LECTURE 1: INTRODUCTION TO DATA MINING

Dr. Dhaval Patel

CSE, IIT-Roorkee

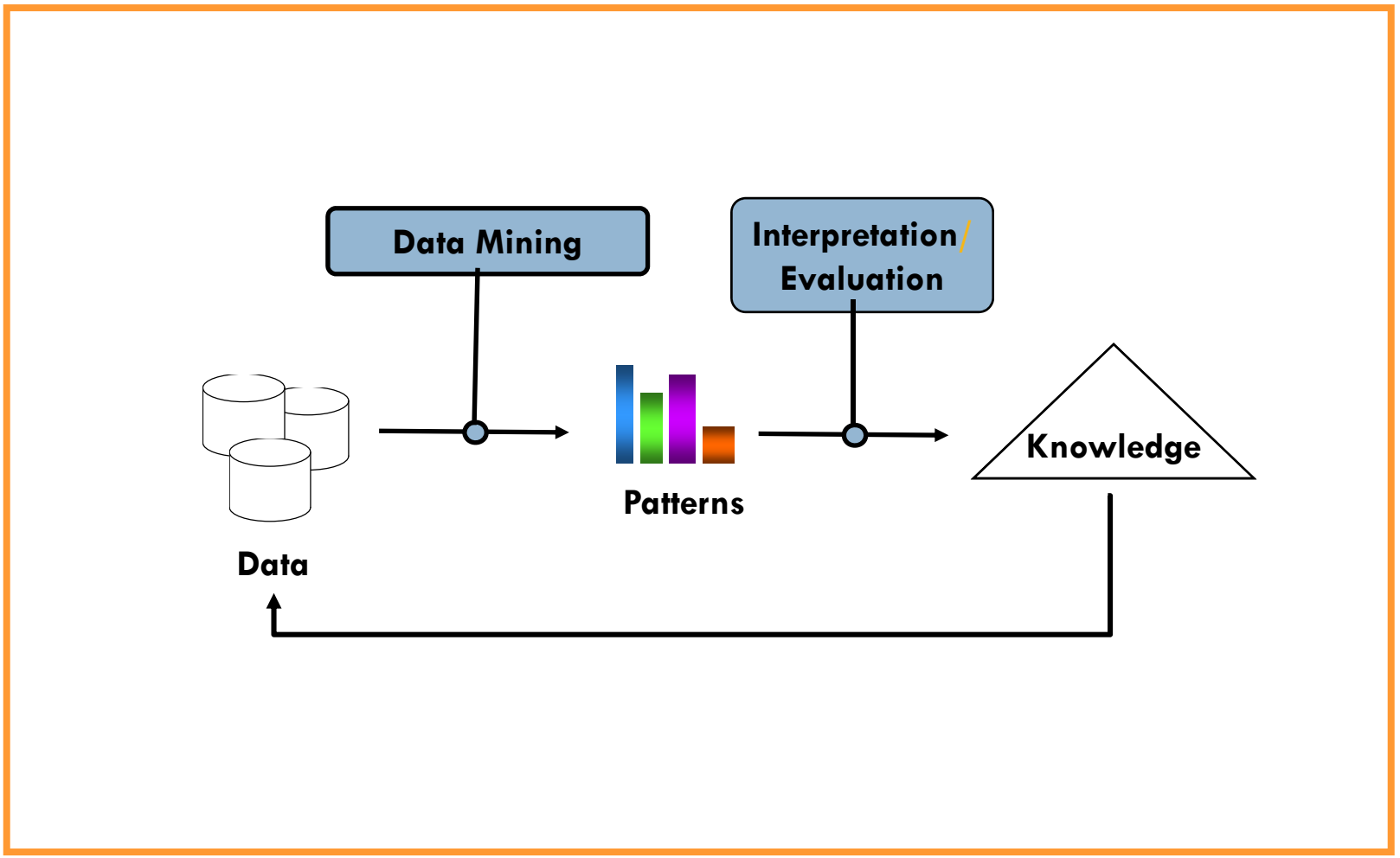
What is data mining?

- Data mining is also called *knowledge discovery and data mining* (KDD)

- Data mining is
  - extraction of useful patterns from **data sources**, e.g., databases, texts, web, image.

- Patterns must be:
  - valid, novel, potentially useful, understandable

Knowledge Discovery in Data: Process

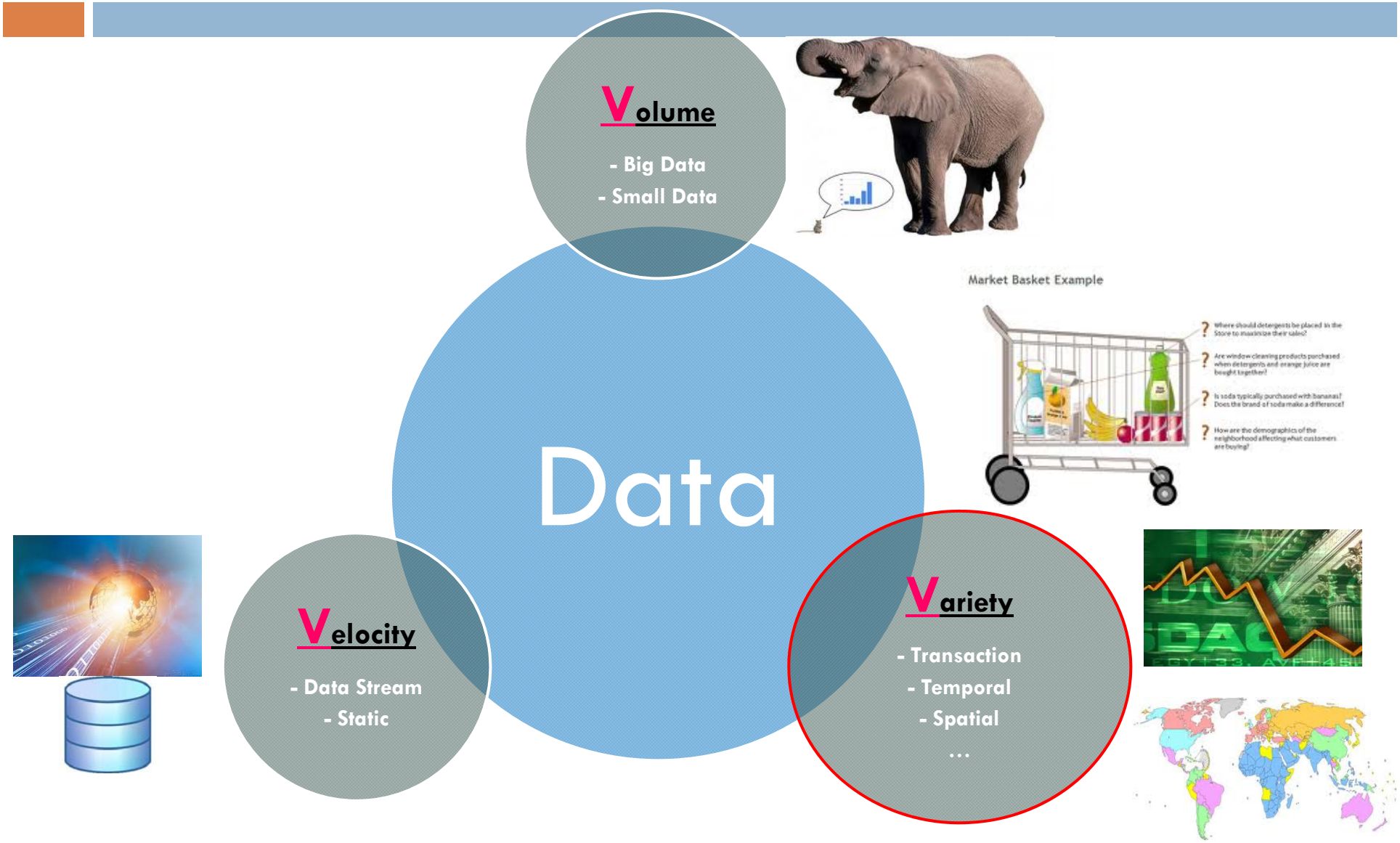


Knowledge Discovery in Data: Process

Example 1: Analysis of purchases in a supermarket

customer1	pizza	beer	cheese	bread	chips
customer2	milk	bread	ham	cigaretts	
customer3	yoghurt	sugar	flour	cornflakes	napkins
customer4	shampoo	beer	chips	newspaper	pizza
xustomer5	chips	coffee	beer	pizza	cream
customer6	jam	rolls	butter	beer	
...					

Knowledge Discovery in Data: Challenges



Outline (Part 1)

- Introduction to Data
  - Transactional Data
  - Temporal Data
  - Spatial & Spatial-Temporal Data
- Data Preprocessing
  - Missing Values
  - Summarization

INTRODUCTION TO DATA

Data Come from Everywhere

But, they have different form

Hospital

Weather Station

Social Media

What is Data?

- Collection of records and their attributes
- An attribute is a characteristic of an object
- A collection of attributes describe an object

Attributes

Objects

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

Types of Data

- Record Data
  - Transactional Data
- Temporal Data
  - Time Series Data
  - Sequence Data
- Spatial & Spatial-Temporal Data
  - Spatial Data
  - Spatial-Temporal Data
- Graph Data
  - Transactional Data
- UnStructured Data
  - Twitter Status Message
  - Review, news article
- Semi-Structured Data
  - Paper Publications Data
  - XML format

Record Data

- Transaction Data

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

Market-Basket Dataset

Data Matrix

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute

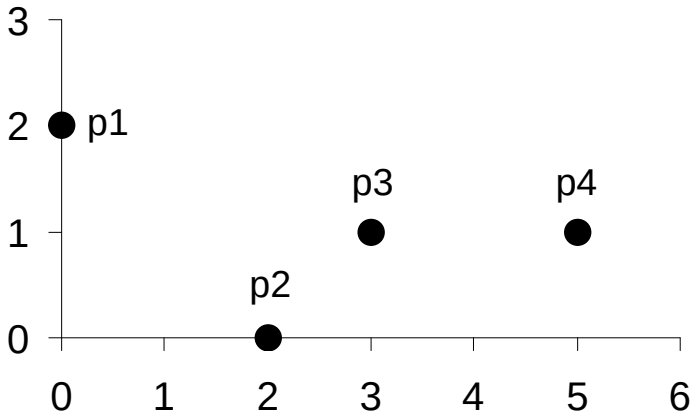
- Such data set can be represented by an m by n matrix, where there are m rows, one for each object, and n columns, one for each attribute

Data Matrix Example for Documents

- Each document becomes a `term' vector,
  - each term is a component (attribute) of the vector,
  - the value of each component is the number of times the corresponding term occurs in the document.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Distance Matrix



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distance Matrix

Temporal Data

- Sequences Data

ID	Symptom Sequence
1	{Night sweat, hypodynamia } →Fever →Achroacytosis→Anemia
2	Night sweat→Fever→Achroacytosis→Anemia
3	Night sweat→Fever→Achroacytosis→Anemia
4	Night sweat→Fever→Achroacytosis→Splenomegalia
5	Night sweat→Fever→Achroacytosis
6	Night sweat→Fever→Anemia
7	Night sweat→Splenomegalia→Anemia
8	Night sweat→Sleepy→Anemia

(Patient Data obtained from Zhang's KDD 06 Paper)

Temporal Data

- Time Series Data

Yahoo Finance Website

Biological Sequence Data

Are Rhesus Monkeys or Gibbons More Closely Related to Humans? DNA and polypeptide sequences from closely related species are more similar to each other than are sequences from more distantly related species. In this exercise, you will look at amino acid sequence data for the β polypeptide chain of hemoglobin, often called β-globin. You will then interpret the data to hypothesize whether the monkey or the gibbon is more closely related to humans.

How Such Experiments Are Done Researchers can isolate the polypeptide of interest from an organism and then determine the amino acid sequence. More frequently, the DNA of the relevant gene is sequenced, and the amino acid sequence of the polypeptide is deduced from the DNA sequence of its gene.

Data from the Experiments In the data below, the letters give the sequence of the 146 amino acids in β-globin from humans, rhesus monkeys, and gibbons. Because a complete sequence would not fit on one line here, the sequences are broken into three segments. The sequences for the three different species are aligned so that you can compare them easily. For example, you can see that for all three species, the first amino acid is V (valine) and the 146th amino acid is H (histidine).

Species	Alignment of Amino Acid Sequences of β-globin				
Human	1	VHLTPEEKSA	VTALWGKVNV	DEVGGEALGR	LLVVYPWTQR	FFESFGDLST
Monkey	1	VHLTPEEKNA	VTTLWGKVNV	DEVGGEALGR	LLLVYPWTQR	FFESFGDLSS
Gibbon	1	VHLTPEEKSA	VTALWGKVNV	DEVGGEALGR	LLVVYPWTQR	FFESFGDLST
Human	51	PDAVMGNPKV	KAHGKKVLGA	FSDGLAHLDN	LKGTFATLSE	LHCDKLHVDP
Monkey	51	PDAVMGNPKV	KAHGKKVLGA	FSDGLNHLDN	LKGTFAQLSE	LHCDKLHVDP
Gibbon	51	PDAVMGNPKV	KAHGKKVLGA	FSDGLAHLDN	LKGTFAQLSE	LHCDKLHVDP
Human	101	ENFRLLGNVL	VCVLAHHFGK	EFTPPVQAAY	QKVVAGVANA	LAHKYH
Monkey	101	ENFKLLGNVL	VCVLAHHFGK	EFTPQVQAAY	QKVVAGVANA	LAHKYH
Gibbon	101	ENFRLLGNVL	VCVLAHHFGK	EFTPQVQAAY	QKVVAGVANA	LAHKYH

Interpret the Data

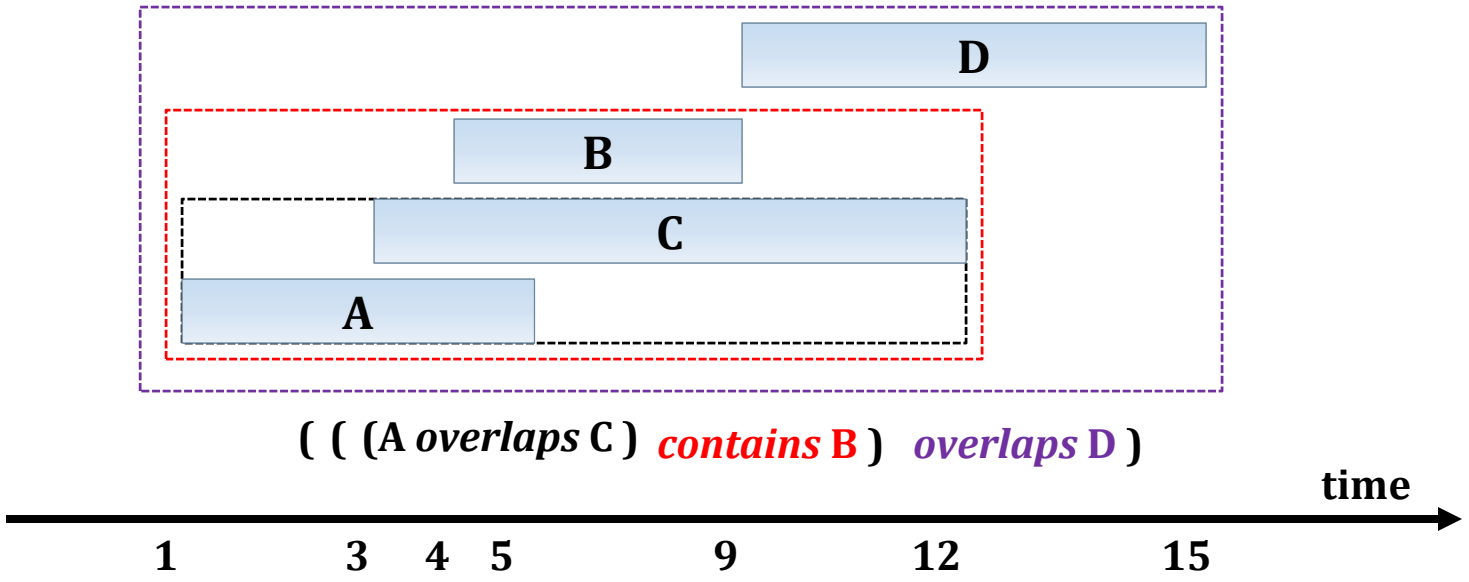
1. Scan the monkey and gibbon sequences, letter by letter, circling any amino acids that do not match the human sequence. (a) How many amino acids differ between the monkey and the human sequences? (b) Between the gibbon and human?
2. For each nonhuman species, what percent of its amino acids are identical to the human sequence of β-globin?
3. Based on these data alone, state a hypothesis for which of these two species is more closely related to humans. What is your reasoning?
4. What other evidence could you use to suppor your hypothesis?

MB A version of this Scientific Skills Exercise can be assigned in MasteringBiology.

Data from Human: http://www.ncbi.nlm.nih.gov/protein/AAA21113.1; rhesus monkey: http://www.ncbi.nlm.nih.gov/protein/122634; gibbon: http://www.ncbi.nlm.nih.gov/protein/122616

Interval Data

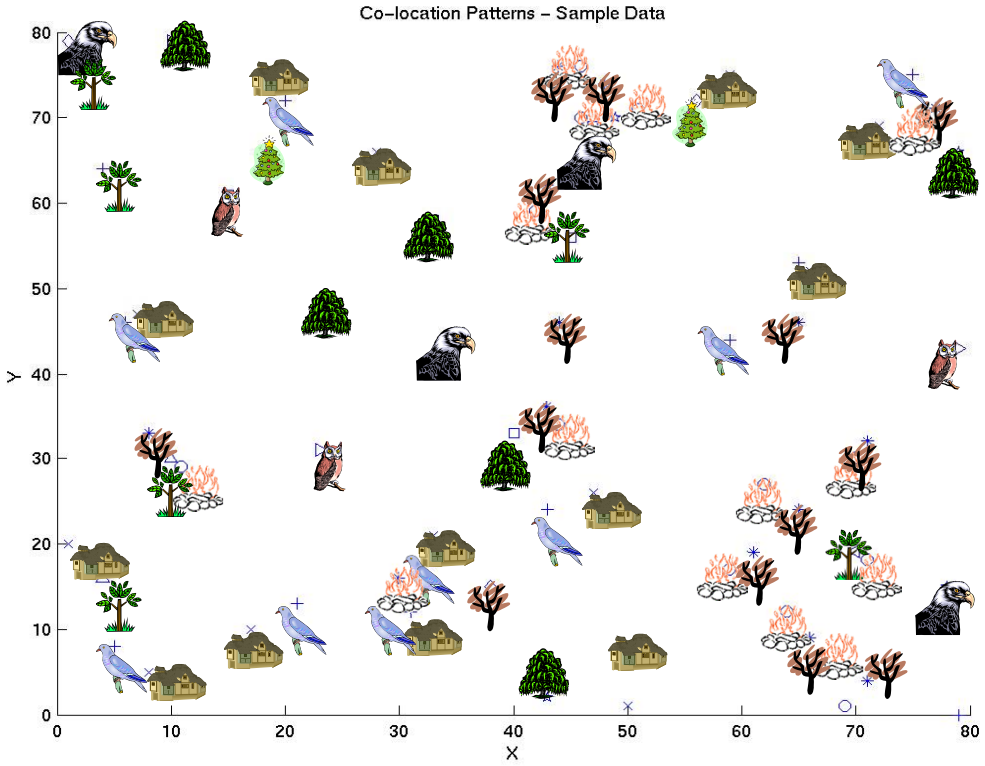
EL= { (A, 1, 5),(C, 3, 12), (B, 4, 9), (D, 9, 15) }



(Interval Patient Data obtained from Amit's M.Tech. Thesis Work)

Spatial & Spatial-Temporal Data

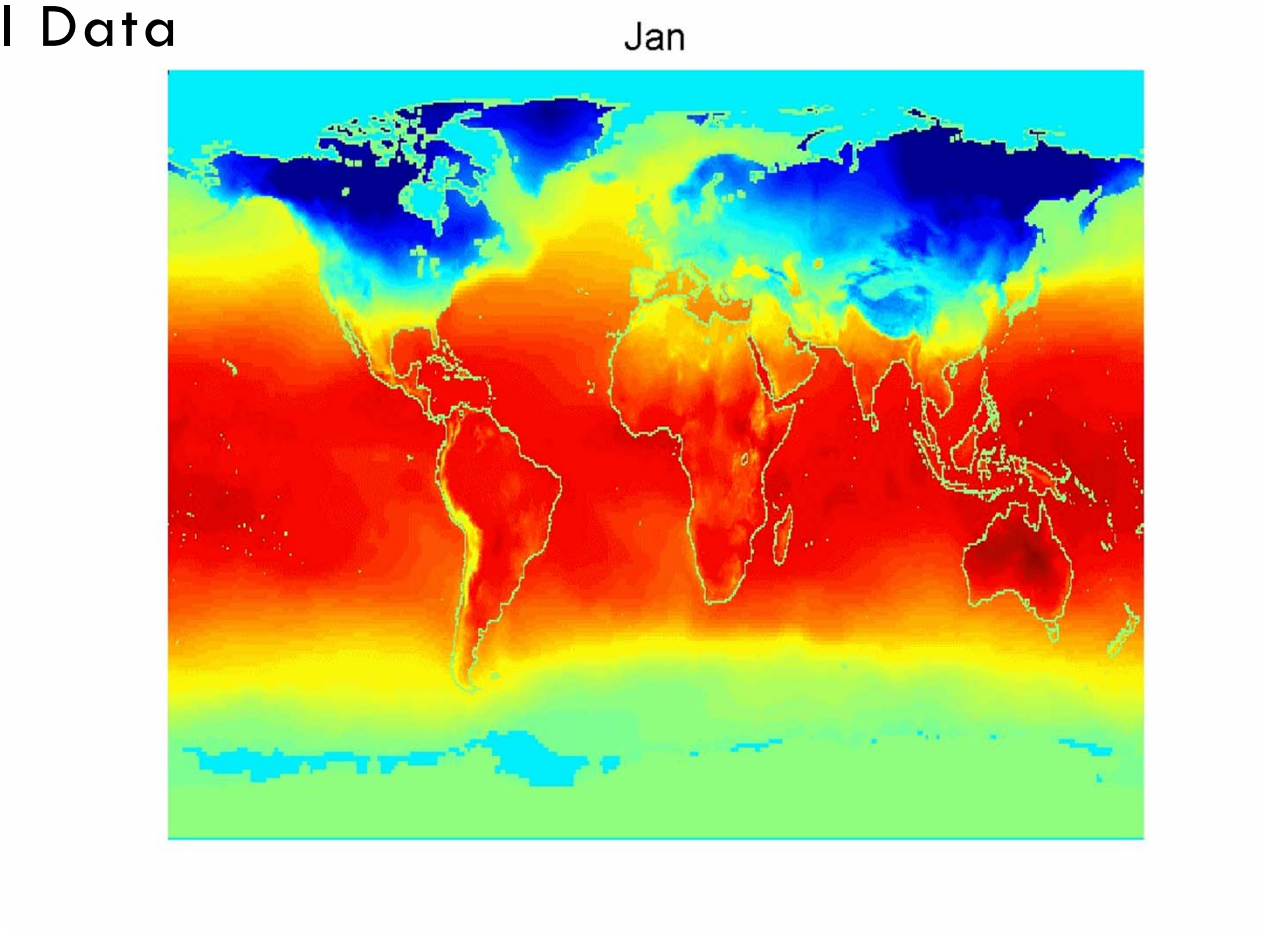
- Spatial Data



(Spatial Distribution of Objects of Various Types : Prof. Shashi Shekhar)

Spatial & Spatial-Temporal Data

- Spatial Data



Average Monthly Temperature of land and ocean

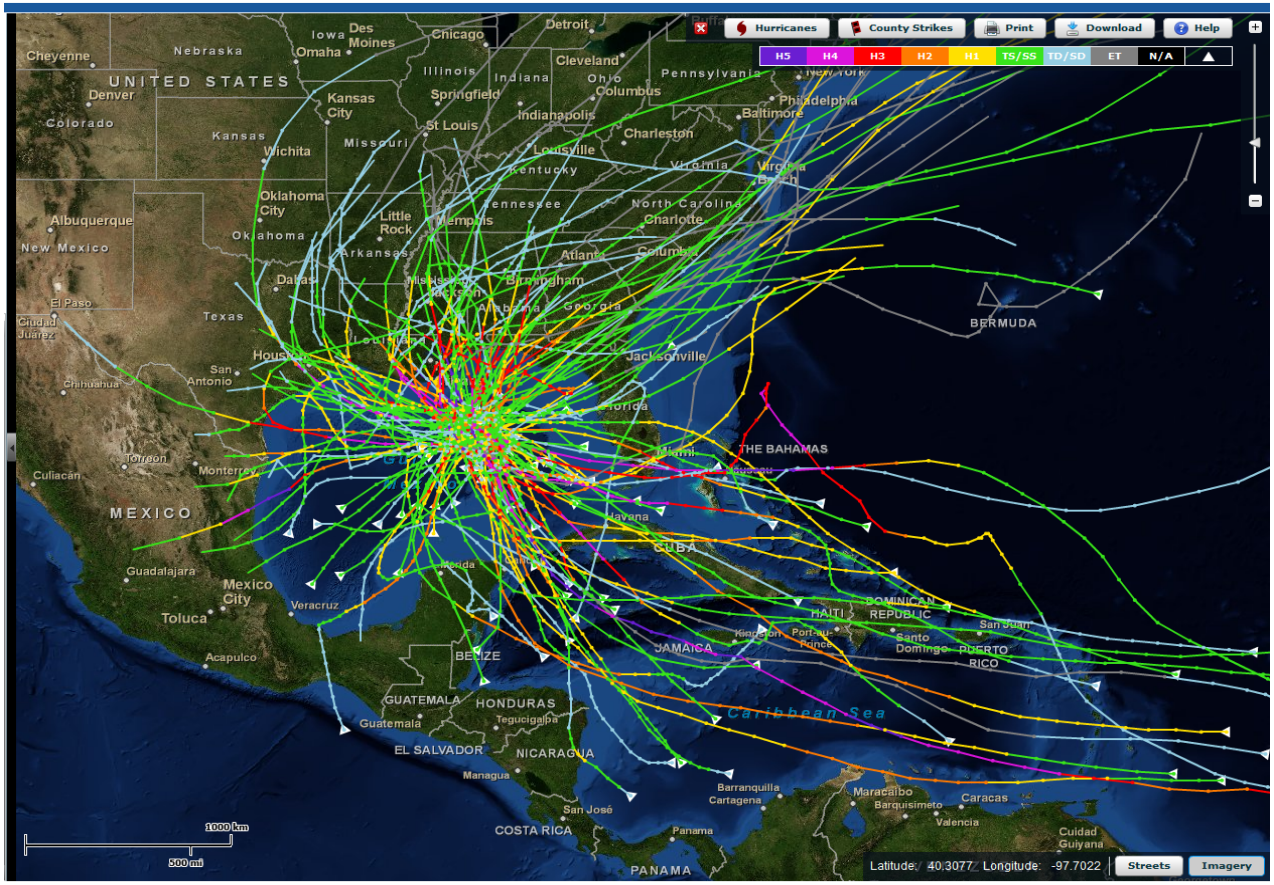
Spatial & Spatial-Temporal Data

- Spatial Data

Dengue Disease Dataset (Singapore)

Spatial & Spatial-Temporal Data

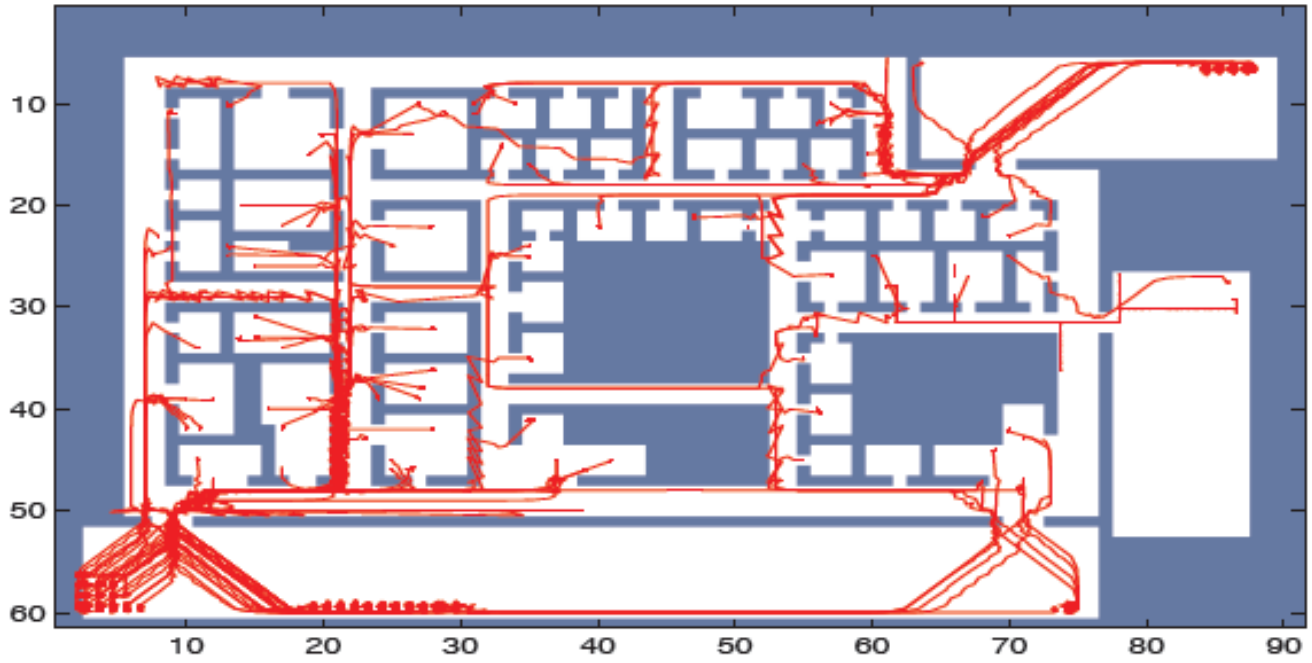
- Trajectory Data: Set of Harricans



http://csc.noaa.gov/hurricanes

Spatial & Spatial-Temporal Data

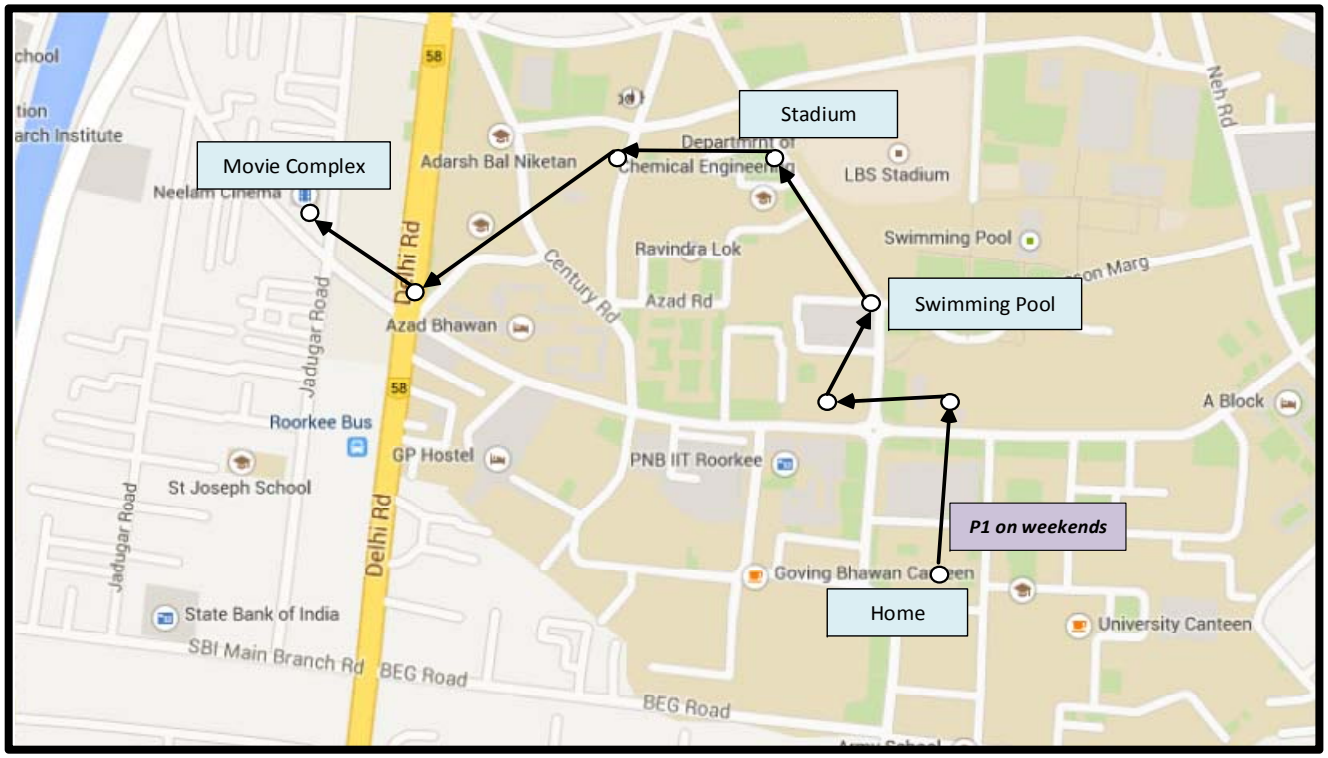
- Trajectory Data: (of 87 users obtained using RFID)



Vast 2008 Challenge – RFID Dataset

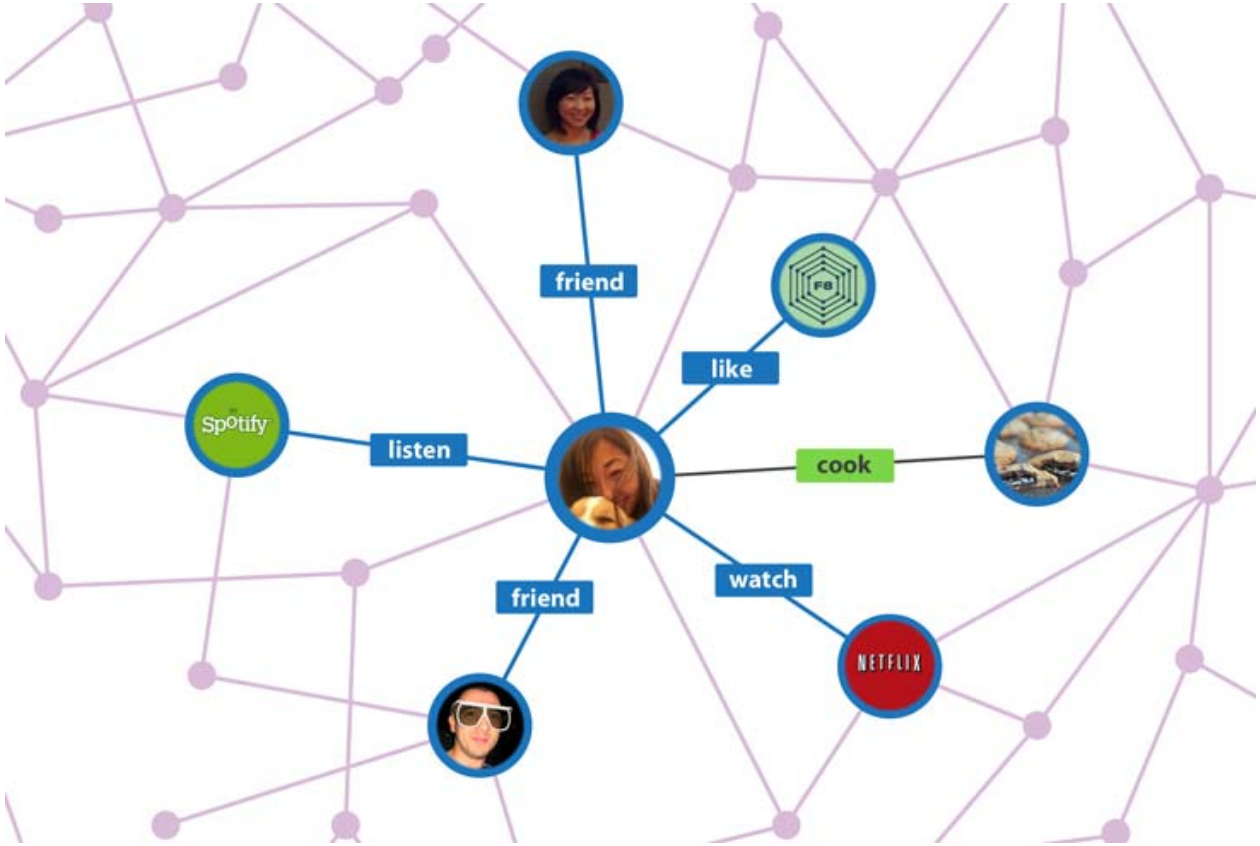
User Movement Data

- *Trajectory*
  - *Movement trail of a user*
  - *Sampling Points: <latitude, longitude, time>*

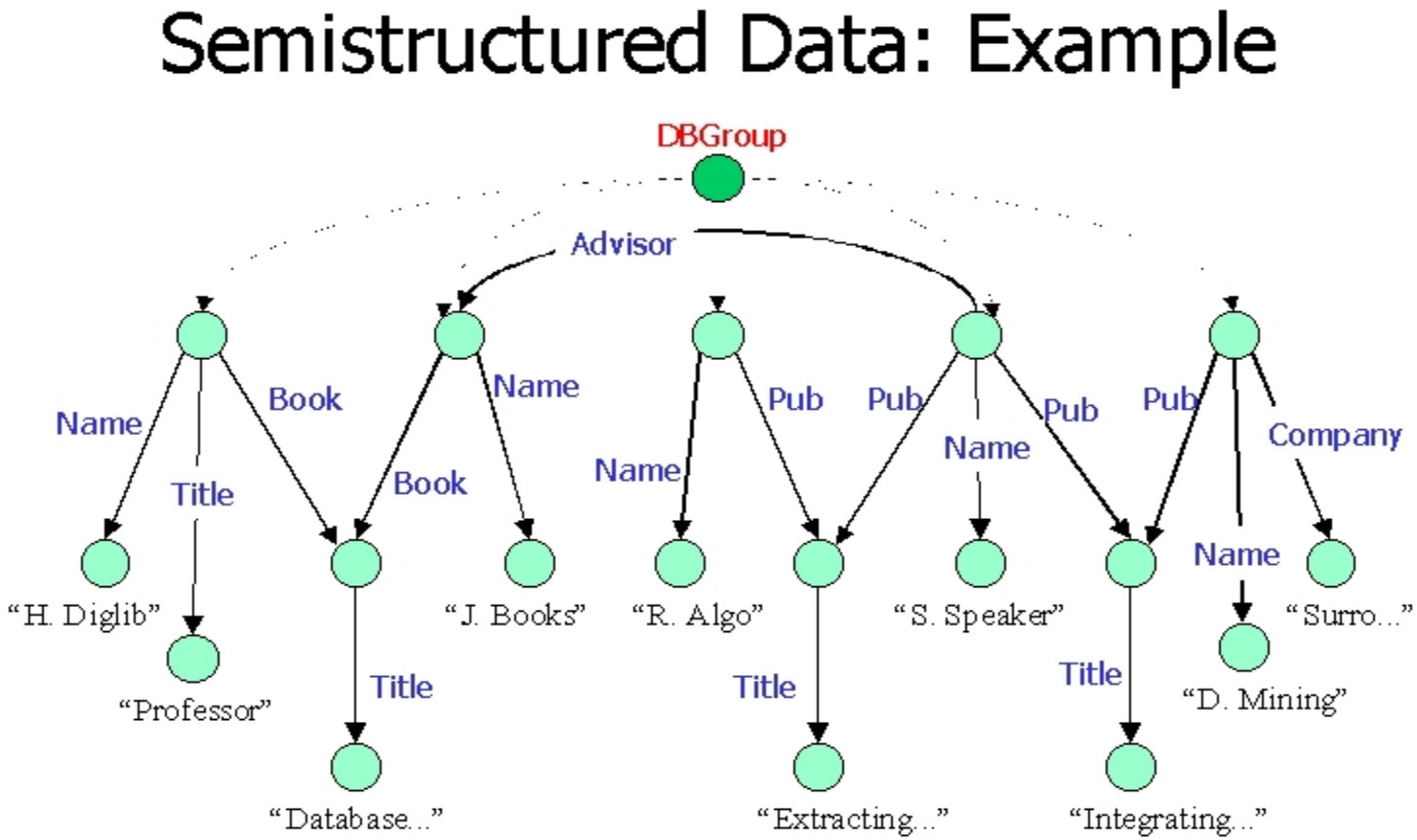


Thanks to Shreyash and Sahoishnu (M.Tech. Students)

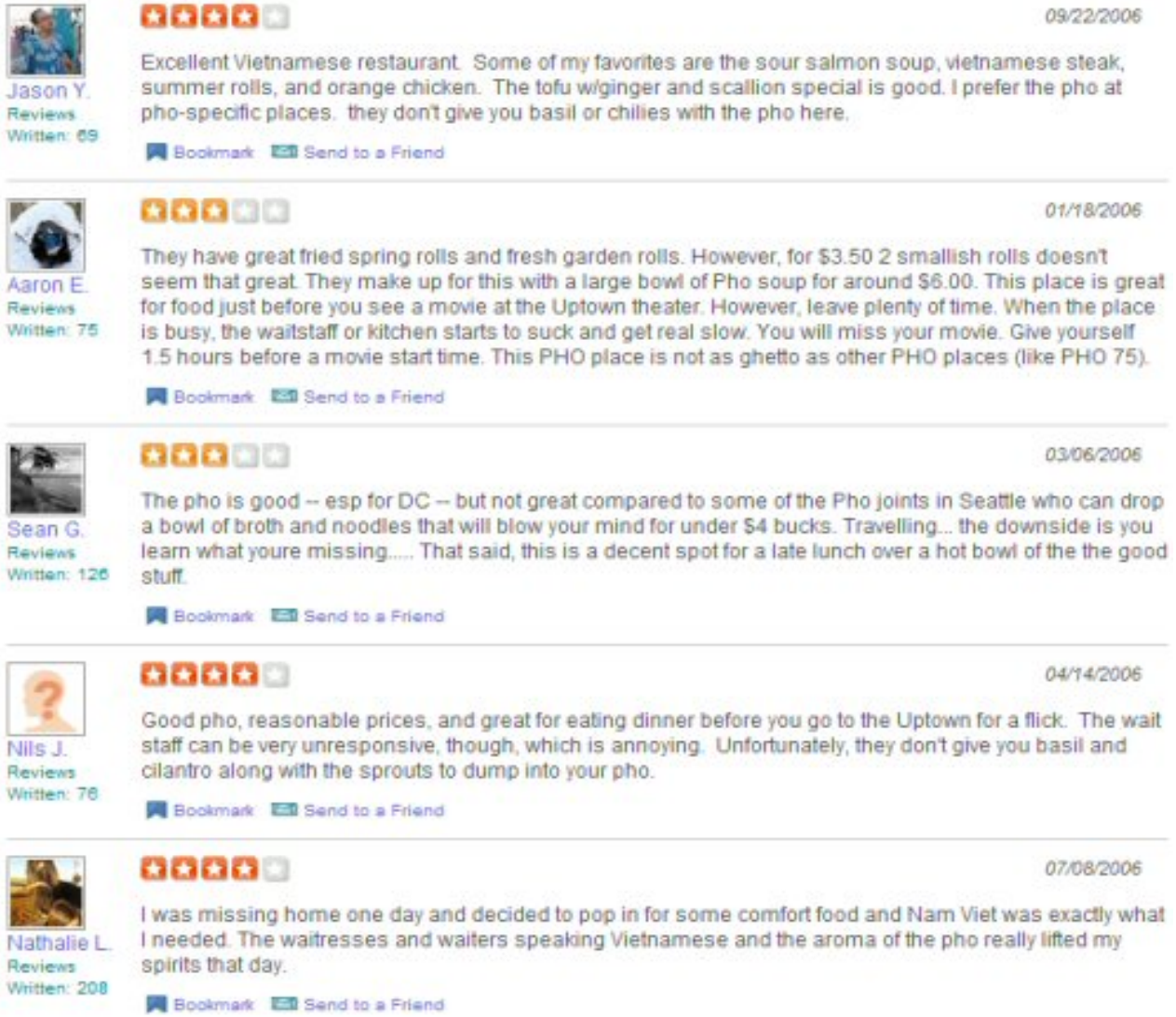
Graph Data



Semi-structured Data



Unstructured Data



Jason Y.
Reviews
Written: 69

09/22/2006

Excellent Vietnamese restaurant. Some of my favorites are the sour salmon soup, vietnamese steak, summer rolls, and orange chicken. The tofu w/ginger and scallion special is good. I prefer the pho at pho-specific places. they don't give you basil or chilies with the pho here.

Bookmark Send to a Friend

Aaron E.
Reviews
Written: 75

01/18/2006

They have great fried spring rolls and fresh garden rolls. However, for $3.50 2 smallish rolls doesn't seem that great. They make up for this with a large bowl of Pho soup for around $6.00. This place is great for food just before you see a movie at the Uptown theater. However, leave plenty of time. When the place is busy, the waitstaff or kitchen starts to suck and get real slow. You will miss your movie. Give yourself 1.5 hours before a movie start time. This PHO place is not as ghetto as other PHO places (like PHO 75).

Bookmark Send to a Friend

Sean G.
Reviews
Written: 126

03/06/2006

The pho is good -- esp for DC -- but not great compared to some of the Pho joints in Seattle who can drop a bowl of broth and noodles that will blow your mind for under $4 bucks. Travelling... the downside is you learn what youre missing..... That said, this is a decent spot for a late lunch over a hot bowl of the the good stuff.

Bookmark Send to a Friend

Nils J.
Reviews
Written: 76

04/14/2006

Good pho, reasonable prices, and great for eating dinner before you go to the Uptown for a flick. The wait staff can be very unresponsive, though, which is annoying. Unfortunately, they don't give you basil and cilantro along with the sprouts to dump into your pho.

Bookmark Send to a Friend

Nathalie L.
Reviews
Written: 208

07/08/2006

I was missing home one day and decided to pop in for some comfort food and Nam Viet was exactly what I needed. The waitresses and waiters speaking Vietnamese and the aroma of the pho really lifted my spirits that day.

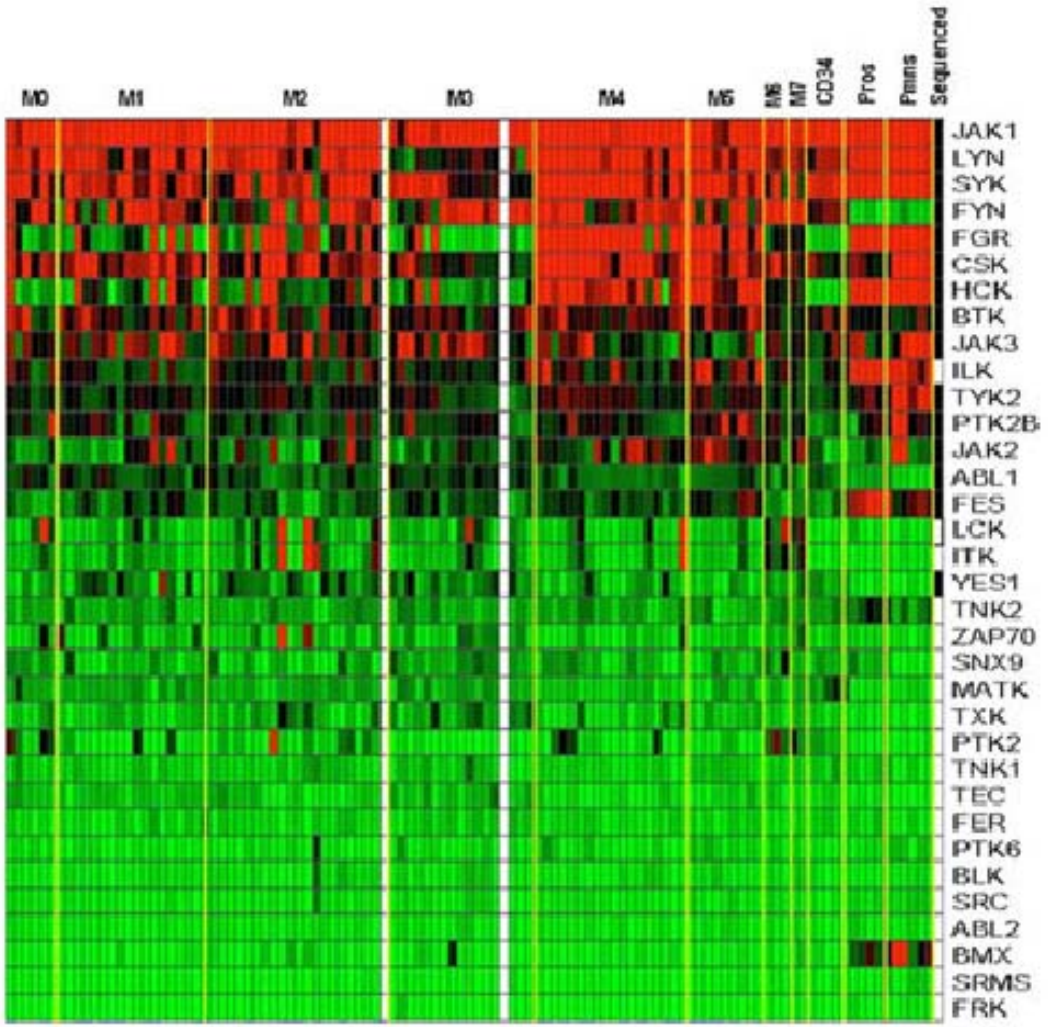
Bookmark Send to a Friend

Data can help us solve specific problems.

How should these pictures be placed into 3 groups?

How should these pictures be placed into groups? How many groups should there be?

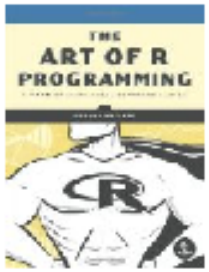
Which genes are associated with a disease? How can expression values be used to predict survival?



What items should Amazon display for me?

Books

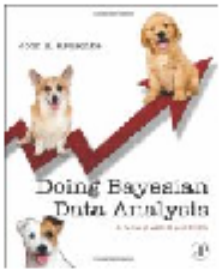
Page 2 of 20



The Art of R Programming
Norman S. Matloff
(22)
Paperback
~~$35.95~~ $24.32
Why recommended?

The Preservation Kitchen
Kate Leahy
(19)
Hardcover
~~$29.99~~ $18.74
Why recommended?

Agatha H. and the Clo...
Phil Foglio
(9)
Hardcover
~~$24.99~~ $16.44
Why recommended?



Doing Bayesian Data A...
John K. Kruschke
(18)
Hardcover
~~$89.95~~ $64.15
Why recommended?

Probabilistic Graphic...
Daphne Koller
(13)
Hardcover
~~$95.00~~ $84.25
Why recommended?

See all recommendations in Books

Is it likely that this stock was traded based on illegal insider information?

Where are the faces in this picture?

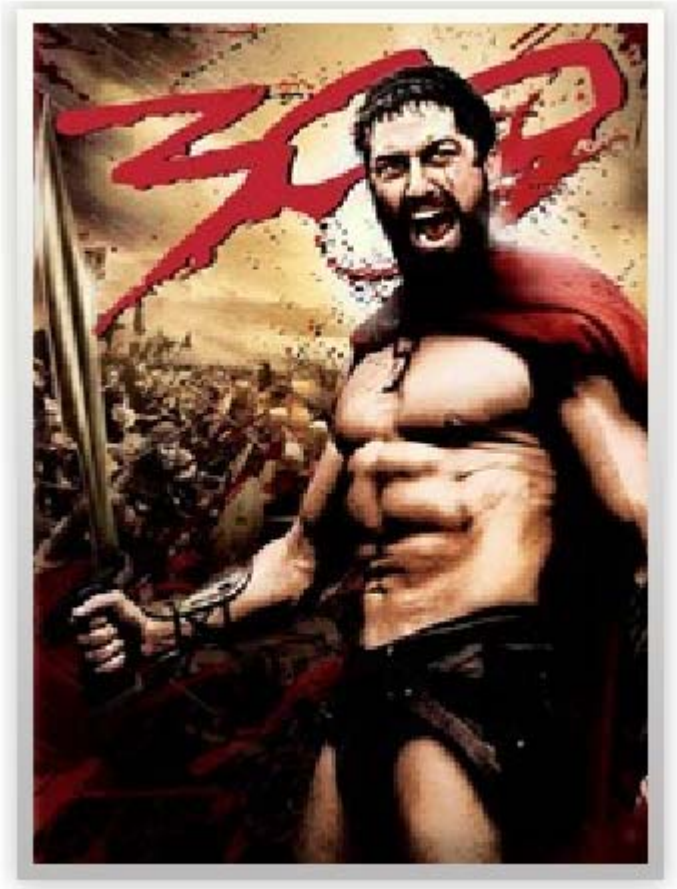
Is this spam?

hi backpackers,

i saw that close to my hotel there is a pub with bowling (it's on market between 9th and 10th avenue). are you up to it? i think it is about 20 years i haven't played... if you like the idea what about 8.30 there?

otherwise any suggestion welcome. i can survive another 20 years without bowling.

Will I like 300?



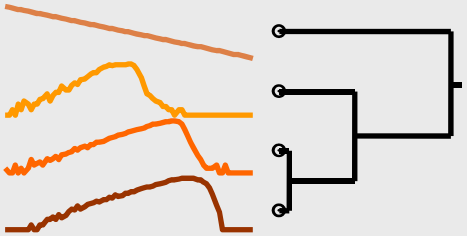
What techniques people apply on data?

- They apply data mining algorithms and discover useful knowledge

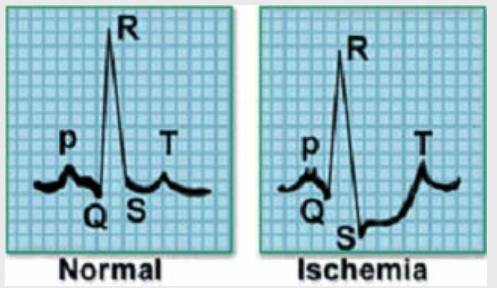
- So, what are the some of the well-known **Data mining Tasks**?
  - Clustering,
  - Classification,
  - Frequent Patterns,
  - Association Rules,
  - ….

What people do with the time series data?

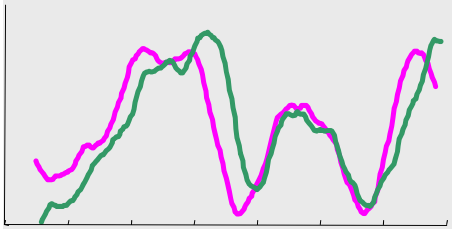
Clustering



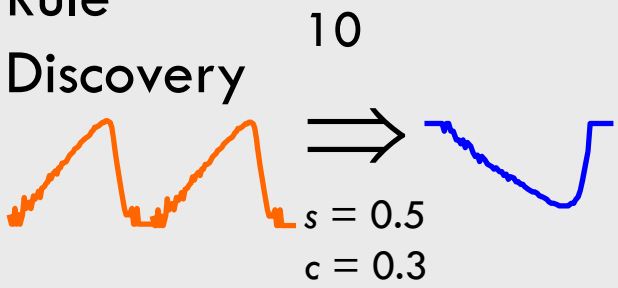
Classification



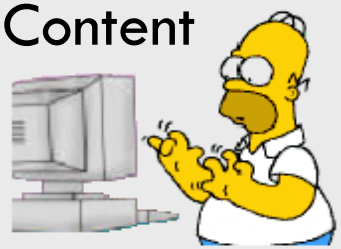
Motif Discovery

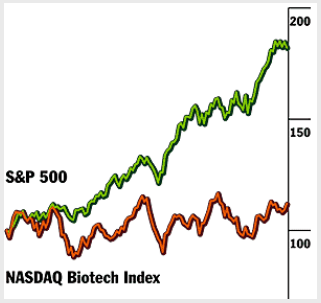


Rule Discovery



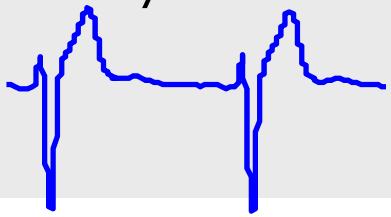
Query by Content



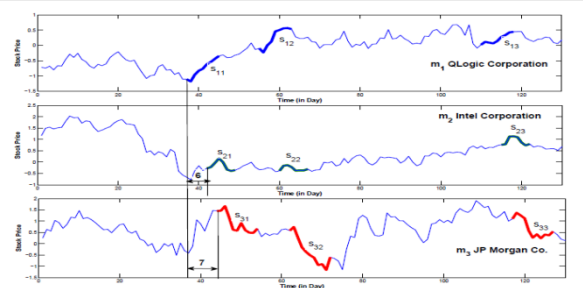


Visualization

Novelty Detection

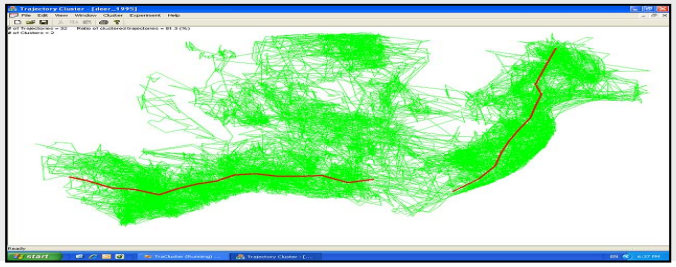


Motif Association

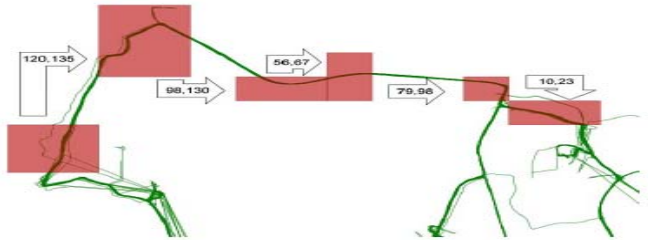


What people do with the trajectory data?

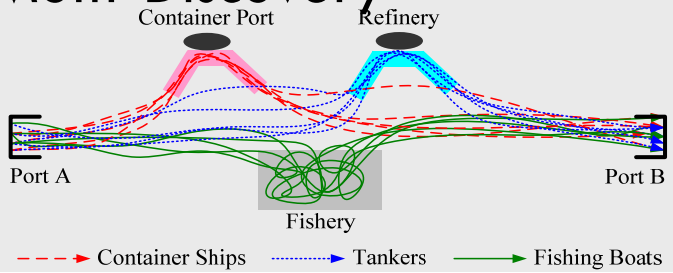
Clustering



Frequent Travel Patterns



Motif Discovery



Prediction

Visualization

Classification

In, Summary

Types of Data

- Transactional Data
- Sequence Data
- Interval Data
- Time Series Data
- Spatial Data
- Spatio-Temporal Data
- Data Set with Multiple Kinds of Data
- ….

Algorithms

Data Mining Methods

- Frequent Pattern Discovery
- Classification
- Clustering
- Outlier Detection
- Statistical Analysis
- …

Activity 1

- Find top 3 recent research activities around the world that are analyzing data. You need to write short summary for each research activities. First three line must follow following format:
  - **Line 1**: Problem they are trying to sole along with dataset they are using
  - **Line 2**: How they are solving the problem
  - **Line 3**: Justify yourself why you rate this work as a top 5 activities
  - Remaining lines… you can think yourself ….

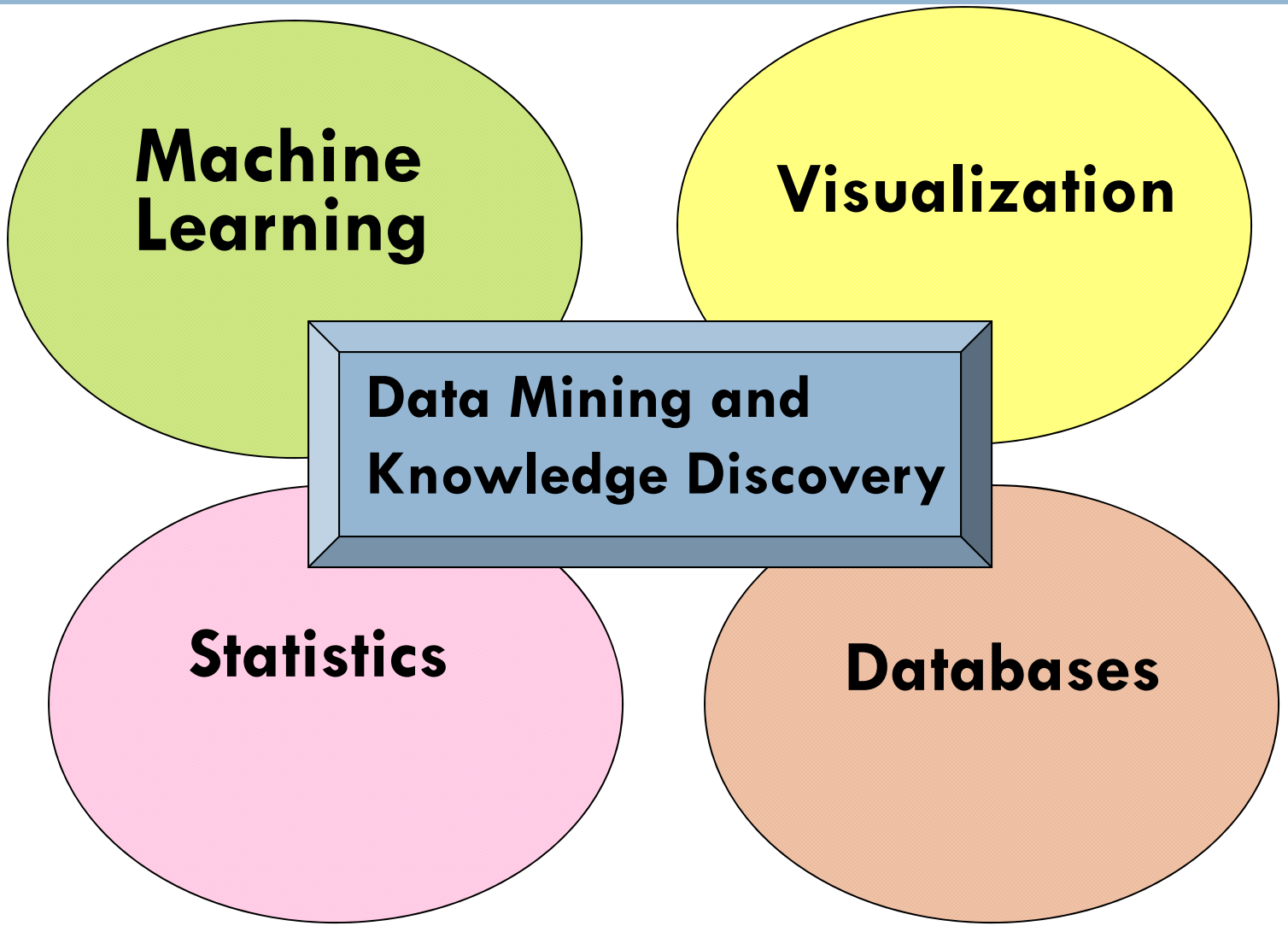
BigN'Smart Research group at IIT-Roorkee is analyzing "YelpReview" Dataset for learning Location-to-activity Tagging. They are applying … . I feel this is an interesting research because …

Activity 2: Why Data Mining ???

- Google
- Facebook
- Netflix
- eHarmony
- FICO
- FlightCaster
- IBM's Watson

Read About Their Story

Related Field

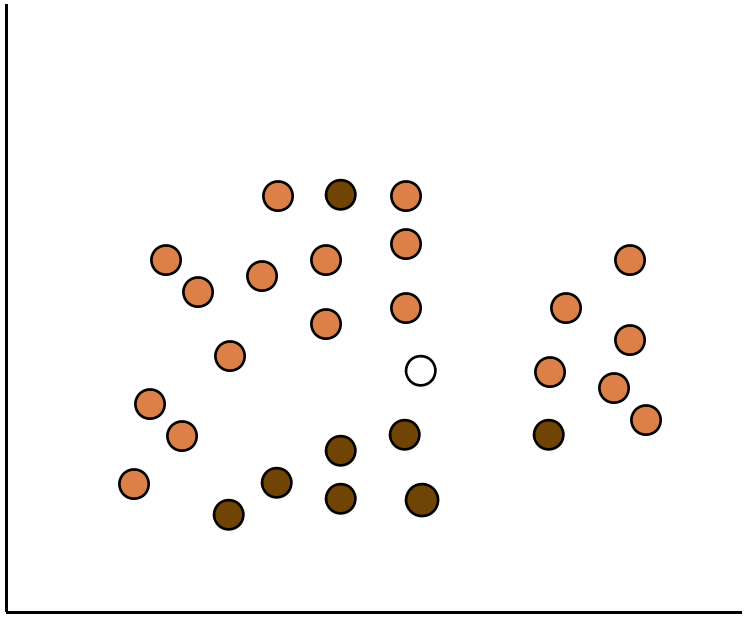


Related Field

- Statistics:
  - more theory-based
  - more focused on testing hypotheses
- Machine learning
  - more heuristic
  - focused on improving performance of a learning agent
  - also looks at real-time learning and robotics – areas not part of data mining
- Data Mining and Knowledge Discovery
  - integrates theory and heuristics
  - focus on the entire process of knowledge discovery, including data cleaning, learning, and integration and visualization of results
- Distinctions are fuzzy

Classification

Learn a method for predicting the instance class from pre-labeled (classified) instances

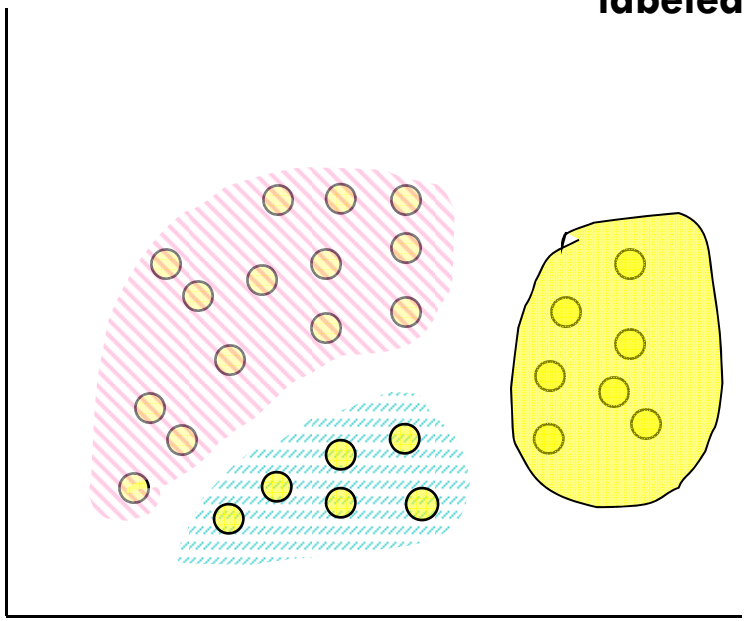


Many approaches: Statistics, Decision Trees, Neural Networks,

…

Clustering

Find "natural" grouping of instances given un-labeled data



Association Rules & Frequent Itemsets

Transactions

TID	Produce
1	MILK, BREAD, EGGS
2	BREAD, SUGAR
3	BREAD, CEREAL
4	MILK, BREAD, SUGAR
5	MILK, CEREAL
6	BREAD, CEREAL
7	MILK, CEREAL
8	MILK, BREAD, CEREAL, EGGS
9	MILK, BREAD, CEREAL

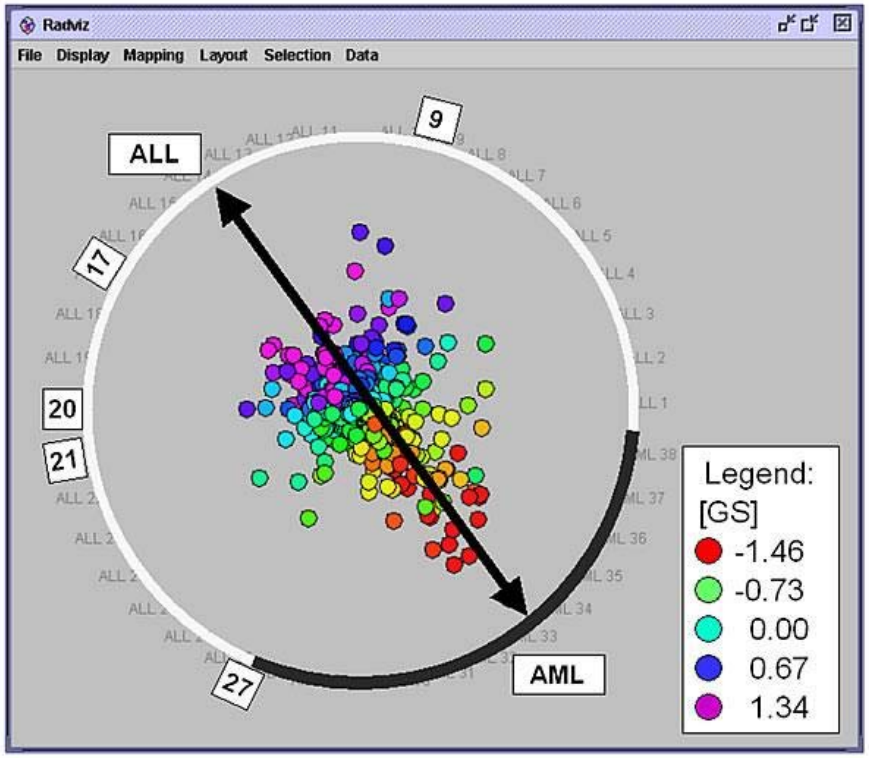
Frequent Itemsets:

Milk, Bread (4)
Bread, Cereal (3)
Milk, Bread, Cereal (2)
…

Rules:
Milk => Bread (66%)

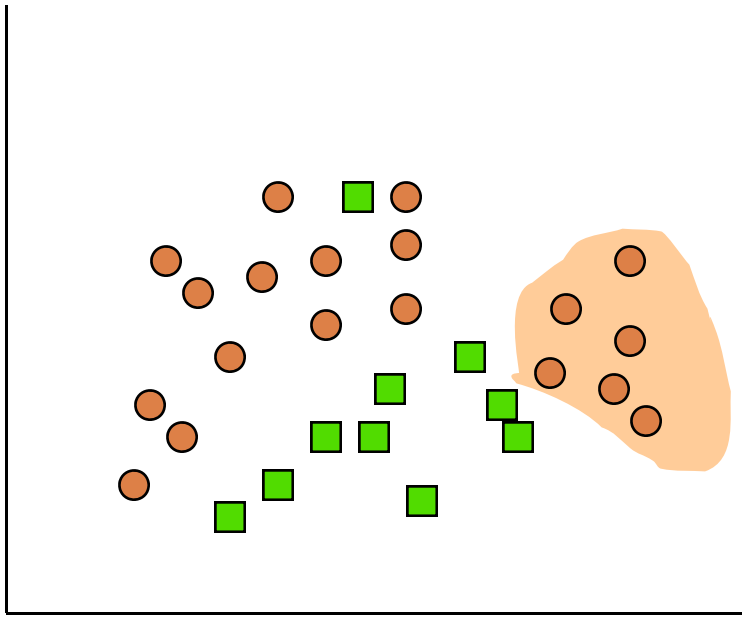
Visualization & Data Mining

- Visualizing the data to facilitate human discovery



- Presenting the discovered results in a visually "nice" way

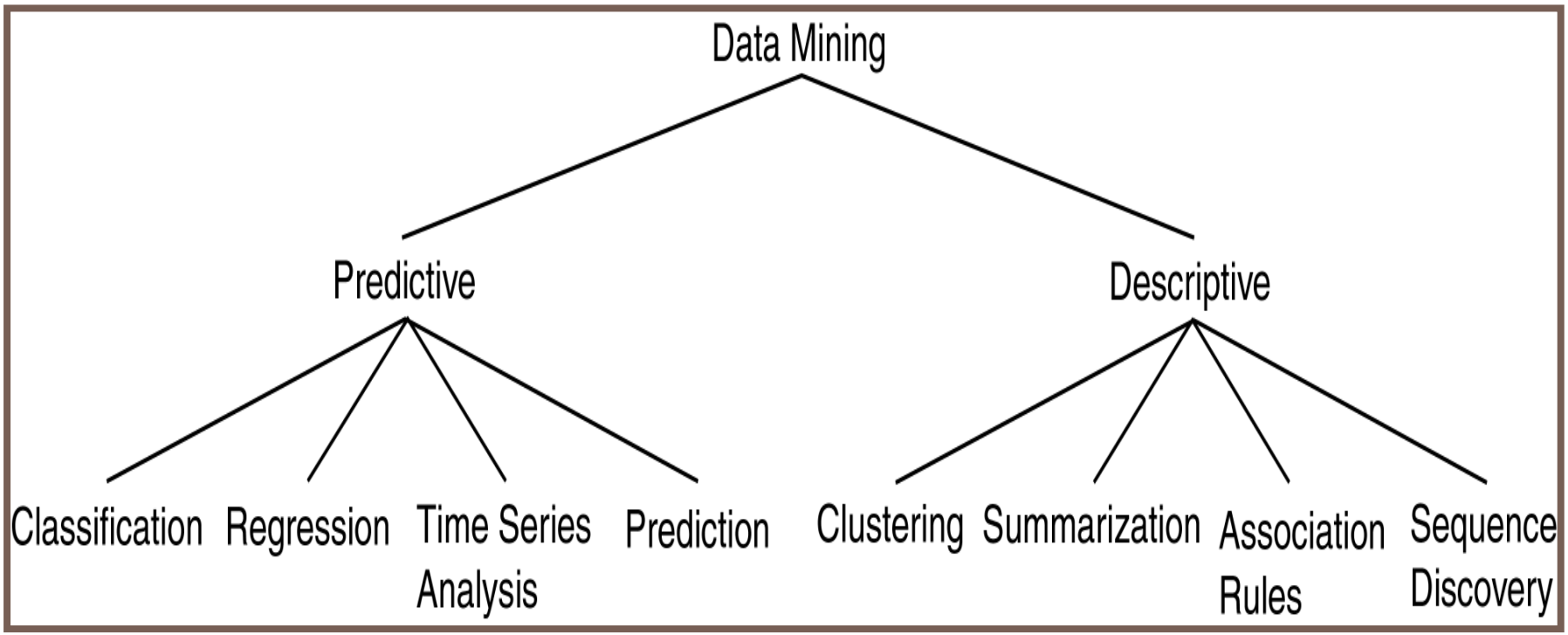
Summarization



- **Describe features of the selected group**
- **Use natural language and graphics**
- **Usually in Combination with Deviation detection or other methods**

Average length of stay in this study area rose 45.7 percent, from 4.3 days to 6.2 days, because ...

Data Mining Models and Tasks



Obtained from Prof. Srini's Lecture notes