MachineLearnAthon - Microlecture
What is Classification?

Classification

15.10.2024

Learning outcomes of today

After successfully completing this micro-lecture, you are able to….

- Identify the type of a machine learning problem
- Understand the stages of building a machine learning pipeline
- Apply the proper steps for machine learning modeling

Agenda for today

- Types of Machine Learning Tasks
- Stages of Supervised Machine Learning Pipeline
- Proper Machine Learning Modeling

Types of Machine Learning Tasks

	Data	No Data
Target		
No Target		

Types of Machine Learning Tasks

	Data	No Data
Target	Supervised Learning • Classification • Regression	
No Target		

Types of Machine Learning Tasks: Supervised Classification

Examples:

- Traffic Sign Recognition
  - Data: Images
  - Target: Labels **(Label is discrete variable)**

Types of Machine Learning Tasks: Supervised Regression

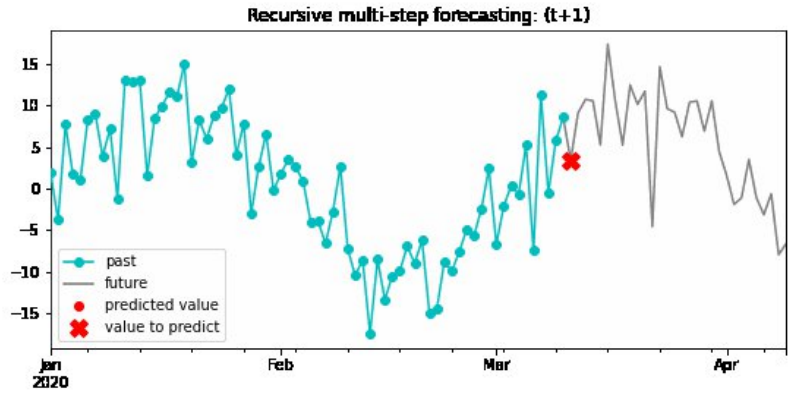
Examples:

- Rental Price Regression
  - Data:
    - No. of Rooms
    - District
    - Floor
    - Area
  - Target:
    - Rental Price **(Target is continuous variable)**

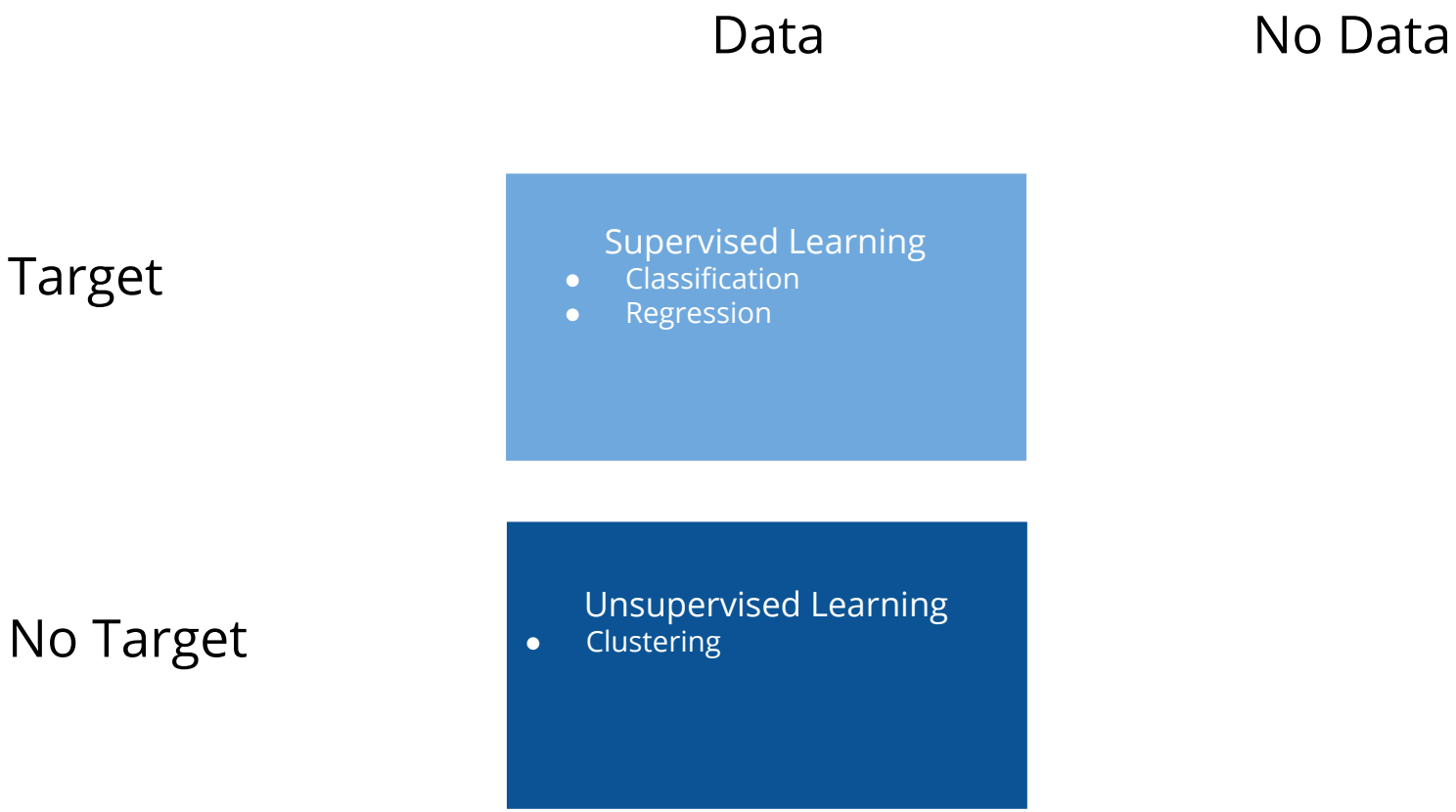
Types of Machine Learning Tasks: Supervised Regression

Examples:

- Weather Forecasting
  - Data:
    - History of rain
    - History of temperature
    - History of Humidity
    - Month of Year
  - Target:
    - Any historical value given the past values to this one. **(Target is continuous variable)**



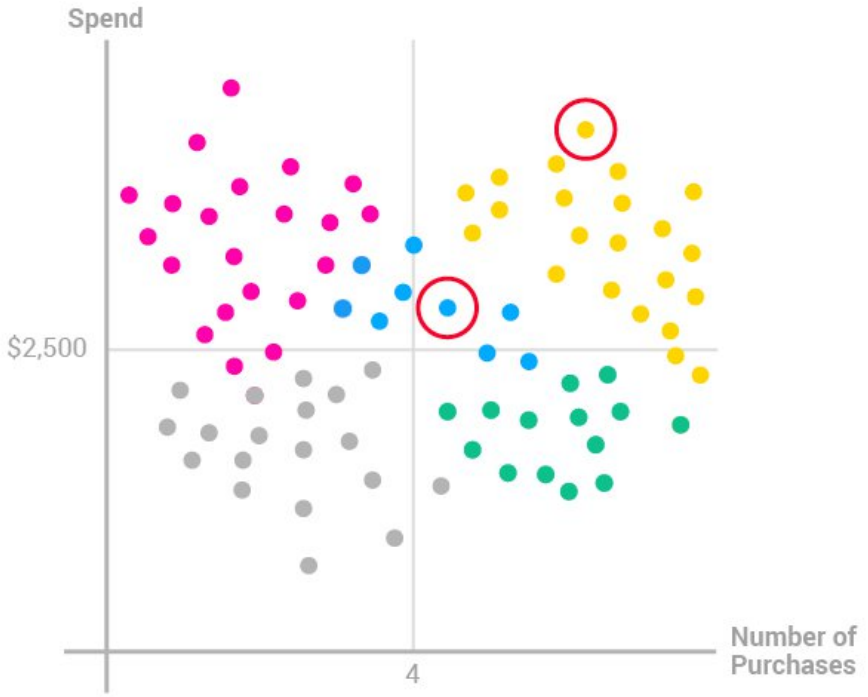
Types of Machine Learning Tasks



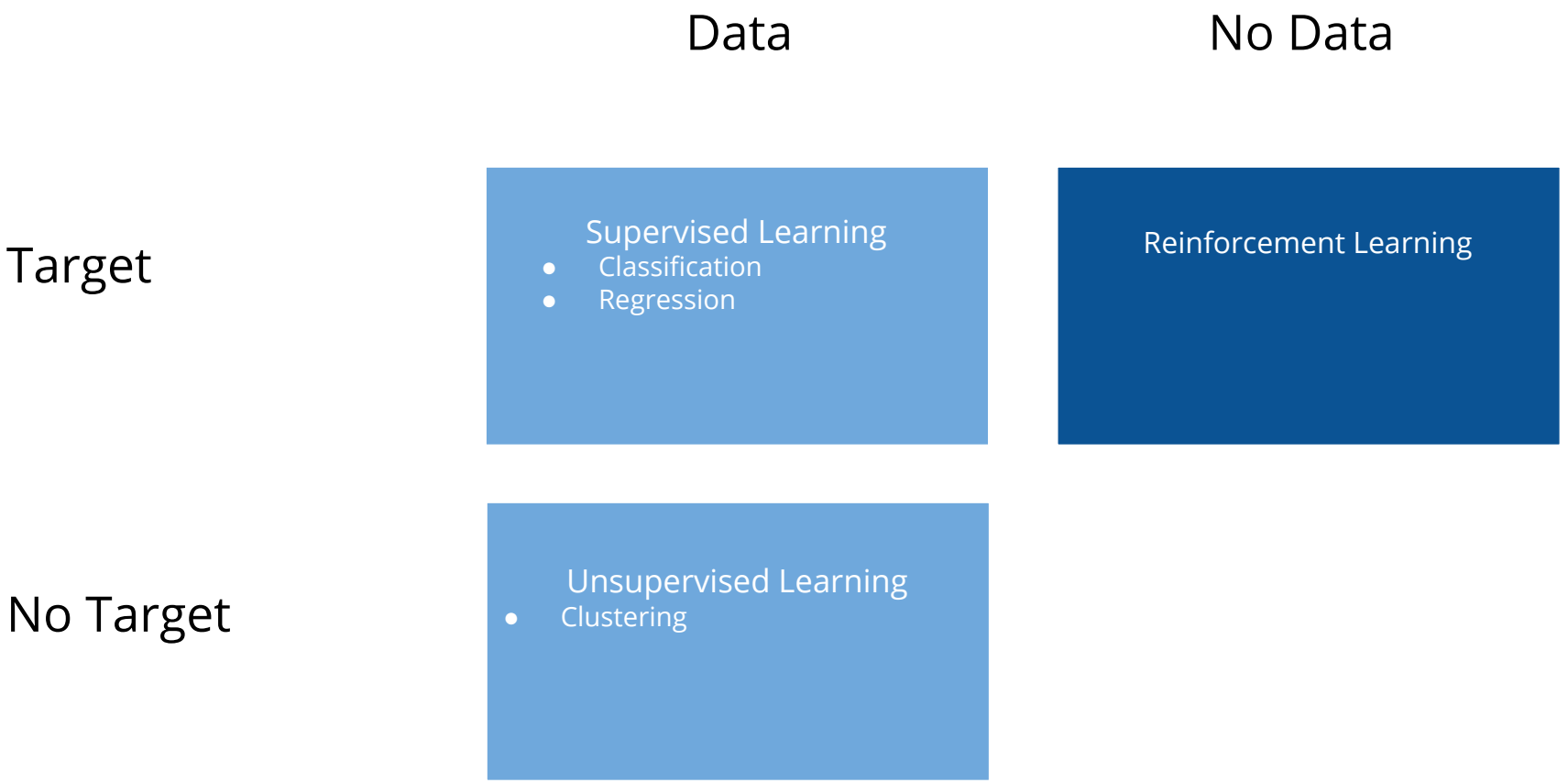
Types of Machine Learning Tasks: Unsupervised

Examples:

- Customer Segmentation
  - Data:
    - No. of Purchases
    - Average Spend



Types of Machine Learning Tasks

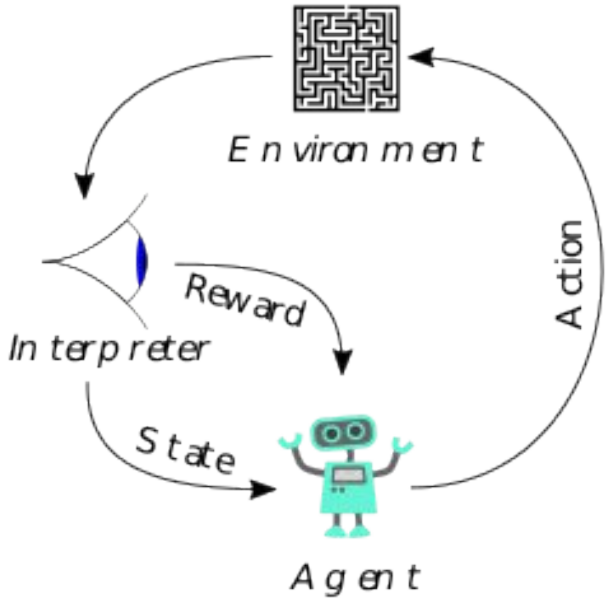


Types of Machine Learning Tasks: Reinforcement Learning

Examples:

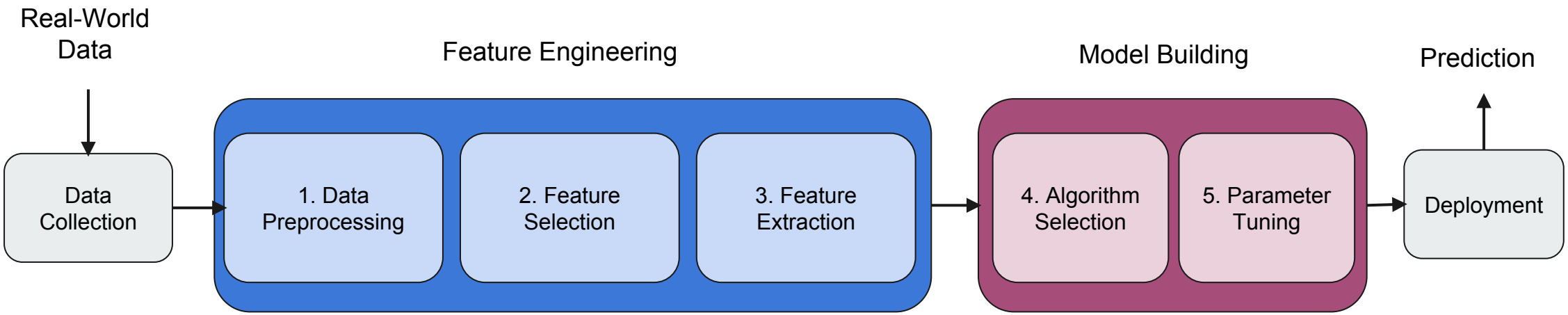
- Ads Recommendation
  - Target: Action
    - User clicks the Ad = Reward
    - User doesn't click the Ad = Punishment

Make Mistakes and learn from them.
Requires Huge amount of Trials = Collecting Data (States)

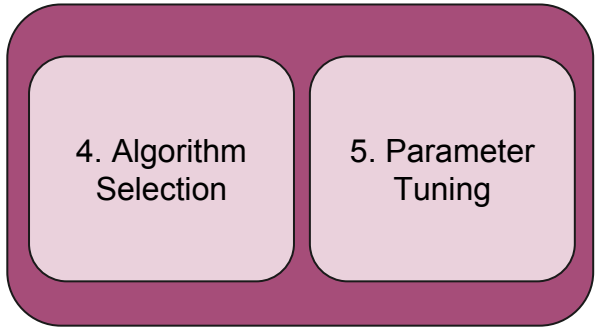


Ref: https://neptune.ai/blog/reinforcement-learning-applications

Supervised ML Pipeline



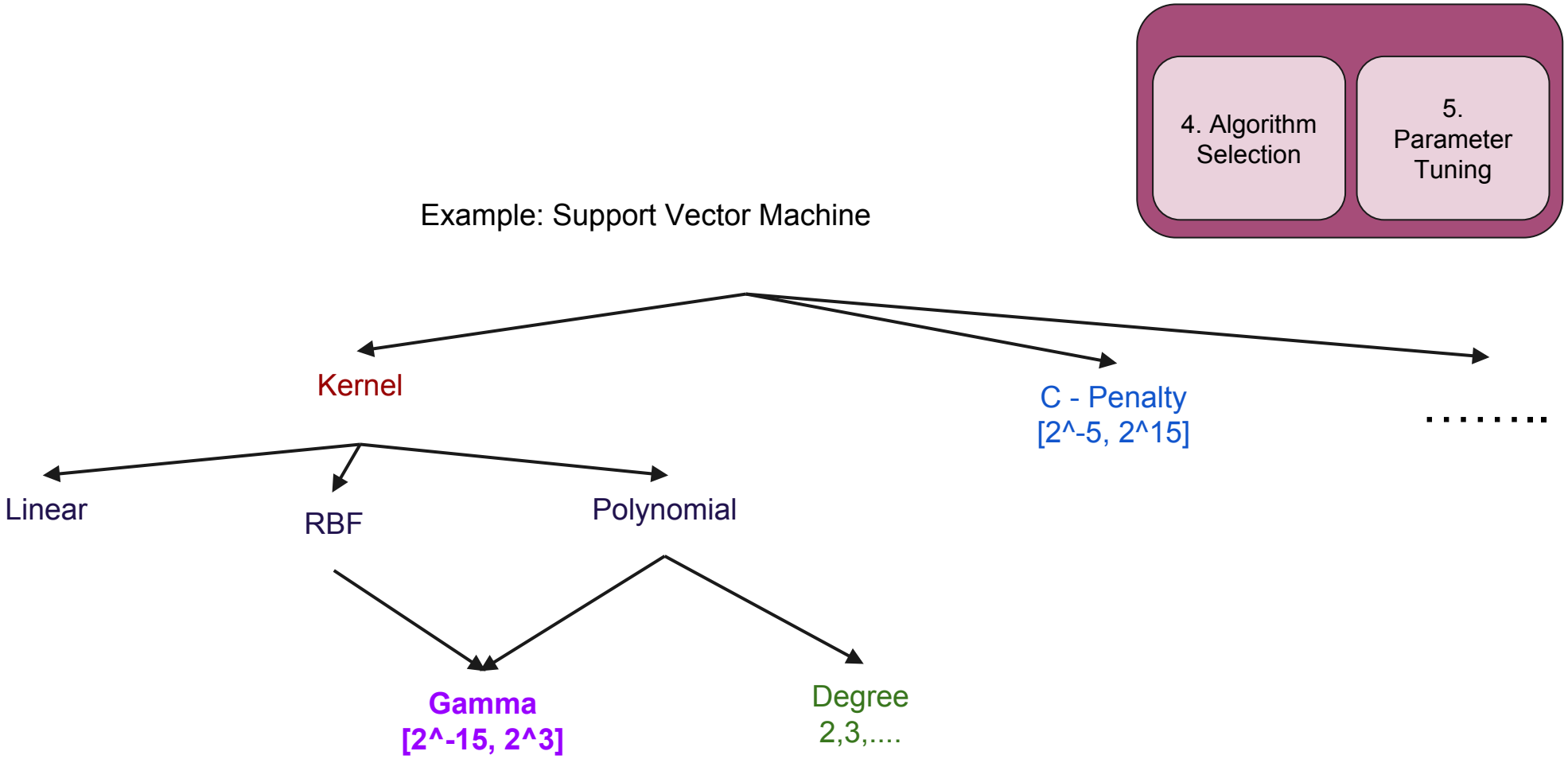
Supervised ML Pipeline



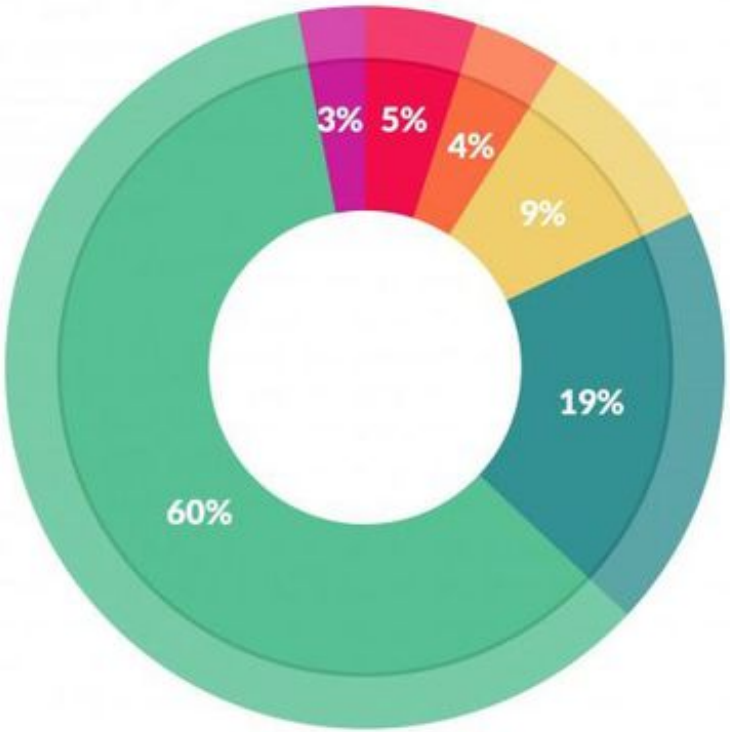
Examples:

- Linear Classification: (Simple Linear Classification, Ridge, Lasso, Simple Perceptron, ….)
- Support Vector Machines
- Decision Tree (ID3, C4.5, C5.0, CART, ….)
- Nearest Neighbors
- Gaussian Processes
- Naive Bayes (Gaussian, Bernoulli, Complement, ….)
- Ensembling: (Random Forest, GBM, AdaBoost, ….)

Supervised ML Pipeline



Supervised ML Pipeline



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

Forbes: Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says

Supervised ML Pipeline

1. Data Preprocessing

Different Scale → Normalization ??

Non-Numeric Values → Encoding ??

Missing Value → Imputation ??

	1	2	3	4	5	6	7	
Name	**Sex**	**Fold**	**Pulse**	**Age**	**Clap**	**Exer**	**Smoke**	**H**
Label								
Type	Factor	Factor	Number	Number	Factor	Factor	Factor	N
Format								
Levels	Female#...	L on R#,...			Left#,#...	Freq#,#...	Heavy#,...	
1	Female	R on L	92	18.25	Left	Some	Never	
2	Male	R on L	104	17.583	Left	None	Regul	
3	Male	L on R	87	16.917	Neither	None	Occas	
4	Male	R on L		20.333	Neither	None	Never	
5	Male	Neither	35	23.667	Right	Some	Never	
6	Female	L on R	64	21	Right	Some	Never	
7	Male	L on R	83	18.833	Right	Freq	Never	
8	Female	R on L	74	35.833	Right	Freq	Never	
9	Male	R on L	72	19	Right	Some	Never	
10	Male	R on L	90	22.333	Right	Some	Never	

Supervised ML Pipeline

1. Data Preprocessing

	1	2	3	4	5	6	7
Name	**Sex**	**Fold**	**Pulse**	**Age**	**Clap**	**Exer**	**Smoke**
Label							
Type	Factor	Factor	Number	Number	Factor	Factor	Factor
Format							
Levels	Female#...	L on R#,...			Left#,#...	Freq#,#...	Heavy#,...
1	Female	R on L	92	18.25	Left	Some	Never
2	Male	R on L	104	17.583	Left	None	Regul
3	Male	L on R	87	16.917	Neither	None	Occas
4	Male	R on L		20.333	Neither	None	Never
5	Male	Neither	35	23.667	Right	Some	Never
6	Female	L on R	64	21	Right	Some	Never
7	Male	L on R	83	18.833	Right	Freq	Never
8	Female	R on L	74	35.833	Right	Freq	Never
9	Male	R on L	72	19	Right	Some	Never
10	Male	R on L	90	22.333	Right	Some	Never

Non-Numeric Values → Encoding??

Example

	Smoke
I1	**Never**
I2	**Never**
I3	**Occas**

Supervised ML Pipeline

1. Data Preprocessing

Examples of Data Preprocessors:

1. Scaling
2. Normalization
3. Standardization
4. Binarization
5. Imputation
6. Deletion
7. One-Hot-Encoding
8. Hashing
9. Discretization

Supervised ML Pipeline

2. Feature Selection

Example: Feature Selection: *Univariate* Feature Selection (Fast):

Best Two Features → They are the same!!

Age	Year of Birth	Diabetes	Blood Pressure	Early Bird/ Night Owl	Smoker	**Mortality (Class Labels)**
20	1999	Yes	Normal	Night Owl	No	**Low**
80	1939	No	Normal	Early Bird	No	**High**

Supervised ML Pipeline

2. Feature Selection

Example: Feature Selection: *Multivariate* Feature Selection (Slow):

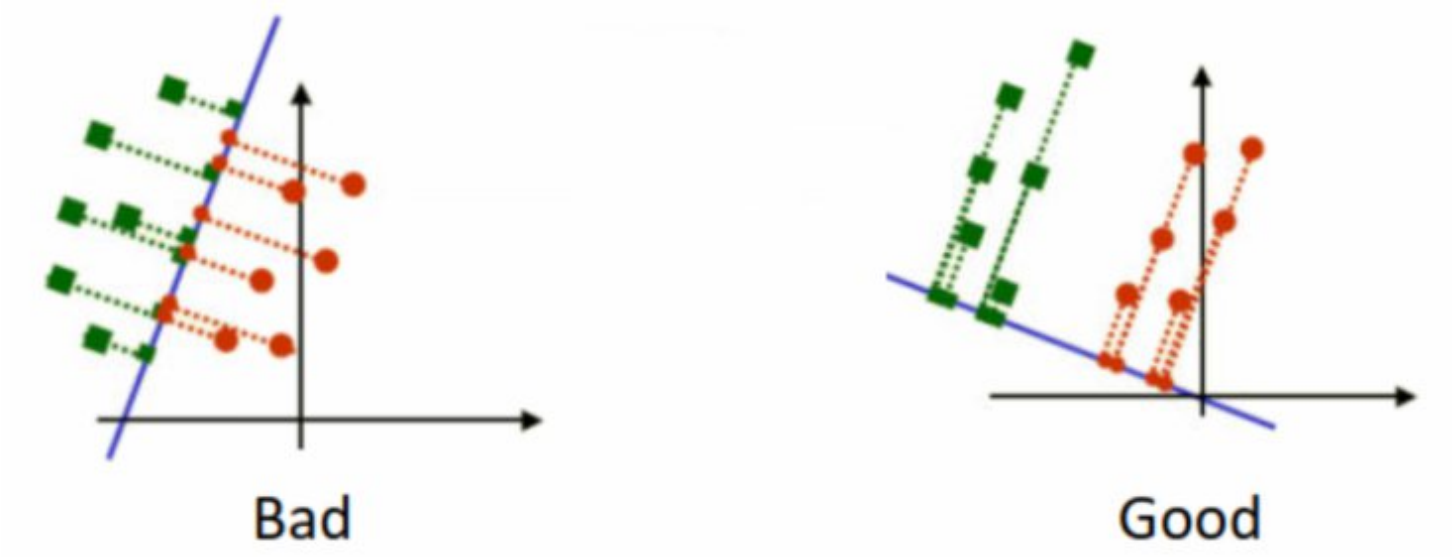
- Are we going to try every possible set of features?
- How many features are enough?

Age	Year of Birth	Diabetes	Blood Pressure	Early Bird/ Night Owl	Smoker	**Mortality (Class Labels)**
20	1999	Yes	Normal	Night Owl	No	**Low**
80	1939	No	Normal	Early Bird	No	**High**

Supervised ML Pipeline

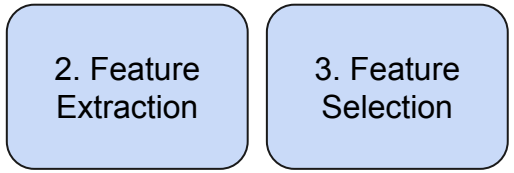
3. Feature Extraction

Example: Feature Extraction: Principal Component Analysis:
How to **reduce** dataset dimensions while keeping as much **variation** as possible



Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4, No. 4, p. 738). New York: springer.

Supervised ML Pipeline



Examples of Feature Extraction:

1. Principal Component Analysis
2. Linear Discriminant Analysis
3. Multiple Discriminant Analysis
4. Independent Component Analysis

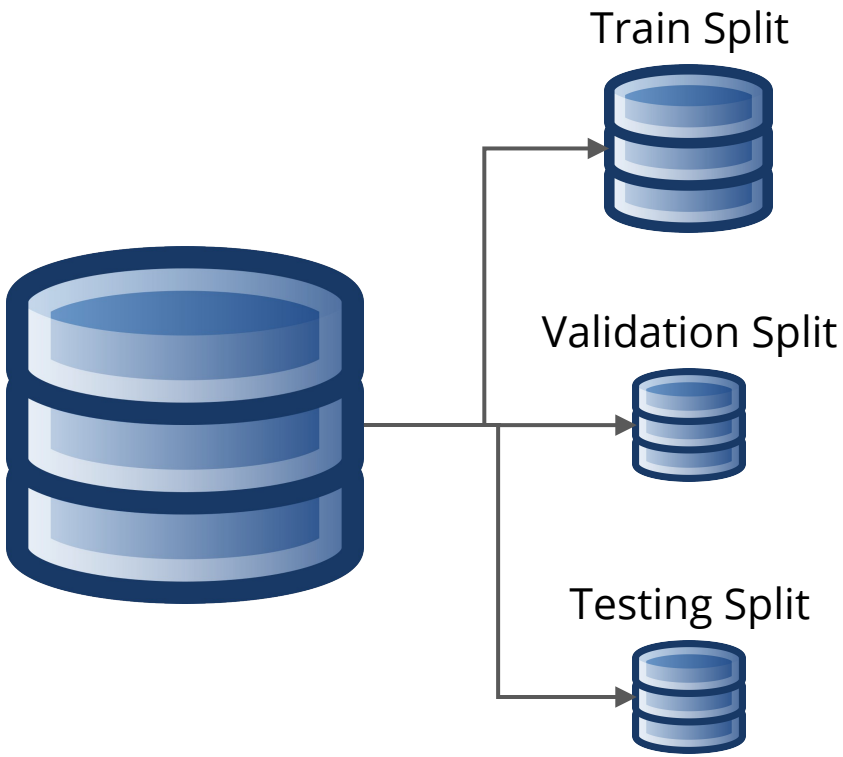
Examples of Univariate Feature Selection:

1. Information Gain
2. Fisher Score
3. Correlations with Target

Examples of Multivariate Feature Selection:

1. Relief
2. Cross-Correlation Feature Selection
3. Branch and Bound
4. Sequential Forward Selection
5. Plus L - Minus R

Proper ML Modeling



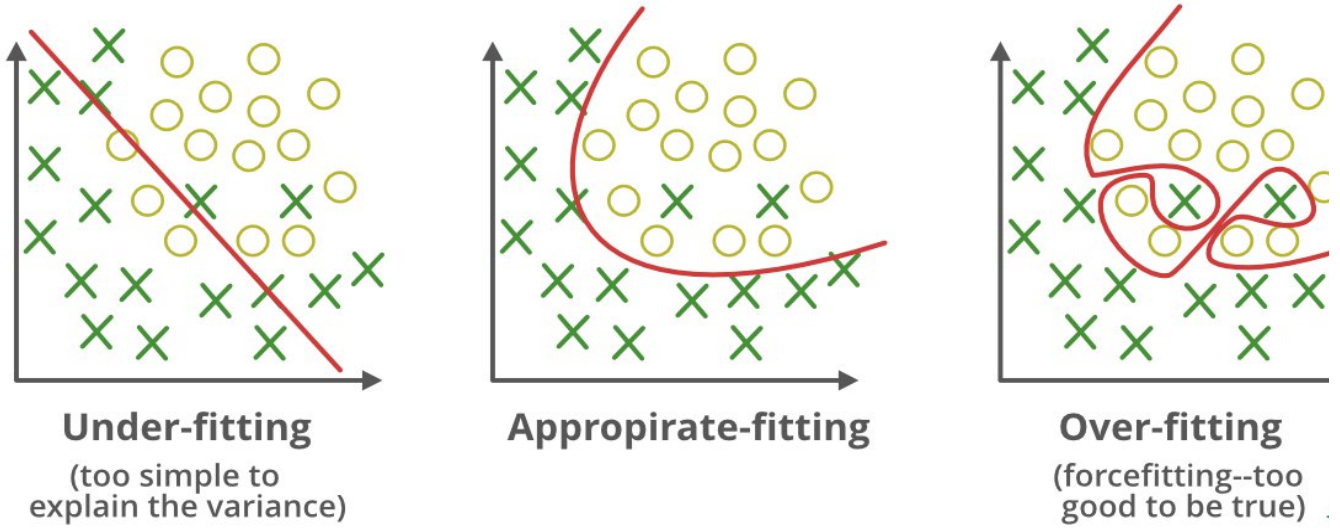
To Train your model weights'

To Tune your model hyper-parameters

To evaluate your model performance on unseen data

Proper ML Modeling

Classification



Bad on Train/Test	Good on Train/Test	Good on Train Bad on Test

Kolluri, J., Kotte, V. K., Phridviraj, M. S. B., & Razia, S. (2020, June). Reducing overfitting problem in machine learning using novel L1/4 regularization method. In *2020 4th international conference on trends in electronics and informatics (ICOEI)(48184)* (pp. 934-938). IEEE.

Recap this lecture

After successfully completing this lecture, you are able to….

- Identify the type of a machine learning problem
- Understand the stages of building a machine learning pipeline
- Apply the proper steps for machine learning modeling

Outlook: What will the tutorial be about?

- In this micro-lecture, we'll demystify classification tasks:
  - the common pitfalls of underfitting and overfitting.
  - You'll see how a properly fitted model finds the right balance, making accurate predictions without memorizing the data.
- By the end, you'll know how to recognize and avoid the traps of poor model fitting to build robust classification models!

The European Commission support for the production of this publication does not constitute an endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.