GENERAL ASSEMBLY

INTRODUCTION TO CORRELATED DATA

Matt Brems, Data Science Immersive

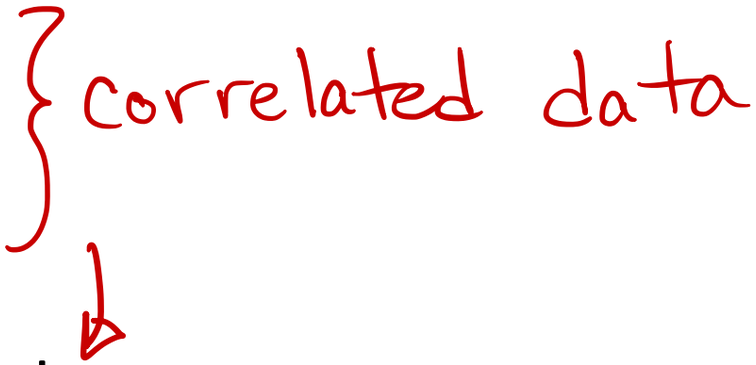
LEARNING OBJECTIVES

By the end of the lesson, students should be able to:

- Describe why our existing modeling tactics don’t work well with correlated data.
- Name three different types of correlated data.
- Identify intuitive ways to detect correlation among observations.
- Describe omitted variable bias.

DATA SCIENCE PROCESS

1. Define the problem.
2. Obtain the data.
3. Explore the data.
4. Model the data.
5. Evaluate the model.
6. Answer the problem.



HOW DO YOU PICK A MODEL?

- When starting new projects, how do you pick which model you want to choose?

I test them all. → Grid Search

Based on type of data. → is Y continuous or discrete?

Depends on our goal!

↳ prediction

↳ inference / interpretability

ASSUMPTIONS

- Underlying each modeling tactic we've used, there have been some assumptions.

- **Parametric modeling** tactics make assumptions about the distributions of our data.

Linear regression ($\varepsilon \sim N(0, \sigma)$) Naive Bayes - Bernoulli - Multinomial - Gaussian

- **Nonparametric modeling**, while not making assumptions about how our data are distributed, still often assume that our observations are independent of each other.

Random Forests k-Nearest Neighbors

- **The most common assumption we'll make in modeling is that our observations are independent of one another**.

INDEPENDENT OBSERVATIONS

- In many cases, this is perfectly reasonable. If I take a random sample of 300 voters, it's rational for me to assume our data are independent.

- Even in cases where this is slightly violated, we'll believe it to be reasonable. If my random sample of 300 voters included two members of the same household, we'd almost certainly proceed with the assumption that our data are independent.

- Unfortunately, it isn't always reasonable for us to assume that our observations are independent of one another.

CORRELATED DATA WEEK

- This week, we're going to talk about "correlated" data, which refers to data that is correlated with itself.

- Examples of correlated data include **time series** data, **spatial** data, and **network** data.

CORRELATED EXAMPLES

- In thinking about the following examples, let’s consider how the data we observe are not independent of one another.

CORRELATED EXAMPLES

Y_t = stock price at time t

Cross-sectional data: data observed at one moment in time

CORRELATED EXAMPLES

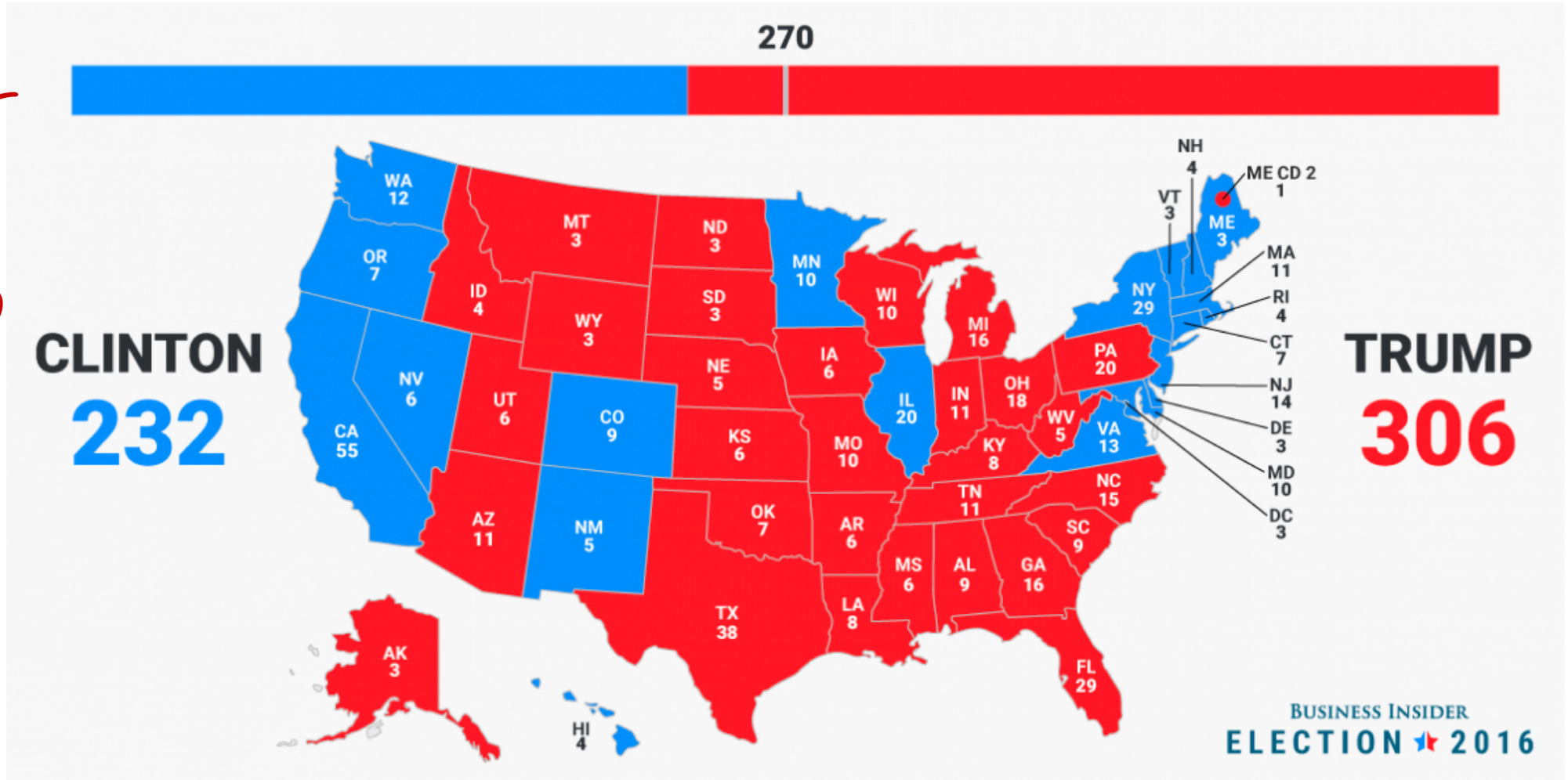
Y_t = # of electoral votes won by Clinton

$Y_{WA} = 12$

$Y_{ID} = 0$

$Y_{MT} = 0$

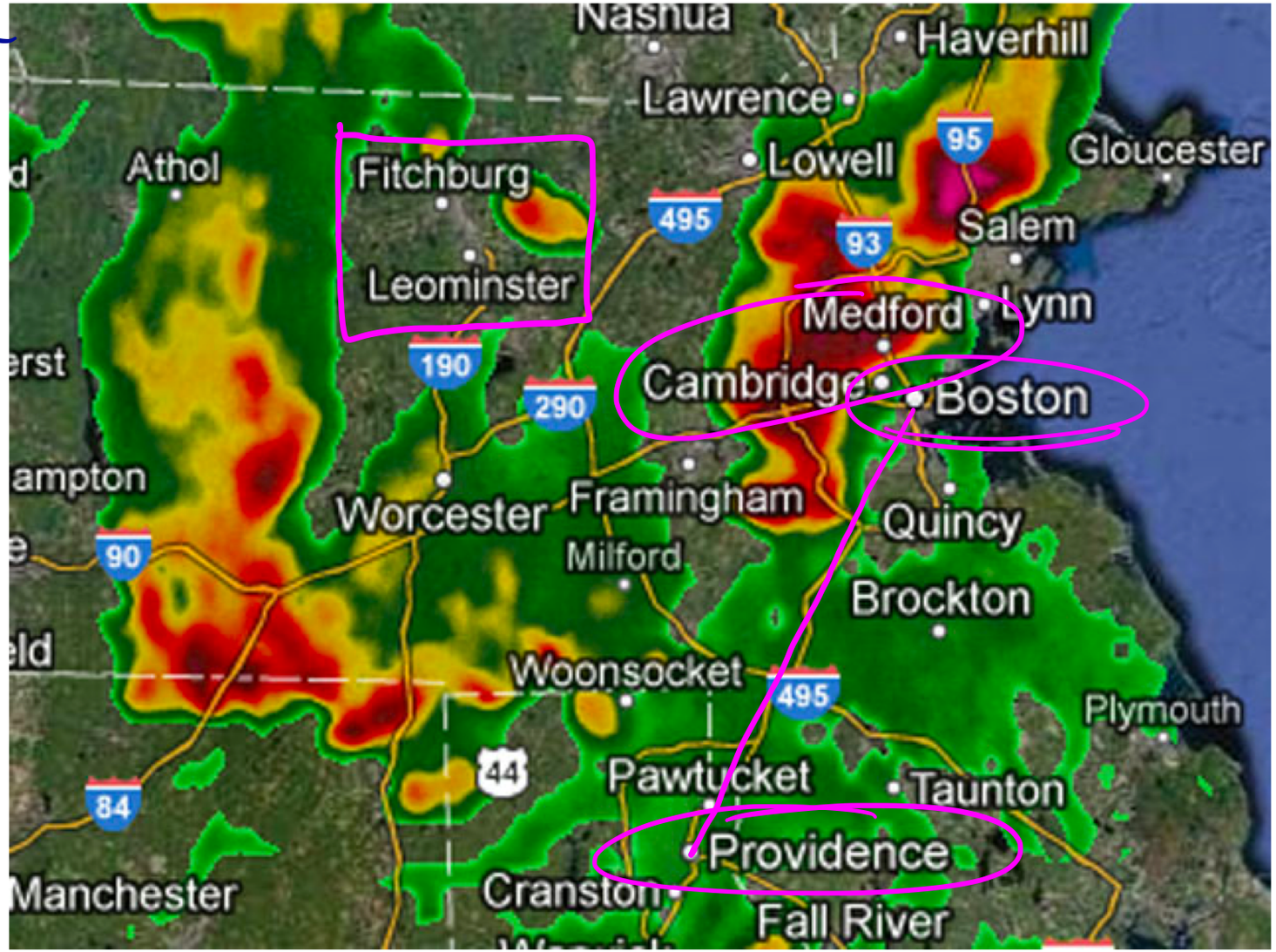
⋮



CORRELATED EXAMPLES

correlation over space

correlation over time



CORRELATED EXAMPLES

- It would be possible for us to consider these data as independent of one another. My DataFrame of stock price data might include stock price as the Y, time and other variables as our X. **Pandas won't throw an error if we try to fit this model**.

- …but we should be able to do better.

GROUP ACTIVITY #1

- Before building a model with the time series data here…

Discuss until :57 after the hour!

- 1. How might I detect temporal dependence of my observations? → across time

- 2. When modeling, how could I try to account for this dependence?

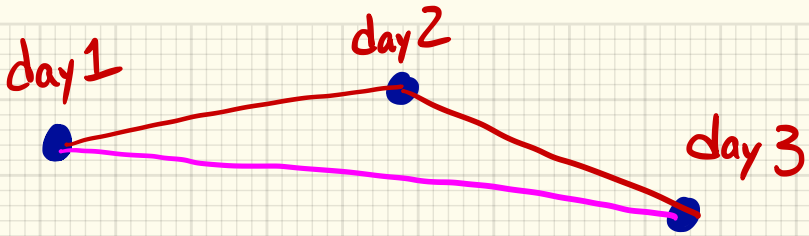
GROUP ACTIVITY #1: DISCUSSION

- 1. How might I detect temporal dependence of my observations?

$Corr(Y_t, Y_{t-1})$ $Corr(Y_t, Y_{t+1})$

- 2. When modeling, how could I try to account for this dependence?

add in a time variable



interpolation

GROUP ACTIVITY #2

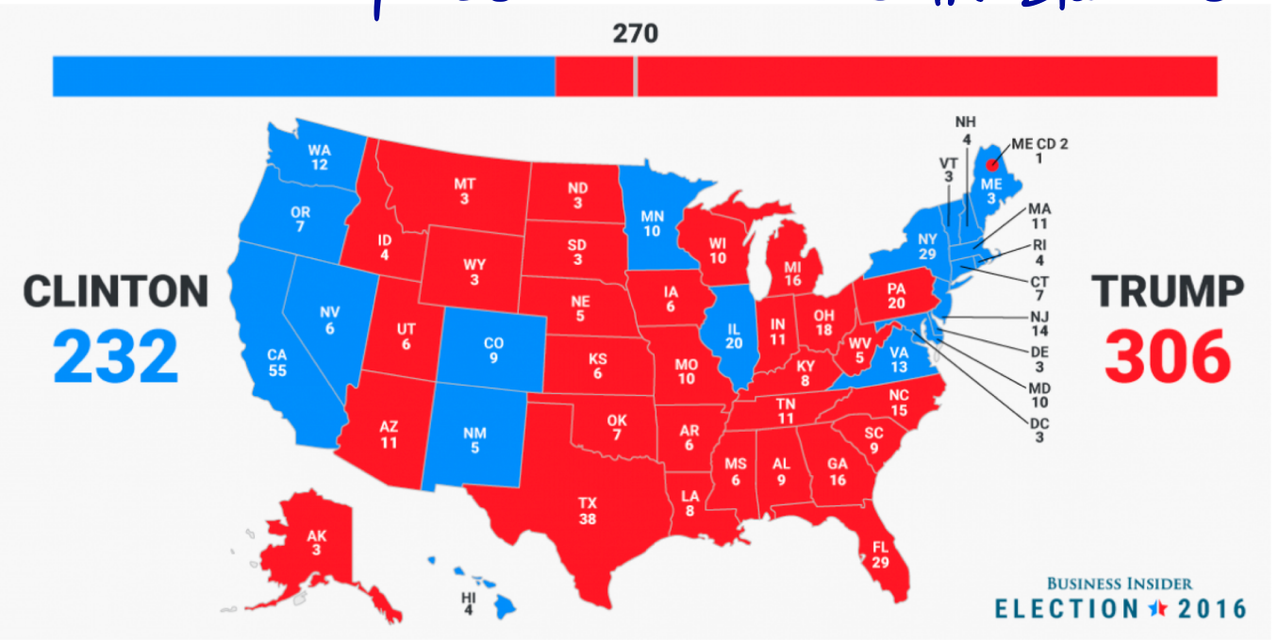
- Before building a model with spatial data here…

Discuss until :10 after hour!

- 1. How might I detect spatial dependence of my observations?

- 2. When modeling, how could I try to account for this dependence?

Y_t = # of electoral votes won by Dem. candidate in state t



GROUP ACTIVITY #2: DISCUSSION

- 1. How might I detect spatial dependence of my observations?

Correlation between state t's results and the results in "neighboring" states. → first degree neighbors
within 200 miles ↵ ↳ which correlations are strong

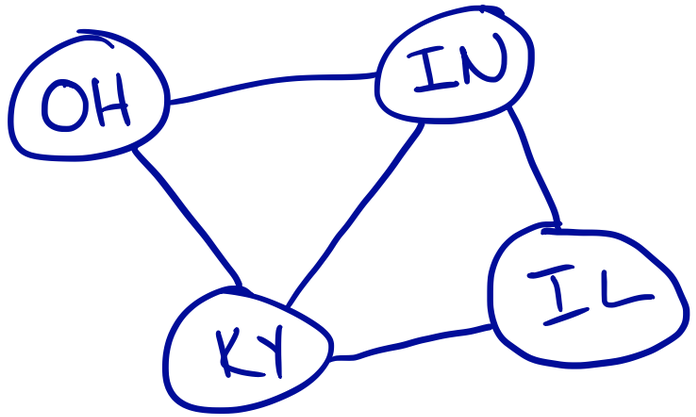
- 2. When modeling, how could I try to account for this dependence?

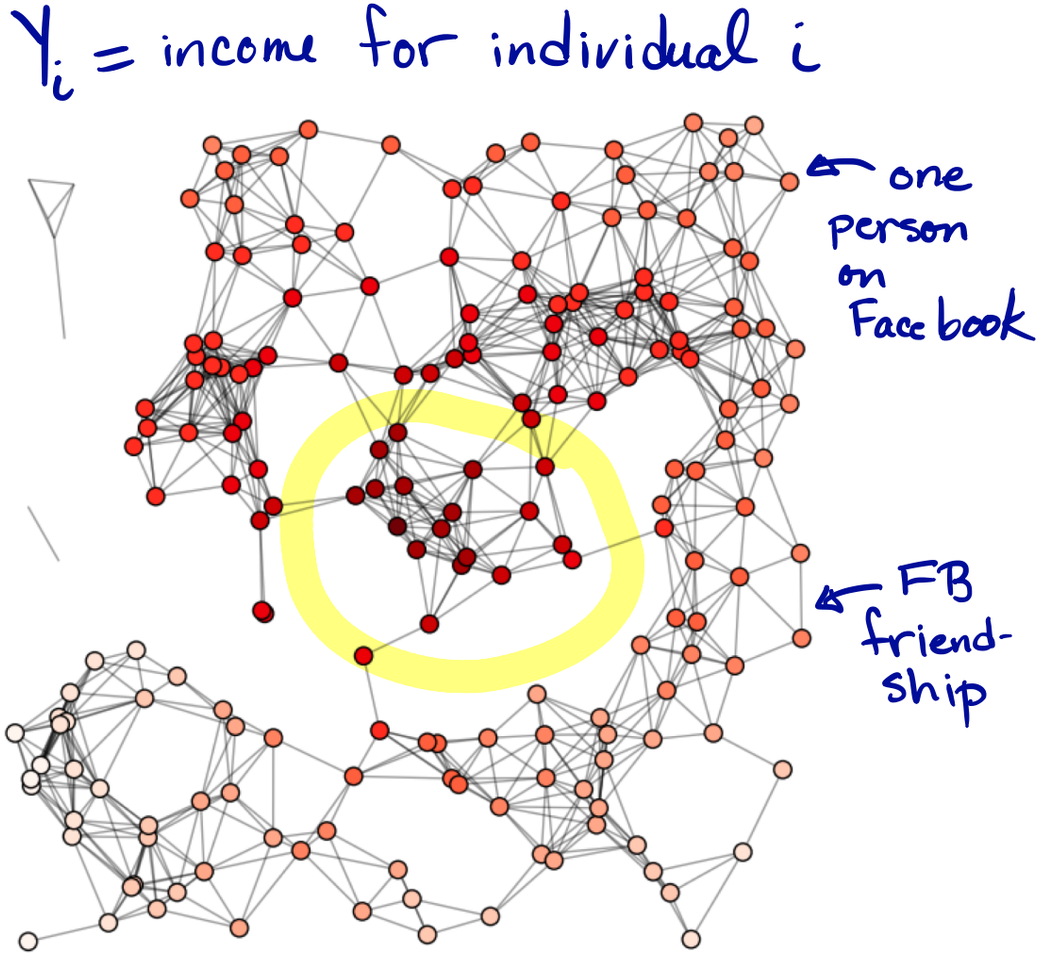
"Bin" the U.S. into larger buckets.
↳ statistical method
↳ split by geographic region

NETWORK EXAMPLE

- Suppose each connection here indicates Facebook friendship.

- I want to fit a model predicting income from these individuals.





MODELS WITH HIGH BIAS

- If our model suffers from high error due to bias, what does this mean?

underfit

model is too simple

model is missing features

OMITTED VARIABLE BIAS

- If your model leaves out important predictors, the bias that occurs is known as **omitted variable bias**. (The bias that exists because we omitted, or left out, an important predictor.)

Real: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ we don't observe this!

Our Model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1$ omitted

OMITTED VARIABLE BIAS

- If your model leaves out important predictors, the bias that occurs is known as **omitted variable bias**. (The bias that exists because we omitted, or left out, an important predictor.)

- In a time series model, that means we might include all of the old values of Y that we think are relevant in predicting the new values of Y!

$$Y_t = \beta_0 + \beta_1 Y_{t-1}$$

autoregressive

- In a spatial model, that means we're going to include all of the values of Y that we think are relevant in predicting other values of Y.

$$Y_{OH} = \beta_0 + \beta_1 Y_{IN} + \beta_2 Y_{PA} + \beta_3 Y_{KY}$$