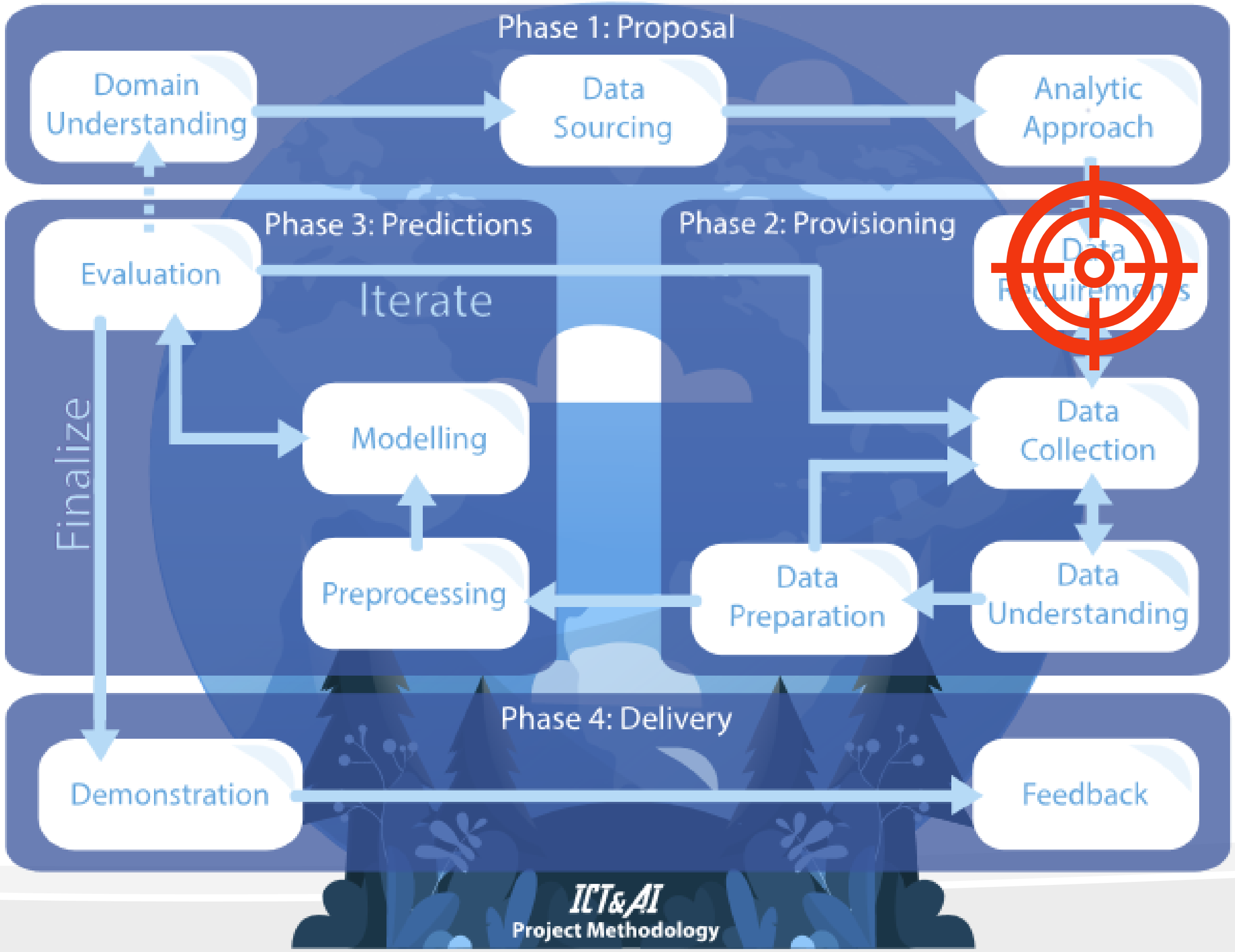
Data Requirements

Your ‘grocery list’

AI project methodology: your roadmap



Target variable, model and domain requirements, data source(s), …
What do you want to achieve (predict)?

What data (quality) is required?
Define a data dictionary

How do you get (generate) and combine your data? Capture the process.

Explore your data (EDA and EDV)

NO DATA

Think before you act..

Shopping list:

☑ Tomatoes
☑ Peppers
☑ Corn
☑ Cucumbers
☑ Eggplant
☑ Lettuce

Make sure you have a common understanding…

What do you put on a grocery list?

- Ingredients: name, brand, alternatives
- Quantity: how much? And how:
- Quality: fresh, canned, frozen, pre-cooke, pre-cut, seasoned, …

Also consider:

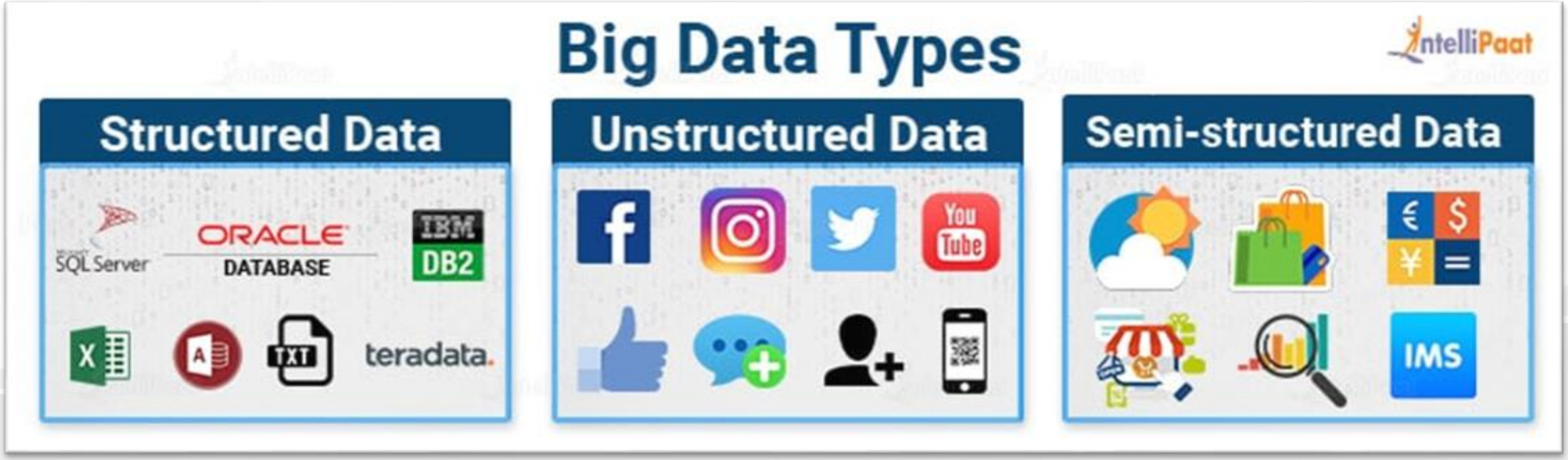
- When do you need it? Check its expiration date
- What’s your budget? Or check out the offers…
- Do you want it in a bag, delivered, …?
- Where to get it, supermarket, fresh market, farm shop, own garden…

Where to start…

1. Identify Data Types
2. List Data Elements
3. Determine Data Volume
4. Define Data Quality Standards
5. Consider Ethical and Legal Aspects
6. Finish with documenting data requirements

Data requirements

1. **Identify Data Types:**
   - **Remember data types explained**
   - **Also think of its origin and form (images, binary, text, numerical,…)**
2. List Data Elements
3. Determine Data Volume

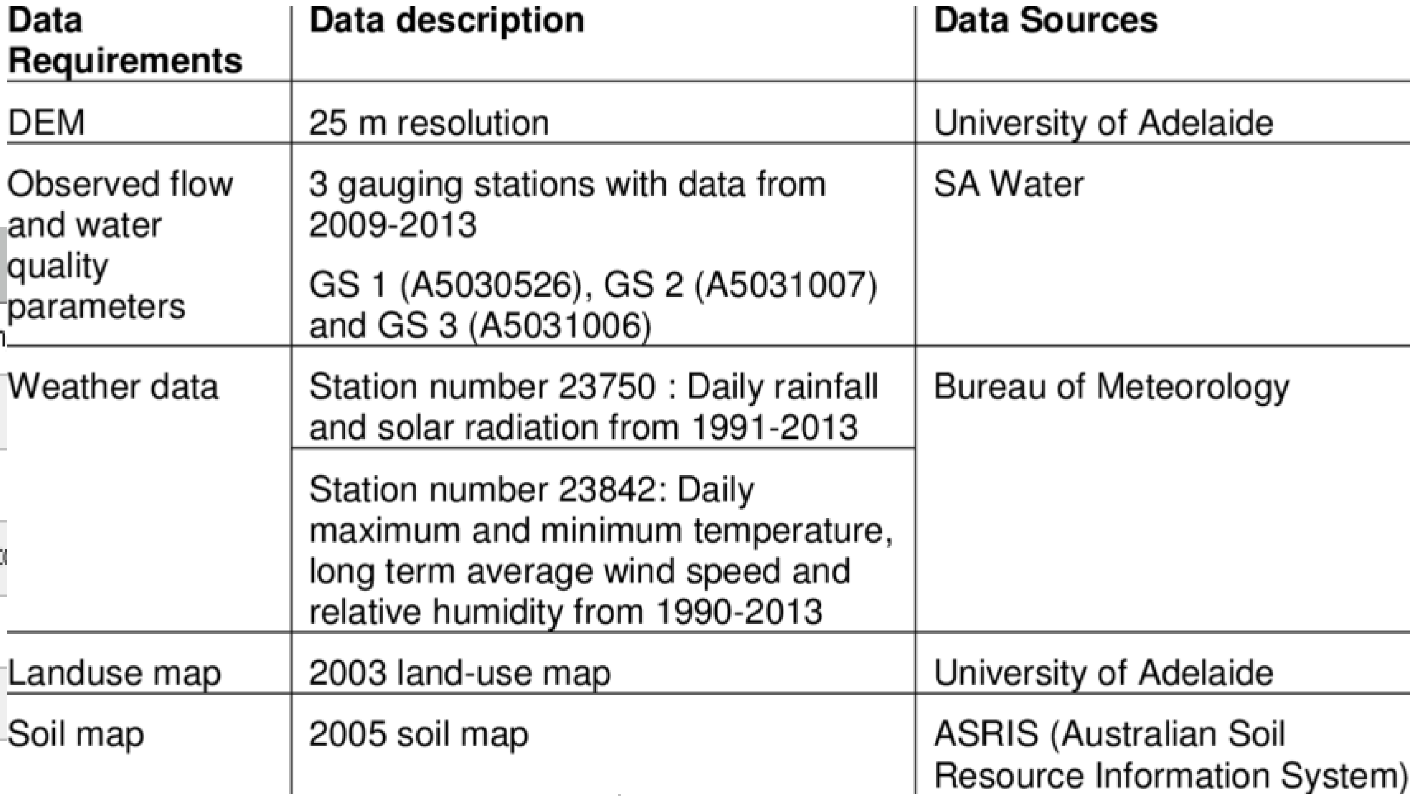


https://www.selecthub.com/big-data-analytics/types-of-big-data-analytics/

Data requirements

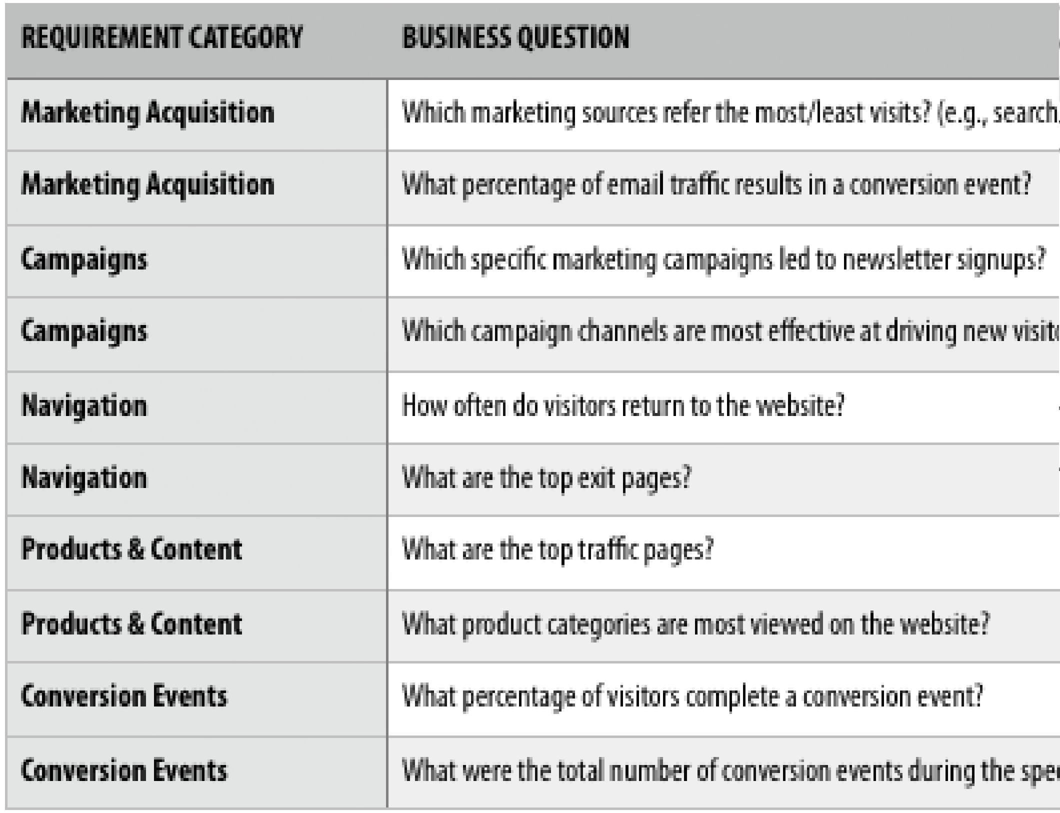
1. Identify Data Types:
2. **List Data Elements**
    1. **Describe content, (possible) values, units/ranges/measures/…**
    2. **(business) rules: mandatory, (restricted) format, relation to other elements…**
    3. **Think of its origin/source (if possible)**
3. Determine Data Volume
4. Define Data Quality Standards
5. Consider Ethical and Legal Aspects
6. Finish with documenting data requirements

Breaking down data elements



Data Requirements	Data description	Data Sources
DEM	25 m resolution	University of Adelaide
Observed flow and water quality parameters	3 gauging stations with data from 2009-2013 GS 1 (A5030526), GS 2 (A5031007) and GS 3 (A5031006)	SA Water
Weather data	Station number 23750 : Daily rainfall and solar radiation from 1991-2013 Station number 23842: Daily maximum and minimum temperature, long term average wind speed and relative humidity from 1990-2013	Bureau of Meteorology
Landuse map	2003 land-use map	University of Adelaide
Soil map	2005 soil map	ASRIS (Australian Soil Resource Information System)

https://www.researchgate.net/figure/Data-requirements-descriptions-and-sources_tbl1_268747018



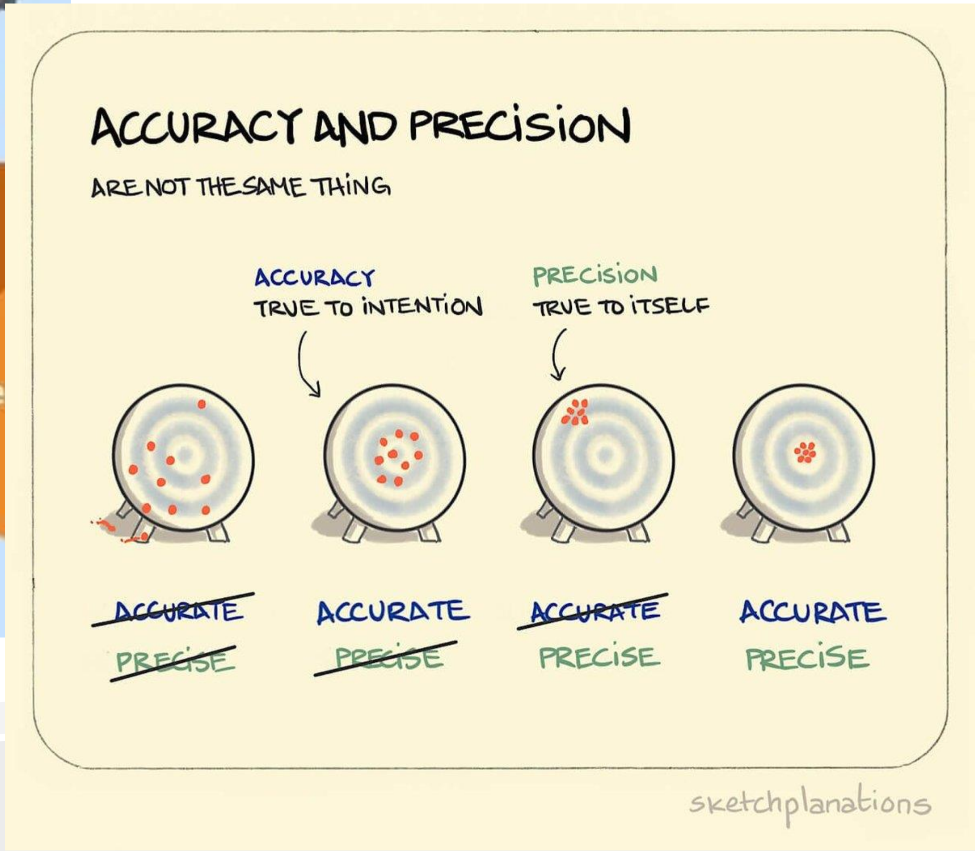
REQUIREMENT CATEGORY	BUSINESS QUESTION
Marketing Acquisition	Which marketing sources refer the most/least visits? (e.g., search
Marketing Acquisition	What percentage of email traffic results in a conversion event?
Campaigns	Which specific marketing campaigns led to newsletter signups?
Campaigns	Which campaign channels are most effective at driving new visit
Navigation	How often do visitors return to the website?
Navigation	What are the top exit pages?
Products & Content	What are the top traffic pages?
Products & Content	What product categories are most viewed on the website?
Conversion Events	What percentage of visitors complete a conversion event?
Conversion Events	What were the total number of conversion events during the specified time period?

https://analyticsdemystified.com/analytics-strategy/bulletproof-business-requirements/

How to describe data elements…

S
M
A
R
T
Specific
Measurable
Attainable
Relevant
Time Base

https://www.toolshero.com/personal-development/smart-goals/

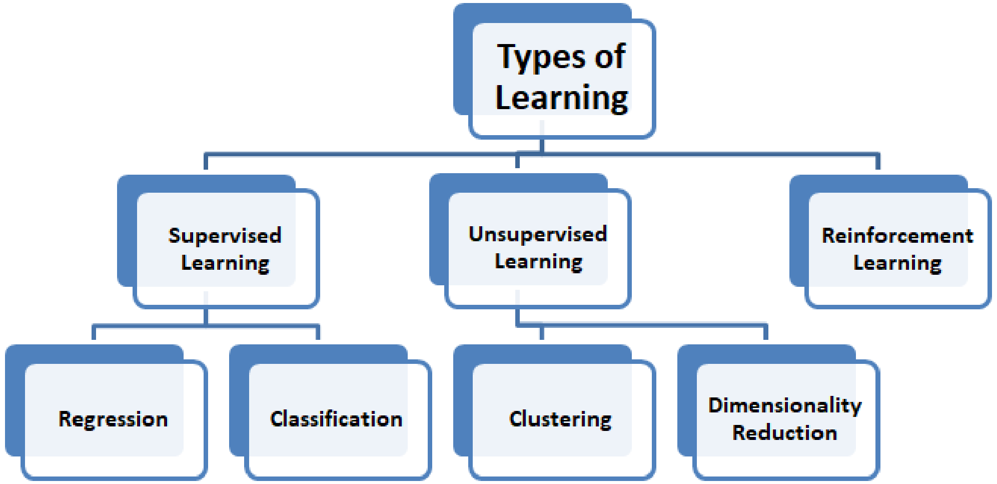


https://twitter.com/sketchplanator/status/10470789454625955

Data requirements

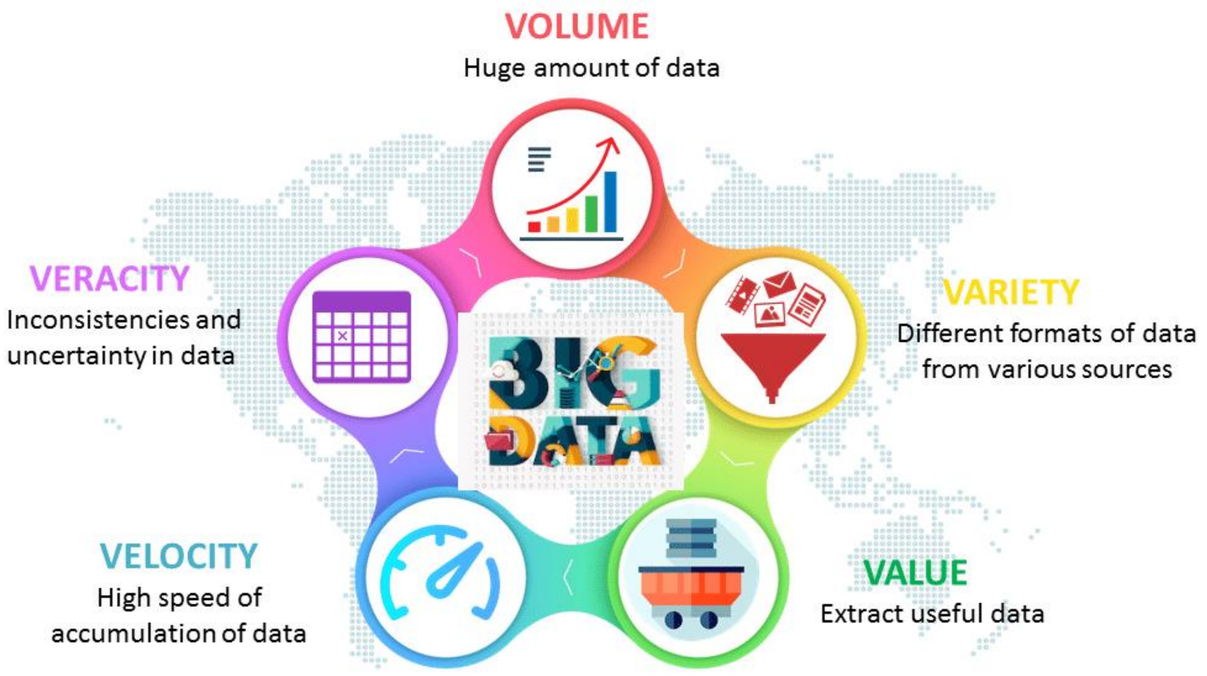
1. Identify Data Types
2. List Data Elements
3. **Determine Data Volume**
   - **What's your analytic approach (specific model needs)**

4. Define Data Quality Standards
5. Consider Ethical and Legal Aspects
6. Finish with documenting data requirements



Data requirements

1. Identify Data Types
2. List Data Elements
3. Determine Data Volume
4. **Define Data Quality Standards**
   - **Remember the V's of big data?**



VOLUME
Huge amount of data

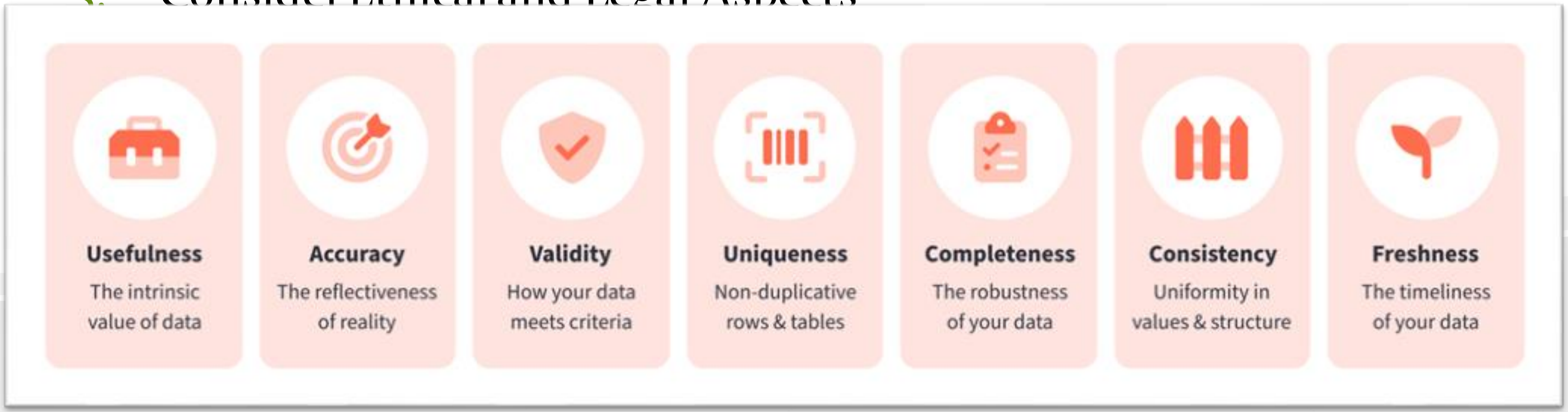
VARIETY
Different formats of data from various sources

VALUE
Extract useful data

VELOCITY
High speed of accumulation of data

VERACITY
Inconsistencies and uncertainty in data

5. Consider Ethical and Legal Aspects



Usefulness
The intrinsic value of data

Accuracy
The reflectiveness of reality

Validity
How your data meets criteria

Uniqueness
Non-duplicative rows & tables

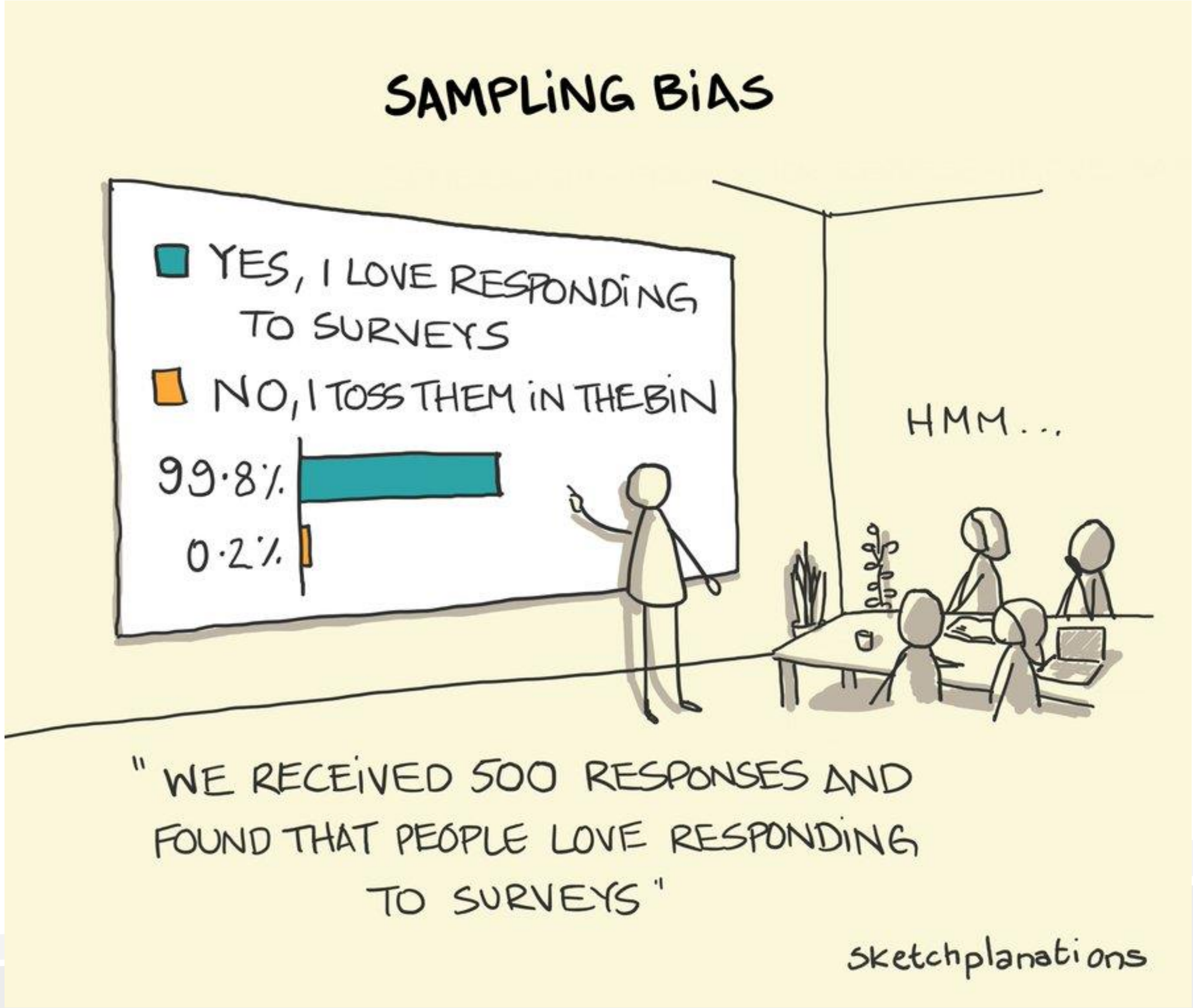
Completeness
The robustness of your data

Consistency
Uniformity in values & structure

Freshness
The timeliness of your data

Data quality

Data is always just a snapshot of reality…



SAMPLING BIAS
🟦 YES, I LOVE RESPONDING TO SURVEYS
🟧 NO, I TOSS THEM IN THE BIN
99.8%
0.2%
HMM…
"WE RECEIVED 500 RESPONSES AND FOUND THAT PEOPLE LOVE RESPONDING TO SURVEYS"
sketchplanations

https://twitter.com/sketchplanator/status/1409175698166763528

Data requirements

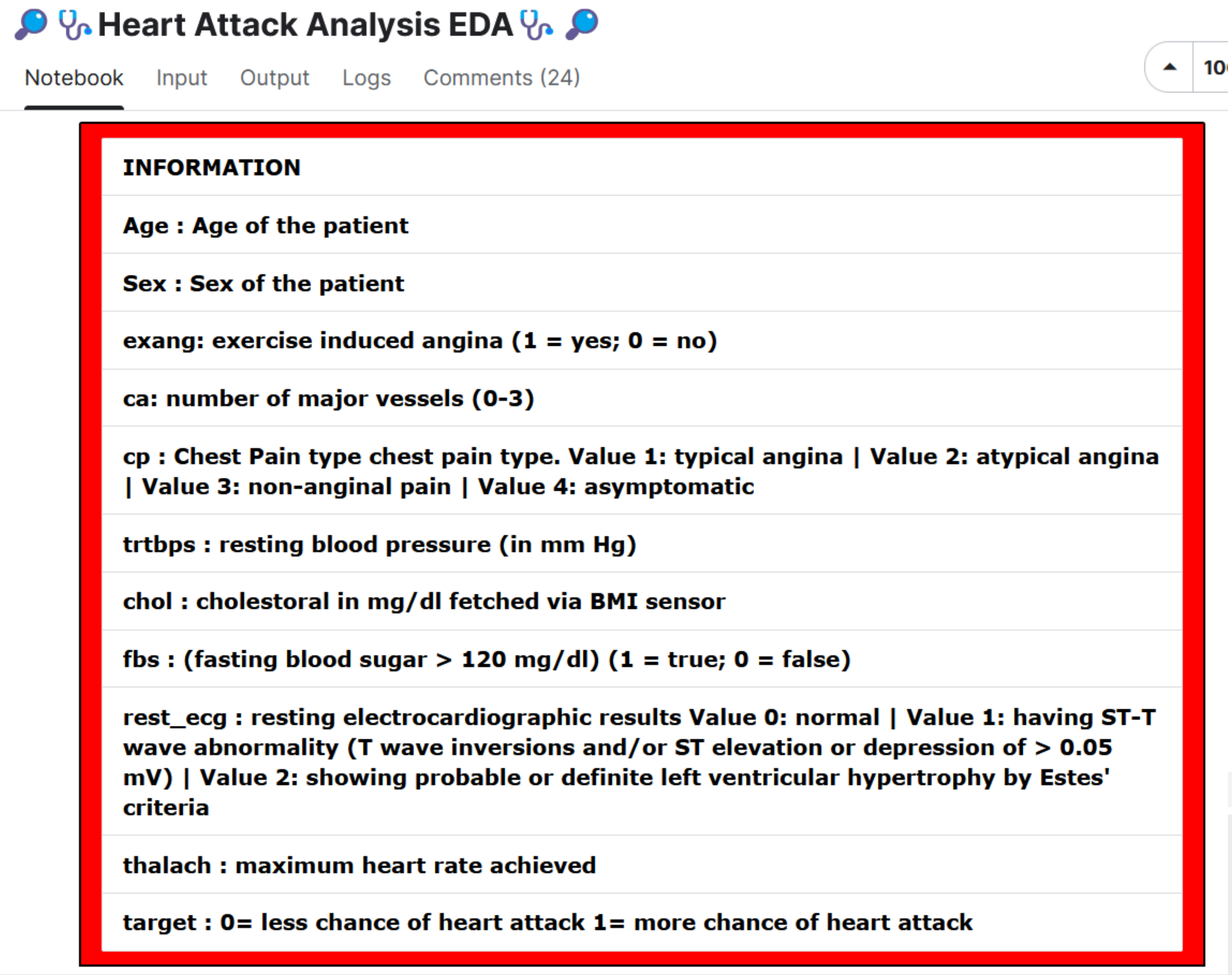
1. Identify Data Types
2. List Data Elements
3. Determine Data Volume
4. Define Data Quality Standards
5. **Consider Ethical and Legal Aspects**
    - **Societal Impact**
    - **It is not just about privacy/GDPR (think of criminal acts, competition, fraud,…)**
6. Finish with documenting data requirements

Data requirements

1. Identify Data Types
2. List Data Elements
3. Determine Data Volume
4. Define Data Quality Standards
5. Consider Ethical and Legal Aspects
6. **Finish with documenting data requirements:**
   - **Data dictionary: a collection of metadata such as object name, data type, size, classification, and relationships with other data assets**
   - **Data catalog/data definitions/.... (business glossary)**

Examples

Existing work on Kaggle…
not all the best example



🔎 🩺 Heart Attack Analysis EDA 🩺 🔎

INFORMATION

Age : Age of the patient

Sex : Sex of the patient

exang: exercise induced angina (1 = yes; 0 = no)

ca: number of major vessels (0-3)

cp : Chest Pain type chest pain type. Value 1: typical angina | Value 2: atypical angina | Value 3: non-anginal pain | Value 4: asymptomatic

trtbps : resting blood pressure (in mm Hg)

chol : cholestoral in mg/dl fetched via BMI sensor

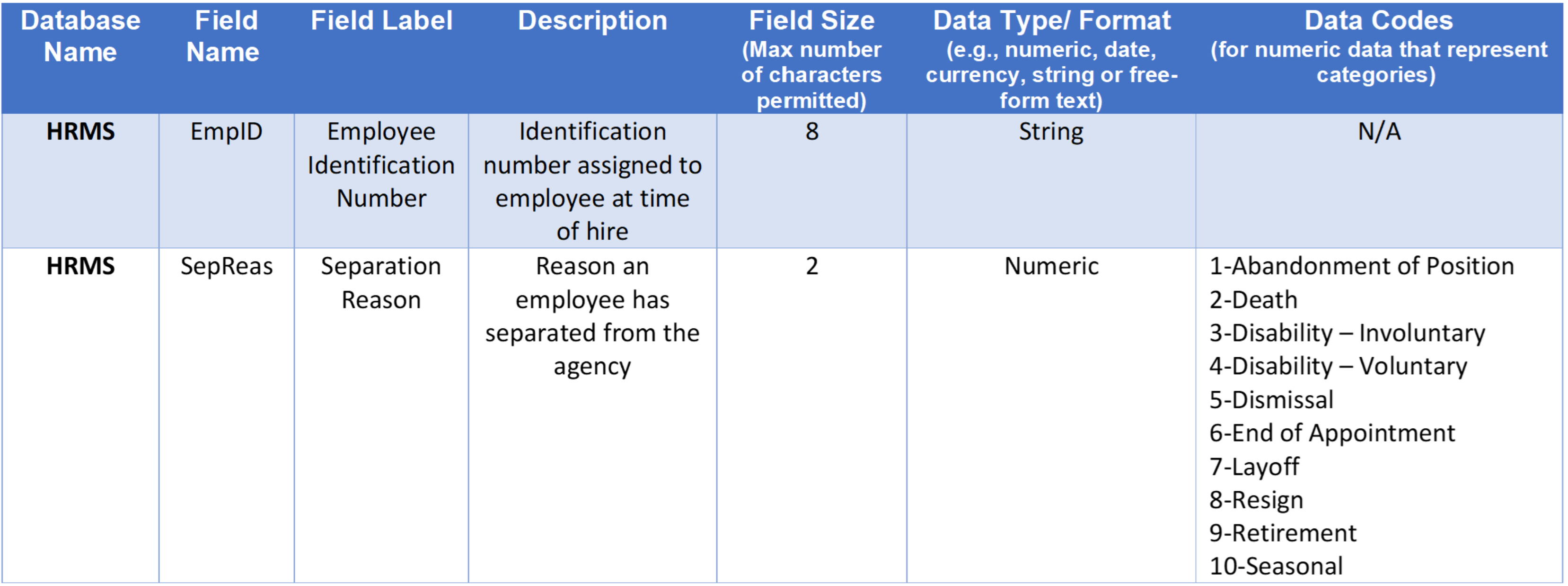
fbs : (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

rest_ecg : resting electrocardiographic results Value 0: normal | Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) | Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria

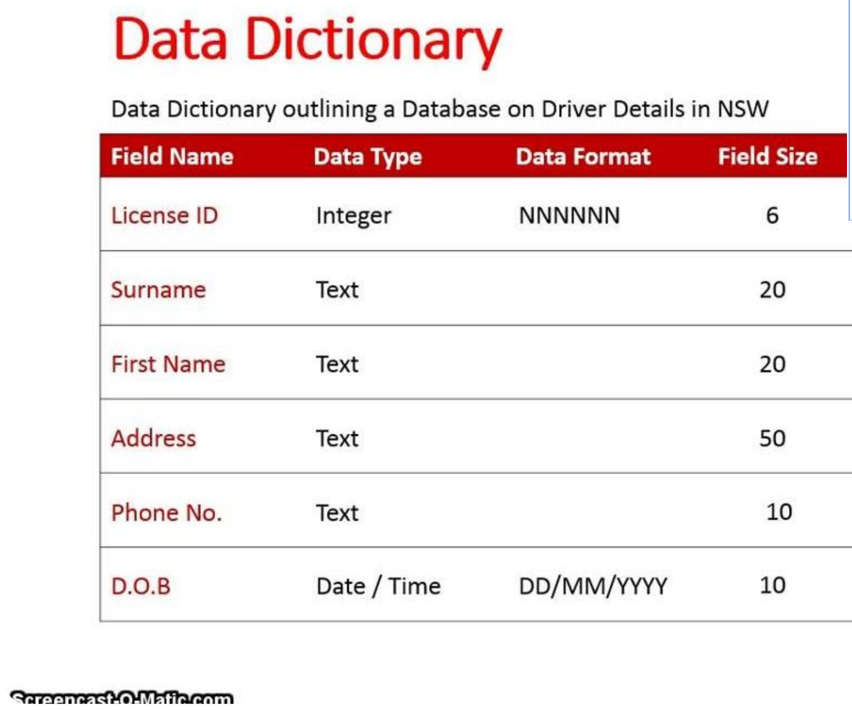
thalach : maximum heart rate achieved

target : 0= less chance of heart attack 1= more chance of heart attack

Examples



Database Name	Field Name	Field Label	Description	Field Size (Max number of characters permitted)	Data Type/ Format (e.g., numeric, date, currency, string or free-form text)	Data Codes (for numeric data that represent categories)
HRMS	EmpID	Employee Identification Number	Identification number assigned to employee at time of hire	8	String	N/A
HRMS	SepReas	Separation Reason	Reason an employee has separated from the agency	2	Numeric	1-Abandonment of Position 2-Death 3-Disability – Involuntary 4-Disability – Voluntary 5-Dismissal 6-End of Appointment 7-Layoff 8-Resign 9-Retirement 10-Seasonal



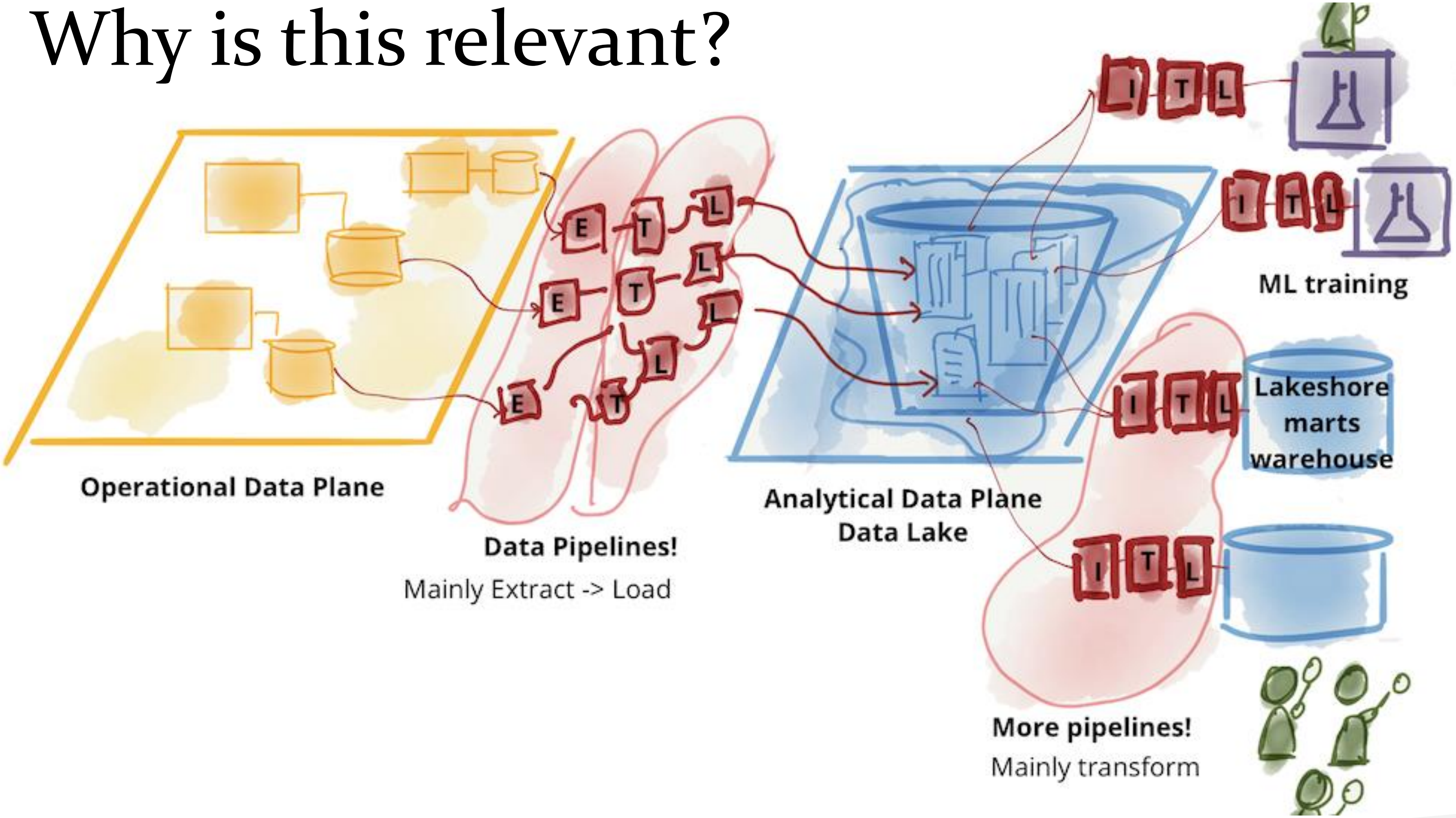
Data Dictionary

Data Dictionary outlining a Database on Driver Details in NSW

Field Name	Data Type	Data Format	Field Size		
License ID	Integer	NNNNNN	6	ID for all drivers	[illegible]
Surname	Text		20	Surname for Driver	Jones
First Name	Text		20	First Name for Driver	Arnold
Address	Text		50	First Name for Driver	11 Rocky st Como 2233
Phone No.	Text		10	License holders contact number	0400111222
D.O.B	Date / Time	DD/MM/YYYY	10	Drivers Date of Birth	08/05/1956

- https://atlan.com/data-catalog-vs-data-dictionary/?ref=/what-is-a-data-dictionary/
- https://www.usgs.gov/data-management/data-dictionaries
- https://help.osf.io/article/217-how-to-make-a-data-dictionary

Why is this relevant?



Operational Data Plane

Data Pipelines!

Mainly Extract -> Load

Analytical Data Plane

Data Lake

ML training

Lakeshore marts warehouse

More pipelines!

Mainly transform

Exercise

- Study ‘Data Requirements’ section in Canvas
- Start the Data Requirements exercise (under ‘Lectures & Exercises’)
- It’s a group project and effort, so discuss your ideas and results with your group
- Submit your groupwork (1 group member) before 12:00 h.

Data requirements

Create your ‘grocery list’ and keep refining/improving it…