GENERAL ASSEMBLY

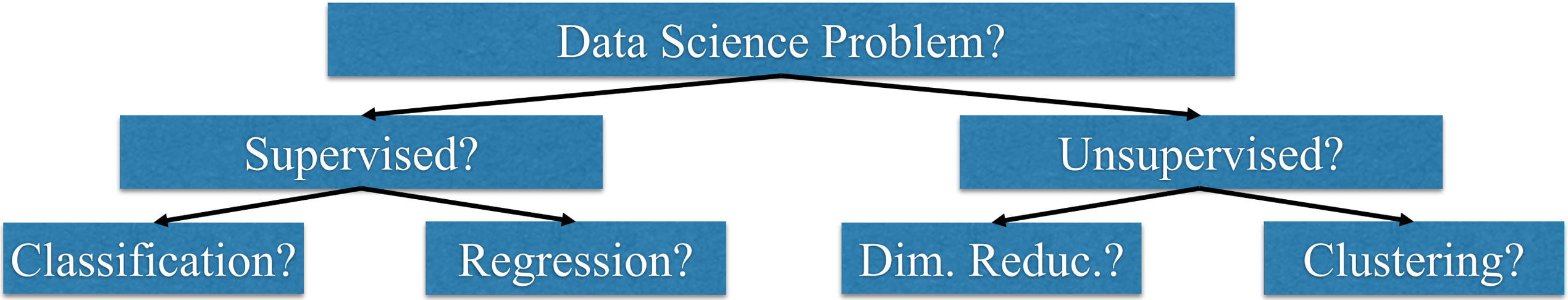
INTRODUCTION TO CORRELATED DATA

Matt Brems, Data Science Immersive

AGENDA

- Review of Existing Tactics
- Correlated Data
- Handling Correlated Data

FRAMING



- K-NN
- Logistic Regression
- SVM
- CART

- K-NN
- Linear Regression
- SVM
- CART

- Feature Elimination
- Feature Extraction
- PCA

- K-means
- DBSCAN
- Hierarchical

FRAMING

- While this hierarchy helps us to understand and organize the different problem-solving methods available, there are other questions we have in mind that help guide our work.

- What is our goal with our project?

- Does this project stand alone, or does it fit in with additional work-related projects?

- Who is using the results of this project?

FRAMING

- While this hierarchy helps us to understand and organize the different problem-solving methods available, there are other questions we have in mind that help guide our work.
- Do we care more about prediction or inference?
- How important are interpretable results?
- How do we control for bias and variance?
- Are we adopting a Bayesian or frequentist approach?
- What are the limitations of my data?

ASSUMPTIONS

- Underlying each modeling tactic we've used, there have been some assumptions.
- Parametric modeling tactics make assumptions about the distributions of our data.
- Nonparametric modeling, while not making assumptions about how our data are distributed, still often assume that our observations are independent of each other.
- In fact, 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.

SPATIOTEMPORAL WEEK

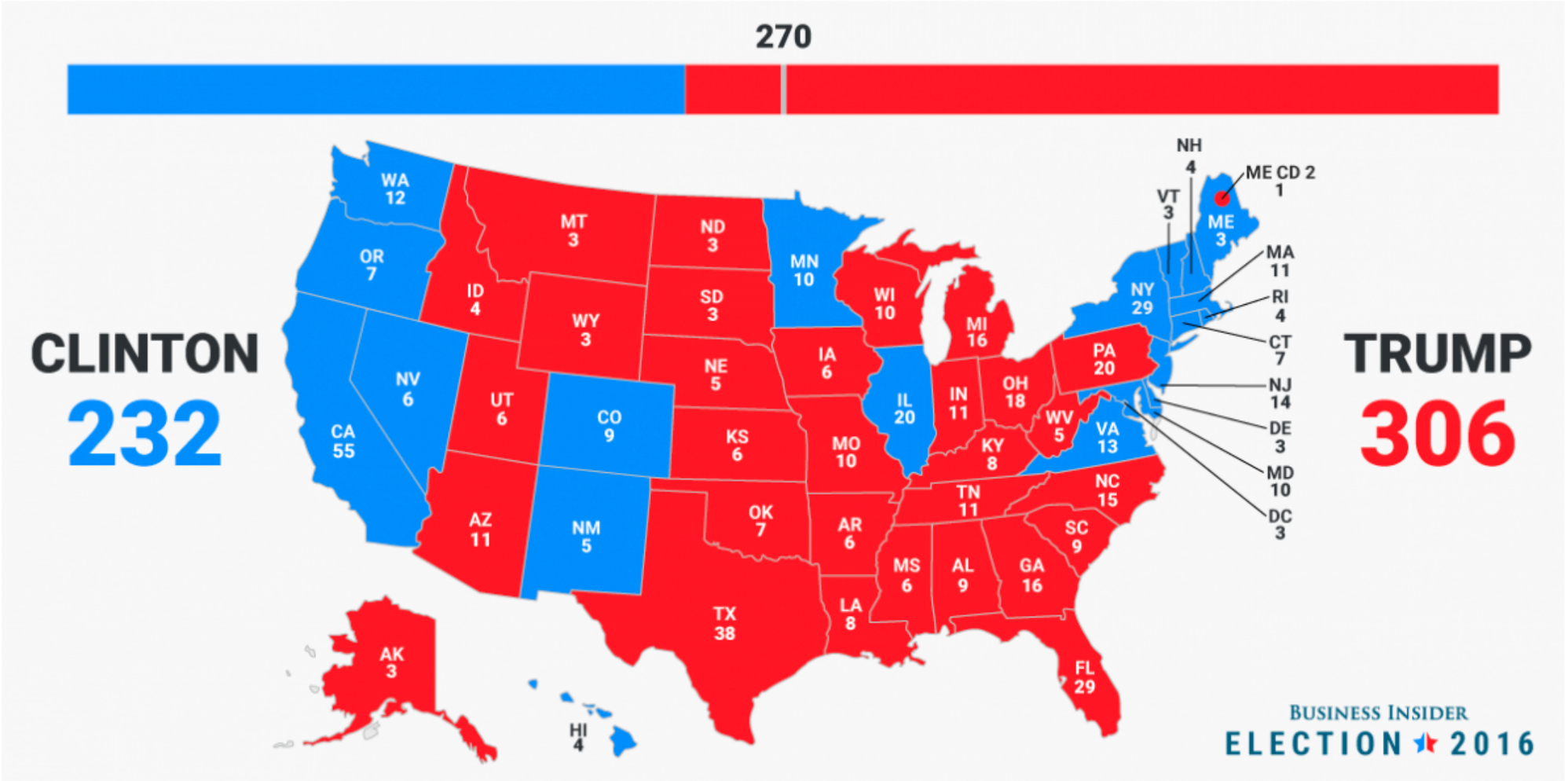
- This week, we're going to talk about "spatiotemporal" data, which just means data that has a space component and a time component.
- In this lecture, we'll introduce both.
- For the first half of the week, we'll spend most of our time on time series data.
- Near the end of the week, we'll discuss spatial data and how to integrate it with temporal data.

SPATIOTEMPORAL WEEK

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

SPATIOTEMPORAL EXAMPLES

SPATIOTEMPORAL EXAMPLES



SPATIOTEMPORAL EXAMPLES

SPATIOTEMPORAL WEEK

- 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…

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

- 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?

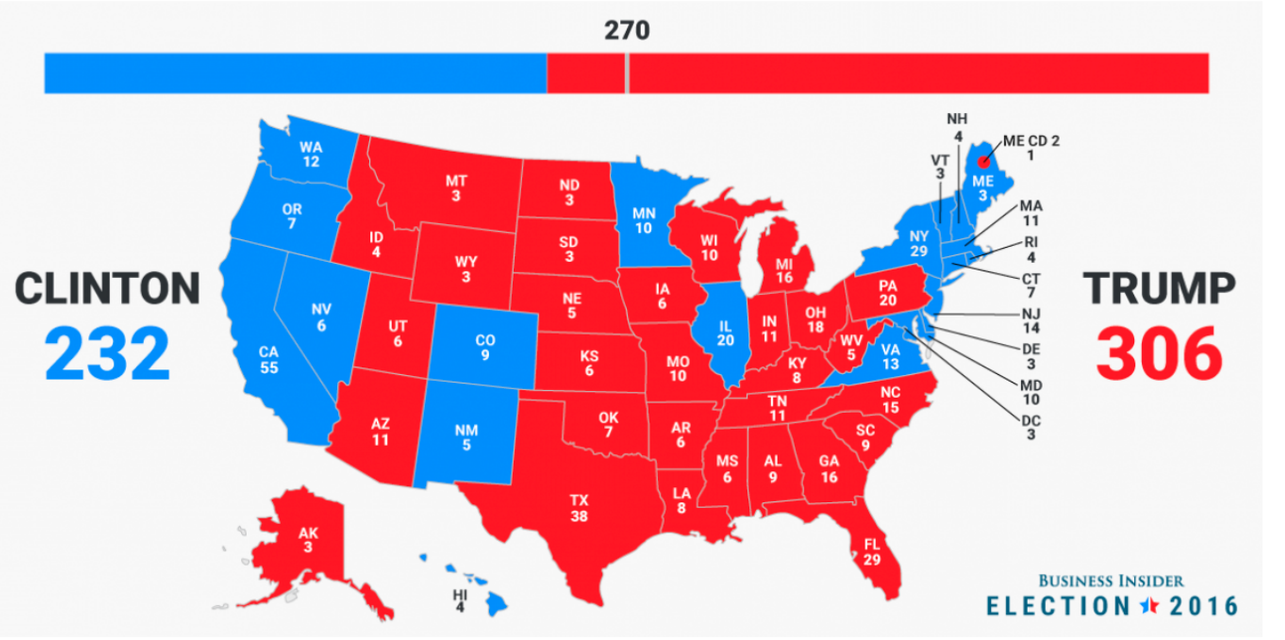
- 2. When modeling, how could I try to account for this dependence?

GROUP ACTIVITY #2

- Before building a model with spatial data here…

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

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



GROUP ACTIVITY #2: DISCUSSION

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

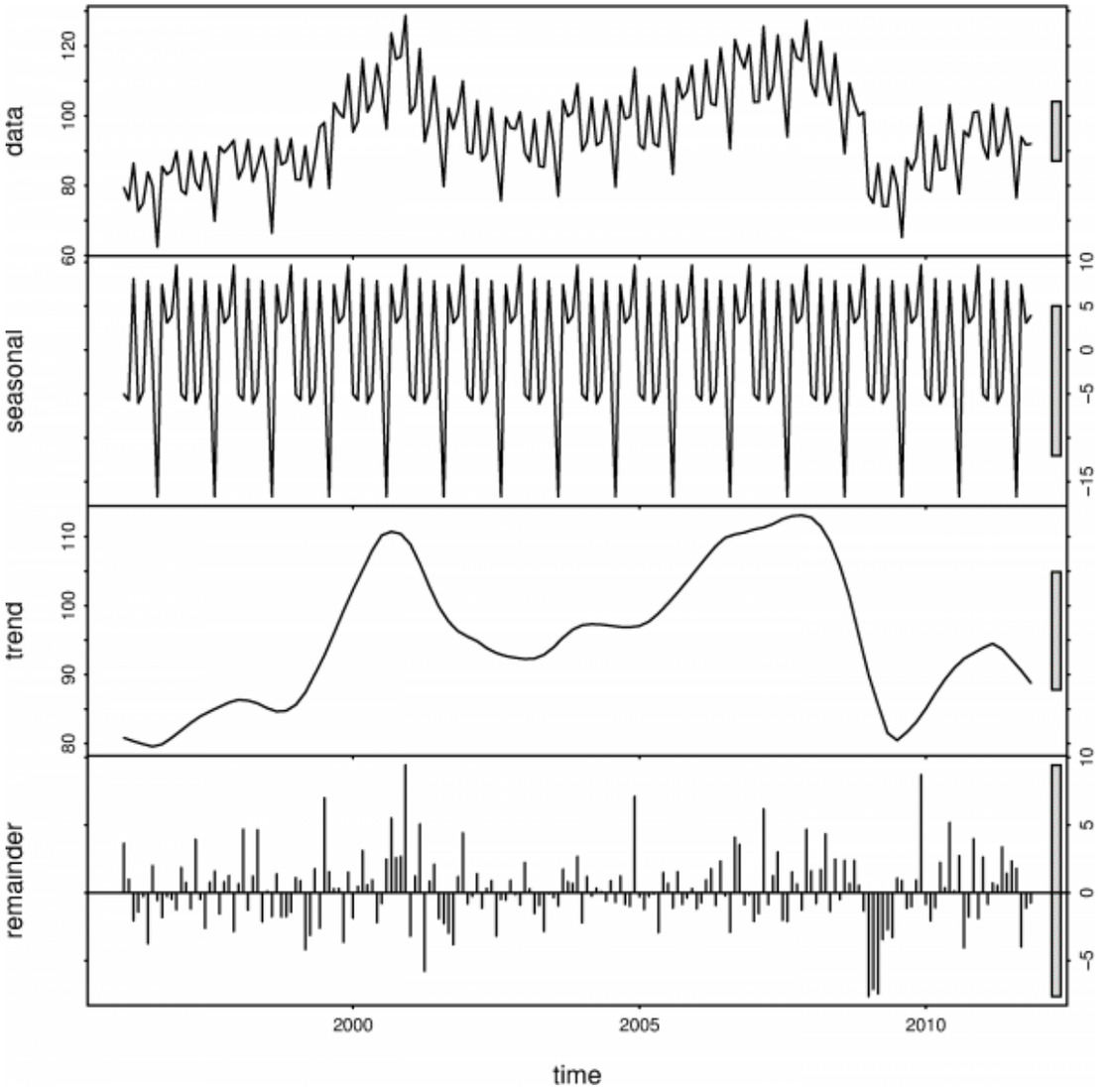
- 2. When modeling, how could I try to account for this dependence?

DECOMPOSITION

- We will often use models to decompose spatiotemporal data into different components.

- Decomposing, in a modeling context, means writing out a model that isolates each component.

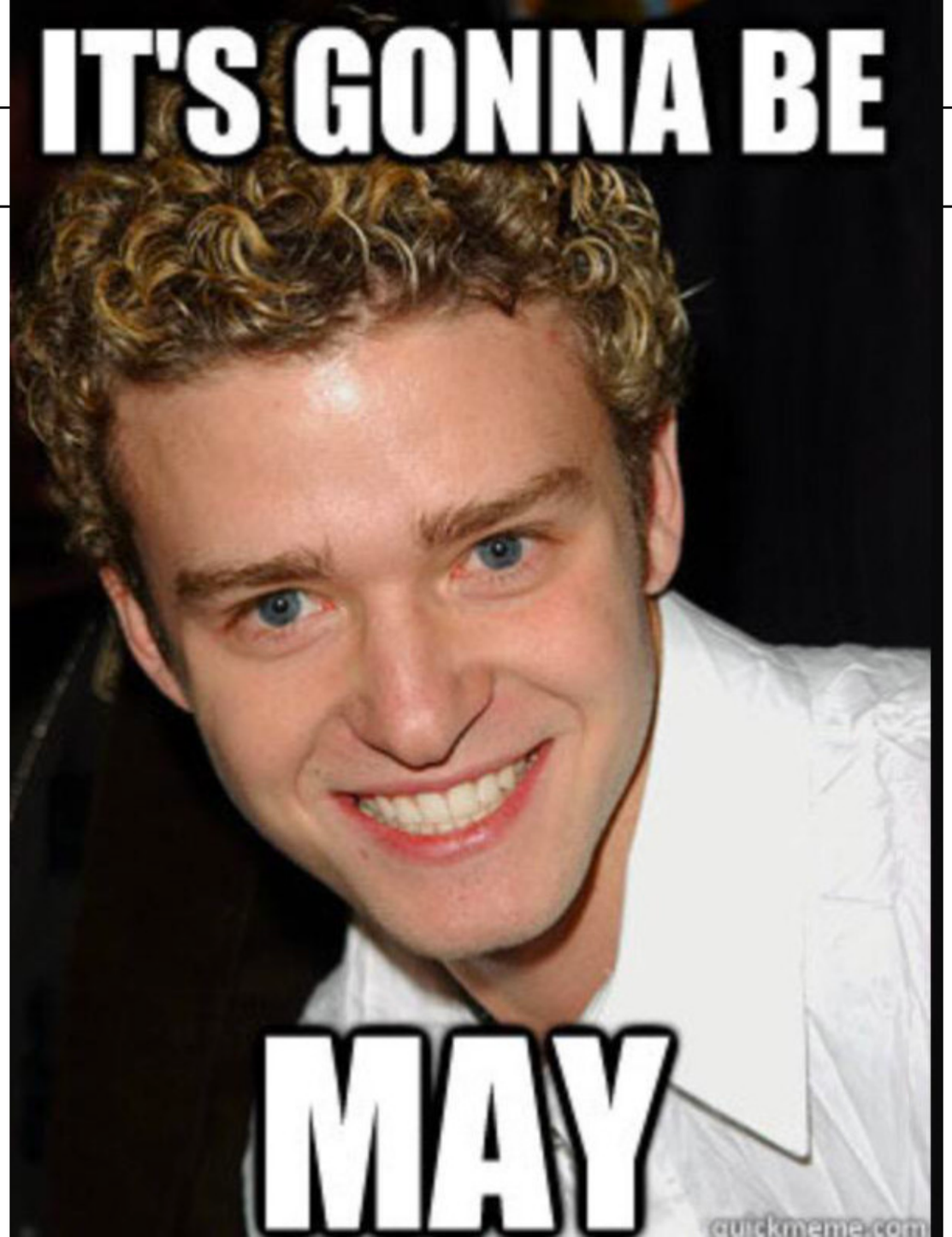
- $data = f(season) + g(trend) + h(noise)$



ADDITIONAL ASSUMPTIONS

- Making additional assumptions can make our analysis easier… but we have to weigh whether or not these assumptions seem realistic.

- **Stationarity** means that the relationship (covariance!) between observation a and observation b depends only on the distance between them.
- Spatial: If we're attempting to predict temperature and believe stationarity to hold true, then we'd say that the relationship between temperatures in D.C. and Baltimore is the same as the relationship between temperatures in Cincinnati and Dayton.
- Temporal: If we're attempting to predict temperature and believe stationarity to hold true, then we'd say that the relationship between temperature on January 1 and February 1 is the same as the relationship between temperature on April 1 and May 1.


IT'S GONNA BE
MAY
quickmeme.com