CS249: ADVANCED DATA MINING

1: Introduction

Instructor: Yizhou Sun

yzsun@cs.ucla.edu

(Instructor for Today's class: Ting Chen)

April 9, 2017

Course Information

- Course homepage: http://web.cs.ucla.edu/~yzsun/classes/2017Spring_CS249/index.html
  - Class Schedule
  - Slides
  - Announcement
  - Assignments
  - …

- Prerequisites
  - You are expected to have background knowledge in data structures, algorithms, basic linear algebra, and basic statistics.
  - You will also need to be familiar with at least one programming language, and have programming experiences.

Meeting Time and Location

- When
  - M&W, 2-4pm
- Where
  - 5436 Boelter Hall

Instructor and TA Information

- Instructor: Yizhou Sun
  - Homepage: http://web.cs.ucla.edu/~yzsun/
  - Email: yzsun@cs.ucla.edu
  - Office: 3531E
  - Office hour: W&M 4-5pm

- TA: Jae LEE
  - Email: jlee734@ucla.edu
  - Office hours: TBD

Grading

- Homework: 30%
- Midterm exam: 30%
- Course project: 35%
- Participation: 5%

Grading: Homework

- Homework: 30%
  - 3 assignments are expected
  - Deadline: 11:59pm of the indicated due date via *ccle* system
    - *Late submission policy:* get original score* $\mathbf{1}(t <= 24)e^{-(ln(2)/12)*t}$ if you are t hours late.
  - No copying or sharing of homework!
    - But you can discuss general challenges and ideas with others
    - *Suspicious cases will be reported to The Office of the Dean of Students*

Grading: Midterm Exam

- Midterm exam: 30%
  - Closed book exam, but you can take a "reference sheet" of A4 size

Grading: Course Project

- Course project: 35%
  - Group project (3-4 people for one group)
  - Goal: Solve an open data mining problem
  - You are expected to present your project to the class, and submit a project report and your code at the end of the quarter

Grading: Participation

- Participation (5%)
  - In-class participation
  - Quizzes
  - Online participation (piazza)
    - piazza.com/ucla/spring2017/comsci249/home

Textbook

- Recommended: Jiawei Han, Micheline Kamber, and Jian Pei. Data Mining: Concepts and Techniques, 3rd edition, Morgan Kaufmann, 2011
- References
  - "Data Mining: The Textbook" by Charu Aggarwal (http://www.charuaggarwal.net/Data-Mining.htm)
  - "Data Mining" by Pang-Ning Tan, Michael Steinbach, and Vipin Kumar (http://www-users.cs.umn.edu/~kumar/dmbook/index.php)
  - "Machine Learning" by Tom Mitchell (http://www.cs.cmu.edu/~tom/mlbook.html)
  - "Introduction to Machine Learning" by Ethem ALPAYDIN (http://www.cmpe.boun.edu.tr/~ethem/i2ml/)
  - "Pattern Classification" by Richard O. Duda, Peter E. Hart, David G. Stork (http://www.wiley.com/WileyCDA/WileyTitle/productCd-0471056693.html)
  - "The Elements of Statistical Learning: Data Mining, Inference, and Prediction" by Trevor Hastie, Robert Tibshirani, and Jerome Friedman (http://www-stat.stanford.edu/~tibs/ElemStatLearn/)
  - "Pattern Recognition and Machine Learning" by Christopher M. Bishop (http://research.microsoft.com/en-us/um/people/cmbishop/prml/)

Goals of the Course

- Know what data mining is and learn the basic algorithms
- Know how to apply algorithms to real-world applications
- Provide a starting course for research in data mining

1. Introduction

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
  - What Kinds of Data Can Be Mined?
  - What Kinds of Patterns Can Be Mined?
  - What Kinds of Technologies Are Used?
  - What Kinds of Applications Are Targeted?
- Content covered by this course

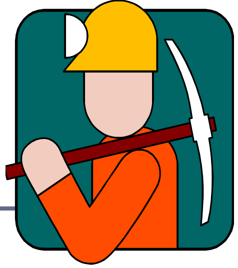
Why Data Mining?

- The Explosive Growth of Data: from terabytes to petabytes
  - Data collection and data availability
    - Automated data collection tools, database systems, Web, computerized society
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Society and everyone: news, digital cameras, YouTube
- <u>We are drowning in data, but starving for knowledge!</u>
- "Necessity is the mother of invention"—Data mining—Automated analysis of massive data sets

1. Introduction

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
  - What Kinds of Data Can Be Mined?
  - What Kinds of Patterns Can Be Mined?
  - What Kinds of Technologies Are Used?
  - What Kinds of Applications Are Targeted?
- Content covered by this course

What Is Data Mining?

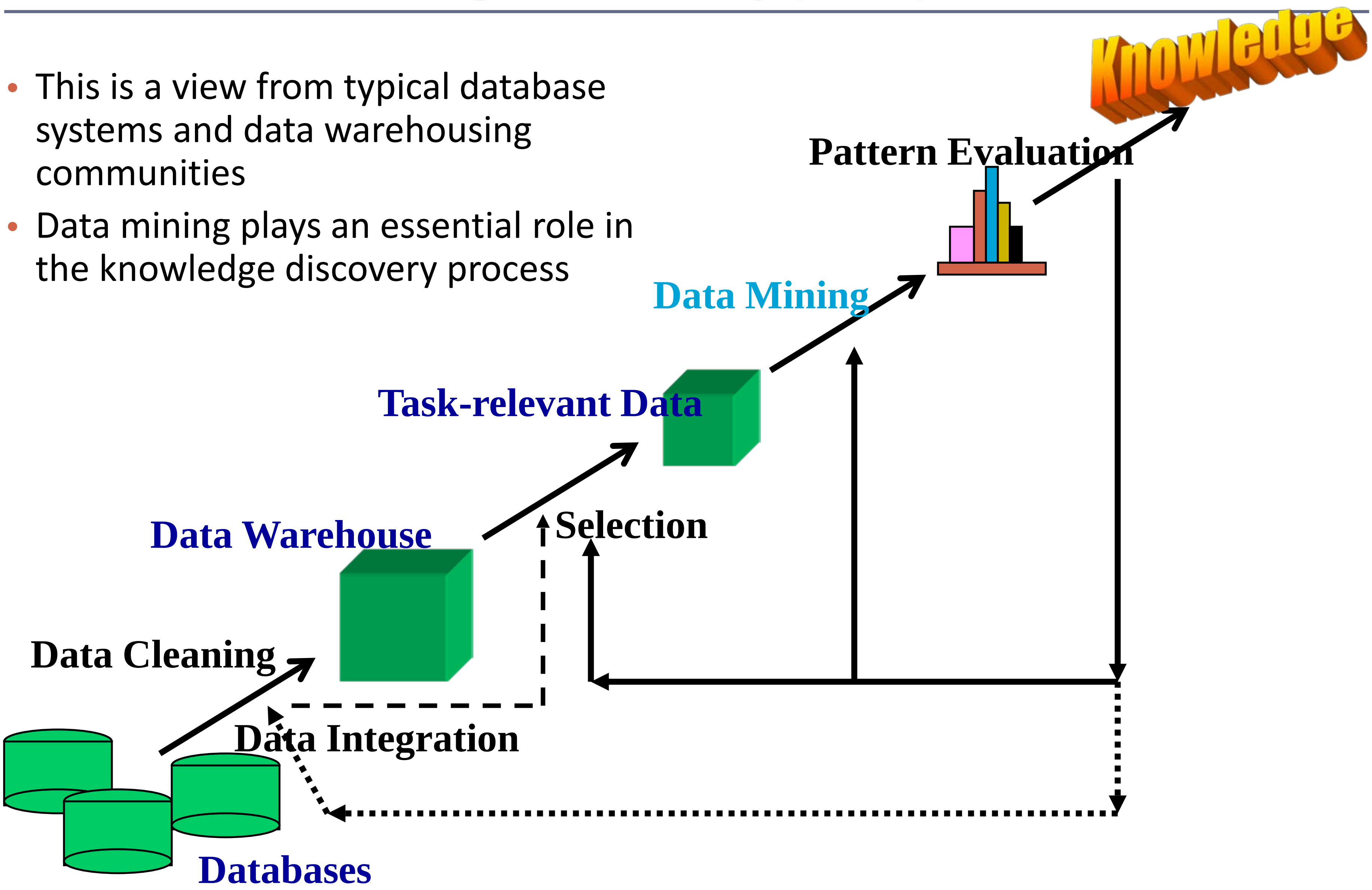


- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data

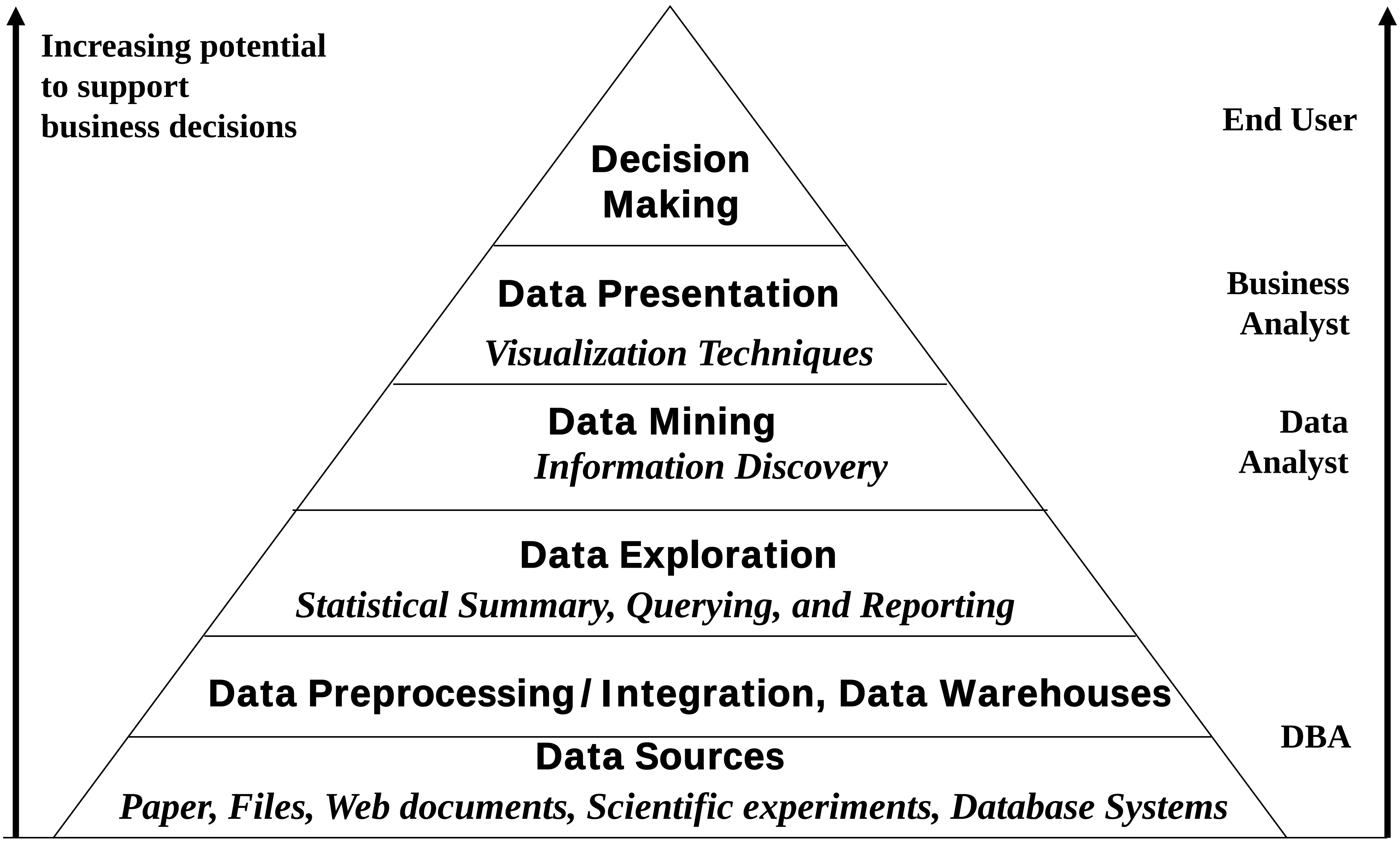
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

Knowledge Discovery (KDD) Process

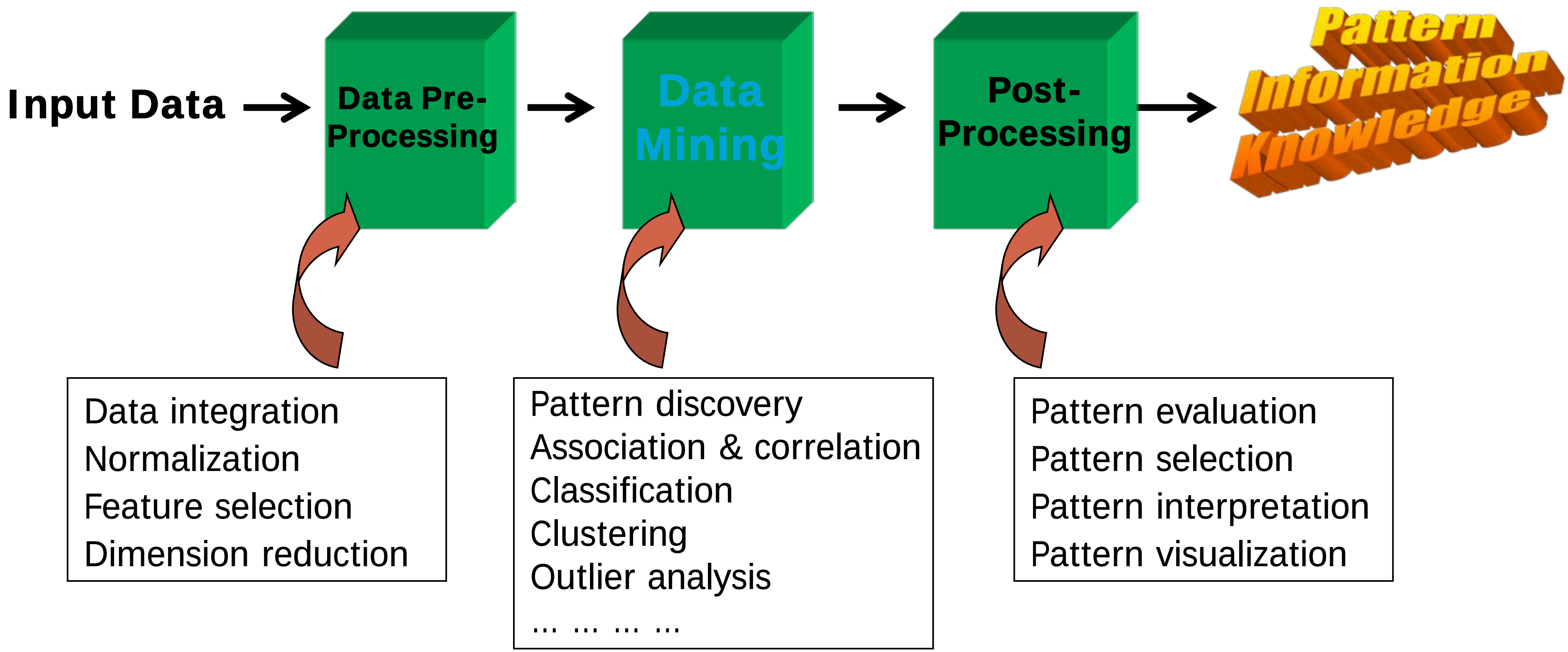
- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process



Data Mining in Business Intelligence



KDD Process: A Typical View from ML and Statistics



- This is a view from typical machine learning and statistics communities

1. Introduction

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
  - What Kinds of Data Can Be Mined?
  - What Kinds of Patterns Can Be Mined?
  - What Kinds of Technologies Are Used?
  - What Kinds of Applications Are Targeted?
- Content covered by this course

Multi-Dimensional View of Data Mining

- **Data to be mined**
  - Database data (extended-relational, object-oriented, heterogeneous, legacy), data warehouse, transactional data, stream, spatiotemporal, time-series, sequence, text and web, multi-media, graphs & social and information networks
- **Knowledge to be mined (or: Data mining functions)**
  - Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
  - Descriptive vs. predictive data mining
  - Multiple/integrated functions and mining at multiple levels
- **Techniques utilized**
  - Data-intensive, data warehouse (OLAP), machine learning, statistics, pattern recognition, visualization, high-performance, etc.
- **Applications adapted**
  - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

1. Introduction

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
  - What Kinds of Data Can Be Mined?
  - What Kinds of Patterns Can Be Mined?
  - What Kinds of Technologies Are Used?
  - What Kinds of Applications Are Targeted?
- Content covered by this course

Matrix Data

	Sex	Race	Height	Income	Marital Status	Years of Educ.	Liberal-ness
R1001	M	1	70	50	1	12	1.73
R1002	M	2	72	100	2	20	4.53
R1003	F	1	55	250	1	16	2.99
R1004	M	2	65	20	2	16	1.13
R1005	F	1	60	10	3	12	3.81
R1006	M	1	68	30	1	9	4.76
R1007	F	5	66	25	2	21	2.01
R1008	F	4	61	43	1	18	1.27
R1009	M	1	69	67	1	12	3.25

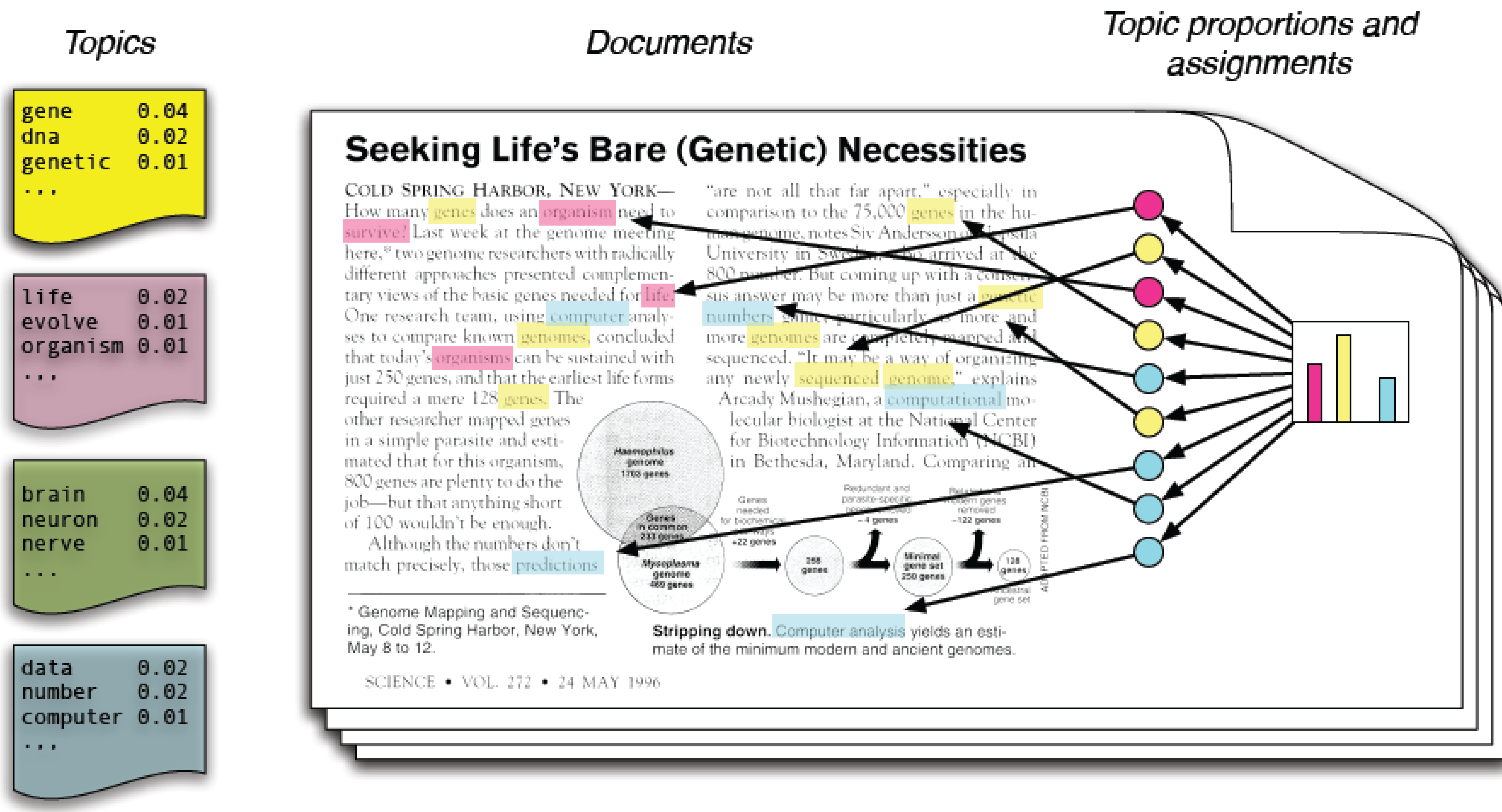
Set Data

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

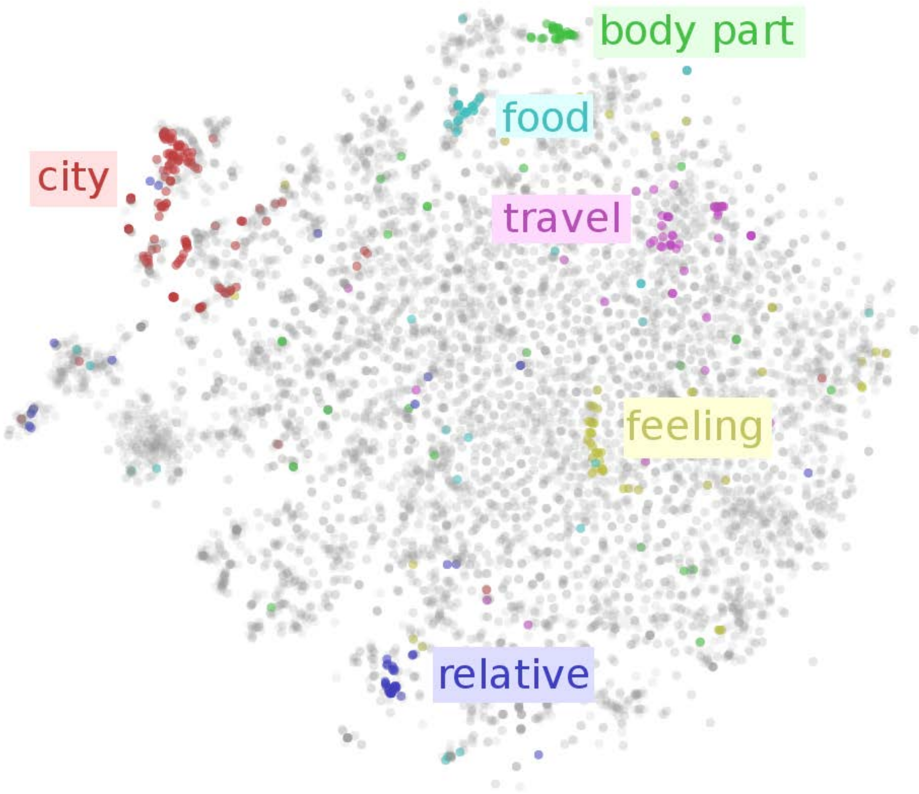
Text Data

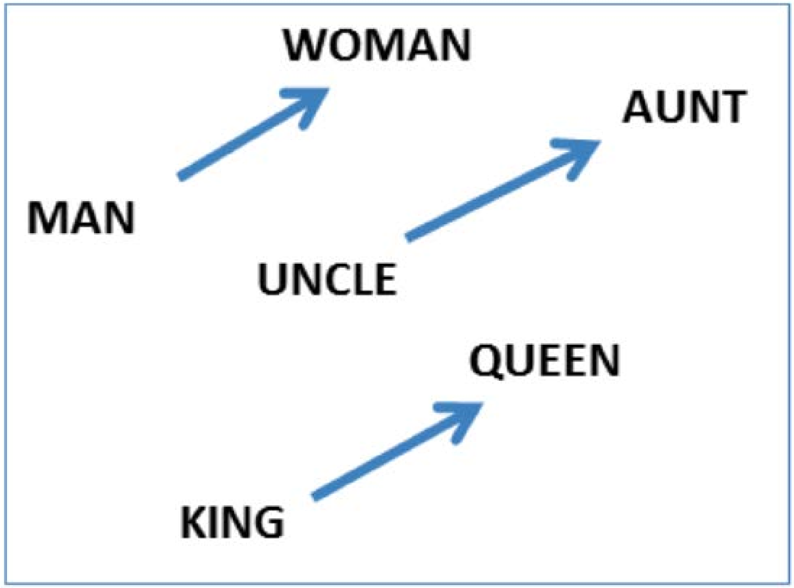
- “Text mining, also referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling (i.e., learning relations between named entities).” –from wiki

Text Data – Topic Modeling



Text Data – Word Embedding





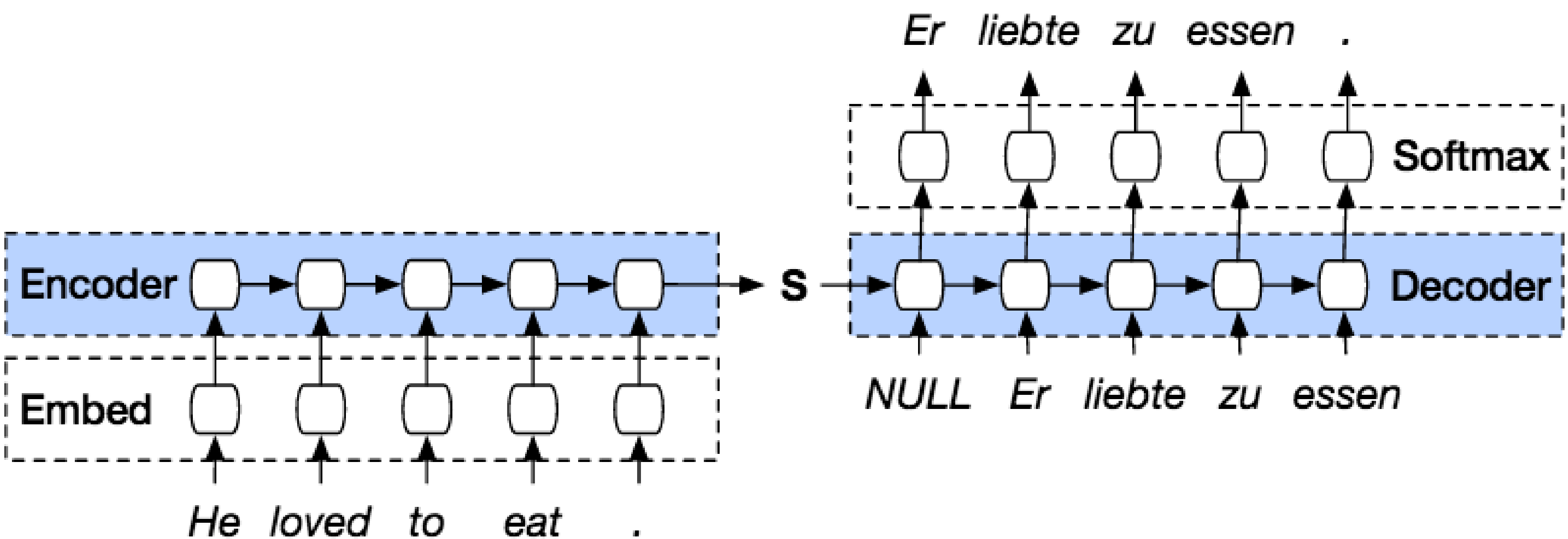
king - man + woman = queen

Sequence Data

SYNTENIC ASSEMBLIES FOR CG15386

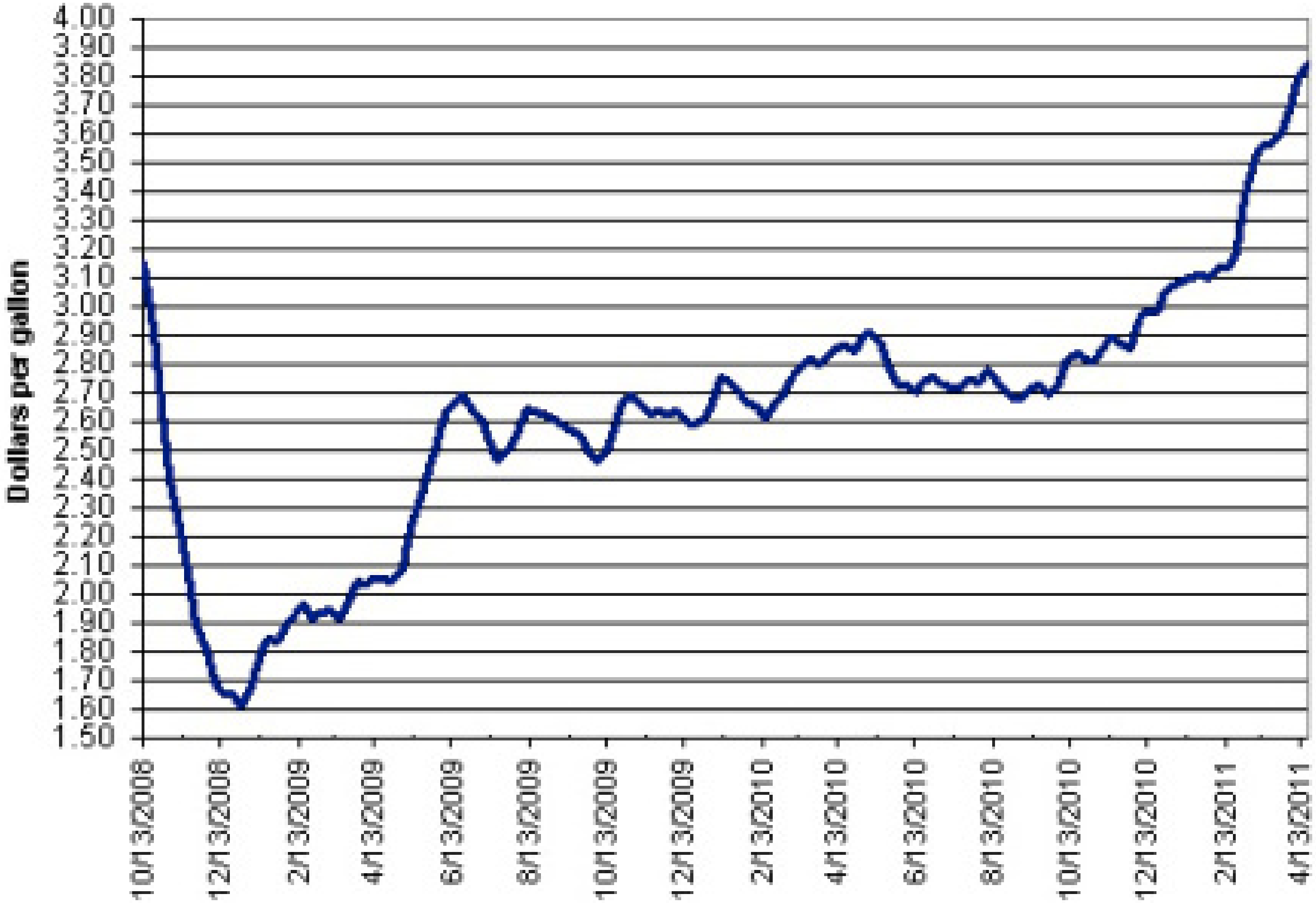
```
MD106   ATGCTTAGTAATCCCTACTTTAAGTCCGTTTTGTGGCTGATTGGCTTCGGAGGAATGGG
NEWC    ATGCTTAGTAATCCTTACTTTAAATCCGTTTTGTGGCTGATTGGCTTCGGAGGAATGGG
W501    ATGCTTAGTAATCCCTACTTTAAGTCCGTTTTGTGGCTGATTGGCTTCGGAGGAATGGG
MD199   ATGCTTAGTAATCCCTACTTTAAGTCCGTTTTGTGGCTGATTGGCTTCGGAGGAATGGG
C1674   ATGCTTAGTAATCCCTACTTTAAGTCCGTTTTGTGGCTGATTGGCTTCGGAGGAATGGG
SIM4    ATGCTTAGTAATCCCTACTTTAAGTCCGTTTTGTGGCTGATTGGCTTCGGAGGAATGGG

MD106   CTACGGCCTAATGGTGCTAACAGAGCCGAACGTCGACAAAATAGAGCGCATCAAAGCCT
NEWC    CTACGGCCTAATGGTGCTAACCGAGCCGAACGTCGACAAAATAGAGCGCATCAAAGCCT
W501    CTACGGCCTAATGGTGCTAACCGAGCCGAACGTCGACAAAATAGAGCGCATCAAAGCCT
MD199   CTACGGCCTAATGGTGCTAACCGAGCCGAACGTCGACAAAATAGAGCGCATCAAAGCCT
C1674   CTACGGCCTAATGGTGCTAACCGAGCCGAACGTCGACAAAATAGAGCGCATCAAAGCCT
SIM4    CTACGGCCTAATGGTGCTAACCGAGCCGAACGTCGACAAAATAGAGCGCATCAAAGCCT

MD106   CCGTTTCAAGTACCAAACTGAGTGCGGATGAGCAGCGAAAGGCTCTGTTTATGAAGAAG
NEWC    CCGTTTCAAGTACCAAACTGAGTGCGGATGAGCAGCGAAAGGCTCTGTTTATGAAGAAG
W501    CCGTTTCAAGTACCAAACTGAGTGCGGATGAGCAGCGAAAGGCTCTGTTTATGAAGAAG
MD199   CCGTTTCAAGTACCAAACTGAGTGCGGATGAGCAGCGAAAGGCTCTGTTTATGAAGAAG
C1674   CCGTTTCAAGTACCAAACTGAGTGCGGATGAGCAGCGAAAGGCTCTGTTTATGAAGAAG
SIM4    CCGTTTCAAGTACCAAACTGAGTGCGGATGAGCAGCGAAAGGCTCTGTTTATGAAGAAG

MD106   CTGCAGGAGGCGTCCACCACCAGTGCCCCAATCTACAGGTCAGCGGCCGAGAAATAG
NEWC    CTGCAGGAGGCGTCCACCACCAGTGCCCCAATCTACAGGTCATCGGCCGAGAAATAG
W501    CTGCAGGAGGCGTCCACCACCACTGCCCCAATCTACAGGTCATCGGCCGAGAAATAG
MD199   CTGCAGGAGGCGTCCACCACCAGTGCCCCAATCTACAGGTCAGCGGCCGAGAAATAG
C1674   CTGCAGGAGGCGTCCACCACCAGTGCCCCAATCTACAGGTCAGCGGCCGAGAAATAG
SIM4    CTGCAGGAGGCGTCCACCACCAGTGCCCCAATCTACAGGTCAGCGGCCGAGAAATAG
```

Sequence Data – Seq2Seq



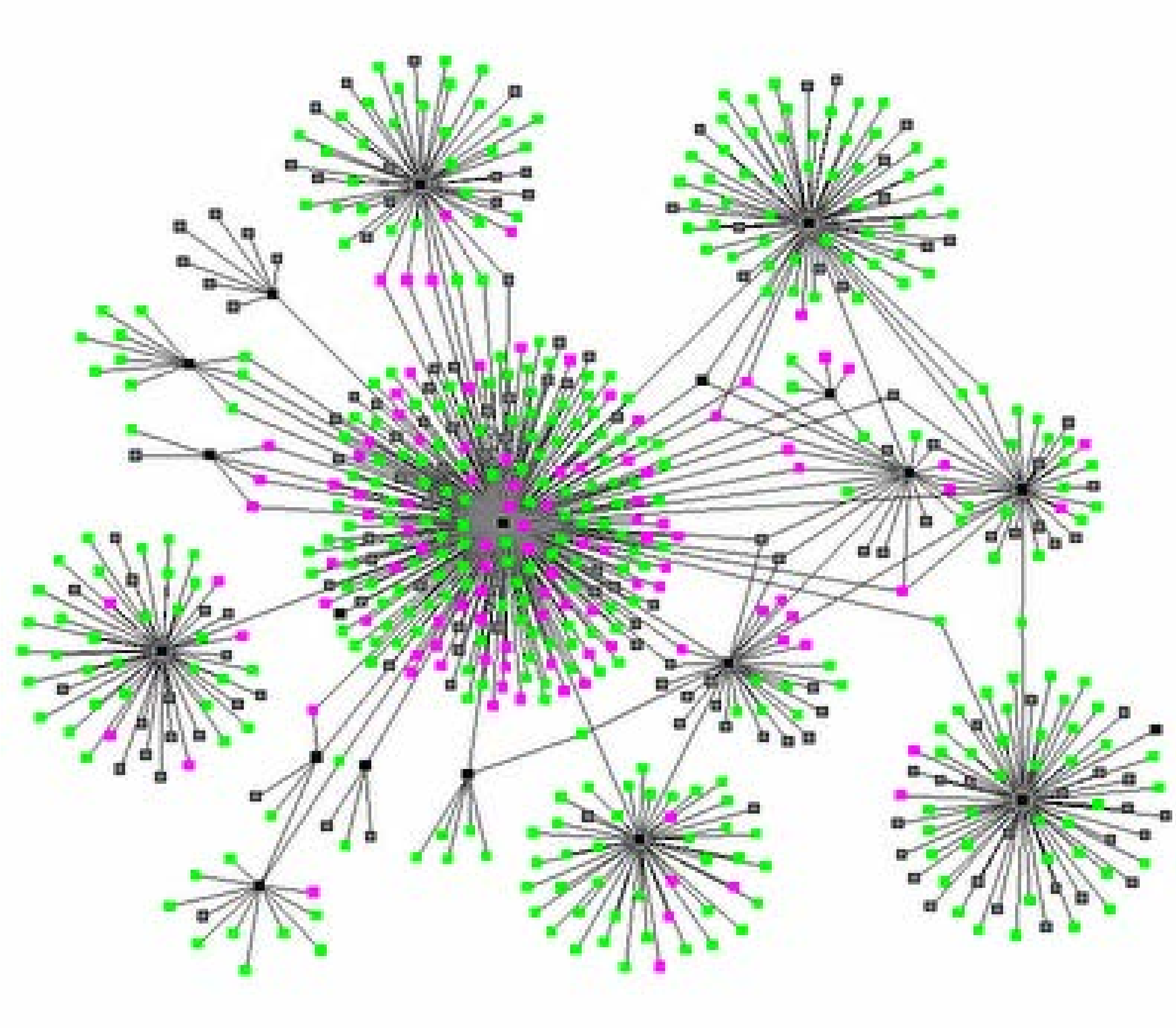
Time Series

Weekly U.S. Retail Gasoline Prices, Regular Grade

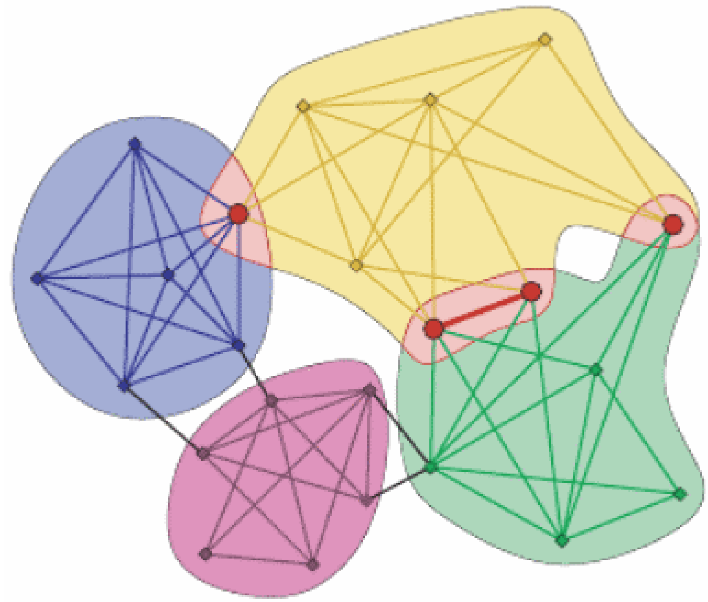


Source: Energy Information Administration

Graph / Network



Graph / Network – Community Detection



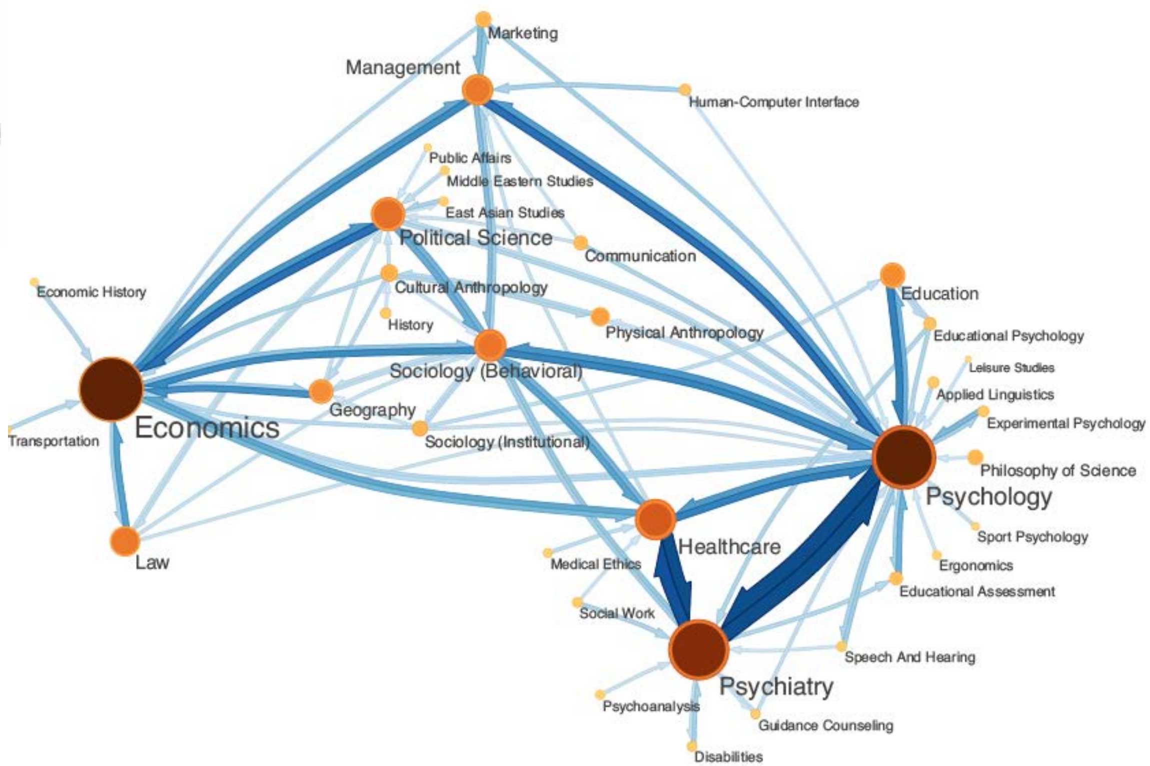


Image Data

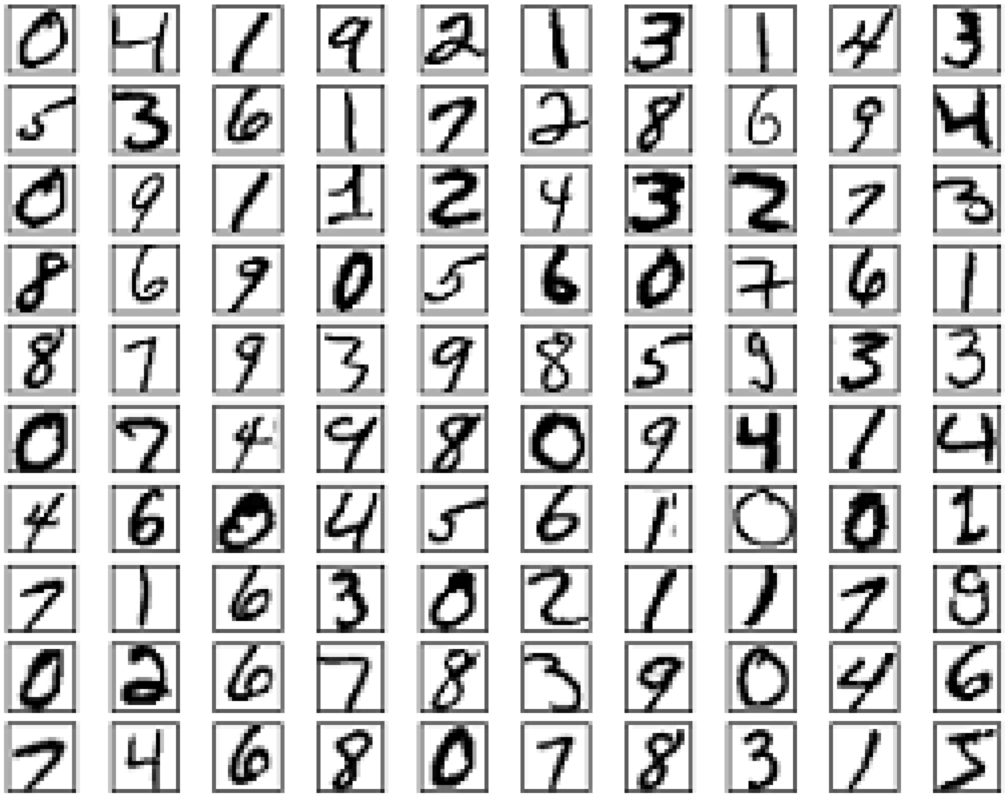


Image Data – Neural Style Transfer

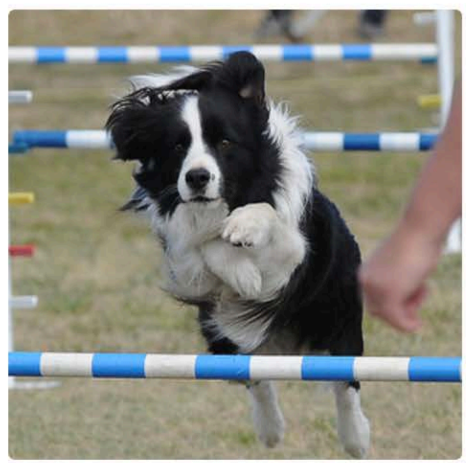
Image Data – Image Captioning

"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

"girl in pink dress is jumping in air."



"black and white dog jumps over bar."

"young girl in pink shirt is swinging on swing."

1. Introduction

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
  - What Kinds of Data Can Be Mined?
  - What Kinds of Patterns Can Be Mined?
  - What Kinds of Technologies Are Used?
  - What Kinds of Applications Are Targeted?
- Content covered by this course

Data Mining Function: Association and Correlation Analysis

- Frequent patterns (or frequent itemsets)
  - What items are frequently purchased together in your Walmart?
- Association, correlation vs. causality
  - A typical association rule
    - Diaper → Beer [0.5%, 75%] (support, confidence)

Data Mining Function: Classification

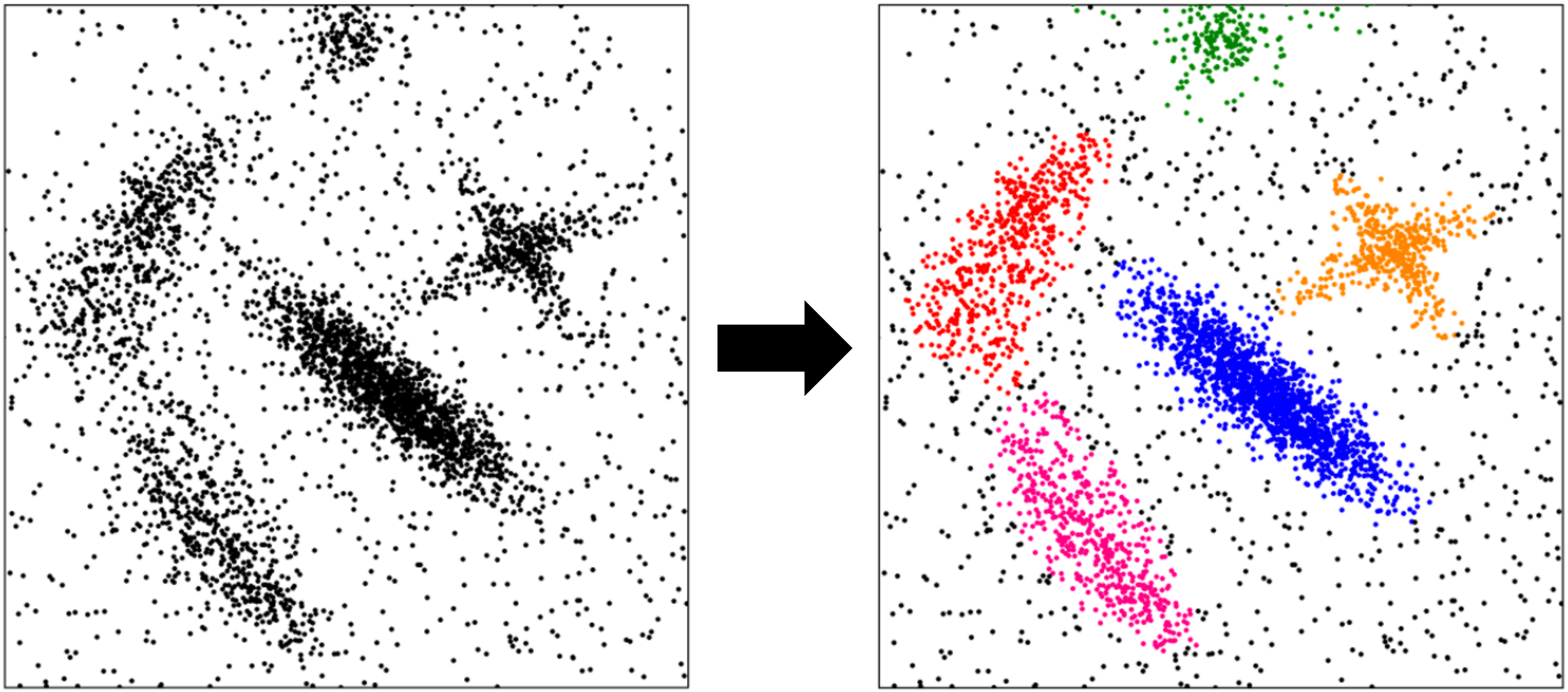
- Classification and label prediction
  - Construct models (functions) based on some training examples
  - Describe and distinguish classes or concepts for future prediction
    - E.g., classify countries based on (climate), or classify cars based on (gas mileage)
  - Predict some unknown class labels
- Typical methods
  - Decision trees, naïve Bayesian classification, support vector machines, neural networks, rule-based classification, pattern-based classification, logistic regression, ...
- Typical applications:
  - Credit card fraud detection, direct marketing, classifying stars, diseases, web-pages, ...

Image Classification Example

Data Mining Function: Cluster Analysis

- Unsupervised learning (i.e., Class label is unknown)
- Group data to form new categories (i.e., clusters), e.g., cluster houses to find distribution patterns
- Principle: Maximizing intra-class similarity & minimizing interclass similarity
- Many methods and applications

Clustering Example



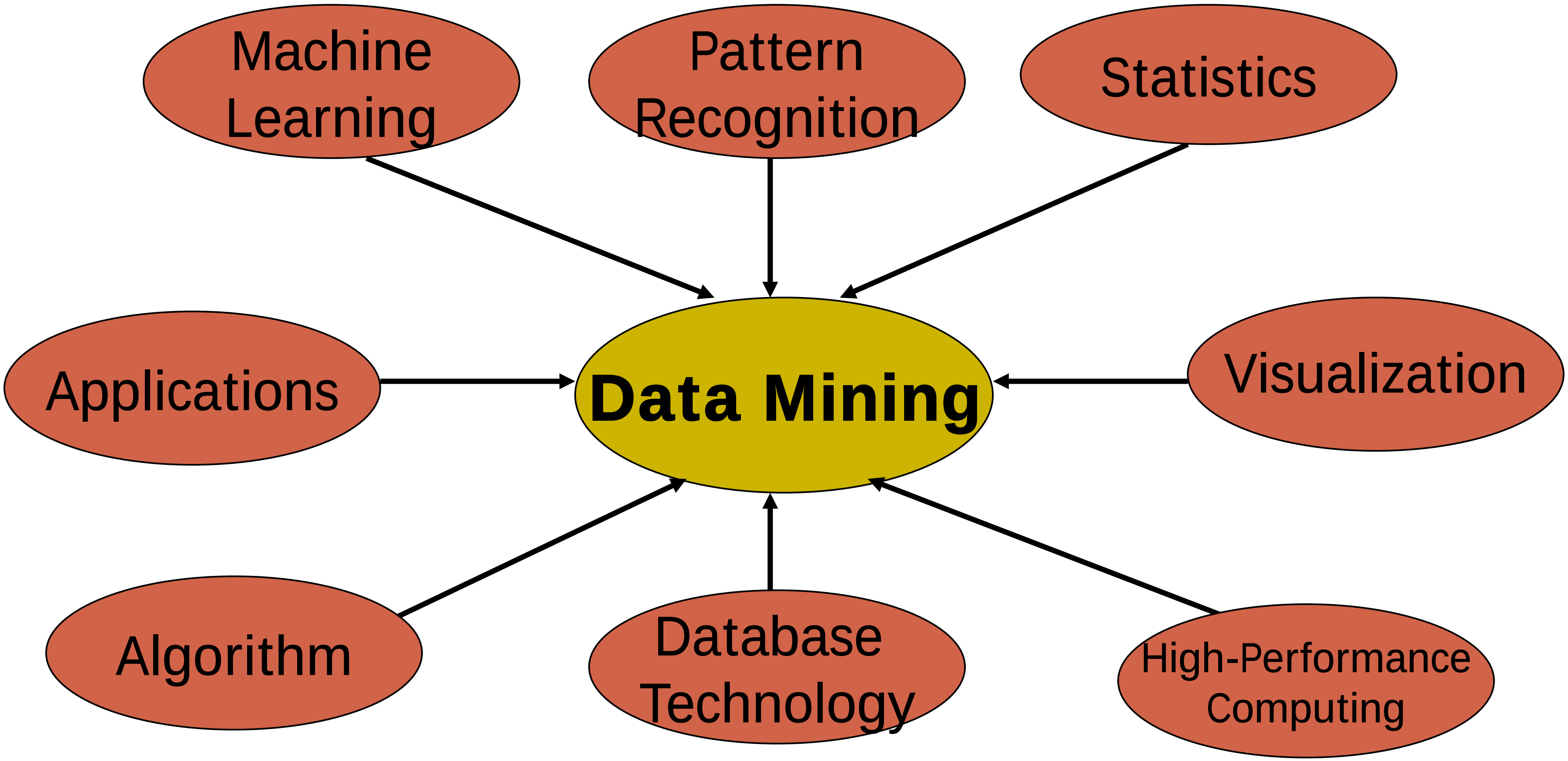
Data Mining Functions: Others

- Prediction
- Similarity search
- Ranking
- Outlier detection
- …

1. Introduction

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
  - What Kinds of Data Can Be Mined?
  - What Kinds of Patterns Can Be Mined?
  - What Kinds of Technologies Are Used?
  - What Kinds of Applications Are Targeted?
- Content covered by this course

Data Mining: Confluence of Multiple Disciplines



1. Introduction

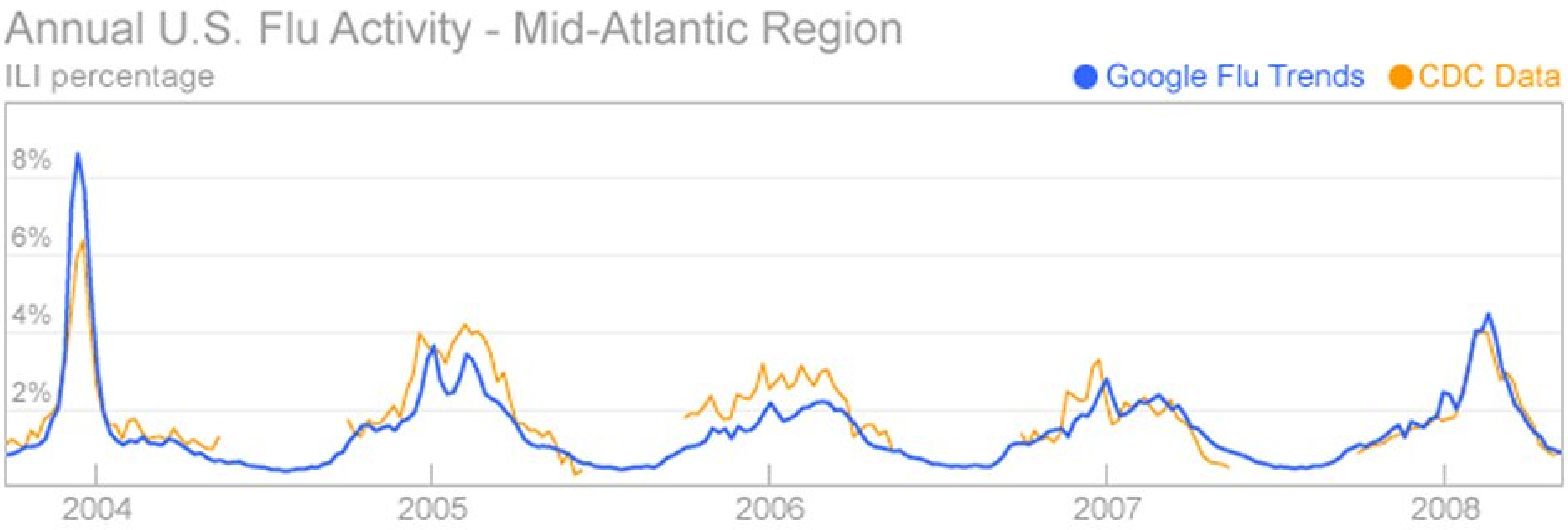
- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
  - What Kinds of Data Can Be Mined?
  - What Kinds of Patterns Can Be Mined?
  - What Kinds of Technologies Are Used?
  - What Kinds of Applications Are Targeted?
- Content covered by this course

Applications of Data Mining

- Web page analysis: from web page classification, clustering to PageRank & HITS algorithms
- Collaborative analysis & recommender systems
- Basket data analysis to targeted marketing
- Biological and medical data analysis: classification, cluster analysis (microarray data analysis), biological sequence analysis, biological network analysis
- Data mining and software engineering (e.g., IEEE Computer, Aug. 2009 issue)
- Social media
- Game

Google Flu Trends

- https://www.youtube.com/watch?v=6111nS66Dpk



NetFlix Prize

- https://www.youtube.com/watch?v=4_e2sNYYfxA

NETFLIX

Netflix Prize — COMPLETED

Home | Rules | Leaderboard | Update

Leaderboard

Showing Test Score. Click here to show quiz score

Display top 20 **leaders.**

Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos				
1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:38:22
3	Grand Prize Team	0.8582	9.90	2009-07-10 21:24:40
4	Opera Solutions and Vandelay United	0.8588	9.84	2009-07-10 01:12:31
5	Vandelay Industries !	0.8591	9.81	2009-07-10 00:32:20
6	PragmaticTheory	0.8594	9.77	2009-06-24 12:06:56
7	BellKor in BigChaos	0.8601	9.70	2009-05-13 08:14:09
8	Dace_	0.8612	9.59	2009-07-24 17:18:43
9	Feeds2	0.8622	9.48	2009-07-12 13:11:51
10	BigChaos	0.8623	9.47	2009-04-07 12:33:59
11	Opera Solutions	0.8623	9.47	2009-07-24 00:34:07
12	BellKor	0.8624	9.46	2009-07-26 17:19:11

Facebook MyPersonality App

- https://www.youtube.com/watch?v=GOZArvMMHKs

PNAS PNAS PNAS

Private traits and attributes are predictable from digital records of human behavior

Michal Kosinski[a,1], David Stillwell[a], and Thore Graepel[b]

[a]Free School Lane, The Psychometrics Centre, University of Cambridge, Cambridge CB2 3RQ United Kingdom; and [b]Microsoft Research, Cambridge CB1 2FB, United Kingdom

Edited by Kenneth Wachter, University of California, Berkeley, CA, and approved February 12, 2013 (received for review October 29, 2012)

We show that easily accessible digital records of behavior, Facebook Likes, can be used to automatically and accurately predict a range of highly sensitive personal attributes including: sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender. The analysis presented is based on a dataset of over 58,000 volunteers who provided their Facebook Likes, detailed demographic profiles, and the results of several psychometric tests. The proposed model uses dimensionality reduction for preprocessing the Likes data, which are then entered into logistic/linear regression to predict individual psychodemographic profiles from Likes. The model correctly discriminates between homosexual and heterosexual men in 88% of cases, African Americans and Caucasian Americans in 95% of cases, and between Democrat and Republican in 85% of cases. For the personality trait "Openness," prediction accuracy is close to the test–retest accuracy of a standard personality test. We give examples of associations between attributes and Likes and discuss implications for online personalization

browsing logs (11–15). Similarly, it has been shown that personality can be predicted based on the contents of personal Web sites (16), music collections (17), properties of Facebook or Twitter profiles such as the number of friends or the density of friendship networks (18–21), or language used by their users (22). Furthermore, location within a friendship network at Facebook was shown to be predictive of sexual orientation (23).

This study demonstrates the degree to which relatively basic digital records of human behavior can be used to automatically and accurately estimate a wide range of personal attributes that people would typically assume to be private. The study is based on Facebook Likes, a mechanism used by Facebook users to express their positive association with (or "Like") online content, such as photos, friends' status updates, Facebook pages of products, sports, musicians, books, restaurants, or popular Web sites. Likes represent a very generic class of digital records, similar to Web search queries, Web browsing histories, and credit card purchases. For example, observing users' Likes related to music

1. Introduction

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
  - What Kinds of Data Can Be Mined?
  - What Kinds of Patterns Can Be Mined?
  - What Kinds of Technologies Are Used?
  - What Kinds of Applications Are Targeted?
- Content covered by this course

Course Content

- Basics in data mining
  - Prediction and classification
  - Clustering
- Advanced topics
  - Text Mining
    - Topic models
    - Word embedding
  - Recommender Systems
    - Collaborative filtering, matrix factorization
  - Information network mining
    - PageRank, Spectrum clustering, label propagation, link prediction, network embedding

Methods to Learn

	Matrix Data	Text Data	Recommender System	Graph & Network
Classification	Decision Tree; Naïve Bayes; Logistic Regression SVM; NN			Label Propagation
Clustering	K-means; hierarchical clustering; DBSCAN; Mixture Models; kernel k-means	PLSA; LDA	Matrix Factorization	SCAN; Spectral Clustering
Prediction	Linear Regression GLM		Collaborative Filtering	
Ranking				PageRank
Feature Representation		Word embedding		Network embedding

Where to Find References? DBLP, CiteSeer, Google

- Data mining and KDD (SIGKDD: CDROM)
  - Conferences: ACM-SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, PAKDD, etc.
  - Journal: Data Mining and Knowledge Discovery, KDD Explorations, ACM TKDD
- Database systems (SIGMOD: ACM SIGMOD Anthology—CD ROM)
  - Conferences: ACM-SIGMOD, ACM-PODS, VLDB, IEEE-ICDE, EDBT, ICDT, DASFAA
  - Journals: IEEE-TKDE, ACM-TODS/TOIS, JIIS, J. ACM, VLDB J., Info. Sys., etc.
- AI & Machine Learning
  - Conferences: Machine learning (ML), AAAI, IJCAI, COLT (Learning Theory), CVPR, NIPS, etc.
  - Journals: Machine Learning, Artificial Intelligence, Knowledge and Information Systems, IEEE-PAMI, etc.
- Web and IR
  - Conferences: SIGIR, WWW, CIKM, etc.
  - Journals: WWW: Internet and Web Information Systems,
- Statistics
  - Conferences: Joint Stat. Meeting, etc.
  - Journals: Annals of statistics, etc.
- Visualization
  - Conference proceedings: CHI, ACM-SIGGraph, etc.
  - Journals: IEEE Trans. visualization and computer graphics, etc.

Recommended Reference Books

- E. Alpaydin. Introduction to Machine Learning, 2nd ed., MIT Press, 2011
- S. Chakrabarti. Mining the Web: Statistical Analysis of Hypertex and Semi-Structured Data. Morgan Kaufmann, 2002
- R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, 2ed., Wiley-Interscience, 2000
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley & Sons, 2003
- U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996
- U. Fayyad, G. Grinstein, and A. Wierse, Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd ed. , 2011
- T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., Springer, 2009
- B. Liu, Web Data Mining, Springer 2006
- T. M. Mitchell, Machine Learning, McGraw Hill, 1997
- Y. Sun and J. Han, Mining Heterogeneous Information Networks, Morgan & Claypool, 2012
- P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Wiley, 2005
- S. M. Weiss and N. Indurkhya, Predictive Data Mining, Morgan Kaufmann, 1998
- I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2nd ed. 2005