Data Mining [CS 484]

Spring 2019

Carlotta Domeniconi
Associate Professor
Dept. of Computer Science
George Mason University
email: carlotta@cs.gmu.edu

Website: www.cs.gmu.edu/~carlotta

Slides are adapted from the available book slides developed by Tan, Steinbach and Kumar

Roadmap for Today

- Welcome and Introduction
- Class Policies and Syllabus
  Grading, Assignments, Exam, Project, Policies
- Introduction to Data Mining
  Examples, Motivation, Definitions, Methods

Class Logistics

CS 484 Data Mining

General Information

- Professor: Carlotta Domeniconi
  - Office: ENG, Rm 4424
  - Email: carlotta@cs.gmu.edu
- Office hours: TR 1:30-2:30pm, or by appointment, or stop by
- http://www.cs.gmu.edu/~carlotta/teaching//CS-484-s19/info.html (follow the link "Schedule")
- Visit the class webpage often!

General Information

- GTA: Priya Mani
  - Office: Rm. 4456
  - Email: pmani@masonlive.gmu.edu
- Office hours: T 3 - 4pm; W 3:30 – 4:30pm

Objective of the course

- The course covers key concepts and algorithms at the core of data mining. Topics include: classification, clustering, association analysis, and anomaly detection.

- Technical tools from… linear algebra, probability, and statistics, optimization. Programming experience is expected.

Course Format

- Lectures by me: no laptops, please!
- Participation: attendance is highly recommended!
- Four Assignments: individual effort (unless specified otherwise). Lots of programming!
- One exam: in class and closed book
- Project: proposal, presentation, report
- We will set up a Piazza account to enable questions/discussions

Important Dates

- **April 9**: Exam

- **March 7**: Project pitch
- **March 21:** Project proposal
- **May 2:** Project presentation
- **May 9:** Project report (No final)

The final grade is based on…

- Assignments: 40% (10% each assignment)
- Exam: 20%
- Project: 40%

- Project: (proposal, presentation, paper)
  - Proposal: 5%
  - Presentation: 10%
  - Report: 25%

- Extra credit: participation; winners of competitions

Resources

- **Weka**: open source Java package implementing many learning algorithms for data mining tasks. It contains tools for data pre-processing, classification, clustering, association rules, and visualization.

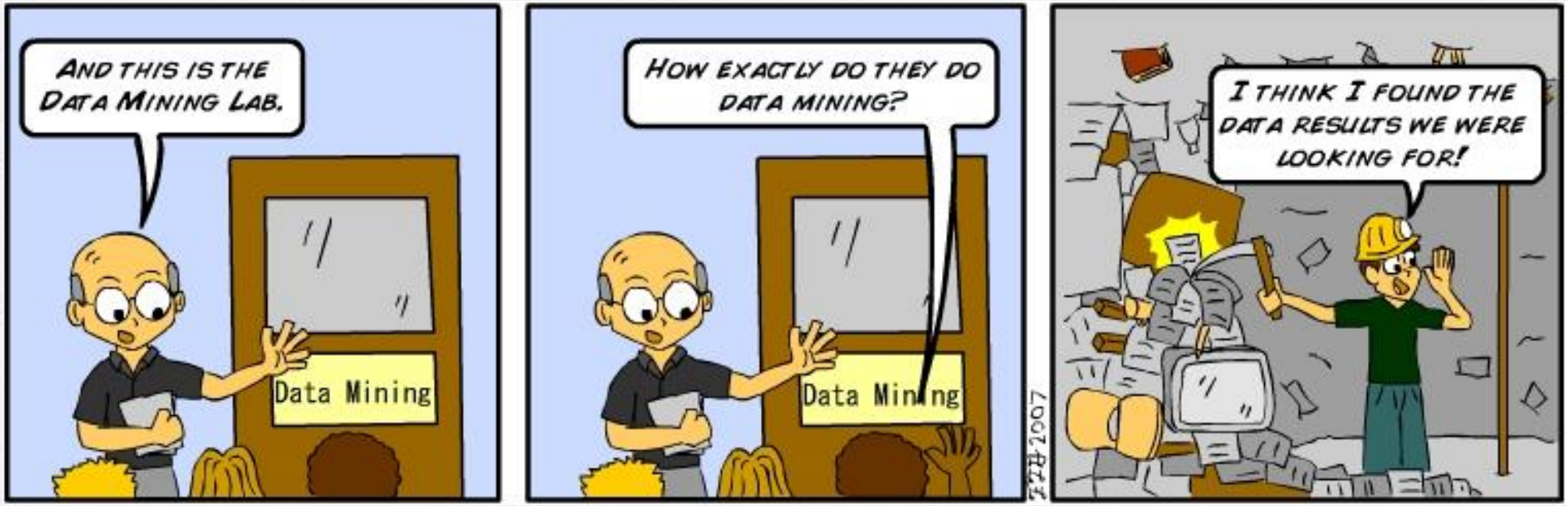
- **MEKA**: A Multi-label Extension to WEKA: http://meka.sourceforge.net/

- **MALLET** [Machine Learning for LanguagE Toolkit]: http://mallet.cs.umass.edu/

Resources (cont.ed)

- **scikit-learn**: Machine Learning in Python
http://scikit-learn.org/stable/

What is data mining?



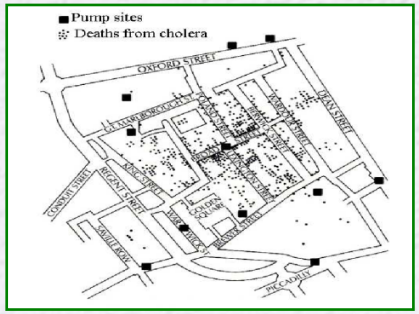
- Any example?
- Definition?

Large-scale Data is Everywhere!

- There has been enormous data growth in both commercial and scientific databases due to advances in data generation and collection technologies
- New mantra
  - Gather whatever data you can whenever and wherever possible.
- Expectations
  - Gathered data will have value either for the purpose collected or for a purpose not envisioned.

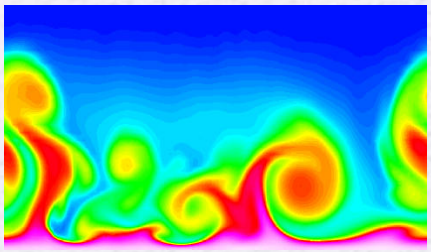
Homeland Security

Business Data



Geo-spatial data

Sensor Networks



Computational Simulations

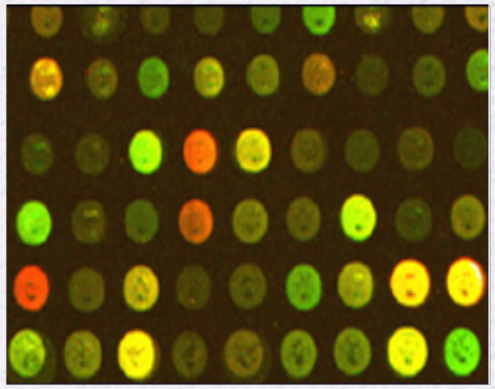
Why Data Mining? Commercial Viewpoint

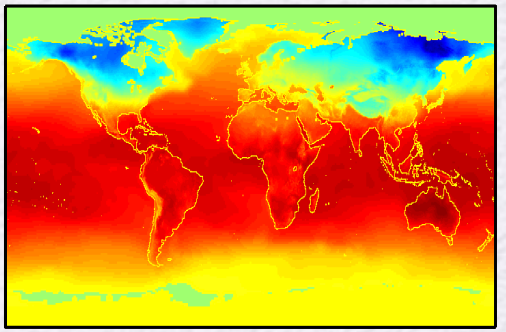
- Lots of data is being collected and warehoused
  - Web data
    - Yahoo has 2PB web data
    - Facebook has 400M active users
  - purchases at department/ grocery stores, e-commerce
    - Amazon records 2M items/day
  - Bank/Credit Card transactions

- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better, customized services for an edge (e.g. in Customer Relationship Management)

Why Data Mining? Scientific Viewpoint

- Data collected and stored at enormous speeds
  - remote sensors on a satellite
    - NASA EOSDIS archives over 1-petabytes of earth science data / year
  - telescopes scanning the skies
    - Sky survey data
  - High-throughput biological data
  - scientific simulations
    - terabytes of data generated in a few hours





- Data mining helps scientists
  - in automated analysis of massive datasets
  - In hypothesis formation

Mining Scientific Data - Fields

- Past decade has seen a huge growth of interest in mining data in a variety of scientific domains

- Astroinformatics
- Neuroinformatics
- Quantum Informatics
- Health Informatics
- Evolutionary Informatics
- Veterinary Informatics
- Organizational Informatics
- Pharmacy Informatics
- Social Informatics
- Ecoinformatics
- Geoinformatics
- Chemo Informatics

My Favorite Data Mining Examples

- Amazon.com, Google News, Netflix, Target
  - Personal Recommendations.
  - Profile-based advertisements.
  - How Target figured out a teen girl was pregnant before her father did!
- Spam Filters/Priority Inbox
  - Nigerian 419 scam; Spanish Prisoner scam; Black money scam.
- Scientific Discovery
  - Grouping patterns in sky.
  - Inferring complex life science processes.
  - Weather forecasting.

Americans Warned to Be Vigilant in Europe After Terror Threat

Bloomberg - Mark Drajem - **27 minutes ago**

Analysis: Terrorism alerts reflect evolving militant threat Reuters

U.S. Travel Alert Seeks VigilanceAcross Europe Wall Street Journal

USA Today - Voice of America - ABC Online - Los Angeles Times - Wikipedia: 2010 European terror plot

all 2,673 news articles »

Emanuel Says He's Preparing Run For Chicago Mayor

NPR - **1 hour ago**

It won't be easy for Emanuel, Chicago analysts warn USA Today

As Emanuel returns home to run for mayor, many in Chicago say he has a lot to ... Washington Post

Chicago Tribune - St. Louis Post-Dispatch - FOXNews - ABC News

all 544 news articles »

Rand Paul now says he'd support Mitch McConnell as Senate GOP leader

Los Angeles Times - Kathleen Hennessey - **53 minutes ago**

Kentucky Race for Senate Puts Focus on National Policies Wall Street Journal

Paul ties rival to Obama in Senate debate Washington Times

FOXNews - WFPL - WKYT - The Associated Press

all 293 news articles »

As Mourners Were Honoring Tyler Clementi, News Came of a Fifth Suicide

ABC News - Jeremy Hubbard - **7 hours ago**

Tyler Clementi suicide: Reaction is swift and widespread Christian Science Monitor

NJ school holds vigil for student who killed self The Associated Press

New York Daily News - New York Post - The Star-Ledger - NJ.com - New York Times - Wikipedia: Suicide of Tyler Clem

all 3,863 news articles »

Polls tighten as elections approach. Good news for Democrats? Maybe.

Christian Science Monitor - Brad Knickerbocker - **4 hours ago**

Parties Scramble to Adapt To Shifting Political Scene Wall Street Journal

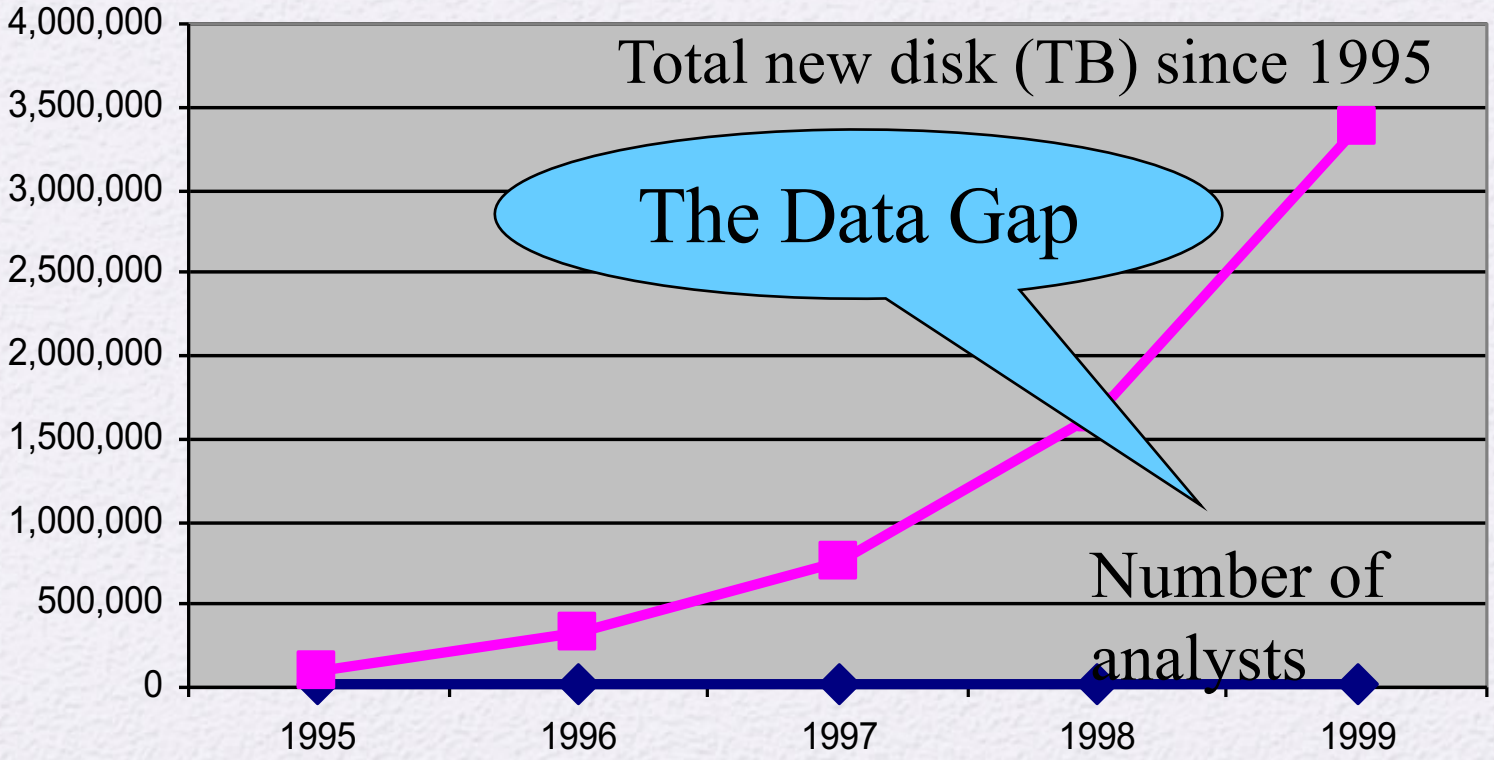
With election losses certain, Democrats discuss the way forward CNN International

The Associated Press - USA Today - Bloomberg - CQPolitics.com

all 915 news articles »

Mining Large Data Sets - Motivation

- There is often information "hidden" in the data that is not readily evident
- Human analysts may take weeks to discover useful information
- Much of the data is never analyzed at all



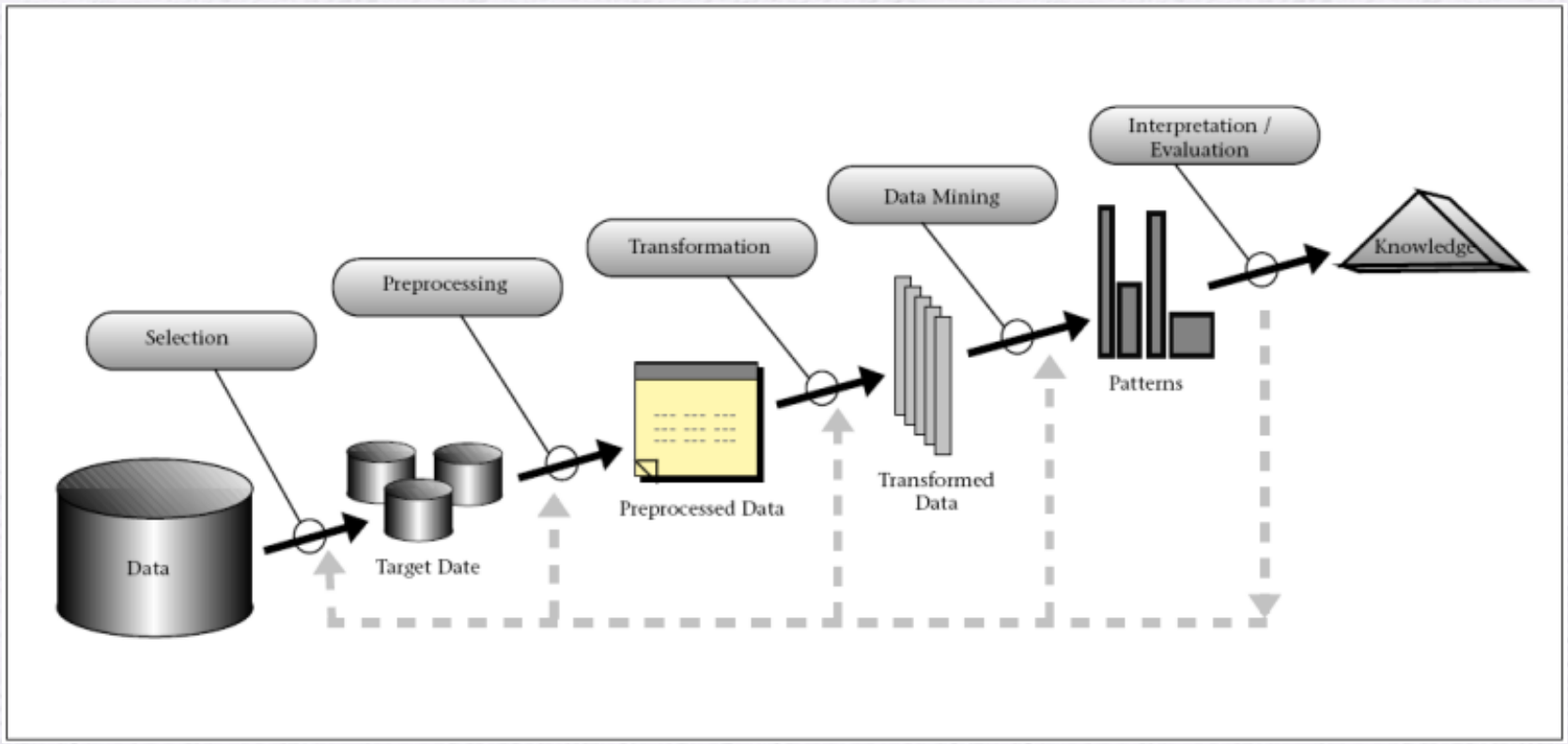
From: R. Grossman, C. Kamath, V. Kumar, "Data Mining for Scientific and Engineering Applications"

Data Mining Definitions

- Non-trivial extraction of implicit, previously unknown and potentially useful information from data (normally large databases).
- Exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns.
- Part of the Knowledge Discovery in Databases Process.

KDD Process

CONVERTING RAW DATA TO USEFUL INFORMATION.



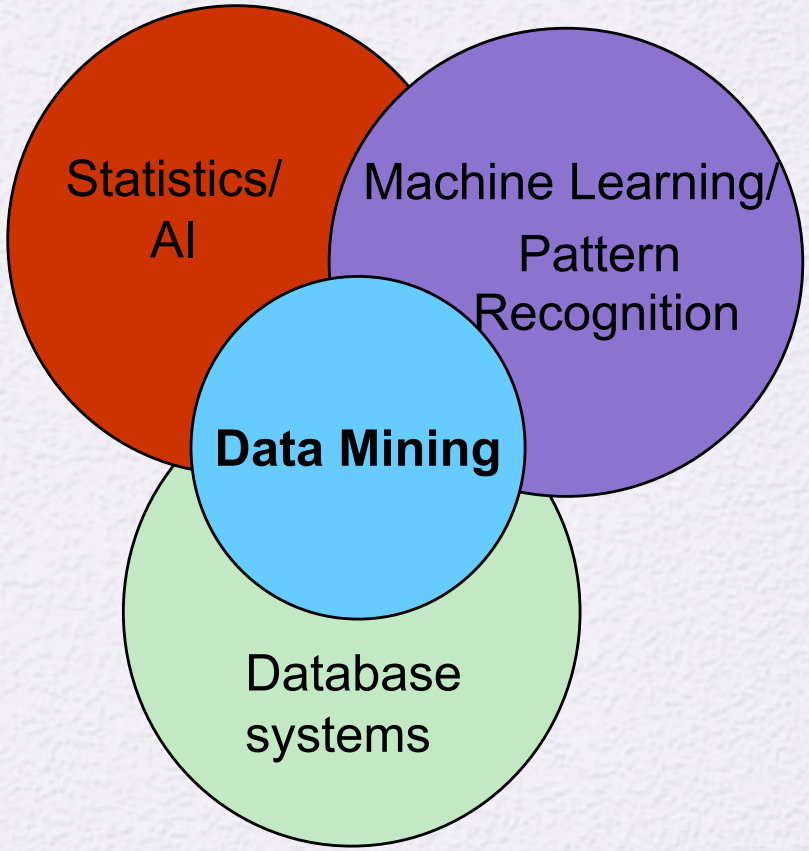
Fayyad 1996
http://liris.cnrs.fr/abstract/abstract.html

What is (not) Data Mining?

- What is not Data Mining?
  - Look up phone number in phone directory
  - Query a Web search engine for information about "Amazon"
- What is Data Mining
  - Certain names are more prevalent in certain US locations (O'Brien, O'Rurke, O'Reilly… in Boston area)
  - Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com,

Origins of Data Mining

- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems
- Traditional Techniques may be unsuitable due to
  - Enormity of data
  - High dimensionality of data
  - New types of data
  - Heterogeneous, distributed nature of data



Data Mining Tasks (1)

- **Predictive Tasks**
  - Use some variables to predict unknown or future values of other variables.
- **Descriptive Tasks**
  - Find human-interpretable patterns that describe the data; e.g.: groups, correlations, trends, anomalies.

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Data Mining Tasks (2)

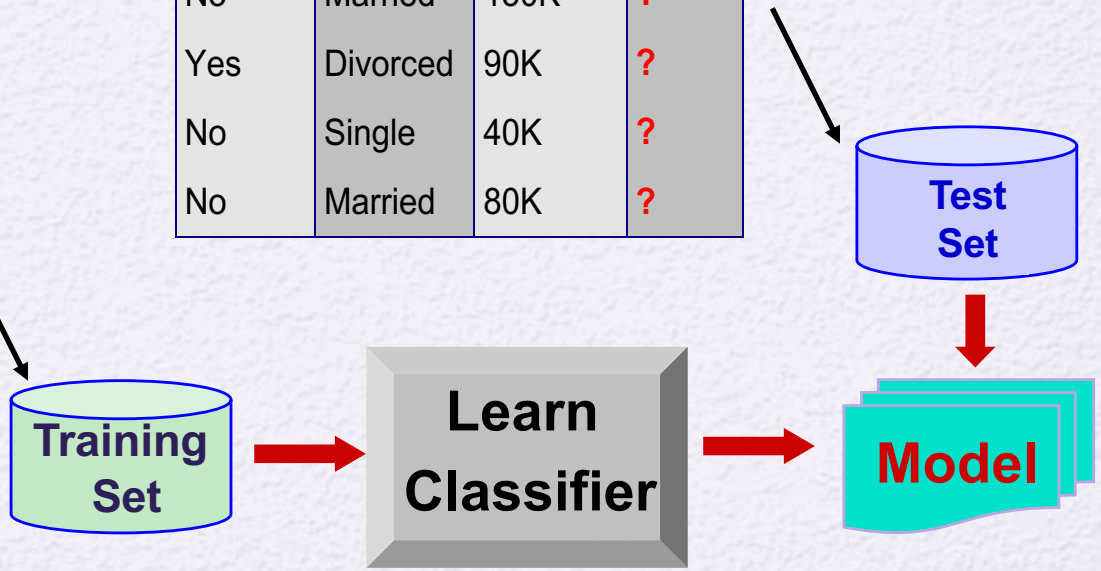
- Classification [Predictive]
- Regression [Predictive]
- Clustering [Descriptive]
- Anomaly (Outlier) Detection [Descriptive]
- Association Rule Discovery [Descriptive]

Classification: Example (1)

categorical categorical continuous class

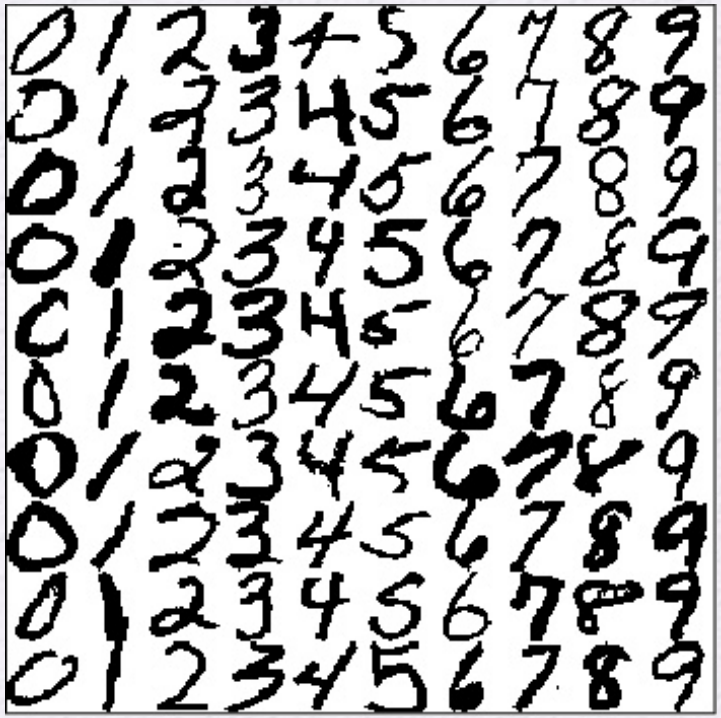
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?

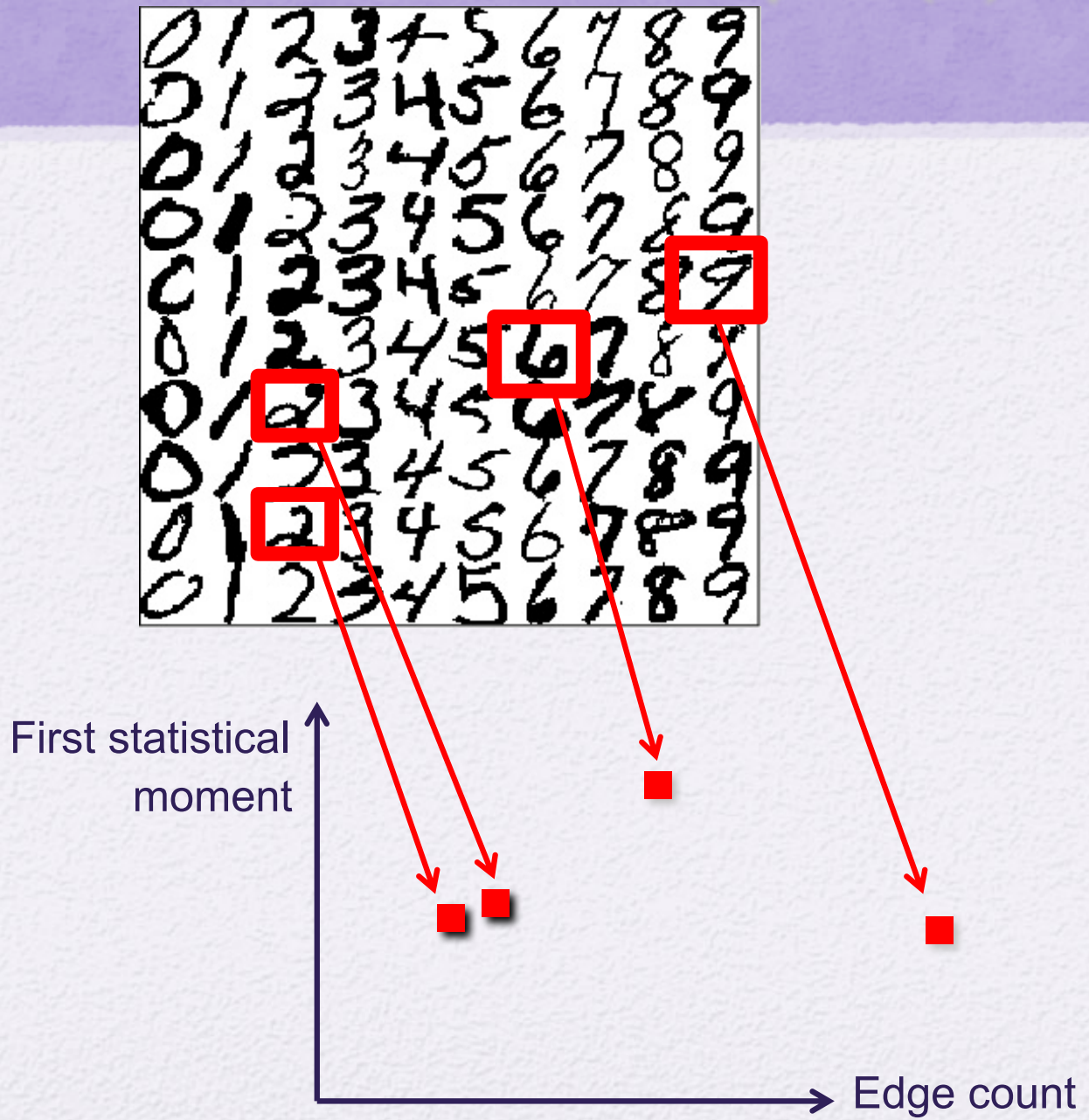


Classification: Example (2)

- You've been given a set of *N* pictures of digits. For each picture, you're told the digit number
- Discover a set of rules which, when applied to pictures you've never seen, correctly identifies the digits in those pictures



Classification: Example (2)



Classification: Definition

- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Classification: Direct Marketing

- Direct Marketing
  - Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.
  - Approach:
    - Use the data for a similar product introduced before.
    - We know which customers decided to buy and which decided otherwise. This {*buy, don't buy*} decision forms the *class attribute*.
    - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
      - Type of business, where they stay, how much they earn, etc.
    - Use this information as input attributes to learn a classifier model.

From [Berry & Linoff] Data Mining Techniques, 1997

Classification: Your Turn

- Fraud Detection
  - Goal: Predict fraudulent cases in credit card transactions.
  - Approach:
    - What kind of data will you try to get?
    - Can you say something about the characteristics of the data?
    - What kind of pitfalls you might run into?

Classification: Fraud Detection

- Fraud Detection
  - Goal: Predict fraudulent cases in credit card transactions.
  - Approach:
    - Use credit card transactions and the information on account-holder as attributes.
      - When does a customer buy, what does he buy, how often he pays on time, etc
    - Label past transactions as fraud or fair transactions. This forms the class attribute.
    - Learn a model for the class of the transactions.
    - Use this model to detect fraud by observing credit card transactions on an account.

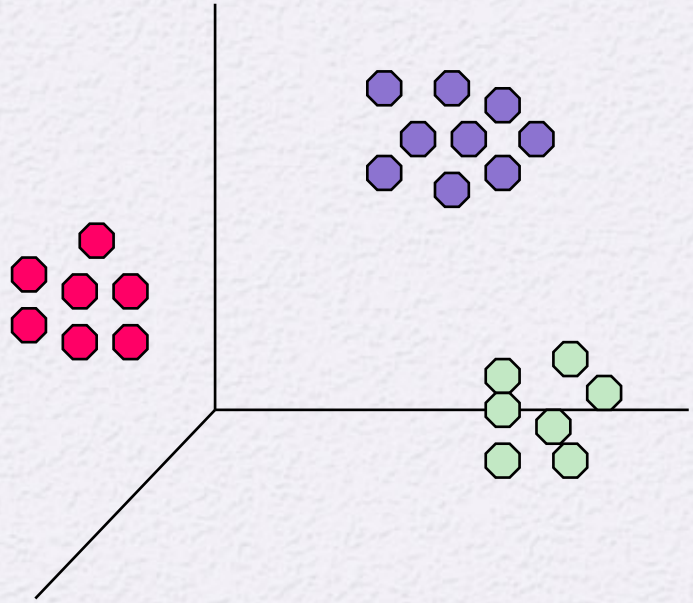
Clustering

- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - Euclidean Distance if attributes are continuous.
  - Other Problem-specific Measures.

Illustrating Clustering

Intracluster distances are minimized

Intercluster distances are maximized

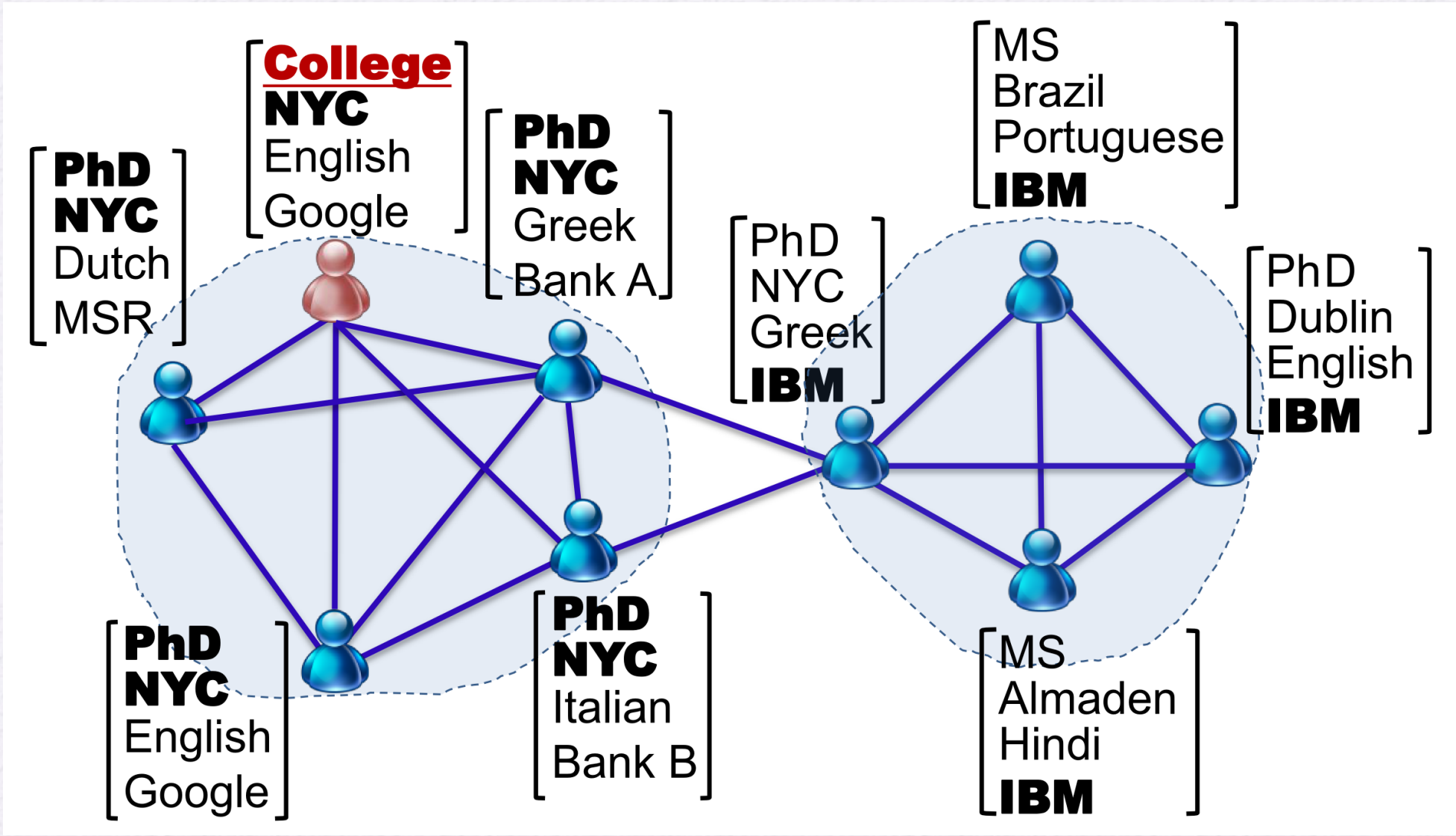


- Euclidean Distance Based Clustering in 3-D space.

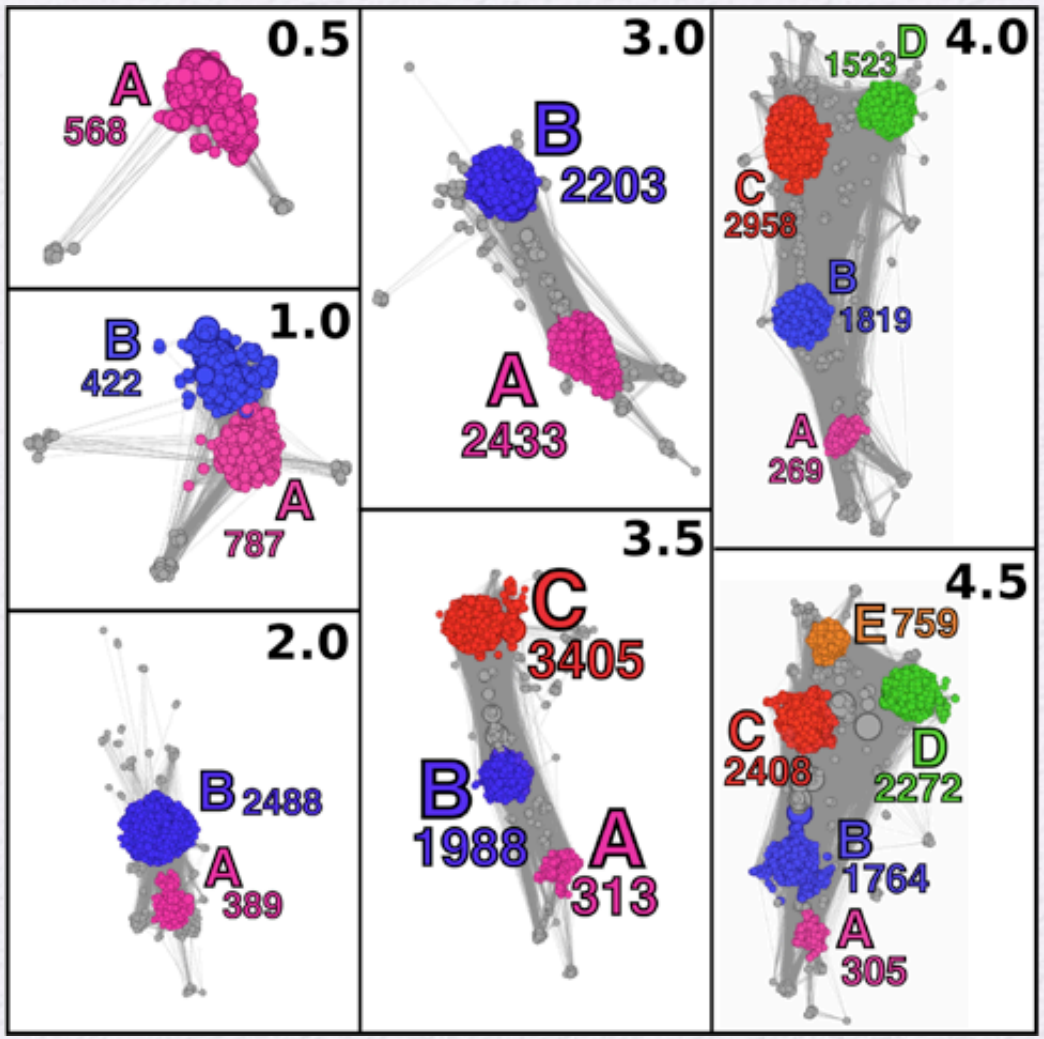
Clustering: Documents

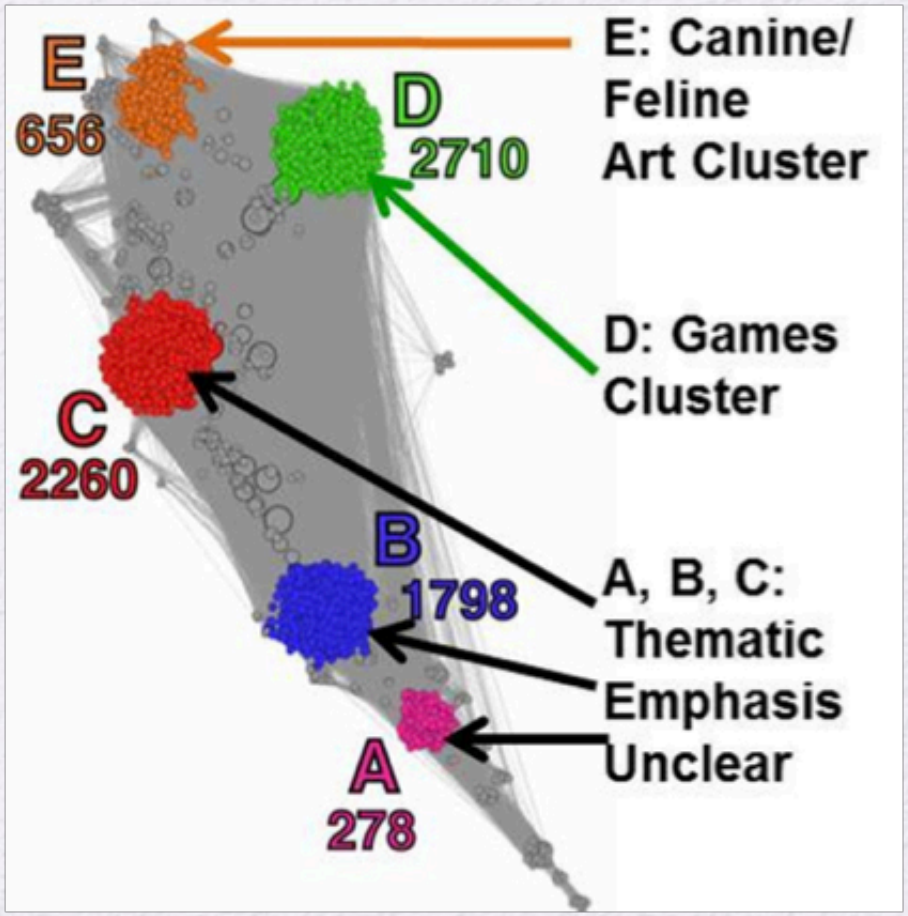
- Document Clustering:
  - Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
  - Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
  - Gain: Information Retrieval can utilize the clusters to relate a new document or search terms in clustered documents.

Clustering: community discovery in graphs

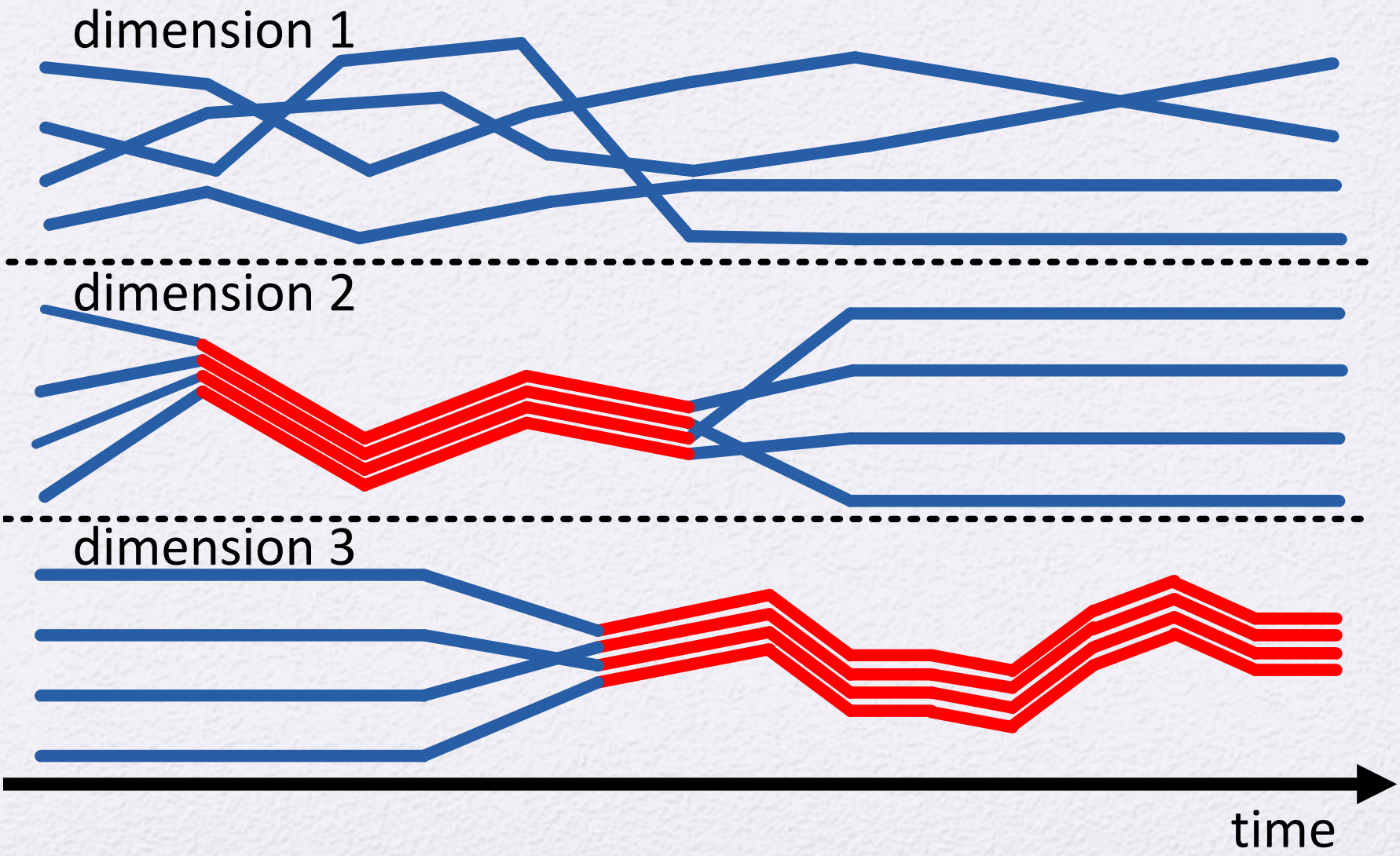


Clustering: community discovery in graphs





Clustering: time series



Let's Think!

- Differences between classification and clustering?

Association Rule Discovery: Definition

- Given a set of records each of which contain some number of items from a given collection;
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

Market Basket Analysis

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

Rules Discovered:

{Diapers} --> {Milk}

Urban Legend …

- Classic Association Rule Example:
  - If a customer buys diaper and milk, then he is very likely to buy beer.
  - Any plausible explanations? ☺

Association Rule Discovery: Application 1

- Marketing and Sales Promotion:
  - Let the rule discovered be
    {Bagels, … } --> {Potato Chips}
  - Potato Chips as consequent => Can be used to determine what should be done to boost its sales.
  - Bagels in the antecedent => Can be used to see which products would be affected if the store discontinues selling bagels.

Association Rule Discovery: Application 2

- Supermarket shelf management.
  - Goal: To identify items that are bought together by sufficiently many customers.
  - Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.
  - Wal-mart, Target, and departmental store managers are big into this.
  - All your purchases get processed & analyzed in a warehouse.

Regression

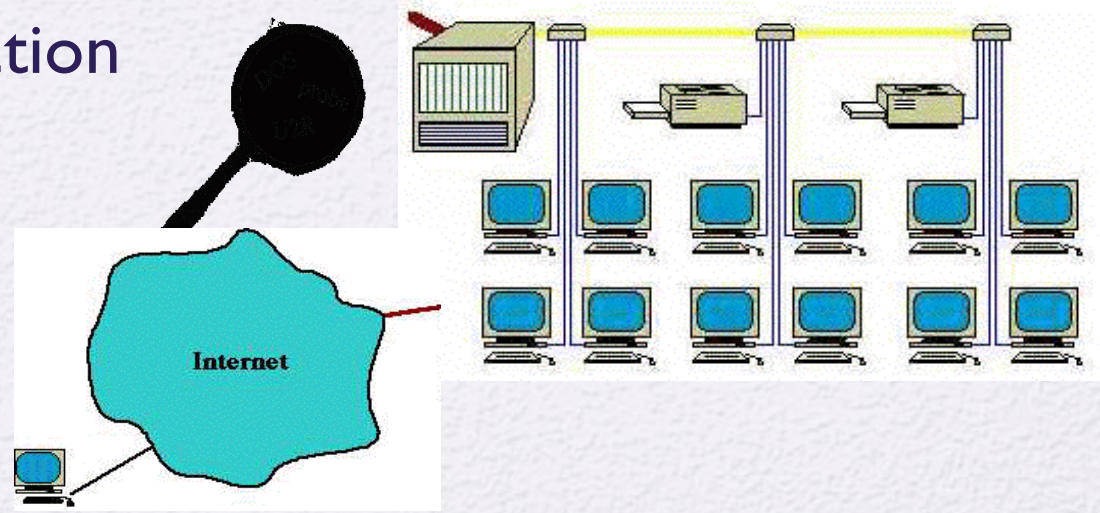
- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Also called density estimation.
- Greatly studied in statistics, neural network fields.
- Examples:
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.

Regression: Example

- A CMU team trained a neural network to drive a car by feeding in many pictures of roads, plus the *value* according to which the graduate student was turning the steering wheel at the time.
- Eventually the neural network learned to predict the correct value for previously unseen pictures of roads.

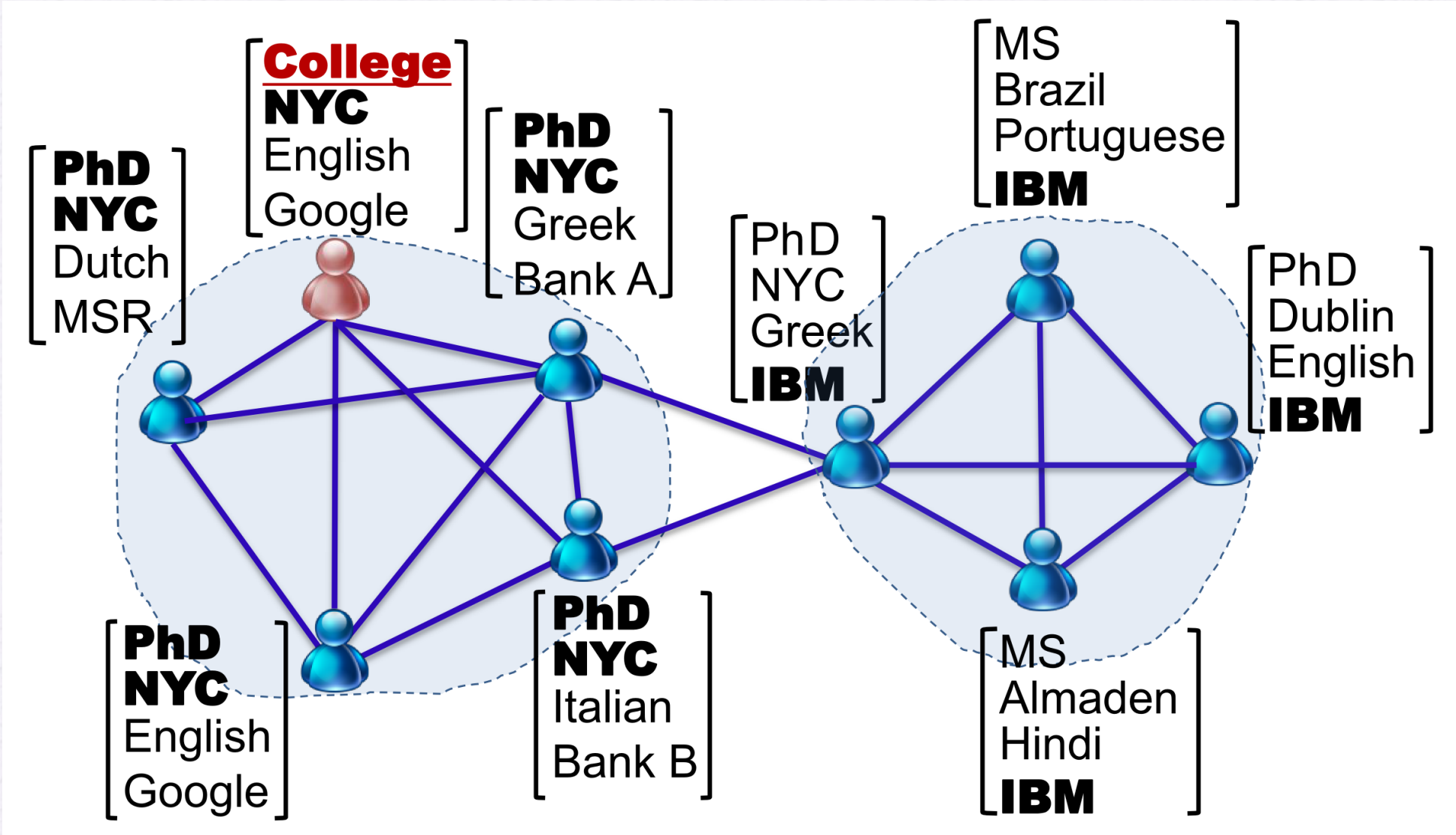
Outlier/Anomaly Detection

- Detect significant deviations from normal behavior
- Applications:
  - Credit Card Fraud Detection
  - Network Intrusion Detection



Anomaly detection in images

Anomaly detection in graphs



What else can Data Mining do?

Dilbert

Challenges of Data Mining

- Scalability
- Dimensionality
- Complex and Heterogeneous Data
- Data Quality: missing data and noise
- Data Ownership and Distribution
- Privacy Preservation
- Streaming Data