Unifying Data Science

Nick Eubank

By the end of this class, you will be able to:

1. Understand how data science tools relate to one another using a unified conceptual framework,
2. Answer causal questions
   Does X cause Y?
3. Execute a data science project from conception to delivery

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.:

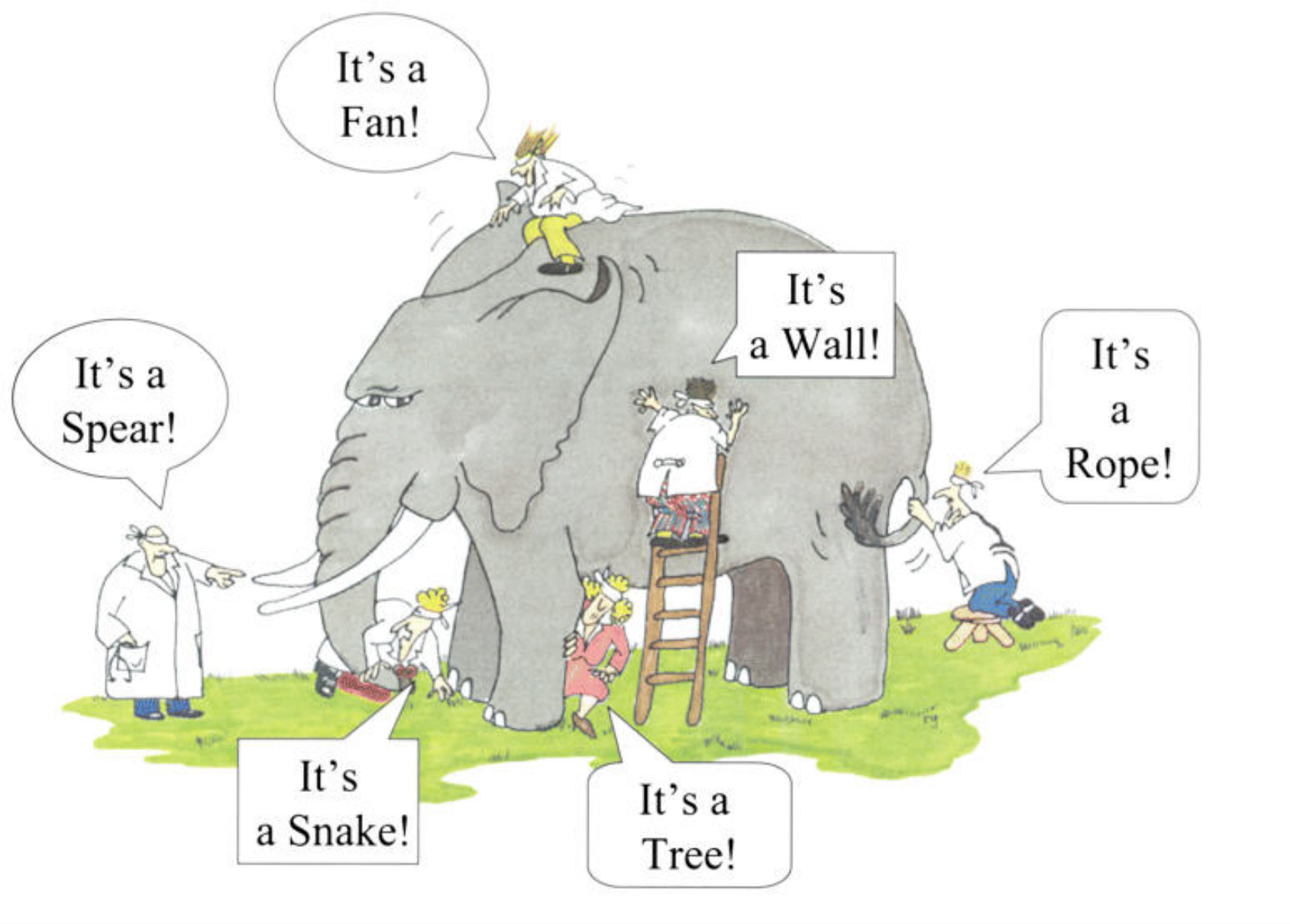
Where are we today?

Very little cross-pollination across silos

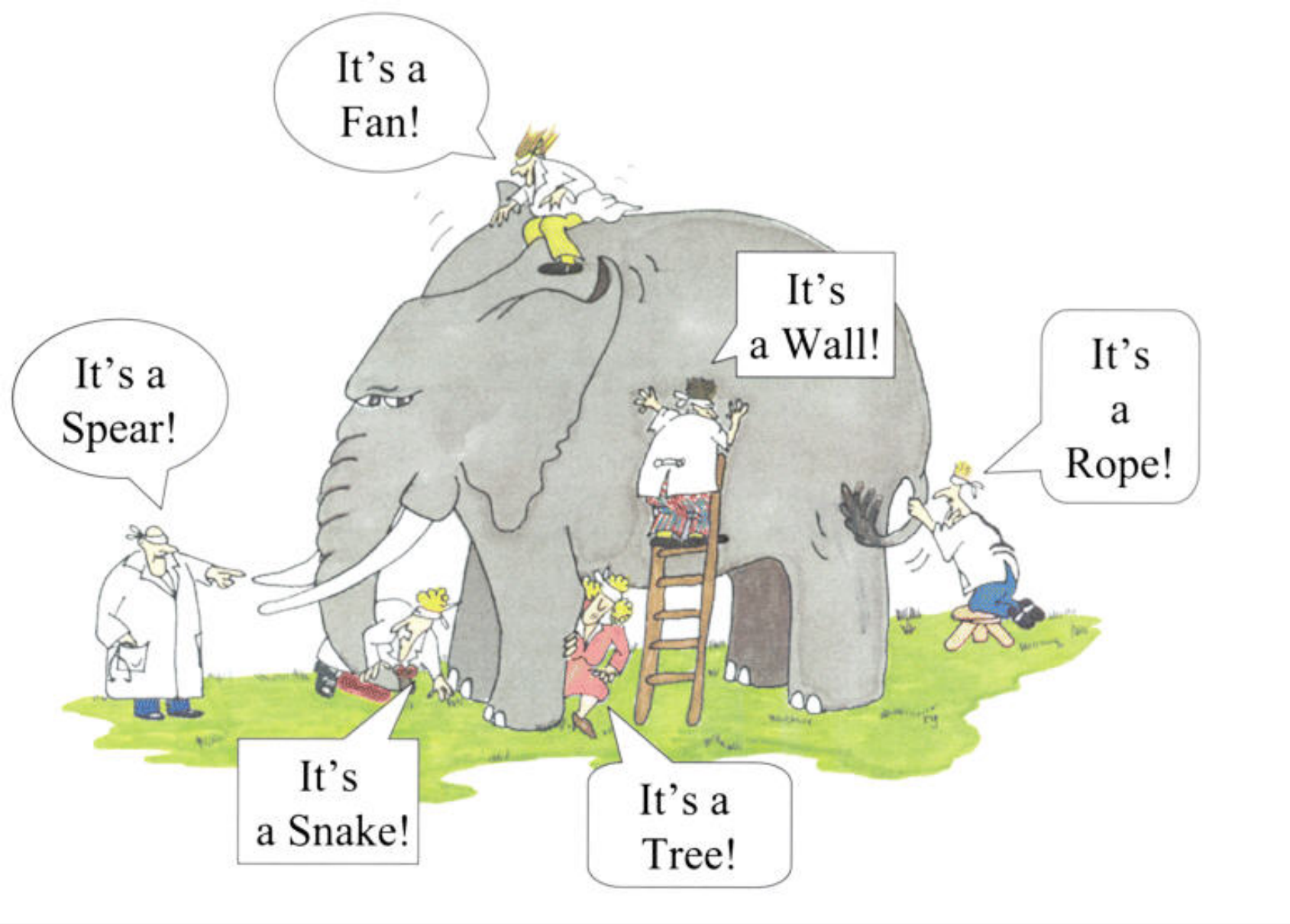
- Lots of duplication of development.
- Every silo has its own vocabulary.
- Each silo has focused on the aspects most relevant to their applications. e.g.:
  - CS likes to classify things and make predictions, don't care how model works
  - Social scientists like to make causal statements, don't care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

This Class

1. Introduce a taxonomy of questions

   Descriptive, causal, predictive
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

   Descriptive, causal, predictive
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
   Descriptive, causal, predictive
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
  e.g. experiments, matching, regression, differences-in-differences
- Other approaches to descriptive analysis

This Class

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

This Class

1. Introduce taxonomy of questions
2. Discuss descriptive questions
3. Learn causal inference
4. Discuss prediction

This Class

1. Introduce taxonomy of questions
   Practice generating questions
2. Discuss descriptive questions

3. Learn causal inference

4. Discuss prediction

This Class

1. Introduce taxonomy of questions
   Practice generating questions
2. Discuss descriptive questions
   Relatively brief
3. Learn causal inference

4. Discuss prediction

This Class

1. Introduce taxonomy of questions
   Practice generating questions
2. Discuss descriptive questions
   Relatively brief
3. Learn causal inference
   Deep dive – $\sim$ half the semester
4. Discuss prediction

This Class

1. Introduce taxonomy of questions
   Practice generating questions
2. Discuss descriptive questions
   Relatively brief
3. Learn causal inference
   Deep dive – $\sim$ half the semester
4. Discuss prediction
   Relative merits of supervised machine learning v. causal methods

Data Science Project

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

- Teams of 3-4, grouped by interest and experience
- On topic of your own choosing

Data Science Project

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

- Teams of 3-4, grouped by interest and experience
- On topic of your own choosing

$\rightarrow$ Nice portfolio piece

Data Science Project

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

- Teams of 3-4, grouped by interest and experience
- On topic of your own choosing

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

Who Are We?

I am a social scientist

Who Are We?

I am a social scientist

- PhD in Political Economy, Masters in Economics, BA in Economics and Political Science
- Research on international development, social networks, election administration, gerrymandering

Who Are We?

I am a social scientist

- PhD in Political Economy, Masters in Economics, BA in Economics and Political Science
- Research on international development, social networks, election administration, gerrymandering

Zeren Li (TA)

Who Are We?

I am a social scientist

- PhD in Political Economy, Masters in Economics, BA in Economics and Political Science
- Research on international development, social networks, election administration, gerrymandering

Zeren Li (TA)

- PhD Candidate in Political Science
- Studies Chinese politics
- Strong background in causal inference and machine learning

Last Notes On This Class

Last Notes On This Class

- First time this class has been taught
  (Familiar material; but new audience)

Last Notes On This Class

- First time this class has been taught
  (Familiar material; but new audience)
- We'll do several evaluations over the semester of how things are going, and adjust as we go.

Last Notes On This Class

- First time this class has been taught
  (Familiar material; but new audience)
- We'll do several evaluations over the semester of how things are going, and adjust as we go.
- First few weeks won't be full representative.

Last Notes On This Class

- First time this class has been taught
  (Familiar material; but new audience)
- We'll do several evaluations over the semester of how things are going, and adjust as we go.
- First few weeks won't be full representative.
  - If you're deciding whether to take this class, I'd suggest buying *Mostly Harmless Econometrics* and skimming a few chapters to get a sense of material we'll focus on for much of semester.

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 quizzes are likely to be a regular feature of the class.

Things to Know

- Course site: `http://www.unifyingdatascience.org`
  Contents subject to change!
- Readings are *incredibly* important.
  Reading quizzes are likely to be a regular feature of the 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