Machine Learning for Data Science (CS4786)
Lecture 1

Tu-Th 11:40AM to 12:55 PM

Holister B14

Instructor : Karthik Sridharan

Welcome the first lecture!

The Awesome TA's

TA's:

1. Geoff Pleiss
2. Davis Wertheimer
3. Valts Blukis
4. Andrew Mullen
5. Tianwei Huang

TA Consultants:

1. Keelan Cosgrove
2. Michelle Yuan
3. Siddarth Reddy
4. Junia George
5. Mukund Sudarshan
6. Claire Liang
7. Patrick Nicholson

Course Information

- Course webpage is the official source of information: http://www.cs.cornell.edu/Courses/cs4786/2016fa

- Join Piazza: https://piazza.com/class/irw7q5gjfex48m

- TA office hours will start from next week

- While the course is not coding intensive, you will need to do some light coding.

Sylabus

1. Dimensionality Reduction:
    1. Principal Component Analysis (PCA)
    2. Canonical Component Analysis (CCA)
    3. Random Projections
    4. Kernel Methods/Kernel PCA
2. Clustering and More:
    1. Single Link Clustering
    2. K-means Algorithm
    3. Spectral Clustering
    4. Gaussian Mixture Models and Other Mixture Models
    5. Latent Dirichlet Allocation
3. Probabilistic Modeling and Graphical Models:
    1. MLE Vs MAP Vs Bayesian Methods
    2. EM Algorithm
    3. Graphical Models
        1. Hidden Markov Models
        2. Exact Inference: Variable Elimination, Belief Propogation
        3. Learning in Graphical Models
        4. Approximate Inference

Course Information

- Six Assignments worth 60% of the grades, done individually.
- Two competitions worth 40% of the grade, done in groups of at most 4
- TA office hours will start from next week
- Course is not coding intensive, light coding needed though (language your choice)

Assignments

- Diagnostic assignment 0 is out: for our calibration.
  - Students who want to take course for credit need to submit this, **only then you will be added to CMS**.
  - Hand in your assignments beginning of class on August 30th.
  - **Has to be done individually**
  - Write your **full name** and **net id** on the first page of the hand-in. You will be added to cms based on this.

Assignments

- Besides the diagnostic assignments, there are 6 other assignments **to be done individually**
- The 6 assignments are worth 60% of your grades.
- Rough timeline:

1. Assignment 1: Out: September 1st Due: September 8th
2. Assignment 2: Out: September 13th Due: September 20th
3. Assignment 3: Out: September 22nd Due: September 29th
4. Assignment 4: Out: September 29th Due: October 6th
5. Assignment 5: Out: October 18th Due: October 25th
6. Assignment 6: Out: November 3rd Due: November 15th

Competitions

- 2 competition/challenges, worth 40% of total course grade
  - Competition I: Clustering challenge (Due Mid Oct)
  - Competition II: Graphical Model centric challenge (Due Nov end)

- Will be hosted on "In class Kaggle"!

- 40% of the competition grades for kaggle score/performance

- 60% of the competition grades for report.

- Mid competition, a one page preliminary report (to be submitted individually) explaining work done so far by each individual in the group. Worth 10% of th competition grade.

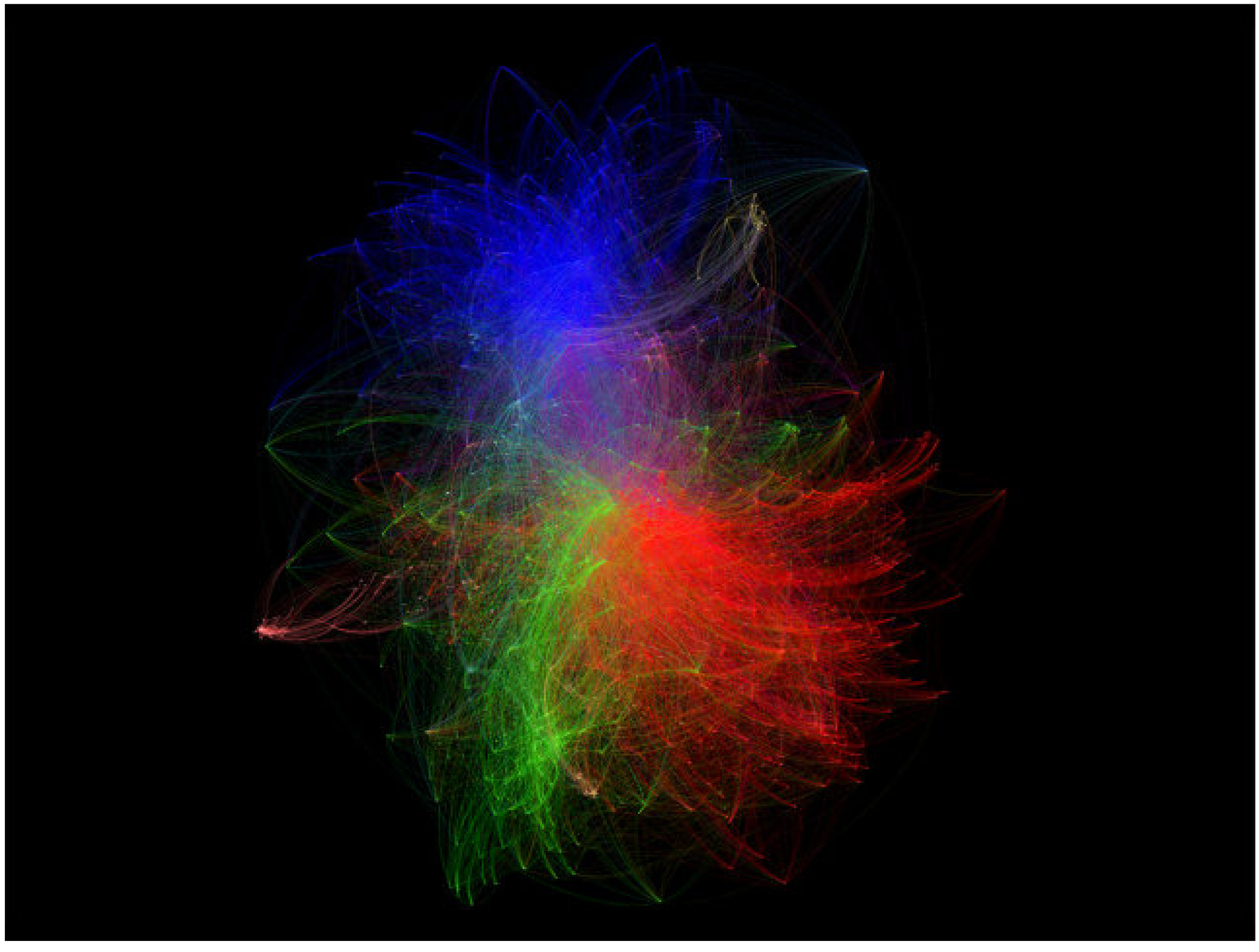
- Groups of size at most 4.

Lets get started …

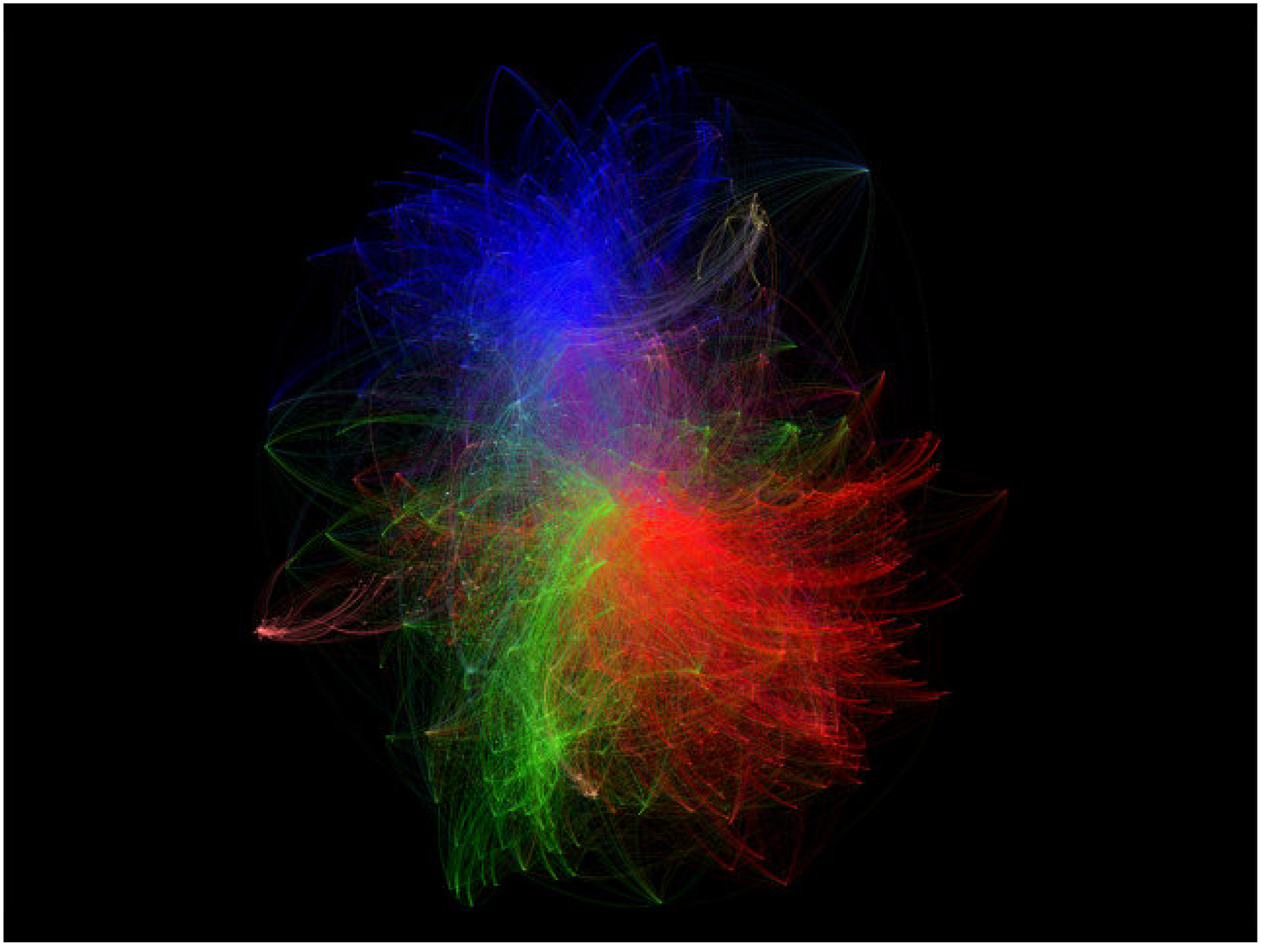
Data Deluge

- Each time you use your credit card: who purchased what, where and when
- Netflix, Hulu, smart TV: what do different groups of people like to watch
- Social networks like Facebook, Twitter, . . . : who is friends with who, what do these people post or tweet about
- Millions of photos and videos, many tagged
- Wikipedia, all the news websites: pretty much most of human knowledge

Guess?



Social Network of Marvel Comic Characters!



by Cesc Rosselló, Ricardo Alberich, and Joe Miro from the University of the Balearic Islands

What can we learn from all this data?

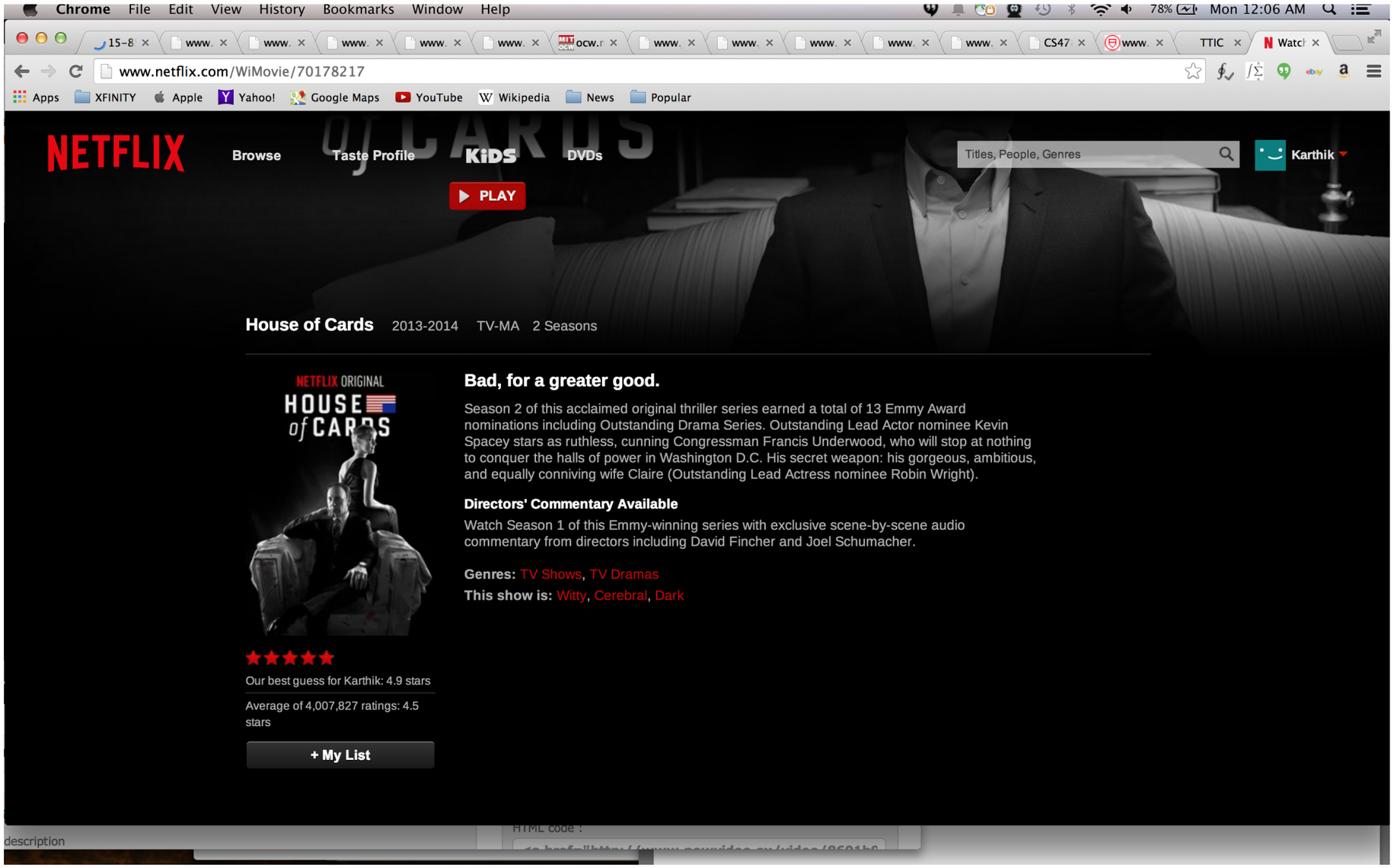
What is Machine Learning?

Use data to automatically learn to perform tasks better.

Close in spirit to T. Mitchell's description

Where is it used ?

Movie Rating Prediction



Where is it used ?

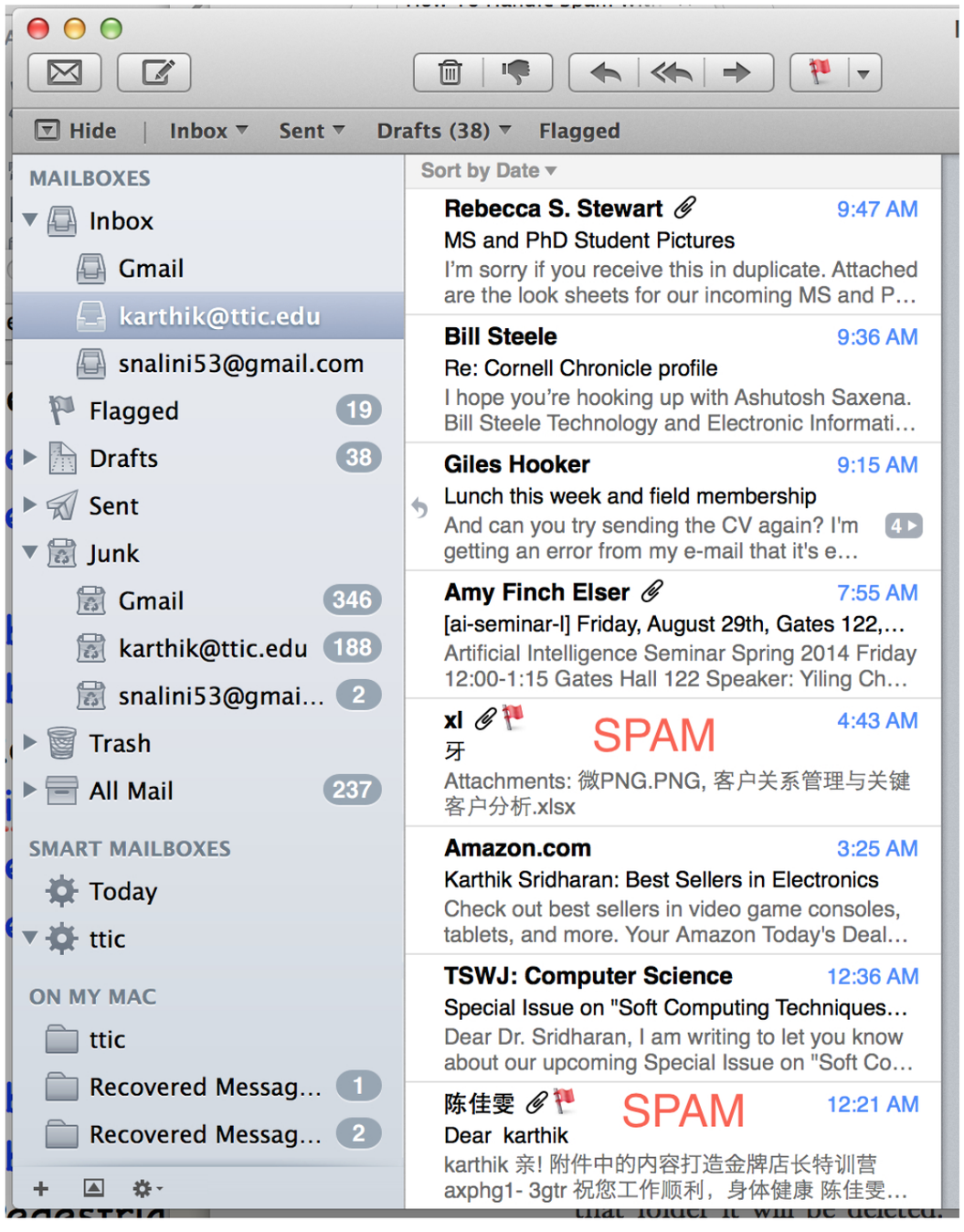
Pedestrian Detection

Where is it used ?

Market Predictions

Where is it used ?

Spam Classification



More Applications

- Each time you use your search engine
- Autocomplete: Blame machine learning for bad spellings
- Biometrics: reason you shouldn't smile
- Recommendation systems: what you may like to buy based on what your friends and their friends buy
- Computer vision: self driving cars, automatically tagging photos
- Topic modeling: Automatically categorizing documents/emails by topics or music by genre
- …

Topics we will cover

unsupervised learning

1. Dimensionality Reduction:
   Principal Component Analysis (PCA), Canonical Component Analysis (CCA), Random projections, Compressed Sensing (CS), …

2. Clustering and Mixture models:
   k-means clustering, gaussian mixture models, single-link clustering, spectral clustering, …

3. Probabilistic Modeling & Graphical Models:
   Probabilistic modeling, MLE Vs MAP Vs Bayesian approaches, inference and learning in graphical models, Latent Dirichlet Allocation (LDA), Hidden Markov Models (HMM), …

Unsupervised Learning

Given (unlabeled) data, find useful information, pattern or structure

- Dimensionality reduction/compression : compress data set by removing redundancy and retaining only useful information
- Clustering: Find meaningful groupings in data
- Topic modeling: discover topics/groups with which we can tag data points

Dimensionality Reduction

- You are provided with n data points each in $\mathbb{R}^d$
- Goal: Compress data into n, points in $\mathbb{R}^K$ where $K << d$

Dimensionality Reduction

- You are provided with n data points each in $\mathbb{R}^d$
- Goal: Compress data into n, points in $\mathbb{R}^K$ where $K << d$
    - Retain as much information about the original data set
    - Retain desired properties of the original data set
- Eg. PCA, compressed sensing, ...

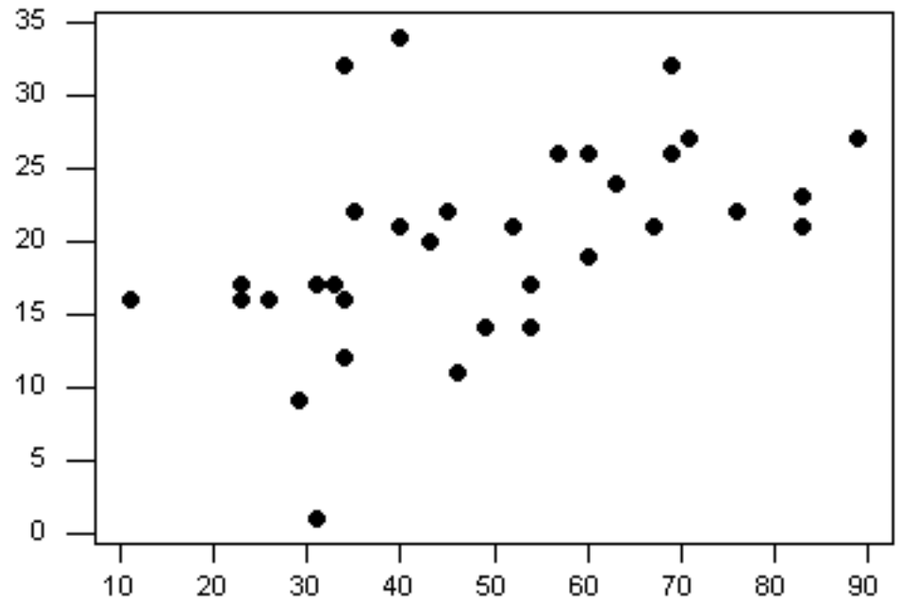
Principal Component Analysis (PCA)

Turk & Pentland'91

Eigen Face:

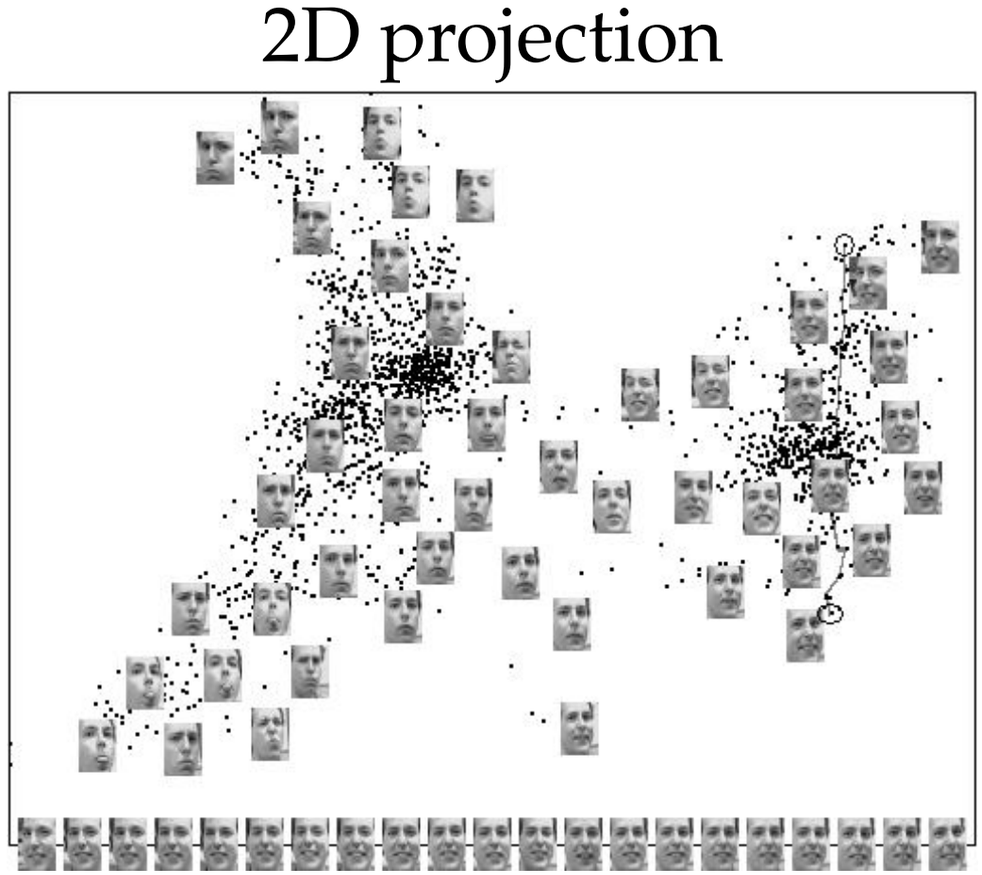
- Write down each data point as a linear combination of small number of basis vectors
- Data specific compression scheme
- One of the early successes: in face recognition: classification based on nearest neighbor in the reduced dimension space

Canonical Component Analysis (PCA)



- Extract common information between multiple sources views
- Noise specific to only one or subset of views is automatically filtered
- Success story: Speaker/speach recognition using both audio and video data

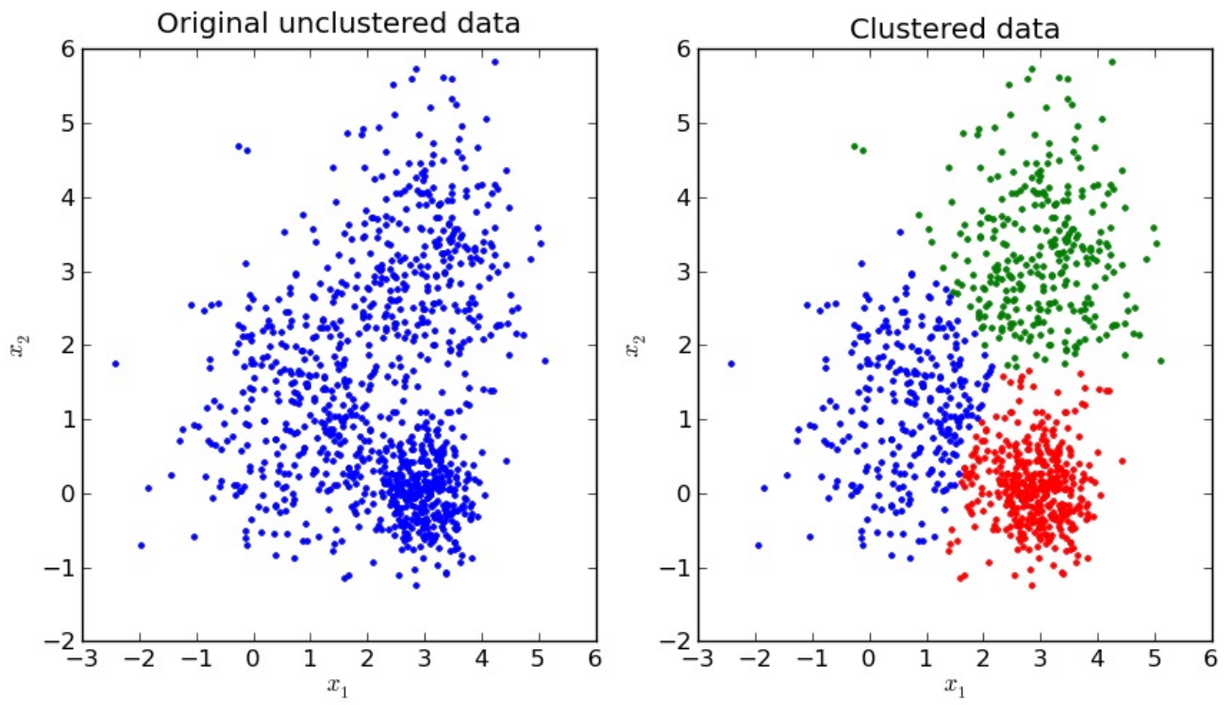
Data Visualization



- Help visualize data (in relation to each other)
- Preserve relative distances among data-points (at least close by ones)

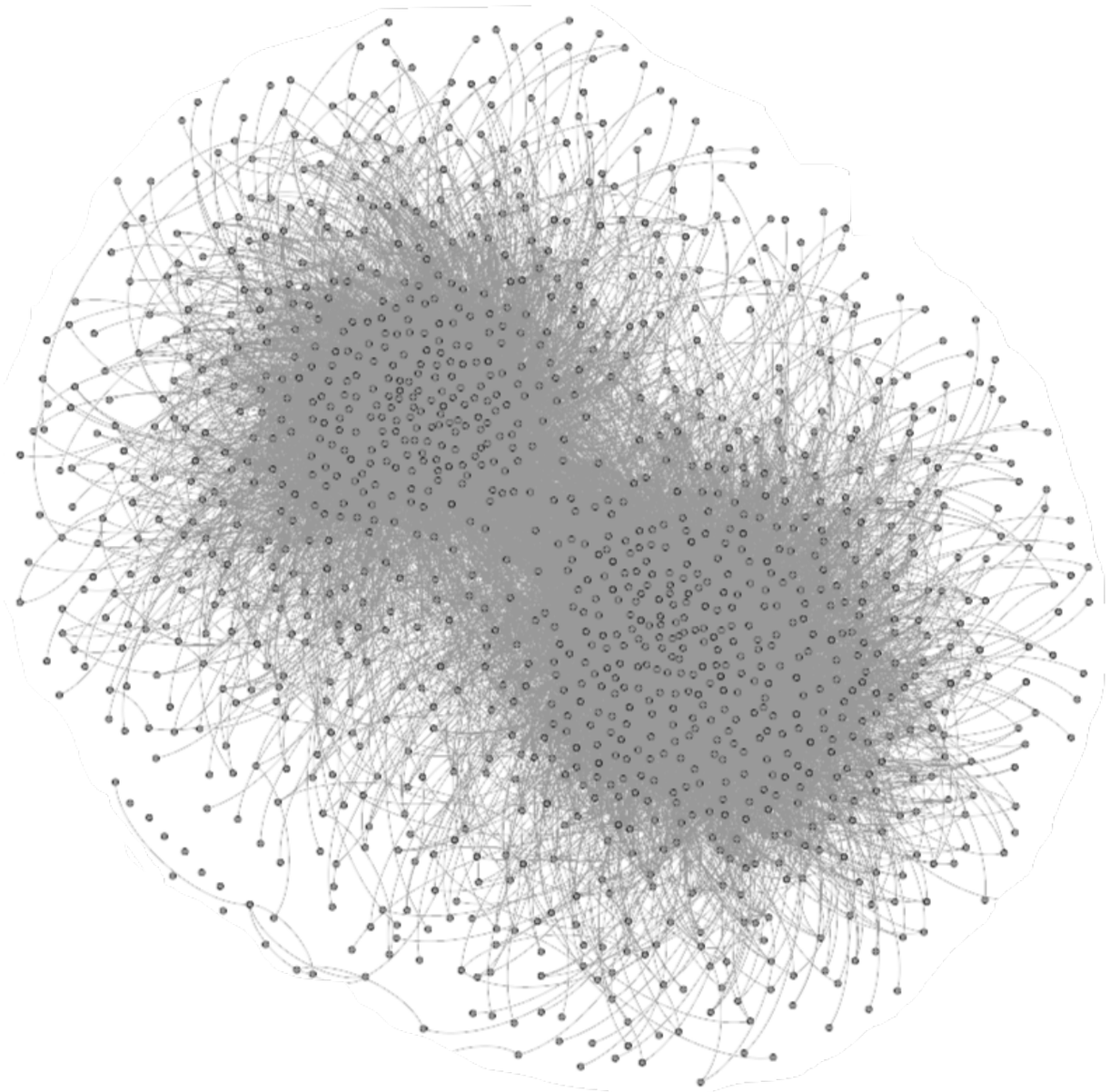
Clustering

K-means clustering

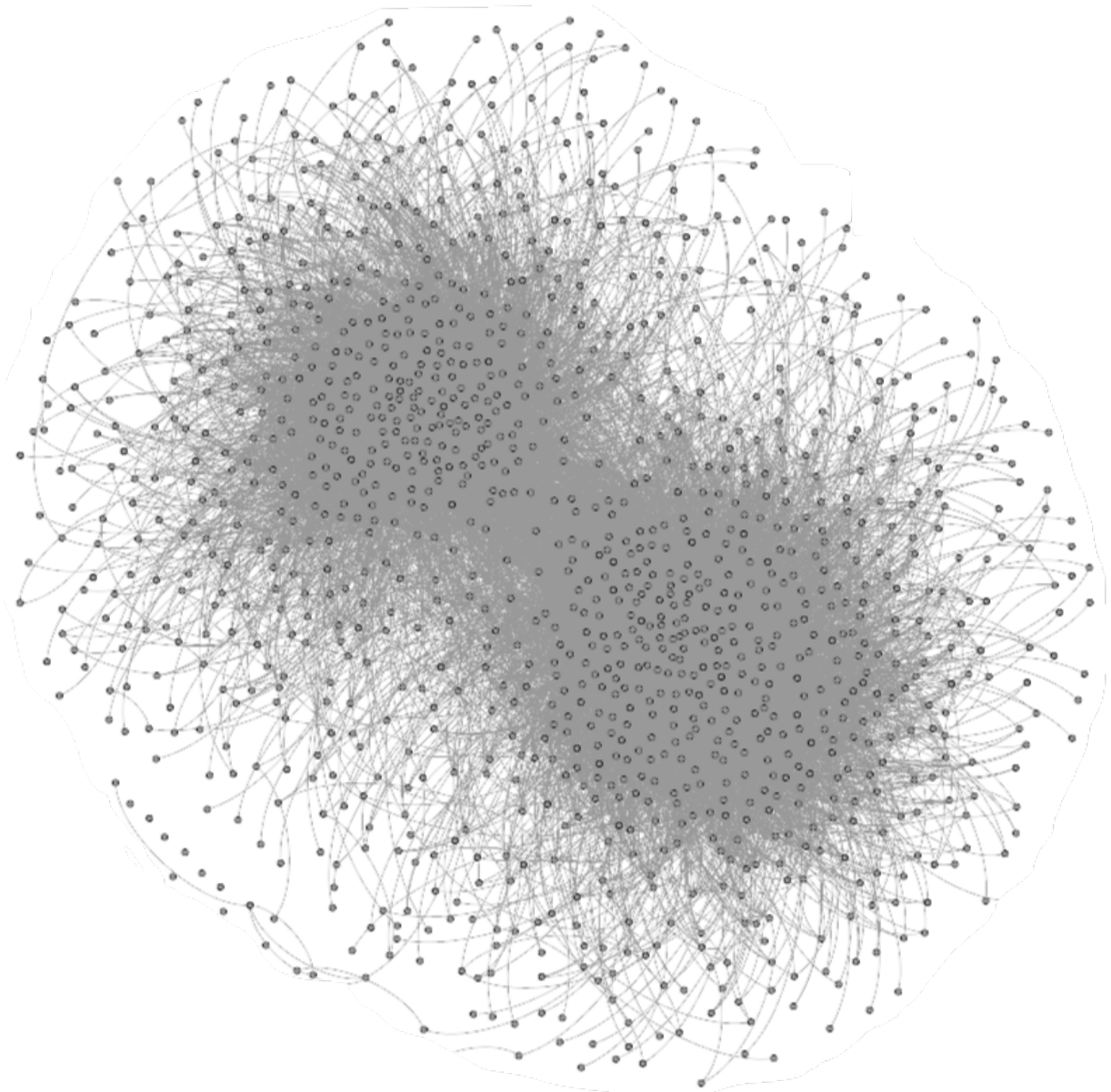


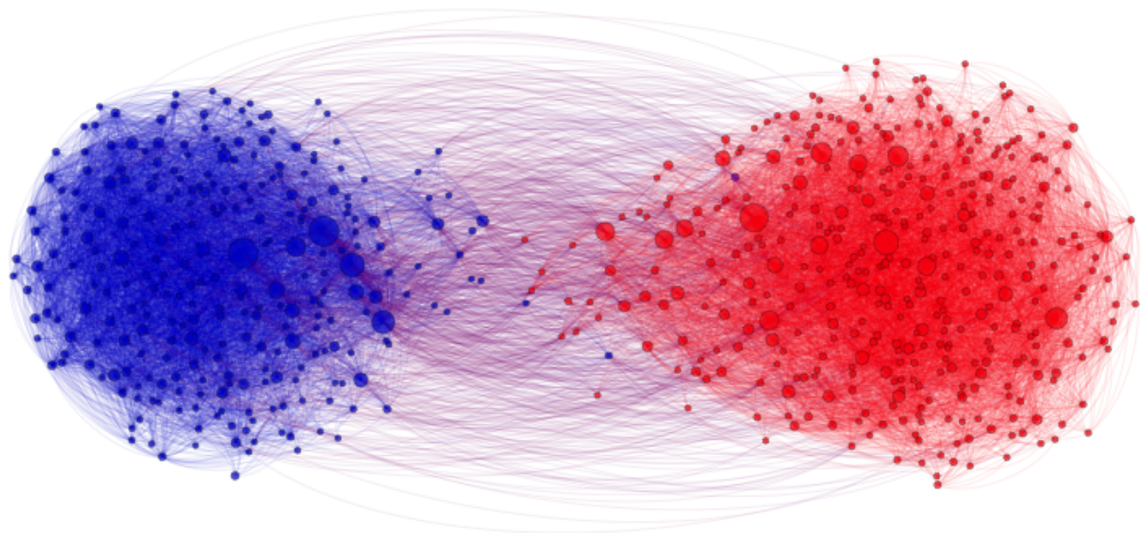
- Given just the data points group them in natural clusters
- Roughly speaking
  - Points within a cluster must be close to each other
  - Points between clusters must be separated
- Helps bin data points, but generally hard to do

Tell me who your friends are . . .



Tell me who your friends are . . .



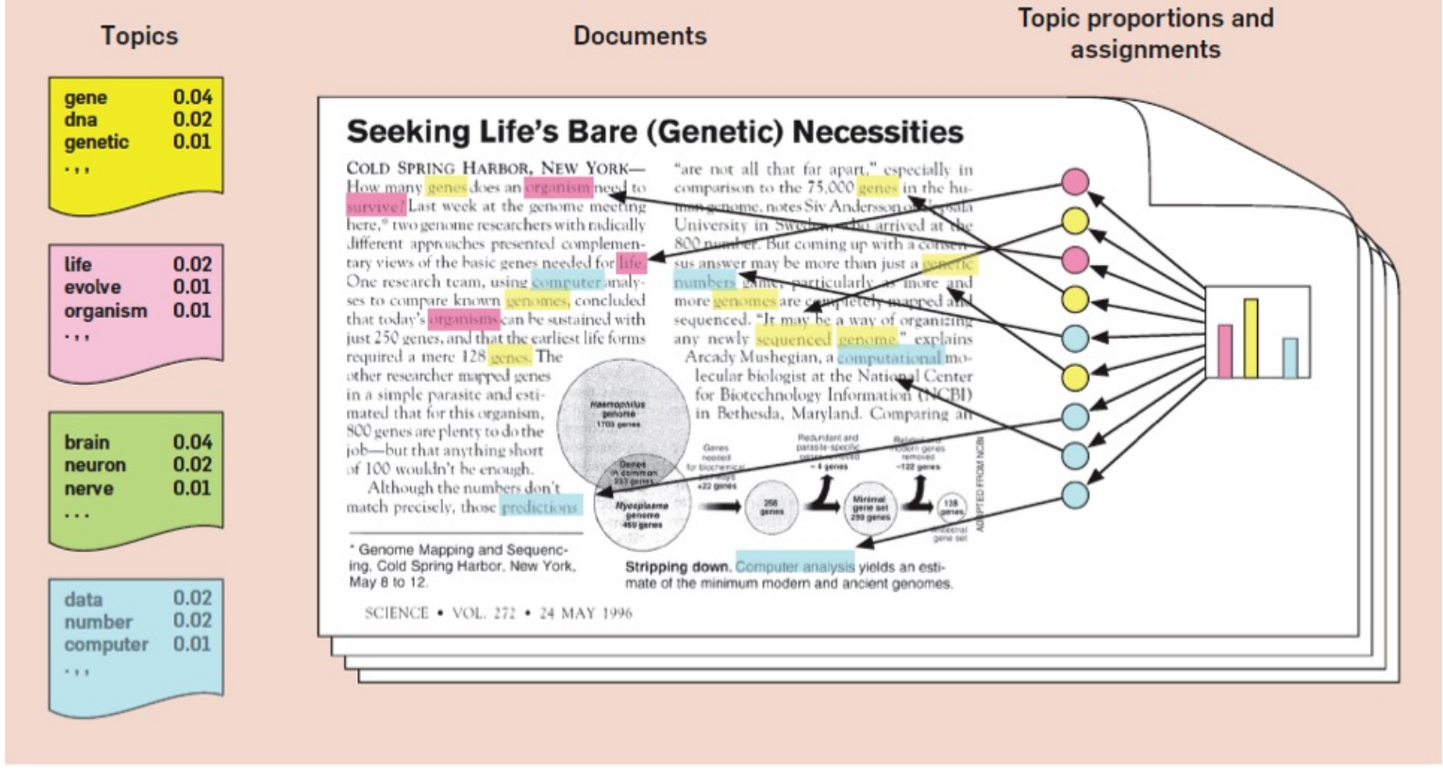


- Cluster nodes in a graph.
- Analysis of social network data.

Topic Modelling

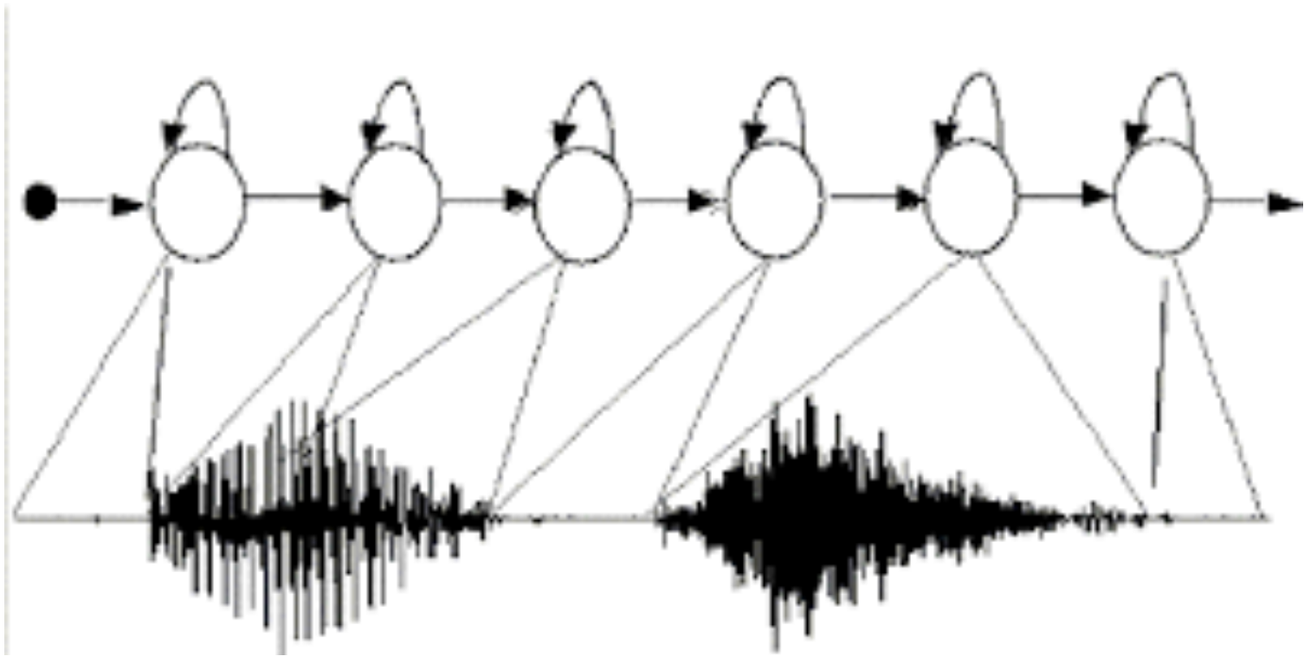
Blei, Ng & Jordan'06

Figure 1. The intuitions behind latent Dirichlet allocation. We assume that some number of "topics," which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.



- Probabilistic generative model for documents
- Each document has a fixed distribution over topics, each topic is has a fixed distribution over words belonging to it
- Unlike clustering, groups are non-exclusive

Hidden Markov Model



- Speech data is a stream of data flowing in
- Only makes sense to consider entire stream not each bit alone
- Hidden markov models, capture our belief that we produce sound based on phoneme we think of
- Phonemes in right sequence model what we want to say

What we won't cover

- Feature extraction is a problem/domain specific art, we won't cover this in class
- We won't cover optimization methods for machine learning
- Implementation tricks and details won't be covered
- There are literally thousands of methods, we will only cover a few!

What you can take home

- How to think about a learning problem and formulate it
- Well known methods and how and why they work
- Hopefully we can give you an intuition on choice of methods/approach to try out on a given problem

Dimensionality Reduction

Given data $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ compress the data points in to low dimensional representation $\mathbf{y}_1, \ldots, \mathbf{y}_n \in \mathbb{R}^K$ where $K << d$

Why dimensionality reduction?

- For computational ease
  - As input to supervised learning algorithm
  - Before clustering to remove redundant information and noise
- Data visualization
- Data compression
- Noise reduction

Dimensionality Reduction

Desired properties:

1. Original data can be (approximately) reconstructed
2. Preserve distances between data points
3. "Relevant" information is preserved
4. Redundant information is removed
5. Models our prior knowledge about real world

Based on the choice of desired property and formalism we get different methods

Sneak Peek

- Linear projections
- Principle component analysis