Big Data Hadoop and Spark Developer

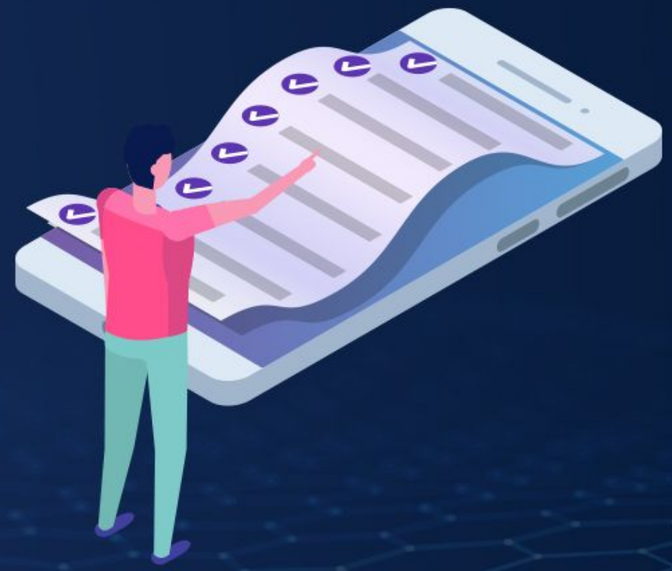
simplilearn

Spark MLlib: Modeling Big Data with Spark

Learning Objectives

By the end of this lesson, you will be able to:

- Identify the skills required to become a data scientist and data analyst
- Define analytics in Spark and list the types of analytics
- Describe the machine learning algorithms
- Define MLlib and MLlib pipeline



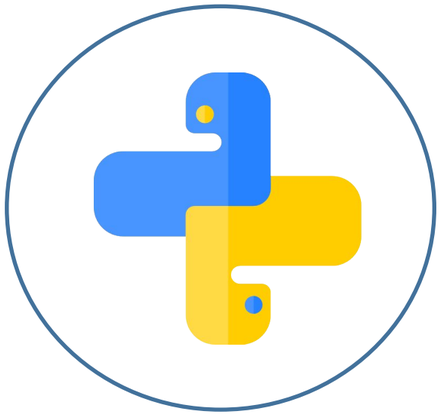
DATA AND
ARTIFICIAL INTELLIGENCE

Role of Data Scientist and Data Analyst in Big Data

Who Is a Data Scientist?

A data scientist is the person who gathers data from multiple sources and applies machine learning, predictive analytics, and sentiment analysis to extract critical information from the collected data sets.

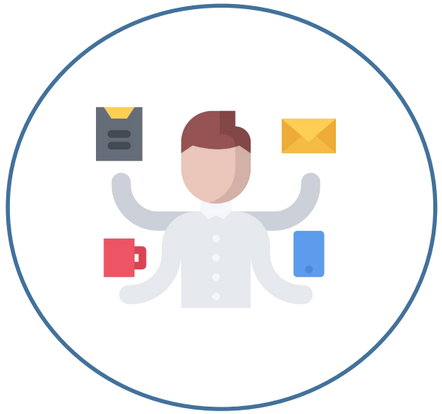
Skills Required to Become a Data Scientist



Knowledge of Python and R

Knowledge of Machine Learning

Experience in SQL



Understanding of Multiple Analytics Function

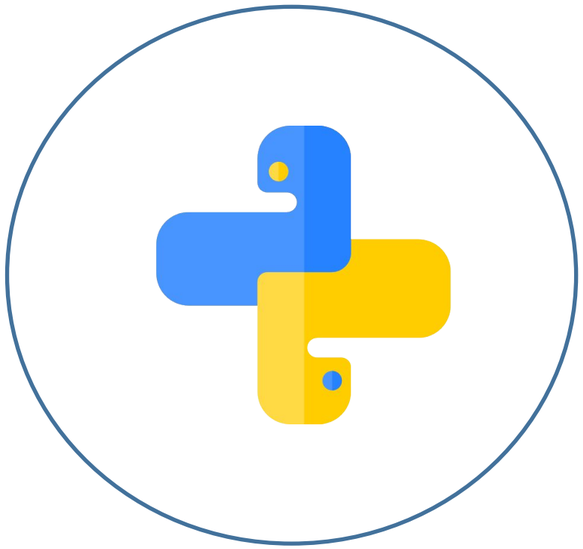
Ability to Work with Unstructured Data

Who Is a Data Analyst?

> A data analyst is the person who can do basic descriptive statistics, visualize data, and communicate data points for conclusions.

Skills Required to Become a Data Analyst

Knowledge of Mathematical Statistics



Understanding of R and Python

Data Wrangling

Understanding of Pig and Hive

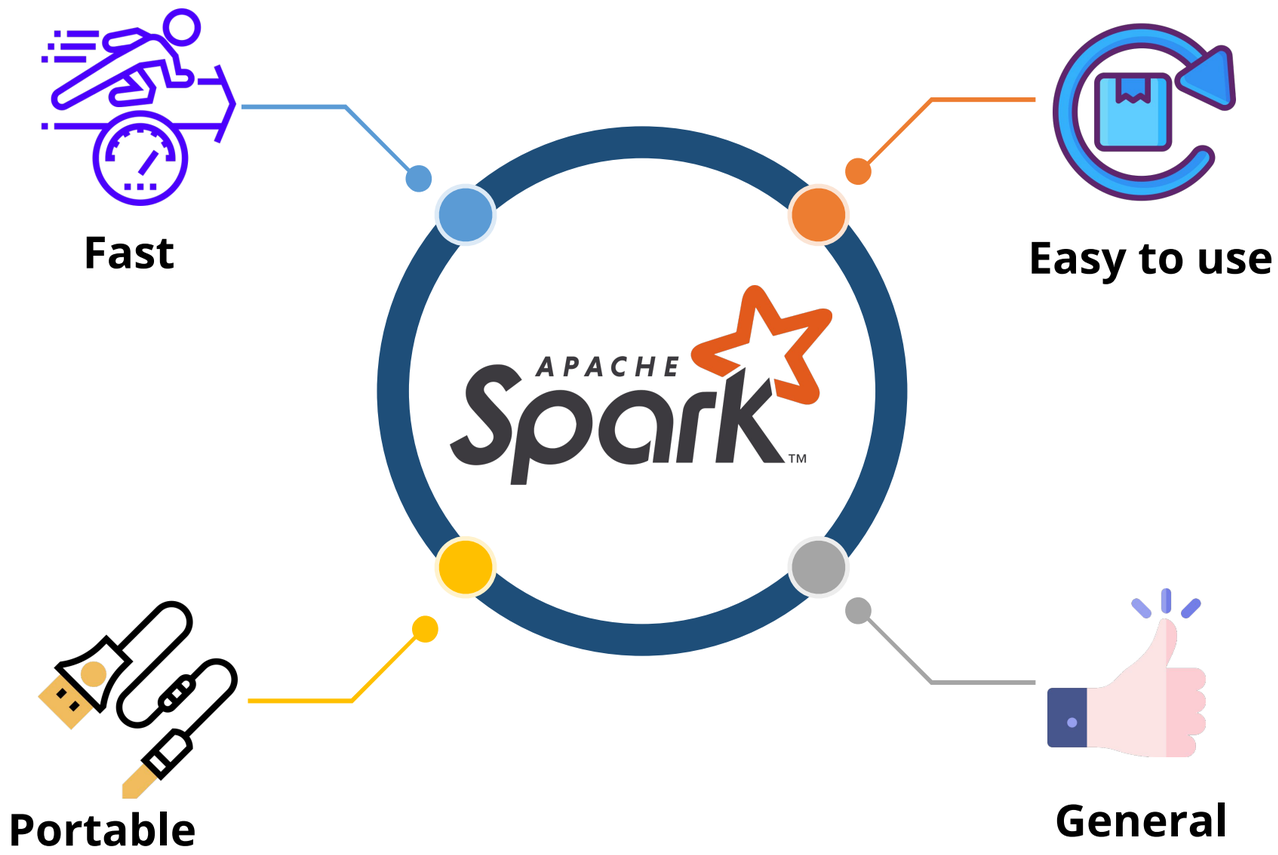
DATA AND
ARTIFICIAL INTELLIGENCE

Analytics in Spark

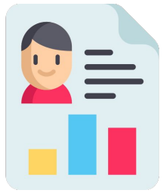
Analytics in Spark

Apache Spark is a unified analytics engine for large-scale data processing.

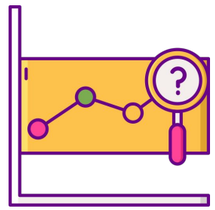
The following are the benefits of analytics in Spark:



Types of Analytics



Descriptive Analytics

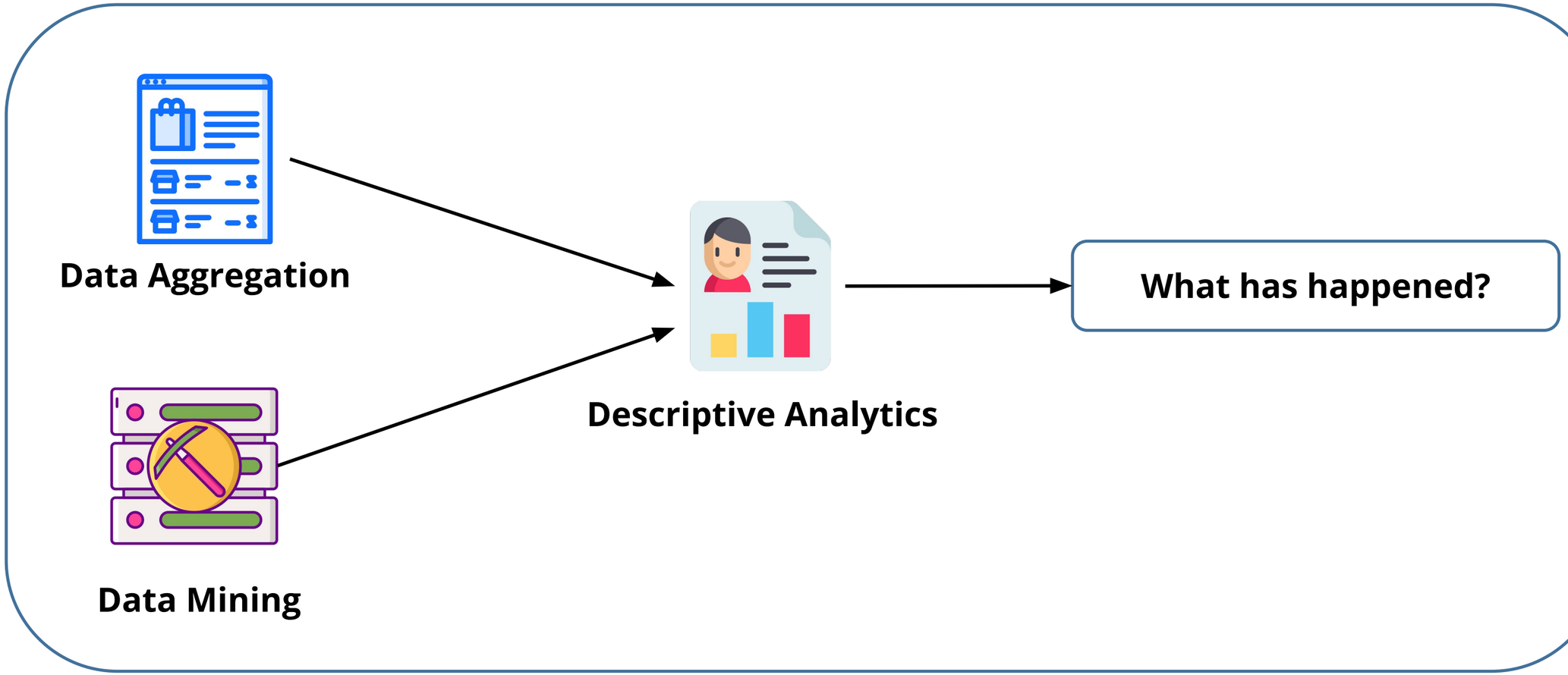


Predictive Analytics

Prescriptive Analytics

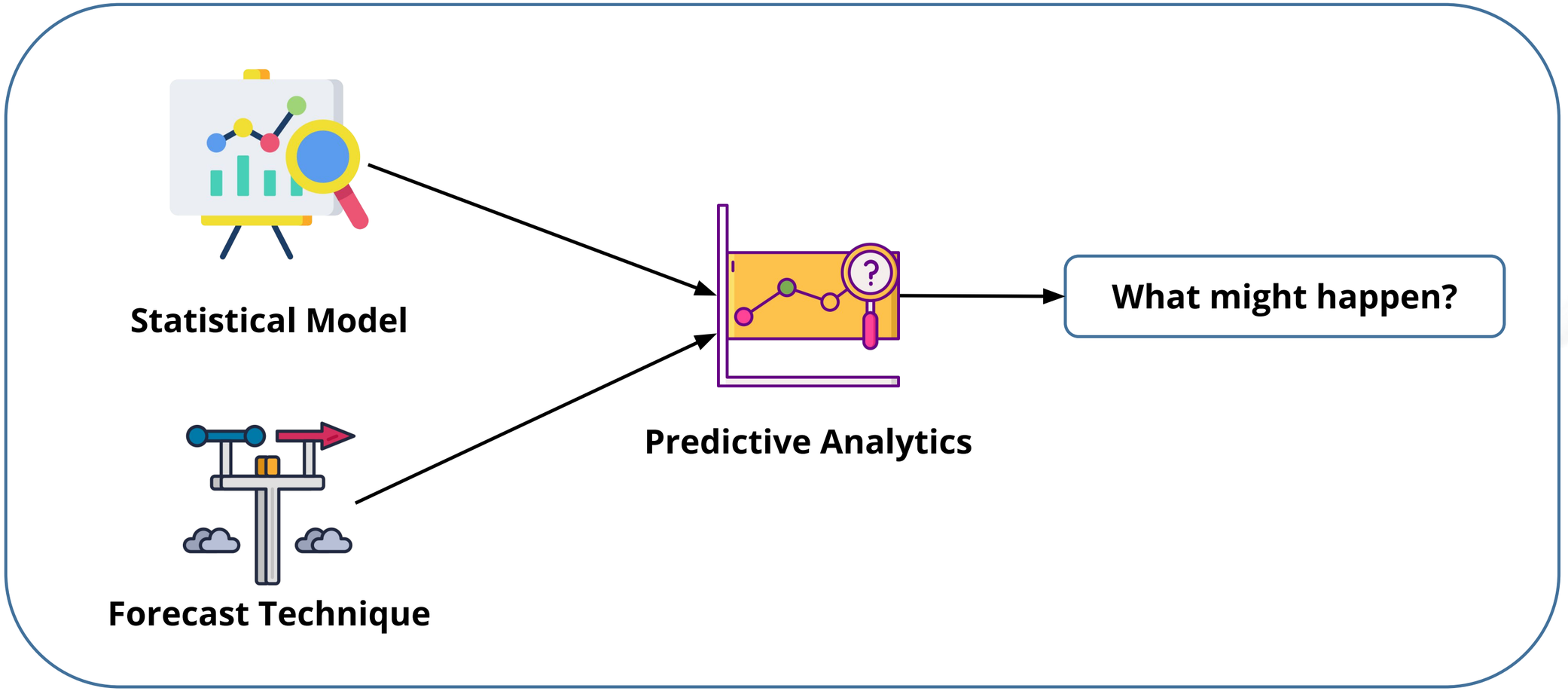
Descriptive Analytics

The type of analytics that describes the past and answers the question: “What has happened?”.



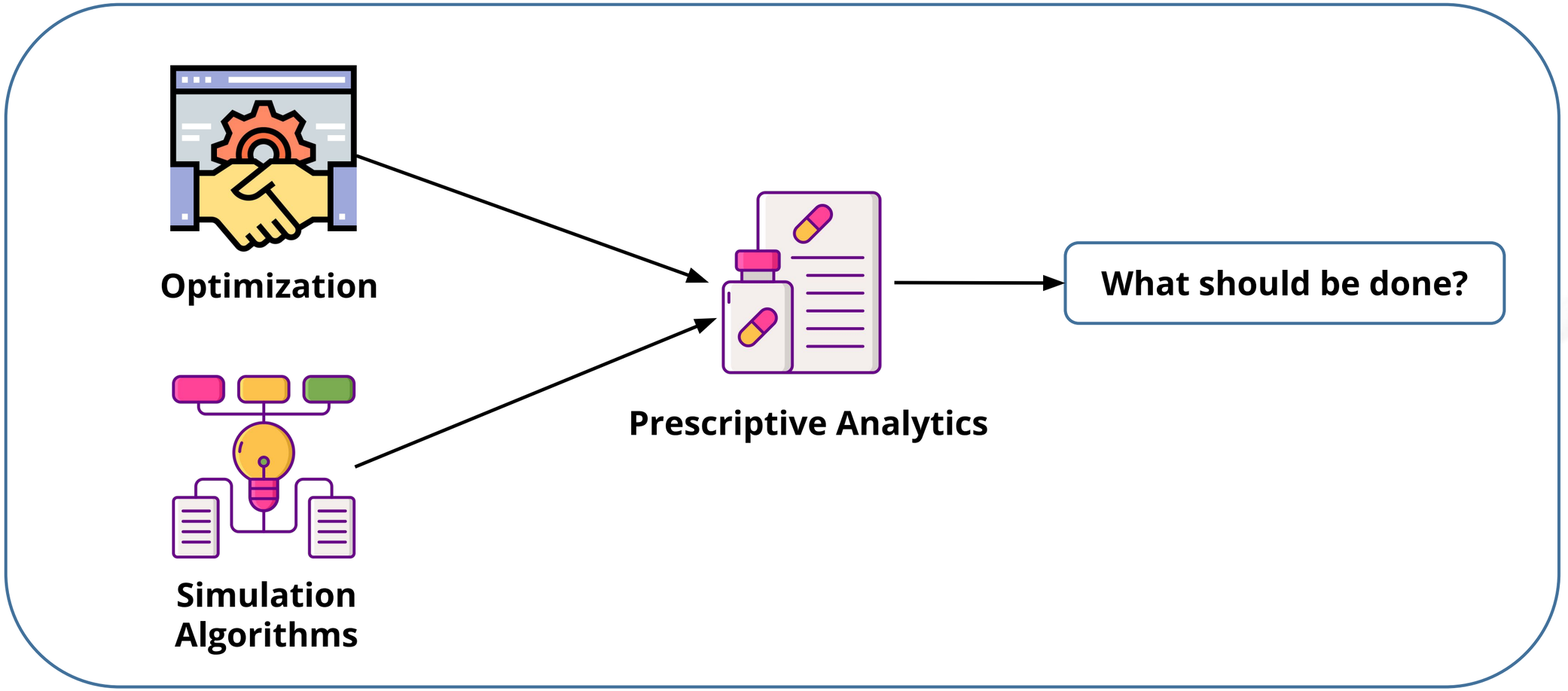
Predictive Analytics

The type of analytics that has the ability to understand the future and answer the question: “What might happen?”.



Prescriptive Analytics

The type of analytics that is used to advice the users on possible outcome and answer the question: “What should be done?”.



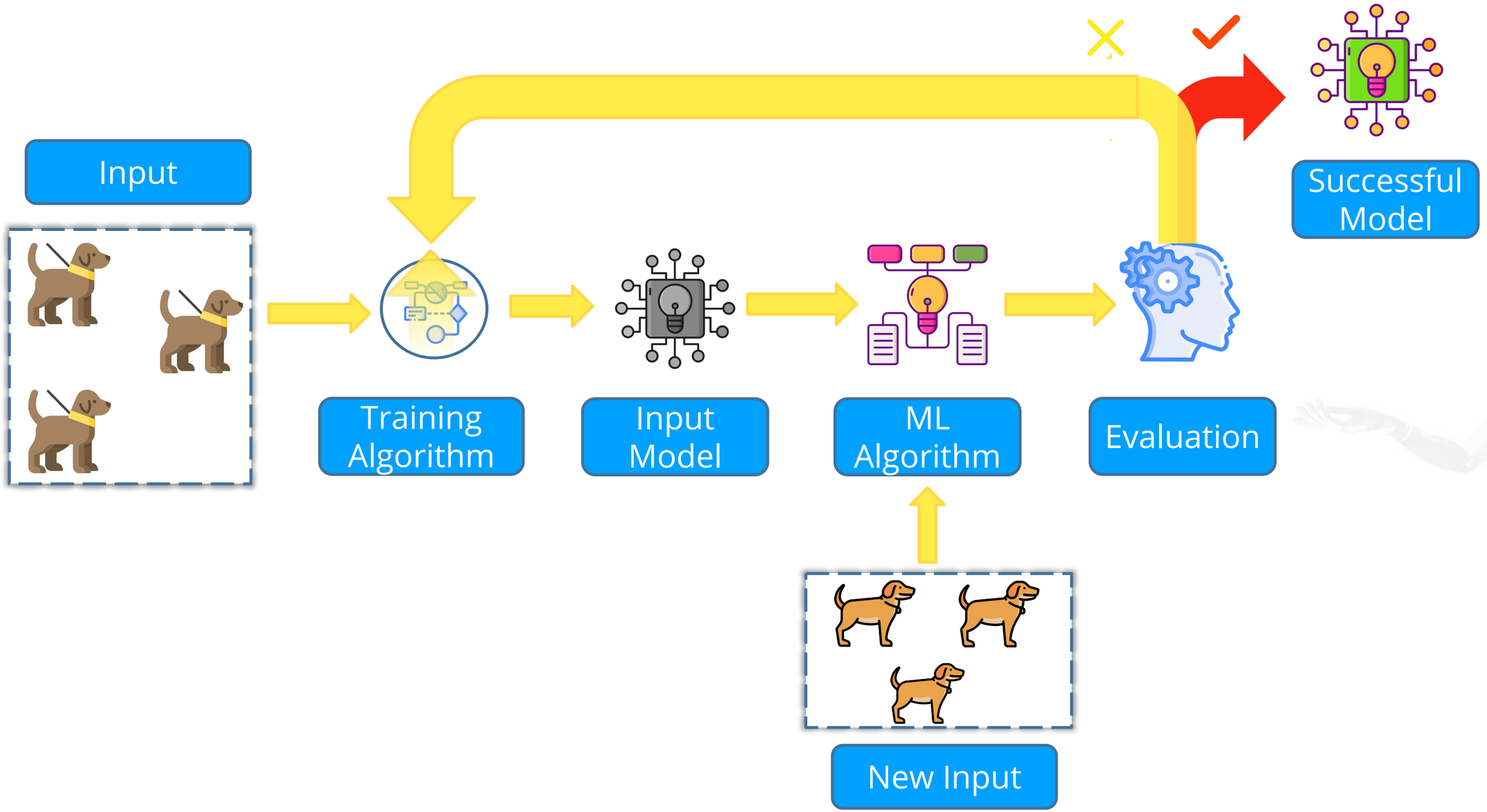
DATA AND
ARTIFICIAL INTELLIGENCE

Machine Learning

What Is Machine Learning?

> The capability of Artificial Intelligence systems to learn by extracting patterns from data is known as Machine Learning.

What Is Machine Learning?



Relationship between Machine Learning and Data Science

Data Science and Machine Learning go hand in hand.
Data Science helps evaluate data for Machine Learning algorithms.

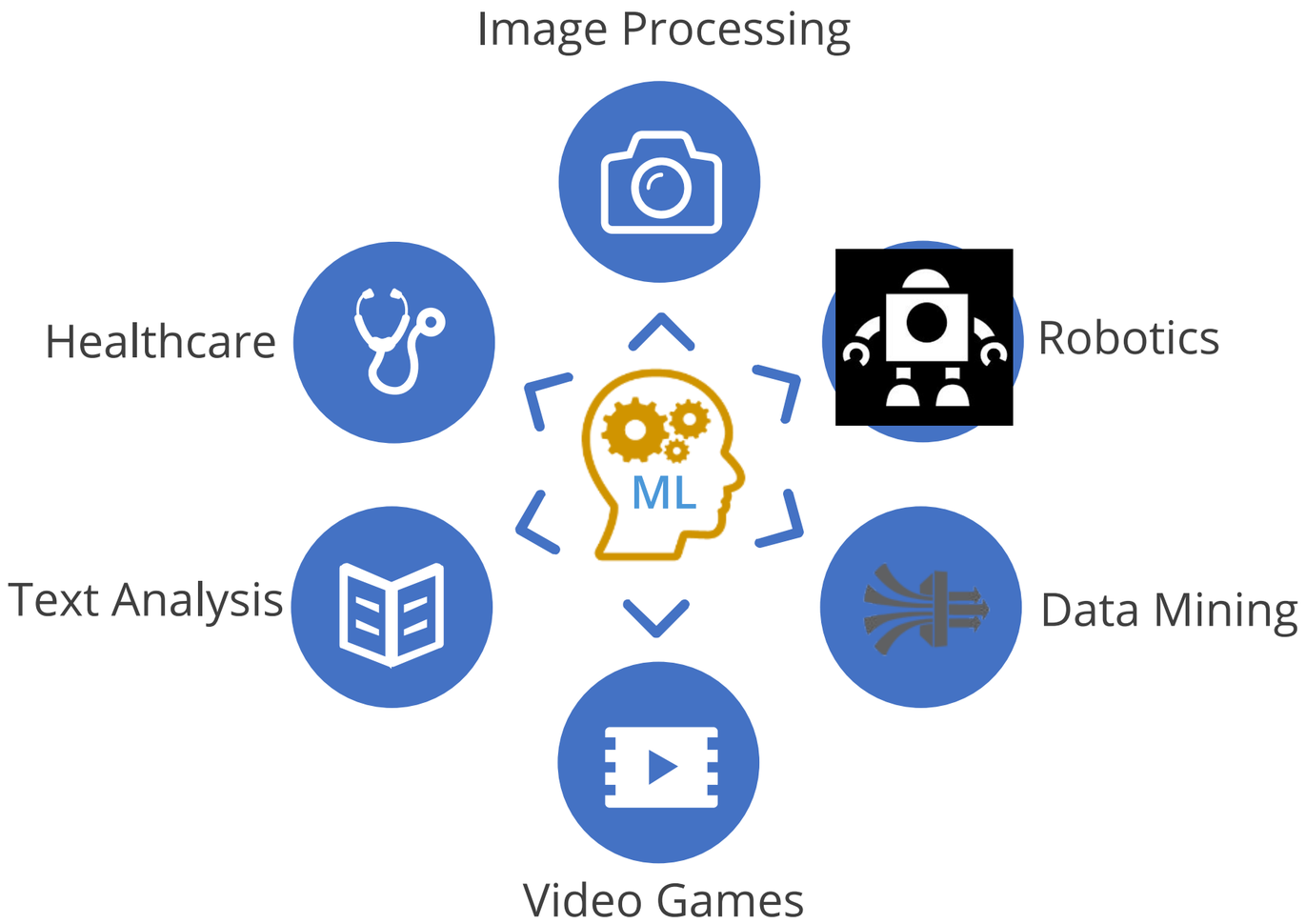
Large-Scale Machine Learning

Large-scale machine learning involves large data which has large number of training, features, or classes.

Large-Scale Machine Learning Tools

Applications of Machine Learning

Machine learning is used in various fields such as:



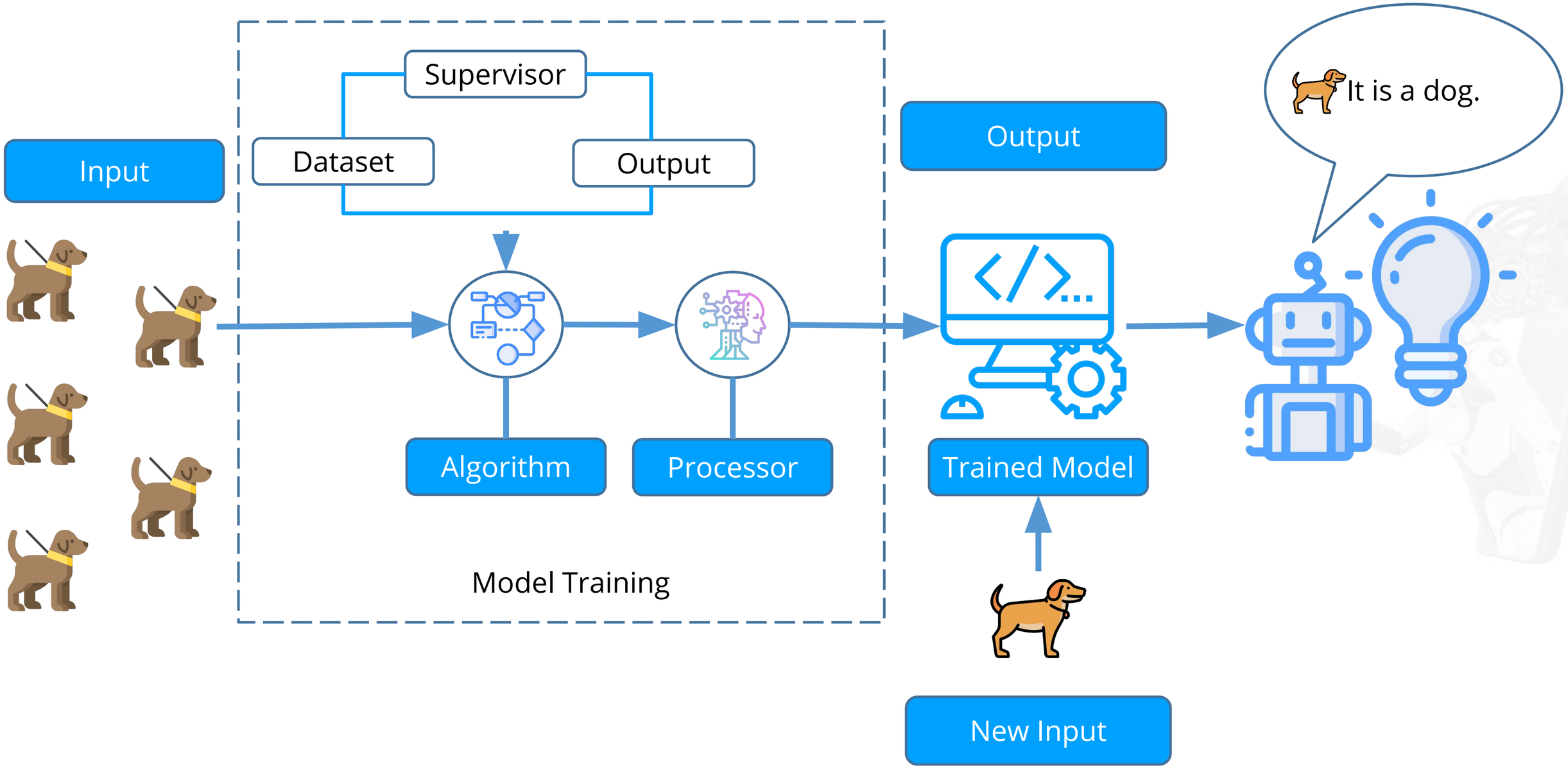
Types of ML Algorithms

- Supervised Learning
- Unsupervised Learning
- Semi-Supervised Learning
- Reinforcement Learning

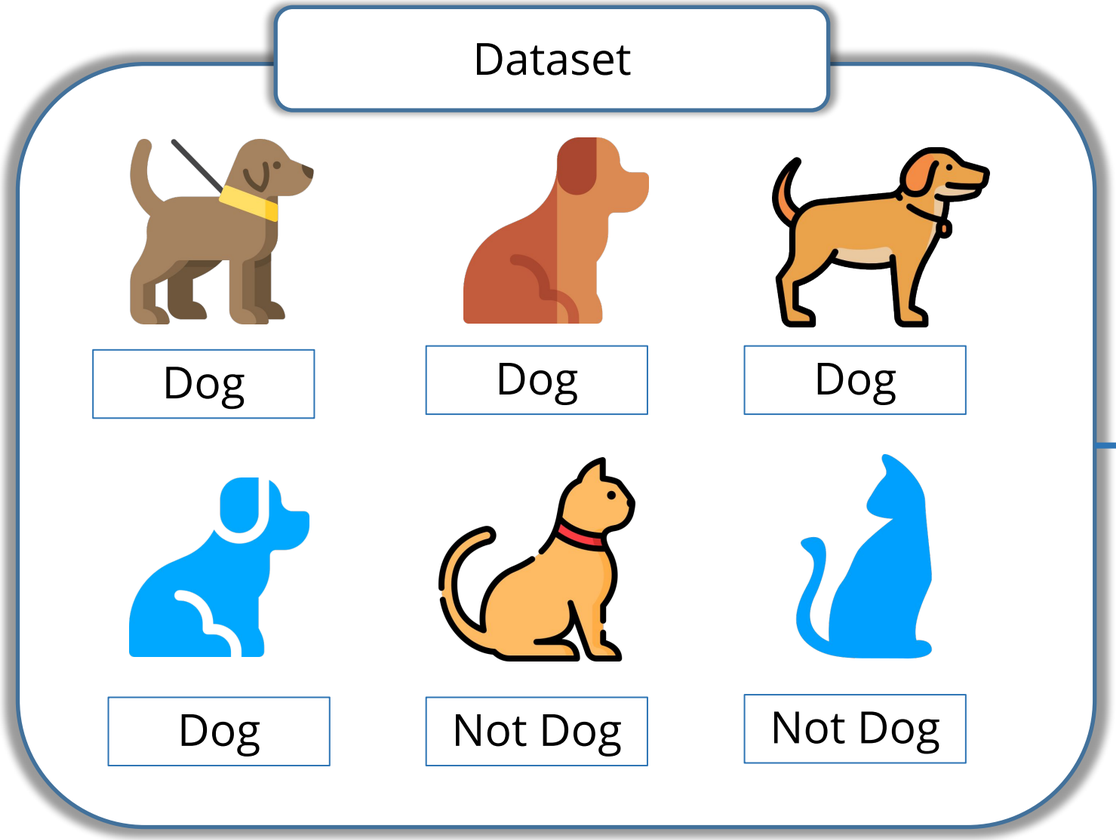
DATA AND
ARTIFICIAL INTELLIGENCE

Supervised Learning

Supervised Learning



Supervised Learning: Example



Supervised Learning

Predictive Model

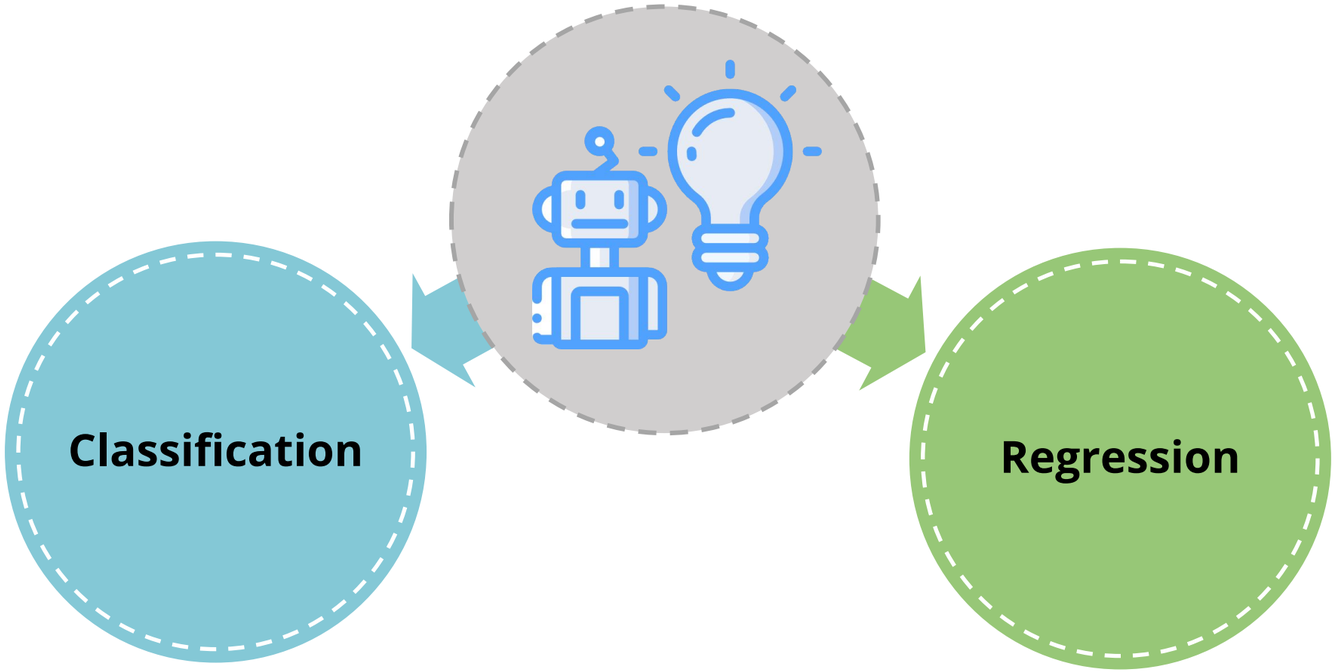
Predictive Model

Dog

Supervised Learning: Example

Netflix uses **supervised learning** algorithms to recommend users the shows they may watch based on the viewing history and ratings by similar classes of users.

Supervised Learning Algorithms



Classification with Real-World Problem

Duration: 10 mins

Problem Statement: In this demonstration, you will perform classification with real-world problem.

Access: Click on the **Practice Labs** tab on the left side panel of the LMS. Copy or note the username and password that is generated. Click on the **Launch Lab** button. On the page that appears, enter the username and password in the respective fields, and click **Login**.

Linear Regression with Real-World Problem

Duration: 10 mins

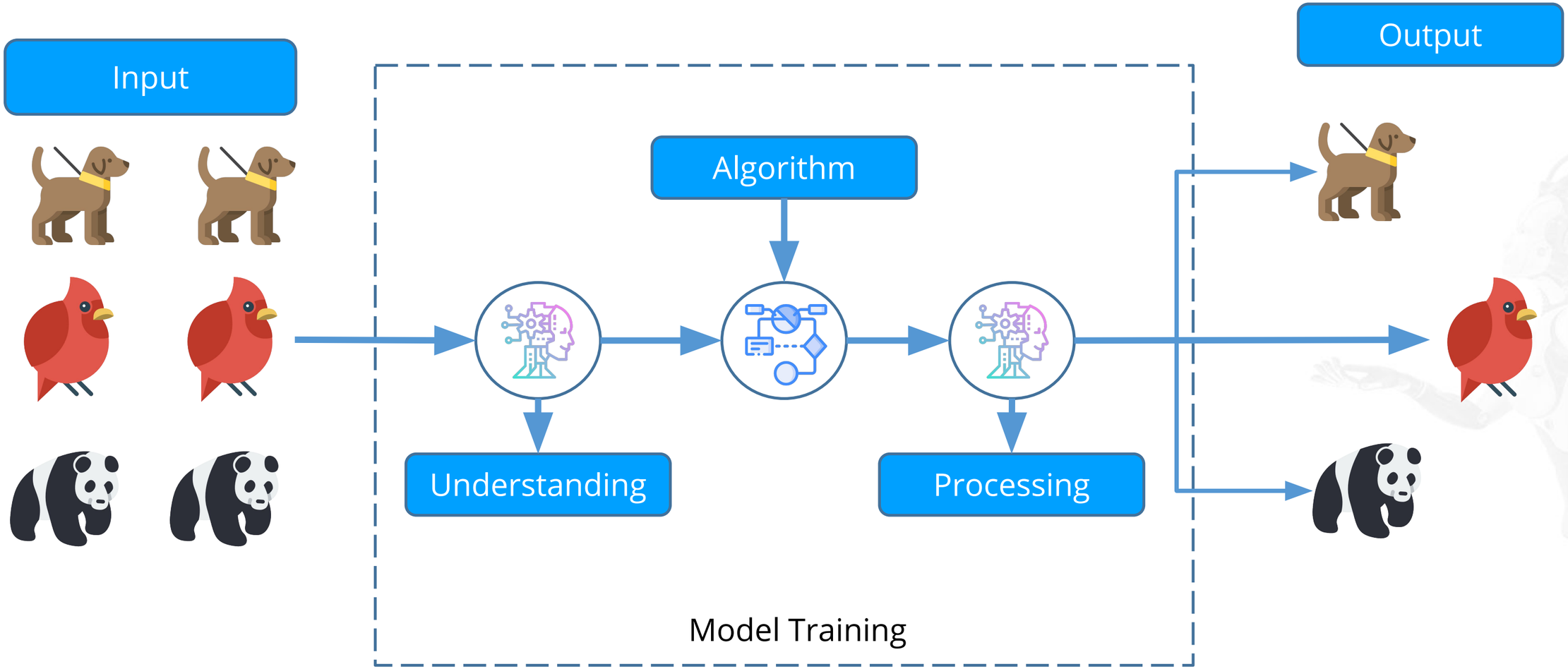
Problem Statement: In this demonstration, you will perform linear regression with real-world problem.

Access: Click on the **Practice Labs** tab on the left side panel of the LMS. Copy or note the username and password that is generated. Click on the **Launch Lab** button. On the page that appears, enter the username and password in the respective fields, and click **Login**.

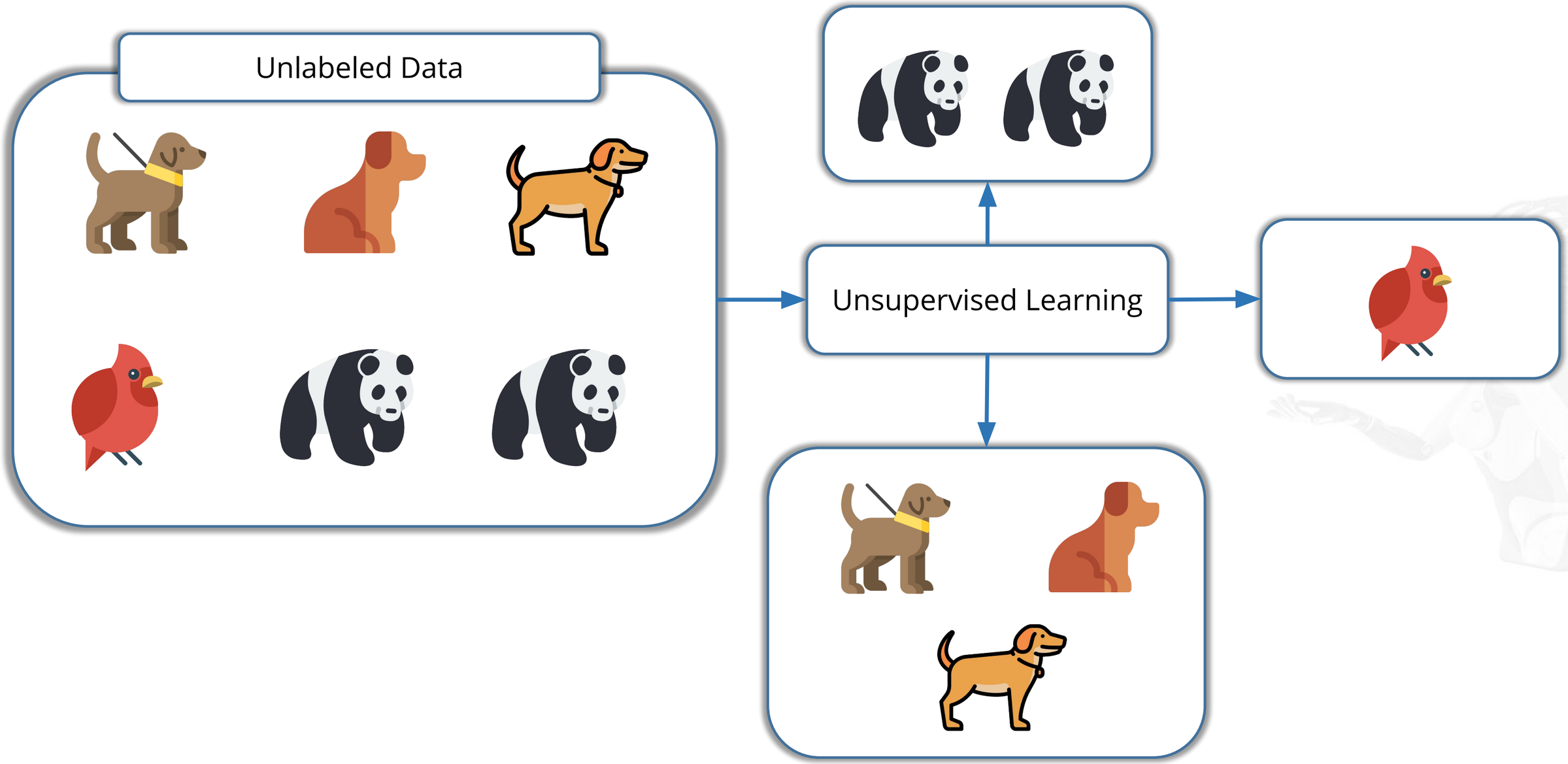
DATA AND
ARTIFICIAL INTELLIGENCE

Unsupervised Learning

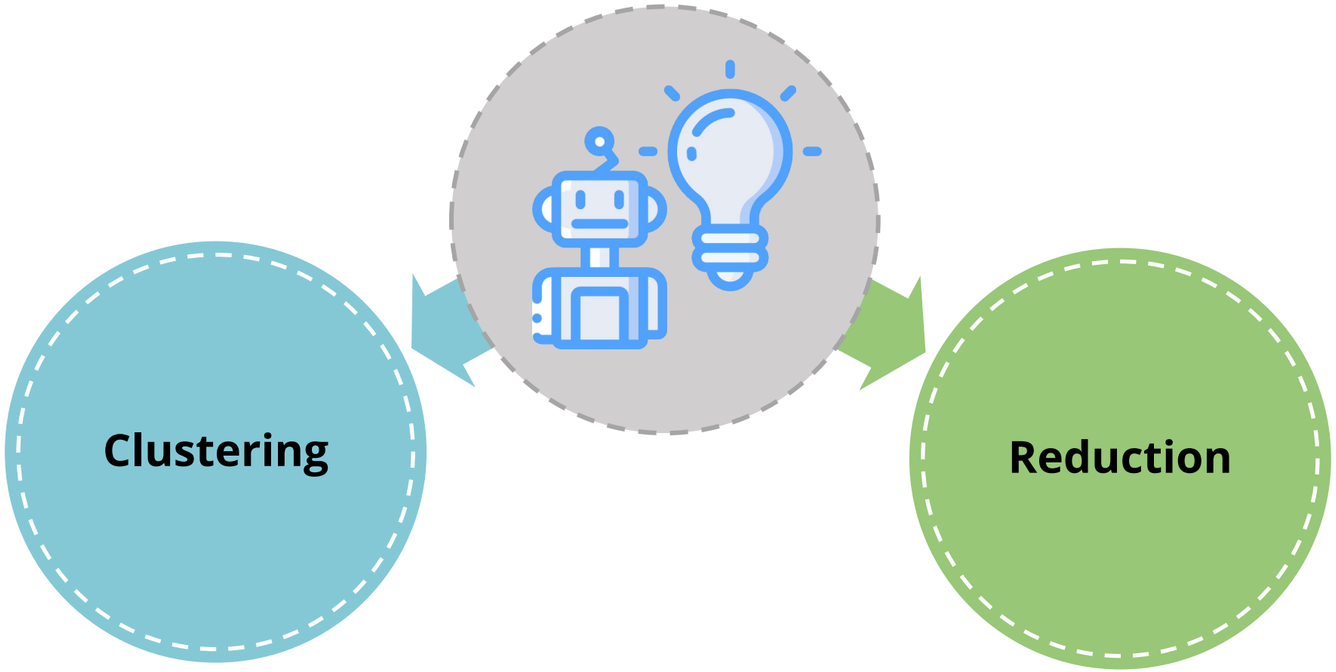
Unsupervised Learning



Unsupervised Learning: Example



Unsupervised Learning Algorithms



Unsupervised Learning with Real-World Problem

Duration: 10 mins

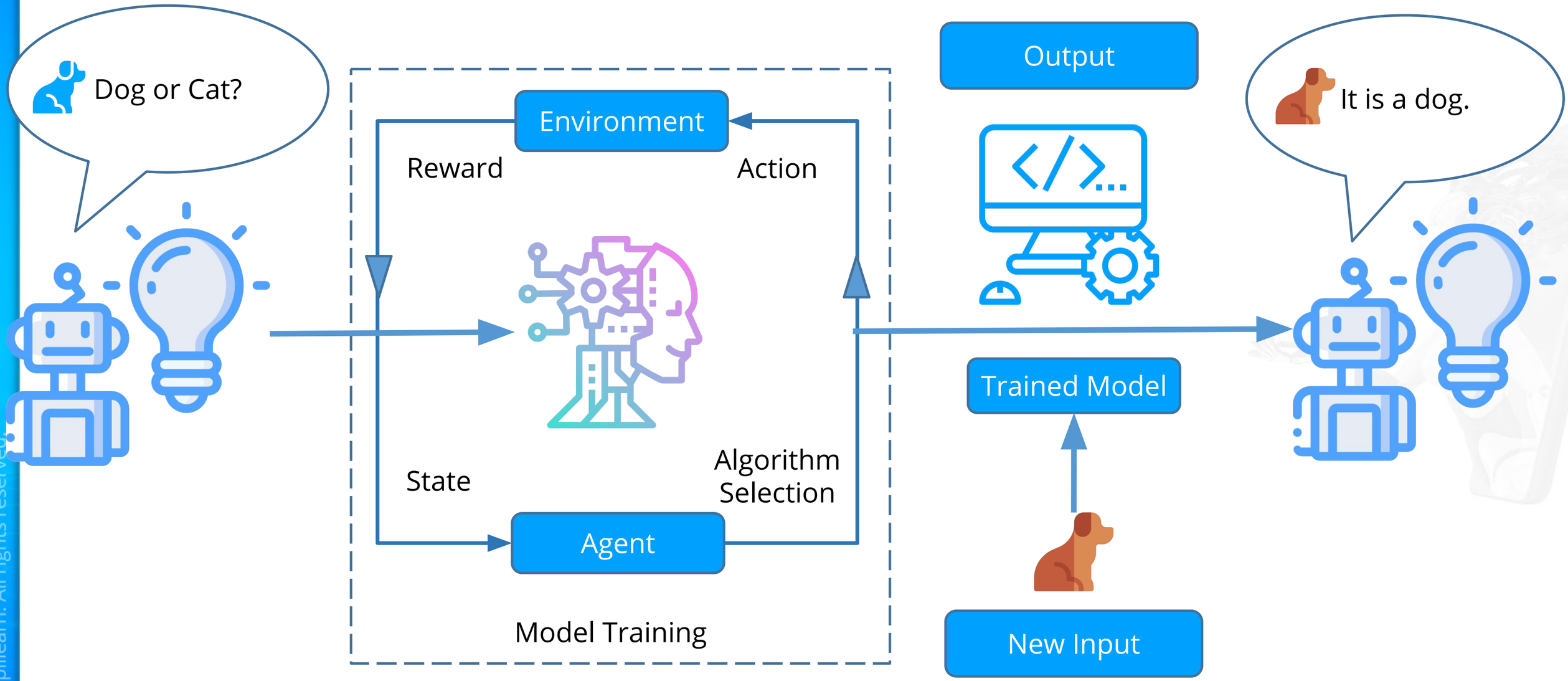
Problem Statement: In this demonstration, you will perform unsupervised learning with real-world problem.

Access: Click on the **Practice Labs** tab on the left side panel of the LMS. Copy or note the username and password that is generated. Click on the **Launch Lab** button. On the page that appears, enter the username and password in the respective fields, and click **Login**.

DATA AND
ARTIFICIAL INTELLIGENCE

Reinforcement Learning

Reinforcement Learning



Reinforcement Learning: Example

1. Before Conditioning

Response

Salivation

Food

Unconditioned Stimulus

Unconditioned Response

2. Before Conditioning

Response

Bell

Salivation

Neutral Stimulus

No Conditioned Response

3. During Conditioning

Response

Salivation

Food

Bell

Unconditioned Response

4. After Conditioning

Response

Bell

Salivation

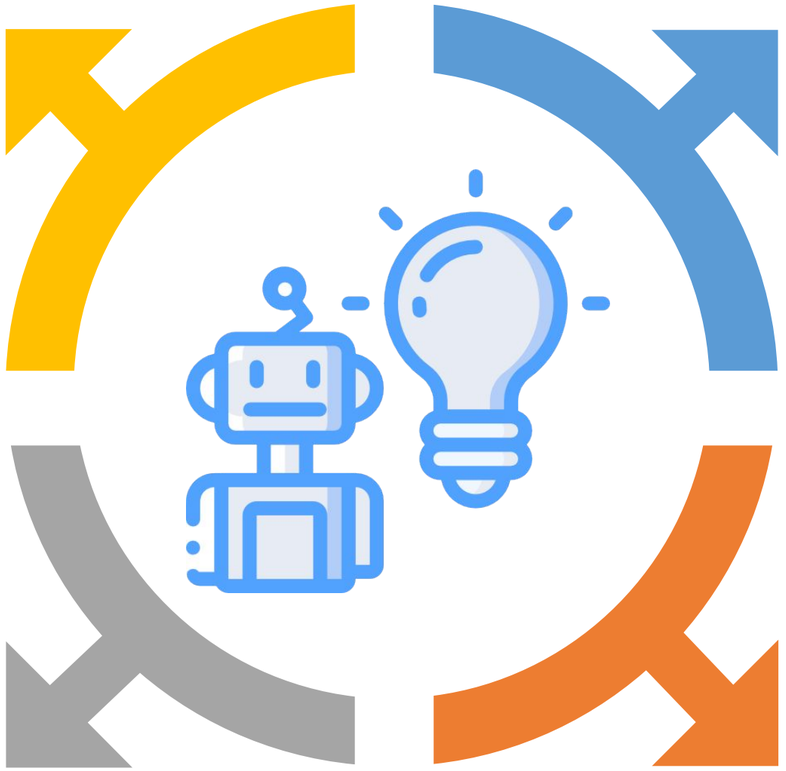
Conditioned Stimulus

Conditioned Response

Reinforcement Learning Algorithms

Q-Learning

Temporal Difference (TD)



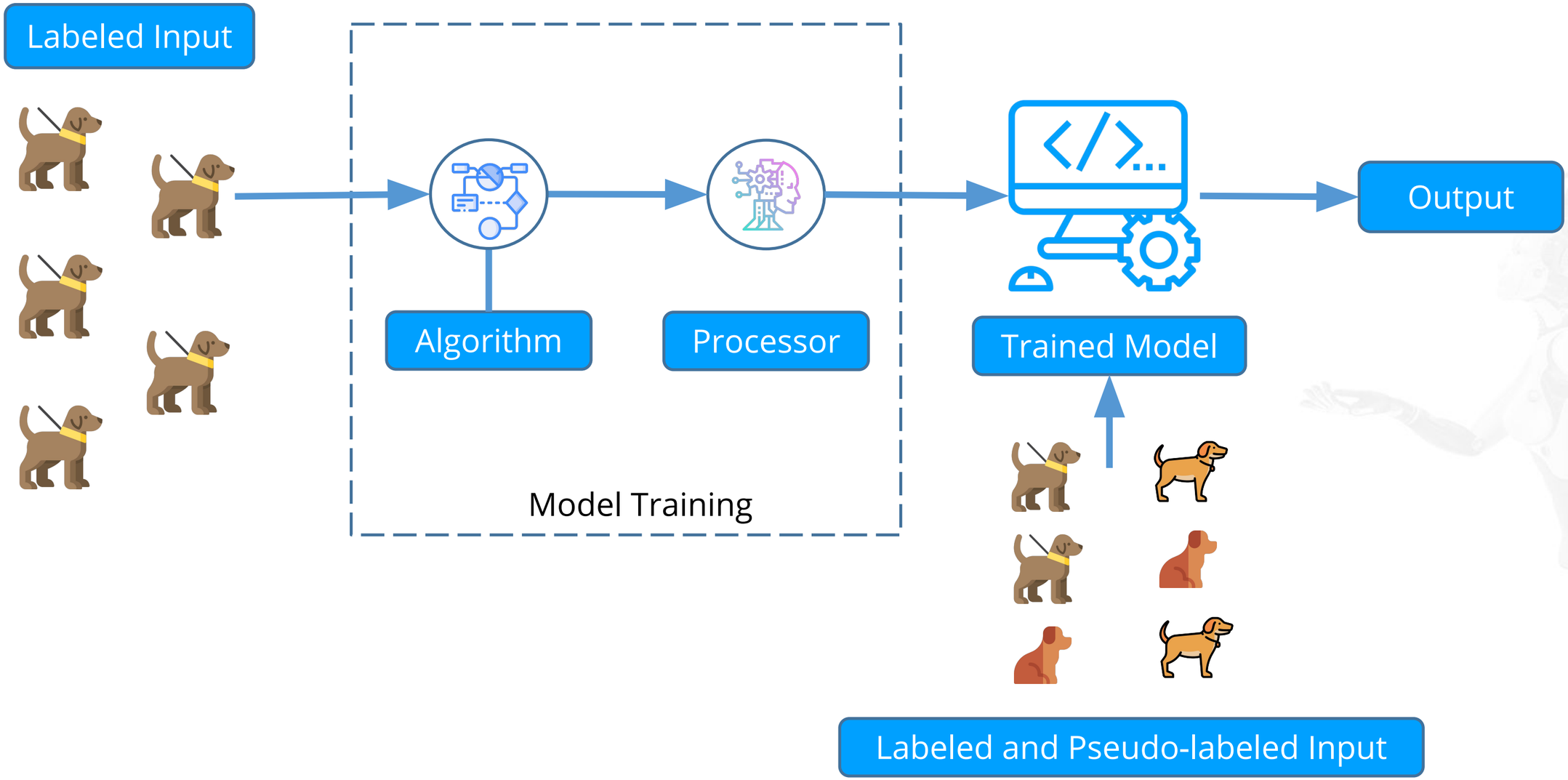
Monte-Carlo Tree Search (MCTS)

Asynchronous Actor-Critic Agents (A3C)

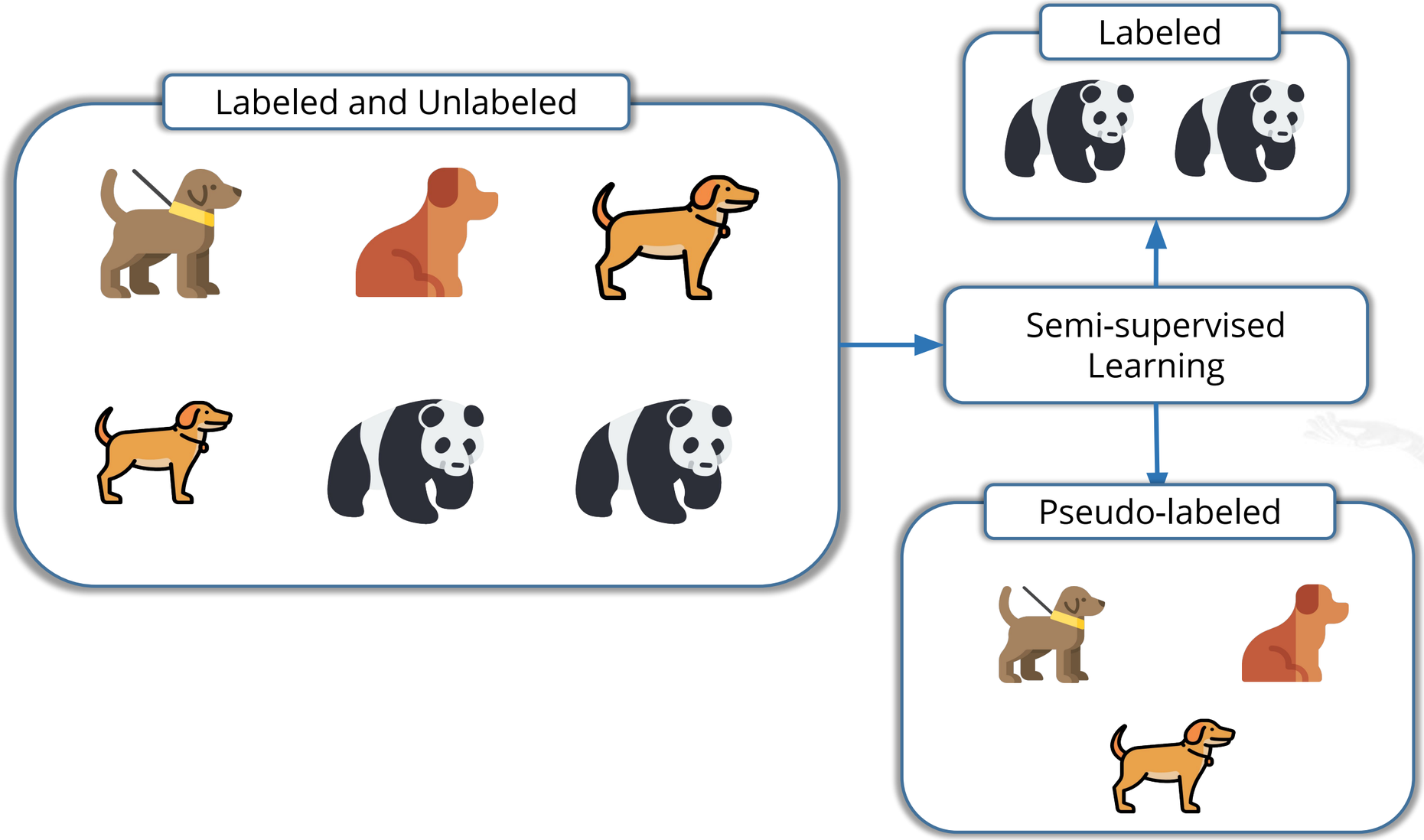
DATA AND
ARTIFICIAL INTELLIGENCE

Semi-Supervised Learning

Semi-Supervised Learning



Semi-Supervised Learning: Example

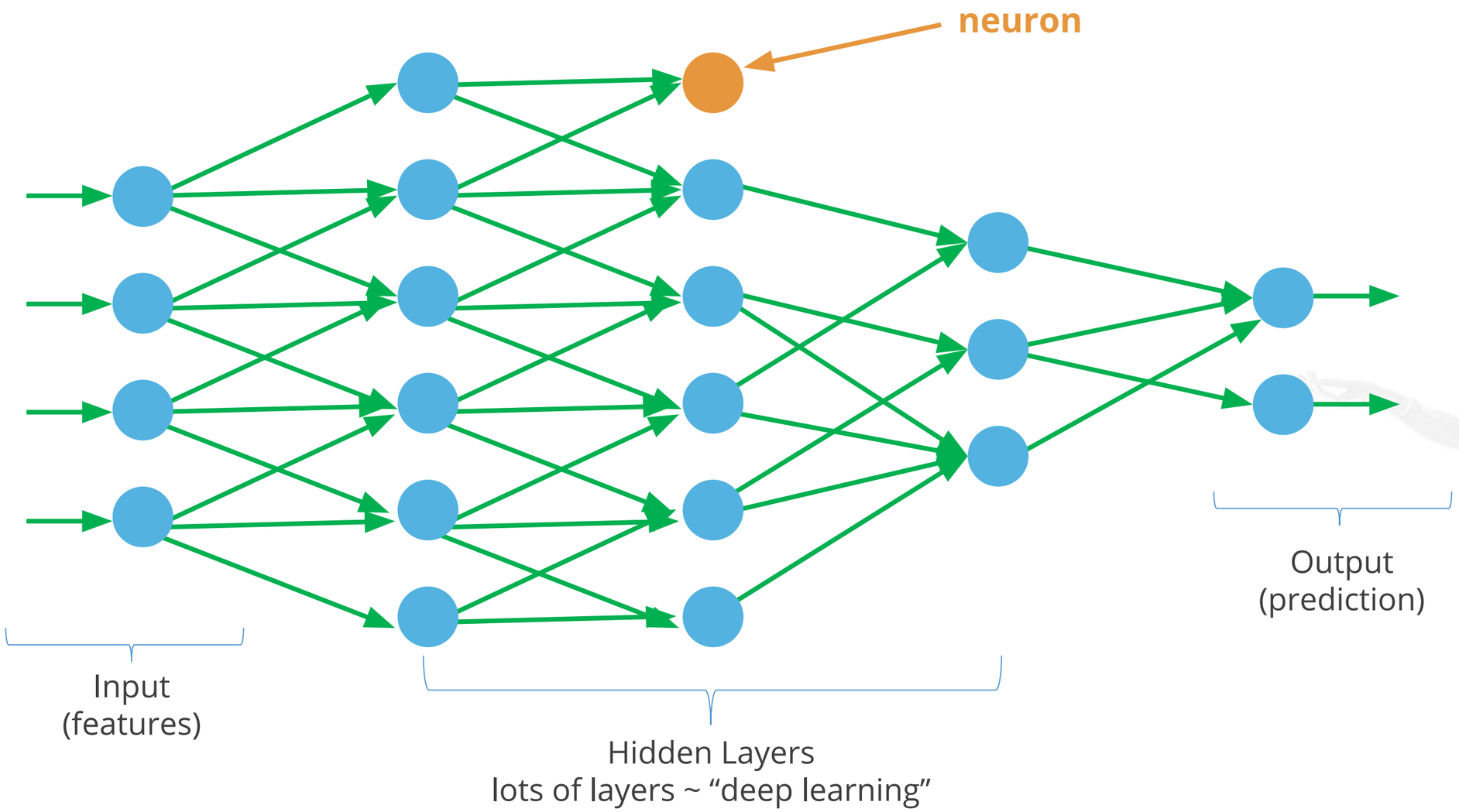


DATA AND
ARTIFICIAL INTELLIGENCE

Deep Learning

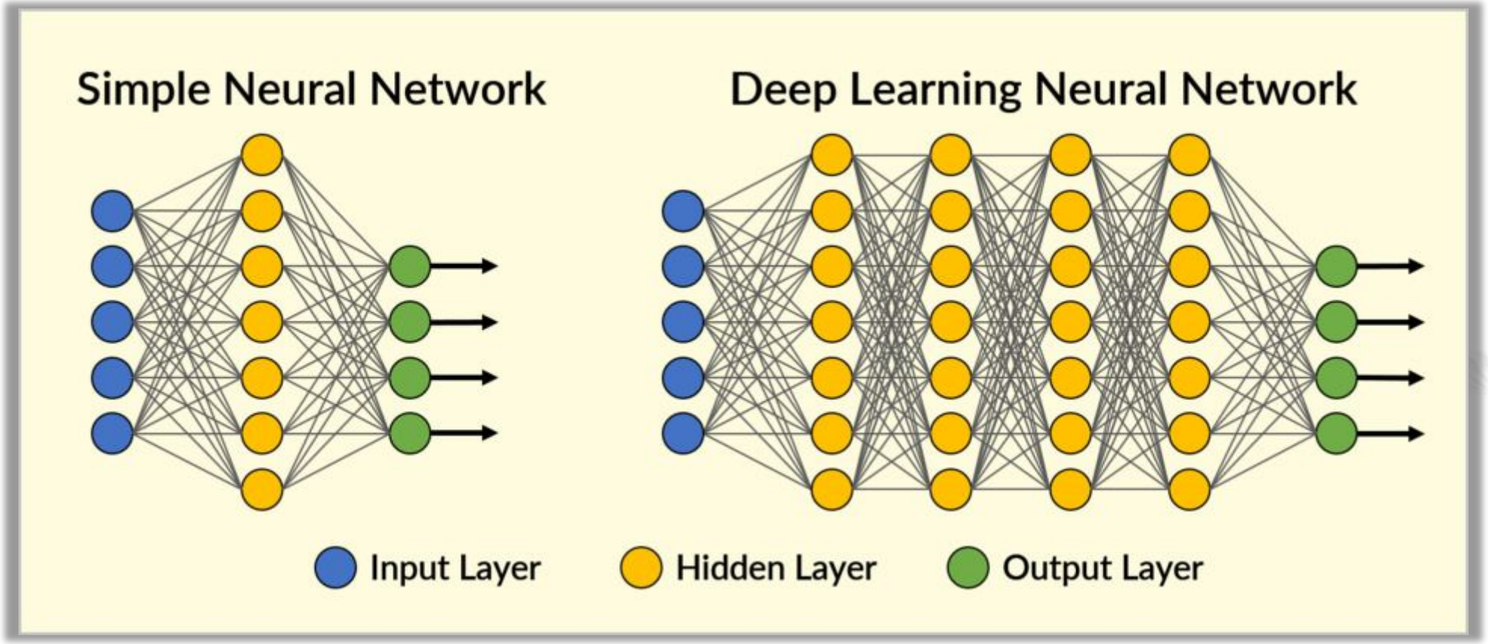
Deep Learning

The “deep” in Deep Learning refers to the large number of layers of neurons that help to learn various representations of data.



Network Architecture

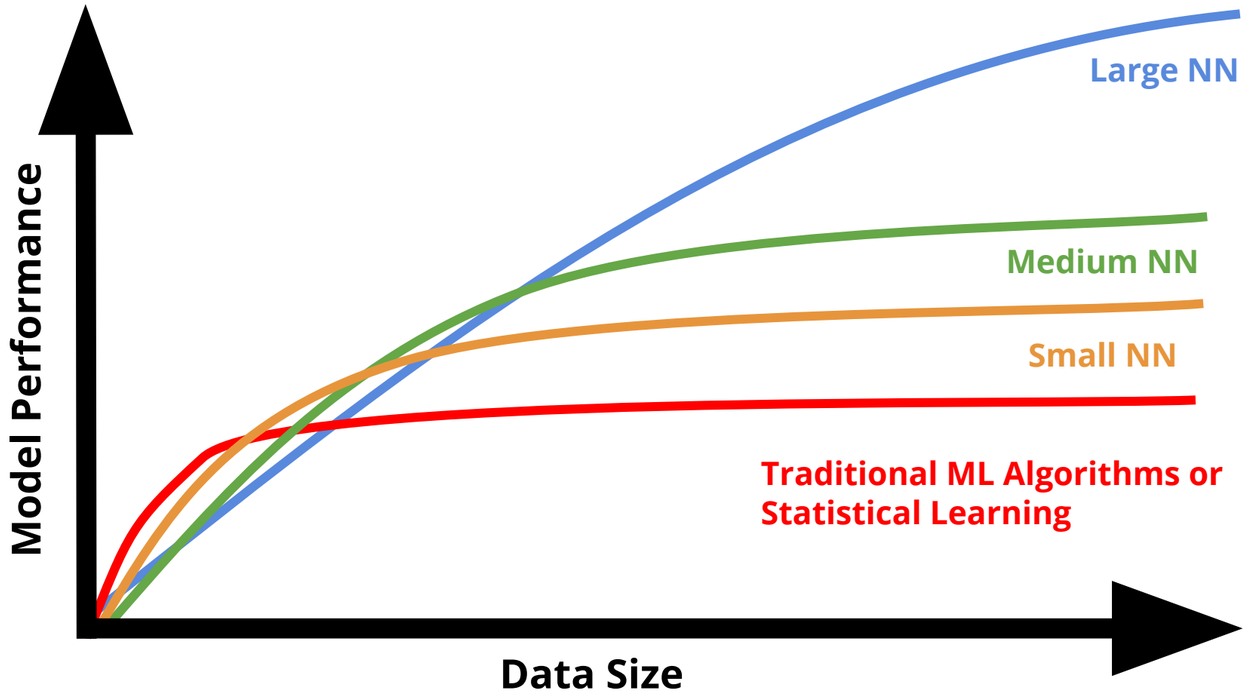
One can programmatically set the number and type of neural layers and the number of neurons comprising each layer.



Shallow neural networks consist of only 1 or 2 hidden layers.

Performance

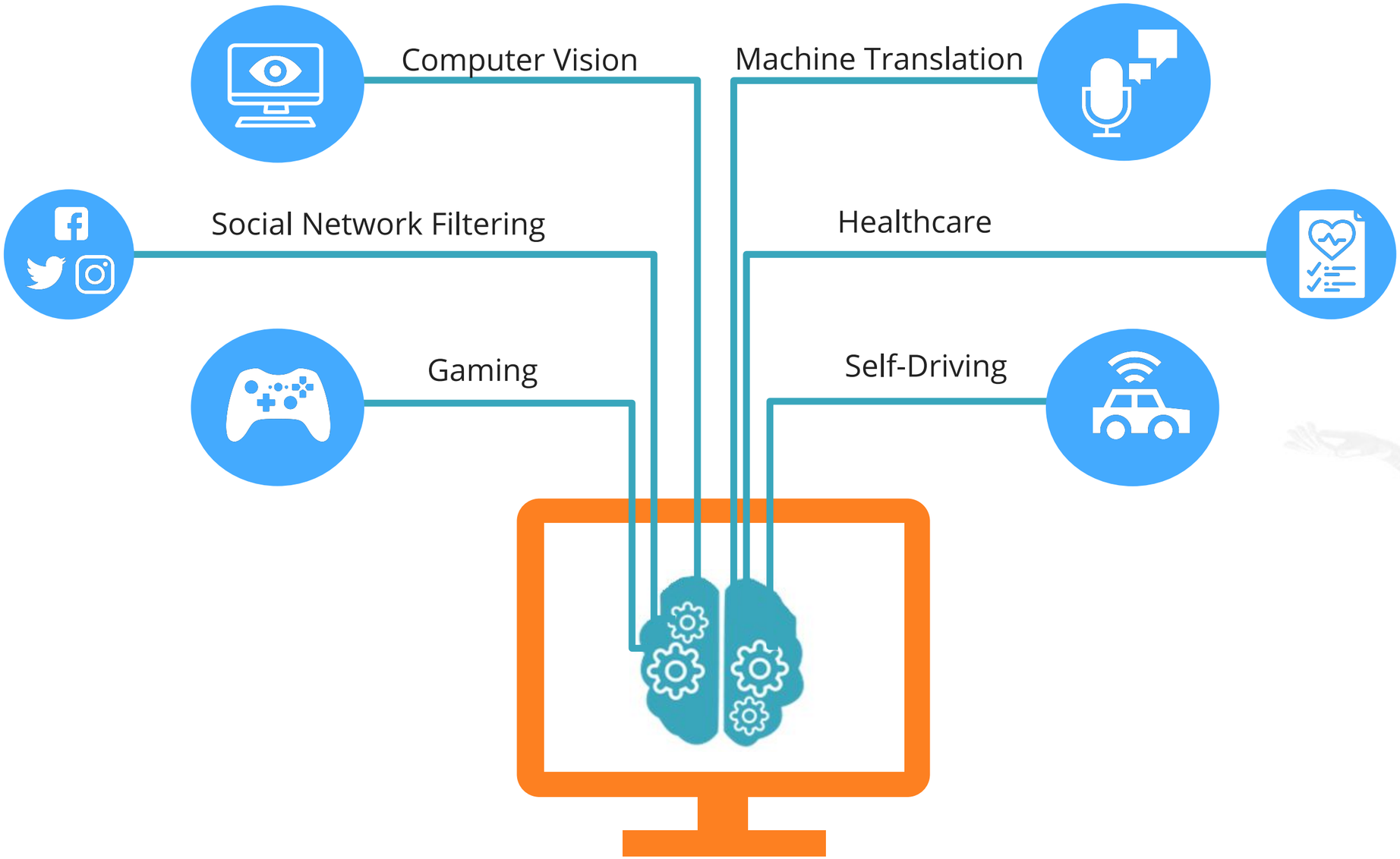
Deep AI models have greater precision than conventional techniques but, require more information to train and attain this precision.



DATA AND
ARTIFICIAL INTELLIGENCE

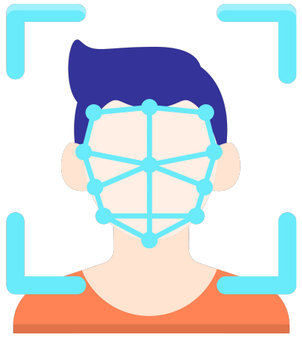
Applications of Deep Learning

Applications



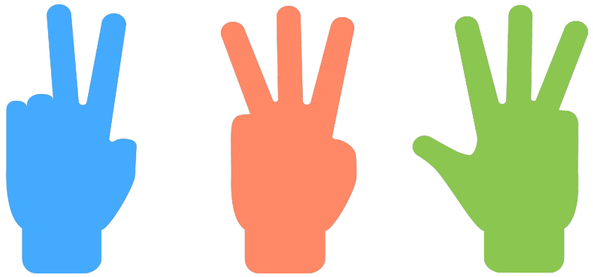
Computer Vision

The field of computer vision uses deep learning neural network methods to solve challenging problems such as:



Facial Recognition

Augmented Reality



Gesture Recognition

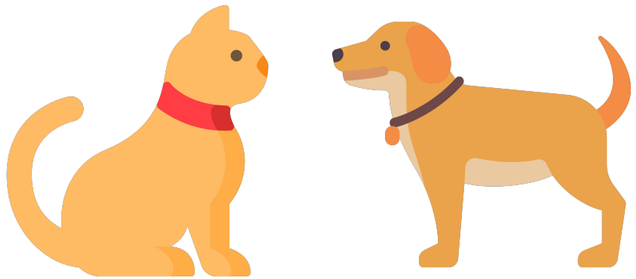
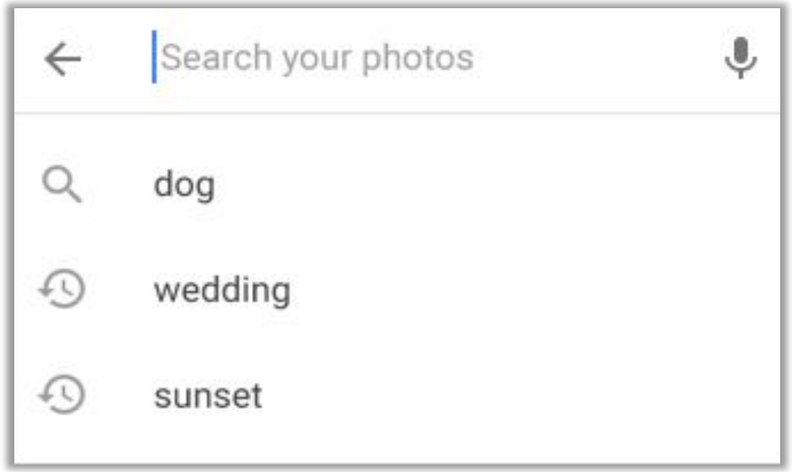


Image Classification

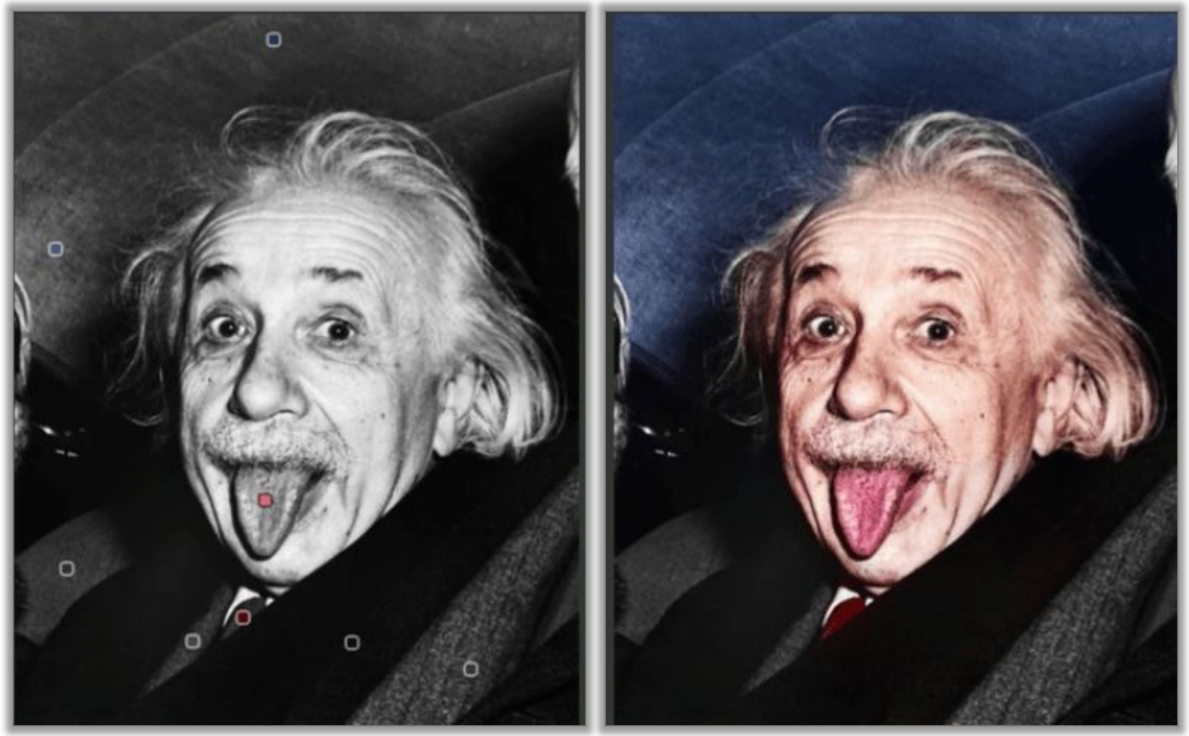
Computer Vision

Advances in image recognition have made it easy to search or automatically organize collections of photos with no identifying tags.



Computer Vision

It is now possible to restore old, black and white images and generate artificial videos with accurate lip syncs.

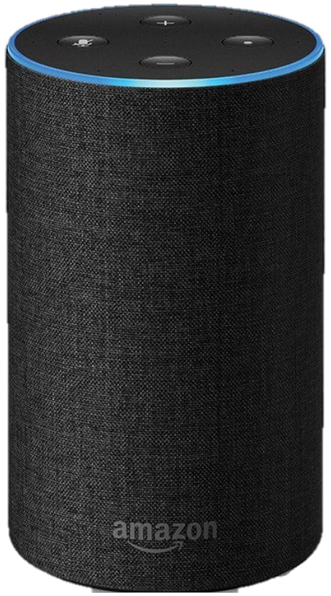


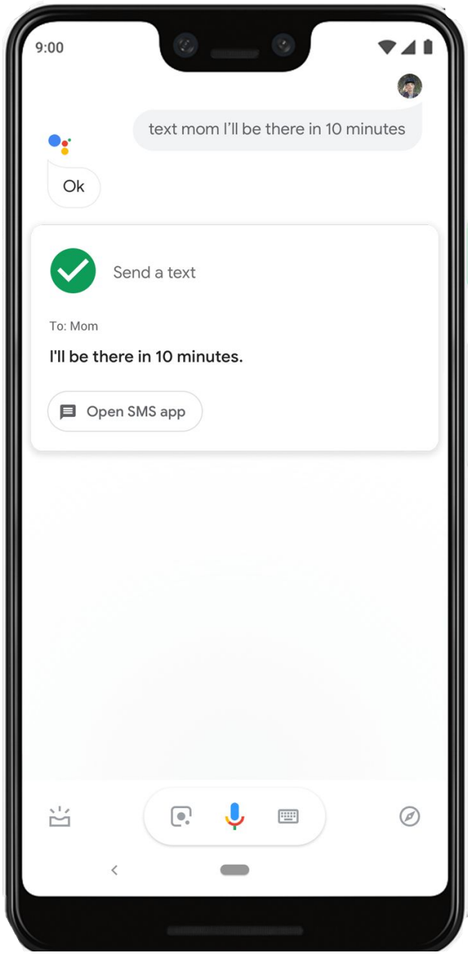
Automatic Colorization

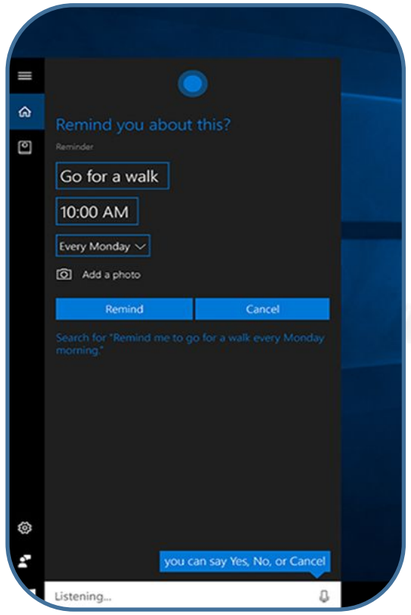
Deep Fakes

Machine Translation

Advanced neural network algorithms are used to synthesize text into a variety of voices and languages.



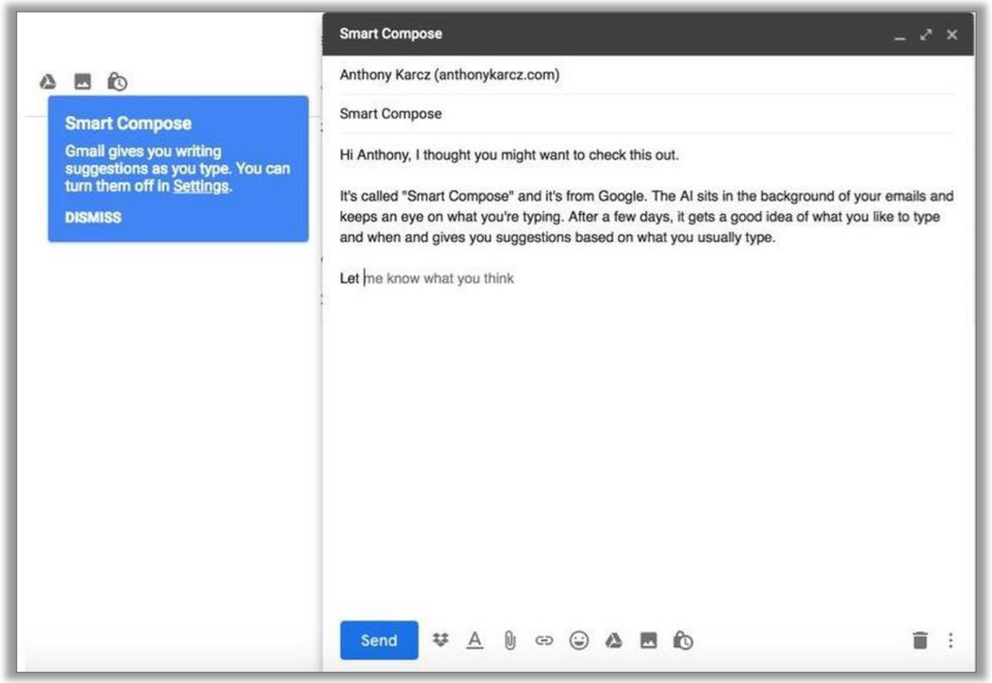




Recent stats show that speech recognition functions on digital assistants work better than before and customers have almost tripled their use of speech interfaces.

Machine Translation

Automatic text generation is used to generate customized texts and translate foreign languages.



Gmail Autocomplete

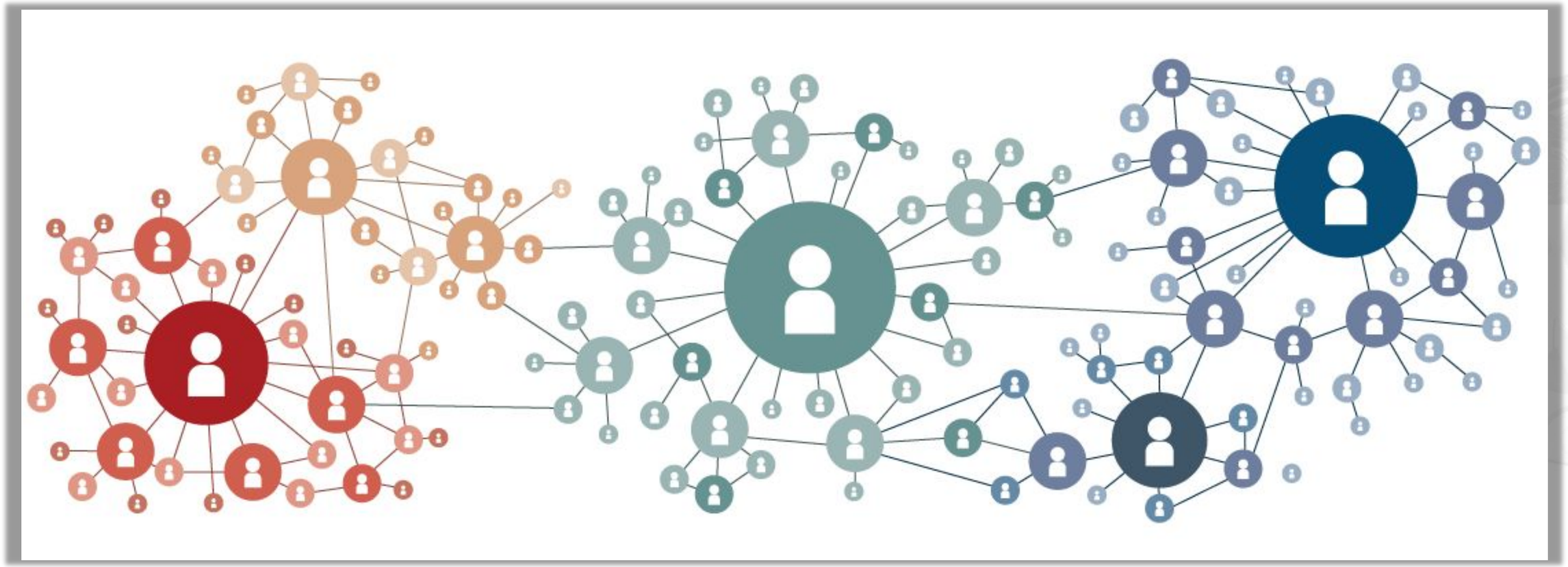
Auto-Translate

Healthcare

DL methods can draw correlations by analyzing millions of data points and enable clinical physicians to keep pace with current research and treatment protocols.

Deep Learning Meets Social Networks

The extensive use of social media platforms creates exciting potential for using neural network models to develop network representations.



Gaming

DL is used to dynamically render finer graphic details in games and help game developers to focus on effective storytelling rather than the game graphics.

Dynamic Graphics Rendering

Neural Net Playing FIFA

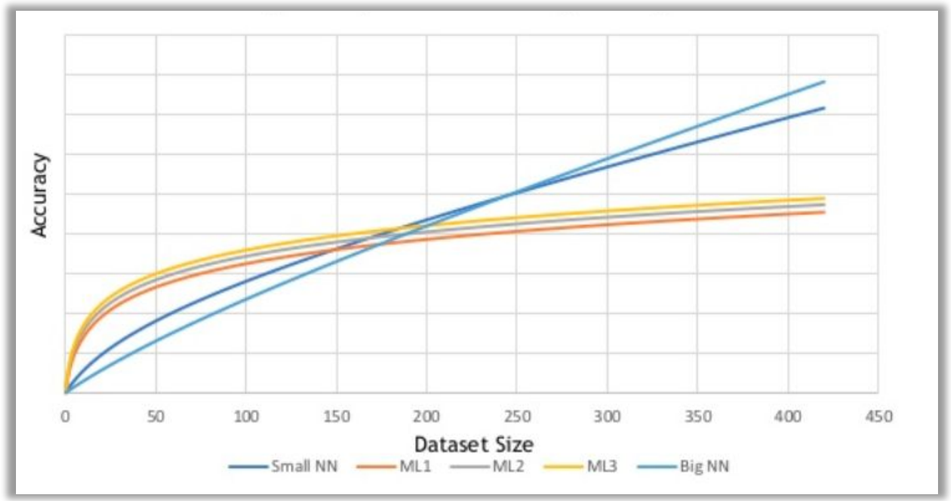
Development of human-like bots is one of the promising DL tools to enrich the gaming experience.

DATA AND
ARTIFICIAL INTELLIGENCE

Challenges in Deep Learning

Training Data

Deep learning gives accurate predictions with right quantity and quality of training data.



Accuracy increasing with dataset size

Building a DL model involves time-consuming tasks like collecting and labeling the training data.

Effective Learning and Teaching

Humans can learn from very few examples whereas, machines need thousands or millions of examples.

We cannot give every possible labeled sample of a problem space to a DL model.
So, the DL model generalizes or interpolates to classify any data not contained in its original dataset.

Understanding Context

Deep learning doesn't understand context very well as it lacks pure perception.

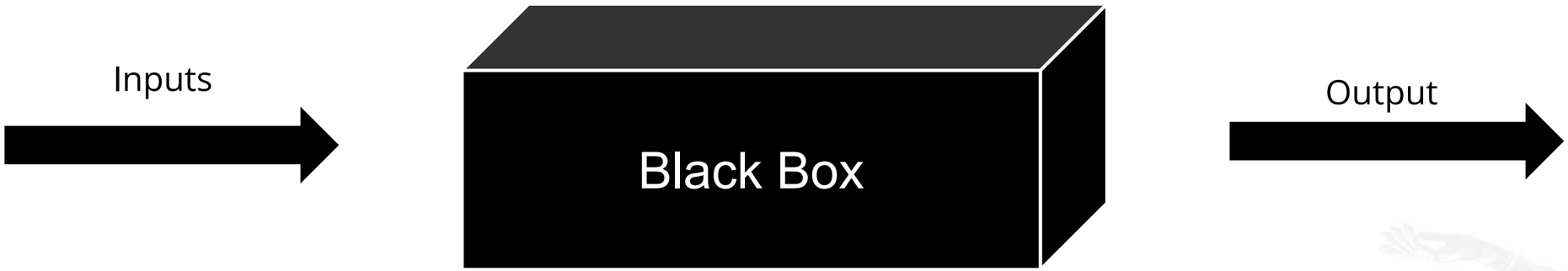
What's on people's minds?

Mirrors

image classification ≠ understanding the scene

The Black Box Problem

Neural networks do not rely on rules established in advance.
They find out patterns and correlations without exposing the reasons leading to them.



Implementation is "opaque" (black)

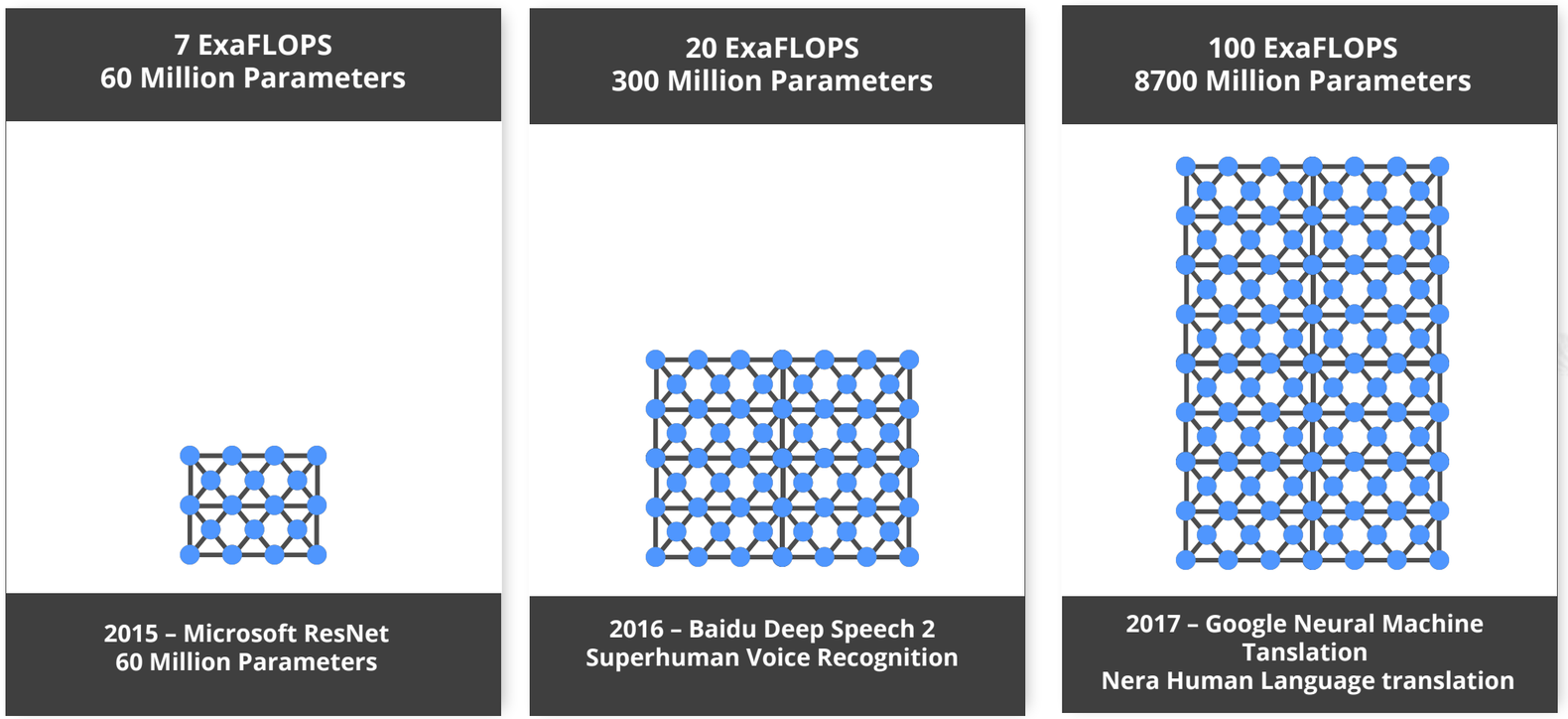
Human users like to understand how a system has arrived at a given decision, as decisions are potential liabilities for domains such as finance and medicine.

Large Size Model and Complexity

State-of-the-art deep learning models are multi-gigabytes in size and continue getting larger.

NEURAL NETWORK COMPLEXITY IS EXPLODING

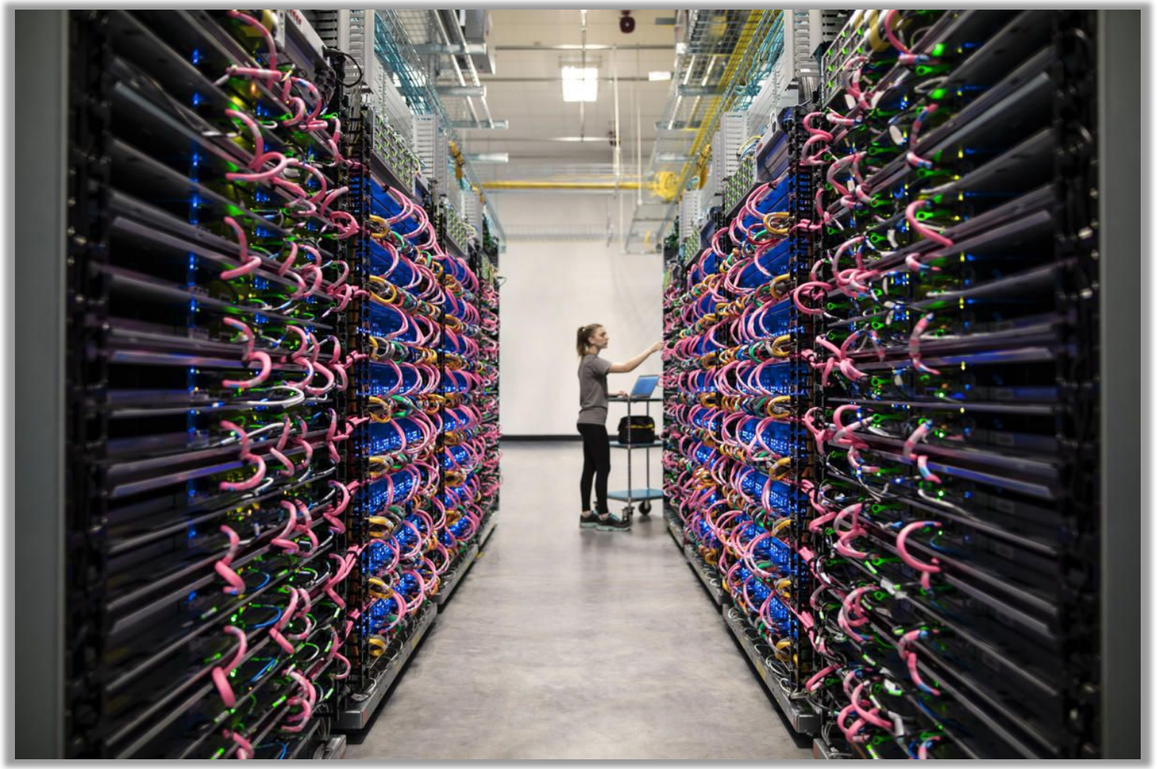
To Tackle Increasingly Complex Challenges



Number of parameters are directly proportional to the amount of information absorbed by the neural net.

Infrastructure

Majority of the useful deep learning problems activate the entire neural network for each batch size thereby blowing up the computation costs.



Therefore, models are loaded and scaled on multiple machines.

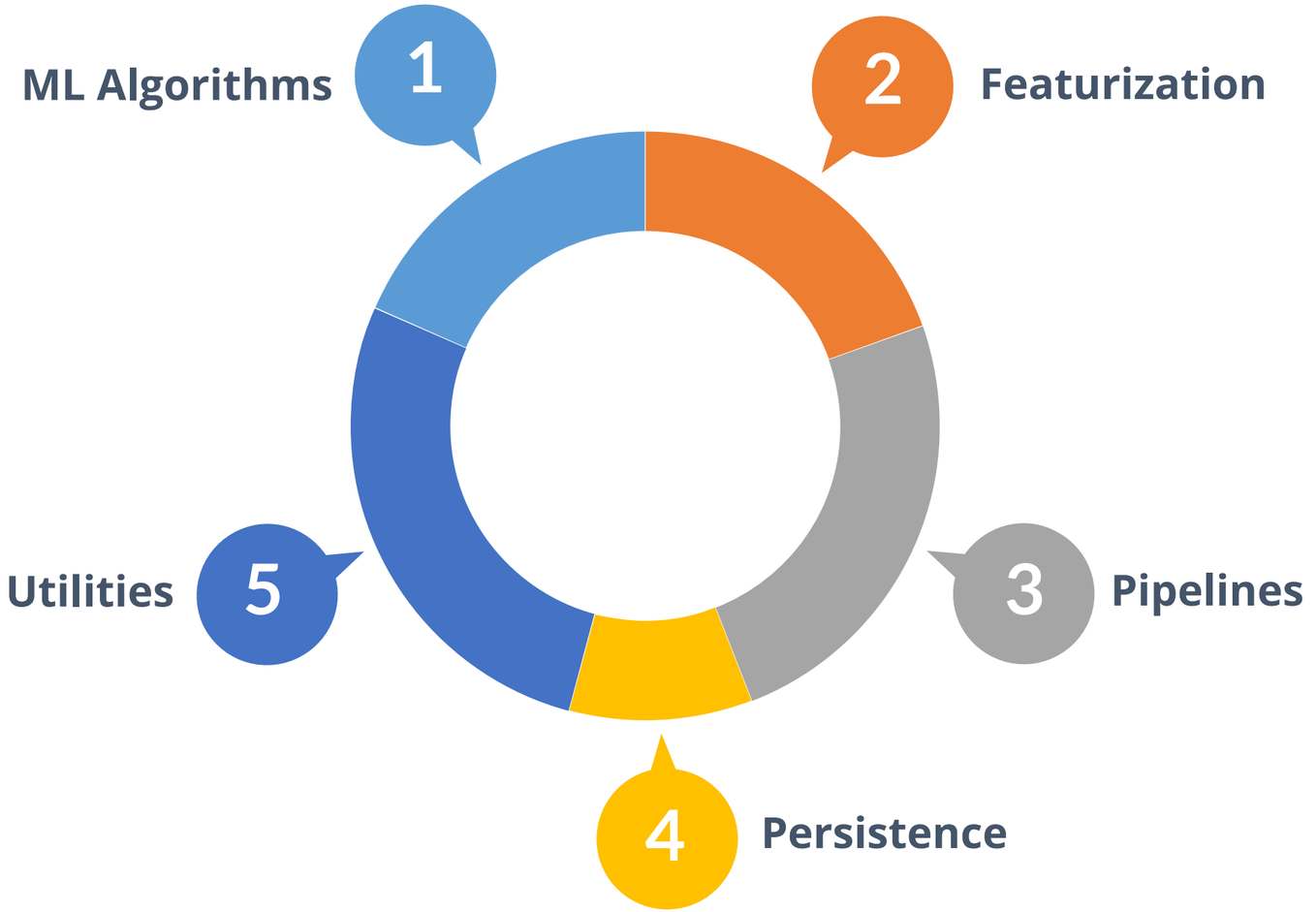
DATA AND
ARTIFICIAL INTELLIGENCE

Overview of MLlib

What Is MLlib?

MLlib is Spark's scalable machine learning library consisting of common learning algorithms, utilities, and optimization primitives.

MLlib Tools

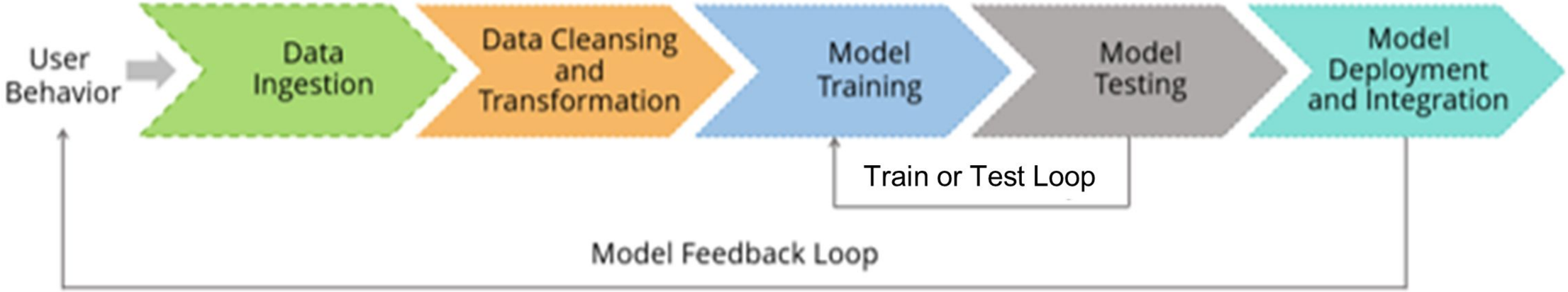


MLlib Algorithms

DATA AND
ARTIFICIAL INTELLIGENCE

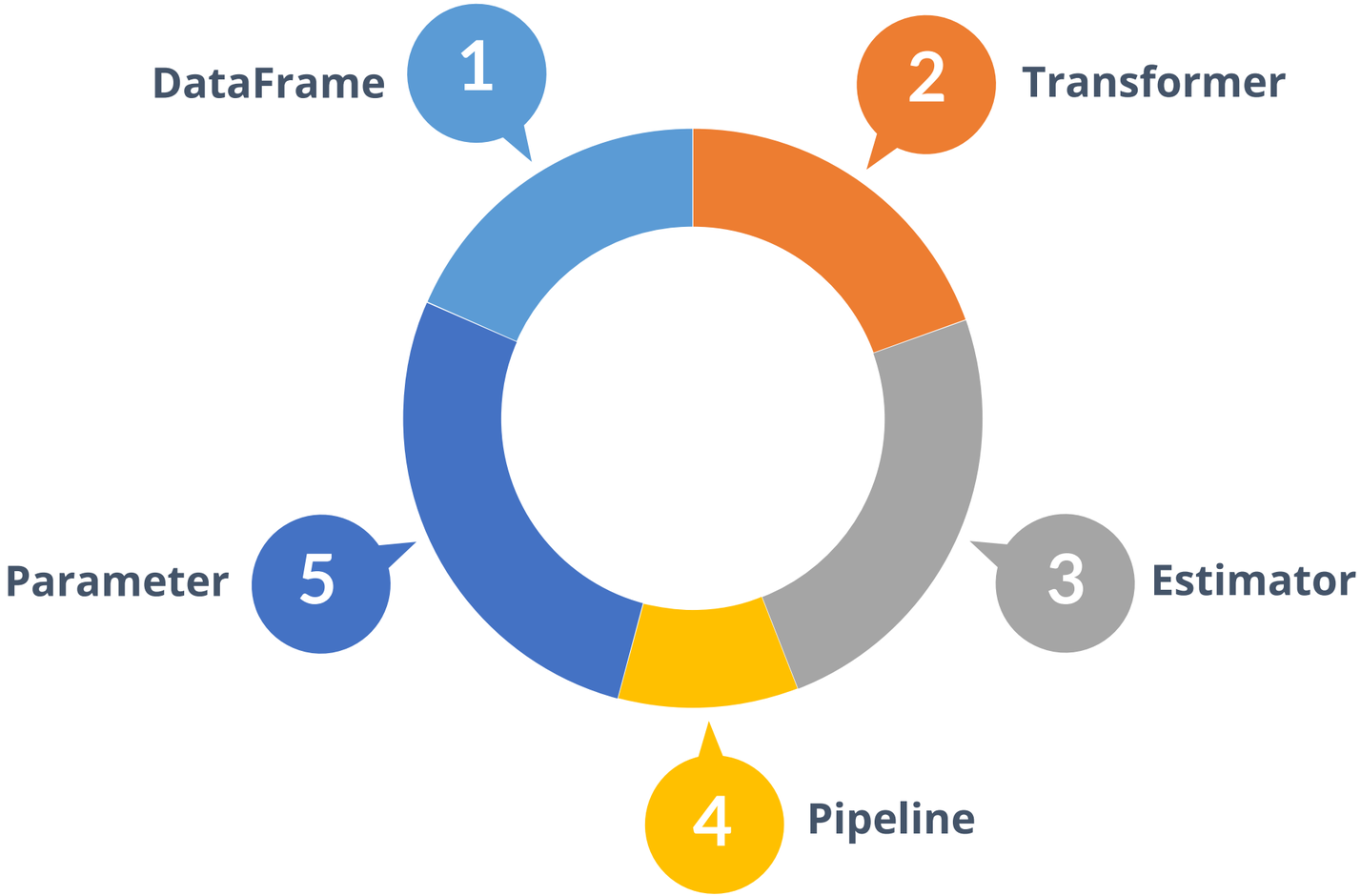
MLlib Pipelines

Machine Learning Pipeline



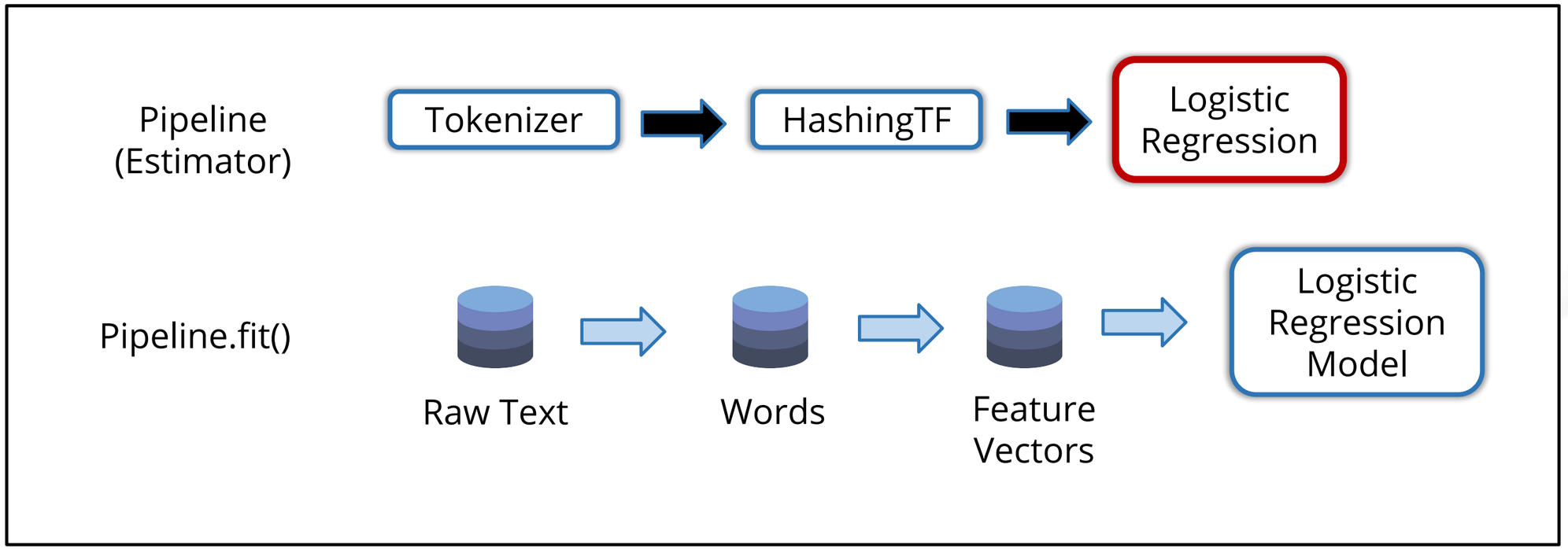
MLlib Pipelines

ML Pipelines provide a uniform set of high-level APIs that help users create and tune practical machine learning pipelines.

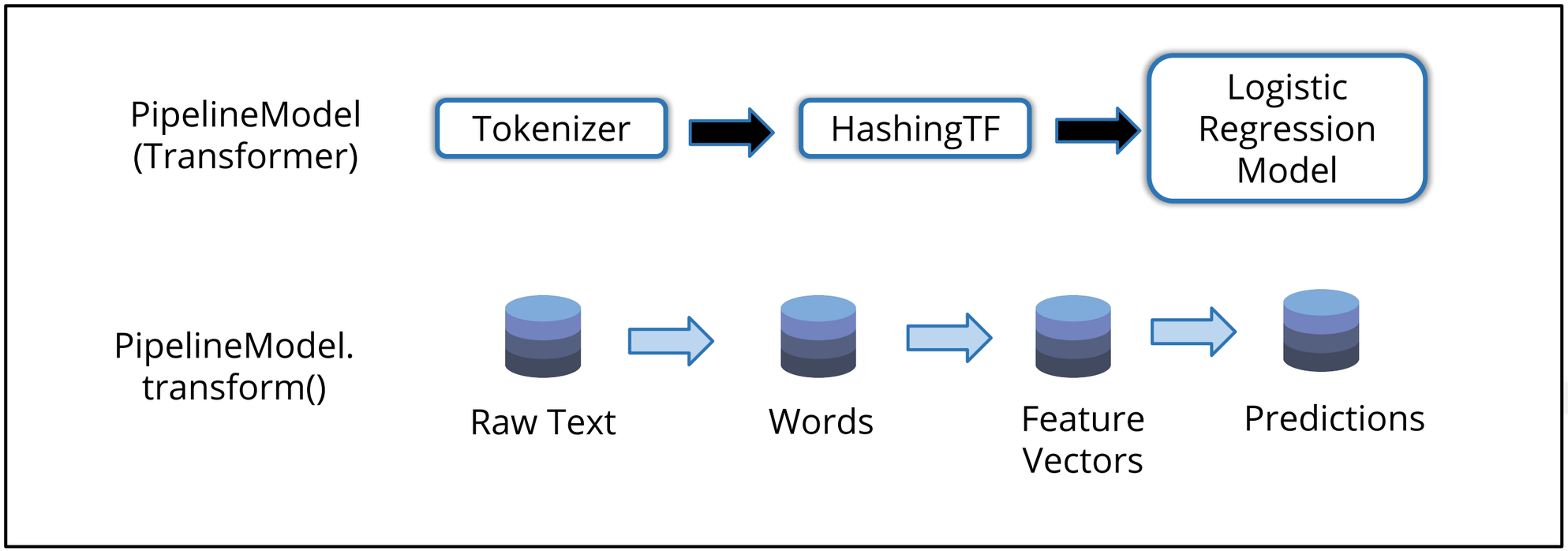


Working of Pipeline

Pipeline is specified as a sequence of stages, and each stage is either a Transformer or an Estimator.



Working of Pipeline

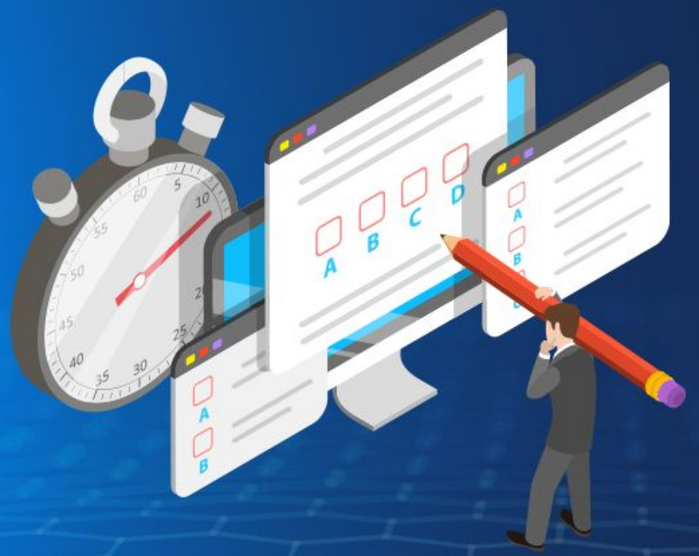


Key Takeaways

You are now able to:

- Identify the skills required to become a data scientist and data analyst
- Define analytics in Spark and list the types of analytics
- Describe the machine learning algorithms
- Define MLlib and MLlib pipeline

DATA AND
ARTIFICIAL INTELLIGENCE



Knowledge Check

Knowledge Check 1

Which of the following skills are required to become a data scientist?

a. Knowledge of Python and R

b. Ability to work with unstructured data

c. Experience in SQL

d. All of the above

Knowledge Check 1

Which of the following skills are required to become a data scientist?

a. Knowledge of Python and R

b. Ability to work with unstructured data

c. Experience in SQL

d. All of the above

The correct answer is **d**

Knowledge of Python and R, ability to work with unstructured data, and experience in SQL are required to become a data scientist.

Knowledge Check 2	**Which of the following types of analytics describes the past and answers the question, “What has happened?”?**

a. Descriptive analytics

b. Predictive analytics

c. Prescriptive analytics

d. None of the above

Knowledge Check 2

Which of the following types of analytics describes the past and answers the question, “What has happened?”?

a. Descriptive analytics

b. Predictive analytics

c. Prescriptive analytics

d. None of the above

The correct answer is **a**

Descriptive analytics describes the past and answers the question, “What has happened?”.

Knowledge Check 3

Which machine learning algorithm develops a self-sustained system based on the interaction between the environment and the learning agent?

a. Supervised learning

b. Unsupervised learning

c. Reinforcement learning

d. None of the above

Knowledge Check 3

Which machine learning algorithm develops a self-sustained system based on the interaction between the environment and the learning agent?

a. Supervised learning

b. Unsupervised learning

c. Reinforcement learning

d. None of the above

The correct answer is **c**

Reinforcement learning algorithm develops a self-sustained system based on the interaction between the environment and the learning agent.

Knowledge Check 4

Which MLlib algorithm is a statistical process for estimating the relationships among variables?

a. Classification

b. Regression

c. Clustering

d. Optimization

Knowledge Check 4

Which MLlib algorithm is a statistical process for estimating the relationships among variables?

a. Classification

b. Regression

c. Clustering

d. Optimization

The correct answer is **b**

Regression algorithm is a statistical process for estimating the relationships among variables.

Lesson-End Project

Problem Statement: The Blue Nile is one of the largest diamond retail e-commerce companies in the world. The company's revenue numbers are good this year. Recently a distributor was shutting down its business due to his financial instability and he wants to sell all of the diamonds in an open auction. This is a great opportunity for the company to expand its diamond inventory and wants to participate in the bidding. To make sure your bid is mostly accurate you must predict the correct price of the diamond.
To predict the price you should have a machine learning model and to build that you need to have the correct data first. You have collected the diamond data with all the possible diamond features concerned with deciding the price of the diamond.

Lesson-End Project

You have created a diamond.csv file which has the following details:
1. Index: counter
2. Carat: Carat weight of the diamond
3. Cut: Describe the cut quality of the diamond. Quality in increasing order: Fair, Good, Very Good, Premium, Ideal
4. Color: Color of the diamond, with D being the best and J the worst
5. Clarity: How obvious are the inclusions are in the diamond: (in order from best to worst, FL = flawless, I3= level 3 inclusions)
6. Depth: depth %: The height of the diamond, measured from the culet to the table, divided by its average girdle diameter
7. Table: table%: The width of the diamond's table expressed as a percentage of its average diameter
8. Price: the price of the diamond
9. X: length mm
10. Y: width mm
11. Z: depth mm

As a business analyst of the company, you are assigned the task of recommending the bid amount for the diamond that the company should bid for using the model built by the analytics team.

DATA AND
ARTIFICIAL INTELLIGENCE