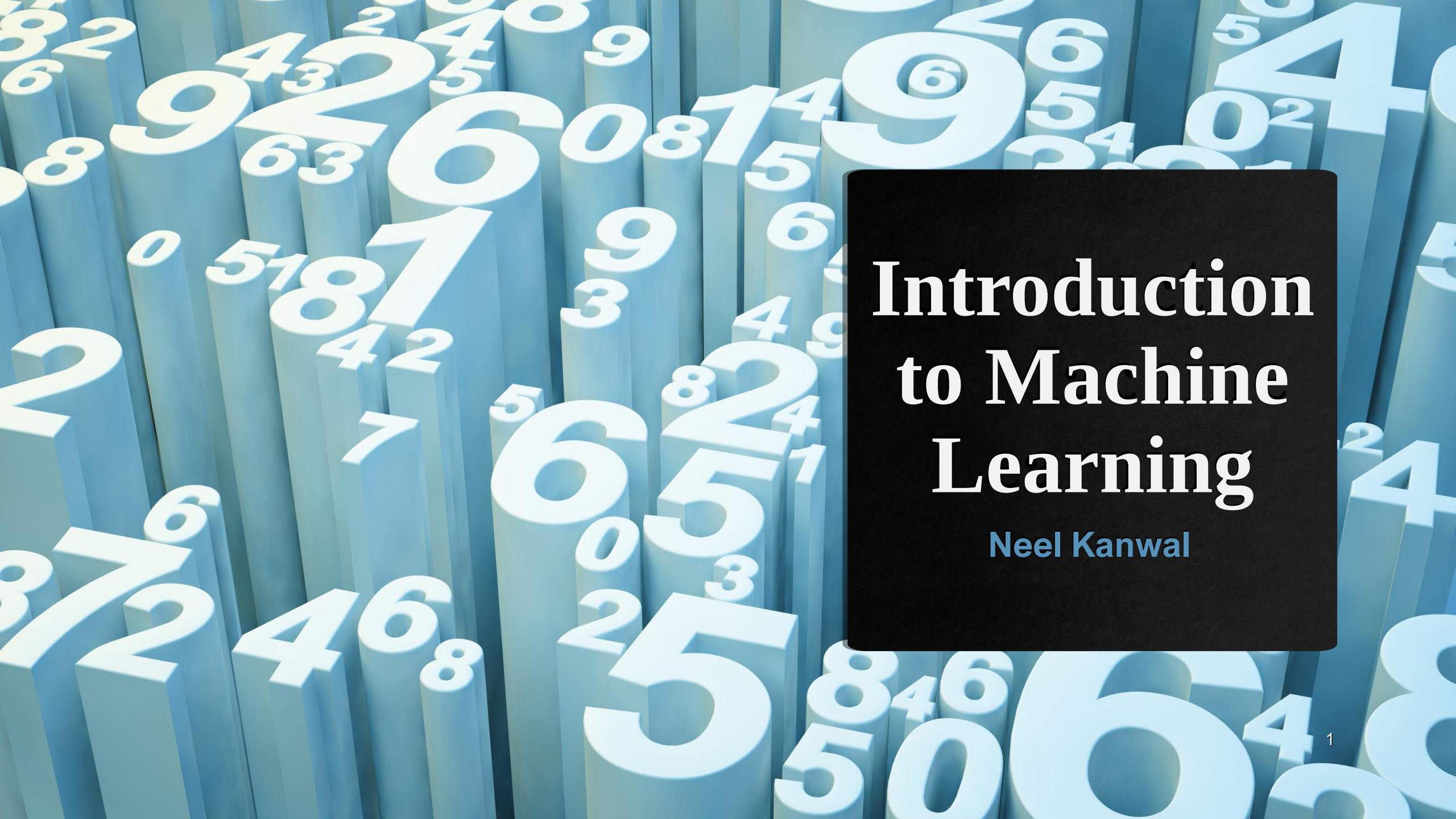


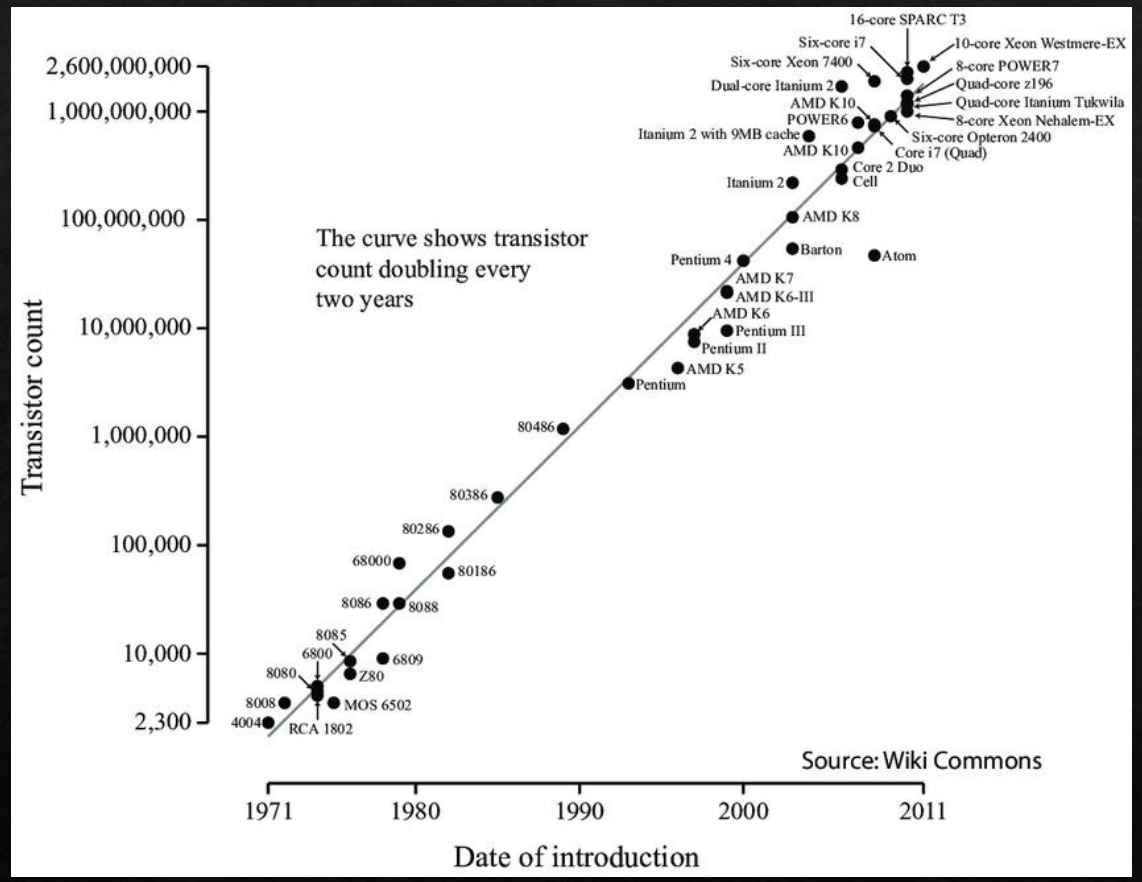
Introduction to Machine Learning

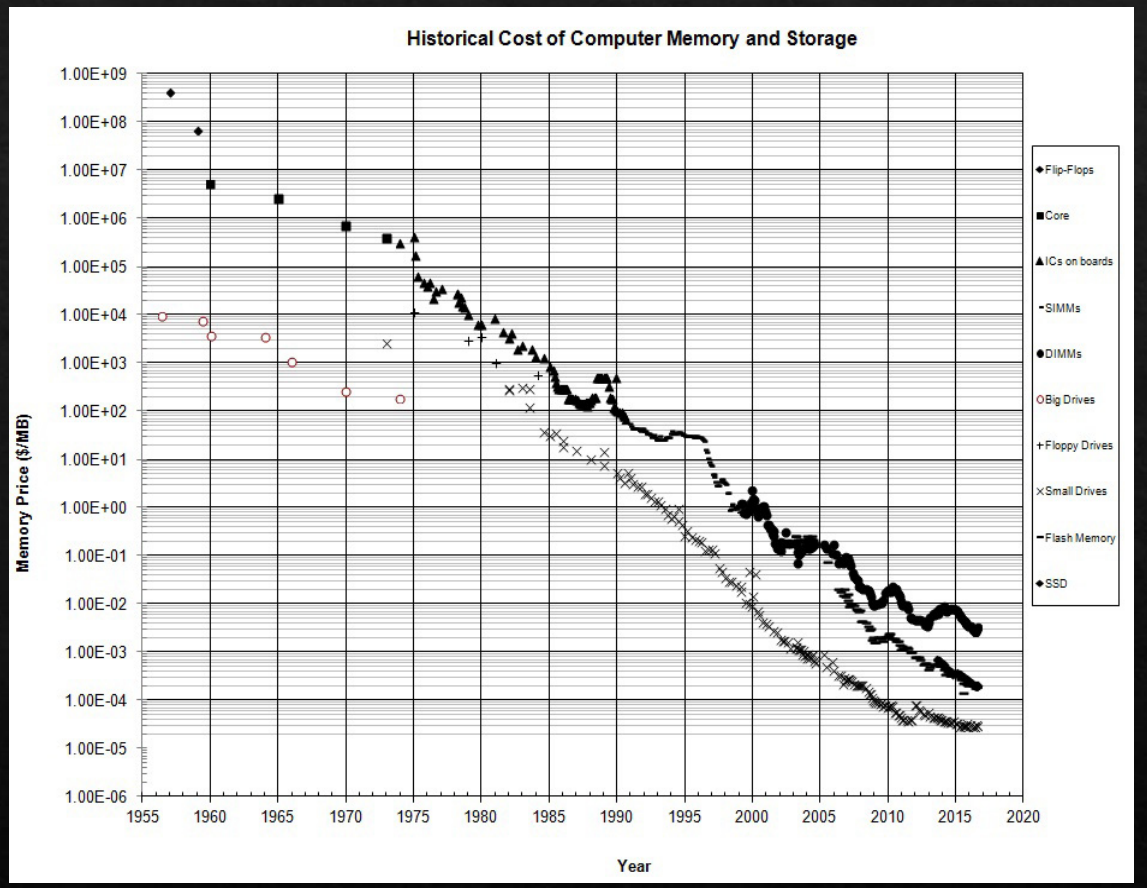
Neel Kanwal

Trend of digitalization

- Amount of data every day is growing at a fast pace.
- Global digital transformation market to grow to $1009.8 billion by 2025 from $469.8 in 2020, CAGR of 16.5%.
- Availability of data makes it easier to analyze trends/patterns.
- Big data makes it hard to analyze manually, extremely time-consuming for some tasks.
- Requires computational and storage resources.

Extended Computation and Storage



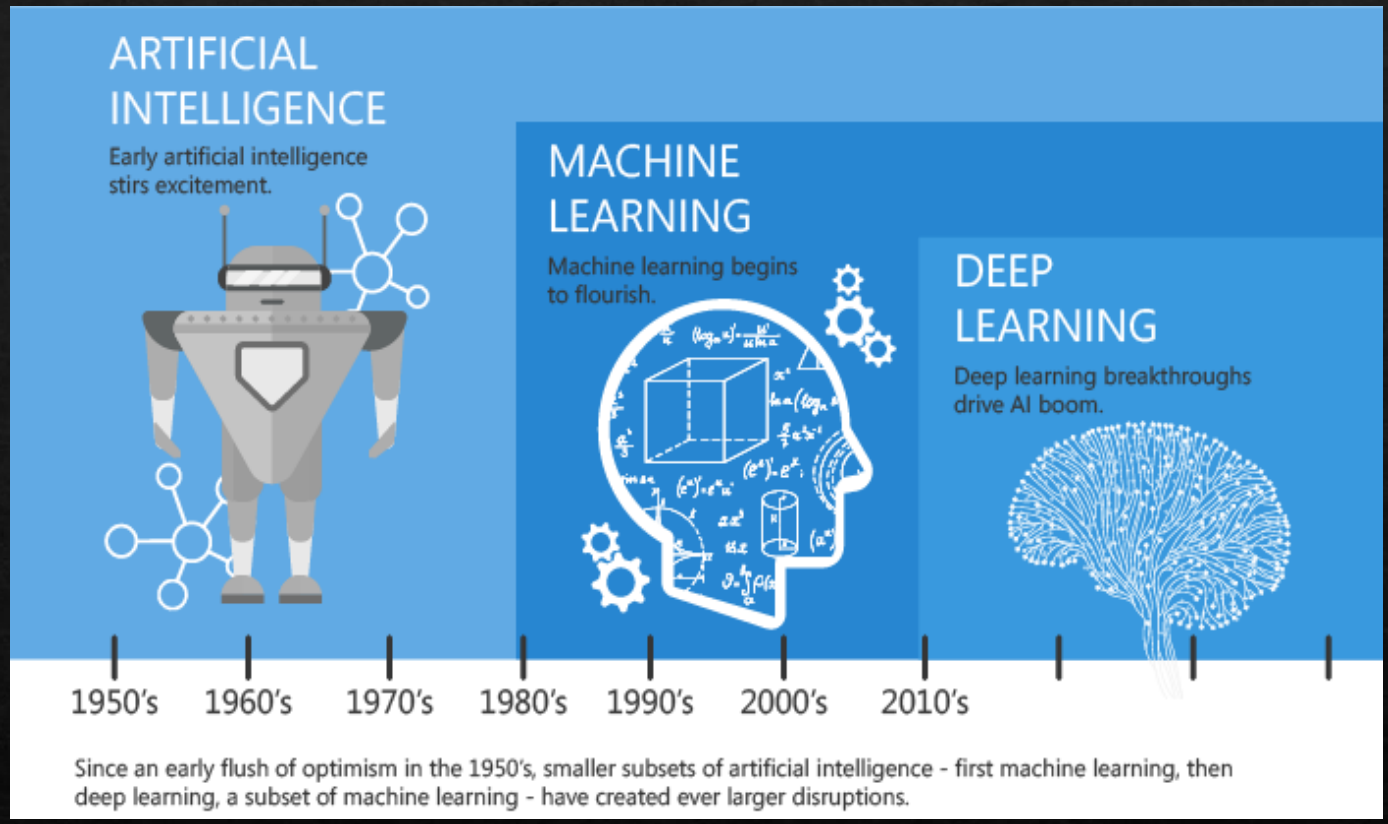


•DOI: 10.2298/FUEE0403285S

https://arstechnica.com/gadgets/2016/11/how-cheap-ram-changes-computing/

What is artificial intelligence?

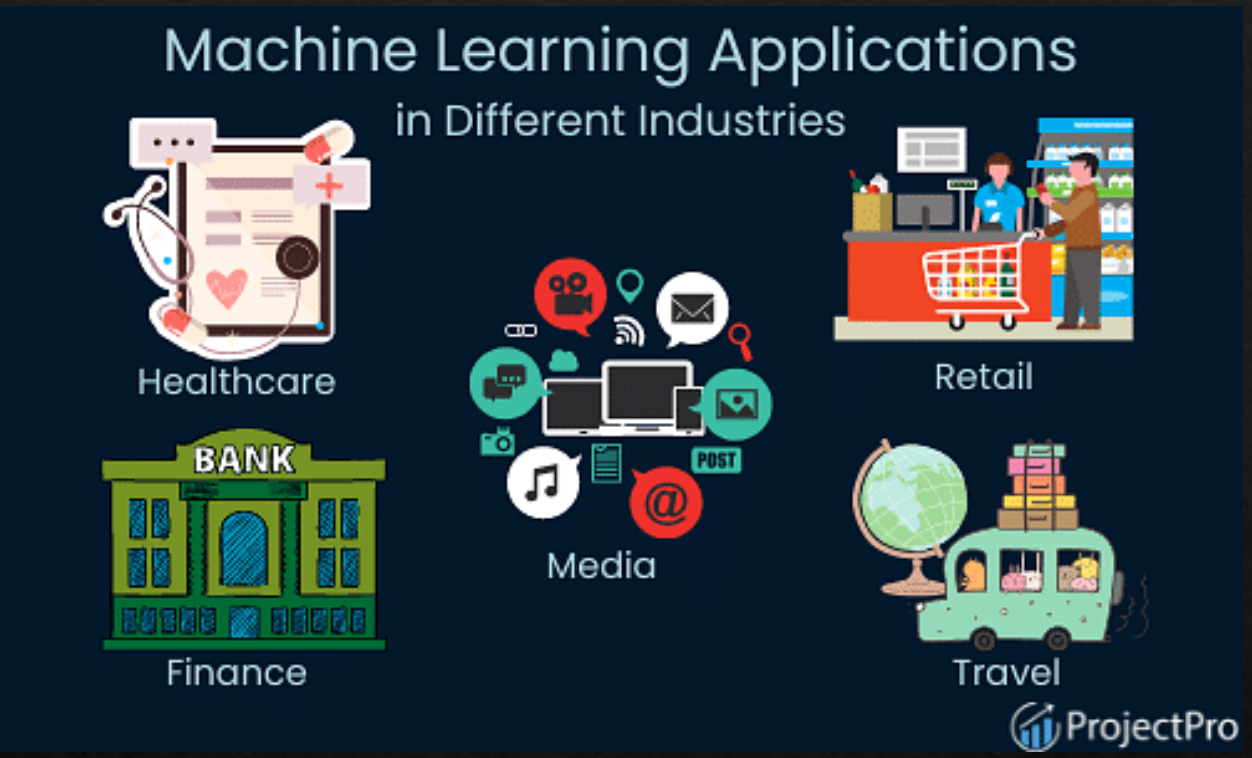
- AI is an umbrella term for any computer program that does something smart.
- Machine learning is a subset of AI.
- ML uses computing-based systems to make sense of data.
- ML extracting patterns, fitting data to functions, classifying data, etc
- Deep learning is a subset of machine learning.



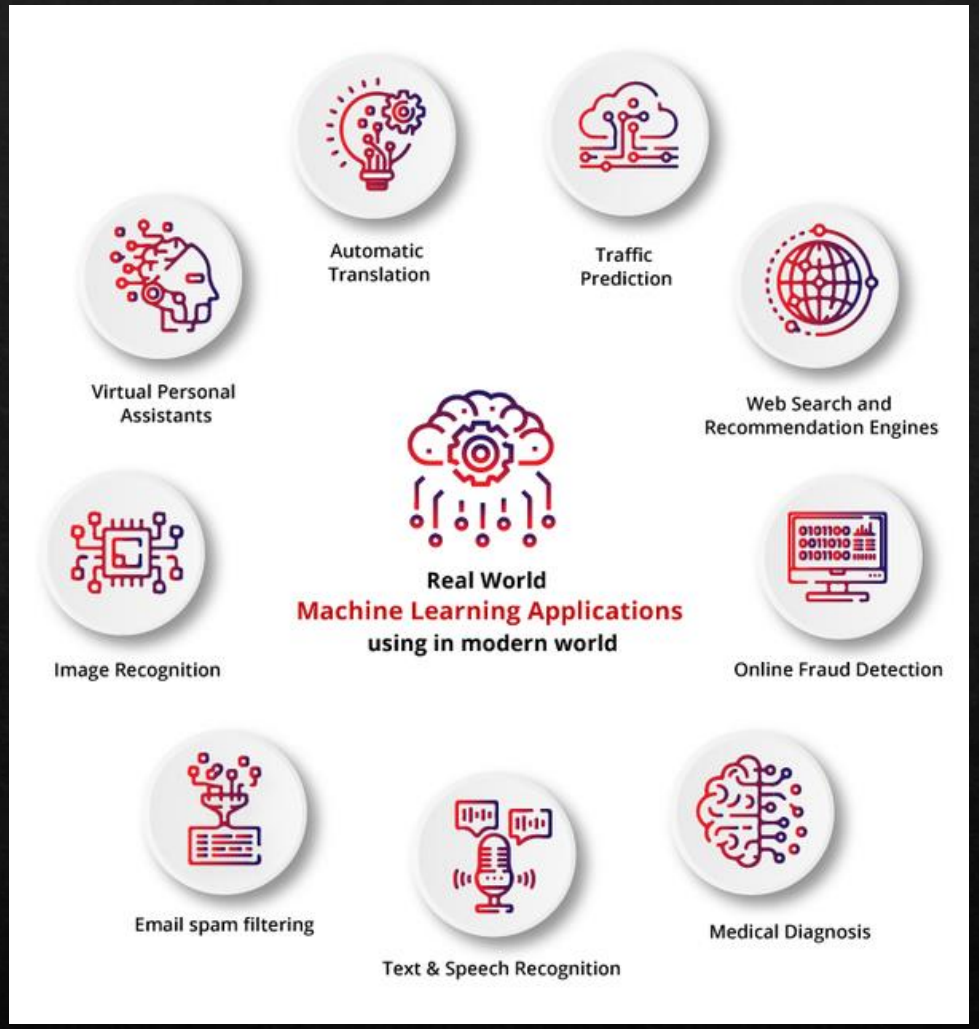
https://www.linkedin.com/pulse/artificial-intelligence-ai-vs-machine-learning-deep-natarajan-siva/

Machine Learning

- In 1959, Arthur Samuel, a pioneer in the field of machine learning (ML) defined it as the “field of study that gives computers the ability to learn without being explicitly programmed”.
- Tom Mitchell provides a more modern definition: “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.”
- Learning is the process of updating the parameters of the model so the model performs better.
- Bridges theoretical computer science and real noise data.



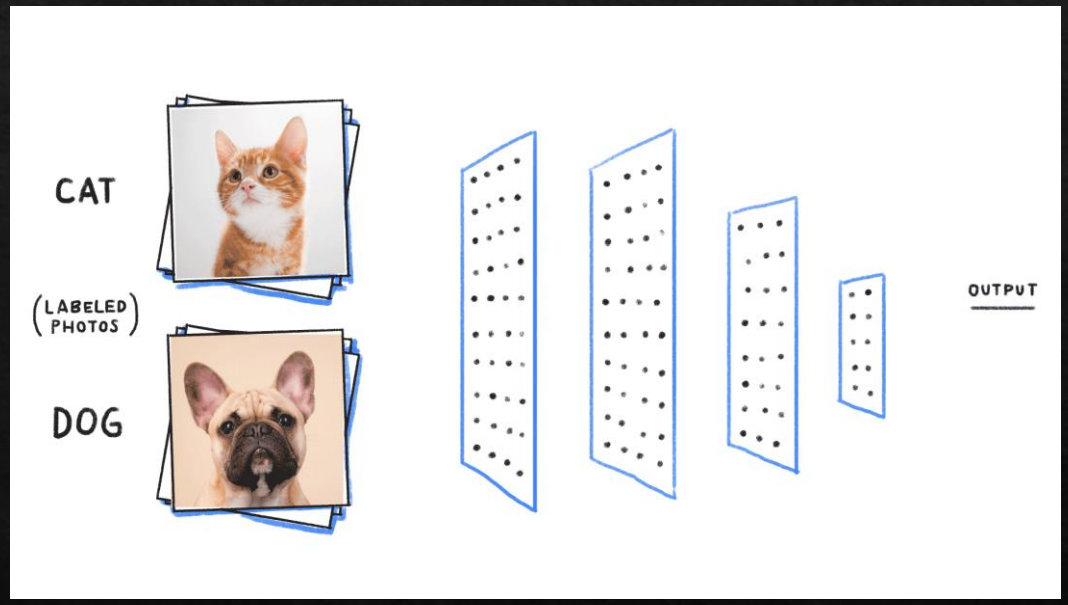
https://www.projectpro.io/article/10-awesome-machine-learning-applications-of-today/364



https://studymachinelearning.com/applications-of-machine-learning/

Developing a machine learning model.

- Requires data (Image, numbers/matrices, text, etc.)
  - Image: Computer Vision, Text: Natural Language Processing
- Understanding the data (Preprocessing)
- Availability of Labels.
- Task (Classification, Regression).
- Selecting a Model/algorithm.
- Evaluation criteria.

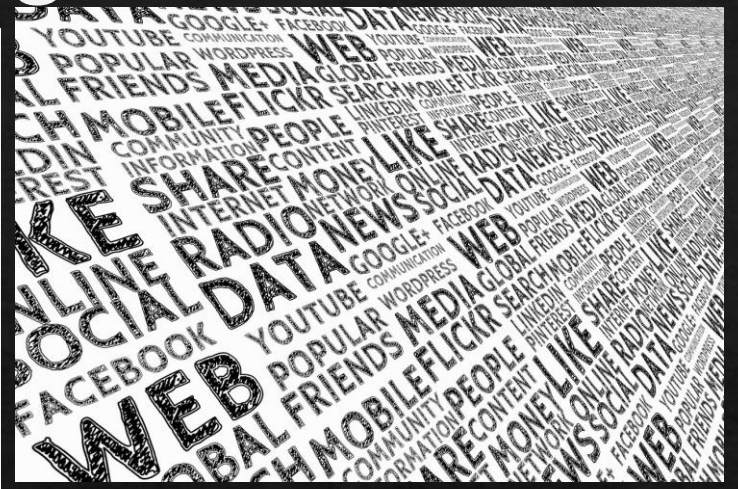


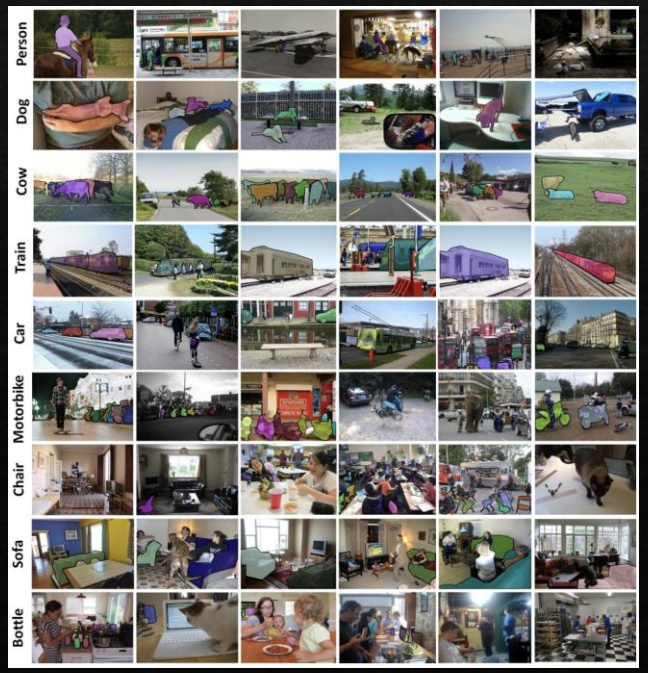
https://medium.com/analytics-vidhya/different-types-of-machine-learning-algorithm-b4f76b5730fd

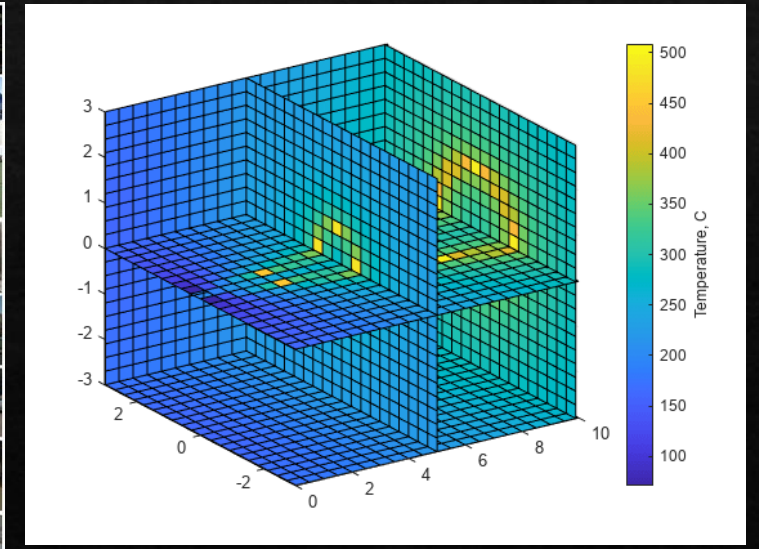
Developing a machine learning model.

- Requires data (Image, numbers/matrices, text, etc.).
- Understanding the data (Preprocessing)
- Availability of Labels
- Task (Classification, Regression).
- Selecting a Model/algorithm.
- Evaluation criteria.

	A	B	C	D
1	First Name	Last Name	Age	Salary
2	Jon	Smith	36	26500
3	Helen	Mirren	22	21000
4	David	Cameron	29	39000
5	Brad	Pitt	52	45000
6	Anna	Starolsky	41	22500
7	Peter	Piper	20	31500
8	David	Duck	19	15700
9	Julie	Walters	33	19000

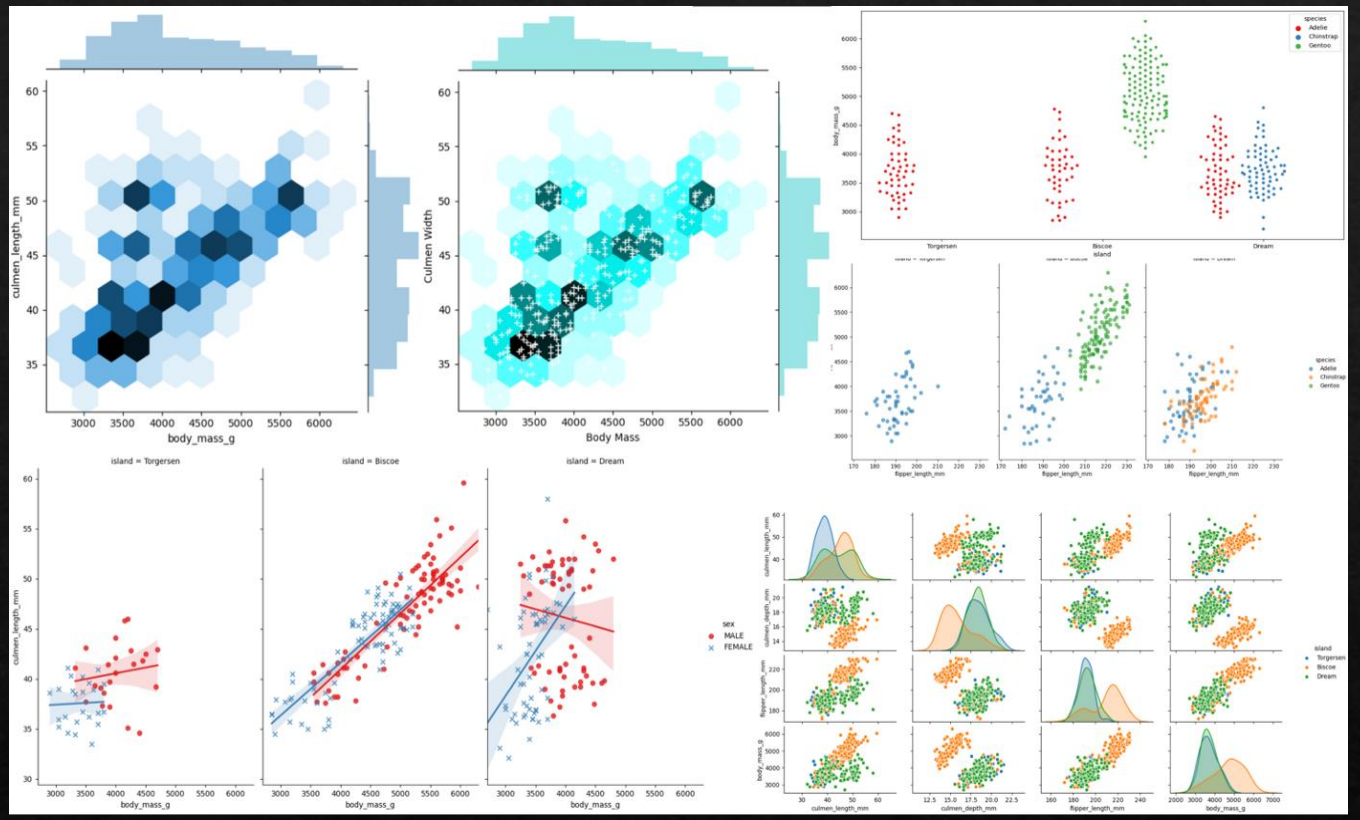






Developing a machine learning model.

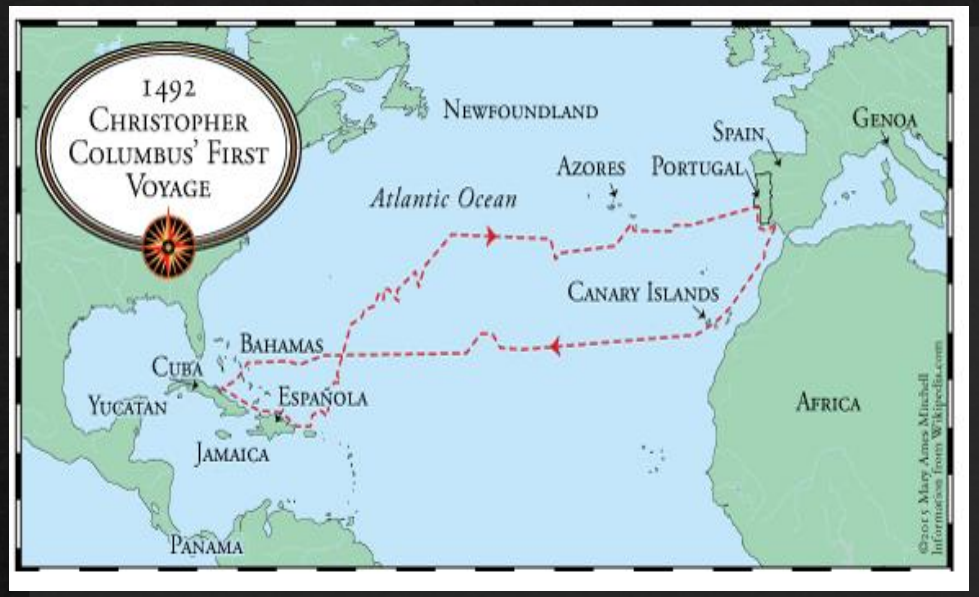
- Requires data (Image, numbers/matrices, text, etc.).
- Understanding the data (Preprocessing)
- Availability of Labels
- Task (Classification, Regression).
- Selecting a Model/algorithm.
- Evaluation criteria.



https://towardsdatascience.com/5-advanced-visualisation-for-exploratory-data-analysis-eda-c8eafeb0b8cb

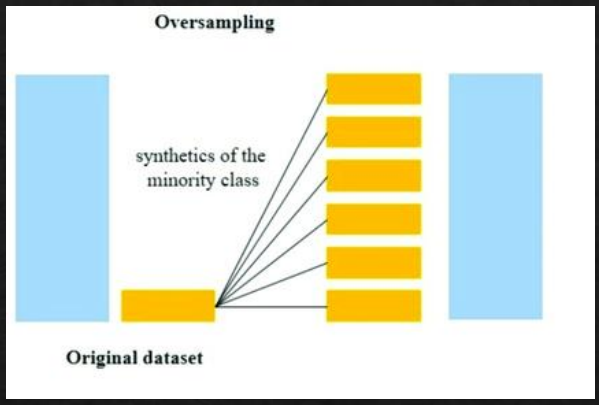
Why Preprocessing is important?

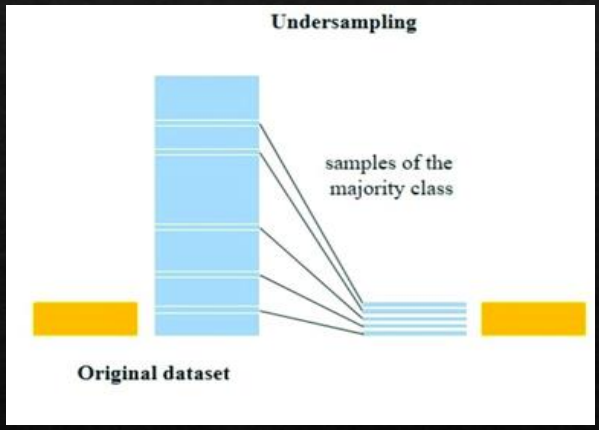
- Wrong use of data may lead to disasters.
- Trans-Atlantic Voyage 1492: Cristopher Columbus forgot to convert Arabic miles (By Alfrangus) to Roman miles and landed in the Americas rather than Asia.
- Mars Climate Orbiter (1999): Orbiter never made it because of disintegration and not converting data to a metric system.
- In 2018: Amazon's AI recruitment tool discarded woman applicants due to bias against women due to the misrepresentation of women in their data.
- Refined data may lead to more accurate decisions.
- In the clinical domain, it becomes more important to diagnose correctly for proper treatment.
- Preprocessing is a necessary step before feeding data to the algorithm.

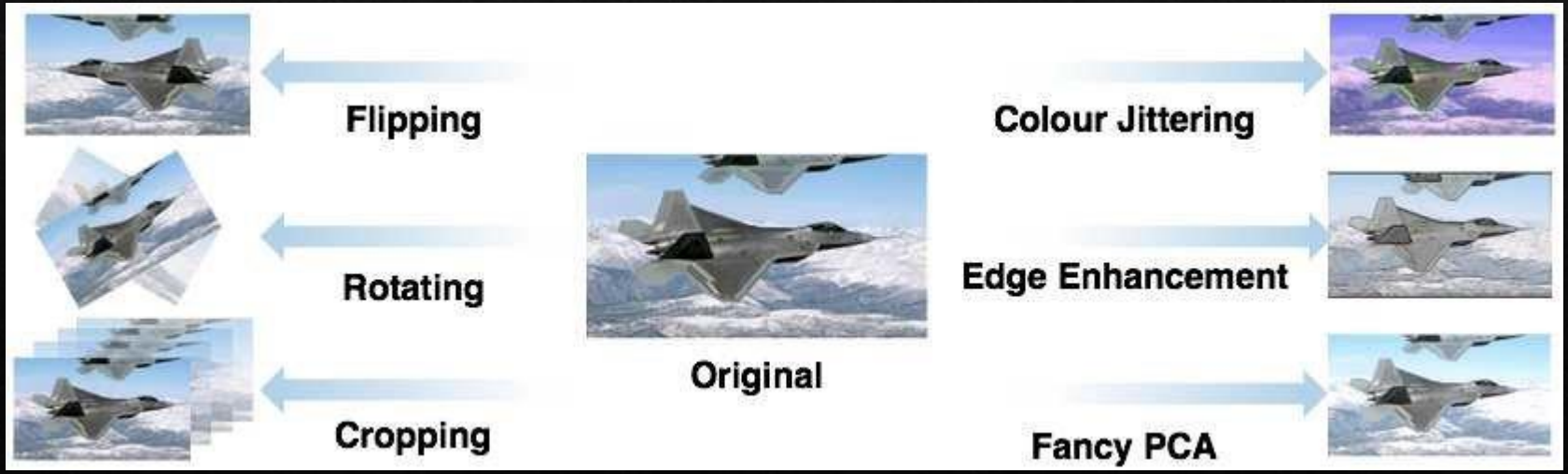


Data Preprocessing

- Class Imbalance - > Skewed distribution
- Solution: Under-sampling majority class or oversampling minority class.
- Data Augmentation
- Synthetic Data -> GAN (State-of-the-art)



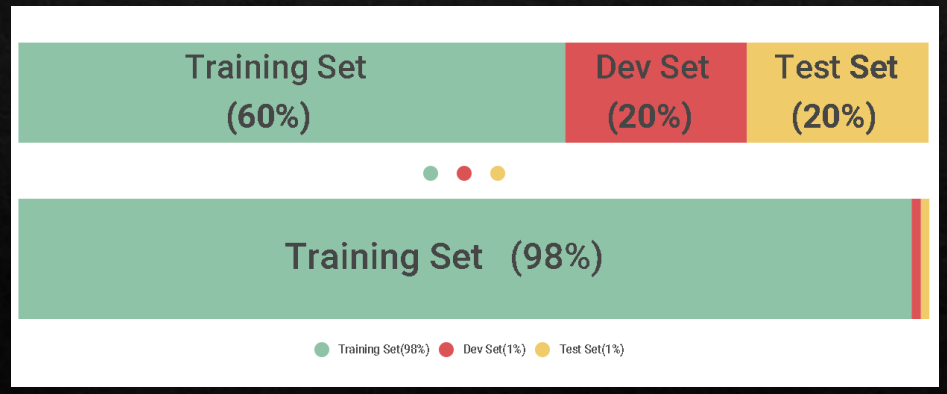




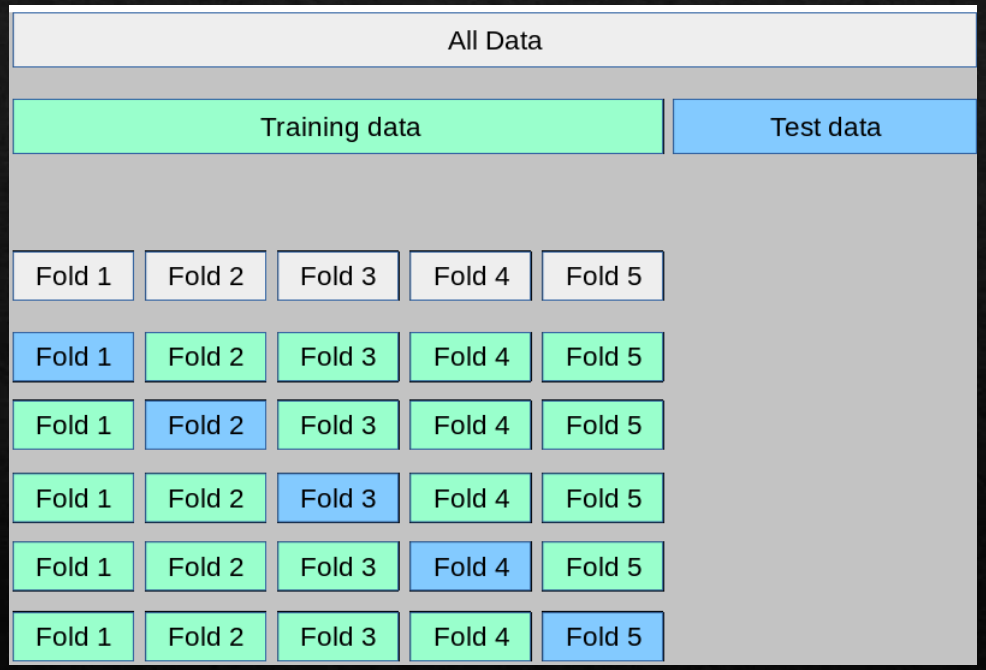
https://www.researchgate.net/publication/319210096_Improving_Deep_Learning_using_Generic_Data_Augmentation

Data Preprocessing

- Data Scientists at IBM spend 80% of their time in preprocessing.
- Preprocessing steps: Quality, Authenticity (Noiseless), Missing Values, Smoothing, transformation.
- Splitting dataset (Training, Validation, Test).
- Cross-Validation



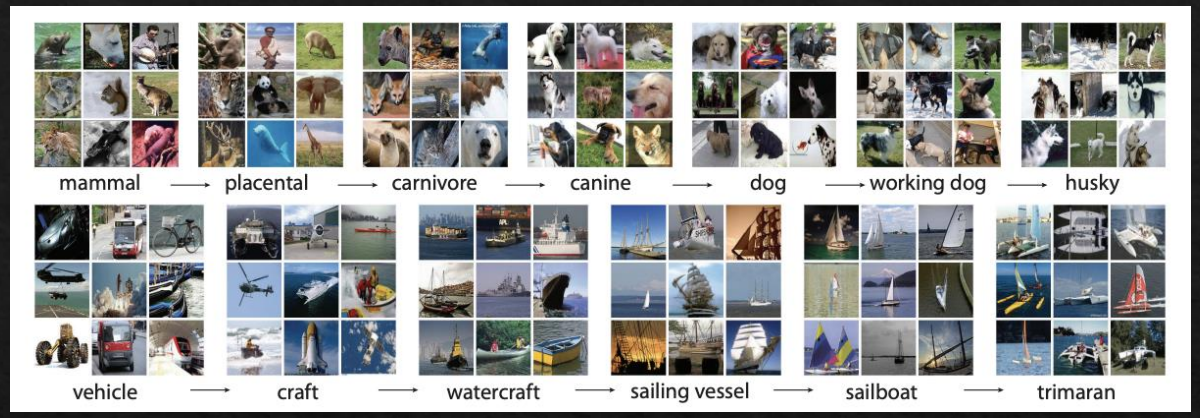
https://towardsdatascience.com/data-splitting-technique-to-fit-any-machine-learning-model-c0d7f3f1c790



https://www.ibm.com/cloud/blog/ibm-data-catalog-data-scientists-productivity

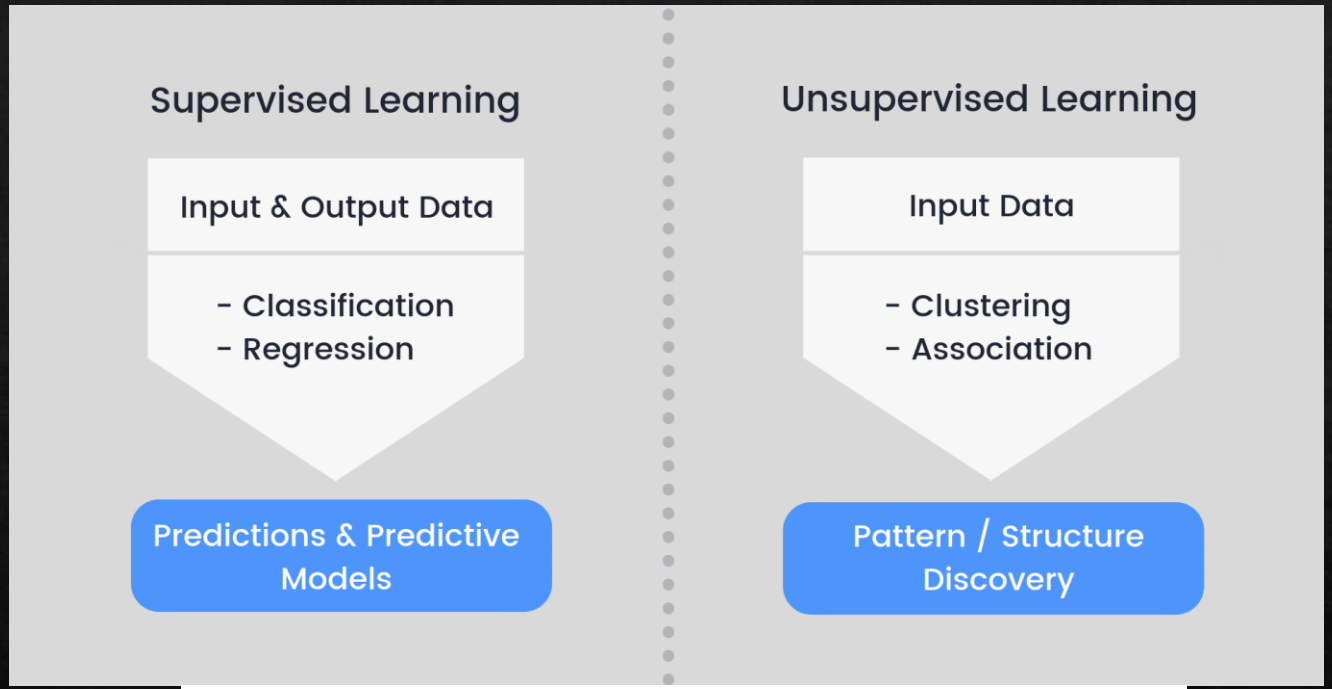
Developing a machine learning model.

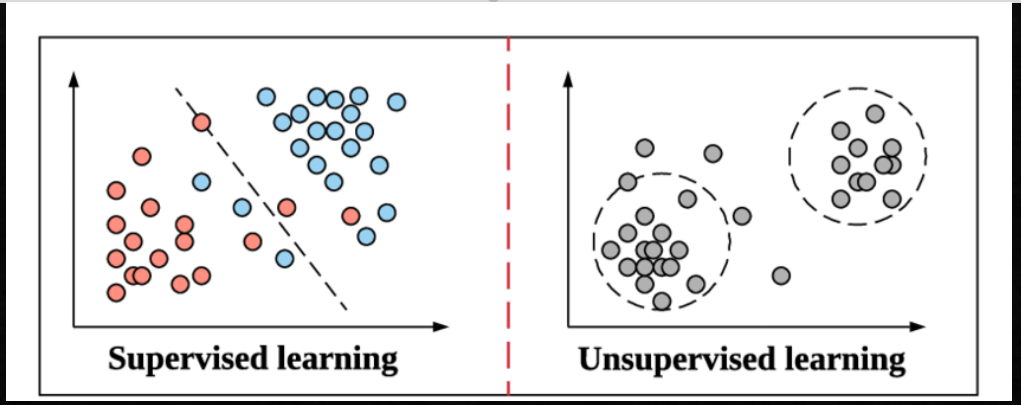
- Requires data (Image, numbers/matrices, text, etc.).
- Understanding the data (Preprocessing)
- Availability of Labels
- Task (Classification, Regression).
- Selecting a Model/algorithm.
- Evaluation criteria.



Developing a machine learning model.

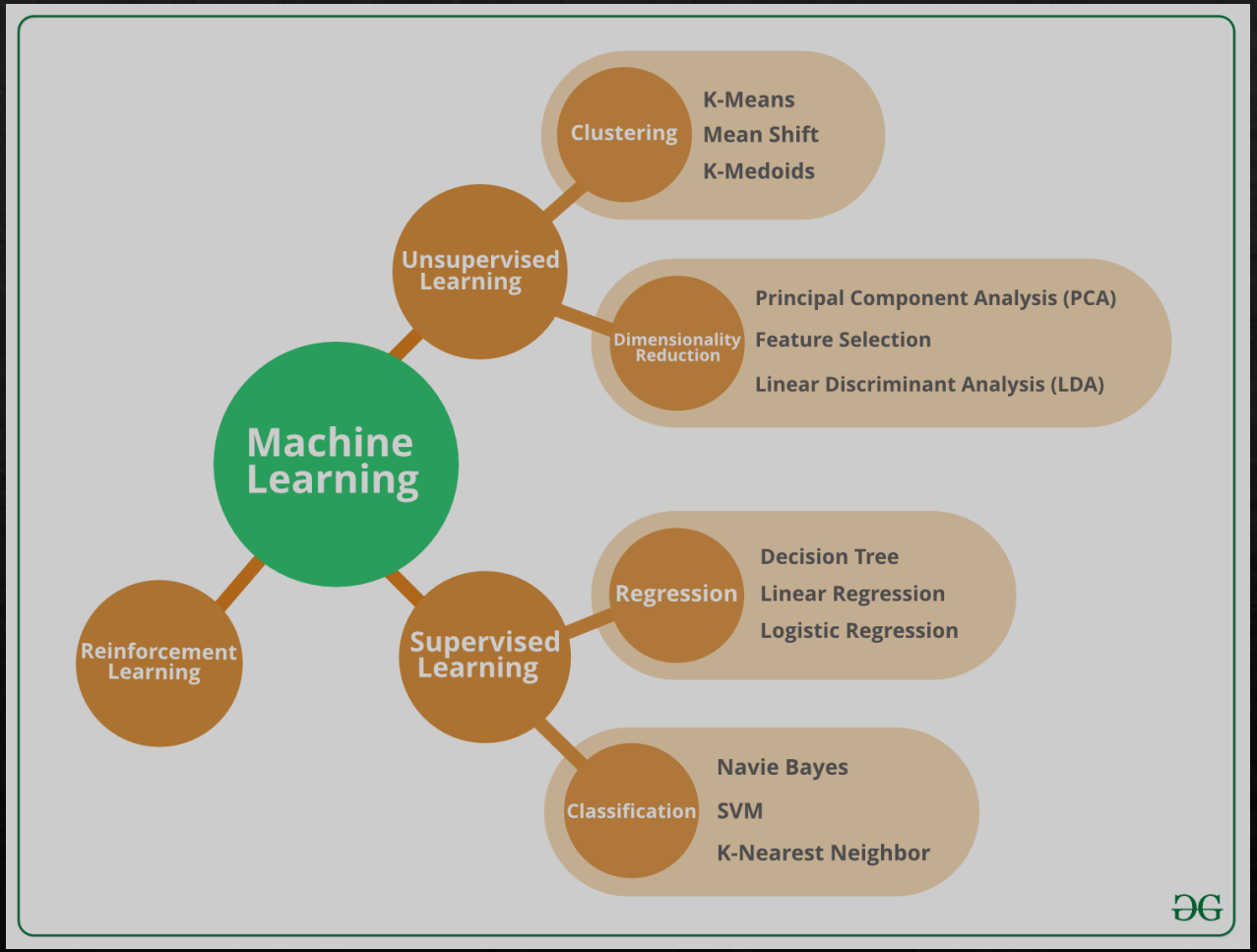
- Requires data (Image, numbers/matrices, text, etc.).
- Understanding the data (Preprocessing)
- Availability of Labels
- Task (Classification, Regression).
- Selecting a Model/algorithm.
- Evaluation criteria.





Developing a machine learning model.

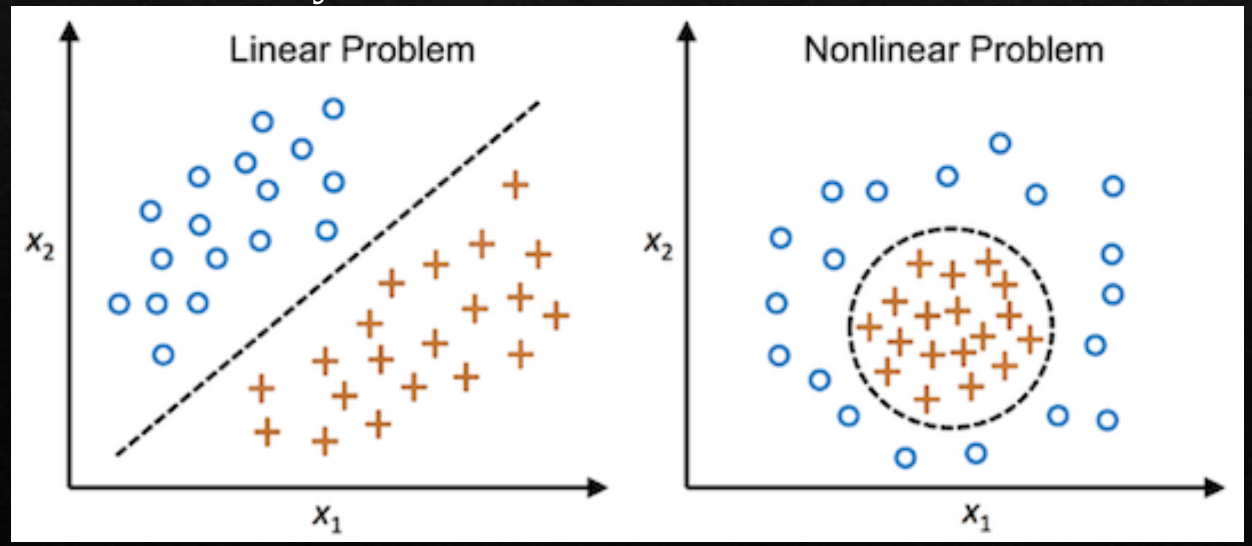
- Requires data (Image, numbers/matrices, text, etc.).
- Understanding the data (Preprocessing)
- Availability of Labels
- Task (Classification, Regression).
- Selecting a Model/algorithm.
- Evaluation criteria.



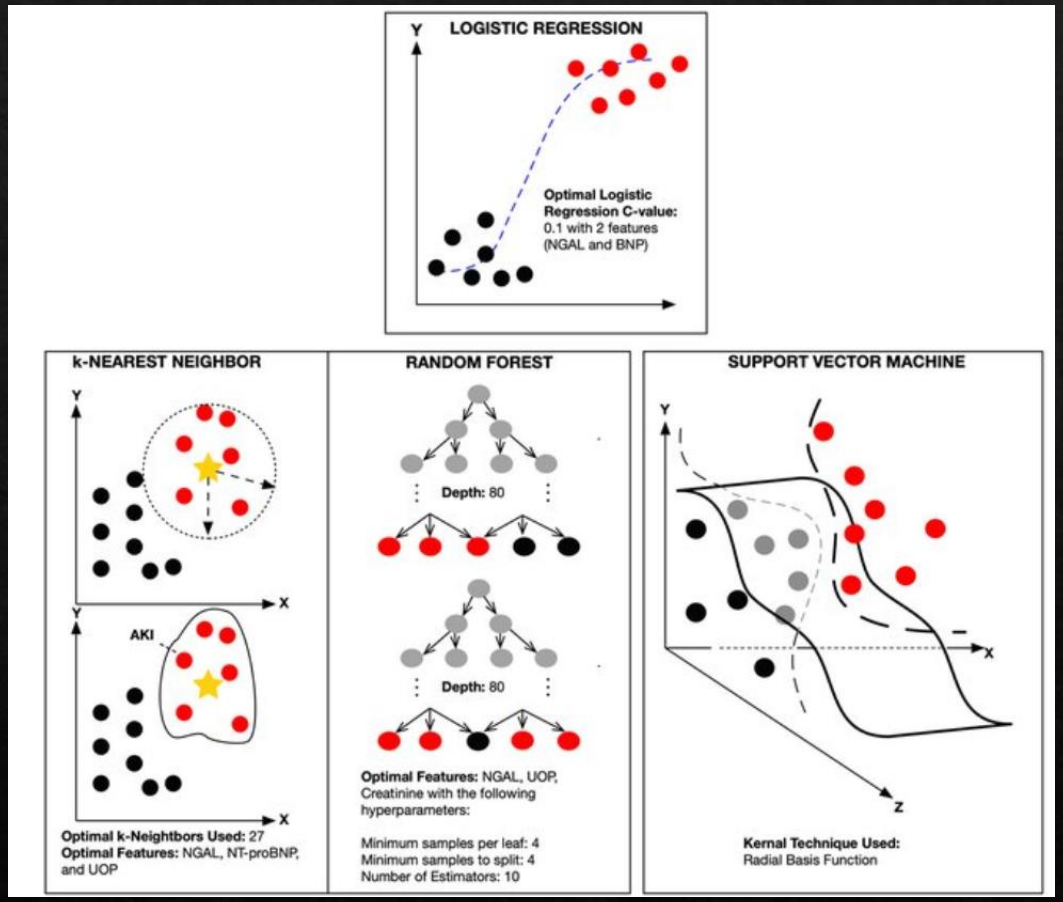
https://medium.com/analytics-vidhya/different-types-of-machine-learning-algorithm-b4f76b5730fd

Selecting a ML Model

- KNN performs well on a small dataset and classifies based on nearest neighbors.
- SVM is less computationally demanding than KNN, and requires a Kernel, works well with bigger and linearly separated datasets.
- Random forest is training data size intensive and relatively slow.



https://towardsdatascience.com/decision-trees-and-random-forests-df0c3123f991



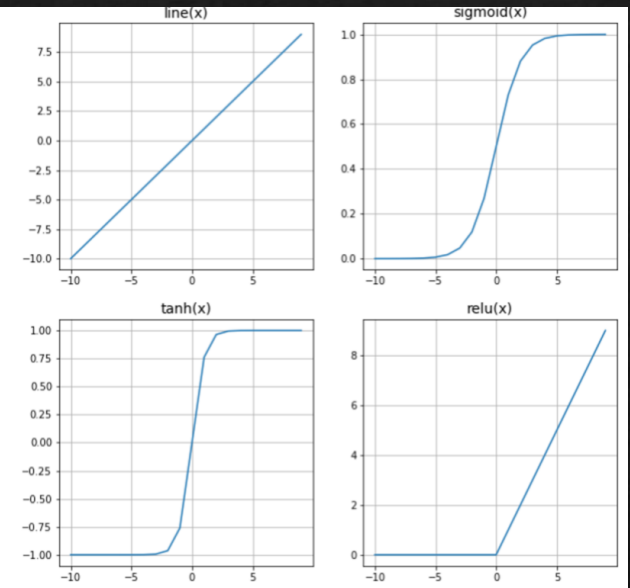
DOI: 10.1038/s41598-019-57083-6

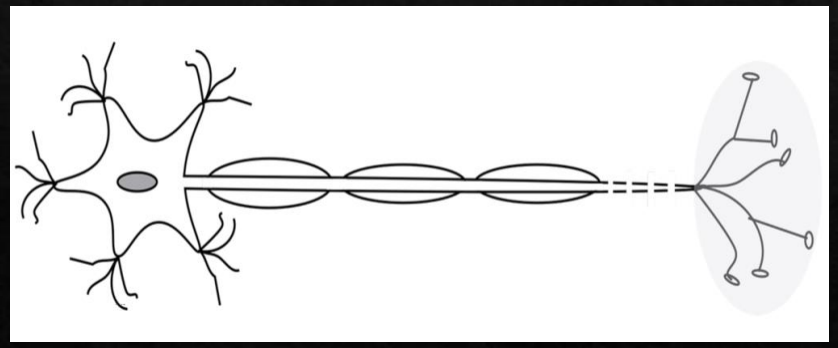
Artificial Neural Network

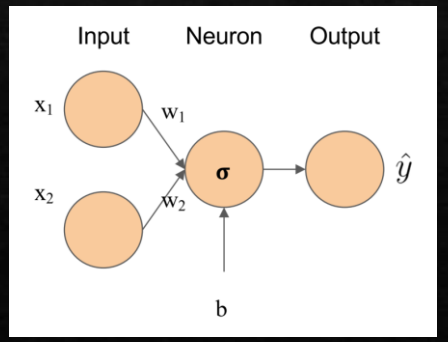
- Neural Networks, more accurately known as Artificial Neural Networks, is inspired by the human brain, mimicking the way that biological neurons signal to one another.
- Activation function add non-linearity before passing the output to another neuron.

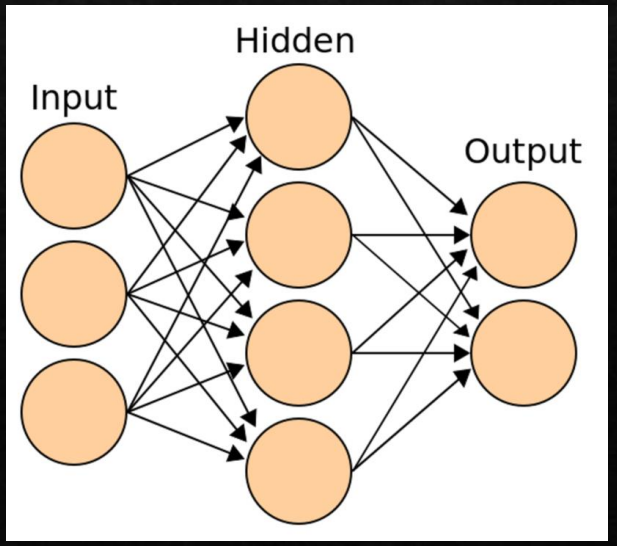
Types of activation function:

- Linear: f(x) = x
- Sigmoid: $\frac{1}{1+e^{-x}}$
- Tanh: $\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$
- Relu: $max(0,x)$





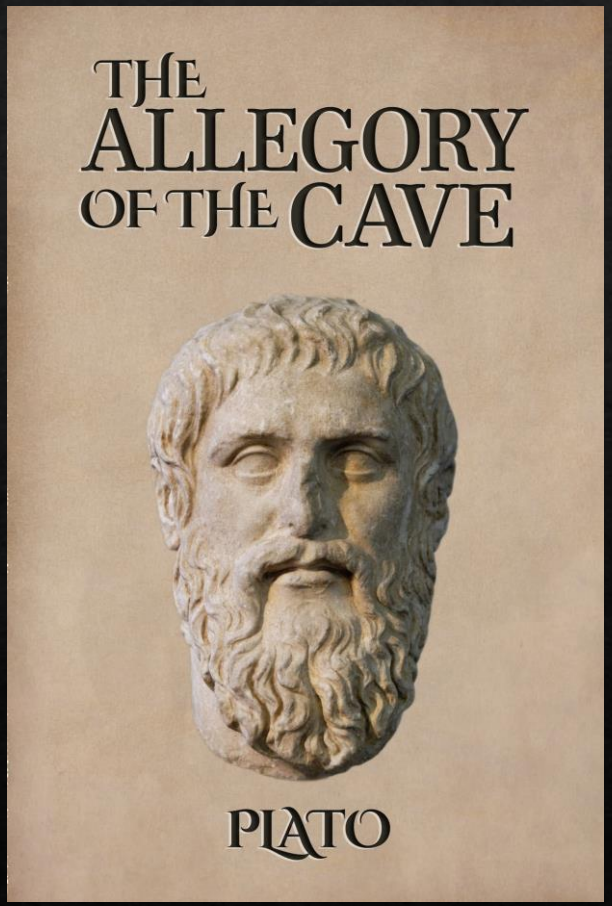




Unsupervised Learning

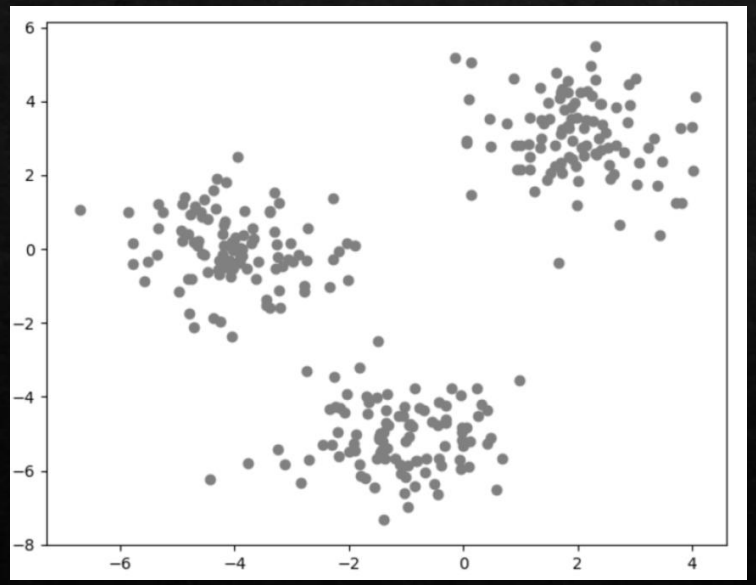
- Problem: 1. What to learn? 2. How to learn?
- Data carries hidden patterns.
- Model finds the underlying patterns in the data.

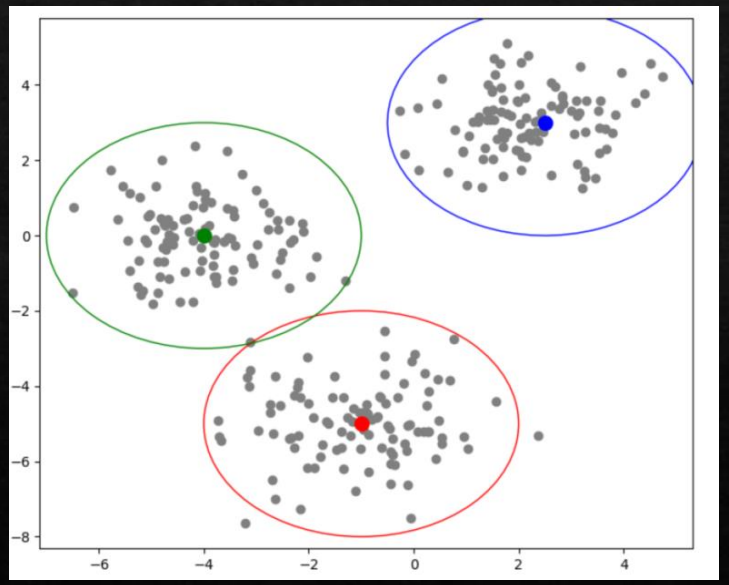
https://www.thoughtco.com/the-allegory-of-the-cave-120330

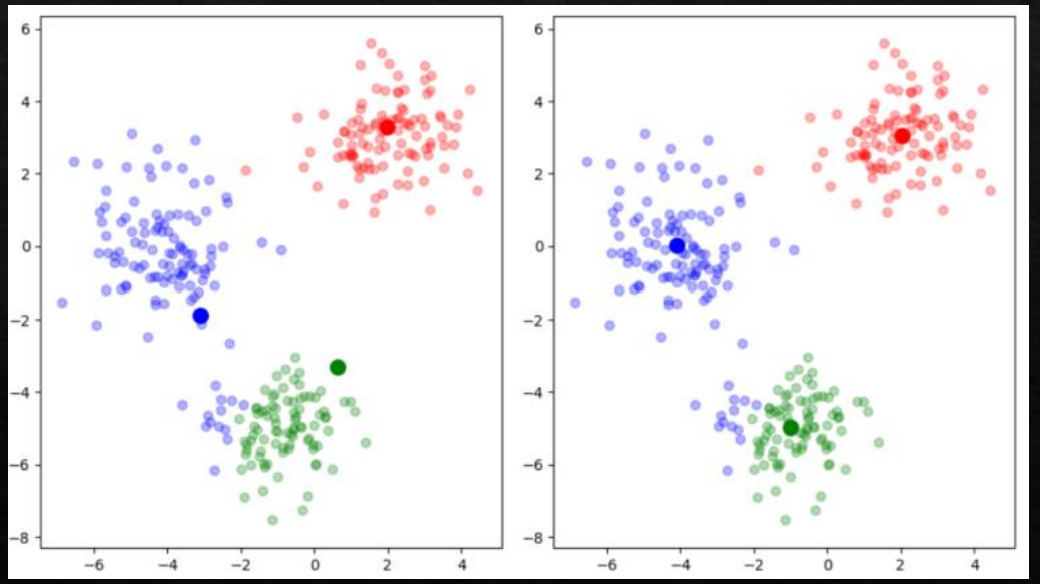


Unsupervised Algorithms

- Problem: 1. What to learn? 2. How to learn?
- **K means Clustering**: K-denotes clusters, assigns clusters centers randoms at first, and computes new clusters based on the previous center iteratively.
- Don’t confuse KNN (supervised) with K-means (unsupervised).

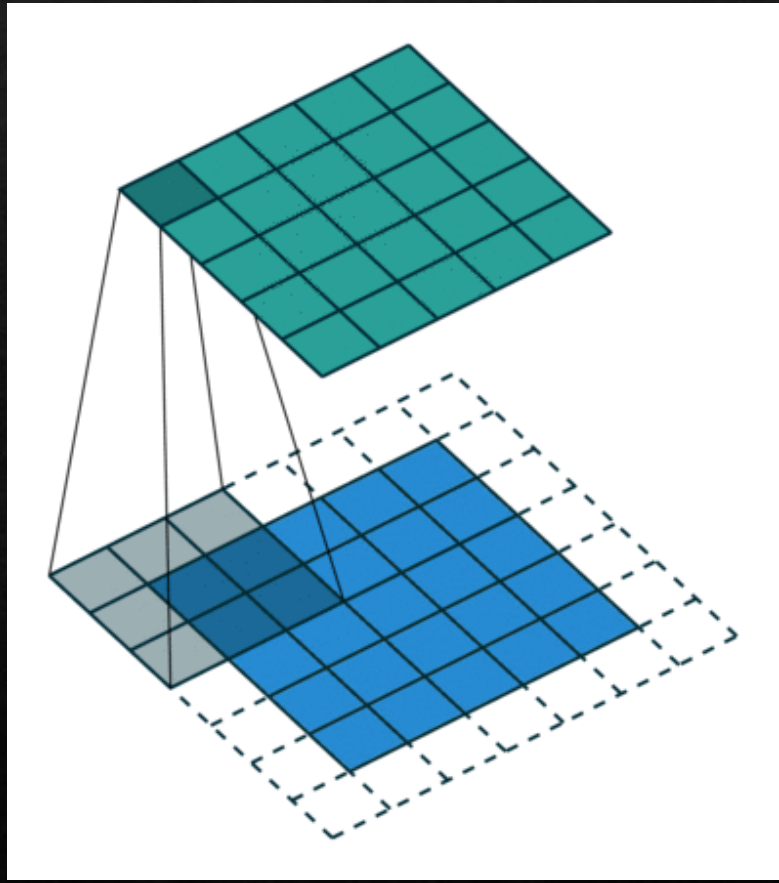


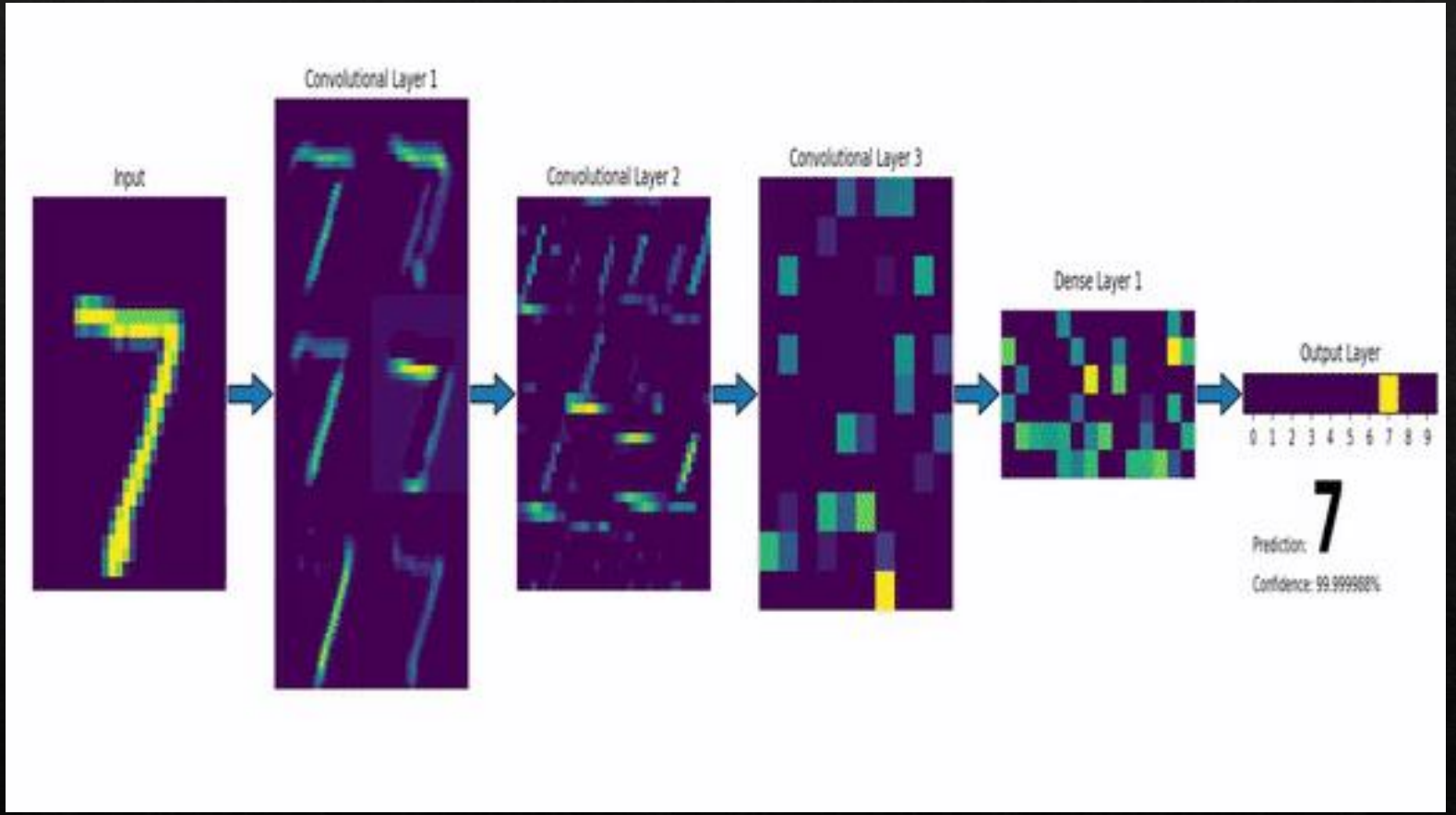




https://becominghuman.ai/comprehending-k-means-and-knn-algorithms-c791be90883d

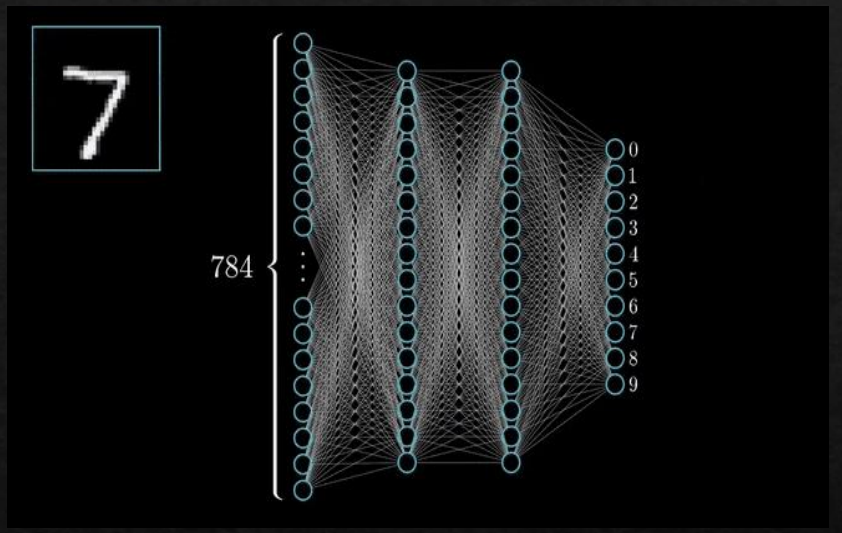
Convolutional Neural Networks

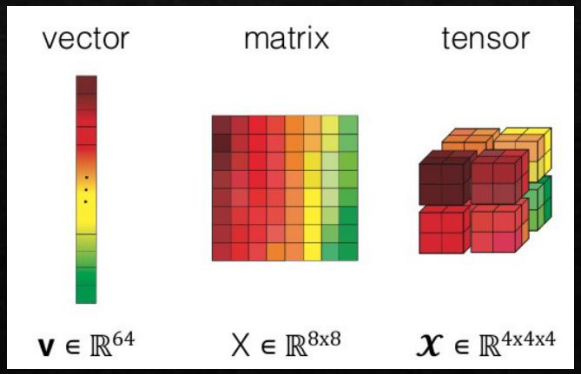




Training ML Model

- Neural Networks approximate a function, a suitable model with a number of hidden layers
- Complex models may not converge.
- Training hyperparameters (learning rate, optimizer, etc.)
  - **Batch size:** dividing subset further based on available memory.
  - **Epoch:** One iteration over the subset.
  - **Learning rate:** Tuning parameter to update weights.
  - **Optimizer:** to update the weights and biases. (Adam, SGD, Adagrad etc.)
  - **Loss function:** Objective function to minimize over every epoch (CrossEntropy, MAE etc.)
- Type of hardware (CPU vs. GPU vs. TPU)
  - **CPU:** primary hardware to run all arithmetic, logic, control input/output
  - **GPU:** can handle parallel computing, high data throughput
  - **TPU:** powerful custom-built processors to run the project made on a specific framework

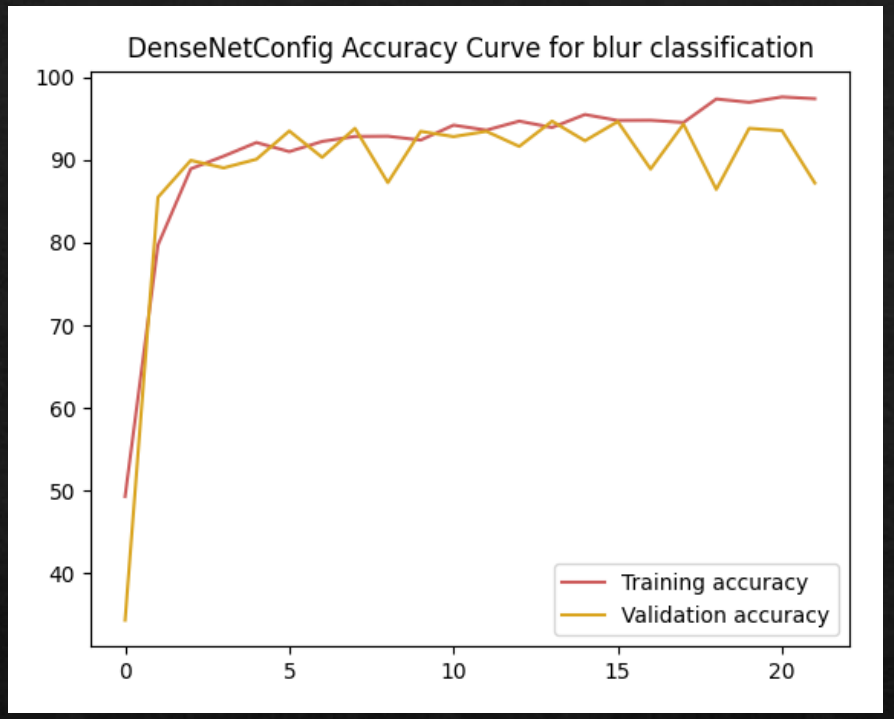




https://medium.com/analytics-vidhya/numpy-on-gpu-tpu-efb8d367020a

Evaluating ML Models

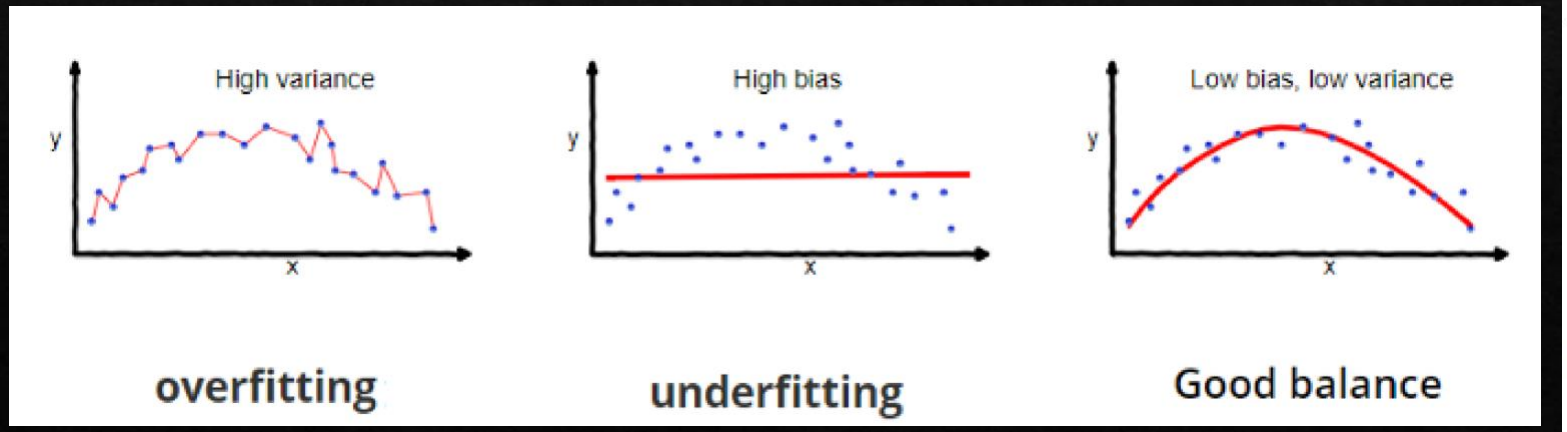
- Metrics: Confusion Matrix, F1
- Best models selected on Validation.
- Generalization: Real-World Data.
- Adversarial Attacks.
- Learning Curves (Training and Validation accuracy/losses)

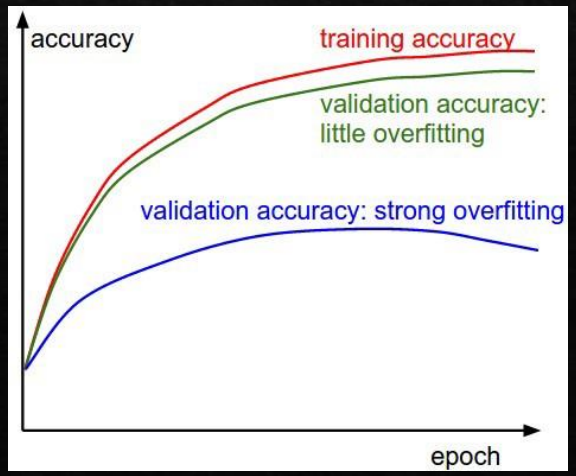


		Ground truth +	Ground truth -	
Predicted	+	True positive (TP)	False positive (FP)	Precision = TP / (TP + FP)
Predicted	-	False negative (FN)	True negative (TN)	
		Recall = TP / (TP + FN)		Accuracy = (TP + TN) / (TP + FP + TN + FN)

Generalization Error

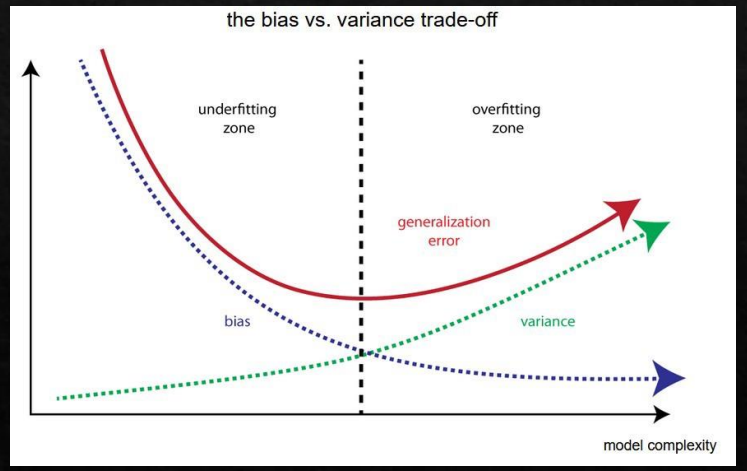
- Bias is the simplifying assumptions made by the model to make the target function easier to approximate. Error due to inaccurate assumptions/simplifications made by the model.
- Variance is the amount that the estimate of the target function will change given different training data.
  - **Underfitting**: model is too “simple” to represent all the relevant class characteristics
  - **Overfitting**: model is too “complex” and fits irrelevant characteristics (noise) in the data



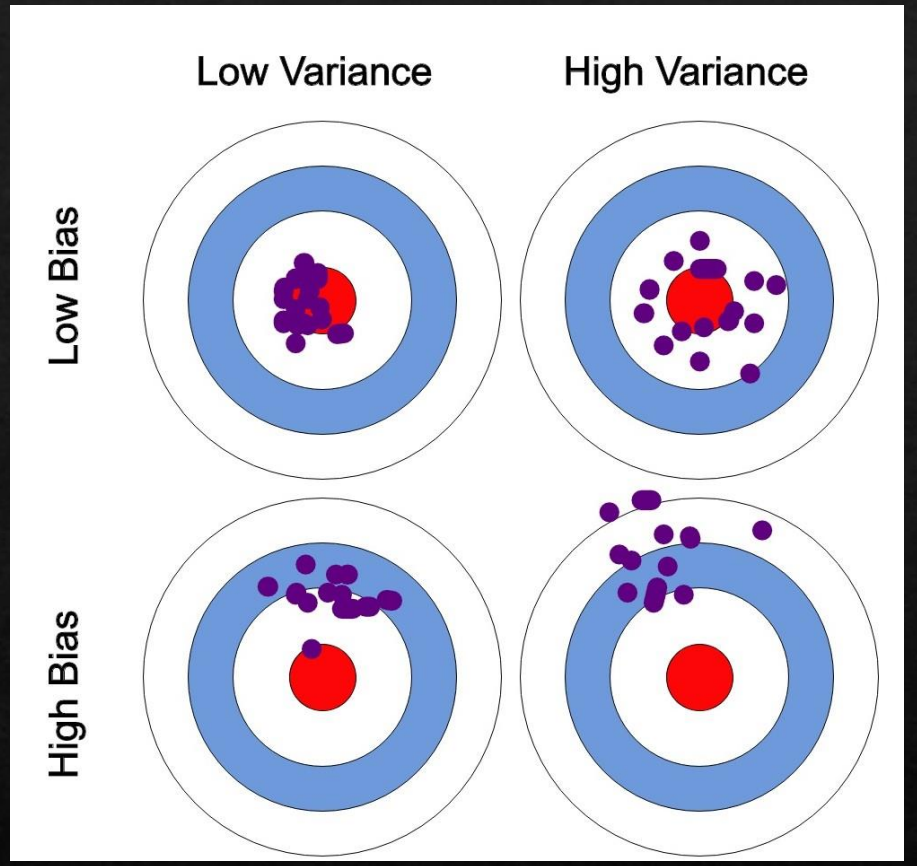


Bias-Variance Tradeoff

- Models with too few parameters are inaccurate because of a large bias (not enough flexibility).
  - High bias and low variance
  - High training error and high test error
- Models with too many parameters are inaccurate because of a large variance (too much sensitivity to the sample).
  - Low bias and high variance
  - Low training error



https://towardsdatascience.com/bias-and-variance-but-what-are-they-really-ac539817e171



https://nvsyashwanth.github.io/machinelearningmaster/bias-variance/

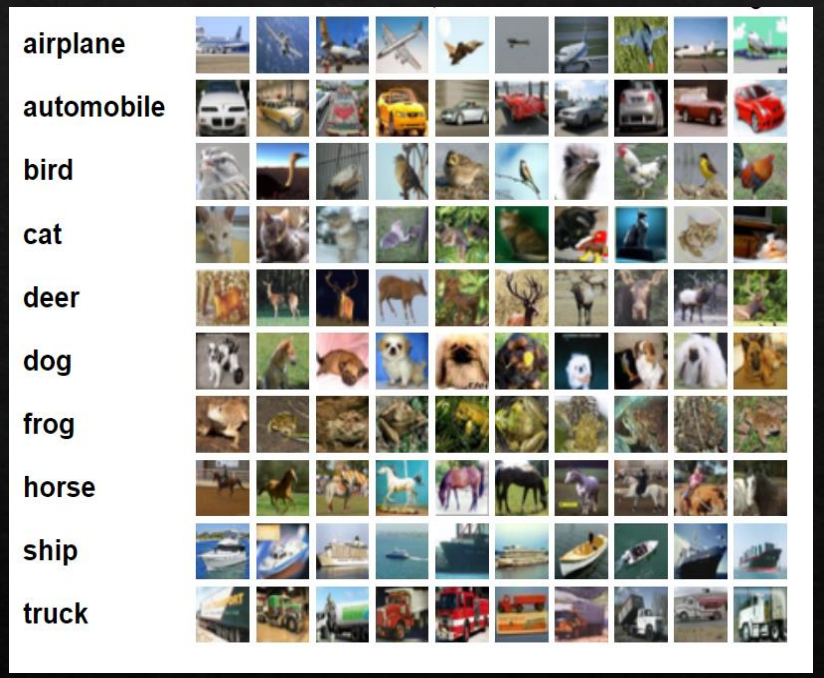
Benchmarks Image Classification

- MNIST

60000 small squares 28*28, grayscale

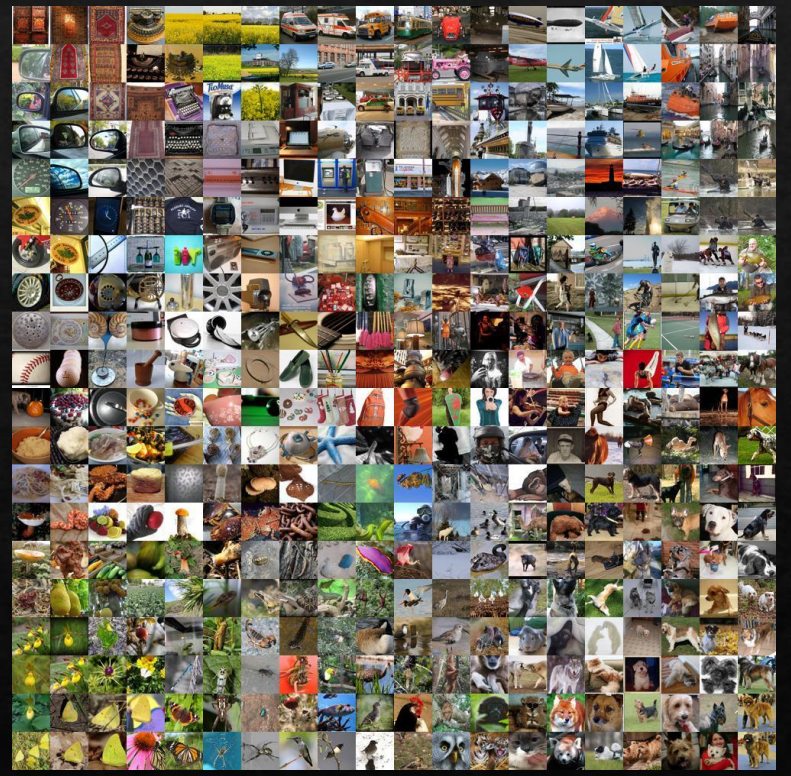
CIFAR-10

60000 color images, 32*32



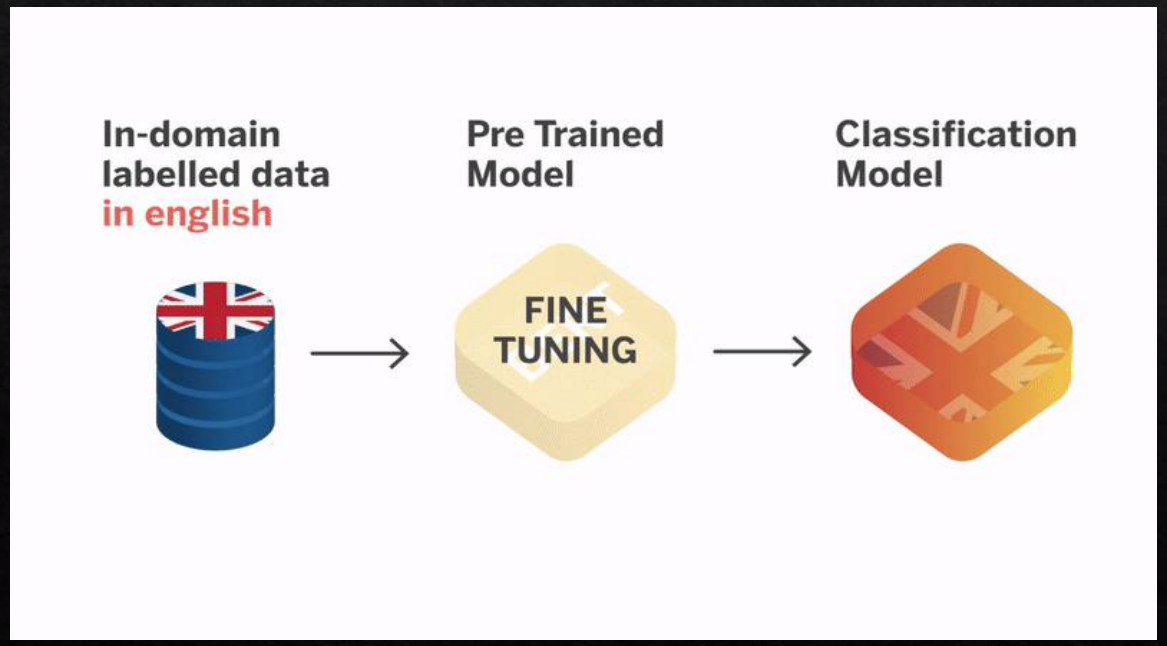
ImageNet

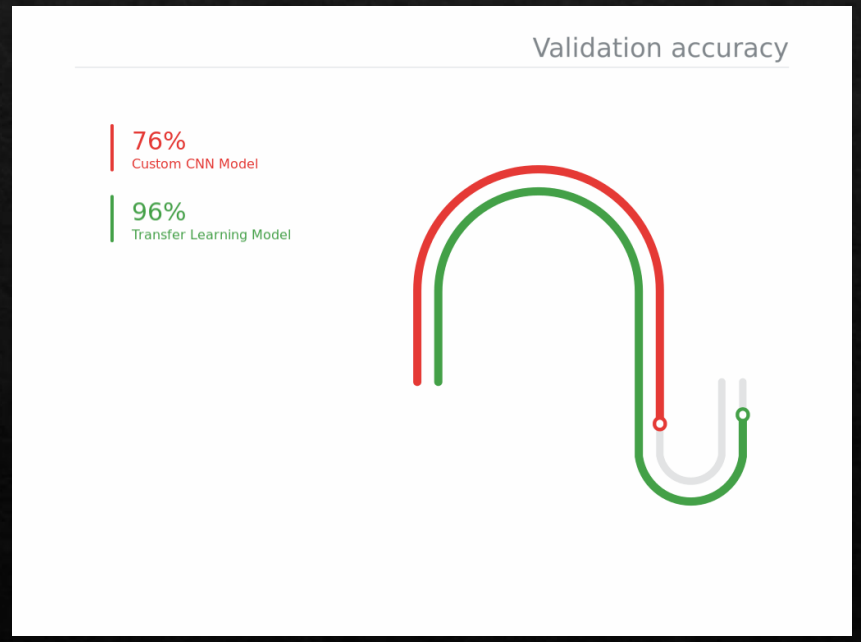
14,197,122 images, 469*387, Object Detection, Image Segmentation



Transfer Learning

- Traditional ML: Isolated task; knowledge is learned from the data for the task.
- Transfer Learning: Use the previous knowledge to learn a new task.
- Fine Tuning: Freezing a part of a network and updating the classifier.





https://deepnote.com/@jhon-smith-flores/Transfer-Learning-864f7d51-84f9-4d43-baa0-6194de7943de

Machine Learning Algorithm Cheat Sheet

This cheat sheet helps you choose the best machine learning algorithm for your predictive analytics solution.
Your decision is driven by both the nature of your data and the goal you want to achieve with your data.

Microsoft Azure

What do you want to do?

Extract information from text

Text Analytics

Derives high-quality information from text

Answers questions like: What info is in this text?

- Latent Dirichlet Allocation ← Unsupervised topic modeling, group texts that are similar
- Extract N-Gram Features from Text ← Creates a dictionary of n-grams from a column of free text
- Feature Hashing ← Converts text data to integer encoded features using the Vowpal Wabbit library
- Preprocess Text ← Performs cleaning operations on text, like removal of stop-words, case normalization
- Word2Vector ← Converts words to values for use in NLP tasks, like recommender, named entity recognition, machine translation

Predict values

Regression

Makes forecasts by estimating the relationship between values

Answers questions like: How much or how many?

- Fast Forest Quantile Regression ← Predicts a distribution
- Poisson Regression ← Predicts event counts
- Linear Regression ← Fast training, linear model
- Bayesian Linear Regression ← Linear model, small data sets
- Decision Forest Regression ← Accurate, fast training times
- Neural Network Regression ← Accurate, long training times
- Boosted Decision Tree Regression ← Accurate, fast training times, large memory footprint

Generate recommendations

Recommenders

Predicts what someone will be interested in

Answers the question: What will they be interested in?

- Use the Train Wide & Deep Recommender module ← Hybrid recommender, both collaborative filtering and content-based approach
- SVD Recommender ← Collaborative filtering, better performance with lower cost by reducing dimensionality

Discover structure

Clustering

Separates similar data points into intuitive groups

Answers questions like: How is this organized?

- K-Means ← Unsupervised learning

Find unusual occurrences

Anomaly Detection

Identifies and predicts rare or unusual data points

Answers the question: Is this weird?

- One Class SVM ← Under 100 features, aggressive boundary
- PCA-Based Anomaly Detection ← Fast training times

Predict between several categories

Multiclass Classification

Answers complex questions with multiple possible answers

Answers questions like: Is this A or B or C or D?

- Multiclass Logistic Regression ← Fast training times, linear model
- Multiclass Neural Network ← Accuracy, long training times
- Multiclass Decision Forest ← Accuracy, fast training times
- One-vs-All Multiclass ← Depends on the two-class classifier
- One-vs-One Multiclass ← Depends on binary classifier, less sensitive to an imbalanced dataset with larger complexity
- Multiclass Boosted Decision Tree ← Non-parametric, fast training times and scalable

Predict between two categories

Two-Class Classification

Answers simple two-choice questions, like yes or no, true or false

Answers questions like: Is this A or B?

- Two-Class Support Vector Machine ← Under 100 features, linear model
- Two-Class Averaged Perceptron ← Fast training, linear model
- Two-Class Decision Forest ← Accurate, fast training
- Two-Class Logistic Regression ← Fast training, linear model
- Two-Class Boosted Decision Tree ← Accurate, fast training, large memory footprint
- Two-Class Neural Network ← Accurate, long training times

Classify images

Image Classification

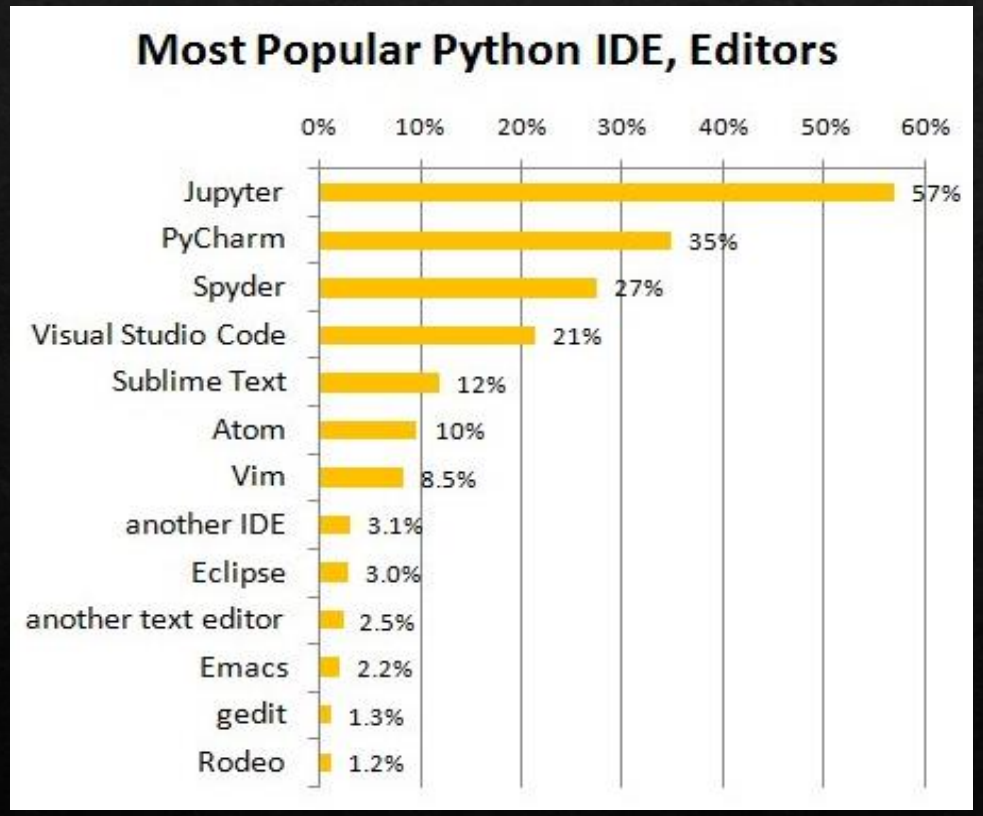
Classifies images with popular networks

Answers questions like: What does this image represent?

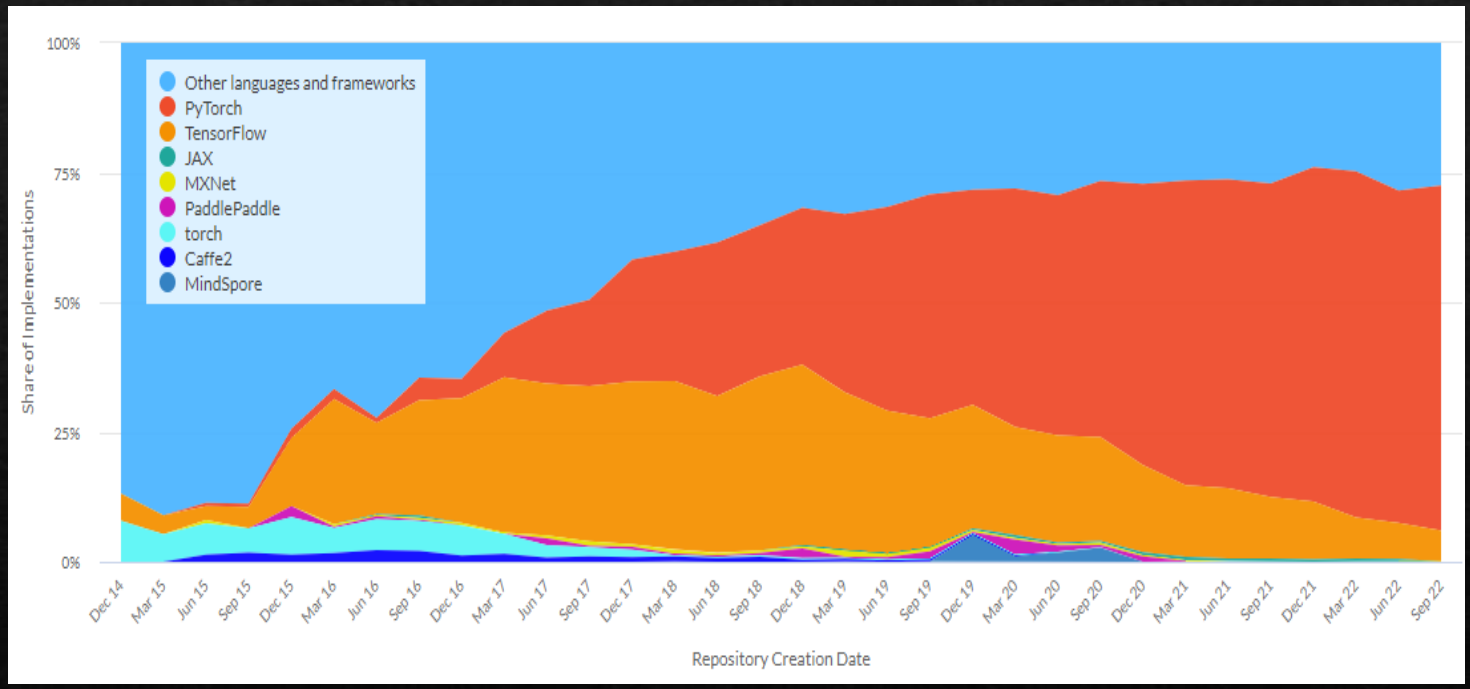
- ResNet ← Modern deep learning neural network
- DenseNet ← Modern deep learning neural network

Managing a ML Project

- Integrated Development Environment (IDE).
- DL Framework (Pytorch vs Tensorflow).



https://www.kdnuggets.com/2018/12/most-popular-python-ide-editor.html



https://paperswithcode.com/trends

Thank you

University of Stavanger