Introduction to ISOM5270

Instructor: Jing Wang

Department of ISOM

Spring 2023

HKUST
BUSINESS SCHOOL
香港科大商學院

About Me

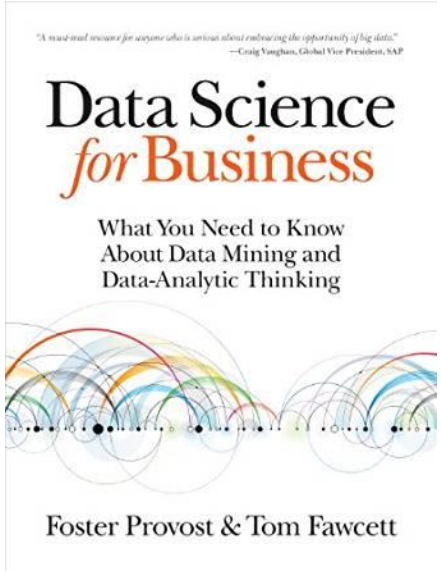
- Name: **Jing Wang**
- Position: Associate Professor, ISOM
- Education: PhD from NYU Stern Business School (Major: IS) and Bachelor from Tsinghua University (Major: CS)
- Work Experience: Yahoo! Research and HP Labs
- Research Interests: Crowdsourcing, Data Mining, Social Networks, Online Labor Markets, Crowdfunding, etc.

Course Materials

- All the materials (e.g., lecture slides, readings, guidance) will be posted on Canvas course website.

- Optional textbook:
  - ***Data Science for Business: What you need to know about data mining and data-analytic thinking***, O'Reilly Media, 2013

    by **Foster Provost** and Tom Fawcett
  - HKUST Library has online access



Course Structure

- Learn through…
  - Mainly lectures with some in-class discussion and exercises.
  - Hands-on problem-solving using Python in lab session (very important for learning).

Software

The most popular Python distribution

A web-based interactive computing platform that combines live code, equations, narrative text, visualizations, etc.

A free Jupyter notebook environment that runs entirely in the cloud

Visual Studio Code

A lightweight but powerful source code editor for building and debugging modern web and cloud applications.

Grading Components

- Class Participation: 10%
- Homework Assignments (2): 20%
- Group Project: 30%
- Final Exam: 40%

Project Guideline

- Form groups of 4-5 students
- General goal: In this project, you will apply the data mining techniques you learned in the class to solve a real-world business problem.
  - Feb 11: Group formation
  - Feb 25: Submit project idea for approval
  - Mar 29: Submit project report

Important Dates

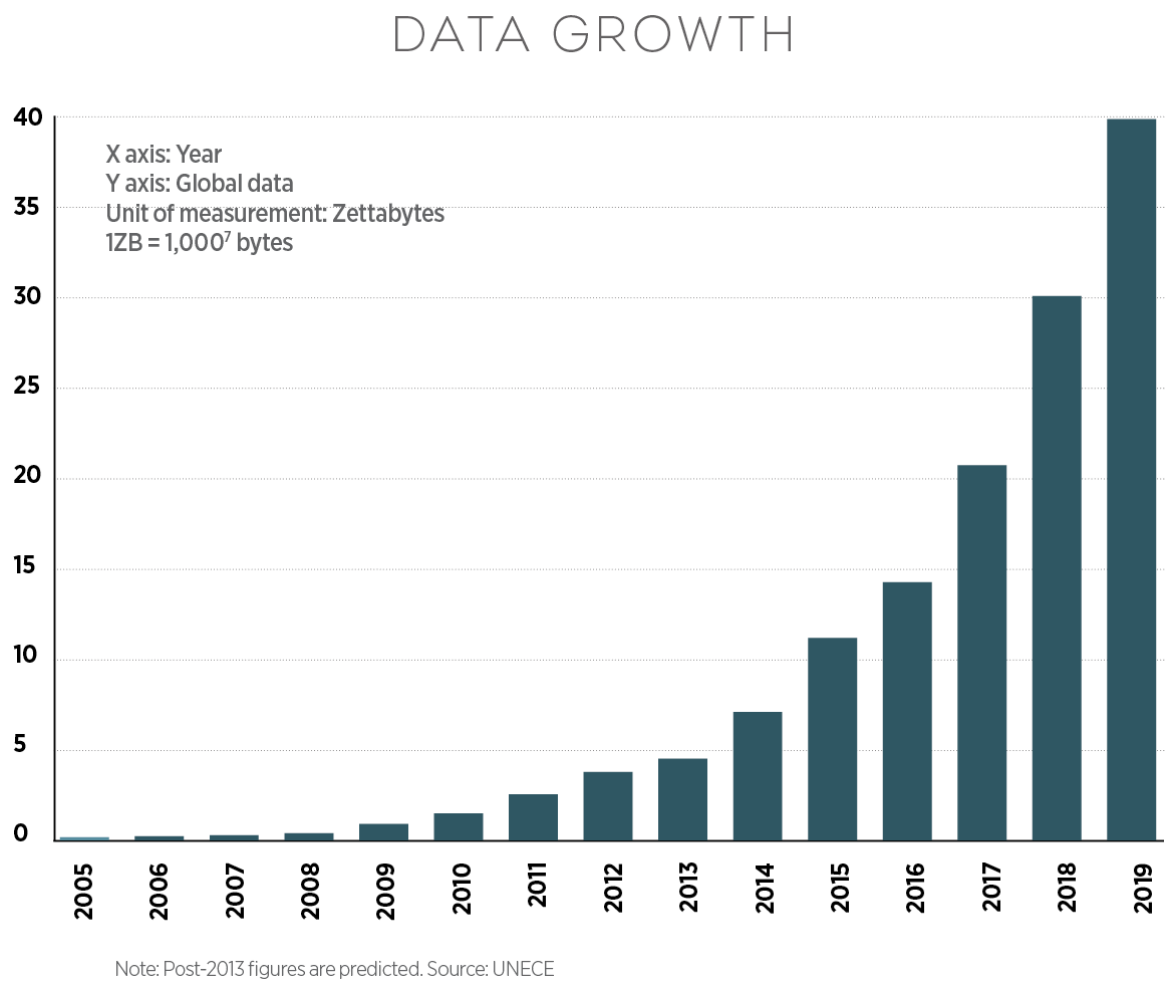
- **Feb 11:** Group formation due
- **Feb 25:** Project idea due
- **Feb 25**: Assignment 1 due
- **Mar 18:** Assignment 2 due
- **Mar 25**: Final exam
- **Mar 29**: Project report due

What If I Have Questions?

- Instructor: Prof. Jing WANG
  - **Email:** jwang@ust.hk Begin subject: [ISOM5270] …
  - **Phone:** 3469-2125
  - **Office Hours:** By appointment
  - **Office Location:** LSK 4044
- Teaching Assistant: Sophie GU
  - **Email:** imsophie@ust.hk
  - **Phone:** 2358-7653
  - **Office Hours:** By appointment
  - **Office Location:** LSK 4065

Data Analytics

- Data is generated at an extraordinary rate.

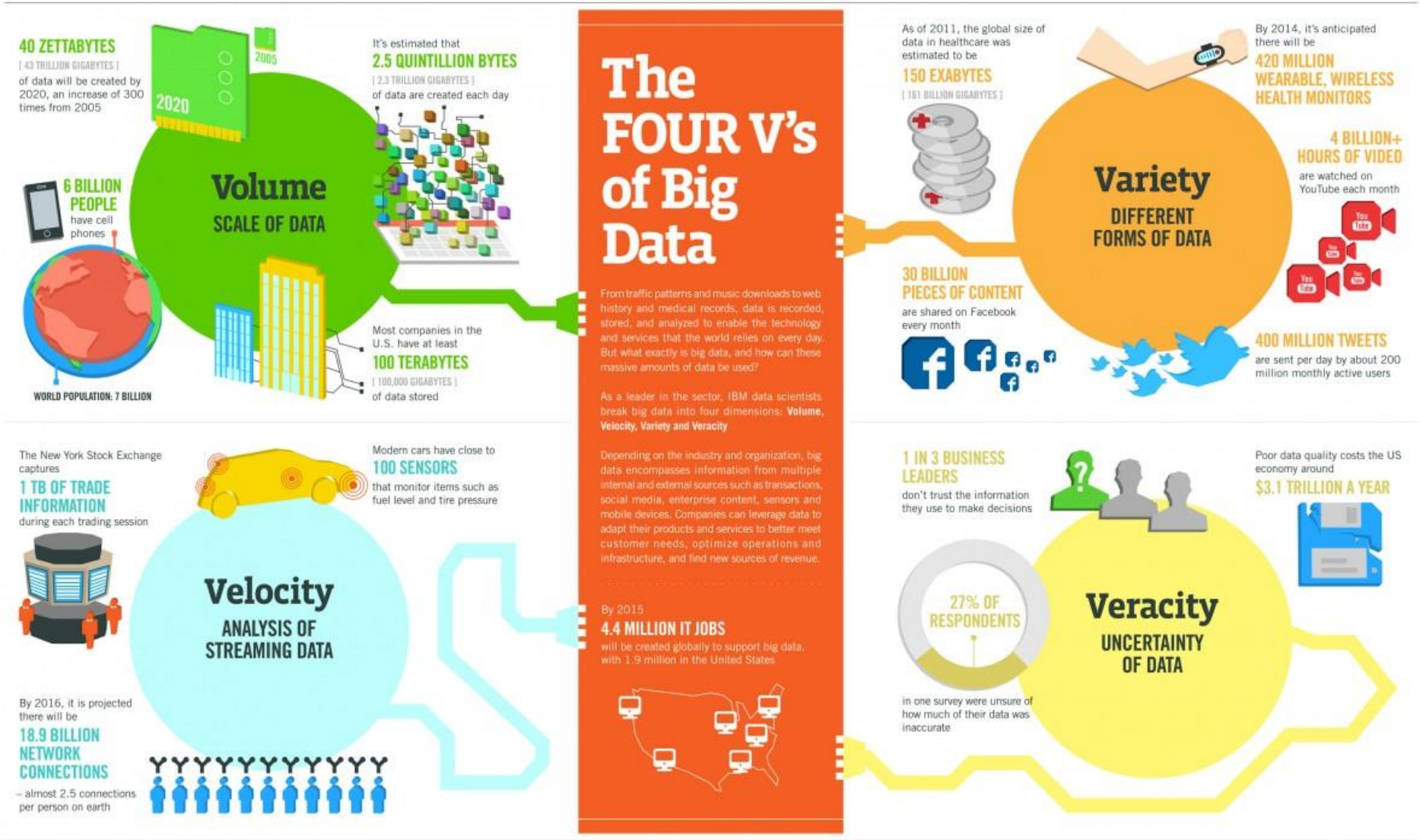


Note: Post-2013 figures are predicted. Source: UNECE

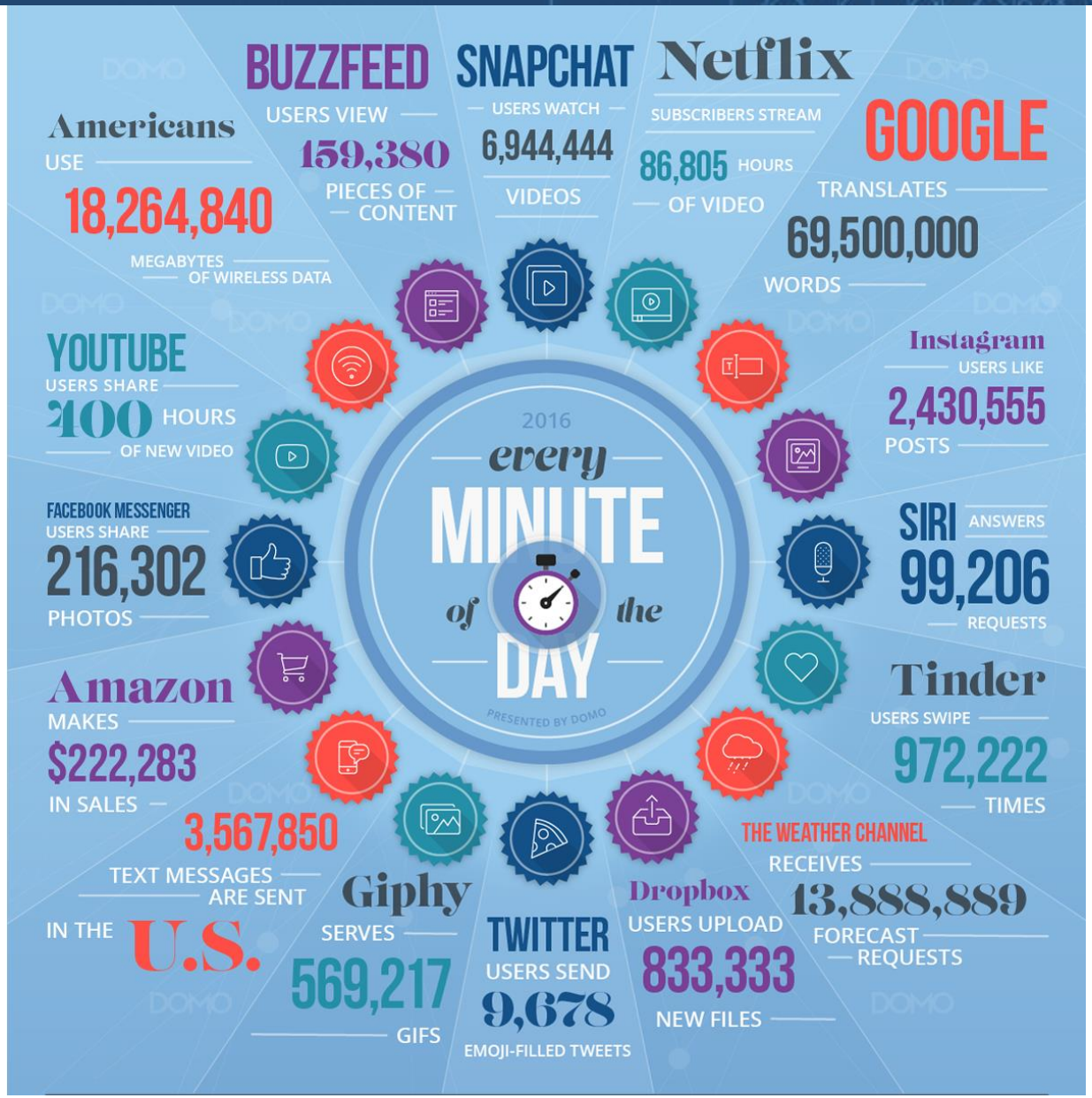
What is Big Data?

Eg. Stock market
Tesla - self driving

The Four V's of Big Data

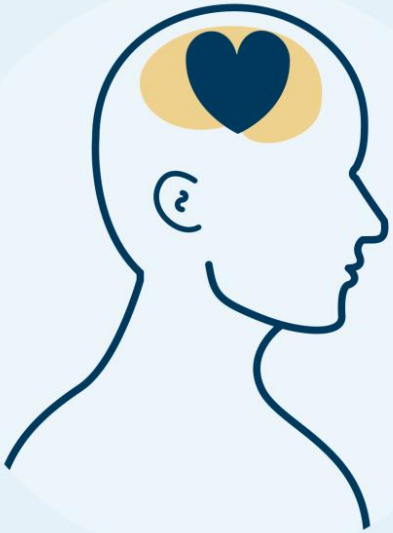


Data Never Sleeps



Data Driven vs. Human Intuition

Intuition Driven Decision Making



Relies on Gut Feeling

Make best guesses

Relies heavily on previous experience

Inherently risky

Corrective

Data Driven Decision Making

Relies on Facts

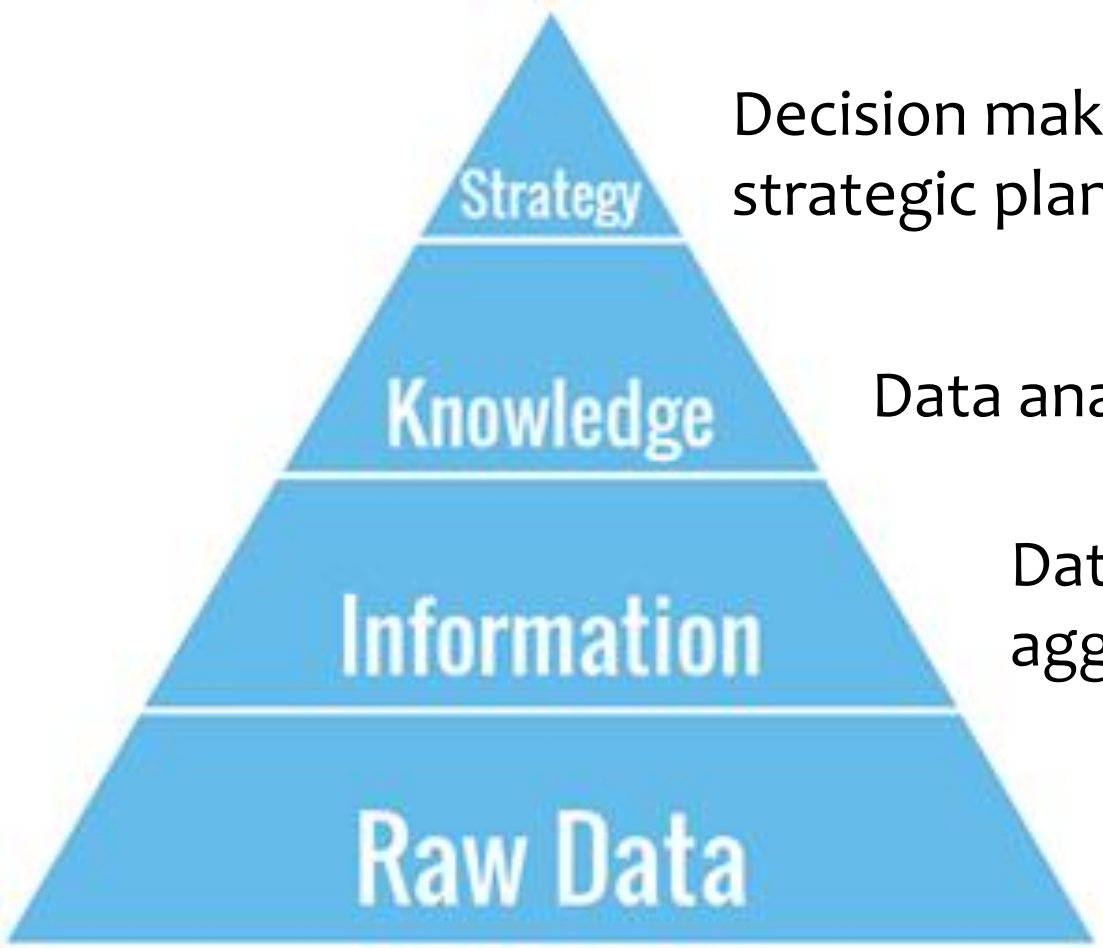
Choices are tested

Inspired by previous experience

Risk aversing

Directive

Business Intelligence Pyramid



Decision making and strategic planning

Data analysis

Data retrieval and aggregation

Data management and storage

An Example: Customer Retention

- Which customers should they target with a special offer, prior to contract expiration?

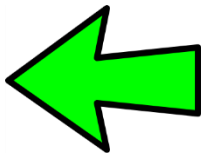
Traditional Solutions

- Offer incentives to every customer before contract expires
- Contact each customer to probe propensity to terminate contract
- Send offers to customers with high contract value

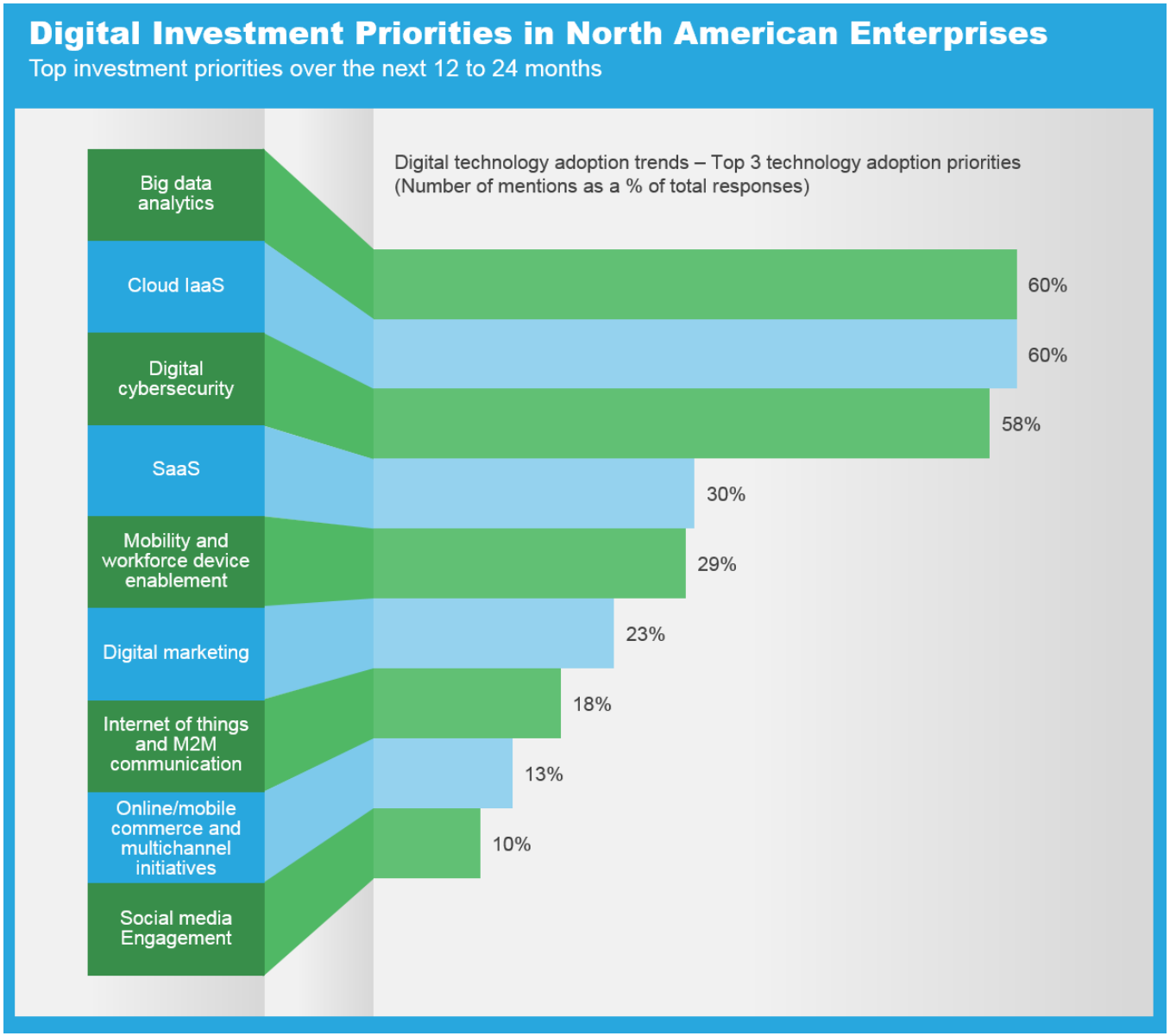
Are these traditional solutions cost effective? Why?

Ideal - Targeting customer with high contract value, and high possibility to leave

Big Data Solution



IT Investment Priorities



Application of Big Data Analytics

Retail

- CRM – Customer Scoring
- Store Siting and Layout
- Fraud Detection / Prevention
- Supply Chain Optimization

Advertising & Public Relations

- Demand Signaling
- Ad Targeting
- Sentiment Analysis
- Customer Acquisition

Financial Services

- Algorithmic Trading
- Risk Analysis
- Fraud Detection
- Portfolio Analysis

Media & Telecommunications

- Network Optimization
- Customer Scoring
- Churn Prevention
- Fraud Prevention

Manufacturing

- Product Research
- Engineering Analytics
- Process & Quality Analysis
- Distribution Optimization

Energy

- Smart Grid
- Exploration

Government

- Market Governance
- Counter-Terrorism
- Econometrics
- Health Informatics

Healthcare & Life Sciences

- Pharmaco-Genomics
- Bio-Informatics
- Pharmaceutical Research
- Clinical Outcomes Research

Why Big Data for You?

Data Scientist:

The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data.

by Thomas H. Davenport and D.J. Patil

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

Skills Possessed by Modern Data Scientists

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing package e.g. R
- Databases SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

Course Objectives

- You will learn
  - Various data mining models, i.e., how to find patterns
  - Hands-on experience with Python (sample codes and instructions)
  - Analytical thinking in various business examples
- You will not learn
  - Data management, big data technologies
  - Business/managerial planning
  - Python programming in a comprehensive manner

Please note: this is essentially a data mining course. The emphasis is on the mastery of the concepts and techniques rather than programming.

Planned Topics

- Data mining basics
- Data understanding and preparation
- Decision tree learning
- Overfitting and model evaluation
- Cost-sensitive learning
- Linear and logistic regression
- Naïve Bayes classifier
- Text Mining
- Feature selection
- Model selection
- Association rule learning
- Clustering methods
- K-nearest neighbors
- Recommender system using collaborative filtering
- Support vector machines
- Ensemble learning
- Neural networks and deep learning

Background Survey

Data Mining Basics

Instructor: Jing Wang

Department of ISOM

Spring 2023

HKUST
BUSINESS SCHOOL
香港科大商學院

What is Data Mining?

- **Data mining** (**knowledge discovery** from data)
  - Automatic extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from large amount of data
  - Involves methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.

Non-Trivial Data Mining Results

- Beers and diapers were often bought together by customers.

Young couple might drink beers to relief pressure

- Phoenix is not a good place for selling golf clubs, despite the many golf courses nearby.

more careful and organize

- People who buy small pad that adhere to the bottom of chair legs (to protect the floor) are more likely to be good credit risk.

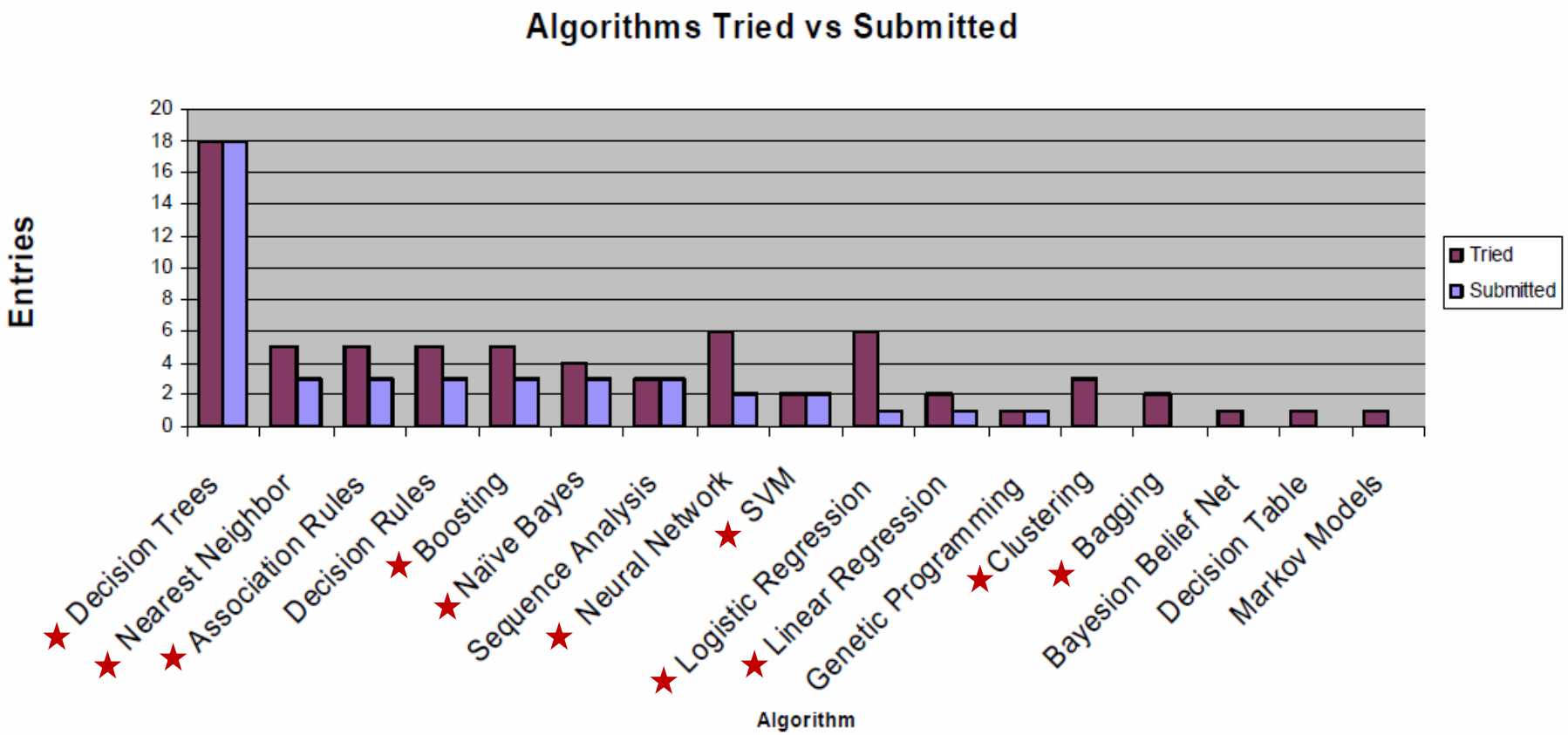
- Vegetarians tend to miss fewer flights.

Vegetarians need to preorder meal, more likely to show up

Common Data Mining Tasks

- Classification and class probability estimation
  - Determine which class an individual belongs to
- Regression
  - Estimate the numerical value of some variable for an individual
- Similarity matching
  - Identify similar individuals based on known attributes
- Clustering
  - Group individuals together by their similarity
- Co-occurrence grouping (frequent itemset mining)
  - Find associations between entities based on transactions involving them.

Commonly Used Induction Algorithms



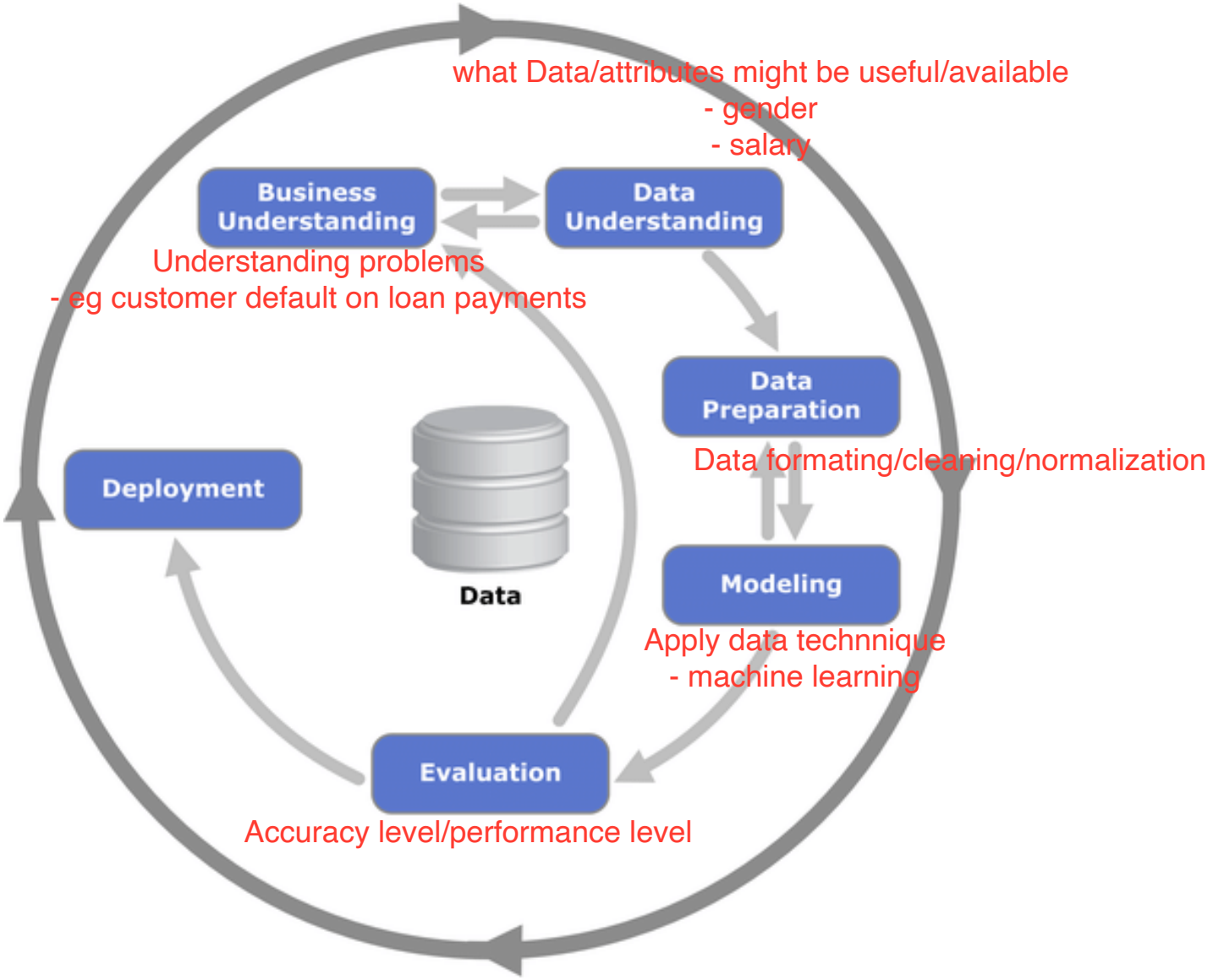
Revisit: Customer Retention

- Which customers should they target with a special offer, prior to contract expiration?

Example Data Mining Solutions

- Decision tree technique
  - If Education = ‘high’ and Gender = ‘male’, then customer is likely to churn.
    客户流失
- Logistic regression technique
  - Calculate the probability of churning given the features of a customer.
- Nearest neighbor technique
  - Calculate how similar a customer is to existing churning customers.

A Process View to Data Mining



Data Mining Basic Terminologies

- Data, target variable, model
- Supervised vs. unsupervised learning
- Classification vs. regression
- Training vs. testing
- Mining phase vs. using phase

Data

- Example (Instance)
  - A fact or a data point; described by a set of attributes (fields, columns, variables, features).
- A data set:
  - A set of examples
  - A sample/subset of the universe

Variables

Name	Balance	Age	Default
Mike	123,000	50	No
Mary	51,100	40	Yes
Bill	68,000	55	No
Jim	74,000	46	Yes
Dave	23,000	44	No
Anne	100,000	50	Yes

One example/instance

Exercise

- Can you name a few attributes of the following?
  - A stock code, mark cap, company, industry etc
  - An apartment size, range, location

Target Variable

- A special variable that is the interest/target of the task.

- Equivalent statistics terminology:
  - Attributes: variables
  - Target variable: predict/estimate result
    dependent variable
    outcome variable

Target variable

Name	Balance	Age	Default
Mike	123,000	50	No
Mary	51,100	40	Yes
Bill	68,000	55	No
Jim	74,000	46	Yes
Dave	23,000	44	No
Anne	100,000	50	Yes

Types of Attributes/Variables (I)

- Numerical variable (quantitative data)
  - Discrete variable: has only a finite or countably infinite set of values (often integer variables)
    - Example: the number of items bought by a customer (e.g., 12)
  - Continuous variable: has real numbers as attribute values
    - Example: the time that the customer spends (e.g., 16.49 min)

https://www.youtube.com/watch?v=6IdJ1aPFDCs

Types of Attributes/Variables (II)

- Categorical variable (qualitative data)
  - Ordinal variable: has categories that can be meaningfully ordered
    - Example: course grade (A, B, C, D, …)
  - Nominal variable: the categories have no meaningful orderings
    - Example: location region (Sai Kung, Sha Tin, Wan Chai, etc.)

In-Class Exercise

- The size of a company (#employees) is a Discrete variable.
- The average height of students taking this class is a Continuous variable.
- The country that an individual lives in is a Nominal variable.
- Education level (less than high school, high school, bachelor, master, doctoral) is a Ordinal variable.

Unstructured Data

- Merrill Lynch cited a rule of thumb that somewhere around 80-90% of all potentially usable business information may originate in unstructured form (text, image, social networks, etc).
  - Product and hotel reviews
  - Blogs, forums and other social media
  - Voice of the customer data
  - Machine logs
  - Web logs
  - …

Unstructured data: Text (0/1 Representation)

--- Each entry in the table represents a document.

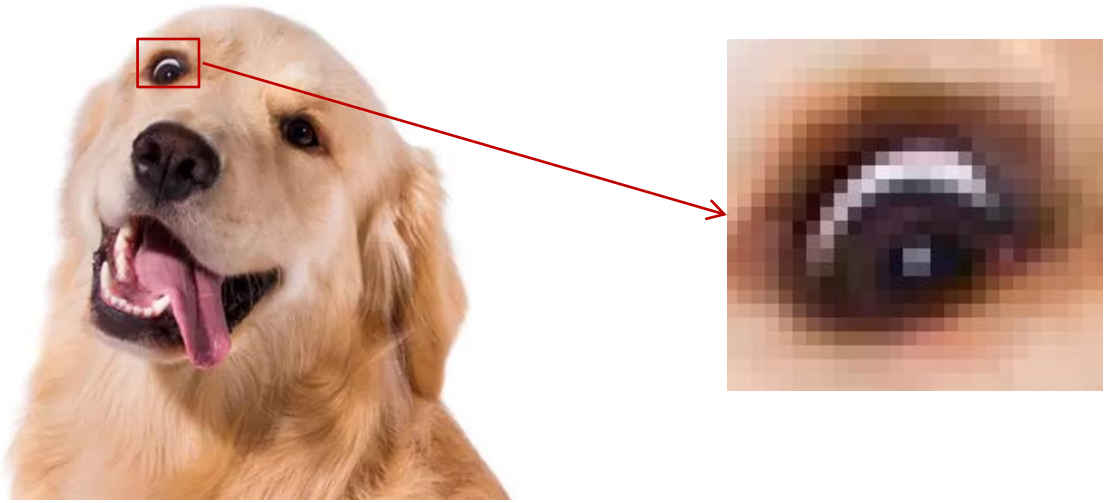
--- Attribute describes **whether or not** a word appears in the document.

On November 15, 2011, the Board of Directors (the "Board") of Apple Inc. (the "Company") appointed Robert A. Iger to the Board. Mr. Iger will serve on the Audit and Finance Committee of the Board.

	word				
	Apple	board	appoint	loss	…
doc 1	1	1	1	0	
doc 2	0	0	0	1	
…	…	…	…	…	…

…On a GAAP basis, the Company reported a net loss of $356 million … We remain intensely focused on helping our financial institution and retail clients through this difficult period …

Unstructured data: Image



- What are the features of an image?
- One tongue, two eyes, one nose?
- Computer represents an image using RGB pixels. An 640*480 image consists of 307,200 pixels. Each pixel is a RGB tuple value between 0~255. e.g. (255,0,0) is red.

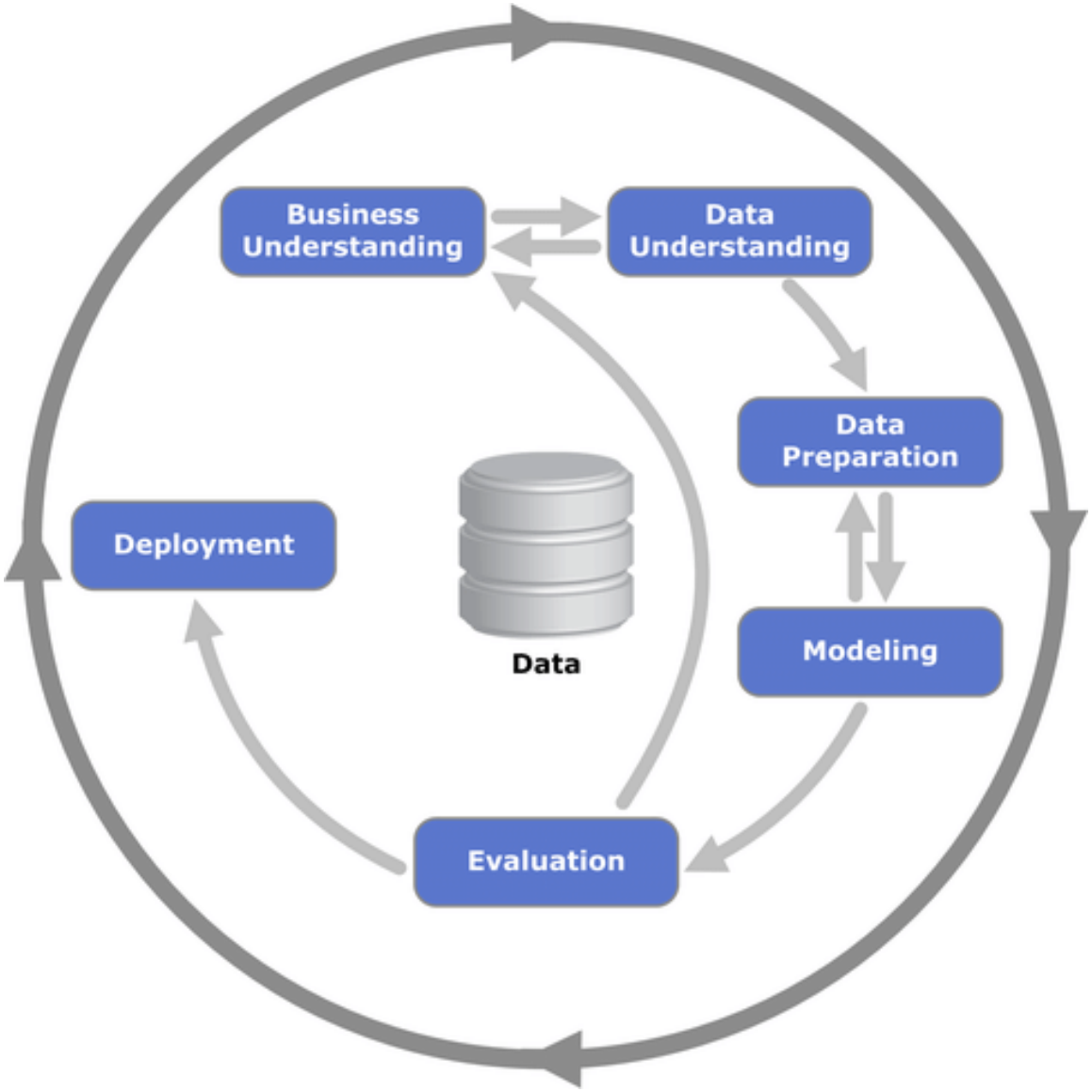
red, geen, blue

Unstructured data: Network

connection

- Your network properties, such as your neighbors, your neighbors' neighbor, your "centrality"

Data Mining Process Revisit



Model

- A model is:
  - A pattern.
  - A summarization of relationships in the data.
  - A simplified representation of reality created to serve specific purpose.

- Some examples
  - IF Balance >= 50K AND Age > 45

    THEN Default = ‘no’

    ELSE Default = ‘yes’

Learner

- A learner or inducer or algorithm
  - A method or algorithm used to generalize a model from a set of examples.

Name	Balance	Age	Default
Mike	123,000	50	No
Mary	51,100	40	Yes
Bill	68,000	55	No
Jim	74,000	46	Yes
Dave	23,000	44	No
Anne	100,000	50	Yes

Learner: induces a pattern from examples

IF Balance >= 50K AND Age > 45
THEN Default = 'no'
ELSE Default = 'yes'

In practice, people use model and learner interchangeably. But they are different.

Supervised vs. Unsupervised Learning

- **Supervised learning (prediction)**: learns a model that predicts target outcome based on a set of other attributes (i.e., training data where target value is known).
  - Stock price prediction (numerical target variable)
  - Credit card default (binary target variable)
- **Unsupervised learning (relationship mining)**: finds relationships in the data without reference to target variable.
  - Beer and diaper

Key: is there a target that we are trying to predict?

No

Classification vs. Regression

- The difference is the type of target variable:
  - Classification: categorical target variable
    - Is this customer "loyal" or "likely to terminate contract"?
    - Is a credit card use "legitimate" or "fraudulent"?
  - Regression: numerical target
    - How much a customer is going to spend?
    - What is the credit score of a customer?

classify group/categories

Targeting variable is a number

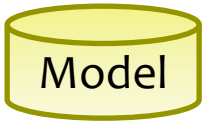
- Both are supervised learning!

Predictive DM/Modeling: the Philosophy

Data you already have

Name	Balance	Age	Default
Mike	123,000	50	No
Mary	51,100	40	Yes
Bill	68,000	55	No
Jim	74,000	46	Yes
Dave	23,000	44	No
Anne	100,000	50	Yes

Build/Evaluate model



Apply to new data

Data you will have

…	…	…	…
Henry	61,100	35	???
Amy	68,000	52	???
Allen	22,000	21	???
Tom	123,000	60	???
Jane	100,000	47	???

Model Evaluation

- Supervised Learning
  - Ground truth: Yes 知道结果后可以evaluate model sucessful
  - Evaluation: predictive performance

- Unsupervised Learning
  - Ground truth: No
  - Evaluation: intelligibility

没办法知道final result，因为没有target，no absolute true

Model Training vs. Model Testing

- Problem:
  - After learning a model, can we have an estimate on how well the model would perform on new data?
- Solution: split data into two parts
  - Training data to learn the model.
  - Testing data to evaluate performance of learned model on "new" data.
  - Never ever use testing data to learn your model!

don't use training data to evaluate your model

Why do we want to split data into two parts?

Training data use to evaluate = 100% accuracy overfitting

Data Splitting for Training and Testing

Name	Balance	Age	Default
Mike	123,000	50	No
Mary	51,100	40	Yes
Bill	68,000	55	No
Jim	74,000	46	Yes
Dave	23,000	44	No
Anne	100,000	50	Yes

Training data

Name	Balance	Age	Default
Mike	123,000	50	No
Mary	51,100	40	Yes
Bill	68,000	55	No
Jim	74,000	46	Yes

Testing data (Hold-out data)

Name	Balance	Age	Default
Dave	23,000	44	No
Anne	100,000	50	Yes

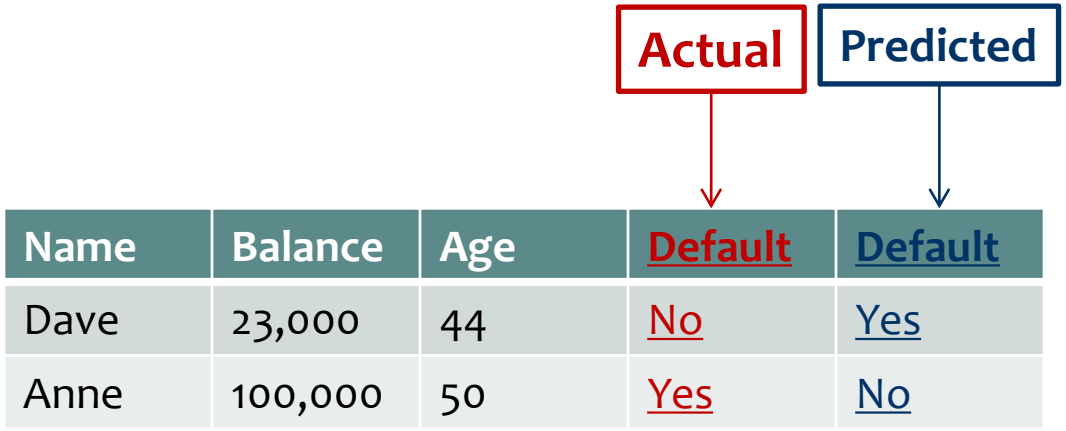
Model Training (on Training Data)

Name	Balance	Age	Default
Mike	123,000	50	No
Mary	51,100	40	Yes
Bill	68,000	55	No
Jim	74,000	46	Yes

Learner: induces a pattern from examples

IF Balance >= 50K AND Age > 45
THEN Default = 'no'
ELSE Default = 'yes'

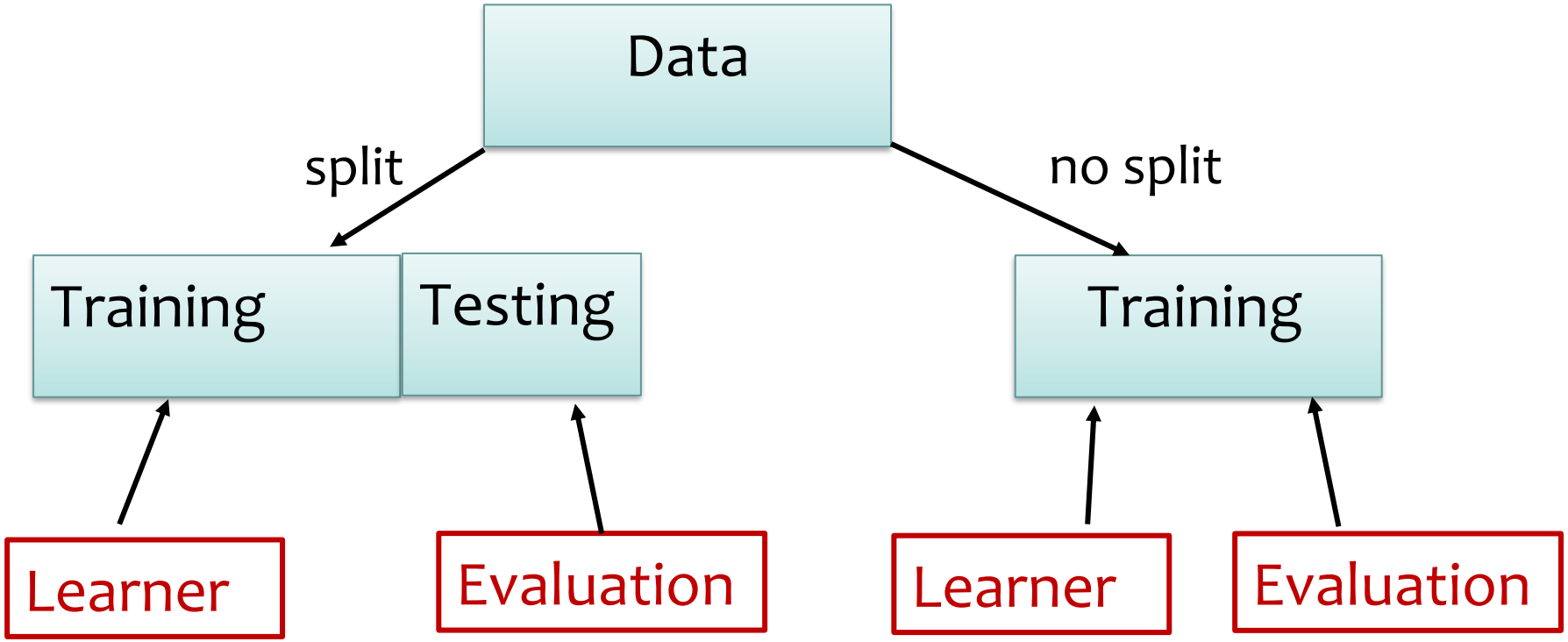
Model Testing (on Testing/Hold-Out Data)



Name	Balance	Age	Default	Default
Dave	23,000	44	No	Yes
Anne	100,000	50	Yes	No

IF Balance >= 50K AND Age > 45
THEN Default = 'no'
ELSE Default = 'yes'

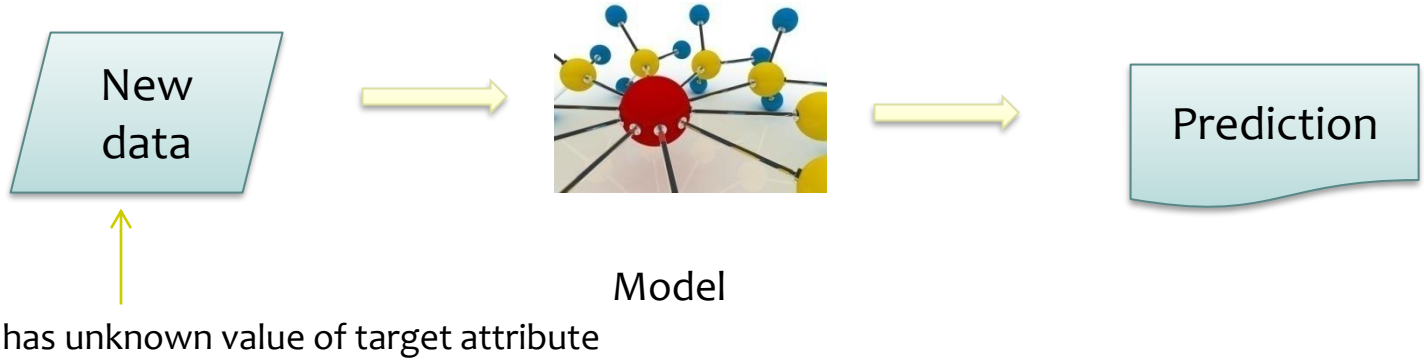
Process for Supervised/Unsupervised Learning



Supervised Learning

Unsupervised Learning

Data Mining Phase vs. Use Phase (Supervised)



In-Class Exercise

- TelCo, a major telecommunications firm, wants to investigate its problem with customer attrition, or "churn"
- This is a saturated market; a large proportion of cell-phone customers leave when their contracts expire.

Supervised learning

Q: Which customers should they target with a special offer, prior to contract expiration?

- Try to come up with a data-driven solution to the problem.
- Use the concepts you learned today.
- Lay out a step-by-step plan (high level).

Step-by-Step Plan

Step 1: Target variable (terminate or not)
Step 2: Explanatory variables (demographic attributes - gender, age, income, region; Relationship/activities)
Step 3: tables -> dataset with target variable and explanatory
Step 4: split into training - build the model; testing - evaluate the model -> find best model
Step 5: predit for those customers, yes - terminate; no - not terminate -> rank confidency
Step 6: Send offers top 10% who are likely to churn

Data Understanding and Preparation

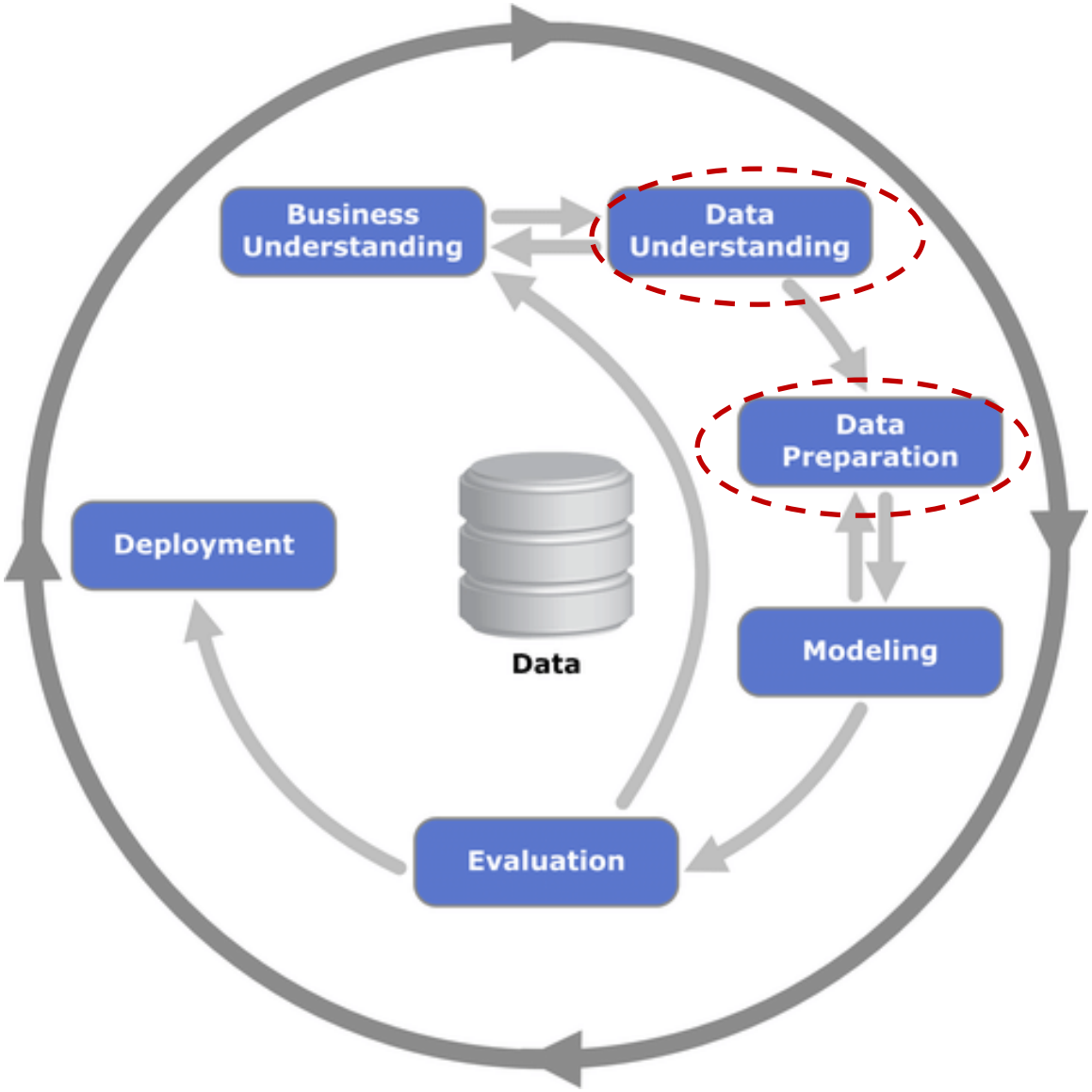
Instructor: Jing Wang

Department of ISOM

Spring 2023

HKUST
BUSINESS SCHOOL
香港科大商學院

A Process View to Data Mining



Where are the data from?

- Company sales database
- Company customer database
- Survey data
- Government public data
- Third-party data provider
- Social media data
- …

https://data.gov.hk/en/

Example: Zillow

- Zillow is the most popular online real estate information site in the US.

  https://www.zillow.com/

- Zestimates: a tool for estimating home value.

To Describe the Dataset

- What do your records represent?
- What does each attribute mean?
- What type of attributes?
  - Categorical (e.g., nominal, ordinal)
  - Numerical (e.g., discrete, continuous)
  - Text

Boston Housing Dataset

https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html

1. CRIM - per capita crime rate by town
2. ZN - proportion of residential land zoned for lots over 25,000 sq.ft.
3. INDUS - proportion of non-retail business acres per town.
4. CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
5. NOX - nitric oxides concentration (parts per 10 million)
6. RM - average number of rooms per dwelling
7. AGE - proportion of owner-occupied units built prior to 1940
8. DIS - weighted distances to five Boston employment centres
9. RAD - index of accessibility to radial highways
10. TAX - full-value property-tax rate per $10,000
11. PTRATIO - pupil-teacher ratio by town
12. B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
13. LSTAT - % lower status of the population
14. MEDV - Median value of owner-occupied homes in $1000's

Data Understanding: Explore the Dataset

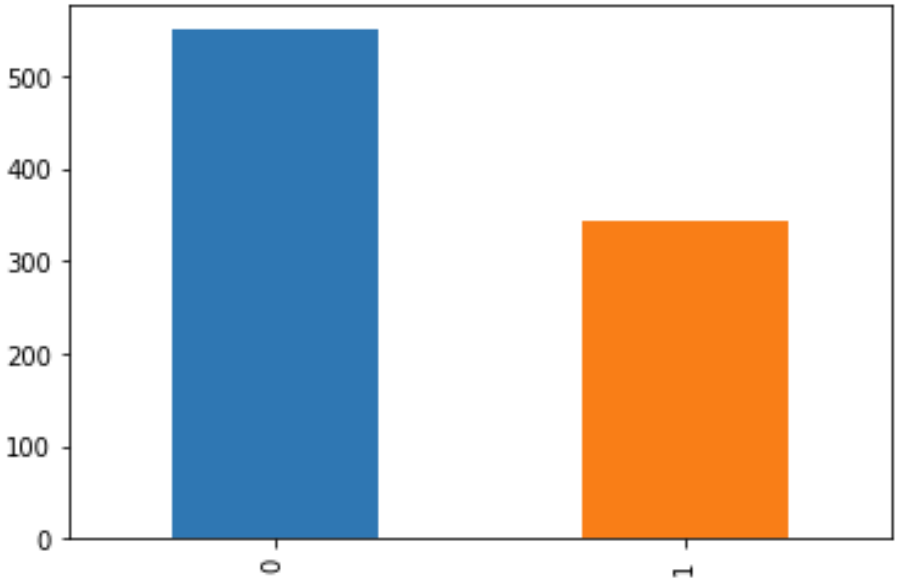
- Preliminary investigation of the data to better understand its specific characteristics; Help in selecting appropriate data mining algorithms
- Things to look at
  - Class imbalance
  - Dispersion of data attribute values
  - Skewness, outliers, missing values
  - Correlation analysis
- Visualization tools are important
  - Histograms, box plots
  - Scatter plots

Class Balance

- Many datasets have a discrete (binary) attribute class
  - What is the frequency of each class?
  - Is there a considerable less frequent class?
- Sometimes, classes have very unequal frequency
  - Medical diagnosis: 90% healthy, 10% disease
  - Online purchase: 99% don't buy, 1% buy
  - Fraud detection: 99.9% transactions are not fraudulent
- Data mining algorithms may give poor results due to class imbalance problem
  - Identify the problem in an initial phase

Bar Charts

- A bar chart presents categorical data with rectangular bars with heights proportional to the values that they represent.



Didn't survive Survived

Useful Statistics

- Discrete attributes
  - Frequency of each value (bar charts)
  - Mode = value with highest frequency
- Continuous attributes
  - Range of values, i.e., min and max
  - Mean (average)
  - Median
  - Skewed distribution

Five-Number Summary

- A set of descriptive statistics that provide information about a dataset: (min, Q1, Q2, Q3, max)
  - Minimum (min), lower quartile (Q1), median (Q2), upper quartile (Q3), maximum (max)

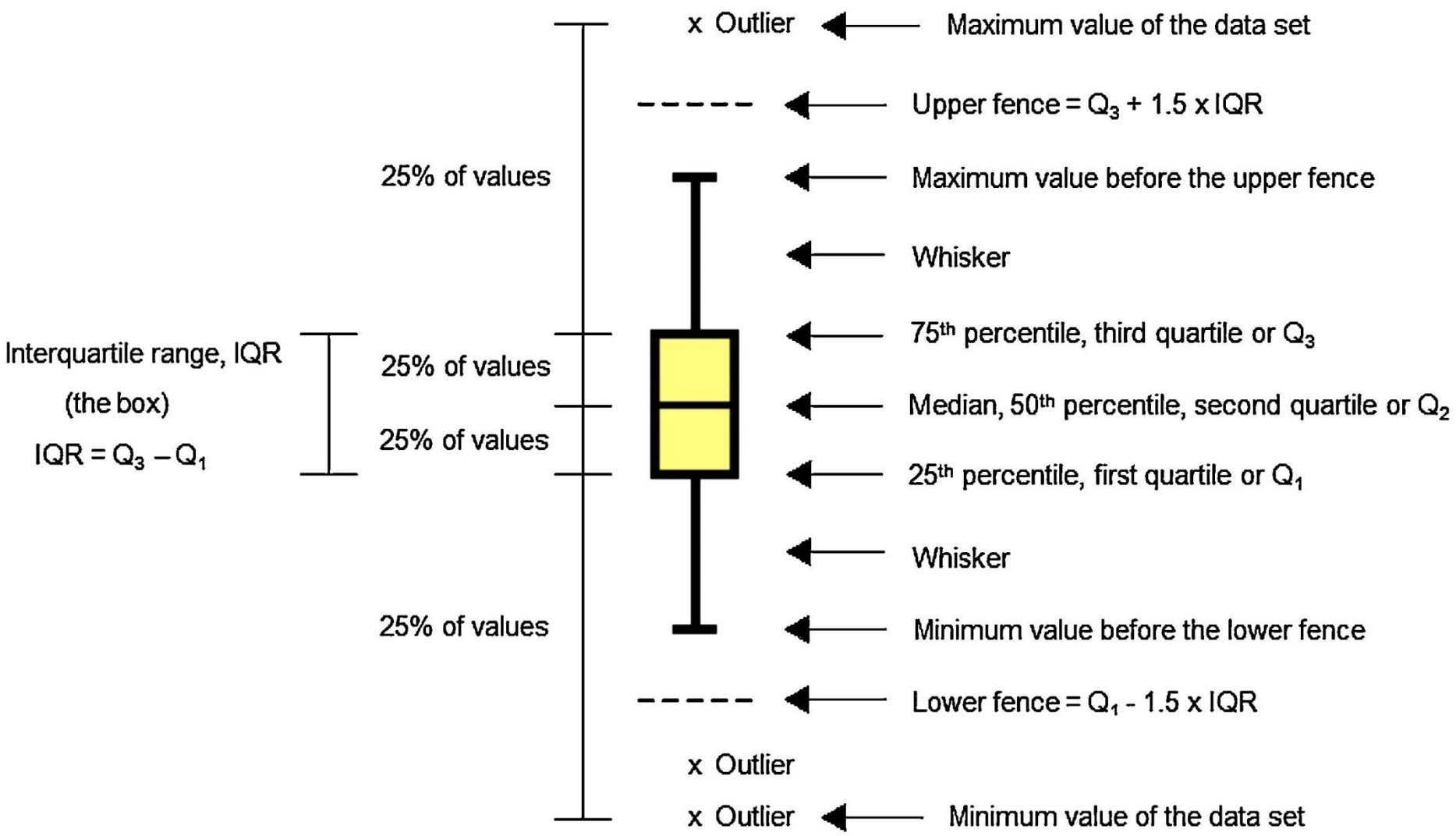
Attribute values: 6 47 49 15 42 41 7 39 43 40 36

Sorted: 6 7 15 36 39 40 41 42 43 47 49

What is the five-number summary of this dataset?

1. Sort the Dataset
2. Find min (6) and max (49)
3. Median 40 (Q2)
4. Q1 15
5. Q3 43

Box Plots

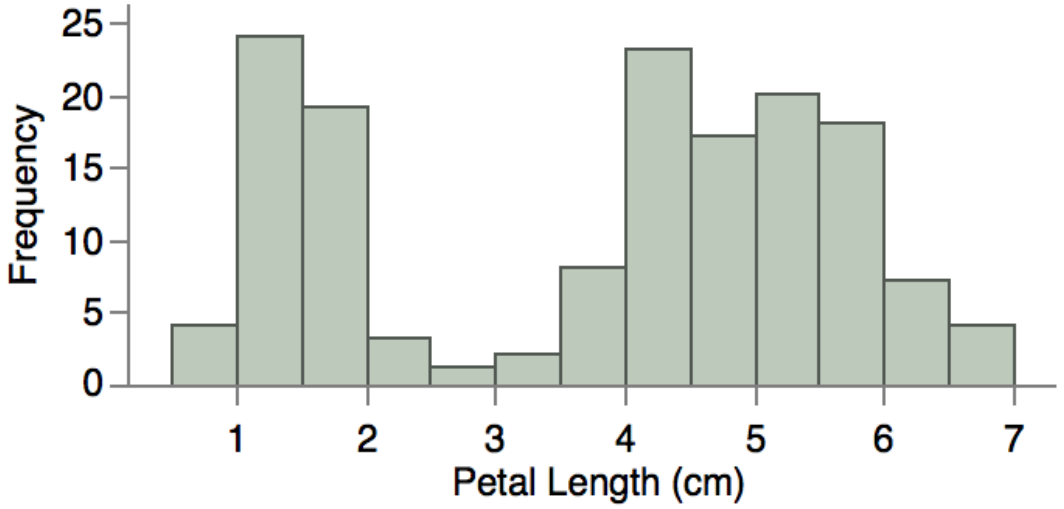

x Outlier
Maximum value of the data set
Upper fence = Q_3 + 1.5 x IQR
25% of values
Maximum value before the upper fence
Whisker
Interquartile range, IQR
(the box)
IQR = Q_3 – Q_1
25% of values
75th percentile, third quartile or Q_3
Median, 50th percentile, second quartile or Q_2
25% of values
25th percentile, first quartile or Q_1
Whisker
25% of values
Minimum value before the lower fence
Lower fence = Q_1 - 1.5 x IQR
x Outlier
x Outlier
Minimum value of the data set

Box Plots

- A box plot can provide useful information about an attribute
  - Sample's range
  - Median
  - Normality of the distribution
  - Skew (asymmetry) of the distribution
  - Plot extreme cases within the sample

Histograms

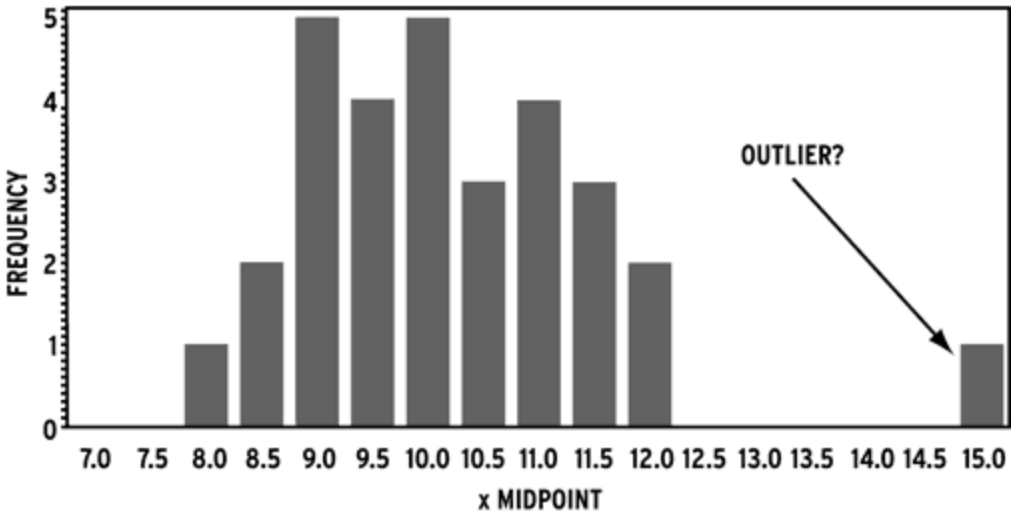
- A histogram is an estimate of the probability distribution of a continuous variable.
  - "Bin" the range of values (i.e., divide the entire range of values into a series of intervals) and count how many values fall into each interval
  - The bins (intervals) must be adjacent, and are often of equal size.



Outliers

- Outliers are values that lie far away from the bulk of data.
  - E.g., anything over 3 standard deviations away from the mean
  - Outliers can be legitimate instances or values

Some algorithms may produce poor results in the presence of outliers (need to identify and remove them)



Correlation Analysis (Numerical Data)

- Correlation is a measure that captures the statistical relationship between two variables.

Pearson's correlation coefficient

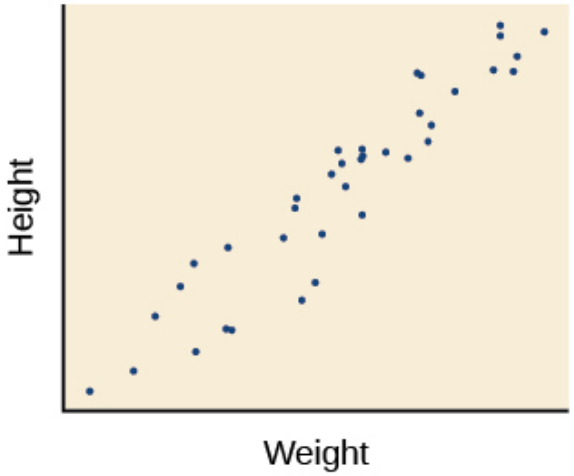
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where:

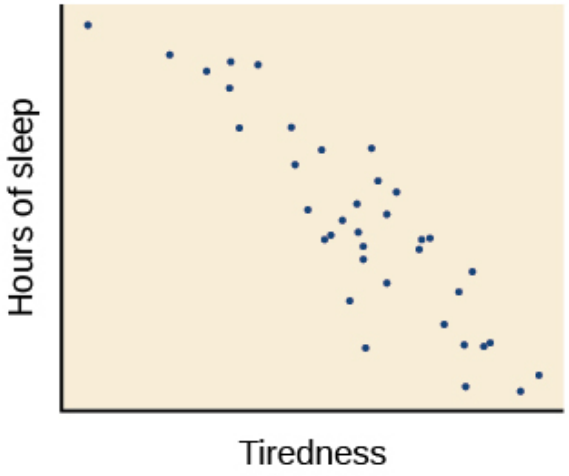
- n is sample size
- x_i, y_i are the individual sample points indexed with *i*
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$

Scatter Plot

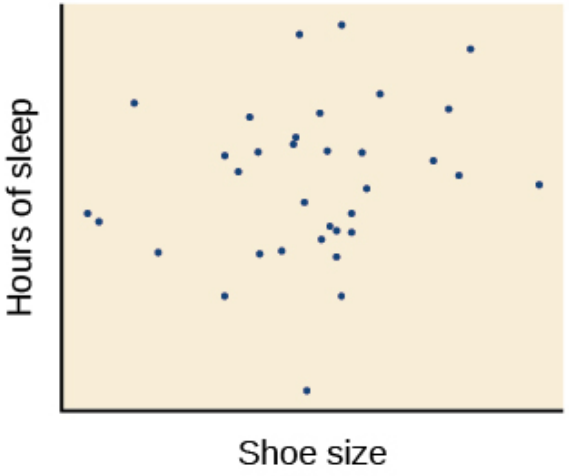
- A scatter plot is a two-dimensional data visualization that shows the correlation between two variables.



Positive Correlation



Negative Correlation



Little/no Correlation

Which one shows positive correlation, negative correlation, and no correlation?

Data Preprocessing

- Data is often collected for unspecified applications
  - Data may have quality problems that need to be addressed before applying a data mining technique
    - Noise and outliers
    - Missing values
  - Preprocessing may be needed to make data more suitable for data mining.

"If you want to find gold dust, move the rocks out of the way first!"

Data Preprocessing

- Data transformation might be needed
  - Handling missing values
  - Handling categorical variables
  - Feature transformation
    - e.g., log transformation
    - Normalization (back to this when clustering is discussed)
  - Feature discretization
  - Feature selection

Missing Values

- Data is not always available
- Missing data may be due to various reasons
  - Data not entered due to misunderstanding
  - Certain data may not be considered important at the time of entry
  - Deleted accidentally
- Missing data may carry some information content
  - A credit application may carry information by noting which field the applicant did not complete

How to Handle Missing Values: Remove

- Remove data instances that have missing value (may affect a lot records)
  - Okay if not more than 5% of the records
- Remove attributes with missing values (may leave out important features)

How to Handle Missing Values: Infer

- Use a global constant to fill in the missing value
  - e.g., “unknown”. (May create a new class!)
- Use the average (or most frequent) value to fill in the missing value
- Use the attribute mean (or most frequent value) for all samples belonging to the same class to fill in the missing value
- Other more sophisticated (data mining) methods
  - e.g., finding the k neighbors nearest to the point and fill in the average (or most frequent) value

Missing Values

- No matter what techniques you use to conquer the problem, it comes at a price. **The more guessing** you have to do, **the further away from the real data** the database becomes. Thus, in turn, it can affect the accuracy and validation of the mining results.

Handling Categorical Variables

- Categorical variable
  - Size: small, medium, large
  - Industry: Finance, IT, Marketing, etc...
- Some data mining algorithms can support categorical values without further manipulation but there are many more algorithms that do not.

Handling Categorical Variables

- Automobile dataset [link]
- Some variables in the dataset are categorical

	symboling	normalized_losses	make	fuel_type	aspiration	num_doors	body_style	drive_wh
0	3	NaN	alfa-romero	gas	std	two	convertible	rwd
1	3	NaN	alfa-romero	gas	std	two	convertible	rwd
2	1	NaN	alfa-romero	gas	std	two	hatchback	rwd
3	2	164.0	audi	gas	std	four	sedan	fwd
4	2	164.0	audi	gas	std	four	sedan	4wd

One-Hot-Encoding

- Convert each category value into a new (dummy) column and assigns a 1 or 0 value to the column.
  - e.g., drive_wheels: rwd, fwd, 4wd

drive_wheels
rwd
rwd
rwd
fwd
4wd

wheel_rwd	wheel_fwd	wheel_4wd
1	0	0
1	0	0
1	0	0
0	1	0
0	0	1

Any disadvantage of using this strategy?

If the categorical value have many category, you will need to create too many attributes

Ordinal Encoding

- Suitable for ordinal variables
  - Preserve ordinal relationships

Assumption: the distance between poor and good, good and very good, very good and excellent is the same.
However, the distance can not be accurately measure

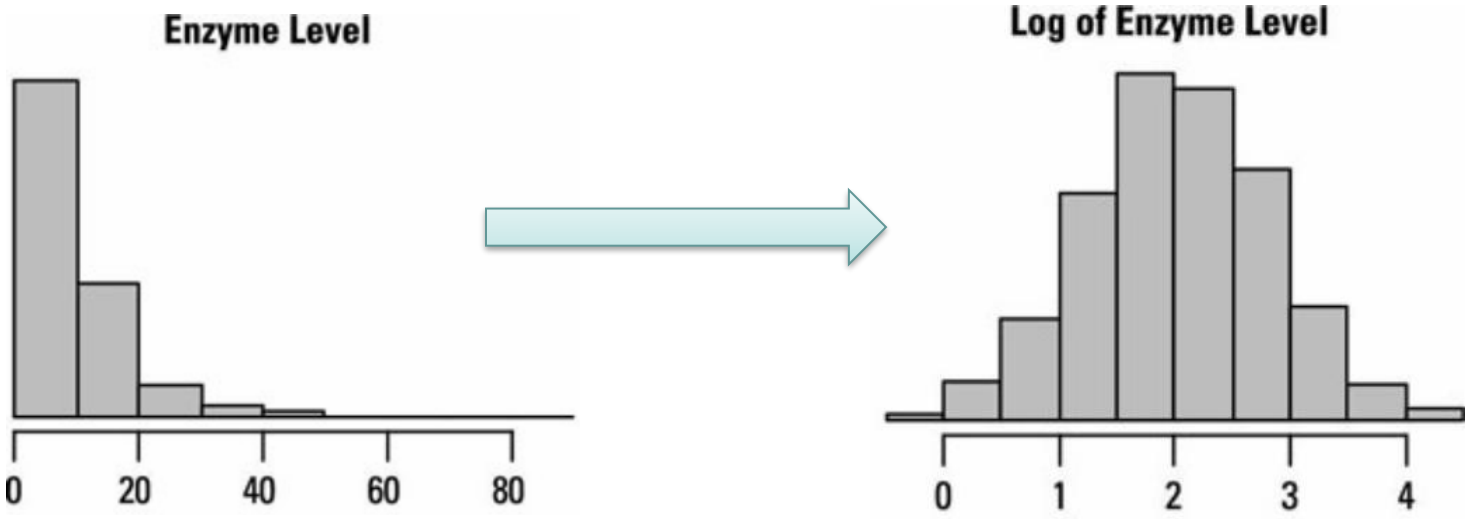
```
#Creating dictionary for mapping the ordinal numberical value
Cust_Rating_dict = {'Poor' : 1, 'Good': 2, 'Very Good': 3, 'Excellent': 4}
```

Customer_Rating
Poor
Good
Very Good
Excellent
Very Good
Excellent

Customer_Rating
1
2
3
4
3
4

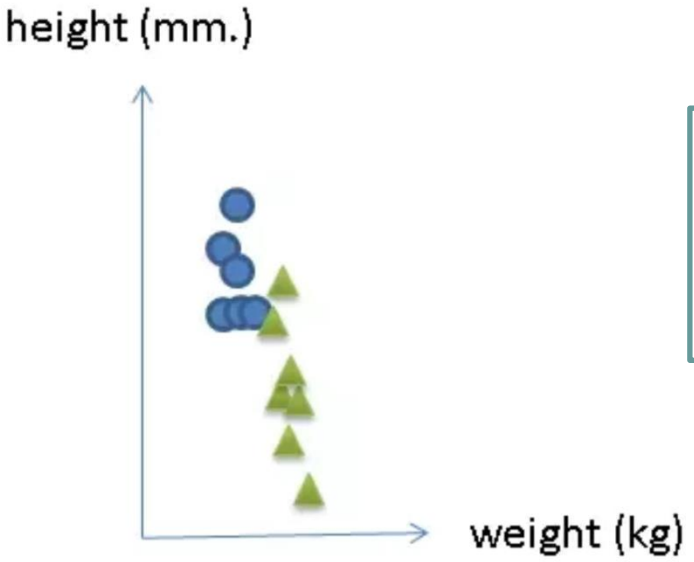
Feature Transformation: Taking Log

- A function is applied to each value of an attribute
  - E.g., Use $\log x$ to transform data that has a highly skewed distribution into data that are less skewed



Feature Normalization

- Normalization (standardization) helps to prevent that attributes with large ranges outweigh attributes with small ranges. It brings all variables to the same scale.



Difference in one coordinate (in this example, weight) is insignificant compared to a change in the other coordinate (height).

Feature Normalization: min-max

- Rescaling (min-max normalization):

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

range x' = [0,1]

- Suppose that the minimum and maximum values for the income variable are $12,000 and $98,000, respectively. By min-max normalization, a value of $73,600 for income is transformed to?

x' = (73600-12000)/(98000-12000)=

Feature Normalization: z-score

- Standardization

$$x' = \frac{x - \bar{x}}{\sigma}$$

  - where $\bar{x}$ is the mean of feature values, and σ is the standard deviation of feature values.

x' = (73600-54000)/16000

- Suppose that the mean and standard deviation of the values for the income variable are \$54,000 and \$16,000, respectively. With z-score normalization, a value of \$73,600 for income is transformed to ?

Questions

Which of the following normalization method(s) may transform the original variable to a negative value: min-max, z-score, or both, or neither?

z-score

min-max: ranging from 0-1

Do you apply normalization on training or testing set?

YES, if normalization applied on training set, then we need to normalize the testing set

Feature Discretization

- To transform a continuous attribute into a discrete attribute
  - Some data mining algorithms only work with discrete attributes (without extension)
    - e.g., decision trees, naïve Bayes
  - Better results may be obtained with discretized attributes

Feature Discretization: Binning

- Unsupervised discretization
  - Equal-interval binning
    - Divide attribute values into *N* intervals of equal size
    - The width of intervals: *(max-min)/N*
    - The most straightforward way; but outliers may dominate presentation; cannot handle skewed data well.
- Attribute values: {0, 4, 12, 16, 18, 24, 26, 28}. Can you put these values into 3 bins using equal-interval binning?

[0, (28-0)/3); [(28-0)/3, 2((28-0)/3))…..

Feature Selection

- Purpose:
  - Many data mining algorithms work better if the number of attributes is lower
  - Simplified model is easier to interpret.
  - Reduce storage requirement and training time.
  - Reduce overfitting and achieve better generalization performance.
- Techniques:
  - To be discussed in Feature Selection.

Lab: Data Preparation

- Files needed
  - Data_preparation.ipynb (Python file)
  - titanic_train.csv (dataset)
  - titanic_test.csv (dataset)