Feature Selection in Machine Learning for BioMedical Data

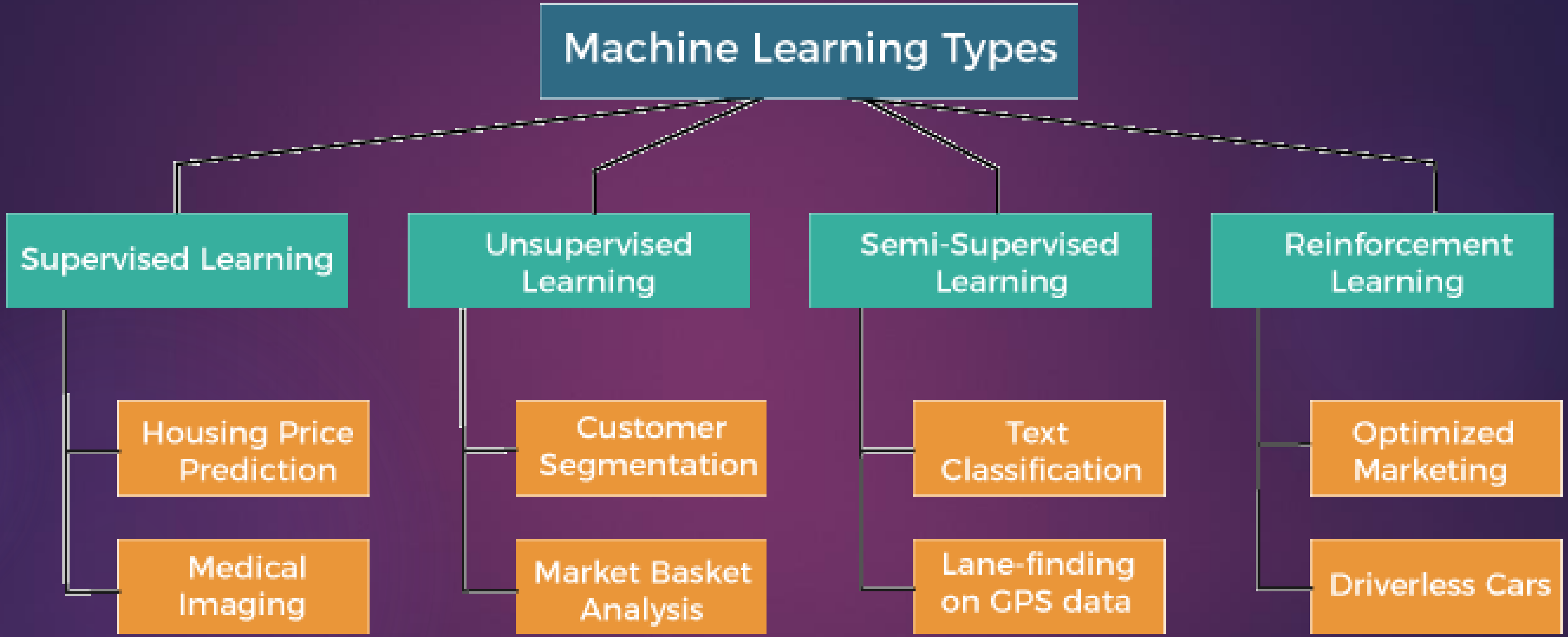
Nov 25 2024

Giảng viên: TS. Lưu Phúc Lợi
Luu.p.loi@googlemail.com

Content of Lecture 1

1. Introduction to Feature Selection
2. Filter method
3. Wrapper methods
4. Embedded method

Introduction to Feature Selection



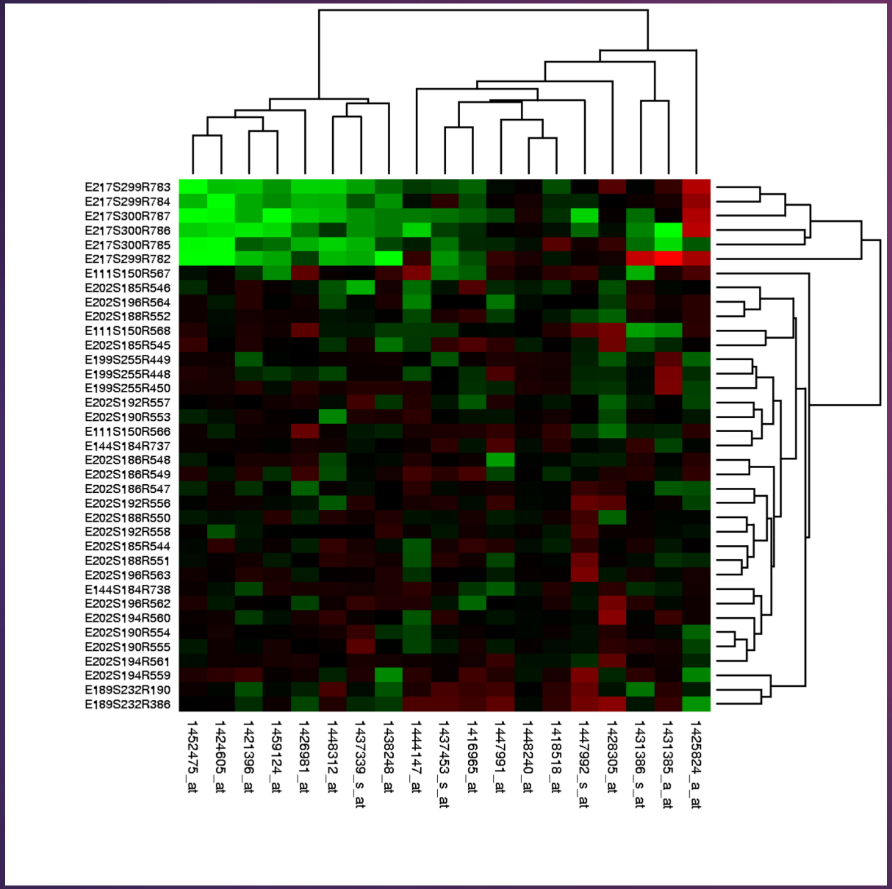
Machine Learning Types

- Supervised Learning
  - Housing Price Prediction
  - Medical Imaging
- Unsupervised Learning
  - Customer Segmentation
  - Market Basket Analysis
- Semi-Supervised Learning
  - Text Classification
  - Lane-finding on GPS data
- Reinforcement Learning
  - Optimized Marketing
  - Driverless Cars

Supervise Learning: regression or classification

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{\mathbb{Y}} = \underbrace{\begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & & \ddots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}}_{\mathbb{X}} \underbrace{\begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_p \end{bmatrix}}_{\theta}$$

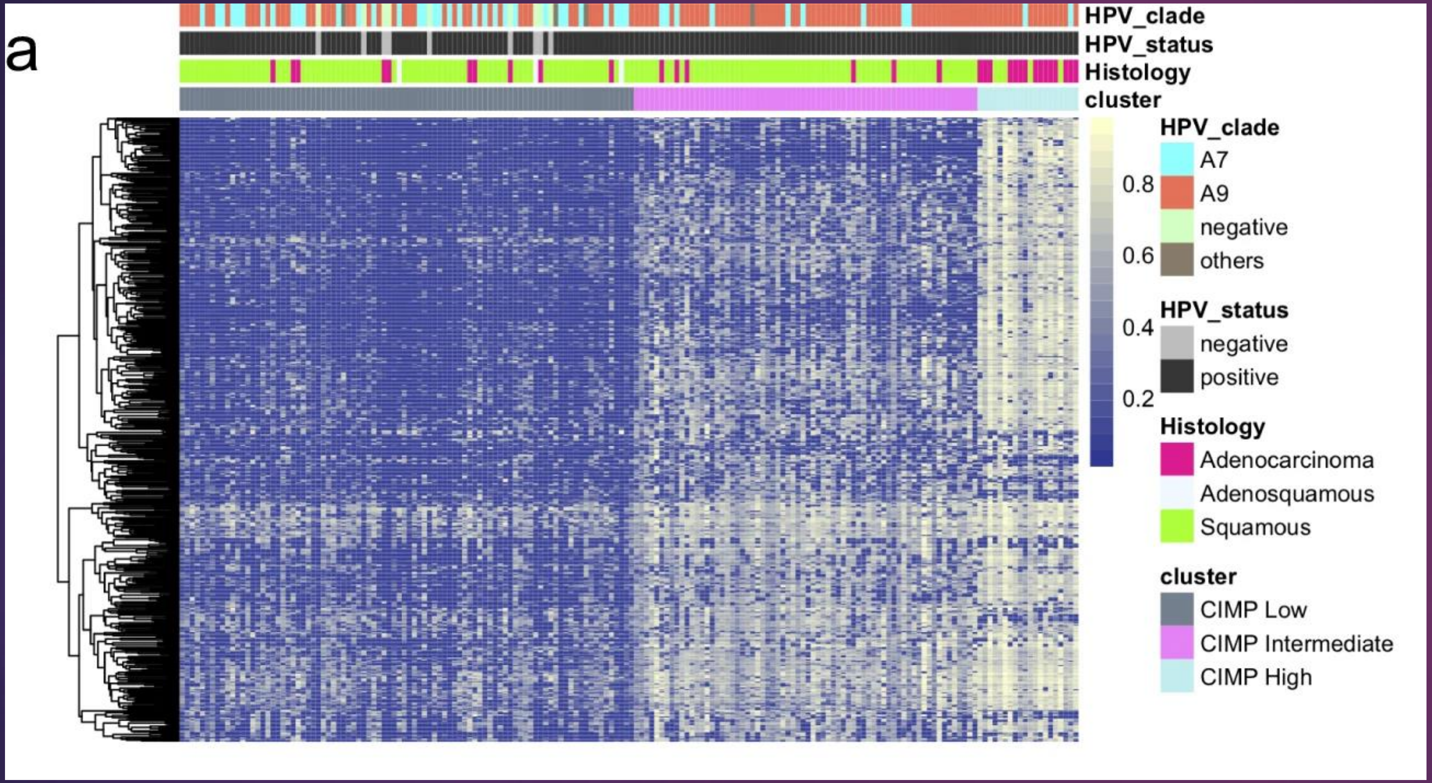
BioMedical data: gene expression with p >> n



n = number of samples 6, 10, 100, 1k
P = number of genes 20k

https://en.wikipedia.org/wiki/Gene_expression_profiling

BioMedical data: DNA methylation (p >> n)

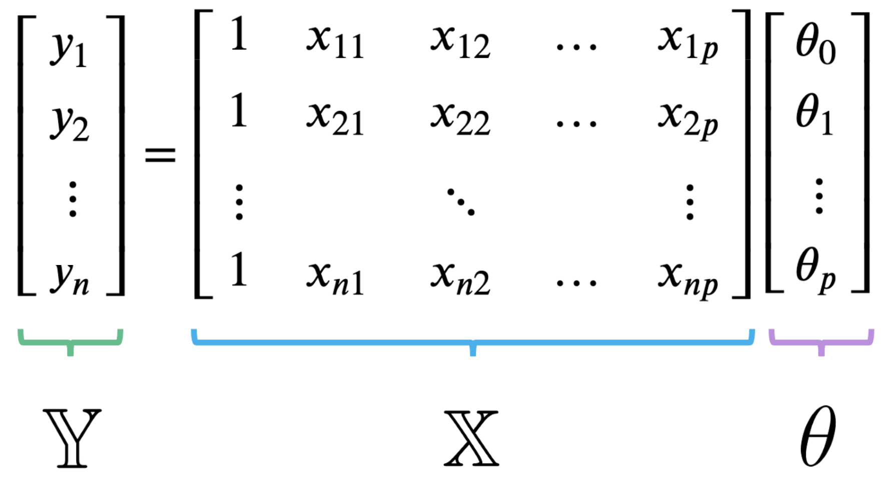


n = number of samples 6, 10, 100, 1k
P = number of CpG 28M

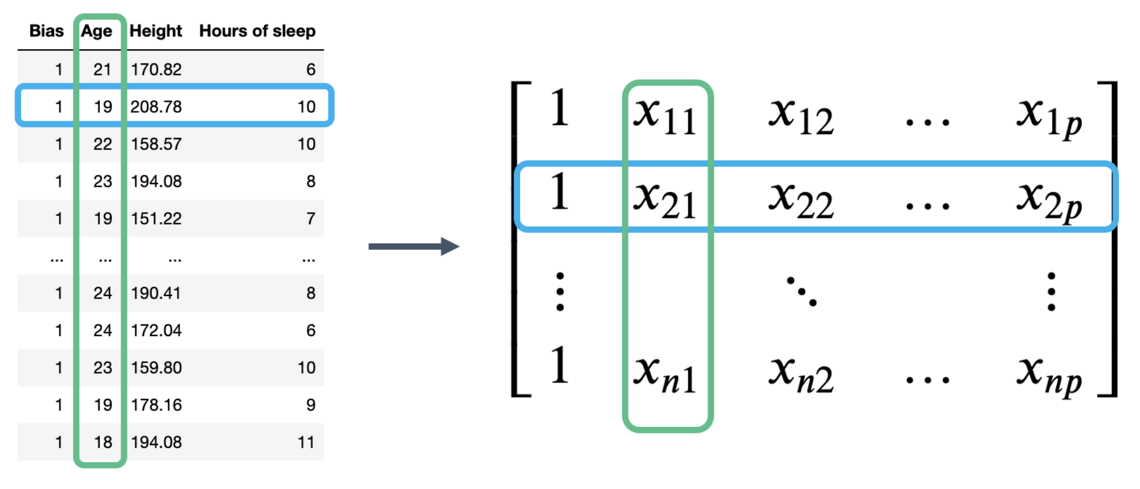
https://www.nature.com/articles/s41392-019-0081-6/figures/1

What is feature/variable selection?

- Find the features (variables/columns) in X which are important for predicting, and remove the features that are not
- Give:



$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{\mathbb{Y}} = \underbrace{\begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & & \ddots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}}_{\mathbb{X}} \underbrace{\begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_p \end{bmatrix}}_{\theta}$$



Bias	Age	Height	Hours of sleep
1	21	170.82	6
1	19	208.78	10
1	22	158.57	10
1	23	194.08	8
1	19	151.22	7
...	...	...	...
1	24	190.41	8
1	24	172.04	6
1	23	159.80	10
1	19	178.16	9
1	18	194.08	11

$$\longrightarrow \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & & \ddots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

- Find the columns in X which are important for predicting y

https://ds100.org/course-notes/ols/ols.html

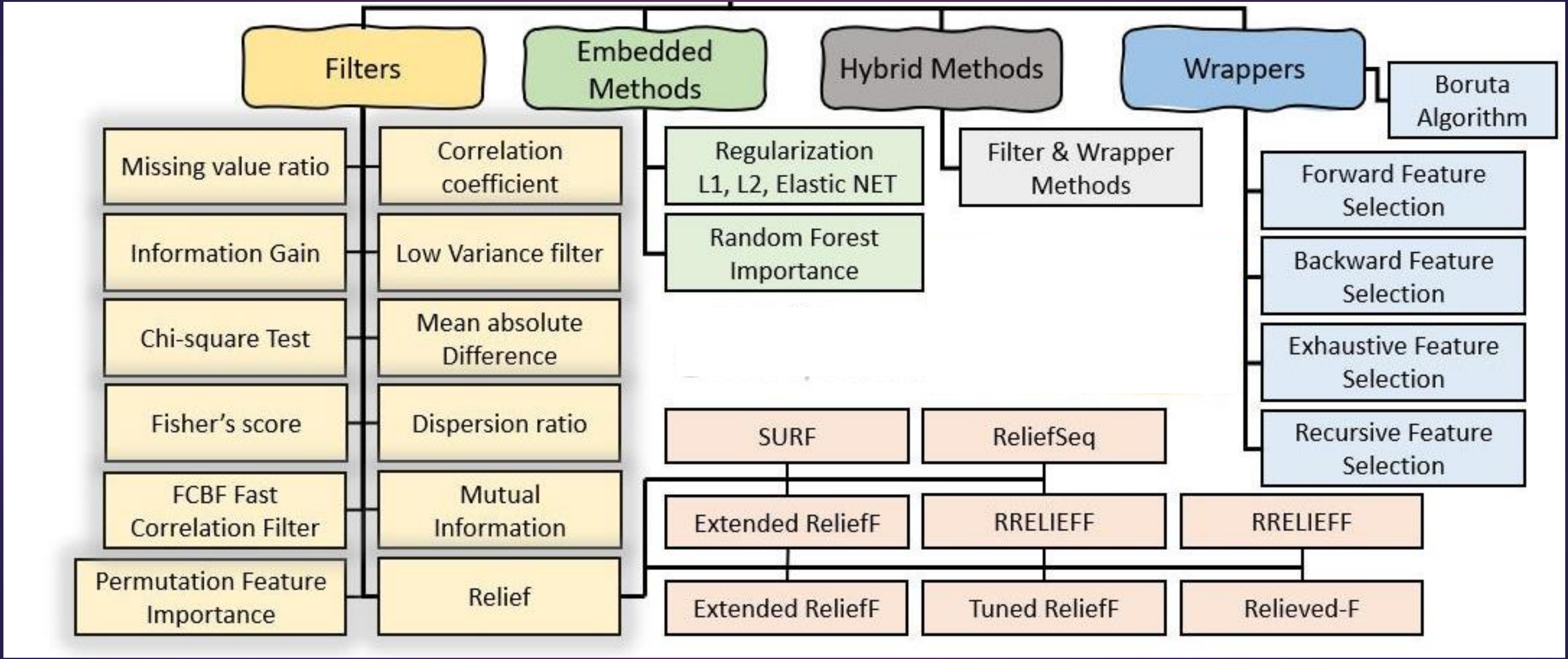
https://ubc-cs.github.io/cpsc330-2023W1/lectures/13_feature-engineering-selection.html

Why feature selection?

- Interpretability: Models are more interpretable with fewer features. If you get the same performance with 10 features instead of 500 features, why not use the model with smaller number of features?
- Computation: Models fit/predict faster with fewer columns.
- Data collection: What type of new data should I collect? It may be cheaper to collect fewer columns.
- Fundamental tradeoff: Can I reduce overfitting by removing useless features?
- Feature selection can often result in better performing (less overfit), easier to understand, and faster model.

https://ubc-cs.github.io/cpsc330-2023W1/lectures/13_feature-engineering-selection.html

How do we carry out feature selection? Supervised Feature Selection

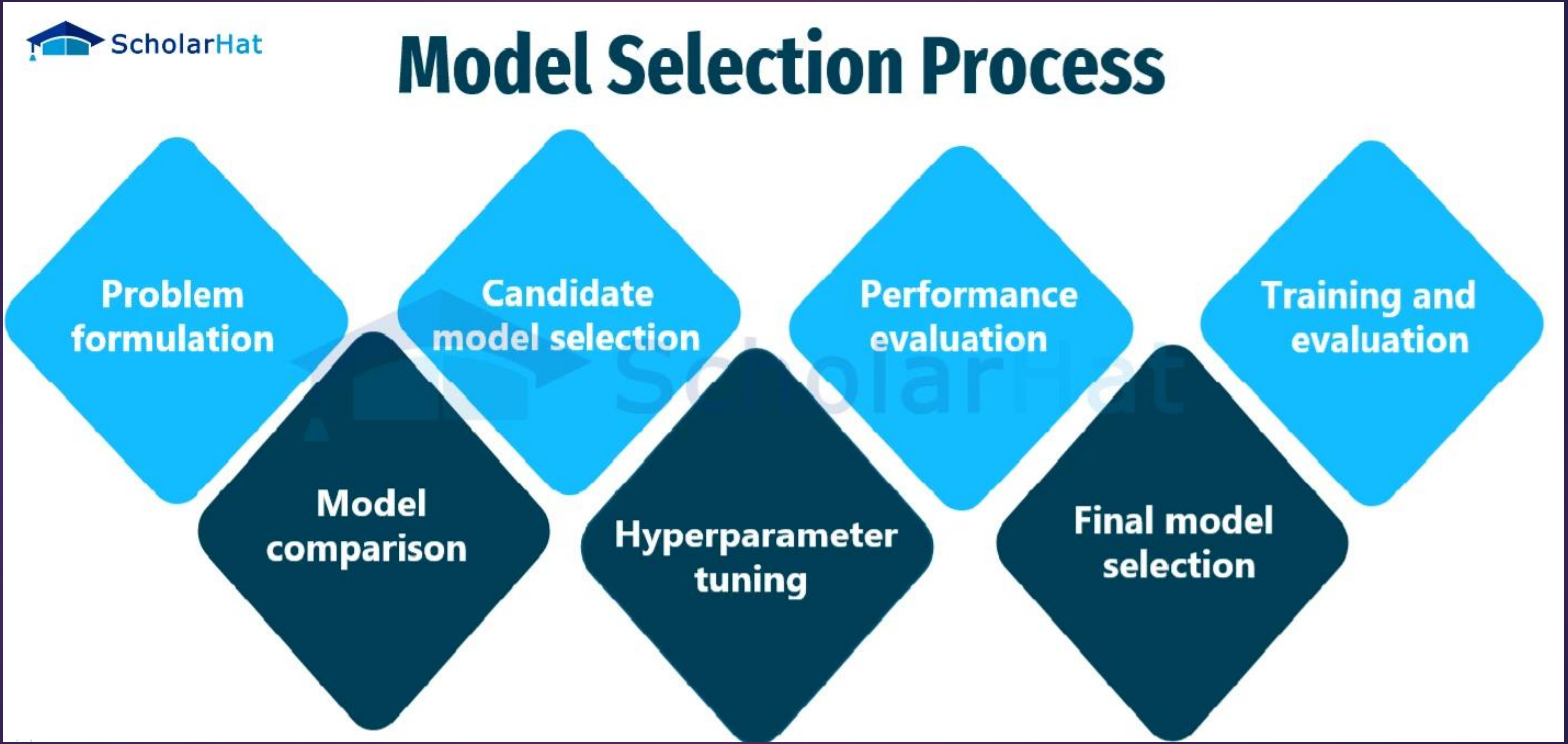


https://medium.com/analytics-vidhya/feature-selection-extended-overview-b58f1d524c1c

Model Selection vs Feature Selection

- Feature Selection is a part of Model Selection

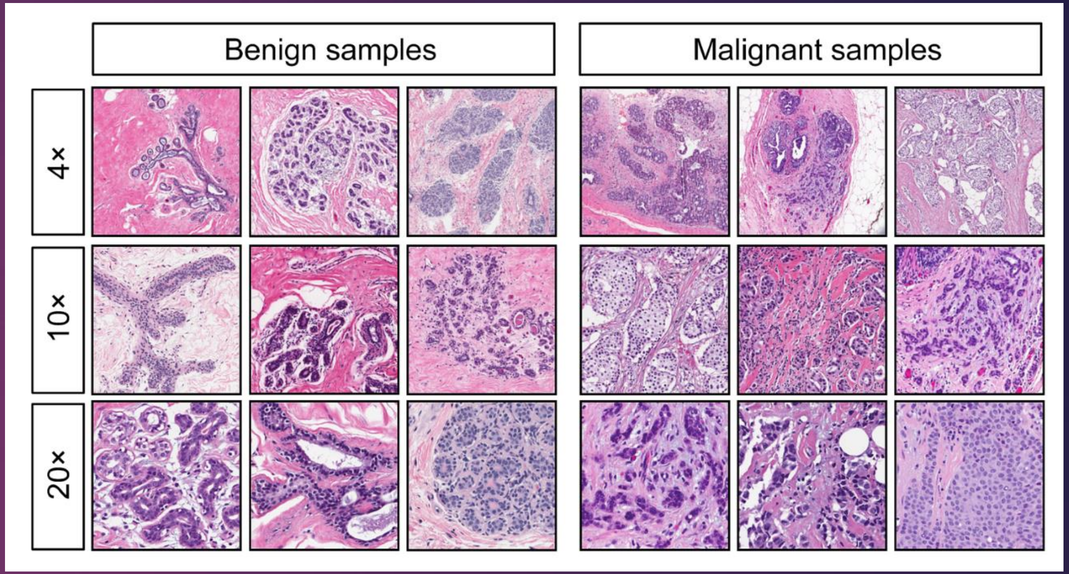
Model selection: steps



https://www.scholarhat.com/tutorial/machinelearning/model-selection-for-machine-learning

Model selection: steps

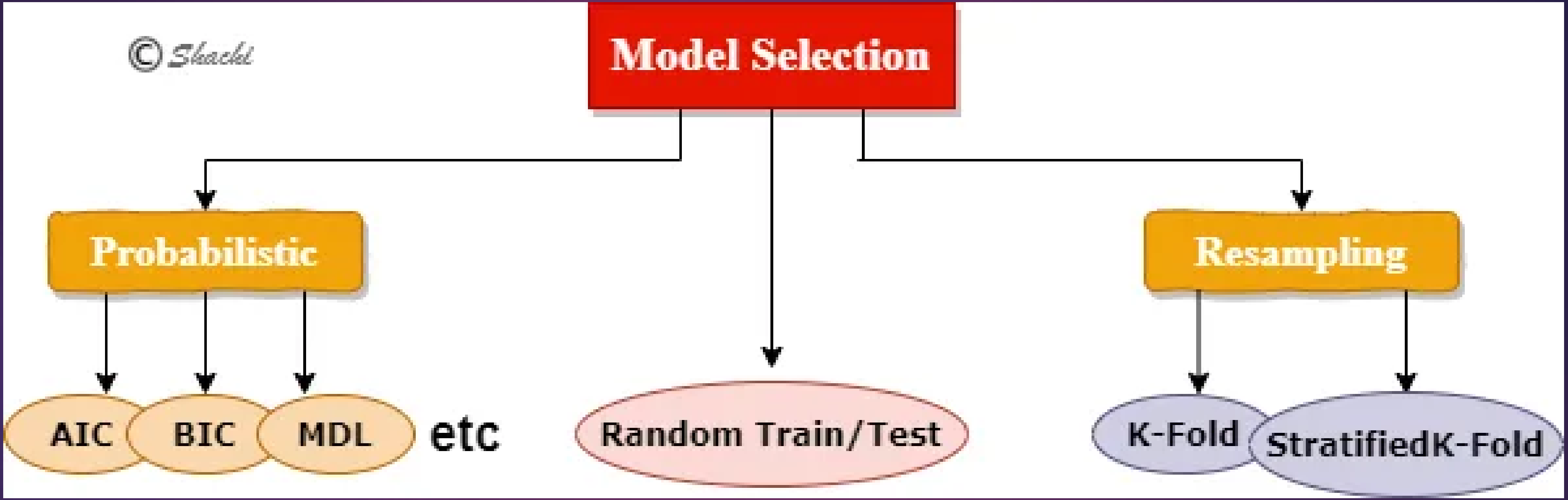
Stage 1: Selecting the regression model forms
Stage 2: Selecting the regression model and the independent variables
Stage 3: Fitting the model
Stage 4: Examining or validation of the applied model



https://www.shiksha.com/online-courses/articles/model-selection-in-machine-learning-regression/

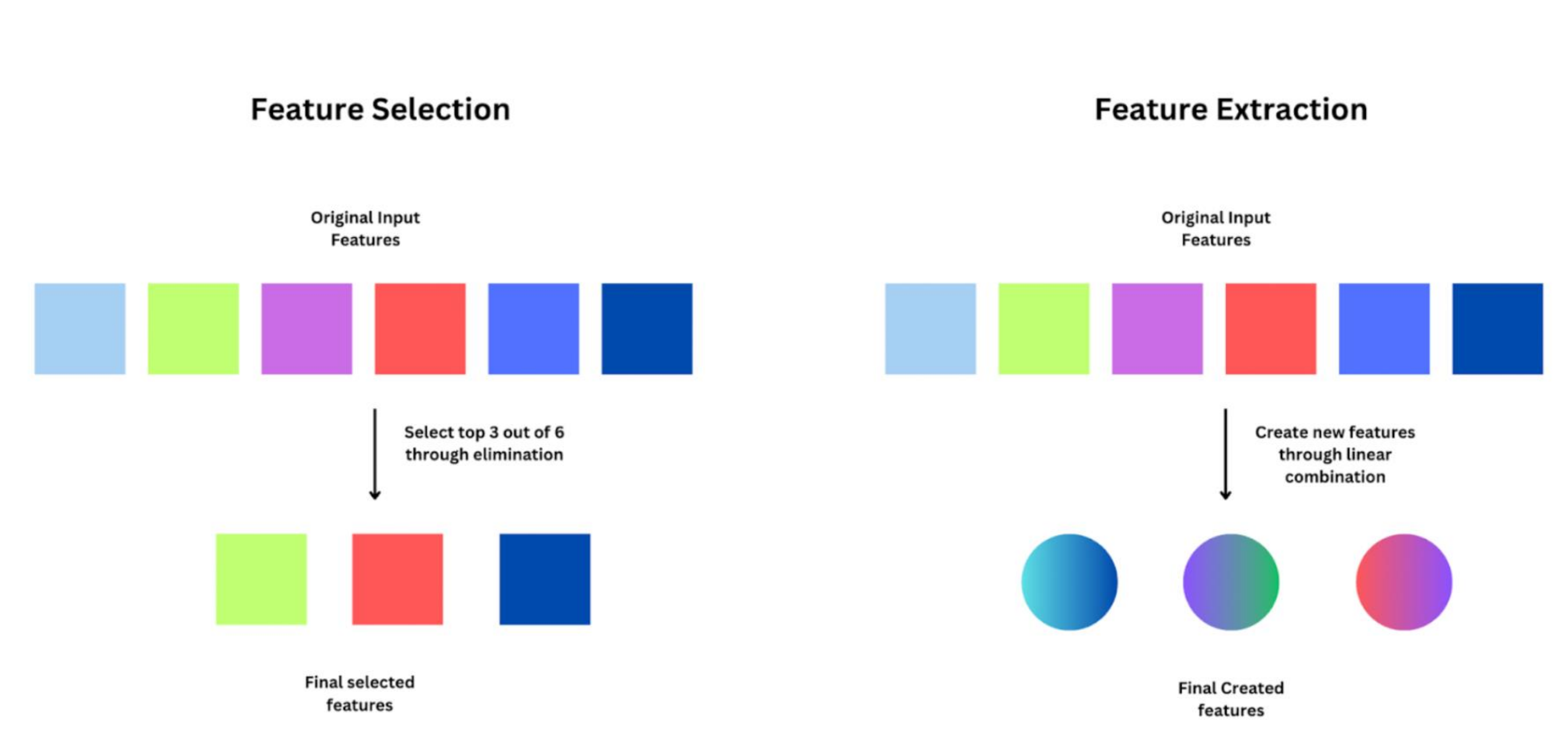
https://medium.com/analytics-vidhya/breast-cancer-diagnostic-dataset-eda-fa0de80f15bd

Model selection: methods



https://medium.com/analytics-vidhya/model-selection-techniques-in-ml-ai-with-python-fdf308d9fa10

Feature Selection vs Feature Extraction/Engineering



https://viso.ai/deep-learning/feature-extraction-in-python/

Xin chân thành cảm ơn!

LUU PHUC LOI, PHD

ZALO: 0901802182

LUU.P.LOI@GOOGLEMAIL.COM