# Data Science for Economists

## Lecture 16: Intro to Data Science

Drew Van Kuiken
University of North Carolina | ECON 370

# Table of contents

# Introduction

# Motivation

We've spent most the class learning R and becoming competent in programming.

Today, we'll do a quick scan of the data science world and the things that R can be used for in the workplace.

Remember that I do not expect you to entirely understand everything we talk about.

- I want you to get a taste so that when you really learn this stuff later, it is easier.

# What Is Data Science?

# What is data science?

Data science is a relatively new "field" that is still evolving.

Wikipedia's definition:

> Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from noisy, structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains. Data science is related to data mining, machine learning and big data.

Another definition:

> Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician. - Josh Wills

Anyone can be a data scientist!

- Economists have a special toolkit that is more important for data science than ever.

# What do data scientists do?

# What do data scientists do?

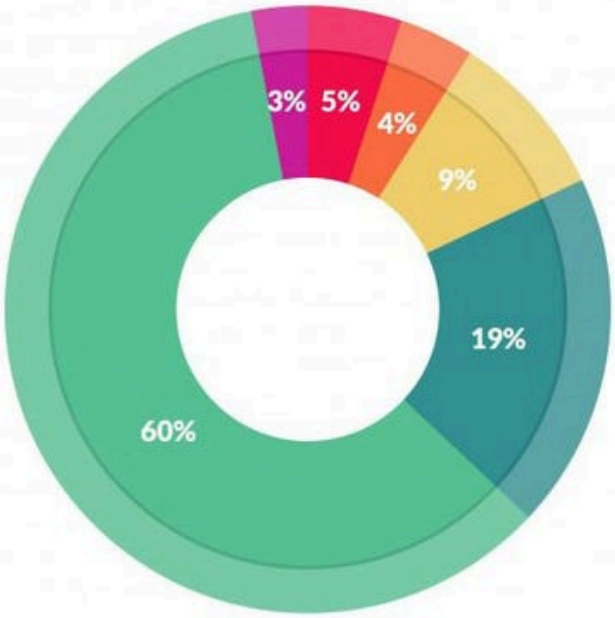
1. Data mining: extracting, wrangling, and storing large amounts of data.
2. Modeling: applying models and ideas from both statistical/machine learning and traditonal statistics to build algorithms to do things too difficult for humans.
3. Software/website development: some data scientists will take the data, algorithms, and insights they develop and integrated them into software or websites.

There are lots of buzzwords in this area. It is important to see through this and not get intimidated.

We've already spent a bit of time talking about data wrangling.

Let's talking about the modeling side.

# What do data scientist do?




**What data scientists spend the most time doing**

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

# Modeling

# Statistics vs Machine Learning

One of the most confusing thing more many people is understanding the difference between machine learning and traditional statistics.

In truth, there isn't much of a difference and there's *lots* of overlap between the two.

- E.g. Ridge and LASSO regression are used in both and many statistical learning algorithms are just "flexible estimation" (nonparametric estimation) in statistics.

The big difference comes down to philosophical differences and objects of interest.

Suppose you have a traditional linear model:

$$\boldsymbol{y}_i = \boldsymbol{x}_i\beta + \varepsilon_i$$

- Statisticians will care about getting an estimate for $\beta$, called $\hat{\beta}$
  - Ideally, $\hat{\beta}$ will have desirable properties.
- Machine learning cares about getting accurate and "precise" estimates of $\boldsymbol{y}_i$, called $\hat{\boldsymbol{y}}_i$

This difference comes down to inference of $\hat{\beta}$ versus prediction of $\hat{\boldsymbol{y}}_i$.

# Inference vs Prediction

Since we do not know the true value of $\beta$ we want to estimate it using an "estimator."

- An estimator is a function $\hat{\beta}$ that takes data $\boldsymbol{Z}$ (for our example, $\boldsymbol{Z} = (y_i, x_i)_{i=1}^N$) as an input and returns an estimate of $\beta$.
- For a given sample $\boldsymbol{Z}$, $\hat{\beta}(\boldsymbol{Z})$ is an estimate for $\beta$. Will just write $\hat{\beta}$ from now on, but remember that estimators are functions of your *data*.

One approach to estimation is to choose an estimator that minimizes the "mean-squared error" (MSE)

- The MSE is the average squared difference between the estimate $\hat{\beta}$ and the true value $\beta$

$$\mathrm{MSE}_\beta(\hat{\beta}) = \boldsymbol{E}[(\hat{\beta} - \beta)^2]$$

However, different estimators can minimize the MSE but have different properties. Using algebra, we can write the MSE like so:

$$\mathrm{MSE}_\beta(\hat{\beta}) = \boldsymbol{Bias}(\hat{\beta}, \beta)^2 + \boldsymbol{Var}(\hat{\beta})$$

This decomposition of the MSE encapsulates the different goals of traditional statistics (inference) and machine learning (prediction).

# Bias-Variance Trade-Off

Suppose we have two estimators $\hat{\beta}$ and $\tilde{\beta}$ that result in *the same* $\mathbf{MSE}_\beta = \bar{m}$

- For estimator $x$, let $\varepsilon(x) = \boldsymbol{Bias}(x, \beta) = \boldsymbol{E}[x] - \beta$ and $\Sigma(x) = \boldsymbol{Var}(x)$.

Since $\varepsilon(\hat{\beta})^2 + \Sigma(\hat{\beta}) = \bar{m}$ and $\varepsilon(\tilde{\beta})^2 + \Sigma(\tilde{\beta}) = \bar{m}$, $\varepsilon(\hat{\beta})^2 + \Sigma(\hat{\beta}) = \varepsilon(\tilde{\beta})^2 + \Sigma(\tilde{\beta})$ or

$$\varepsilon(\hat{\beta})^2 - \varepsilon(\tilde{\beta})^2 = \Sigma(\tilde{\beta}) - \Sigma(\hat{\beta}) \iff (\varepsilon(\hat{\beta}) - \varepsilon(\tilde{\beta}))(\varepsilon(\hat{\beta}) + \varepsilon(\tilde{\beta})) = \Sigma(\tilde{\beta}) - \Sigma(\hat{\beta})$$

1. If $\Sigma(\tilde{\beta}) = \Sigma(\hat{\beta})$, it must be $|\varepsilon(\hat{\beta})| = |\varepsilon(\tilde{\beta})|$.
2. If $\Sigma(\tilde{\beta}) > \Sigma(\hat{\beta})$, it must be $|\varepsilon(\hat{\beta})| > |\varepsilon(\hat{\beta})|$.

**Conclusion**: If one estimator has a lower variance than another and they both have the same MSE, it must be the one with lower variance has a larger bias.

1. For a fixed level of MSE $m(x) = \bar{m}$, $\frac{d\,\Sigma(x)}{d\,\varepsilon(x)} = -2\varepsilon(x)$.
2. $\frac{dm(x)}{d\varepsilon(x)} = 2\varepsilon(x)$ and $\frac{dm(x)}{d\Sigma(x)} = 1$, so $\left|\frac{dm(x)}{d\varepsilon(x)}\right| < \left|\frac{dm(x)}{d\Sigma(x)}\right| \iff |\varepsilon(x)| < \frac{1}{2}$

**Conclusion**: An increase in the bias is substituted with a larger decrease in the variance. As well, accepting some bias increases MSE less than increasing the variance as long as the bias is "small."

# Inference vs Prediction (cont.)

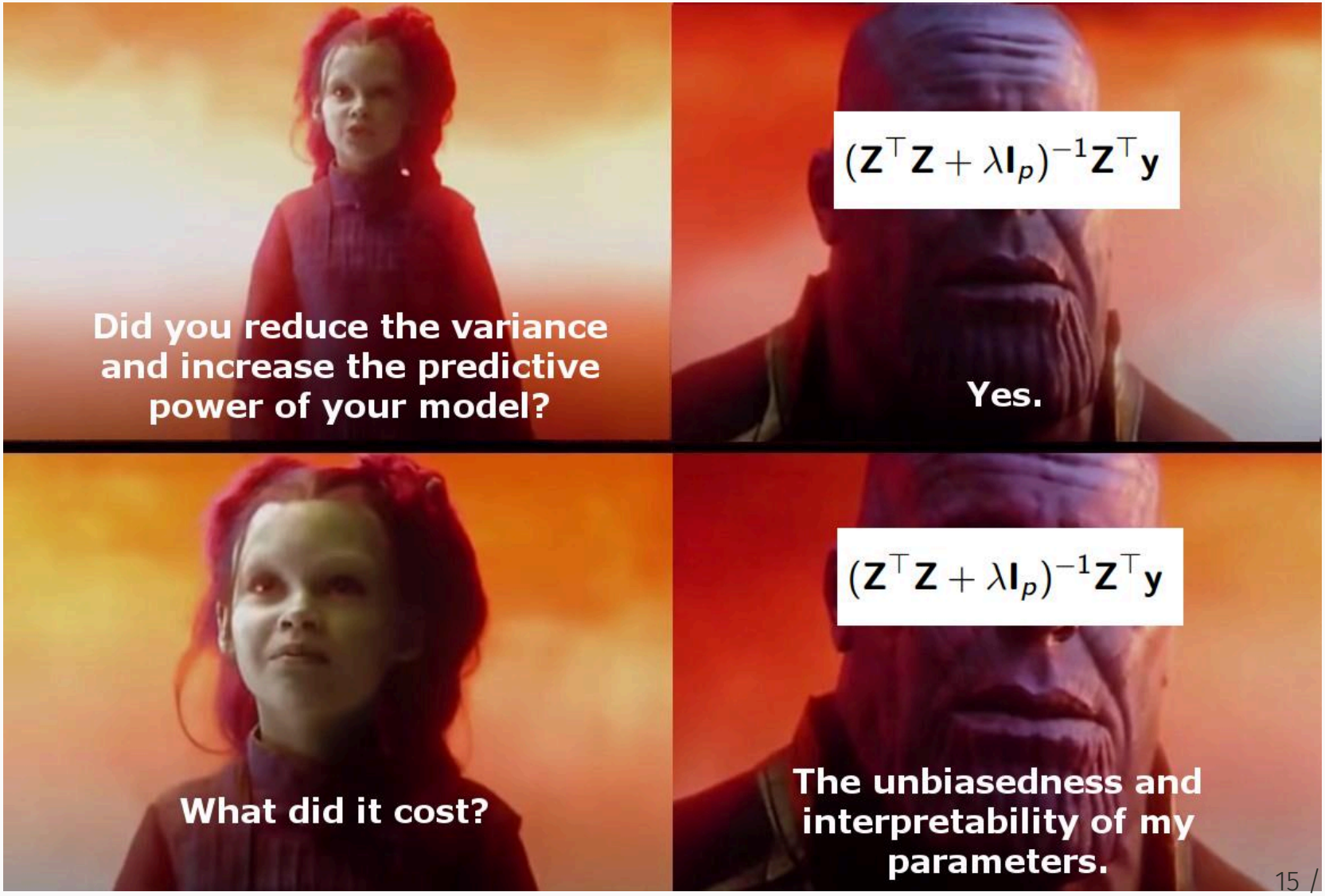
Traditional statistics $\rightarrow$ inference of $\beta$

- We want to try to infer the value of $\beta$ from the data.
- Choose an estimator $\hat{\beta}$ that has low variance and low bias to perform "powerful" statistical tests and make conclusions about the value of $\beta$.
- Usually (though not always) this requires us to choose an unbiased estimator so we aren't "skewing" the estimates $\hat{\beta}$ too much.

Machine Learning $\rightarrow$ predicting $y$

- Estimate some model given "training data" and then predict the value of $y$ given a new value of $x$.
- Call this prediction $\hat{y}$. In our example, $\hat{y} = x\hat{\beta}$
- Usually want the variance of the prediction to be small (noisy predictions are useless).
- In our example, given a new $x$, we have $Var(\hat{y}) = Var(x\hat{\beta}) = x^2 Var(\hat{\beta})$.
- To decrease the variance of $\hat{y}$, we must decrease the variance of $\hat{\beta}$.
- Leads to choosing estimators with some bias if it decreases the variance enough.

# Inference vs Prediction (cont.)



# Inference vs Prediction (cont.)

Prediction Problems:

1. How much will this ad campaign increase revenue?
2. What will traffic on the website be tomorrow?
3. Is this tweet harmful and should it be flagged?
4. What will the price of a home be in a year?

Inference Problems:

1. Will this ad campaign have a causal impact on revenue?
2. Does education have a causal effect on wages?
3. Will changing the Twitter UI cause people to spend more time on the site?
4. What is the causal effect of renovating a kitchen on a home price?

For more on the difference, check out this blog post by r y x, r

# Types of Learning

- Supervised Learning:
  - You have a target variable ("dependent variable") and you would like to learn from function of the features ("independent variables") that explains the target.
- Unsupervised Learning:
  - We observe $X$, but not $y$. While we can't use traditional statistical models, we can still do things like "clustering," classify observations of $X$ based on similarity.

There are two main types of supervised learning:

1. Regression: the target variable, $y$, is continuous and you want to learn a function $f$ of the features $X$ where $y = f(X) + \varepsilon$ where $\varepsilon$ is some error term.

- If $\hat{f}$ is your estimate of $f$ (or "learned function"), then the prediction of $y$, called $\hat{y}$ is $\hat{y} = \hat{f}(X)$

2. Classification: the target variable $y$ is a category (e.g. $y =$ freshman, sophomore, junior, senior) and you want to learn $P(Y = y|X)$ so if you're given a new observation of $X$, you can predict which group it belongs to.

# Types of Learning (cont.)

Below are some examples of each type you may have heard of.

- Regression*: linear regression, Ridge regression, LASSO regression.
- Classification: logistic regression, K-nearest neighbors.
- Unsupervised learning: K-means.

* Don't confuse regression in the learning sense with regression in the statistical sense. While they are similar and have the same name, they are different. When we say linear regression, we are referring to estimating a condition mean $E[y|X]$ with a linear model. Regression in machine learning is any model where $y$ is continuous regardless of what's being estimated.

# What do economists have to bring?

# What do economists have to bring?

As economists, we bring a lot of unique tools to the world of data science.

Our specialty is being able to think carefully about observational data to obtain causal effects or "counterfactuals."

This is where the idea of the "data generating process" comes in handy. How did the data come to us? What economic choices affect how we observe the data?

The Nobel (Memorial) Prize in Economics was award to Angrist, Imbens, and Card in part for their work on developing ideas an econometric framework to think about causal questions seriously.

In fact, Twitter Engineering (@TwitterEng) created a tweet thread about how Angrist's and Imbens's work have influenced their work at Twitter that is worth a read.

Econometricians have contributed to the field of machine learning; see Athey et. al (2019), Athey and Imbens (2016), Nekipelov et. al. (2021), just to name a few.

# What do economists have to bring?

Many of the issues we face in economics are the same issues data scientists are finding cause problems with their models.

- If you train an algorithm for resumes on only white men, what do you think the algorithm will do when it gets a resume from someone not white or male?
- The types of people who select into using Twitter are likely different than those who don't use Twitter. How/when should Twitter keep this in mind when training their algorithms?
- How does racial bias in incarceration rates affect algorithms used to recommend probation or sentencing? (Yes, these exist.)
- How does endogeneity in credit/financial history affect the credit scores assigned to individuals?