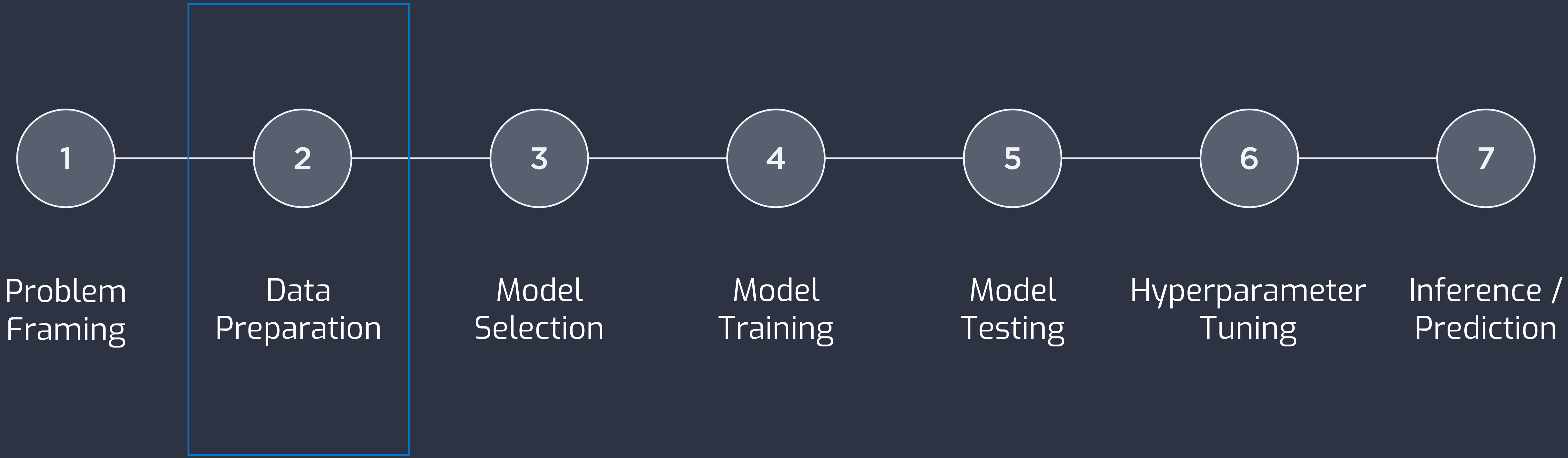
COMP2261 ARTIFICIAL INTELLIGENCE / MACHINE LEARNING

Data Preparation

Dr SHI Lei



1. Problem Framing
2. Data Preparation
3. Model Selection
4. Model Training
5. Model Testing
6. Hyperparameter Tuning
7. Inference / Prediction

Learning Objectives

- Understand what is Data
- Have an overview of data preparation for machine learning

What is Data?

EXAMPLE. to build a happiness predictor

Country	GDP/capita	Happiness Score
Austria	1.376	7.246
Finland	1.340	7.769
France	1.324	6.592
Germany	1.373	6.985
Greece	1.181	5.287
Italy	1.294	6.223
Netherlands	1.396	7.488
Portugal	1.221	5.693
Poland	1.206	6.182
Spain	1.286	6.354
Sweden	1.387	7.343
United Kingdom	1.333	7.054

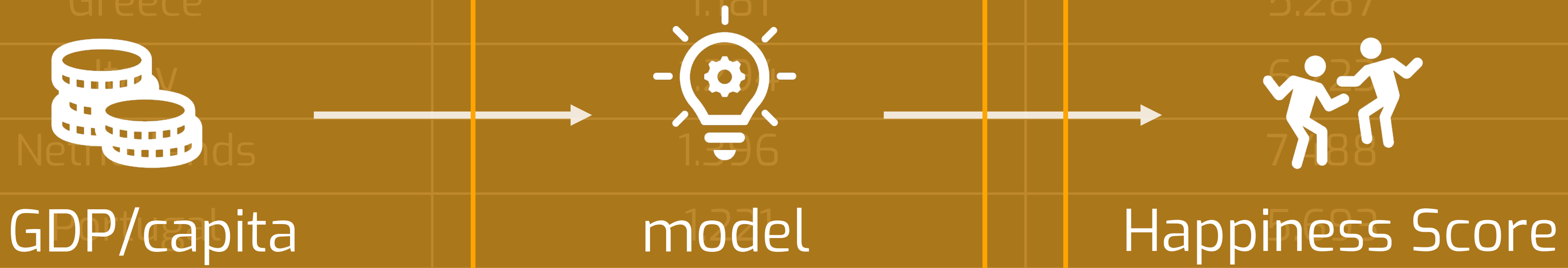
Source: https://www.kaggle.com/unsdsn/world-happiness

EXAMPLE. to build a happiness predictor

input X

output y

Country	GDP/capita	Happiness Score
Austria	1.376	7.246
Finland	1.340	7.769
France	1.324	6.592
Germany	1.373	6.985
Greece	1.181	5.287
Italy	1.294	6.223
Netherlands	1.396	7.488
Portugal	1.221	5.693
Poland	1.206	6.182
Spain	1.286	6.354
Sweden	1.387	7.343
United Kingdom	1.333	7.054



GDP/capita → model → Happiness Score

Source: https://www.kaggle.com/unsdsn/world-happiness

EXAMPLE. to build a happiness predictor

Country	GDP/capita	Generosity	Happiness Score
Austria	1.376	0.244	7.246
Finland	1.340	0.153	7.769
France	1.324	0.111	6.592
Germany	1.373	0.261	6.985
Greece	1.181	0.000	5.287
Italy	1.294	0.158	6.223
Netherlands	1.396	0.322	7.488
Portugal	1.221	0.047	5.693
Poland	1.206	0.117	6.182
Spain	1.286	0.153	6.354
Sweden	1.387	0.267	7.343
United Kingdom	1.333	0.348	7.054

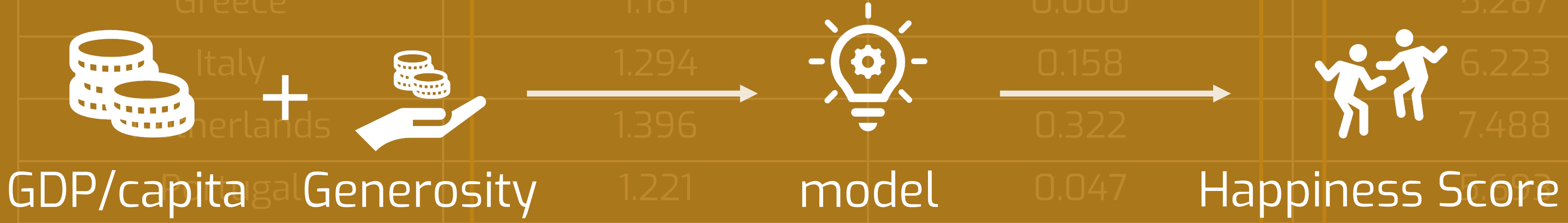
Source: https://www.kaggle.com/unsdsn/world-happiness

EXAMPLE. to build a happiness predictor

input X

output y

Country	GDP/capita	Generosity	Happiness Score
Austria	1.376	0.244	7.246
Finland	1.340	0.153	7.769
France	1.324	0.111	6.592
Germany	1.373	0.261	6.985
Greece	1.181	0.000	5.287
Italy	1.294	0.158	6.223
Netherlands	1.396	0.322	7.488
Portugal	1.221	0.047	5.693
Poland	1.206	0.117	6.182
Spain	1.286	0.153	6.354
Sweden	1.387	0.267	7.343
United Kingdom	1.333	0.348	7.054



GDP/capita + Generosity → model → Happiness Score

Source: https://www.kaggle.com/unsdsn/world-happiness

EXAMPLE. to build a happiness predictor

Country	GDP/capita	Generosity	Happiness Score
Austria	1.376	0.244	7.246
Finland	1.340	0.153	7.769
France	1.324	0.111	6.592
Germany	1.373	0.261	6.985
Greece	1.181	0.000	5.287
Italy	1.294	0.158	6.223
Netherlands	1.396	0.322	7.488
Portugal	1.221	0.047	5.693
Poland	1.206	0.117	6.182
Spain	1.286	0.153	6.354
Sweden	1.387	0.267	7.343
United Kingdom	1.333	0.348	7.054

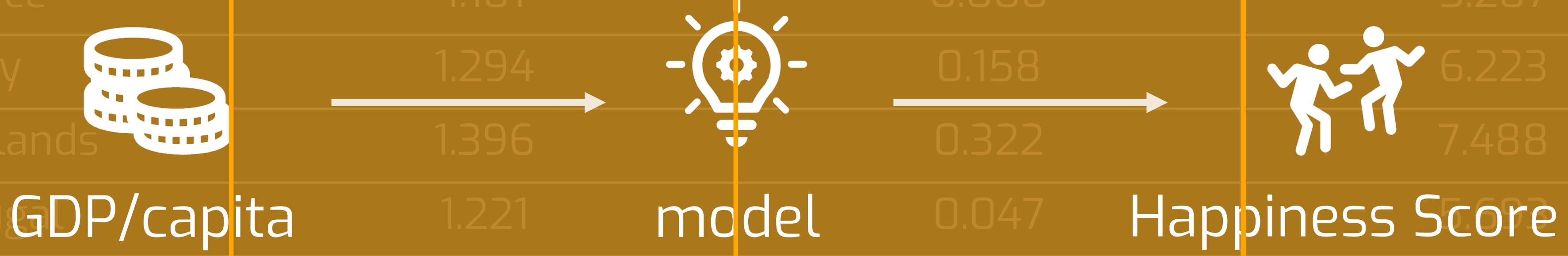
Source: https://www.kaggle.com/unsdsn/world-happiness

EXAMPLE. to build a happiness predictor

output y

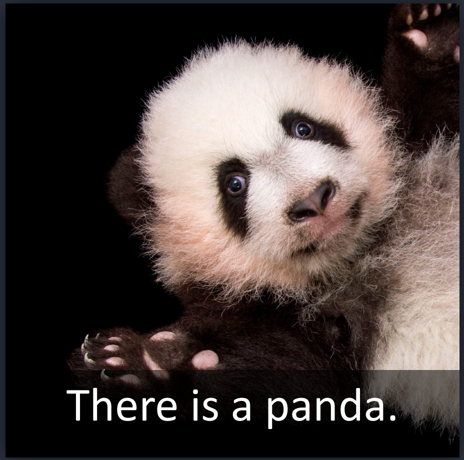
input X

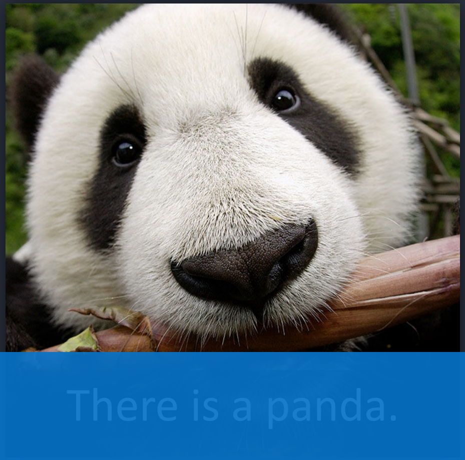
Country	GDP/capita	Generosity	Happiness Score
Austria	1.376	0.244	7.246
Finland	1.340	0.153	7.769
France	1.324	0.111	6.592
Germany	1.373	0.261	6.985
Greece	1.181	0.000	5.287
Italy	1.294	0.158	6.223
Netherlands	1.396	0.322	7.488
Portugal	1.221	0.047	5.693
Poland	1.206	0.117	6.182
Spain	1.286	0.153	6.354
Sweden	1.387	0.267	7.343
United Kingdom	1.333	0.348	7.054

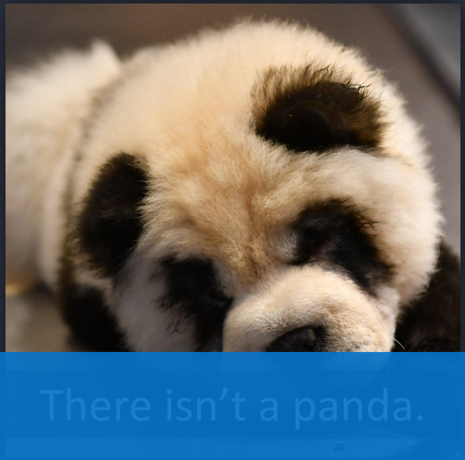


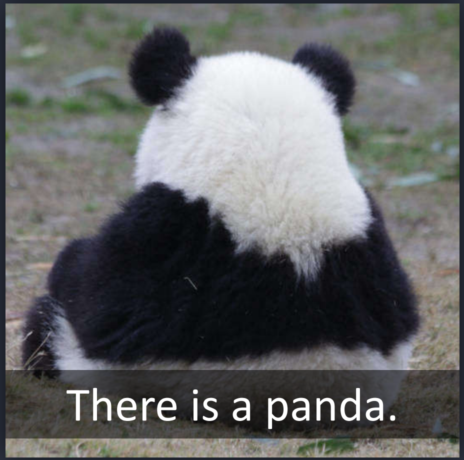
Source: https://www.kaggle.com/unsdsn/world-happiness

EXAMPLE. to build a panda detector

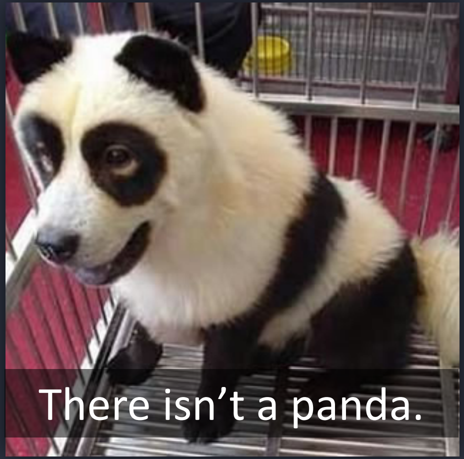

There is a panda.

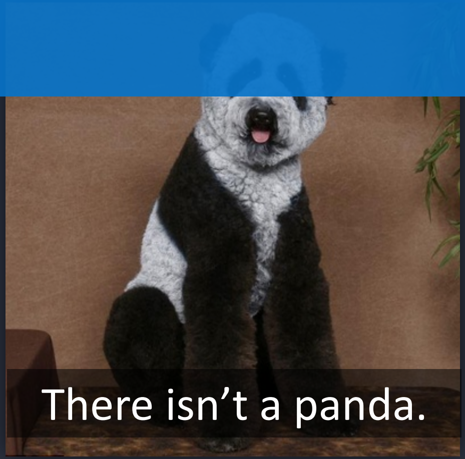

There is a panda.

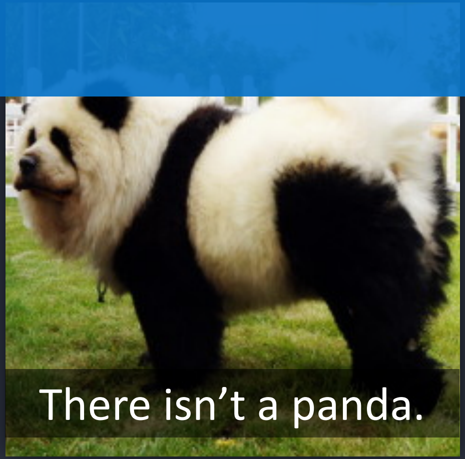

There isn’t a panda.

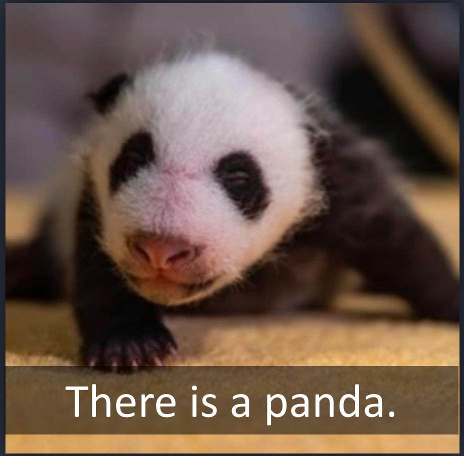

There is a panda.

Manual Labelling


There isn’t a panda.


There isn’t a panda.


There isn’t a panda.

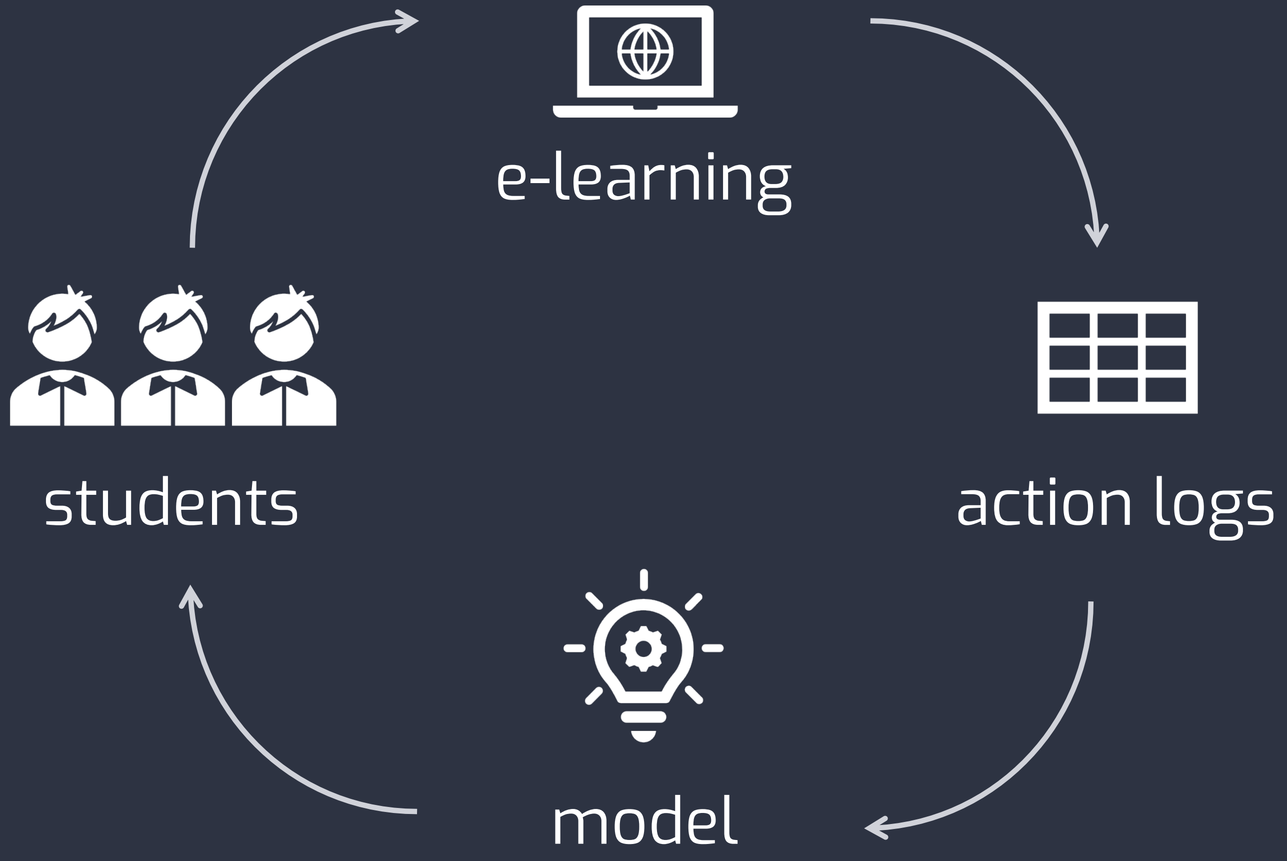

There is a panda.

Manual Labelling

Behaviour Tracking

EXAMPLE. to build a panda detector

Behaviour Tracking



✓ Takeaway Points

- Data preparation is extremely important and takes long time.
- For different problems, same data can be used very differently.
- Manually Labelling data is sometimes necessary.
- Behaviour Tracking can help collect data automatically.