Machine Learning for Data Science (CS4786)
Lecture 1

Tu-Th 10:10 to 11:25 AM
Hollister B14

Instructors : Lillian Lee and Karthik Sridharan

Rough Details About the Course

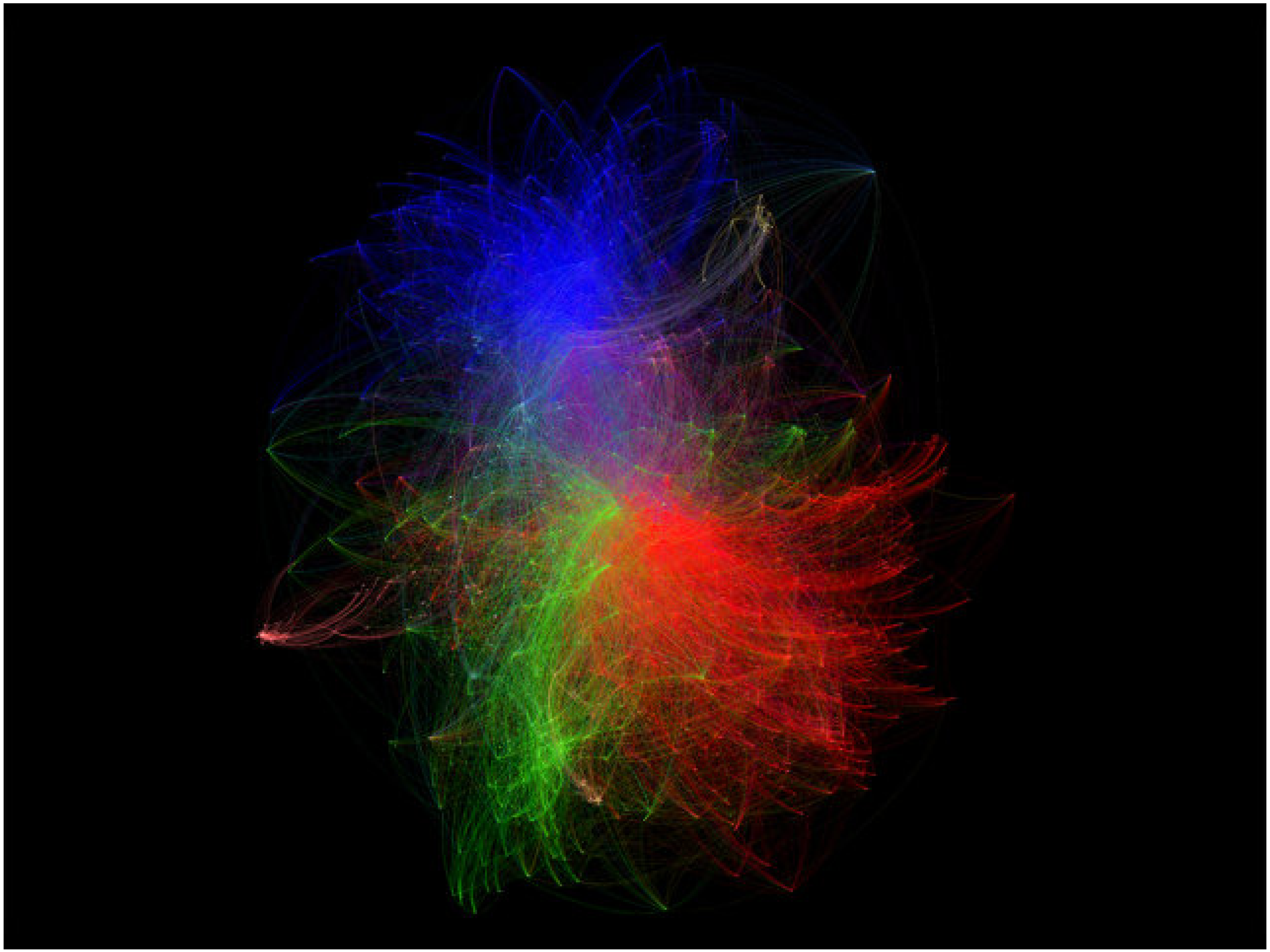
- Diagnostic assignment 0 is out: for our calibration.
  Hand in your assignments beginning of class on 29th Jan.
- We are thinking roughly three assignments
- (Approximately) 2 competition/challenges,
  - Clustering/data visualization challenge
  - Prediction challenge with focus on feature extraction/selection

Lets get started . . .

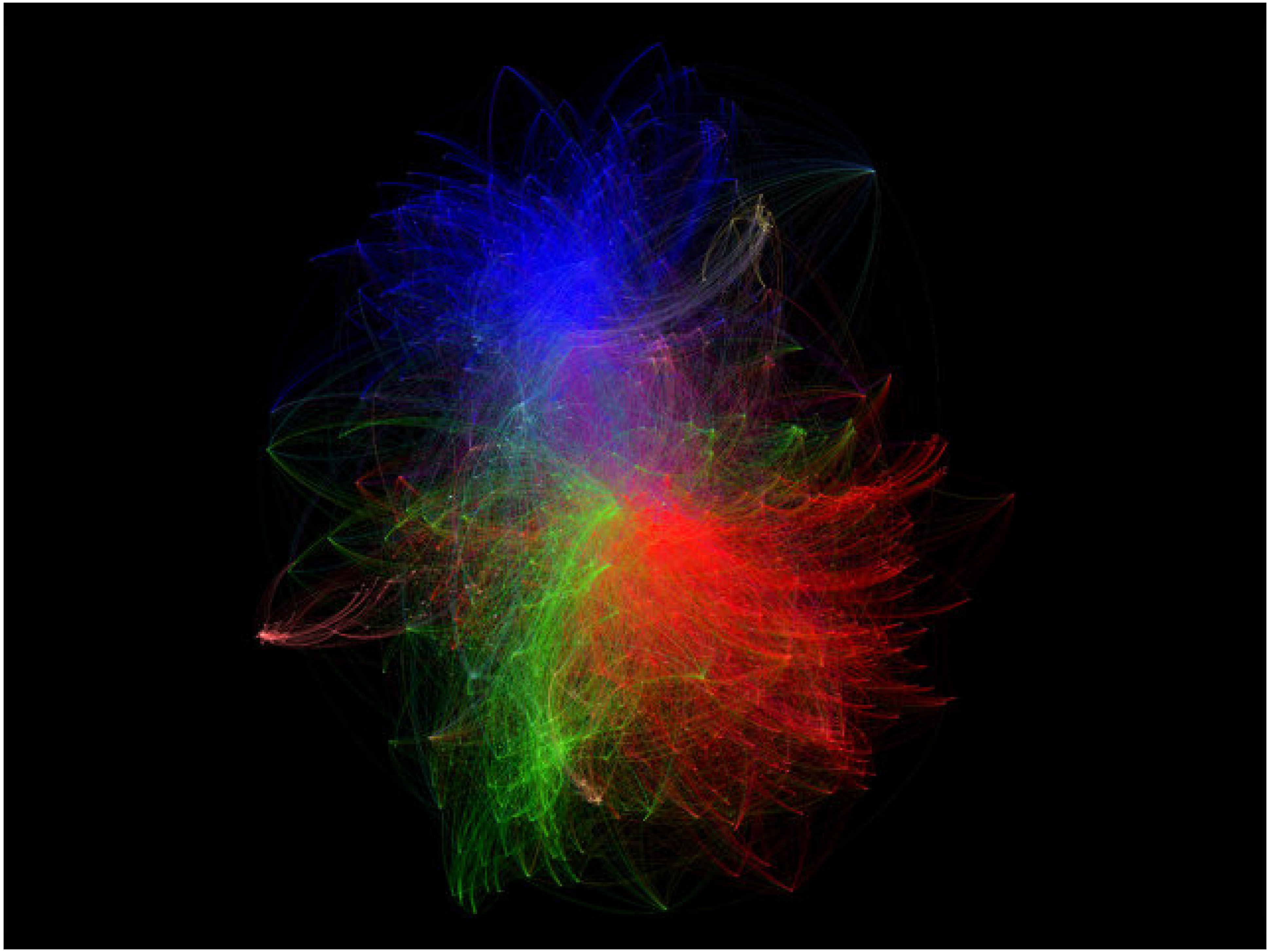
DATA DELUGE

- 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?

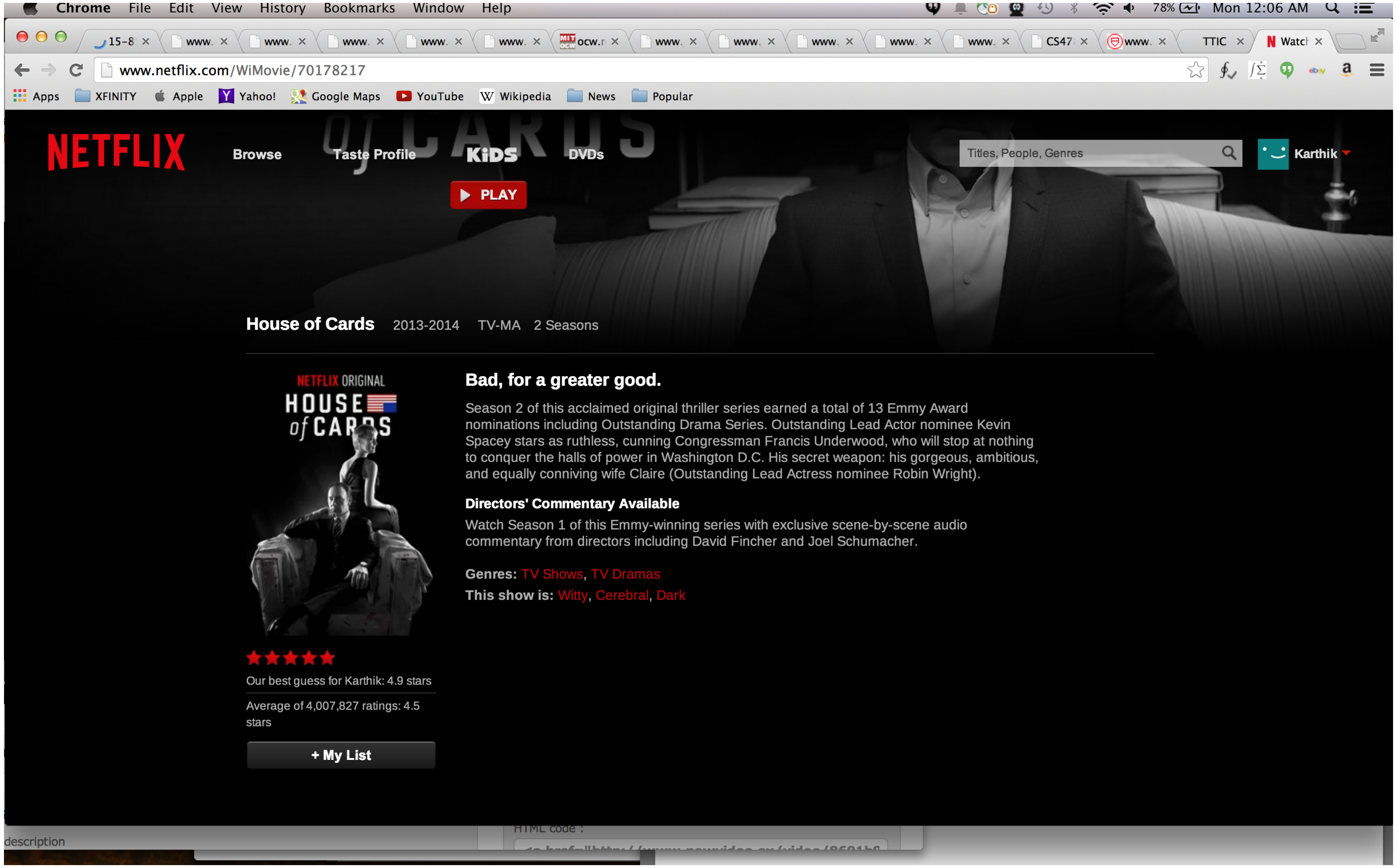
WHAT IS MACHINE LEARNING?

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 ?

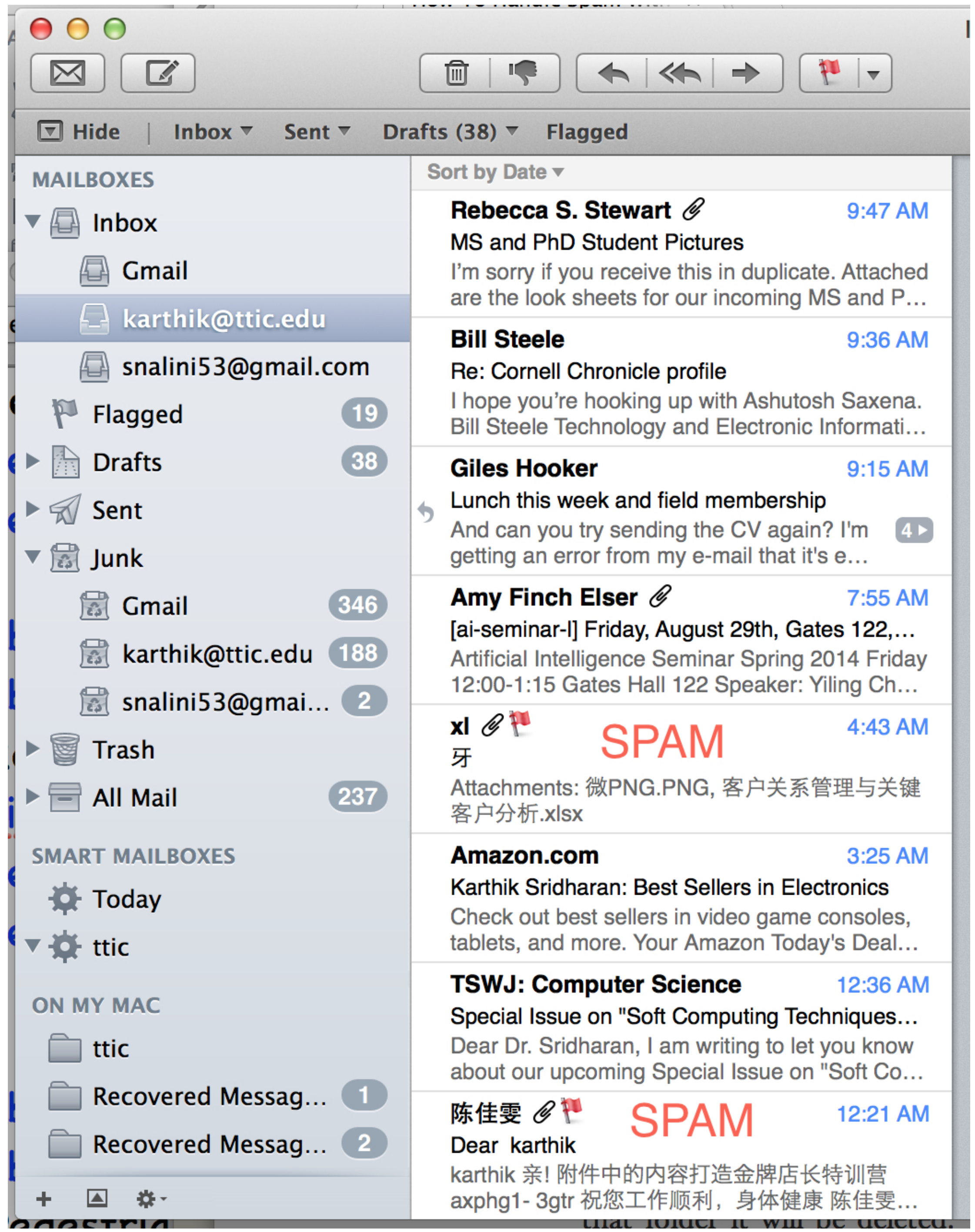
Pedestrian Detection

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 hope to cover

1. Dimensionality Reduction:

2. Clustering and Mixture models:

3. Probabilistic Modeling & Graphical Models:

4. *Some supervised learning:* (if time permits)

Topics we hope to cover

1. Dimensionality Reduction:
   Principal Component Analysis (PCA), Canonical Component Analysis (CCA), Random projections, Compressed Sensing (CS), Independent Component Analysis (ICA), Information-Bottleneck, Linear Discriminant Analysis
2. Clustering and Mixture models:
3. Probabilistic Modeling & Graphical Models:
4. *Some supervised learning:* (if time permits)

Topics we hope to cover

1. Dimensionality Reduction:
   Principal Component Analysis (PCA), Canonical Component Analysis (CCA), Random projections, Compressed Sensing (CS), Independent Component Analysis (ICA), Information-Bottleneck, Linear Discriminant Analysis
2. Clustering and Mixture models:
   k-means clustering, gaussian mixture models, hierarchical clustering, link based clustering
3. Probabilistic Modeling & Graphical Models:
4. *Some supervised learning:* (if time permits)

Topics we hope to cover

1. Dimensionality Reduction:
   Principal Component Analysis (PCA), Canonical Component Analysis (CCA), Random projections, Compressed Sensing (CS), Independent Component Analysis (ICA), Information-Bottleneck, Linear Discriminant Analysis
2. Clustering and Mixture models:
   k-means clustering, gaussian mixture models, hierarchical clustering, link based clustering
3. Probabilistic Modeling & Graphical Models:
   Probabilistic modeling, MLE Vs MAP Vs Bayesian approaches, inference and learning in graphical models, Latent Dirichlet Allocation (LDA)
4. *Some supervised learning:* (if time permits)

Topics we hope to cover

1. Dimensionality Reduction:
   Principal Component Analysis (PCA), Canonical Component Analysis (CCA), Random projections, Compressed Sensing (CS), Independent Component Analysis (ICA), Information-Bottleneck, Linear Discriminant Analysis
2. Clustering and Mixture models:
   k-means clustering, gaussian mixture models, hierarchical clustering, link based clustering
3. Probabilistic Modeling & Graphical Models:
   Probabilistic modeling, MLE Vs MAP Vs Bayesian approaches, inference and learning in graphical models, Latent Dirichlet Allocation (LDA)
4. *Some supervised learning:* (if time permits)
   linear regression, logistic regression, Lasso, ridge regression, neural networks/deep learning, . . .

TOPICS WE HOPE TO COVER

unsupervised learning

1. Dimensionality Reduction:
   Principal Component Analysis (PCA), Canonical Component Analysis (CCA), Random projections, Compressed Sensing (CS), Independent Component Analysis (ICA), Information-Bottleneck, Linear Discriminant Analysis
2. Clustering and Mixture models:
   k-means clustering, gaussian mixture models, hierarchical clustering, link based clustering
3. Probabilistic Modeling & Graphical Models:
   Probabilistic modeling, MLE Vs MAP Vs Bayesian approaches, inference and learning in graphical models, Latent Dirichlet Allocation (LDA)
4. *Some supervised learning:* (if time permits)
   linear regression, logistic regression, Lasso, ridge regression, neural networks/deep learning, . . .

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

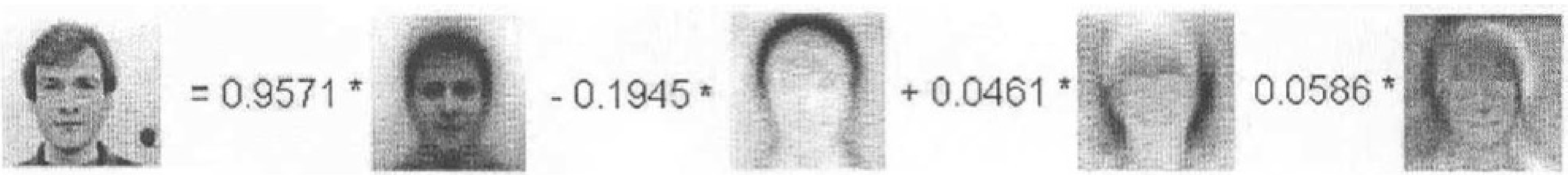
Dimensionality Reduction

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

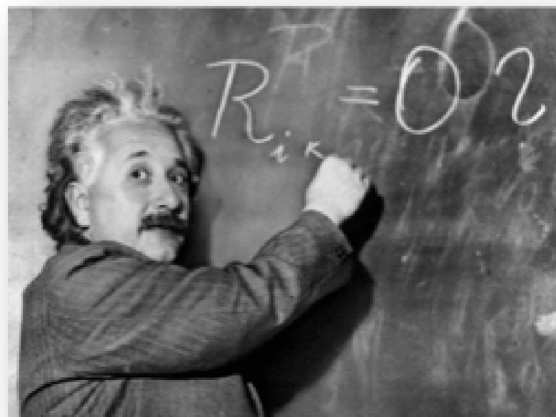


- Write down each data point as a linear combination of small number of basis vectors
- Data specific compression scheme
- One of the early successes: in face recognition: classification based on nearest neighbor in the reduced dimension space

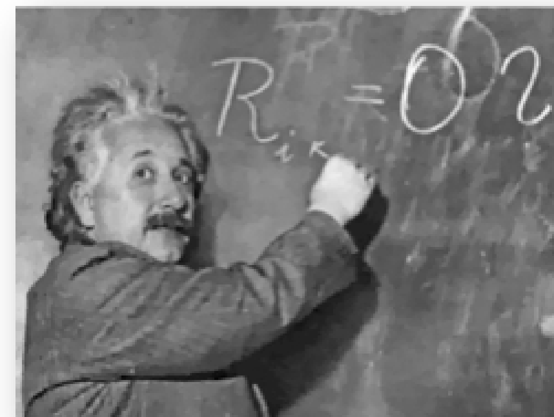
Compressed Sensing

Candes, Tao, Donaho $\approx'$ 04

From Compressive Sensing Camera



Original Target



InView SWIR Reproduction

- Can we compress directly while receiving the input?
- We now have cameras that directly sense/record compressed information …and very fast!
- Time spent only for reconstructing the compressed information
- Especially useful for capturing high resolution MRI's

Independent Component Analysis (ICA)

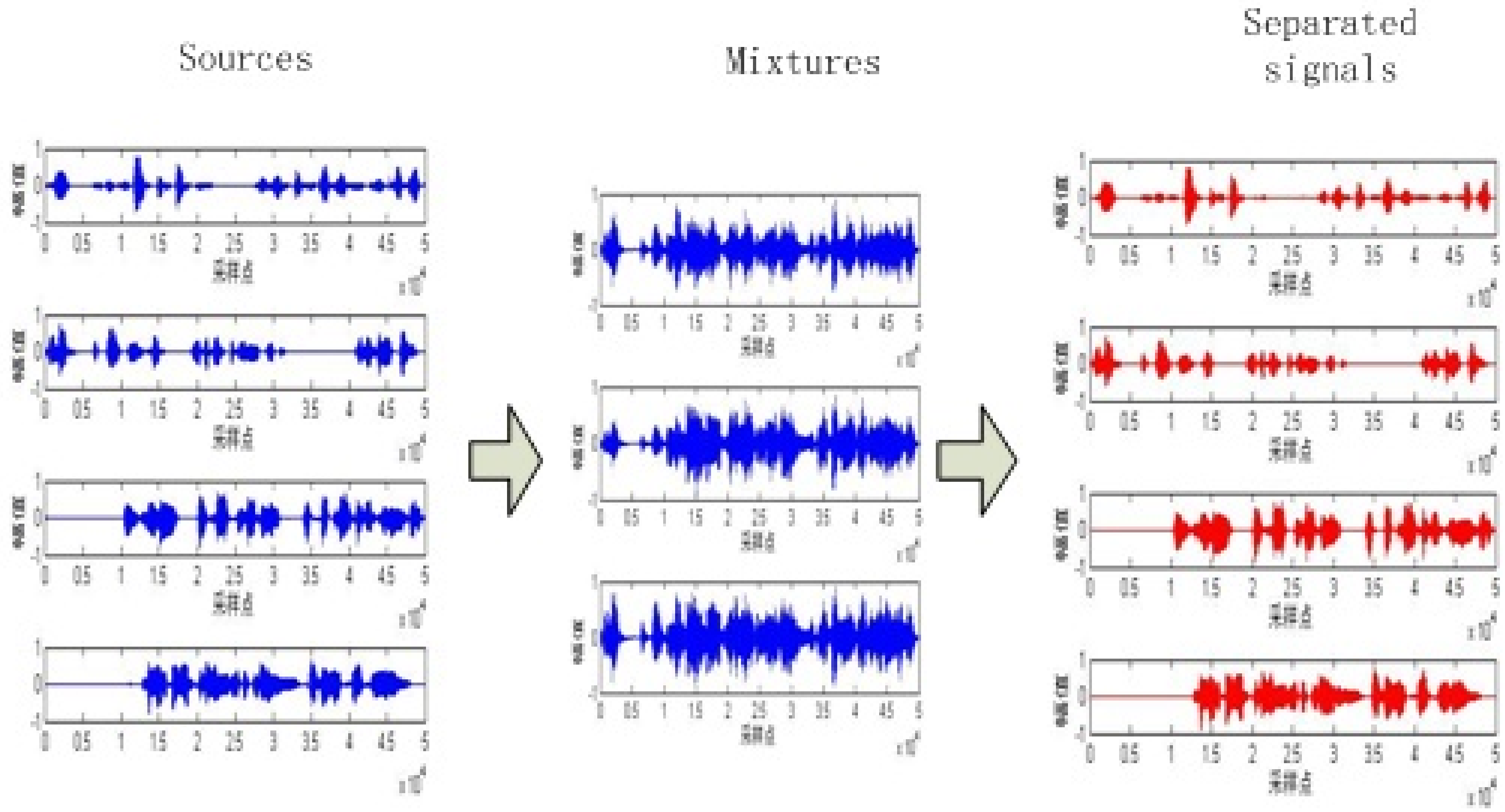
Cocktail Party

- You are at a cocktail party, people are speaking all around you
- But you are still able to follow conversation with your group?
- Can a computer do this automatically?

Independent Component Analysis (ICA)

Bell & Sejnowski '95

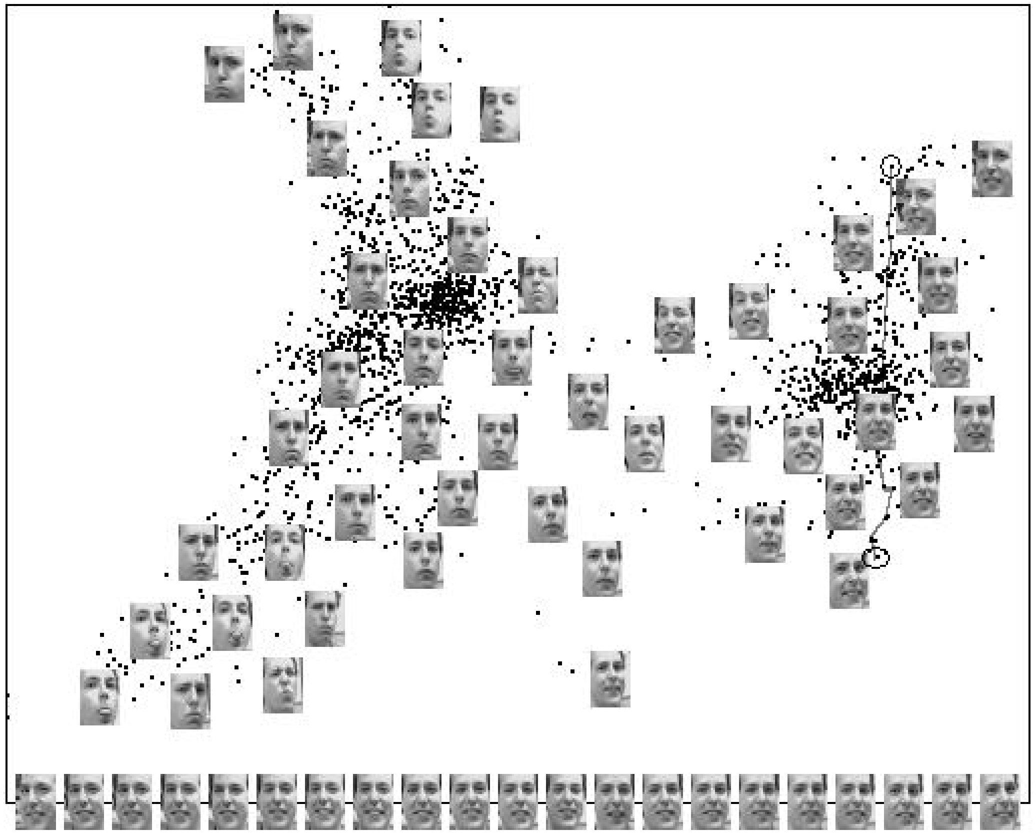
Blind Source Seperation



- Can do this as long as the sources are independent
- Represent data points as linear (or non-linear) combination of independent sources

Data Visualization

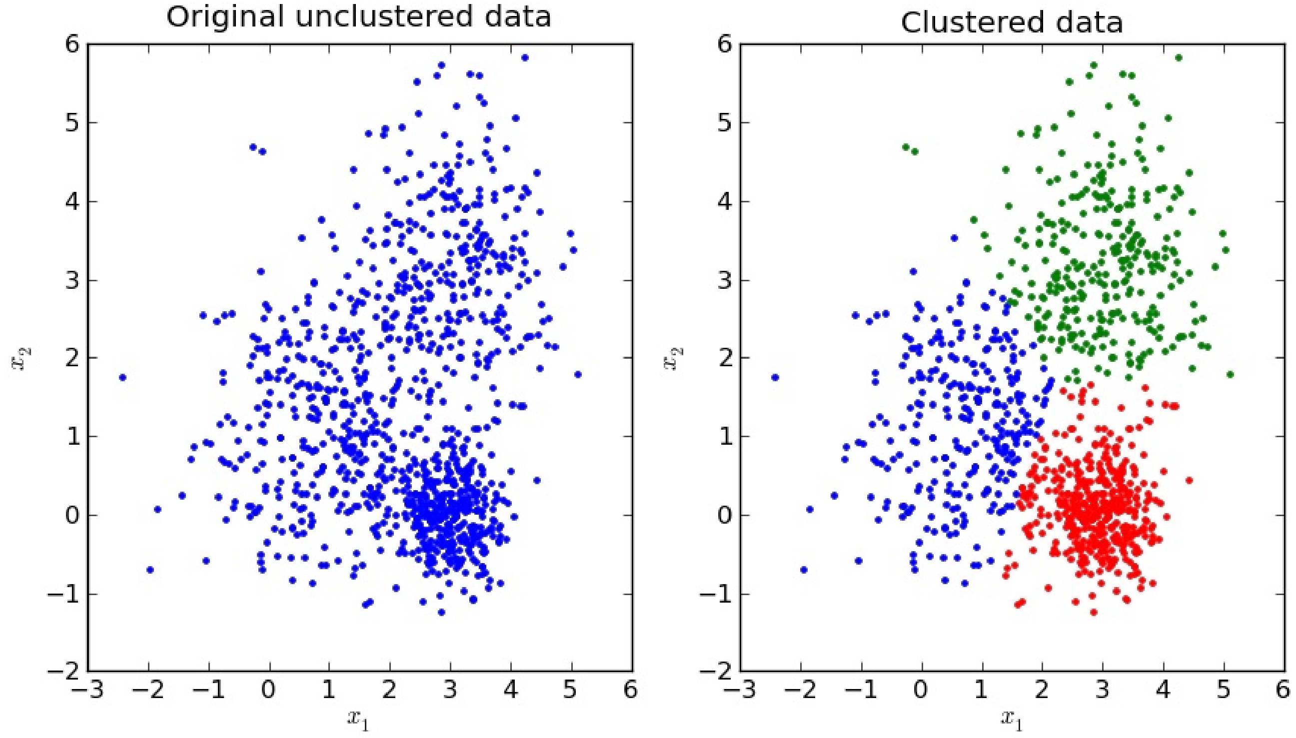
2D projection



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

CLUSTERING

K-means clustering

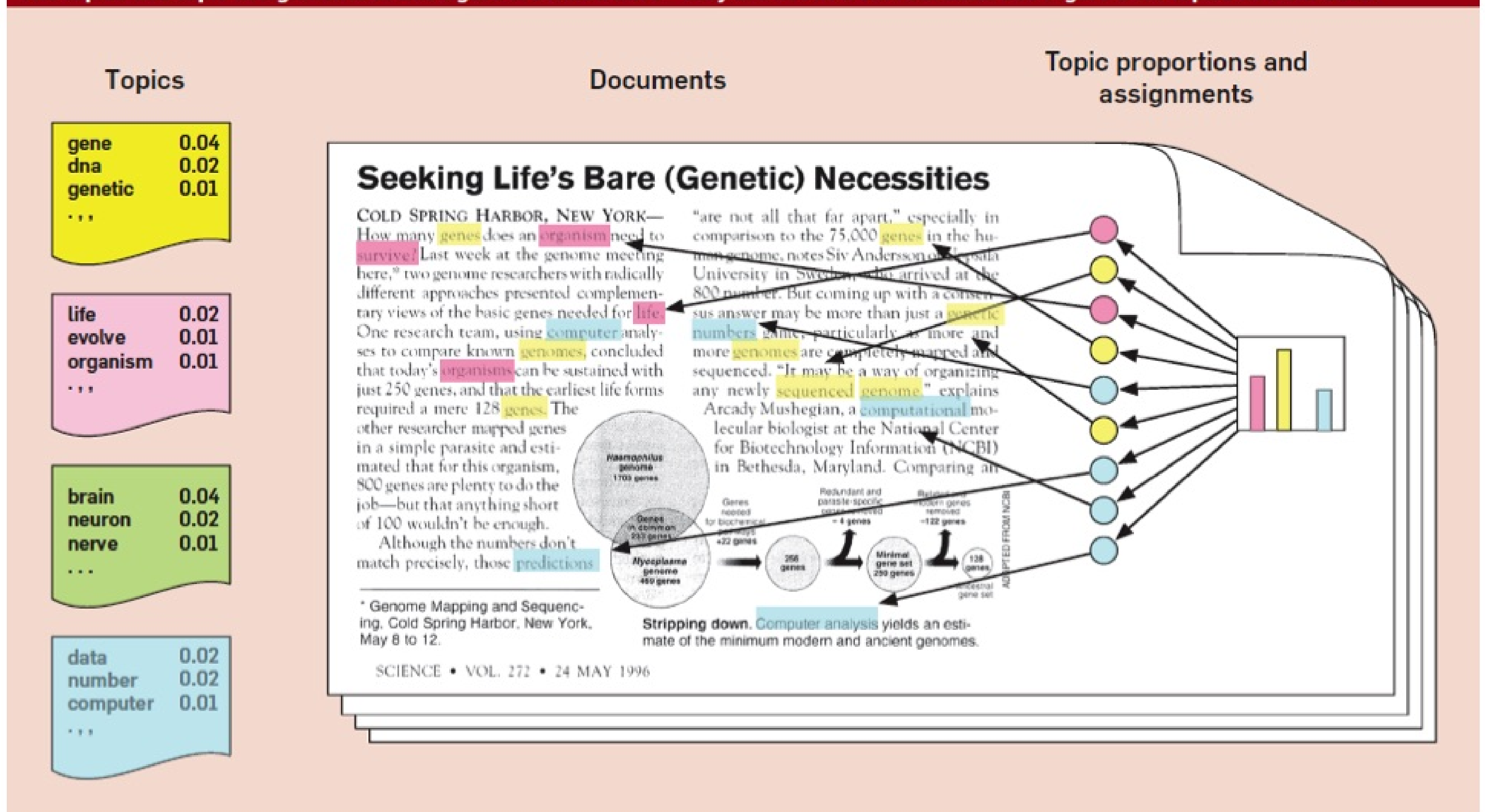


- Given just the data points group them in natural clusters
- Roughly speaking
  - Points within a cluster must be close to each other
  - Points between clusters must be separated
- Helps bin data points, but generally hard to do

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

Supervised Learning

PHOTO: Handout, Subaru

- Training data comes as input output pairs (x, y)
- Based on this data we learn a mapping from input to output space
- Goal: Given new input instance x, predict outcome y accurately based on given training data
- Classification, regression

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