

.....

If you check out on Wikipedia, you will find a fairly abstract definition of machine learning:

“Machine learning explores the study and construction of **algorithms** that can **learn** from and make **predictions** on data. Such algorithms operate by building a model from example inputs **in order to make decisions**, rather than following strictly static program instructions.”

Outline of this lecture

- Machine Learning and Data Mining
- Type of Learning
- Steps of Machine Learning
- Data Munging
- Understanding Data
- Data Munging Tasks
- Descriptive Statistics

.....

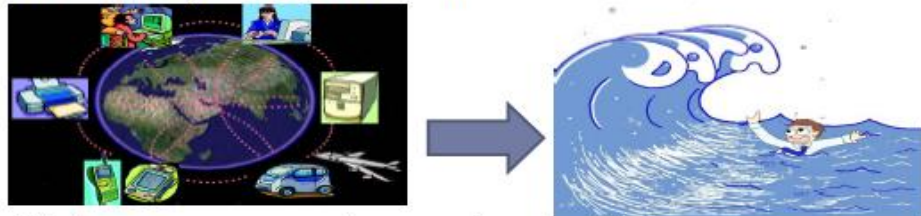
Machine Learning and Data Mining

Why Data Mining?

- ▶ Explosive Growth of Data: from terabytes to petabytes
- ▶ Data Collections and Data Availability
 - Crawlers, database systems, Web, etc.

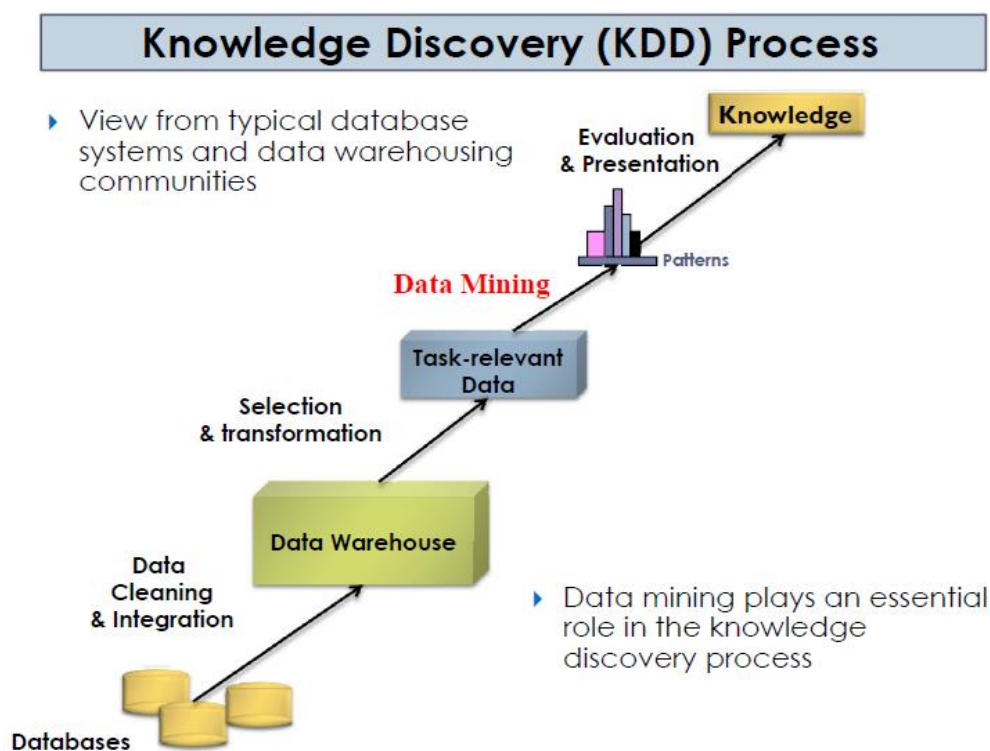
→ Sources

- Business: Web, e-commerce, transactions, etc.
- Science: Remote sensing, bioinformatics, etc.
- Society and everyone: news, YouTube, etc.



- ▶ **Problem:** We are drowning in data, but starving for knowledge!
- ▶ **Solution:** Use Data Mining tools for Automated Analysis of massive data sets

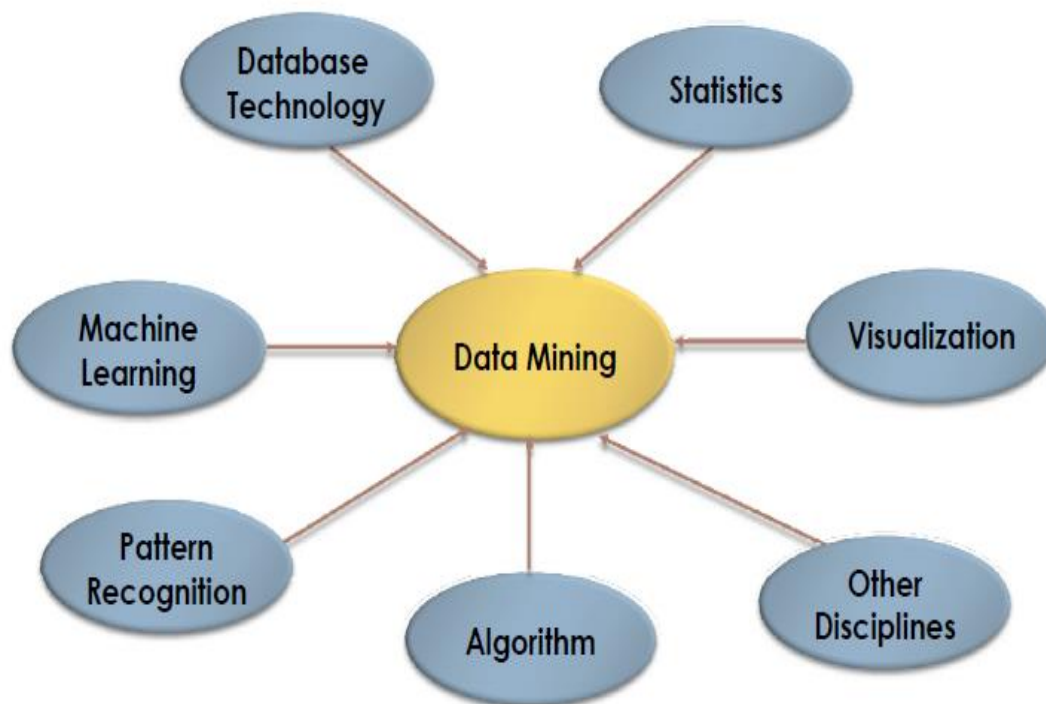
Data mining (knowledge discovery from data)



At the summary:

Data Mining is a process of extracting knowledge from data

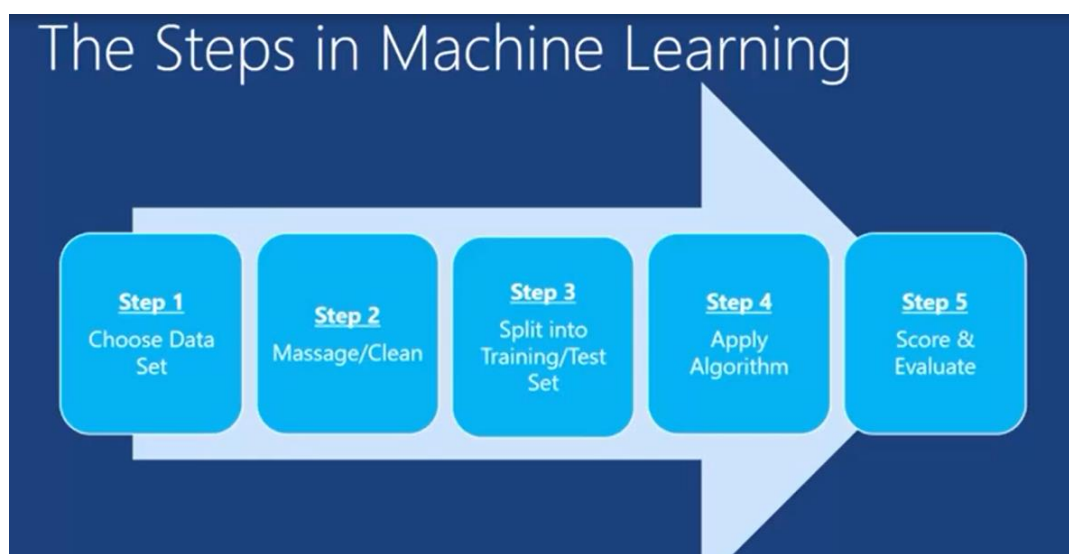
- **Data to be mined** can be of any type: Relational Databases, Advanced databases, etc.
- **Knowledge to be discovered:** Frequent patterns, correlations, associations, classification, prediction, clustering.
- **Techniques to be used:** Statistics, machine learning, visualization, etc.



Types of Learning

- Supervised learning --- where the algorithm generates a function that maps inputs to desired outputs. One standard formulation of the supervised learning task is the classification problem: the learner is required to learn (to approximate the behavior of) a function which maps a vector into one of several classes by looking at several input-output examples of the function.
- Unsupervised learning --- which models a set of inputs: labeled examples are not available.
- Semi-supervised learning --- which combines both labeled and unlabeled examples to generate an appropriate function or classifier.
- Reinforcement learning --- where the algorithm learns a policy of how to act given an observation of the world. Every action has some impact in the environment, and the environment provides feedback that guides the learning algorithm.
- Transduction --- similar to supervised learning, but does not explicitly construct a function: instead, tries to predict new outputs based on training inputs, training outputs, and new inputs.
- Learning to learn --- where the algorithm learns its own inductive bias based on previous experience.

Steps of Machine Learning



Data Munging

- ❖ Data Munging is the transformation of raw data to a useable format.
- ❖ Many datasets are not readily available for analysis.
- ❖ Data needs to be transformed or cleaned first.
- ❖ This process is often the most difficult and the most time consuming.

Understanding Data

price	lotsize	bedrooms	bathrms	stories	driveway	recroom	fullbase
\$42,000.00	5850	3	1	2	yes	no	yes
\$38,500.00	4000	2	1	1	yes	no	no
\$49,500.00	3060	3	1	1	yes	no	no
\$60,500.00	6650	3	1	2	yes	yes	no
	6360	2	1	1	yes	no	no
	4160	3	1	1	yes	yes	yes
\$66,000.00	3880	3	2	2	yes	no	yes
\$69,000.00	4160	3	1	3	yes	no	no
\$83,800.00	4800	3	1	1	yes	yes	yes
\$88,500.00	5500	3	2	4	yes	yes	no
\$90,000.00	7200	3	2	1	yes	no	yes
\$30,500.00	3000	2	1	1	no	no	no
\$27,000.00	1700	15	1	2	yes	7	no

Missing Values →

Non-Numeric

price	lotsize	bedrooms	bathrms	stories	driveway	recroom	fullbase
\$42,000.00	5850	3	1	2	yes	no	yes
\$38,500.00	4000	2	1	1	yes	no	no
\$49,500.00	3060	3	1	1	yes	no	no
\$60,500.00	6650	3	1	2	yes	yes	no
	6360	2	1	1	yes	no	no
	4160	3	1	1	yes	yes	yes
\$66,000.00	3880	3	2	2	yes	no	yes
\$69,000.00	4160	3	1	3	yes	no	no
\$83,800.00	4800	3	1	1	yes	yes	yes
\$88,500.00	5500	3	2	4	yes	yes	no
\$90,000.00	7200	3	2	1	yes	no	yes
\$30,500.00	3000	2	1	1	no	no	no
\$27,000.00	1700	15	1	2	yes	7	no

Outlier

Error

Data Munging Tasks

1. Renaming Variables

T1K5X
\$42,000.00
\$38,500.00
\$49,500.00
\$60,500.00

→

Price
\$42,000.00
\$38,500.00
\$49,500.00
\$60,500.00

2. Data Type conversion

- Depending upon the modeling task at hand and the software, the data may need to be expressed in a specific format in order to process correctly.

Date
January 1st, 2014

→

Date
1/1/2014

3. Encoding Data

- There are times when we need to change the underlying contents in a variable to prepare them for analytics. Ex. Qualitative to Quantitative.

driveway	driveway
yes	1
no	0
yes	1

- If we are using categorical variables, we need to clean them to get rid of non response categories like "I don't know", "no answer", "n/a", etc... We also need to order the encoding of categories (potentially reverse valence) to ensure that models are built and interpreted correctly.

Response	Response	Response	Response
Strongly Agree	Strongly Agree	Strongly Agree	4
Strongly Disagree	Agree	Agree	3
Agree	No Preference	Disagree	2
Disagree	Disagree	Strongly Disagree	1
No Preference	Strongly Disagree		

4. Merging Data Sets

- It is quite rare that you will have a dataset readily constructed for analysis. This may require some data manipulation and merging in order to get the data in the correct form.

ID	price	lotsize	ID	bedrooms	bathrms	stories	garagepl	ID	price	lotsize	bedrooms	bathrms	stories	garagepl
A1234	\$42,000.00	5850	A1234	3	1	2	1	A1234	\$42,000.00	5850	3	1	2	1
A1235	\$38,500.00	4000	A1235	2	1	1	0	A1235	\$38,500.00	4000	2	1	1	0
A1236	\$49,500.00	3060	A1236	3	1	1	0	A1236	\$49,500.00	3060	3	1	1	0
A1237	\$60,500.00	6650	A1237	3	1	2	0	A1237	\$60,500.00	6650	3	1	2	0
A1238	\$41,000.00	6360	A1238	2	1	1	0	A1238	\$41,000.00	6360	2	1	1	0
A1239	\$66,000.00	4160	A1239	3	1	1	0	A1239	\$66,000.00	4160	3	1	1	0
A1240	\$66,000.00	3880	A1240	3	2	2	2	A1240	\$66,000.00	3880	3	2	2	2
A1241	\$69,000.00	4160	A1241	3	1	3	0	A1241	\$69,000.00	4160	3	1	3	0
A1242	\$83,800.00	4800	A1242	3	1	1	0	A1242	\$83,800.00	4800	3	1	1	0
A1243	\$88,500.00	5500	A1243	3	2	4	1	A1243	\$88,500.00	5500	3	2	4	1
A1244	\$90,000.00	7200	A1244	3	2	1	3	A1244	\$90,000.00	7200	3	2	1	3
A1245	\$30,500.00	3000	A1245	2	1	1	0	A1245	\$30,500.00	3000	2	1	1	0
A1246	\$27,000.00	1700	A1246	3	1	2	0	A1246	\$27,000.00	1700	3	1	2	0

5. Imputation

If there are missing values in a column, these cannot be left unattended. We must decide if we want to:

- ❖ Remove the observation from the dataset
- ❖ Calculate a value for the null (impute). This usually is determined with the mean or median, however, a more advanced version can use a multiple linear regression formula.

price
\$ 42,000.00
\$ 38,500.00
\$ 49,500.00
\$ 60,500.00
\$ 66,000.00
\$ 69,000.00
\$ 83,800.00
\$ 88,500.00
\$ 90,000.00
\$ 30,500.00
\$ 27,000.00



price
\$42,000.00
\$38,500.00
\$49,500.00
\$60,500.00
\$58,660.00
\$58,660.00
\$66,000.00
\$69,000.00
\$83,800.00
\$88,500.00
\$90,000.00
\$30,500.00
\$27,000.00

Mean = 58,660

Median = 60,500

6. Handling Anomalous Values



- ❖ Outliers are data points that deviate significantly from the spread or distribution of other similar data points. These can typically be detected through the use of scatterplots.
- ❖ Many times we will delete the entry with an outlier to achieve normality in the dataset.
- ❖ In some instances, an outlier can be imputed but this must be approached with caution.

Descriptive Statistics

Mean

- ❖ The sum of the observations divided by the total number of observations. It is the most common indicator of central tendency of a variable.

$$\bar{X} = \frac{\sum X_i}{n}$$

Median

- ❖ To get the median, we need to sort the data from lowest to highest. The median is the number in the middle of the data

2 2 5 6 7 8 9

Mode

- ❖ Refers to the most frequent of commonly occurring number within the variable.

2 2 5 6 7 8 9

Variance

- ❖ Measures the dispersion of the data from the mean.

$$s^2 = \frac{\sum (X_i - \bar{X})^2}{(n-1)}$$

Standard Deviation

- ❖ The Standard Deviation is the Squared Root of the Variance. This indicates how close the data is to the mean.

$$s = \sqrt{\frac{\sum (X_i - \bar{X})^2}{(n-1)}}$$

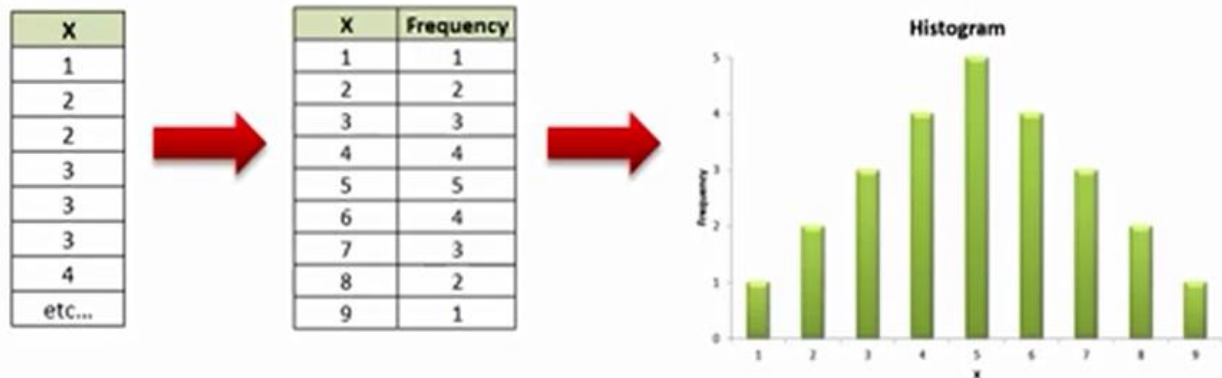
Range

- ❖ Range is a measure of dispersion. It is the simple difference between the largest and smallest values.

2 2 5 6 7 8 9
2 to 9

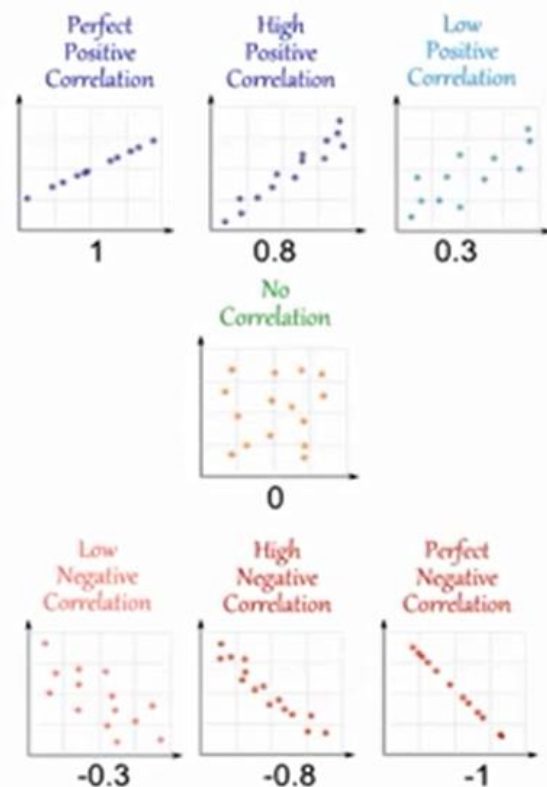
Histogram

- Histograms are a graphical display of data using bars of different heights. This allows us to evaluate the shape of the underlying distribution. Essentially, a histogram is a bar chart that groups numbers into ranges or bins.



Correlations

- When two sets of data are strongly linked together we say they have a high correlation.
- Correlation is **Positive** when the values **increase** together, and
- Correlation is **Negative** when one value **decreases** as the other increases.
- Correlation can have a value ranging from:
 - 1 is perfect positive correlation
 - 0 implies that there is no correlation
 - 1 is perfect negative correlation



In Lab:

- Create local database in C# .
- Add (Chart) tool to your form design,
- Implement some of descriptive statistic, [see page 9](#)
- Read English script, then show the histogram of English letters in it, [see page 10](#).

Next Lecture

Machine Learning Algorithms