

Machine Learning for Data Science (CS4786)

Lecture 1

Tu-Th 11:40AM to 12:55 PM
Klarman Hall KG70

Instructor : Karthik Sridharan

THE AWESOME TA'S

① Caroline Chang

② Jason Chen

③ Cheng Preng Phoo

④ Pulkit Kashyap

⑤ Ayush Sekhari

① Amol Tandel

② Dilip Thiagarajan

③ Lequn Wang

④ Qiantong Xu

⑤ Zhao Zhan

COURSE INFORMATION

- Course webpage is the official source of information:
<http://www.cs.cornell.edu/Courses/cs4786/2017fa>
- Join Piazza: <https://piazza.com/class/j6i8gdcku9x7i8>
- 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 Gaussian Mixture Models and Other Mixture Models
 - ④ Spectral Clustering
- ③ Probabilistic Modeling and Graphical Models:
 - ① MLE Vs MAP Vs Bayesian Methods
 - ② EM Algorithm
 - ③ Graphical Models
 - ① Hidden Markov Models
 - ② Latent Dirichlet Allocation
 - ③ Exact Inference: Variable Elimination, Belief Propagation
 - ④ Learning in Graphical Models
 - ⑤ Approximate Inference

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.**
 - Submit the assignment online via google form in link given by August 29th.
 - **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: August 31st Due: September 7th
 - ② Assignment 2: Out: September 12th Due: September 19th
 - ③ Assignment 3: Out: September 21st Due: September 28th
 - ④ Assignment 4: Out: September 28th Due: October 5th
 - ⑤ Assignment 5: Out: October 17th Due: October 24th
 - ⑥ Assignment 6: Out: November 2nd Due: November 14th

SURVEYS

- Two surveys + final course evaluation
- The surveys provide very valuable information about how course is progressing and helps me improve
- All surveys are **anonymous** and will be on Piazza
- If overall class participation is above 90% on all three I will drop worst assignment for everyone (evaluation on best 5 assignments)

COMPETITIONS

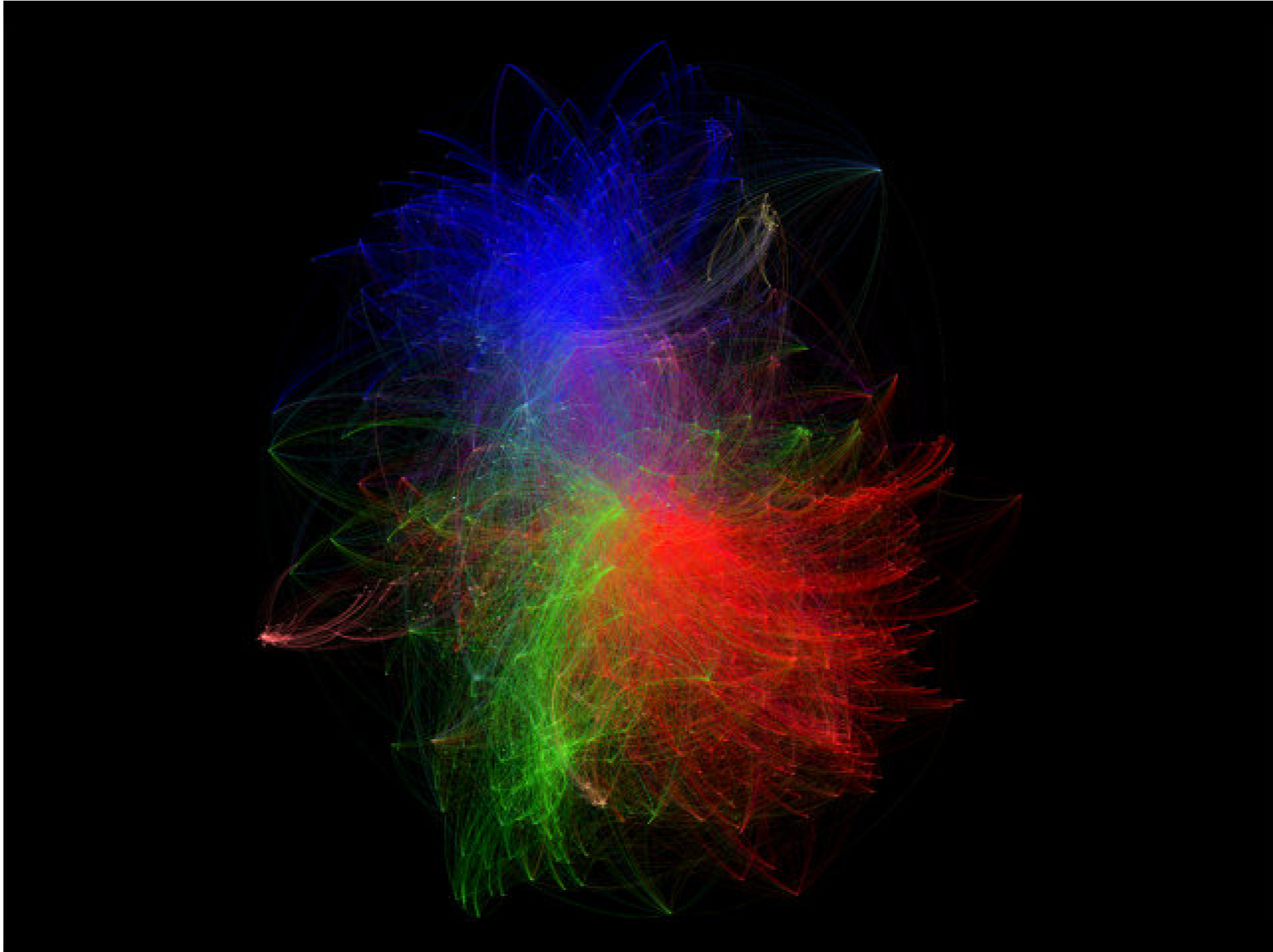
- 2 competition/challenges, worth 40% of total course grade
 - Competition I: Clustering challenge (Due Mid Oct)
 - Competition II: Graphical Model centric challenge (Due Nov end)
- Will be hosted on “In class Kaggle”!
- 40% of the competition grades for kaggle score/performance
- 60% of the competition 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 the competition grade.
- Groups of size at most 4.

Lets get started ...

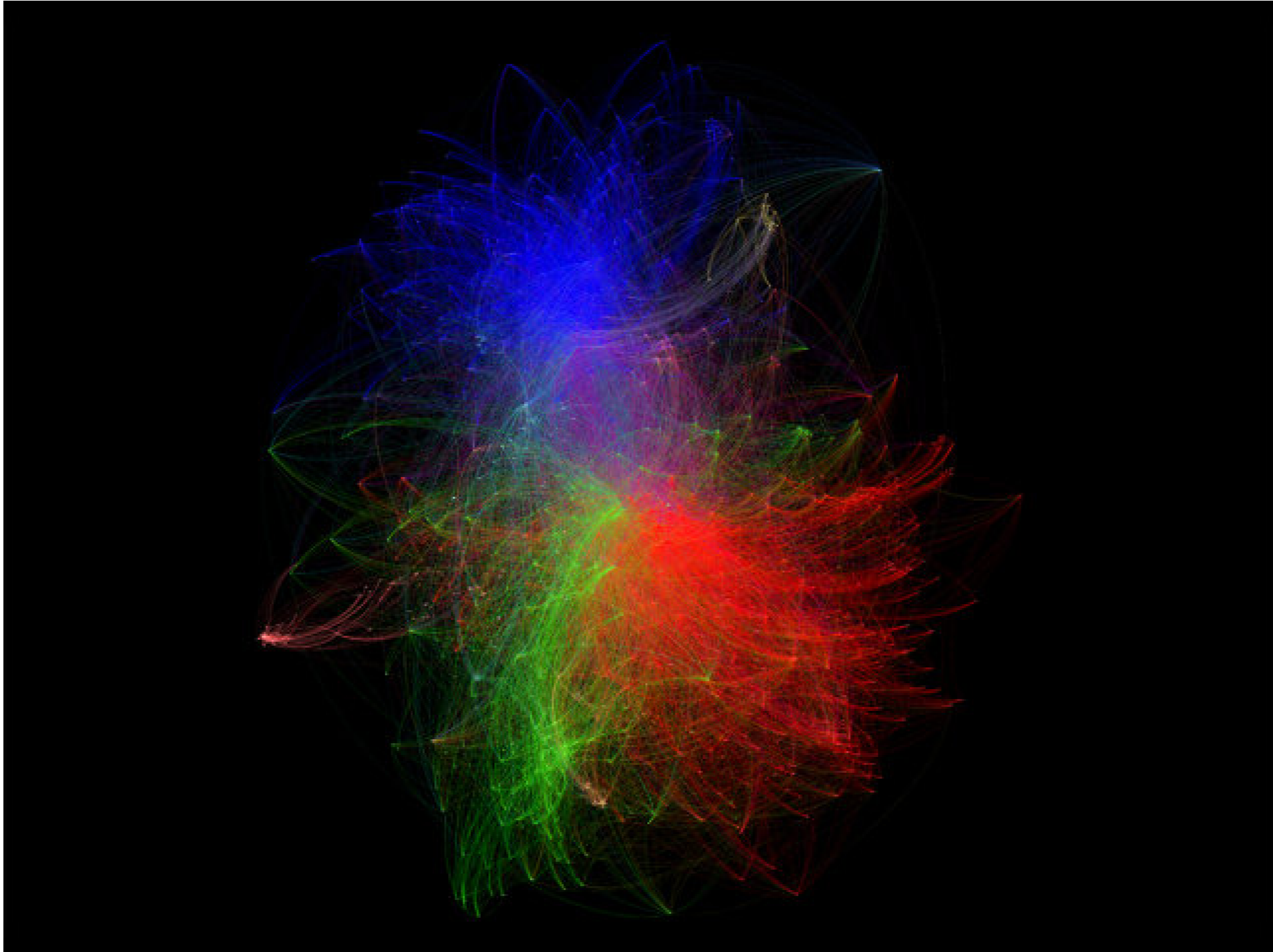
DATA DELUGE

- 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

The screenshot shows the Netflix interface for the TV show 'House of Cards'. At the top, there's a navigation bar with the Netflix logo, 'Browse', 'Taste Profile', 'KIDS', and 'DVDs'. A search bar is on the right with the text 'Titles, People, Genres' and a user profile 'Karthik'. Below the navigation is a large banner for 'House of Cards' with a 'PLAY' button. The main content area includes the show's title 'House of Cards', the years '2013-2014', the rating 'TV-MA', and '2 Seasons'. A description follows: 'Bad, for a greater good. Season 2 of this acclaimed original thriller series earned a total of 13 Emmy Award nominations including Outstanding Drama Series. Outstanding Lead Actor nominee Kevin Spacey stars as ruthless, cunning Congressman Francis Underwood, who will stop at nothing to conquer the halls of power in Washington D.C. His secret weapon: his gorgeous, ambitious, and equally conniving wife Claire (Outstanding Lead Actress nominee Robin Wright)'. Below the description, it says 'Directors' Commentary Available' and 'Watch Season 1 of this Emmy-winning series with exclusive scene-by-scene audio commentary from directors including David Fincher and Joel Schumacher.' The genres are listed as 'TV Shows, TV Dramas' and 'This show is: Witty, Cerebral, Dark'. At the bottom, there are five red stars, a rating of 'Our best guess for Karthik: 4.9 stars', and 'Average of 4,007,827 ratings: 4.5 stars'. A '+ My List' button is also visible.

WHERE IS IT USED ?

Pedestrian Detection



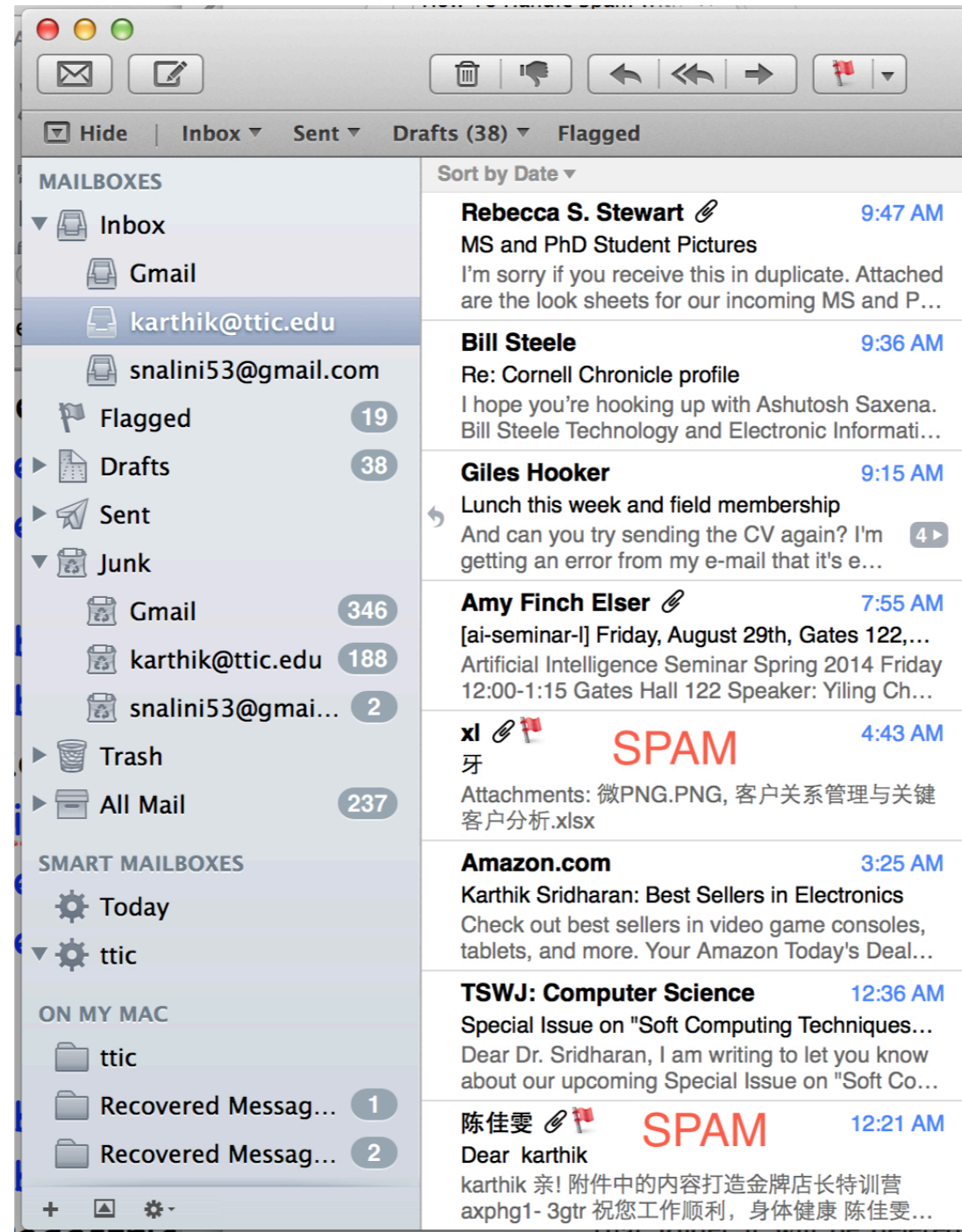
WHERE IS IT USED ?

Market Predictions



WHERE IS IT USED ?

Spam Classification



MORE APPLICATIONS

MORE APPLICATIONS

- Each time you use your search engine

MORE APPLICATIONS

- Each time you use your search engine
- Autocomplete: Blame machine learning for bad spellings

MORE APPLICATIONS

- Each time you use your search engine
- Autocomplete: Blame machine learning for bad spellings
- Biometrics: reason you shouldn't smile

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

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

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

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

TOPICS WE WILL COVER

- 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, ...

TOPICS WE WILL COVER

- 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), ...

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

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

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

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 $K \ll d$

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 $K \ll d$
 - 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:



PRINCIPAL COMPONENT ANALYSIS (PCA)

Turk & Pentland'91

Eigen Face:



- Write down each data point as a linear combination of small number of basis vectors

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

PRINCIPAL COMPONENT ANALYSIS (PCA)

Turk & Pentland'91

Eigen Face:

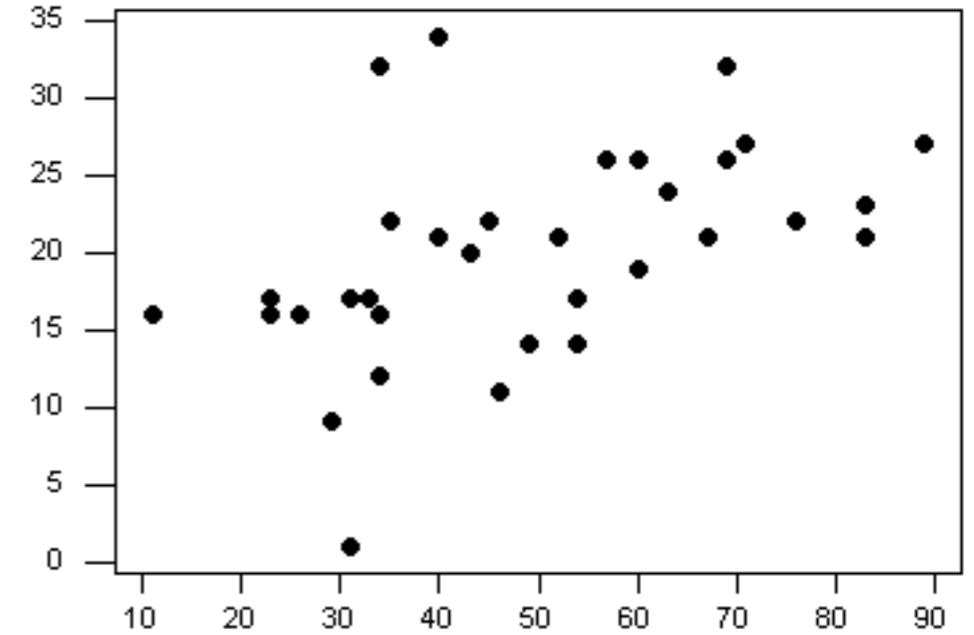
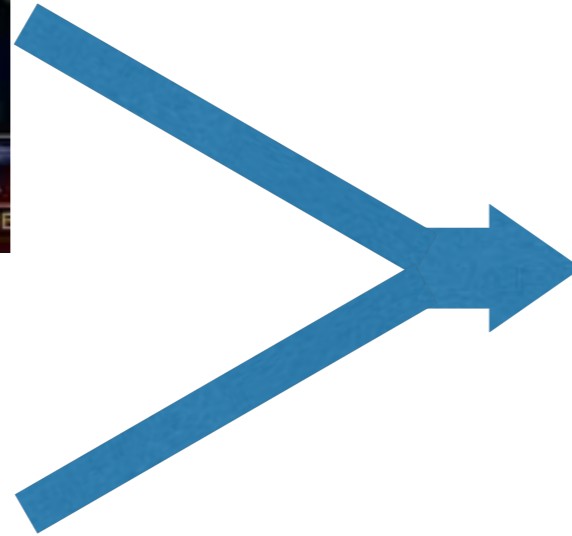


- Write down each data point as a linear combination of small number of basis vectors
- Data specific compression scheme
- One of the early successes: in face recognition: classification based on nearest neighbor in the reduced dimension space

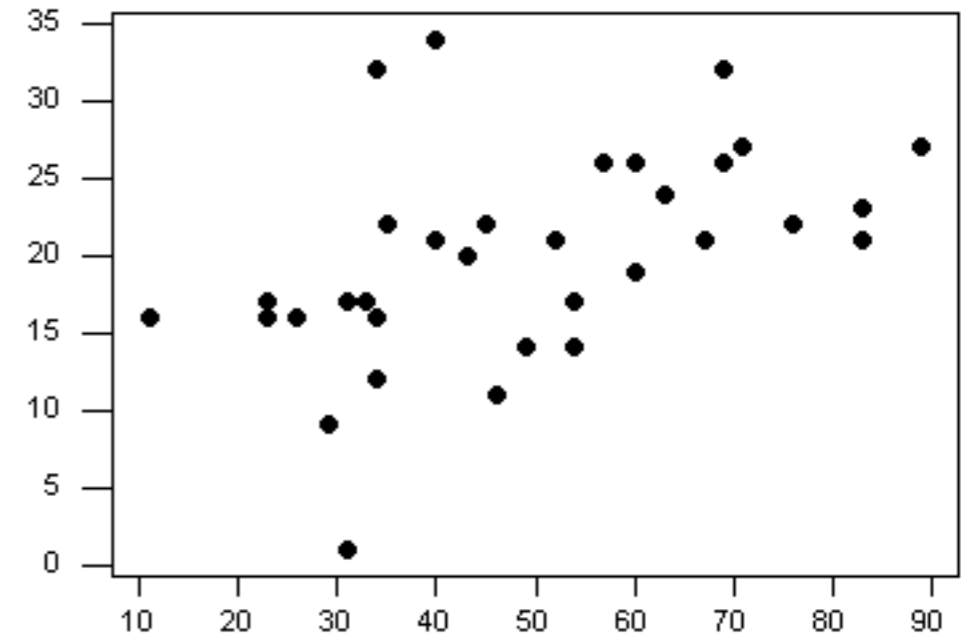
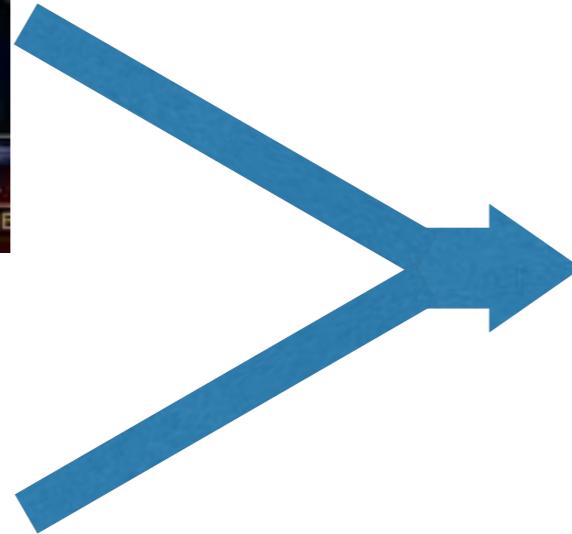
CANONICAL COMPONENT ANALYSIS (PCA)



CANONICAL COMPONENT ANALYSIS (PCA)

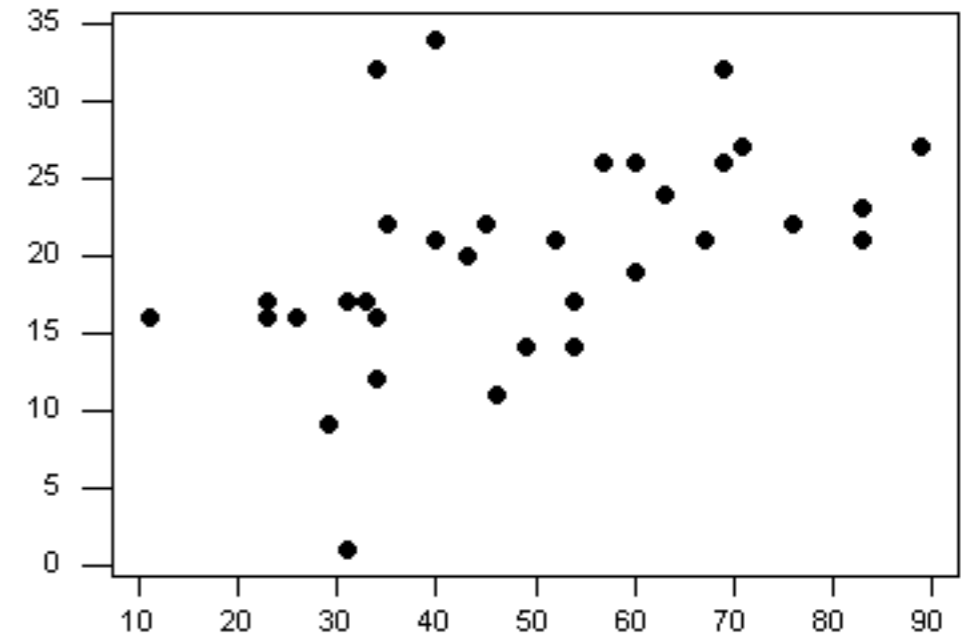
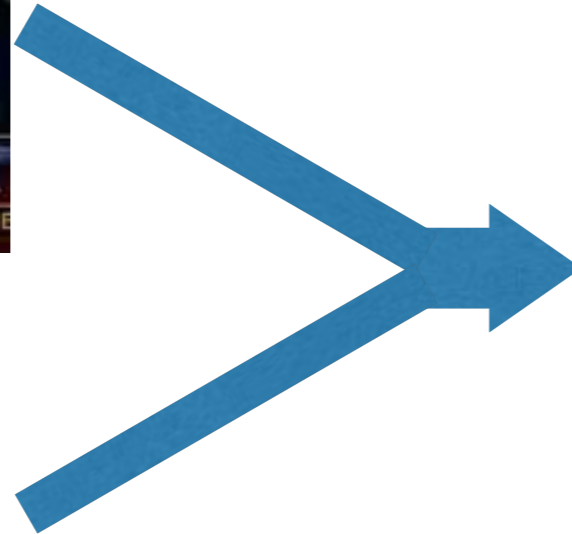


CANONICAL COMPONENT ANALYSIS (PCA)



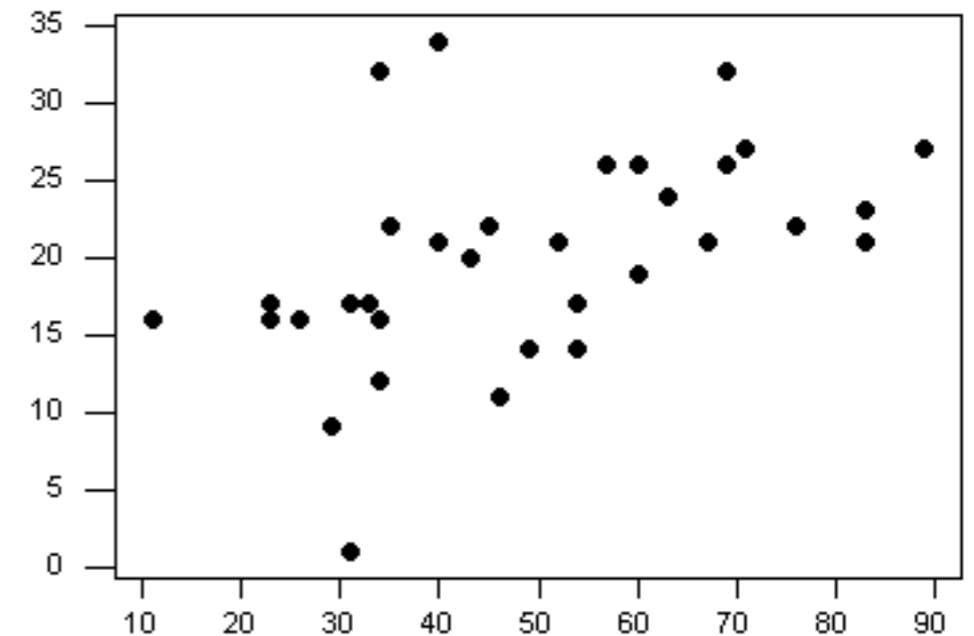
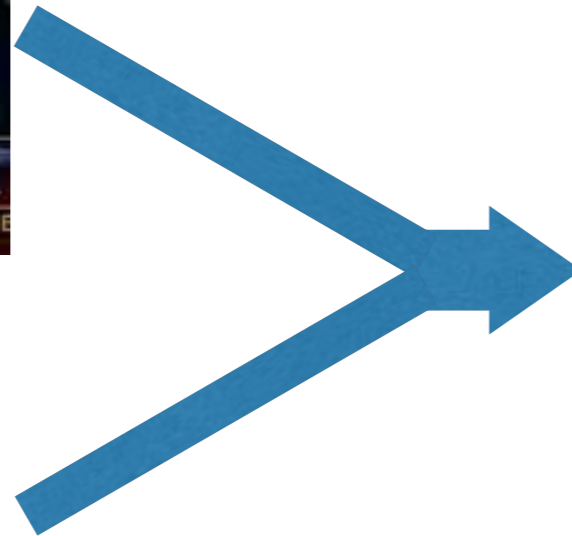
- Extract common information between multiple sources views

CANONICAL COMPONENT ANALYSIS (PCA)



- Extract common information between multiple sources views
- Noise specific to only one or subset of views is automatically filtered

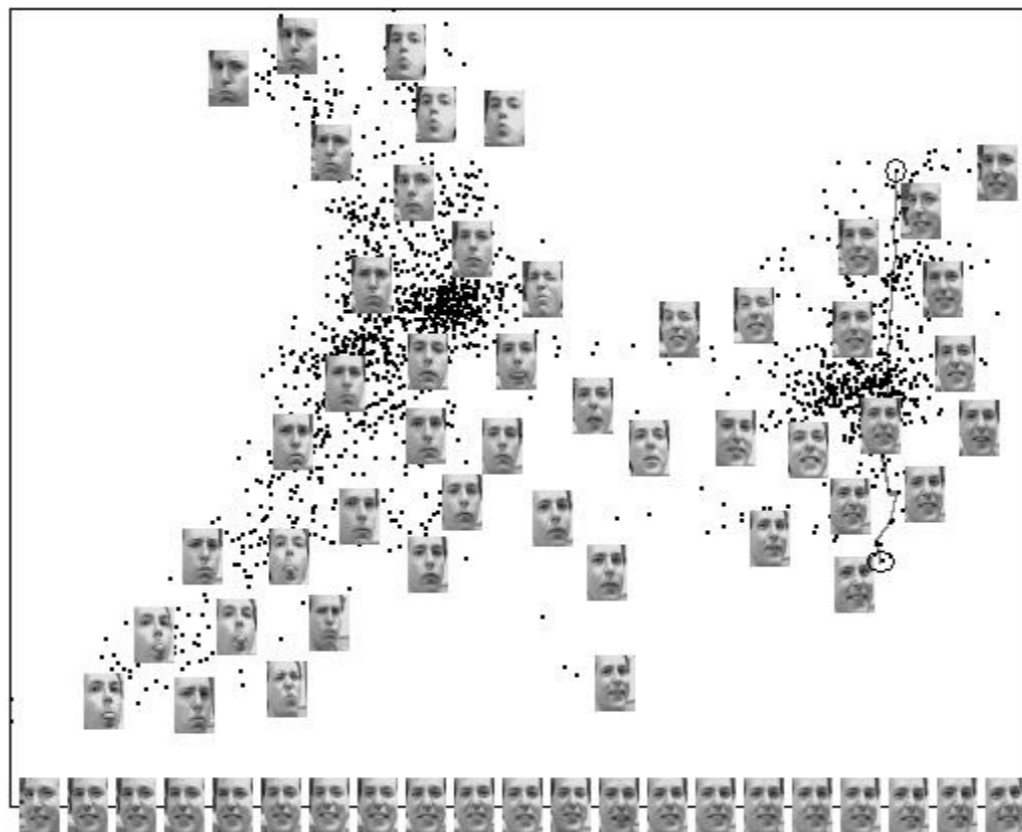
CANONICAL COMPONENT ANALYSIS (PCA)



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

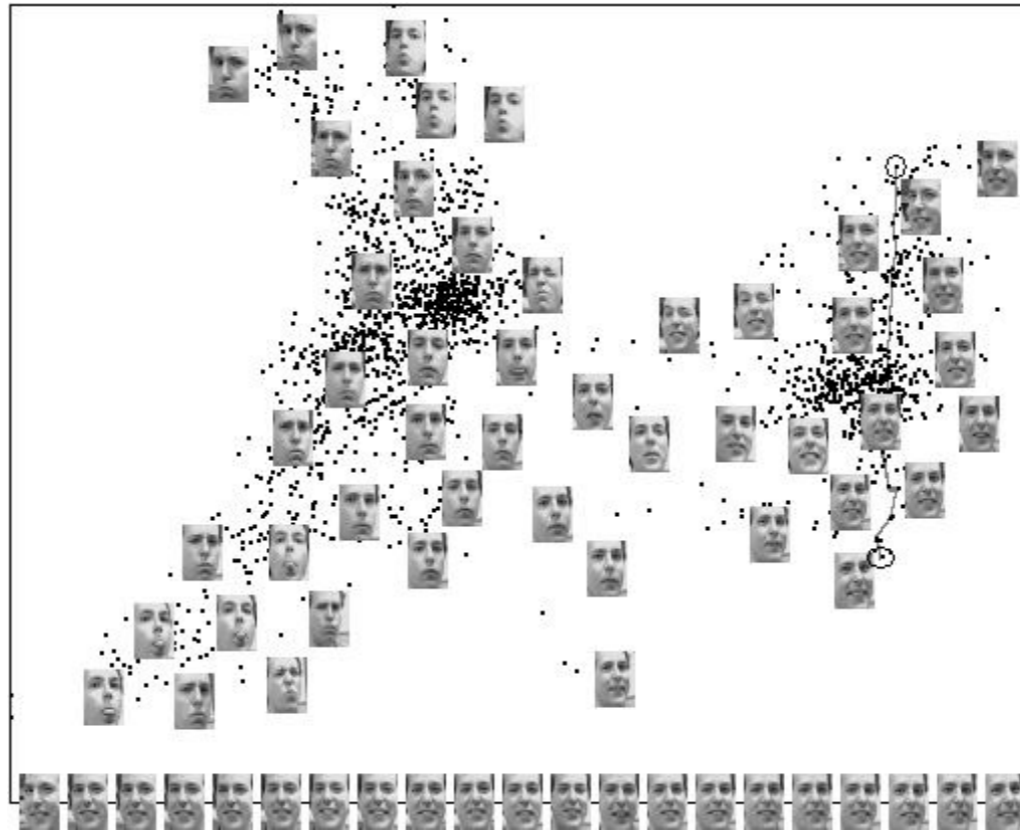
DATA VISUALIZATION

2D projection



DATA VISUALIZATION

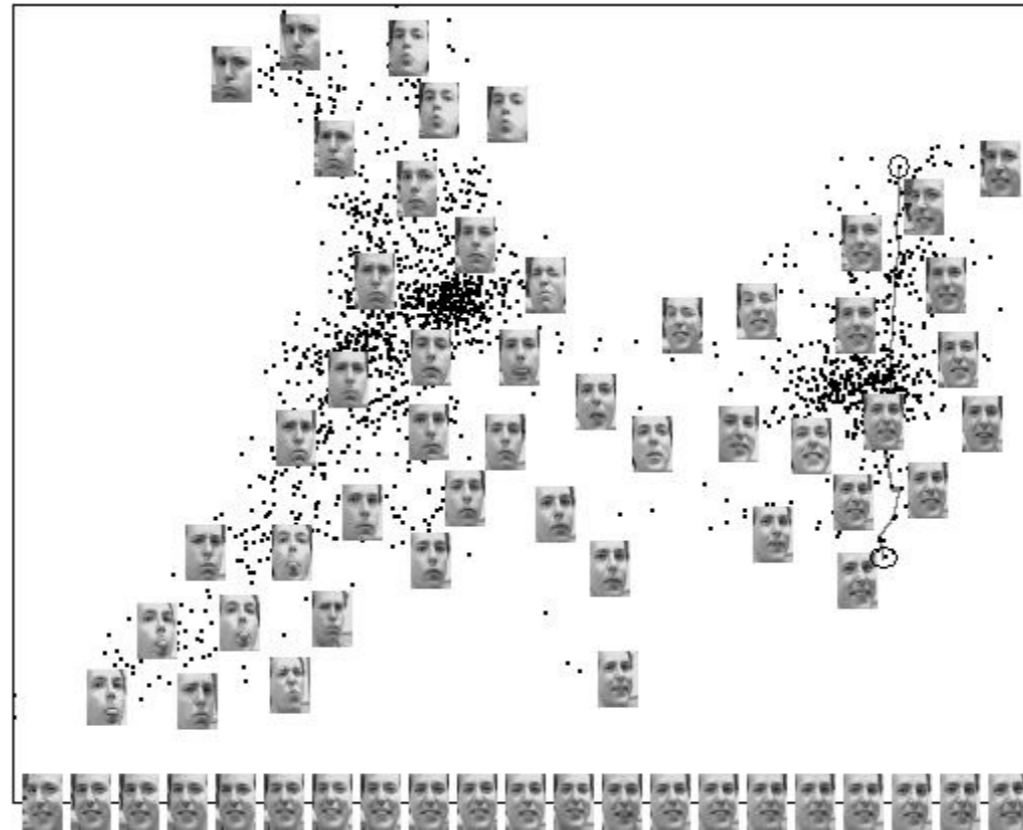
2D projection



- Help visualize data (in relation to each other)

DATA VISUALIZATION

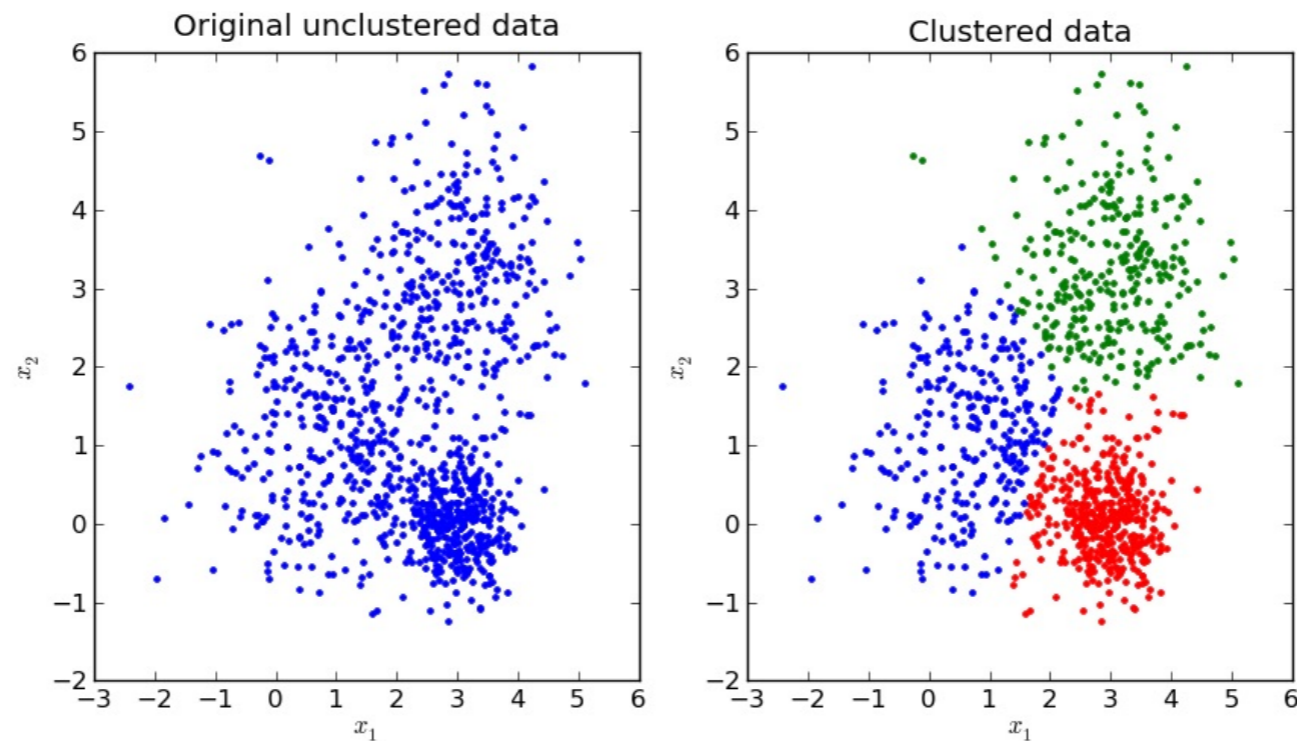
2D projection



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

CLUSTERING

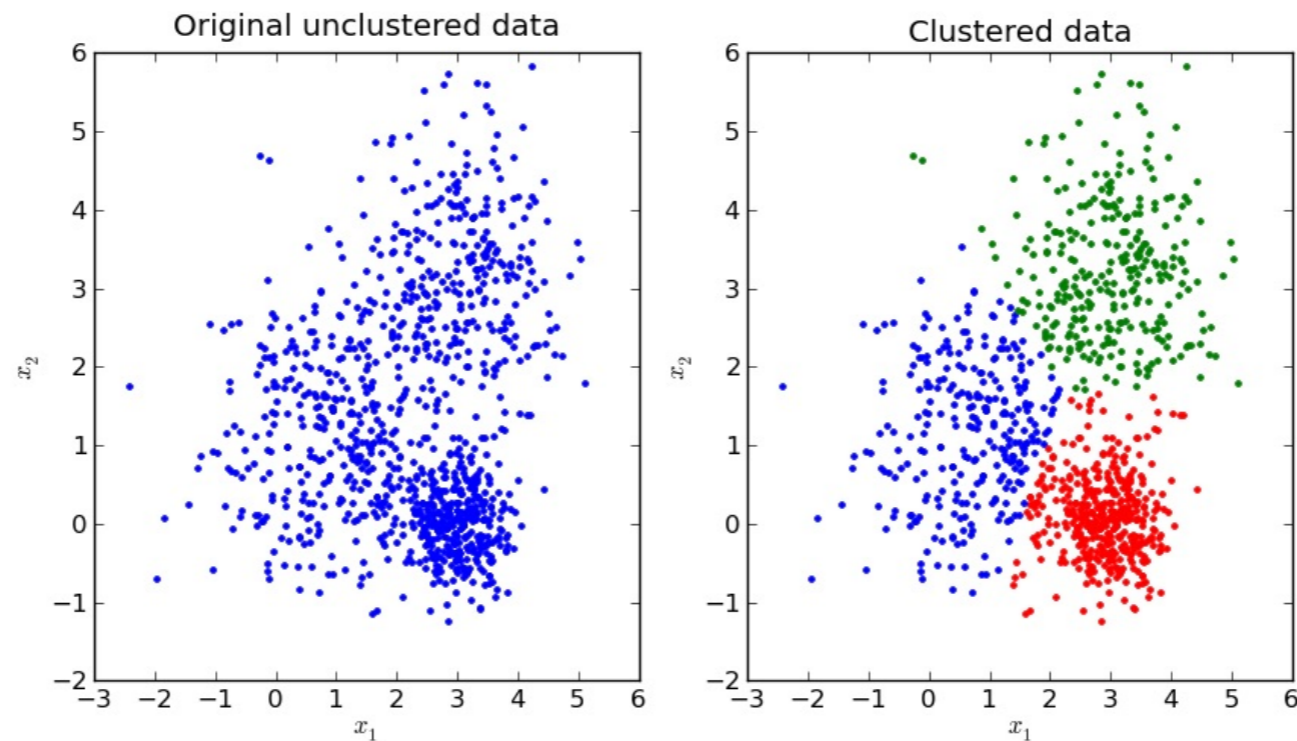
K-means clustering



- Given just the data points group them in natural clusters

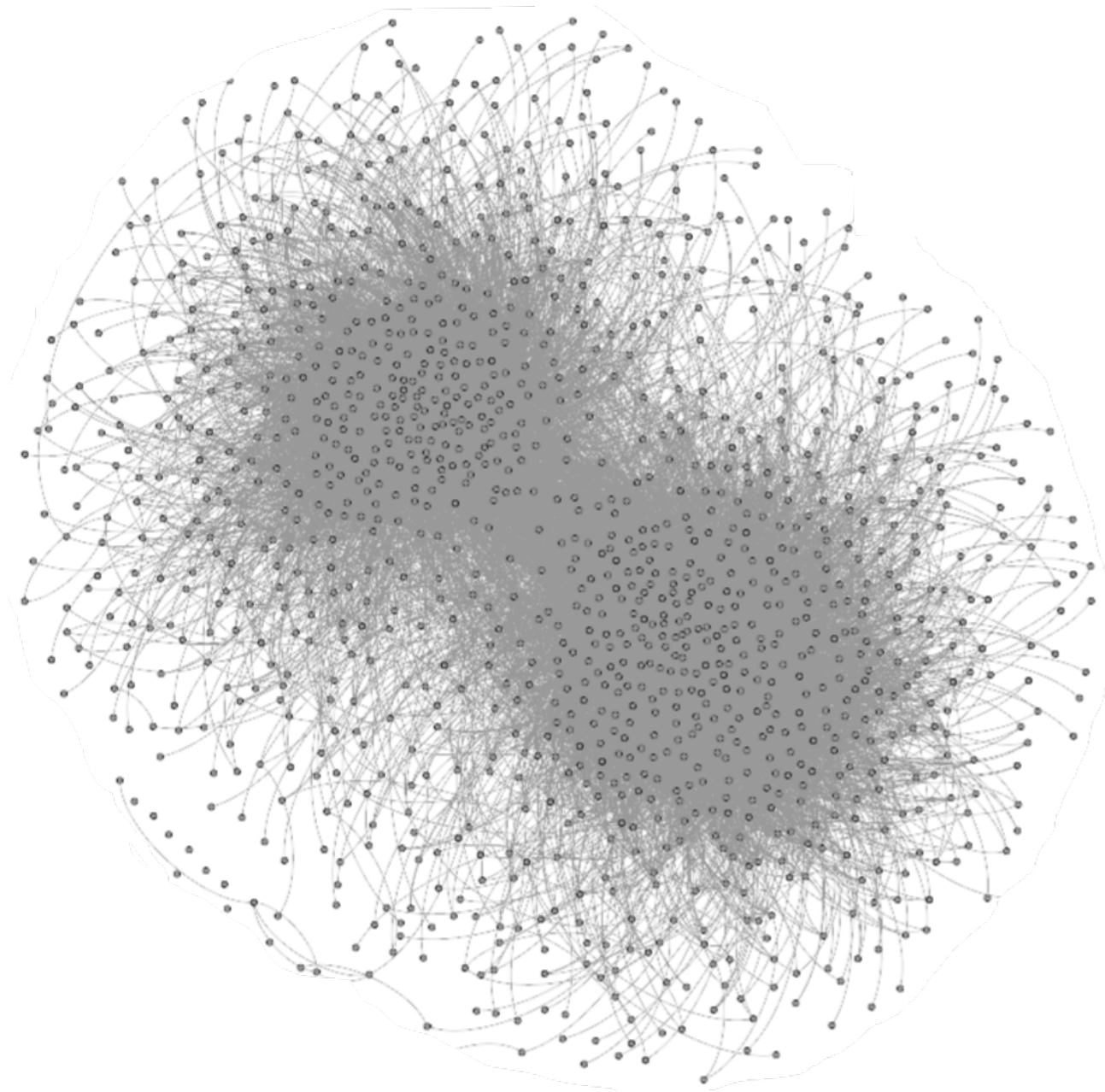
CLUSTERING

K-means clustering

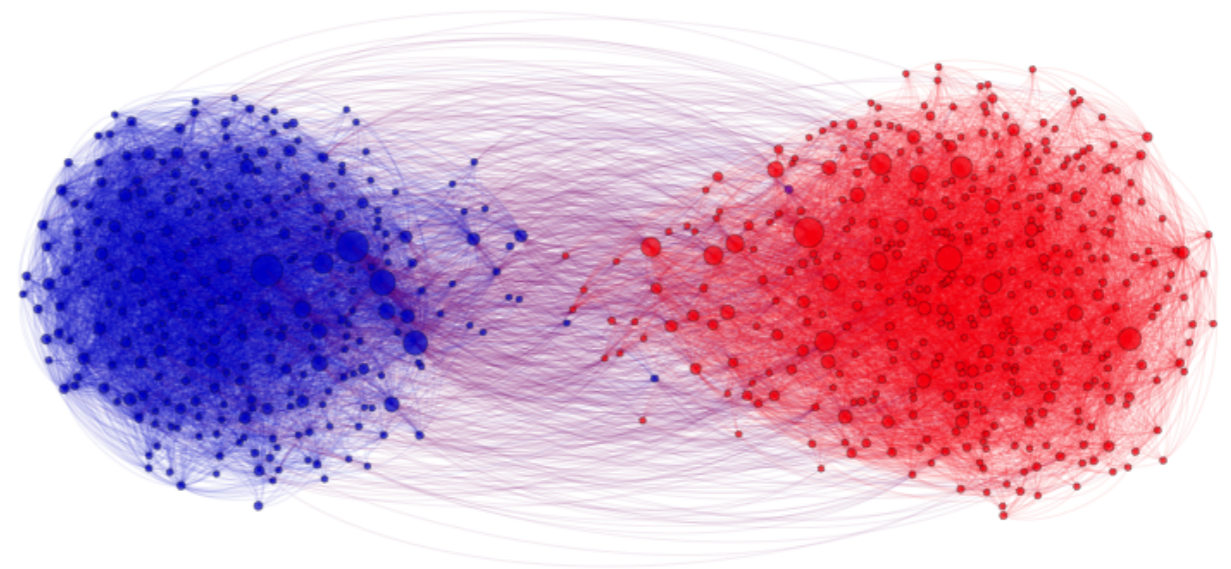
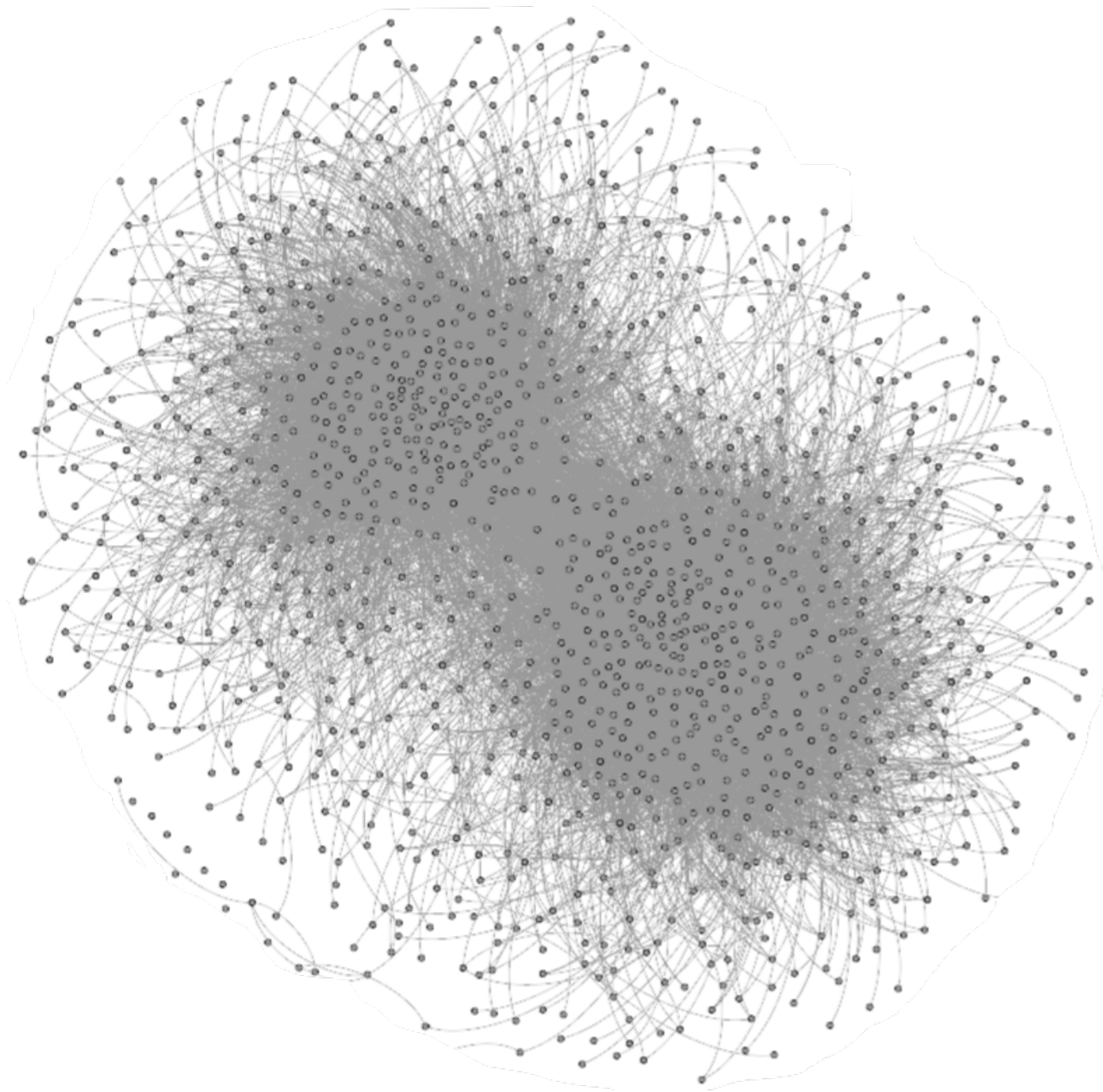


- Given just the data points group them in natural clusters
- Roughly speaking
 - Points within a cluster must be close to each other
 - Points between clusters must be separated
- Helps bin data points, but generally hard to do

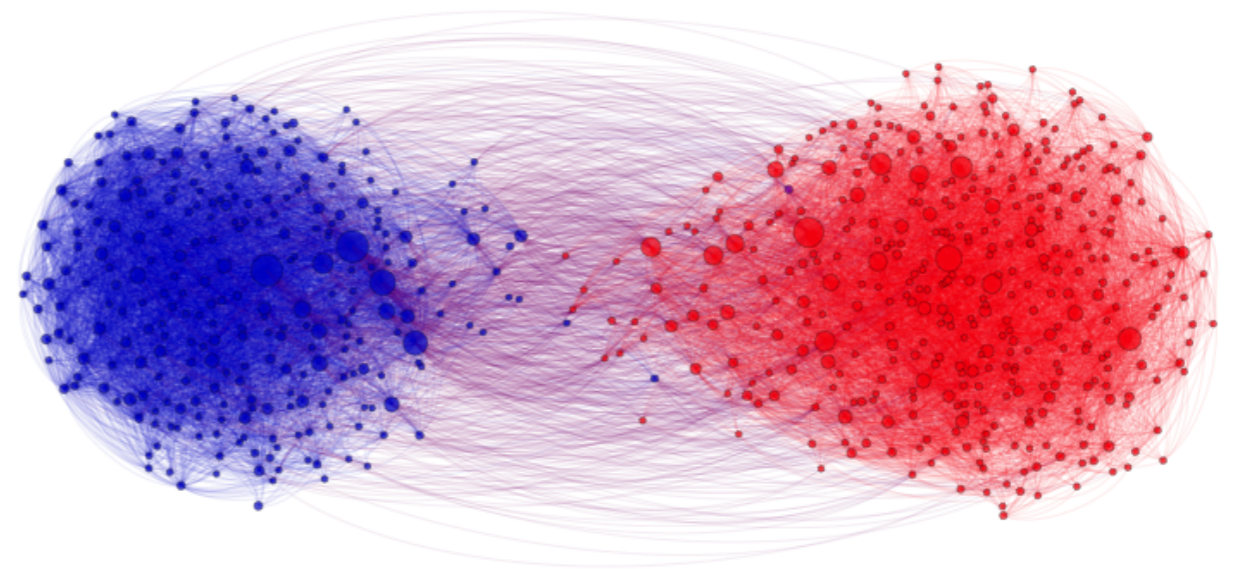
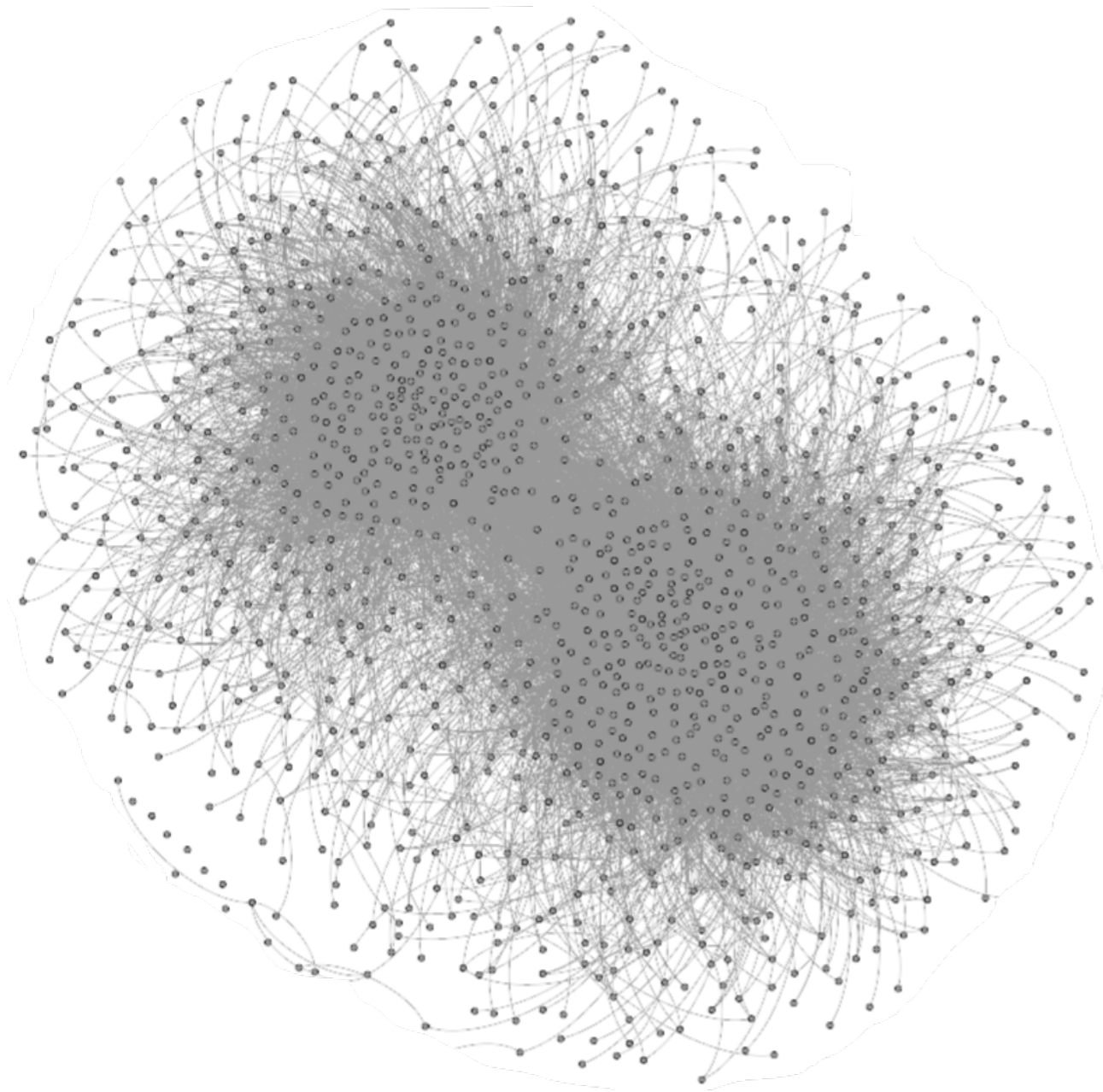
TELL ME WHO YOUR FRIENDS ARE . . .



TELL ME WHO YOUR FRIENDS ARE . . .



TELL ME WHO YOUR FRIENDS ARE . . .

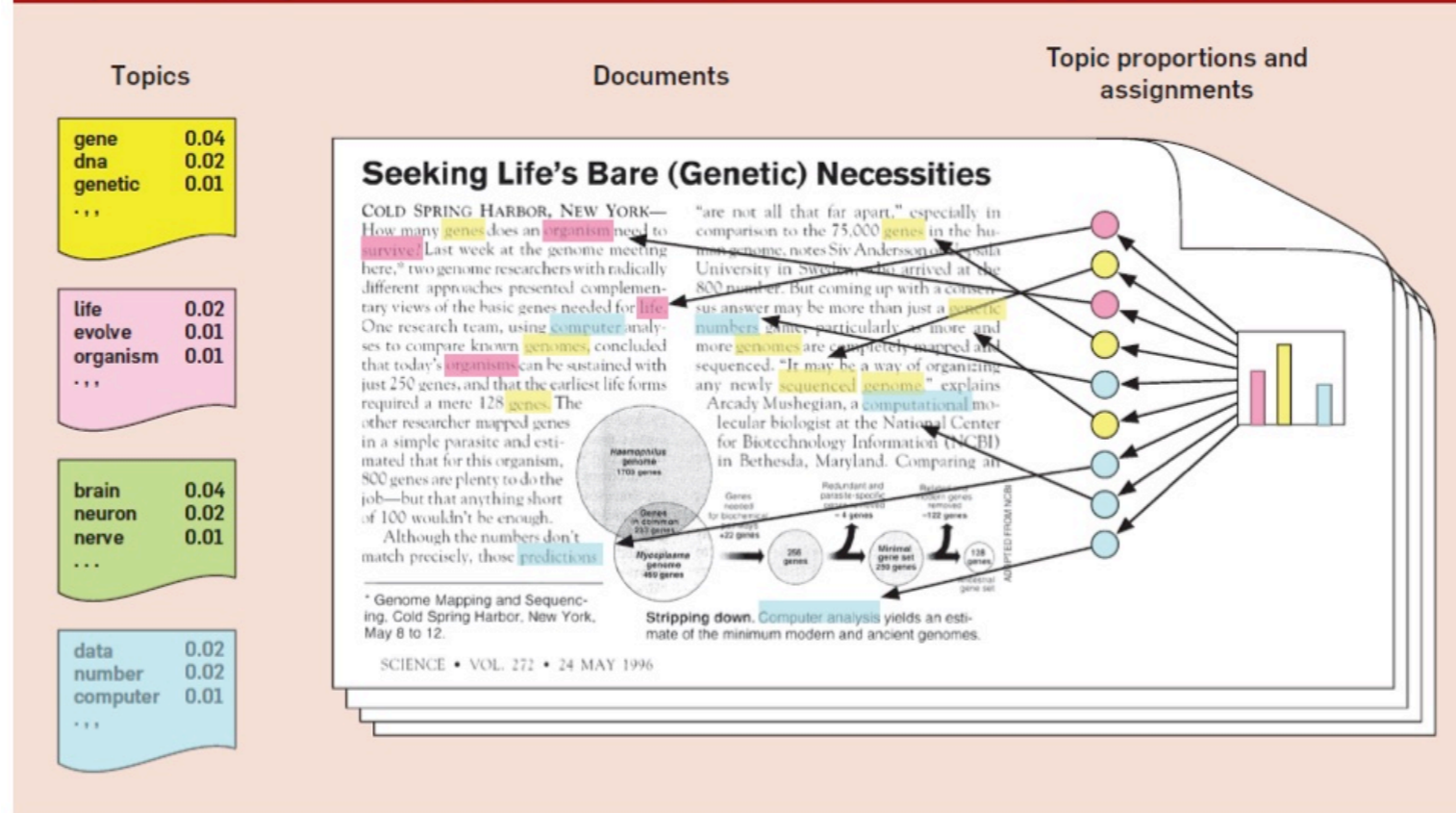


- Cluster nodes in a graph.
- Analysis of social network data.

TOPIC MODELLING

Blei, Ng & Jordan'06

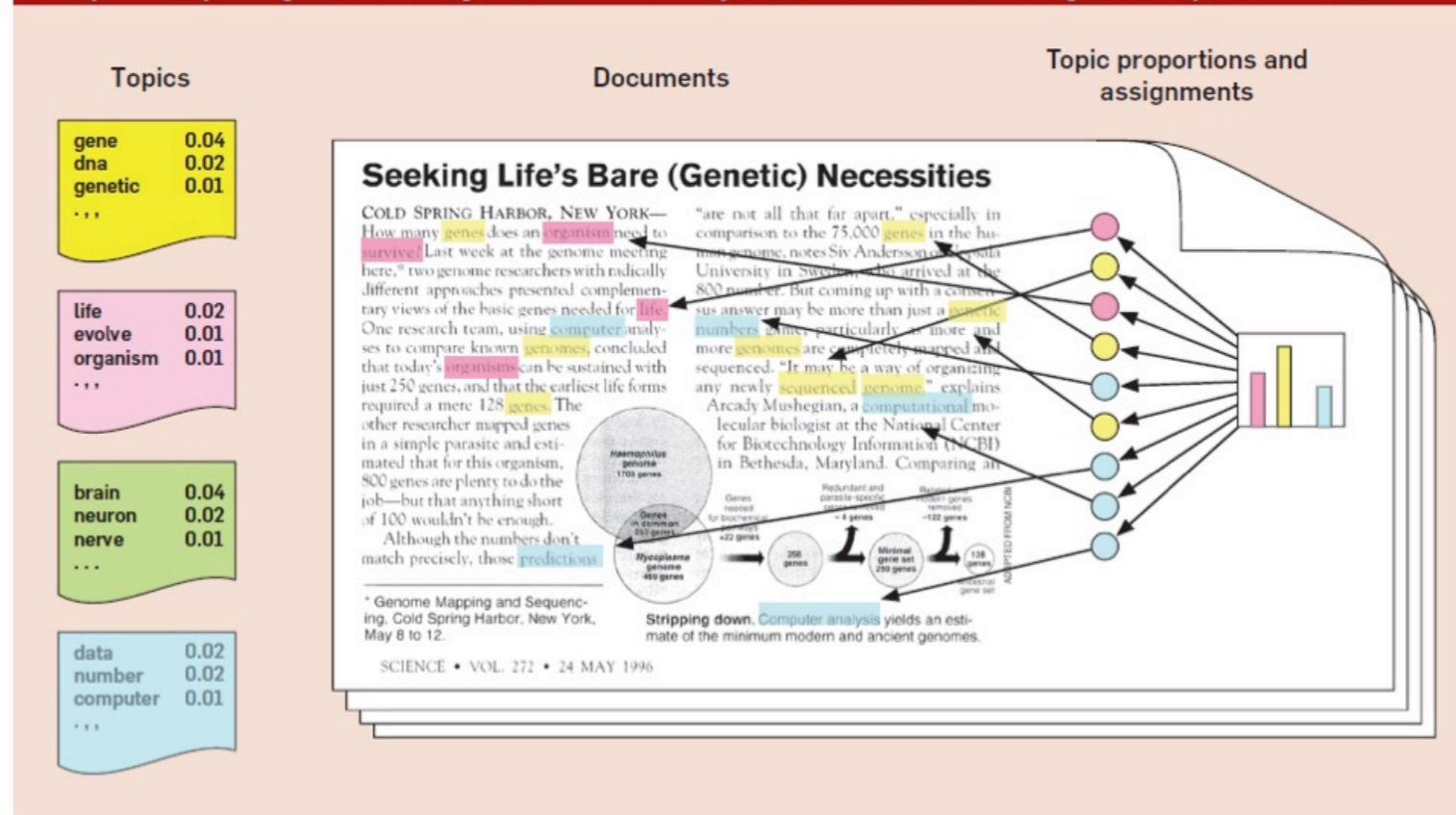
Figure 1. The intuitions behind latent Dirichlet allocation. We assume that some number of "topics," which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); 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.



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.

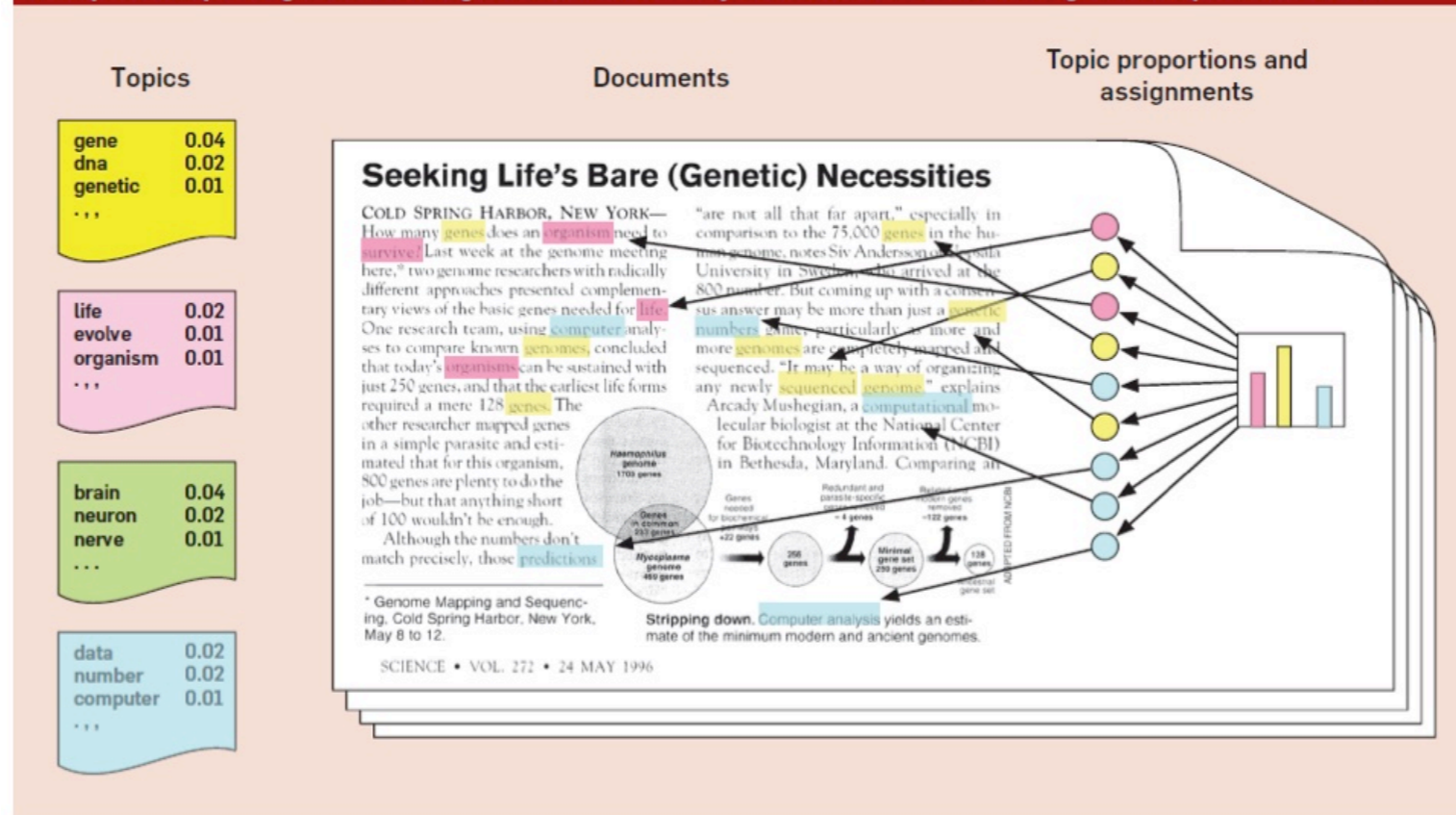


- Probabilistic generative model for documents

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.

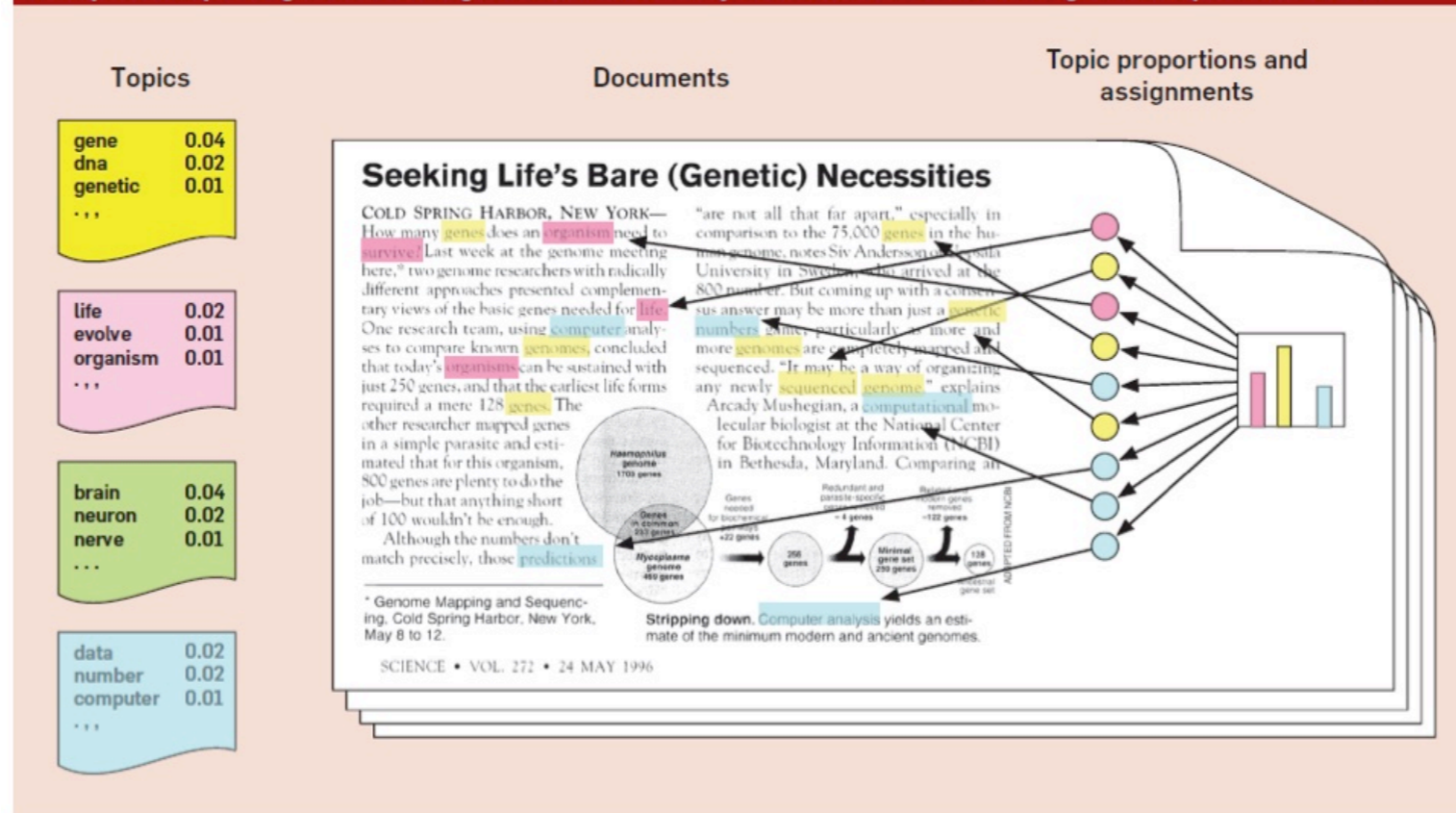


- Probabilistic generative model for documents
- Each document has a fixed distribution over topics, each topic is has a fixed distribution over words belonging to it

TOPIC MODELLING

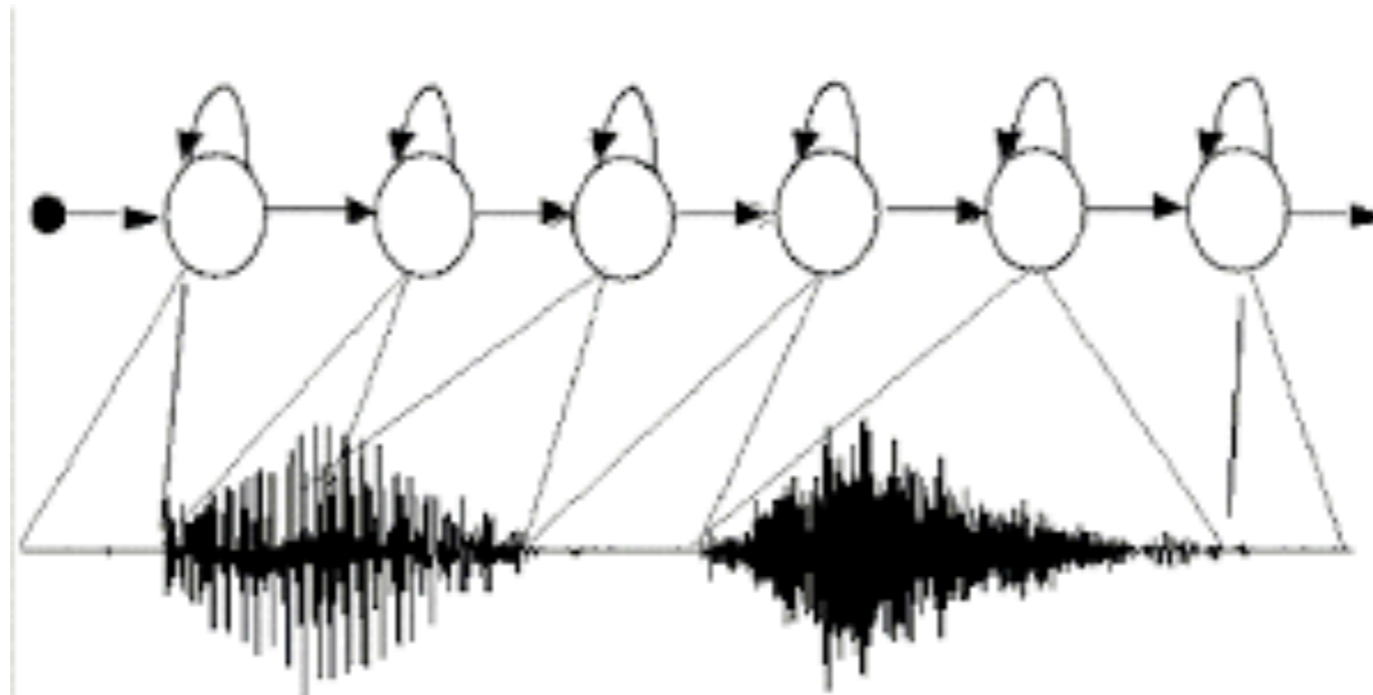
Blei, Ng & Jordan'06

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

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

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

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

WHAT YOU CAN TAKE HOME

- How to think about a learning problem and formulate it

WHAT YOU CAN TAKE HOME

- How to think about a learning problem and formulate it
- Well known methods and how and why they work

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

Clustering

CLUSTERING

- Grouping sets of data points s.t.
 - points in same group are similar
 - points in different groups are dissimilar
- A form of unsupervised classification where there are no predefined labels

CLUSTERING

- Partition data into K disjoint groups

CLUSTERING

- Partition data into K disjoint groups
- Compression or Quantization
 - Compress n points into K representatives/groups

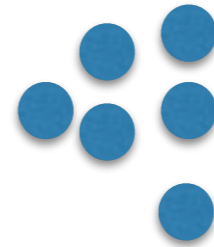
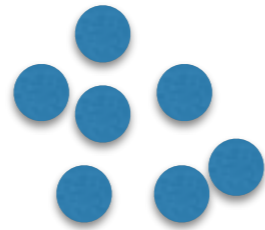
CLUSTERING

- Partition data into K disjoint groups
- Compression or Quantization
 - Compress n points into K representatives/groups
- Visualization or Understanding
 - Taxonomy: Animals Vs plants Vs Microbes, Science Vs Math Vs Social Sciences
 - Segmentation: different types of customers, students etc. Find natural groupings in data

CLUSTERING

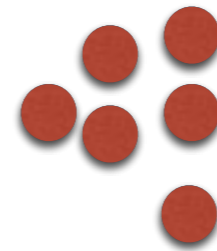
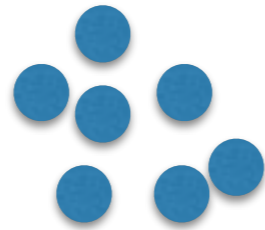
- Partition data into K disjoint groups
- Compression or Quantization
 - Compress n points into K representatives/groups
- Visualization or Understanding
 - Taxonomy: Animals Vs plants Vs Microbes, Science Vs Math Vs Social Sciences
 - Segmentation: different types of customers, students etc. Find natural groupings in data
- What this does not include: items belonging to more than one type

EXAMPLES



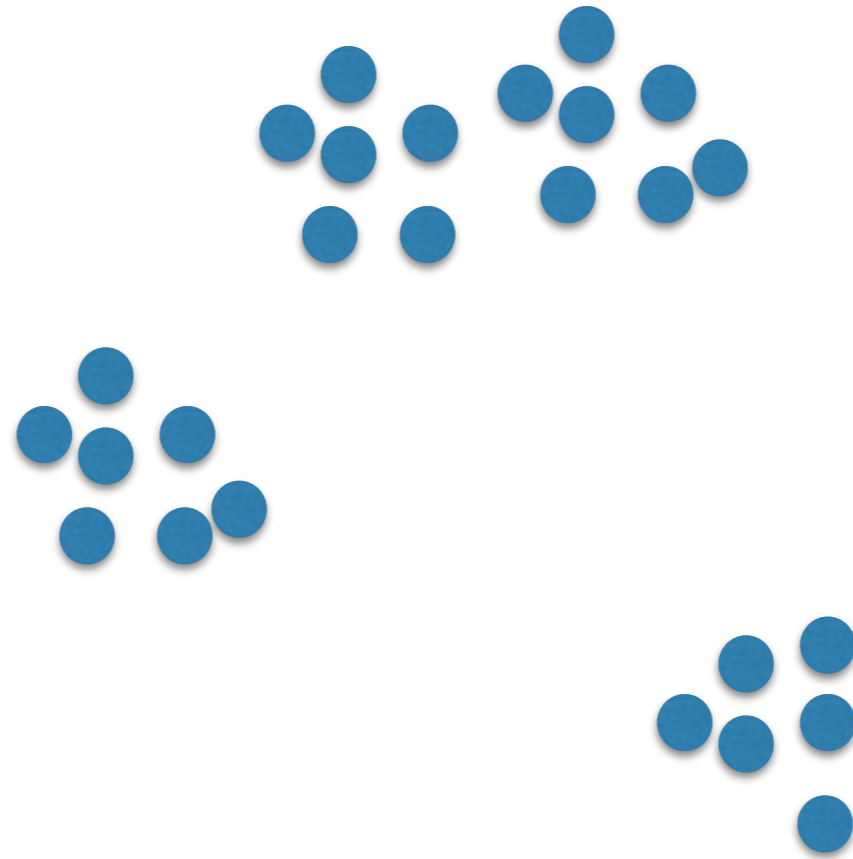
What are the clusters?

EXAMPLES



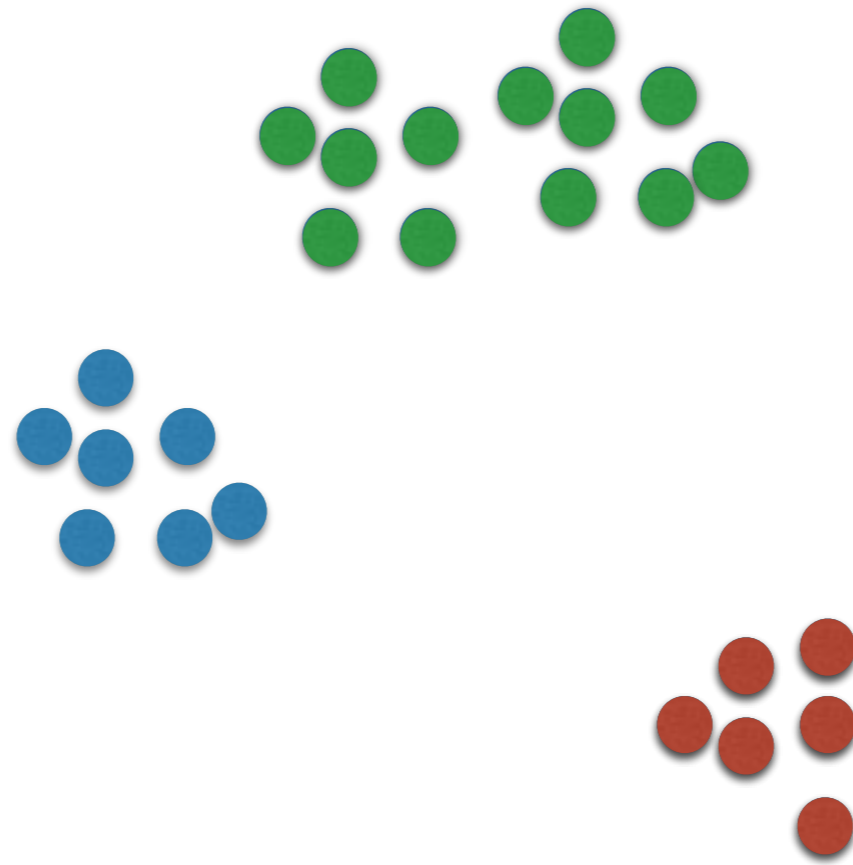
What are the clusters?

EXAMPLES



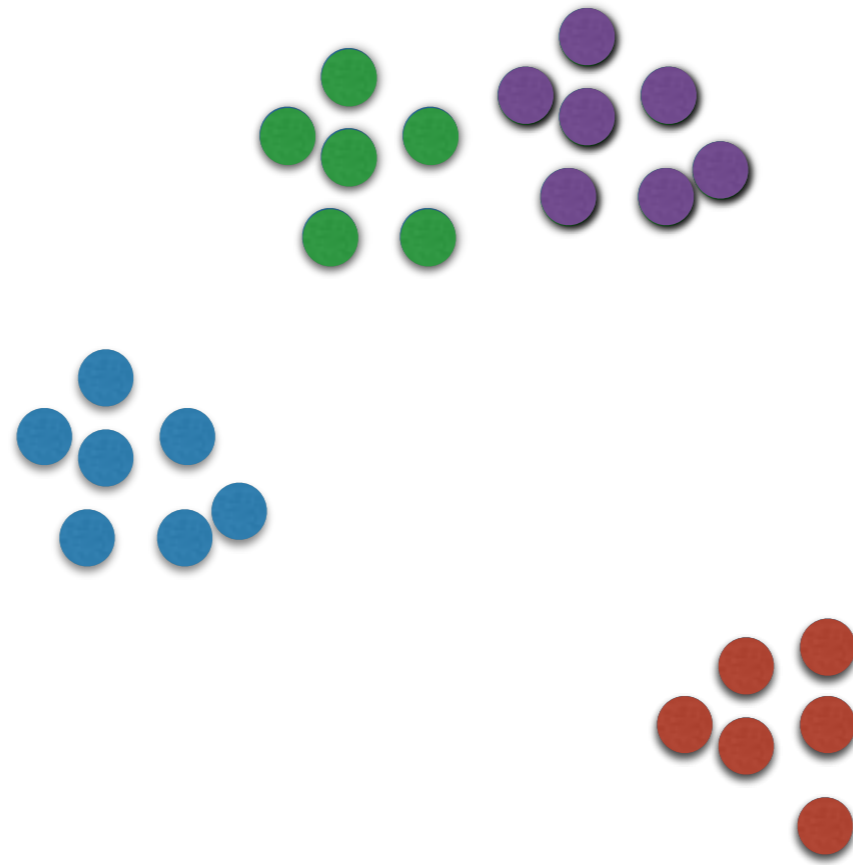
What are the clusters?

EXAMPLES



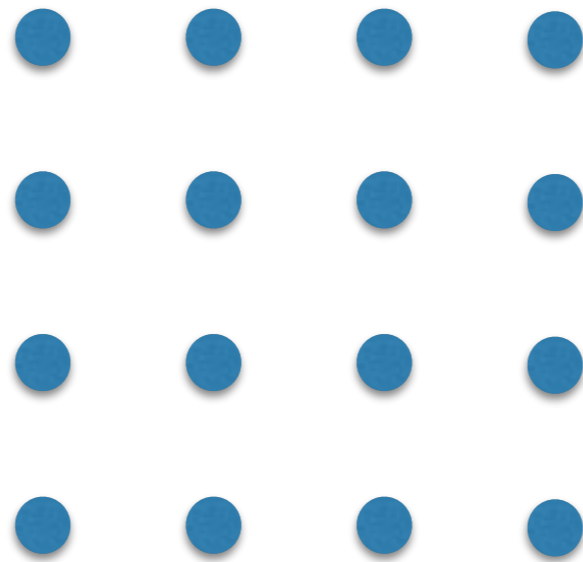
What are the clusters?

EXAMPLES



What are the clusters?

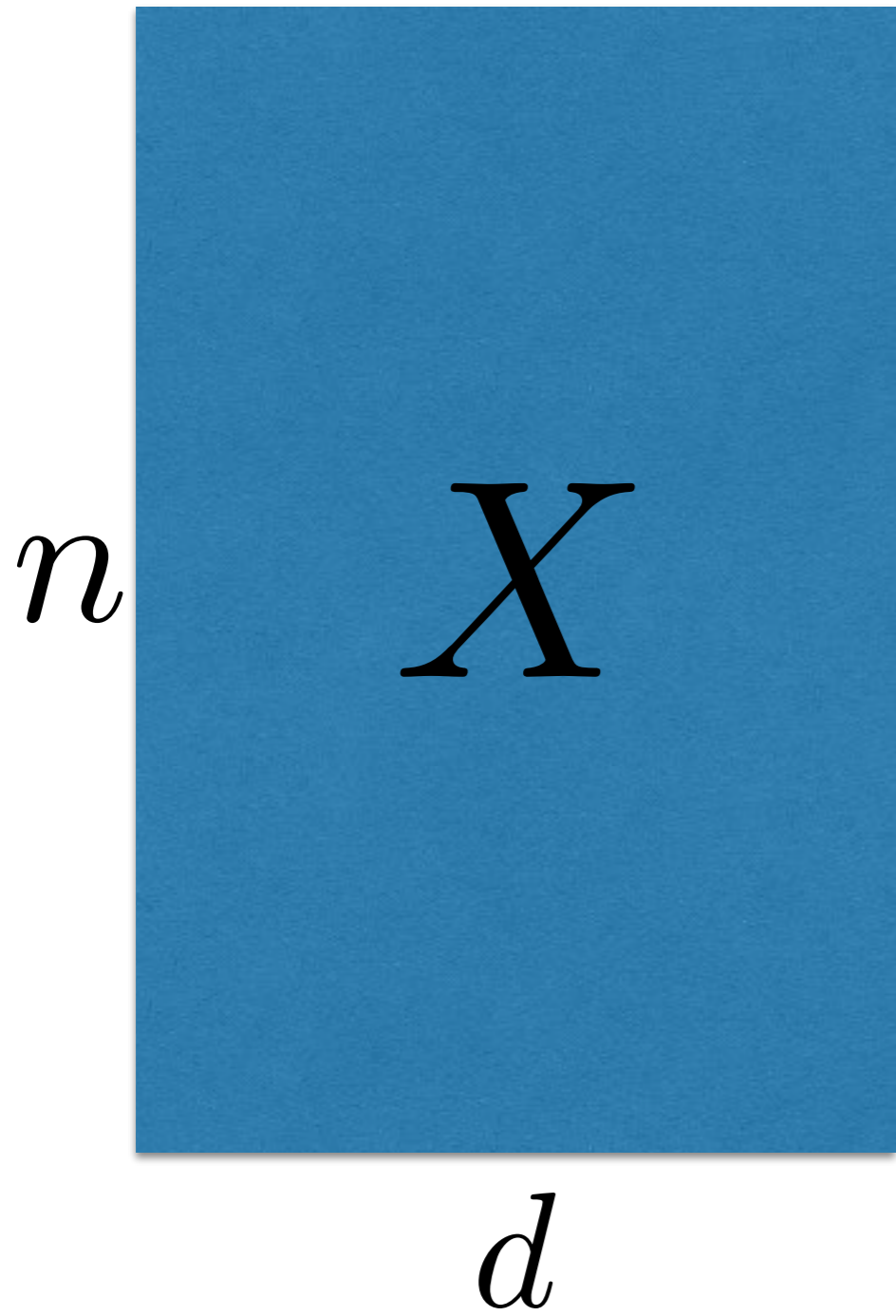
EXAMPLES



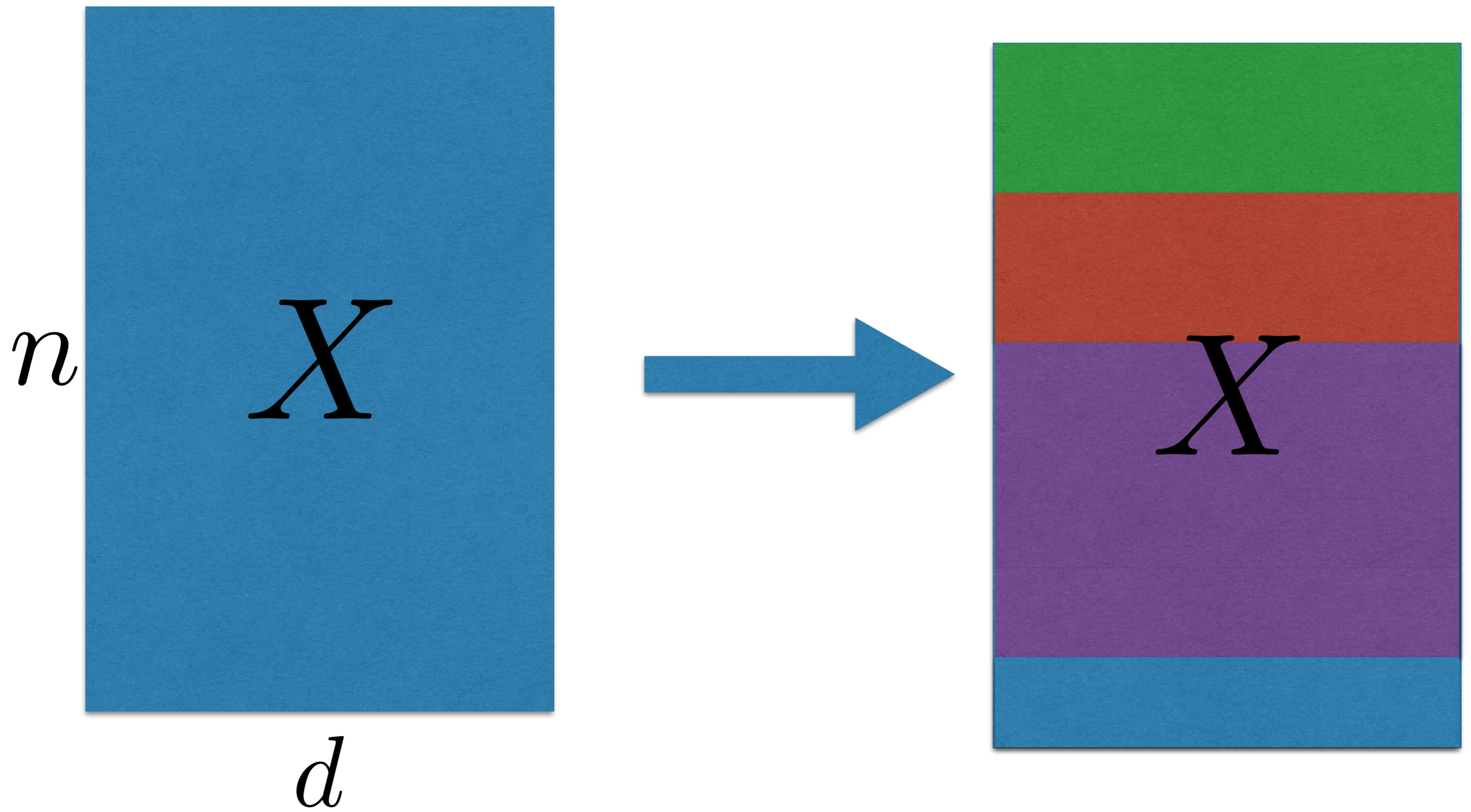
What are the clusters?

CLUSTERING

CLUSTERING



CLUSTERING



Can we formalize criterion/ objectives for clustering?

- . Assume points are represented as vectors
- . Use Euclidean distances for now
- . Similar points in same cluster
- . Points across clusters are dissimilar