INTRODUCTION TO DEEP LEARNING

ABOUT ME

WHO AM I?

- Visiting Assistant Professor
- Education:
  - BS at Oregon State University
  - Masters/PhD at the University of Massachusetts Amherst
    - Reinforcement Learning
  - Postdoc at the University of Alberta
- I started my PhD in 2015 and thought it was the peak deep learning craze

ABOUT ME

WHAT I BELIEVE

- I actually enjoy teaching
- I expect students to try
- I expect everyone to make mistakes
    - Let’s learn from them and fix them when possible

ABOUT ME

WHAT I BELIEVE

- I actually enjoy teaching
- I expect students to try
- I expect everyone to make mistakes
  - Let’s learn from them and fix them when possible

I want to make this a **safe** and **inclusive** environment for everyone to learn.

We should all strive to be **respectful** to eachother.

COURSE INFO

CONTACT AND MORE

Course website: Canvas (merged class for 1678 and 2078 coming)

Contact me: Scott Jordan (scott.jordan@pitt.edu)

- Include [CS1678] in the beginning of your subject

Office: 6105 Sennott Square

Class Mon/Wed 9:30am—10:45am

TA: TBD

Class forum: edstem.org (Check canvas for instructions)

Top Hat: In-class quizzes

WHAT IS DEEP LEARNING?

OVERVIEW

WHAT IS DEEP LEARNING?

OVERVIEW

— Historically, multi-layered neural networks

— “Deep Learning” is (somewhat) a marketing move in 2015

“Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction.” — LeCun, Bengio, and Hinton (2015).

— The optimization of flexible and differentiable function approximators

WHERE IS IT USED

EXAMPLES

— (1998) Handwritten character recognition an check processing

— (Now) Everywhere

WHERE IS IT USED

EXAMPLES

— (1998) Ha

— (Now) Eve

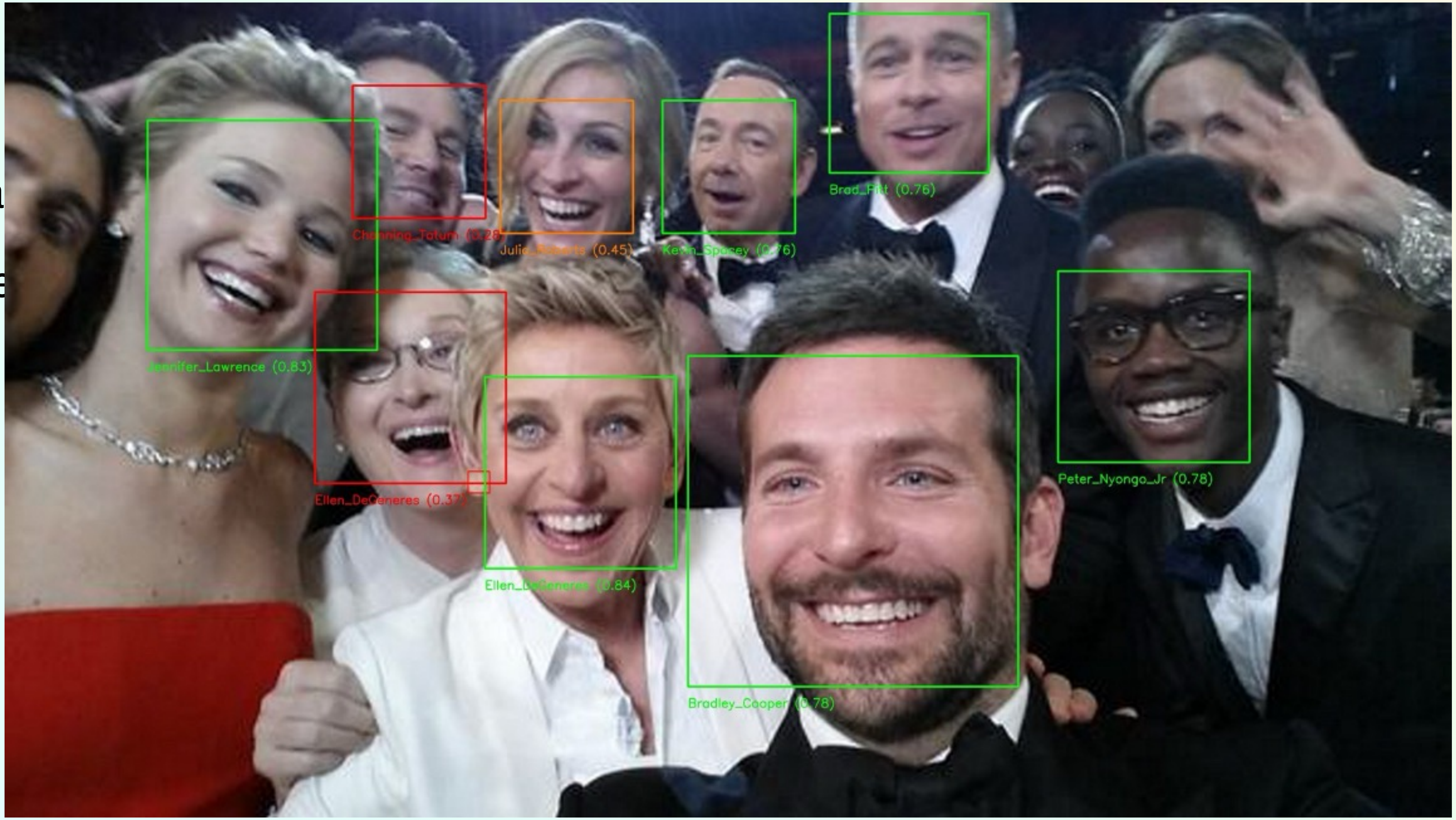
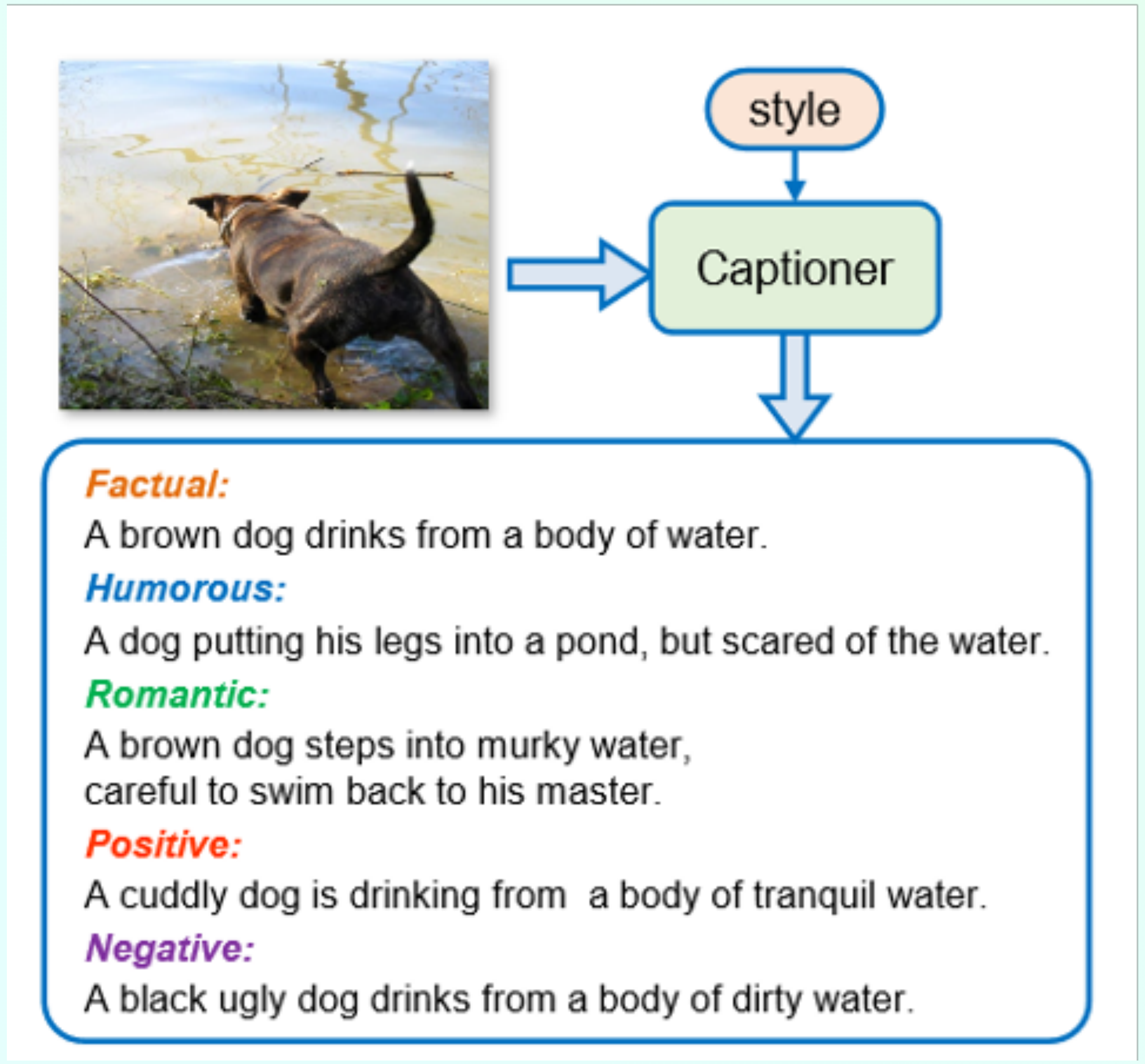
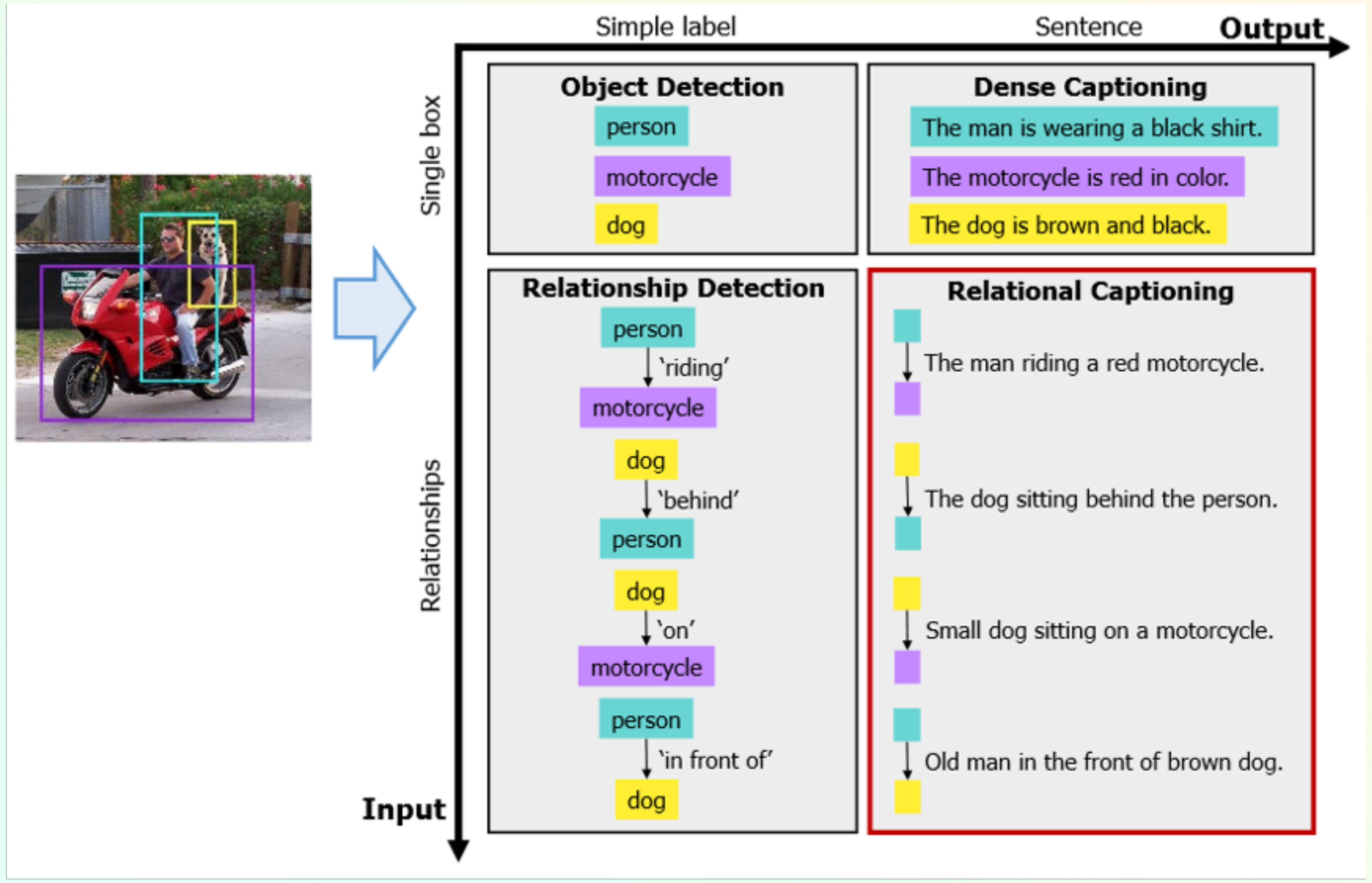


IMAGE CAPTION GENERATION

DL EXAMPLES

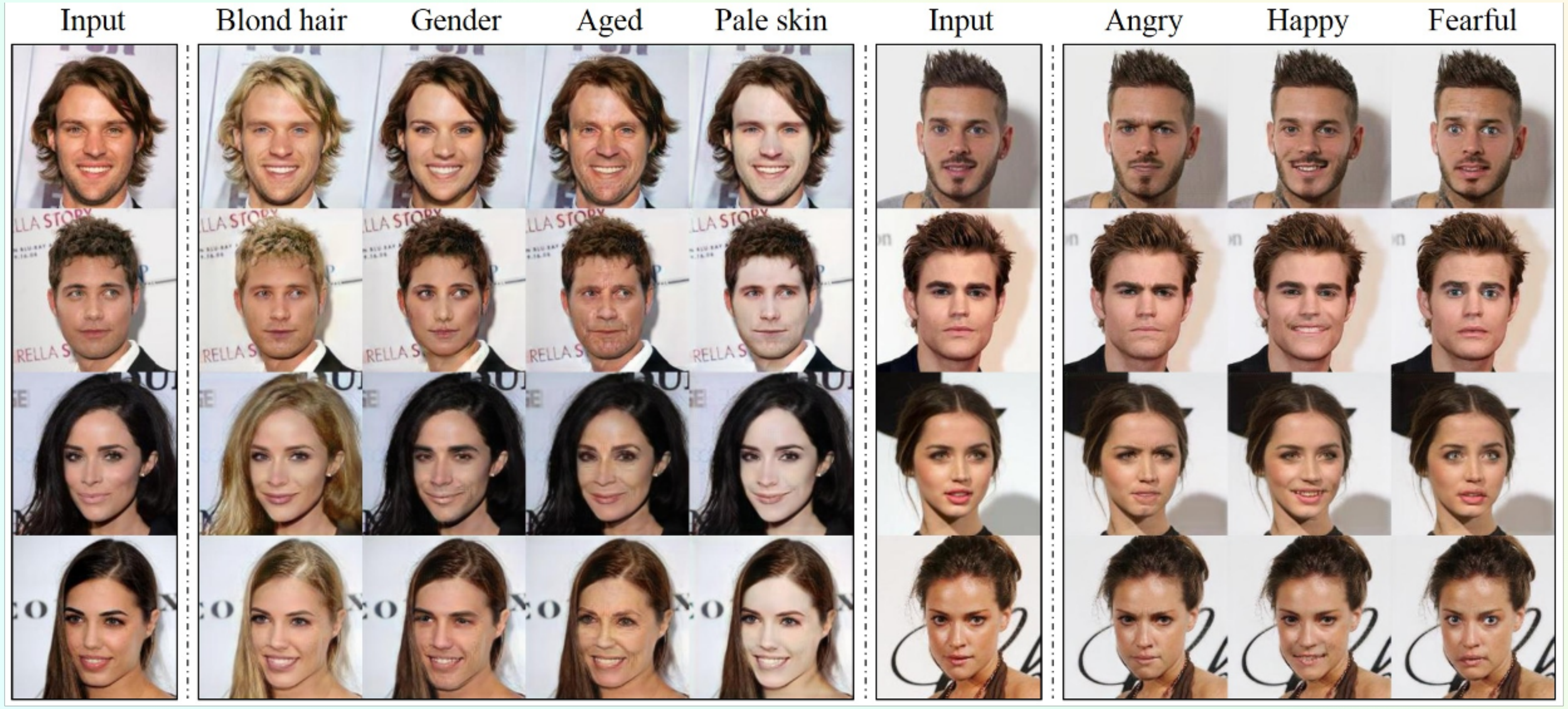




http://openaccess.thecvf.com/content_CVPR_2019/papers/Guo_MSCap_Multi-Style_Image_Captioning_With_Unpaired_Stylized_Text_CVPR_2019_paper.pdf
http://openaccess.thecvf.com/content_CVPR_2019/papers/Kim_Dense_Relational_Captioning_Triple-Stream_Networks_for_Relationship-Based_Captioning_CVPR_2019_paper.pdf

IMAGE GENERATION

DL EXAMPLES



Choi et al., "StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation", CVPR 2018

FAX NEWS GENERATION

DL EXAMPLES

https://www.youtube.com/watch?v=-QvIX3cY4lc

MACHINE TRANSLATION

DL EXAMPLES

SPEACH RECOGNITION

DL EXAMPLES

- “Hey [Siri,Alex,Google]”

TEXT GENERATIOIN

DL EXAMPLES

```
PANDARUS:
Alas, I think he shall be come approached and the day
When little srain would be attain'd into being never fed,
And who is but a chain and subjects of his death,
I should not sleep.

Second Senator:
They are away this miseries, produced upon my soul,
Breaking and strongly should be buried, when I perish
The earth and thoughts of many states.

DUKE VINCENTIO:
Well, your wit is in the care of side and that.

Second Lord:
They would be ruled after this chamber, and
my fair nues begun out of the fact, to be conveyed,
Whose noble souls I'll have the heart of the wars.

Clown:
Come, sir, I will make did behold your worship.

VIOLA:
I'll drink it.
```

CHAT BOTS

DL EXAMPLES

Chat GPT and all the others

- Really, really good text generators compared to previous generation models
- Also, combine extra information from information retrieval (search results)

CODE GENERATION

DL EXAMPLES

CoPilot

```javascript
function sort(arr) {
    const { length } = arr;
    let temp;
    for (let i = 1; i < length; i++) {
        let j = i;
        temp = arr[i];
        while (j > 0 && (arr[j - 1] > temp)) {
            arr[j] = arr[j - 1];
            j--;
        }
        arr[j] = temp;
    }
    return arr;
}
```

convert this to use bubble sort

Accept Discard

Changed 13 lines

```javascript
function bubbleSort(arr) {
    const { length } = arr;
    for (let i = 0; i < length - 1; i++) {
        for (let j = 0; j < length - 1 - i; j++) {
            if (arr[j] > arr[j + 1]) {
                let temp = arr[j];
                arr[j] = arr[j + 1];
                arr[j + 1] = temp;
            }
        }
    }
    return arr;
}
```

QUESTION ANSWERING

DL EXAMPLES

Agrawal et al., "[VQA: Visual Question Answering]", ICCV 2015

APPLICATIONS

DISCUSSION

Why are these tasks challenging?

What are some problems from everyday life that can be helped by deep learning?

What are some of the ethical concerns about using deep learning?

COURSE GOALS

COARSE OVERVIEW

Understand the basics of Deep Learning (math and code)

1. Know how to create your own deep learning system (from scratch)
2. Know how to use common tools
3. Know (some) modern architecture choices for specific applications
4. Know the basics of successfully training deep models
5. (CS 2078) conduct a project to create a solution to a real problem or answer a scientific question

TEXTBOOK

COARSE OVERVIEW

Ian Goodfellow, Yoshua Bengio, Aaron Corville, *Deep Learning*.

Many other good resources out there

I will post links to other reading resources to supplement lectures.

PROGRAMMING

COARSE OVERVIEW

Languages/Frameworks: Python, Numpy, Jax

Numpy tutorial: http://cs231n.github.io/python-numpy-tutorial/

Computing resource: Google Colab (free GPU) — slow for development and debugging.

Debug locally then run full code

COURSE STRUCTURE

COARSE OVERVIEW

- Lecture
- Two Exams
- Homework (different sizes, coding, written, math)
- Participation (quizzes)
- Weekly quiz

COURSE STRUCTURE

COARSE OVERVIEW

Course Project (CS 2078)

- Teams of 2-3 students,
- Proposal, two reports, presentation
- New application or solution
- Answer a scientific question

POLICIES

COARSE OVERVIEW

- 5 Late Days (25% off for each late day after that)
- Get prior consent for excused absence
- If you need accommodations, contact Disability Resources and Services (DRS) and me for any specific requests

  I'm happy to accommodate
- Collaboration — not allowed (except project)
  - Your work must be your own.
  - You may discuss but not share answers or code
- AI — Do not use Chat GPT or other services to generate answers
  - It is obvious and often wrong

See Syllabus for more Info

WORKLOAD

COARSE OVERVIEW

- Some homework will take a long time 6+ hours
- Start early!!!!!! Deep learning can be slow to train
- **Come to office hours** if you are getting lost (or just want to chat more about a topic)

WARNING!!

MATH CONTENT AHEAD

- This course will be **MATH** heavy
- The heart of deep learning is optimizing differentiable functions
- Mathematical formulations are necessary to describe, discuss, and understand what is going on

WARNING!!

MATH CONTENT AHEAD

- This course will be **MATH** heavy
- The heart of deep learning is optimizing differentiable functions
- Mathematical formulations are necessary to describe, discuss, and understand what is going on
- You will need to be able to read and understand expressions with derivatives and matrices

$$\nabla l(w) = \frac{\partial}{\partial W}\frac{1}{n}\sum_{i=1}^{n}\left(x_i W^\top - y_i\right)^2$$

$$\begin{aligned}\frac{\partial l(\theta)}{\partial W^i} &= \frac{\partial l(\theta)}{h^i}\frac{\partial h^i}{\partial W^i}\\ &= \left(\partial_{h^i} l(\theta) \odot \frac{\partial \sigma(z^i)}{\partial z^i}\right)^\top h^{i-1}\end{aligned}$$

EXTRA CREDIT

OPPORTUNITY

Earn 2% on course total:

- Schedule a one-on-one meeting with me for 30 minutes (except for the last week of classes)
- We can discuss anything: course content, research, hobbies, etc
- My goal: get people who do not normally come into office hours for help to come.

WHAT IS DEEP LEARNING

DL EXAMPLES

All the above examples (except some Chat Bots) are just approximating a function

WHAT IS DEEP LEARNING

DL EXAMPLES

All the above examples (except some Chat Bots) are just approximating a function

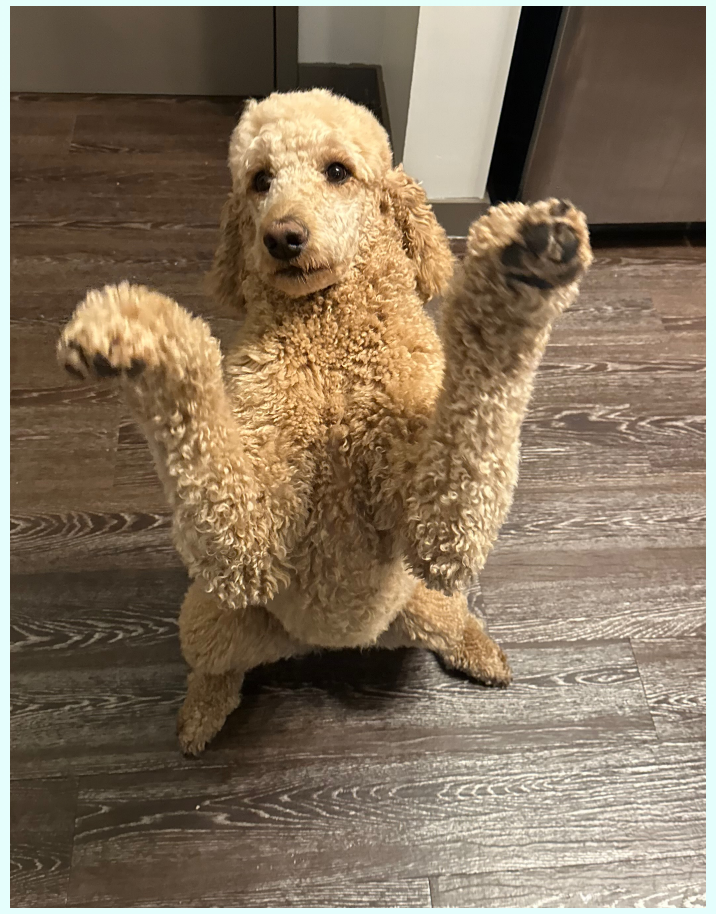
A function $f\colon \mathscr{X} \rightarrow \mathscr{Y}$ is a mapping from some input space $\mathscr{X}$ to some output space $\mathscr{Y}$

APPROXIMATING FUNCTIONS

BASICS OF DEEP LEARNING

Processing

Cat

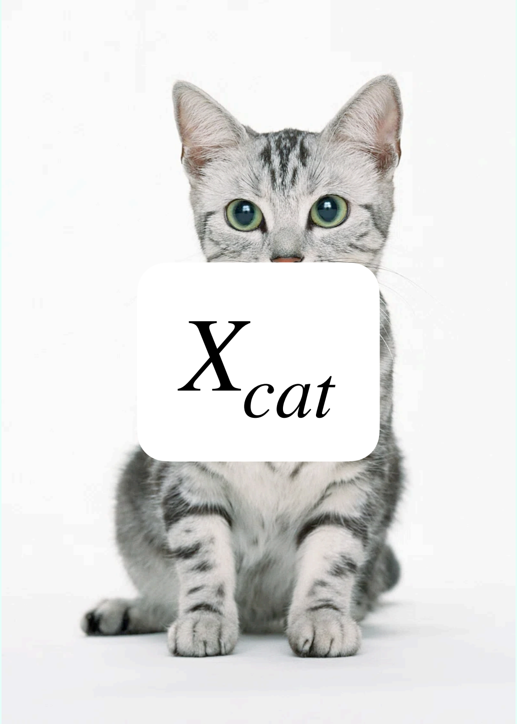


Processing

Dog

APPROXIMATING FUNCTIONS

BASICS OF DEEP LEARNING

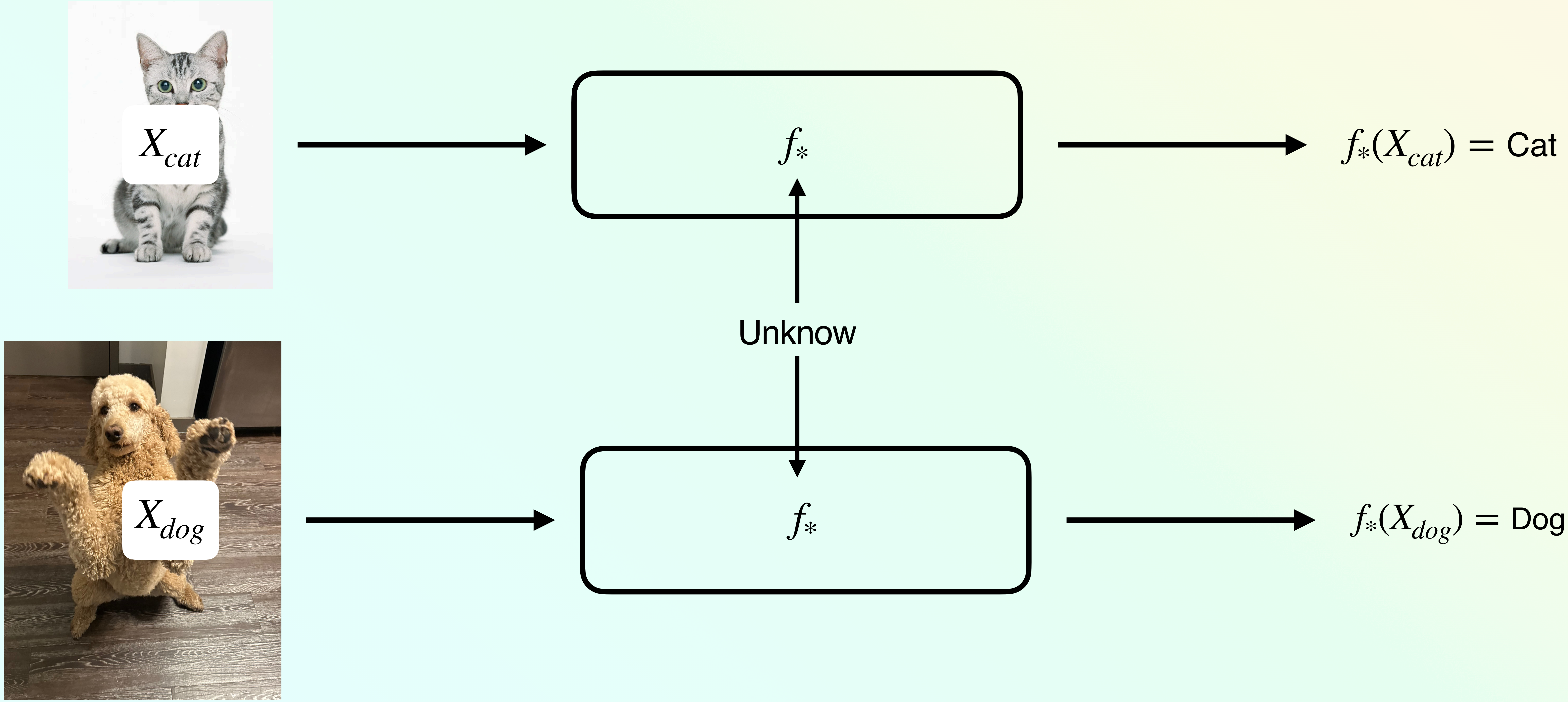


$\longrightarrow$ $\boxed{f_*}$ $\longrightarrow$ $f_*(X_{cat}) =$ Cat

$\longrightarrow$ $\boxed{f_*}$ $\longrightarrow$ $f_*(X_{dog}) =$ Dog

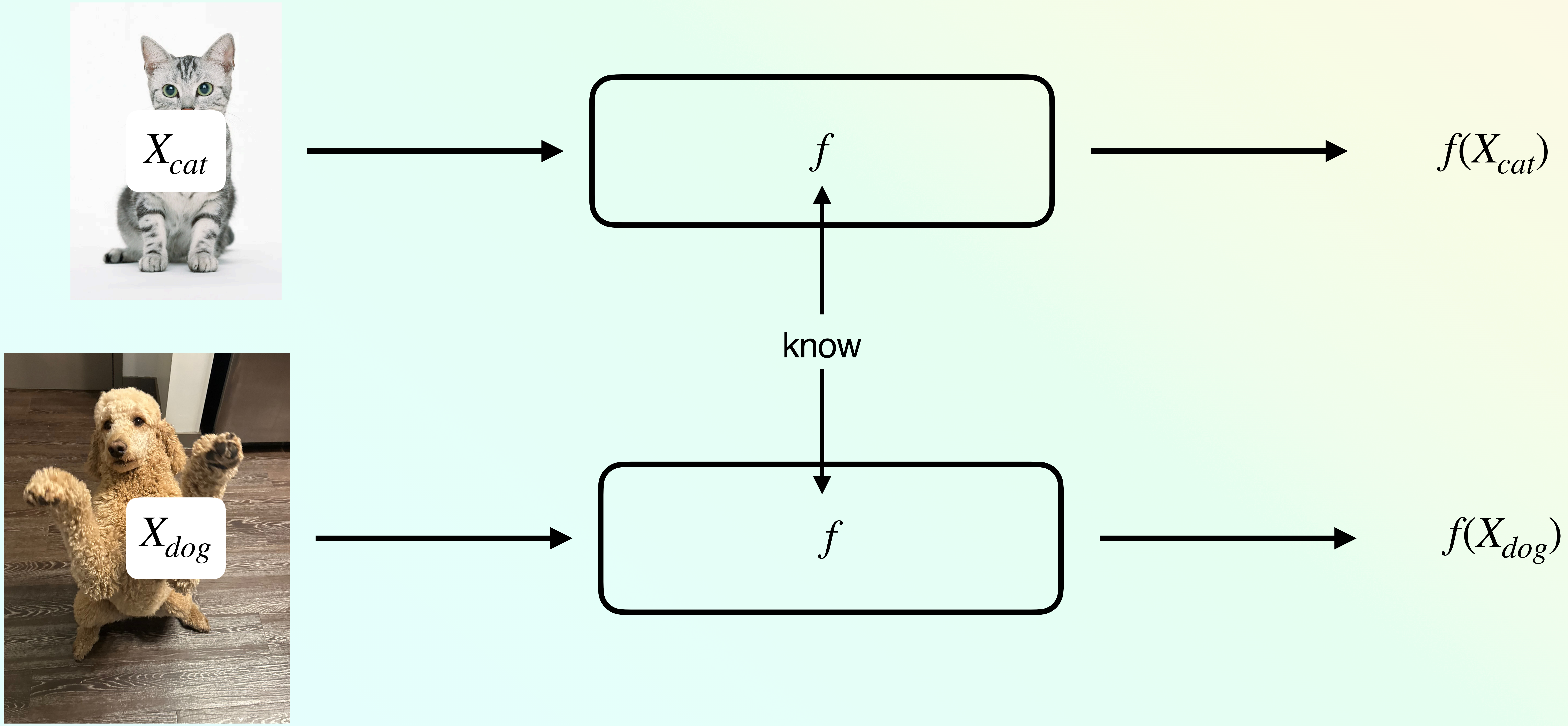
APPROXIMATING FUNCTIONS

BASICS OF DEEP LEARNING



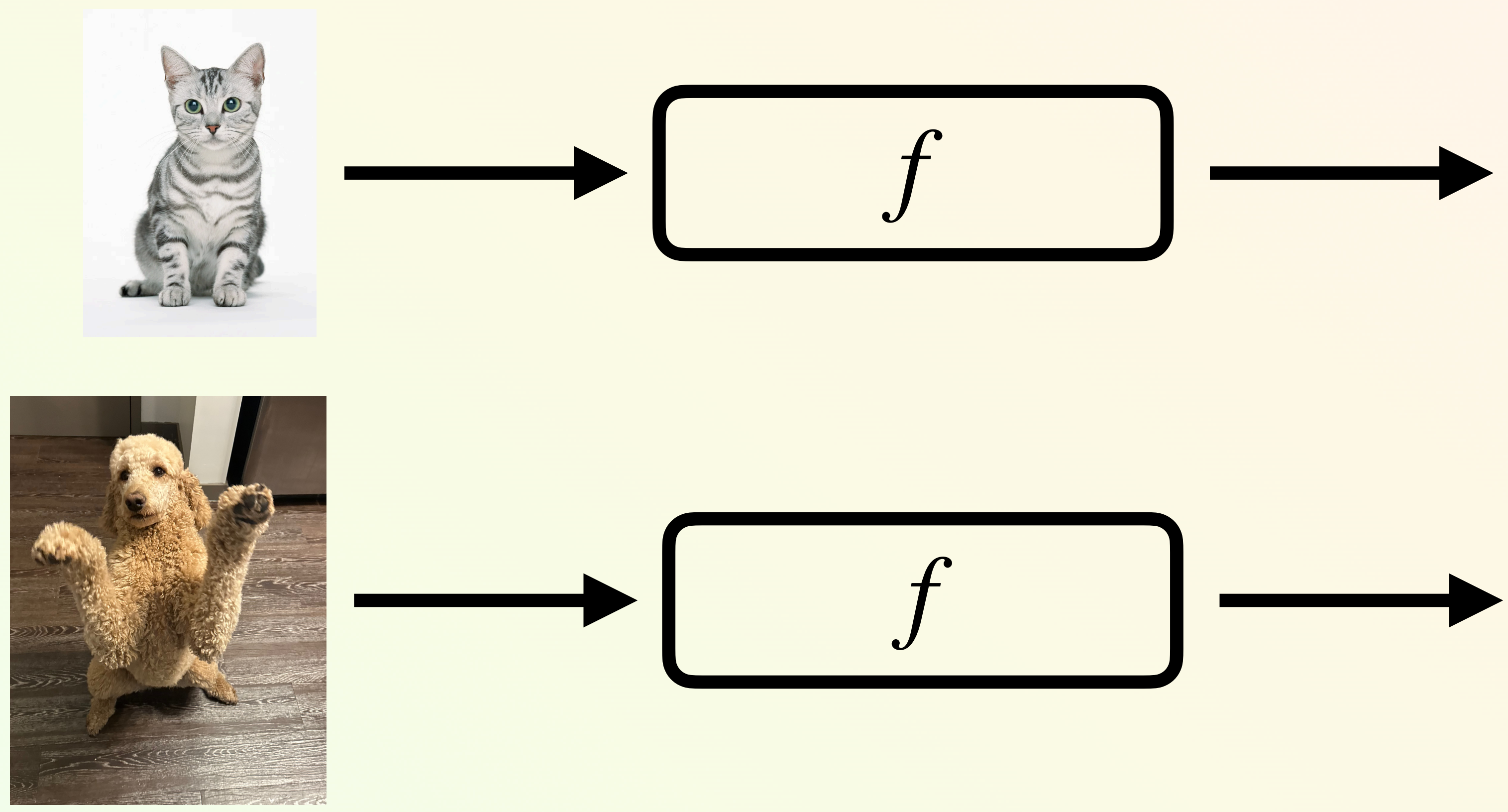
APPROXIMATING FUNCTIONS

BASICS OF DEEP LEARNING



APPROXIMATING FUNCTIONS

BASICS OF DEEP LEARNING



Find f such that

$$||f_*(X) - f(X)|| \leq \epsilon$$

Outputs of f are close to f_*

$f_*(X_{cat})$ is Cat so f should output Cat with high certainty

MACHINE TRANSLATION

DL EXAMPLES

APPROXIMATING FUNCTIONS

MACHINE TRANSLATION

$X =$ spanish phrase

$f_*(X) =$ english phrase

$f_*(\text{hola}) =$ hello

Collect many examples of Spanish-to-English phrases

Find f that will map Spanish phrases to the corresponding English phrases

TEXT GENERATIOIN

DL EXAMPLES

```
PANDARUS:
Alas, I think he shall be come approached and the day
When little srain would be attain'd into being never fed,
And who is but a chain and subjects of his death,
I should not sleep.

Second Senator:
They are away this miseries, produced upon my soul,
Breaking and strongly should be buried, when I perish
The earth and thoughts of many states.

DUKE VINCENTIO:
Well, your wit is in the care of side and that.

Second Lord:
They would be ruled after this chamber, and
my fair nues begun out of the fact, to be conveyed,
Whose noble souls I'll have the heart of the wars.

Clown:
Come, sir, I will make did behold your worship.

VIOLA:
I'll drink it.
```

APPROXIMATING FUNCTIONS

TEXT GENERATION

$X_{1:n} =$ ASCII characters of an n character passage

Example: Entire Shakespeare play

$X_k =$ first k characters of $X_{1:n}$

$Y_k =$ last n-k characters of $X_{1:n}$

minimize $||f(X_k) - Y_k||$ for all k

APPROXIMATING FUNCTIONS

DEEP LEARNING BASICS

Try to approximate any function just need examples of (X,Y)

- Often need LOTS of examples

Deep learning is not magic: models only learn what they need to, not what you WANT them to.

REVIEW

LINEAR ALGEBRA

Need a consistent language to represent data

Linear Algebra represents multiple variables in groups as vectors and matrics

Lets us abstract many “low level” operations

LINEAR ALGEBRA

BASICS

$x \in \mathbb{R}^3$ is a column vector of elements, $A \in \mathbb{R}^{2,3}$ is the matrix

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

$$A = \begin{bmatrix} A_{1,1} & A_{1,2} & A_{1,3} \\ A_{2,1} & A_{2,2} & A_{1,3} \end{bmatrix}$$

LINEAR ALGEBRA

PROPERTIES

Linearity

$\alpha x = [\alpha x_1, \alpha x_2, \alpha x_3]^\top$

$$\alpha A = \begin{bmatrix} \alpha A_{1,1} & \alpha A_{1,2} & \alpha A_{1,3} \\ \alpha A_{2,1} & \alpha A_{2,2} & \alpha A_{1,3} \end{bmatrix}$$

$\top$ is the transpose operator that flips the rows and columns, e.g., $x^\top \in \mathbb{R}^{1,3}$ and $A^\top \in \mathbb{R}^{3,2}$

$x \in \mathbb{R}^3$ and $y \in \mathbb{R}^3$

$$x + y = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ x_3 + y_3 \end{bmatrix}$$

$A \in \mathbb{R}^{2,3}$ and $B \in \mathbb{R}^{2,3}$

$$A + B = \begin{bmatrix} A_{1,1} + B_{1,1} & A_{1,2} + B_{1,2} & A_{1,3} + B_{1,3} \\ A_{2,1} + B_{2,1} & A_{2,2} + B_{2,2} & A_{2,3} + B_{2,3} \end{bmatrix}$$

LINEAR ALGEBRA

USEFUL OPERATIONS

Inner product: $\sum_{i=1}^{3} x_i y_i = x^\top y = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$

Matrix-vector multiplication

$$Ax = \begin{bmatrix} A_{11} & A_{1,2} & A_{1,3} \\ A_{21} & A_{2,2} & A_{2,3} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} A_{1,1}x_1 + A_{1,2}x_2 + A_{1,3}x_3 \\ A_{2,1}x_1 + A_{2,2}x_2 + A_{2,3}x_3 \end{bmatrix} = \begin{bmatrix} A_{1,\cdot}x \\ A_{2,\cdot}x \end{bmatrix}$$

LINEAR ALGEBRA

USEFUL OPERATIONS

Matrix-Matrix multiplication

$A \in \mathbb{R}^{2\times3}, B \in \mathbb{R}^{3\times2}$

$$AB = \begin{bmatrix} A_{1,1} & A_{1,2} & A_{1,3} \\ A_{2,1} & A_{2,2} & A_{2,3} \end{bmatrix} \begin{bmatrix} B_{1,1} & B_{1,2} \\ B_{2,1} & B_{2,2} \\ B_{3,1} & B_{3,2} \end{bmatrix} = \begin{bmatrix} A_{1,1}B_{1,1} + A_{1,2}B_{2,1} + A_{1,3}B_{3,1} & A_{1,1}B_{1,2} + A_{1,2}B_{2,2} + A_{1,3}B_{3,2} \\ A_{2,1}B_{1,1} + A_{2,2}B_{2,1} + A_{2,3}B_{3,1} & A_{2,1}B_{1,2} + A_{2,2}B_{2,2} + A_{2,3}B_{3,2} \end{bmatrix}$$

LINEAR ALGEBRA

EXAMPLE

Represent parameters of a line $f(x) = mx + b$

Vector of parameters $w = \begin{bmatrix} m \\ b \end{bmatrix}$

Vector representation of inputs $\mathbf{x} = \begin{bmatrix} x \\ 1 \end{bmatrix}$

Vector representation of f: $f\colon \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$

$$f(\mathbf{x}, w) = \mathbf{x}^\top w = \mathbf{x}_1 w_1 + \mathbf{x}_2 w_2 = mx + b(1)$$

LINEAR ALGEBRA

EXAMPLE

Predict a person's chance of heart attack, stroke, and death.

$$f_{\text{heart}}(\text{age}) = m_1\text{age} + b_1, \quad f_{\text{stroke}}(\text{age}) = m_2\text{age} + b_2, \quad f_{\text{death}}(\text{age}) = m_3\text{age} + b_3$$

$$W \in \mathbb{R}^{2,3} = \begin{bmatrix} m_1 & m_2 & m_3 \\ b_1 & b_2 & b_3 \end{bmatrix}, x = \begin{bmatrix} \text{age} & 1 \end{bmatrix}$$

$$f_{\text{diseases}}(x, W) = xW = \begin{bmatrix} m_1\text{age} + b_1 & m_2\text{age} + b_2 & m_3\text{age} + b_3 \end{bmatrix}$$

LINEAR ALGEBRA

EXAMPLE

Predict a person's chance of heart attack, stroke, and death.

$$f_{\text{heart}}(\text{age}) = m_1\text{age} + b_1, \quad f_{\text{stroke}}(\text{age}) = m_2\text{age} + b_2, \qquad f_{\text{death}}(\text{age}) = m_3\text{age} + b_3$$

Predictions for multiple people

$$x = \begin{bmatrix} 21 & 1 \\ 35 & 1 \\ 60 & 1 \end{bmatrix}$$

$$f_{\text{diseases}}(x, W) = xW = \begin{bmatrix} m_1 21 + b_1 & m_2 21 + b_2 & m_3 21 + b_3 \\ m_1 35 + b_1 & m_2 35 + b_2 & m_3 35 + b_3 \\ m_1 60 + b_1 & m_2 60 + b_2 & m_3 60 + b_3 \end{bmatrix} = \begin{bmatrix} f_{\text{heart}}(21) & f_{\text{stroke}}(21) & f_{\text{death}}(21) \\ f_{\text{heart}}(35) & f_{\text{stroke}}(35) & f_{\text{death}}(35) \\ f_{\text{heart}}(60) & f_{\text{stroke}}(60) & f_{\text{death}}(60) \end{bmatrix}$$

CALCULUS

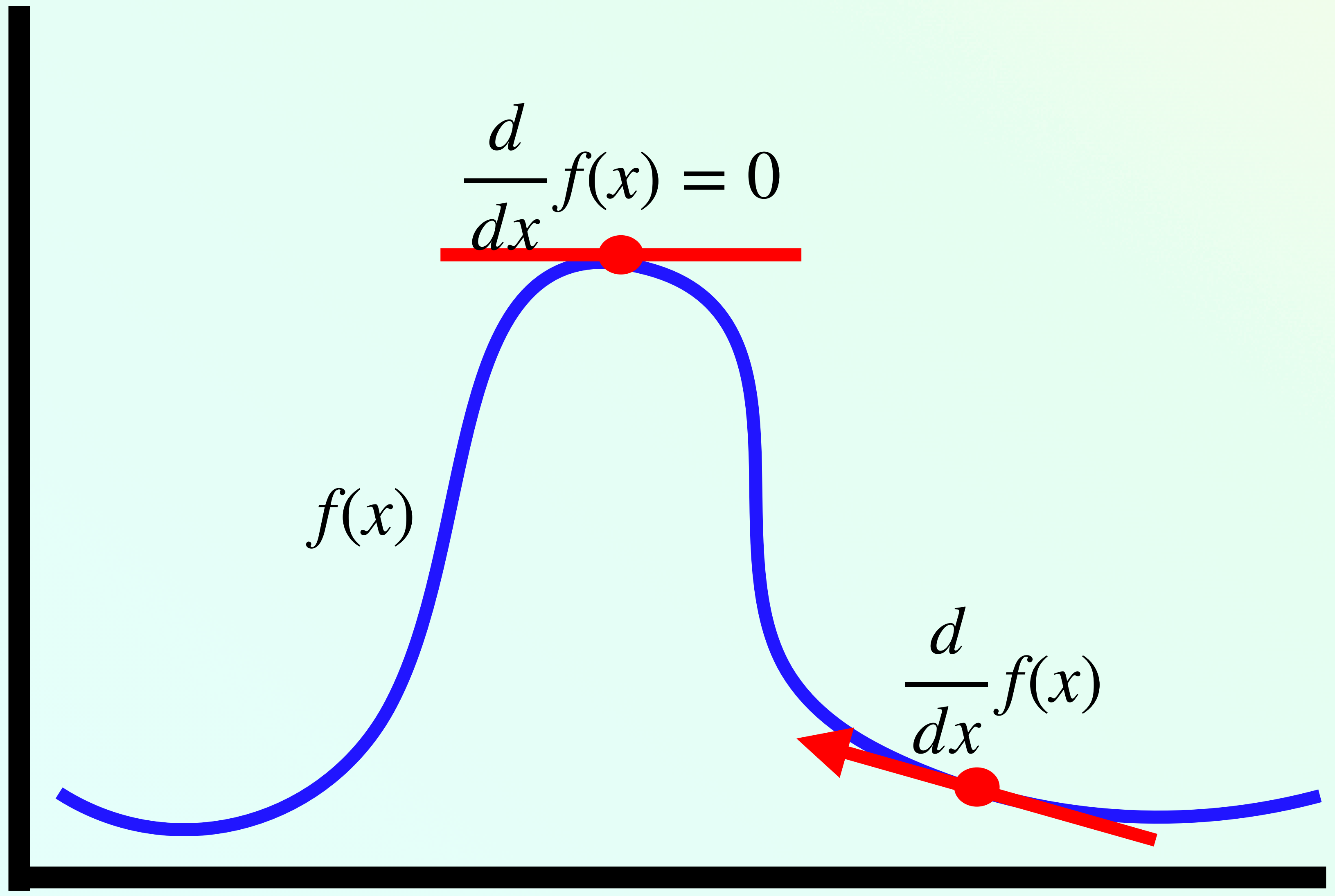
WHY

1. Take derivatives (how fast a function changes)
2. Helpful for finding the minimum and maximum of a function
3. Used in (stochastic) gradient descent

CALCULUS

DEFINITIONS

A derivative of a function gives the slope (tangent) of that function

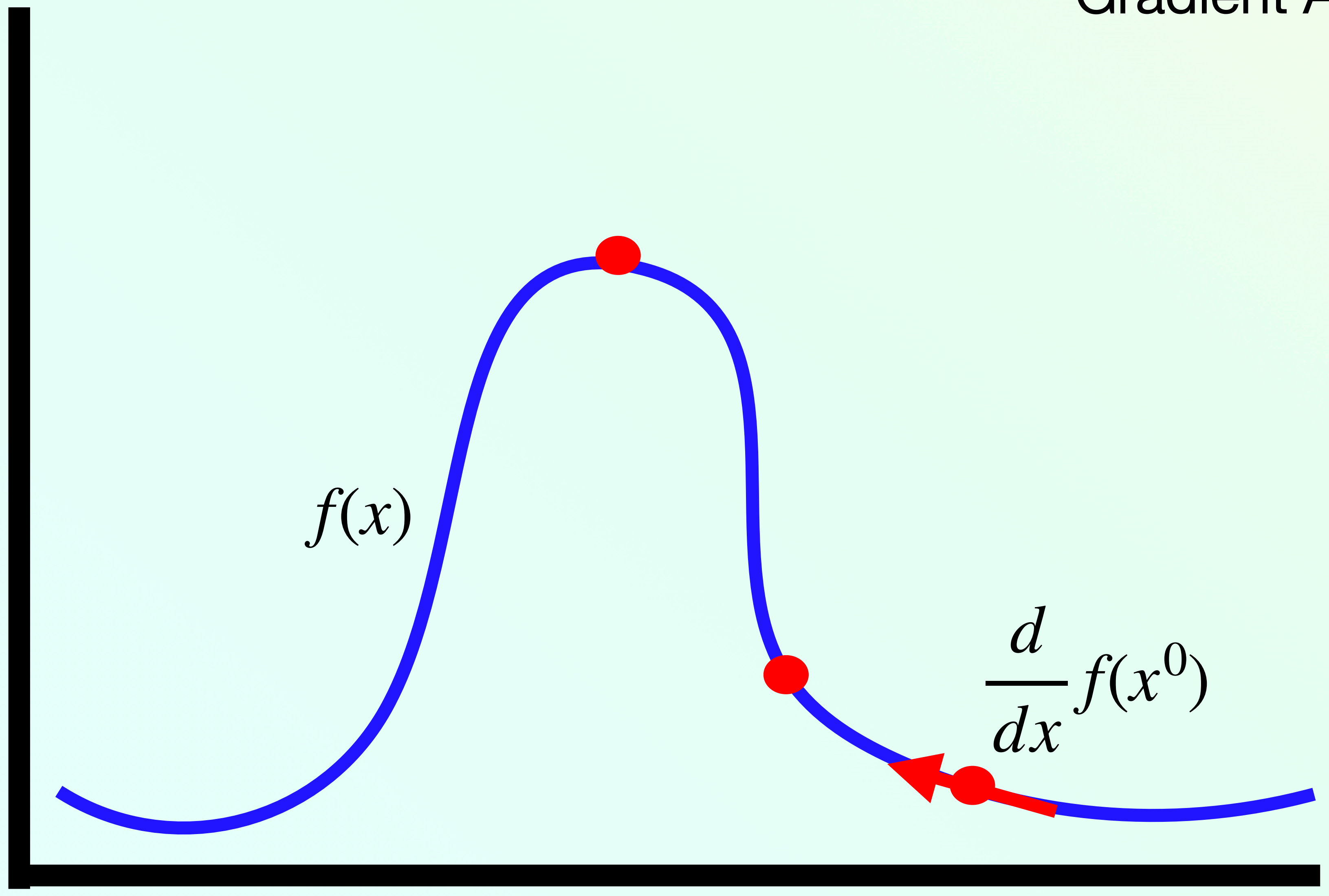


CALCULUS

OPTIMIZATION

Derivative says how to change the input to the function to increase that function

Gradient Ascent: $x^{i+1} \leftarrow x^i + \eta \frac{d}{dx} f(x^i)$

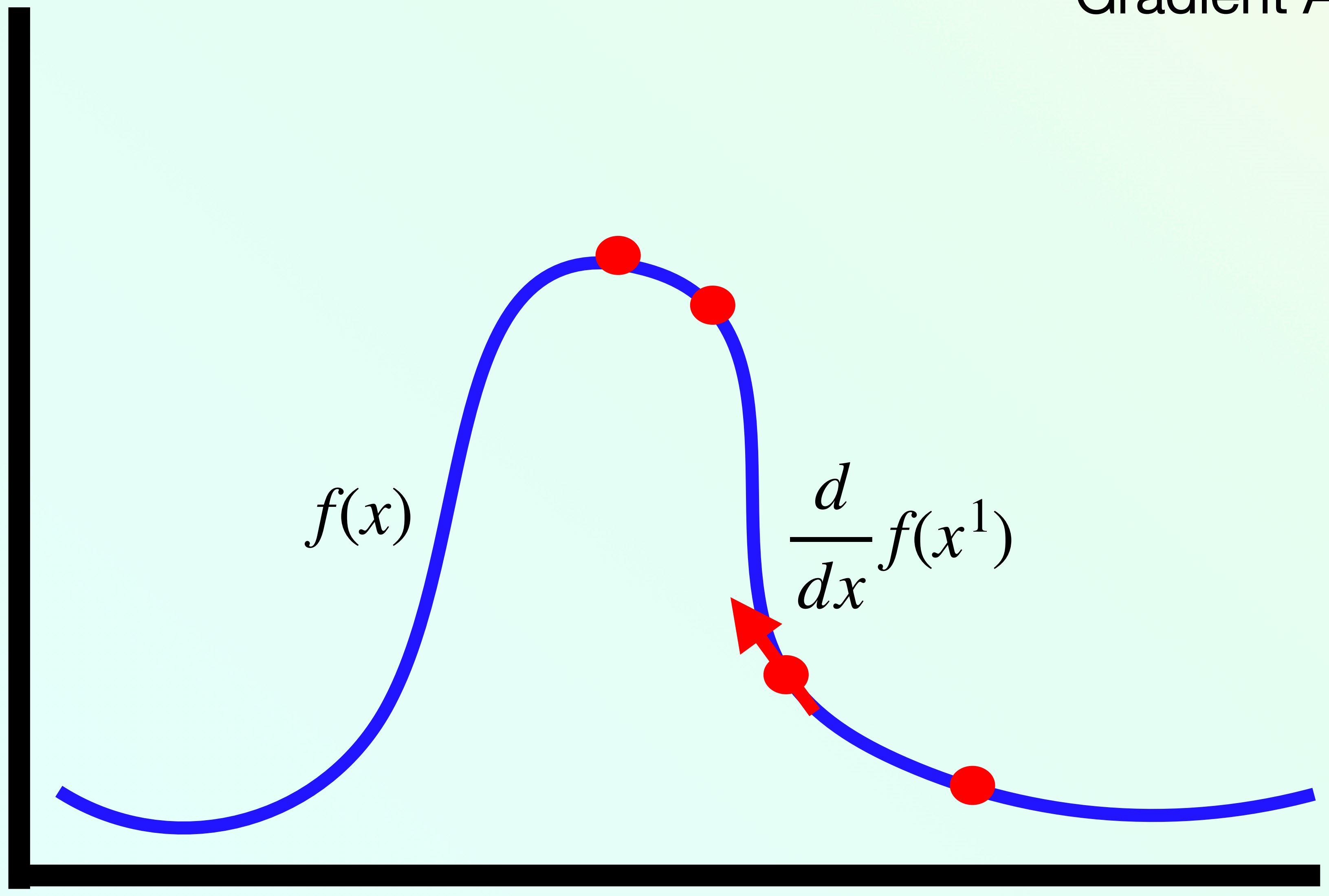


CALCULUS

OPTIMIZATION

Derivative says how to change the input to the function to increase that function

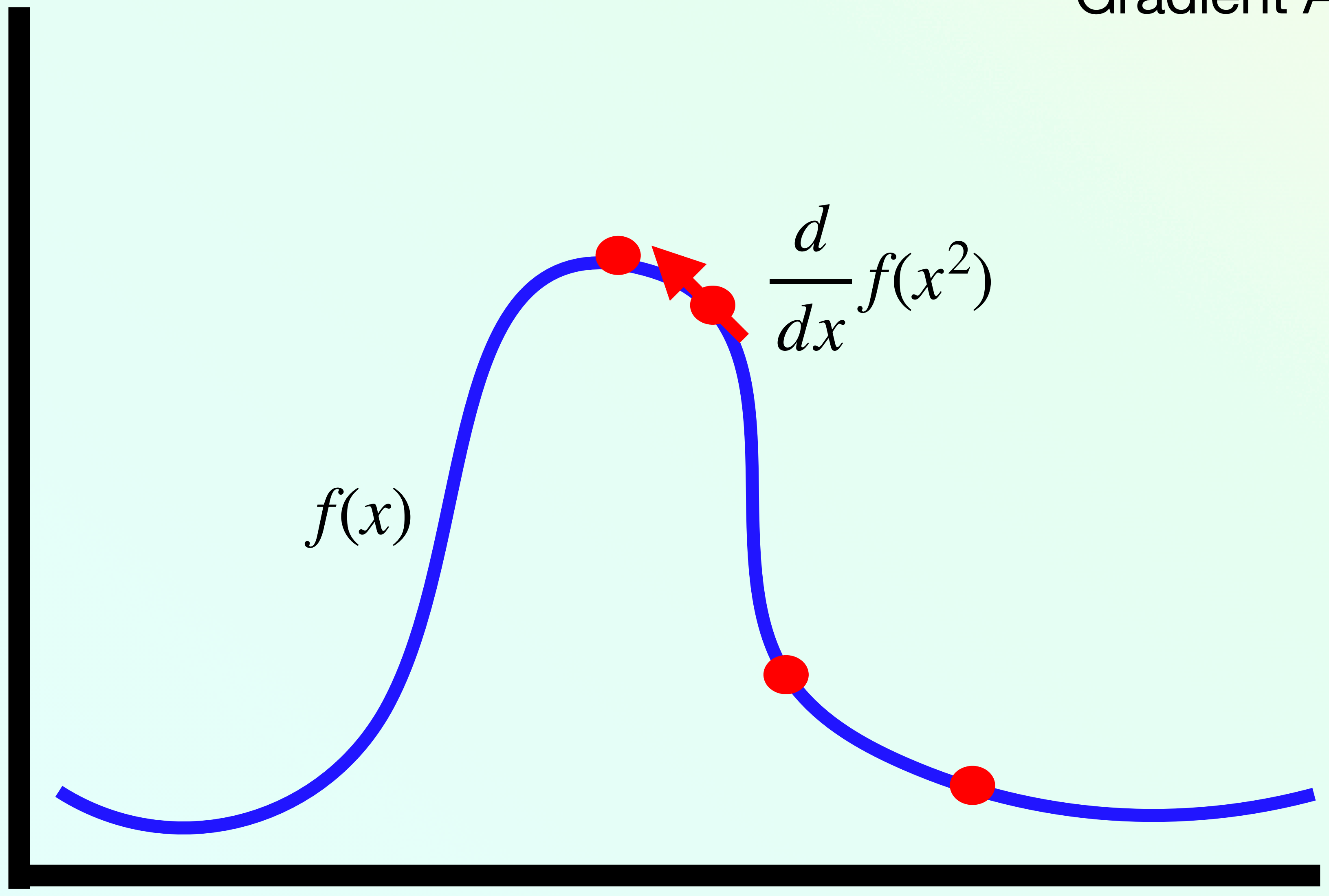
Gradient Ascent: $x^{i+1} \leftarrow x^i + \eta \frac{d}{dx} f(x^i)$



CALCULUS

OPTIMIZATION

Derivative says how to change the input to the function to increase that function



Gradient Ascent: $x^{i+1} \leftarrow x^i + \eta \frac{d}{dx} f(x^i)$

CALCULUS

DEFINITIONS: MULTIVARIATE

Partial derivatives of a multivariate function $f(x, y) = x^2 + y^2$ are

$$\frac{\partial}{\partial x} f(x, y) = \frac{\partial}{\partial x} x^2 + \frac{\partial}{\partial x} y^2 = \frac{\partial}{\partial x} x^2$$

Treat all variables as constants for partial derivatives if they are not in the denominator of the derivative and apply typical differentiation rules.

$\frac{\partial}{\partial x}$ — treat y as a constant, $\frac{\partial}{\partial y}$ — treat x as a constant

The gradient of a function is all partial derivatives of that function's inputs, i.e.,

$$\nabla f(x, y) = \begin{bmatrix} \frac{\partial}{\partial x} f(x, y) & \frac{\partial}{\partial y} f(x, y) \end{bmatrix}$$

CALCULUS

DEFINITIONS: MULTIVARIATE

The gradient of a function is all partial derivatives of that function's inputs, i.e.,

$$\nabla f(x, y) = \begin{bmatrix} \frac{\partial}{\partial x} f(x, y) \\ \frac{\partial}{\partial y} f(x, y) \end{bmatrix}$$

CALCULUS

WITH LINEAR ALGEBRA

$$w \in \mathbb{R}^3, x \in \mathbb{R}^3 \quad f(w, x) = x^\top w = \sum_{i=1}^{3} x_i w_i$$

$$\frac{\partial}{\partial w} f(w, x) = \begin{bmatrix} \frac{\partial f(w, x)}{\partial w_1} \\ \frac{\partial f(w, x)}{\partial w_2} \\ \frac{\partial f(w, x)}{\partial w_3} \end{bmatrix}$$

CALCULUS

WITH LINEAR ALGEBRA

$$w \in \mathbb{R}^3, x \in \mathbb{R}^3 \quad f(w,x) = x^\top w = \sum_{i=1}^{3} x_i w_i$$

$$\frac{\partial}{\partial w} f(w,x) = \begin{bmatrix} \frac{\partial f(w,x)}{\partial w_1} \\ \frac{\partial f(w,x)}{\partial w_2} \\ \frac{\partial f(w,x)}{\partial w_3} \end{bmatrix} \qquad \frac{\partial f(w,x)}{\partial w_1} = \frac{\partial}{\partial w_1} \sum_{i=1}^{3} w_i x_i = \frac{\partial}{\partial w_1} w_1 x_1 + \frac{\partial}{\partial w_1} w_2 x_2 + \frac{\partial}{\partial w_1} w_3 x_3 = \frac{\partial}{\partial w_1} w_1 x_1 = x_1$$

CALCULUS

WITH LINEAR ALGEBRA

$w \in \mathbb{R}^3, x \in \mathbb{R}^3 \quad f(w,x) = x^\top w = \sum_{i=1}^{3} x_i w_i$

$$\frac{\partial}{\partial w} f(w,x) = \begin{bmatrix} \frac{\partial f(w,x)}{\partial w_1} \\ \frac{\partial f(w,x)}{\partial w_2} \\ \frac{\partial f(w,x)}{\partial w_3} \end{bmatrix} \qquad \frac{\partial f(w,x)}{\partial w_1} = \frac{\partial}{\partial w_1} \sum_{i=1}^{3} w_i x_i = \frac{\partial}{\partial w_1} w_1 x_1 + \frac{\partial}{\partial w_1} w_2 x_2 + \frac{\partial}{\partial w_1} w_3 x_3 = \frac{\partial}{\partial w_1} w_1 x_1 = x_1$$

$$\frac{\partial}{\partial w} f(w,x) = \begin{bmatrix} \frac{\partial f(w,x)}{\partial w_1} \\ \frac{\partial f(w,x)}{\partial w_2} \\ \frac{\partial f(w,x)}{\partial w_3} \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = x$$

CALCULUS

WITH LINEAR ALGEBRA

$$W \in \mathbb{R}^{3,2}, x \in \mathbb{R}^3 \quad f(W,x) = x^\top W = \left[\sum_{i=1}^3 x_i W_{i,1} \quad \sum_{i=1}^3 x_i W_{i,2}\right] = \left[x^\top W_{\cdot,1} \quad x^\top W_{\cdot,2}\right]$$

CALCULUS

WITH LINEAR ALGEBRA

$$W \in \mathbb{R}^{3,2}, x \in \mathbb{R}^3 \quad f(x, W) = x^\top W = \left[\sum_{i=1}^3 x_i W_{i,1} \quad \sum_{i=1}^3 W_{i,2} x_i\right] = \left[x^\top W_{\cdot,1} \quad x^\top W_{\cdot,2}\right]$$

If this is hard for you to track, write out the operations by hand to see that they are true.

You need to learn to think and read in matrix notation.

Practicing by hand helps build this connection (necessary)

CALCULUS

WITH LINEAR ALGEBRA

$$W \in \mathbb{R}^{3,2}, x \in \mathbb{R}^3 \quad f(x, W) = x^\top W = \left[\sum_{i=1}^3 x_i W_{i,1} \quad \sum_{i=1}^3 W_{i,2} x_i\right] = \left[x^\top W_{\cdot,1} \quad x^\top W_{\cdot,2}\right]$$

$$g(x, W) = \sum_{k=1}^{2} f(x, W)_{1,k}$$

CALCULUS

WITH LINEAR ALGEBRA

$$W \in \mathbb{R}^{3,2}, x \in \mathbb{R}^3 \quad f(x, W) = x^\top W = \left[\sum_{i=1}^3 x_i W_{i,1} \quad \sum_{i=1}^3 W_{i,2} x_i\right] = \left[x^\top W_{\cdot,1} \quad x^\top W_{\cdot,2}\right]$$

$$g(x, W) = \sum_{k=1}^{2} f(x, W)_{1,k}$$

$$\frac{\partial}{\partial W} g(x, W) = \begin{bmatrix} \frac{\partial g(x,W)}{\partial W_{11}} & \frac{\partial g(x,W)}{\partial W_{12}} \\ \frac{\partial g(x,W)}{\partial W_{21}} & \frac{\partial g(x,W)}{\partial W_{22}} \\ \frac{\partial g(x,W)}{\partial W_{31}} & \frac{\partial g(x,W)}{\partial W_{32}} \end{bmatrix} = \begin{bmatrix} \frac{\partial g(x,W)}{\partial W_{\cdot,1}} & \frac{\partial g(x,W)}{\partial W_{\cdot,2}} \end{bmatrix}$$

CALCULUS

WITH LINEAR ALGEBRA

$$W \in \mathbb{R}^{3,2}, x \in \mathbb{R}^3 \quad f(x, W) = x^\top W = \left[\sum_{i=1}^{3} x_i W_{i,1} \quad \sum_{i=1}^{3} W_{i,2} x_i\right] = \left[x^\top W_{\cdot,1} \quad x^\top W_{\cdot,2}\right]$$

$$g(x, W) = \sum_{k=1}^{2} f(x, W)_{1,k}$$

$$\frac{\partial}{\partial W_{\cdot,1}} g(x, W) = \frac{\partial}{\partial W_{\cdot,1}} \left(x^\top W_{\cdot,1} + x^\top W_{\cdot,2}\right) = \frac{\partial}{\partial W_{\cdot,1}} x^\top W_{\cdot,1} + \frac{\partial}{\partial W_{\cdot,1}} x^\top W_{\cdot,2} = \frac{\partial}{\partial W_{\cdot,1}} x^\top W_{\cdot,1} = x$$

CALCULUS

WITH LINEAR ALGEBRA

$$W \in \mathbb{R}^{3,2}, x \in \mathbb{R}^3 \quad f(x, W) = x^\top W = \left[\sum_{i=1}^3 x_i W_{i,1} \quad \sum_{i=1}^3 W_{i,2} x_i\right] = \left[x^\top W_{\cdot,1} \quad x^\top W_{\cdot,2}\right]$$

$$g(x, W) = \sum_{k=1}^{2} f(x, W)_{1,k}$$

$$\frac{\partial}{\partial W_{\cdot,1}} g(x, W) = \frac{\partial}{\partial W_{\cdot,1}} \left(x^\top W_{\cdot,1} + x^\top W_{\cdot,2}\right) = \frac{\partial}{\partial W_{\cdot,1}} x^\top W_{\cdot,1} + \frac{\partial}{\partial W_{\cdot,1}} x^\top W_{\cdot,2} = \frac{\partial}{\partial W_{\cdot,1}} x^\top W_{\cdot,1} = x$$

$$\frac{\partial}{\partial W_{\cdot,2}} g(x, W) = x$$

CALCULUS

WITH LINEAR ALGEBRA

$W \in \mathbb{R}^{3,2}, x \in \mathbb{R}^3 \quad f(x, W) = x^\top W = \left[\sum_{i=1}^{3} x_i W_{i,1} \quad \sum_{i=1}^{3} W_{i,2} x_i\right] = \left[x^\top W_{\cdot,1} \quad x^\top W_{\cdot,2}\right]$

$$g(x, W) = \sum_{k=1}^{2} f(x, W)_{1,k}$$

$$\frac{\partial}{\partial W_{\cdot,1}} g(x, W) = x, \ \frac{\partial}{\partial W_{\cdot,2}} g(x, W) = x$$

$$\frac{\partial}{\partial W} g(x, W) = \begin{bmatrix} \frac{\partial g(x, W)}{\partial W_{11}} & \frac{\partial g(x, W)}{\partial W_{12}} \\ \frac{\partial g(x, W)}{\partial W_{21}} & \frac{\partial g(x, W)}{\partial W_{22}} \\ \frac{\partial g(x, W)}{\partial W_{31}} & \frac{\partial g(x, W)}{\partial W_{32}} \end{bmatrix} = \begin{bmatrix} \frac{\partial g(x, W)}{\partial W_{\cdot,1}} & \frac{\partial g(x, W)}{\partial W_{\cdot,2}} \end{bmatrix} = [x \quad x] = \begin{bmatrix} x_1 & x_1 \\ x_2 & x_2 \\ x_3 & x_3 \end{bmatrix}$$

PROBABILITY

RANDOM VARIABLE

A random variable X is a function from possible outcomes in a *sample space* to a *measurable space.* (Wikipedia)

Example: Flipping a coin

- Sample space: The set $\{\text{heads}, \text{tails}\}$
- Measurable space: The set $\{-1,1\}$
- $X\colon \{\text{heads}, \text{tails}\} \to \{-1,1\}$

PROBABILITY

RANDOM VARIABLE

Example: Rolling a 20-sided die twice and taking the sum

- Sample space: The set $\{(1,1), (1,2), \cdots, (2,1), \cdots, (20,20)\}$
- Measurable space: The set $\{2,3,\cdots,40\}$
- $X : \{(1,1), (1,2), \cdots, (2,1), \cdots, (20,20)\} \rightarrow \{2,3,\cdots,40\}$

PROBABILITY

RANDOM VARIABLE

Example: D&D Sword (roll for damage)

- Sample space: The set $\{1,2,3,\cdots,20\}$
- Measurable space: The set $\{1,2,4\}$
- $X : \{1,2,3,\cdots,20\} \rightarrow \{1,2,4\}$

PROBABILITY

DEFINITION

$\Pr(A)$ is a measure of how likely is the event A

$$\Pr(A) = \frac{\text{\# occurances of } A}{\text{\# of all possible outcomes}} \in [0,1]$$

Dice roll $X : \{1,2,\cdots,20\} \to \{1,2,\cdots,20\}$

Event $A := X \in [15,20]$

$$\Pr(A) = \frac{6}{20}$$

PROBABILITY

NOTE

$\Pr(X)$ is undefined because X is not an event

PROBABILITY

PROPERTIES

Normalization $\sum_{e \in \mathcal{S}} \Pr(e) = 1$

Additivity $\Pr(A \cup B) = \Pr(A) + \Pr(B)$ if $A \cap B = \{\}$

PROBABILITY

EXPECTATION

An expectation is the mean or average value of a random variable.

Let the random variable $X \in \mathcal{X}$

The mean of X is

$$\mathbf{E}[X] = \sum_{x \in \mathcal{X}} \Pr(X = x)x$$

For some function $f\colon \mathcal{X} \to \mathbb{R}$

$$\mathbf{E}[f(X)] = \sum_{x \in \mathcal{X}} \Pr(X = x)f(x)$$

PROBABILITY

EXPECTATION: PROPERTIES

Linearity of expectation

$$\mathbf{E}[X + Y] = \mathbf{E}[X] + \mathbf{E}[Y]$$

$$\mathbf{E}[\alpha X] = \alpha \mathbf{E}[X]$$

PROBABILITY

COMMON MANIPULATIONS OF PROBABILITIES

Conditional Probability: $\Pr(A \mid B) = \dfrac{\Pr(A, B)}{\Pr(B)}$

PROBABILITY

COMMON MANIPULATIONS OF PROBABILITIES

Conditional Probability: $\Pr(A \,|\, B) = \dfrac{\Pr(A, B)}{\Pr(B)}$

Two random variables X and Y

$$\Pr(Y = y \,|\, X = x) = \frac{\Pr(Y = y, X = x)}{\Pr(X = x)}$$

PROBABILITY

MARIGINALIZATION

Two random variables X and Y

$\Pr(Y = y)$ — called the marginal probability (considers all possible X)

$$\Pr(Y = y) = \sum_{x \in \mathcal{X}} \Pr(Y = y, X = x)$$

PROBABILITY

MARIGINALIZATION

Two random variables X and Y

$\Pr(Y = y)$ — called the marginal probability (considers all possible X)

$$\Pr(Y = y) = \sum_{x \in \mathcal{X}} \Pr(X = x) \Pr(Y = y \mid X = x)$$

PROBABILITY

CONDITIONAL EXPECTATION

$$\mathbf{E}[Y|X=x] \doteq \sum_{y \in \mathcal{Y}} \Pr(Y=y|X=x)y$$

NEXT CLASS

1. Watch Canvas for a forum invite
2. Watch for a review assignment

Next Class — Brief Review and Function Approximation

Bring pen and paper (or other writing device)