# Welcome to Unifying Data Science!

Nick Eubank

# Three Goals of the Course

By the end of this course, you will:

# Three Goals of the Course

By the end of this course, you will:

1. Understanding how different approaches to data science relate to one another, and know when to employ different toolsets.
   *The "Unifying" in Unifying Data Science*

# Three Goals of the Course

By the end of this course, you will:

1. Understanding how different approaches to data science relate to one another, and know when to employ different toolsets.
   *The "Unifying" in Unifying Data Science*
2. Be able to critically evaluate causal claims, and develop research designs to answer causal questions.
   *Causal Inference*

# Three Goals of the Course

By the end of this course, you will:

1. Understanding how different approaches to data science relate to one another, and know when to employ different toolsets.
   *The "Unifying" in Unifying Data Science*
2. Be able to critically evaluate causal claims, and develop research designs to answer causal questions.
   *Causal Inference*
3. Execute a data science project from conception to delivery
   *Complete with step-by-step models*

# Part One: Unifying Data Science

# How did Data Science become a thing?

# How did Data Science become a thing?

- Academic research is organized into silos:

# How did Data Science become a thing?

- Academic research is organized into silos:
    - Computer Science
    - Statistics
    - Economics
    - Political Science
    - Engineering

# How did Data Science become a thing?

- Academic research is organized into silos:
  - Computer Science
  - Statistics
  - Economics
  - Political Science
  - Engineering

$\Rightarrow$ Development of new tools occurred *within* each silo.

# Where are we today?

Very little cross-pollination across silos

## Where are we today?

Very little cross-pollination across silos

- Lots of duplication of development.

# Where are we today?

Very little cross-pollination across silos

- Lots of duplication of development.
- Every silo has its own vocabulary.

# Where are we today?

Very little cross-pollination across silos

- Lots of duplication of development.
- Every silo has its own vocabulary.
- Each silo has focused on the aspects most relevant to their applications. e.g.:

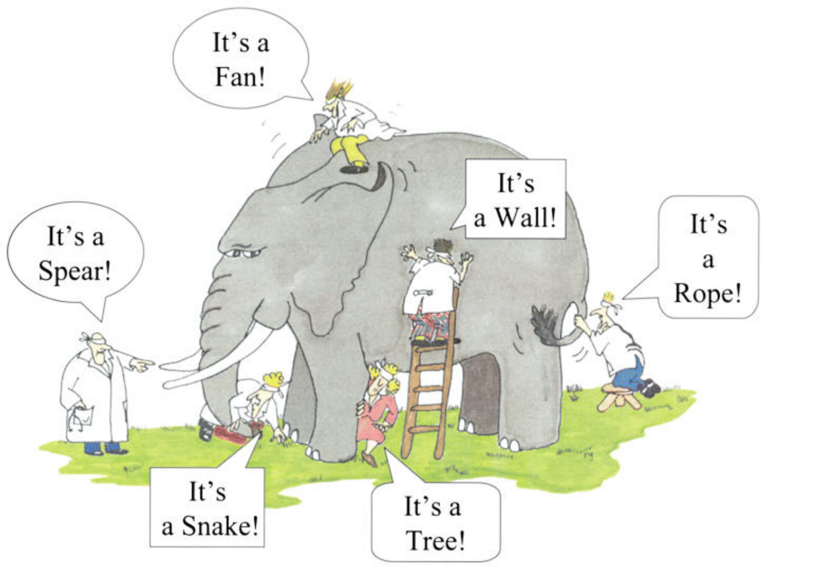
# Where are we today?

Very little cross-pollination across silos

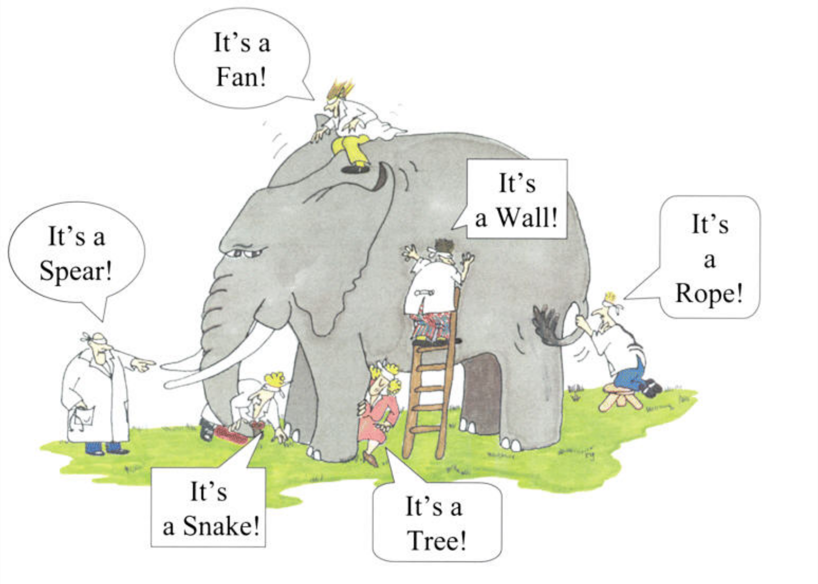
- Lots of duplication of development.
- Every silo has its own vocabulary.
- Each silo has focused on the aspects most relevant to their applications. e.g.:
    - CS likes to classify things and make predictions, don't care how model works
    - Social scientists like to make causal statements, don't care about predictive power

# Blind Men and the Elephant

# Blind Men and the Elephant



# Blind Men and the Elephant




⇒ This is where data science is *now*.

# What do I think Data Science should be?

# What do I think Data Science should be?

An effort to unify the development of quantitative methods

# What do I think Data Science should be?

An effort to unify the development of quantitative methods
$\rightarrow$ Recognize the elephant

# This Class

Discipline of learning how best to answer questions using quantitative data.

# This Class

1. Introduce a taxonomy of questions

   Exploratory, passive-predictive, causal

# This Class

1. Introduce a taxonomy of questions
   Exploratory, passive-predictive, causal
2. For each class of questions, we will discuss:
   - Intrinsic challenges to answering each class of questions
   - What tools are best suited to each type of question

# This Class

1. Introduce a taxonomy of questions
   Exploratory, passive-predictive, causal
2. For each class of questions, we will discuss:
   - Intrinsic challenges to answering each class of questions
   - What tools are best suited to each type of question

By the end of the course, you should know when to reach for...

# This Class

1. Introduce a taxonomy of questions
   Exploratory, passive-predictive, causal
2. For each class of questions, we will discuss:
   - Intrinsic challenges to answering each class of questions
   - What tools are best suited to each type of question

By the end of the course, you should know when to reach for...

- Unsupervised machine learning
- Supervised machine learning
- Range of causal inference techniques
- Other approaches to exploratory analysis

# This Class

The tool you use should be dictated by the question you seek to answer

# Who Are We?

I am a empirical / computational social scientist

# Who Are We?

I am a empirical / computational social scientist

- PhD in Political Economy, Masters in Economics, BA in Economics and International Relations
- Research on criminal justice, policing, social networks, election administration, gerrymandering, and (in days gone by) international development.

## Who Are We?

I am a empirical / computational social scientist

- PhD in Political Economy, Masters in Economics, BA in Economics and International Relations
- Research on criminal justice, policing, social networks, election administration, gerrymandering, and (in days gone by) international development.

(But I have a pretty strong CS background for a social scientist.)

# Who Are We?

I am a empirical / computational social scientist

- PhD in Political Economy, Masters in Economics, BA in Economics and International Relations
- Research on criminal justice, policing, social networks, election administration, gerrymandering, and (in days gone by) international development.

(But I have a pretty strong CS background for a social scientist.)

Erika Fox & Clarissa Ache Cabello (TA)

# Who Are We?

I am a empirical / computational social scientist

- PhD in Political Economy, Masters in Economics, BA in Economics and International Relations
- Research on criminal justice, policing, social networks, election administration, gerrymandering, and (in days gone by) international development.

(But I have a pretty strong CS background for a social scientist.)

Erika Fox & Clarissa Ache Cabello (TA)

- MIDS Second Year Students
- *Extremely* good at this

# Who Are We?

I am a empirical / computational social scientist

- PhD in Political Economy, Masters in Economics, BA in Economics and International Relations
- Research on criminal justice, policing, social networks, election administration, gerrymandering, and (in days gone by) international development.

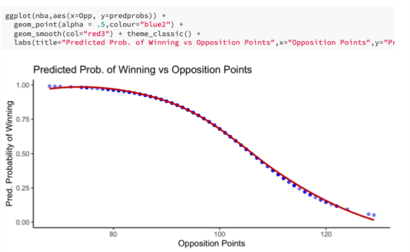
(But I have a pretty strong CS background for a social scientist.)

Erika Fox & Clarissa Ache Cabello (TA)

- MIDS Second Year Students
- *Extremely* good at this
- Causal inference is a discipline that people spend their careers studying, so they are terrific resources, but also be aware you may hit questions they redirect to me.

# Part Two: Causal Inference

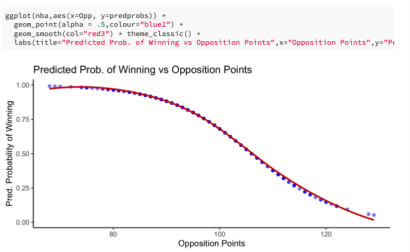
# Modeling and Representation of Data




You learned a *lot*:

- Model selection
- Interpreting BIC, AIC, AUC, R-squared
- Residual Plots

# Modeling and Representation of Data




You learned a *lot*:

- Model selection
- Interpreting BIC, AIC, AUC, R-squared
- Residual Plots

$\Rightarrow$ Develop model to faithfully represent patterns in the data

# Causal Inference

We're focused on what comes next.

# Causal Inference

We're focused on what comes next.

*Assume* our model faithfully represents the data.

# Causal Inference

We're focused on what comes next.

*Assume* our model faithfully represents the data.

$\Rightarrow$ Given those models, what can we conclude **about the world**?

# Causal Inference

We're focused on what comes next.

*Assume* our model faithfully represents the data.

$\Rightarrow$ Given those models, what can we conclude **about the world**?

Suppose we find a correlation between car advertising and consumer spending across neighborhoods in North Carolina.

# Causal Inference

We're focused on what comes next.

*Assume* our model faithfully represents the data.

$\Rightarrow$ Given those models, what can we conclude **about the world**?

Suppose we find a correlation between car advertising and consumer spending across neighborhoods in North Carolina.

- Does that imply that more advertising would increase spending further?

## Causal Inference

We're focused on what comes next.

*Assume* our model faithfully represents the data.

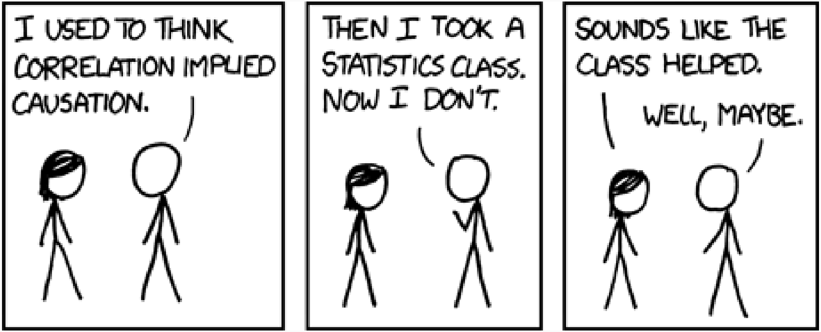
$\Rightarrow$ Given those models, what can we conclude **about the world**?

Suppose we find a correlation between car advertising and consumer spending across neighborhoods in North Carolina.

- Does that imply that more advertising would increase spending further?
  In other words, based on this model, do we think advertising is *causing* more consumer spending?

# Causal Inference

Correlation does not imply causation


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

# Causal Inference

Does advertising cause increased consumer spending?

# Causal Inference

Does advertising cause increased consumer spending?

- "Well, correlation does not imply causation, so I can't say."

## Causal Inference

Does advertising cause increased consumer spending?

- "Well, correlation does not imply causation, so I can't say."
- "Well, correlation does not imply causation, *but* yea probably."

# Causal Inference

Correlation does not *necessary* imply causation,

# Causal Inference

Correlation does not *necessary* imply causation,

- but when certain assumptions are met, correlation *does* imply causation.

## Causal Inference

Correlation does not *necessary* imply causation,

- but when certain assumptions are met, correlation *does* imply causation.

By learning the *assumptions* that are required for a correlation to be a good estimate of a causal effect, you can:

# Causal Inference

Correlation does not *necessary* imply causation,

- but when certain assumptions are met, correlation *does* imply causation.

By learning the *assumptions* that are required for a correlation to be a good estimate of a causal effect, you can:

- Evaluate whether those assumptions are likely to be met,

## Causal Inference

Correlation does not *necessary* imply causation,

- but when certain assumptions are met, correlation *does* imply causation.

By learning the *assumptions* that are required for a correlation to be a good estimate of a causal effect, you can:

- Evaluate whether those assumptions are likely to be met,
- Come up with different research designs whose assumptions *would* be met.

# Causal Inference

**Modeling**

Developing Model to Faithfully Represent Data

$$\Downarrow$$

**Inference**

Interpreting Model Parameters

**Modeling**

Developing Model to Faithfully Represent Data

$\Downarrow$

**Inference**

Interpreting Model Parameters

# Causal Inference

While we will use a "statistical framework" (Potential Outcomes Framework) to help us be rigorous in our thinking...

# Causal Inference

While we will use a "statistical framework" (Potential Outcomes Framework) to help us be rigorous in our thinking...

There are **NO** statistical tests that will tell you if your model is estimating a true causal effect.

# Causal Inference

While we will use a “statistical framework” (Potential Outcomes Framework) to help us be rigorous in our thinking...

There are **NO** statistical tests that will tell you if your model is estimating a true causal effect.

- *Fundamental Problem of Causal Inference*

# Causal Inference

While we will use a "statistical framework" (Potential Outcomes Framework) to help us be rigorous in our thinking...

There are **NO** statistical tests that will tell you if your model is estimating a true causal effect.

- *Fundamental Problem of Causal Inference*

Causal inference is unavoidably about:

- Critical thinking
- Case knowledge

# Causal Inference

After introducting Potential Outcomes, we'll explore a range of causal research techniques:

# Causal Inference

After introducting Potential Outcomes, we'll explore a range of causal research techniques:

- Experiments (Randomized-Control Trials, or RCTs)

## Causal Inference

After introducting Potential Outcomes, we'll explore a range of causal research techniques:

- Experiments (Randomized-Control Trials, or RCTs)
- Linear Regressions as Causal Tools

# Causal Inference

After introducting Potential Outcomes, we'll explore a range of causal research techniques:

- Experiments (Randomized-Control Trials, or RCTs)
- Linear Regressions as Causal Tools
- Matching

# Causal Inference

After introducting Potential Outcomes, we'll explore a range of causal research techniques:

- Experiments (Randomized-Control Trials, or RCTs)
- Linear Regressions as Causal Tools
- Matching
- Differences in Differences

## Causal Inference

After introducting Potential Outcomes, we'll explore a range of causal research techniques:

- Experiments (Randomized-Control Trials, or RCTs)
- Linear Regressions as Causal Tools
- Matching
- Differences in Differences
- Natural Experiments

## Causal Inference

After introducting Potential Outcomes, we'll explore a range of causal research techniques:

- Experiments (Randomized-Control Trials, or RCTs)
- Linear Regressions as Causal Tools
- Matching
- Differences in Differences
- Natural Experiments

$\Rightarrow$ Each of these designs will provide causal estimates if certain assumptions are met,

# Causal Inference

After introducting Potential Outcomes, we'll explore a range of causal research techniques:

- Experiments (Randomized-Control Trials, or RCTs)
- Linear Regressions as Causal Tools
- Matching
- Differences in Differences
- Natural Experiments

⇒ Each of these designs will provide causal estimates if certain assumptions are met,

- But it will always be is up to you, the researcher, to evaluate whether those assumptions are reasonable!

# Causal Inference

By the end of this course, you will:

- Understand why causal inference is hard,

# Causal Inference

By the end of this course, you will:

- Understand why causal inference is hard,
- Be able to critically evaluate causal evidence collected by others,

# Causal Inference

By the end of this course, you will:

- Understand why causal inference is hard,
- Be able to critically evaluate causal evidence collected by others,
- Articulate causal questions,

# Causal Inference

By the end of this course, you will:

- Understand why causal inference is hard,
- Be able to critically evaluate causal evidence collected by others,
- Articulate causal questions,
- And develop research designs to answer those questions.

# Part Three: Your Data Science Project

# Data Science Project

Over semester, you will also develop a data science project from start-to-finish

- Teams of 3-4,
- On topic of your own choosing.
- Only rule: it has to be causal.

# Data Science Project

Over semester, you will also develop a data science project from start-to-finish

- Teams of 3-4,
- On topic of your own choosing.
- Only rule: it has to be causal.

$\rightarrow$ Nice portfolio piece

# Data Science Project

Over semester, you will also develop a data science project from start-to-finish

- Teams of 3-4,
- On topic of your own choosing.
- Only rule: it has to be causal.

$\rightarrow$ Nice portfolio piece
$\rightarrow$ MIDS first-years: Capstone with training wheels

# Data Science Project

Introducing in stages:

# Data Science Project

Introducing in stages:

- Stakeholder management

# Data Science Project

Introducing in stages:

- Stakeholder management
- Backwards Design

# Data Science Project

Introducing in stages:

- Stakeholder management
- Backwards Design
- Workflow Management

# Data Science Project

Introducing in stages:

- Stakeholder management
- Backwards Design
- Workflow Management
- Presenting to Different Audiences

# Data Science Project

Introducing in stages:

- Stakeholder management
- Backwards Design
- Workflow Management
- Presenting to Different Audiences
- Giving Feedback

## Things to Know

- Course site: `http://www.unifyingdatascience.org`
  Contents subject to change!

# Things to Know

- Course site: `http://www.unifyingdatascience.org`
  Contents subject to change!
- Readings are *incredibly* important.

## Things to Know

- Course site: `http://www.unifyingdatascience.org`
  Contents subject to change!
- Readings are *incredibly* important.
- Reading reflections for every reading.
  Due **7 am morning of class.**

# Things to Know

- Course site: `http://www.unifyingdatascience.org`
  Contents subject to change!
- Readings are *incredibly* important.
- Reading reflections for every reading.
  Due **7 am morning of class.**
- If you don't know git or github, you'll want to learn that early.
  - Data Camp and Practical Data Science tools will be made available
  - Workshops hosted by Library

# If you have issues...

# If you have issues...

- With the course material,

# If you have issues...

- With the course material,
- With the course design,

# If you have issues...

- With the course material,
- With the course design,
- With learning online,

# If you have issues...

- With the course material,
- With the course design,
- With learning online,
- With the isolation associated with COVID-life,

## If you have issues...

- With the course material,
- With the course design,
- With learning online,
- With the isolation associated with COVID-life,
- *Or anything else...*

## If you have issues...

- With the course material,
- With the course design,
- With learning online,
- With the isolation associated with COVID-life,
- *Or anything else...*

**Talk to me!**

Phew. That's it!

Questions?