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 $\uparrow$
2. Computational power $\uparrow$

$\Rightarrow$ 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?

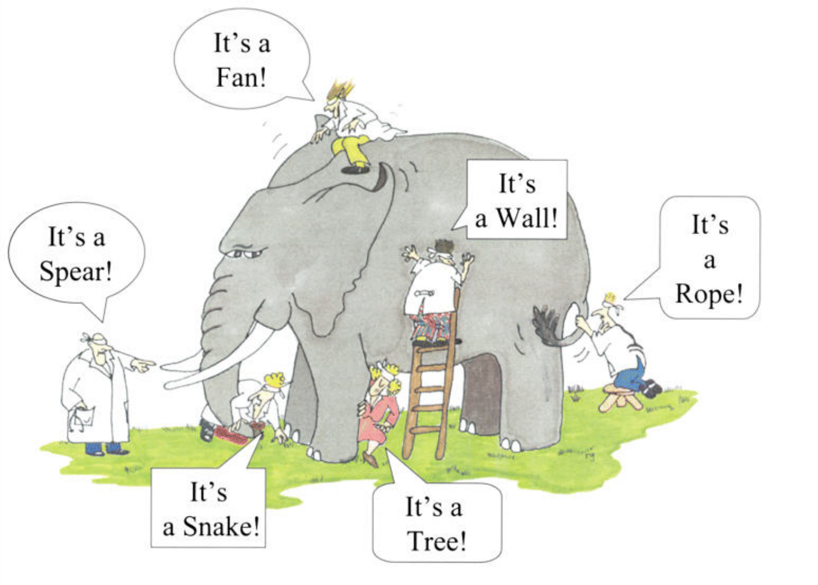
Very little cross-pollination across silos

- 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.:

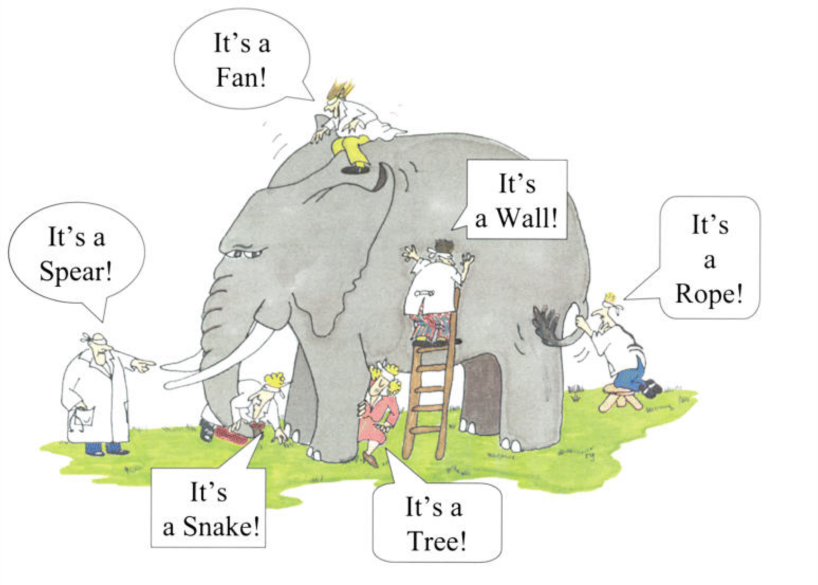
Where are we today?

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
→ 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. Within MIDS, you will do lots of both!

Bootcamp

Definitely emphasizes the Software Engineering side of Data Science.

Bootcamp

Definitely emphasizes the Software Engineering side of Data Science.

- Drew and Genevieve are *extremely* talented software engineers, and will be providing a rigorous foundation for your future programming endeavors.

Bootcamp

Definitely emphasizes the Software Engineering side of Data Science.

- Drew and Genevieve are *extremely* talented software engineers, and will be providing a rigorous foundation for your future programming endeavors.
- Even if you have programmed before, please be open to what they teach.
    - LOTS of industry experience feeds into their recommendations.
    - Great opportunity to break some bad habits.

Bootcamp & LLMs

Bootcamp & LLMs

- It is a *really* bad idea to use LLMs at this stage in your education.

Bootcamp & LLMs

- It is a *really* bad idea to use LLMs at this stage in your education.
- Why? In short:

Bootcamp & LLMs

- It is a *really* bad idea to use LLMs at this stage in your education.
- Why? In short:
    - For real data science applications, LLMs are valuable but error-prone tools and require careful supervision.

Bootcamp & LLMs

- It is a *really* bad idea to use LLMs at this stage in your education.
- Why? In short:
    - For real data science applications, LLMs are valuable but error-prone tools and require careful supervision.
        - Essentially, they are like undergraduate RAs, good at drafting code, but not fully trustworthy.

Bootcamp & LLMs

- It is a *really* bad idea to use LLMs at this stage in your education.
- Why? In short:
    - For real data science applications, LLMs are valuable but error-prone tools and require careful supervision.
        - Essentially, they are like undergraduate RAs, good at drafting code, but not fully trustworthy.
    - But for beginner programming exercises, they are deceptively powerful, and so using them is like learning multiplication using a calculator — it deprives you of the opportunity to develop the intuitive "number sense" required to do advanced work.

Bootcamp & LLMs

- It is a *really* bad idea to use LLMs at this stage in your education.
- Why? In short:
    - For real data science applications, LLMs are valuable but error-prone tools and require careful supervision.
        - Essentially, they are like undergraduate RAs, good at drafting code, but not fully trustworthy.
    - But for beginner programming exercises, they are deceptively powerful, and so using them is like learning multiplication using a calculator — it deprives you of the opportunity to develop the intuitive “number sense” required to do advanced work.

If you use LLMs now, you will never develop the skills needed to supervise them effectively for real projects! So please don’t cheat yourself.

Bootcamp & LLMs

- It is a *really* bad idea to use LLMs at this stage in your education.
- Why? In short:
    - For real data science applications, LLMs are valuable but error-prone tools and require careful supervision.
        - Essentially, they are like undergraduate RAs, good at drafting code, but not fully trustworthy.
    - But for beginner programming exercises, they are deceptively powerful, and so using them is like learning multiplication using a calculator — it deprives you of the opportunity to develop the intuitive “number sense” required to do advanced work.

If you use LLMs now, you will never develop the skills needed to supervise them effectively for real projects! So please don’t cheat yourself.

(More detailed memo to follow on Slack)

emacs & VS Code

You have two editor options for Bootcamp:

- emacs
- Visual Studio Code (VS Code)

emacs & VS Code

You have two editor options for Bootcamp:

- emacs
- Visual Studio Code (VS Code)

Consider talking to your TAs and older students when deciding.