Statistical Learning: Introduction

Jesse Gronsbell
Department of Statistical Sciences, University of Toronto

Intro: What is statistical learning?

What do you think data science is?

What about machine learning?

What about statistical learning?

Key definitions

- **Machine learning (ML)**: Development of models and algorithms from data
  - **Deep learning**: A subfield of ML focused on algorithms modeled after the human brain
- **Statistical learning**: Branch of applied statistics focusing on statistical models and uncertainty quantification
- **Data Science**: Extraction of knowledge from data using a toolbox made up of mathematical, statistical, engineering, and machine learning techniques

https://web.stanford.edu/~hastie/TALKS/SLBD_new.pdf

Key definitions

- **Machine learning (ML)**: Development of models and algorithms from data
  - **Deep learning**: A subfield of ML focused on algorithms modeled after the human brain

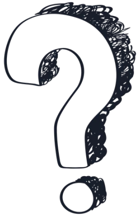
- **Statistical learning**: Branch of applied statistics focusing on statistical models and uncertainty quantification

- **Data Science**: Extraction of knowledge from data using a toolbox made up of mathematical, statistical, engineering, and machine learning techniques

Through different lenses, they all aim to **make sense of data**!

https://web.stanford.edu/~hastie/TALKS/SLBD_new.pdf

Why the differences?



Let's talk data

In the most familiar setting (i.e. **supervised learning**) data is made up of two primary ingredients:

1. **Outcome/Output/Target/Label(s)**: The variable(s) that you want to understand better

2. **Covariate/Input/Feature(s)**: The variable(s) that can potentially tell you something about your outcome(s)

Let’s talk data

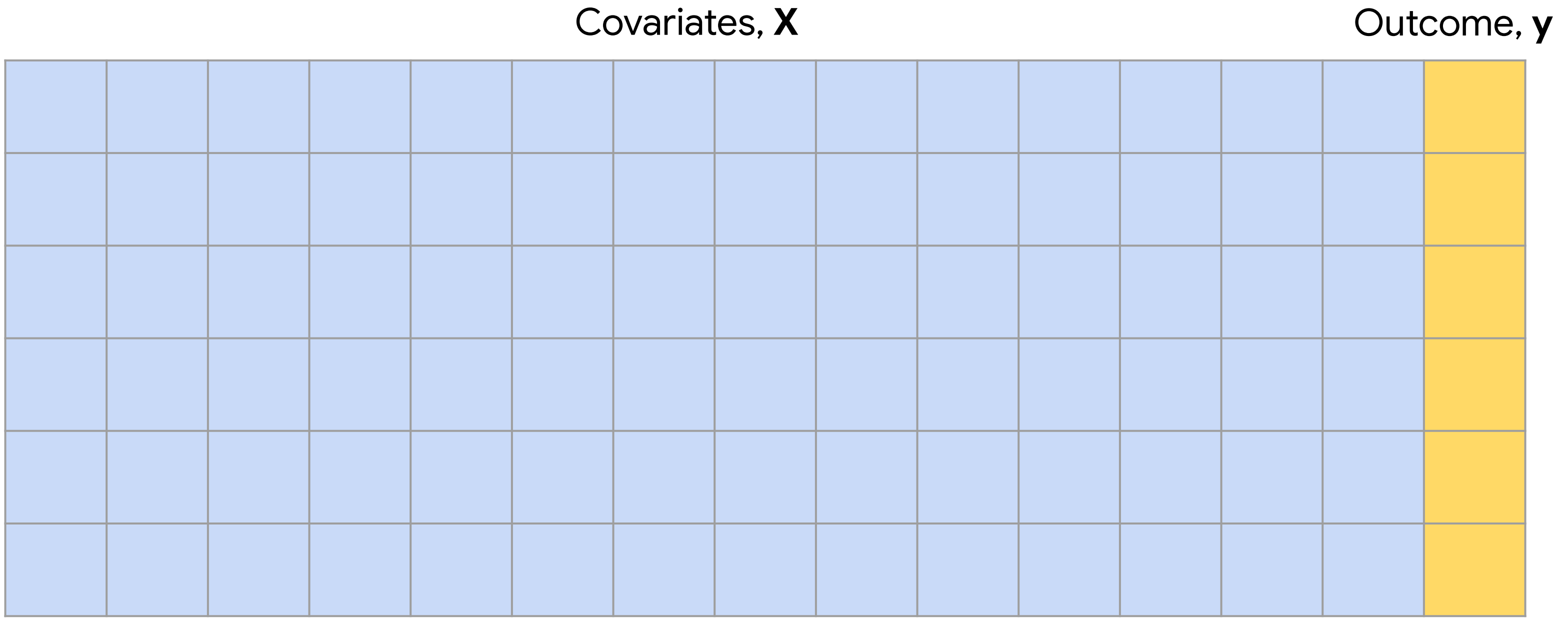
In the most familiar setting (i.e. **supervised learning**) data is made up of two primary ingredients:

1. **Outcome/Output/Target/Label(s)**: The variable(s) that you want to understand better

2. **Covariate/Input/Feature(s)**: The variable(s) that can potentially tell you something about your outcome(s)

Outcome and covariate are most commonly used in statistics so we’ll use them!

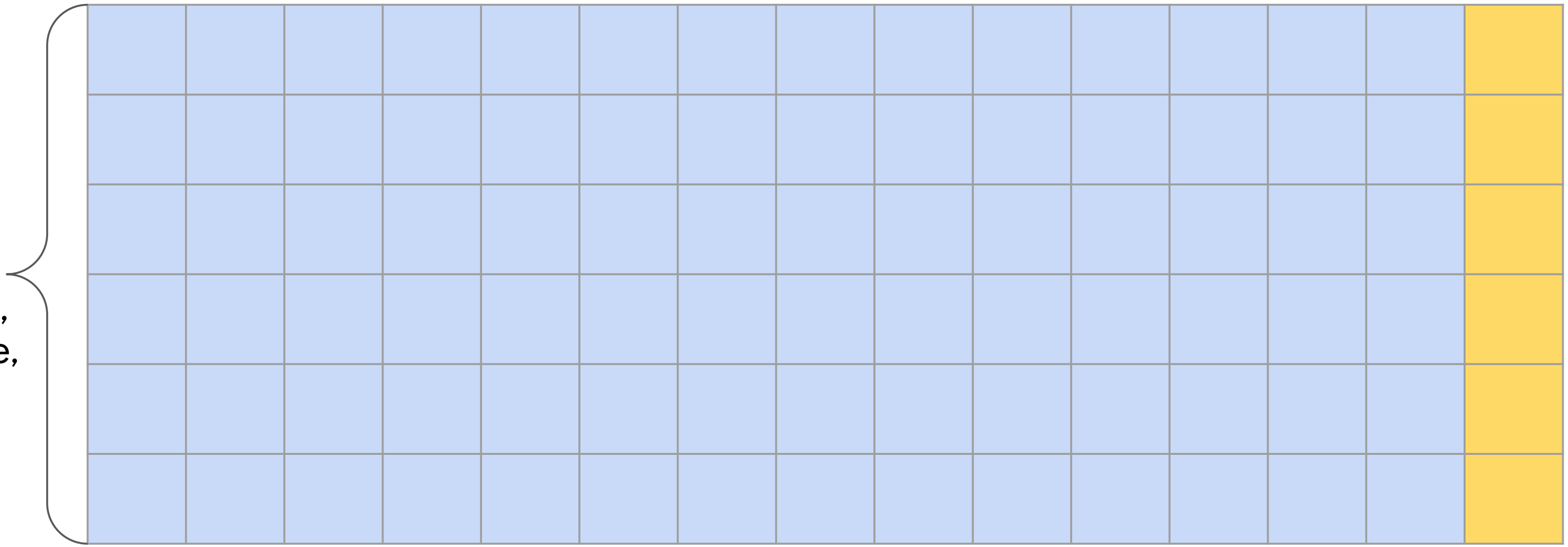
Let's look at a *data matrix*



Let's look at a *data matrix*

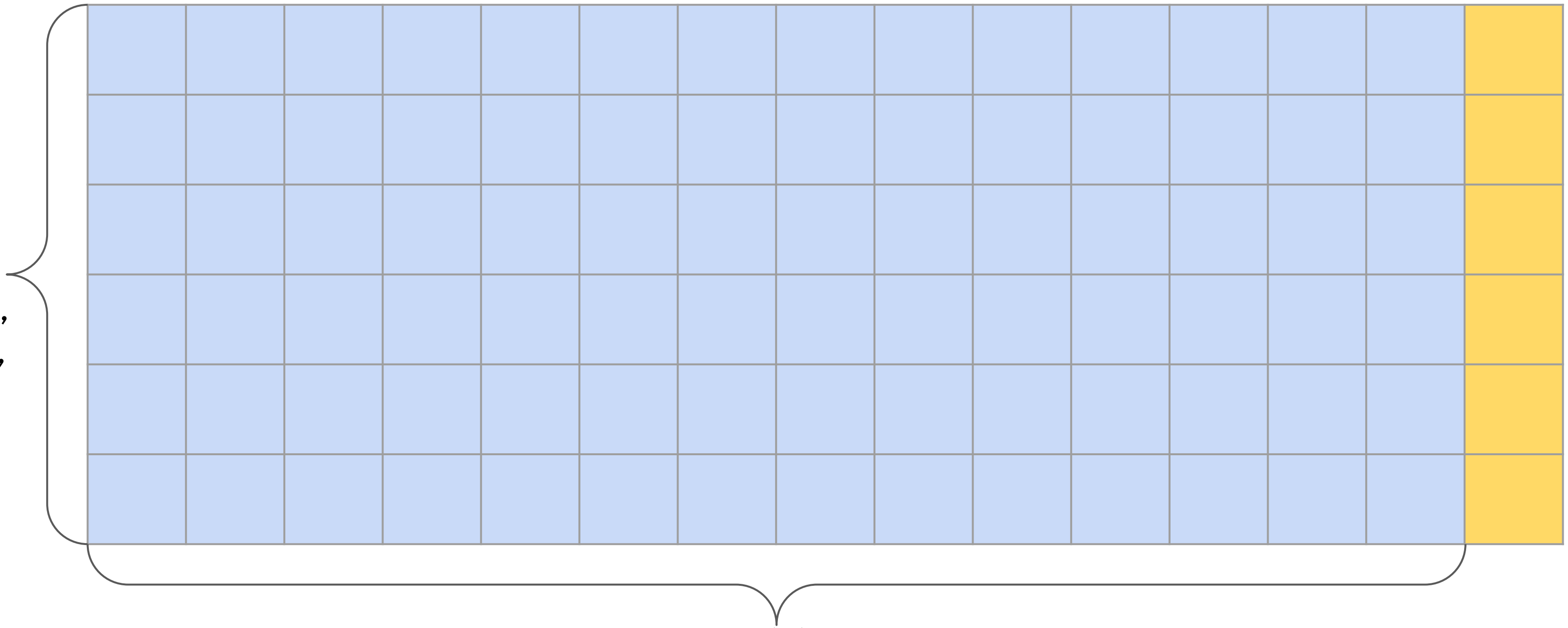
Covariates, **X**

Outcome, **y**



<u>Rows</u> (n): contain the units of analysis, e.g. people, images, sentences

Let's look at a *data matrix*



Example data matrix: Penguin dataset

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
[1,]	3	1	42.0	13.5	210	4150	1	2007
[2,]	3	1	54.3	15.7	231	5650	2	2008
[3,]	2	2	42.4	17.3	181	3600	1	2007
[4,]	3	1	48.6	16.0	230	5800	2	2008
[5,]	2	2	47.0	17.3	185	3700	1	2007
[6,]	3	1	50.4	15.7	222	5750	2	2009
[7,]	1	2	36.0	17.8	195	3450	1	2009
[8,]	1	2	41.3	20.3	194	3550	2	2008
[9,]	1	2	39.6	18.8	190	4600	2	2007
[10,]	3	1	49.6	16.0	225	5700	2	2008
[11,]	3	1	46.9	14.6	222	4875	1	2009
[12,]	1	2	40.2	17.1	193	3400	1	2009
[13,]	3	1	46.8	16.1	215	5500	2	2009
[14,]	3	1	49.4	15.8	216	4925	2	2009
[15,]	1	1	38.2	18.1	185	3950	2	2007

p = ?
n = ?

Example data matrix: Penguin dataset

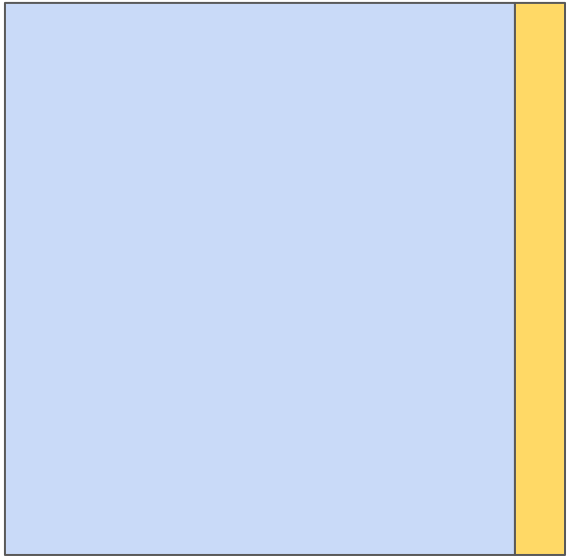
	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
[1,]	3	1	42.0	13.5	210	4150	1	2007
[2,]	3	1	54.3	15.7	231	5650	2	2008
[3,]	2	2	42.4	17.3	181	3600	1	2007
[4,]	3	1	48.6	16.0	230	5800	2	2008
[5,]	2	2	47.0	17.3	185	3700	1	2007
[6,]	3	1	50.4	15.7	222	5750	2	2009
[7,]	1	2	36.0	17.8	195	3450	1	2009
[8,]	1	2	41.3	20.3	194	3550	2	2008
[9,]	1	2	39.6	18.8	190	4600	2	2007
[10,]	3	1	49.6	16.0	225	5700	2	2008
[11,]	3	1	46.9	14.6	222	4875	1	2009
[12,]	1	2	40.2	17.1	193	3400	1	2009
[13,]	3	1	46.8	16.1	215	5500	2	2009
[14,]	3	1	49.4	15.8	216	4925	2	2009
[15,]	1	1	38.2	18.1	185	3950	2	2007

$p = 7$
$n = 15$

We need the terminology because data comes in many forms

Tall data

$p < n$

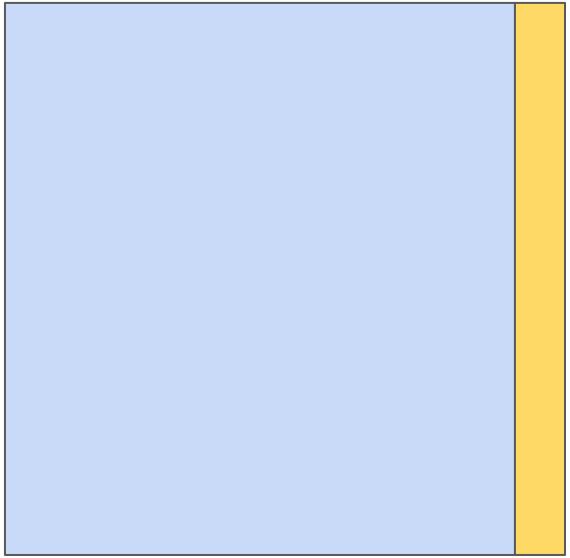


The data of early **statistical learning**

We need the terminology because data comes in many forms

Tall data

$p < n$



The data of early **statistical learning**

Wide data

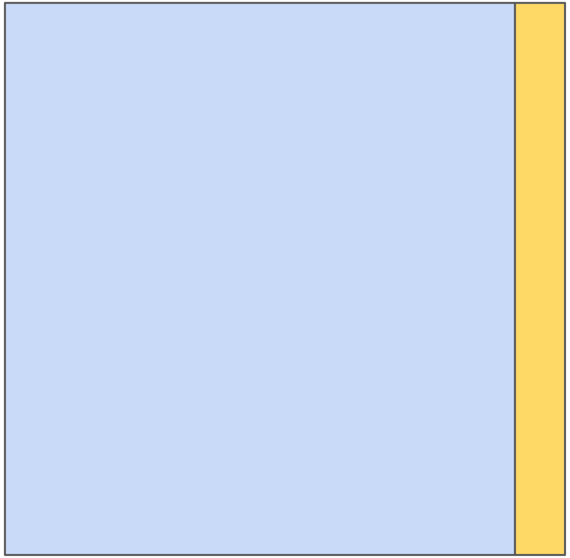
$p > n$

The data of later **statistical learning**

We need the terminology because data comes in many forms

Tall data

$p < n$



The data of early **statistical learning**

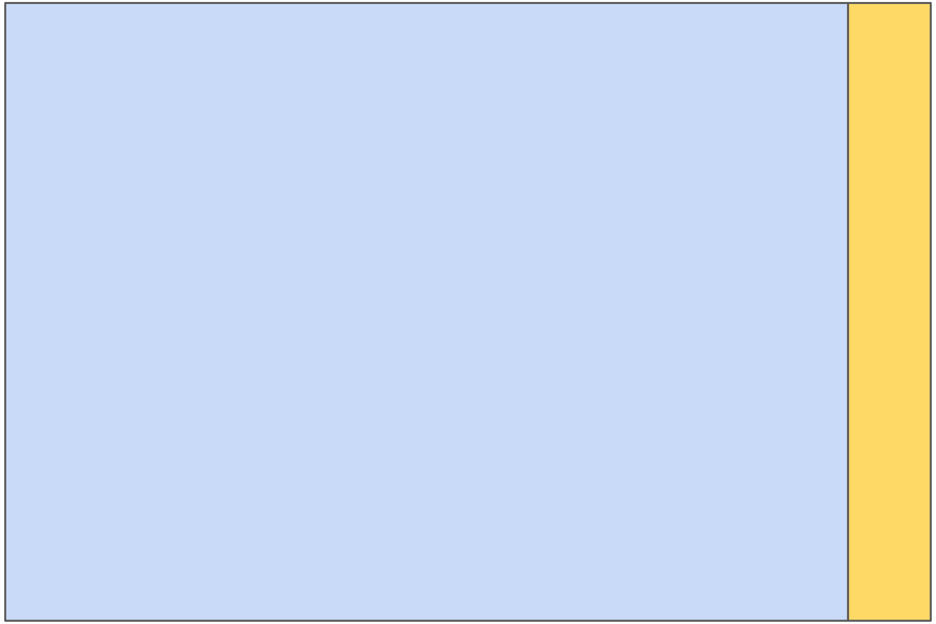
Wide data

$p > n$

The data of later **statistical learning**

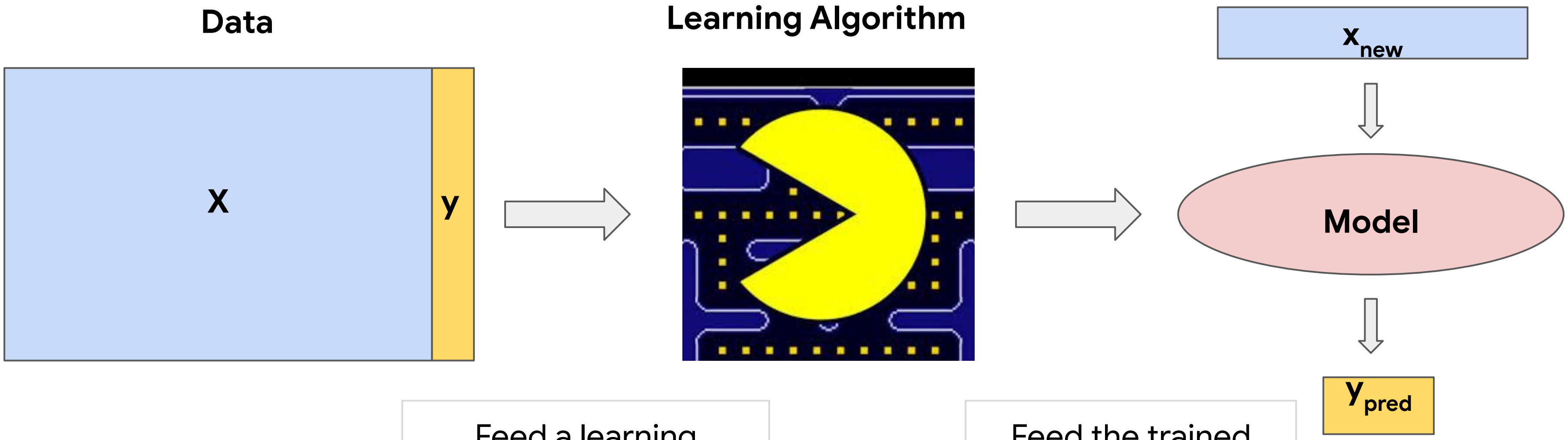
We need the terminology because data comes in many forms

The 'biggie'
big p and/or big n



The data of **modern statistical learning + machine learning + data science**

Now matter what you call it, this is what you do



Feed a learning algorithm your data for ***estimation*** or ***training*** of the model.

Feed the trained model new data to make a ***prediction*** about the outcome.

We are going to play with image classification

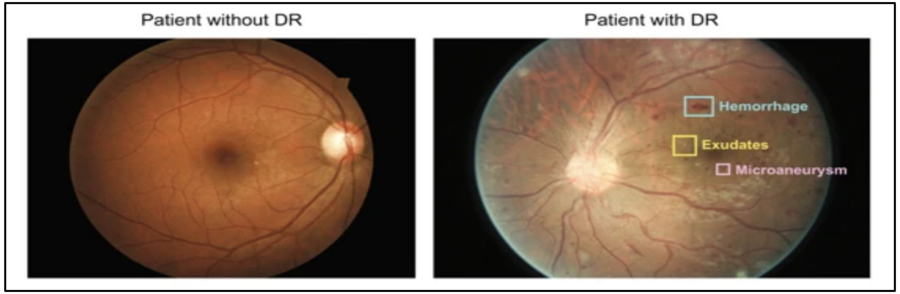
Chihuahua or blueberry muffin?

We are going to play with image classification

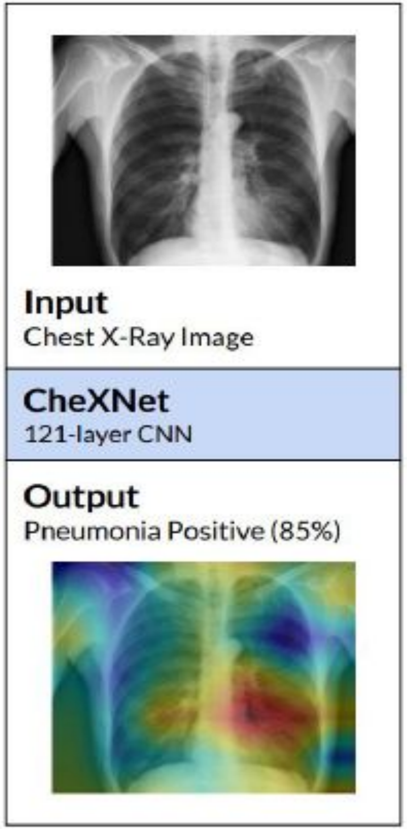
Chihuahua or blueberry muffin?

Labradoodle or fried chicken?

Many real life of examples of this...



Deep learning algorithm predicts diabetic retinopathy progression in individual patients



CheXNet: Radiologist-level pneumonia detection on chest x-rays with deep learning

How we’ll get there

- Learn the recipe for cooking up a statistical learning model
- Review some techniques for modeling
- Summarize methods for model evaluation