Machine Learning Algorithms

With Python

How to NOT learn Machine Learning

1. Get really good at Python programming and Python syntax.
2. Deeply study the underlying theory and parameters for machine learning algorithms in

scikit-learn.

3. Avoid or lightly touch on all of the other tasks needed to complete a real project.

Making Predictions

- This class focuses on the Machine Learning Sub-Field called “predictive modeling”
- It is the field of Machine Learning that is used the most in industry

The Ernest.Net Method (6 of the 9 steps)

1. **Define Problem**: Investigate and characterize the problem in order to better understand

the goals of the project.

2. **Analyze Data**: Use descriptive statistics and visualization to better understand the data you have available.

3. **Prepare Data**: Use data transforms in order to better expose the structure of the prediction problem to modeling algorithms.

4. **Evaluate Algorithms**: Design a test harness to evaluate a number of standard algorithms

on the data and select the top few to investigate further.

5. **Improve Results**: Use algorithm tuning and ensemble methods to get the most out of well-performing algorithms on your data.

6. **Present Results**: Finalize the model, make predictions and present results.

Outline of the Course

. Section 1: Python Ecosystem for Machine Learning.

. Section 2: Python and SciPy Crash Course.

. Section 3: Load Datasets from CSV.

. Section 4: Understand Data With Descriptive Statistics. (Analyze Data)

. Section 5: Understand Data With Visualization. (Analyze Data)

. Section 6: Pre-Process Data. (Prepare Data)

1.2. Machine Learning in Python 4

. Section 7: Feature Selection. (Prepare Data)

. Section 8: Resampling Methods. (Evaluate Algorithms)

. Section 9: Algorithm Evaluation Metrics. (Evaluate Algorithms)

. Section 10: Spot-Check Classication Algorithms. (Evaluate Algorithms)

. Section 11: Spot-Check Regression Algorithms. (Evaluate Algorithms)

. Section 12: Model Selection. (Evaluate Algorithms)

. Section 13: Pipelines. (Evaluate Algorithms)

. Section 14: Ensemble Methods. (Improve Results)

. Section 15: Algorithm Parameter Tuning. (Improve Results)

. Section 16: Model Finalization. (Present Results)

Datasets used in the class

- (Iris flowers dataset) : This is a quick pass through the project steps without much tuning or optimizing on a dataset that is widely used as the hello world of machine learning.
- (Boston House Price dataset) : Work through each step of the project process with a regression problem.
- (Sonar dataset) : Work through each step of the project process using all of the methods on a binary classification problem.

Learning Outcomes

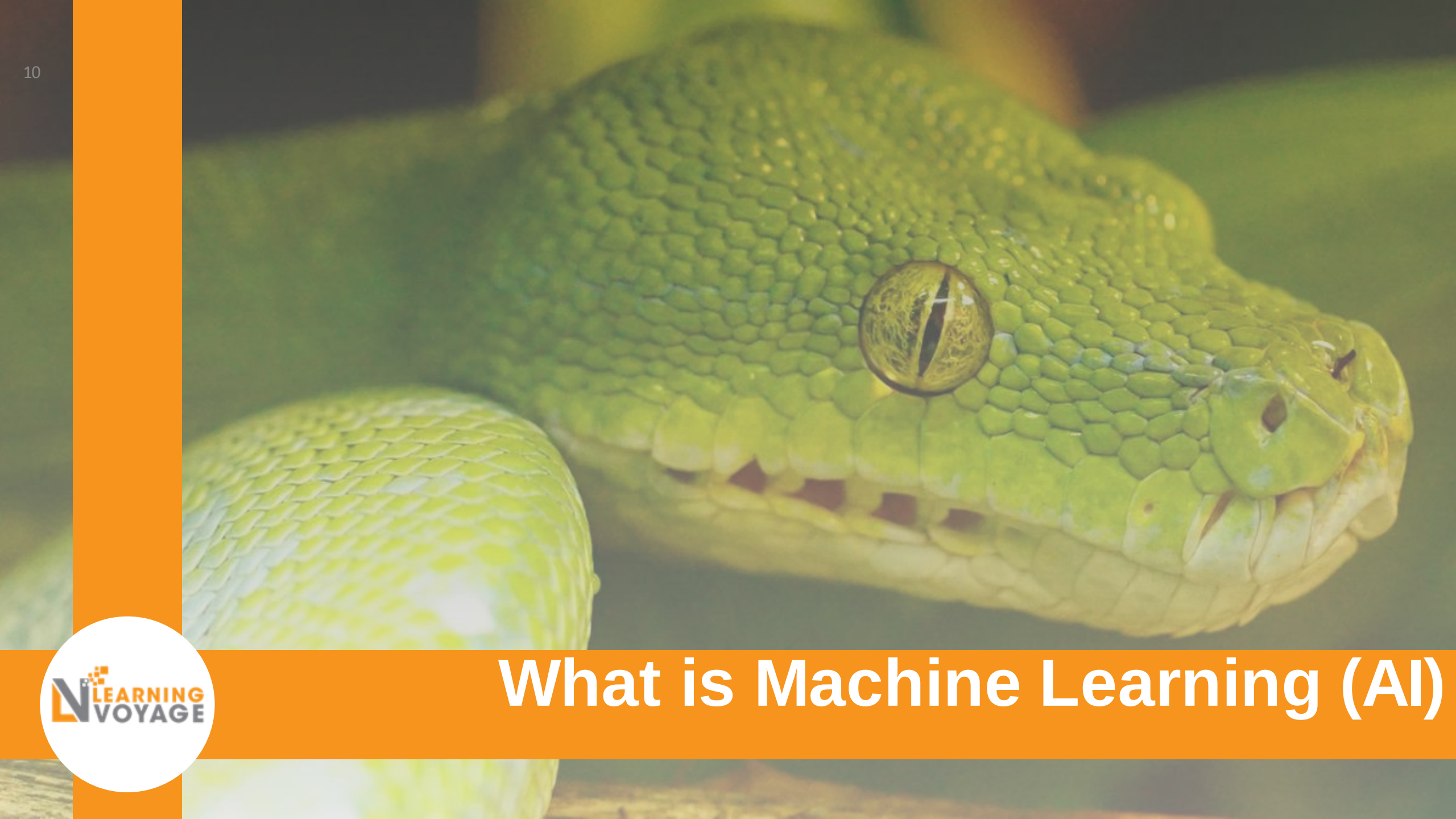
- How to work through a small to medium sized dataset end-to-end.
- How to deliver a model that can make accurate predictions on new data.
- How to complete all subtasks of a predictive modeling problem with Python.
- How to learn new and different techniques in Python and SciPy.
- How to get help with Python machine learning.

What this course is NOT

- This is not a machine learning college course
- This is not a deep algorithm or math class.
- This is not a Python programming class

Let's get started

- Day 1:
  - Part I : What is AI/Machine Learning
  - Part II : Data, Data, Data
- Day 2:
  - Machine Learning Project Template
  - Work through several ML use cases end to end



What is Machine Learning (AI)

Learning Goals

1.1 Background on Artificial Intelligence and Machine Learning

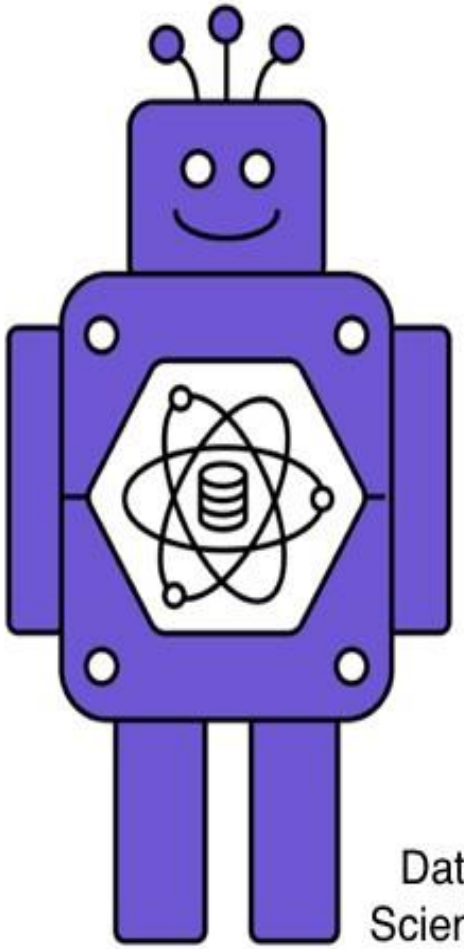
1.2 Types of Machine Learning

1.3 Explore Real-World Use Cases

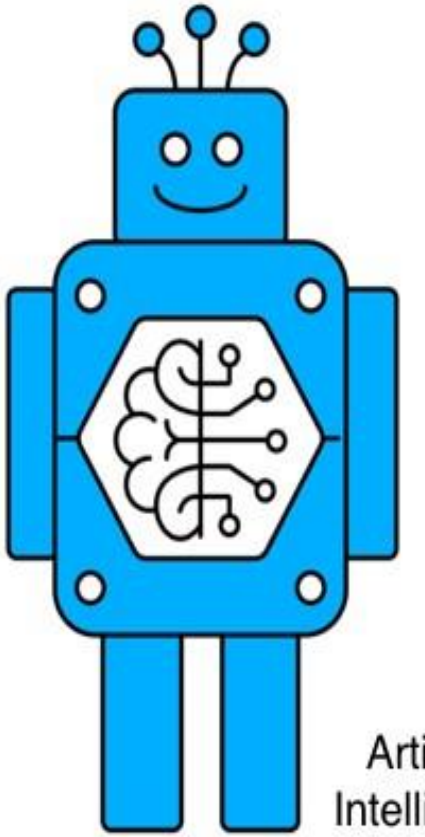
The Future

https://vimeo.com/279683304

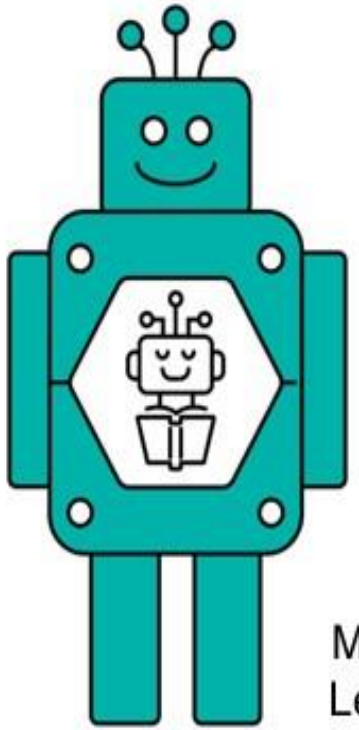
The Data Science/Machine Learning Landscape



Data Science



Artificial Intelligence



Machine Learning

What is Data Science?

- Extract insights from data
- Analyzes large volumes of data of any format/type
- Applicable across all industries

Includes:

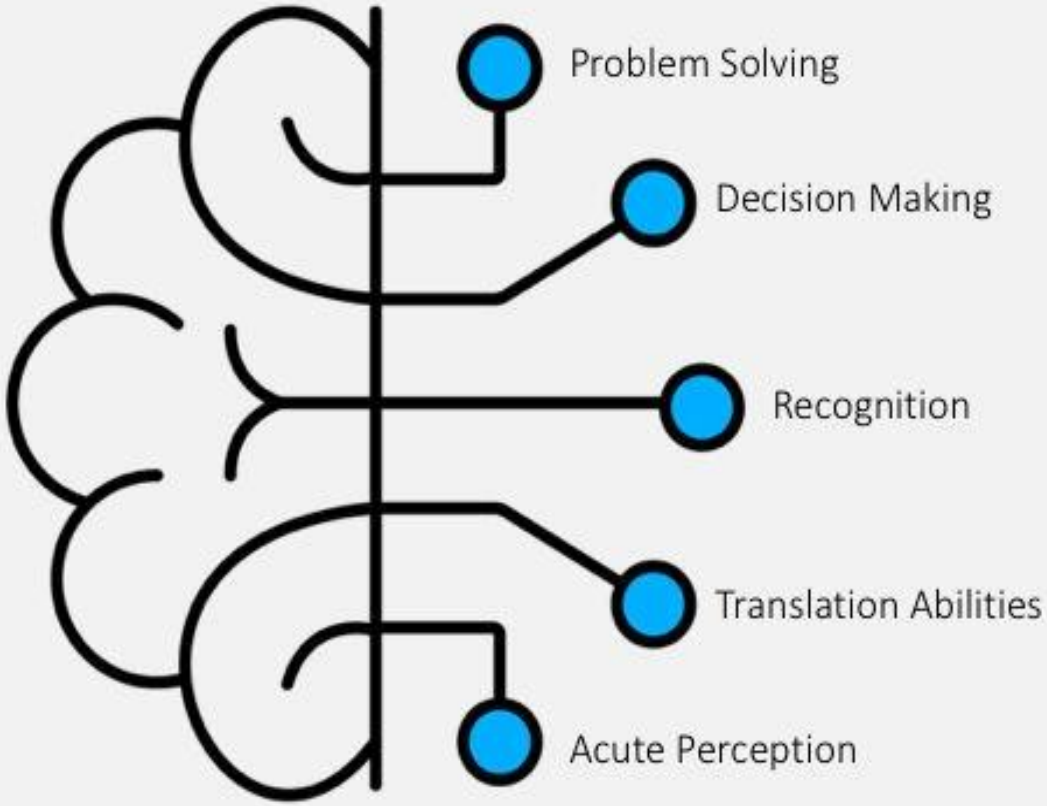
- Artificial Intelligence (AI)
- Machine Learning (ML)

What is Artificial Intelligence?

- Simulation of intelligent human behavior

Includes:

- Symbolic AI and Expert Systems
- AI Planning
- Machine Learning



Symbolic Artificial Intelligence

Symbols are physical patterns, or representations, of an abstract concept that get combined into structures, and then are manipulated to create new expressions

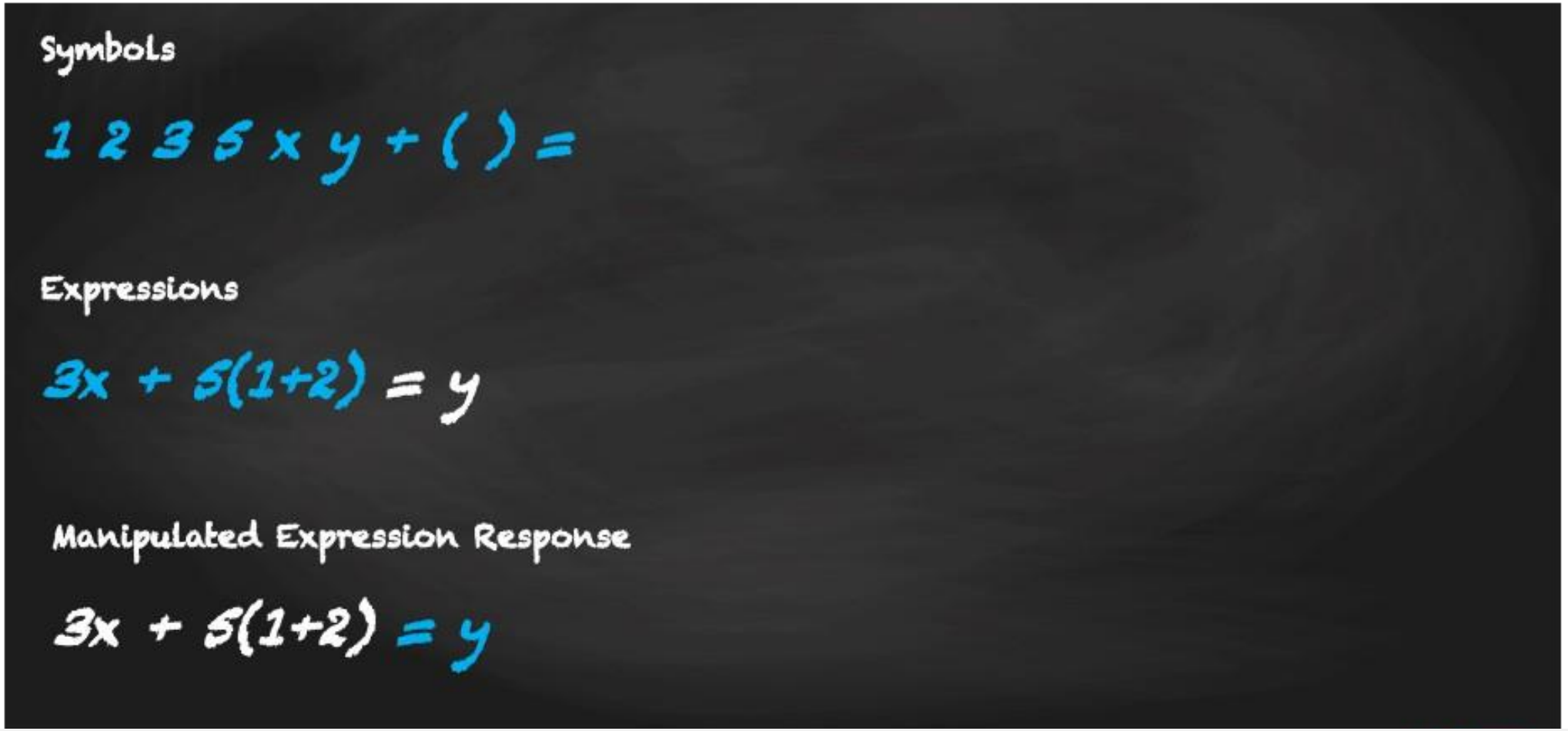
Physical Symbol Systems (Formal Systems)

Symbols
- Encoded in our brains
- Example: This joining of two perpendicular lines is a plus sign

Thoughts
- Structures or expressions
- Example: The plus sign means to add something

Thinking
- Applying the symbol and the structure together
- Example: 1 + 2 = 3

Physical Symbol System Example: Algebra



Physical Symbol System Example: Formal Logic

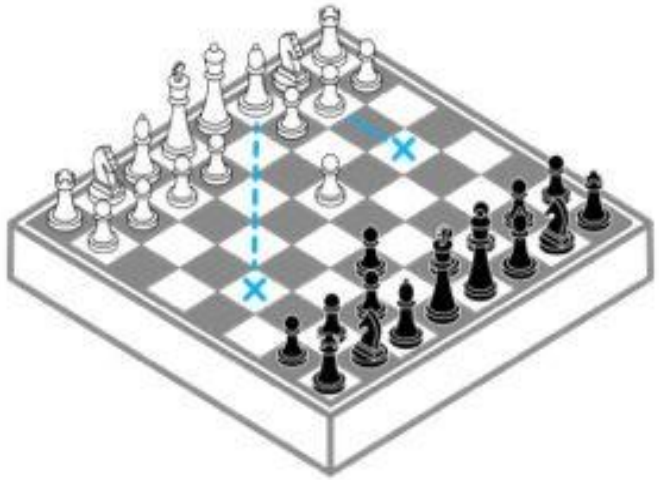
Symbols:
If, or, not, and, then

Expressions:
If ..., then...
Or true false statements

Manipulated Expression Response
If ..., then..., this…

Physical Symbol System Example: Chess

Symbols
Defined number of pieces

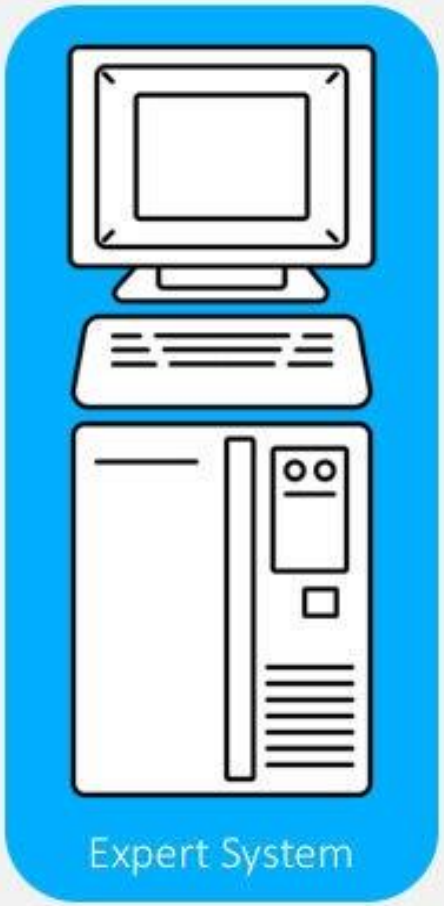


Expressions/Structures
Legal chess moves

Manipulated Expression Response
Resulting position of the pieces on the board after each move

Expert Systems

An expert system is a computer system that is designed to solve problems by imitating the decision making abilities of a human expert



Input Data from Trainer

Knowledge Base
- Input data
- Production/If-Then rules

Inference Engine
- Applies knowledge base rules to the data
- Derives new facts

New Facts

AI Planning

Another branch of classic AI is automated planning and scheduling, also referred to as AI Planning

Artificial Intelligence Examples

AI varies greatly in its potential complexity. Every day we get closer and closer to this type of reality. Mainly, this is because we are now teaching machines how to learn, and grow, on their own.

A pile of if-then statements

Complex statistical model mapping raw sensory data to symbolic categories

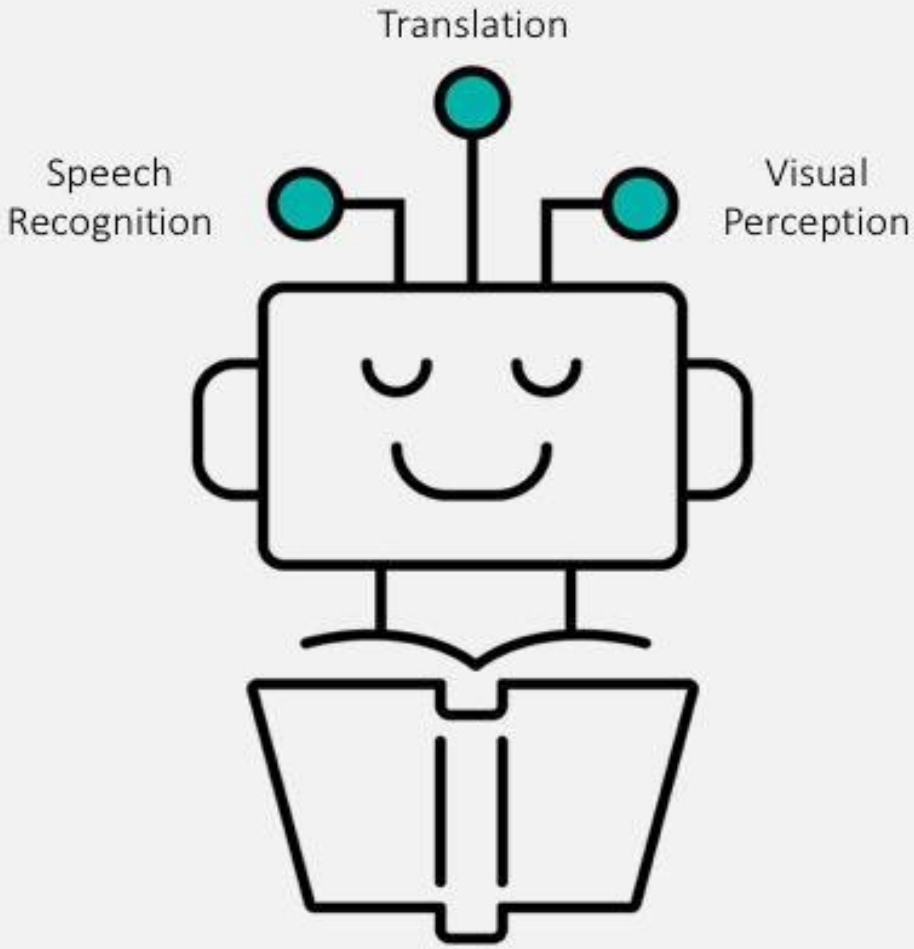
Super-fancy computer systems, specialized robots, and advanced androids

What is Machine Learning?

- Machines are taught to learn from data and make decisions
- Provide quick results for fast decision making

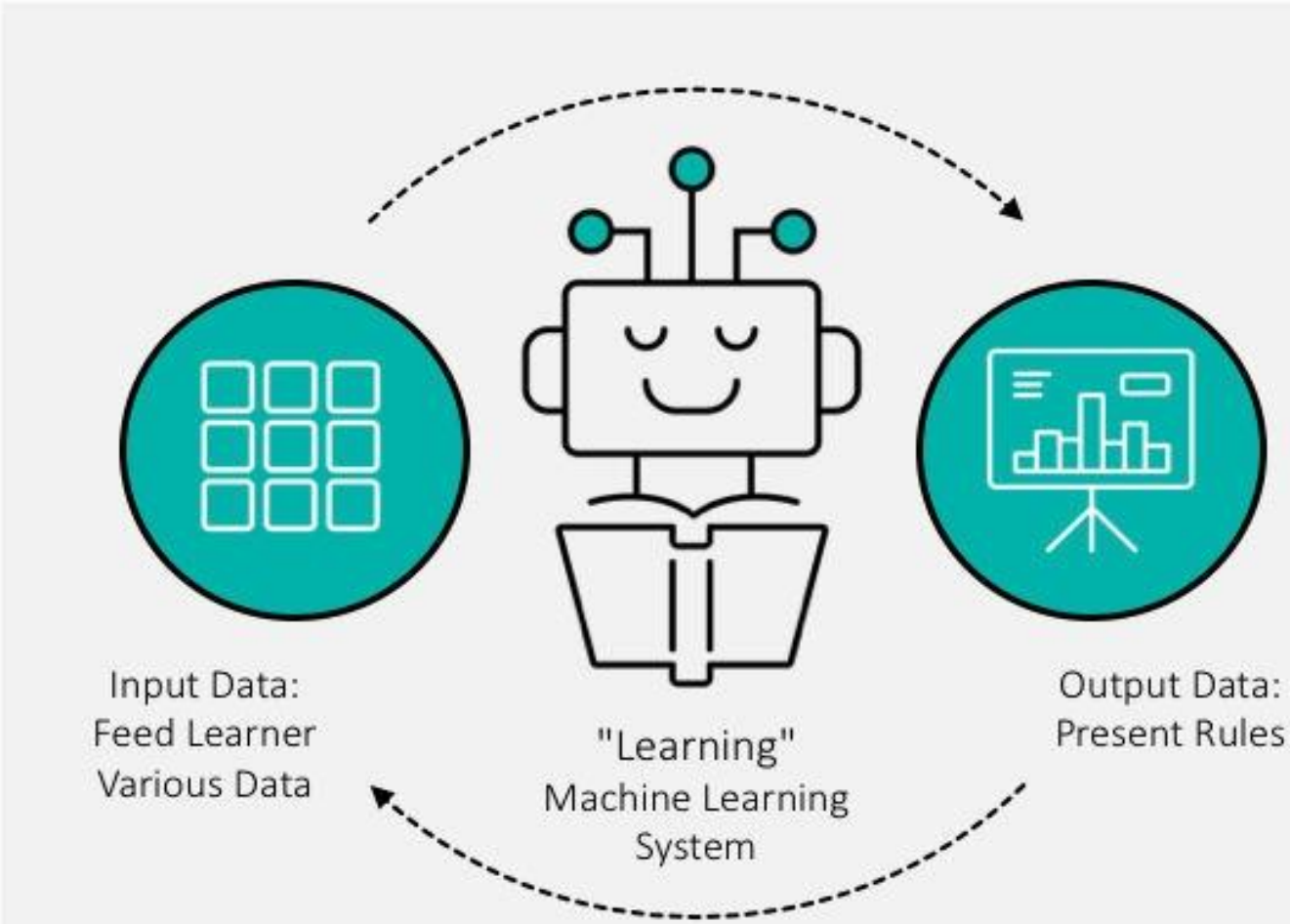
Includes:

- Deep Learning



Why is it Important?

- Learns autonomously through a dynamic feedback loop
- Increasingly self-healing, self-organizing, and self-architecting

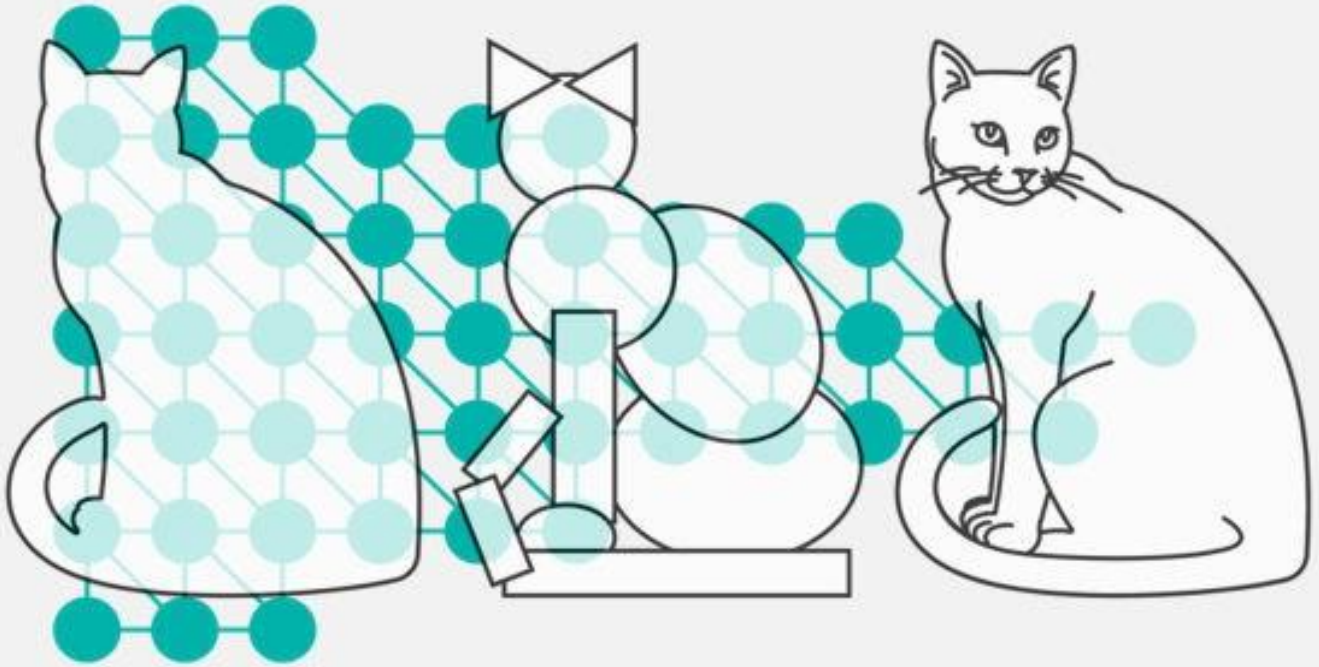


What is Deep Learning?

- Learning is done in a hierarchy of layers
- Modeled after the brain's neural networks
- "Deep" describes the number of layers used

Known for:

- Speech and image recognition
- Language processing



First Layer: Outlines

Next Layer: Shapes

Final Layer: Features

Machine Learning Today

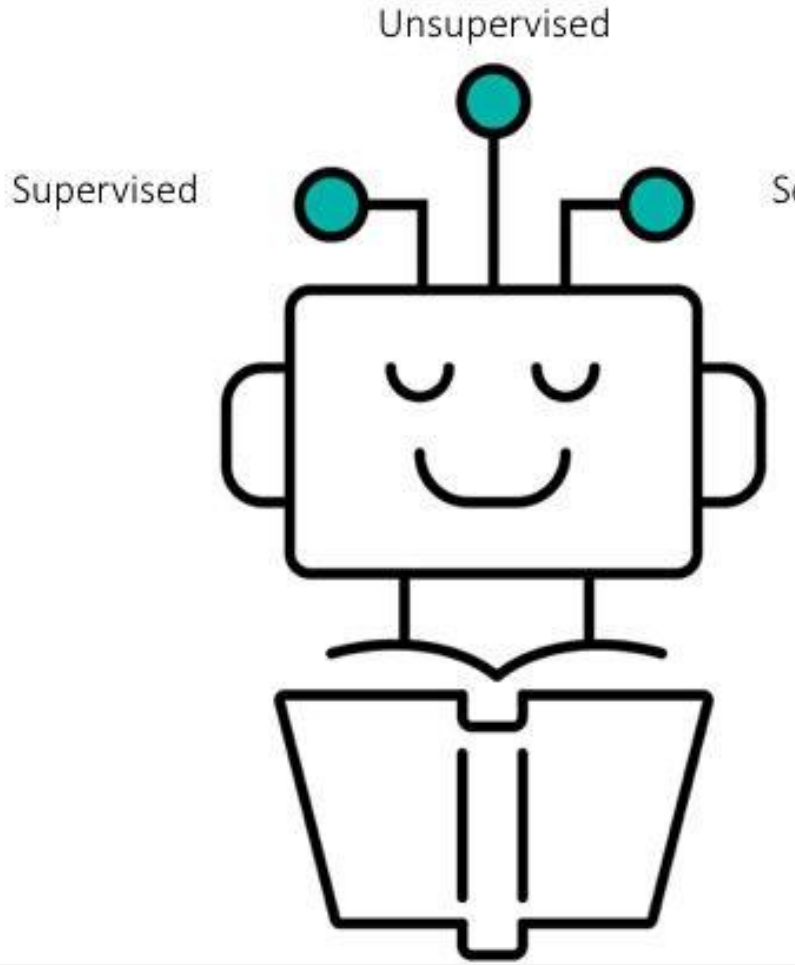
Learning Goals

1.1 Background on Artificial Intelligence and Machine Learning

1.2 Types of Machine Learning

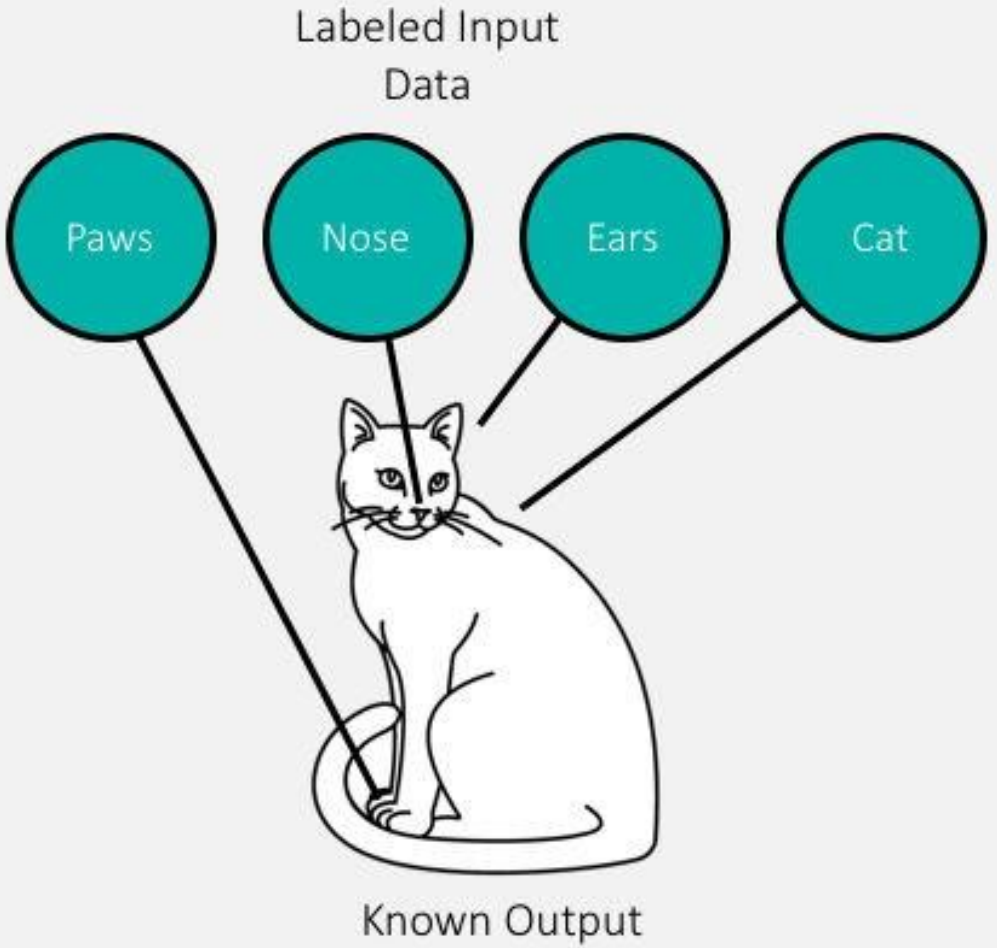
1.3 Explore Real-World Use Cases

Types of Machine Learning

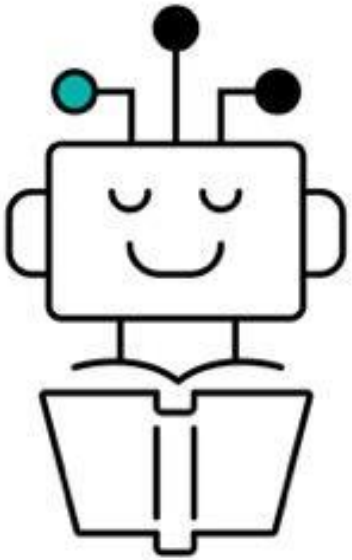


Supervised Learning: Defined

- Uses labeled training data
- Learns relationships between given inputs to a given output
- Desired output must be part of the labeled data, to be known



Supervised Learning: How it Works

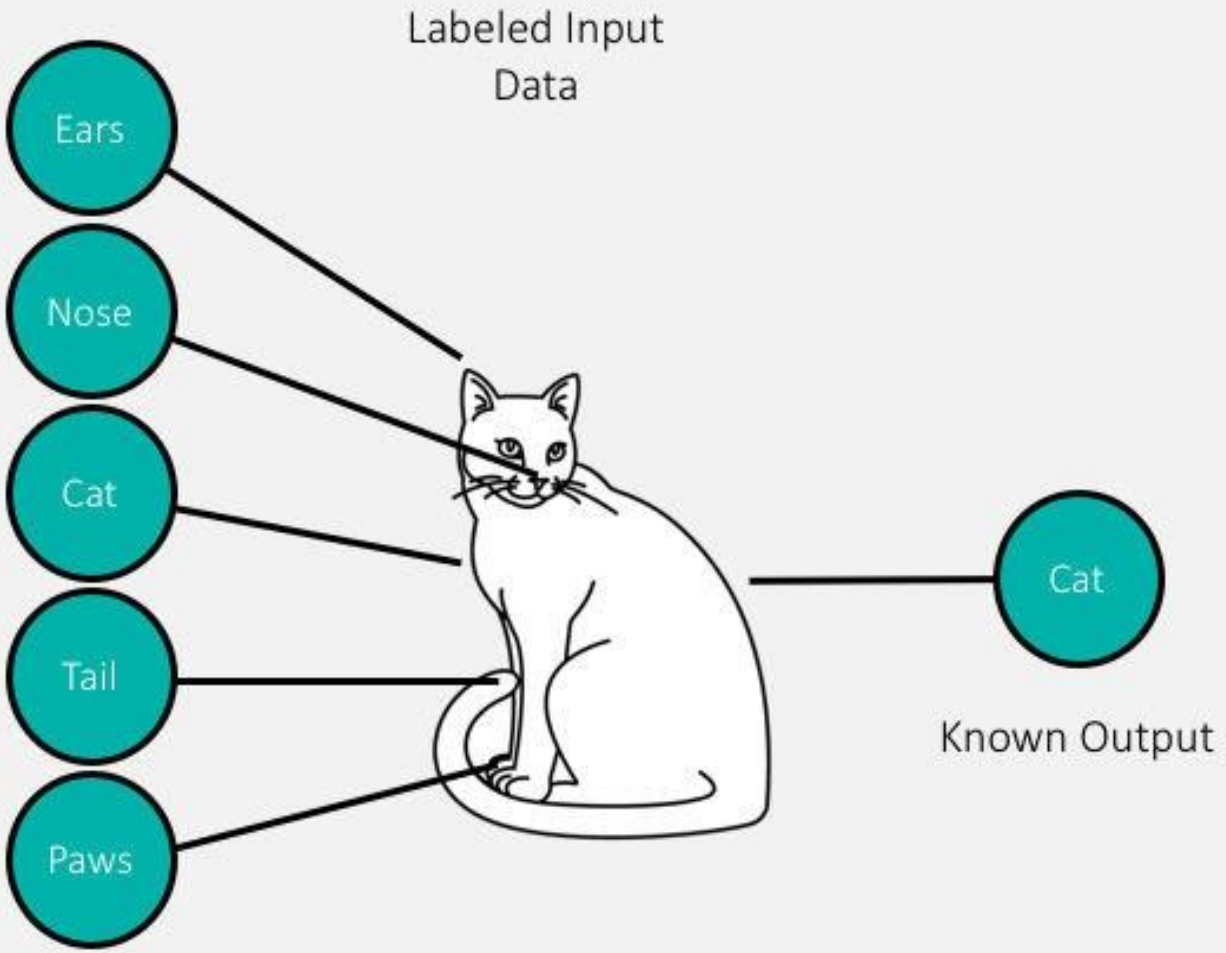


Supervised Learning

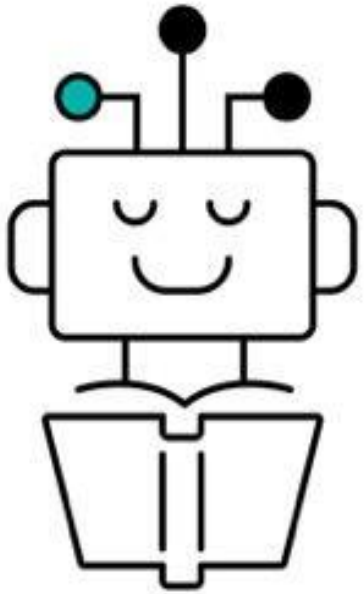
1. Load labeled input data
2. Train model on the data: connection to input variables and output is made
3. Apply new data to algorithm
4. Provides output

Supervised Learning: How it Works Example

Supervised learning uses labeled training data to train machines to learn relationships between given inputs to a given output



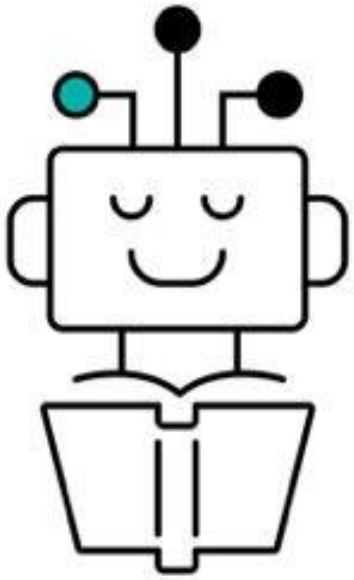
Supervised Learning: Pros and Cons



Supervised Learning

Pros	Cons
Very clear objective	Often labor-intensive
Easy to measure accuracy	Limited data to work with
Controlled training	Limited insights

Supervised Algorithms: Classification

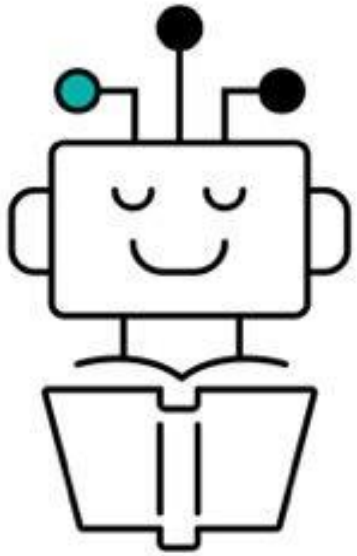


Supervised Learning

Two general supervised learning categories:

Type	Algorithm or Task
Supervised	Classification (used to predict a categorical result)
Supervised	Regression (used to predict the output value given the input value)

Supervised Algorithms: Classification Use Cases



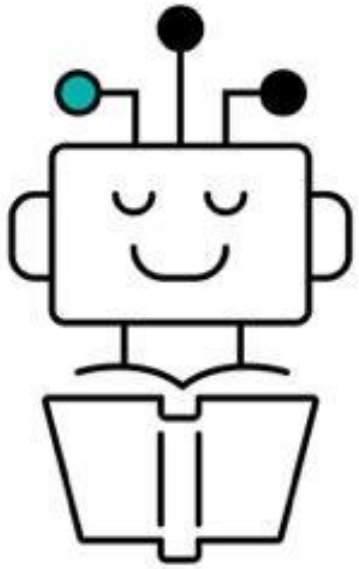
Supervised Learning

Use case examples for supervised learning algorithms include these simple binary classification problems:

Fraud Detection

Spam Filtering

Supervised Algorithms: Regression

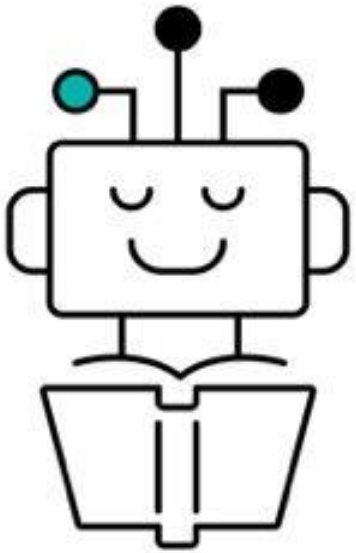


Supervised Learning

Regression algorithms are used to predict a continuous numeric result.

Type	Algorithm or Task
Supervised	Classification (used to predict a categorical result)
Supervised	Regression (used to predict the output value given the input value)

Supervised Algorithms: Regression Use Cases



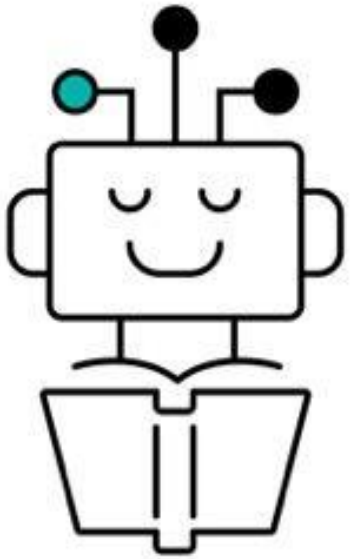
Supervised Learning

Use case example for a supervised learning regression problem include:

Credit Card Model Assessment

Customer/Employee Churn

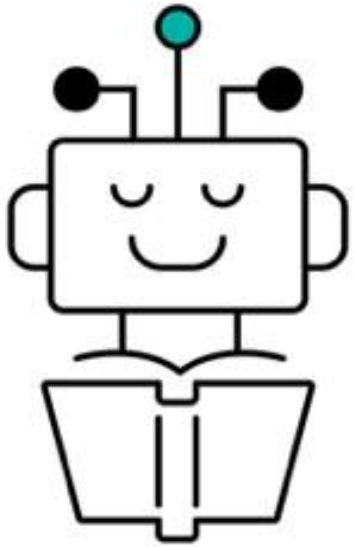
Supervised Algorithms Table



Supervised Learning

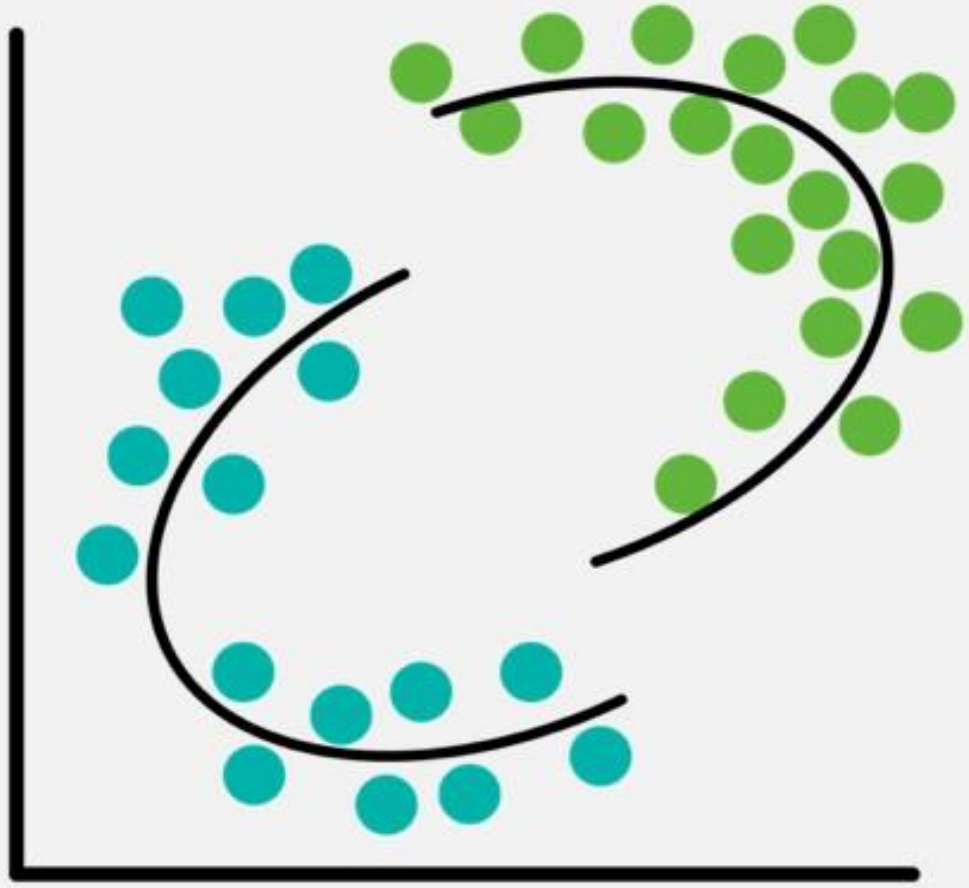
Type	Algorithm or Task
Classification	Naïve Bayes
Classification	Logistic Regression
Classification	Support Vector Machines (SVM)
Regression	Linear Regression
Both	Decision Trees/Random Forest
Both	K-Nearest Neighbor (KNN)
Both	Gradient Boosting Algorithms (GBA)

Unsupervised Learning Defined

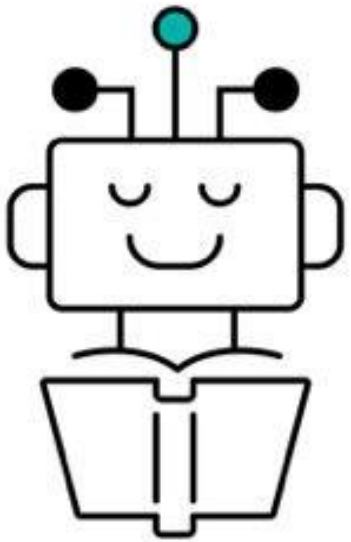


Unsupervised Learning

Unsupervised learning uses raw, unlabeled data and has no knowledge of the output label.



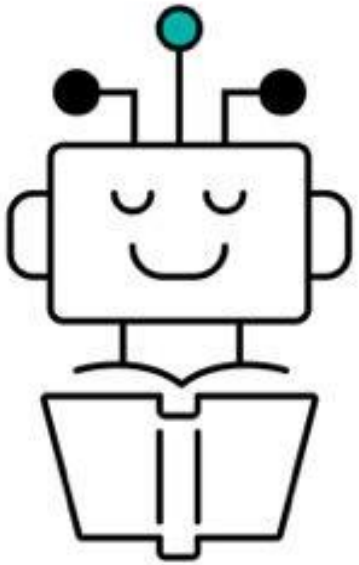
Unsupervised Learning: How it Works



Unsupervised Learning

1. Load unlabeled raw data
2. Algorithm infers patterns from the data on its own
3. Algorithm identifies groups of data that exhibit similar patterns
4. Provides output

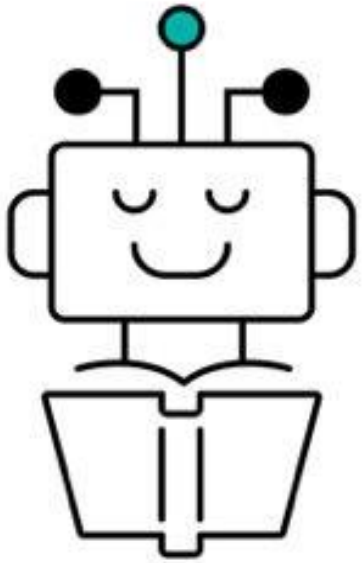
Unsupervised Learning: Pros and Cons



Unsupervised Learning

Pros	Cons
Very fast to start	Difficult to measure accuracy
Disruptive insights	Requires more experience
	Curse of dimensionality

Unsupervised Algorithms: Common Use Case

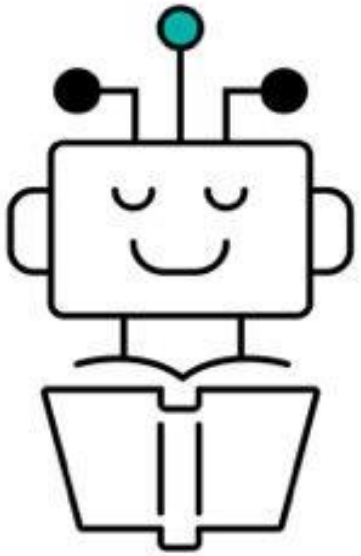


Unsupervised Learning

Use case example for an unsupervised learning problem includes cluster analysis

Security Anomaly Detection

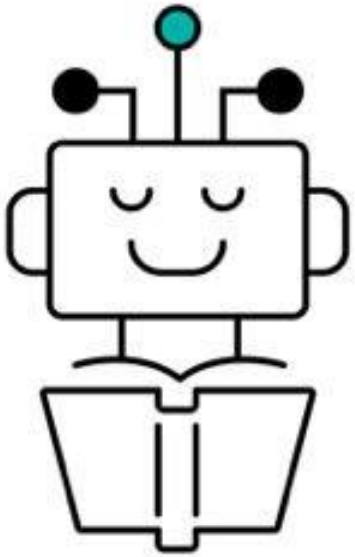
Unsupervised Learning: Use Case Example



Unsupervised Learning

Using Anomaly Detection

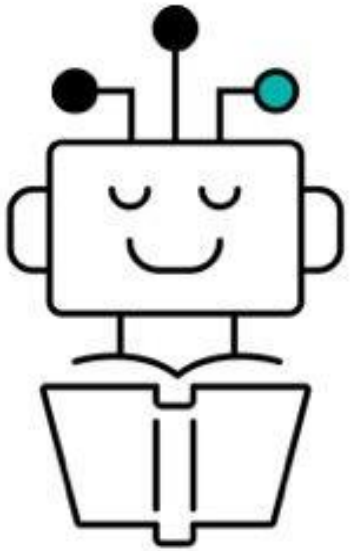
Unsupervised Algorithms Table



Unsupervised Learning

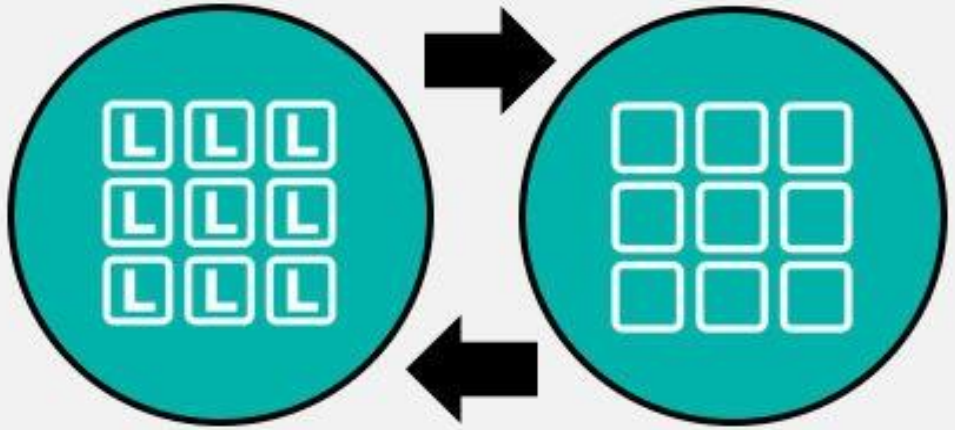
Type	Algorithm or Task
Unsupervised	K-Means: Cluster Analysis
Unsupervised	Association Rule Learning
Unsupervised	Dimensionality Reduction Techniques (PCA, SVD)

Semi-supervised Learning Defined



Semi-supervised Learning

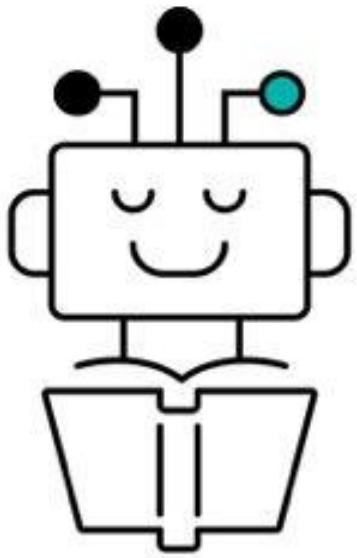
Semi-supervised learning uses a combination of supervised labeled training data with unsupervised methods



Labeled Data

Unlabeled Data

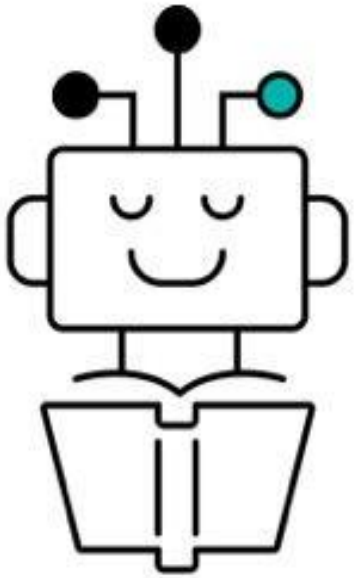
Semi-supervised Learning: How it Works



Semi-supervised Learning

1. Load labeled input training data
2. Model is trained on the data
3. Present unlabeled raw data
4. Algorithm infers classifiers for unlabeled data on its own
5. High confidence data is added to labeled training data set
6. Algorithm progressively adapts and learns

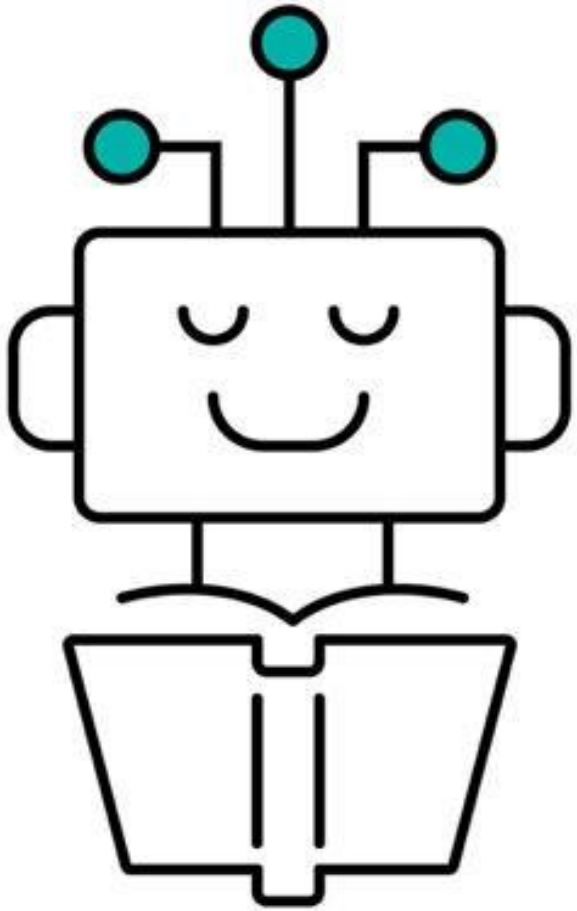
Semi-supervised Algorithms Table



Semi-supervised Learning

Type	Algorithm or Task
Semi-supervised	Self-training Algorithms
Semi-supervised	Generative Model – Gaussian Mixture model
Semi-supervised	Graph Based Algorithms – label propagation

Available Machine Learning Libraries



- Scikit-learn
- Deeplearning4j
- Tensorflow
- Spark MLlib
- Keras
- Torch
- Caffe
- MXnet
- XGBoost
- h2o

Learning Goals

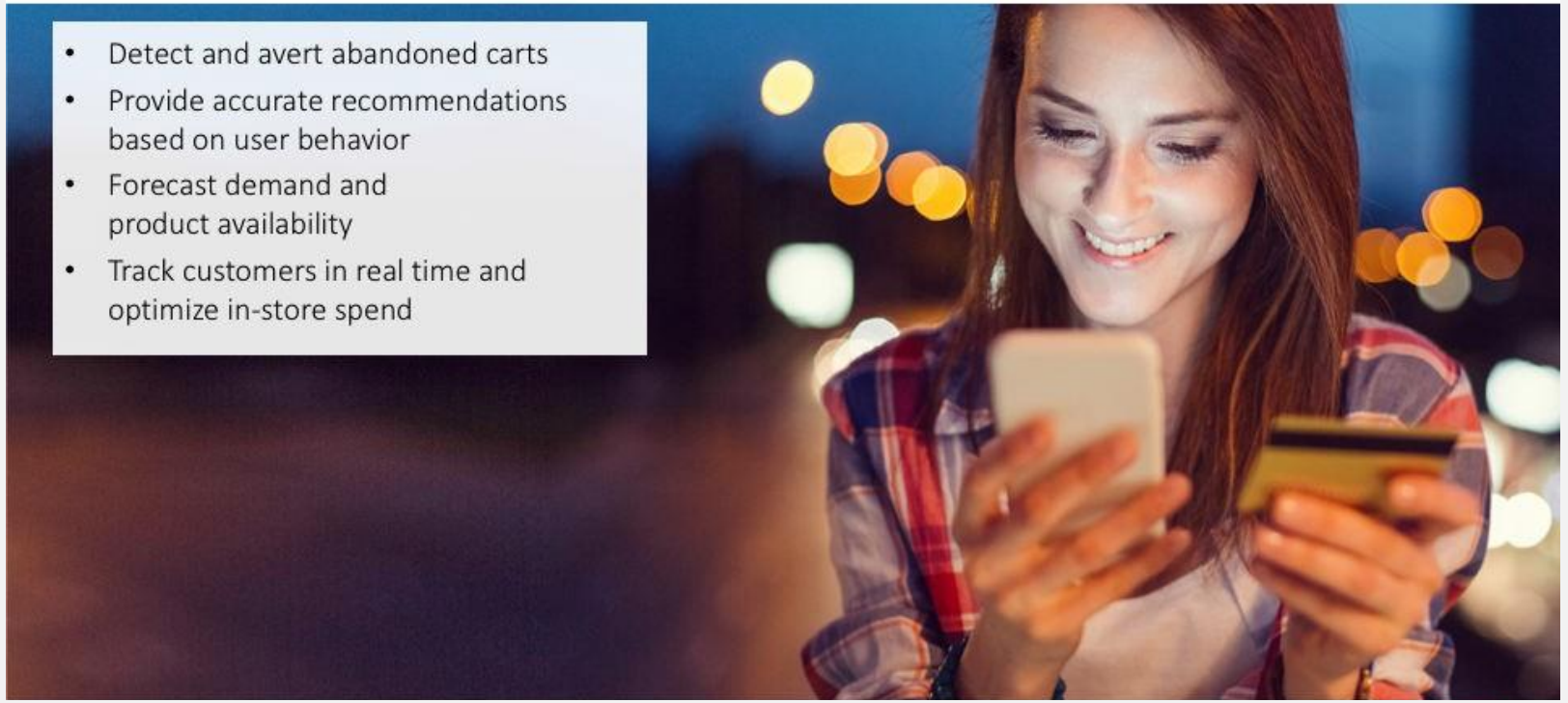
1.1 Background on Artificial Intelligence and Machine Learning

1.2 Types of Machine Learning

1.3 Explore Real-World Use Cases

Common Machine Learning Use Cases

Use Case: Retail

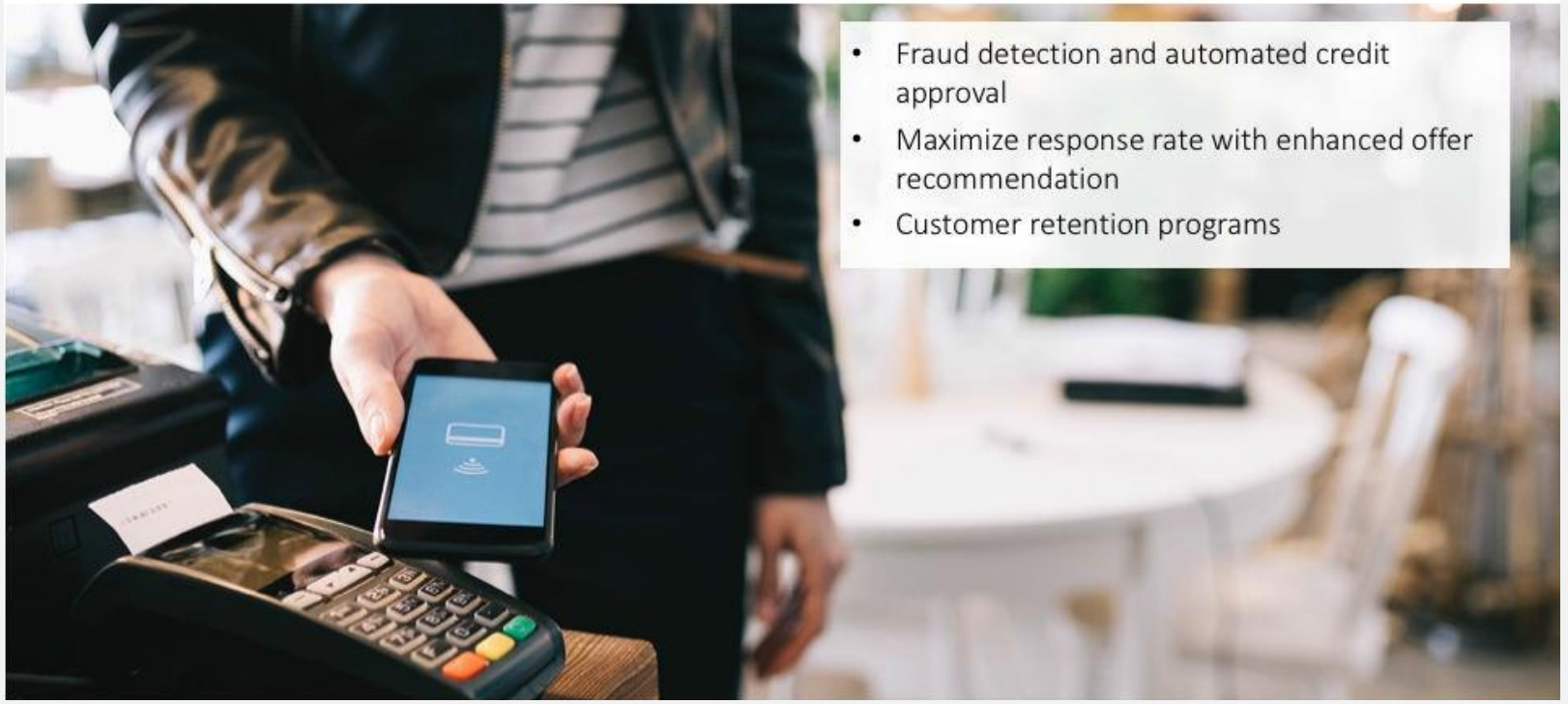


Use Case: Retail Example

Recommendation systems predict a user's preference for, or rating of, an item

Collaborative Filtering (Recommendation):

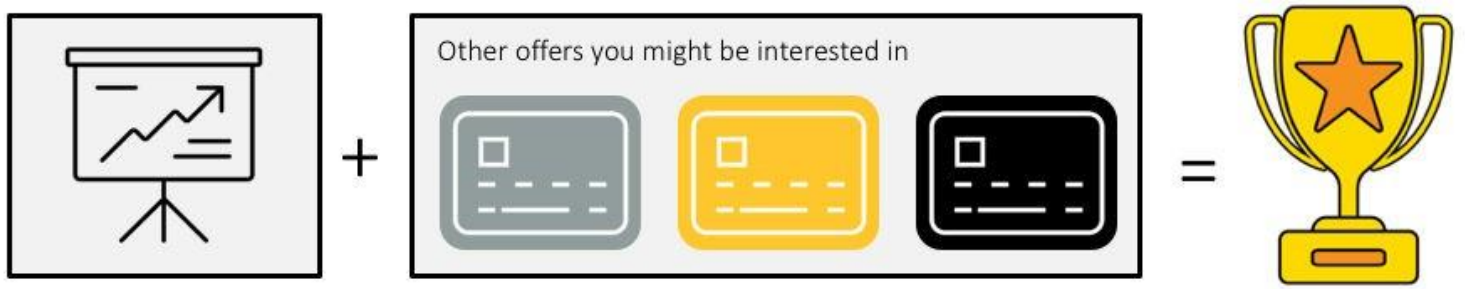
Use Case: Finance



- Fraud detection and automated credit approval
- Maximize response rate with enhanced offer recommendation
- Customer retention programs

Use Case: Finance Example

Customer Retention Program



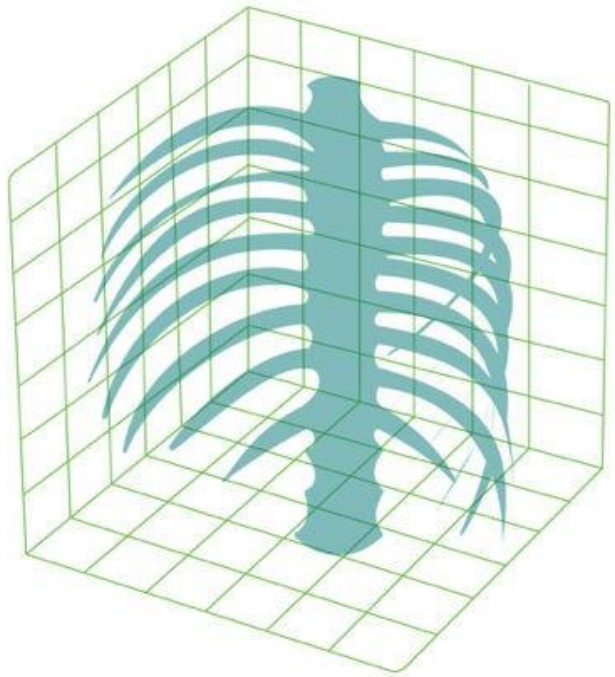
Predictive Classification Model

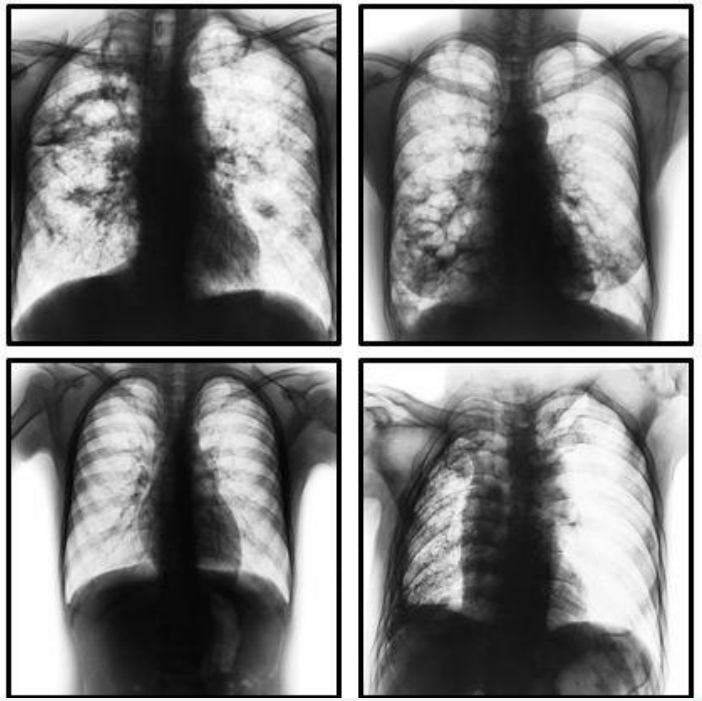
Recommender Model

Customer Retention Program

Use Case: Healthcare

Use Case: Healthcare Example



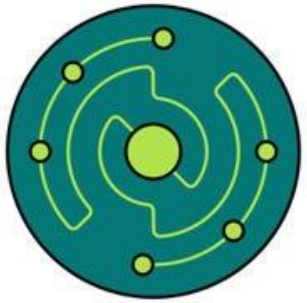


Use Case: Manufacturing

- Efficiency improvements and quality control
- Averts employee churn
- Anomaly detection with real-time streaming data from IoT sensors for predictive maintenance

Use Case: Manufacturing Example

Robot
Arm



Robot
Sensor

Health
Status

Trigger
Maintenance

Use Case: Self-Driving Car

- Train models on varied terrains and speeds using streaming input data
- Classify and label data to train static vs. moving object recognition
- Combines learning types and multiple algorithms

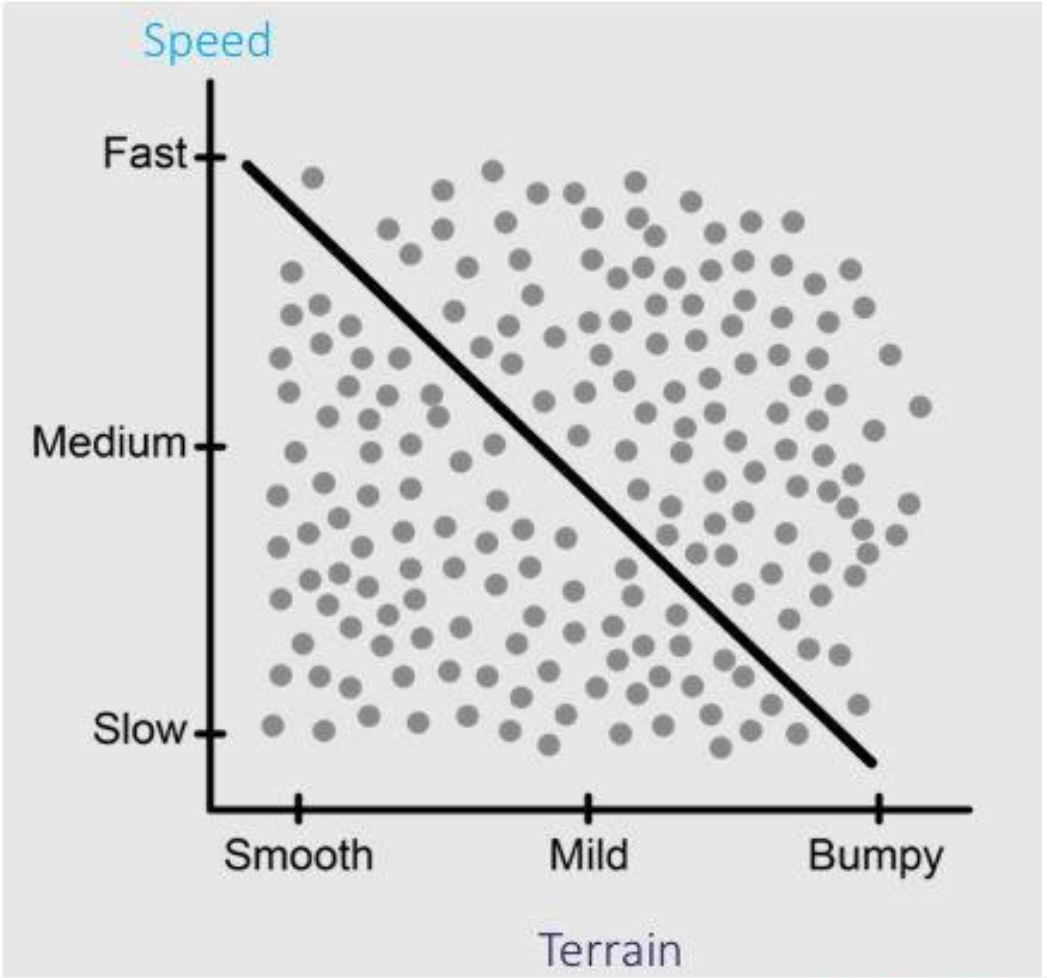
Use Case: Self-Driving Car Example

Streaming Data

Speed

Streaming Data

Terrain



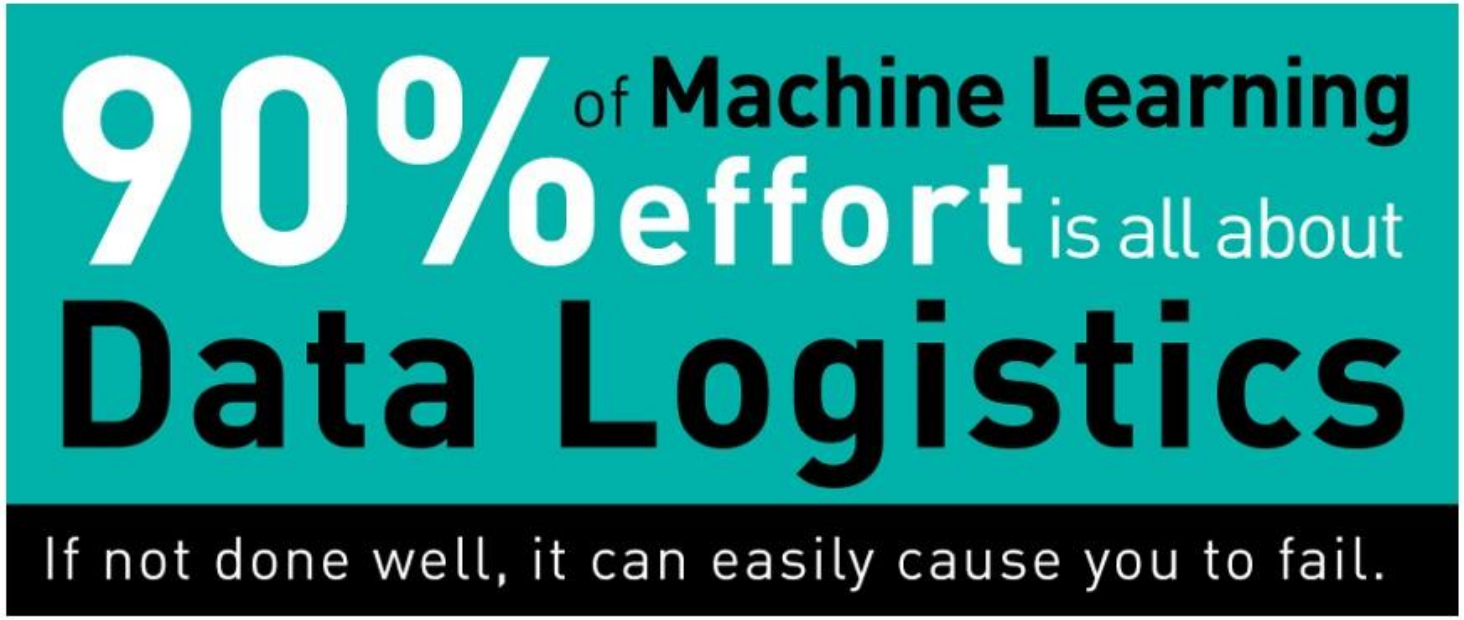
Learning Goals

1.1 Background on Artificial Intelligence and Machine Learning

1.2 Types of Machine Learning

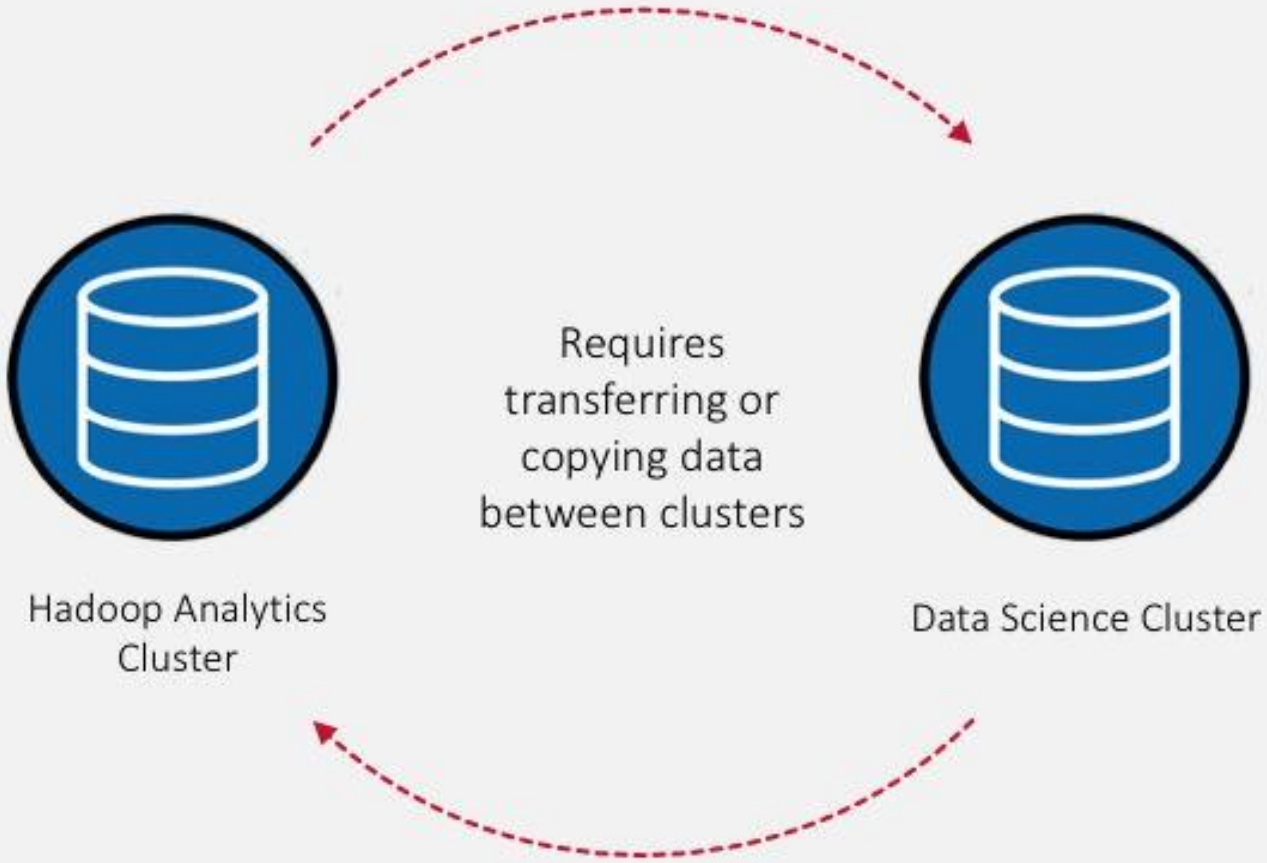
1.3 Explore Real-World Use Cases

Data Management Logistics



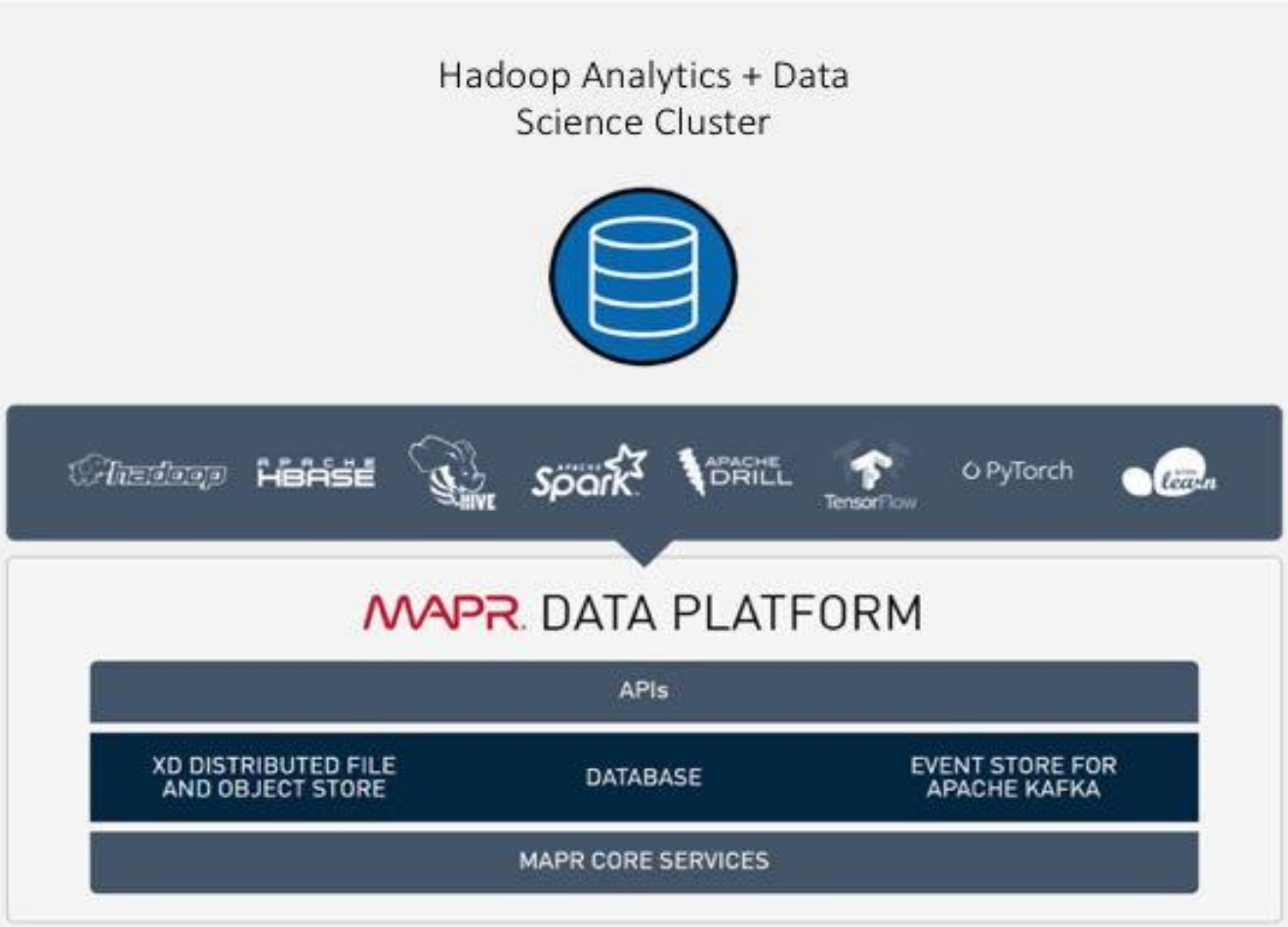
Problem: Hadoop Data Access

- Data Science libraries may not be compatible with HDFS; cannot communicate with traditional Hadoop clusters
- Separate clusters for ML workflows and Hadoop storage and processing
  - Requires data transfer/copying
  - Expense of additional clusters
  - Increased processing time impacts real-time results
  - Potential forked data



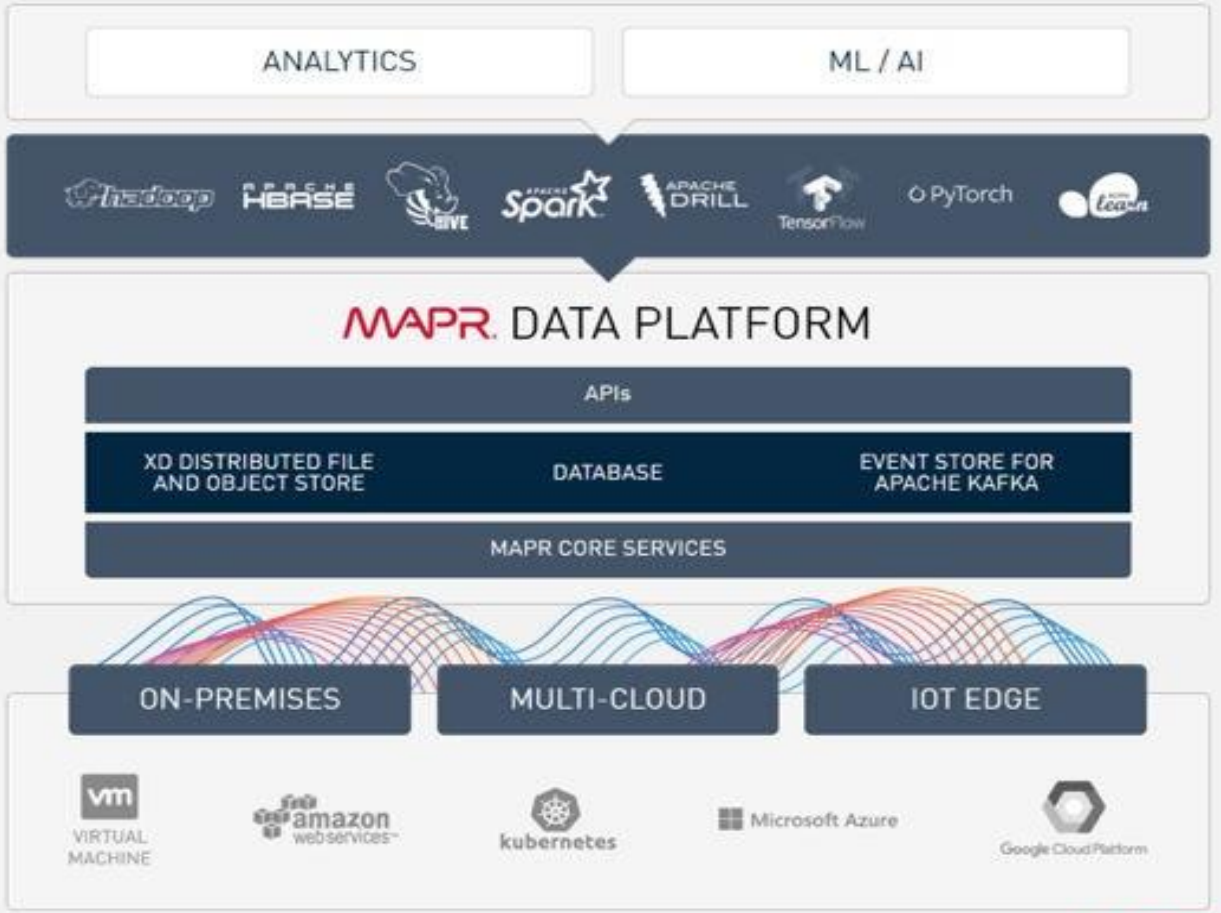
MapR Solution: Open APIs

- Diverse set of open APIs supports all types of data science
  - POSIX compliant APIs
  - Kafka
  - Spark
  - Amazon S3
- Single cluster, no data movement



MapR Data Platform

- Foundational technology to support ML data logistics
- Global namespace
- Scalable, flexible enterprise storage
- Flexible database capabilities
- Real-time, global event streaming



MapR Data Storage Solutions: MapR XD

- Distributed file and object store
- POSIX compliant file system
- Supports large binary files like images or video

MapR XD Distributed File and Object Store

MapR Data Storage Solutions: MapR Database

- Binary, JSON document database
- NoSQL datastore
- Fully compatible with the HBase API

MapR Data Storage Solutions: MapR Event Streaming

- Kafka API-based publish/subscribe messaging system
- Real-time data streaming
- Replayable and replicable streams

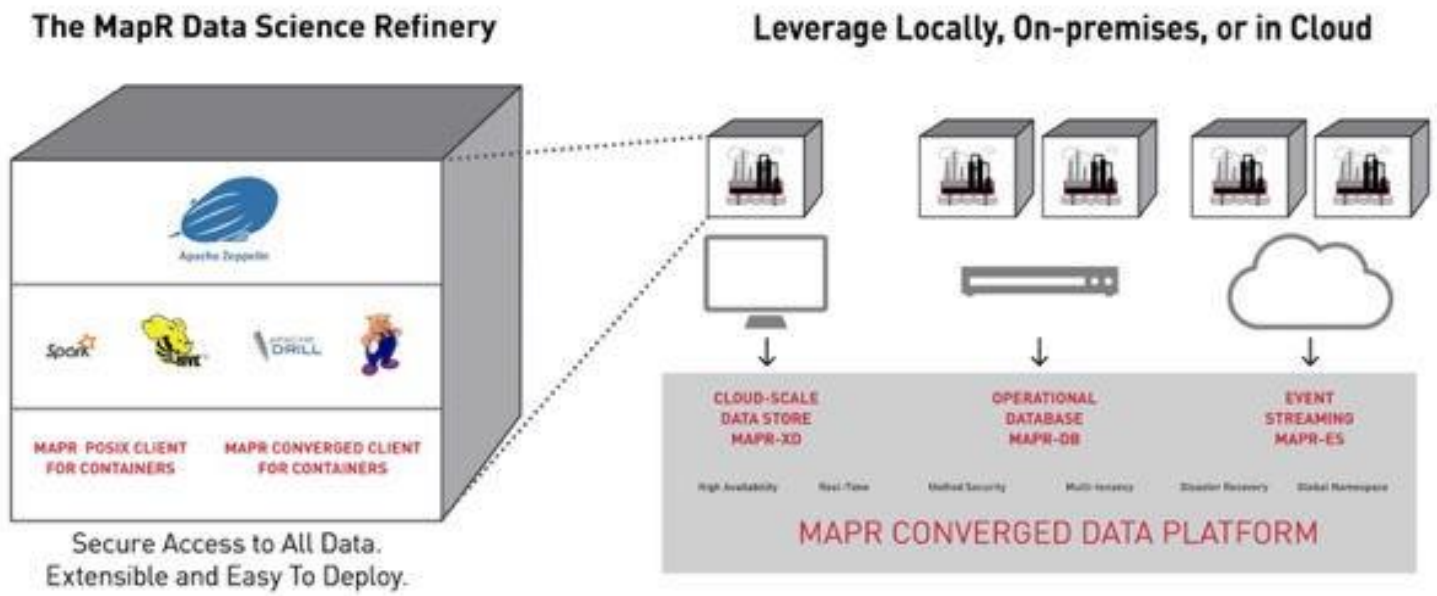
MapR Event Store
for Apache Kafka

MapR Data Storage Solutions: Cloud Integration

- Easily integrates with any cloud storage solution
- Uses the same tools in-cloud as on-premise

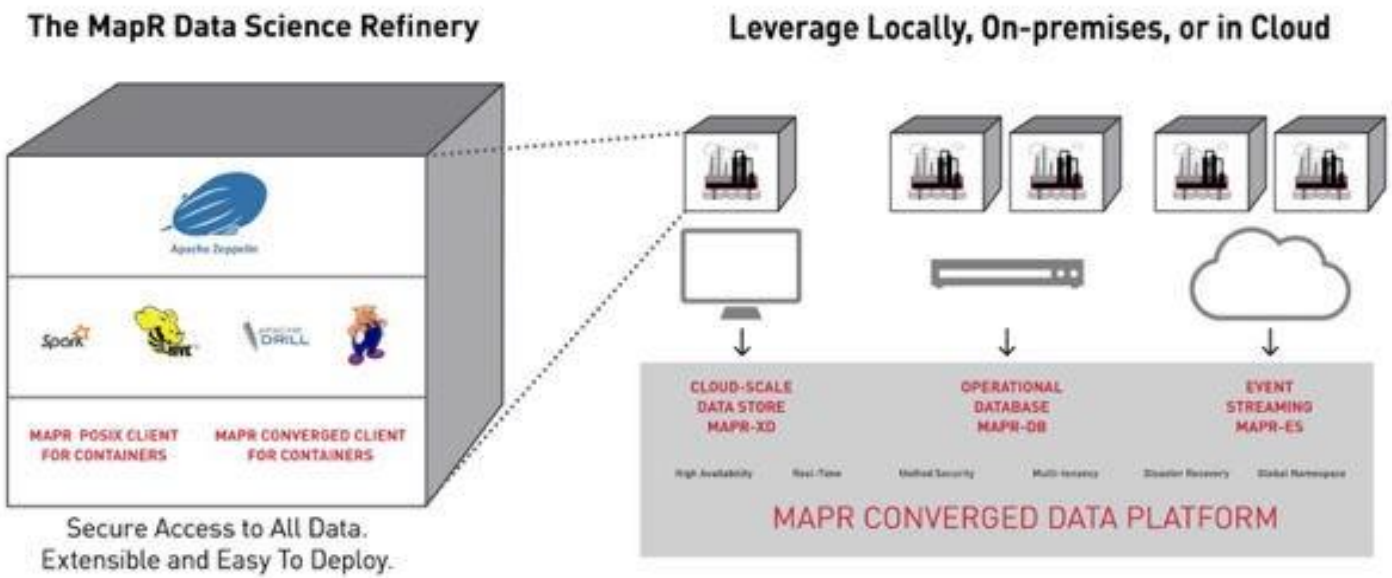
MapR Data Science Refinery

- Preconfigured solution works for Data Science Teams
- Data Science Toolkit is easy to deploy and scalable
- Containerized data science provides predictable environment, while separating compute and storage



MapR Data Science Refinery

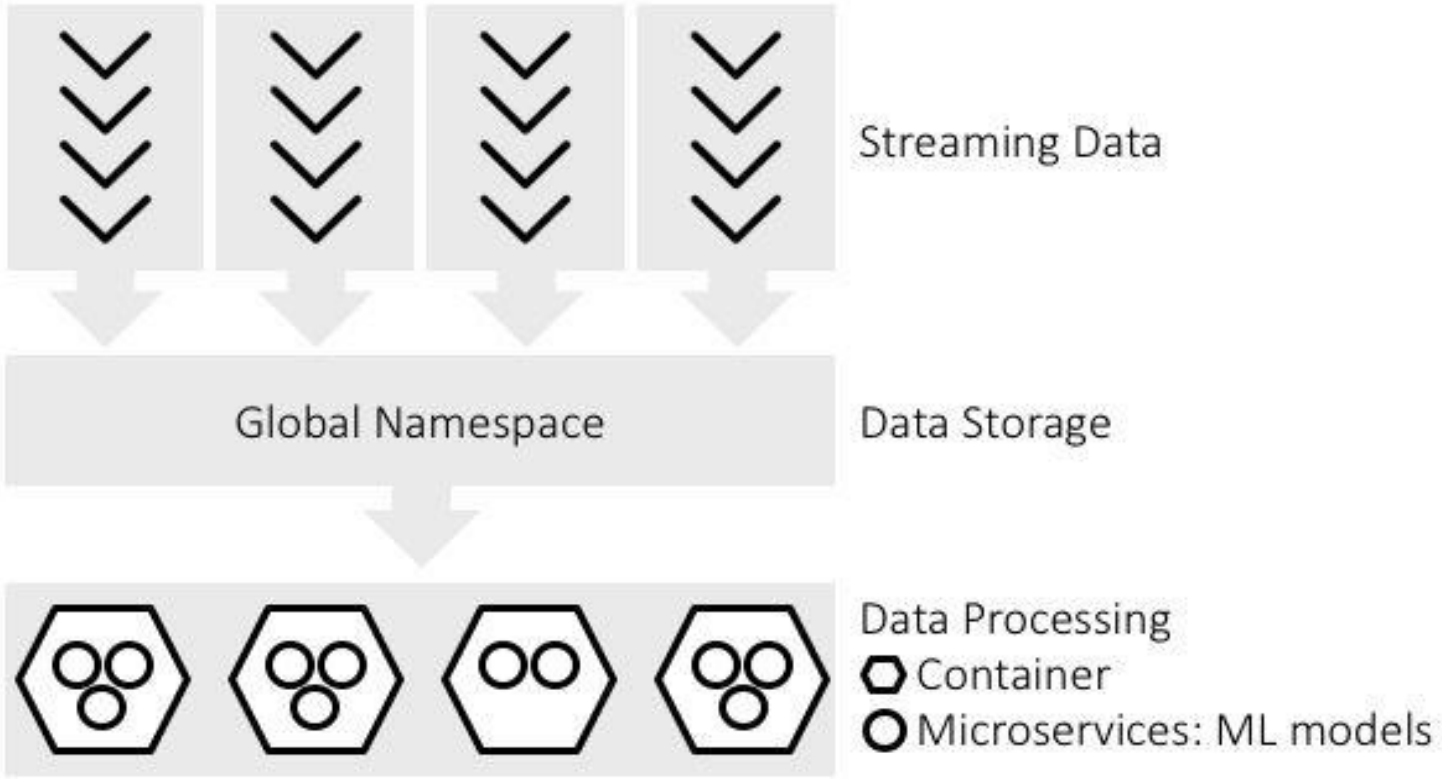
- Global namespace enables secure, built-in collaboration abilities
- Preconfigured Docker containers leverage MapR as a persistent data store
- Apache Zeppelin offers preconfigured visualization capabilities



ML for Enterprise Operations

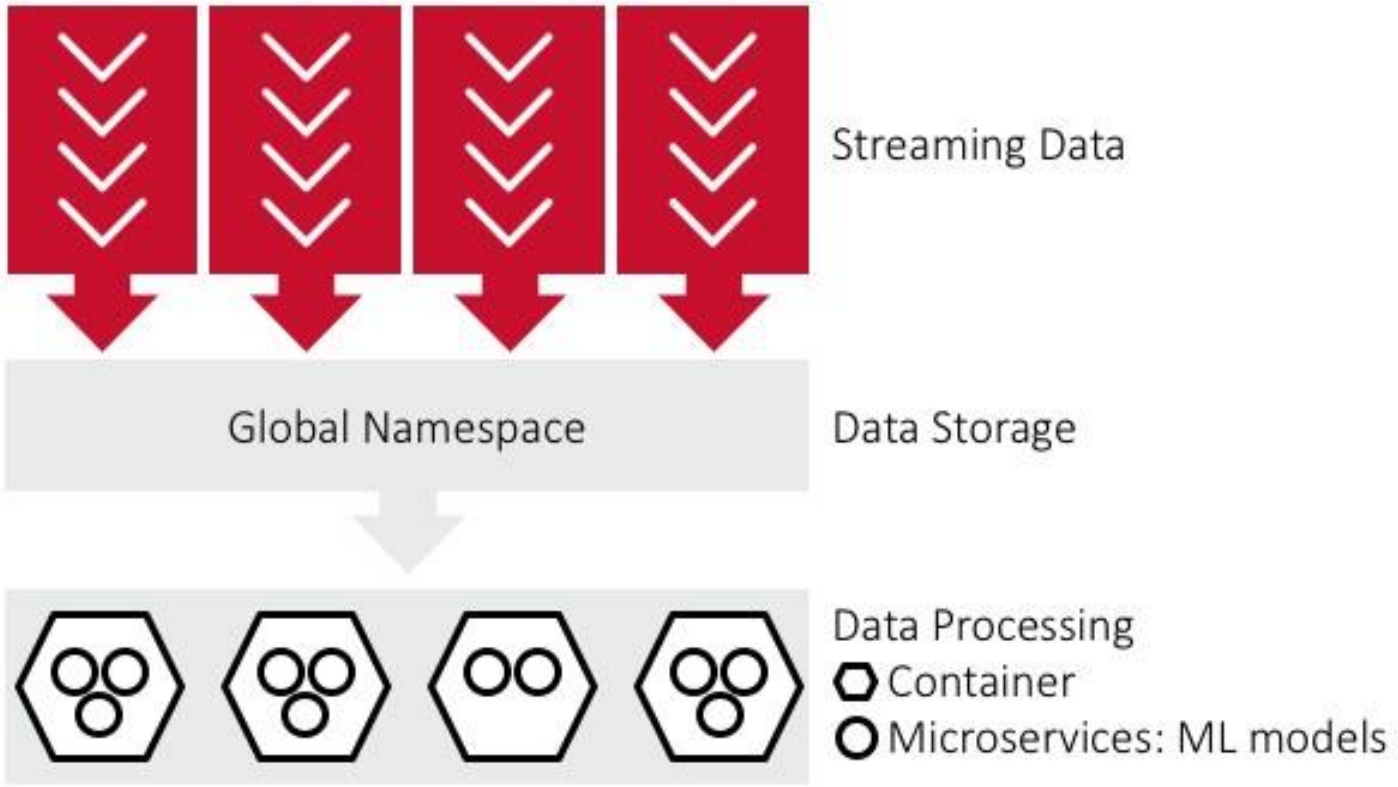
Rendezvous Architecture

- Proposed design to handle data logistics for a wide range of ML use cases
- Foundation for enterprise-grade data logistics management
- Continuous integration system for machine learning models



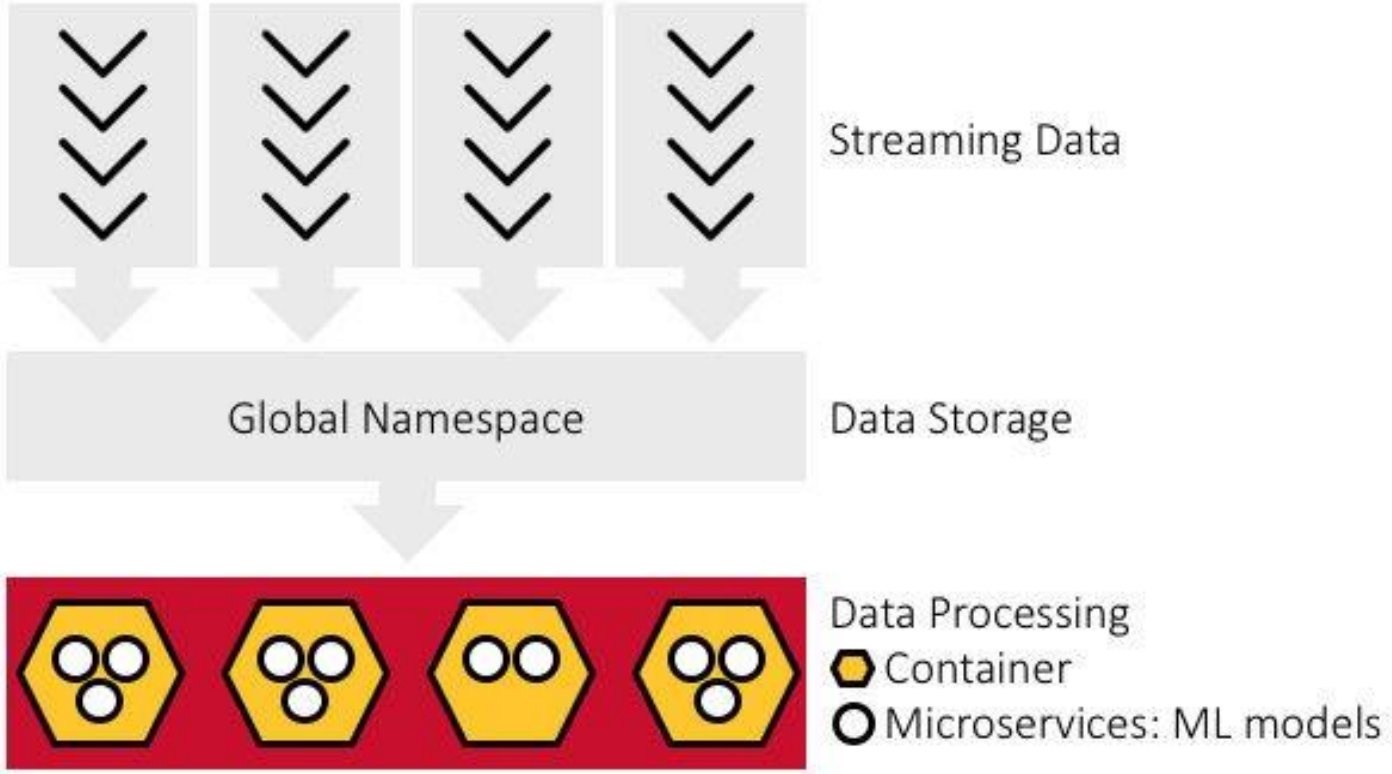
Rendezvous Architecture: Streaming Data

1. Takes advantage of streaming data



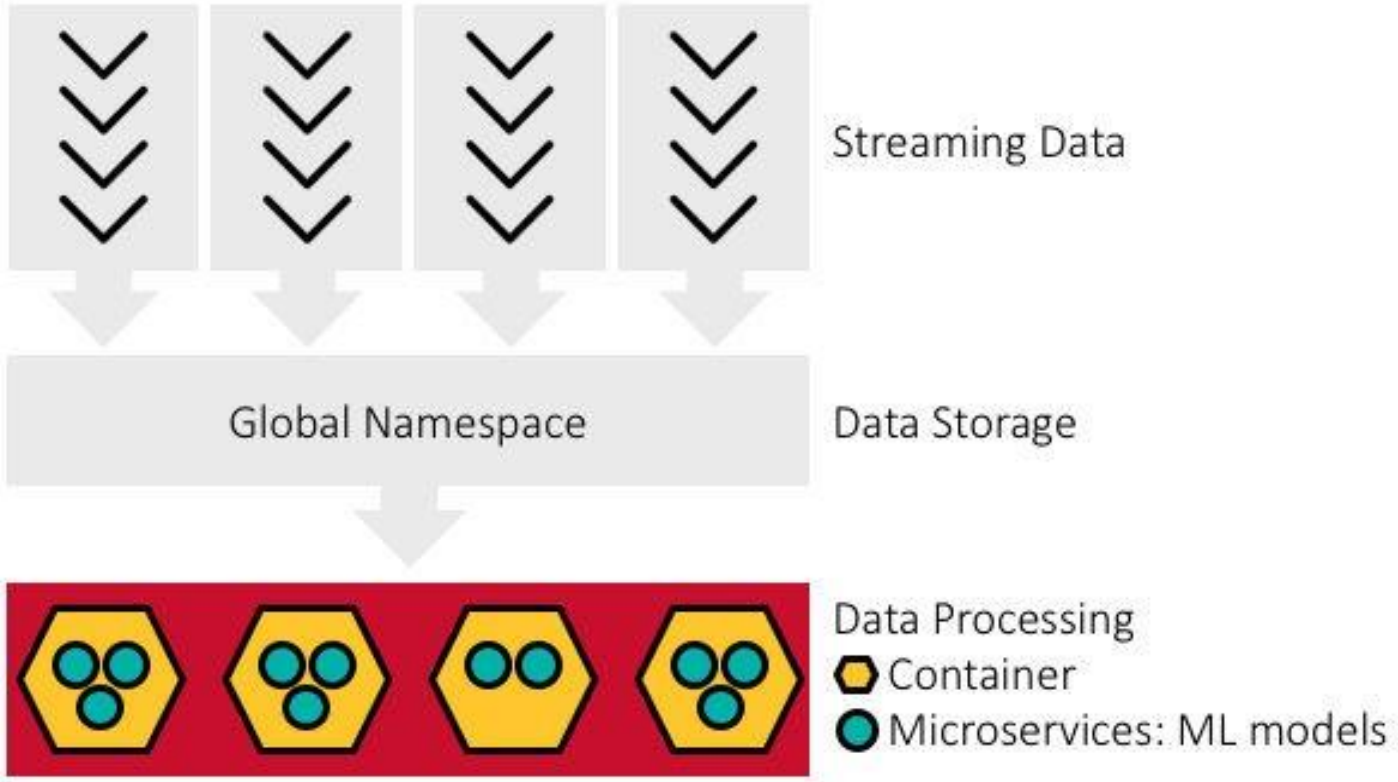
Rendezvous Architecture: Containers

1. Takes advantage of streaming data
2. Predictable environment by using containers

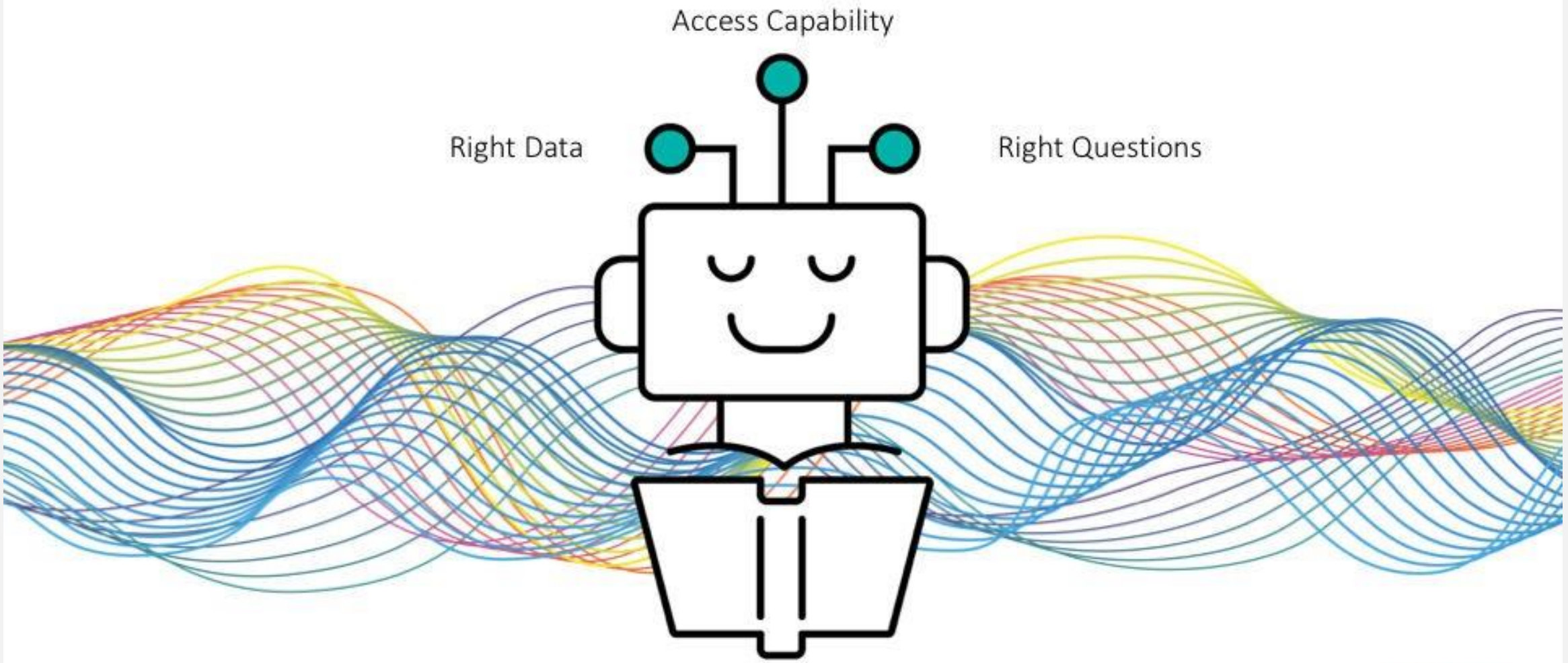


Rendezvous Architecture: Microservices Approach

1. Takes advantage of streaming data
2. Predictable environment by using containers
3. Flexible, streaming microservices approach



Machine Learning Success




MAPR academy
Build Machine Learning Projects
Lesson 1: Prepare Your Project

Learning Goals

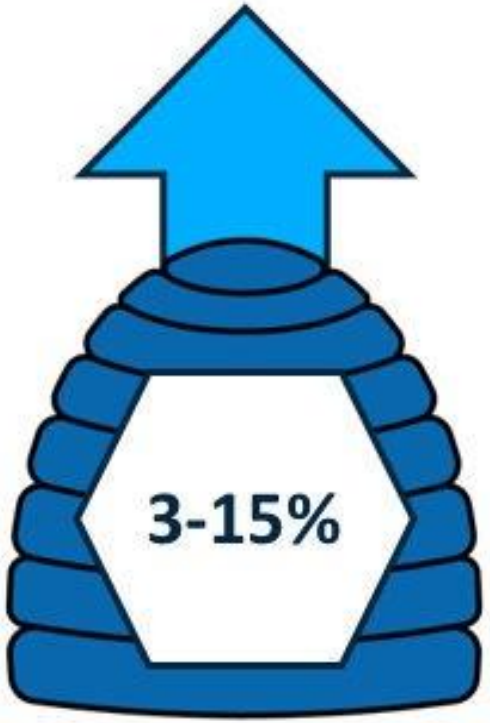
1.1 AI/ML Overview

1.2 Step 1: Identify Business Need and Plan

1.3 Step 2: Data Management

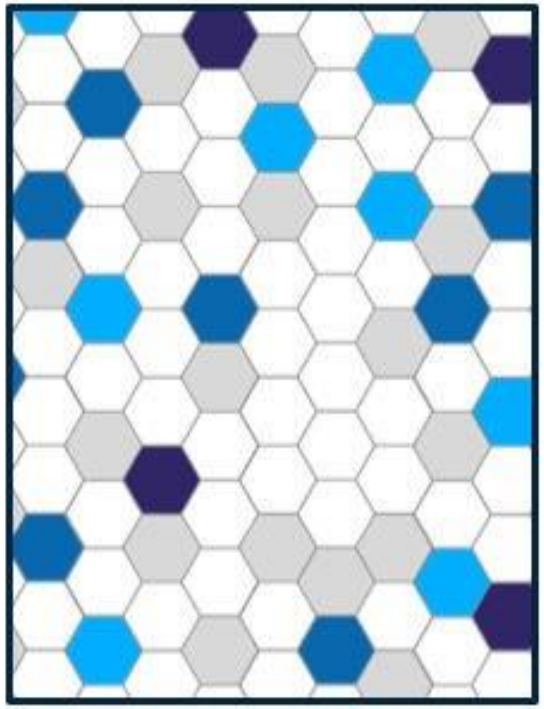
1.4 Step 3: Feature Selection and Engineering

Artificial Intelligence Adopters



Profit Advantage

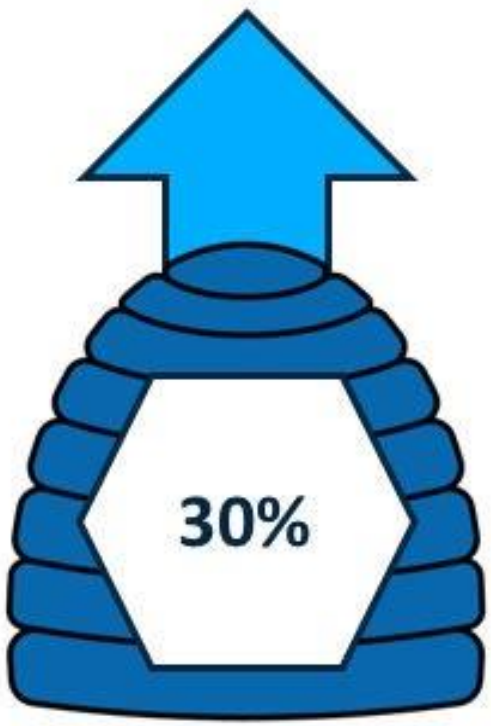
Overall Revenue



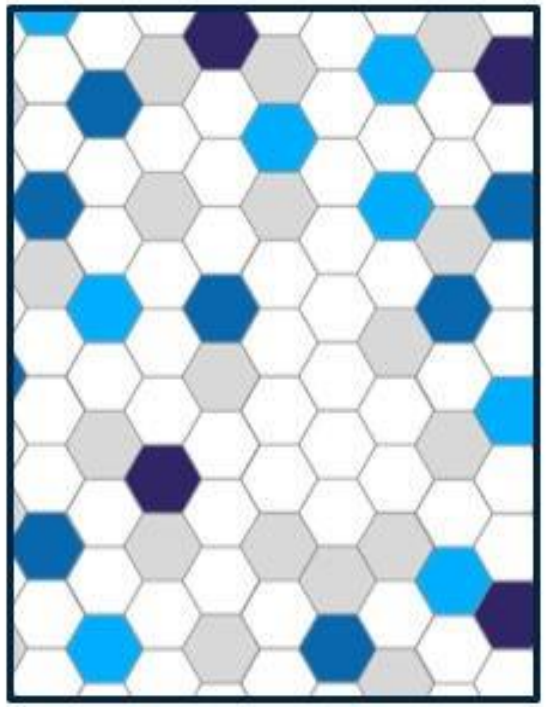
Production

Artificial Intelligence Adopters

Profit Advantage

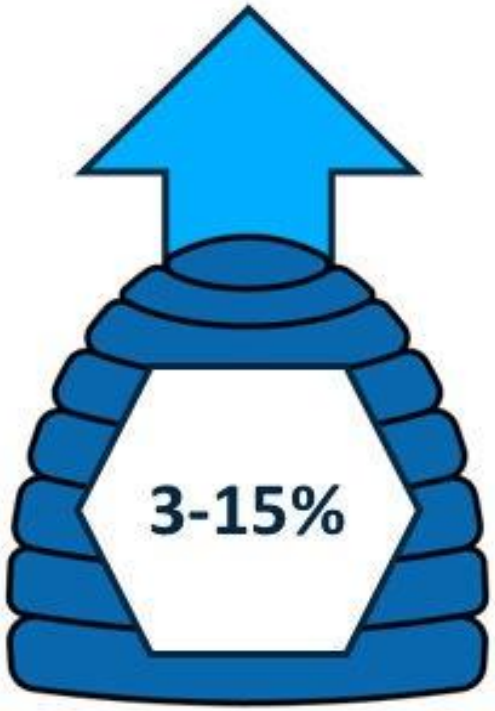


Overall Revenue

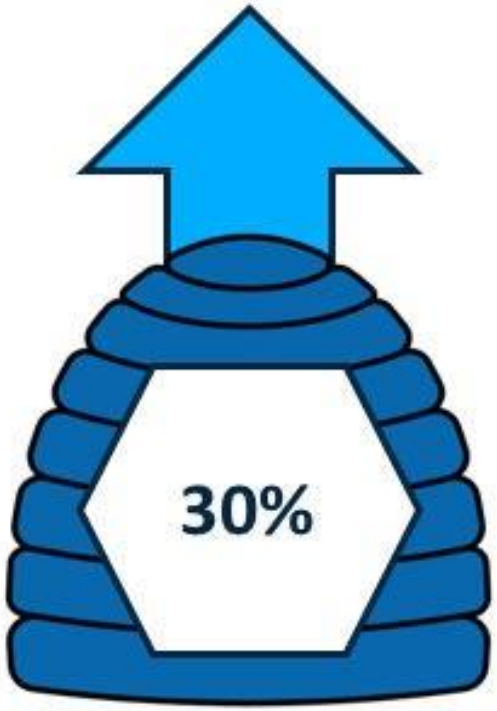


Production

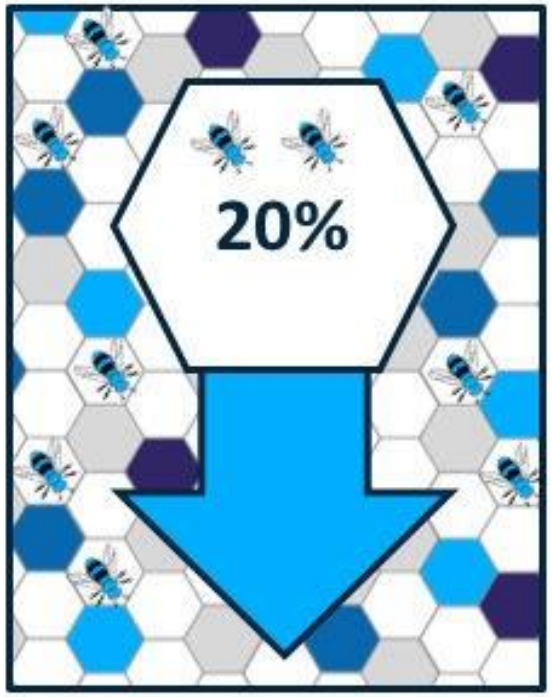
Artificial Intelligence Adopters



Profit Advantage

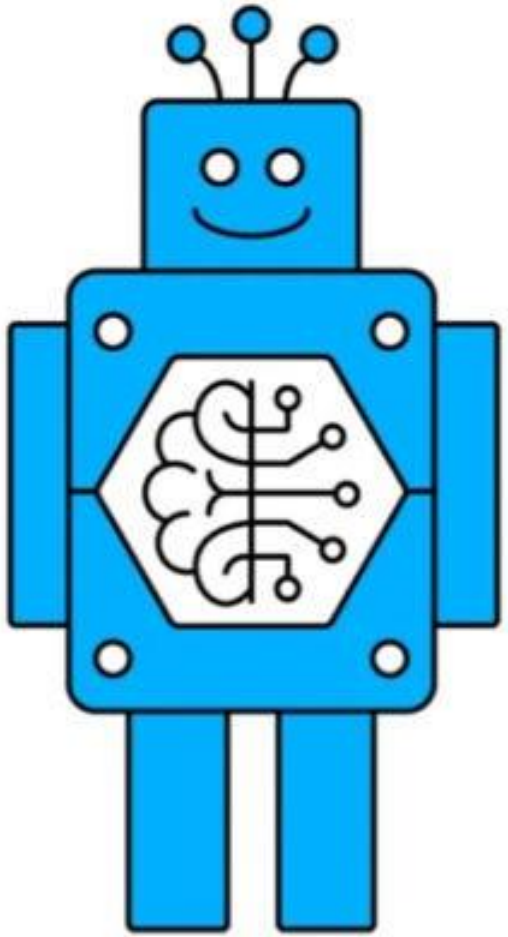


Overall Revenue



Production

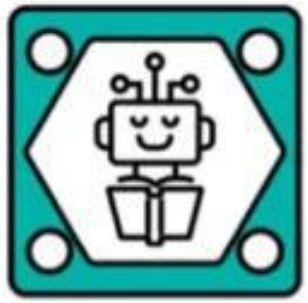
The Data Science Landscape



AI Planning

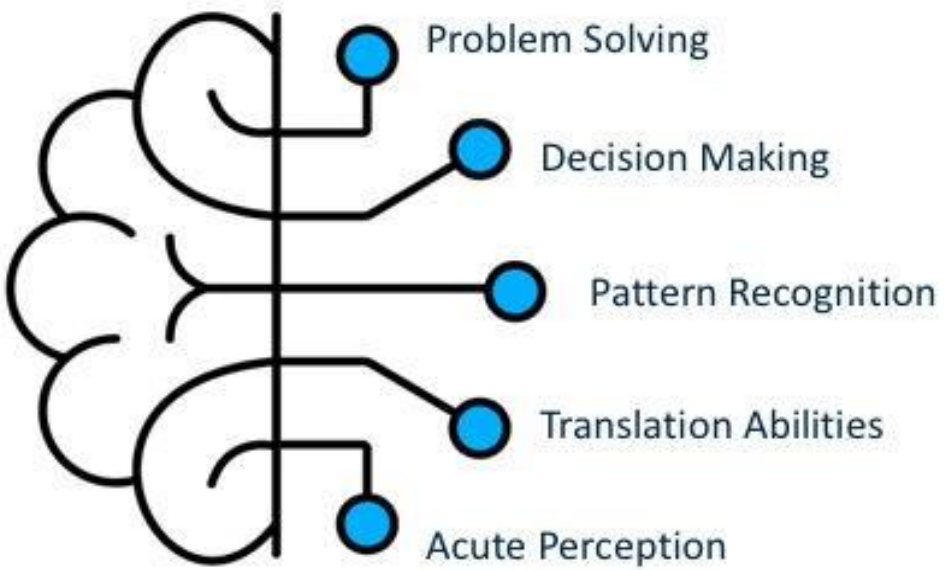
Symbolic Logic
Expert System

Machine Learning

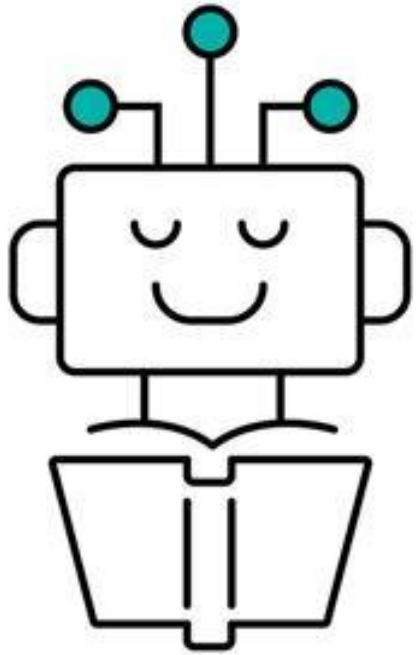


| Review: AI vs. ML

Artificial Intelligence

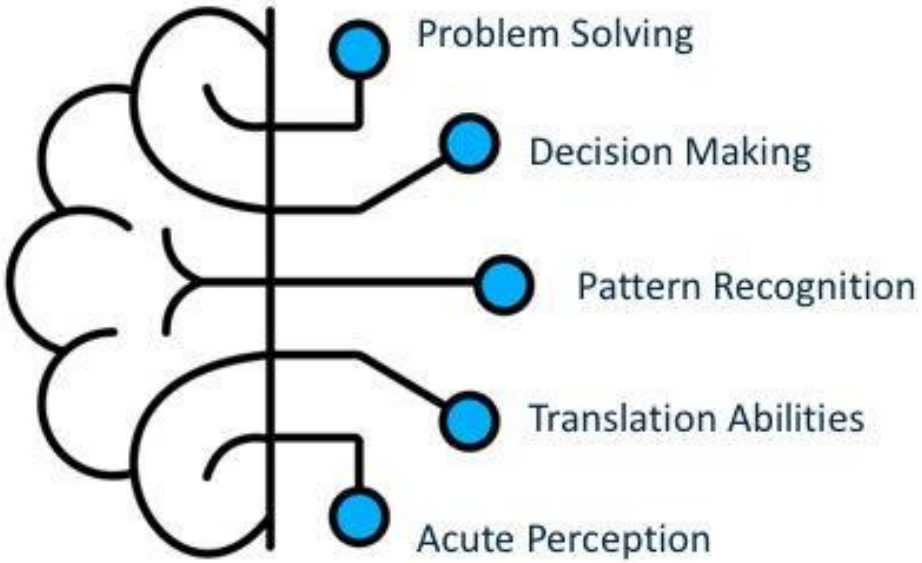


Machine Learning



Review: AI vs. ML

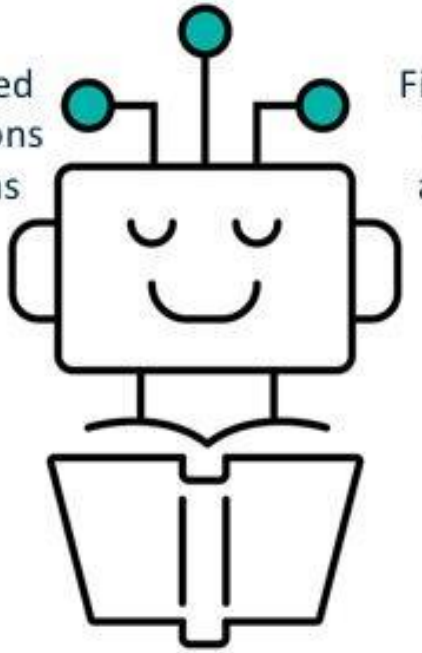
Artificial Intelligence



Machine Learning

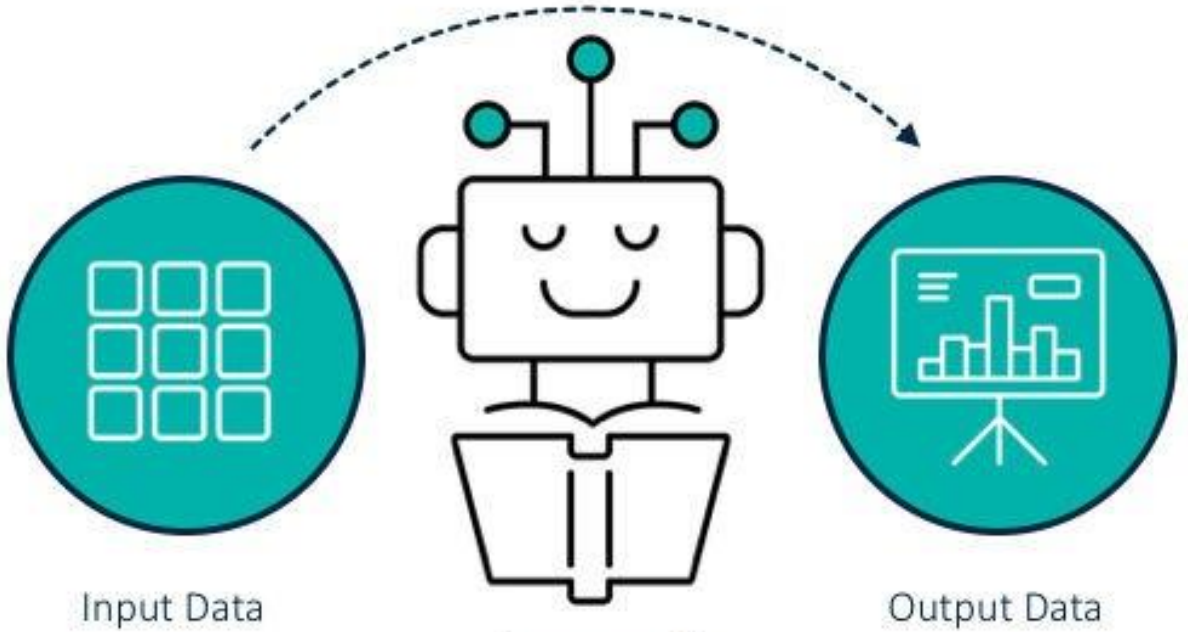
Taught to learn from data and make decisions

Makes calculated recommendations and predictions



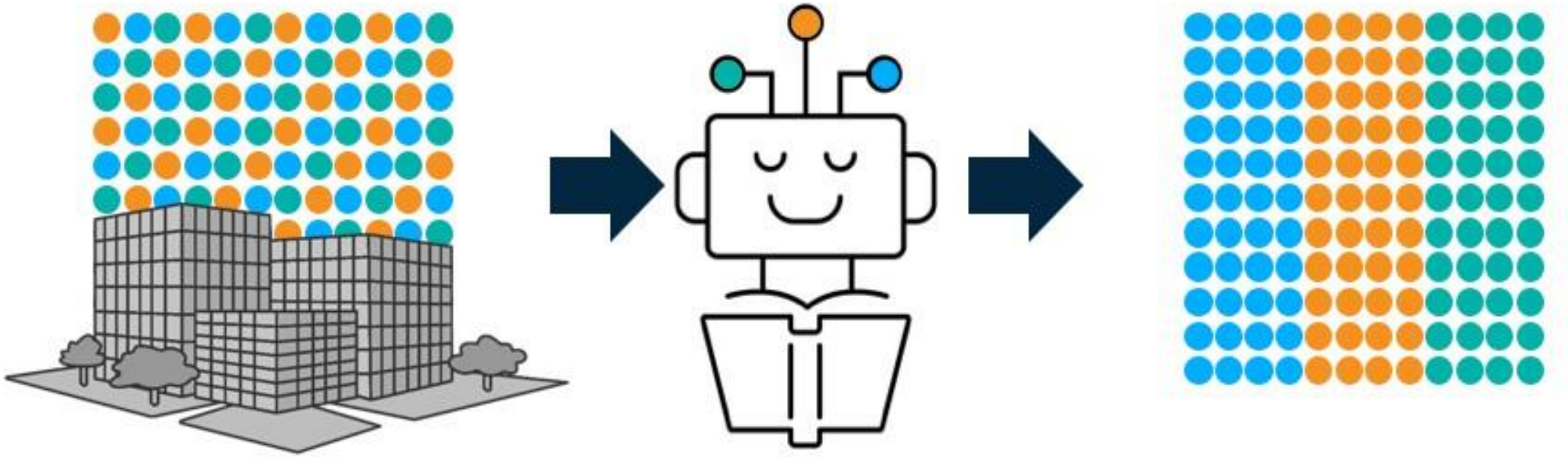
Finds patterns and makes associations

Review: AI vs. ML



Dynamic Feedback Loop

Machine Learning Today



Enterprise Data

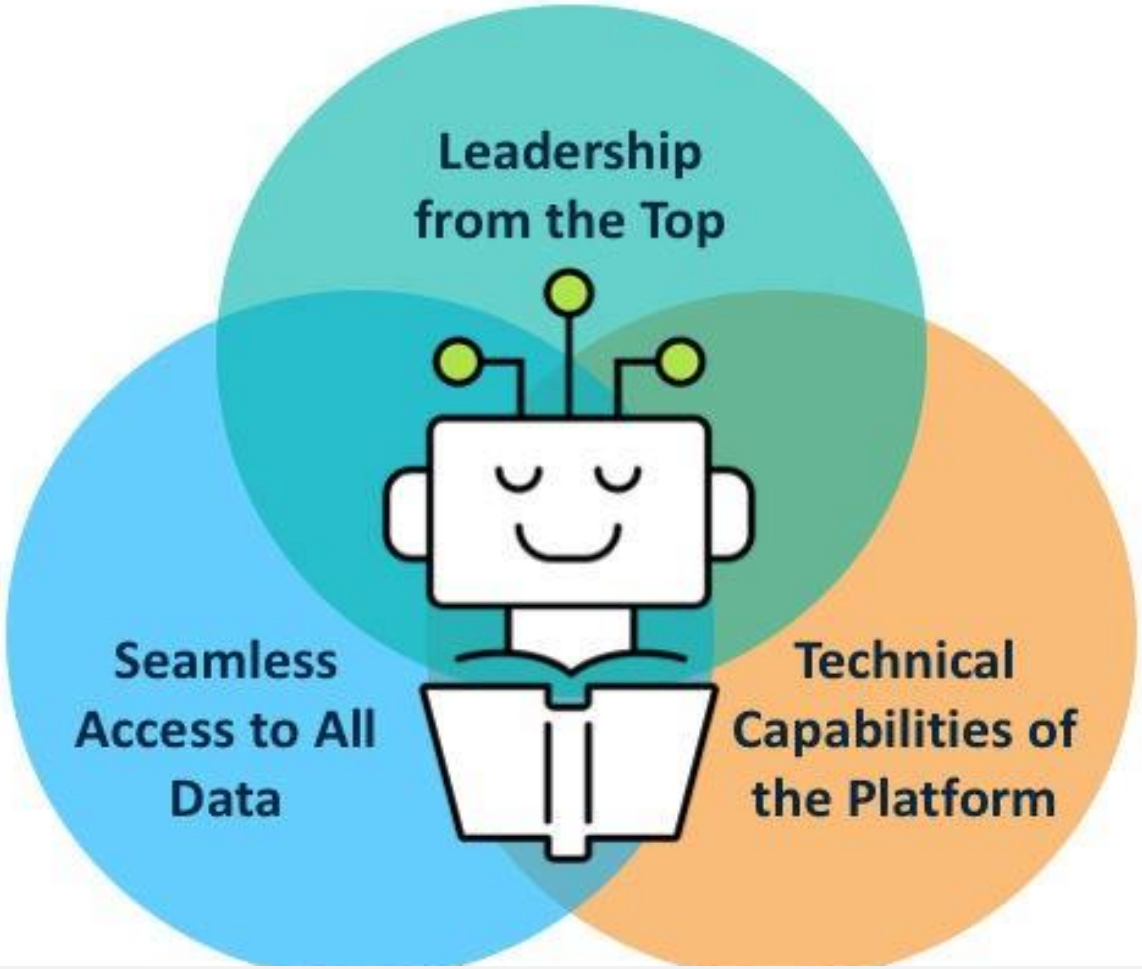
Machine Learning

Tagging and Classification

Machine Learning Today

- Used in every industry
- Applied to any scenario where data is analyzed
- Machine-driven insights drive actionable results

A Strong Foundation



Learning Goals

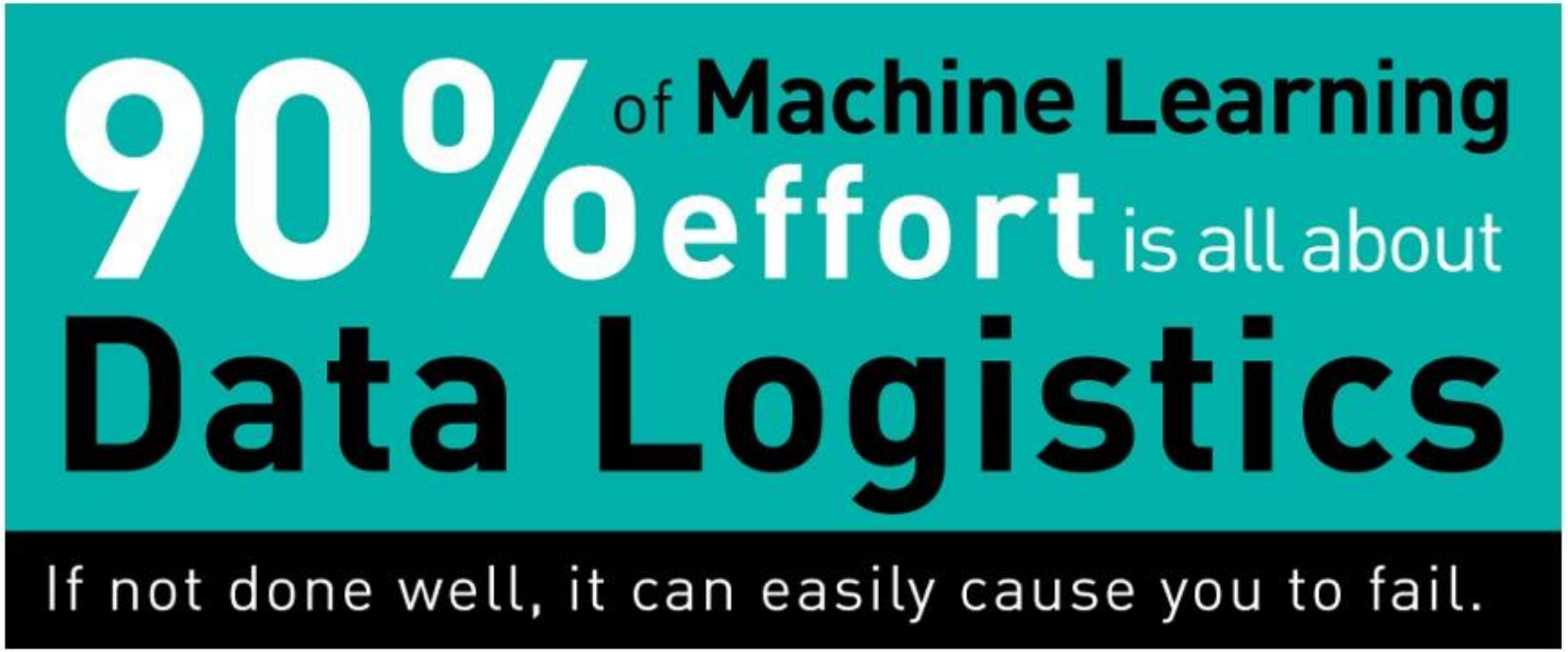
1.1 AI/ML Overview

1.2 Step 1: Identify Business Need and Plan

1.3 Step 2: Data Management

1.4 Step 3: Feature Selection and Engineering

Data Management Logistics



Project Plan Workflow

Part I (Preparation)			Part II (Implementation)		
Plan	Prepare	Features	Framework	Model	Production
1	2	3	4	5	6
Identify business need and create a project plan	Manage and prepare data	Select and engineer features for the model	Select algorithms and frameworks	Train and validate model	Implement and monitor project
					

1 Identify Business Need

How will the business derive value?

- Define project goal
- Break down broad goals to make them specific
- Determines success criteria

Determine the following:

1 Examples by Industry

Healthcare

Manufacturing

Finance

Retail

1 Examples by Industry - Healthcare

- Improve patient care
- Improve accuracy of medical diagnoses
- Improve accuracy detection on skin disorders
- Improve accuracy detection using image recognition on skin lesions by X%

1 Examples by Industry - Manufacturing

Improve business efficiency and reduce costs

Detect and prevent unexpected machine line maintenance

Install new IoT sensors to monitor performance, identifying potential instances of performance degradation

Reduce machine line downtime by X%

1 Examples by Industry - Finance

Reduce operational costs and
improve customer experience

Reduce response times of reported fraud behavior with program automation

Create ML application that automates system communications to report anomalous behavior, then block/reverse charges upon verification

Reduce total processing time by X%

1 Examples by Industry – Retail

- Increase sales and customer engagement
- Offer product recommendations and extend cross-promotional sales
- Create a robust recommendation engine
- Increase recommended product checkout rates by X%

1 Example Use Case – AllPets Retail Supplier

1 Example Use Case – Recommendation Engines

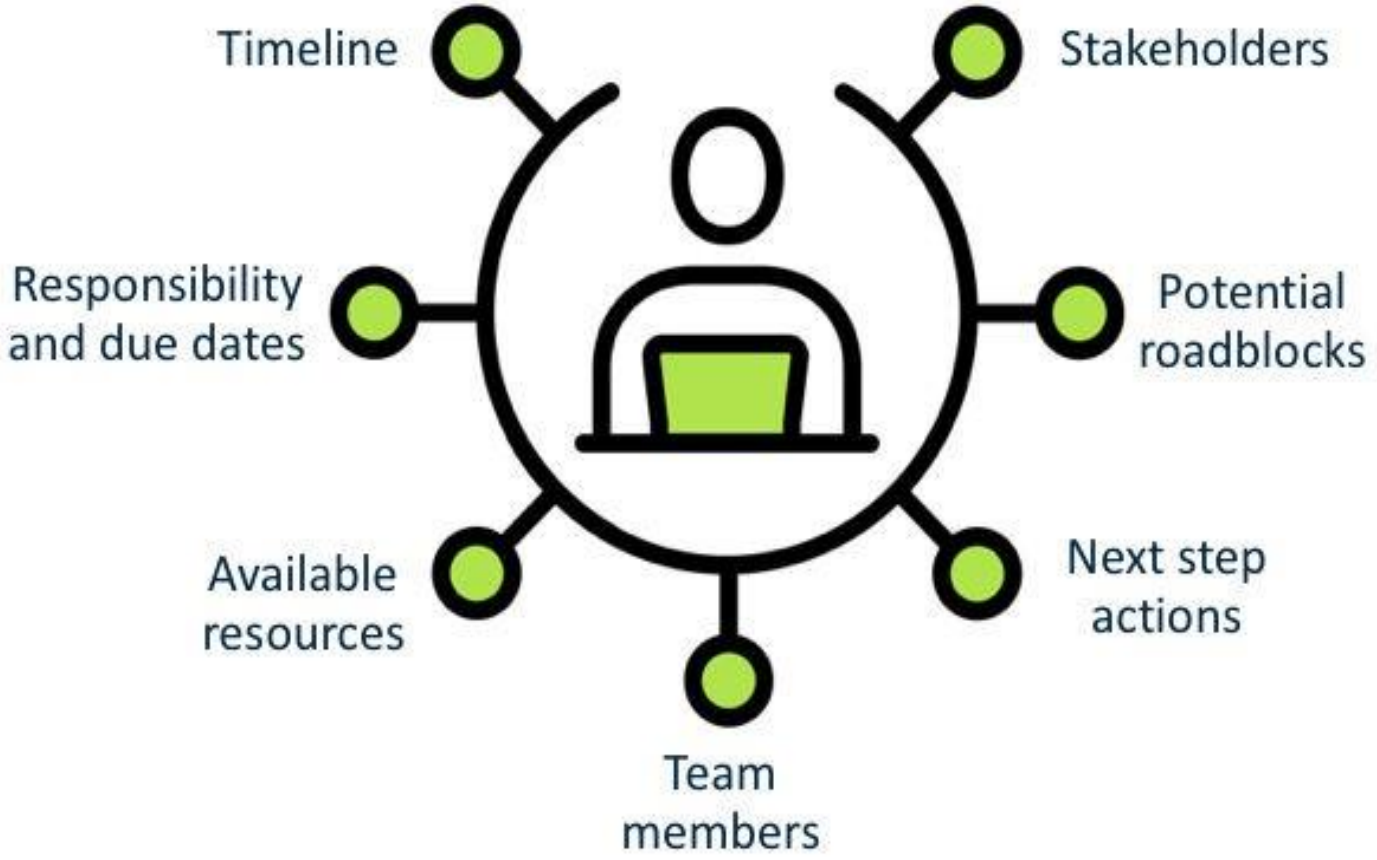
Content-Based Engines

- Item attribute data

Collaborative Filtering Engines

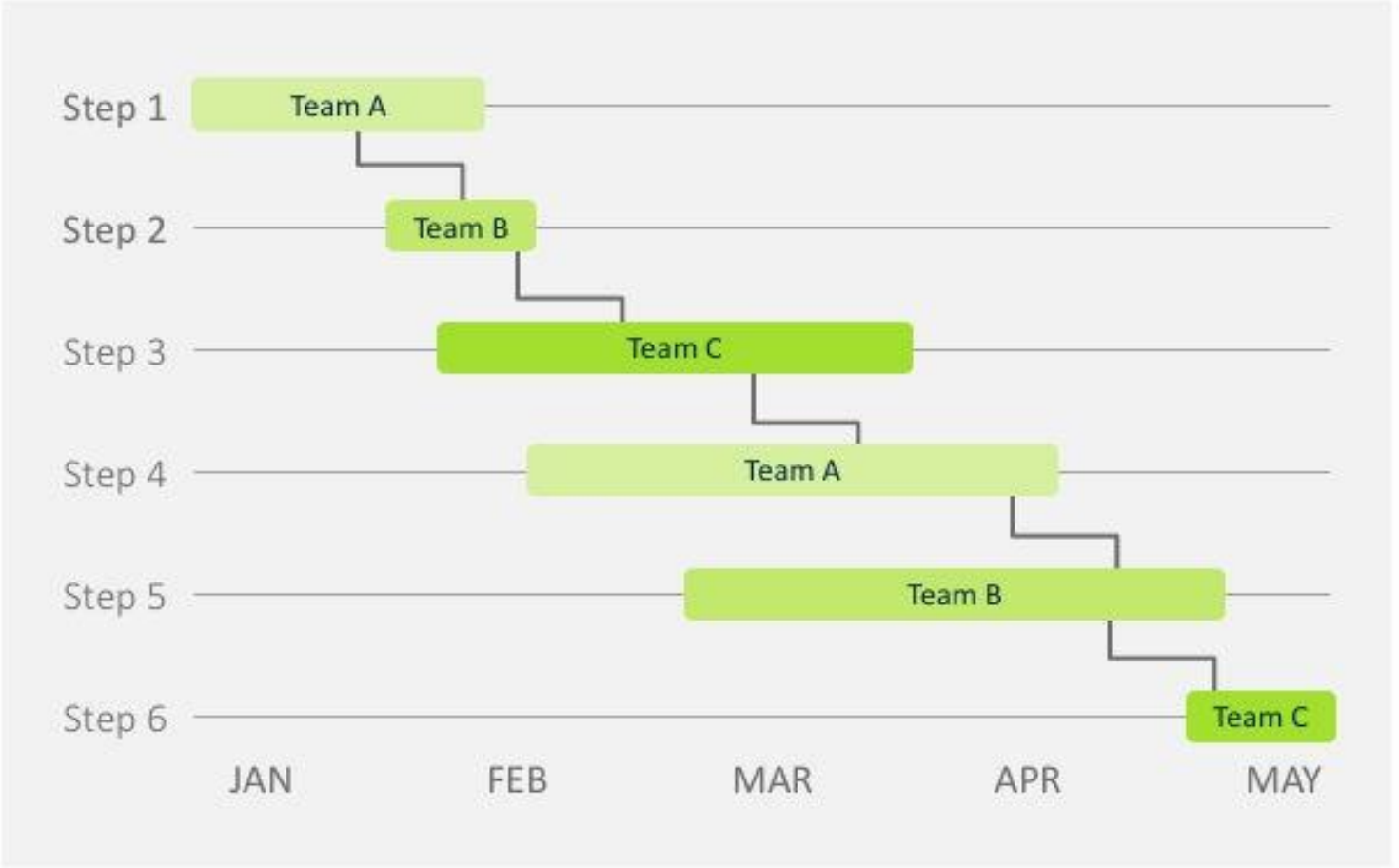
- User interactions

1 Establish a Solid Project Plan



1 Establish a Solid Project Plan

- Pre-defined launch tasks
- Incorporate all post-launch requirements



1 Machine Learning Planning Specifics

What tools will produce the best **results**?

What methods will **deliver** the most value?

Allocate enough time for:

- Selection of algorithms and frameworks
- Feature engineering
- Data preparation and transformation

1 Other ML Considerations

Privacy and Regulations

- What are the privacy concerns?
- Do you need to document reasoning behind decisions?

Cold Start Projects

- Starting without the required data
- Can buy or simulate data

Surrogate or Approximation Models

- Exploration with data you might want to model
- Expected outcomes are unknown

Learning Goals

1.1 AI/ML Overview

1.2 Step 1: Identify Business Need and Plan

1.3 Step 2: Data Management

1.4 Step 3: Feature Selection and Engineering

Project Plan Workflow: Step 2

Part I (Preparation)		
Plan	Prepare	Features
1	2	3
Identify business need and create a project plan	Manage and prepare data	Select and engineer features for the model
		

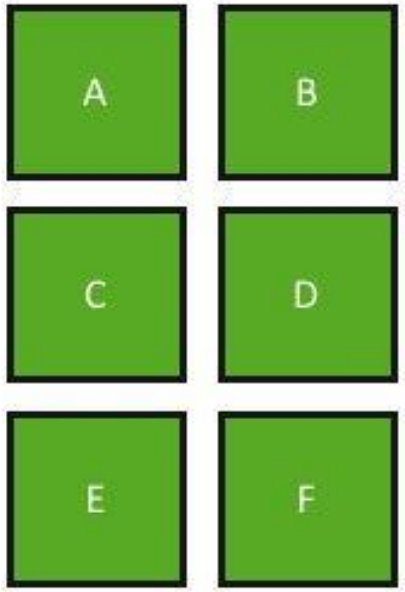
Main Areas of Data Management

- Collect
- Explore
- Assess

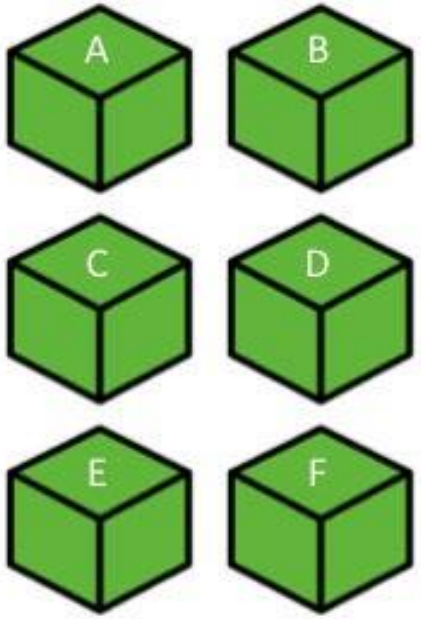
- Pre-process
- Clean
- Format

- Transform

2 Data Management – Collect Data



Identify datasets from various sources



Compile and retrieve datasets

Organize datasets

2 Data Management – Explore and Assess Data

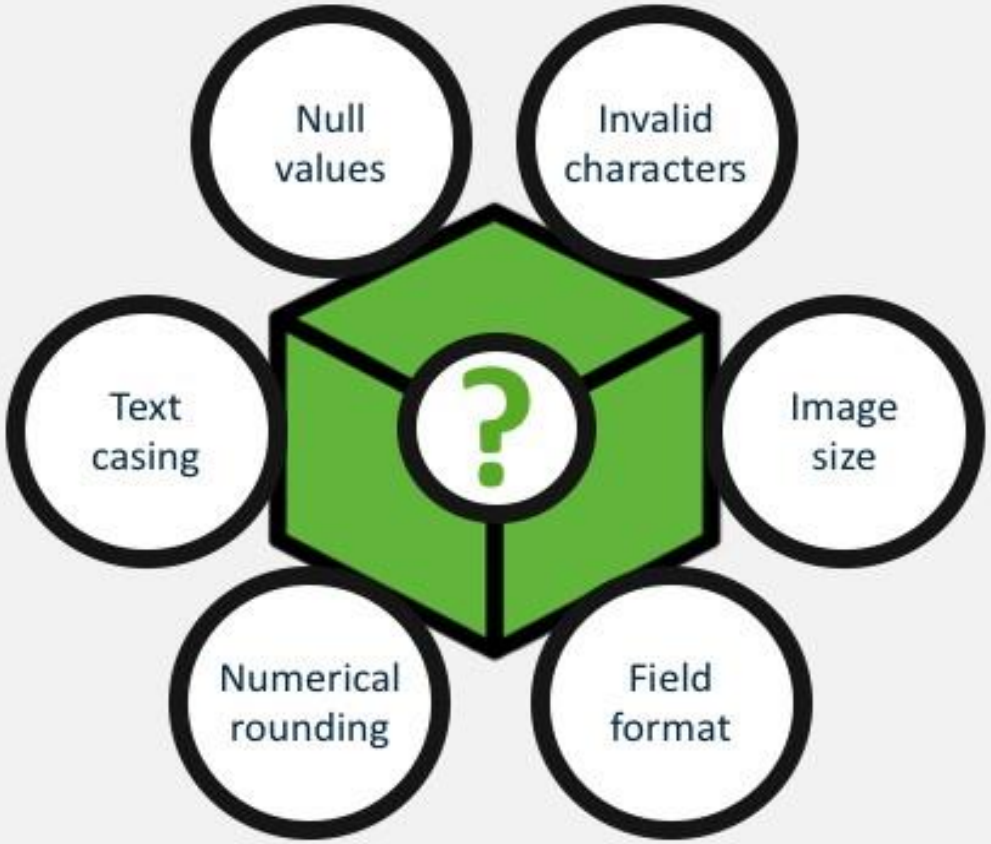
Considerations:

- Unstructured vs. Structured data formats
- Missing values, invalid characters % ! ^ ;*_
- Numerical rounding or text casing
- Image/video resolution, rotation, sizes

2 Data Preparation – Pre-Process Data

- Addresses questions raised during exploration phase
- Integrity of original data should be protected
- Never overwrite original datasets
- Use snapshots, mirrors, or version control to protect raw data

Pre-processing Examples



2 Data Pre-processing – Example Use Case

Client Name	Transaction ID	Purchase Date	Purchase TimeStamp	Address Field 1	State	Zip	Client Email
Sara Kendy	10294..	10/24/18	12:44:32	1011 First A.	California	94402	smkend@y...
sara kendy	10283..	06/4/17	2:50:42%00	1011 1st Av..	CA	94402	smkend@y...
Chip Ray	10291..	2018/1/4	04:03:33>	22 Beverly..	ND	-	chips@gma..
D. Little	10290..	2/22/17	12:54:02	848 Gamin..	CO	-	lild@hotmai..
Jon Allen	10292..	4/5/2018	08:06:55	447 Pieland.	CAL	90523	jonizkewl@h.
Jon allen	10293..	09/16/18	^7:^3:20	7 Butan Ro..	MN	40528	jonizkewl@h.

2 Data Pre-processing – Example Use Case

Client Name	Transaction ID	Purchase Date	Purchase TimeStamp	Address Field 1	State	Zip	Client Email
sara kendy	10294..	10/24/18	12:44:32	1011 First A.	CA	94402	smkend@y...
sara kendy	10283..	06/04/17	02:50:42	1011 1st Av..	CA	94402	smkend@y...
chip ray	10291..	01/04/18	04:03:33	22 Beverly..	ND	0	chips@gma..
d little	10290..	02/22/17	12:54:02	848 Gamin..	CO	0	lild@hotmai..
jon allen	10292..	04/05/18	08:06:55	447 Pieland.	CA	90523	jonizkewl@h.
jon allen	10293..	09/16/18	07:03:20	7 Butan Ro..	MN	40528	jonizkewl@h.

2 Data Transformations: Unstructured to Structured Data

Data Type: Text
Transformation: Vectorizing

Unstructured Structured

Data Type: Image
Transformation: Labeling

Unstructured Structured

Data Type: Audio
Transformation: Signal Processing

Unstructured Structured

Data Type: Video
Transformation: Labeling

Unstructured Structured

2 Data Transformations: Dimensionality Reduction

Dimensionality Reduction Algorithms:

- Principal Component Analysis (PCA)
- Linear Discriminant Analysis (LDA)
- Singular value Decomposition (SVD)

Learning Goals

1.1 AI/ML Overview

1.2 Step 1: Identify Business Need and Plan

1.3 Step 2: Data Management

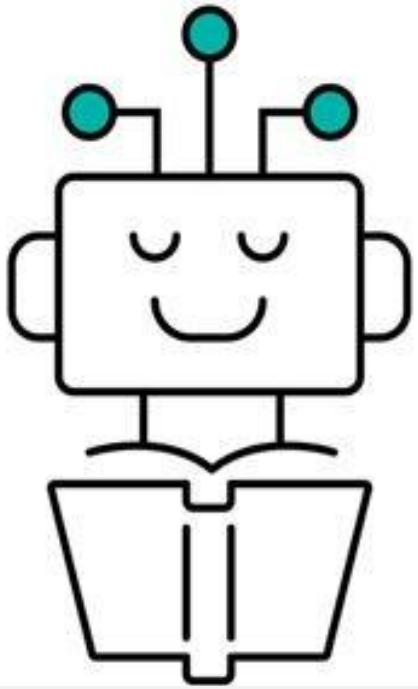
1.4 Step 3: Feature Selection and Engineering

Project Plan Workflow: Step 3

Part I (Preparation)		
Plan	Prepare	Features
1	2	3
Identify business need and create a project plan	Manage and prepare data	Select and engineer features for the model
		

3 Feature Selection

Select the features from your data that will provide the most value to your project. The model may work perfectly, but if the wrong features were selected for training, it won't produce useful results.



Consider the intention of your original project goal.

Examples of goals that may require very different features for selection:

- Increasing cross-sells
- Increasing click rates
- Increasing page visit durations

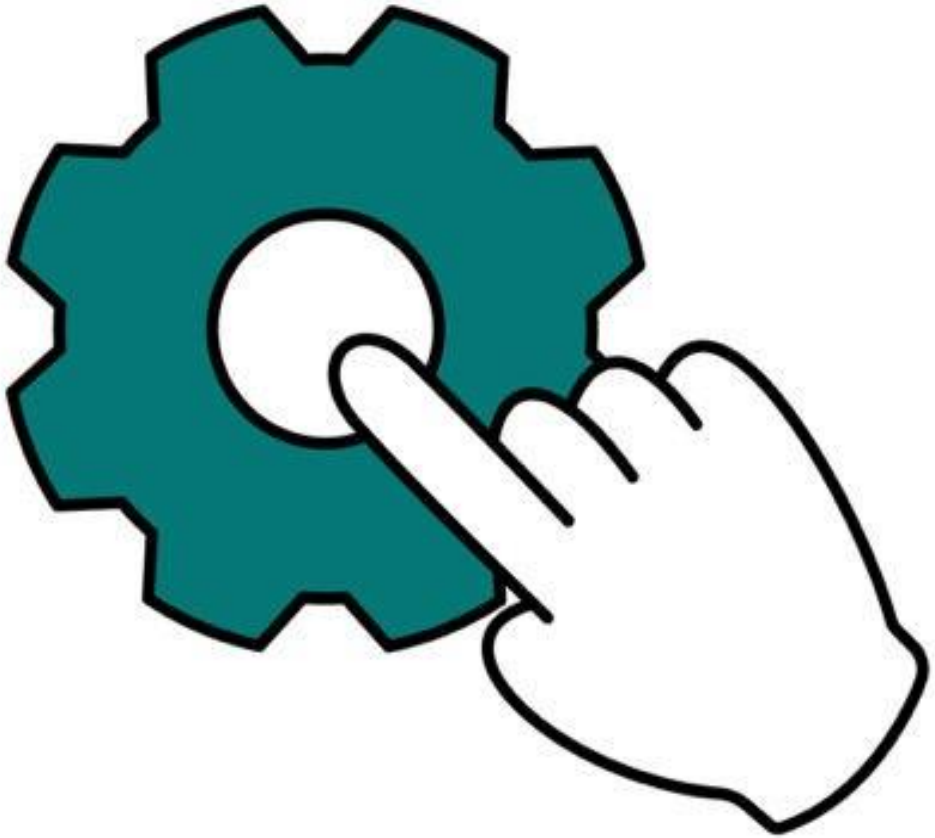
3 Feature Selection

Feature Selection Tips	Examples
Useful vs. useless features	Do you need to keep all physical characteristic data, or just some? Do you need all user data like shipping information?
Avoid redundant features	Measurement fields showing in multiple formats: in/cm, lbs/kg
Keep it simple	Age vs. date of birth fields, latitude/longitude coordinates vs. simplified address or zip code

3 Feature Selection

Client Name	Transaction ID	Purchase Date	Purchase TimeStamp	Address Field 1	State	Zip	Client Email	Product(s)
sara kendy	10294..	10/24/18	12:44:32	1011 First A.	CA	94402	smkend@y...	Dog XL Bed Com..
sara kendy	10283..	06/04/17	02:50:42	1011 1st Av..	CA	94402	smkend@y...	10Pk Bone Flavo..
chip ray	10291..	01/04/18	04:03:33	22 Beverly..	ND	0	chips@gma..	Premium Wate..
d little	10290..	02/22/17	12:54:02	848 Gamin..	CO	0	lild@hotmai..	AquaSafe Fish B..
jon allen	10292..	04/05/18	08:06:55	447 Pieland.	CAL	40528	jonizkewl@h.	PurrSnickety toy
jon allen	10293..	09/16/18	07:03:20	7 Butan Ro..	MN	40528	jonizkewl@h.	Cat XL Bee Costu..

3 Feature Selection - Example Use Case



Potential Features to Train Recommendation Model

- Unique users and items
- User purchase history
- Product run rates
- Any available ratings or product reviews by user
- User viewing history (web logs)

3 Feature Selection – Example Use Case: Reviews

★★ "Loved the product but the color didn't look good on me."

★★★★ "I haven't worn it yet, but it arrived the next day!"

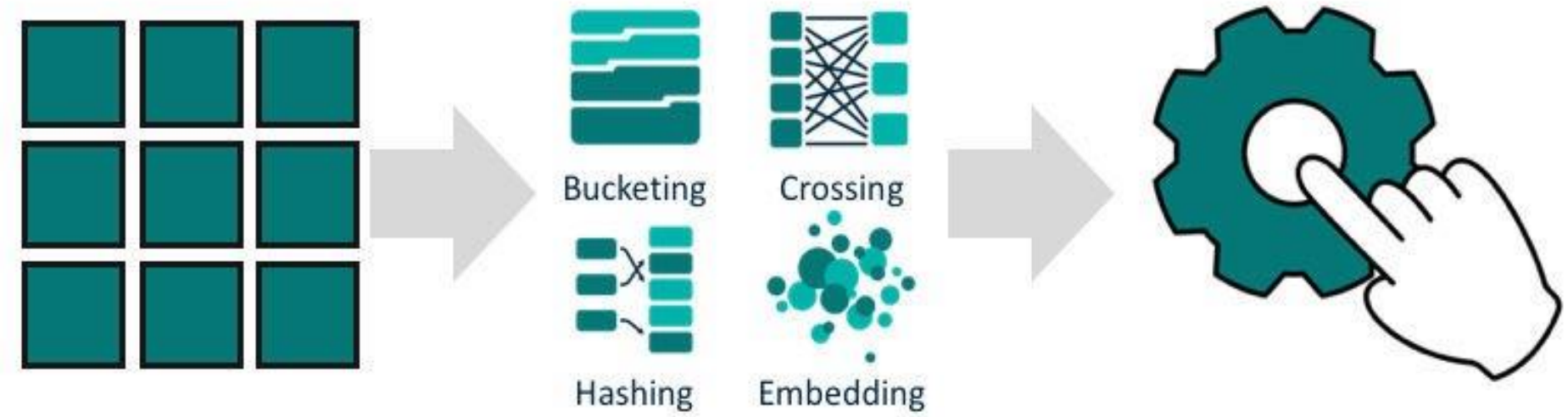
★ "The box was crushed when it arrived."

★★ "I never actually ordered this, my computer crashed while I was ordering."

★ "I bought this for my boyfriend, and then he broke up with me."

3 Feature Engineering

Feature engineering is the transformation of raw data into inputs for your algorithm, along with creating composite features.




MAPR academy
Build Machine Learning Projects
Lesson 2: Implement to Production

Learning Goals

2.1 Step 4: Algorithm and Framework Selection

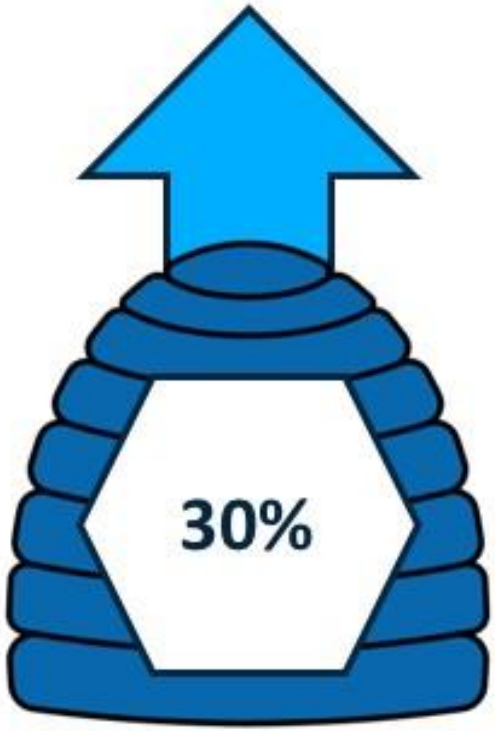
2.2 Step 5: Model Training and Validation

2.3 Step 6: Implementation to Production and Monitoring

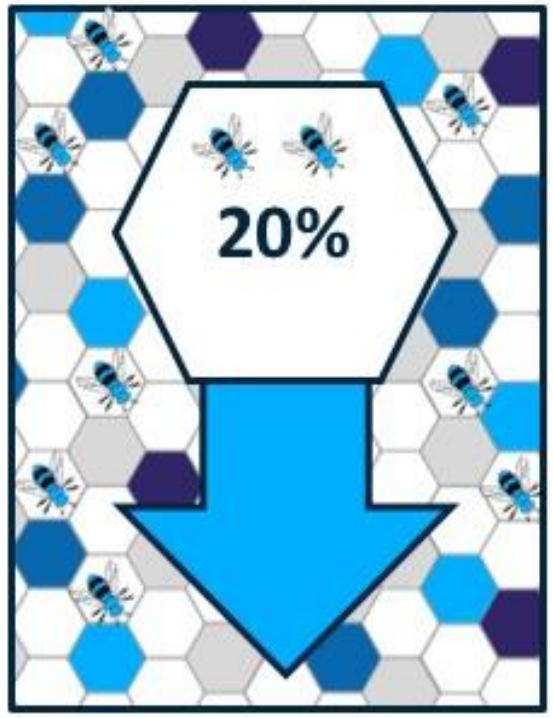

Data Management Logistics
90% of Machine Learning effort is all about
Data Logistics
If not done well, it can easily cause you to fail.

Artificial Intelligence Adopters

Profit Advantage



Overall Revenue



Production

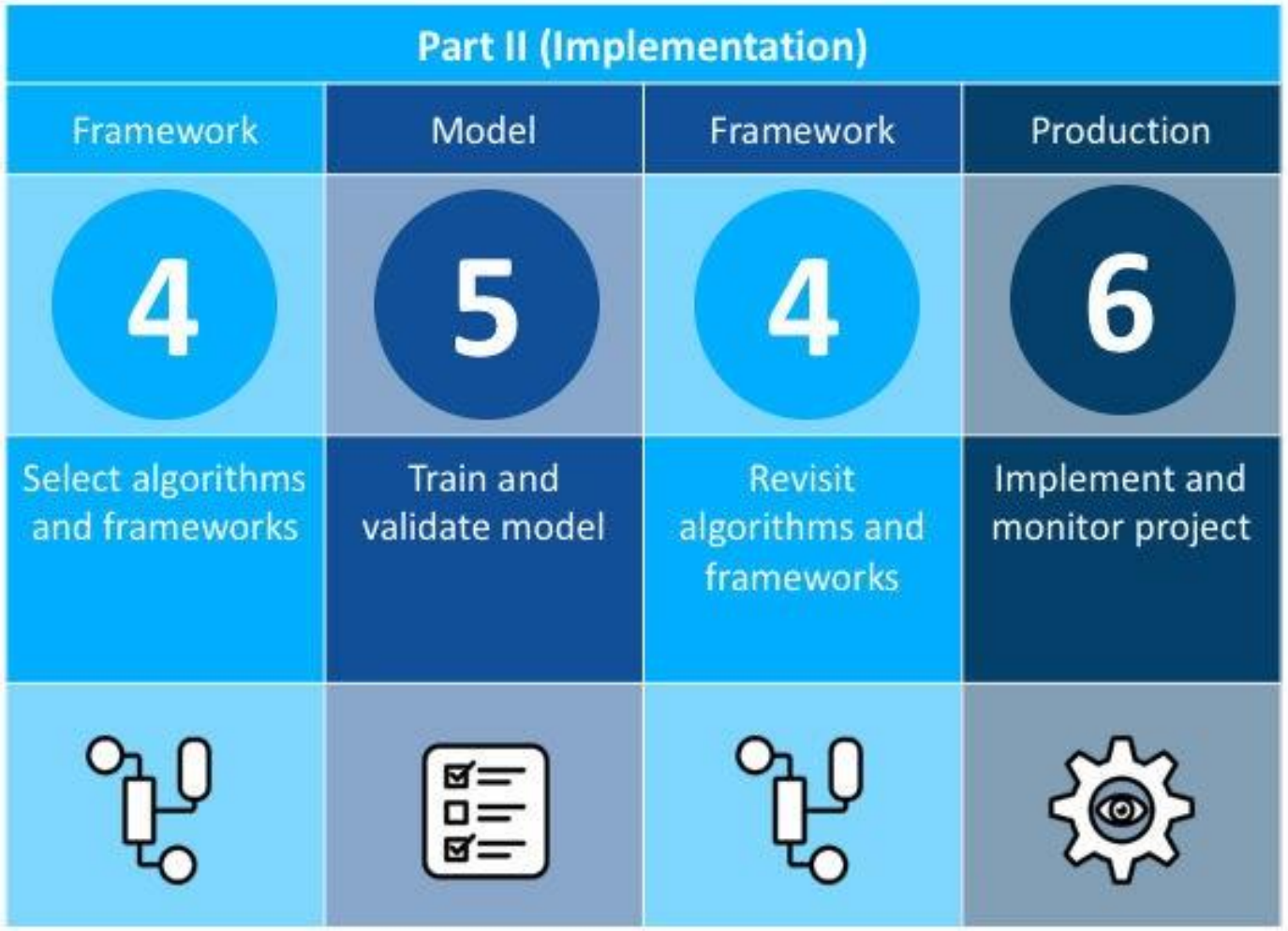
Full Project Plan Workflow

Part I (Preparation)			Part II (Implementation)		
Plan	Prepare	Features	Framework	Model	Production
1	2	3	4	5	6
Identify business need and create a project plan	Manage and prepare data	Select and engineer features for the model	Select algorithms and frameworks	Train and validate model	Implement and monitor project
					

Project Plan Workflow: Step 4

Part II (Implementation)		
Framework	Model	Production
4	5	6
Select algorithms and frameworks	Train and validate model	Implement and monitor project
		

Project Plan Workflow



- Not always linear
- Steps may intermingle or become iterative

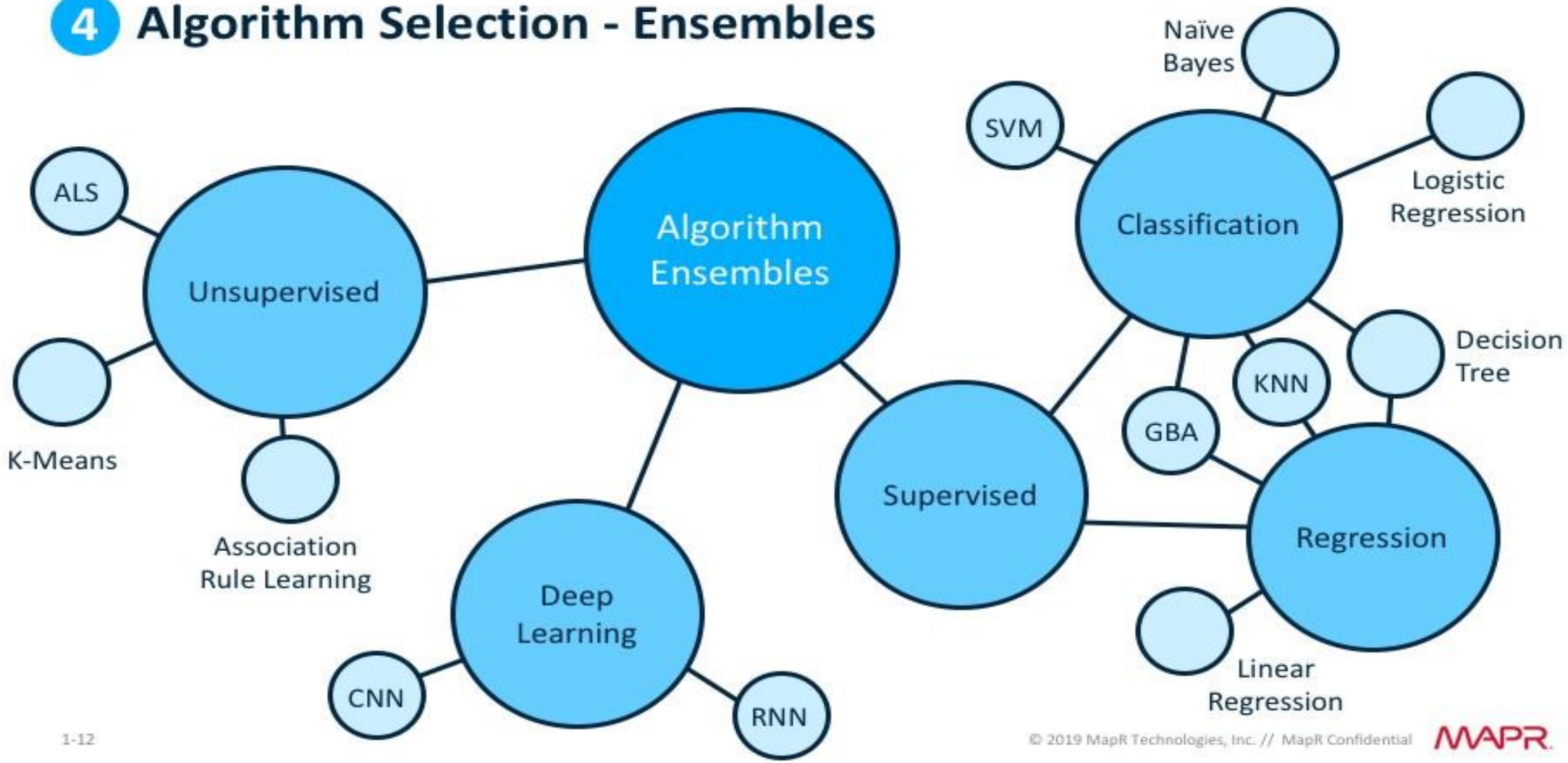
4 Algorithm and Framework Selection

Learning Method	Algorithm Type	Algorithm Name
Supervised	Classification	Naïve Bayes
Supervised	Classification	Logistic Regression
Supervised	Classification	Support Vector Machines (SVM)
Supervised	Regression	Linear Regression
Supervised	Classification/Regression	Decision Trees/Random Forest
Supervised	Classification/Regression	K-Nearest Neighbor (KNN)
Supervised	Classification/Regression	Gradient Boosting Algorithms (GBA)
Unsupervised	Unsupervised	K-Means: Cluster Analysis
Unsupervised	Unsupervised	Association Rule Learning
Unsupervised	Unsupervised	Alternating Least Squares (ALS)

4 Algorithm and Framework Selection

Algorithm Type	Learning Method	Algorithm Name
Deep Learning	Supervised/Unsupervised	Recurrent Neural Network (RNN)
Deep Learning	Supervised/Unsupervised	Convolutional Neural Network (CNN)

4 Algorithm Selection - Ensembles



MAPR

4 Framework Selection

- Determine which ML frameworks to use
- Tools and frameworks should allow for flexibility in your pipeline
- This is only a subset of available libraries
- Will continue to change and grow as the field evolves

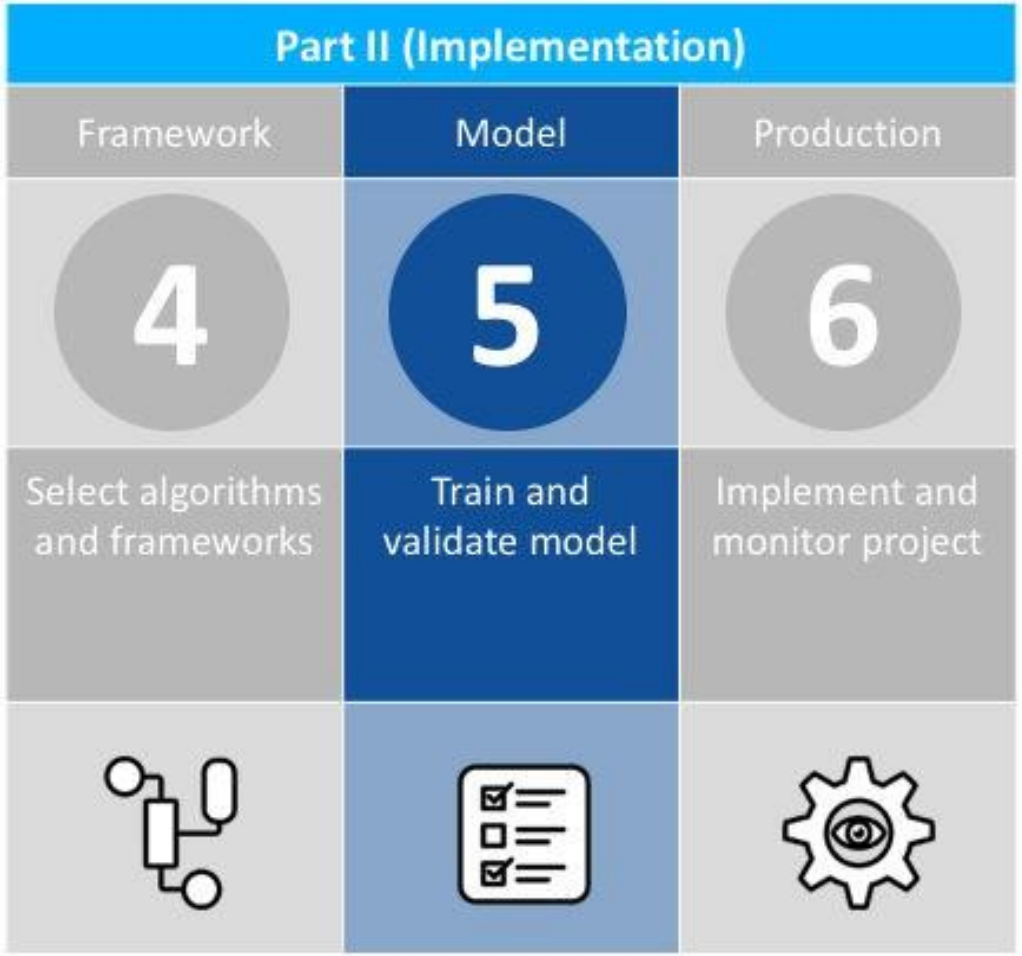
Learning Goals

2.1 Step 4: Algorithm and Framework Selection

2.2 Step 5: Model Training and Validation

2.3 Step 6: Implementation to Production and Monitoring

Project Plan Workflow: Step 5



Steps to Train and Validate Model

A. Split data into Train vs. Test datasets: Load "Train" dataset

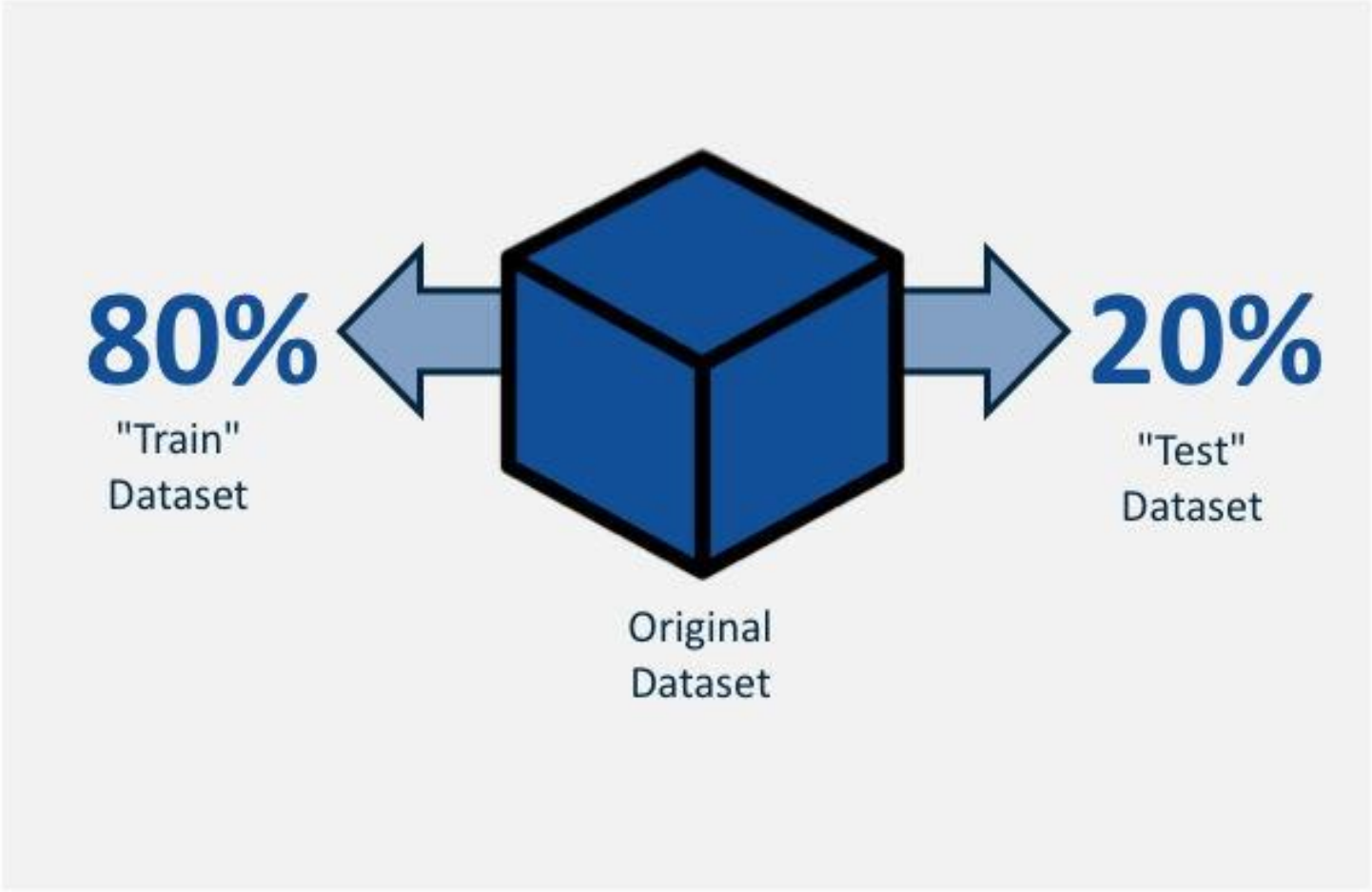
B. Create decoy model

C. Model is trained on "Train" dataset

D. Validate and test the model by loading the "Test" dataset

5 Supervised Model Training and Validation

A. **Split data into Train vs. Test datasets: Load "Train" dataset**
B. Create decoy model
C. Model is trained on "Train" dataset
D. Validate and test the model by loading the "Test" dataset



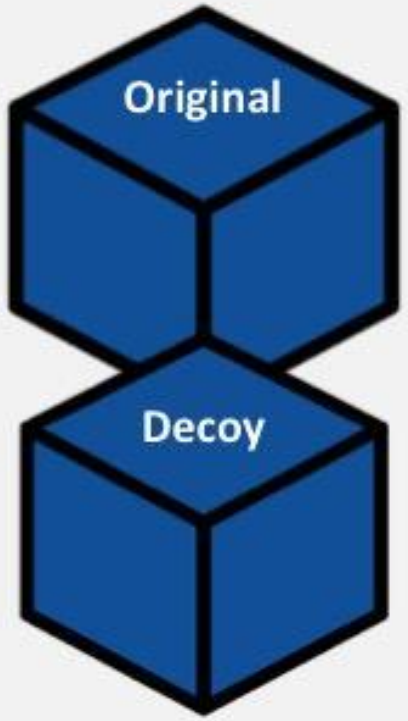
5 Supervised Model Training and Validation

A. Split data into Train vs. Test datasets: Load "Train" dataset

B. Create decoy model

C. Model is trained on "Train" dataset

D. Validate and test the model by loading the "Test" dataset



Decoy "Training" Model

- Preserves specific parameters of the original training of a model
- Accepts exact same training data as the model, but it doesn't actually do anything.
- Saves the information as an archive that can be referenced later, if needed.

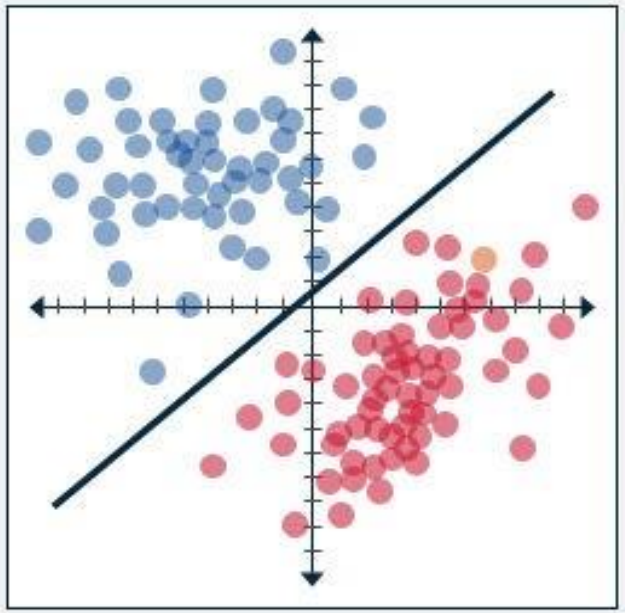
5 Supervised Model Training and Validation

A. Split data into Train vs. Test datasets: Load "Train" dataset

B. Create decoy model

C. **Model is trained on "Train" dataset**

D. Validate and test the model by loading the "Test" dataset



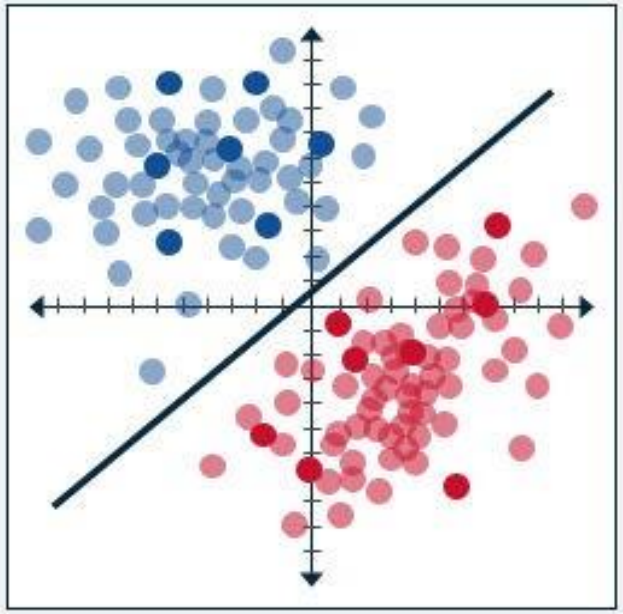
5 Supervised Model Training and Validation

A. Split data into Train vs. Test datasets: Load "Train" dataset

B. Create decoy model

C. Model is trained on "Train" dataset

D. Validate and test the model by loading the "Test" dataset



5 Supervised Model Training and Validation

A. Split data into Train vs. Test datasets: Load "Train" dataset

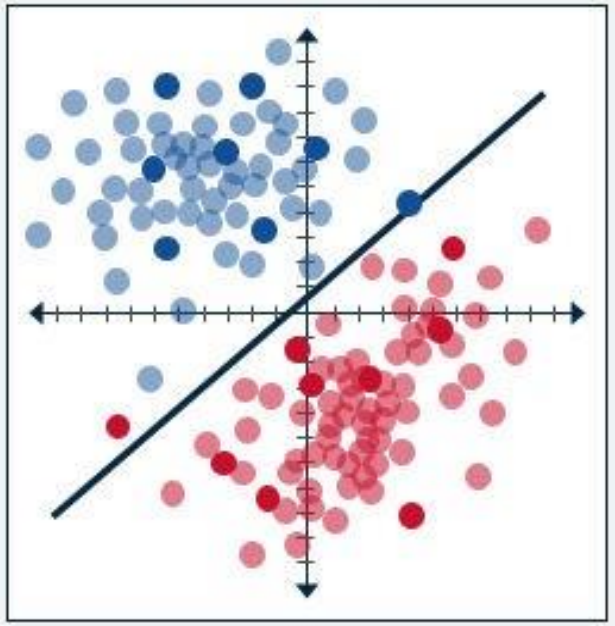
B. Create decoy model

C. Model is trained on "Train" dataset

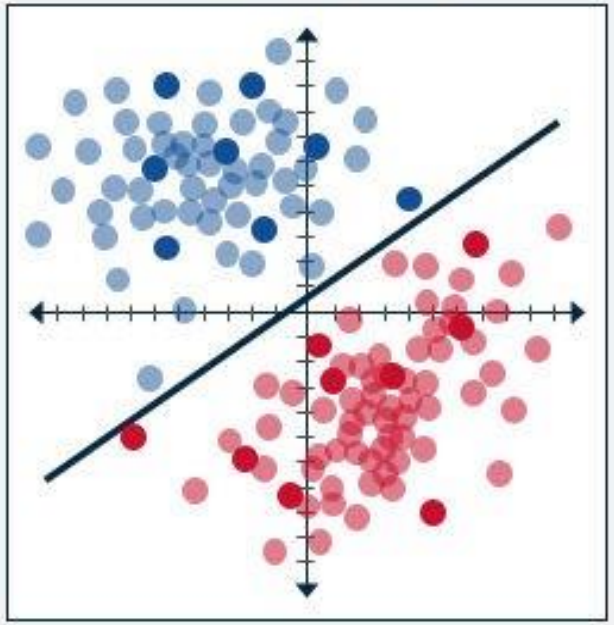
D. Validate and test the model by loading the "Test" dataset

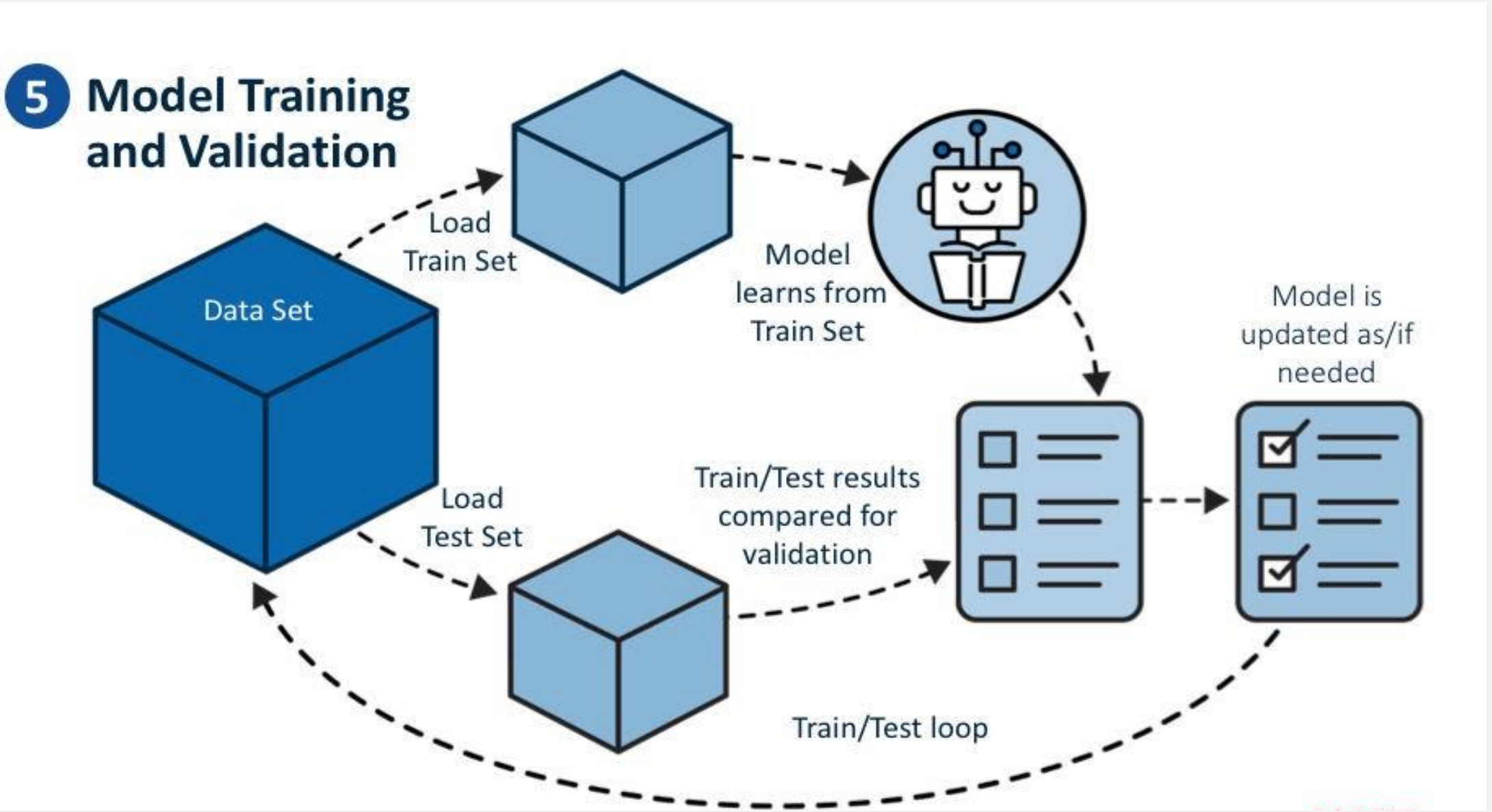
Validate: Retrain/update model parameters until required accuracy levels are met.

Trained Model



Validated Model




5 Model Training and Validation
Data Set
Load Train Set
Model learns from Train Set
Model is updated as/if needed
Load Test Set
Train/Test results compared for validation
Train/Test loop

Learning Goals

2.1 Step 4: Algorithm and Framework Selection

2.2 Step 5: Model Training and Validation

2.3 Step 6: Implementation to Production and Monitoring

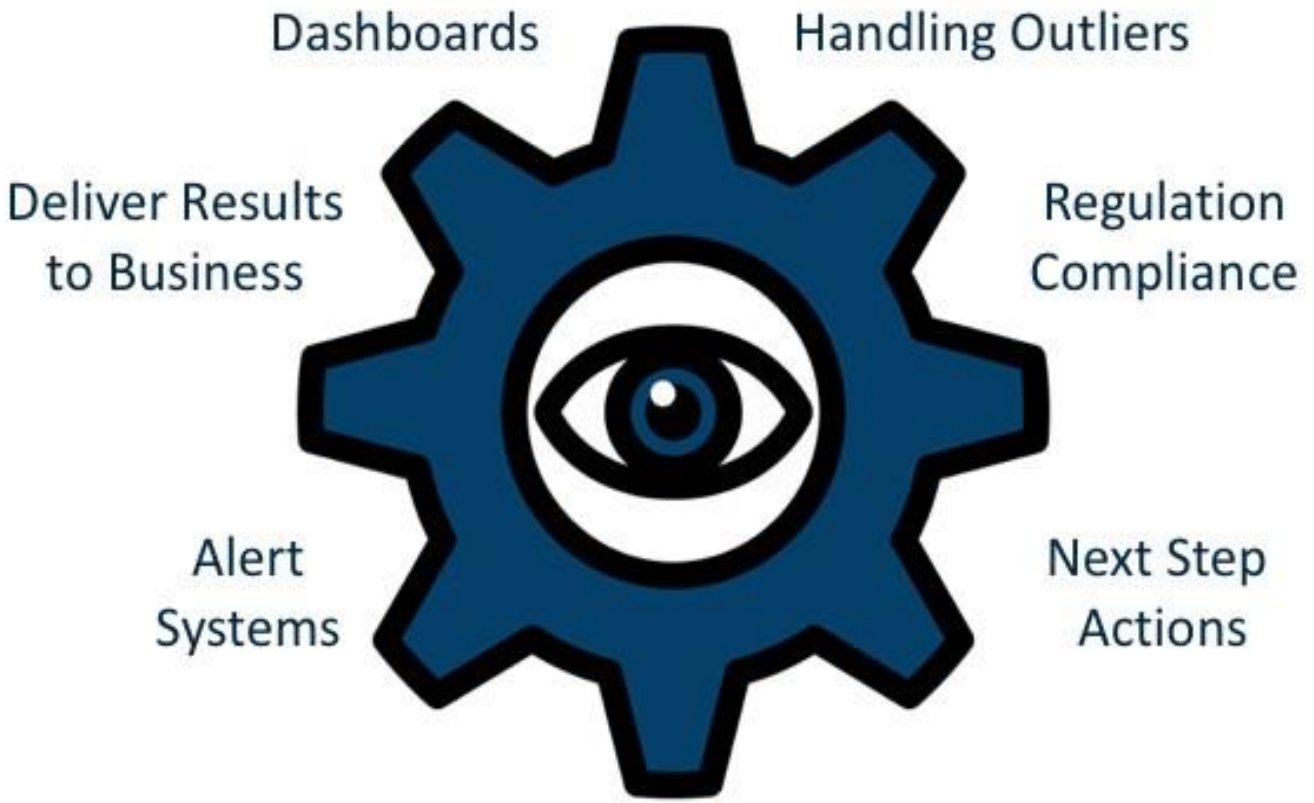
Project Plan Workflow: Step 6

Part II (Implementation)		
Framework	Model	Production
4	5	6
Select algorithms and frameworks	Train and validate model	Implement and monitor project
		

The Full Project Plan Workflow

Part I (Preparation)			Part II (Implementation)		
Plan	Prepare	Features	Framework	Model	Production
1	2	3	4	5	6
Identify business need and create a project plan	Manage and prepare data	Select and engineer features for the model	Select algorithms and frameworks	Train and validate model	Implement and monitor project
					

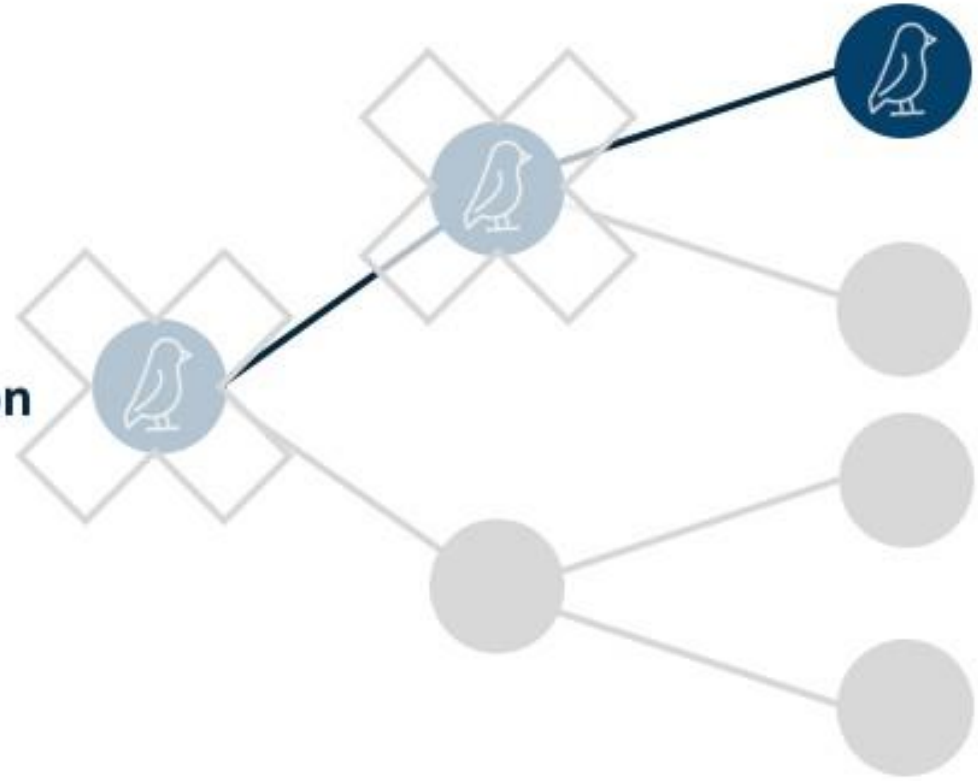
6 Implementation to Production and Monitoring



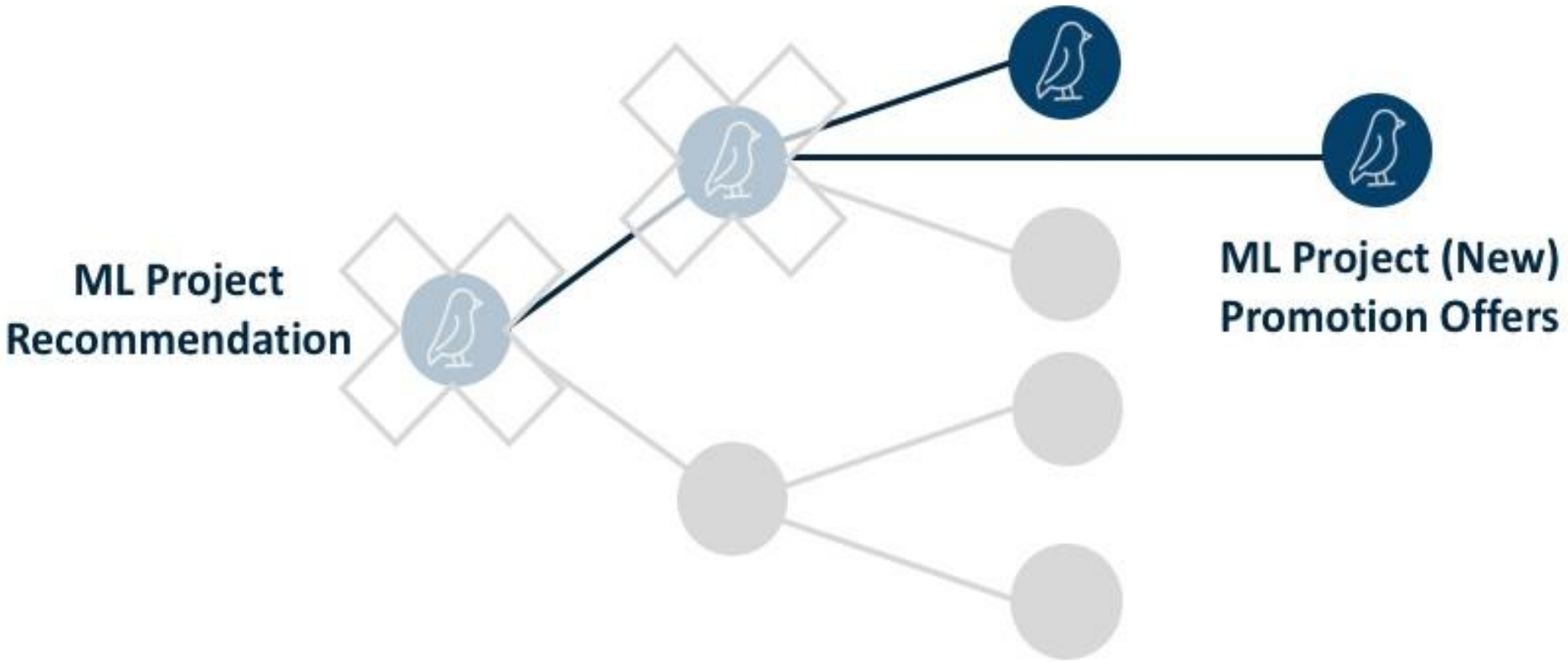
6 Implementation: Launch Plan

6 Monitoring Results

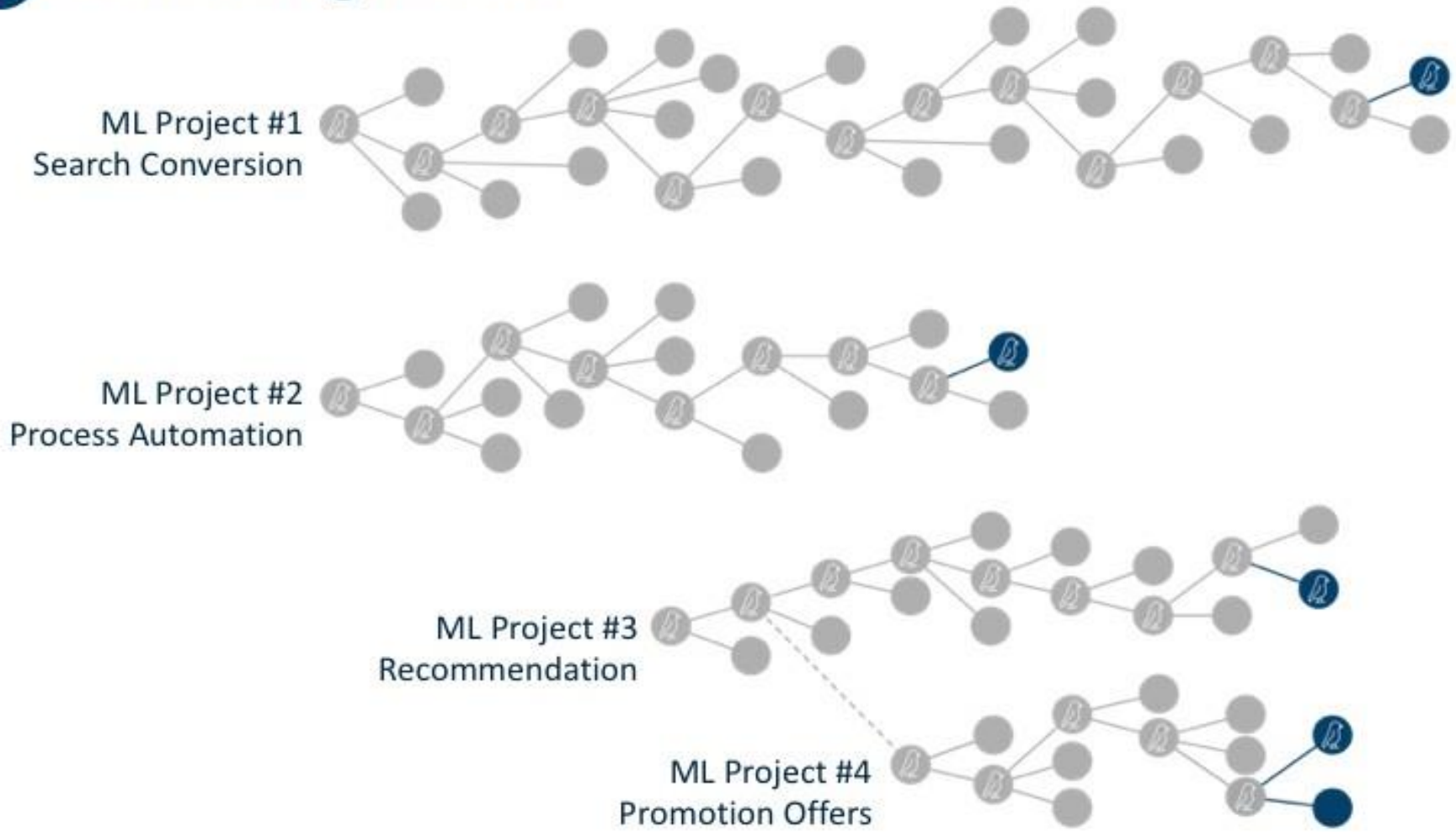
ML Project Recommendation



6 Monitoring Results



6 Monitoring Results



Learning Goals

2.1 Step 4: Algorithm and Framework Selection

2.2 Step 5: Model Training and Validation

2.3 Step 6: Implementation to Production and Monitoring

A Strong Foundation