Lecture 11: Clustering Introduction and Projects
Machine Learning

Andrew Rosenberg

March 12, 2010

Last Time

- Junction Tree Algorithm
  - Efficient Marginals in Graphical Models

Today

- Clustering
- Project Details

Clustering

- Clustering is an **unsupervised** Machine Learning application
- The task is to group similar entities into groups.

We do this all the time







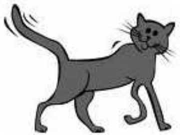
We do this all the time

We do this all the time

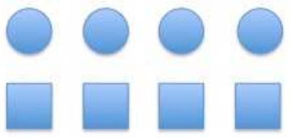




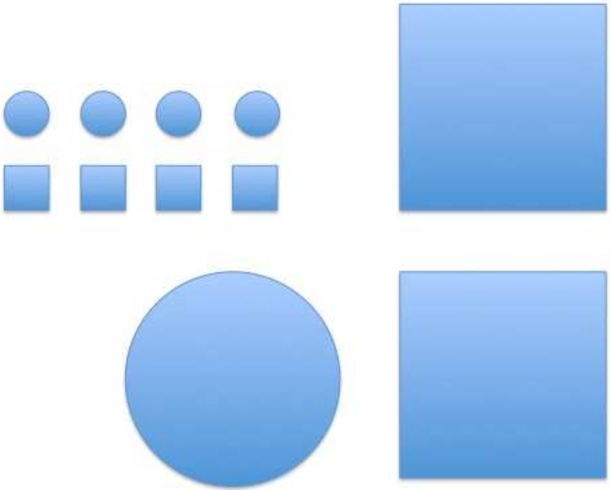


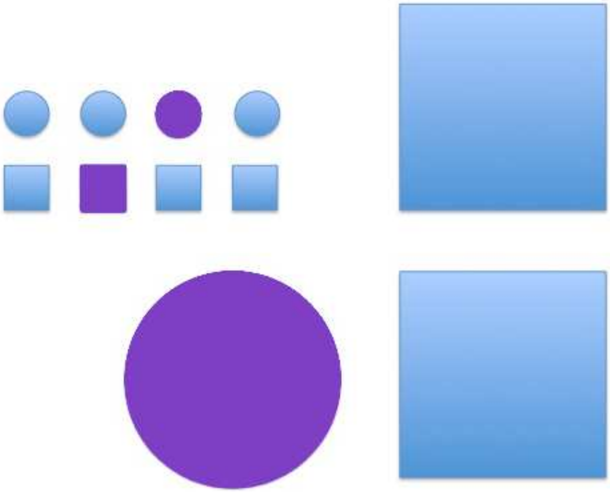
We can do this in many dimensions



We can do this to many degrees



We can do this in many dimensions



We do this all the time

How do we set this up computationally?

In Machine Learning, we optimize objective functions to find the best solution.

- Maximum Likelihood (for Frequentists)
- Maximum A Posteriori (for Bayesians)
- Empirical Risk Minimization
- Loss function Minimization

What makes a good cluster?
How do we define loss or likelihood in a clustering solution?

Cluster Evaluation

- **Intrinsic** Evaluation
  - Evaluate the compactness of the clusters
- **Extrinsic** Evaluation
  - Compare the results to some **gold standard** – labeled data.
  - (Not covered today)

Intrinsic Evaluation

- Intercluster Variability (IV)
  - How different are the data points within the same cluster
- Extracluster Variability (EV)
  - How different are the data points that are in distinct clusters

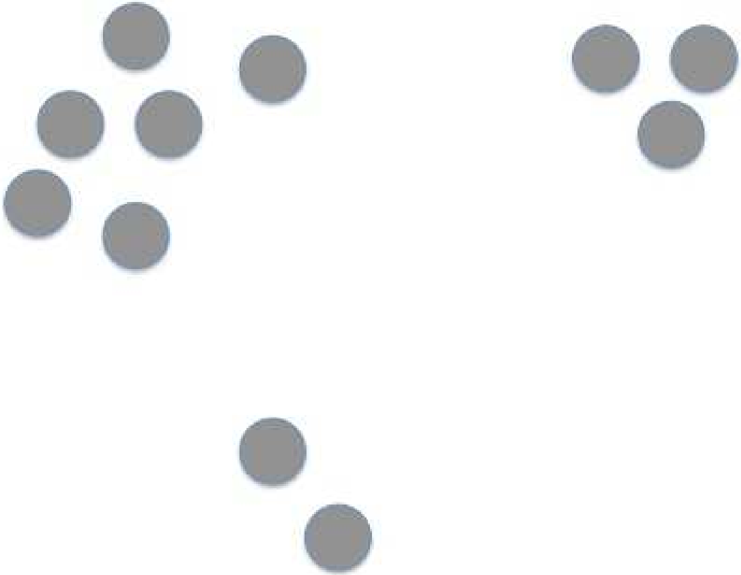
Minimize IV while maximizing EV.
Minimize $\frac{IV}{EV}$

$$IV = \sum_{C} \sum_{x \in C} d(x, c)$$

$$d(x, c) = ||x - c||$$

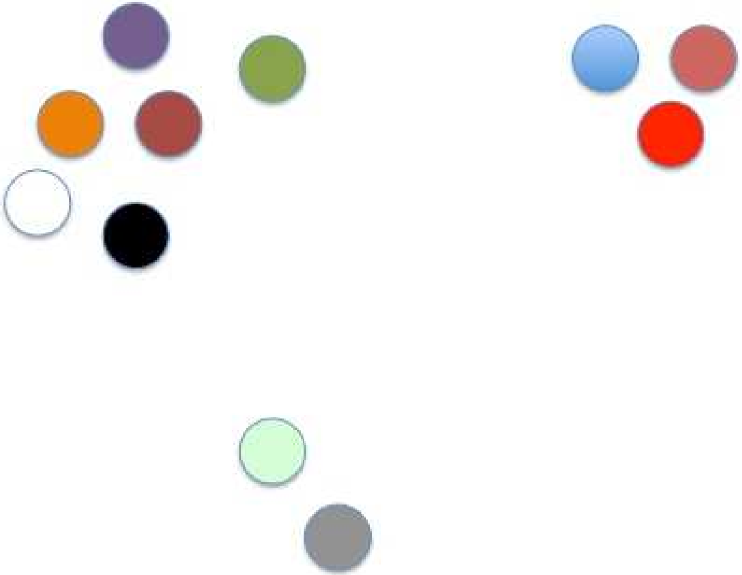
Degenerate Clustering Solutions

One Cluster



Degenerate Clustering Solutions

N Clusters

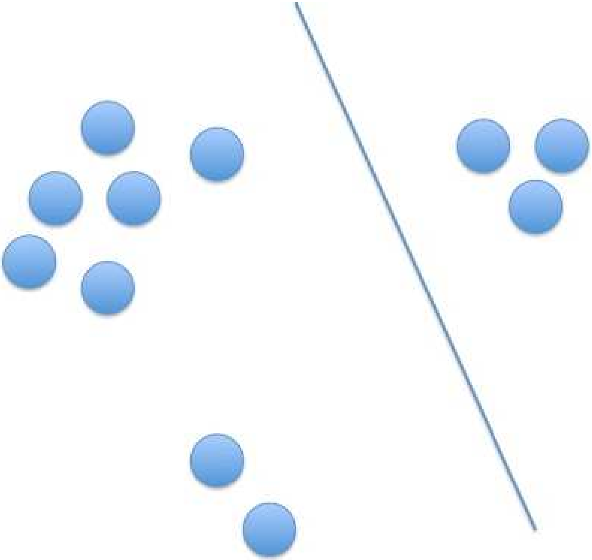


Clustering Approaches

- Hierarchical Clustering
- Partitional Clustering

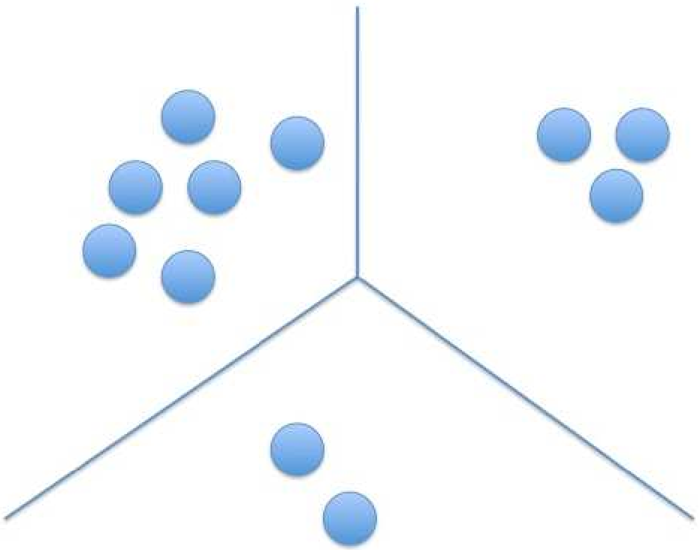
Hierarchical Clustering

Recursive Partitioning

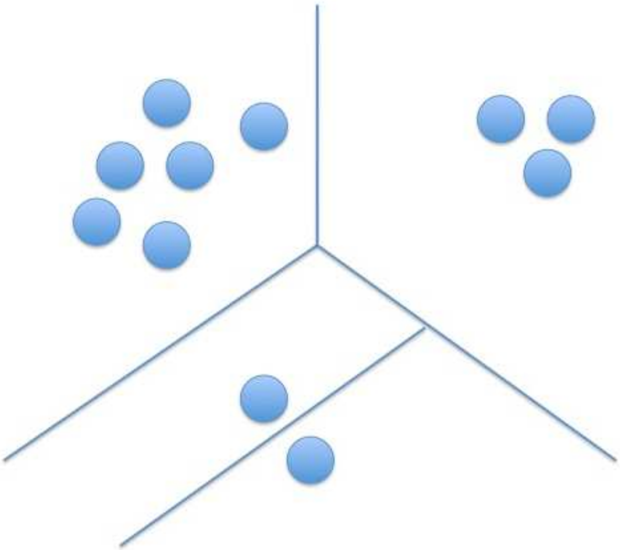


Hierarchical Clustering

Recursive Partitioning

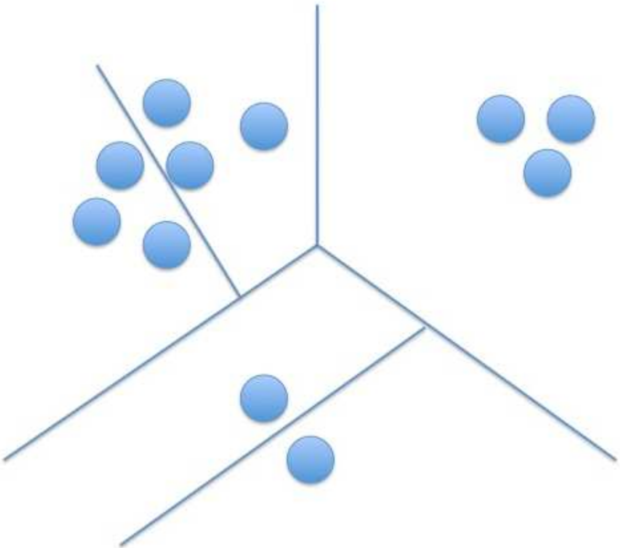


Recursive Partitioning

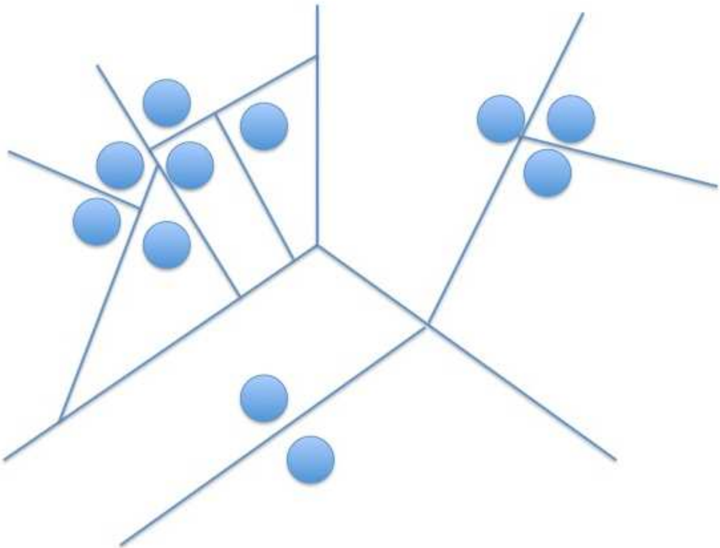


Hierarchical Clustering

Recursive Partitioning

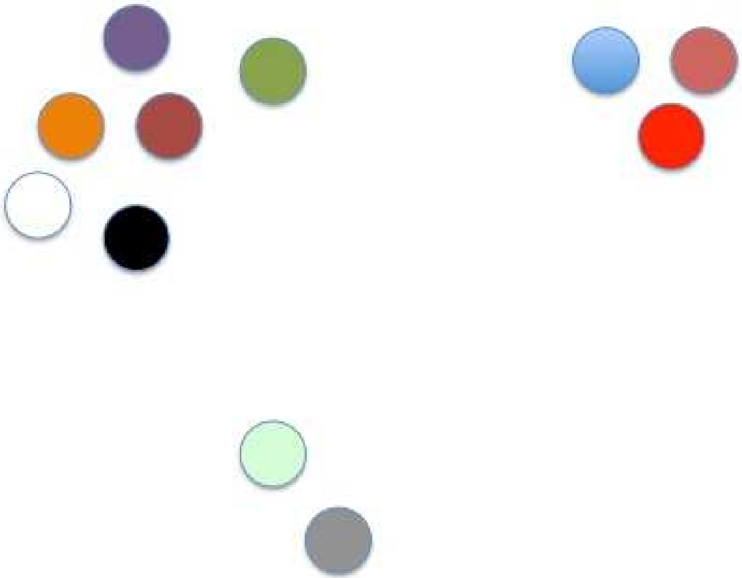


Recursive Partitioning



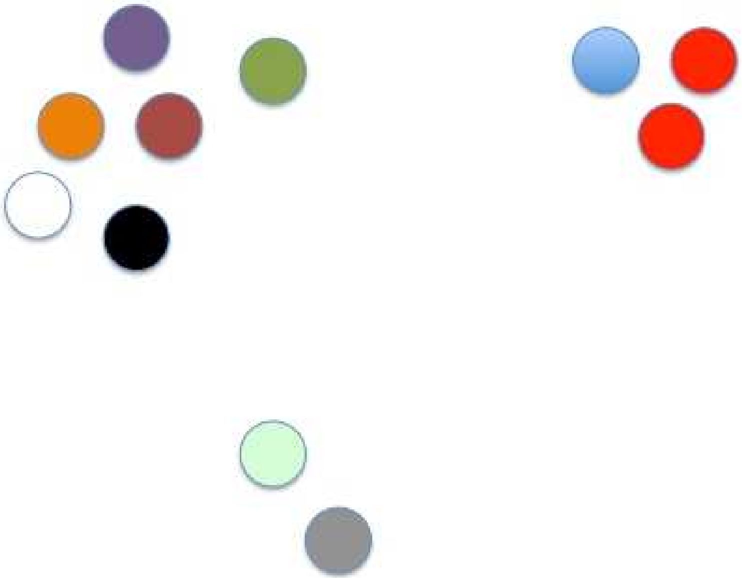
Hierarchical Clustering

Agglomerative Clustering

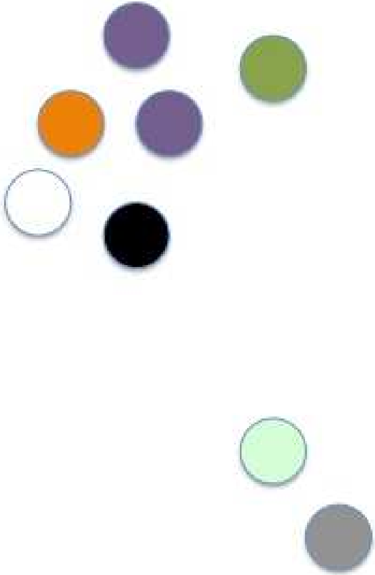


Hierarchical Clustering

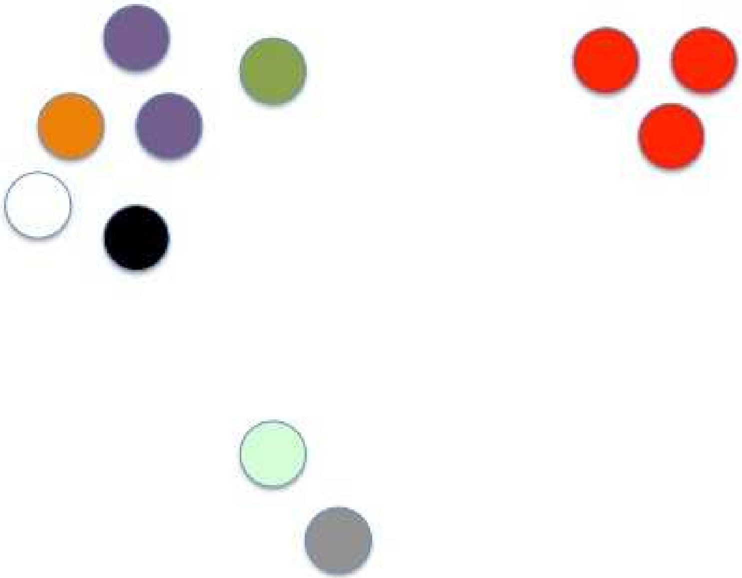
Agglomerative Clustering



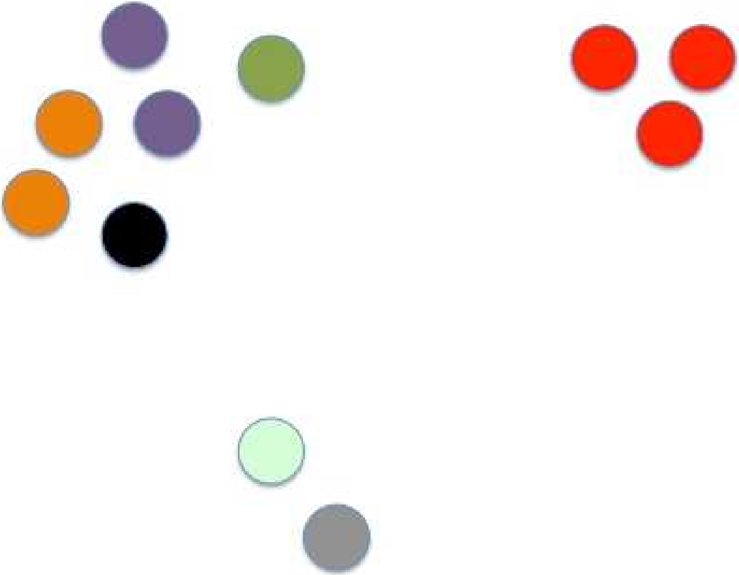
Agglomerative Clustering



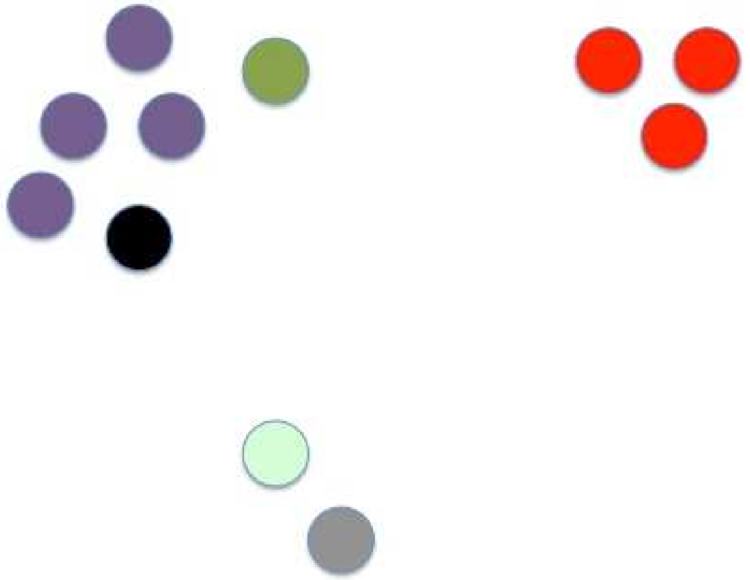
Agglomerative Clustering



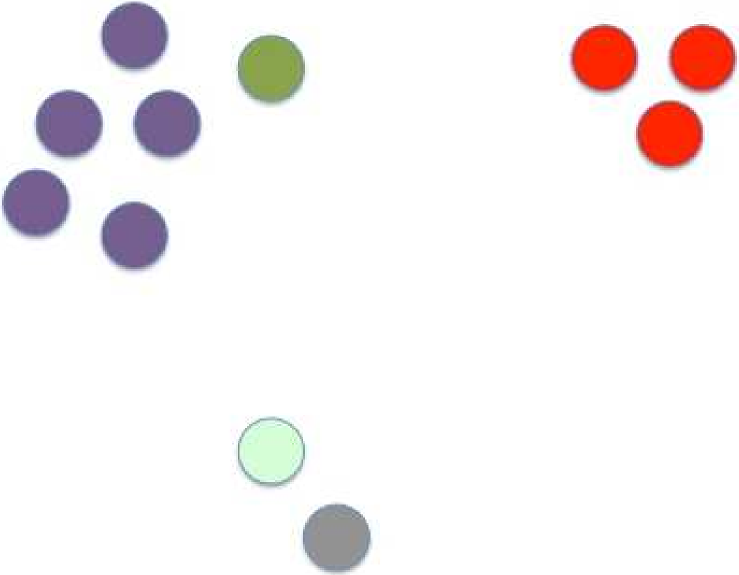
Agglomerative Clustering



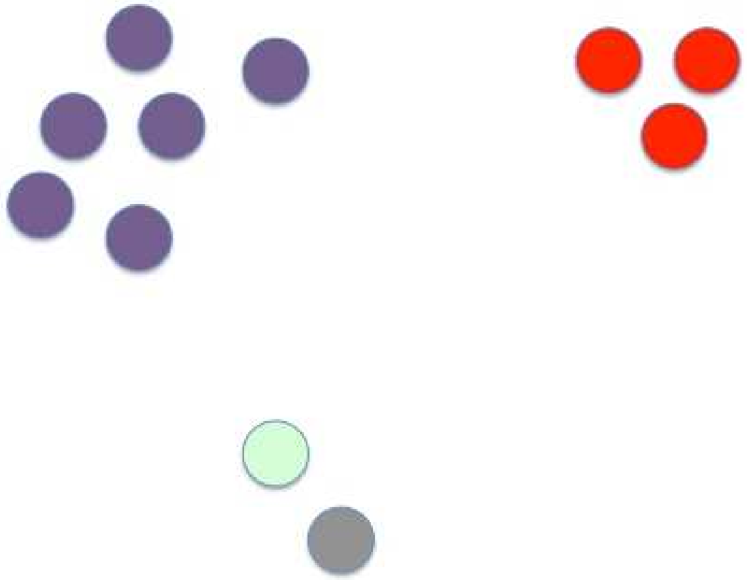
Agglomerative Clustering



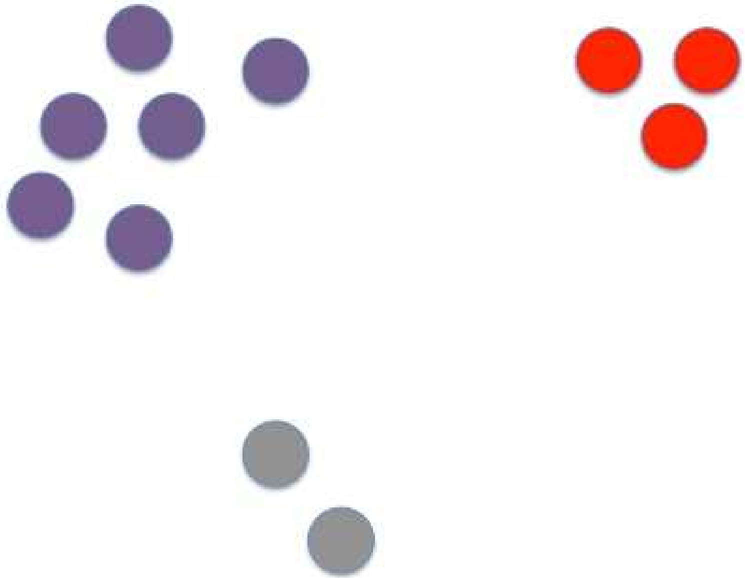
Agglomerative Clustering



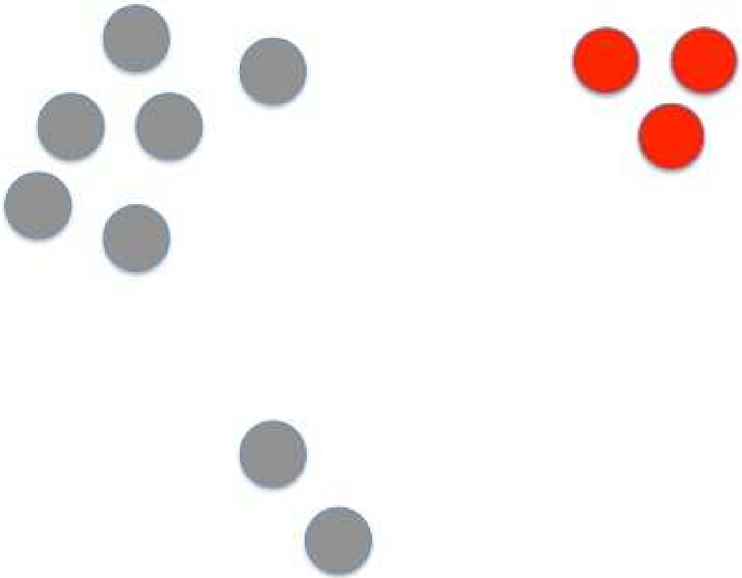
Agglomerative Clustering



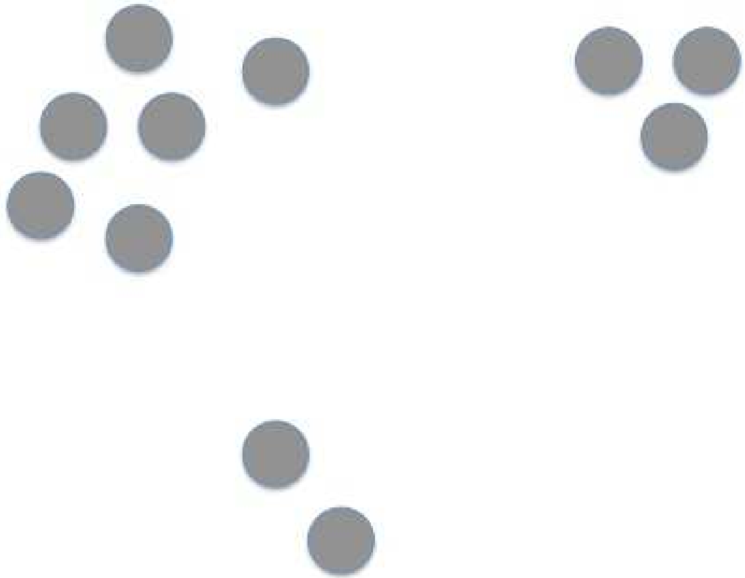
Agglomerative Clustering



Agglomerative Clustering



Agglomerative Clustering

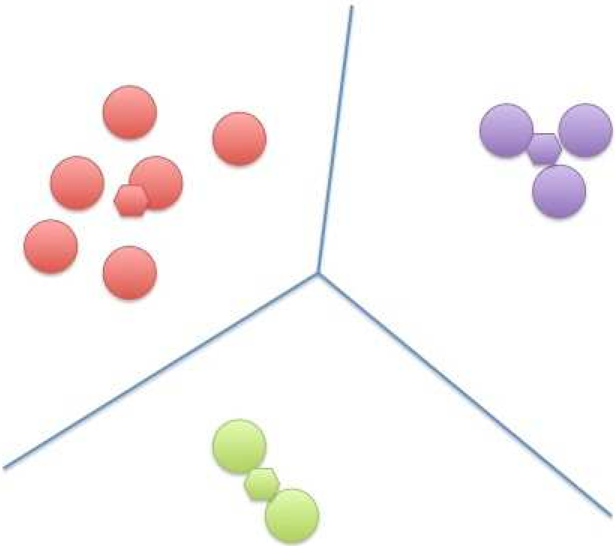


K-Means Clustering

K-Means clustering is a **Partitional** Clustering Algorithm.

- Identify different partitions of the space for a fixed number of clusters
- Input: a value for K – the number of clusters.
- Output: the K centers of clusters – **centroids**

K-Means Clustering



K-Means Clustering

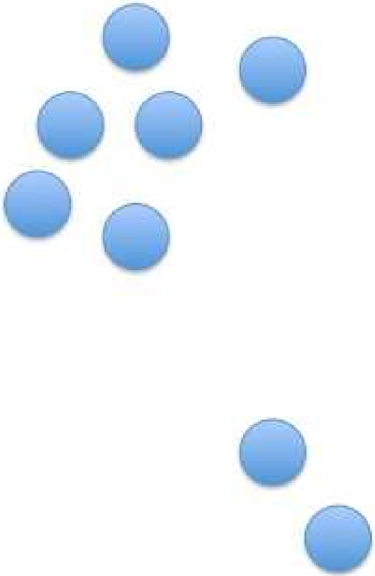
Algorithm:

- Given an integer K specifying the number of clusters.
- Initialize K cluster **centroids**
  - Select K points from the data set at random
  - Select K points from the space at random
- For each point in the data set, assign it to the cluster whose center it is closest to.
  - $\text{argmin}_{C_i} d(x, C_i)$
- Update the centroid based on the points assigned to the cluster.
  - $c_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$
- If any data point has changed clusters, repeat.

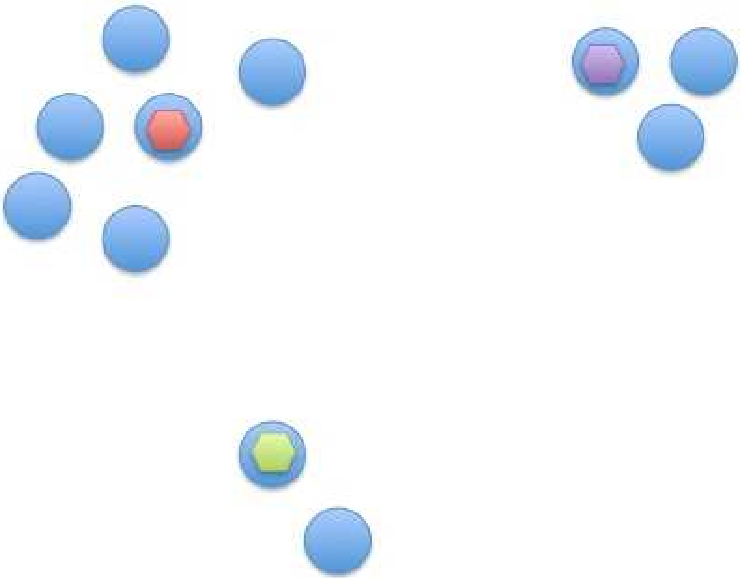
Why does K-Means Work?

- When an assignment is changed, the sum of squared distances of the data point to its assigned cluster is reduced.
  - IV is reduced.
- When a cluster centroid is moved the sum of squared distances of the data points within that cluster is reduced
  - IV is reduced.
- At convergence we have found a local minimum of IV

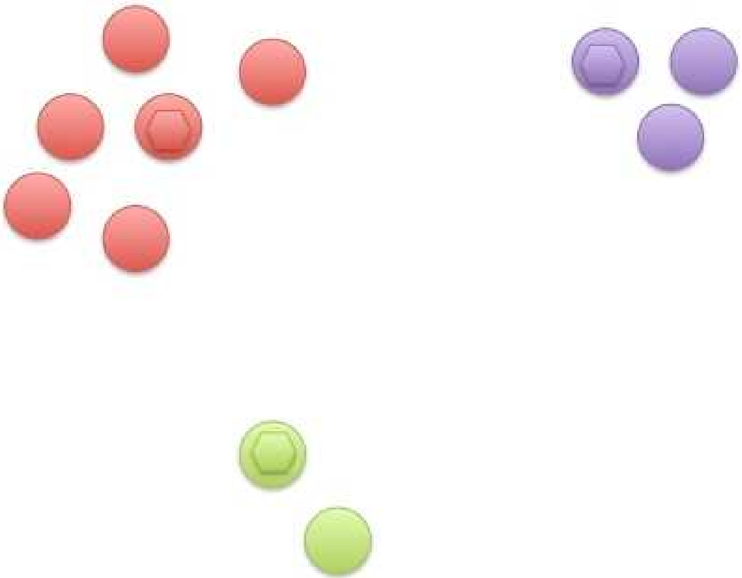
K-Means Clustering



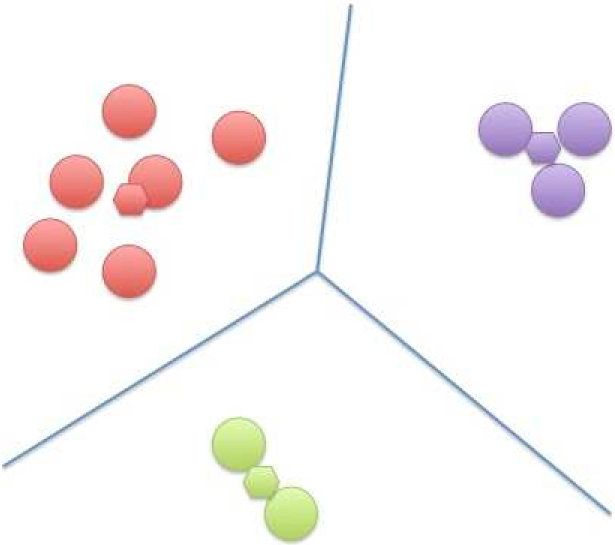
K-Means Clustering



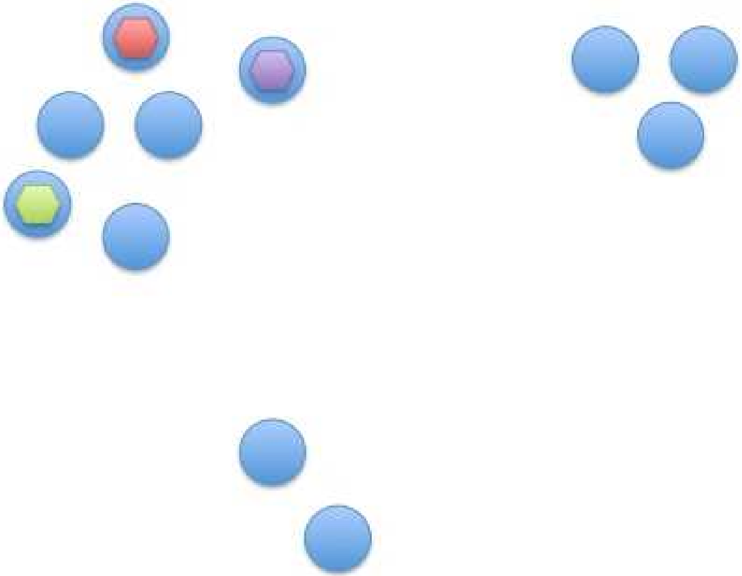
K-Means Clustering



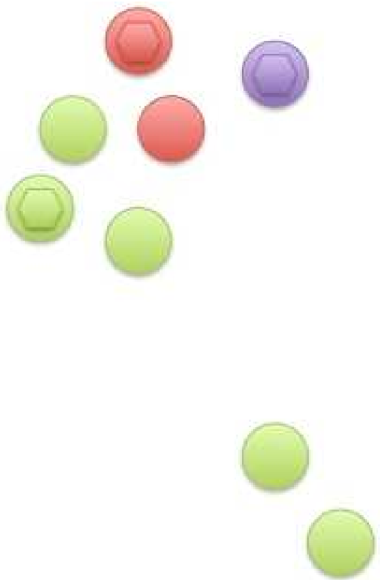
K-Means Clustering



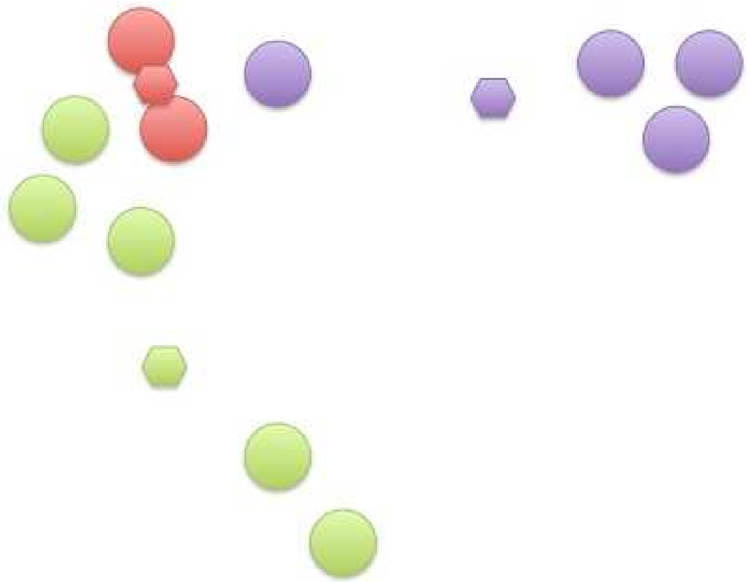
K-Means Clustering



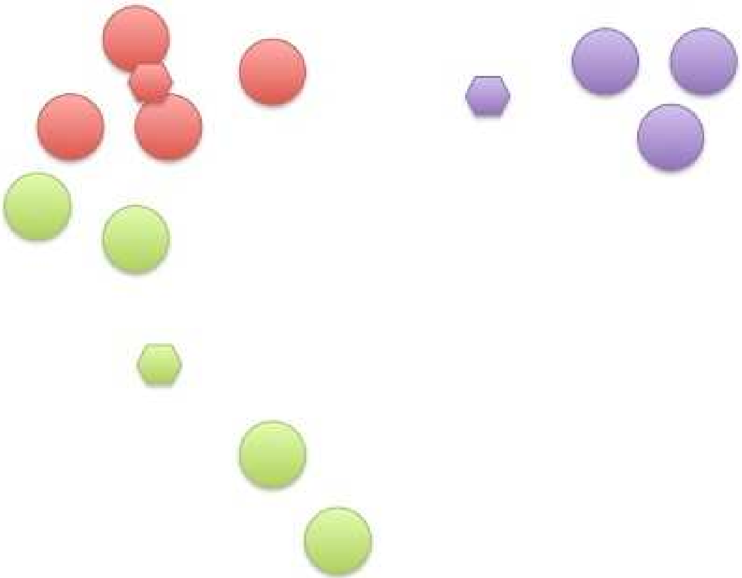
K-Means Clustering



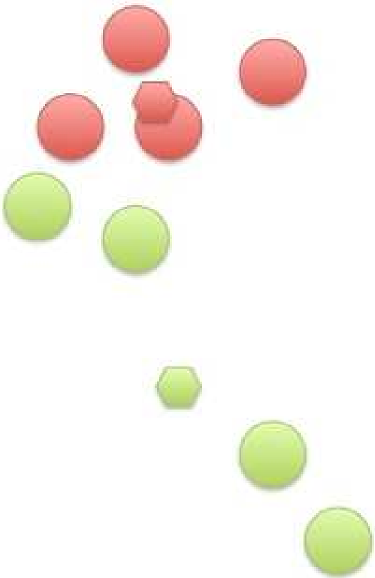
K-Means Clustering



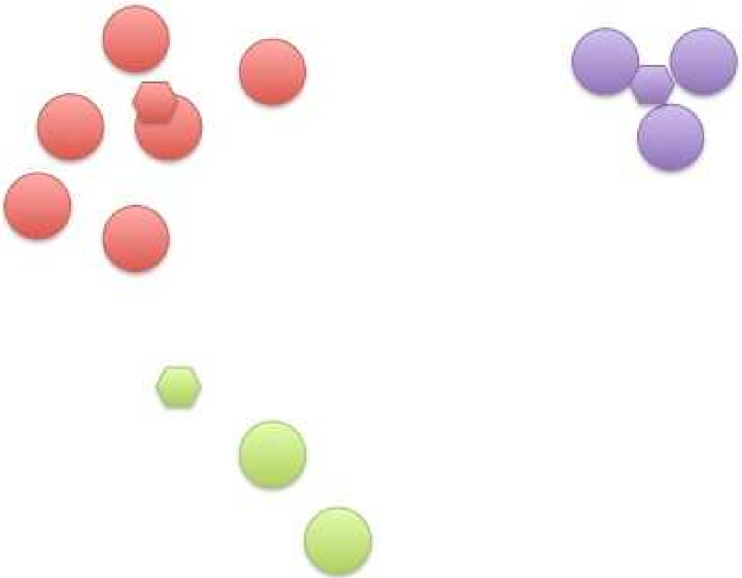
K-Means Clustering



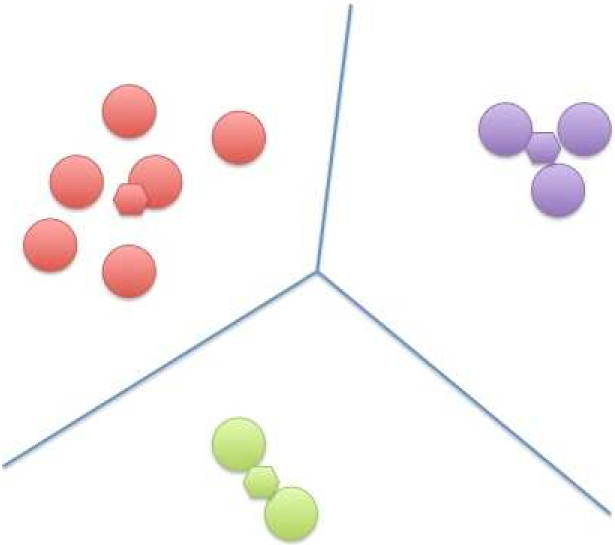
K-Means Clustering



K-Means Clustering



K-Means Clustering



Soft K-Means

- In K-means, we forced every data point to be the member of exactly one cluster.
- We can relax this constraint.

$$p(x, C_i) = \frac{d(x, c_i)}{\sum_j d(x, c_j)}$$

$$p(x, C_i) = \frac{\exp\{-d(x, c_i)\}}{\sum_j \exp\{-d(x, c_j)\}}$$

Based on minimizing entropy of cluster assignment.

- We still define a cluster by a centroid, but we calculate the centroid as a **weighted** center of **all** the data points.

$$c_i = \frac{\sum_x x \cdot p(x, C_i)}{\sum_x p(x, C_i)}$$

- Convergence is based on a stopping threshold rather than changing assignments.

Potential Problems with K-Means

Optimal?

- K-means approaches a local minimum, but this is not guaranteed to be globally optimal.
- Could you design an approach which is globally optimal?

Consistent?

- Different starting clusters can lead to different cluster solutions

Potential Problems with K-Means

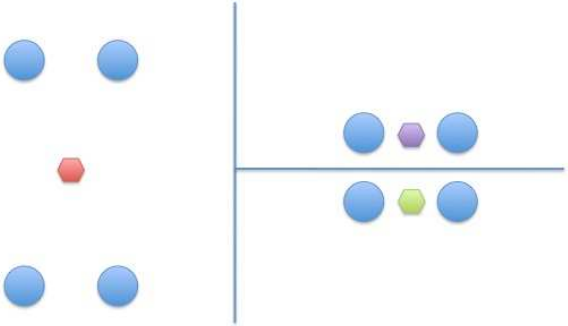
Optimal?

- K-means approaches a local minimum, but this is not guaranteed to be globally optimal.
- Could you design an approach which is globally optimal?
- Sure, in NP.

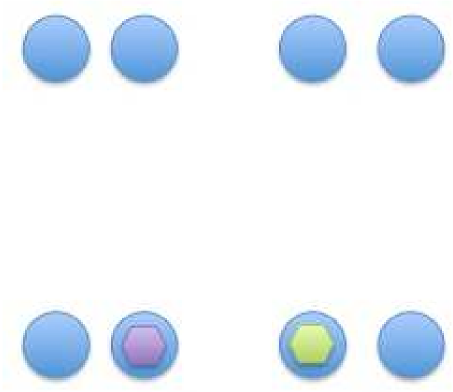
Consistent?

- Different starting clusters can lead to different cluster solutions

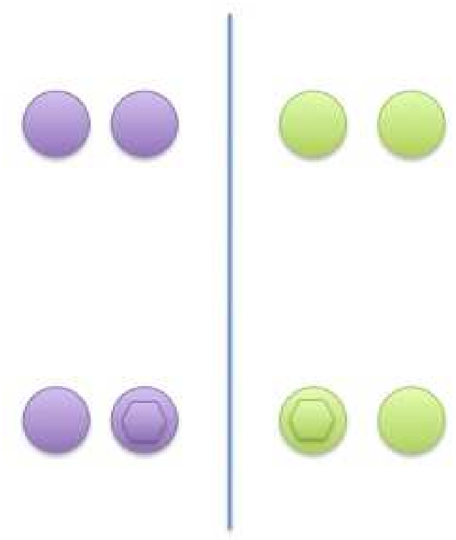
Suboptimality in K-Means



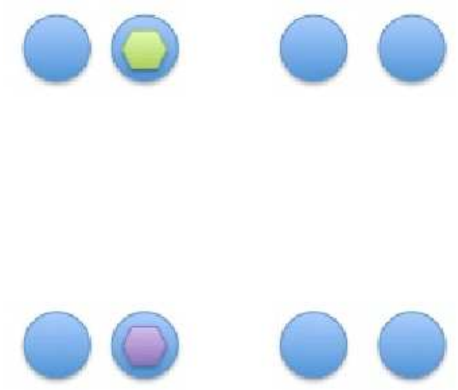
Inconsistency in K-Means



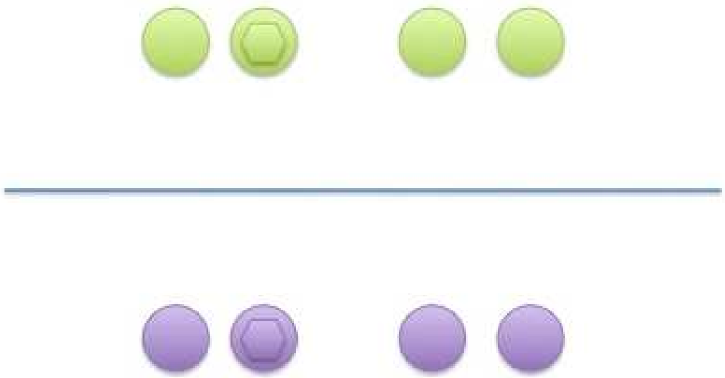
Inconsistency in K-Means



Inconsistency in K-Means



Inconsistency in K-Means



More Clustering

- K-Nearest Neighbors
- Gaussian Mixture Models
- Spectral Clustering

We will return to these.

The Project

- Research Paper
- Project

Research Paper

- 8-10 pages
- Reporting on work in 4-5 papers.
- Scope:
  - One application area
  - **or** One technique

Research Paper: Application Areas

Identify an application that has made use of machine learning and discuss how.

Graphics

- Object Recognition
- Optical Character Recognition
- Superresolution
- Segmentation

Natural Language Processing

- Parsing
- Sentiment Analysis
- Information Extraction

Speech

- Recognition
- Synthesis
- Discourse Analysis
- Intonation

Game Playing

- Scrabble
- Craps
- Prisoner's Dilemma

Financials

- Stock Prediction

Review Systems

- Amazon
- Netflix
- Facebook

Research Paper: Techniques

Identify a machine learning technique. Describe its use and variants.

- L1-regularization
- Non-linear Kernels
- Loopy Belief Propagation
- Non-parametric Belief Propagation
- Soft-Decision Trees
- Analysis of Neural Network Hidden Layers
- Structured Learning
- Generalized Expectation
- Evaluation Measures
    - Cluster Evaluation
    - Semi-supervised Evaluation
- Graph Embedding
- Dimensionality Reduction
- Feature Selection
- Graphical Model Construction
- Non-parametric Bayesian Methods
- Latent Dirichlet Allocation

Project

- Run a Machine Learning Experiment
  - Identify a problem/task data set.
  - Implement one or more ML algorithm
  - Evaluate the approach.
- Write a Report of the Experiment
  - 4 pages including references.
  - Abstract – 1 paragraph summarizing the experiment
  - Introduction – describe the Problem
  - Data – Describe the data set, features extracted, etc.
  - Method – Describe the algorithm/approach
  - Results – Present and discuss results
  - Conclusion – Summarize the experiment and results.

Project Ideas: Tasks

Projects can take any combination of Tasks and Approaches

- Graphics
  - Object Classification
  - Facial Recognition
  - Fingerprint Identification
  - Optical Character Recognition
  - Handwriting recognition
    - (for languages/character systems other than English...)
- Language
  - Topic Classification
  - Sentiment Analysis
  - Speech Recognition
  - Speaker Identification
  - Punctuation Restoration
  - Semantic Segmentation
  - Recognition of Emotion, Sarcasm, etc.
  - SMS Text normalization
  - Chat participant identification
  - Twitter classification/threading

- Games
  - Chess
  - Checkers
  - Poker (Poker Academy Pro)
  - Blackjack
- Recommenders (Collaborative Filtering)
  - Netflix
  - Courses
  - Jokes
  - Books
  - Facebook?
- Video Classification
  - Motion classification
  - Segmentation

Bye

- Next
  - Hidden Markov Models
  - Viterbi Decoding