

2143488 BIG DATA
AND ARTIFICIAL
INTELLIGENCE
DR. JING TANG

DATA ACQUISITION & PREPROCESSING

SOURCE OF DATA (1)

- **Operational Systems:** raw form of data containing errors and being scattered over several tables and files
- **Data Warehouses & Data Marts:** continuously collected, cleaned, summarized core data
- Need to do ETL , takes time
- **Online Analytical Processing (OLAP):** cleaned and organized data cubes behind BI apps, which enable to view data from multidimensions
organized data , data warehouse , data mart
and takes time
- **Surveys:** most expensive source of data, which has limited amount, but directly address the objectives
but you get exactly what you want.

SOURCE OF DATA (2)

- **Household and Demographic Databases:** second-hand structured data from external sources
- **Other External Unstructured Databases:** social media, etc.
 - Huge amount of data are needed
 - Labelling data is tedious and error prone

Sentimental analysis

. classify : positive, neutral , negative

. challenge : public data set → fast → but too generic don't know if they are accurate

label data: time consuming

don't label data: not accurate

VARIABLE TYPES

- **Nominal or Categorical Variables:** Lack the properties of order, scale, or distance between values. i.e., Gender, House Type, Credit Card Type *has no order*
- **Ordinal or Rank or Ordered Variables:** Categorical variables with order. i.e., Income Level, Rank *has order*
- **Discrete Variables:** Countable and finite values with order, scale, and distance. i.e., Num of Products
- **Continuous Variables:** Infinite values with order, scale distance. i.e., Time, Height

CROSS-SECTIONAL DATA: IDENTIFIER, VARIABLE, RECORD

The diagram illustrates the structure of cross-sectional data. At the top, three red boxes are labeled "Identifier", "Variables", and "Records". Red arrows point from these labels to their corresponding components in a table below. The table has a header row with columns labeled "id", "var1", "var2", "var3", "var4", and "var5". The body of the table contains 7 rows, each representing a "Record". A blue bracket on the left side of the first row is labeled "record".

id	var1	var2	var3	var4	var5	Records
1	7.3	32.27	0.1	Yes	Male	"record"
2	8.28	40.68	0.56	No	Female	
3	3.35	5.62	0.55	Yes	Female	
4	4.08	62.8	0.83	Yes	Male	
5	9.09	22.76	0.26	No	Female	
6	8.15	90.85	0.23	Yes	Female	
7	7.59	54.94	0.42	Yes	Male	

CROSS-SECTIONAL TIME SERIES DATA (PANEL DATA):

	id	year	var1	var2	var3
Group 1	1	2000	7	74.03	0.55
	1	2001	2	4.6	0.44
	1	2002	2	25.56	0.77
Group 2	2	2000	7	59.52	0.05
	2	2001	2	16.95	0.94
	2	2002	9	1.2	0.08
Group 3	3	2000	9	85.85	0.5
	3	2001	3	98.85	0.32
	3	2002	3	69.2	0.76

DATA FORMAT: ASCII (AMERICAN STANDARD CODE FOR INFORMATION INTERCHANGE)

- Delimited: data is separated by comma, tab or space
 - *.csv (comma-separated value), *.txt (tab-separated), *.prn (space-separated) , excel
 - All data package can read these formats
- Record form (or fixed): data is structured by fixed blocks.
No column headings is available, so it needs a codebook to figure out the layout of the data
 - *.dat, *.txt
 - Special packages: SPSS, SAS...

DATA ROLLUP & DATA INTEGRATION

- **Data Rollup:** Summarize data *by monthly, yearly, etc.*

- Categories to Variables

id	2000_v1	2000_v2	2000_v3	2001_v1	2001_v2	2001_v3	2002_v1	2002_v2	2002_v3
1001	Ross	Sharma		1001	72				
1002	Cary	Carlisle		1002	7				
1003	Stephan	Stroup		1003	20				
1004	Corean	Knott		1004	10				
1005	Guadalupe	Hernandez		1005	87				
1006	Carroll	Courtney		1006	27				
1007	Terry	Starr		1007	59				
1008	Rosevelt	Pruett		1008	83				
1009	Zelma	Roddy		1009	53				
1010	Elmina	Fay		1010	41				

- Rollup with sums, averages, and counts

- **Data Integration:** Merge or Concatenation Multiple

Tables

Player Id	First Name	Last Name	High Score
1001	Ross	Sharma	72
1002	Cary	Carlisle	7
1003	Stephan	Stroup	20
1004	Corean	Knott	10
1005	Guadalupe	Hernandez	87
1006	Carroll	Courtney	27
1007	Terry	Starr	59
1008	Rosevelt	Pruett	83
1009	Zelma	Roddy	53
1010	Elmina	Fay	41



Player Id	First Name	Last Name	High Score
1001	Ross	Sharma	72
1002	Cary	Carlisle	7
1003	Stephan	Stroup	20
1004	Corean	Knott	10
1005	Guadalupe	Hernandez	87
1006	Carroll	Courtney	27
1007	Terry	Starr	59
1008	Rosevelt	Pruett	83
1009	Zelma	Roddy	53
1010	Elmina	Fay	41

A	B	C
1	Item	Qty.
2	Sweets	20
3	Biscuits	100
4	Ice-cream	50
5	Juice	95

A	B	C
1	Item	Qty.
2	Cakes	120
3	Croissants	70
4	Appe pies	75
5	Doughnuts	200

A	B	C
1	Item	Qty.
2	Strawberry	120
3	Bilberry	70
4	Raspberry	85
5	Blackberry	110

Summary sheet		
1	Item	Qty.
2	Sweets	20
3	Biscuits	100
4	Ice-cream	50
5	Juice	95
6	Lollipops	90
7	Cakes	120
8	Croissants	70
9	Appe pies	75
10	Doughnuts	200
11	Pastry	150
12	Strawberry	120
13	Bilberry	70
14	Raspberry	85
15	Blackberry	110
16	Honeyberry	90

CHECK AFTER DATA ACQUISITION:

- 1) Make sure variables are in columns and observations in rows.
- 2) Make sure you have all variables you need.
delete unnecessary variable
- 3) Make sure there is at least one id.
- 4) If times series make sure you have the time (i.e., years, months) you want to include in your study.
- 5) Make sure to make a back-up copy of your original dataset.
Separate original data, work in a copy file.
- 6) Have the codebook handy

ADJUST DATA FORMAT

- Make sure variables are in their expected format.
 - Numeric should be numeric
 - Text should be text.

Var. 1(Numeric)	Var. 2 (String)
1	1
2	2
2	2
4	4
3	3

Cannot do statistic
with Var. 2, rather
than frequencies

WHY PREPROCESS DATA?

- Data in the real world is dirty
 - **Incomplete**: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
 - **Noisy**: containing errors or outliers
 - **Inconsistent**: containing discrepancies in codes or names
- No quality data, no quality mining results!
 - **Quality decisions** must be based on **quality data**
 - Data warehouse needs **consistent integration** of quality data

EXAMPLES OF PROBLEMS FROM INACCURATE DATA

- **Marketing:** An ad campaign using low quality data and reaching out to users with irrelevant offers. This not only reduces customer satisfaction but also misses a significant sales opportunity.
- **Sales:** A sales representative failing to contact previous customers, because of not having their complete, accurate data.
- **Compliance:** Any online business receiving penalties from the government by not meeting data privacy rules for its customers. Facebook could be receiving such a penalty in the wake of Cambridge Analytica scandal.
- **Operations:** Configuring robots and other production machines based on low quality operational data, can cause major problems for manufacturing companies

BENEFITS OF DATA PREPROCESSING

- Streamlined business practices
- Increased productivity
- Faster sales cycle
- Better decisions



HIGH QUALITY DATA

- **Validity:** meets defined business rules or constraints
- **Accuracy:** confirms to a standard or a true value
- **Completeness**
- **Consistency:** equivalency of measures across systems
- **Uniformity:** the same units of measure
- **Traceability:** being able to find the source of the data
- **Timeliness:** recent data

MAJOR TASKS IN DATA PREPROCESSING (1)

- **Data Cleaning:** Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data Integration:** Integration of multiple databases, data cubes, or files, especially to solve inconsistency
- **Data Transformation:** Normalization and aggregation
- **Data Reduction:** Obtains reduced representation in volume but produces the same or similar analytical results
- **Data Discretization:** Part of data reduction but with particular importance, especially for numerical data

Data Cleaning

- Ex: Questionnaire answers
 - ↳ may report reversed question to see if the person paid attention
 - ↳ remove invalid entries (ex: answered all 5)
- Identify and remove outliers
 - ↳ need to check if the outlier is a true record or a possible error
 - ↳ need to check if the outlier is related to the research topic or not.
 - ↳ not always to remove outliers
- Remove inconsistencies
 - ↳ suppose the address for online transaction
 - ↳ definition for "address" might not be "home_address", but rather the most-frequently-used address
 - ↳ ex: ordering for someone else.

Data Cleaning

- Ex: Questionnaire answers
 - ↳ may repeat/revert questions to see if the person paid attention
 - ↳ remove invalid entries (ex: answered all 5)
- Identify and remove outliers
 - ↳ need to check if the outlier is a true record or a possible error
 - ↳ need to check if the outlier is related to the research topic or not.
 - ↳ not always to remove outliers
- Remove inconsistencies
 - ↳ suppose the address for online transaction
 - ↳ definition for "address" might not be "home_address", but rather the most-frequently-used address
 - ↳ ex: ordering for someone else.

Data Integration

- can't realistically load all data to process at once
- Ex: 7-11
 - Branch 1A: collect transactions data one each day into an excel sheet.
 - Suppose you want to know the sales, you gonna vertically concat them
- You might need to merge the data when some info are same, and some columns are different
 - But you have to be careful about column names such as "ID" from different tables
 - Should provide unique, meaningful names!

Data Transformation

- normalization, aggregation
- ex: stocks analysis
 - maybe you want to do daily/monthly forecasting
 - in this case, you need to aggregate data into daily-based, etc.

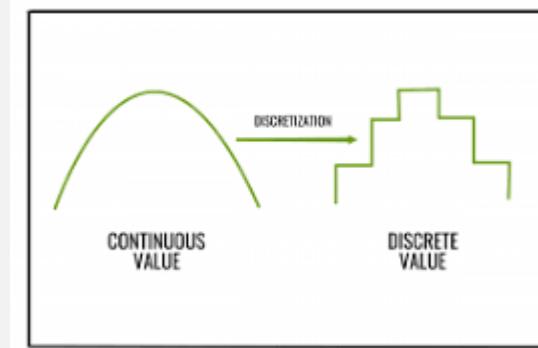
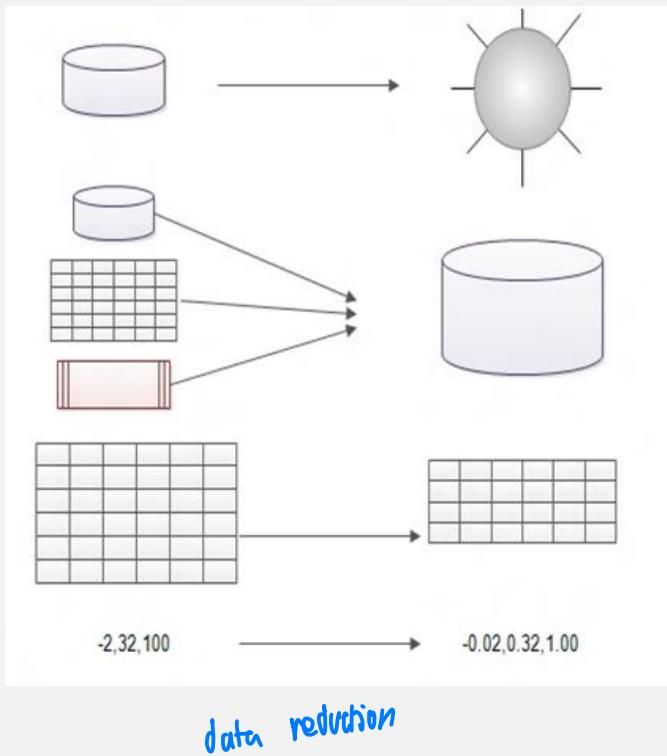
Data Reduction

- Data too big, complex
- limit age: should be in range 16-46
 - ↳ 90% of women are pregnant in this age range
- education label
 - ↳ not going significant if you're lower than university level → 0, university and up → 1

Discretize data

- Midterm:
 - ↳ give graph then ask how to
 - ↳ what the graph means
 - ↳ give example to explain why to perform some actions with the data

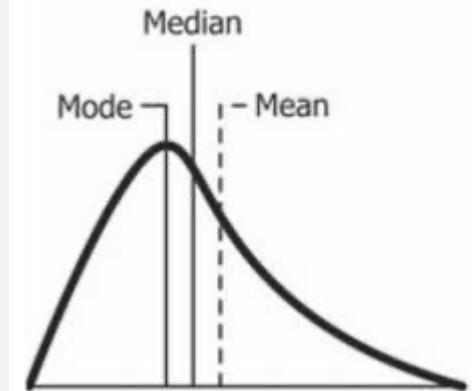
MAJOR TASKS IN DATA PREPROCESSING (2)



DESCRIPTIVE DATA SUMMARIZATION: OVERALL PICTURE OF DATA FOR SUCCESSFUL DATA PREPROCESSING

- **Location** : Central value of the variable

- Mean
↳ works well if the data is a normal distribution
- Median
Star should come to your mind first , mean is very sensitive to outliers
- Mode: unimodal, bimodal, trimodal, etc.
↳ too sensitive based on how you process the data



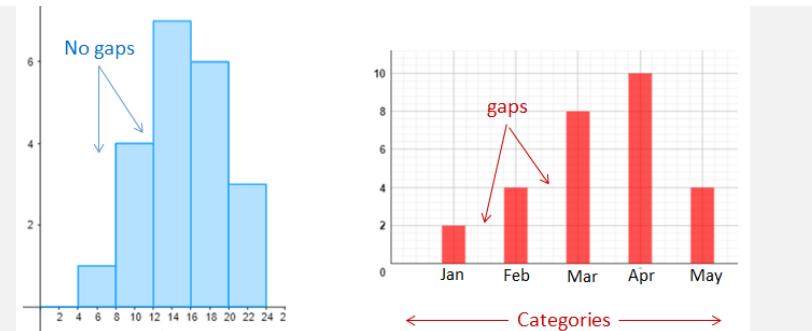
- **Variability**: the spread of the data from the center value

- Variance
- Standard Deviation
- Range

GRAPHICS DISPLAYS (1)

- Histograms vs. Bar Chart**

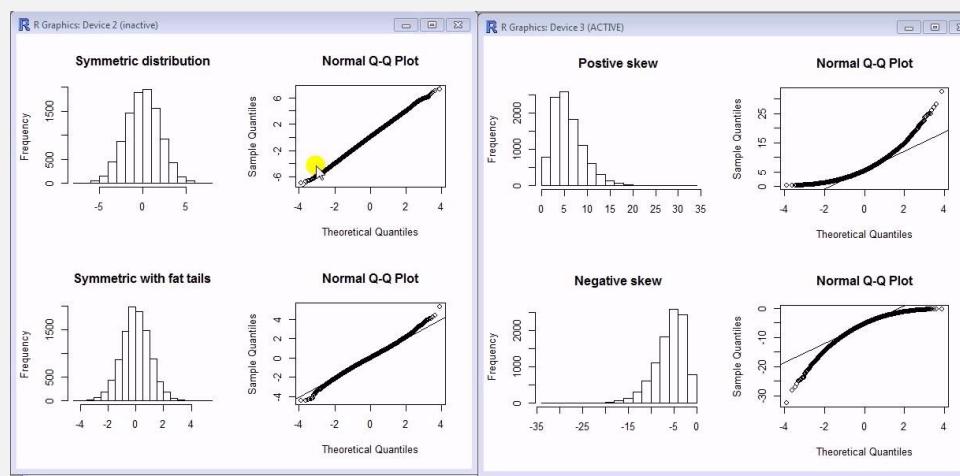
- Univariable distribution



- Quantile-Quantile (q-q) plots**

- Compare two distributions

- used for checking the type of distributions

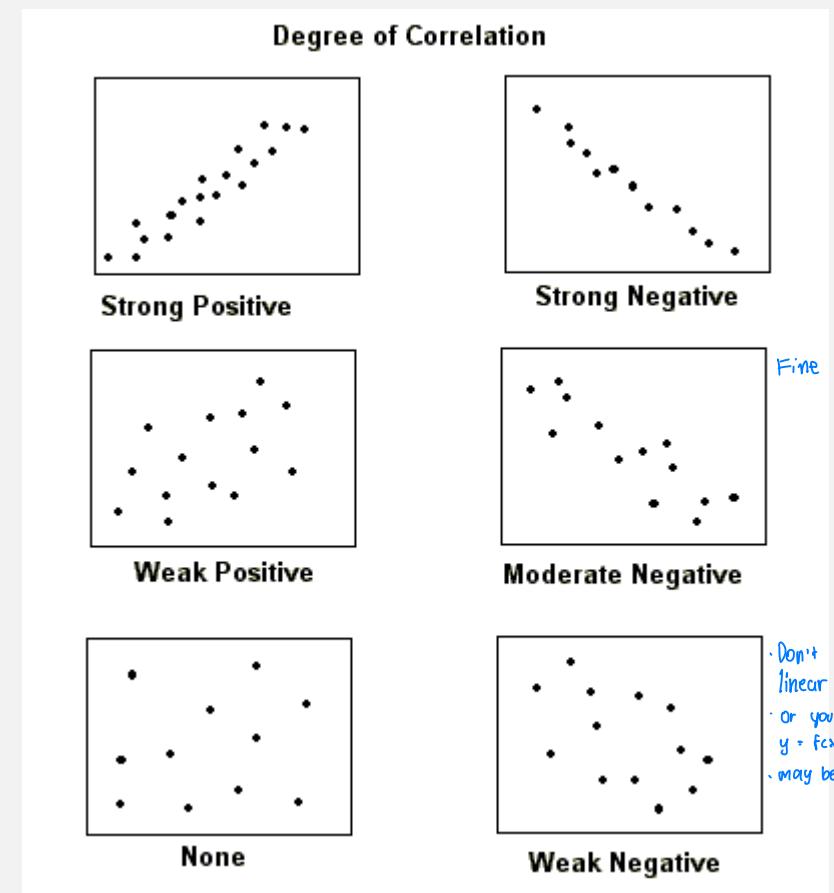


GRAPHICS DISPLAYS (2)

- Scatter plots

- Correlation of two variables

• You can reduce many data away



DATA CLEANING

- 1) Fill in missing values
- 2) Identify outliers and smooth out noisy data
- 3) Correct inconsistent data

MISSING DATA

- Data is not always available
 - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
 - Equipment malfunction
 - Inconsistent with other recorded data and thus deleted
 - Data not entered due to misunderstanding
 - Certain data may not be considered important at the time of entry
 - Not register history or changes of the data
- Missing data may need to be **inferred**

HOW TO HANDLE MISSING DATA?

- **Ignore** the tuple: usually done when class label is missing (assuming the tasks in classification—not effective when the percentage of missing values per attribute varies considerably)
- **Fill in** the missing value **manually**: tedious + infeasible?
- Use a **global constant** to fill in the missing value: e.g., “unknown”, a new class?! *NULL, 0, etc.*
- Use the **attribute mean** to fill in the missing value
↳ good.
↳ not good for time series
- Use the **most probable value** to fill in the missing value:
↳ second best inference- based such as Bayesian formula or decision tree

BE CAREFUL WITH NON-RESPONSE

- Getting rid of the **non-response categories**
 - ‘do not know’, ‘no answer’, ‘no applicable’, ‘not sure’, ‘refused’, etc.
 - higher values like 99, 999, 9999, etc. (or in some cases negative values)

. tab var1

Status of Nat'l Eco	Freq.	Percent	Cum.
Very well	149	10.85	10.85
Fairly well	670	48.80	59.65
Fairly badly	348	25.35	85.00
Very badly	191	13.91	98.91
Not sure	12	0.87	99.78
Refused	3	0.22	100.00
Total	1,373	100.00	

. tab var1_rec

RECODE of var1 (Status of Nat'l Eco)	Freq.	Percent	Cum.
= Very badly	191	14.06	14.06
Fairly badly	348	25.63	39.69
Fairly well	670	49.34	89.03
► Very well	149	10.97	100.00
Total	1,358	100.00	

NOISY DATA

- Noise: random error or variance in a measured variable
- Incorrect attribute values may be due to
 - Faulty data collection instruments
 - Data entry problems
 - Data transmission problems
 - Technology limitation
 - Inconsistency in naming convention
- Other data problems which require data cleaning
 - Duplicate records
 - Incomplete data
 - Inconsistent data



HOW TO HANDLE NOISY DATA?

- Binning method:
 - 1) Sort data and partition into (equi-depth) bins
 - 2) Smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Clustering
 - Detect and remove outliers
- Combined computer and human inspection
 - detect suspicious values and check by human
- Regression
 - smooth by fitting the data into regression functions

SIMPLE DISCRETIZATION METHODS: BINNING

- **Equal-width** (distance) partitioning:
 - It divides the range into N intervals of equal size: uniform grid
 - if A and B are the lowest and highest values of the attribute, the width of intervals will be: $W = (B-A)/N$.
 - The most straightforward
 - But outliers may dominate presentation
 - Skewed data is not handled well.
- **Equal-depth** (frequency) partitioning:
 - It divides the range into N intervals, each containing approximately same number of samples
 - Good data scaling
 - Managing categorical attributes can be tricky.

EXAMPLE 1: BINNING METHODS FOR DATA SMOOTHING

- Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
- Partition into (equi-depth) bins:
 - Bin 1: 4, 8, 9, 15
 - Bin 2: 21, 21, 24, 25
 - Bin 3: 26, 28, 29, 34
- Smoothing by bin means:
 - Bin 1: 9, 9, 9, 9
 - Bin 2: 23, 23, 23, 23
 - Bin 3: 29, 29, 29, 29
- Smoothing by bin boundaries:
 - Bin 1: 4, 4, 4, 15
 - Bin 2: 21, 21, 25, 25
 - Bin 3: 26, 26, 26, 34

DATA INTEGRATION

- Data integration: combines data from multiple sources into a coherent store
- **Schema** integration
 - Integrate metadata from different sources
 - Entity identification problem: identify real world entities from multiple data sources, e.g., A.cust-id \equiv B.cust-#
- Detecting and resolving data **value** conflicts
 - For the same real-world entity, attribute values from different sources are different
 - Possible reasons: different representations, different scales, e.g., metric vs. British units

HANDLING REDUNDANT DATA

- Redundant data occur often when integration of multiple databases
 - The same attribute may have different names in different databases
 - Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

DATA TRANSFORMATION

- Smoothing: remove noise from data
- Aggregation: summarization, data cube construction
- Generalization: concept hierarchy climbing
- Normalization: scaled to fall within a small, specified range
 - min-max normalization
 - z-score normalization
 - normalization by decimal scaling

DATA TRANSFORMATION: NORMALIZATION

- Min-Max Normalization *want data from 0 → 1*

$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{newmax}_A - \text{newmin}_A) + \text{newmin}_A$$

- Z-score Normalization

$$v' = \frac{v - \text{mean}_A}{sd_A}$$

- Normalization by decimal scaling *change super big data to smaller data*

$$v' = \frac{v}{10^j}$$

where j is the smallest integer such that $\text{Max}(|v'|) < 1$

DATA REDUCTION STRATEGIES

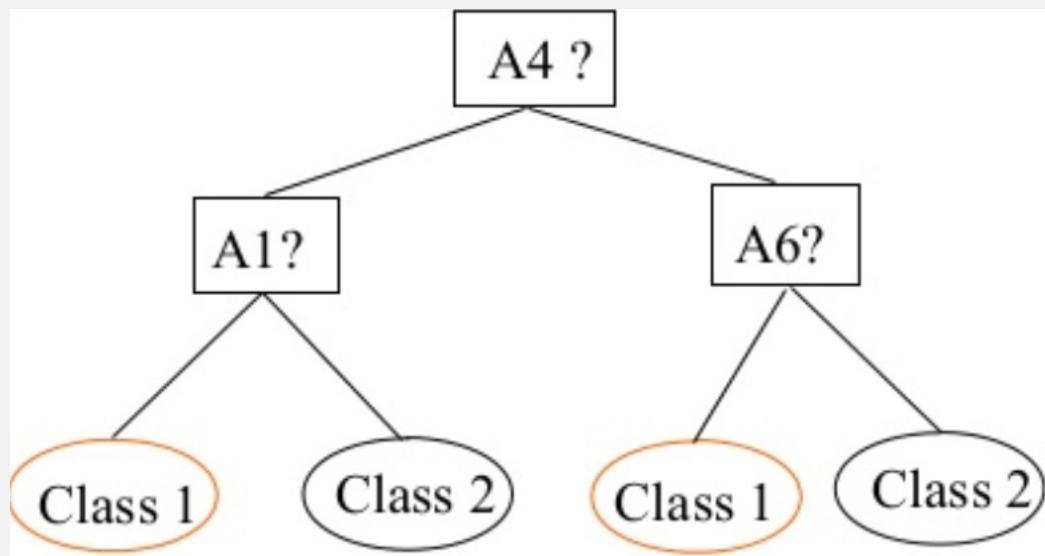
- Warehouse may store terabytes of data: Complex data analysis/mining may take a very long time to run on the complete data set
- Data reduction
 - Obtains a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results
- Data reduction strategies
 - Data cube aggregation
 - Dimensionality reduction
 - Numerosity reduction
 - Discretization and concept hierarchy generation

DIMENSIONALITY REDUCTION

- Feature selection (i.e., attribute subset selection):
 - Select a minimum set of features such that the probability distribution of different classes given the values for those features is as close as possible to the original distribution given the values of all features
 - Reduce # of patterns in the patterns, easier to understand

EXAMPLE 2: DECISION TREE INDUCTION

- Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$
- Reduced attribute set: $\{A_1, A_4, A_6\}$



HISTOGRAMS

- A popular data reduction technique
- Divide data into buckets and store average (sum) for each bucket
- Can be constructed optimally in one dimension using dynamic programming
- Related to quantization problems

DISCRETIZATION

- A Three types of attributes:
 - Nominal — values from an unordered set (i.e., names, color)
 - Ordinal — values from an ordered set (i.e., ranking)
 - Continuous — real numbers (i.e., scores, times)
- Discretization:
 - Divide the range of a continuous attribute into intervals
 - Some classification algorithms only accept categorical attributes.
 - Reduce data size by discretization
 - Prepare for further analysis

DISCRETIZATION AND CONCEPT HIERARCHY

- Discretization
 - Reduce the number of values for a given **continuous attribute** by dividing the range of the attribute into **intervals**. Interval labels can then be used to replace actual data values.
- Concept hierarchies
 - Reduce the data by collecting and replacing **low level concepts** (such as numeric values for the attribute age) by **higher level concepts** (such as young, middle-aged, or senior).

DISCRETIZATION FOR NUMERIC DATA

- Binning
- Histogram Analysis
- Clustering Analysis

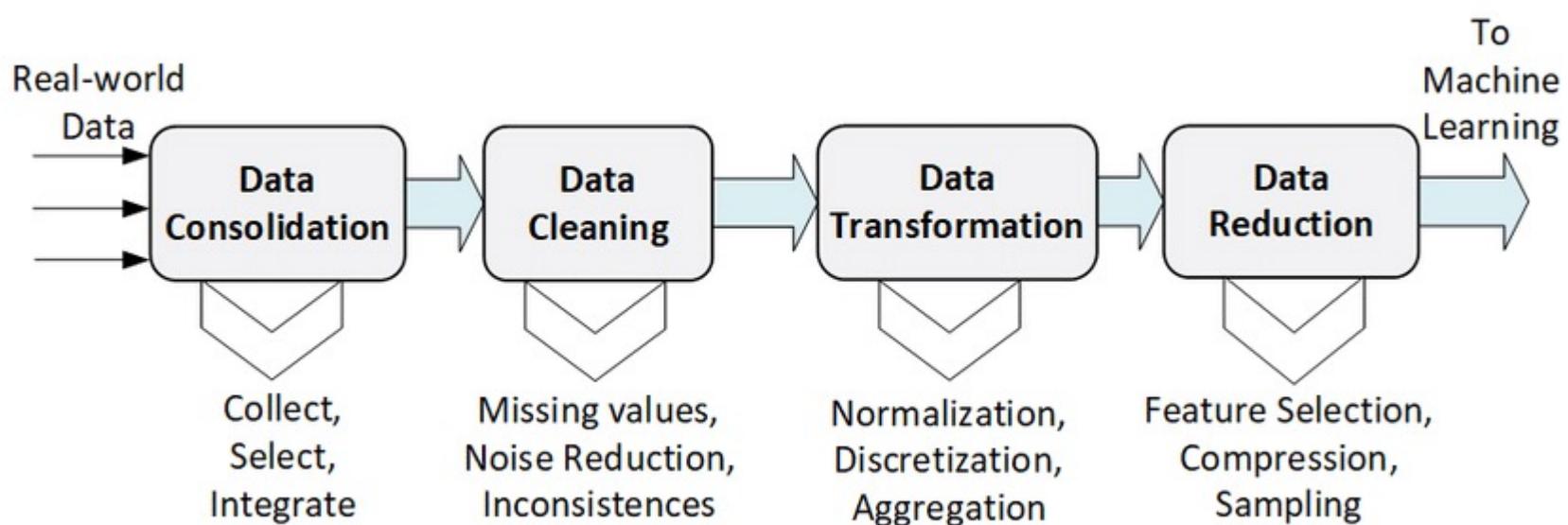
REPLACE WIDE TO LONG

Quarter	2016	2017	2018	2019
1	2.1	2.9	3.1	3.4
2	2.2	2.9	3	3.5
3	2.4	2.8	3.3	3.6
4	2.3	2.7	3.2	3.5



Quarter	Year	value
1	2016	2.1
2	2016	2.2
3	2016	2.4
4	2016	2.3
1	2017	2.9
2	2017	2.9
3	2017	2.8
4	2017	2.7
1	2018	3.1
2	2018	3
3	2018	3.3
4	2018	3.2
1	2019	3.4
2	2019	3.5
3	2019	3.6
4	2019	3.5

SUMMARY



HW 2: SUMMARIZE THE ANALYSIS IN A PPT WITH BRIEF EXPLANATION

- Download 2006 microdata survey about housing for the state of Idaho using “HW2 Housing Survey.csv” and load the data into Excel\RapidMiner\Python. The code book is “HW2 Housing Survey.pdf”.
 - Please answer the following question: (better in pictures)
 - How many properties are worth \$1,000,000 or more?
 - How many people recorded in a house on average?
 - Draw a graph to show the relationship between the property value and the number of persons recorded?
 - Normalize family income into a range (0~1). Compare before vs after in histogram.
 - Create 5 bins for family income.
- / go check which column means property price?
- perform in code only → 80%
→ slides or pdf to explain results in 1 sentence
→ submit 2 files
- don't need to screenshot code,
- submit code as individual file
- & Focus on submitting powerpoint*