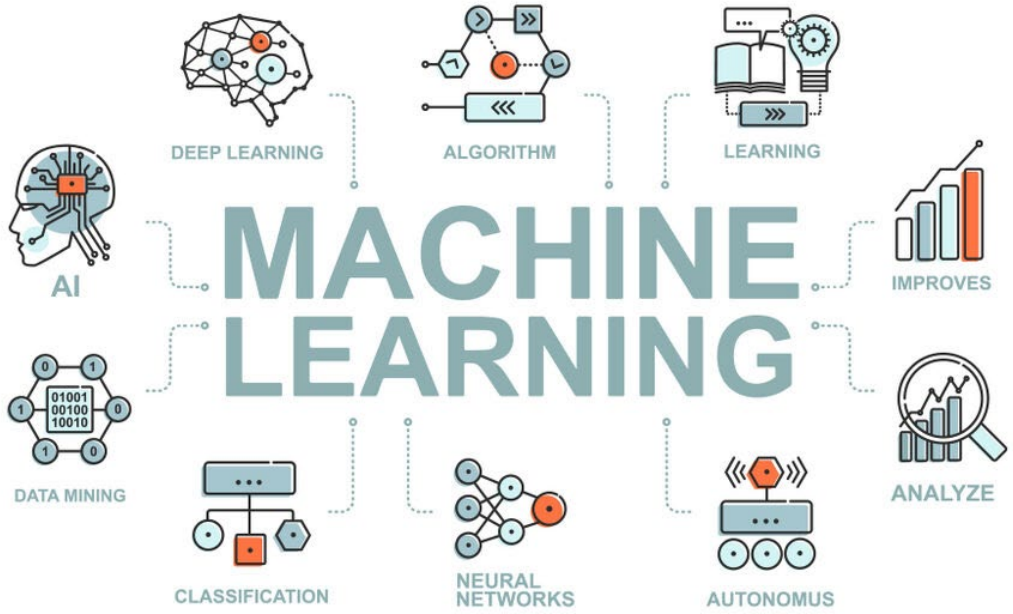
TONI MONLEON . SECTION OF STATISTICS. FACULTY OF BIOLOGY. UB
4/2023

Presentation. Definitions and concepts in ML.

Machine Learning



Toni Monleón-Getino, amonleong@ub.edu
04/25/2023

Definitions and concepts in ML

- What is meant by Machine Learning?
- History
- Types of problems to solve:
  - Regression
  - Classification
  - Supervised and unsupervised methods
- What are the training sets
- Validation techniques (Training, test, validation)
- Model errors: Bias & variance
- What is overfitting?
- Techniques for dealing with overfitting (e.g. Cross validation)
- Evaluation metrics

Free use material, for educational use only. Efforts have been made to cite all sources (own, books, internet)

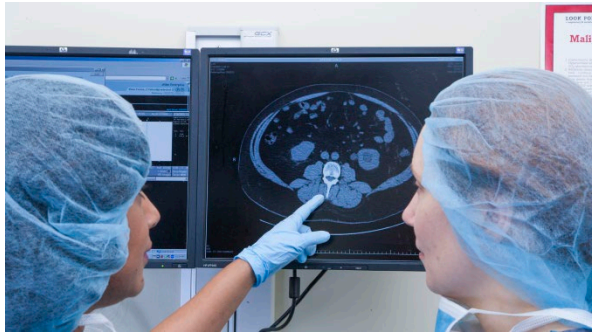
What is meant by Machine Learning?

- Why “Learn” ?
  - Machine learning is programming computers to optimize a performance criterion using example data or past experience.
  - There is no need to “learn” to calculate a mean
  - Learning is used when:
    - Human expertise does not exist (COVID19 pandemia),
    - Humans are unable to explain their expertise (medic diagnosis)
    - Solution changes in time (routing on a computer network)
    - Solution needs to be adapted to particular cases (user biometrics)

Learning to Machine Learning

- What We Talk About “Learning”:
  - Learning general models from a data of particular examples
  - Data is cheap and abundant (data warehouses, data marts); knowledge is expensive and scarce.
  - Example in retail: extract image characteristics to improve diagnostics (see examples in Kaggle https://www.kaggle.com/datasets)
  - Build a model that is a good and useful approximation to the data.

- Machine Learning:
  - Optimize a performance criterion using example data or past experience.
  - Role of Statistics: Inference from a sample
  - Role of Computer science: Efficient algorithms to:
    - Solve the optimization problem
    - Representing and evaluating the model for inference



Machine Learning: definitions

- Machine learning (ML) is a field of inquiry devoted to understanding and building methods that 'learn', that is, methods that leverage data to improve performance on some set of tasks.
- It is seen as a part of artificial intelligence.
- Machine learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so.
- Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, speech recognition, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.
- A subset of machine learning is closely related to computational statistics, which focuses on making predictions using computers, but not all machine learning is statistical learning.
- The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning.
- Data mining is a related field of study, focusing on exploratory data analysis through unsupervised learning. Some implementations of machine learning use data and neural networks in a way that mimics the working of a biological brain.
- In its application across business problems, machine learning is also referred to as predictive analytics
- Machine learning programs can perform tasks without being explicitly programmed to do so. It involves computers learning from data provided so that they carry out certain tasks. For simple tasks assigned to computers, it is possible to program algorithms telling the machine how to execute all steps required to solve the problem at hand; on the computer's part, no learning is needed. For more advanced tasks, it can be challenging for a human to manually create the needed algorithms. In practice, it can turn out to be more effective to help the machine develop its own algorithm, rather than having human programmers specify every needed step.

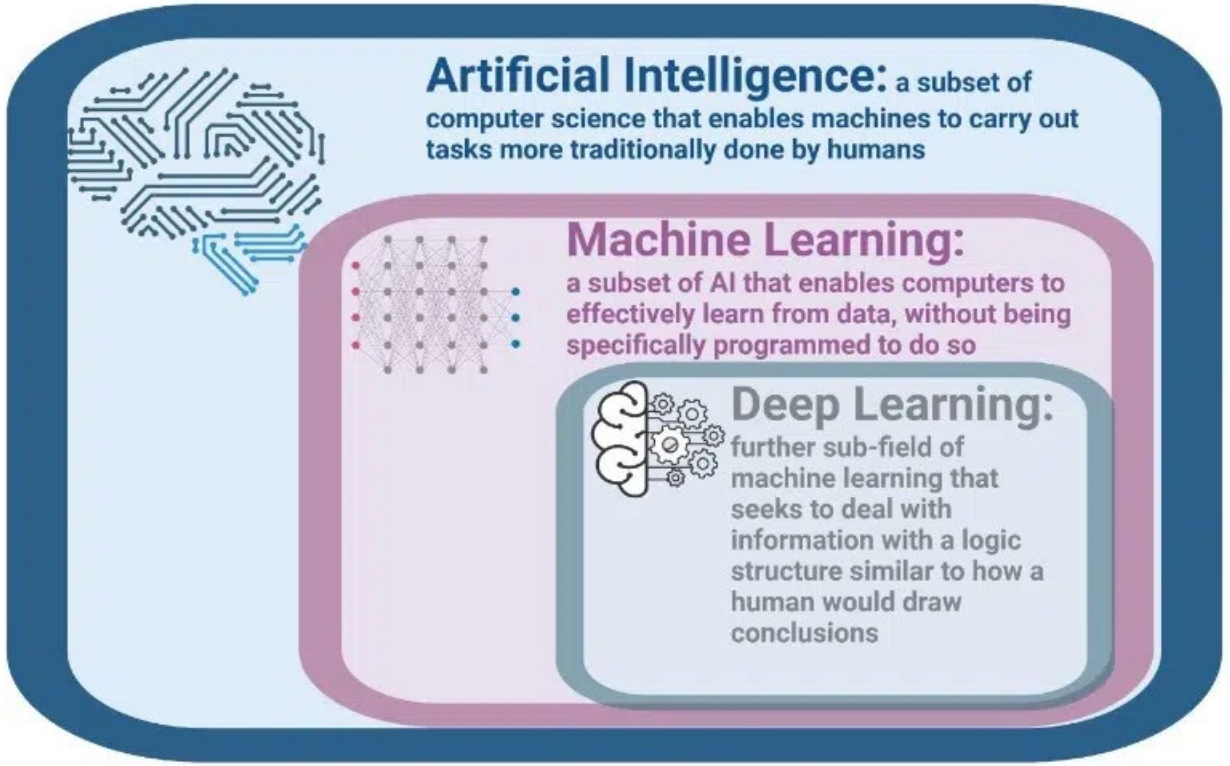
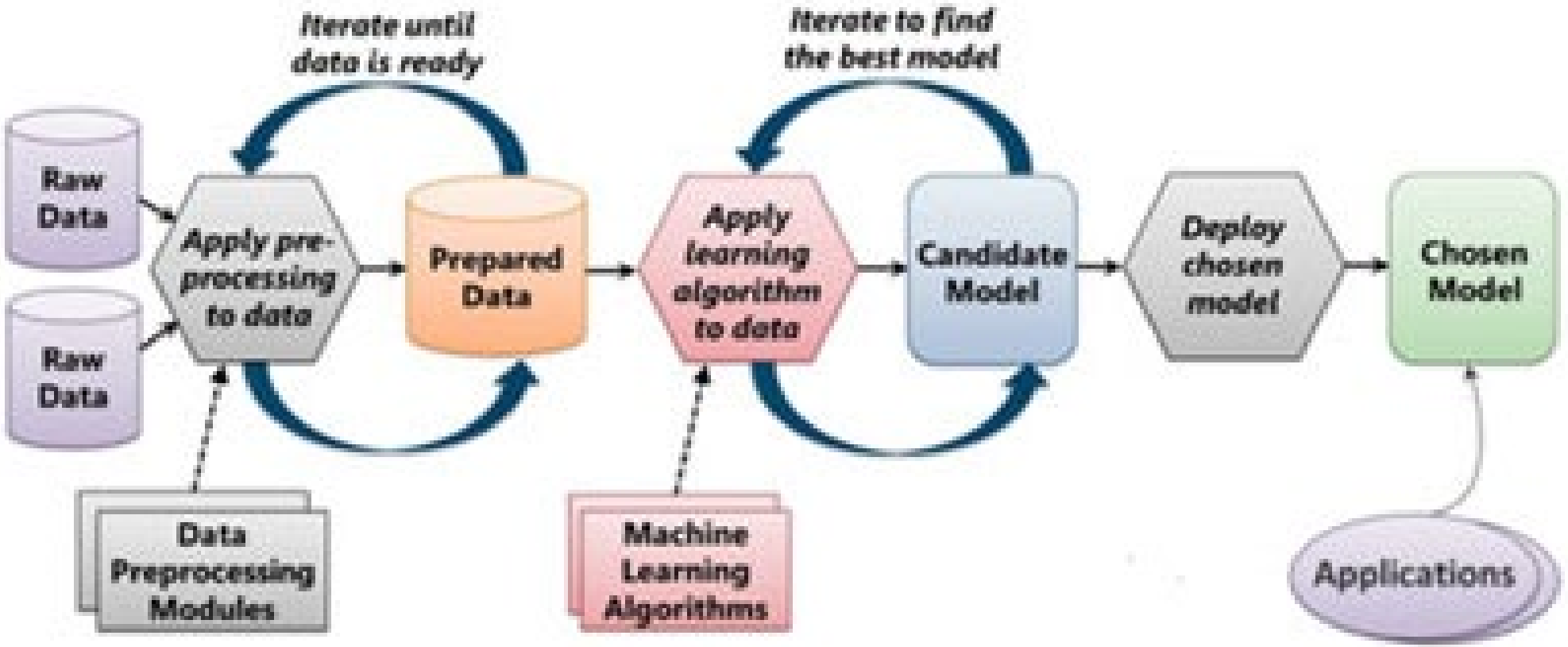


Figure 1: The three key terms in AI and how they are related

Internet

The Machine Learning Process

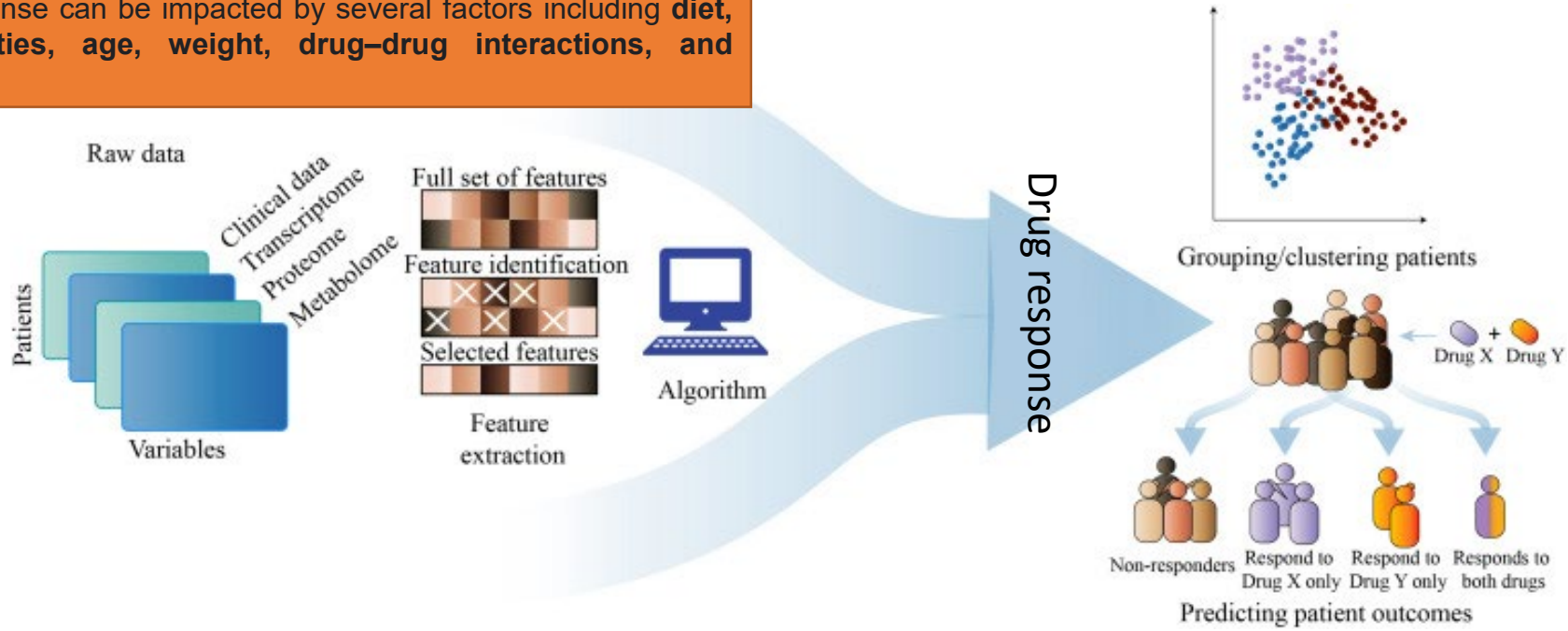


From "Introduction to Microsoft Azure" by David Chappell

Example of ML in translational medicine

Drug response can be impacted by several factors including **diet, comorbidities, age, weight, drug–drug interactions, and genetics**

https://www.thelancet.com/article/S2352-3964%2819%2930549-3/fulltext

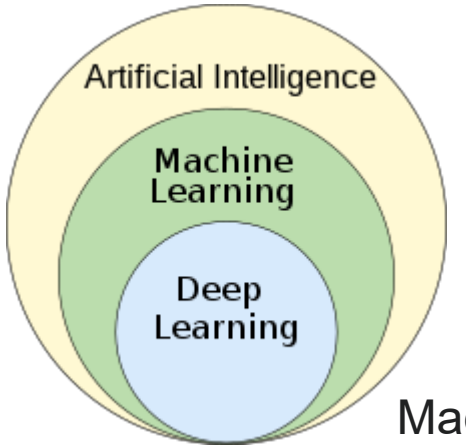


Process of AI/ML in translational medicine. A number of high-throughput assays generate data from many patient samples. Datasets are then structured into machine-readable format and potentially important variables are identified using an ML algorithm. The algorithm will learn relationships between the variables and may perform intelligent tasks such as grouping patients or predicting their outcomes.

History

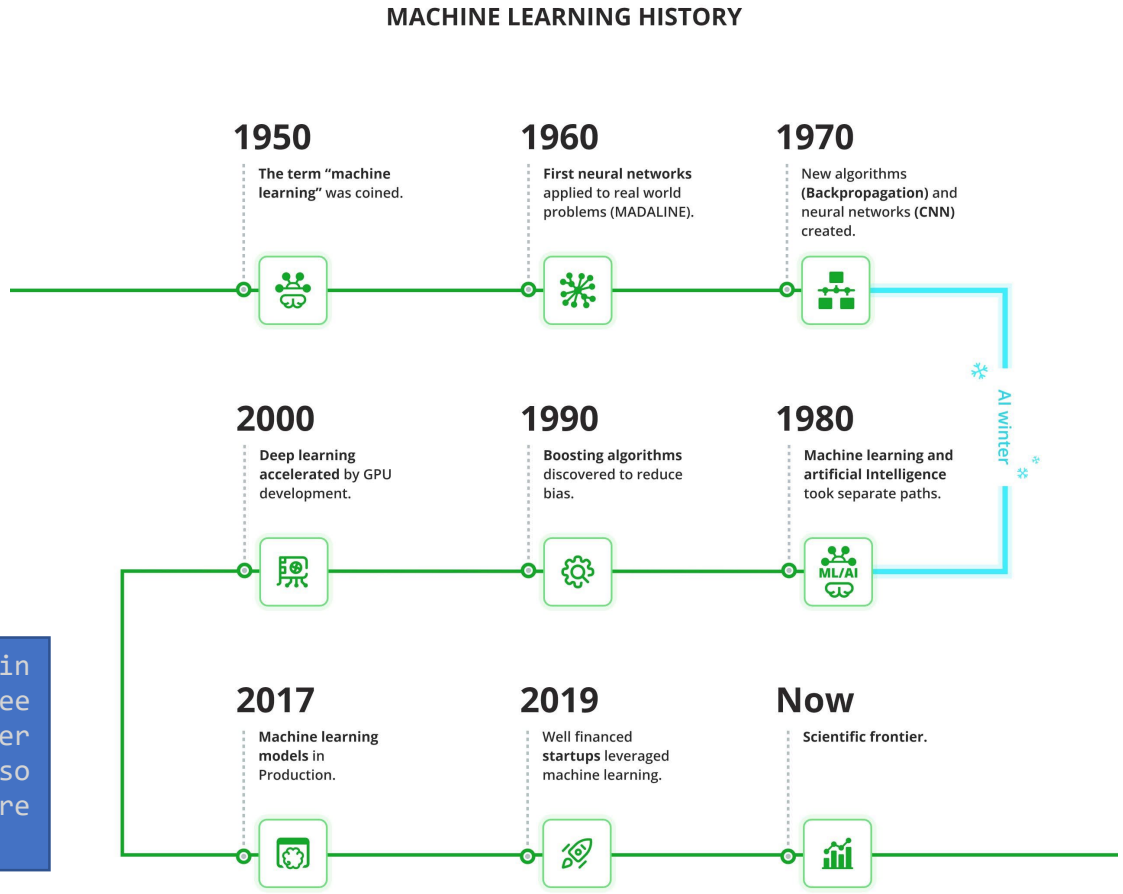
Machine learning was **first conceived from the mathematical modeling of neural networks**. A paper by logician Walter Pitts and neuroscientist Warren McCulloch, published in 1943, attempted to mathematically map out thought processes and decision making in human cognition

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



Machine learning as subfield of AI[23]

The term machine learning was coined in 1959 by Arthur Samuel, an IBM employee and pioneer in the field of computer gaming and artificial intelligence. Also the synonym self-teaching computers were used in this time period.



https://mobidev.biz/blog/future-machine-learning-trends-impact-business

Types of problems to solve

- **Categories of Machine Learning**
  - At the most fundamental level, machine learning can be categorized into two main types: **supervised learning** and **unsupervised learning**
  - **Supervised learning:** involves somehow modeling the relationship between measured features of data and some label associated with the data; once this model is determined, it can be used to apply labels to new, unknown data.
    - This is further subdivided into **classification tasks** and **regression tasks. The labels:**
      - in classification, the labels are discrete categories (ex: disease1, disease2,…)
      - while in regression, the labels are continuous quantities (ex: biomarker concentration 1.23, 1.45, 2,…)
    - We will see examples of both types of supervised learning in the following section.

  - **Unsupervised learning:** involves modeling the features of a dataset without reference to any label, and is often described as "letting the dataset speak for itself."
  - These models include tasks such as **clustering** and **dimensionality reduction**.
    - Clustering algorithms identify distinct groups of data, while dimensionality reduction algorithms search for more succinct representations of the data.
    - We will see examples of both types of unsupervised learning in the following section.

  - In addition, there are so-called **semi-supervised learning** methods, which falls somewhere between supervised learning and unsupervised learning:
    - Semi-supervised learning methods are often useful when only incomplete labels are available.

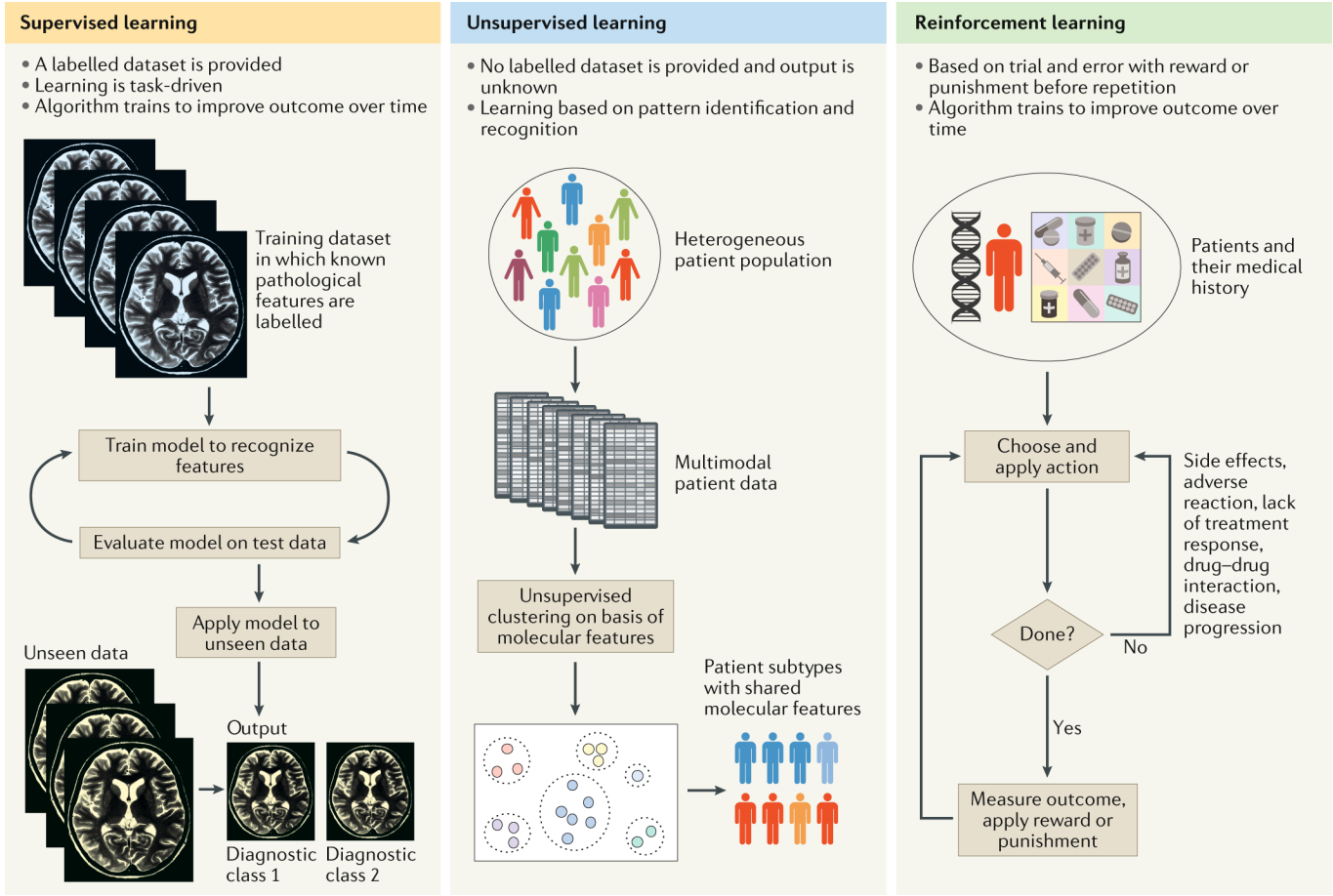
From Python Data Science Handbook](http://shop.oreilly.com/product/0636920034919.do) by Jake VanderPlas; the content is available [on GitHub](https://github.com/jakevdp/PythonDataScienceHandbook).

Machine Learning Categories

- At the most fundamental level, machine learning can be categorized into two main types: supervised learning and unsupervised learning

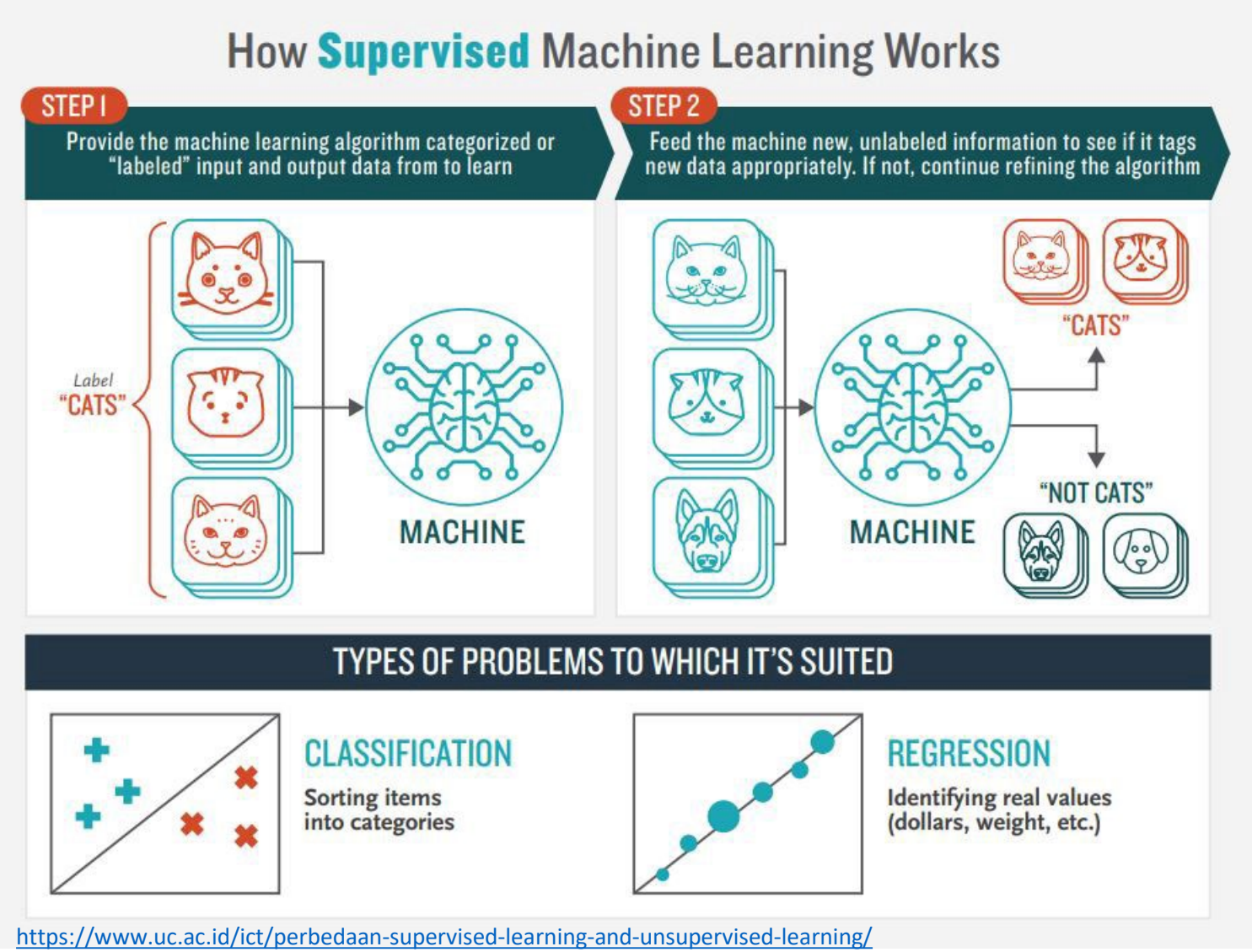
  - **Supervised learning** involves somehow modeling the relationship between measured features of data and some label associated with the data; once this model is determined, it can be used to apply labels to new, unknown data.

  - **Unsupervised learning** involves modeling the features of a dataset without reference to any label, and is often described as "letting the dataset speak for itself."



https://www.nature.com/articles/s41582-020-0377-8

Supervised machine learning works & Types of problems to solve

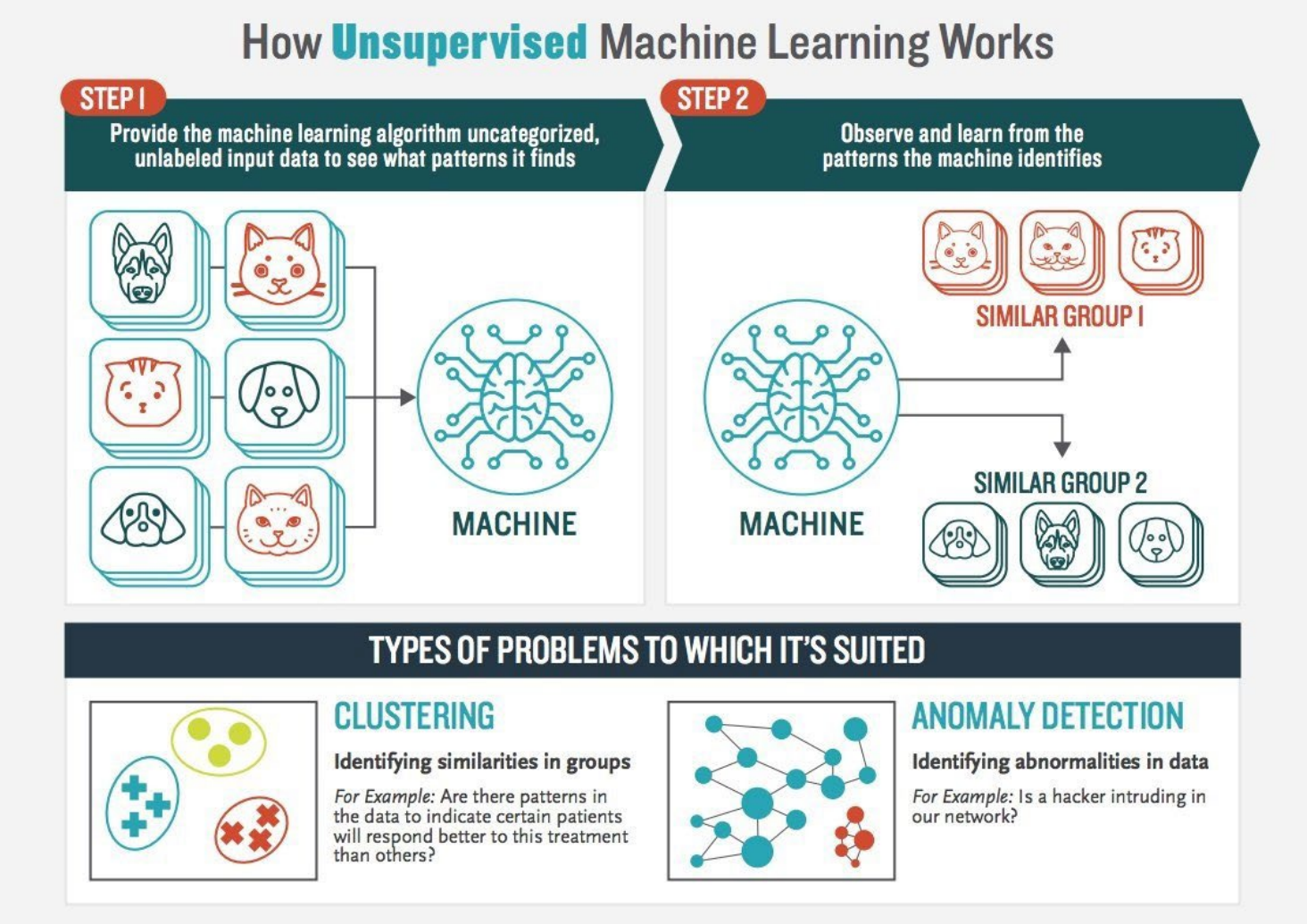


https://www.uc.ac.id/ict/perbedaan-supervised-learning-and-unsupervised-learning/

- This is further subdivided into classification tasks and regression tasks:
  - in classification, the labels are discrete categories,
  - while in regression, the labels are continuous quantities.

- We will see examples of both types of supervised learning in the following section.

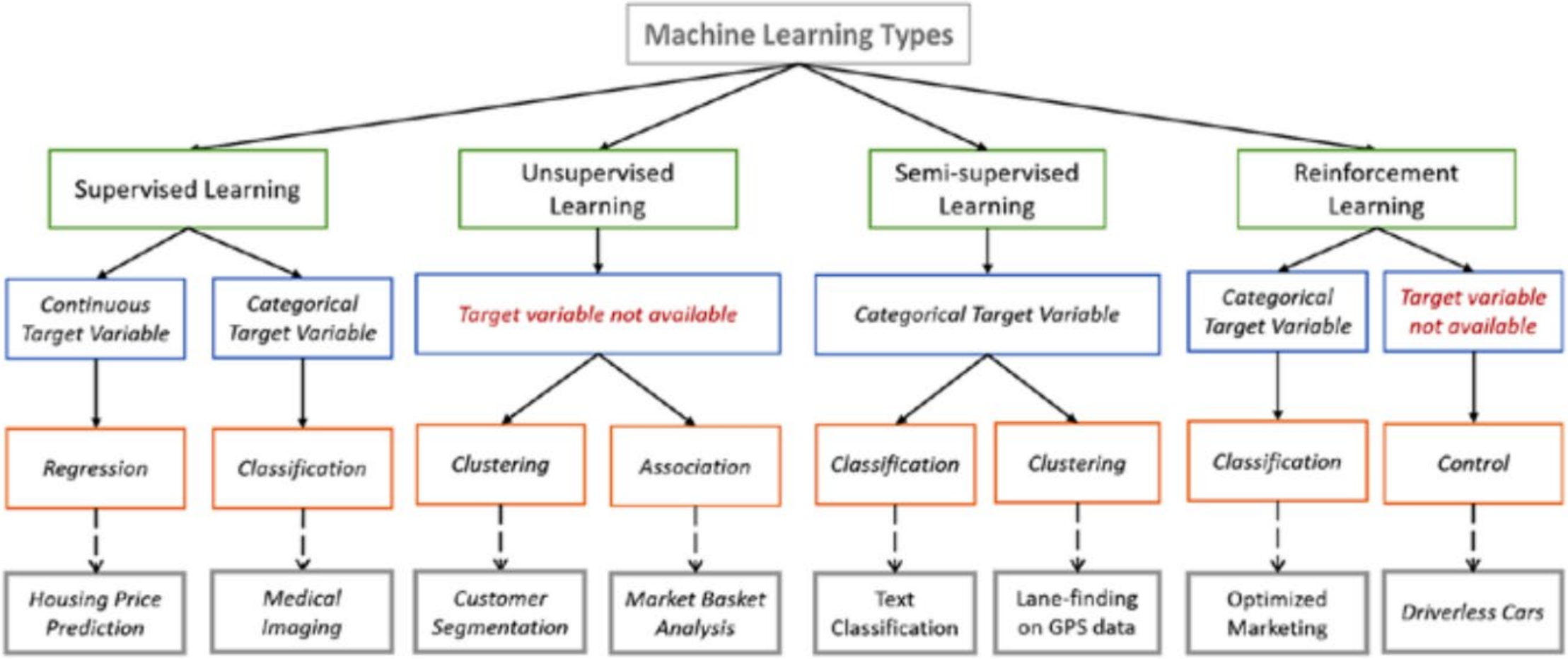
Unsupervised machine learning works & Types of problems to solve



- These models include tasks such as clustering and dimensionality reduction.
- Clustering algorithms identify distinct groups of data, while dimensionality reduction algorithms search for more succinct representations of the data.
- We will see examples of both types of unsupervised learning in the following section.

- In addition, there are so-called *semi-supervised learning methods, which falls somewhere between supervised learning and unsupervised learning.
- Semi-supervised learning methods are often useful when only incomplete labels are available

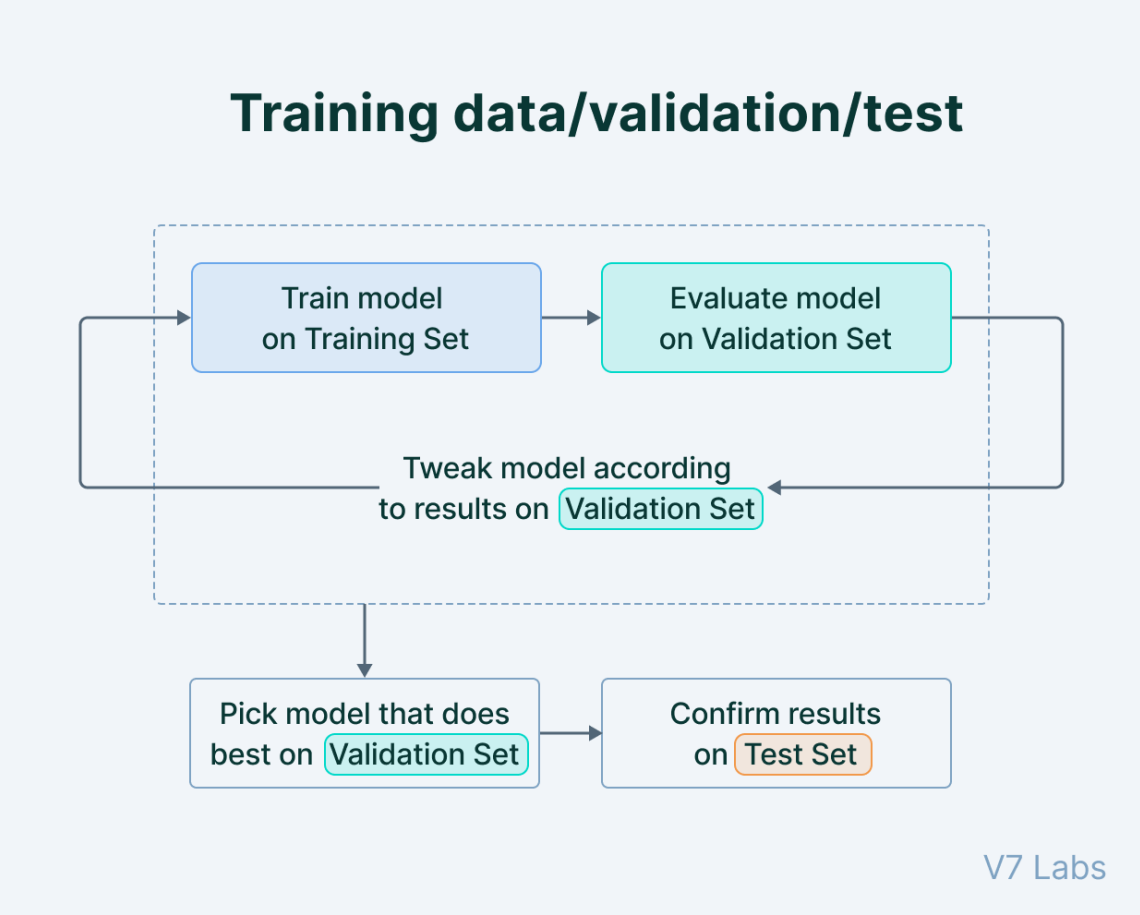
Algorithms in ML



Interactive map: https://chart-studio.plotly.com/create/?fid=SolClover:40#/

What are the training sets

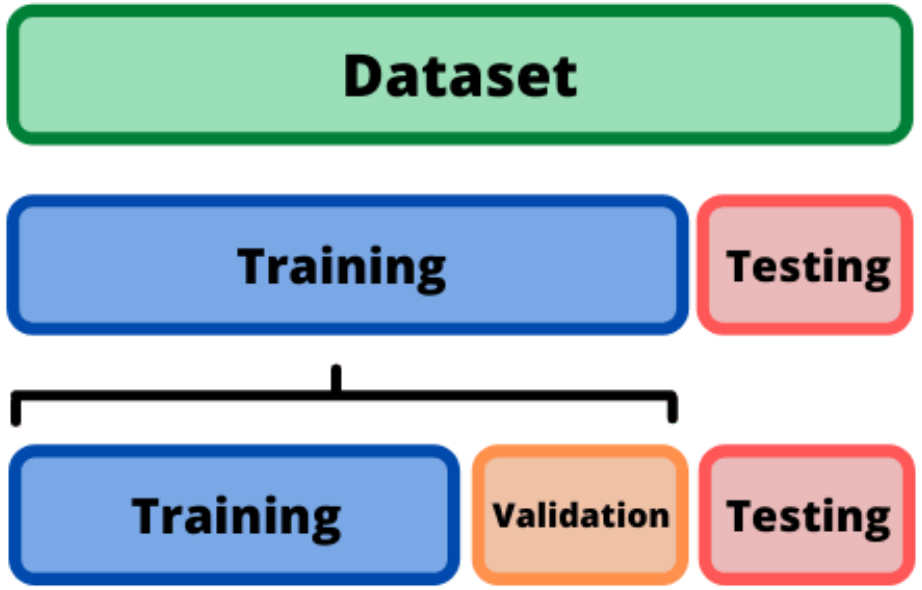
- The creation of different samples and splits in the dataset helps us judge the true model performance.
- The dataset split ratio depends on the number of samples present in the dataset and the model.
- Some common inferences that can be derived on dataset split include:
  - If there are several hyperparameters to tune, the machine learning model requires a larger validation set to optimize the model performance. Similarly, if the model has fewer or no hyperparameters, it would be easy to validate the model using a small set of data.
  - If a model use case is such that a false prediction can drastically hamper the model performance—like falsely predicting cancer—it's better to validate the model after each epoch to make the model learn varied scenarios.



https://www.v7labs.com/blog/train-validation-test-set

Training and Test data

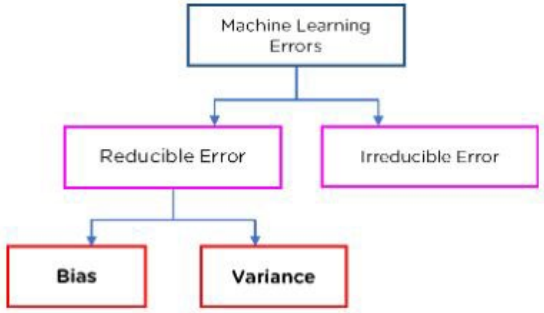
- In Data Science, training data and testing data are two major roles. Evaluating the performance of a built model is just as significant as training and building the model because a model with unevaluated performance may produce false predictions and lead to serious complications. In order to prevent such situations and to ensure the accuracy of the predictions, you must test and validate the model well enough
- To build and evaluate the performance of a machine learning model, we usually break our dataset into two distinct datasets. These two datasets are the training data and test data. Let's have a closer look at each of these sub-datasets.
- Training data:
  - Training data are the sub-dataset which we use to train a model. These datasets contain data observations in a particular domain. Algorithms study the hidden patterns and insights which are hidden inside these observations and learn from them. The model will be trained over and over again using the data in the training set machine learning and continue to learn the features of this data. Later we can deploy the trained model and have accurate predictions over new data. These predictions will be based on the learnings from the training dataset.
- Test data:
  - In Machine learning Test data is the sub-dataset that we use to evaluate the performance of a model built using a training dataset. Although we extract both train and test data from the same dataset, the test dataset should not contain any training dataset data. The purpose of creating a model is to predict unknown results. The test data is used to check the performance, accuracy, and precision of the model created using training data

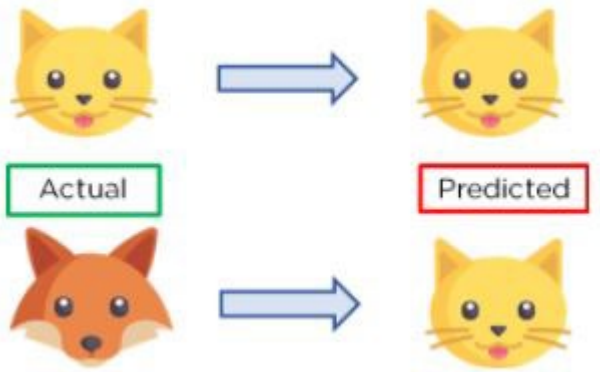


https://sdsclub.com/how-to-train-and-test-data-like-a-pro/

Evaluation metrics: Errors

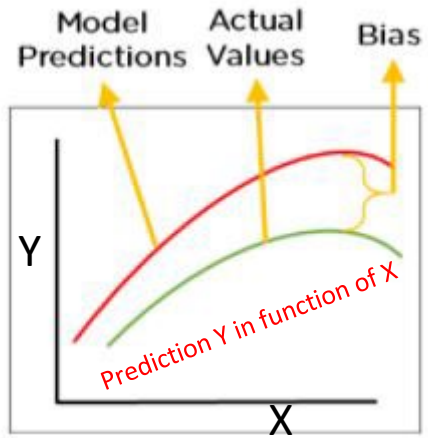
- There are two main types of errors present in any machine learning model: a)Reducible Errors and b)Irreducible Errors:
  - Irreducible errors are errors which will always be present in a machine learning model, because of unknown variables, and whose values cannot be reduced.
  - Reducible errors are those errors whose values can be further reduced to improve a model. They are caused because our model's output function does not match the desired output function and can be optimized.
  - We can further divide reducible errors into two:
    - Bias
    - Variance
- **Bias:**
  - To make predictions, our model will analyze our data and find patterns in it.
  - Using these patterns, we can make generalizations about certain instances in our data.
  - Our model after training learns these patterns and applies them to the test set to predict them.
  - Bias is the difference between our actual and predicted values.
  - Bias is the simple assumptions that our model makes about our data to be able to predict new data.
  - When the Bias is high, assumptions made by our model are too basic, the model can't capture the important features of our data. This means that our model hasn't captured patterns in the training data and hence cannot perform well on the testing data too. If this is the case, our model cannot perform on new data and cannot be sent into production.
- **Variance:**
  - We can define variance as the model's sensitivity to fluctuations in the data (noise).
  - Variance is the very opposite of Bias: During training, it allows our model to 'see' the data a certain number of times to find patterns in it. If it does not work on the data for long enough, it will not find patterns and bias occurs. On the other hand, if our model is allowed to view the data too many times, it will learn very well for only that data. It will capture most patterns in the data, but it will also learn from the unnecessary data present, or from the noise.
  - Our model may learn from noise. This will cause our model to consider trivial features as important.





Example of Variance

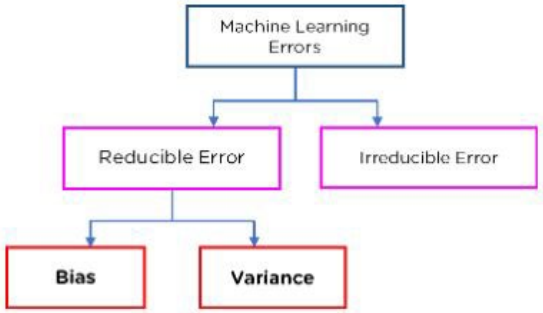
We can see that our model has learned extremely well for our training data, which has taught it to identify cats. But when given new data, such as the picture of a fox, our model predicts it as a cat, as that is what it has learned. This happens when the Variance is high, our model will capture all the features of the data given to it, including the noise, will tune itself to the data, and predict it



Example of Bias

https://www.simplilearn.com/tutorials/machine-learning-tutorial/bias-and-variance

Evaluation metrics: Bias & Variance



Cheat Sheet – Bias-Variance Tradeoff

What is Bias? $bias = \mathbb{E}[f'(x)] - f(x)$

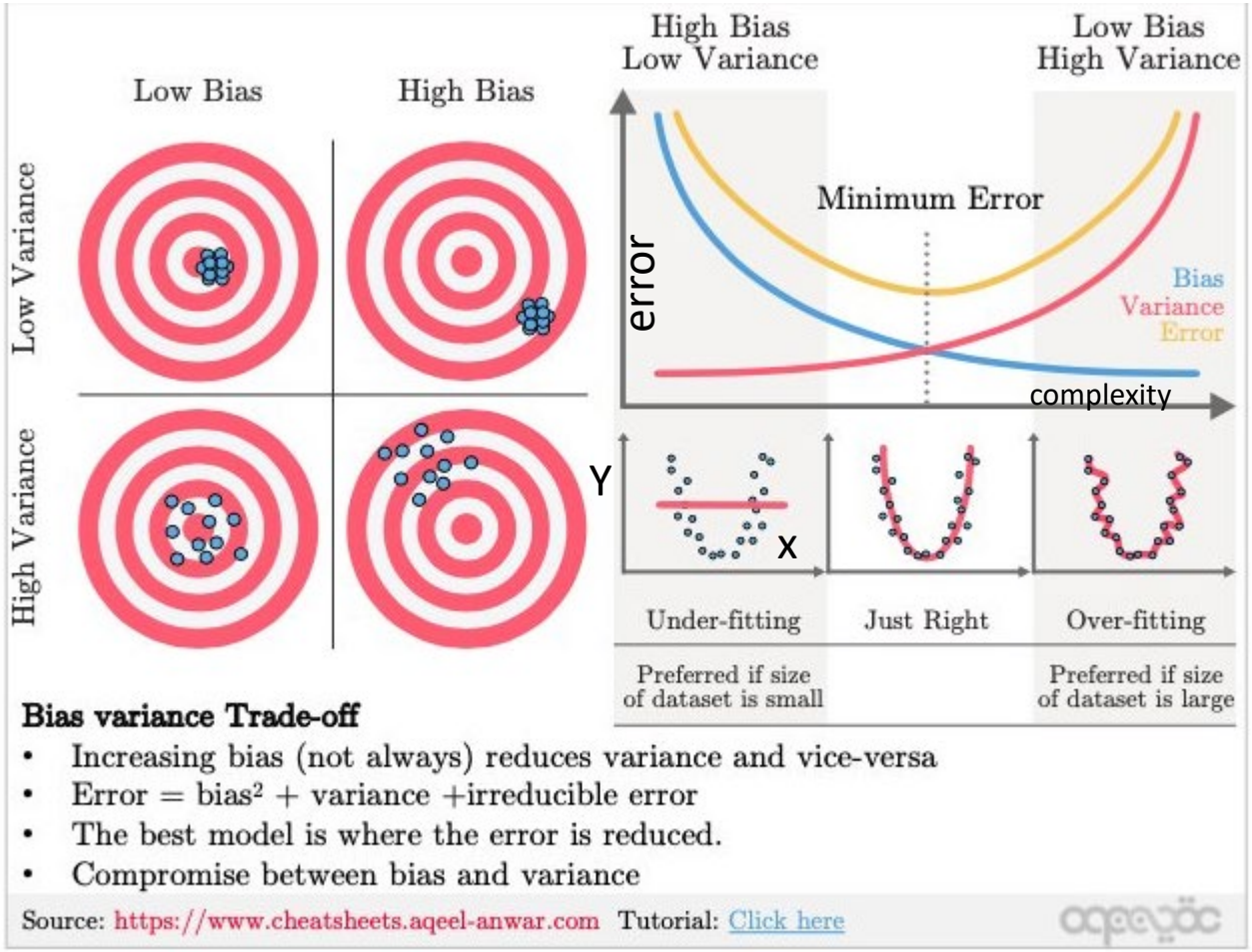
- Error between average model prediction and ground truth
- The bias of the estimated function tells us the capacity of the underlying model to predict the values

What is Variance? $variance = \mathbb{E}\left[(f'(x) - \mathbb{E}[f'(x)])^2\right]$

- Average variability in the model prediction for the given dataset
- The variance of the estimated function tells you how much the function can adjust to the change in the dataset

High Bias
- → Overly-simplified Model
- → Under-fitting
- → High error on both test and train data

High Variance
- → Overly-complex Model
- → Over-fitting
- → Low error on train data and high on test
- → Starts modelling the noise in the input



Bias variance Trade-off

- Increasing bias (not always) reduces variance and vice-versa
- Error = bias² + variance +irreducible error
- The best model is where the error is reduced.
- Compromise between bias and variance

Source: https://www.cheatsheets.aqeel-anwar.com Tutorial: Click here

Evaluation metrics: Bias

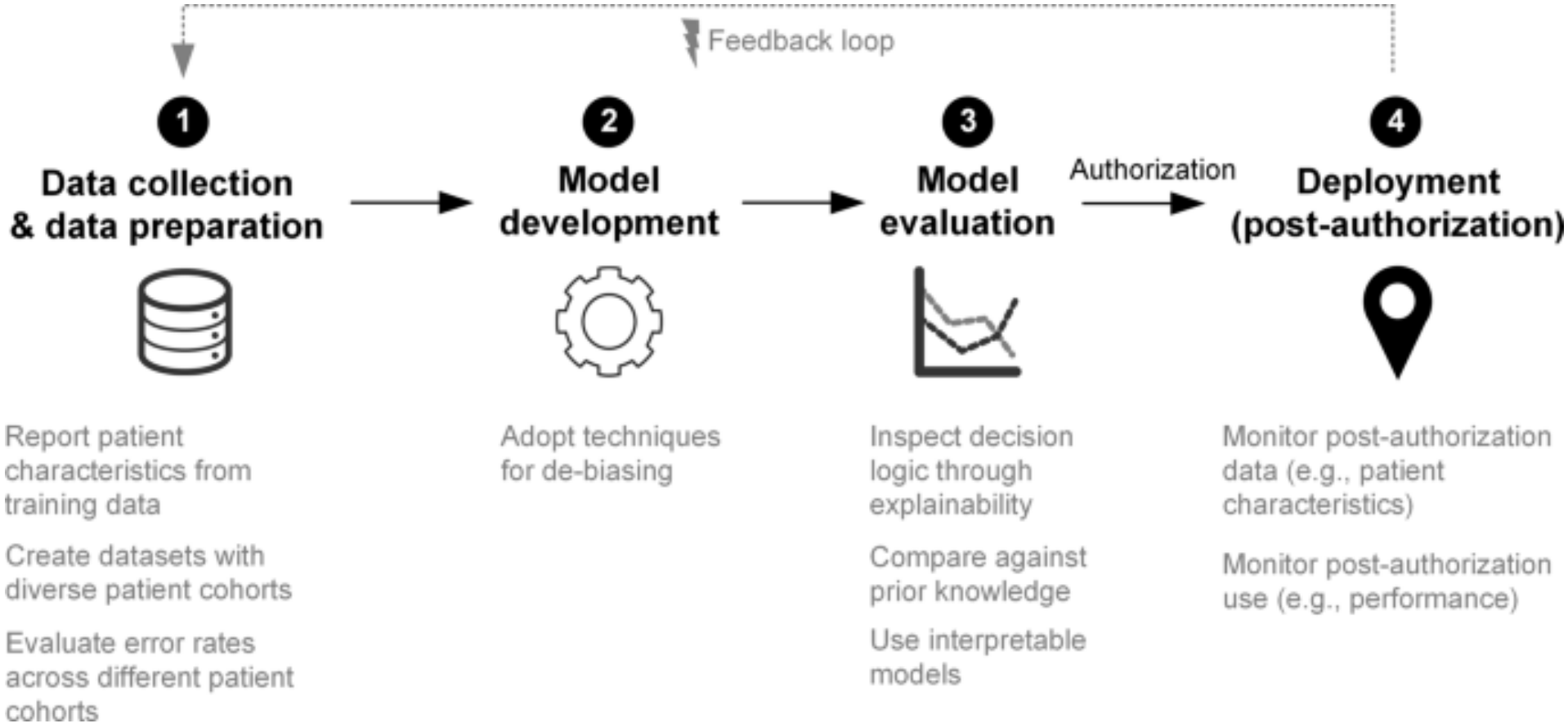
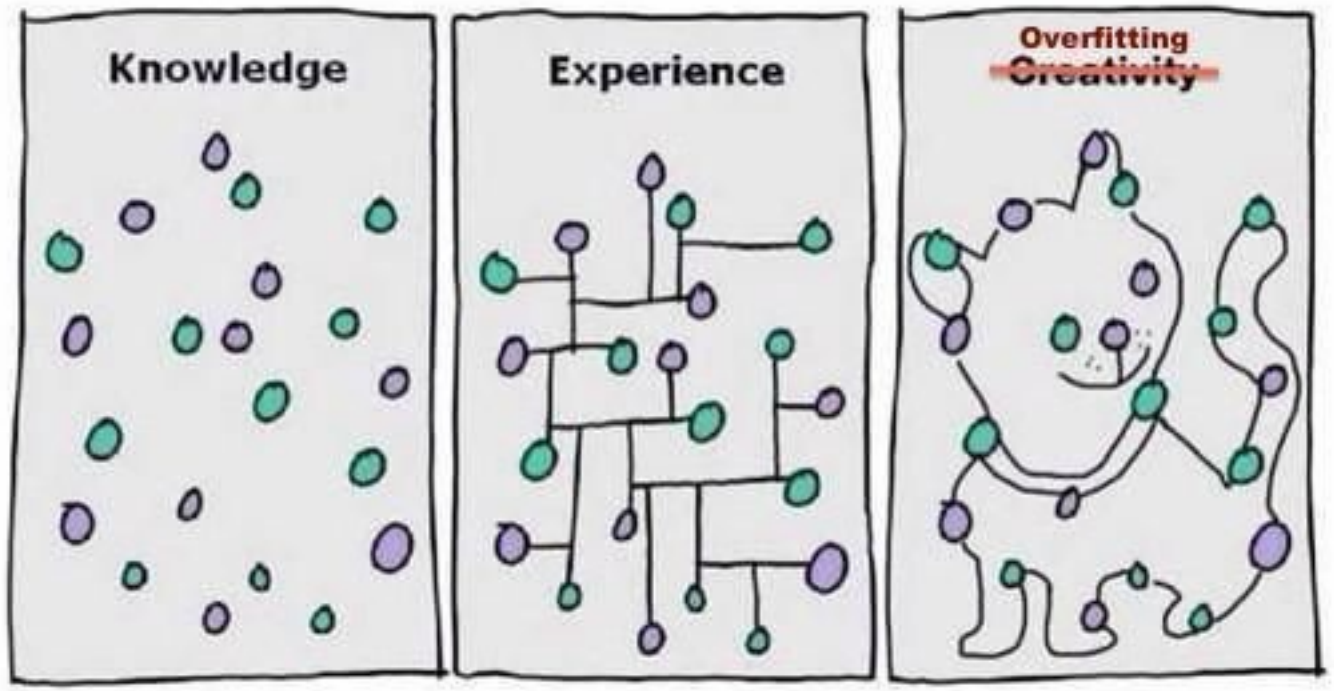
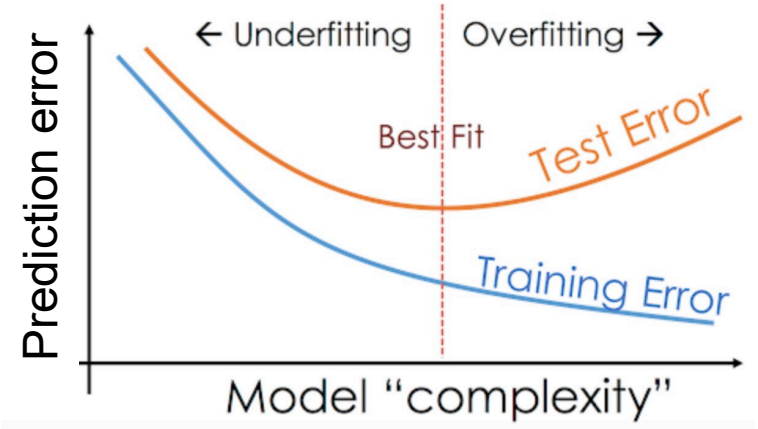


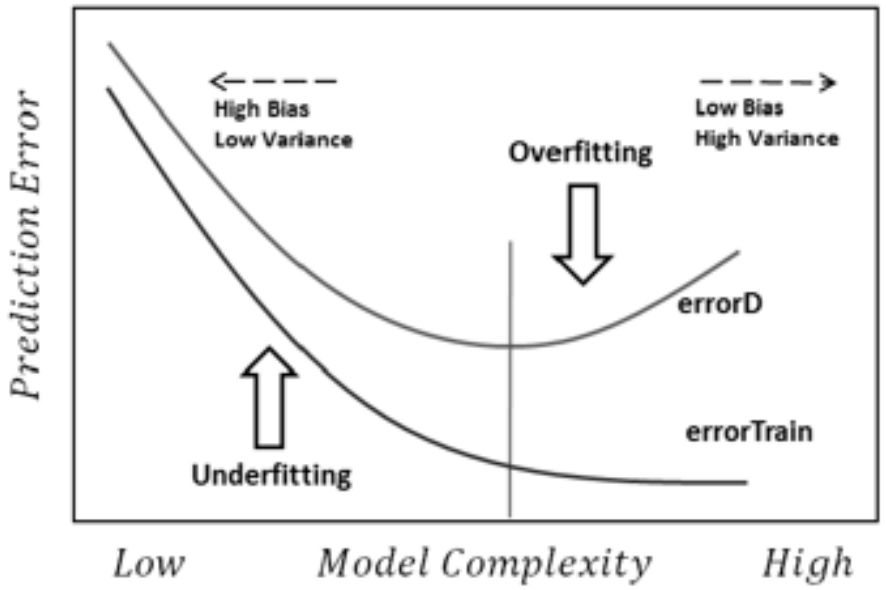
Diagram outlining proposed solutions on how to mitigate bias across the different development steps of ML-based systems for medical applications: (1) Data collection and data preparation, (2) Model development, (3) Model evaluation, and (4) Deployment.

https://www.nature.com/articles/s43856-021-00028-w

Model Complexity

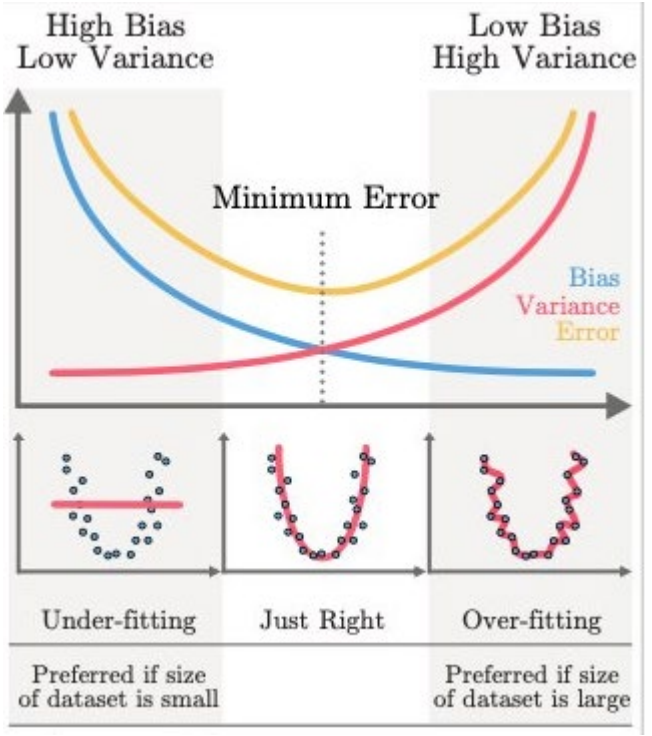


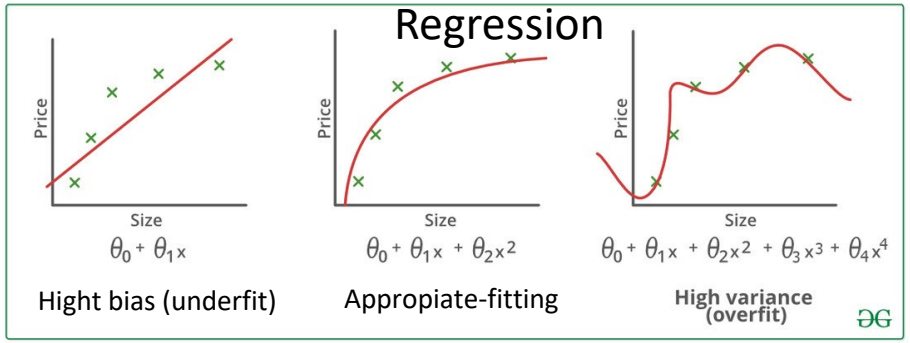


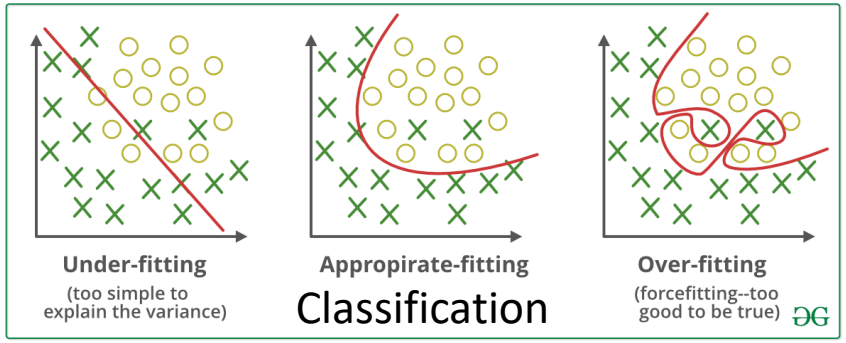


What is overfitting?

- Overfitting occurs when our machine learning model tries to cover all the data points, or more than the required data points present in the given dataset. Because of this, the model starts caching noise and inaccurate values present in the dataset, and all these factors reduce the efficiency and accuracy of the model.
- Reasons for Overfitting are as follows:
  - High variance and low bias
  - The model is too complex
  - The size of the training data
- Techniques to reduce overfitting:
  - Increase training data.
  - Reduce model complexity.
  - Early stopping during the training phase (have an eye over the loss over the training period as soon as loss begins to increase stop training).
  - Ridge Regularization and Lasso Regularization
  - Use dropout for neural networks to tackle overfitting.





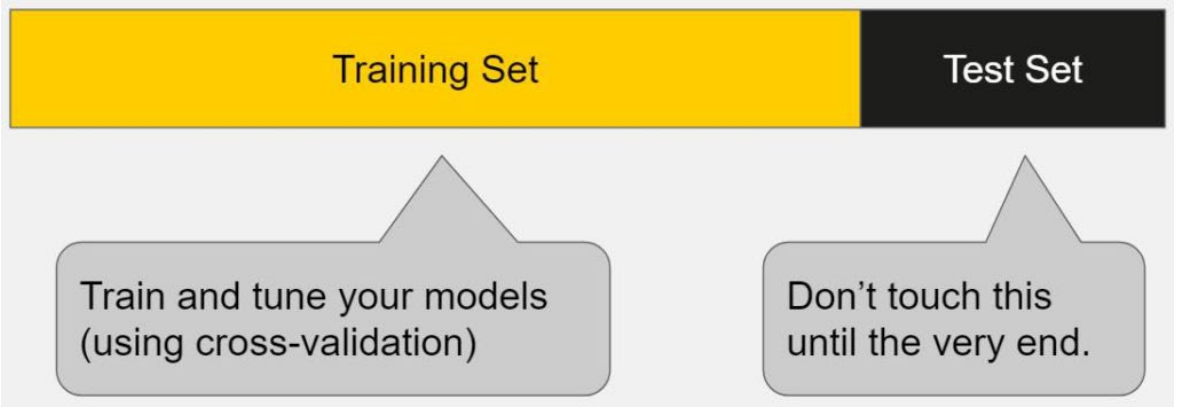


Overfitting examples

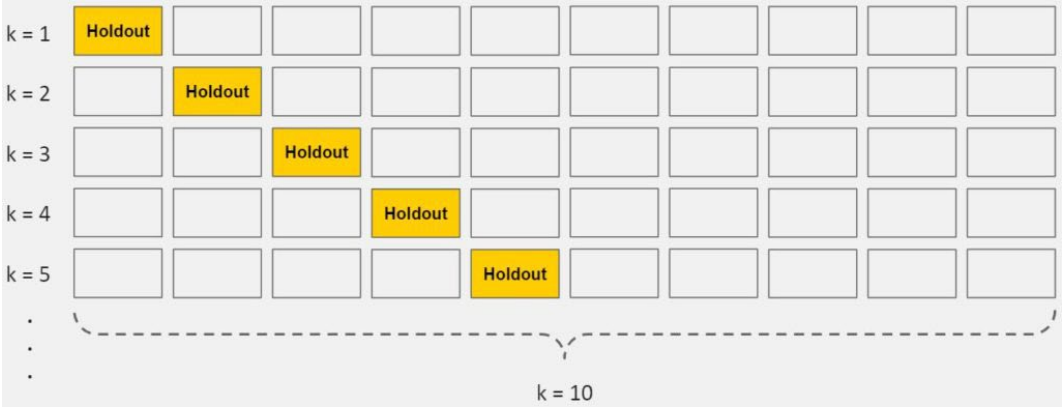
https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/

Cross validation

- How to Prevent Overfitting in Machine Learning?
- To address this, we can split our initial dataset into separate training and test subsets.
- Detecting overfitting is useful, but it doesn't solve the problem. Fortunately, you have several options to try: like use cross-validation
- Cross-validation:
  - Cross-validation is a powerful preventative measure against overfitting.
  - The idea is clever: Use your initial training data to generate multiple mini train-test splits. Use these splits to tune your model.
  - In standard k-fold cross-validation, we partition the data into k subsets, called folds. Then, we iteratively train the algorithm on k-1 folds while using the remaining fold as the test set (called the "holdout fold").
  - Cross-validation allows you to tune hyperparameters with only your original training set. This allows you to keep your test set as a truly unseen dataset for selecting your final model.
  - We have another article with a more detailed breakdown of cross-validation in https://elitedatascience.com/overfitting-in-machine-learning



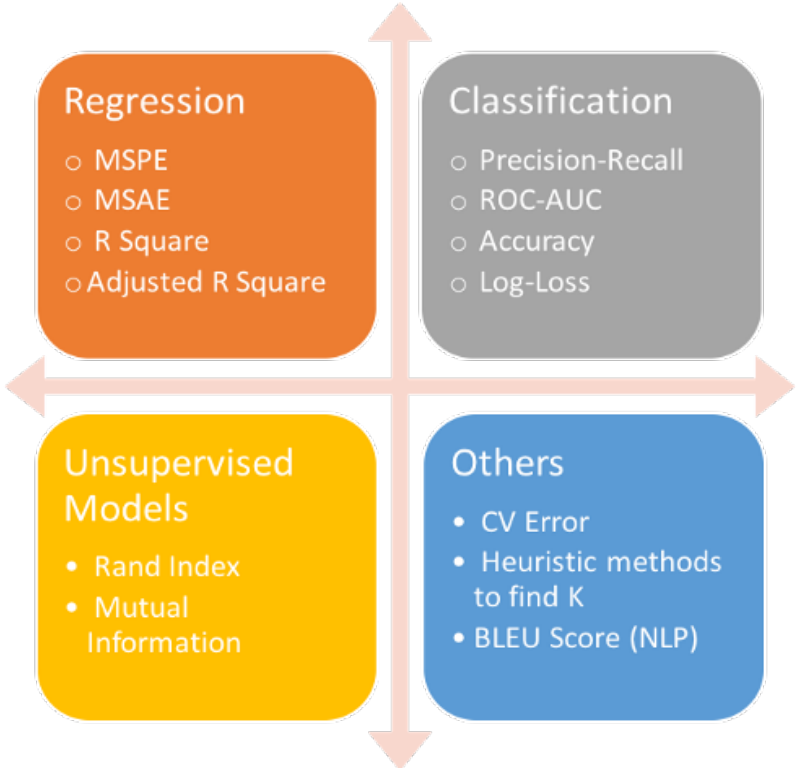
K-Fold Cross-Validation



https://elitedatascience.com/overfitting-in-machine-learning

Evaluation metrics: Model performance

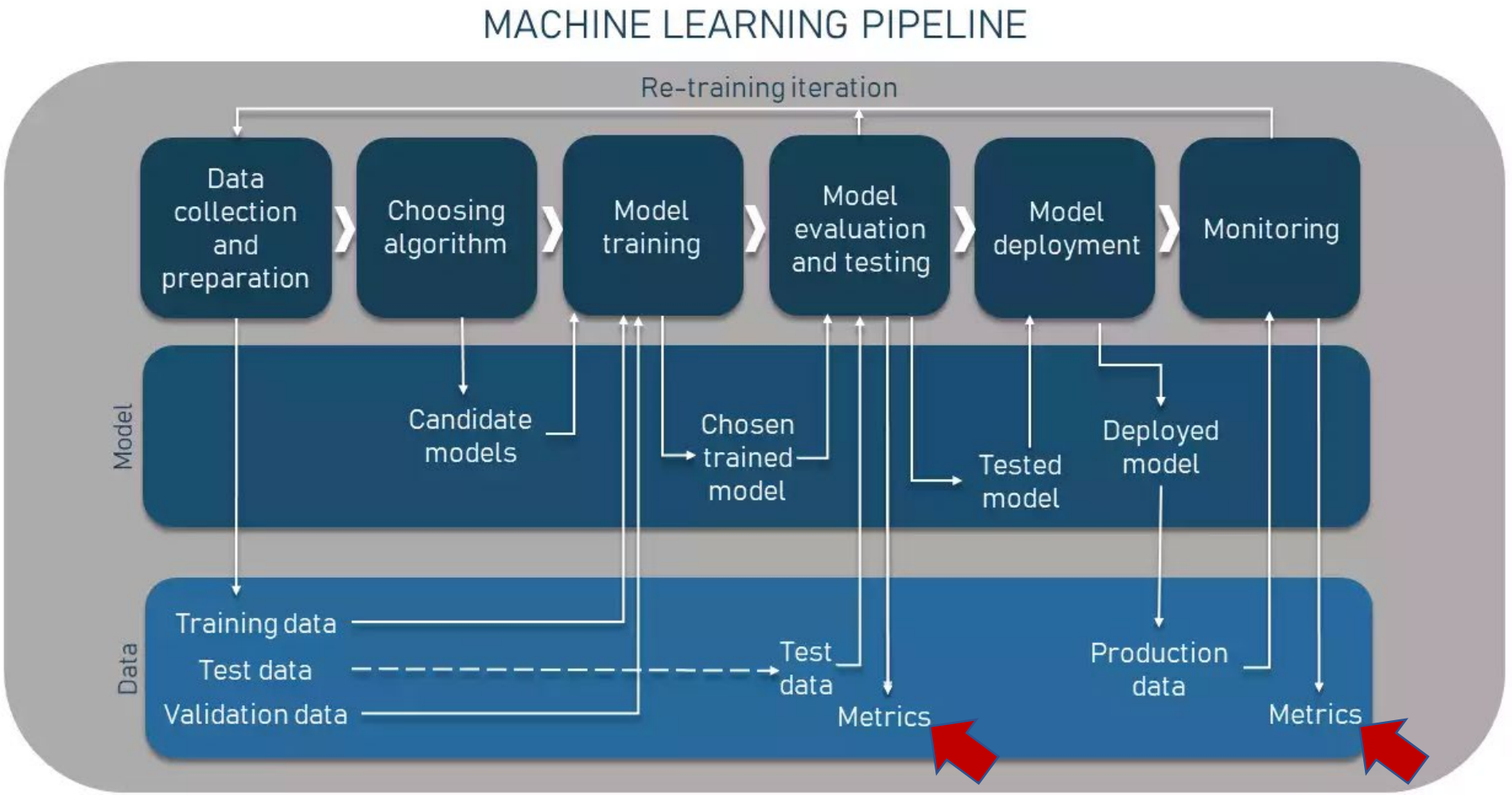
- The most important task in building any ML model is to evaluate its performance. So, the question arises that how would one measure the success of a machine learning model?
- Evaluation metrics are tied to machine learning tasks.
- There are different metrics for the tasks of classification and regression.
- Some metrics, like precision-recall, are useful for multiple tasks.
- Classification and regression are examples of supervised learning, which constitutes a majority of machine learning applications.
- Using different metrics for performance evaluation, we should be able to improve our model's overall predictive power before we roll it out for production on unseen data.
- Without doing a proper evaluation of the Machine Learning model by using different evaluation metrics, and only depending on accuracy, can lead to a problem when the respective model is deployed on unseen data and may end in poor predictions.



https://www.kdnuggets.com/2018/06/right-metric-evaluating-machine-learning-models-2.html

https://www.analyticsvidhya.com/blog/2021/07/metrics-to-evaluate-your-classification-model-to-take-the-right-decisions/

Metrics: When do metrics to evaluate models?

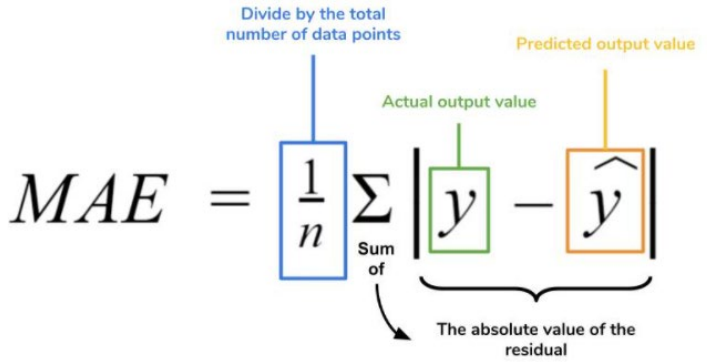


https://www.altexsoft.com/blog/machine-learning-metrics/

altexsoft

Evaluation metrics: Regression problems

- **MAE**: Mean Absolute Error is the average of the difference between the Original Values and the Predicted Values. It gives us the measure of how far the predictions were from the actual output. However, they don't gives us any idea of the direction of the error i.e. whether we are under predicting the data or over predicting the data.
- **MSE**: Mean Squared Error(MSE) is quite similar to Mean Absolute Error, the only difference being that MSE takes the average of the **square** of the difference between the original values and the predicted values. The advantage of MSE being that it is easier to compute the gradient, whereas Mean Absolute Error requires complicated linear programming tools to compute the gradient.
- **RMSE**: Root Mean Squared Error is the extension of MSE that allows you to get rid of the squared error by calculating the square root of the MSE result.
- **R^2**: In statistics, the coefficient of determination, denoted R^2 or r^2 and pronounced "R squared", is the proportion of the variation in the dependent variable that is predictable from the independent variable(s). R^2 explains the degree to which your input variables explain the variation of your output / predicted variable. So, if R-square is 0.8, it means 80% of the variation in the output variable is explained by the input variables. So, in simple terms, higher the R squared, the more variation is explained by your input variables and hence better is your model. However, the problem with R-squared is that it will either stay the same or increase with addition of more variables, even if they do not have any relationship with the output variables.



$$\text{MSE} = \frac{1}{N} \times \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

N - number of data samples
y_i - actual data value
$\hat{y}_i$ - predicted data value

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

N - number of data samples
y_i - actual data value
$\hat{y}_i$ - predicted data value

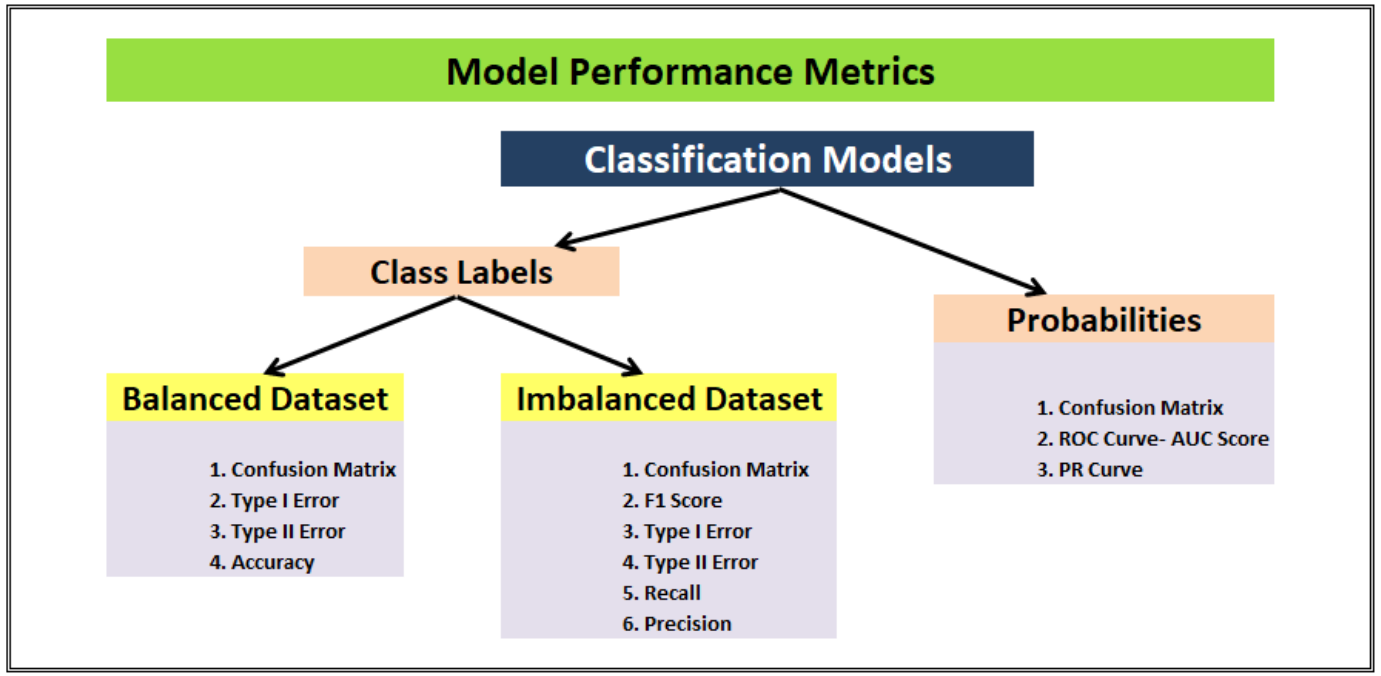
$$\hat{R}^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

https://www.researchgate.net/publication/342009715_Estandarizacion_de_metricas_de_rendimiento_para_clasificadores_Machine_y_Deep_Learning

https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

https://www.kdnuggets.com/2018/06/right-metric-evaluating-machine-learning-models-2.html

Metrics: Classification problems



		Actual class: Positive	Actual class: Negative	
Predicted class: Positive		**TP:** True Positive	**FP:** False Positive (Type I Error)	**Precision:** TP / (TP + FP)
Predicted class: Negative		**FN:** False Negative (Type II Error)	**TN:** True Negative	**Negative Predictive Value:** TN / (TN+FN)
		Recall or Sensitivity: TP / (TP + FN)	**Specificity:** TN / (TN + FP)	**Accuracy:** TP + TN / (TP + TN + FP + FN)

•True Positive (TP) — a class is predicted true and is true in reality (patients that are sick and diagnosed sick);
•True Negative (TN) — a class is predicted false and is false in reality (patients that are healthy and diagnosed healthy);
•False Positive (FP) — a class is predicted true but is false in reality (patients that are healthy but diagnosed sick); and
•False Negative (FN) — a class is predicted false but is true in reality (patients that are sick but diagnosed healthy)

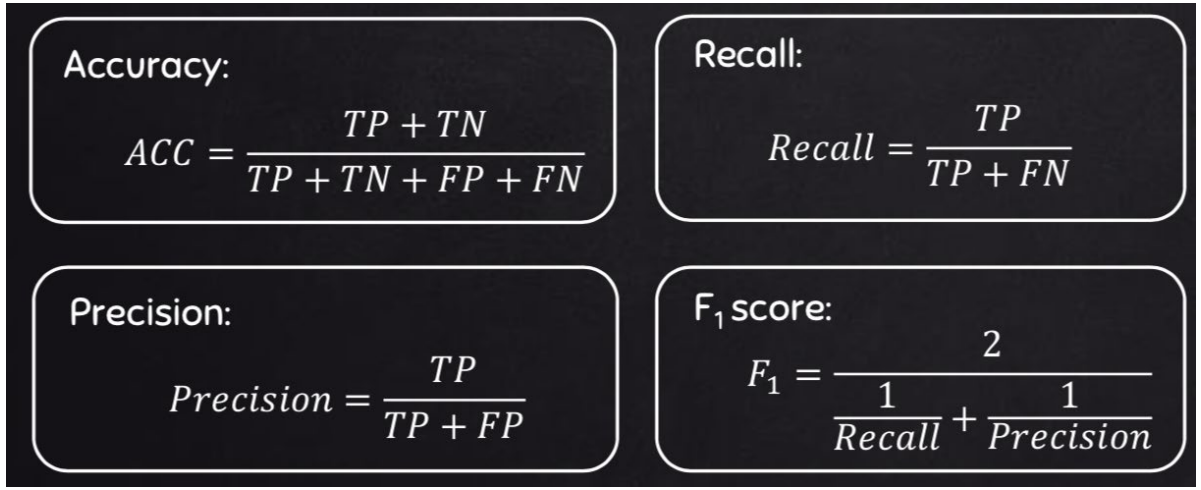
https://towardsdatascience.com/top-10-model-evaluation-metrics-for-classification-ml-models-a0a0f1d51b9

https://www.analyticsvidhya.com/blog/2020/12/decluttering-the-performance-measures-of-classification-models/

https://www.researchgate.net/publication/342009715_Estandarizacion_de_metricas_de_rendimiento_para_clasificadores_Machine_y_Deep_Learning

https://www.kdnuggets.com/2018/06/right-metric-evaluating-machine-learning-models-2.html

Evaluation metrics: Classification problems



- **The confusion matrix** is a core element that can be used to measure the performance of the ML classification model but it's not considered a metric. By nature, it is a table with two dimensions showing actual values and predicted values. Say, we need to make a classifier that diagnoses patients as sick and healthy.
- Accuracy: is used to calculate the proportion of the total number of predictions that were correct. It is the number of correct predictions divided by the total number of predictions.
- Precision: shows what proportion out of all positive predictions was correct. To calculate it, you divide the number of correct positive results (TP) by the total number of all positive results (TP + FP) predicted by the classifier.
- Recall: shows a proportion of correct positive predictions out of all positives a model could have made. To calculate it, you divide all True Positives by the sum of all True Positives and False Negatives in the dataset. In this way, recall provides an indication of missed positive predictions, unlike the precision metric we explained above.
- F1: tries to find the balance between precision and recall by calculating their harmonic mean. It is a measure of a test's accuracy where the highest possible value is 1. This indicates perfect precision and recall.

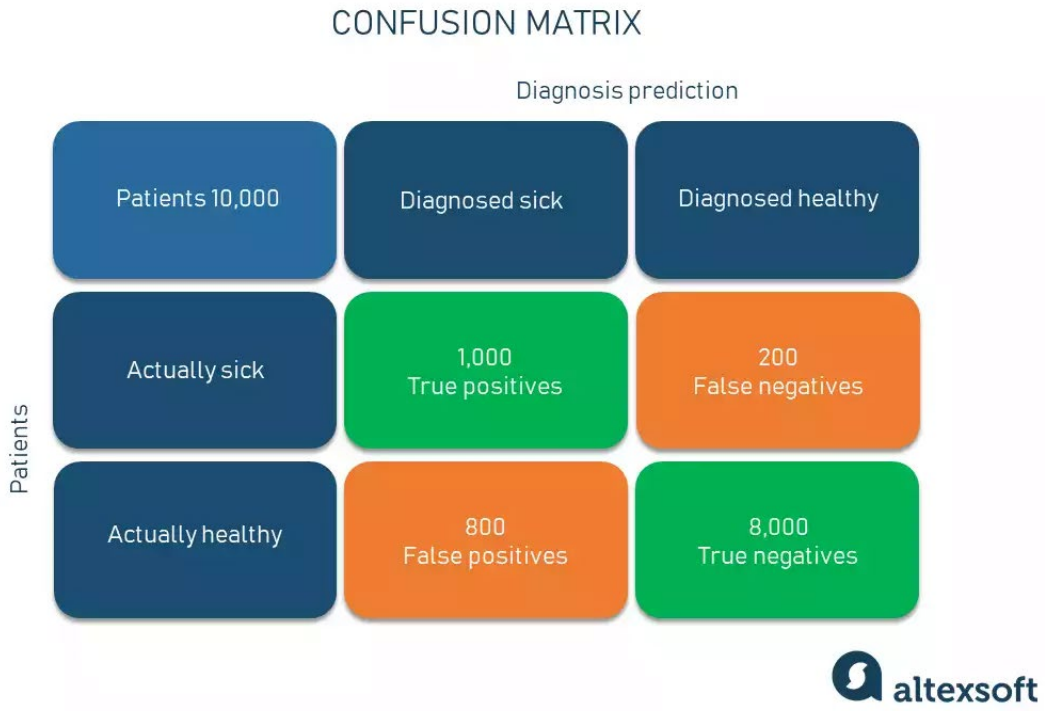
https://www.researchgate.net/publication/342009715_Estandarizacion_de_metricas_de_rendimiento_para_clasificadores_Machine_y_Deep_Learning

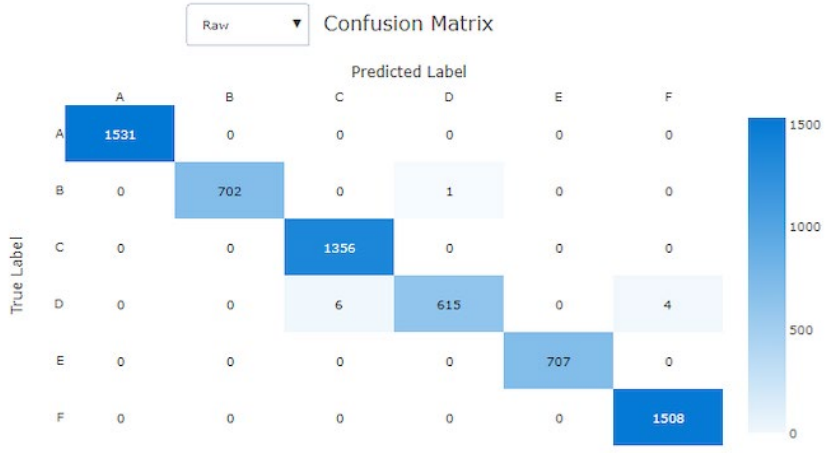
https://www.kdnuggets.com/2020/04/performance-evaluation-metrics-classification.html

https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

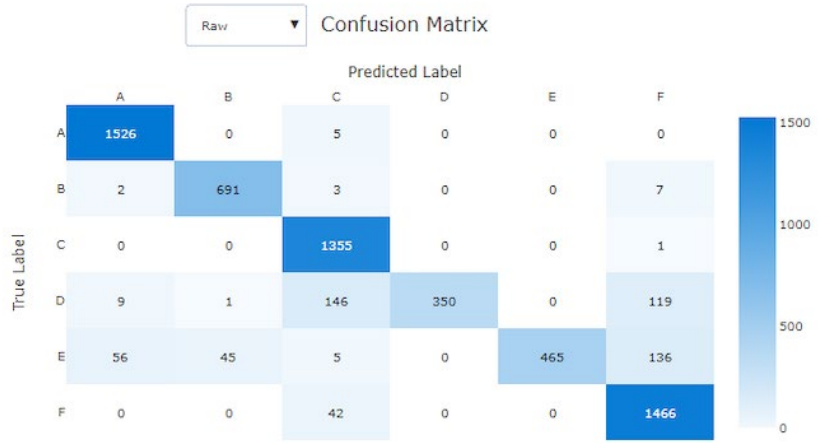
https://www.kdnuggets.com/2018/06/right-metric-evaluating-machine-learning-models-2.html

Confussion matrix: examples





Confusion matrix for a Bad model



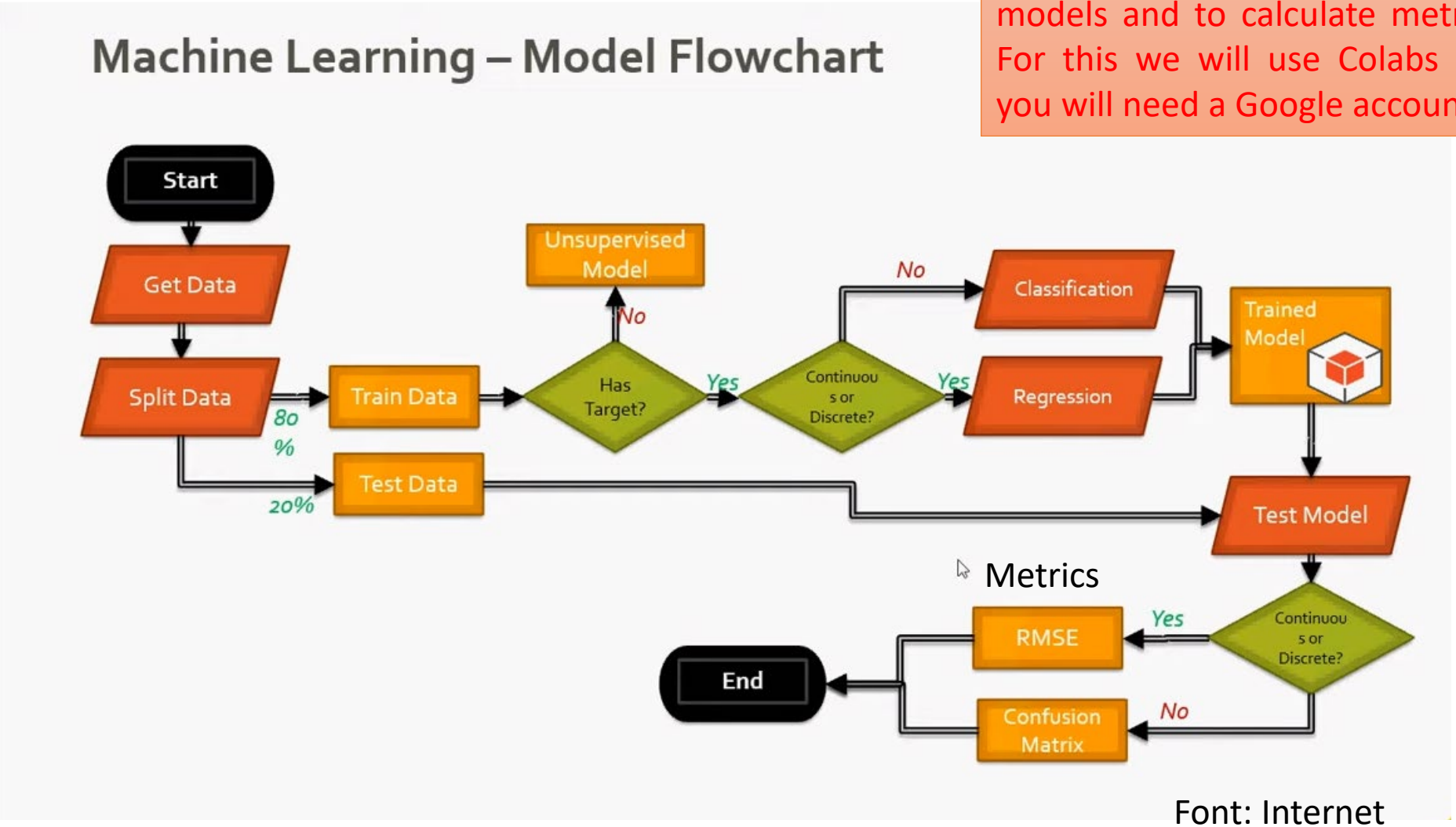
Confusion matrix for a Good model

Confussion examples with multiple classes

https://learn.microsoft.com/es-es/azure/machine-learning/how-to-understand-automated-ml

Conclusion: ML Workflow

Next day we are going to practice, to make some ML models and to calculate metrics. For this we will use Colabs and you will need a Google account



Font: Internet

Resources: Datasets

- Kaggle repository: https://www.kaggle.com/datasets
- UCI Repository: http://www.ics.uci.edu/~mlearn/MLRepository.html
- UCI KDD Archive: http://kdd.ics.uci.edu/summary.data.application.html
- Statlib: http://lib.stat.cmu.edu/
- Delve: http://www.cs.utoronto.ca/~delve/
- https://www.researchgate.net/publication/342009715_Estandarizacion_de_metricas_de_rendimiento_para_clasificadores_Machine_y_Deep_Learning
- https://www.kdnuggets.com/2020/04/performance-evaluation-metrics-classification.html
- https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

Resources: Journals

- Journal of Machine Learning Research www.jmlr.org
- Machine Learning
- Neural Computation
- Neural Networks
- IEEE Transactions on Neural Networks
- IEEE Transactions on Pattern Analysis and Machine Intelligence
- Annals of Statistics
- Journal of the American Statistical Association

Resources: Conferences

- International Conference on Machine Learning (ICML)
  - ICML05: http://icml.ais.fraunhofer.de/
- European Conference on Machine Learning (ECML)
  - ECML05: http://ecmlpkdd05.liacc.up.pt/
- Neural Information Processing Systems (NIPS)
  - NIPS05: http://nips.cc/
- Uncertainty in Artificial Intelligence (UAI)
  - UAI05: http://www.cs.toronto.edu/uai2005/
- Computational Learning Theory (COLT)
  - COLT05: http://learningtheory.org/colt2005/
- International Joint Conference on Artificial Intelligence (IJCAI)
  - IJCAI05: http://ijcai05.csd.abdn.ac.uk/
- International Conference on Neural Networks (Europe)
  - ICANN05: http://www.ibspan.waw.pl/ICANN-2005/
- ...