



DEEP  
LEARNING  
INSTITUTE



DLI Accelerated Data Science Teaching Kit

# Lecture 3.1 - Introduction to Data Pre-processing

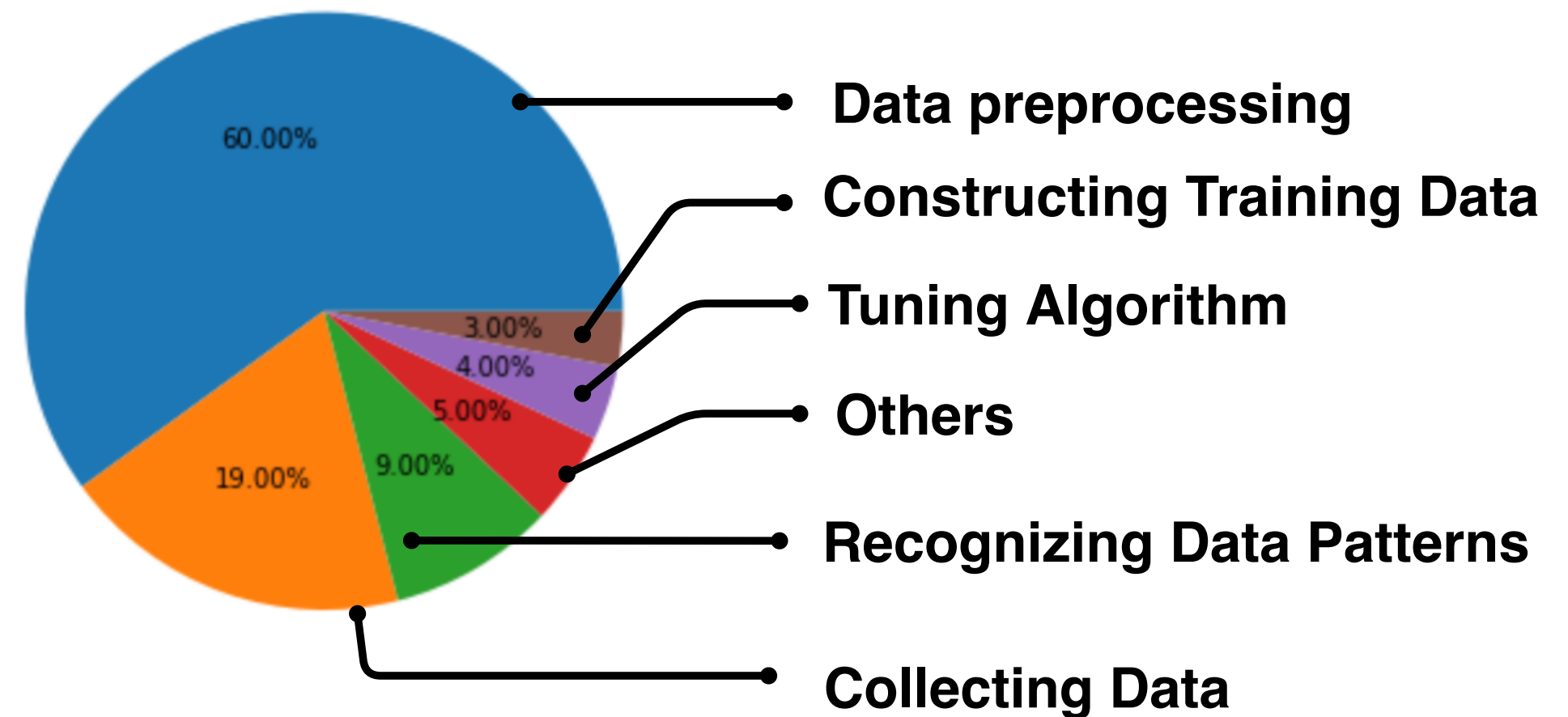


The Accelerated Data Science Teaching Kit is licensed by NVIDIA, Georgia Institute of Technology, and Prairie View A&M University under the [Creative Commons Attribution-NonCommercial 4.0 International License](https://creativecommons.org/licenses/by-nc/4.0/).

# Data Pre-processing

Where do data scientists spend their most of time for data analytics?

- Data scientists spend more than 50% of their time on **Data preprocessing**.
- **Collecting data** is the second most time-consuming component.
- **Tuning algorithm** occupies a small part.



# Why Do We Need Data Pre-processing?

Data is rarely clean and often has data quality issues.

- **Incomplete:** data lacks attributes or contains missing values.
- **Noisy:** data contains incorrect records such as outliers.
- **Inconsistent:** data contains conflicting records or discrepancies.

# Why Do We Need Data Pre-processing?

Data from the real world is dirty.

- **Missing values:** some attributes in the collected data would have blank or NULL values.
- **Invalid Values:** some well-know attributes such as gender may have incorrect values.
- **Uniqueness:** repeated values of the same identifiers.
- **Misspellings:** incorrectly written values
- ....

# Why Do We Need Data Pre-processing?

Data from the real world is dirty.

- Missing values
- Invalid Values
- Uniqueness
- Misspellings
- ....

#	Id	Name	Birthday	Gender	IsTeacher?	#Students	Country	City
1	111	John	31/12/1990	M	0	0	Ireland	Dublin
2	222	Mery	15/10/1978	F	1	15	Iceland	
3	333	Alice	19/04/2000	F	0	0	Spain	Madrid
4	444	Mark	01/11/1997	M	0	0	France	Paris
5	555	Alex	15/03/2000	A	1	23	Germany	Berlin
6	555	Peter	1983-12-01	M	1	10	Italy	Rome
7	777	Calvin	05/05/1995	M	0	0	Italy	Italy
8	888	Roxane	03/08/1948	F	0	0	Portugal	Lisbon
9	999	Anne	05/09/1992	F	0	5	Switzerland	Geneva
10	101010	Paul	14/11/1992	M	1	26	Ytali	Rome

Annotations:

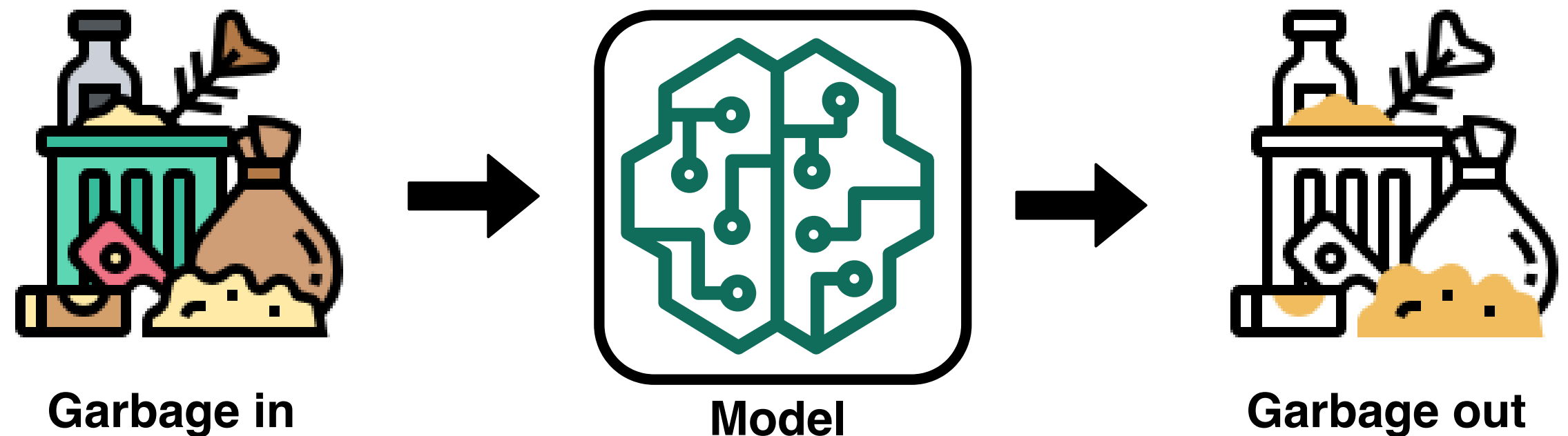
- Missing values: Arrow pointing to the empty City cell in row 2.
- Invalid values: Arrow pointing to the 'A' in the Gender cell of row 5.
- Misfielded values: Arrow pointing to the 'Italy' in the City cell of row 7.
- Misspellings: Arrow pointing to the 'Ytali' in the Country cell of row 10.
- Uniqueness: Arrow pointing to the '555' Id values in rows 5 and 6.
- Formats: Arrow pointing to the '1983-12-01' Birthday in row 6.
- Attribute dependencies: Arrow pointing to the '5' in the #Students cell of row 9.

An example of dirty data

# Goal of Data Pre-processing

To avoid “Garbage in, Garbage out” for data analytics

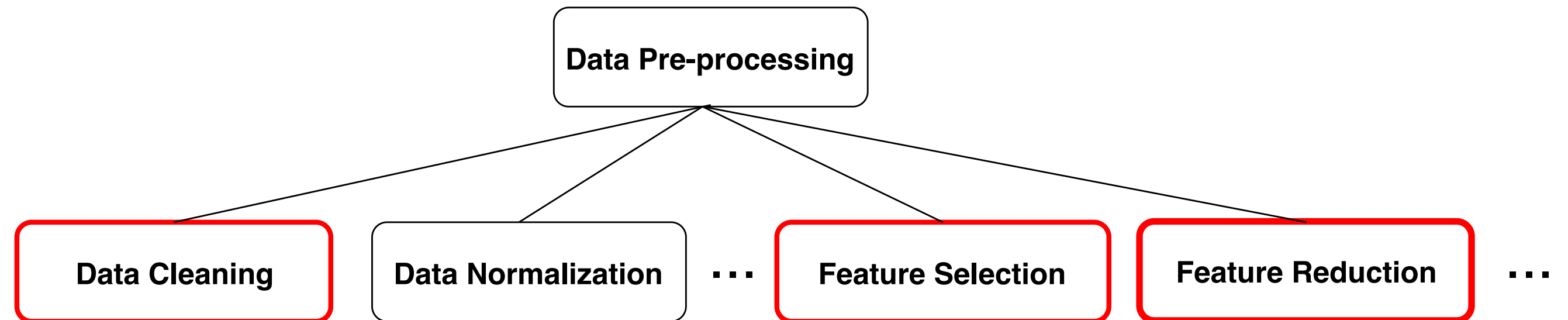
- Dirty data as input
- Useless results as output
- Meaningless work in model construction
- Failed projects/tasks



# Goal of Data Pre-processing

To avoid “Garbage in, Garbage out” for data analytics

- Data Cleaning and/or Statistical Preprocessing
- Feature Selection
- Feature Reduction
- ....







DEEP  
LEARNING  
INSTITUTE



DLI Accelerated Data Science Teaching Kit

# Thank You