

INT247

Machine Learning Foundations

Lecture #4.0

Data Pre-processing

Data

- Machine learning depends largely on test data.
- A large amount of data is required for ML.



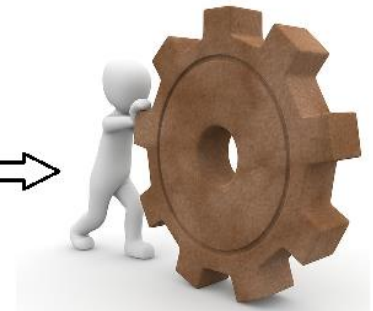
Raw Data



Data Cleaning



Feature Engineering



Model Building

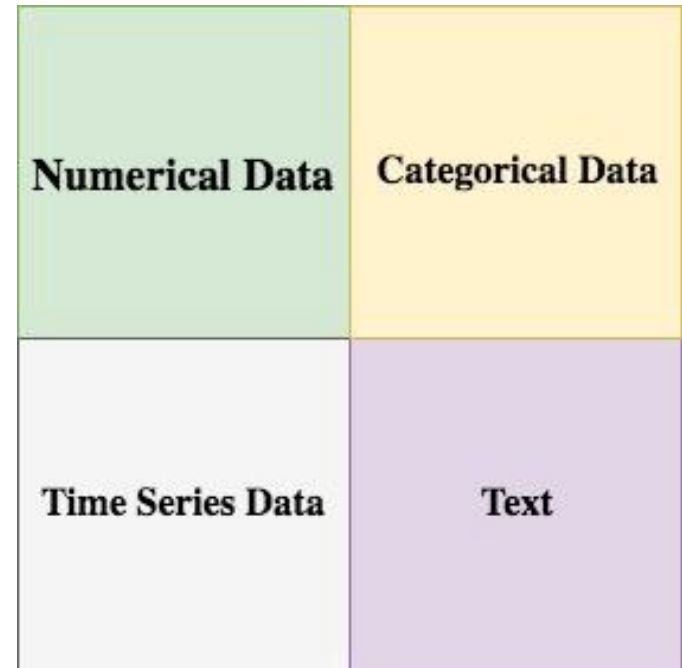


Data Preprocessing

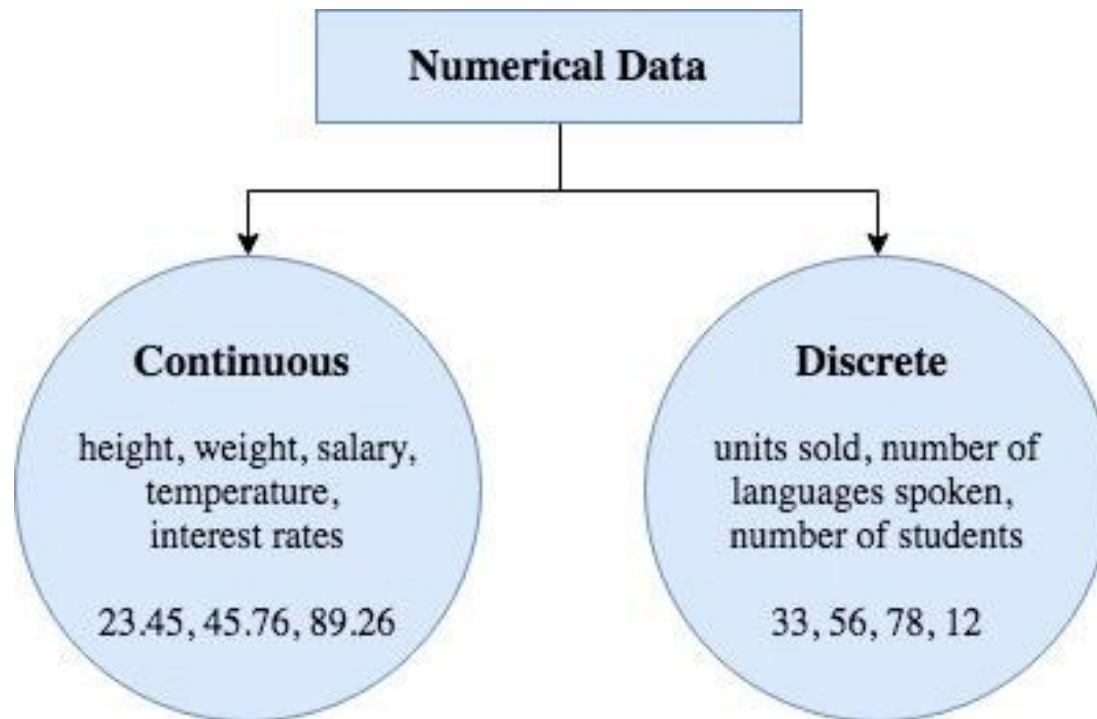
Different types of data in ML

Data can be categorized into 4 basic types:

- Numerical Data
- Categorical Data
- Time Series Data
- Text



Numerical Data



Categorical Data



beginner



intermediate



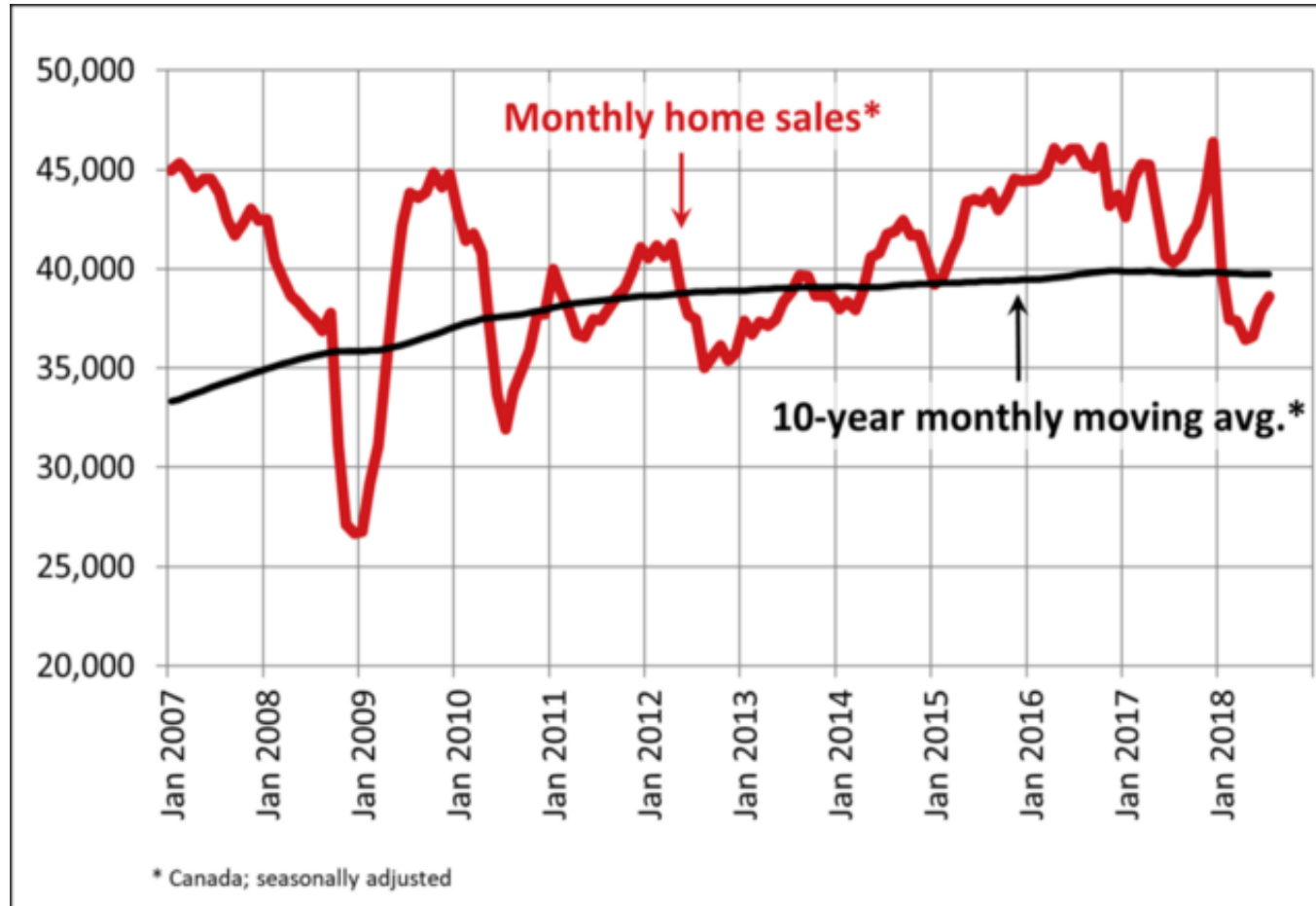
advanced

Population Bins

0 – 10 million, 10 – 100 million, 100 – 500 million, > 500 million

Ordinal Data

Time-Series Data





Data Preparation Process

- The process of data preparation comprises the following:
 - Data Selection
 - Data Pre-processing
 - Data Transformation



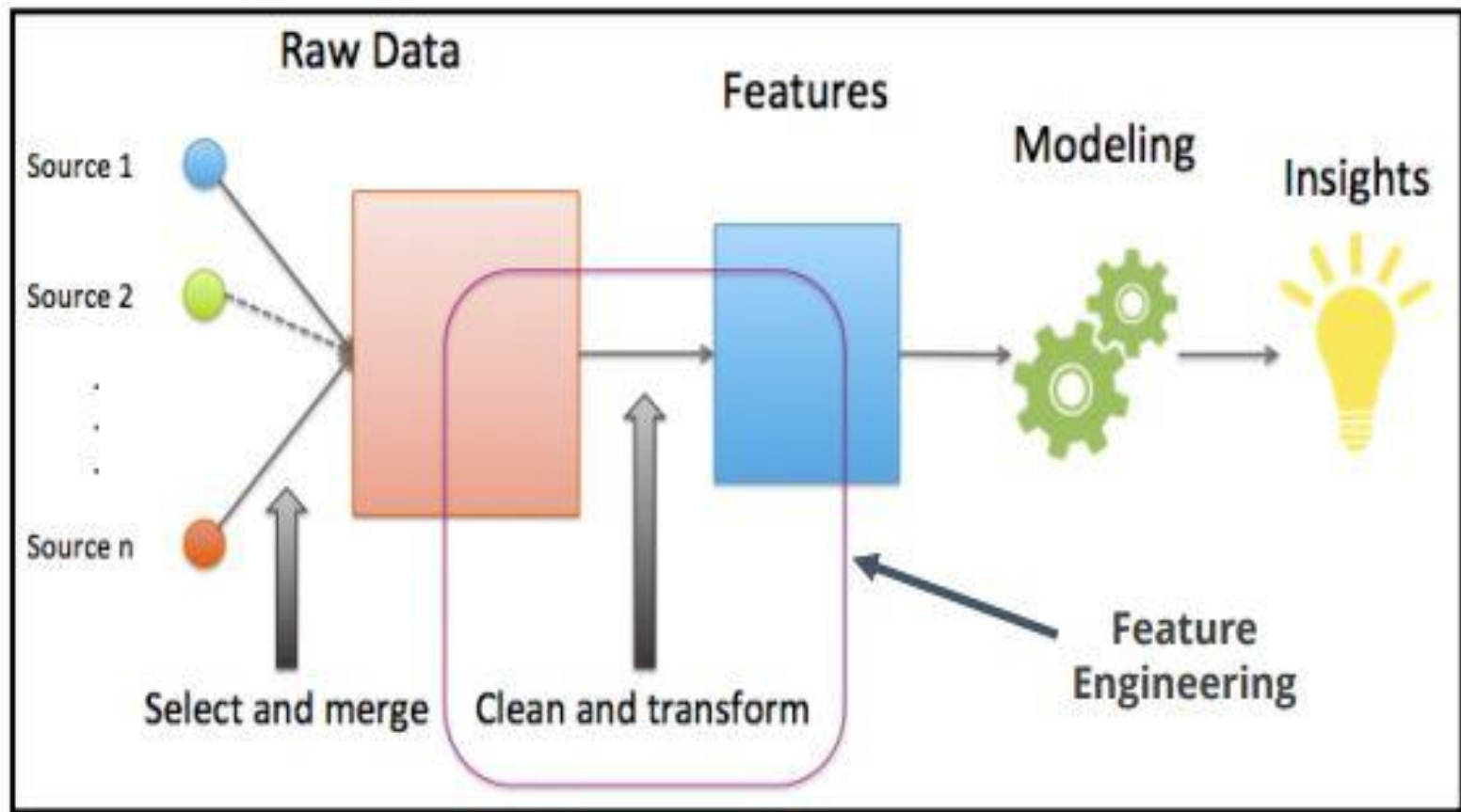
Data Selection

Steps involved in Data Selection involves:

- **Selecting only a subset of available data.**
- **The selected sample must be an accurate representation of the entire population.**
- **Some data can be derived or simulated from the available data if required.**
- **Data not relevant to the problem at hand can be excluded.**

Feature Engineering

- Transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data.



Framework of Feature Engineering


- **Frame your problem:**
 - Can you frame your problem in a way that machine learning could be useful. Eg: prediction.
- **Understand your data:**
 - What data will be most helpful to understand and generate a better understanding of the problem.
- **Frame your feature goals:**
 - What are you optimizing for?
 - Iteration speed
 - Model performance
- **Test, Iterate, Test Again:**
 - Check your choices for robustness.
 - Validate

Aspects of Feature Engineering

- Feature Selection
- Feature Extraction
- Feature Addition
- Feature Filtering

Example: flight date time vs status

- Status of flight depends on the hour of the day, not on the date-time.

	Date_Time_Combined	Status		Hour_Of_Day	Status
0	2018-02-14 20:40	Delayed		20	Delayed
1	2018-02-15 10:30	On Time		10	On Time
2	2018-02-14 07:40	On Time		7	On Time
3	2018-02-15 18:10	Delayed		18	Delayed
4	2018-02-14 10:20	On Time		10	On Time

Creating new feature “Hour_of_Day” is the feature engineering.

Feature Selection vs Feature Extraction vs Feature Engineering

- ***Feature selection*** is essential for creating the dataset.
- ***Feature extraction*** applies automatic methods like PCA for constructing new features.
- ***Feature engineering*** deals with the manual construction of features from raw data.

COMING UP
