Chap1: Machine Learning in Security: An Overview #1

January 6, 2023

Devesh C Jinwala,
Professor, SVNIT and Adjunct Prof., CSE, IIT Jammu
Department of Computer Science and Engineering,
Sardar Vallabhbhai National Institute of Technology, SURAT

Chap 1: An Overview of Machine Learning in Security: Topics

- Introduction to the Course Contents, Review of the Basic Machine Learning Concepts. Foundations of Machine Learning for Security: Artificial Intelligence and Machine Learning.
  Review of the ML techniques. Machine Learning problems viz. Classification, Regression, Clustering, Association rule learning, Structured output, Ranking. Linear Regression. Logistics Regression and Bayesian Classification. Support Vector Machines, Decision Tree and Random Forest, Neural Networks, DNNs , Ensemble learning. Principal Components Analysis. Un-supervised learning algorithms: K-means for clustering problems, K-NN (k nearest neighbours). Apriori algorithm for association rule learning problems. Generative vs Discriminative learning. [4 hours]

What is Machine Learning?

What is Machine Learning?

- Learning is the process of building a scientific model after discovering knowledge from a sample data set or data sets.

What is Machine Learning?

- Learning is the process of building a scientific model after discovering knowledge from a sample data set or data sets.
- Machine learning

What is Machine Learning?

- Learning is the process of building a scientific model after discovering knowledge from a sample data set or data sets.
- Machine learning
  - is considered to be the process of applying a computing-based resource to implement learning algorithms.

What is Machine Learning?

- Learning is the process of building a scientific model after discovering knowledge from a sample data set or data sets.
- Machine learning
  - is considered to be the process of applying a computing-based resource to implement learning algorithms.
  - refers algorithms and processes that *learn* in the sense of being able to generalize past data and experiences in order to predict future outcomes.

What is Machine Learning?

- Learning is the process of building a scientific model after discovering knowledge from a sample data set or data sets.
- Machine learning
  - is considered to be the process of applying a computing-based resource to implement learning algorithms.
  - refers algorithms and processes that *learn* in the sense of being able to generalize past data and experiences in order to predict future outcomes.
  - at its core, is a set of mathematical techniques, implemented on computer systems, that enables a process of *information mining, pattern discovery, and drawing inferences from data.*

What is Machine Learning?

- Learning is the process of building a scientific model after discovering knowledge from a sample data set or data sets.
- Machine learning
  - is considered to be the process of applying a computing-based resource to implement learning algorithms.
  - refers algorithms and processes that *learn* in the sense of being able to generalize past data and experiences in order to predict future outcomes.
  - at its core, is a set of mathematical techniques, implemented on computer systems, that enables a process of *information mining, pattern discovery, and drawing inferences from data.*
  - An Example.....classifying animals as reptiles or mammals.... and the feature *gives birth to live offspring.*

What is Machine Learning?

- Learning is the process of building a scientific model after discovering knowledge from a sample data set or data sets.
- Machine learning
  - is considered to be the process of applying a computing-based resource to implement learning algorithms.
  - refers algorithms and processes that *learn* in the sense of being able to generalize past data and experiences in order to predict future outcomes.
  - at its core, is a set of mathematical techniques, implemented on computer systems, that enables a process of *information mining, pattern discovery, and drawing inferences from data.*
  - An Example.....classifying animals as reptiles or mammals.... and the feature *gives birth to live offspring.*

What is Machine Learning?

- Learning is the process of building a scientific model after discovering knowledge from a sample data set or data sets.
- Machine learning
  - is considered to be the process of applying a computing-based resource to implement learning algorithms.
  - refers algorithms and processes that *learn* in the sense of being able to generalize past data and experiences in order to predict future outcomes.
  - at its core, is a set of mathematical techniques, implemented on computer systems, that enables a process of *information mining, pattern discovery, and drawing inferences from data.*
  - An Example.....classifying animals as reptiles or mammals.... and the feature *gives birth to live offspring.*

Formal definition

- Formally, machine learning is defined as the complex computation process of automatic pattern recognition and intelligent decision making based on training sample data.

What is Machine Learning (ML) ?...

- This also means that we need to equip the machine with the ability to mimic human behavior.

What is Machine Learning (ML) ?...

- This also means that we need to equip the machine with the ability to mimic human behavior.
- ML is concerned with giving computers the ability to perform a task without being explicitly commanded/programmed.

What is Machine Learning (ML) ?...

- This also means that we need to equip the machine with the ability to mimic human behavior.
- ML is concerned with giving computers the ability to perform a task without being explicitly commanded/programmed.
  - As an example, suppose we want to classify the emails into promotional and non-promotional emails.

What is Machine Learning (ML) ?...

- This also means that we need to equip the machine with the ability to mimic human behavior.
- ML is concerned with giving computers the ability to perform a task without being explicitly commanded/programmed.
    - As an example, suppose we want to classify the emails into promotional and non-promotional emails.
        - For example, classify an email as promotional if it contains the words "Discount", "Sale", or "Free Gift".

What is Machine Learning (ML) ?...

- This also means that we need to equip the machine with the ability to mimic human behavior.
- ML is concerned with giving computers the ability to perform a task without being explicitly commanded/programmed.
    - As an example, suppose we want to classify the emails into promotional and non-promotional emails.
        - For example, classify an email as promotional if it contains the words "Discount", "Sale", or "Free Gift".
        - Classify an email as non-promotional if the email address includes ".gov" or ".edu".

What is Machine Learning (ML) ?...

- This also means that we need to equip the machine with the ability to mimic human behavior.
- ML is concerned with giving computers the ability to perform a task without being explicitly commanded/programmed.
  - As an example, suppose we want to classify the emails into promotional and non-promotional emails.
    - For example, classify an email as promotional if it contains the words “Discount”, “Sale”, or “Free Gift”.
    - Classify an email as non-promotional if the email address includes “.gov” or “.edu”.
  - How would one do so in the conventional programming ?

What is Machine Learning (ML) ?...

- This also means that we need to equip the machine with the ability to mimic human behavior.
- ML is concerned with giving computers the ability to perform a task without being explicitly commanded/programmed.
  - As an example, suppose we want to classify the emails into promotional and non-promotional emails.
    - For example, classify an email as promotional if it contains the words "Discount", "Sale", or "Free Gift".
    - Classify an email as non-promotional if the email address includes ".gov" or ".edu".
  - How would one do so in the conventional programming ?
  - What are the issues with the approach followed in the conventional programming ?

What is Machine Learning (ML) ?...

- This also means that we need to equip the machine with the ability to mimic human behavior.
- ML is concerned with giving computers the ability to perform a task without being explicitly commanded/programmed.
    - As an example, suppose we want to classify the emails into promotional and non-promotional emails.
        - For example, classify an email as promotional if it contains the words "Discount", "Sale", or "Free Gift".
        - Classify an email as non-promotional if the email address includes ".gov" or ".edu".
    - How would one do so in the conventional programming ?
    - What are the issues with the approach followed in the conventional programming ?
        - it is challenging to come up with the rules. How ?

What is Machine Learning? Motivation

- From the other angle, why would the ML based approach be required in such a scenario ?

What is Machine Learning? Motivation

- From the other angle, why would the ML based approach be required in such a scenario ?
- ML can super-charge the classification program by identifying each email's unique attributes.

What is Machine Learning? Motivation

- From the other angle, why would the ML based approach be required in such a scenario ?
- ML can super-charge the classification program by identifying each email's unique attributes.
- It would derive robust rules to automate the classification process, thereby preventing the need for manually engineered rules.

What is Machine Learning? Motivation

- From the other angle, why would the ML based approach be required in such a scenario ?
- ML can super-charge the classification program by identifying each email's unique attributes.
- It would derive robust rules to automate the classification process, thereby preventing the need for manually engineered rules.
  - But, for a machine to do that, we need to provide it with data.

What is Machine Learning? Motivation

- From the other angle, why would the ML based approach be required in such a scenario ?
- ML can super-charge the classification program by identifying each email's unique attributes.
- It would derive robust rules to automate the classification process, thereby preventing the need for manually engineered rules.
  - But, for a machine to do that, we need to provide it with data.
- Thus, the goal is for the machine to learn the rules directly from the data, using what are known as ML algorithms.

What is Machine Learning? Motivation

- From the other angle, why would the ML based approach be required in such a scenario ?
- ML can super-charge the classification program by identifying each email's unique attributes.
- It would derive robust rules to automate the classification process, thereby preventing the need for manually engineered rules.
  - But, for a machine to do that, we need to provide it with data.
- Thus, the goal is for the machine to learn the rules directly from the data, using what are known as ML algorithms.
- ML can autonomously derive robust rules to automate the classification process,

What is Machine Learning? Motivation

- From the other angle, why would the ML based approach be required in such a scenario ?
- ML can super-charge the classification program by identifying each email's unique attributes.
- It would derive robust rules to automate the classification process, thereby preventing the need for manually engineered rules.
  - But, for a machine to do that, we need to provide it with data.
- Thus, the goal is for the machine to learn the rules directly from the data, using what are known as ML algorithms.
- ML can autonomously derive robust rules to automate the classification process,
  - thereby preventing the need for manually engineered rules.

What is Machine Learning? Motivation

- From the other angle, why would the ML based approach be required in such a scenario ?
- ML can super-charge the classification program by identifying each email's unique attributes.
- It would derive robust rules to automate the classification process, thereby preventing the need for manually engineered rules.
  - But, for a machine to do that, we need to provide it with data.
- Thus, the goal is for the machine to learn the rules directly from the data, using what are known as ML algorithms.
- ML can autonomously derive robust rules to automate the classification process,
  - thereby preventing the need for manually engineered rules.
- Again, how could it derive robust rules to automate the classification ?

What is Machine Learning? Motivation...

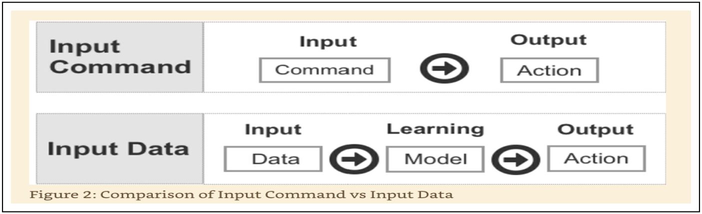

Figure: Comparison of Input command vs Input Data [Src: Oliver Theobald]

- It would do so, using what are known as machine learning algorithms.

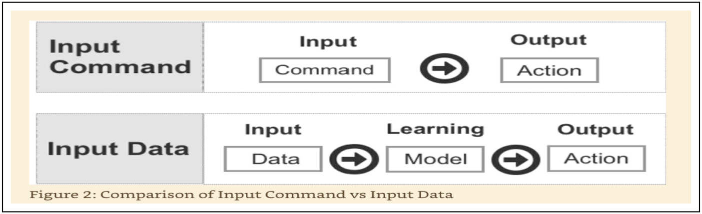

Figure: Comparison of Input command vs Input Data [Src: Oliver Theobald]

- It would do so, using what are known as machine learning algorithms.
- The formulas and procedures - derived from mathematical concepts

What is Machine Learning? Motivation...

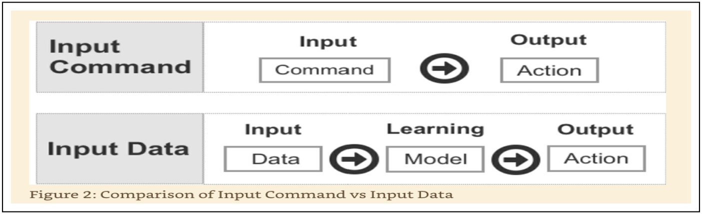

Figure: Comparison of Input command vs Input Data [Src: Oliver Theobald]

- It would do so, using what are known as machine learning algorithms.
- The formulas and procedures - derived from mathematical concepts
  - for appropriate ML task say classification, would be built into the ML algorithms.

What is Machine Learning? Motivation...

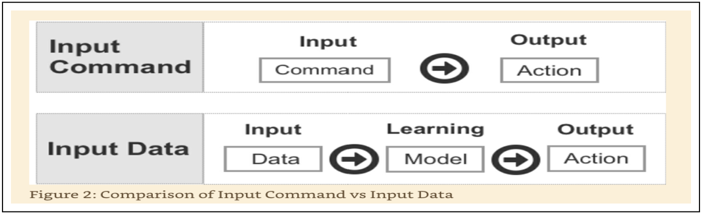

Figure: Comparison of Input command vs Input Data [Src: Oliver Theobald]

- It would do so, using what are known as machine learning algorithms.
- The formulas and procedures - derived from mathematical concepts
  - for appropriate ML task say classification, would be built into the ML algorithms.
- The ML algorithms would be implemented in programming code to perform calculations on our data

.......continued

What is Machine Learning? Motivation...

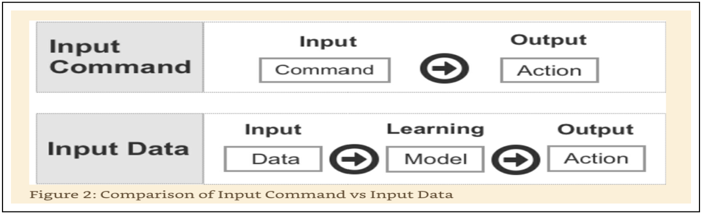

Figure: Comparison of Input command vs Input Data [Src: Oliver Theobald]

- It would do so, using what are known as machine learning algorithms.
- The formulas and procedures - derived from mathematical concepts
  - for appropriate ML task say classification, would be built into the ML algorithms.
- The ML algorithms would be implemented in programming code to perform calculations on our data
  - after which the algorithm/program typically generates an output known as a model.

......continued

What is Machine Learning? Motivation...

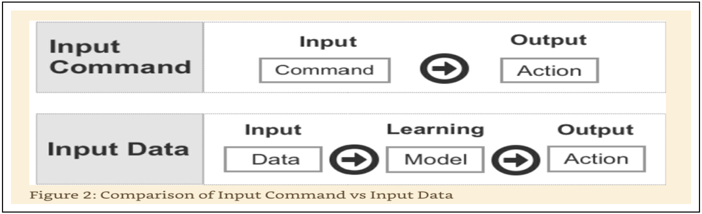

Figure: Comparison of Input command vs Input Data [Src: Oliver Theobald]

......*continued*

- The process of generating the model is known as training the model.

What is Machine Learning? Motivation...

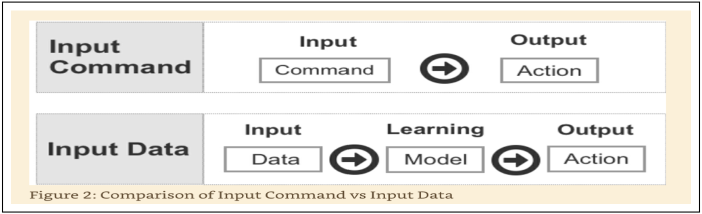

Figure: Comparison of Input command vs Input Data [Src: Oliver Theobald]

......*continued*

- The process of generating the model is known as training the model.
- This model describes the rules, numbers, and any other algorithm-specific data structures that our machine learned from the data.

What is Machine Learning? Motivation...

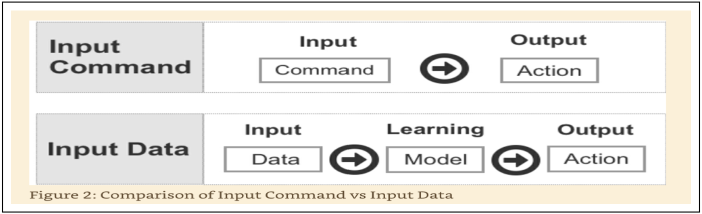

Figure: Comparison of Input command vs Input Data [Src: Oliver Theobald]

......*continued*

- The process of generating the model is known as training the model.
- This model describes the rules, numbers, and any other algorithm-specific data structures that our machine learned from the data.
- Our machine can then use the model to perform its task.

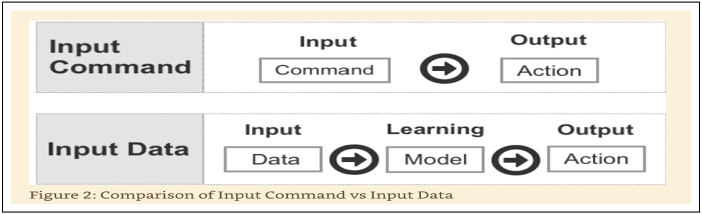

Figure: Comparison of Input command vs Input Data [Src: Oliver Theobald]

......*continued*

- The process of generating the model is known as training the model.
- This model describes the rules, numbers, and any other algorithm-specific data structures that our machine learned from the data.
- Our machine can then use the model to perform its task.
- Thus, a key characteristic of ML is the concept of self-learning.

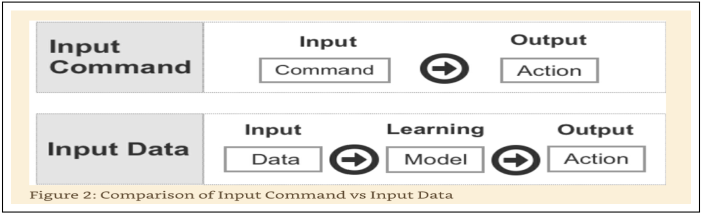

Figure: Comparison of Input command vs Input Data [Src: Oliver Theobald]

......*continued*

- The process of generating the model is known as training the model.
- This model describes the rules, numbers, and any other algorithm-specific data structures that our machine learned from the data.
- Our machine can then use the model to perform its task.
- Thus, a key characteristic of ML is the concept of self-learning.
  - i.e. the application of statistical modeling to detect patterns and improve performance based on data and empirical information; all without direct programming commands.

What is Machine Learning? Motivation...

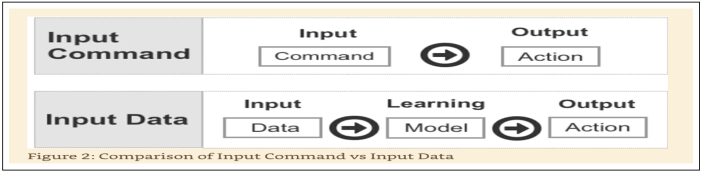

Figure: Comparison of Input command vs Input Data [Src: Oliver Theobald]

Self learning, in summary

- ML uses data as input to build a decision model.

What is Machine Learning? Motivation...

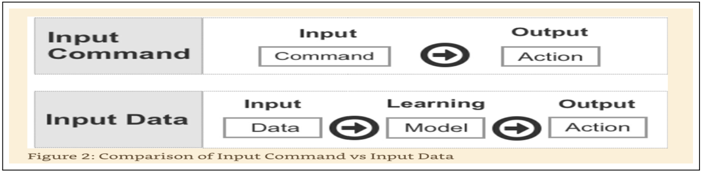

Figure: Comparison of Input command vs Input Data [Src: Oliver Theobald]

Self learning, in summary

- ML uses data as input to build a decision model.
- decisions are generated by deciphering relationships and patterns in the data using probabilistic reasoning, trial and error, and other computationally-intensive techniques.

What is Machine Learning? Motivation...

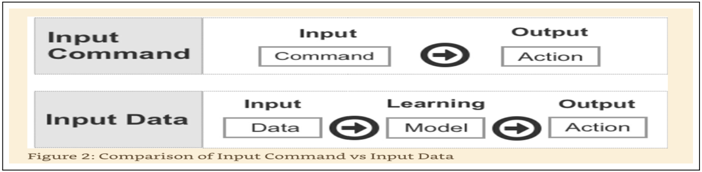

Figure: Comparison of Input command vs Input Data [Src: Oliver Theobald]

Self learning, in summary

- ML uses data as input to build a decision model.
- decisions are generated by deciphering relationships and patterns in the data using probabilistic reasoning, trial and error, and other computationally-intensive techniques.
- means that the output of the decision model is determined by the contents of the input data rather than any pre-set rules defined by a human programmer.

What is Machine Learning? Motivation...

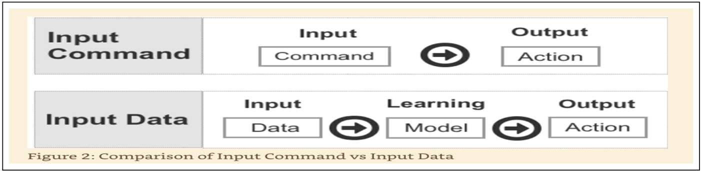

Figure: Comparison of Input command vs Input Data [Src: Oliver Theobald]

Self learning, in summary

- ML uses data as input to build a decision model.
- decisions are generated by deciphering relationships and patterns in the data using probabilistic reasoning, trial and error, and other computationally-intensive techniques.
- means that the output of the decision model is determined by the contents of the input data rather than any pre-set rules defined by a human programmer.
- the human programmer is still responsible

What is Machine Learning? Motivation...

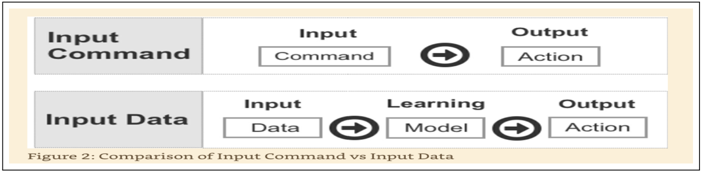

Figure: Comparison of Input command vs Input Data [Src: Oliver Theobald]

Self learning, in summary

- ML uses data as input to build a decision model.
- decisions are generated by deciphering relationships and patterns in the data using probabilistic reasoning, trial and error, and other computationally-intensive techniques.
- means that the output of the decision model is determined by the contents of the input data rather than any pre-set rules defined by a human programmer.
- the human programmer is still responsible
  - for feeding the data into the model,

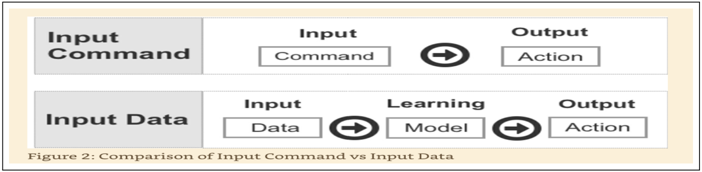

Figure: Comparison of Input command vs Input Data [Src: Oliver Theobald]

Self learning, in summary

- ML uses data as input to build a decision model.
- decisions are generated by deciphering relationships and patterns in the data using probabilistic reasoning, trial and error, and other computationally-intensive techniques.
- means that the output of the decision model is determined by the contents of the input data rather than any pre-set rules defined by a human programmer.
- the human programmer is still responsible
  - for feeding the data into the model,
  - selecting an appropriate algorithm and

What is Machine Learning? Motivation...

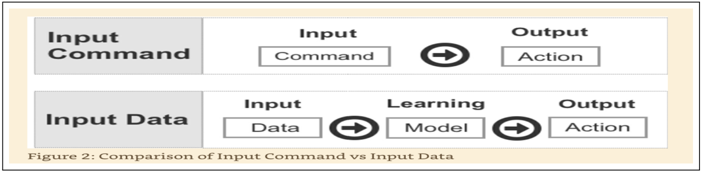

Figure: Comparison of Input command vs Input Data [Src: Oliver Theobald]

Self learning, in summary

- ML uses data as input to build a decision model.
- decisions are generated by deciphering relationships and patterns in the data using probabilistic reasoning, trial and error, and other computationally-intensive techniques.
- means that the output of the decision model is determined by the contents of the input data rather than any pre-set rules defined by a human programmer.
- the human programmer is still responsible
  - for feeding the data into the model,
  - selecting an appropriate algorithm and
  - tweaking its settings (called hyperparameters) in a bid to reduce prediction error,

Lineage of ML

Anatomy of Machine Learning: Where does it fit in

- It

Figure: Lineage of ML [Src: Oliver Theobald]

Relationship between data related fields

- Data Science comprises methods and systems to extract knowledge and insights from data with the aid of computers

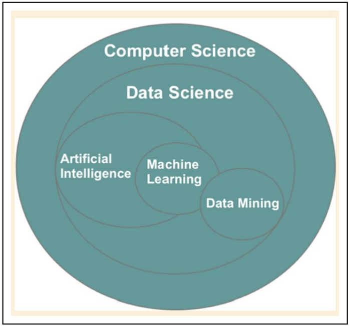

Figure: Visual representation of the relationship between data-related fields

footnoteTheobald, Oliver. Machine Learning for Absolute Beginners

Relationship between data related fields

- Data Science comprises methods and systems to extract knowledge and insights from data with the aid of computers
- AI, encompasses the ability of machines to perform intelligent and cognitive tasks

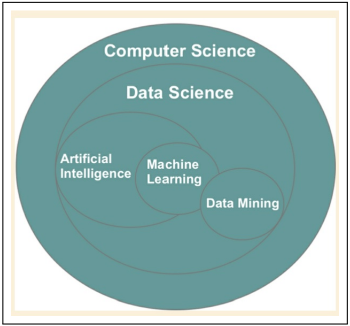

Figure: Visual representation of the relationship between data-related fields

footnoteTheobald, Oliver. Machine Learning for Absolute Beginners

Relationship between data related fields

- Data Science comprises methods and systems to extract knowledge and insights from data with the aid of computers
- AI, encompasses the ability of machines to perform intelligent and cognitive tasks
- A simile:

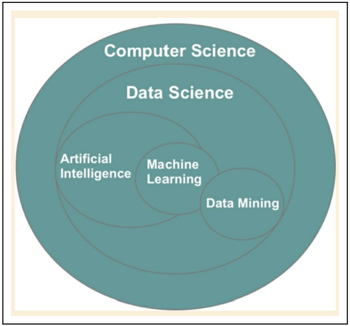

Figure: Visual representation of the relationship between data-related fields

footnoteTheobald, Oliver. Machine Learning for Absolute Beginners

- Data Science comprises methods and systems to extract knowledge and insights from data with the aid of computers
- AI, encompasses the ability of machines to perform intelligent and cognitive tasks
- A simile:
    - Industrial Revolution → an era of machines simulating physical tasks,

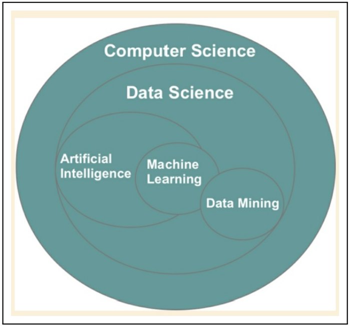

Figure: Visual representation of the relationship between data-related fields

footnoteTheobald, Oliver. Machine Learning for Absolute Beginners

Relationship between data related fields

- Data Science comprises methods and systems to extract knowledge and insights from data with the aid of computers
- AI, encompasses the ability of machines to perform intelligent and cognitive tasks
- A simile:
  - Industrial Revolution $\rightarrow$ an era of machines simulating physical tasks,
  - AI $\rightarrow$ development of machines capable of simulating cognitive abilities.

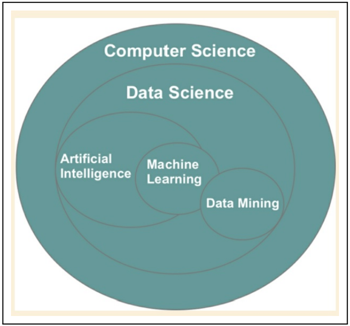

Figure: Visual representation of the relationship between data-related fields

footnoteTheobald, Oliver. Machine Learning for Absolute Beginners

Relationship between data related fields

- Data Science comprises methods and systems to extract knowledge and insights from data with the aid of computers
- AI, encompasses the ability of machines to perform intelligent and cognitive tasks
- A simile:
  - Industrial Revolution → an era of machines simulating physical tasks,
  - AI → development of machines capable of simulating cognitive abilities.
- AI includes the subfields search and planning, reasoning and knowledge representation, perception, natural language processing (NLP), and machine learning

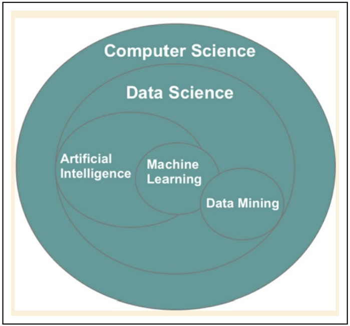

Figure: Visual representation of the relationship between data-related fields

footnoteTheobald, Oliver. Machine Learning for Absolute Beginners

Machine learning overlaps with data mining

- ML overlaps with data mining—a discipline based on discovering and unearthing patterns in large datasets.

Technique	Input is Known	Output is Known	Methodology
Data Mining	✓		Analyzes inputs to generate an unknown output.
Supervised Learning	✓	✓	Analyzes combinations of known inputs and outputs to predict future outputs based on new input data.
Unsupervised Learning	✓		Analyzes inputs to generate an output—algorithms may differ from data mining.
Reinforcement Learning		✓	Randomly trials a high number of input variables to produce a desired output.

Table 1: Comparison of techniques based on the utility of input and output data/variable

Figure: Visual representation of the relationship between data-related fields[a]

[a] Theobald, Oliver. Machine Learning for Absolute Beginners

Machine learning overlaps with data mining

- ML overlaps with data mining—a discipline based on discovering and unearthing patterns in large datasets.
- Both techniques rely on inferential methods, i.e. predicting outcomes based on other outcomes and probabilistic reasoning,

Technique	Input is Known	Output is Known	Methodology
Data Mining	✓		Analyzes inputs to generate an unknown output.
Supervised Learning	✓	✓	Analyzes combinations of known inputs and outputs to predict future outputs based on new input data.
Unsupervised Learning	✓		Analyzes inputs to generate an output—algorithms may differ from data mining.
Reinforcement Learning		✓	Randomly trials a high number of input variables to produce a desired output.

Table 1: Comparison of techniques based on the utility of input and output data/variable

Figure: Visual representation of the relationship between data-related fields[a]

[a] Theobald, Oliver. Machine Learning for Absolute Beginners

Machine learning overlaps with data mining

- ML overlaps with data mining—a discipline based on discovering and unearthing patterns in large datasets.
- Both techniques rely on inferential methods, i.e. predicting outcomes based on other outcomes and probabilistic reasoning,
- both draw from a similar assortment of algorithms includingprincipal component analysis, regression analysis, decision trees, and clustering techniques

Technique	Input is Known	Output is Known	Methodology
Data Mining	✓		Analyzes inputs to generate an unknown output.
Supervised Learning	✓	✓	Analyzes combinations of known inputs and outputs to predict future outputs based on new input data.
Unsupervised Learning	✓		Analyzes inputs to generate an output—algorithms may differ from data mining.
Reinforcement Learning		✓	Randomly trials a high number of input variables to produce a desired output.

Table 1: Comparison of techniques based on the utility of input and output data/variable

Figure: Visual representation of the relationship between data-related fields[a]

[a] Theobald, Oliver. Machine Learning for Absolute Beginners

Machine learning overlaps with data mining

- ML emphasizes the incremental process of self-learning and automatically detecting patterns through experience derived from exposure to data,

Technique	Input is Known	Output is Known	Methodology
Data Mining	✓		Analyzes inputs to generate an unknown output.
Supervised Learning	✓	✓	Analyzes combinations of known inputs and outputs to predict future outputs based on new input data.
Unsupervised Learning	✓		Analyzes inputs to generate an output—algorithms may differ from data mining.
Reinforcement Learning		✓	Randomly trials a high number of input variables to produce a desired output.

Table 1: Comparison of techniques based on the utility of input and output data/variables

Figure: Visual representation of the relationship between data-related fields[a]

[a] Theobald, Oliver. Machine Learning for Absolute Beginners

Machine learning overlaps with data mining

- ML emphasizes the incremental process of self-learning and automatically detecting patterns through experience derived from exposure to data,
- whereas, data mining is a less autonomous technique of extracting hidden insight

Technique	Input is Known	Output is Known	Methodology
Data Mining	✓		Analyzes inputs to generate an unknown output.
Supervised Learning	✓	✓	Analyzes combinations of known inputs and outputs to predict future outputs based on new input data.
Unsupervised Learning	✓		Analyzes inputs to generate an output—algorithms may differ from data mining.
Reinforcement Learning		✓	Randomly trials a high number of input variables to produce a desired output.

Table 1: Comparison of techniques based on the utility of input and output data/variables

Figure: Visual representation of the relationship between data-related fields[a]

[a] Theobald, Oliver. Machine Learning for Absolute Beginners

Machine learning overlaps with data mining

- ML emphasizes the incremental process of self-learning and automatically detecting patterns through experience derived from exposure to data,
- whereas, data mining is a less autonomous technique of extracting hidden insight
  - it seeks out patterns and relationships that are yet to be mined - well-suited for understanding large datasets with complex patterns

Technique	Input is Known	Output is Known	Methodology
Data Mining	✓		Analyzes inputs to generate an unknown output.
Supervised Learning	✓	✓	Analyzes combinations of known inputs and outputs to predict future outputs based on new input data.
Unsupervised Learning	✓		Analyzes inputs to generate an output—algorithms may differ from data mining.
Reinforcement Learning		✓	Randomly trials a high number of input variables to produce a desired output.

Table 1: Comparison of techniques based on the utility of input and output data/variables

Figure: Visual representation of the relationship between data-related fields[a]

[a] Theobald, Oliver. Machine Learning for Absolute Beginners

Machine learning overlaps with data mining

- ML emphasizes the incremental process of self-learning and automatically detecting patterns through experience derived from exposure to data,
- whereas, data mining is a less autonomous technique of extracting hidden insight
  - it seeks out patterns and relationships that are yet to be mined - well-suited for understanding large datasets with complex patterns
- An Example: Excavation operation on sites by two different team of archaeologists...

Technique	Input is Known	Output is Known	Methodology
Data Mining	✓		Analyzes inputs to generate an unknown output.
Supervised Learning	✓	✓	Analyzes combinations of known inputs and outputs to predict future outputs based on new input data.
Unsupervised Learning	✓		Analyzes inputs to generate an output—algorithms may differ from data mining.
Reinforcement Learning		✓	Randomly trials a high number of input variables to produce a desired output.

Table 1: Comparison of techniques based on the utility of input and output data/variables

Figure: Visual representation of the relationship between data-related fields[a]

[a] Theobald, Oliver. Machine Learning for Absolute Beginners

What Machine Learning is Not?

- Artificial intelligence

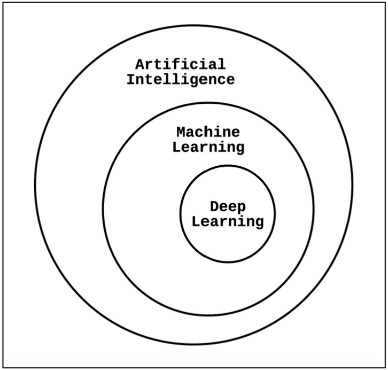

Figure: ML is part of AI[a]

[a] Clarence and Chio, ML and Security..., O'Reilly Media

What Machine Learning is Not?

- Artificial intelligence
  - indicates algorithmic solutions to complex problems typically solved by humans.

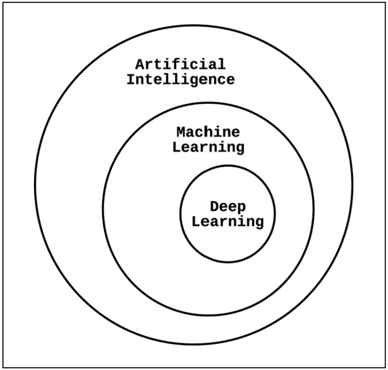

Figure: ML is part of AI[a]

[a] Clarence and Chio, ML and Security..., O'Reilly Media

What Machine Learning is Not?

- Artificial intelligence
  - indicates algorithmic solutions to complex problems typically solved by humans.
  - systems are loosely defined to be machine-driven decision engines that can achieve near-human-level intelligence.

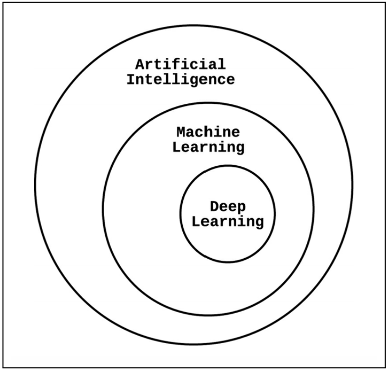

Figure: ML is part of AI[a]

[a] Clarence and Chio, ML and Security..., O'Reilly Media

What Machine Learning is Not?

- Artificial intelligence
  - indicates algorithmic solutions to complex problems typically solved by humans.
  - systems are loosely defined to be machine-driven decision engines that can achieve near-human-level intelligence.
- Machine learning

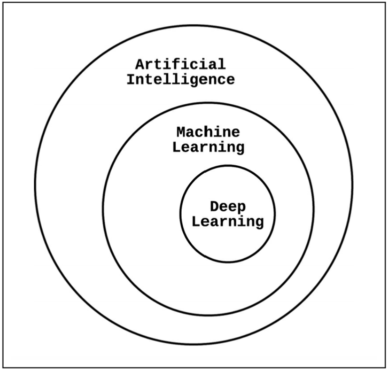

Figure: ML is part of AI[a]

[a] Clarence and Chio, ML and Security..., O'Reilly Media

What Machine Learning is Not?

- Artificial intelligence
  - indicates algorithmic solutions to complex problems typically solved by humans.
  - systems are loosely defined to be machine-driven decision engines that can achieve near-human-level intelligence.
- Machine learning
  - is a core building block for AI.

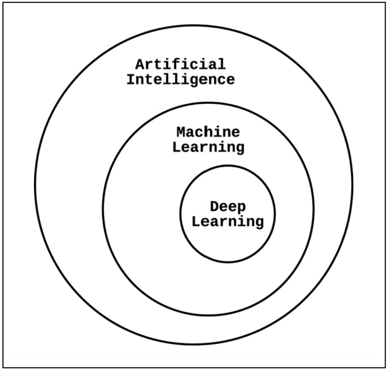

Figure: ML is part of AI[a]

[a] Clarence and Chio, ML and Security..., O'Reilly Media

What Machine Learning is Not?

- Artificial intelligence
  - indicates algorithmic solutions to complex problems typically solved by humans.
  - systems are loosely defined to be machine-driven decision engines that can achieve near-human-level intelligence.
- Machine learning
  - is a core building block for AI.
  - helps us create AI, but is not the only way to achieve it.

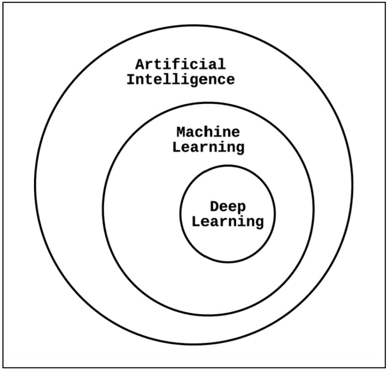

Figure: ML is part of AI[a]

[a] Clarence and Chio, ML and Security..., O'Reilly Media

What Machine Learning is Not?

- Artificial intelligence
  - indicates algorithmic solutions to complex problems typically solved by humans.
  - systems are loosely defined to be machine-driven decision engines that can achieve near-human-level intelligence.
- Machine learning
  - is a core building block for AI.
  - helps us create AI, but is not the only way to achieve it.
  - refers to statistical learning algorithms that are able to create generalizable abstractions (models) by seeing and dissecting a dataset.

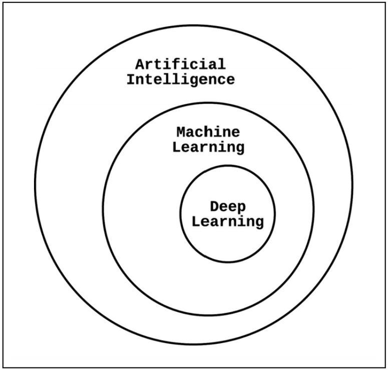

Figure: ML is part of AI[a]

[a] Clarence and Chio, ML and Security..., O'Reilly Media

What Machine Learning is Not?

- Artificial intelligence
  - indicates algorithmic solutions to complex problems typically solved by humans.
  - systems are loosely defined to be machine-driven decision engines that can achieve near-human-level intelligence.
- Machine learning
  - is a core building block for AI.
  - helps us create AI, but is not the only way to achieve it.
  - refers to statistical learning algorithms that are able to create generalizable abstractions (models) by seeing and dissecting a dataset.
- e.g. self-driving car's functions in a self-driving system

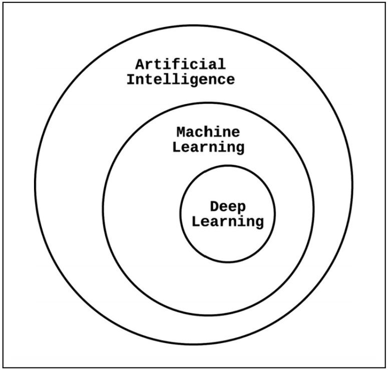

Figure: ML is part of AI[a]

[a] Clarence and Chio, ML and Security..., O'Reilly Media

Data Mining

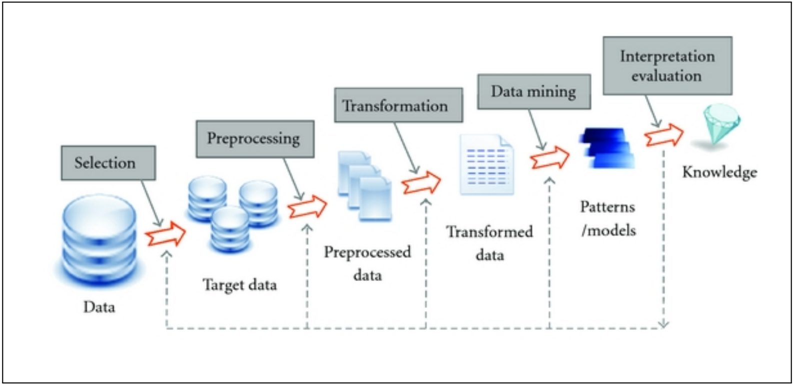

Figure: Data Mining

Data Mining...

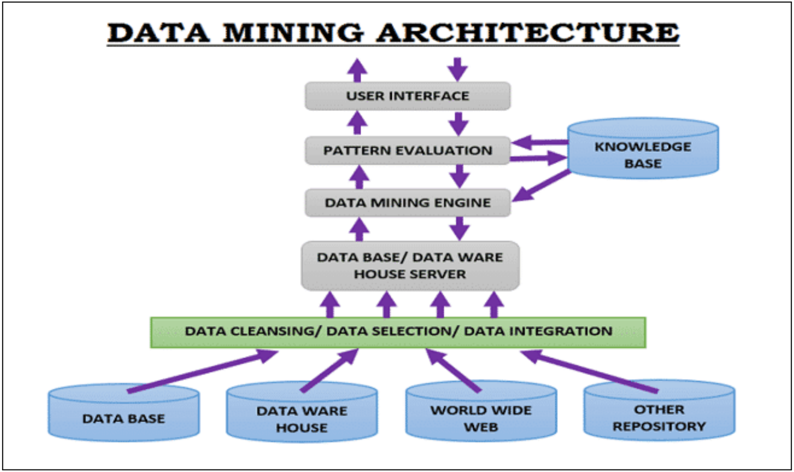

Figure: Data Mining

Data mining requires machine learning

Data mining to machine learning

- the goal is to develop of predictive models that enable a real-time cyber response after a sequence of cybersecurity processes,

Data mining requires machine learning

Data mining to machine learning

- the goal is to develop of predictive models that enable a real-time cyber response after a sequence of cybersecurity processes,
- the response must include real-time data sampling, selection, analysis and query, and mining peta-scale data, with the goal to

Data mining requires machine learning

Data mining to machine learning

- the goal is to develop of predictive models that enable a real-time cyber response after a sequence of cybersecurity processes,
- the response must include real-time data sampling, selection, analysis and query, and mining peta-scale data, with the goal to
  - to classify and detect attacks and intrusions on a computer network.

Data mining requires machine learning

Data mining to machine learning

- the goal is to develop of predictive models that enable a real-time cyber response after a sequence of cybersecurity processes,
- the response must include real-time data sampling, selection, analysis and query, and mining peta-scale data, with the goal to
  - to classify and detect attacks and intrusions on a computer network.
  - learning user patterns and/or behaviors - that is critical for intrusion detection and attack predictions

Data mining requires machine learning

Data mining to machine learning

- the goal is to develop of predictive models that enable a real-time cyber response after a sequence of cybersecurity processes,
- the response must include real-time data sampling, selection, analysis and query, and mining peta-scale data, with the goal to
  - to classify and detect attacks and intrusions on a computer network.
  - learning user patterns and/or behaviors - that is critical for intrusion detection and attack predictions
- can identify and describe structural patterns in the data automatically and theoretically explain data and predict patterns.

Data mining requires machine learning

Data mining to machine learning

- the goal is to develop of predictive models that enable a real-time cyber response after a sequence of cybersecurity processes,
- the response must include real-time data sampling, selection, analysis and query, and mining peta-scale data, with the goal to
  - to classify and detect attacks and intrusions on a computer network.
  - learning user patterns and/or behaviors - that is critical for intrusion detection and attack predictions
- can identify and describe structural patterns in the data automatically and theoretically explain data and predict patterns.
- automatic and theoretic learning require complex computation that calls for abundant machine-learning algorithms.

Overview of ML tasks and Examples

Two of the most common tasks that ML models perform

- Three of the most common tasks ML models perform are
  - classification - e.g., classifying emails into promotional and non-promotional
  - prediction - e.g., predicting stock prices.
  - regression - e.g. predicting how much a used car would sell for given historical data on recent used car sales in the area

Two of the most common tasks that ML models perform

- Three of the most common tasks ML models perform are
  - classification - e.g., classifying emails into promotional and non-promotional
  - prediction - e.g., predicting stock prices.
  - regression - e.g. predicting how much a used car would sell for given historical data on recent used car sales in the area
- There are other tasks that include
- making recommendations,

Two of the most common tasks that ML models perform

- Three of the most common tasks ML models perform are
  - classification - e.g., classifying emails into promotional and non-promotional
  - prediction - e.g., predicting stock prices.
  - regression - e.g. predicting how much a used car would sell for given historical data on recent used car sales in the area
- There are other tasks that include
- making recommendations,
- image recognition, and

Two of the most common tasks that ML models perform

- Three of the most common tasks ML models perform are
  - classification - e.g., classifying emails into promotional and non-promotional
  - prediction - e.g., predicting stock prices.
  - regression - e.g. predicting how much a used car would sell for given historical data on recent used car sales in the area
- There are other tasks that include
- making recommendations,
- image recognition, and
- natural language processing

Two of the most common tasks that ML models perform

- Three of the most common tasks ML models perform are
  - classification - e.g., classifying emails into promotional and non-promotional
  - prediction - e.g., predicting stock prices.
  - regression - e.g. predicting how much a used car would sell for given historical data on recent used car sales in the area
- There are other tasks that include
- making recommendations,
- image recognition, and
- natural language processing
- transcription

Two of the most common tasks that ML models perform

- Three of the most common tasks ML models perform are
  - classification - e.g., classifying emails into promotional and non-promotional
  - prediction - e.g., predicting stock prices.
  - regression - e.g. predicting how much a used car would sell for given historical data on recent used car sales in the area
- There are other tasks that include
- making recommendations,
- image recognition, and
- natural language processing
- transcription
- machine translation

Two of the most common tasks that ML models perform

- Three of the most common tasks ML models perform are
  - classification - e.g., classifying emails into promotional and non-promotional
  - prediction - e.g., predicting stock prices.
  - regression - e.g. predicting how much a used car would sell for given historical data on recent used car sales in the area
- There are other tasks that include
- making recommendations,
- image recognition, and
- natural language processing
- transcription
- machine translation
- anomaly detection

Two of the most common tasks that ML models perform

- Three of the most common tasks ML models perform are
  - classification - e.g., classifying emails into promotional and non-promotional
  - prediction - e.g., predicting stock prices.
  - regression - e.g. predicting how much a used car would sell for given historical data on recent used car sales in the area
- There are other tasks that include
- making recommendations,
- image recognition, and
- natural language processing
- transcription
- machine translation
- anomaly detection
- synthesis & sampling

Two of the most common tasks that ML models perform

- Three of the most common tasks ML models perform are
  - classification - e.g., classifying emails into promotional and non-promotional
  - prediction - e.g., predicting stock prices.
  - regression - e.g. predicting how much a used car would sell for given historical data on recent used car sales in the area
- There are other tasks that include
  - making recommendations,
  - image recognition, and
  - natural language processing
  - transcription
  - machine translation
  - anomaly detection
  - synthesis & sampling
  - estimation of probability density and probability mass function

Two of the most common tasks that ML models perform

- Three of the most common tasks ML models perform are
  - classification - e.g., classifying emails into promotional and non-promotional
  - prediction - e.g., predicting stock prices.
  - regression - e.g. predicting how much a used car would sell for given historical data on recent used car sales in the area
- There are other tasks that include
- making recommendations,
- image recognition, and
- natural language processing
- transcription
- machine translation
- anomaly detection
- synthesis & sampling
- estimation of probability density and probability mass function
- similarity matching

Two of the most common tasks that ML models perform

- Three of the most common tasks ML models perform are
  - classification - e.g., classifying emails into promotional and non-promotional
  - prediction - e.g., predicting stock prices.
  - regression - e.g. predicting how much a used car would sell for given historical data on recent used car sales in the area
- There are other tasks that include
  - making recommendations,
  - image recognition, and
  - natural language processing
  - transcription
  - machine translation
  - anomaly detection
  - synthesis & sampling
  - estimation of probability density and probability mass function
  - similarity matching
  - co-occurrence grouping

Two of the most common tasks that ML models perform

- Three of the most common tasks ML models perform are
  - classification - e.g., classifying emails into promotional and non-promotional
  - prediction - e.g., predicting stock prices.
  - regression - e.g. predicting how much a used car would sell for given historical data on recent used car sales in the area
- There are other tasks that include
- making recommendations,
- image recognition, and
- natural language processing
- transcription
- machine translation
- anomaly detection
- synthesis & sampling
- estimation of probability density and probability mass function
- similarity matching
- co-occurrence grouping
- causal modeling

Two of the most common tasks that ML models perform

- Three of the most common tasks ML models perform are
  - classification - e.g., classifying emails into promotional and non-promotional
  - prediction - e.g., predicting stock prices.
  - regression - e.g. predicting how much a used car would sell for given historical data on recent used car sales in the area
- There are other tasks that include
- making recommendations,
- image recognition, and
- natural language processing
- transcription
- machine translation
- anomaly detection
- synthesis & sampling
- estimation of probability density and probability mass function
- similarity matching
- co-occurrence grouping
- causal modeling
- link profiling and so on....

The key development phases in ML used to solve the different tasks

- Following are the key development phases that are used to solve the different tasks listed in the previous slide.

The key development phases in ML used to solve the different tasks

- Following are the key development phases that are used to solve the different tasks listed in the previous slide.
- These form the key phases of the machine learning models' (MLM) development lifecycle.

The key development phases in ML used to solve the different tasks

- Following are the key development phases that are used to solve the different tasks listed in the previous slide.
- These form the key phases of the machine learning models' (MLM) development lifecycle.
    - Data gathering

The key development phases in ML used to solve the different tasks

- Following are the key development phases that are used to solve the different tasks listed in the previous slide.
- These form the key phases of the machine learning models' (MLM) development lifecycle.
    - Data gathering
    - Data preprocessing

The key development phases in ML used to solve the different tasks

- Following are the key development phases that are used to solve the different tasks listed in the previous slide.
- These form the key phases of the machine learning models' (MLM) development lifecycle.
  - Data gathering
  - Data preprocessing
  - Exploratory data analysis (EDA)

The key development phases in ML used to solve the different tasks

- Following are the key development phases that are used to solve the different tasks listed in the previous slide.
- These form the key phases of the machine learning models' (MLM) development lifecycle.
  - Data gathering
  - Data preprocessing
  - Exploratory data analysis (EDA)
  - Feature engineering including feature creation/extraction, feature selection, dimensionality reduction

The key development phases in ML used to solve the different tasks

- Following are the key development phases that are used to solve the different tasks listed in the previous slide.
- These form the key phases of the machine learning models' (MLM) development lifecycle.
    - Data gathering
    - Data preprocessing
    - Exploratory data analysis (EDA)
    - Feature engineering including feature creation/extraction, feature selection, dimensionality reduction
    - Training machine learning models

The key development phases in ML used to solve the different tasks

- Following are the key development phases that are used to solve the different tasks listed in the previous slide.
- These form the key phases of the machine learning models' (MLM) development lifecycle.
  - Data gathering
  - Data preprocessing
  - Exploratory data analysis (EDA)
  - Feature engineering including feature creation/extraction, feature selection, dimensionality reduction
  - Training machine learning models
  - Model / Algorithm selection

- Following are the key development phases that are used to solve the different tasks listed in the previous slide.
- These form the key phases of the machine learning models' (MLM) development lifecycle.
  - Data gathering
  - Data preprocessing
  - Exploratory data analysis (EDA)
  - Feature engineering including feature creation/extraction, feature selection, dimensionality reduction
  - Training machine learning models
  - Model / Algorithm selection
  - Testing and matching

The key development phases in ML used to solve the different tasks

- Following are the key development phases that are used to solve the different tasks listed in the previous slide.
- These form the key phases of the machine learning models' (MLM) development lifecycle.
  - Data gathering
  - Data preprocessing
  - Exploratory data analysis (EDA)
  - Feature engineering including feature creation/extraction, feature selection, dimensionality reduction
  - Training machine learning models
  - Model / Algorithm selection
  - Testing and matching
  - Model monitoring

The key development phases in ML used to solve the different tasks

- Following are the key development phases that are used to solve the different tasks listed in the previous slide.
- These form the key phases of the machine learning models' (MLM) development lifecycle.
  - Data gathering
  - Data preprocessing
  - Exploratory data analysis (EDA)
  - Feature engineering including feature creation/extraction, feature selection, dimensionality reduction
  - Training machine learning models
  - Model / Algorithm selection
  - Testing and matching
  - Model monitoring
  - Model retraining

A Few Example Applications of some ML tasks

How the power of ML can be exploited....e.g. for analyzing YouTube viewing habits

- Say in one application, the decision model identifies a significant relationship among data scientists who like watching cat videos

A Few Example Applications of some ML tasks

How the power of ML can be exploited....e.g. for analyzing YouTube viewing habits

- Say in one application, the decision model identifies a significant relationship among data scientists who like watching cat videos
  - the machine analyzes which videos data scientists enjoy watching on YouTube based on user engagement

A Few Example Applications of some ML tasks

How the power of ML can be exploited....e.g. for analyzing YouTube viewing habits

- Say in one application, the decision model identifies a significant relationship among data scientists who like watching cat videos
  - the machine analyzes which videos data scientists enjoy watching on YouTube based on user engagement
  - user engagement is measured in likes, subscribes, and repeat viewing.

A Few Example Applications of some ML tasks

How the power of ML can be exploited....e.g. for analyzing YouTube viewing habits

- Say in one application, the decision model identifies a significant relationship among data scientists who like watching cat videos
  - the machine analyzes which videos data scientists enjoy watching on YouTube based on user engagement
  - user engagement is measured in likes, subscribes, and repeat viewing.
- the model developed would be specific to this application, as compared to any other, with different objectives.....

A Few Example Applications of some ML tasks

How the power of ML can be exploited....e.g. for analyzing YouTube viewing habits

- Say in one application, the decision model identifies a significant relationship among data scientists who like watching cat videos
  - the machine analyzes which videos data scientists enjoy watching on YouTube based on user engagement
  - user engagement is measured in likes, subscribes, and repeat viewing.
- the model developed would be specific to this application, as compared to any other, with different objectives.....

A Few Example Applications of some ML tasks

How the power of ML can be exploited....e.g. for analyzing YouTube viewing habits

- Say in one application, the decision model identifies a significant relationship among data scientists who like watching cat videos
  - the machine analyzes which videos data scientists enjoy watching on YouTube based on user engagement
  - user engagement is measured in likes, subscribes, and repeat viewing.
- the model developed would be specific to this application, as compared to any other, with different objectives.....

...........*this is continued*

A Few Examples of ML tasks

How the power of ML can be exploited....analyzing YouTube viewing habits

- Say in another application, the decision model identifies patterns among the physical traits of baseball players and

A Few Examples of ML tasks

How the power of ML can be exploited....analyzing YouTube viewing habits

- Say in another application, the decision model identifies patterns among the physical traits of baseball players and
- identifies their likelihood of winning the season's Most Valuable Player (MVP) award.

A Few Examples of ML tasks

How the power of ML can be exploited....analyzing YouTube viewing habits

- Say in another application, the decision model identifies patterns among the physical traits of baseball players and
- identifies their likelihood of winning the season's Most Valuable Player (MVP) award.
    - here, the machine assesses the physical attributes of previous baseball MVPs among other features such as age and education.

A Few Examples of ML tasks

How the power of ML can be exploited....analyzing YouTube viewing habits

- Say in another application, the decision model identifies patterns among the physical traits of baseball players and
- identifies their likelihood of winning the season's Most Valuable Player (MVP) award.
  - here, the machine assesses the physical attributes of previous baseball MVPs among other features such as age and education.
- Here, this model is different from the one in the previous slide.....and hence produces different output. But, at no stage the decision model is programmed to produce the two required outcomes.

A Few Examples of ML tasks

How the power of ML can be exploited....analyzing YouTube viewing habits

- Say in another application, the decision model identifies patterns among the physical traits of baseball players and
- identifies their likelihood of winning the season's Most Valuable Player (MVP) award.
    - here, the machine assesses the physical attributes of previous baseball MVPs among other features such as age and education.
- Here, this model is different from the one in the previous slide.....and hence produces different output. But, at no stage the decision model is programmed to produce the two required outcomes.
- the model only decodes the complex patterns in the input data, and uses machine learning to find connections without human help.

A Few Examples of ML tasks

How the power of ML can be exploited....analyzing YouTube viewing habits

- Say in another application, the decision model identifies patterns among the physical traits of baseball players and
- identifies their likelihood of winning the season's Most Valuable Player (MVP) award.
    - here, the machine assesses the physical attributes of previous baseball MVPs among other features such as age and education.
- Here, this model is different from the one in the previous slide.....and hence produces different output. But, at no stage the decision model is programmed to produce the two required outcomes.
- the model only decodes the complex patterns in the input data, and uses machine learning to find connections without human help.
- this also means that a related dataset collected from another time period, with fewer or greater data points, might push the model to produce a slightly different output.

ML: Improving predictions based on experience

- ML also has the ability to improve predictions based on experience; like mimicking the way humans base their decisions on experience.

ML: Improving predictions based on experience

- ML also has the ability to improve predictions based on experience; like mimicking the way humans base their decisions on experience.
- For the purpose, ML utilizes exposure to data to improve its decision making.

ML: Improving predictions based on experience

- ML also has the ability to improve predictions based on experience; like mimicking the way humans base their decisions on experience.
- For the purpose, ML utilizes exposure to data to improve its decision making.
- The exposure to the data points

ML: Improving predictions based on experience

- ML also has the ability to improve predictions based on experience; like mimicking the way humans base their decisions on experience.
- For the purpose, ML utilizes exposure to data to improve its decision making.
- The exposure to the data points
  - provides experience and enables the model to familiarize itself with patterns in the data.

ML: Improving predictions based on experience

- ML also has the ability to improve predictions based on experience; like mimicking the way humans base their decisions on experience.
- For the purpose, ML utilizes exposure to data to improve its decision making.
- The exposure to the data points
  - provides experience and enables the model to familiarize itself with patterns in the data.
  - deepens the model's understanding of patterns, including the significance of changes in the data,

ML: Improving predictions based on experience

- ML also has the ability to improve predictions based on experience; like mimicking the way humans base their decisions on experience.
- For the purpose, ML utilizes exposure to data to improve its decision making.
- The exposure to the data points
  - provides experience and enables the model to familiarize itself with patterns in the data.
  - deepens the model's understanding of patterns, including the significance of changes in the data,
  - and based on the same, allows constructing an effective self-learning model.

ML: Improving predictions based on experience

- ML also has the ability to improve predictions based on experience; like mimicking the way humans base their decisions on experience.
- For the purpose, ML utilizes exposure to data to improve its decision making.
- The exposure to the data points
  - provides experience and enables the model to familiarize itself with patterns in the data.
  - deepens the model's understanding of patterns, including the significance of changes in the data,
  - and based on the same, allows constructing an effective self-learning model.
- A common example of self learning is a system for detecting spam email messages.

ML: Improving predictions based on experience

- ML also has the ability to improve predictions based on experience; like mimicking the way humans base their decisions on experience.
- For the purpose, ML utilizes exposure to data to improve its decision making.
- The exposure to the data points
  - provides experience and enables the model to familiarize itself with patterns in the data.
  - deepens the model's understanding of patterns, including the significance of changes in the data,
  - and based on the same, allows constructing an effective self-learning model.
- A common example of self learning is a system for detecting spam email messages.
- Let us primarily investigate the scenario.....

Another Example: An ML model for detecting spam email messages.

- Initial data used to develop a model - model learns to flag emails as spams

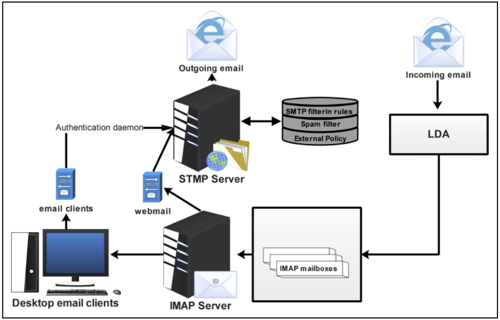

Figure: Spam Mail Detection using ML[a]

[a] Eman M.Bahgat et al

Another Example: An ML model for detecting spam email messages.

- Initial data used to develop a model - model learns to flag emails as spams
- this is based on the identified suspicious subject lines and body text containing keywords from the mails flagged by users as spam n the past.

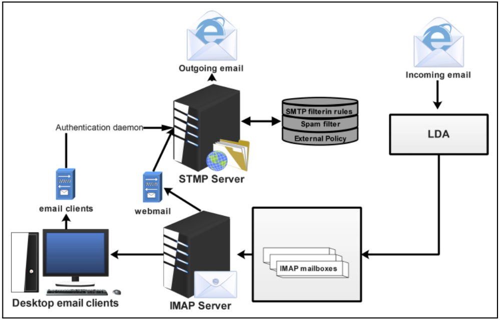

Figure: Spam Mail Detection using ML[a]

[a] Eman M.Bahgat et al

Another Example: An ML model for detecting spam email messages.

- Initial data used to develop a model - model learns to flag emails as spams
- this is based on the identified suspicious subject lines and body text containing keywords from the mails flagged by users as spam n the past.
  - e.g. words like dear friend, free, invoice, PayPal, Viagra, casino, payment, bankruptcy, and winner

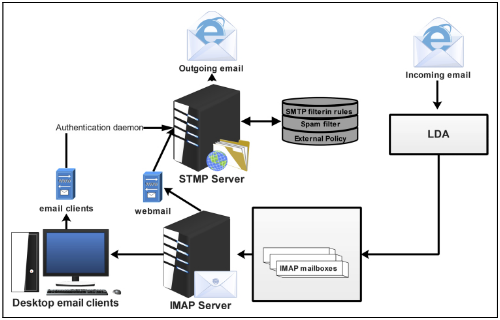

Figure: Spam Mail Detection using ML[a]

[a] Eman M.Bahgat et al

Another Example: An ML model for detecting spam email messages.

- Initial data used to develop a model - model learns to flag emails as spams
- this is based on the identified suspicious subject lines and body text containing keywords from the mails flagged by users as spam n the past.
  - e.g. words like dear friend, free, invoice, PayPal, Viagra, casino, payment, bankruptcy, and winner
- however, as more data is analyzed, the model might also find exceptions and incorrect assumptions

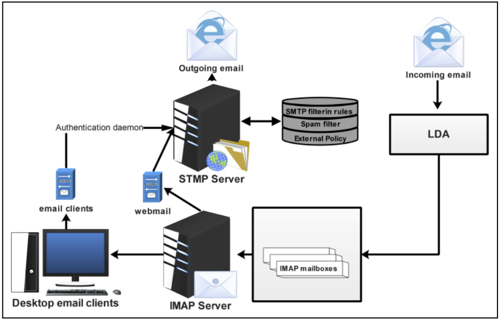

Figure: Spam Mail Detection using ML[a]

[a] Eman M.Bahgat et al

Another Example: An ML model for detecting spam email messages.

- Initial data used to develop a model - model learns to flag emails as spams
- this is based on the identified suspicious subject lines and body text containing keywords from the mails flagged by users as spam n the past.
  - e.g. words like dear friend, free, invoice, PayPal, Viagra, casino, payment, bankruptcy, and winner
- however, as more data is analyzed, the model might also find exceptions and incorrect assumptions
- this could render the model susceptible to bad predictions.

.....*continued*

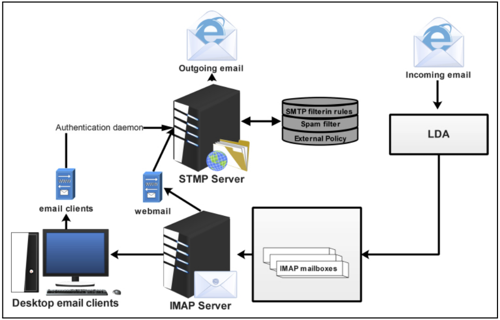

Figure: Spam Mail Detection using ML[a]

[a] Eman M.Bahgat et al

Another Example: An ML model for detecting spam email messages...

- The false positives (or even the false negatives) predicted by the model

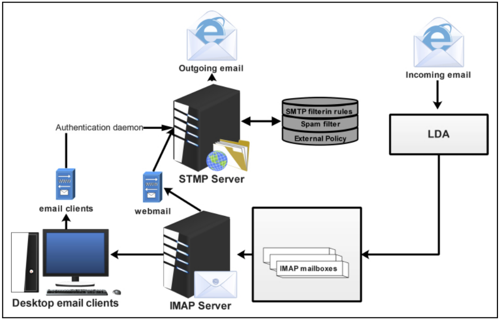

Figure: Spam Mail Detection using ML[a]

[a] Eman M.Bahgat et al

- The false positives (or even the false negatives) predicted by the model
  - depend on the quality and the quantity of the data supplied during the training

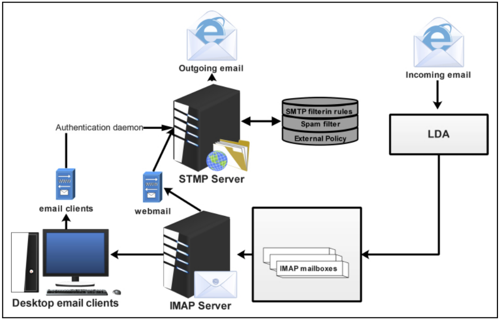

Figure: Spam Mail Detection using ML[a]

[a] Eman M.Bahgat et al

Another Example: An ML model for detecting spam email messages...

- The false positives (or even the false negatives) predicted by the model
  - depend on the quality and the quantity of the data supplied during the training
  - e.g. if there is limited data to reference its decision, the email subject viz. "PayPal has received your payment for Casino Royale purchased on eBay." might be wrongly classified as spam

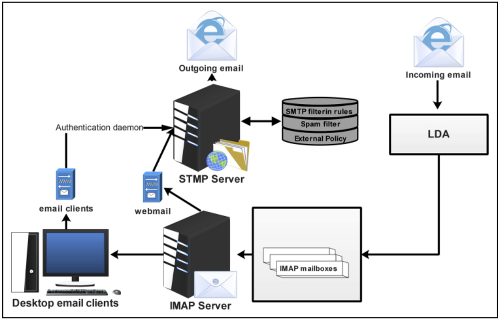

Figure: Spam Mail Detection using ML[a]

[a] Eman M.Bahgat et al

Another Example: An ML model for detecting spam email messages...

- The false positives (or even the false negatives) predicted by the model
  - depend on the quality and the quantity of the data supplied during the training
  - e.g. if there is limited data to reference its decision, the email subject viz. "PayPal has received your payment for Casino Royale purchased on eBay." might be wrongly classified as spam
- traditional programming is highly susceptible to this problem

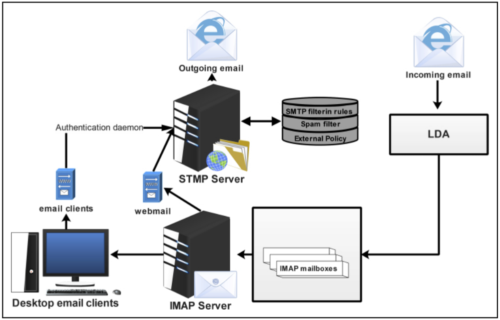

Figure: Spam Mail Detection using ML[a]

[a] Eman M.Bahgat et al

Another Example: An ML model for detecting spam email messages...

- The false positives (or even the false negatives) predicted by the model
  - depend on the quality and the quantity of the data supplied during the training
  - e.g. if there is limited data to reference its decision, the email subject viz. "PayPal has received your payment for Casino Royale purchased on eBay." might be wrongly classified as spam
- traditional programming is highly susceptible to this problem
  - this is due to the rigidly defined pre-set rules.

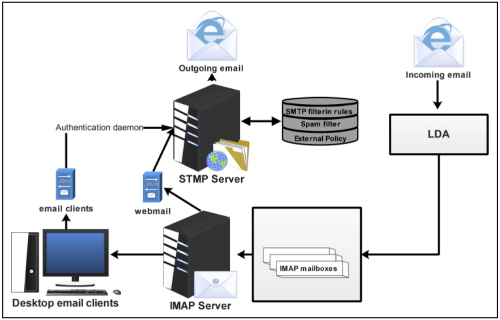

Figure: Spam Mail Detection using ML[a]

[a] Eman M.Bahgat et al

- ML, however, with exposure to data as a way to refine the model, adjusts weak assumptions,

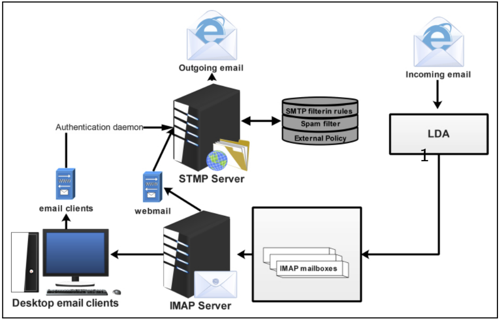

Figure: Spam Mail Detection using ML

Another Example: An ML model for detecting spam email messages...

- ML, however, with exposure to data as a way to refine the model, adjusts weak assumptions,
- ML thereby, responds appropriately to unique data points such as the scenario just described.

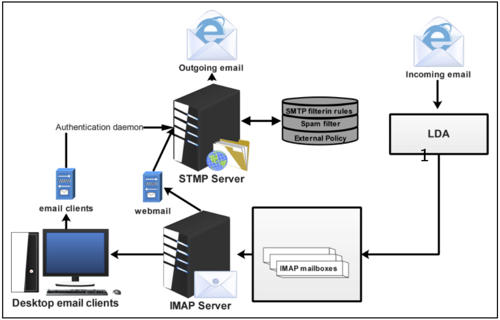

Figure: Spam Mail Detection using ML

Another Example: An ML model for detecting spam email messages...

- ML, however, with exposure to data as a way to refine the model, adjusts weak assumptions,
- ML thereby, responds appropriately to unique data points such as the scenario just described.
- Does more data lead to better predictions ?

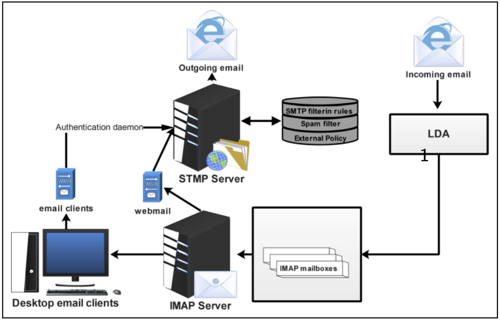

Figure: Spam Mail Detection using ML

- ML, however, with exposure to data as a way to refine the model, adjusts weak assumptions,
- ML thereby, responds appropriately to unique data points such as the scenario just described.
- Does more data lead to better predictions ?
- No

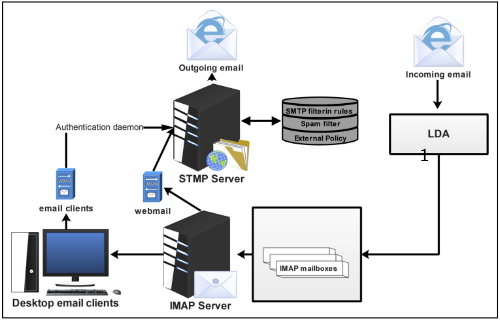

Figure: Spam Mail Detection using ML

- ML, however, with exposure to data as a way to refine the model, adjusts weak assumptions,
- ML thereby, responds appropriately to unique data points such as the scenario just described.
- Does more data lead to better predictions ?
- No
- while data is used to source the self-learning process, more data do not always equate to better decisions;

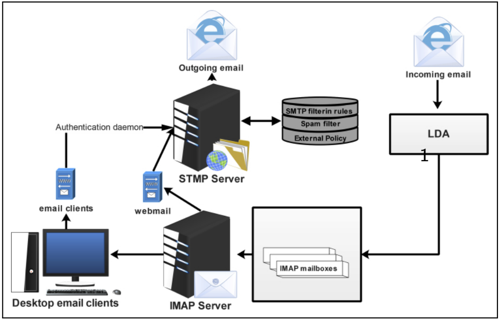

Figure: Spam Mail Detection using ML

ML tasks: more data do not always better decisions

- While data is used to source the self-learning process, more data do not always equate to better decisions;

ML tasks: more data do not always better decisions

- While data is used to source the self-learning process, more data do not always equate to better decisions;
- the input data must be relevant, to realize the objective.

ML tasks: more data do not always better decisions

- While data is used to source the self-learning process, more data do not always equate to better decisions;
- the input data must be relevant, to realize the objective.
- in Data and Goliath: The Hidden Battles to Collect Your Data and Control Your World, Bruce Schneir writes that,

ML tasks: more data do not always better decisions

- While data is used to source the self-learning process, more data do not always equate to better decisions;
- the input data must be relevant, to realize the objective.
- in Data and Goliath: The Hidden Battles to Collect Your Data and Control Your World, Bruce Schneir writes that,
  - “When looking for the needle, the last thing you want to do is pile lots more hay on it.”

ML tasks: more data do not always better decisions

- While data is used to source the self-learning process, more data do not always equate to better decisions;
- the input data must be relevant, to realize the objective.
- in Data and Goliath: The Hidden Battles to Collect Your Data and Control Your World, Bruce Schneir writes that,
  - "When looking for the needle, the last thing you want to do is pile lots more hay on it."
- that is, adding irrelevant data can be counter-productive to achieving a desired result.

ML tasks: more data do not always better decisions

- While data is used to source the self-learning process, more data do not always equate to better decisions;
- the input data must be relevant, to realize the objective.
- in Data and Goliath: The Hidden Battles to Collect Your Data and Control Your World, Bruce Schneir writes that,
  - "When looking for the needle, the last thing you want to do is pile lots more hay on it."
- that is, adding irrelevant data can be counter-productive to achieving a desired result.
- In addition, the amount of input data should be compatible with the processing resources and the available time.

ML dataset: Training Data & Test Data

In ML, the input data is typically split into training data and test data.

- Training data

ML dataset: Training Data & Test Data

In ML, the input data is typically split into training data and test data.

- Training data
  - it is the initial reserve of data used to develop the model and the subset of original dataset - fed into the ML model to discover, learn patterns.

ML dataset: Training Data & Test Data

In ML, the input data is typically split into training data and test data.

- Training data
  - it is the initial reserve of data used to develop the model and the subset of original dataset - fed into the ML model to discover, learn patterns.
  - is generally larger in size compared to the testing dataset.

ML dataset: Training Data & Test Data

In ML, the input data is typically split into training data and test data.

- Training data
  - it is the initial reserve of data used to develop the model and the subset of original dataset - fed into the ML model to discover, learn patterns.
  - is generally larger in size compared to the testing dataset.
  - is well known to the model, as it is used to train the model

ML dataset: Training Data & Test Data

In ML, the input data is typically split into training data and test data.

- Training data
  - it is the initial reserve of data used to develop the model and the subset of original dataset - fed into the ML model to discover, learn patterns.
  - is generally larger in size compared to the testing dataset.
  - is well known to the model, as it is used to train the model
  - e.g. in the spam email detection example, false-positives similar to the PayPal auto-response message - discussed earlier - might be detected from the training data

ML dataset: Training Data & Test Data

In ML, the input data is typically split into training data and test data.

- Training data
  - it is the initial reserve of data used to develop the model and the subset of original dataset - fed into the ML model to discover, learn patterns.
  - is generally larger in size compared to the testing dataset.
  - is well known to the model, as it is used to train the model
  - e.g. in the spam email detection example, false-positives similar to the PayPal auto-response message - discussed earlier - might be detected from the training data
    - then, modifications must then be made to the model, e.g., email notifications issued from the sending address "payments@paypal.com" should be excluded from spam filtering.

ML dataset: Training Data & Test Data

In ML, the input data is typically split into training data and test data.

- Training data
  - it is the initial reserve of data used to develop the model and the subset of original dataset - fed into the ML model to discover, learn patterns.
  - is generally larger in size compared to the testing dataset.
  - is well known to the model, as it is used to train the model
  - e.g. in the spam email detection example, false-positives similar to the PayPal auto-response message - discussed earlier - might be detected from the training data
    - then, modifications must then be made to the model, e.g., email notifications issued from the sending address "payments@paypal.com" should be excluded from spam filtering.
    - thus, the model can then be trained to automatically detect these errors (by analyzing historical examples of spam messages and deciphering their patterns) without direct human interference.

ML dataset: Training Data & Test Data

In ML, the input data is typically split into training data and test data.

- Training data
  - it is the initial reserve of data used to develop the model and the subset of original dataset - fed into the ML model to discover, learn patterns.
  - is generally larger in size compared to the testing dataset.
  - is well known to the model, as it is used to train the model
  - e.g. in the spam email detection example, false-positives similar to the PayPal auto-response message - discussed earlier - might be detected from the training data
    - then, modifications must then be made to the model, e.g., email notifications issued from the sending address "payments@paypal.com" should be excluded from spam filtering.
    - thus, the model can then be trained to automatically detect these errors (by analyzing historical examples of spam messages and deciphering their patterns) without direct human interference.
    - subsequently, after developing the model based on patterns extracted from the training data one can test the model on the remaining data, known as the test data.

- Testing data

ML dataset: Training Data & Test Data

- Testing data
  - is used to check the accuracy of the model.

ML dataset: Training Data & Test Data

- Testing data
  - is used to check the accuracy of the model.
  - is the unseen data used to test the ML model.

- Testing data
  - is used to check the accuracy of the model.
  - is the unseen data used to test the ML model.
  - thus, is used t to evaluate the performance and progress of the training of ML algorithms and adjust or optimize it for improved results.

ML dataset: Training Data & Test Data

- Testing data
  - is used to check the accuracy of the model.
  - is the unseen data used to test the ML model.
  - thus, is used t to evaluate the performance and progress of the training of ML algorithms and adjust or optimize it for improved results.
  - must represent the actual dataset.

ML dataset: Training Data & Test Data

- Testing data
  - is used to check the accuracy of the model.
  - is the unseen data used to test the ML model.
  - thus, is used t to evaluate the performance and progress of the training of ML algorithms and adjust or optimize it for improved results.
  - must represent the actual dataset.
  - must be large enough to generate meaningful predictions

ML dataset: Training Data & Test Data

- Testing data
  - is used to check the accuracy of the model.
  - is the unseen data used to test the ML model.
  - thus, is used t to evaluate the performance and progress of the training of ML algorithms and adjust or optimize it for improved results.
  - must represent the actual dataset.
  - must be large enough to generate meaningful predictions
  - thus the model already "knows" the training data - but how it performs on new test data will lead to know if it's working accurately or if it requires more training data to perform

ML dataset: Training Data & Test Data

- Testing data
  - is used to check the accuracy of the model.
  - is the unseen data used to test the ML model.
  - thus, is used t to evaluate the performance and progress of the training of ML algorithms and adjust or optimize it for improved results.
  - must represent the actual dataset.
  - must be large enough to generate meaningful predictions
  - thus the model already "knows" the training data - but how it performs on new test data will lead to know if it's working accurately or if it requires more training data to perform
  - hence, test data provides a final, real-world check of an unseen dataset to confirm that the machine learning algorithm was trained effectively.

ML dataset: Training Data & Test Data

- Testing data
  - is used to check the accuracy of the model.
  - is the unseen data used to test the ML model.
  - thus, is used t to evaluate the performance and progress of the training of ML algorithms and adjust or optimize it for improved results.
  - must represent the actual dataset.
  - must be large enough to generate meaningful predictions
  - thus the model already "knows" the training data - but how it performs on new test data will lead to know if it's working accurately or if it requires more training data to perform
  - hence, test data provides a final, real-world check of an unseen dataset to confirm that the machine learning algorithm was trained effectively.
- normally, there is a split of 80% for training and 20% for testing dataset.

ML dataset: Training, Validation & Testing Data Sets

- Very importantly note that model performance depends on how the dataset are splitted in the model building.

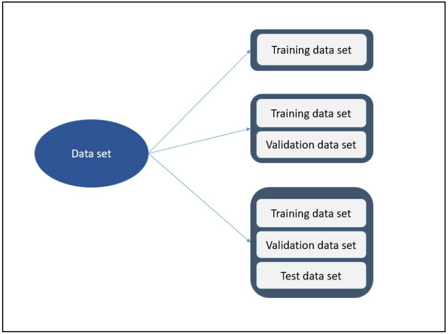

Figure: Three Splits of ML dataset

ML dataset: Training, Validation & Testing Data Sets

- Very importantly note that model performance depends on how the dataset are splitted in the model building.
- Hence, sometimes the dataset is viewed to be splitted into

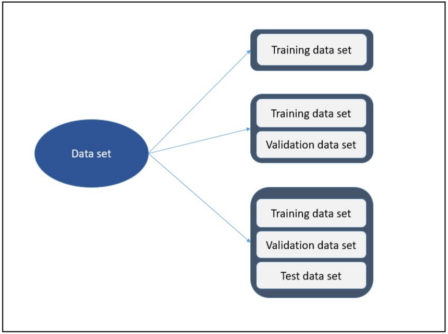

Figure: Three Splits of ML dataset

ML dataset: Training, Validation & Testing Data Sets

- Very importantly note that model performance depends on how the dataset are splitted in the model building.
- Hence, sometimes the dataset is viewed to be splitted into
  - training data,

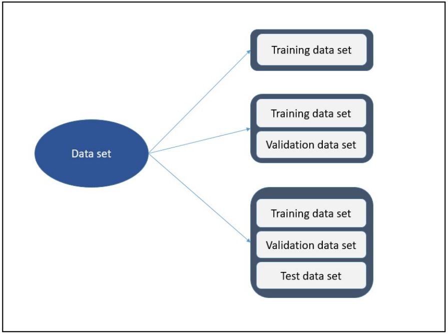

Figure: Three Splits of ML dataset

ML dataset: Training, Validation & Testing Data Sets

- Very importantly note that model performance depends on how the dataset are splitted in the model building.
- Hence, sometimes the dataset is viewed to be splitted into
  - training data,
  - test data and

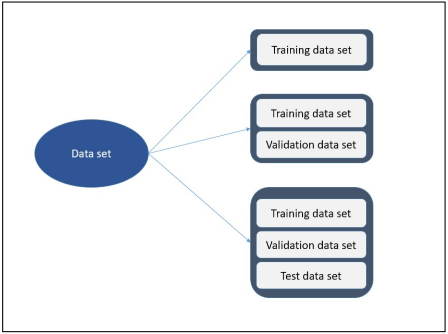

Figure: Three Splits of ML dataset

ML dataset: Training, Validation & Testing Data Sets

- Very importantly note that model performance depends on how the dataset are splitted in the model building.
- Hence, sometimes the dataset is viewed to be splitted into
  - training data,
  - test data and
  - validation data.

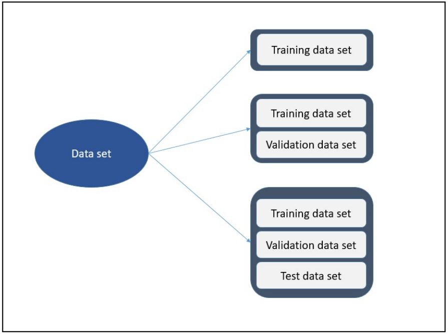

Figure: Three Splits of ML dataset

ML dataset: Training, Validation & Testing Data Sets

- Very importantly note that model performance depends on how the dataset are splitted in the model building.
- Hence, sometimes the dataset is viewed to be splitted into
  - training data,
  - test data and
  - validation data.
- Note that all three are typically split from one large dataset....

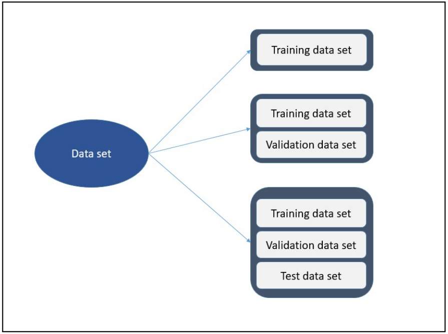

Figure: Three Splits of ML dataset

ML dataset: Training, Validation & Testing Data Sets

- Very importantly note that model performance depends on how the dataset are splitted in the model building.
- Hence, sometimes the dataset is viewed to be splitted into
    - training data,
    - test data and
    - validation data.
- Note that all three are typically split from one large dataset....
- However, each one typically has its own distinct use in ML modeling.

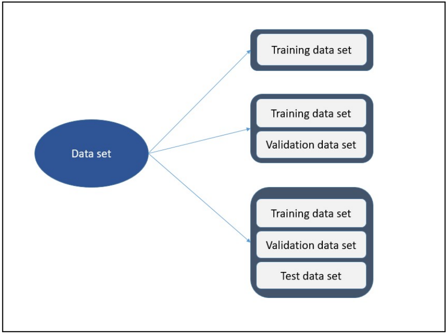

Figure: Three Splits of ML dataset

ML dataset: Training, Validation & Testing Data Sets

- Very importantly note that model performance depends on how the dataset are splitted in the model building.
- Hence, sometimes the dataset is viewed to be splitted into
  - training data,
  - test data and
  - validation data.
- Note that all three are typically split from one large dataset....
- However, each one typically has its own distinct use in ML modeling.
- But first, reviewing the meaning/semantics of each dataset (again).....

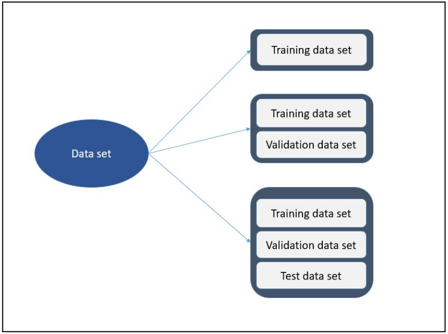

Figure: Three Splits of ML dataset

ML dataset: Training, Validation & Testing Data Sets...

Training dataset

- As seen already, this type of data builds up the machine learning algorithm.

ML dataset: Training, Validation & Testing Data Sets...

Training dataset

- As seen already, this type of data builds up the machine learning algorithm.
- The data scientist feeds the algorithm input data, which corresponds to an expected output.

ML dataset: Training, Validation & Testing Data Sets...

Training dataset

- As seen already, this type of data builds up the machine learning algorithm.
- The data scientist feeds the algorithm input data, which corresponds to an expected output.
- The model evaluates the data repeatedly to learn more about the data's behavior

ML dataset: Training, Validation & Testing Data Sets...

Training dataset

- As seen already, this type of data builds up the machine learning algorithm.
- The data scientist feeds the algorithm input data, which corresponds to an expected output.
- The model evaluates the data repeatedly to learn more about the data's behavior
- And then the model adjusts itself to serve its intended purpose.

ML dataset: Training, Validation & Testing Data Sets...

Training dataset

- As seen already, this type of data builds up the machine learning algorithm.
- The data scientist feeds the algorithm input data, which corresponds to an expected output.
- The model evaluates the data repeatedly to learn more about the data's behavior
- And then the model adjusts itself to serve its intended purpose.

ML dataset: Training, Validation & Testing Data Sets...

Training dataset

- As seen already, this type of data builds up the machine learning algorithm.
- The data scientist feeds the algorithm input data, which corresponds to an expected output.
- The model evaluates the data repeatedly to learn more about the data's behavior
- And then the model adjusts itself to serve its intended purpose.

Testing dataset

- After the model is built, testing data once again validates that it can make accurate predictions.

ML dataset: Training, Validation & Testing Data Sets...

Training dataset

- As seen already, this type of data builds up the machine learning algorithm.
- The data scientist feeds the algorithm input data, which corresponds to an expected output.
- The model evaluates the data repeatedly to learn more about the data's behavior
- And then the model adjusts itself to serve its intended purpose.

Testing dataset

- After the model is built, testing data once again validates that it can make accurate predictions.
- if training and validation data include labels to monitor performance metrics of the model, the testing data should be unlabeled.

ML dataset: Training, Validation & Testing Data Sets...

Training dataset

- As seen already, this type of data builds up the machine learning algorithm.
- The data scientist feeds the algorithm input data, which corresponds to an expected output.
- The model evaluates the data repeatedly to learn more about the data's behavior
- And then the model adjusts itself to serve its intended purpose.

Testing dataset

- After the model is built, testing data once again validates that it can make accurate predictions.
- if training and validation data include labels to monitor performance metrics of the model, the testing data should be unlabeled.
- Test data provides a final, real-world check of an unseen dataset to confirm that the ML algorithm was trained effectively.

Validation dataset

- during training, validation data infuses new data into the model that it hasn't evaluated before.

ML dataset: Training, Validation & Testing Data Sets...

Validation dataset

- during training, validation data infuses new data into the model that it hasn't evaluated before.
- provides the first test against unseen data, allowing data scientists to evaluate how well the model makes predictions based on the new data.

Validation dataset

- during training, validation data infuses new data into the model that it hasn't evaluated before.
- provides the first test against unseen data, allowing data scientists to evaluate how well the model makes predictions based on the new data.
- not all data scientists use validation data, but it can provide some helpful information to optimize hyperparameters, which influence how the model assesses data.

ML dataset: Implications of using differing datasets

Model built using just training data set

- gets highly biased to the dataset.

ML dataset: Implications of using differing datasets

Model built using just training data set

- gets highly biased to the dataset.
- most likely won't be able to generalize on unseen data, unless the dataset used for training represented the entire population.

ML dataset: Implications of using differing datasets

Model built using just training data set

- gets highly biased to the dataset.
- most likely won't be able to generalize on unseen data, unless the dataset used for training represented the entire population.
- thus, overfits the training dataset.

ML dataset: Implications of using differing datasets

Model built using just training data set

- gets highly biased to the dataset.
- most likely won't be able to generalize on unseen data, unless the dataset used for training represented the entire population.
- thus, overfits the training dataset.

ML dataset: Implications of using differing datasets

Model built using just training data set

- gets highly biased to the dataset.
- most likely won't be able to generalize on unseen data, unless the dataset used for training represented the entire population.
- thus, overfits the training dataset.

Model built with training & validation data set:

- when evaluated on validation dataset, the model performs much better than the earlier model trained using entire dataset.

ML dataset: Implications of using differing datasets

Model built using just training data set

- gets highly biased to the dataset.
- most likely won't be able to generalize on unseen data, unless the dataset used for training represented the entire population.
- thus, overfits the training dataset.

Model built with training & validation data set:

- when evaluated on validation dataset, the model performs much better than the earlier model trained using entire dataset.
- however, when trained for long time, the model gets biased.

ML dataset: Implications of using differing datasets

Model built using just training data set

- gets highly biased to the dataset.
- most likely won't be able to generalize on unseen data, unless the dataset used for training represented the entire population.
- thus, overfits the training dataset.

Model built with training & validation data set:

- when evaluated on validation dataset, the model performs much better than the earlier model trained using entire dataset.
- however, when trained for long time, the model gets biased.
- basically, the hyperparameters get changed appropriately in each iteration such that model performs better with validation data set.

ML dataset: Implications of using differing datasets

Model built using just training data set

- gets highly biased to the dataset.
- most likely won't be able to generalize on unseen data, unless the dataset used for training represented the entire population.
- thus, overfits the training dataset.

Model built with training & validation data set:

- when evaluated on validation dataset, the model performs much better than the earlier model trained using entire dataset.
- however, when trained for long time, the model gets biased.
- basically, the hyperparameters get changed appropriately in each iteration such that model performs better with validation data set.
- Thus, the validation dataset details get leaked into training dataset.

Model built with training, validation & test data set:

- uses the third dataset split from the original dataset which is kept hidden from training and evaluation process.

Model built with training, validation & test data set:

- uses the third dataset split from the original dataset which is kept hidden from training and evaluation process.
- thus, have a greater likelihood of generalizing on unseen dataset than earlier two cases mentioned above.

Categories of ML methods

Categories of Machine Learning methods/mechanisms

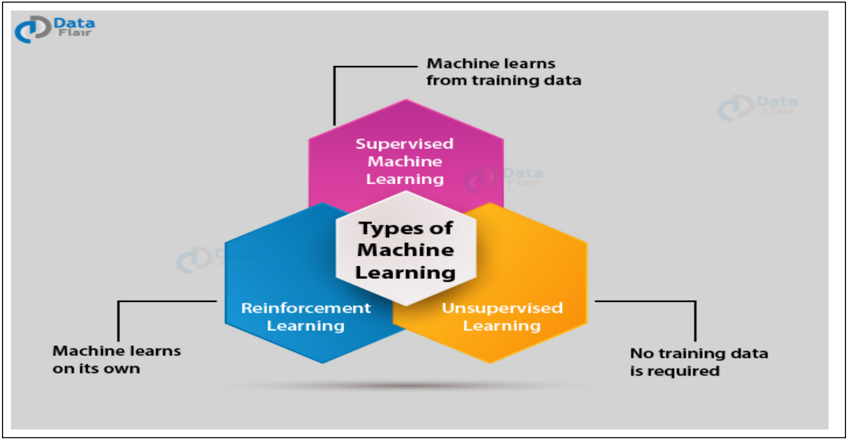

Figure: Machine Learning Techniques [2]

ML methods based on training patterns

- ML methods - training patterns - classifier model.

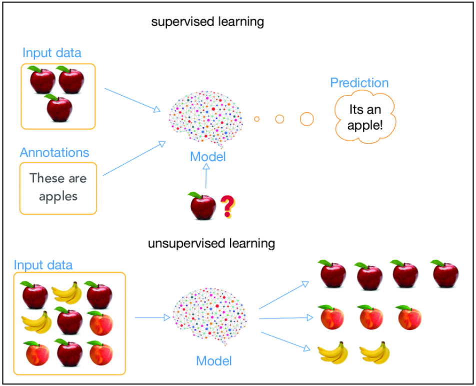

Figure: Supervised and Unsupervised Learning

ML methods based on training patterns

- ML methods - training patterns - classifier model.
- Classifier model can be parametric or non-parametric

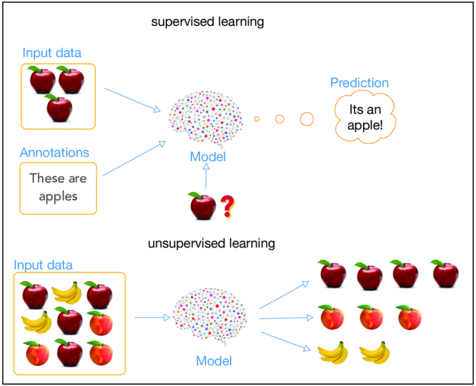

Figure: Supervised and Unsupervised Learning

ML methods based on training patterns

- ML methods - training patterns - classifier model.
- Classifier model can be parametric or non-parametric
- The goal of using ML algorithms is to reduce the classification error on the given training sample data.

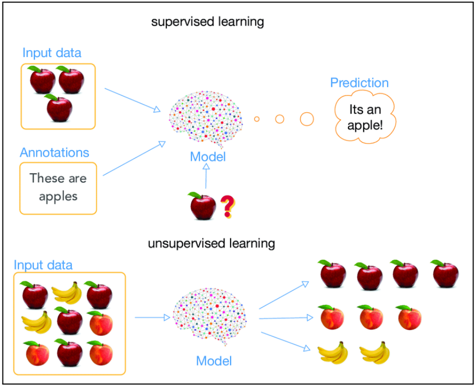

Figure: Supervised and Unsupervised Learning

ML methods based on training patterns

- ML methods - training patterns - classifier model.
- Classifier model can be parametric or non-parametric
- The goal of using ML algorithms is to reduce the classification error on the given training sample data.
- ML algorithms are categorized into supervised learning and unsupervised learning based on

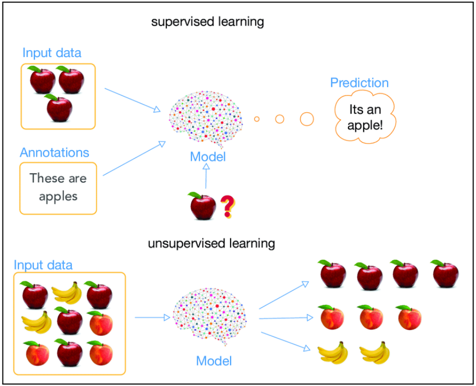

Figure: Supervised and Unsupervised Learning

ML methods based on training patterns

- ML methods - training patterns - classifier model.
- Classifier model can be parametric or non-parametric
- The goal of using ML algorithms is to reduce the classification error on the given training sample data.
- ML algorithms are categorized into supervised learning and unsupervised learning based on
  - how ML algorithms treat the input and output variables.

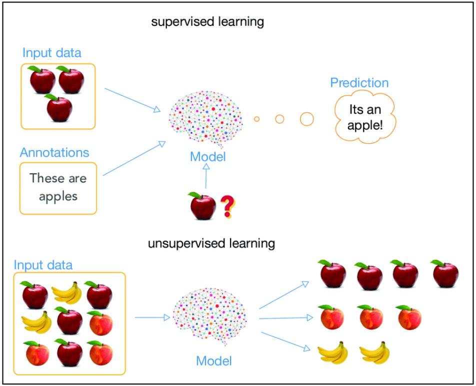

Figure: Supervised and Unsupervised Learning

ML methods based on training patterns

- ML methods - training patterns - classifier model.
- Classifier model can be parametric or non-parametric
- The goal of using ML algorithms is to reduce the classification error on the given training sample data.
- ML algorithms are categorized into supervised learning and unsupervised learning based on
  - how ML algorithms treat the input and output variables.
  - on the availability of training data and the desired outcome of the learning algorithms.

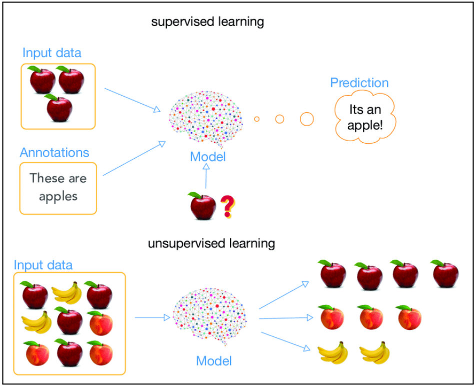

Figure: Supervised and Unsupervised Learning

Supervised Learning

Supervised learning methods

- imitates our own ability to extract patterns from known examples and use that extracted insight to engineer a repeatable outcome.

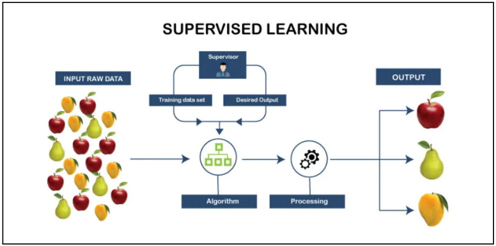

Figure: Supervised Learning

Supervised learning methods

- imitates our own ability to extract patterns from known examples and use that extracted insight to engineer a repeatable outcome.
- the process of understanding a known input-output combination to learn the underlying patterns is the focus here.

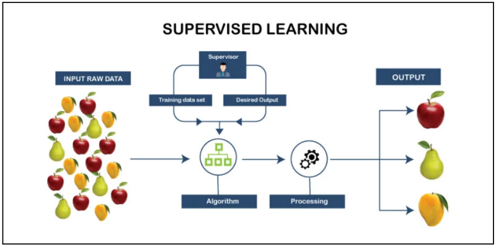

Figure: Supervised Learning

Supervised learning methods

- imitates our own ability to extract patterns from known examples and use that extracted insight to engineer a repeatable outcome.
- the process of understanding a known input-output combination to learn the underlying patterns is the focus here.
- the supervised learning ML model analyzes and deciphers the relationship between input and output data to learn the underlying patterns.

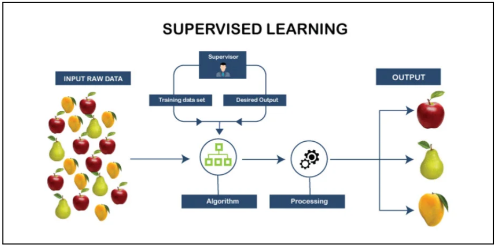

Figure: Supervised Learning

Supervised Learning...

Supervised learning methods

- The example of how Toyota designed their first car prototype from the Chevrolet car.

	Input	Input	Input	Output
	Card Brand	**Mileage (km)**	**Year of Make**	**Price (USD)**
Car 1	Lexus	51715	2012	15985
Car 2	Lexus	7980	2013	19600
Car 3	Lexus	82497	2012	14095
Car 4	Lexus	85199	2011	12490
Car 5	Audi	62948	2008	13985

Table 2: Extract of a used car dataset

Figure: Supervised Learning

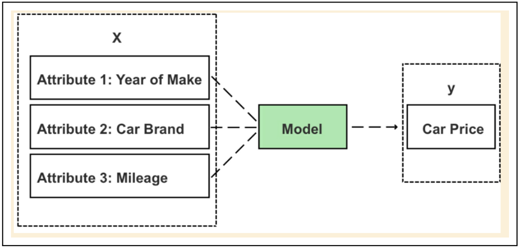

Figure: Supervised Learning

Supervised Learning...

Supervised learning methods

- The example of how Toyota designed their first car prototype from the Chevrolet car.
- created its first vehicle prototype after taking apart a Chevrolet car and unlocking its design - which was kept secret by Chevrolet in America

	Input	Input	Input	Output
	Card Brand	**Mileage (km)**	**Year of Make**	**Price (USD)**
Car 1	Lexus	51715	2012	15985
Car 2	Lexus	7980	2013	19600
Car 3	Lexus	82497	2012	14095
Car 4	Lexus	85199	2011	12490
Car 5	Audi	62948	2008	13985

Table 2: Extract of a used car dataset

Figure: Supervised Learning

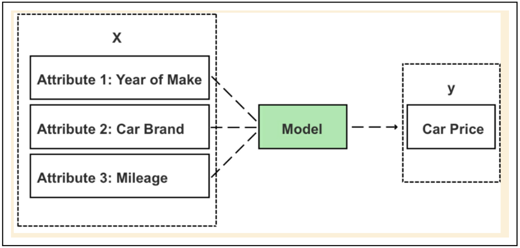

Figure: Supervised Learning

Supervised learning methods

- The example of how Toyota designed their first car prototype from the Chevrolet car.
- created its first vehicle prototype after taking apart a Chevrolet car and unlocking its design - which was kept secret by Chevrolet in America
- this process of understanding a known input-output combination is what is seen in supervised learning.

	Input Card Brand	Input Mileage (km)	Input Year of Make	Output Price (USD)
Car 1	Lexus	51715	2012	15985
Car 2	Lexus	7980	2013	19600
Car 3	Lexus	82497	2012	14095
Car 4	Lexus	85199	2011	12490
Car 5	Audi	62948	2008	13985

Table 2: Extract of a used car dataset

Figure: Supervised Learning

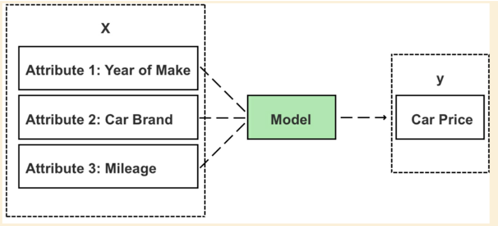

Figure: Supervised Learning

Supervised learning methods

- The example of how Toyota designed their first car prototype from the Chevrolet car.
- created its first vehicle prototype after taking apart a Chevrolet car and unlocking its design - which was kept secret by Chevrolet in America
- this process of understanding a known input-output combination is what is seen in supervised learning.
- the model analyzes and deciphers the relationship between input and output data to learn the underlying patterns.

	Input	Input	Input	Output
	Card Brand	Mileage (km)	Year of Make	Price (USD)
Car 1	Lexus	51715	2012	15985
Car 2	Lexus	7980	2013	19600
Car 3	Lexus	82497	2012	14095
Car 4	Lexus	85199	2011	12490
Car 5	Audi	62948	2008	13985

Table 2: Extract of a used car dataset

Figure: Supervised Learning

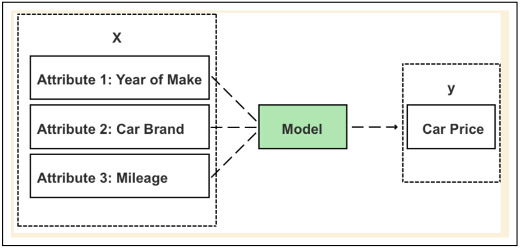

Figure: Supervised Learning

Supervised Learning...

Supervised learning methods

- The example of how Toyota designed their first car prototype from the Chevrolet car.
- created its first vehicle prototype after taking apart a Chevrolet car and unlocking its design - which was kept secret by Chevrolet in America
- this process of understanding a known input-output combination is what is seen in supervised learning.
- the model analyzes and deciphers the relationship between input and output data to learn the underlying patterns.
- Input data → independent variable (uppercase "X"), Output data → dependent variable (lowercase "y").

	Input	Input	Input	Output
	Card Brand	**Mileage (km)**	**Year of Make**	**Price (USD)**
Car 1	Lexus	51715	2012	15985
Car 2	Lexus	7980	2013	19600
Car 3	Lexus	82497	2012	14095
Car 4	Lexus	85199	2011	12490
Car 5	Audi	62948	2008	13985

Table 2: Extract of a used car dataset

Figure: Supervised Learning

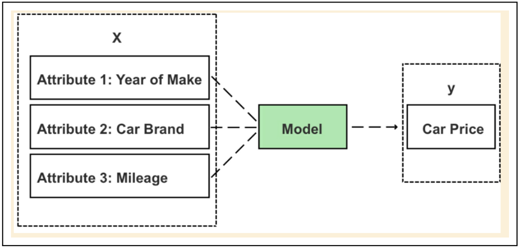

Figure: Supervised Learning

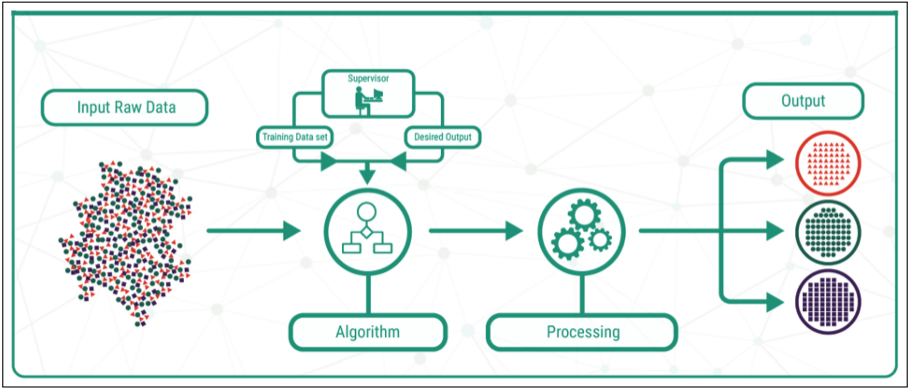

Figure: Supervised Learning

1

[1] https://www.crayondata.com/machine-learning-explained-understanding-supervised-unsupervised-and-reinforcement-learning/

Supervised Learning algorithms usecases...

The most common use cases of supervised learning are as follows:

- Spam detection - discussed before
- Bioinformatics
  - used for in storage of biological information of human beings that includes – fingertips, iris textures, eyes, swabs, and so on.
  - every time one wants to unlock your devices, it asks to authenticate either through fingertips or facial recognition.
- Object Recognitions
  - captcha - where one has to choose multiple images as per the instruction to get confirmed that one is a human.

Supervised Learning algorithms...

Supervised Learning algorithms

- are categorized based on the structures and objective functions of learning algorithms.
- are commonly characterized by the two types of problems viz. Classification and Regression
- Popular categorizations of the algorithms include
  - Linear and Logistic Regression
  - Artificial Neural Network (ANN),
  - Support Vector Machine (SVM), and
  - Decision trees.
- adopt a Bayesian approach to knowledge discovery, using probabilities of previously observed events to infer the probabilities of new events.

Supervised Learning algorithms...: Advantages

- are categorized based on the structures and objective functions of learning algorithms.
- permits one unmistakable with regards to the meaning of the marks/labels
- outcomes delivered by the directed strategy are more precise and dependable as compared to those of other procedures of AI.

Supervised Learning algorithms...: Disadvantages

- are categorized based on the structures and objective functions of learning algorithms. Hence
  - Computation time is vast for supervised learning.
  - Unwanted data downs efficiency - requires a ton of calculation time for preparing.
  - Pre-processing of data is no less than a big challenge.
  - Always in need of updates.
  - Anyone can overfit supervised algorithms easily.

Active Learning i.e. Smart Data Labeling with ML

- In ML, Data Labeling (DaL) is the process of identifying raw data (images, text files, videos, etc.) and

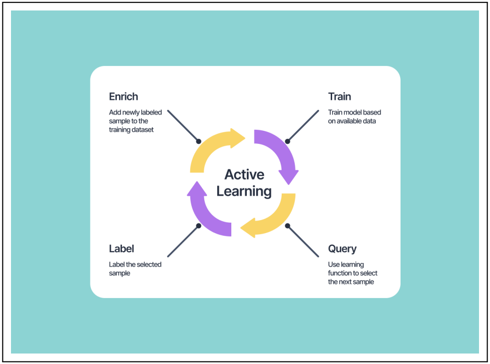

Figure: Active Learning for Smart Labeling

Active Learning i.e. Smart Data Labeling with ML

- In ML, Data Labeling (DaL) is the process of identifying raw data (images, text files, videos, etc.) and
  - adding one or more meaningful and informative labels to provide context so that an ML model can learn from it.

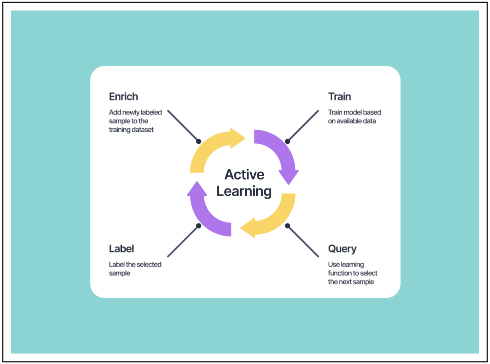

Figure: Active Learning for Smart Labeling

Active Learning i.e. Smart Data Labeling with ML

- In ML, Data Labeling (DaL) is the process of identifying raw data (images, text files, videos, etc.) and
  - adding one or more meaningful and informative labels to provide context so that an ML model can learn from it.
- In conventional DaL, label tags are attached to data points by a human who is an in-house labeler or outsourced personnel.

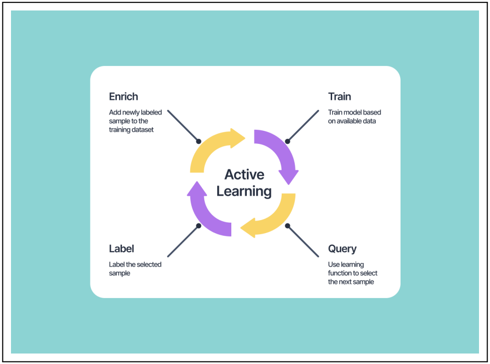

Figure: Active Learning for Smart Labeling

Active Learning i.e. Smart Data Labeling with ML

- In ML, Data Labeling (DaL) is the process of identifying raw data (images, text files, videos, etc.) and
  - adding one or more meaningful and informative labels to provide context so that an ML model can learn from it.
- In conventional DaL, label tags are attached to data points by a human who is an in-house labeler or outsourced personnel.
- However, considering the massive volume of data, manual labeling can be time-consuming, costly, and difficult to coordinate.

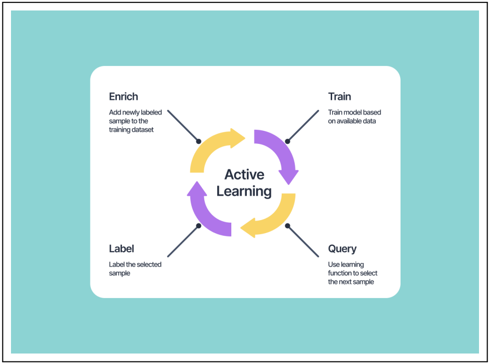

Figure: Active Learning for Smart Labeling

Massive volume of data discourages manual labeling of the data...

- Therefore, smart labeling or automatic labeling is employed

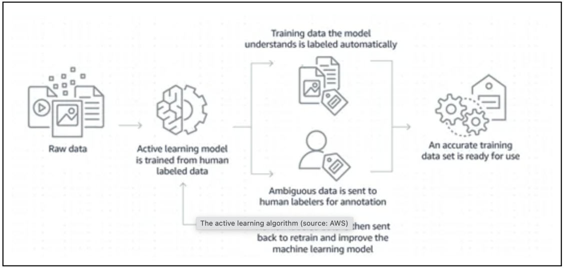

Figure: Smart Labeling

[Ref: https://aws.amazon.com/sagemaker/data-labeling/what-is-data-labeling/]

Massive volume of data discourages manual labeling of the data...

- Therefore, smart labeling or automatic labeling is employed
- here, a separate ML model can be trained to understand raw data

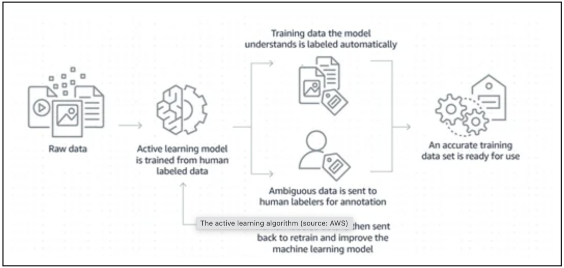

Figure: Smart Labeling

[Ref: https://aws.amazon.com/sagemaker/data-labeling/what-is-data-labeling/]

Massive volume of data discourages manual labeling of the data...

- Therefore, smart labeling or automatic labeling is employed
- here, a separate ML model can be trained to understand raw data
- and then, output appropriate label tags.

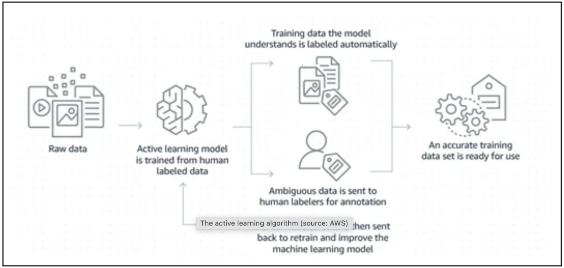

Figure: Smart Labeling

[Ref: https://aws.amazon.com/sagemaker/data-labeling/what-is-data-labeling/]

An ethical credit scoring system for banks and financial institutions

- Banking the unbanked i.e. developing credit rating for those who do not have a credit cards and hence no formal credit score.

An ethical credit scoring system for banks and financial institutions

- Banking the unbanked i.e. developing credit rating for those who do not have a credit cards and hence no formal credit score.
- In one of the implementations, transactions made by different account numbers, the region, mode of transaction, etc, the per capita income per area and the job title of the account numbers was used to develop such a system.

Supervised Learning algorithms...: Real world Applications...

An ethical credit scoring system for banks and financial institutions

- Banking the unbanked i.e. developing credit rating for those who do not have a credit cards and hence no formal credit score.
- In one of the implementations, transactions made by different account numbers, the region, mode of transaction, etc, the per capita income per area and the job title of the account numbers was used to develop such a system.

An ethical credit scoring system for banks and financial institutions

- Banking the unbanked i.e. developing credit rating for those who do not have a credit cards and hence no formal credit score.
- In one of the implementations, transactions made by different account numbers, the region, mode of transaction, etc, the per capita income per area and the job title of the account numbers was used to develop such a system.

Understanding Youth Sentiments Through Artificial Intelligence

- a real world application in which a Data Analysis pipeline was developed for sentiment analysis

An ethical credit scoring system for banks and financial institutions

- Banking the unbanked i.e. developing credit rating for those who do not have a credit cards and hence no formal credit score.
- In one of the implementations, transactions made by different account numbers, the region, mode of transaction, etc, the per capita income per area and the job title of the account numbers was used to develop such a system.

Understanding Youth Sentiments Through Artificial Intelligence

- a real world application in which a Data Analysis pipeline was developed for sentiment analysis
- this was to understand youth sentiments, analyzing aspirations, fears, and thoughts of the youth through scraping the web and youth-led media.

Medical applications

- an application was developed to anticipate patient danger (like the high-hazard patient etc.) or for foreseeing the likelihood of a congestive cardiovascular breakdown.

Medical applications

- an application was developed to anticipate patient danger (like the high-hazard patient etc.) or for foreseeing the likelihood of a congestive cardiovascular breakdown.

Medical applications

- an application was developed to anticipate patient danger (like the high-hazard patient etc.) or for foreseeing the likelihood of a congestive cardiovascular breakdown.

Public safety application

- a tool was built for analysing and classifying cases of sexual abuse in the workplace to identify patterns of such behaviors.

Unsupervised Learning methods

- here, one does not have to direct the model with pre-labeled input/output data.

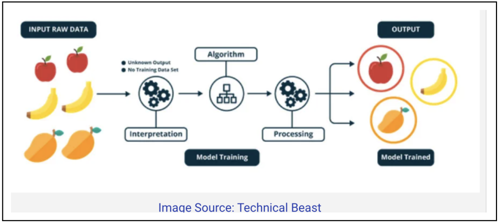

Figure: Un-Supervised Learning

Unsupervised Learning...

Unsupervised Learning methods

- here, one does not have to direct the model with pre-labeled input/output data.
- it permits the model to chip away at its own to find examples and data that was beforehand undetected.

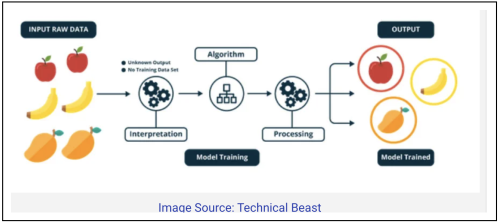

Figure: Un-Supervised Learning

Unsupervised Learning...

Unsupervised Learning methods

- here, one does not have to direct the model with pre-labeled input/output data.
- it permits the model to chip away at its own to find examples and data that was beforehand undetected.
- are the ones where no target or label of the data is given in sample data.

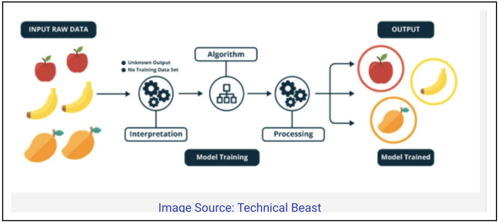

Figure: Un-Supervised Learning

Unsupervised Learning...

Unsupervised Learning methods

- here, one does not have to direct the model with pre-labeled input/output data.
- it permits the model to chip away at its own to find examples and data that was beforehand undetected.
- are the ones where no target or label of the data is given in sample data.
- are designed to summarize the key features of the data and to form the natural clusters of input patterns given a particular cost function.

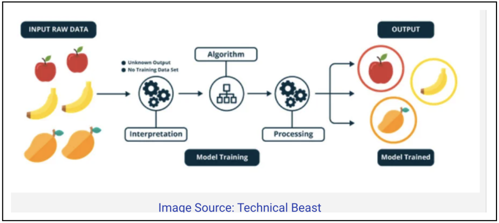

Figure: Un-Supervised Learning

Unsupervised Learning...

Unsupervised Learning methods

- thus, it draw abstractions from unlabeled datasets and apply these to new data.

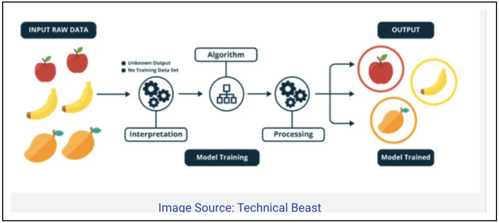

Figure: Un-Supervised Learning

Unsupervised Learning methods

- thus, it draw abstractions from unlabeled datasets and apply these to new data.
- the example methods include

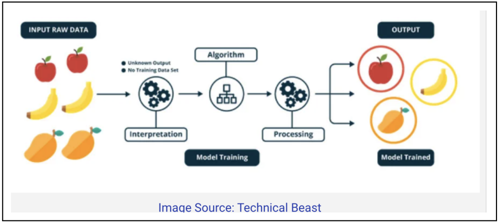

Figure: Un-Supervised Learning

Unsupervised Learning methods

- thus, it draw abstractions from unlabeled datasets and apply these to new data.
- the example methods include
  - Hierarchical, K-Means clustering.

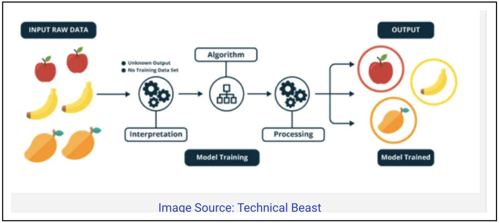

Figure: Un-Supervised Learning

Unsupervised Learning...

Unsupervised Learning methods

- thus, it draw abstractions from unlabeled datasets and apply these to new data.
- the example methods include
  - Hierarchical, K-Means clustering.
  - K-NN (k nearest neighbours).

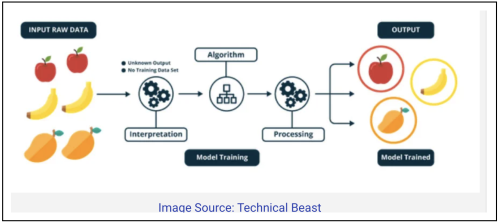

Figure: Un-Supervised Learning

Unsupervised Learning...

Unsupervised Learning methods

- thus, it draw abstractions from unlabeled datasets and apply these to new data.
- the example methods include
  - Hierarchical, K-Means clustering.
  - K-NN (k nearest neighbours).
  - Principal Component Analysis.

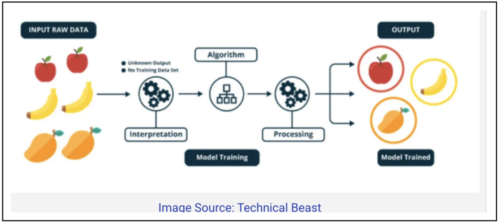

Figure: Un-Supervised Learning

Unsupervised Learning methods

- thus, it draw abstractions from unlabeled datasets and apply these to new data.
- the example methods include
    - Hierarchical, K-Means clustering.
    - K-NN (k nearest neighbours).
    - Principal Component Analysis.
    - Singular Value Decomposition.

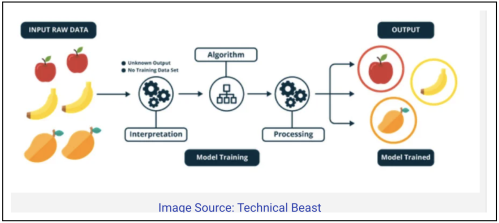

Figure: Un-Supervised Learning

Unsupervised Learning methods

- thus, it draw abstractions from unlabeled datasets and apply these to new data.
- the example methods include
  - Hierarchical, K-Means clustering.
  - K-NN (k nearest neighbours).
  - Principal Component Analysis.
  - Singular Value Decomposition.
  - Independent Component Analysis.

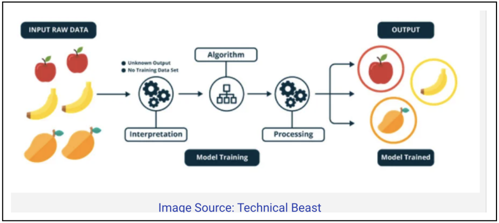

Figure: Un-Supervised Learning

Unsupervised Learning...

Unsupervised Learning methods

- thus, it draw abstractions from unlabeled datasets and apply these to new data.
- the example methods include
  - Hierarchical, K-Means clustering.
  - K-NN (k nearest neighbours).
  - Principal Component Analysis.
  - Singular Value Decomposition.
  - Independent Component Analysis.
  - Self-organization map.

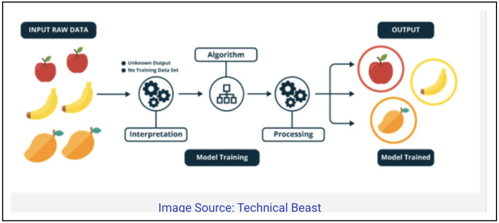

Figure: Un-Supervised Learning

Unsupervised Learning methods

- thus, it draw abstractions from unlabeled datasets and apply these to new data.
- the example methods include
  - Hierarchical, K-Means clustering.
  - K-NN (k nearest neighbours).
  - Principal Component Analysis.
  - Singular Value Decomposition.
  - Independent Component Analysis.
  - Self-organization map.
- are difficult to evaluate, because does not have an explicit teacher i.e. does not have labeled data for testing.

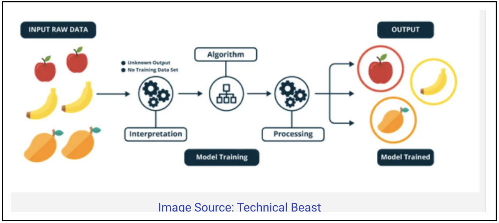

Figure: Un-Supervised Learning

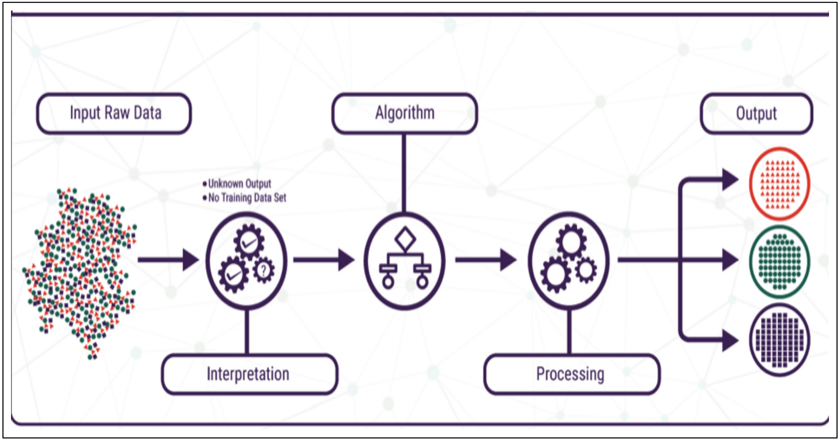

Figure: Un-Supervised Learning

1

[1] https://www.crayondata.com/machine-learning-explained-understanding-supervised-unsupervised-and-reinforcement-learning/

Un-Supervised Learning algorithms...: Advantages

- are categorized based on the structures and objective functions of learning algorithms.
- less intricacy in correlation with administered learning
- nobody is needed to comprehend and afterward name i.e. label the information inputs
- it is frequently simpler to get unlabeled information

[1]

[1] https://omdena.com/blog/supervised-and-unsupervised-machine-learning/

Un-Supervised Learning algorithms...: Dis-advantages

- less exactness of the outcomes.
- the consequences of the investigation can't be found out

[1]

[1] https://omdena.com/blog/supervised-and-unsupervised-machine-learning/

Un-Supervised Learning algorithms...: Real-world Applications

- An Anomaly detection system developed using USML.
  - The system is capable of capturing sudden vegetation changes, which can be used as an alert mechanism to provide immediate relief to the people and communities in need.
- Besides, USML is generally used for

- Optical character recognition (OCR)
- Search engines
- Computer vision
- Classifying DNA sequences
- Detecting fraud, e.g., credit card and internet
- Medical diagnosis
- Natural language processing
- Speech and handwriting recognition
- Economics and finance
- Recommendation engines, such as those used by Netflix and Amazon

Supervised & Un-Supervised Learning algorithms

- Supervised learning = uses labeled data
- Unsupervised learning = uses unlabeled data.
- Well the main difference is that supervised learning uses off-line analysis whereas unsupervised learning uses real-time analysis of data.
- In SL, the number of classes is known but in unsupervised learning the number of classes is unknown.
- The results of supervised learning are accurate and reliable,
- on the other hand, the results of unsupervised learning are moderate, accurate, and reliable.

Supervised & Un-Supervised Learning algorithms

Parameters	Supervised machine learning	Unsupervised machine learning
Input Data	Algorithms are trained using labeled data.	Algorithms are used against data that is not labeled
Computational Complexity	Simpler method	Computationally complex
Accuracy	Highly accurate	Less accurate
No. of classes	No. of classes is known	No. of classes is not known
Data Analysis	Uses offline analysis	Uses real-time analysis of data
Algorithms used	Linear and Logistics regression, Random forest, Support Vector Machine, Neural Network, etc.	K-Means clustering, Hierarchical clustering, Apriori algorithm, etc.

Figure: Machine Learning

Reinforcement Learning

- Unlike SL and USL, reinforcement learning builds its prediction model by gaining feedback from random trial and error and leveraging insight from previous iterations.
- the goal is to achieve a specific goal (output) by randomly trialling a vast number of possible input combinations and grading their performance
- can best be explained by using a video game analogy
- algorithms are set to train the model based on continuous learning.
- a standard reinforcement learning model has measurable performance criteria where outputs are graded.
    - In the case of self-driving vehicles, avoiding a crash earns a positive score, and in the case of chess, avoiding defeat likewise receives a positive assessment.

Q Learning

- is a a specific algorithmic example of reinforcement learning
- understand through the Pac-Man game, as follows......
- Three main components
  - states could be the challenges, obstacles or pathways that exist in the video game
  - "A" - could depict the set of possible actions to respond to these states limited to left, right, up, and down movements, as well as multiple combinations thereof.
  - "q" - could depict the the model's starting value and has an initial value of "0."
- as the game progresses, two main things happen
  - Q drops as negative things occur after a given state/action.
  - Q increases as positive things occur after a given state/action.
- In Q-learning, the machine learns to match the action for a given state that generates or preserves the highest level of Q
- the model records its results (rewards and penalties) and how they impact its Q level and stores those values to inform and optimize its future actions.
- this is computationally expensive

An Overview of ML tasks

...to be continued

B l a n k

B l a n k