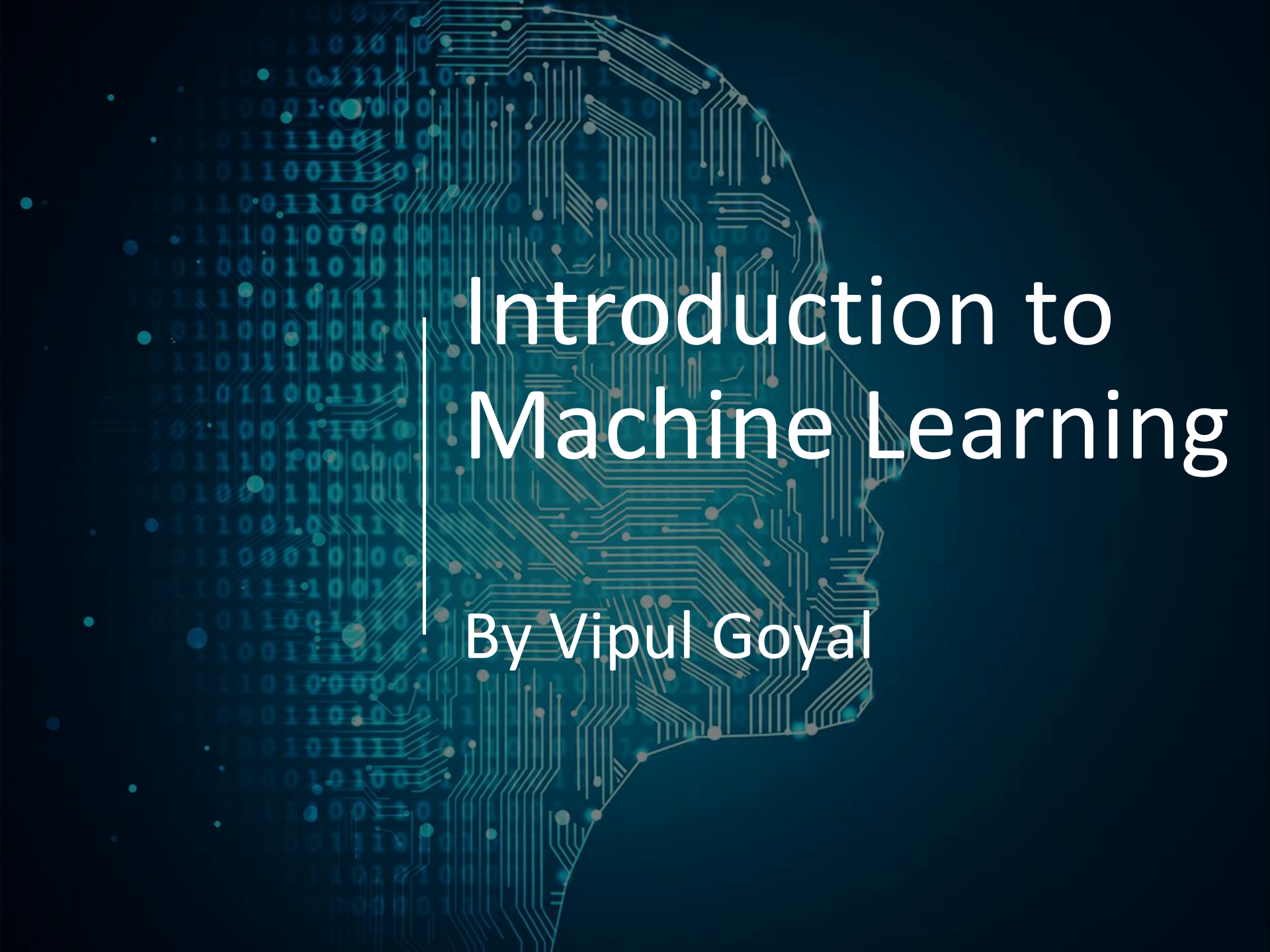

Introduction to
Machine Learning
By Vipul Goyal

Some Basics

My Name: Vipul Goyal

Home page: http://www.cs.cmu.edu/~goyal/

My research interests:

- Cryptography, cybersecurity, blockchains
- Machine learning/AI
  - More specifically: privacy preserving machine learning

About this Course

- Basic first course on Machine Learning with a focus on privacy aspects and game playing

- Target focus: beginning college or high school students
  - Will cover most background material needed

- What this course is NOT:
  - Advanced machine learning course

Course Basics

- Lecture sessions every week
- Mute your mic during the lectures
- Slides will be provided immediately after the lecture
- If you have any questions during the lectures, please type them in the chat window
- We will take periodic breaks and I will answer the typed questions and we can have a longer discussion

Big Data

- Widespread use of personal computers and wireless communication leads to “big data”
- We are both producers and consumers of data
- Data is not random, it has structure, e.g., customer behavior
- We need “big theory” to extract that structure from data for:

  (a) Understanding the process

  (b) Making predictions for the future

Machine Learning

- Machine learning is programming computers to optimize a performance criterion using example data or past experience.
- There is no need to “learn” to calculate payroll
- Learning is used when:
  - Human expertise does not exist (navigating on Mars)
  - Humans are unable to explain their expertise (face recognition, speech recognition)
  - Solution changes in time (self-driving cars)

Machine Learning is Everywhere

- Web Search
- Network Routing
- Speech Recognition
- Spam Filters
- Machine Translation .....

Definition of Machine Learning

Arthur Samuel (1959): Machine Learning is the field of study that gives the computer the ability to learn without being explicitly programmed.

A. L. Samuel*

Some Studies in Machine Learning
Using the Game of Checkers. II—Recent Progress

Definition of Machine Learning

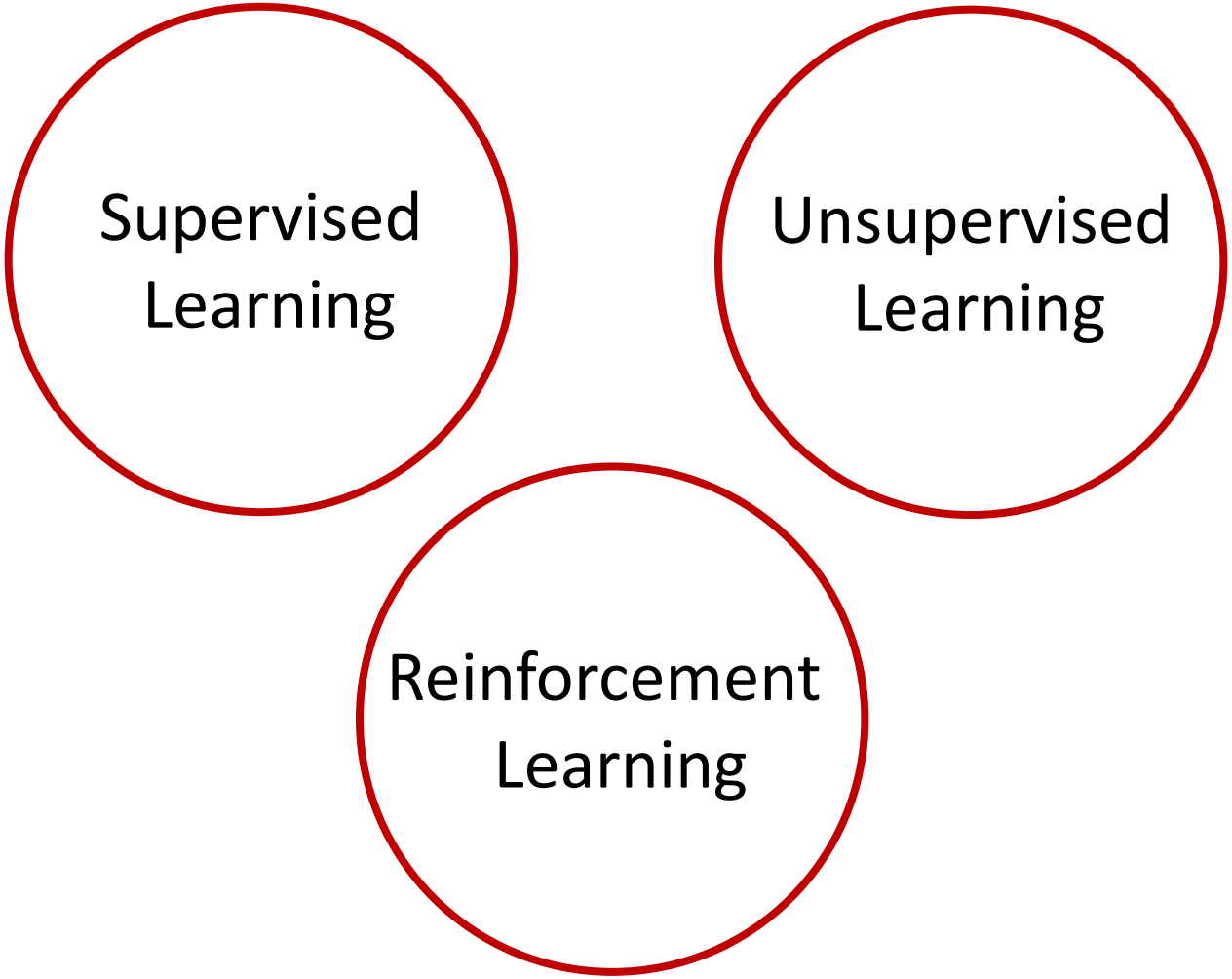
Tom Mitchell (1998): a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

Experience (data): games played by the program (with itself or others)

Performance measure: winning rate

Types of Machine Learning

(A Simplified View)



Types of Learning

- Supervised Learning
  - Classification
  - Regression
- Unsupervised Learning
- Reinforcement Learning

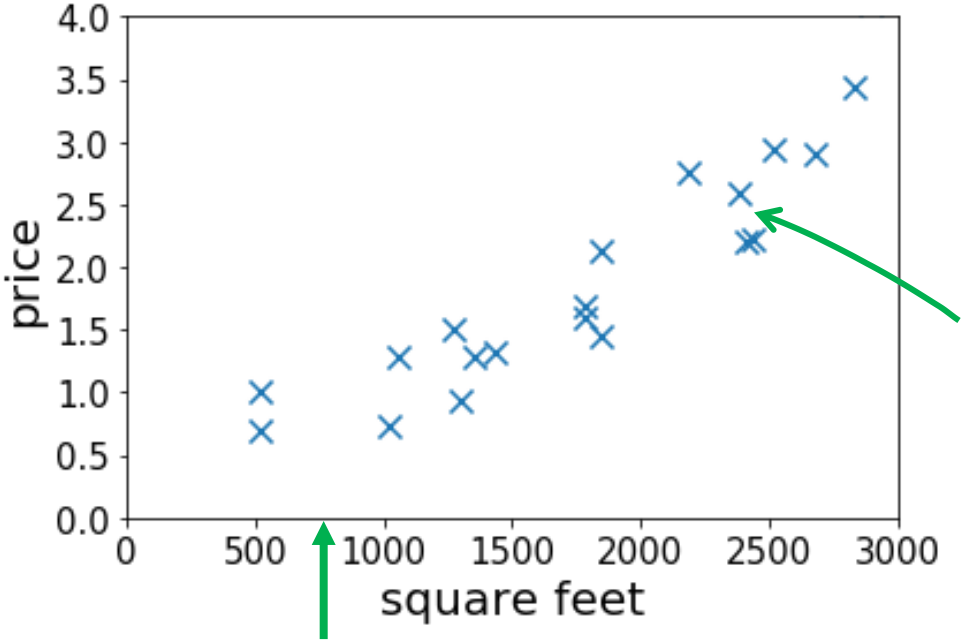
Supervised Learning

Housing Price Prediction

- Given: a dataset that contains m samples

$$\left(x^{(1)}, y^{(1)}\right), \ldots \left(x^{(m)}, y^{(m)}\right)$$

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

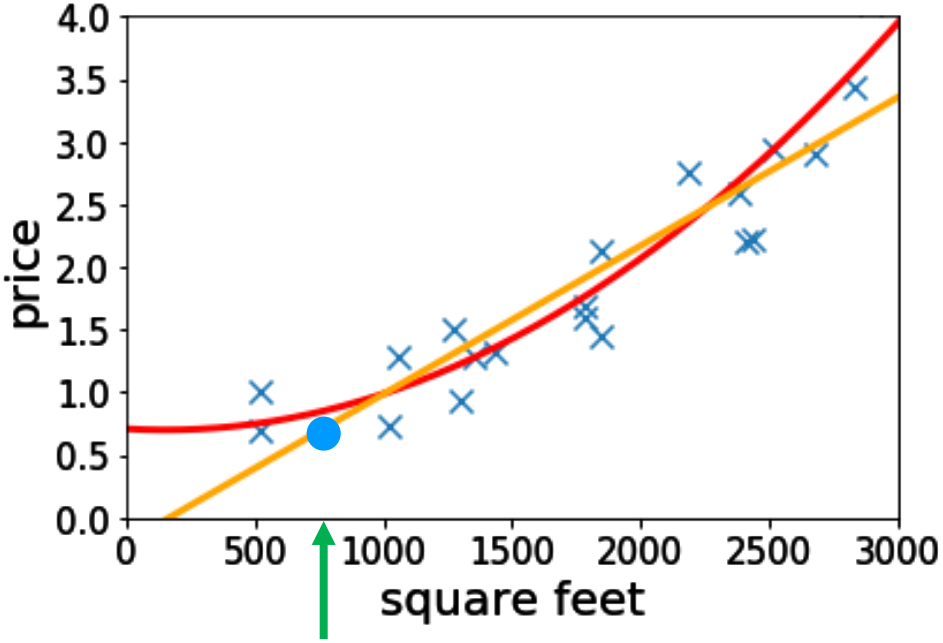


Housing Price Prediction

- Given: a dataset that contains m samples

$$\left(x^{(1)}, y^{(1)}\right), \ldots \left(x^{(m)}, y^{(m)}\right)$$

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



$x = 800$

$y = ?$

More Features

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

(size, lot size) → price

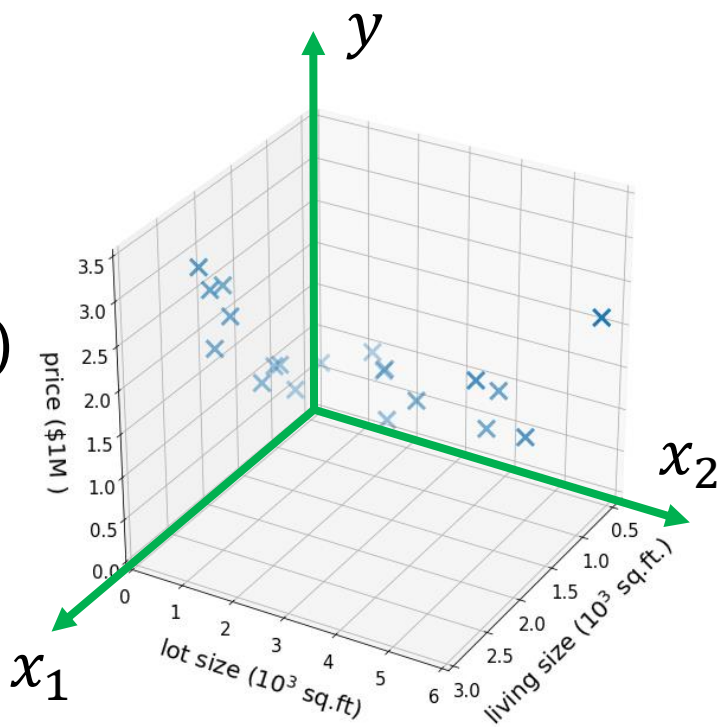
features/input $x \in \mathbb{R}^2$

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

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

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

- "Supervision" refers to $y^{(1)}, \ldots, y^{(m)}$



High-dimensional Features

- $x \in \mathbb{R}^d$ for large d

- E.g.,

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

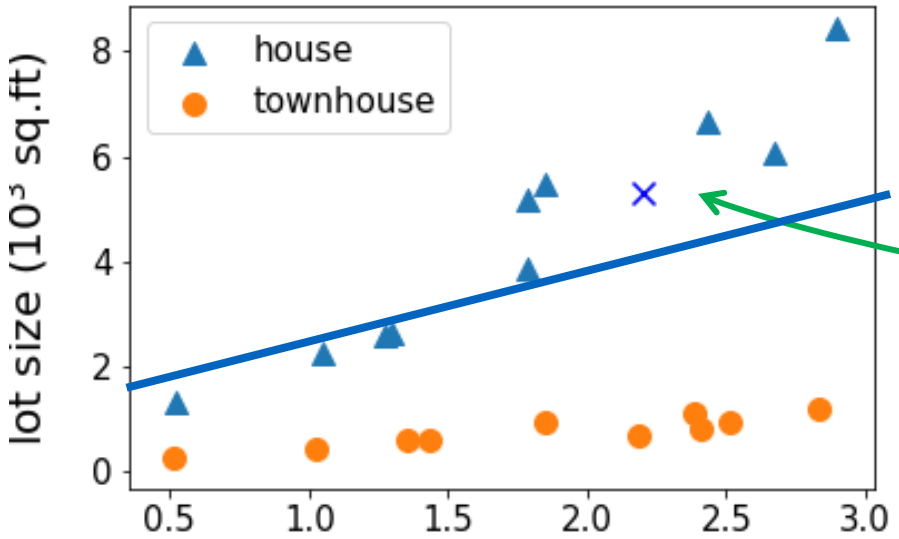
--- living size
--- lot size
--- year built
--- condition
--- zip code
$\vdots$

$\rightarrow$ y --- price

Regression vs Classification

- regression: if $y \in \mathbb{R}$ is a continuous variable, e.g., price prediction
- classification: the label y is a discrete variable
- e.g., the task of predicting the types of residence

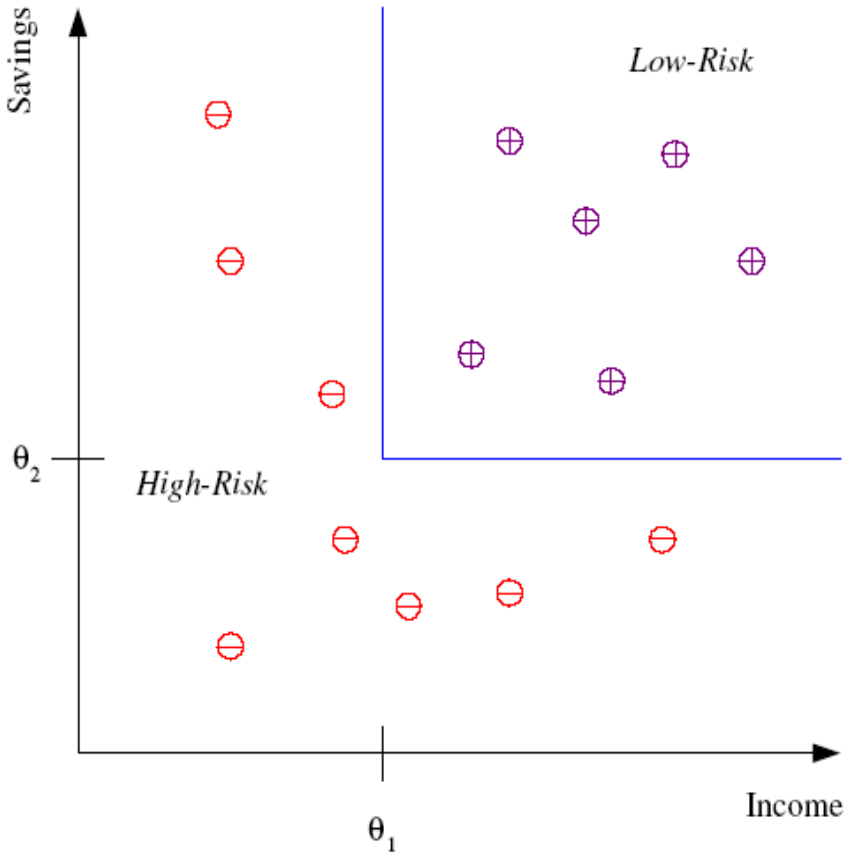
(size, lot size) → house or townhouse?



y = house or townhouse?

Classification Examples

- Example: Credit scoring
- Differentiating between low-risk and high-risk customers from their *income* and *savings*



Discriminant: IF *income* > θ_1 AND *savings* > θ_2
THEN low-risk ELSE high-risk

Supervised Learning in Computer Vision

- Image Classification

- x = raw pixels of the image, y = the main object

Classification Examples

x



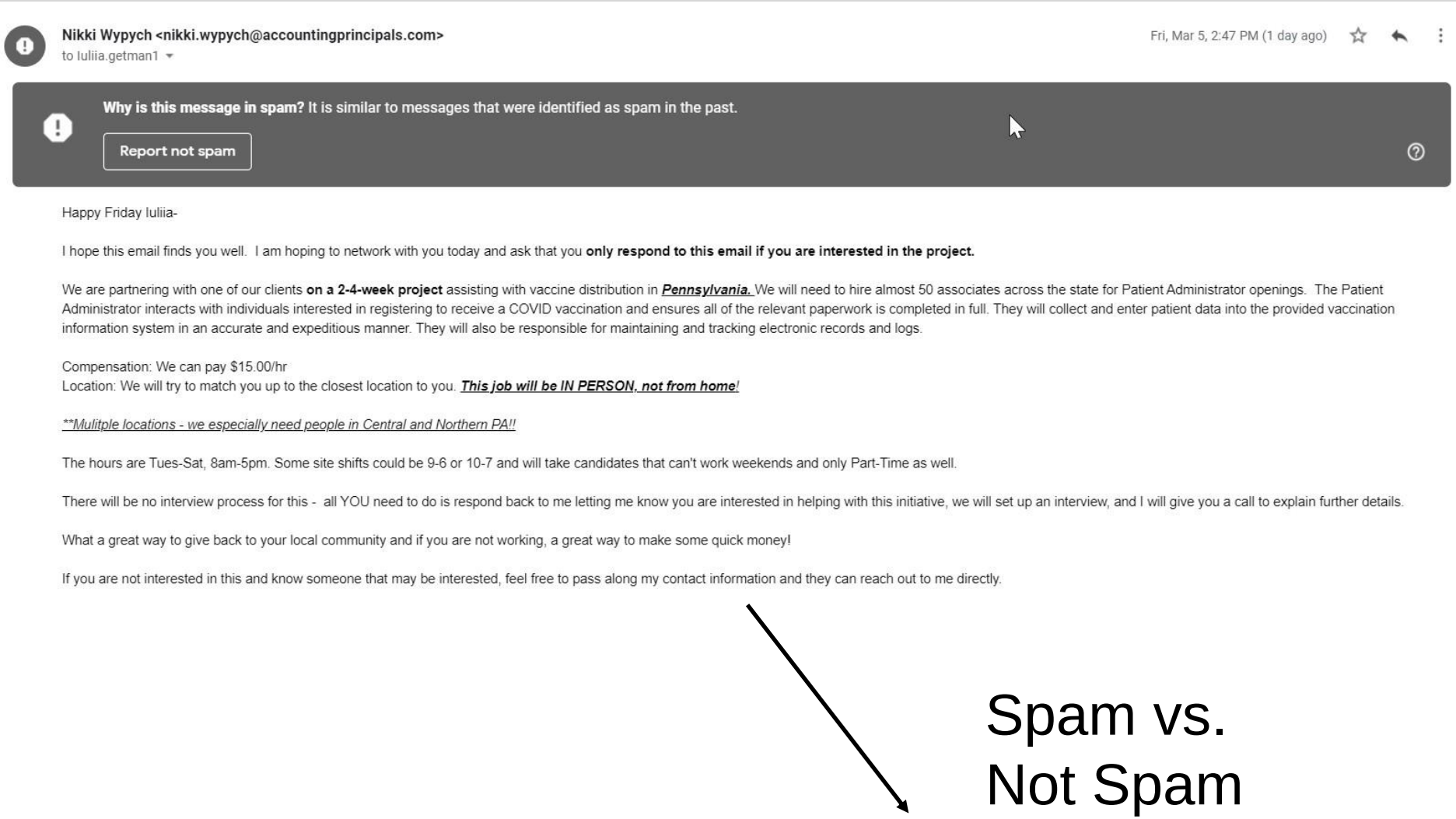



y

giraffe giraffe giraffe llama llama llama

In Humans: Is the given picture a Boy or girl?


Spam Filtering



Nikki Wypych <nikki.wypych@accountingprincipals.com>
to Iuliia.getman1

Fri, Mar 5, 2:47 PM (1 day ago)

Why is this message in spam? It is similar to messages that were identified as spam in the past.

Report not spam

Happy Friday Iuliia-

I hope this email finds you well. I am hoping to network with you today and ask that you **only respond to this email if you are interested in the project.**

We are partnering with one of our clients **on a 2-4-week project** assisting with vaccine distribution in ***Pennsylvania.*** We will need to hire almost 50 associates across the state for Patient Administrator openings. The Patient Administrator interacts with individuals interested in registering to receive a COVID vaccination and ensures all of the relevant paperwork is completed in full. They will collect and enter patient data into the provided vaccination information system in an accurate and expeditious manner. They will also be responsible for maintaining and tracking electronic records and logs.

Compensation: We can pay $15.00/hr
Location: We will try to match you up to the closest location to you. ***This job will be IN PERSON, not from home!***

***Mulitple locations - we especially need people in Central and Northern PA!!*

The hours are Tues-Sat, 8am-5pm. Some site shifts could be 9-6 or 10-7 and will take candidates that can't work weekends and only Part-Time as well.

There will be no interview process for this - all YOU need to do is respond back to me letting me know you are interested in helping with this initiative, we will set up an interview, and I will give you a call to explain further details.

What a great way to give back to your local community and if you are not working, a great way to make some quick money!

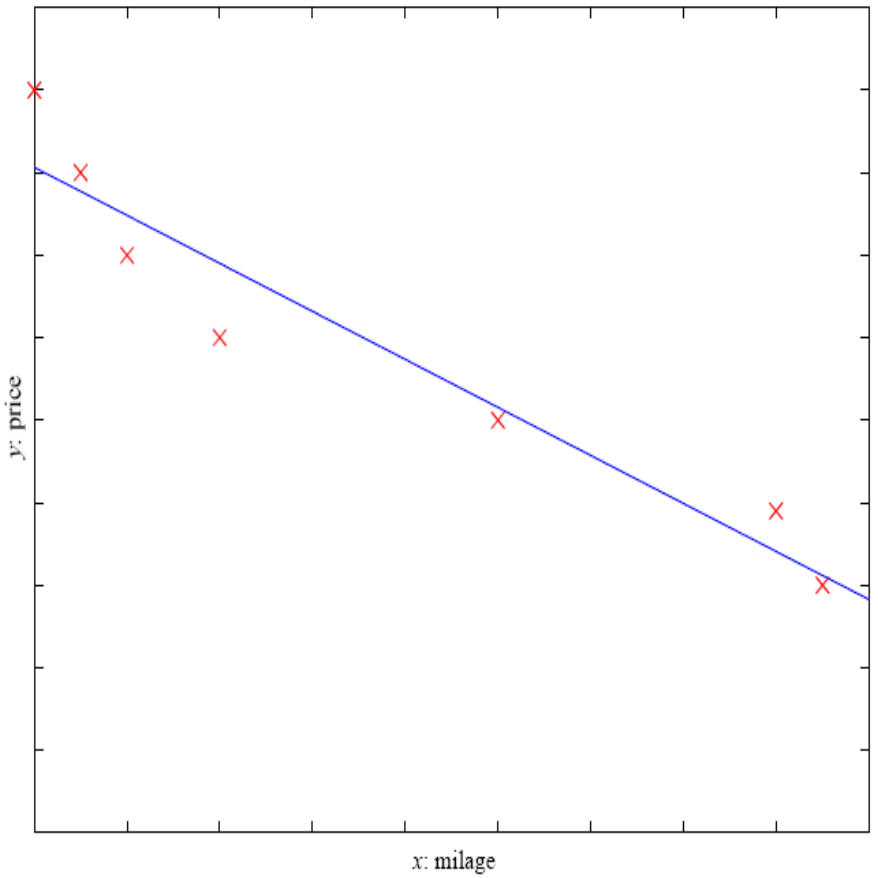
If you are not interested in this and know someone that may be interested, feel free to pass along my contact information and they can reach out to me directly.

Spam vs.
Not Spam

Regression Examples

- Example: Price of a used car
- *x* : car attributes

  y : price



Regression Example

Self Driving car: Compute angle of the steering wheel

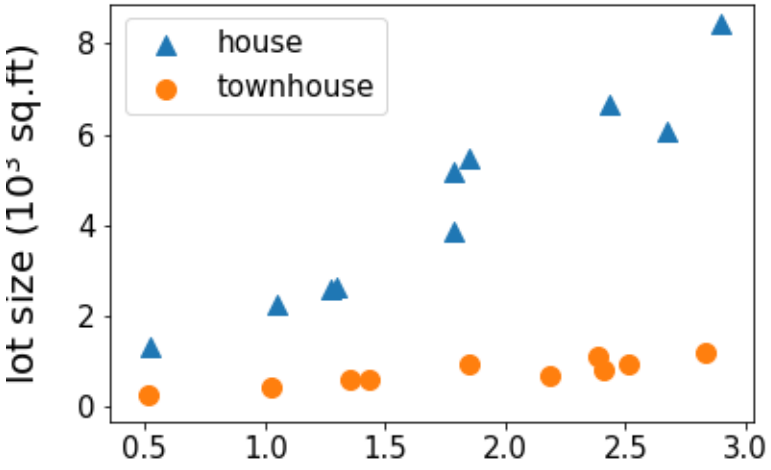
Stock Market Prediction

Unsupervised Learning

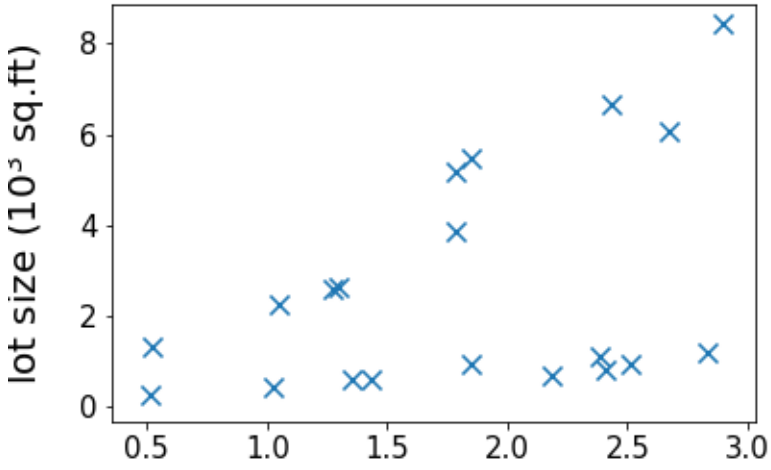
Unsupervised Learning

- Dataset contains no labels: $x^{(1)}, \ldots x^{(m)}$
- Goal (vaguely-posed): to find interesting structures in the data

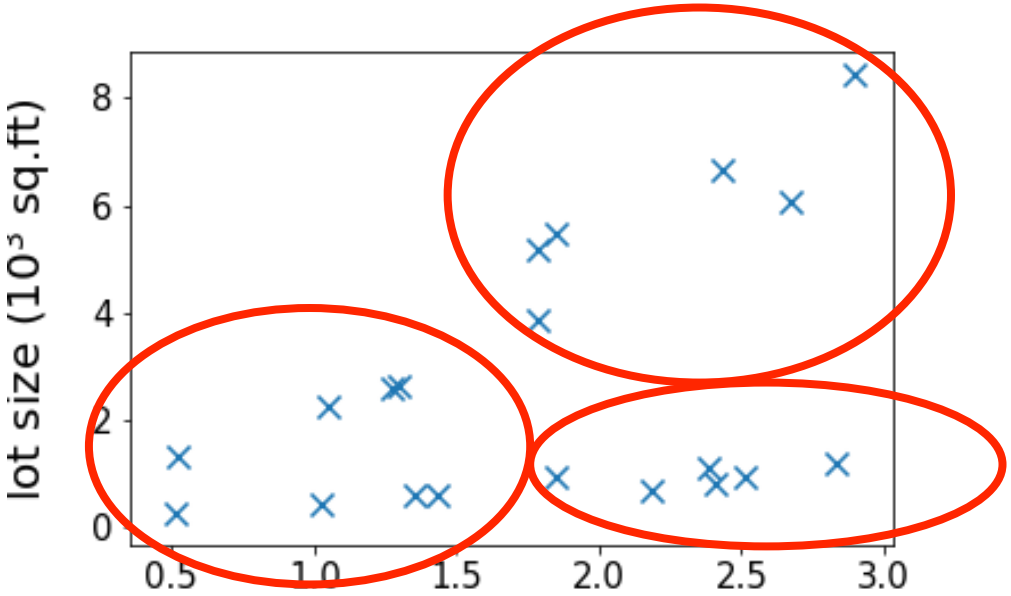
supervised



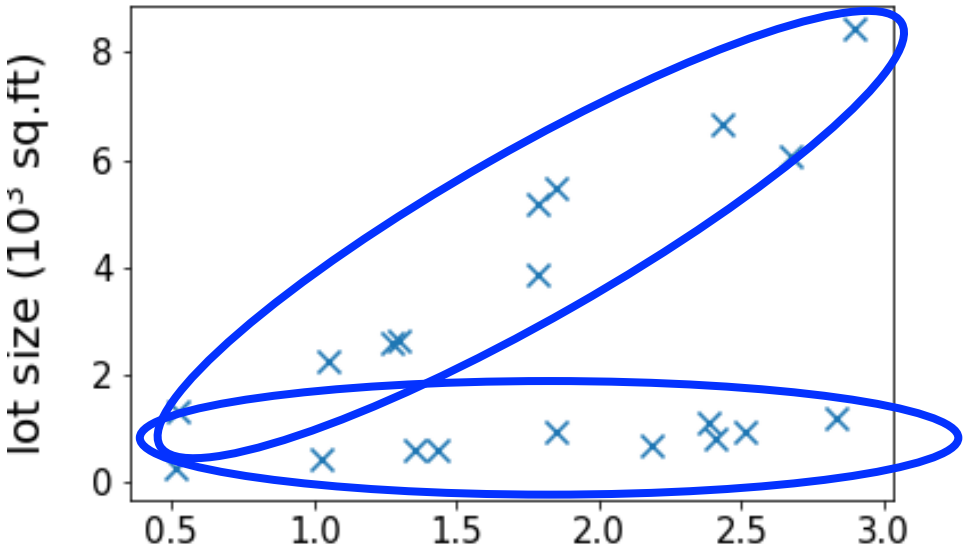
unsupervised



Clustering



Clustering

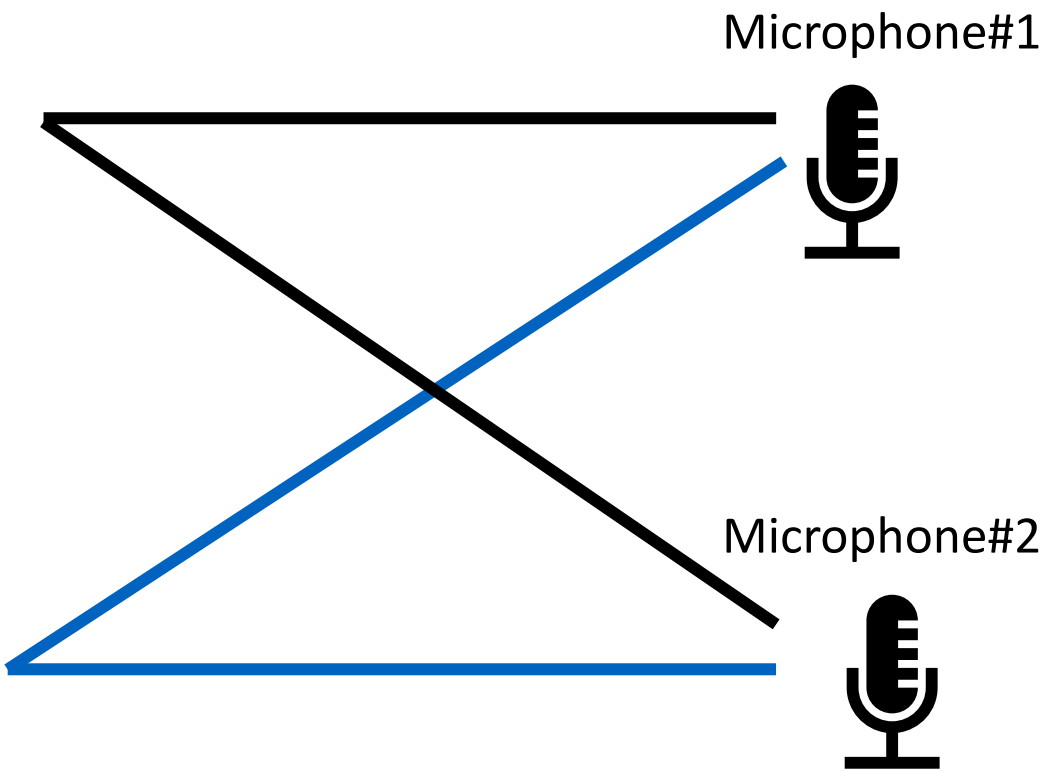


Cocktail Party Problem

- Many people talking at the same time

Speaker #1

Speaker #2



The two speeches will be shifted in M1 and M2

Google News

Headlines

More Headlines

COVID-19 news: See the latest coverage of the coronavirus (COVID-19)

Here's How the Senate Pared Back Biden's Stimulus Plan

The New York Times · 1 hour ago

- Third stimulus check is in COVID relief bill | USA TODAY
  USA TODAY · 1 hour ago
- Senate passes $1.9 trillion Covid relief bill, including $1,400 stimulus checks, with no Republican support
  NBC News · 6 hours ago
- Biden's historic victory for America -- no thanks to GOP
  CNN · 4 hours ago · Opinion
- Senate Democrats eke out 50-49 COVID-19 relief bill victory
  The Week · 5 hours ago

View Full Coverage

Amanda Gorman says she was "tailed" by security guard on her way home: "This is the reality of black girls"

CBS News · 7 hours ago

- Amanda Gorman, inaugural poet, 'tailed' by security guard on her walk home
  CNN · 8 hours ago

View Full Coverage

Jacksonville

Rain
55° F

Today	Sun	Mon	Tue	Wed
57°F 47°F	61°F 45°F	62°F 47°F	67°F 53°F	72°F 56°F

C | F | K

More on weather.com

Fact check

Did Kyrsten Sinema Bring Cake to the Senate and Vote Against Raising Minimum Wage?
Snopes.com

Fact Check: Are COVID-Positive Migrants Allowed to Cross Southern Border Into US?
Newsweek

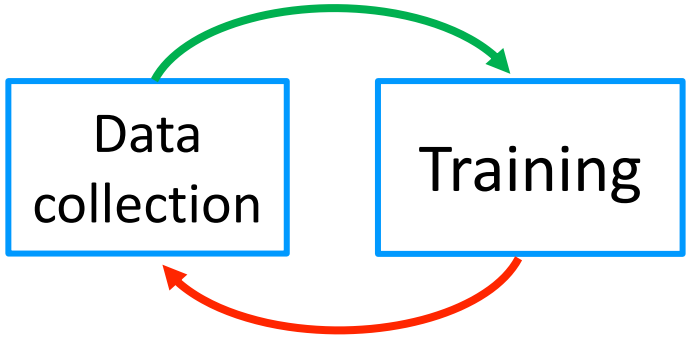
Fitzgerald overstates claim on pork in COVID-19 relief bill
PolitiFact

Reinforcement Learning

Reinforcement Learning

- The algorithm can collect data interactively

Try the strategy and collect feedbacks



Improve the strategy based on the feedbacks

Credit Decisions

Credit history

Accept/Deny

- Applicant applies for credit
- Our algorithm gets various pieces of data as input: income, housing status (rent/own), past payment history, education ….
- Algorithm decides yes/no

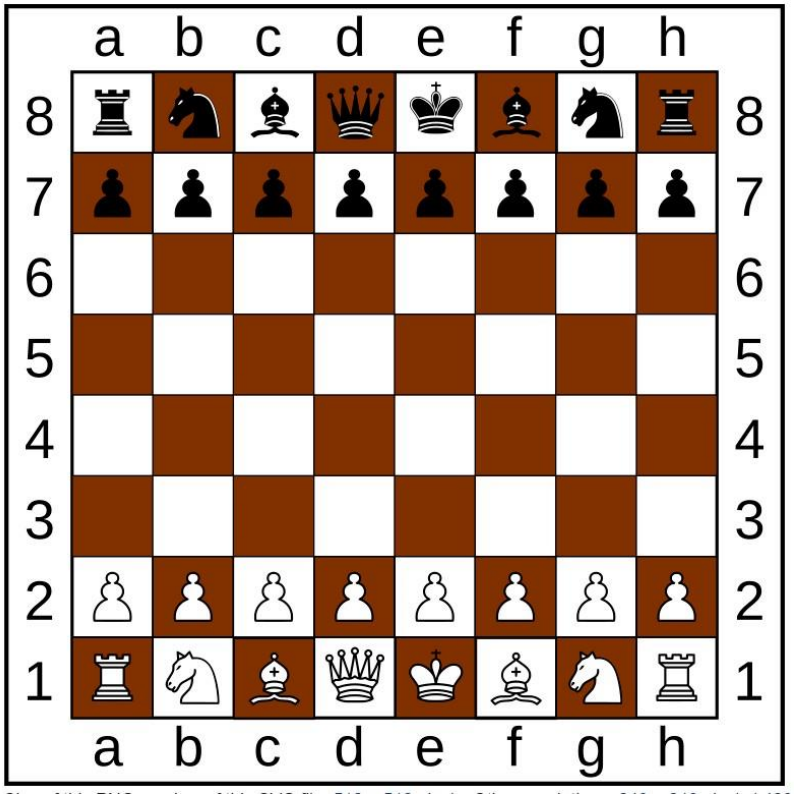
Credit Decisions

Credit history

Accept/Deny

- Feedback given much later (did the candidate fail to pay back on time?)
- Refine algorithm, repeat!

Automated Game Playing



- Start with a strategy: win/lose
- Refine algorithm, repeat!

Questions?