

Lecture 01 - Introduction to Data Mining

Fall 2019

Department of Computer Science and Statistics

Faculty of Mathematics

K. N. Toosi University of Technology

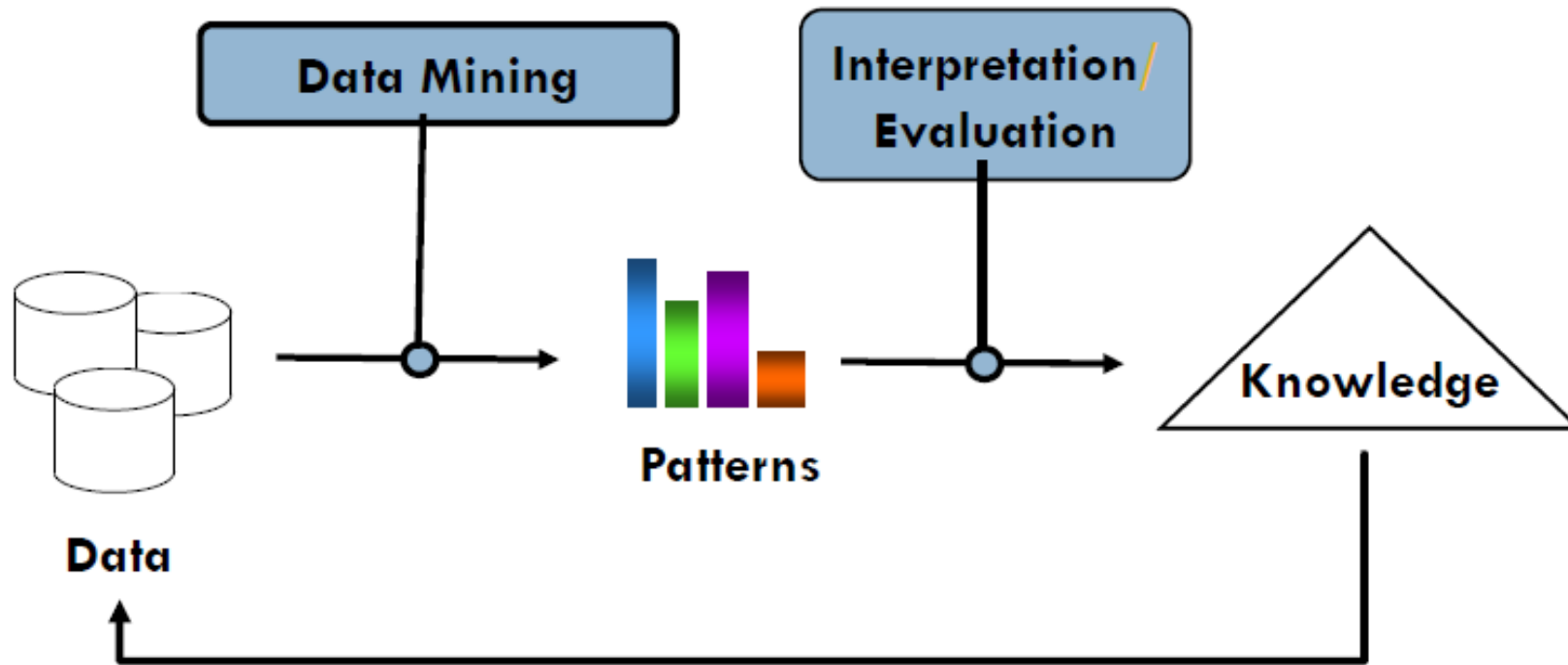
Sources for this lecture

- Text book: Data Mining, Concepts and Techniques, Chapter 1
- Text book: Mining Massive Data Sets, Chapter 1
- Slides by Jiawei Han
- Slides by Dhaval Patel (Data Mining Course)
- Various online webpages

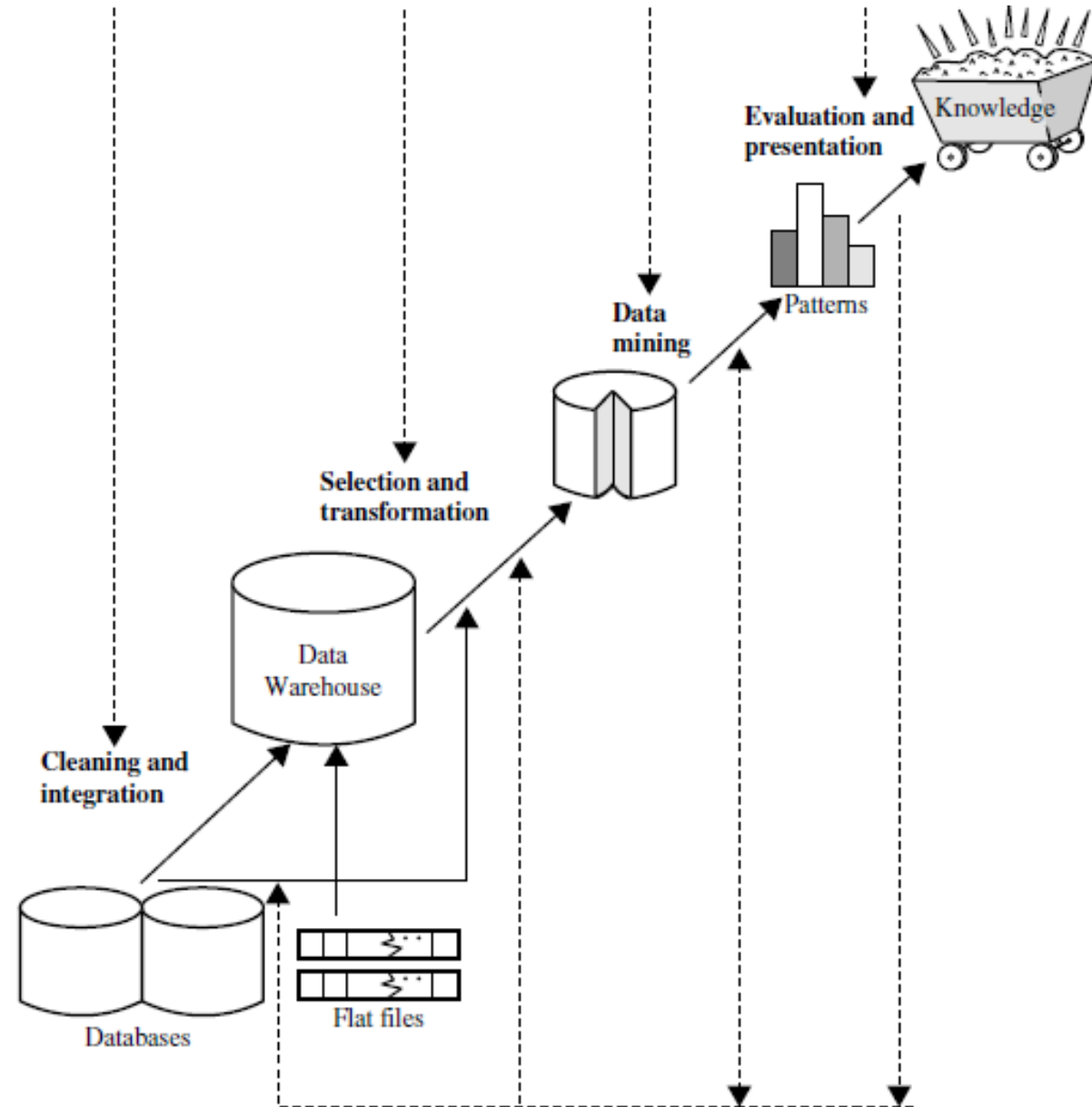
What is Data Mining?

- **Database approach:** Data Mining (aka Knowledge Discovery in Databases KDD) is **extraction of useful patterns from data sources** (databases, text, web, images, ...)
- **Statistics + Machine learning approach:** Data Mining is **discovery of models for data**
- **Big data approach:** Data mining is automated analysis of **massive data sets**

Data Mining Cycle (Data Cycle)



Data Mining Cycle



What do we mean by Data?

Introduction to Data

- Different Sources / Different forms

Examples:

- Businesses
- Social Media
- Weather Stations
- Hospitals



Forms of Data, examples:

- Record Data (Transactional Databases)
- Temporal Data (Time-series data, Sequence Data)
- Spatial, Spatial Temporal Data
- Graph Data
- Unstructured Data (Facebook status, News Articles)
- Semi-Structured Data (Publications, XML format)
- Data Matrix (Terms and Documents)

Record Data

Attributes

<i>Tid</i>	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

Objects

Temporal Data

- Sequence 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→Splénomegalia
5	Night sweat→Fever→Achroacytosis
6	Night sweat→Fever→Anemia
7	Night sweat→Splénomegalia→Anemia
8	Night sweat→Sleepy→Anemia

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

Temporal Data

- Time-series data



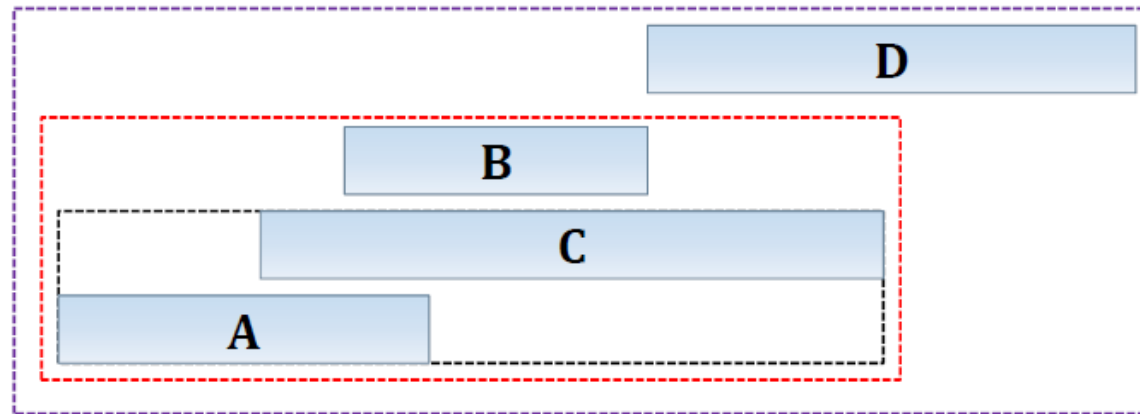
Temporal Data

- Biological Sequence Data

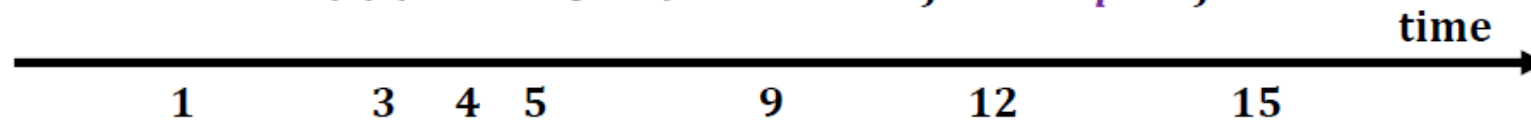
Species	Alignment of Amino Acid Sequences of β -globin					
Human	1	VHLTPEEKSA	VTALWGKLVNV	DEVGGEALGR	LLVVYPWTQR	FFESFGDLST
Monkey	1	VHLTPEEKNA	VTTLWGKLVNV	DEVGGEALGR	LLLVYPWTQR	FFESFGDLSS
Gibbon	1	VHLTPEEKSA	VTALWGKLVNV	DEVGGEALGR	LLVVYPWTQR	FFESFGDLST
Human	51	PDAVMGNPKV	KAHGKKVLGA	FSDGLAHLDN	LKGTFAQLSE	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

Interval Data

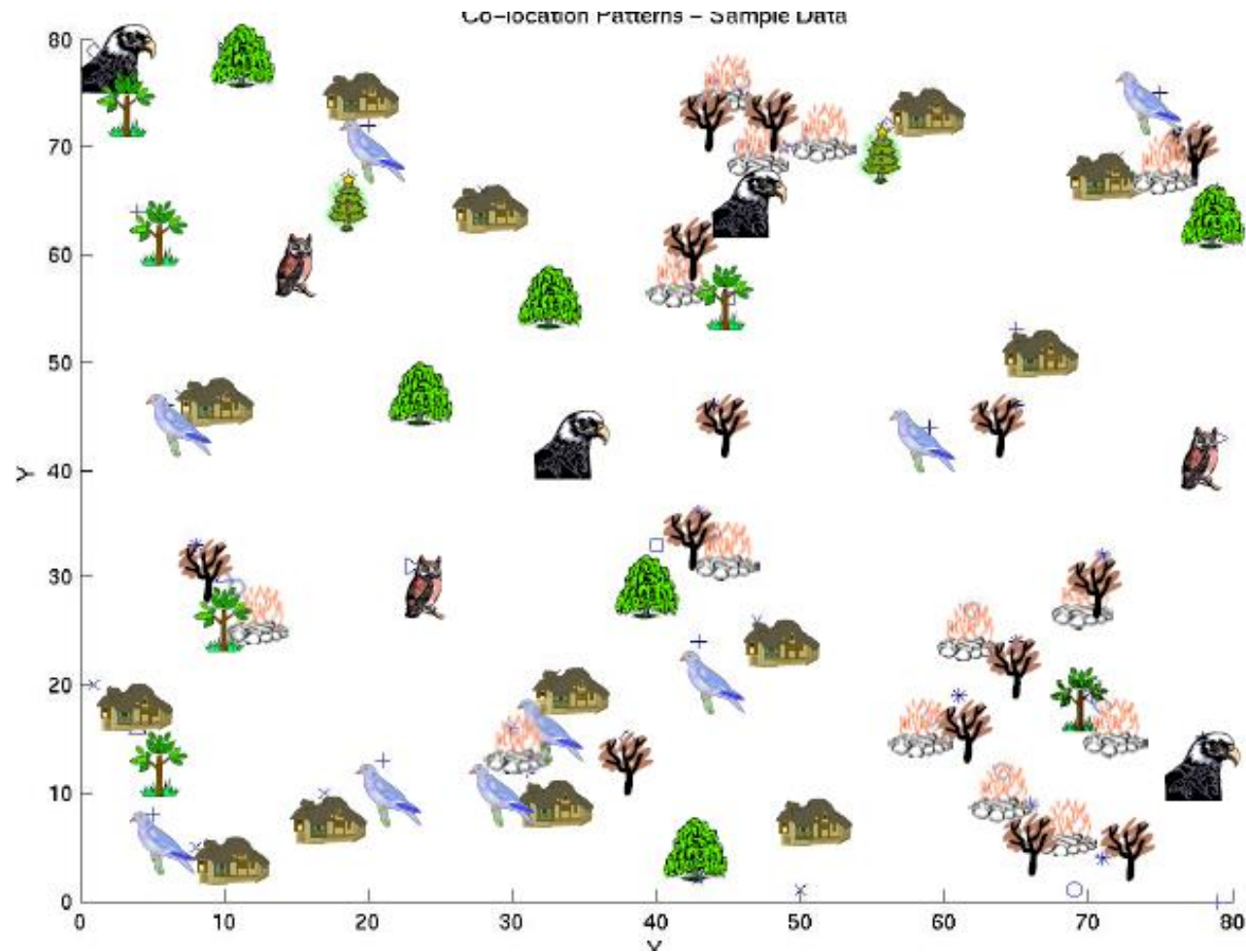
$$EL = \{ (A, 1, 5), (C, 3, 12), (B, 4, 9), (D, 9, 15) \}$$



$((A \text{ overlaps } C) \text{ contains } B) \text{ overlaps } D$

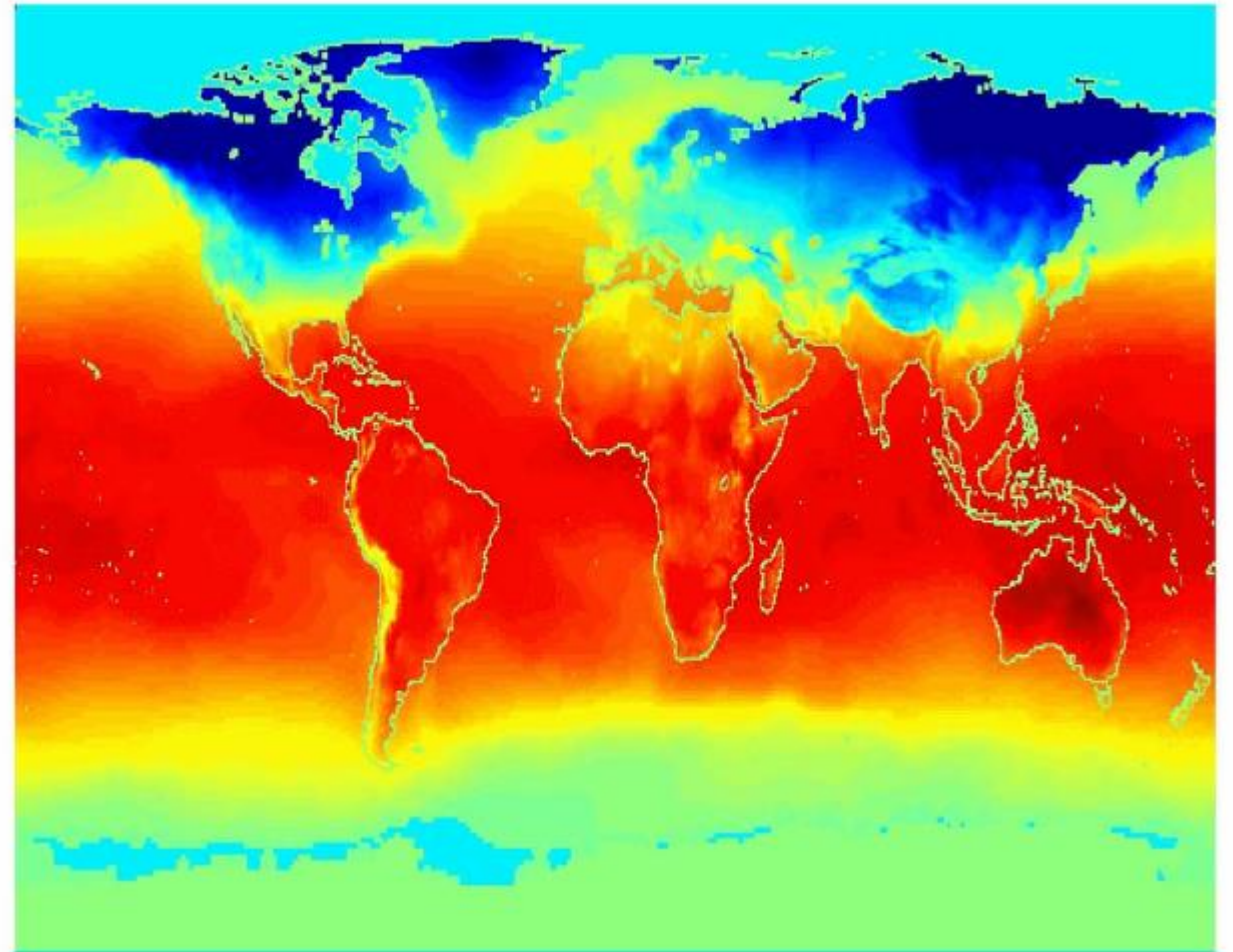


Spatial Data

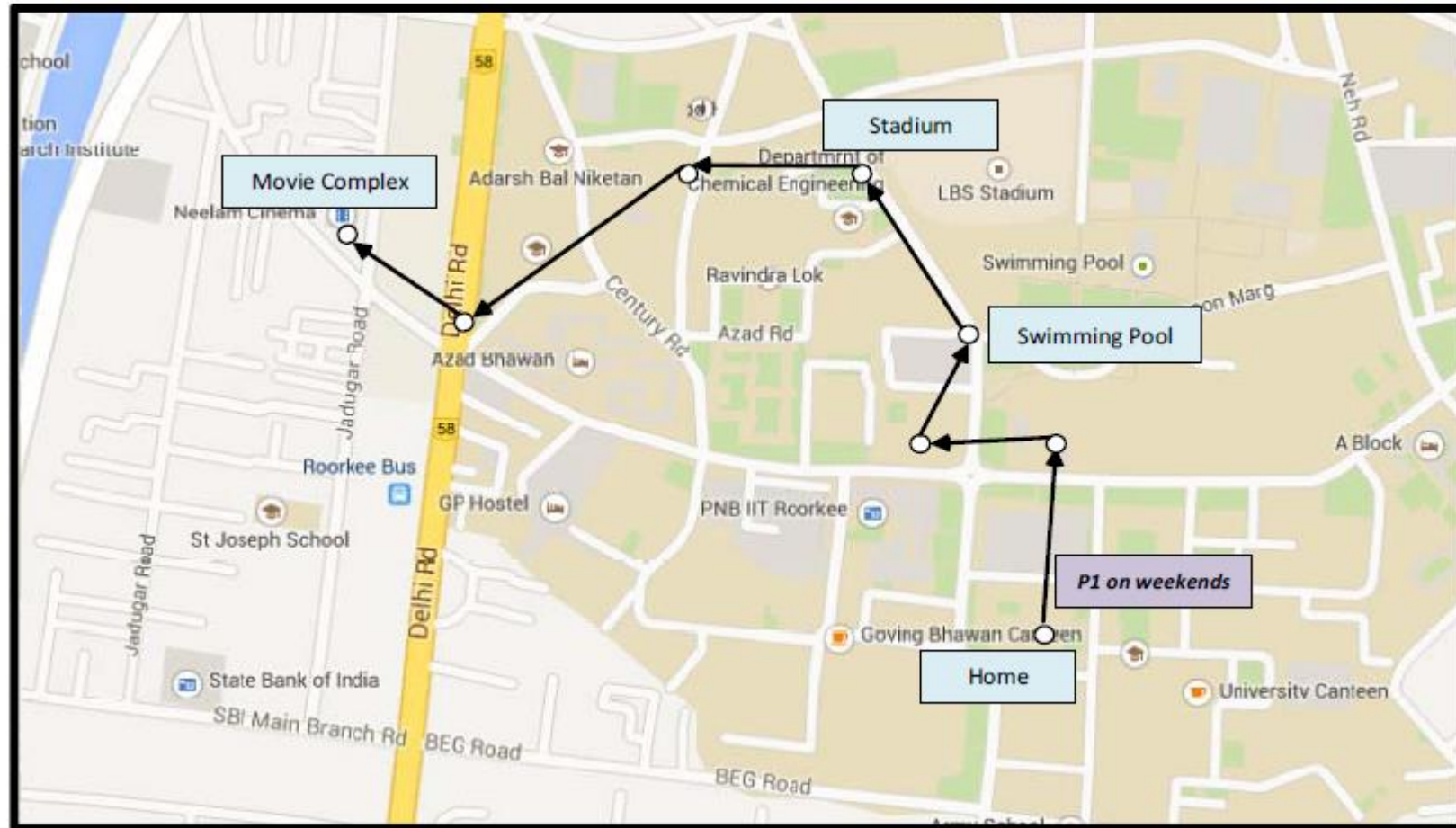


Spatial Data

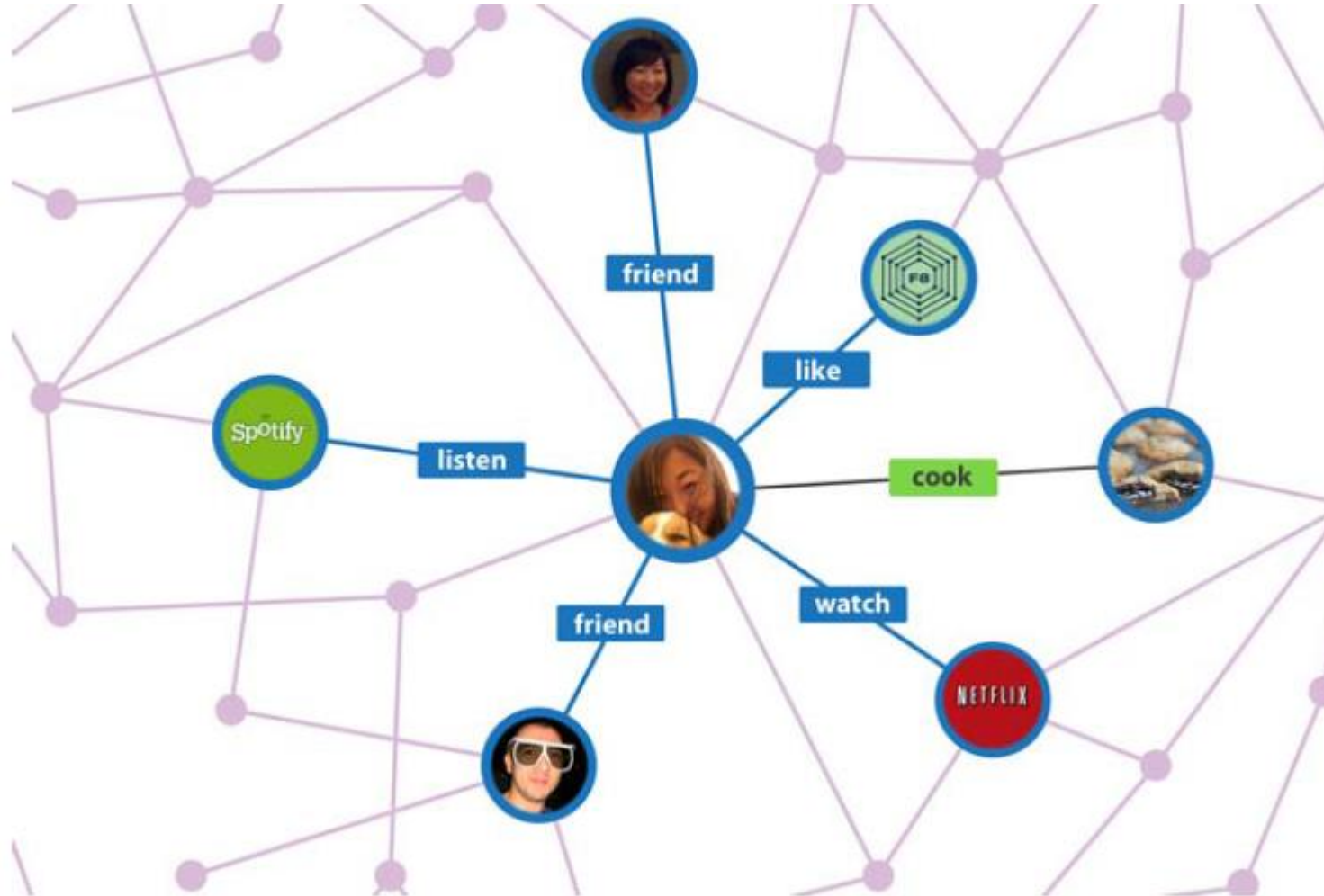
- Average monthly temperature of ocean and land



Trajectory Data



Graph Data



Semi-Structured/ Unstructured Data

```
- <accounts>
- <account gm_recid="000000091{{>2R" gm_accountno="ELAN SOFTWARE CORP.">
- <properties>
- <property name="Accnt Mngr" db_name="KEY4">
  <property_string>C.Stott</property_string>
</property>
- <property name="Actionon" db_name="ACTIONON">
  <property_string>20021109</property_string>
</property>
- <property name="Address" db_name="ADDRESS1">
  <property_string>1150 Kelly Johnson Boulevard</property_string>
</property>
- <property name="Asst" db_name="SECR">
  <property_string>Jane</property_string>
</property>
- <property name="Business" db_name="KEY2">
  <property_string>Comp. Sfw. Dev.</property_string>
</property>
- <property name="Callbkfreq" db_name="CALLBKRFREQ">
  <property_string>0</property_string>
</property>
+ <property name="City" db_name="CITY">
- <property name="Company" db_name="COMPANY">
  <property_string>FrontRange Solutions Inc.</property_string>
</property>
- <property name="Contact" db_name="CONTACT">
  <property_string>Patrick B. Hillyard</property_string>
</property>
- <property name="Contact Type" db_name="KEY1">
  <property_string>Internal</property_string>
```

06/17/2014 ★★★★★

“Absolutely the best selection in the area! I always recommend this place to my friends.”

 Jocelyn Y.

06/16/2014 ★★★★★

“I have yet to find a better store when you need knowledgeable salespeople and fast service. I stopped by here on my way...”

 Kate S.

06/15/2014 ★★★★★

“They never fail to deliver. I called on a whim before placing an order on Amazon, and I'm glad I did...”

 Will B.

Data Matrix

- Each Document is a **term vector**

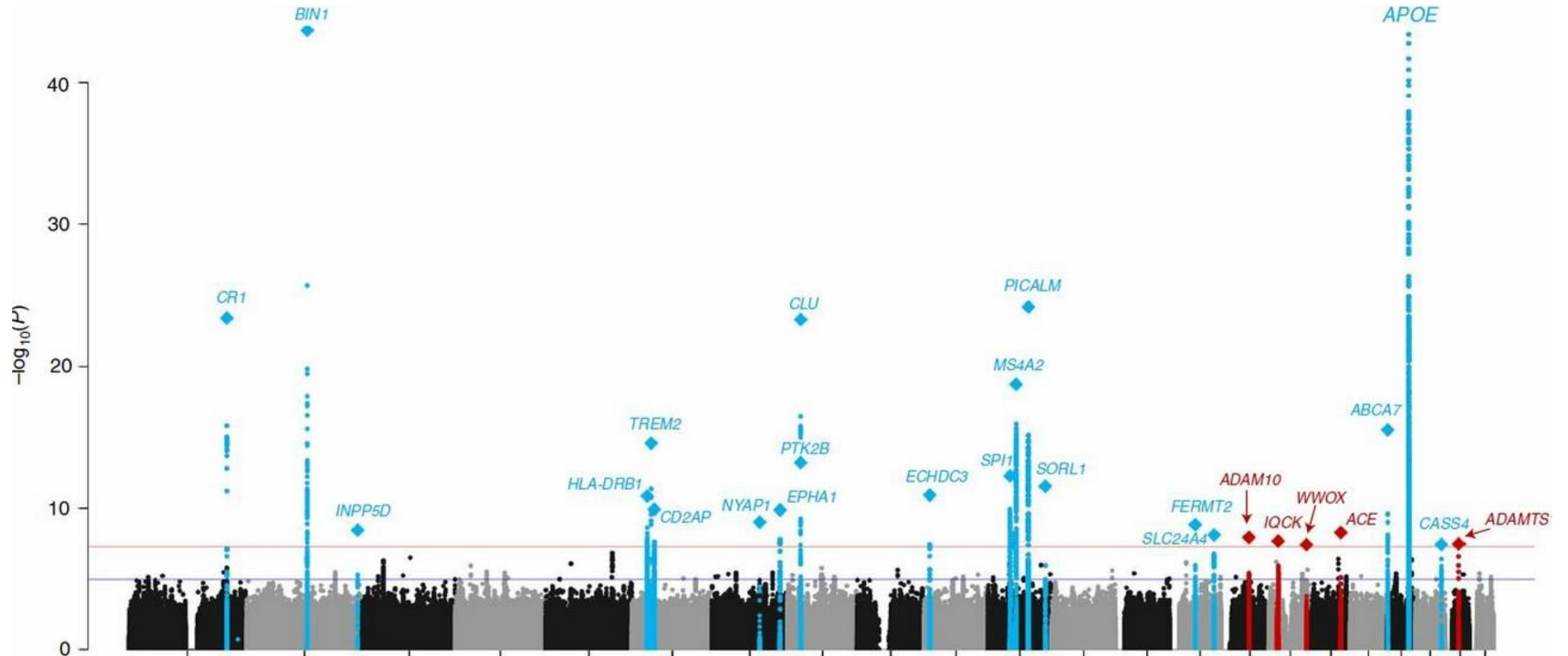
	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Some Real World Problems

Group pictures



Relations between genes and diseases



What ads Google should display for me?

Google

میز ناهار خوری

EKBATAN TOWN شهرک اکباتان Tehran DISTRICT 13 منطقه ۱۳ Forest Park Map data ©2019

Rating ▾ Hours ▾


فروشگاه اینترنتی ایران میز
4.2 ★★★★★ (14) · خدمات تجارت الکترونیک
Tehran Province, Tehran · 021 2225 3925
Open 24 hours
Their website mentions میز ناهار خوری

مبلمان پردیس مبل کلاسیک راحتی ناهارخوری میز
No reviews · مرکز خرید
Tehran Province, Pardis · 0912 221 1361

زوفا چوب
4.9 ★★★★★ (16) · فروشگاه مبل
Tehran Province, Tehran · 021 2632 7696
Closed · Opens 10AM
Their website mentions میز ناهار خوری

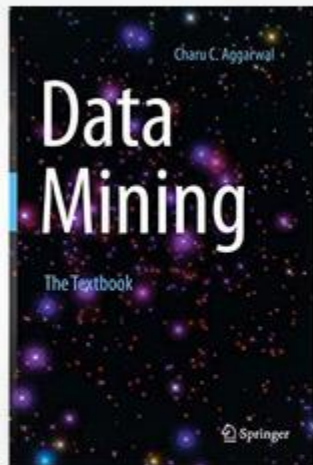
ranMiz.com

SOLESTAN کوی گلستان
Inteqal Blvd
Soroush



What items should Amazon display for me?

Recommended for you

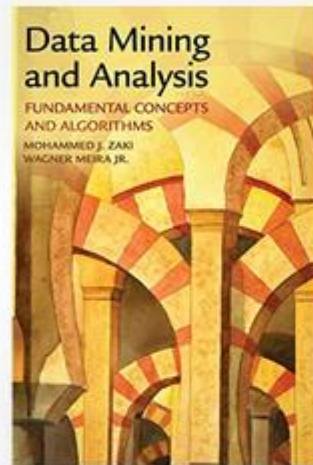


\$60⁴⁶

~~\$89.99~~ ✓prime

Data Mining: The Textbook

★★★★★ 12



\$50⁴¹

~~\$69.99~~ ✓prime

Data Mining and Analysis:
Fundamental Concepts and...

★★★★★ 11

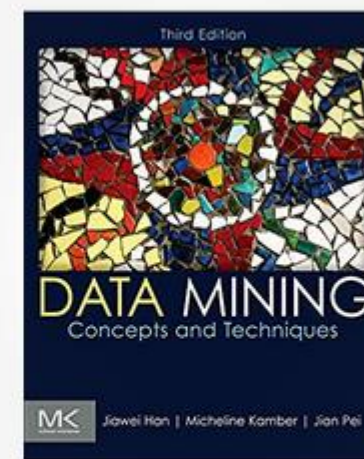


\$62⁶¹

~~\$79.99~~ ✓prime

Machine Learning for Text

★★★★★ 7

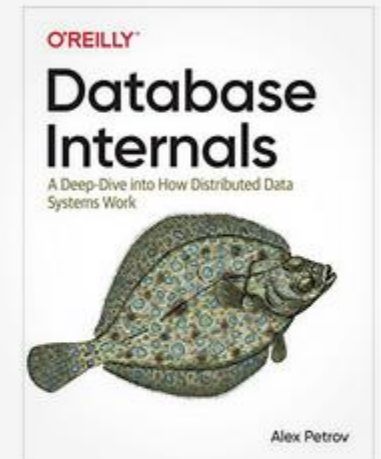


\$50⁶⁴

~~\$74.95~~ ✓prime

Data Mining: Concepts and
Techniques (The Morgan...

★★★★★ 52



\$41⁹⁹

~~\$59.99~~ ✓prime

Database Internals: A Deep
Dive into How Distributed D...

Blur faces in the 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.

Well-known Data Mining Tasks

- **Frequent items & Association Rules**
- **Clustering**
- Classification
- **Outlier detection**
- Data Visualization
- Summarization
- Finding Specific Patterns (Periodicity)

Association Rules and 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)
...

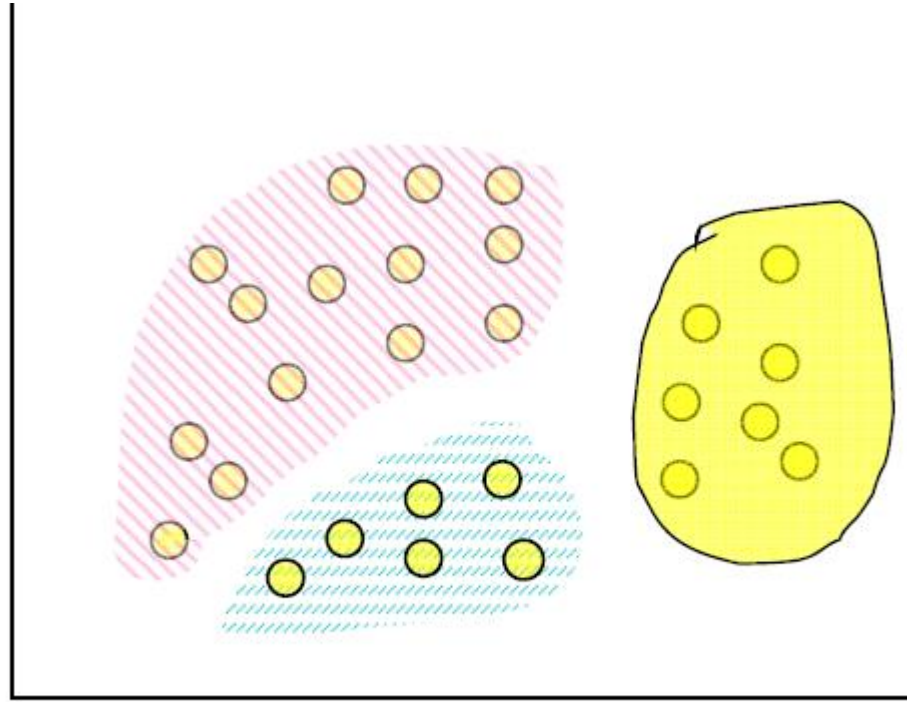


Rules:

Milk => Bread (66%)

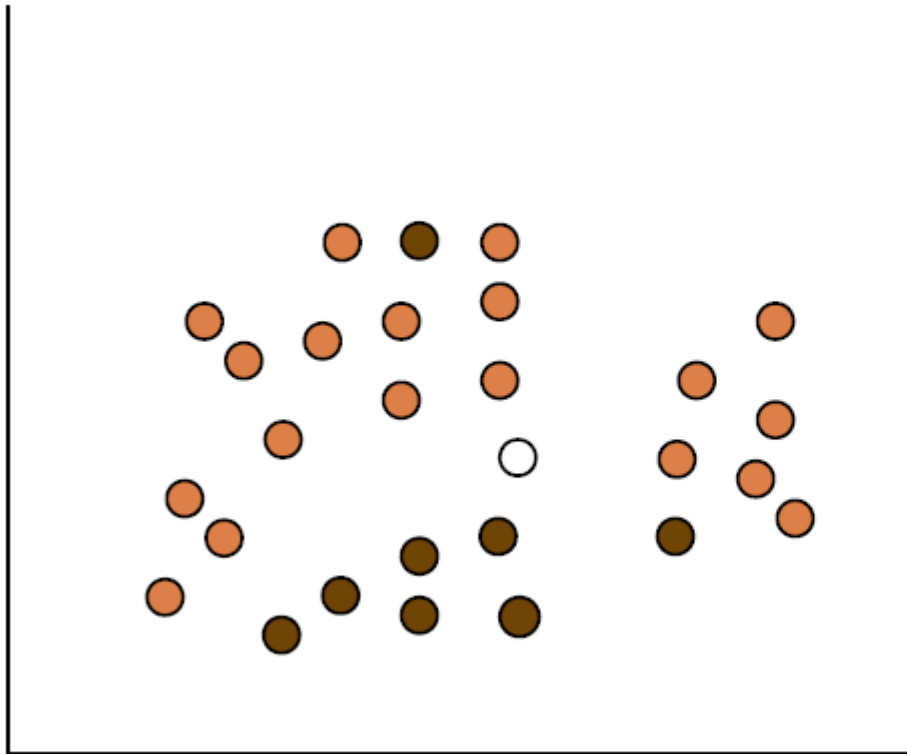
Clustering

- Find “natural” groupings of the given instances



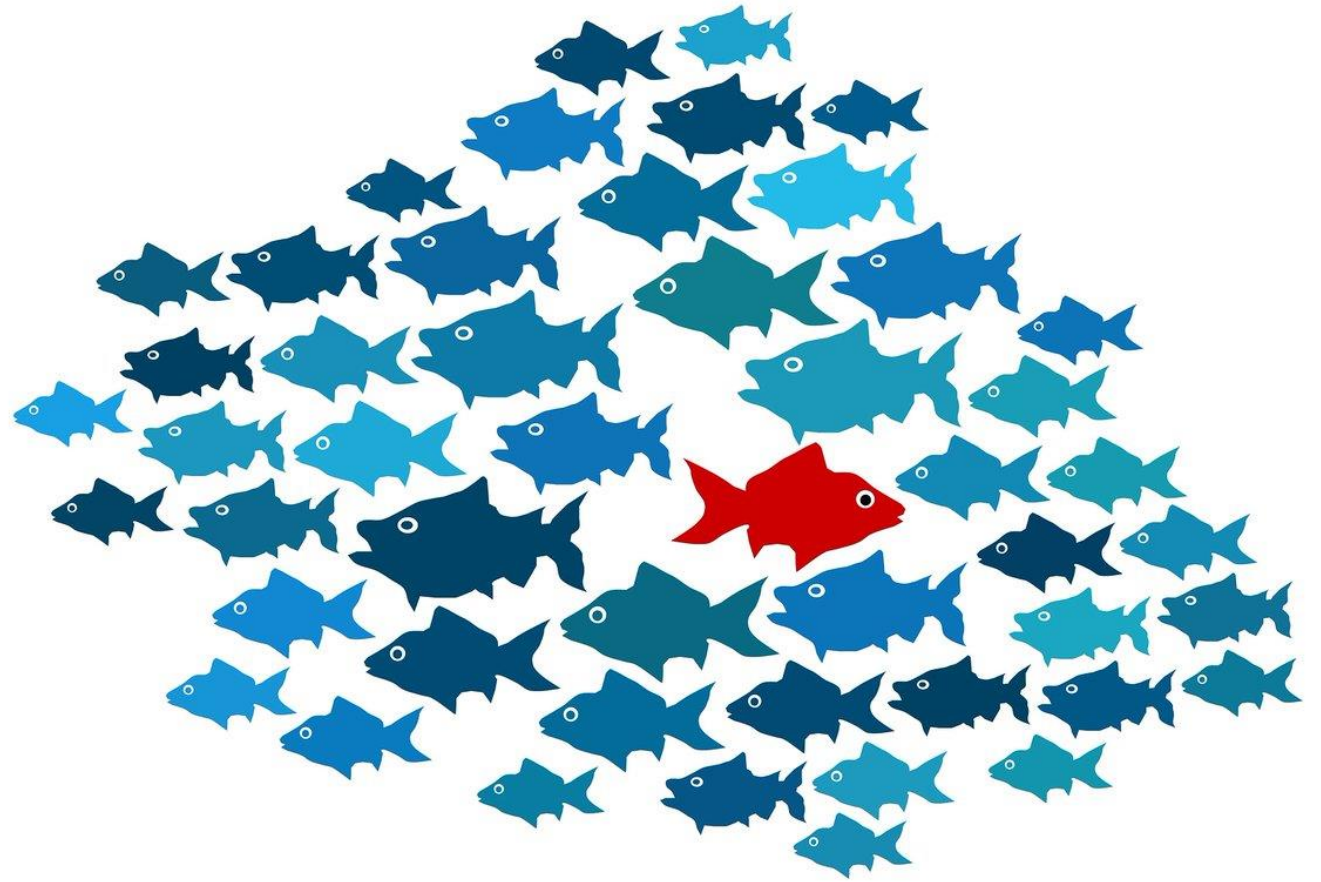
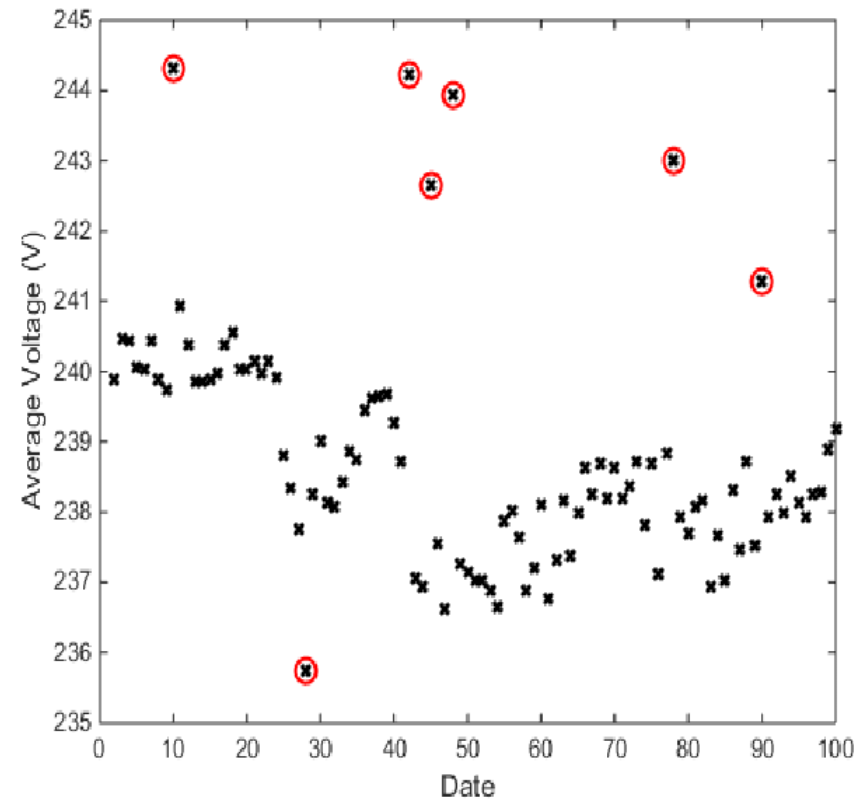
Classification

- Predict the label of the new data based on pre-given instances

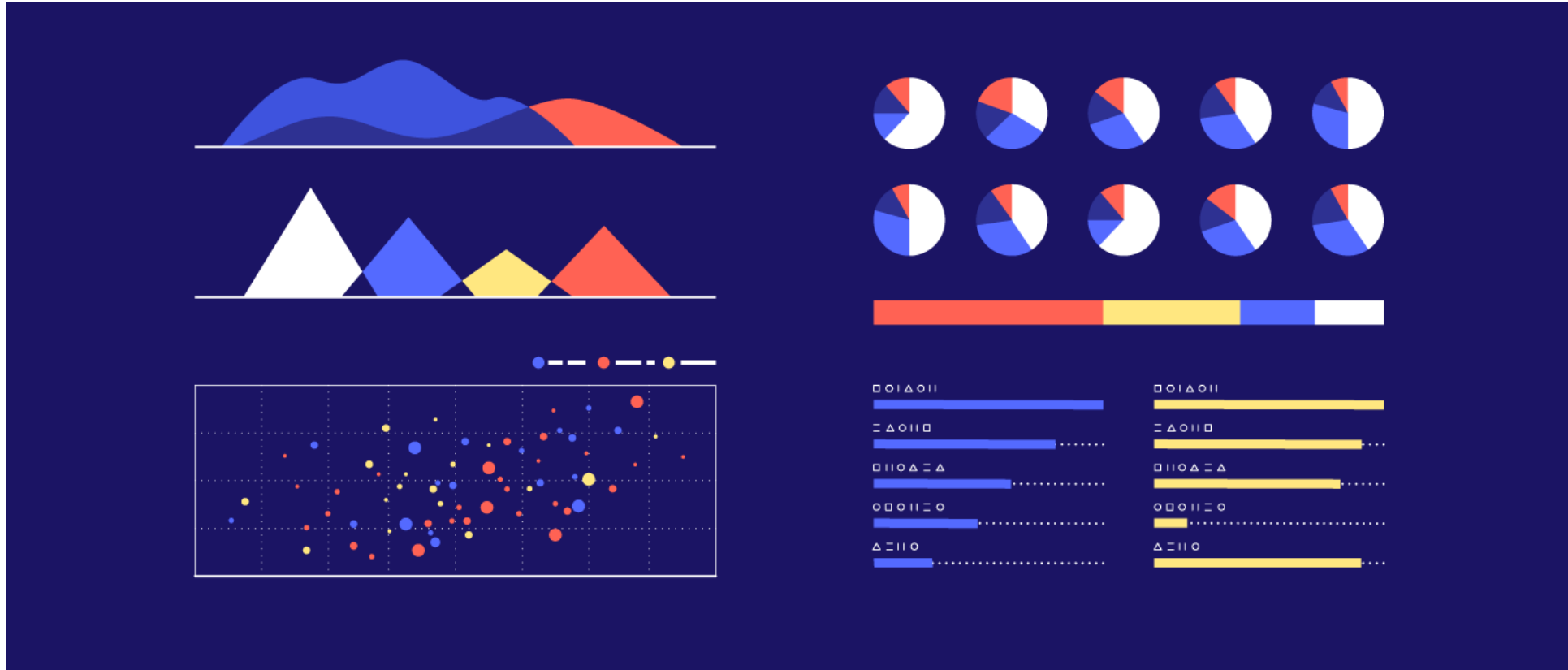


Many approaches: Statistics,
Decision Trees, Neural
Networks,
...

Outlier Detection



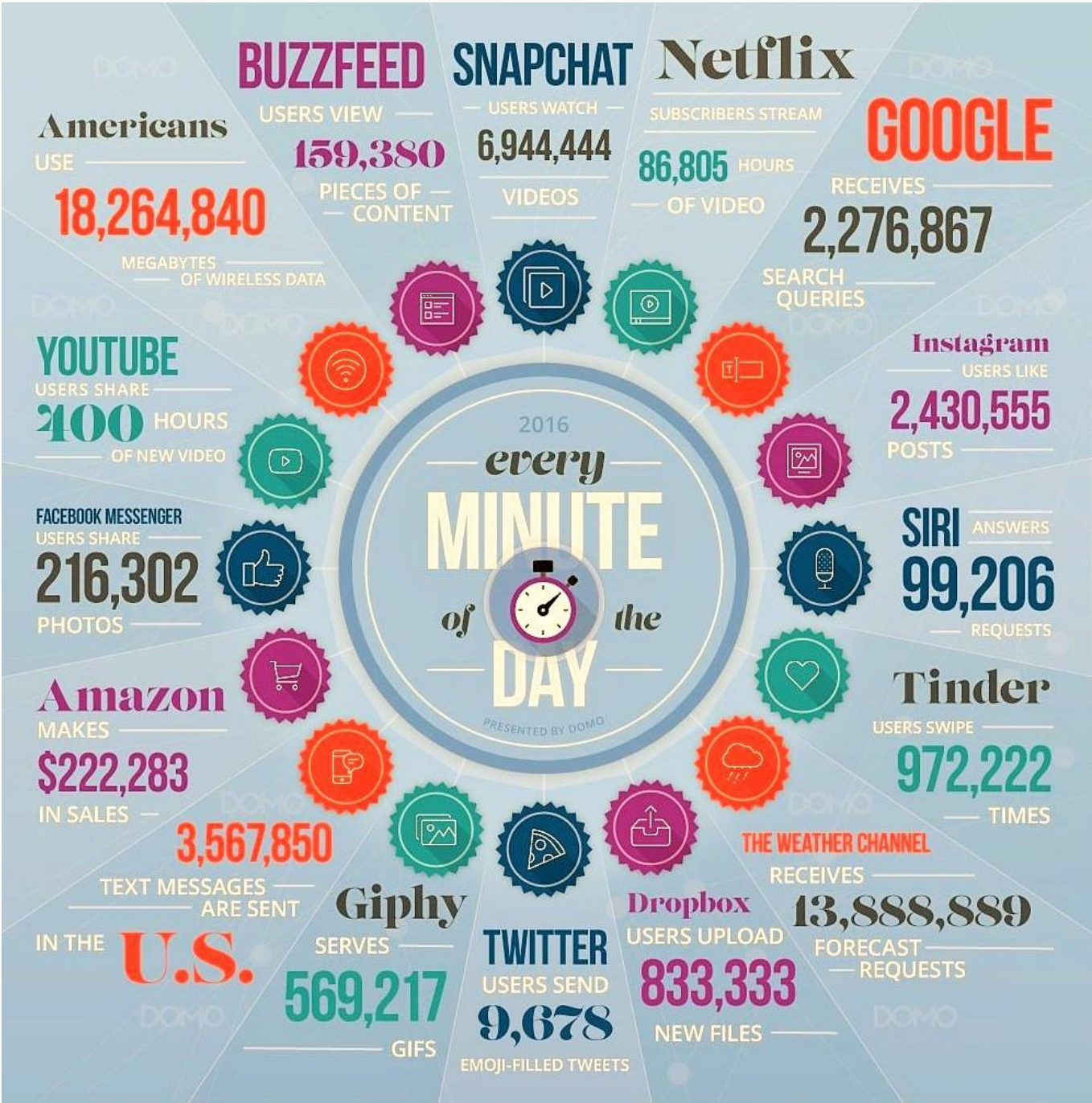
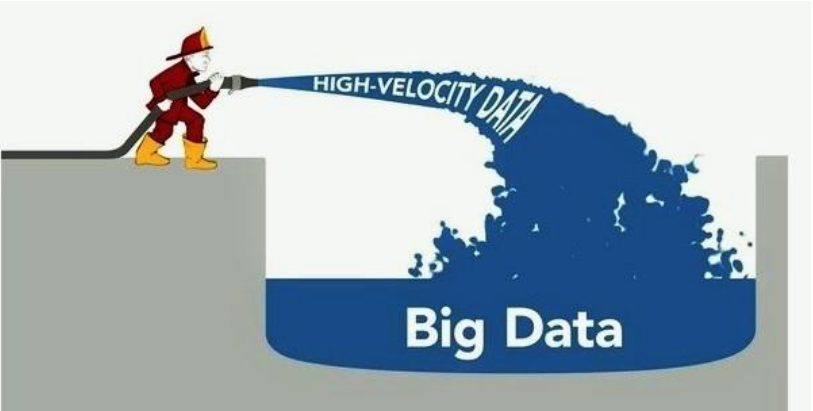
Data Visualization



Challenges in Data Mining

- Data is growing (Big data): Computational Challenges
- Data is distributed, changing, evolving: Data Streams
- Real World problems are complex: Text summarization
- Bonferroni Principle (Statistics)
- Privacy and ethical issues

Explosive Growth of Data



DATA NEVER SLEEPS 4.0

How much data is generated every minute? In the fourth annual edition of Data Never Sleeps, newcomers like Giphy and Facebook Messenger illustrate the rise of our multimedia messaging obsession, while veterans like Youtube and Snapchat highlight our insatiable appetite for video. Just how many GIFs, videos, and emoji-filled Tweets flood the internet every minute? See for yourself below.

Research On Data Mining

Journals and Conferences on Data Mining

- 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.

Next lecture:

- Statistical Description of Data
- Visualization