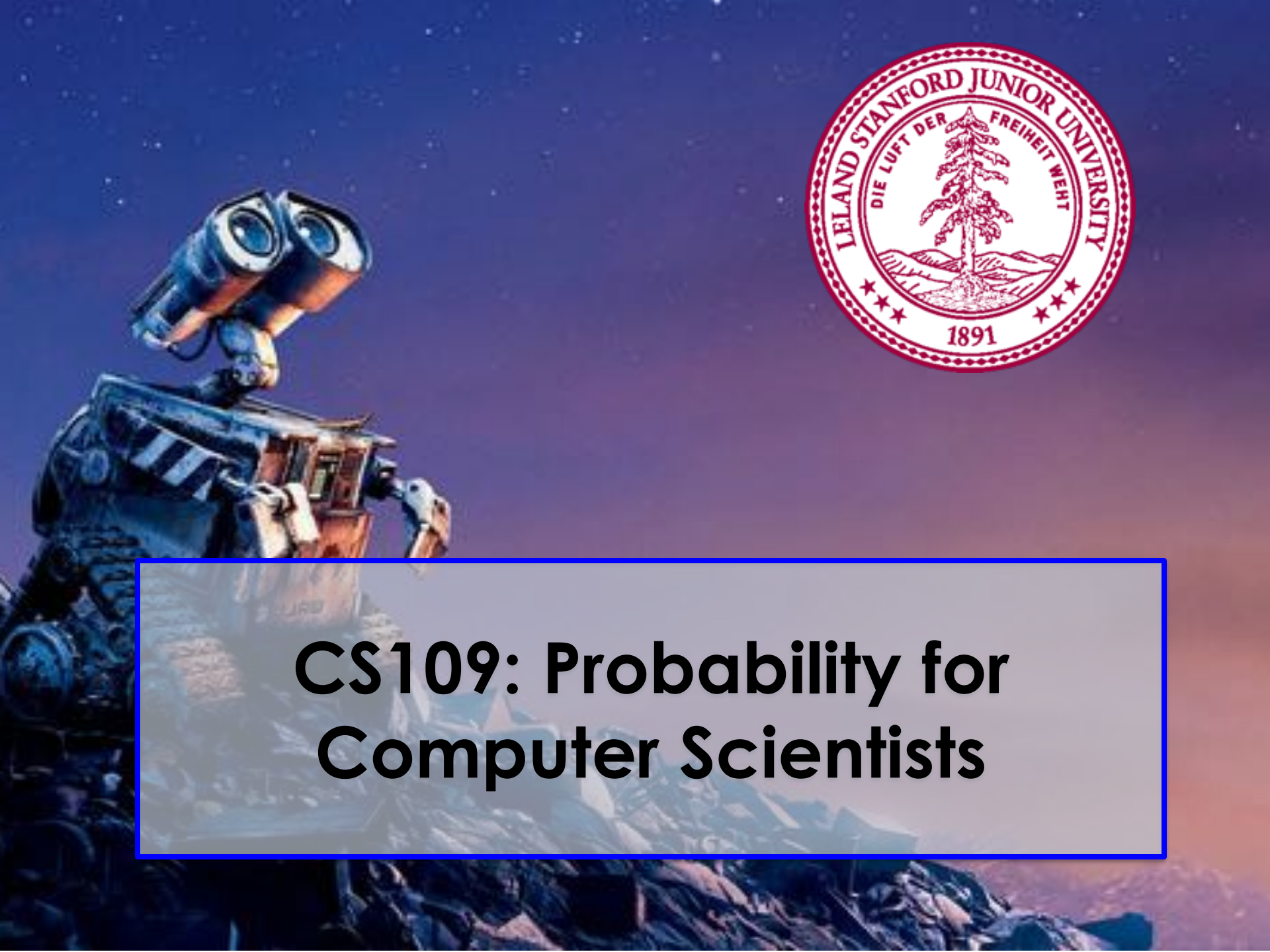
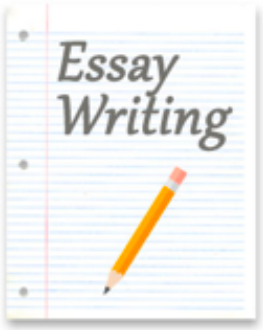


CS109: Probability for Computer Scientists

Noah Arthurs



STANFORD UNIVERSITY
SLE
STRUCTURED
LIBERAL
EDUCATION

Tycho

Teaching Team

We love questions!

=

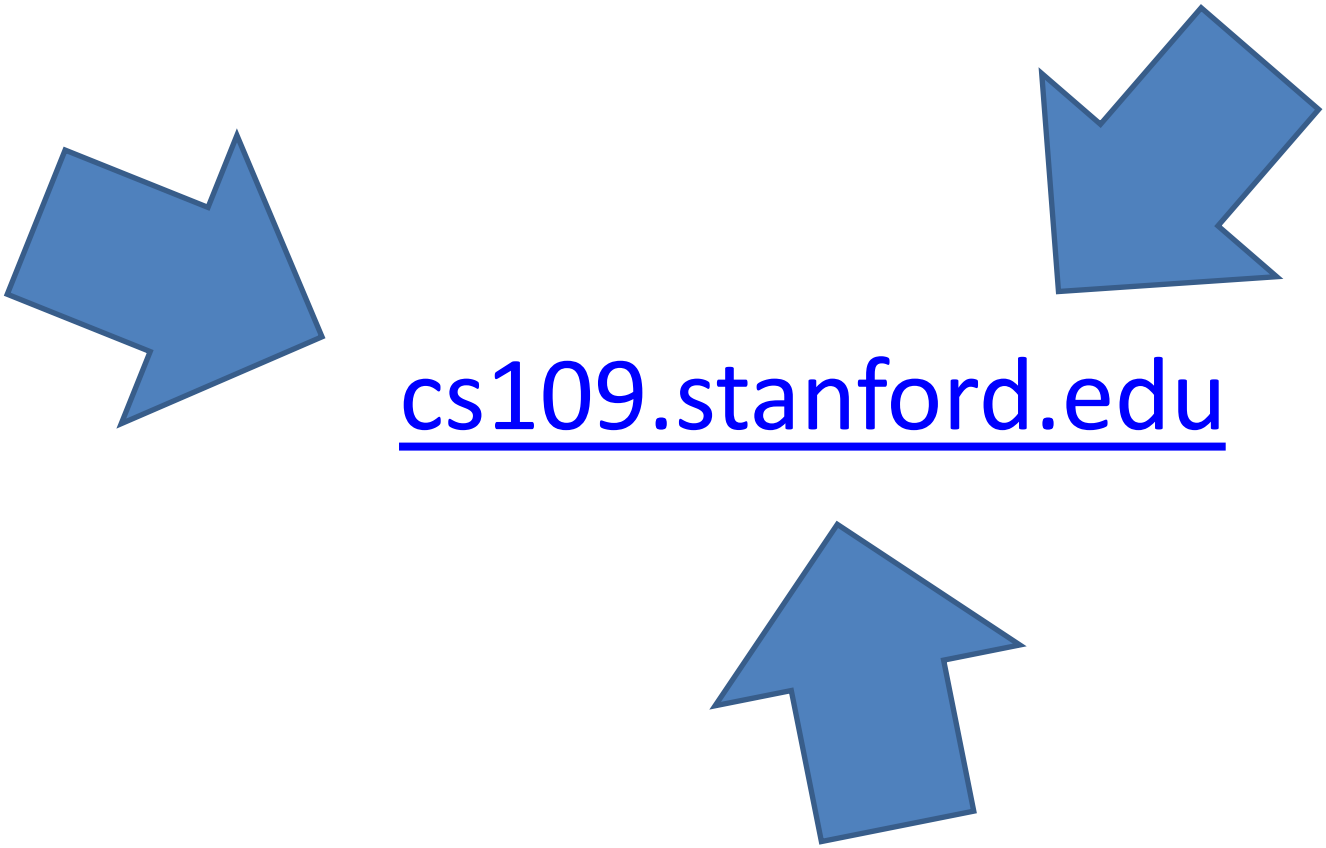
This Lecture

1. Course Mechanics
2. Why you should take CS109
3. Counting!

Course mechanics

(this is a light version. Please read the handout for details).

Essential Information



cs109.stanford.edu

Are you in the right place?

Prereqs

What you really need:

CS106B/X (important):

- Recursion
- Hash Tables
- Binary Trees
- Programming

CS103 (ok as a corequisite):

- Proof techniques (induction)
- Set theory
- Math maturity

Math 51 or CME 100 (important)

- Multivariate differentiation
- Multivariate integration
- Basic facility with linear algebra (vectors)

Coding in CS109

Review session on Thursday

Staff Contact

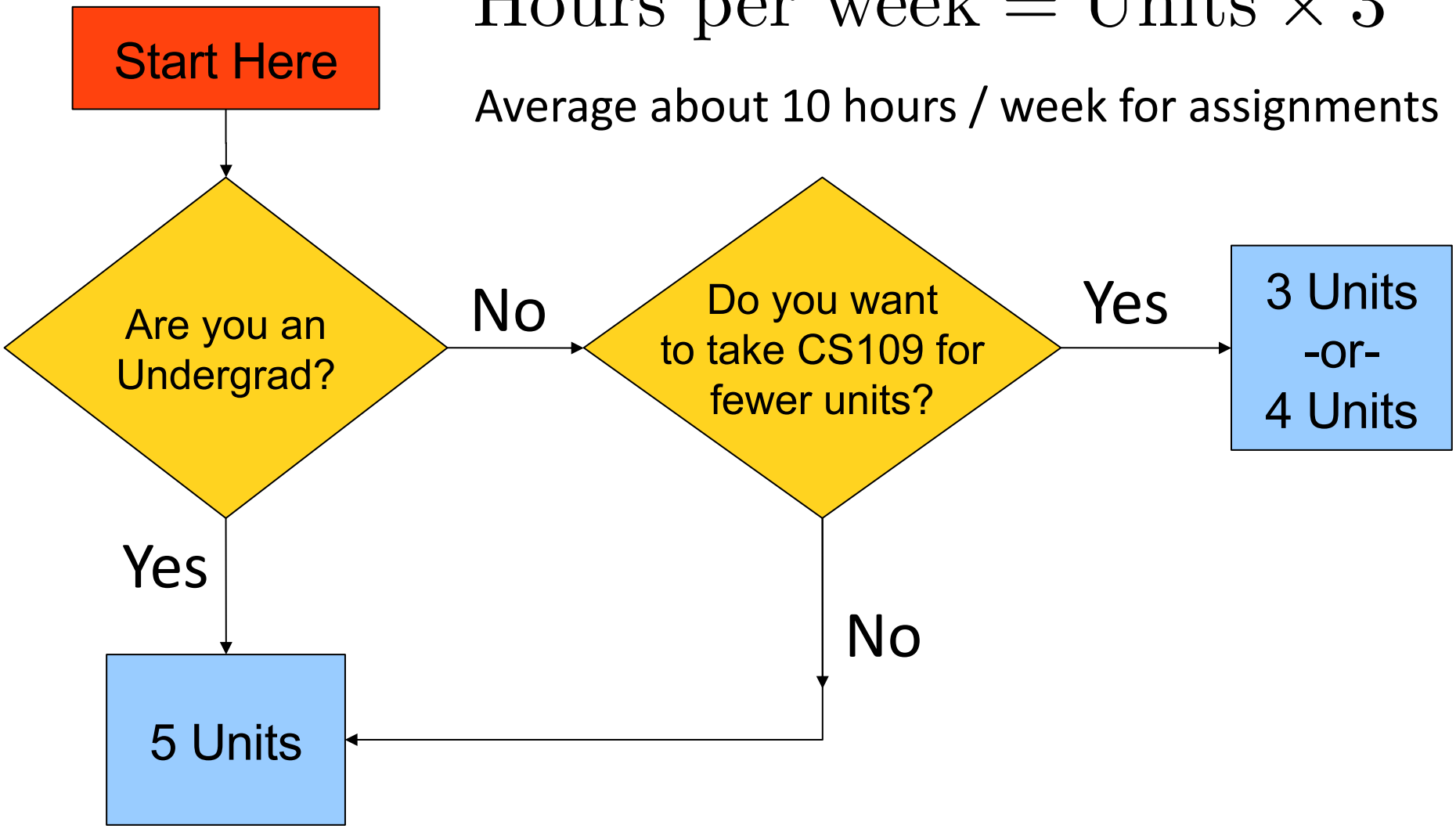
- Post to Piazza for clarification
- Go to Working Office Hours
- Email cs109-sum-1819-staff@lists.stanford.edu

- Email Noah or go to his office for course level issues.

CS109 Units

$$\text{Hours per week} = \text{Units} \times 3$$

Average about 10 hours / week for assignments



Videotaped*

*but there are many reasons you should come to lecture

Sections

Thursdays 2:30-3:20PM in Gates B03

Participation

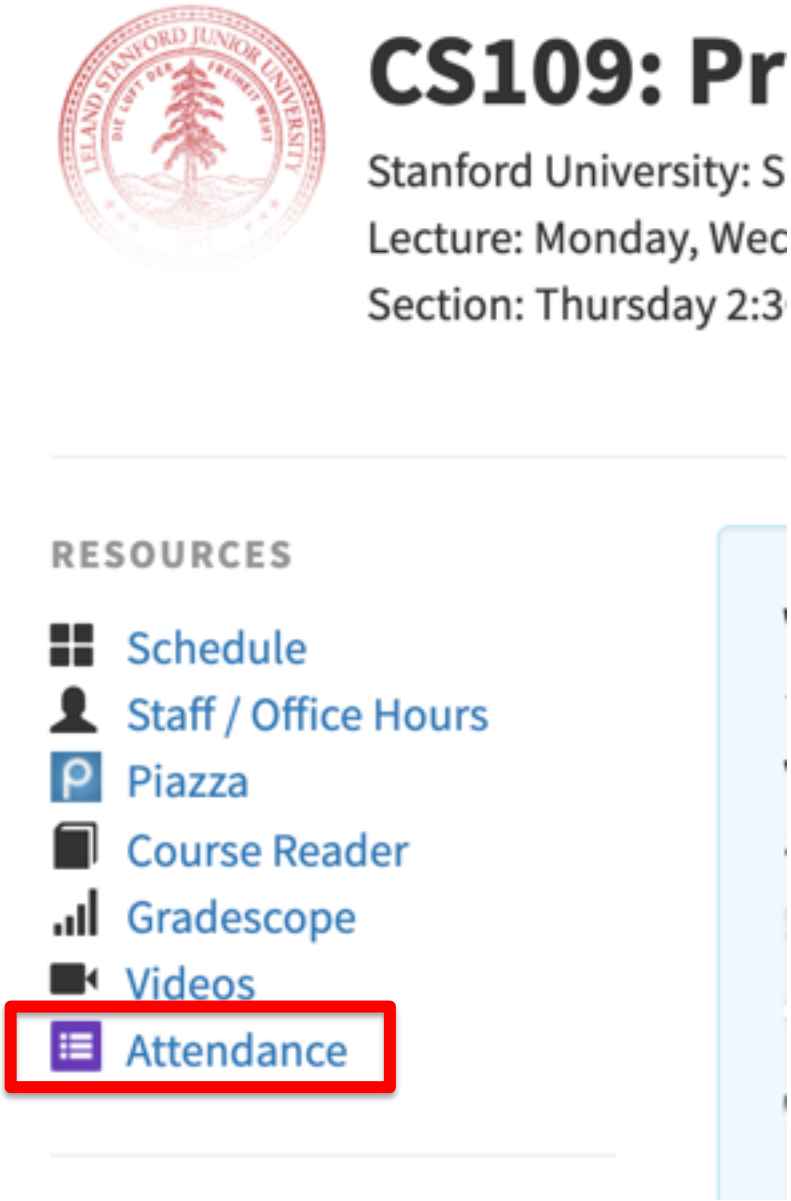
If you:

- Attend every lecture,
- Attend every section after this week,
- Fill out mid-quarter feedback,

then you get full credit for 7% of your grade and your final counts for less.

This is incremental, so every bit of participation helps your final grade.

Taking Attendance



In-person code: on board during lecture (not on video)

SCPD/remote: on slides in video

Class Breakdown

45% **6 Assignments**

20% **Midterm**
Tuesday July 23rd, 7-9pm

28-35% **Final**
Friday August 16th, 3:30-6:30pm

0-7% **Participation**

Late Days

2

Textbook

Chris Piech
CS 109

Lecture Notes #1
Sept 24, 2018

Counting

Based on a handout by Mehran Sahami with examples by Peter Norvig

Although you may have thought you had a pretty good grasp on the notion of counting at the age of three, it turns out that you had to wait until now to learn how to really count. Aren't you glad you took this class now?! But seriously, below we present some properties related to counting which you may find helpful in the future.

Counting is important in the world of computer science for a few reasons. In order to understand probability on a fundamental level, it is useful to first understand counting. Moreover, while computers are fast, some problems require so much work that they would take an unreasonable amount of time to complete. We can use counting theory to reason about complexity.

1 Sum Rule

Sum Rule of Counting
If the outcome of an experiment can either be one of m outcomes or one of n outcomes, where none of the outcomes in the set of m outcomes is the same as the any of the outcomes in the set of n outcomes, then there are $m + n$ possible outcomes of the experiment.

Email me if…

- You have an OAE letter or need accommodations of any kind.
- You absolutely cannot make the regular midterm time (note that we are not offering alternate finals)
- You are NOT an SCPD student but you cannot make it to lectures and/or sections.

My email is narthurs@stanford.edu.

Story of Modern AI

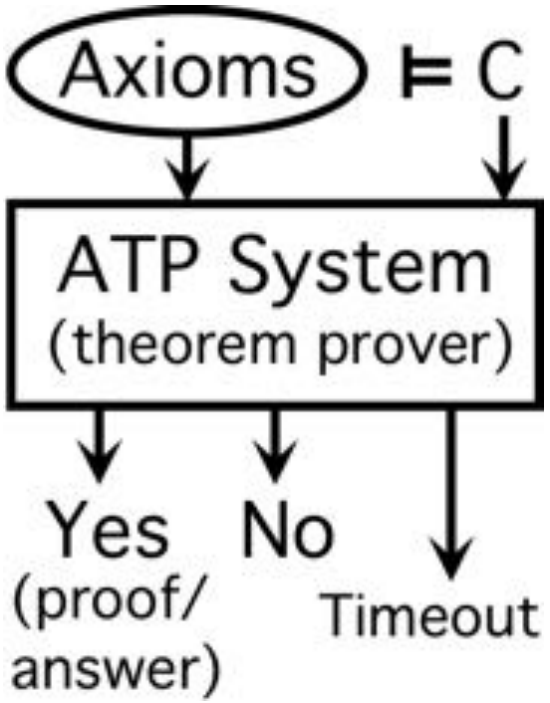
Modern AI
or, How we learned to combine probability and programming

Brief History

Early Optimism 1950

1952

1955



Early Optimism 1950

"Machines will be capable, within twenty years, of doing any work a man can do."
–Herbert Simon, 1952

Underwhelming Results 1950s to 1980s

The world is too complex

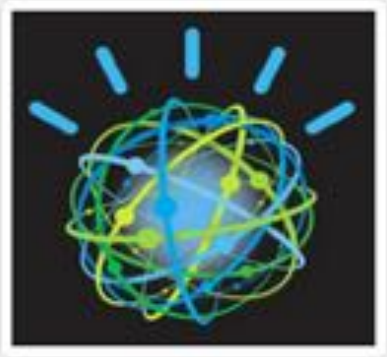

BRACE YOURSELVES
WINTER IS COMING

Something is going on in the world of AI

Big Milestones Pt 1

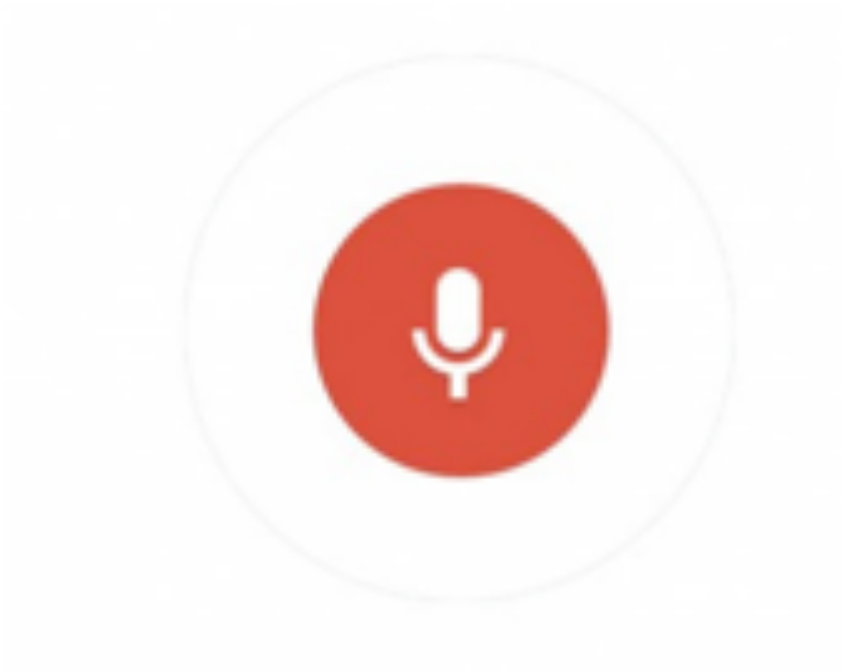
1997 Deep Blue

2005 Stanley



2011 Watson

Told Speech Was 30 Years Out



Almost perfect…

The Last Remaining Board Game

Computers Making Art

Self Driving Cars

What is going on?

[suspense]

Focus on one problem

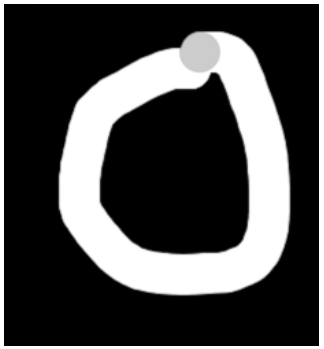
Computer Vision

Classification

That is a picture of a **one**

Classification

That is a picture of a **zero**



Classification

That is a picture of an **zero**

* It doesn't have to be correct all of the time

Can you do it?

What number is this?

What number is this?

How about now?

What a computer sees

0	0	1	0	1	0	1	0	0	0	1	1	1	0	1
1	0	0	1	0	1	1	1	0	1	0	0	0	0	0
1	1	1	0	1	0	0	1	1	0	0	1	0	1	0
1	1	1	1	1	0	0	0	0	0	1	1	0	1	1
0	0	0	1	1	0	0	1	0	0					
1	0	0	1	1	0	0	0	1	0					
1	1	0	1	1	0	0	1	1	0					
1	0	1	0	0	1	0	0	1	0					
0	0	0	0	1	0	1	0	1	1					
0	1	1	0	0	0	0	0	1	1					
0	0	1	0	1	1	1	0	0	0					
0	1	1	1	0	1	0	0	1	0					
1	1	0	0	0	0	0	0	0	0					
0	0	0	0	0	0	0	0	1	1					
0	0	1	1	1	0	1	0	1	1					

What a human sees

Very hard to Program


```
public class HarryHat extends ConsoleProgram {

    public void run() {
        println("Todo: Write program");
    }

}
```

Two Great Ideas

1. Probability from Examples

2. Artificial Neurons

Two Great Ideas

1. Probability from Examples

2. Artificial Neurons

1. Probability From Examples

When Does the Magic Happen?

Lots of Data + Sound Probability

Machine Learning

Basically just a rebranding of statistics and probability.

Vision is Hard

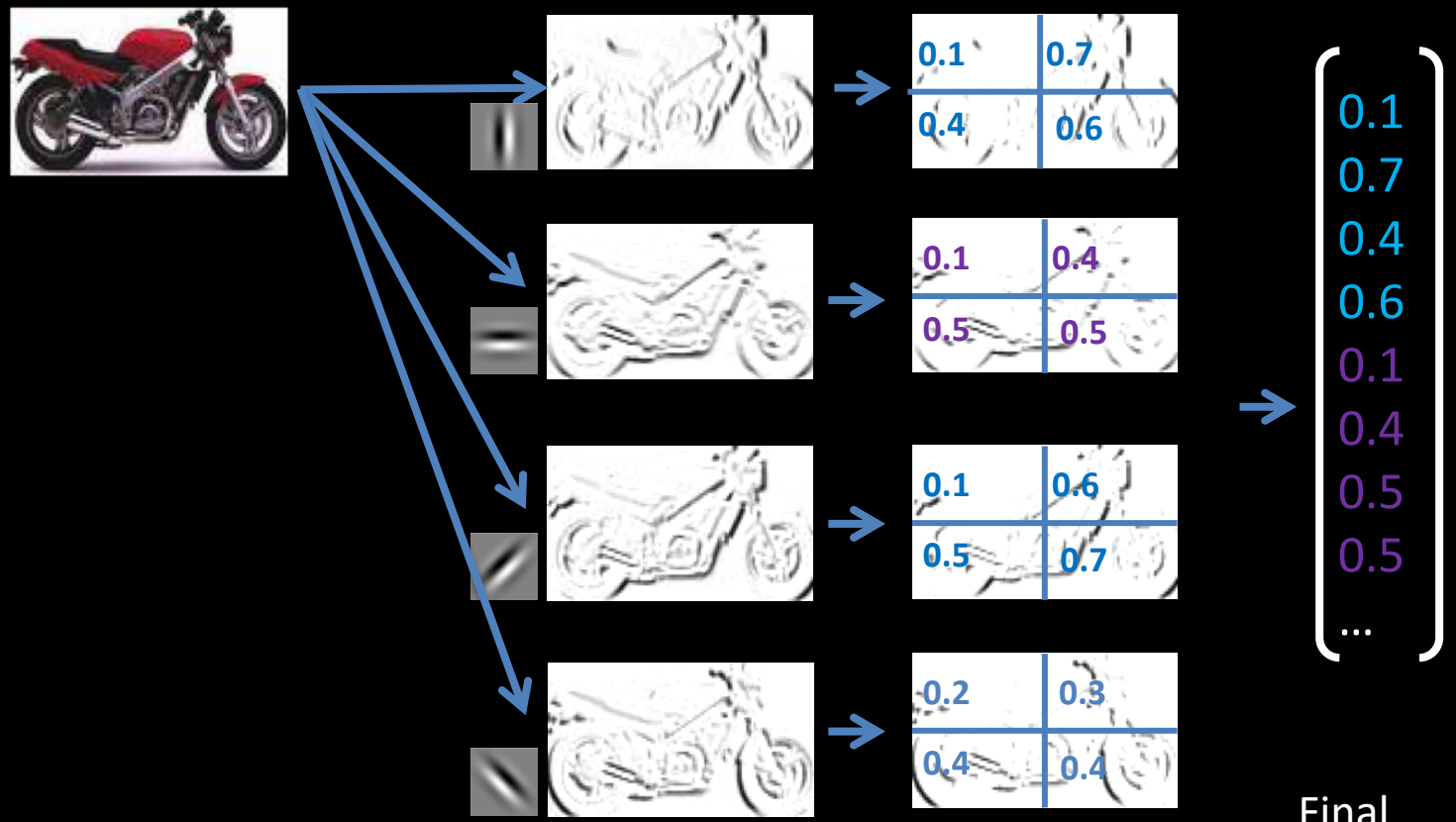
You see this:

But the camera sees this:

194	210	201	212	199	213	215	195	178	158	182	209
180	189	190	221	209	205	191	167	147	115	129	163
114	126	140	188	176	165	152	140	170	106	78	88
87	103	115	154	143	142	149	153	173	101	57	57
102	112	106	131	122	138	152	147	128	84	58	66
94	95	79	104	105	124	129	113	107	87	69	67
68	71	69	98	89	92	98	95	89	88	76	67
41	56	68	99	63	45	60	82	58	76	75	65
20	43	69	75	56	41	51	73	55	70	63	44
50	50	57	69	75	75	73	74	53	68	59	37
72	59	53	66	84	92	84	74	57	72	63	42
67	61	58	65	75	78	76	73	59	75	69	50

[Andrew Ng]

Human Designed Features



[Andrew Ng]

Some Great Thinkers

Daphne Koller

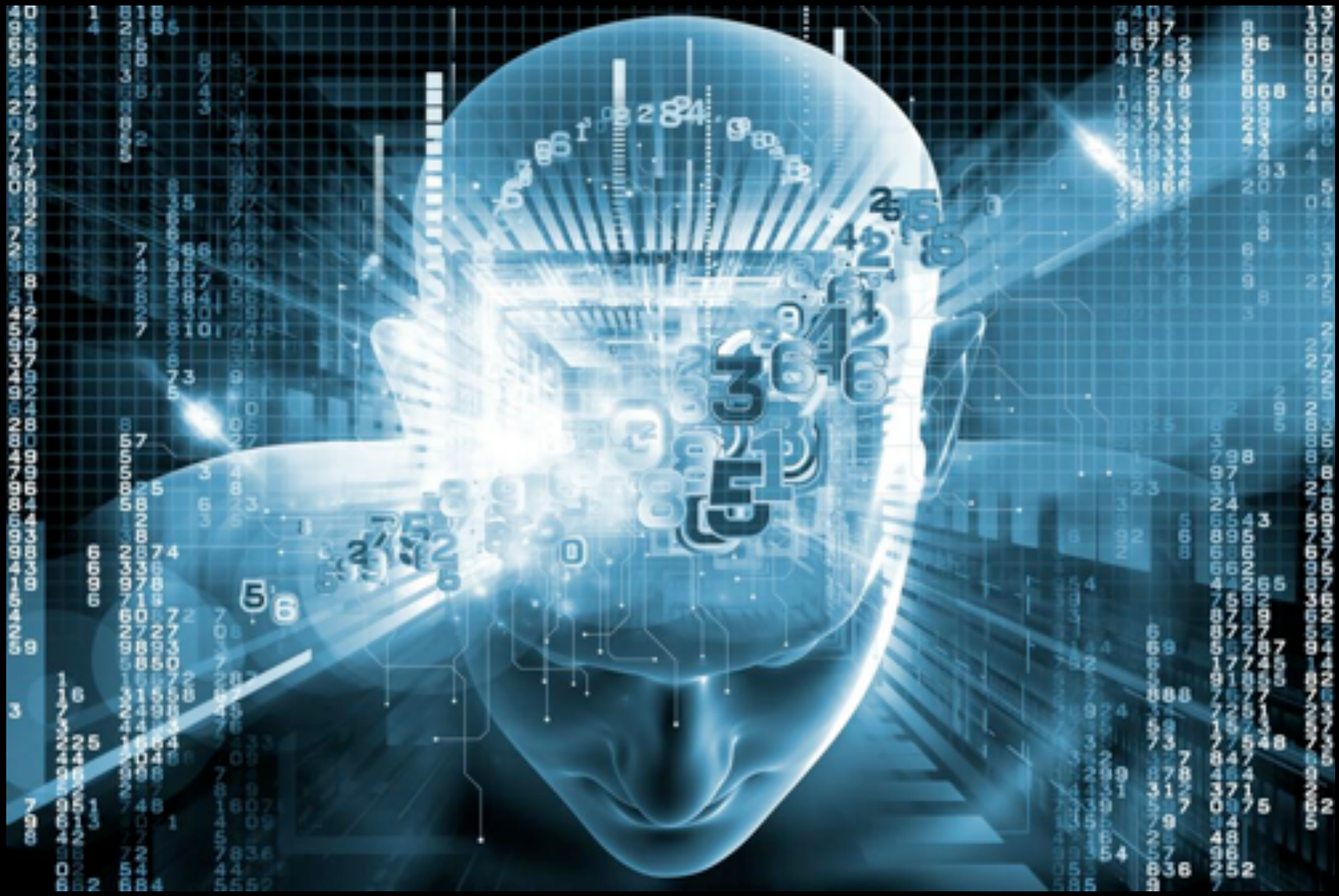
Straight ML Not Perfect…

Two Great Ideas

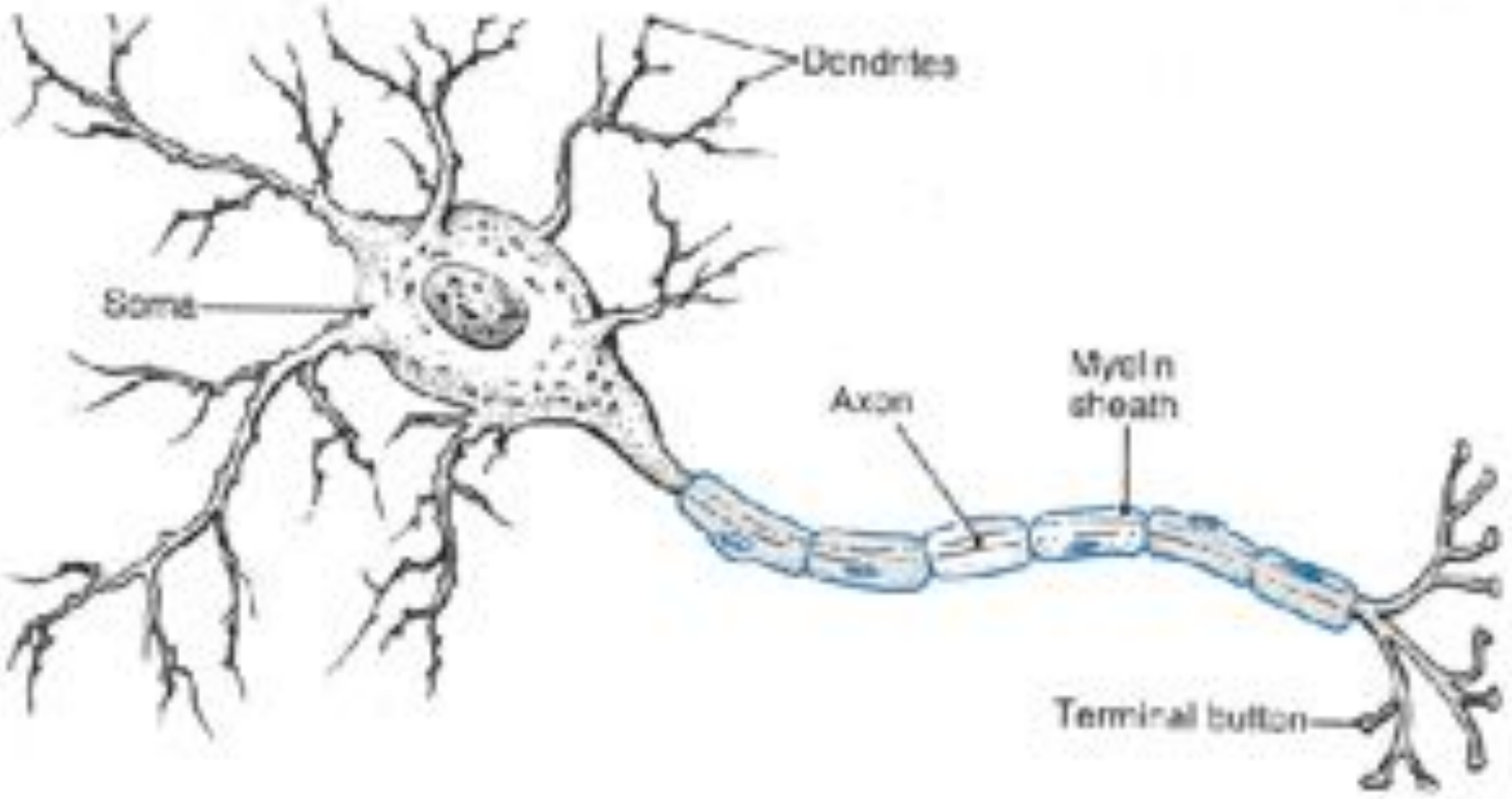
1. Probability from Examples

2. Artificial Neurons

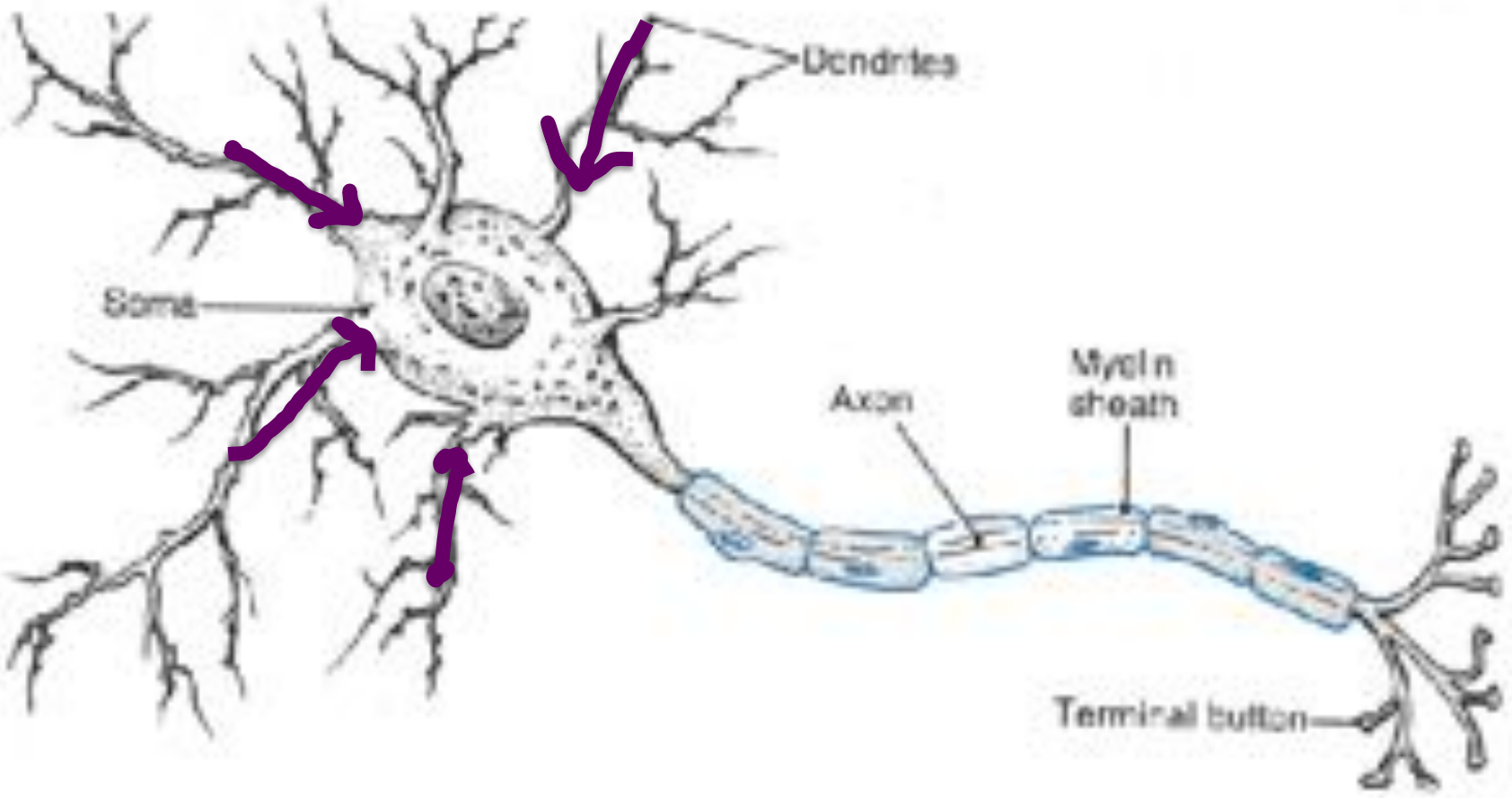
2. Artificial Neurons



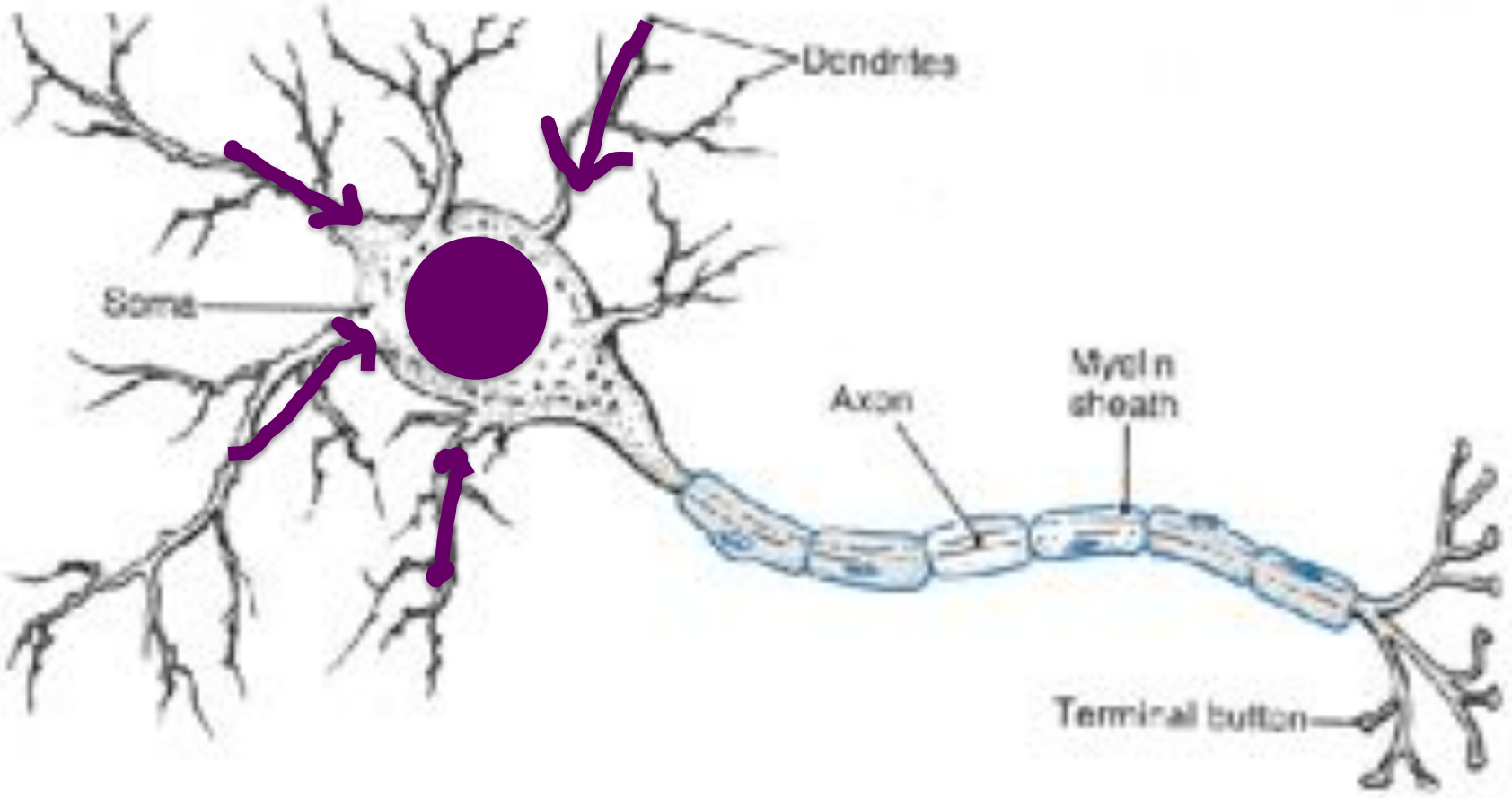
Neuron



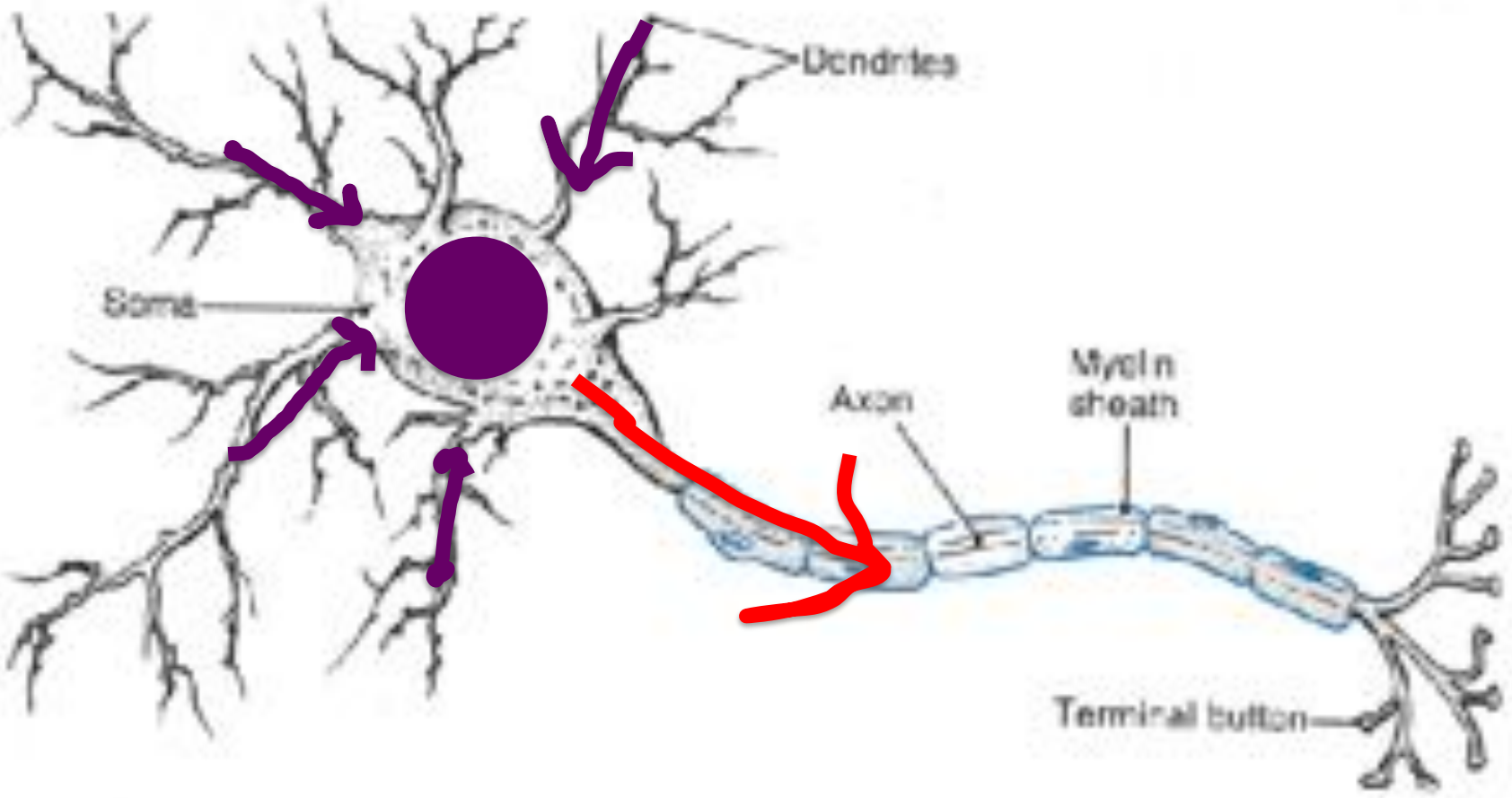
Neuron



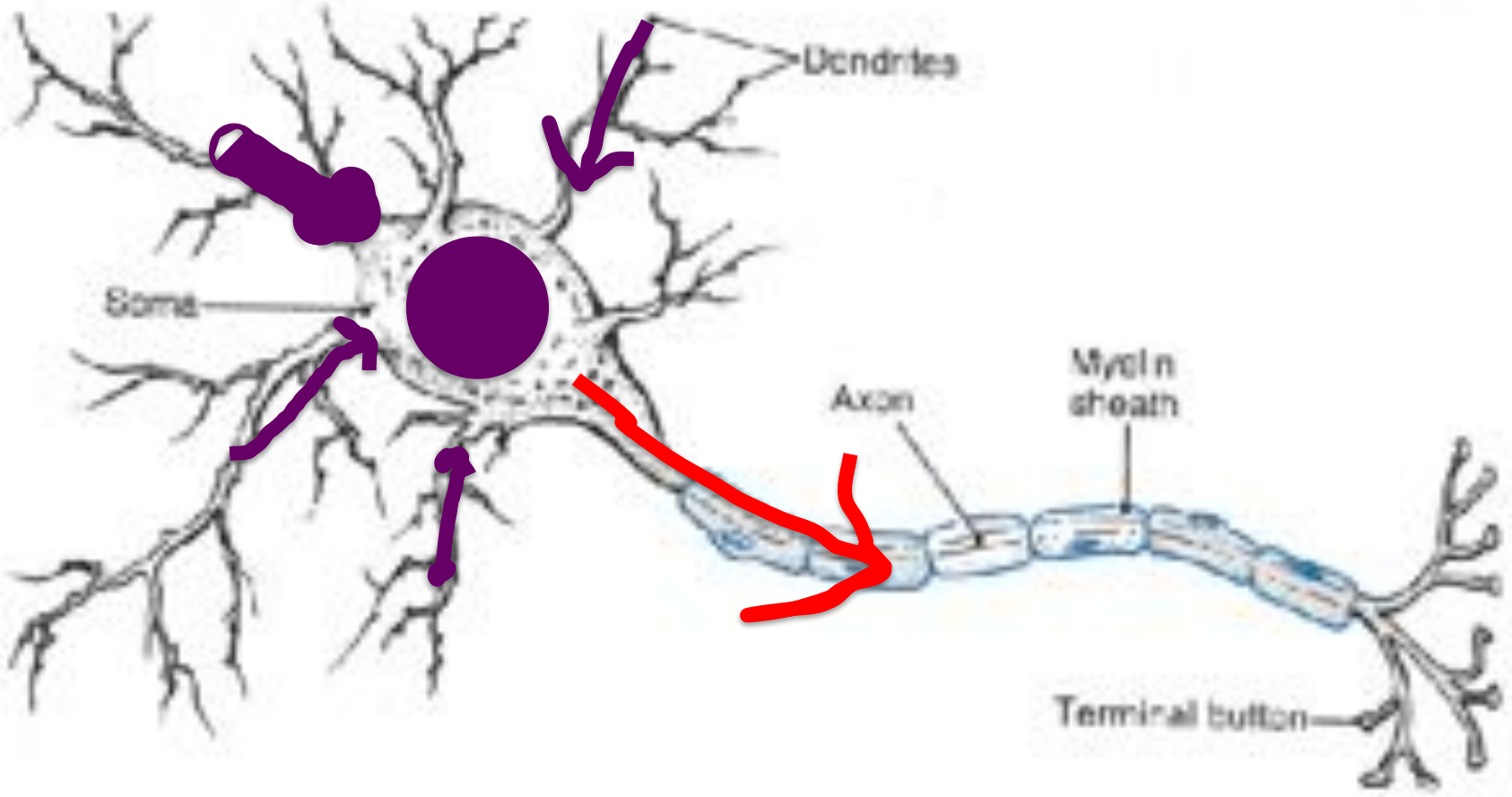
Neuron



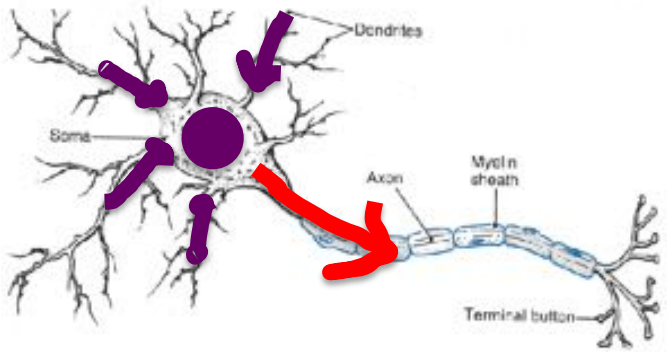
Neuron

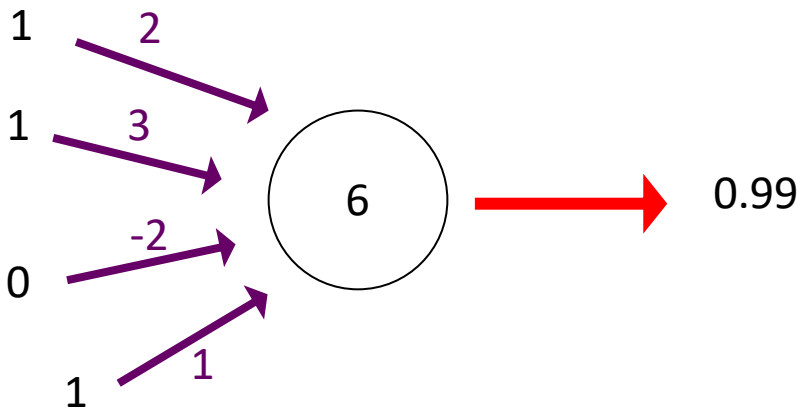


Some Inputs are More Important



Artificial Neuron

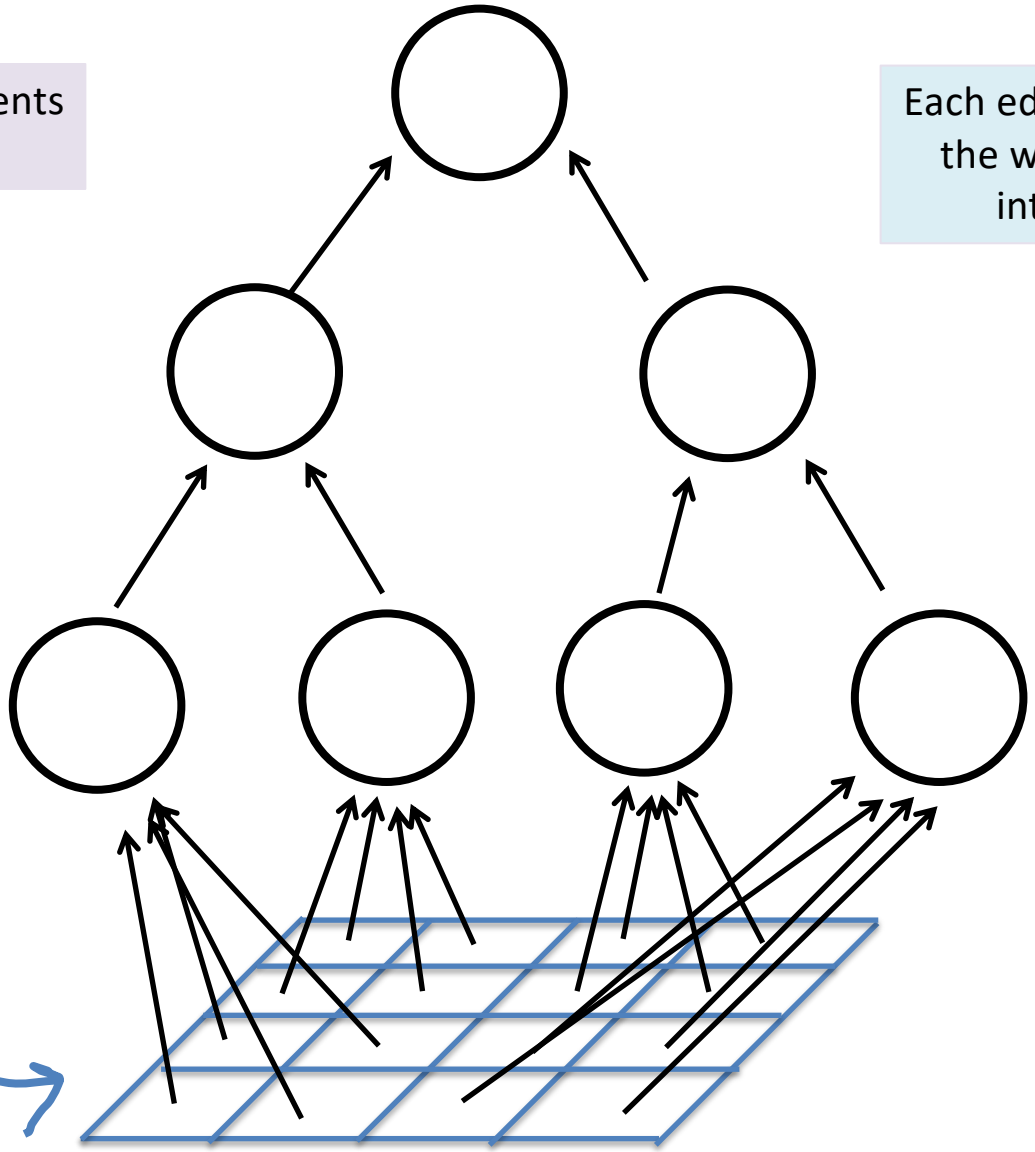




Neural Network

Each node represents a neuron

Each edge represents the weight of the interaction

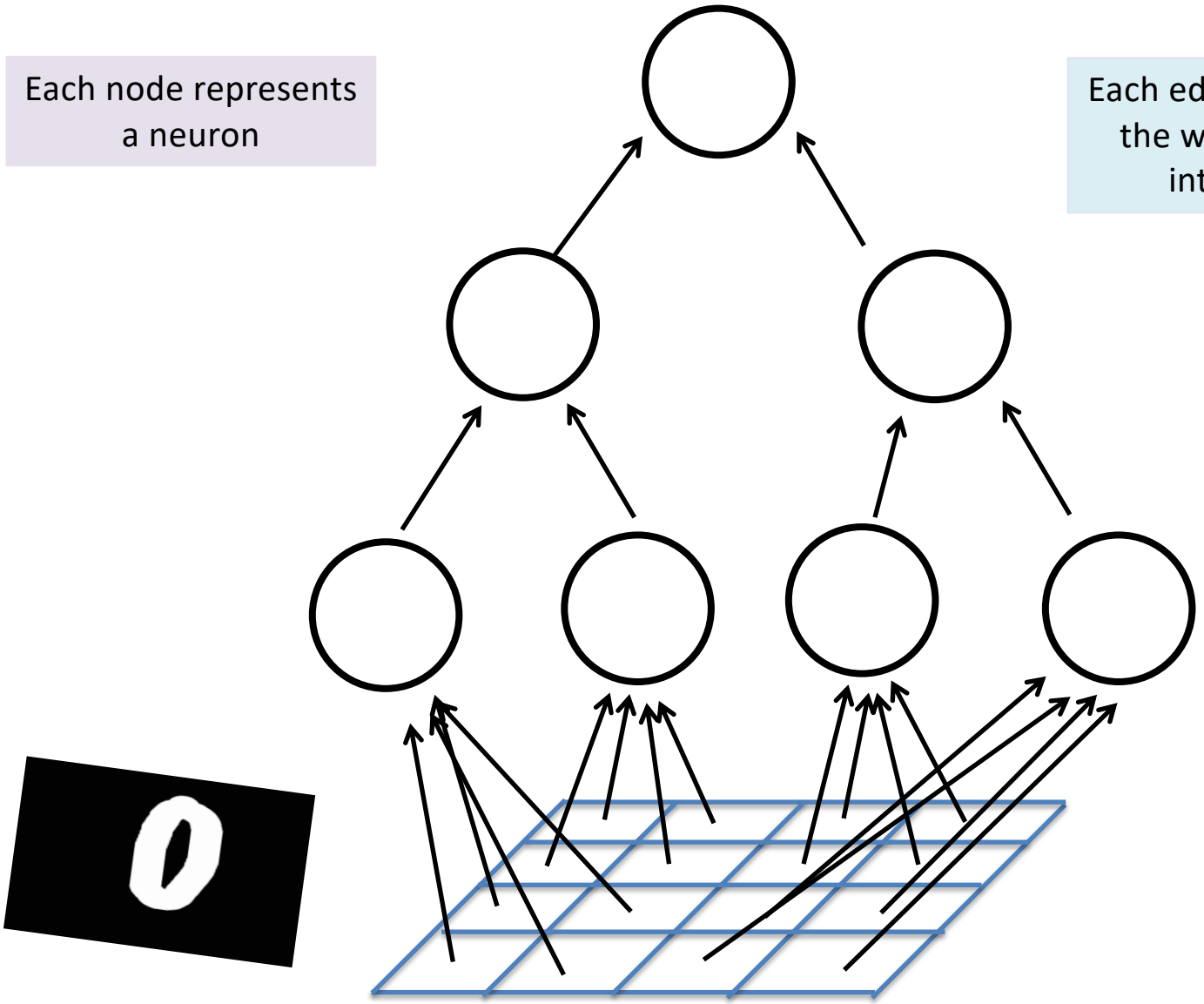


Pixels

Neural Network

Each node represents a neuron

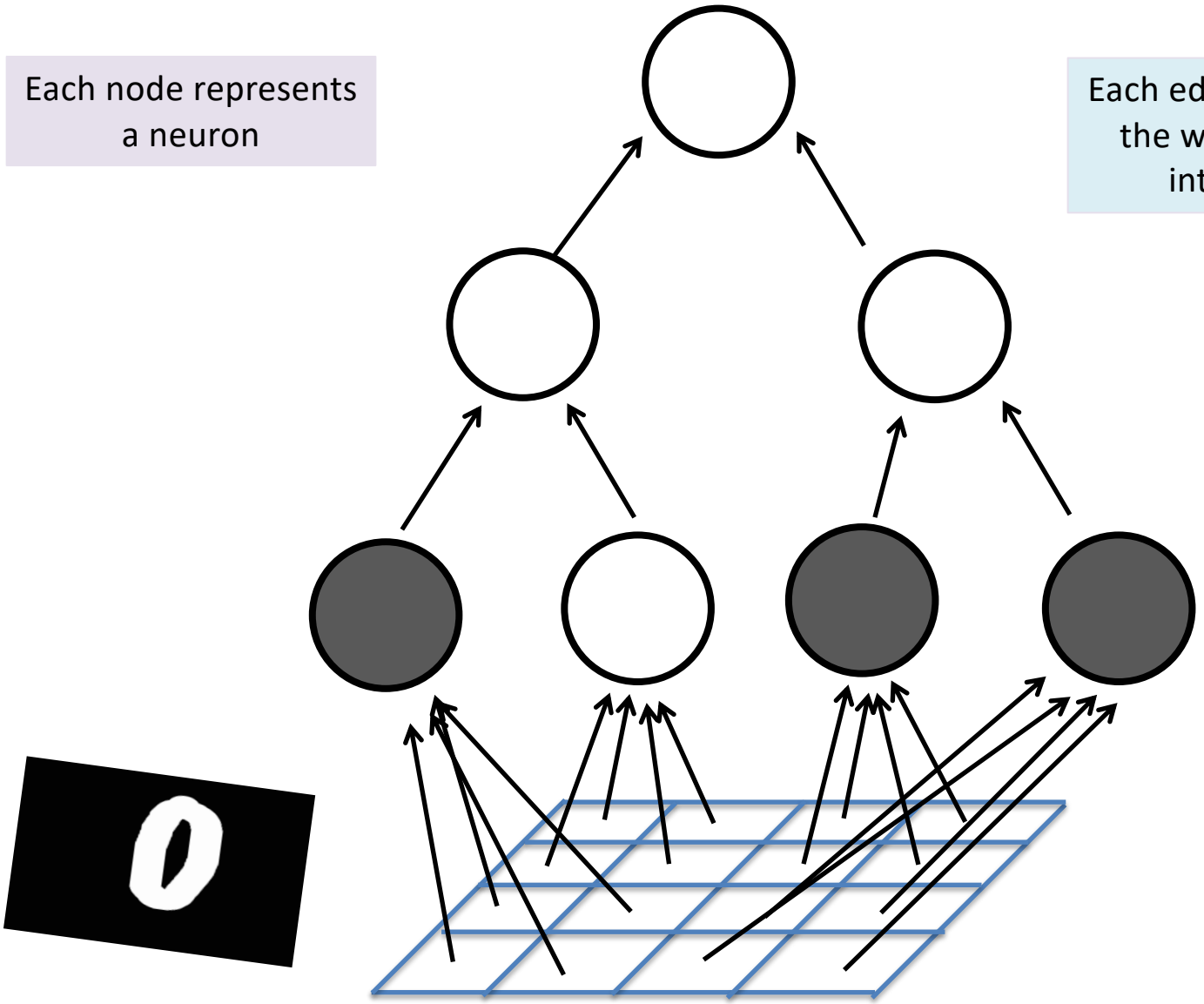
Each edge represents the weight of the interaction



Neural Network

Each node represents a neuron

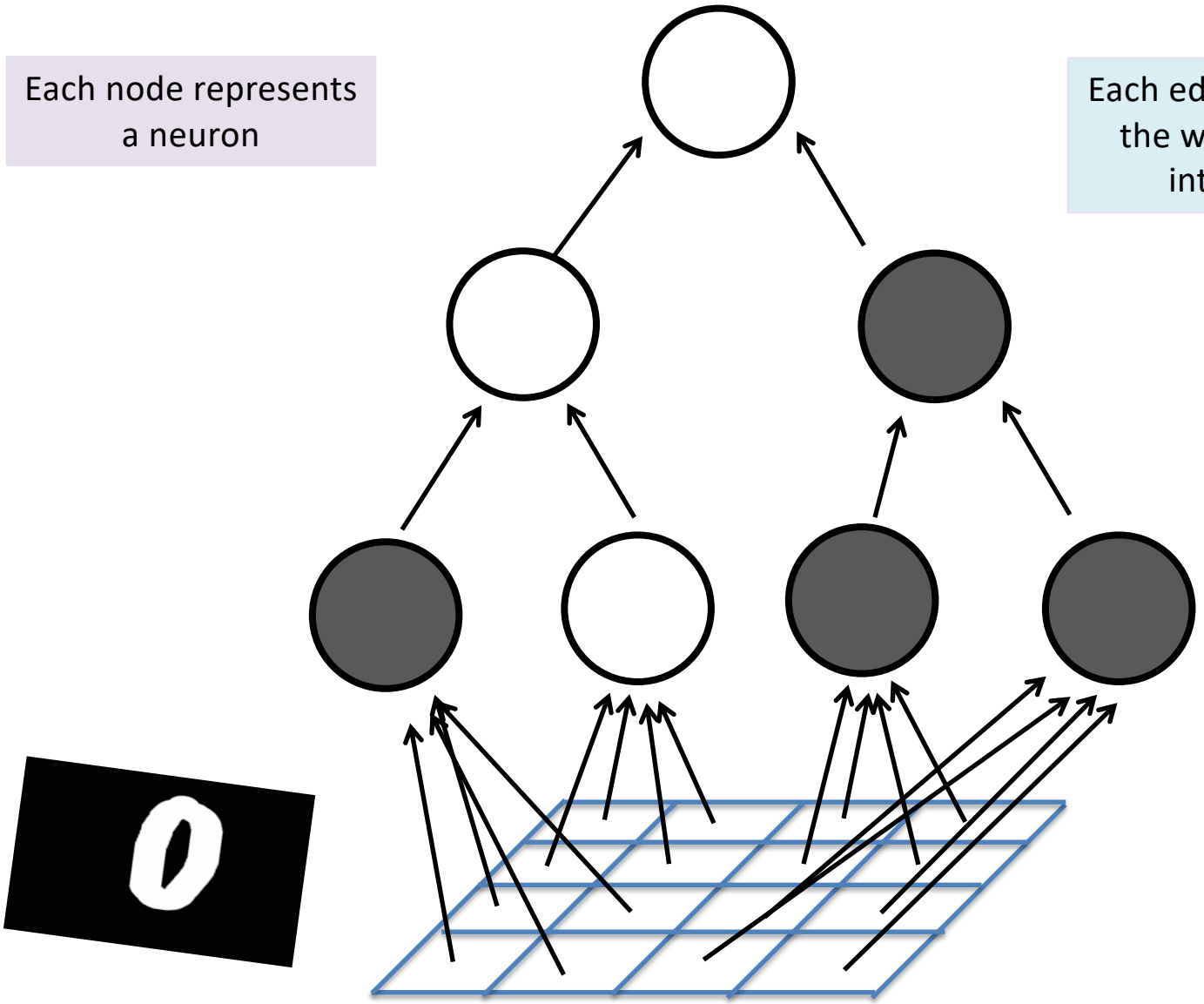
Each edge represents the weight of the interaction



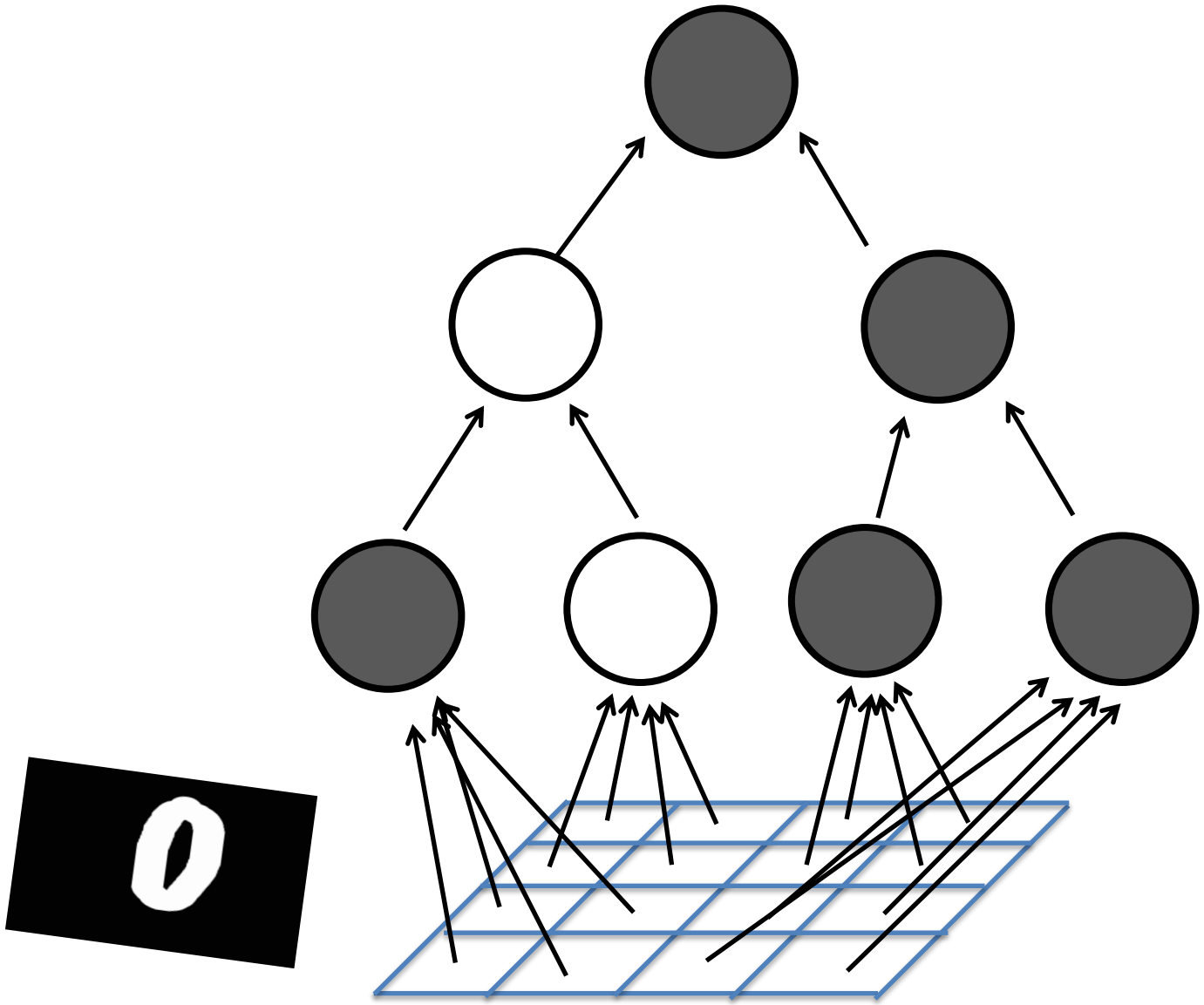
Neural Network

Each node represents a neuron

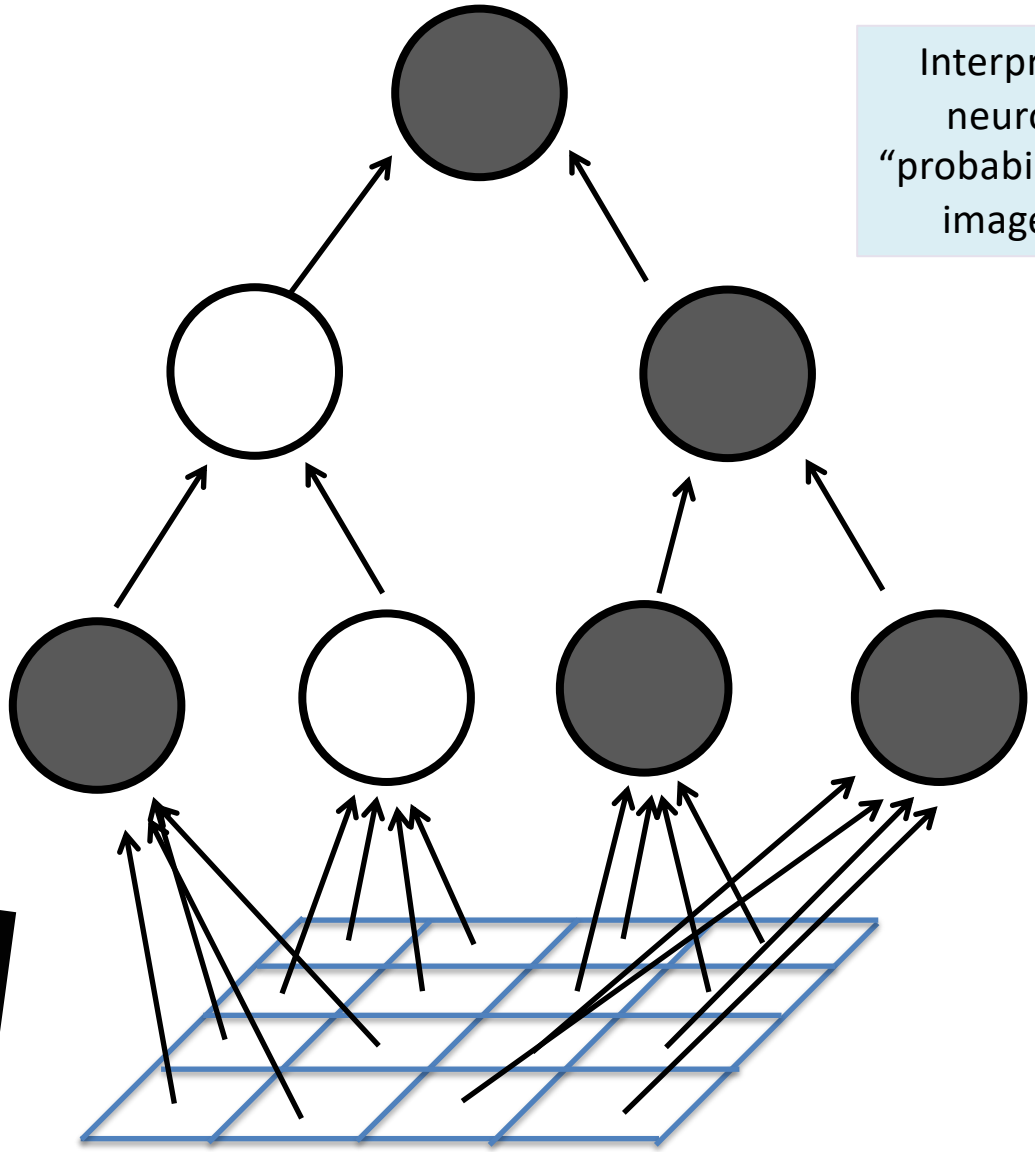
Each edge represents the weight of the interaction



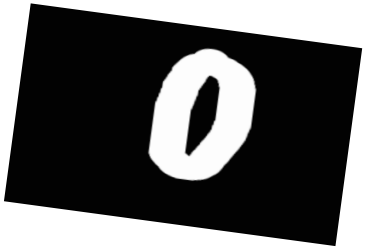
Neural Network



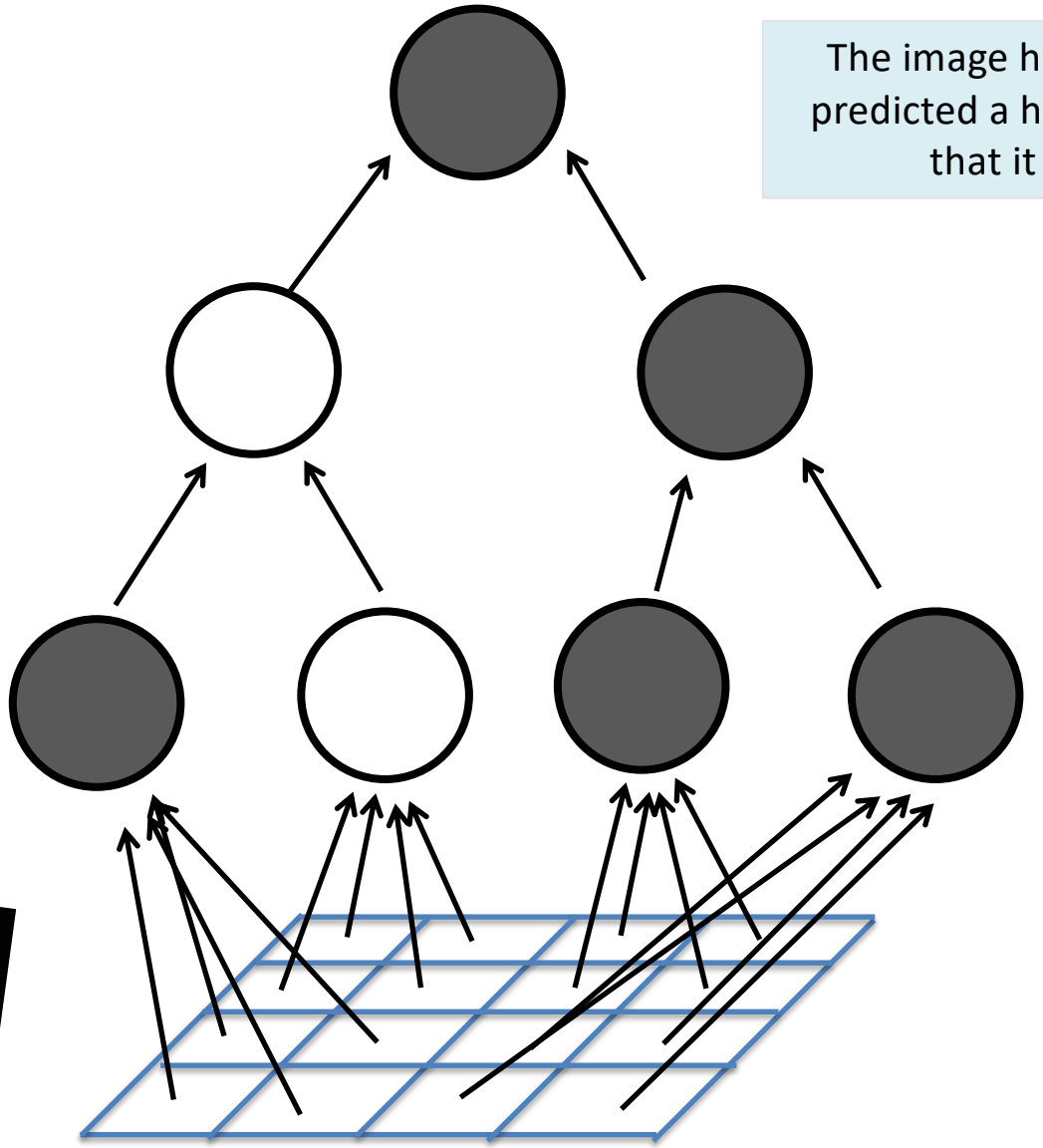
Neural Network



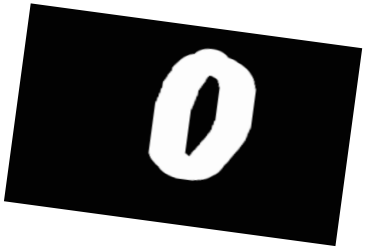
Interpret the last neuron as the "probability" that the image is of a 1



Neural Network



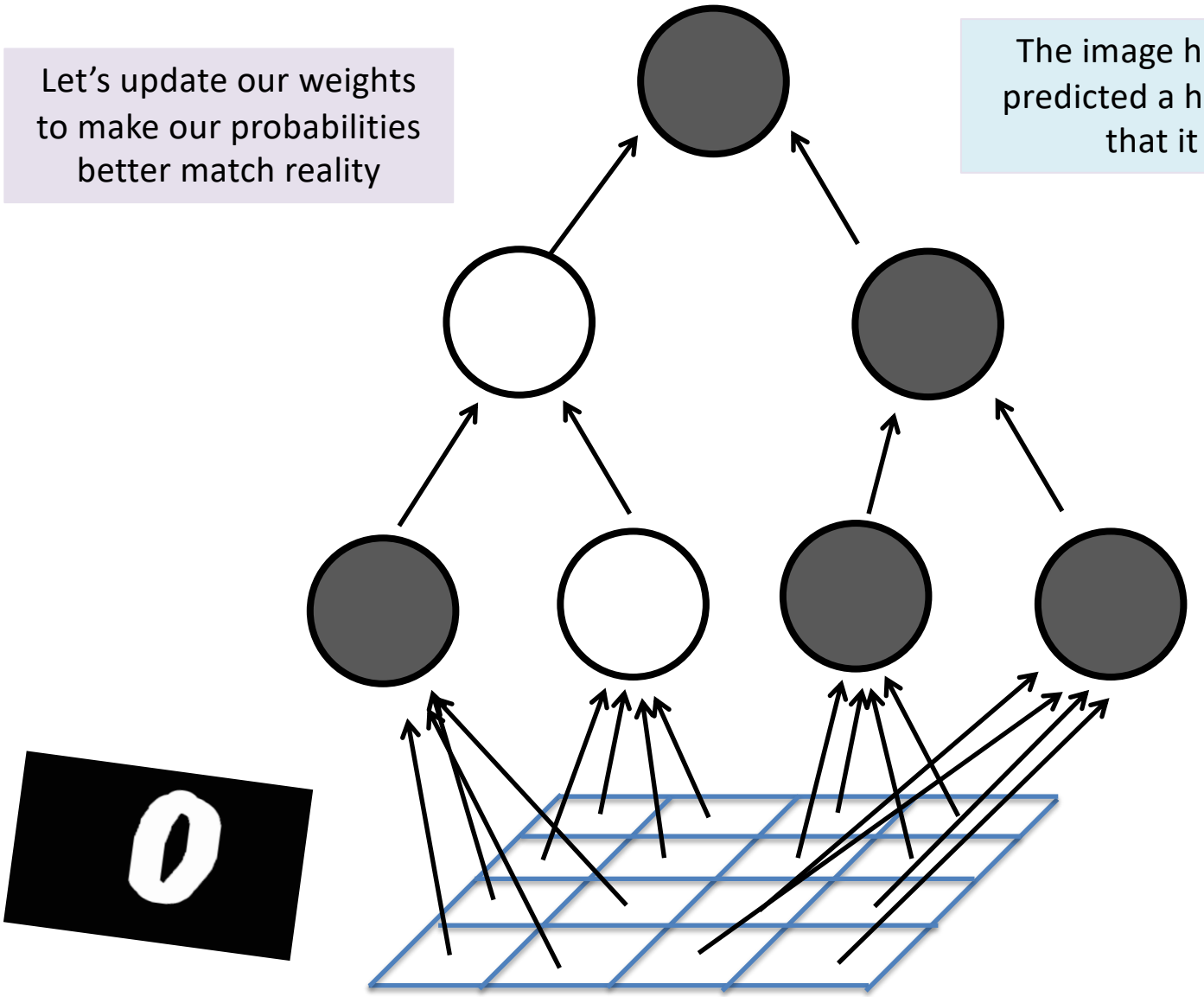
The image had a 0 but we predicted a high probability that it was a 1



Neural Network

Let's update our weights to make our probabilities better match reality

The image had a 0 but we predicted a high probability that it was a 1

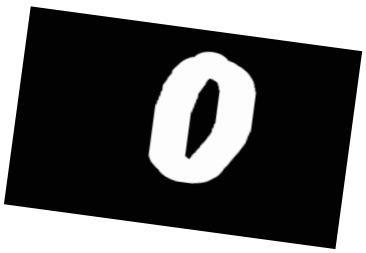


Neural Network

Let's update our weights to make our probabilities better match reality

The image had a 0 but we predicted a high probability that it was a 1

Update the parameter at each connection



Change network to max probability of the right answer

$$P(Y = 1|X = \mathbf{x}) = \hat{\mathbf{y}} \qquad \hat{y} = \sigma\left(\sum_{j=0}^{m_h} \mathbf{h}_j \theta_j^{(\hat{y})}\right)$$

For one datum

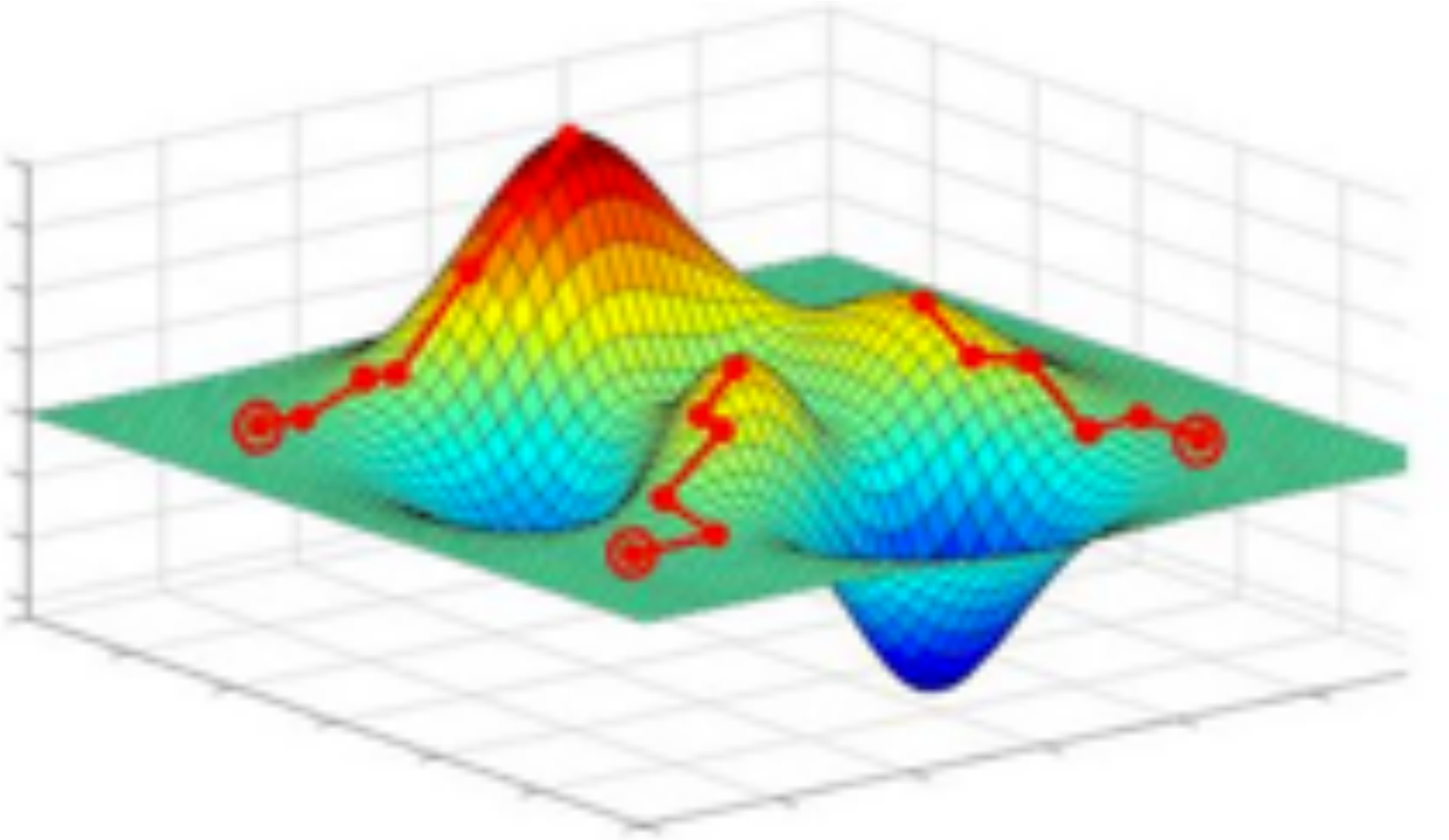
$$P(Y = y|X = \mathbf{X}) = (\hat{y})^y (1 - \hat{y})^{1-y}$$

For IID data

$$L(\theta) = \prod_{i=1}^{n} P(Y = y^{(i)}|X = \mathbf{x}^{(i)})$$

$$= \prod_{i=1}^{n} (\hat{y}^{(i)})^{y^{(i)}} \cdot \left[1 - (\hat{y}^{(i)})\right]^{(1-y^{(i)})}$$

Gradient Ascent



Walk uphill and you will find a local maxima
(if your step size is small enough)

Gradient of output layer params

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

$$\hat{y} = \sigma\left(\sum_{j=0}^{m_h} \mathbf{h}_j \theta_j^{(\hat{y})}\right)$$

$$\frac{\partial \hat{y}}{\partial \theta_i^{(\hat{y})}} = \sigma\left(\sum_{j=0}^{m_h} \mathbf{h}_j \theta_j^{(\hat{y})}\right)\left[1 - \sigma\left(\sum_{j=0}^{m_h} \mathbf{h}_j \theta_j^{(\hat{y})}\right)\right] \cdot \frac{\partial}{\partial \theta_i^{(\hat{y})}} \sum_{j=0}^{m_h} \mathbf{h}_j \theta_j^{(\hat{y})}$$

$$= \hat{y}[1 - \hat{y}] \cdot \frac{\partial}{\partial \theta_i^{(\hat{y})}} \sum_{j=0}^{m_h} \mathbf{h}_j \theta_j^{(\hat{y})}$$

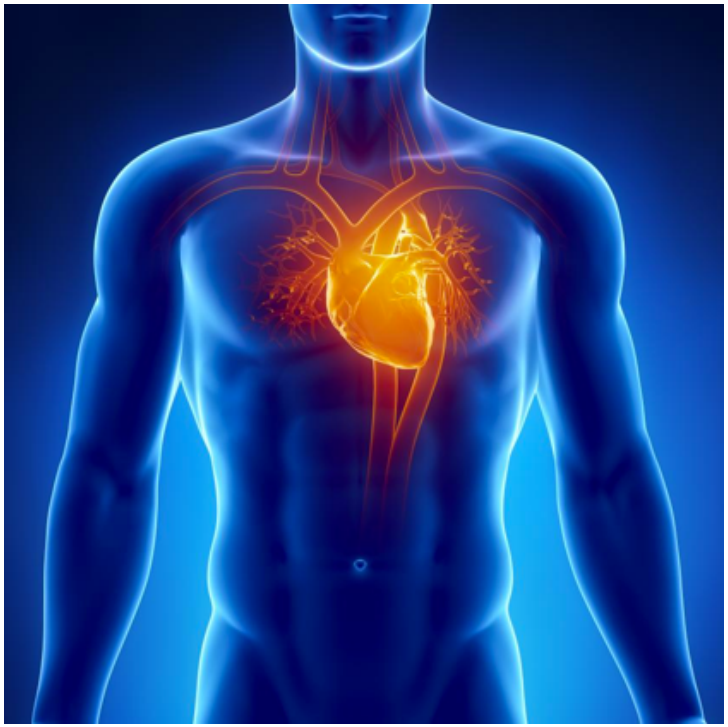
$$= \hat{y}[1 - \hat{y}] \cdot h_i$$

You will be able to do this.

Where you will be by the end of class

CS View of Probability

Heart



Ancestry

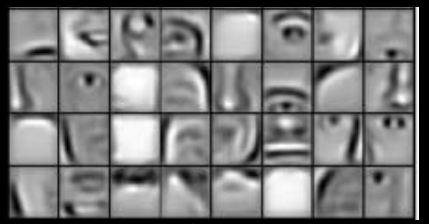
Netflix

When you train,
something really neat happens

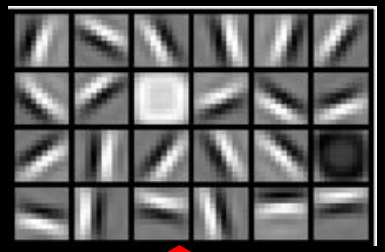
Visualize the Weights

Training set: Aligned images of faces.

object models



object parts (combination of edges)



edges

pixels

[Honglak Lee]

Google Brain

Google Brain

1 Trillion Artificial Neurons

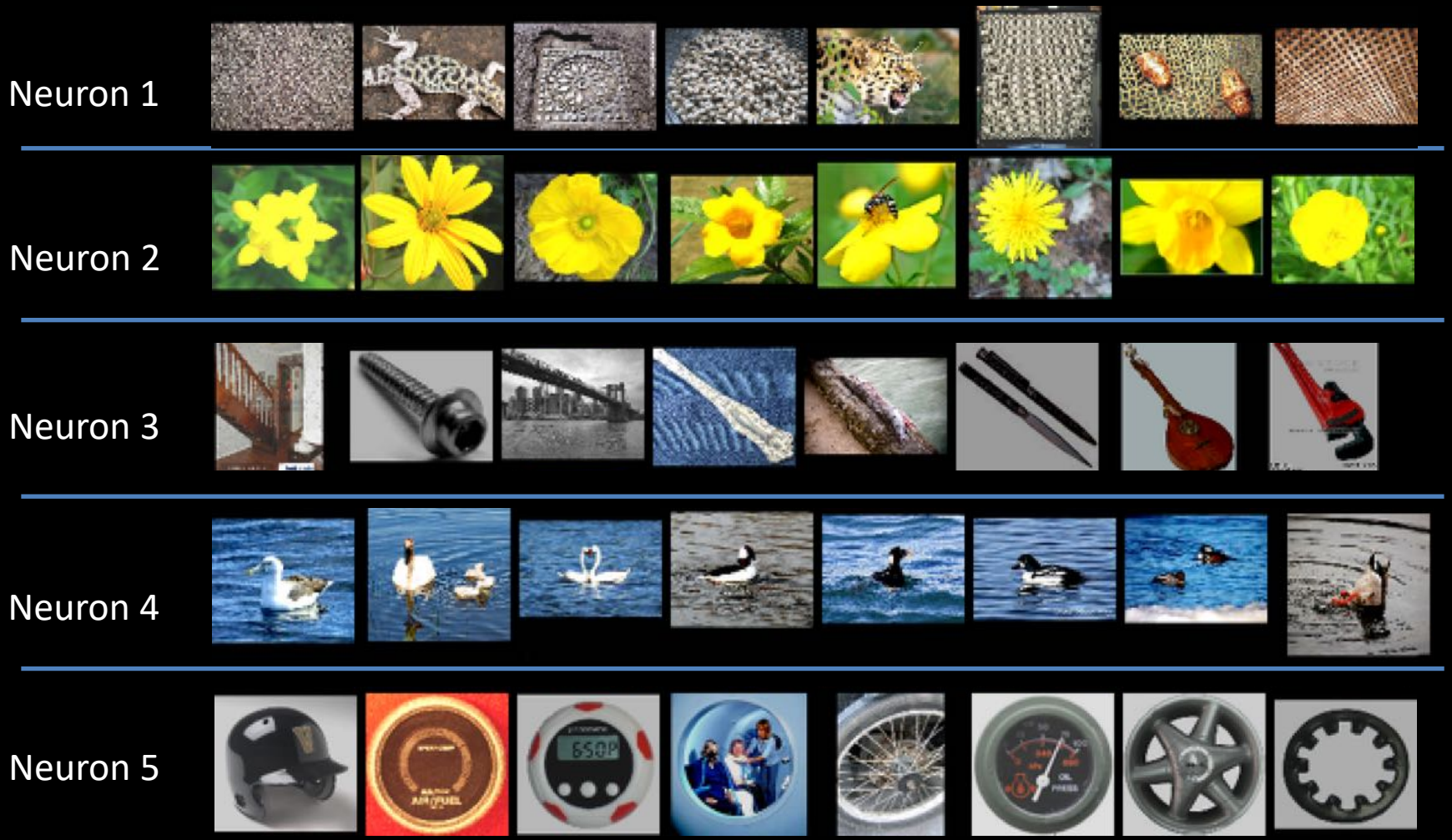
A Neuron That Fires When It Sees Cats

Top stimuli from the test set

Optimal stimulus
by numerical optimization

Le, et al., *Building high-level features using large-scale unsupervised learning*. ICML 2012

Other Neurons



Le, et al., *Building high-level features using large-scale unsupervised learning*. ICML 2012

ImageNet Classification

…
smoothhound, smoothhound shark, Mustelus mustelus
American smooth dogfish, Mustelus canis
Florida smoothhound, Mustelus norrisi
whitetip shark, reef whitetip shark, Triaenodon obseus
Atlantic spiny dogfish, Squalus acanthias
Pacific spiny dogfish, Squalus suckleyi
hammerhead, hammerhead shark
smooth hammerhead, Sphyrna zygaena
smalleye hammerhead, Sphyrna tudes
shovelhead, bonnethead, bonnet shark, Sphyrna tiburo
angel shark, angelfish, Squatina squatina, monkfish
electric ray, crampfish, numbfish, torpedo
smalltooth sawfish, Pristis pectinatus
guitarfish
roughtail stingray, Dasyatis centroura
butterfly ray
eagle ray
spotted eagle ray, spotted ray, Aetobatus narinari
cownose ray, cow-nosed ray, Rhinoptera bonasus
manta, manta ray, devilfish
Atlantic manta, Manta birostris
devil ray, Mobula hypostoma
grey skate, gray skate, Raja batis
little skate, Raja erinacea
…

Stingray

Mantaray

ImageNet Classification

0.005%	1.5%	?
Random guess	Pre Neural Networks	GoogLeNet

Le, et al., *Building high-level features using large-scale unsupervised learning*. ICML 2012

ImageNet Classification

0.005%
Random guess

1.5%
Pre Neural Networks

43.9%
GoogLeNet

Szegedy et al, Going Deeper With Convolutions, CVPR 2015

ImageNet Classification

0.005%

Random guess

1.5%

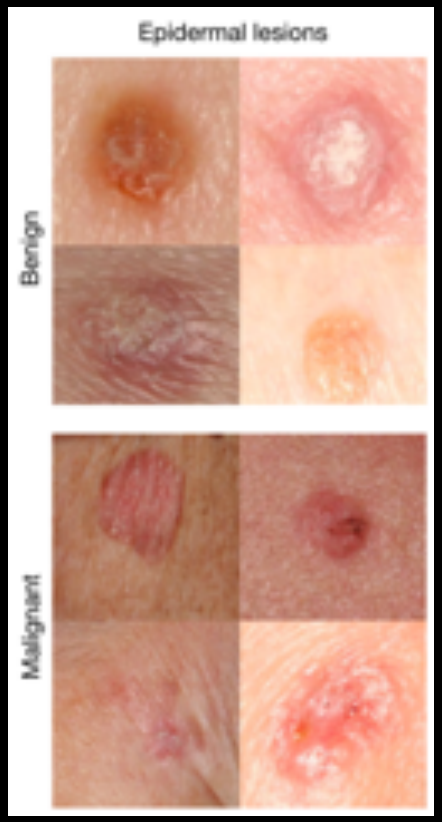
Pre Neural Networks

82.7%

NASNet

https://arxiv.org/pdf/1707.07012.pdf

Where is this useful?



A machine learning algorithm performs **better than** the best dermatologists.

Developed in 2017, at Stanford.

Esteva, Andre, et al. "Dermatologist-level classification of skin cancer with deep neural networks." *Nature* 542.7639 (2017): 115-118.

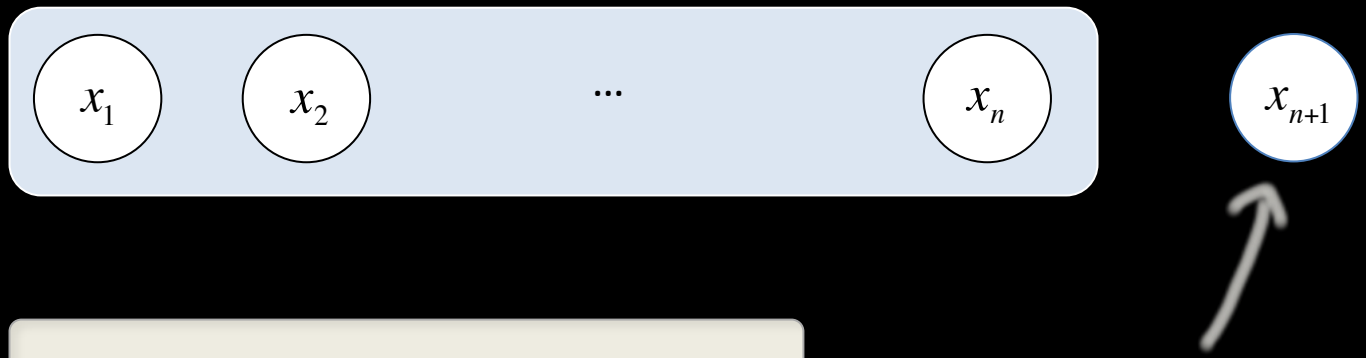
Autonomous Scale Education

KHAN ACADEMY

50 thousand students

Knowledge Tracing for Feedback

Given *n* historical answers:



Predict the next one

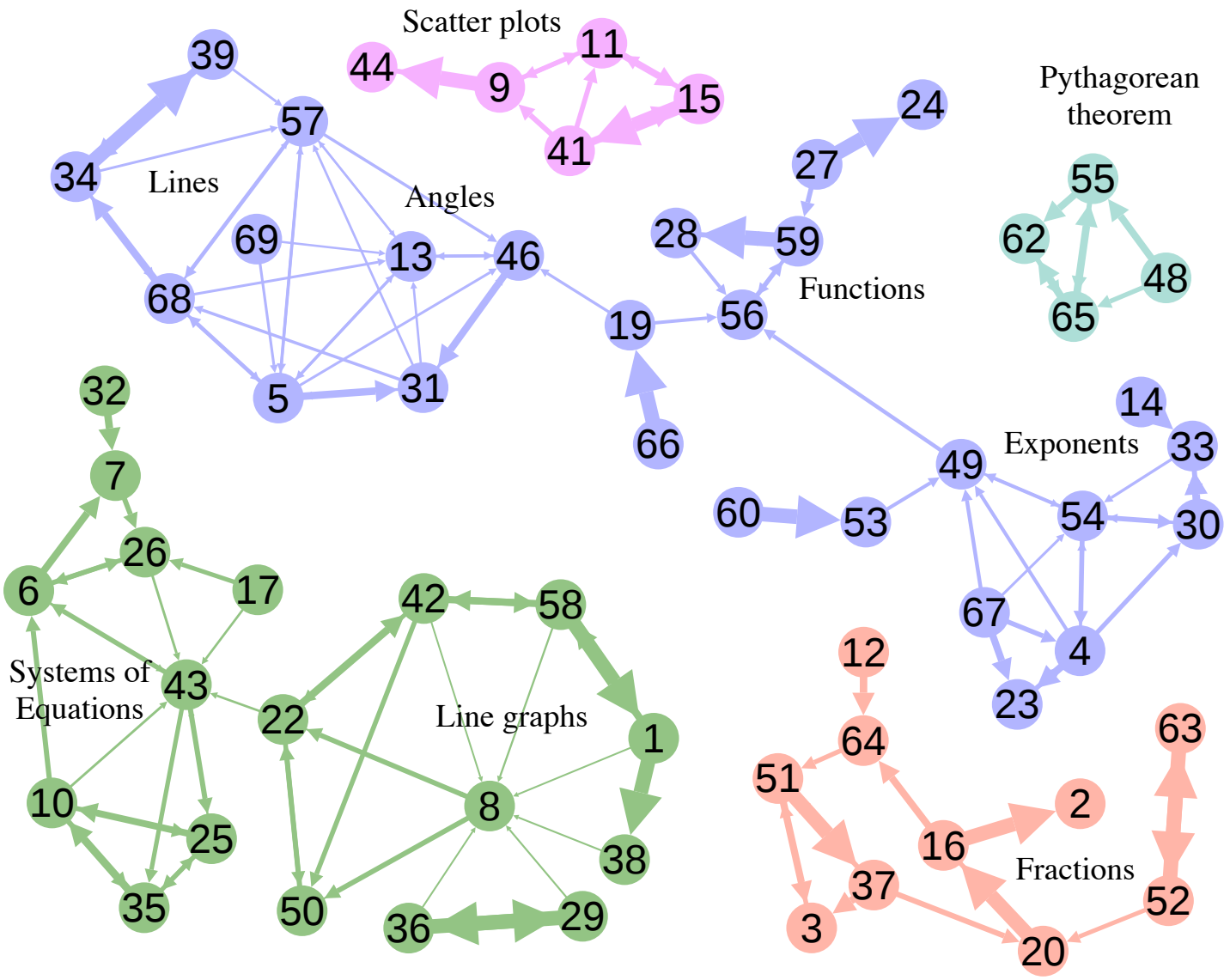
Answer is a tuple:

$$x_i = \{q_i, a_i\}$$

Question

Student response

Learns Concept Relationships



Not once, but twice, AI was revolutionized by people who understood probability theory.

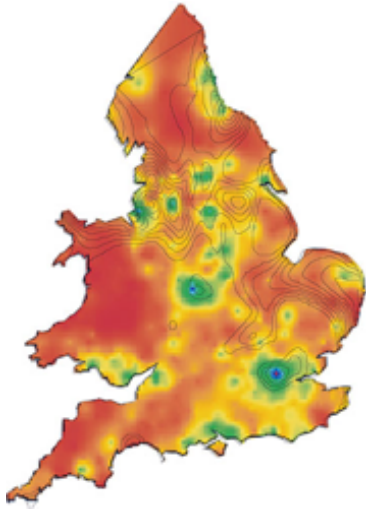
End of Story

Except it isn’t the end of the story…

Probability is more than just machine learning

Abundance of Important Problems

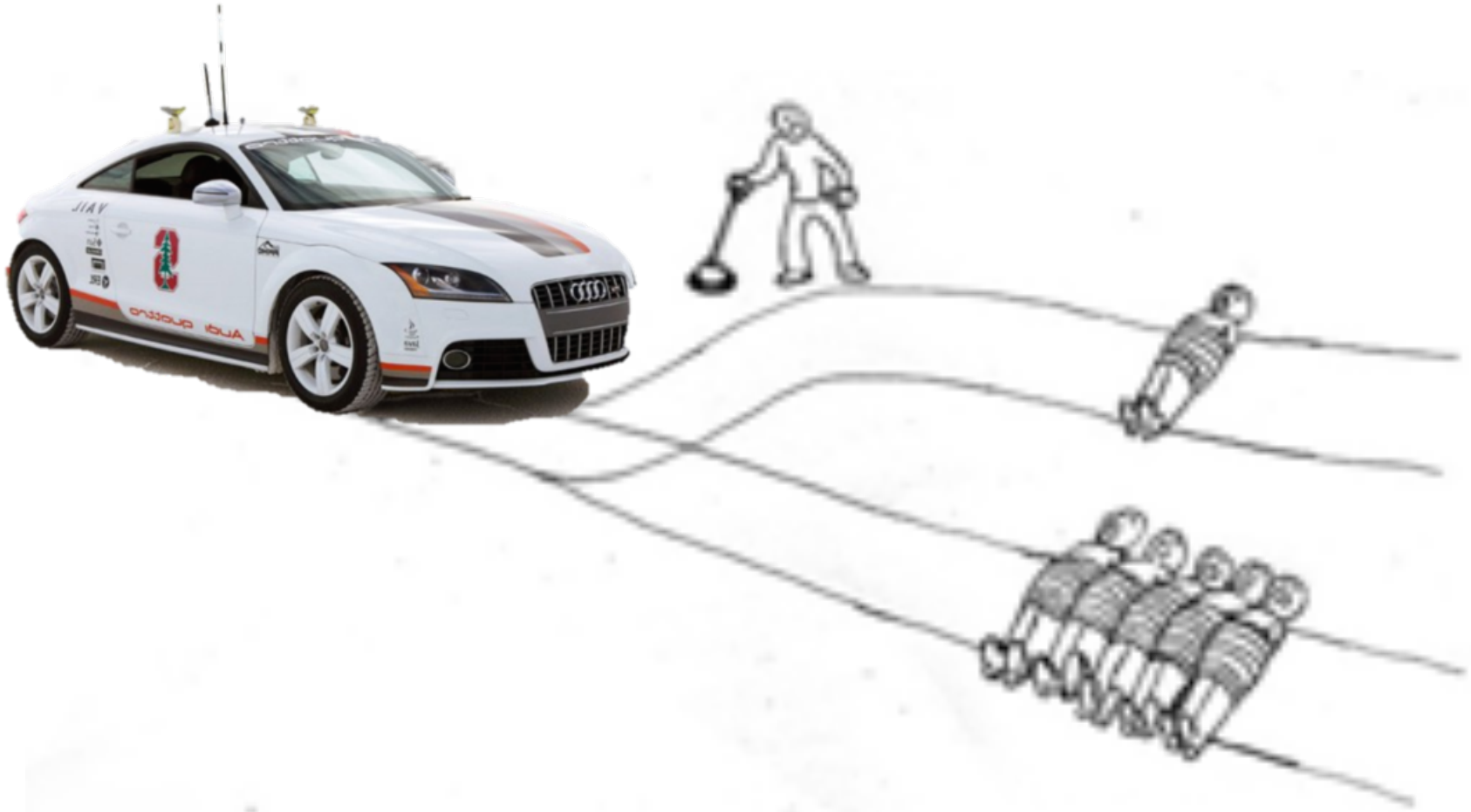
Medicine and Probability



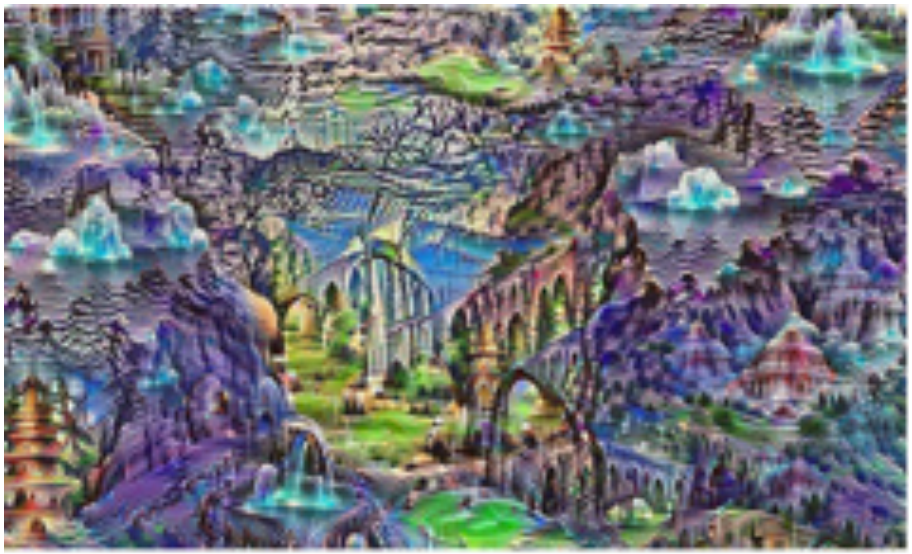
Autocomplete

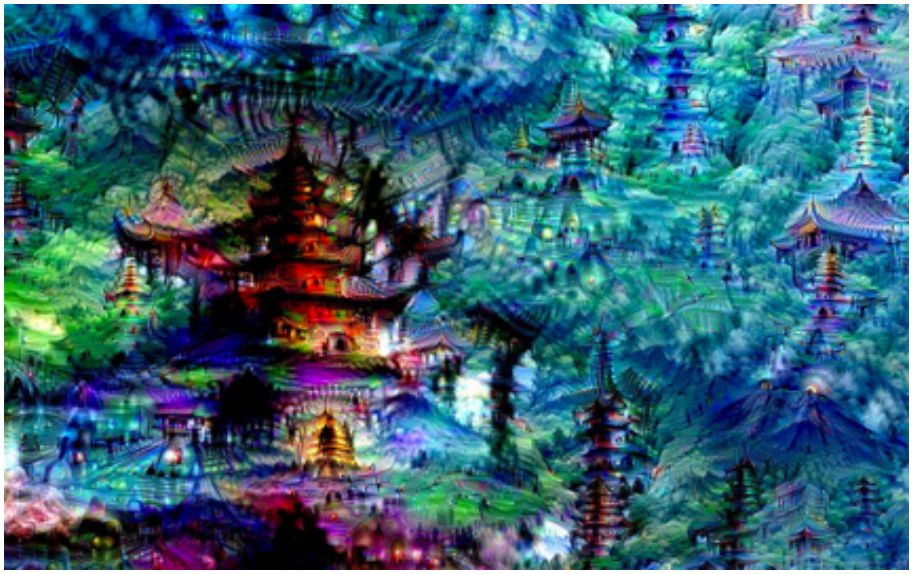
Probability in Practice

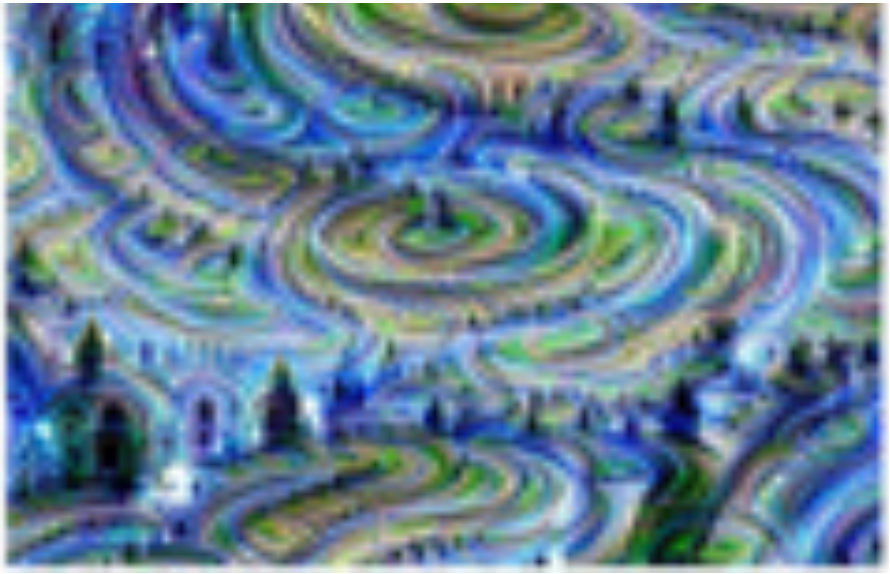
Philosophy and Probability



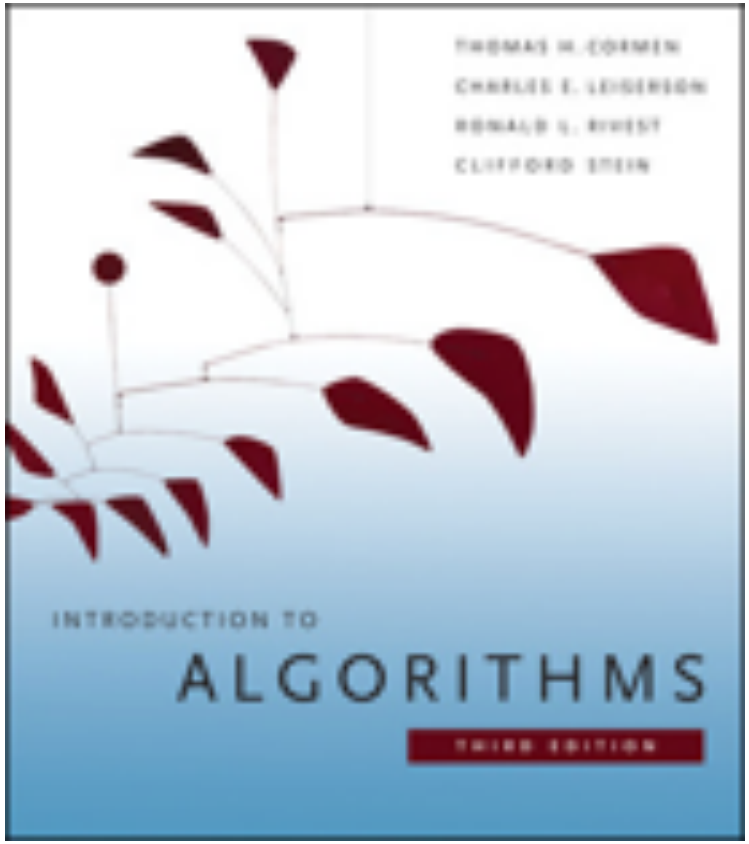
Art and Probability



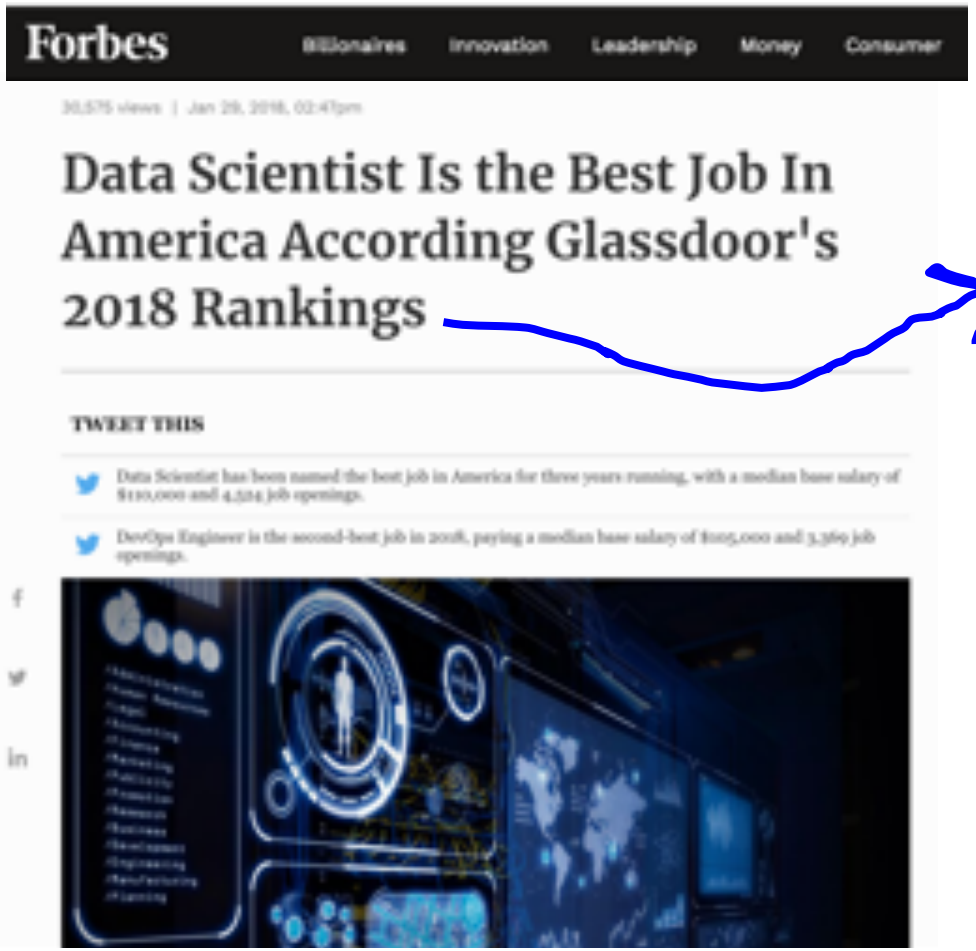




Probabilistic Analysis of Algorithms



#1 Most Desired Skill in Industry



Job Score is based on:

- Earning potential
- Number of jobs
- Job satisfaction rating

"Data science and machine learning are generating more jobs than candidates right now, making these two areas the *fastest growing employment areas*."

9.8 times more jobs than five years ago.

LinkedIn's 2017 U.S. Emerging Jobs Report

#1 Most Desired Skill in Academia

Most CS PhD students list their highest desiderata upon graduation as:

"Better understanding of probability"

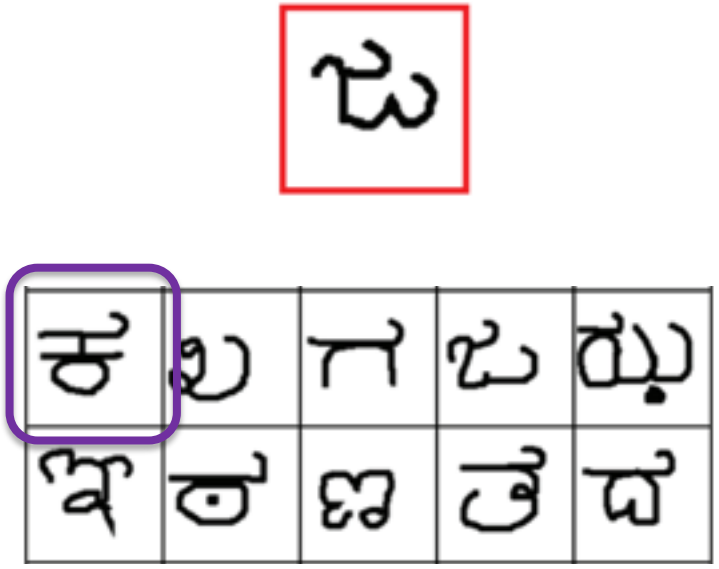
Open Problem: One Shot Learning

B Lake, R Salakhutdinov, J Tenenbaum. Science 2015.
Human-level concept learning through probabilistic program induction.

Current deep learning methods are not enough to move the needle as far as we want, **especially on socially relevant problems** that often do not have the benefit of massive public datasets. The best new ideas are coming from probability theory

Open Problem: One Shot Learning

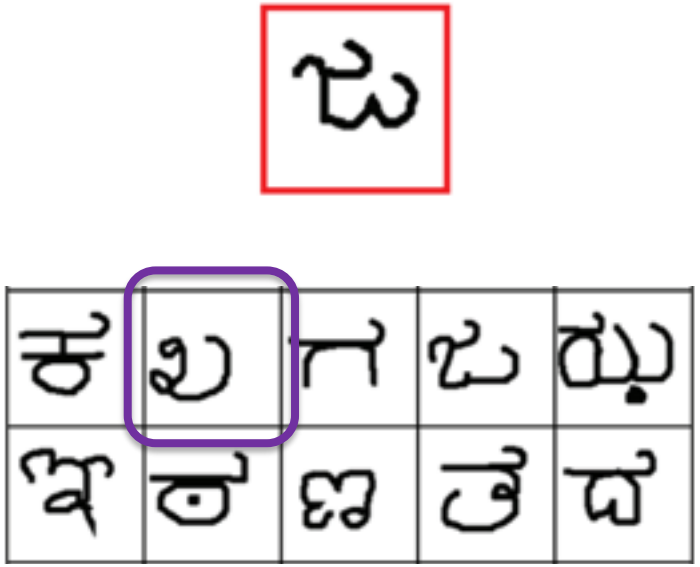
B Lake, R Salakhutdinov, J Tenenbaum. Science 2015.
Human-level concept learning through probabilistic program induction.



Current deep learning methods are not enough to move the needle as far as we want, **especially on socially relevant problems** that often do not have the benefit of massive public datasets. The best new ideas are coming from probability theory

Open Problem: One Shot Learning

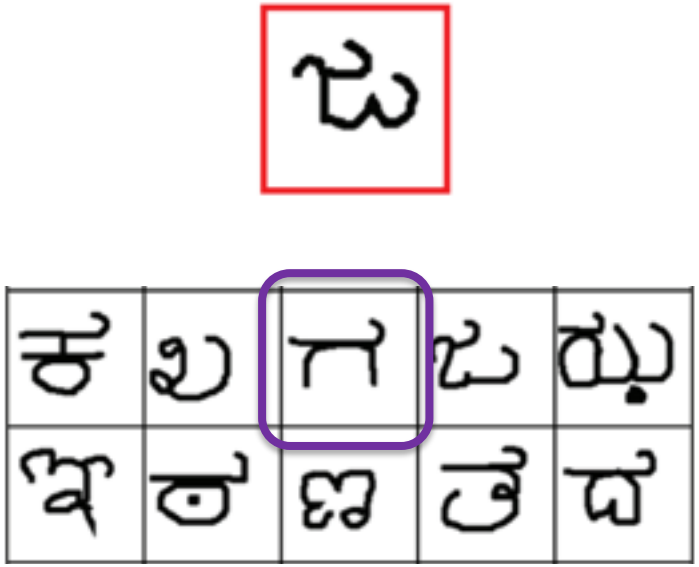
B Lake, R Salakhutdinov, J Tenenbaum. Science 2015.
Human-level concept learning through probabilistic program induction.



Current deep learning methods are not enough to move the needle as far as we want, **especially on socially relevant problems** that often do not have the benefit of massive public datasets. The best new ideas are coming from probability theory

Open Problem: One Shot Learning

B Lake, R Salakhutdinov, J Tenenbaum. Science 2015.
Human-level concept learning through probabilistic program induction.



Current deep learning methods are not enough to move the needle as far as we want, **especially on socially relevant problems** that often do not have the benefit of massive public datasets. The best new ideas are coming from probability theory

Open Problem: One Shot Learning

B Lake, R Salakhutdinov, J Tenenbaum. Science 2015.
Human-level concept learning through probabilistic program induction.

Current deep learning methods are not enough to move the needle as far as we want, **especially on socially relevant problems** that often do not have the benefit of massive public datasets. The best new ideas are coming from probability theory

Foundation for your future

But its not always intuitive

Zika Test

A patient has a positive Zika test.

What is the probability they have zika?

- *0.8% of people have zika*
- *Test has 90% positive rate for people with zika*
- *Test has 7% positive rate for people without zika*

The right answer is 9%

Probability = Important + Needs Study

Delayed gratification

CS109 View of Probability

Teach you how to write programs
that most people are not able to write.

Stretch!

SCPD Attendance Code in lecture video.

Lets dive in…

Counting