

Practical Data Science: Wrangling Data and Answering Questions

Nick Eubank

What is Data Science?

What is Data Science?

1. What (in theory) do we think Data Science should be?

What is Data Science?

1. What (**in theory**) do we think Data Science should be?
2. What (**empirically**) is Data Science?

What (in theory) should Data Science be?

What (in theory) should Data Science be?

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

What (in theory) should Data Science be?

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

- Question-first approach

What (in theory) should Data Science be?

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

- Question-first approach
- The tool you use should be dictated by the question you seek to answer

What (empirically) is Data Science?

How did Data Science become a thing?

Over the past several decades:

1. Availability of data ↑
2. Computational power ↑

How did Data Science become a thing?

Over the past several decades:

1. Availability of data ↑
2. Computational power ↑

⇒ Huge proliferation and increase in sophistication of computational methods

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

⇒ 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?

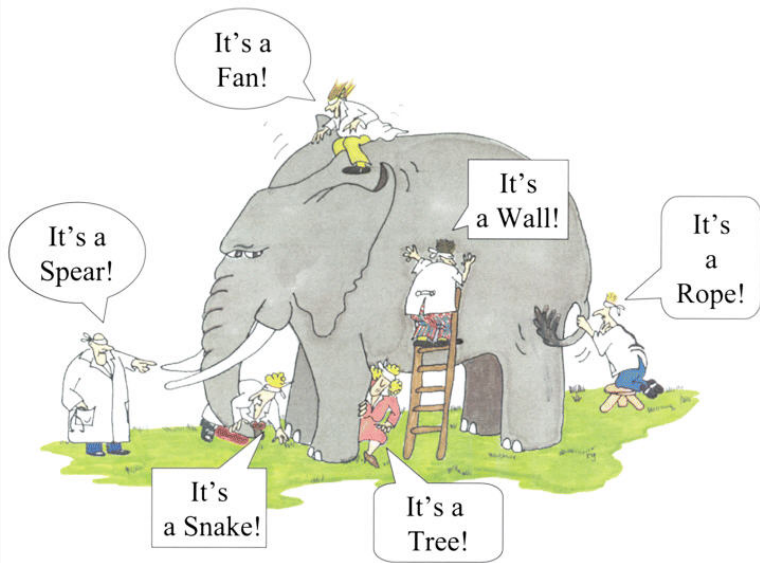
Very little cross-pollination across silos

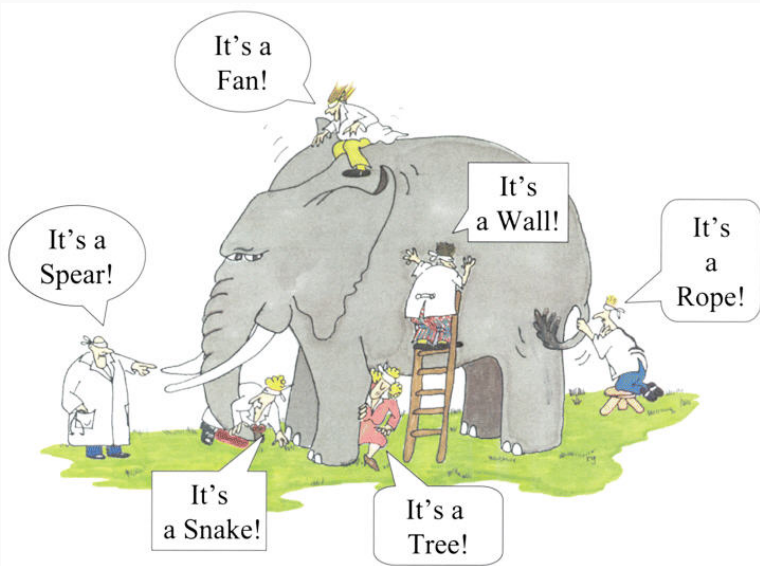
- Lots of duplication of development.
- Every silo has its own vocabulary.
- Each silo has focused on the aspects most relevant to their applications. e.g.:

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





⇒ This is where we are *now*.

What is (empirically) Data Science?

What is (empirically) Data Science?

An effort to unify the development of quantitative methods

What is (empirically) Data Science?

An effort to unify the development of quantitative methods

→ Recognize the elephant

Why does this matter to you?

- Most current researchers learned their skills in a silos.

Why does this matter to you?

- Most current researchers learned their skills in a silos.
In many ways, *you will have better perspective than your professors.*
- Important not just technically, but also when it comes to advice.

Why does this matter to you?

- Most current researchers learned their skills in a silos.
In many ways, *you will have better perspective than your professors.*
- Important not just technically, but also when it comes to advice.
 - Recognize that your professors' conception of "data science" *may not match yours.*

Why does this matter to you?

- Most current researchers learned their skills in a silos.
In many ways, *you will have better perspective than your professors.*
- Important not just technically, but also when it comes to advice.
 - Recognize that your professors' conception of "data science" *may not match yours.*
 - Also just good life advice: scientists are *very unscientific* when it comes to career advice!

Areas of Data Science

Software Engineering DS

Data Analysis DS

Areas of Data Science

Software Engineering DS

- Recommendation engines
- Financial trading algorithms
- Self-driving cars

Data Analysis DS

Areas of Data Science

Software Engineering DS

- Recommendation engines
- Financial trading algorithms
- Self-driving cars

Data Analysis DS

- Impact of policy change
- Effectiveness of health interventions
- Plan political campaigns

Areas of Data Science

Software Engineering DS

- Recommendation engines
- Financial trading algorithms
- Self-driving cars

Data Analysis DS

- Impact of policy change
- Effectiveness of health interventions
- Plan political campaigns

Nearly all data scientists will use some of both sets of skills.

Areas of Data Science

Software Engineering DS

- Recommendation engines
- Financial trading algorithms
- Self-driving cars

Data Analysis DS

- Impact of policy change
- Effectiveness of health interventions
- Plan political campaigns

Nearly all data scientists will use some of both sets of skills.
In this class, we will be focused on the **Data Analysis** flavor of Data Science.

This Class

Part 1: Everything that comes *before* analysis

This Class

Part 1: Everything that comes *before* analysis

- Importing data of lots of forms

This Class

Part 1: Everything that comes *before* analysis

- Importing data of lots of forms
- Cleaning dirty data and developing practices to protect yourself

This Class

Part 1: Everything that comes *before* analysis

- Importing data of lots of forms
- Cleaning dirty data and developing practices to protect yourself
- Merging data from different sources

This Class

Part 1: Everything that comes *before* analysis

- Importing data of lots of forms
- Cleaning dirty data and developing practices to protect yourself
- Merging data from different sources

→ Start to finish wrangling portion of data science project.

This Class

Part 1: Everything that comes *before* analysis

- Importing data of lots of forms
- Cleaning dirty data and developing practices to protect yourself
- Merging data from different sources

→ Start to finish wrangling portion of data science project.

Part 2: Picking tools

This Class

Part 1: Everything that comes *before* analysis

- Importing data of lots of forms
- Cleaning dirty data and developing practices to protect yourself
- Merging data from different sources

→ Start to finish wrangling portion of data science project.

Part 2: Picking tools

- Developing an *answerable* question

This Class

Part 1: Everything that comes *before* analysis

- Importing data of lots of forms
- Cleaning dirty data and developing practices to protect yourself
- Merging data from different sources

→ Start to finish wrangling portion of data science project.

Part 2: Picking tools

- Developing an *answerable* question
- Understanding how your question dictates the best tool to use.

This Class

Part 1: Everything that comes *before* analysis

- Importing data of lots of forms
- Cleaning dirty data and developing practices to protect yourself
- Merging data from different sources

→ Start to finish wrangling portion of data science project.

Part 2: Picking tools

- Developing an *answerable* question
- Understanding how your question dictates the best tool to use.
- Plan your project using backwards design

This Class

Part 1: Everything that comes *before* analysis

- Importing data of lots of forms
- Cleaning dirty data and developing practices to protect yourself
- Merging data from different sources

→ Start to finish wrangling portion of data science project.

Part 2: Picking tools

- Developing an *answerable* question
- Understanding how your question dictates the best tool to use.
- Plan your project using backwards design

→ Develop a research question and outline a full project.

This Class

By the end of this class:

This Class

By the end of this class:

- Find and organize data *on your own*.

This Class

By the end of this class:

- Find and organize data *on your own*.
- Understand how different tools relate to one another.

This Class

By the end of this class:

- Find and organize data *on your own*.
- Understand how different tools relate to one another.
- Know how to approach organizing a full project.

But... What about Python?

But... What about Python?

Don't worry. We'll do lots of Python (esp numpy & pandas).

But... What about Python?

Don't worry. We'll do lots of Python (esp numpy & pandas).
But many of the take-aways aren't Python specific.

But... What about Python?

Don't worry. We'll do lots of Python (esp numpy & pandas).
But many of the take-aways aren't Python specific.

For students who are struggling with (vanilla) Python:

- The books we will use include Python refreshers,
- All students will have full access to DataCamp if they want it,
- I can provide additional resources.

But... What about Python?

Don't worry. We'll do lots of Python (esp numpy & pandas).
But many of the take-aways aren't Python specific.

For students who are struggling with (vanilla) Python:

- The books we will use include Python refreshers,
- All students will have full access to DataCamp if they want it,
- I can provide additional resources.

(But this class doesn't pick up where Drew left off)

But... What about Python?

Don't worry. We'll do lots of Python (esp numpy & pandas).
But many of the take-aways aren't Python specific.

For students who are struggling with (vanilla) Python:

- The books we will use include Python refreshers,
- All students will have full access to DataCamp if they want it,
- I can provide additional resources.

(But this class doesn't pick up where Drew left off)

My job is to meet you where you are.

About Me

I am a social scientist

About Me

I am a social scientist

- PhD in Political Economy

About Me

I am a social scientist

- PhD in Political Economy
- Master in Economics

About Me

I am a social scientist

- PhD in Political Economy
- Master in Economics
- BA in Economics and Political Science

- Looking for evidence of polling place manipulation in North Carolina

My Research

- Looking for evidence of polling place manipulation in North Carolina
- Developing methods of measuring Gerrymandering in the US.

My Research

- Looking for evidence of polling place manipulation in North Carolina
- Developing methods of measuring Gerrymandering in the US.
- Testing theories about how social networks shape political behavior using cell-phone meta-data to map social networks of entire countries (Zambia and Venezuela).

My Research

- Looking for evidence of polling place manipulation in North Carolina
- Developing methods of measuring Gerrymandering in the US.
- Testing theories about how social networks shape political behavior using cell-phone meta-data to map social networks of entire countries (Zambia and Venezuela).
- Studying whether political elites in the US South turned to using incarceration to prevent black voters from exercising political influence after the Voting Rights Act removed their ability to use Jim Crow restrictions.

What that means for you

What that means for you

I've worked with *a lot* of messy data in many contexts.

What that means for you

I've worked with *a lot* of messy data in many contexts.

I have unusually broad exposure to data science:

What that means for you

I've worked with *a lot* of messy data in many contexts.

I have unusually broad exposure to data science:

- Within academic circles, I've worked with economists, statisticians, computer scientists, and political scientists.

What that means for you

I've worked with *a lot* of messy data in many contexts.

I have unusually broad exposure to data science:

- Within academic circles, I've worked with economists, statisticians, computer scientists, and political scientists.
- I've done **policy consulting** (World Bank, Center for Global Development), **public outreach** (Op-Ed in *The Guardian*), and **legal consulting** (various voting rights cases).

What that means for you

I've worked with *a lot* of messy data in many contexts.

I have unusually broad exposure to data science:

- Within academic circles, I've worked with economists, statisticians, computer scientists, and political scientists.
- I've done **policy consulting** (World Bank, Center for Global Development), **public outreach** (Op-Ed in *The Guardian*), and **legal consulting** (various voting rights cases).

If this sounds like your “flavor” of data science, I'm happy to talk about career options in this domain.

www.practicaldatascience.org