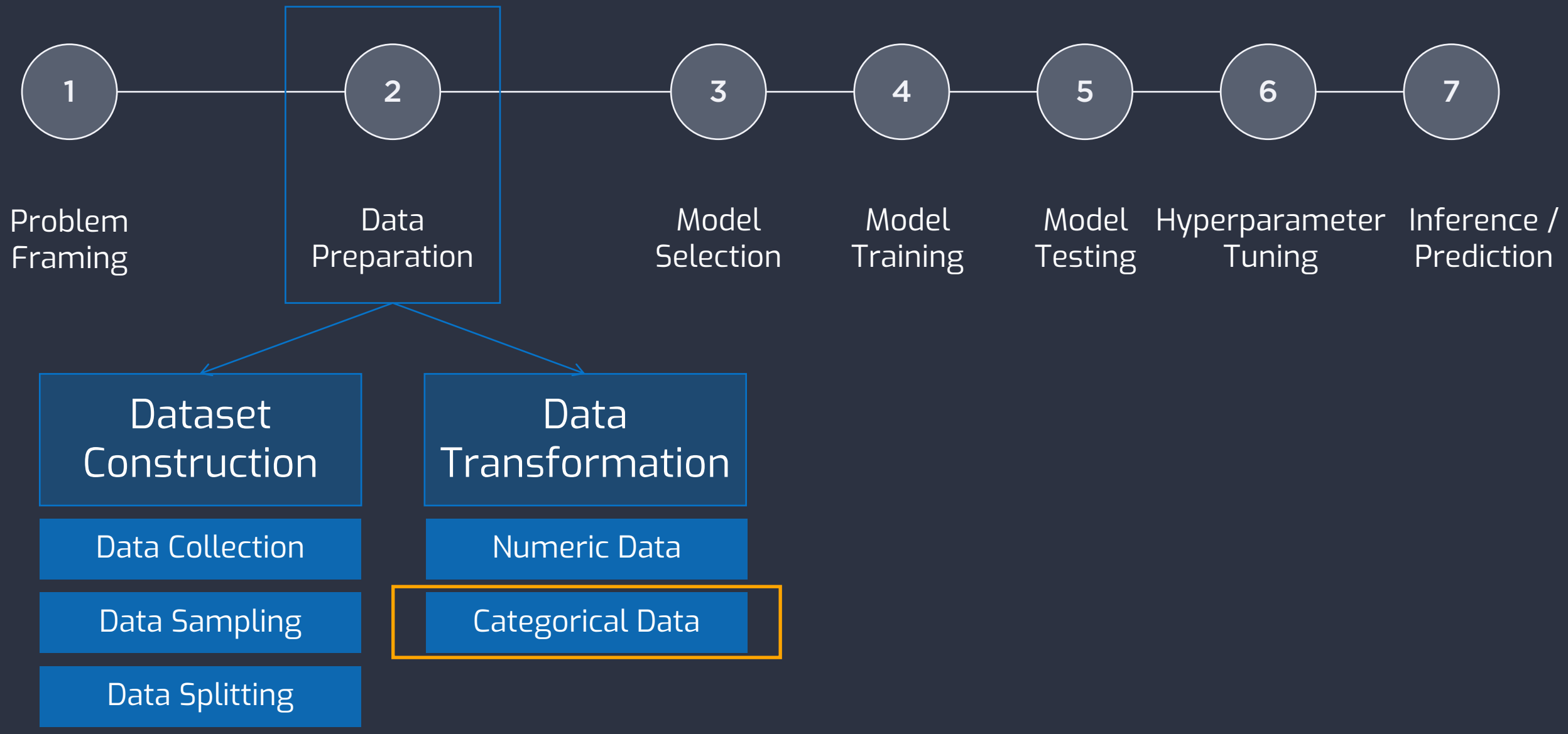
COMP2261 ARTIFICIAL INTELLIGENCE / MACHINE LEARNING

Transforming Categorical Data

Dr SHI Lei

Durham University

learning lab


1
2
3
4
5
6
7
Problem Framing
Data Preparation
Model Selection
Model Training
Model Testing
Hyperparameter Tuning
Inference / Prediction
Dataset Construction
Data Collection
Data Sampling
Data Splitting
Data Transformation
Numeric Data
Categorical Data

Learning Objectives

- Understand the challenges of working with categorical data.
- Understand differences between nominal and ordinal data.
- Learn various encoding techniques.

What is Categorical Data?

Categorical Data

- Categorical data is discrete.
- Each value of a categorical feature is called a category or a group, it could be represented as a string.

e.g.

Role: student | lecturer | admin

Two types of categorical data: Nominal and Ordinal

Categorical Data - Nominal (labelled)

- Nominal data is also called labelled data.
- Values are not comparable or calculatable.
- e.g. race, gender, sexuality, religion, nationality, blood type…

Labour vs Conservative longer, heavier, warmer??

Liberal Democrats + Scottish National = ??

Democratic Unionist - Sinn Féin = ??

Categorical Data - Ordinal

- Ordinal data is with a set order or scale to it.
- Values are comparable.
- e.g. initial | intermediate | advanced

  XS | S | M | L | XL

  First | Second | Third

- There is an ordered relationship between categories but this order does not have a standard scale to be measured.

~~intermediate = initial + 1~~ ~~Third = First x 3~~

Categorical Data

- Sometimes, we must use categorical data instead of numeric data to present certain features even if they contain integer values.

Categorical Data

- Machine learning algorithms require features to be numbers.
- Encoding categorical data in order to be able to fit and evaluate models.
- Encoding techniques

Ordinal Encoding	One-Hot Encoding
Feature Crossing	Feature Hashing

Ordinal Encoding

Ordinal Encoding

- To transform categorical features into ordinal numbers in ordered set, i.e. to assign a sequence of numerical values as per the order of data.
- In most cases, we map each unique label to an integer value.
- We can only use ordinal encoding when there is a known relationship between the categories.
- If we map inherent ordinality of an ordinal feature to a wrong scale, there will be a very negative impact on model's performance.

EXAMPLE. Customer feedback survey

Customer rating for service, in an ordered relationship

Very poor	Poor	Fair	Good	Very good

(e.g. good is better than fair; fair is better than poor…)

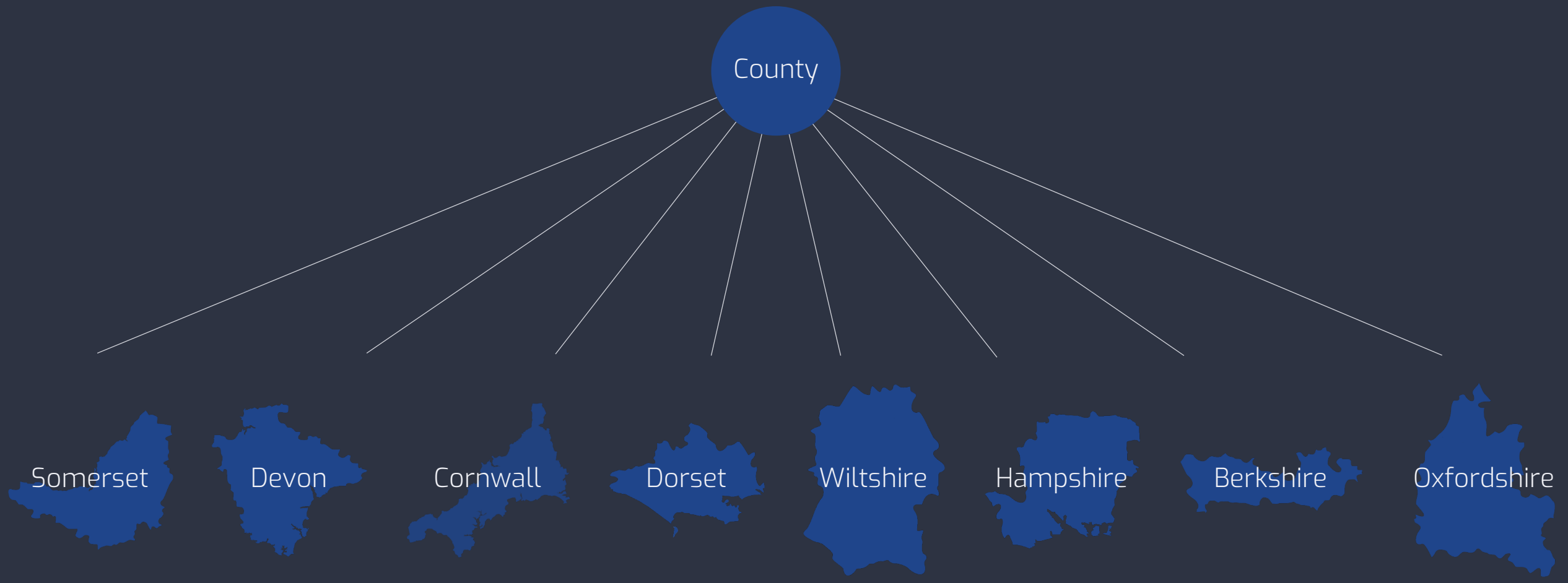
EXAMPLE. Customer feedback survey

Customer rating for service, in an ordered relationship

Very poor	Poor	Fair	Good	Very good
1	2	3	4	5

(e.g. good is better than fair; fair is better than poor…)

EXAMPLE. Counties of England



One-Hot Encoding

One-Hot Encoding

EXAMPLE. Counties of England

County
Berkshire
Cornwall
Devon
Dorset
Hampshire
Oxfordshire
Somerset
Wiltshire

One-Hot Encoding

EXAMPLE. Counties of England

County	Berkshire	Cornwall	Devon	Dorset	Hampshire	Oxfordshire	Somerset	Wiltshire
Berkshire	1	0	0	0	0	0	0	0
Cornwall	0	1	0	0	0	0	0	0
Devon	0	0	1	0	0	0	0	0
Dorset	0	0	0	1	0	0	0	0
Hampshire	0	0	0	0	1	0	0	0
Oxfordshire	0	0	0	0	0	1	0	0
Somerset	0	0	0	0	0	0	1	0
Wiltshire	0	0	0	0	0	0	0	1

One-Hot Encoding

EXAMPLE. Counties of England

Out-Of-Vocabulary Bin (OOV bin)

Berkshire	Cornwall	Devon	Dorset	Hampshire	Oxfordshire	Somerset	Wiltshire	Other
1	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0
0	0	0	0	1	0	0	0	0
0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	1

One-Hot Encoding

EXAMPLE. Counties of England

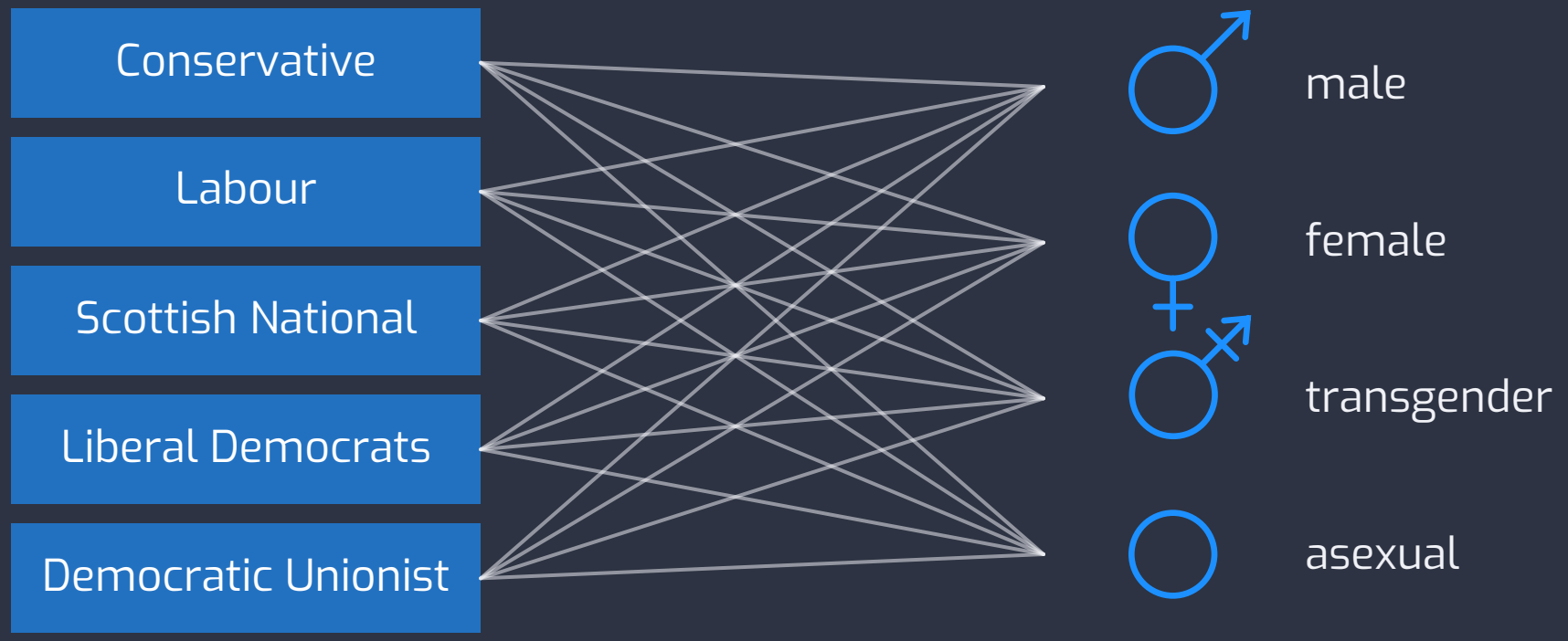
Cornwall	$\longrightarrow$	[0, 1, 0, 0, 0, 0, 0, 0, 0]	
Dorset	$\longrightarrow$	[0, 0, 0, 1, 0, 0, 0, 0, 0]	
Somerset	$\longrightarrow$	[0, 0, 0, 0, 0, 0, 1, 0, 0]	
Durham	$\longrightarrow$	[0, 0, 0, 0, 0, 0, 0, 0, 1]	("other")

Feature Crossing

Feature Crossing

EXAMPLE. Party and Gender

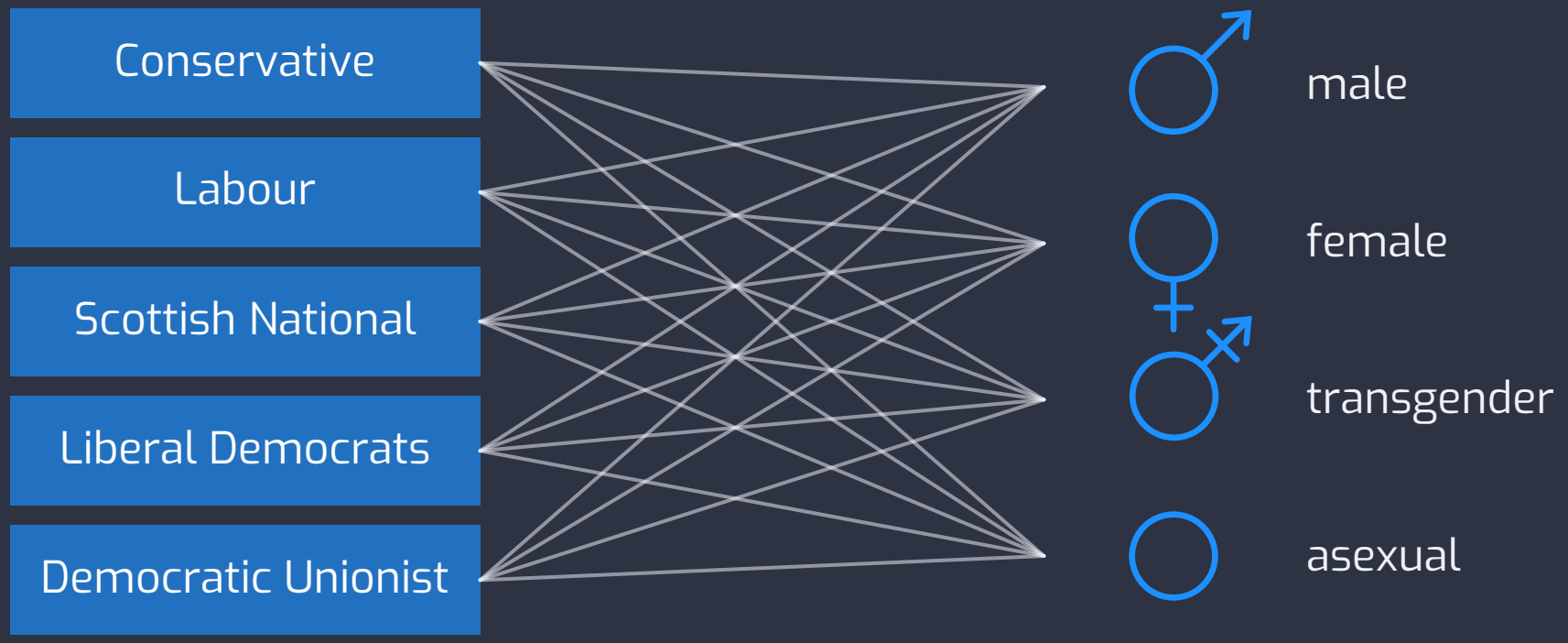
For each cross, we create a new true/false feature



Feature Crossing

EXAMPLE. Party and Gender

For each cross, we create a new true/false feature



Party_Gender

Conservative & male

Conservative & female

Conservative & transgender

Conservative & asexual

Labour & male

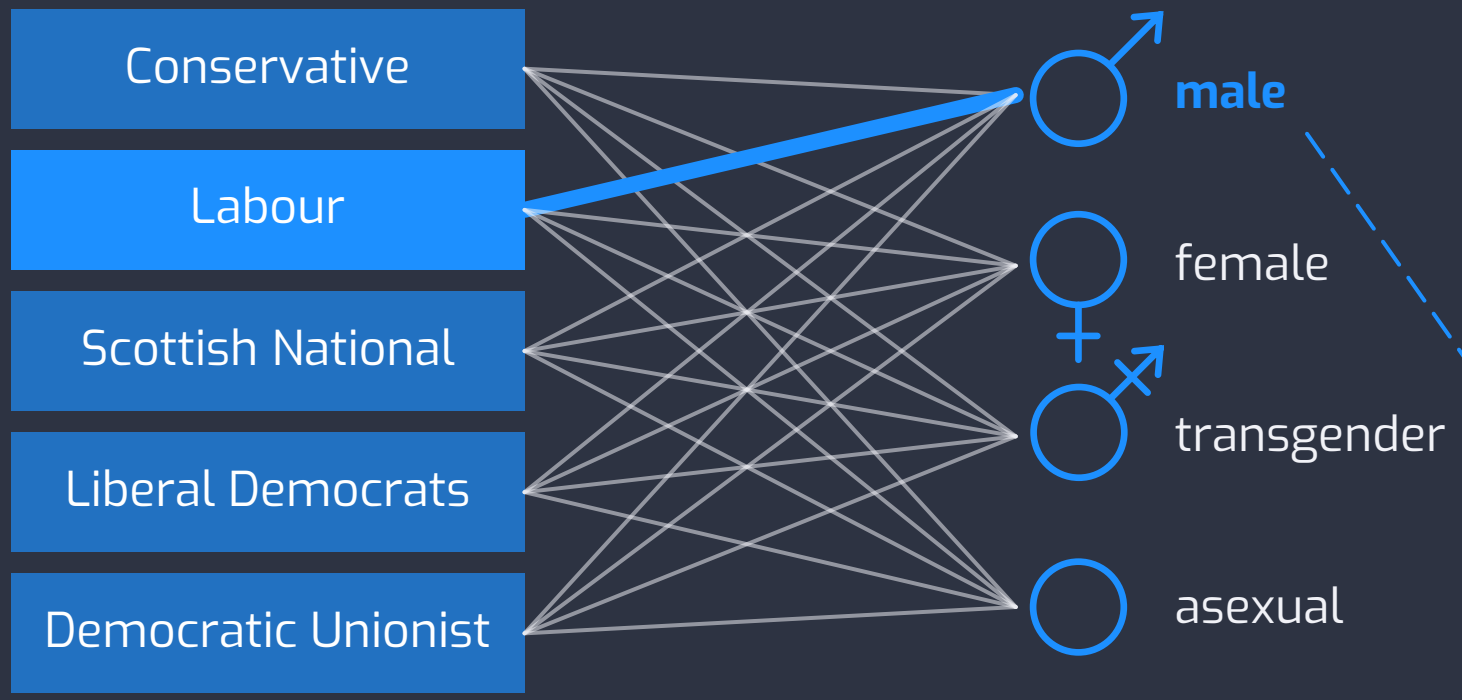
Labour & female

...

Feature Crossing

EXAMPLE. Party and Gender

For each cross, we create a new true/false feature



Party_Gender

Conservative & male

Conservative & female

Conservative & transgender

Conservative & asexual

Labour & male ——→ **(0,0,0,0,1,0,0,...)**

Labour & female

...

What if we have a categorical feature with a very large vocabulary?

Feature Hashing

Feature Hashing

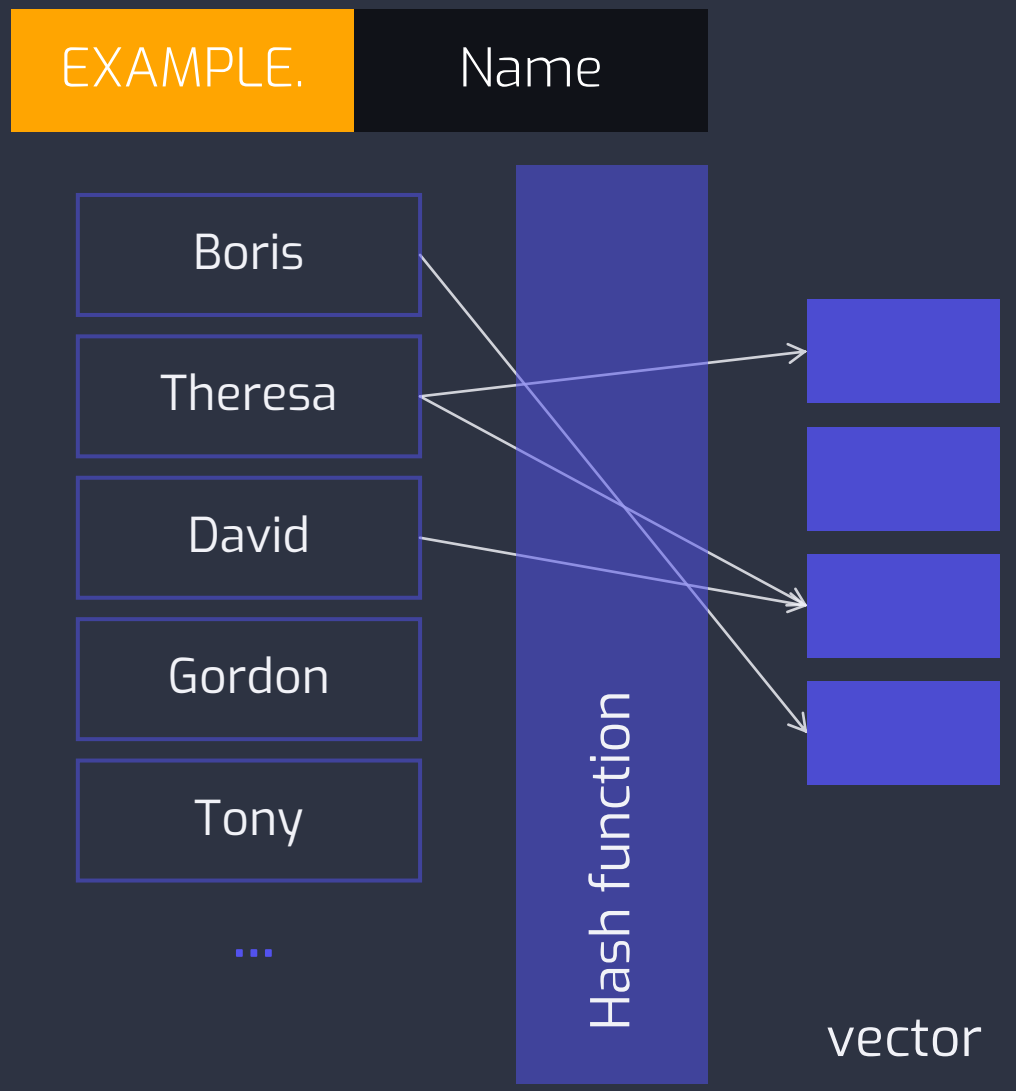
- It turns arbitrary features into a sparse binary vector.
- It can be very efficient by having a standalone hash function to make dataset easier to work with, since we don't have to build a vocabulary, which could be a very long list.

Feature Hashing

EXAMPLE.	Name

Boris	Harold	Neville	Archibald
Theresa	Edward	Stanley	William
David	Harold	Ramsay	Benjamin
Gordon	Alec	Bonar	Edward
Tony	Harold	Herbert	John
John	Anthony	Henry	George
Margaret	Winston	Arthur	Charles
James	Clement	Robert	...

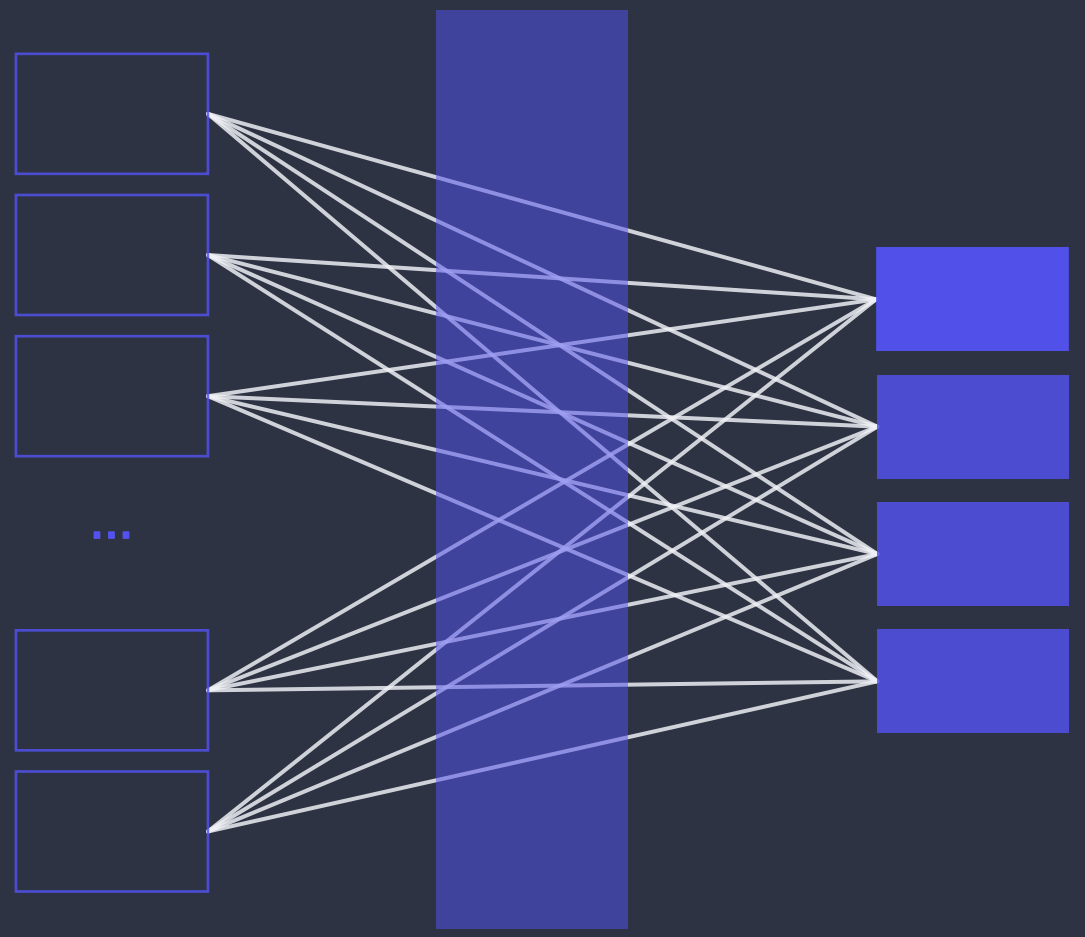
Feature Hashing



- The output range is smaller than the input range, so multiple categories may be mapped to the same hash output.
- Similar to one-hot encoding but with a control over the output dimensions, i.e., it limits the maximum number of possibilities.
- There could be collisions, i.e. different items area mapped to the same value.

Feature Hashing

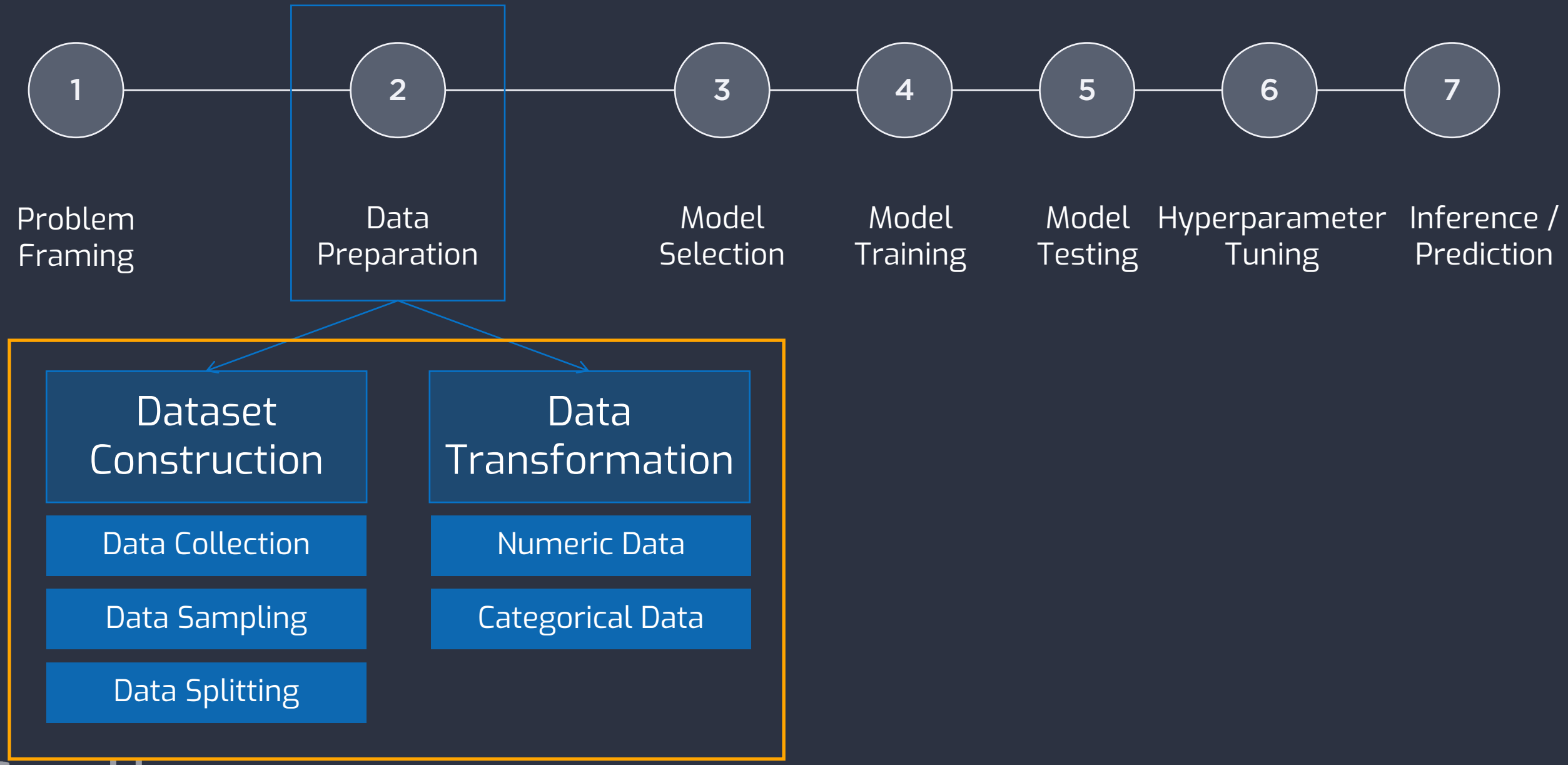
EXAMPLE.	Name

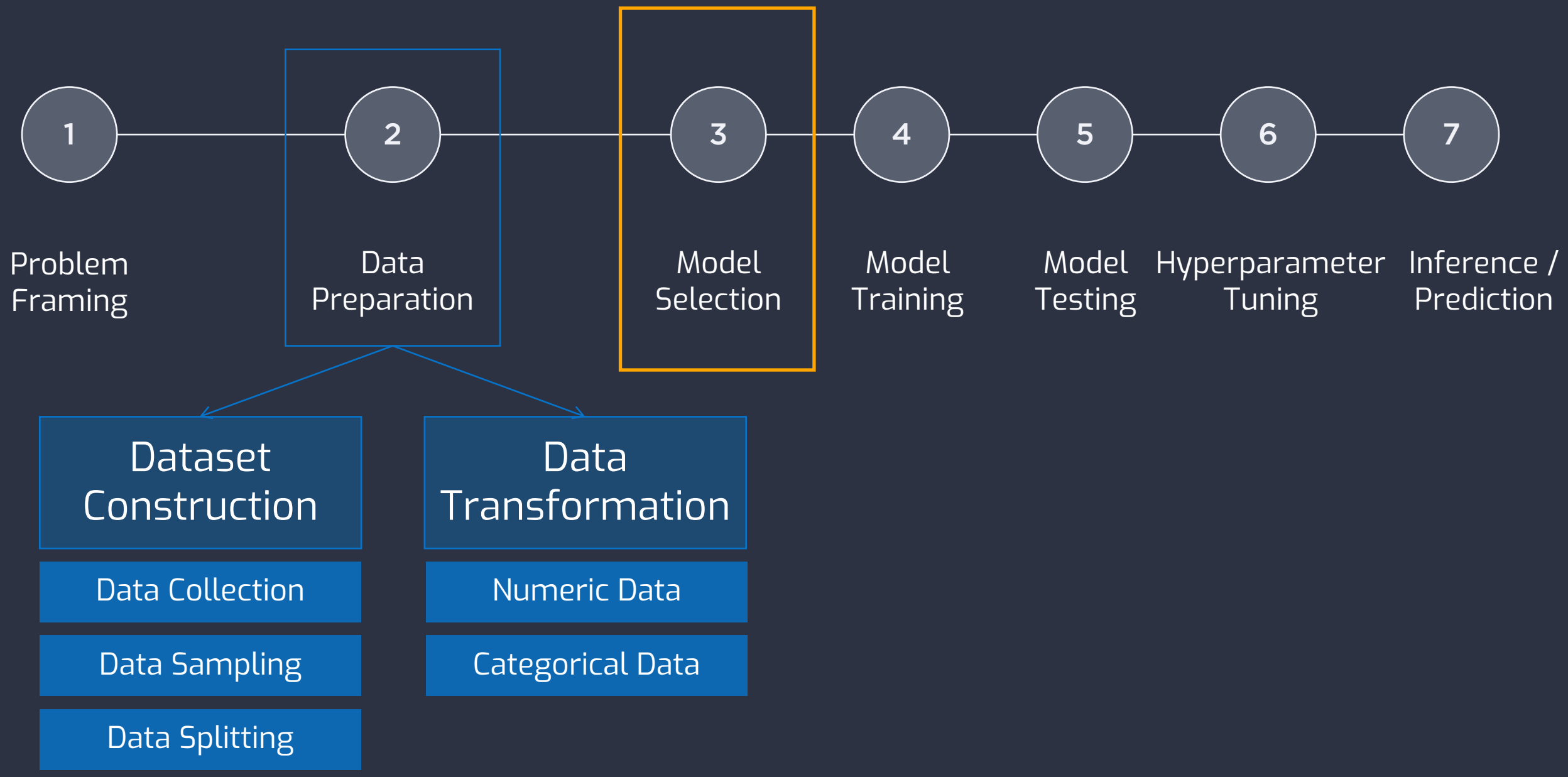


- It can be used to limit the maximum number of possibilities -> saving time!
- It can be used to limit memory usage, at the cost of adding some noise to our training data.

✓ Takeaway Points

- Categorical data needs to be transformed into numbers for learning algorithms.
- Ordinal categorical data can be transformed using ordinal encoding – values are comparable but not computable.
- Nominal categorical data can be transformed using one-hot encoding, feature crossing, and feature hashing.


1
2
3
4
5
6
7
Problem Framing
Data Preparation
Model Selection
Model Training
Model Testing
Hyperparameter Tuning
Inference / Prediction
Dataset Construction
Data Collection
Data Sampling
Data Splitting
Data Transformation
Numeric Data
Categorical Data


1
2
3
4
5
6
7
Problem Framing
Data Preparation
Model Selection
Model Training
Model Testing
Hyperparameter Tuning
Inference / Prediction
Dataset Construction
Data Collection
Data Sampling
Data Splitting
Data Transformation
Numeric Data
Categorical Data