# Lecture 000

## Why are we here?

Edward Rubin

# Admin

# Admin

## In-class today

- *Course website:* https://github.com/edrubin/EC524W26/
- Syllabus (on website)

## TODO list

- *Today:* Sign up for Kaggle
- Upcoming readings:
  - ISL Ch1–Ch2
  - Prediction Policy Problems by Kleinberg *et al.* (2015)
- *Assignment:* This week? (Getting to know prediction and Kaggle)

# What's the goal?

# What's the goal?

## What's different?

We've got a whole class on **prediction**. Why?

# What's the goal?

## What's different?

We've got a whole class on **prediction**. Why?

Up to this point, we've focused on causal **identification/inference** of $\beta$, *i.e.*,

$$\mathbf{Y}_i = \mathbf{X}_i\beta + \boldsymbol{u}_i$$

meaning we want an unbiased (consistent) and precise estimate $\hat{\beta}$.

# What's the goal?

## What's different?

We've got a whole class on **prediction**. Why?

Up to this point, we've focused on causal **identification/inference** of $\beta$, *i.e.*,

$$\mathbf{Y}_i = \mathbf{X}_i\beta + \boldsymbol{u}_i$$

meaning we want an unbiased (consistent) and precise estimate $\hat{\beta}$.

With **prediction**, we shift our focus to accurately estimating outcomes.

In other words, how can we best construct $\hat{\mathbf{Y}}_i$?

# What's the goal?

## ... so?

So we want "nice"-performing estimates $\hat{y}$ instead of $\hat{\beta}$.

Q Can't we just use the same methods (*i.e.*, OLS)?

# What's the goal?

## ... so?

So we want "nice"-performing estimates $\hat{y}$ instead of $\hat{\beta}$.

**Q** Can't we just use the same methods (*i.e.*, OLS)?

**A** It depends.

# What's the goal?

## ... so?

So we want "nice"-performing estimates $\hat{y}$ instead of $\hat{\beta}$.

**Q** Can't we just use the same methods (*i.e.*, OLS)?

**A** It depends. How well does your **linear**-regression model approximate the underlying data? (And how do you plan to select your model?)

# What's the goal?

## ... so?

So we want "nice"-performing estimates $\hat{y}$ instead of $\hat{\beta}$.

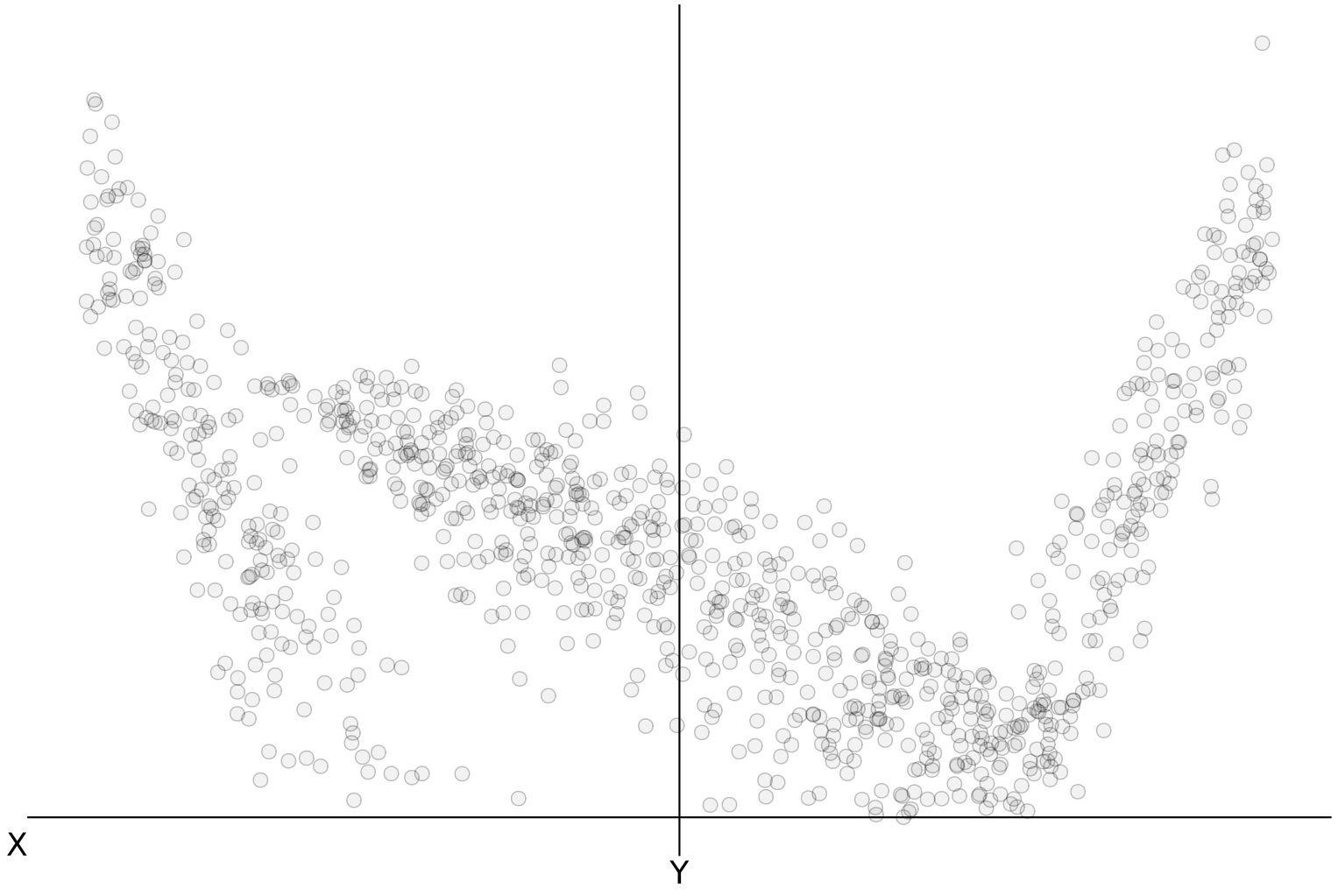
Q Can't we just use the same methods (*i.e.*, OLS)?

A It depends. How well does your **linear**-regression model approximate the underlying data? (And how do you plan to select your model?)

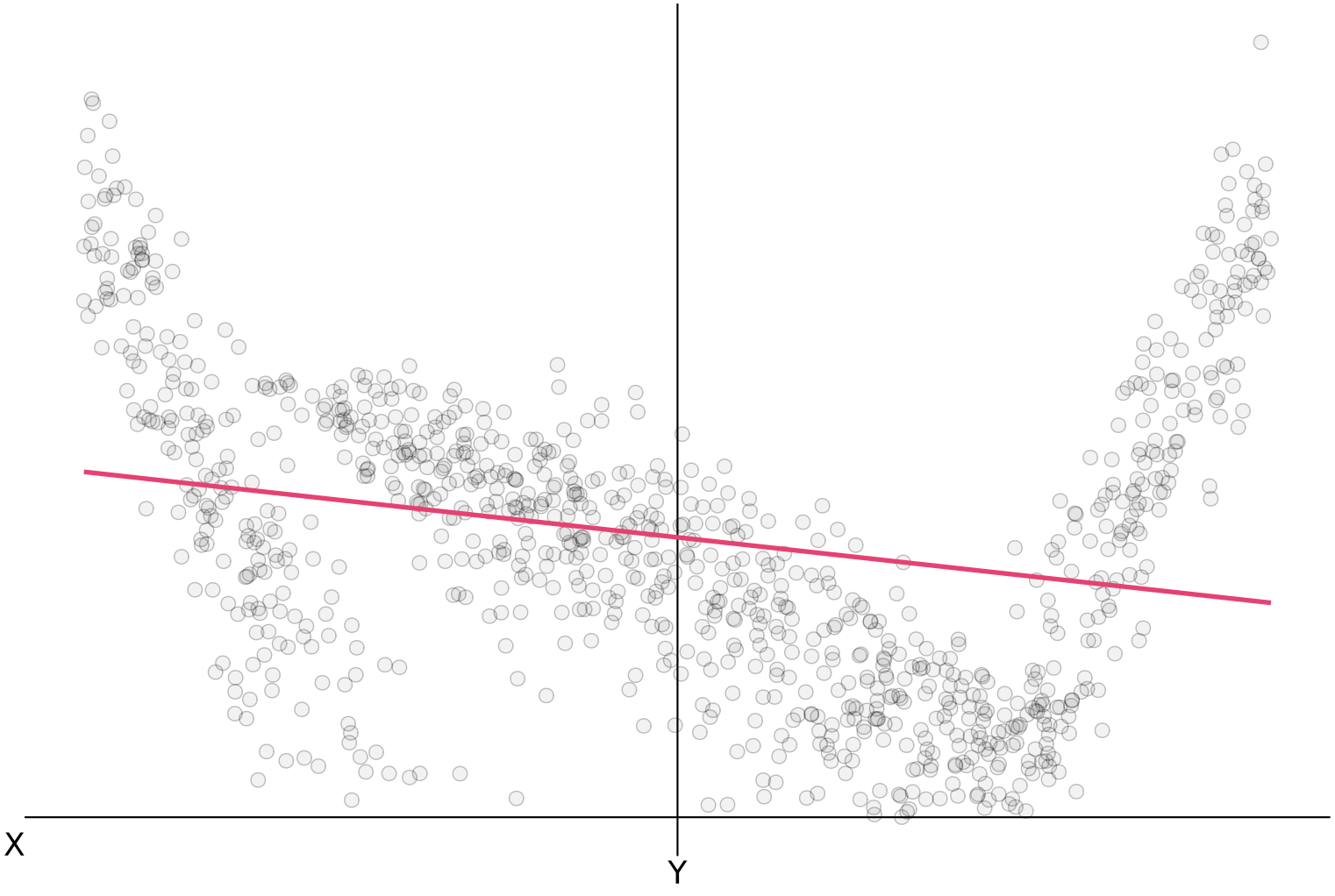
*Recall* Least-squares regression is a great **linear** estimator.

Data can be tricky[†]—as can understanding many relationships.

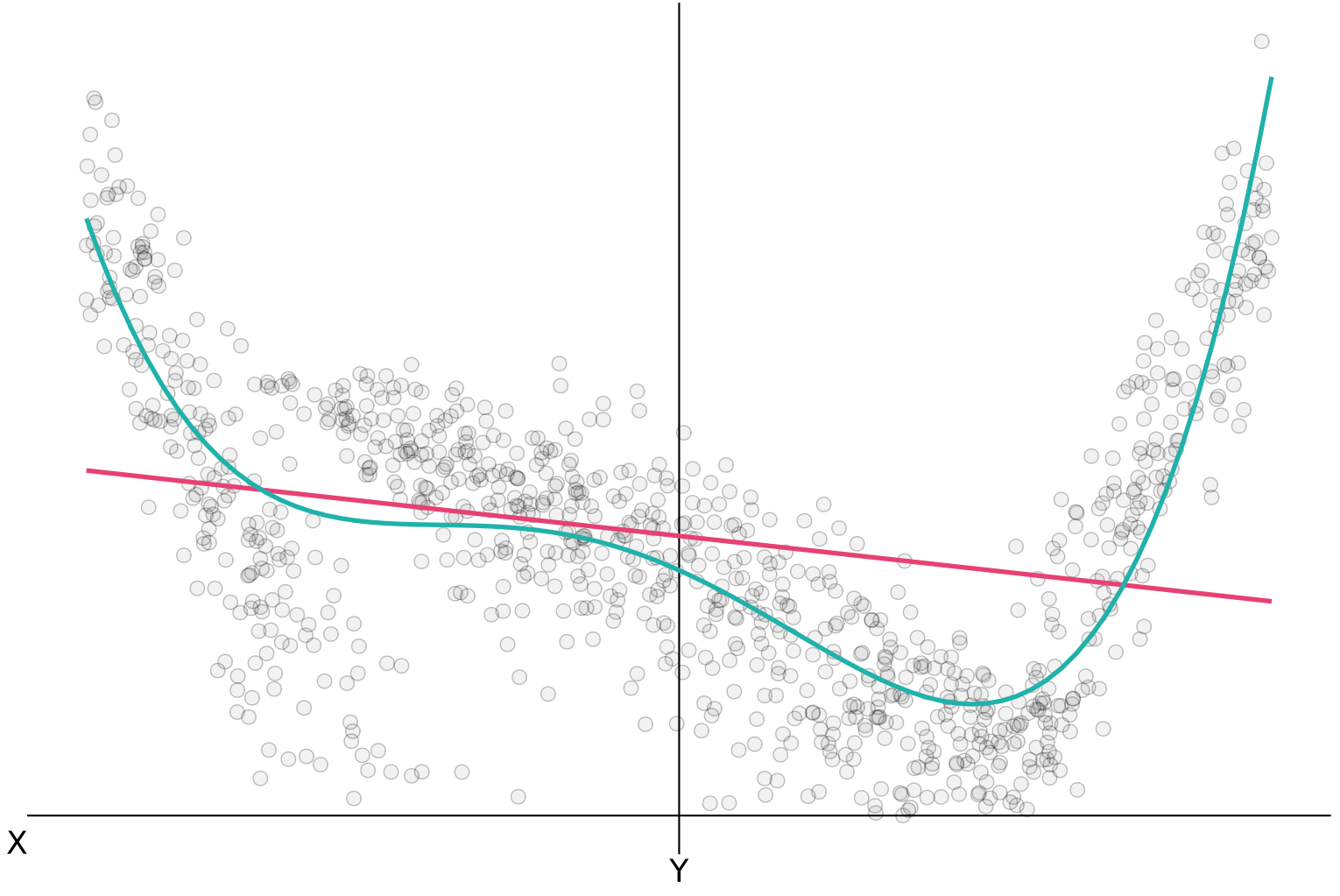
† A typo previously had this slide saying "Data data be tricky", which I really like. "Tricky" might mean nonlinear... or many other things...


X
Y

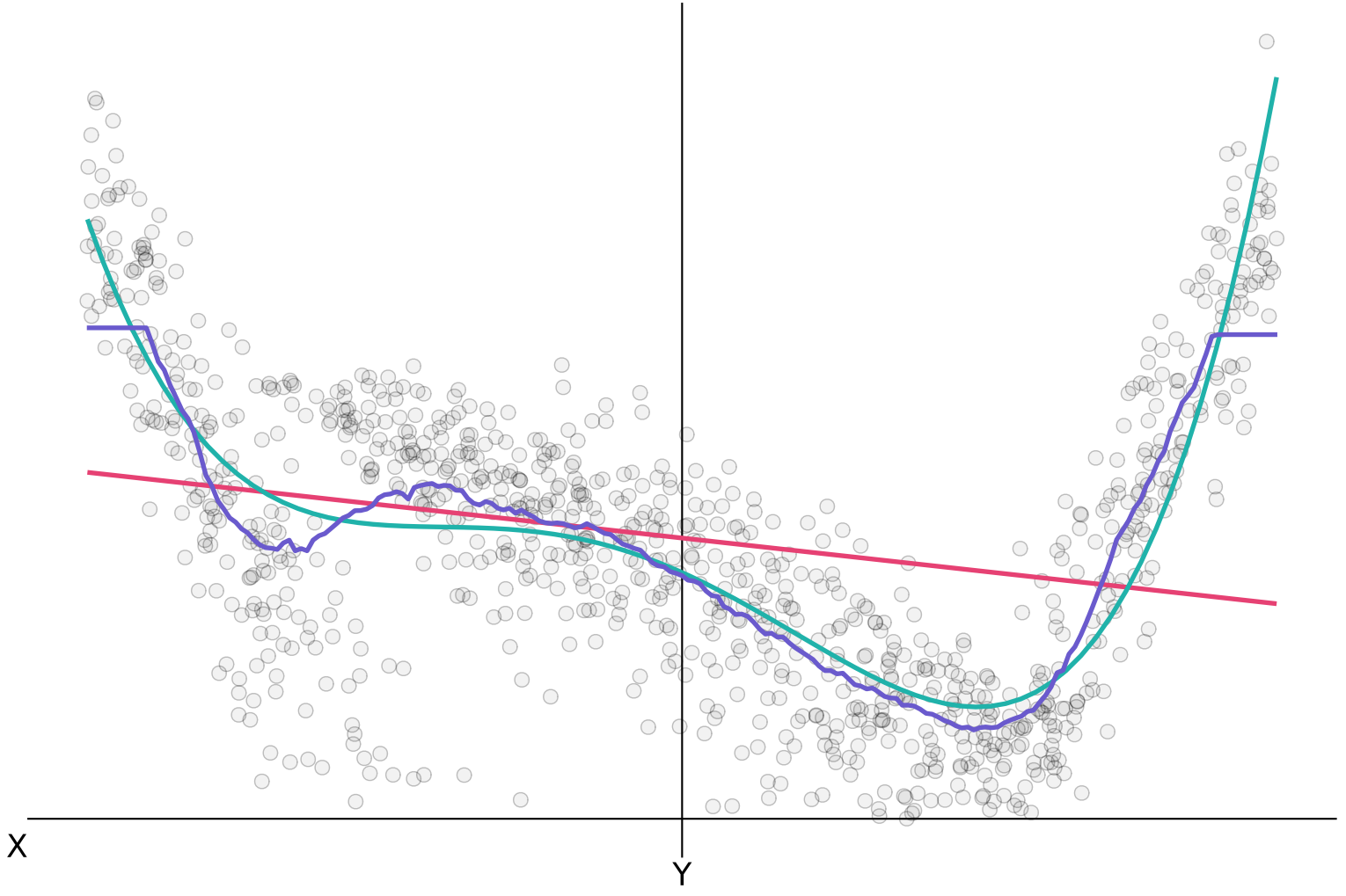
# Linear regression

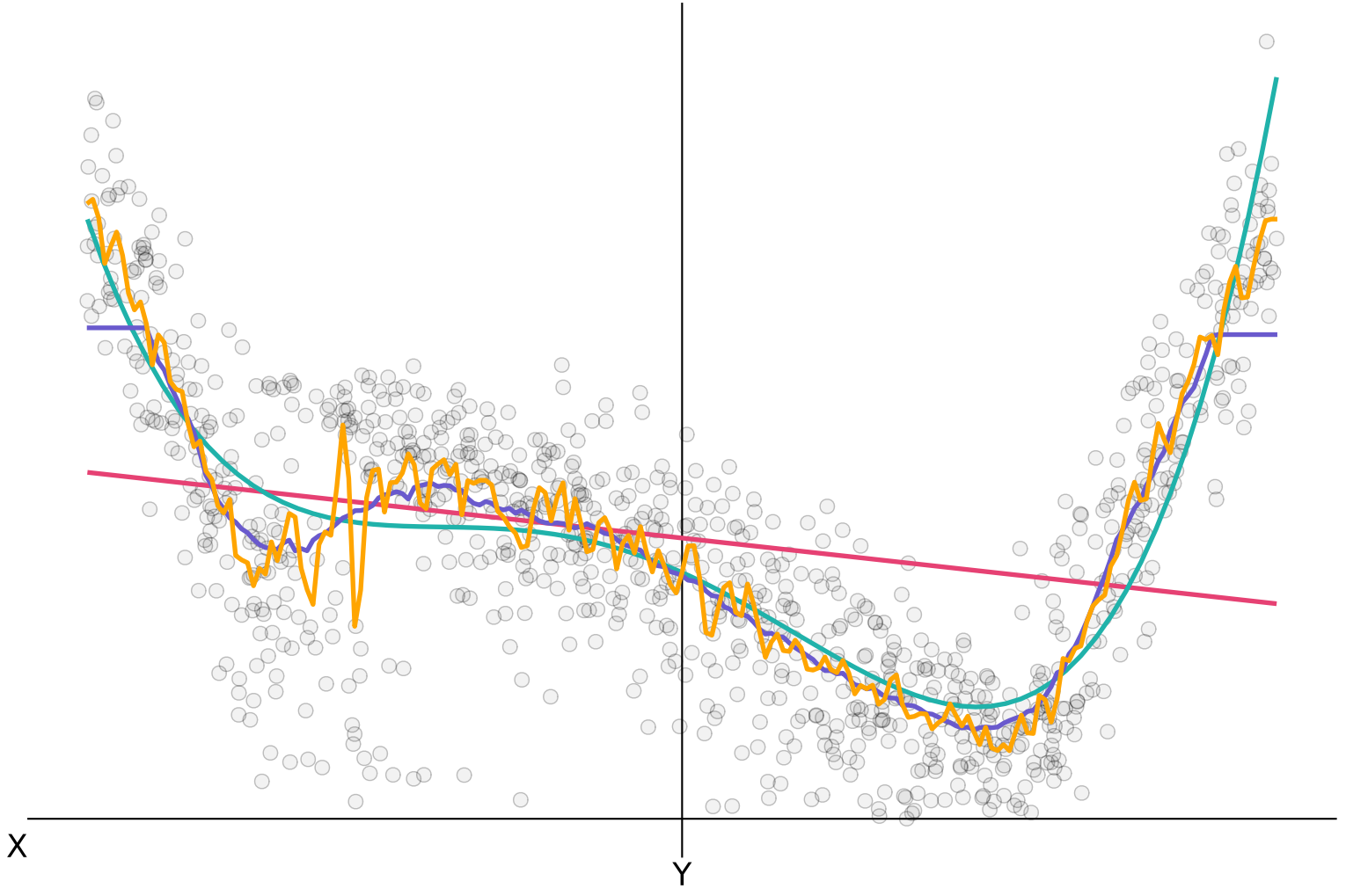


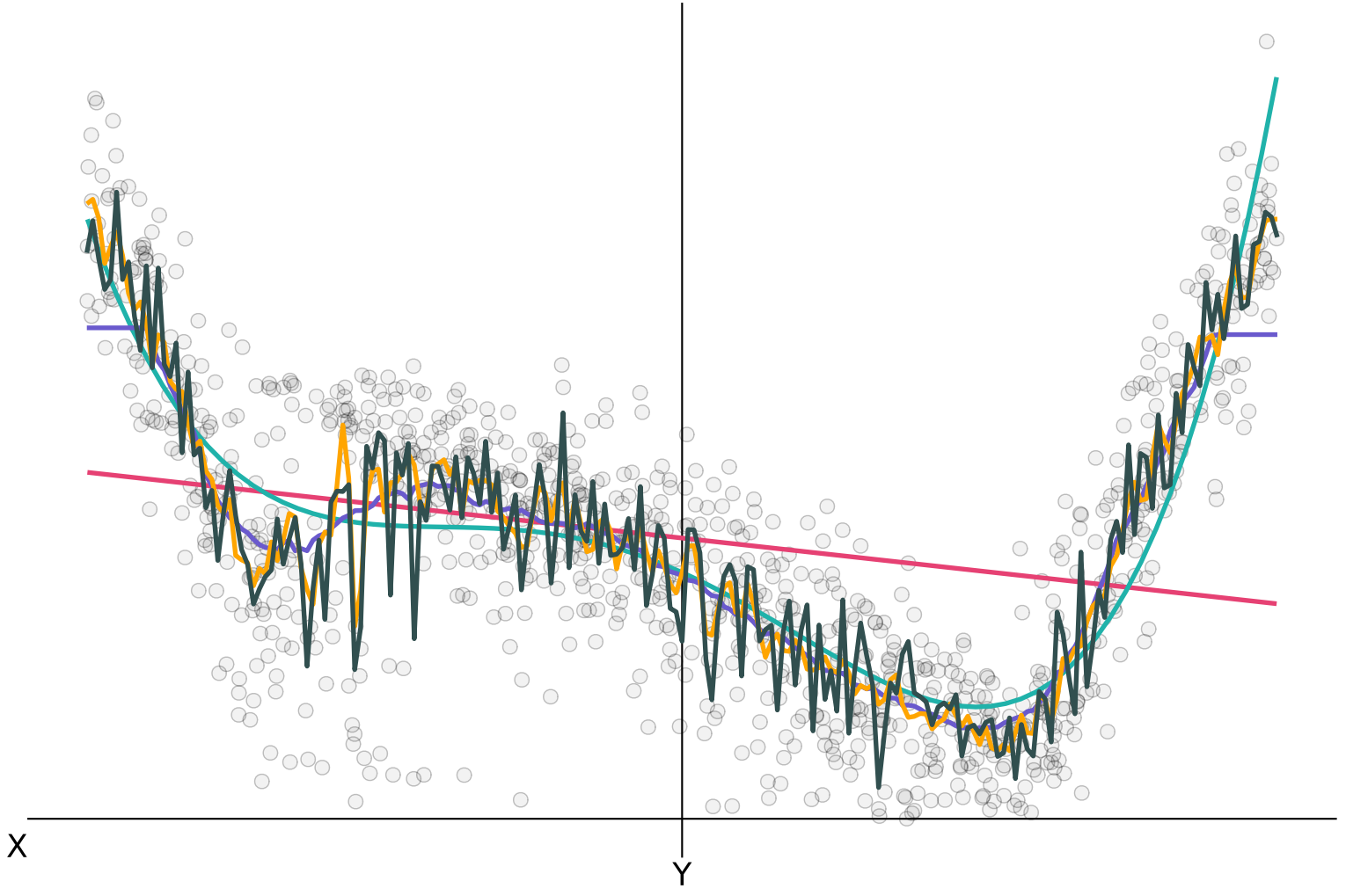
Linear regression, linear regression $\left(x^4\right)$



Linear regression, linear regression $\left(x^4\right)$, KNN (100)




Linear regression, linear regression $(x^4)$, KNN (100), KNN (10)
X
Y


Linear regression, linear regression $(x^4)$, KNN (100), KNN (10), random forest
X
Y

*Note* That example only had one predictor...

# What's the goal?

## Tradeoffs

In prediction, we constantly face many tradeoffs, *e.g.*,

- **flexibility** and **parametric structure** (and interpretability)
- performance in **training** and **test** samples
- **variance** and **bias**

# What's the goal?

## Tradeoffs

In prediction, we constantly face many tradeoffs, *e.g.*,

- **flexibility** and **parametric structure** (and interpretability)
- performance in **training** and **test** samples
- **variance** and **bias**

As your economic training should have predicted, in each setting, we need to **balance the additional benefits and costs** of adjusting these tradeoffs.

# What's the goal?

## Tradeoffs

In prediction, we constantly face many tradeoffs, *e.g.*,

- **flexibility** and **parametric structure** (and interpretability)
- performance in **training** and **test** samples
- **variance** and **bias**

As your economic training should have predicted, in each setting, we need to **balance the additional benefits and costs** of adjusting these tradeoffs.

Many machine-learning (ML) techniques/algorithms are crafted to optimize with these tradeoffs, but the practitioner (you) still needs to be careful.

# What's the goal?

There are many reasons to step outside the world of linear regression...

# What's the goal?

There are many reasons to step outside the world of linear regression...

**Multi-class** classification problems

- Rather than {0,1}, we need to classify $y_i$ into 1 of K classes
- *E.g.*, ER patients: {heart attack, drug overdose, stroke, nothing}

# What's the goal?

There are many reasons to step outside the world of linear regression...

**Multi-class** classification problems

- Rather than {0,1}, we need to classify $y_i$ into 1 of K classes
- *E.g.*, ER patients: {heart attack, drug overdose, stroke, nothing}

**Text analysis** and **image recognition**

- Comb though sentences (pixels) to glean insights from relationships
- *E.g.*, detect sentiments in tweets or roof-top solar in satellite imagery

# What's the goal?

There are many reasons to step outside the world of linear regression...

**Multi-class** classification problems

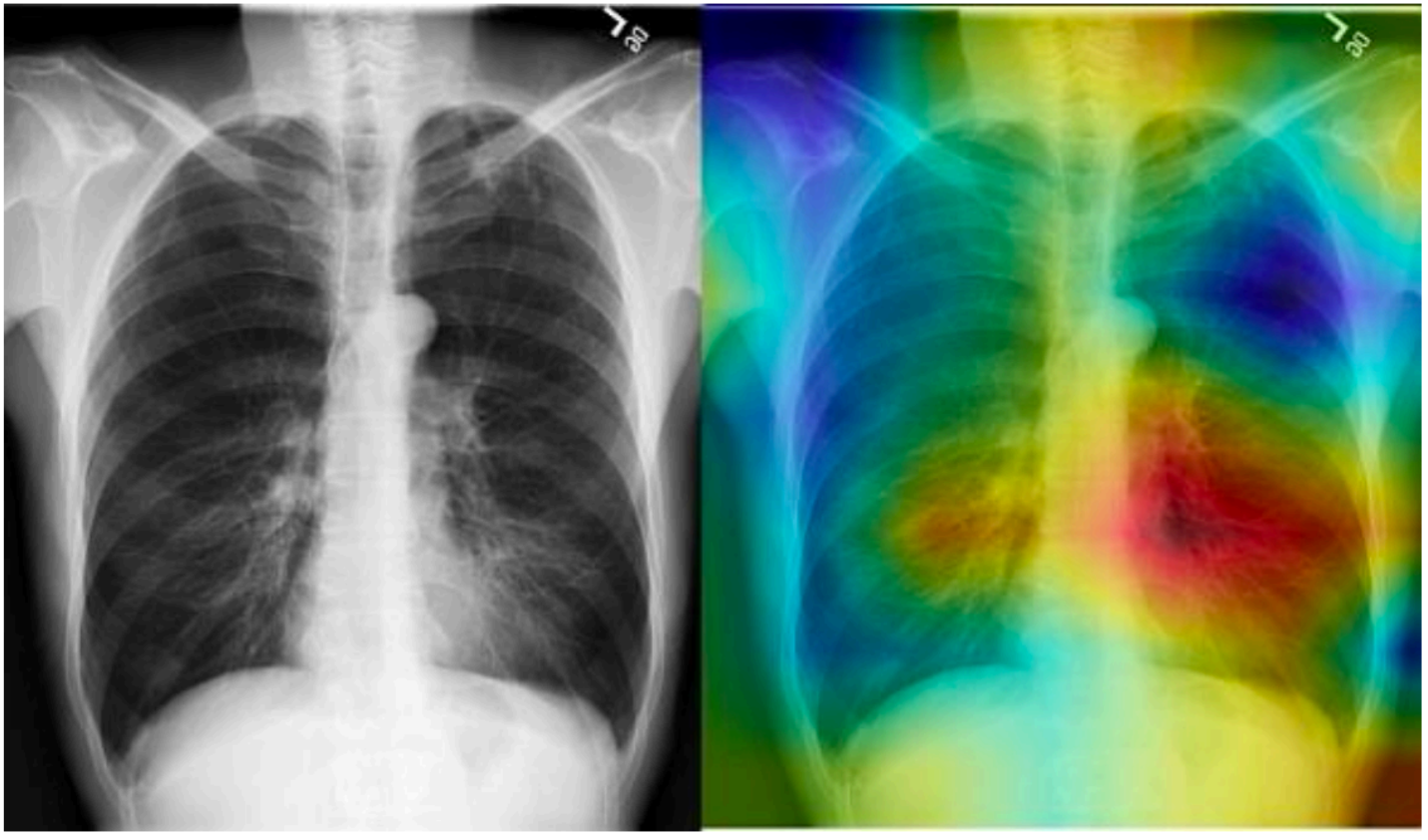
- Rather than {0,1}, we need to classify $y_i$ into 1 of K classes
- *E.g.*, ER patients: {heart attack, drug overdose, stroke, nothing}

**Text analysis** and **image recognition**

- Comb though sentences (pixels) to glean insights from relationships
- *E.g.*, detect sentiments in tweets or roof-top solar in satellite imagery

**Unsupervised learning**

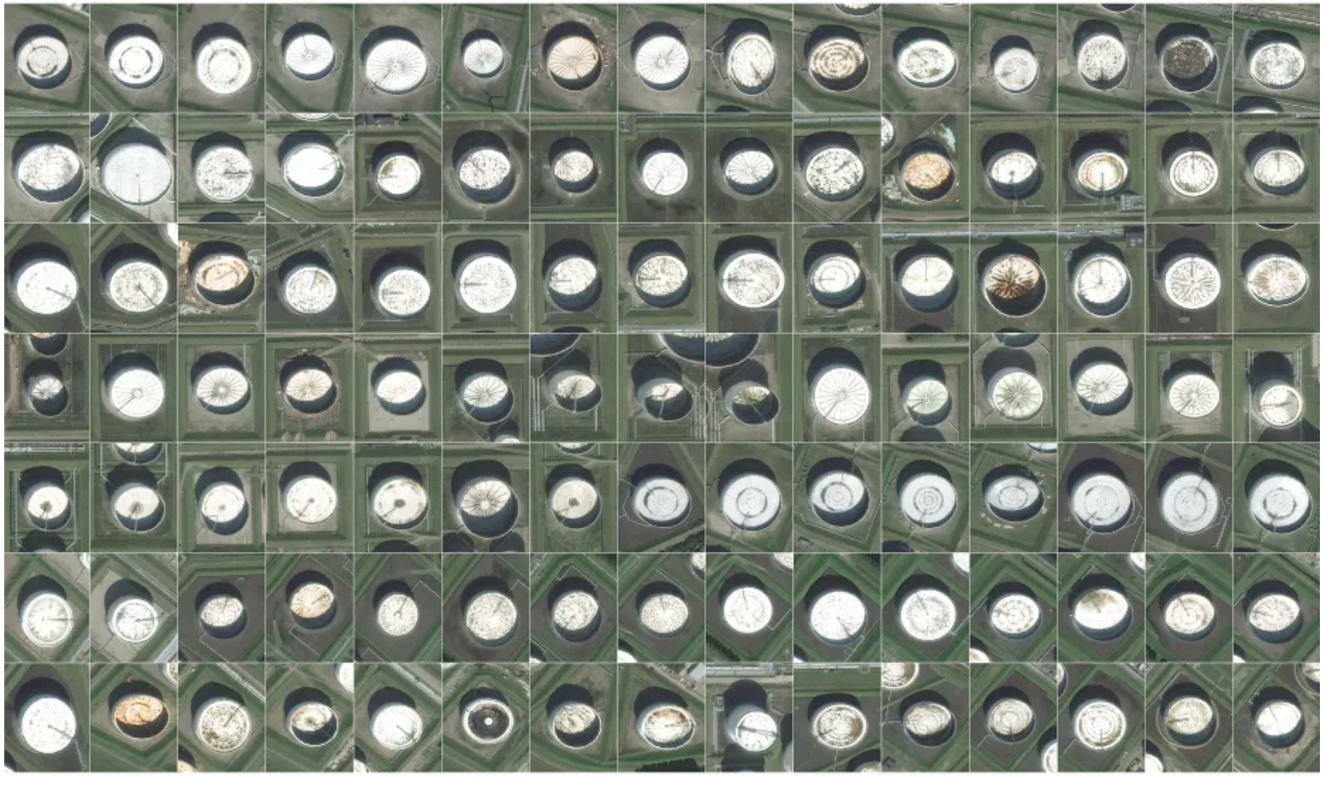
- You don't know groupings, but you think there are relevant groups
- *E.g.*, classify spatial data into groups



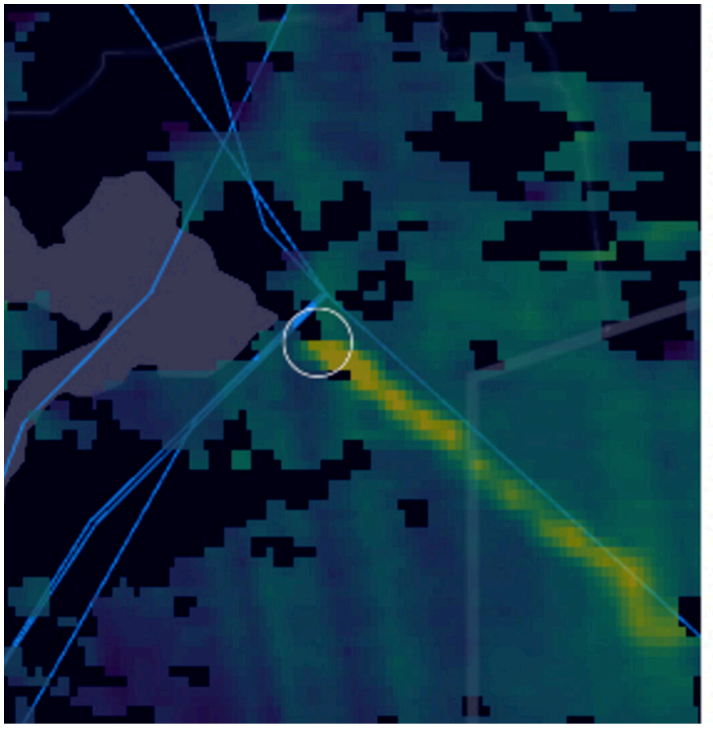
**Stanford University (Stanford, CA ) researchers have developed a deep-learning algorithm that can evaluate chest X-ray images for signs of disease at a level exceeding practicing radiologists.**

# Parking Lot Vehicle Detection Using Deep Learning

# How AI Can Calculate Our Oil Surplus...From Space



ORBITAL INSIGHT/DIGITALGLOBE



# Monitoring methane emissions from gas pipelines

THE NEW YORKER

A REPORTER AT LARGE OCTOBER 14, 2019 ISSUE

# The Next Word

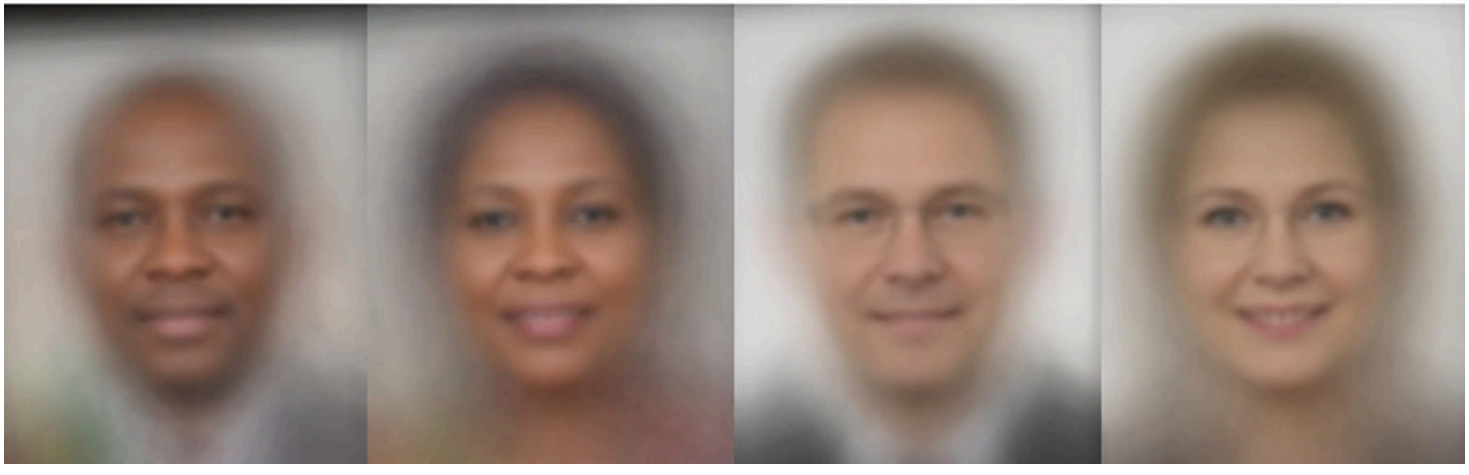
*Where will predictive text take us?*

Text by John Seabrook

And of course...

OpenAI, ChatGPT, Copilot, Claude, Midjourney ...

| Gender Classifier | Darker Male | Darker Female | Lighter Male | Lighter Female | Largest Gap |
|---|---|---|---|---|---|
| Microsoft | 94.0% | 79.2% | 100% | 98.3% | 20.8% |
| FACE++ | 99.3% | 65.5% | 99.2% | 94.0% | 33.8% |
| IBM | 88.0% | 65.3% | 99.7% | 92.9% | 34.4% |



# Takeaways?

Any main takeaways/thoughts from these examples?

# Takeaways?

Any main takeaways/thoughts from these examples?

*Mine*

- **interactions and nonlinearities** likely matter;
- **engineering** features/variables can be important;
- *related:* we might not even know **which features** that matter;
- **flexibility** is huge—but we still want to avoid **overfitting**.

# Takeaways?

Any main takeaways/thoughts from these examples?

*Mine*

- **interactions and nonlinearities** likely matter;
- **engineering** features/variables can be important;
- *related:* we might not even know **which features** that matter;
- **flexibility** is huge—but we still want to avoid **overfitting**.

*Ideal* We want to

- flexibly approximate complex relationships...
- while still generalizing to new data...
- and understanding weaknesses/biases.

Bonus for interpretability. (Computation might matter too.)

*Next time* Start formal building blocks of prediction.

# Sources

Sources (articles) of images

- Deep learning and radiology
- Parking lot detection
- *New Yorker* writing
- Oil surplus
- Methane leaks
- Gender Shades

# Table of contents