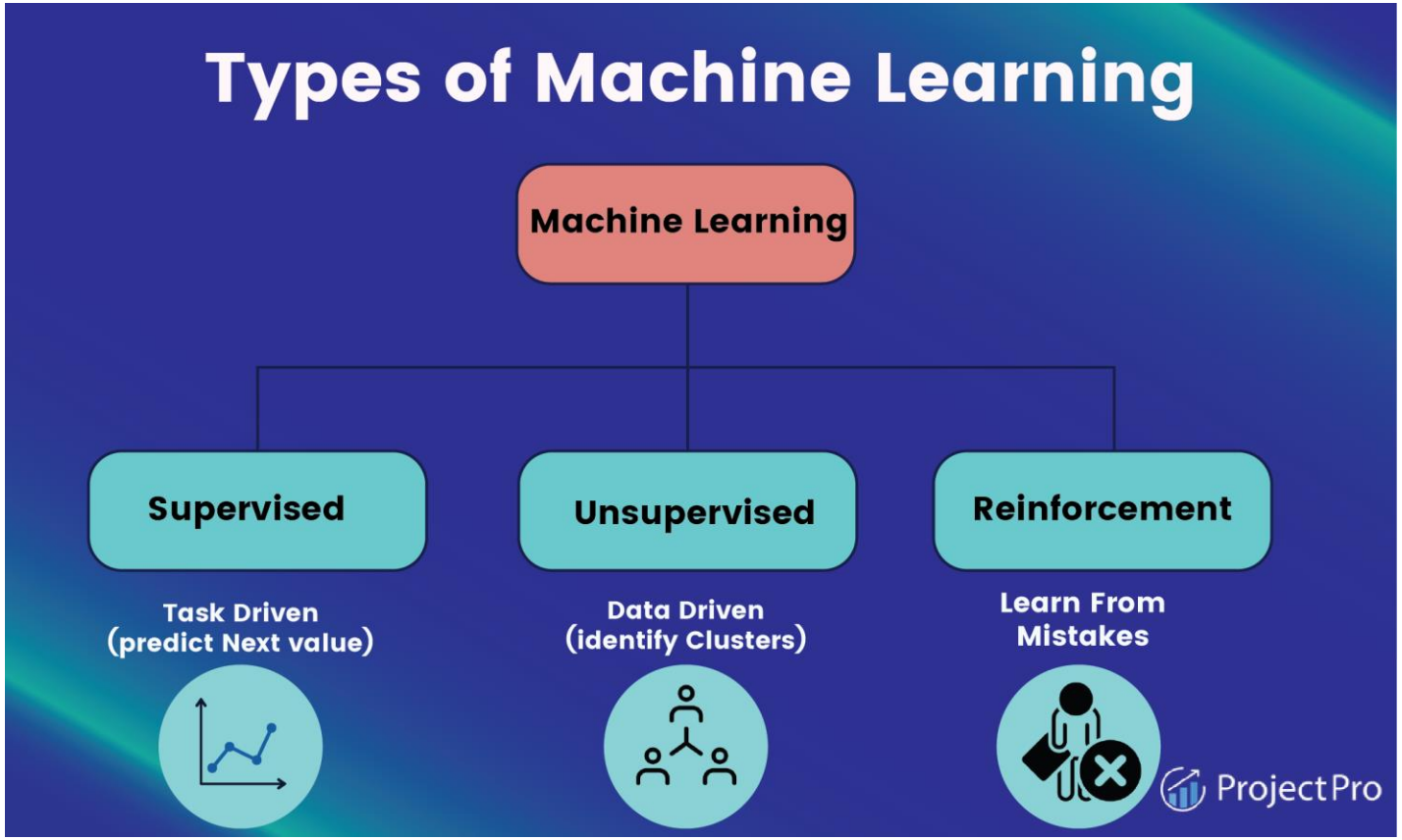
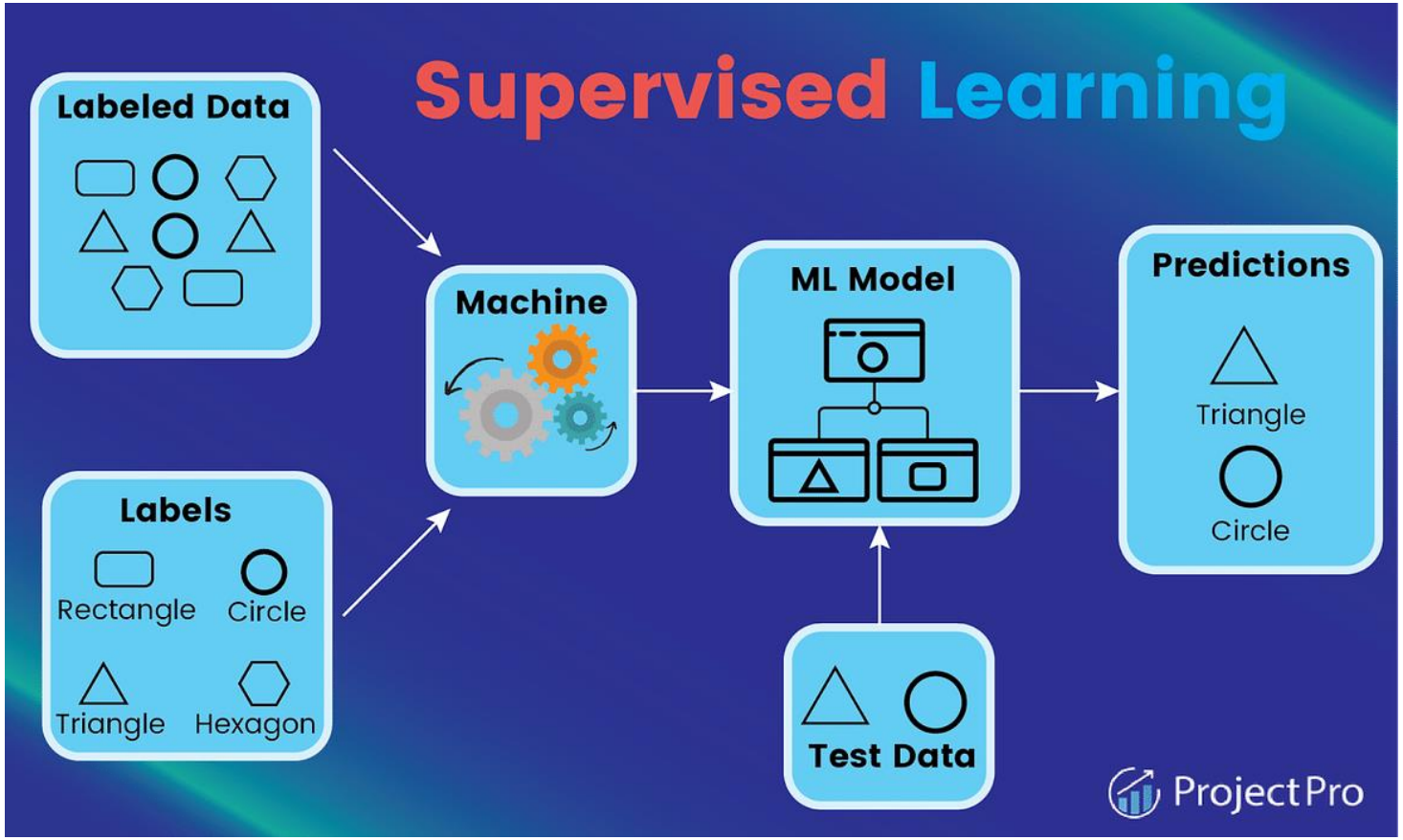
THE BASICS OF MACHINE LEARNING

Eago, Kai Yang
Assistant Professor
Dept of Government and Public Administration
Lingnan University, HK
February 12, 2025

MACHINE LEARNING



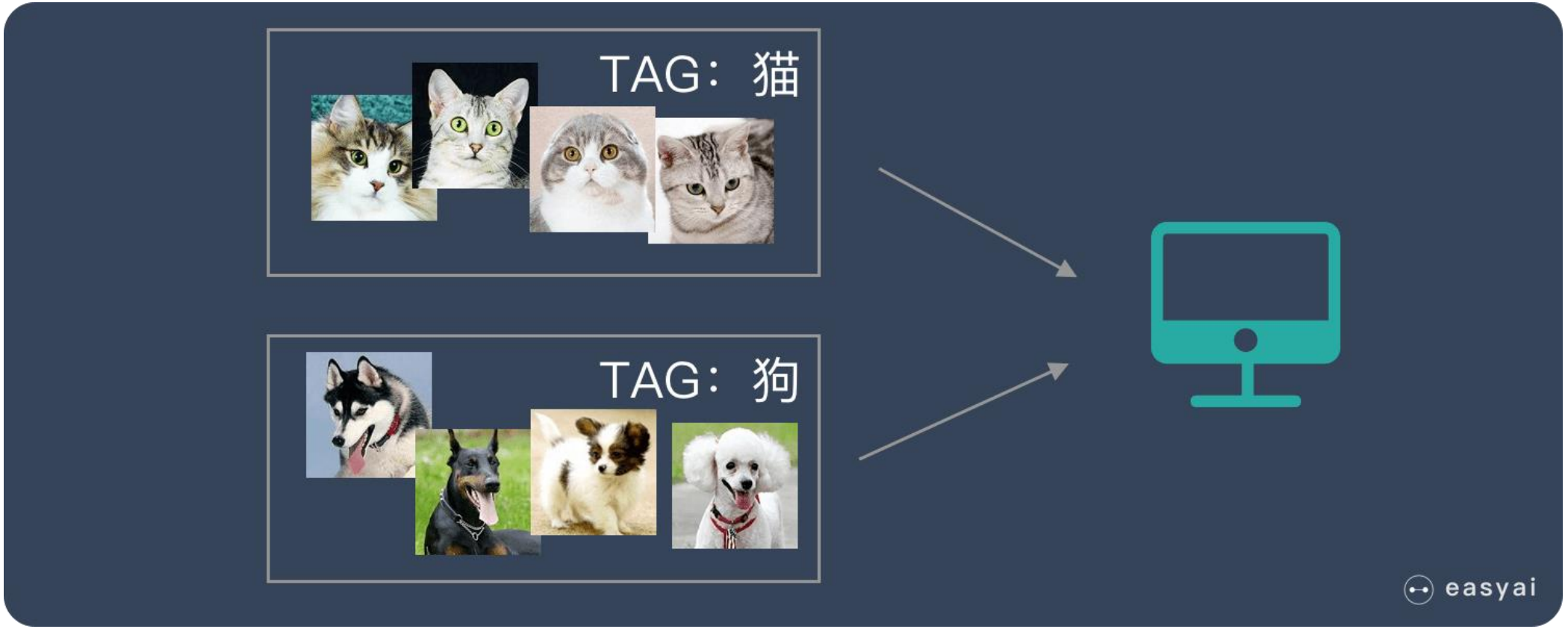
I. SUPERVISED LEARNING 監督學習



1.1 CLASSIFICATION

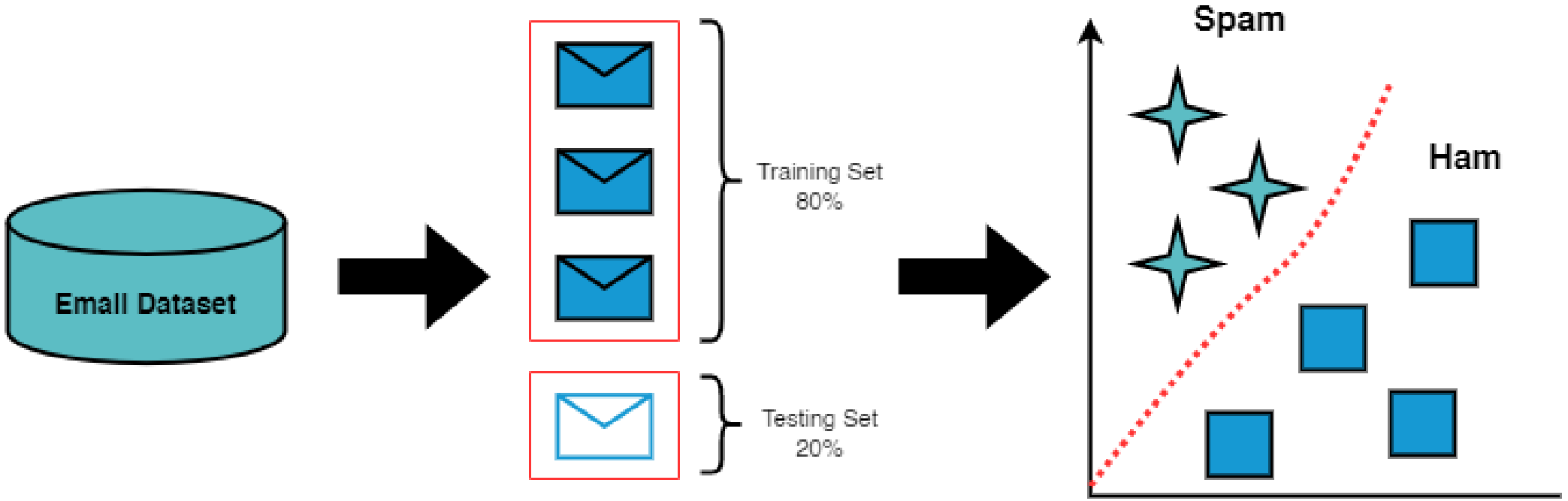
- Predicts the label of a class
- Predict the dataset's categories
- Example: "Yes" or "No“
- Commonly Used Algorithms:
- Decision Tree Algorithm
- Logistic Regression
- Random Forest Algorithm
- Support Vector Machine Algorithm

IMAGE DETECTION



FACIAL RECOGNITION

SPAM DETECTION



Instance Gathering

Training and Testing

Classification

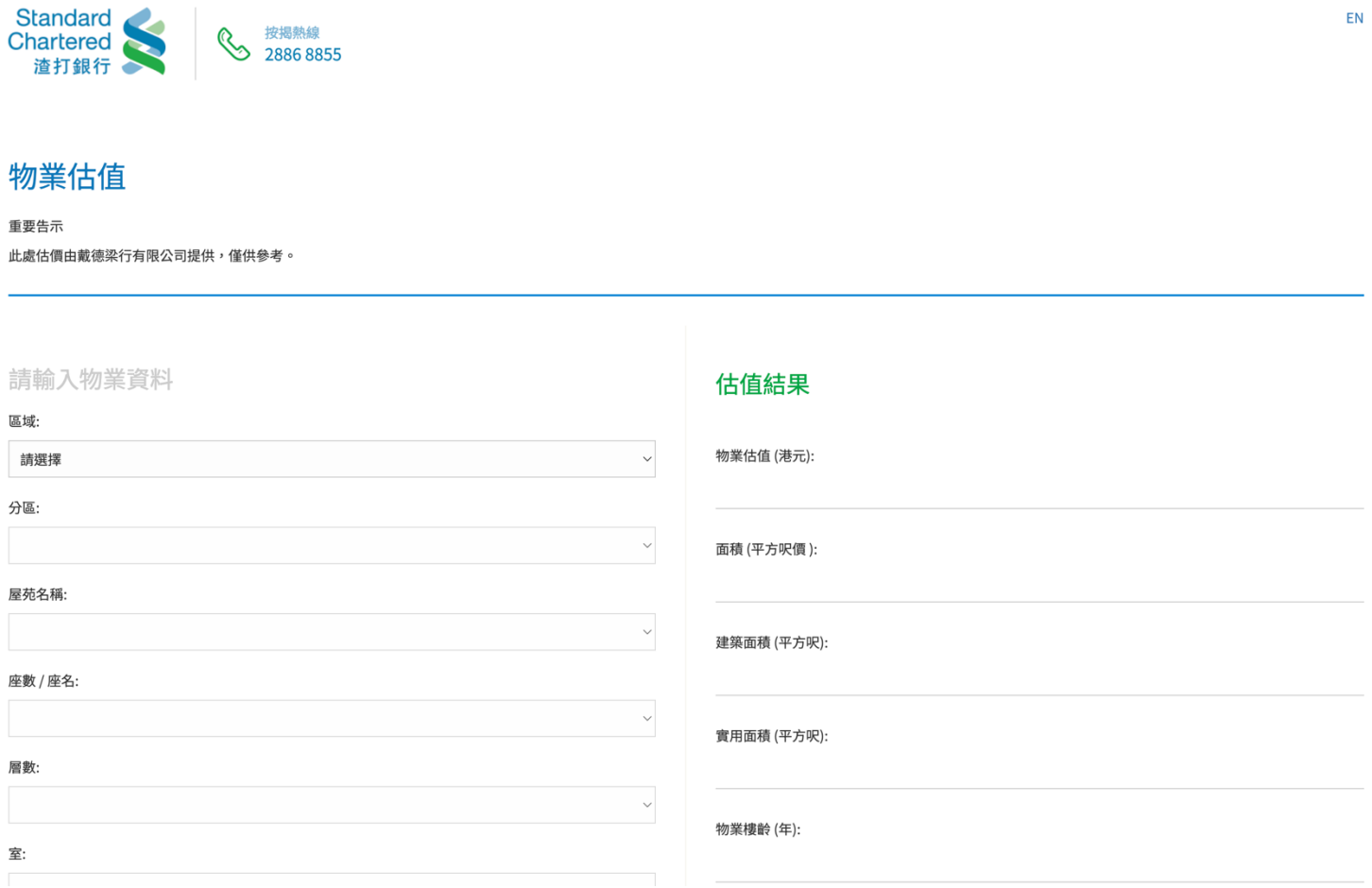
1.2 REGRESSION

- Predicts the numerical label/continuous variables

- Example: weather prediction

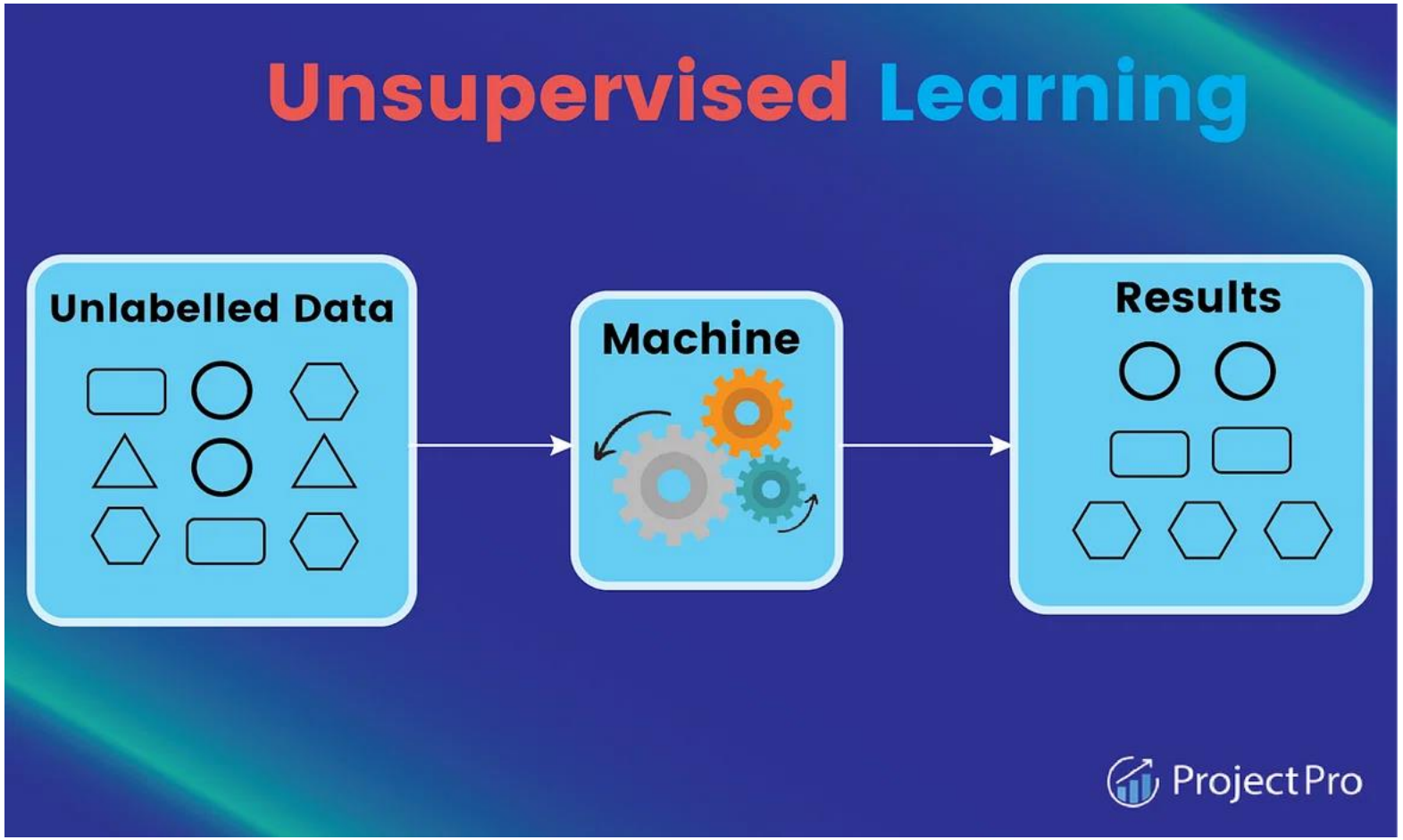
- Commonly used algorithms:
- Decision Tree Algorithm
- Lasso Regression
- Multivariate Regression Algorithm
- Simple Linear Regression Algorithm

HOUSE PRICE PREDICTION



https://www.sc.com/hk/zh/others/property-valuation.html

2. UNSUPERVISED LEARNING非監督學習



2.1 ASSOCIATION

- Finds relations between (independent) variables in a large dataset
- Goal: discover and map data dependent on the other to produce maximum profit

- Supermarket Market Basket Analysis

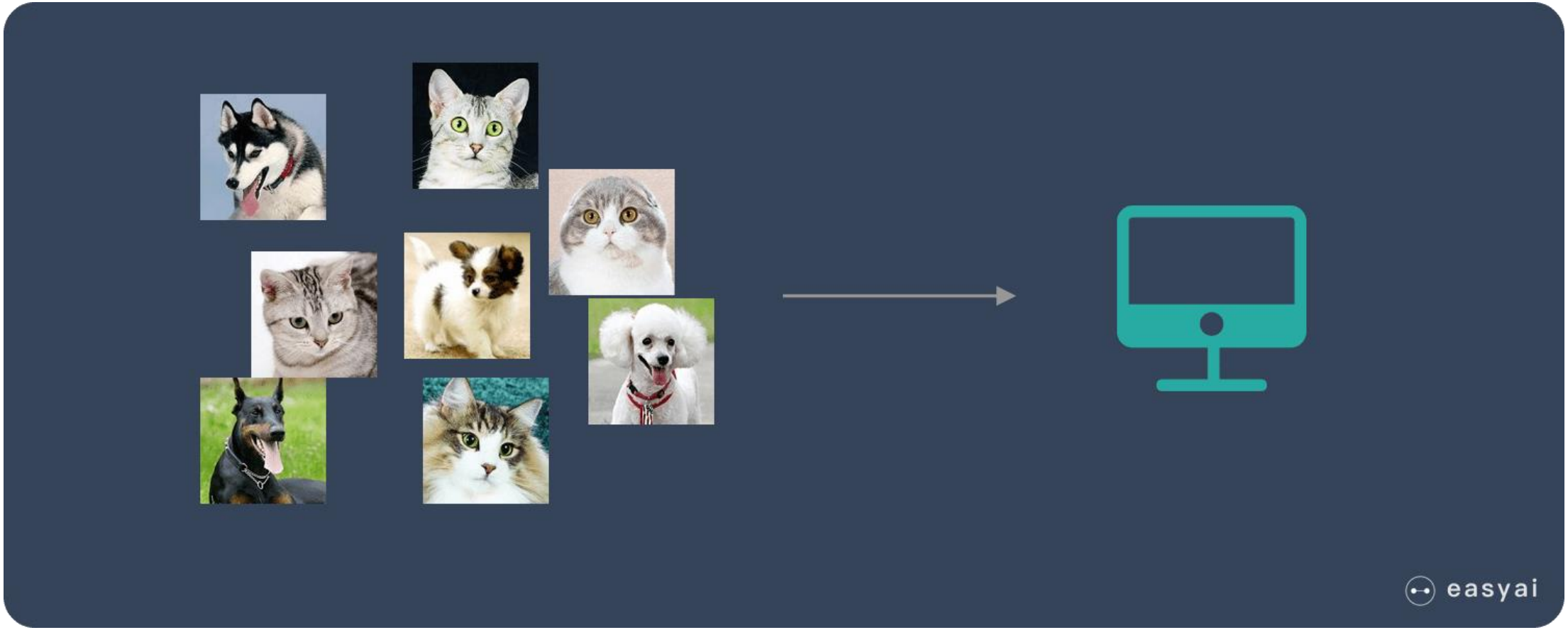
- 30% of transactions contain bread, butter, and milk, and
- 80% of customers who bought bread and butter also bought milk

Transaction	Items Purchased
1	Milk, Egg, Bread, Butter
2	Milk, Butter, Egg, Ketchup, Butter
3	Bread, Butter, Ketchup
4	Milk, Bread, Butter
5	Bread, Butter, Cookies
6	Milk, Bread, Butter, Cookies
7	Milk, Cookies
8	Milk, Bread, Butter
9	Bread, Butter, Egg, Cookies
10	Milk, Butter, Bread
11	Milk, Bread
12	Milk, Bread, Cookies, Ketchup

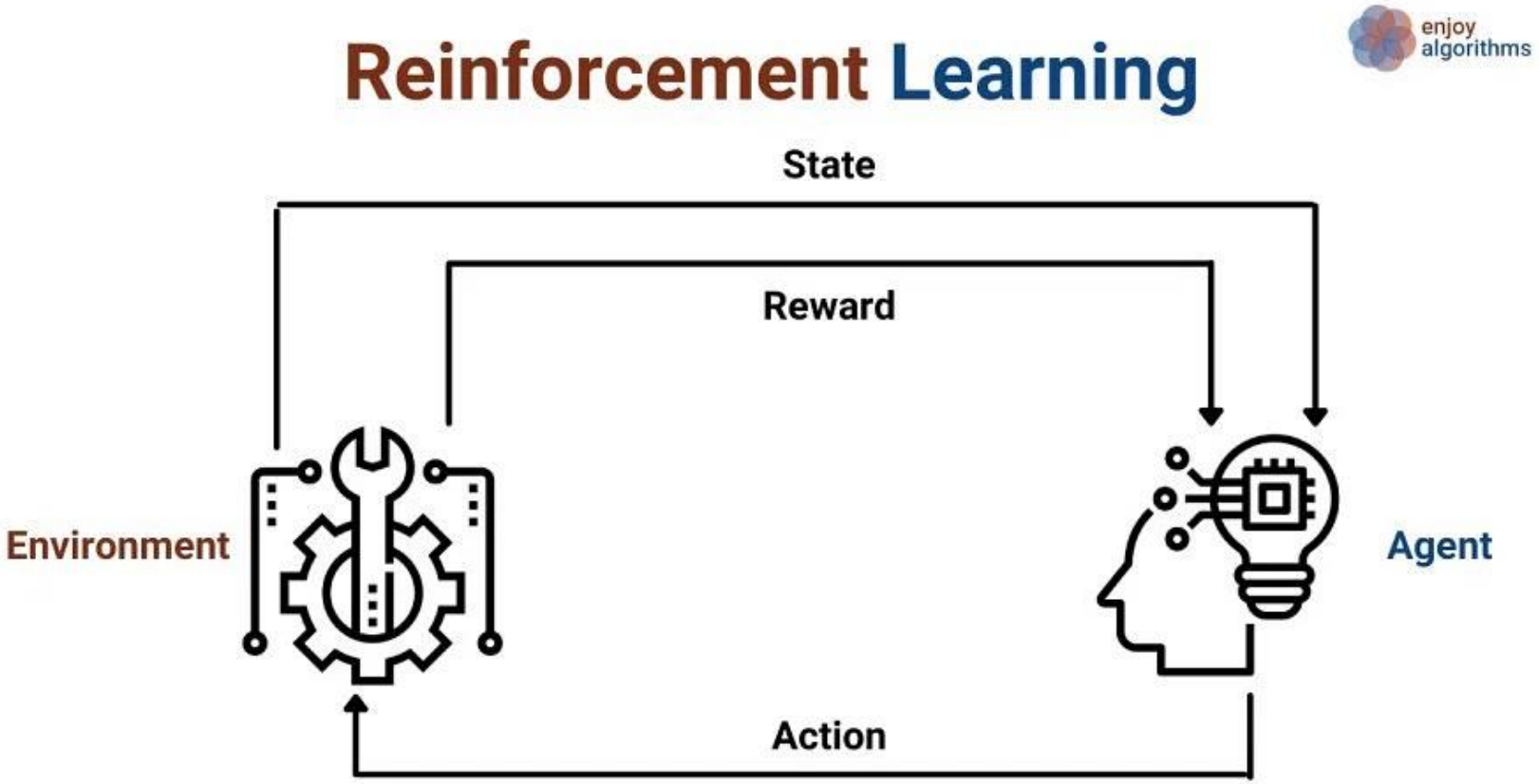
2.2 CLUSTERING

- A method of grouping each set of similar objects into a cluster
- Goal: discover inherent groups from the dataset
- Example: retail marketing

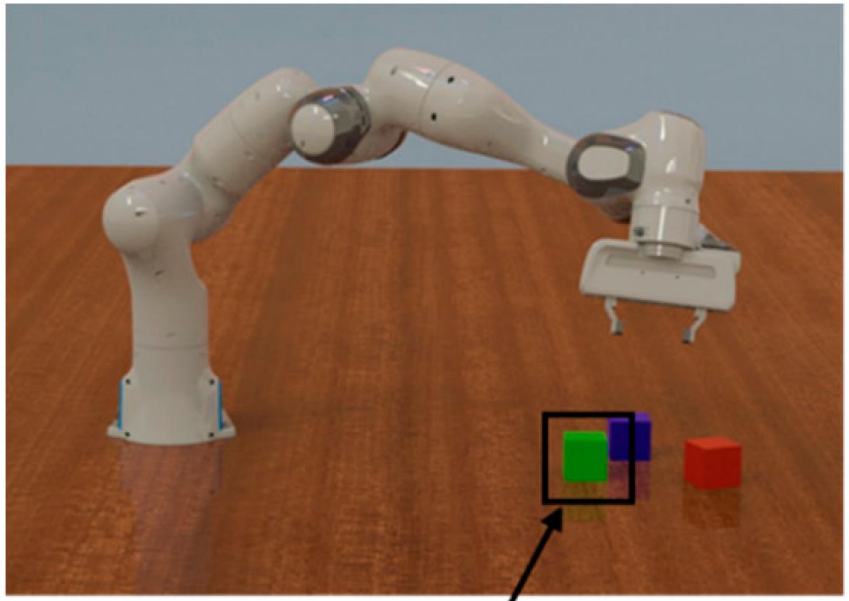
IMAGE CLUSTERING WITHOUT REFERENCE



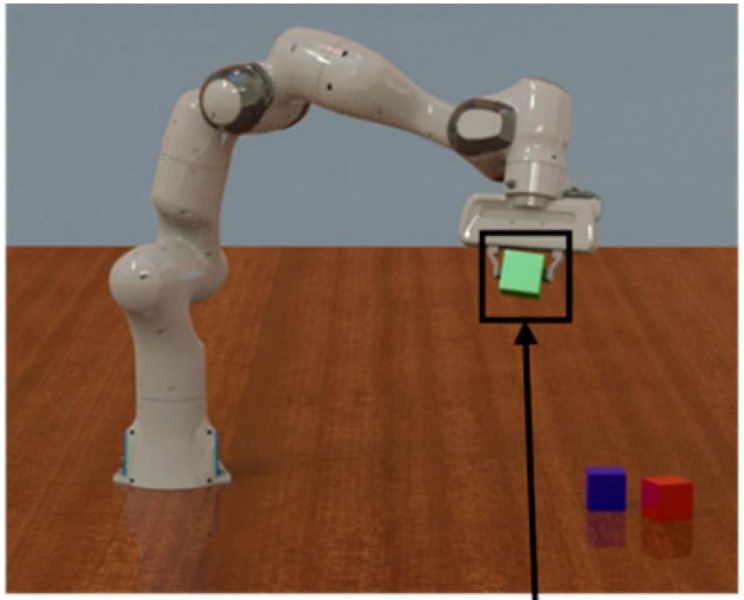
3. REINFORCEMENT LEARNING強化學習



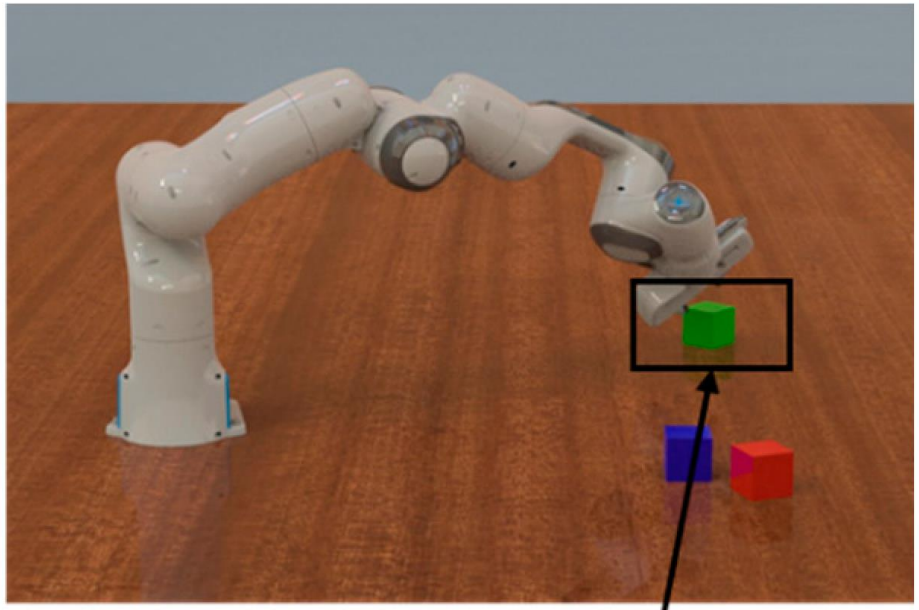
ROBOTICS TRAINING



Target Object

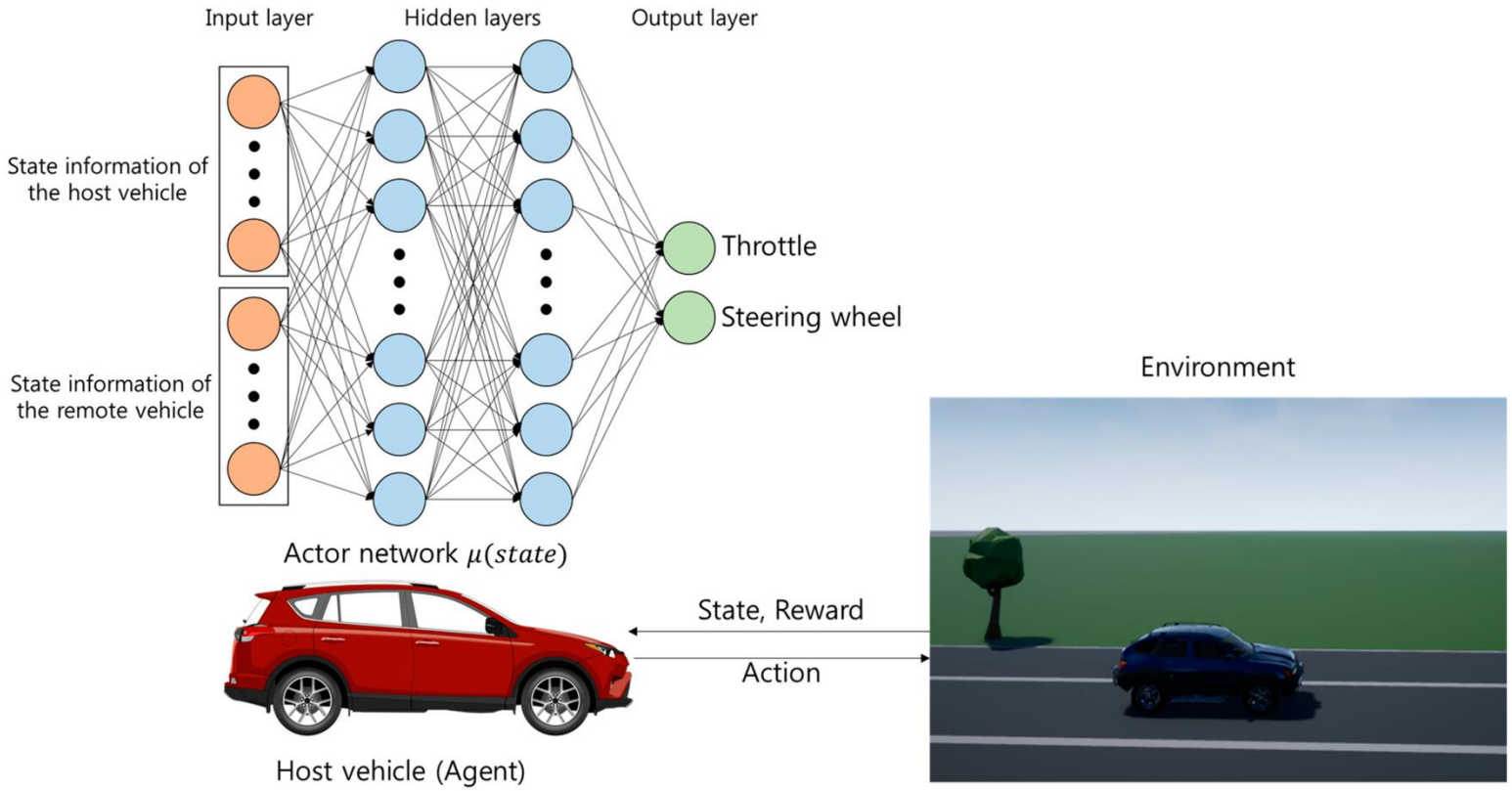


Target Selected



Successful Placement

SELF-DRIVING AUTOMOBILES



CHATGPT (SUPERVISED+REINFORCEMENT LEARNING)

Step 1

Collect demonstration data and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3.5 with supervised learning.

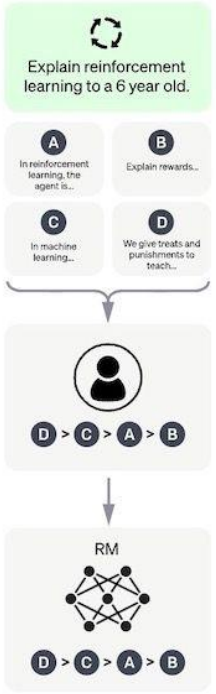
Step 2

Collect comparison data and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

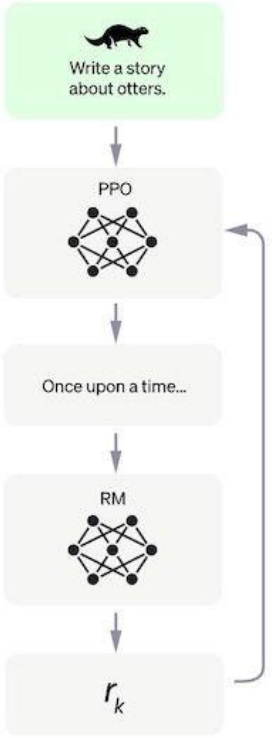
A new prompt is sampled from the dataset.

The PPO model is initialized from the supervised policy.

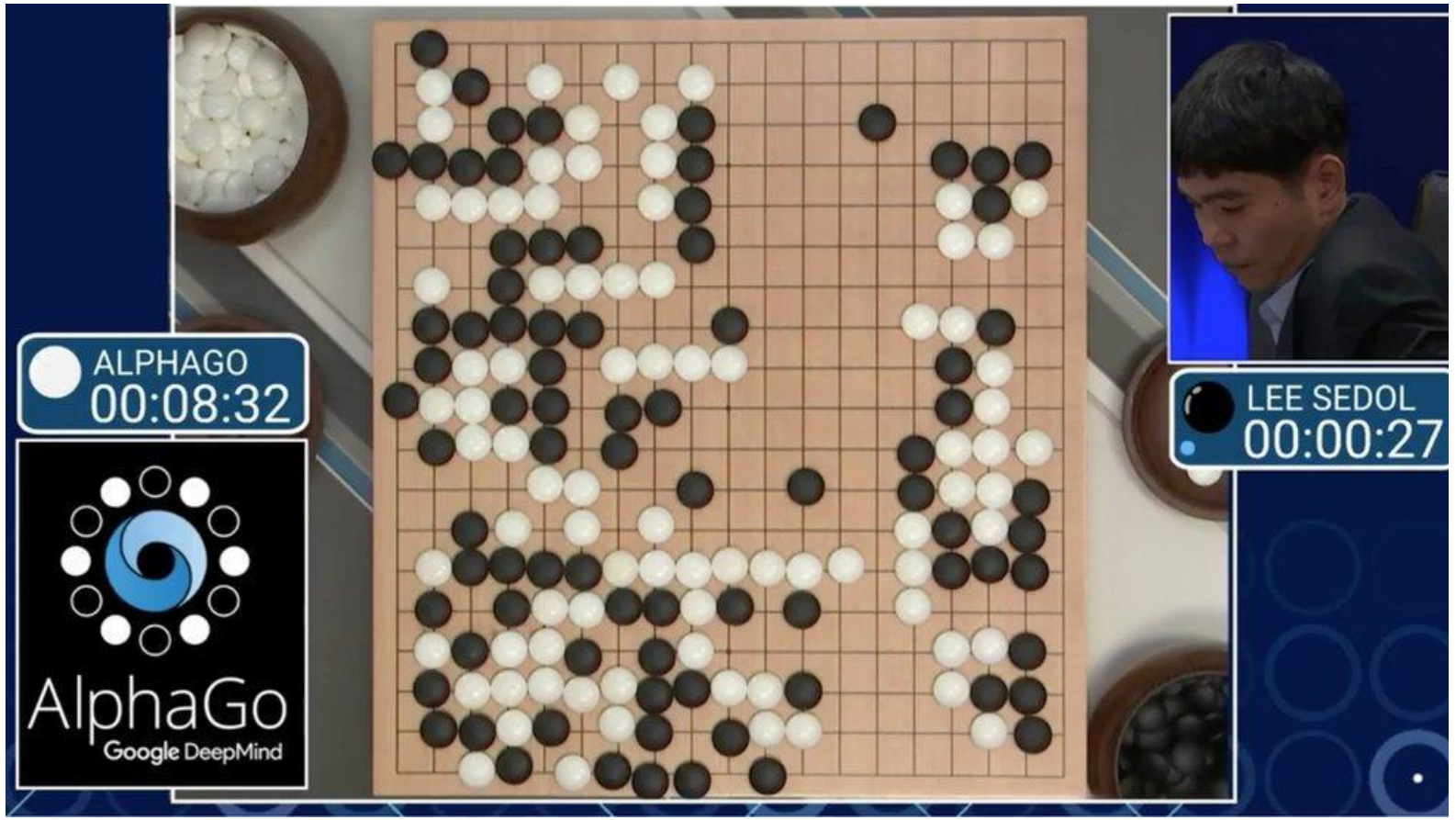
The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



CHESS-PLAYING



- Introduction to the Virtual Issue: Machine Learning in Political Science
- https://www.cambridge.org/core/services/aop-file-manager/file/5c348274e401b41903dae11b/PAN-VSI-Intro-0119-Machine-learning.pdf

- Machine Learning for Social Science: An Agnostic Approach
- https://www.annualreviews.org/doi/pdf/10.1146/annurev-polisci-053119-015921?download=true

REFERENCE LIST

- Brady, Henry E., and David Collier, eds. *Rethinking social inquiry: Diverse tools, shared standards*. Rowman & Littlefield Publishers, 2010.
- Creager A.N.H., Lunbeck, J., Wise, Matthew N. *Science without laws: Model systems, cases, exemplary narratives*. Duke University Press, 2007.
- Goertz, Gary, and James Mahoney. *A tale of two cultures: Qualitative and quantitative research in the social sciences*. Princeton University Press, 2012.
- Mahoney, James. "After KKV: The new methodology of qualitative research." *World Politics* 62, no. 1 (2010): 120-147.
- Caldarelli, Guido, Alessandro Chessa, Fabio Pammolli, Gabriele Pompa, Michelangelo Puliga, Massimo Riccaboni, and Gianni Riotta. "A multi-level geographical study of Italian political elections from Twitter data." *PloS one* 9, no. 5 (2014): e95809.
- Crosas, Merce, Gary King, James Honaker, and Latanya Sweeney. "Automating open science for big data." *The ANNALS of the American Academy of Political and Social Science* 659, no. 1 (2015): 260-273.
- González-Bailón, Sandra, and Georgios Paltoglou. "Signals of public opinion in online communication: A comparison of methods and data sources." *The ANNALS of the American Academy of Political and Social Science* 659, no. 1 (2015): 95-107.
- Salganik, Matthew J. *Bit by bit: Social research in the digital age*. Princeton University Press, 2018.
- King, Gary, Robert O. Keohane, and Sidney Verba. *Designing social inquiry: Scientific inference in qualitative research*. Princeton university press, 1994.