AI200: APPLIED MACHINE LEARNING

INTRODUCTION TO MACHINE LEARNING

BROAD IDEA OF MACHINE LEARNING

Real Estate

Real Estate Evaluator

Consider features such as location, floor size

A Machine Learning model could similarly make a prediction of the house’s valuation based on its features.

Evaluation

Predicts house is worth **$xxx,xxxx.xx**

BROAD IDEA OF MACHINE LEARNING

- In machine learning, data comes in the form of:
  - **Outcome** we want to predict and,
  - **Features** that we will use to predict the outcome

Features **Outcome**

S/N	sqm	bedrooms	price
1	100	3	300200
2	120	4	25000
		. . .	
N	105	3	300000

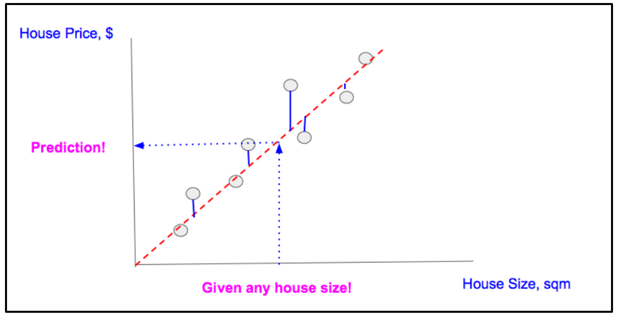
BROAD IDEA OF MACHINE LEARNING

- The idea behind the machine learning approach is to:
  - Train an algorithm/ML model using a dataset for which we do know the **outcome**

	Features		Outcome
S/N	sqm	bedrooms	price
1	100	3	300200
2	120	4	250000
		. . .	
N	105	3	300000

Train Model on Data →

Algorithm / ML Model



BROAD IDEA OF MACHINE LEARNING

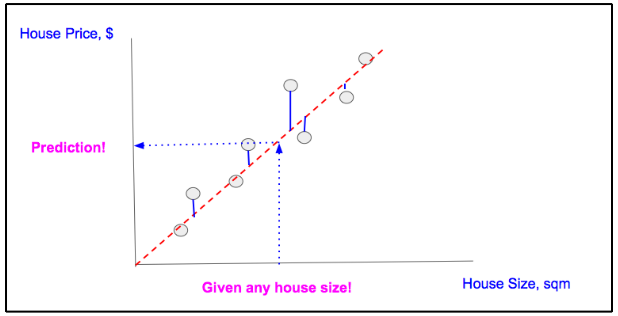
- The idea behind the machine learning approach is to:
  - Train an algorithm/ML model using a dataset for which we do know the **outcome**
  - And then use this algorithm/ML model on the **available features** in the future to make a prediction for when we don't know the **outcome**

Features **Outcome**

S/N	sqm	bedrooms	price
1	100	3	300200
2	120	4	250000
		. . .	
N	105	3	300000

Train Model on Data

Algorithm / ML Model



Predict Outcome for Data

In Future

Available Features **Outcome**

S/N	sqm	bedrooms	price
1	200	4	?
2	210	4	?
		. . .	
N	215	4	?

Predicted Outcome

S/N	sqm	bedrooms	price
1	200	4	400000
2	210	4	410000
		. . .	
N	215	4	420000

COMMONLY USED LINGOS/NOTATIONS IN MACHINE LEARNING

- Throughout the course, you will see us frequently use:
  - y to denote **outcome** and
  - X (x_1, x_2 … x_p) to denote the **features**

- There are many other synonyms for outcomes and features, which sometimes is what makes ML so confusing. For this course, and for your future correspondence with other Data Scientists, stick to **Outcome (y)** and **Features (*x*₁, *x*₂ … *x*ₚ)**. This is the most used terminology.

Features (X_1, X_2…X_p) / Predictors / Covariates / Explanatory Variable / Independent Variable

	X_1	X_2	y
S/N	sqm	bedrooms	price
1	100	3	300200
2	120	4	250000
		. . .	
N	105	3	300000

Outcome (y) / Response Variable / Dependent Variable / Target Variable

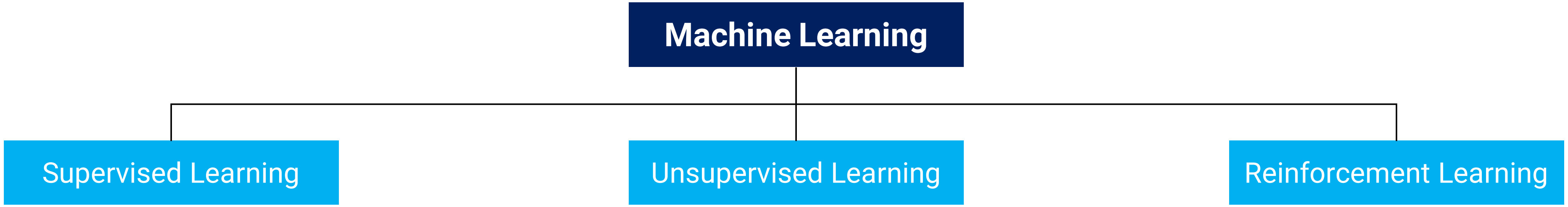
OVERVIEW & LITERATURE OF MACHINE LEARNING: LEARNING SPECTRUM

- Earlier, we provided a general intuition for understanding Machine Learning. However, the truth is there are numerous algorithms/ML model each with their own intricacies, and they solve very different kind of problems.

- As such, there is a need to be able to categorize this wide range of algorithms and ML models based on their use cases. In the next slide, we provide a synthesis of how various ML terms fit into the big picture.

OVERVIEW & LITERATURE OF MACHINE LEARNING: LEARNING SPECTRUM

1. Spectrum of supervision



Machine Learning

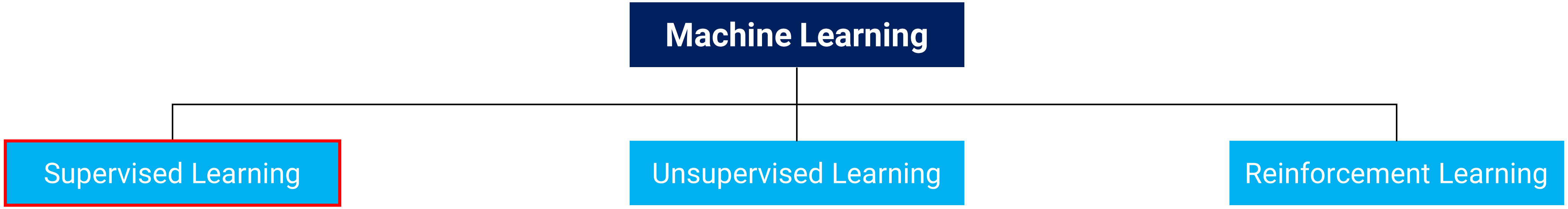
Supervised Learning

Unsupervised Learning

Reinforcement Learning

OVERVIEW & LITERATURE OF MACHINE LEARNING: LEARNING SPECTRUM

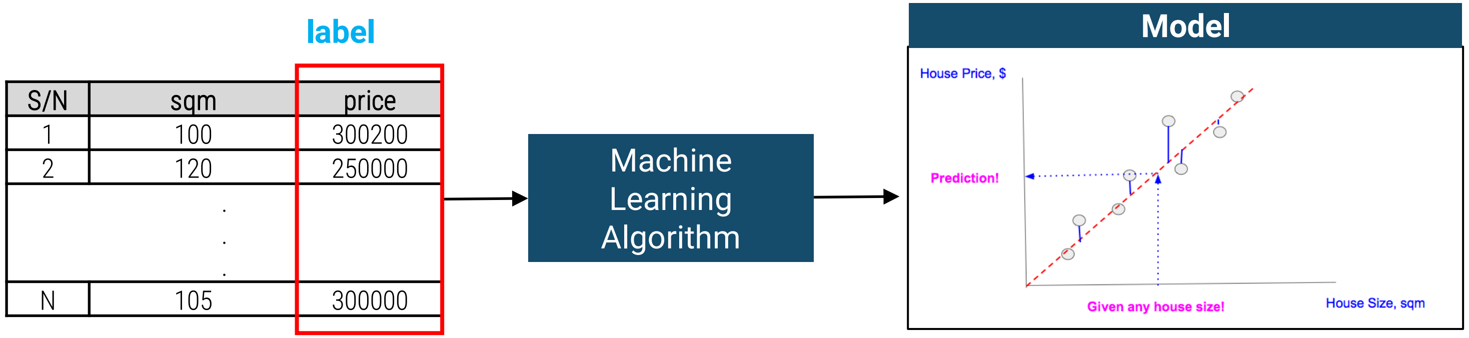
1. Spectrum of supervision



Supervised Learning

Supervised learning is a type of ML where the model is provided with **labeled** training data. This labeled data acts as a teacher and in training the model. Once the model gets trained it can start making a prediction or decision when new data is given to it. Label is the one thing that we are interested in predicting.

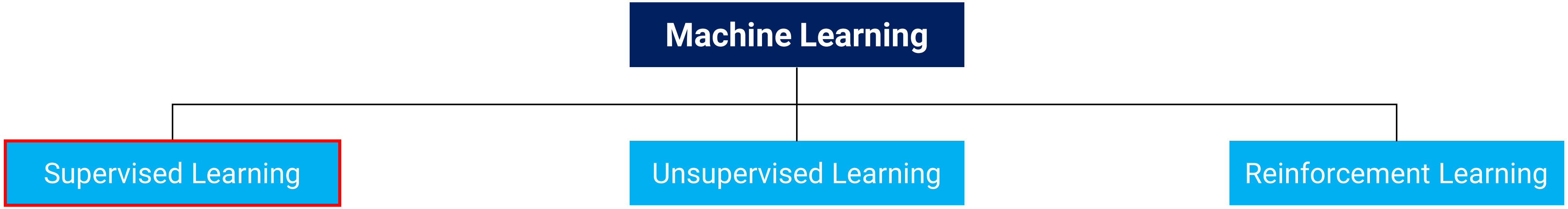
Training



S/N	sqm	price
1	100	300200
2	120	250000
	. . .	
N	105	300000

OVERVIEW & LITERATURE OF MACHINE LEARNING: LEARNING SPECTRUM

1. Spectrum of supervision



Supervised Learning

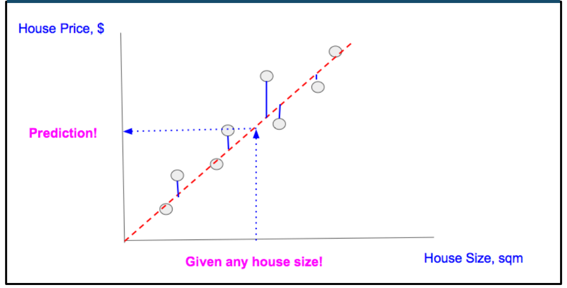
Supervised learning is a type of ML where the model is provided with **labeled** training data. This labeled data acts as a teacher and in training the model. Once the model gets trained it can start making a prediction or decision when new data is given to it. Label is the one thing that we are interested in predicting.

Training

S/N	sqm	price
1	100	300200
2	120	250000
.	.	.
N	105	300000

Machine Learning Algorithm

Model



Prediction

no label

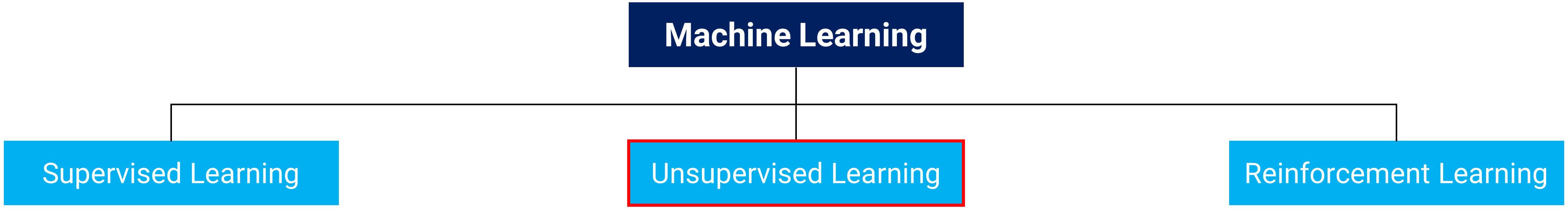
S/N	sqm	price
1	40	?
2	50	?

Predicted outcome

S/N	sqm	price
1	40	170000
2	50	175000

OVERVIEW & LITERATURE OF MACHINE LEARNING: LEARNING SPECTRUM

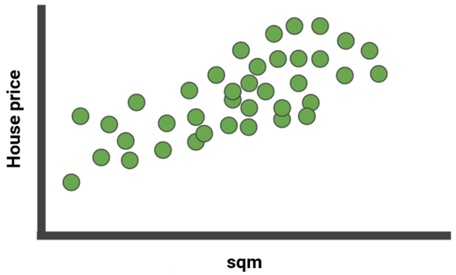
1. Spectrum of supervision



Unsupervised Learning

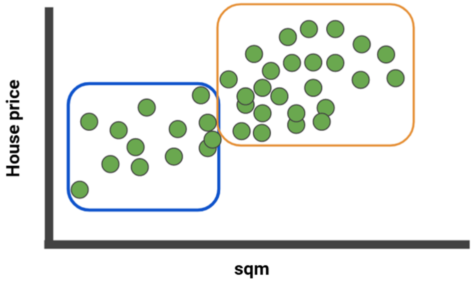
In unsupervised learning, the goal is to identify meaningful patterns in the data. To accomplish this, the machine must learn from an **unlabeled** data set. In other words, the model has no hints how to categorize each piece of data and must infer its own rules for doing so.

S/N	sqm	price
1	100	300200
2	120	250000
	. . .	
N	105	300000



Machine Learning Algorithm

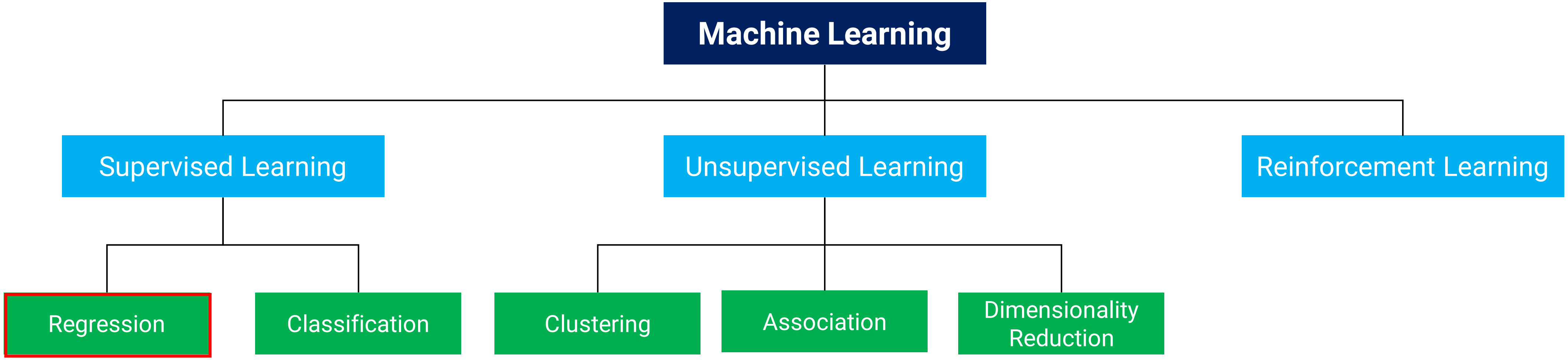
S/N	sqm	price	Cluster
1	100	300200	2
2	120	250000	1
	. . .		
N	105	300000	2



OVERVIEW & LITERATURE OF MACHINE LEARNING: TYPES OF PROBLEMS

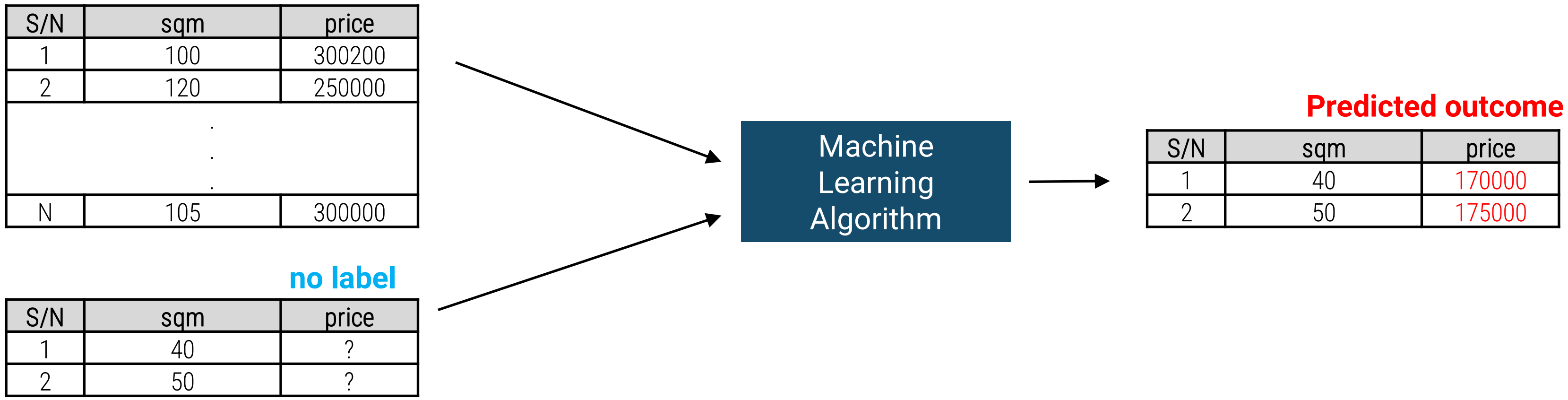
1. Spectrum of supervision

2. Types of problems



Regression

A regression problem is when the output variable is a real or continuous value, such as “price” or “weight”. In layman terms - predict numerical data on a continuous scale.



Training

S/N	sqm	price
1	100	300200
2	120	250000
	. . .	
N	105	300000

Prediction

no label

S/N	sqm	price
1	40	?
2	50	?

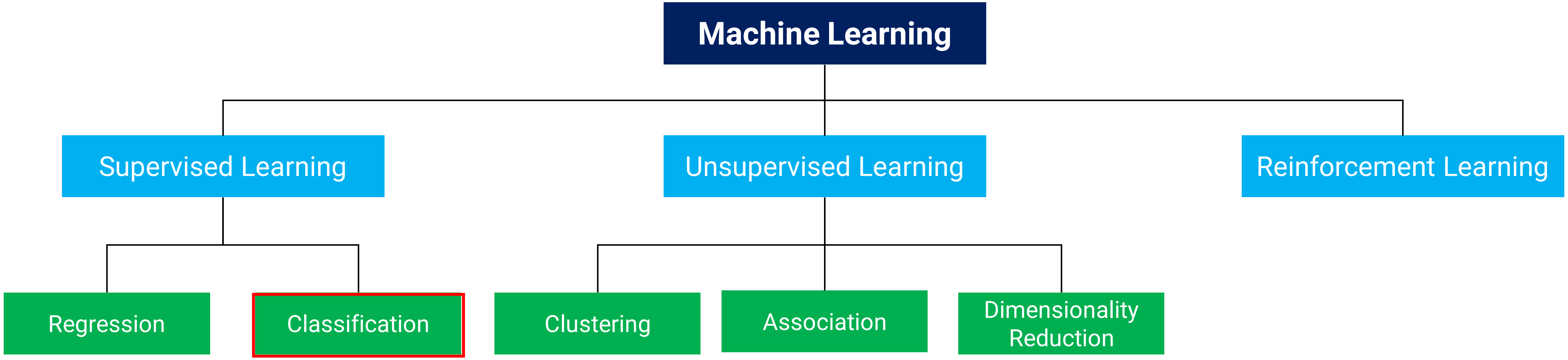
Predicted outcome

S/N	sqm	price
1	40	170000
2	50	175000

OVERVIEW & LITERATURE OF MACHINE LEARNING: TYPES OF PROBLEMS

1. Spectrum of supervision

2. Types of problems



Machine Learning

Supervised Learning

Unsupervised Learning

Reinforcement Learning

Regression

Classification

Clustering

Association

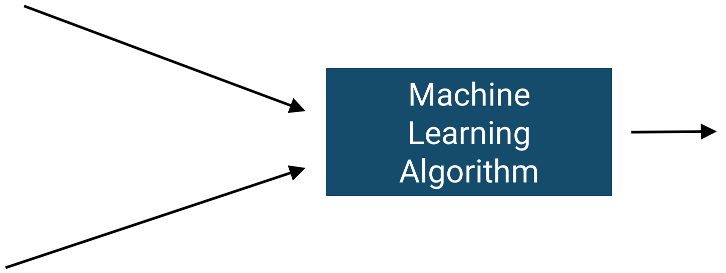
Dimensionality Reduction

Classification

A classification problem is when the output variable is a category, such as “red” or “blue” or “disease” and “no disease”. In layman terms - predict a discrete category.

Training

S/N	sqm	price	Sold?
1	100	300200	0
2	120	250000	1
	⋮		
N	105	300000	1



Machine Learning Algorithm

Predicted outcome

S/N	sqm	price	Sold?
1	40	150000	1
2	50	155000	0

For classification problems, labels are always assigned a numerical value representing the label.

Prediction

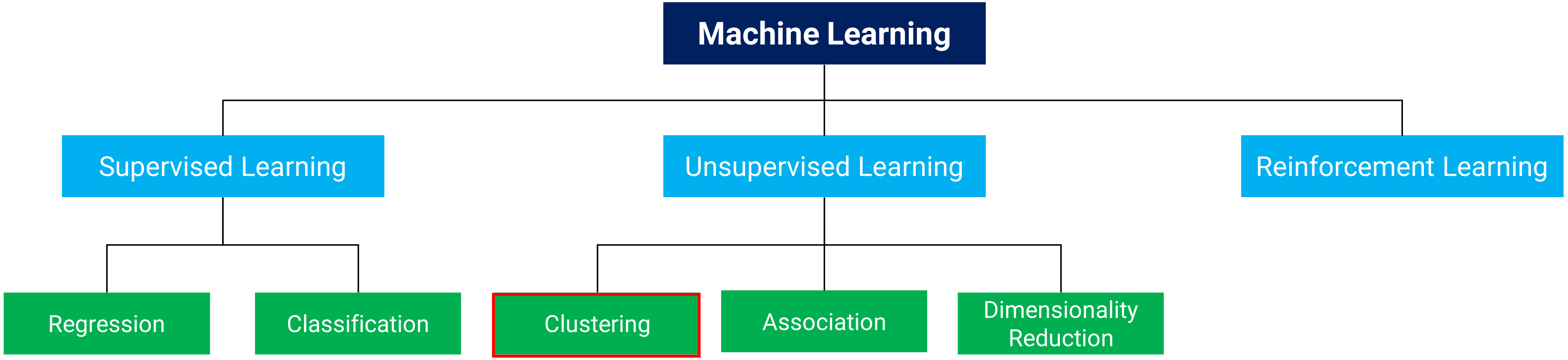
no label

S/N	sqm	price	Sold?
1	40	150000	?
2	50	155000	?

OVERVIEW & LITERATURE OF MACHINE LEARNING: TYPES OF PROBLEMS

1. Spectrum of supervision

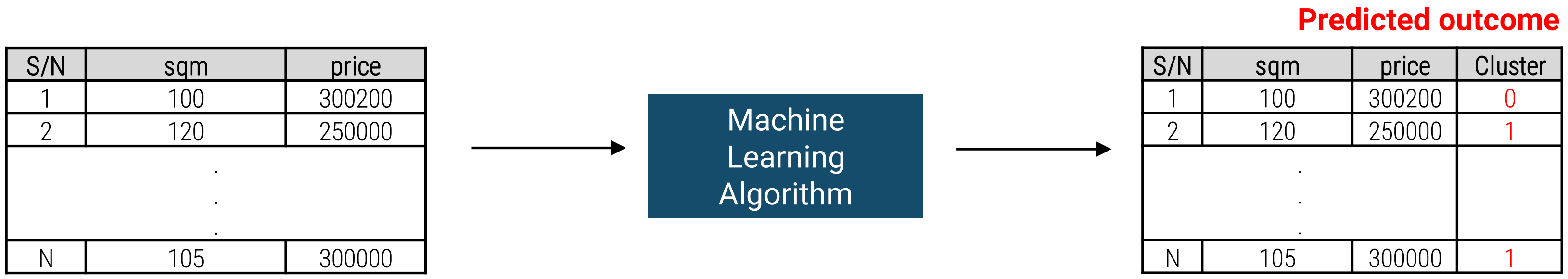
2. Types of problems



Clustering

Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). In layman terms – group similar examples.

Training

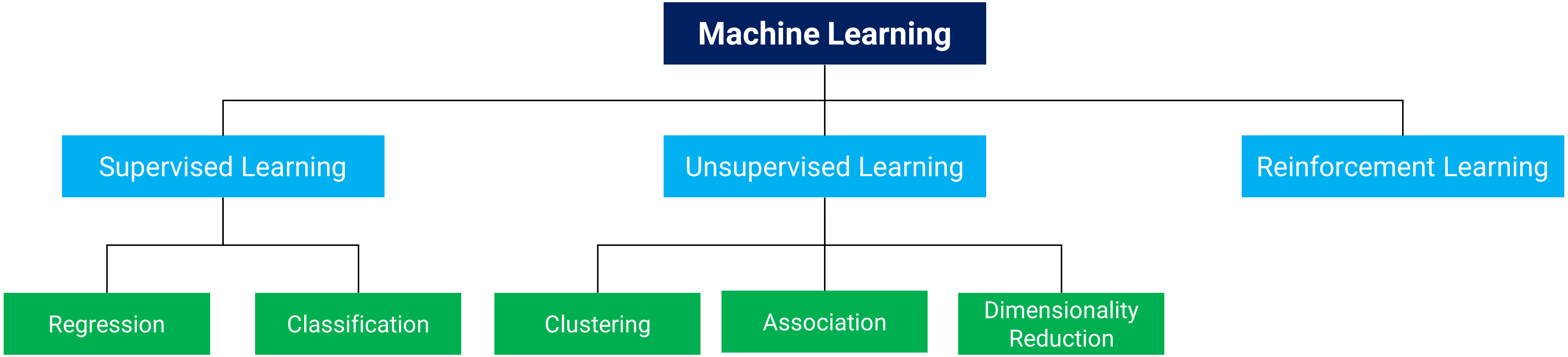


S/N	sqm	price
1	100	300200
2	120	250000
	⋮	
N	105	300000

S/N	sqm	price	Cluster
1	100	300200	0
2	120	250000	1
	⋮		
N	105	300000	1

The difference between classification and clustering is that in clustering, there is no label provided at all.

OVERVIEW & LITERATURE OF MACHINE LEARNING: TYPES OF MODELS



Machine Learning

1. Spectrum of supervision

Supervised Learning

Unsupervised Learning

Reinforcement Learning

2. Types of problems

Regression

Classification

Clustering

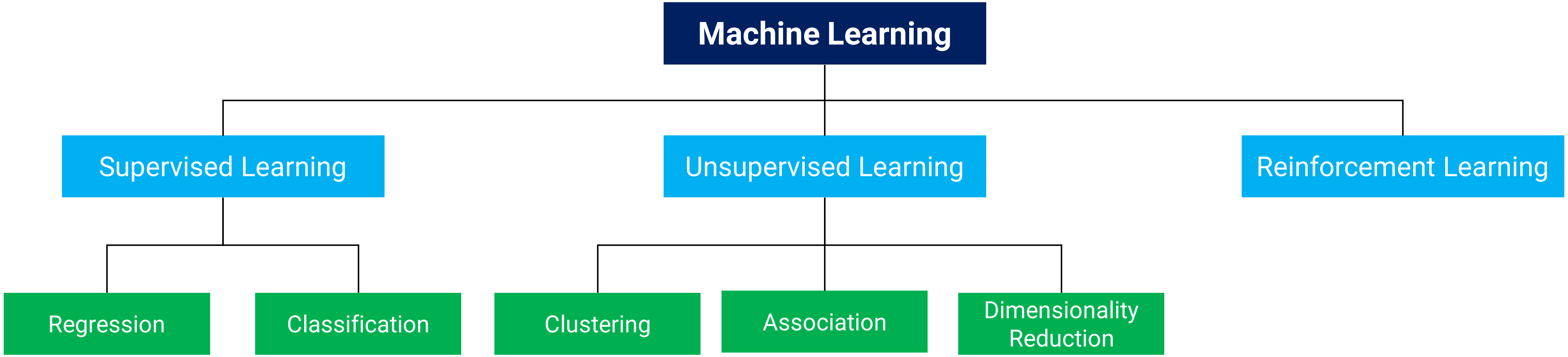
Association

Dimensionality Reduction

3. Types of algorithm / models

Regression	Classification	Clustering	Association	Dimensionality Reduction
Linear regression Polynomial regression Neural Networks Decision Tree Random Forest XGBoost	Logistic Regression KNN Naïve Bayes SVM Neural Networks Decision Tree Random Forest XGBoost	Neural Networks K-Means Hierarchical Clustering Hidden Markov Models Gaussian Mixture	Apriori Algorithm FP-Growth Algorithm Eclat Algorithm	Neural Networks SVD PCA

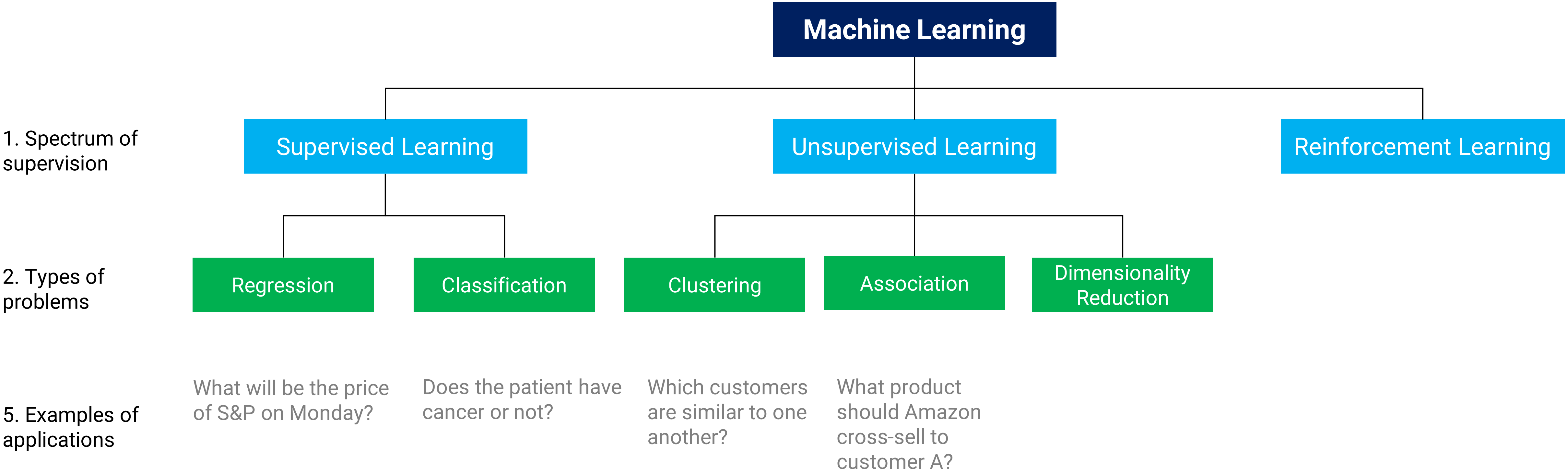
OVERVIEW & LITERATURE OF MACHINE LEARNING: TYPES OF MODELS



Machine Learning

1. Spectrum of supervision	Supervised Learning		Unsupervised Learning			Reinforcement Learning
2. Types of problems	Regression	Classification	Clustering	Association	Dimensionality Reduction	
3. Types of algorithm / models	Linear regression Polynomial regression Neural Networks Decision Tree Random Forest XGBoost	Logistic Regression KNN Naïve Bayes SVM Neural Networks Decision Tree Random Forest XGBoost	Neural Networks K-Means Hierarchical Clustering Hidden Markov Models Gaussian Mixture	Apriori Algorithm FP-Growth Algorithm Eclat Algorithm	Neural Networks SVD PCA	
4. Types of applications	Forecasting Predictions Process optimization New Insights	Fraud Detection Image Classification Customer Retention Diagnostics	Recommendations Targeted Marketing Customer Segmentation	Market Basket Analysis	Feature Elicitation Structure Discovery Meaningful Compression Big Data Visualisation	Real-time Decisions Game AI Learning Tasks Skill Acquisition Robot Navigation

OVERVIEW & LITERATURE OF MACHINE LEARNING: TYPES OF MODELS



Machine Learning

1. Spectrum of supervision

Supervised Learning

Unsupervised Learning

Reinforcement Learning

2. Types of problems

Regression

Classification

Clustering

Association

Dimensionality Reduction

5. Examples of applications

What will be the price of S&P on Monday?

Does the patient have cancer or not?

Which customers are similar to one another?

What product should Amazon cross-sell to customer A?

STATE OF MACHINE LEARNING TODAY

- The realm of machine learning can be overwhelming with so many types of algorithm / ML models and use cases.
- We can liken this to a hawker center:
    - Every hawker center has many types of cuisine (Chinese, Indian, western…etc.)
    - Every type of cuisine has many different types of dishes (chicken rice, carrot cake…etc.)
    - There are several stalls selling the same dish in a hawker center (Ah beng chicken rice, Tian Tian chicken rice…etc.)
- Given our limited time and stomach space, how do we optimize our experience?

AI200's APPROACH TO ML

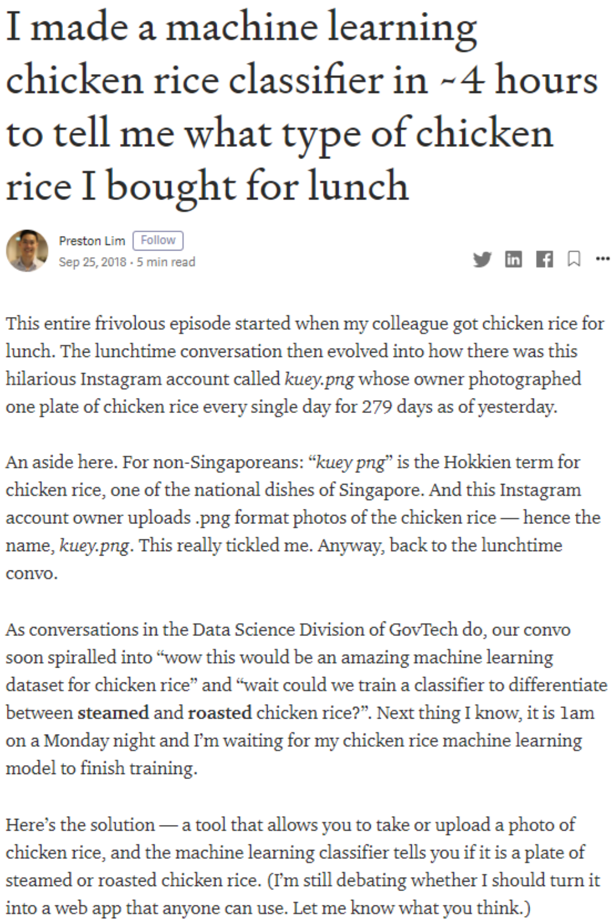
- To optimize our learning experience, we need to be targeted and will focus on:
  - Common ML problems (regression, classification & clustering)
  - Baseline models for each of the ML problems (linear regression, random forest, K-Means), where we will learn:
    - The intuition behind the model (and not the math),
    - The implementation of the model using Scikit-learn
  - How to measure performance of the model, and tune the model to improve performance
  - Application of the models in common but high-impact use cases (e.g. Recommendation Systems for E-Commerce)
  - Key considerations as we go about modelling:
    - Generalizability
    - Bias-Variance Tradeoff
    - Interpretability-Flexibility Tradeoff
    - Feature Engineering

AI200’s APPROACH TO ML

- In a hawker centre, the goal of a discerning connoisseur is to find the best chicken rice rather than try as many different chicken rice as possible. Similarly, in the context of ML, rather than blindly learning as many ML models as we can, we aim to go deep for select important aspects in ML.

Golden words: Be a connoisseur and not a glutton when it comes to both ML and food!

SOMEONE IS MORE OBSESSED WITH CHICKEN RICE THAN US



I made a machine learning chicken rice classifier in ~4 hours to tell me what type of chicken rice I bought for lunch

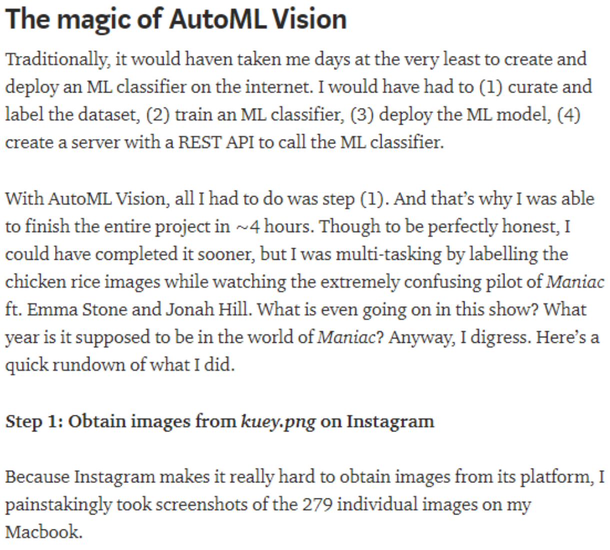
Preston Lim Follow
Sep 25, 2018 · 5 min read

This entire frivolous episode started when my colleague got chicken rice for lunch. The lunchtime conversation then evolved into how there was this hilarious Instagram account called *kuey.png* whose owner photographed one plate of chicken rice every single day for 279 days as of yesterday.

An aside here. For non-Singaporeans: "*kuey png*" is the Hokkien term for chicken rice, one of the national dishes of Singapore. And this Instagram account owner uploads .png format photos of the chicken rice — hence the name, *kuey.png*. This really tickled me. Anyway, back to the lunchtime convo.

As conversations in the Data Science Division of GovTech do, our convo soon spiralled into "wow this would be an amazing machine learning dataset for chicken rice" and "wait could we train a classifier to differentiate between **steamed** and **roasted** chicken rice?". Next thing I know, it is 1am on a Monday night and I'm waiting for my chicken rice machine learning model to finish training.

Here's the solution — a tool that allows you to take or upload a photo of chicken rice, and the machine learning classifier tells you if it is a plate of steamed or roasted chicken rice. (I'm still debating whether I should turn it into a web app that anyone can use. Let me know what you think.)



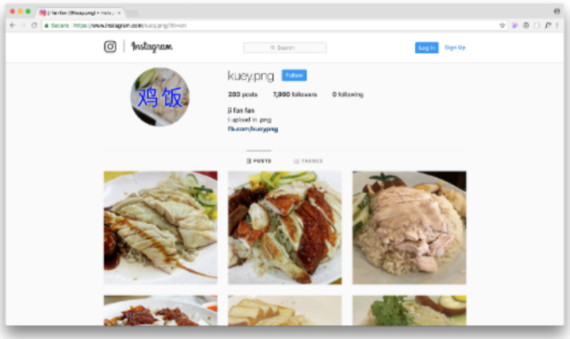
The magic of AutoML Vision

Traditionally, it would haven taken me days at the very least to create and deploy an ML classifier on the internet. I would have had to (1) curate and label the dataset, (2) train an ML classifier, (3) deploy the ML model, (4) create a server with a REST API to call the ML classifier.

With AutoML Vision, all I had to do was step (1). And that's why I was able to finish the entire project in ~4 hours. Though to be perfectly honest, I could have completed it sooner, but I was multi-tasking by labelling the chicken rice images while watching the extremely confusing pilot of *Maniac* ft. Emma Stone and Jonah Hill. What is even going on in this show? What year is it supposed to be in the world of *Maniac*? Anyway, I digress. Here's a quick rundown of what I did.

Step 1: Obtain images from *kuey.png* on Instagram

Because Instagram makes it really hard to obtain images from its platform, I painstakingly took screenshots of the 279 individual images on my Macbook.



Golden words:
Alternatively, you can become a ML Food Connoisseur ☺

Source: https://blog.usejournal.com/i-made-a-machine-learning-chicken-rice-classifier-in-4-hours-to-tell-me-what-type-of-chicken-rice-e9b1af4aa069