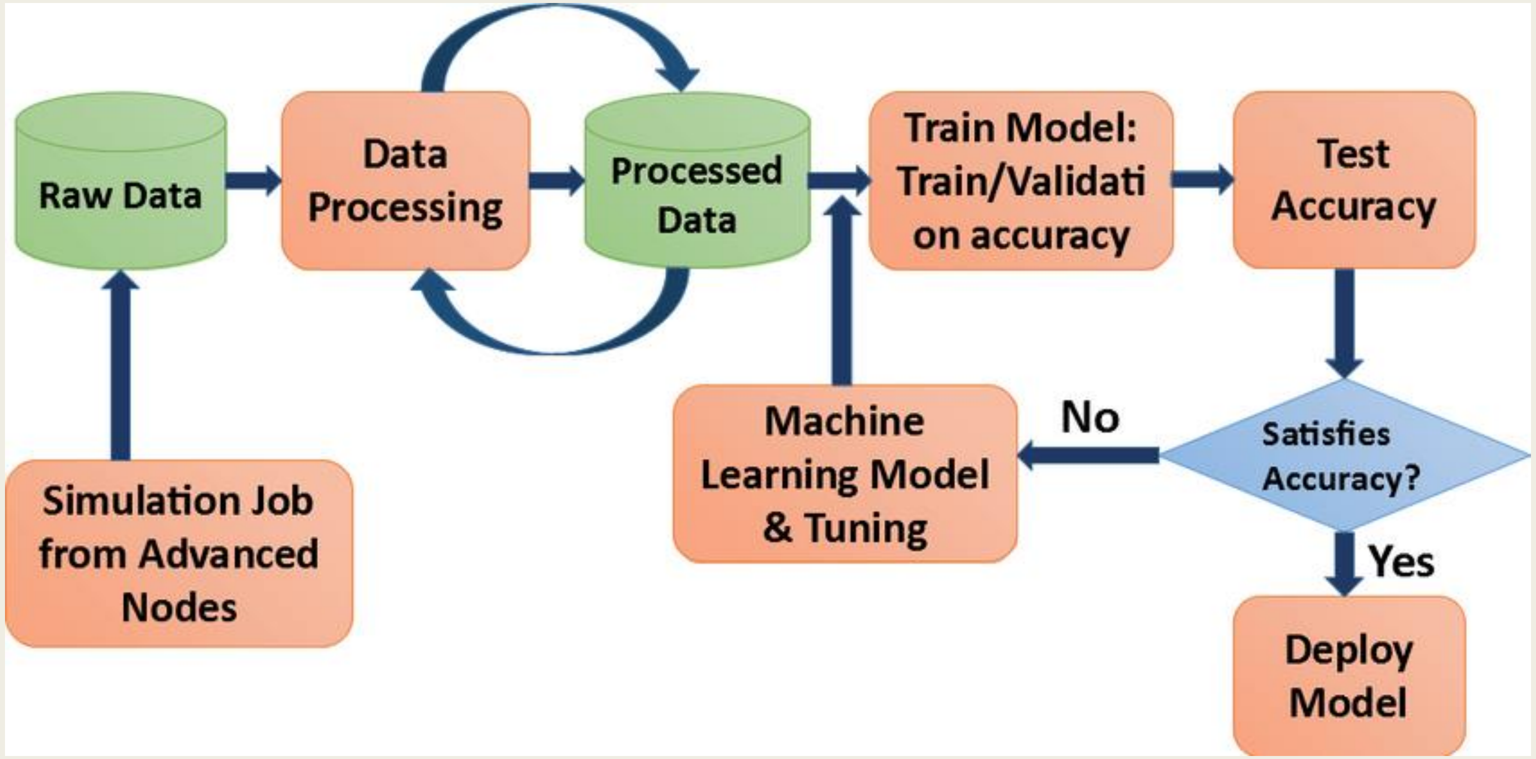
MACHINE LEARNING WORKFLOW

AI Summer School

University of Tehran

Machine Learning Workflow



Raw Data

Data Processing

Processed Data

Train Model: Train/Validati on accuracy

Test Accuracy

Satisfies Accuracy?

No

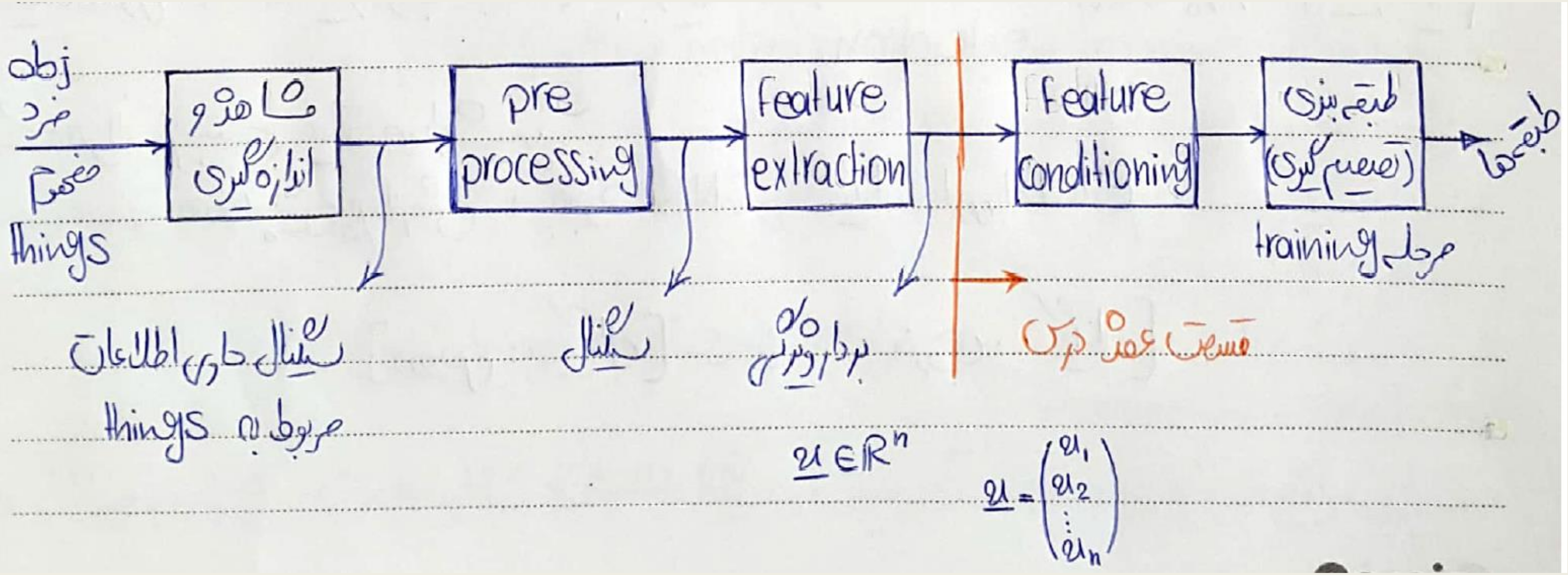
Yes

Machine Learning Model & Tuning

Simulation Job from Advanced Nodes

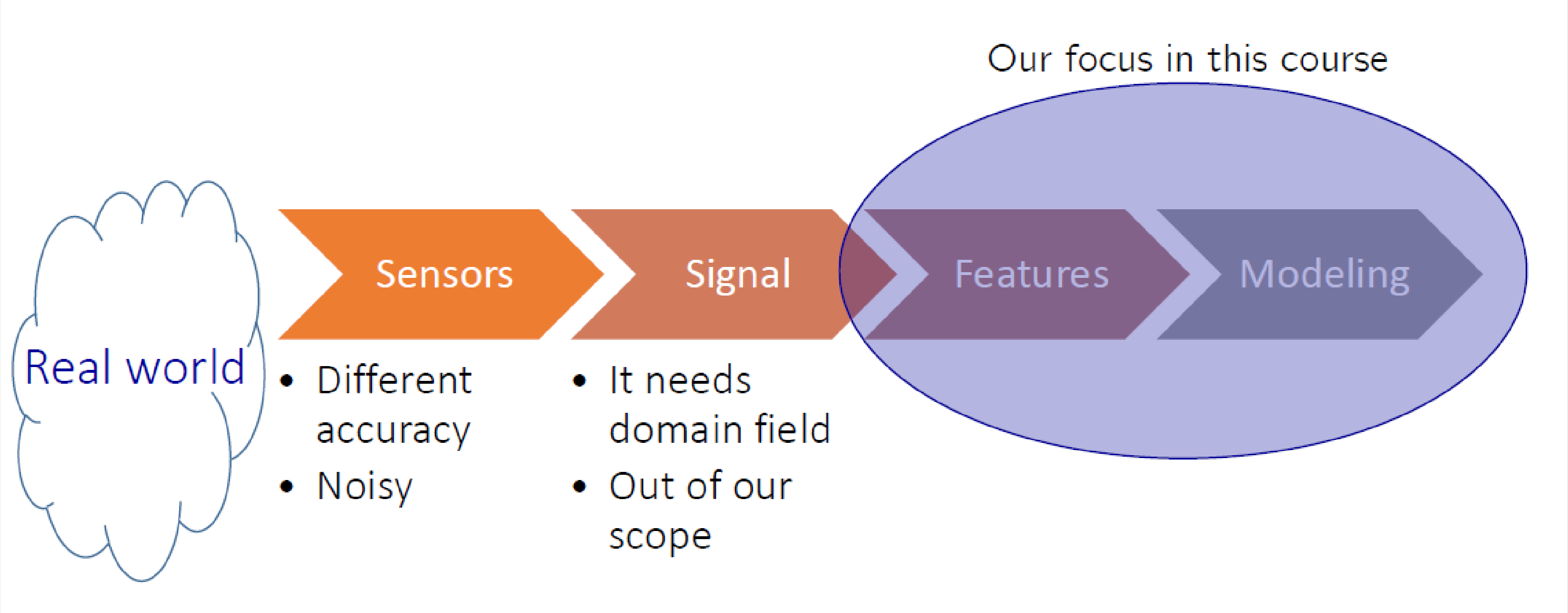
Deploy Model

Machine Learning Workflow



Data Gathering

- Useful features
- Dataset size
- Data collection process

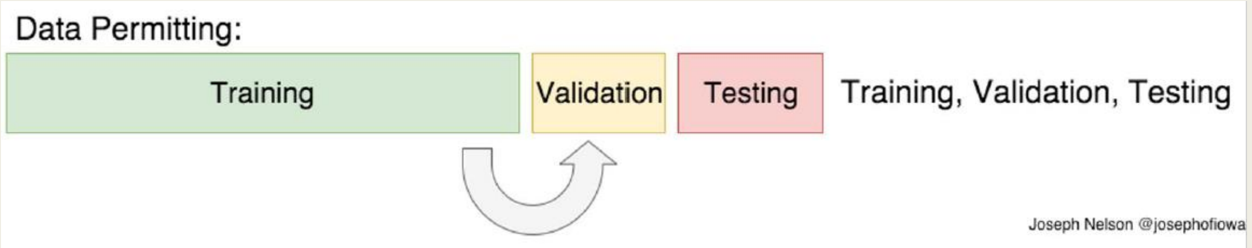


Data Processing

- Train / Validation / Test split

- Data Visualization
  - *Class label distribution*
  - *Feature values distribution*
  - *Correlation between features*
  - *Separability of classes*

- Rescaling



Data Permitting:

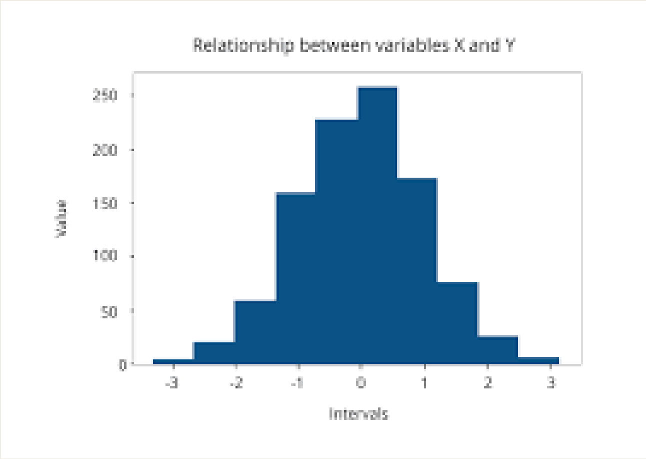
Training

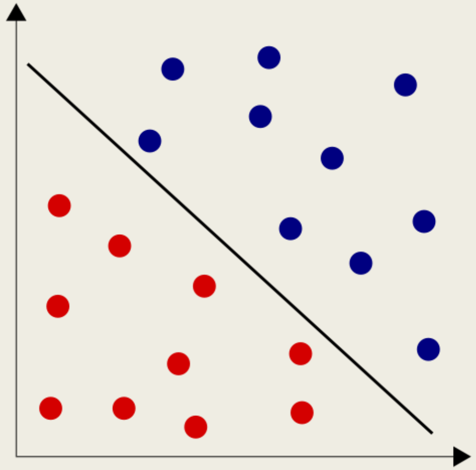
Validation

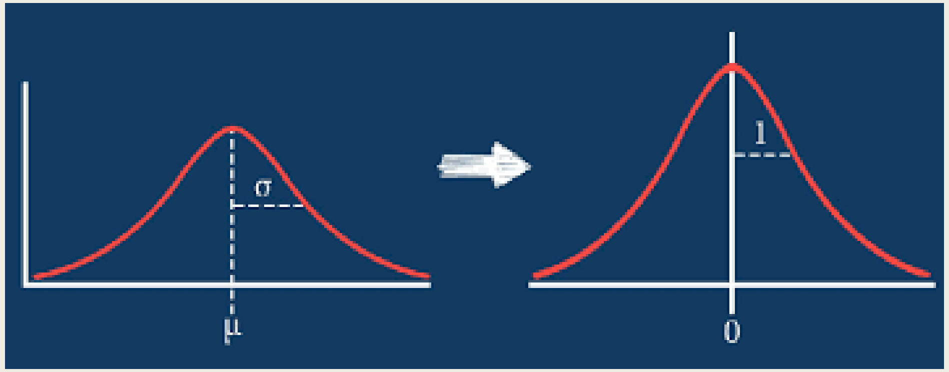
Testing

Training, Validation, Testing

Joseph Nelson @josephofiowa





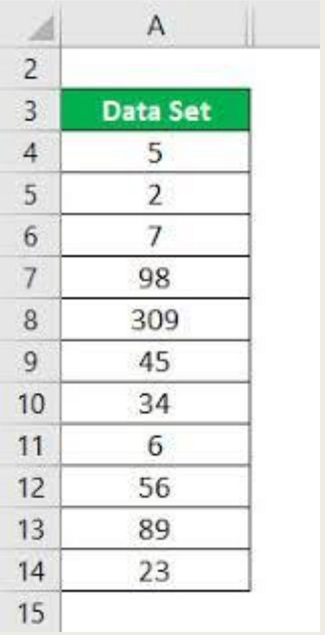


Data Processing

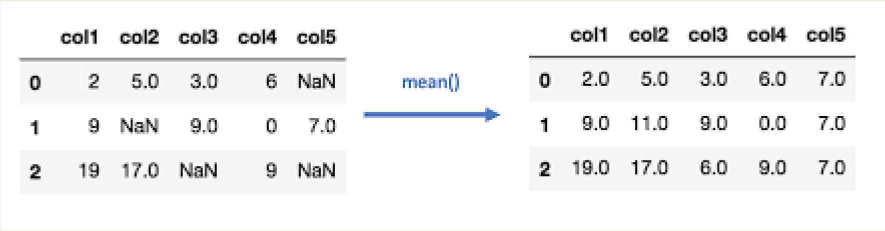
- Error correction
  - *Handling missing values*
  - *Handling inconsistent values*
  - *Handing outliers*

- Handling categorical features

- Feature engineering



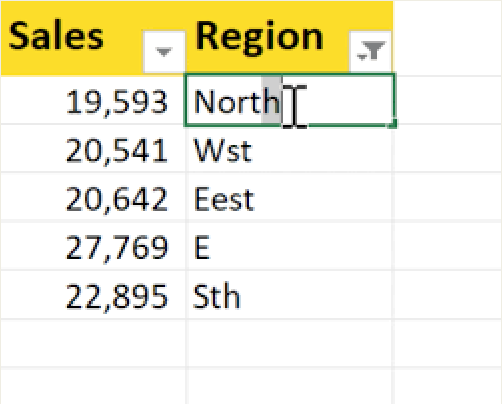
	A
2	
3	Data Set
4	5
5	2
6	7
7	98
8	309
9	45
10	34
11	6
12	56
13	89
14	23
15	



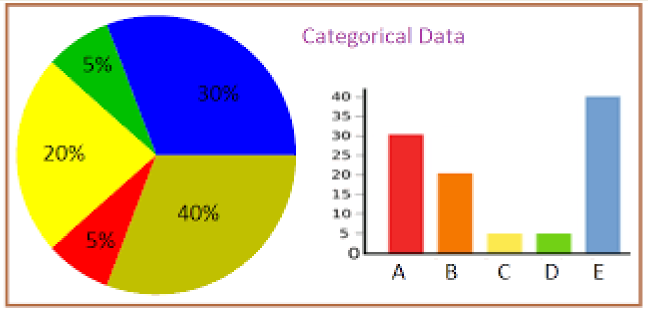
	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

mean()

	col1	col2	col3	col4	col5
0	2.0	5.0	3.0	6.0	7.0
1	9.0	11.0	9.0	0.0	7.0
2	19.0	17.0	6.0	9.0	7.0



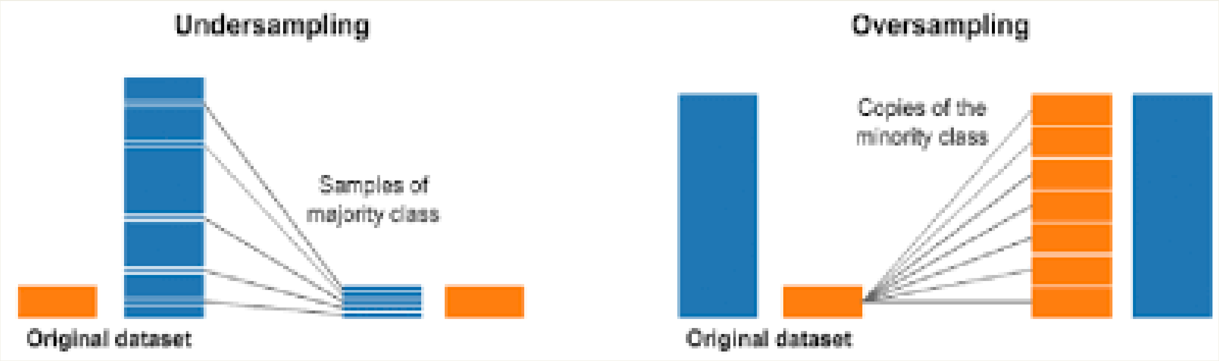
Sales	Region
19,593	North
20,541	Wst
20,642	Eest
27,769	E
22,895	Sth

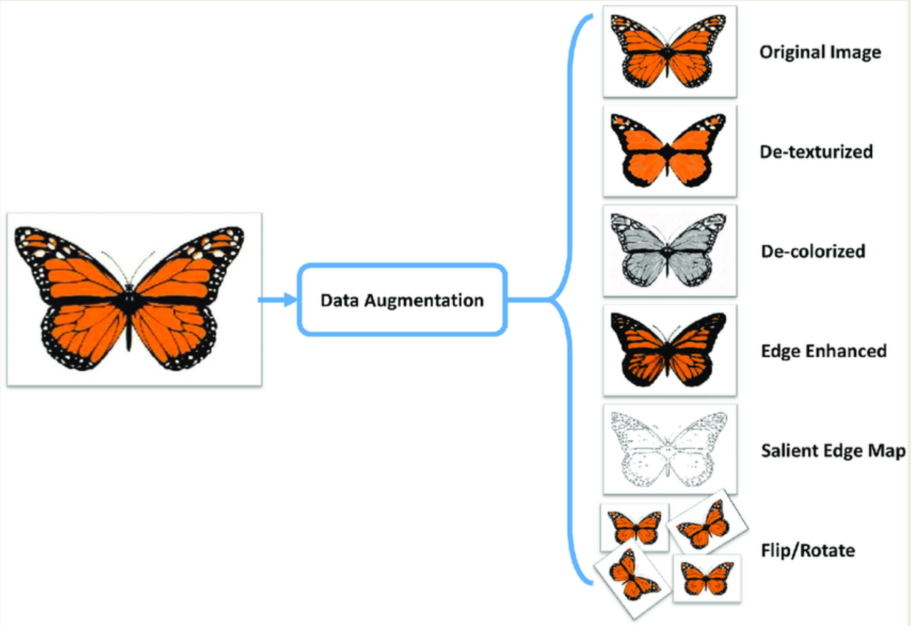


Data Processing

- Dealing with imbalanced classes:
  - *Oversampling*
  - *Undersampling*

- Data augmentation





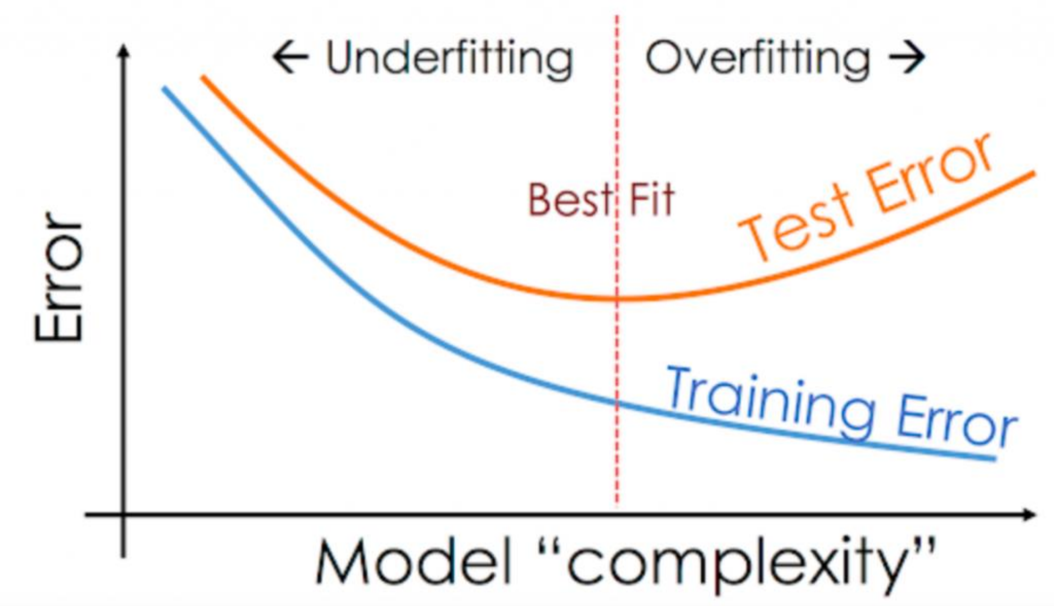
Data Processing

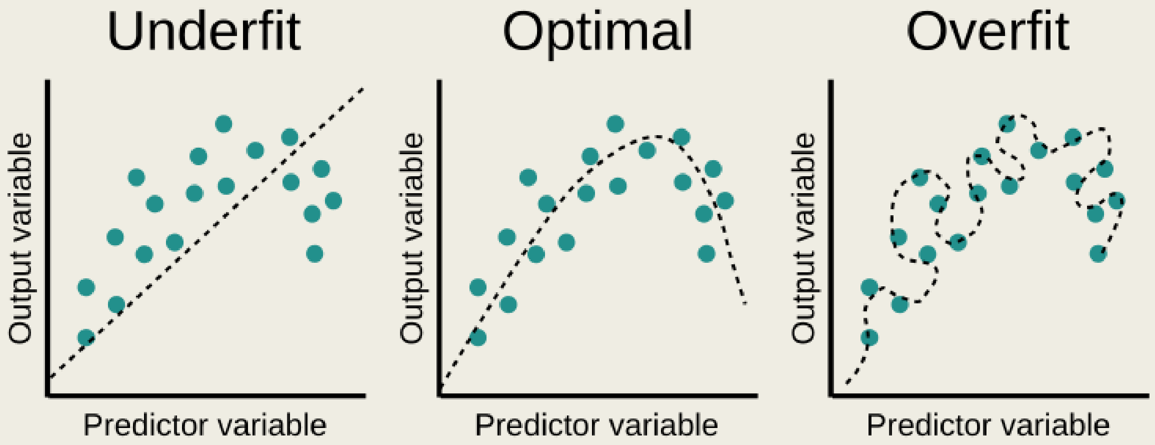
- Preprocessing for text:
  - *Normalization*
  - *Tokenization*
  - *Stop word deletion*
  - *Stemming*
  - *…*
- Preprocessing for image:
  - *Resizing*
  - *Converting the format: RGB, HSB, gray scale*
  - *Normalization*
  - *….*

Training and Evaluation

- Model selection
- Hyper parameter tuning
- Selecting proper evaluation metric
- Fighting with overfitting

Error Analysis


← Underfitting
Overfitting →
Error
Best Fit
Test Error
Training Error
Model "complexity"


Underfit
Optimal
Overfit
Output variable
Predictor variable
Output variable
Predictor variable
Output variable
Predictor variable