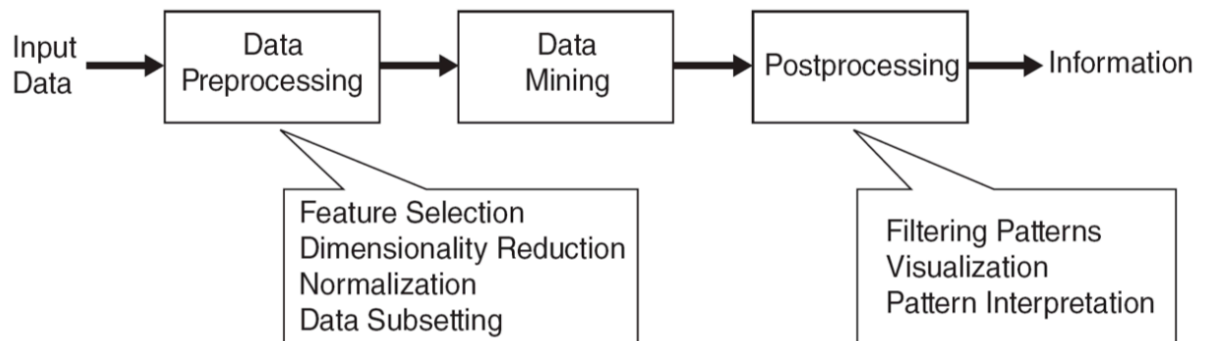


# MA 440 Data Mining

## Lecture 1: Introduction

### 1. Definition

**Data Mining:** Data mining is the process of extracting and discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal of extracting information (with intelligent methods) from a data set and transforming the information into a comprehensible structure for further use (1).



### 2. Why Data Mining ?

#### 2.1 Commercial Viewpoint

Lots of data is being collected and warehoused

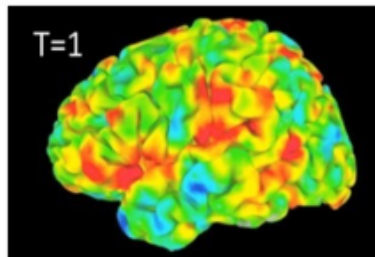
1. **Website data:** Google has Peta Bytes of web data. Facebook has billions of active users
2. **Purchases:** at department/ grocery stores, e-commerce, Amazon handles millions of visits/day
3. **Online Transactions:** Bank/Credit Card transactions



## 2.2 Scientific Viewpoint

Data is stored at enormous speed through

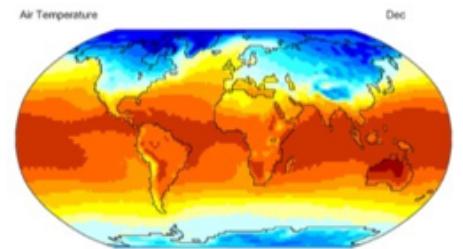
1. Scientific simulations/models
2. Telescopes and sensors on satellites
3. Earth system observations



**fMRI Data from Brain**



**Sky Survey Data**

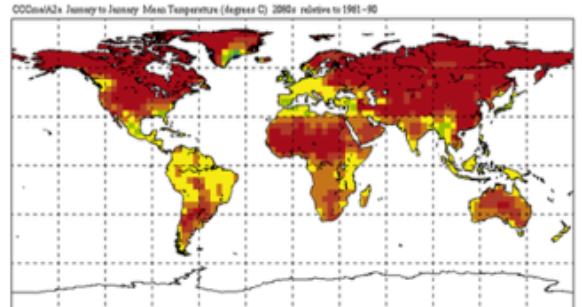


**Surface Temperature of Earth**

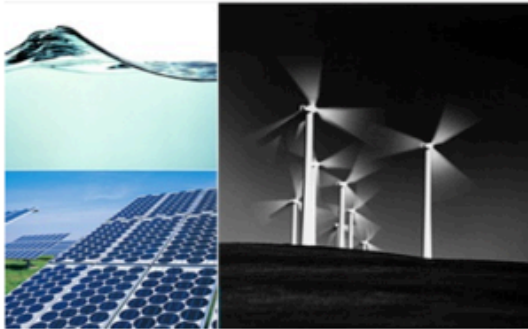
## 2.3. Opportunity to solve society's major problems



**Improving health care and reducing costs**



**Predicting the impact of climate change**



**Finding alternative/ green energy sources**



**Reducing hunger and poverty by increasing agriculture production**

## 2.4 Traditional statistical model not suitable any more

The scientific needs of information retrieval from modern datasets cannot be fulfilled by traditional statistical practices due to the following characteristics of these datasets:

1. Large Scale
2. High dimensional
3. Heterogeneous
4. Complex
5. Distributed

## 3. Data mining tasks

There are two main tasks:

1. **Supervised tasks:** You have some samples ( $X$ /independent variables) for which you know the labels ( $Y$ /dependent variable) and you want to predict the label ( $y$ ) at a new data sample( $x$ ). *For example you know the max-temperature ( $Y$ ) for the past month ( $X$ ) and you want to predict the temperature ( $y$ ) for the next day ( $x$ )*
2. **Unsupervised tasks:** You have some samples ( $X$ ) but no associated label( $Y$ ). Your goal is to find patterns in these samples that can be used to make decisions. *For example finding clusters of users on Amazon that buy similar items and then suggesting a new item to a customer which a similar customer has bought before.*

## 4. Further categorization of supervised learning tasks

### Regression

1. Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
2. Predicting sales amounts of new product based on advertising expenditure

### Classification

1. Classifying credit card transactions as legitimate or fraudulent
2. Classifying land covers (water bodies, urban areas, forests, etc.) using satellite data
3. Categorizing news stories as finance, weather, entertainment, sports, etc
4. Predicting tumor cells as benign or malignant

## 5. Further categorization of unsupervised learning tasks

## Clustering

1. Custom profiling for targeted marketing
2. Group related documents for browsing
3. Group stocks with similar price fluctuations

## Association

1. If a customer bought *milk* then they are likely to buy *eggs*
2. A customer who viewed movie Avengers 1, will likely view the movie Avengers 2

## Anomaly detection

1. Credit card fraud detection
2. Detecting changes in global forest cover

## References

1. [https://en.wikipedia.org/wiki/Data\\_mining\\_\(https://en.wikipedia.org/wiki/Data\\_mining\)](https://en.wikipedia.org/wiki/Data_mining_(https://en.wikipedia.org/wiki/Data_mining))
2. [https://www-users.cse.umn.edu/~kumar001/dmbook/index.php\\_\(https://www-users.cse.umn.edu/~kumar001/dmbook/index.php\)](https://www-users.cse.umn.edu/~kumar001/dmbook/index.php_(https://www-users.cse.umn.edu/~kumar001/dmbook/index.php))

In [ ]: