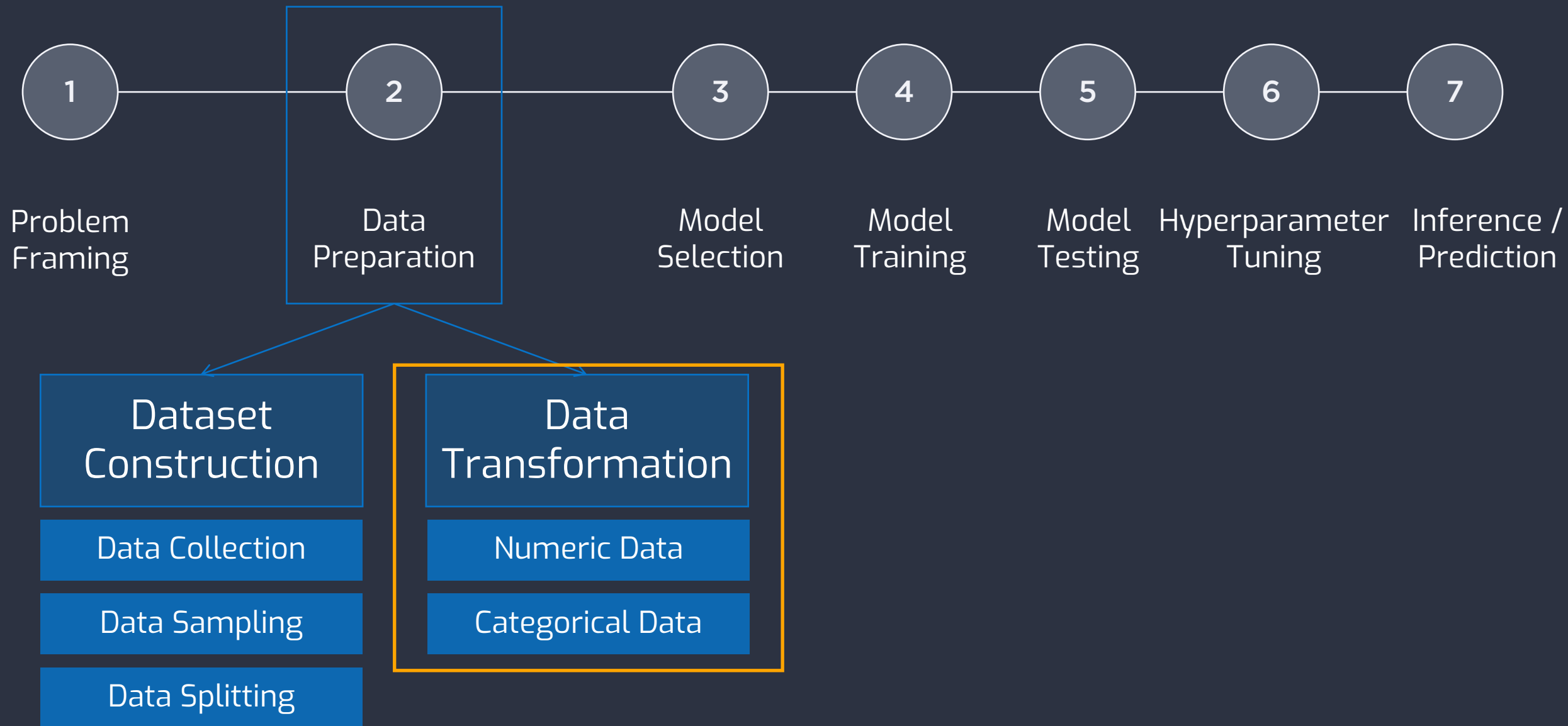


COMP2261 ARTIFICIAL INTELLIGENCE / MACHINE LEARNING

# Data Transformation

Dr SHI Lei



## Learning Objectives

- Understand what is Data Transformation
- Understand why we need Data Transformation

In Machine Learning projects, we need Data to train models.



Data in real world doesn't come in a nice format that can be directly used.



Raw / source data (directly collected from a source)

- From an electronic sensor which measures physical input  
e.g. light, temperature, humidity, pressure, radiation level...

- From human inputs

→ e.g. texts, handwritings, voices... →



- May contain useful info & human, machine, instrument errors.
- May be in different colloquial formats, unformatted, uncoded or suspected such as outliers.
- May be in XML, JSON, relational database records, etc.

We need to put those heterogeneous data sources together and create Feature vectors from it for learning algorithms.

# Feature Engineering

The process of selecting and extracting features from raw data.



art



1010  
1010

science / engineering

Machine learning engineers often spend 70%~80% of their time in data preparation phase before modelling.

# Data Transformation



# Data Transformation

EXAMPLE. transforming raw data to feature vector

## Raw Data (in JSON)

```
0: {  
  student_info: {  
    first_name: Bruce,  
    last_name: Lee,  
    age: 18,  
    sex: male,  
    height: 1.72,  
    num_module: -1  
    ...  
  }  
}
```



Feature Engineering

## Feature Vector

```
[  
  32.0,  
  96.0,  
  18.0,  
  0.0,  
  1.72,  
  -1.0,  
  0.0,  
  3.142,  
  ...  
]
```

Feature engineering: the process of selecting and extracting features from raw data.

# Data Transformation

## Feature

A specific representation on top of raw data.

It is an individual, measurable attribute shared by all of the independent observations on which analysis or prediction is to be done.

*An **attribute** could only be a **feature** if it is useful to the model; and in another word, a feature being a feature other than being an attribute is because it is in the context of a specific problem and can contribute to solving the problem.*

# Main reasons to transform features

## Mandatory transformations

For data compatibility  
e.g., we convert non-numeric features into numeric, so that we can do matrix multiplication; we resize the inputs to a fixed size, because linear models have a fixed number of inputs.

## Optional quality transformations

For better model performance  
e.g., tokenisation of text features; or normalise numeric features.

# Before feature transformation

Need to explore and clean up data.

e.g., explore several rows of data and check basic statistics or fix missing entries.



Data can look one way in basic statistics and another when in graph

Visualise data in graphs and charts such as scatter plots, lines and histograms.



## ✓ Takeaway Points

- Data may come from various sources e.g. sensors, human, etc.
- Data may be in various formats, unformatted, uncoded, etc.
- Data may contain useful info and human, machine errors.
- Need to put heterogeneous data sources together and create feature vectors from it for learning algorithms
- Transforming raw data to feature vector - feature engineering.
- Need to explore and clean up data before data transformation.