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 ↑

$\Rightarrow$ 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

$\Rightarrow$ 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?

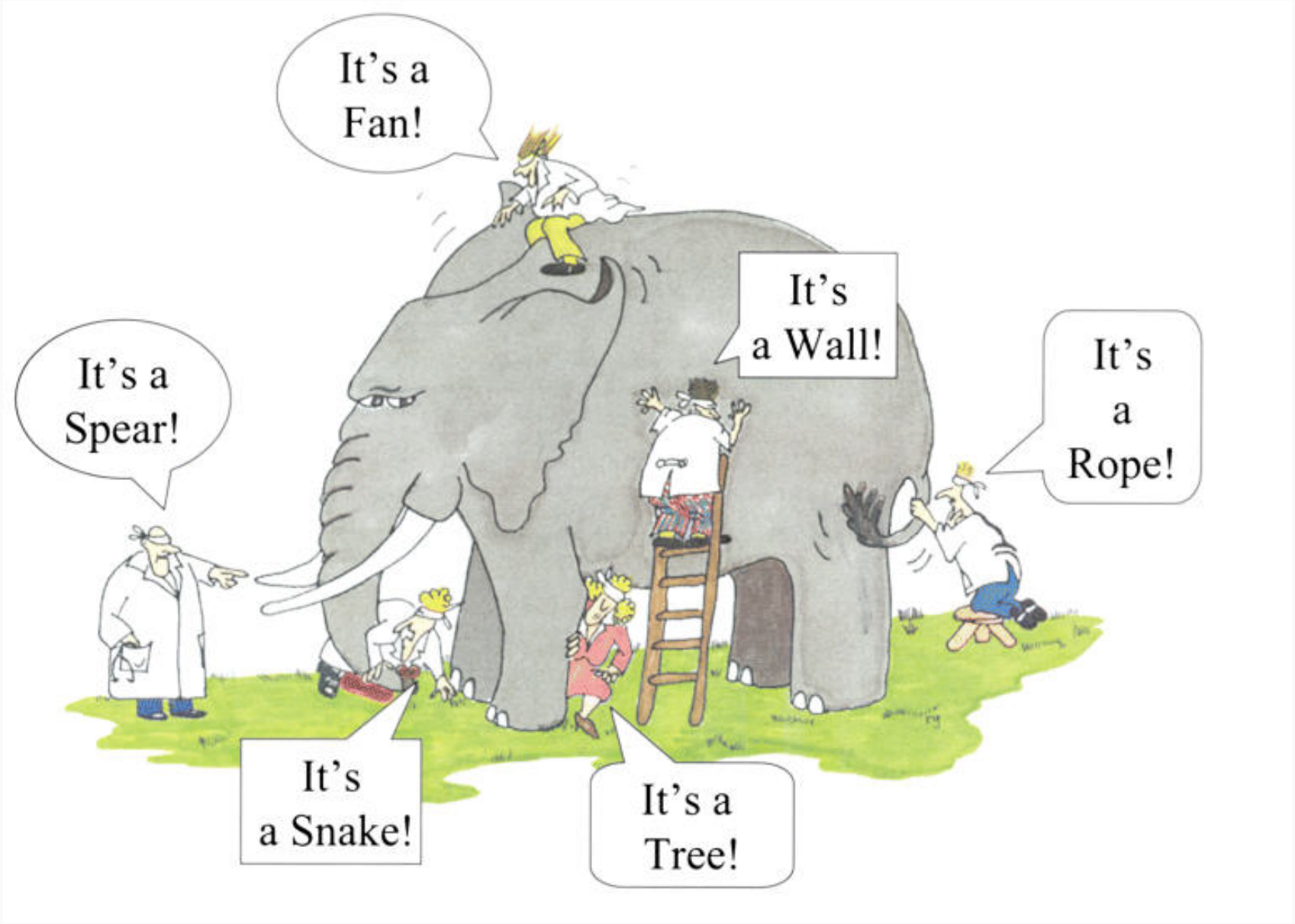
Very little cross-pollination across silos

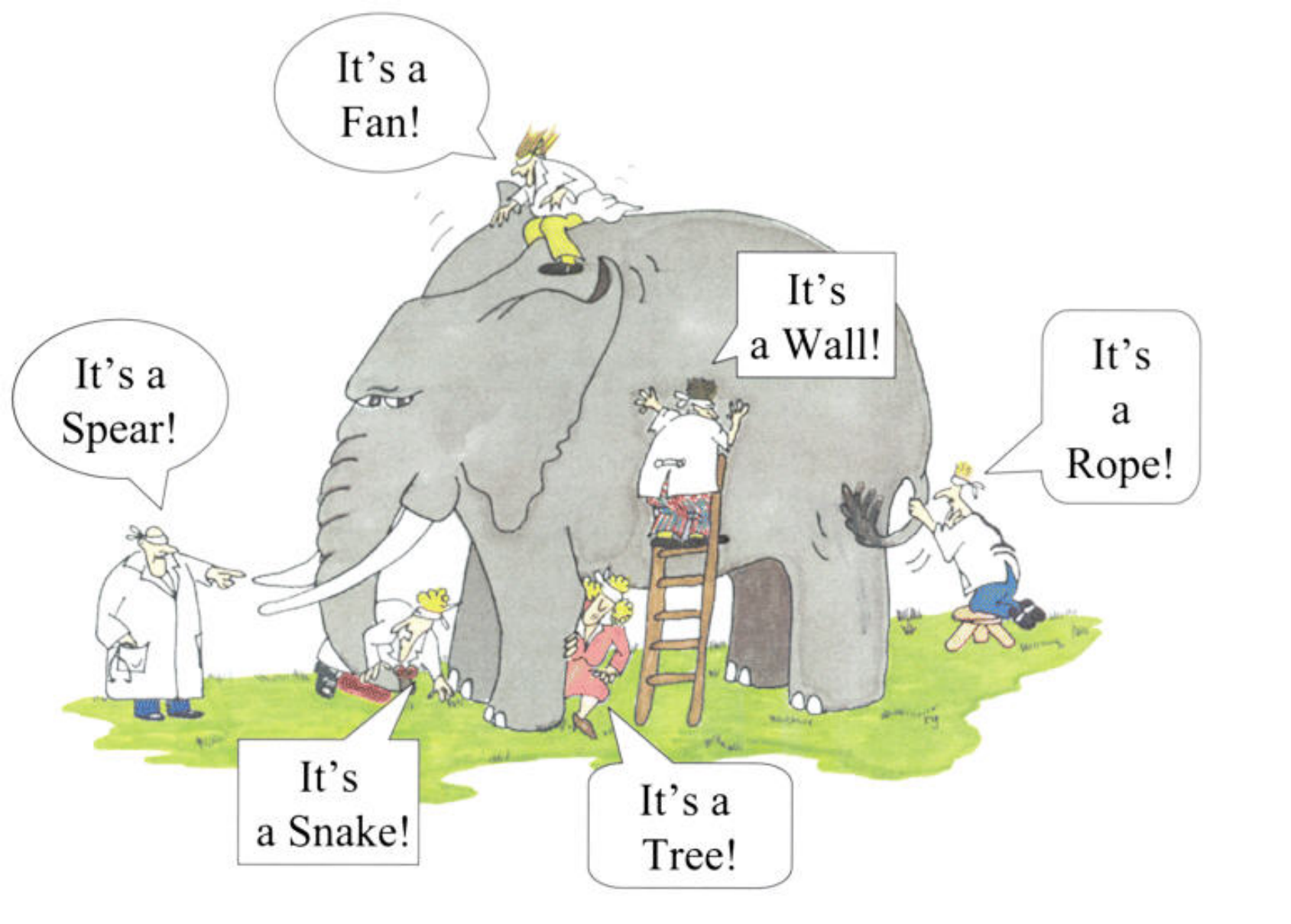
- Lots of duplication of development.
- Every silo has its own vocabulary.
- Each silo has focused on the aspects most relevant to their applications. e.g.:

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


It's a Fan!
It's a Spear!
It's a Wall!
It's a Rope!
It's a Snake!
It's a Tree!



⇒ 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

$\rightarrow$ 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.

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

"The Python Class"

This Class

"The Python Class"

- *Data Science* Python

This Class

"The Python Class"

- *Data Science* Python
    - numpy, pandas, scikit-learn, statsmodels, geopandas

“The Everything-That-Comes-*Before*-Data-Analysis Class”

“The Everything-That-Comes-*Before*-Data-Analysis Class”

80% of a data scientist’s time is spent marshaling data.

“The Everything-That-Comes-*Before*-Data-Analysis Class”

80% of a data scientist’s time is spent marshaling data.

In this course, you will learn:

“The Everything-That-Comes-*Before*-Data-Analysis Class”

80% of a data scientist’s time is spent marshaling data.

In this course, you will learn:

- about different data formats and how to work with them,

“The Everything-That-Comes-*Before*-Data-Analysis Class”

80% of a data scientist’s time is spent marshaling data.
In this course, you will learn:

- about different data formats and how to work with them,
- to work with real, messy, error-ridden data,

“The Everything-That-Comes-*Before*-Data-Analysis Class”

80% of a data scientist’s time is spent marshaling data.
In this course, you will learn:

- about different data formats and how to work with them,
- to work with real, messy, error-ridden data,
- best practices for data science workflow management and collaboration,

“The Everything-That-Comes-*Before*-Data-Analysis Class”

80% of a data scientist’s time is spent marshaling data.
In this course, you will learn:

- about different data formats and how to work with them,
- to work with real, messy, error-ridden data,
- best practices for data science workflow management and collaboration,
- about what makes “big data” big and how to work with it,

“The Everything-That-Comes-*Before*-Data-Analysis Class”

80% of a data scientist’s time is spent marshaling data.
In this course, you will learn:

- about different data formats and how to work with them,
- to work with real, messy, error-ridden data,
- best practices for data science workflow management and collaboration,
- about what makes “big data” big and how to work with it,
- how to work with geospatial data, and

“The Everything-That-Comes-*Before*-Data-Analysis Class”

80% of a data scientist’s time is spent marshaling data.
In this course, you will learn:

- about different data formats and how to work with them,
- to work with real, messy, error-ridden data,
- best practices for data science workflow management and collaboration,
- about what makes “big data” big and how to work with it,
- how to work with geospatial data, and
- how to develop a data science project using backwards design.

“The Everything-That-Comes-*Before*-Data-Analysis Class”

80% of a data scientist’s time is spent marshaling data.
In this course, you will learn:

- about different data formats and how to work with them,
- to work with real, messy, error-ridden data,
- best practices for data science workflow management and collaboration,
- about what makes “big data” big and how to work with it,
- how to work with geospatial data, and
- how to develop a data science project using backwards design.

All through hands-on experience.

This Class

So yes, we will be working *in* Python,

This Class

So yes, we will be working *in* Python,
but this isn't a class *about* Python.

This Class

So yes, we will be working *in* Python,
but this isn't a class *about* Python.
$\Rightarrow$ Emphasis on generalizable data science skills

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 to clean, merge, and manipulate real-world data.

This Class

By the end of this class:

- Find and organize data *on your own*.
- Understand how to clean, merge, and manipulate real-world data.
- Know how to approach organizing a full project.

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

My Research

My Research

- 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