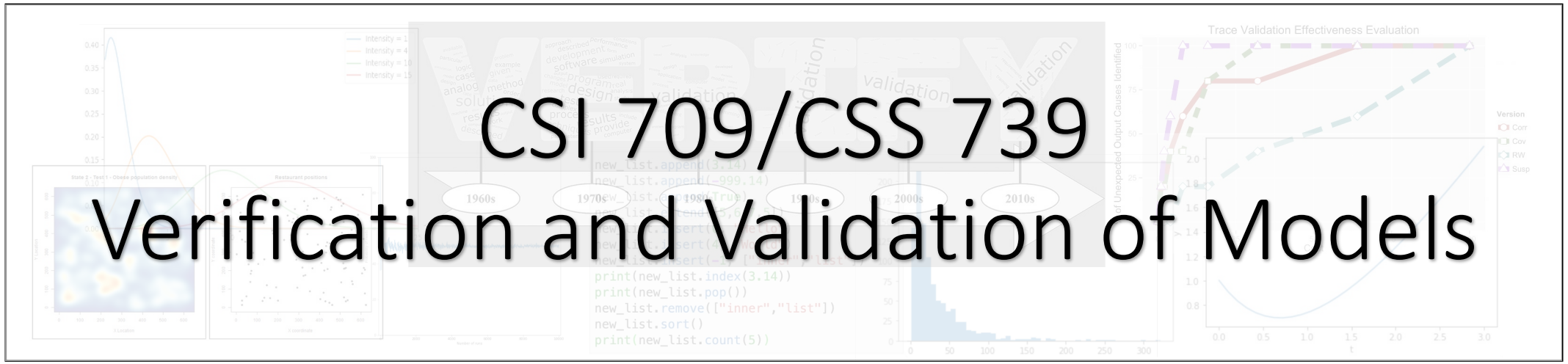


CSI 709/CSS 739
Verification and Validation of Models

Validation of Machine Learning Models (Basics)

Dr. Hamdi Kavak
Computational and Data Sciences Department

http://www.hamdikavak.com
hkavak@gmu.edu

GEORGE MASON UNIVERSITY

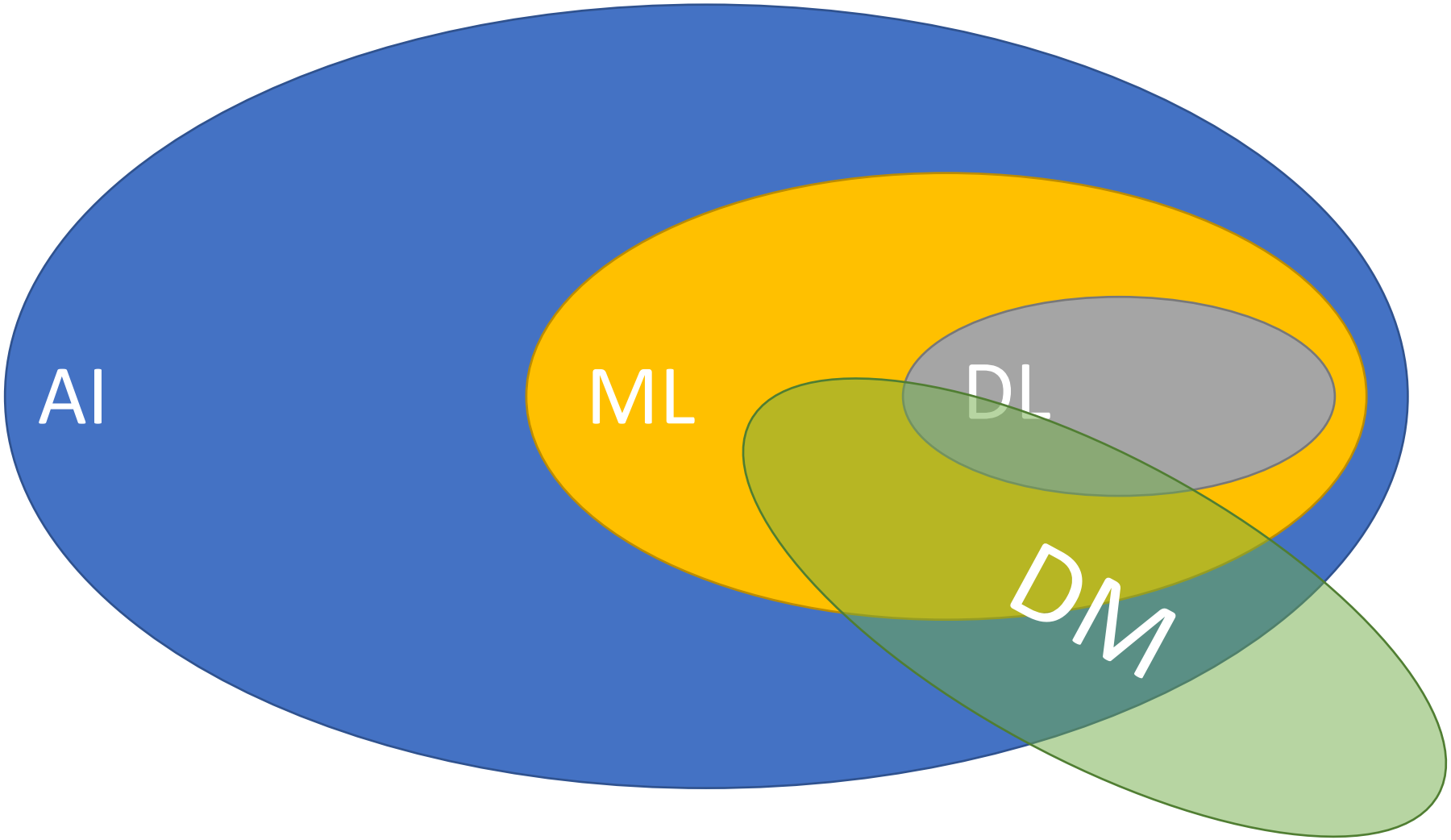
Center for Social Complexity

Machine learning basics

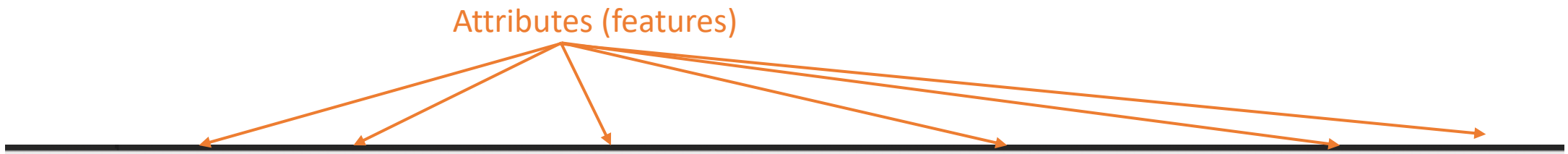
Terminology

- **Artificial intelligence (AI):** “the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings” (Copeland, 2020).
- **Machine learning (ML):** “… discipline concerned with the implementation of computer software that can learn autonomously” (Hosch, 2021).
- **Data mining (DM):** “the process of discovering interesting and useful patterns and relationships in large volumes of data” (Clifton, 2019).

AI vs. ML vs. DM vs. Deep Learning (DL)



Data sets

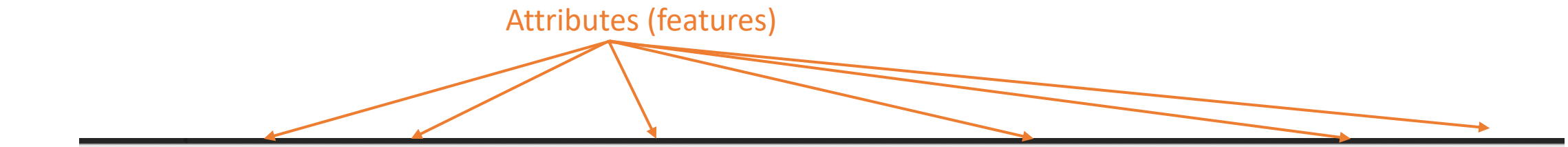


Instances

A	B	C	D	E	F	G
	checkin_ratio	end_of_day_ratio	end_of_inactive_day_ratio	kilometer_distance_to_most_checked_in	midnight_ratio	page_rank
0	0.087	0.000	0.000	6.834	0.000	0.023
1	0.043	0.000	0.000	8.457	0.000	0.033
2	0.435	0.667	0.667	0.000	1.000	0.294
3	0.087	0.000	0.000	0.326	0.000	0.123
4	0.130	0.000	0.000	0.707	0.000	0.176
5	0.043	0.000	0.000	1.620	0.000	0.073
6	0.043	0.000	0.000	0.380	0.000	0.085
7	0.043	0.000	0.000	0.806	0.000	0.096
8	0.043	0.000	0.000	1.506	0.000	0.073
9	0.043	0.333	0.333	1.343	0.000	0.023

Places a person visits: https://github.com/hamdikavak/home-location-prediction/blob/master/data/training_test_set_anonymized.csv

Data sets



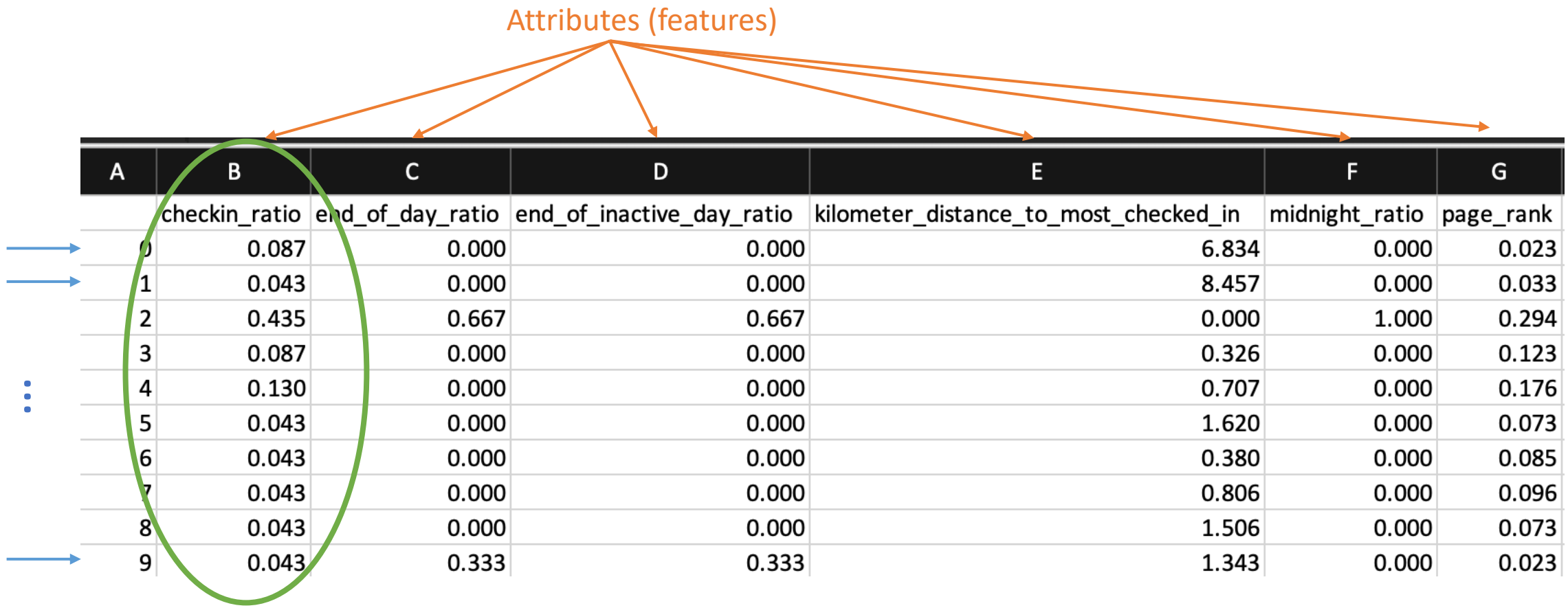
Instances

A	B	C	D	E	F	G
	checkin_ratio	end_of_day_ratio	end_of_inactive_day_ratio	kilometer_distance_to_most_checked_in	midnight_ratio	page_rank
0	0.087	0.000	0.000	6.834	0.000	0.023
1	0.043	0.000	0.000	8.457	0.000	0.033
2	0.435	0.667	0.667	0.000	1.000	0.294
3	0.087	0.000	0.000	0.326	0.000	0.123
4	0.130	0.000	0.000	0.707	0.000	0.176
5	0.043	0.000	0.000	1.620	0.000	0.073
6	0.043	0.000	0.000	0.380	0.000	0.085
7	0.043	0.000	0.000	0.806	0.000	0.096
8	0.043	0.000	0.000	1.506	0.000	0.073
9	0.043	0.333	0.333	1.343	0.000	0.023

Each **instance** here is a visited place from an individual

Places a person visits: https://github.com/hamdikavak/home-location-prediction/blob/master/data/training_test_set_anonymized.csv

Data sets



Places a person visits: https://github.com/hamdikavak/home-location-prediction/blob/master/data/training_test_set_anonymized.csv

Data sets w/ classes

Is this place home?

A	B	C	D	E	F	G	H
	checkin_ratio	end_of_day_ratio	end_of_inactive_day_ratio	kilometer_distance_to_most_checked_in	midnight_ratio	page_rank	is_home
0	0.087	0.000	0.000	6.834	0.000	0.023	FALSE
1	0.043	0.000	0.000	8.457	0.000	0.033	FALSE
2	0.435	0.667	0.667	0.000	1.000	0.294	TRUE
3	0.087	0.000	0.000	0.326	0.000	0.123	FALSE
4	0.130	0.000	0.000	0.707	0.000	0.176	FALSE
5	0.043	0.000	0.000	1.620	0.000	0.073	FALSE
6	0.043	0.000	0.000	0.380	0.000	0.085	FALSE
7	0.043	0.000	0.000	0.806	0.000	0.096	FALSE
8	0.043	0.000	0.000	1.506	0.000	0.073	FALSE
9	0.043	0.333	0.333	1.343	0.000	0.023	FALSE

Places a person visits: https://github.com/hamdikavak/home-location-prediction/blob/master/data/training_test_set_anonymized.csv

CLASSICAL MACHINE LEARNING

Data is pre-categorized or numerical

Data is not labeled in any way

SUPERVISED

Predict a category

Predict a number

CLASSIFICATION

«Divide the socks by color»

REGRESSION

«Divide the ties by length»

UNSUPERVISED

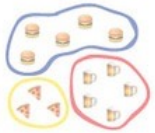
Divide by similarity

Find hidden dependencies

Identify sequences

CLUSTERING

«Split up similar clothing into stacks»



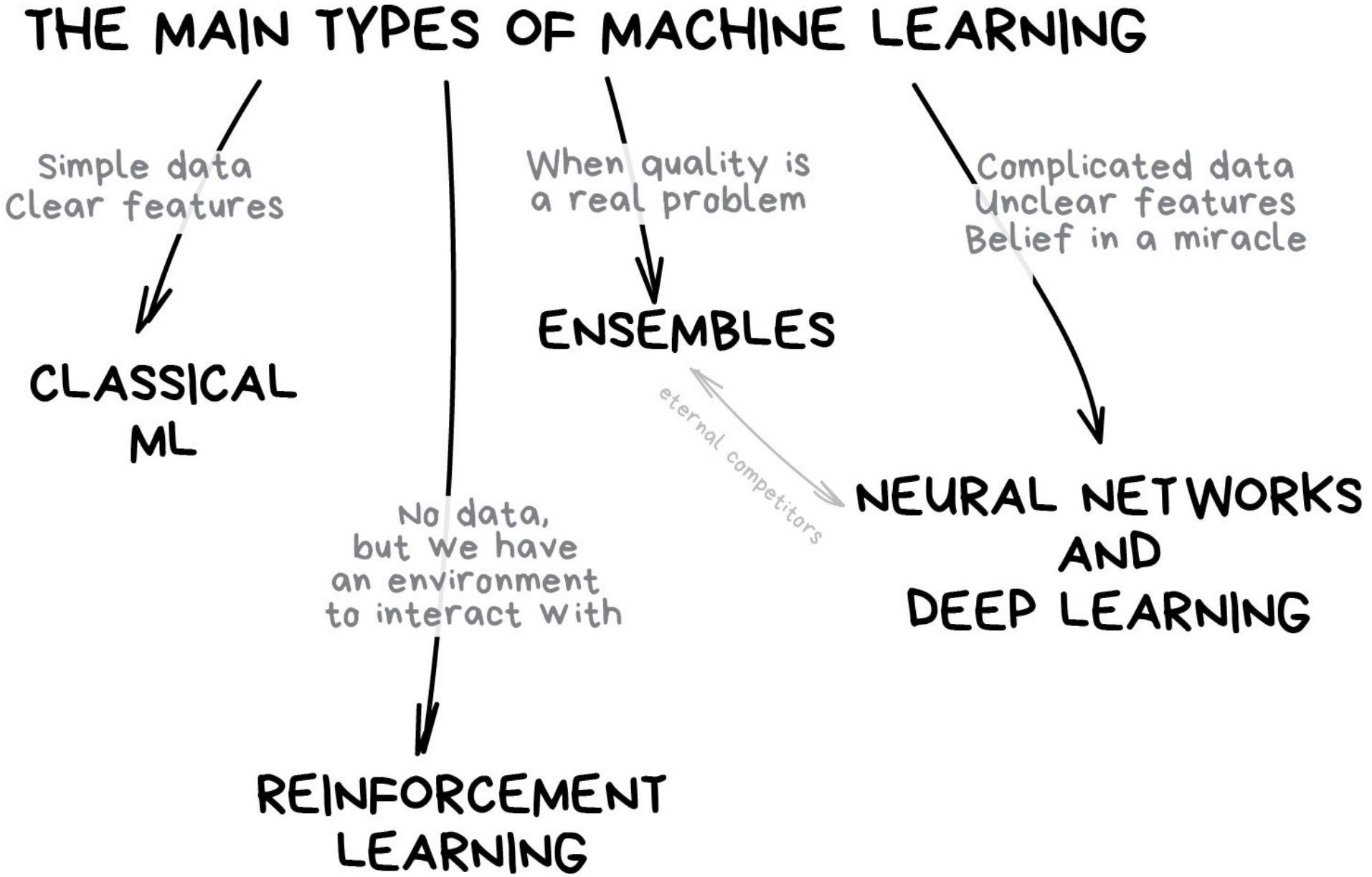
ASSOCIATION

«Find what clothes I often wear together»

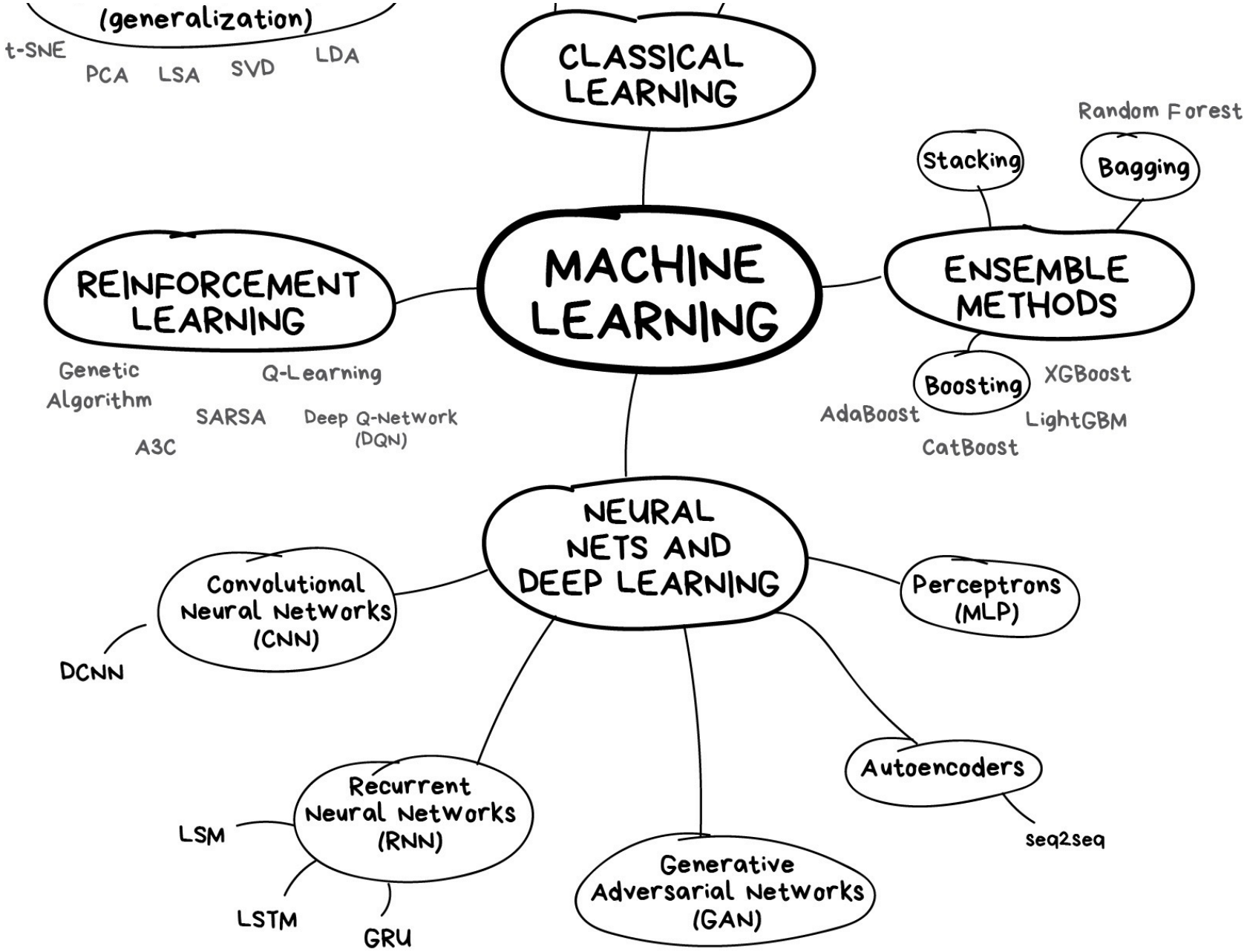
DIMENSION REDUCTION (generalization)

«Make the best outfits from the given clothes»

Source: https://vas3k.com/blog/machine_learning/index.html



Source: https://vas3k.com/blog/machine_learning/index.html



Source: https://vas3k.com/blog/machine_learning/index.html

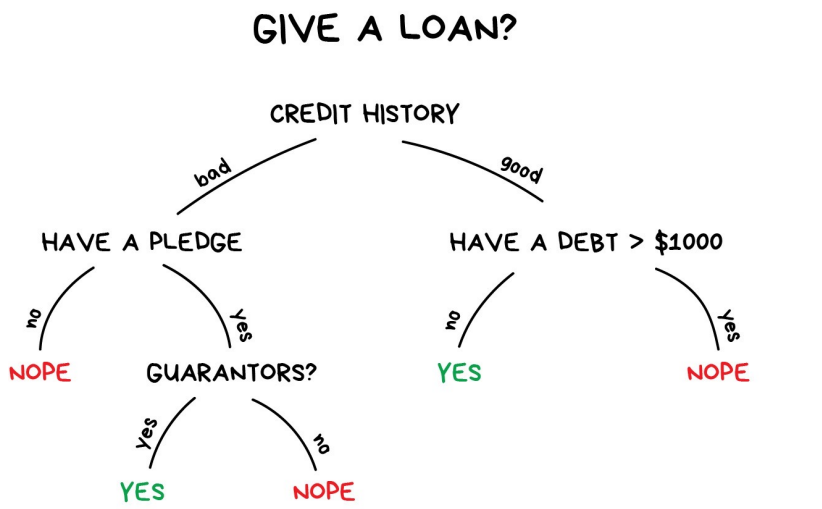
Supervised learning

Supervised learning

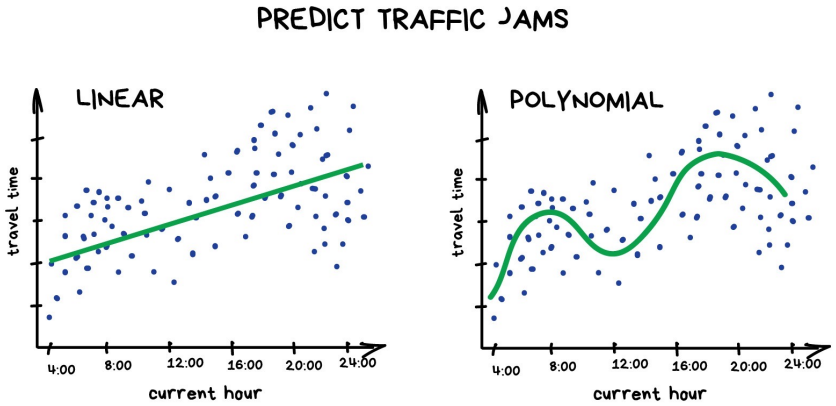
- Category prediction (i.e., classification)
  - Task is to assign instances to a discrete class
  - Two classes: binary classification
  - Three or more classes: multiclass classification
  - E.g.:
    - Fraud detection, spam detection, document classification, sentiment prediction
- Numerical prediction (i.e., regression)
  - Task is to assign instances to a numerical value
  - E.g.:
    - Population, stock price, house price, vaccine acceptance

Supervised learning

- Category prediction (i.e., classification)
  - Task is to assign instances to a discrete class
  - Two classes: binary classification
  - Three or more classes: multiclass classification
  - E.g.:
    - Fraud detection, spam detection, document classification, sentiment prediction, …
- Numerical prediction (i.e., regression)
  - Task is to assign instances to a numerical value
  - E.g.:
    - Population, stock price, house price, vaccine acceptance, …



DECISION TREE



REGRESSION

Source: https://vas3k.com/blog/machine_learning/index.html

Some supervised learning techniques

- Support vector machines
- Nearest neighbor classification
- Decision tree
- Random forest
- XGBoost
- Neural networks

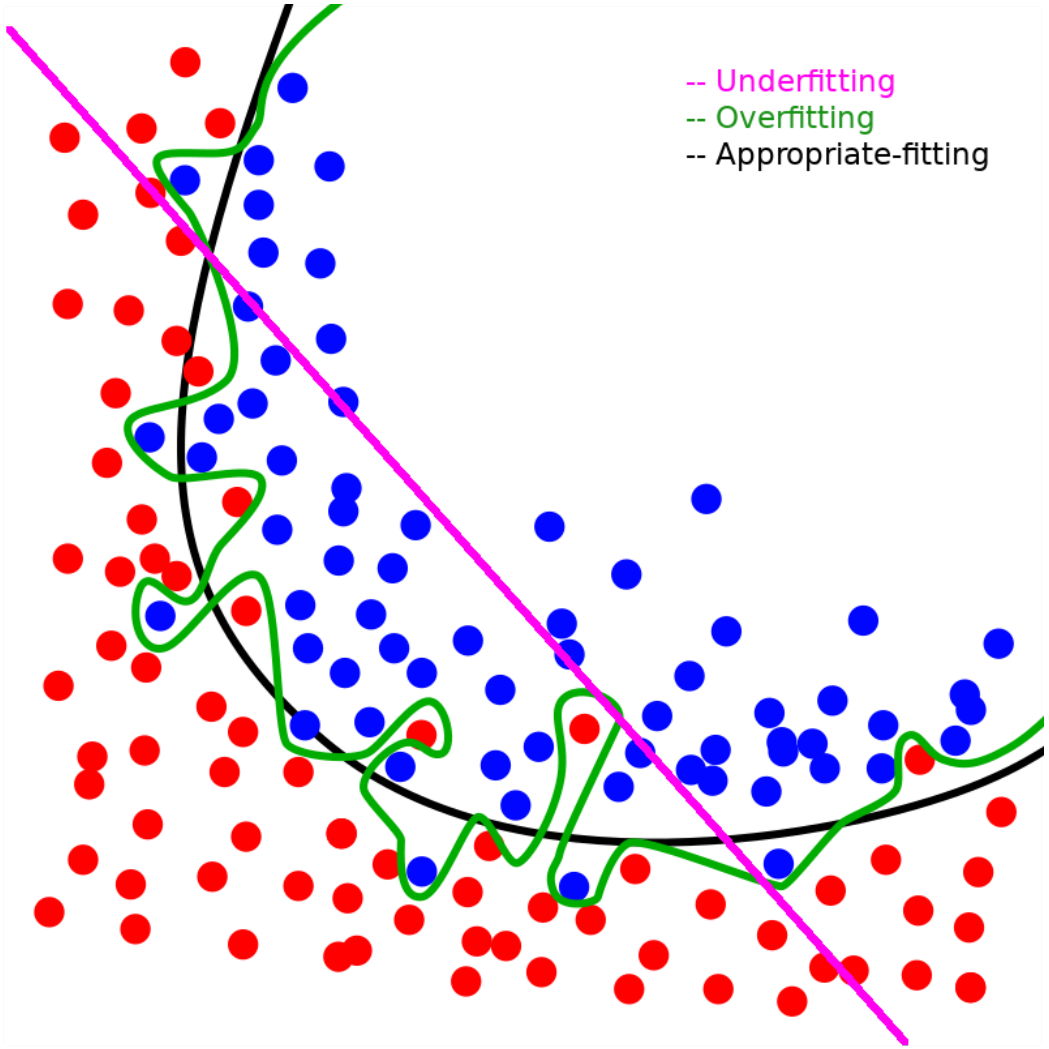
Model training parameters

- Input data (instances)
  - E.g.: training data
- Model parameters
  - E.g.,: weights in a neural network, coefficients in a regression model
- Hyperparameters
  - Configuration parameters of the ML algorithm
  - E.g.: how many hidden layers, learning rate, regularization parameter

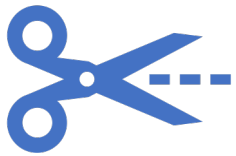
Evaluating supervised models

Underfit vs. overfit

- Carefully analyze the model's outputs to evaluate whether they are meeting the goals that we set up for it.



Preventing overfit



Data splits

Train/test

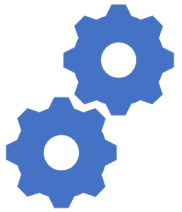
Train/validation/test

Cross validation

Improve data

Collect more

Remove noise



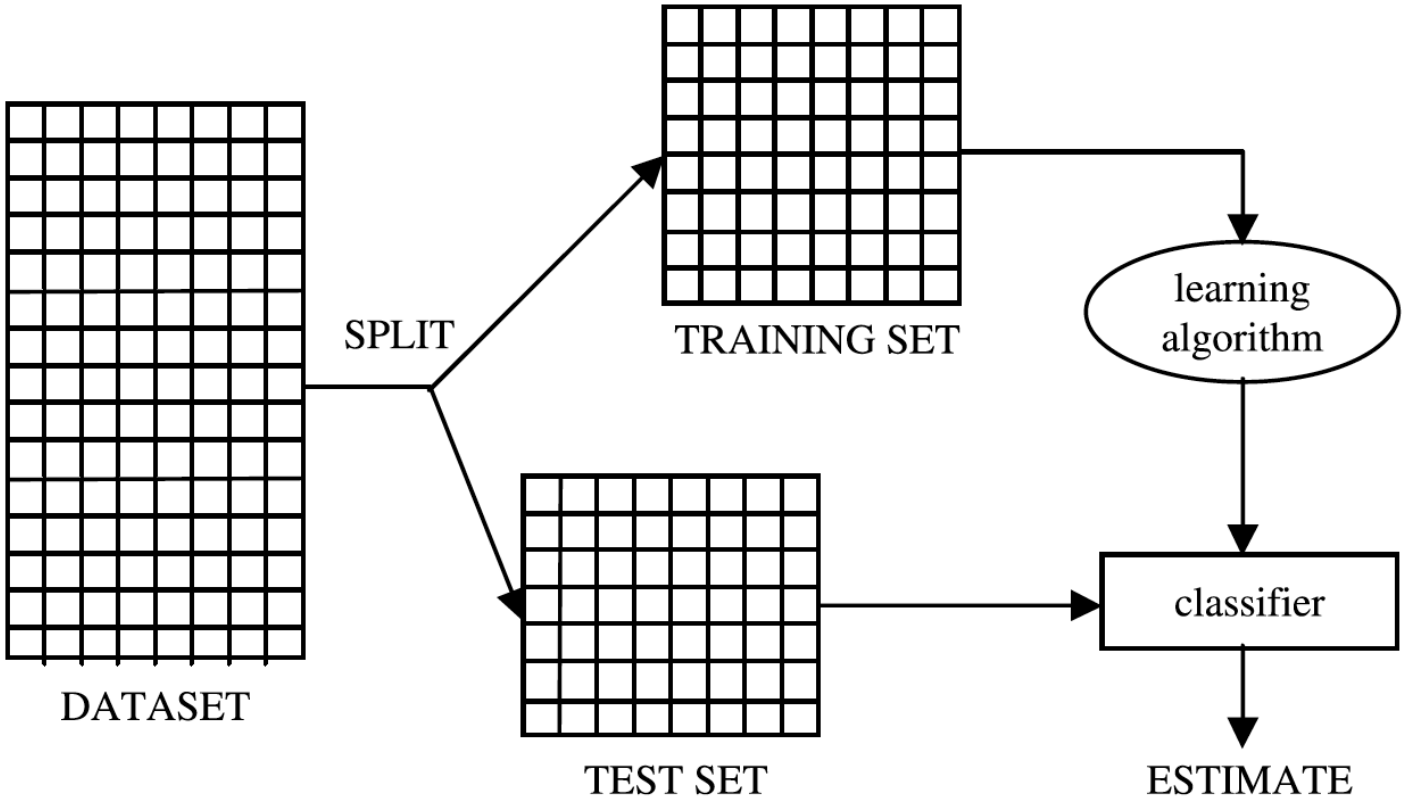
Model update

Use a simpler model

Regularization

Holdout testing

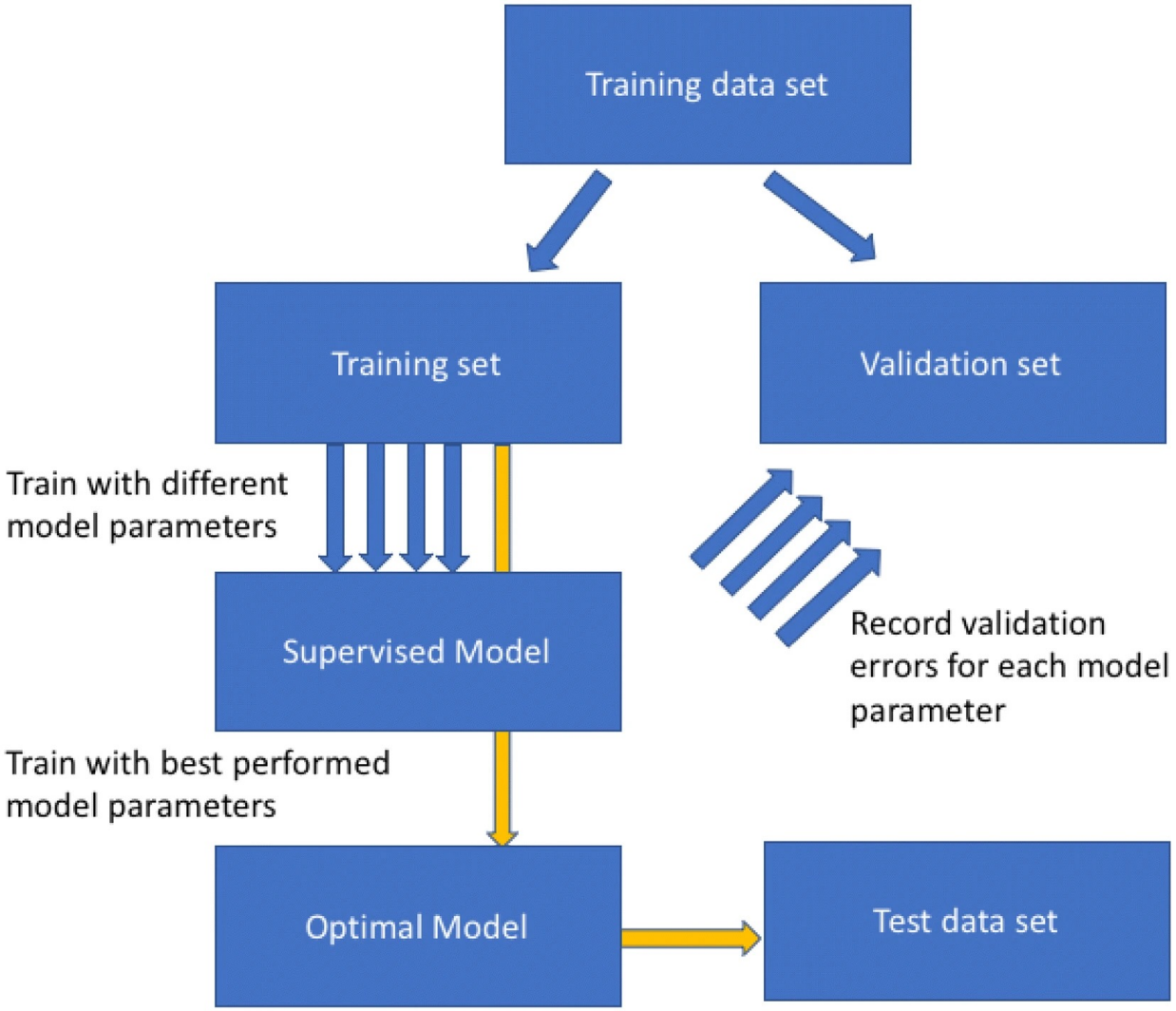
- Train/test split



Source: Bramer, M. (2016). *Principles of data mining* (3rd edition). London: Springer.

Holdout testing

- Train/validation/test split

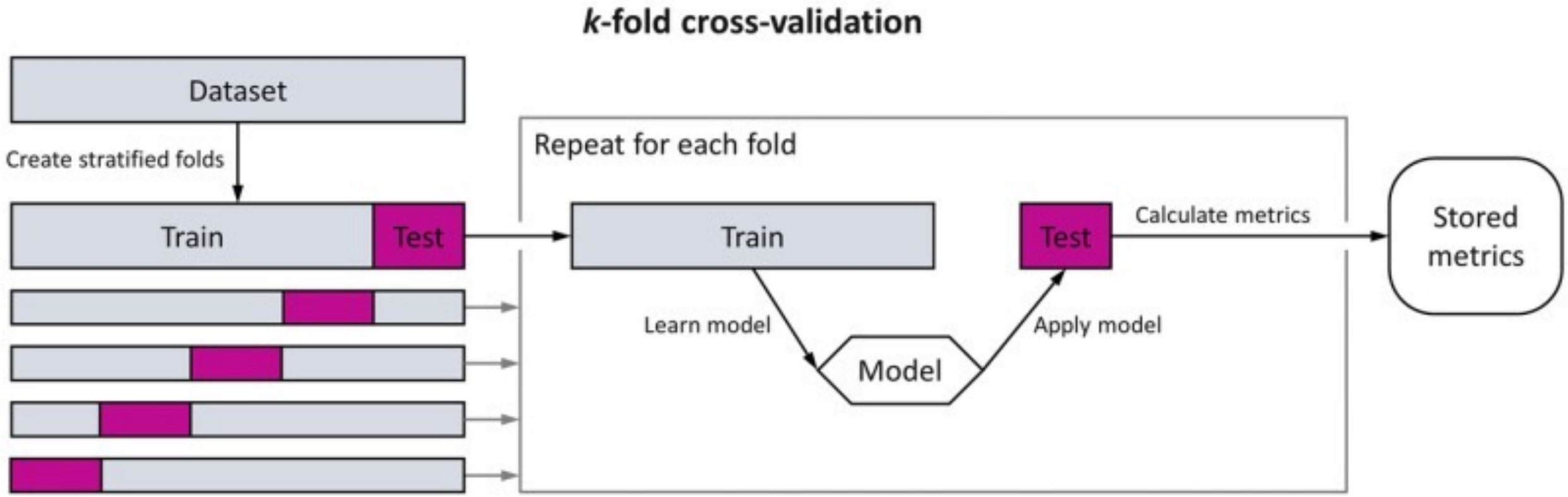


Source: Xu, Y., & Goodacre, R. (2018). On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of Analysis and Testing*, *2*(3), 249-262.

Cross validation

- When number of instances is small, you want to have less variance in model predictions.
- Often, we use *k-fold cross-validation*
  - Divide *N* instances into *k* equal folds
  - Hold each fold as a testing data and train the model using the remaining *k-1* folds
  - Measure the performance across folds

Cross validation



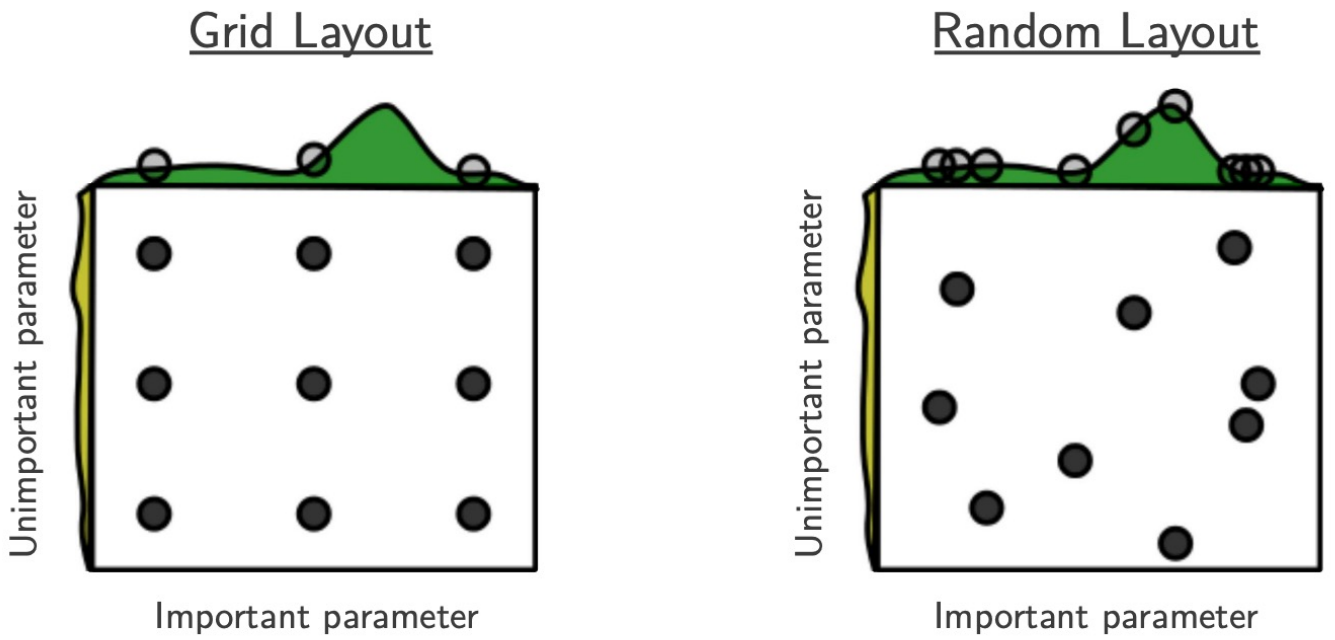
Source: Dankers FJWM, Traverso A, Wee L, et al. Prediction Modeling Methodology. 2018 Dec 22. In: Kubben P, Dumontier M, Dekker A, editors. Fundamentals of Clinical Data Science [Internet]. Cham (CH): Springer; 2019. Chapter 8. Available from: https://www.ncbi.nlm.nih.gov/books/NBK543534/ doi: 10.1007/978-3-319-99713-1_8

How about “Nested Cross Validation”?

You will see it in the second presentation today.

Hyperparameter optimization

- The process of finding hyperparameters that improves model fit.

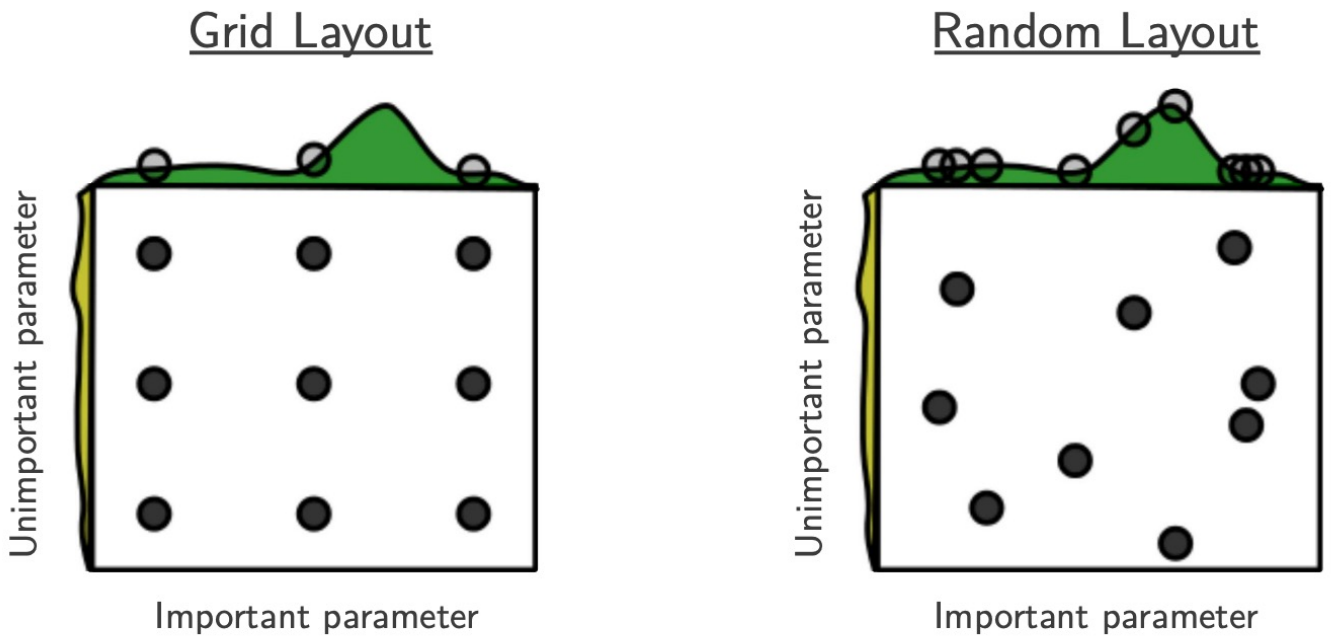


Source: Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, *13*(2).

Hyperparameter optimization

- The process of finding hyperparameters that improves model fit.

Does this process remind you of anything from our previous classes?

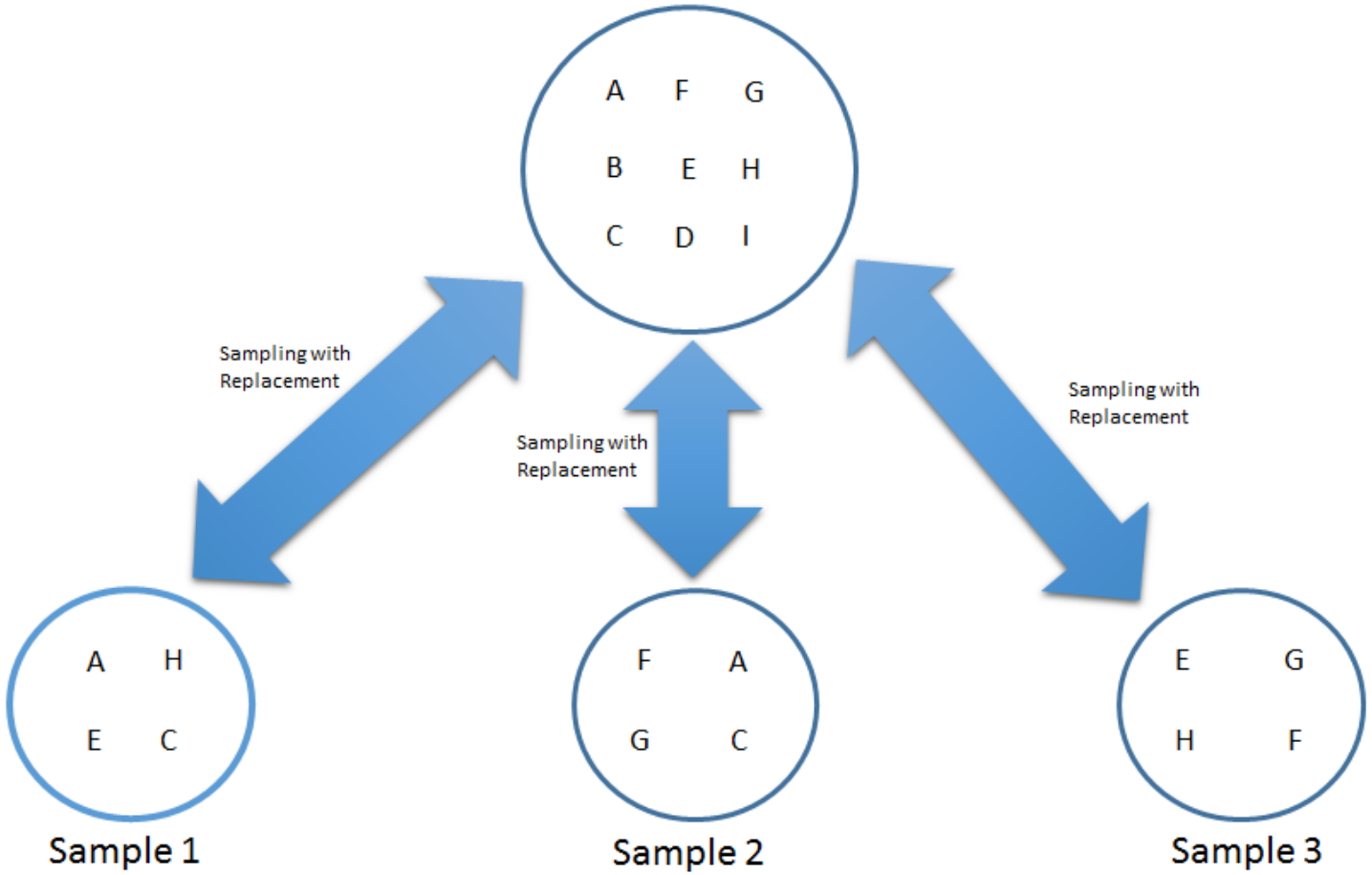


Source: Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, *13*(2).

Bootstrapping

- In cross validation, each instance will be used once.
- Bootstrapping allows sampling the dataset with replacement.
- In general, it's not more robust than cross validation.
- Often used in training/testing ensemble ML models.

Bootstrapping

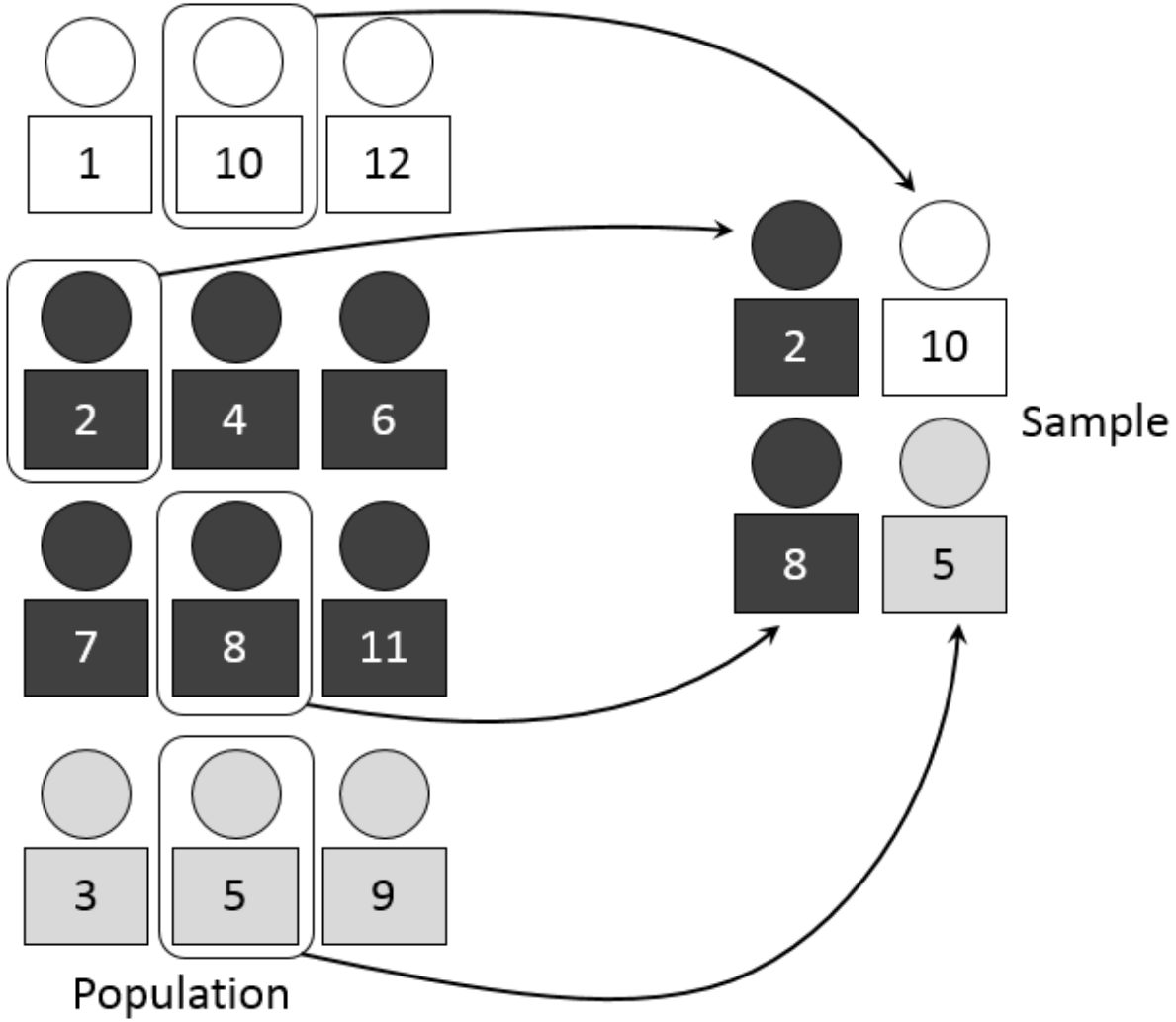


Source: Kumar, R. (2019). *Machine Learning Quick Reference: Quick and essential machine learning hacks for training smart data models*. Packt Publishing Ltd.

Stratified sample

- It is used to eliminate bias in the dataset.
- Assumes that **you know** the true underlying population distributions and your test does not follow that distribution (hence biased).
- Not specific to ML tasks but it's useful

How to create a stratified sample



Source: https://www.netquest.com/blog/en/random-sampling-stratified-sampling

Imbalanced data cases

- For instance, you have a fraud detection dataset with 3% fraud & 97% normal transactions.

Imbalanced data cases

- For instance, you have a fraud detection dataset with 3% fraud & 97% normal transactions.

- What will your ML model do?

Imbalanced data cases

- For instance, you have a fraud detection dataset with 3% fraud & 97% normal transactions.

- What will your ML model do?
- How do you handle challenges?

Imbalanced data cases

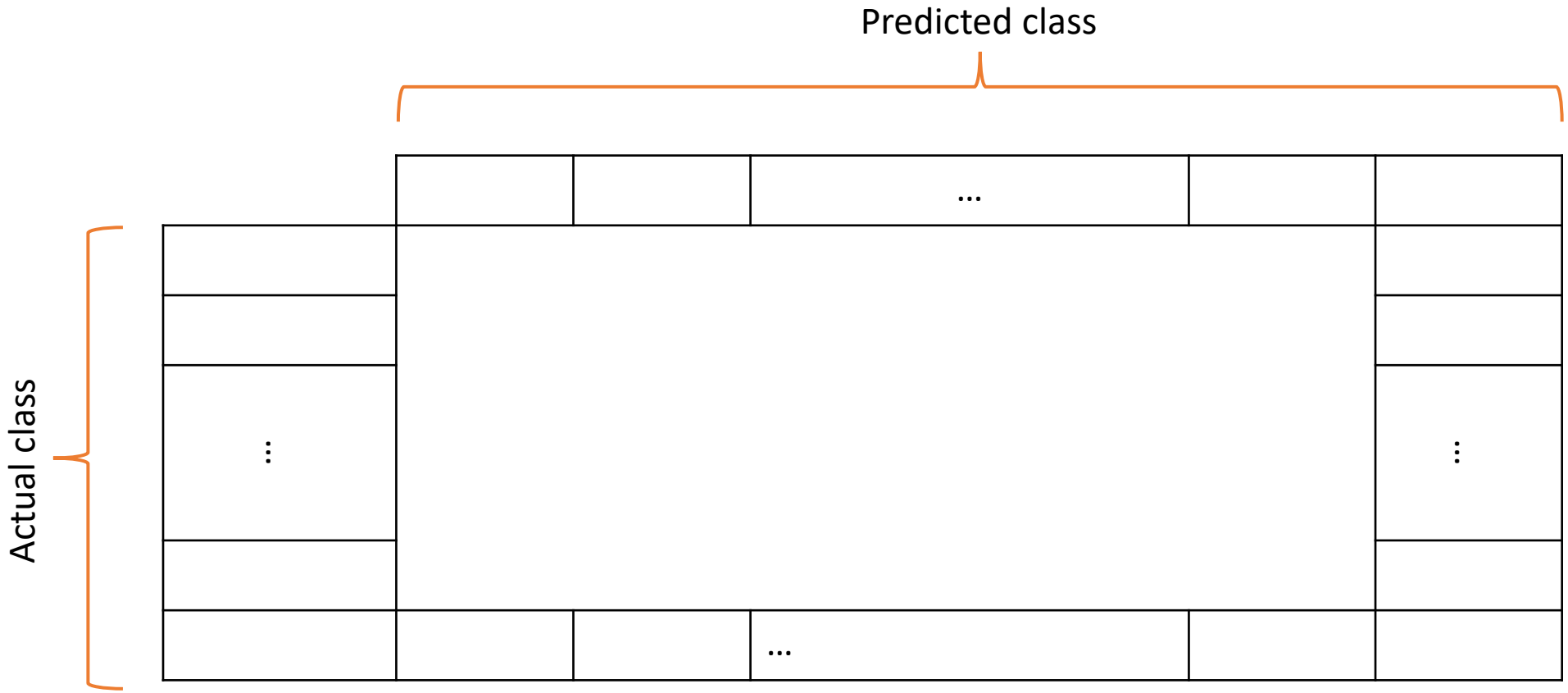
- For instance, you have a fraud detection dataset with 3% fraud & 97% normal transactions.

- What will your ML model do?
- How do you handle challenges?

1. Resample
2. Use model-specific handling of imbalance

Confusion matrix

- Compactly shows a classifier performance



Confusion matrix

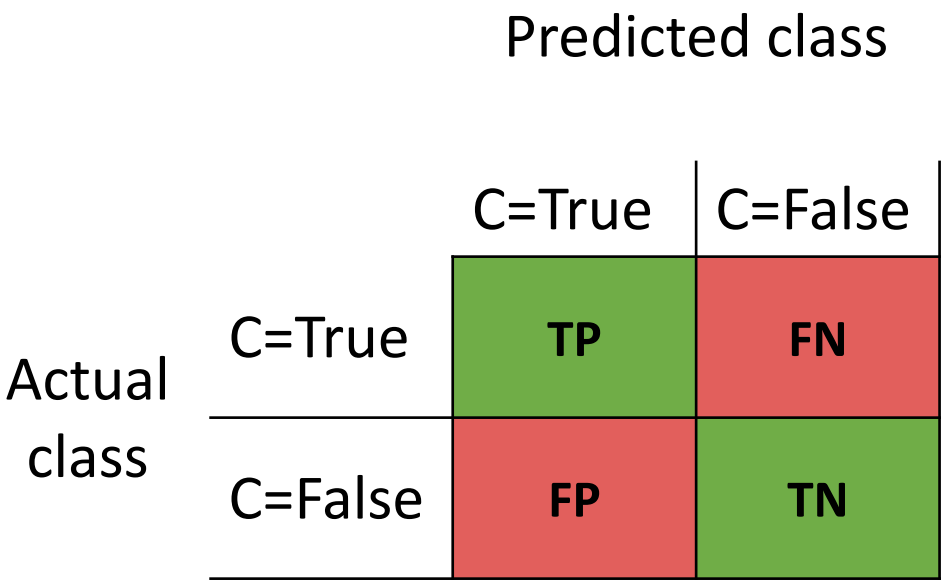
- Examples

Correct classification	Classified as	
	democrat	republican
democrat	81 (97.6%)	2 (2.4%)
republican	6 (11.5%)	46 (88.5%)

Correct classification	Classified as					
	1	2	3	5	6	7
1	52	10	7	0	0	1
2	15	50	6	2	1	2
3	5	6	6	0	0	0
5	0	2	0	10	0	1
6	0	1	0	0	7	1
7	1	3	0	1	0	24

Source: Bramer, M. (2016). *Principles of data mining* (3rd edition). London: Springer.

Confusion matrix: metrics



$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

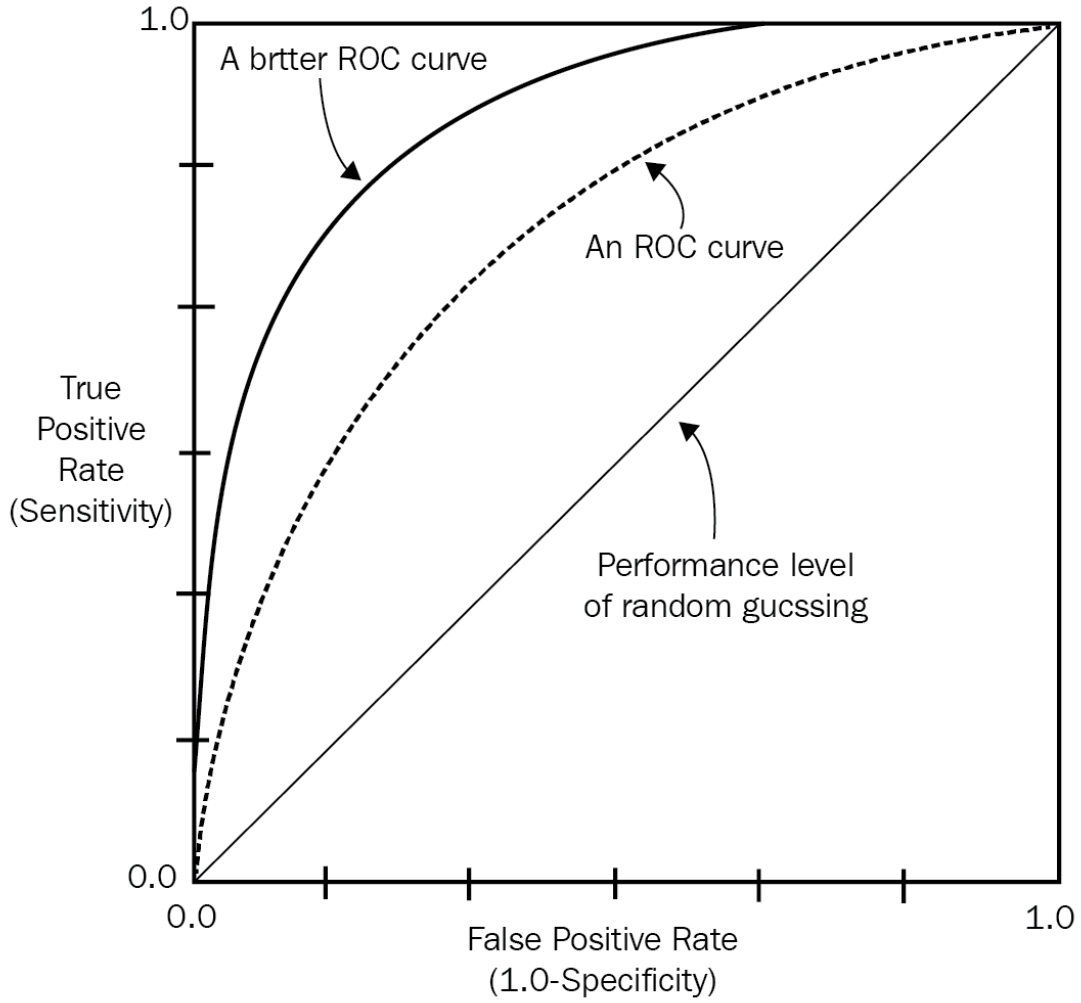
$$Sensitivity\ (recall) = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F1\ score = \frac{2 * Precision * Sensitivity}{Precision + Sensitivity}$$

Receiver operating characteristic (ROC) curve



Source: Ciaburro, G., & Venkateswaran, B. (2017). *Neural Networks with R: Smart models using CNN, RNN, deep learning, and artificial intelligence principles*. Packt Publishing Ltd.

Numerical prediction error measures

- Calculates the difference between predicted ($\hat{Y}$) vs. expected (Y) values.

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i|$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$$

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n} \left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

Sources

- Bramer, M. (2016). *Principles of data mining* (3rd edition). London: Springer.
- Clifton, C. (2019, December 20). *data mining*. *Encyclopedia Britannica*. https://www.britannica.com/technology/data-mining
- Copeland, B. (2020, August 11). *artificial intelligence*. *Encyclopedia Britannica*. https://www.britannica.com/technology/artificial-intelligence
- Hosch, William L.. "machine learning". *Encyclopedia Britannica*, 29 Jul. 2021, https://www.britannica.com/technology/machine-learning. Accessed 25 October 2021.
- Menéndez González V. Evaluating Machine Learning Models [version 1; not peer reviewed]. *F1000Research* 2020, **9**:329 (slides) (https://doi.org/10.7490/f1000research.1117855.1)
- https://cloud.google.com/ai-platform/training/docs/hyperparameter-tuning-overview
- Zheng, A. (2015). *Evaluating machine learning models: a beginner's guide to key concepts and pitfalls*. O'Reilly Media.