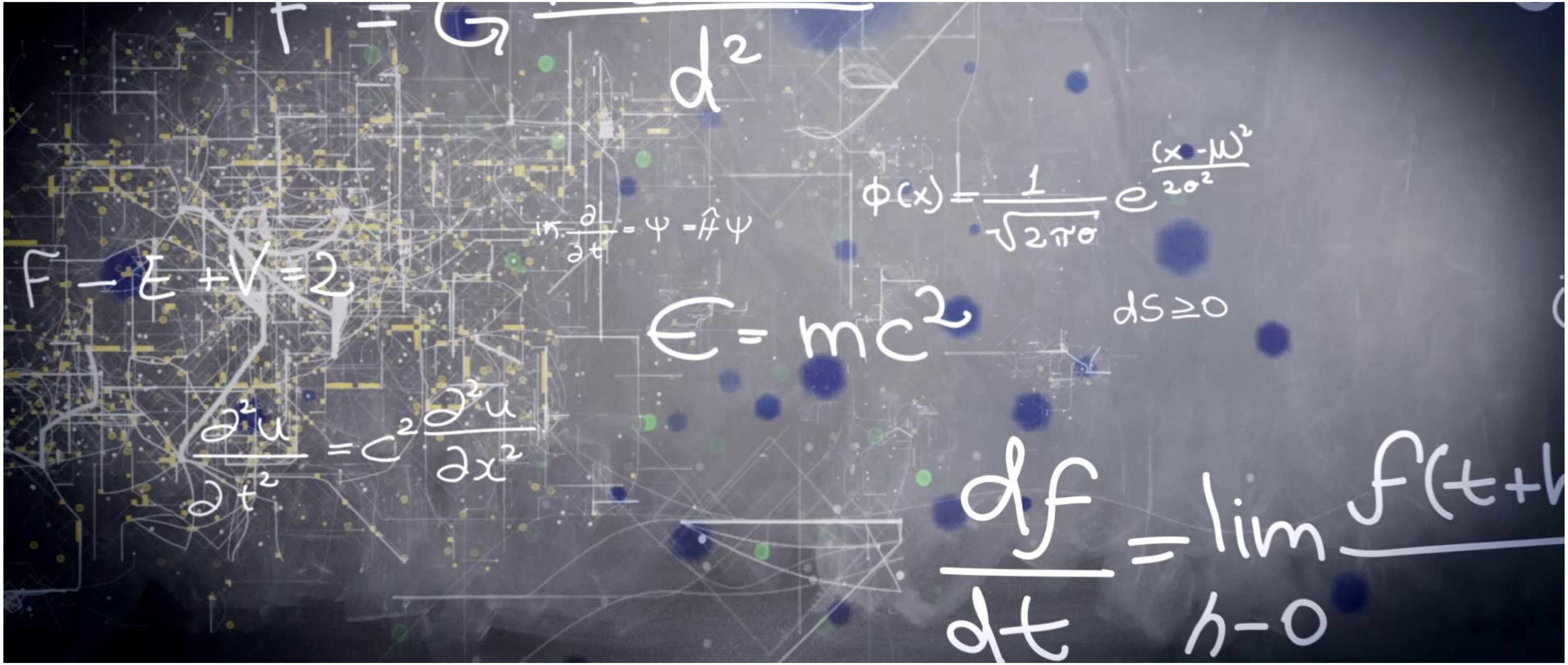
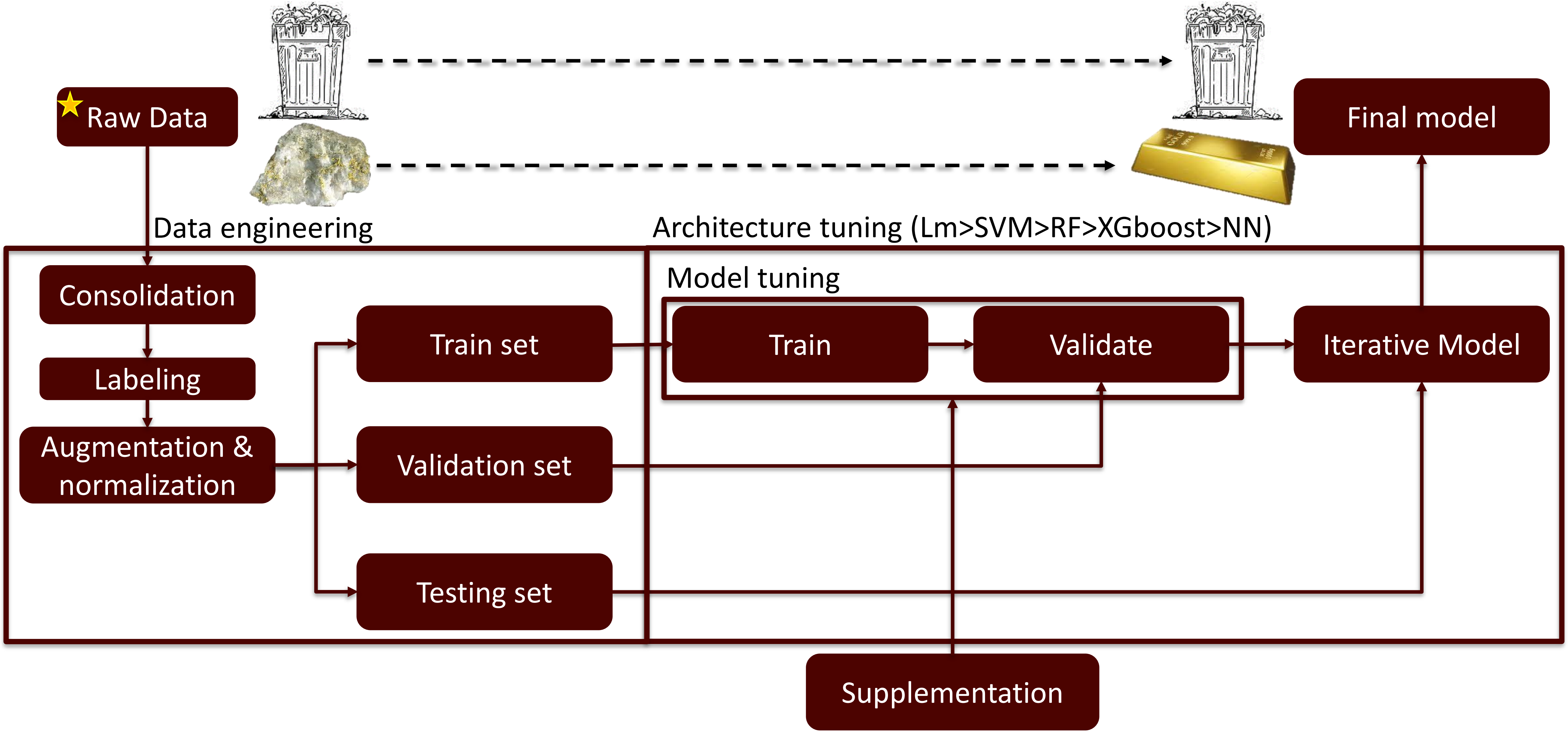
Machine Learning

TEXAS A&M UNIVERSITY

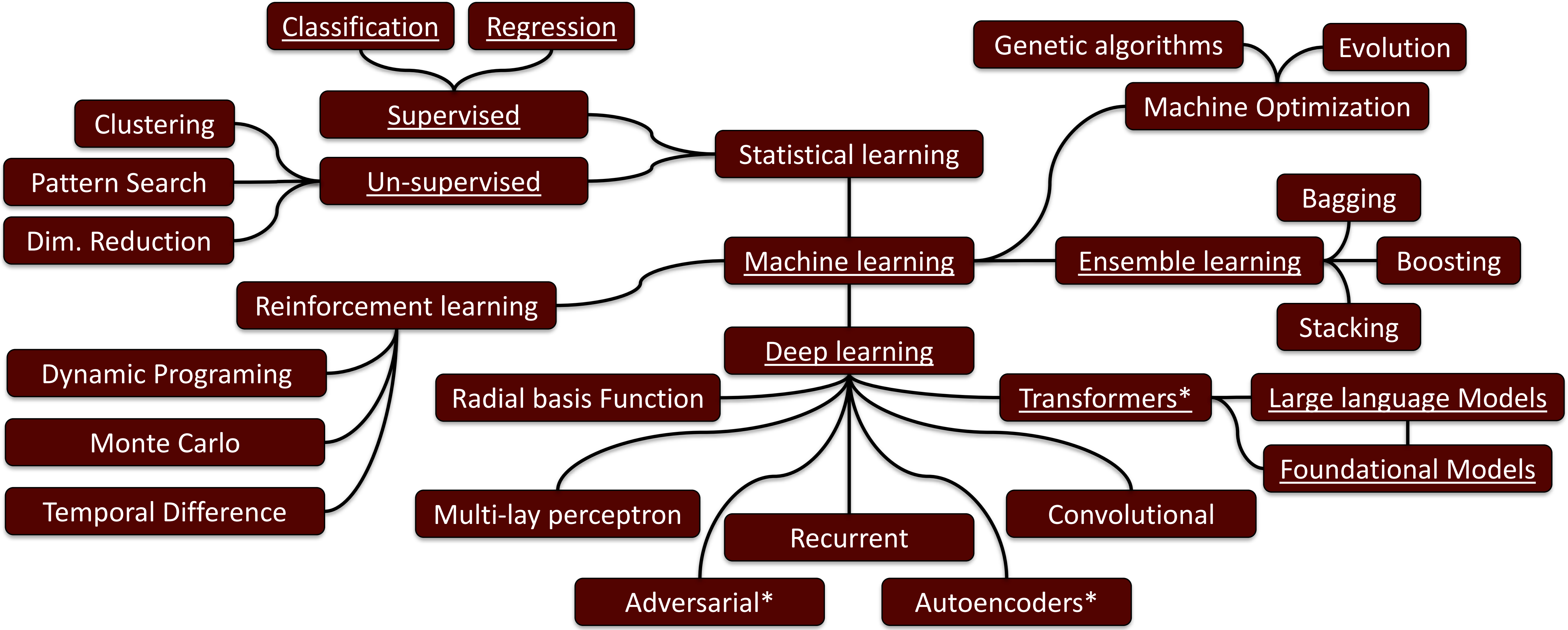


Training a supervised model



A network of ML methods

TEXAS A&M UNIVERSITY

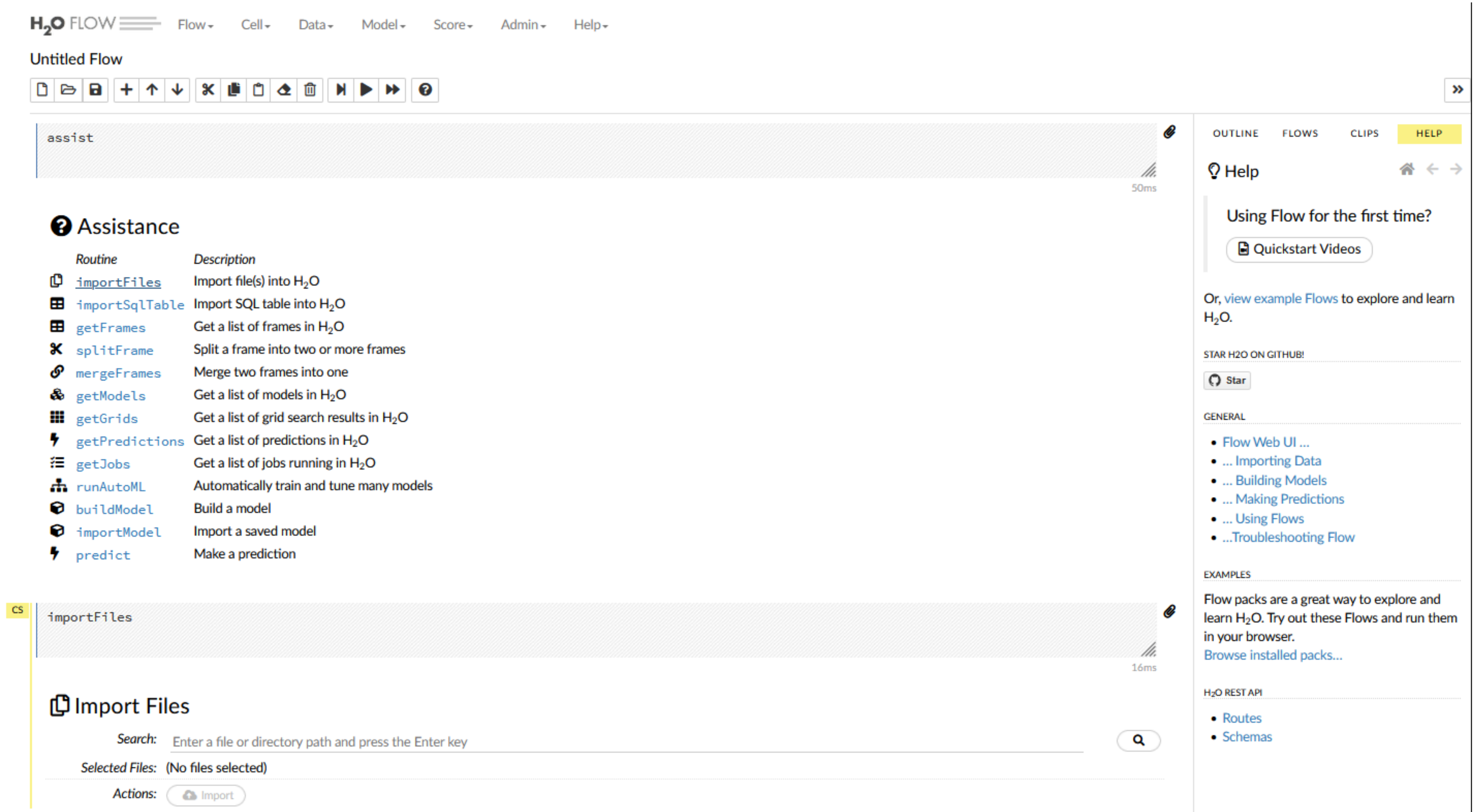


Underline indicts linked information in the appendix. * indicates generative AI

The conventional bar to entry…

Low-code AutoML (H2O Flow)

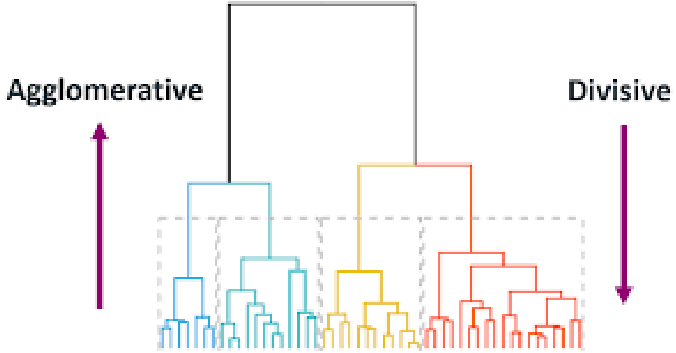
```
Python 3.8.17 (default, Jul  5 2023, 20:44:21) [MSC v.1916 64 bit (AMD64)] :: Anaconda, Inc. on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> import h2o
>>> h2o.init()
```



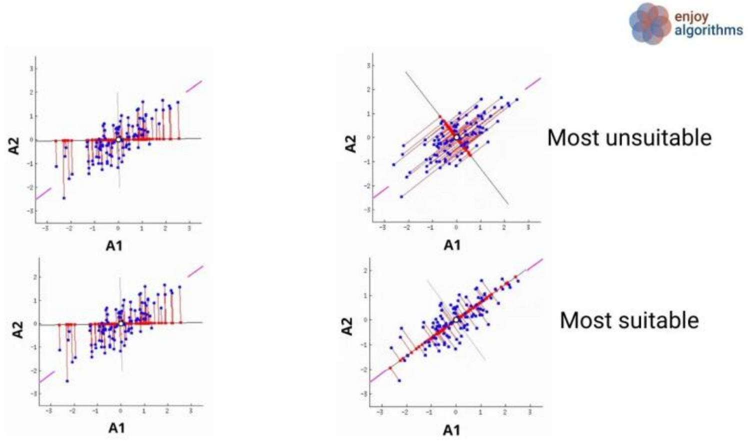
Unsupervised ML

TEXAS A&M UNIVERSITY

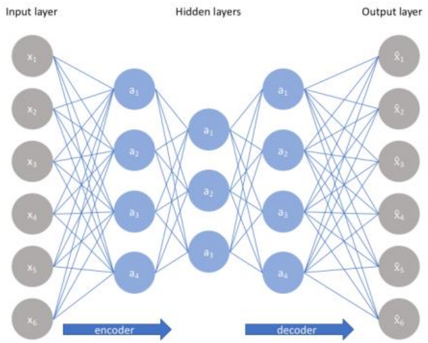
Hierarchical Clustering



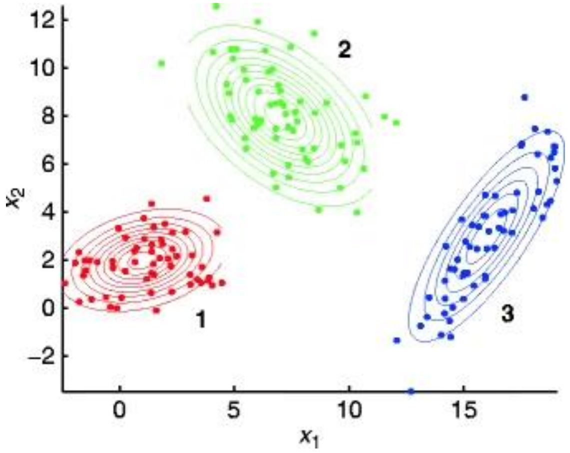
Principle component analysis (PCA)



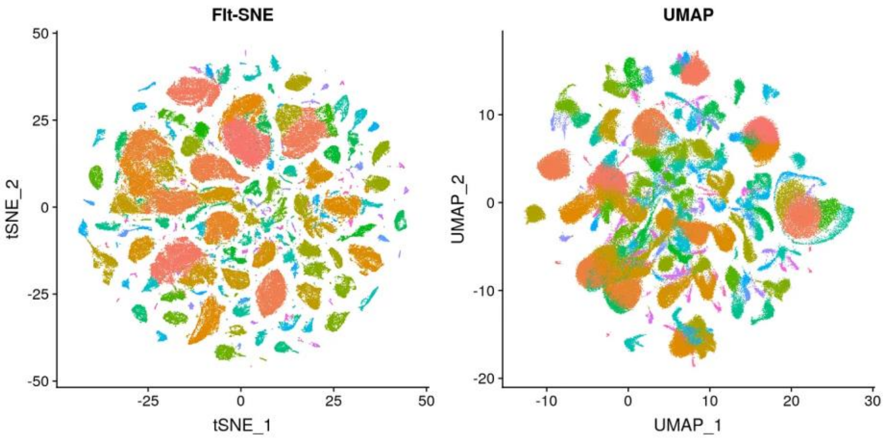
Autoencoders



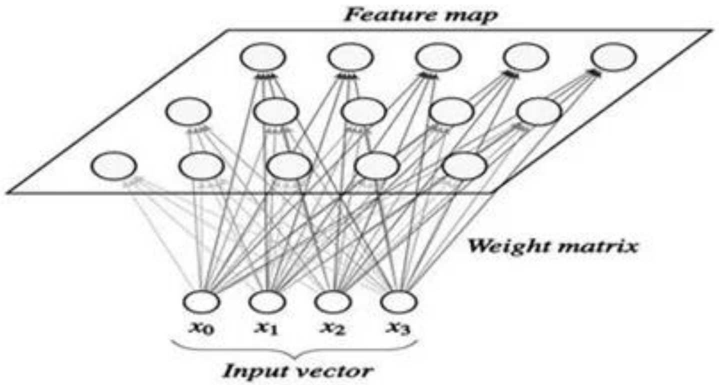
Model-based clustering



t-SNE & UMAP



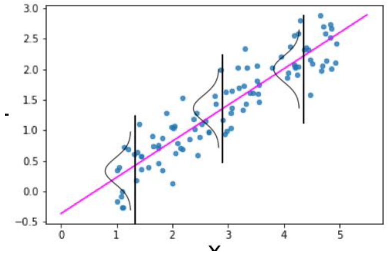
Self organizing maps (SOM)



Citations/info links in images

Conventional supervised ML

GLM



Lasso/Ridge/Elastic network

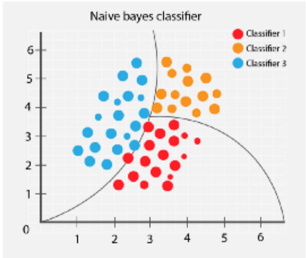
$$\frac{\sum_{i=1}^{n}(y_i - x_i^J\hat{\beta})^2}{2n} + \lambda\left(\frac{1-\alpha}{2}\sum_{j=1}^{m}\hat{\beta}_j^2 + \alpha\sum_{j=1}^{m}|\hat{\beta}_j|\right)$$

Naïve Bayes model

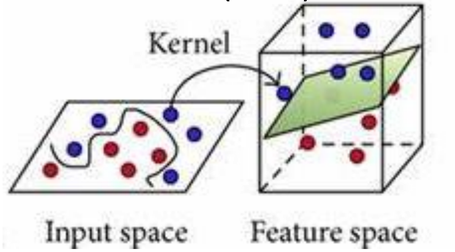
$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

using Bayesian probability terminology, the above equation can be written as

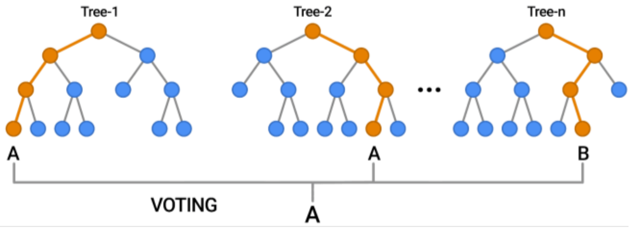
$$\text{Posterior} = \frac{\text{prior x likelihood}}{\text{evidence}}$$



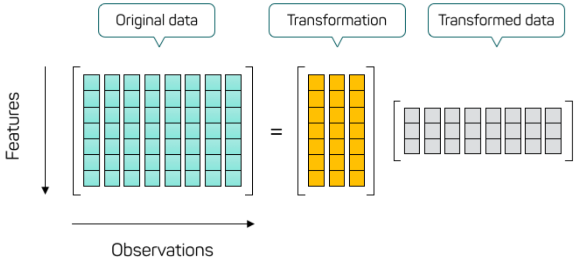
Support vector machine (SVM)



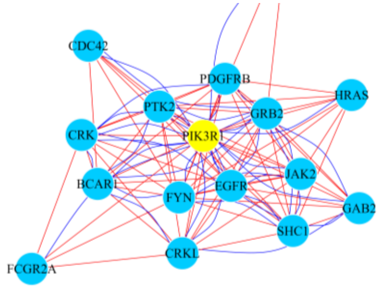
Tree/RF/XGBoost



Recommender systems



Network-based algorithms



Citations/info links in images

Reinforcement learning

Monte-Carlo

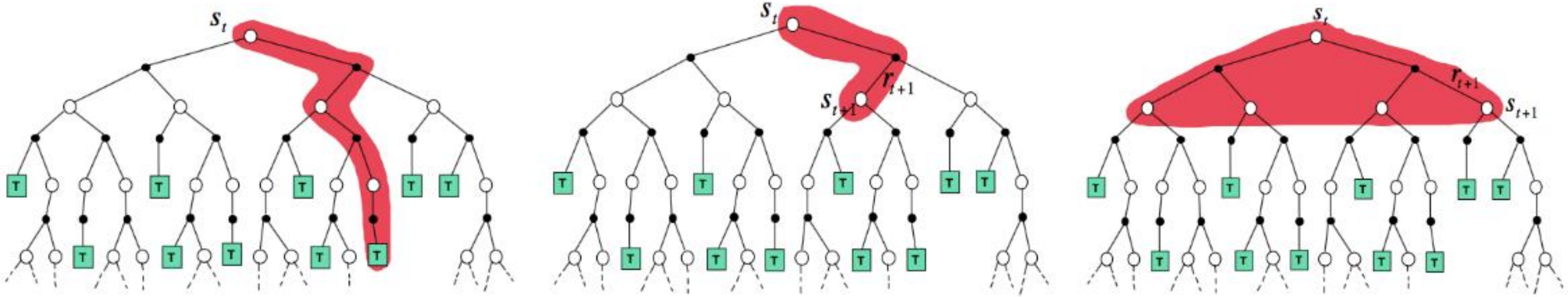
$V(S_t) \leftarrow V(S_t) + \alpha \left(G_t - V(S_t) \right)$

Temporal-Difference

$V(S_t) \leftarrow V(S_t) + \alpha \left(R_{t+1} + \gamma V(S_{t+1}) - V(S_t) \right)$

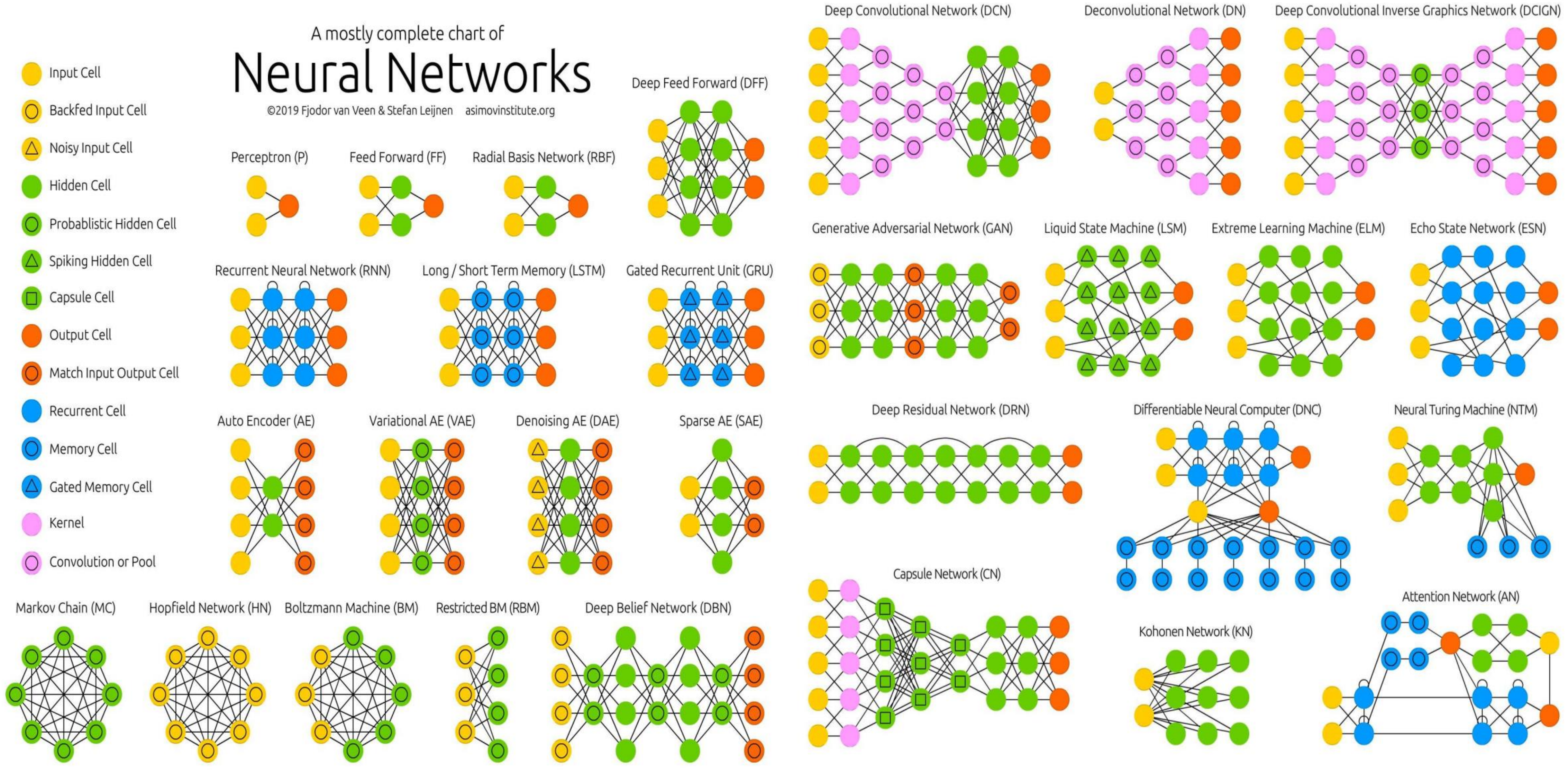
Dynamic Programming

$V(S_t) \leftarrow \mathbb{E}_\pi \left[R_{t+1} + \gamma V(S_{t+1}) \right]$



https://roboticseabass.com/2020/08/02/an-intuitive-guide-to-reinforcement-learning/

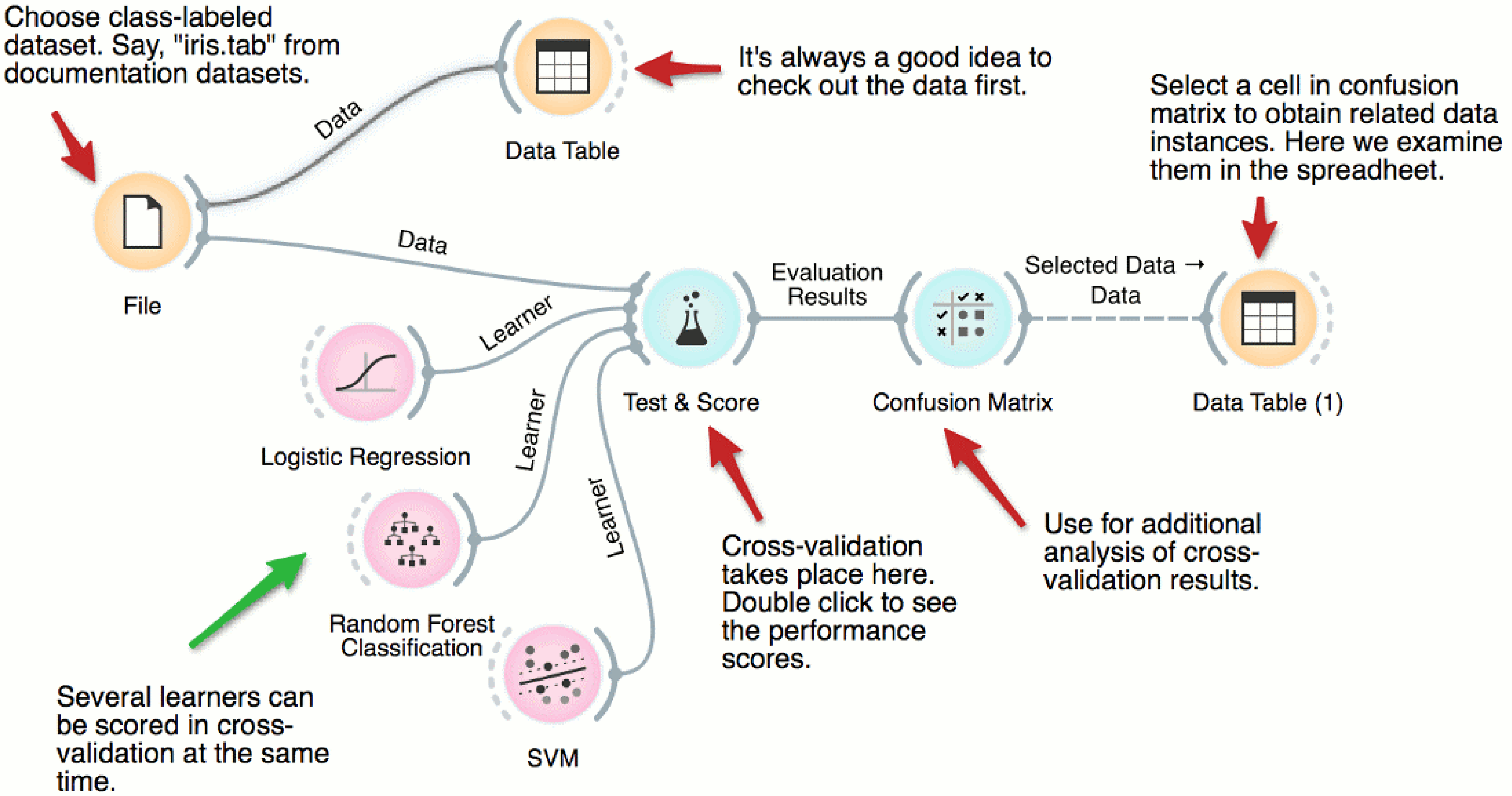
Deep learning models



https://towardsdatascience.com/the-mostly-complete-chart-of-neural-networks-explained-3fb6f2367464

No-Code Generalized ML

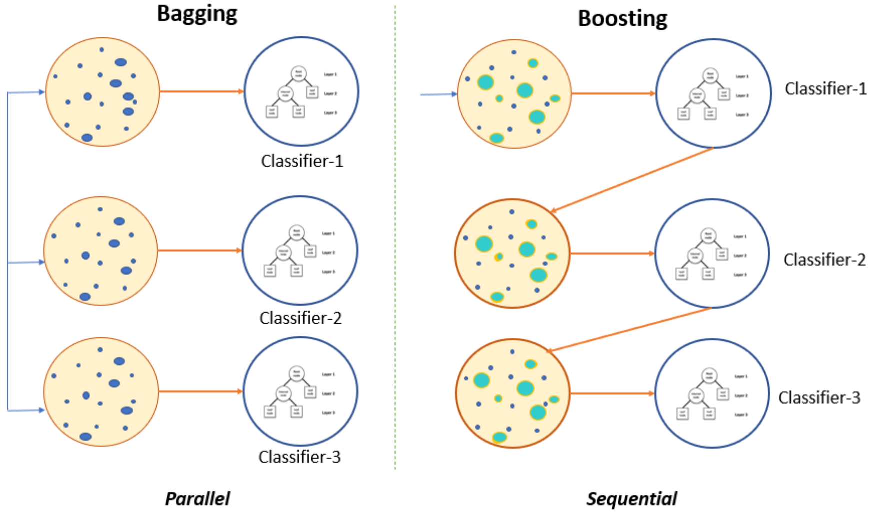
TEXAS A&M UNIVERSITY



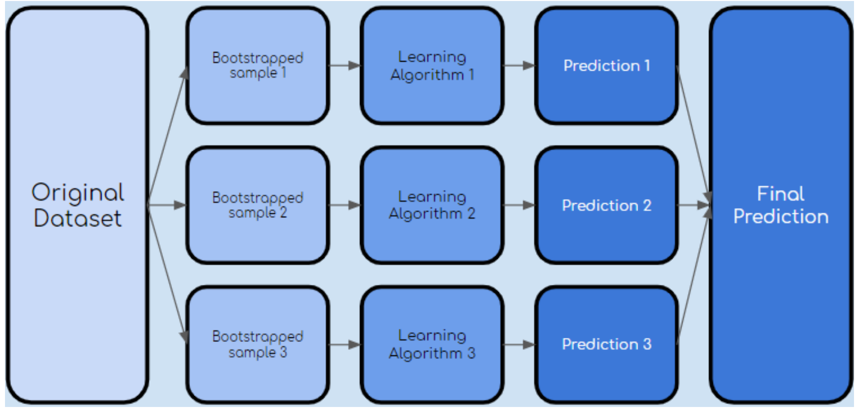
https://orangedatamining.com/workflows/

Ensemble modeling

Bagging and boosting



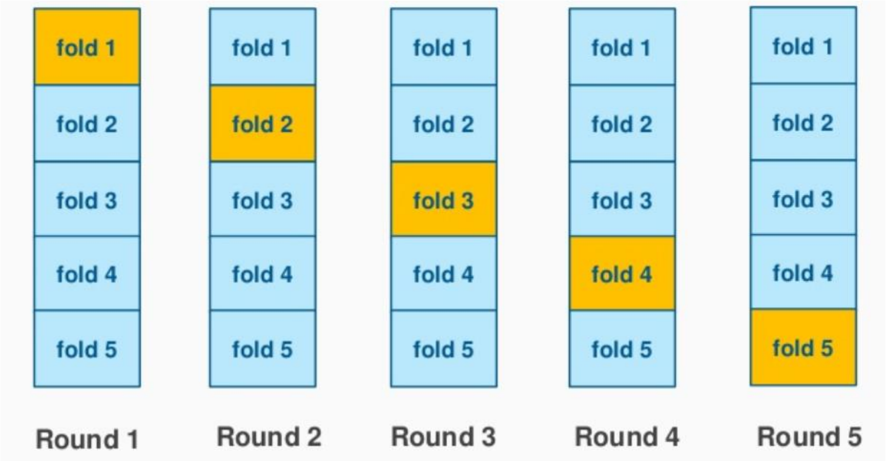
Stacking



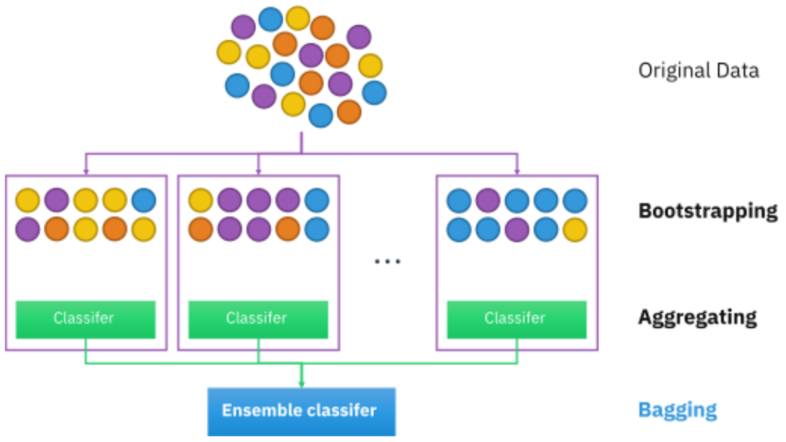
Performance enhancing

Robustness enhancing

Cross-validated

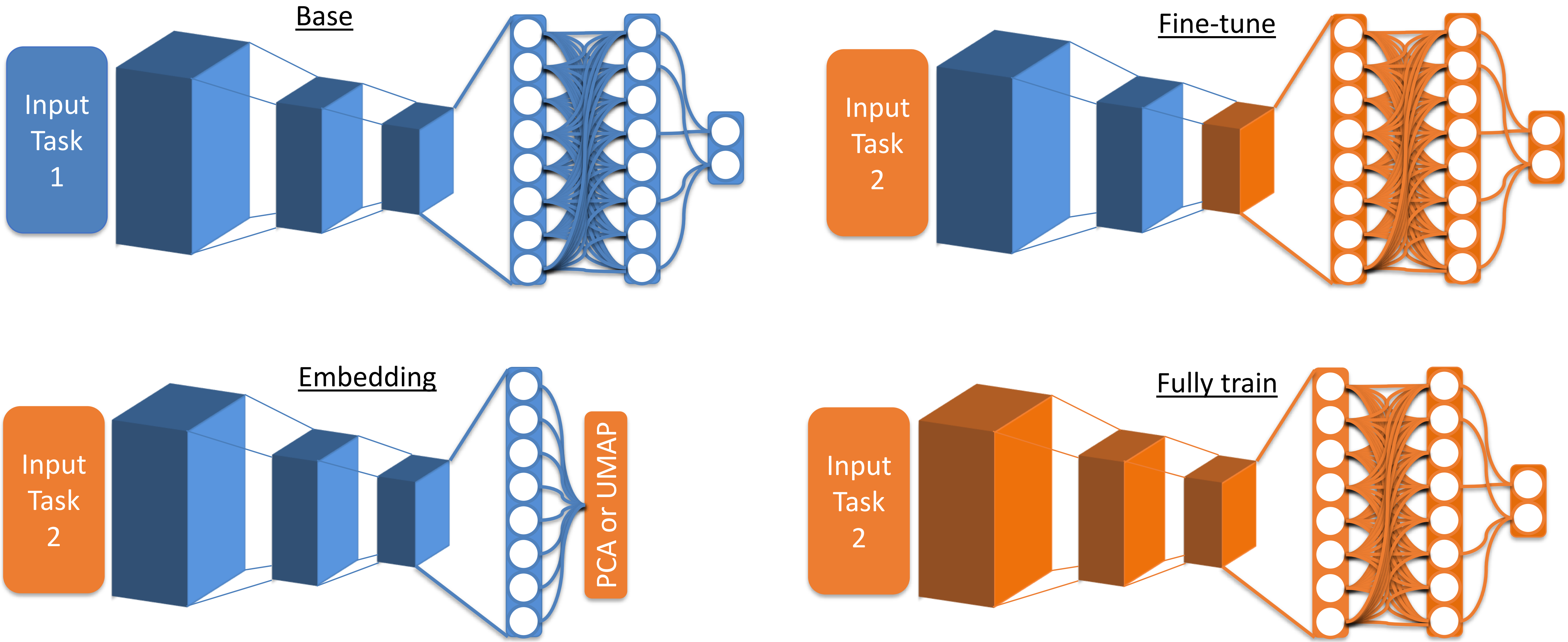


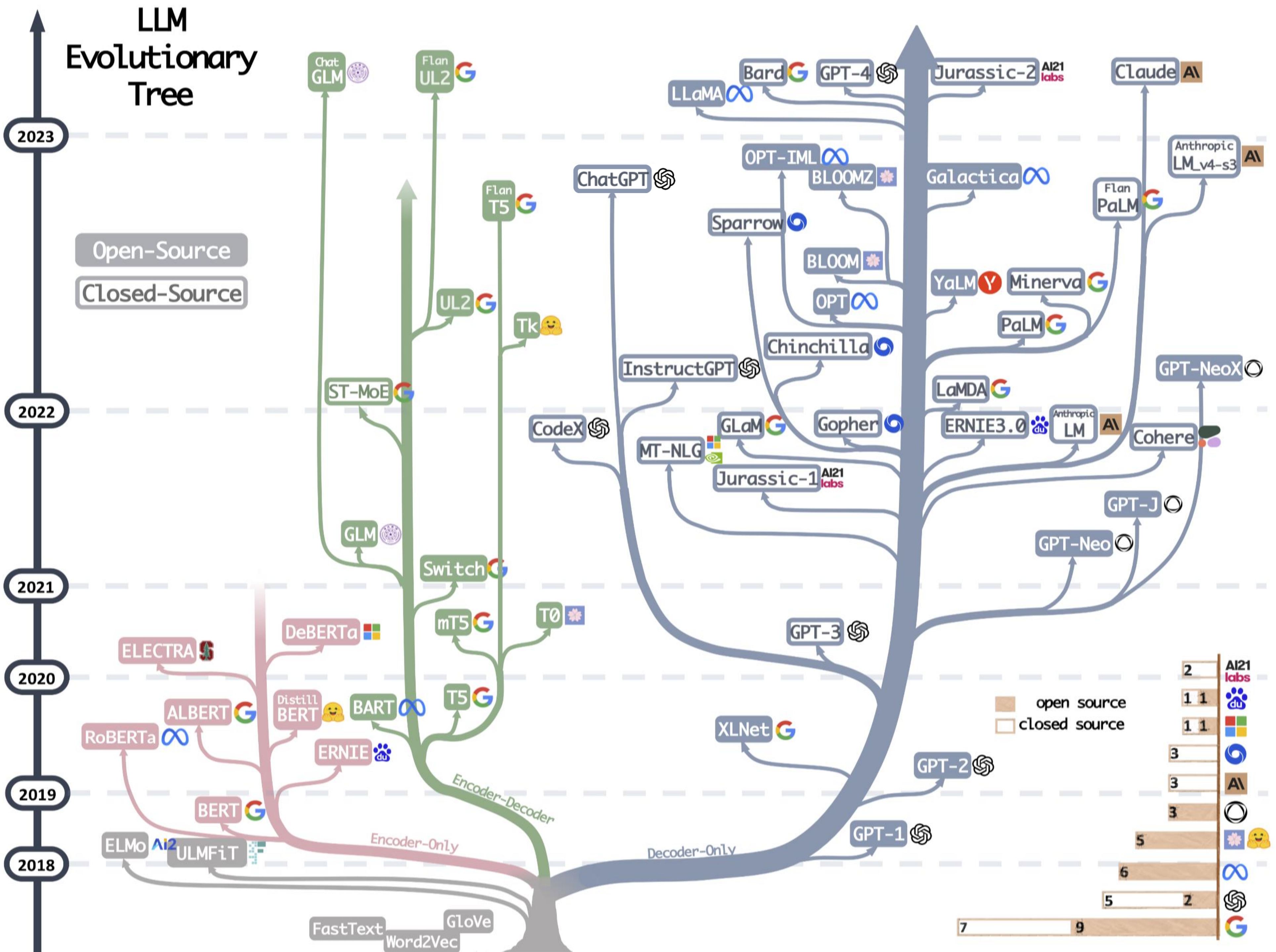
Bootstrapped



https://stats.stackexchange.com/questions/49540/understanding-stratified-cross-validation; https://www.analyticsvidhya.com/blog/2020/02/what-is-bootstrap-sampling-in-statistics-and-machine-learning/

Transfer learning methods





LLM Evolutionary Tree
2023
2022
2021
2020
2019
2018
Open-Source
Closed-Source
Encoder-Only
Encoder-Decoder
Decoder-Only
ChatGLM
Flan UL2
Flan T5
UL2
Tk
ST-MoE
GLM
Switch
mT5
T0
T5
BART
ERNIE
DeBERTa
ELECTRA
ALBERT
Distill BERT
RoBERTa
BERT
ELMo
ULMFiT
FastText
Word2Vec
GloVe
LLaMA
Bard
GPT-4
Jurassic-2
Claude
ChatGPT
OPT-IML
BLOOMZ
Galactica
Anthropic LM_v4-s3
Flan PaLM
Sparrow
BLOOM
YaLM
Minerva
OPT
PaLM
Chinchilla
InstructGPT
GPT-NeoX
LaMDA
CodeX
GLaM
Gopher
ERNIE3.0
Anthropic LM
Cohere
MT-NLG
Jurassic-1
GPT-J
GPT-Neo
GPT-3
XLNet
GPT-2
GPT-1
open source
closed source
2
1 1
1 1
3
3
3
5
6
5 2
7 9

Foundational models

Broad Training Data: trained on extensive datasets, which require substantial computational resources. This training allows them to learn a wide range of tasks and skills during the initial phase.

Self-Supervision: Generally, use self-supervision techniques during training where labels or targets are generated from the data itself, rather than relying solely on human-labeled data.

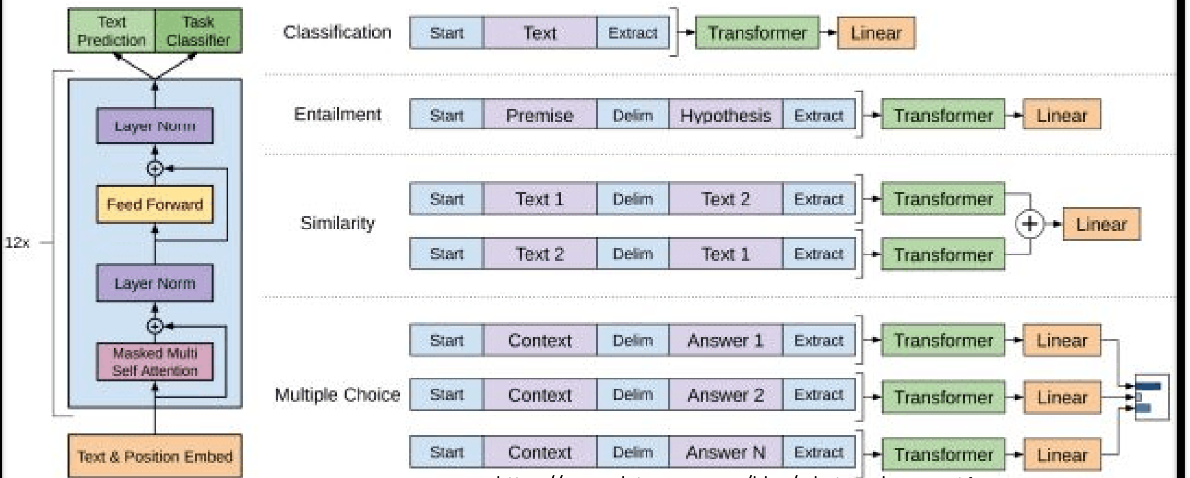
Large Parameter Count: Typically contains at least billions of parameters to enable them to capture complex patterns and relationships in the data.

Applicability Across Contexts: Applicable across a wide range of contexts, can be secondarily fine-tuned for specific tasks with minimal adjustments, making them highly versatile.

Foundation models

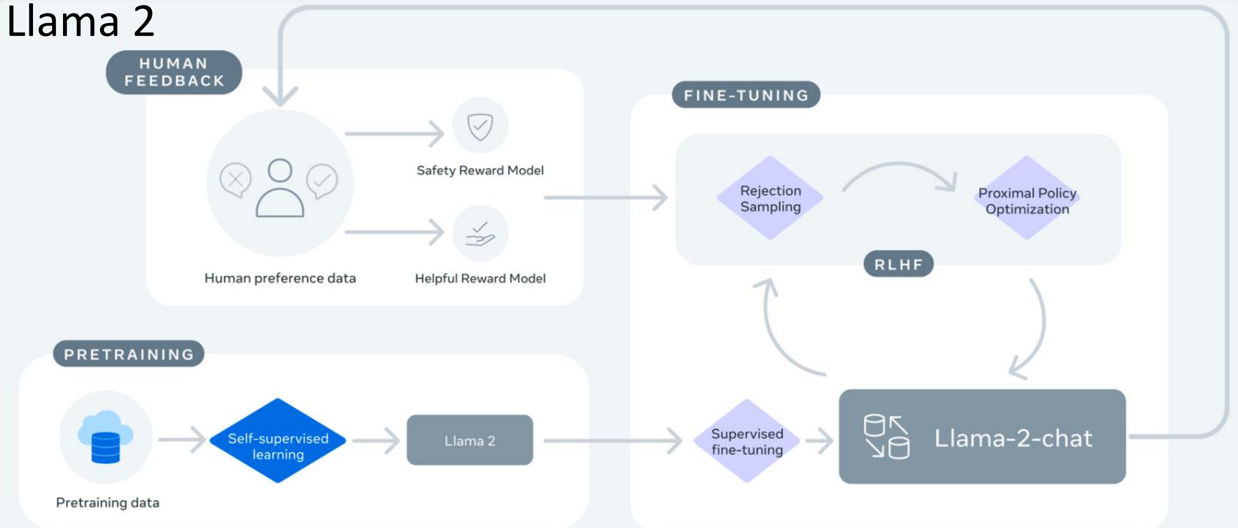
TEXAS A&M UNIVERSITY

Generative Pre-trained Transformer (GPT)



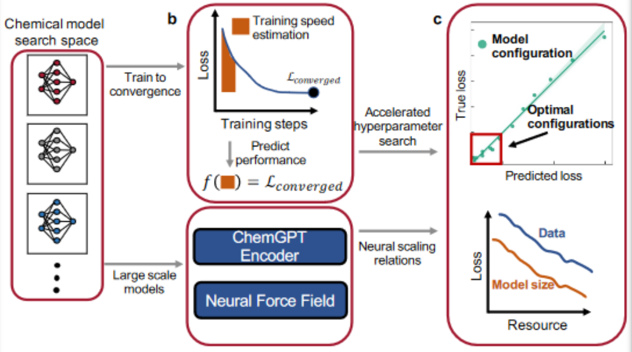
https://www.datacamp.com/blog/what-we-know-gpt4

Llama 2



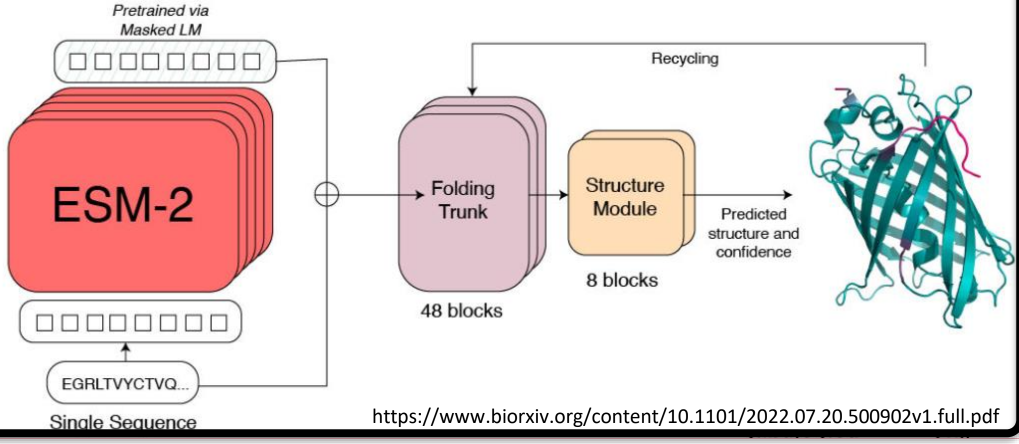
https://huggingface.co/blog/llama2

ChemGPT



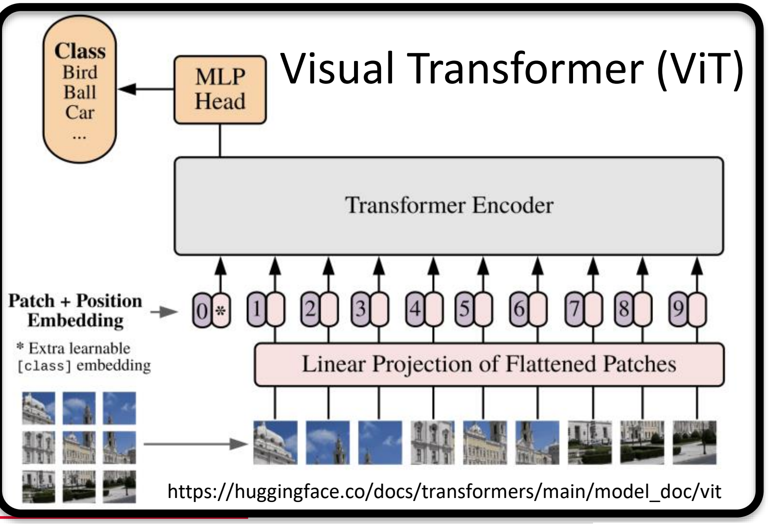
https://chemrxiv.org/engage/chemrxiv/article-details/627bddd544bdd532395fb4b5

ESM2: A large language model of protein structure



https://www.biorxiv.org/content/10.1101/2022.07.20.500902v1.full.pdf

Visual Transformer (ViT)



https://huggingface.co/docs/transformers/main/model_doc/vit

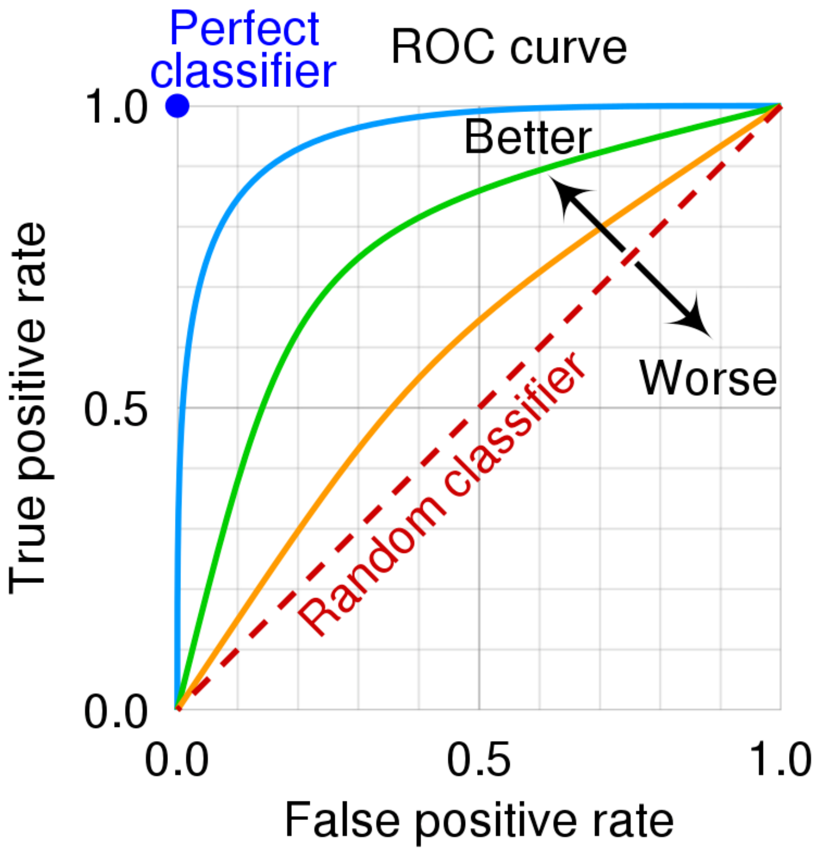
Open source and local deployment demo's: Hugging Face

Model Type	Implementation & Key Features	Use Cases
Memory Networks	- LSTM - External memory for handling long-term dependencies - Read and write operations.	- Question-answering - Dialogue systems
Causal Language Modeling (CLM)	- GPT, Llama - Autoregressive model that predicts sequential tokens. - Unidirectional context.	- Text generation - Summarization
Masked Language Modeling (MLM)	- BERT, RoBERTa - Input tokens are masked - Model predicts context. - Bidirectional context.	- Text classification - Sentiment analysis - Named entity recognition
Sequence-to-Sequence (Seq2Seq)	- T5 - Encoder-decoder architecture. - Handles input-output transformations.	- Machine translation - Summarization - Question-answering

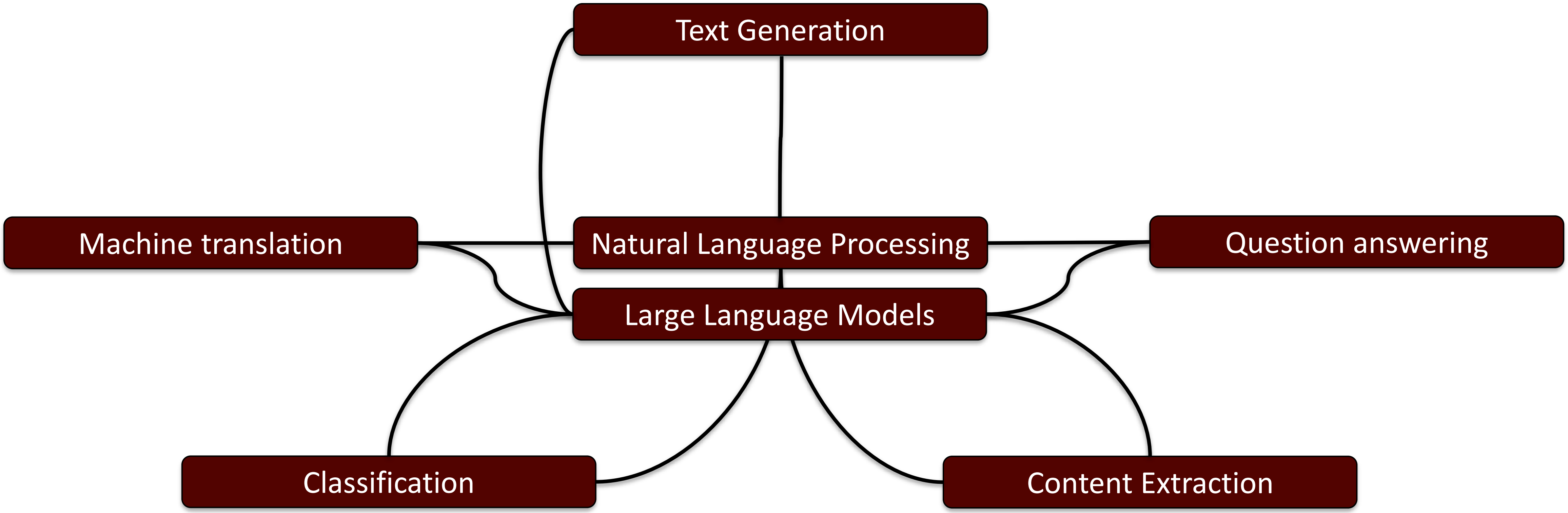
Interpreting the quality of an AI/ML model

TEXAS A&M UNIVERSITY

		Predicted condition		Sources: [1][2][3][4][5][6][7][8][9] view · talk · edit	
	Total population $= P + N$	**Positive (PP)**	**Negative (PN)**	Informedness, bookmaker informedness (BM) $= TPR + TNR - 1$	Prevalence threshold (PT) $= \frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$
Actual condition	**Positive (P)**	**True positive (TP)**, **hit**	**False negative (FN)**, **type II error**, **miss**, **underestimation**	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{TP}{P} = 1 - FNR$	False negative rate (FNR), miss rate $= \frac{FN}{P} = 1 - TPR$
	Negative (N)	**False positive (FP)**, **type I error**, **false alarm**, **overestimation**	**True negative (TN)**, **correct rejection**	False positive rate (FPR), probability of false alarm, fall-out $= \frac{FP}{N} = 1 - TNR$	True negative rate (TNR), specificity (SPC), selectivity $= \frac{TN}{N} = 1 - FPR$
	Prevalence $= \frac{P}{P + N}$	Positive predictive value (PPV), precision $= \frac{TP}{PP} = 1 - FDR$	False omission rate (FOR) $= \frac{FN}{PN} = 1 - NPV$	Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$	Negative likelihood ratio (LR−) $= \frac{FNR}{TNR}$
	Accuracy (ACC) $= \frac{TP + TN}{P + N}$	False discovery rate (FDR) $= \frac{FP}{PP} = 1 - PPV$	Negative predictive value (NPV) $= \frac{TN}{PN} = 1 - FOR$	Markedness (MK), deltaP (Δp) $= PPV + NPV - 1$	Diagnostic odds ratio (DOR) $= \frac{LR+}{LR-}$
	Balanced accuracy (BA) $= \frac{TPR + TNR}{2}$	F_1 score $= \frac{2 PPV \times TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$	Fowlkes–Mallows index (FM) $= \sqrt{PPV \times TPR}$	Matthews correlation coefficient (MCC) $= \sqrt{TPR \times TNR \times PPV \times NPV} - \sqrt{FNR \times FPR \times FOR \times FDR}$	Threat score (TS), critical success index (CSI), Jaccard index $= \frac{TP}{TP + FN + FP}$



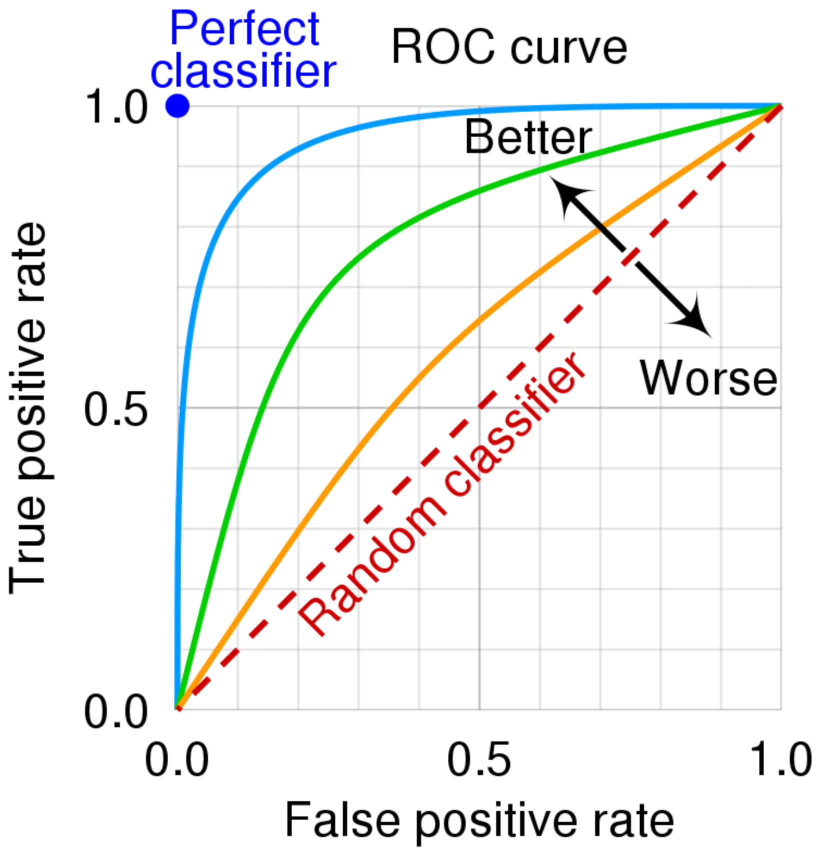
Natural language processing



Interpreting the quality of an AI/ML model

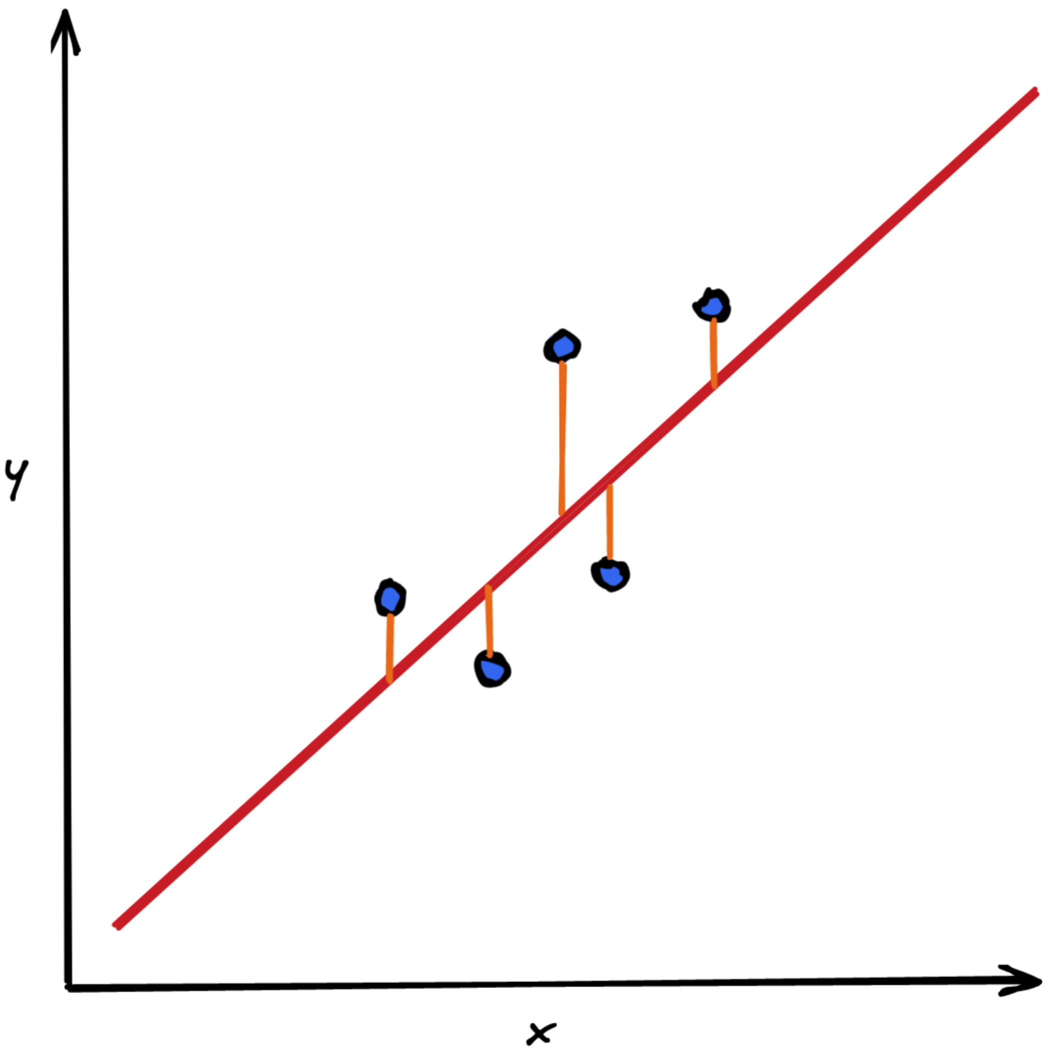
TEXAS A&M UNIVERSITY

		Predicted condition		Sources: [1][2][3][4][5][6][7][8][9] view · talk · edit	
	Total population $= P + N$	**Positive (PP)**	**Negative (PN)**	Informedness, bookmaker informedness (BM) $= TPR + TNR - 1$	Prevalence threshold (PT) $= \frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$
Actual condition	**Positive (P)**	**True positive (TP)**, hit	**False negative (FN)**, **type II error**, **miss**, **underestimation**	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{TP}{P} = 1 - FNR$	False negative rate (FNR), miss rate $= \frac{FN}{P} = 1 - TPR$
Actual condition	**Negative (N)**	**False positive (FP)**, **type I error**, **false alarm**, **overestimation**	**True negative (TN)**, **correct rejection**	False positive rate (FPR), probability of false alarm, fall-out $= \frac{FP}{N} = 1 - TNR$	True negative rate (TNR), specificity (SPC), selectivity $= \frac{TN}{N} = 1 - FPR$
	Prevalence $= \frac{P}{P + N}$	Positive predictive value (PPV), precision $= \frac{TP}{PP} = 1 - FDR$	False omission rate (FOR) $= \frac{FN}{PN} = 1 - NPV$	Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$	Negative likelihood ratio (LR−) $= \frac{FNR}{TNR}$
	Accuracy (ACC) $= \frac{TP + TN}{P + N}$	False discovery rate (FDR) $= \frac{FP}{PP} = 1 - PPV$	Negative predictive value (NPV) $= \frac{TN}{PN} = 1 - FOR$	Markedness (MK), deltaP (Δp) $= PPV + NPV - 1$	Diagnostic odds ratio (DOR) $= \frac{LR+}{LR-}$
	Balanced accuracy (BA) $= \frac{TPR + TNR}{2}$	F_1 score $= \frac{2 PPV \times TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$	Fowlkes–Mallows index (FM) $= \sqrt{PPV \times TPR}$	Matthews correlation coefficient (MCC) $= \sqrt{TPR \times TNR \times PPV \times NPV} - \sqrt{FNR \times FPR \times FOR \times FDR}$	Threat score (TS), critical success index (CSI), Jaccard index $= \frac{TP}{TP + FN + FP}$



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

Evaluating a regressive model



Metric	Benefits	Limitations
R-squared (R^2)	- High values = a good fit - a measure of the proportion of variance	- Influenced by the sample size & # of predictors - May not be reliable when there are outliers or non-linear relationships in the data
Adjusted R-squared (R^2_adj)	- Similar to R^2 Takes into account the number of predictors - Provides a more accurate assessment of the model's performance when there are multiple predictors	- May be less informative when there are only a few predictors in the model
Mean Squared Error (MSE)	- Measures the average squared difference between the predicted and actual values	- Provides a measure of the magnitude of the errors in the model
Mean Absolute Error (MAE)	- Measures the average absolute difference between the predicted and actual values	- Provides a measure of the magnitude of the errors in the model
Root Mean Squared Error (RMSE)	- Measures the square root of the average squared difference between the predicted and actual values	- Provides a measure of the magnitude of the errors in the model