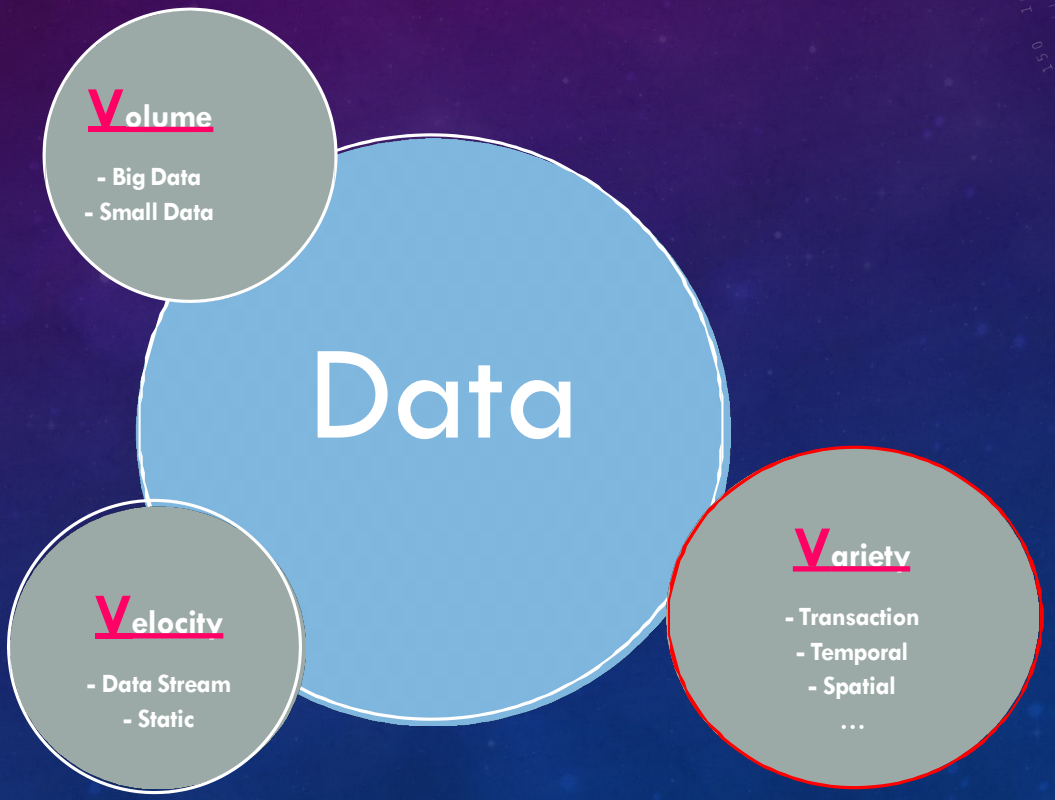
DATA PRE-PROCESSING

Dr K Kotecha
https://www.linkedin.com/in/ketankotecha/

Data: Challenges



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	Re fund	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
  - Road map connecting cities(S)
  - Tracking moving object(SST)
- 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

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

- Tim

Yahoo Finance Website

BIOLOGICAL SEQUENCE DATA

SCIENTIFIC SKILLS EXERCISE

Analyzing Polypeptide 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

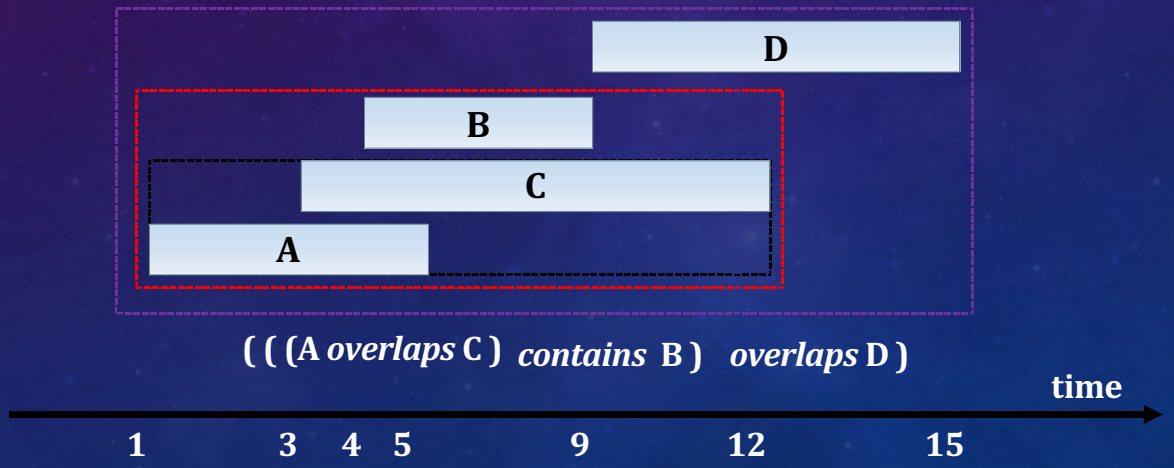
1. Scan the monkey and gibbon sequences, letter by letter, circling any amino acids that do not match the human sequence. (a) How many amino acids differ between the monkey and the human sequences? (b) Between the gibbon and human?
2. For each nonhuman species, what percent of its amino acids are identical 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 support 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

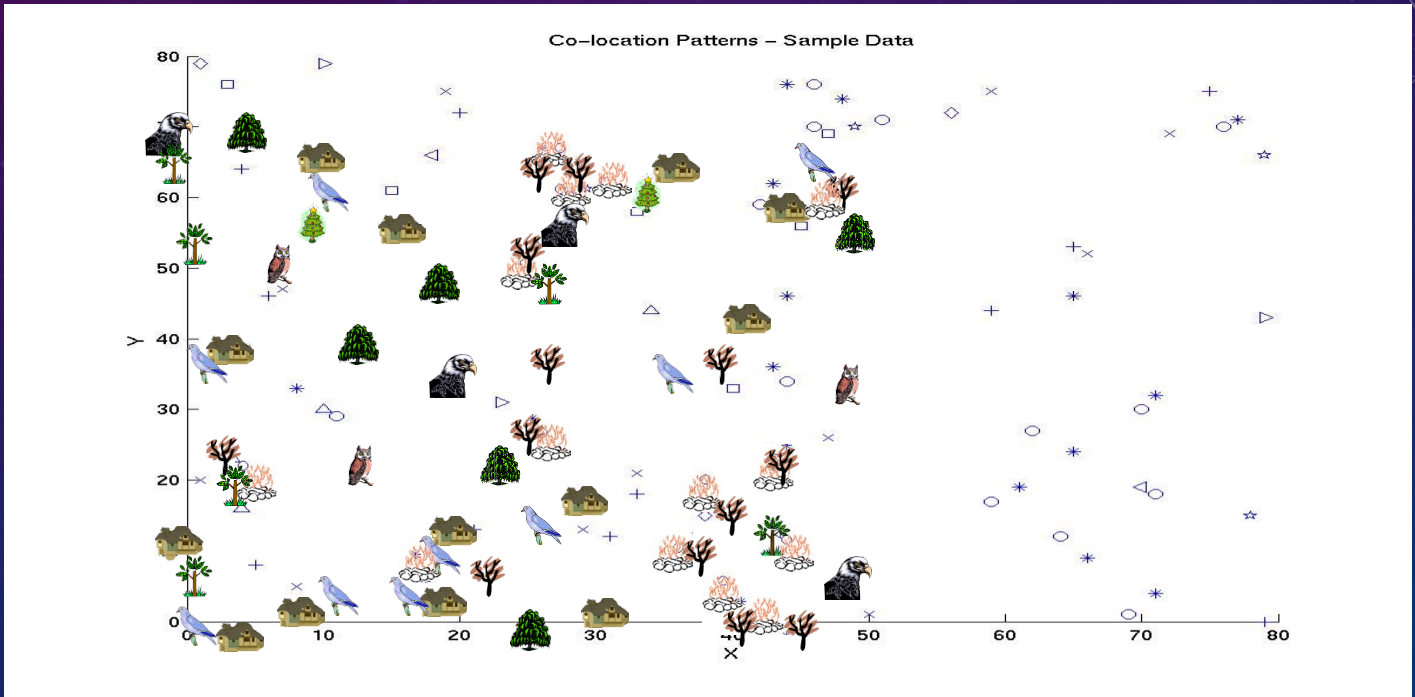
INTERVAL DATA

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

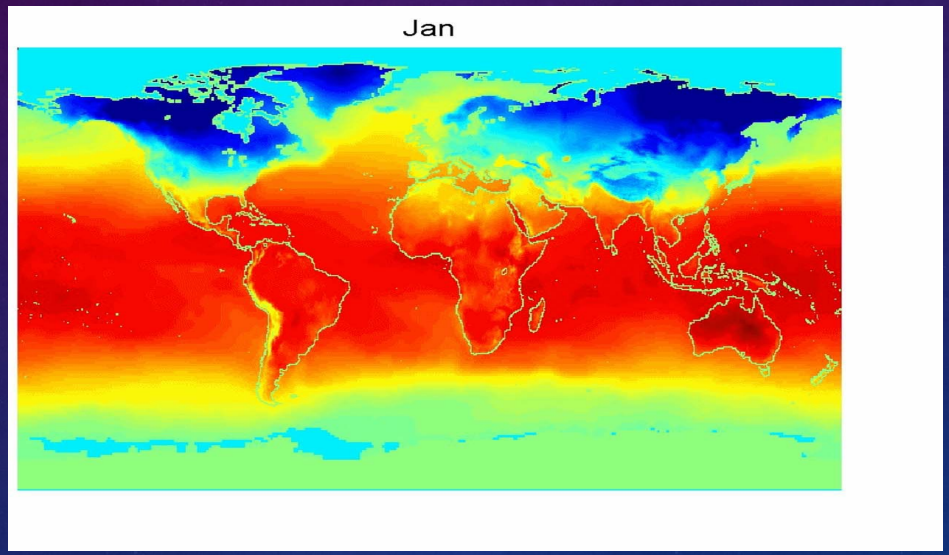


Interval Patient Data

SPATIAL & SPATIAL-TEMPORAL DATA



Spatial & Spatial-Temporal Data



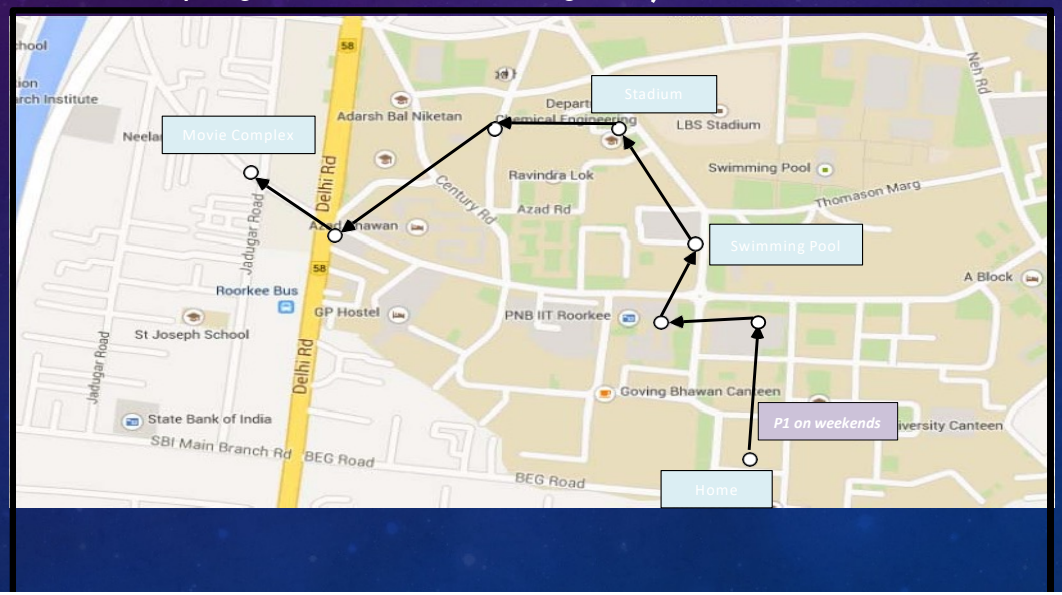
Average Monthly Temperature of land and ocean

Spatial & Spatial-Temporal Data

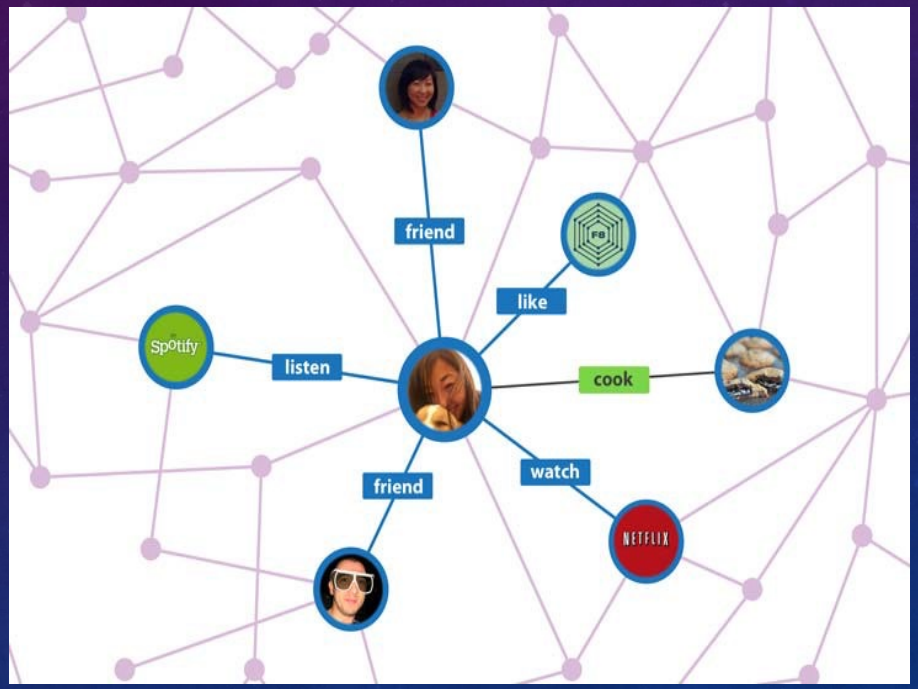
Corona Data set

User Movement Data

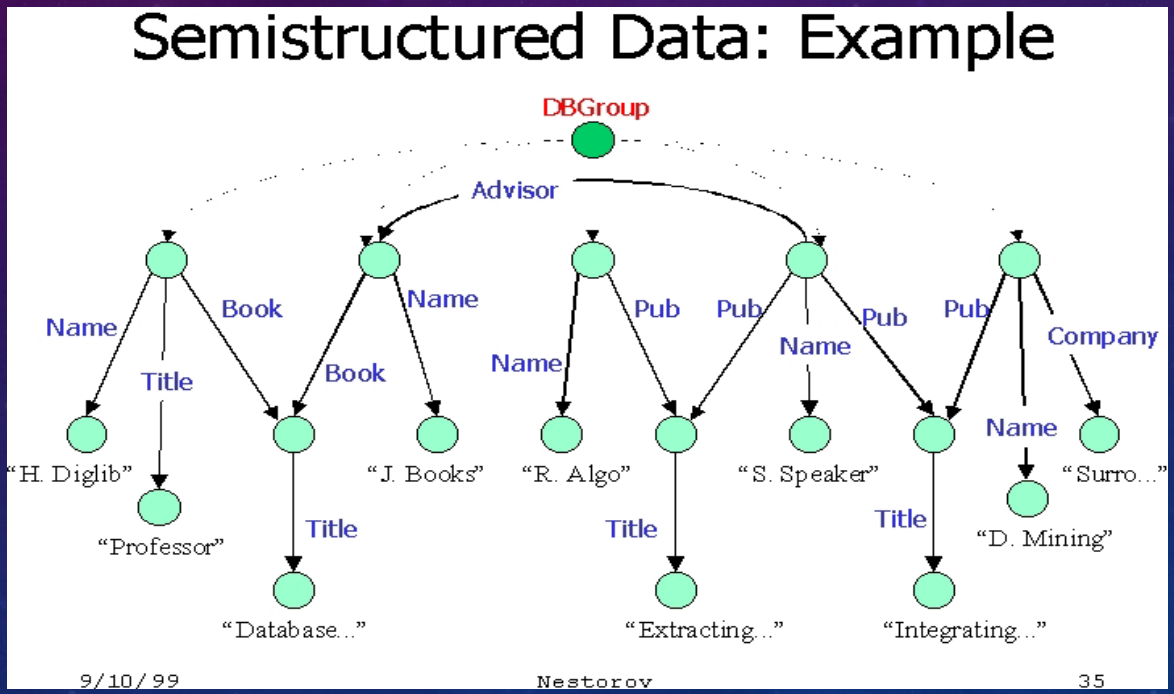
- *Trajectory*
  - *Movement trail of a user*
  - *Sampling Points: <latitude, longitude, time>*



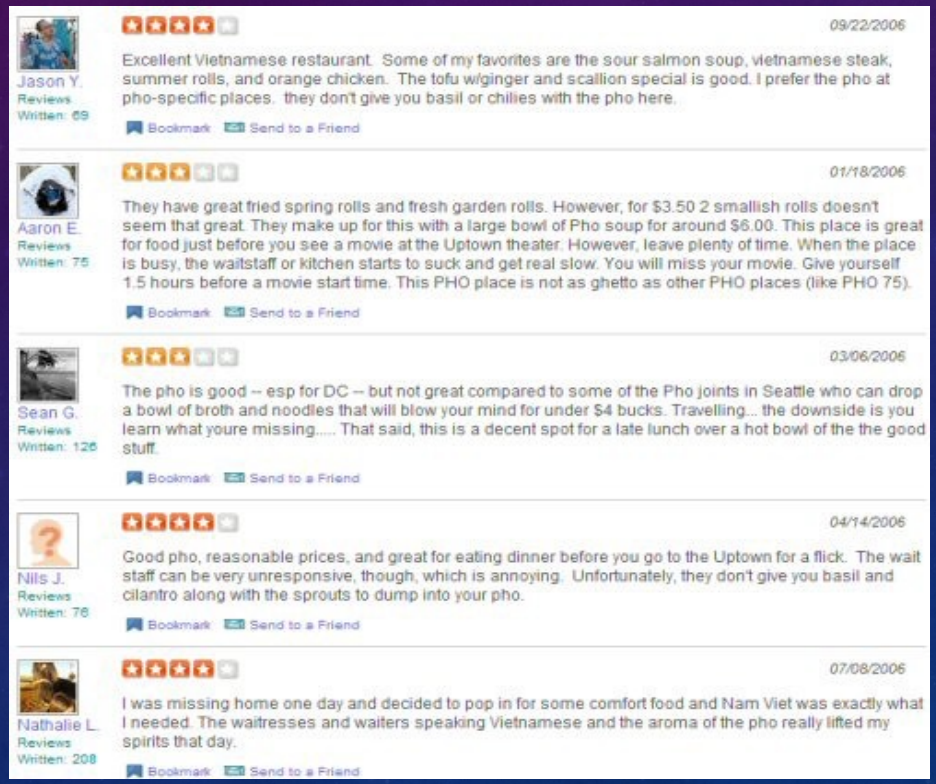
Graph Data



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

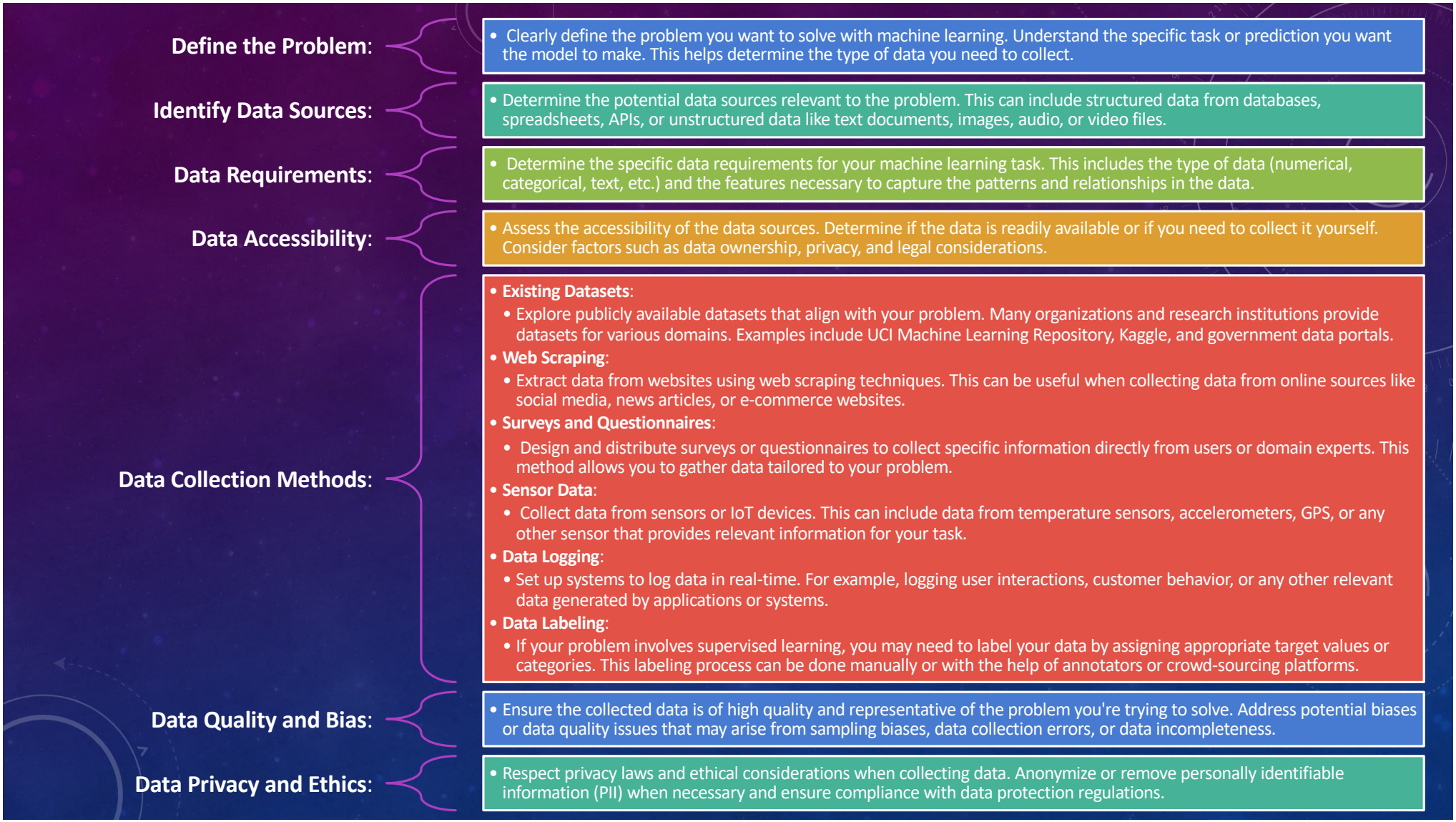
Sean G.
Reviews Written: 126
03/06/2006
The pho is good -- esp for DC -- but not great compared to some of the Pho joints in Seattle who can drop a bowl of broth and noodles that will blow your mind for under $4 bucks. Travelling... the downside is you learn what youre missing..... That said, this is 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

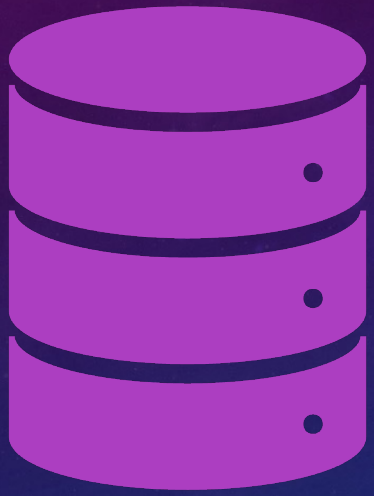
Nathalie L.
Reviews Written: 208
07/08/2006
I was missing home one day and decided to pop in for some comfort food and Nam Viet was exactly what I needed. The waitresses and waiters speaking Vietnamese and the aroma of the pho really lifted my spirits that day.
Bookmark Send to a Friend

Data Collection:	• Relevant and representative data is gathered for the problem at hand. This data serves as the training data for the machine learning model.
Data Pre-processing:	• The collected data is cleaned, transformed, and prepared for analysis. This step may involve handling missing values, removing outliers, and encoding categorical variables.
Feature Extraction and Selection:	• The most informative and relevant features are selected from the dataset or engineered based on domain knowledge. This step aims to represent the data in a way that captures the underlying patterns and relationships.
Model Training:	• The selected machine learning algorithm is trained on the prepared training data. During training, the algorithm learns from the data and adjusts its internal parameters to minimize errors and improve performance.
Model Evaluation:	• The trained model is evaluated using test data that the model has not seen during training. Evaluation metrics such as accuracy, precision, recall, or mean squared error are used to assess the model's performance and generalization capabilities.
Model Deployment	• If the model performs well, it can be deployed to make predictions or take actions on new, unseen data. This can involve integrating the model into a larger system or making it accessible through an API.

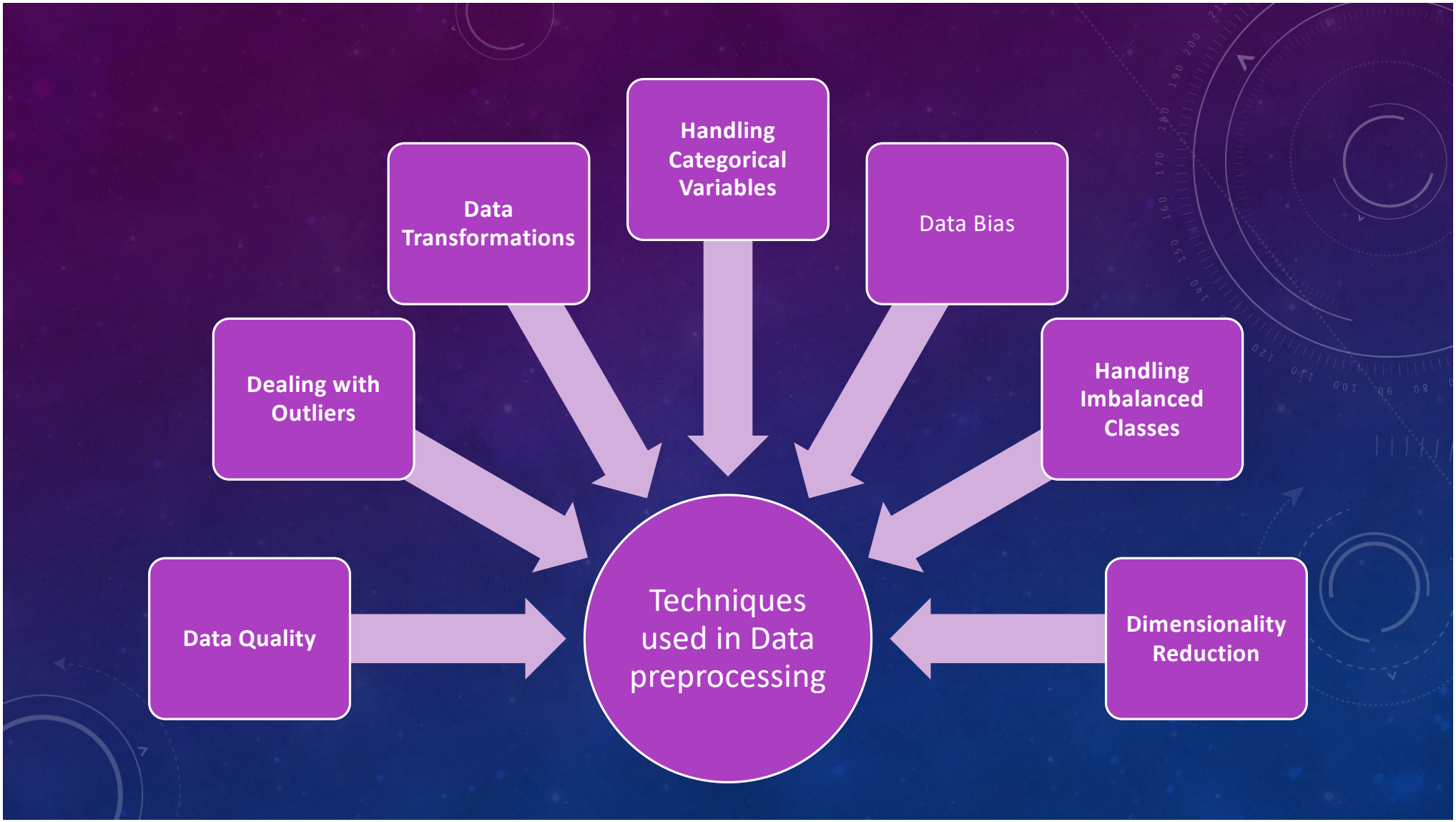

DATA COLLECTION


Define the Problem:
• Clearly define the problem you want to solve with machine learning. Understand the specific task or prediction you want the model to make. This helps determine the type of data you need to collect.
Identify Data Sources:
• Determine the potential data sources relevant to the problem. This can include structured data from databases, spreadsheets, APIs, or unstructured data like text documents, images, audio, or video files.
Data Requirements:
• Determine the specific data requirements for your machine learning task. This includes the type of data (numerical, categorical, text, etc.) and the features necessary to capture the patterns and relationships in the data.
Data Accessibility:
• Assess the accessibility of the data sources. Determine if the data is readily available or if you need to collect it yourself. Consider factors such as data ownership, privacy, and legal considerations.
Data Collection Methods:
• Existing Datasets:
• Explore publicly available datasets that align with your problem. Many organizations and research institutions provide datasets for various domains. Examples include UCI Machine Learning Repository, Kaggle, and government data portals.
• Web Scraping:
• Extract data from websites using web scraping techniques. This can be useful when collecting data from online sources like social media, news articles, or e-commerce websites.
• Surveys and Questionnaires:
• Design and distribute surveys or questionnaires to collect specific information directly from users or domain experts. This method allows you to gather data tailored to your problem.
• Sensor Data:
• Collect data from sensors or IoT devices. This can include data from temperature sensors, accelerometers, GPS, or any other sensor that provides relevant information for your task.
• Data Logging:
• Set up systems to log data in real-time. For example, logging user interactions, customer behavior, or any other relevant data generated by applications or systems.
• Data Labeling:
• If your problem involves supervised learning, you may need to label your data by assigning appropriate target values or categories. This labeling process can be done manually or with the help of annotators or crowd-sourcing platforms.
Data Quality and Bias:
• Ensure the collected data is of high quality and representative of the problem you're trying to solve. Address potential biases or data quality issues that may arise from sampling biases, data collection errors, or data incompleteness.
Data Privacy and Ethics:
• Respect privacy laws and ethical considerations when collecting data. Anonymize or remove personally identifiable information (PII) when necessary and ensure compliance with data protection regulations.

DATA PRE-PROCESSING



- Data preprocessing is an essential step in machine learning that involves transforming raw data into a suitable format for model training. It aims to address issues such as missing values, outliers, inconsistent formats, and other data irregularities that can adversely affect the performance and accuracy of machine learning models.


Handling Categorical Variables
Data Transformations
Data Bias
Dealing with Outliers
Handling Imbalanced Classes
Data Quality
Techniques used in Data preprocessing
Dimensionality Reduction

DATA QUALITY

- Examples of data quality problems:
  - Noise and outliers
  - Missing values
  - Duplicate data

A mistake or a millionaire?

Missing values

Inconsistent duplicate entries

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	10000K	Yes
6	No	NULL	60K	No
7	Yes	Divorced	220K	NULL
8	No	Single	85K	Yes
9	No	Married	90K	No
9	No	Single	90K	No

Idea 1: Skip data points with missing values

x

N = 9, 3 features

Credit	Term	Income	y
excellent	3 yrs	high	safe
fair	?	low	risky
fair	3 yrs	high	safe
poor	5 yrs	high	risky
excellent	3 yrs	low	risky
fair	5 yrs	high	safe
poor	3 yrs	low	risky
poor	3 yrs	?	safe
fair	?	high	safe

Skip data points with missing values

h(x)

N = 6, 3 features

Credit	Term	Income	y
excellent	3 yrs	high	safe
fair	3 yrs	high	safe
poor	5 yrs	high	risky
excellent	3 yrs	low	risky
fair	5 yrs	high	safe
poor	3 yrs	low	risky

Idea 2: Skip features with missing values

x

N = 9, 3 features

Credit	Term	Income	y
excellent	3 yrs	high	safe
fair	?	low	risky
fair	3 yrs	high	safe
poor	?	high	risky
excellent	?	low	risky
fair	5 yrs	high	safe
poor	?	high	risky
poor	?	low	safe
fair	?	high	safe

Skip features with many missing values

h(x)

N = 9, 2 features

Credit	Income	y
excellent	high	safe
fair	low	risky
fair	high	safe
poor	high	risky
excellent	low	risky
fair	high	safe
poor	high	risky
poor	low	safe
fair	high	safe

Idea 2: Imputation/Substitution

N = 9, 3 features

Credit	Term	Income	y
excellent	3 yrs	high	safe
fair	?	low	risky
fair	3 yrs	high	safe
poor	5 yrs	high	risky
excellent	3 yrs	low	risky
fair	5 yrs	high	safe
poor	3 yrs	high	risky
poor	?	low	safe
fair	?	high	safe

Fill in each missing value with a calculated guess

N = 9, 3 features

Credit	Term	Income	y
excellent	3 yrs	high	safe
fair	3 yrs	low	risky
fair	3 yrs	high	safe
poor	5 yrs	high	risky
excellent	3 yrs	low	risky
fair	5 yrs	high	safe
poor	3 yrs	high	risky
poor	3 yrs	low	safe
fair	3 yrs	high	safe

Example: Replace ? with most common value

3 year loans: 4
5 year loans: 2

Credit	Term	Income	y
excellent	3 yrs	high	safe
fair	?	low	risky
fair	3 yrs	high	safe
poor	5 yrs	high	risky
excellent	3 yrs	low	risky
fair	5 yrs	high	safe
poor	3 yrs	high	risky
poor	?	low	safe
fair	?	high	safe

Credit	Term	Income	y
excellent	3 yrs	high	safe
fair	3 yrs	low	risky
fair	3 yrs	high	safe
poor	5 yrs	high	risky
excellent	3 yrs	low	risky
fair	5 yrs	high	safe
poor	3 yrs	high	risky
poor	3 yrs	low	safe
fair	3 yrs	high	safe

Common (simple) rules for purification by imputation

Credit	Term	Income	y
excellent	3 yrs	high	safe
fair	?	low	risky
fair	3 yrs	high	safe
poor	5 yrs	high	risky
excellent	3 yrs	low	risky
fair	5 yrs	high	safe
poor	3 yrs	high	risky
poor	?	low	safe
fair	?	high	safe

Impute each feature with missing values:

1. Categorical features use mode: Most popular value (mode) of non-missing x_i
2. Numerical features use average or median: Average or median value of non-missing x_i

Many advanced methods exist,
e.g., expectation-maximization (EM) algorithm

DATA TRANSFORMATION

Data has an attribute values
Then,

Can we compare these attribute values?

For Example: Compare following two records (1) (5.9 ft, 50 Kg)
(2) (4.6 ft, 55 Kg)
Vs.
(3) (5.9 ft, 50 Kg)
(4) (5.6 ft, 56 Kg)

We need Data Transformation to makes different dimension(attribute) records comparable
…

WHY? TO AVOID BIASING TO WARDS ONE FEATURE

	Student	CGPA	Salary '000
0	1	3.0	60
1	2	3.0	40
2	3	4.0	40
3	4	4.5	50
4	5	4.2	52

to

	Student	CGPA	Salary '000
0	1	-1.184341	1.520013
1	2	-1.184341	-1.100699
2	3	0.416120	-1.100699
3	4	1.216350	0.209657
4	5	0.736212	0.471728

DATA TRANSFORMATION TECHNIQUES

- Normalization: scaled to fall within a small, specified range.
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling

- Discretization

DATA TRANSFORMATION: NORMALIZATION

- min-max normalization $v' = \frac{v - min}{max - min}(new\_max - new\_min) + new\_min$
- z-score normalization

$$v' = \frac{v - mean}{stand\_dev}$$

- normalization by decimal scaling

$$v' = \frac{v}{10^j}$$ Where j is the smallest integer such that Max($|v'|$)<1

MIN MAX NORMALIZATION

Suppose the minimum and maximum value for an attribute profit(P) are Rs. 10, 000 and Rs. 100, 000. We want to plot the profit in the range [0, 1]. Using min-max normalization the value of Rs. 20, 000 for attribute profit can be plotted to: so v' = 0.11

- min-max normalization $v' = \frac{v - min}{max - min}(new\_max - new\_min) + new\_min$

$$\frac{20000 - 10000}{100000 - 10000}(1 - 0) + 0 = 0.11$$

Z- SCORE NORMALIZATION

- Let mean of an attribute P = 60, 000, Standard Deviation = 10, 000, for the attribute P. Using z-score normalization, a value of 85000 for P can be transformed to:

$$\frac{v - mean}{stand\_dev}$$

$$\frac{85000 - 60000}{10000} = 2.50$$

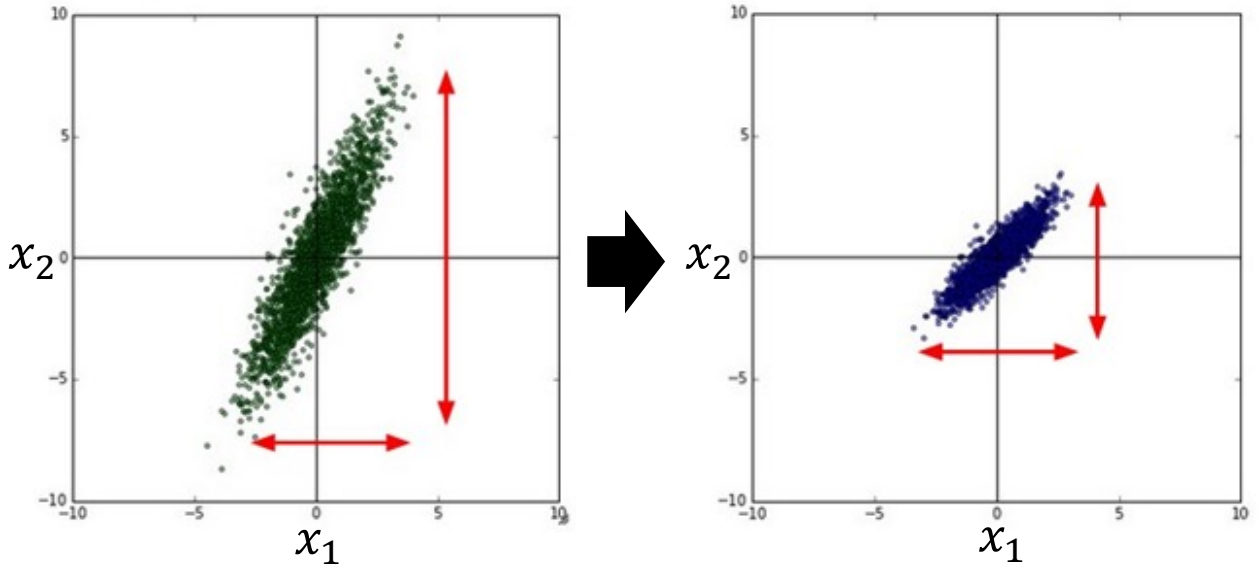
DECIMAL SCALING

- Suppose: Values of an attribute P varies from -99 to 99.
- The maximum absolute value of P is 99.
- For normalizing the values we divide the numbers by 100 (i.e., j = 2) or (number of integers in the largest number) so that values come out to be as 0.98, 0.97 and so on.

$$v' = \frac{v}{10^j} \qquad \text{Where } j \text{ is the smallest integer such that } \text{Max}(|v'|)<1$$

Feature Scaling

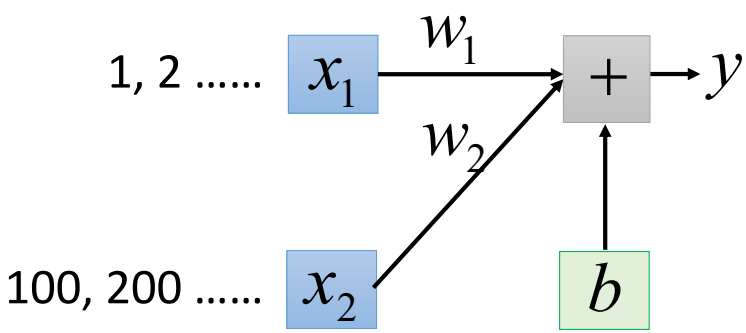
$$y = b + w_1 x_1 + w_2 x_2$$

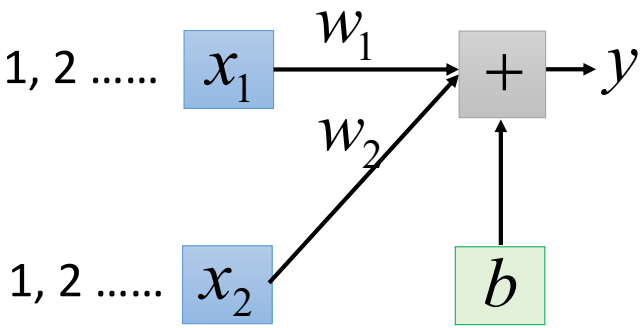


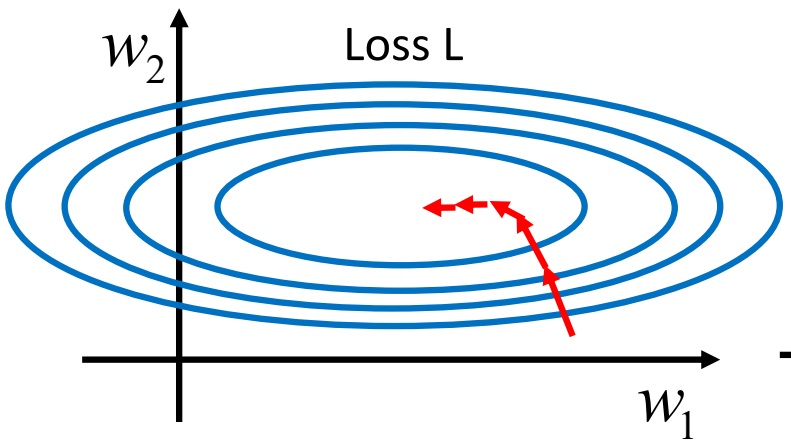
Make different features have the same scaling

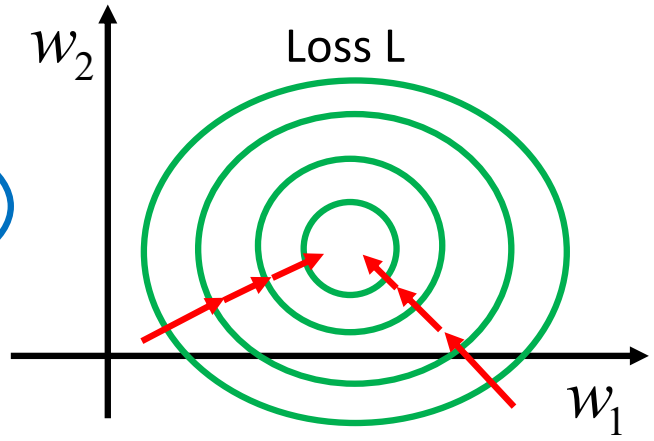
Feature Scaling

$$y = b + w_1 x_1 + w_2 x_2$$









DATA DISCRETIZATION

- Given probabilitites p_1, p_2, .., p_s whose sum is 1, ***Entropy*** is defined as:

$$H(p_1, p_2, \ldots, p_s) = \sum_{i=1}^{s} (p_i log(1/p_i))$$

- Entropy measures the amount of randomness or surprise or uncertainty.
- Only takes into account non-zero probabilities

Discretization

- For each continuous attribute,
  - Sort the attribute on values
  - Choose a split position midway between any two values
  - Linearly scan these values, each time updating the count matrix and computing gini index
  - Choose the split position that has the least gini index

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

Cheat	No		No		No		Yes		Yes		Yes		No		No		No		No	
Taxable Income																				
Sorted Values →	60		70		75		85		90		95		100		120		125		220	
Split Positions →	55		65		72		80		87		92		97		110		122		172		230	
	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
Yes	0	3	0	3	0	3	0	3	1	2	2	1	3	0	3	0	3	0	3	0	3	0
No	0	7	1	6	2	5	3	4	3	4	3	4	3	4	4	3	5	2	6	1	7	0
Gini	0.420		0.400		0.375		0.343		0.417		0.400		*0.300*		0.343		0.375		0.400		0.420	