

# DATA MINING

# LECTURE 1

---

Introduction

# Intro

Instructor: Aris Anagnostopoulos (just Aris)

Web page: <http://aris.me>

Register to the mailing list

Lectures

Book: <http://infolab.stanford.edu/~ullman/mmds.html>

What do you need to know

Homeworks

Office hours

Exams

Collaboration policy

# What is data mining?

- After years of data mining there is still no unique answer to this question.



- A tentative definition:

Data mining is the use of **efficient** techniques for the analysis of **very large** collections of data and the extraction of **useful** and possibly **unexpected** patterns in data.



# Why do we need data mining?

- **Really, really huge amounts of raw data!!**
  - In the digital age, TB of data is generated by the second
    - Mobile devices, digital photographs, web documents.
    - Facebook updates, Tweets, Blogs, User-generated content
    - Transactions, sensor data, surveillance data
    - Queries, clicks, browsing
  - Cheap storage has made possible to maintain this data
- **Need to analyze the raw data to extract knowledge**

# Why do we need data mining?

- “The data is the computer”
  - Large amounts of data can be more powerful than complex algorithms and models
    - Google has solved many Natural Language Processing problems, simply by looking at the data
    - Example: misspellings, synonyms
  - Data is power!
    - Today, the collected data is one of the biggest assets of an online company
      - Query logs of Google
      - The friendship and updates of Facebook
      - Tweets and follows of Twitter
      - Amazon transactions
  - We need a way to harness the collective intelligence

# The data is also very **complex**

- Multiple **types** of data: tables, time series, images, graphs, etc
- **Spatial** and **temporal** aspects
- **Interconnected** data of different types:
  - From the mobile phone we can collect, location of the user, friendship information, check-ins to venues, opinions through twitter, images through cameras, queries to search engines

# Example: transaction data

- Billions of real-life customers:
  - WALMART: 20M transactions per day
  - AT&T 300 M calls per day
  - Credit card companies: billions of transactions per day.
- The point cards allow companies to collect information about specific users

# Example: document data

- Web as a document repository: estimated 50 billions of web pages
- Wikipedia: ~ 4.5 million articles (and counting)
- Online news portals: steady stream of 100's of new articles every day
- Twitter: >500 million tweets every day



# Example: network data

- Web: 50 billion pages linked via hyperlinks
- Facebook: 1.23 billion users
- Twitter: 243 million active users
- Instant messenger: ~1 billion users
- WhatsApp: 250 million users
  
- Blogs: 250 million blogs worldwide, presidential candidates run blogs

# Example: genomic sequences

- <http://www.1000genomes.org/page.php>
- Full sequence of 1000 individuals
- $3 \times 10^9$  nucleotides per person  $\rightarrow 3 \times 10^{12}$  nucleotides
- Lots more data in fact: medical history of the persons, gene expression data

# Example: environmental data

- Climate data (just an example)

<http://www.ncdc.gov/oa/climate/ghcn-monthly/index.php>

- “a database of temperature, precipitation and pressure records managed by the National Climatic Data Center, Arizona State University and the Carbon Dioxide Information Analysis Center”
- “6000 temperature stations, 7500 precipitation stations, 2000 pressure stations”
  - **Spatiotemporal** data

# Example: behavioral data

- Mobile phones today record a large amount of information about the user behavior
  - GPS records position
  - Camera produces images
  - Communication via phone and SMS
  - Text via facebook updates
  - Association with entities via check-ins
- Amazon collects all the items that you browsed, placed into your basket, read reviews about, purchased.
- Google and Bing record all your browsing activity via toolbar plugins. They also record the queries you asked, the pages you saw and the clicks you did.
- Data collected for millions of users on a daily basis

# So, what is Data?

- Collection of data **objects** and their **attributes**
- An attribute is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as **variable**, **field**, **characteristic**, or **feature**
- A collection of attributes describe an object
  - Object is also known as **record**, **point**, **case**, **sample**, **entity**, or **instance**

**Objects**

**Attributes**

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

**Size:** Number of objects

**Dimensionality:** Number of attributes

**Sparsity:** Number of populated object-attribute pairs

# Types of Attributes

- There are different types of attributes
  - **Categorical**
    - Examples: eye color, zip codes, words, rankings (e.g, good, fair, bad), height in {tall, medium, short}
    - **Nominal** (no order or comparison) vs **Ordinal** (order but not comparable)
  - **Numeric**
    - Examples: dates, temperature, time, length, value, count.
    - **Discrete** (counts) vs **Continuous** (temperature)
    - Special case: **Binary** attributes (yes/no, exists/not exists)

# Numeric Record Data

- 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 **n-by-d data matrix**, where there are **n** rows, one for each object, and **d** columns, one for each attribute

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

# Categorical Data

- Data that consists of a collection of records, each of which consists of a **fixed set** of **categorical** attributes

<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	High	No
2	No	Married	Medium	No
3	No	Single	Low	No
4	Yes	Married	High	No
5	No	Divorced	Medium	Yes
6	No	Married	Low	No
7	Yes	Divorced	High	No
8	No	Single	Medium	Yes
9	No	Married	Medium	No
10	No	Single	Medium	Yes



# Document Data

- 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.
  - **Bag-of-words** representation – no ordering

	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

# Transaction Data

- Each record (transaction) is a **set of items**.

<i>TID</i>	<i>Items</i>
1	B read, C oke, M ilk
2	B eer, B read
3	B eer, C oke, D iaper, M ilk
4	B eer, B read, D iaper, M ilk
5	C oke, D iaper, M ilk

- A set of items can also be represented as a **binary vector**, where each attribute is an item.
- A document can also be represented as a **set of words** (no counts)

**Sparsity**: average number of products bought by a customer

# Ordered Data

- Genomic **sequence** data

```
GGTTC CGCCTTCAGCCCCGCGCC  
CGCAGGGCCCGCCCCGCGCCGTC  
GAGAAGGGCCCGCCTGGCGGGCG  
GGGGGAGGCGGGGCCGCCCGAGC  
CCAACCGAGTCCGACCAGGTGCC  
CCCTCTGCTCGGCCTAGACCTGA  
GCTCATTAGGCGGCAGCGGACAG  
GCCAAGTAGAACACGCGAAGCGC  
TGGGCTGCCTGCTGCGACCAGGG
```

- Data is a long **ordered** string

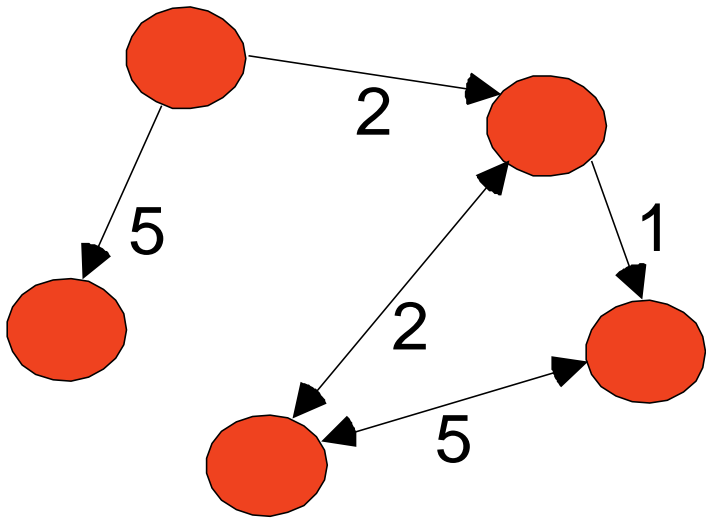
# Ordered Data

- Time series
  - Sequence of ordered (over “time”) numeric values.



# Graph Data

- Examples: Web graph and HTML Links



```
<a href="papers/papers.html#bbbb">  
Data Mining </a>
```

```
<li>
```

```
<a href="papers/papers.html#aaaa">  
Graph Partitioning </a>
```

```
<li>
```

```
<a href="papers/papers.html#aaaa">  
Parallel Solution of Sparse Linear System of Equations </a>
```

```
<li>
```

```
<a href="papers/papers.html#ffff">  
N-Body Computation and Dense Linear System Solvers
```

# Types of data

- **Numeric data**: Each object is a point in a multidimensional space
- **Categorical data**: Each object is a vector of categorical values
- **Set data**: Each object is a set of values (with or without counts)
  - Sets can also be represented as binary vectors, or vectors of counts
- **Ordered sequences**: Each object is an ordered sequence of values.
- **Graph data**

# What can you do with the data?

- Suppose that you are the owner of a supermarket and you have collected billions of **market basket** data. What information would you extract from it and how would you use it?

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Product placement

Catalog creation

Recommendations

- What if this was an online store?

# What can you do with the data?

- Suppose you are a search engine and you have a **toolbar log** consisting of
  - pages browsed,
  - queries,
  - pages clicked,
  - ads clicked

Ad click prediction

Query reformulations

each with a **user id** and a **timestamp**. What information would you like to get out of the data?



# What can you do with the data?

- Suppose you are a stock broker and you observe the fluctuations of multiple stocks over time. What information would you like to get out of your data?

