The Fundamentals of Machine Learning

Tim Pengajaran

Mata Kuliah Machine Learning

Jurusan Teknologi Informasi Tahun 2021

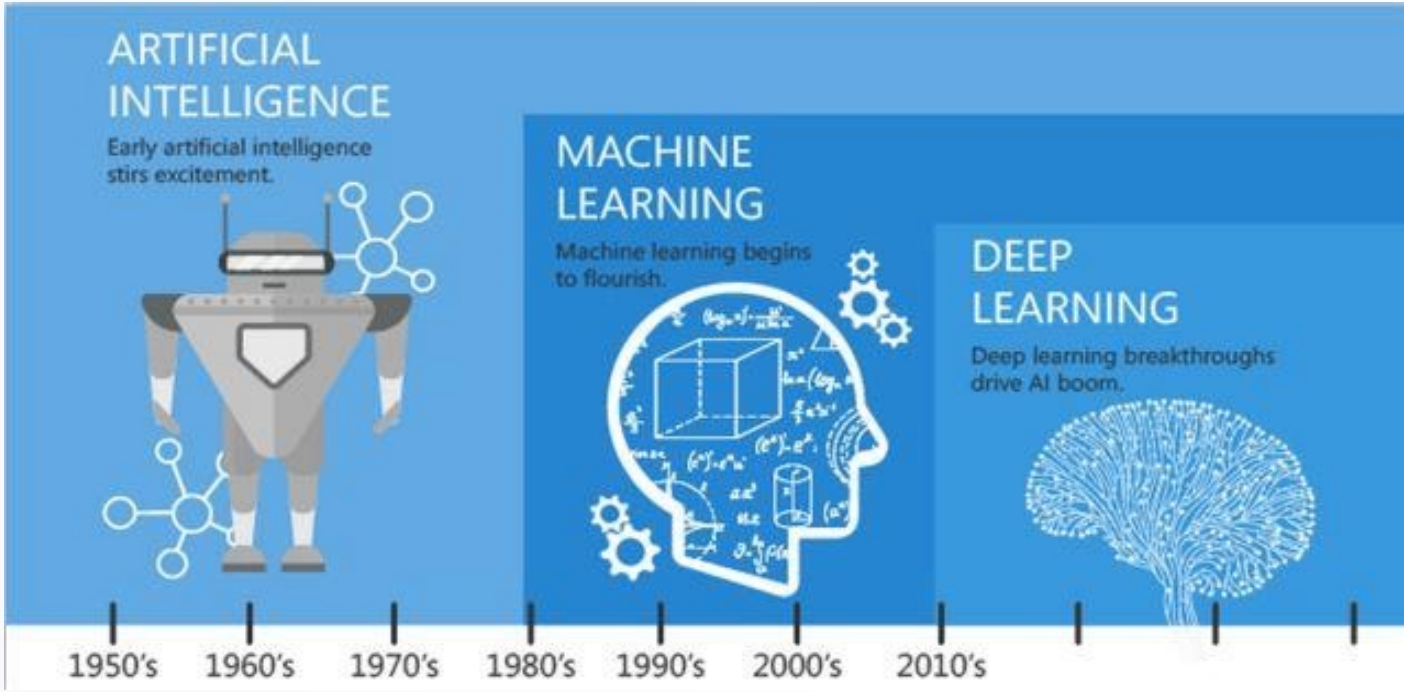
Disclaimer

- This presentation material, including examples, images, references are provided for informational and explanation assistance only
- The names of actual products and companies mentioned here in, if any, may be the trademarks of their respective owners
- Credits shall be given to the images taken from the open-source and cannot be used for promotional activities

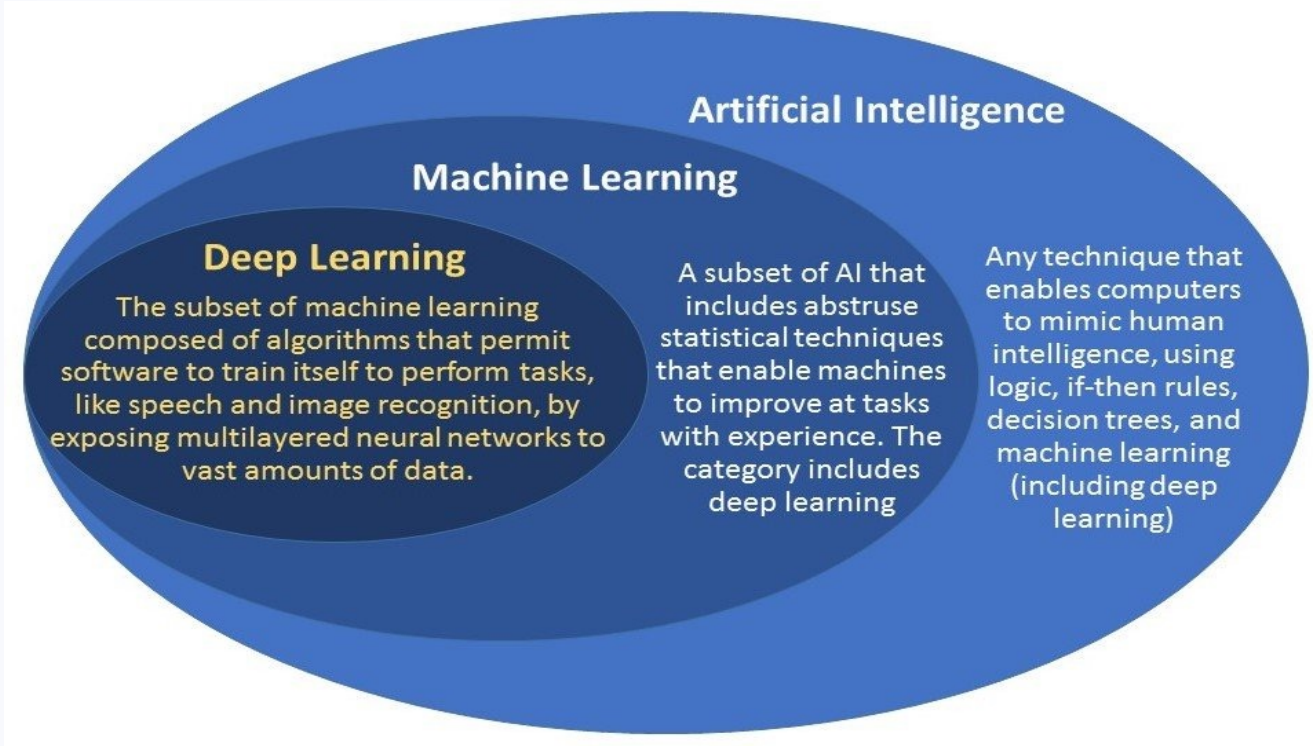
Outline

- The **Relation** between Artificial Intelligence and Machine Learning
- **What** is Machine Learning?
- **Why** Machine Learning
- **The Most Essential Concept** of Machine Learning
- **How** Machine Learning becomes Intelligent?
- **How to Measure** the Performance of Machine Learning?
- An Introduction **to Scikit-learn** and Other Tools
- A Brief on **Google Colaboratory**

The Relation between AI and ML – The Emergence



The Relation between AI and ML – The Level



The Relation between AI and ML – The Tree

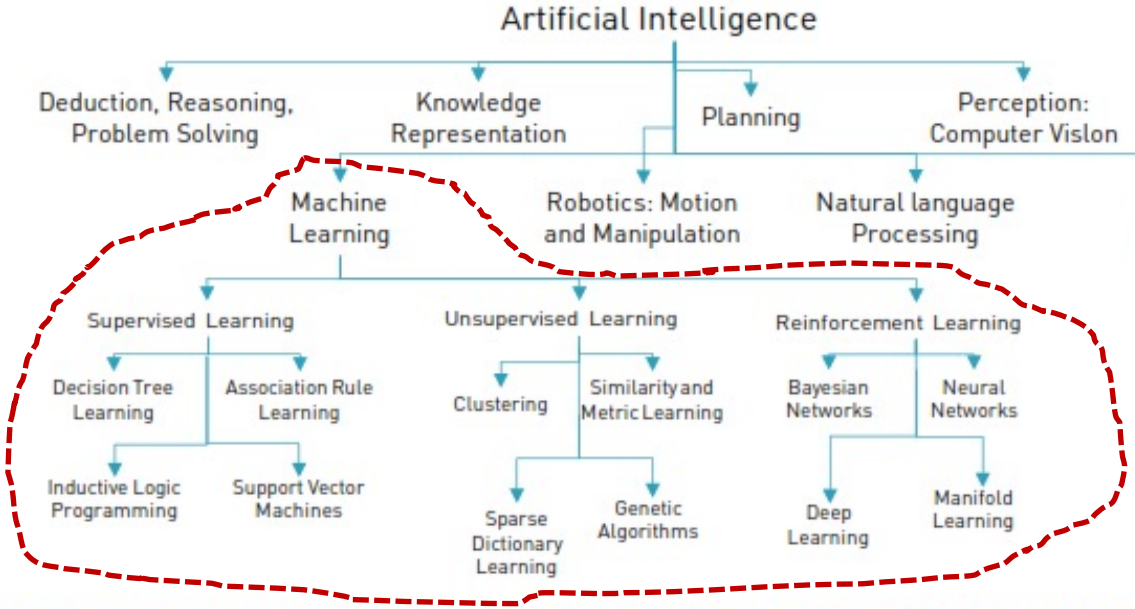
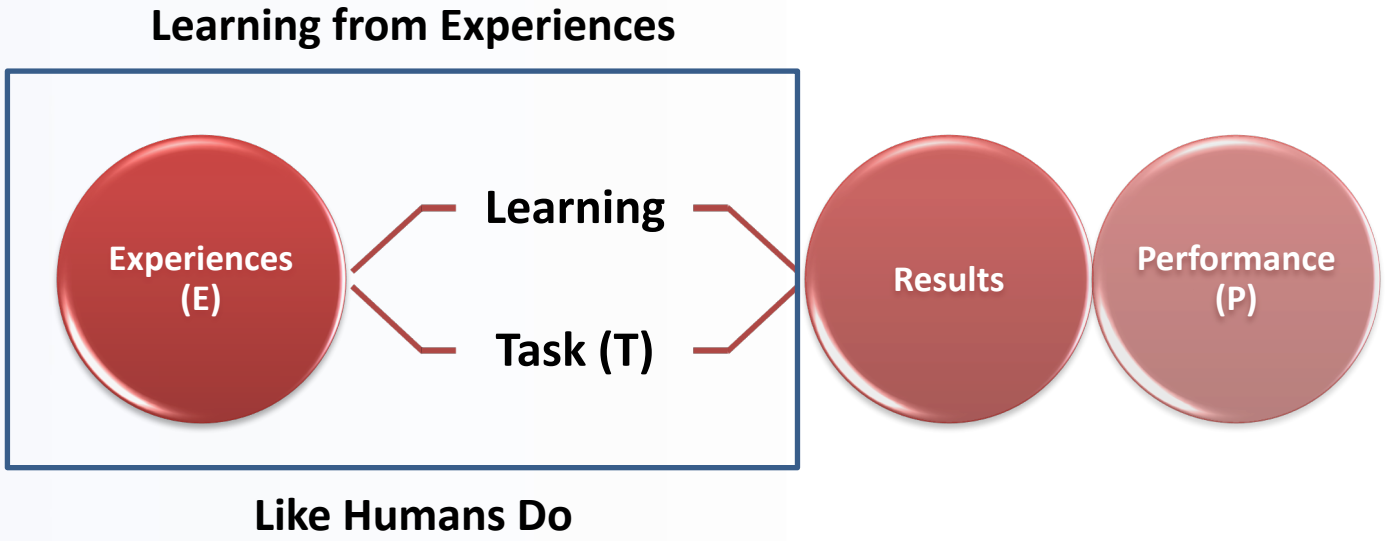


FIGURE 5: AN OVERVIEW OF NOTABLE APPROACHES AND DISCIPLINES IN AI AND MACHINE LEARNING[101]

What is Machine Learning?

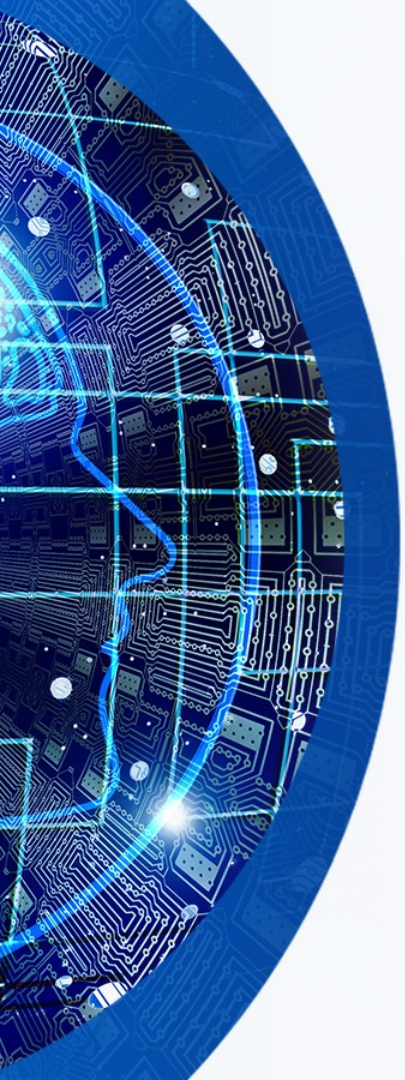
- The design and study of **software artifacts that use past experience** to inform future decisions.
- The study of programs that **learn from data**.
- A program can be said to **learn from experience 'E' with respect to some class of tasks 'T'** and **performance measure 'P'** **(Tom Mitchell's Concept)**

The Most Essential Concept of Machine Learning



How Machine Learning becomes Intelligent?

Learning develops Knowledge makes up Intelligence



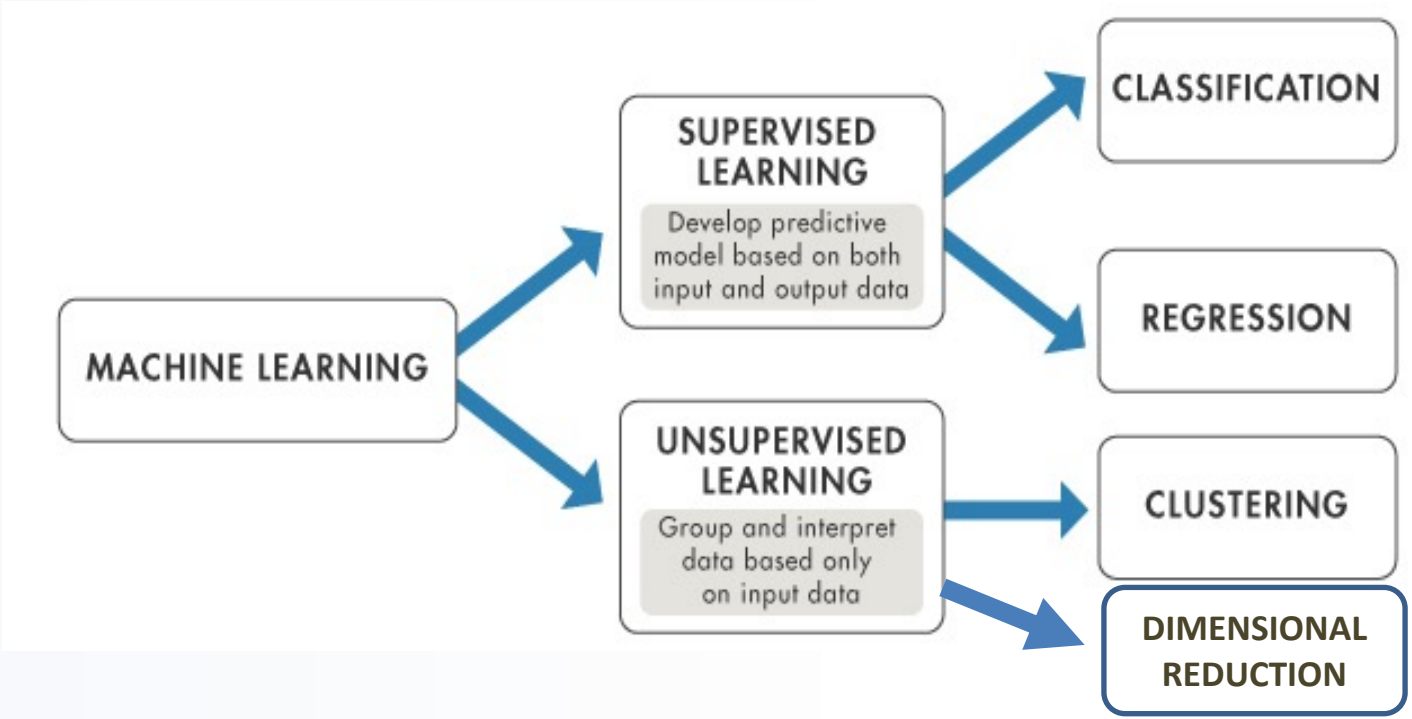
Learning in Machine Learning

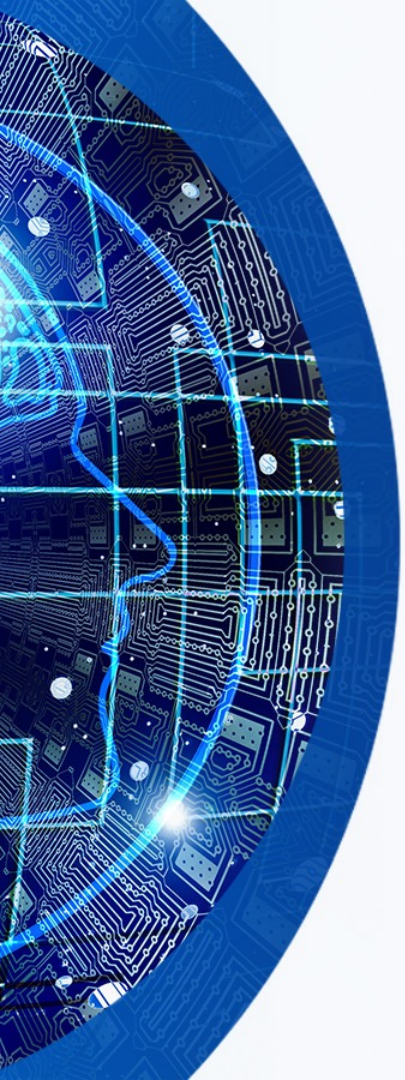
- Focus on **Supervised and Unsupervised Learning**

- **Supervised Learning**
  - A program **predicts an output for an input** by learning from pairs of labeled inputs and outputs.
  - The program **learns from examples of the "right or correct answers“**.

- **Unsupervised Learning**
  - A program **does not learn from labeled data**.
  - The program **attempts to discover patterns** in data.
  - It **cannot give label** to the inputs.

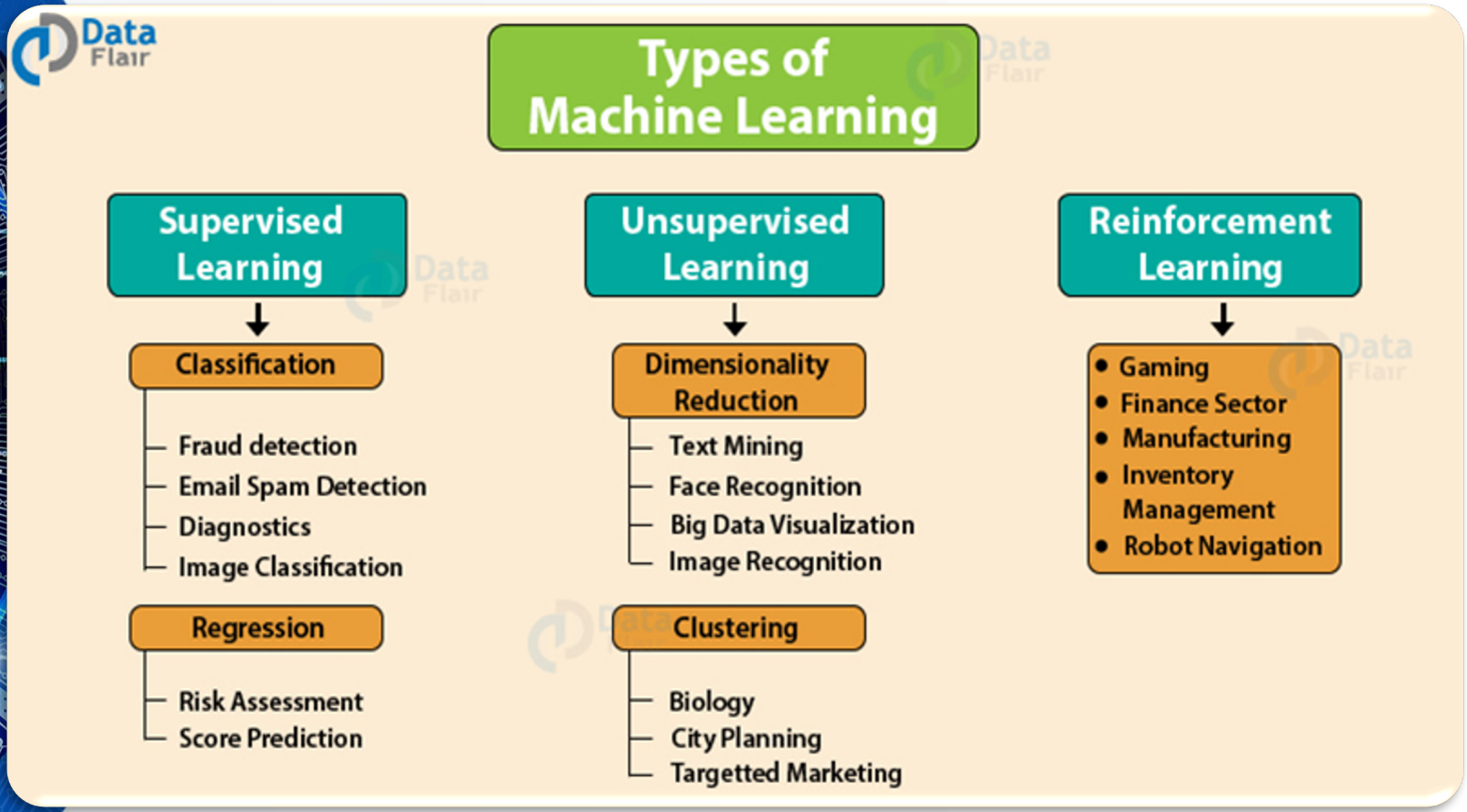
Machine Learning Tasks – Generic



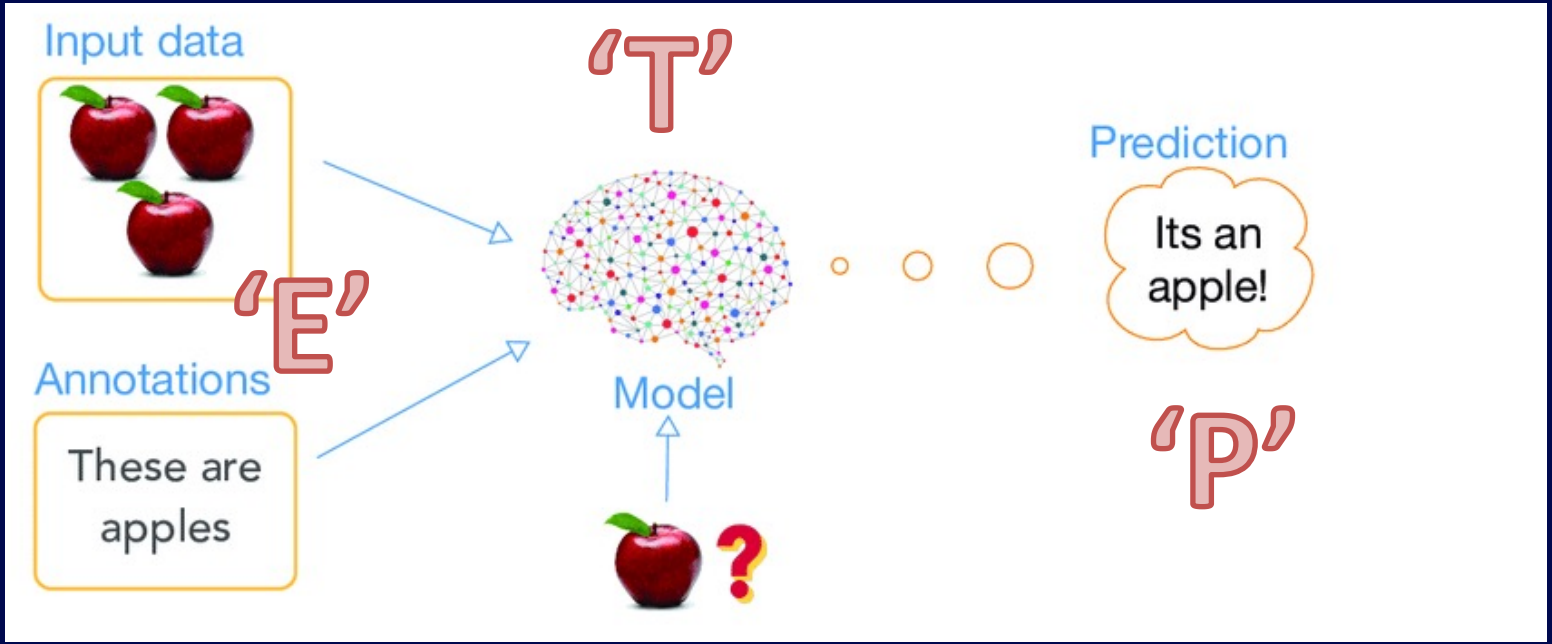


Machine Learning Tasks

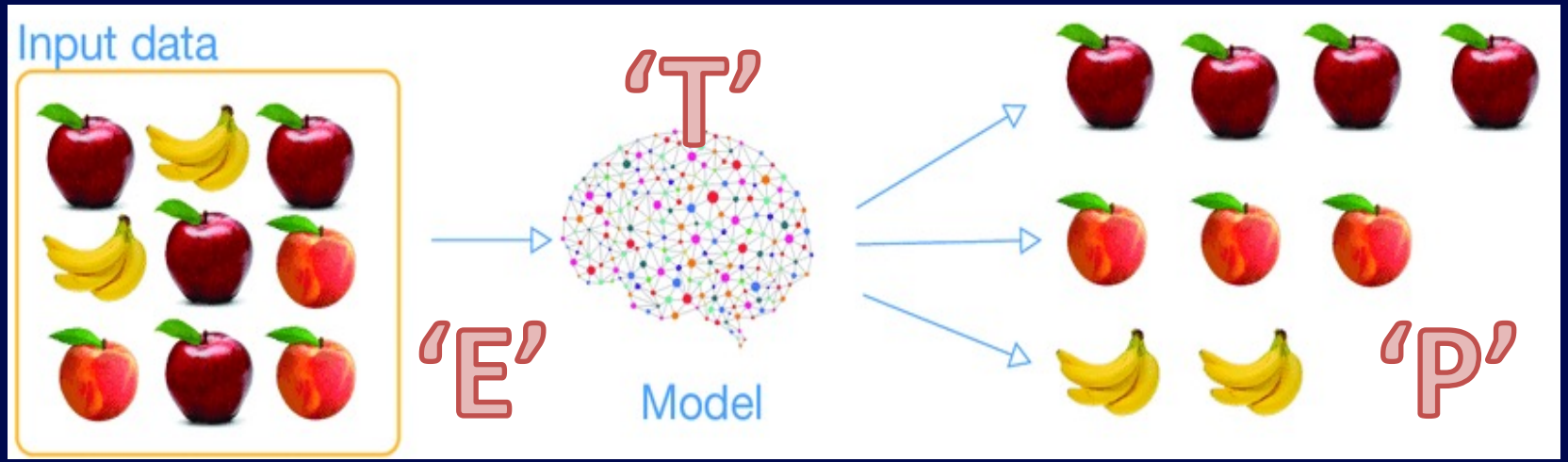
- **Dimensional Reduction**
  - The process of **discovering the features** that account for the greatest changes in the response variable.
  - Can also be used to **visualize data**.
  - **Unsupervised** approach.

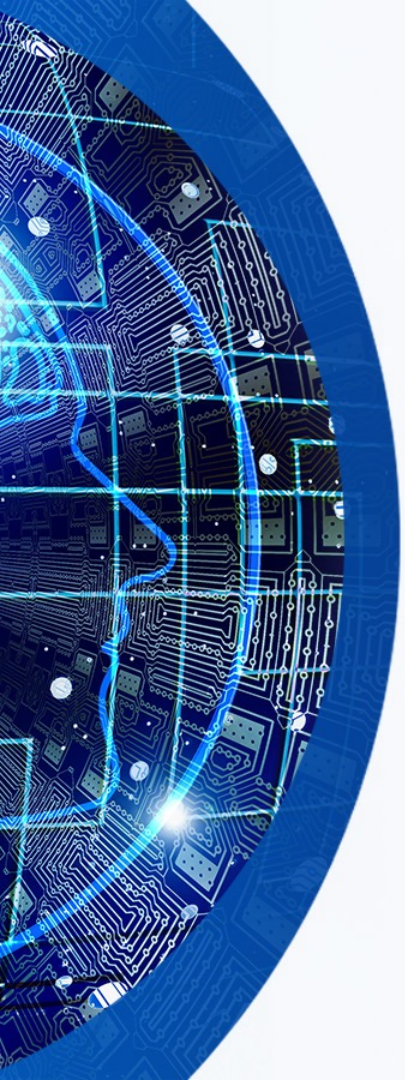

Data Flair
Types of Machine Learning
Supervised Learning
Classification
Fraud detection
Email Spam Detection
Diagnostics
Image Classification
Regression
Risk Assessment
Score Prediction
Unsupervised Learning
Dimensionality Reduction
Text Mining
Face Recognition
Big Data Visualization
Image Recognition
Clustering
Biology
City Planning
Targetted Marketing
Reinforcement Learning
Gaming
Finance Sector
Manufacturing
Inventory Management
Robot Navigation

The Concept of Supervised Learning



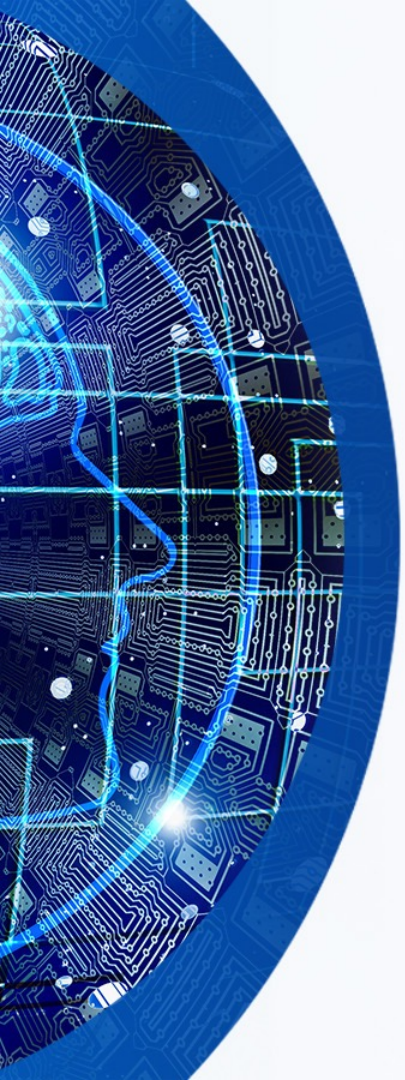
The Concept of Unsupervised Learning





Training Data, Testing Data, and Validation Data

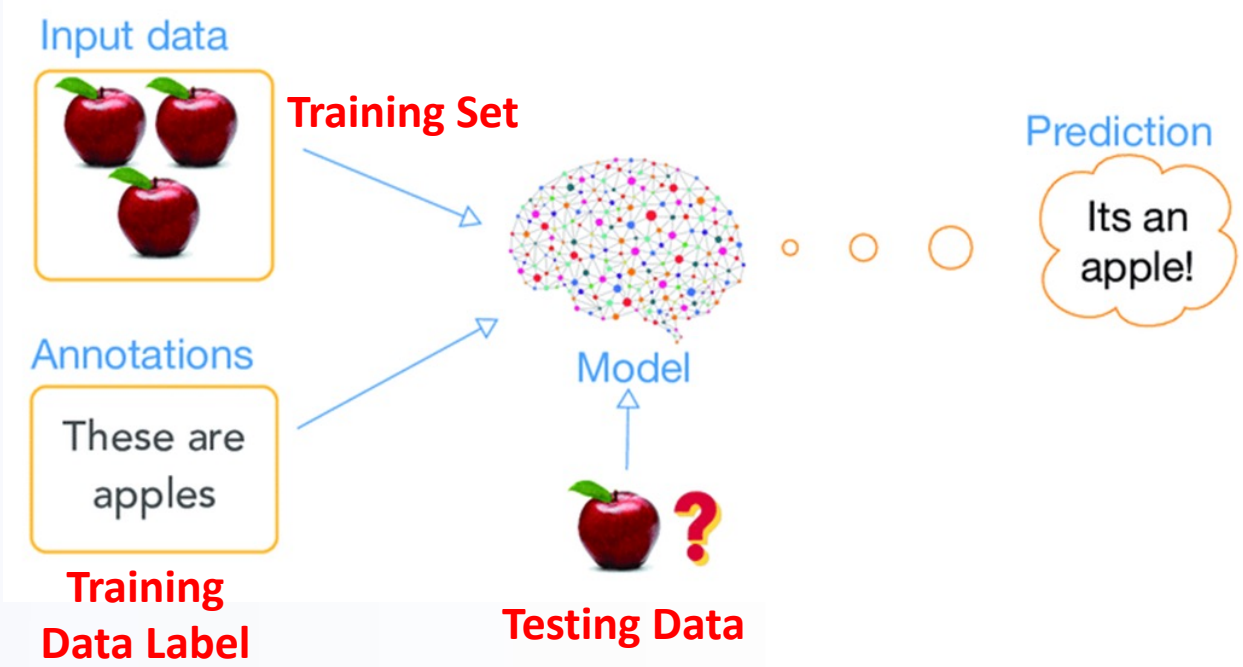
- **Training Data**
  - A training set is **a collection of observations**.
- **Testing Data**
  - The test set is **a similar collection of observations**.
  - **Used to evaluate the performance of the model** using some performance metric.
  - **No observations from the training set** are included in the test set.
- **Validation Data or Hold-out Set**
  - **Used to tune variables called hyperparameters** that control how the algorithm learns from the training data, or **the properties** that govern the entire training process.
  - **Still evaluated on the test set** to provide an estimate of its performance in the real world.



Training Data, Testing Data, and Validation Data – The Rule

- **It is common to partition a single set** of supervised observations into training, validation, and test sets.
- **No requirements for the sizes** of the partitions.
- **Vary according to the amount of data** available.
- It is common to allocate between **fifty and seventy-five percent of the data to the training set**.
- **Ten to twenty-five percent of the data to the test set**.
- The **remainder** to the **validation set**.

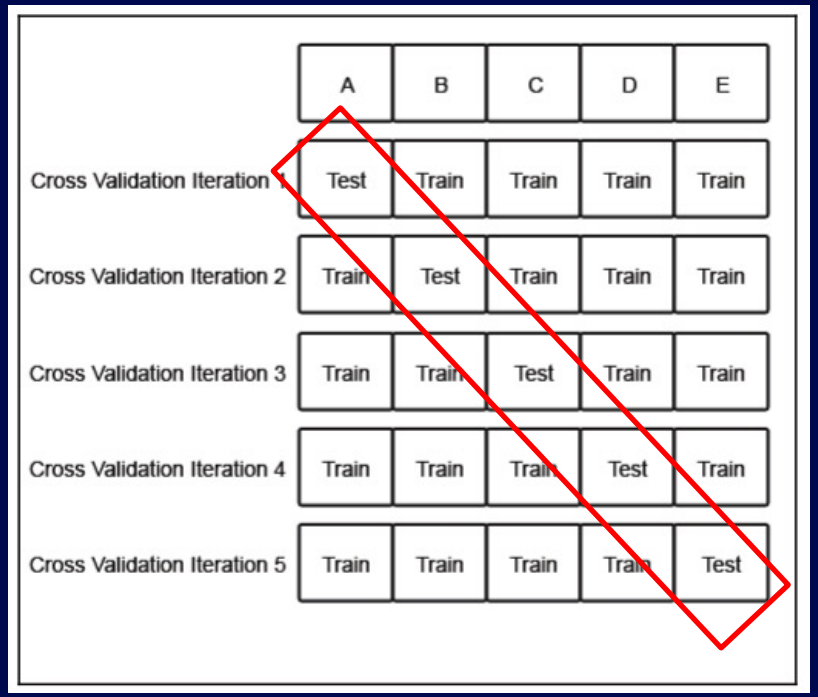
Training Data and Testing Data



Validating Scarce Data

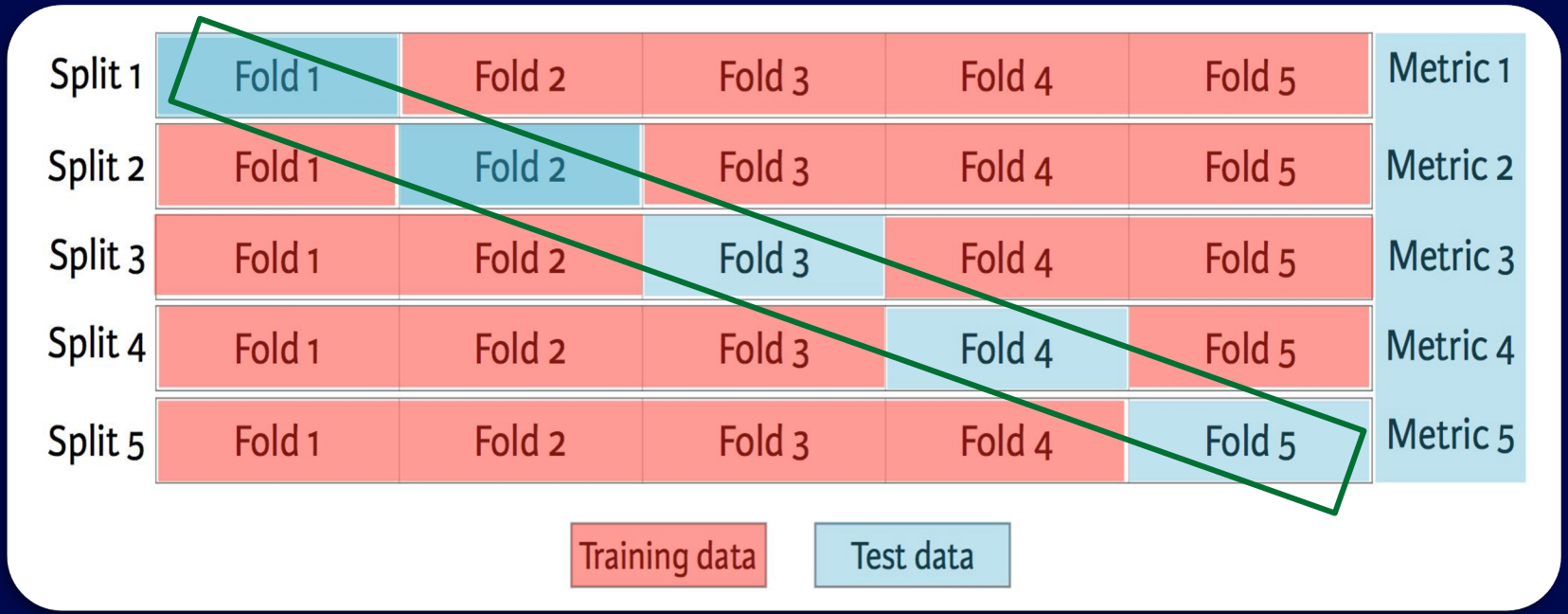
- **Cross-validation technique**
  - Can be used **to train and validate a model** on the same data.
  - The training data is **partitioned**.
  - **The model is trained using all but one of the partitions, and tested on the remaining partition**.
  - The partitions are then **rotated several times** so that the model is trained and evaluated on all of the data.
  - The **mean of the model's scores** on each of the partitions is a better estimate of performance in the real world than an evaluation using a single training/testing split.

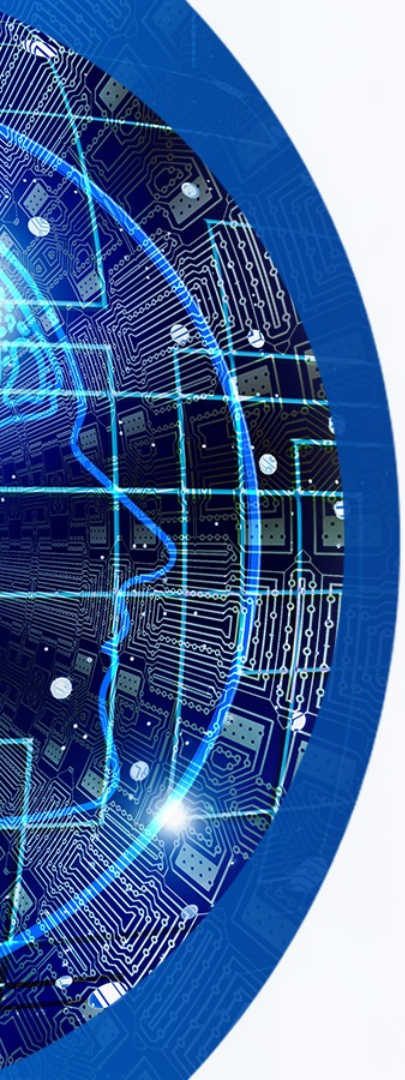
Operating Cross-Validation



- The original dataset is **partitioned into five subsets of equal size** labeled A through E.
- Initially the model is **trained on partitions** B through E, and tested on partition A.
- In the next iteration, the model is trained on partitions A, C, D, and E, and tested on partition B.
- **The partitions are rotated** until models have been trained and tested on all of the partitions.
- Cross-validation provides **a more accurate estimate** of the model's performance than testing a single partition of the data.

Operating Cross-Validation

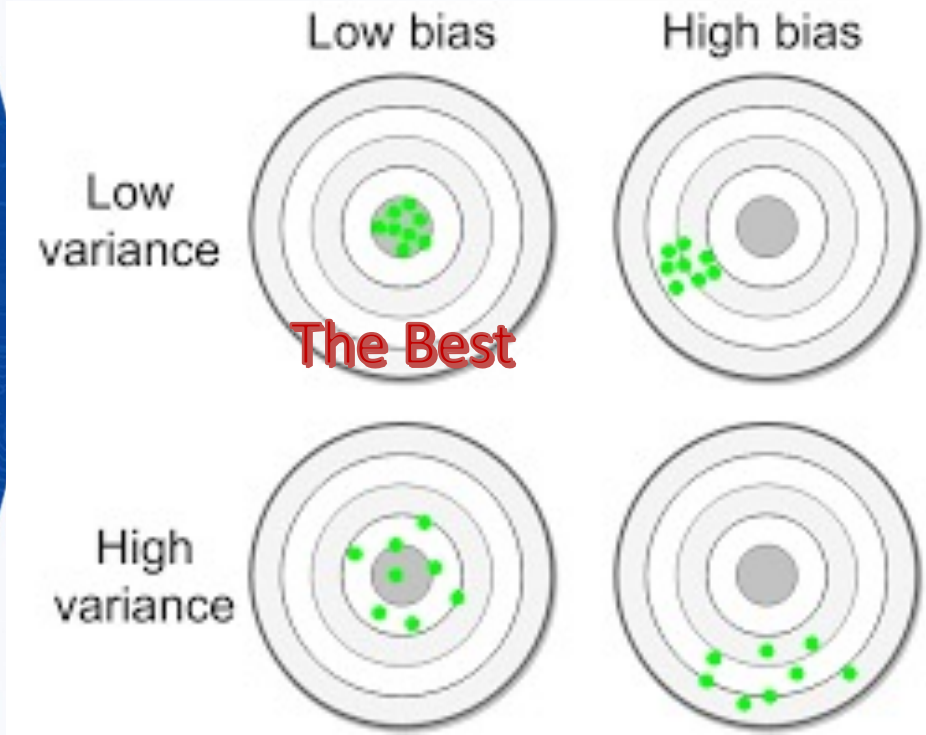




How to Measure the Performance of Machine Learning?

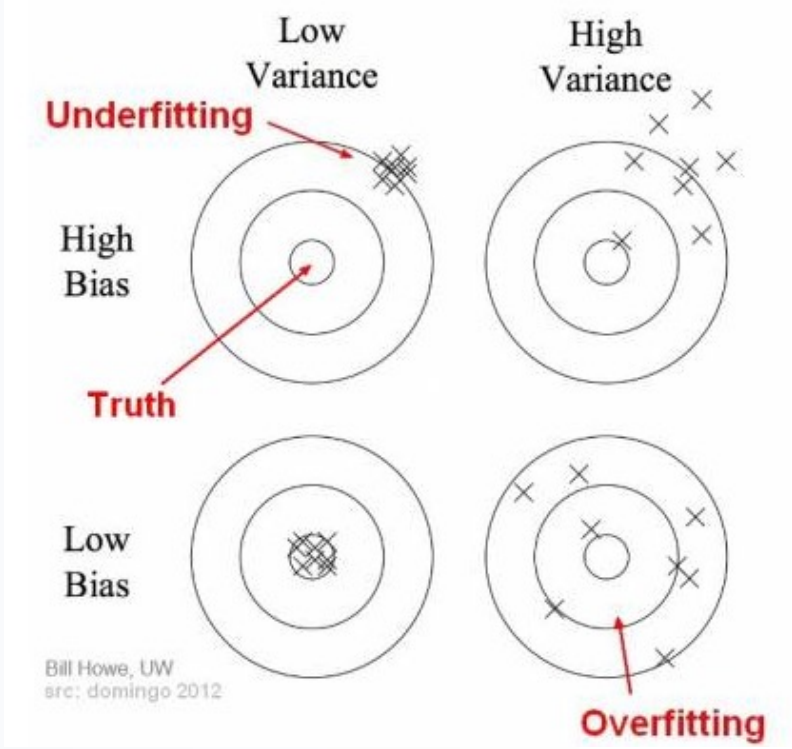
- Performance metrics measure **the amount of prediction error.**
- **Bias**
  - An error from **erroneous assumptions** in the learning algorithm.
- **Variance**
  - A type of error that occurs due to **a model's sensitivity** to small fluctuations in the training set.

Types of Bias and Variance

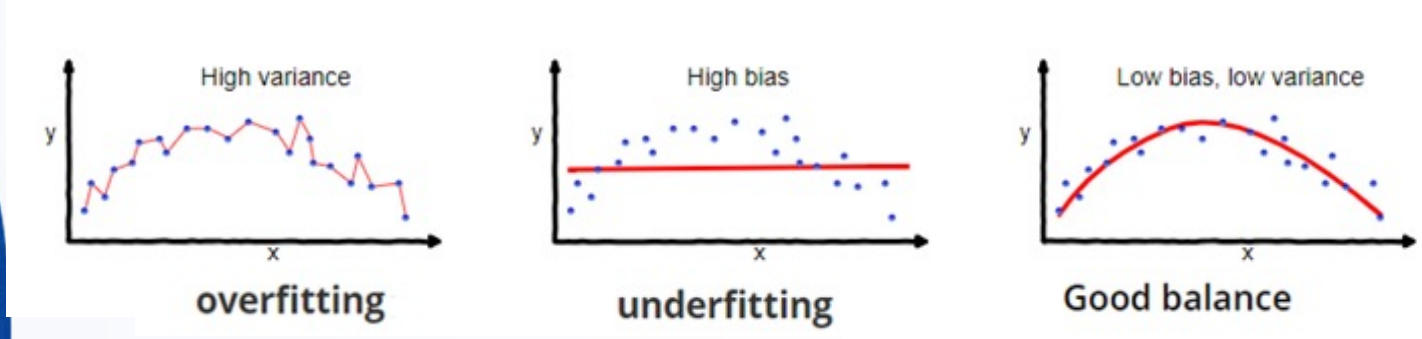


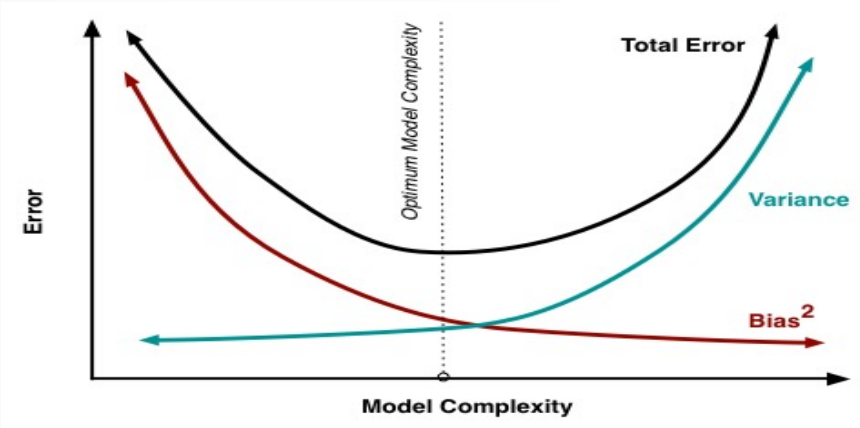
- A model with **high variance over-fits** the training data.
- A model with **high bias under-fits** the training data.
- So, find **the most proper technique** to reduce bias and variance (**trade-off**).

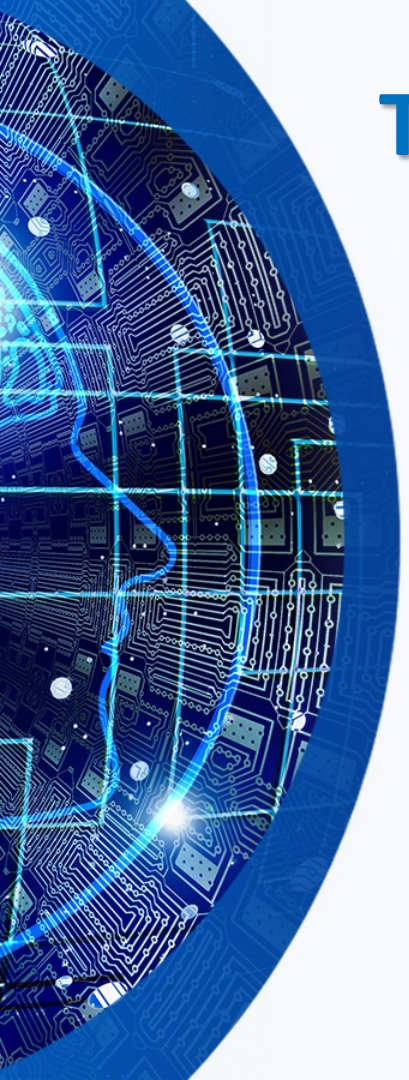
More View on Bias and Variance



Over, Underfitting and the Trade-Off





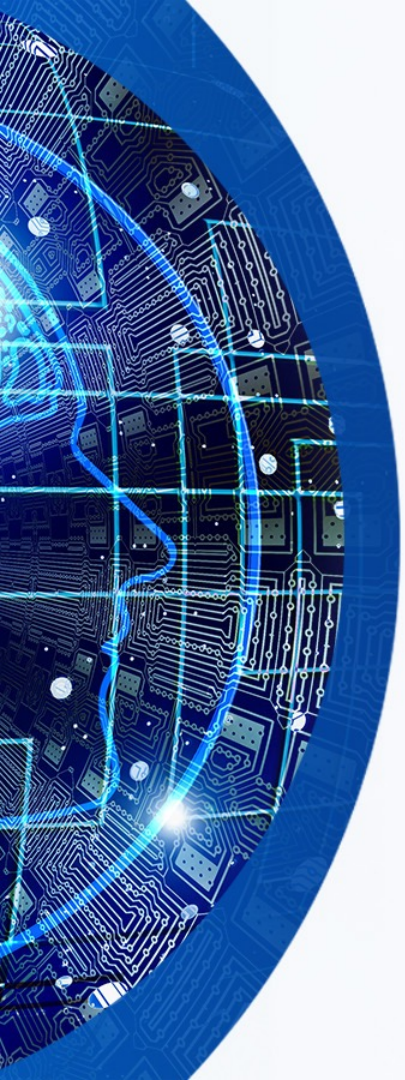


Techniques to Reduce High Variance

1. **Get more training examples** because a larger the dataset is more probable to get a higher predictions.

2. **Try smaller sets of features** because you are overfitting.

3. **Try increasing lambda**, so you can not overfit the training set as much. The higher the lambda, the more the regularization applies, for Linear Regression with regularization.



Techniques to Reduce High Bias

1. **Try getting additional features**, you are generalizing the datasets.

2. **Try adding polynomial features**, make the model more complicated.

3. **Try decreasing lambda**, so you can try to fit the data better. The lower the lambda, the less the regularization applies, for Linear Regression with regularization.

The Accuracy of Prediction

- **The number of correctly predicted data points** out of all the data points.
- **Called as the fraction of instances** that were classified correctly.
- An **intuitive measure** of the program's performance.
- **Does measure the program's performance**, but it does not differentiate.
- Often **used along with Precision (P) and Recall (R)**, which are other metrics that use various ratios of **true/false positives/negatives,** namely:
  - **TP = True Positive**
  - **TN = True Negative**
  - **FP = False Positive**
  - **FN = False Negative**

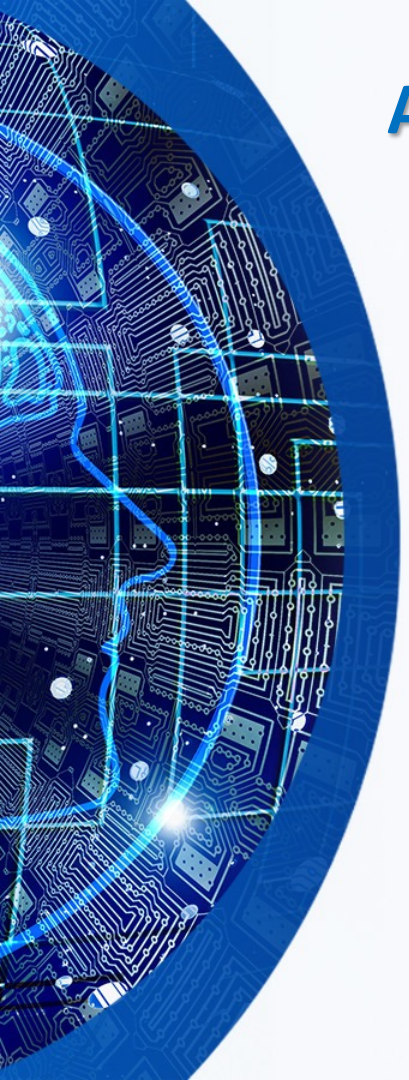
The Accuracy of Prediction

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \qquad P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN}$$

- Can be calculated by the help of **Confusion Matrix**.
  - ✓ **True Positive (TP)** – The case in which the system predicted YES and the actual output was also YES.
  - ✓ **True Negative (TN)** – The case in which the system predicted NO and the actual output was NO.
  - ✓ **False Positive (FP)** – The case in which the system predicted YES and the actual output was NO.
  - ✓ **False Negative (FN)** – The case in which the system predicted NO and the actual output was YES.

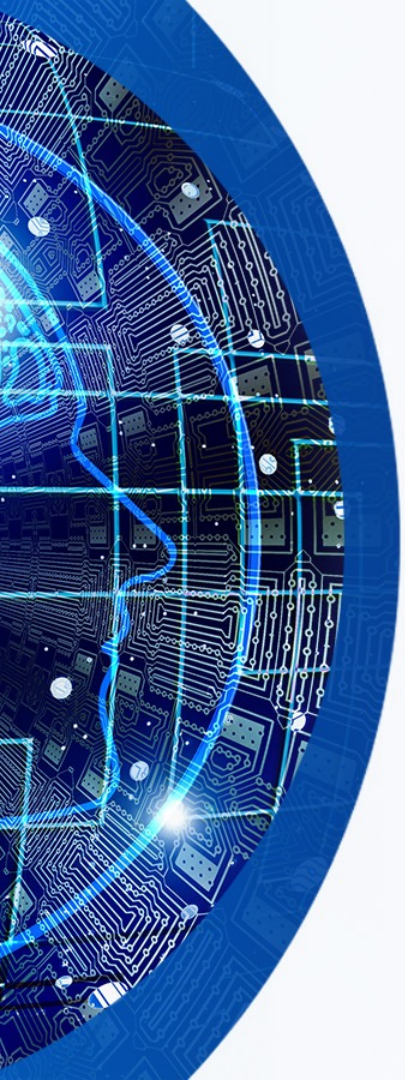
n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

Try to calculate this table



An Introduction to Scikit-learn and Other Tools

- **Scikit-learn**
  - **Focused on Machine Learning.**
  - Built on the popular Python libraries **NumPy and SciPy**.
  - **A library in Python** that provides many unsupervised and supervised **learning** algorithms.
  - Provides a range of supervised and unsupervised learning algorithms via **a consistent interface** in Python.
  - A **free machine learning library** for Python.

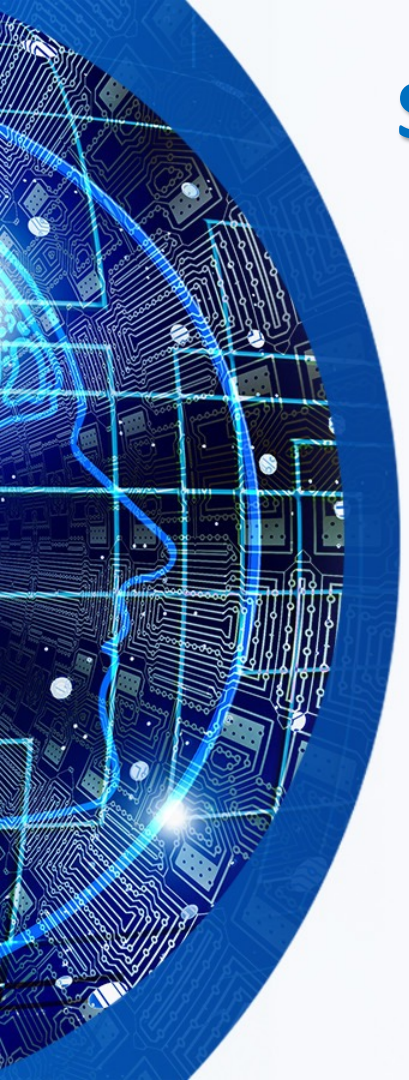


Scikit-learn Supports Various Libraries

- **NumPy**: For any work with matrices, especially **math operations.**
- **SciPy**: Scientific and technical **computing**.
- **Matplotlib**: Data **visualisation**.
- **IPython**: **Interactive console** for Python.
- **Sympy**: **Symbolic** mathematics.
- **Pandas**: **Data** handling, manipulation, and analysis.

Scikit-learn Supports Various Algorithms

- **Regression:** **Fitting** linear and non-linear models
- **Clustering:** Unsupervised **classification**.
- **Decision Trees:** Tree **induction** and pruning for both **classification** and **regression** tasks.
- **Neural Networks:** End-to-end training for both **classification and regression**. Layers can be easily defined in a tuple.
- **SVMs:** for learning **decision boundaries**.
- **Naive Bayes**: Direct **probabilistic modelling**.



Scikit-learn Supports Various Algorithms

- **Ensemble Methods:** Boosting, Bagging, Random Forest, Model voting and averaging.
- **Feature Manipulation**: Dimensionality reduction, feature selection, feature analysis.
- **Outlier Detection:** For detecting outliers and rejecting noise.
- **Model selection and validation:** Cross-validation, Hyperparameter tuning, and metrics.

More on Scikit-learn

- Machine Learning in Python - https://scikit-learn.org/stable/

- An introduction to machine learning with scikit-learn - https://scikit-learn.org/stable/tutorial/basic/tutorial.html#

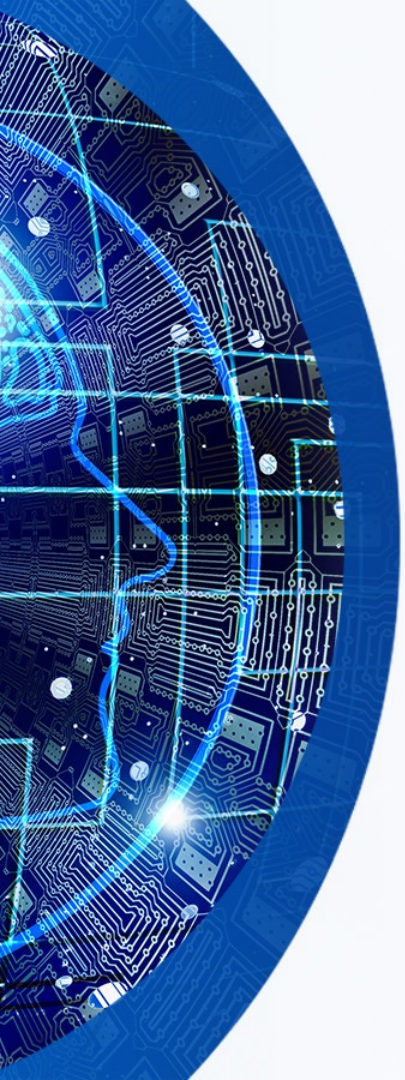
Platform Scikit-learn

- Windows
- Ubuntu Linux
- Mac OS

Using

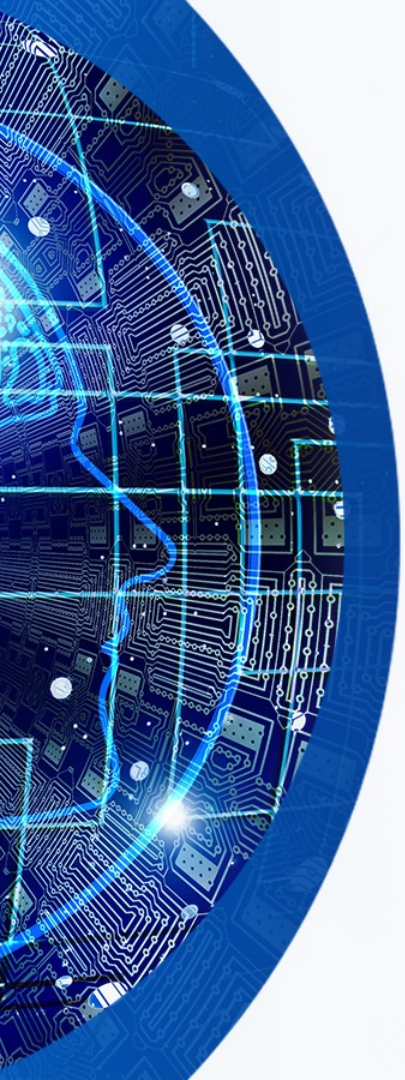
- Anaconda
- Pip
- Google Colab

See the guide in https://scikit-learn.org/stable/install.html#install-official-release



A Brief on Google Colab

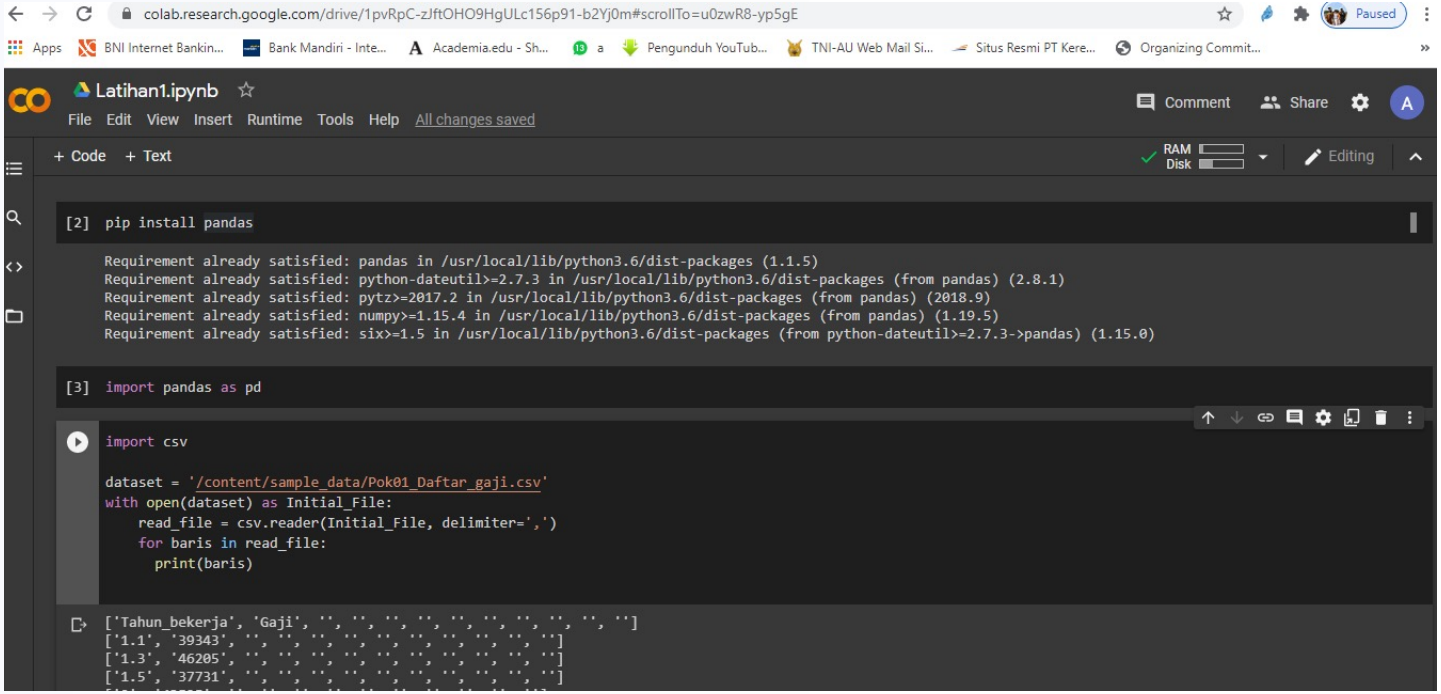
- **A free online cloud-based Jupyter notebook environment** that allows us to train our machine learning and deep learning models on CPUs, GPUs, and TPUs.
  - CPU – Central Processing Unit.
  - GPU – Graphic Processing Unit.
  - TPU – Tensor Processing Unit.



Installing and Practicing Google Colab

- Takes only 5 minutes.
- Follow online instruction such as in:
  - https://medium.com/@dede.brahma2/cara-menggunakan-google-colaboratory-5f5e4393ac2f
- Learn to code Python. There are abundant resources such as:
  - https://colab.research.google.com/github/cs231n/cs231n.github.io/blob/master/python-colab.ipynb#scrollTo=DL5sMSZ9L9eq

An Example of Python Coding in Colab



Thank You for Today
Keep Spirit!