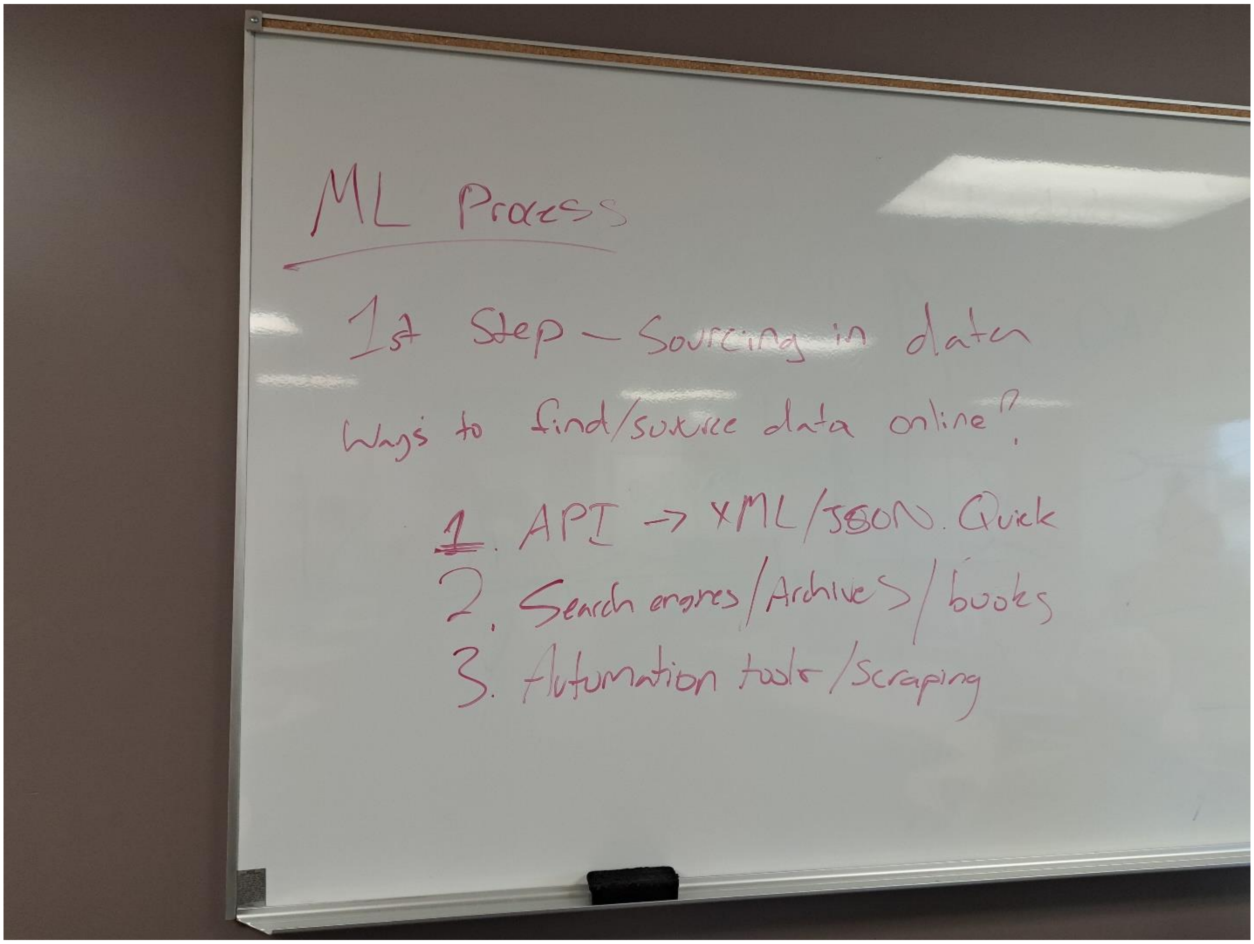
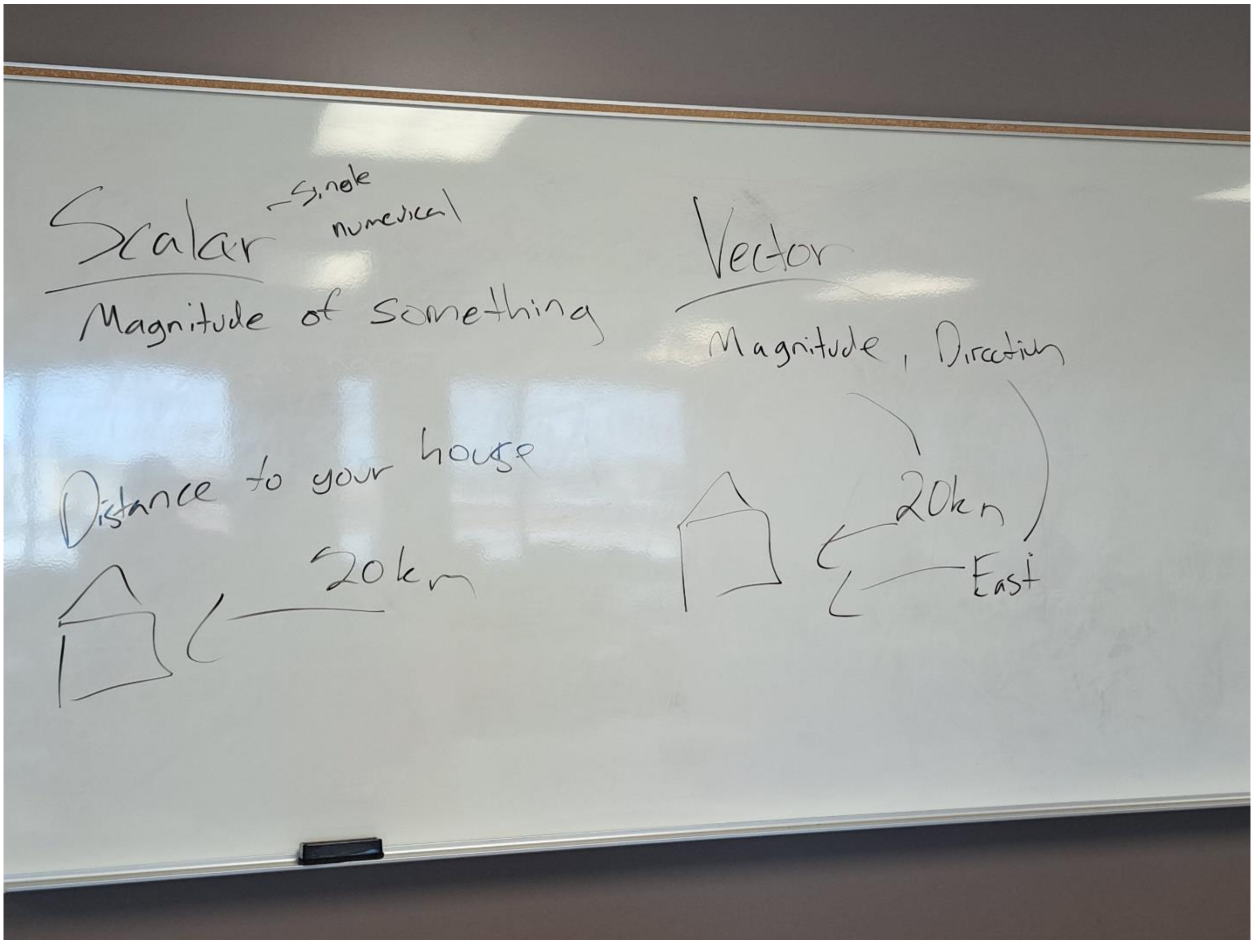
Day 6 Lecture Notes



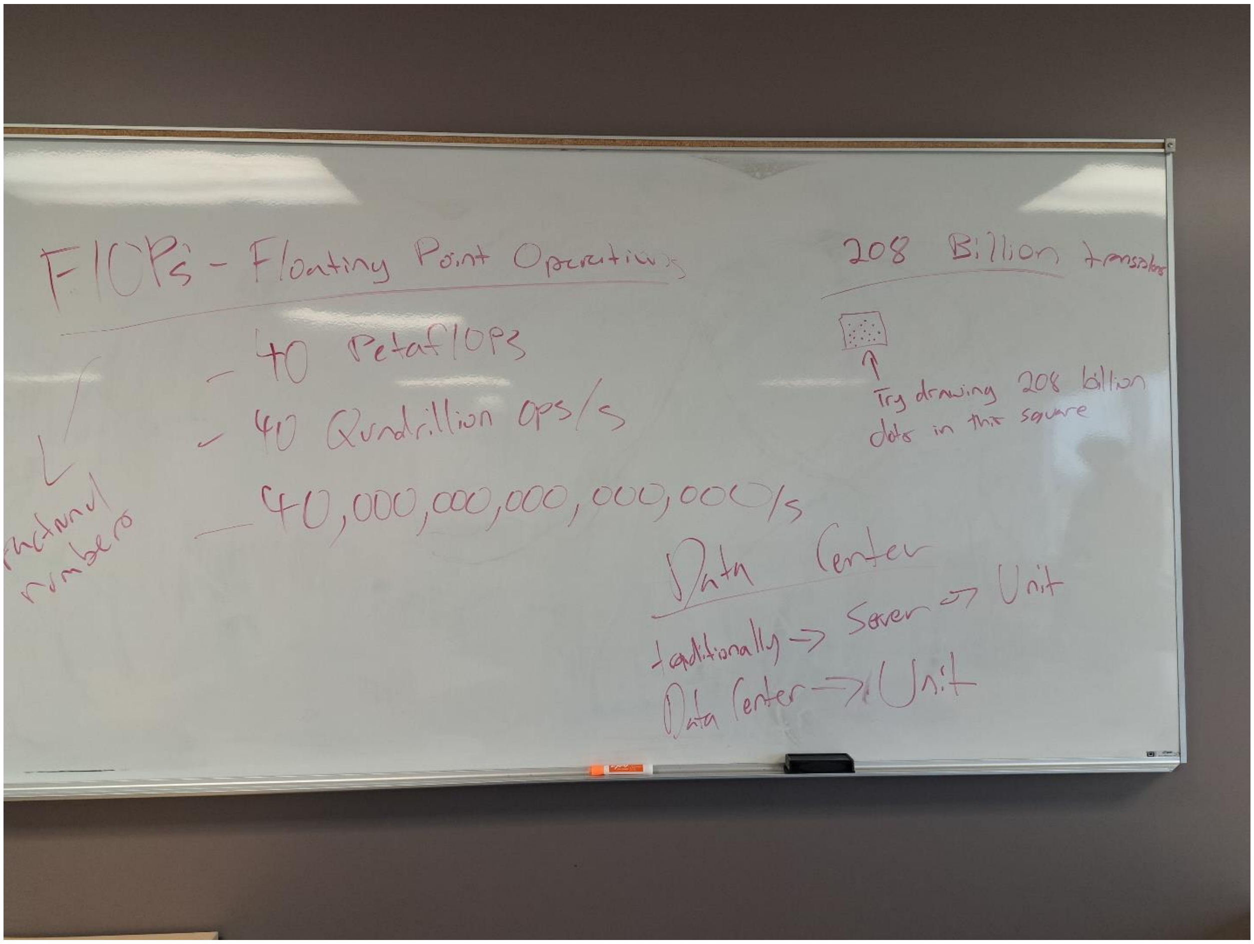
Please see "Web Scraping and Data Extraction with Python." The first step in the Machine Learning process is sourcing data. Use API's when available as these are faster. Many platforms use scraping technology to get regularly updates on archived or near real-time data.

In the code lab we use three packages –CSV to write data to comma separated values files; selenium for automated browser tasks; beautiful soup for parsing html (or xml). There is also a lecture on these located in the same project folder.



We are also introduced to the linear algebraic concepts of vectors, scalars, matrices, & tensors. These concepts also exist in physics but carry the same meaning. In ML, we typically convert labels/features to numerical form to make training and inference infinitely more efficient. The data is usually in 2d tabular form but can be made up of 1d rows or columns or multidimensional arrays. Please see our Blackboard code lab interface (on the main page) for more information and illustrations.

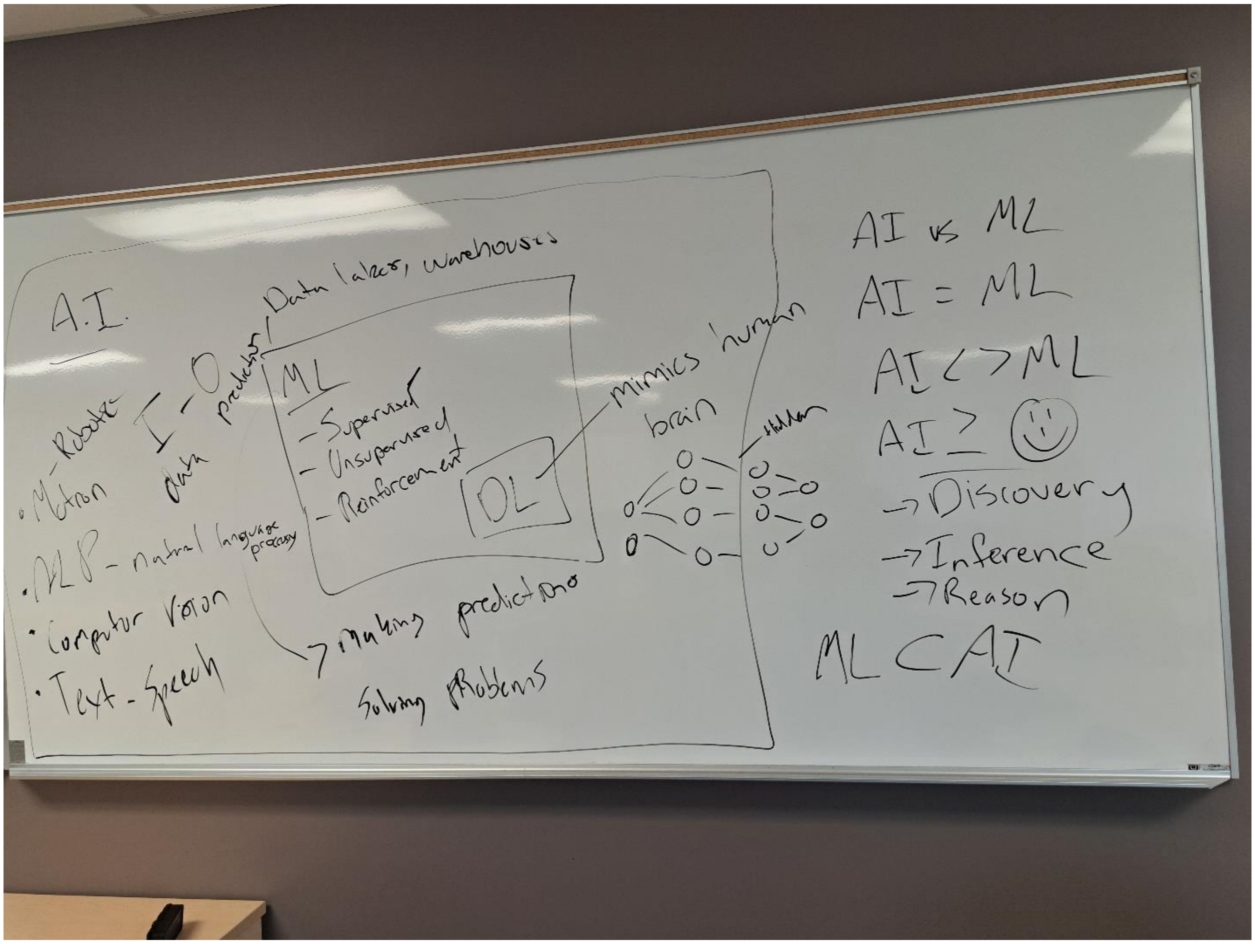
Day 5 Lecture Notes



Discussed specifications of Nvidia's Blackwell chip. Try drawing 208 billion dots in a 1.5inch x 1.5inch square. This is how many transistors are inside their latest chip. 40 quadrillion floating-point operations per second. Combined with the nvlink technology, which is 900Gigs/s. This means that traditional data centers with separate servers will now act as machine learning hives of interconnected parallel processing.

Please see "Importing a CSV & Shortcuts in Jupyter Notebooks – Code Lab." Machine Learning uses a lot of tabular data which needs to be processed & saved. Python makes this task relatively easy. 4

Day 1 lecture notes.



In day one we learn that Machine Learning is a subset of AI (as illustrated in the Venn diagram above). This means that Machine Learning is, in fact, artificial intelligence.

Does it include everything in A.I.? Short answer is no. The aim in AI is to ultimately make a model equivalent to a human, or even surpass its abilities (re: Turing Test). This involves discovery, inference, and reasoning.

Major fields inside A.I. include motion for robotics, computer vision, text-to-speech, and natural language processing (NLP), among others. These endeavours will overlap into ML.

Machine learning, in contrast, involves making predictions based off large volumes of data. It takes inputs into a trained model engine and outputs predictions. Ultimately, we are gleaning insights into problems.

ML can be subcategorized into four main facets:

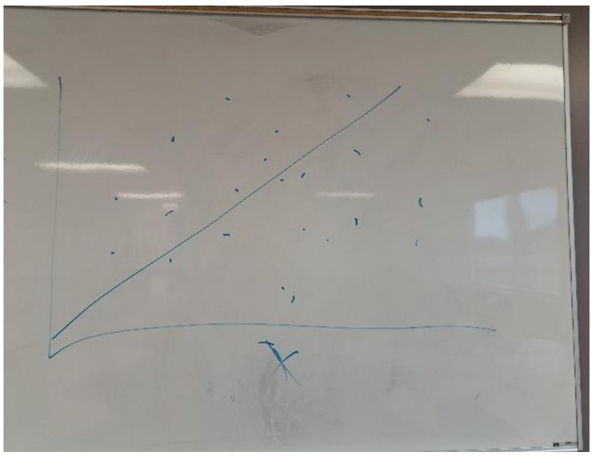
- Supervised learning – where data has labels and tackles regression or classification problems.
- Unsupervised learning – where we glean insights into unlabeled data using methods like clustering, decision trees, random forests, etc.
- Reinforcement learning – a highly optimized model achieved by many iterations and optimization loops.
- Deep learning- uses neural networks.

Deep learning involves neural networks that have many hidden processing layers. We learn more about this later in the course. This setup mimics the human brain's neurons and synapses, making it highly efficient and accurate.

When training models, it is typical to split the data into training and test sets. To maintain accuracy, convention is to split 80/20 or 70/30 (training/testing sets), respectively.

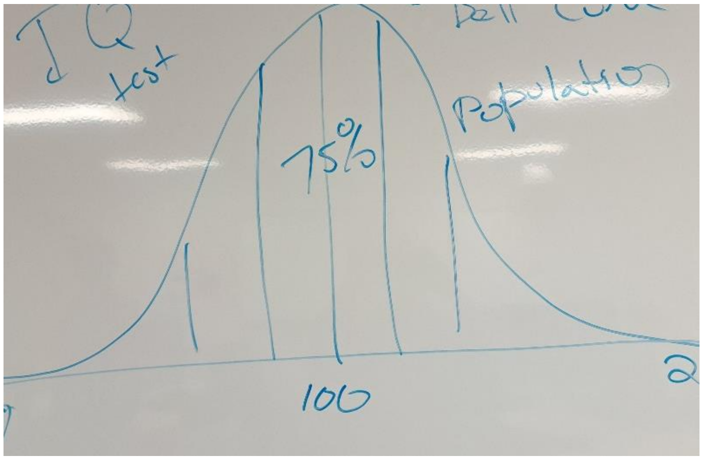
Day 2 Lecture Notes

The accuracy of a machine learning model's predictions can ultimately be influenced by the initial hypotheses we make, which are assertions or presumptions we believe to be true. It's crucial to carefully qualify and validate these assertions to ensure the reliability of the model's predictions.

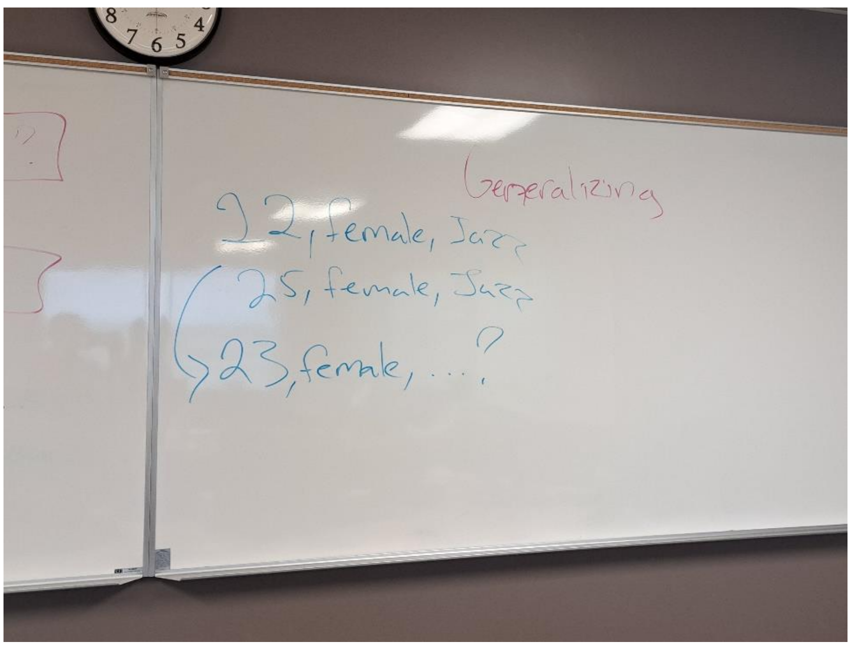


Regression tackles supervised learning problems. We typically have an independent and dependent variable. Independent variables are features like # of bedrooms, square feet, price, etc. We use these features or labels to predict the dependent variable (AKA target variable). In addition to linear regression, we learn about logistic regression (classification task) later in the course.

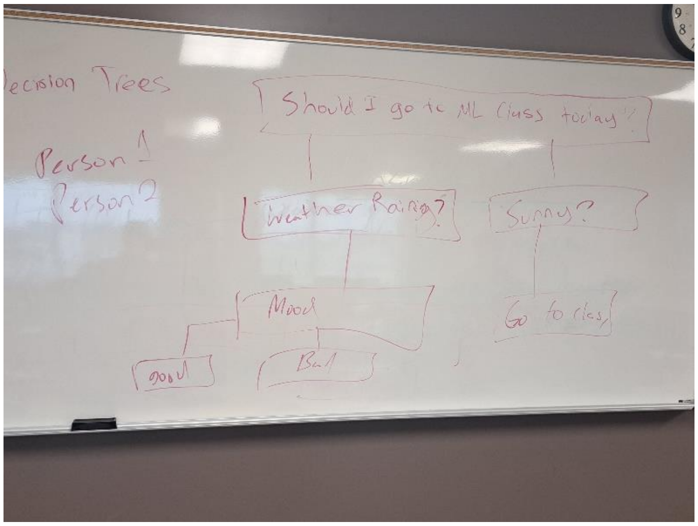
In clustering tasks (unsupervised learning with no labels), we can plot a decision boundary demarcating the data points into respective sets. This typically happens in 2dimensional space or arrays with x & y axes, but you can have multidimensional arrays (with a lot of features/labels).



Gaussian distribution holds that given a large enough dataset, values can be distilled into a bell-curve like distribution, where the bulk falls within the mean (standard deviation). We use the example of IQ test scores. It is statistically probable that most of the populations' IQ falls within this second quartile range. This holds true for a lot of data, which means we can apply it to ML to make predictions.



Generalization in ML means even if data is "unseen" (there exists no sample data to match to a target variable prediction), we can still have a prediction that is plausible. In the above example, despite having no sample data on 23-year-old females, we can predict from surrounding sample data that jazz is the favoured genre of music. This is the crux of generalization, albeit based upon a very simplistic hypothesis that would breakdown in most cases.



Decision-trees are perhaps the most intuitive of ML models or algorithms, as they are fundamentally IF-THEN control flow statements. Adding more branches (forks) or leaf nodes will bring considerably higher accuracy to the final prediction.

Underfitting occurs when a model is too simple to capture the underlying structure of the data.

Overfitting occurs when a model is too complex and learns to capture noise or random fluctuations in the training data.

Red Deer Polytechnic

Google Colab, Jupyter Notebook, & Kaggle

Module 1

by Todd Nash

Red Deer Polytechnic

kaggle

Google Colab, Jupyter Notebook, and Kaggle are all platforms/tools commonly used in the field of data science, machine learning, and scientific computing. Each serves specific purposes and has unique features.

Google Colaboratory

jupyter

by Todd Nash

Red Deer Polytechnic

Jupyter Notebook is an open-source, interactive web application that allows you to create and share documents that contain live code, equations, visualizations, and narrative text. It supports various programming languages, but it is widely used with Python.

Features:

- Interactive and exploratory data analysis with inline code execution.
- Supports multiple programming languages (Python, R, Julia, etc.).
- Rich support for visualizations using Matplotlib, Seaborn, and other libraries.
- Extensive ecosystem with extensions and integrations.

by Todd Nash

Red Deer Polytechnic

Google Colab:

A free, cloud-based platform provided by Google that allows users to write and execute Python code in a collaborative environment. It is built on top of Jupyter Notebooks and provides free access to GPU resources, making it suitable for machine learning tasks.

Features:

- Integration with Google Drive for easy storage and sharing of notebooks.
- Pre-installed libraries and access to GPU and TPU resources.
- Collaboration features, allowing multiple users to work on the same notebook simultaneously.

No setup required, as it runs entirely in the browser.

by Todd Nash

Red Deer Polytechnic

Kaggle is a platform for data science competitions, collaboration, and learning.

Features:

- Datasets: A vast collection of datasets available for exploration and analysis.
- Kernels: A cloud-based computational environment where you can write and execute code in Python or R, and share your analysis with others.
- Competitions: Hosts machine learning competitions with real-world datasets and problems.

by Todd Nash

Red Deer Polytechnic

Choosing the Right Platform:

Google Colab: Ideal for collaborative and lightweight tasks. Offers free GPU access.

Jupyter Notebook: Widely used for local development and interactive data analysis.

Kaggle: Great for participating in data science competitions, accessing datasets, and collaborative analysis.

by Todd Nash

Red Deer Polytechnic

Use Cases of Notebooks:

- Data Exploration: Notebooks are excellent for interactively exploring datasets, visualizing trends, and understanding data distributions.
- Data Analysis: Analysts and data scientists use notebooks to perform exploratory data analysis (EDA), statistical analysis, and hypothesis testing.
- Machine Learning: Notebooks are widely used for developing and testing machine learning models, documenting experiments, and sharing results.
- Education: Notebooks are used as educational tools to teach programming, data science, and machine learning.

Red Deer Polytechnic

Note:

GitHub has native support for rendering Jupyter Notebooks, which means you can view them right on their website. Developers will always want to use git for version control.

by Todd Nash

Introduction to Machine Learning

- Machine Learning (ML) is an automation process of extracting information from a data set, with the goal of predicting future data streams.
- Two important types of learning for scientific applications are supervised and unsupervised learning. In unsupervised learning:
    - The idea is to find previously unknown patterns in the data.
    - An example of this is the Support Vector Machine, or SVM, which often seeks to classify data into subsets corresponding to previously unknown patterns.
- In supervised learning, the idea is to develop a method that learns from a training data set and predicts the type of behavior from future data.

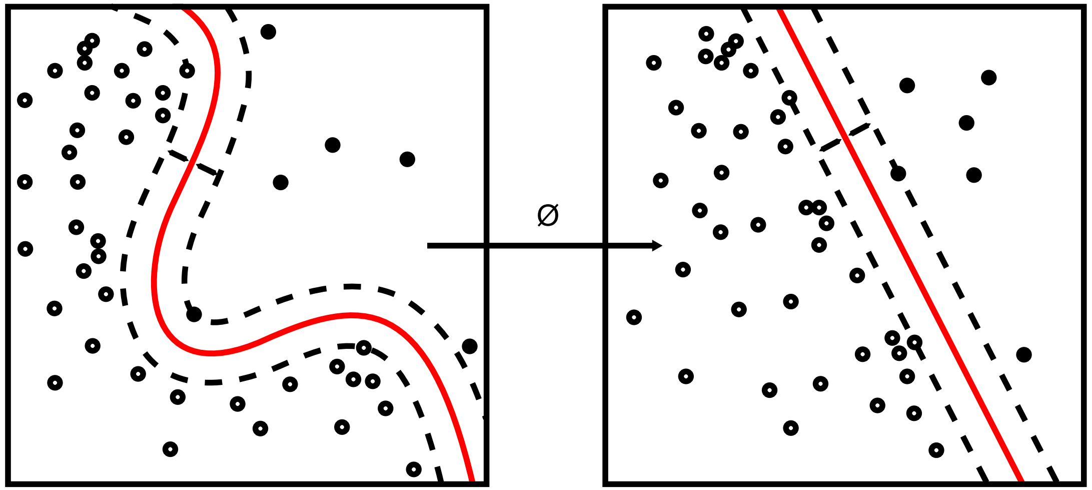
Machine Learning: Examples

	Supervised Learning	Unsupervised Learning
Goal	A program that performs as well as humans	To find structure in the data
Task	Well defined (e.g., a target function to constrain results)	Not well defined or could be undefined
Past Experience	Training data set provided by a human	None supplied
Performance	Error/accuracy with respect to the task	Unable to evaluate
Subcategories	Classification and Regression	Clustering/Dimensional Reduction (including feature extraction/selection)

Cost Function, Fitness Biology, Free Energy

Introduction to Machine Learning

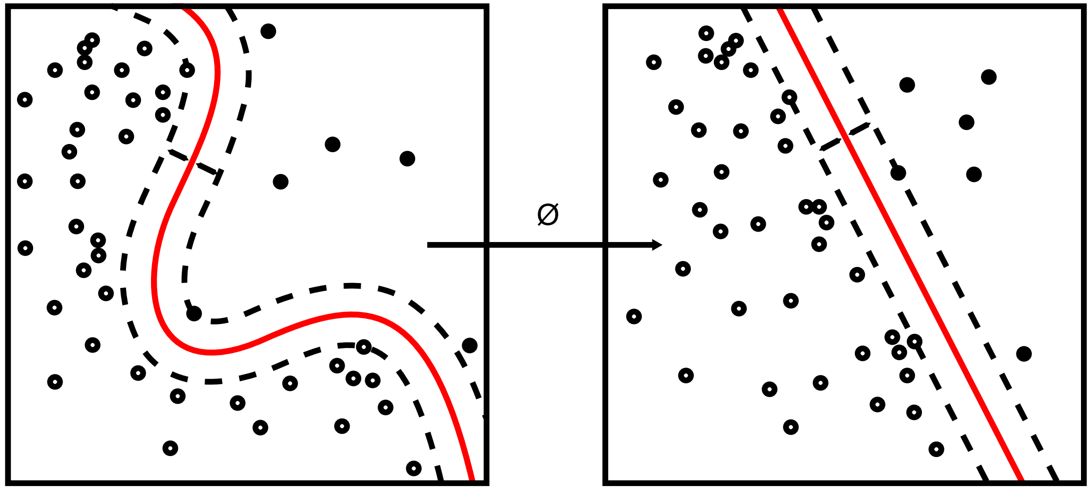
- Machine learning (ML) is also the study of computer algorithms that can improve automatically through experience and by the use of data.
- It is seen as a part of artificial intelligence.
- Machine learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so.
- Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, speech recognition, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.



- Examples of from a linear support vector machine and its learned decision boundary.
- The width of the dashed lines is termed the "margin".

Introduction to Machine Learning

- A subset of machine learning is closely related to computational statistics, which focuses on making predictions using computers
- Not all machine learning is statistical learning.
- The study of mathematical optimization involves methods, theory and applications to the field of machine learning.
- Data mining is a related field of study, focusing on exploratory data analysis through unsupervised learning.
- Some implementations of machine learning use data and neural networks in a way that mimics the working of a biological brain.
- In its application across science and business problems, machine learning is also referred to as predictive analytics.



- Examples of from a linear support vector machine and its learned decision boundary.
- The width of the dashed lines is termed the "margin".

Applications

- Agriculture
- Anatomy
- Adaptive website
- Affective computing
- Astronomy
- Banking
- Bioinformatics
- Brain–machine interfaces
- Cheminformatics
- Citizen science
- Computer networks
- Computer vision
- Credit-card fraud detection
- Data quality
- DNA sequence classification
- Economics
- Financial market analysis[75]
- General game playing
- Handwriting recognition
- Information retrieval
- Insurance
- Internet fraud detection
- Knowledge graph embedding
- Linguistics
- Machine learning control
- Machine perception
- Machine translation
- Marketing
- Medical diagnosis
- Natural language processing
- Natural language understanding
- Online advertising
- Optimization
- Recommender systems
- Robot locomotion
- Search engines
- Sentiment analysis
- Sequence mining
- Software engineering
- Speech recognition
- Structural health monitoring
- Syntactic pattern recognition
- Telecommunication
- Theorem proving
- Time-series forecasting
- User behavior analytics
- Behaviorism

Overview

- Learning algorithms work on the basis that strategies, algorithms, and inferences that worked well in the past are likely to continue working well in the future.
- These inferences can be obvious, such as "since the sun rose every morning for the last 10,000 days, it will probably rise tomorrow morning as well". (Except: Black Swans)
- They can be nuanced, such as "X% of families have geographically separate species with color variants, so there is a Y% chance that undiscovered variations exist".
- Machine learning programs can perform tasks without being explicitly programmed to do so.
- It involves computers learning from data provided so that they carry out certain tasks.
- For simple tasks assigned to computers, it is possible to program algorithms telling the machine how to execute all steps required to solve the problem at hand
- For more advanced tasks, it can be challenging for a human to manually create the needed algorithms.
- In practice, it can turn out to be more effective to help the machine develop its own algorithm, rather than having human programmers specify every needed step.

Overview

- The discipline of machine learning employs various approaches to teach computers to accomplish tasks where no fully satisfactory algorithm is available.
- In cases where vast numbers of potential answers exist, one approach is to label some of the correct answers as valid with a label or target variable Y, then used these in a supervised learning application.
- The labels are also considered to be the target variable, i.e., the result that is to be learned
- Often labels are paired with a data vector, usually termed a feature vector X
- This can then be used as training data for the computer to improve the algorithm(s) it uses to determine correct answers.
- In other applications, unsupervised learning can be carried out, for example in problems involving grouping data into clusters

Main Point:

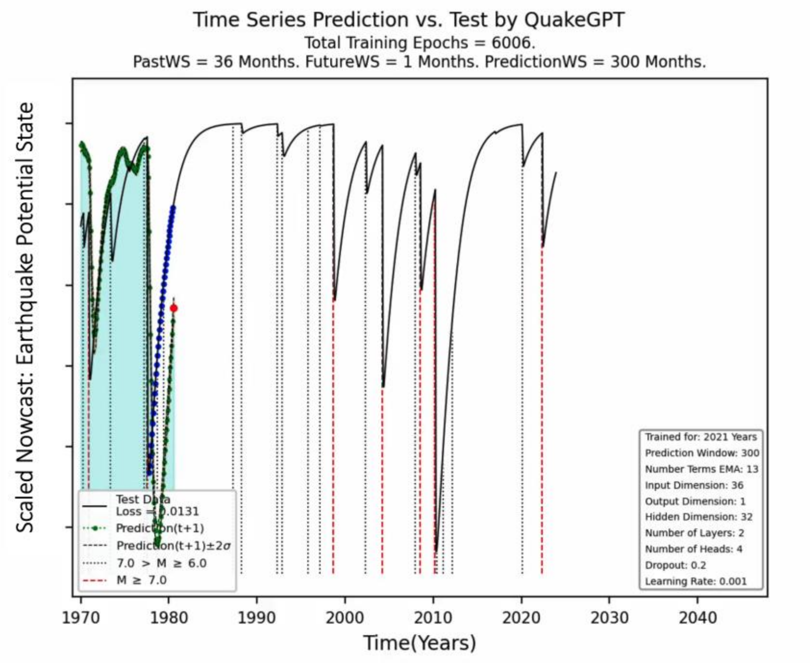
X = Feature Vector

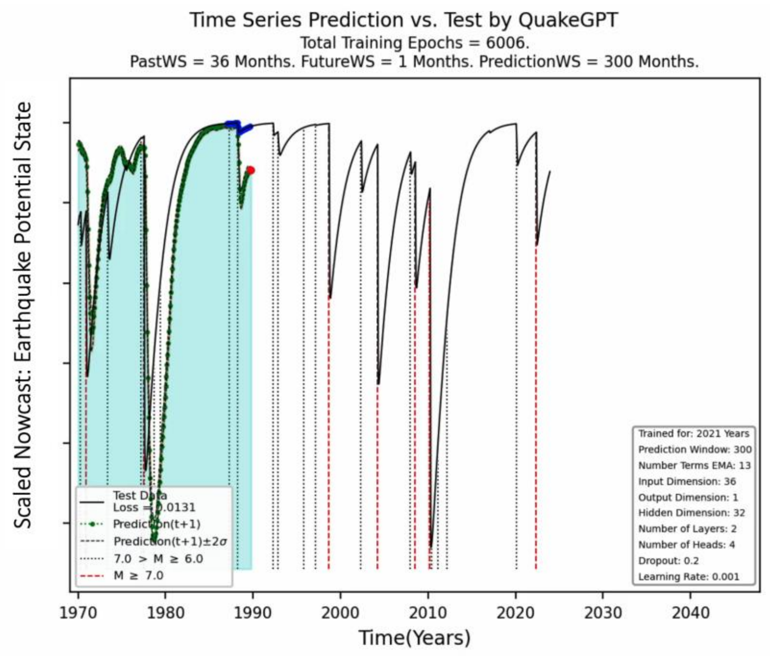
Y = Label or Target Variable

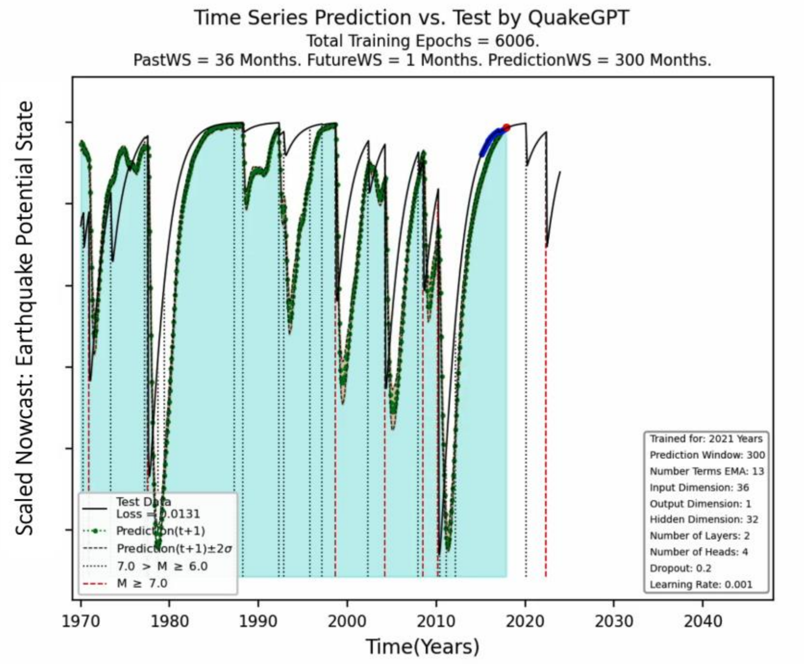
Examples of Feature Vectors and Targets for Time Series Forecasting

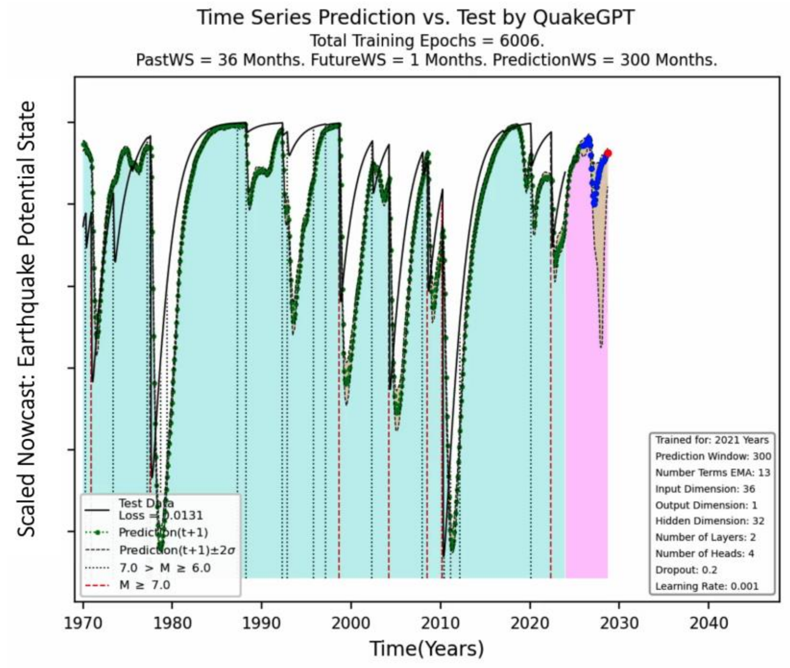
Feature Vector in Blue.

Target Value in Red









History

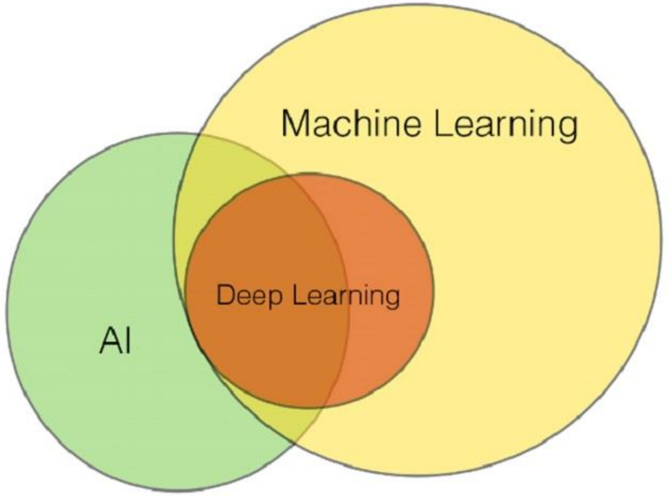
- The term machine learning was coined in 1959 by Arthur Samuel, an American IBM employee and pioneer in the field of computer gaming and artificial intelligence.
- Also the synonym self-teaching computers was used in this time period.
- A representative book of the machine learning research during the 1960s was Nilsson's book on Learning Machines, dealing mostly with machine learning for pattern classification.
- Current book by Andriy Bukov, “The Hundred Page Machine Learning Book”
- Interest related to pattern recognition continued into the 1970s, as described by Duda and Hart in 1973.
- In 1981 a report was given on using teaching strategies so that a neural network learns to recognize 40 characters (26 letters, 10 digits, and 4 special symbols) from a computer terminal.

History

- Tom M. Mitchell provided a widely quoted, more formal definition of the algorithms studied in the machine learning field:
- "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."
- This definition of the tasks in which machine learning is concerned offers a fundamentally operational definition rather than defining the field in cognitive terms.
- Modern day machine learning has two objectives:
  - One is to classify data based on models which have been developed
  - The other purpose is to make predictions for future outcomes based on these models.
- A hypothetical algorithm specific to classifying data may use computer vision of skin moles coupled with supervised learning in order to train it to classify the cancerous moles.
- Whereas, a machine learning algorithm for stock trading may inform the trader of future potential predictions.

Artificial Intelligence: History

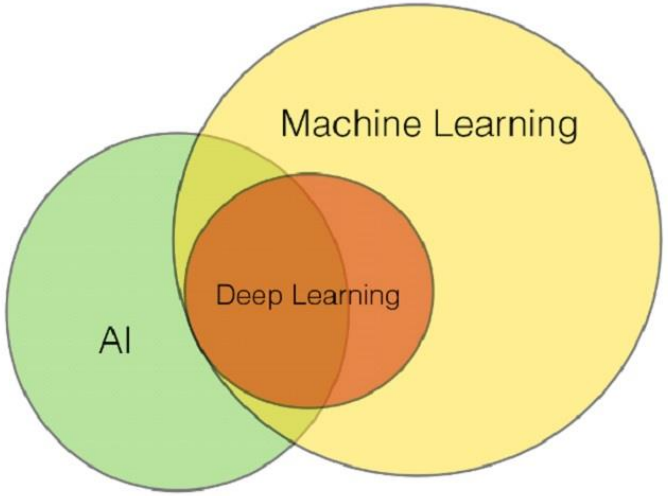
- As a scientific endeavor, machine learning grew out of the quest for artificial intelligence.
- In the early days of AI as an academic discipline, some researchers were interested in having machines learn from data.
- They attempted to approach the problem with various symbolic methods, as well as what was then termed "neural networks"
- These were mostly perceptrons and other models that were later found to be reinventions of the generalized linear models of statistics.
- Probabilistic reasoning was also employed, especially in automated medical diagnosis.
- However, an increasing emphasis on the logical, knowledge-based approach caused a rift between AI and machine learning.



Part of machine learning as subfield of AI or part of AI as subfield of machine learning

Artificial Intelligence: History

- Probabilistic systems were plagued by theoretical and practical problems of data acquisition and representation.
- By 1980, expert systems had come to dominate AI, and statistics was out of favor.
- Work on symbolic/knowledge-based learning did continue within AI, leading to inductive logic programming, but the more statistical line of research was now outside the field of AI proper, in pattern recognition and information retrieval.
- Neural networks research had been abandoned by AI and computer science around the same time.
- This line, too, was continued outside the AI/CS field, as "connectionism", by researchers from other disciplines including Hopfield (Hopfield-Tank algorithm), Rumelhart and Hinton.
- Their main success came in the mid-1980s with the reinvention of backpropagation.



Part of machine learning as subfield of AI or part of AI as subfield of machine learning

“Artificial Intelligence” vs. “Machine Learning”

- Machine learning (ML), reorganized as a separate field, started to flourish in the 1990s.
- The field changed its goal from achieving artificial intelligence to tackling solvable problems of a practical nature.
- It shifted focus away from the symbolic approaches it had inherited from AI, and toward methods and models borrowed from statistics and probability theory.
- The difference between ML and AI is frequently misunderstood.
- ML learns and predicts based on passive observations, whereas AI implies an agent interacting with the environment (you) to learn and take actions that maximize its chance of successfully achieving its (or your) goals.
- Others have the view that not all ML is part of AI, but only an 'intelligent subset' of ML should be considered AI.

“Data Mining” vs. “Machine Learning”

- Machine learning and data mining often employ the same methods and overlap significantly
- Machine learning focuses on prediction, based on known properties learned from the training data
- Data mining focuses on the discovery of (previously) unknown properties in the data (this is the analysis step of knowledge discovery in databases).
- Data mining uses many machine learning methods, but with different goals
- Machine learning employs data mining methods as "unsupervised learning" or as a preprocessing step to improve learner accuracy.
- In machine learning, performance is usually evaluated with respect to the ability to reproduce known knowledge (target or label variable), while in knowledge discovery and data mining (KDD) the key task is the discovery of previously unknown knowledge.
- Evaluated with respect to known knowledge, an uninformed (unsupervised) method will easily be outperformed by other supervised methods, while in a typical KDD task, supervised methods cannot be used due to the unavailability of training data.

Optimization and Generalization

- Machine learning also has intimate ties to optimization
- Many learning problems are formulated as minimization of some loss function on a training set of examples.
- Loss functions express the discrepancy between the predictions of the model being trained and the actual problem instances
- For example, in classification, one wants to assign a label to instances, and models are trained to correctly predict the pre-assigned labels of a set of examples.

Optimization
(supervised)

- The difference between optimization and machine learning arises from the goal of generalization
- While optimization algorithms can minimize the loss on a training set, machine learning is concerned with minimizing the loss on unseen samples.
- Characterizing the generalization of various learning algorithms is an active topic of current research, especially for deep learning algorithms.

Generalization or Classification
(unsupervised)

“Statistics” vs. “Machine Learning”

- Machine learning and statistics are closely related fields in terms of methods, but distinct in their principal goal
- Statistics draws population inferences from a sample, while machine learning finds generalizable predictive patterns.
- According to some, the ideas of machine learning, from methodological principles to theoretical tools, have had a long pre-history in statistics.
- They also suggested the term data science as a placeholder to call the overall field.
- Others distinguished two statistical modeling paradigms, data model and algorithmic model, wherein "algorithmic model" means machine learning algorithms like Random Forest.
- Some statisticians have adopted methods from machine learning, leading to a combined field that they call statistical learning.

Learning

- A core objective of a learner is to generalize from its experience.
- Generalization in this context is the ability of a learning machine to perform accurately on new, unseen examples/tasks after having experienced a learning data set.
- The training examples come from some generally unknown probability distribution (considered representative of the space of occurrences)
- The learner has to build a general model about this space that enables it to produce sufficiently accurate predictions in new cases.
- The computational analysis of machine learning algorithms and their performance is a branch of theoretical computer science known as computational learning theory.
- Because training sets are finite and the future is uncertain, learning theory usually does not yield guarantees of the performance of algorithms.
- Instead, probabilistic bounds on the performance are quite common.
- The bias–variance decomposition is one way to quantify generalization error.

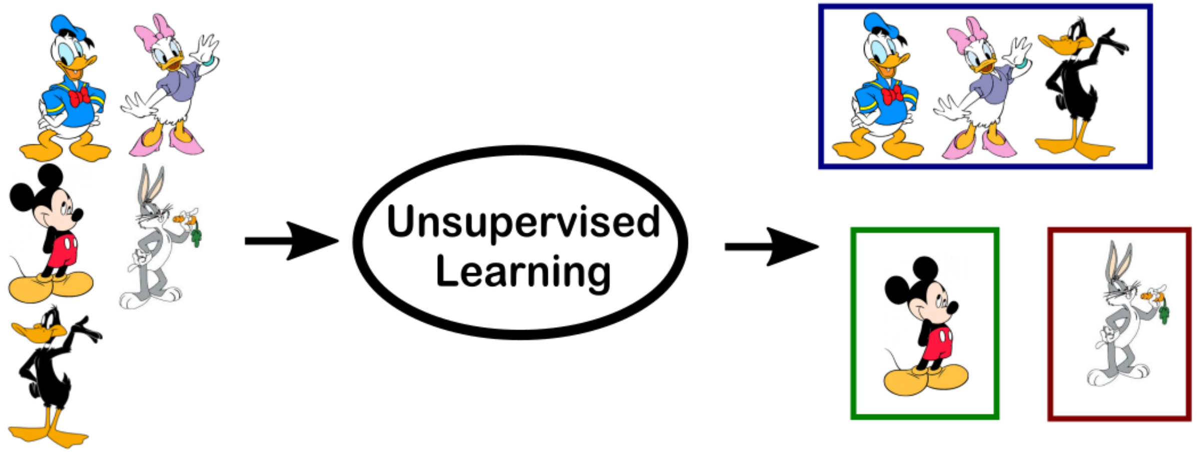
Complexity

- For the best performance in the context of generalization, the complexity of the hypothesis should match the complexity of the function underlying the data.
- If the hypothesis is less complex than the function, then the model has under fitted the data.
- If the complexity of the model is increased in response, then the training error decreases.
- But if the hypothesis is too complex, then the model is subject to overfitting and generalization will be poorer.
- In addition to performance bounds, learning theorists study the time complexity and feasibility of learning.
- In computational learning theory, a computation is considered feasible if it can be done in polynomial time.
- An algorithm is said to be of polynomial time if its running time is upper bounded by a polynomial expression in the size of the input for the algorithm, that is, $T(n) = O(n^k)$ for some positive constant k
- There are two kinds of time complexity results:
    - Positive results show that a certain class of functions can be learned in polynomial time.
    - Negative results show that certain classes cannot be learned in polynomial time

Approaches

- Machine learning approaches are traditionally divided into three broad categories, depending on the nature of the "signal" or "feedback" available to the learning system:
  - Supervised learning: The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs.
  - Unsupervised learning: No labels are given to the learning algorithm, leaving it on its own to find structure in its input. => Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning).
  - Reinforcement learning: A computer program interacts with a dynamic environment in which it must perform a certain goal (such as driving a vehicle or playing a game against an opponent). => As it navigates its problem space, the program is provided feedback that's analogous to rewards, which it tries to maximize.

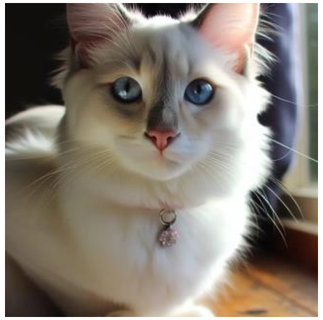
Machine Learning: Supervised and Unsupervised



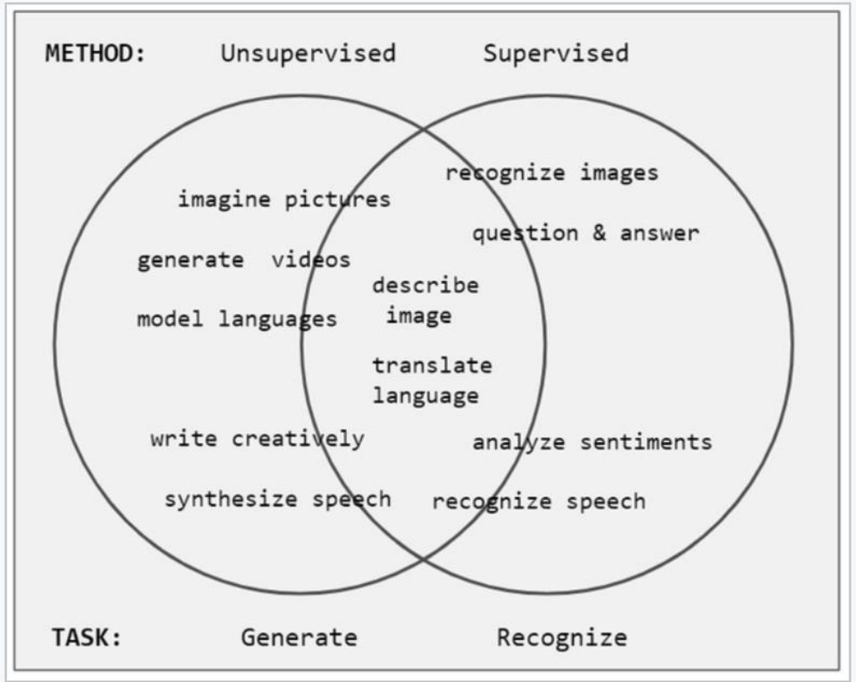
	Supervised Learning	Unsupervised Learning
Goal	A program that performs as well as humans	To find structure in the data
Task	Well defined (e.g., a target function to constrain results)	Not well defined or could be undefined
Past Experience	Training data set provided by a human	None supplied
Performance	Error/accuracy with respect to the task	Unable to evaluate
Subcategories	Classification and Regression	Clustering/Dimensional Reduction (including feature extraction/selection)

Types of Tasks

Cat image generated by DALL-E from Openai.com
Example of Generative Modeling



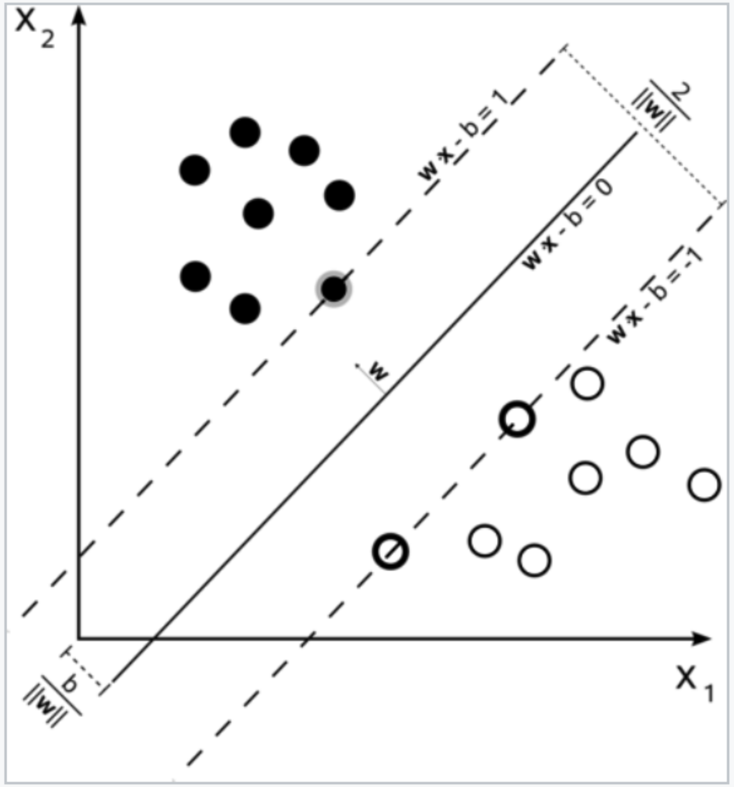
- Generative modeling is an unsupervised learning task in machine learning that involves automatically discovering and learning the regularities or patterns in input data in such a way that the model can be used to generate or output new examples that plausibly could have been drawn from the original dataset.
- Discriminative models, also referred to as conditional models, are a class of logistical models used for classification or regression. They distinguish decision boundaries through observed data, such as pass/fail, win/lose, alive/dead or healthy/sick.
    - Typical generative model approaches include naive Bayes classifiers, Gaussian mixture models, variational autoencoders, generative adversarial networks and others.
    - Typical discriminative models include logistic regression (LR), conditional random fields (CRFs) (specified over an undirected graph), decision trees, and many others.



Tendency for a task to employ Supervised vs. Unsupervised methods. The separation can be blurred.

Supervised Learning

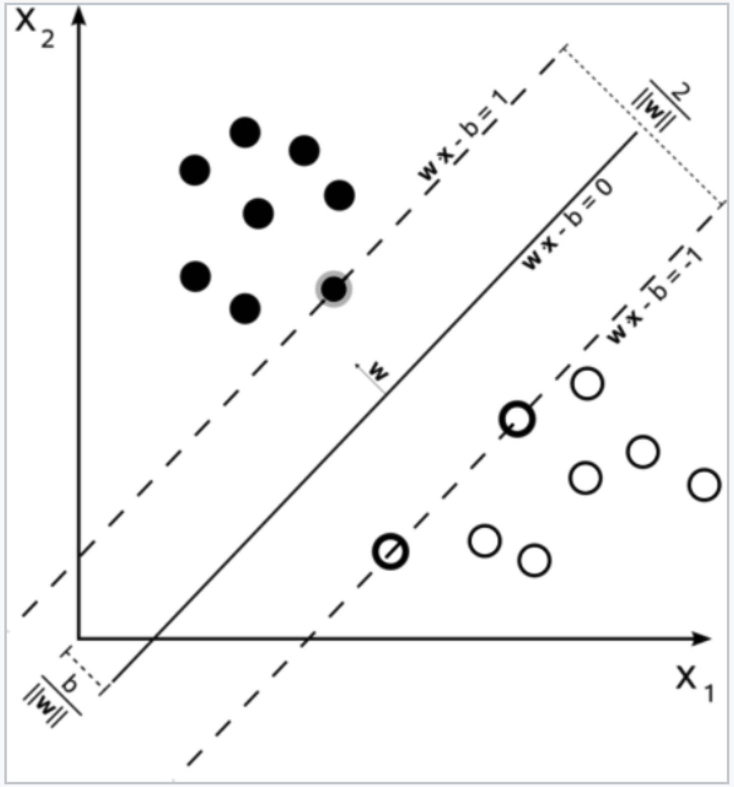
- Supervised learning algorithms build a mathematical model of a set of data that contains both the inputs and the desired outputs.
- The data is known as training data and consists of a set of training examples.
- Each training example has one or more inputs and the desired output, also known as a supervisory signal.
- In the mathematical model, each training example is represented by an array or vector X, sometimes called a feature vector, and the training data is represented by a matrix.
- Desired output is represented by a label or target Y
- Through iterative optimization of an objective function, supervised learning algorithms learn a function that can be used to predict the output associated with new inputs.
- An optimal function will allow the algorithm to correctly determine the output for inputs that were not a part of the training data.
- An algorithm that improves the accuracy of its outputs or predictions over time is said to have learned to perform that task.



A support-vector machine is a supervised learning model that divides the data into regions separated by a linear boundary. Here, the linear boundary divides the black circles from the white.

Supervised Learning

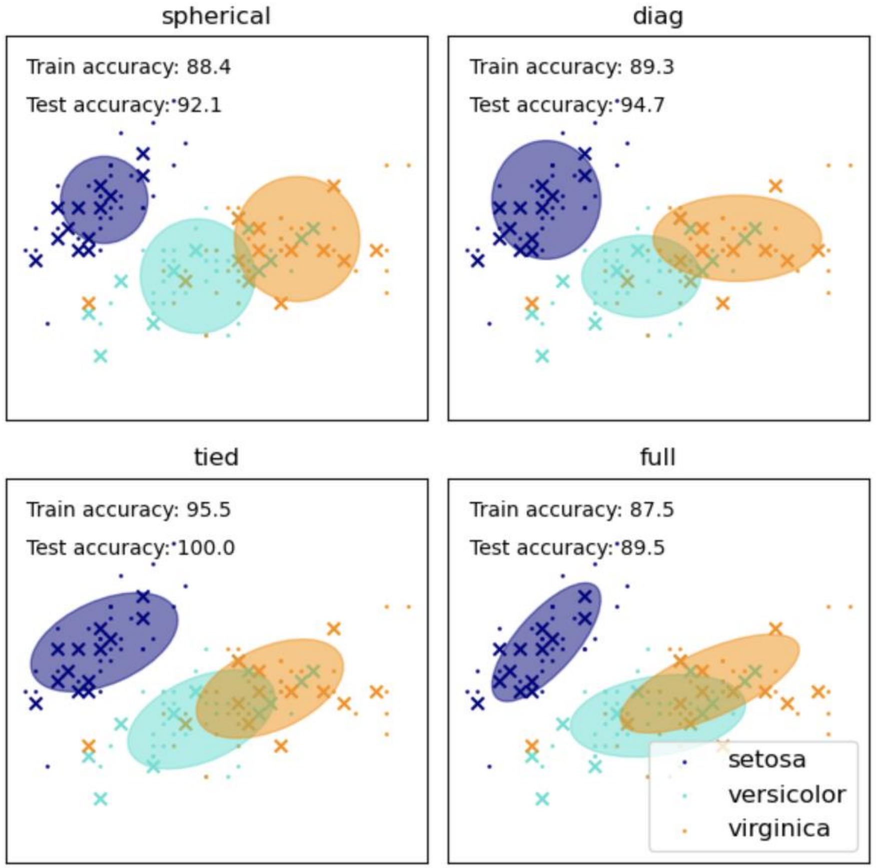
- Types of supervised learning algorithms include active learning, optimization, classification and regression (fitting a line to data).
- Classification algorithms are used when the outputs are restricted to a limited set of values, and regression algorithms are used when the outputs may have any numerical value within a range.
- As an example, for a classification algorithm that filters emails, the input would be an incoming email, and the output would be the name of the folder in which to file the email.
- Similarity learning is an area of supervised machine learning closely related to regression and classification, but the goal is to learn from examples using a similarity function that measures how similar or related two objects are.
- It has applications in ranking, recommendation systems, visual identity tracking, face verification, and speaker verification.



A support-vector machine is a supervised learning model that divides the data into regions separated by a linear boundary. Here, the linear boundary divides the black circles from the white.

Unsupervised Learning

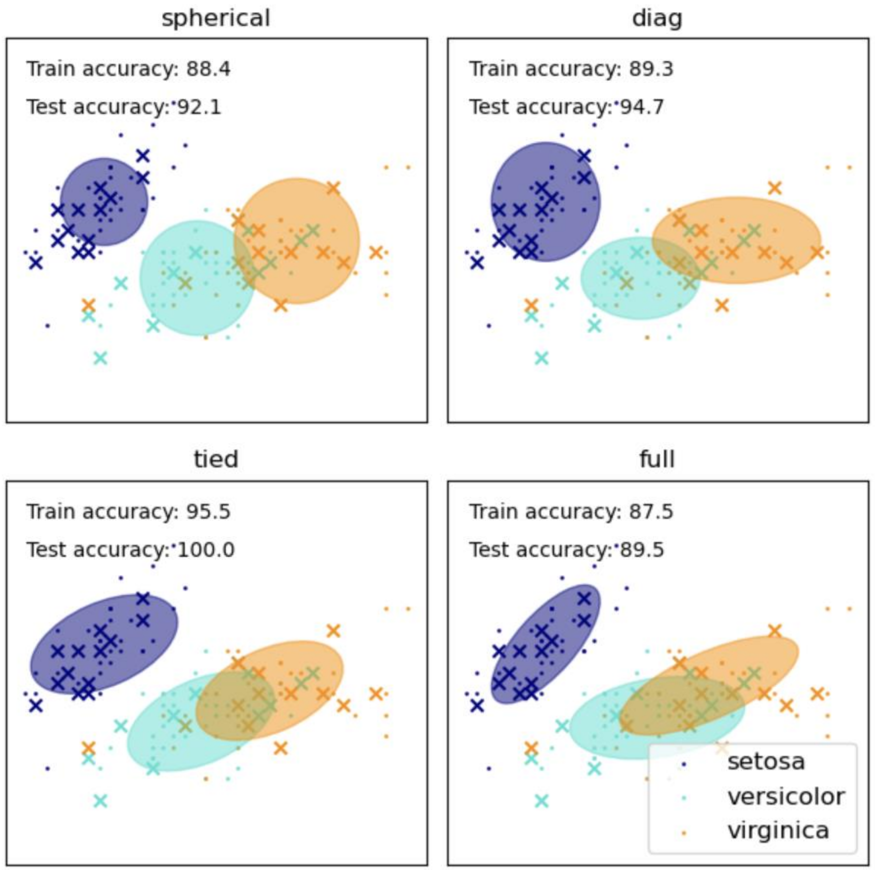
- Unsupervised learning algorithms take a set of data that contains only inputs, and find structure in the data, like grouping or clustering of data points.
- The algorithms, therefore, learn from test data that has not been labeled, classified or categorized.
- Unsupervised learning algorithms must first self-discover any naturally occurring patterns in that training data set
- Instead of responding to feedback, unsupervised learning algorithms identify commonalities in the data and react based on the presence or absence of such commonalities in each new piece of data.
- A central application of unsupervised learning is in the field of density estimation in statistics, such as finding the probability density function.
- Though unsupervised learning encompasses other domains involving summarizing and explaining data features.



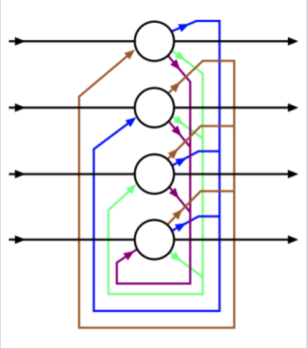
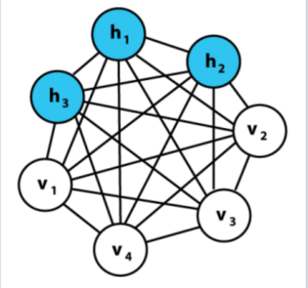
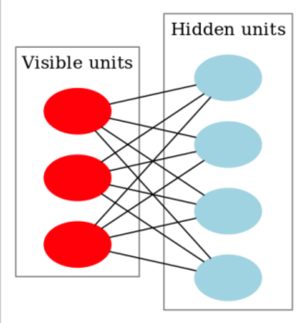
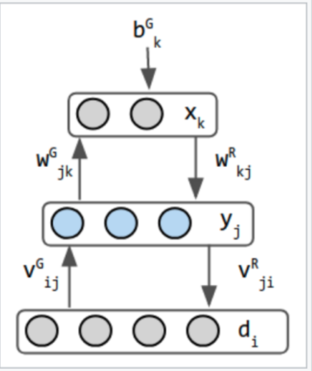
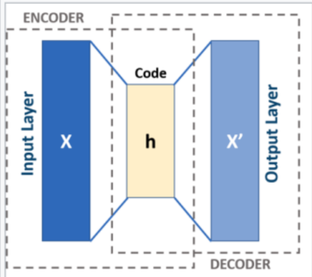
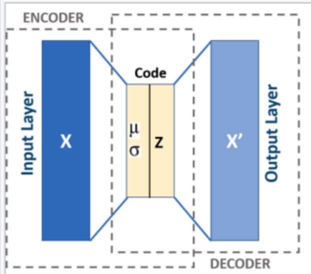
Unsupervised learning in a Gaussian Mixture Model

Unsupervised Learning: Cluster Analysis

- Cluster analysis is the assignment of a set of observations into subsets (called clusters) so that observations within the same cluster are similar according to one or more predesignated criteria, while observations drawn from different clusters are dissimilar.
- Different clustering techniques make different assumptions on the structure of the data, often defined by some similarity metric and evaluated, for example, by internal compactness, or the similarity between members of the same cluster, and separation, the difference between clusters.
- Other methods are based on estimated density and graph connectivity.



Supervised Learning: Examples of Neural Networks

Hopfield	Boltzmann	RBM	Helmholtz	Autoencoder	VAE
 A network based on magnetic domains in iron with a single self-connected layer.	 2 layers. Uses symmetric 2-way weights. Following Boltzmann's thermodynamics, individual probabilities give rise to macroscopic energies.	 Restricted Boltzmann Machine. This is a Boltzmann machine where lateral connections within a layer are prohibited to make analysis tractable.	 Instead of the bidirectional symmetric connection of a Boltzmann machine, we have separate one-way connections to form a loop. It does both generation and discrimination.	 A feed forward network that aims to find a good middle layer representation of its input world.	 Applies Variational Inference to the Autoencoder. The middle layer is a set of means & variances for Gaussian distributions.

In Boltzmann machines, Energy plays the role of the Cost function. An energy function is a macroscopic measure of a network's state. This analogy with physics is inspired by Ludwig Boltzmann's analysis of a gas' macroscopic energy from the microscopic probabilities of particle motion $p \propto e^{E/kT}$, where k is the Boltzmann constant and T is temperature. In the RBM network the relation is $p = e^{-E} / Z$,[2] where p & E vary over every possible activation pattern and $Z = \sum_{AllPatterns} e^{-E(pattern)}$. To be more precise, $p(a) = e^{-E(a)} / Z$, where a is an activation pattern of all neurons (visible and hidden). Hence, early neural networks bear the name Boltzmann Machine. Paul Smolensky calls -E the Harmony. A network seeks low energy which is high Harmony.

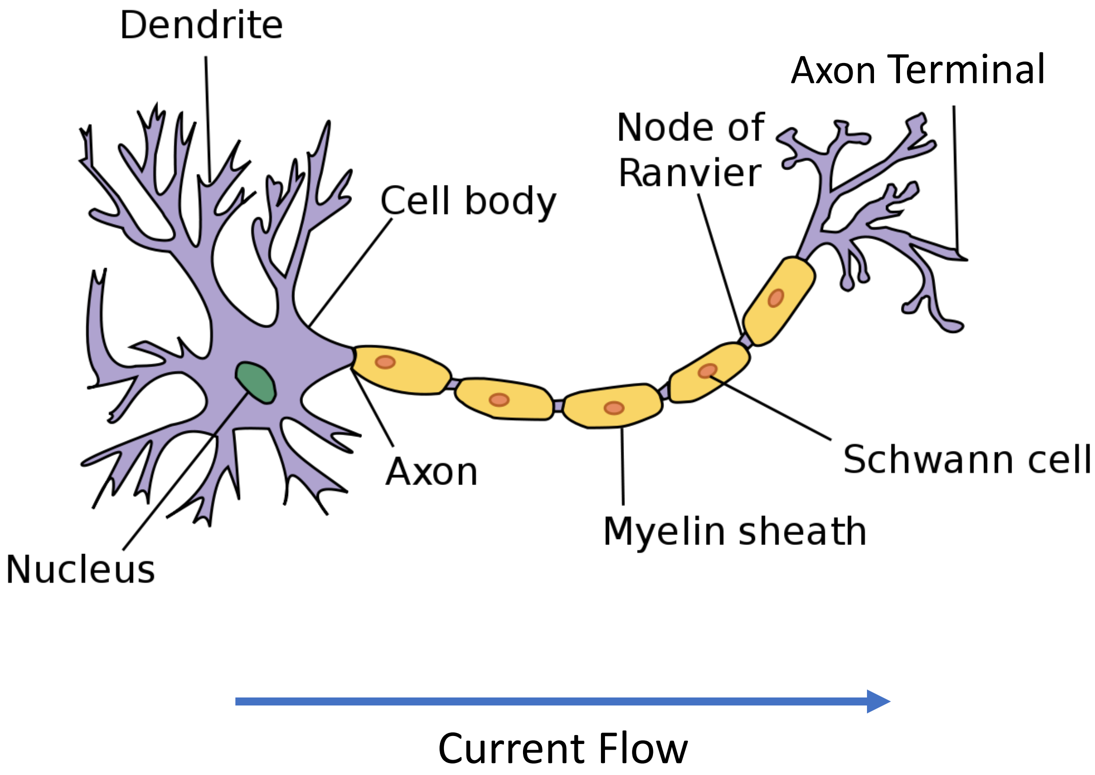
History of Neural Network Modeling

1969	Perceptrons by Minsky & Papert shows a perceptron without hidden layers fails on XOR
1970s	(approximate dates) AI winter I
1974	Ising magnetic model proposed by WA Little for cognition
1980	Fukushima introduces the neocognitron, which is later called a convolution neural network. It is mostly used in SL, but deserves a mention here.
1982	Ising variant Hopfield net described as CAMs and classifiers by John Hopfield.
1983	Ising variant Boltzmann machine with probabilistic neurons described by Hinton & Sejnowski following Sherington & Kirkpatrick's 1975 work.
1986	Paul Smolensky publishes Harmony Theory, which is an RBM with practically the same Boltzmann energy function. Smolensky did not give an practical training scheme. Hinton did in mid-2000s
1995	Schmidthuber introduces the LSTM neuron for languages.
1995	Dayan & Hinton introduces Helmholtz machine
1995-2005	(approximate dates) AI winter II
2013	Kingma, Rezende, & co. introduced Variational Autoencoders as Bayesian graphical probability network, with neural nets as components.

A Biological Neuron

Ended Here 3/4/2024

- The way an actual neuron works involves the accumulation of electric charge, which when exceeding a particular value causes the pre-synaptic neuron to discharge across the axon and stimulate the post-synaptic neuron.
- Humans have billions of neurons (10^{10} to 10^{11}) which are interconnected and can produce incredibly complex firing patterns.
- A typical neuron is connected to 10,000 others by means of synaptic junctions
- The capabilities of the human brain are incredible compared to what we can do even with state-of-the-art neural networks.
- We will likely not see neural networks mimicking the function of the human brain anytime soon.



Semi-Supervised Learning

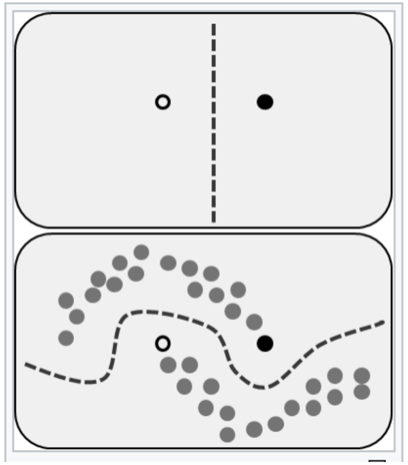
Semi-supervised learning falls between unsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data).

Some of the training examples are missing training labels

Yet unlabeled data, when used in conjunction with a small amount of labeled data, can produce a considerable improvement in learning accuracy.

In weakly supervised learning, the training labels are noisy, limited, or imprecise

However, these labels are often cheaper to obtain, resulting in larger effective training sets.



- An example of the influence of unlabeled data in semi-supervised learning.
- The top panel shows a decision boundary we might adopt after seeing only one positive (white circle) and one negative (black circle) example.
- The bottom panel shows a decision boundary we might adopt if, in addition to the two labeled examples, we were given a collection of unlabeled data (gray circles).
- This could be viewed as performing clustering and then labeling the clusters with the labeled data, pushing the decision boundary away from high-density regions, or learning an underlying one-dimensional manifold where the data reside.

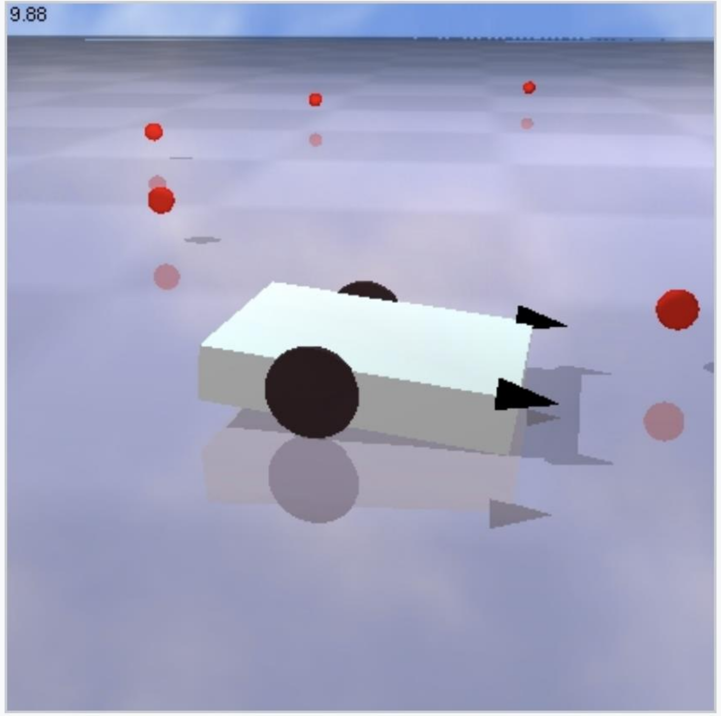
Reinforcement Learning

- Reinforcement learning is an area of machine learning concerned with how software agents ought to take actions in an environment so as to maximize some notion of cumulative reward.
- Due to its generality, the field is studied in many other disciplines, such as game theory, control theory, operations research, information theory, simulation-based optimization, multi-agent systems, swarm intelligence, statistics and genetic algorithms.
- In machine learning, the environment is typically represented as a Markov Decision Process (MDP).
- MDP provides a mathematical framework for modeling decision making in situations where outcomes are partly random and partly under the control of a decision maker, and is useful in optimization problems
- Reinforcement learning algorithms do not assume knowledge of an exact mathematical model of the MDP and are used when exact models are infeasible.
- Reinforcement learning algorithms are used in autonomous vehicles or in learning to play a game against a human opponent such as Chess or Go

Artificial Life

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

- Artificial life (often abbreviated ALife or A-Life) is a field of study wherein researchers examine systems related to natural life, its processes, and its evolution, through the use of simulations with computer models, robotics, and biochemistry.
- The discipline was named by Christopher Langton, an American theoretical biologist, in 1986.
- In 1987 Langton organized the first conference on the field, in Los Alamos, New Mexico
- There are three main kinds of Alife, named for their approaches:
  - Soft, from software
  - Hard, from hardware
  - Wet, from biochemistry
- Artificial life researchers study traditional biology by trying to recreate aspects of biological phenomena.

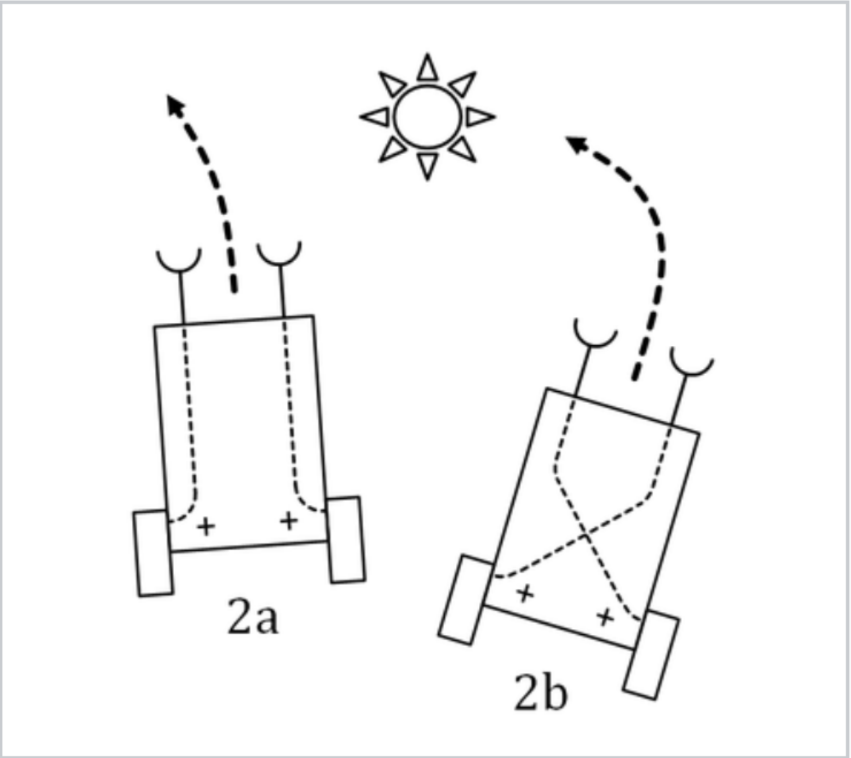


A Braitenberg vehicle simulation, programmed in breve, an artificial life simulator

https://en.wikipedia.org/wiki/Braitenberg_vehicle
(see Video there)

Braitenberg Vehicle
https://en.wikipedia.org/wiki/Braitenberg_vehicle

- A Braitenberg vehicle is an agent that can autonomously move around based on its sensor inputs.
- It has primitive sensors that measure some stimulus at a point, and wheels (each driven by its own motor) that function as actuators or effectors.
- In the simplest configuration, a sensor is directly connected to an effector, so that a sensed signal immediately produces a movement of the wheel.
- Depending on how sensors and wheels are connected, the vehicle exhibits different behaviors (which can be goal-oriented).
- This means that, depending on the sensor-motor wiring, it appears to strive to achieve certain situations and to avoid others, changing course when the situation changes.



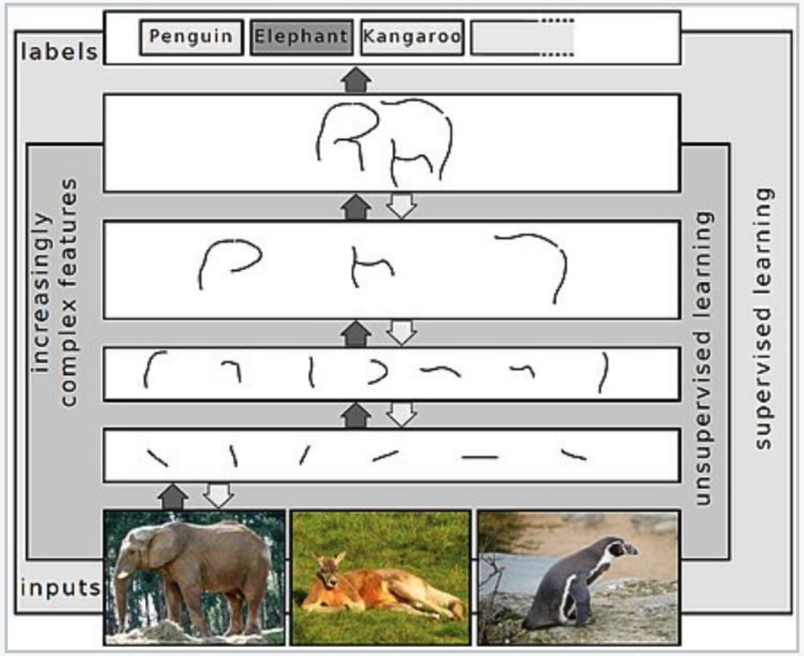
Vehicles 2a, 2b

Deep Learning
https://en.wikipedia.org/wiki/Deep_learning

- Deep learning (also known as deep structured learning) is part of a broader family of machine learning methods based on artificial neural networks.
- The adjective "deep" in deep learning refers to the use of multiple hidden layers in the network.
- By contrast, “wide shallow” networks have few hidden layers (perceptrons) but many neurons in those layers
- Deep learning is a modern variation which is concerned with an unbounded number of layers of bounded size, which permits practical application and optimized implementation, while retaining theoretical universality under mild conditions.
- In deep learning the layers are also permitted to be heterogeneous and to deviate widely from biologically informed models
- Deep-learning architectures such as deep neural networks, deep belief networks, deep reinforcement learning, recurrent neural networks and convolutional neural networks have been applied to fields including computer vision, speech recognition, natural language processing, machine translation, bioinformatics, drug design, medical image analysis, material inspection and board game programs

Deep Learning

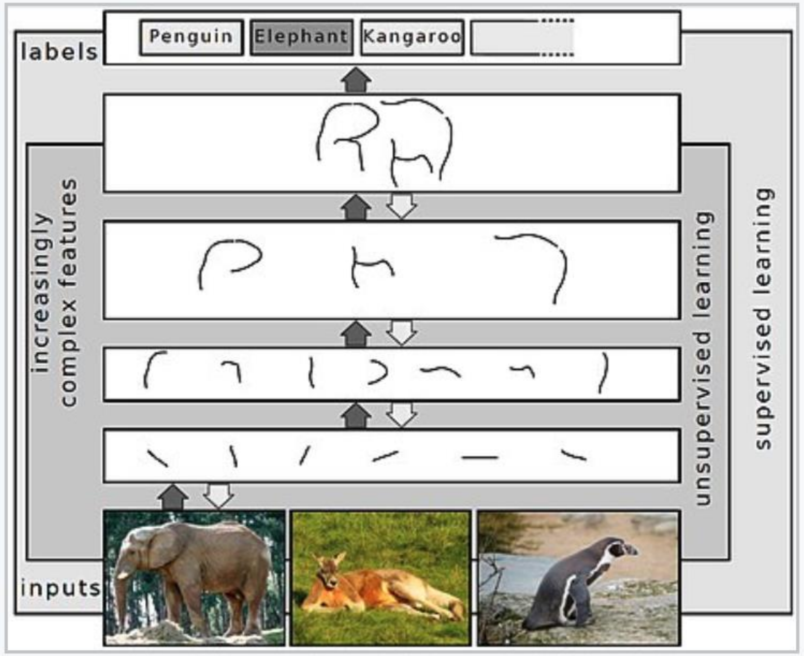
- Deep learning is a class of machine learning algorithms that uses multiple layers to progressively extract higher-level features from the raw input.
- For example, in image processing, lower layers may identify edges, while higher layers may identify the concepts relevant to a human such as digits or letters or faces.
- Most modern deep learning models are based on artificial neural networks, specifically convolutional neural networks (CNN)s
- But they can also include propositional formulas or latent variables organized layer-wise in deep generative models such as the nodes in deep “Boltzmann machines”.
- The Boltzmann machine is based on a spin-glass model with an external field (i.e., a stochastic Ising magnetic spin model)



Representing Images on Multiple Layers of Abstraction in Deep Learning[11]

Deep Learning

- In deep learning, each level learns to transform its input data into a slightly more abstract and composite representation.
- In an image recognition application, the raw input may be a matrix of pixels
  - The first representational layer may abstract the pixels and encode edges
  - The second layer may compose and encode arrangements of edges
  - The third layer may encode a nose and eyes
  - The fourth layer may recognize that the image contains a face.
- Importantly, a deep learning process can learn which features to optimally place in which level on its own.
- This does not completely eliminate the need for hand-tuning
- For example, varying numbers of layers and layer sizes can provide different degrees of abstraction.



Representing Images on Multiple Layers of Abstraction in Deep Learning[11]

Applications of Deep Learning: Speech Recognition

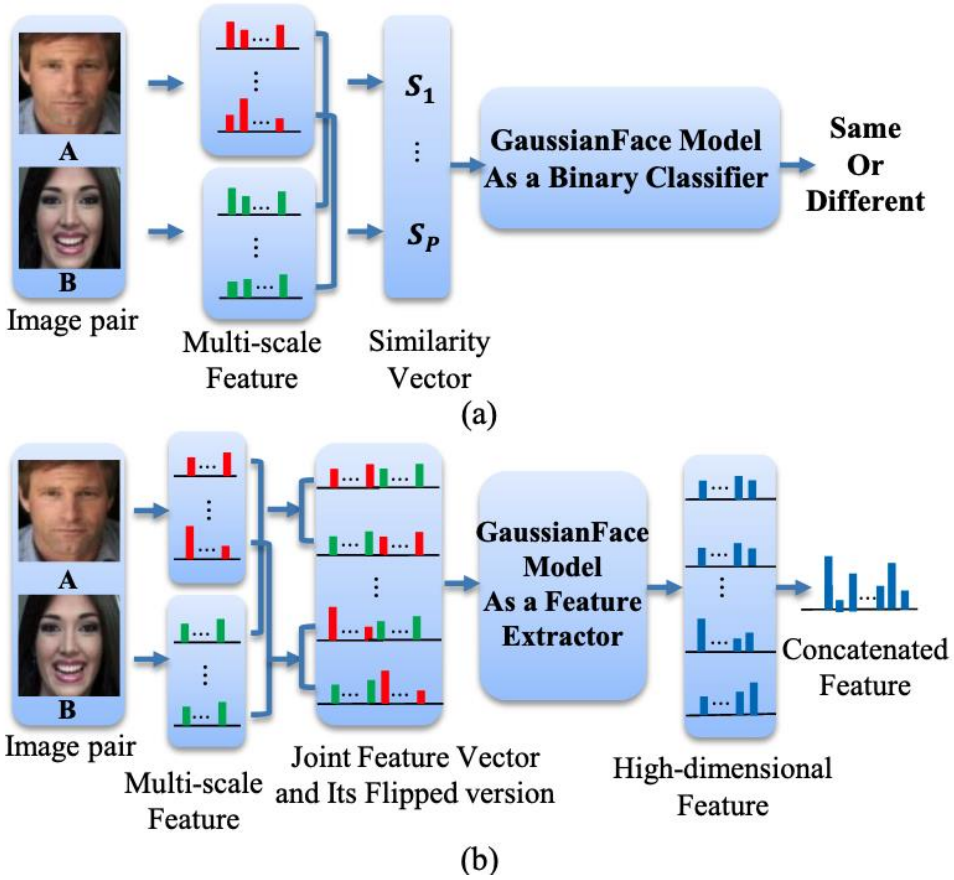
- Large-scale automatic speech recognition is the first and most convincing successful case of deep learning.
- LSTM RNNs can learn "Very Deep Learning" tasks that involve multi-second intervals containing speech events separated by thousands of discrete time steps, where one time step corresponds to about 10 ms.
- LSTM with forget gates is competitive with traditional speech recognizers on certain tasks.
- The initial success in speech recognition was based on small-scale recognition tasks.
- The data set contains 630 speakers from eight major dialects of American English, where each speaker reads 10 sentences.
- Its small size lets many configurations be tried. The error rates listed below, including these early results and measured as percent phone error rates (PER), have been summarized since 1991.

Method	Percent phone error rate (PER) (%)
Randomly Initialized RNN[134]	26.1
Bayesian Triphone GMM-HMM	25.6
Hidden Trajectory (Generative) Model	24.8
Monophone Randomly Initialized DNN	23.4
Monophone DBN-DNN	22.4
Triphone GMM-HMM with BMMI Training	21.7
Monophone DBN-DNN on fbank	20.7
Convolutional DNN[135]	20.0
Convolutional DNN w. Heterogeneous Pooling	18.7
Ensemble DNN/CNN/RNN[136]	18.3
Bidirectional LSTM	17.8
Hierarchical Convolutional Deep Maxout Network[137]	16.5

LSTM RNN = "Long-Short Term Memory Recurrent Neural Network" (similar to a Finite State Machine)

Applications of Deep Learning: Image Recognition

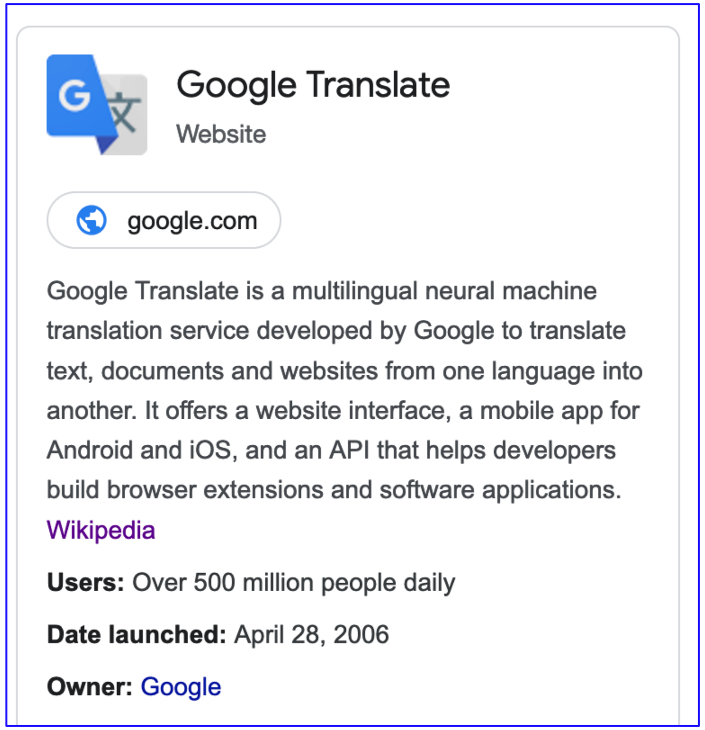
- A common evaluation set for image classification is the MNIST database data set.
- The MNIST database (Modified National Institute of Standards and Technology database) is a large database of handwritten digits that is commonly used for training various image processing systems
- MNIST is composed of handwritten digits and includes 60,000 training examples and 10,000 test examples.
- Its small size lets users test multiple configurations.
- Deep learning-based image recognition has become "superhuman", producing more accurate results than human contestants.
- This first occurred in 2011 in recognition of traffic signs, and in 2014, with recognition of human faces, surpassing human level face recognition
- Deep learning-trained vehicles now interpret 360° camera views.

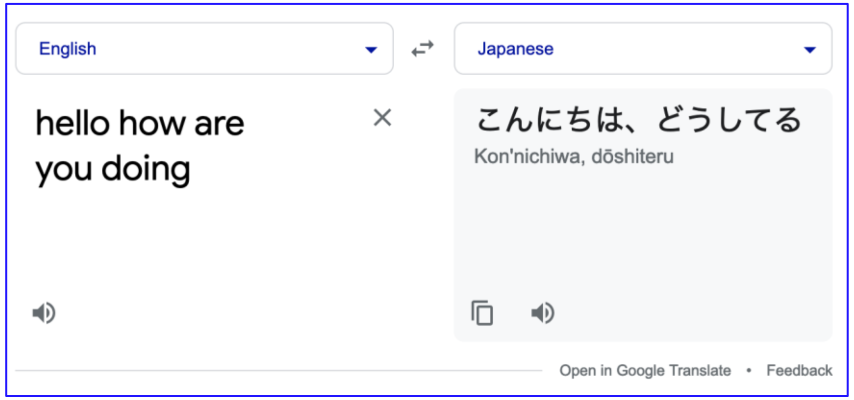


https://arxiv.org/pdf/1404.3840.pdf

Applications of Deep Learning: Natural Language Processing

- Neural networks have been used for implementing language models since the early 2000s.
- LSTM helped to improve machine translation and language modeling.
- Google Translate (GT) uses a large end-to-end long short-term memory (LSTM) network
- Google Neural Machine Translation (GNMT) uses an example-based machine translation method in which the system "learns from millions of examples."
- It translates "whole sentences at a time, rather than pieces.
- Google Translate supports over one hundred languages.
- The network encodes the "semantics of the sentence rather than simply memorizing phrase-to-phrase translations".
- GT uses English as an intermediary between most language pairs.





Other Topics in Deep Learning

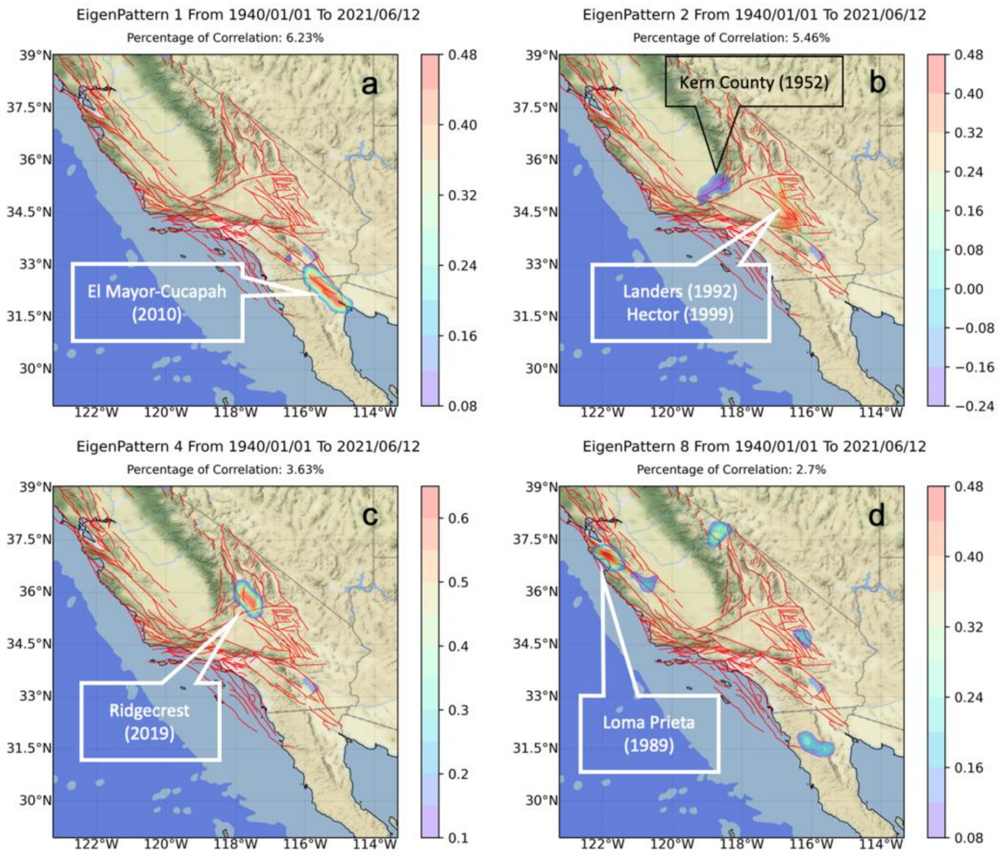
- Applications of artificial intelligence
- Comparison of deep learning software
- Compressed sensing
- Differentiable programming
- Echo state network
- List of artificial intelligence projects
- Liquid state machine
- List of datasets for machine learning research
- Reservoir computing
- Sparse coding

Feature Learning

- Several learning algorithms aim at discovering better representations of the inputs provided during training.
- Classic examples include Principal Component Analysis (PCA) and cluster analysis.
- Feature learning algorithms, also called representation learning algorithms, often attempt to preserve the information in their input but also transform it in a way that makes it useful, often as a pre-processing step before performing classification or predictions.
- This replaces manual feature engineering and allows a machine to both learn the features and use them to perform a specific task.
- Feature learning can be either supervised or unsupervised.
- In supervised feature learning, features are learned using labeled input data: Examples include artificial neural networks, multilayer perceptrons, and supervised dictionary learning.
- In unsupervised feature learning, features are learned with unlabeled input data.
- Examples include dictionary learning, independent component analysis, autoencoders, matrix factorization and various forms of clustering.

Principal Component Analysis (PCA)
Dimensionality Reduction

- Dimensionality reduction is a process of reducing the number of random variables under consideration by obtaining a set of principal variables.
- In other words, it is a process of reducing the dimension of the feature set, also called "number of features".
- Most of the dimensionality reduction techniques can be considered as either feature elimination or extraction.
- One of the popular methods of dimensionality reduction is Principal Component Analysis (PCA).
- PCA involves changing higher-dimensional data (e.g., 3D) to a smaller space (e.g., 2D).
- This results in a smaller dimension of data (2D instead of 3D), while keeping all original variables in the model without changing the data.



An example of PCA from JBR et al. (2021)

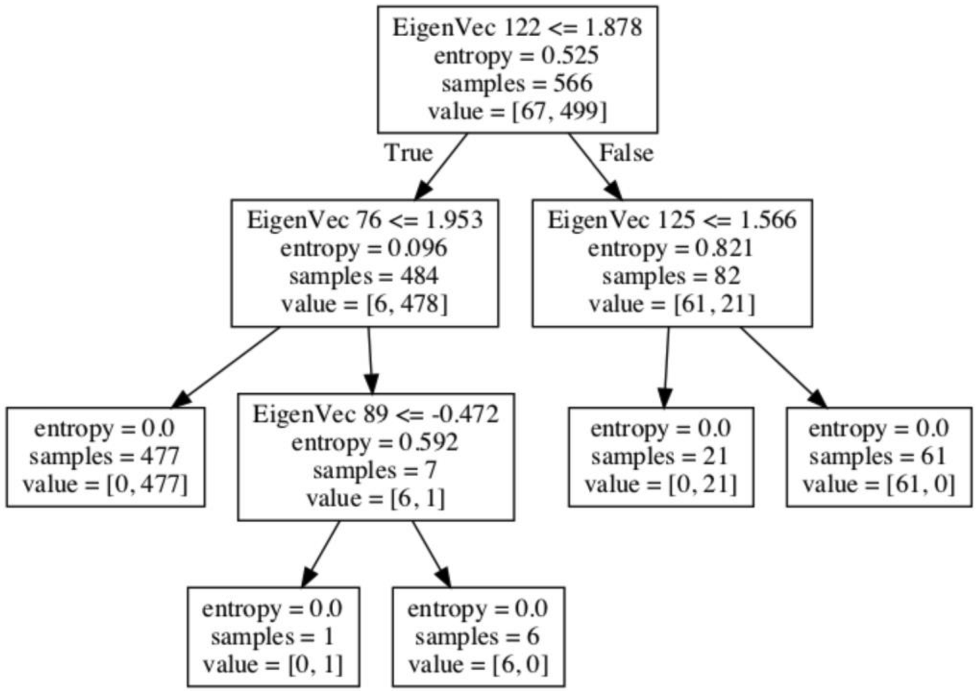
Anomaly Detection

- In data mining, anomaly detection, also known as outlier detection, is the identification of rare items, events or observations which raise suspicions by differing significantly from the majority of the data.
- Typically, the anomalous items represent an issue such as bank fraud, a structural defect, medical problems or errors in a text.
- Anomalies are referred to as outliers, novelties, noise, deviations and exceptions.
- In particular, in the context of abuse and network intrusion detection, the interesting objects are often not rare objects, but unexpected bursts of inactivity.
- This pattern does not adhere to the common statistical definition of an outlier as a rare object, and many outlier detection methods (in particular, unsupervised algorithms) will fail on such data unless it has been aggregated appropriately.
- Instead, a cluster analysis algorithm may be able to detect the micro-clusters formed by these patterns.

Random Forest is a standard decision tree algorithm for timeseries analysis

Decision Trees

- Decision tree learning uses a decision tree as a predictive model to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves).
- It is one of the predictive modeling approaches used in statistics, data mining, and machine learning.
- Tree models where the target variable can take a discrete set of values are called classification trees;
- in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels.
- Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees.
- In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making.
- In data mining, a decision tree describes data, but the resulting classification tree can be an input for decision making.



A typical decision tree used in classifying feature vectors

Regression Analysis

- Regression analysis encompasses a large variety of statistical methods to estimate the relationship between input variables and their associated features.
- Its most common form is linear regression, where a single line is drawn to best fit the given data according to a mathematical criterion such as ordinary least squares.
- The latter is often extended by regularization (mathematics) methods to mitigate overfitting and bias, as in ridge regression.
- When dealing with non-linear problems, go-to models include
  - Polynomial regression (for example, used for trendline fitting in Microsoft Excel),
  - Logistic regression (often used in statistical classification)
  - Or even kernel regression, which introduces non-linearity to implicitly map input variables to higher-dimensional space.

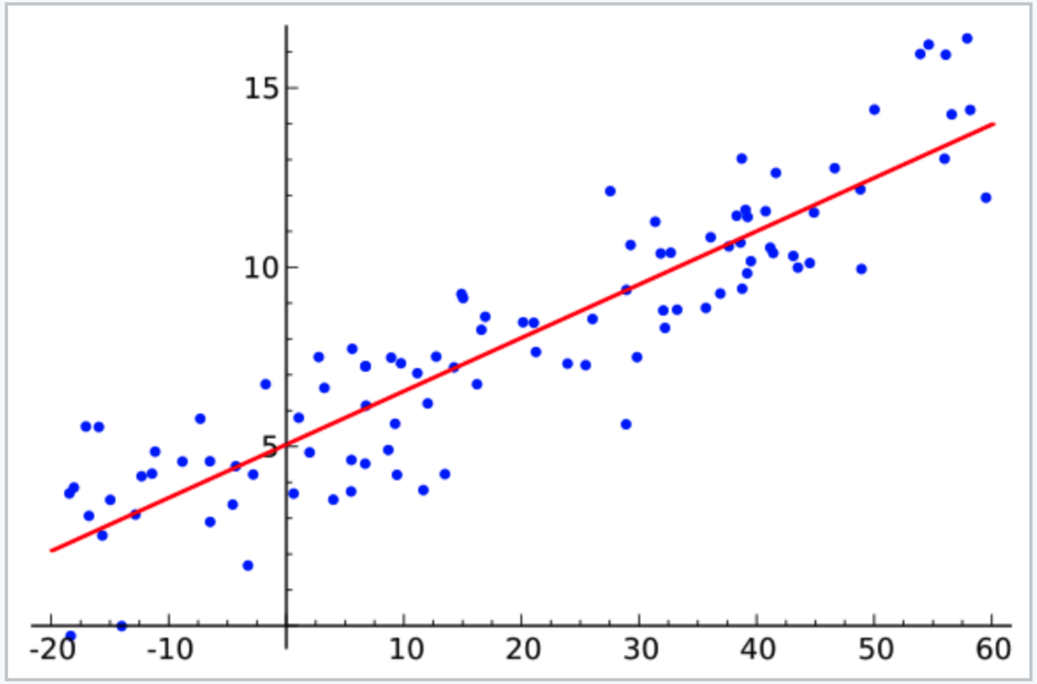


Illustration of linear regression on a data set.

Other Applications

There are many applications for machine learning, including:

- Agriculture
- Anatomy
- Adaptive website
- Affective computing
- Astronomy
- Banking
- Bioinformatics
- Brain–machine interfaces
- Cheminformatics
- Citizen science
- Computer networks
- Computer vision
- Credit-card fraud detection
- Data quality
- DNA sequence classification
- Economics
- Financial market analysis[75]
- General game playing
- Handwriting recognition
- Information retrieval
- Insurance
- Internet fraud detection
- Knowledge graph embedding
- Linguistics
- Machine learning control
- Machine perception
- Machine translation
- Marketing
- Medical diagnosis
- Natural language processing
- Natural language understanding
- Online advertising
- Optimization
- Recommender systems
- Robot locomotion
- Search engines
- Sentiment analysis
- Sequence mining
- Software engineering
- Speech recognition
- Structural health monitoring
- Syntactic pattern recognition
- Telecommunication
- Theorem proving
- Time-series forecasting
- User behavior analytics
- Behaviorism

An Interesting Python Resource (for machine learning)

machine learning +

Blogs | Resources | University | Search | Learn Free

Matplotlib | Python | Visualization

Top 50 matplotlib Visualizations – The Master Plots (with full python code)

November 28, 2018 by Selva Prabhakaran

A compilation of the Top 50 matplotlib plots most useful in data analysis and visualization. This list lets you choose what visualization to show for what situation using python's matplotlib and seaborn library. **Create Powerful Visualizations using Python with my FREE 9-Day-Video-Course.**

Introduction

The charts are grouped based on the 7 different purposes of your visualization objective. For example, if you want to picturize the relationship between 2 variables, check out the plots under the 'Correlation' section. Or if you want to show how a value changed over time, look under the 'Change' section and so on. An effective chart is one which:

1. Conveys the right and necessary information without distorting facts.
2. Simple in design, you don't have to strain in order to get it.
3. Aesthetics support the information rather than overshadow it.
4. Not overloaded with information.

__Related Posts: __ Matplotlib Full Tutorial Matplotlib Subplots **Become a high paid data scientist** with my structured Machine Learning Career Path. *Includes access to all my current and future courses of Machine Learning, Deep Learning and Industry Projects. With 24x7 query support.*

Related Posts

- Matplotlib Tutorial – A Complete Guide
- Matplotlib – Complete overview – Part 2
- Matplotlib Histogram
- Bar Plot in Python
- Matplotlib Line Plot
- Matplotlib Pyplot
- Python Boxplot
- Python Scatter Plot
- Subplots Python (Matplotlib)
- Top 50 matplotlib Visualizations

https://www.machinelearningplus.com/plots/top-50-matplotlib-visualizations-the-master-plots-python/

Credits

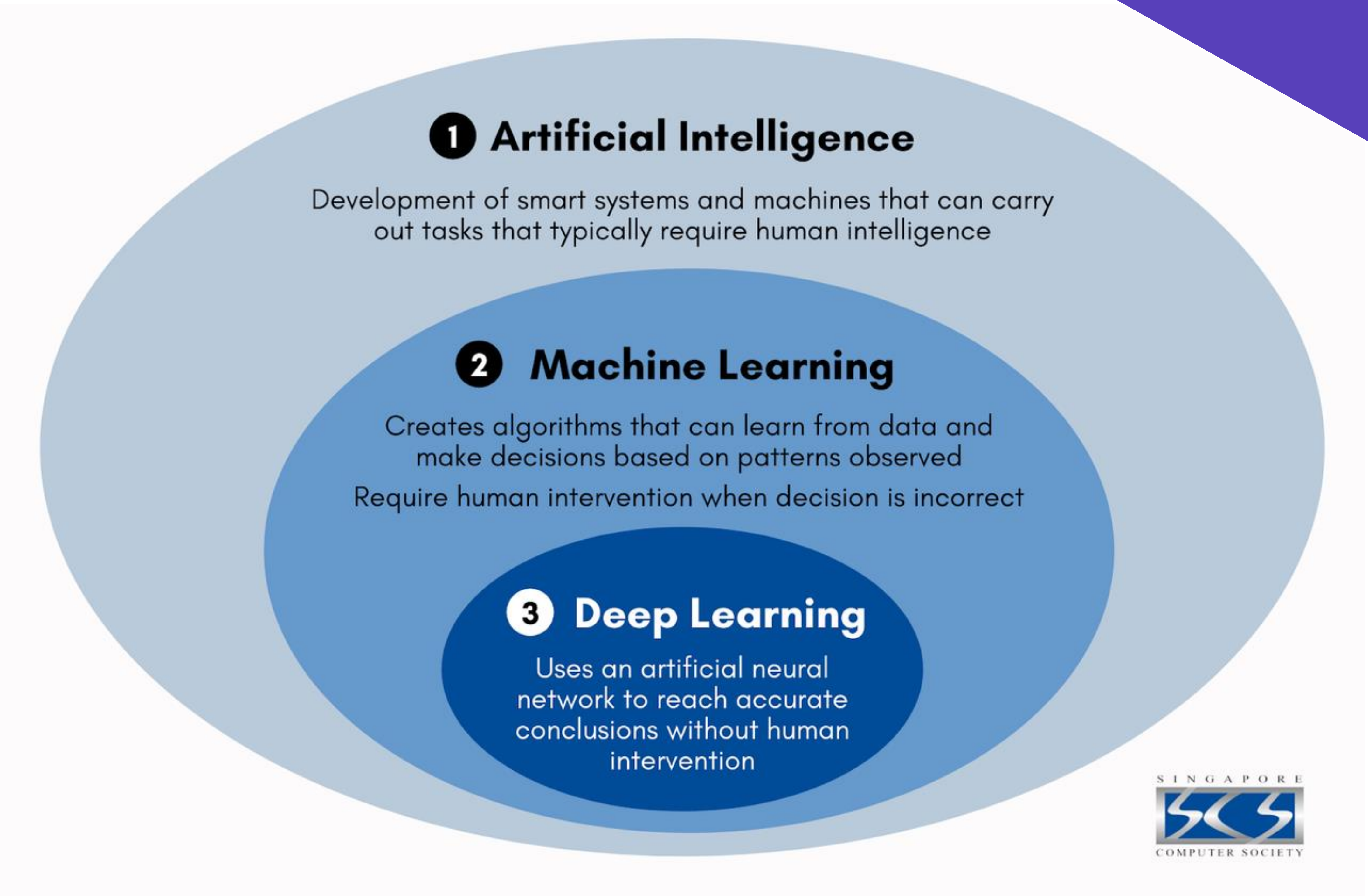
- https://en.wikipedia.org/wiki/Machine_learning
- https://en.wikipedia.org/wiki/Supervised_learning
- https://en.wikipedia.org/wiki/Unsupervised_learning
- https://en.wikipedia.org/wiki/Deep_learning
- https://www.machinelearningplus.com/plots/top-50-matplotlib-visualizations-the-master-plots-python/

Introduction to Machine Learning

Module 1

by Todd Nash

Red Deer
Polytechnic



Machine learning is a branch of AI, based on the concept that machines and systems can analyze and understand data, and learn from it and make decisions with minimal to zero human intervention.

by Todd Nash

Red Deer Polytechnic

Machine Learning Timeline

- **1950** — Alan Turing publishes paper asking 'Can machines think?'
- **1959** — Arthur Samuel coined the name machine learning
- **1985** — Neural network that can self-teach and pronounce 20k words developed
- **1997** — IBM's deep blue computer wins game of chess from a human
- **2012** — Google creates deep neural network to recognize images of humans and cats from videos
- **2014** — First chatbot 'Euren Goostman' clears the Turing test

  Deepface by Facebook recognizes human faces like humans do
- **till 2023** — Numerous advancements: self-driving cars, personal assistants, weather prediction, recommendation systems, traffic analysis

by Todd Nash

Red Deer Polytechnic

Key benefits of AI/ML adoption

Boost revenue

Encourages customisation and personalisation of services for customers, boosting revenue

Reduce cost

Assist in building efficiencies in processes, reducing overall cost

Reduce error

Automation processes help reduce error and improve resource utilisation

Unleash potential

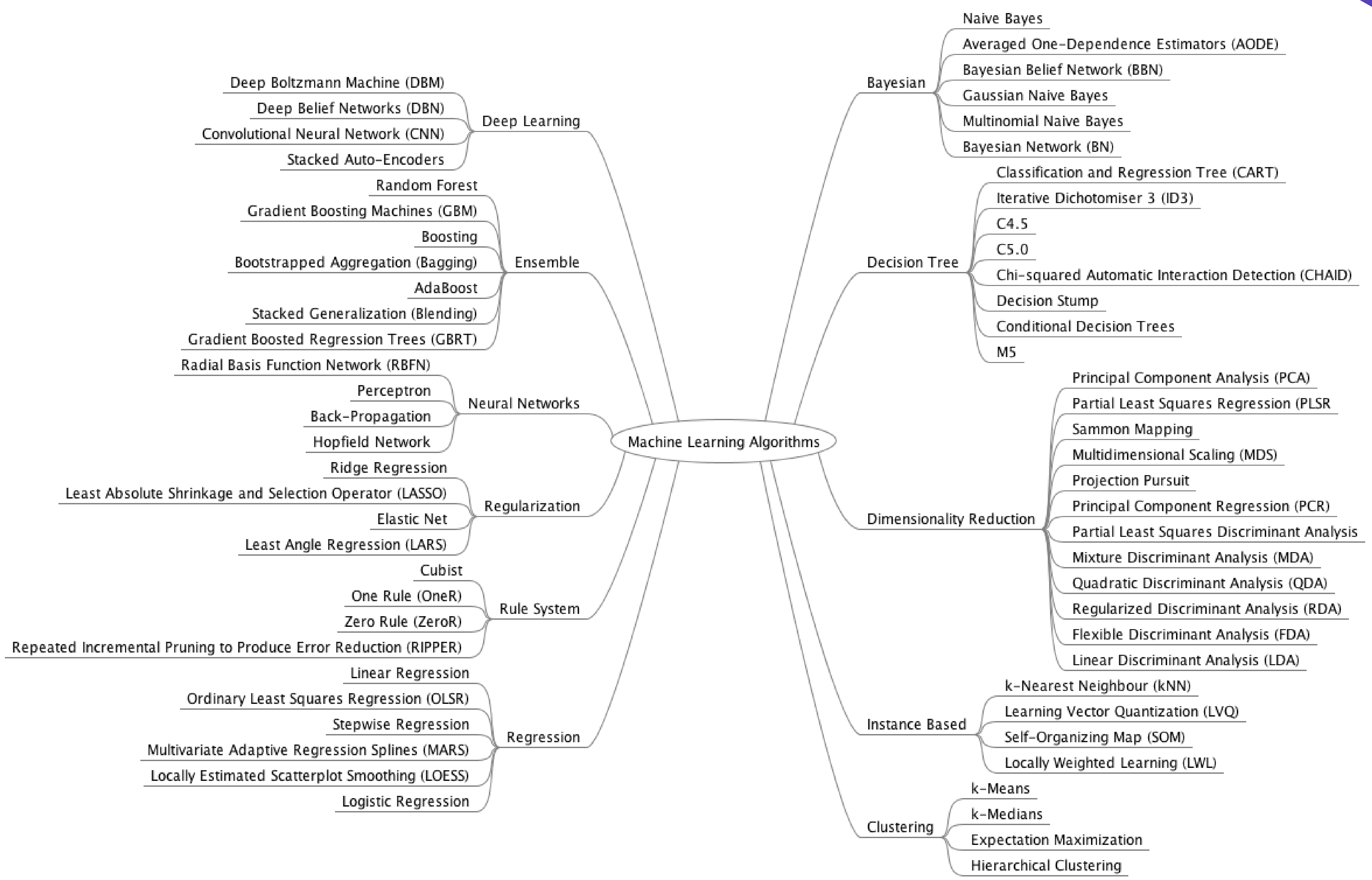
Help discover unrealised opportunities by enabling processing & generation of insights from vast datasets

by Todd Nash

Red Deer Polytechnic

ML in our daily lives.

by Todd Nash



What is Machine Learning, exactly?

by Todd Nash

Red Deer Polytechnic

Cat

or

Dog?

How ML works…

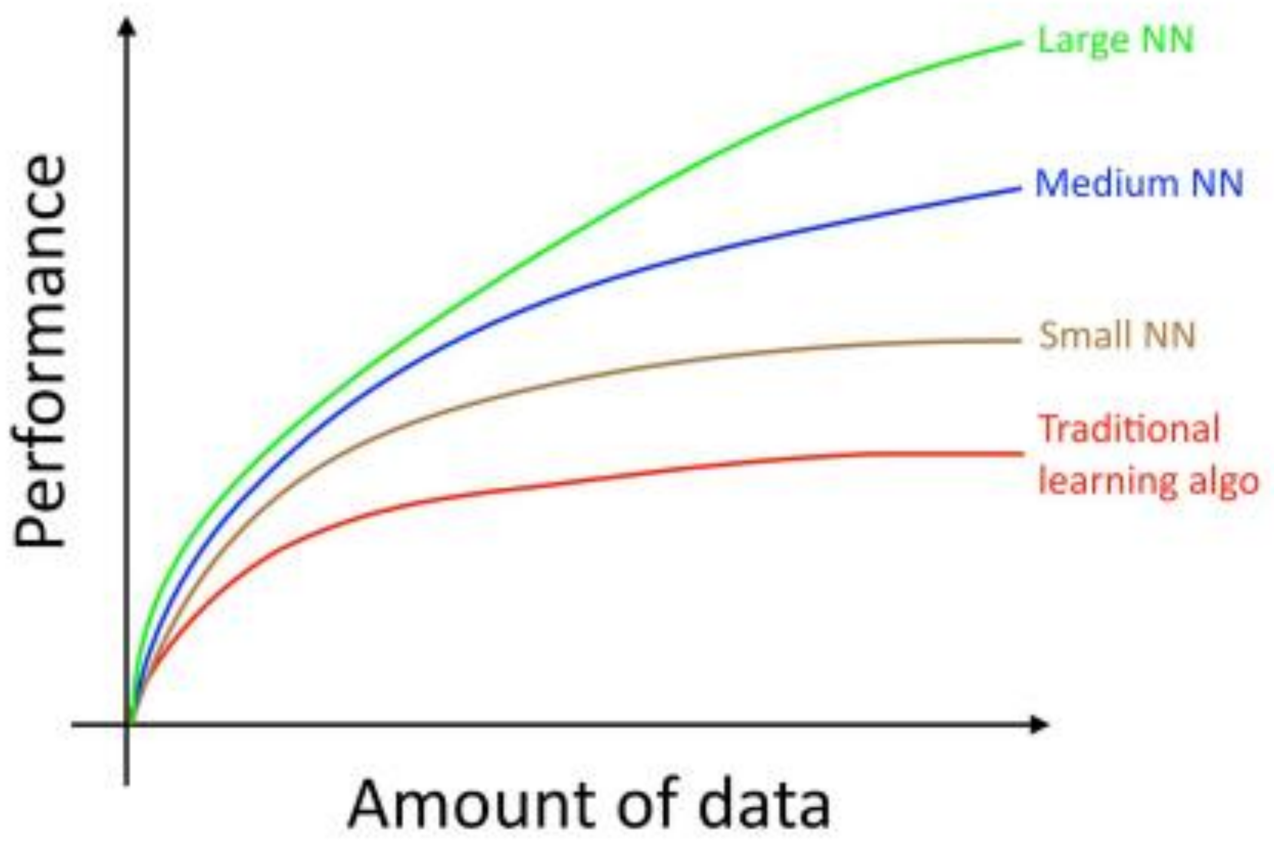
by Todd Nash

Scale drives the Machine Learning process

Many of the ideas of deep learning (neural networks) have been around for decades. Why are these ideas taking off now?

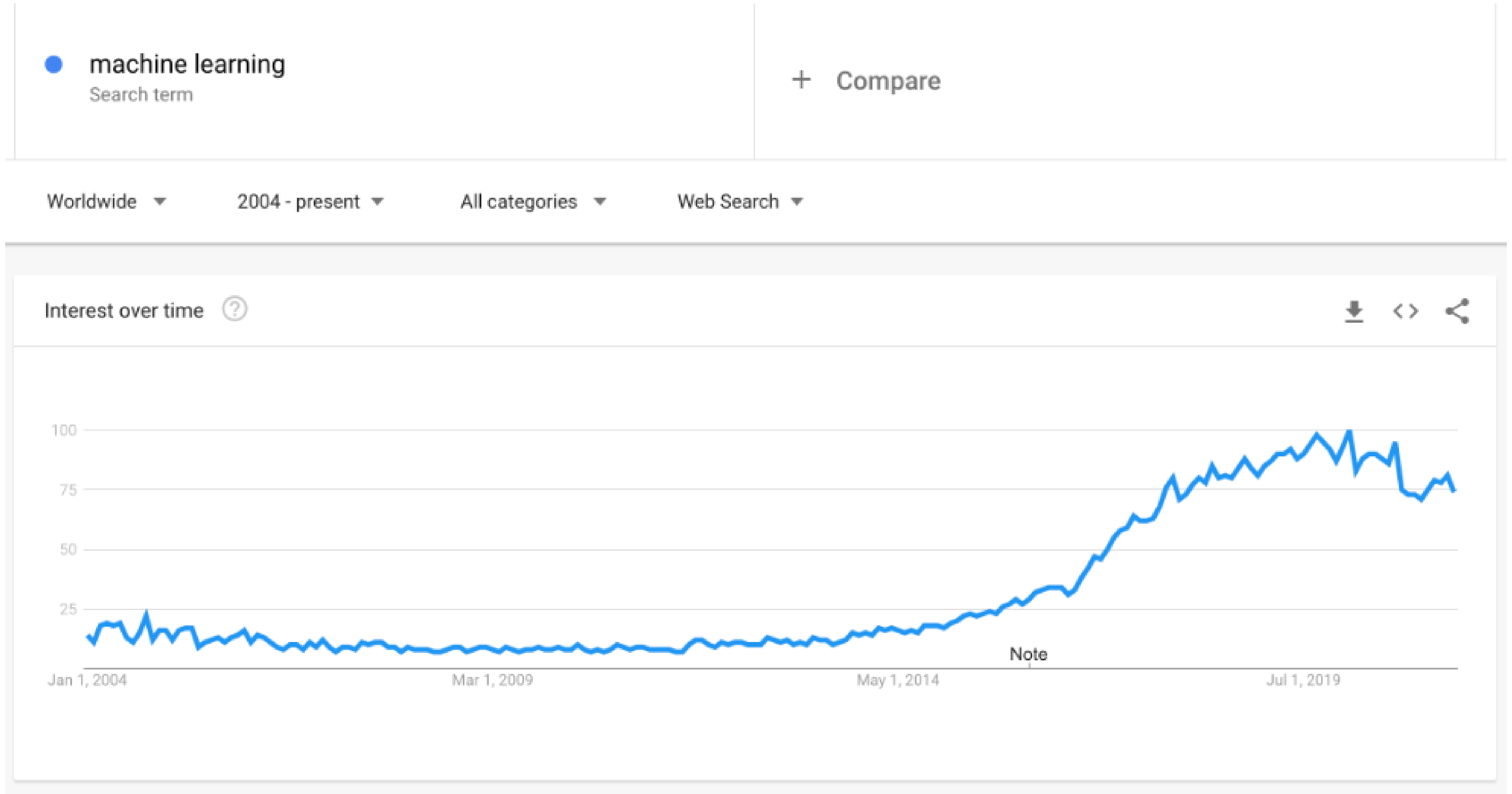
Two of the biggest drivers of recent progress have been:

- **Data availability.**
- **Computational scale.**



by Todd Nash

The hype curve



Google Trends shows the recent 'hype curve' of the term 'machine learning'

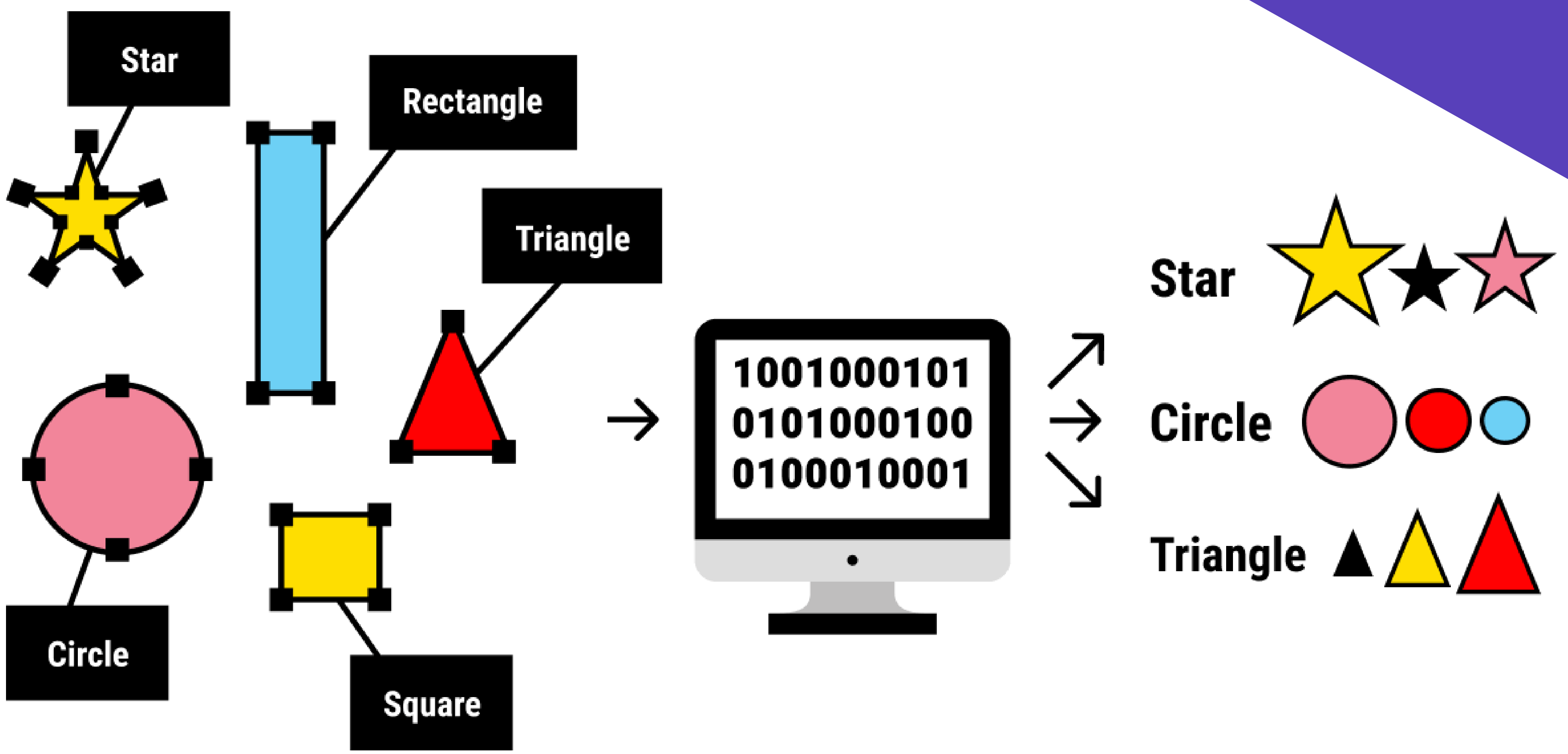
by Todd Nash

Basic Terminology

by Todd Nash

Red Deer Polytechnic

Labels

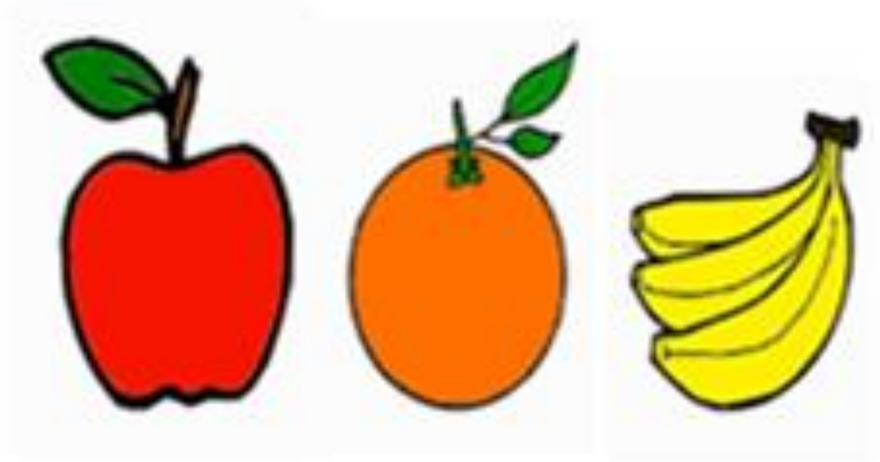


A label is the thing we're predicting (the y variable in simple linear regression). The label could be the future price of wheat, the kind of animal shown in a picture, the meaning of an audio clip, or just about anything.

by Todd Nash

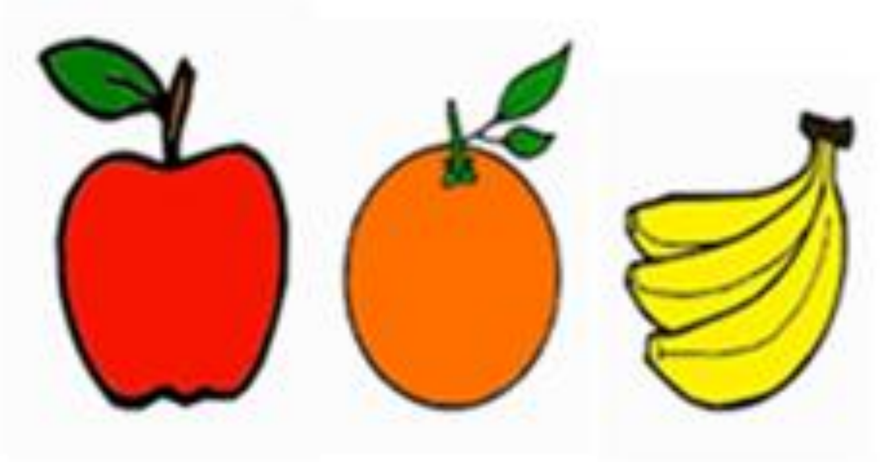
Red Deer Polytechnic

LABELED DATA (BASED ON TASTE)



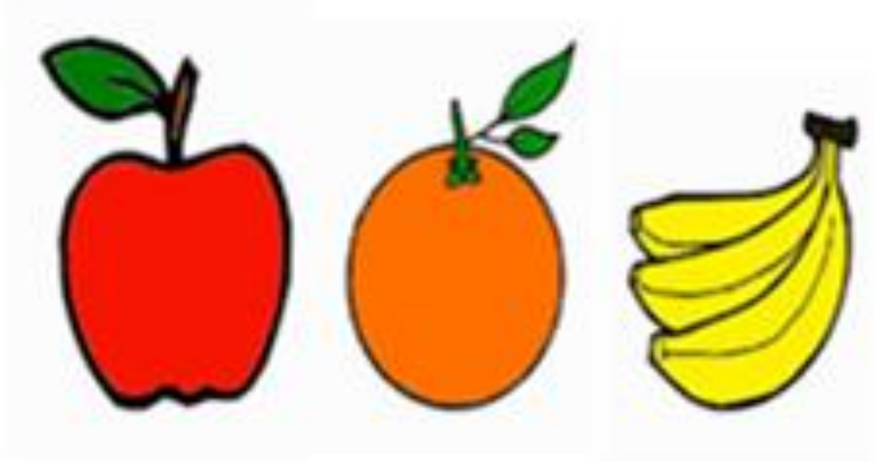
Apple **Orange** **Banana**

LABELED DATA (BASED ON COLOUR)



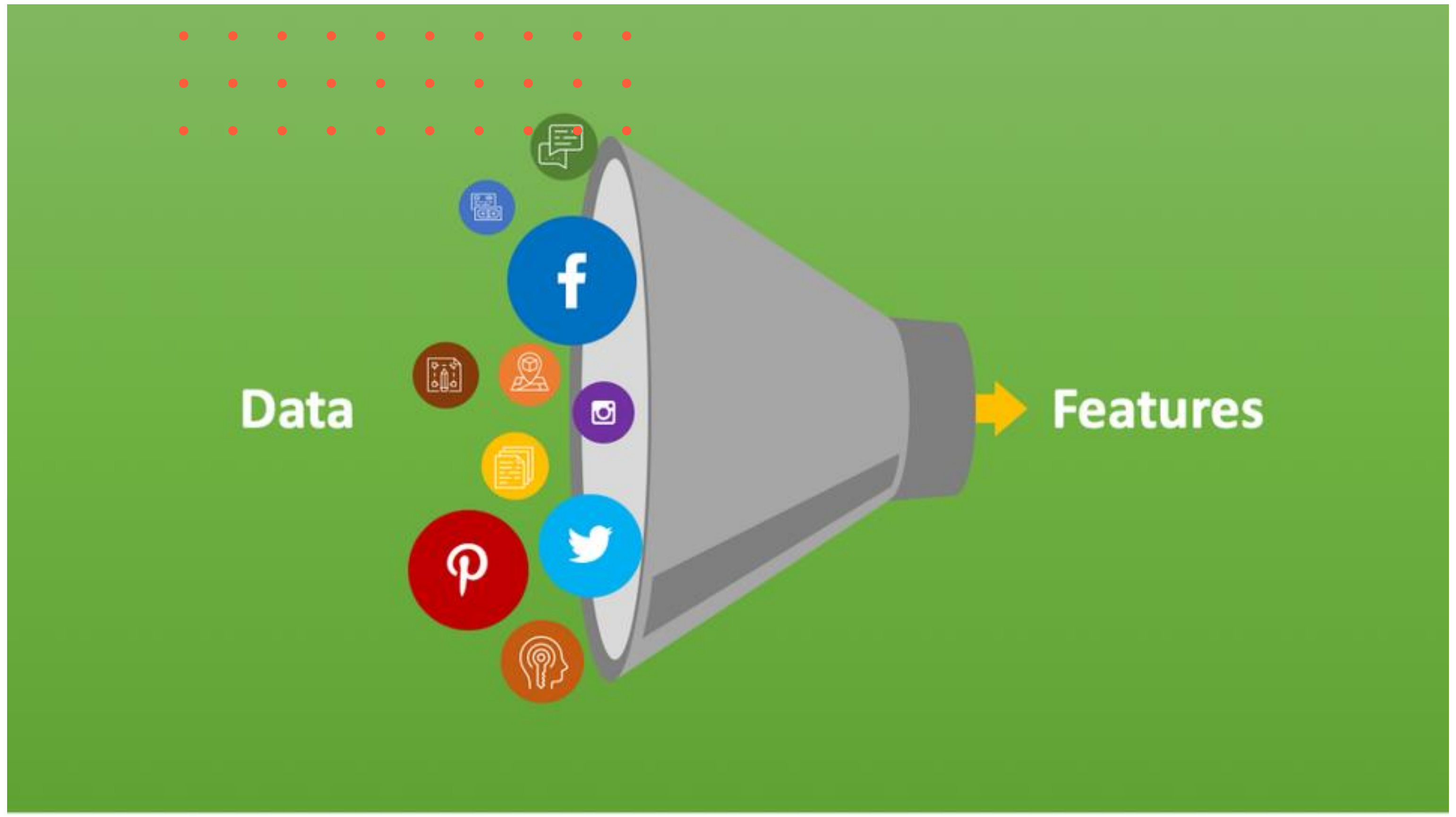
Red **Orange** **Yellow**

UNLABELED DATA



by Todd Nash

Features



A feature is an input variable (the x variable in simple linear regression). A simple machine learning project might use a single feature, while a more sophisticated machine learning project could use millions of features.

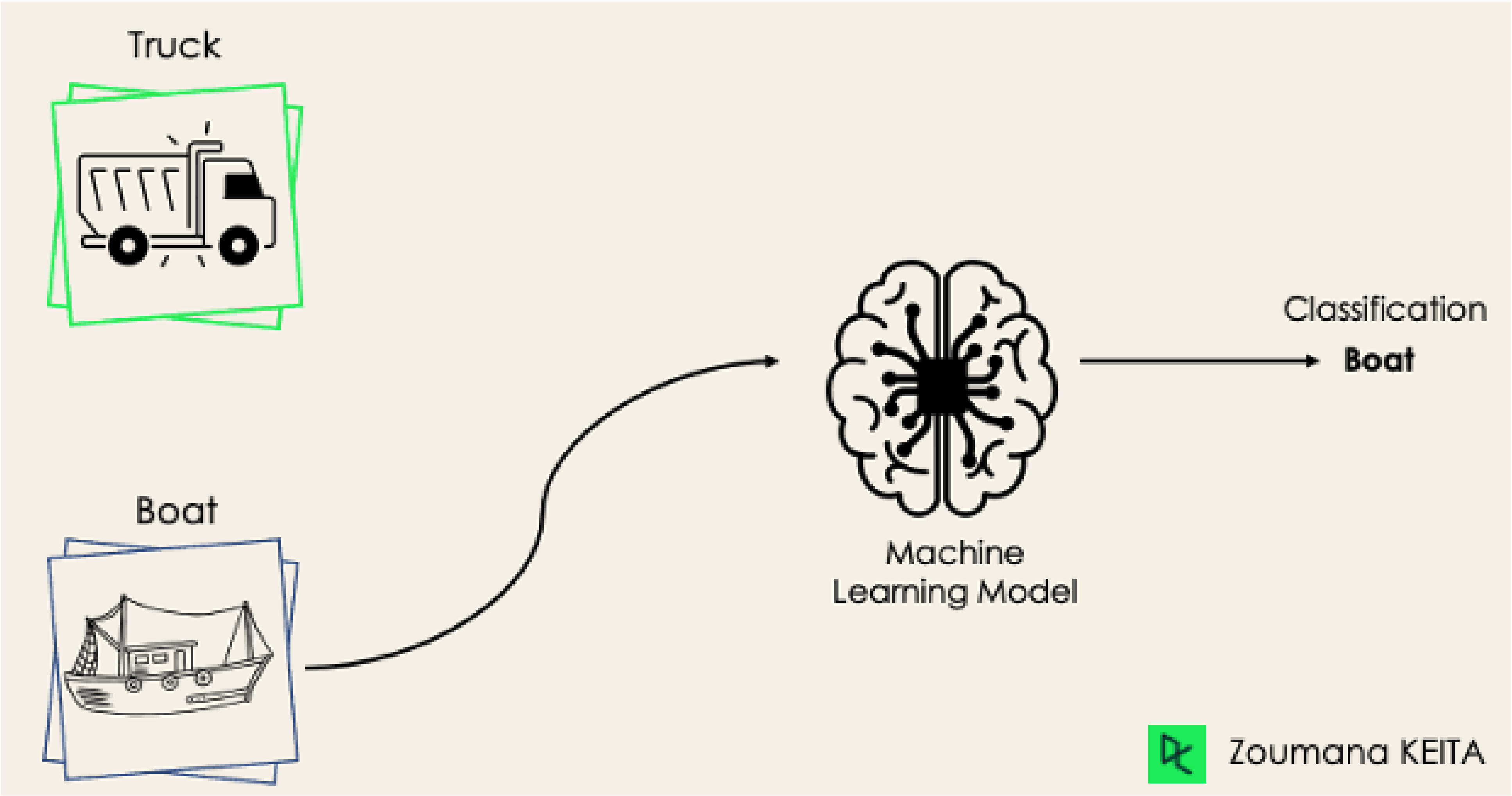
by Todd Nash

Classification

Objective: In classification, the goal is to assign input data points to predefined categories or classes.

Output: The output is a discrete label or category.

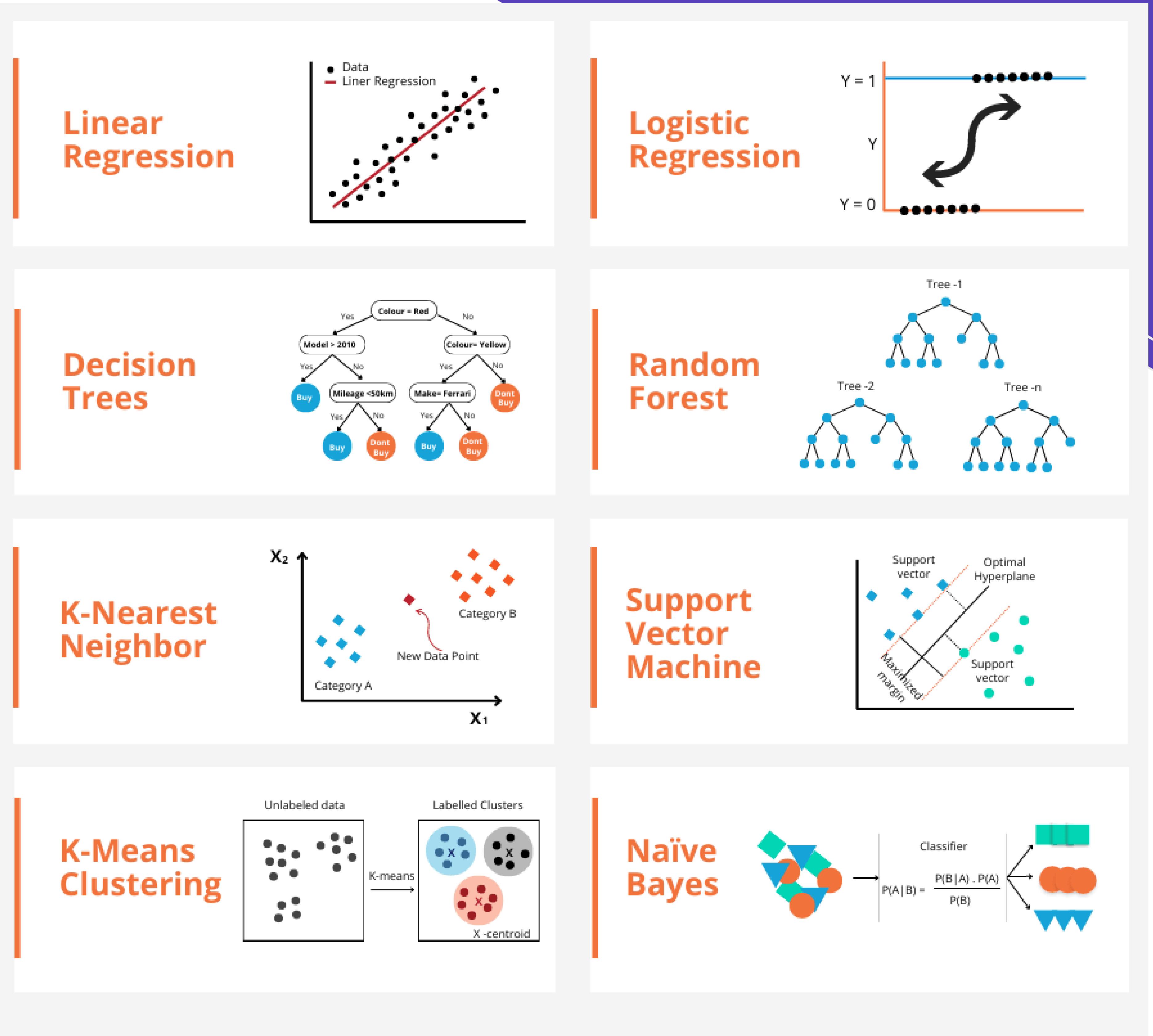
Example: Predicting whether an email is spam or not, classifying images of digits as numbers 0-9, or determining whether a patient has a particular disease based on medical test results.



by Todd Nash

Model

A model is a mathematical representation or algorithm that captures patterns and relationships in data. The primary purpose of a model is to make predictions or decisions based on input data. Models are trained on historical data, learning from examples, and are then used to make predictions on new, unseen data.



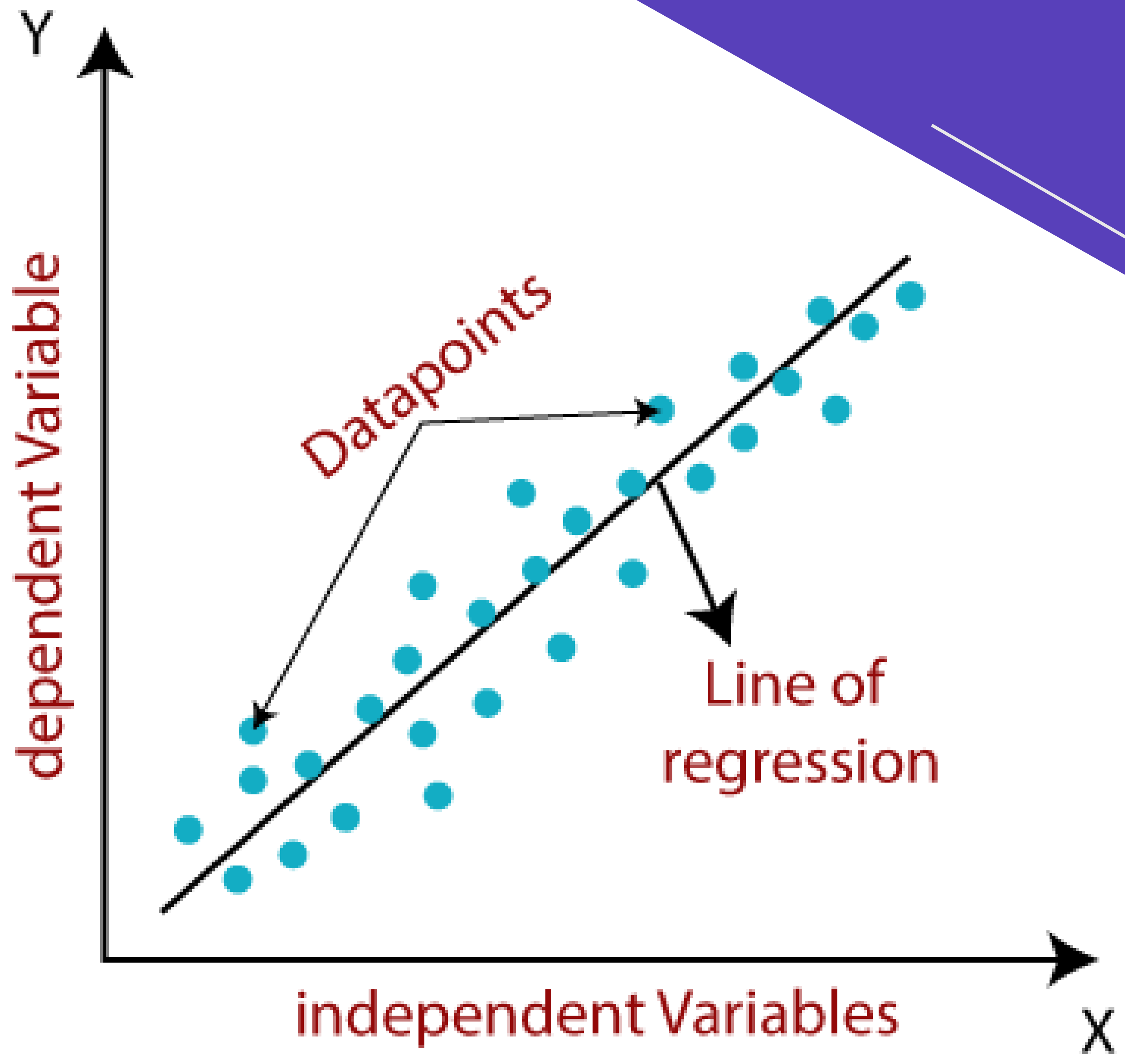
by Todd Nash

Regression

Objective: In regression, the goal is to predict a continuous numerical output or value.

Output: The output is a quantity, often a real number.

Example: Predicting house prices based on features like square footage and number of bedrooms, forecasting stock prices, or estimating a person's age based on demographic information.



by Todd Nash

Red Deer Polytechnic

Training

Objective: The primary goal of training in machine learning is to equip a model with the ability to recognize patterns and relationships within a dataset, enabling it to make predictions or classifications.

Output: The model, during training, learns to produce an output based on input features.

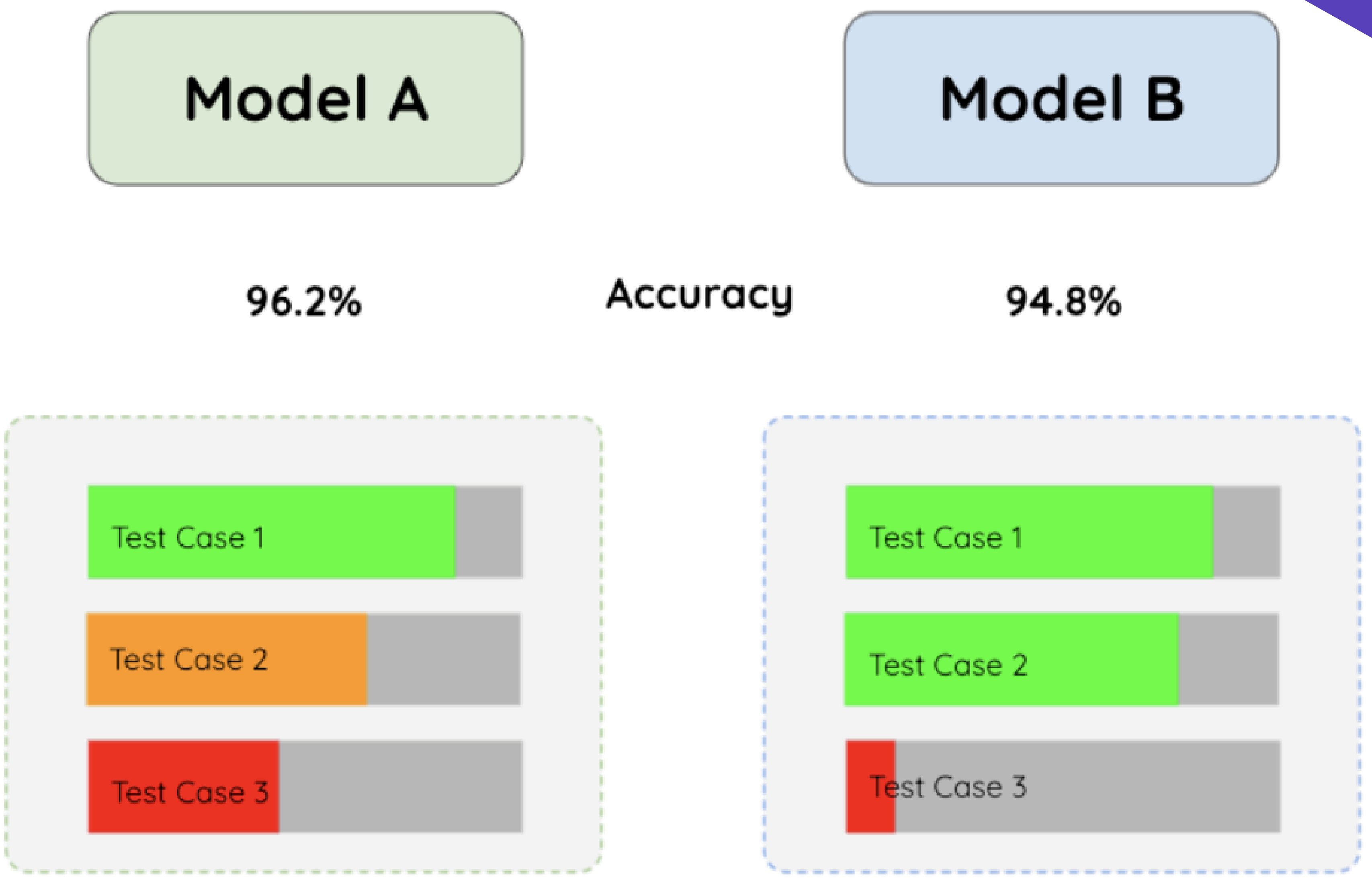
Example: Consider a scenario where a machine learning model is trained to distinguish between images of cats and dogs. Through exposure to labeled examples, the model learns to associate features in the images with the correct categories, enabling it to make accurate predictions on new, unseen images.

by Todd Nash

Testing

Objective: The primary objective of testing in machine learning is to evaluate the performance and generalization ability of a trained model on new, unseen data.

Test Dataset: A separate dataset, distinct from the training and validation sets, is used for testing. This dataset contains instances the model has not encountered during training or validation.



by Todd Nash

Types of Machine Learning

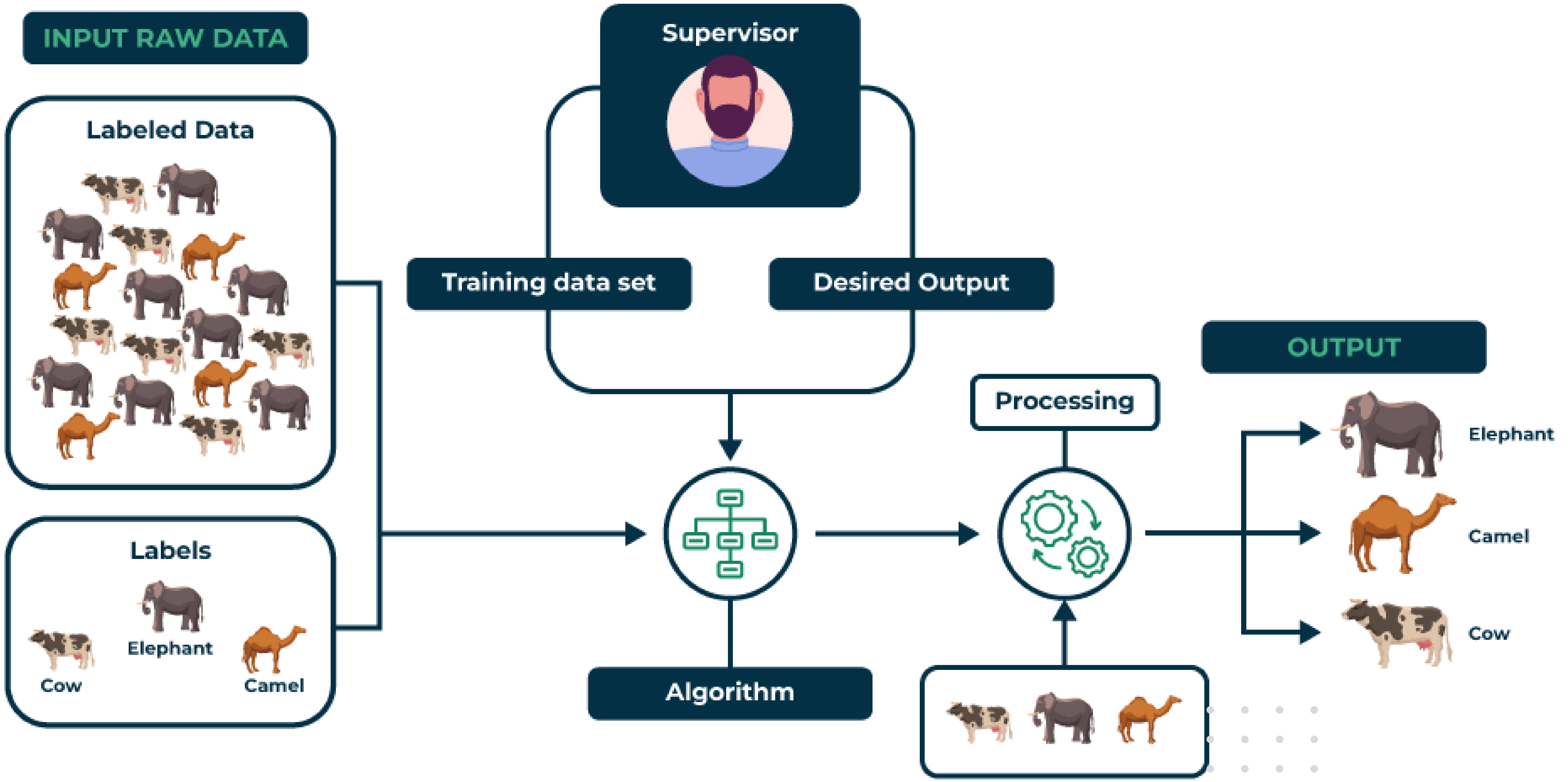
When talking about machine learning basics, you must know that it is comprised of three different types:

Supervised learning: You supervise the machine while training it to work on its own. This requires labeled training data.

Unsupervised learning: There is training data, but it won't be labeled.

Reinforcement learning: The system learns on its own through many iterations.

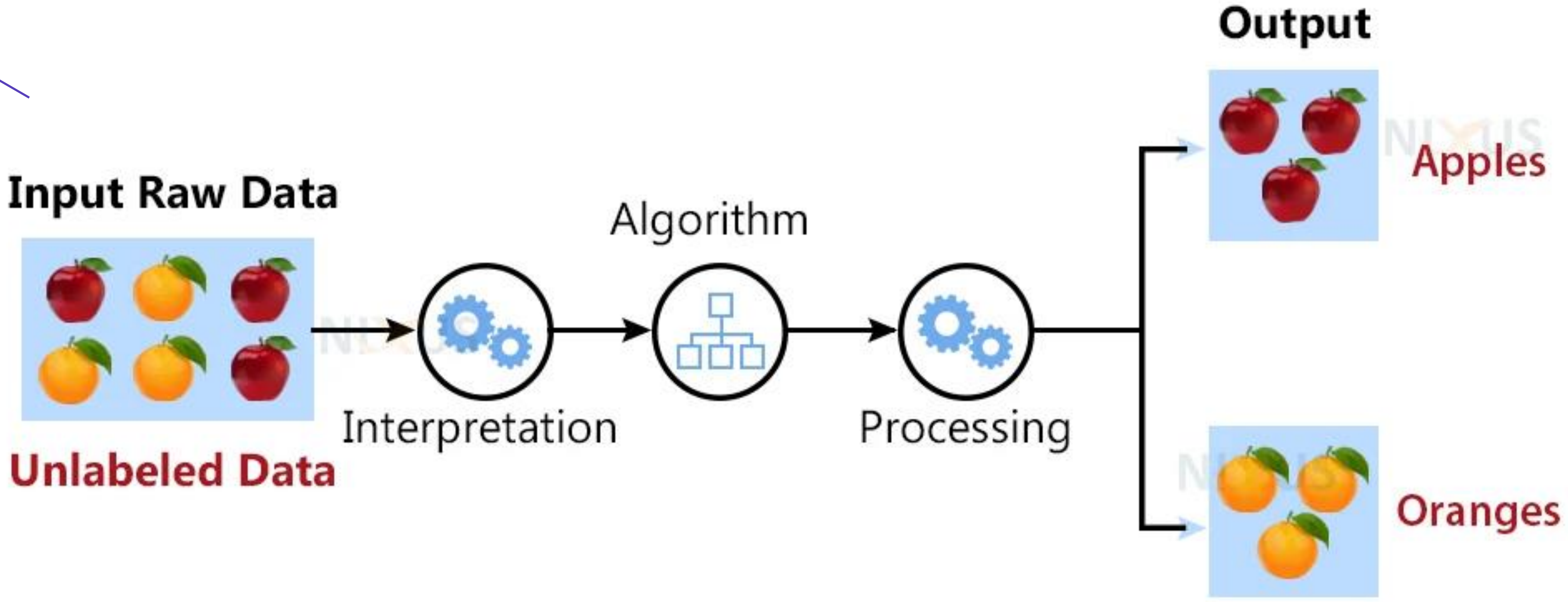
Supervised Learning



by Todd Nash

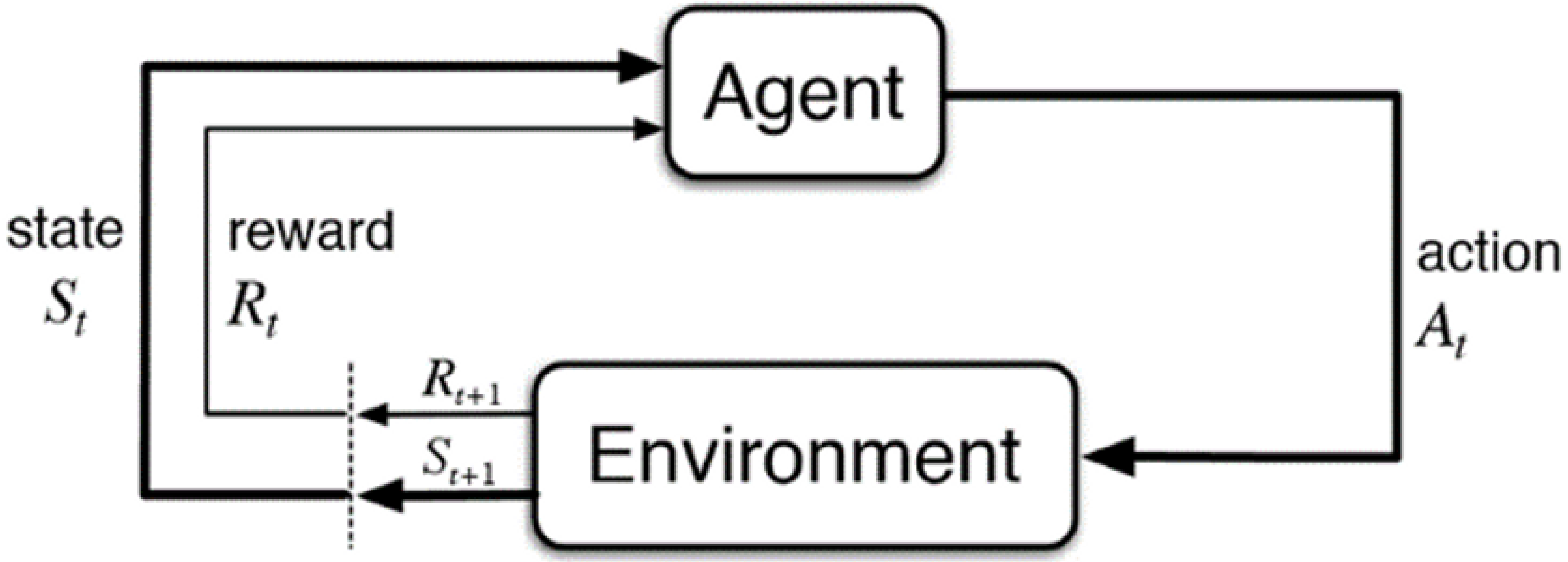
Red Deer Polytechnic

Unsupervised Machine Learning



by Todd Nash

Reinforcement Learning



by Todd Nash

Red Deer Polytechnic

Machine learning models and their training algorithms

Supervised learning

Data scientists provide input, output and feedback to build model (as the definition).

EXAMPLE ALGORITHMS:

Linear regressions

- Sales forecasting.
- Risk assessment.

Support vector machines

- Image classification.
- Financial performance comparison.

Decision trees

- Predictive analytics.
- Pricing.

Unsupervised learning

Use deep learning to arrive at conclusions and patterns through unlabeled training data.

EXAMPLE ALGORITHMS:

Apriori

- Sales functions.
- Word associations.
- Searcher.

K-means clustering

- Performance monitoring.
- Searcher intent.

Artificial neural networks

- Generate new, synthetic data.
- Data mining and pattern recognition.

Semi-supervised learning

Builds a model through a mix of labeled and unlabeled data, a set of categories, suggestions and exampled labels.

EXAMPLE ALGORITHMS:

Generative adversarial networks

- Audio and video manipulation.
- Data creation.

Self-trained Naïve Bayes classifier

- Natural language processing.

Reinforcement learning

Self-interpreting but based on a system of rewards and punishments learned through trial and error, seeking maximum reward.

EXAMPLE ALGORITHMS:

Q-learning

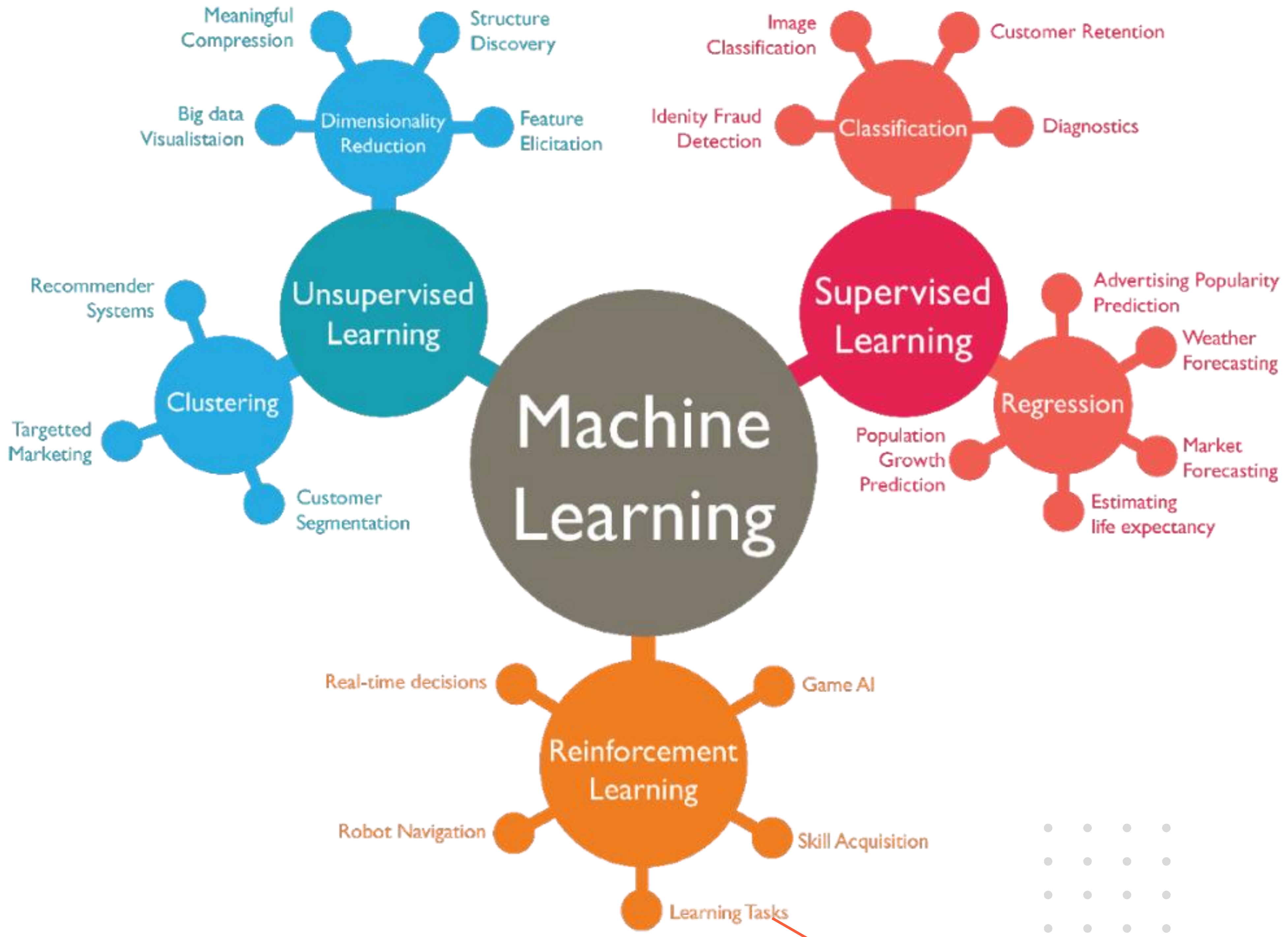
- Policy creation.
- Consumption reduction.

Model-based value estimation

- Linear tasks.
- Estimating parameters.

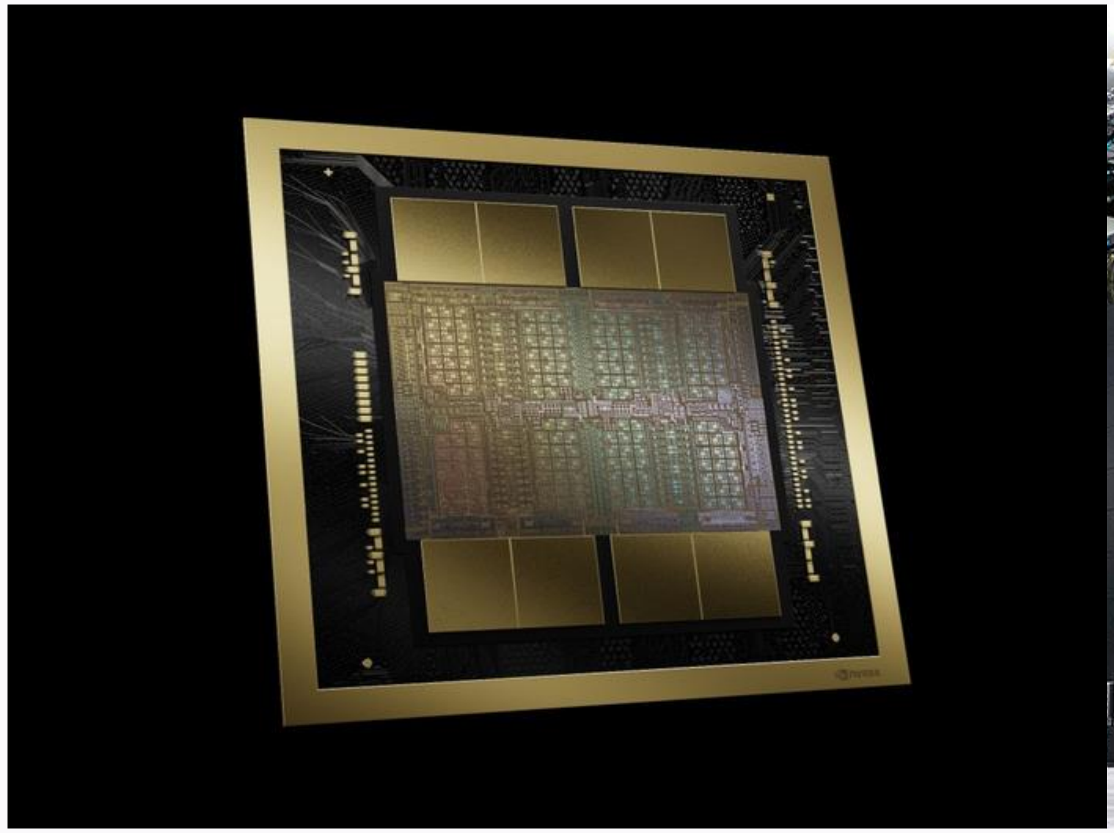
by Todd Nash

How do you choose the right Machine Learning solution to use?



by Todd Nash

Red Deer
Polytechnic

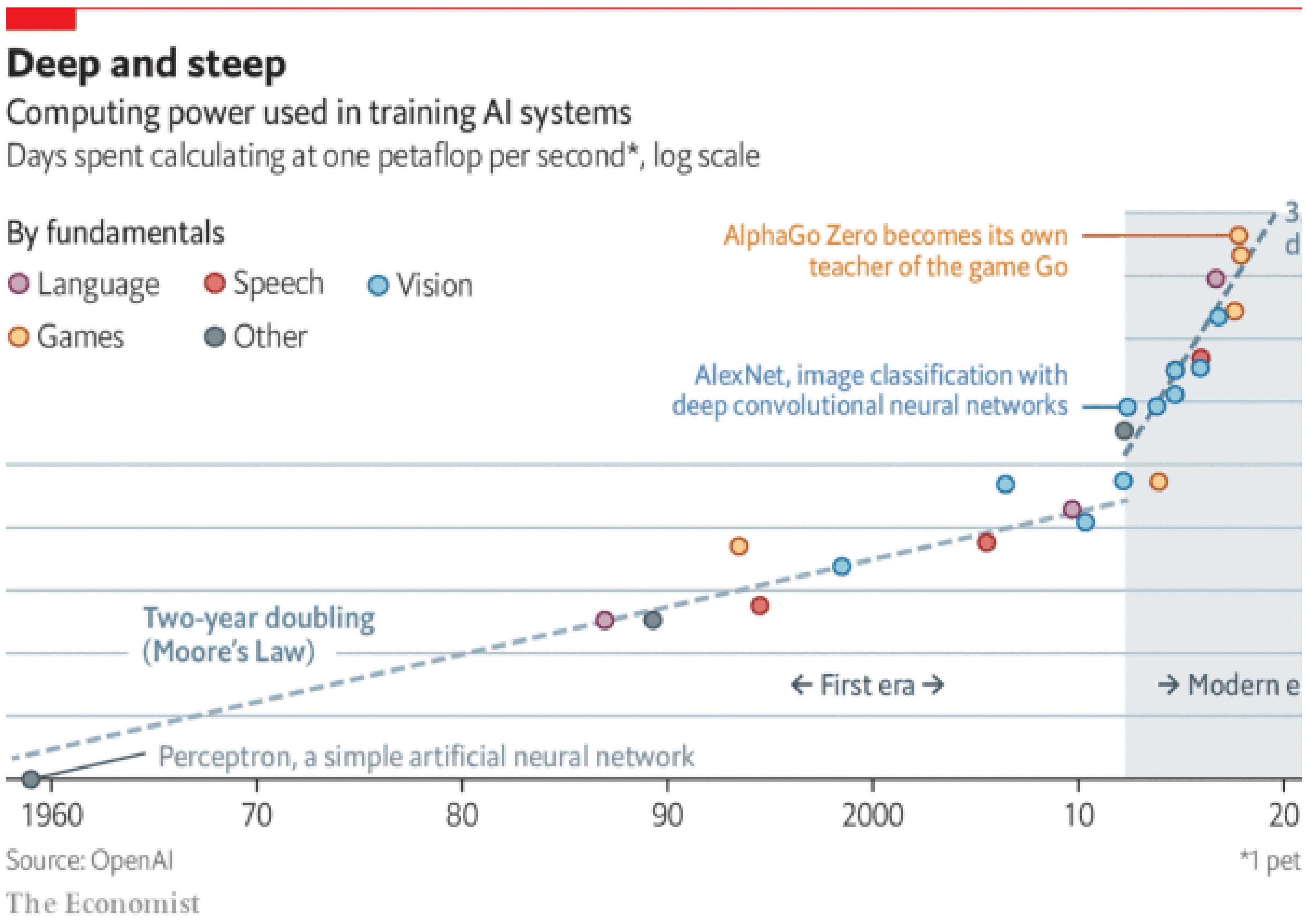


Machine Learning Hardware Accelerators & Architectures

by Todd Nash

Why Use Hardware Accelerators for Machine Learning?

- Increasing demand for computational power in machine learning tasks
- Traditional processors' limitations in handling complex algorithms
- Need for parallel processing and specialized architectures

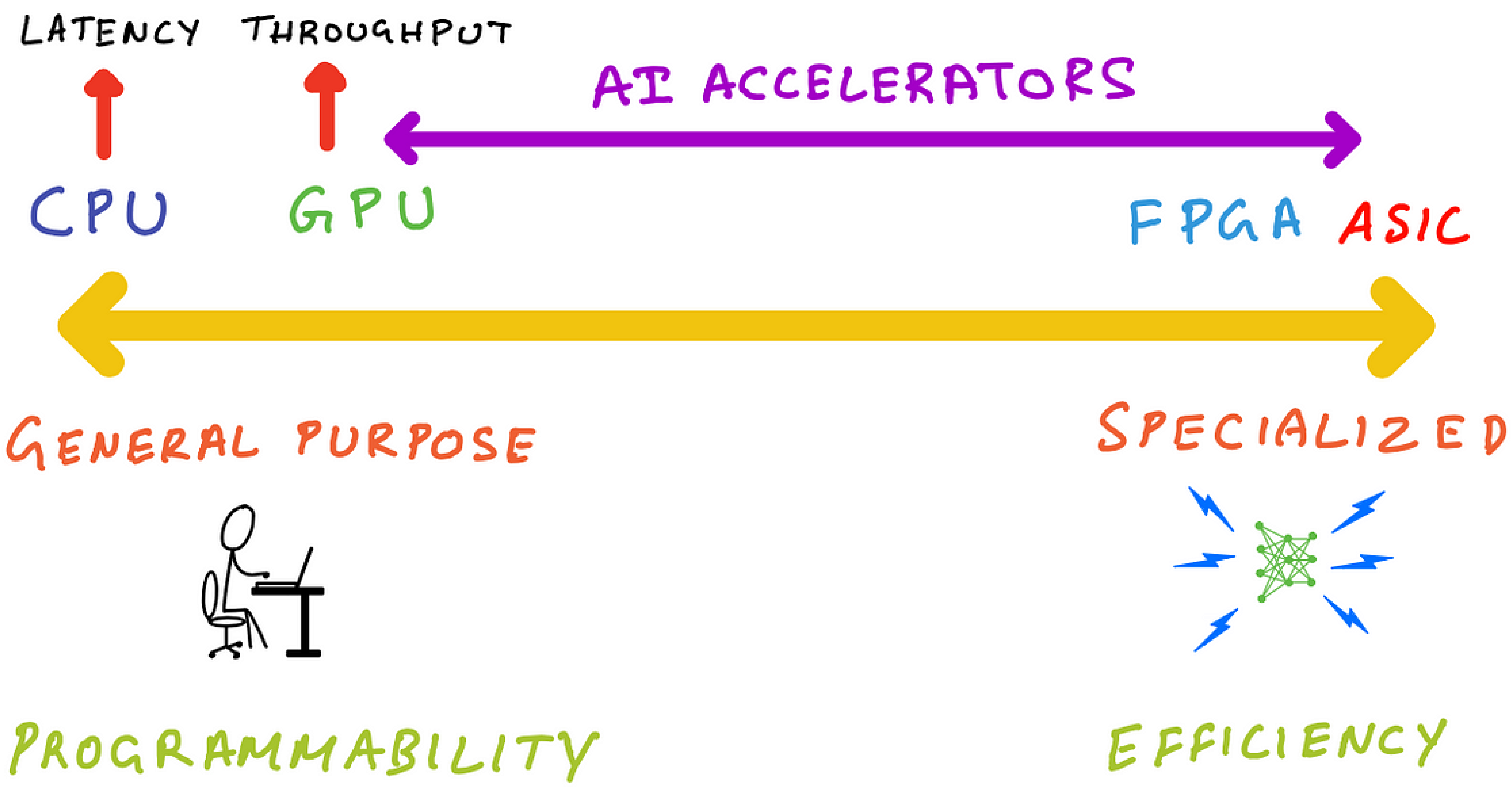


by Todd Nash

Types of Machine Learning Hardware Accelerators

- CPU (Central Processing Unit)
- GPU (Graphics Processing Unit)
- FPGA (Field-Programmable Gate Array)
- ASIC (Application-Specific Integrated Circuit)
- TPU (Tensor Processing Unit)

ML Accelerators



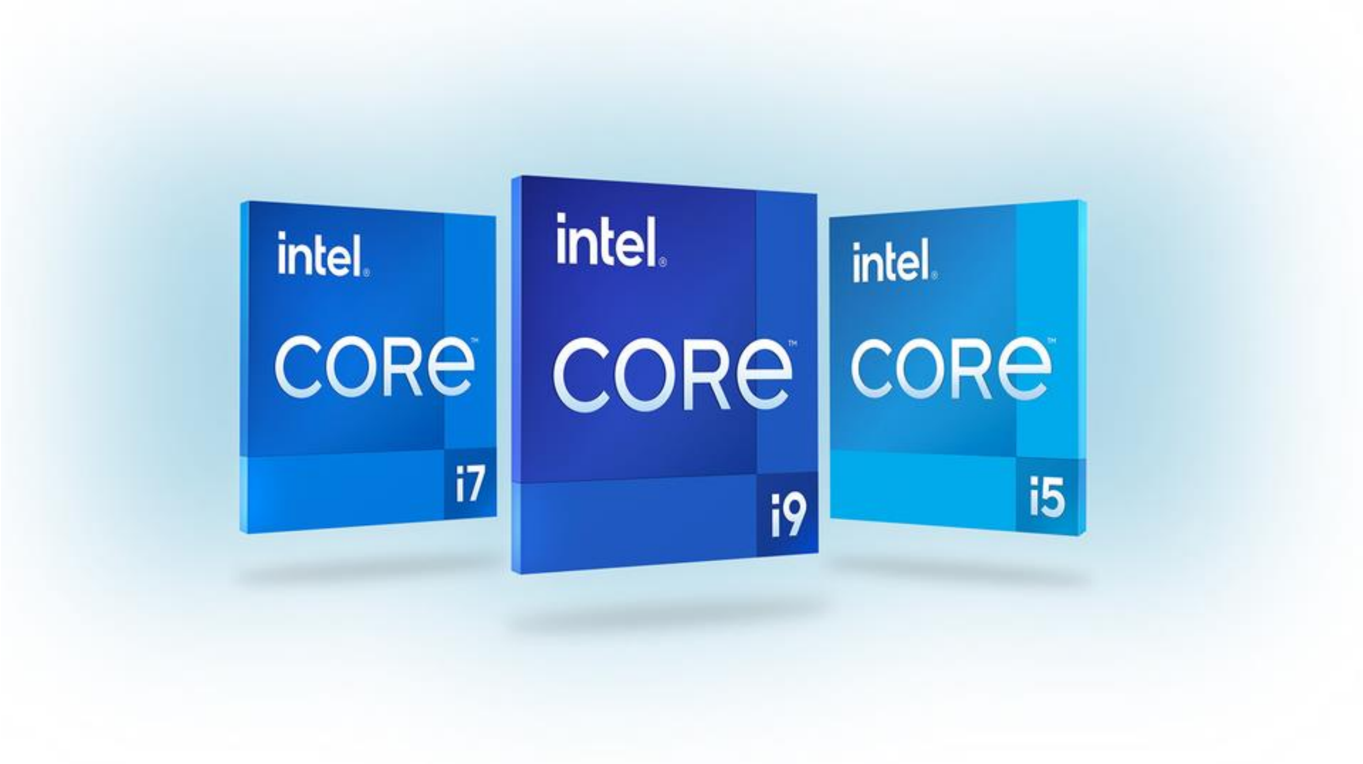
by Todd Nash

CPUs (Central Processing Units)

The **CPU, or Central Processing Unit**, is the primary component of a computer system responsible for executing instructions and performing calculations.

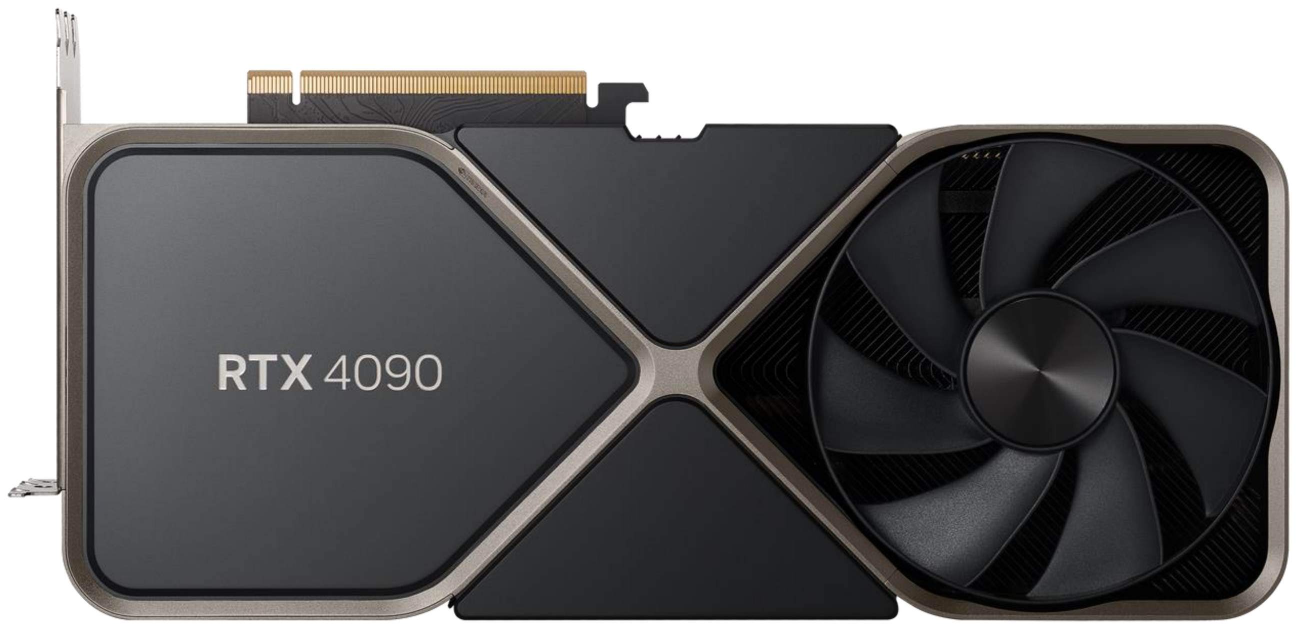
Functionality:

- Executes instructions fetched from memory.
- Handles arithmetic and logical operations.
- Manages input/output operations.
- Coordinates activities of other hardware components.



by Todd Nash

GPU (Graphics Processing Unit)



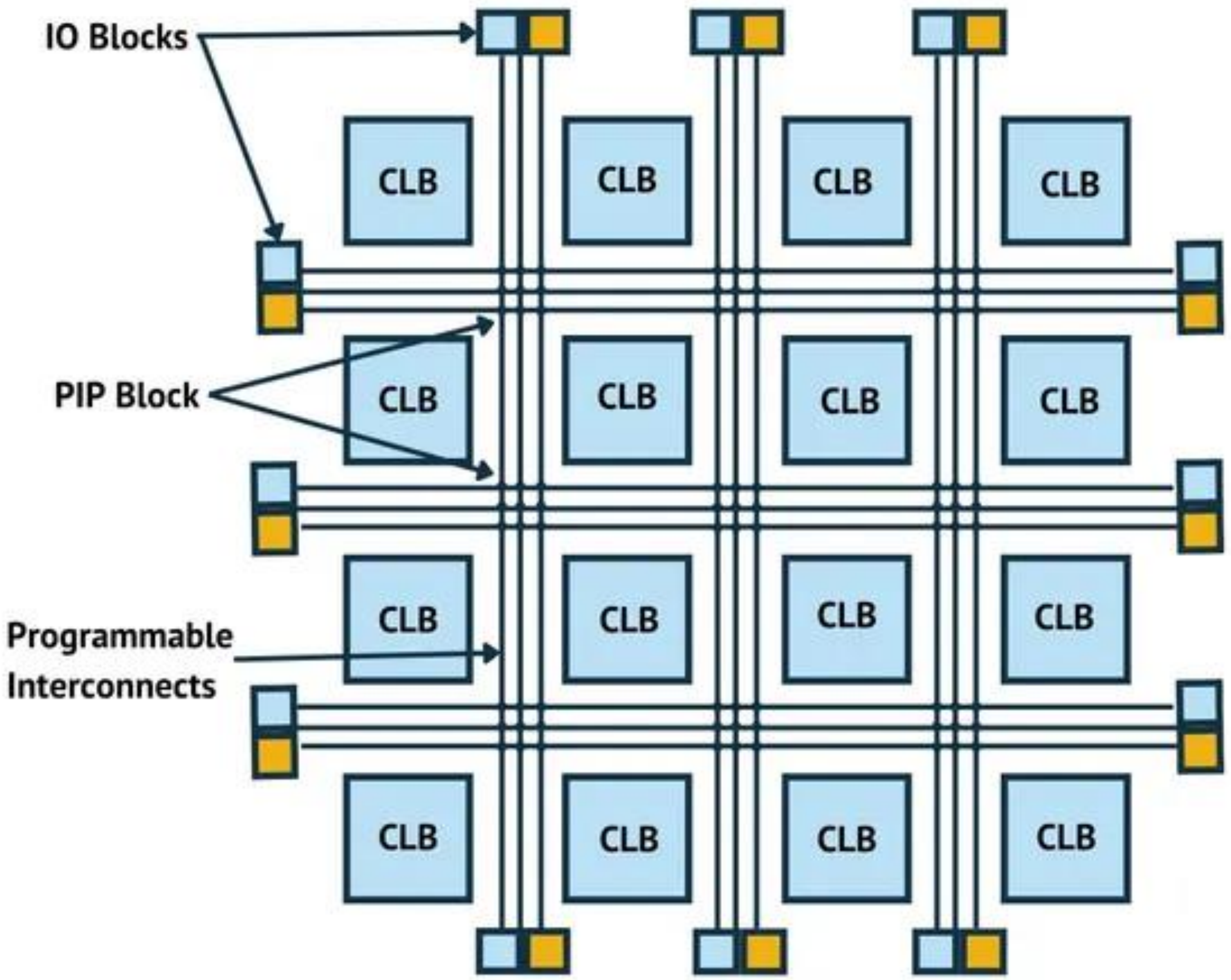
A **GPU, or Graphics Processing Unit,** is a specialized electronic circuit designed to rapidly manipulate and alter memory to accelerate the creation of images in a frame buffer intended for output to a display device.

Functionality:

- Originally designed for rendering graphics in video games and multimedia applications.
- Highly parallel architecture with thousands of cores.
- Optimized for parallel processing and *floating-point calculations.
- Efficient handling of matrix operations, crucial in many machine learning algorithms.

by Todd Nash

FPGA (Field-Programmable Gate Array)



FPGA, or Field-Programmable Gate Array, is a type of integrated circuit that can be configured and programmed after manufacturing.

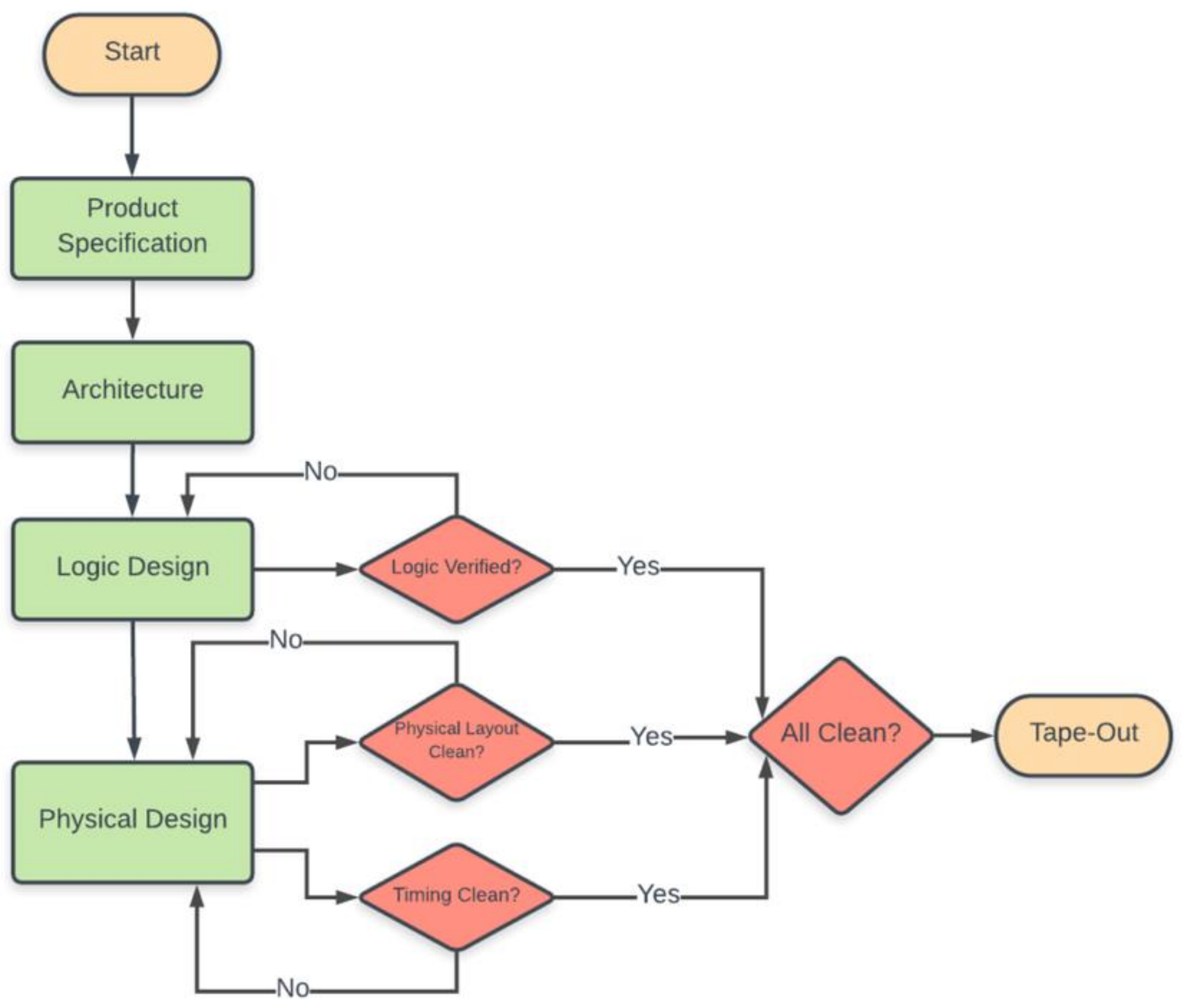
Functionality:

- FPGA consists of an array of configurable logic blocks and interconnects.
- Logic blocks can be programmed to perform various functions, such as arithmetic operations, digital signal processing, and memory storage.
- Interconnects allow for flexible routing of signals betwen logic blocks.

by Todd Nash

Red Deer Polytechnic

ASIC (Application-Specific Integrated Circuit)



ASIC, or Application-Specific Integrated Circuit, is a type of integrated circuit customized for a specific application or task.

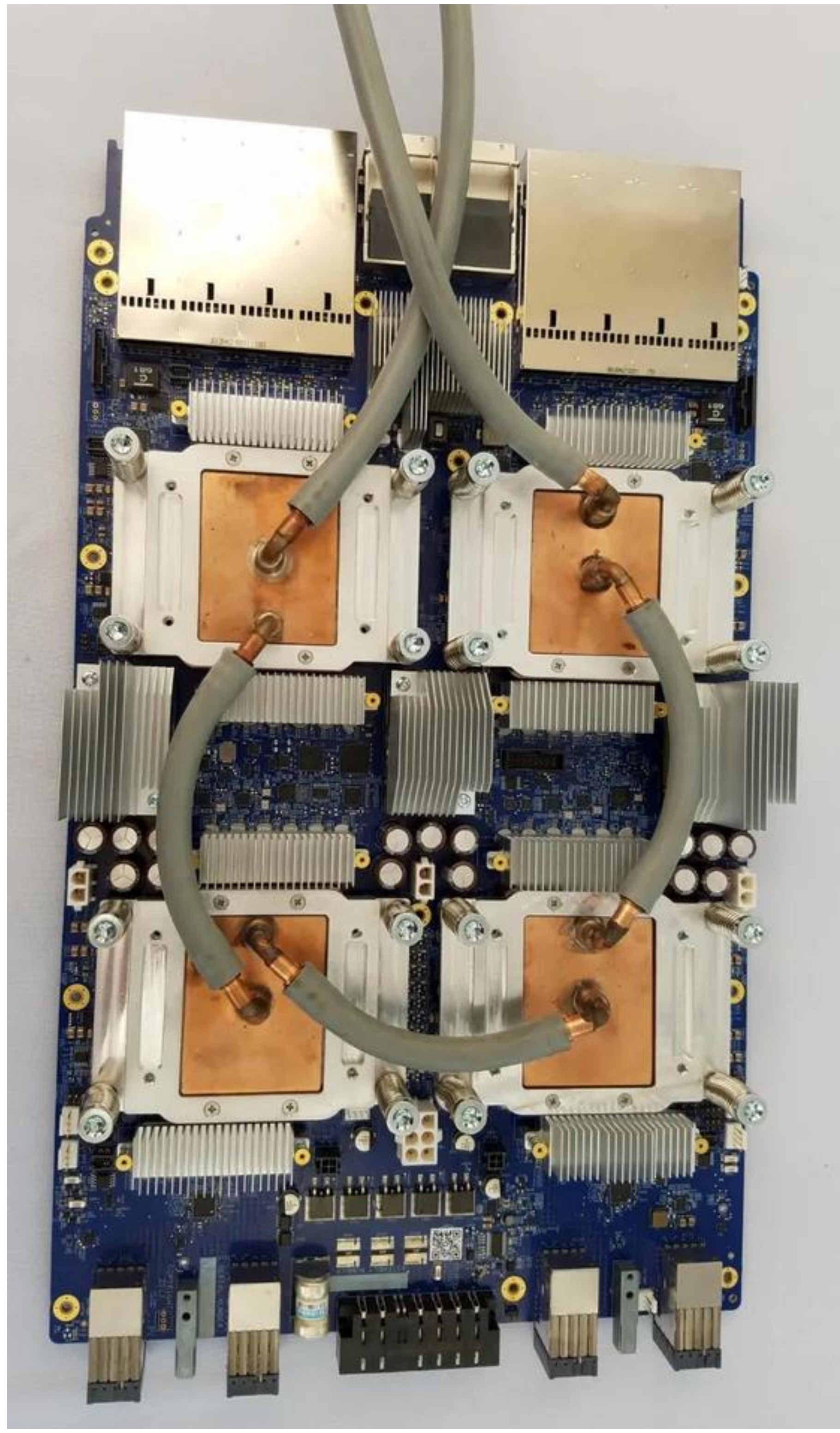
Functionality:

- ASICs are designed and optimized for a particular function or set of functions, such as cryptographic operations, signal processing, or machine learning inference.
- Unlike general-purpose processors, ASICs do not have programmable components and are tailored to perform predefined tasks efficiently.

by Todd Nash

Red Deer Polytechnic

TPU (Tensor Processing Unit)



TPU, or Tensor Processing Unit, is a custom-built ASIC designed by Google specifically for accelerating machine learning workloads.

Functionality:

- Optimized for performing tensor operations, which are fundamental to deep learning algorithms.
- Specialized hardware accelerators for matrix multiplication, convolution, and other common operations in neural networks.
- Designed to work seamlessly with TensorFlow and other deep learning frameworks.

by Todd Nash

Red Deer Polytechnic

Comparisons

CPU
- Small models
- Small datasets
- Useful for design space exploration

GPU
- Medium-to-large models, datasets
- Image, video processing
- Application on CUDA or OpenCL

TPU
- Matrix computations
- Dense vector processing
- No custom TensorFlow operations

FPGA
- Large datasets, models
- Compute intensive applications
- High performance, high perf./cost ratio

by Todd Nash

Red Deer Polytechnic

	CPU	GPU
Function	Generalized component that handles main processing functions of a server	Specialized component that excels at parallel computing
Processing	Designed for serial instruction processing	Designed for parallel instruction processing
Design	Fewer, more powerful cores	More cores than CPUs, but less powerful than CPU cores
Best suited for	General purpose computing applications	High-performance computing applications

by Todd Nash

Red Deer Polytechnic

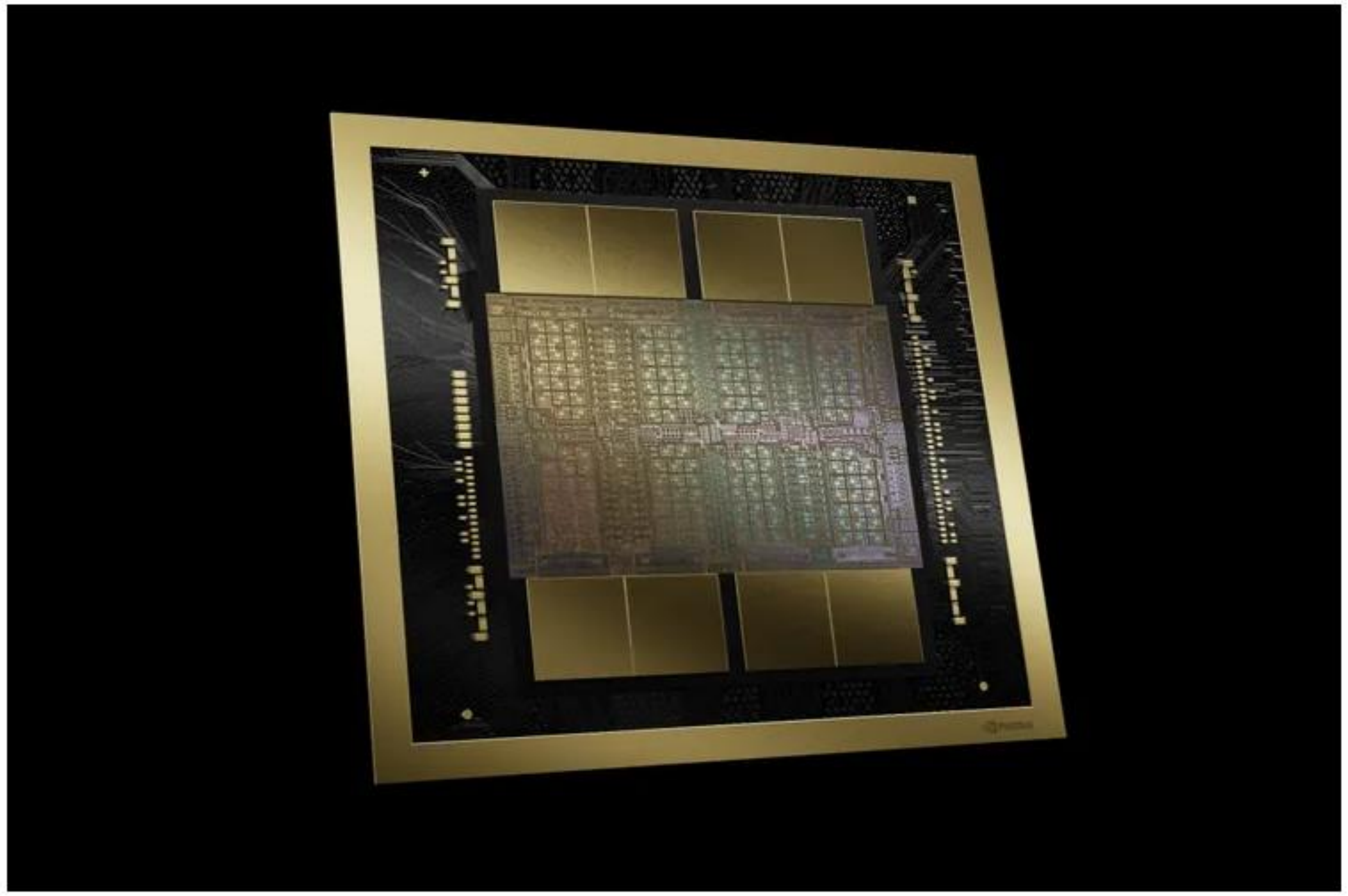
- Each type of machine learning hardware accelerator offers unique advantages and trade-offs.
- Selection depends on factors such as workload characteristics, performance requirements, cost considerations, and ecosystem support.
- Hybrid approaches combining multiple accelerators may provide optimal performance and flexibility for diverse machine learning tasks.

by Todd Nash

Red Deer Polytechnic

NVIDIA / TECH / ARTIFICIAL INTELLIGENCE

Nvidia reveals Blackwell B200 GPU, the ‘world’s most powerful chip’ for AI



The Blackwell B200 GPU. Image: Nvidia

/ ‘Built to democratize trillion-parameter AI.’

By Sean Hollister, a senior editor and founding member of The Verge who covers gadgets, games, and toys. He spent 15 years editing the likes of CNET, Gizmodo, and Engadget.

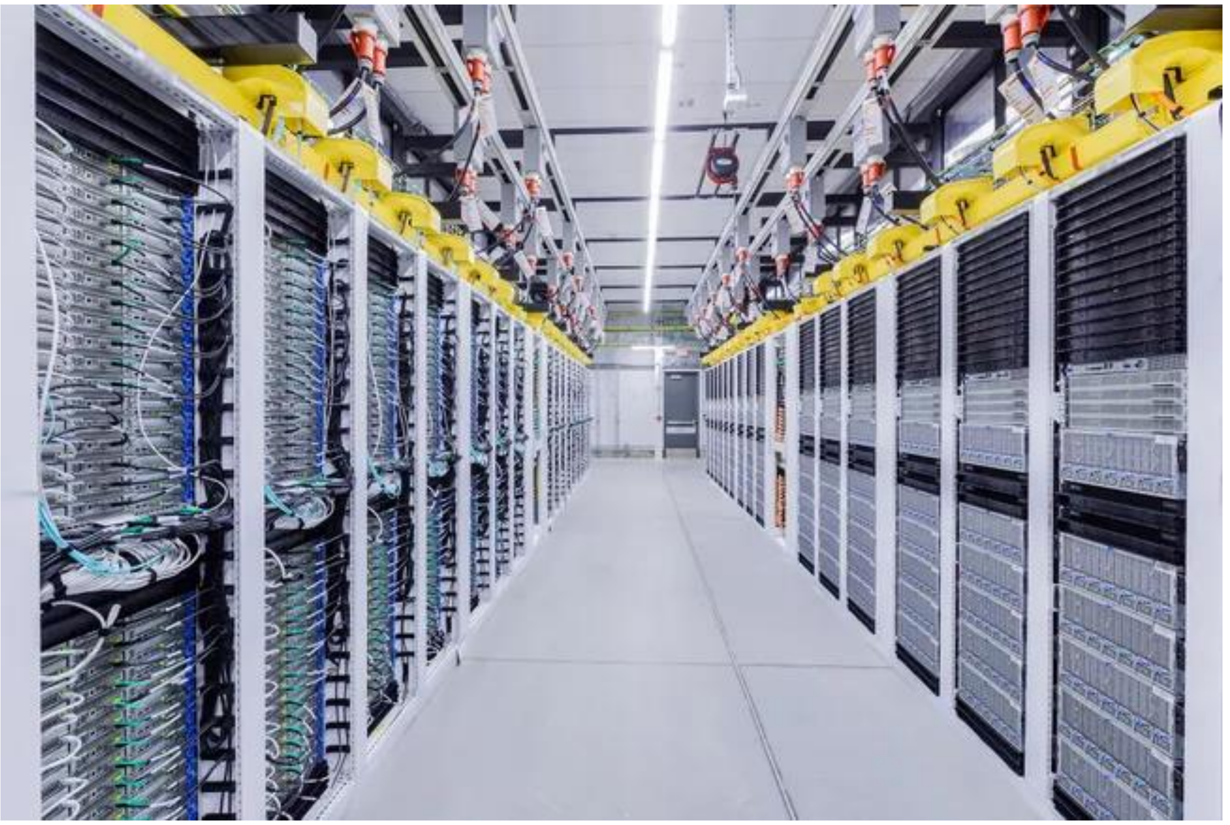
Mar 18, 2024, 2:39 PM MDT

53 Comments (53 New)

by Todd Nash

Red Deer Polytechnic

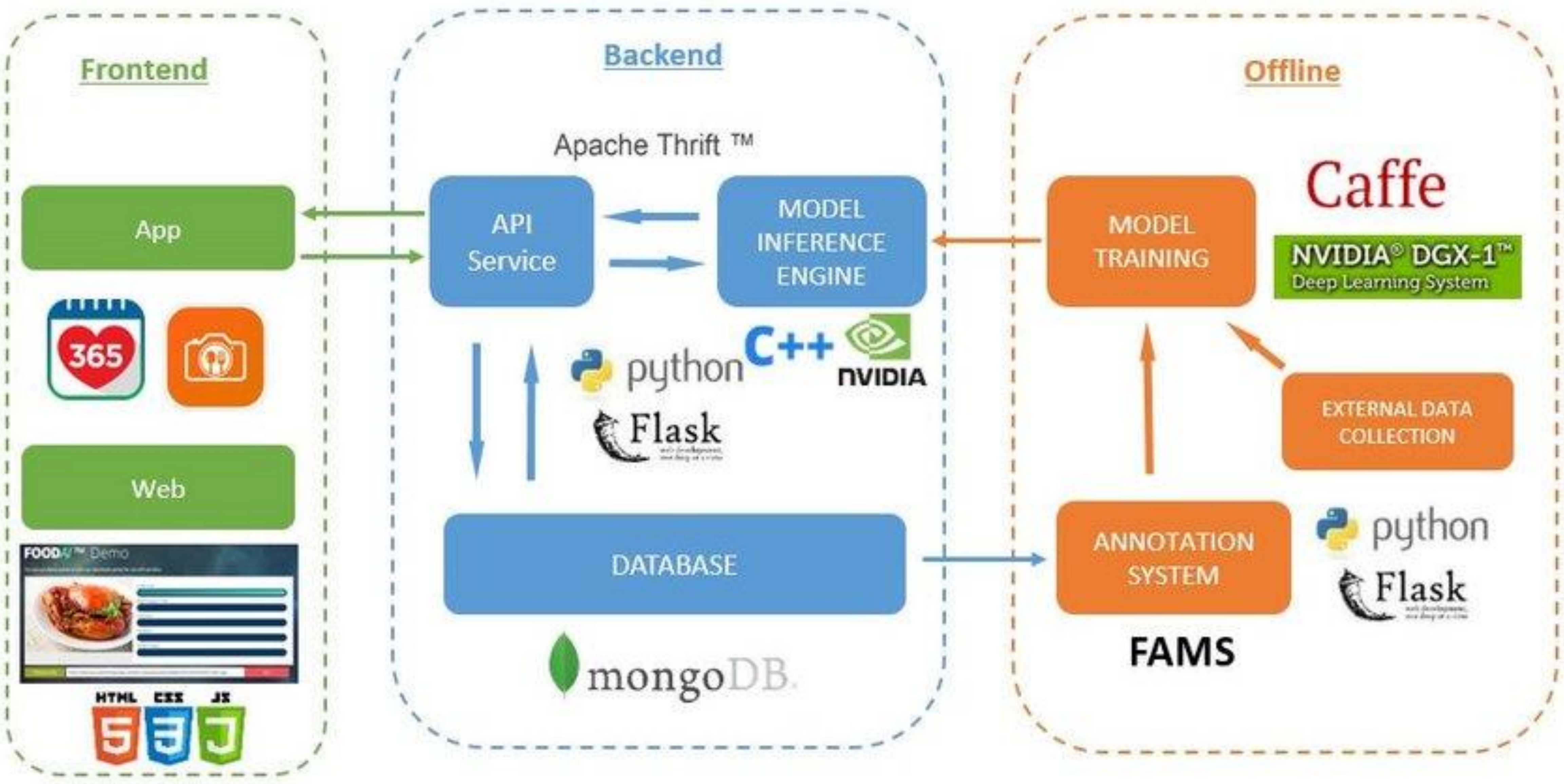
“OpenAI's ChatGPT is 'only possible' with Microsoft Azure,” says Microsoft.



Microsoft teamed up with NVIDIA, linking together thousands of NVIDIA A100 GPUs leveraging the firm's proprietary 400 GB/s Quantum-2 Infiniband networking technology. Working together, these GPUs can process and create inference and context for absolutely vast amounts of data — the kind of data needed to take the dumb chatbots of yesteryear to the eerily natural-sounding responses we see on Bing Chat and ChatGPT today.

by Todd Nash

Common Architectures



AI architectures typically refer to the underlying infrastructure and deployment setups used to support AI and machine learning applications.

by Todd Nash

Red Deer Polytechnic

Common Architectures

On-Premises Infrastructure:
- AI deployed in organization's data center.
- Full control over hardware and data.
- Requires significant upfront investment.

Cloud-Based Architecture:
- AI on third-party cloud platforms like AWS, GCP, Azure.
- Scalable, flexible, and accessible.
- Requires internet, raises data security concerns.

Edge Computing Architecture:
- AI on local devices or edge servers.
- Enables real-time inference with low latency.
- Ideal for IoT, autonomous vehicles, etc.

Hybrid Architecture:
- Combination of cloud and edge computing.
- Distributes AI workloads for flexibility.
- Requires careful orchestration.

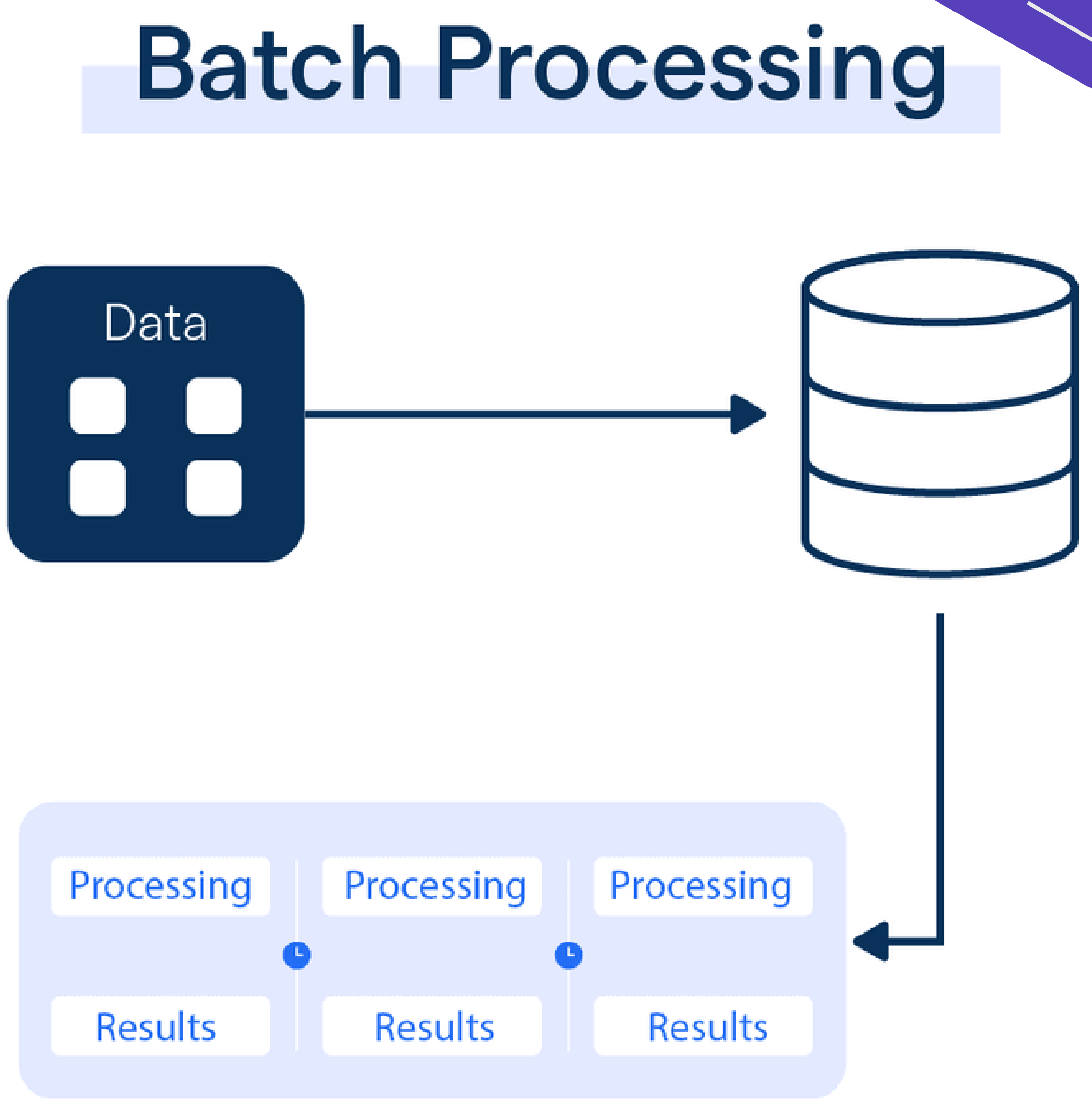
Serverless Architecture:
- AI deployed as stateless functions (FaaS - function as a service).
- Offers cost efficiency and scalability.
- Introduces resource usage constraints.

by Todd Nash

Deployment Strategies

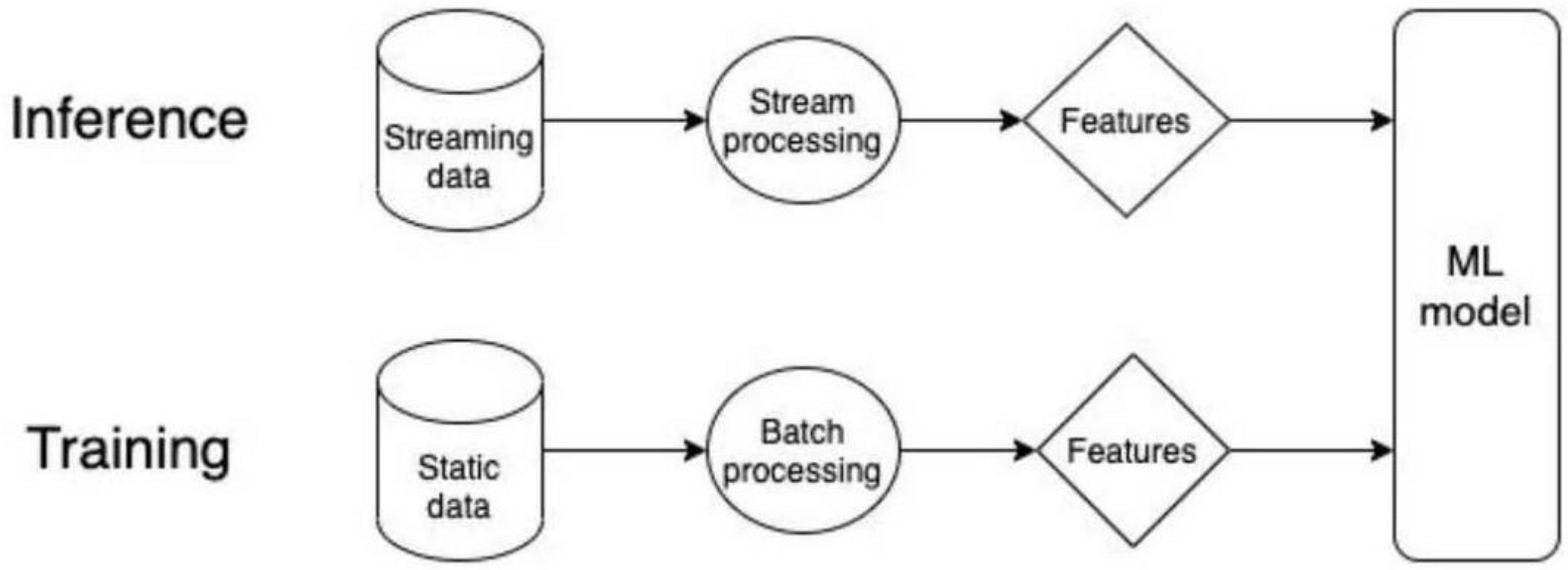
Batch Processing

- In this configuration, data is collected over a period of time, and machine learning models are trained periodically on the accumulated data.
- Once trained, the models are deployed to make predictions or classifications on new data.
- Commonly used for offline, non-real-time applications where latency is not critical, such as batch data analysis, recommendations, or periodic reporting.



by Todd Nash

Deployment Strategies



Real-Time Streaming

- In real-time streaming deployment, data is processed continuously as it arrives, and machine learning models make predictions or classifications in real-time.
- This setup is suitable for applications that require immediate responses to incoming data, such as fraud detection, anomaly detection, or real-time monitoring systems.
- Requires a stream processing framework like Apache Kafka, Apache Flink, or Apache Spark Streaming to handle continuous data streams.

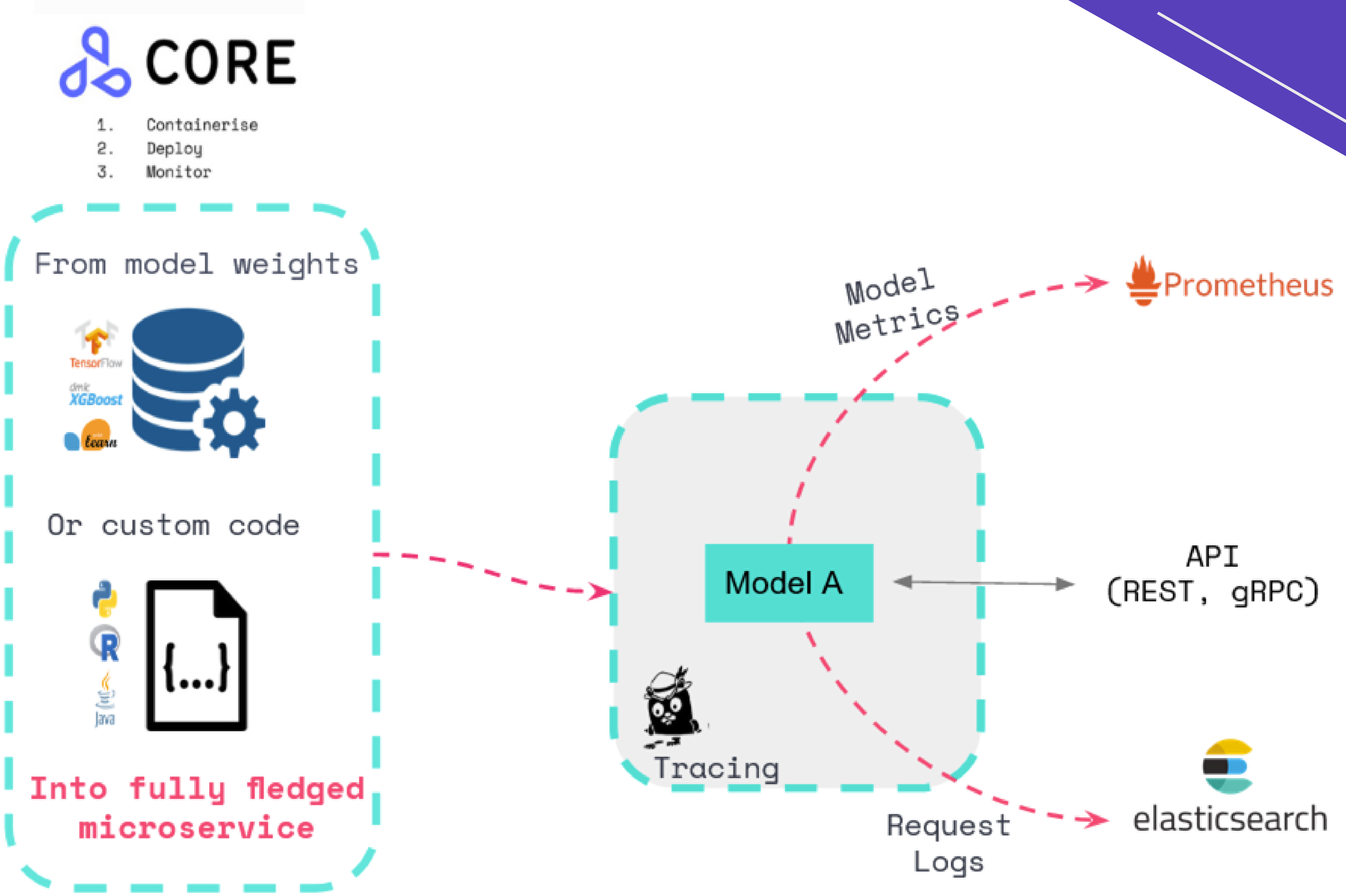
by Todd Nash

Deployment Strategies

Red Deer Polytechnic

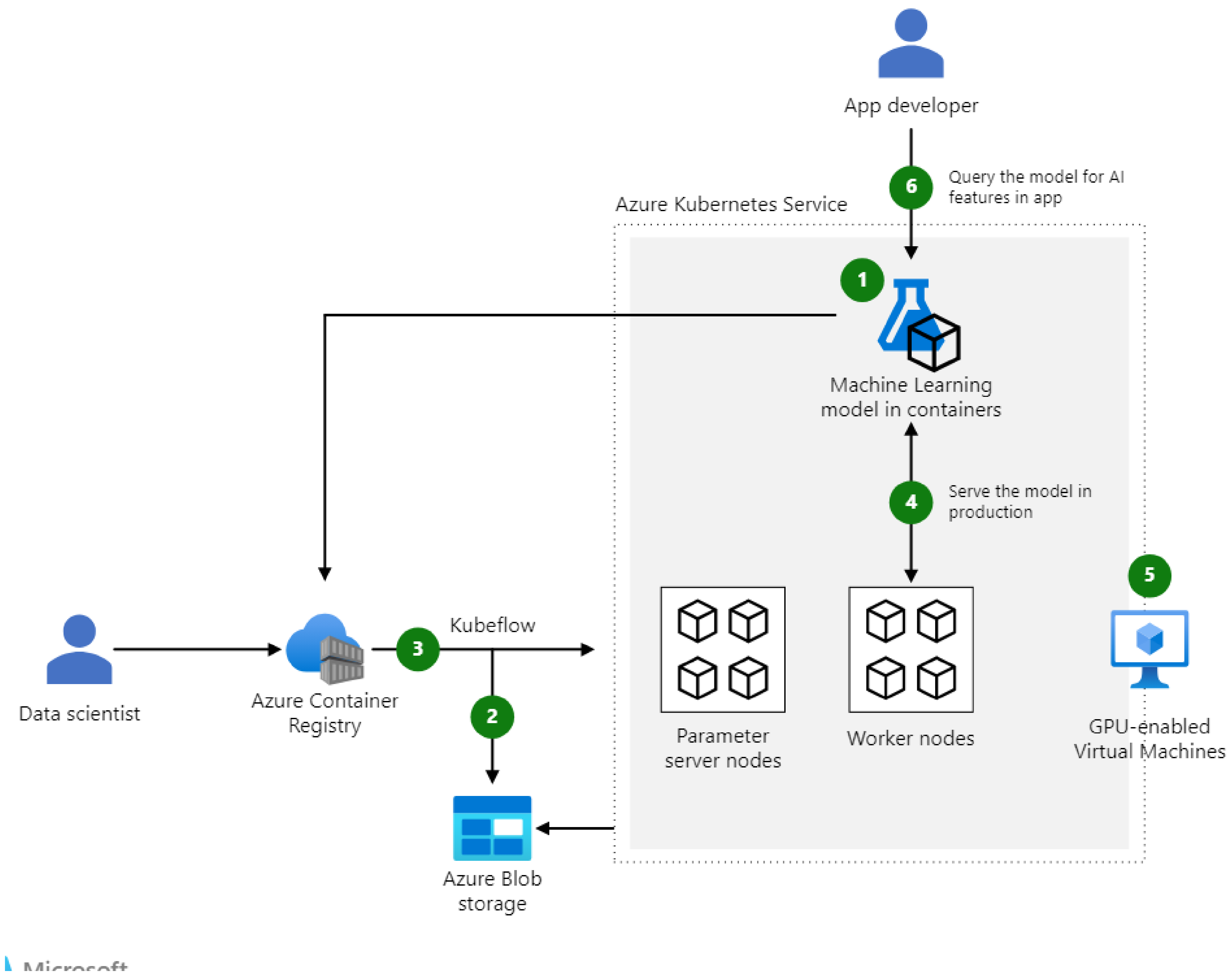
Microservices Architecture

- Machine learning models are deployed as microservices, which are small, independent, and loosely coupled services that communicate with each other via APIs.
- Each microservice may serve a specific function, such as data preprocessing, feature extraction, model inference, or post-processing.
- Enables scalability, flexibility, and ease of maintenance, as individual components can be updated or replaced without affecting the entire system.



by Todd Nash

Deployment Strategies



Containerization

- Packaging machine learning models, dependencies, and runtime environments into lightweight containers using technologies like Docker or Kubernetes.
- Containers provide isolation, portability, and consistency across different environments, making it easier to deploy models consistently across development, testing, and production environments.
- Allows for efficient resource utilization and scaling, especially in cloud or hybrid cloud environments.

by Todd Nash

Red Deer Polytechnic

ChatGPT-4, by the numbers.

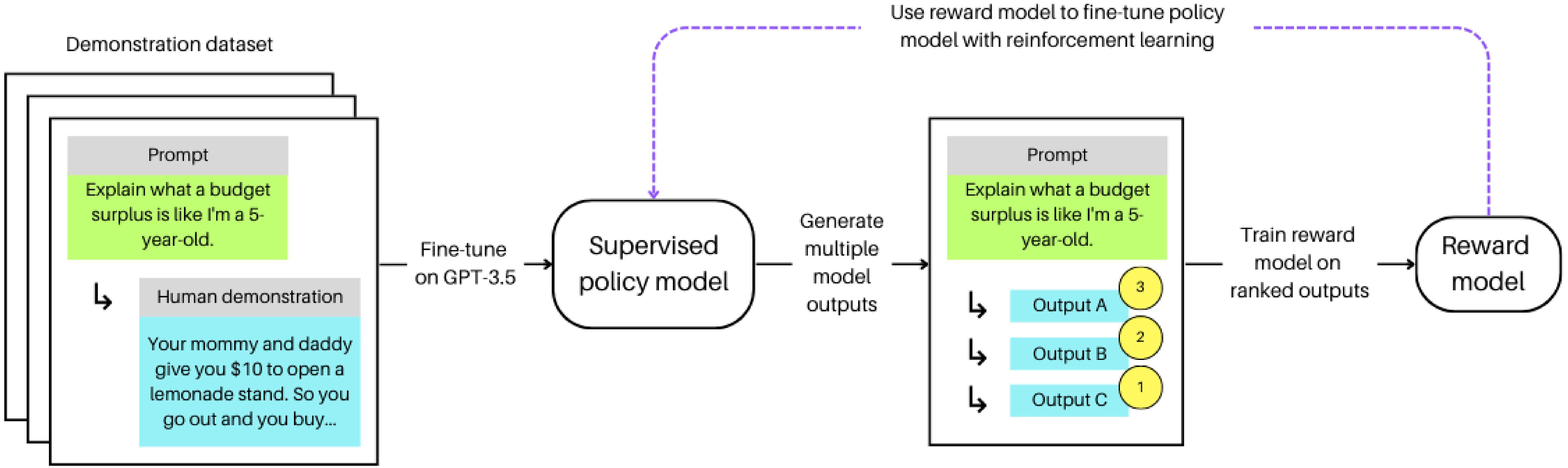
	GPT-3.5	GPT-4
Date of Training data	Up to **Sep 2021***	Up to **Sep 2021**
Type of learning methods	Unsupervised LLM	Supervised LLM
Number of parameters	**175 billons**	> **1 trillion** **
Type of Inputs	**Text**	**Text** and **Images**
Maximum Context length for request (Tokens)	Up to **4,096** ***	Up to **32,786**
Number of words can be process at once	Up to **3,000**	Up to **24,000**
Number of words in memory during a conversation	Up to **8,000**	Up to **64,000**

** October 2019 (text-curie-001, GPT 3 not GPT 3.5), June 2021 (text-davinci-002/003), & September 2021 (gpt-35-turbo or ChatGPT)*

*** Estimation based on assumptions.* **** text-davinci-002/003 (4,096 tokens), others (2,049 tokens)*

by Todd Nash

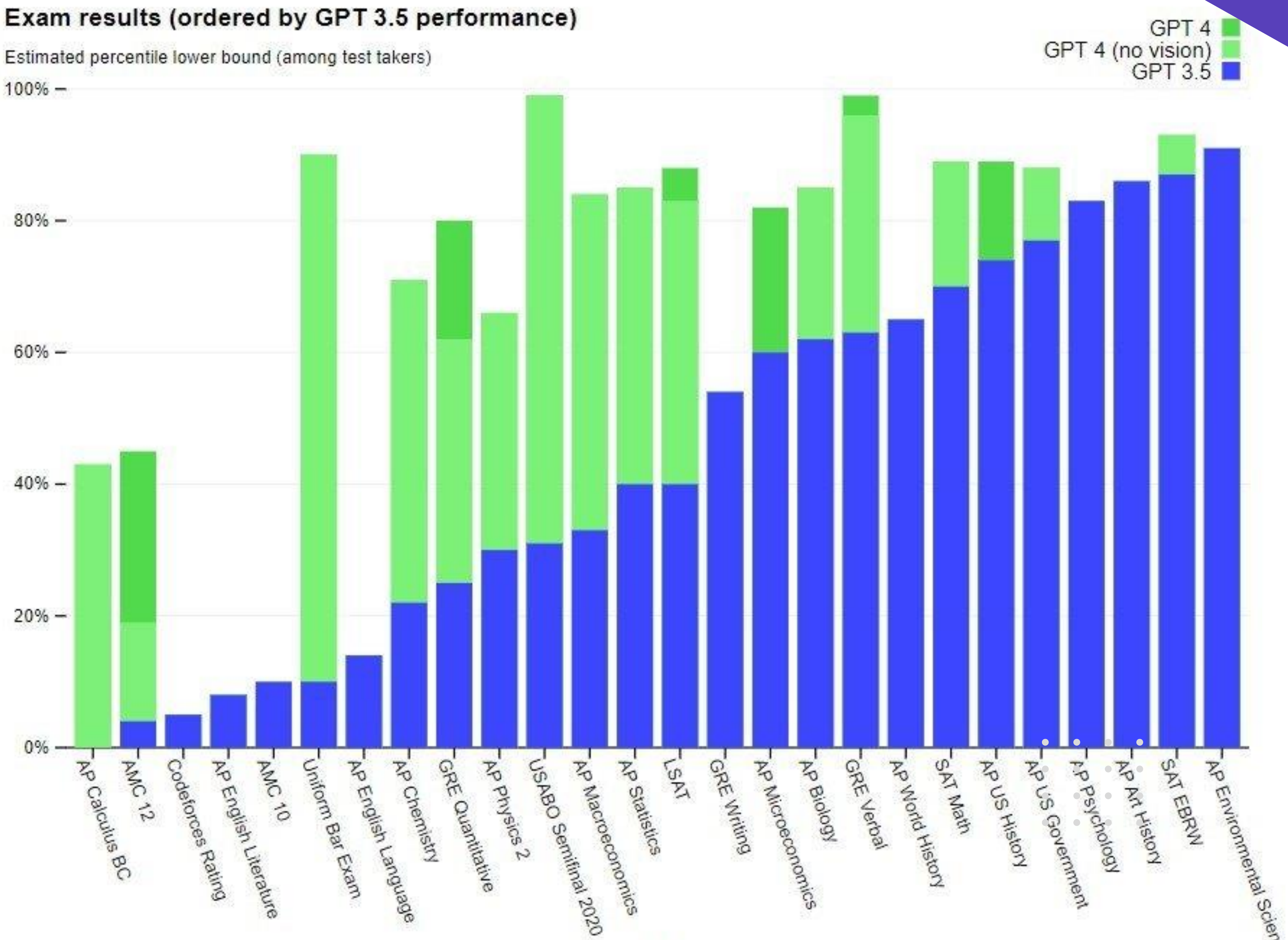
Training ChatGPT



by Todd Nash

Red Deer Polytechnic

ChatGPT Metrics



by Todd Nash