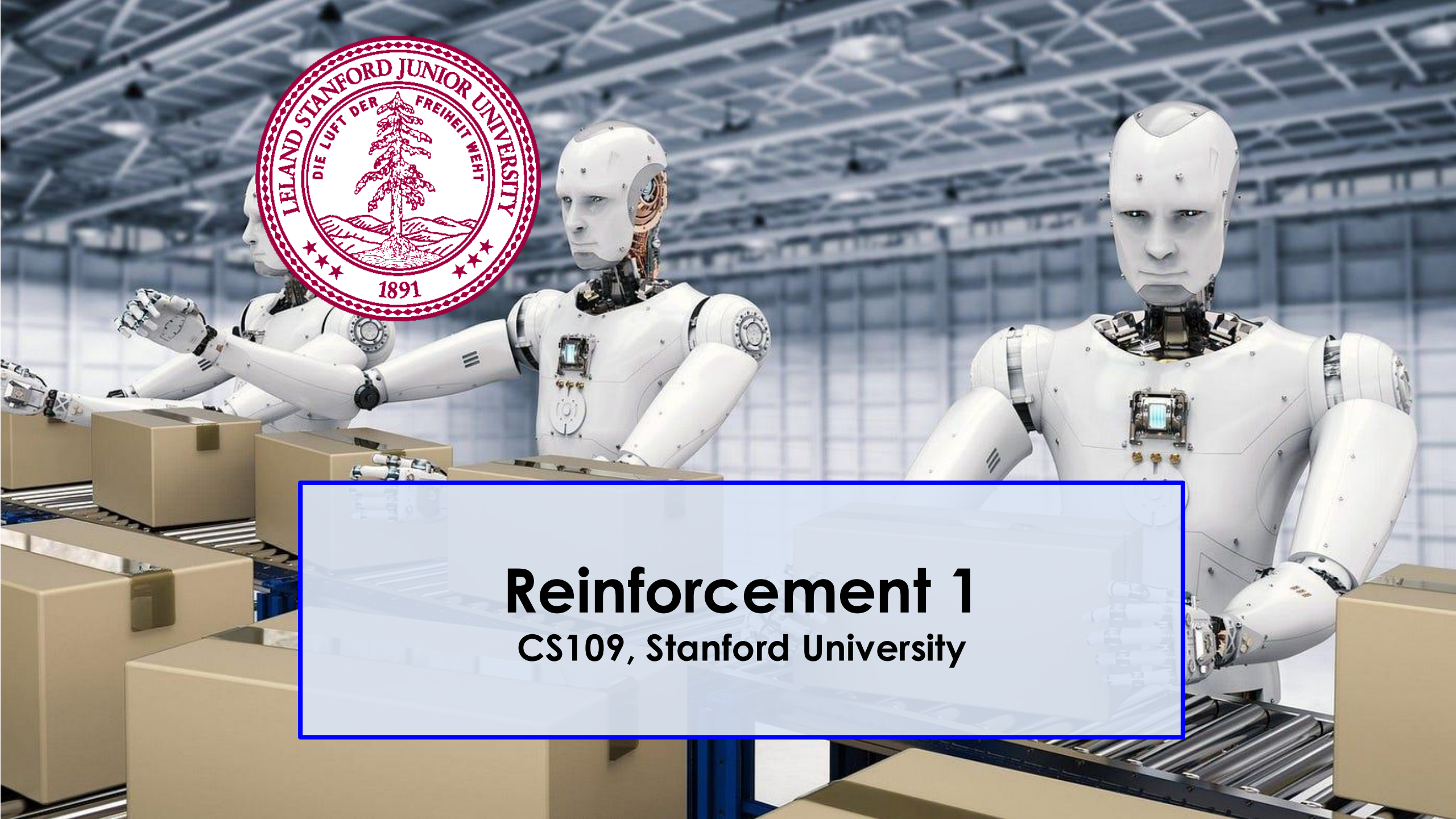
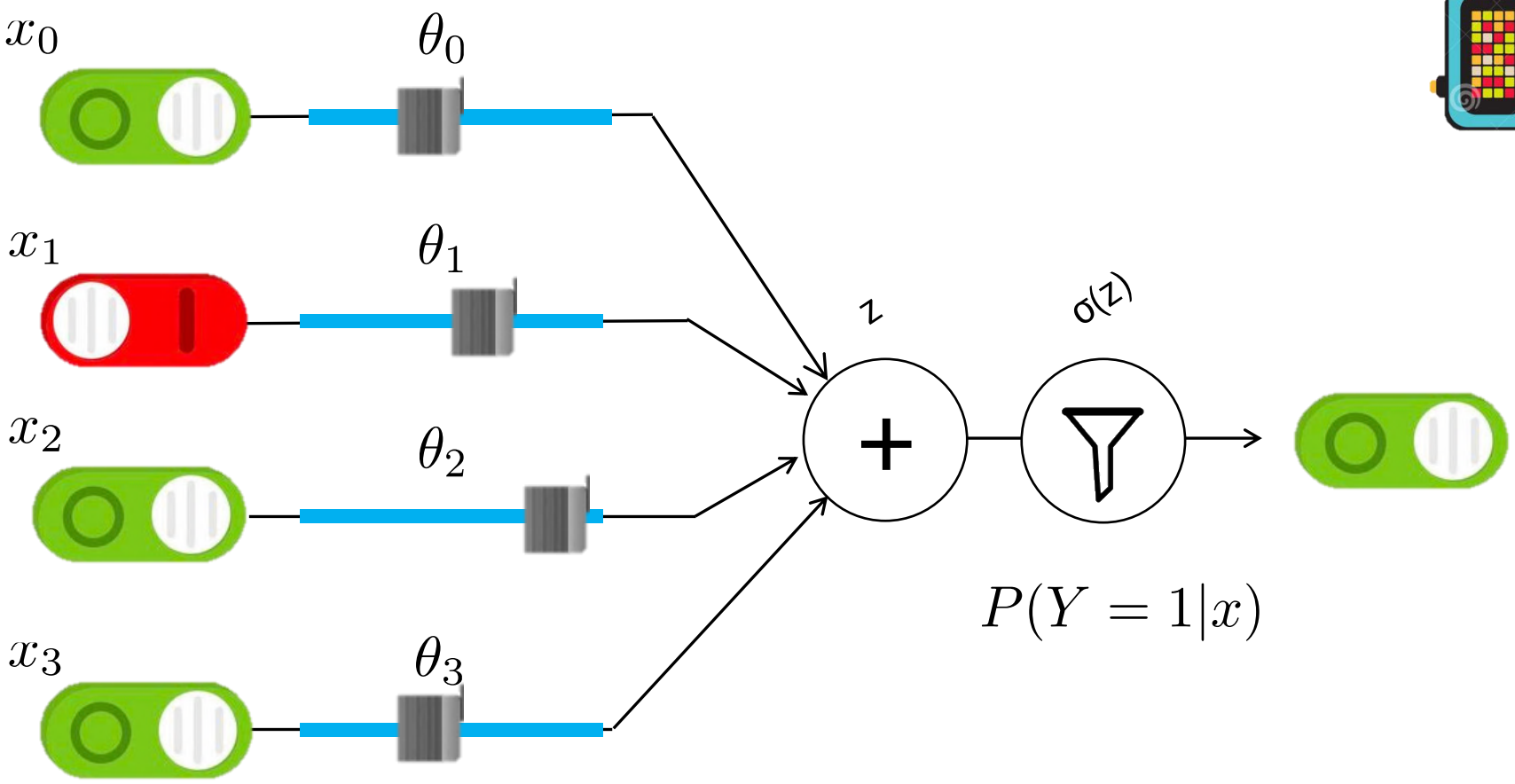


Reinforcement 1

CS109, Stanford University

Review

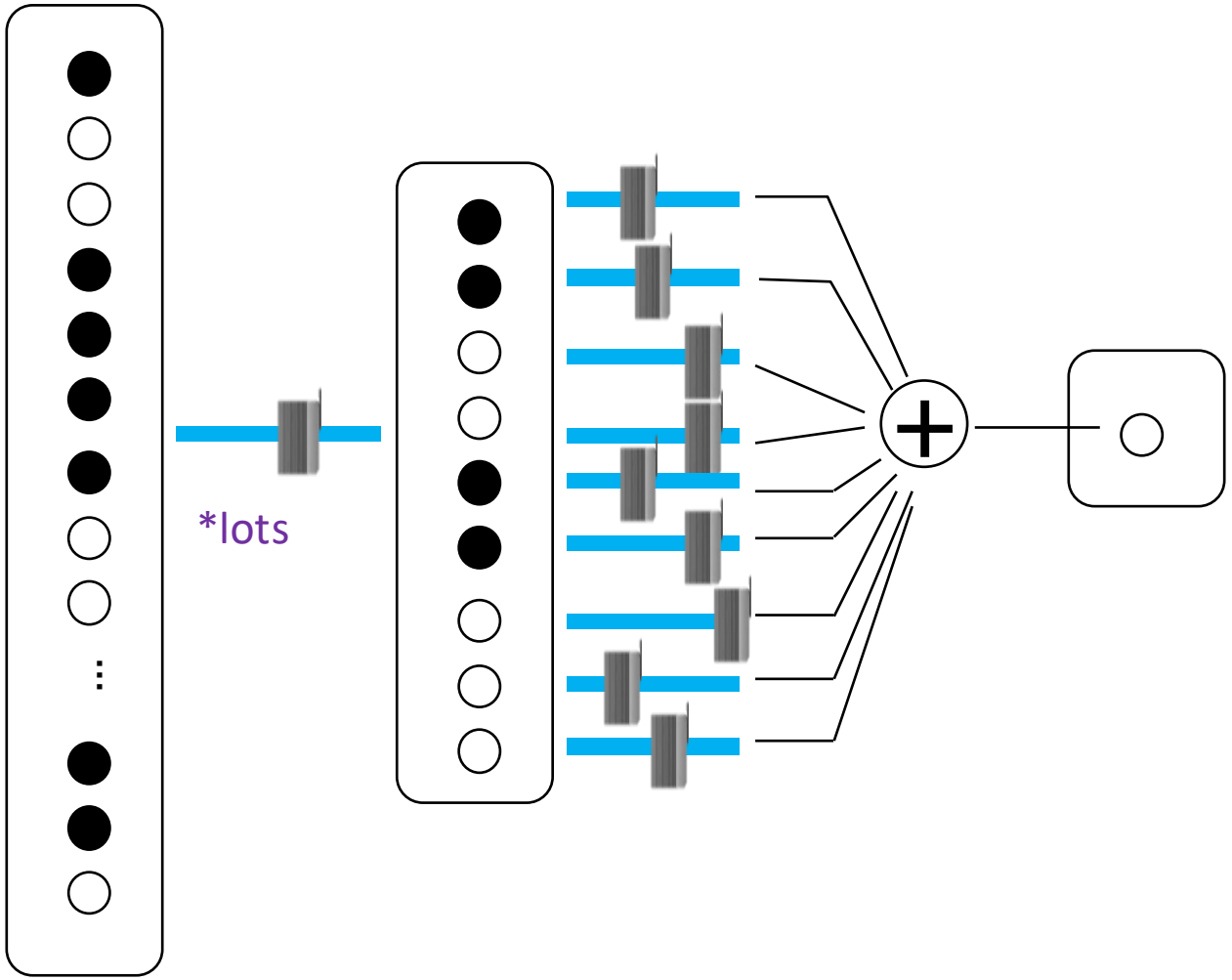
Logistic Regression



$$P(Y = 1|\mathbf{X} = \mathbf{x}) = \sigma\Big(\sum_i \theta_i x_i\Big)$$

Deep Learning

We Can Put Neurons Together



These are the features for one training example.

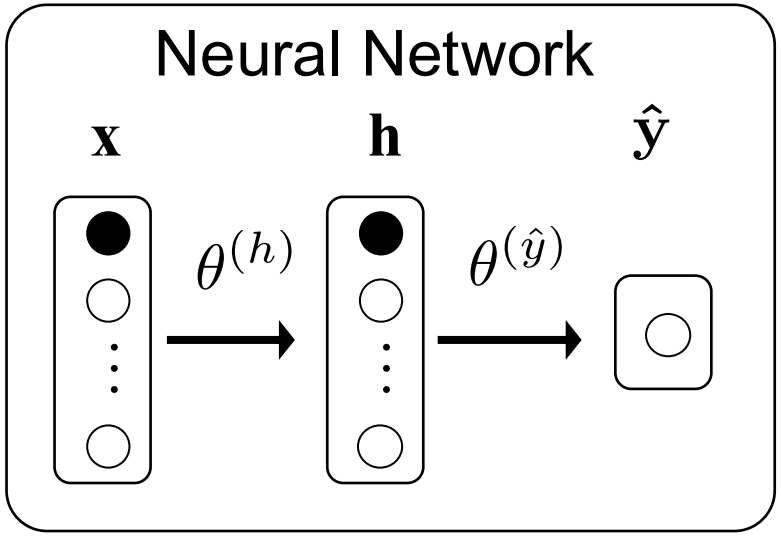
Derivative Goals

Loss with respect to output layer params

$$\frac{\partial LL(\theta)}{\partial \theta_i^{(\hat{y})}}$$

Loss with respect to hidden layer params

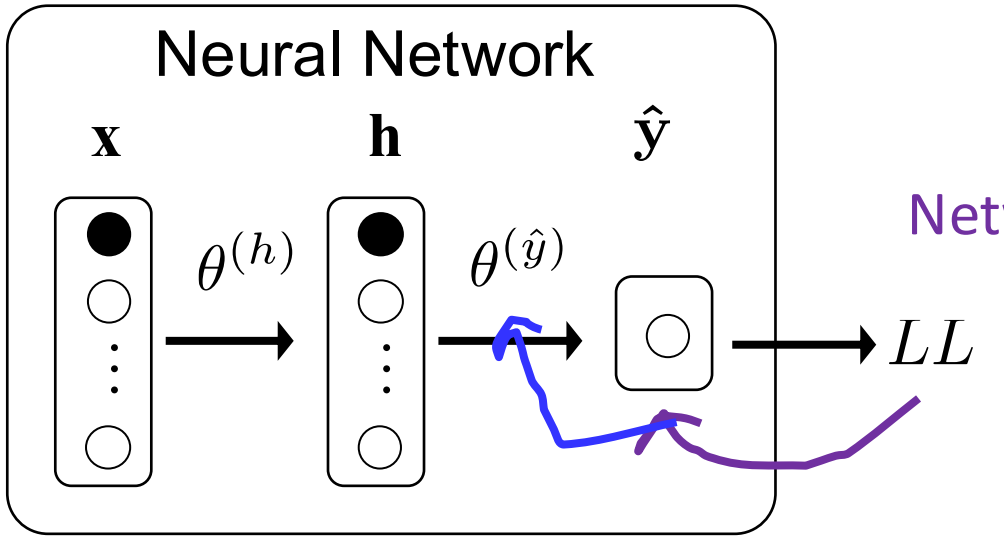
$$\frac{\partial LL(\theta)}{\partial \theta_{i,j}^{(h)}}$$



Chain Rule Example 1

$$\frac{\partial LL(\theta)}{\partial \theta_i^{(\hat{y})}}$$

Goal



Network

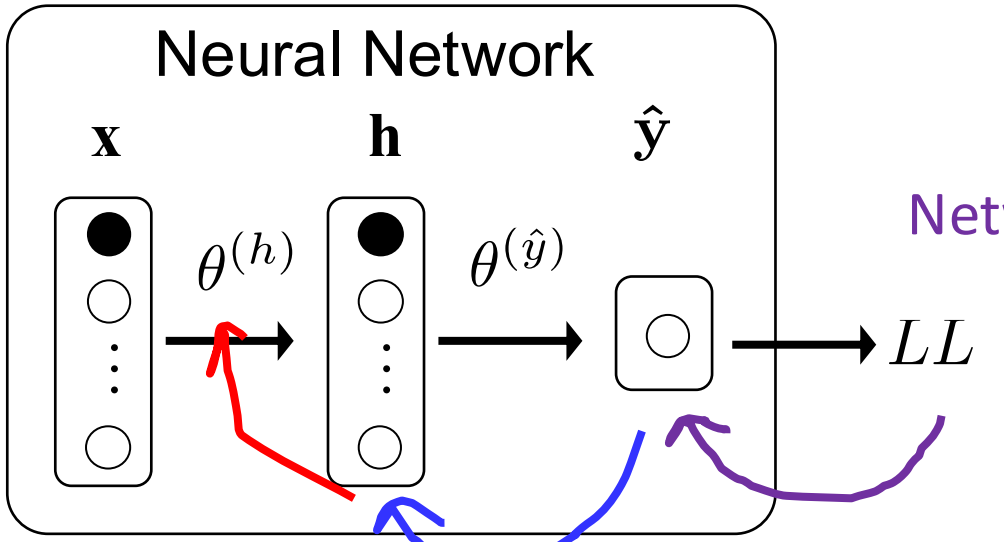
$$\frac{\partial LL(\theta)}{\partial \theta_i^{(\hat{y})}} = \frac{\partial LL}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial \theta_i^{(\hat{y})}}$$

Decomposition

Chain Rule Example 2

$$\frac{\partial LL(\theta)}{\partial \theta_{i,j}^{(h)}}$$

Goal



Network

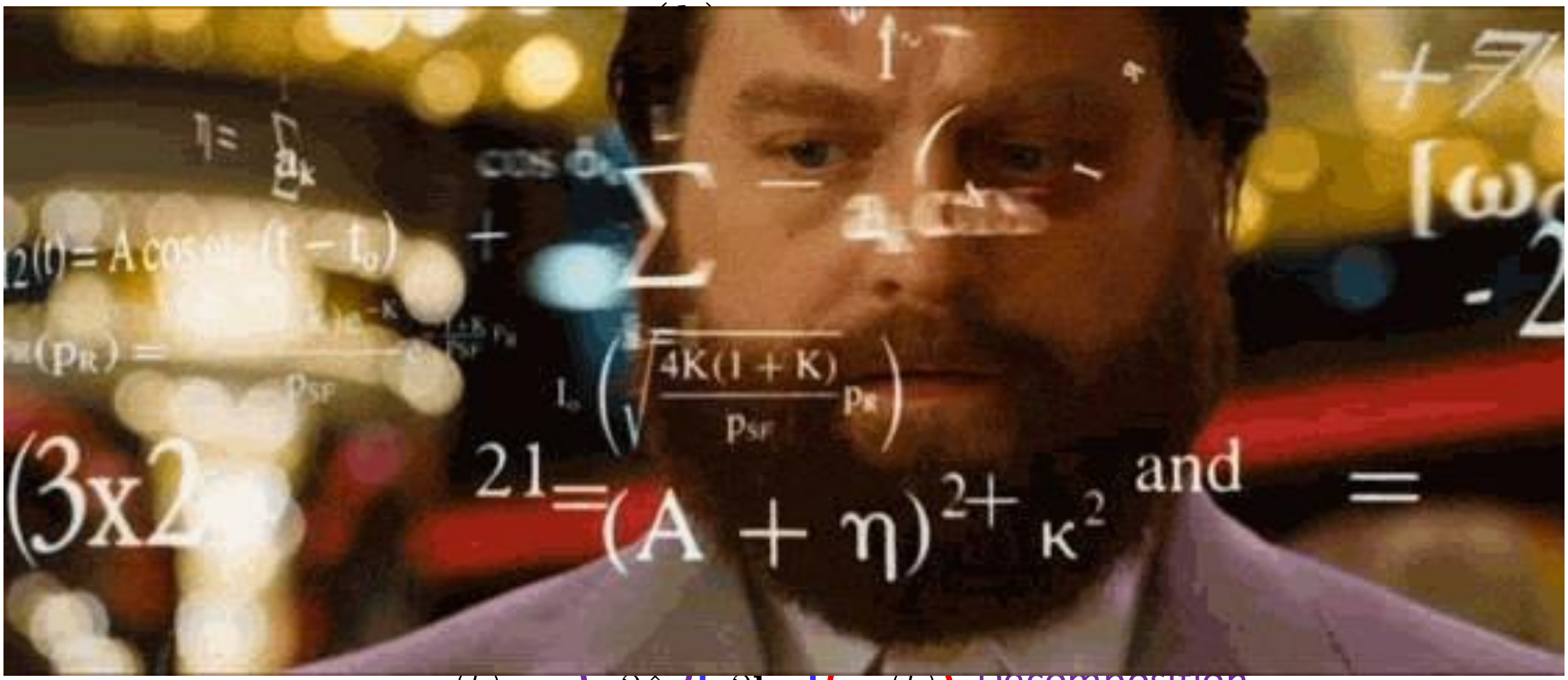
$$\frac{\partial LL(\theta)}{\partial \theta_{i,j}^{(h)}} = \frac{\partial LL}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial \mathbf{h}_j} \cdot \frac{\partial \mathbf{h}_j}{\partial \theta_{i,j}^{(h)}}$$

Decomposition

Chain Rule Example 2

$$\frac{\partial LL(\theta)}{\partial \theta_{i,j}^{(h)}}$$

Goal



$$\frac{\partial \hat{y}}{\partial \mathbf{h}_j} \frac{\partial \mathbf{h}_j}{\partial \theta_{i,j}^{(h)}}$$

Decomposition

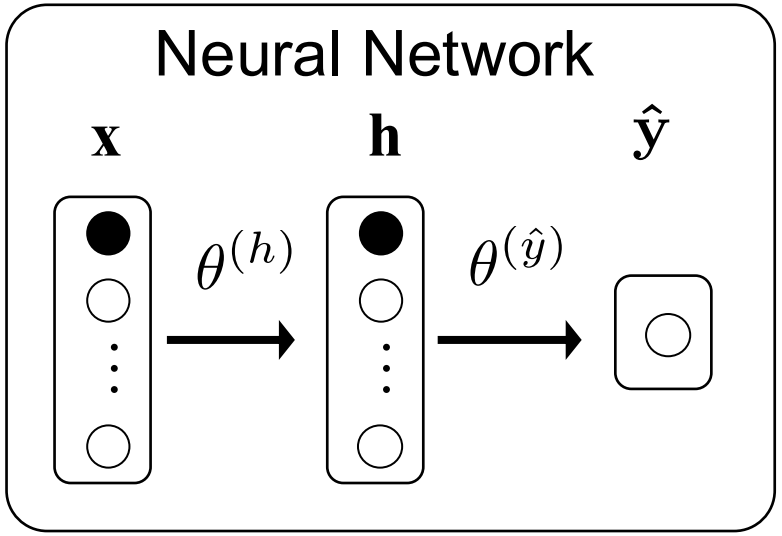
Deep Learning

Loss with respect to output layer params

$$\frac{\partial LL(\theta)}{\partial \theta_i^{(\hat{y})}}$$

Loss with respect to hidden layer params

$$\frac{\partial LL(\theta)}{\partial \theta_{i,j}^{(h)}}$$



Make it Simple

$$\frac{\partial LL(\theta)}{\partial \theta_i^{(\hat{y})}} = \text{  } \cdot \text{  }$$

$$\text{  } = \frac{y}{\hat{y}} - \frac{(1-y)}{(1-\hat{y})}$$

$$\text{  } = \hat{y}[1-\hat{y}] \cdot h_i$$

Make it Simple

$$\frac{\partial LL(\theta)}{\partial \theta_{i,j}^{(h)}} = \text{  }$$

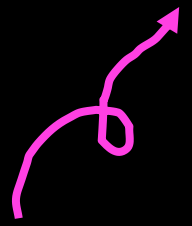
$$\text{  } = \frac{y}{\hat{y}} - \frac{(1-y)}{(1-\hat{y})}$$

$$\text{  } = \hat{y}[1-\hat{y}]\theta_j^{(\hat{y})}$$

$$\text{  } = \mathbf{h}_j[1-\mathbf{h}_j]\mathbf{x}_j$$

You will see this math in many future courses!

You will see this math in many future courses!



Backpropagation

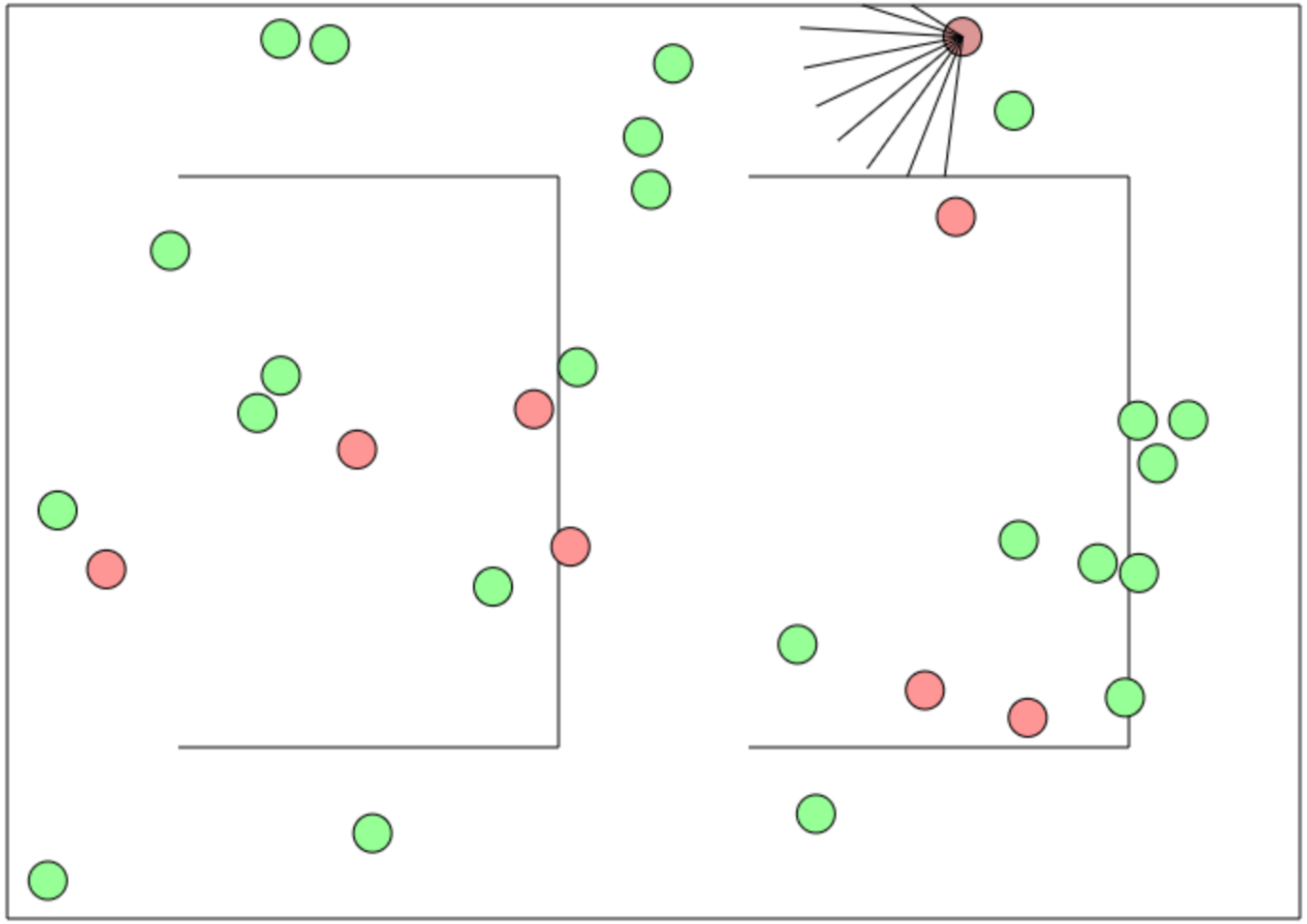
Seeing and working through deep learning derivatives will help you better understand derivative for logistic regression.

Need to know for CS109

But wait….

Is all of ML Classification?

Lets start training a Critter



http://cs.stanford.edu/people/karpathy/convnetjs/demo/rldemo.html

Types of Machine Learning Tasks

Muli-Class Classification

Regression

Reinforcement Learning

Generation

Types of Machine Learning Tasks

Muli-Class Classification

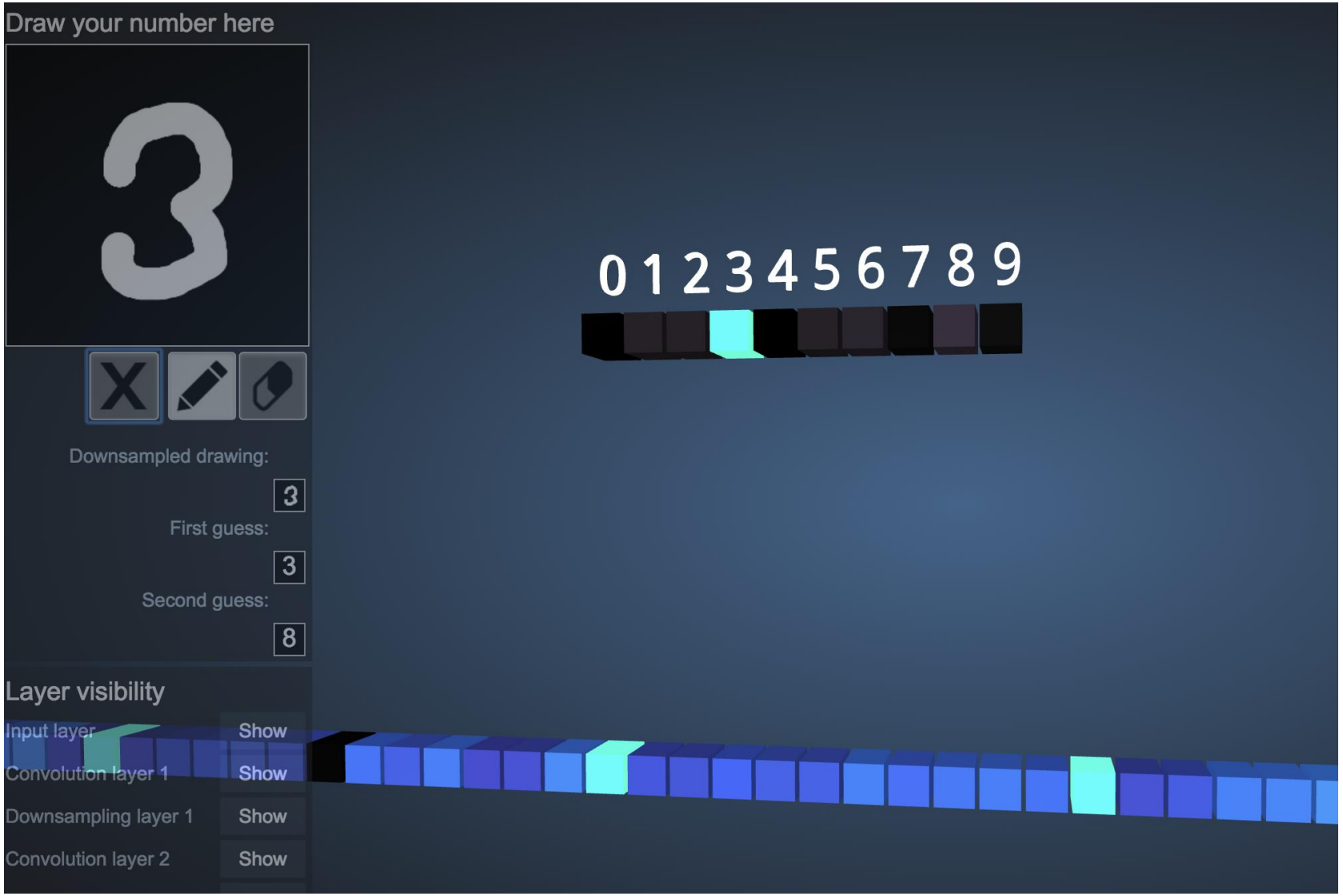
Regression

Reinforcement Learning

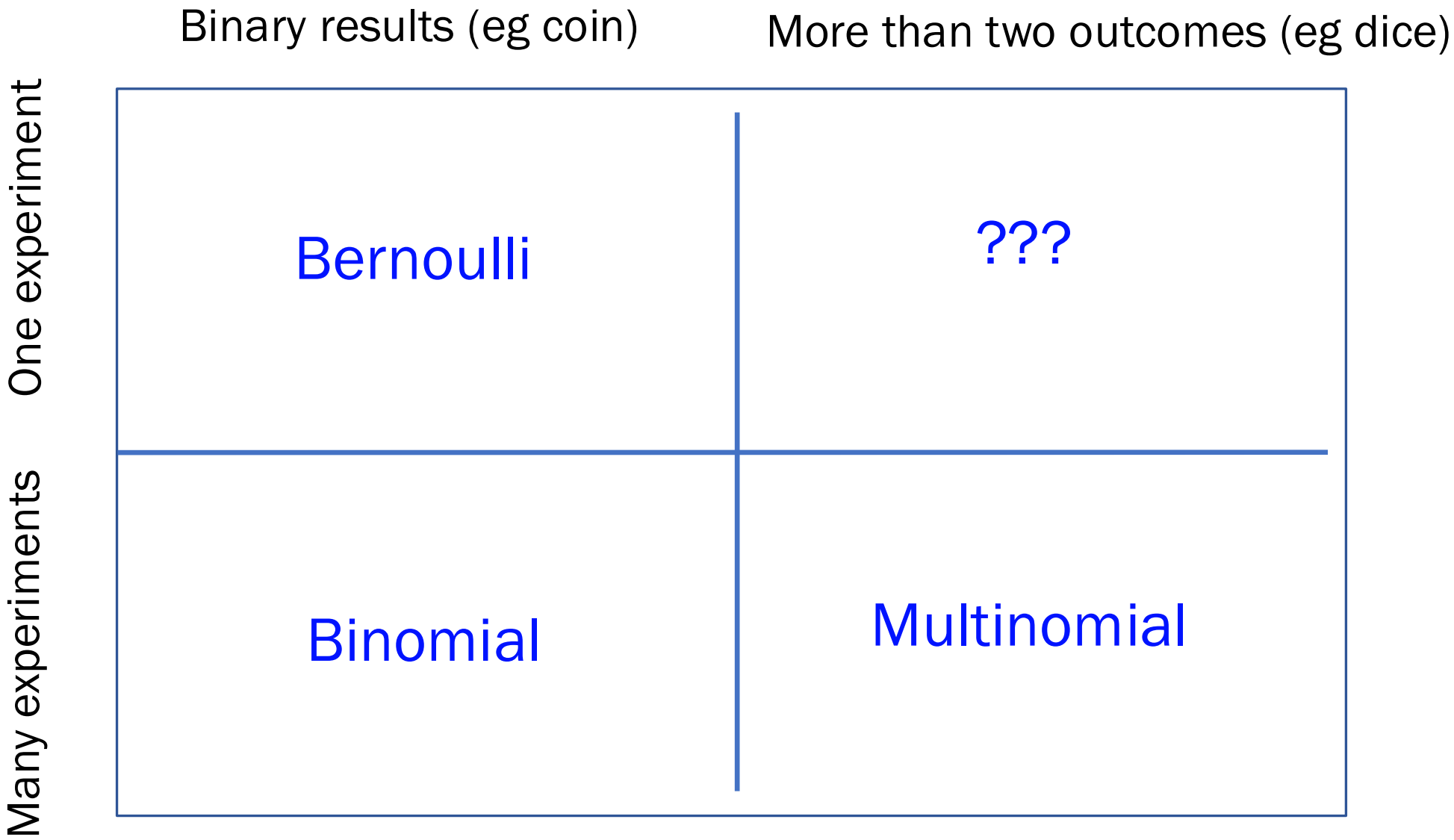
Generation

Beyond Binary Classification

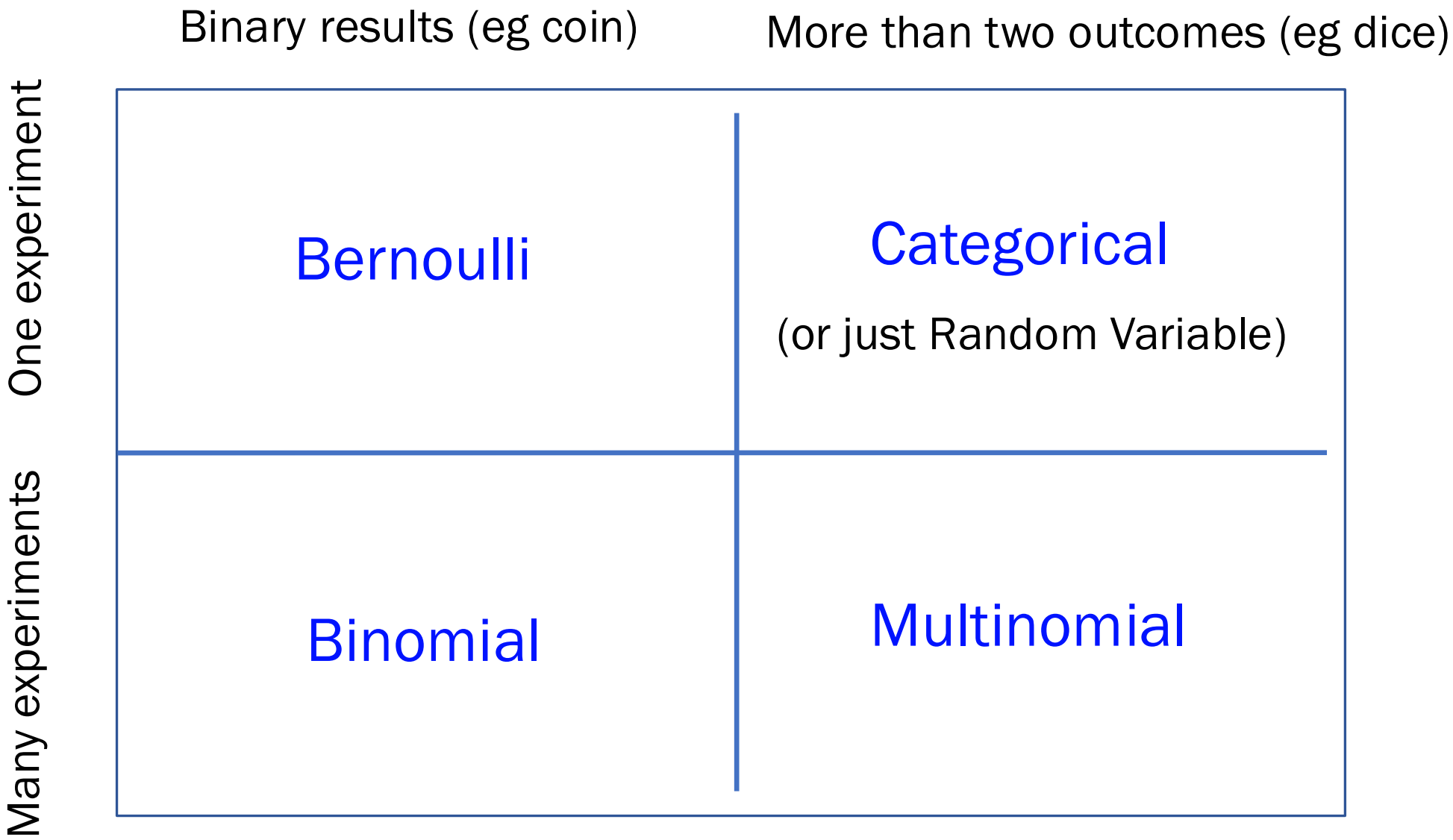
Multiple Outputs



The Categorical



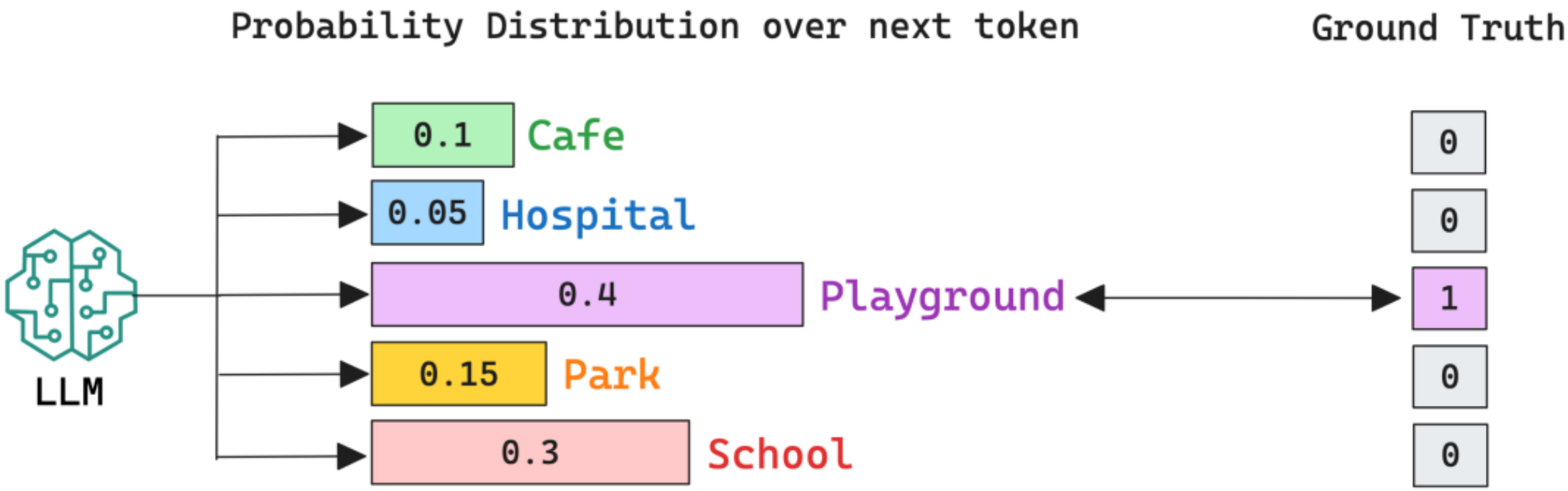
The Categorical



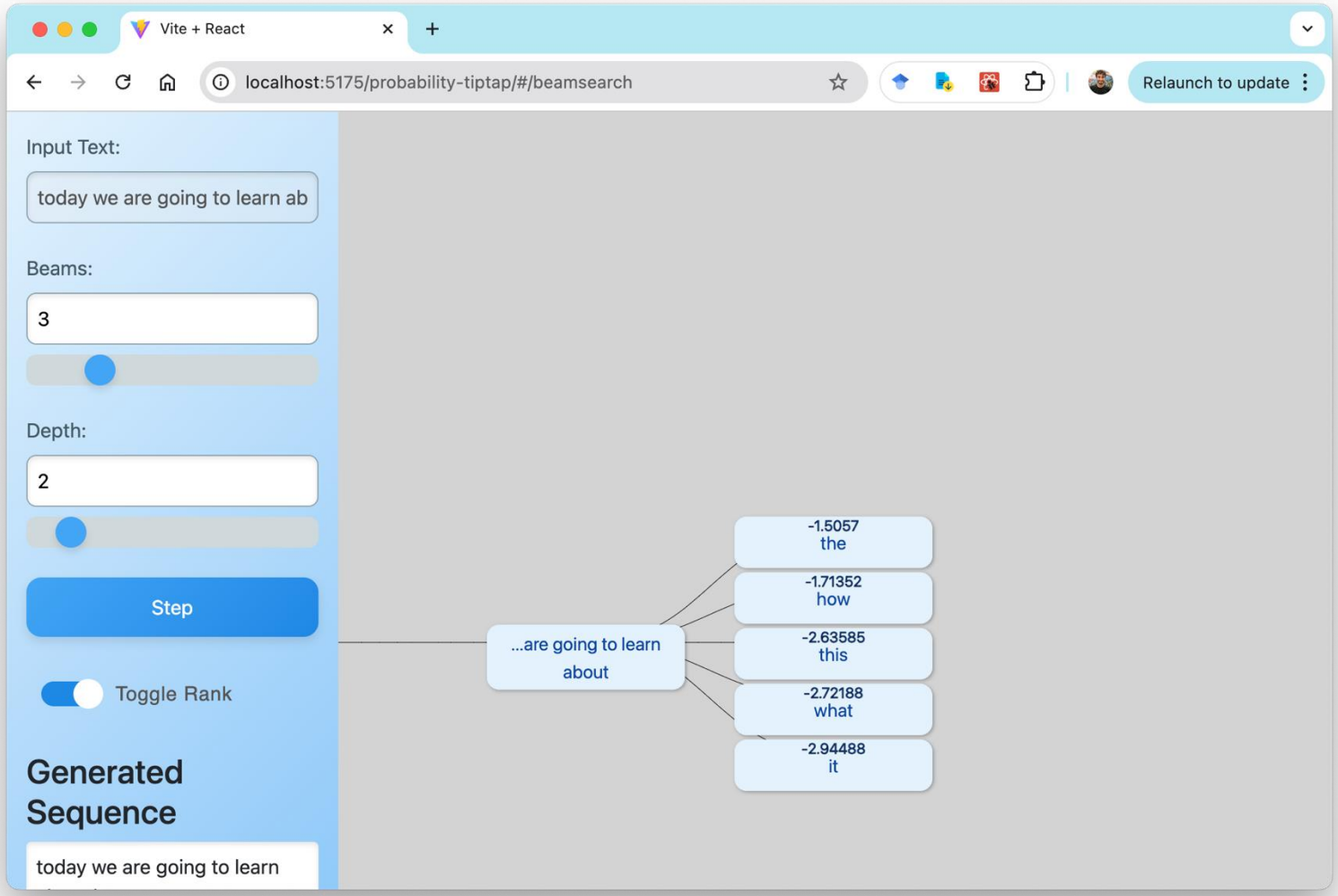
Categorial Example from Class: Thinking of an Animal

Notice that these are not numbers

Output of an LLM is a Categorical for Next Token

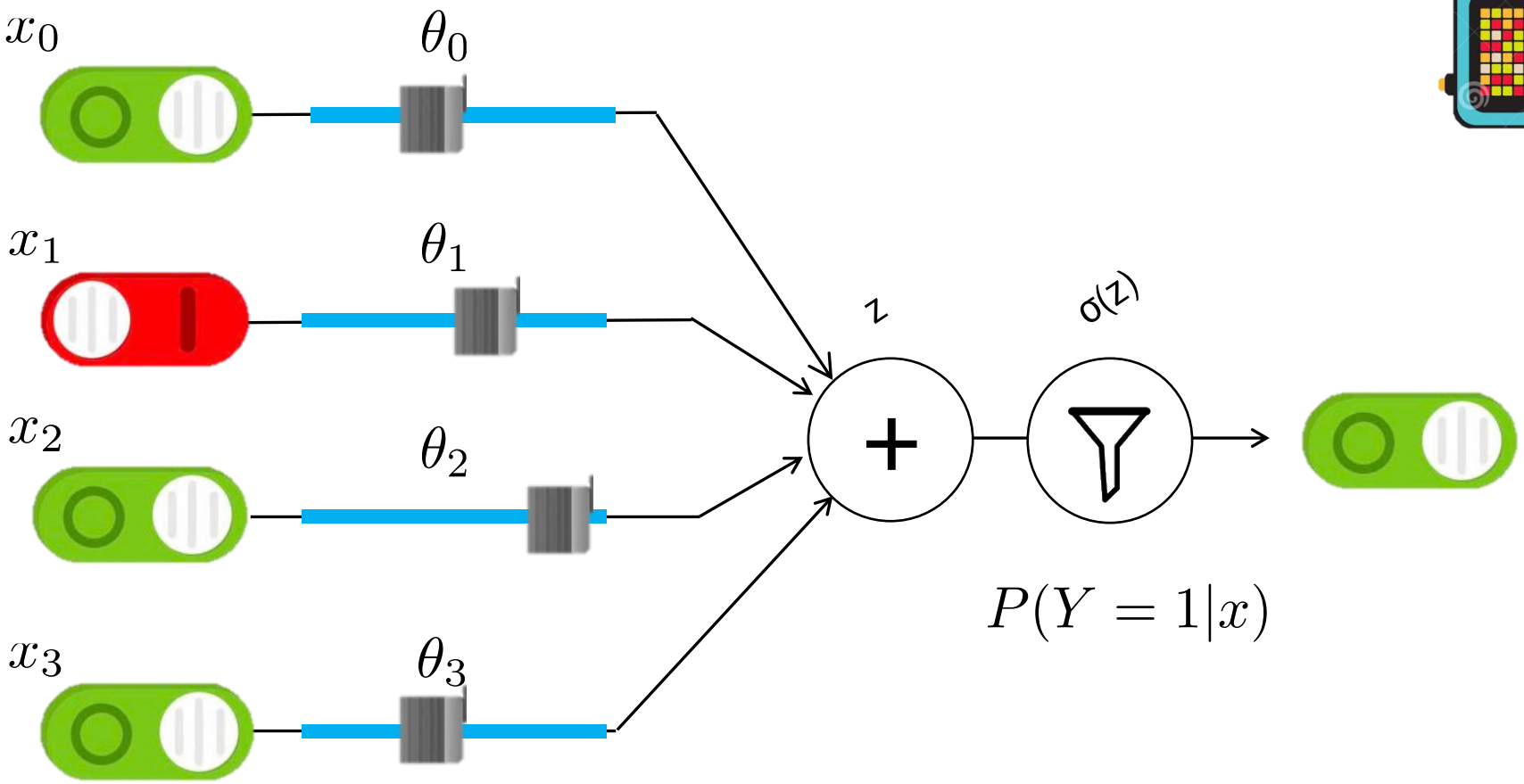


Output of an LLM is a Categorical for Next Token



Visualization is thanks to Justin Blumencranz and Adam Boswell

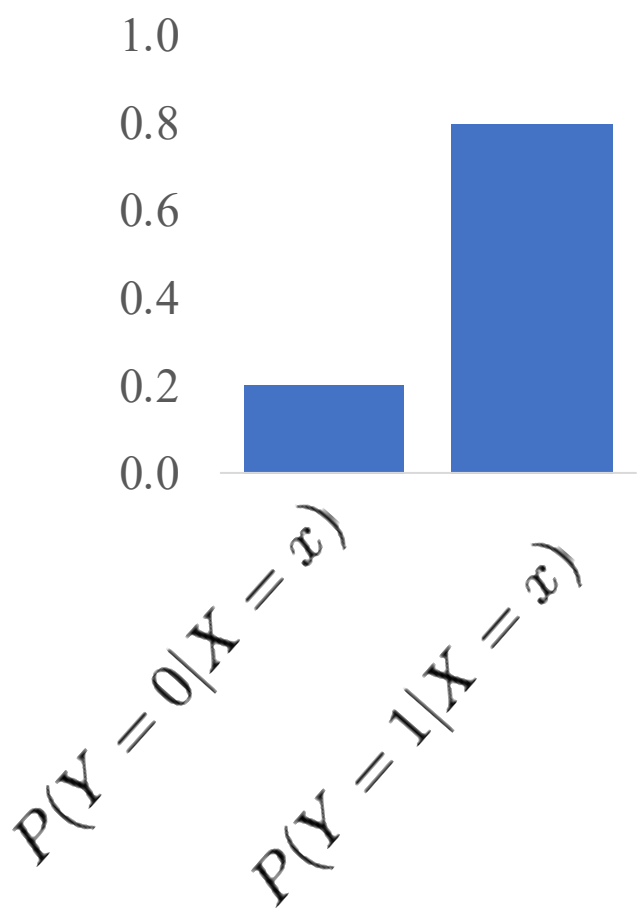
Logistic Regression to Predict a Categorical?



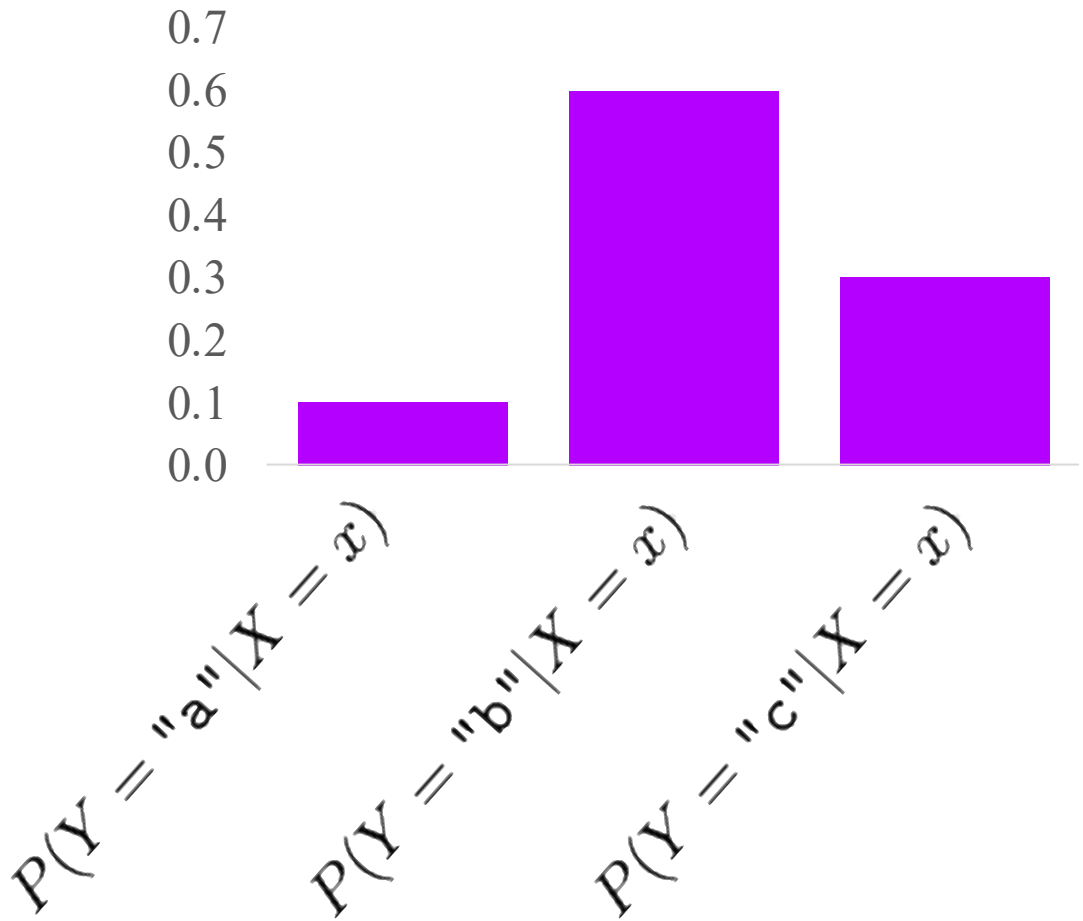
$$P(Y = 1|\mathbf{X} = \mathbf{x}) = \sigma\Big(\sum_i \theta_i x_i\Big)$$

Logistic Regression to Predict a Categorical?

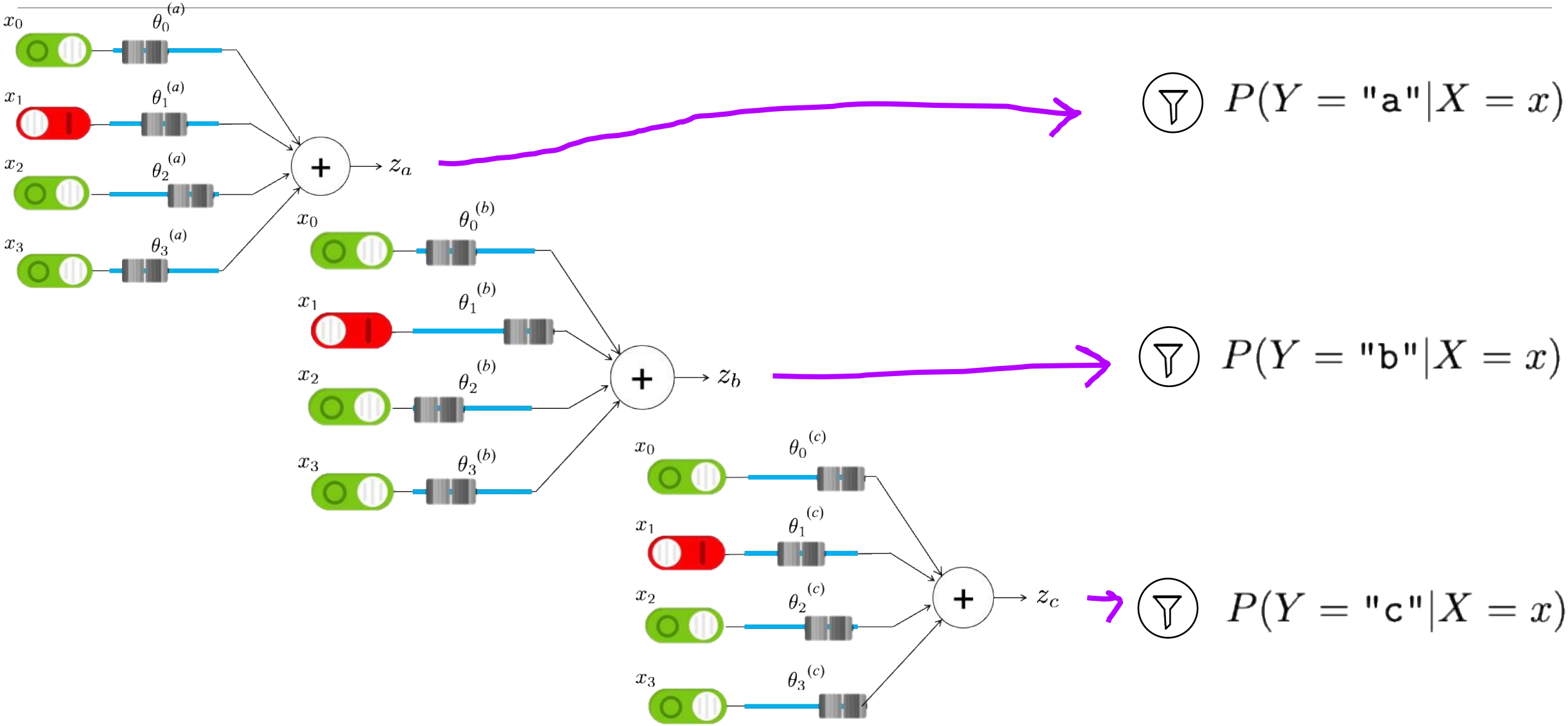
Standard Logistic Regression



Multi-Class Logistic Regression



Logistic Regression to Predict a Categorical



Types of Machine Learning Tasks

Muli-Class Classification

Regression

Today

Reinforcement Learning

Generation

Regression: Predicting Real Numbers

	Opposing team ELO	Points in last game	...	At Home?	Output
					# Points
Game 1	84	105		1	120
Game 2	90	102		0	95
		⋮			⋮
Game *n*	74	120		0	115

Same Notation for Training Data

Training Data: assignments all random variables **X** and Y

Assume IID data:

n training datapoints

$$(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), \ldots (\mathbf{x}^{(n)}, y^{(n)})$$

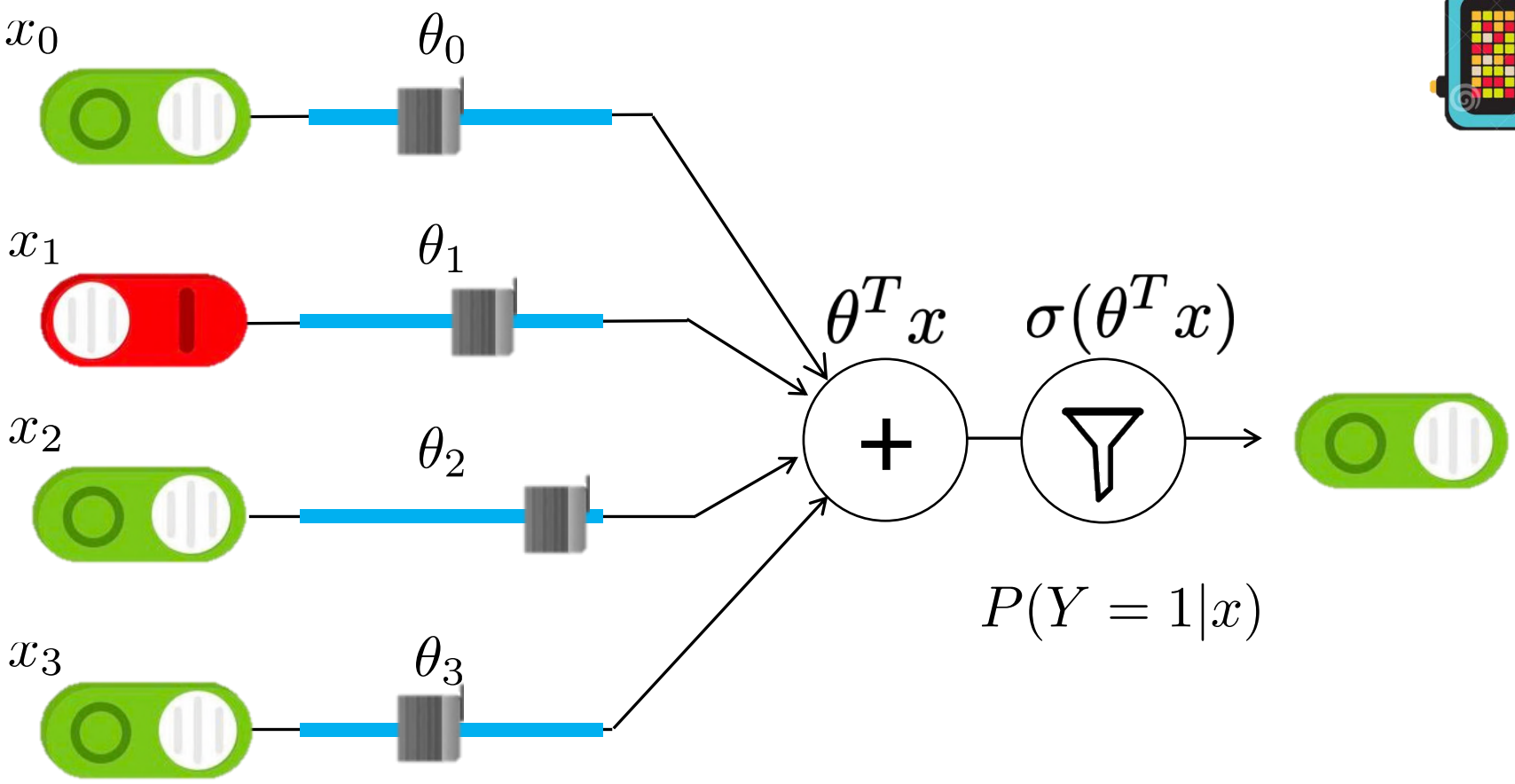
$$m = |\mathbf{x}^{(i)}|$$

Each datapoint has m features and a single output

Regression: Predicting Real Numbers

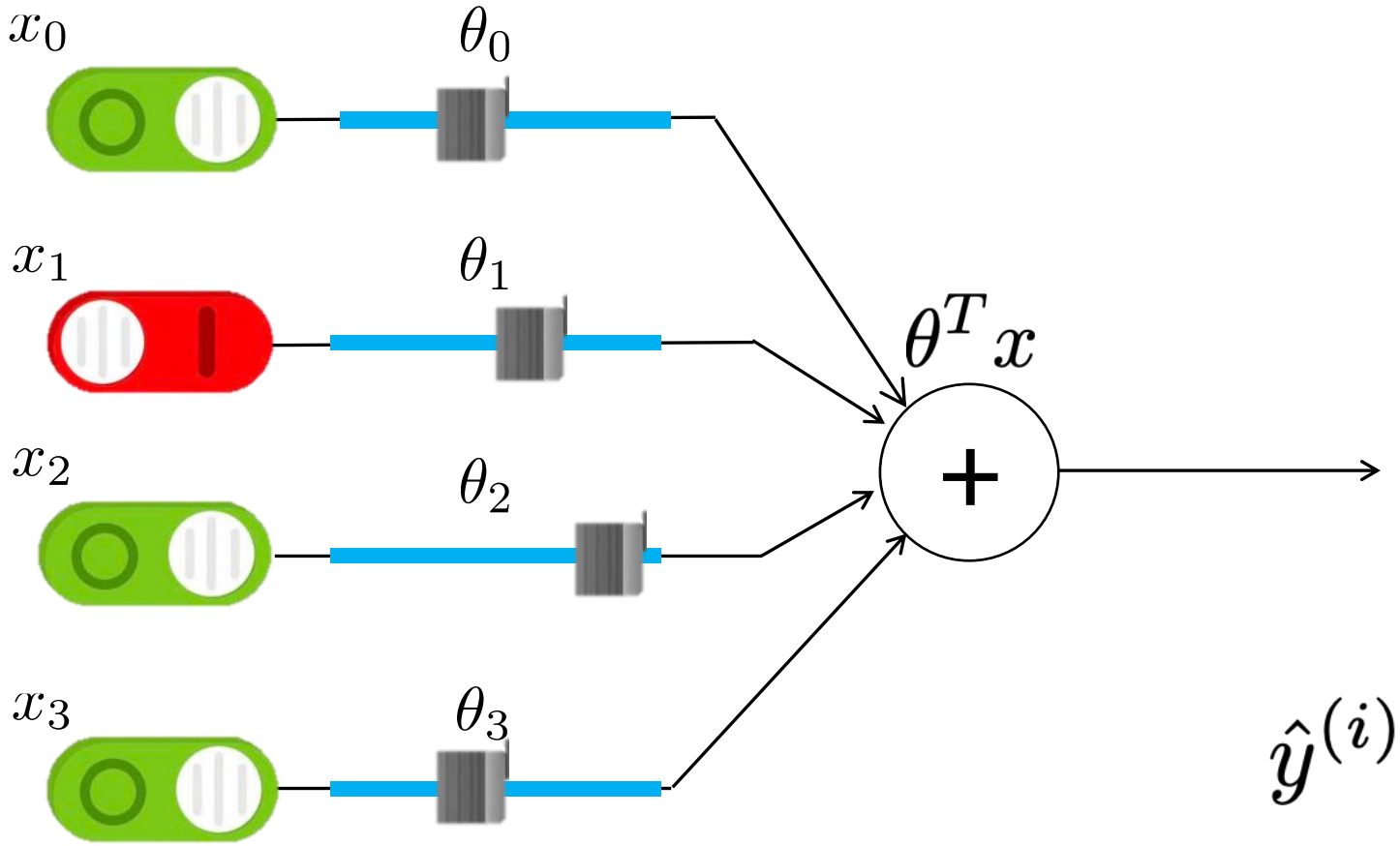
	Opposing team ELO	Points in last game	...	At Home?	Output
					# Points
Game 1	84	105		1	120
Game 2	90	102		0	95
		⋮			⋮
Game *n*	74	120		0	115

Logistic Regression



$$P(Y = 1|\mathbf{X} = \mathbf{x}) = \sigma\Big(\sum_i \theta_i x_i\Big)$$

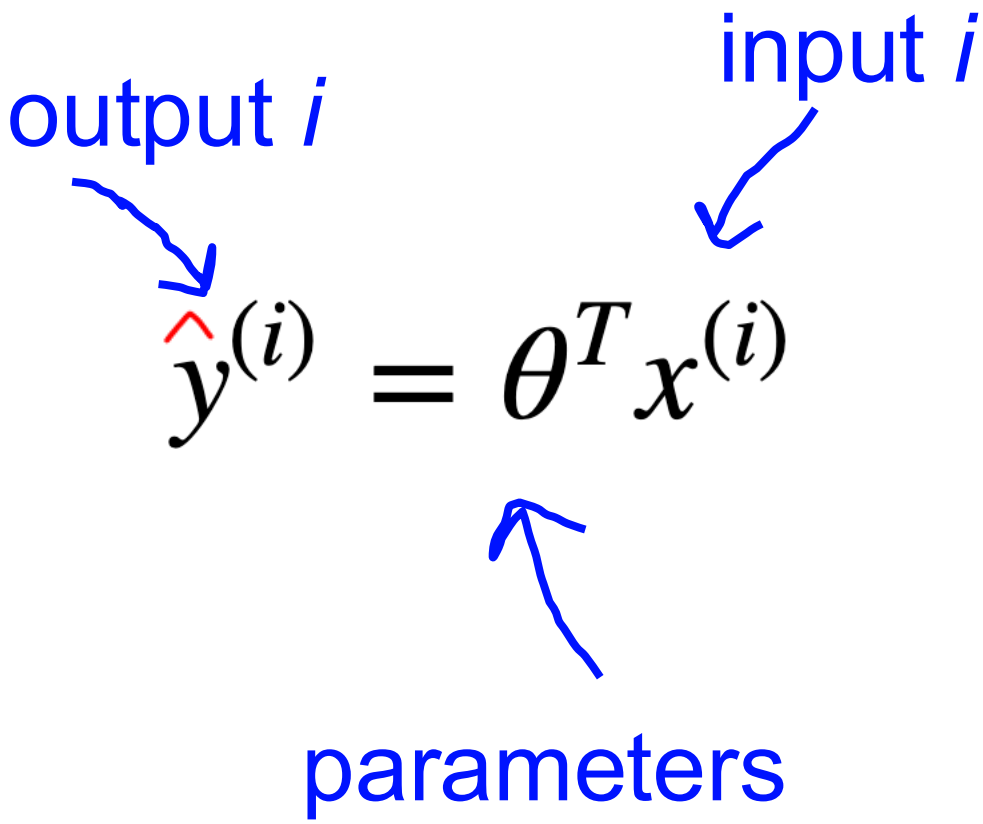
Linear Regression



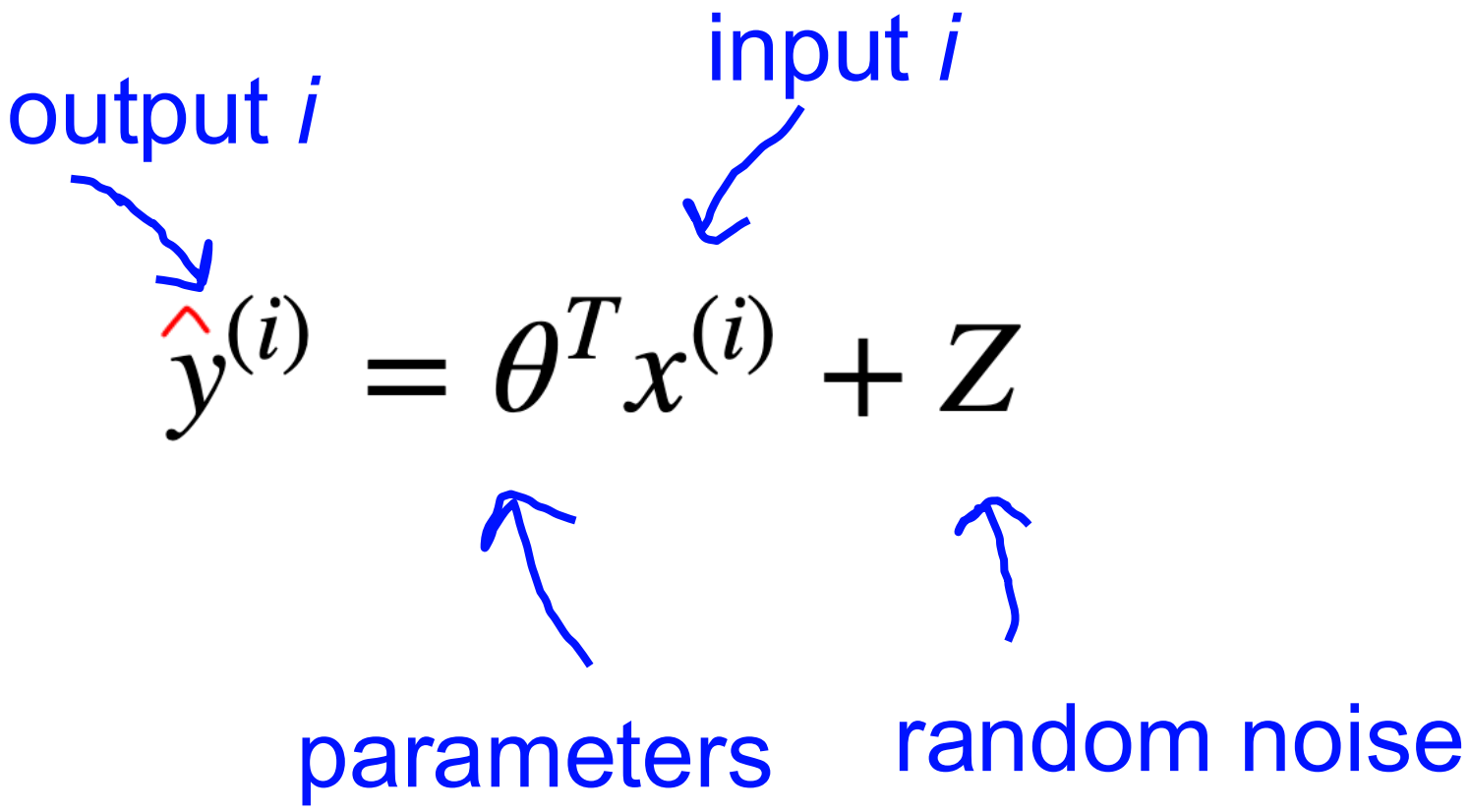
$$\hat{y}^{(i)} = \theta^T x^{(i)} + Z$$

$$Z \sim N(0, \sigma^2)$$

Linear Regression Model



Linear Regression Model



Linear Regression Model

output *i*

input *i*

$$\hat{y}^{(i)} = \theta^T x^{(i)} + Z$$

parameters

random noise

Noise is *N* with mean 0

$$Z \sim N(0, \sigma^2)$$

Linear Regression Model

output *i*

noise

noise is *N* with mean 0

$$\hat{y}^{(i)} = \theta^T x^{(i)} + Z$$

$$Z \sim N(0, \sigma^2)$$

1. Linear Transform

If X is a Normal such that $X \sim N(\mu, \sigma^2)$ and Y is a linear transform of X such that $Y = aX + b$ then Y is also a Normal where:

$$Y \sim N(a\mu + b, a^2\sigma^2)$$

Linear Regression Model

output *i*

noise

$$\hat{y}^{(i)} = \theta^T x^{(i)} + Z$$

noise is *N* with mean 0

$$Z \sim N(0, \sigma^2)$$

Output is normal too:

$$\hat{y}^{(i)} \sim N(\theta^T x^{(i)}, \sigma^2)$$

Log Likelihood

Assume: $\hat{y}^{(i)} \sim N(\theta^T x^{(i)}, \sigma^2)$

Data: $(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), \ldots (\mathbf{x}^{(n)}, y^{(n)})$

$$LL(\theta) = \sum_{i=1}^{n} \log \left[f(y^{(i)}) \right]$$

$$= \sum_{i=1}^{n} \log \left[\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y^{(i)}-\mu}{\sigma}\right)^2} \right]$$

Normal distribution PDF

$$= \sum_{i=1}^{n} \log \left[\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y^{(i)}-\theta^T x^{(i)}}{\sigma}\right)^2} \right]$$

Substitute in the mean

$$= \sum_{i=1}^{n} \log \left[\frac{1}{\sigma\sqrt{2\pi}} \right] - \frac{1}{2} \left(\frac{y^{(i)} - \theta^T x^{(i)}}{\sigma} \right)^2$$

Apply the log

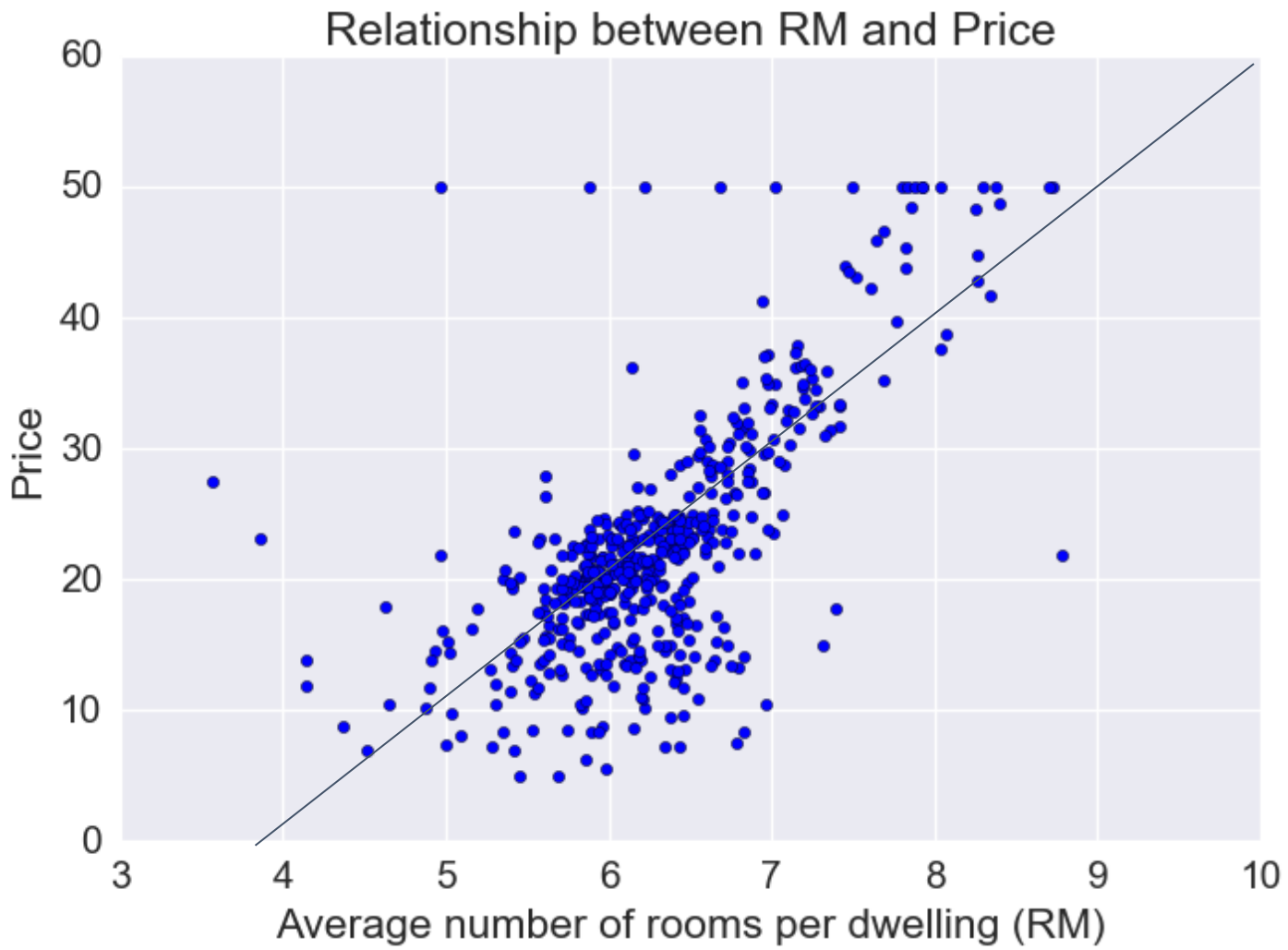
Optimization

Log likelihood: $$LL(\theta) = \sum_{i=1}^{n} \log\left[\frac{1}{\sigma\sqrt{2\pi}}\right] - \frac{1}{2}\left(\frac{y^{(i)} - \theta^T x^{(i)}}{\sigma}\right)^2$$

$$\begin{aligned}
\underset{\theta}{\text{argmax}}\, LL(\theta) &= \underset{\theta}{\text{argmax}} \sum_{i=1}^{n} \log\left[\frac{1}{\sigma\sqrt{2\pi}}\right] - \frac{1}{2}\left(\frac{y^{(i)} - \theta^T x^{(i)}}{\sigma}\right)^2 \\
&= \underset{\theta}{\text{argmax}} - \sum_{i=1}^{n} \frac{1}{2}\left(\frac{y^{(i)} - \theta^T x^{(i)}}{\sigma}\right)^2 \qquad \text{Simplify} \\
&= \underset{\theta}{\text{argmax}} - \sum_{i=1}^{n} \left(y^{(i)} - \theta^T x^{(i)}\right)^2 \qquad \text{Simplify}
\end{aligned}$$

Hey it's the sum of squared errors!

Linear Regression with one Input



Derivative is Necessary for Gradient Ascent

$$\frac{\partial LL(\theta)}{\partial \theta_j} = -\sum_{i=1}^{n} \frac{\partial}{\partial \theta_j} \left(y^{(i)} - \theta^T x^{(i)} \right)^2$$

Derivative of a sum

$$= -\sum_{i=1}^{n} 2\left(y^{(i)} - \theta^T x^{(i)} \right)(-x_j^{(i)})$$

Chain rule

$$= \sum_{i=1}^{n} 2\left(y^{(i)} - \theta^T x^{(i)} \right) \cdot x_j^{(i)}$$

Simplify

Types of Machine Learning Tasks

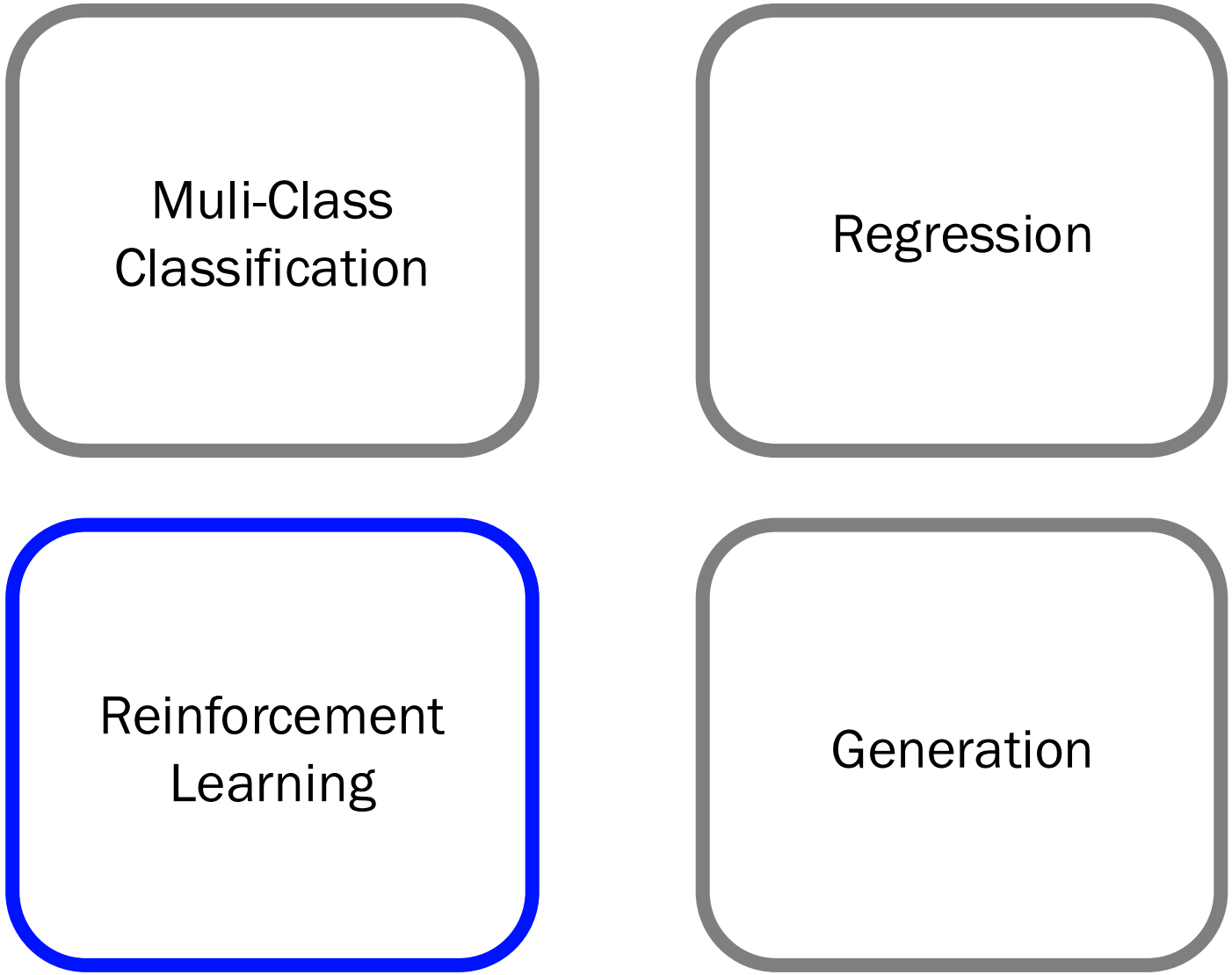
Muli-Class Classification

Regression

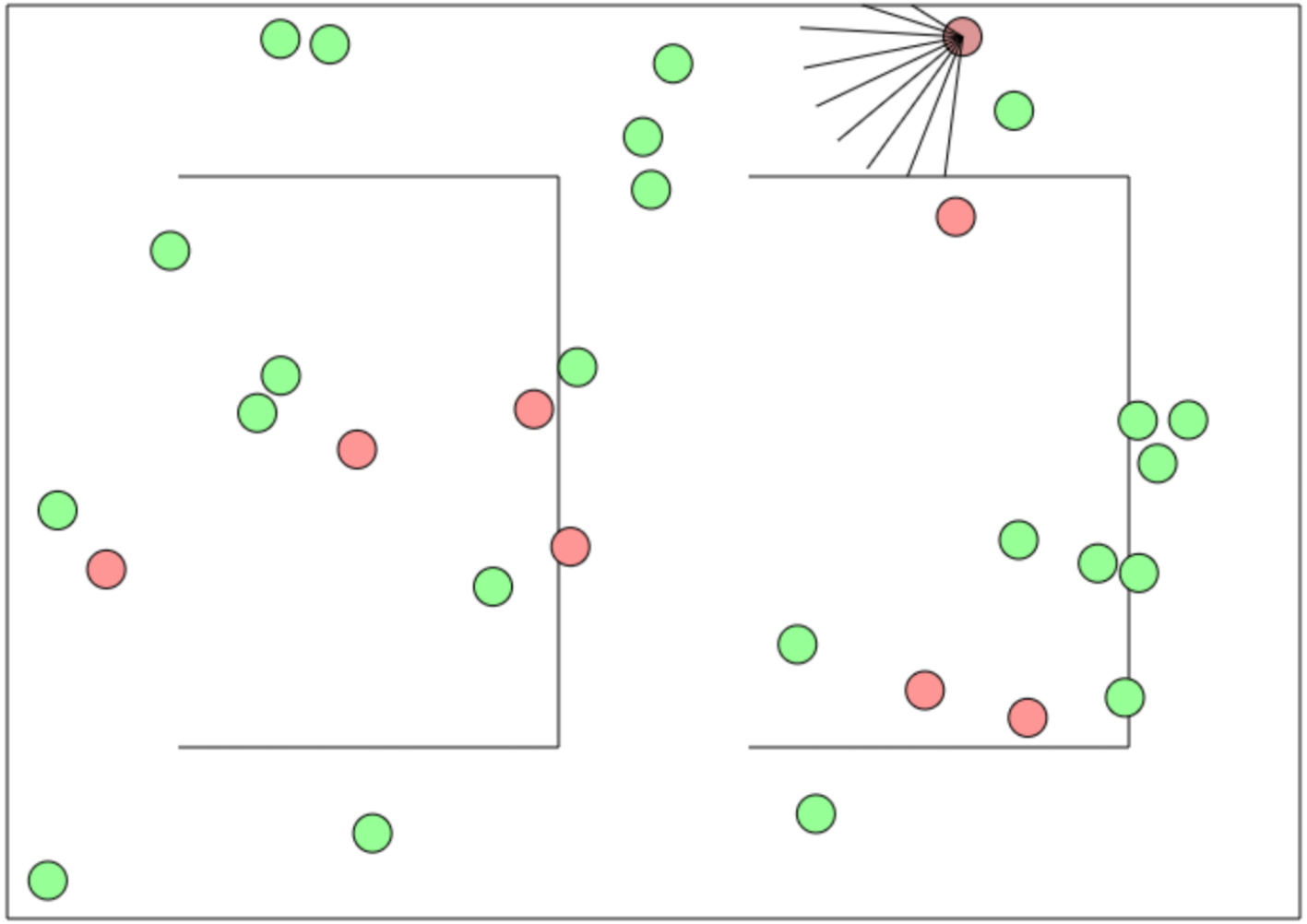
Reinforcement Learning

Generation

Types of Machine Learning Tasks

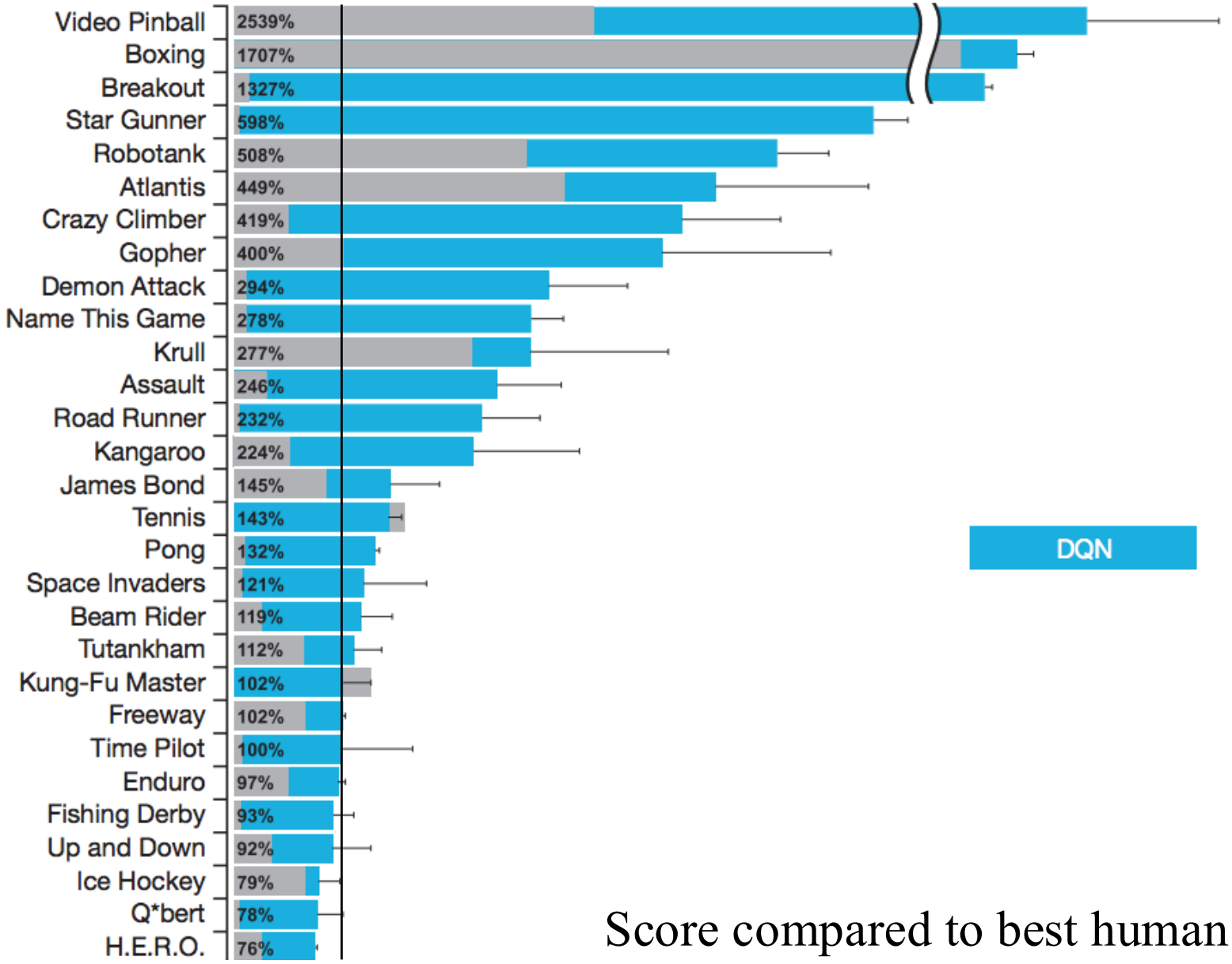


Deep Reinforcement Learning



http://cs.stanford.edu/people/karpathy/convnetjs/demo/rldemo.html

Deep Mind Atari Games

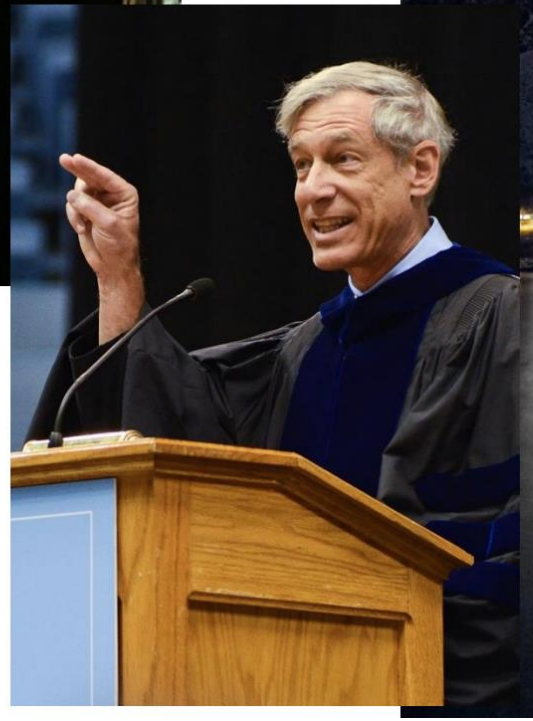


Score compared to best human

Review

Night Sight

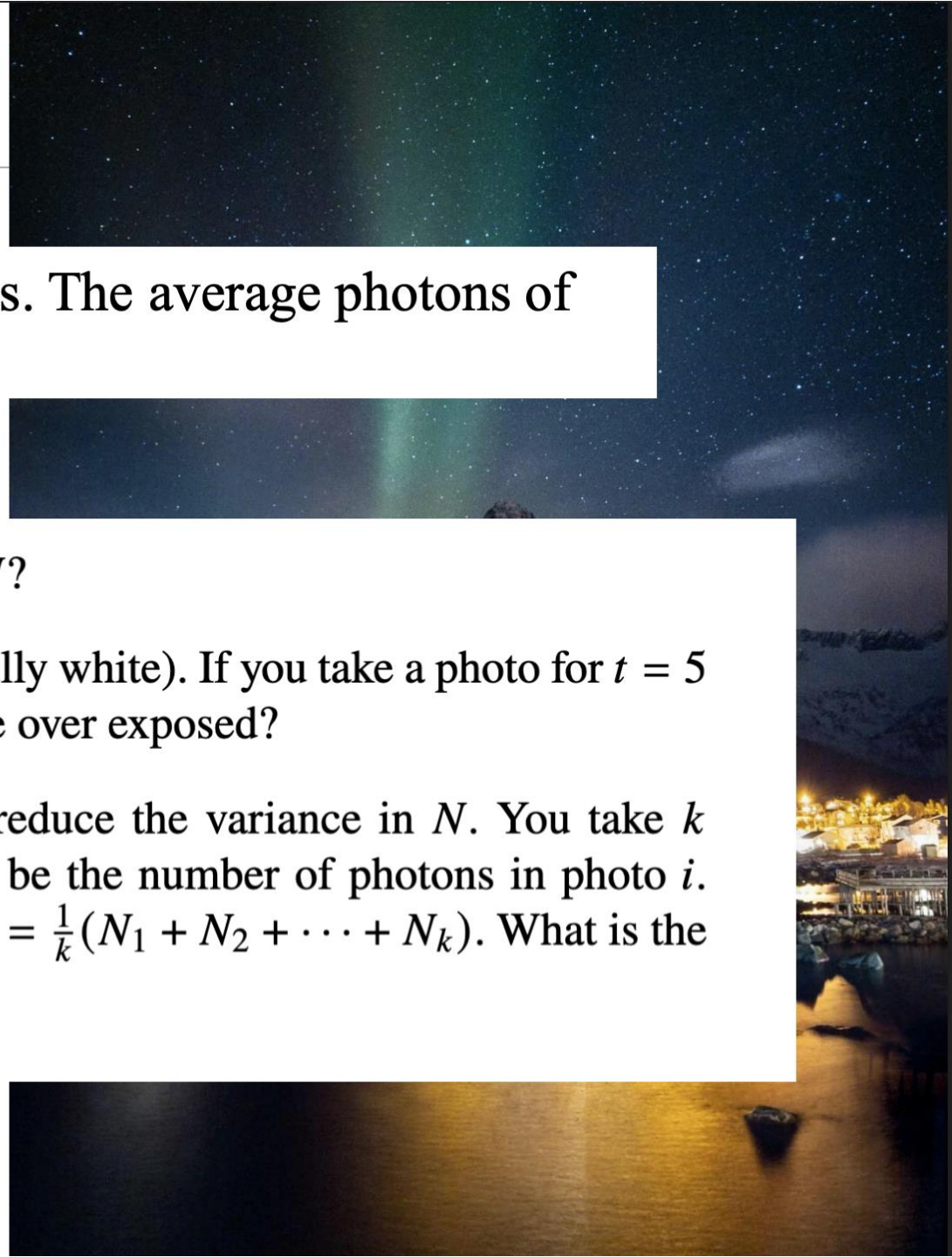
Night Sight



Mark Levoy, Stanford Emeritus Professor

https://static.googleusercontent.com/media/hdrplusdata.org/en//hdrplus.pdf

Night Sight



Light hits sensors on a camera as a Poisson process. The average photons of light that hit a given pixel in one second is r.

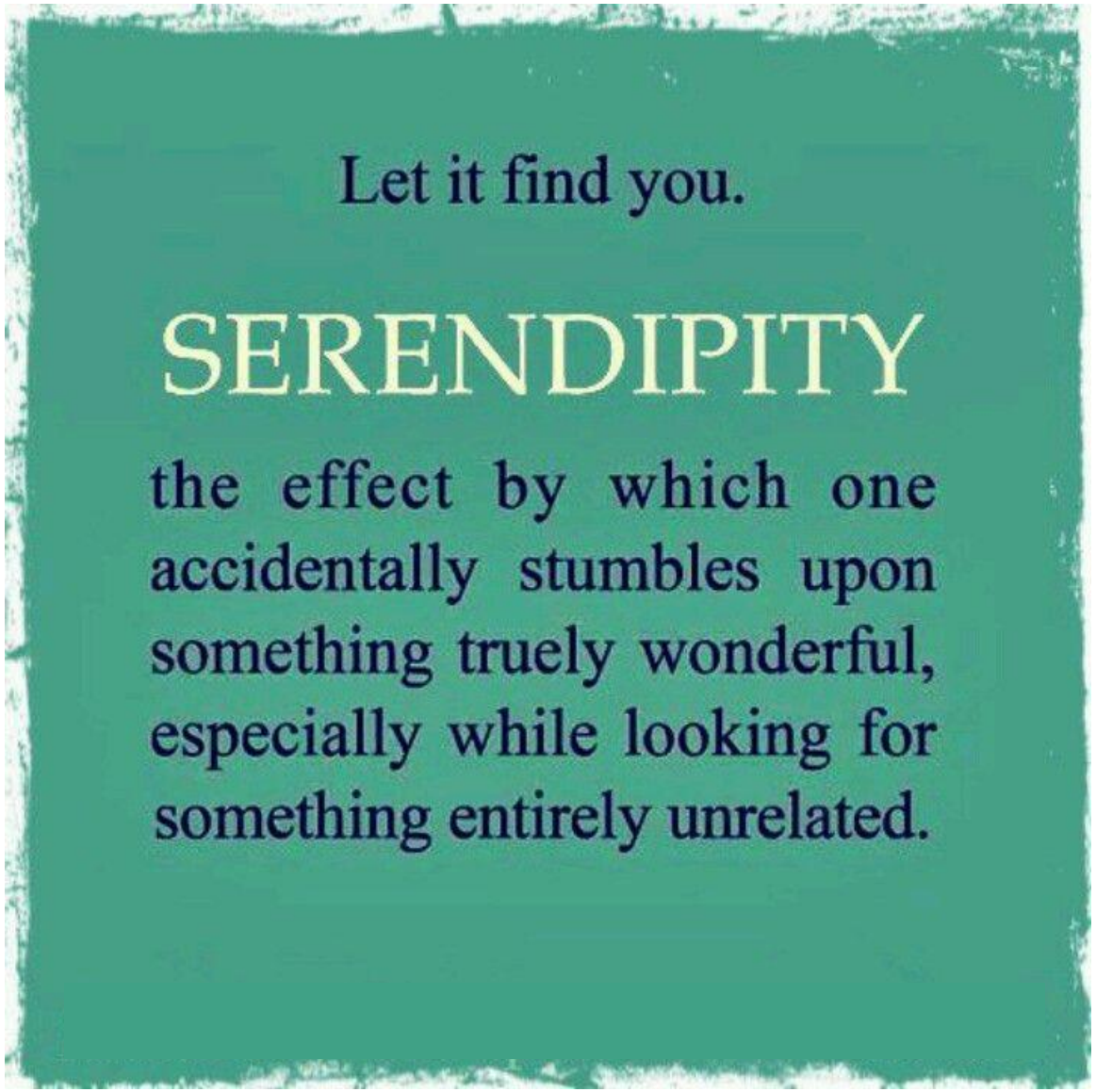
Let N be the number of photons that hit our pixel.

a. A photo is taken for t seconds. What is the variance of N?

b. If $N > 100$, the pixel will be over exposed (and will be fully white). If you take a photo for $t = 5$ seconds, what is the probability that a given pixel will be over exposed?

c. If you take several photos and average them, you can reduce the variance in N. You take k photos, each with exposure length of t seconds. Let N_i be the number of photons in photo i. The average number of photons across the k photos is $N = \frac{1}{k}(N_1 + N_2 + \cdots + N_k)$. What is the variance of N?

Serendipity

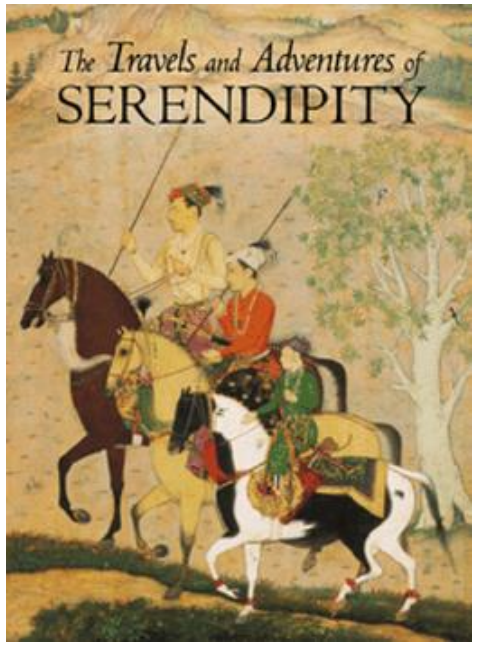

Let it find you.
SERENDIPITY
the effect by which one accidentally stumbles upon something truely wonderful, especially while looking for something entirely unrelated.

WHEN YOU MEET YOUR BEST FRIEND

Somewhere you didn't expect to.

Serendipity

- Say the population of Stanford is 17,000 people
  - You are friends with 100
  - Walk into a room, see 450 random people.
  - What is the probability that you see someone you know?
  - Assume you are equally likely to see each person at Stanford



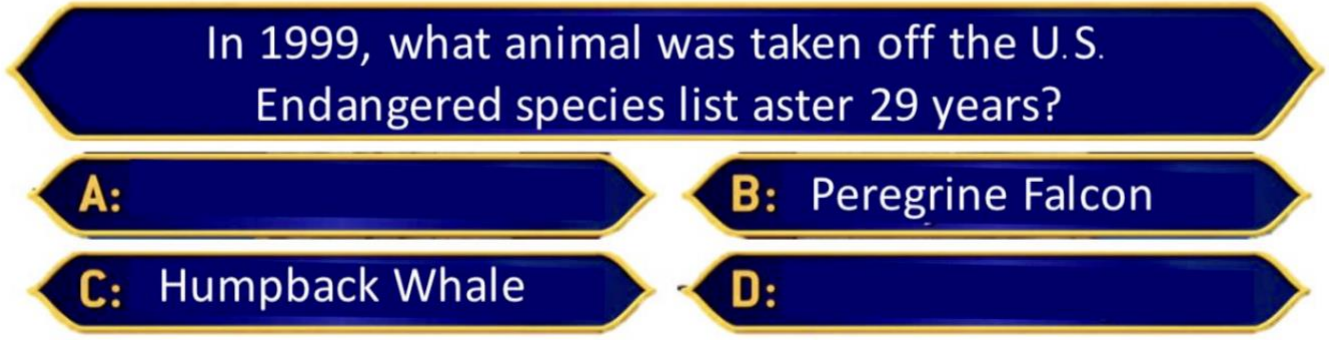
Many times it is easier to calculate $P(E^C)$.

Wisdom of the Crowds

Wisdom of the Crowds

There are two answers for each audience member to chose from, a Correct answer and an Incorrect answer.

- 10% of the audience are knowledgeable about the problem (call them experts). An expert votes for the Correct answer with a probability of 0.7, otherwise they vote for the Incorrect answer.
- 90% of the audience are not knowledgeable (call them non-experts). A non-expert votes randomly with equal likelihood between the Correct answer and the Incorrect answer.



Wisdom of the Crowds

200 People in Audience

There are two answers for each audience member to chose from, a Correct answer and an Incorrect answer.

- 10% of the audience are knowledgeable about the problem (call them experts). An expert votes for the Correct answer with a probability of 0.7, otherwise they vote for the Incorrect answer.
- 90% of the audience are not knowledgeable (call them non-experts). A non-expert votes randomly with equal likelihood between the Correct answer and the Incorrect answer.

a. What is the probability that exactly k of the experts vote for the Correct answer? You may assume that k is a number between 0 and 20 inclusive.

b. If exactly k of the experts vote for the Correct answer, what is the probability that the Correct answer will get at least 101 votes? (hint: the Correct answer needs at least 101 - k more votes from the non-experts).

c. Write an expression for the exact probability that the Correct answer will get at least 101 votes.

d. Use an approximation to estimate the probability that the Correct answer gets at least 101 votes. You may leave your answer in terms of roots and/or values that could be looked up from the ϕ table. For full credit your approximation calculation should *not* include a summation or integral.

Quant Interview

Quant Interview

100 cupcakes.
Scenario 1: 51 are blue.
Scenario 2: 49 are blue.

a) You look at one cupcake. It is blue.

b) You look at three cupcakes (with replacement). Two are blue.

Quant Interview - Bonus

100 cupcakes.
Scenario 1: 51 are blue.
Scenario 2: 49 are blue.

a) You look at one cupcake. It is blue.

b) You look at three cupcakes (with replacement). Two are blue.

c) You look at three cupcakes (without replacement). Two are blue.

Quant Interview: Bonus

c. A hypergeometric is a random variable for the number of successes if you remove items **without** replacement from a fixed population. If $X \sim \text{Hypergeom}(t, k, n)$, the PMF is

$$\mathrm{P}(X = x) = \frac{\binom{k}{x}\binom{t-k}{n-x}}{\binom{t}{n}}$$

Where:

- t is the population size,
- k is the number of "success" items in the population,
- n is the number of draws (without replacement),
- X counts the number of successes observed in those n draws.

Three cupcakes are drown without replacement and two are blue. What is the probability that the majority were blue?