

Data Collection & Preprocessing

CSGE603130 - Kecerdasan Artifisial dan Sains Data Dasar
Semester Genap 2022/2022

Siti Aminah, Dinial Utami

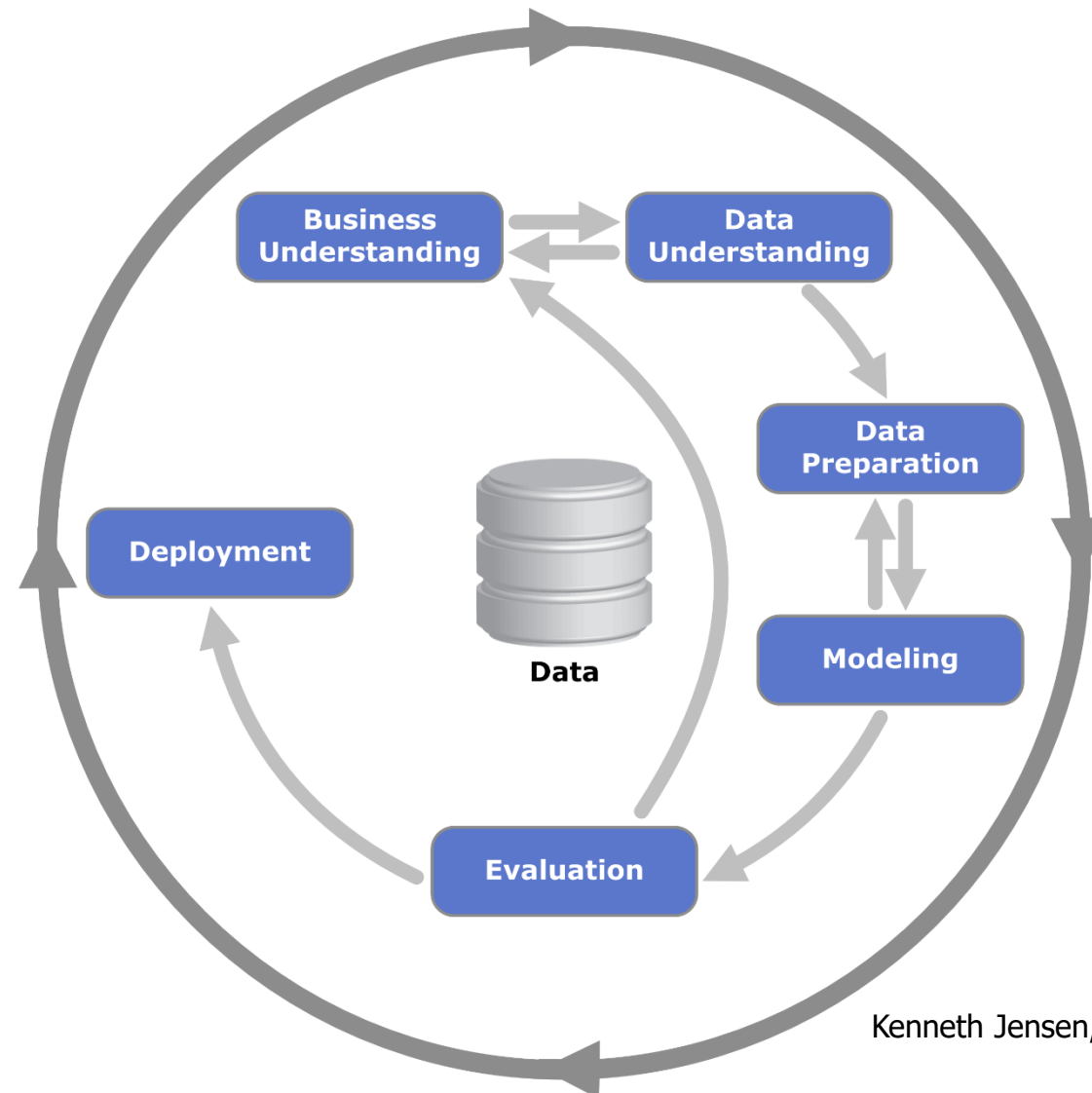


UNIVERSITAS
INDONESIA

Veritas, Probitas, Iustitia

FACULTY OF
**COMPUTER
SCIENCE**

CRISP-DM: Cross-industry standard process for data mining



Kenneth Jensen, CC BY-SA 3.0, via Wikimedia Commons

Data Collection

Data Collection

- Data collection is the **process of accumulating data that's required to solve a problem statement** (that formulated in business understanding phase).
- Two main methods:

Primary Data Collection

- Interviews
- Observations
- Surveys & Questionnaires
- Focus Group
- Oral Histories

Secondary Data Collection

- Internet
- Government Archives
- Libraries

Open Data

- The idea behind open data is that some data should be freely available in a public domain that can be used by anyone as they wish, without restrictions from copyright, patents, or other mechanisms of control
- Open Data Principles:
 - Public
 - Accessible
 - Reusable
 - Complete
 - Timely
 - Managed Post-Release

Some Open Data Resources

- Kaggle (www.kaggle.com)
- UCI Machine Learning Repository (<https://archive.ics.uci.edu/ml/index.php>)
- Nasdaq, Financial Data (<https://data.nasdaq.com/>)
- Data.gov (<https://www.data.gov/>)
- World Bank Data (<https://www.data.gov/>)
- Grouplens, University of Minnesota (<https://grouplens.org/>)
- Google Dataset Search (<https://datasetsearch.research.google.com/>)
- Open ML (<https://www.openml.org/search?type=data>)

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

[View ALL Data Sets](#)

Welcome to the UC Irvine Machine Learning Repository!

We currently maintain 622 data sets as a service to the machine learning community. You may [view all data sets](#) through our searchable interface. For a general overview of the Repository, please visit our [About page](#). For information about citing data sets in publications, please read our [citation policy](#). If you wish to donate a data set, please consult our [donation policy](#). For any other questions, feel free to [contact the Repository librarians](#).



In Collaboration With:



Latest News:







- 09-24-2018:** Welcome to the new Repository admins Dheeru Dua and Efi Karra Taniskidou!
- 04-04-2013:** Welcome to the new Repository admins Kevin Bache and Moshe Lichman!
- 03-01-2010:** [Note](#) from donor regarding Netflix data
- 10-16-2009:** Two new data sets have been added.
- 09-14-2009:** Several data sets have been added.
- 03-24-2008:** New data sets have been added!
- 06-25-2007:** Two new data sets have been added: UJI Pen Characters, MAGIC Gamma Telescope

Featured Data Set: [UJI Pen Characters \(Version 2\)](#)









Task: Classification
Data Type: Multivariate, Sequential
Instances: 11640

Newest Data Sets:

- 06-05-2021:**  [Average Localization Error \(ALE\) in sensor node localization process in WSNs](#)
- 05-25-2021:**  [9mers from culpdb](#)
- 05-18-2021:**  [TamilSentiMix](#)
- 05-02-2021:**  [Accelerometer](#)
- 04-21-2021:**  [Synchronous Machine Data Set](#)
- 04-21-2021:**  [Synchronous Machine Data Set](#)

Most Popular Data Sets (hits since 2007):

- 4495834:**  [Iris](#)
- 2392021:**  [Adult](#)
- 1849949:**  [Wine](#)
- 1783353:**  [Wine Quality](#)
- 1770064:**  [Heart Disease](#)
- 1673153:**  [Bank Marketing](#)

DATA PRODUCTS | PUBLISHERS

Browse

Filters

- ☐ Premium
- ☐ Free

Asset Class

- ☐ Equities
- ☐ Currencies
- ☐ Interest Rates & Fixed Income
- ☐ Options
- ☐ Indexes
- ☐ Mutual Funds & ETFs
- ☐ Real Estate
- ☐ Venture Capital & Private Equity
- ☐ Economy & Society
- ☐ Energy
- ☐ Agriculture

CORE FINANCIAL DATA | ALTERNATIVE DATA | ESG DATA HUB

Search bar containing the text "banking".

There are 99 databases with data on 'banking'

End of Day US Stock Prices

Professional-grade EOD stock prices, dividends, adjustments and splits for publicly-traded US stocks.

PREMIUM | HAS SAMPLE DATA

EOD TICKERS	
TICKER	COMPANY_NAME
COLB	Columbia Banking System Inc.
WBK	Westpac Banking Corporation
SBCF	Seacoast Banking Corporation of Florida

Jakarta Open Data Portal



Wikidata



WIKIDATA

Main page

Item [Discussion](#)

Indonesia (Q252)

republic in Southeast Asia
Republic of Indonesia | NKRI | Negara Kesatuan

<https://www.wikidata.org/wiki/Q252>

official language



Indonesian

▼ 2 references

reference URL

http://badanbahasa.kemdikbud.go.id/lamanbahasa/sites/default/files/UU_2009_24.pdf

stated in

Constitution of Indonesia

section, verse, or paragraph

36

anthem



Indonesia Raya

▼ 1 reference

reference URL

http://badanbahasa.kemdikbud.go.id/lamanbahasa/sites/default/files/UU_2009_24.pdf


motto






Bhinneka Tunggal Ika

▼ 0 references

Wikidata

 Wikidata Query Service

[Examples](#) [Help](#) [More tools](#)



```
1 SELECT ?countryLabel ?anthemLabel
2 WHERE {
3   ?country wdt:P31 wd:Q3624078 .
4   ?country wdt:P85 ?anthem .
5   SERVICE wikibase:label { bd:serviceParam wikibase:language "en" }
6 }
```

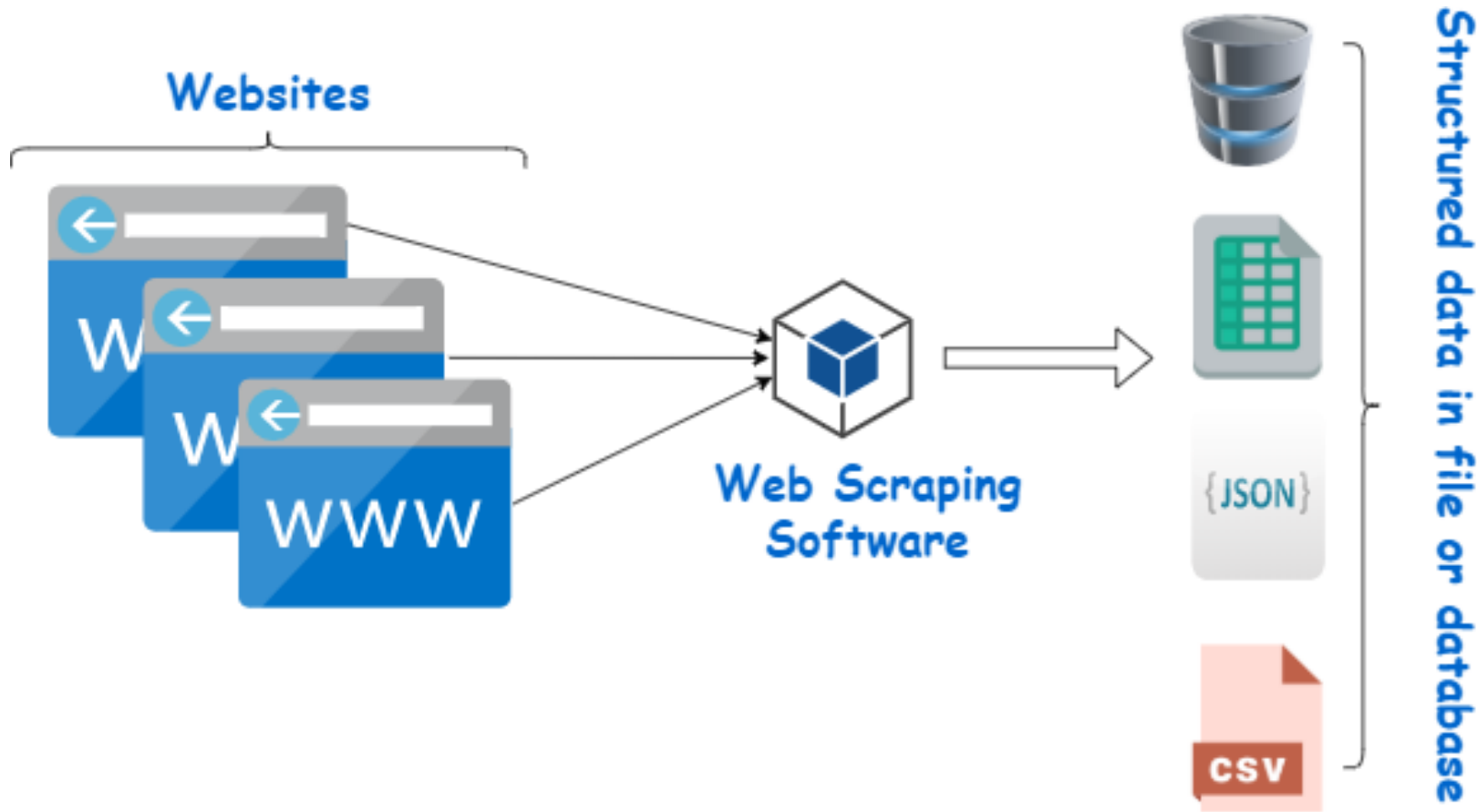
countryLabel	anthemLabel
Indonesia	Indonesia Raya
India	Jana Gana Mana
Madagascar	Ry Tanindrazanay malala ô!
São Tomé and Príncipe	Independência total

• • • • •

Social Media Data

- Social media has become a gold mine for collecting data to analyze for research or marketing purposes.
- This is facilitated by the Application Programming Interface (API) that social media companies provide to researchers and developers.
- Think of the API as a set of rules and methods for asking and sending data.
- For various data-related needs (e.g., retrieving a user's profile picture), one could send API requests to a particular social media service. This is typically a programmatic call that results in that service sending a response in a structured data format, such as an XML.

Web Scrapping



Web scraping software examples:

- BeautifulSoup
- Scrapy

Multimodal Data

- We are living in a world where more and more devices exist – from lightbulbs to cars – and are getting connected to the Internet, creating an emerging trend of the Internet of Things (IoT).
- These devices are generating and using much data, but not all of which are “traditional” types (numbers, text).
- When dealing with such contexts, we may need to collect and explore multimodal (different forms) and multimedia (different media) data such as images, music and other sounds, gestures, body posture, and the use of space.

Commonly Used Data Format

- CSV (Comma-Separated Values)
 - the most common import and export format for spreadsheets and databases.
- TSV (Tab-Separated Values)
 - use for raw data and can be imported into and exported from spreadsheet software.
- XML (eXtensible Markup Language).
 - designed to be both human- and machinereadable, and can thus be used to store and transport data.
- RSS (Really Simple Syndication)
 - a format used to share data between services, and which was defined in the 1.0 version of XML
- JSON (JavaScript Object Notation)
 - a lightweight data-interchange format.
 - It is not only easy for humans to read and write, but also easy for machines to parse and generate.
 - It is based on a subset of the JavaScript Programming Language, Standard ECMA-262, 3rd Edition – December 1999

Commonly Used Data Format

```
treat,before,after,diff  
No Treatment,13,16,3  
No Treatment,10,18,8  
No Treatment,16,16,0  
Placebo,16,13,-3  
Placebo,14,12,-2  
Placebo,19,12,-7  
Seroxat (Paxil),17,15,-2  
Seroxat (Paxil),14,19,5  
Seroxat (Paxil),20,14,-6
```

CSV

```
Name<TAB>Age<TAB>Address  
Ryan<TAB>33<TAB>1115 W Franklin  
Paul<TAB>25<TAB>Big Farm Way  
Jim<TAB>45<TAB>W Main St  
Samantha<TAB>32<TAB>28 George St
```

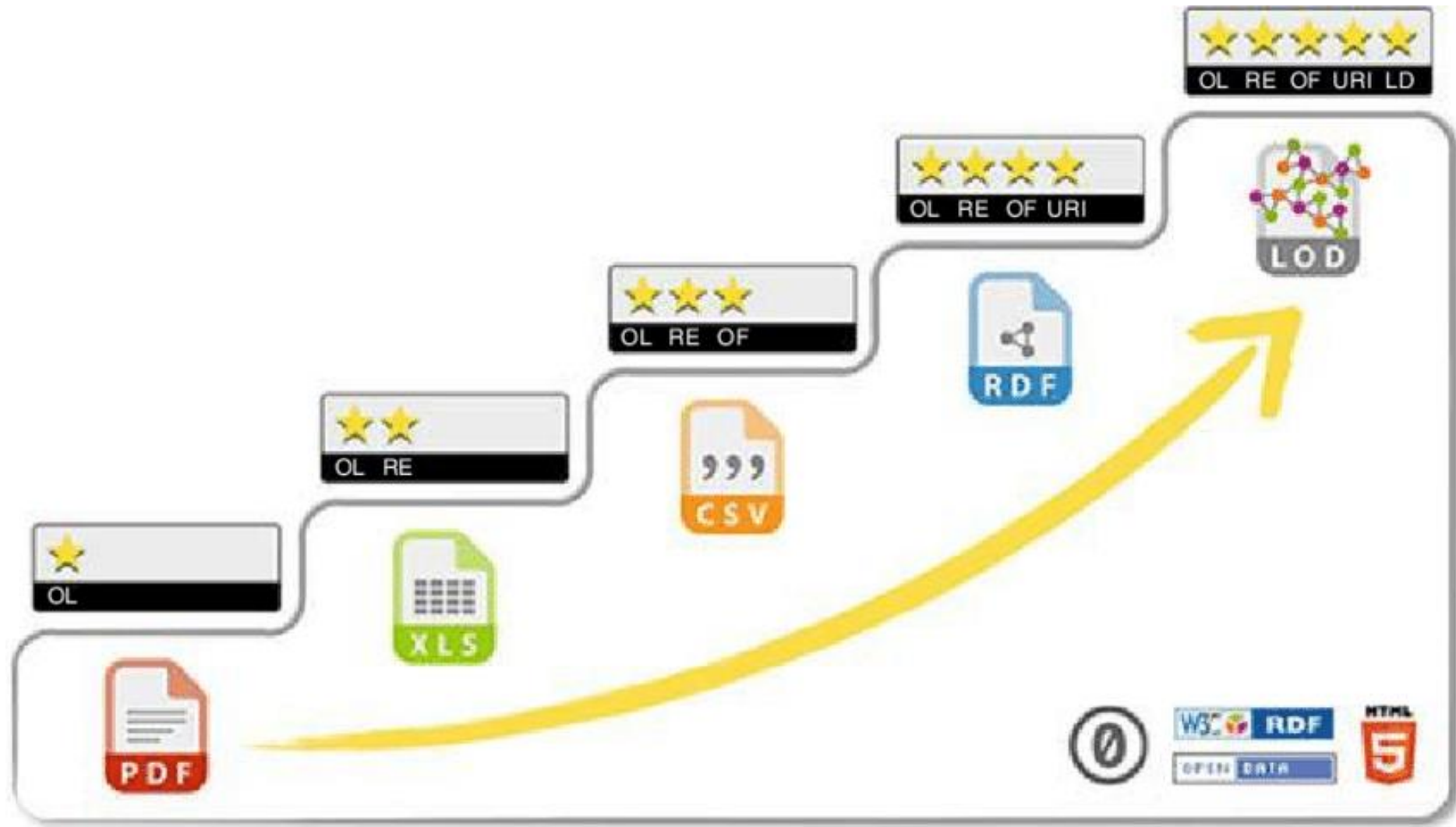
TSV

Commonly Used Data Format

```
<?xml version="1.0" encoding="UTF-8"?>
<bookstore>
  <book category="information science" cover="hardcover">
    <title lang="en">Social Information Seeking</title>
    <author>Chirag Shah</author>
    <year>2017</year>
    <price>62.58</price>
  </book>
  <book category="data science" cover="paperback">
    <title lang="en">Hands-On Introduction to Data
      Science</title>
    <author>Chirag Shah</author>
    <year>2019</year>
    <price>50.00</price>
  </book>
</bookstore>
```

XML

5-Star Data



Data Preprocessing

"Dirty" Data (?)

Data Cleaning

Data Integration

Data Transformation

Data Selection

Data Reduction

Data Discretization

Data Balancing

“Dirty” Data (?)

- Data in the real world is often “**dirty**”; means that they need to be “**cleaned up**” before they can be used for a desired purpose.



Source: <https://skillzme.com/nitty-gritty-dirty-data-infographic/>

An illustration of problems in (survey) data

- ❑ Respondents only answering a portion of questions
- ❑ Respondents not meeting our target criteria
- ❑ Respondents speeding through our survey
- ❑ Straight-line respondents
- ❑ Respondents giving unrealistic answers
- ❑ Respondents giving contradictory responses

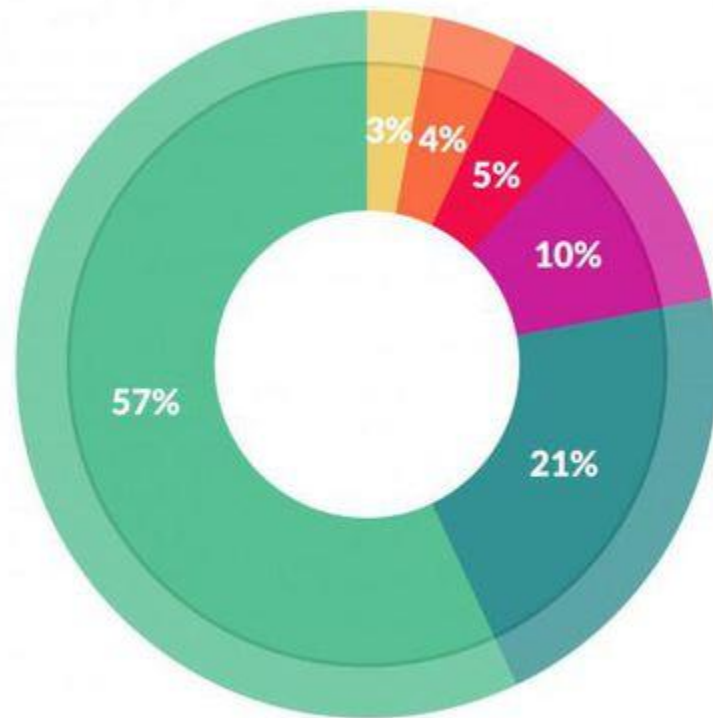
Garbage In



Garbage Out



-
- Data can be “dirty” due to some factors:
 - **Incomplete:** When some of the attribute values are lacking, certain attributes of interest are lacking, or attributes contain only aggregate data.
 - **Noisy:** When data contains errors or outliers. E.g.: extreme values (extremely low or high) that can severely affect the dataset’s range.
 - **Inconsistent:** Data contains discrepancies in codes or names. E.g.: the “Name” column for registration records of employees contains values other than alphabetical letters.



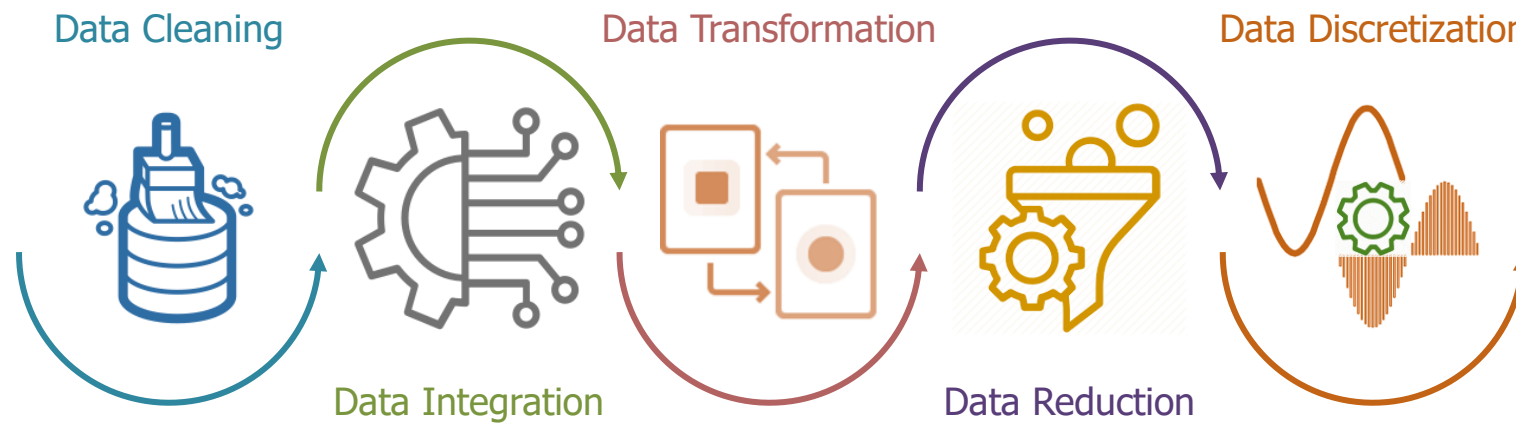
What's the least enjoyable part of data science?

- Building training sets: 10%
- Cleaning and organizing data: 57%
- Collecting data sets: 21%
- Mining data for patterns: 3%
- Refining algorithms: 4%
- Other: 5%

Source:

<https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/?sh=7c14b58f6f63>

- So, what should we do to handle such “dirty” data?
- Any technique involved in **“treating dirty data”/any act transforming data obtained from the real world into a form that can readily be used for a desired purpose** is called **Data Preparation/Data Pre-processing**.



Data Cleaning



- Since there are several reasons why data could be “dirty,” there are just as many ways to “clean” it.
- For this discussion, we will look at three key methods that describe ways in which data may be “cleaned,” or better organized, or scrubbed of potentially incorrect, incomplete, or duplicated information.



- **Data Munging:**

- Often, the data is not in a format that is easy to work with.
- Thus, we need to convert it to something more suitable for a computer to understand.
- The approaches to take are all about **manipulating or wrangling (or munging) the data** to turn it into something that is more convenient or desirable.

Consider the following text recipe. "Add two diced tomatoes, three cloves of garlic, and a pinch of salt in the mix."



Table 2.2 Wrangled data for a recipe.

Ingredient	Quantity	Unit/size
Tomato	2	Diced
Garlic	3	Cloves
Salt	1	Pinch

• Handling Missing Data:

- Sometimes data may be in the right format, but some of the values are missing.
- Missing values occur when no data value is stored for a variable (feature) in an observation.
- Usually “missing value: in dataset” appears as “?”, “N/A”, 0 or just a blank cell.

	symboling	normalized- losses:	make	fuel- type	aspiration	num-of- doors	body-style	drive- wheels	engine- location	wheel-base
0	3	?	alfa-romero	gas	std	two	convertible	rwd	front	88.6
1	3	?	alfa-romero	gas	std	two	convertible	rwd	front	88.6
2	1	?	alfa-romero	gas	std	two	hatchback	rwd	front	94.5
3	2	164	audi	gas	std	four	sedan	fwd	front	99.8
4	2	164	audi	gas	std	four	sedan	4wd	front	99.4

Source: <https://archive.ics.uci.edu/ml/datasets/automobile>

-
- Although **there is no single answer that always works for all scenarios**, these are the typical options you can consider :
 - Check with the person/group that collected the data
 - Drop the missing value (drop the whole variable or single data entry with missing value)
 - Replace the missing value
 - Leave it as missing value

	symboling	normalized- losses:	make	fuel- type	aspiration	num-of- doors	body-style	drive- wheels	engine- location	wheel-base
0	3	?	alfa-romero	gas	std	two	convertible	rwd	front	88.6
1	3	?	alfa-romero	gas	std	two	convertible	rwd	front	88.6
2	1	?	alfa-romero	gas	std	two	hatchback	rwd	front	94.5
3	2	164	audi	gas	std	four	sedan	fwd	front	99.8
4	2	164	audi	gas	std	four	sedan	4wd	front	99.4
Average		122								

- Replace missing value by the average value of the entire variable:
 - As an example, suppose we have some entries that have missing values for the 'normalized-losses' column, and the column average for entries with data is 122. While there is no way for us to get an accurate guess of what the missing values under the 'normalized-losses' column should have been, you can **approximate their values** using **the average value of the column, 122**.

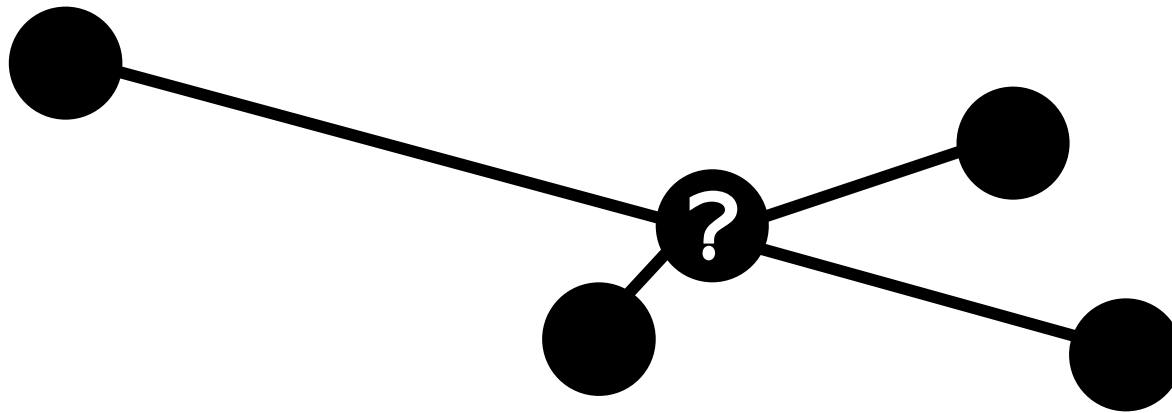
- Replace missing value by frequency

- For a variable like 'fuel-type', there isn't an "average" fuel type, since the variable values are not numbers. In this case, **one possibility is to try using the mode – the most common, like "gasoline"**.

-
- Replace missing value by zero or other constant

	col1	col2	col3	col4	col5		col1	col2	col3	col4	col5	
0	2	5.0	3.0	6	NaN	df.fillna(0)	0	2	5.0	3.0	6	0.0
1	9	NaN	9.0	0	7.0		1	9	0.0	9.0	0	7.0
2	19	17.0	NaN	9	NaN		2	19	17.0	0.0	9	0.0

-
- Replace missing value by nearest neighbor



-
- Replace missing value by nearest neighbor

X	Y
10	5
11	?
30	1

-
- Replace missing value by nearest neighbor

X	Y
10	5
11	5
30	1

- Replace missing value based on other function.

- Sometimes we may find another way to guess the missing data. This is usually because **the data gatherer knows something additional about the missing data.**
- For example, he may know that the missing values tend to be **old cars, and the normalized losses of old cars are significantly higher than the average vehicle.**

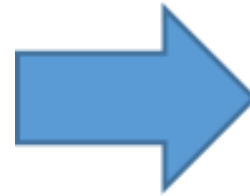
• **Data Formatting:**

- Data is usually collected from different places, by different people, which may be stored in different formats.
- Data formatting means bringing data into a common standard of expression that allows users to make meaningful comparisons.
- Data formatting ensures that data is consistent and easily understandable.

Non-formatted data:

- confusing
- hard to aggregate
- hard to compare

City
NY
New York
N.Y
N.Y



City
New York
New York
New York
New York

Formatted data:

- more clear
- easy to aggregate
- easy to compare

• **Handling Noisy Data:**

- **Noisy Data** may be a result of faulty data collection instruments, data entry problems, or technology limitations.
- E.g.: a digital thermometer measures temperature to one decimal point, but the storage system ignores the decimal points. This is obviously a big deal when this is used to measure human temperatures. The system will consider both 99.4°F (means you are fine) and 99.8°F (means you have a fever) temperatures as 99°F, then it fails to differentiate between healthy and sick persons!

• Handling Noisy Data:

- **Regression**: smooth by fitting the data into regression functions.
- **Clustering**: detect and remove outliers
- **Combined computer and human inspection**: detect suspicious values and check by human (e.g., deal with possible outliers)

Data Integration



- To be as efficient and effective for various data analyses as possible, data from various sources commonly needs to be integrated.
- The following steps describe how to integrate multiple databases or files.
 - Combine data from multiple sources into a coherent storage place.
 - Engage in schema integration or combine metadata from different sources. E.g., A.cust-id \equiv B.cust-#.

-
- ...
 - Detect and resolve data value conflicts.
 - For the same real world entity, attribute values from different sources are different.
 - Reasons for this conflict could be different representations or different scales; for example, metric vs. British units.
 - Entity identification problem: identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton.

- ...

- Address redundant data in data integration. Redundant data is commonly generated in the process of integrating multiple databases. For example:
 - The same attribute may have different names in different databases.
 - One attribute may be a “derived” attribute in another table; for example, annual revenue.
 - Correlation analysis such as Pearson’s correlation and Chi-Square test may detect instances of redundant data:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

$$\chi^2 = \sum \frac{(\text{Expected} - \text{Observed})^2}{\text{Expected}}$$

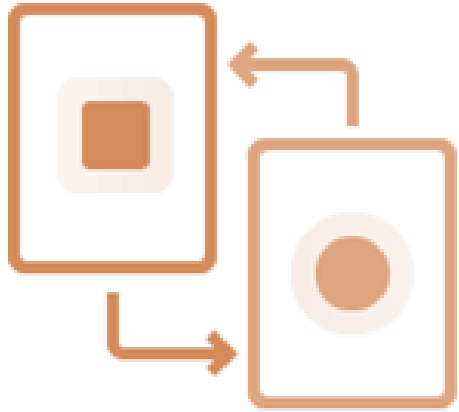
-
- One attribute may be a “derived” attribute in another table:

Person ID	is_male	is_female
1	1	0
2	1	0
3	1	0
4	0	1
5	1	0
6	1	0
7	0	1

- What is the correlation between the two variables?

	x	y					
Person ID	is_male	is_female	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(x_i - \bar{y})^2$
1	1	0	0.285714	-0.28571	-0.081632653	0.08163265	0.08163265
2	1	0	0.285714	-0.28571	-0.081632653	0.08163265	0.08163265
3	1	0	0.285714	-0.28571	-0.081632653	0.08163265	0.08163265
4	0	1	-0.71429	0.714286	-0.510204082	0.51020408	0.51020408
5	1	0	0.285714	-0.28571	-0.081632653	0.08163265	0.08163265
6	1	0	0.285714	-0.28571	-0.081632653	0.08163265	0.08163265
7	0	1	-0.71429	0.714286	-0.510204082	0.51020408	0.51020408
	\bar{x}	\bar{y}			$\sum(x_i - \bar{x})(y_i - \bar{y})$	$\sum(x_i - \bar{x})^2$	$\sum(y - \bar{y})^2$
	0.71429	0.285714			-1.428571429	1.42857143	1.42857143
				r	-1		

Data Transformation



- Data must be transformed so it is consistent and readable (by a system). Some processes include:
 - **Smoothing**: Remove noise from data.
 - **Aggregation**: Summarization, data cube construction.
 - **Generalization**: Concept hierarchy climbing.
 - **Normalization**: Scaled to fall within a small, specified range and aggregation.
 - **Attribute or feature construction**: new attributes constructed from the given ones.

Not-normalized

age	income
20	100000
30	20000
40	500000



age	income
0.2	0.2
0.3	0.04
0.4	1

Normalized

• Data Normalization:

- Consider a dataset containing two features: "age" and "income", where "age" ranges from 0-100, while "income" ranges from 0-20,000 and higher.
- When we do further analysis, like linear regression, for example, the attribute "income" will intrinsically influence the result more, due to its larger value, but this doesn't necessarily mean it is more 'important' as a predictor.

- **Data Normalization Methods:**

- Simple Feature Scaling:

- $x_{new} = \frac{x_{old}}{x_{max}}$

- Min-Max:

- $x_{new} = \frac{x_{old} - x_{min}}{x_{max} - x_{min}}$

- Z-score:

- $x_{new} = \frac{x_{old} - \mu}{\sigma}$

Data Selection

- ❑ Select subsets of our data based on some criteria
 - > Column selection: Pick specific columns of our dataset
 - > Row selection: Pick specific rows of our dataset

Data Selection

Column Selection

No	Employee ID	Name	Position	Company	Monthly Salary (in IDR)
1	E012	Andi	Software Engineer	PT ABC	7000000
2	E123	Budi	Web Developer	PT ABC	6000000
3	E321	Ani	HR Manager	PT ABC	6000000
4	E222	Endang	CTO	PT ABC	12000000
5	E555	Sarah	CEO	PT ABC	15000000
6	Z012	Boy	Software Engineer	PT DEF	8000000
7	Z123	Tom	Web Developer	PT DEF	7000000
8	Z321	Julia	HR Manager	PT DEF	7000000
9	Z222	Dedy	CTO	PT DEF	13000000
10	Z555	Sinta	CEO	PT DEF	16000000



Position	Monthly Salary (in IDR)
Software Engineer	7000000
Web Developer	6000000
HR Manager	6000000
CTO	12000000
CEO	15000000
Software Engineer	8000000
Web Developer	7000000
HR Manager	7000000
CTO	13000000
CEO	16000000

Data Selection

Row Selection

No	Employee ID	Name	Position	Company	Monthly Salary (in IDR)
1	E012	Andi	Software Engineer	PT ABC	7000000
2	E123	Budi	Web Developer	PT ABC	6000000
3	E321	Ani	HR Manager	PT ABC	6000000
4	E222	Endang	CTO	PT ABC	12000000
5	E555	Sarah	CEO	PT ABC	15000000
6	Z012	Boy	Software Engineer	PT DEF	8000000
7	Z123	Tom	Web Developer	PT DEF	7000000
8	Z321	Julia	HR Manager	PT DEF	7000000
9	Z222	Dedy	CTO	PT DEF	13000000
10	Z555	Sinta	CEO	PT DEF	16000000



No	Employee ID	Name	Position	Company	Monthly Salary (in IDR)
4	E222	Endang	CTO	PT ABC	12000000
5	E555	Sarah	CEO	PT ABC	15000000
9	Z222	Dedy	CTO	PT DEF	13000000
10	Z555	Sinta	CEO	PT DEF	16000000

Retain rows with Salary of at least
10000000 (10 million)

Data Reduction



- Data reduction is a key process in which a **reduced representation of a dataset that produces the same or similar analytical** results is obtained.
- A **database/data warehouse** may store **terabytes** of data. **Complex data analysis** may take a **very long time** to run on the complete data set.

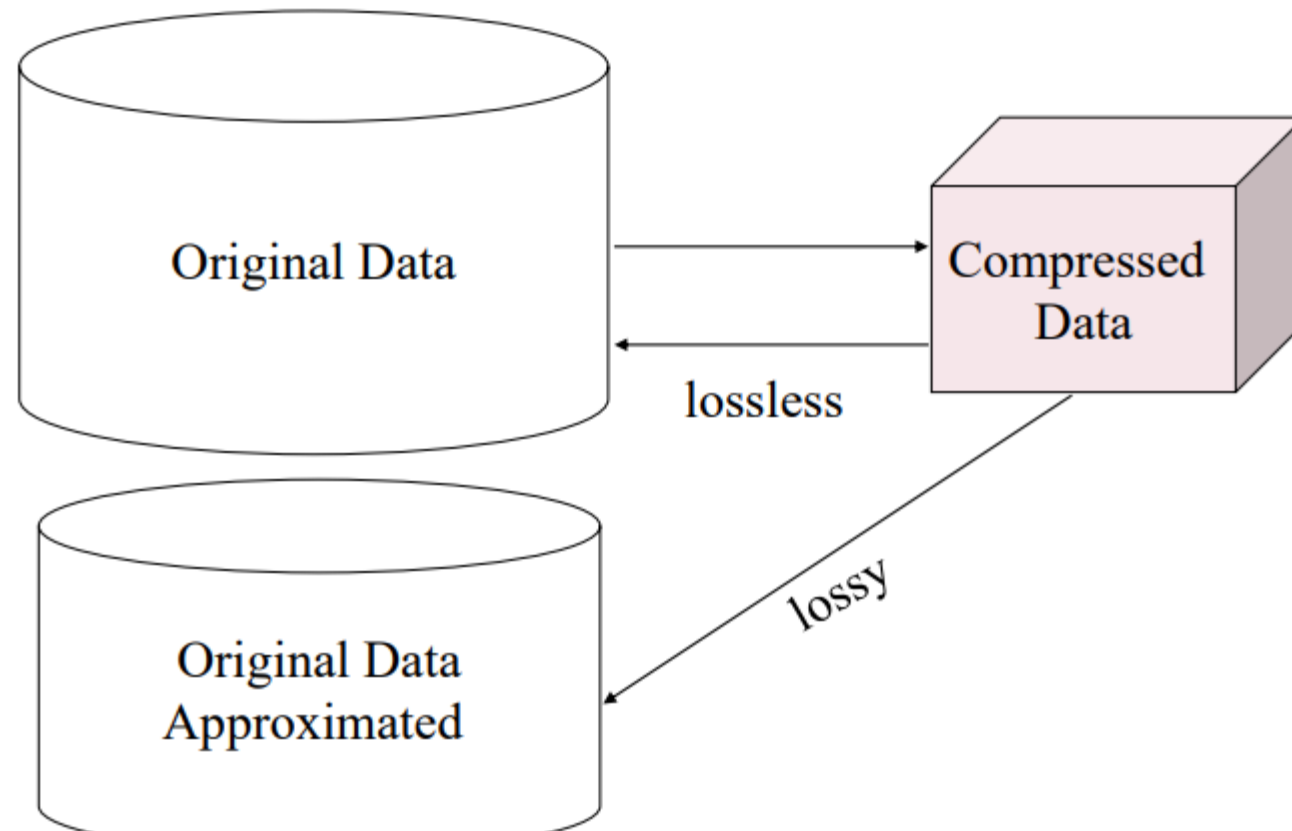
-
- Data reduction strategies:
 - Data Cube Aggregation
 - Dimensionality reduction
 - Wavelet transforms
 - Principal Components Analysis (PCA)
 - Feature subset selection
 - Feature creation
 - Numerosity reduction (some simply call it: Data Reduction)
 - Regression and Log-Linear Models
 - Histograms, clustering, sampling
 - Data cube aggregation – Data compression
 - Data compression

• **Data Cube Aggregation:**

- The lowest level of a data cube is the aggregated data for an individual entity of interest.
- To do this, use the smallest representation that is sufficient to address the given task.
- In other words, we reduce the data to its more meaningful size and structure for the task at hand.
- Queries regarding aggregated information should be answered using data cube, when possible.

- **Data Compression:**

- String compression
 - There are extensive theories and well-tuned algorithms.
 - Typically lossless, but only limited manipulation is possible without expansion.
- Audio/video compression
 - Typically lossy compression, with progressive refinement.
 - Sometimes small fragments of signal can be reconstructed without reconstructing the whole.
- Time sequence is not audio
 - Typically short and vary slowly with time.
- Dimensionality and numerosity reduction may also be considered as forms of data compression



Data Discretization



- We are often dealing with data that are collected from processes that are continuous, such as temperature, ambient light, and a company's stock price.
- But sometimes we need to convert these continuous values into more manageable parts.

- **A Type of Data Discretization:**

- Binning: first sort data and partition into (equal-frequency) bins and then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.

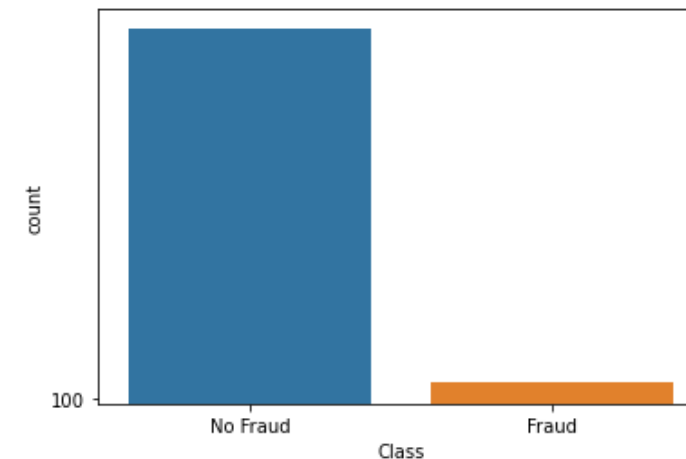
price
13495
16500
18920
41315
5151
6295
...



price	price-binned
13495	Low
16500	Low
18920	Medium
41315	High
5151	Low
6295	Low
...	...

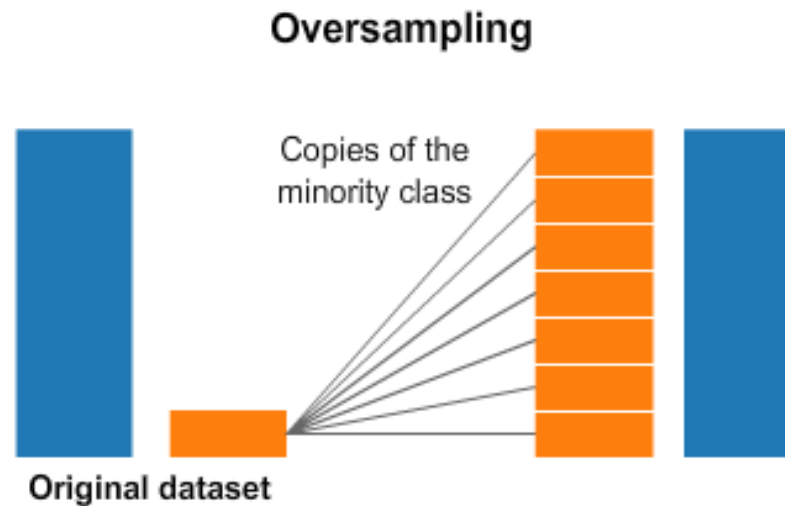
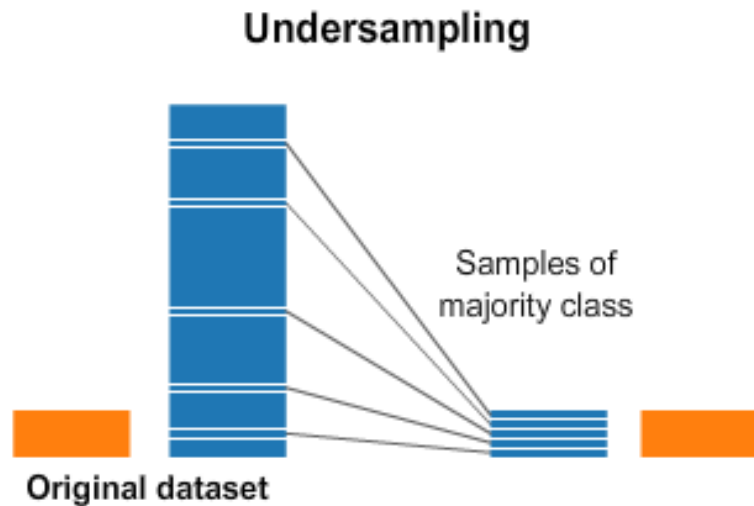
Data Balancing

- Imbalanced data might lead to biased data analysis
- In this case, the class distribution needs to be adjusted
- Examples of imbalanced data:
 - Fraud detection
 - Transportation failure
 - Medical disease



Data Balancing

- How to balance the data?



- We'll cover more detail in topic 12.

Summary

- Data collection and pre-processing are key steps in data science
- Garbage-In, Garbage-Out
- Data in the real world is often “**dirty**”; means that they need to be “**cleaned up**” before they can be used for a desired purpose.
- Data preparation includes:
 - data cleaning: missing/noisy value, outliers
 - data integration: entity identification problem, remove redundancies, detect inconsistencies
 - data transformation: smoothing, normalization, etc
 - data selection
 - data reduction: dimensionality reduction, compression, etc
 - data discretization
 - data balancing

A Hands-on Example

Table 2.3 Excessive wine consumption and mortality data.

#	Country	Alcohol	Deaths	Heart	Liver
1	Australia	2.5	785	211	15.30000019
2	Austria	3.000000095	863	167	45.59999847
3	Belg. and Lux.	2.900000095	883	131	20.70000076
4	Canada	2.400000095	793	NA	16.39999962
5	Denmark	2.900000095	971	220	23.89999962
6	Finland	0.800000012	970	297	19
7	France	9.100000381	751	11	37.90000153
8	Iceland	-0.800000012	743	211	11.19999981
9	Ireland	0.699999988	1000	300	6.5
10	Israel	0.600000024	-834	183	13.69999981

-
- The dataset consists of the following attributes:
 - Name of the country from which sample obtained
 - Alcohol consumption measured as litres of wine, per capita
 - Number of deaths from alcohol consumption, per 100,000 people
 - Number of heart disease deaths, per 100,000 people
 - Number of deaths from liver diseases, also per 100,000 people
 - Could you detect any noise/missing in the data presented in Table 2.3?
 - How could we handle such a noise?
 - What kind of data preparation steps can we use to remove the noise or to fill missing data?

#	Country	Alcohol	Deaths	Heart	Liver
1	Australia	2.5	785	211	15.30000019
2	Austria	3.000000095	863	167	45.59999847
3	Belg. and Lux.	2.900000095	883	131	20.70000076
4	Canada	2.400000095	793	NA	16.39999962
5	Denmark	2.900000095	971	220	23.89999962
6	Finland	0.800000012	970	297	19
7	France	9.100000381	751	11	37.90000153
8	Iceland	-0.800000012	743	211	11.19999981
9	Ireland	0.699999988	1000	300	6.5

- We note that the wine consumption value for Iceland per capita is -0.800000012 (red box). However, wine consumption values per capita cannot be negative. Therefore, it must be a faulty entry and we should change the alcohol consumption for Iceland to 0.800000012 .

3	Belg. and Lux.	2.900000095	883	131	20.70000076
4	Canada	2.400000095	793	NA	16.39999962
5	Denmark	2.900000095	971	220	23.89999962
6	Finland	0.800000012	970	297	19
7	France	9.100000381	751	11	37.90000153
8	Iceland	-0.800000012	743	211	11.19999981
9	Ireland	0.699999988	1000	300	6.5
10	Israel	0.600000024	-834	183	13.69999981

- Using the same logic, the number of deaths for Israel should be converted from -834 to 834.
- As we can see in the dataset, we have missing values (represented by NA – purple box) of the number of cases of heart disease for Canada and number of cases of heart and liver disease for Spain.
- So, for Canada, we can fill the missing data (NA) using the average number of heart diseases over all countries (185).

Table 2.4 Wine consumption vs. mortality data after data cleaning.

#	Country	Alcohol	Deaths	Heart	Liver
1	Australia	2.5	785	211	15.30000019
2	Austria	3.000000095	863	167	45.59999847
3	Belg. and Lux.	2.900000095	883	131	20.70000076
4	Canada	2.400000095	793	185	16.39999962
5	Denmark	2.900000095	971	220	23.89999962
6	Finland	0.800000012	970	297	19
7	France	9.100000381	751	11	37.90000153
8	Iceland	0.800000012	743	211	11.19999981
9	Ireland	0.699999988	1000	300	6.5
10	Israel	0.600000024	834	183	13.69999981
11	Italy	27.900000095	775	107	42.20000076
12	Japan	1.5	680	36	23.20000076
13	Netherlands	1.799999952	773	167	9.199999809
14	New Zealand	1.899999976	916	266	7.699999809
15	Norway	0.0800000012	806	227	12.19999981
16	Spain	6.5	724	185	20.27

Table 2.5 Data about alcohol consumption and health from various States in India.

#	Name of the State	Alcohol consumption	Heart disease	Fatal alcohol-related accidents
1	Andaman and Nicobar Islands	1.73	20,312	2201
2	Andhra Pradesh	2.05	16,723	29,700
3	Arunachal Pradesh	1.98	13,109	11,251
4	Assam	0.91	8532	211,250
5	Bihar	3.21	12,372	375,000
6	Chhattisgarh	2.03	28,501	183,207
7	Goa	5.79	19,932	307,291

- Now let us assume we have another dataset (fictitious) collected from a different source, which is about alcohol consumption and number of related fatalities across various states of India, as shown in Table 2.5.

-
- Here is what the dataset contains:
 - Name of the State.
 - Liters of alcohol consumed per capita.
 - Number of fatal heart diseases, measured per 1,000,000 people.
 - Number of fatal accidents related to alcohol per 1,000,000 people
 - Could we find any benefit from the data in Table 2.5?
 - How can we process the data in Table 2.5 such they add some benefits to the data in Table 2.3?

-
- Now we can use this dataset to integrate the attributes for India into our original dataset.
 - To do this, we calculate the total alcohol consumption for the country of India as an average of alcohol consumption for all the states, which is 2.95.
 - Similarly, we can calculate the fatal heart diseases per 100,000 people for India as 171 (approximated to the nearest integer value).
 - Since we do not have any source for the number of total deaths or the number of fatal liver diseases for India, we are going to handle these the same way we previously addressed any missing values. The resultant dataset is shown in Table 2.6.

Table 2.6 Wine consumption and associated mortality after data integration.

#	Country	Alcohol	Deaths	Heart	Liver
1	Australia	2.5	785	211	15.30000019
2	Austria	3.000000095	863	167	45.59999847
3	Belg. and Lux.	2.900000095	883	131	20.70000076
4	Canada	2.400000095	793	185	16.39999962
5	Denmark	2.900000095	971	220	23.89999962
6	Finland	0.800000012	970	297	19
7	France	9.100000381	751	11	37.90000153
8	Iceland	0.800000012	743	211	11.19999981
9	Ireland	0.699999988	1000	300	6.5
10	Israel	0.600000024	834	183	13.69999981
11	Italy	27.900000095	775	107	42.20000076
12	Japan	1.5	680	36	23.20000076
13	Netherlands	1.799999952	773	167	9.199999809
14	New Zealand	1.899999976	916	266	7.699999809
15	Norway	0.0800000012	806	227	12.19999981
16	Spain	6.5	724	185	20.27
17	Sweden	1.600000024	743	207	11.19999981
18	Switzerland	5.800000191	693	115	20.29999924
19	UK	1.299999952	941	285	10.30000019
20	US	1.200000048	926	199	22.10000038
21	West Germany	2.700000048	861	172	36.70000076
22	India	2.950000000	750	171	20.27

-
- **Data Transformation: smoothing**, removing noise from data, summarization, generalization, and normalization.
 - As we can see, in our data the wine consumption per capita for Italy is unusually high, whereas the same for Norway is unusually low. So, chances are these are outliers.
 - In this case we will replace the value of wine consumption for Italy with 7.900000095.
 - Similarly, for Norway we will use the value of 0.800000012 in place of 0.0800000012.
 - We are treating both of these potential errors as "equipment error" or "entry error," which resulted in an extra digit for both of these countries (extra"2" in front for Italy and extra"0" after the decimal point for Norway).

Table 2.7 Wine consumption and associated mortality dataset after data transformation.

#	Country	Alcohol	Deaths	Heart	Liver
1	Australia	2.5	785	211	15.30000019
2	Austria	3.000000095	863	167	45.59999847
3	Belg. and Lux.	2.900000095	883	131	20.70000076
4	Canada	2.400000095	793	185	16.39999962
5	Denmark	2.900000095	971	220	23.89999962
6	Finland	0.800000012	970	297	19
7	France	9.100000381	751	11	37.90000153
8	Iceland	0.800000012	743	211	11.19999981
9	Ireland	0.699999988	1000	300	6.5
10	Israel	0.600000024	834	183	13.69999981
11	Italy	7.900000095	775	107	42.20000076
12	Japan	1.5	680	36	23.20000076
13	Netherlands	1.799999952	773	167	9.199999809
14	New Zealand	1.899999976	916	266	7.699999809
15	Norway	0.800000012	806	227	12.19999981

-
- **Data Reduction:** aims at producing a reduced representation of the dataset that can be used to obtain the same or similar analytical results.
 - For our example, the sample is relatively small, with only 22 rows.
 - Now imagine that we have values for all 196 countries in the world, and the geospatial values, for which the attribute values are available, are stated.
 - In that case, the number of rows is large, and, depending on the limited processing and storage capacity you have at your disposal, it may make more sense to round up the alcohol consumption per capita to two decimal places.

Table 2.8 Wine consumption and associated mortality dataset after data reduction.

#	Country	Alcohol	Deaths	Heart	Liver
1	Australia	2.50	785	211	15.3
2	Austria	3.00	863	167	45.6
3	Belg. and Lux.	2.90	883	131	20.7
4	Canada	2.40	793	185	16.4
5	Denmark	2.90	971	220	23.9
6	Finland	0.80	970	297	19.0
7	France	9.10	751	11	37.9
8	Iceland	0.80	743	211	11.2
9	Ireland	0.70	1000	300	6.5
10	Israel	0.60	834	183	13.7
11	Italy	7.90	775	107	42.2
12	Japan	1.50	680	36	23.2

-
- **Data Discretization:** depending on the model you want to build, you may have to discretize the attribute values into binary or categorical types.
 - For example, you may want to discretize the wine consumption per capita into four categories – less than or equal to 1.00 per capita (represented by 0), more than 1.00 but less than or equal to 2.00 per capita (1), more than 2.00 but less than or equal to 5.00 per capita (2), and more than 5.00 per capita (3). The resultant dataset should look like that shown in Table 2.9

Table 2.9 Wine consumption and mortality dataset at the end of pre-processing.

#	Country	Alcohol	Deaths	Heart	Liver
1	Australia	2	785	211	15.3
2	Austria	2	863	167	45.6
3	Belg. and Lux.	2	883	131	20.7
4	Canada	2	793	185	16.4
5	Denmark	2	971	220	23.9
6	Finland	0	970	297	19.0
7	France	3	751	11	37.9
8	Iceland	0	743	211	11.2
9	Ireland	0	1000	300	6.5
10	Israel	0	834	183	13.7
11	Italy	3	775	107	42.2
12	Japan	1	680	36	23.2
13	Netherlands	1	773	167	9.2
14	New Zealand	1	916	266	7.7
15	Norway	0	806	227	12.2
16	Spain	3	724	185	20.3
17	Sweden	1	743	207	11.2
18	Switzerland	3	693	115	20.3
19	UK	1	941	285	10.3
20	US	1	926	199	22.1
21	West Germany	2	861	172	36.7
22	India	2	750	171	20.3

References & Credits

- A Hands on Tutorial to Data Science, Chirag Shah, A, 2020
 - Principles of Data Science, Sinan Ozdemir, December 2016
 - IBM Data Science Training Materials and cognitiveclass.ai
 - David Forsyth, Probability and Statistics for Computer Science, Springer International Publishing AG 2018
 - Siti Aminah & Dhimas Arief Darmawan, Data Preparations, Data Sains Semester Genap 2020/2021
 - Fariz Darari, Data Collection & Preparation, KASDD Semester Gasal 2021/2022
-
- Gambar dan tangkapan layar hanya untuk kebutuhan penjelasan
 - Hak cipta tetap ada pada pemilik aslinya.

Wish You Success

