The Privacy Expert's Guide To Artificial Intelligence and Machine Learning

OCTOBER 2018

FUTURE OF PRIVACY FORUM

TABLE OF CONTENTS

INTRODUCTION

Advanced algorithms, machine learning (ML), and artificial intelligence (AI) are appearing across digital and technology sectors from healthcare to financial institutions, and in contexts ranging from voice-activated digital assistants, to traffic routing, identifying at-risk students, and getting purchase recommendations on various online platforms Embedded in new technologies like autonomous cars and smart phones to enable cutting edge features, AI is equally being applied to established industries such as agriculture and telecomm to increase accuracy and efficiency. Moving forward, machine learning is likely to be the foundation of many of the products and services in our daily lives, becoming unremarkable in much the same way that electricity faded from novelty to background during the industrialization of modern life 100 years ago.

Understanding AI and its underlying algorithmic processes presents new challenges for privacy officers and others responsible for data governance in companies ranging from retailers to cloud service providers. In the absence of targeted legal or regulatory obligations, AI poses new ethical and practical challenges for companies that strive to maximize consumer benefits while preventing potential harms.

For privacy experts, AI is more than just Big Data on a larger scale. Artificial Intelligence is differentiated by its interactive qualities—systems that collect new data in real time via sensory inputs (touchscreens, voice, video or camera inputs), and adapt their responses and subsequent functions based on these inputs. The unique features of AI and ML include not just big data's defining characteristic of tremendous amounts of data, but the additional uses, and most importantly, the multi-layered processing models developed to harness and operationalize that data. AI-driven applications offer beneficial services and research opportunities, but pose potential harms to individuals and groups when not implemented with a clear focus on protecting individual rights and personal information. The scope of impact of these systems means it is critical that associated privacy concerns are addressed early in the design cycle, as lock-in effects make it more difficult to later modify harmful design choices. The design must include on-going monitoring and review as well, as these systems are literally built to morph and adapt over time. Intense privacy reviews must occur for existing systems as well, as design decisions entrenched in current systems impact future updates built upon these models. As AI and ML programs are applied across new and existing industries, platforms, and applications, policymakers and corporate privacy officers will want to ensure that individuals are treated with respect and dignity, and retain the awareness, discretion and controls necessary to control their own information.

This guide explains the technological basics of AI and ML systems at a level of understanding useful for non-programmers, and addresses certain privacy challenges associated with the implementation of new and existing ML-based products and services.

What IS Artificial Intelligence?

There are many different definitions for artificial intelligence and remarkably little agreement on one standard description. Some examples include:

- "The science of making machines do things that would require intelligence if done by men."
  —*Marvin Minsky*
- "An umbrella term for the science of making machines smart."
  —*Royal Society*
- "The field of computer science dedicated to solving cognitive problems commonly associated with human intelligence, such as learning, problem solving, and pattern recognition."
  —*Amazon*
- "At base, for a system to exhibit artificial intelligence, it should be able to learn in some manner and then take actions based on that learning. These actions are new behaviors or features of the system evolved from the learnings."
  —*Omar Abdelwahed*

IN THE BEGINNING:
ALGORITHMS

Algorithms are not new. An algorithm is simply a step-by-step set of directions to accomplish a particular task, or achieve a pre-identified outcome. In a digital context, an algorithm is a series of instructions written by a programmer for software to follow. Typically, it details a defined series of steps to be followed verbatim. It may include sequences designed to apply only in the presence of specific facts or circumstances, but can only do so to the extent that those circumstances are foreseen and incorporated in advance. As described below, traditional algorithms can then be developed as the building blocks of machine learning and AI, but at their most basic, they are a piece of software programmed to repeat one process over and over, and in that original mode can only be changed by the intervention of a human programmer.

INTO THE FUTURE:
ARTIFICIAL INTELLIGENCE

While there is no one-size-fits-all definition, AI is generally understood to be a field of computer science focused on designing systems using algorithmic techniques, somewhat inspired by knowledge of the human brain, and capable of performing tasks that, if performed by a human, would be said to require intelligence. AI systems are intended to address challenges such as cognitive learning, problem solving, and pattern recognition. For a system to be said to exhibit artificial intelligence, it should be able to directly perceive its environment, evaluate and adapt to the data received, and respond by editing its own processes to, ideally, achieve better, more reliable outputs or predictions. Artificial intelligence can be further divided into two categories: general (or strong or full) AI and narrow (or weak or specialized) AI.

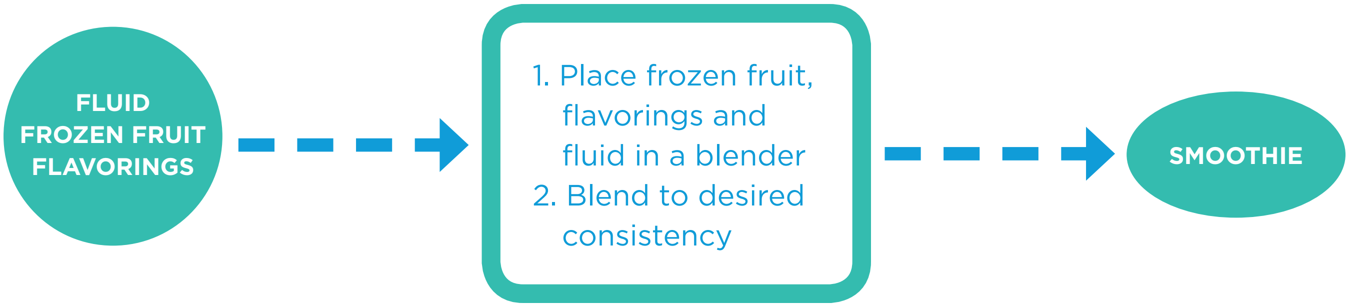

FLUID
FROZEN FRUIT
FLAVORINGS
1. Place frozen fruit, flavorings and fluid in a blender
2. Blend to desired consistency
SMOOTHIE

Figure 1.

A simple recipe is an algorithm. If digitized, the model can only complete the specified steps if provided with the required inputs.

GENERAL ARTIFICIAL INTELLIGENCE

General AI would be a system that is functionally equal (or superior to) human intelligence, and describes the notion of machines that can exhibit the full range of human cognitive abilities. The ability to generalize knowledge or skills—take the experiential value from one field and apply it in a different context—is so far strictly a human achievement.[1] There are many different proposed ways to evaluate whether a machine has achieved general AI (see box below). At this time, general AI is primarily a theoretical possibility, as current intelligent systems achieve superhuman performance only on narrowly defined tasks, and require at least some human engagement to be reapplied in a new context. Nevertheless, an enormous amount of research is underway toward achieving general AI.[2]

Some scientists, and many others who follow this field, believe that once machines reach a level of general AI, they will develop an accelerated ability to improve themselves, leading to subsequent, exponential increases in capabilities, with the possible outcome that the resulting AI-based entity will far surpass the abilities of humankind.

Tests for Confirming General AI[4]

Determining when a machine achieves human-level intelligence is a difficult task and there are competing ideas as to how to measure it. Below are some of the proposed tests for confirming when a machine attains general AI.

The Turing Test (Turing)
A machine and a human both converse, unseen, while a second human listens in and must determine which is machine, and which is human. The machine passes the test if it can fool the evaluator a specified percentage of the time.

Other Notional Tests:[5]

- **The Coffee Test (Wozniak)**
  A machine is required to enter an average American home and figure out how to make coffee: find the coffee machine, find the coffee, add water, find a mug, and brew the coffee by pushing the proper buttons.
- **The Robot College Student Test (Goertzel)**
  A machine enrolls in a university, taking and passing the same classes that humans would, and obtaining a degree.
- **The Employment Test (Nilsson)**
  A machine works an economically important job, performing at least as well as humans in the same job.
- **The Flat Pack Furniture Test (Severyns)**
  A machine is required to unpack and assemble an item of flat-packed furniture. It has to read the instructions and assemble the item as described, correctly installing all fixtures.

The fictional Voight-Kampff test from Blade Runner, *designed to detect replicants by measuring bodily responses to emotionally evocative questions.*

Scientists call this **artificial superintelligence** (ASI) and the moment of AI self-development is referred to as the **singularity**. If ASI is attained, the level and nature of its growth may impact human civilizations in ways that are entirely unpredictable at present, but which many in this community fear would be detrimental to the safety or dignity of humans. General AI is the most popular type found in science fiction, but whether the singularity will ever occur, and if so, whether it would be within years, decades, or centuries, is a matter of extensive debate. Regardless, it is of less immediate concern for policymakers and privacy officers focused on the current impacts on individuals and society of narrow AI and machine learning, as discussed below.[3]

NARROW ARTIFICIAL INTELLIGENCE

Narrow AI is artificial intelligence which applies only to one particular task. Narrow AI differs from general AI in that it is limited to solving specific types of pre-identified problems. For example, AlphaGo is a computer program which uses AI to play the board game Go. AlphaGo exhibits many traits of intelligence, including the ability to recognize new patterns, improve its functionality over time, and independently implement original strategies. While AlphaGo has defeated the best Go player in the world, it does not exhibit intelligence for any other task besides playing Go.[6] It cannot play a different game, like chess, or poker, without being retrained under the different rules and settings. Even more, it cannot apply its strategic understanding of Go to a wargaming scenario, unless that context were to follow the exact rules and terms of Go. Almost all public references to, or discussions about, current uses of artificial intelligence refer to narrow AI.

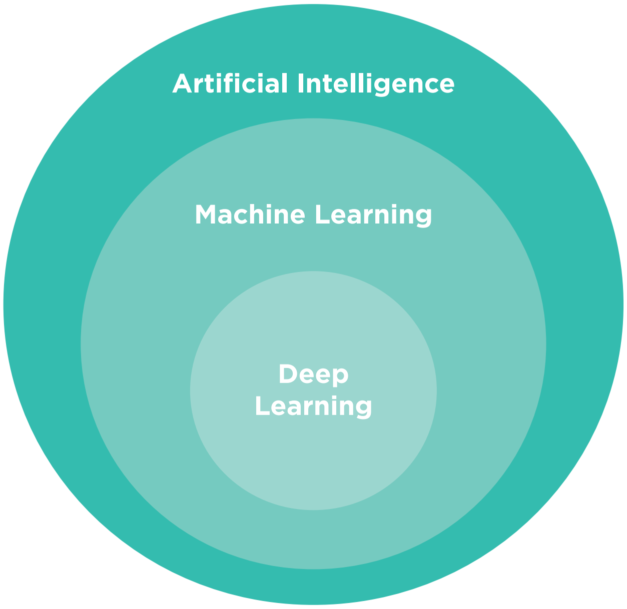


Figure 2.

Machine Learning is just one possible method of designing artificial intelligence. Likewise, Deep Learning is just one method of machine learning.

Other examples of narrow AI currently in use include

- Mapping apps that predict traffic flows and recommend routes based on real-time conditions and updates
- Autopilot systems in commercial airline flights
- Email spam filters
- “Robo-readers” that do an initial read and review of student essays
- Facial recognition systems that can identify unique individuals based on algorithmic facial scans

Importantly, it is not at all clear that narrow AI will just keep “getting better” and eventually expand into general AI. In fact, some researchers believe that the advent of general AI—if or when it happens—will not be based on current narrow AI techniques made more powerful or sophisticated, but will result from some other yet-to-be-discovered programmatic advancement on an entirely different path.[7]

IN THE PRESENT: MACHINE LEARNING

One specific subset of approaches to achieve AI is **Machine Learning**. Today, when companies or the media talk about the development and implementation of AI, they are likely talking about systems run on machine learning models, using algorithmic building blocks. As defined by Arthur Samuel, machine learning is the "field of study that gives computers the ability to learn without being explicitly programmed.[8]" ML involves teaching a computer to identify and recognize patterns by example, rather than programming specific, predetermined rules. This is accomplished in three steps[9]:

1. Evaluating extremely large sets of training data.[10]
2. Learning patterns from the training data, as evidenced through processing testing data.
3. Classifying new, previously unseen, data based on the patterns the system has identified.

What do we mean when we talk about a program "learning?" A machine learning program is different from an ordinary algorithm (represented by a computer program). Traditionally—for the case of non-learning algorithms—a computer programmer would explicitly define the rules (the code) according to which the algorithm was to behave. For maximum usefulness, this requires the programmer to envision and code for every scenario which the respective algorithm might encounter, resulting in program code that would be both unwieldy and often insufficient, as new or unforeseen situations continuously arise. Particularly for scenarios in which the rules for optimal decisions might not be clear, or which are difficult for a human programmer to articulate (e.g. identifying human faces on photographs), the usefulness of hard-coded algorithms reaches its limit.

Machine learning constitutes an alternative approach to providing decision-making systems with the rules that determine their outputs, such as predictions or classifications. In the case of self-learning systems, the human programmer leaves it up to the system to identify (and often continuously update) these rules by itself. Machine learning, therefore, enables an algorithm by itself to identify patterns (set the rules) underlying the decision-making processes, initially by being trained with data, then tested with more data, and finally applied in the real world. Machine learning enables a system to improve its output, or the accuracy of its predictions, with continued experience.

A human programmer will still be involved in this learning process, influencing its outcome through decisions such as shaping the specific learning algorithm created, selecting the training data provided to the system, and other design features and settings. But once operational, machine learning is powerful precisely because it can adapt and improve its ability to classify new data without direct human intervention—in other words, **the quality of output improves with its own experience**.[11] This difference:

- Allows the creation of more powerful and comprehensive algorithms,
- Means ML systems may be retooled and adapted to new and different tasks,[12]
- Applies whether or not models use or generate personal information (many do not).

What is easy for humans is frequently hard for machines, and the reverse is equally true. Chess is hard for many humans, but relatively easy for machines. A baby can easily recognize faces, with a fair degree of sophistication, while machines find this extremely challenging. Machines can store, index, reference and search an almost limitless number of "facts," but it is difficult to teach them to collate those facts in a way that supports a prior supposition.

Consider, humans can easily tell the difference between apples and oranges with little more than a glance, but programming a computer to do so is difficult. Conventional programming based only on algorithms would require accounting of an almost impossibly large number of characteristics. Even if a fairly reliable algorithm for identifying apples and oranges could be created through conventional programming, the system would be fragile and inflexible. Adjusting it to consider unrelated forms of data, to discriminate among types of apples or oranges, or to account for outliers in color or size, would all require enormous amounts of coding that would still eventually be insufficient. And if one wanted to identify pears instead, the code would have to be extensively updated.

Big Data and Artificial Intelligence

"The algorithm without data is blind. Data without algorithms is dumb."
—CNIL

The recent development of machine learning systems is almost entirely driven by the availability of extremely large data sets, and the general ease of both collecting and managing this data. Immense quantities of data, along with the exponential increase in computing power, have enabled the last decade's surge of "smart" systems—that is, systems that continuously improve themselves by learning from vast amounts of data including identifiers, attributes or behaviors.

Unfortunately, the privacy challenges of big data are well-known—the increased difficulty in protecting or screening out personal data; the increased difficulty in deidentifying data within datasets; and the greatly increased possibilities for reidentifying individuals based on comparing data across data sets.[14]

In addition, the need for large amounts of data during development as "training data" creates consent concerns for individuals who might have agreed to provide personal data in a particular commercial or research context, without understanding or expecting it to be further used for new algorithmic design and development.

In addition to the technical and procedural protections discussed later, a data-focused solution in some contexts is the ability to create "synthetic data." This is an artificial data set, including the actual data on no "real" individuals, but which mirrors in characteristics and proportional relationships all the statistical aspects of the original dataset. This is an area still under consideration and development.[15]

DATA MINIMIZATION[16]

The traditional privacy principle of "data minimization" must be carefully considered and when regulated, handled thoughtfully in light of the huge quantities of data required to develop ML models. Data minimization refers to the practice of limiting the collection of personal information to that which is directly relevant and necessary to accomplish a specified purpose. From a technical standpoint, more data (a larger sample of records) is frequently essential for sufficient training, in particular to prevent biased outputs. However, from a privacy perspective, eliminating irrelevant information remains as important as ever.

AI/ML systems are necessarily designed to handle the large amounts of data required to identify patterns and connections and provide precise outputs and analysis. And since ML-based systems are not limited by a human capacity to process and understand patterns in the underlying data, more data significantly improves the diversity (referring to the data type or attribute variety) required for training systems that can ultimately provide great advancements in applications ranging from medical research to personalized services.

Nevertheless, the data minimization principle remains applicable, and requires that the data collected is appropriate and limited to what is necessary for the particular purpose. Even in machine learning contexts, including unnecessary data remains a risk, both to the accuracy and efficiency of the system, and to the individuals from whom the data is derived. Ensuring the exclusion of irrelevant data prevents algorithms from identifying correlations that lack significance, or are coincidental. Where possible, pseudonymisation or homomorphic encryption techniques[18] may protect an individual's identity and help limit the exposure of their personal data. The opportunities presented by these applications must be balanced against the necessary protections for individuals.

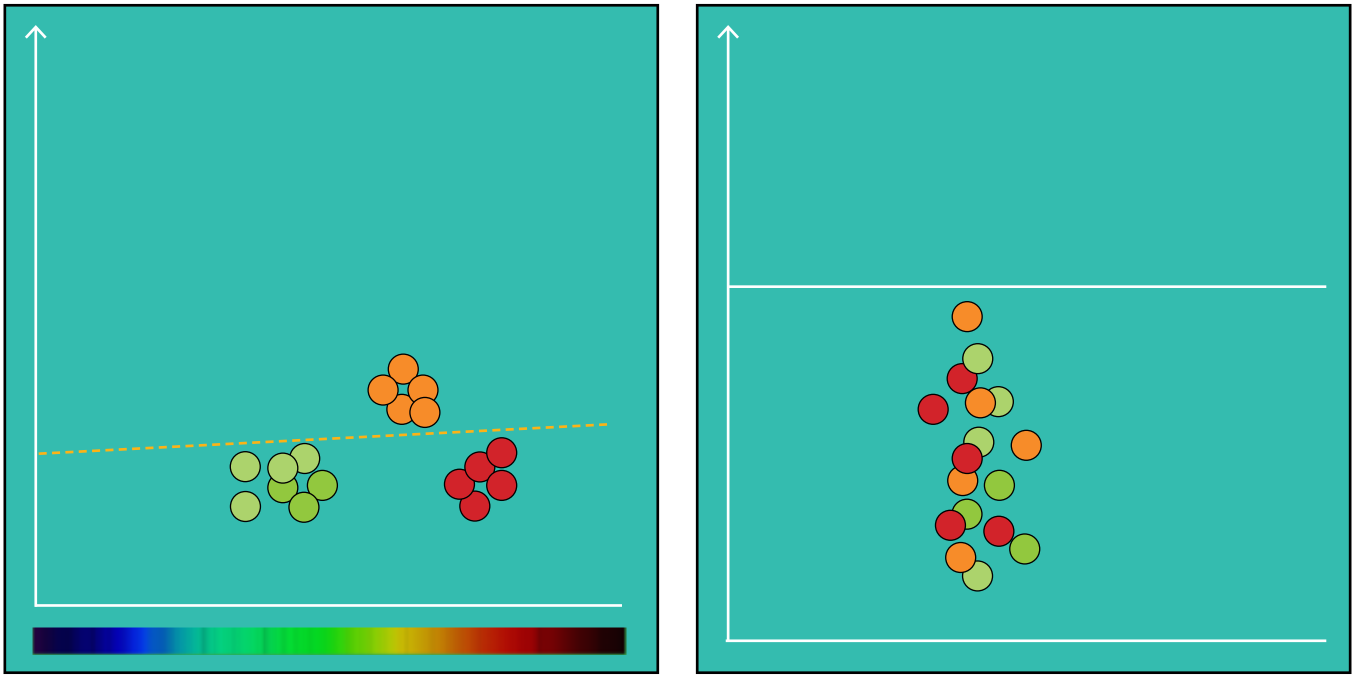


Figure 3.

Dividing apples and oranges and others based on the feature of color results in a meaningful output.[19]

Dividing apples and oranges based on the feature of size makes distinguishing them much harder.[20]

In contrast, a machine learning system is designed to account for multiple variations, and to adapt to new elements. This allows for a far greater number of factors to be considered, and offers greater precision with significantly less detailed inputs. Once a machine learning program has been trained, the method it uses to recognize patterns in new data is called a model. The model is what is used to perform tasks moving forward, with new data. Identifying apples and oranges would be accomplished by employing a model which had been trained for this task. One model may potentially be (re)trained for a different task by using different data, without the need to rewrite the entire code. For example, if one wanted to identify pears instead of apples, new training data containing pears would be used to train the system—but there would be less need to recode the entire process.

In general, training data is provided to a ML system, from which it learns how to recognize specific patterns. However, a machine learning model is only as effective as the quality of its training. It takes huge amounts of data to enable a computer to reach reliable outputs on new data without human involvement. The patterns that a machine learning system identifies depend upon the goals of the program. For instance, returning to our example, if one wishes to distinguish between images of apples and images of oranges, images consisting of apples and oranges must be used and the computer is trained to differentiate between them, as well as to understand that some images might be neither. It does not simultaneously differentiate whether the fruit is on the tree, in a package, or on a table, although those are patterns which may also be trained for, and "learned."

The elements of the data which are used to detect patterns can vary. These elements, called features, are the attributes that either the programmer identifies, or that the machine extrapolates, and will have an effect on the target variable. Deciding what features are used in the machine learning process is very important to the success of the model. For instance, in distinguishing between images of fruit, color could be an important feature. Other features may not matter, or may even decrease the usefulness of the model. The size of an apple or orange has little relevance in distinguishing them from each other. In some AI applications, the AI program decides which features are most effective. The number of features that a machine learning system uses in its model is referred to as how many dimensions it has.

Some examples of Machine Learning programs in common use include:[21]

- Voice-operated personal assistants
- Ridesharing apps (predict wait times, demand, and surge pricing)
- Smart email categorization tools, sorting non-spam email into sub-levels in an inbox
- Plagiarism detectors, particularly for identifying plagiarism from non-digitized sources
- Identifying at-risk students
- Mobile check deposits (combination of AI and ML)
- Credit card transaction fraud detection
- Purchase (or viewing) recommendations

There are three main categories of ways to train a machine learning model: **supervised learning**, **unsupervised learning**, and **reinforcement learning**.

SUPERVISED LEARNING

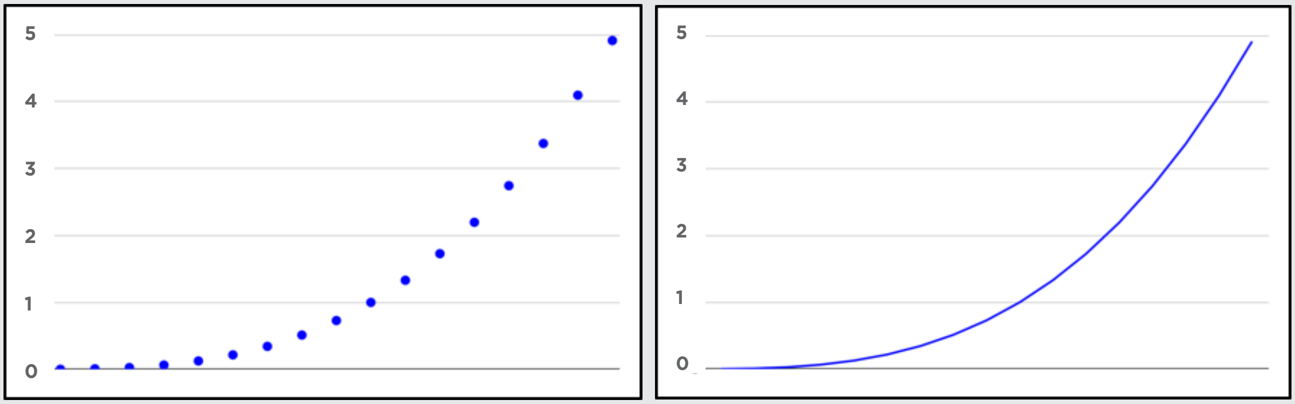
One way to train machine learning systems is by providing training examples in the form of labeled data. Any time the data is labeled ("this is an orange" or "this is an apple") before being processed through the system, it is considered "supervised learning." The system is then instructed on how the labeled data is to be categorized. In this manner, an algorithm is created which learns how to identify specific features that can be applied to previously unseen data. For example, images of fruit which are labeled red and spherical are also labeled as apples. Images labeled orange and spherical are also labeled as oranges. After the system has viewed enough images already labeled as apple or oranges, it can apply the model to new image data, which it hasn't seen before. The hope is that the model can identify apples and oranges from the new images based on patterns it detected in the training data.

There are many different types of evaluative tools which can be used by a supervised machine learning model. Some of the most common are **regression**, **classification**, and **decision trees**.

Discrete and Continuous Variables

Continuous variables are variables which have values without gaps between them, such as height or weight. Continuous variables have an infinite number of values. For example, weight can be measured in any increment between 180 and 181 pounds.

Contrast this with **discrete variables**, those with a finite number of possible values, such as the value of a die or number of people in a room. There cannot be a fraction of a person present, so there is a finite number of output values. On a chart, continuous variables are represented by a line, while discrete variables are represented by points plotted on it.



Discrete variables

Continuous variables

Figure 4.

The concept of continuous and discrete variables is applicable in Classification and Regression models.

REGRESSION

Regression is an existing statistical tool for evaluating data, that can also be used to support ML models. A regression-based model attempts to estimate a value based on the input data, when the target variable is measurable by a continuous metric. Simplistically, an example of regression would be predicting how the consumption of apples affects life expectancy.[22] Using regression, a machine learning system can be fed training data which shows how many apples per day are consumed by an individual and how long each individual lives. Both years and apples may exist in portions less than one (e.g. every half an apple per day extends life expectancy by one month), so these metrics are continuous. These two data fields (apples per day and life expectancy) are the **features** used by this model.

A machine learning system essentially teaches itself from the training data exactly how much the consumption of apples affects life expectancy, allowing it to make predictions about the target variable Y (life expectancy) based upon just the input data X (number of apples eaten per day). As the system encounters new training data, it iteratively improves upon its approximation of Y.

Examples of ML-based regression analysis include predicting the number of people who will click on an ad based on the ad content combined with data about the user's prior online behavior; predicting the number of traffic accidents based on road conditions and speed limit; or predicting the selling price of real estate based on its location, size, and condition.[23]

Subject	Average number of apples eaten per day	Age at death
1	1	65
2	0.5	67
3	1.5	70
4	2	65
5	0.5	60
6	2	75
7	2.5	73
8	2	71
…	…	…
N	X	Y

FIgure 5.
An example of X and Y data.

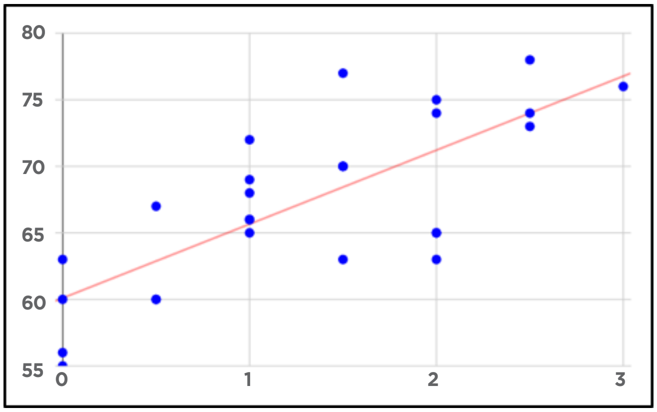


FIgure 6.
An example of linear regression.

It is a defining feature of linear regression that it produces continuous numerical predictions that do not have to be whole numbers. Thus it is most useful for cases where the desired variable output can be any value on the spectrum—a price for a product, a measure of distance, overall revenue, age, and so on.

How well a model reacts to new data after being trained is called its **generalization ability**.[24] The better a model can achieve correct outputs from new data, the more useful it is. Sometimes, when a model has been trained by data that is overly specific, it will fail to effectively generalize for new data. This is called **overfitting**. When the training data is overinclusive of features to derive the general rule, which are then incorporated into the model to evaluate each new instance, the rule becomes perfect for past data, but useless moving forward.

"Overfitting means trying to be too smart. When predicting the success of a new song by a known artist, you can look at the track record of the artist's earlier songs, and come up with a rule like 'if the song is about love, and includes a catchy chorus, it will be top-20'. However, maybe there are two love songs with catchy choruses that didn't make the top-20, so you decide to continue the rule '...except if Sweden or yoga are mentioned' to improve your rule. This could make your rule fit the past data perfectly, but it could in fact make it work worse on future test data."[25]

Machine learning models are vulnerable to overfitting because they can easily identify and accommodate a large variety of rules in combination to "perfectly" describe past events. Balance is required, and it takes experience and skill to design and train a model with sufficient data for effective and unbiased generalization ability, while avoiding overfitting.

CLASSIFICATION

Sometimes, a model is required to sort data into discrete categories. Classification is the method of assigning data to the class to which they most likely belong. Classification is accomplished through the use of supervised learning, with the categories created from labelled data.[26] For example, distinguishing between images of apples and oranges is classification, where the user defines the classes as “apples” and “oranges.” The model is trained with labelled data to give the probability that each new image is an apple or an orange and classifies them into a category. A more sophisticated version would include an “other” category, such that data presented which did not meet a specified threshold of similarity to either apples or oranges, would be rejected from either label. Without that option, the model would put every new image into either apple or orange, even if the similarity was miniscule.

Classification differs from regression in that it attempts to predict a *discrete* class label,[27] where regression models predicted relationships using a continuous metric.[28] Classification is about predicting one specific label (a discrete value) and regression is about predicting a relationship between the variables, usually expressed as a quantity (on a continuous scale).[29]

There are actually several kinds of classification models. **Logistic regression** is one model, which gives the probability of the target variable belonging to a certain class.[30] This can include

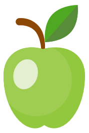
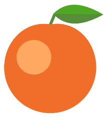
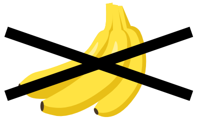
Model	No. of Classes	Classes
LOAN APPLICATION	2	“approved” or “denied”
APPLES AND ORANGES	3	 apple  orange  neither
HANDWRITTEN DIGITS	10	0 1 2 3 4 5 6 7 8 9
FACIAL RECOGNITION	>>1*	
		* A whole

FIgure 7.

Examples of classification models.

binary classification when there are only two possible classes, but it can be scaled up to an enormous number of groups.

Logistic regression calculates the chance that each input belongs to a certain category. For example, a logistic regression model could be designed to estimate the probability that each image it receives is an image of an apple. The same model could also be used to estimate the probability of each image being an orange. Generally, if a logistic regression model determines the probability is greater than 50 percent, the decision is true.[31] Thus, in this example if the model returns a 20 percent chance of an image being an apple and a 70 percent chance that it is an image of an orange, it would classify it as an orange and not an apple. If the probability of an image is less than 50 percent for both classes (apples and oranges), it will classify it as neither.[32]

Examples of Logistic Regression:

1. Binary Logistic Regression
   The categorical response has only two possible outcomes.
   Example: Spam or Not Spam

2. Multinomial Logistic Regression
   Three or more categories without ordering.
   Examples: Predicting which food is preferred more (pescatarian, vegetarian, vegan);
   an autonomous car determining what a particular traffic sign is (speed limit, stop, pedestrian crossing)

3. Ordinal Logistic Regression
   Three or more categories with ordering.
   Example: Movie rating from 1 to 5

DECISION TREES

Another important method of classification is **decision trees**. Decision trees can be used for regression (to predict a continuous quantity) or classification (to classify a discrete value). Decision trees work a lot like the game Twenty Questions. In Twenty Questions, good players start by asking questions that eliminate a large number of incorrect answers and continue by asking increasingly specific questions until the correct answer is determined. A decision tree is like the set of questions asked in Twenty Questions that efficiently leads a machine learning system to the correct answer.[33]

The first step in using a decision tree is deciding what features should be used. As with any machine learning system, features should be chosen for their predictive quality. For example, if one wishes to predict if a subject will live to age eighty[34], relevant features could include, gender, marital status, socioeconomic status, ethnicity, and BMI. Other features, such as eye color, have little or no predictive value.[35]

Decision trees begin by identifying the one feature that will give the best initial split of the data.[36] This is determined by testing each feature independently. The feature that singly best predicts a correct classification becomes the first **root node**. For example, if gender is the most important variable for predicting if a subject reaches age eighty, the data is split into two groups: males and females. Each of these data groups are a **leaf node**.

The process then repeats again with each leaf node. Each feature is tested for its predictive quality within each new group (leaf node), further splitting the data into smaller groups. So, for example, females may be divided into two groups—those with a BMI of less than 25 and those with a BMI of 25 or greater. This branching continues until specified criteria are met.[37]

Subject	Gender	Status	Salary	Ethnicity	BMI	Live to 80?
1	Male	Single	$40,000	White	24.5	No
2	Female	Married	$60,000	Black	26.0	Yes
...	...	...	...	...	...	...

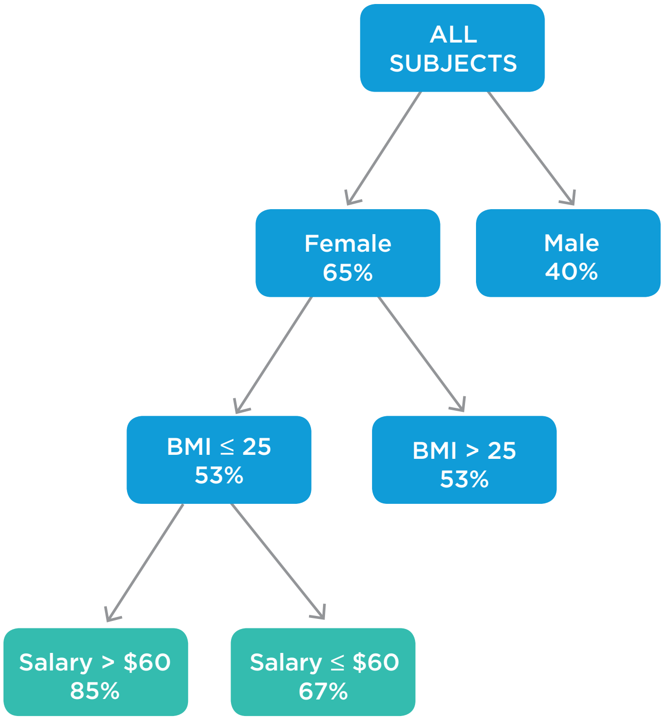


FIgure 8.
Example of a decision tree.

When the decision tree is completed, a model should exist which can efficiently determine which combination of features have the greatest impact on the target variable. For instance, using this example, the user knows that if they focus on females, with a BMI ≤ 25, and a salary > $60,000, there should be a very high (85%) chance they will live past eighty years of age. The decision tree is tested with new training data to determine its accuracy. Corrections can be made by eliminating features which do not have sufficient predictive quality.

Decision trees have many advantages over other forms of machine learning. They work well with large datasets and are simple to understand in concept, although the actual number of variables, and leaf nodes can become extremely complex. It is also easier to determine how a decision tree model arrives at its conclusions, or estimate the likely decision pathway of a particular output.

Fairness and "Bias" Concerns

Supervised learning requires massive amounts of training data. Bias that is present in the training data becomes inherent in the model it produces.

Example 1: Researchers recently used a deep learning technique to train a system to distinguish pictures of huskies from pictures of wolves. When these techniques were applied, each time the AI identified the picture as a wolf, the scientists discovered that it was paying attention to the presence or absence of snow in the image. The systems "learned" that one factor in whether it was a wolf was the presence of snow. Thus, the system relied on a false correlation, deciding that every time there was snow, it was a wolf. In a supervised learning application (with labeled images), this result might be the result of a biased dataset that included few or no pictures of wolves in grass, or dogs in snow.[38]

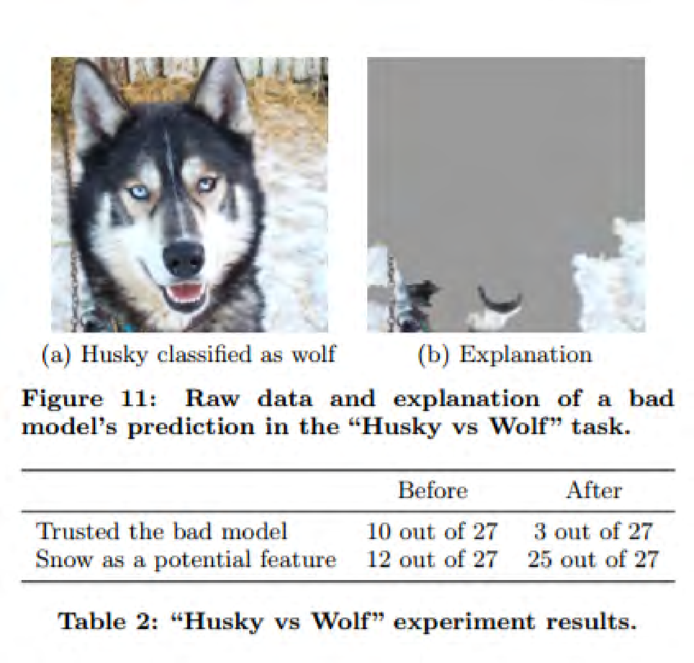

(a) Husky classified as wolf (b) Explanation

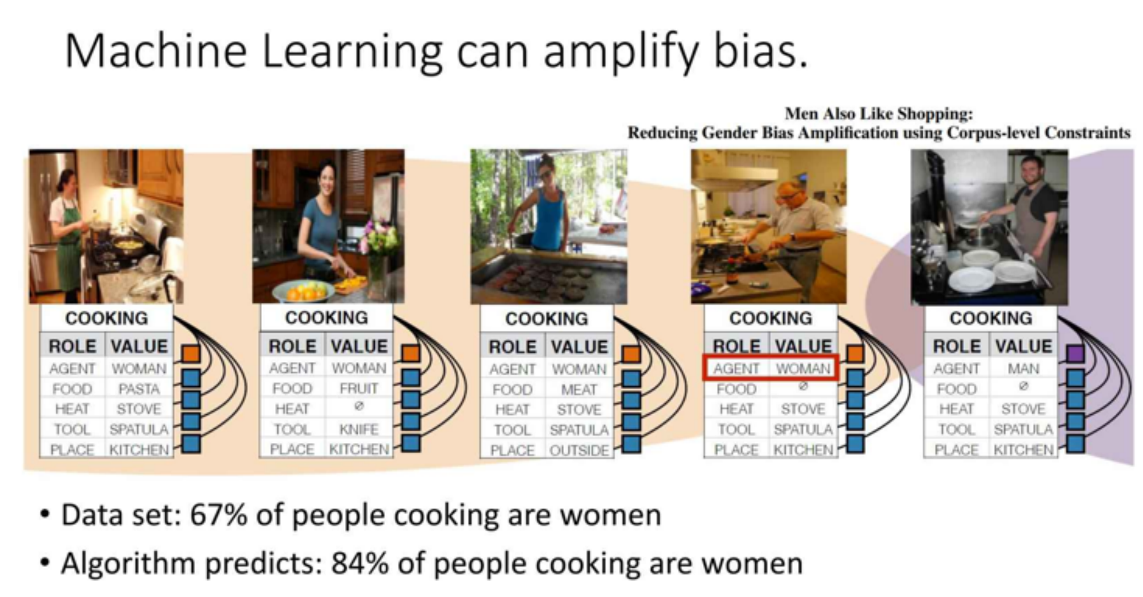
Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27

Table 2: "Husky vs Wolf" experiment results.

(Fairness and bias concerns continued on next page)

- **Example 2:** Researchers found that, largely based on digital images, computers learned to link “male” and “man” with STEM fields, whereas “woman” or “female” were linked with household tasks.[39]



Machine Learning can amplify bias.

Men Also Like Shopping:
Reducing Gender Bias Amplification using Corpus-level Constraints

COOKING	
ROLE	VALUE
AGENT	WOMAN
FOOD	PASTA
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

COOKING	
ROLE	VALUE
AGENT	WOMAN
FOOD	FRUIT
HEAT	∅
TOOL	KNIFE
PLACE	KITCHEN

COOKING	
ROLE	VALUE
AGENT	WOMAN
FOOD	MEAT
HEAT	STOVE
TOOL	SPATULA
PLACE	OUTSIDE

COOKING	
ROLE	VALUE
AGENT	WOMAN
FOOD	∅
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

COOKING	
ROLE	VALUE
AGENT	MAN
FOOD	∅
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

- Data set: 67% of people cooking are women
- Algorithm predicts: 84% of people cooking are women

Bias can occur in two ways (and can occur in both at the same time). First, as in the example here, is bias in the data—either the training data, or in the “real world” data that is supplied to the model once in operation.

Second, bias can occur based on the selection of factors, and the weighting assigned, during the design of the underlying system. Decisions made when defining categories, identifying fields, and establishing relationships impact both the efficiency of the model, and what outputs will be generated. When the outcome is identifying patterns over time, some variations or range of accuracy may be acceptable. However, when the systems are generating recommendations that impact individuals—such as with credit evaluations or sentencing guidelines—any inherent bias in the weighting of various factors or choosing which ones to use has real-world results that must be fair and defensible.

Example 3: ProPublica did a well-publicized project that showed that black defendants were almost twice as likely as white defendants to be rated as “high risk” for reoffending, even after equalizing for prior arrests, age, and gender. The risk score was then used to set bond levels, to determine incarceration decisions, and affect sentence length. A separate validation study found the “risk of recidivism” score was correct in ~68% of both blacks and whites, although whites were much more likely to be labeled low risk. Since race wasn’t a data point, this study also demonstrates the challenge of finding a model that doesn’t create a proxy for race (or other eliminated factor)—such as poverty, joblessness, and social marginalization.[40]

The impact of certain sensitive data such as race or gender may affect outcomes in unforeseen or undesirable ways. While the goal of machine learning programs is frequently to create more objective evaluation or analysis and fairer outcomes, if the training data is inherently biased because of its source (that is, if the human-run process from which it was derived included bias in the collection or correlation of the data included in the data set) then those biases will carry over into the machine learning system. It takes expert understanding and evaluation to prevent or discern this, including programmers who are able to use more algorithms to “audit” or evaluate the testing outcomes and ensure that such distortions do not occur.

UNSUPERVISED LEARNING

While supervised learning models are created from labelled training data which teach them to find the patterns that the user has specified, **unsupervised learning** finds patterns on its own. Since the data that unsupervised learning uses is unlabelled, it can be more difficult to evaluate its accuracy. However, unsupervised learning can provide valuable insights into the underlying structure of a dataset not possible with supervised learning processes.

There are many different approaches to using unsupervised learning. One of the most common approaches is clustering.

CLUSTERING

Clustering is the task of grouping a set of data together in such a way that the data points in the same cluster are more similar to each other than to those in other clusters.41 In other words, clustering attempts to create meaningful distinctions between groups of data without human involvement in creating the groups. Clustering can be used to find groups which have not been explicitly labeled in the data.

Clustering isn't just one algorithm—it can be accomplished with a wide variety of different mathematical approaches. **K-means clustering** is probably the most well-known clustering algorithm. The k-means approach tries to group data into k groups, by creating a centroid for each group, which acts as the heart of each cluster.[42] The centroids are placed randomly at first, but the k-means algorithm quickly adjusts them to better capture clusters of data.

Examples: Netflix uses clustering to make movie recommendations by identifying which new movies are related to movies that the user has previously watched. Amazon uses clustering to provide other products a user may be interested in by determining which products are often bought together.

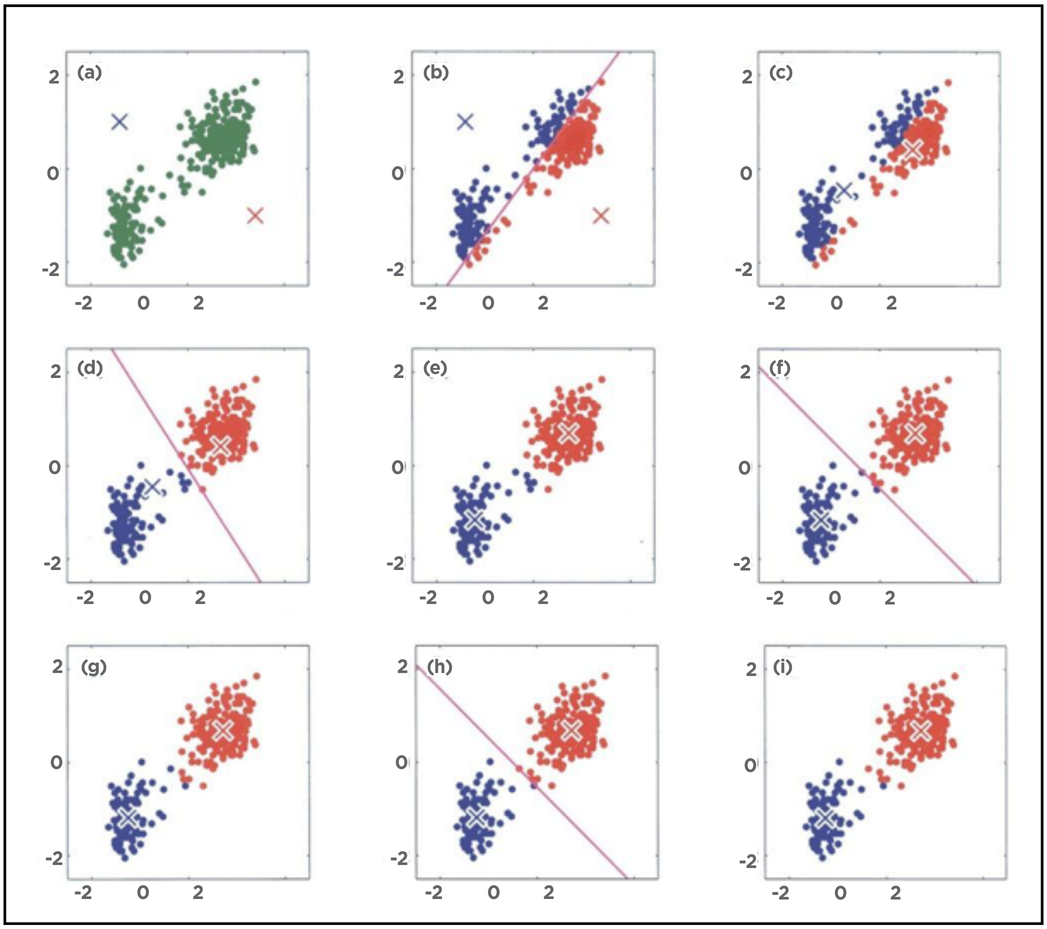


Figure 9.

Illustration of k-means clustering iteratively improving the position of the cluster centroids (X).[43]

Clustering algorithms can inform telephone companies as to where to put new towers by helping discover where users are most concentrated.

Grocery stores may use their "loyalty card" data to define customers in various related or overlapping clusters—"fresh vegetables and seafood," or "frozen dinners and paper products". The machine would generate the overlapping clusters (those who frequently bought both frozen dinners and paper products together); the user (the company representative using the program) would add the labels to manage and track the various groups.

Another example of unsupervised learning is *generative modeling*. This technique has gained traction recently, as a deep learning technique called "generative adversarial networks" (GANs) has been developed. For example, given many photographs of people's faces, a generative model could then generate new artificial samples: real-looking but artificial images of people's faces.[44] This is achieved through an iterative process between two networks ("adversaries"). The generative network produces "fake" images, which the adversarial network, having been trained on real images, tries to distinguish as fakes. Every time it rejects a fake, the generative model gets better, and the cycle continues until eventually the generated images are almost indistinguishable from real ones.

Explainability and Transparency

While not interchangeable, these two concepts are related. Transparency is generally the ability to interpret the functioning of machine learning models, provide explanations for specific algorithmic outputs, and describe or provide frameworks for conducting ethical and legal reviews and audits. It might include providing data subjects with process details, informing them when a decision was reached via an automated system, and providing an understanding of the safeguards and rights available to them.[45]

Explainability describes the privacy risk discussions that need to occur when decisions are made in the development and design stages of a ML model. When decisions made by the model will affect individuals or members of a group, and are reached based solely on the model output, without human review or intervention, there is the potential for problems.

The reality is that many cutting edge machine learning models include multi-layer neural networks (described below) which follow as an internal process under which a particular outcome may not be able to be understood in a mathematically detailed, replicable way even by the data scientists and designers. But the multi-layer feature in fact may make the model "better" in terms or relevance or applicability. This is recognized as the accuracy vs. interpretability tradeoff—sometimes called the "black box" problem.[46]

There are a variety of proposed solutions for this challenge.

- Programs which can follow the steps and describe more fully what the model is doing.[47]
- Risk management structures that use other ways to evaluate the model, the data, and the output to ensure accuracy and reliability.[48]

REINFORCEMENT LEARNING

A third type of machine learning is called **reinforcement learning**. Reinforcement learning is when a system learns by trial-and-error through "reward and punishment" with no sets of existing data provided to it. This contrasts with supervised and unsupervised learning because neither labeled data, nor examples of desired outcomes are provided. Instead, the system is given a "reward" signal for when it accomplishes what the designer wants, or a step that advances the process toward the outcome the designer described. When the system does something wrong (fails to efficiently advance toward the desired outcome), it is simply not rewarded. Thus, there are no datasets of "training data" for a model built on this type of program. The "data" is generated because the system is run many, many times, to allow it to try a large variety of different strategies and determine all the multiple ways to reach the "reward" outcome. If a particular strategy is efficient in accomplishing the goal, the system will be encouraged to explore it further. If it isn't, the strategy is either revised and improved, or abandoned. Proceeding in this manner, the system incrementally improves itself through countless iterative cycles.

In 2013, researchers used reinforcement learning to train a system to play Atari games.[50] To train the system, the researchers could have used supervised learning, feeding it thousands of recorded games played by skilled humans. However, this would have required an enormous dataset of games to achieve real proficiency and would have produced results that mirrored the best human achievement levels. Reinforcement learning, on the other hand, rewarded the system for performing well in terms of winning, regardless of whether it mirrored human decision, and allowed it to develop its own strategies. Using this method, the system was able to outperform human players, developed strategies and techniques that were distinct from those commonly employed by human players, and required no training data.

Though reinforcement algorithms may take much longer to reach a "trained" state, they have a wider variety of applications, and potentially, become better at learning, or more efficient at reaching the desired outcomes. Because they do not need labelled data, the opportunity for "teacher bias" is eliminated. And because they learn by doing, rather than by ingesting unsupervised training data, they are less likely to reflect the bias inherent in historical patterns or or human systems.[51]

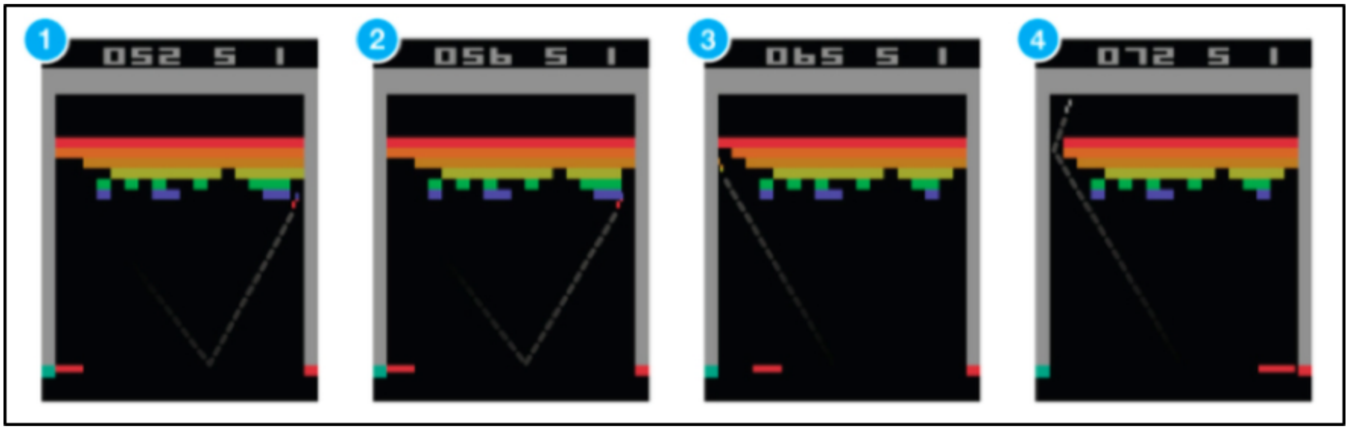


FIgure 10.

Through reinforcement learning, a machine learning system was trained to outperform humans on the Atari game, Breakout. https://youtu.be/TmPfTpjtdgg

NEURAL NETWORKS

In a human brain, a neural network consists of a large number of cells, neurons, that receive and transmit signals to each other. The neurons are very simple processors of information, consisting of a cell body and "wires" (dendrites) that connect the neurons to each other across a gap (synapse). Most of the time, they do nothing but sit still and wait to react to incoming signals. The human brain processes information with millions of these interconnected neurons. When a neuron fires (sends a signal) in the brain, it causes other neurons to fire which collectively contribute to a single output.Each neuron operates in parallel with other neurons and each neuron is influenced by other connected neurons.[52]

Neural networks in computers are a method of machine learning which is designed to loosely emulate the way the human brain works. Neural networks use multiple layers of **perceptrons**, which are analogous to biological neurons. Just like neurons, perceptrons are influenced by the "on" or "off" status of the other perceptrons to which they are connected. They are capable of modeling and processing nonlinear relationships between inputs and outputs, in parallel.[53]

In a traditional computer, information is processed in a central processor (CPU) which can only do one process at a time (albeit extremely fast). The CPU retrieves data to be processed from the computer's memory, and stores the result. Thus, data storage and processing are handled by two separate components of the computer: the memory and the CPU. In neural networks, the system consists of a large number of perceptrons, each of which can process information on its own so that instead of having a CPU process each piece of information one after the other, the perceptrons process vast amounts of information simultaneously. Data can also be stored short term in the perceptrons themselves (as the switch is "on" or "off"), and longer term in the connections between them.[54]

In the simplest model, neural networks are comprised of three layers: the input layer, a single hidden layer, and an output layer. Each layer can contain multiple neurons. Models can become increasingly complex, capable of abstraction and problem solving, by increasing the number of hidden layers, the number of neurons in each layer, and the number of connections between neurons.[55]

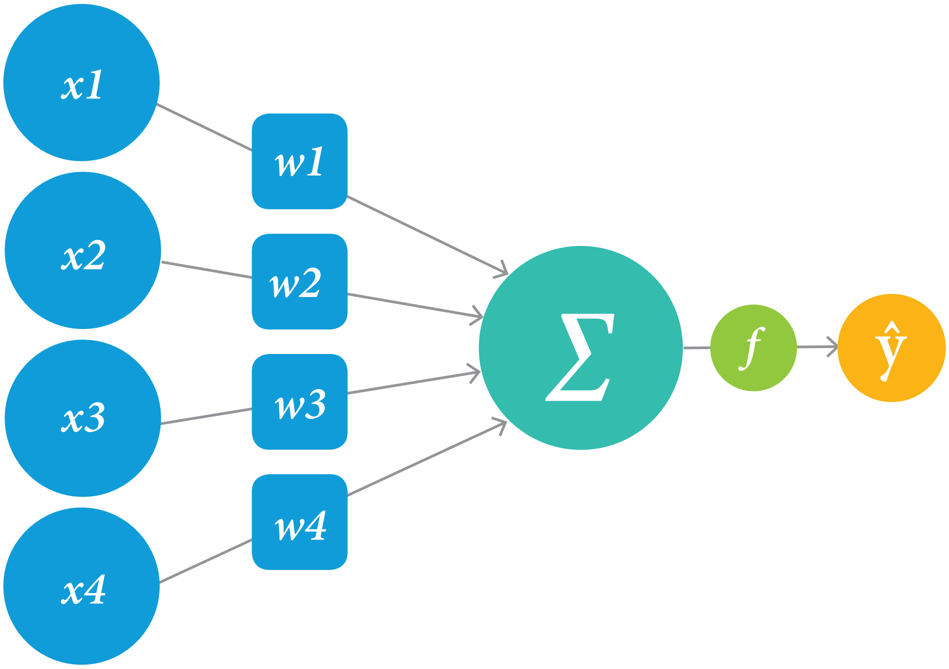


Figure 11.

*Diagram of a perceptron (**x**=input, **w**=weight **Σ**=sum of all weighted inputs and bias, **f**=unit step function, **ŷ**=outcome).*

Neural Network Process:

- **Data input** – Data input—the sets of information submitted into the neural network for processing. In a simplistic example, consider your grocery cart—into which you load varying amounts of items—fresh produce might be measured by weight, along with processed food in containers at a set size and price, as well as some non-edible goods.
- **Weights** - this is the relative predictive value assigned to the different data inputs. If we keep the analogy of the grocery basket, the weighted value would be the price. A high value item will be a larger percent of the total bill than a smaller one. Likewise, something of lower value that appears multiple times will make up a higher percentage of the final cost. Perhaps we want to ensure that at least a set percentage of the cost (or set dollar value) is devoted to food as opposed to cleaning supplies or paper goods. This might be considered the "intercept" value.[56]
- **Output** - is determined by the linear progression and the activation function, either to create a prediction or a decision. The car identifies the traffic sign as a stop sign, and so the car stops.
- **Adjusting weights/threshold values for specific results** - Weights may be adjusted to modify the outputs if they are determined to be sub-optimal, either in accuracy, or to achieve desired levels, or due to reflecting bias, as judged by the programmer and user. A neural network maybe have millions or billions of weights. Only enormous amounts of computing power have made it possible to optimize or adjust data at this scope.[57]

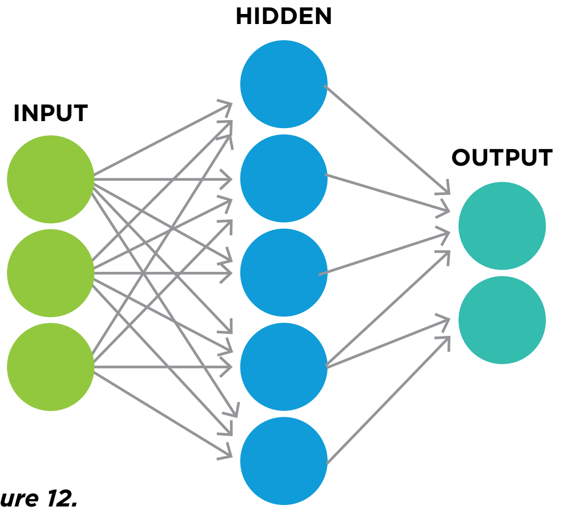


Figure 12.
Neural network process

DEEP LEARNING

Deep learning is a subset of neural networks. It describes certain types of neural networks that operate on very "raw" input data, processing the data through many layers of nonlinear transformations to achieve a target output. It can be supervised or unsupervised. Deep learning techniques include several "layers" of perceptrons connected in a network so that the input to the system is passed through each one of them in turn, and in some instances, passed through certain layers multiple times. Some common applications of Deep Learning include:

- Natural Language Processing (voice-enabled software)
- Image recognition (applicable in social "bulletin boards," online retail, photo-organizing software, and visual description programs for visually impaired users)
- Machine Translation (real-time translation of signs, text, or speech)[59]

As described above, **hidden Layers** are neuron nodes stacked in between the input layer and the output layer, allowing neural networks to complete more complicated processes, or "learn" harder tasks. All neurons on the same layer get inputs from neurons on the previous layer, and feed their output to the next layer. The iteration of data through these layers as the machine learning model proceeds and makes its own edits and adjustments is not always visible or replicable, even to the programmer. This is where design, testing, and audit protections must be applied to ensure accuracy, and support trust and reliance on a particular process.[60] Research is ongoing in this area to provide sufficient analysis and confidence measures for these models.

Especially when dealing with even more complex systems with multiple hidden layers and large numbers of factors, it becomes harder to judge bias in outputs by linking back to inputs to determine what the model is actually learning. Much current research is designed to provide model evaluations by linking and correlating *outputs* or *predictions* back to the *input data*. While this is an understandable and reliable alternative for linear models, it remains challenging for interpreting the results from deep learning networks. The two main approaches that are currently applied are either gradient-based or attention-based.[61]

In gradient-based methods, “the gradients of the target concept calculated in a backward pass are used to produce a map that highlights the important regions in the input for predicting the target concept. This is typically applied in the context of computer vision.”[62] *Backpropagation* is a method used in artificial neural networks to calculate a gradient that is needed in the calculation of the weights to be used in the network. Related to gradient descent and neural networks, this procedure is used to repeatedly adjust the weights so as to minimize the difference between actual output and desired output.

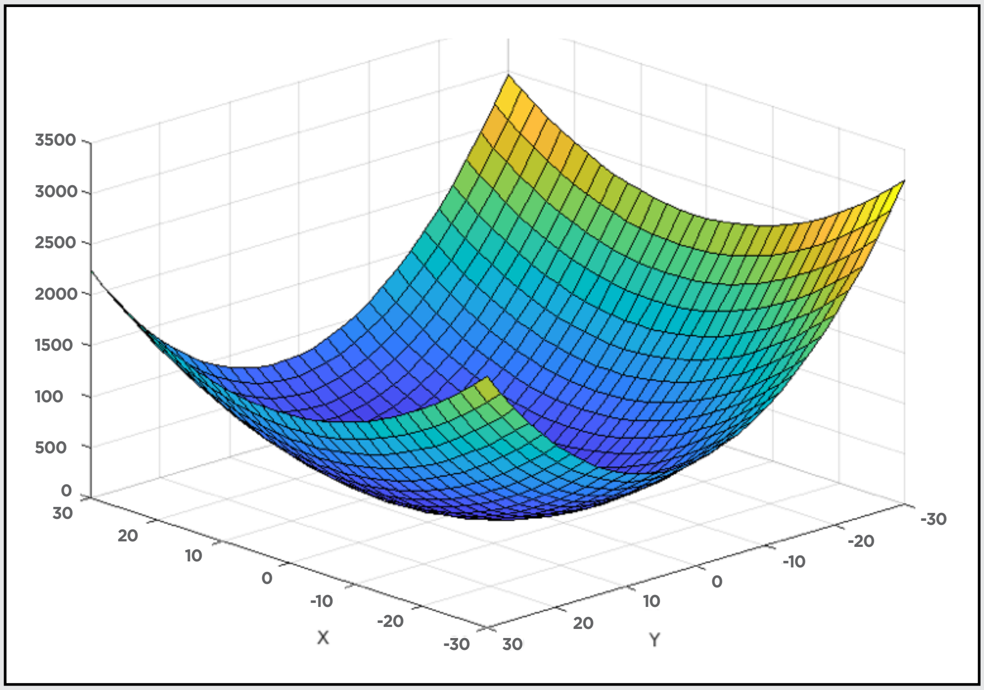
Attention-based methods “are typically used with sequential data (e.g. text data). In addition to the normal weights of the network, attention weights are trained that act as ‘input gates’.”[64] These attention weights control how much each of the different elements or features influences the final output. Besides providing a viewpoint to increase interpretability of model outputs, attention based-methods also generally lead to more correct results, more quickly.

Gradient Descent

Gradient descent is one of the most commonly used algorithms in machine learning. The goal of gradient descent is to minimize the cost function of a model by iteratively adjusting parameter values. In a linear regression model, the cost function measures how “good” a line fits to the data. In other words, gradient descent uses calculus to improve a regression model’s predictive power.

To illustrate how gradient descent works, an analogy is often used of a man trying to get down a mountain in a heavy fog. Since the path down the mountain is not visible, he can only examine the steepness of the terrain immediately adjacent to himself. Using the method of gradient descent, the man proceeds downhill in the direction with the steepest descent. Continuing in this manner, the man will eventually reach the bottom of the mountain.

Sometimes, however, the steepness of the mountain is not easily determined. In some places, the man must spend a considerable amount of time measuring which way is the steepest descent. Thus, the man must decide how often he will take measurements before traveling downhill. If he stops every couple of feet to measure, he won’t get to the bottom in a reasonable amount of time. On the other hand, moving in one direction without adequate measurement, could mean the man might move wrong way. Gradient descent provides the answer as to how often he should take measurements in the most efficient way possible.



A visualization of gradient descent.

LIME

There are a number of exploratory processes being developed to provide understanding or interpretability of extremely complex machine learning models. One of those furthest along is "Local Interpretable Model-agnostic Explanations" (or LIME), a general framework that aims to make the predictions of 'any' machine learning model more interpretable.[65]

In order to remain model-independent, LIME works by modifying the input to the model locally. So instead of trying to understand the entire model at the same time, a specific input instance is modified and the impact on the predictions are monitored. In the context of text classification, this means that some of the words are e.g. replaced, to determine which elements of the input impact the predictions.

First proposed in 2016 by computer scientists at the University of Washington, LIME purports to explain the predictions of any classifier.[66]

Privacy Framing for AI Issues – "FAT"

The discussion of AI privacy implications has largely been framed around the concepts of Fairness, Accountability, and Transparency (frequently shortened to "FAT").

- **Fairness:** Ensure that algorithmic decisions do not create discriminatory or unjust impacts when comparing across different demographics or affected communities and individuals.
- **Accountability:** Algorithmically informed decisions have the potential for significant social impact, and must be designed and implemented in publicly accountable ways, such as an obligation to report, explain and justify specific decisions as well as mitigate negative impacts and potential harms.
- **Transparency:** ensuring that the personal data collected, the model's function and process, and the role of the model's output in final decision-making are explainable to the user or individual impacted by the decision.

Taken together, these concepts represent some of the unique aspects of machine learning models that raise the most concern.

CONCLUSION

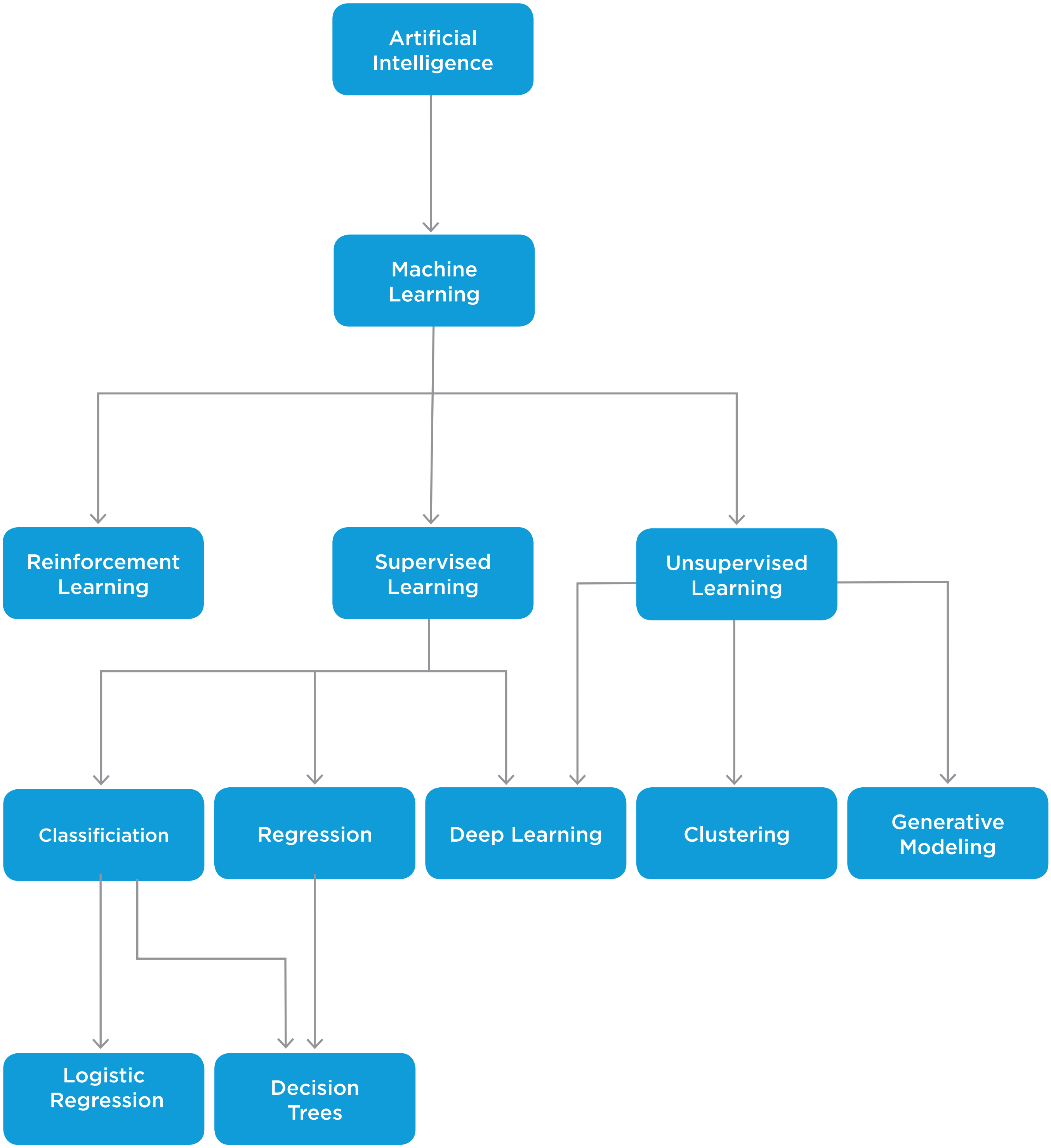
There is growing recognition that using AI and machine learning models raises novel privacy challenges. For policymakers to ensure non-discrimination, due process, and defensibility in decision-making, they must first understand the technology underlying these new features, products, and services. While the benefits are great, the potentially discriminatory impact of machine learning necessitates careful oversight and further technical research into the dangers of encoded bias, or undue opacity in automated decisions. As we look to the future of AI and Machine Learning, it is important to remember that while these systems and models will be invaluable in enabling us to evaluate and benefit from the unfathomable amount of data available, they do not yet represent "intelligence" in any way that correlates to humans. It is incumbent on the people involved in designing, managing, implementing, and regulating these systems to retain accountability—assessing the balance of benefits and risks, and seeking to ensure the best possible outcomes for all.

GLOSSARY

Algorithm	A process or set of rules to be followed in calculations or other problem-solving operations, especially by a computer.
Artificial intelligence	The capability of computer processing systems to perform functions otherwise defined as those requiring human intelligence to accomplish.
Backpropagation	The primary algorithm for performing gradient descent on neural networks.
Bias	The tendency of a process to over- or under-estimate the value (weight, influence) of one or more factors or parameters; may apply to the initial data collection or the processing applied to it
Centroid	The center of a cluster as determined by a k-means or k-median algorithm.
Classification	A type of machine learning model for distinguishing among two or more discrete classes.
Continuous variable	A variable that has an infinite number of possible values.
Decision tree	A decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility.
Deep learning	A subfield of machine learning that uses a collection or sequence of algorithms to model higher-level abstractions in data. Based on neural network models with more than one hidden layer.
Discrete variable	A variable that has an finite number of possible values.
Explainability	The ability to describe or achieve human understanding of why a particular ML model arrived at a specific output or prediction.
Features	Individual measurable properties or characteristics; variables that can help build an accurate predictive model. Attributes of data used to train a machine learning system.
General artificial intelligence	The intelligence of a machine that could successfully perform any intellectual task that a human being can.
Generalization	For ML, refers to a model's ability to make correct predictions on new, previously unseen data as opposed to the data used to train the model.
Gradient Descent	An optimization procedure for ML algorithms, used to train deep neural networks, that finds the optimal “weights” to reduce prediction error
Hidden layers	A processing layer in a neural network between the input layer and the output layer. There may be one or more hidden layers in a given network.
K-means Clustering	A popular clustering algorithm that groups examples in unsupervised learning.
Logistic regression	A model that generates a probability for each possible discrete label value in classification problems.
Machine learning	A field of computer science which designs algorithmic based systems to learn from presented data rather than being explicitly programmed. A program or system that trains a predictive model from input data.
Minimization	A privacy principle for data collection that requires the practice of limiting the collection of personal information to only that which is relevant and necessary to accomplish the specified purpose.
Model	A mathematical representation of the real world. The representation of what an ML system has learned from the training data, which is then used to make predictions.
Narrow artificial intelligence	An application of artificial intelligence which replicates (or surpasses) human intelligence for a dedicated purpose or application.
Neural network	A variation of a machine learning model that, taking inspiration from the human brain, is composed of multiple processing layers.

Overfitting	Creating a model that matches the training data so closely that the model fails to make correct predictions on new data.
Perceptron	A component of neural networks designed to mimic the neurons in a human brain.
Regression	A type of model that estimates relationships among variables, focusing on the relationship between a dependent variable and one or more independent variables; outputs are continuous values.
Reinforcement learning	A type of dynamic programming that trains algorithms using a system of reward and punishment. Uses no existing data for training.”
Singularity	The moment at which machines become “smart” enough to take over their own development, and driving exponential or “runaway” advancement.
Supervised learning	Training a model from input data and its corresponding, previously assigned labels.
Transparency	The ability to explain or understand the inputs, outputs, functions, and data collection and use of a particular process, including the rights and relationship to the data subject and user.
Unsupervised learning	Training a model to find patterns using an unlabeled data set.

TAXONOMY OF TERMS



ENDNOTES

1 Łukasz Kaiser & Aidan N. Gomez, *MultiModel: Multi-Task Machine Learning Across Domains*, Google AI Blog, June 21, 2017, https://ai.googleblog.com/2017/06/multimodel-multi-task-machine-learning.html. There are examples of "MultiModel" learning that draws from multiple networks simultaneously and may solve problems across multiple domains. At present these are limited to sensory inputs from multiple sources, and are a "first step towards the convergence of vision, audio and language understanding into a single network."

2 Seth Baum, *A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy*, Global Catastrophic Risk Institute Working Paper 17-1 (Nov. 16, 2017). As of 2017, over forty organizations worldwide are doing active research on general AI.

3 See Vincent C. Müller & Nick Bostrom, Fut*ure Progress in Artificial Intelligence: A Survey of Expert Opinion, in Fundamental Issues of Artificial Intelligence 553–571* (Springer ed., 2016); Ray Kurzweil, The Singularity Is Near: When Humans Transcend Biology 10–11 (2005).

4 Luke Muehlhauser, *What is AGI?*, Machine Intelligence Research Institute, Aug. 11, 2013, https://intelligence.org/2013/08/11/what-is-agi/.

5 *Id*.

6 Deep Mind Research, *AlphaGo*, https://deepmind.com/research/alphago.

7 Peter Voss, *From Narrow to General AI*, Medium, October 3, 2017, https://medium.com/intuitionmachine/from-narrow-to-general-ai-e21b568155b9.

8 Arthur L. Samuel, *Some Studies in Machine Learning Using the Game of Checkers I, in Computer Games I 335–365* (Springer ed., 1988).

9 Jason Mayes, *Jason's Machine Learning 101*, Google (Nov. 2017), https://goo.gl/5Wd2vy.

10 Some types of ML systems do not required training data. These do still have a "learning" or training stage, however, and will be discussed later.

11 David Kelnar, *The Fourth Industrial Revolution: A Primer on Artificial Intelligence (AI),* Medium (Dec. 2, 2016), https://medium.com/mmc-writes/the-fourth-industrial-revolution-a-primer-on-artificial-intelligence-ai-ff5e7fffcae1.

12 Mayes, *supra* note 9, at 10.

13 Scott Berinato, *There's No Such Thing as Anonymous Data*, Harvard Business Review (February 9, 2015) https://hbr.org/2015/02/theres-no-such-thing-as-anonymous-data.

14 DataFloq, *The Re-Identification of Anonymous People with Big Data* (February 10, 2011) https://datafloq.com/read/re-identifying-anonymous-people-with-big-data/228.

15 Applied AI, *Synthetic Data: An Introduction & 10 Tools*, (June 2018 update), https://blog.appliedai.com/synthetic-data/

16 Datatilsynet, *Artificial Intelligence and Privacy* (Jan. 2018), https://www.datatilsynet.no/globalassets/global/english/ai-and-privacy.pdf

17 Bernard Marr, *Why Data Minimization Is An Important Concept In The Age of Big Data*, Data Science Central (Apr. 5, 2016), https://www.datasciencecentral.com/profiles/blogs/why-data-minimization-is-an-important-concept-in-the-age-of-big.

18 Matthew Green, *A Very Casual Introduction to Fully Homomorphic Encryption* (January 2, 2012), https://blog.cryptographyengineering.com/2012/01/02/very-casual-introduction-to-fully/.

19 Mayes, *supra* note 9, at 18. For purposes of this example, all the apples are red. In reality, apples might be many colors, and so color alone would not be definitive for the model to reach a conclusion.

20 Mayes, *supra* note 9, at 18.

21 Gautam Narula, *Everyday Examples of Artificial Intelligence and Machine Learning*, TechEmergence (Sept. 16, 2018), https://www.techemergence.com/everyday-examples-of-ai/.

22 This example would actually be manageable with traditional statistical regression formulas. It is used here for ease of understanding, but an ML-based model would leverage regression beyond what could be handled by standard statistical tools.

23 See University of Helsinki, *Elements of AI*, https://course.elementsofai.com/.

24 Ethem Alpaydin, *Machine Learning 39-42* (MIT Press Essential Knowledge Series ed., 2016).

25 University of Helsinki, *Types of Machine Learning*, https://course.elementsofai.com/4/1.

26 Unsupervised classification is considered clustering. See *infra*.

27 Jason Brownlee, Diffe*rence Between Classification and Regression in Machine Learning*, Machine Learning Mastery (Dec. 11, 2017), https://machinelearningmastery.com/classification-versus-regression-in-machine-learning/.

28 *Id.*

29 *Id.*

30 Don't be confused by the term "logistic regression." This type of learning falls under classification.

31 Kirill Fuchs, *Machine Learning: Classification Models*, Fuzz (Mar. 28, 2017), https://medium.com/fuzz/machine-learning-classification-models-3040f71e2529.

32 As described above, this is dependent on the specific training process: some models will "force" an output, regardless of low probability, into one of the two available choices, if those are the model's only allowable outcomes. Others will have the ability as described here to have a third "doesn't sufficiently match" choice.

33 Peter Jeffcock, *Decision Trees in Machine Learning*, *Simplified*, Oracle Big Data (Apr. 3, 2017), https://medium.com/oracledevs/decision-trees-in-machine-learning-simplified-abc916c8b22b.

34 Note this is a classification model with two classes: 1) subjects who live to age eighty and, 2) subjects who did not. It could be reworked to become a regression model, by predicting a continuous value, such as life expectancy. In that case, the model would output a predicted longevity instead of putting each subject into a discrete class.

ENDNOTES (cont'd)

35 The relevant features should never be assumed. In many cases, the value of ML models is that they are able to find correlations and predictive relationships between features not previously known to be associated. Prior studies or other objective analysis should be the preferred basis for intentionally deleting otherwise possibly relevant data fields.

36 Jeffcock, *supra* note 33.

37 Usually when a minimum number of records from the training data is not sorted into a node.

38 See *Dogs, Wolves, Data Science, and Why Machines Must Learn Like Humans Do*, Hackernoon (Aug. 28, 2017), https://hackernoon.com/dogs-wolves-data-science-and-why-machines-must-learn-like-humans-do-41c43bc7f982; Marco Tulio Ribeiro, Sameer Singh & Carlos Guestrin, *"Why Should I Trust You?": Explaining the Predictions of Any Classifier* (Cornell University Library ed., 2016), https://arxiv.org/abs/1602.04938.

39 Tom Simonite, *Machines Taught By Photos Learn a Sexist View of Women*, Wired (Aug. 21, 2017), https://www.wired.com/story/machines-taught-by-photos-learn-a-sexist-view-of-women/

40 Julia Angwin, Jeff Larson, Surya Mattu & Lauren Kirchner, *Machine Bias*, ProPublica (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

41 Ethem Alpaydin, *Machine Learning 112-117*, (MIT Press Essential Knowledge Series ed., 2016).

42 Vishal Maini, *Machine Learning for Humans*, Medium (Aug. 19, 2017), https://medium.com/machine-learning-for-humans/why-machine-learning-matters-6164faf1df12.

43 Dendroid, *K-means Clustering Algorithm Explained* (May, 2011), http://dendroid.sk/2011/05/09/k-means-clustering/.

44 See University of Helsinki, *supra*, note 23.

45 Datatilsynet, *supra* note 12.

46 Stuart Shirrell, *The Privacy Pro's Guide to Explainability in Machine Learning*, International Association of Privacy Professionals (Apr. 13, 2018), https://iapp.org/news/a/the-privacy-pros-guide-to-explainability-in-machine-learning/.

47 Adele Peters, *This Tool Lets You See—and Correct—the Bias In An Algorithm*, Fast Company (June 12, 2018), https://www.fastcompany.com/40583554/this-tool-lets-you-see-and-correct-the-bias-in-an-algorithm; Kush Varshney, *Introducing AI Fairness 360*, IBM (Sept. 19, 2018), https://www.ibm.com/blogs/research/2018/09/ai-fairness-360/; Miro Dudík, et al., Machine Learning for Fair Decisions, Microsoft Research Blog (July 17, 2018), https://www.microsoft.com/en-us/research/blog/machine-learning-for-fair-decisions/; Matt Wood, *Thoughts On Machine Learning Accuracy*, Amazon Web Services Blog (July 27, 2018), https://aws.amazon.com/blogs/aws/thoughts-on-machine-learning-accuracy/; Jerome Pesenti, *AI at F8 2018: Open Frameworks and Responsible Development*, Facebook Code (May 2, 2018), https://code.fb.com/ml-applications/ai-at-f8-2018-open-frameworks-and-responsible-development/.

48 Andrew Burt, Brenda Leong, Stuart Shirrell & Xiangnong Wang, *Beyond Explainability: A Practical Guide to Managing Risk in Machine Learning Models 4–5* (Future of Privacy Forum & Immuta eds., 2018).

49 Mayes, *supra* note 9, at 27.

50 Volodymyr Mnih, et al., *Playing Atari with Deep Reinforcement Learning* (Cornell University Library ed., 2013), https://arxiv.org/pdf/1312.5602.pdf.

51 Alpaydin, *supra* note 24, at 134.

52 Alex Castrounis, *Artificial Intelligence, Deep Learning, and Neural Networks, Explained*, KDnuggets (Oct. 19, 2016), https://www.kdnuggets.com/2016/10/artificial-intelligence-deep-learning-neural-networks-explained.html.

53 *Id.*

54 *Id.*

55 *Id.*

56 This example describes the weighting process for a linear regression, which is the first step in a neural network. But obviously a network isn't required if this were the whole process. So a neural network activates functions beyond the basic regression model. Likewise, weights may not be known, but have to be learned from using the model with the test data.

57 Alex Castrounis, *Artificial Intelligence, Deep Learning, and Neural Networks, Explained*, KDnuggets (Oct. 19, 2016), https://www.kdnuggets.com/2016/10/artificial-intelligence-deep-learning-neural-networks-explained.html.

58 Ethem Alpaydin, Machine Learning 17-20, 104-109 (MIT Press Essential Knowledge Series ed., 2016).

59 Jason Mayes, *Jason's Machine Learning 101*, Google (Nov. 2017), https://goo.gl/5Wd2vy.

60 Burt et al., *supra* note 48.

61 Lars Hulstaert, *Interpreting Machine Learning Models, Towards Data Science* (Feb. 20, 2018), https://towardsdatascience.com/interpretability-in-machine-learning-70c30694a05f.

62 *Id.*

63 Nahua Kang, *Introducing Deep Learning and Neural Networks—Deep Learning for Rookies (1), Towards Data Science* (Jun 18, 2017), https://towardsdatascience.com/introducing-deep-learning-and-neural-networks-deep-learning-for-rookies-1-bd68f9cf5883.

64 Hulstaert, supra note 61

65 Ribeiro et al., *supra* note 38.

66 Ribeiro et al., *supra* note 38.

About FPF: Future of Privacy Forum is a nonprofit organization that serves as a catalyst for privacy leadership and scholarship, advancing principled data practices in support of emerging technologies. FPF brings together industry, academics, consumer advocates, and other thought leaders to explore the challenges posed by technological innovation and develop privacy protections, ethical norms, and workable business practices.

1400 Eye Street NW | Suite 450 | Washington, DC 20005 | FPF.org

FUTURE OF PRIVACY FORUM