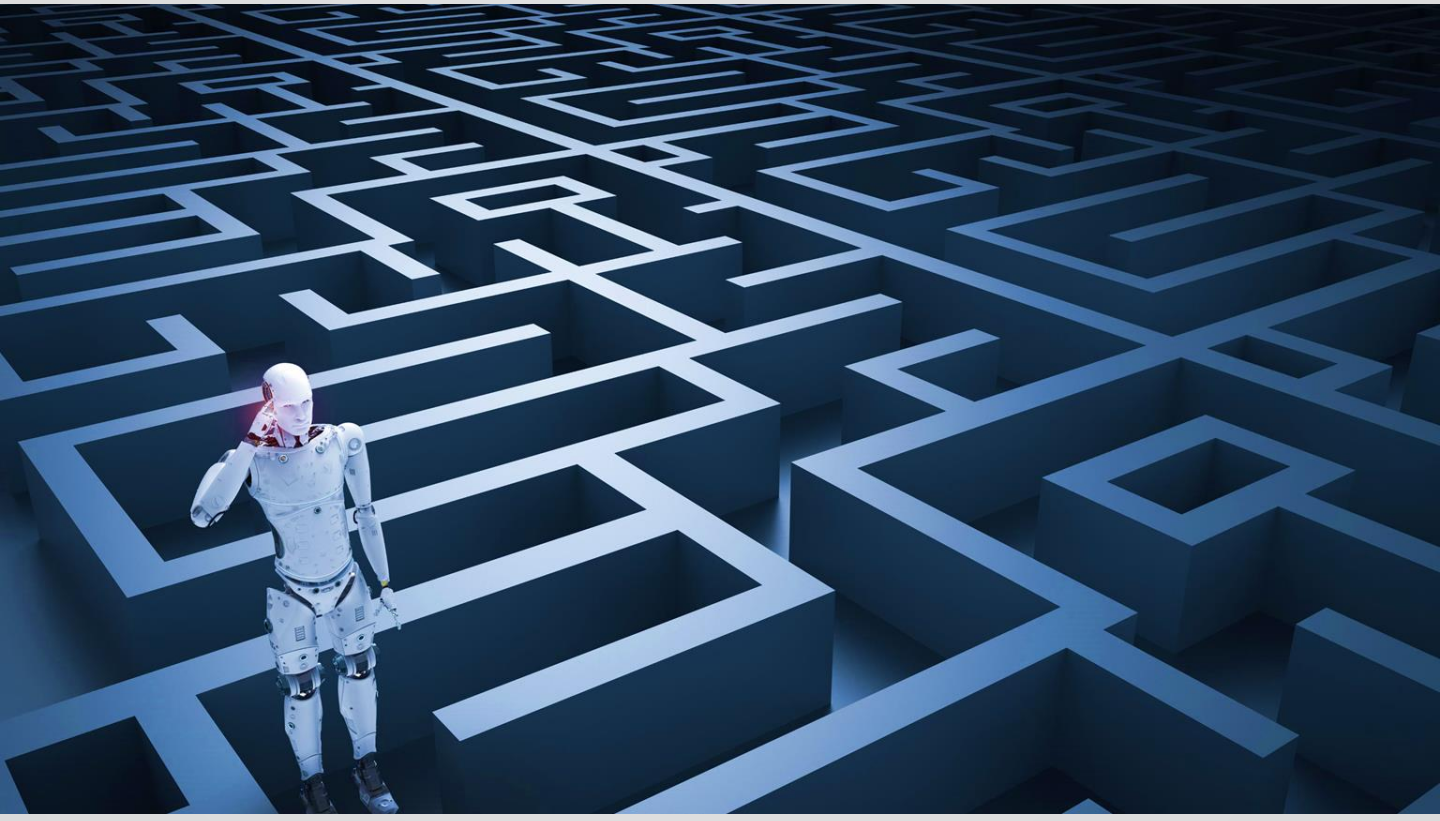
DAU Course Guidebook

DAU SWE0057
What is Machine Learning (ML)?



Version 1.1

Note to Students:

This DAU course guidebook acts is an adjunct educational aid for students participating in online or video-based study. Its main function is to assist with the extraction of images from course videos, aiding students who find visual references helpful during note-taking.

Students in need of fully accessible resources are encouraged to refer to the formal course and accompanying 508-compliant transcript.

Important: This guidebook is not a substitute for the formal course content, and it does not include all information presented in the course or that might be tested during exams or other evaluations.

For questions or feedback regarding this course or artificial intelligence, please reach out to DAU's Software Engineering Team at agile@dau.edu.

DAU

DAU SWE0057
What is Machine Learning (ML)?

Table of Contents*

*The purpose of this supplemental course document is to provide convenient access to important visuals shown in course videos and to facilitate notetaking. It does not contain all content taught within the course.

Machine Learning Definition

Artificial Intelligence (AI) is a broad field of computer science, defined at the highest level as the ability of machines to perform tasks that normally require human intelligence[1].

Machine Learning (ML) a sub-domain of AI, is a set of computing techniques that identify patterns in large data sets – enabling classification, prediction, and improvement over time with exposure to new data. In short, ML is the subset of AI that learns patterns from large data sets.

Artificial Intelligence (AI): [a broad field of Computer Science]

The ability of machines to perform tasks that normally require human intelligence.

Machine Learning (ML): [a sub-domain of Artificial Intelligence]

Computing techniques that identify patterns in large data sets – enabling classification, prediction, and improvement over time with exposure to new data.

1. **AI ≠ ML.** ML is the subset of AI that learns patterns from large datasets.
2. **ML is a "how" rather than a "what."** ML is a multi-faceted problem-solving approach – a potential solution to a requirement, not the requirement itself.
3. **ML is not actually human-like.** ML developers are inspired by cognitive science to approximate human-like learning and analysis capabilities. ML is advanced software, running on advanced hardware, that outputs statistical models as digital 1s and 0s.

Figure 1: Machine Learning definition with clarification for Acquisition.

Lack of a universally accepted AI definition and guidance leads to ambiguity in ML research and information. Three considerations for acquisition professionals:

- AI ≠ ML. ML is a subset of AI that learns patterns from vast datasets.
- ML is a potential solution to a requirement, and not the requirement itself.
- ML is advanced software and hardware but is not technically close to equaling human intelligence.

[1]For an in-depth review of AI definitions in play for DoD, refer to the DAU Course ***What is AI?***

01 ML Definitions & Key Concepts (cont.)

How Machine Learning *Learns*

Machine Learning aims to emulate human learning by leveraging data. To gain insights into ML's objectives, it is beneficial to examine three fundamental steps in human learning processes.

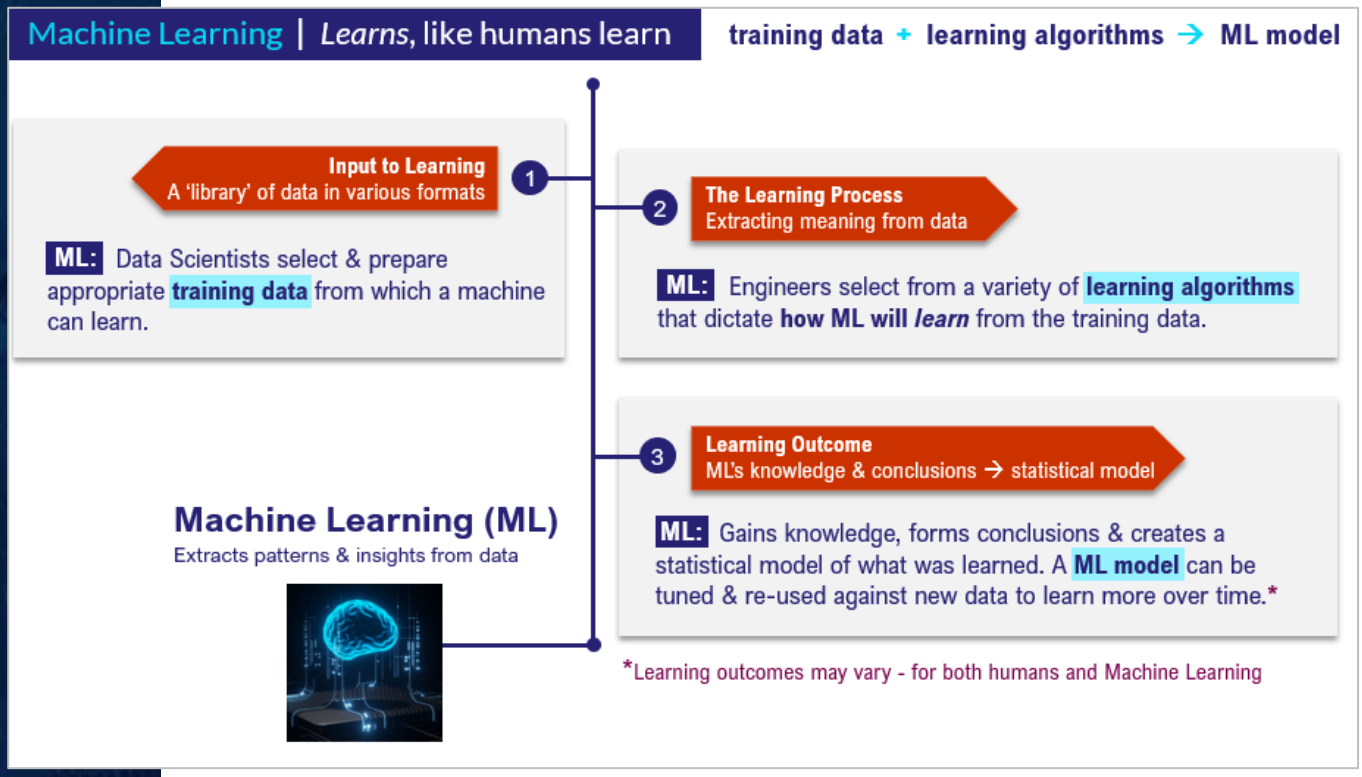


Figure 3: Three basic steps to both human learning and machine learning

(3) fundamental steps to <u>human learning</u>:

- **Step 1 - The Input to Learning:** Scholars do significant up-front work to prepare a library of knowledge and learning materials from which students can learn. The more time spent up-front getting this step right, the better a student's chance of successful learning.
- **Step 2 - The Learning Process:** How students extract meaning from the learning materials. Teachers use a variety of instructional methods and students can learn using varied learning styles.
- **Step 3 - The Learning Outcome:** When a student has acquired knowledge and gains proficiency, they are able to draw conclusions and form *mental models* representing what was learning and how to apply it to future situations.

01 ML Definitions & Key Concepts (cont.)

How Machine Learning *Learns*

Machine Learning endeavors to replicate the three stages of human learning in the following way:

(3) fundamental steps to machine learning:

- **ML Step 1 - The Input to Learning:** Data Scientists select and prepare appropriate **training data** from which a machine can learn. Significant time is spent cleansing, formatting, labeling, and pre-testing the data. Like human learning, the more quality effort expended up-front getting this step right, typically the better ML's chance of successful learning.
- **ML Step 2 - The Learning Process:** ML engineers select from a variety of **learning algorithms** that dictate how ML will learn from the training data.
- **ML Step 3 - The Learning Outcome:** ML gains knowledge, forms conclusions and creates a statistical model of what was learned. A **machine learning model** can be tuned and re-used against new data to learn more over time.

Algorithms – Traditional Software vs. Machine Learning

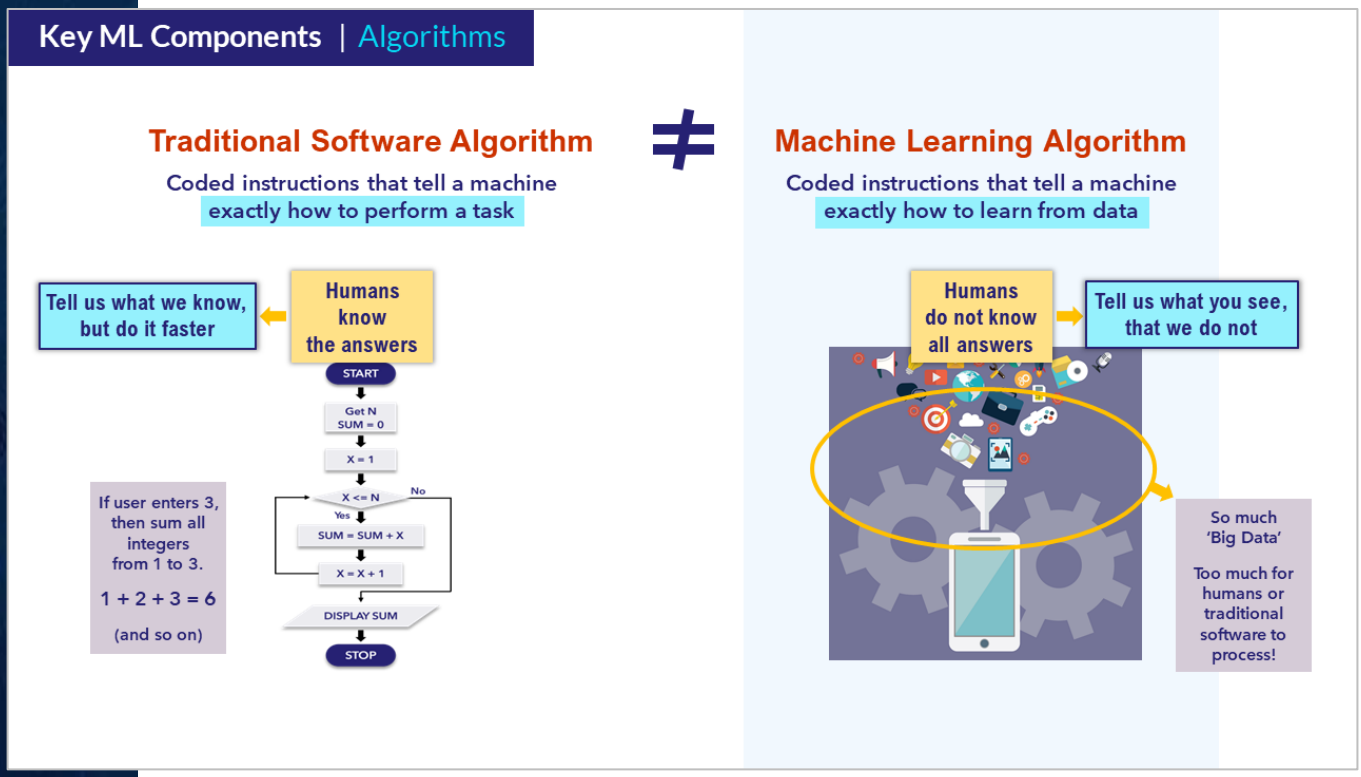


Figure 3: Differences between Traditional Software Algorithms and ML Algorithms

Learning algorithms are significantly different than traditional software algorithms.

Traditional software algorithms reinforce existing knowledge and execute tasks more efficiently.

A traditional software algorithm is a pre-defined set of instructions that tells a computer **exactly how to execute all tasks**. Humans create all the instructions, and they possess prior knowledge of all possible outcomes.

Machine Learning algorithms offer new insights that were previously unknown and execute tasks more intelligently.

A machine learning algorithm is a pre-defined set of instructions that tells a computer **exactly how to learn and how to express outcomes**. Humans do not possess prior knowledge of all possible outcomes. While humans do provide the data and the learning algorithms, we do not know up-front exactly what the computer will learn.

01 ML Definitions & Key Concepts (cont.)

Algorithms – Traditional Software vs. Machine Learning

	Traditional Software	vs.	Machine Learning Software
Algorithmic logic	Exactly how to perform the task(s)		How to learn from data (and) how to model the learning outcome
Programmed instructions	Yes - 100%		Little to no programming beyond how to learn (and) how to model
All output answers are known up-front	Yes		No
Data types and formats	• Smaller, targeted datasets • Labeling required • Prefers least data possible		• Larger, varied datasets • Labeling optional • Prefers all relevant data
Learns and improves over time without re-coding	No		Yes - when provided new datasets to study
Uncovers novel clusters, classifications or predictions	No		Yes
Considered to be intelligent?	Efficient - but not intelligent		Intelligent and efficient

The whole point of ML!

Figure 4: Traditional Software Algorithm vs. Machine Learning Algorithm

The purpose of ML is to analyze vast data and provide useful and trustworthy insights we do not already possess. When executed well, ML should uncover novel clusters, classifications or predictions and it is capable of learning more over time when exposed to new data.

Use the appropriate algorithmic tool for the job. Ensure you start with a problem that needs to be solved with ML.

> ***Best Practice:*** *Do not use AI or ML where it is not required and/or where traditional software can adequately meet the requirement.*

Implementing ML is not a straightforward task. It involves a substantial learning curve and skilled team effort in data preparation, ML engineering, computing resources, and testing. A primary reason for AI/ML project failure is the unnecessary use of AI/ML.

Machine Learning Algorithms

Prior to discussing learning algorithms, it is helpful to understanding the high-level machine learning process and the concept of data labeling.

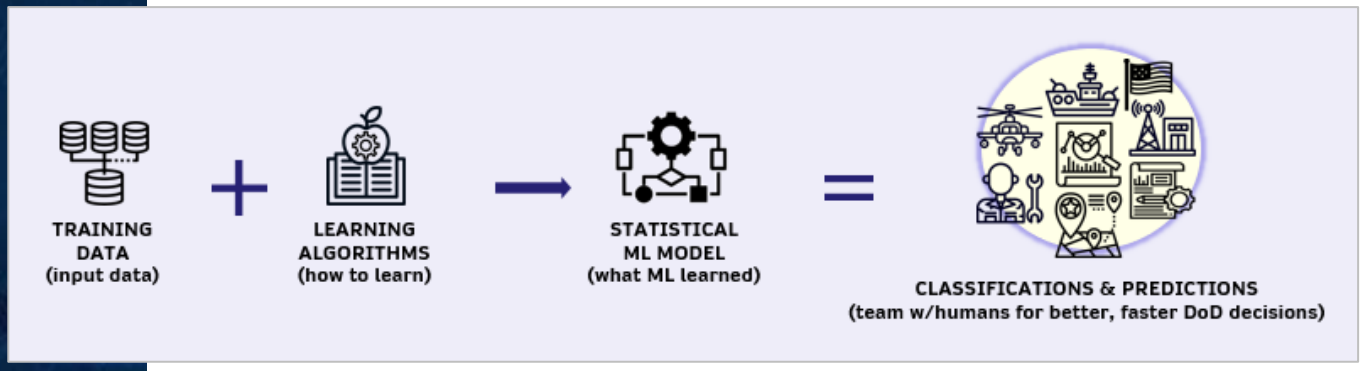


Figure 5: High-level view of the machine learning process

The high-level machine learning process:

- First, gather and prepare the **training data** for ML to study.
- Next, instruct the machine on the appropriate **learning algorithm(s)** to facilitate learning.
- Once trained on the data, ML generates output in the form of a statistical representation called a **Machine Learning Model**.
- Finally, tune and use this model to train on new data with the goal of creating enhanced insights and decision support.

Data – Unlabeled vs. Labeled

Data that is used by software comes in diverse forms and conditions. The presence of labels in data influences the choice of learning algorithm for ML engineers. Data labeling involves assigning tags or categories to each data point in a dataset. Depending on requirements, labeled data is often, but not always, preferred.

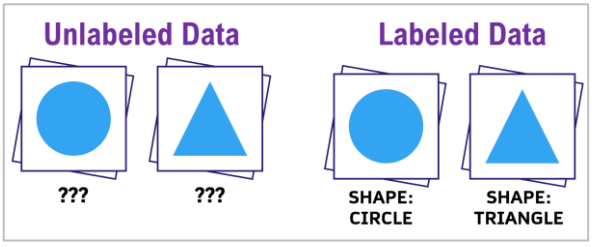


Figure 6: Unlabeled data vs. labeled data

02 ML Algorithms (cont.)

Four Types of Machine Learning Algorithms

There are **four major types of learning algorithms**, also sometimes referred to as *the four types of Machine Learning*.

Machine Learning Engineers choose from a variety of learning algorithm(s) to dictate how ML will learn from the training data. Criteria such as labeled data, problem complexity, available training data, desired accuracy, and computational resources guide the selection.

(4) Types of Learning Algorithms

Also called the **(4) Types of Machine Learning**

	Learning Algorithms	Training Data Condition/Problem Scenario
1	**Supervised Learning**	**Labeled training data.** ML system learns from labeled data by associating inputs with corresponding desired outputs. **DoD use case:** Friend or foe detection from radar data – a well-studied problem with years of labeled data.
2	**Unsupervised Learning**	**Unlabeled training data.** ML independently explores unlabeled data without answers to find hidden patterns or structures. **DoD use case:** Analyzing satellite data for space junk when we don't know what to look for.
3	**Semi-Supervised Learning**	**Labeled & unlabeled data.** Leverages a small amount of labeled data and a larger amount of unlabeled data to improve learning accuracy and efficiency. Useful when obtaining labeled data is costly or time-consuming. **DoD use case:** Cyber threat analysis where access to large quantities of labeled data is problematic.
4	**Reinforcement Learning**	**Learns by trial and error.** Involves and agent interacting with an environment and learning through trial and error to maximize a reward signal. **DoD use case:** Training autonomous drones (agents) to interact with a real-world environment.

Figure 7: Four learning algorithm types with data and usage information

- **Supervised Learning:** **Uses labeled training data.** The ML system learns from labeled data by associating inputs with corresponding desired outputs.
- **Unsupervised Learning:** **Uses unlabeled training data.** The ML system independently explores unlabeled data without un-front answers in order to uncover hidden patterns or structures.
- **Semi-supervised Learning:** **Uses a combo of labeled & unlabeled data.** Leverages a small amount of labeled data and a larger amount of unlabeled data to improve learning accuracy and efficiency. Useful when obtaining labeled data is costly or time-consuming.
- **Reinforcement Learning:** **Focuses less on initial training data.** Learns by trial and error. Involves an agent interacting with an environment and learning through trial and error to maximize a reward signal.

Learning Algorithms Complexity

Within the four types of learning algorithms, there are well over a hundred types of computational learning algorithm available for ML engineers to use. This course does not cover ML algorithms in-depth. However, acquisition planners should be aware that complexity requires data scientists, ML engineers, and ML testers with strong skills in math and statistics, programming, data handling and pre-processing, and ML algorithms.

ML engineers typically leverage existing algorithms from libraries like TensorFlow, PyTorch, and Scikit-learn, customizing them based on statistical principles and project requirements. Figure 8 depicts a few examples among many available algorithms.

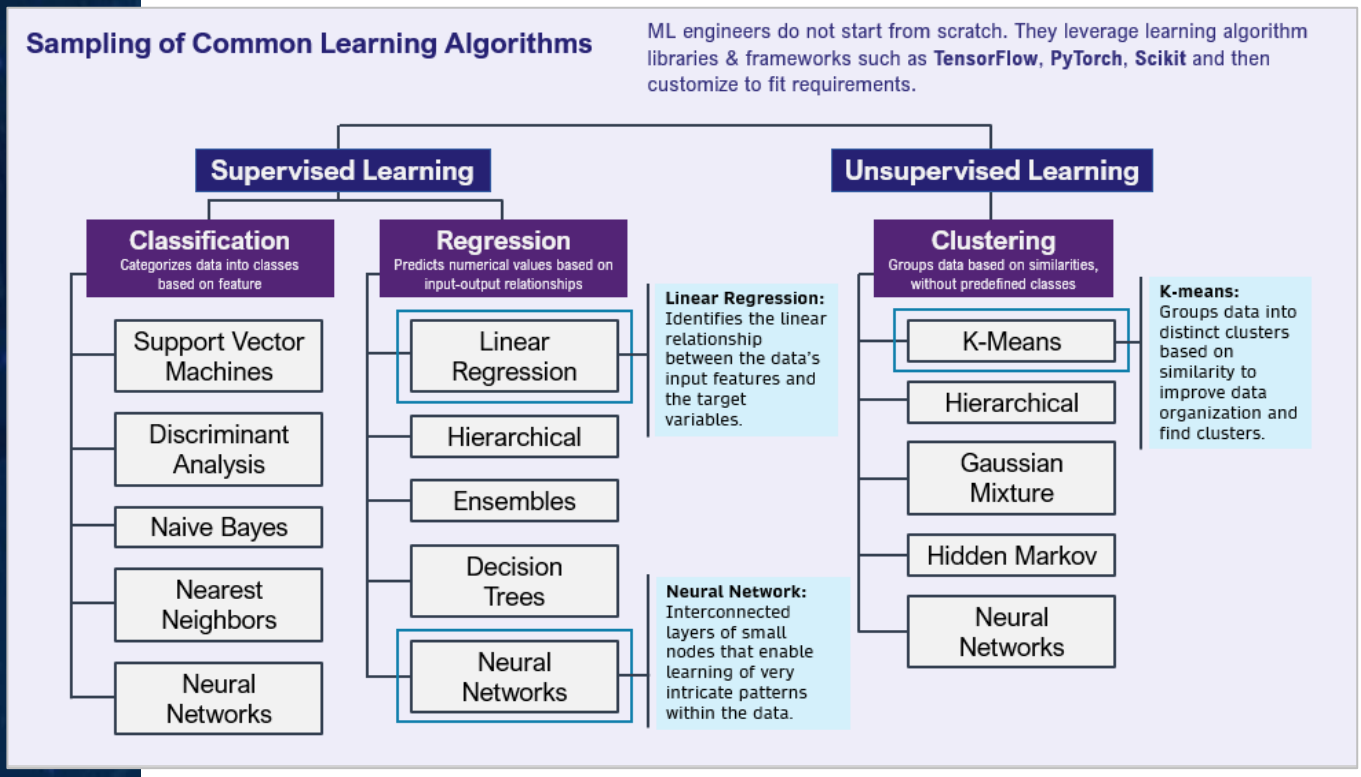


Figure 8: Small sample of common learning algorithms

03 ML Operational Process

Machine Learning Operations (MLOps)

MLOps is the practice supporting development, deployment, and management of machine learning models.

MLOps generally sits on top of DevSecOps and interoperates with multiple DataOps organizations and data sources. The primary focus of MLOps is automating and monitoring the entire ML lifecycle, ensuring efficient and seamless operations throughout the process.

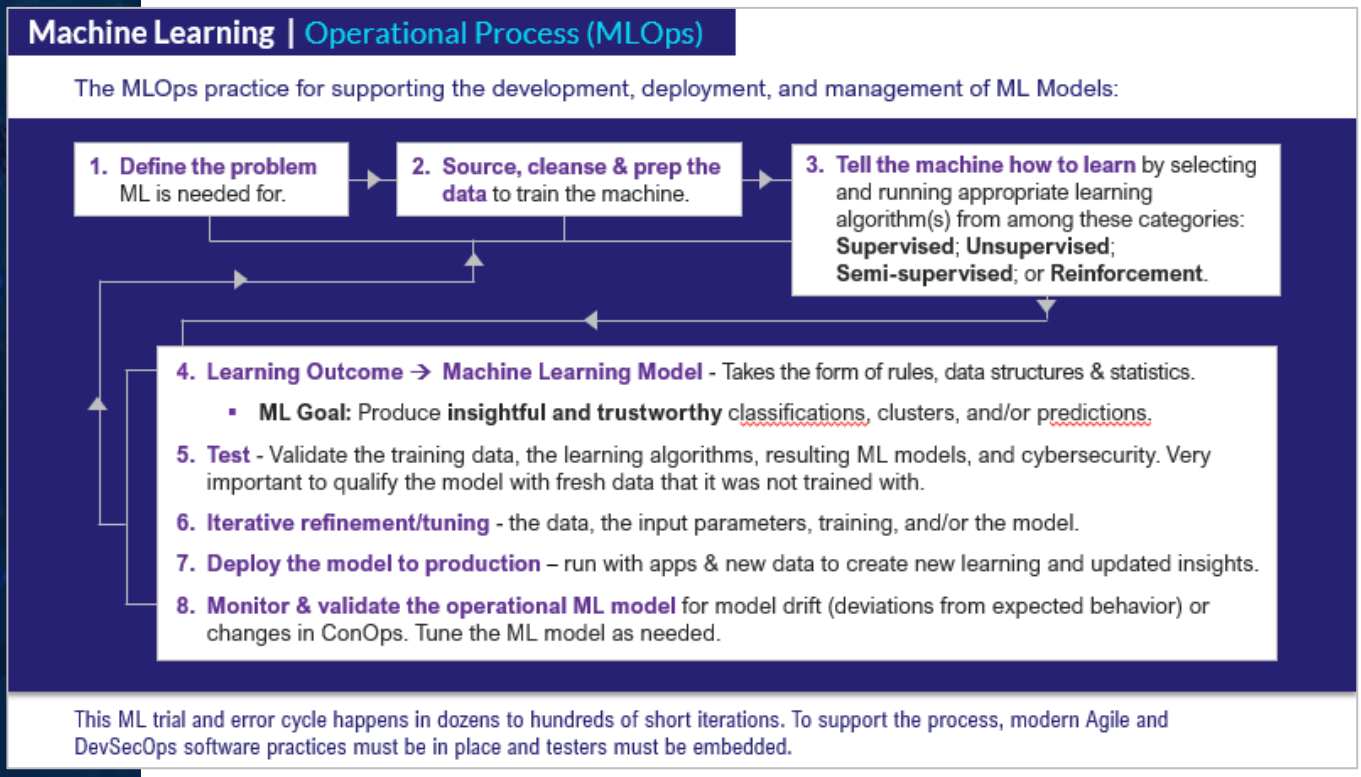


Figure 9: Functional view of the high-level MLOps process

8-Step Functional View of MLOps

The MLOps process involves a functional list of eight steps, presented linearly for better comprehension. In practice, these steps overlap and repeat numerous times during data and model tuning and testing.

- **Step 1:** Define the problem that requires the use of machine learning.
- **Step 2:** Source, cleanse, prepare and test the training data from which ML will learn.

03 ML Operational Process (cont.)

Functional View of MLOps (cont.)

- **Step 3:** Dictate how ML will learn by selecting, tuning and running the appropriate learning algorithm(s) against the training data.
- **Step 4:** Evaluate ML model output for valuable and trustworthy results.
- **Step 5:** Test and validate the ML model.
- **Step 6:** Refine and tune training data, learning algorithms, and ML models as needed.
- **Step 7:** Deploy the ML model to production.
- **Step 8:** Monitor and validate the operational ML model for model drift (deviations from expected behavior), changes in ConOps and cybersecurity.

MLOps Co-exists with DataOps and DevSecOps

To ensure a comprehensive and streamlined approach to building and deploying ML models, MLOps collaborates with DataOps and DevSecOps, to establish end-to-end workflows for DoD's software and data systems, DataOps ensures data availability, quality, and reliability for model training and evaluation, while DevSecOps automates the software development lifecycle (SDLC) and integrates cybersecurity with development, testing, deployment and operations.

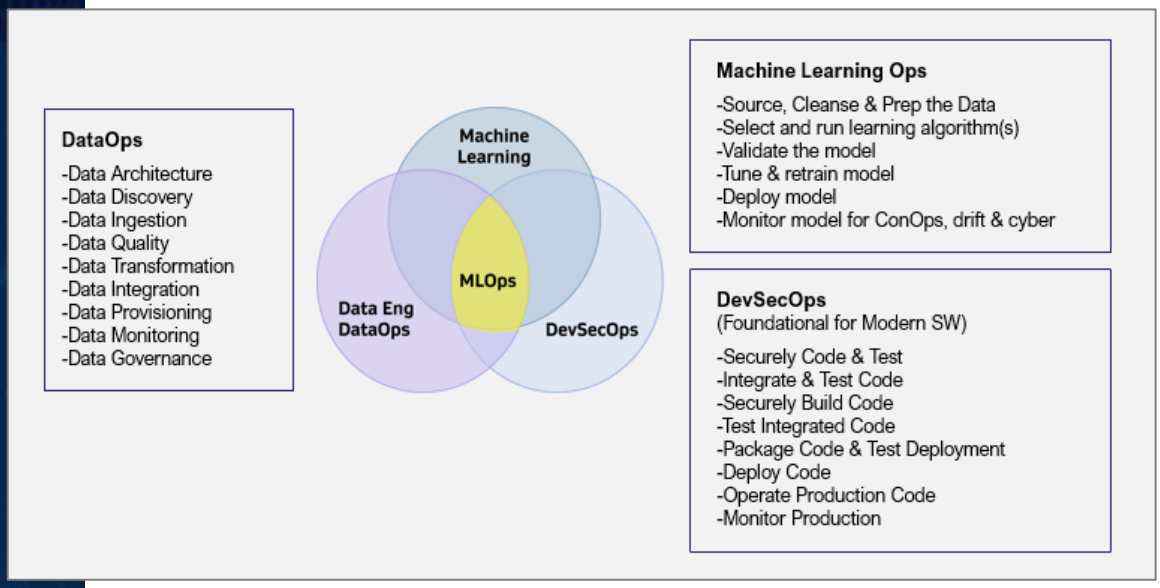


Figure 9. MLOps co-exists with DataOps and DevSecOps

04 Key Takeaways

Video Part 1: Machine Learning Definition & Key Concepts

1. **Definition:** Machine Learning (ML) is a sub-domain of Artificial Intelligence (AI) that identifies patterns in large datasets, enabling classification, prediction, and improvement over time.
    - **In DoD acquisition:** ML techniques should be considered as a potential solution for needs requiring the analysis of high-value data, gaining deeper insights, and addressing gaps where traditional software is insufficient.
    - **ML is a distinct sub-domain of AI:** operates independently or as a data analysis backend for other AI sub-domains. While ML draws inspiration from cognitive science, it does not aim to replicate human capabilities, but rather works alongside humans to enhance decision-making with valuable information.

2. **How ML learns:** ML learns from data through three stages that resemble stages of human learning.
    - **Stage 1: The input to learning:** Data scientists select and prepare **training data** from which ML will learn. Data preparation takes significant up-front time and includes cleansing, formatting, labeling, and testing. Similar to humans doing up-front prep to create learning materials to help students learn successfully.
    - **Stage 2: The learning process:** ML engineers select from among a variety **learning algorithms** that dictate how ML will learn from the training data. Similar to the way humans use diverse techniques for teaching and learning.
    - **Stage 3: The learning outcome:** A statistical representation of what the system has *learned*, called a **Machine Learning Model**, can be applied to new data to improve performance over time. Similar to the way human mental models can improve with new learning and may be adaptable to new topics.

04 Key Takeaways (cont.)

Video Part 1: Machine Learning Definition & Key Concepts (cont.)

3. **Algorithms:** ML algorithms differ from traditional algorithms in their problem-solving approach.
    i. **ML algorithms provide insights, predictions, and classifications** based on learned patterns rather than deterministic outcomes from pre-coded rules.
    ii. **ML algorithms learn patterns and rules from large datasets,** making them effective in handling complex and unstructured data, while traditional algorithms may struggle with such data.
    iii. **ML algorithms can adapt and improve over time with exposure to new data**, while traditional algorithms remain static.

Video Part 2: Machine Learning Use Cases

4. **ML Value:** Machine Learning (ML) provides value by analyzing large datasets, identifying patterns, and generating insights that surpass traditional software capabilities, making it highly relevant for DoD to consider in situations requiring efficient processing of high-value data and deeper insights beyond current capabilities.
    i. **DoD Data Volume:** It is estimated DoD generates up to 20 petabytes of data per day. A petabyte is roughly equivalent to 1,000 terabytes, 1.1 billion megabytes, or 500 billion pages of standard text.
    ii. **Definition: Edge Data:** Data that is generated in the field or at points that are far removed from centralized processing.

04 Key Takeaways (cont.)

SWE0057 ML Use Case Inventory

- **Financial fraud detection:** ML can analyze vast amounts of financial data in real-time to detect patterns and anomalies, enabling timely identification of fraudulent transactions and enhancing prevention measures.
- **Improved medical diagnostics:** ML algorithms can analyze medical imaging scans with high precision, aiding in the early detection of cancer and facilitating more effective treatment planning.
- **Facial recognition:** ML powered facial recognition systems can be used for positive purposes, such as enhancing security measures, assisting in the airline boarding process, and assisting in law enforcement.
- **Decreased military attrition:** ML has been used to examine recruiting data to predict key attributes of Marines who are more likely to stay in uniform.
- **Improved military situational awareness:** Air Force & CDAO are testing ML-based 'smart sensor' capability on MQ-9 to more intelligently process surveillance data at the edge, on the aircraft.
- **Better autonomous decisions on Mars:** NASA implemented ML to empower Mars Rover to recognize Martian rock formations that need further examination.
- **Improved military IT Helpdesk Service:** Navy is using ML-powered, hybrid-AI conversational digital-assistant to improve responses to routine service requests.
- **Improved navigation across the world:** Google's ML processes an excess of 20 petabytes of data to quickly and consistently analyze, plan, predict, recommend, inform, and re-plan automobile, bike, and foot traffic world-wide.

04 Key Takeaways (cont.)

Video Part 3: ML Learning Algorithms

5. **Recap of the High-level Machine Learning Process:**

   training data + learning algorithm → ML model = insight

   i. **Input to Learning:** Gather and prepare **training data** that ML will study.

   ii. **Learning Process:** Select an appropriate **learning algorithm** that tells ML how to learn.

   iii. **Learning Outcome:** ML generates a statistical representation of what it learned in the form of a **ML model**.

   iv. **Extended Learning:** Tune the ML model and expose it to new data to generate improved insights and decision support over time.

6. **Data Labeling:** a process of assigning tags or categories to each data point in a dataset. When ML engineers select appropriate learning algorithms to dictate how ML learns from the training data, the availability of **labeled data** significantly impacts the algorithm selection.

7. **Four Types of Learning Algorithms (or) Four Types of ML:** Each of the learning algorithm types represents a distinct approach to learning from data and has its own characteristics and applications. ML engineers select the most suitable algorithm based on the nature of the problem and state of the available data.

   i. **Supervised Learning:** ML learns from **labeled data** by associating inputs with corresponding desired outputs. Used for classification and regression, where the goal is to predict specific outcomes based on input data.

04 Key Takeaways (cont.)

Part 3: ML Learning Algorithms (cont.)

ii. **Unsupervised Learning:** Learns from **unlabeled data** by detecting patterns or structures within the data. Used for clustering, anomaly detection, dimensionality reduction, enabling insights into data without predefined labels.

iii. **Semi-supervised Learning:** Combines elements of supervised and unsupervised learning. Leverages a **small amount of labeled data and a larger amount of unlabeled data** to improve learning accuracy and efficiency. Beneficial when obtaining labeled data is costly or time-consuming.

iv. **Reinforcement Learning:** Involves **an agent interacting with an environment and learning through trial and error** to maximize a reward signal. Used for dynamic decision-making problems, game playing, robotics, and autonomous systems.

Part 4: ML Operations (MLOps)

8. **MLOps:** The agile practice supporting development, deployment and management of machine learning models, focusing on automating and monitoring the entire ML lifecycle, represented here in eight, high-level steps:

i. **Problem Definition:** Identify the problem that requires a machine learning solution.

ii. **Data Preparation:** Source, cleanse, and prepare the training data.

iii. **Algorithm Selection and Execution:** Choose and run the appropriate learning algorithms.

iv. **Model Output:** The machine learning model generates insights in the form of rules, data structures, and statistical analysis, providing classifications, clusters, and predictions.

Part 4: ML Operations (MLOps) (cont.)

v. **Test:** Continuous testing throughout the ML process. Embedded testers validate the training data, learning algorithm, model results, and cybersecurity. Models must be tested on new data (and not what it was trained on) prior to going to production.

vi. **Iterative Refinement/Tuning**: Based on the model's performance in steps 4-5, refine the data, input parameters, and/or the model itself through retraining - iterating until the desired results are achieved (dozens to hundreds of times)

vii. **Deployment:** Package and deploy the ML model to production.

viii. **Monitoring and Validation**: Continuously monitor the operational ML model in production, ensuring it behaves as expected and detects any deviations known as **model drift**. Special attention is given to cybersecurity, safeguarding against **data injection attacks**.

9. Holistic Modern Software Processing Environment:

MLOps iterations will happen dozens to hundreds of times for a single model. The timeframe for iterations is hours or days (as opposed to weeks or months).

MLOps requires embedded developers and testers and agile processes that are tailored to unique aspects of ML tuning, testing, and monitoring. Must run in a modern software environment that includes DevSecOps as a foundation, coupled with the ability to interface to a variety of DataOps organizations and data sources.

[End of Course Guidebook]