香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

COMP 5212
Machine Learning

Instructor: Junxian He

Course website: https://jxhe.github.io/teaching/comp5212s26

Teaching Team & Office Hours

Instructor: Junxian He

TA1: Kashun Shum (ksshumab@connect.ust.hk)

TA2: Junlong Li (jlini@cse.ust.hk)

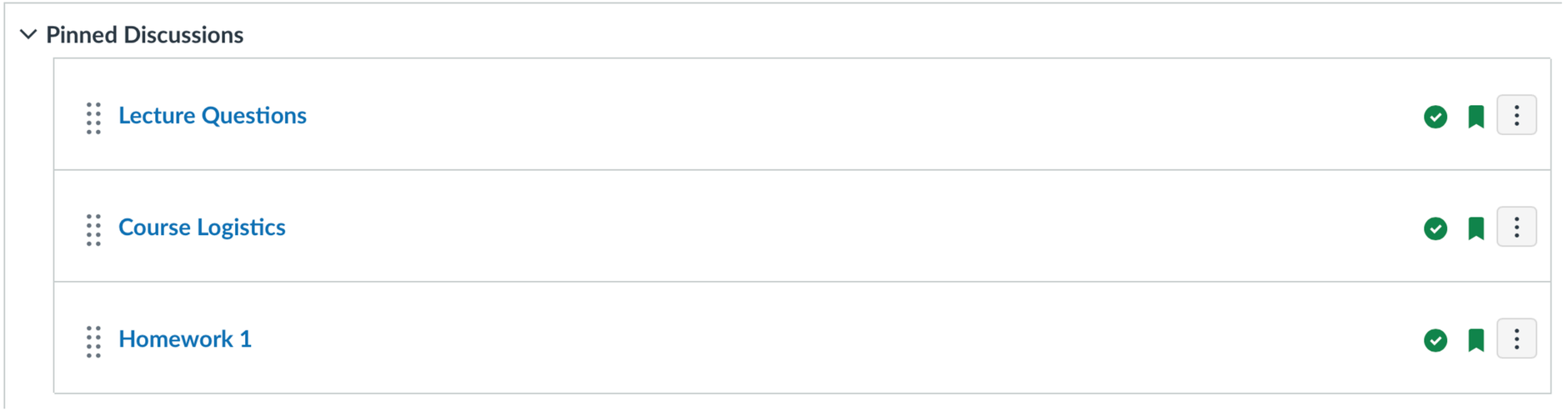
TA3: Hongxiang Li (hlihg@connect.ust.hk)

This is an in-person Class

- Lectures are recorded, but the videos will be released only twice — during the midterm and before the final

- Lecture slides will be available before each lecture

Communication and Discussion



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%)

- 4 assignments (50%)
  - 3 Written + (optional) programing assignments, will be mostly written (12% each)
  - 1 programming-only assignment (14%)
  - 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 + programing assignments (4 * 9%)
  - 1 programming-only assignment
  - 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

- The course quota has been increased to 80 as we are monitoring it, we will try to let most of the students on WL in

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

- Audit is not allowed

- If you just want to sit in the class and been added to Canvas, this is fine

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

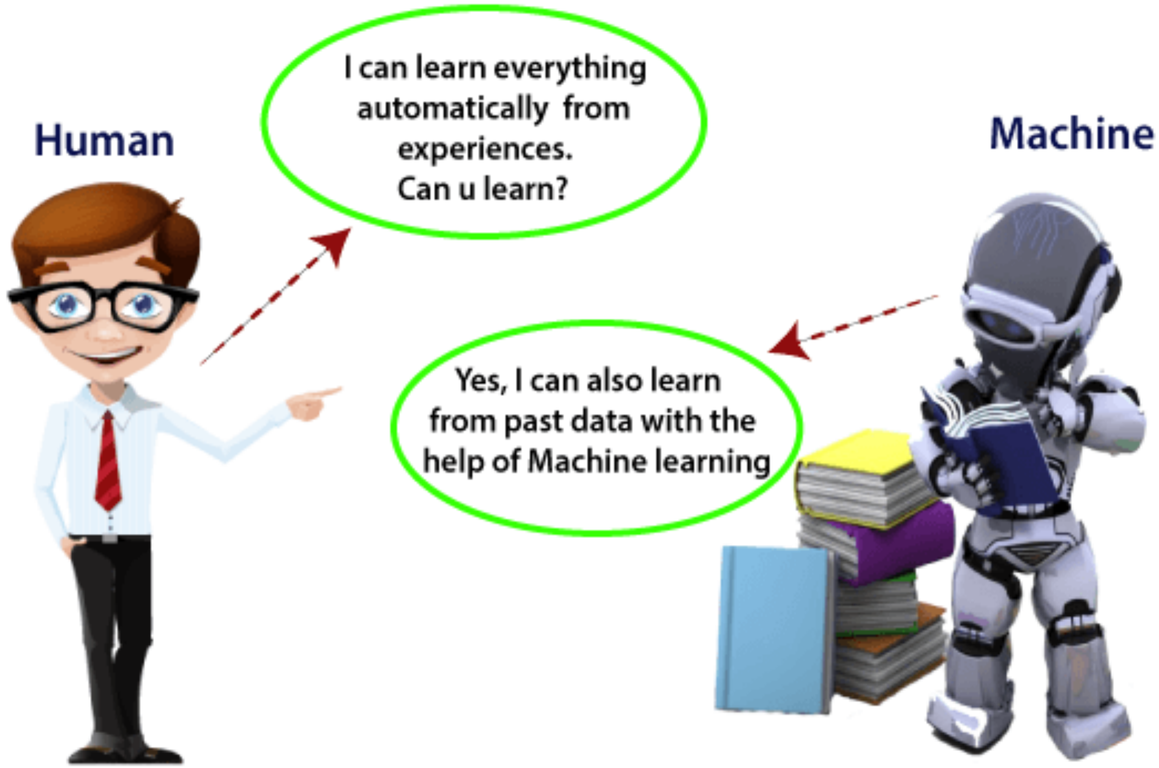
This is NOT an easy course

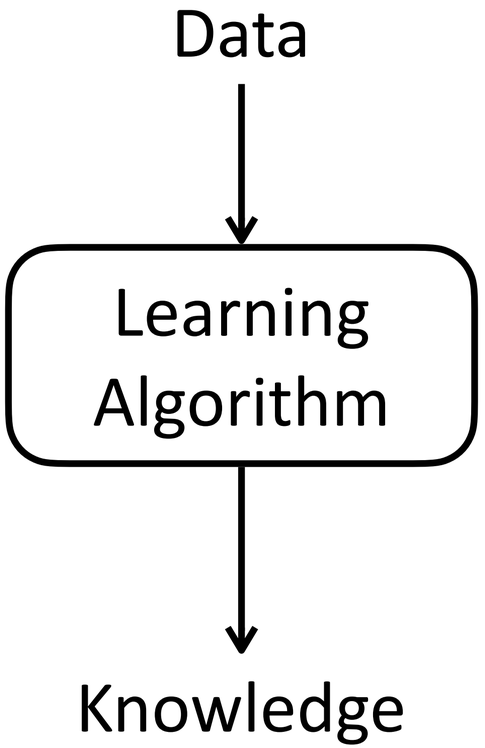
- Overall difficulty comparable to the graduate machine learning course at other top universities like CMU, workload may be slightly lighter
- Homework is the most important and may be difficult sometimes, exams are significantly easier
- It is wrong to regard graduate courses as nothing useful / unhelpful 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

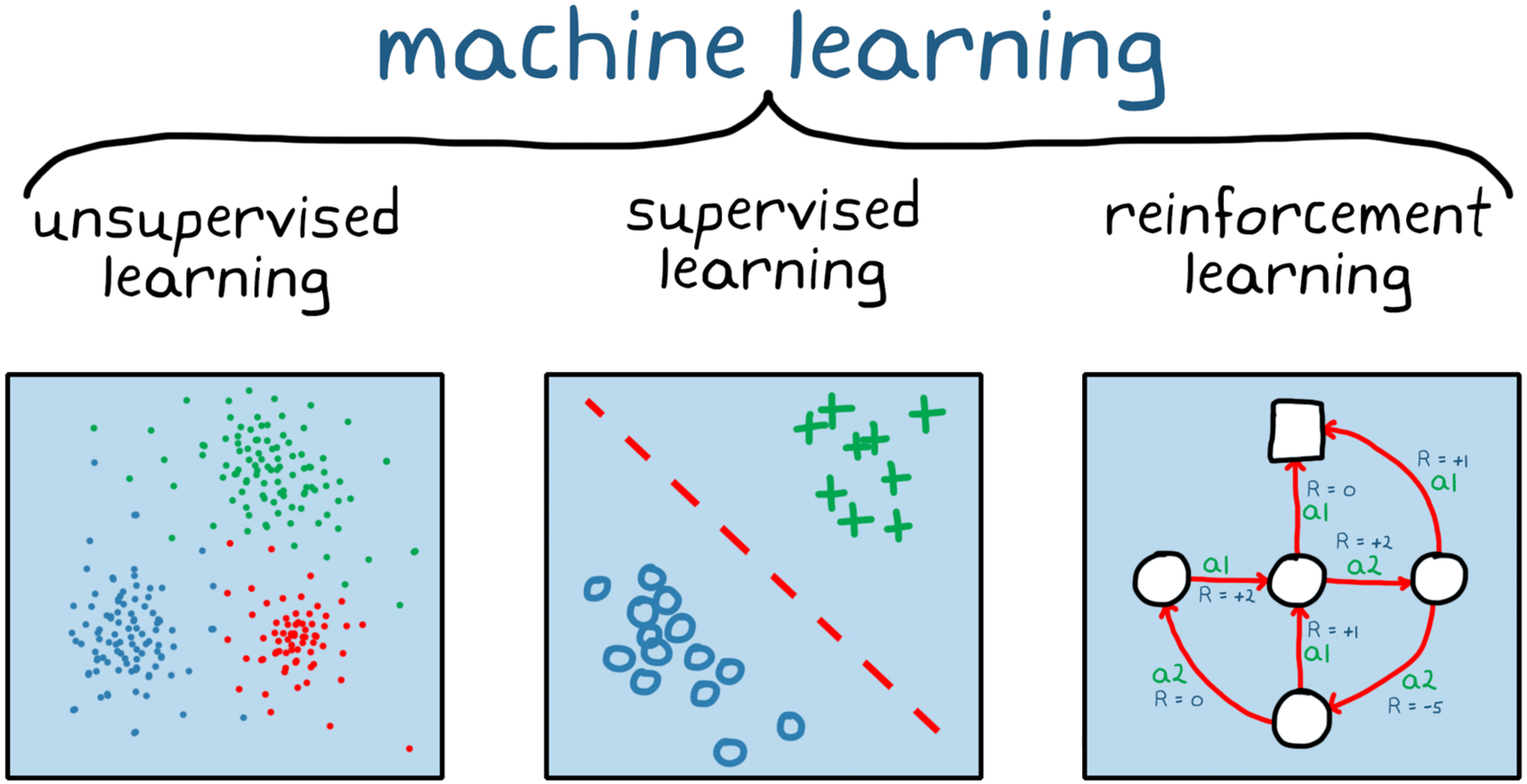




Machine Learning is Trending 🚀🔥

- 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



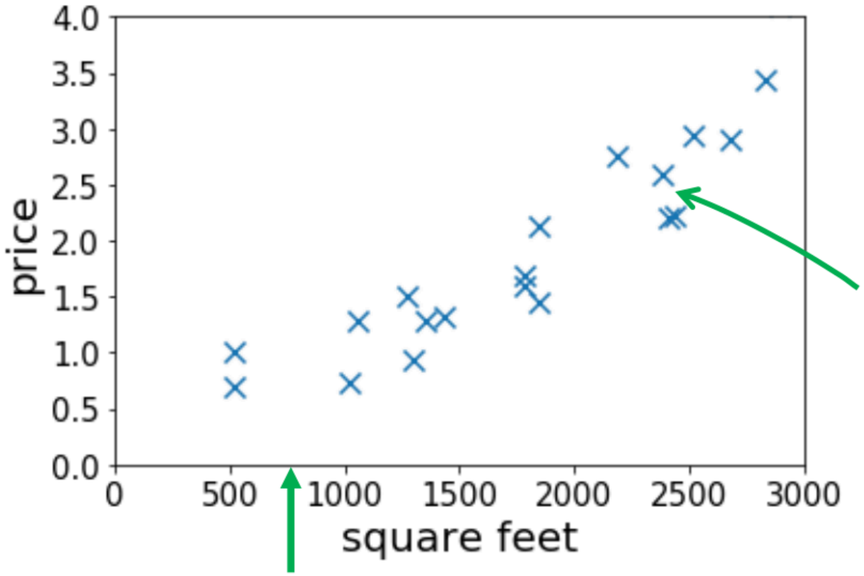
Supervised Learning

Housing Price Prediction

- Given: a dataset that contains n samples

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

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



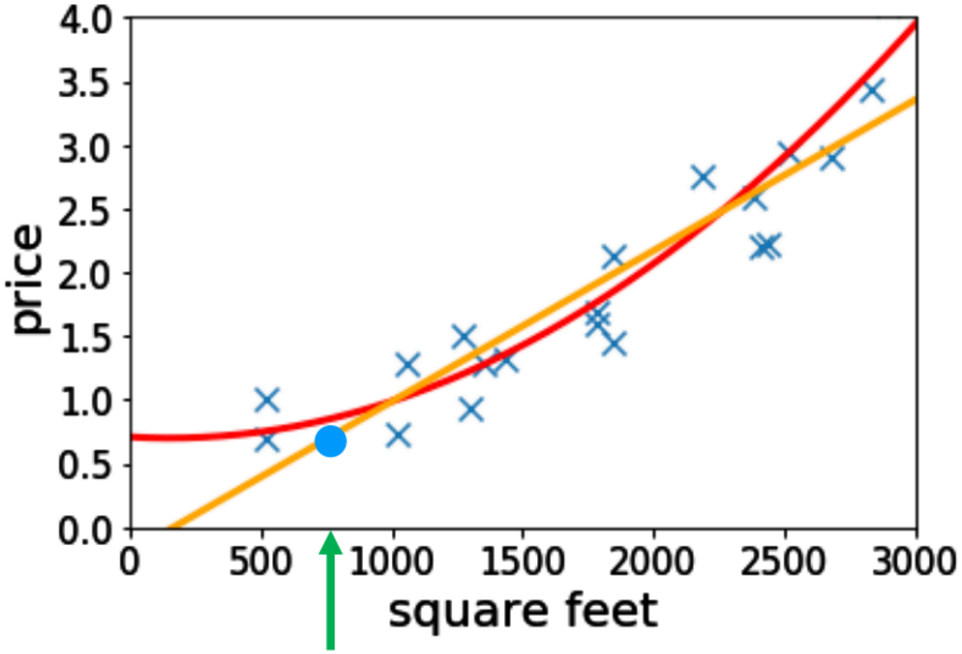
Example from Stanford CS229

Housing Price Prediction

- Given: a dataset that contains n samples

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

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



More Features

- Suppose we also know the lot size

Example from Stanford CS229

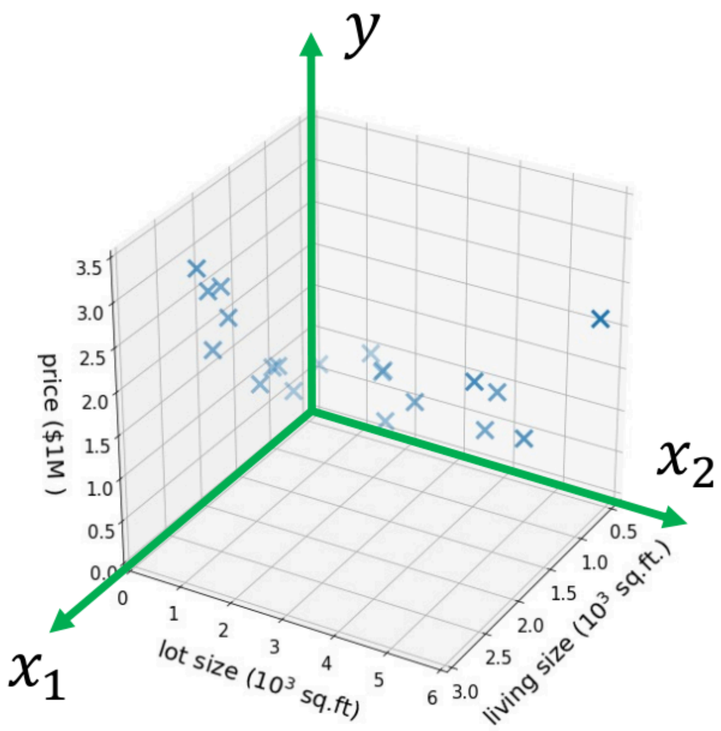
More Features

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

(size, lot size) → price

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

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



More Features

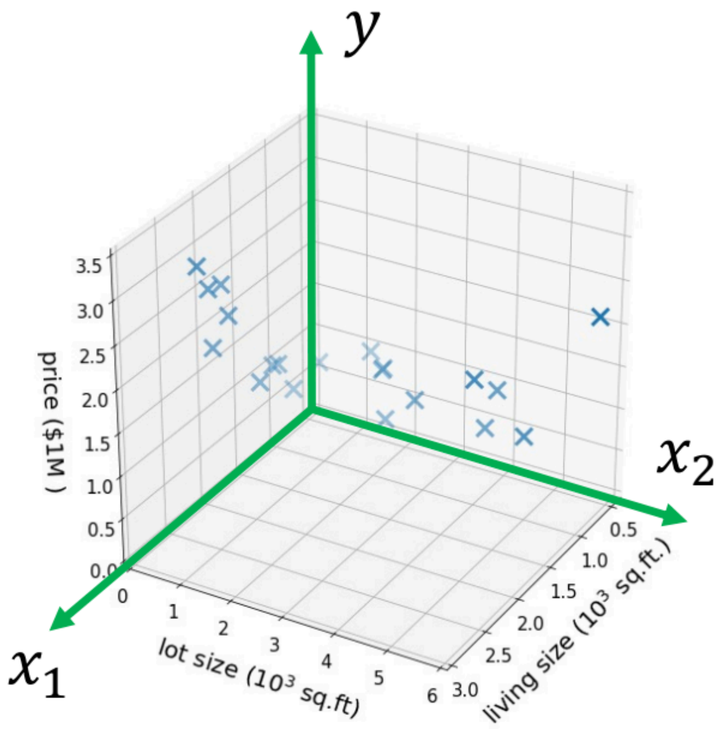
- Suppose we also know the lot size
- Task: find a function that maps

(size, lot size) → price

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

- Dataset: $\left(x^{(1)}, y^{(1)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right)$

where $x^{(i)} = (x_1^{(i)}, x_2^{(i)})$



High-dimensional Features

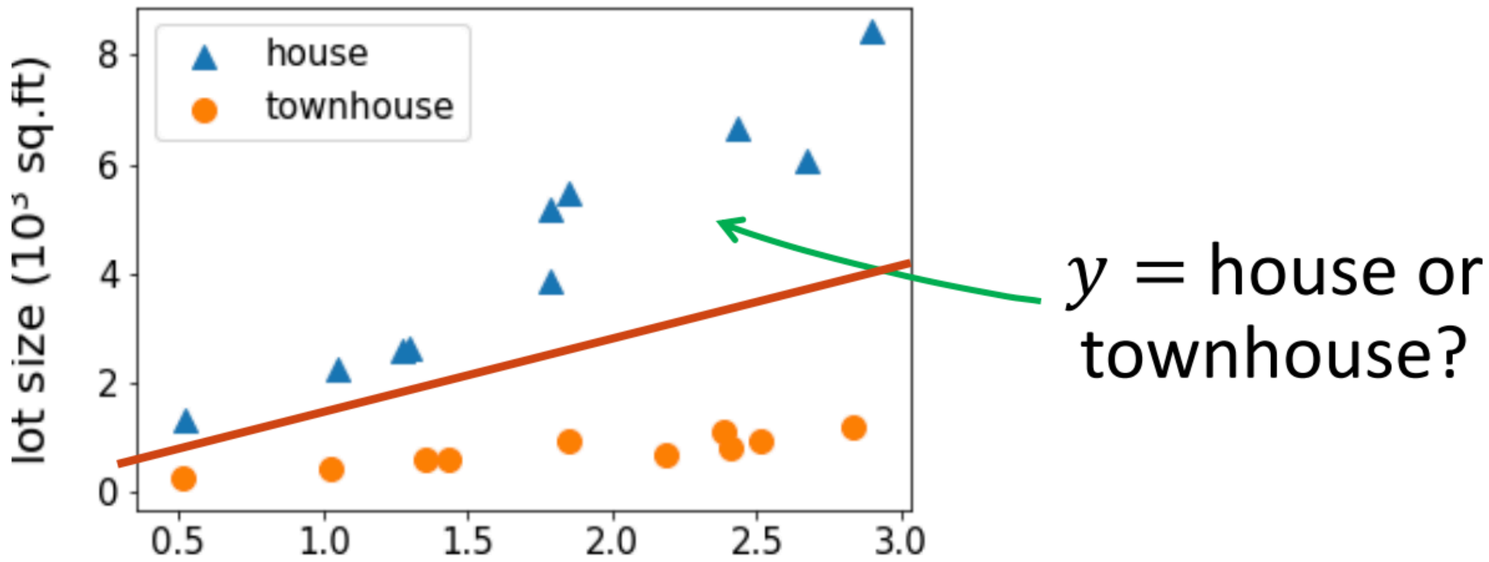
$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ \vdots \\ \vdots \\ x_d \end{bmatrix}$$

--- living size
--- lot size
--- # floors
--- condition
--- zip code
⋮

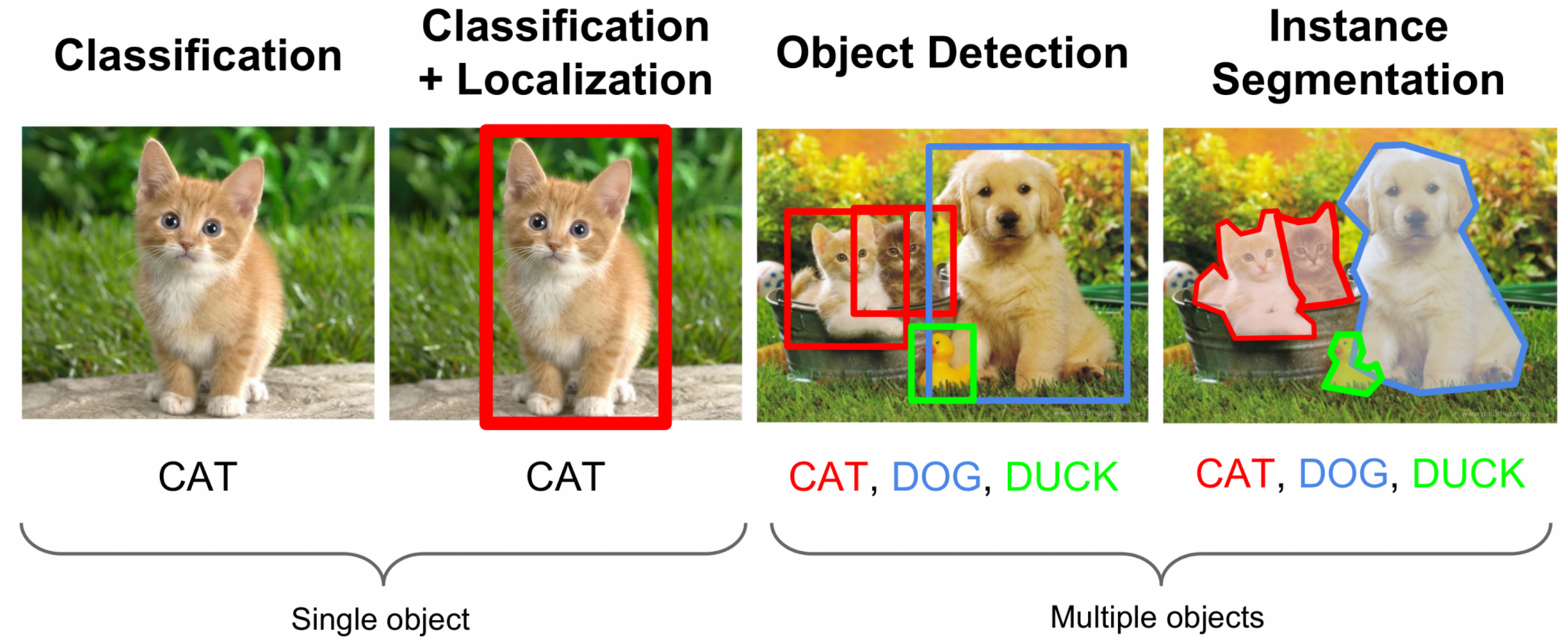
→ y --- price

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



Supervised Learning in Natural Language Processing

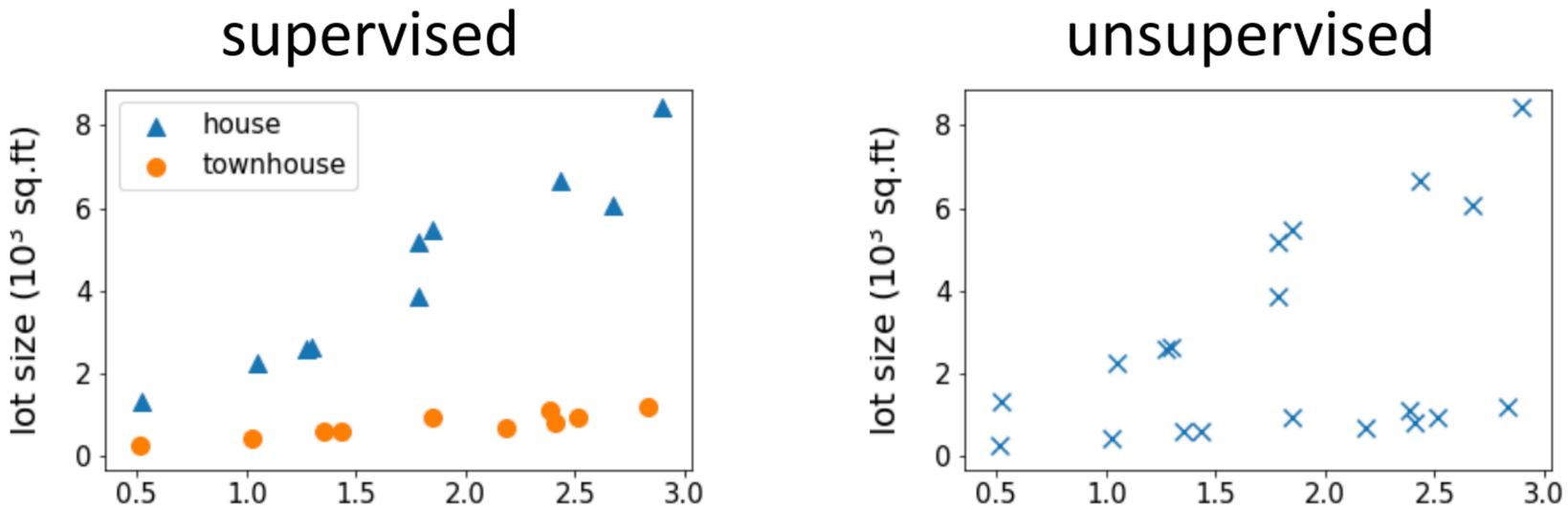
Hey Siri

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

Unsupervised Learning

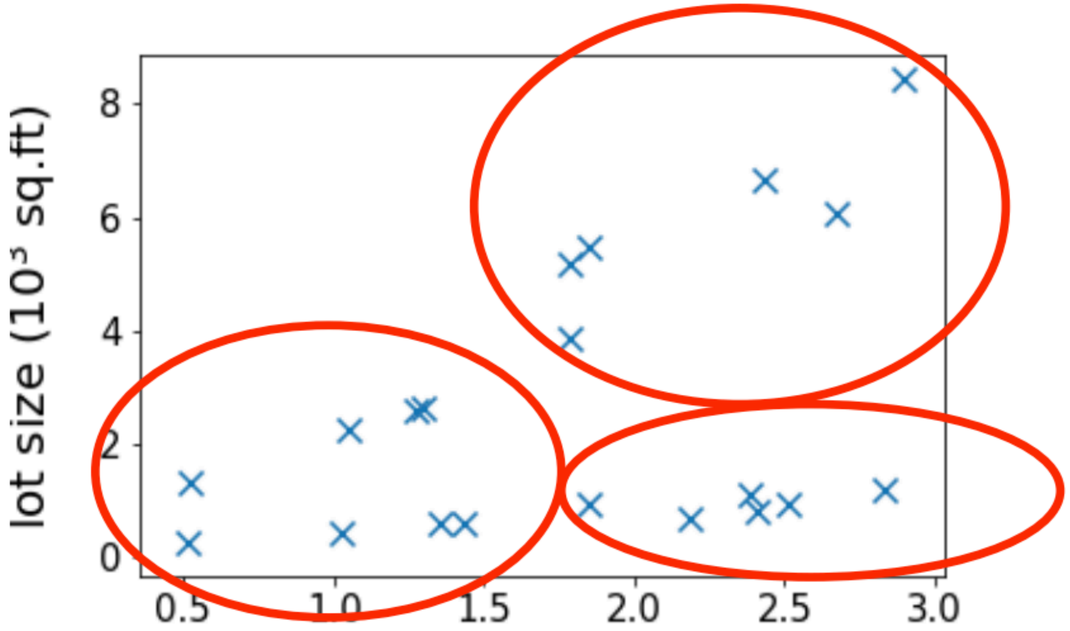
Unsupervised Learning

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

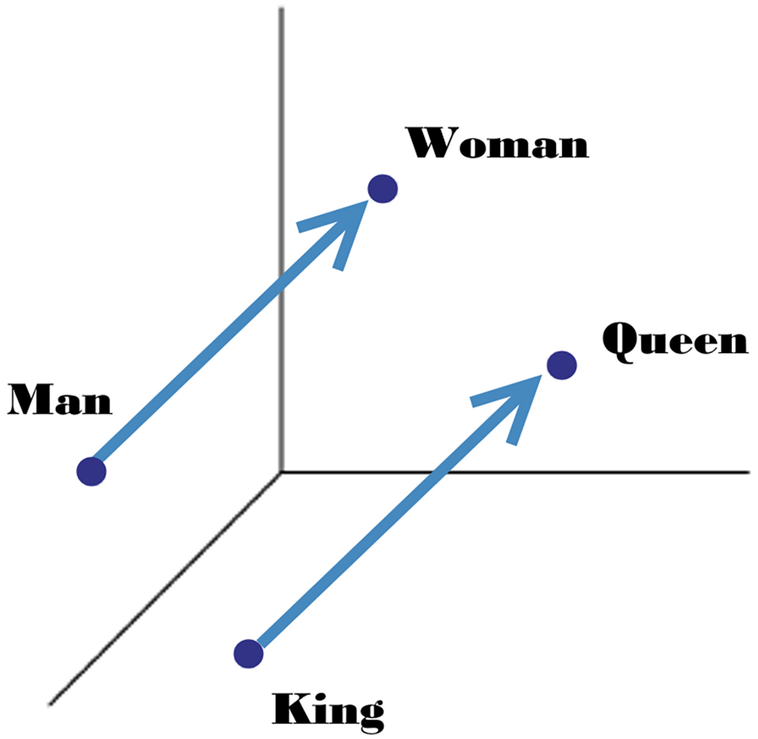


Example from Stanford CS229

Clustering

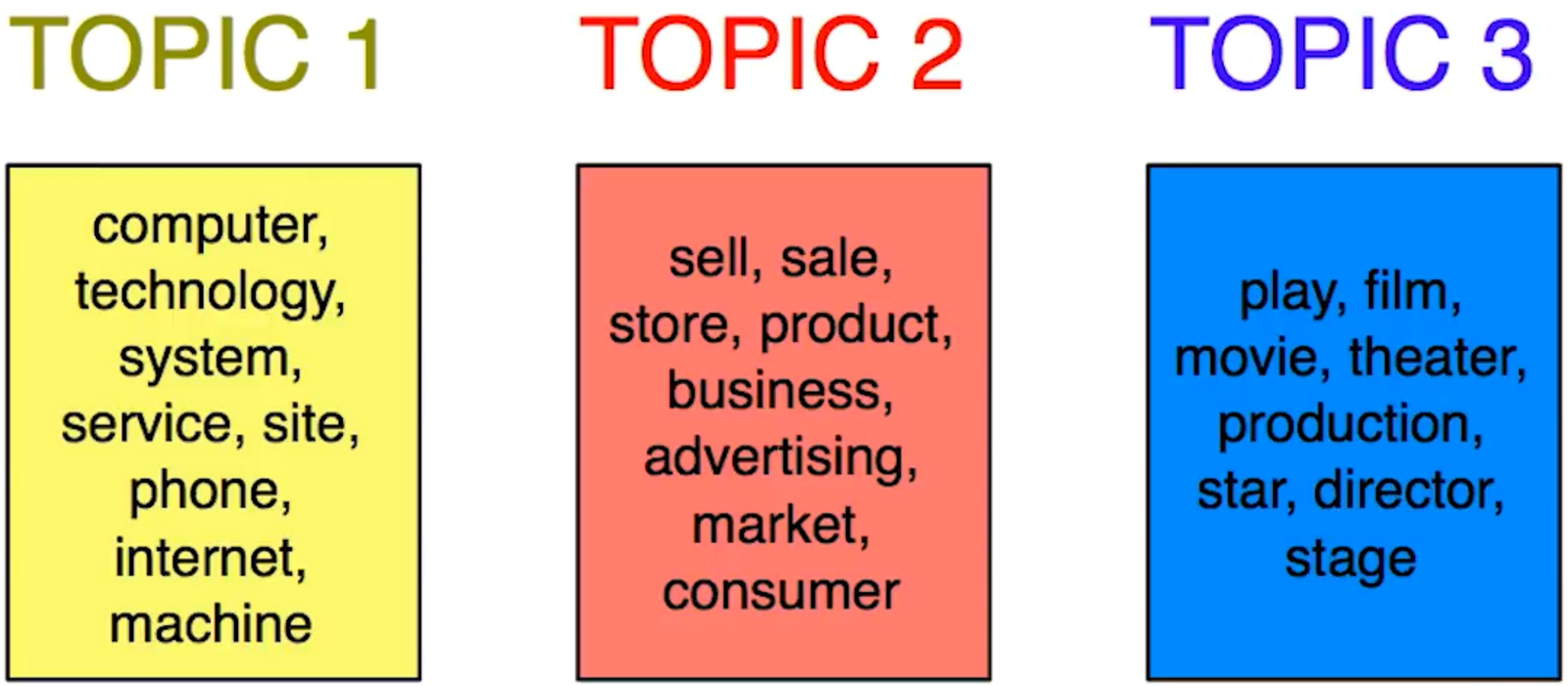


Word Embeddings

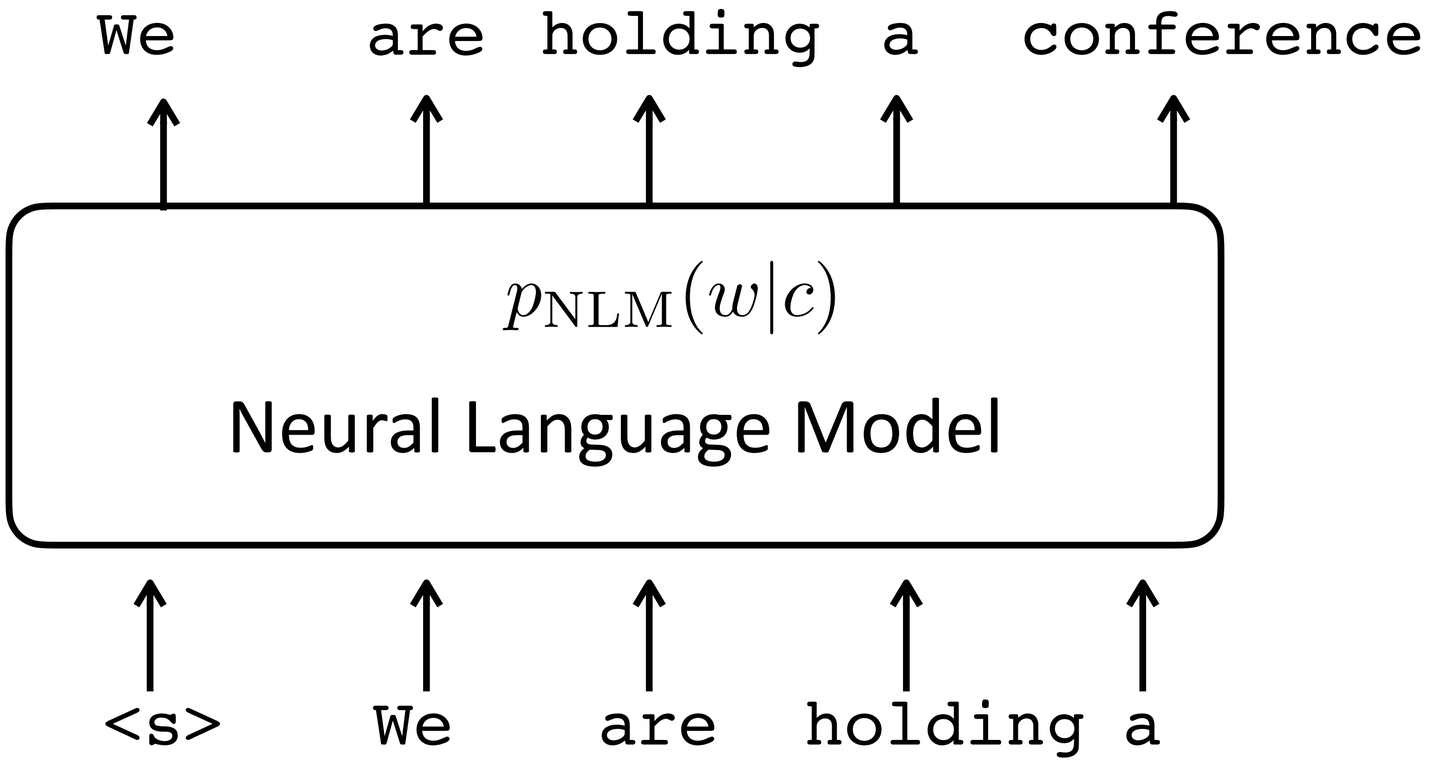


Word2vec [Mikolov et al'13]

Topic Models



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.

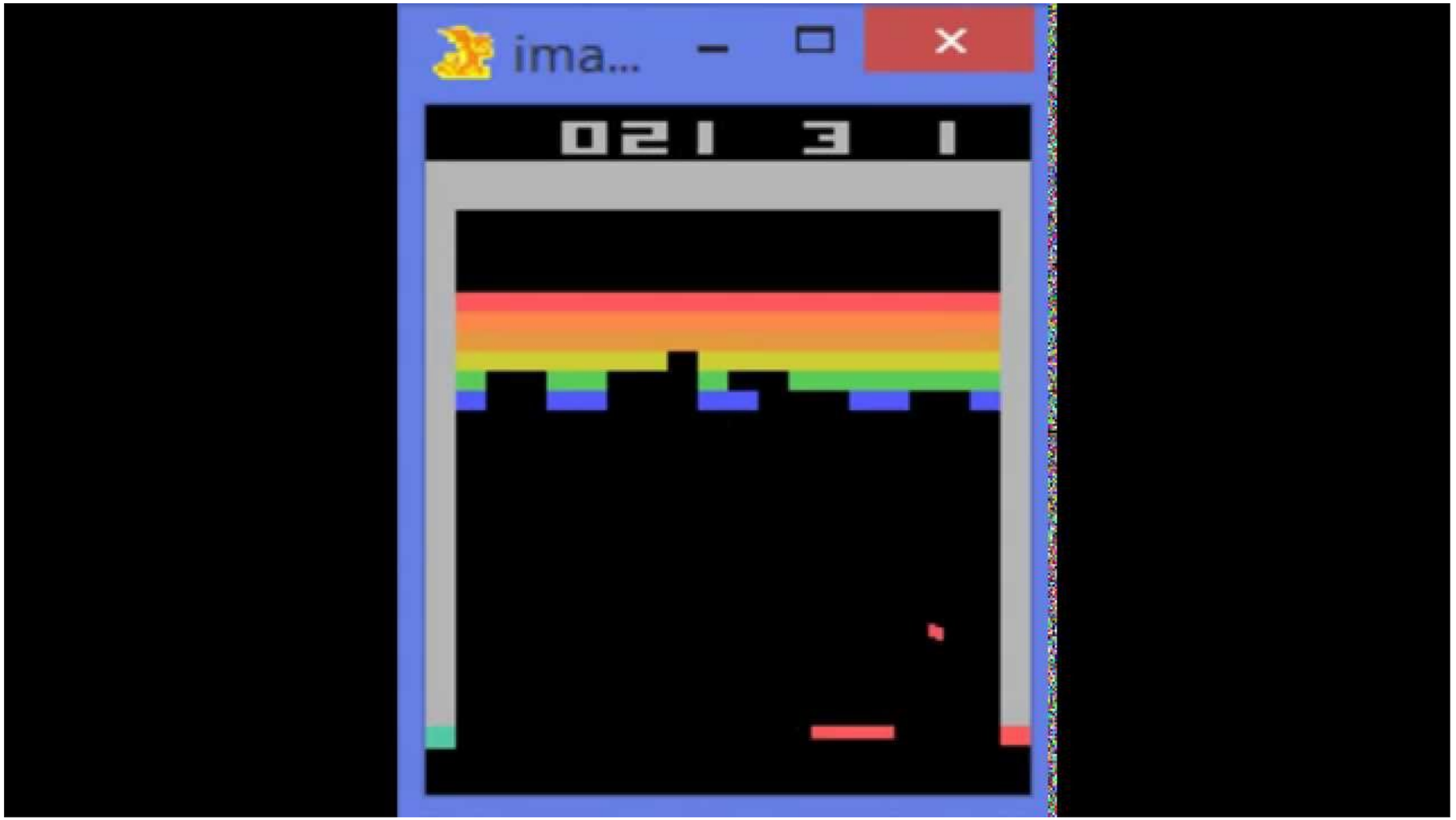
Context →	Q: What is 95 times 45? A:
Target Completion →	4275

Brown et al.'20

Reinforcement Learning

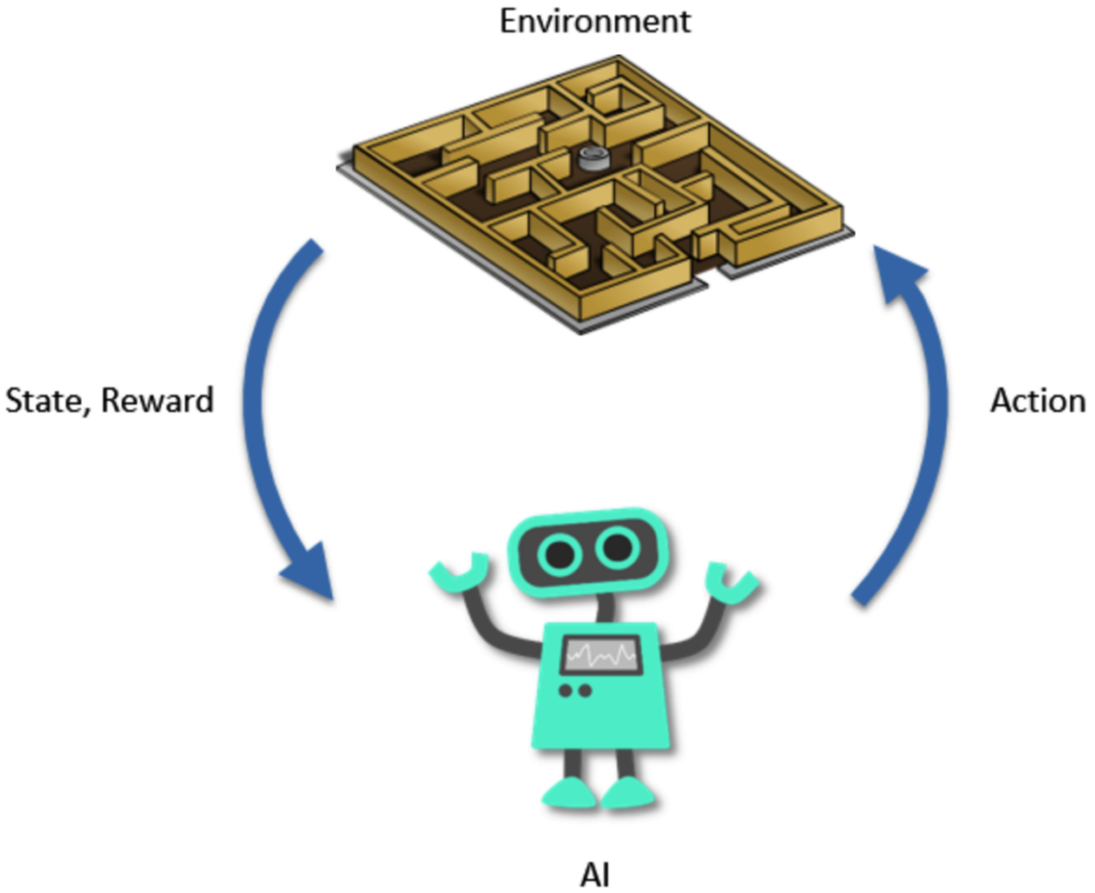
AlphaGo

Atari Breakout Game



Reinforcement Learning

- RL can collect data interactively



Thank You!
Questions?