

COMP 5212

Machine Learning

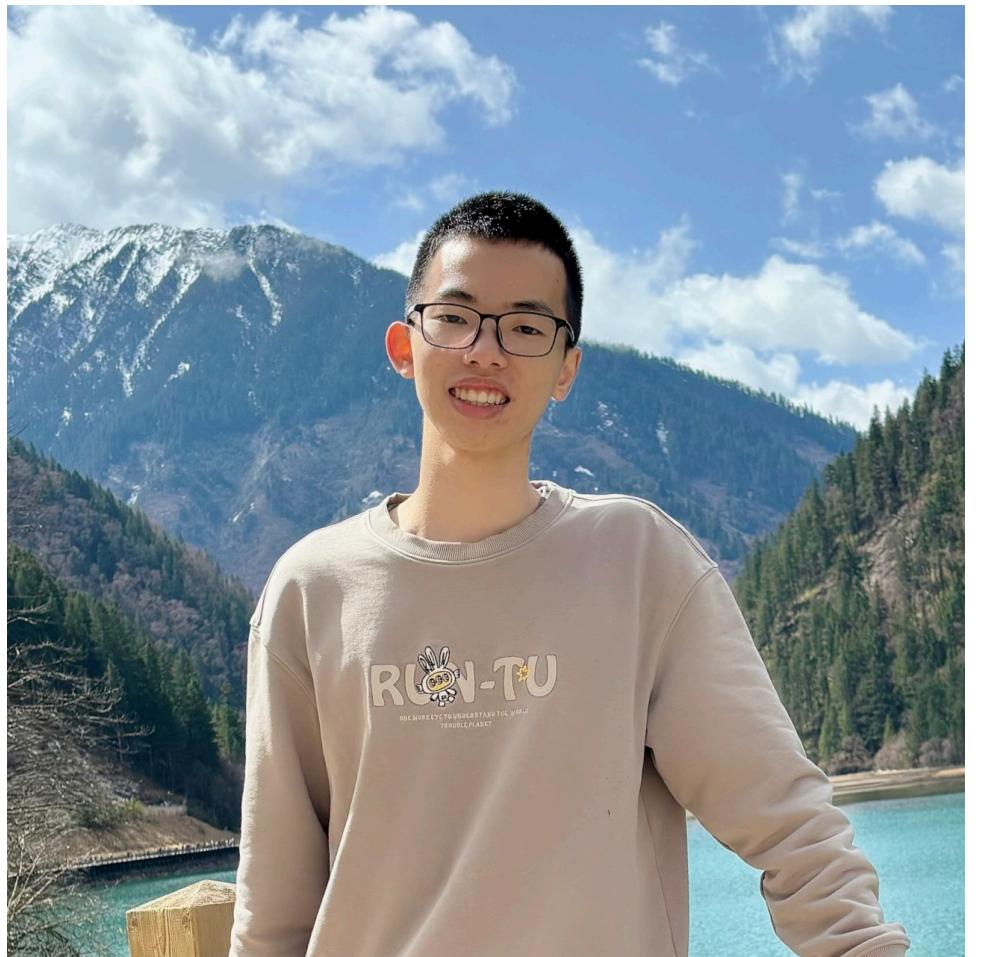
Instructor: Junxian He

Course website: <https://jxhe.github.io/teaching/comp5212f24>

Teaching Team & Office Hours

Instructor: Junxian He

Office Hour: Wed 430pm - 530pm, Room 3512



TA1: Yuzhen Huang

Office Hour: Wed 9am-10am



TA2: Jinghan Zhang

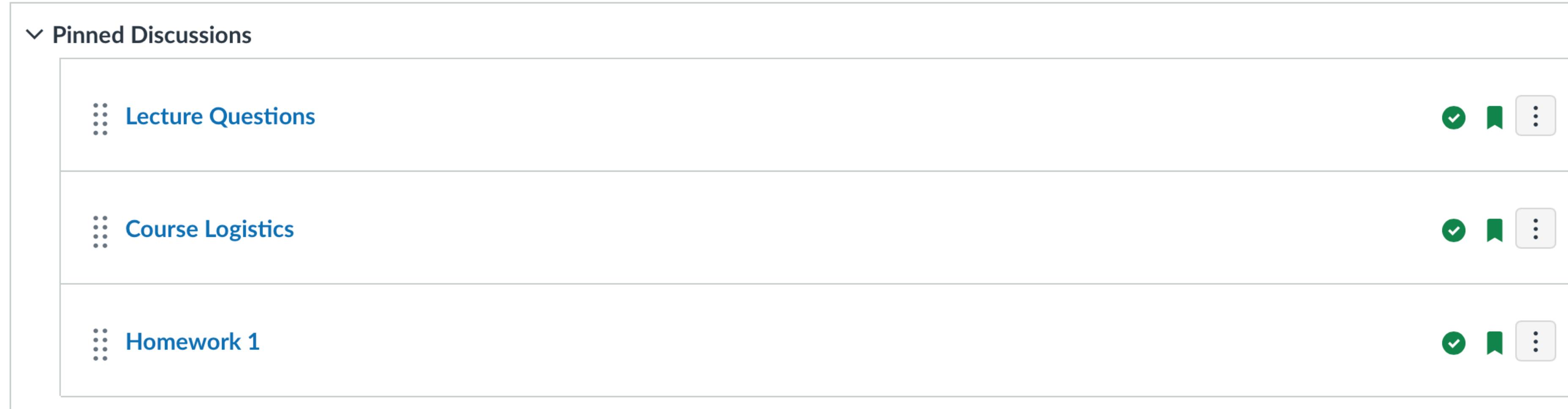
Office Hour: Fri 6pm-7pm

common place of SENG (near Lift 27/28, Room 2580)

This is an in-person Class

- Lectures are recorded, but the videos will be released only twice — during the midterm and before the final
- I may release the respective lecture videos in special cases, e.g., the lectures around Lunar New Year Eve
- Lecture slides will be available before each lecture

Communication and Discussion



A screenshot of the 'Pinned Discussions' section in a Canvas course navigation menu. The section is titled 'Pinned Discussions' with a dropdown arrow. It contains three pinned items, each represented by a blue square icon with three vertical dots, followed by the discussion title in blue text, and a set of three icons in a grey box: a green checkmark, a green bookmark, and a three-dot menu.

Discussion Title	Icons
Lecture Questions	Green checkmark, Green bookmark, Three-dot menu
Course Logistics	Green checkmark, Green bookmark, Three-dot menu
Homework 1	Green checkmark, Green bookmark, Three-dot menu

Please direct all technical questions on Canvas, and do not directly contact the instructor or TAs for technical questions. We will answer your questions on Canvas promptly

Canvas allows anonymous communication

Pre-requisite

- Probability
 - Distribution, random variable, expectation, conditional distribution, variance
- Linear algebra
 - Matrix multiplication
- Python programming

A self-check exam: https://www.cs.cmu.edu/~aarti/Class/10701_Spring23/Intro_ML_Self_Evaluation_new.pdf

Grading

- Attendance (10%)
- 5 assignments (50%)
 - 4 Written + (optional) programming assignments, will be mostly written
 - 1 programming-only assignment (Kaggle alike competition)
 - 3 free late days in total, for additional late days, 20% penalization applied for each day late
 - No assignment will be accepted more than 3 days late
- Mid-term Exam, open-note (20%)
- Final exam, open-note (20%)

Attendance

- Occasional quiz questions
- 80% of attendance will give you full grade
- Correctness of quiz answers does not influence attendance grading

Assignments

- 5 assignments (50%)
 - 4 Written + programming assignments ($4 * 9\%$)
 - 1 programming-only assignment (14%, Kaggle alike competition)
 - 3 free late days in total, for additional late days, 20% penalization applied for each day late
 - No assignment will be accepted more than 3 days late

Honor Code

Do's

- Form study groups to discuss (e.g. homework)
- Write down the homework solutions independently
- Write down the names of people with whom you've discussed the homework
- You are encouraged to use generative AI (e.g. ChatGPT) to **assist** you

Don'ts

- Copy, refer to, or look at solutions from previous years, online, or others
- Copy ChatGPT's answers
- Longer versions on the course website

We have zero tolerance — in the case of honor code violation for a single time, you will fail this course directly

Waiting List & Audit

- We will let everyone on the waiting list in as long as the total enrollment does not exceed 90 (currently 60 enroll + 130 WL)

Many students will drop after the first two weeks, or after we release the first homework

- Audit is allowed, and will be graded in the same criterion

More Info on Course Website

- Canvas is the main platform for announcement, discussion, homework submission
- Recorded videos on canvas
- Syllabus, slides and relevant reading materials
- Detailed course logistics

Topics to be Covered

Introduction
Math basics
Linear Regression
Logistic regression, Exponential Family
Generalized linear models, Kernel Methods
Kernel methods, SVM
SVM
SVM
The National Day Holiday
Generative Models
Naive Bayes, MLE, MAP
Generalization, bias-variance tradeoff
Clustering, EM
Expectation Maximization
PCA
Mid-term exam

Probabilistic Graphical Models
HMM
HMM
Neural Networks, backpropagation
Neural architectures
Transformer, Variational autoencoder
Variational autoencoder
GANs, Reinforcement Learning
Reinforcement Learning
Language models, pretraining
Large language models

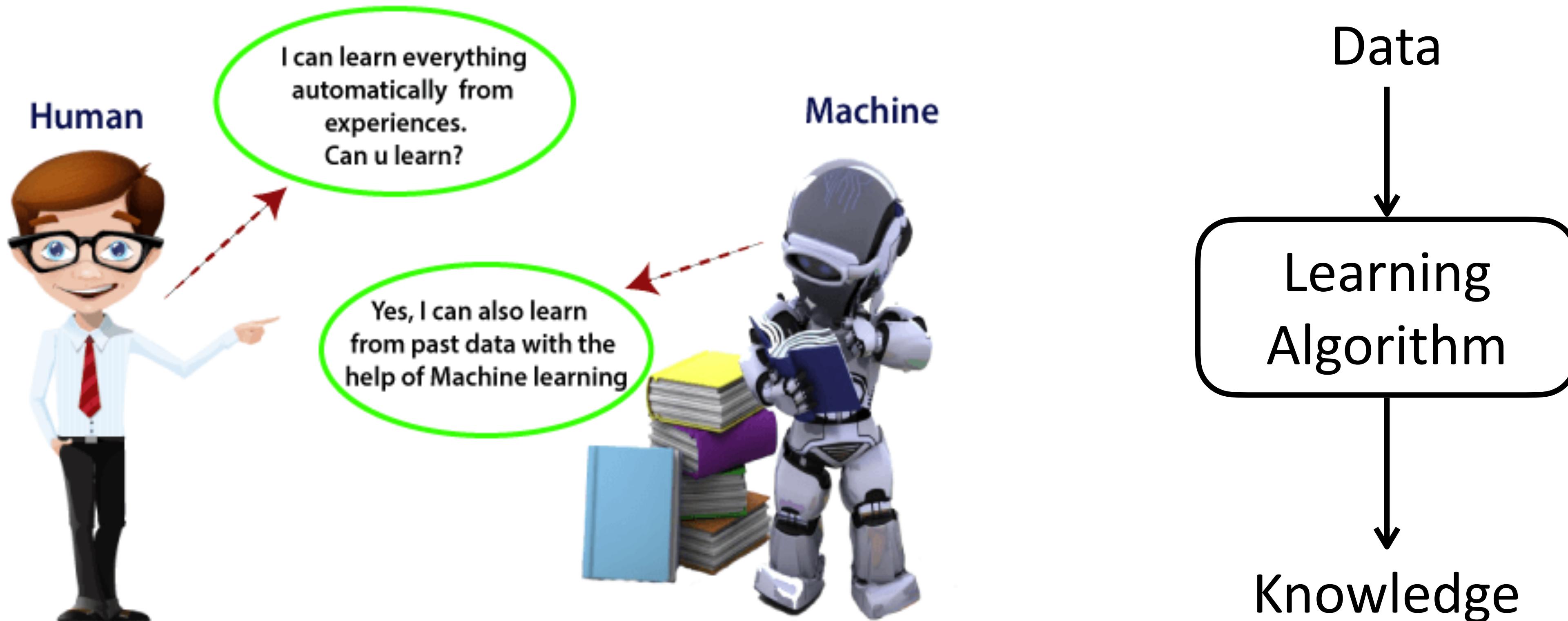
This is NOT an easy course

- Overall difficulty comparable to the graduate machine learning course at other top universities like CMU, workload may be slightly lighter
- Homework is the most important and may be difficult sometimes, exams are significantly easier
- It is wrong to regard graduate courses as nothing useful / unhelpful for your research. Most great PhD students in the US take courses and homework seriously

We'll see more examples in this course on how basic machine learning knowledge motivates great researches nowadays, even though for deep learning

Overview of Machine Learning

What is Machine Learning

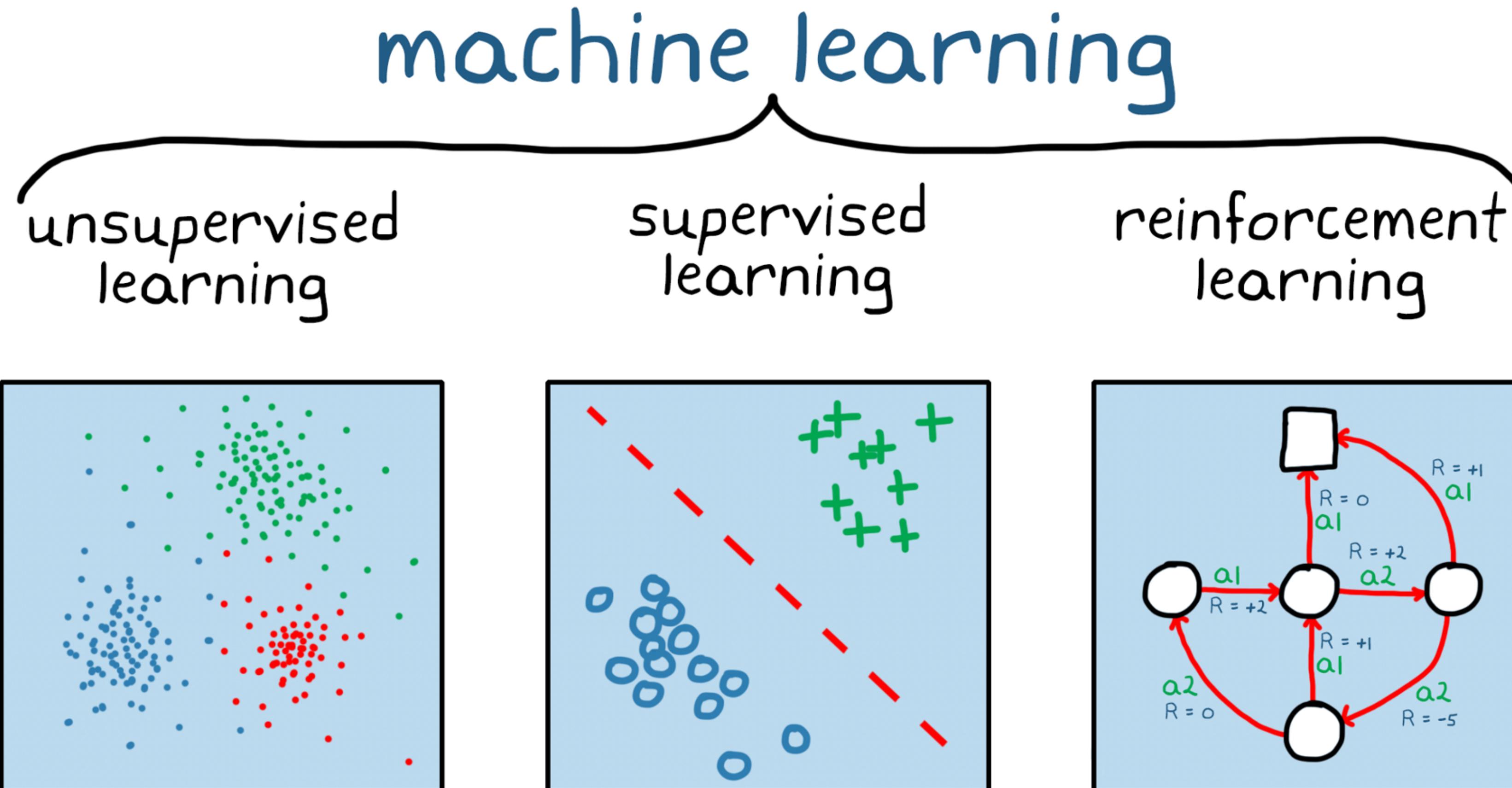


Machine Learning is Trending



- Everywhere — wide application
 - Finance, data scientist, medical diagnosis, translation, self-driving....
- Foundation of artificial intelligence — one of the most important technology for the society in the next 10s of years
 - ChatGPT, large language model, large multimodal model

Taxonomy of Machine Learning



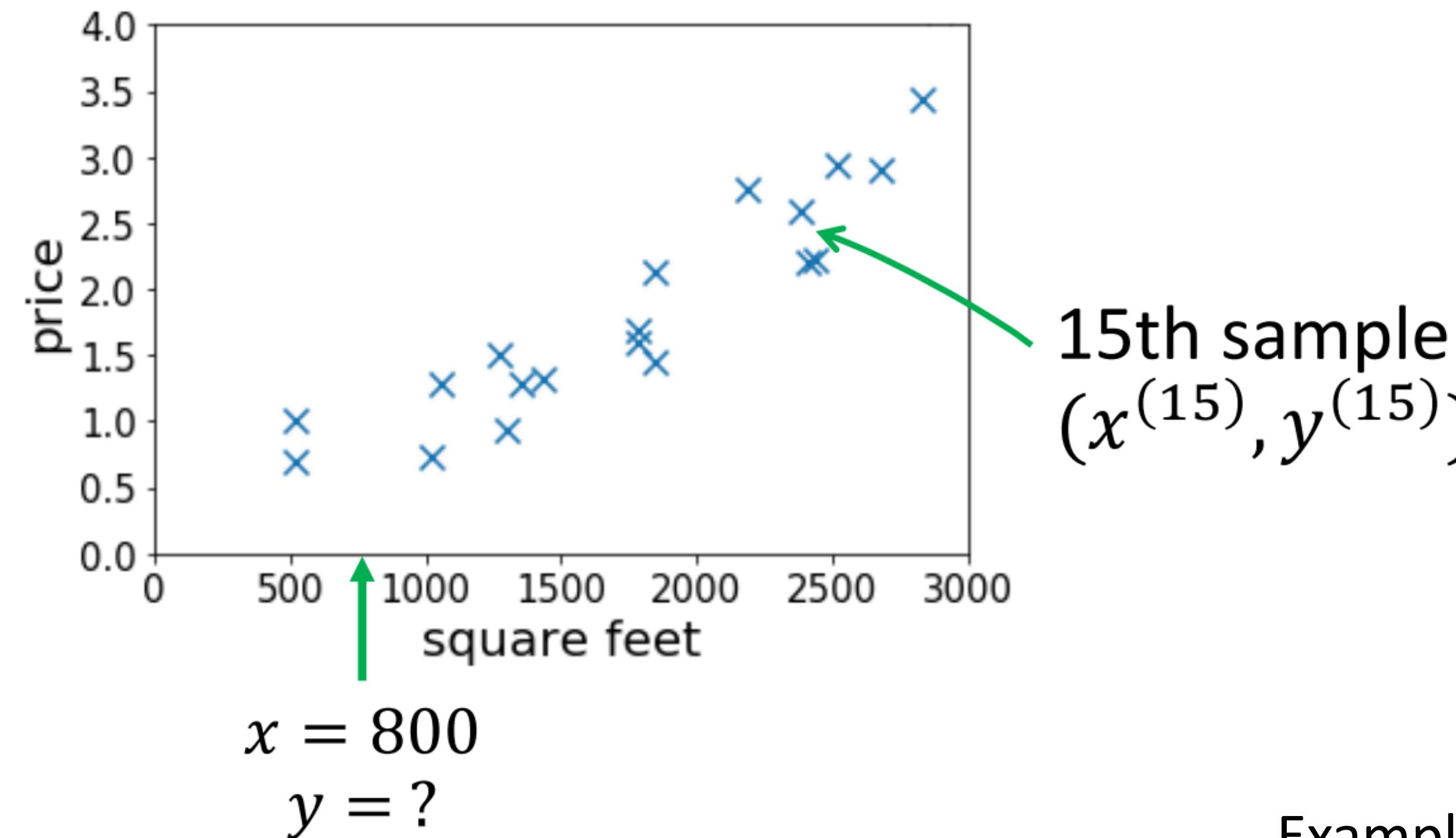
Supervised Learning

Housing Price Prediction

- Given: a dataset that contains n samples

$$(x^{(1)}, y^{(1)}), \dots (x^{(n)}, y^{(n)})$$

- Task: if a residence has x square feet, predict its price?

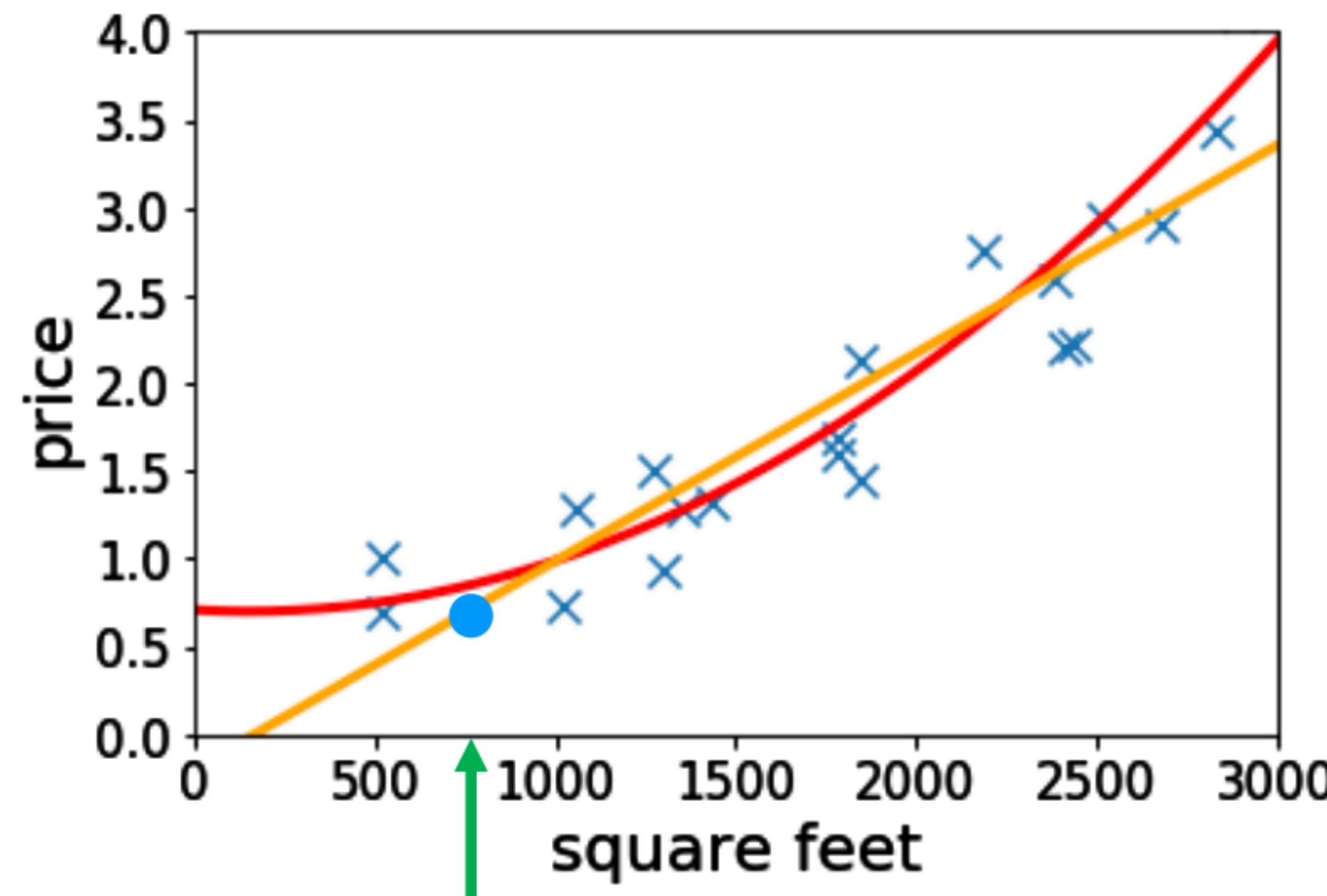


Housing Price Prediction

- Given: a dataset that contains n samples

$$(x^{(1)}, y^{(1)}), \dots (x^{(n)}, y^{(n)})$$

- Task: if a residence has x square feet, predict its price?



More Features

- Suppose we also know the lot size

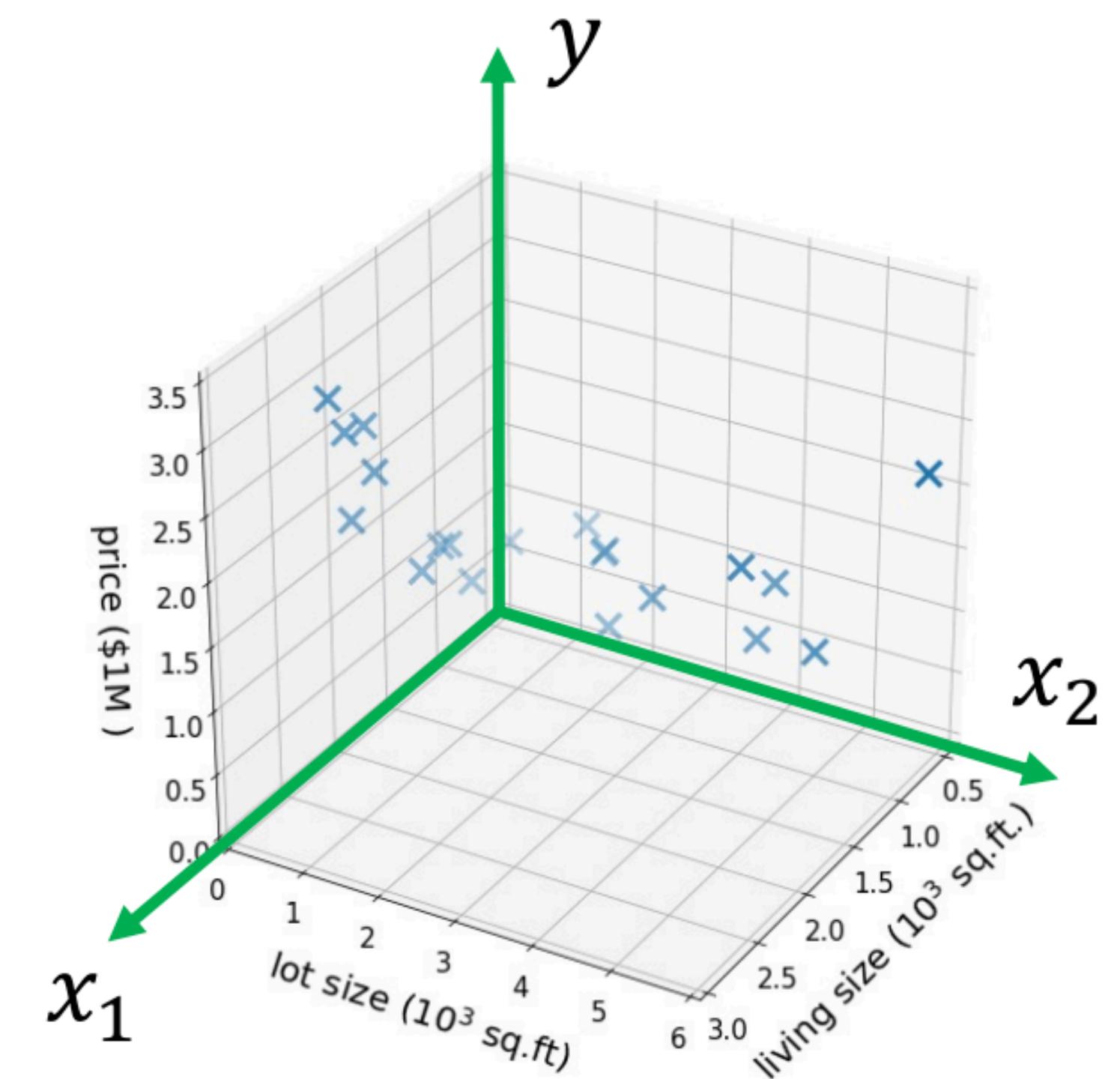
More Features

- Suppose we also know the lot size
- Task: find a function that maps

$(\text{size}, \text{lot size}) \rightarrow \text{price}$

features/input
 $x \in \mathbb{R}^2$

label/output
 $y \in \mathbb{R}$

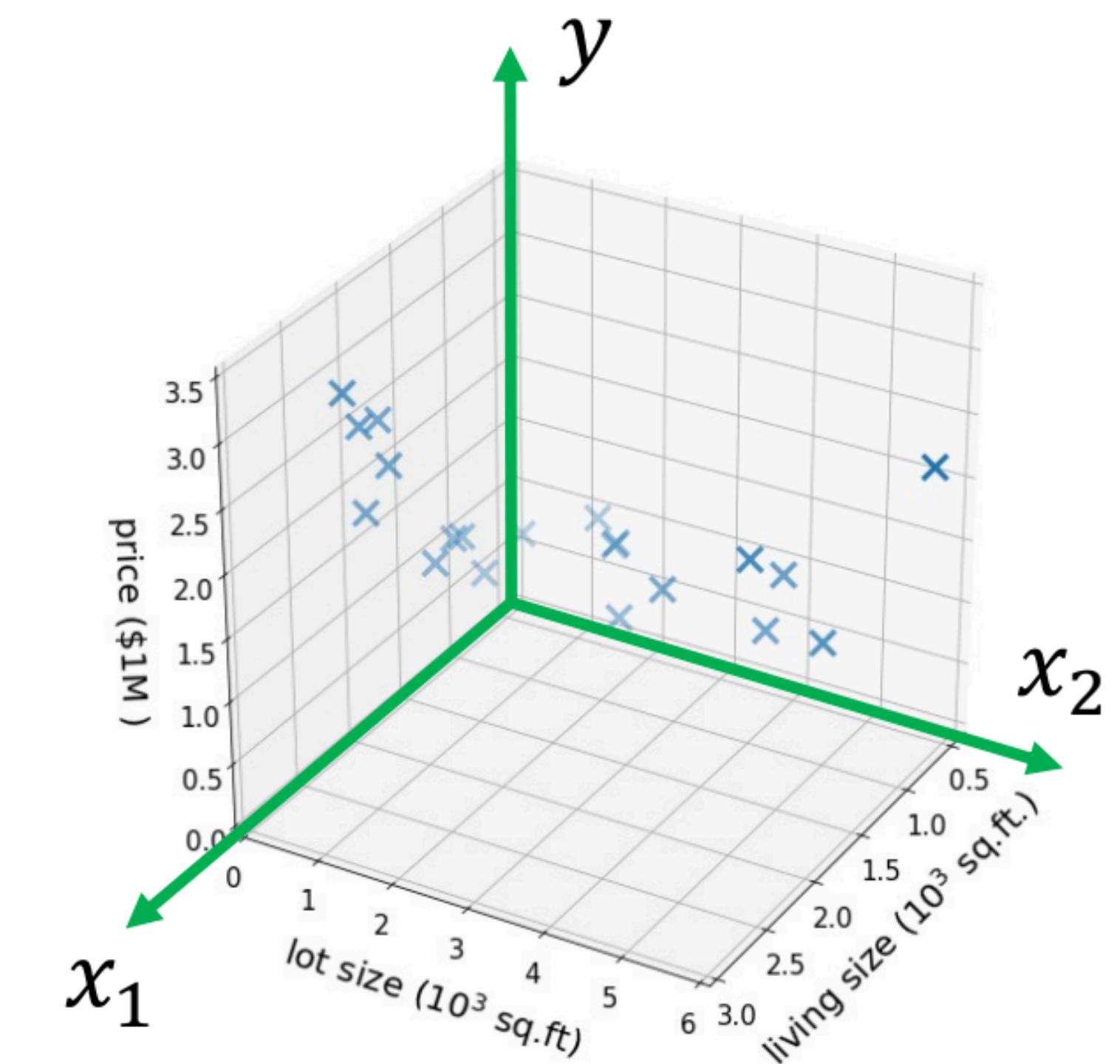


More Features

- Suppose we also know the lot size
 - Task: find a function that maps

The diagram illustrates a machine learning model's architecture. On the left, the input is represented as a pair of values: '(size, lot size)'. A green brace underlines these two values, indicating they are the 'features/input'. An arrow points from this input pair to the right. To the right of the arrow, the output is labeled 'price', which is also underlined by a green brace, indicating it is the 'label/output'. Below the input, the mathematical expression $x \in \mathbb{R}^2$ is shown, specifying the input space. Below the output, the mathematical expression $y \in \mathbb{R}$ is shown, specifying the output space.

- Dataset: $(x^{(1)}, y^{(1)}), \dots, (x^{(n)}, y^{(n)})$

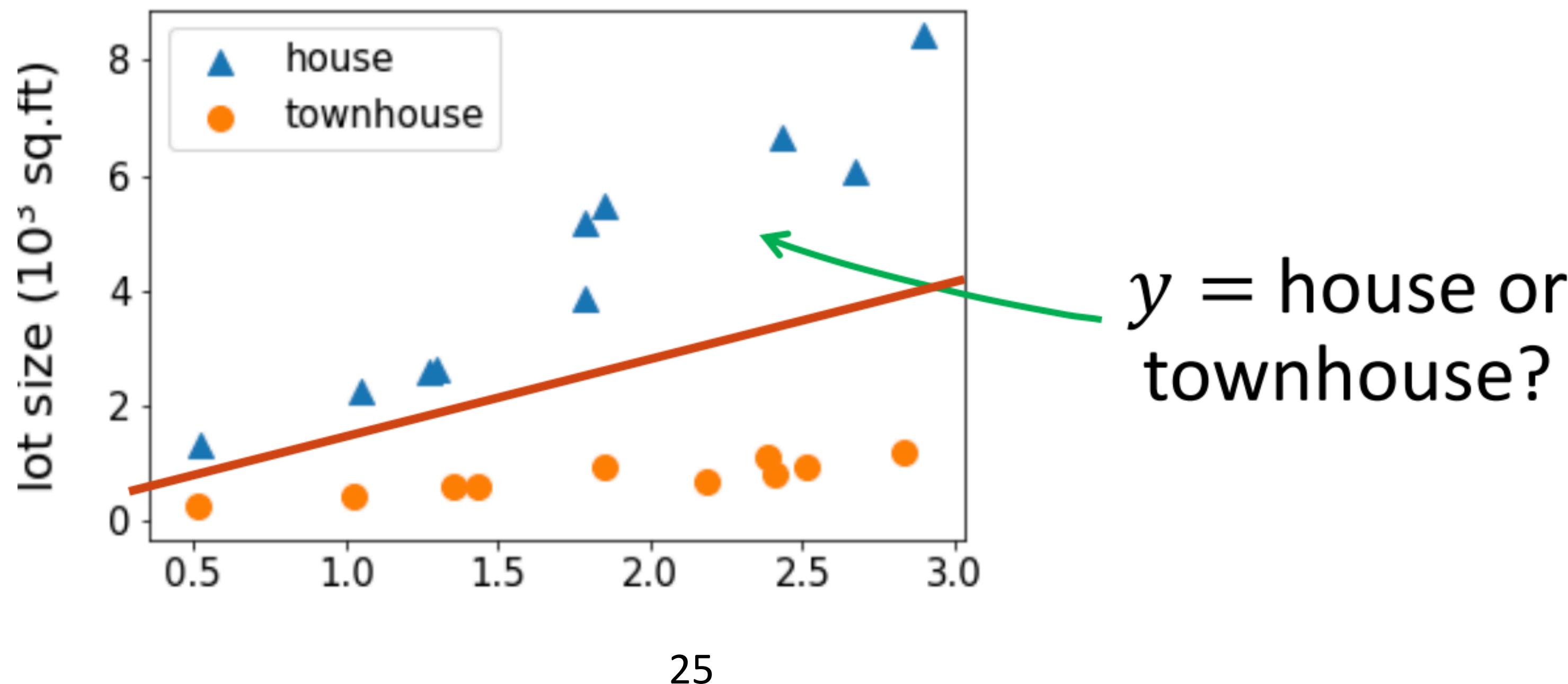


High-dimensional Features

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ \vdots \\ x_d \end{bmatrix} \begin{array}{l} \text{--- living size} \\ \text{--- lot size} \\ \text{--- # floors} \\ \text{--- condition} \\ \text{--- zip code} \\ \vdots \end{array} \longrightarrow y \text{ --- price}$$

Regression vs Classification

- Regression: if $y \in \mathbb{R}$ is a continuous variable
 - E.g., price prediction
- Classification: the label is a discrete variable
 - E.g., predicting the types of residence



Supervised Learning in Computer Vision

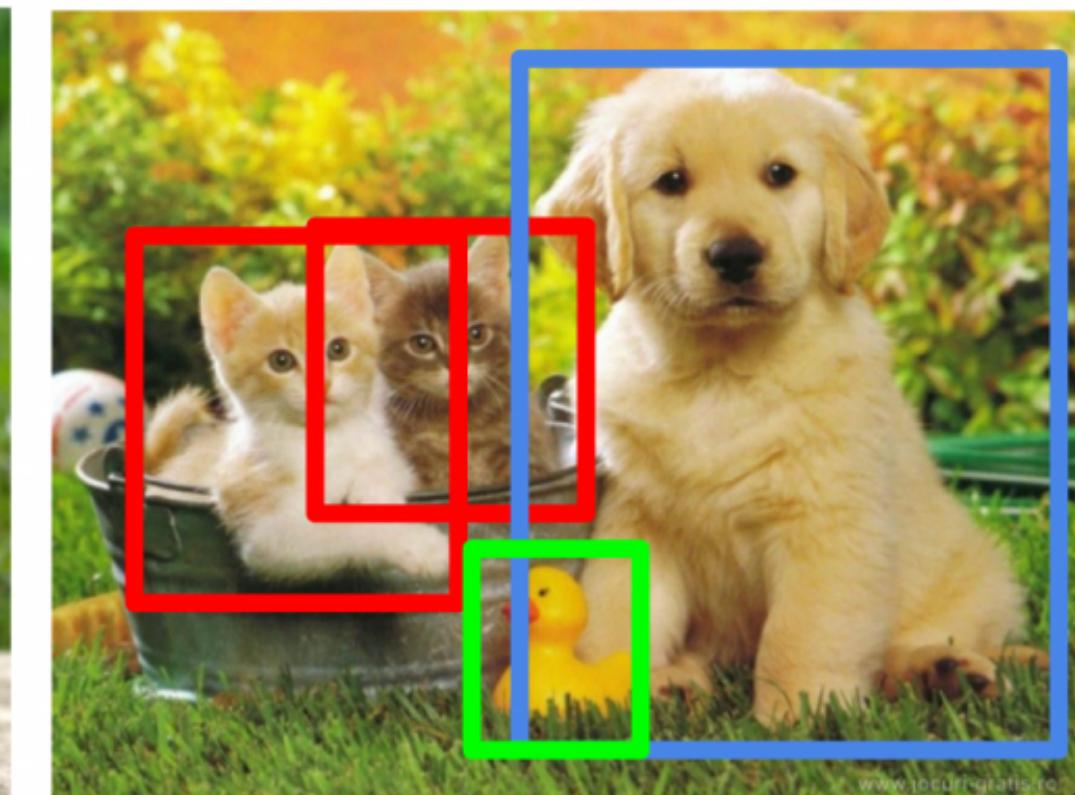
Classification



Classification + Localization



Object Detection



Instance Segmentation



CAT

CAT

CAT, DOG, DUCK

CAT, DOG, DUCK

Single object

Multiple objects

Supervised Learning in Natural Language Processing



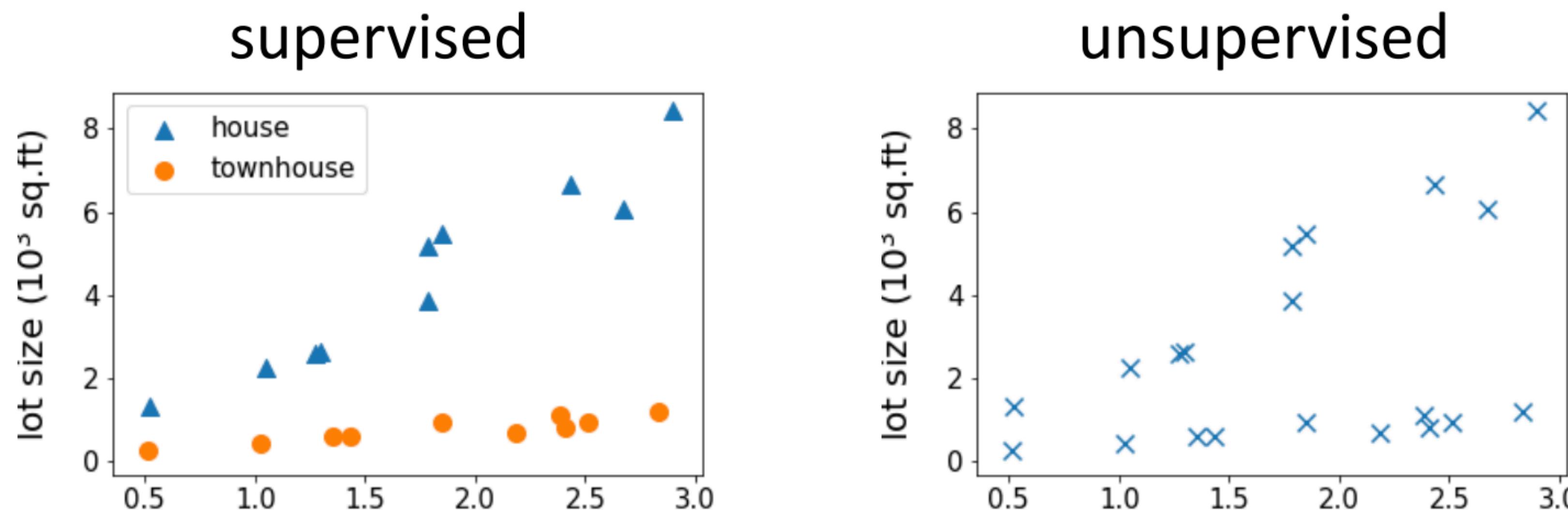
Hey Siri

- This course will only cover basic and fundamental things about ML

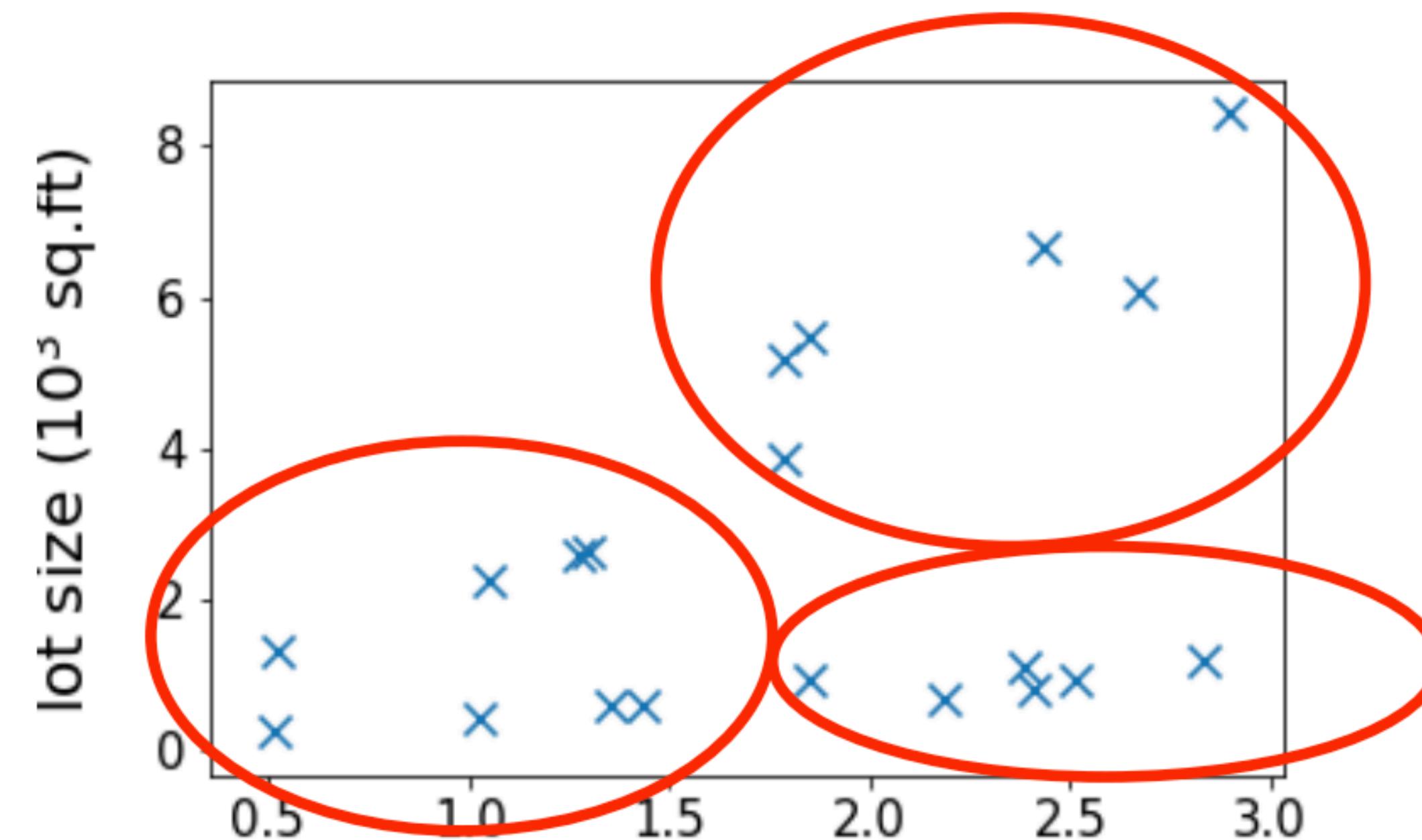
Unsupervised Learning

Unsupervised Learning

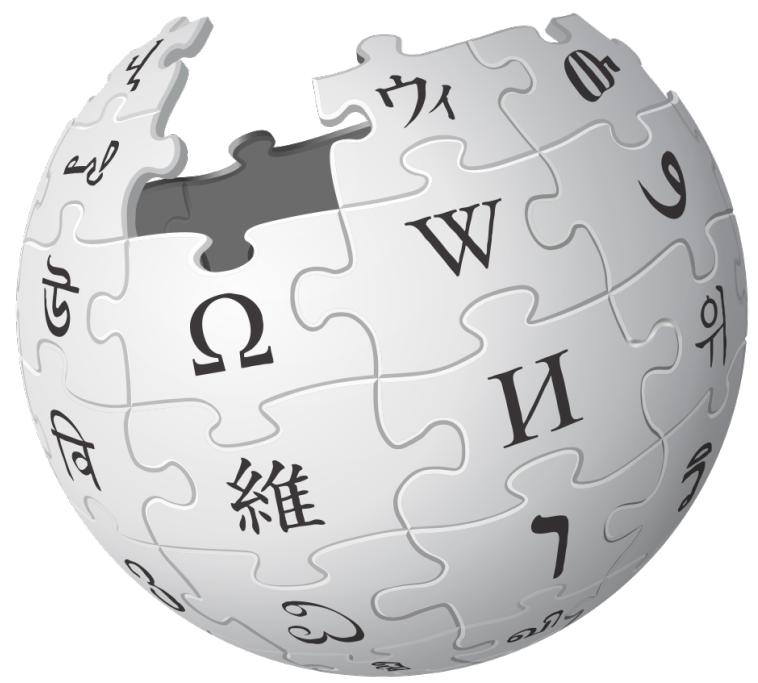
- Dataset contains no labels $x^{(1)}, \dots x^{(n)}$
- Typically very vague goal: to find interesting structures in the data



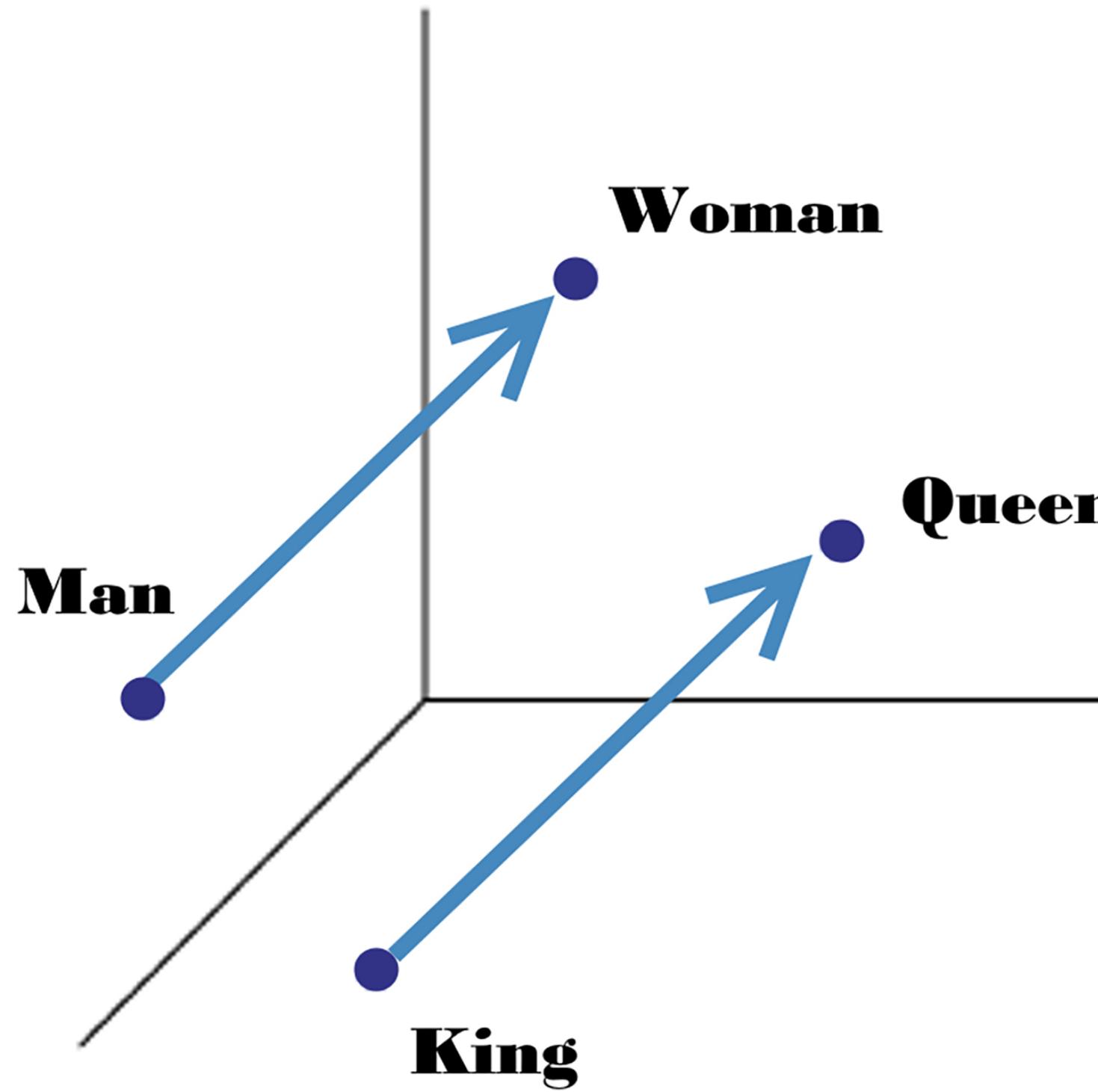
Clustering



Word Embeddings



WIKIPEDIA
The Free Encyclopedia



Word2vec [Mikolov et al'13]

Topic Models

TOPIC 1

computer,
technology,
system,
service, site,
phone,
internet,
machine

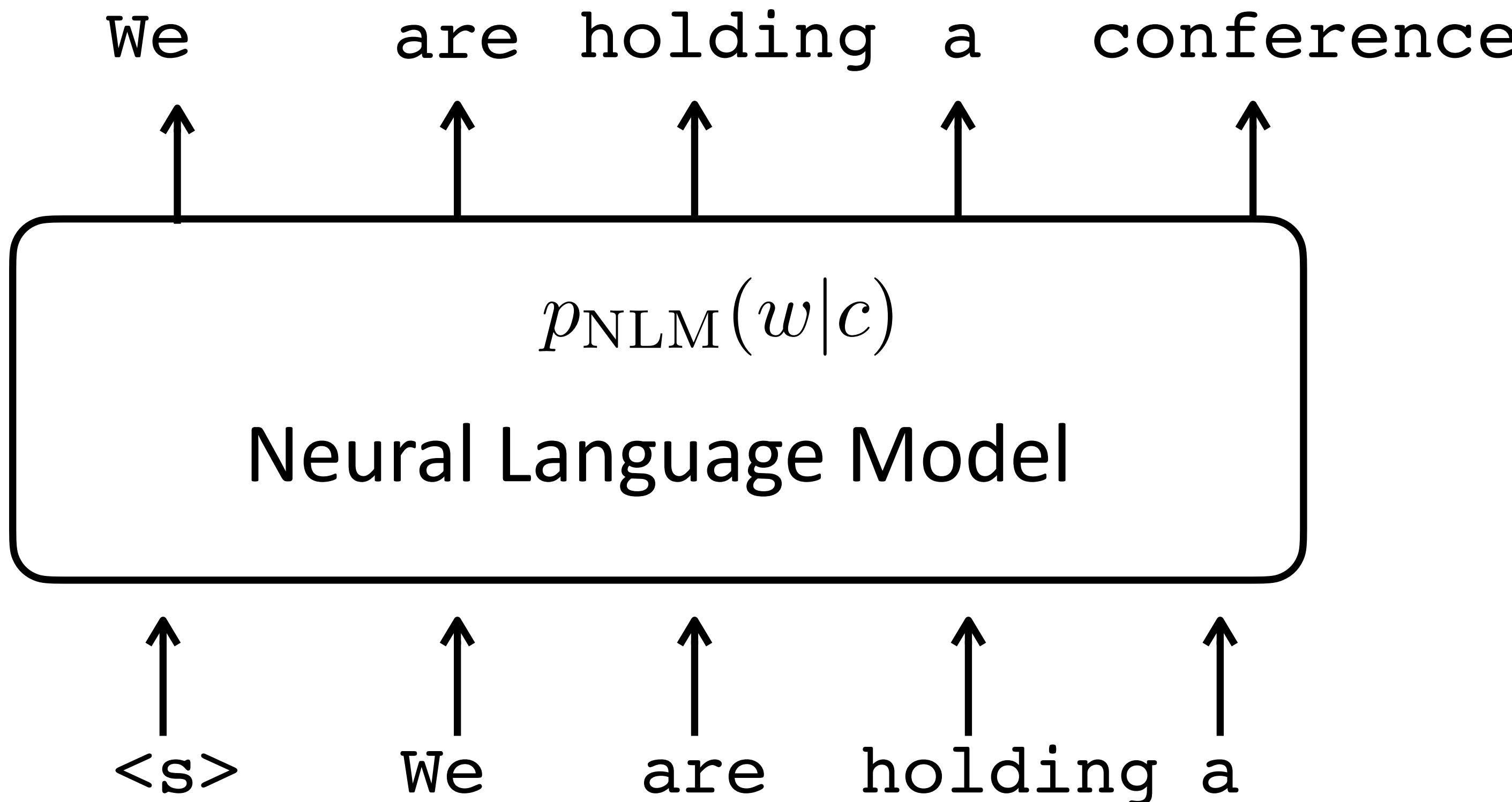
TOPIC 2

sell, sale,
store, product,
business,
advertising,
market,
consumer

TOPIC 3

play, film,
movie, theater,
production,
star, director,
stage

Language Models



Large Language Models

SYSTEM PROMPT
(HUMAN-WRITTEN)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

MODEL COMPLETION
(MACHINE-WRITTEN,
10 TRIES)

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

Large Language Models

Context → Please unscramble the letters into a word, and write that word:
taefed =

Target Completion → defeat

Context → L'analyse de la distribution de fréquence des stades larvaires d'I. verticalis dans une série d'étangs a également démontré que les larves mâles étaient à des stades plus avancés que les larves femelles. =

Target Completion → Analysis of instar distributions of larval *I. verticalis* collected from a series of ponds also indicated that males were in more advanced instars than females.

Context → Q: What is 95 times 45?
A:

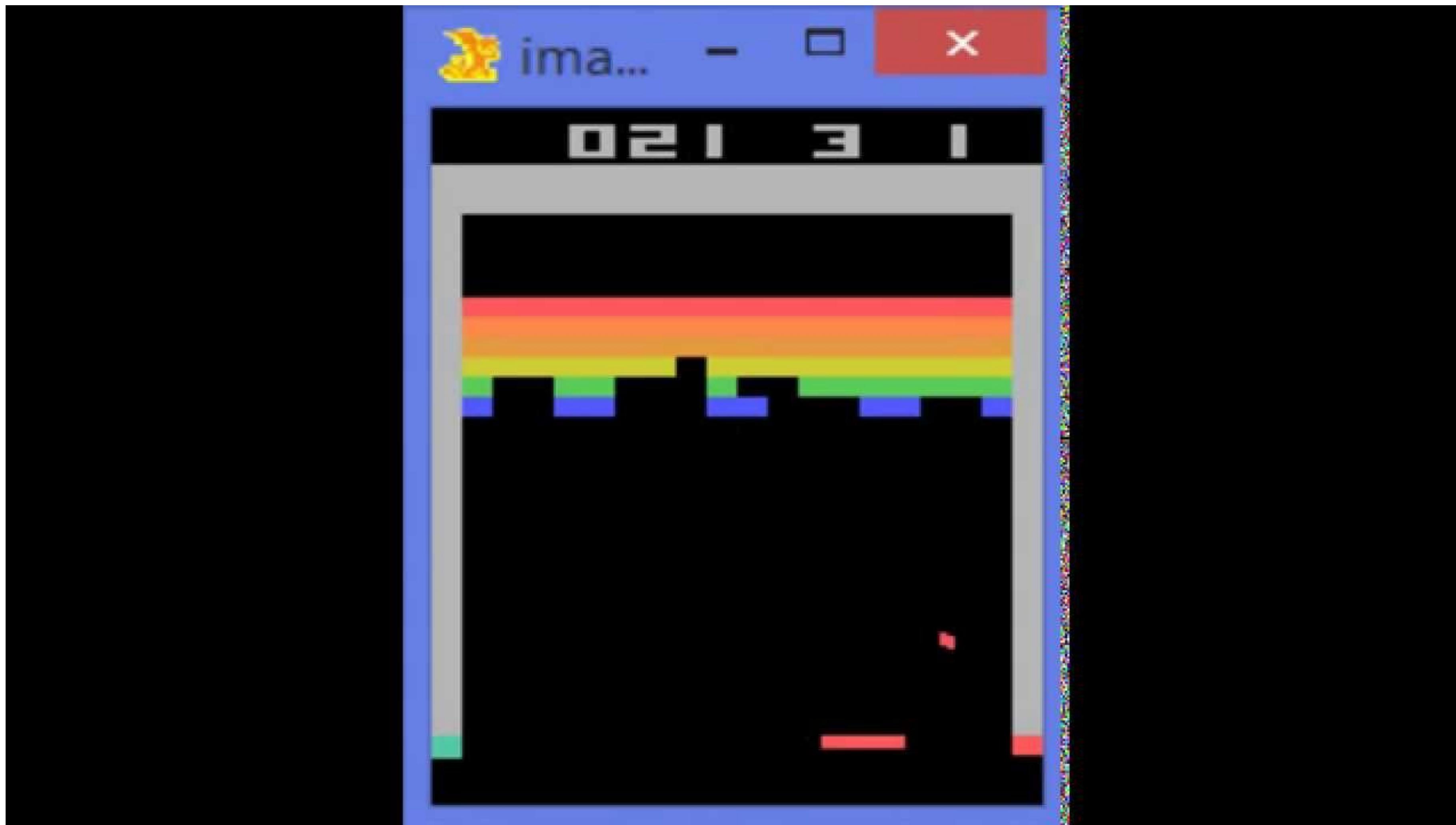
Target Completion → 4275

Reinforcement Learning

AlphaGo

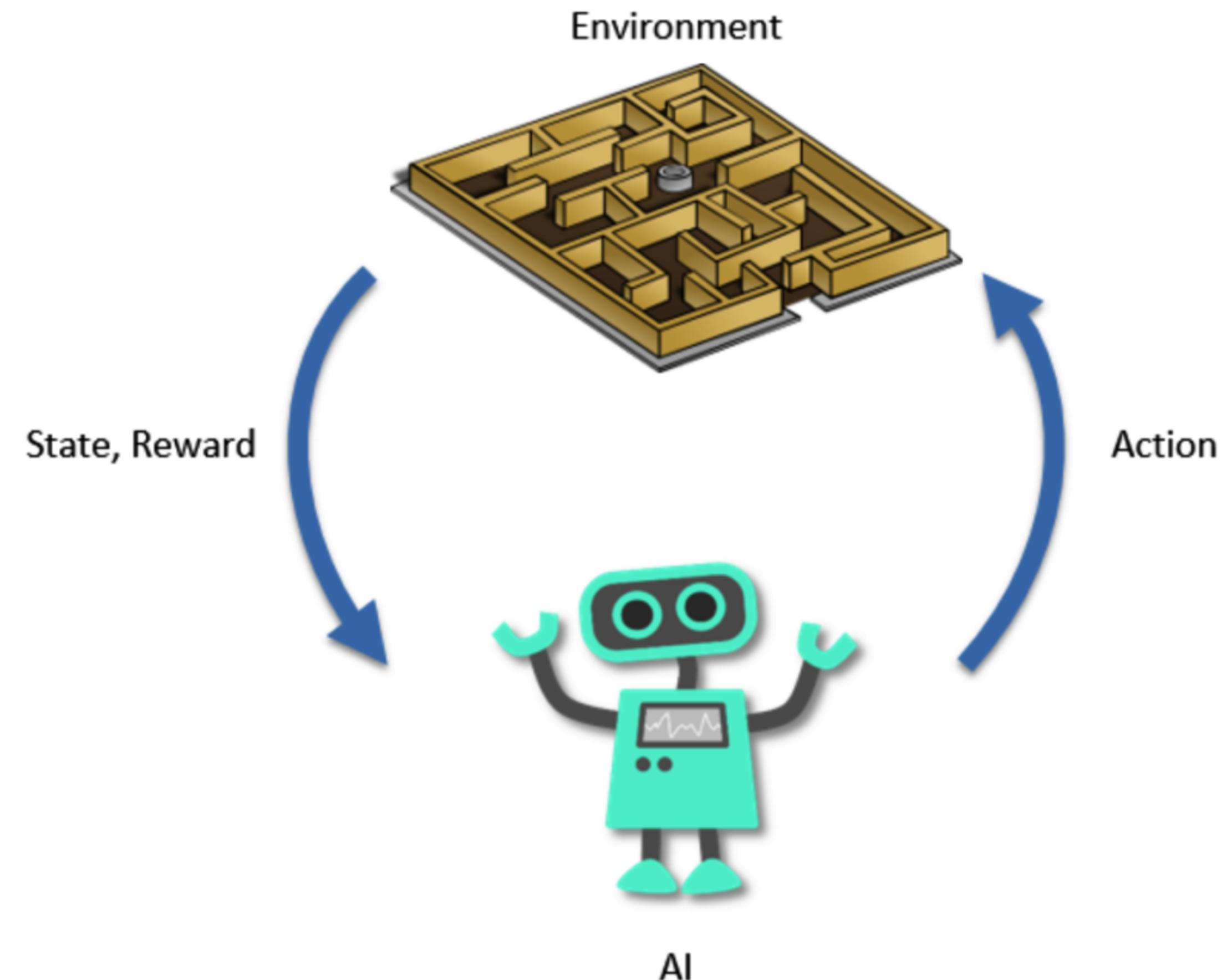


Atari Breakout Game



Reinforcement Learning

- RL can collect data interactively



**Thank You!
Questions?**