BBS654
Data Mining

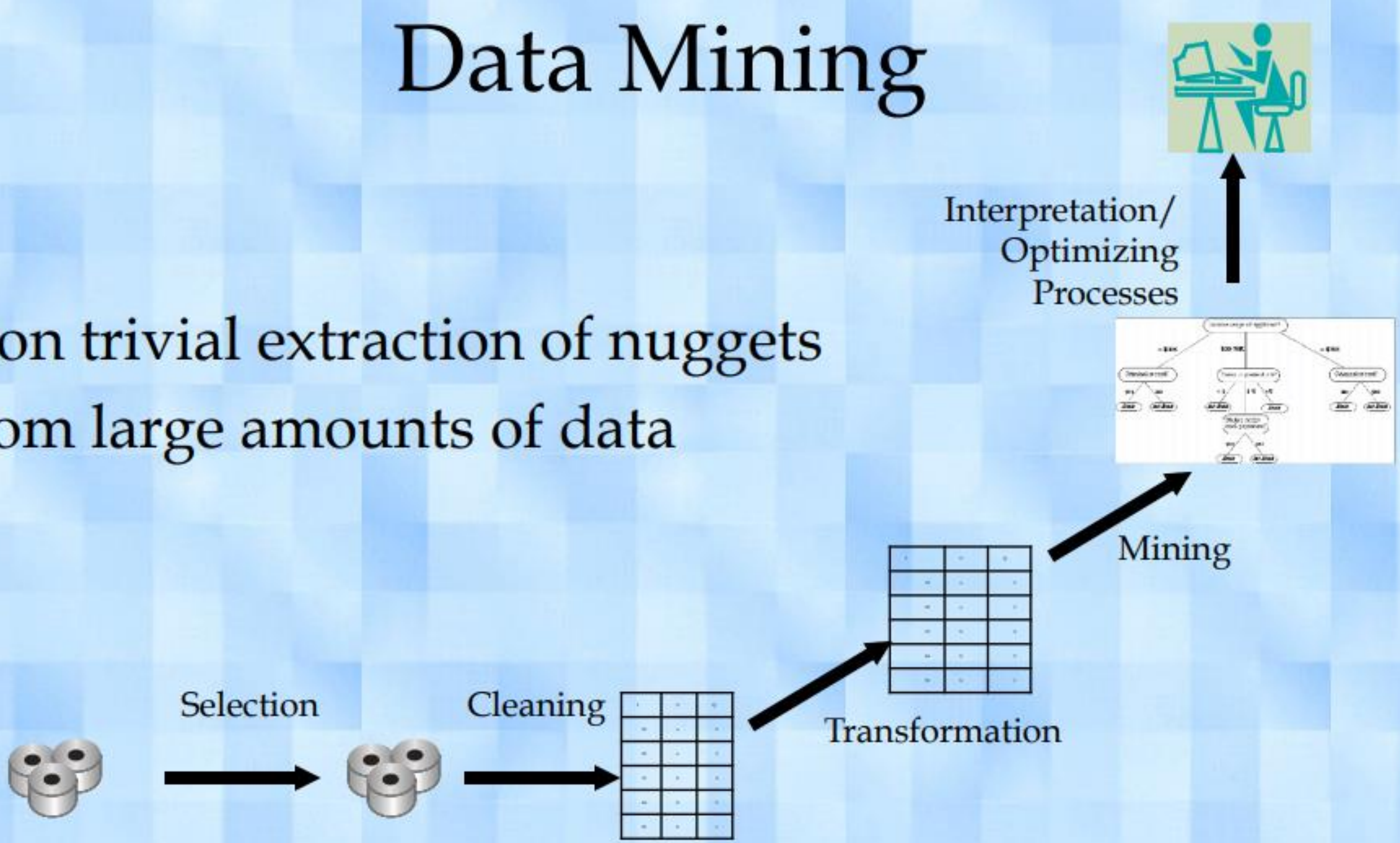
Pinar Duygulu

Slides are adapted from

Nazli Ikizler, Sanjay Ranka

Data Mining

Non trivial extraction of nuggets from large amounts of data



Selection

Cleaning

Transformation

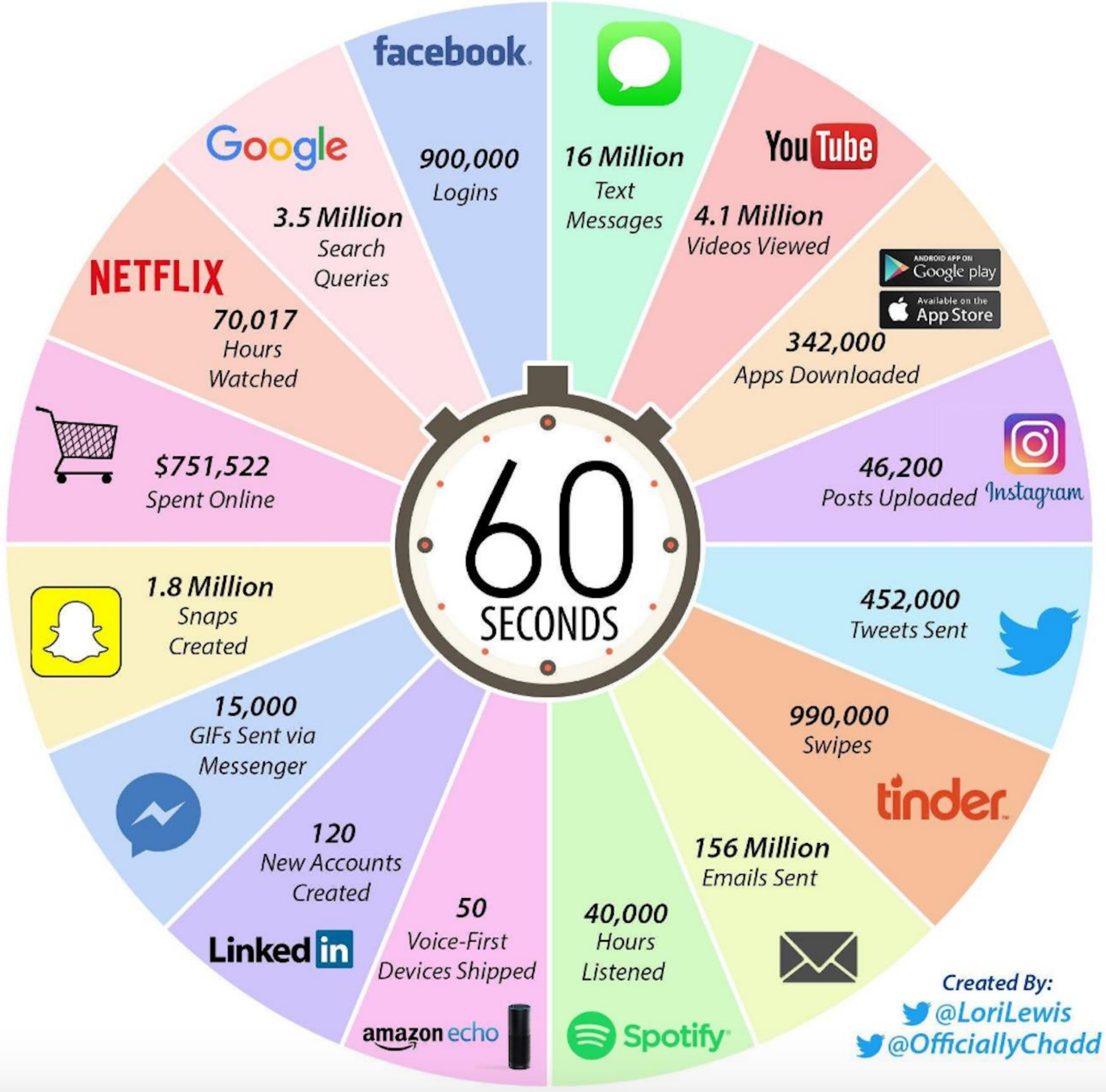
Mining

Interpretation/
Optimizing
Processes

There are lots of data around

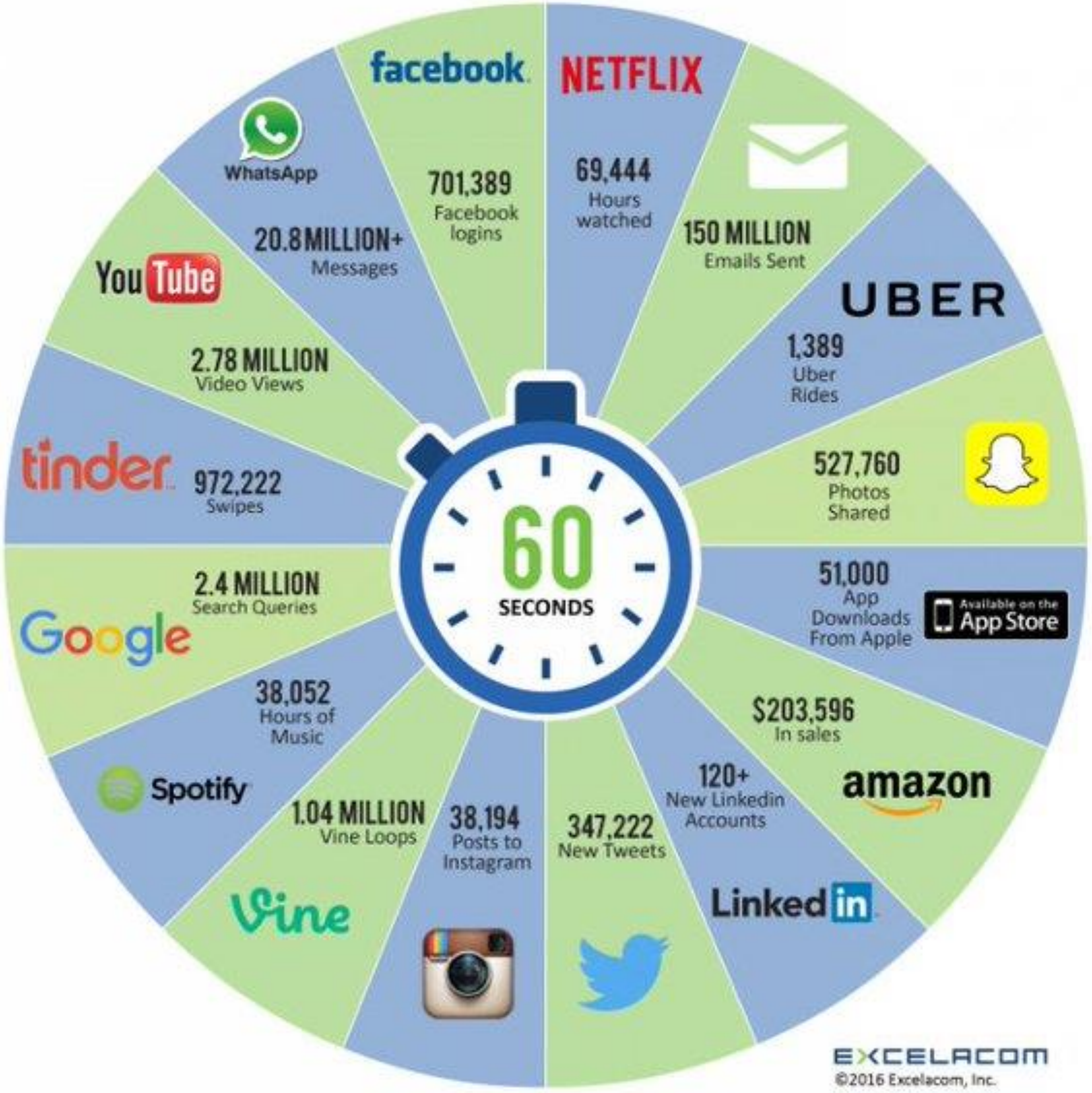
- Web (~50 billion pages) (indexed by Google)
- Online social networks (Facebook has 1.86 billion users - 2016)
- Recommendation systems (93.8 million subscribers on Netflix)
- Wikipedia has 5.33 million articles in English, 40 million articles in 293 languages and counting
- Genomic sequences: 310^9 nucleotides per individual for 1000 people --> 310^12 nucleotided...+ medical history + census information

2017 This Is What Happens In An Internet Minute



Created By:
@LoriLewis
@OfficiallyChadd

2016 What happens in an INTERNET MINUTE?



facebook
701,389 Facebook logins

NETFLIX
69,444 Hours watched

150 MILLION Emails Sent

UBER
1,389 Uber Rides

527,760 Photos Shared

51,000 App Downloads From Apple
Available on the App Store

$203,596 In sales
amazon

120+ New Linkedin Accounts
LinkedIn

347,222 New Tweets

38,194 Posts to Instagram

1.04 MILLION Vine Loops
Vine

38,052 Hours of Music
Spotify

2.4 MILLION Search Queries
Google

tinder
972,222 Swipes

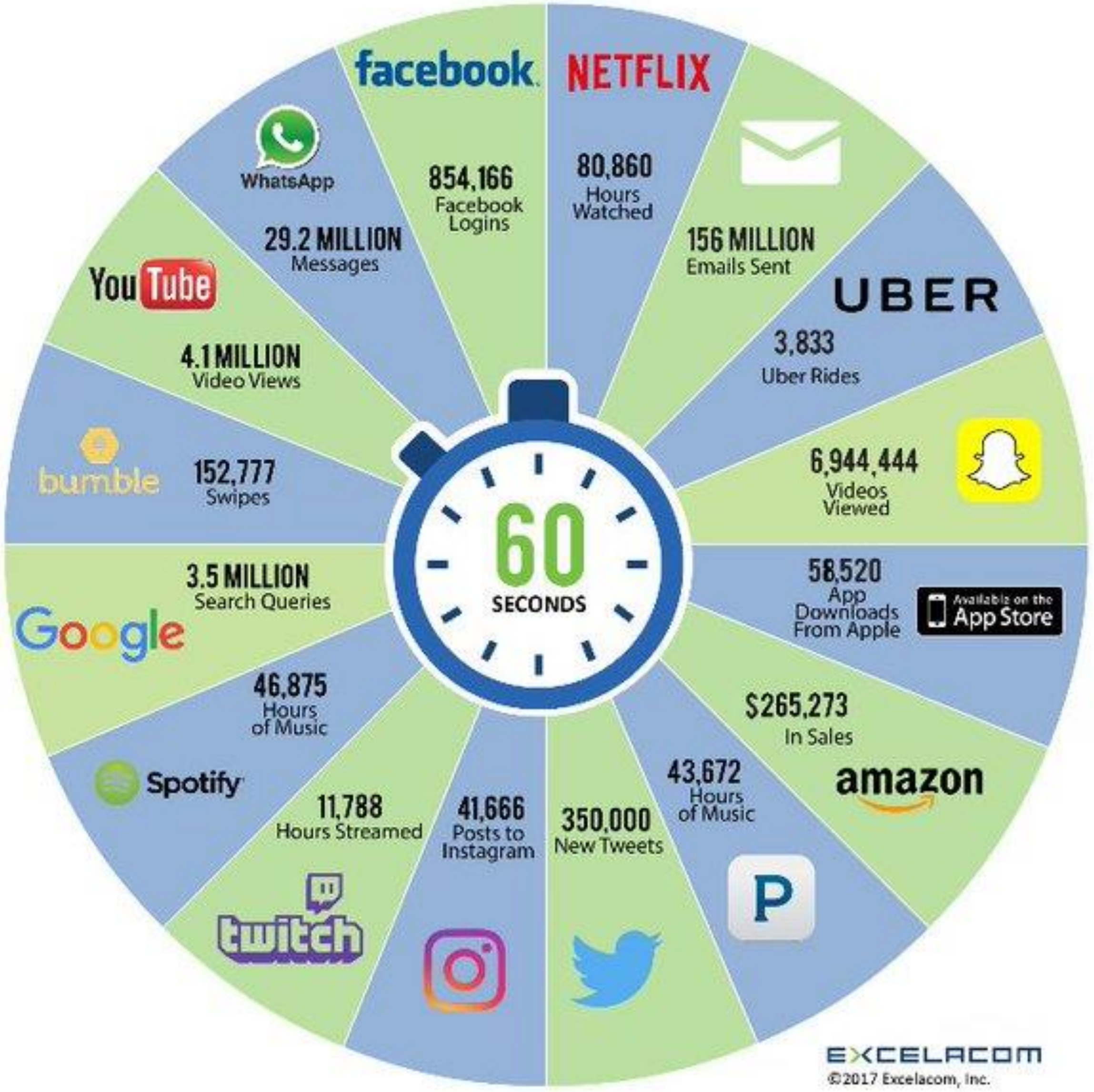
YouTube
2.78 MILLION Video Views

WhatsApp
20.8 MILLION+ Messages

60 SECONDS

EXCELACOM

2017 What happens in an INTERNET MINUTE?



facebook
854,166 Facebook Logins

NETFLIX
80,860 Hours Watched

156 MILLION Emails Sent

UBER
3,833 Uber Rides

6,944,444 Videos Viewed

58,520 App Downloads From Apple
Available on the App Store

$265,273 In Sales
amazon

43,672 Hours of Music

350,000 New Tweets

41,666 Posts to Instagram

11,788 Hours Streamed
twitch

46,875 Hours of Music
Spotify

3.5 MILLION Search Queries
Google

bumble
152,777 Swipes

YouTube
4.1 MILLION Video Views

WhatsApp
29.2 MILLION Messages

60 SECONDS

EXCELACOM

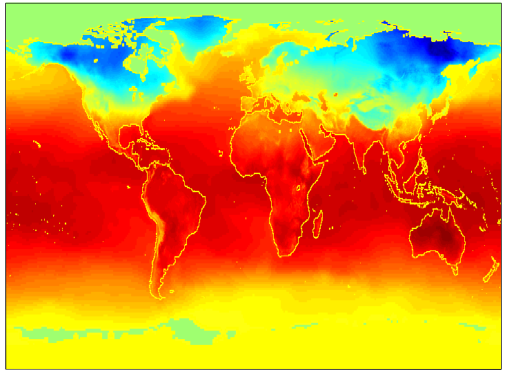
Why Mine Data? – Commercial Viewpoint

- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - purchases at department/ grocery stores
  - Bank/Credit Card transactions

- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better customized services

Why Mine Data? –Scientific Viewpoint

- Data collected and stored at enormous speeds (GB/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data



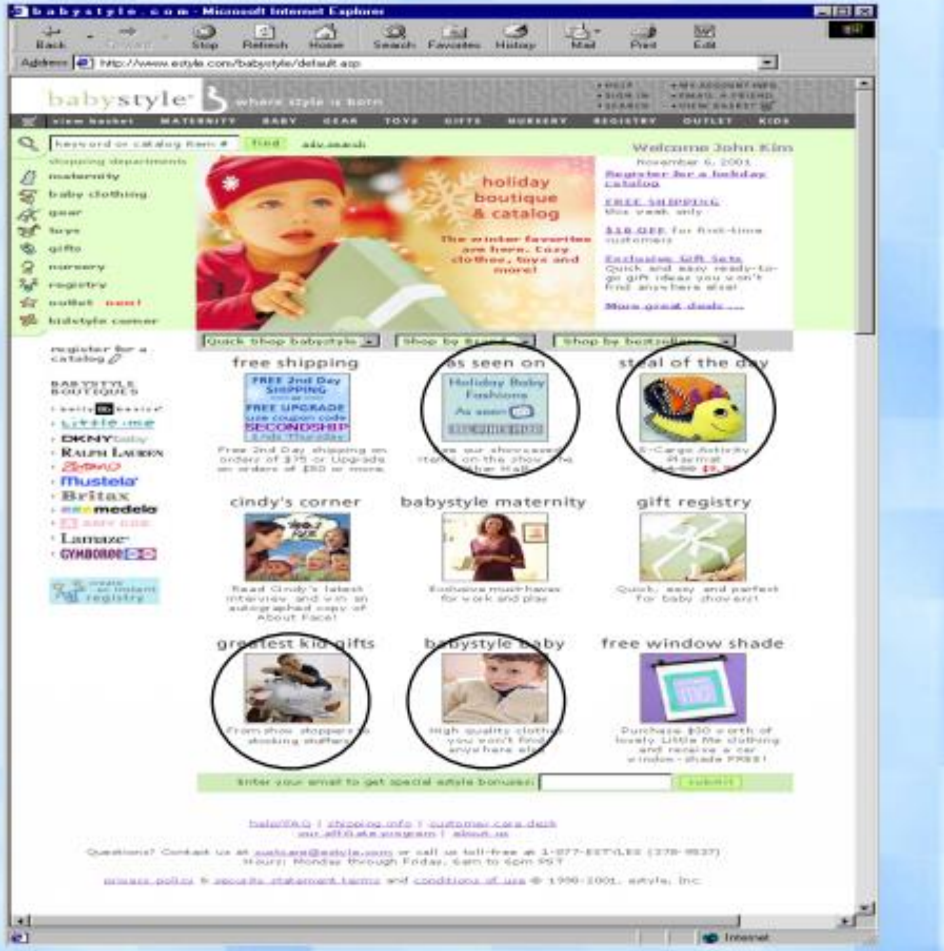
- Traditional techniques infeasible for raw data
- Data mining may help scientists
  - in classifying and segmenting data
  - in Hypothesis Formation

POPULAR SCIENCE
THE FUTURE NOW
THE CONTROL CENTERS
Using Data to Feed the World, Solve Cold Cases, Battle Malware, Predict Our Fate
OFFICER ALGORITHM
Can a Crime Be Prevented Before It Begins?
NEW WAYS OF SEEING
A Gallery of Extraordinary Infographics
PLUS
Juan Enriquez Reprograms Life
James Gleick Unsplits the Bit
AND Lawrence Weschler Questions the Cloud
SPECIAL ISSUE
DATA IS POWER
HOW INFORMATION IS DRIVING THE FUTURE

Data contains value and knowledge

Why is Data Mining prevalent? Quality and richness of data collected in improving

- Retailers
  - Scanner data is much more accurate than other means
- E-Commerce
  - Rich data on consumer browsing
- Science
  - Accuracy of sensors is improving



Welcome John Kim
November 6, 2001

Register for a holiday catalog

FREE SHIPPING this week only

$10 OFF for first-time customers

Exclusive Gift Sets
Quick and easy ready-to-go gift ideas you won't find anywhere else!

More great deals ...

The winter favorites are here. Cozy clothes, toys and more!

Free 2nd Day shipping on orders of $75 or Upgrade on orders of $50 or more.

Read Cindy's latest interview and win an autographed copy of About Face!

Exclusive must-haves for work and play

Quick, easy and perfect for baby showers!

High quality clothes you won't find anywhere else

Purchase $30 worth of lovely Little Me clothing and receive a car window-shade FREE!

Enter your email to get special estyle bonuses:

help/FAQ | shipping info | customer care desk
our affiliate program | about us

Questions? Contact us at custcare@estyle.com or call us toll-free at 1-877-ESTYLES (378-9537)
Hours: Monday through Friday, 6am to 6pm PST

privacy policy & security statement terms and conditions of use © 1998-2001, estyle, Inc.

Data Mining

- **But to extract the knowledge data needs to be**
  - **Stored**
  - **Managed**
  - **And ANALYZED ← this class**

What is Data Mining?

- **Given lots of data**
- **Discover patterns and models that are:**
  - **Valid:** hold on new data with some certainty
  - **Useful:** should be possible to act on the item
  - **Unexpected:** non-obvious to the system
  - **Understandable:** humans should be able to interpret the pattern

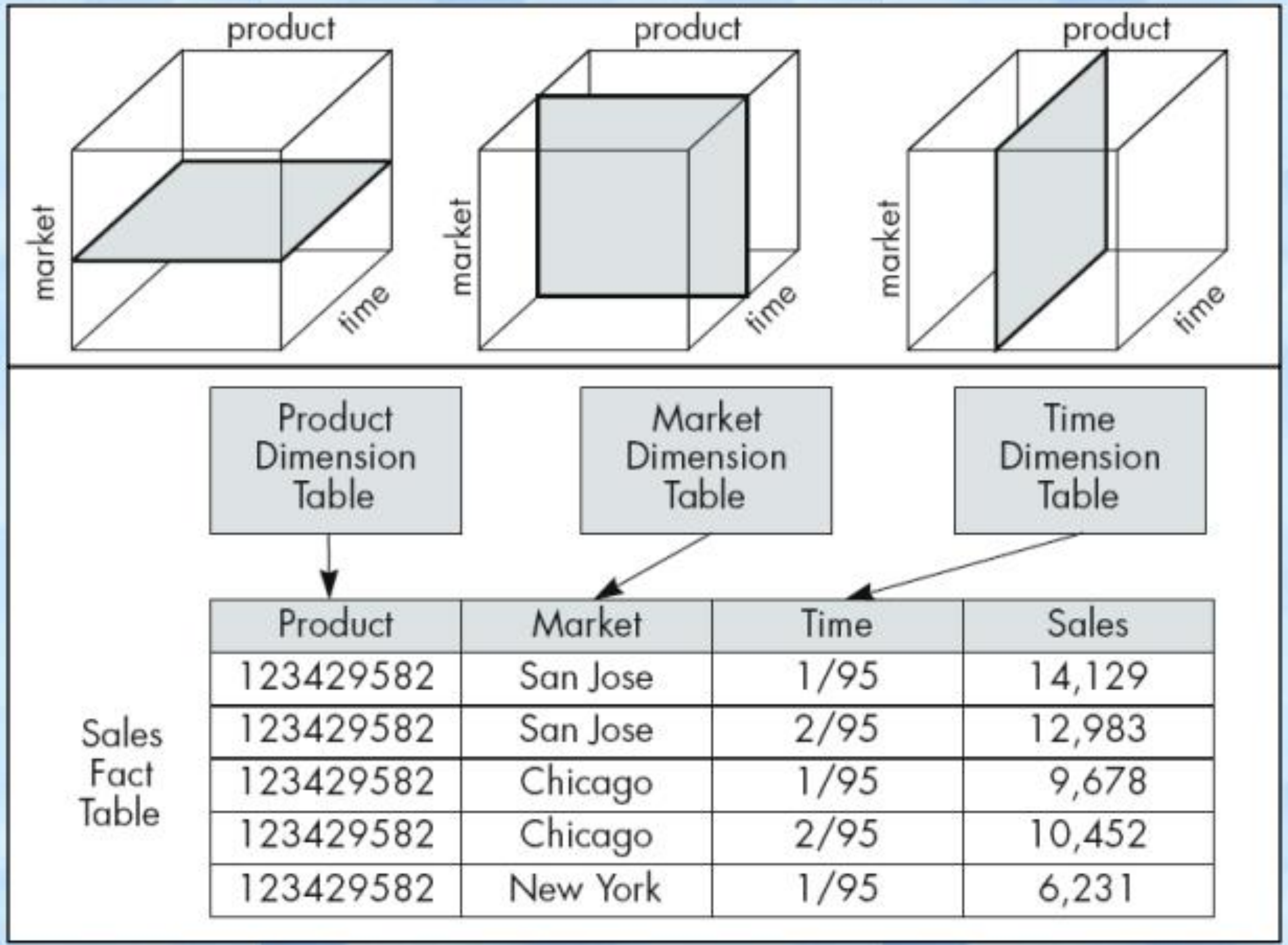
What Is Data Mining?

- 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, information harvesting, business intelligence, etc.

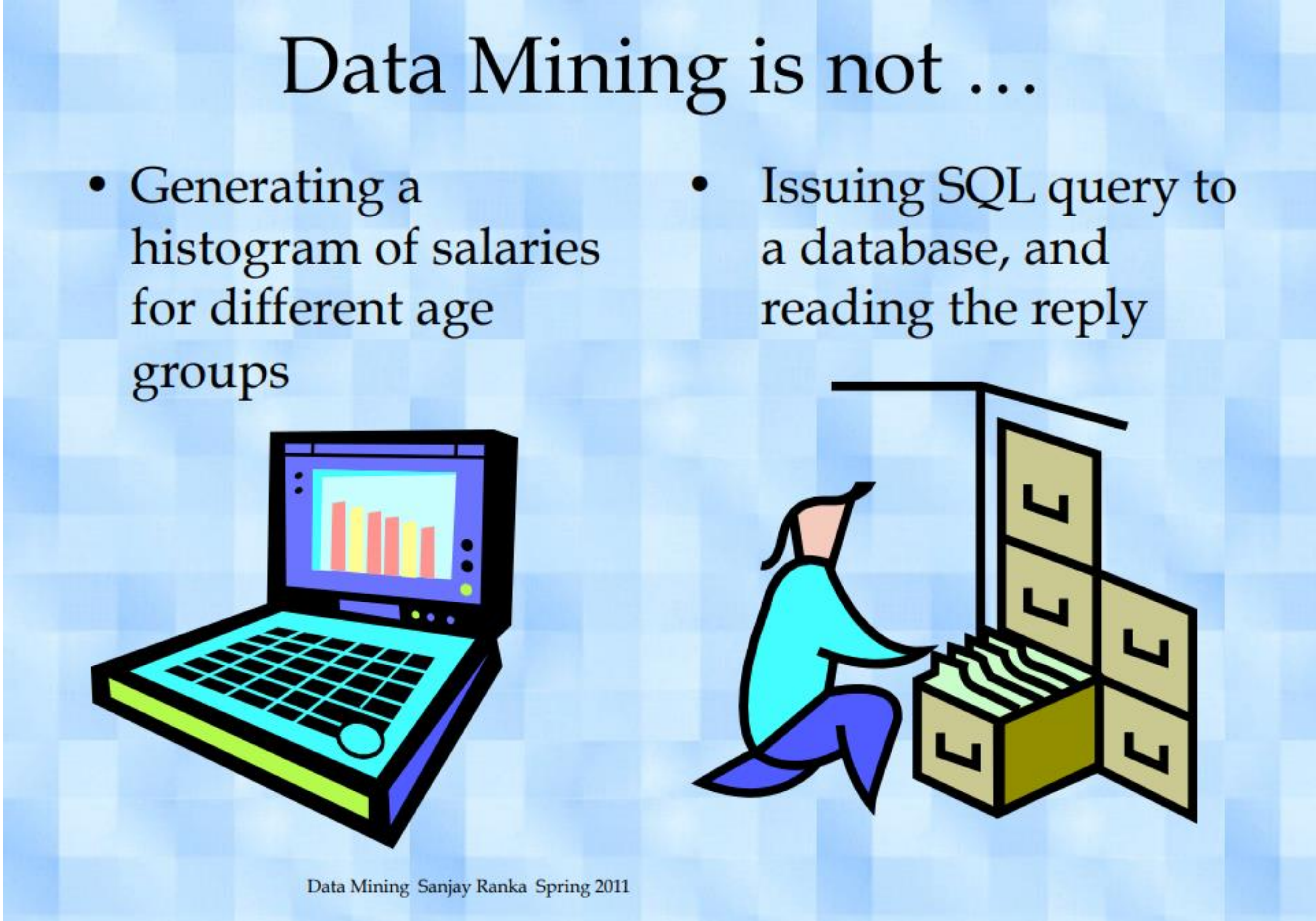
Data Mining is not …

- Generating multidimensional cubes of a relational table



Product	Market	Time	Sales
123429582	San Jose	1/95	14,129
123429582	San Jose	2/95	12,983
123429582	Chicago	1/95	9,678
123429582	Chicago	2/95	10,452
123429582	New York	1/95	6,231

Source: Multidimensional OLAP vs. Relational OLAP by Colin White



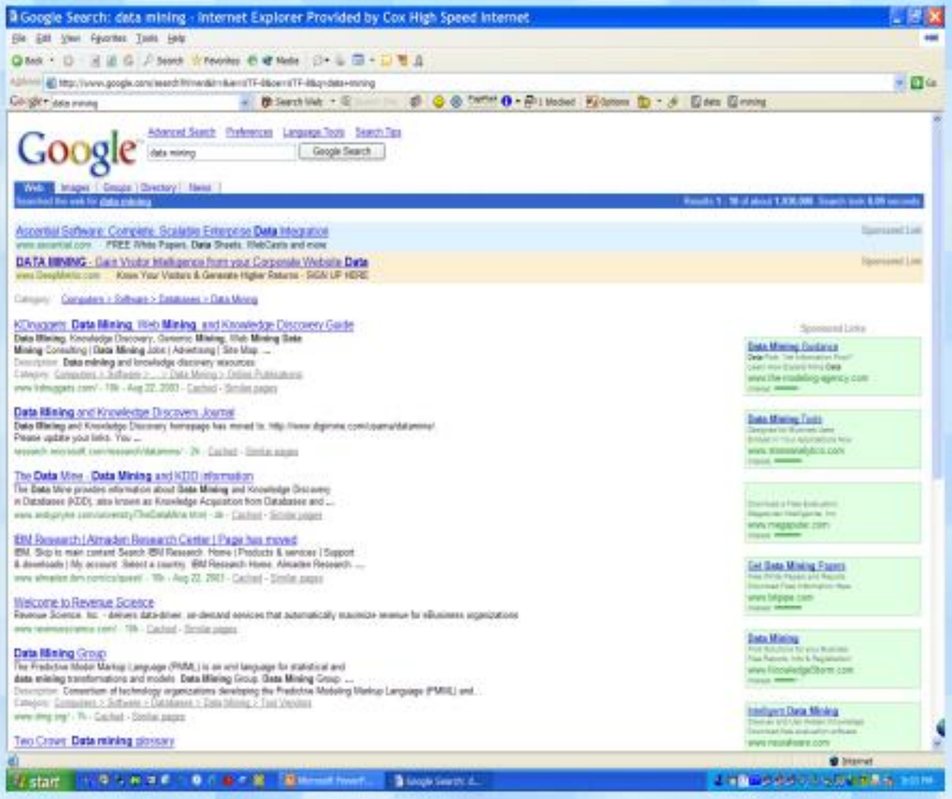
Data Mining is not …

- Generating a histogram of salaries for different age groups
- Issuing SQL query to a database, and reading the reply

Data Mining is not …

- Searching for a phone number in a phone book

- Searching for keywords on Google



Google Search: data mining - Internet Explorer Provided by Cox High Speed Internet
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.google.com/search?hl=en&ie=UTF-8&oe=UTF-8&q=data+mining
Go
Google data mining
Search Web
Options
data mining
Advanced Search Preferences Language Tools Search Tips
data mining
Google Search
Web Images Groups Directory News
Internet
start
Microsoft PowerP...
Google Search: d...

Searched the web for data mining. Results 1 - 10 of about 1,830,000. Search took 0.09 seconds.

Ascential Software: Complete, Scalable Enterprise Data Integration — Sponsored Link
www.ascential.com FREE White Papers, Data Sheets, WebCasts and more

DATA MINING - Gain Visitor Intelligence from your Corporate Website Data — Sponsored Link
www.DeepMetrix.com Know Your Visitors & Generate Higher Returns - SIGN UP HERE

Category: Computers > Software > Databases > Data Mining

KDnuggets: Data Mining, Web Mining, and Knowledge Discovery Guide
Data Mining, Knowledge Discovery, Genomic Mining, Web Mining Data Mining Consulting | Data Mining Jobs | Advertising | Site Map ...
Description: Data mining and knowledge discovery resources
Category: Computers > Software > ... > Data Mining > Online Publications
www.kdnuggets.com/ - 19k - Aug 22, 2003 - Cached - Similar pages

Data Mining and Knowledge Discovery Journal
Data Mining and Knowledge Discovery homepage has moved to: http://www.digimine.com/usama/datamine/ Please update your links. You ...
research.microsoft.com/research/datamine/ - 2k - Cached - Similar pages

The Data Mine - Data Mining and KDD information
The Data Mine provides information about Data Mining and Knowledge Discovery in Databases (KDD), also known as Knowledge Acquisition from Databases and ...
www.andypryke.com/university/TheDataMine.html - 3k - Cached - Similar pages

IBM Research | Almaden Research Center | Page has moved
IBM. Skip to main content Search IBM Research. Home | Products & services | Support & downloads | My account. Select a country. IBM Research Home. Almaden Research ...
www.almaden.ibm.com/cs/quest/ - 19k - Aug 22, 2003 - Cached - Similar pages

Welcome to Revenue Science
Revenue Science, Inc. - delivers data-driven, on-demand services that automatically maximize revenue for eBusiness organizations
www.revenuescience.com/ - 19k - Cached - Similar pages

Data Mining Group
The Predictive Model Markup Language (PMML) is an xml language for statistical and data mining transformations and models. Data Mining Group. Data Mining Group ...
Description: Consortium of technology organizations developing the Predictive Modeling Markup Language (PMML) and...
Category: Computers > Software > Databases > Data Mining > Tool Vendors
www.dmg.org/ - 7k - Cached - Similar pages

Two Crows: Data mining glossary

Sponsored Links

Data Mining Guidance
Data Rich, "Yet Information Poor?"
Learn How Experts Mine Data
www.the-modeling-agency.com

Data Mining Tools
Designed for Business Users
[illegible]
www.[illegible]analytics.com

Download a Free Evaluation
Megaputer Intelligence, Inc
www.megaputer.com

Get Data Mining Papers
Free White Papers and Reports
Download Free Information Now
www.bitpipe.com

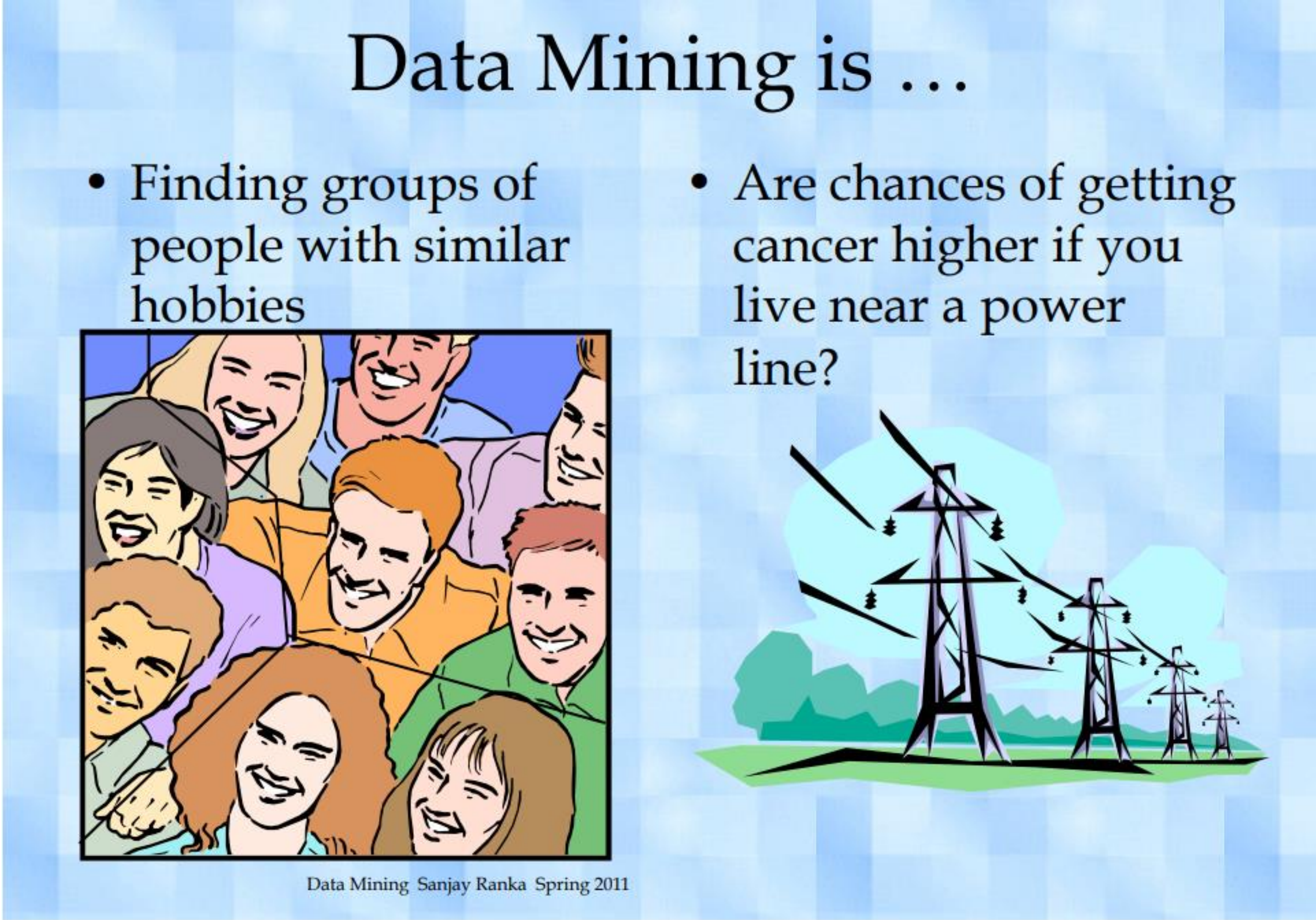
Data Mining
Find Solutions for your Business
Free Reports, Info & Registration
www.KnowledgeStorm.com

Intelligent Data Mining
Discover and Use Hidden Knowledge
Download free evaluation software
www.[illegible].com

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”

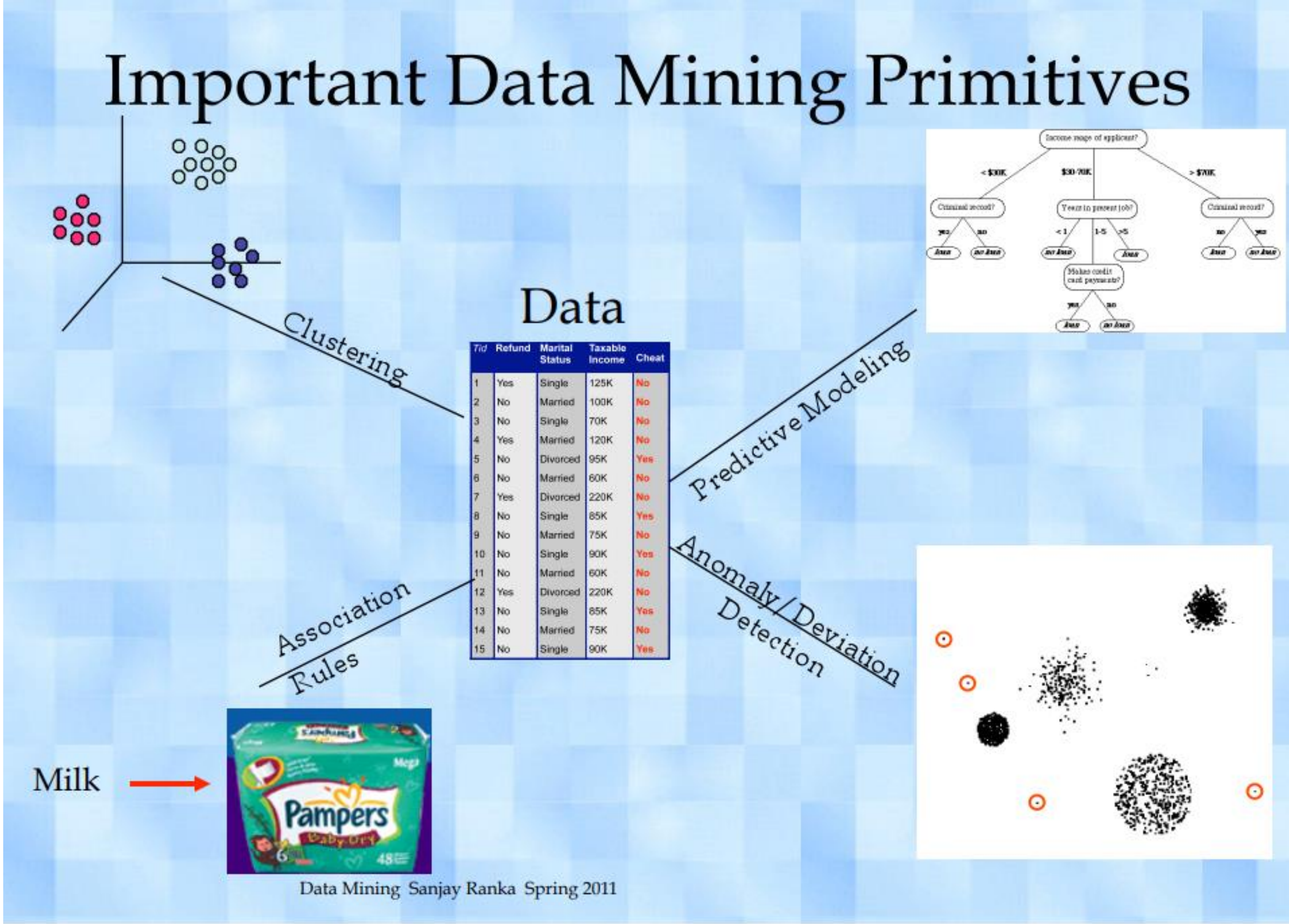
- **What is Data Mining?**
  - Certain names are more prevalent in certain US locations (O’Brien, O’Rurke, O’Reilly… in Boston area)
  - Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com,)



Data Mining is …

- Finding groups of people with similar hobbies
- Are chances of getting cancer higher if you live near a power line?

Data Mining Sanjay Ranka Spring 2011



Important Data Mining Primitives

Data

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
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes

Data Mining Tasks

- **Descriptive methods**
  - Find human-interpretable patterns that describe the data
    - **Example:** Clustering
- **Predictive methods**
  - Use some variables to predict unknown or future values of other variables
    - **Example:** Recommender systems

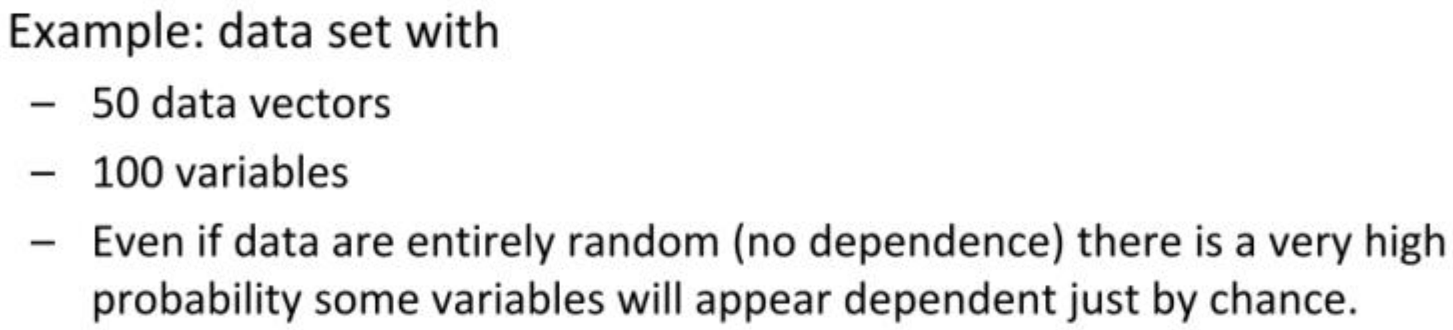
Meaningfulness of Analytic Answers

- **A risk with “Data mining” is that an analyst can “discover” patterns that are meaningless**
- Statisticians call it **Bonferroni’s principle**:
  - Roughly, if you look in more places for interesting patterns than your amount of data will support, you are bound to find crap

DOGBERT CONSULTS
YOU NEED TO DO DATA MINING TO UNCOVER HIDDEN SALES TRENDS.
www.dilbert.com scottadams@aol.com
IF YOU MINE THE DATA HARD ENOUGH, YOU CAN ALSO FIND MESSAGES FROM GOD.
© 2000 United Feature Syndicate, Inc. 4/3/00
...SALES TO LEFT-HANDED SQUIRRELS ARE UP...AND GOD SAYS YOUR TIE DOESN'T GO WITH THAT SHIRT.

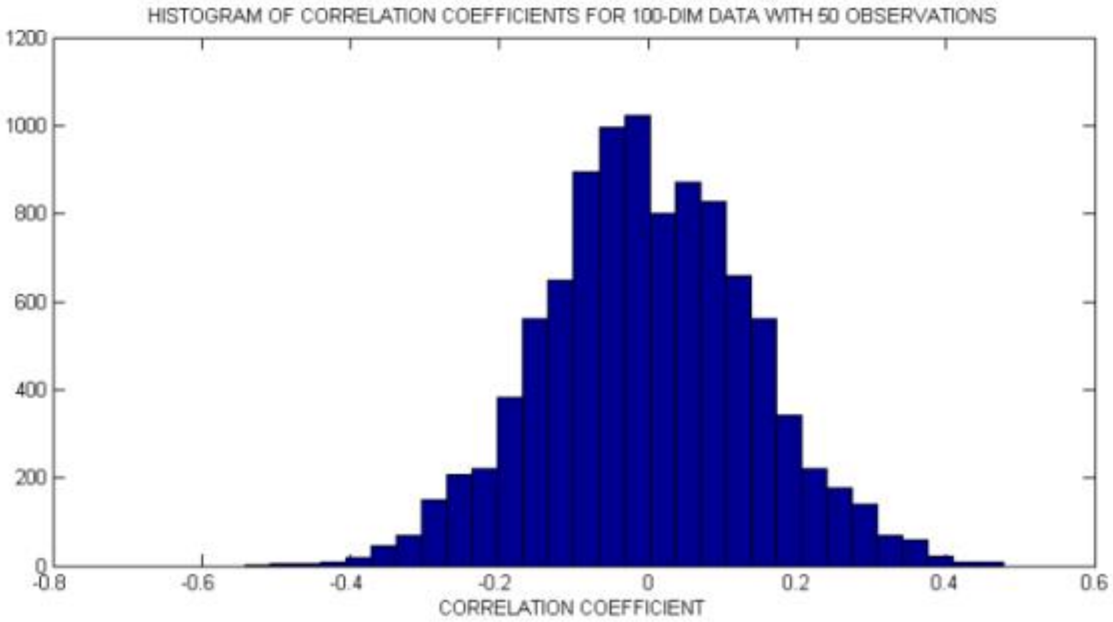
Example of “Data Fishing (Data Dredging)”

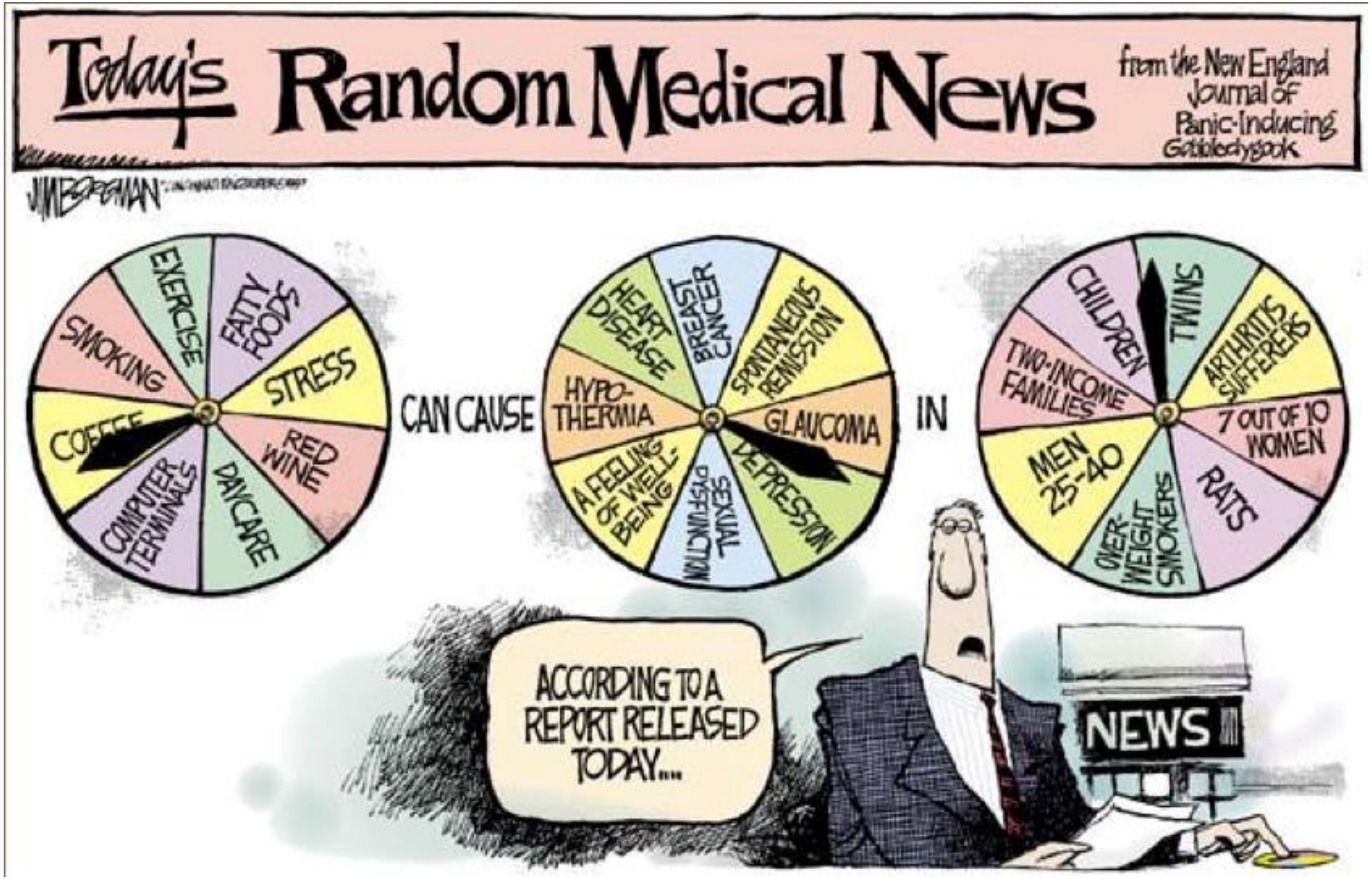
- seeking more information from a data set than it contains



Example: data set with

- 50 data vectors
- 100 variables
- Even if data are entirely random (no dependence) there is a very high probability some variables will appear dependent just by chance.



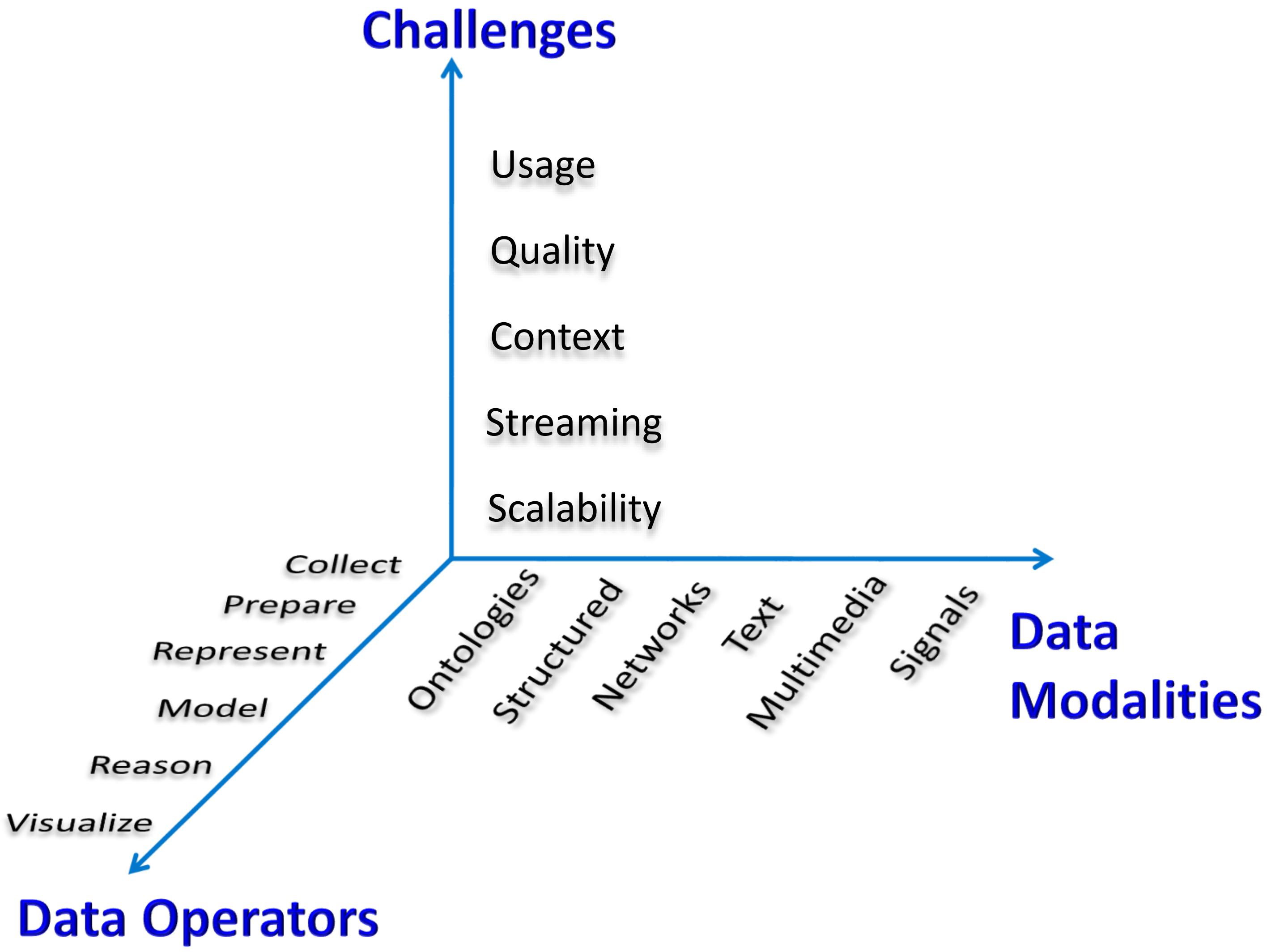

Today's Random Medical News
from the New England Journal of Panic-Inducing Gobbledygook
JIM BORGMAN
SMOKING
EXERCISE
FATTY FOODS
STRESS
RED WINE
DAYCARE
COMPUTER TERMINALS
COFFEE
CAN CAUSE
HEART DISEASE
BREAST CANCER
SPONTANEOUS REMISSION
GLAUCOMA
DEPRESSION
SEXUAL DYSFUNCTION
A FEELING OF WELL-BEING
HYPO-THERMIA
IN
CHILDREN
TWINS
ARTHRITIS SUFFERERS
7 OUT OF 10 WOMEN
RATS
OVER-WEIGHT SMOKERS
MEN 25-40
TWO-INCOME FAMILIES
ACCORDING TO A REPORT RELEASED TODAY...
NEWS

Meaningfulness of Analytic Answers

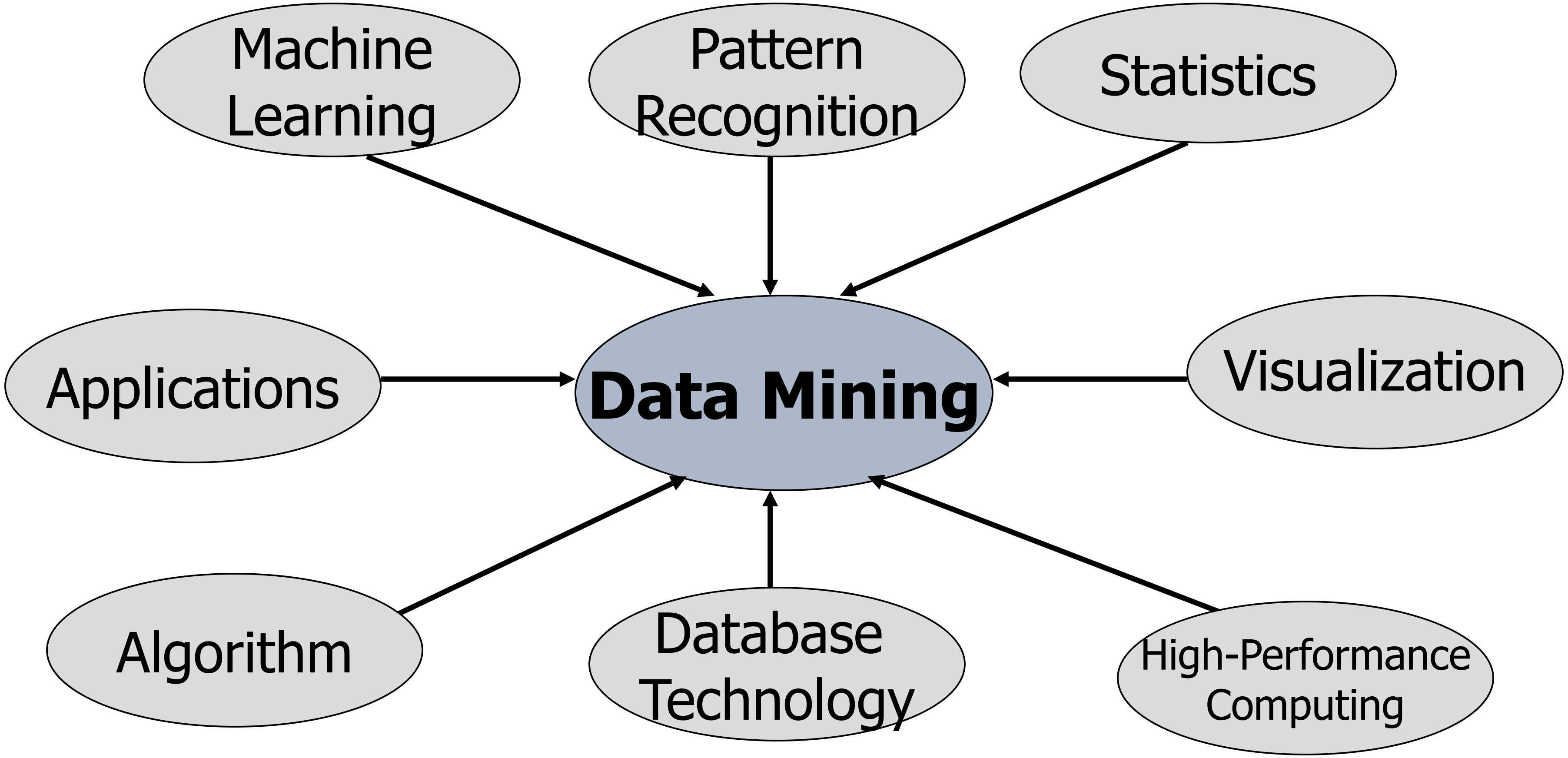
Example:

- We want to find (unrelated) people who **at least twice have stayed at the same hotel on the same day**
  - 10^9 people being tracked
  - 1,000 days
  - Each person stays in a hotel 1% of time (1 day out of 100)
  - Hotels hold 100 people (so 10^5 hotels)
  - **If everyone behaves randomly (i.e., no terrorists) will the data mining detect anything suspicious?**
- **Expected number of "suspicious" pairs of people:**
  - 250,000
  - … too many combinations to check – we need to have some additional evidence to find "suspicious" pairs of people in some more efficient way

What matters when dealing with data?



Data Mining: Confluence of Multiple Disciplines



Data Mining vs Statistics

- The goal is similar
- Different types of methods
- In data mining, one investigates lots of possible hypothesis
- Data mining is more exploratory data analysis
- In data mining, there are much larger datasets – algorithmics/scalability is an issue

Data mining vs Machine Learning

- Machine learning methods are used for data mining
  - Classification, clustering
- Amount of the data makes the difference
  - Data mining deals with much larger datasets and scalability becomes an issue
- Data mining has more modest goals
  - Automating various tedious tasks, not aiming at human performance in discovery
  - Helping users, not replacing them

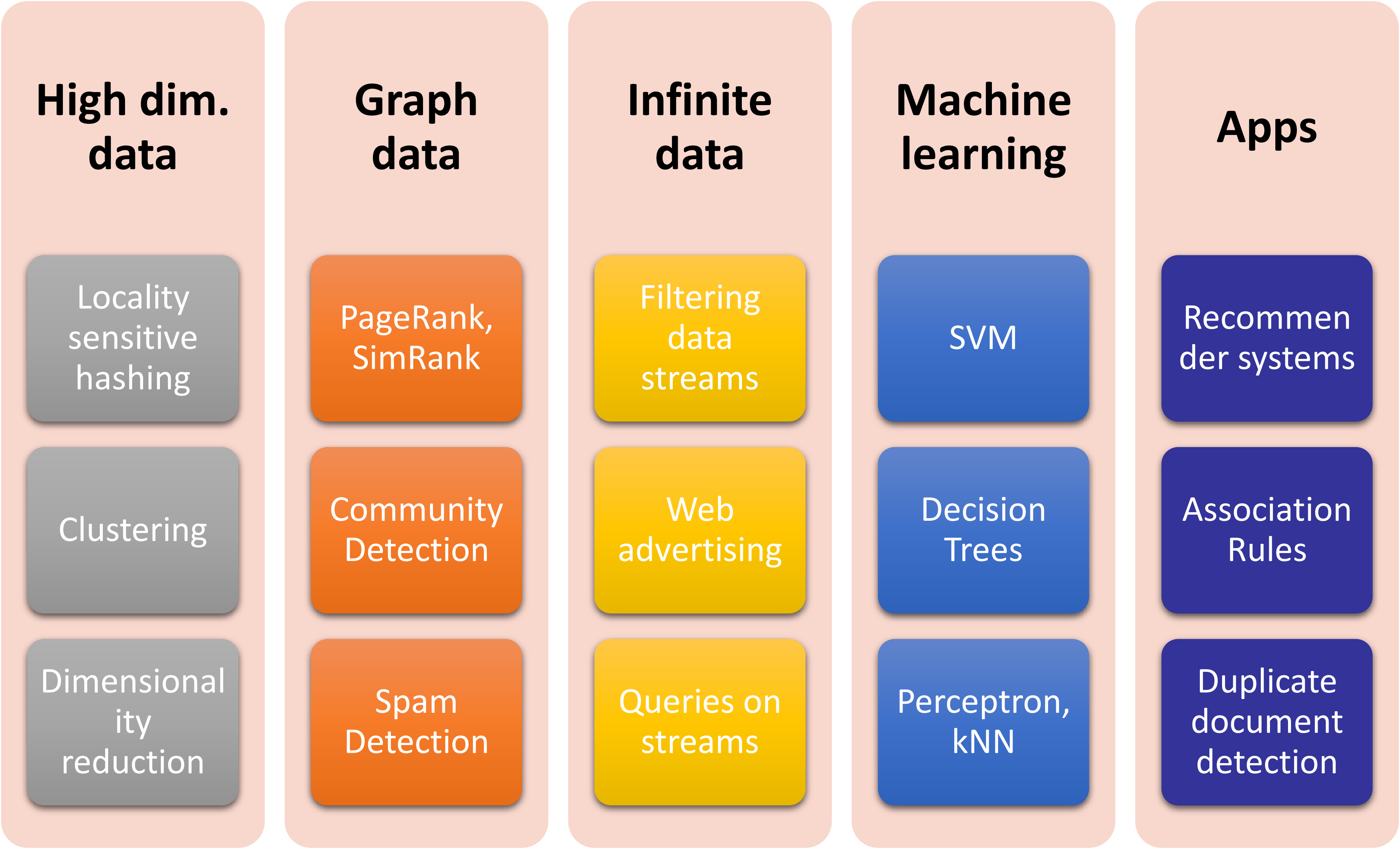
What can data-mining methods do?

- Rank web-query results
  - What are the most relevant web-pages to the query: “Student housing in Hacettepe”?
- Find groups of entities that are similar (clustering)
  - Find groups of facebook users that have similar friends/interests
  - Find groups of customers / amazon users that buy similar products
- Find good recommendations for users
  - Recommend facebook users new friends/groups
  - Recommend amazon customers new books

What will we learn?

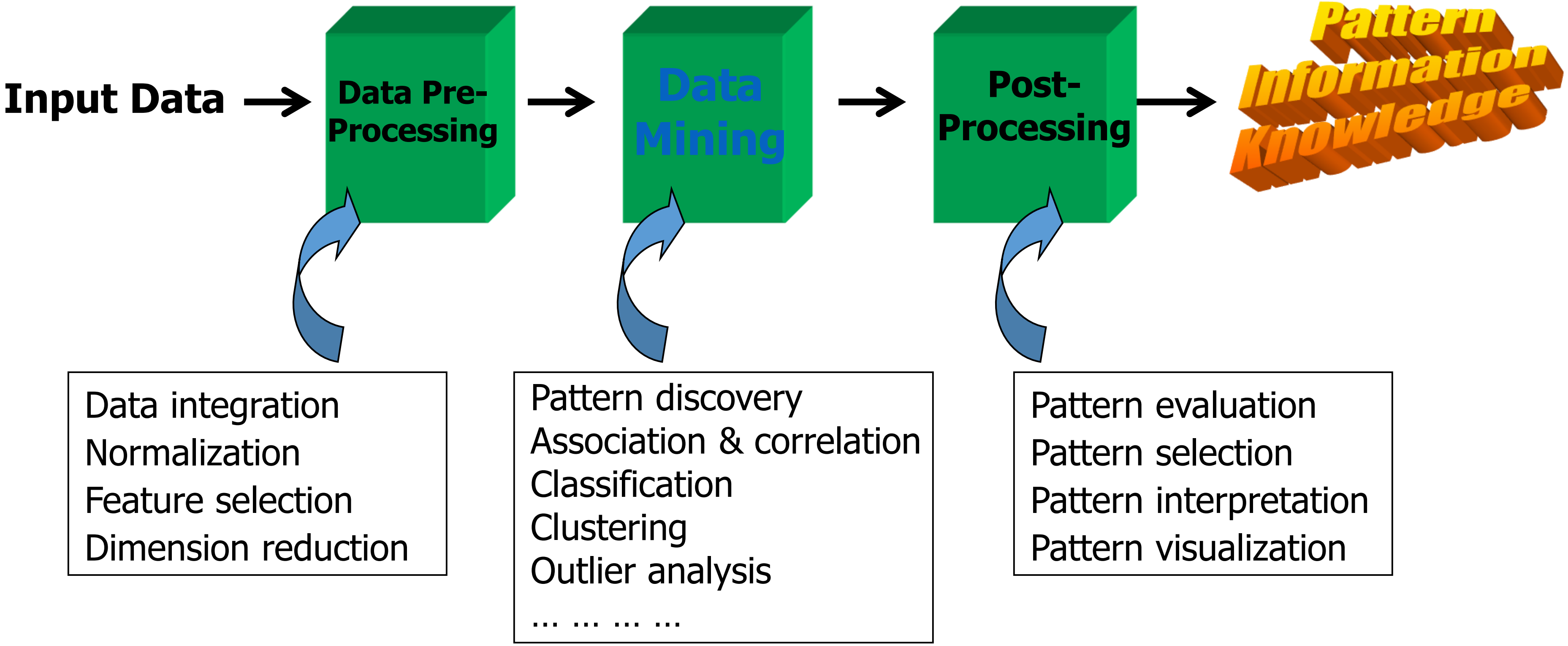
- **We will learn to mine different types of data:**
  - Data is high dimensional
  - Data is a graph
  - Data is infinite/never-ending
  - Data is labeled
- **We will learn to solve real-world problems:**
  - Recommender systems
  - Market Basket Analysis
  - Spam detection
  - Duplicate document detection

How It All Fits Together



High dim. data	Graph data	Infinite data	Machine learning	Apps
Locality sensitive hashing	PageRank, SimRank	Filtering data streams	SVM	Recommender systems
Clustering	Community Detection	Web advertising	Decision Trees	Association Rules
Dimensionality reduction	Spam Detection	Queries on streams	Perceptron, kNN	Duplicate document detection

KDD Process: A Typical View from ML and Statistics



Input Data → Data Pre-Processing → Data Mining → Post-Processing → Pattern Information Knowledge

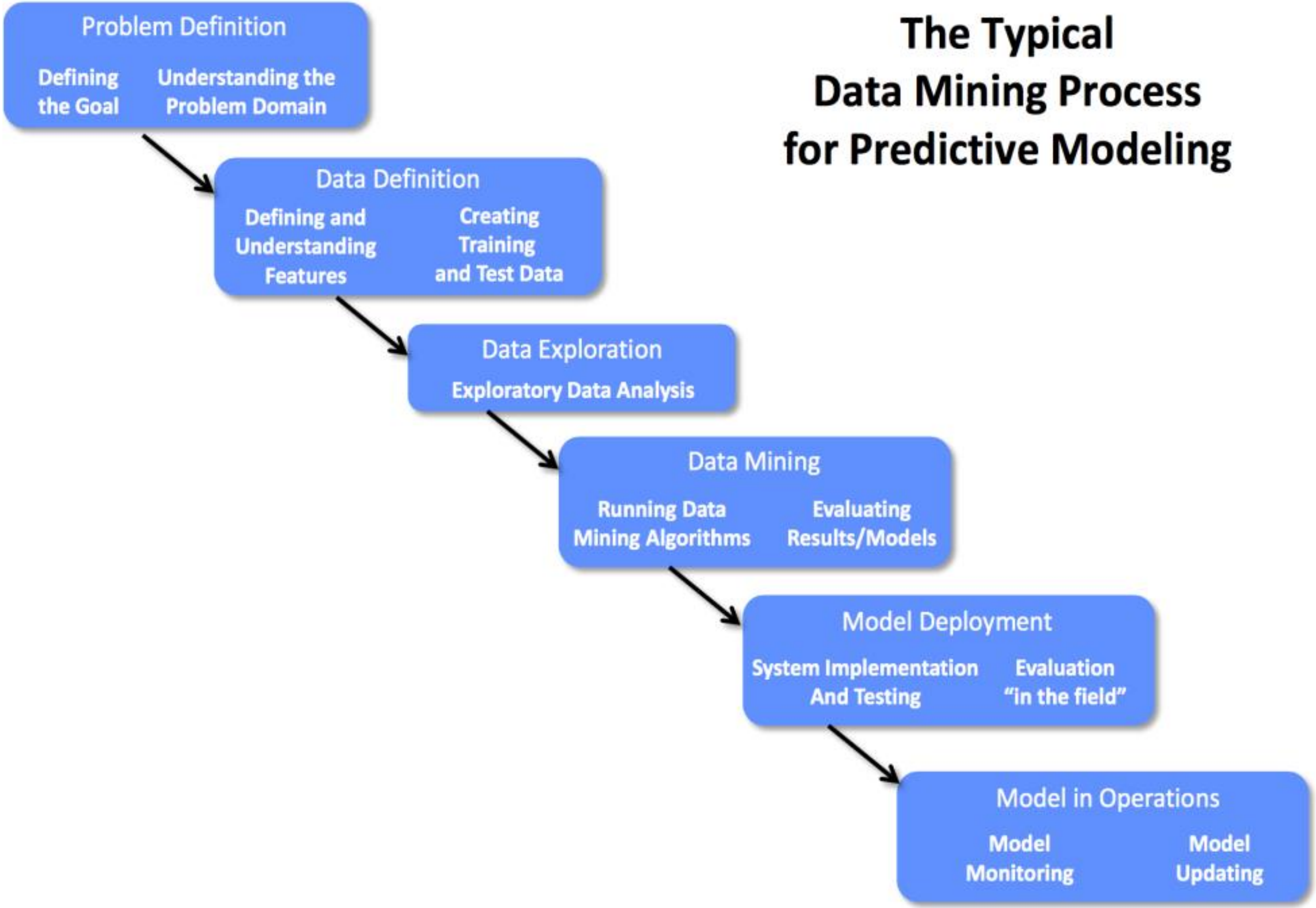
Data integration
Normalization
Feature selection
Dimension reduction

Pattern discovery
Association & correlation
Classification
Clustering
Outlier analysis
… … … …

Pattern evaluation
Pattern selection
Pattern interpretation
Pattern visualization

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

The Typical Data Mining Process for Predictive Modeling



Problem Definition

Defining the Goal

Understanding the Problem Domain

Data Definition

Defining and Understanding Features

Creating Training and Test Data

Data Exploration

Exploratory Data Analysis

Data Mining

Running Data Mining Algorithms

Evaluating Results/Models

Model Deployment

System Implementation And Testing

Evaluation "in the field"

Model in Operations

Model Monitoring

Model Updating

Data Mining Tasks...

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]

Classification

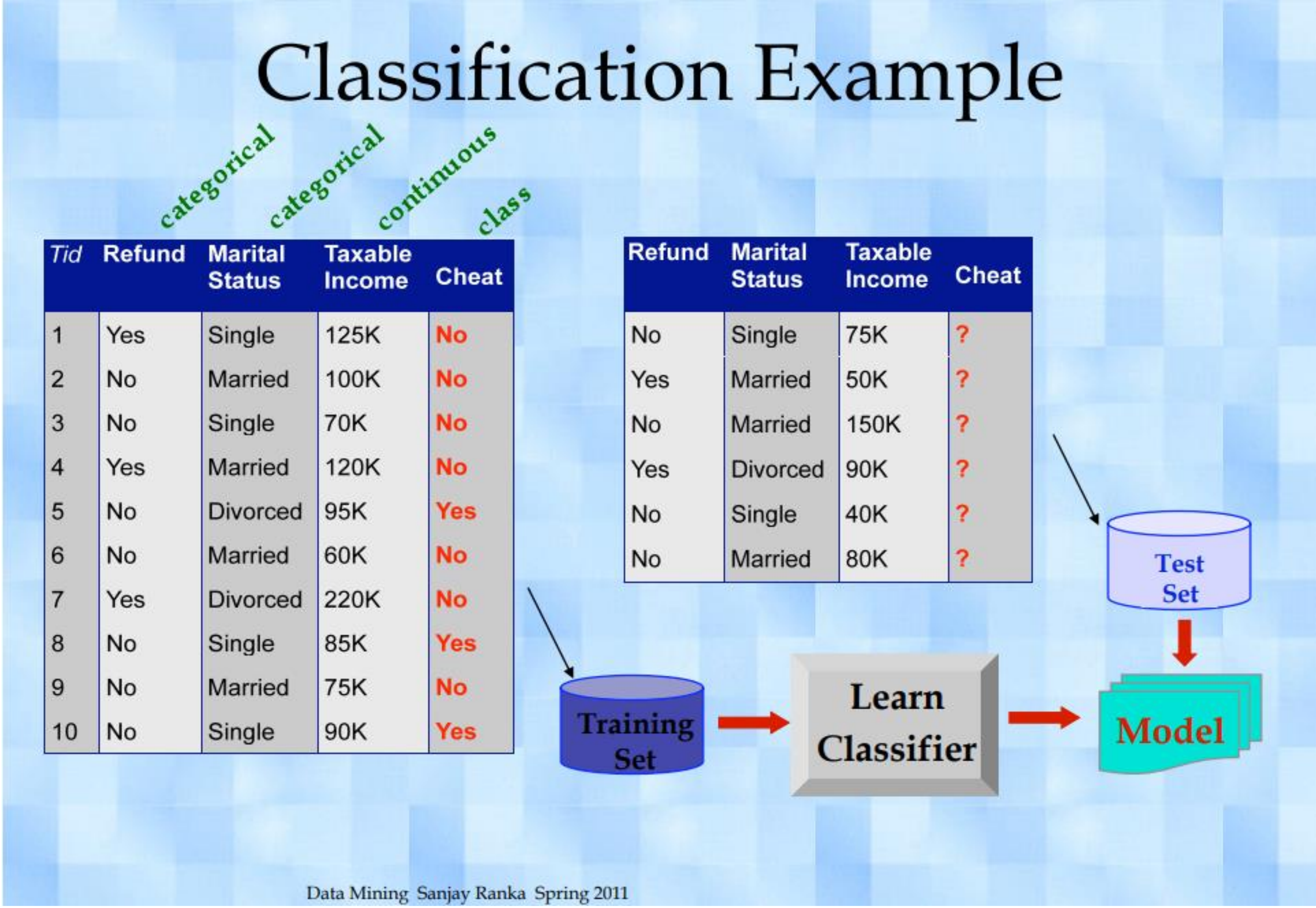
- Given a set of records (called the training set)

- Each record contains a set of attributes. One of the attributes is the class

- Find a model for the class attribute as a function of the values of other attributes

• Goal: Previously unseen records should be assigned to a class as accurately as possible

– Usually, the given data set is divided into training and test set, with training set used to build the model and test set used to validate it. The accuracy of the model is determined on the test set.



Classification Example

	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	?

Training Set → Learn Classifier → Model

Test Set → Model

Classification

- Fraud Detection
  - Goal: Predict fraudulent cases in credit card transactions.
  - Approach:
    - Use credit card transactions and the information on its 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.

Classification

Customer Churn

- Goal: To predict whether a customer is likely to be lost to a competitor
- Approach:

- Use detailed record of transaction with each of the past and current customers, to find attributes

> How often does the customer call, Where does he call, What time of the day does he call most, His financial status, His marital status, etc. (Important Information: Expiration of the current contract).

- Label the customers as {churn, not churn} – Find a model for Churn

Regression

- Predict the value of a given continuous valued variable based on the values of other variables, assuming a linear or non-linear model of dependency
- Extensively studied in the fields of Statistics and Neural Networks
- Examples

– Predicting sales numbers of a new product based on advertising expenditure

– Predicting wind velocities based on temperature, humidity, air pressure, etc

– Time series prediction of stock market indices

Clustering

- Market Segmentation
- Goal: To subdivide a market into distinct subset of customers where each subset can be targeted with a distinct marketing mix
- Approach:

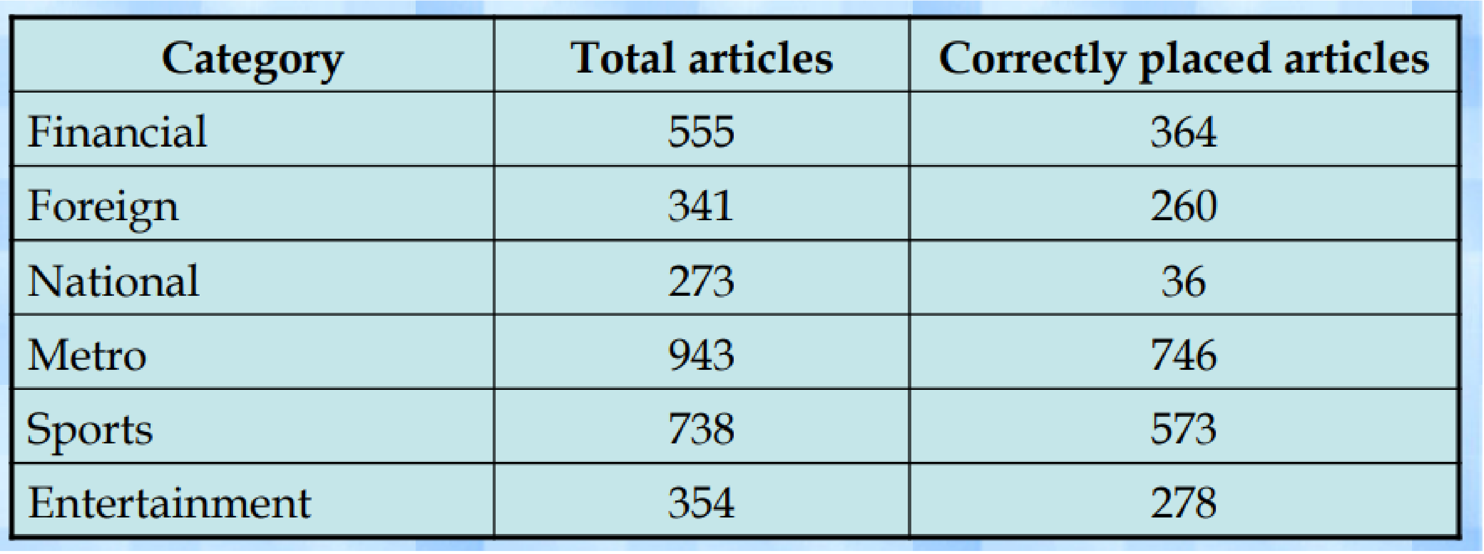
– Collect different attributes of customers based on their geographical and lifestyle related information

– Find clusters of similar customers

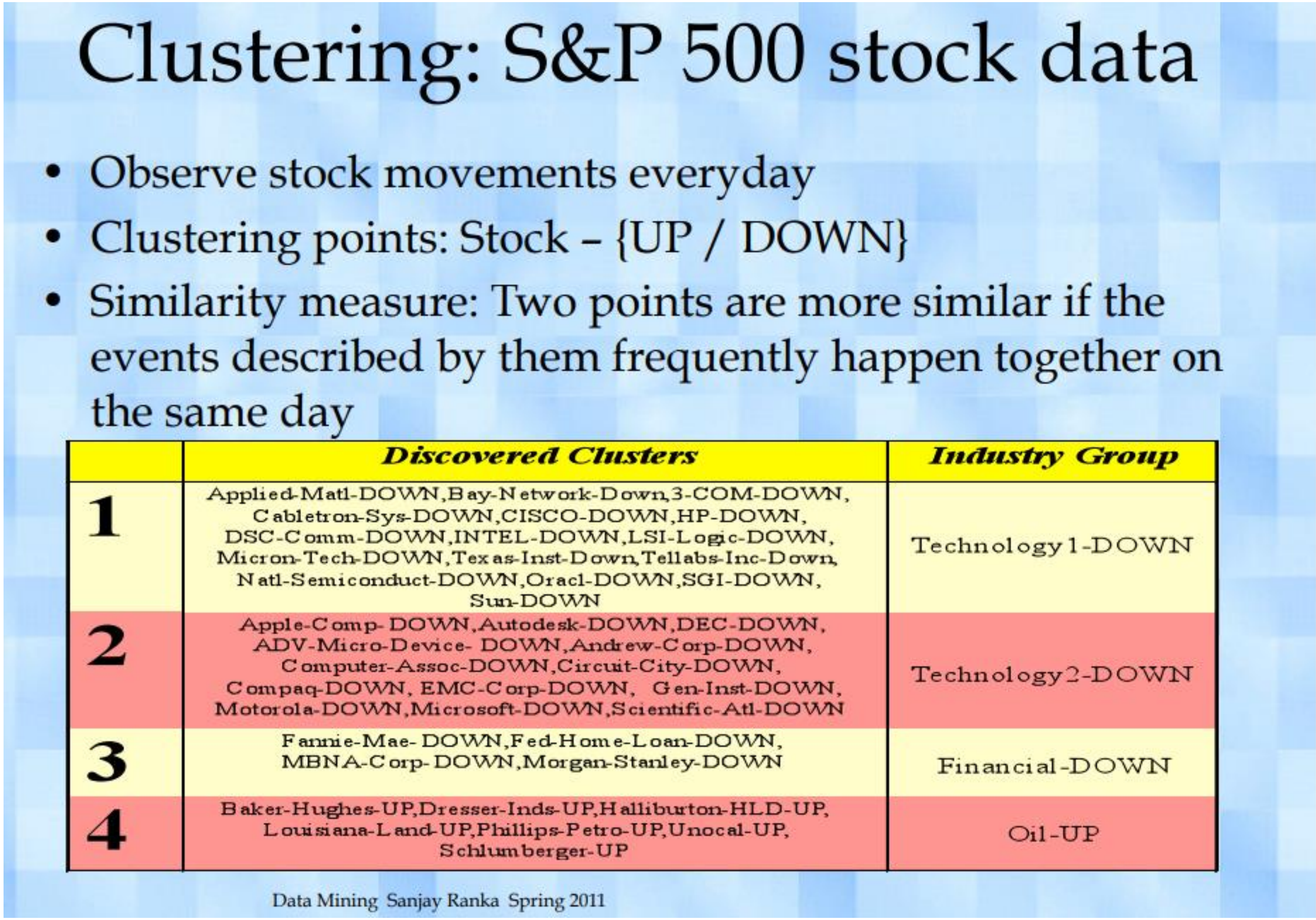
– Measure the clustering quality by observing the buying patterns of customers in the same cluster vs. those from different clusters

Clustering

- **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.



Category	Total articles	Correctly placed articles
Financial	555	364
Foreign	341	260
National	273	36
Metro	943	746
Sports	738	573
Entertainment	354	278



Clustering: S&P 500 stock data

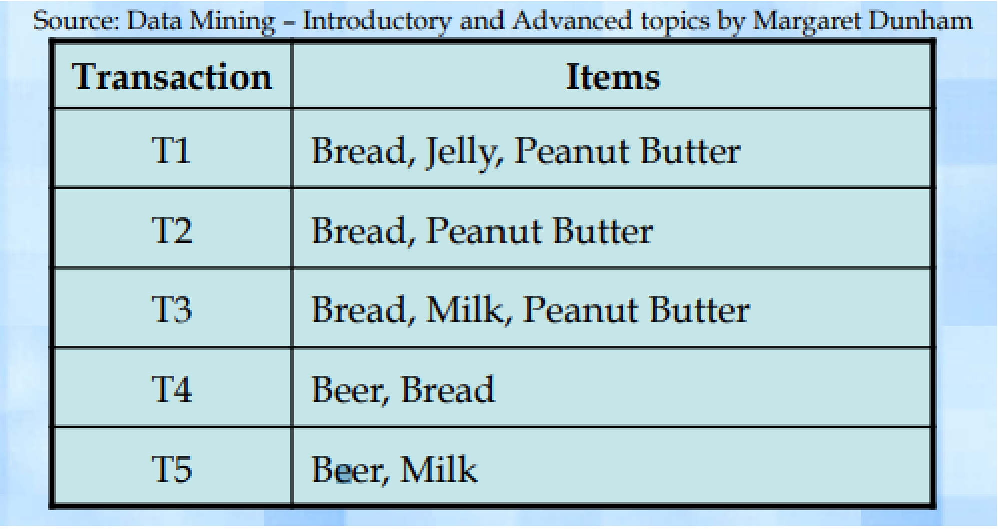
- Observe stock movements everyday
- Clustering points: Stock – {UP / DOWN}
- Similarity measure: Two points are more similar if the events described by them frequently happen together on the same day

	Discovered Clusters	*Industry Group*
1	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device- DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae- DOWN,Fed-Home-Loan-DOWN, MBNA-Corp- DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP	Oil-UP

Data Mining Sanjay Ranka Spring 2011

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.



Source: Data Mining – Introductory and Advanced topics by Margaret Dunham

Transaction	Items
T1	Bread, Jelly, Peanut Butter
T2	Bread, Peanut Butter
T3	Bread, Milk, Peanut Butter
T4	Beer, Bread
T5	Beer, Milk

- Some rules discovered –
- Bread -> Peanut Butter
- Peanut Butter -> Bread
- Jelly -> Peanut Butter

Association Rule Discovery: Super market shelf management

- Goal: To identify items that are bought concomitantly by a reasonable fraction of customers so that they can be shelved appropriately based on business goals.
- Data Used: Point-of-sale data collected with barcode scanners to find dependencies among products
- Example

– If a customer buys Jelly, then he is very likely to buy Peanut Butter.

– So don't be surprised if you find Peanut Butter next to Jelly on an aisle in the super market. Also, salsa next to tortilla chips.

Sequential Pattern Discovery: Definition

- • Given is a set of objects, with each object associated with its own timeline of events, find rules that predict strong sequential dependencies among different events

- Telecommunication alarm logs

(Inverter_Problem Excessive_Line_Current) (Rectifier_Alarm) -> (Fire_Alarm)

- Point of sale transaction sequences – Computer bookstore

(Intro_to_Visual_C) (C++ Primer) -> (Perl_For_Dummies, Tcl_Tk)
Athletic apparel store •(Shoes) (Racket, Racket ball) -> (Sports_Jacket)

Deviation / Anomaly Detection

- • Some data objects do not comply with the general behavior or model of the data. Data objects that are different from or inconsistent with the remaining set are called outliers
- Outliers can be caused by measurement or execution error. Or they represent some kind of fraudulent activity.
- Goal of Deviation / Anomaly Detection is to detect significant deviations from normal behavior

Deviation: Credit Card Fraud Detection

- • Goal: To detect fraudulent credit card transactions
- Approach:
- Based on past usage patterns, develop model for authorized credit card transactions
- Check for deviation form model, before authenticating new credit card transactions
- Hold payment and verify authenticity of “doubtful” transactions by other means (phone call, etc.)

Structure and Network Analysis

- **Graph mining**
  - Finding frequent subgraphs (e.g., chemical compounds), trees (XML), substructures (web fragments)
- **Information network analysis**
  - Social networks: actors (objects, nodes) and relationships (edges)
    - e.g., author networks in CS, terrorist networks
  - Multiple heterogeneous networks
    - A person could be multiple information networks: friends, family, classmates, …
  - Links carry a lot of semantic information: Link mining
- **Web mining**
  - Web is a big information network: from PageRank to Google
  - Analysis of Web information networks
    - Web community discovery, opinion mining, usage mining, …

Major Challenges in Data Mining

- Efficiency and scalability of data mining algorithms
- Parallel, distributed, stream, and incremental mining methods
- Handling high-dimensionality
- Handling noise, uncertainty, and incompleteness of data
- Incorporation of constraints, expert knowledge, and background knowledge in data mining
- Pattern evaluation and knowledge integration

Major Challenges in Data Mining

- Mining diverse and heterogeneous kinds of data: e.g., bioinformatics, Web, software/system engineering, information networks
- Application-oriented and domain-specific data mining
- Invisible data mining (embedded in other functional modules)
- Protection of security, integrity, and privacy in data mining

Conferences and Journals on Data Mining

- KDD Conferences
  - ACM SIGKDD Int. Conf. on Knowledge Discovery in Databases and Data Mining (KDD)
  - SIAM Data Mining Conf. (SDM)
  - (IEEE) Int. Conf. on Data Mining (ICDM)
  - Conf. on Principles and practices of Knowledge Discovery and Data Mining (PKDD)
  - Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD)

- Other related conferences
  - ACM SIGMOD
  - VLDB
  - (IEEE) ICDE
  - WWW, SIGIR
  - ICML, CVPR, NIPS
- Journals
  - Data Mining and Knowledge Discovery (DAMI or DMKD)
  - IEEE Trans. On Knowledge and Data Eng. (TKDE)
  - KDD Explorations
  - ACM Trans. on KDD

Topics to be covered (tentative)

- Introduction to data mining
- Data preprocessing
- Finding similar entities
- Clustering
- Classification
- Frequent pattern mining
    - Frequent itemsets and association rules
- Sequence Mining
    - Time-series data
- Link analysis ranking
- Applications
    - Recommendation systems, etc.

Materials

- **Books:**
  - **P.-N. Tan, M. Steinbach, V. Kumar: Introduction to Data Mining. Addison-Wesley, 2006.**
  - **A. Rajaraman and J. Ullman: Mining of Massive Datasets. Cambridge University Press, 2012.**
  - Jiawer Han and Micheline Kamber: Data Mining: Concepts and Techiques. Second Edition. Morgan Kaufmann Publishers, March 2006
- Research papers (pointers will be provided)

Grading

- Exam 40%
- Homeworks and Project 60%
- Attendance and participation are required