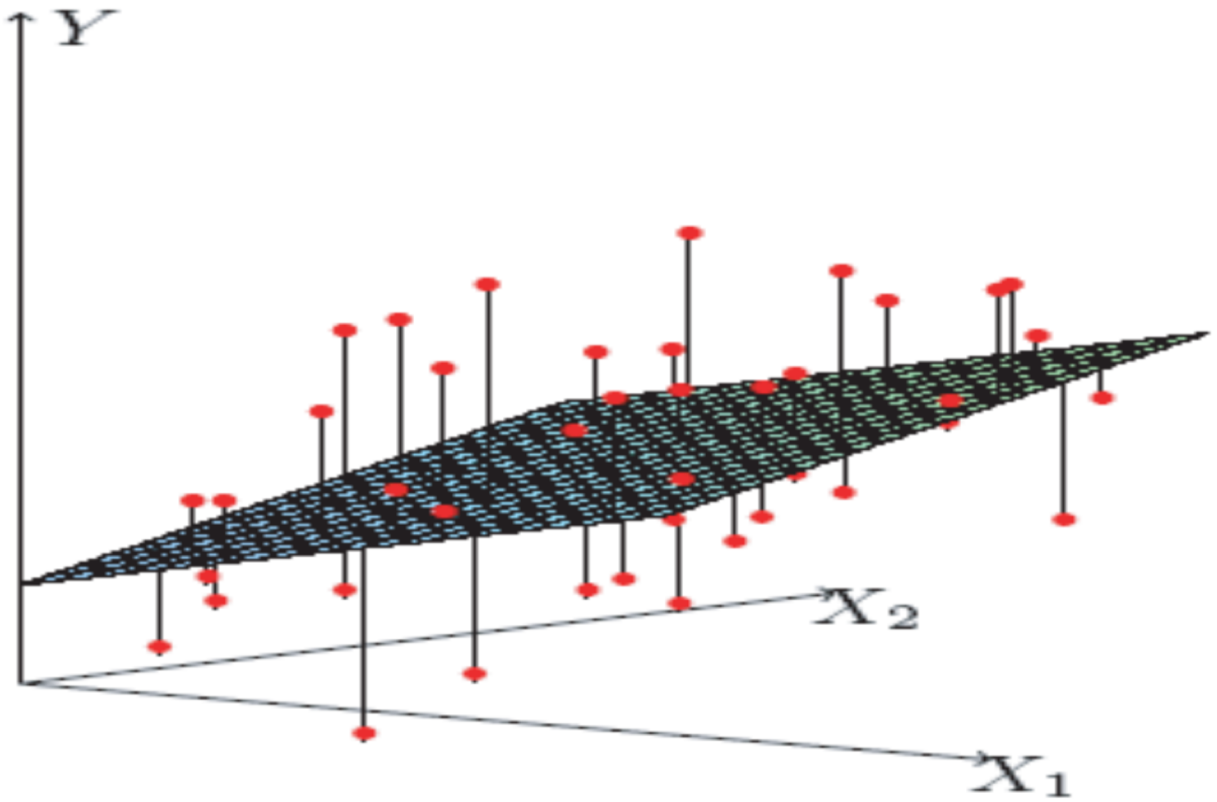


Figure: Linear least square fitting for $x \in \mathbb{R}^2$ and $t \in \mathbb{R}$. We seek the linear function of x that minimizes the sum of square residuals for y.

Piecewise linear regression

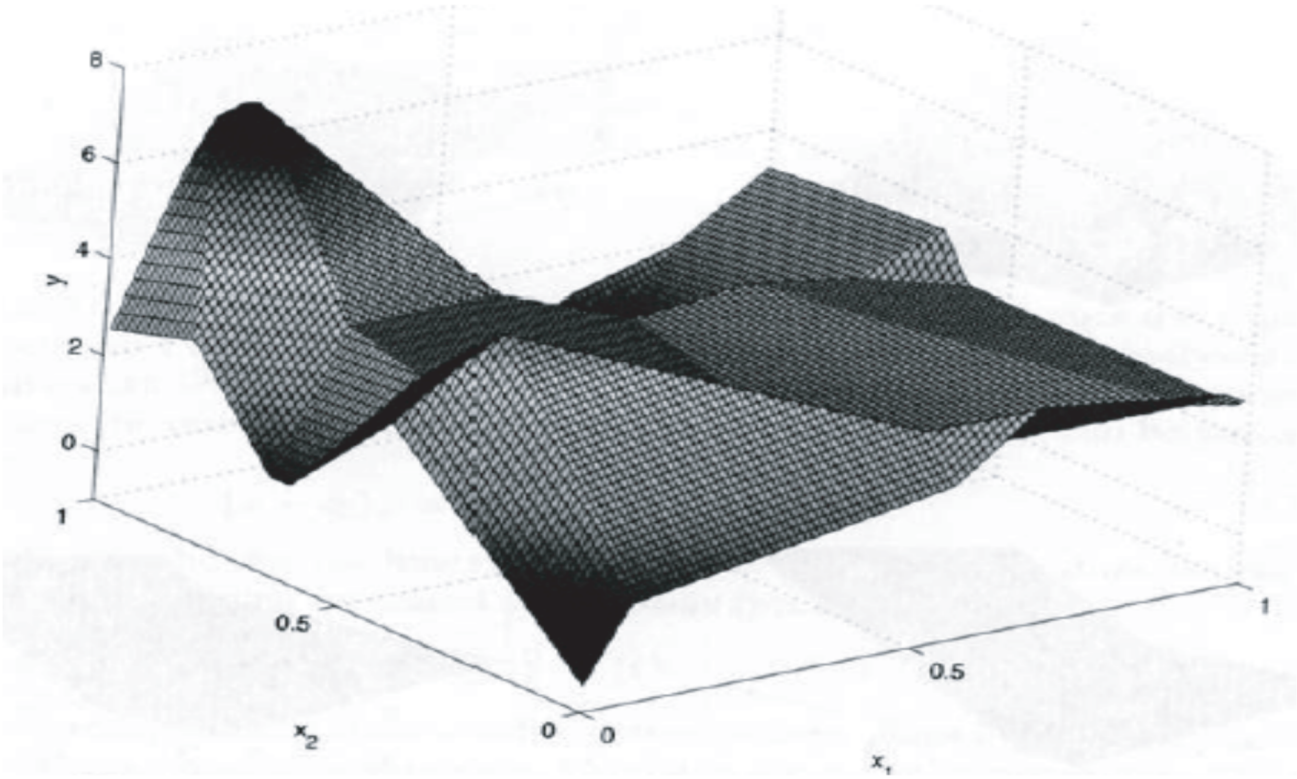


Figure: Piecewise linear regression function

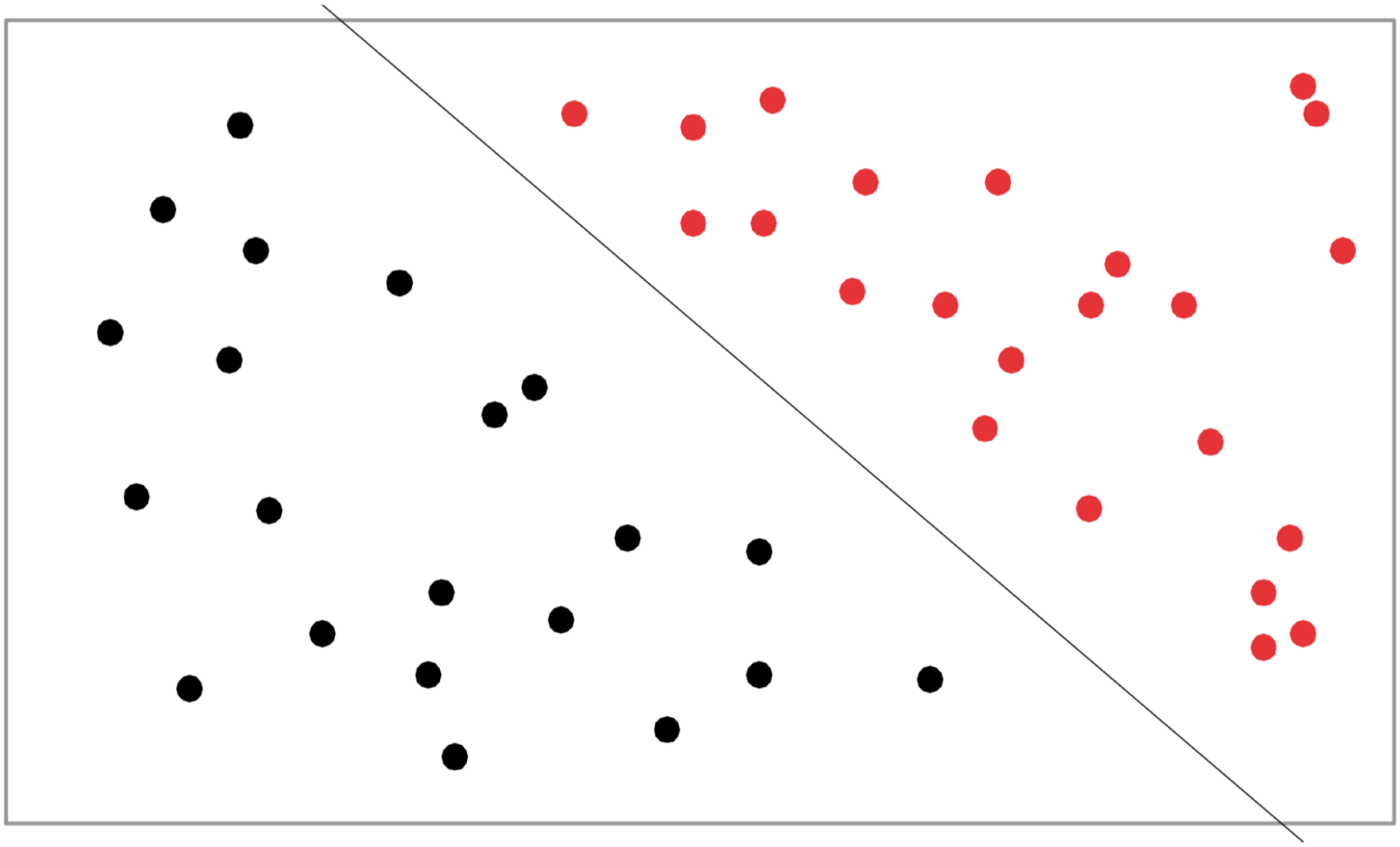


Figure: Linear classifier where $x_n \in \mathbb{R}^2$ and $t_n \in \{0, 1\}$

Handwritten digit recognition

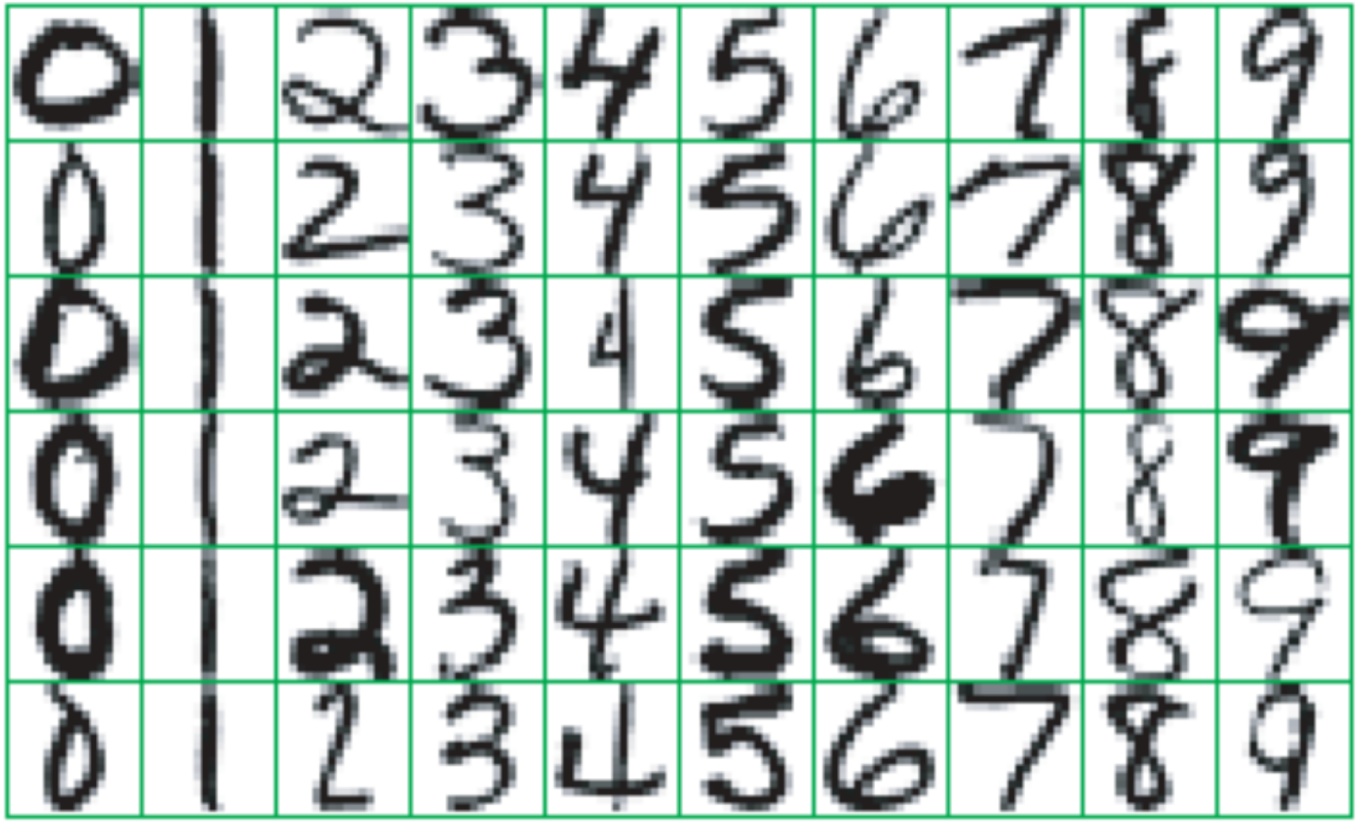


Figure: Examples of handwritten digits from US postal employes

Practical example of supervised learnings

- Email spam filtering (feature vector = "bag of words")
- Webpage classification
- Detecting credit card fraud
- Credit scoring
- Face detection in images

Unsupervised learnings

- Unsupervised learning: we are given training data $\{x_n\}$.
- Aim is to produce a model or build useful representations for x modeling the distribution of the data,
  - clustering,
  - data association,
  - dimensionality reduction,
  - structure learning.

Hard and Soft Clustering

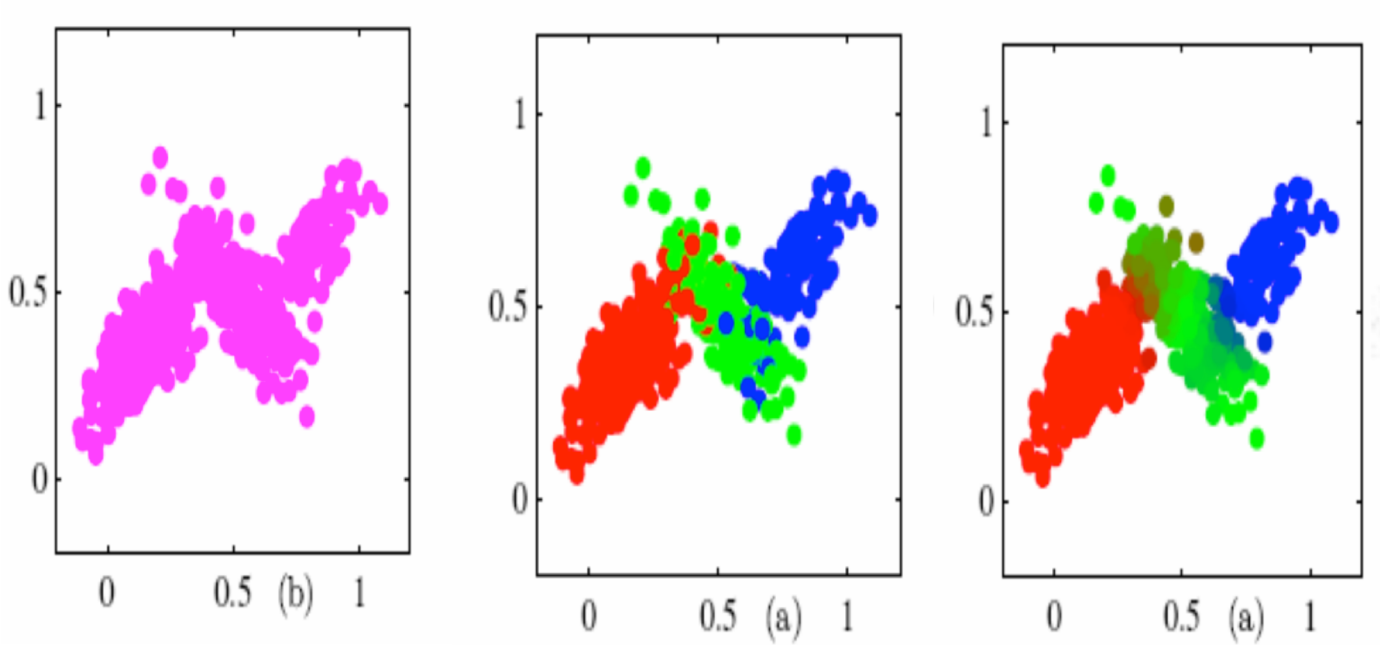


Figure: Data (left), hard clustering (middle), soft clustering (right)

Hierarchical Clustering

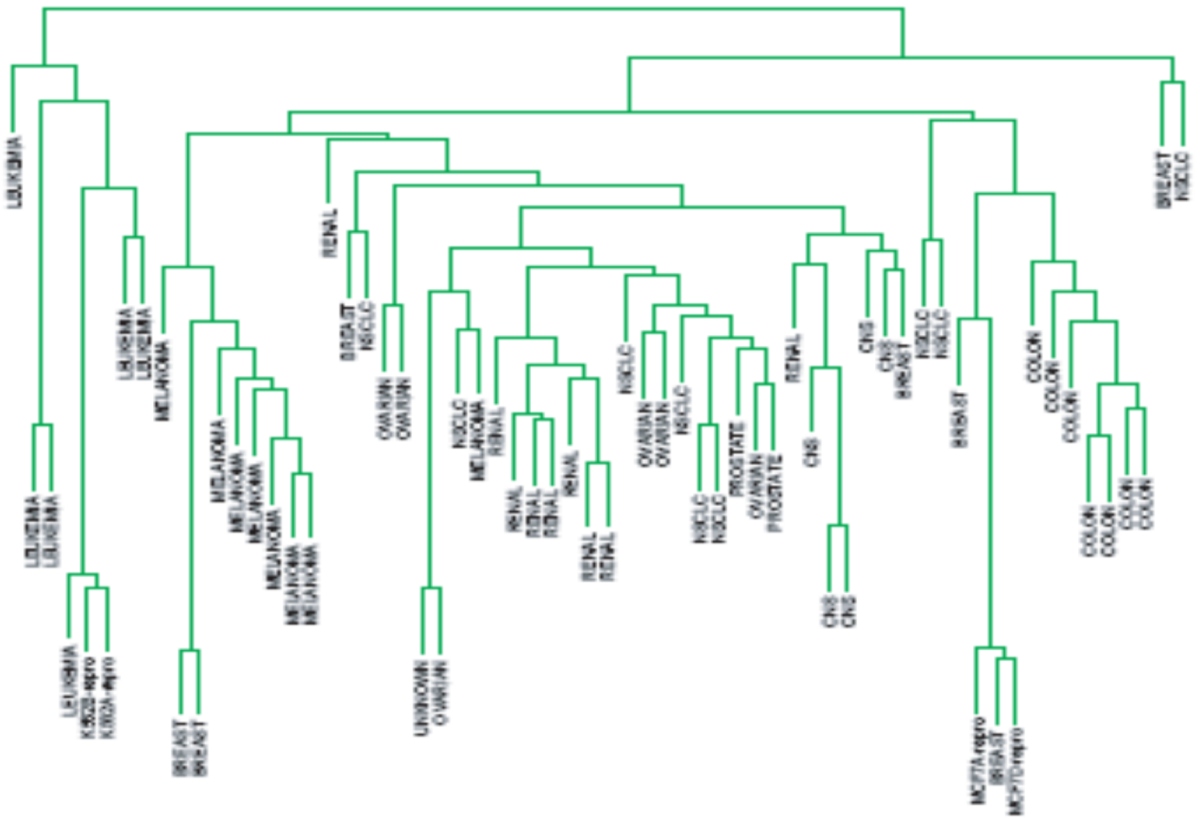


Figure: Dendogram from agglomerative hierarchical clustering with average linkage to the human microarray data

Finite Mixture Models

- Finite Mixture of Gaussians

$$f\left(x|\,\theta\right) = \sum_{i=1}^{k} p_i \mathcal{N}\left(x; \mu_i, \sigma_i^2\right)$$

where $\theta = \left\{\mu_i, \sigma_i^2, p_i\right\}_{i=1,\ldots,k}$ is estimated from some data $\left(x_1, \ldots, x_n\right)$.

- How to estimate the parameters?
- How to estimate the number of components k?

Dimensionality reduction

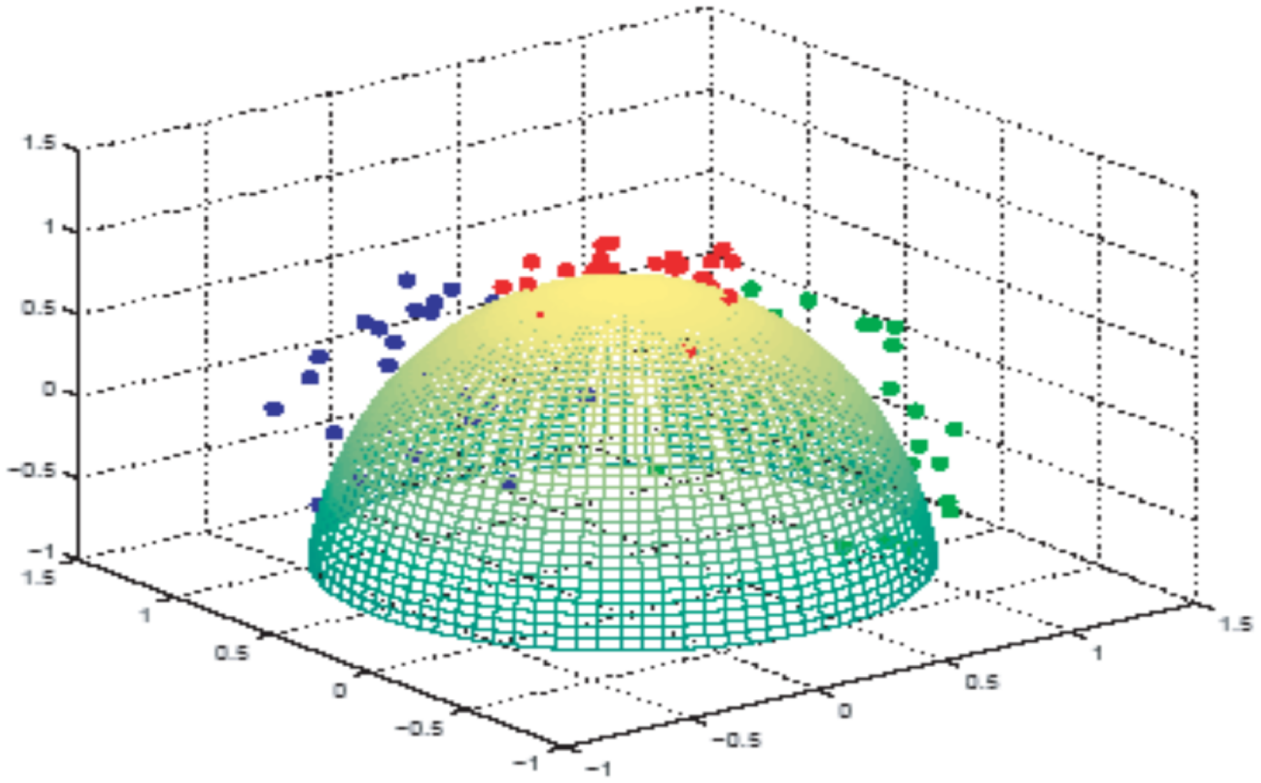


Figure: Simulated data in three classes, near the surface of a half-sphere

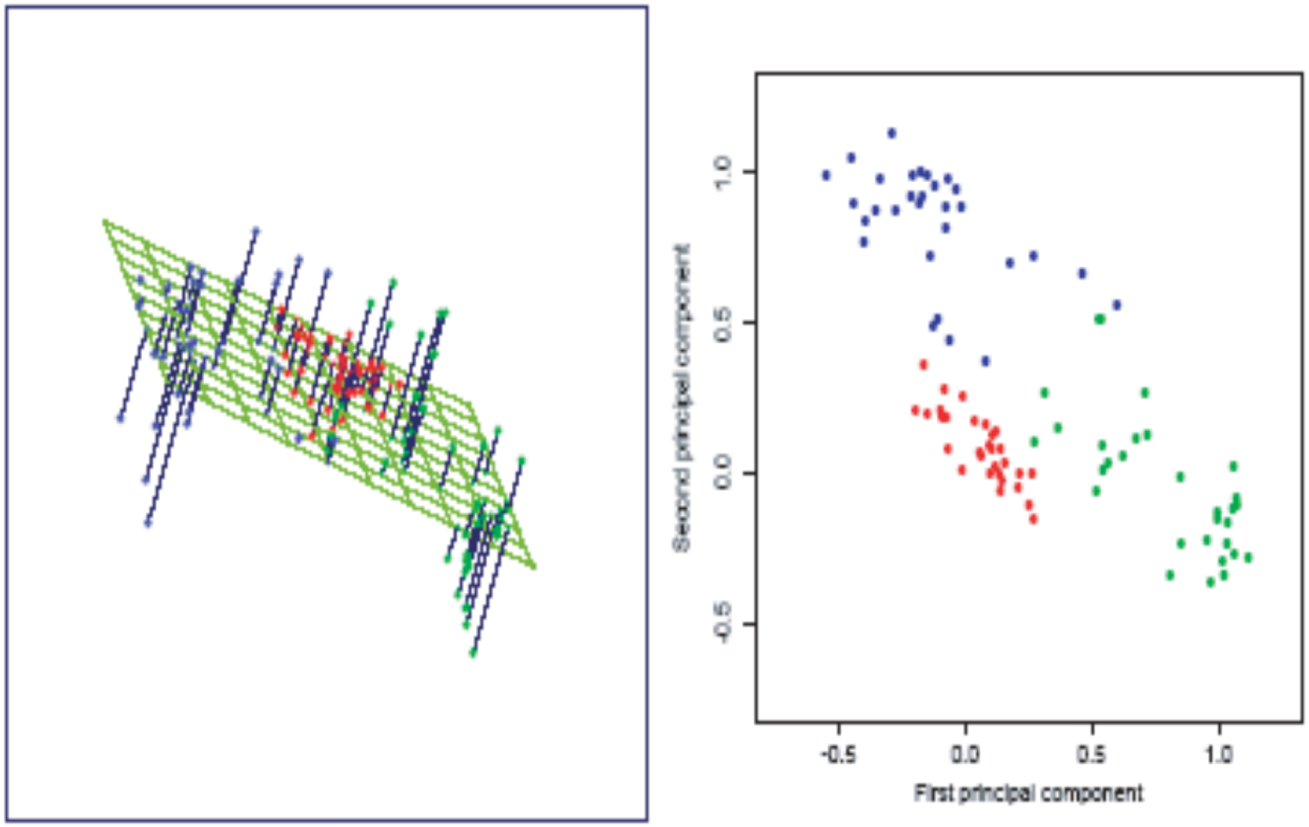
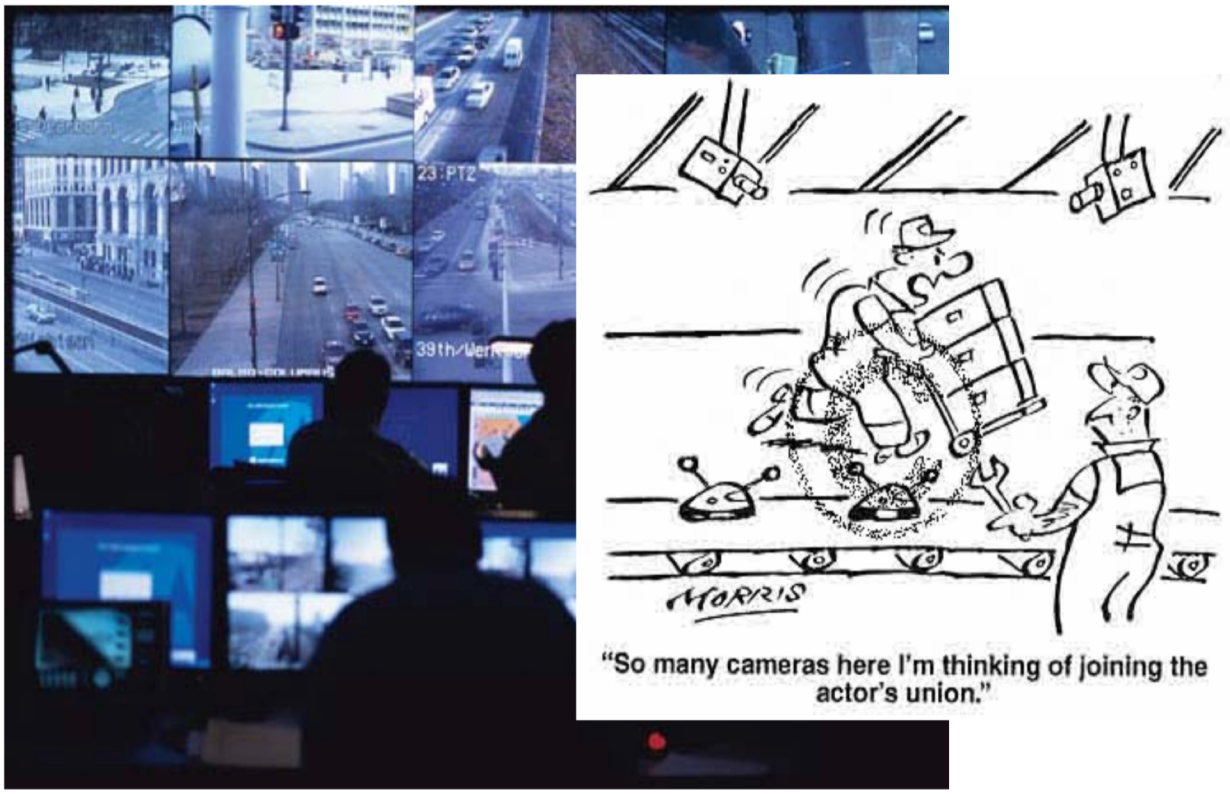


Figure: The best rank-two linear approximation to the half-sphere data. The right panel shows the projected points with coordinates given by the first to principal components of the data

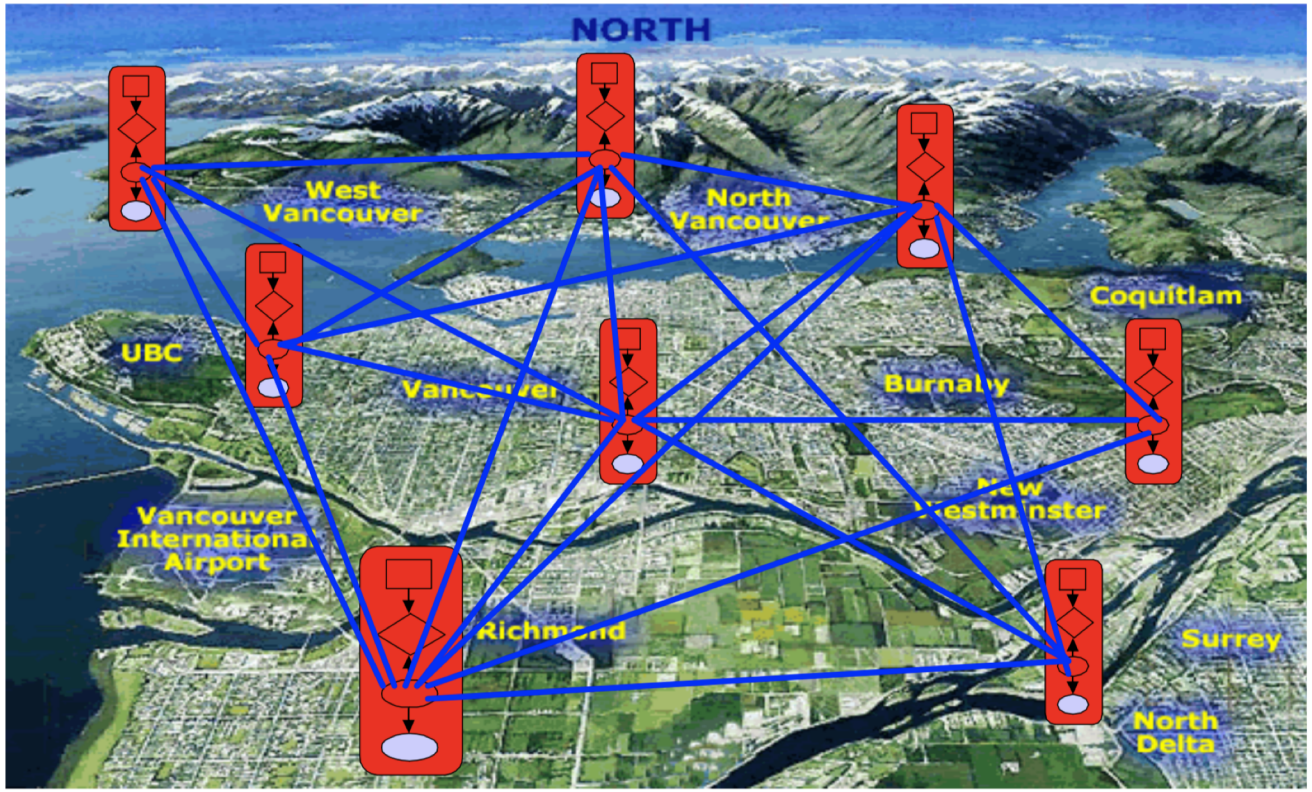
Active learning and surveillance

- In network with thousands of cameras, which camera views should be presented to the human operator?



Active learning and sensor networks

- How doe we optimally choose among a subset of sensors in order to obtain the best understanding of the world while minimizing resource expenditure (power, bandwidth, distractions)?



Active learning example