Data Mining Techniques

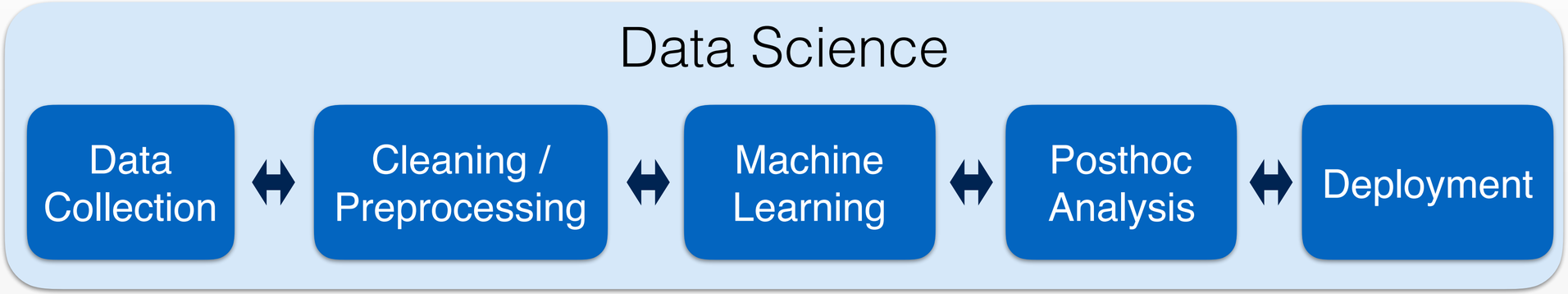
DS 5230 - Summer 2021

Course Overview

Shantanu Jain

What is Data Mining?

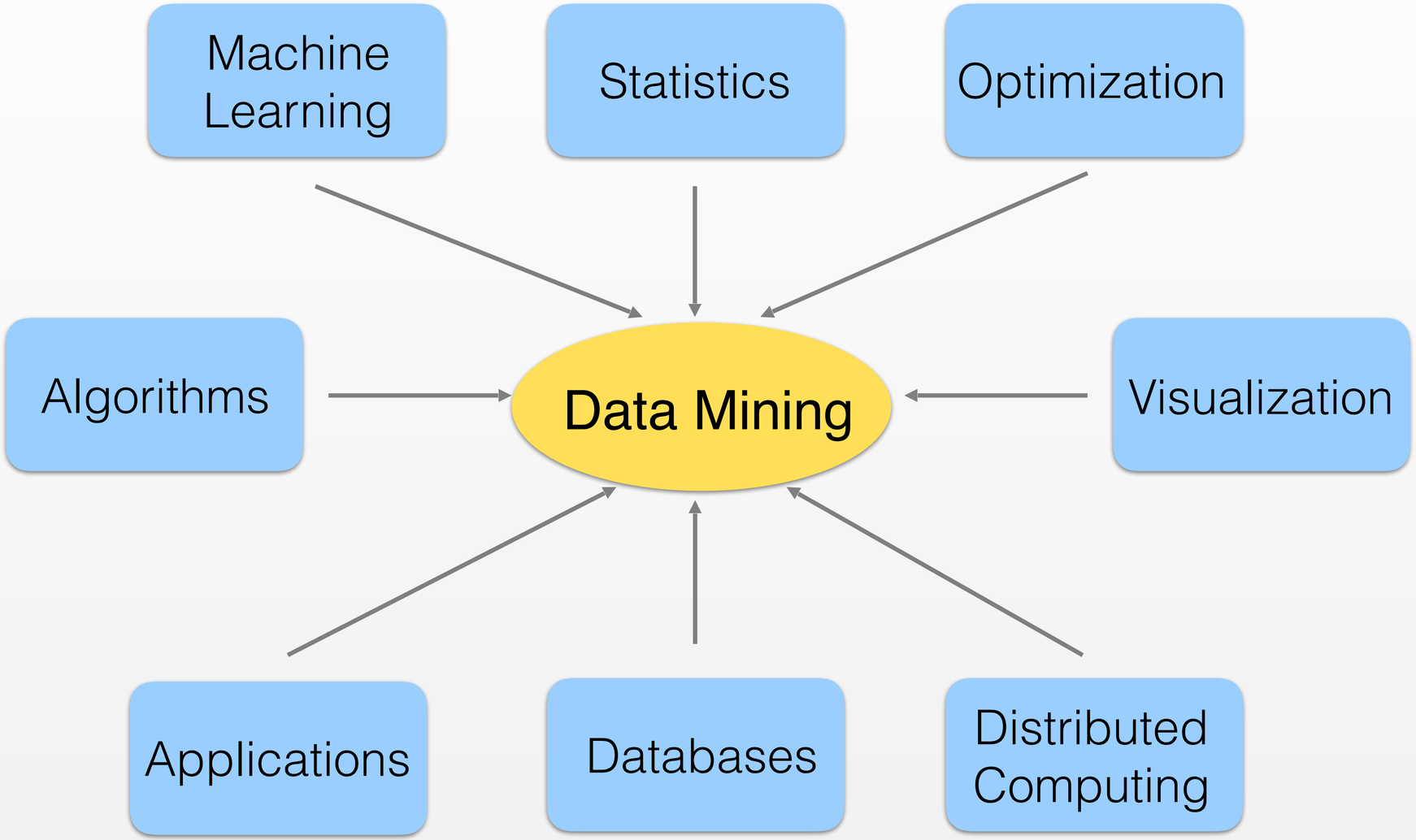
Machine Learning Perspective



API Sensors Scraper	Integration Normalization Feature selection Dimension reduction	Association Rules Classification Clustering Outlier detection	Evaluation Visualization Interpretation	Pipelining Scaling

(*slide adapted from Nate Derbinsky*)

Intersection of Disciplines



(*slide adapted from Han et al. Data Mining Concepts and Techniques*)

Applications

- Technology:

  Self driving car, Intelligent personal assistant, Biometrics

- Business:

  Online Advertisement, A/B testing, Customer segmentation, Dynamic pricing, Product recommendation, Resource allocation.

- Science:

  Climate Modeling, Astronomy, Chemical discovery.

- Medicine:

  Drug discovery, Disease diagnosis, Understanding the mechanism of a disease.

- Policy:

  Use yelp reviews to assign food inspectors to restaurants. Assisting judges to decide if bail could be granted to a defendant. Decide flights to which countries or cities should be cancelled to control the spread of an epidemic, with tolerable impact on commerce.

Data Source

- Single database
  - Find relevant features
- Multiple databases
  - Requires data integration
- Data stream
  - Sensor data, e-commerce transactions, web searches
- Unstructured and/or semi-structured
  - Tweets, images, videos, requires heavy data-preprocessing
- Data might not exist
  - Might require a clinical trial or a survey.

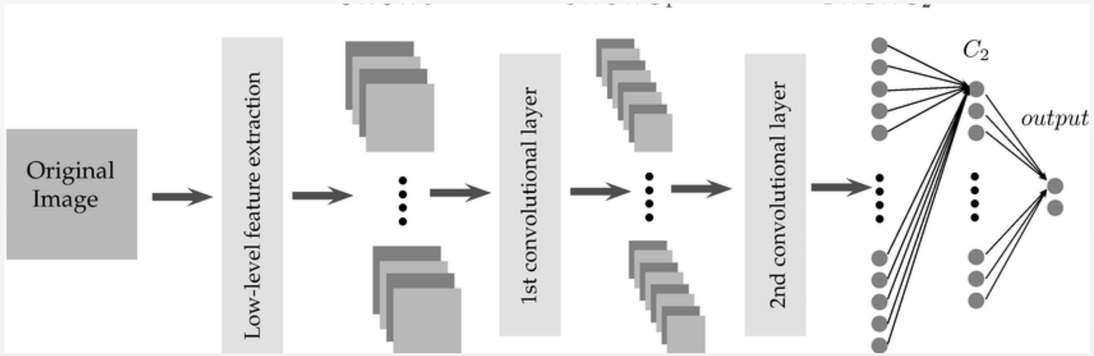
Algorithm

Model Specification

Linear Regression

$$\hat{y} = w_0 + w_1 x_1 + w_2 x_2$$

Deep Neural Network



Convolutional Neural Network

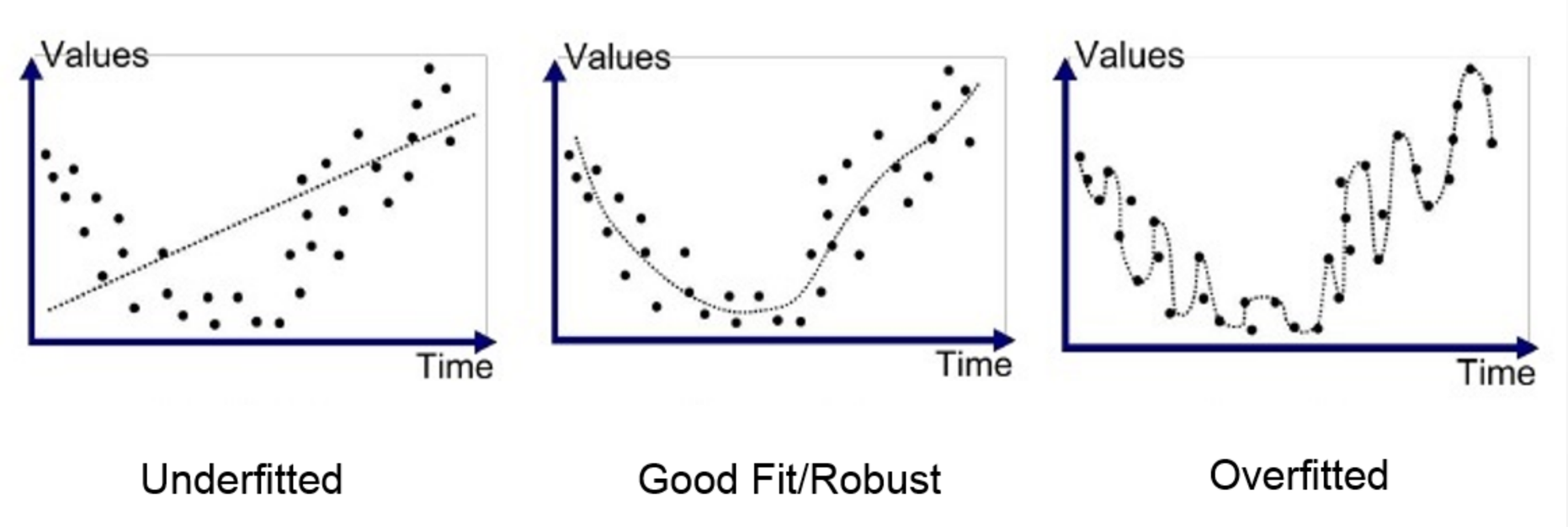
Parameter Fitting

$$\hat{\boldsymbol{w}} = \text{argmin}_{\boldsymbol{w}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

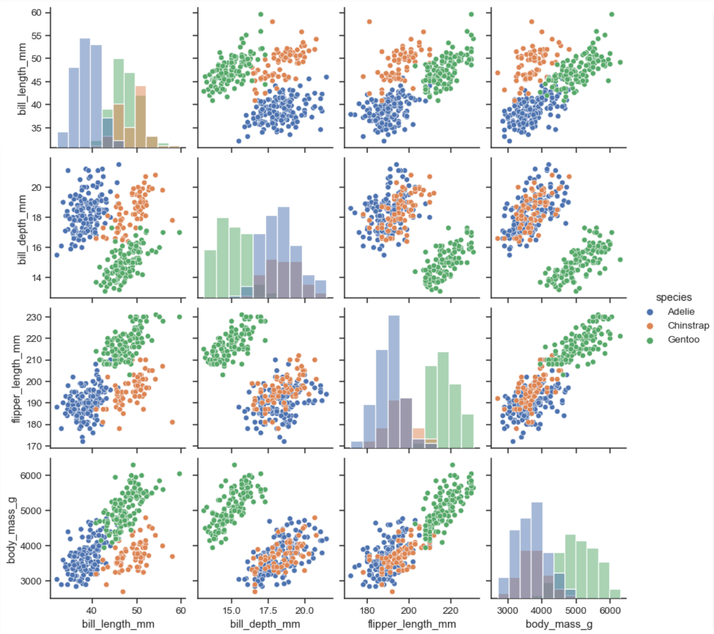
$$= (X^T X)^{-1} X^T y$$

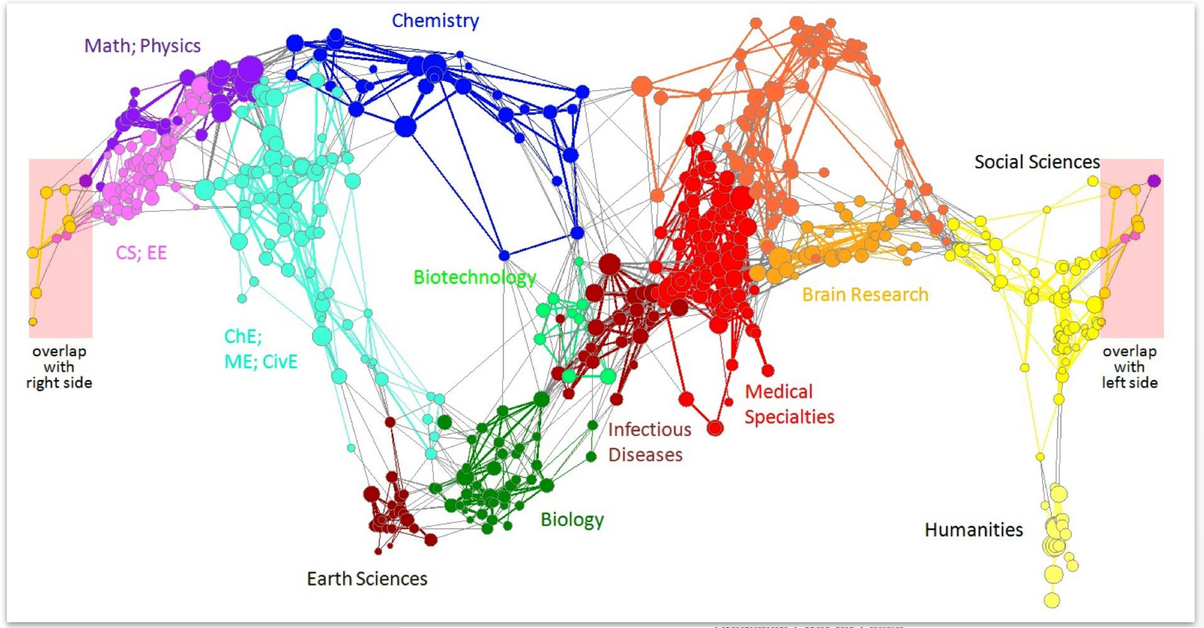
Back Propagation with Stochastic/Batch Gradient Descent.

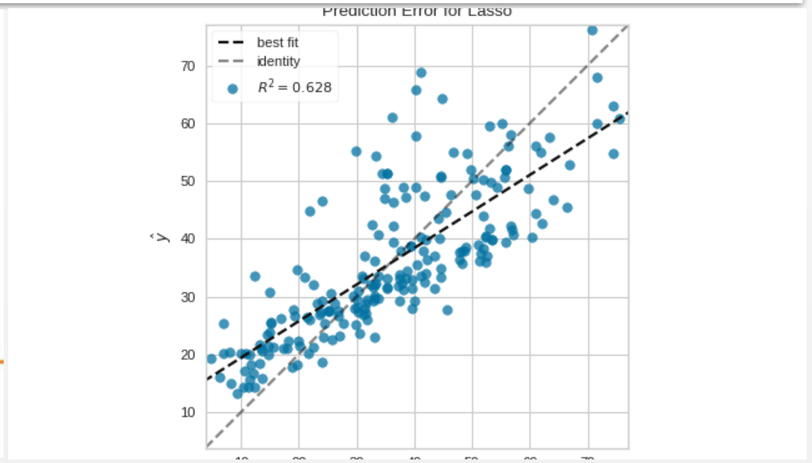
Underfitting/Overfitting



Visualization







Machine Learning Methods

Supervised Learning

Given *labeled* examples, learn to make predictions for *unlabeled* examples. *Example:* Image classification.

Unsupervised Learning (This Course)

Given *unlabeled* examples learn to identify structure. *Example:* Community detection in social networks.

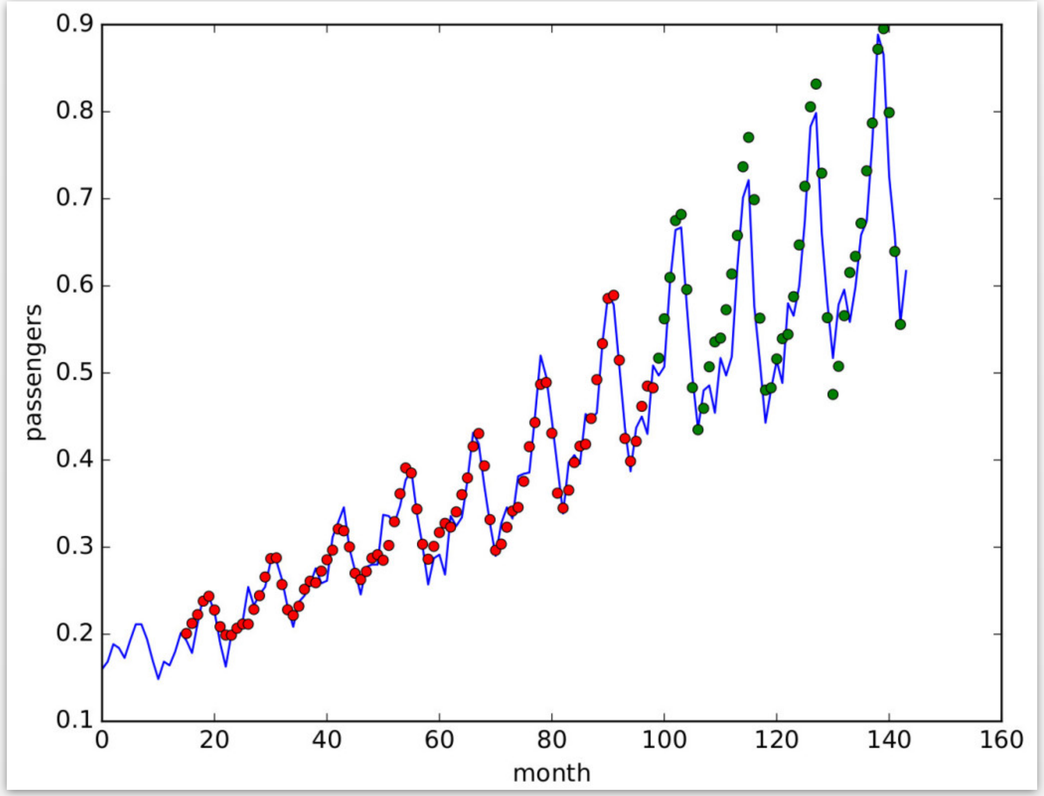
Reinforcement Learning

Learn to take *actions* that maximize future *reward.* *Example:* Targeting advertisements.

Supervised Learning: Regression

Goal: Use past labels (red) to learn trends that generalize to future data points (green)

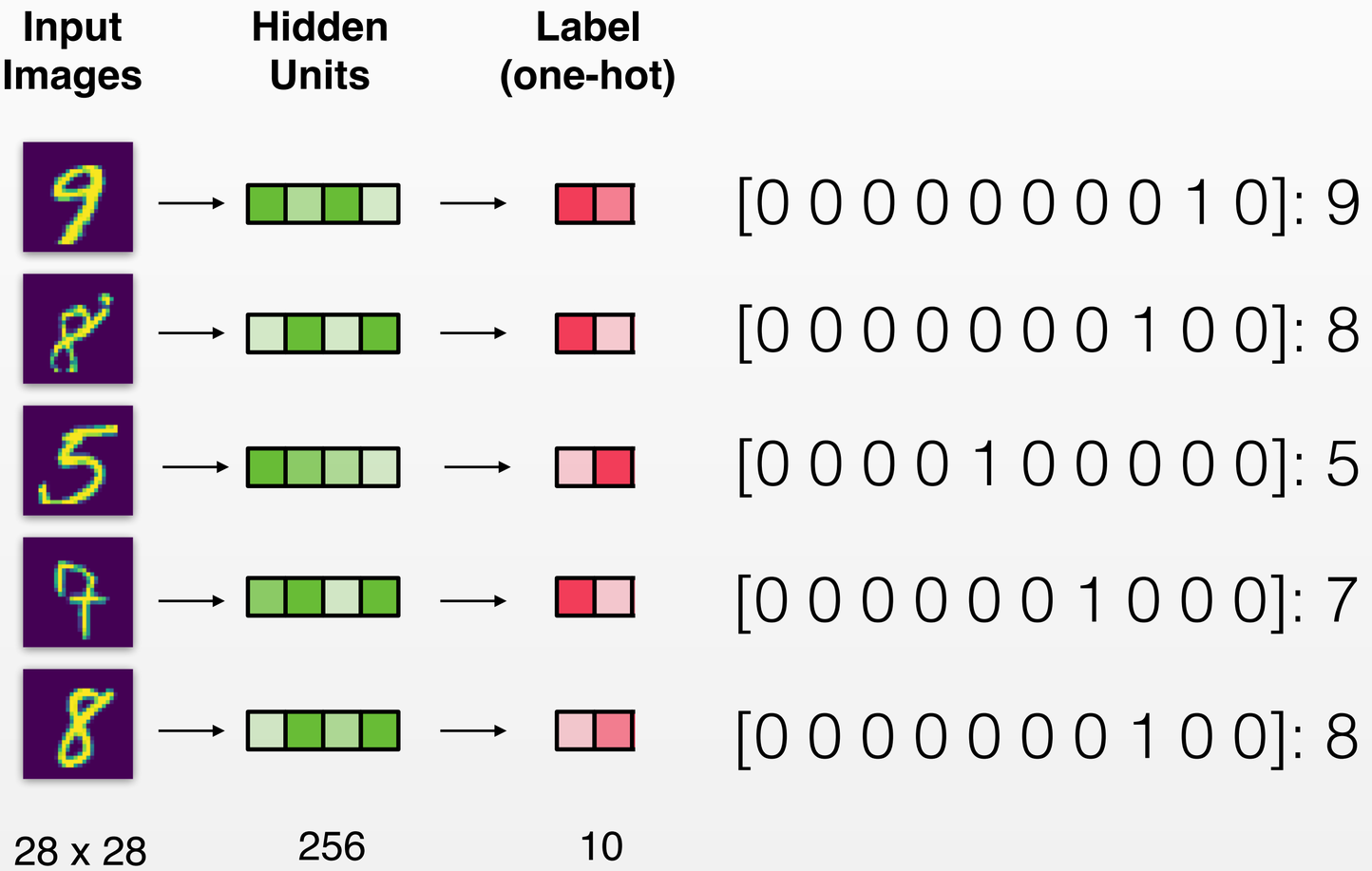
Time-series Data



source: https://am241.wordpress.com/tag/time-series/

Supervised Learning: Classification

Goal: Predict a *discrete* label.



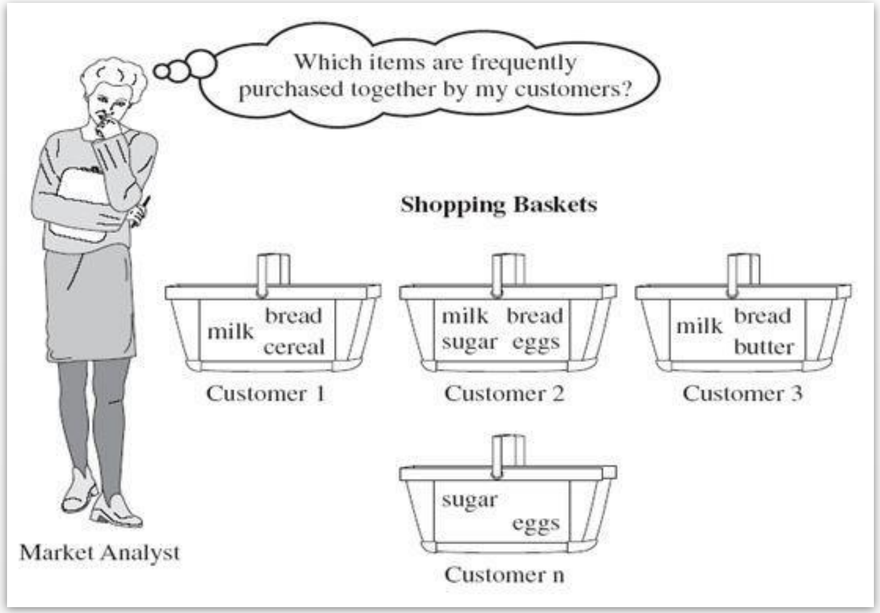
Unsupervised Learning

Goal: Can we make predictions in absence of labels?

Methods in this Course:

- Frequent Itemsets and Association rule mining
- Dimensionality Reduction
- Clustering
- Topic Modeling
- Community Detection
- Link Analysis
- Recommender Systems

Association Rule Mining



Baskets of items

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Association Rules

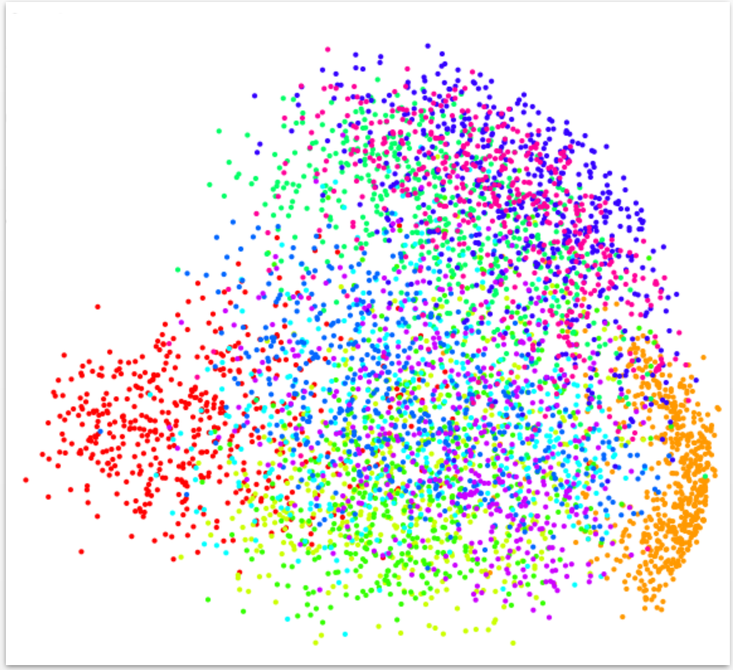
{Milk} --> {Coke}
{Diaper, Milk} --> {Beer}

Dimensionality Reduction

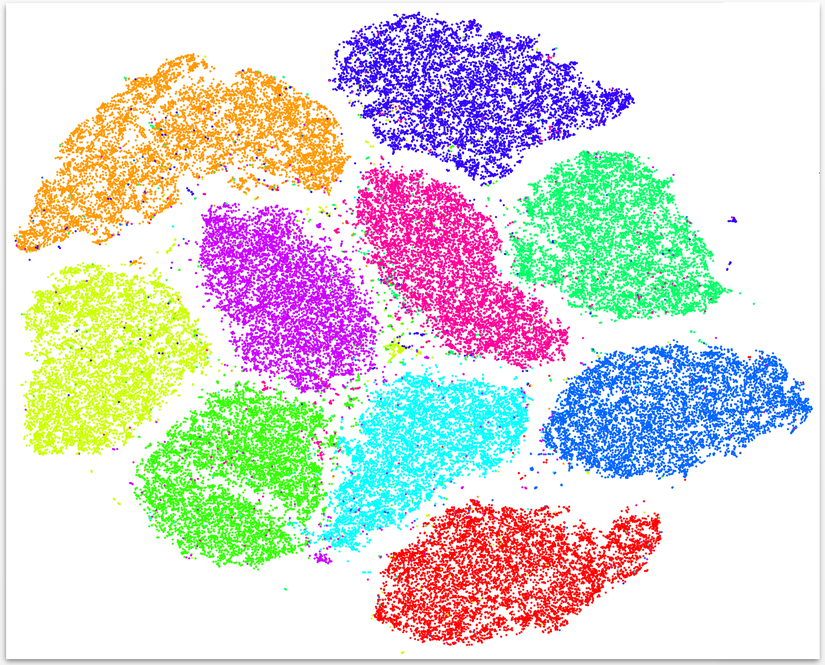
Goal: Map high dimensional data onto lower-dimensional data in a manner that preserves *distances/similarities*

Input Images

PCA (Linear)



TSNE (Non-linear)

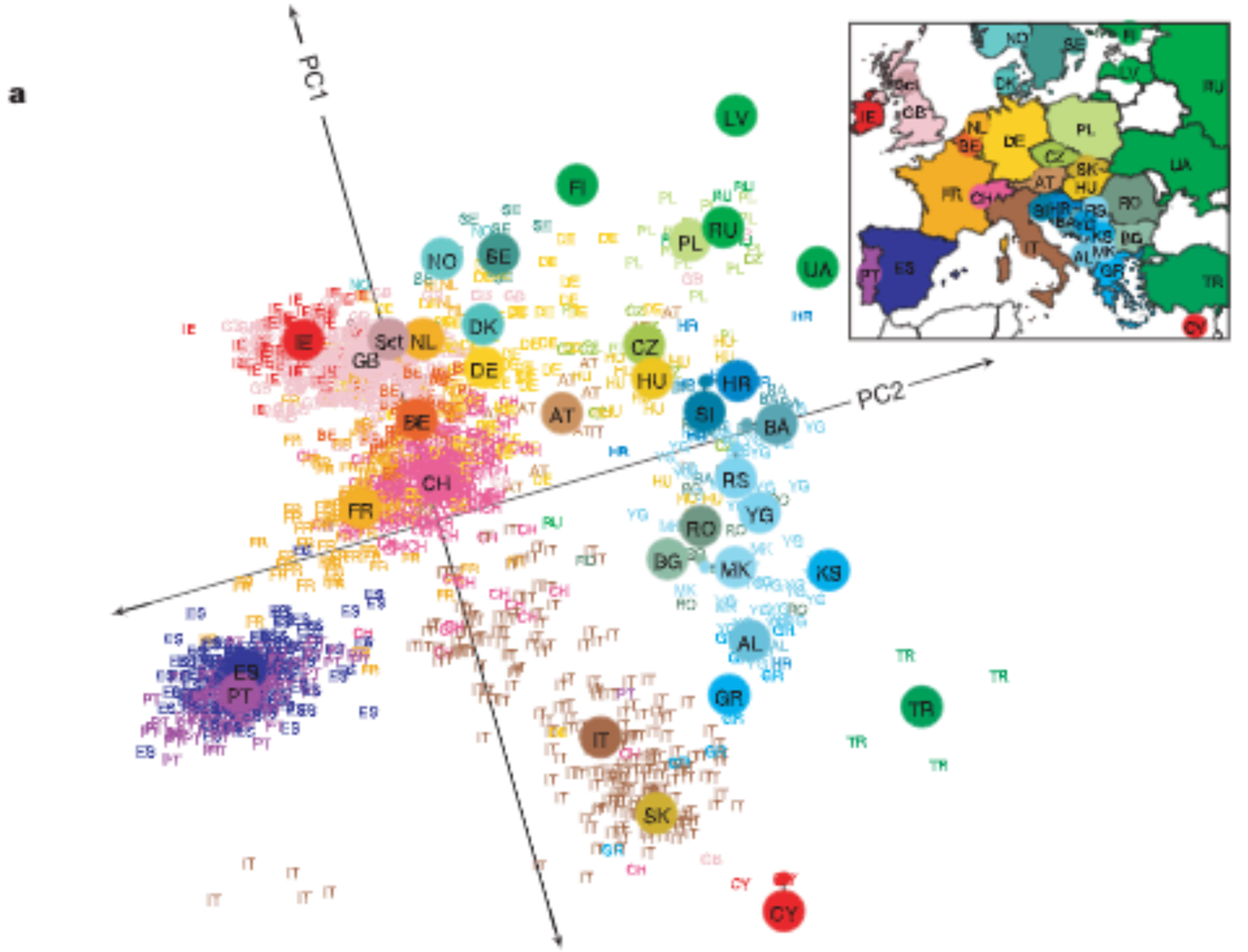


Published: 31 August 2008

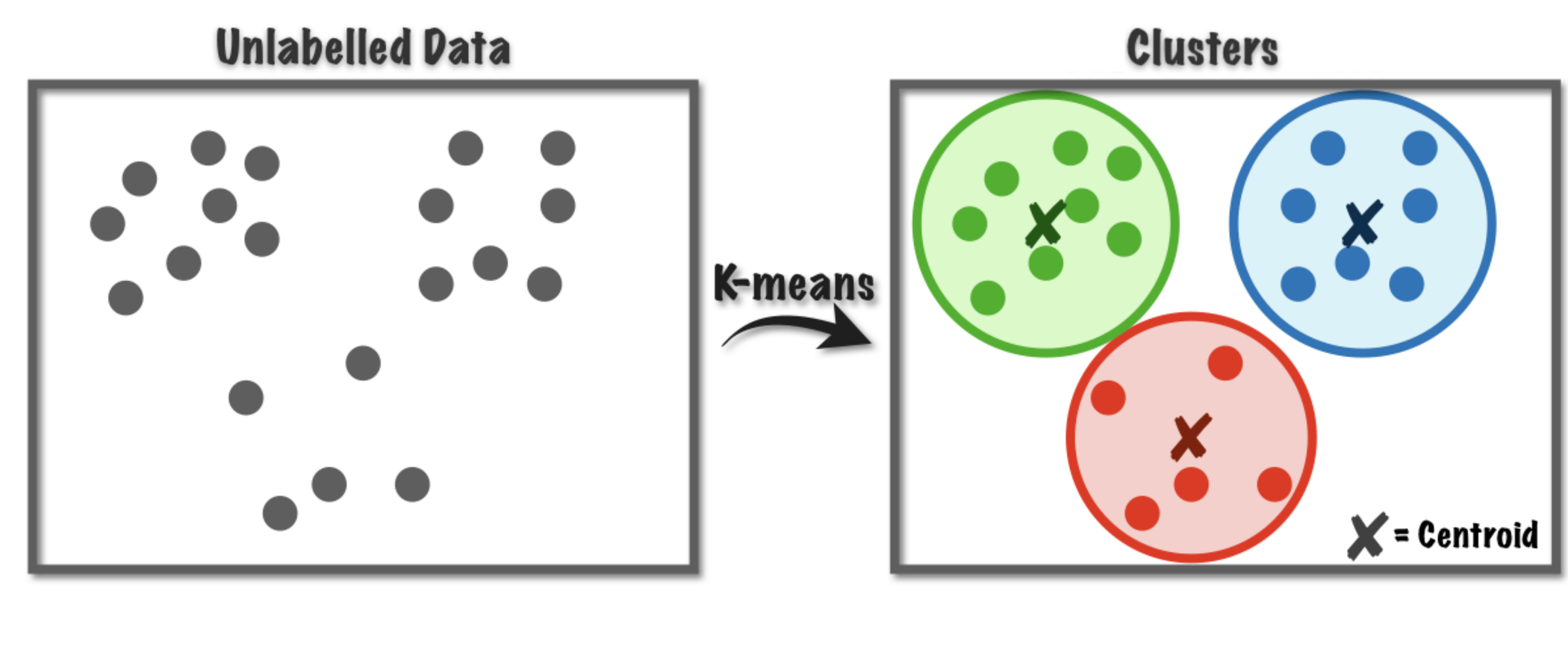
Genes mirror geography within Europe

John Novembre ✉, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R. Boyko, Adam Auton, Amit Indap, Karen S. King, Sven Bergmann, Matthew R. Nelson, Matthew Stephens & Carlos D. Bustamante

Nature **456**, 98–101(2008) | Cite this article

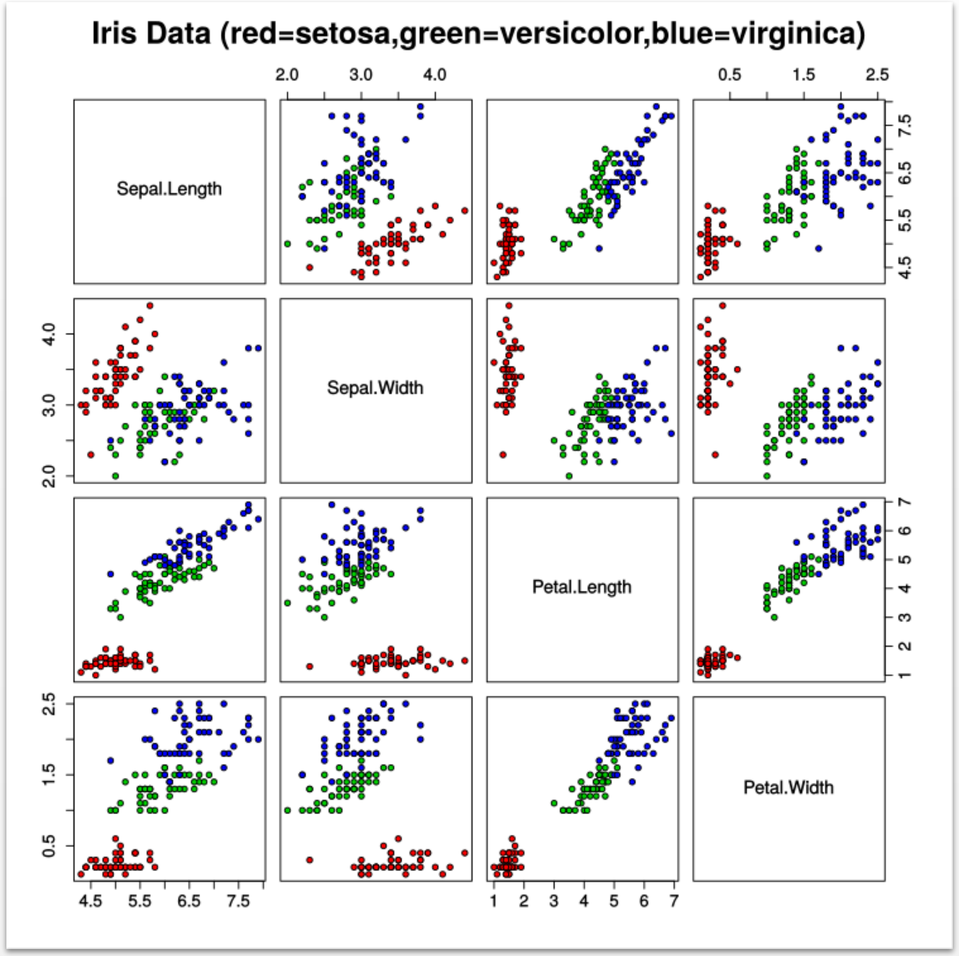


Clustering



Clustering

Example: Iris Data



Iris Setosa

Iris versicolor

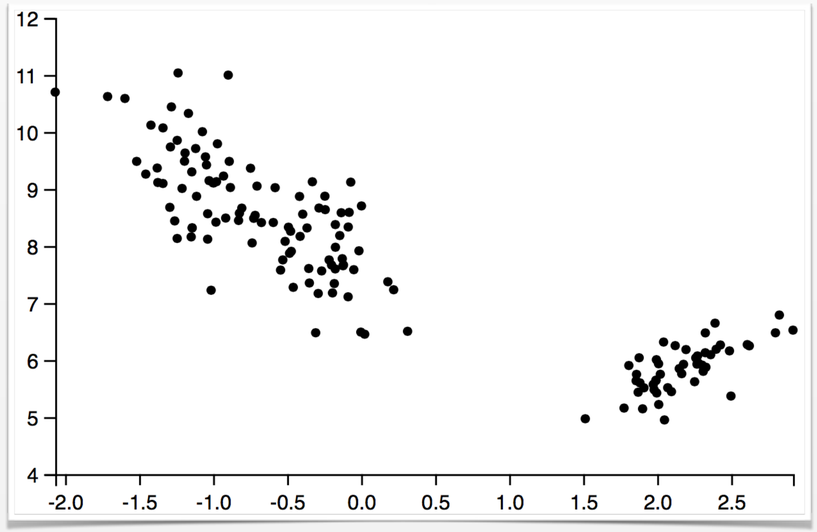
Iris virginica

https://en.wikipedia.org/wiki/Iris_flower_data_set

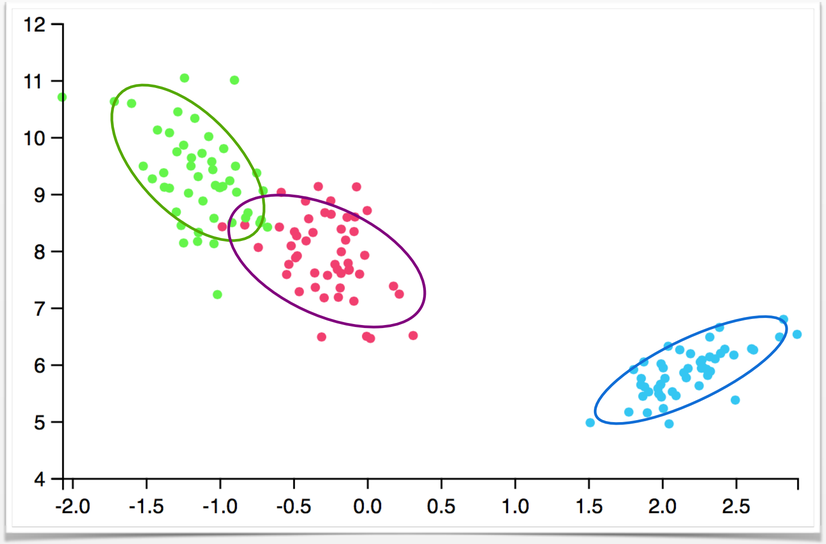
Clustering

Goal: Learn categories of examples
(i.e. classification without labels)

Iris Data (after PCA)



Inferred Clusters



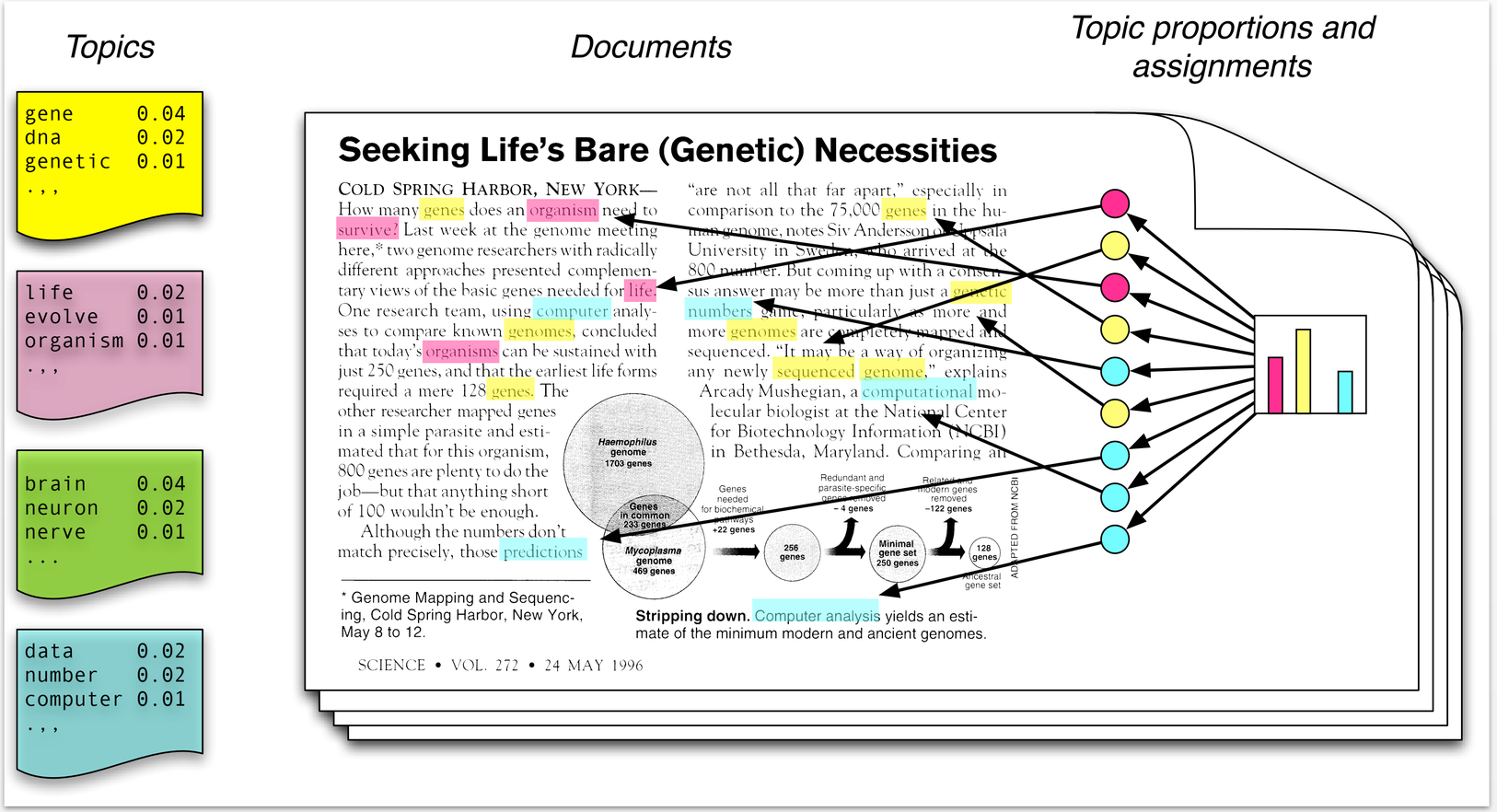
Clustering

- **Market segmenting and customized sales promotion campaigns.**
- **Discovering new concepts:**
  - In Biology, genes can be clustered together to discover new functions of a gene.
- **Medicine**
  - Clustering the patients based on their symptoms would reveal the different types of depression.
- **Increasing diversity of results:**
  - In Information retrieval, clustering can be applied to improve diversity results form a search query.
- **Summarize:**
  - Return a a few representative images from a huge corpus of images.
- **Downstream data analysis:**
  - Apply classification on representative points from each clustered
- **Compression:**
  - Vector quantization. Store or transmit the cluster representatives instead of all data points

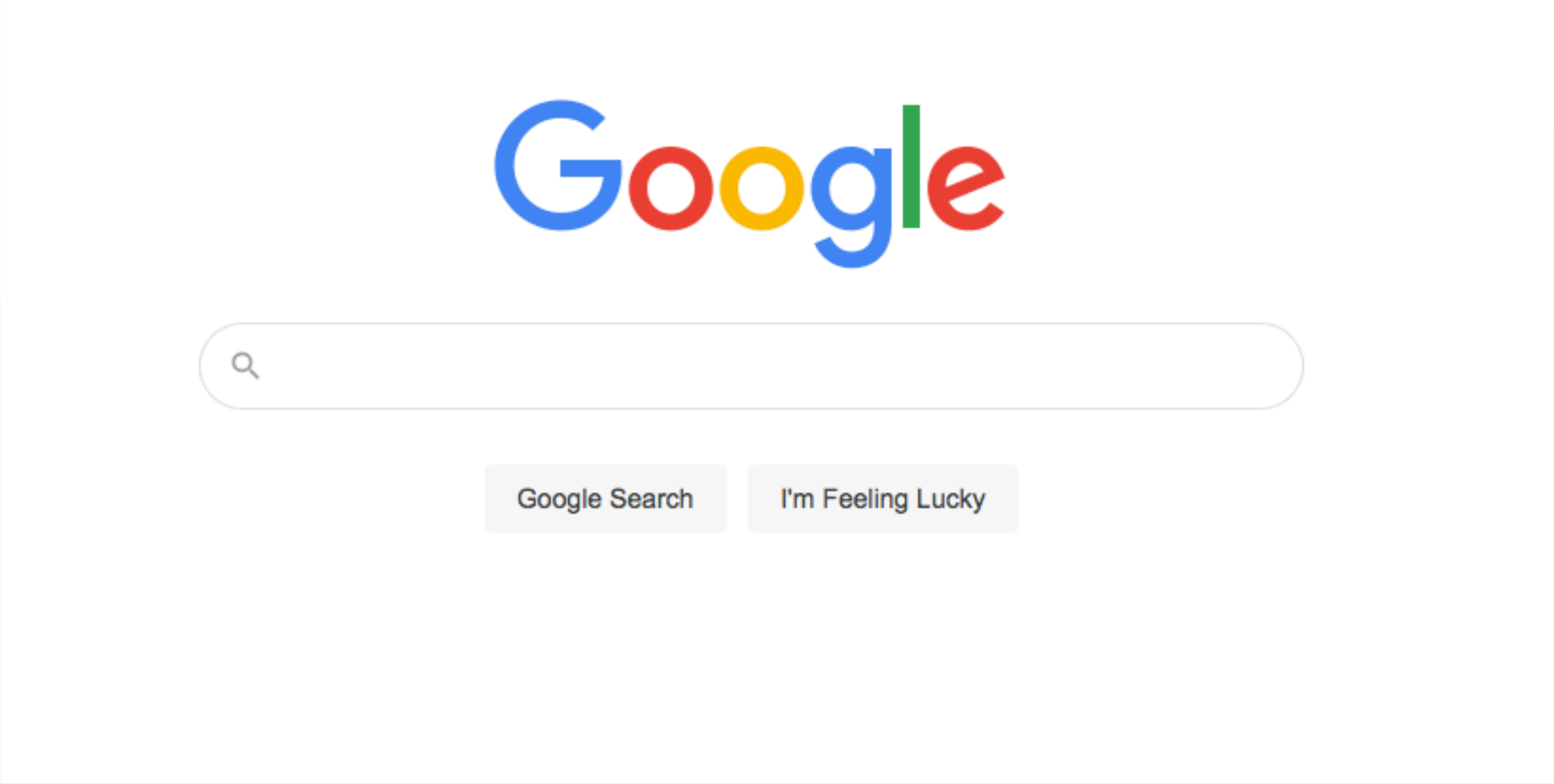
Topic Models

Goal: Learn topics (categories of words)
and quantify topic frequency for each document

- Scientific papers on google scholar.
- Articles from a news outlet.
- The web pages hosted on some domain name.
- The set of all books on amazon.com.
- All the email communications in an organization.

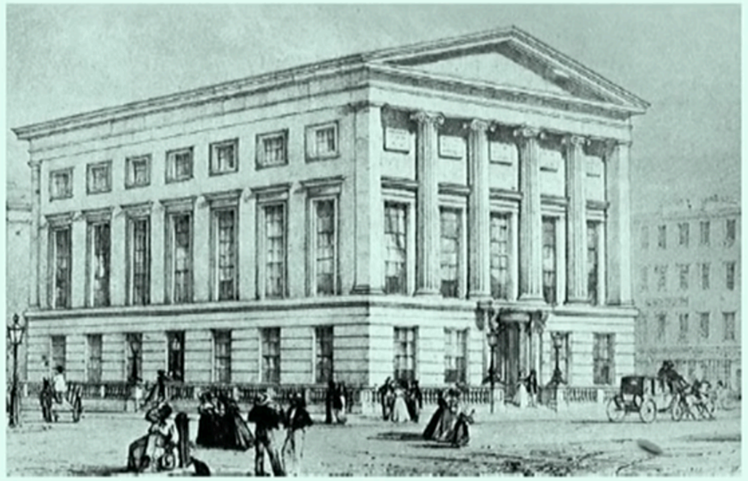


Improving Search Engine

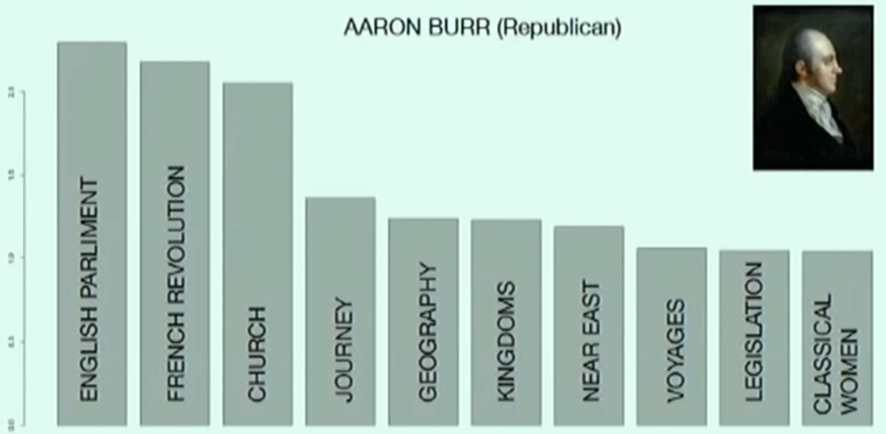


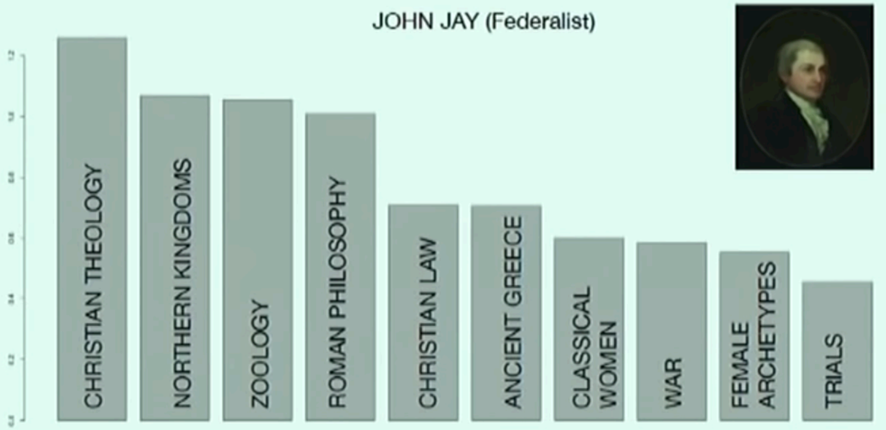
The documents scoring higher on the topics related to the query are relevant.

Sociology



- The New York Society library is the first library in New York City (1754)
- Mark Hoffman and Peter Bearman (sociology) are using collaborative topic models to explore the usage patterns of important figures in U.S. History
- The data
  - 1789 – 1806
  - 847 users (people like Aaron Burr, John Jay, Alexander Hamilton, etc.)
  - 2,327 items (items like The Prince)
  - 33M words; vocabulary of 8,000





Discovering Events

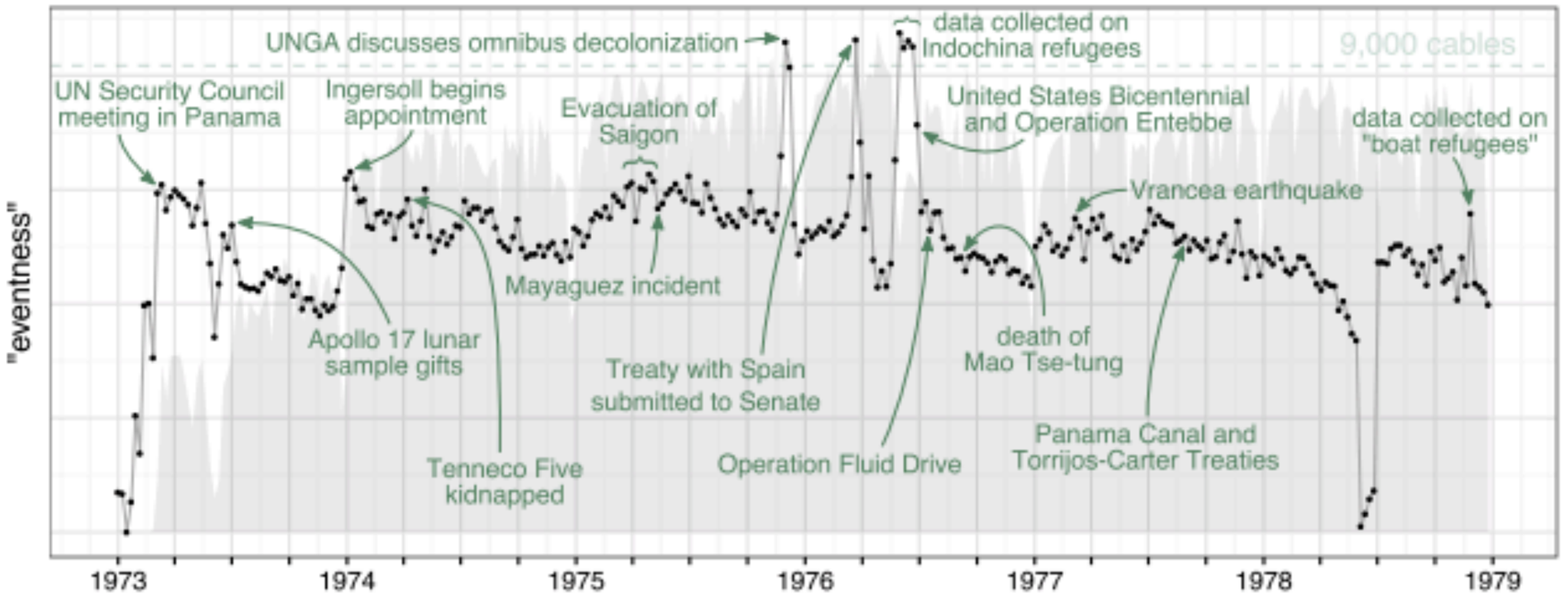
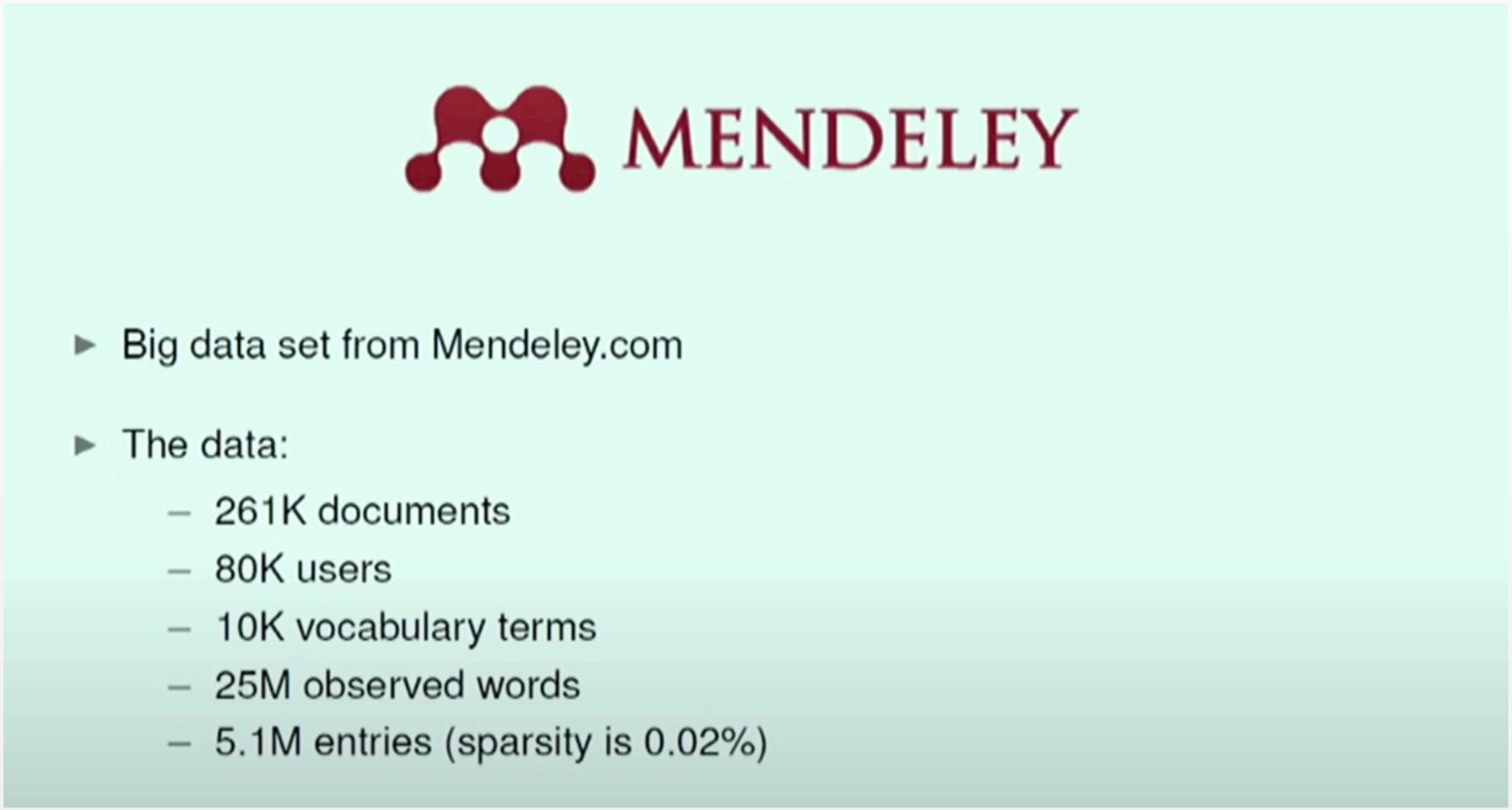


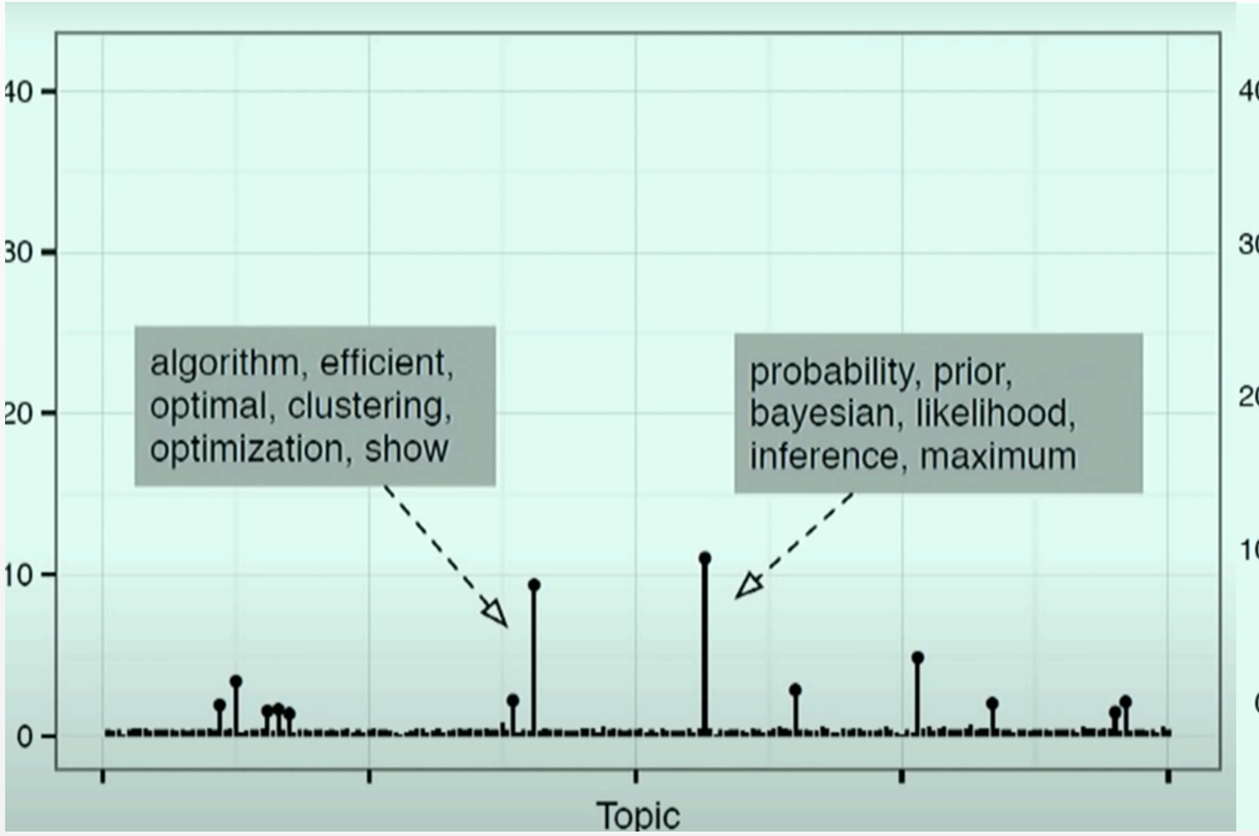
Figure 1: Capsule's analysis (described in detail in section 5) of two million cables from the National Archives' corpus. The y-axis represents a loose measure of "eventness" (equation (5)). The gray background depicts the number of cables sent over time.

Analysis of 2M diplomatic cables reveals interesting events

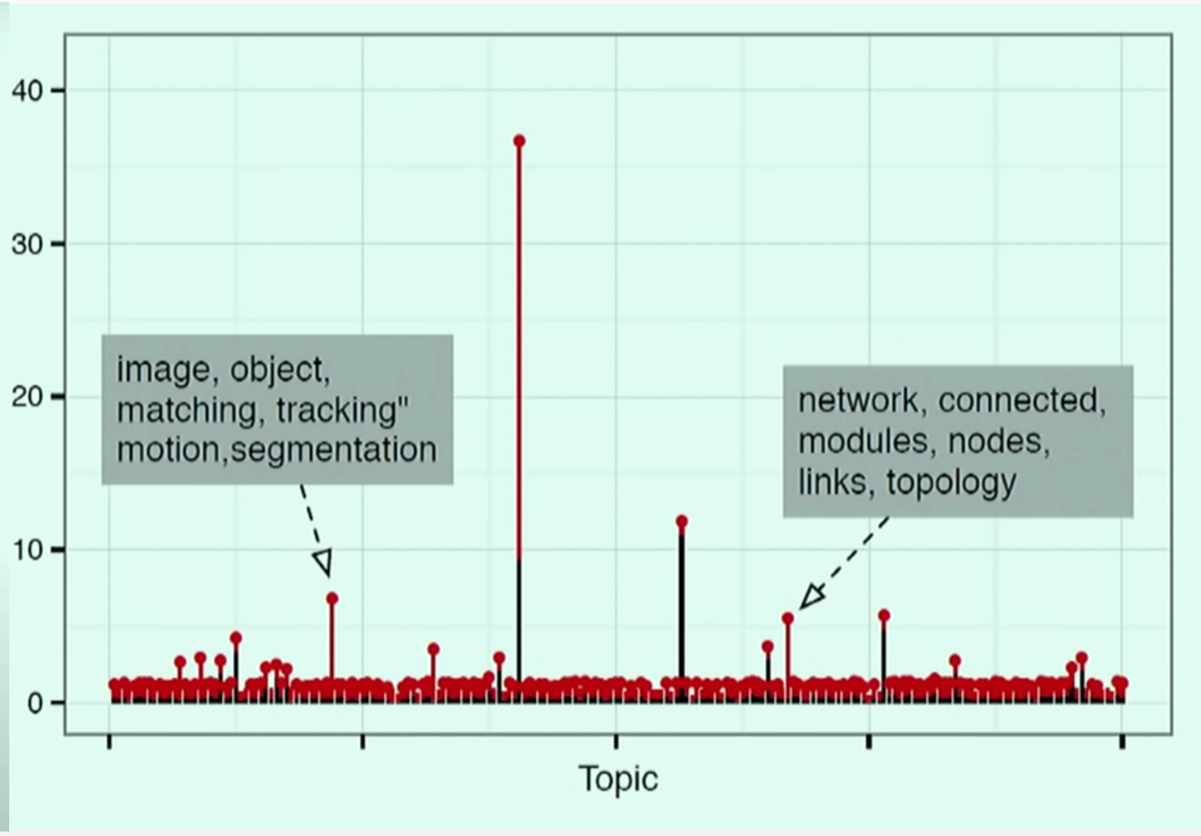
Scientific paper Recommendation engine



Scientific paper Recommendation engine



Topics associated to the EM Algorithm paper without user information

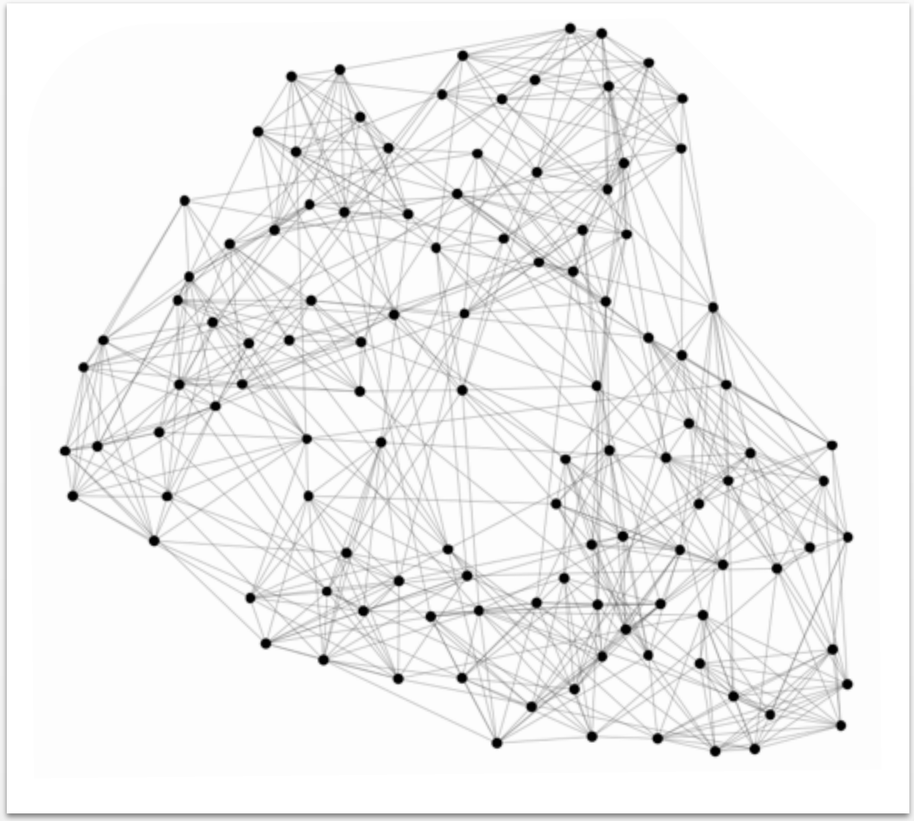


Topics associated to the EM Algorithm paper with user information

Dempster et al., Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society (1977)

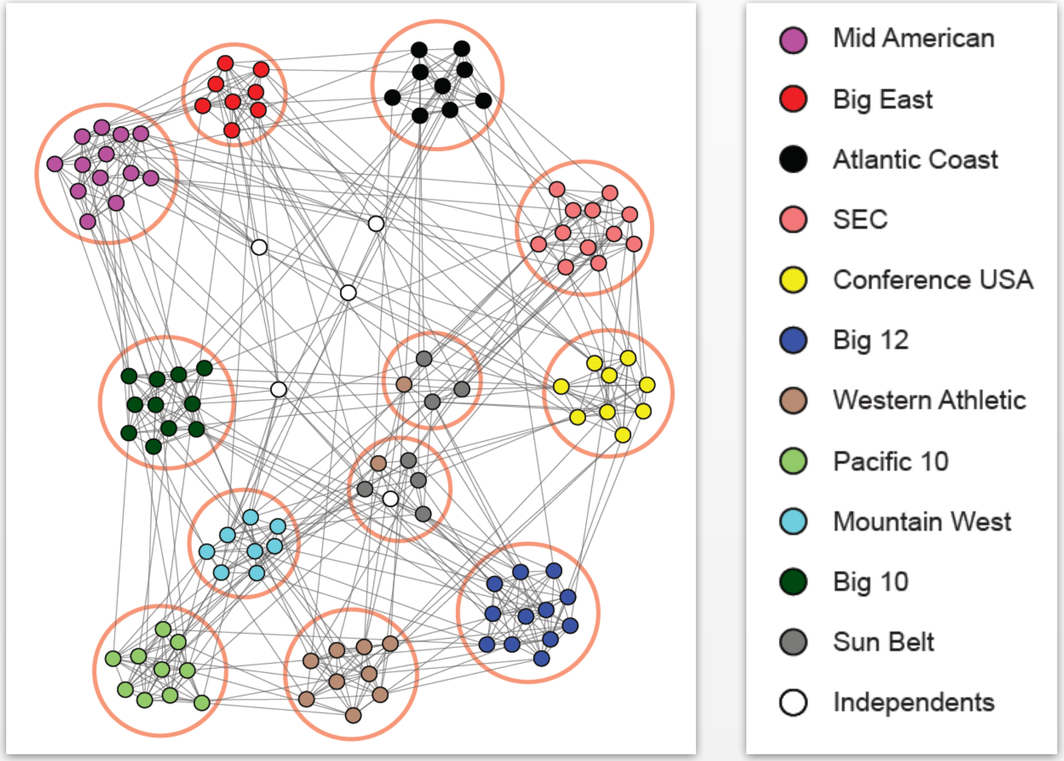
Community Detection

Goal: Identify groups of connected nodes
(i.e. clustering on graphs)



Community Detection

Goal: Identify groups of connected nodes
(i.e. clustering on graphs)

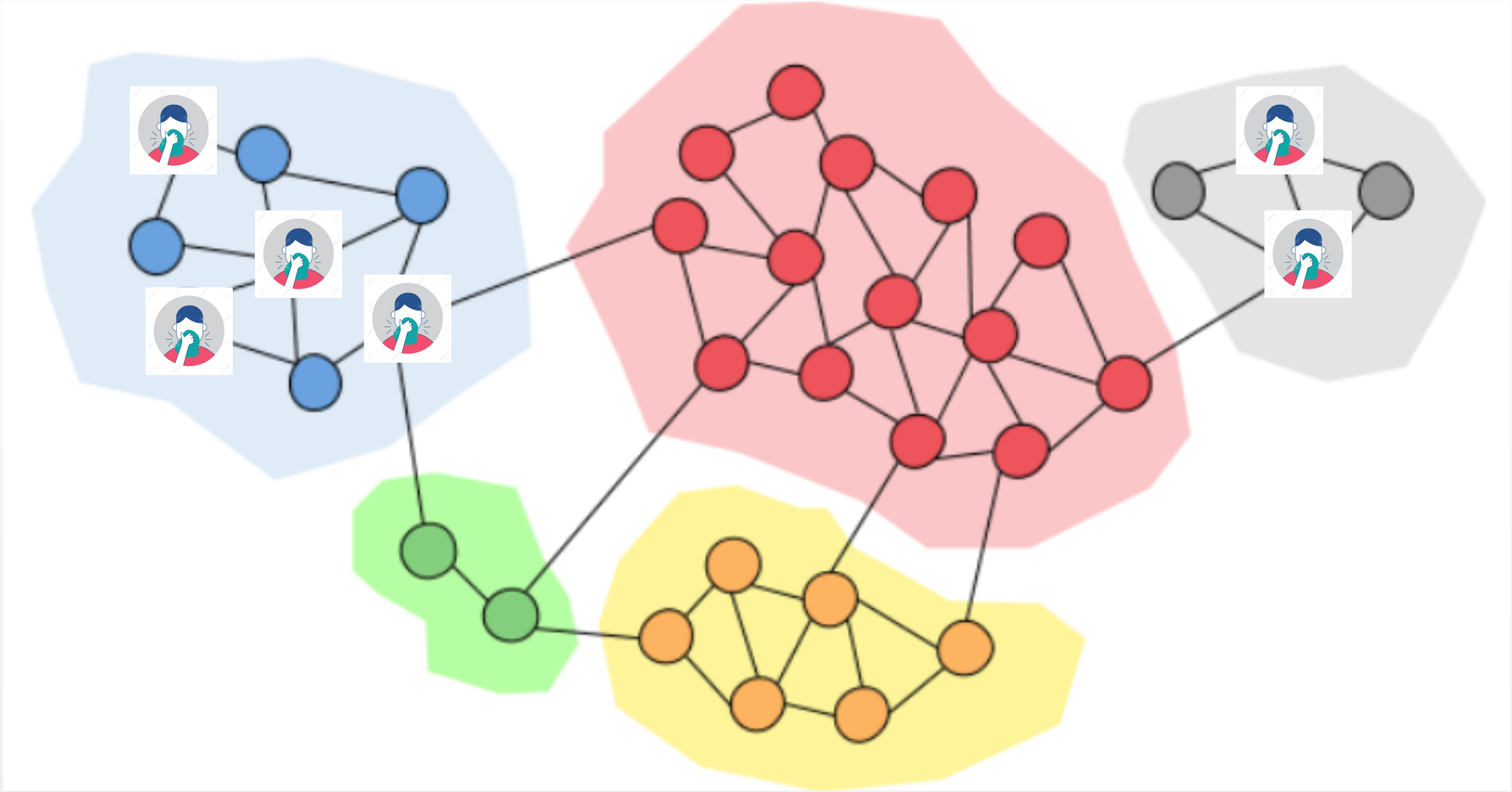


Nodes: Football Teams, *Edges:* Matches,
Communities: Conferences

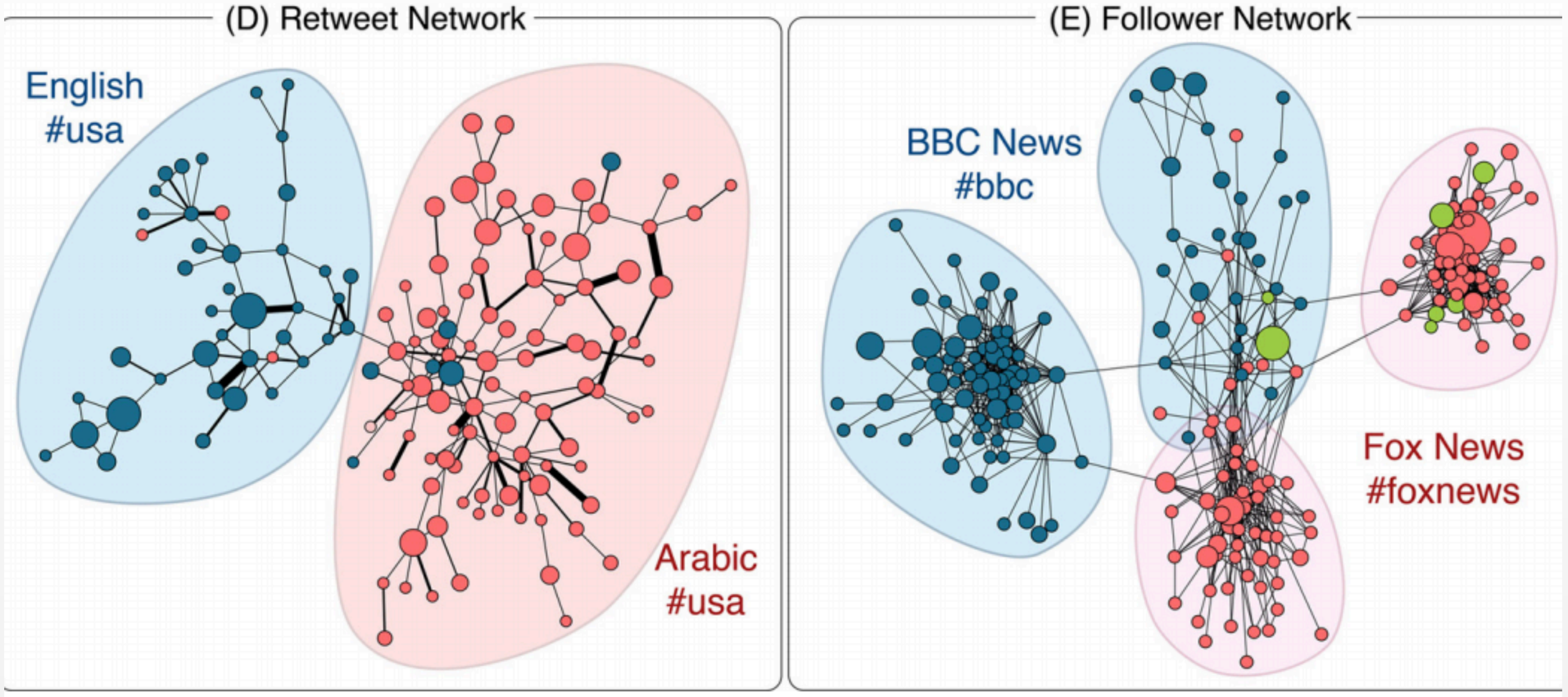
Market segmentation

Detect Fake Users

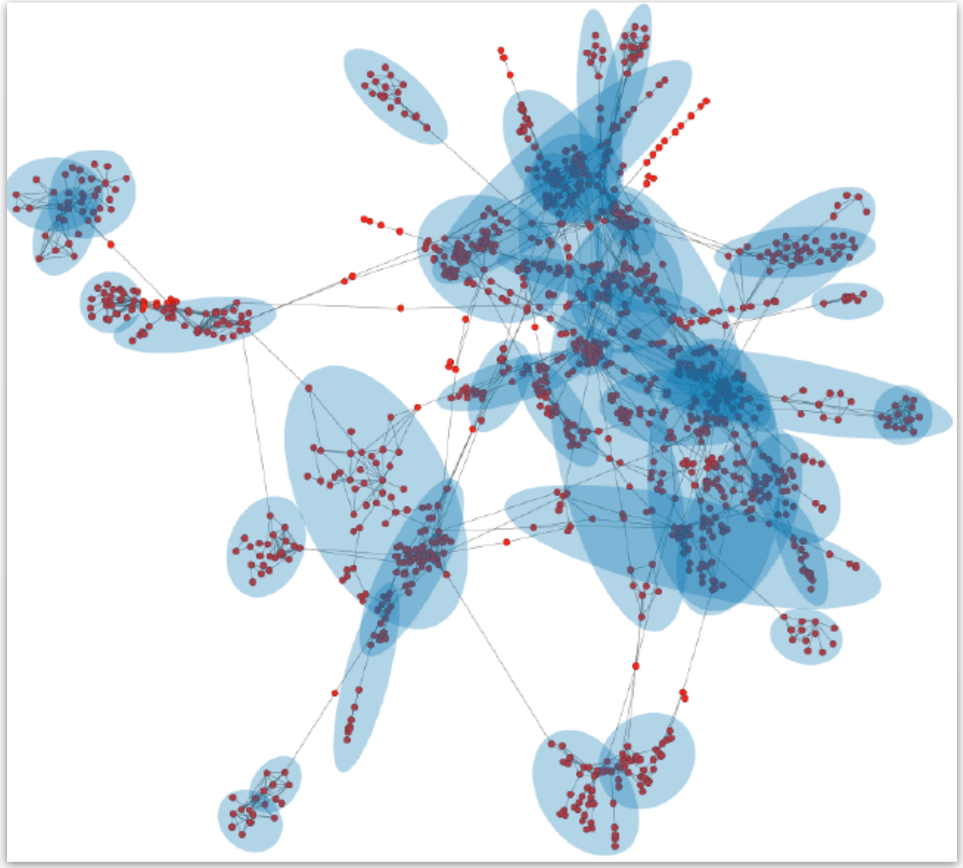
Epidemiology



Social Contagion



Communities: Protein-Protein Interactions

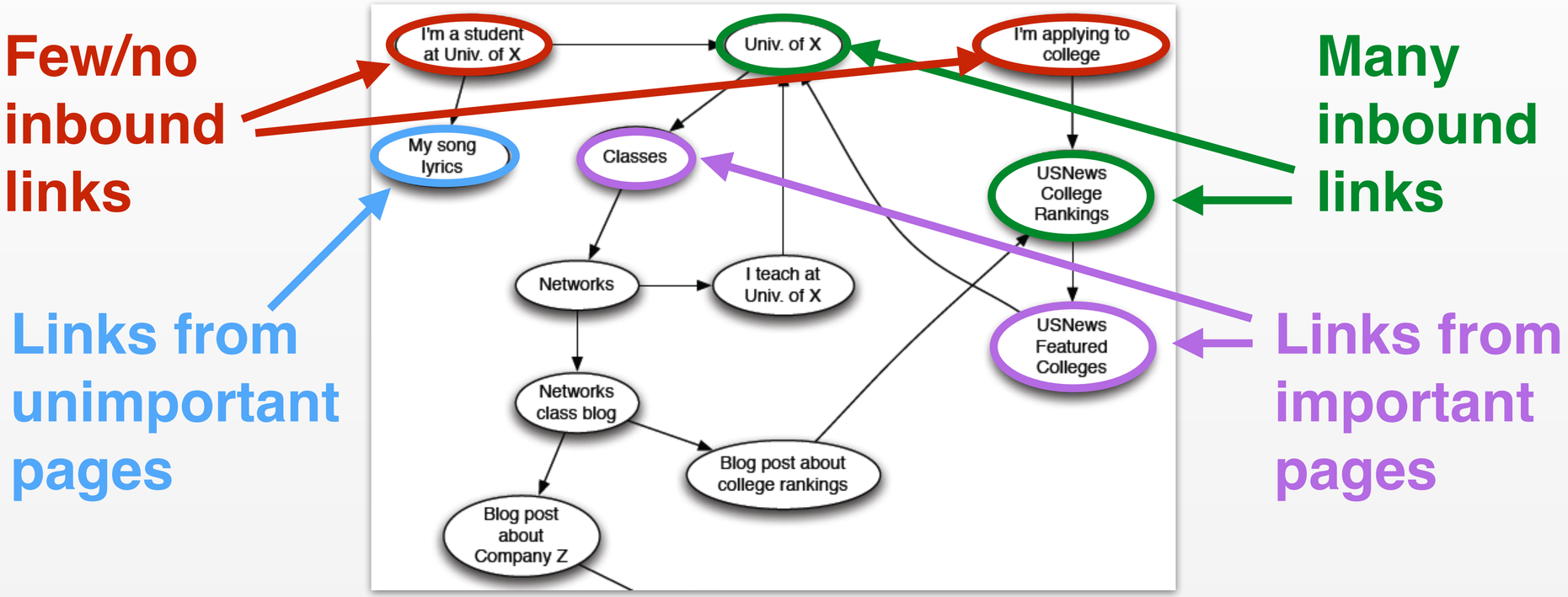


Nodes: Proteins, *Edges:* Physical interactions,
Communities: Functional Modules

(Adapted from: Mining of Massive Datasets, http://www.mmds.org)

Link Analysis

Goal: Predict which website is the most authoritative.



- Pages with **more inbound links** are more **important**
- Inbound **links from important pages** carry **more weight**

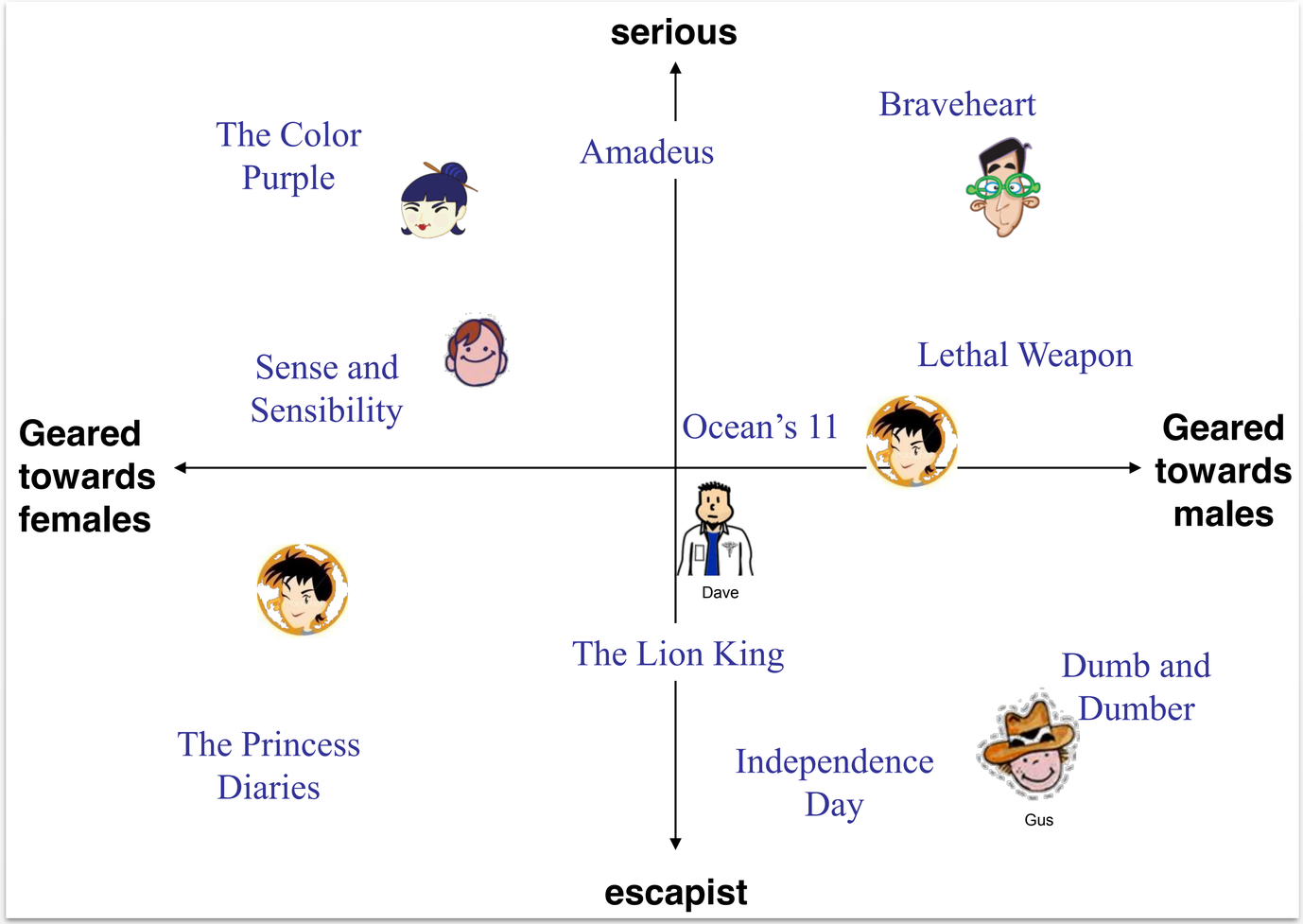
(adapted from:: Mining of Massive Datasets, http://www.mmds.org)

Recommender Systems

Movie	Alice (1)	Bob (2)	Carol (3)	Dave (4)
Love at last	5	5	0	0
Romance forever	5	?	?	0
Cute puppies of love	?	4	0	?
Nonstop car chases	0	0	5	4
Swords vs. karate	0	0	5	?

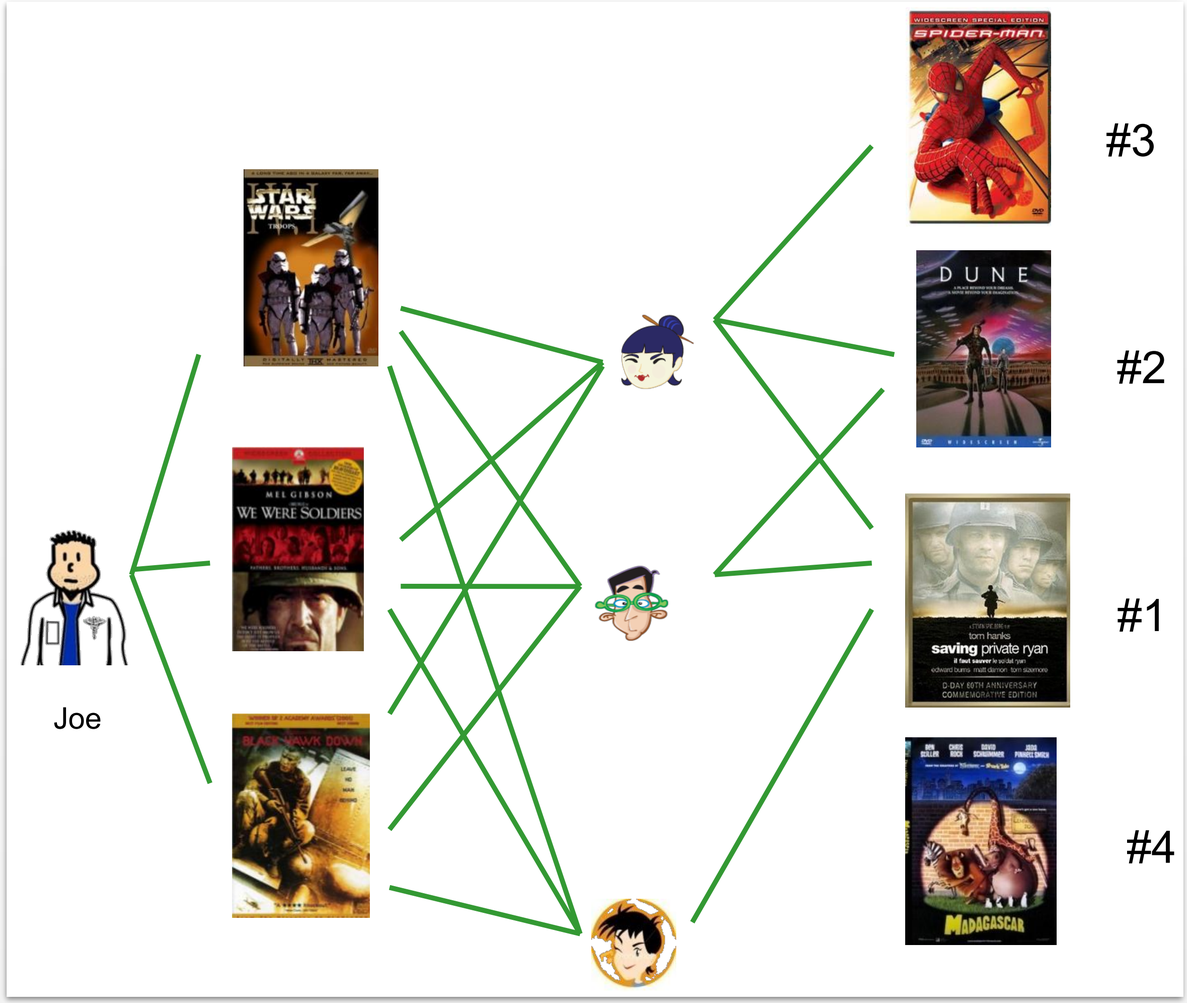
- *Task*: Predict user preferences for unseen items
- *Content-based filtering*: Model user/item features
- *Collaborative filtering*: Implicit similarity of users items

Content-based Filtering



Idea: Predict rating using ***user*** and ***item*** features

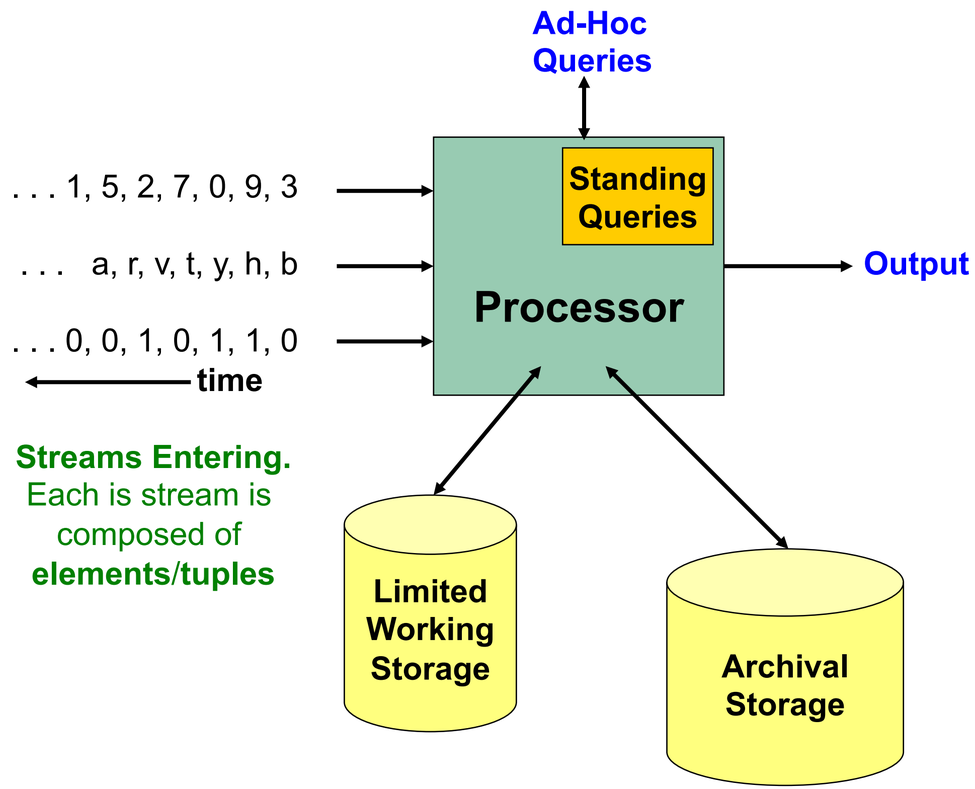
Collaborative Filtering



Idea: Predict rating based on similarity to other users

Mining on Data Streams

- Input elements enter at a rapid rate, at one or more input ports (i.e., streams).
  - Google queries
  - Twitter or Facebook status updates
- The system cannot store the entire stream
- Q: How do you make calculations about the stream using a limited amount of (secondary) memory?



Queries on Stream

- Common Types of Queries:
  - Sampling data from a stream
    - Construct a random sample
  - Queries over sliding windows
    - Number of items of type ***x*** in the last ***k*** elements of the stream
  - Filtering a data stream
    - Select elements with property ***x*** from the stream
  - Counting distinct elements
    - Number of distinct elements in last ***k*** elements of the stream
  - Estimating moments
    - Estimating frequency/surprise

Applications

- Mining query streams
  - Google wants to know what queries are more frequent today than yesterday
- Mining click streams
  - Amazon wants to know which products are getting more clicks this week.
- Mining social network news feeds
  - E.g., look for trending topics on Twitter, Facebook

Applications

- Sensor Data
  - Million temperature sensors deployed in the ocean. Keeping track of average temperature in past 24 hours or past month.
- Telephone call records
  - Compute talk time used in the current billing cycle. Monitor network usage for settlements with other wireless networks.
- IP packets monitored at a switch
  - Gather information for optimal routing
  - Detect denial-of-service attacks

Course Objectives

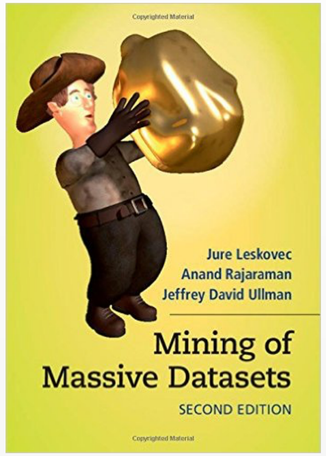
1. Lectures: Understand data mining methods

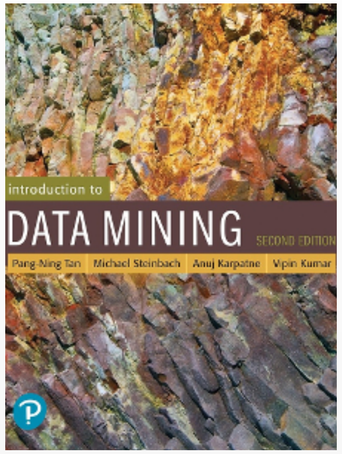
- Mathematical/algorithmic definitions
- When should each method be used?
- What are some limitations of each method?

2. Homework Problems: Use data mining methods

- Implement methods
- Use methods in existing libraries
- Visualize results, evaluate effectiveness
- Prove mathematical results

Textbooks

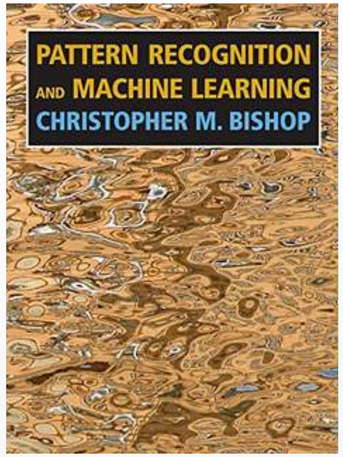




Leskovec, Rajaraman, Ullman

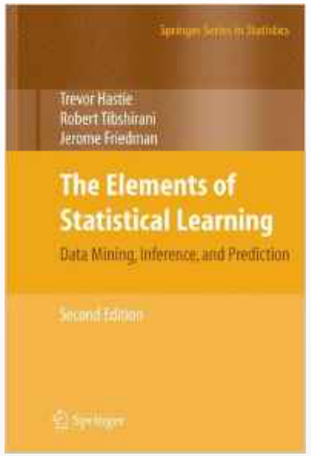
Tan, Steinbach, Kumar

Data Mining



Bishop

Machine Learning



Hastie, Tribshirani, Friedman

Statistics

Grading

- Homework: **40%**
- Midterm: **25%**
- Course Project: **30%**
- Class Participation: **5%**

Homework Problems

- 4 problem sets, 40% of grade.
- Python encouraged, but can use any language (within reason – TA must be able to run your code)
- **Discussion is encouraged, but submissions must be completed <u>individually</u>** (absolutely **no** sharing of code)
- Submission via one **pdf** file and one **zip** file on Canvas by **11.59pm** on day of deadline.
- Penalty for late submission.

# Late days	1	2	3	4	5	6
Penalty Factor	0.9	0.75	0.6	0.4	0.2	0

- Please follow *submission guidelines.*

Project Goals

- Select a dataset / and the data mining task(s).
- Perform exploratory analysis and pre-processsing
- Think of a reasonable baseline method
- Apply one or more algorithms (you can use libraries)
- Critically evaluate results/ visualization
- Submit a report and present project

Project Deadlines

- **10 June**: Form teams of 2-3 people
- **20 June**: Submit abstract (1 paragraph)
- **10 July**: Milestone 1 (exploratory analysis)
- **05 Aug**: Milestone 2 (statistical analysis)
- **12 Aug**: Project Presentations, Draft Report
- **20 Aug**: Final Report (8 pages maximum for the main text. A supplement can be added with more figures and tables).

Class Participation

1. Ask questions during the lecture and the office hours
2. Help others on Piazza.

Class participation is used to adjust grade upwards
(at the discretion of the instructor)

Self-evaluation

After Midterm and Last week of class

- What was your favorite topic?
- What parts were easy / difficult ?
- *List 3 students that contributed to your understanding of the material*

Office hours

Shantanu: ***Tue/Fri: 5:30 - 7:00 pm***

Alina: ***Mon/Thur: 9:00 - 10:30 am***