FIT5196 DATA WRANGLING

Week 1

Introduction to Data Wrangling

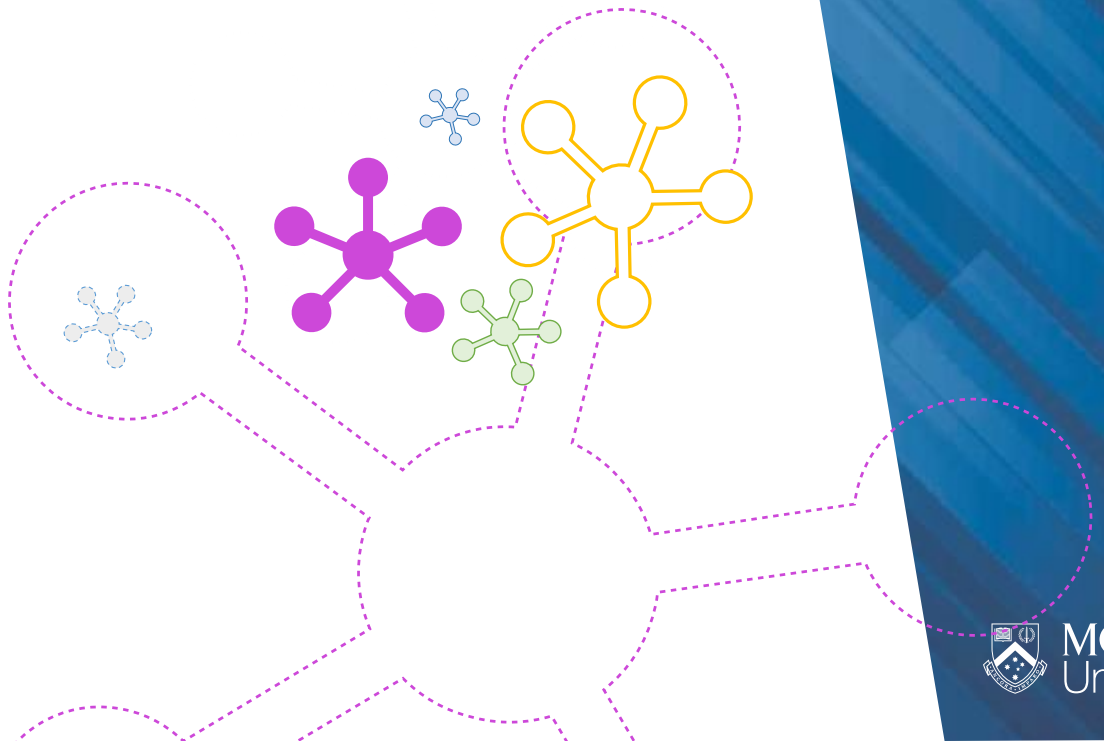
By Jackie Rong

Faculty of Information Technology

Monash University

Outline

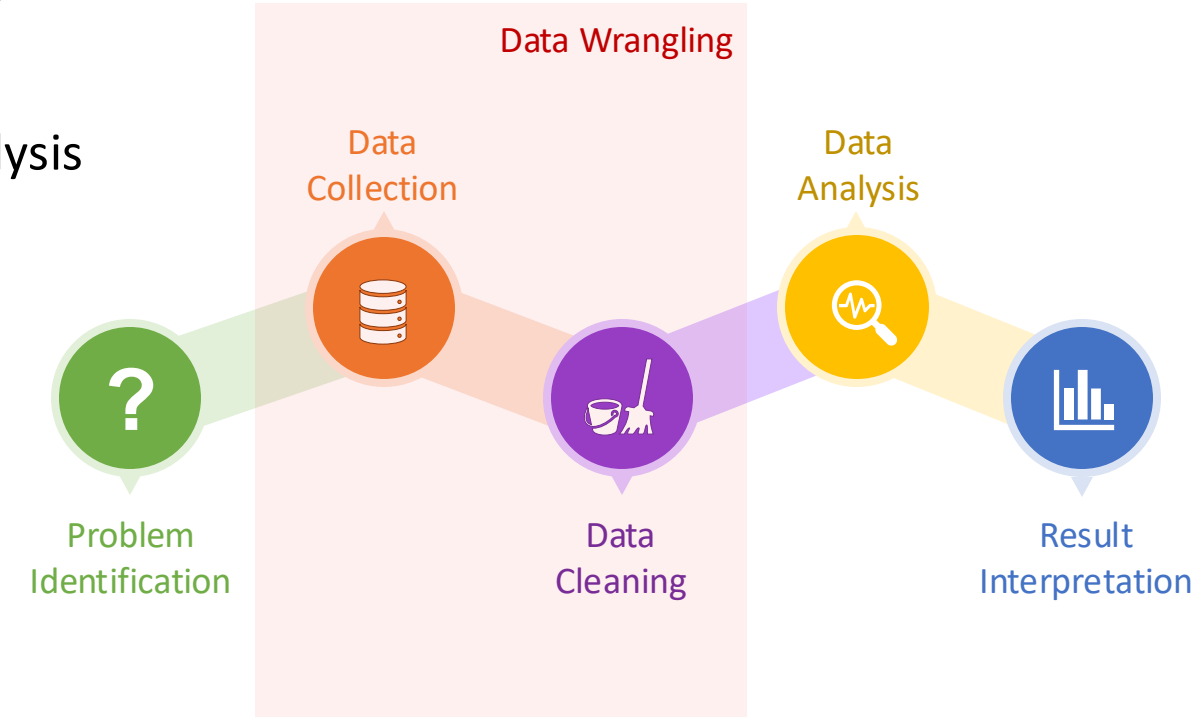
- What is Data Wrangling?
- Why need to do Data Wrangling?
- Challenges in Data Wrangling
- Data Wrangling Process & Tasks
- Programming Language & Environments



MONASH University

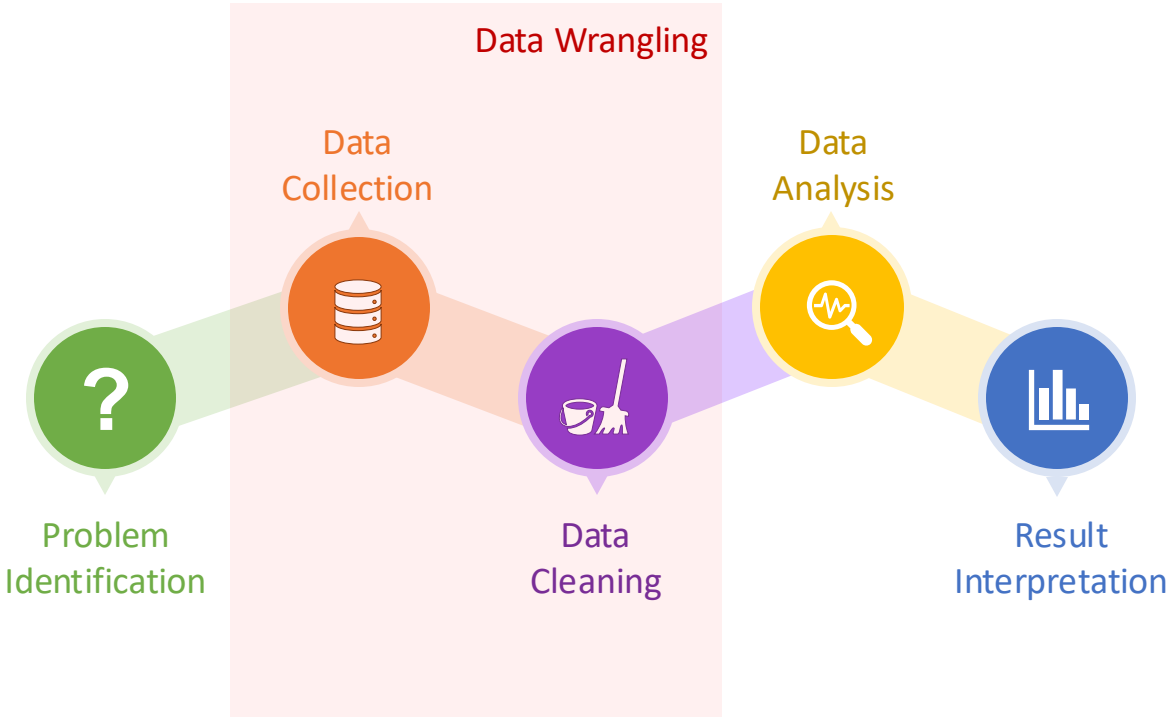
Data Wrangling

- **Data wrangling** is a critical step in the data analysis process.
- **Data wrangling** is a preparatory step for data analysis.
- **Data wrangling** is essential for ensuring that data analysis leads to accurate and actionable insights.
- **Data wrangling** is the process of making data useful.



Data Wrangling

- **Data Wrangling** is the process of acquiring, cleaning, structuring, and enriching raw data into a format that is directly usable for analysis.



Why need to do Data Wrangling?

- **Data wrangling** is essential for ensuring that data analysis leads to accurate and actionable insights.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-famil	White	Male	2174	0	40	United-States	<=50K
2	50	Self-emp-no	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
3	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-famil	White	Male	0	0	40	United-States	<=50K
4	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
5	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
6	37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
7	49	Private	160187	9th	5	Married-spouse-abser	Other-service	Not-in-famil	Black	Female	0	0	16	Jamaica	<=50K
8	52	Self-emp-no	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K
9	31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-famil	White	Female	14084	0	50	United-States	>50K
10	42	Private	159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178	0	40	United-States	>50K
11	37	Private	280464	Some-colleg	10	Married-civ-spouse	Exec-managerial	Husband	Black	Male	0	0	80	United-States	>50K
12	30	State-gov	141297	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Isl	Male	0	0	40	India	>50K
13	23	Private	122272	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	0	0	30	United-States	<=50K
14	32	Private	205019	Assoc-acdm	12	Never-married	Sales	Not-in-famil	Black	Male	0	0	50	United-States	<=50K
15	40	Private	121772	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Isl	Male	0	0	40	?	>50K
16	34	Private	245487	7th-8th	4	Married-civ-spouse	Transport-moving	Husband	Amer-Indian	Male	0	0	45	Mexico	<=50K
17	25	Self-emp-no	176756	HS-grad	9	Never-married	Farming-fishing	Own-child	White	Male	0	0	35	United-States	<=50K
18	32	Private	186824	HS-grad	9	Never-married	Machine-op-inspc	Unmarried	White	Male	0	0	40	United-States	<=50K
19	38	Private	28887	11th	7	Married-civ-spouse	Sales	Husband	White	Male	0	0	50	United-States	<=50K
20	43	Self-emp-no	292175	Masters	14	Divorced	Exec-managerial	Unmarried	White	Female	0	0	45	United-States	>50K
21	40	Private	193524	Doctorate	16	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	60	United-States	>50K
22	54	Private	302146	HS-grad	9	Separated	Other-service	Unmarried	Black	Female	0	0	20	United-States	<=50K
23	35	Federal-gov	76845	9th	5	Married-civ-spouse	Farming-fishing	Husband	Black	Male	0	0	40	United-States	<=50K
24	43	Private	117037	11th	7	Married-civ-spouse	Transport-moving	Husband	White	Male	0	2042	40	United-States	<=50K
25	59	Private	109015	HS-grad	9	Divorced	Tech-support	Unmarried	White	Female	0	0	40	United-States	<=50K

The ”census income" data set from [UCI machine learning data repository]

Why need to do Data Wrangling?

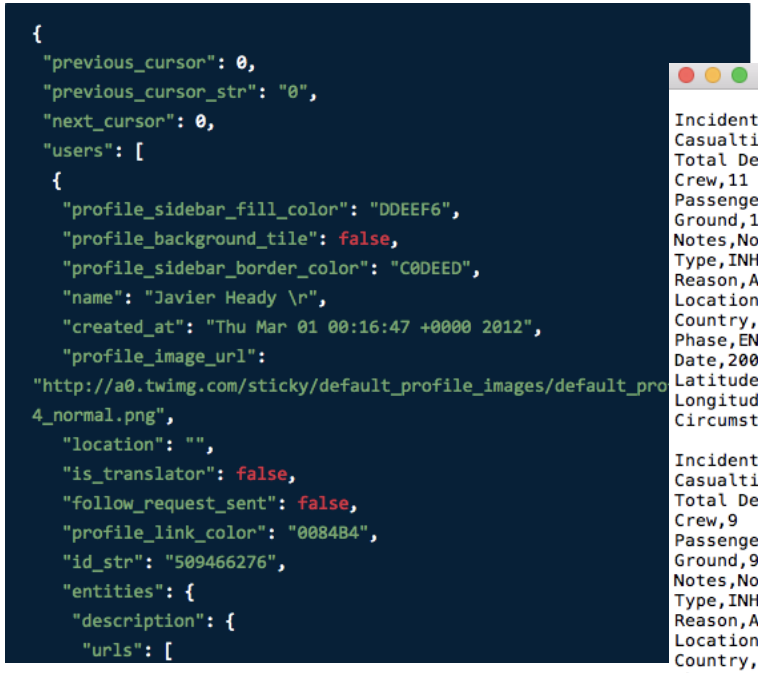
- **Data wrangling** is essential for ensuring that data analysis leads to accurate and actionable insights.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	b	30.83	0	u	g	w	v	1.25	t	t	1	f	g	202	0	+
2	a	58.67	4.46	u	g	q	h	3.04	t	t	6	f	g	43	560	+
3	a	24.5	0.5	u	g	q	h	1.5	t	f	0	f	g	280	824	+
4	b	27.83	1.54	u	g	w	v	3.75	t	t	5	t	g	100	3	+
5	b	20.17	5.625	u	g	w	v	1.71	t	f	0	f	s	120	0	+
6	b	32.08	4	u	g	m	v	2.5	t	f	0	t	g	360	0	+
7	b	33.17	1.04	u	g	r	h	6.5	t	f	0	t	g	164	31285	+
8	a	22.92	11.585	u	g	cc	v	0.04	t	f	0	f	g	80	1349	+
9	b	54.42	0.5	y	p	k	h	3.96	t	f	0	f	g	180	314	+
10	b	42.5	4.915	y	p	w	v	3.165	t	f	0	t	g	52	1442	+
11	b	22.08	0.83	u	g	c	h	2.165	f	f	0	t	g	128	0	+
12	b	29.92	1.835	u	g	c	h	4.335	t	f	0	f	g	260	200	+
13	a	38.25	6	u	g	k	v	1	t	f	0	t	g	0	0	+
14	b	48.08	6.04	u	g	k	v	0.04	f	f	0	f	g	0	2690	+
15	a	45.83	10.5	u	g	q	v	5	t	t	7	t	g	0	0	+
16	b	36.67	4.415	y	p	k	v	0.25	t	t	10	t	g	320	0	+
17	b	28.25	0.875	u	g	m	v	0.96	t	t	3	t	g	396	0	+
18	a	23.25	5.875	u	g	q	v	3.17	t	t	10	f	g	120	245	+
19	b	21.83	0.25	u	g	d	h	0.665	t	f	0	t	g	0	0	+
20	a	19.17	8.585	u	g	cc	h	0.75	t	t	7	f	g	96	0	+
21	b	25	11.25	u	g	c	v	2.5	t	t	17	f	g	200	1208	+
22	b	23.25	1	u	g	c	v	0.835	t	f	0	f	s	300	0	+
23	a	47.75	8	u	g	c	v	7.875	t	t	6	t	g	0	1260	+
24	a	27.42	14.5	u	g	x	h	3.085	t	t	1	f	g	120	11	+
25	a	41.17	6.5	u	g	q	v	0.5	t	t	3	t	g	145	0	+

The "credit approval" data set from UCI machine learning data repository

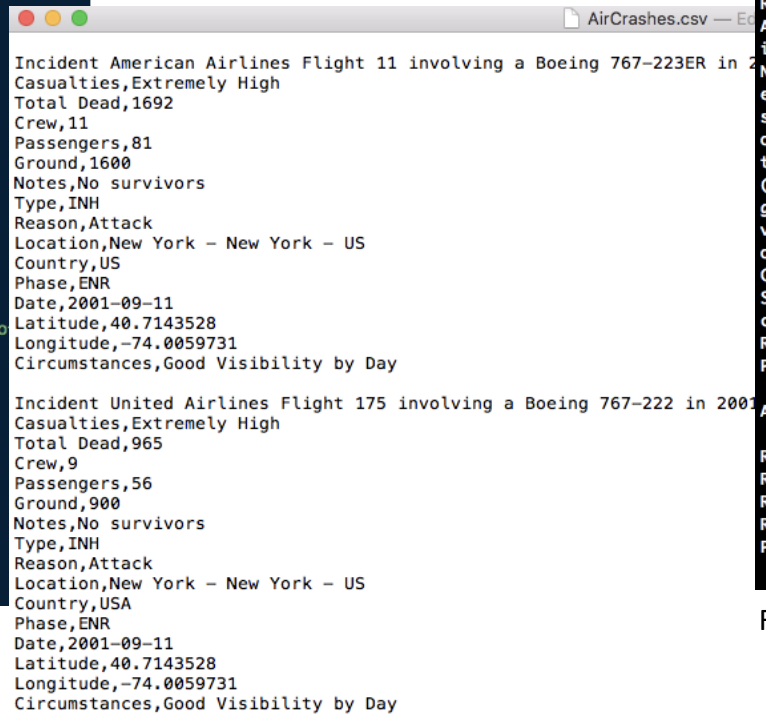
Why need to do Data Wrangling?

- What does raw data really look like?



```
{
 "previous_cursor": 0,
 "previous_cursor_str": "0",
 "next_cursor": 0,
 "users": [
  {
   "profile_sidebar_fill_color": "DDEEF6",
   "profile_background_tile": false,
   "profile_sidebar_border_color": "C0DEED",
   "name": "Javier Heady \r",
   "created_at": "Thu Mar 01 00:16:47 +0000 2012",
   "profile_image_url":
"http://a0.twimg.com/sticky/default_profile_images/default_pro
4_normal.png",
   "location": "",
   "is_translator": false,
   "follow_request_sent": false,
   "profile_link_color": "0084B4",
   "id_str": "509466276",
   "entities": {
    "description": {
     "urls": [
```

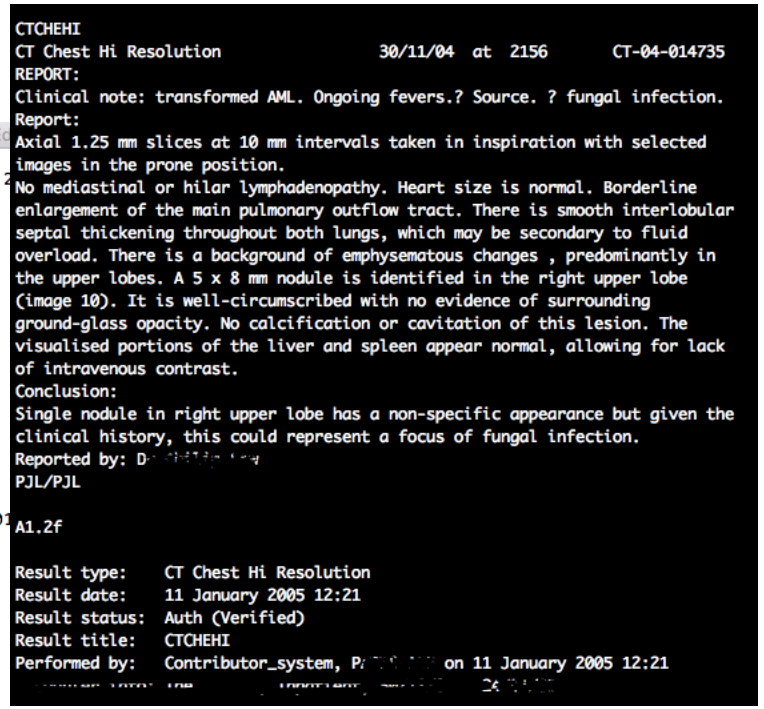
Posts extracted from Twitter
https://dev.twitter.com/rest/reference/get/blocks/list



AirCrashes.csv

```
Incident American Airlines Flight 11 involving a Boeing 767-223ER in 2
Casualties,Extremely High
Total Dead,1692
Crew,11
Passengers,81
Ground,1600
Notes,No survivors
Type,INH
Reason,Attack
Location,New York - New York - US
Country,US
Phase,ENR
Date,2001-09-11
Latitude,40.7143528
Longitude,-74.0059731
Circumstances,Good Visibility by Day

Incident United Airlines Flight 175 involving a Boeing 767-222 in 2001
Casualties,Extremely High
Total Dead,965
Crew,9
Passengers,56
Ground,900
Notes,No survivors
Type,INH
Reason,Attack
Location,New York - New York - US
Country,USA
Phase,ENR
Date,2001-09-11
Latitude,40.7143528
Longitude,-74.0059731
Circumstances,Good Visibility by Day
```

Airline Crash dataset from Wikipedia
https://en.wikipedia.org/wiki/List_of_accidents_and_incidents_involving_commercial_aircraft#2001



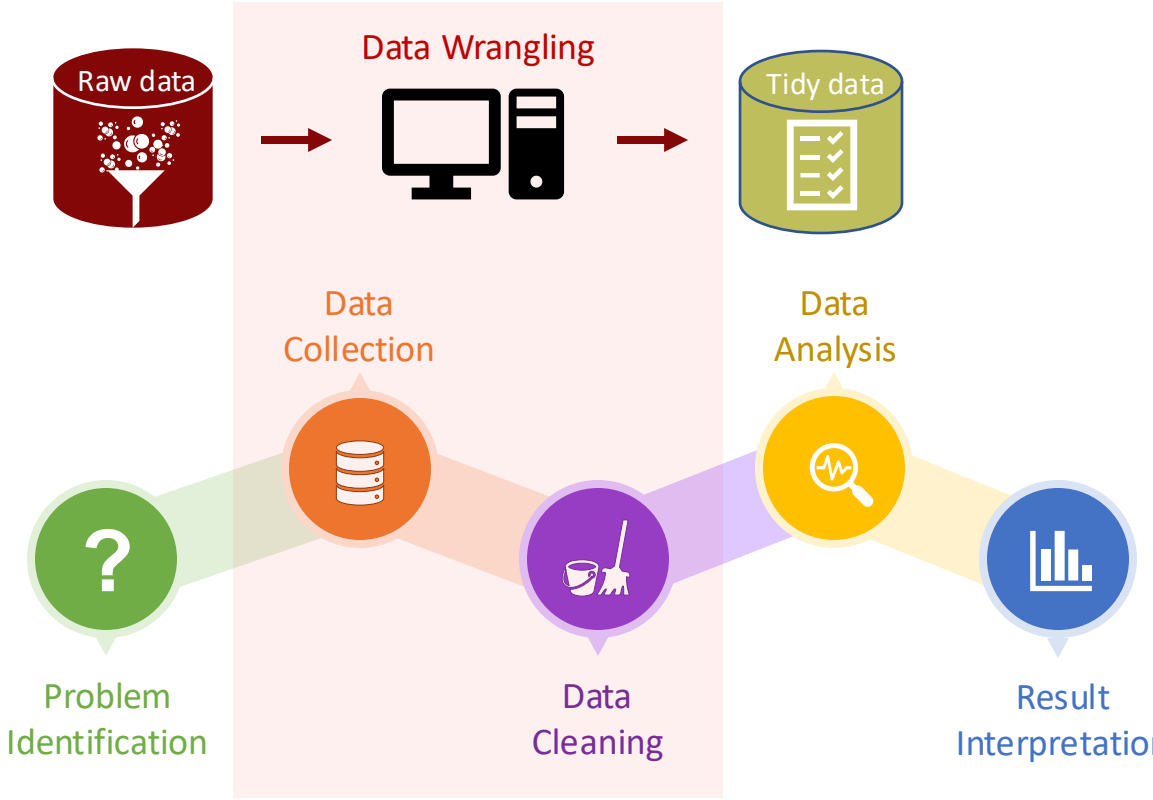
```
CTCHEHI
CT Chest Hi Resolution                30/11/04  at  2156       CT-04-014735
REPORT:
Clinical note: transformed AML. Ongoing fevers.? Source. ? fungal infection.
Report:
Axial 1.25 mm slices at 10 mm intervals taken in inspiration with selected
images in the prone position.
No mediastinal or hilar lymphadenopathy. Heart size is normal. Borderline
enlargement of the main pulmonary outflow tract. There is smooth interlobular
septal thickening throughout both lungs, which may be secondary to fluid
overload. There is a background of emphysematous changes , predominantly in
the upper lobes. A 5 x 8 mm nodule is identified in the right upper lobe
(image 10). It is well-circumscribed with no evidence of surrounding
ground-glass opacity. No calcification or cavitation of this lesion. The
visualised portions of the liver and spleen appear normal, allowing for lack
of intravenous contrast.
Conclusion:
Single nodule in right upper lobe has a non-specific appearance but given the
clinical history, this could represent a focus of fungal infection.
Reported by: D
PJL/PJL

A1.2f

Result type:    CT Chest Hi Resolution
Result date:    11 January 2005 12:21
Result status:  Auth (Verified)
Result title:   CTCHEHI
Performed by:   Contributor_system, P  on 11 January 2005 12:21
```

Fungal disease CT report

Goals of Data Wrangling

- The goals of data wrangling are multifaceted, aiming to simplify data analysis and maximize the value extracted from the data.
  - Improving Data Quality
  - Data Formatting and Standardization
  - Simplifying Access to Data
  - Enriching Data
  - Reducing Data Complexity
  - Facilitating Data Integration
  - Increasing Analytical Efficiency
  - Supporting Decision Making

Data + Wrangling + Analysis
= Data Product (Knowledge)



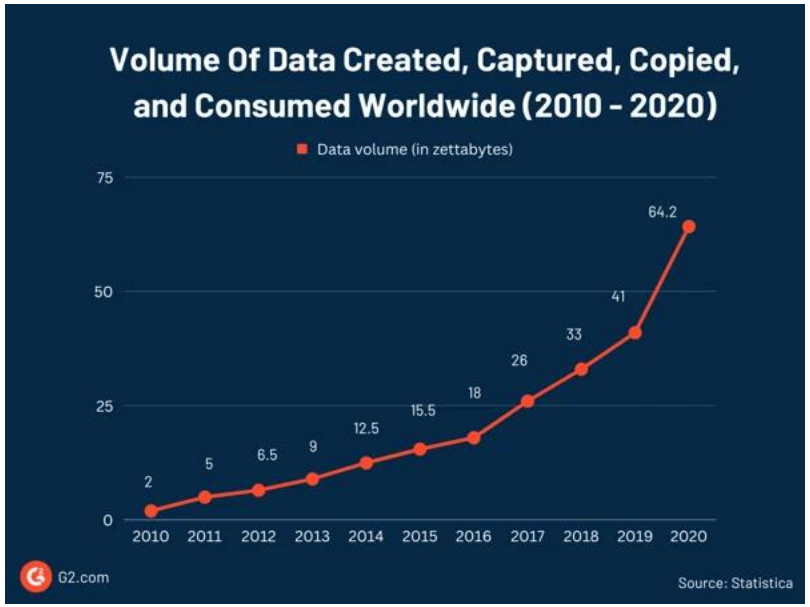
Challenges in Data Wrangling

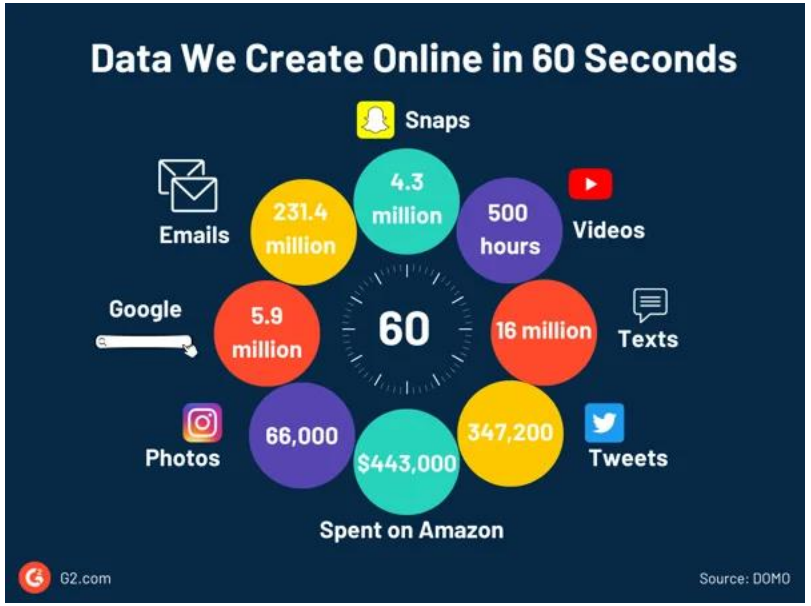
- Challenges arise from
  - the nature of the data itself,
  - the complexity of data sources, and
  - the goals of the data analysis projects.

Challenges in Data Wrangling

- Challenges arise from the nature of the data itself, the complexity of data sources, and the goals of the data analysis projects.
  - **Volume of Data & Scalability**

As in 18 zeroes, as in more than 2,500,000,000,000,000,000 bytes (or 2.5 quintillions) of data are created *each day*.





77+ Surreal Big Data Statistics To Map Growth in 2024, https://www.g2.com/articles/big-data-statistics

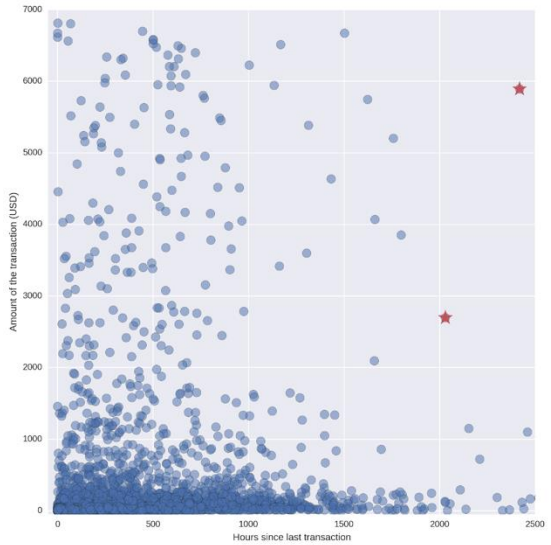
Challenges in Data Wrangling

- Challenges arise from the nature of the data itself, the complexity of data sources, and the goals of the data analysis projects.
  - Volume of Data & Scalability
  - **Data Quality Issues**

Mr. Mark John	33	21-08-1985	180	M	0433010010	Mel,VIC
Mr. Chris, Peter	34	21-Sep-1982	?	Fale	0000000000	Syd, NSW
Ethan Steedman	36	01/01/82	17o	M	0388886789	Mel,VIC

```
32,1,1,95,0,?,0,127,0,.7,1,?,?,1
34,1,4,115,0,?,?,154,0,.2,1,?,?,1
35,1,4,?,0,?,0,130,1,?,?,?,7,3
36,1,4,110,0,?,0,125,1,1,2,?,6,1
38,0,4,105,0,?,0,166,0,2.8,1,?,?,2
38,0,4,110,0,0,0,156,0,0,2,?,3,1
38,1,3,100,0,?,0,179,0,-1.1,1,?,?,0
38,1,3,115,0,0,0,128,1,0,2,?,7,1
38,1,4,135,0,?,0,150,0,0,?,?,3,2
38,1,4,150,0,?,0,120,1,?,?,?,3,1
40,1,4,95,0,?,1,144,0,0,1,?,?,2
```

The Switzerland heart disease data set from
UCI machine learning data repository



Where the data problems come from?

- Manual entry errors
- Malfunction of measurement devices
- Data sources follow different conventions, formats, or data models

Challenges in Data Wrangling

- Challenges arise from the nature of the data itself, the complexity of data sources, and the goals of the data analysis projects.
  - Volume of Data & Scalability
  - Data Quality Issues
  - **Data from Diverse Sources**

Other formats: CSV, Excel, PDF, PNG, JPGE, ……


```
{
  "meta" : {
    "view" : {
      "id" : "tdvh-n9dv",
      "name" : "Melbourne bike share",
      "attribution" : "City of Melbourne, Australia",
      "averageRating" : 0,
      "category" : "Transport & Movement",
      "createdAt" : 1428898164,
      "description" : "Melbourne Bike Share is a joint RACV/Victoria
      "displayType" : "table",
      "downloadCount" : 1314,
      "indexUpdatedAt" : 1453946128,
      "licenseId" : "CC_30_BY_AUS",
      "newBackend" : false,
      "numberOfComments" : 0,
      "oid" : 11003321,
      "publicationAppendEnabled" : true,
      "publicationDate" : 1429672791,
      "publicationGroup" : 2657856,
```


JavaScript Object Notation (JSON):


```
<response>
    <row>
        <row _id="155" _uuid="7C09387D-9E6C-4B42-9041-9A98B88F54
            <id>2</id>
            <featurename>Harbour Town - Docklands Dve - Dockland
            <terminalname>60000</terminalname>
            <nbbikes>9</nbbikes>
            <nbemptydoc>14</nbemptydoc>
            <uploaddate>1453986006</uploaddate>
            <coordinates human_address="{&amp;quot;address&amp;quot;:&amp;qu
                        latitude="-37.814022" longitude="144.93
        </row>
        <row _id="156" _uuid="52739A59-E034-436B-A613-E7A5F62448
            <id>4</id>
            <featurename>Federation Square - Flinders St / Swans
            <terminalname>60001</terminalname>
            <nbbikes>15</nbbikes>
            <nbemptydoc>7</nbemptydoc>
            <uploaddate>1453986006</uploaddate>
            <coordinates human_address="{&amp;quot;address&amp;quot;:&amp;qu
                        latitude="-37.817523" longitude="144.96
```


Extensible Markup Language (XML)

Challenges in Data Wrangling

- Challenges arise from the nature of the data itself, the complexity of data sources, and the goals of the data analysis projects.
  - Volume of Data & Scalability
  - Data Quality Issues
  - Data from Diverse Sources
  - **Complexity of Data Structures**

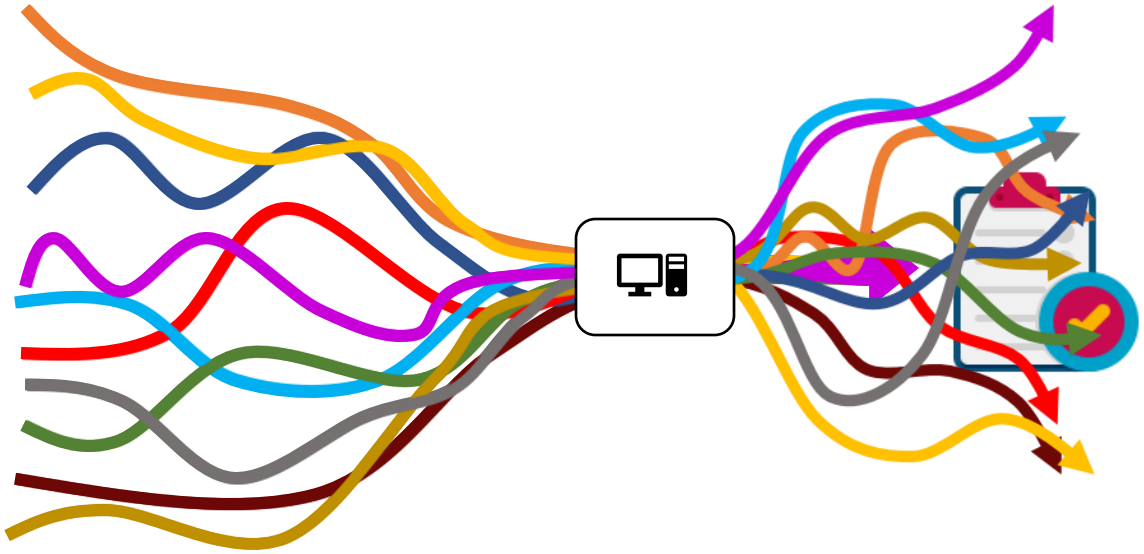
UNSTRUCTURED DATA

STRUCTURED DATA

https://blog.kensho.com/structured-vs-unstructured-data-what-you-need-to-know-f1e7ce61cd1e

Challenges in Data Wrangling

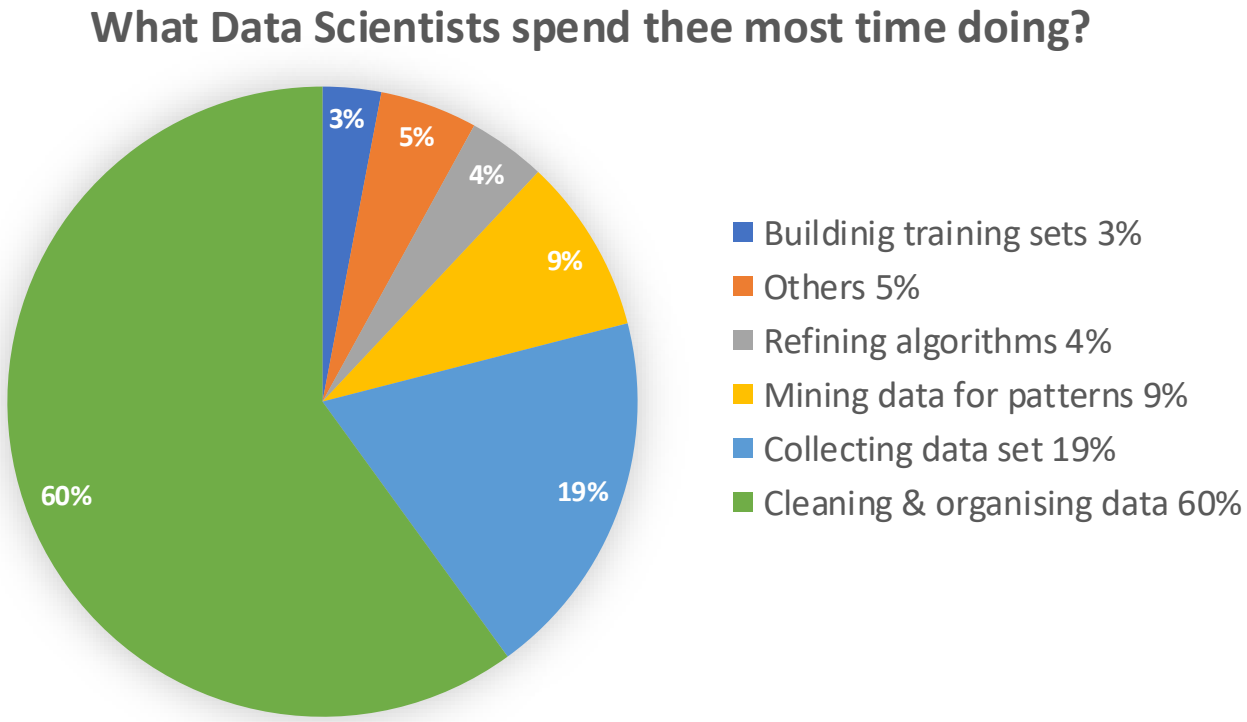
- Challenges arise from the nature of the data itself, the complexity of data sources, and the goals of the data analysis projects.
  - Volume of Data & Scalability
  - Data Quality Issues
  - Data from Diverse Sources
  - Complexity of Data Structures
  - **Lack of Standardization & Interpretability**



icons created by Flat Icons – Flaticon, https://www.flaticon.com/free-icons/iso

Challenges in Data Wrangling

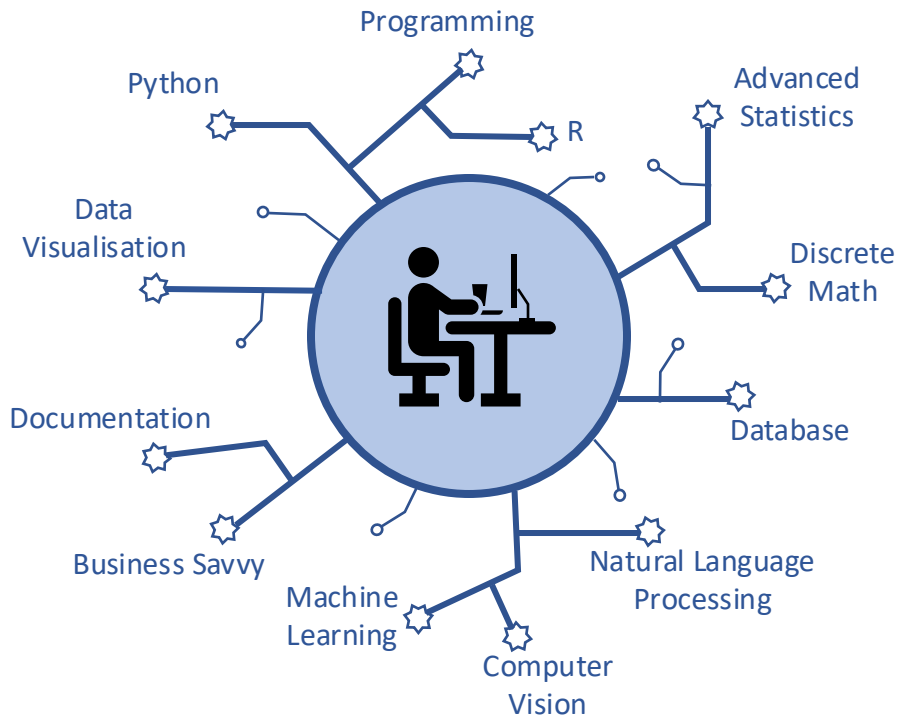
- Challenges arise from the nature of the data itself, the complexity of data sources, and the goals of the data analysis projects.
  - Volume of Data & Scalability
  - Data Quality Issues
  - Data from Diverse Sources
  - Complexity of Data Structures
  - Lack of Standardization & Interpretability
  - **Highly Time-Consuming**



Sarih, Houda & Tchangani, Ayeley & Medjaher, Kamal & PERE, Eric. (2019). Data preparation and preprocessing for broadcast systems monitoring in PHM framework. 1444-1449. 10.1109/CoDIT.2019.8820370.

Challenges in Data Wrangling

- Challenges arise from the nature of the data itself, the complexity of data sources, and the goals of the data analysis projects.
  - Volume of Data & Scalability
  - Data Quality Issues
  - Data from Diverse Sources
  - Complexity of Data Structures
  - Lack of Standardization & Interpretability
  - Highly Time-Consuming
  - **Skill and Tool Requirements**



11 Data Scientist Skills Employers Want to See in 2022, https://bootcamp.berkeley.edu/blog/data-scientist-skills/

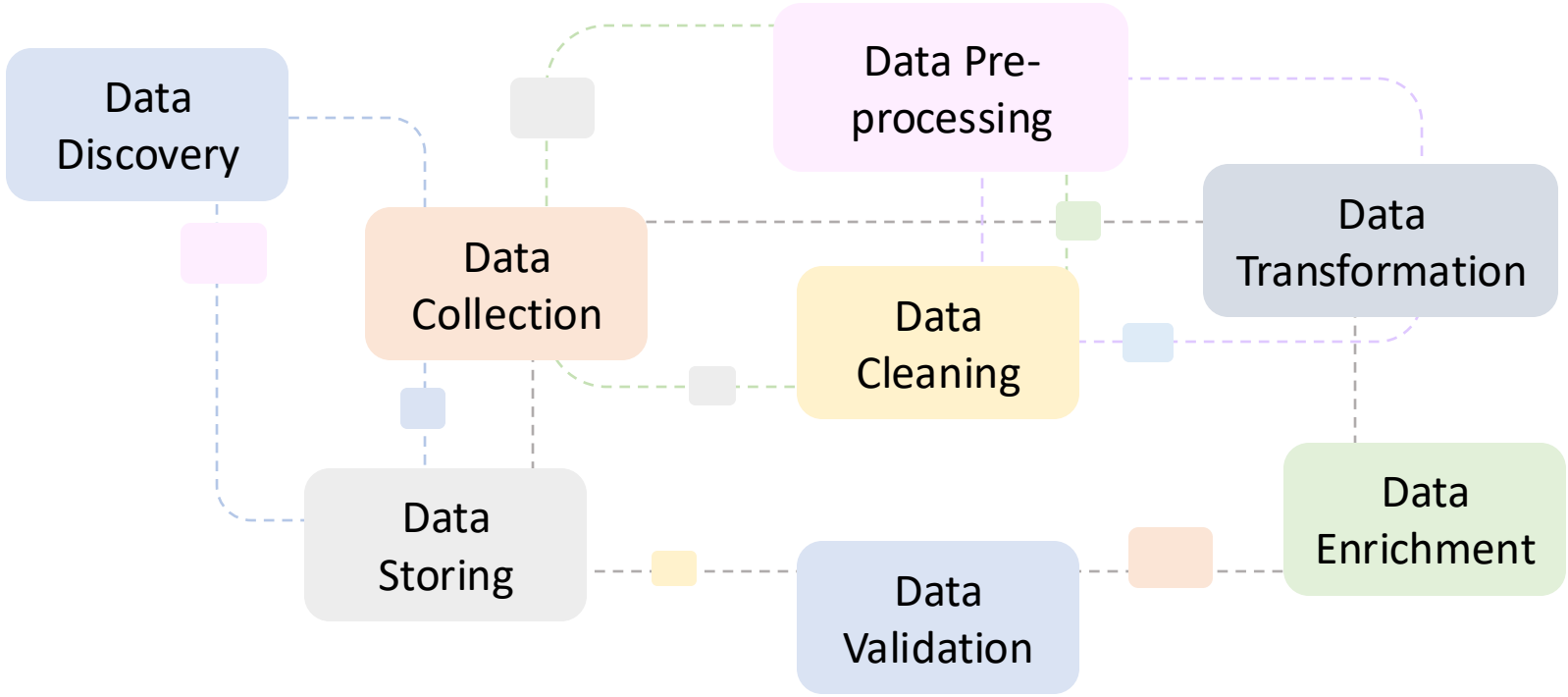
Challenges in Data Wrangling

- Challenges arise from the nature of the data itself, the complexity of data sources, and the goals of the data analysis projects.
  - Volume of Data & Scalability
  - Data Quality Issues
  - Data from Diverse Sources
  - Complexity of Data Structures
  - Lack of Standardization & Interpretability
  - Highly Time-Consuming
  - Skill and Tool Requirements
  - **Data Privacy and Security**

Andrew Kamau, https://blog.google/products/chrome/5-tips-to-stay-safer-online-with-chrome/

Data Wrangling Tasks

- **Data Wrangling** is the process of acquiring, cleaning, structuring, and enriching raw data into a format that is directly usable for analysis.

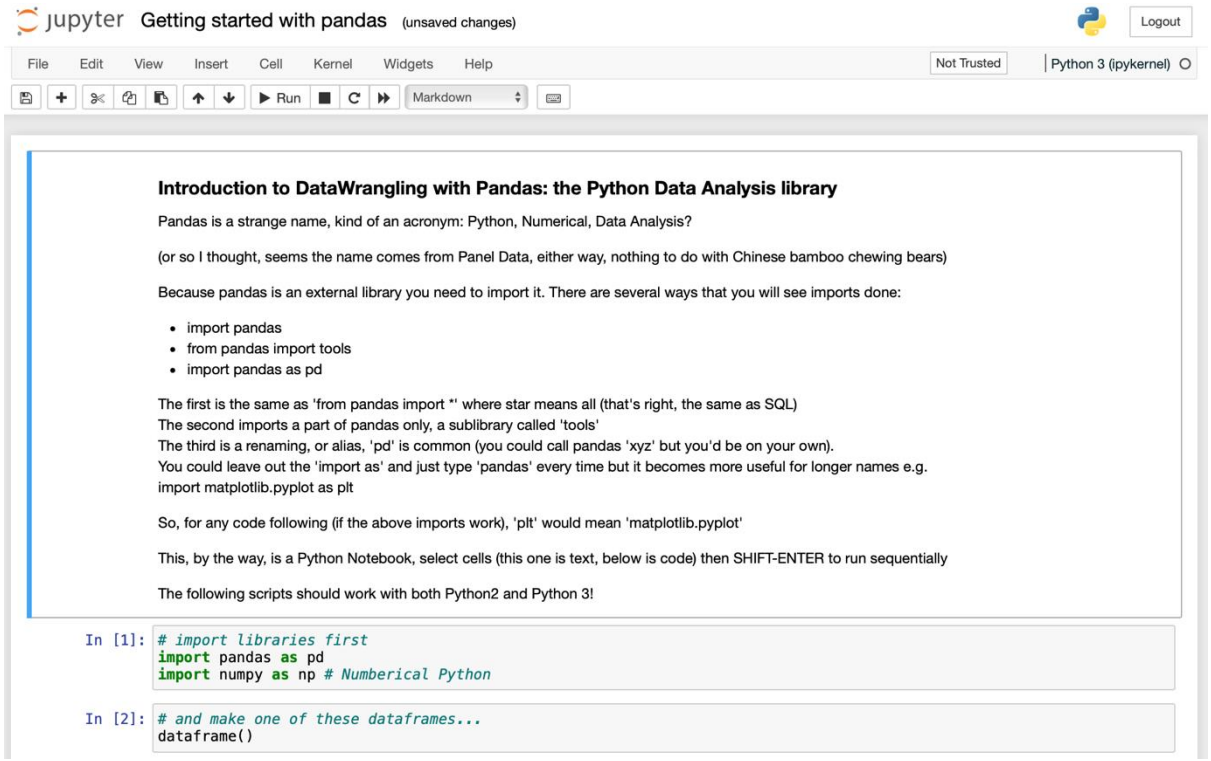


Programming Language & Environment

- Programming language: Python 3.12
  - A scripting language that is easy to get started with and it also comes with a large number of libraries that can be used in data wrangling tasks
  - Major libraries used in this units include (but not limited to)
    - Pandas: a library that provides high-level data structures and manipulation tools that are designed to make data processing fast and easy in Python
    - NLTK: a platform for building Python programs to work with human language data
    - BeautifulSoup: a simple and efficient library for navigating, searching, and modifying HTML and XML documents.
    - Scipy: a fundamental library for scientific computing.
    - scikit-learn: an efficient Python library for data mining and data analysis.

Programming Language & Environment

- Programming environment: Jupyter notebook, Anaconda (optional)
  - The Jupyter Notebook is a web application that allows you to create and share documents that contain live code, equations, visualisations and explanatory text.



Summary & To-do List

- Please please download and read materials provided on Moodle.
- Set up your programming environment by installing Anaconda, Python and Jupyter Notebook or check out Google Colab using your Monash account.
- Last but not least,
  - Choose FIT5196 wisely.
  - Use the discussion (Ed) forum in a proper way and with respect!

- Next Week: Data Wrangling Process & Tasks