AI SUMMIT

COLUMBUS, OH • OCTOBER 25–27, 2022

ML CONCEPTS 101

I WAS SUDDENLY THROWN ONTO A PROJECT AND I DON'T UNDERSTAND THESE WORDS…

AFDO RAPS HEALTHCARE PRODUCTS COLLABORATIVE

AFDO RAPS

AI SUMMIT
COLUMBUS, OH • OCTOBER 25–27, 2022

Lacey Harbour
Regulatory Manager,
Thermo Fisher

History

- RA
- QA
- Clinical Lab
- Clinical Studies
- R&D

Not an AI Expert

Topics

- What is AI
- Data
- Learning/ Training Types
- Modeling
- Tasks/ Algorithm
- Training, Testing, and Validation
- Evaluation
- Challenges
- Platform and Pipeline

AI SUMMIT
COLUMBUS, OH • OCTOBER 25–27, 2022

Mentimeter Question

How familiar are you with AI/ML Concepts?

- Expert (Data Scientist, Modeler, etc.)
- 2+ ML Projects (Well Versed End User)
- 1 ML Projects (Beginner End User)
- Building Business Case for first ML Project
- Completely New professionally, but Understand Concepts
- Completely New Professionally and Conceptually
- AI Summit Working Team Member

AI SUMMIT
COLUMBUS, OH • OCTOBER 25–27, 2022

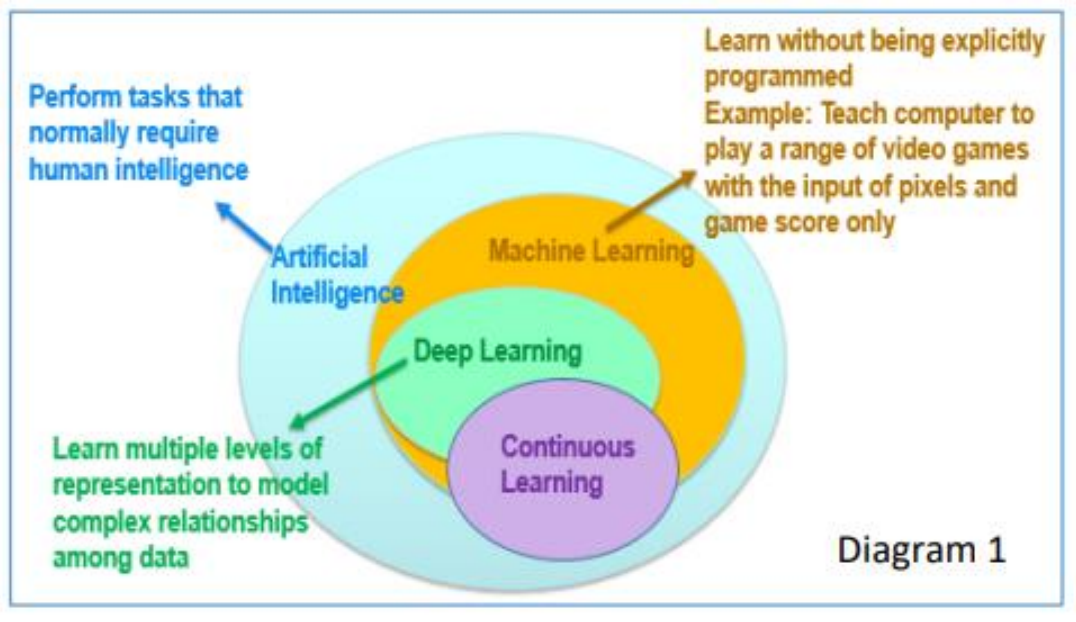
Mentimeter Question:

Are there any concepts you would like support clarifying?

Starting with the obvious…

What is AI?

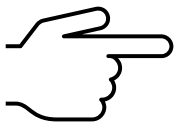
What is AI/ML/CLS/Deep Learning?



Perspectives and Good Practices for AI and Continuous Learning Systems in Healthcare, GMLP Team, 2018

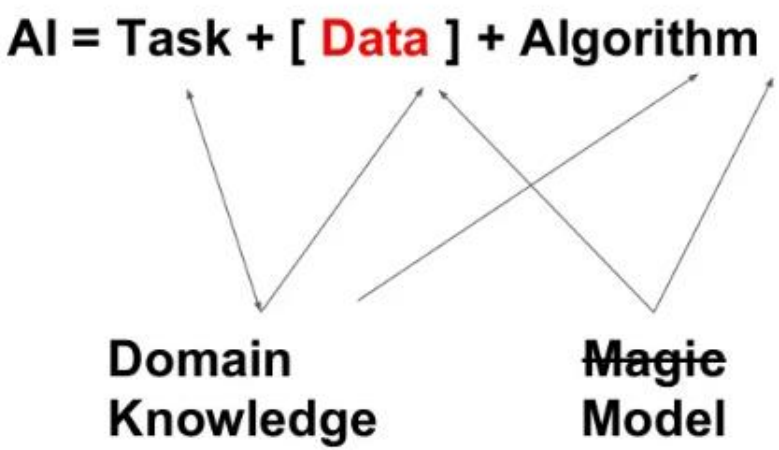
What is ML?

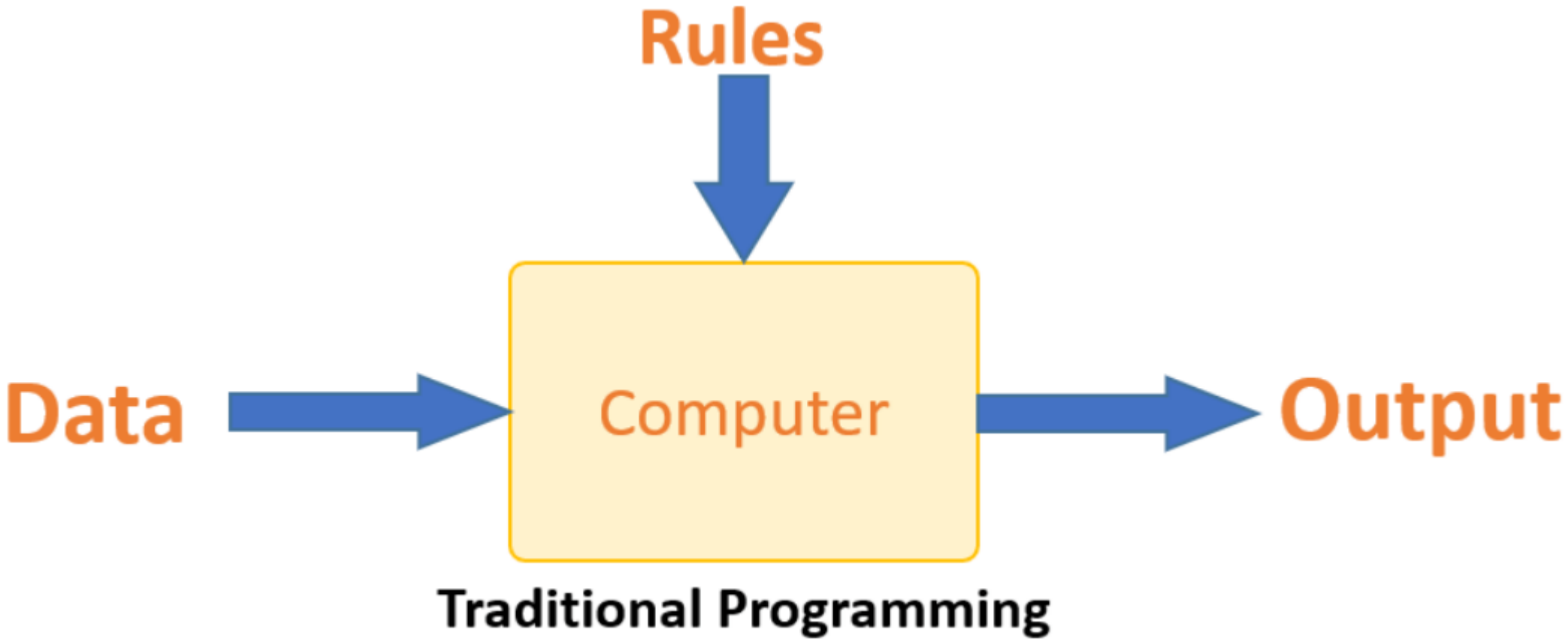
- *"Machine Learning at its most basic is the practice of using algorithms to parse data, learn from it, and then make a determination or prediction about something in the world." — **Nvidia***



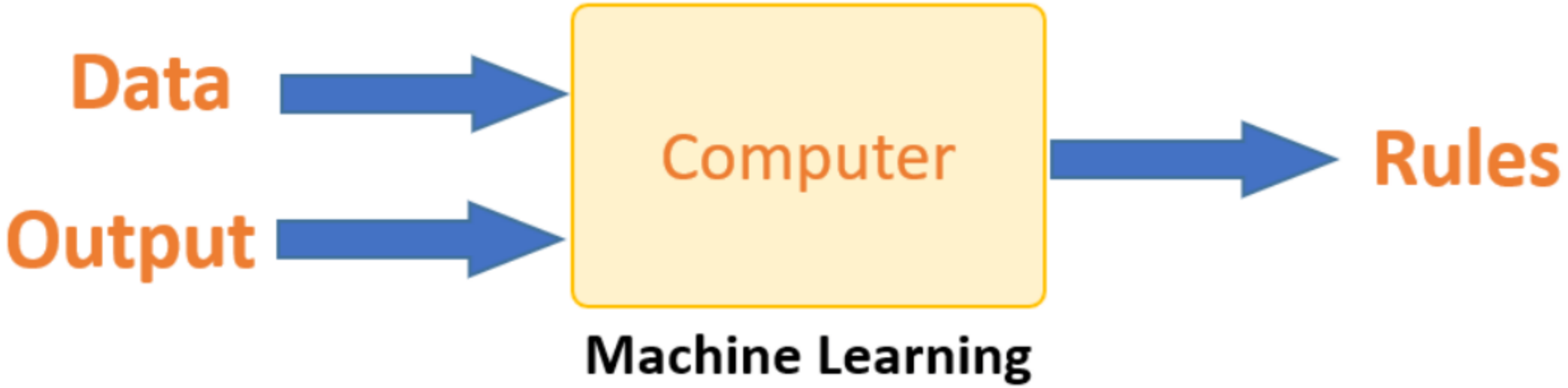
Machine learning is not magic. You need to have a question appropriate data (and lots of it), data context, training, and monitoring.

Artificial Intelligence (AI) and Machine Learning (ML)



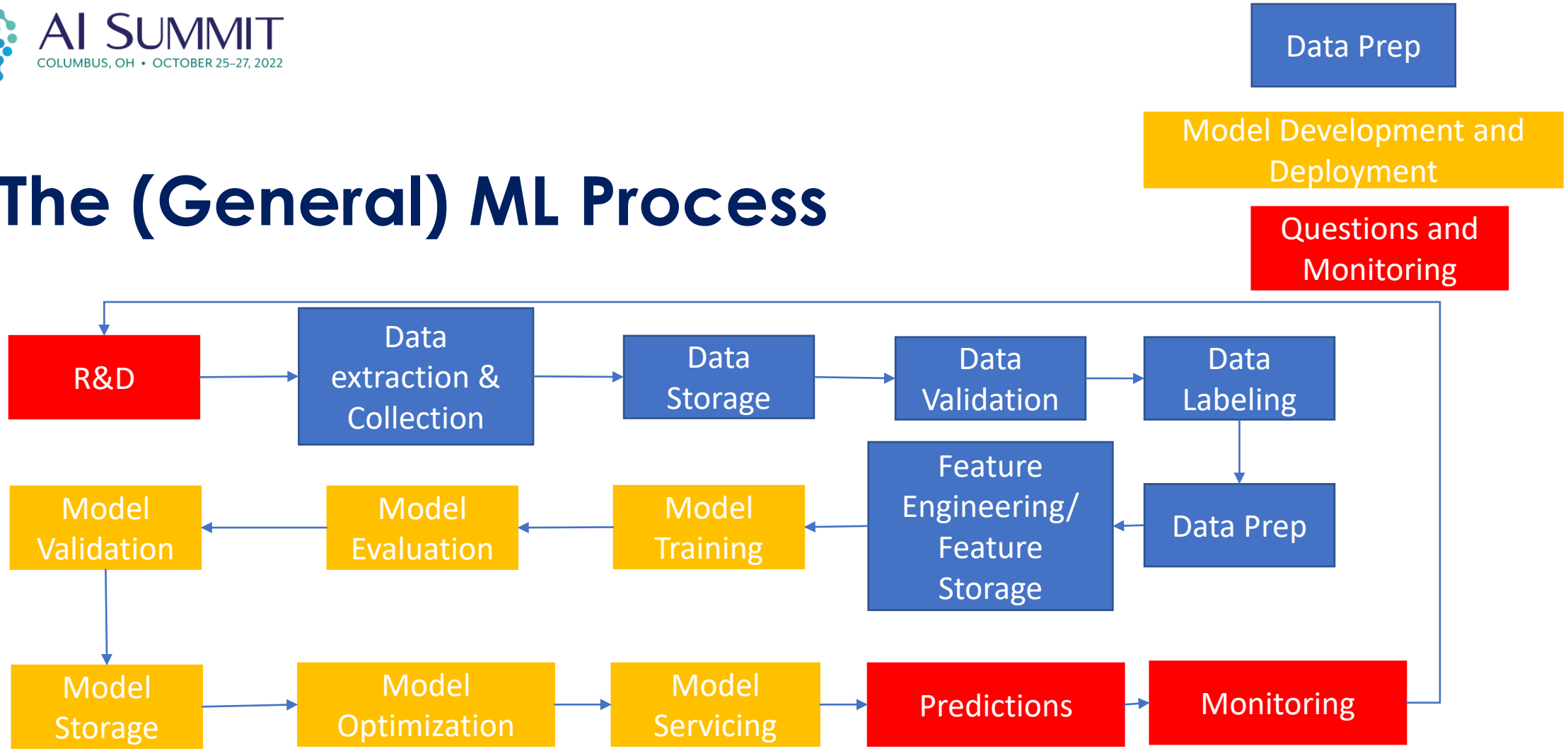


*Guru99

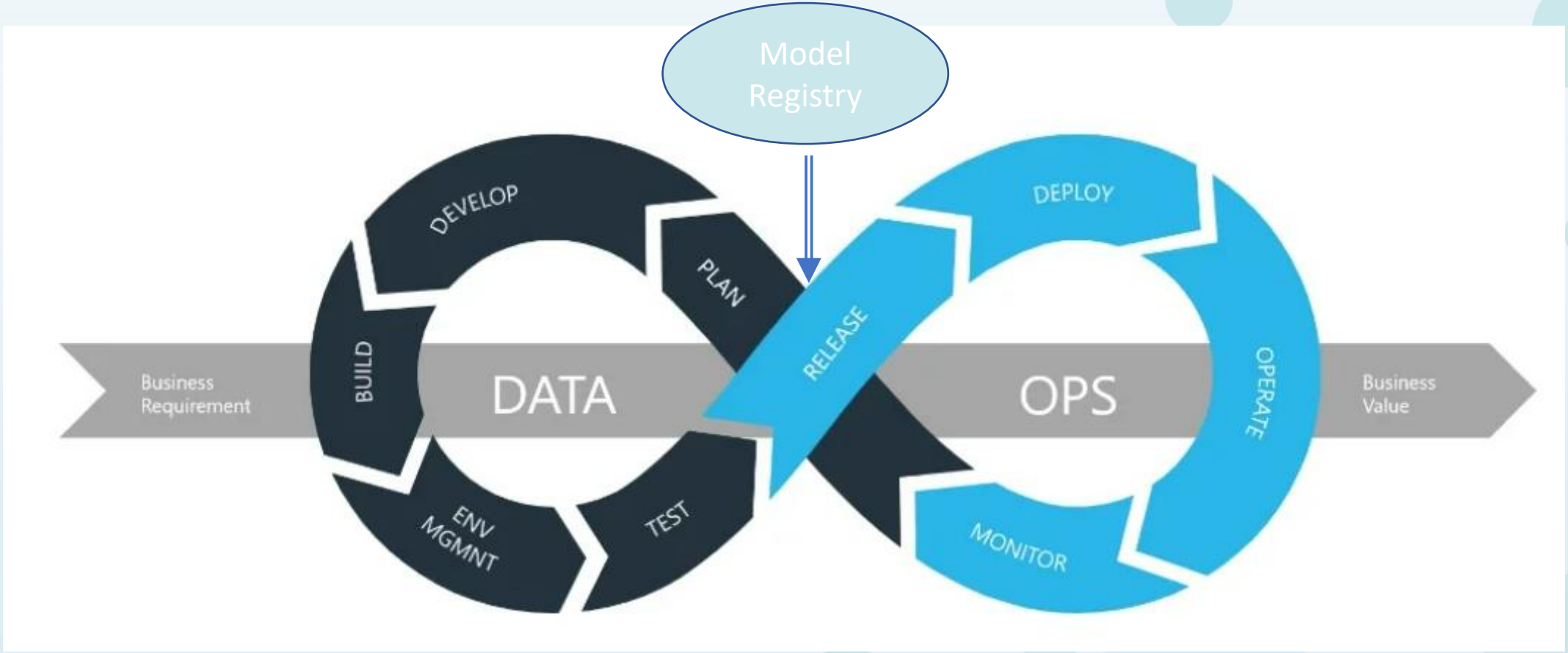


*Guru99

The (General) ML Process



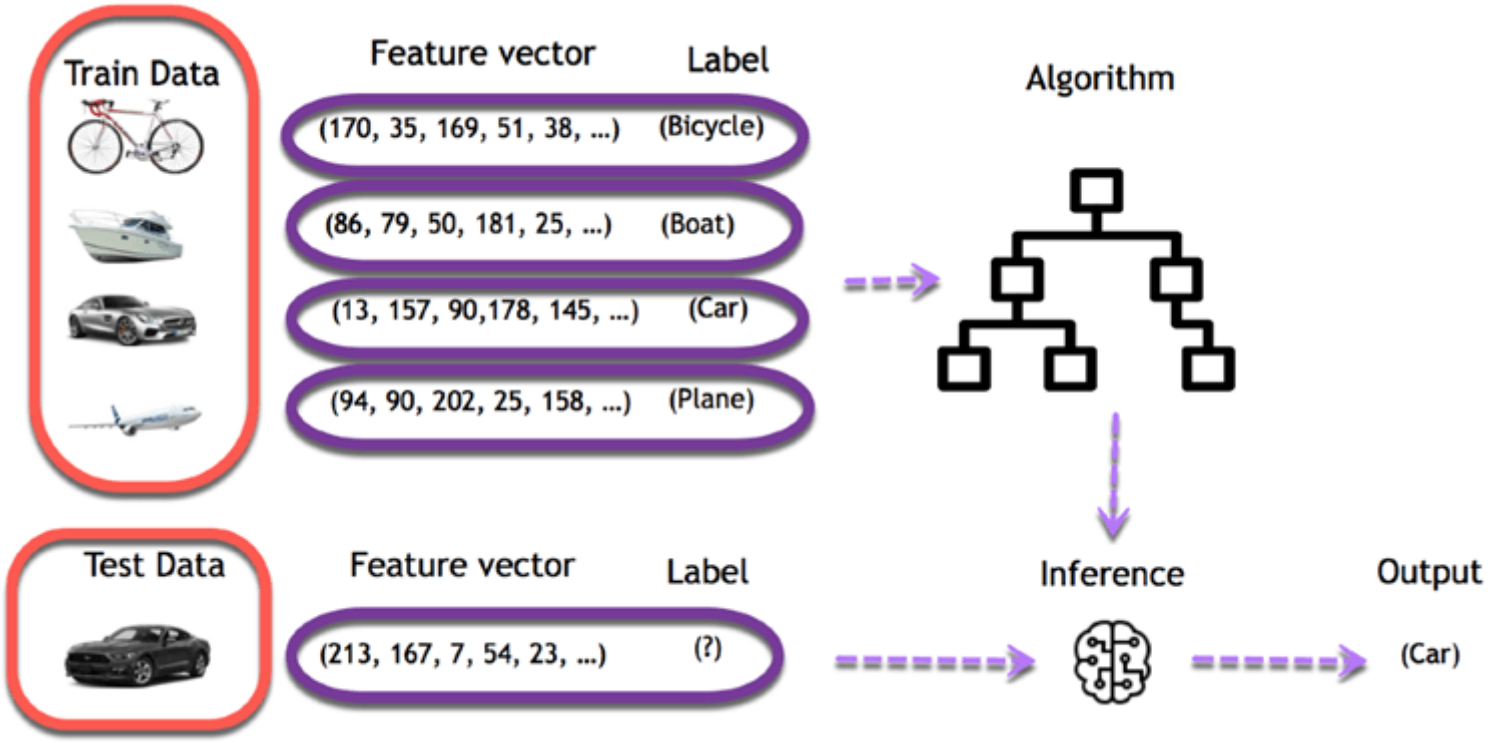
Quality of the Data impacts the Quality of the Model


Model
Registry
DEVELOP
PLAN
BUILD
ENV
MGMNT
TEST
RELEASE
DEPLOY
OPERATE
MONITOR
Business
Requirement
DATA
OPS
Business
Value

Basic ML Building Blocks

- **Data**: This refers to observations of real-world phenomena.
- **Features**: These are groups of data points which can be used to train machine learning algorithms. The are descriptive.
- **Label/Target:** This is what the model is trying to predict. e.g. (whether the animal is a dog or a cat).
- **Classifier**: A classifier is an ML algorithm which uses the features of an object to try identify the class it belongs to.

Area (m2)	No. of Stories	Distance to city (km)	Price (LKR)
500	1	5	500,000.00
525	2	5	1,000,000.00



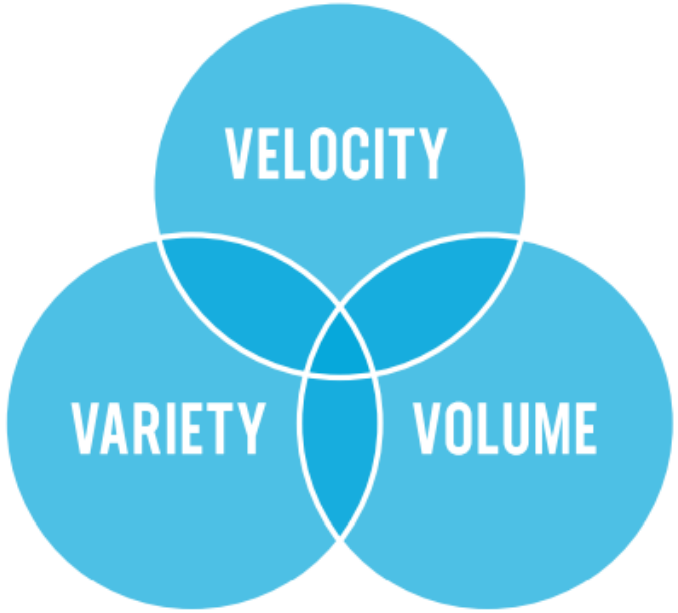
*Guru99 and Nvidia

Data

The "Vs" of Data

- Volume: How much data is there? If you have too little data, the application will not be able to robustly learn.
- Velocity: How quickly is the data being created? If you have too little data, how long will it be before you have an adequate amount of data? Is the data no longer relevant?
- Velocity Δ: Is the velocity of data creation accelerating or decelerating? Are the changes foreseeable?
- Variety: How many data sources are there? Does this application rely on only one source of information? Is the data adequately representative of the intended use or user group? If there are different data sources, what are the unique characteristics of those sources. For example, although patient data may be pulled from multiple hospital locations, are those hospitals part of the same network?
- Veracity: Why do you think you can trust the data? It is not uncommon to spend a significant amount of time cleaning up data before it is used.
- Validity: Are the data values correct? Are they timely? Describe the protocol used to collect this data, explaining 'When', 'How' and 'by Whom' data was gathered
- Viability: How is the data relevant to the use case?
- Volatility: How often does the data change? Describe how long is it relevant, how long to store the data before archiving or deleting.

The traditional 3 V's of Big Data



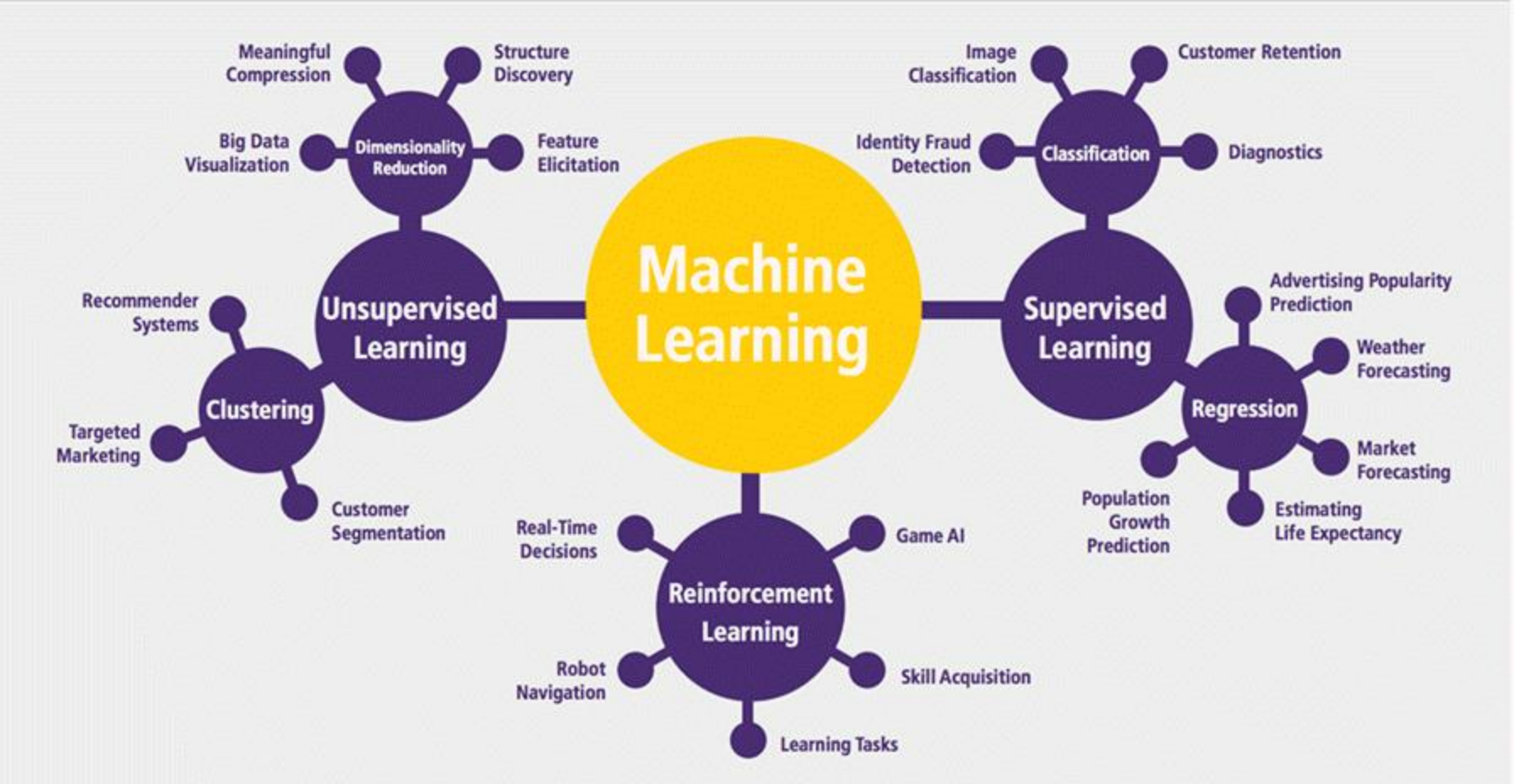
Data Structure

- **Structured data** as data that resides in in a fixed field within a record or file
- **Unstructured data** as information that doesn't reside in a traditional row-column database.
  - Examples of this can be from diagnostic images, hand-written triage notes, e-mail documents, word processing documents, PDF, PPTs, videos, photos, audio files, blogs and more.
- **Semi-structured data** refers to data that is partially organized by tags and markers in a fashion that is accessible by ML analytic tools.
  - Examples of this include XML documents, Word metadata files, e-mail sending and receiving data, tags on photos, and NoSQL.

"Data directly contributes to the performance of the algorithm and should be carefully managed and controlled. The quality and diversity of the data used to train or retrain an AI application should be subjected to proper data management, control and governance."

-GMLP Team, 2019

Learning/ Training Types



*Guru99 and Nvidia

Supervised Learning

- Supervised learning happens when we use a data set ***with labels*** to train the model. These labels or desired outputs are also being known as "***Supervisory Signal***" in the machine learning world since it supervise the model, this kind of output is expected when these kind of inputs are given.
  - Ex: House prices example. Those were past house prices and their features to train the model. Therefore, it belongs to Supervised Learning.
- **Algorithms used in Supervised Learning:**
  - Linear Regression
  - Logistic Regression
  - Support Vector Machine (SVM)
  - K-Nearest Neighbor (KNN)
  - Decision Trees
  - Naive Bayes

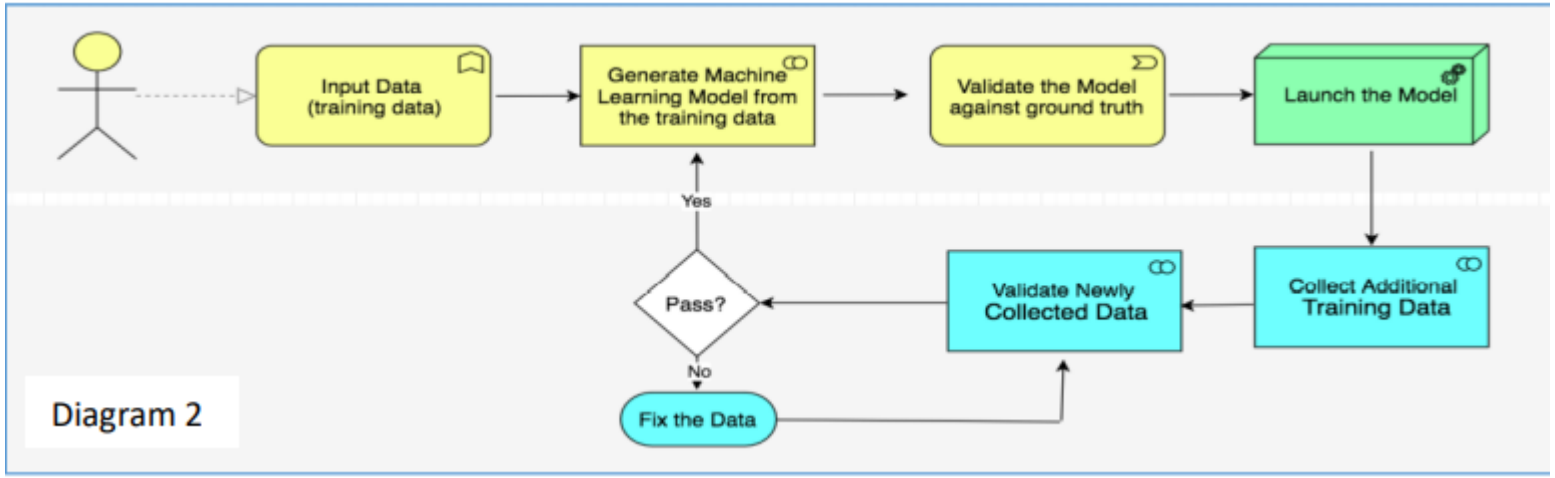
Unsupervised

- Unsupervised Learning happen when we are using a data set ***without labels*** to train a model. It is up to the model to identify commonalities of data and presence or absence of those commonalities to derive at classes.
  - Example: Biological grouping or clustering of species groups based on genetic markers.
- Unsupervised learning does not have a target or response variable, but rather finds patterns or relationships between the inputs.
- **Algorithms used in Unsupervised Learning:**
  - Clustering
  - Anomaly Detection
  - Neural Networks
  - Expectation Maximization Algorithm (EM)
  - Principle Component Analysis (PCA)
  - Independent Component Analysis

Self-supervised learning

- A major hurdle of supervised machine learning is acquiring enough labelled data on which to train a ML model.
- Self-supervised learning is a technique through which an algorithm creates labels from the data itself, without human supervision.
- Self-supervised learning can be used for image and text generation tasks – for example, by removing elements of a sentence or picture and training the AI to fill in the blanks

Continuously Learning



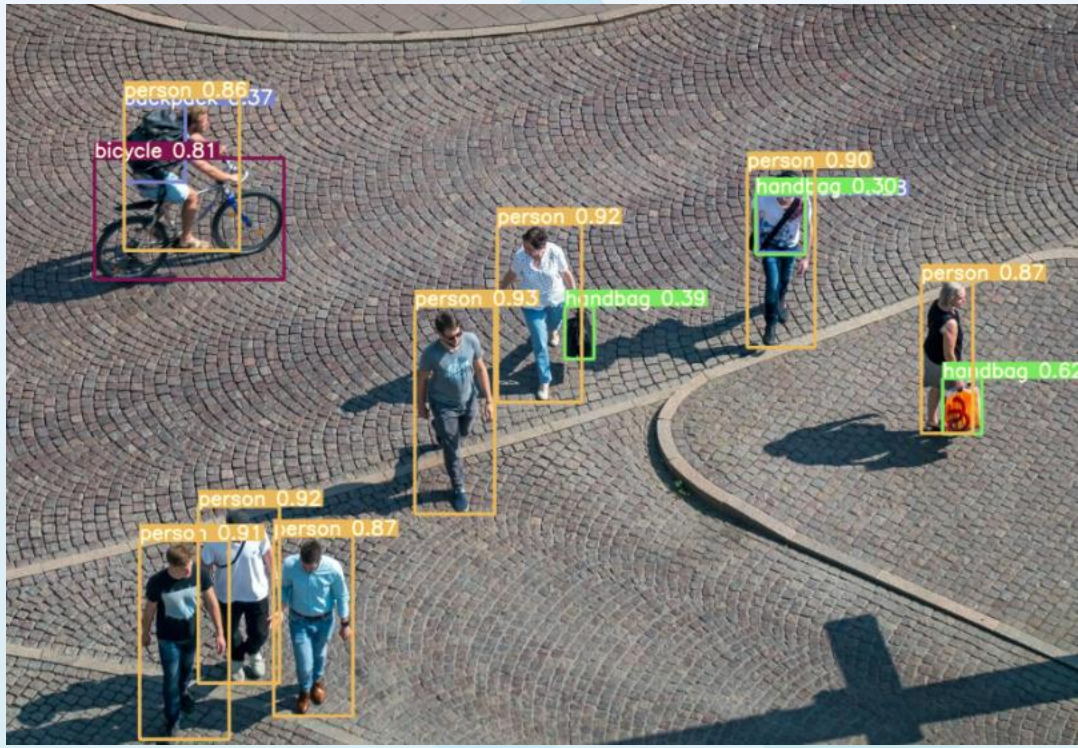
Algorithm keeps learning as humans do.

Human-in-the-loop AI

- We would expect students to ask questions to clarify understanding during training sessions – training an AI system is no different.
- Human-in-the-loop AI systems siphon off portions of validation data for human review, especially where prediction confidence is low or prediction error is high. During development, the AI system can receive targeted feedback (additional labelled data) on which to continue training and in a live environment can defer marginal predictions to a human for manual consideration
- Where it needs extra help, you ensure it remains flexible to deal with uncertainty, has the ability to adapt quickly to new scenarios and finds blind-spots in prediction capability.

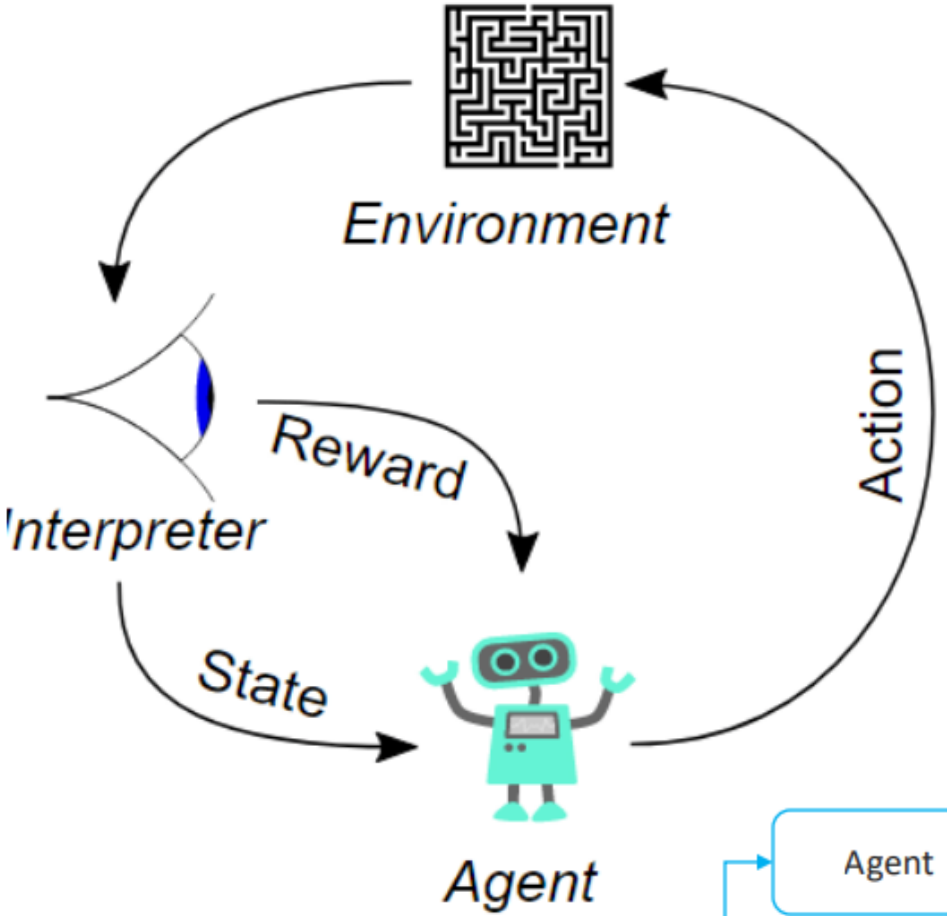
Deep Learning

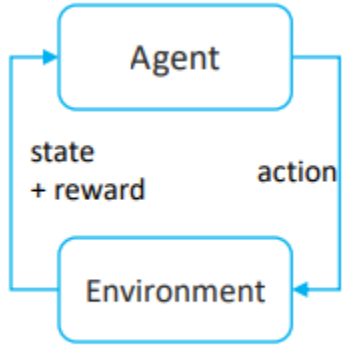
- A model that is used to perform tasks based upon layers of algorithms and massive amounts of data.



Reinforcement Learning

- Think of an infant.
- He does not know that fire is dangerous. He touches it and burn himself. But he also learn fire can burn him. Next time he only take hand near to fire, but does not touch. Still the heat burn him. Then he understand even too close is not good with regards to fire. If we relate this to Machine Learning, our system is like this infant.
- It is called as “Agent”. The fire is the “Environment” and the feeling of burning is similar to “Feedback”





The RL interaction loop

Creative AI/ ‘generative modelling’ (GANs)

- Models can be trained to output completely new images that are indistinguishable from those in the training set.
- GANs consist of two competing models that are trained in parallel - the generator (that creates new images), and the discriminator (that tries to guess if the image is real or fake)

Modeling

What is a Model?

- AI Model: An algorithm which is based on a dataset through which it can arrive at a decision. It should be built to recognize patterns, make conclusions, and predictions.
- ML Model: is narrower than an AI model as it could be said to be used to make predictions.
- DL Model: subfield of machine learning concerned with algorithms inspired by the structure and function of the brain called artificial neural networks

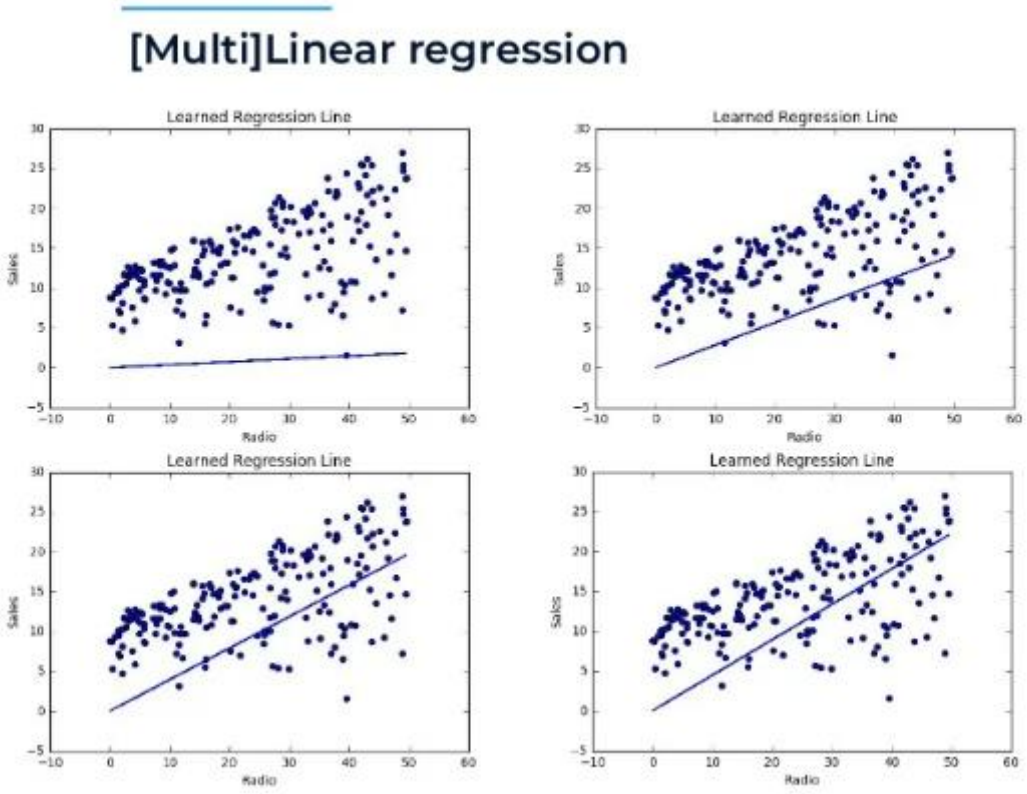
Artifact

- Term that is used to describe the output created by the training process.
- Output could be a fully trained model, a model checkpoint, or a file created during the training process.

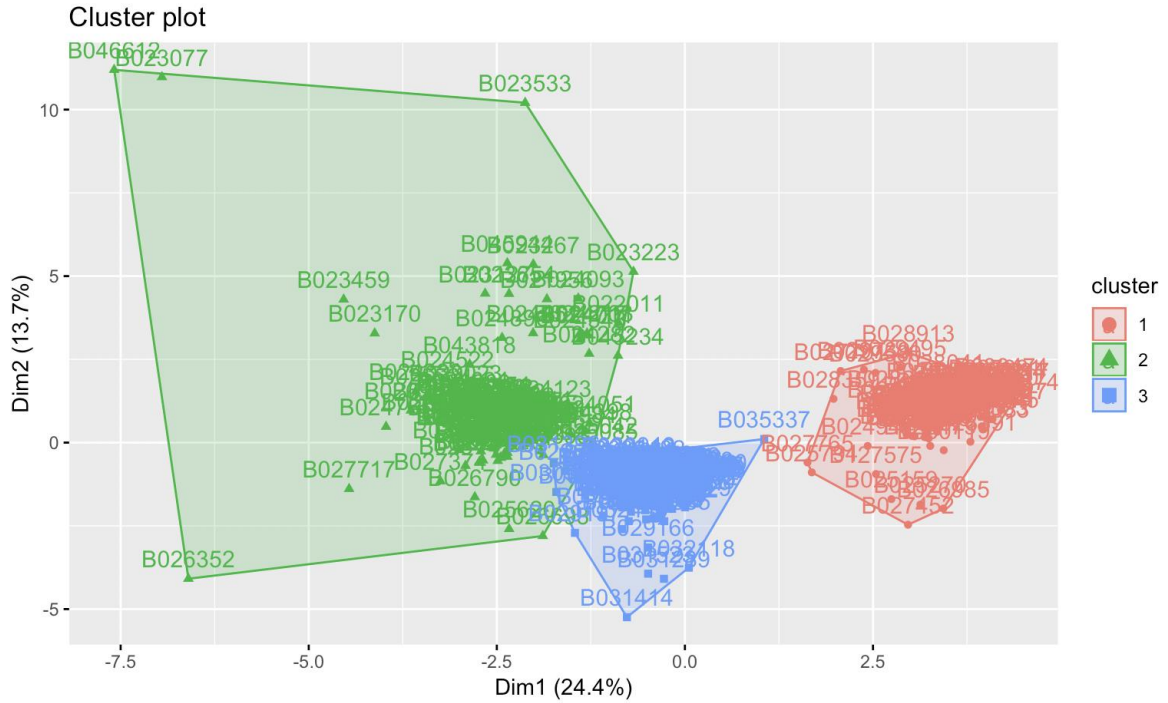
Tasks/ Algorithms

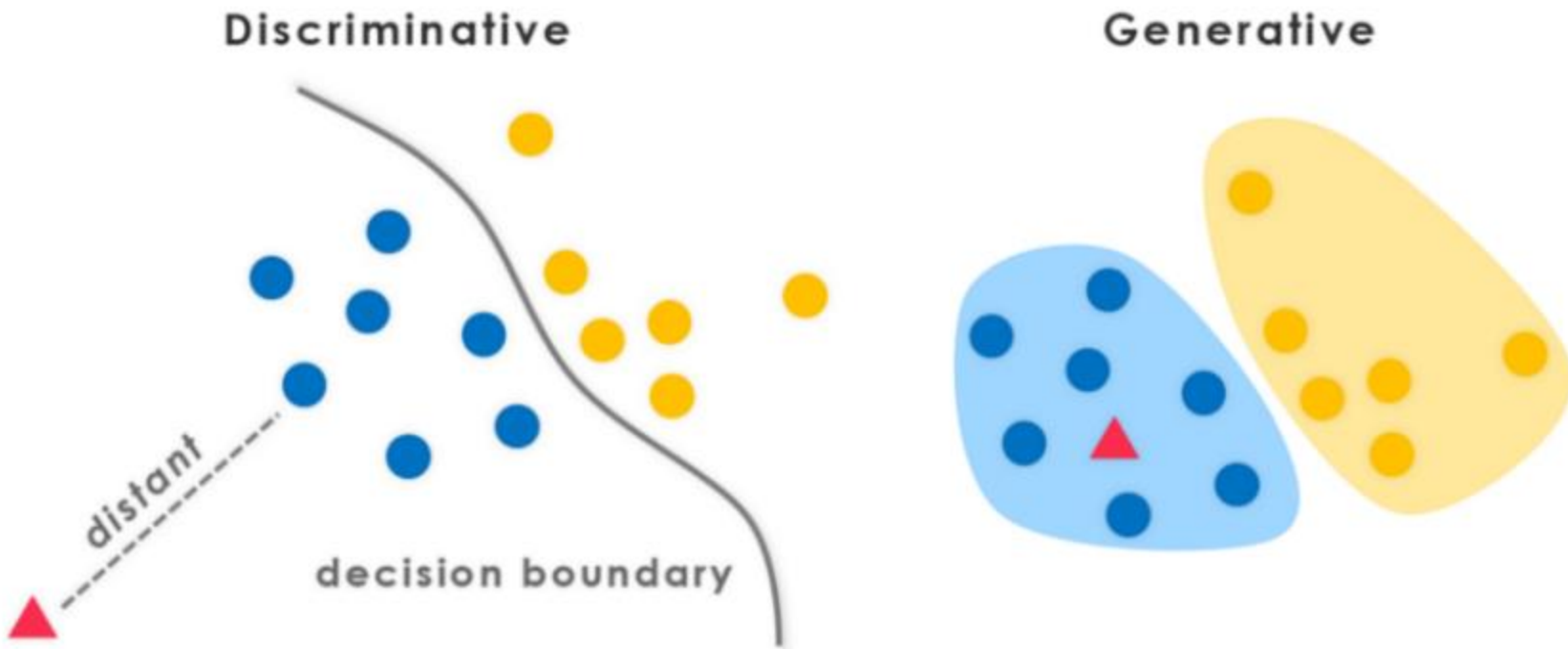
Linear Regression

- Linear regression attempts to define the relationship between multiple variables by fitting a linear equation to a dataset. The output of a linear regression model can then be used to estimate the value of missing points in the dataset.



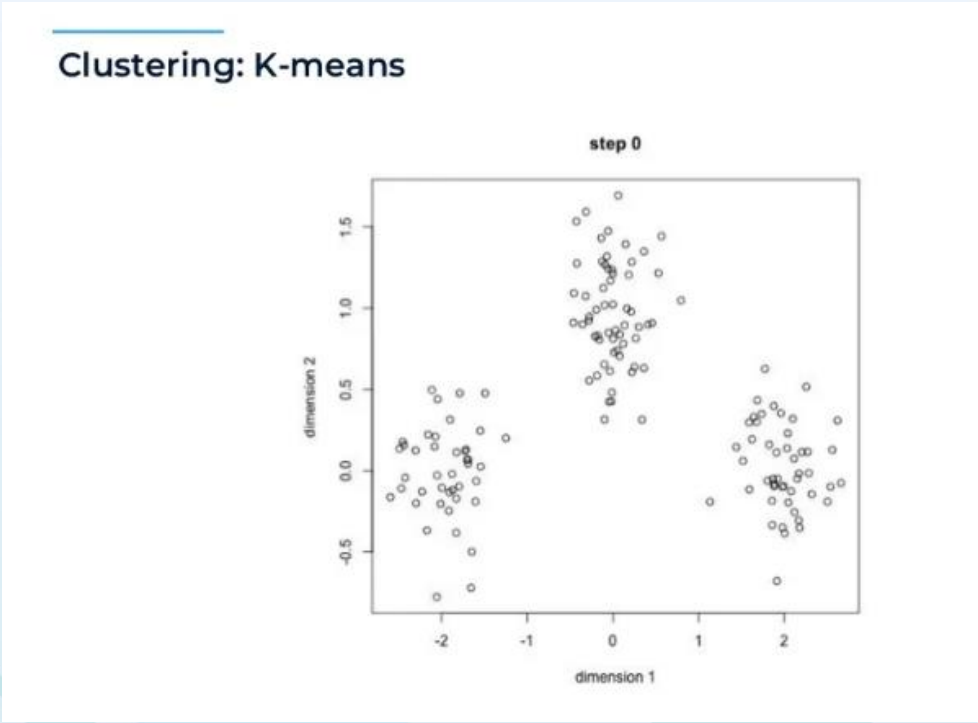
Clustering….


Cluster plot
B046612
B023077
B023533
B023223
B023459
B023170
B022011
B043818
B028913
B035337
B027717
B026790
B026352
B029766
B031414
cluster
1
2
3
Dim2 (13.7%)
10
5
0
-5
-7.5
-5.0
-2.5
0.0
2.5
Dim1 (24.4%)


Discriminative
distant
decision boundary
Generative

K Means

- The k-means algorithm is used to separate a dataset into k different clusters (where k is some integer).
- Start by randomly choosing k points (called centroids) in space, and assigning each point to the closest centroid.
- Next, calculate the mean of all the points that have been assigned to the same centroid. This mean value then becomes the cluster's new centroid. We repeat the algorithm until it converges, i.e. the position of the centroids does not change.



K nearest

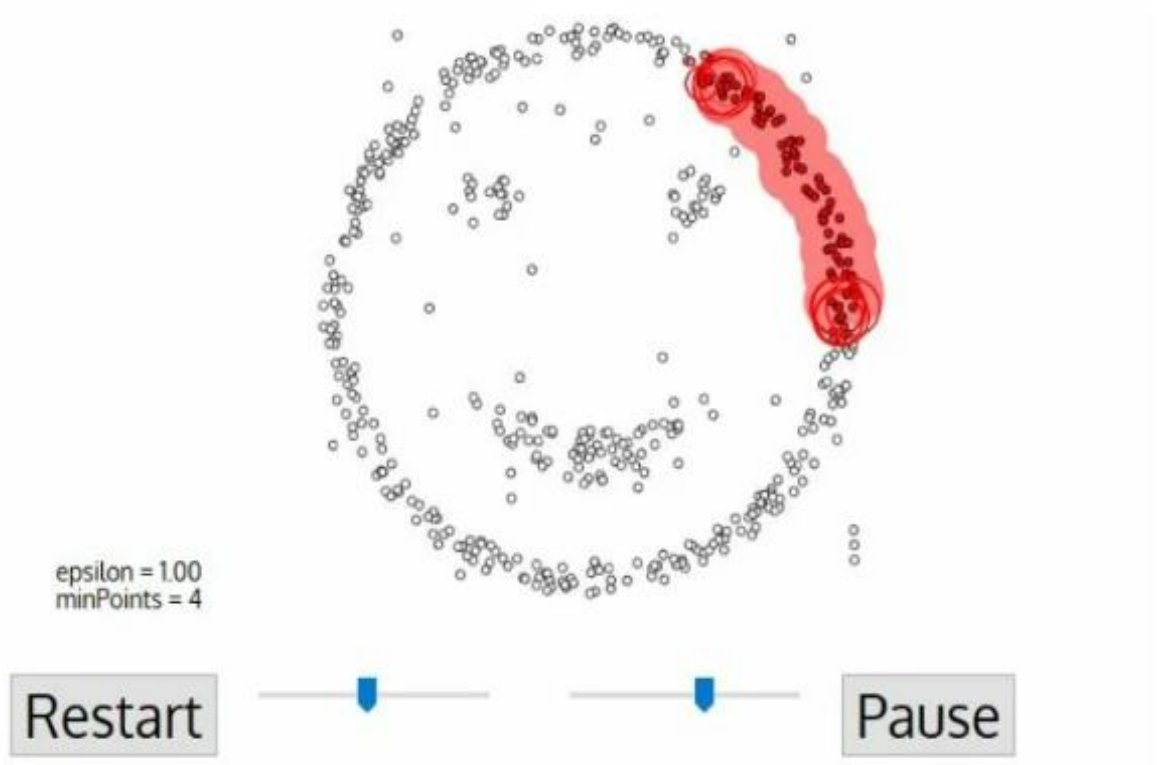
- k-nearest neighbors: The k-nearest neighbors algorithm is used to classify data points based on the classification of their k nearest neighbors (where k is some integer).
- For example, if we have k = 5, then for each new data point, we will give it the same classification as the majority(or the plurality) of its closest neighbors in the data set.

Hierarchy Clustering

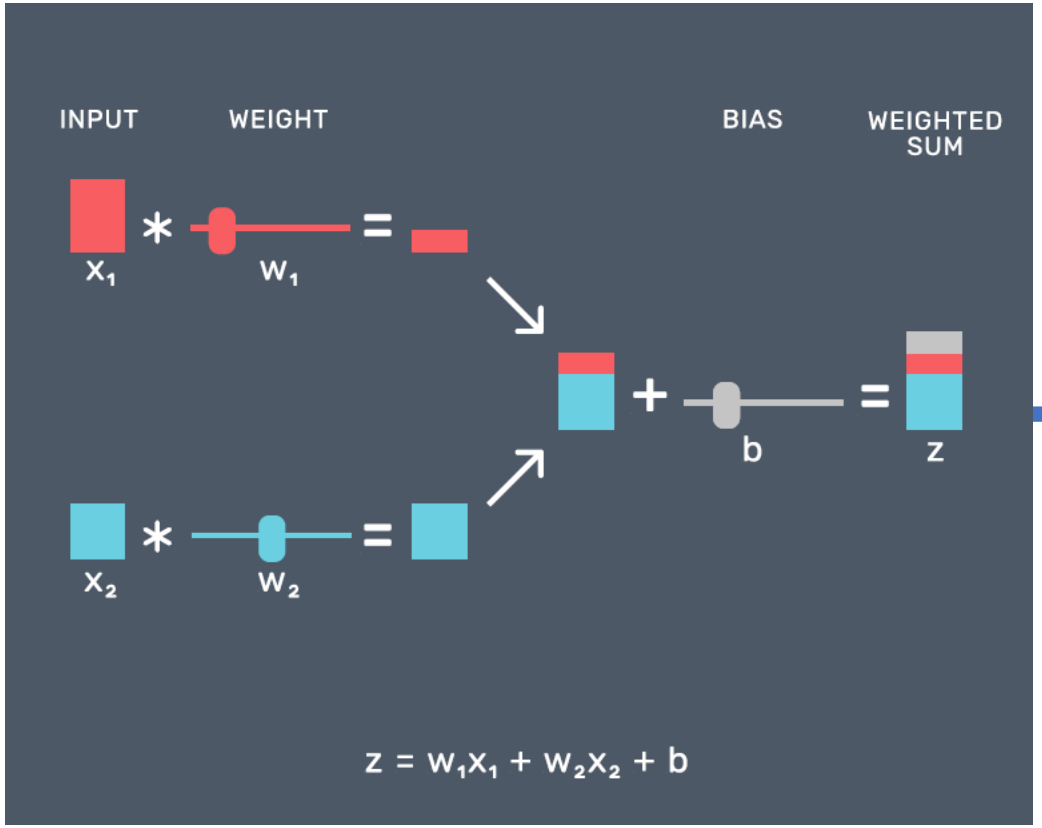
- Agglomerative Clustering (involving decomposition of cluster using bottom-up strategy)
- Divisive Clustering (involving decomposition of cluster using top-down strategy)

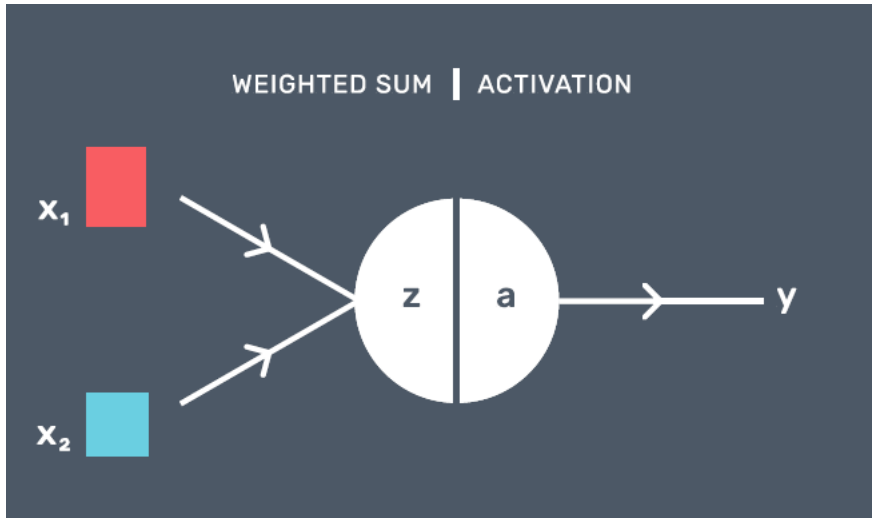
DBSCAN

- Used to discover clusters of different shapes and sizes from a large amount of data, which is containing noise and outliers.

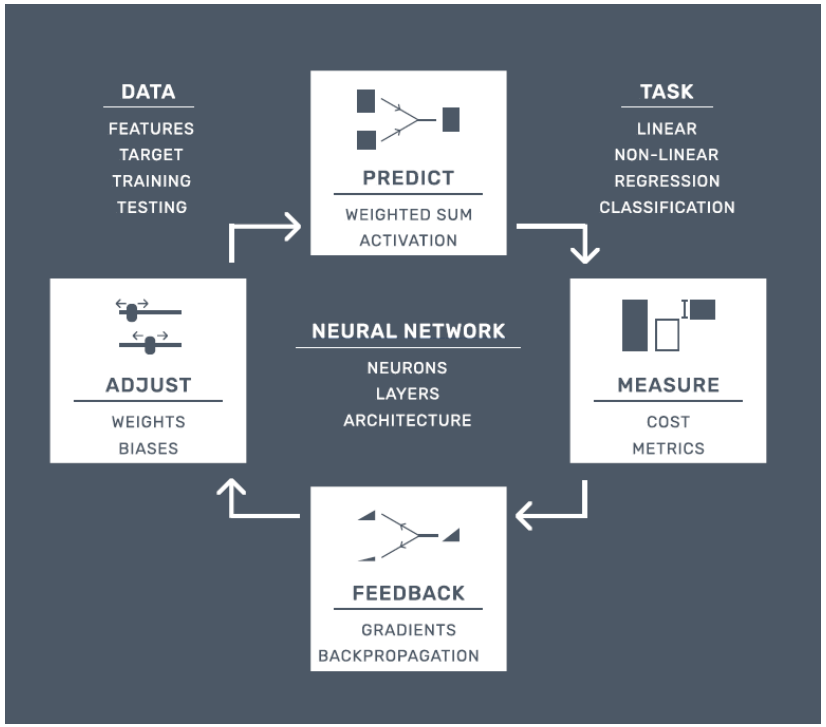


Neural Networks

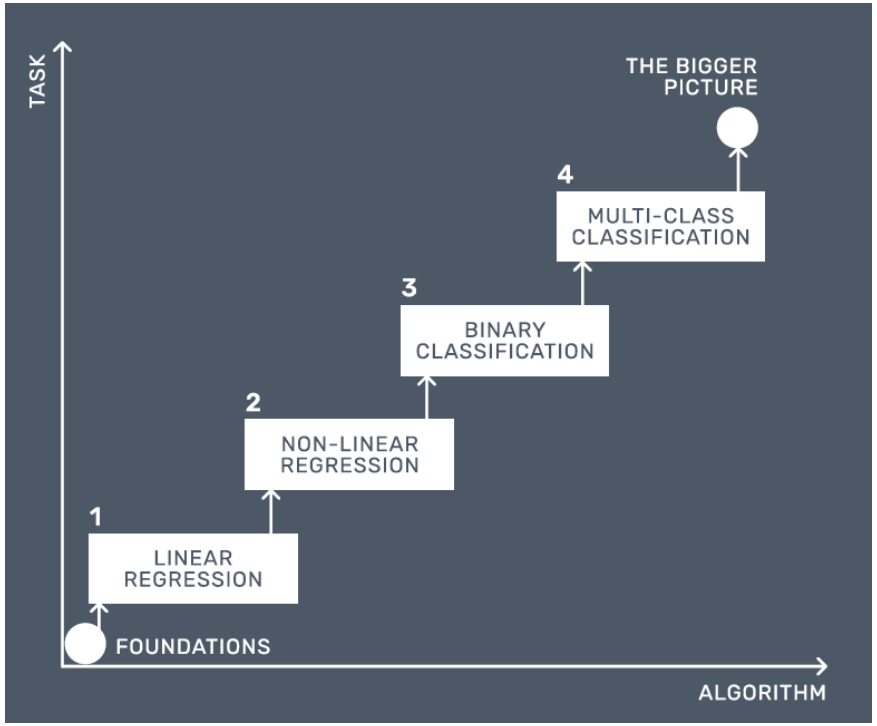


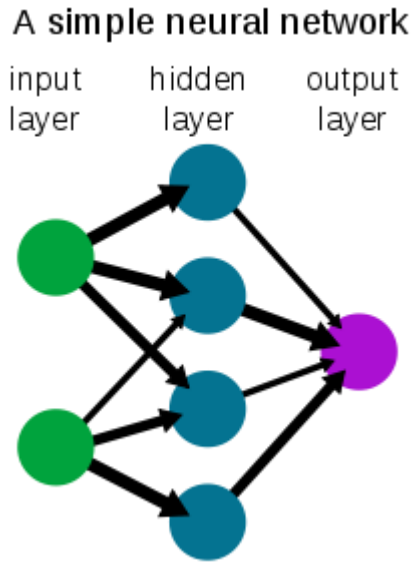


Neural Networks



Neural Networks





Ensemble Learning

- For both Supervised and Unsupervised learning
- Layering of algorithms to get better results
- Use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone
- May be more efficient at improving overall accuracy for the same increase in compute, storage, or communication resources by using that increase on two or more methods, than would have been improved by increasing resource use for a single method.
- Fast algorithms such as decision trees are commonly used in ensemble methods (for example, random forests), although slower algorithms can benefit from ensemble techniques as well.

Boosting Algorithms

- Type of Ensemble Learning
- It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees.
- When a decision tree is the weak learner, the resulting algorithm is called gradient-boosted trees; it usually outperforms random forest.
- A gradient-boosted trees model is built in a stage-wise fashion as in other boosting methods, but it generalizes the other methods by allowing optimization of an arbitrary differentiable loss function.

Training, Testing, and Validation

Training, Testing, and Validation

- **Training Dataset**: The sample of data used to fit the model.
- **Validation Dataset**: The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters. The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration.
- **Test Dataset**: The sample of data used to provide an unbiased evaluation of a final model fit on the training dataset.

Training

- Learning algorithms go through a period of training, cross validation, and testing before being placed in use.
- Like all statistical models, performance depends on how well the data set used for training is representative of the actual environment of use. During use, the ML algorithm collects additional data, which can be collated and used (offline) to repeat the original cycle of testing and validation. The original ML algorithm can then be replaced with the “new” algorithm with improved performance. This is sometimes called batch learning.

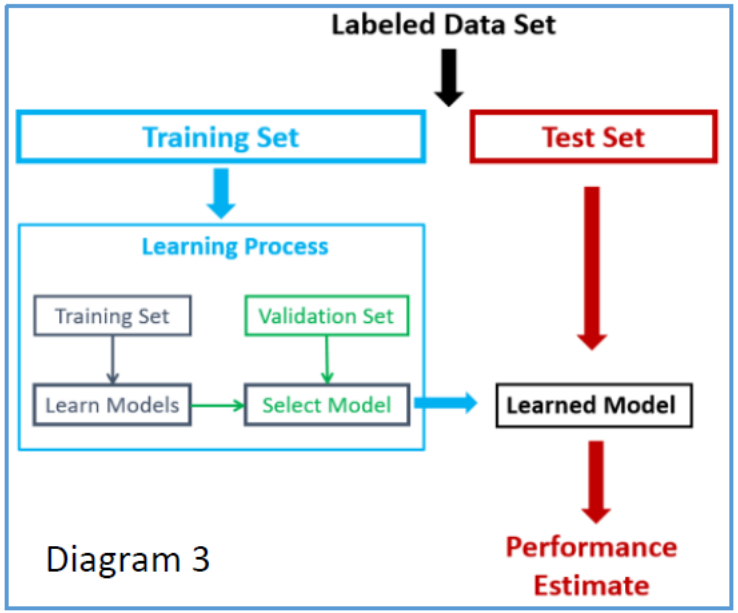


Diagram 3

- Perspectives and Good Practices for AI and Continuously Learning Systems in Healthcare, GMLP Team, 2018

Cross Validation

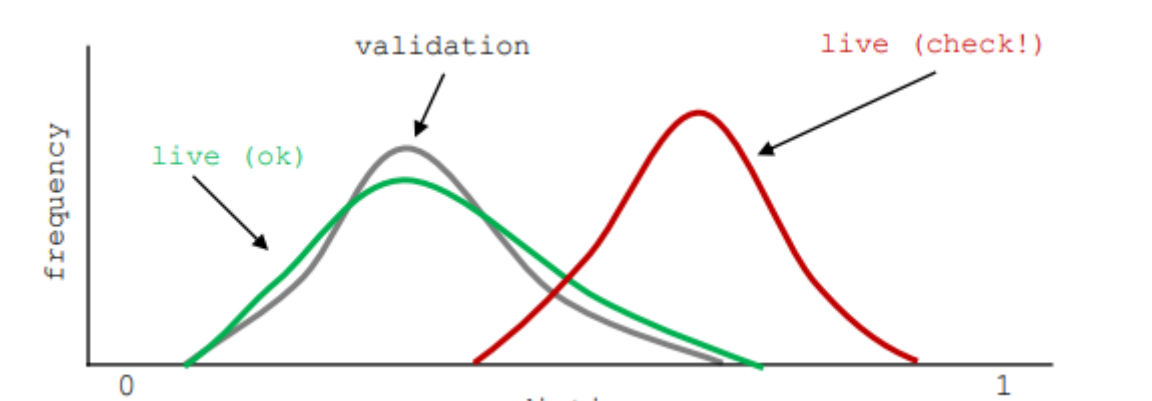
- Limits sampling bias
- Prevents overfitting in a predictive model where the amount of data may be limited
- Types:
  - Non-Exhaustive Methods
  - Hold Out method
  - K fold
  - More

Epochs, Batch Size, & Iterations

- In many cases, it is not possible to feed all the training data into an algorithm in one pass due to the size of the dataset and memory limitations.
- **Epoch** elapses when an entire dataset is passed forward and backward through the neural network exactly one time. If the entire dataset cannot be passed into the algorithm at once, it must be divided into mini-batches.
- **Batch size** is the total number of training samples present in a single min-batch.
- An **iteration** is a single gradient update (update of the model's weights) during training. The number of iterations is equivalent to the number of batches needed to complete one epoch.
- So, if a dataset includes 1,000 images split into mini-batches of 100 images, it will take 10 iterations to complete a single epoch.

Prediction validation

- The distribution of predictions in the live test set should approximately match the distribution of predictions in the validation set, as shown below. If this isn't the case, then the data used to train the model isn't representative of the new data being predicted in the live environment and the predictions may therefore not be accurate.



Evaluation

How do you know that your model is performing well?

- Evaluation Metrics:
  - Accuracy & Loss
  - Precision
  - Confusion Matrix
  - AUC (Area Under ROC curve)
  - Mean Absolute Error (MAE)
  - Root Mean Square Error (RMSE)
  - R Square
  - F1 Score
  - Recall & Sensitivity
  - Specificity
  - Within-group sum of squares (WGSS)

AUC (Area under the ROC Curve)

- Very important
- It is a performance measurement for a classification problem at various thresholds settings.
- The ROC Curve measures how accurately the model can distinguish between two things (e.g. determine if the subject of an image is a dog or a cat).
- AUC measures the entire two-dimensional area underneath the ROC curve.
- This score gives us a good idea of how well the classifier will perform.

Accuracy and Loss

- Accuracy
  - Accuracy is the count of predictions where the predicted value is equal to the true value.
  - It is binary (true/false) for a particular sample.
  - Communicated in %.
- Loss
  - A loss function (cost function) takes into account the probabilities or uncertainty of a prediction based on how much the prediction varies from the true value.
  - Communicated as a summation of the errors made for each sample in training or validation sets.
  - Loss is often used in the training process to find the "best" parameter values for the model
  - Ex: Log Loss, Cross-Entropy Loss, Likelihood loss and mean square error

Confusion Matrix

- Describes the performance of a classification model (or "classifier").
- By table, compares predicted and actual values.
- The basic components of the table are as follows:
  - True positives (TP): The prediction was yes, and the true value is yes
  - True negatives (TN): The prediction was no, and the true value is no
  - False positives (FP): The prediction was yes, but the true value was no
  - False negatives (FN): The prediction was no, but the the true value is yes

Confusion Matrix Supports other Evaluators

Metric	Formula	Definition
Accuracy	(TP+TN)/(TP+TN+FP+FN)	Percentage of total items classified correctly
Precision	TP/(TP+FP)	How accurate the positive predictions are
Recall/Sensitivity	TP/(TP+FN)	True positive rate (eg to asses false positive rate)
Specificity	TN/(TN+FP)	True negative rate (eg to assess false negative rate)
F1 score	2TP/(2TP+FP+FN)	A weighted average of precision and recall

Challenges

Alert

- Scripts deployed through the platform that can be used to alert data scientists and engineers to workflows that need attention.
- For example, if the latest predictions do not match the distribution of past predictions, then rather than overwriting the latest values, the process can terminate early and send a process report to the engineer in charge of the workflow for analysis.

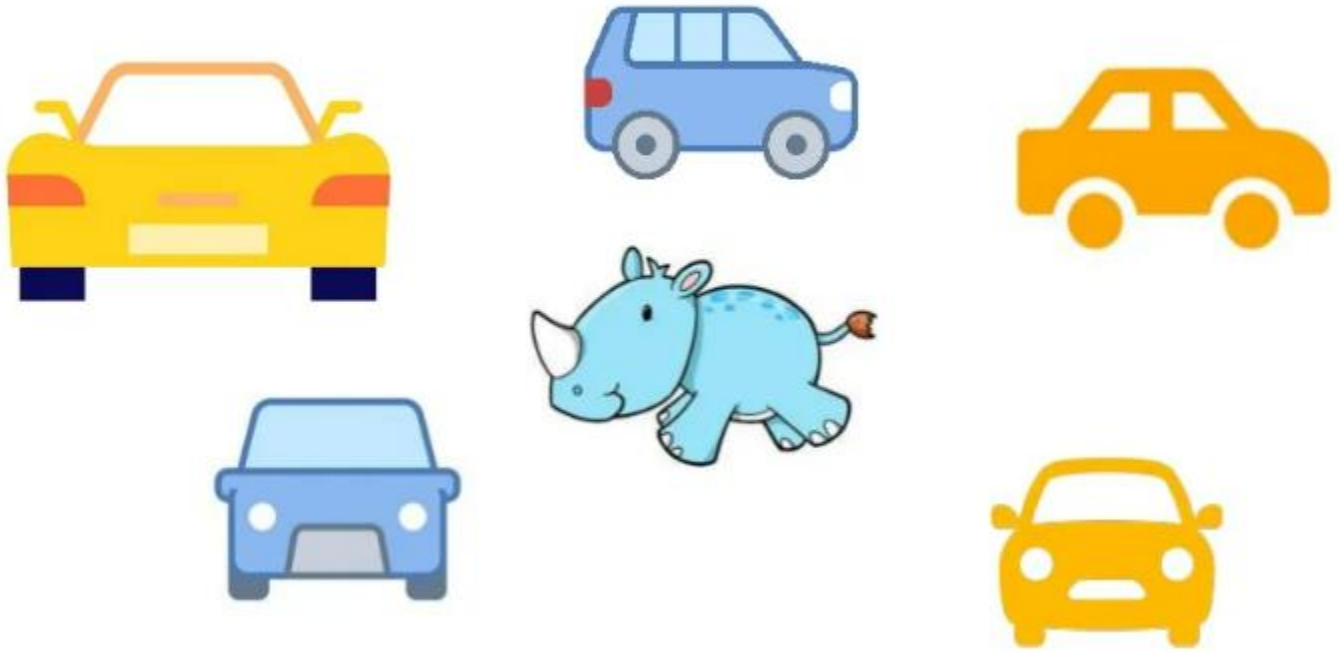
Data

- Missing data
- Distribution
- Drift

Overfitting and Underfitting

- Overfitting means that your model makes not accurate predictions. In this case, train error is very small and validation/test error is large.
- Underfitting means that your model makes accurate, but initially incorrect predictions. In this case, train error is large and validation/test error is large.

Anomaly Detection



Concept Drift

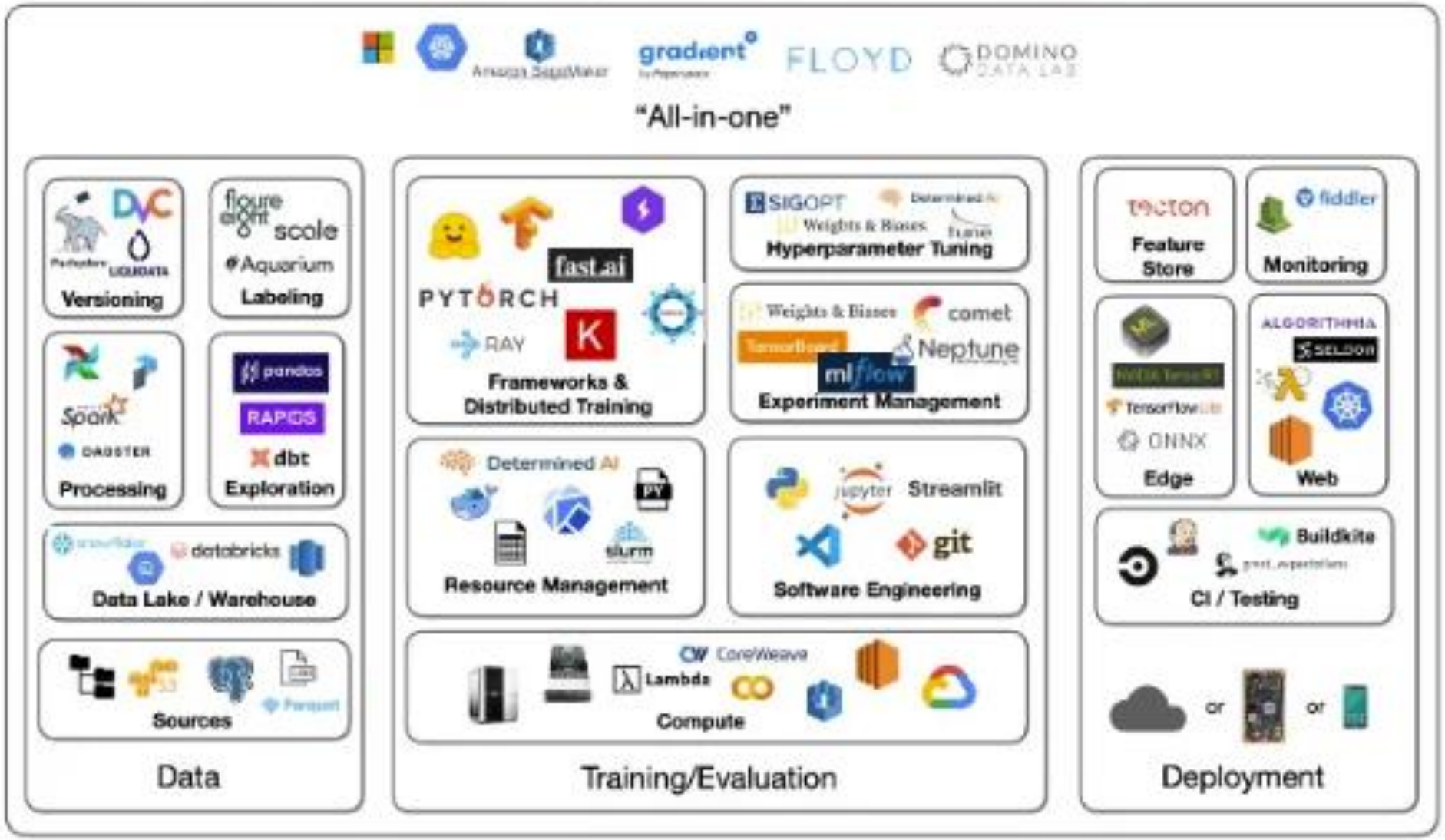
Bias in Development

- In the development process inadvertent bias could be introduced by the project team members due to a lack of awareness of the issues of other groups and stakeholders that may not be directly represented in the demographics of the project team or the data (e.g., race, gender, comorbidities, etc.).
- Inadvertent bias could introduce risk to the user or service provider from bad (i.e. incomplete) data.
- Selection bias and exclusion bias
  - Different societal groups could be over or underrepresented in the data used for development.
- The bias introduced through lack of awareness of the issues of different user groups could also manifest as training bias

Bias in Data

- Data Bias
- Data bias can be introduced during the data gathering process.
- Over the lifecycle of product development bias could be introduced by product changes, software complexity, human resources, and inadequate management of data and programming risks3 .
- There is the potential for inadvertent incorporation of unconscious societal bias into the data set. The demographics associated with the source of the data set can potentially introduce bias. In defining the AI application consider which demographic the product serves, whether there is a population match in alignment with the intended application(s), and whether recalibration or compensation technique(s) were used to adapt input data to align correctly with the target demographic setting.

Platform & Pipeline


gradient
FLOYD
DOMINO DATA LAB
"All-in-one"
Versioning
Labeling
Aquarium
Processing
Exploration
dbt
Data Lake / Warehouse
databricks
Sources
Data
fast.ai
PYTORCH
RAY
Frameworks & Distributed Training
SIGOPT
Weights & Biases
Hyperparameter Tuning
comet
Neptune
mlflow
Experiment Management
Determined AI
slurm
Resource Management
jupyter
Streamlit
git
Software Engineering
CoreWeave
Lambda
Compute
Training/Evaluation
tecton
Feature Store
fiddler
Monitoring
TensorflowLite
ONNX
Edge
ALGORITHMIA
SELDON
Web
Buildkite
CI / Testing
or
or
Deployment

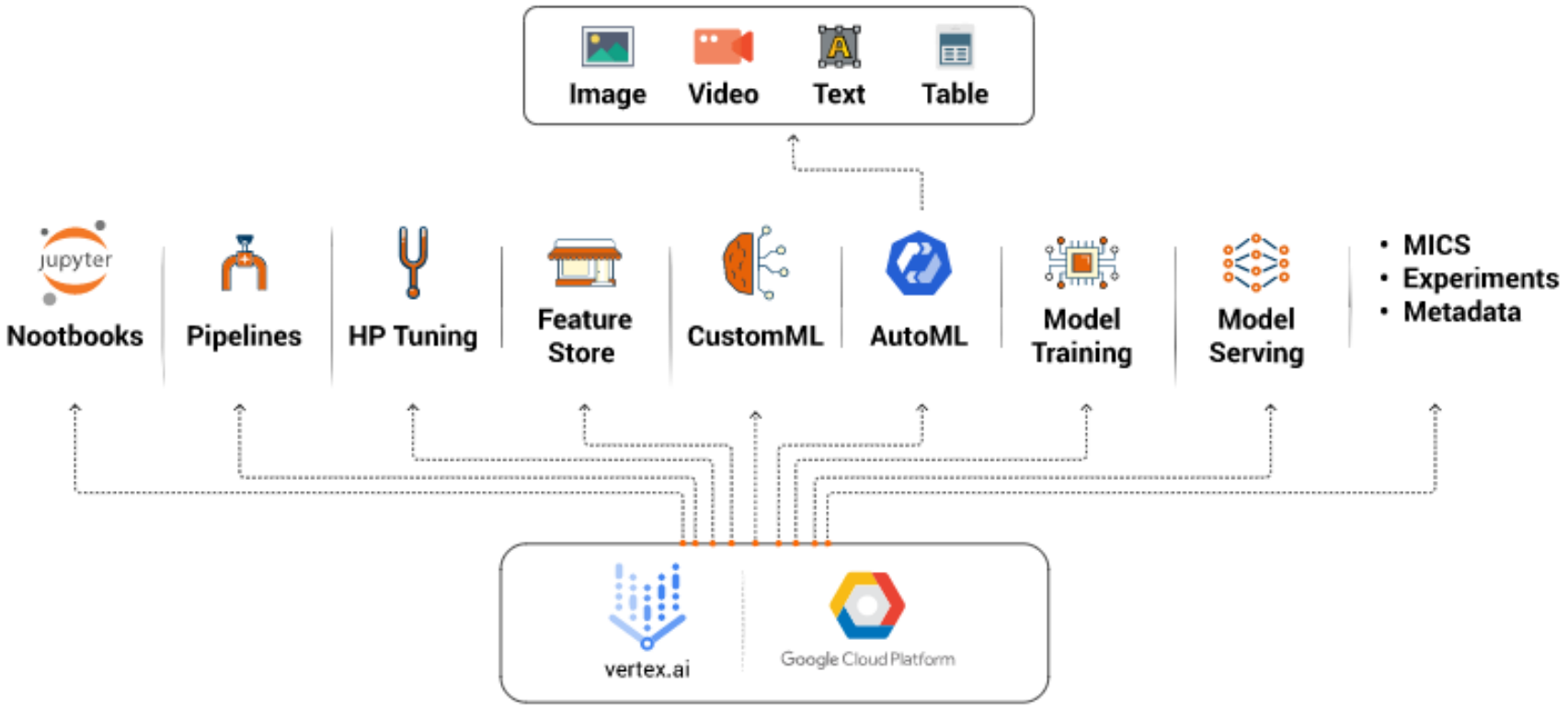
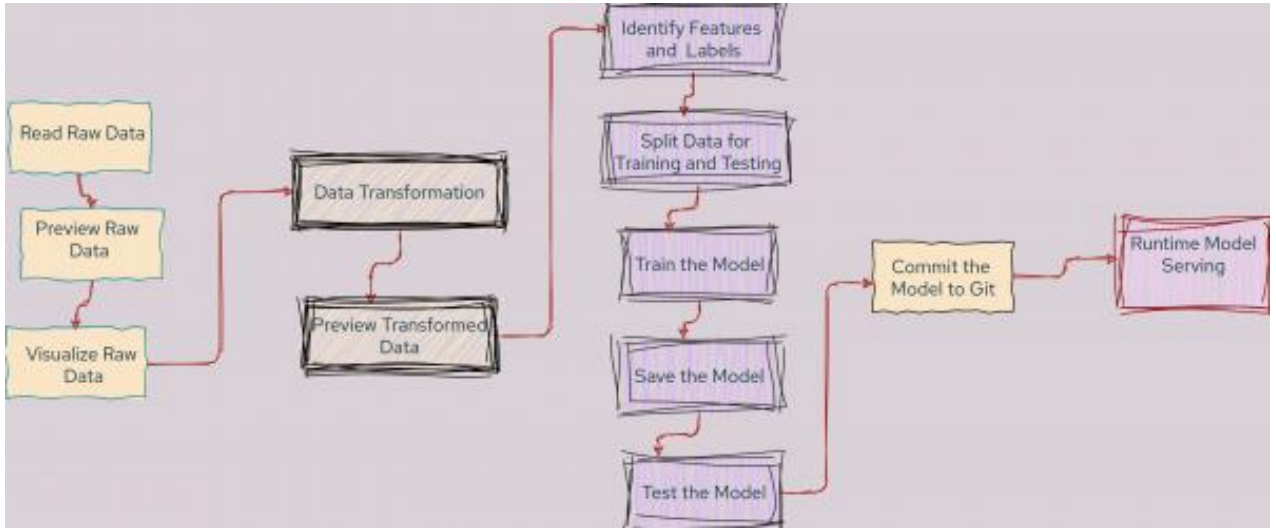


Image
Video
Text
Table
jupyter
Nootbooks
Pipelines
HP Tuning
Feature Store
CustomML
AutoML
Model Training
Model Serving
• MICS
• Experiments
• Metadata
vertex.ai
Google Cloud Platform

Notebooks

- Provide an interactive computational environment for developing data science applications. Jupyter notebooks combine software code, computational output, explanatory text, and rich content in a single document. Notebooks allow in-browser editing and execution of code and display computation results
- Jupyter notebooks



GitHub & GitLab

- Git: Keep track of changes to source code over time (most common). Code change control
- GitHub or GitLab → privately upload codebases using Git

Container

- Standard unit of software that packages up code and all its dependencies, so the application runs quickly and reliably from one computing environment to another
- Virtual machines
- Ex is Docker

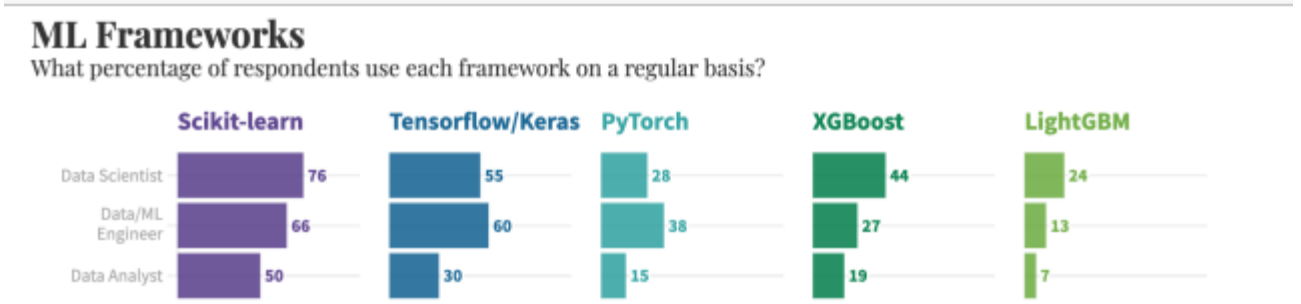
ETL

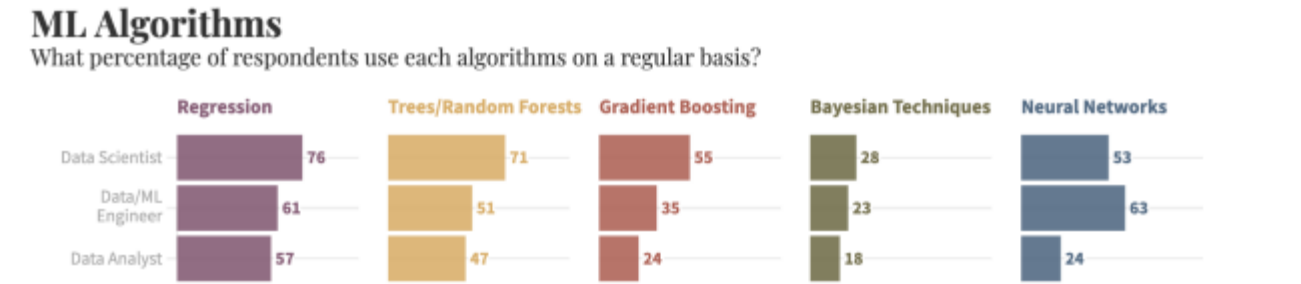
- Extract, Transform, Load
- It refers to taking data from one or multiple sources such as a database, transforming it in some way as needed, and loading it into a data warehouse.

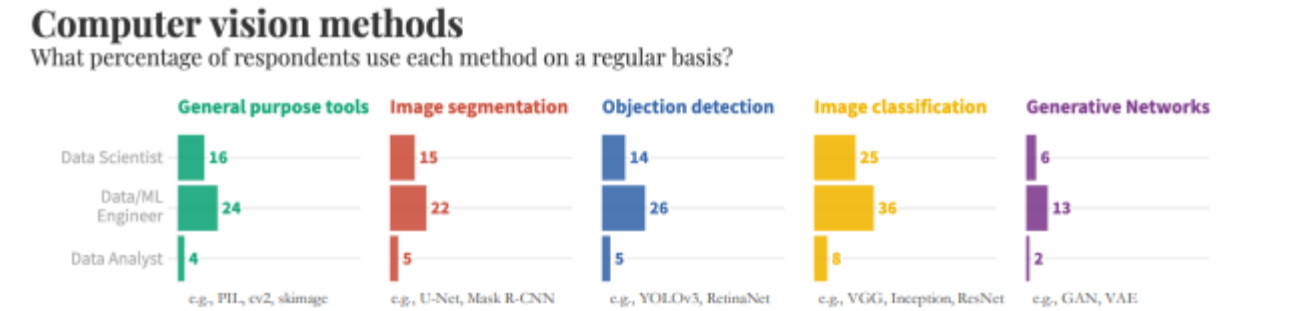
AI SUMMIT
COLUMBUS, OH • OCTOBER 25–27, 2022

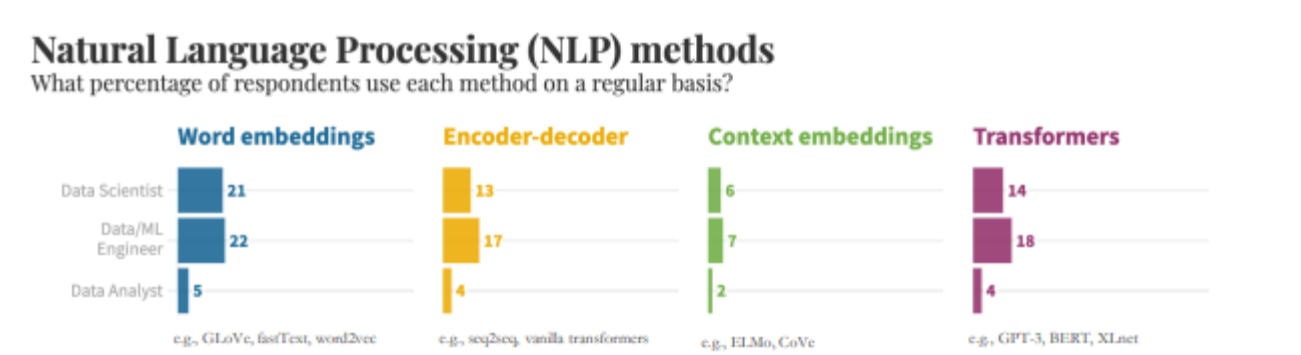
Tools: Coding

- Python
- R
- SQL
- Bash









Data source: https://www.kaggle.com/c/kaggle-survey-2020

AI SUMMIT
COLUMBUS, OH • OCTOBER 25–27, 2022

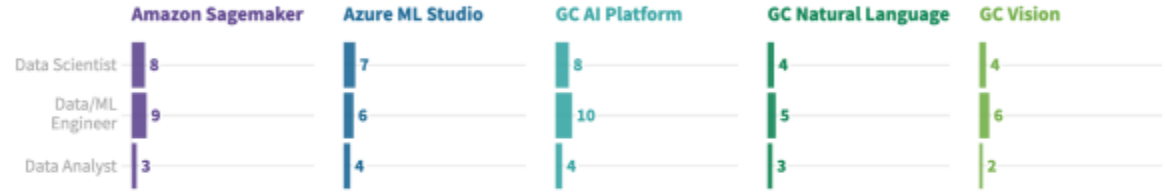
Cloud and Storage

Cloud computing platforms

What percentage of respondents use each platform on a regular basis?

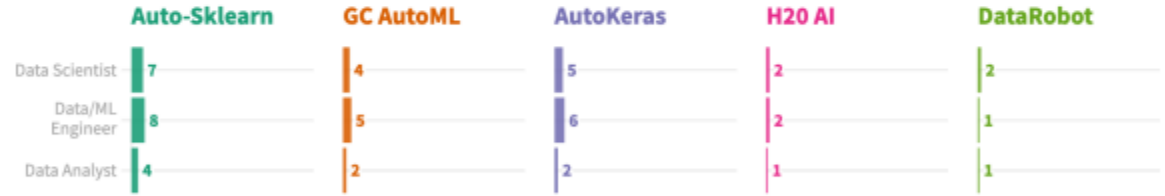
Cloud ML products

What percentage of respondents use each product on a regular basis?



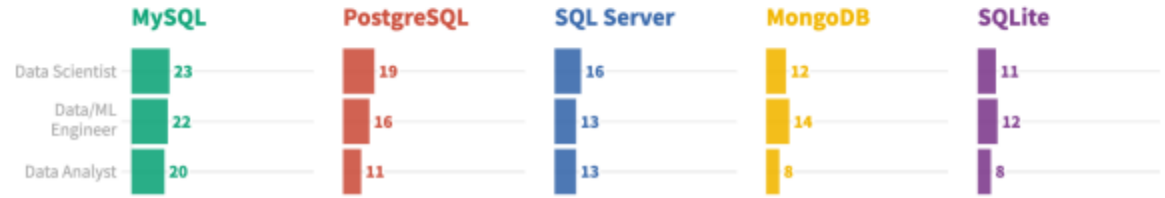
AutoML tools

What percentage of respondents use each tool on a regular basis?



Databases

What percentage of respondents use each database on a regular basis?



Data source: https://www.kaggle.com/c/kaggle-survey-2020

AI SUMMIT
COLUMBUS, OH • OCTOBER 25–27, 2022

Return to Mentimeter Question (below) results:

Are there any concepts you would like support clarifying?

AI SUMMIT
COLUMBUS, OH • OCTOBER 25–27, 2022

Mentimeter Question:

Are there any additional concepts you would like support clarifying?

AI SUMMIT
COLUMBUS, OH • OCTOBER 25–27, 2022

Mentimeter Question:

What is your top concern about AI/ML?