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/comp5212s24

Teaching Team & Office Hours

Instructor: Junxian He

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

TA1: Kashun Shum

Office Hour: Wed 2pm-3pm, outside room 3657

TA2: Chi Xu

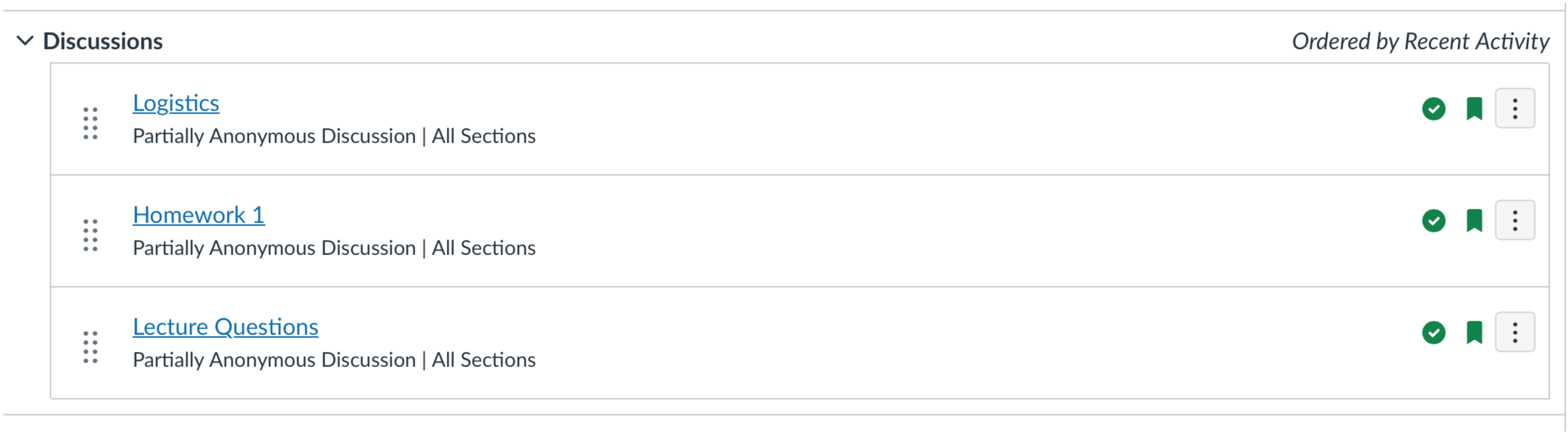
Office Hour: Fri 6pm-7pm, Room 4033

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 after each lecture

Communication and Discussion



Pre-requisite

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

This is not an easy course! It may be mathematically intense for you.
Difficulty is often rewarding :)

Grading

- Attendance (10%)

- 5 assignments (50%)
  - 4 Written + programing assignments
  - 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 + programing 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 + 20 WL)

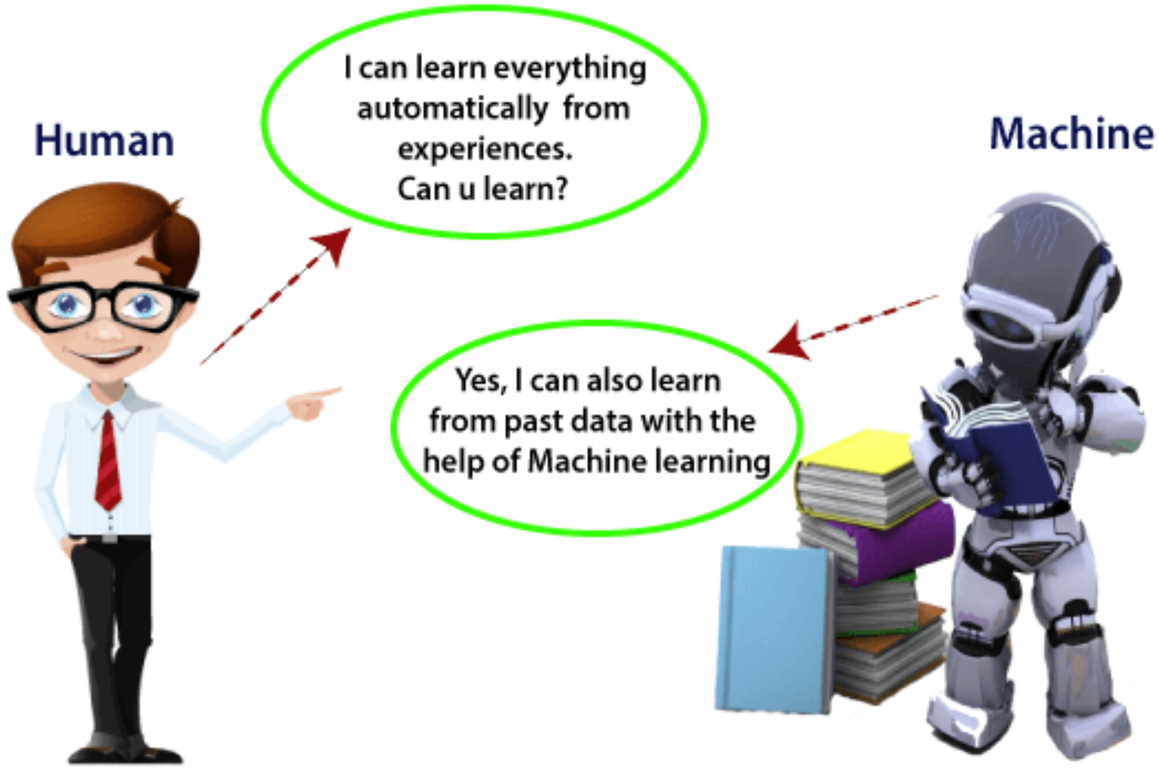
- Audit is allowed, and will be graded in the same criterion

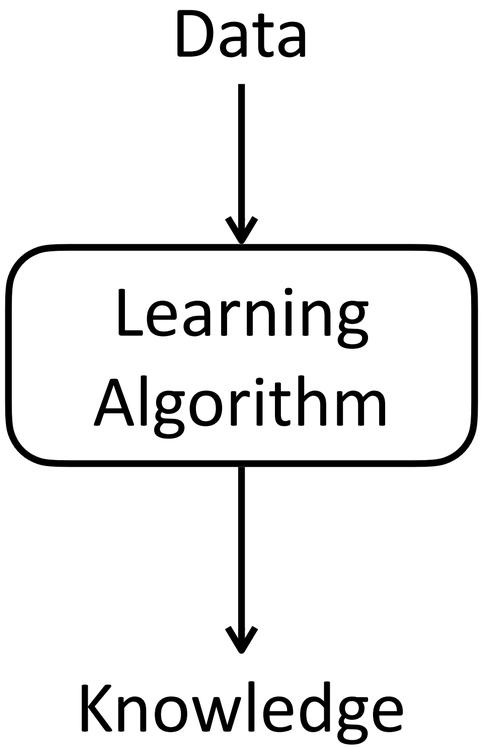
More Info on Course Website

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

Overview of Machine Learning

What is Machine Learning

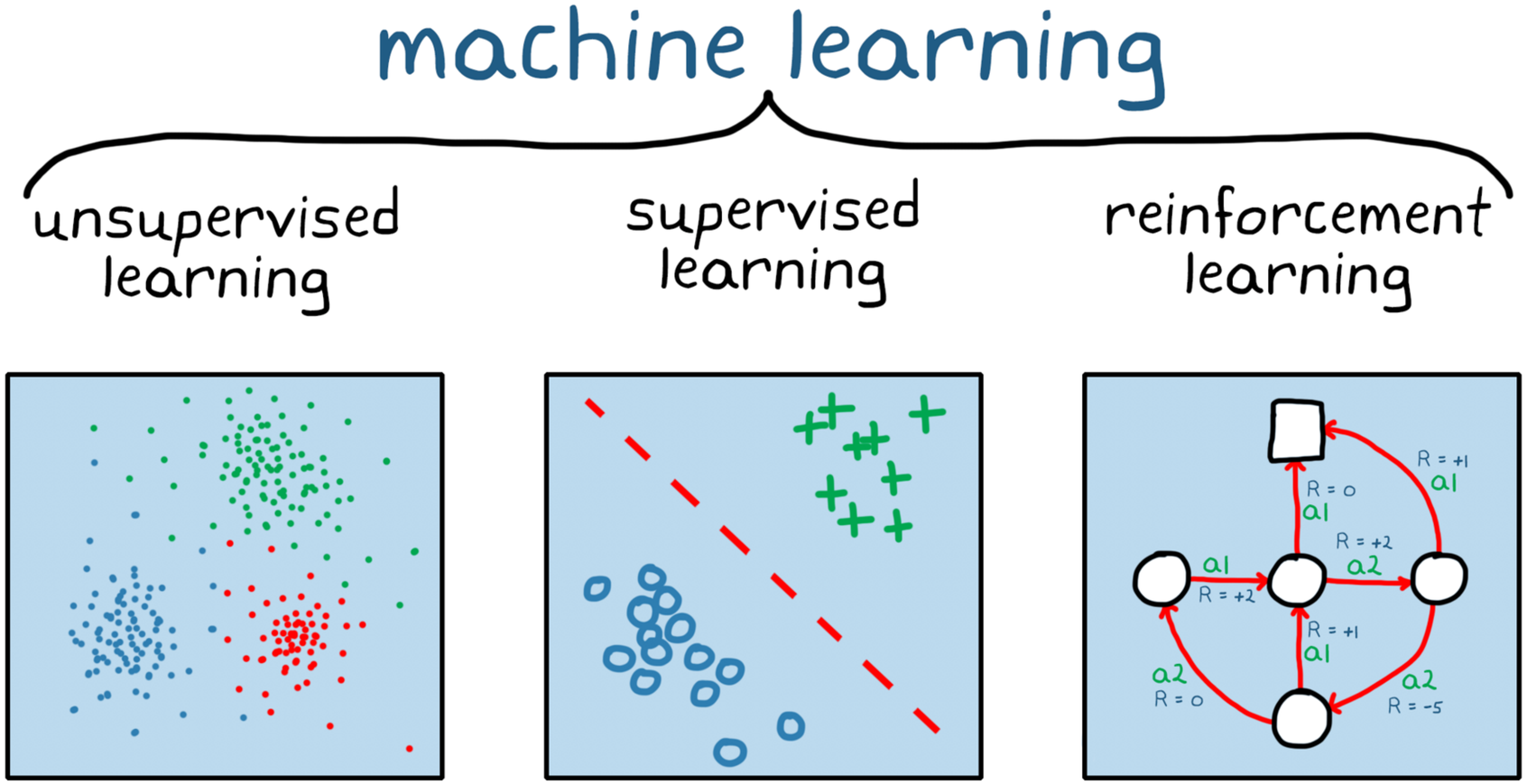




Machine Learning is Trending 🚀🔥

- 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



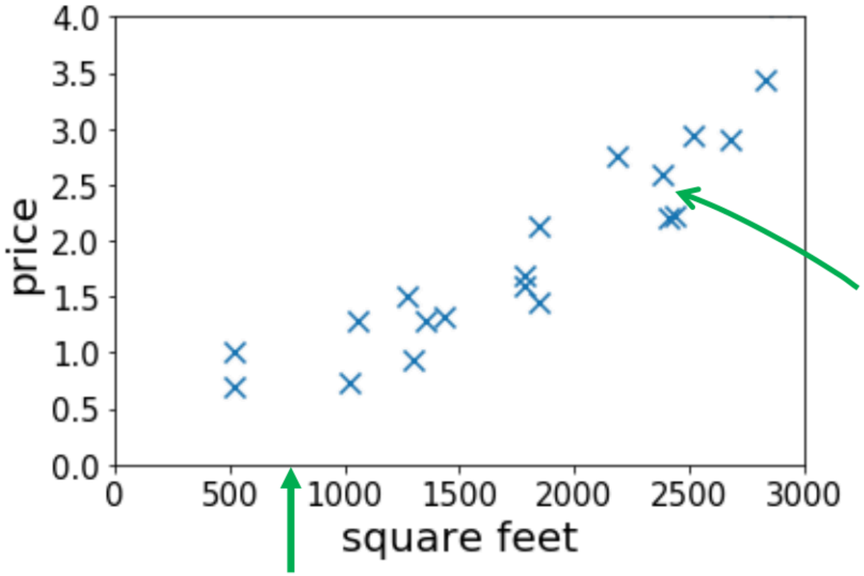
Supervised Learning

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?



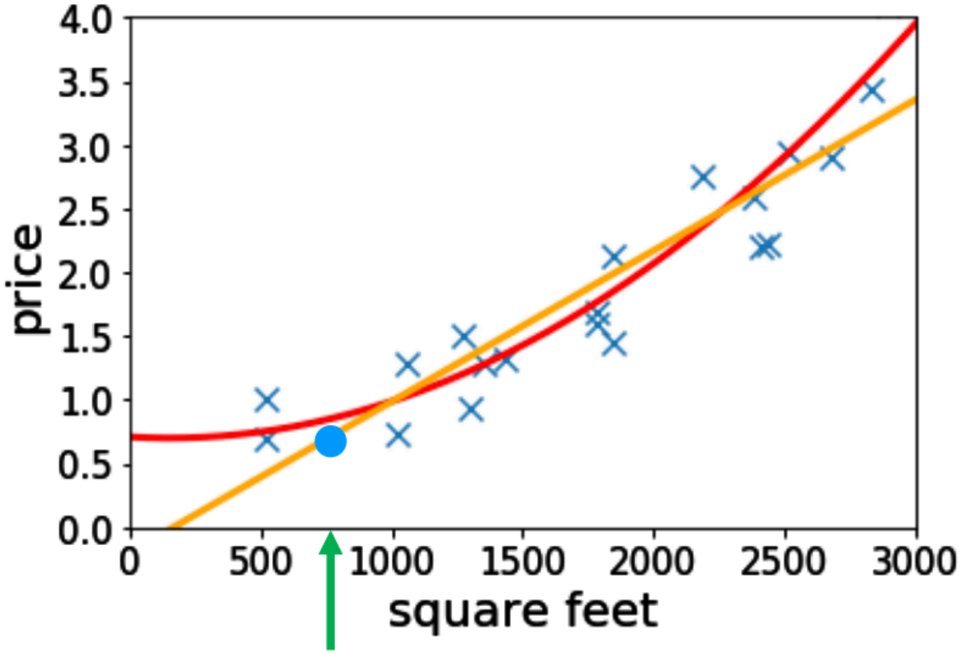
Example from Stanford CS229

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

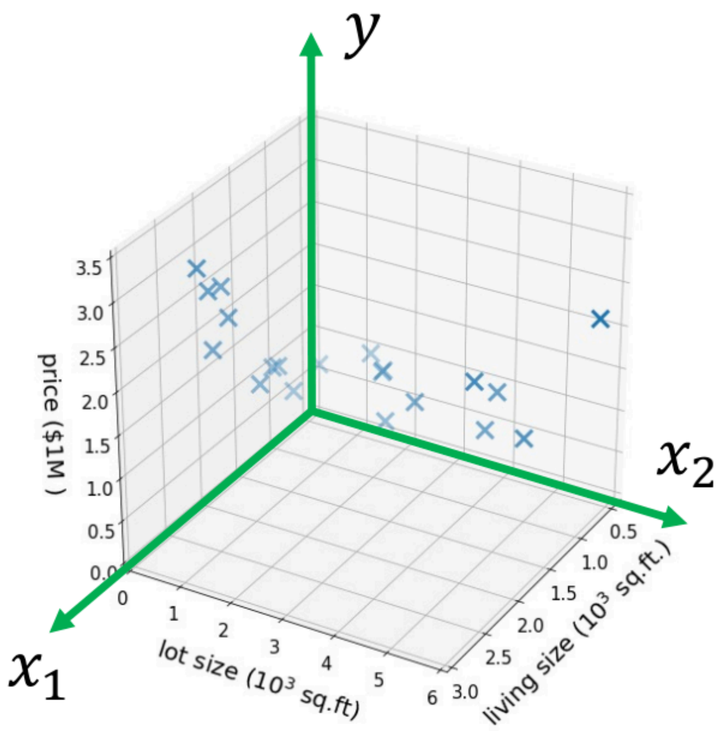
Example from Stanford CS229

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

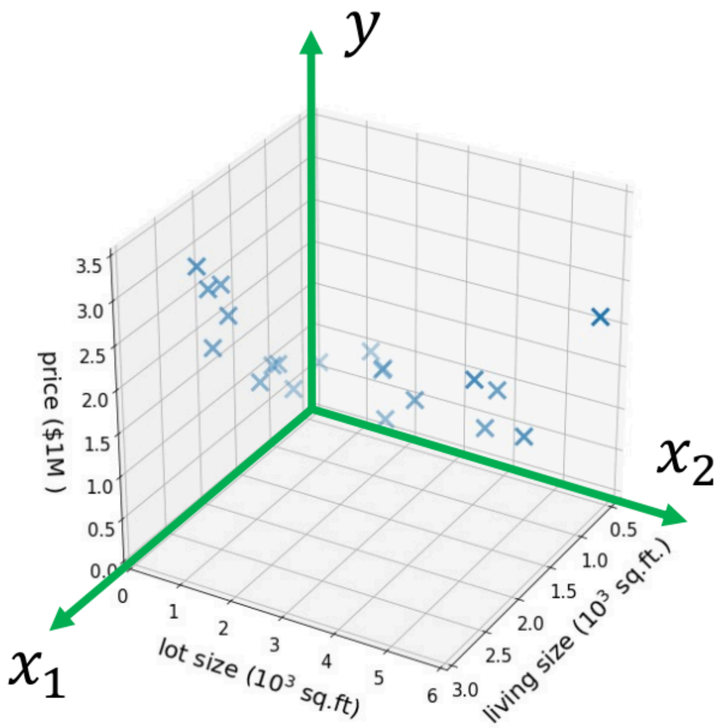
- Suppose we also know the lot size
- Task: find a function that maps

(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

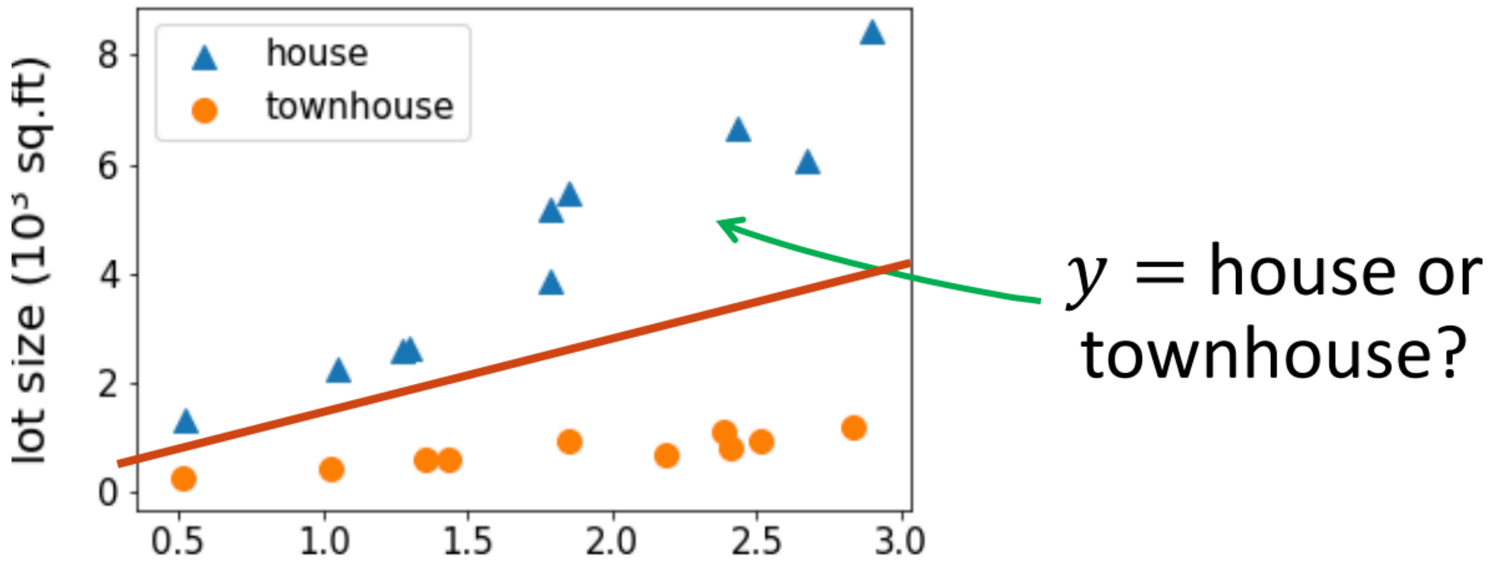
$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ \vdots \\ \vdots \\ x_d \end{bmatrix}$$

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

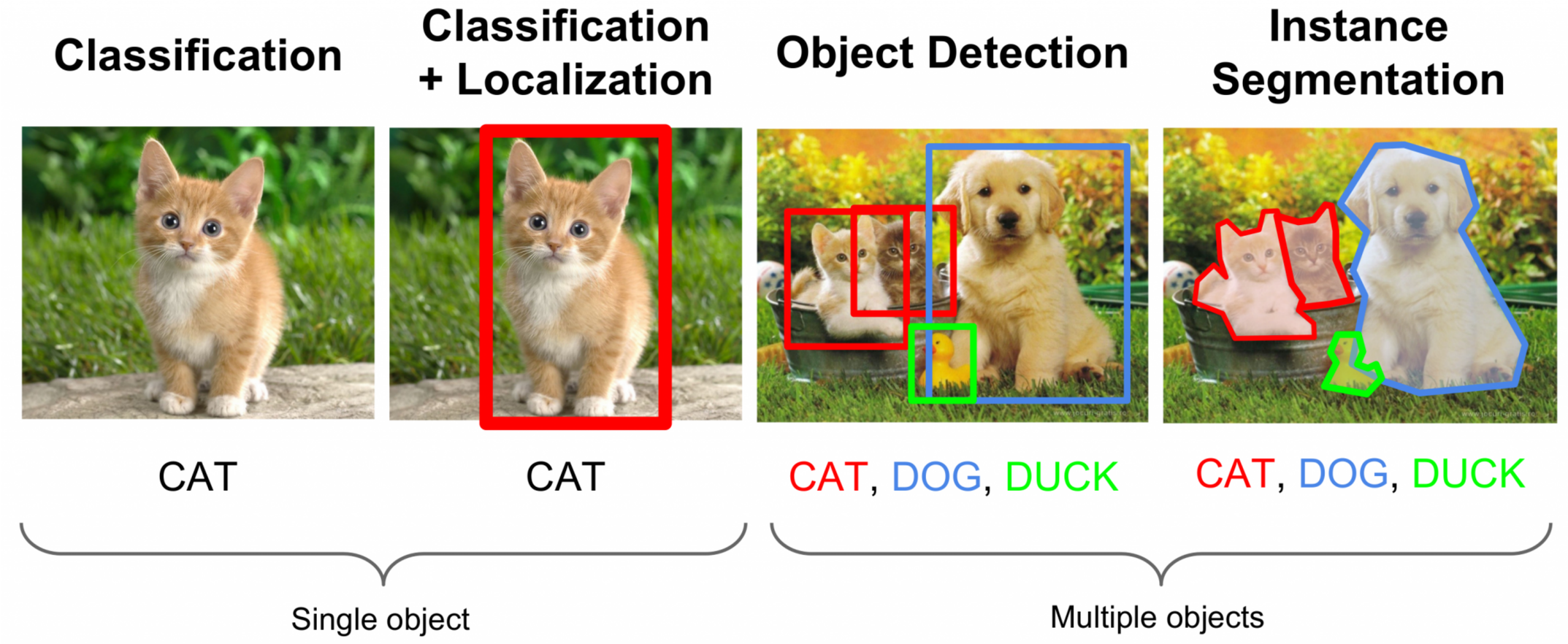
→ y --- price

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

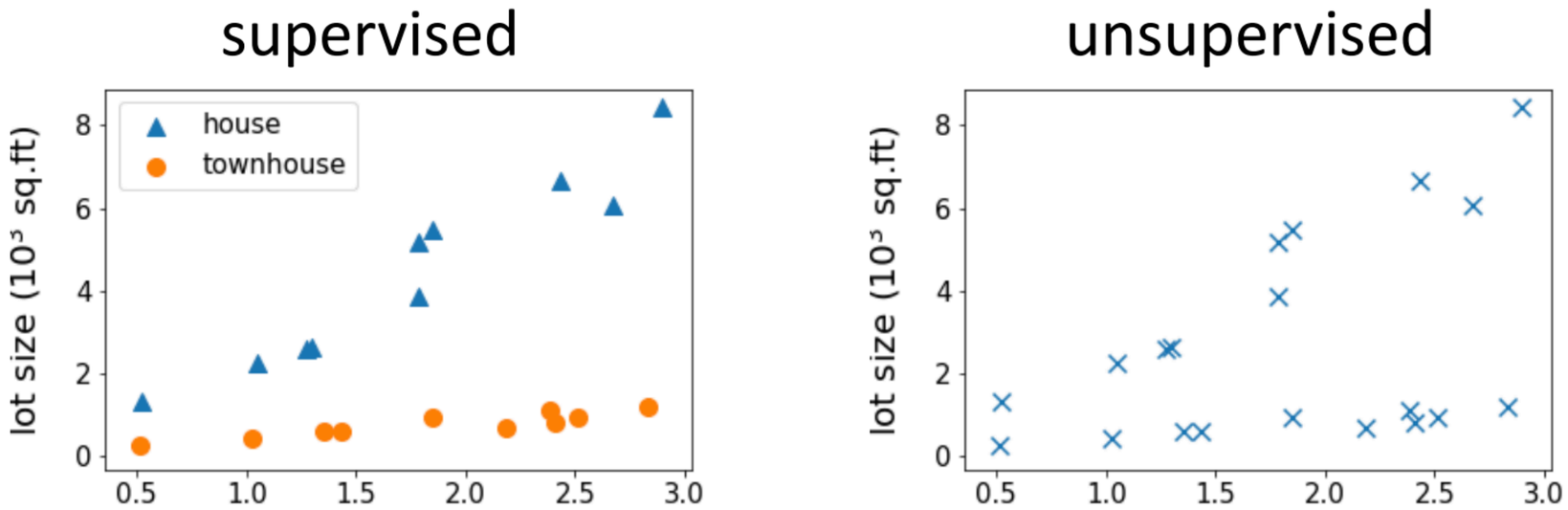
Hey Siri

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

Unsupervised Learning

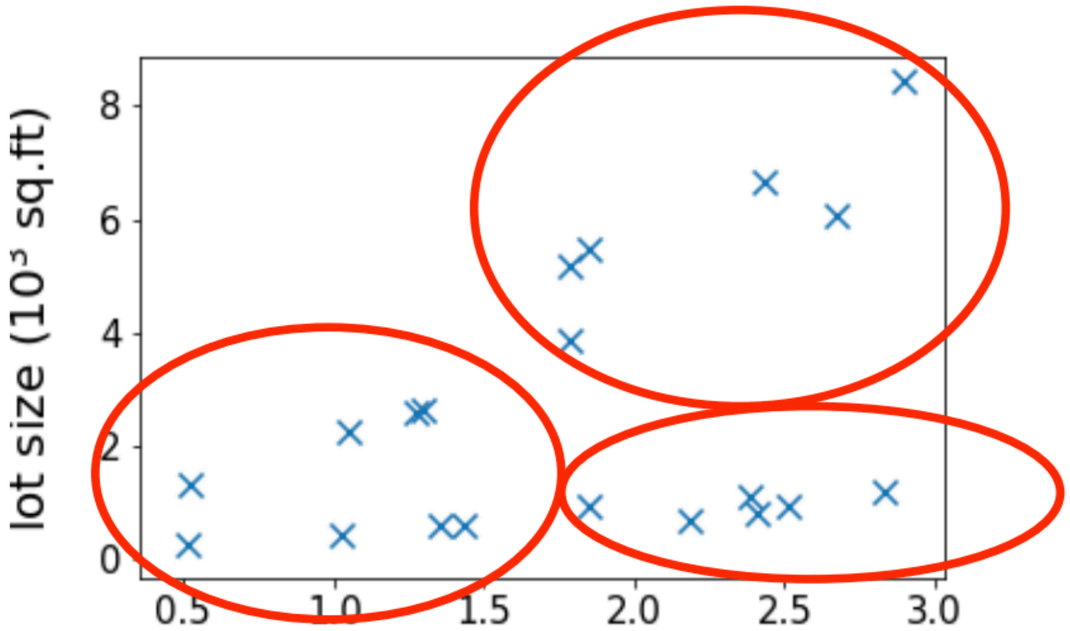
Unsupervised Learning

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

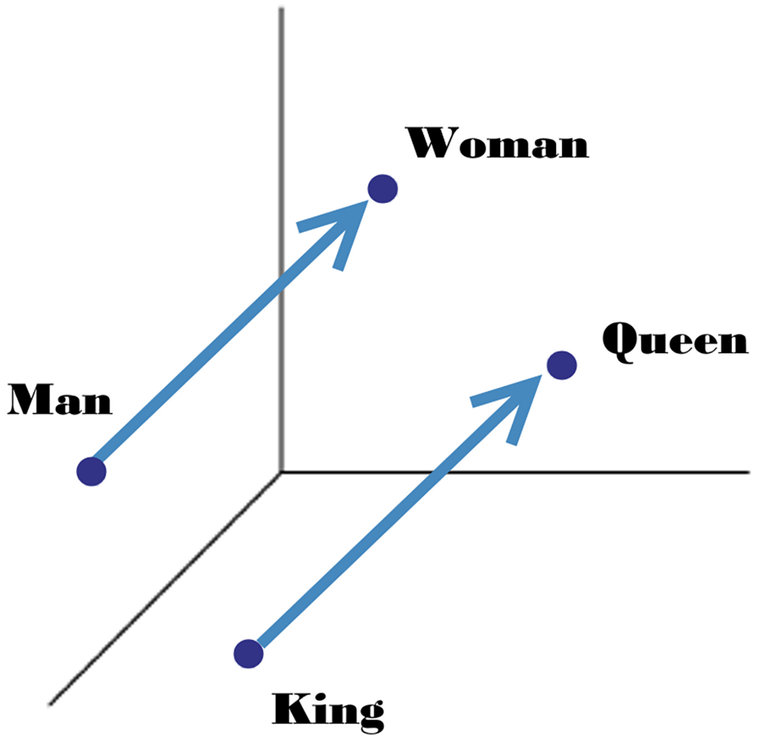


Example from Stanford CS229

Clustering

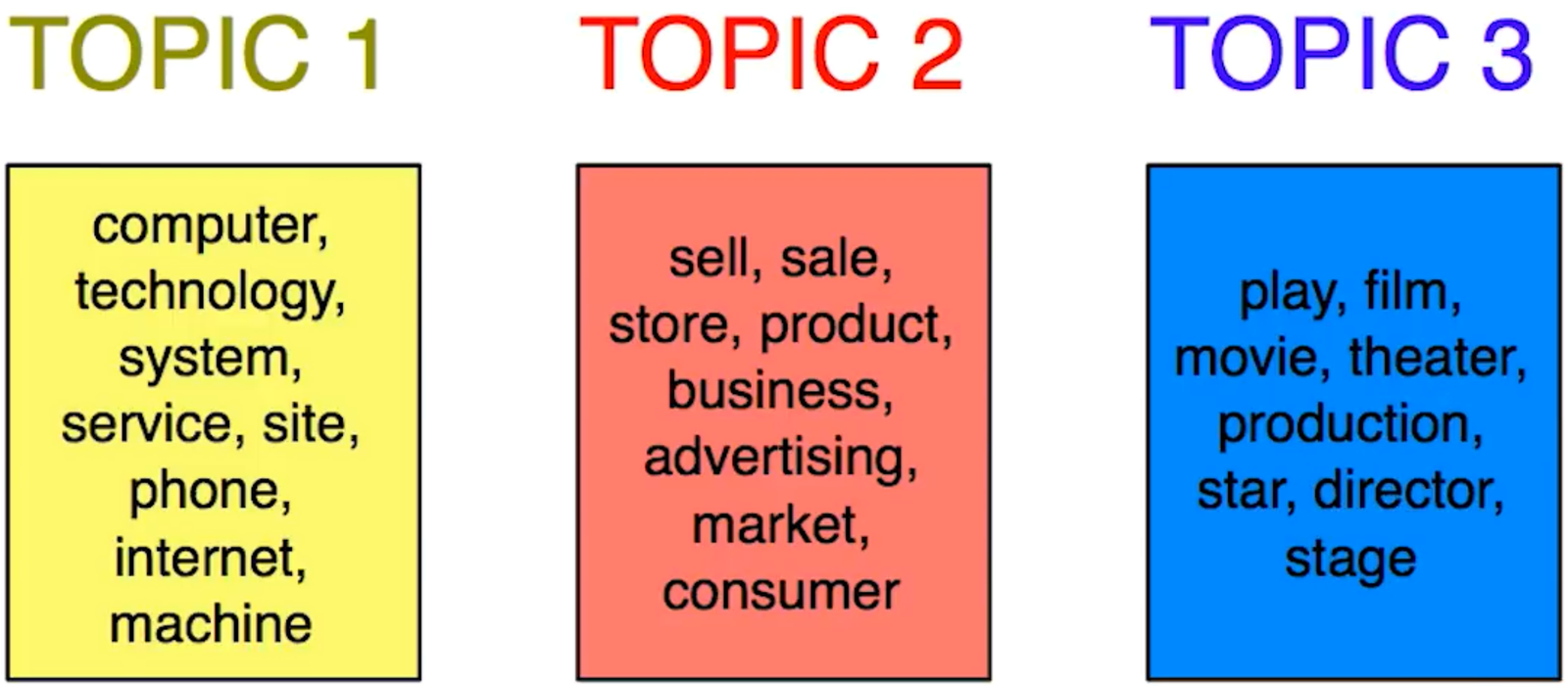


Word Embeddings

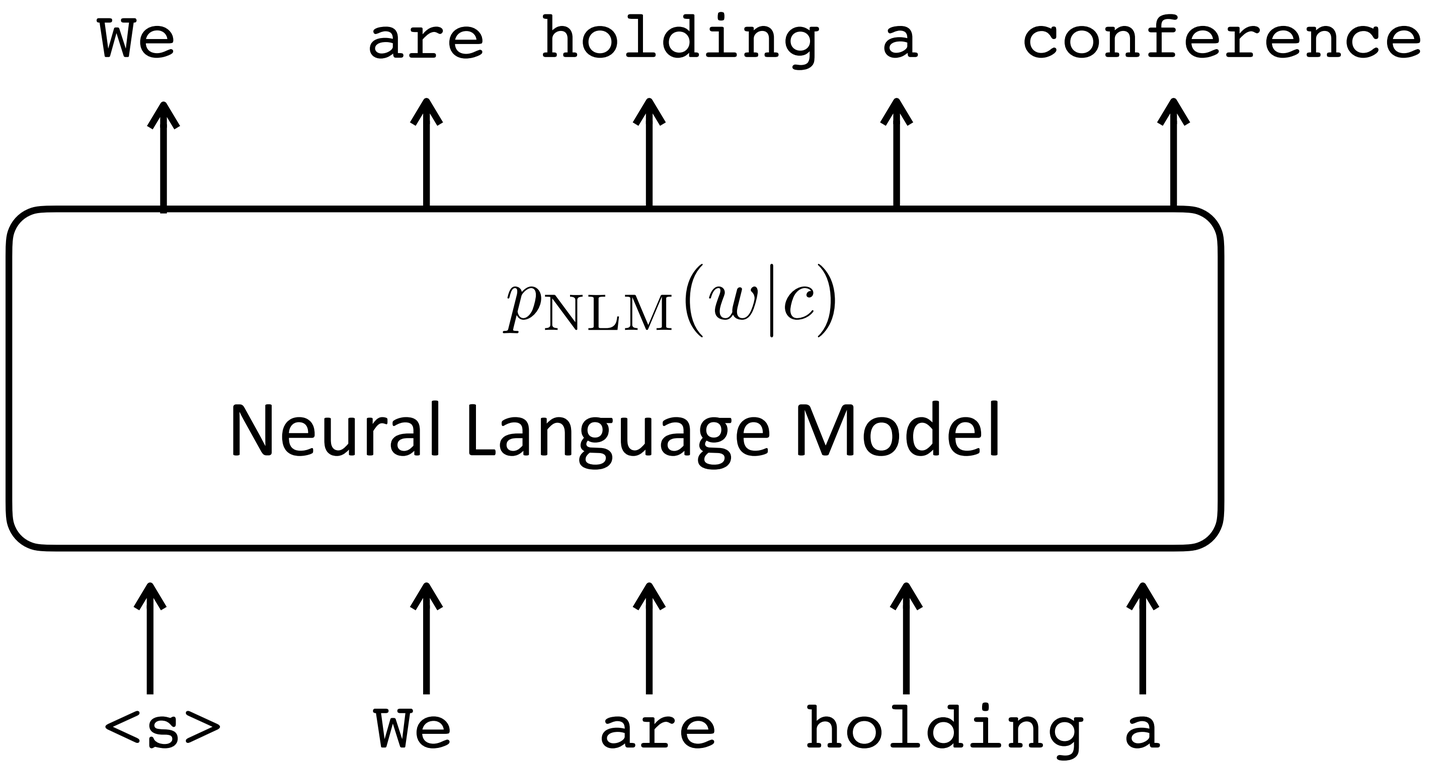


Word2vec [Mikolov et al'13]

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.

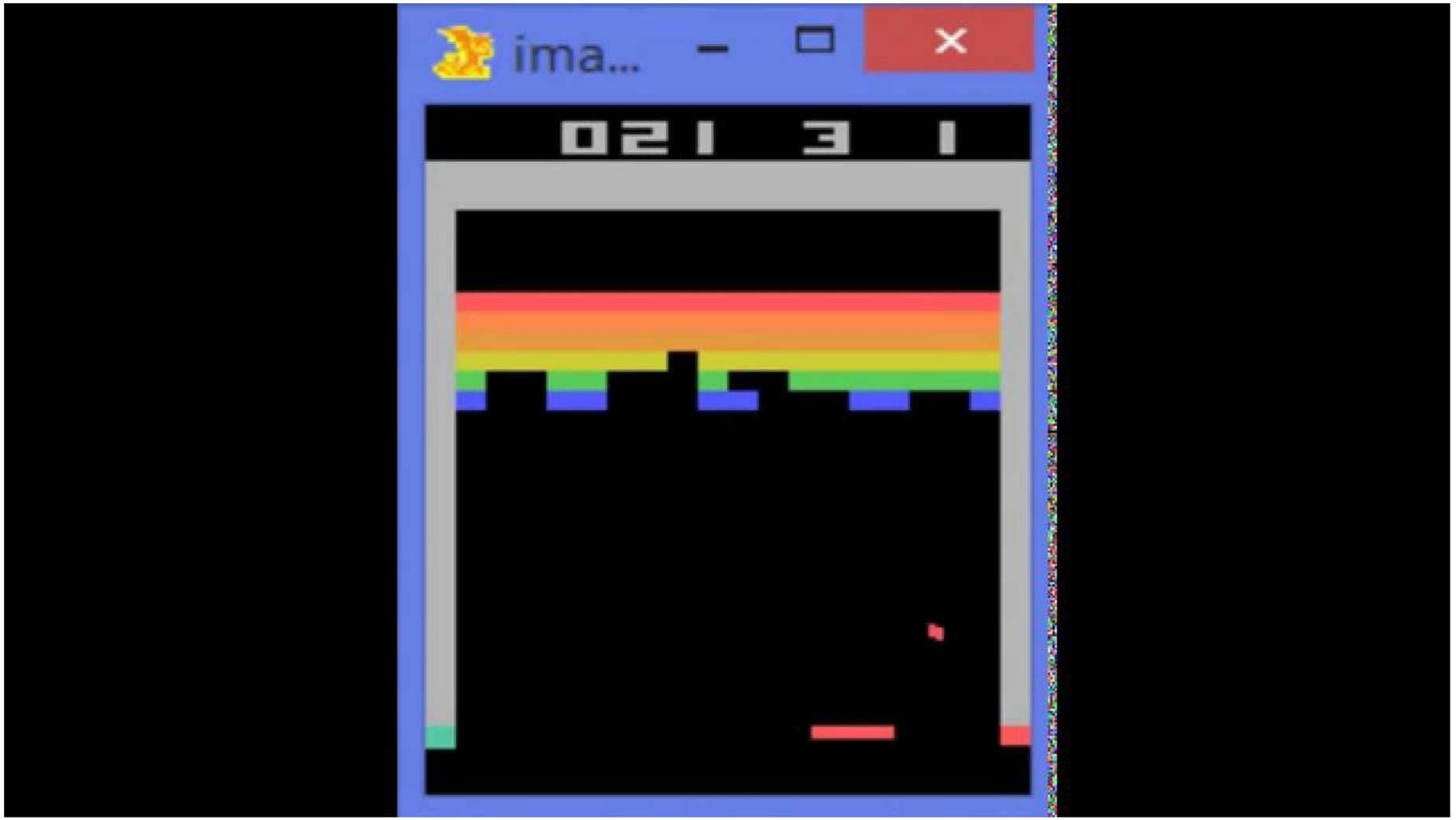
Context →	Q: What is 95 times 45? A:
Target Completion →	4275

Brown et al.'20

Reinforcement Learning

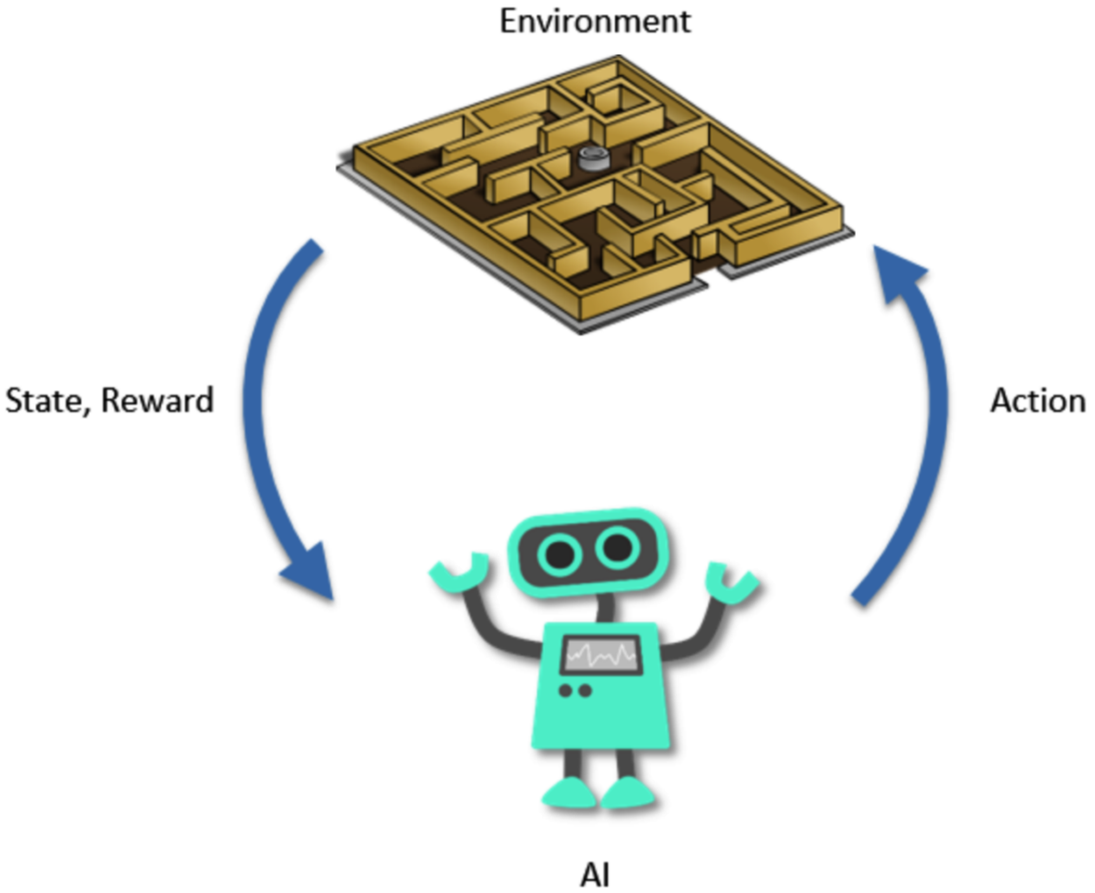
AlphaGo

Atari Breakout Game



Reinforcement Learning

- RL can collect data interactively



Other Topics to Cover

Introduction
Math basics
Supervised learning basics
Logistic regression
Generalized linear models, classification
Kernel methods
SVM
Naive Bayes
MLE, MAP
Gradient descent, SGD, Newton's method
Generalization, bias-variance tradeoff
Clustering
Expectation Maximization
PCA/ICA
mid-term exam
Probabilistic Graphical Models
HMM

Neural Networks, backprop
Neural Networks, architectures
Neural architectures
Variational autoencoder
Generative adversarial networks
Reinforcement Learning
Labor day
Languge models
Pretraining
Large language models

Thank You!
Questions?