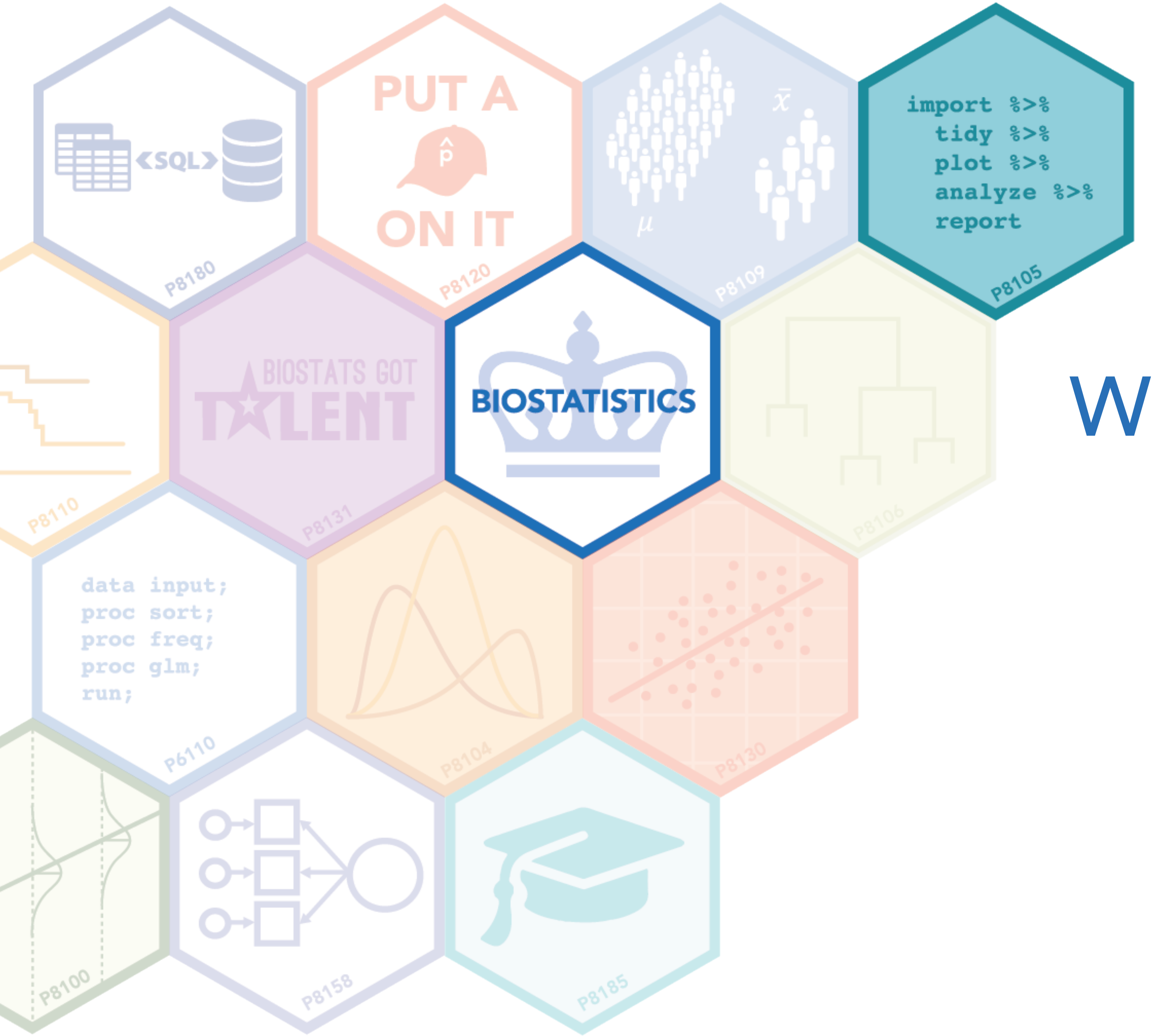


```
import %>%
  tidy %>%
  plot %>%
  analyze %>%
  report
```

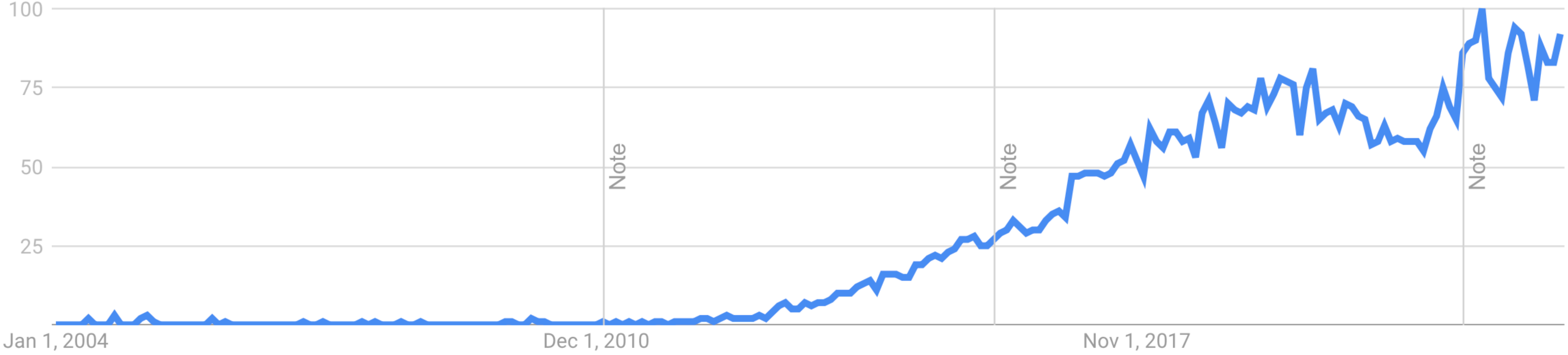
```
data input;
proc sort;
proc freq;
proc glm;
run;
```

# WHAT IS DATA SCIENCE?

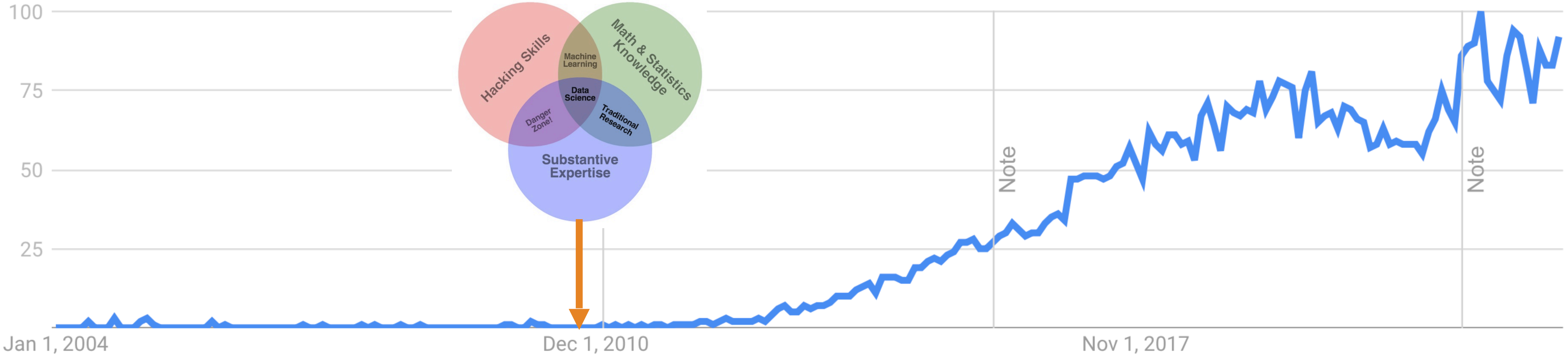
Jeff Goldsmith, PhD

Department of Biostatistics

# Data science is pretty new



# Data science is pretty new

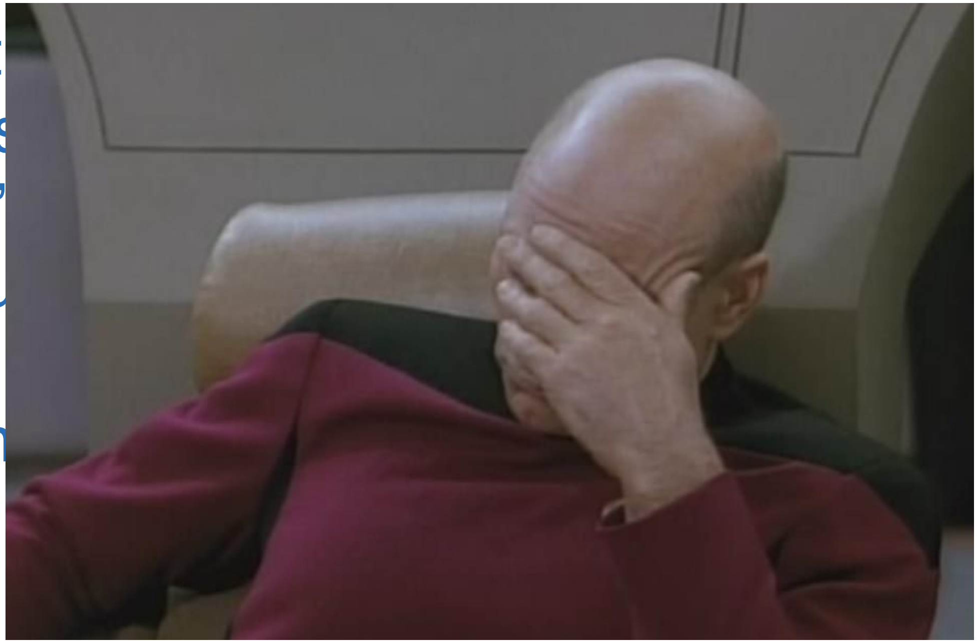


# Some not great definitions

- Data science = statistics
- Data science = computer science
- Data science = machine learning
- Data science = statistics + computer science + machine learning
- Data scientists are big data wranglers
- “A data scientist is just a sexier word for statistician.” –Nate Silver
- “A data scientist is a better computer scientist than a statistician and is a better statistician than a computer scientist.”
- “A data scientist is a statistician who is useful” – Hadley Wickham
- A data scientist is a *good* statistical analyst
- A data scientist is a statistician who codes in python

# Some not great definitions

- Data science = statistics
- Data science = computer science
- Data science = machine learning
- Data science = statistics + computer science + machine learning
- Data scientists are big data wranglers
- "A data scientist is just a sexier word for stat
- "A data scientist is a better computer scientis better statistician than a computer scientist."
- "A data scientist is a statistician who is usefu
- A data scientist is a *good* statistical analyst
- A data scientist is a statistician who codes in



# Maybe pictures will help?

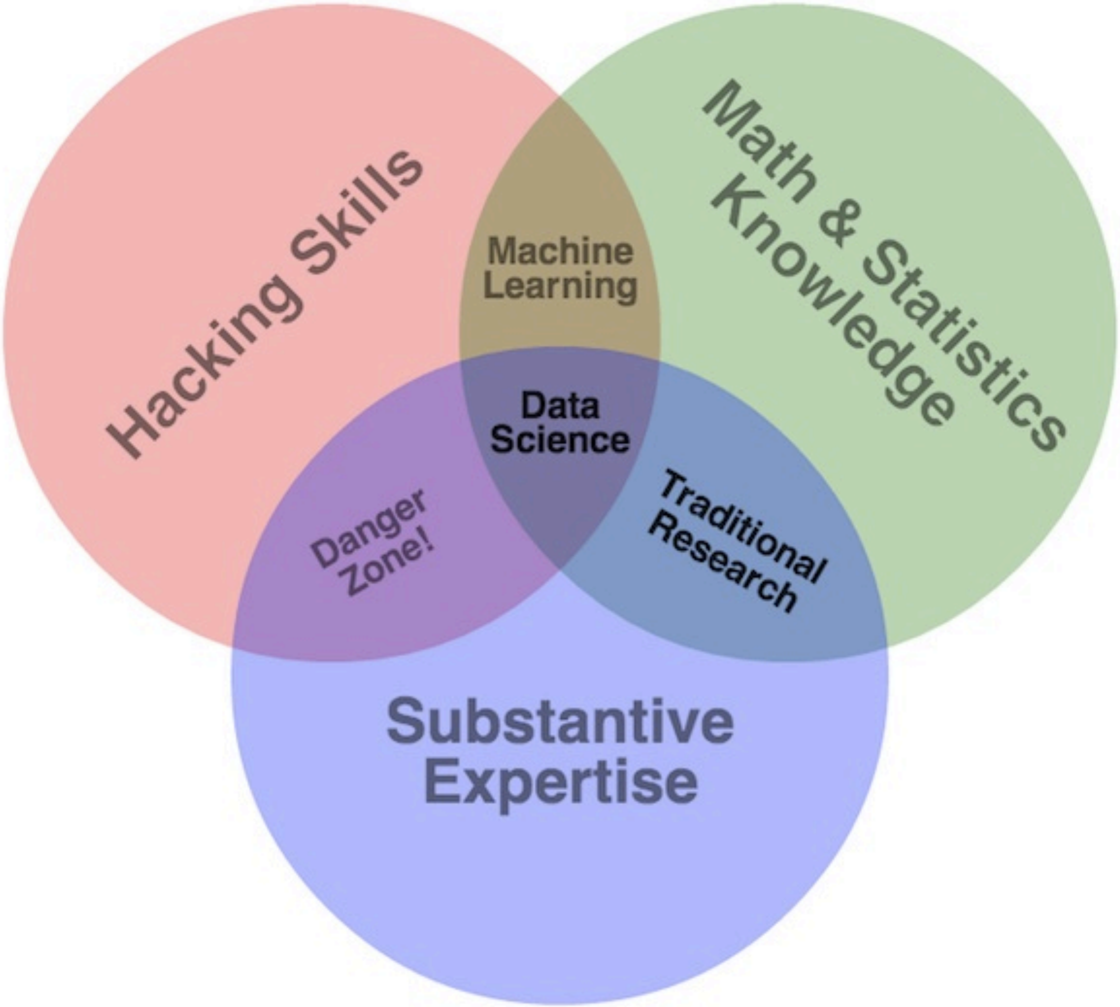
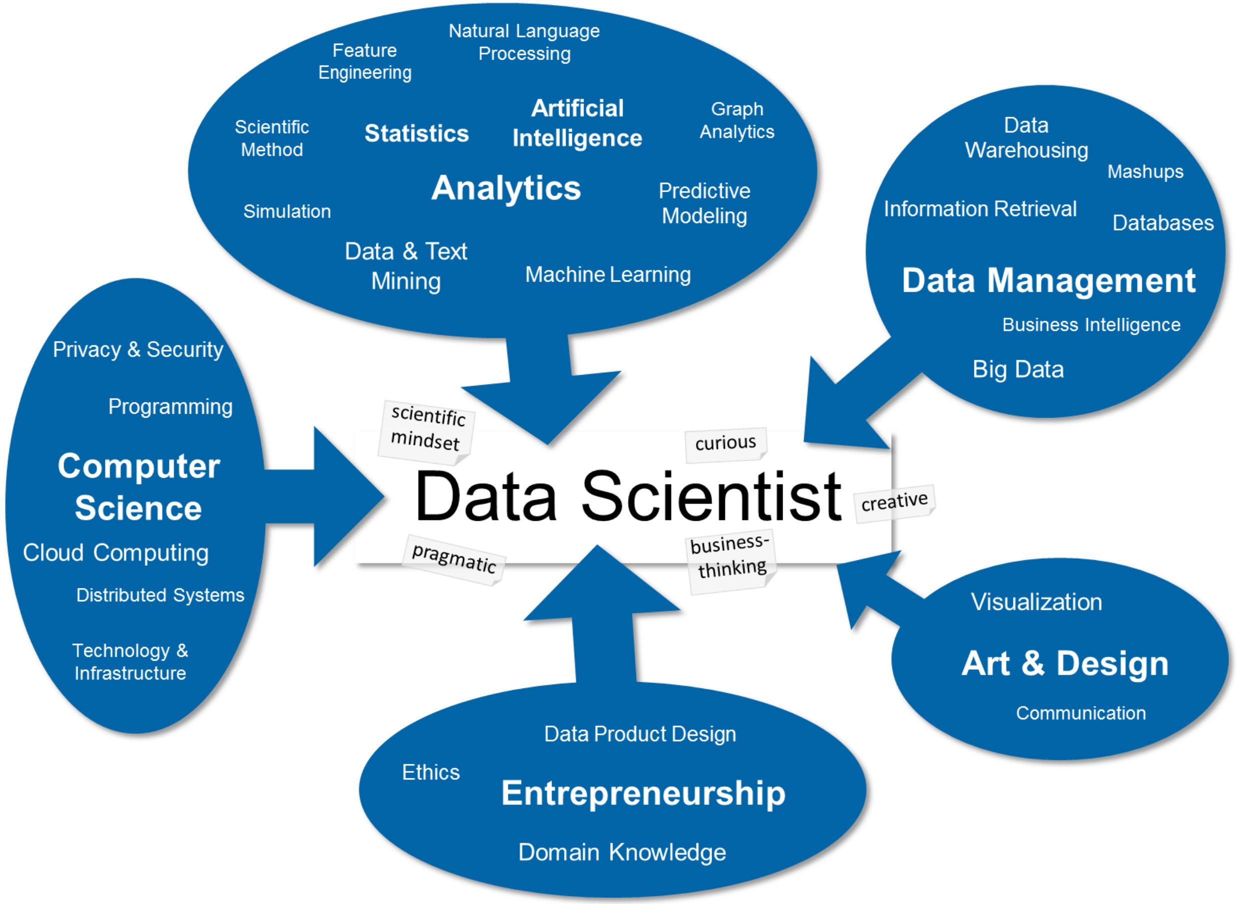


Image from Drew Conway

# Maybe pictures will help?



https://blog.zhaw.ch/datascience/the-data-science-skill-set/

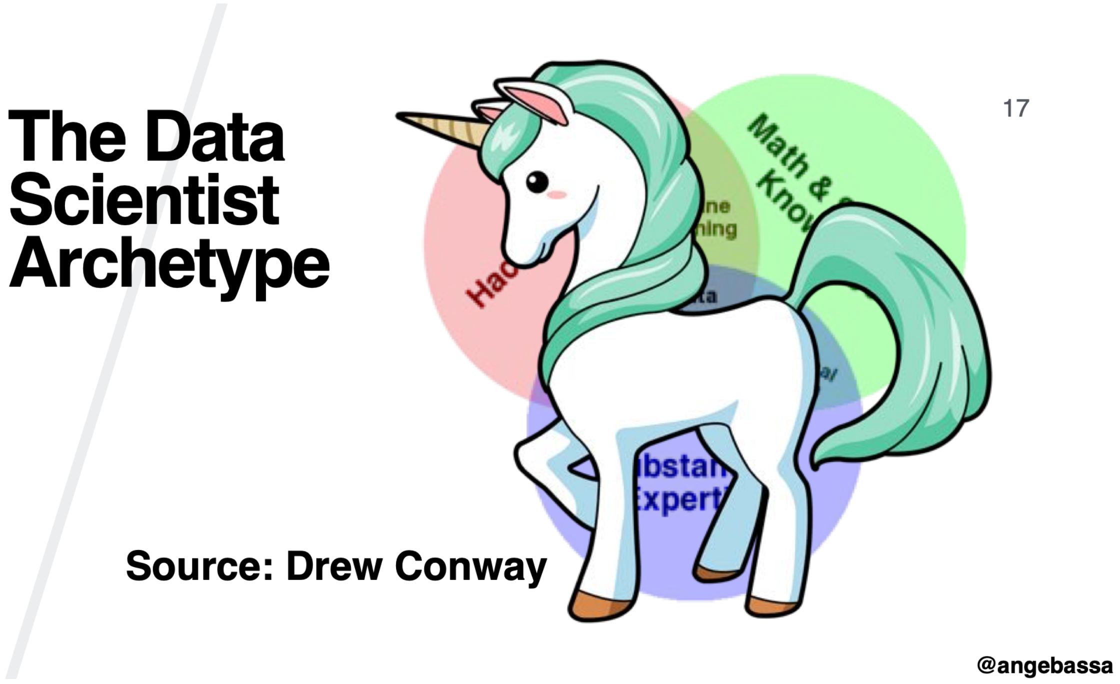
# Why these definitions are bad

- "Data science is just …" definitions miss the point
  - If data science is just statistics (or machine learning, or computer science, or engineering) we wouldn't need a new term, let alone a new discipline
  - The popularity of "data science" suggests that there's a newly recognized need

- "A data scientist is a *good* " whatever definitions aren't helpful
  - They're almost deliberately judgmental
  - A good definition doesn't depend on opinions
  - There are "data scientists" in each discipline, but some very good statisticians / computer scientists / etc aren't "data scientists"

# Why these definitions are bad

- “Data science is the combination of these 40 skills …” are unrealistic



The Data Scientist Archetype

Source: Drew Conway

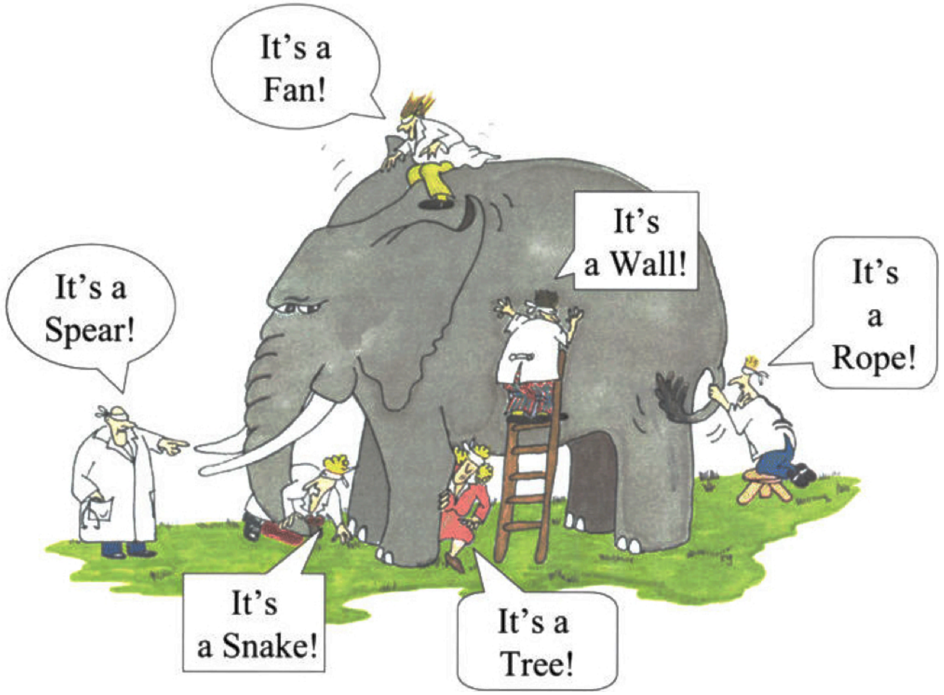
17

@angebassa

https://www.youtube.com/watch?v=b9ZLXwAuUyw&app=desktop

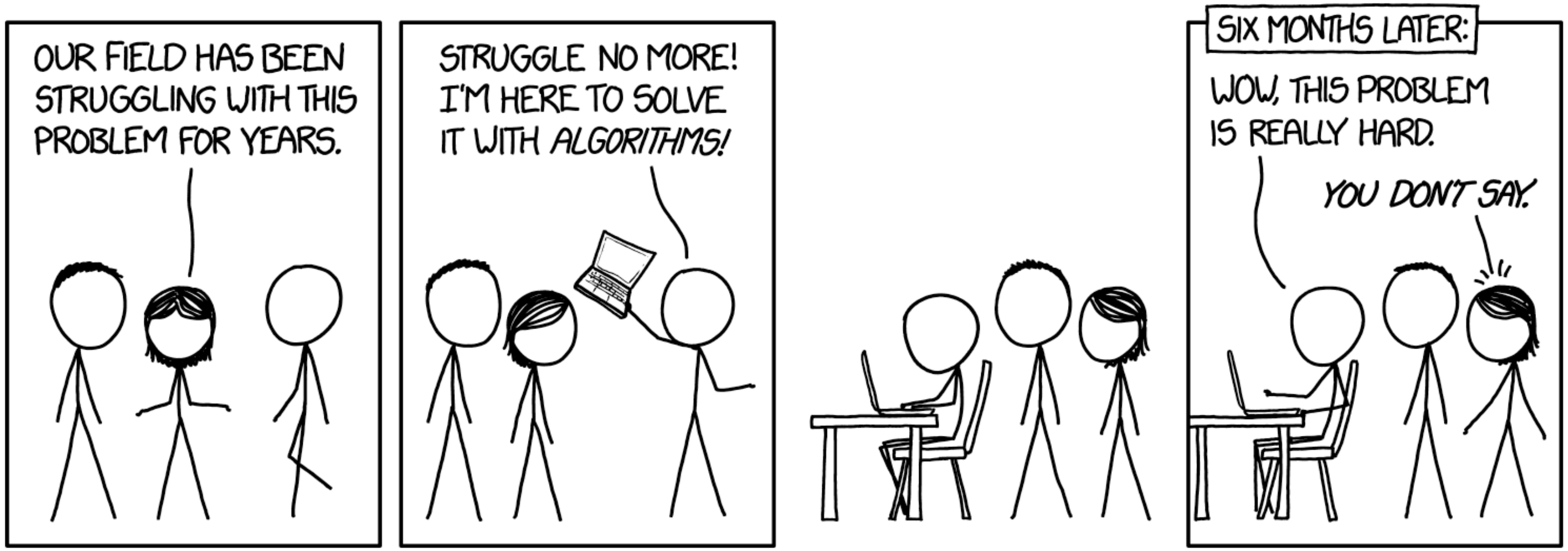
# Why these definitions are good

- Kinda like the blind men and the elephant – no one perspective is completely right or completely wrong, but piling them all up isn't right either
- They give a sense of what is valued by the data science community – using data in a principled way and coding well



# Why these definitions are good

- Data science is interdisciplinary
  - You do need a breadth of skills
  - You also need a particular mindset – curiosity and engagement is critical
  - You need some domain knowledge to be successful


OUR FIELD HAS BEEN STRUGGLING WITH THIS PROBLEM FOR YEARS.
STRUGGLE NO MORE! I'M HERE TO SOLVE IT WITH *ALGORITHMS!*
SIX MONTHS LATER:
WOW, THIS PROBLEM IS REALLY HARD.
*YOU DON'T SAY.*

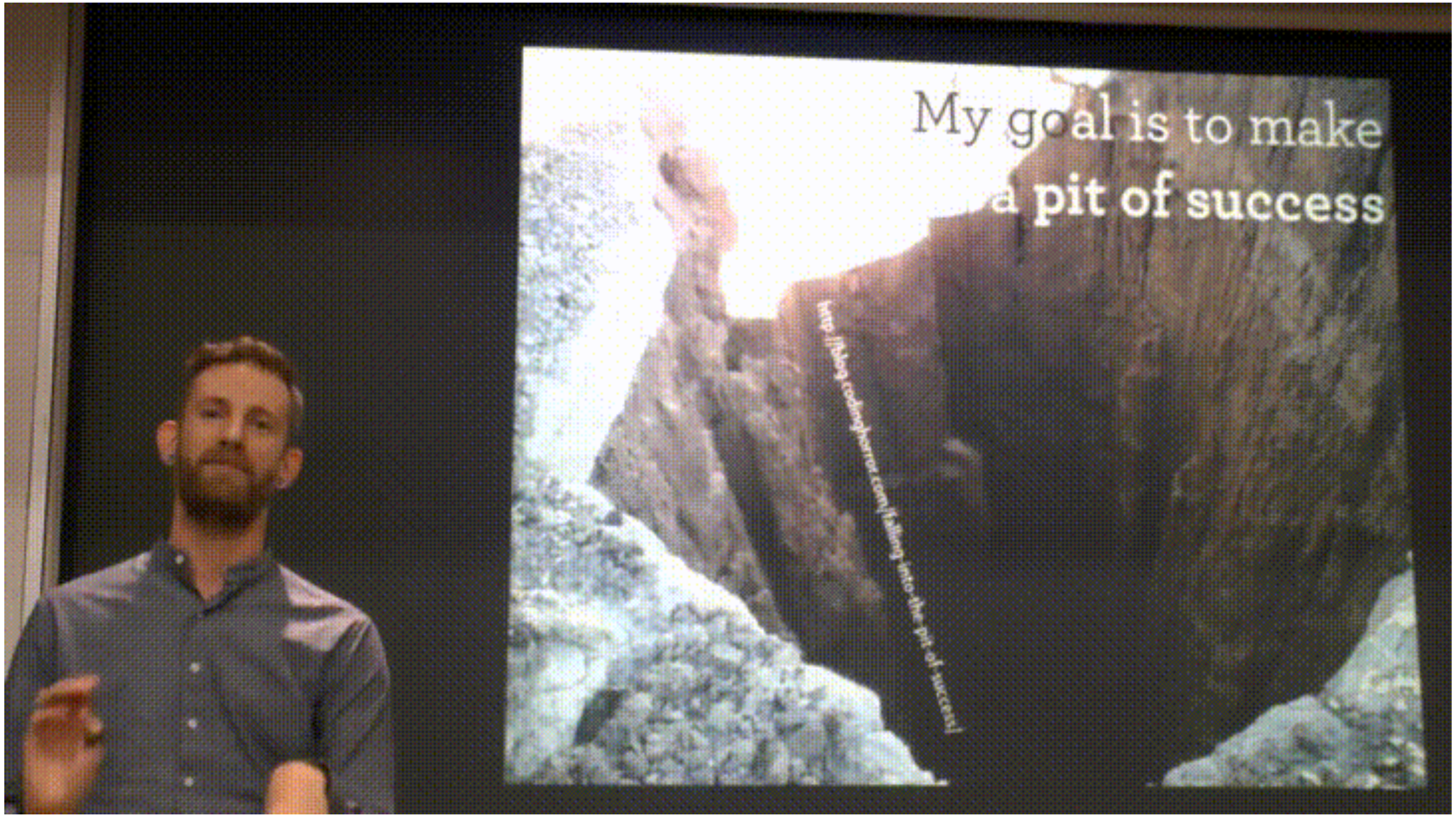
https://www.xkcd.com/1831/

# For the purpose of this class:

Data science is the study of formulating and rigorously answering questions using a data-centric process that emphasizes clarity, reproducibility, effective communication, and ethical practices.

- We'll focus mostly on process; how to formulate and answer questions through analyses are the focus of other courses
- This is also a "bad" definition, in that it doesn't explain where data science came from

# ISI 2017



My goal is to make a pit of success

http://blog.codinghorror.com/falling-into-the-pit-of-success/

# First question from the audience

“What is the point of ‘data science’? Aren’t *we* **already** data scientists?”

# First question from the audience

“What is the point of ‘data science’? Aren’t *we* **already** data scientists?”

# Response from Hadley Wickham (roughly)

“A data scientist is a statistician who’s useful”

# Response from Hadley Wickham (roughly)

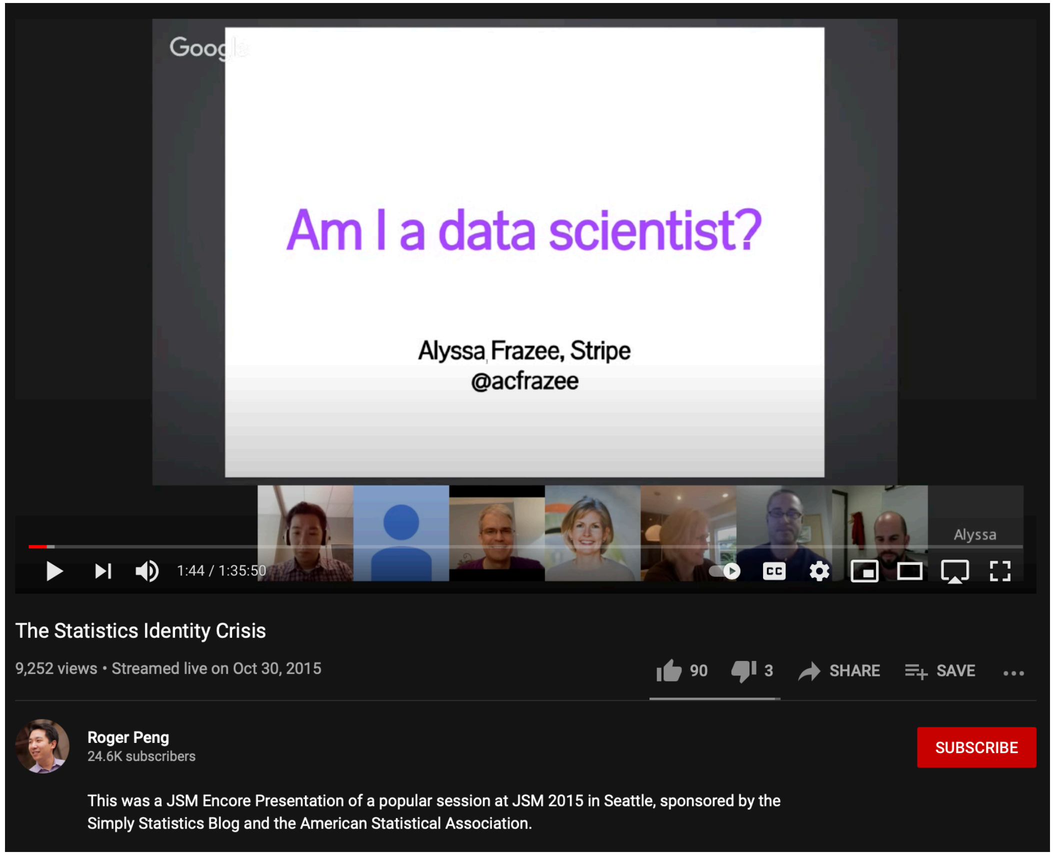
“A data scientist is a statistician who’s useful”

# That question is understandable

- It’s easy, in 2021, to forget what the statistical identity crisis phase was like
- But that was a whole thing, for quite a while

# That question is understandable

- It’s easy, in 2021, to forget what the statistical identity crisis phase was like
- But that was a whole thing, for quite a while



Am I a data scientist?

Alyssa Frazee, Stripe
@acfrazee

The Statistics Identity Crisis

9,252 views • Streamed live on Oct 30, 2015

Roger Peng
24.6K subscribers

This was a JSM Encore Presentation of a popular session at JSM 2015 in Seattle, sponsored by the Simply Statistics Blog and the American Statistical Association.

# What made “data science” happen

- Data science emerged in parallel to (at least) six broad trends:
  - Big data
  - Emphasis on prediction
  - Reproducibility crisis in science
  - Interdisciplinary research
  - Diversity, equity, and inclusion
  - Everything should be on the internet

- These weren’t new in 2012 and aren’t unique to data science
- … but they had a big impact on the “data science” perspective

# Connotation >> definition

- Core data science values aren’t built into the definition, but were critical to the valence of “data science”

# Public Health Data Science

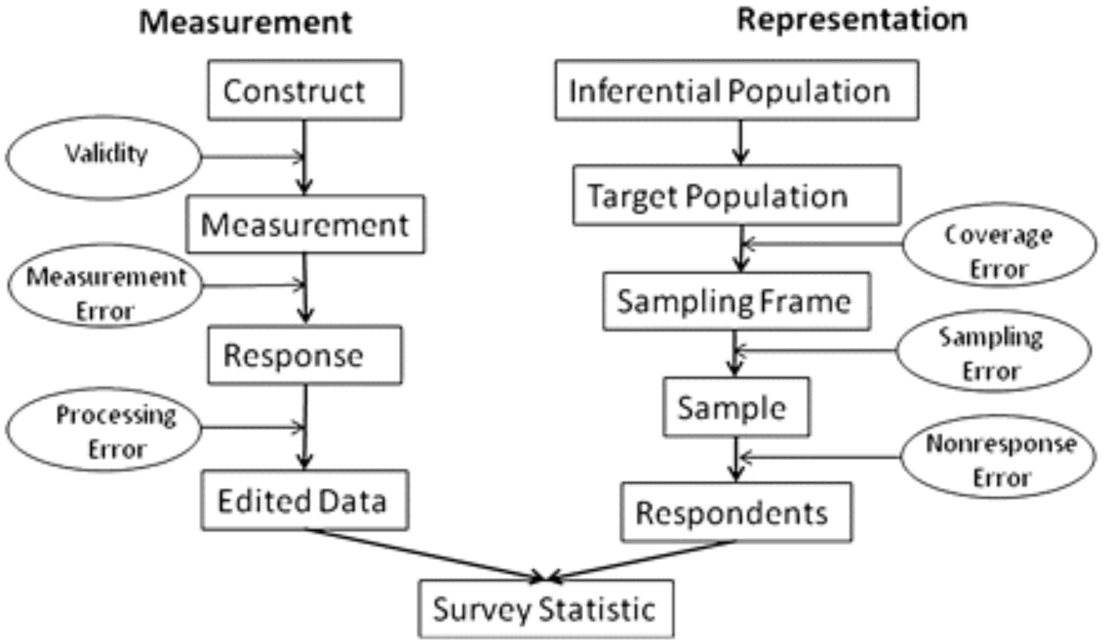
[Public health] data science is the study of formulating and rigorously answering questions [in order to advance health and well-being] using a data-centric process that emphasizes clarity, reproducibility, effective communication, and ethical practices.

# “Public Health” is the important part

- Public health training emphasizes some elements that are critical data science thinking and work:
  - Study design
  - Sampling process
  - Measurement process
  - Desire vs ability to infer causation
  - Cross-disciplinary collaboration
  - Engagement with data ethics
  - Public dissemination and dialog

# “Public Health” is the important part

- Public health training emphasizes some elements that are critical data science thinking and work:
  - Study design
  - Sampling process
  - Measurement process
  - Desire vs ability to infer causation
  - Cross-disciplinary collaboration
  - Engagement with data ethics
  - Public dissemination and dialog



From “Total Survey Error: Past, Present, and Future” (Groves and Lyberg)
*via “Data Alone Isn’t Ground Truth” by Angela Bassa*

# How to learn data science

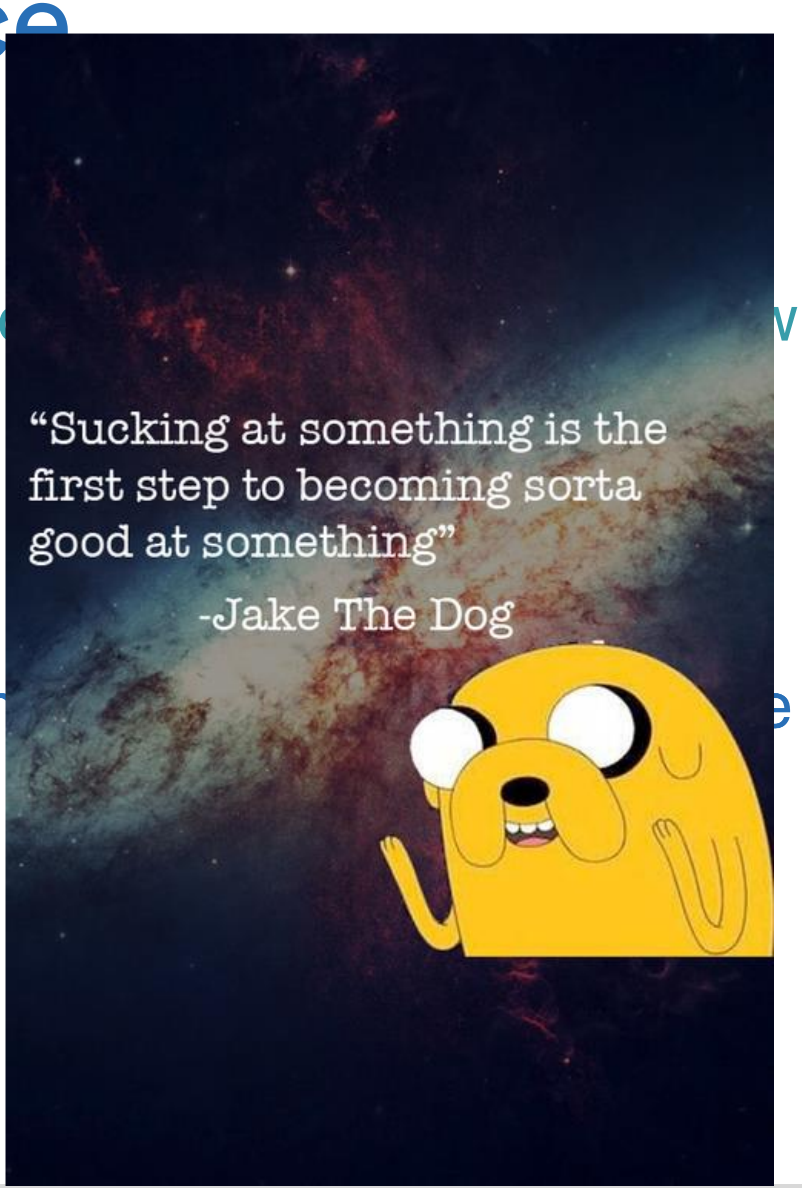
- Build a broad knowledge base
- Don’t be embarrassed by what you don’t know
  - Corollary: don’t be a jerk to people who don’t know what you know
- Ask questions (well) and keep learning

- Pretty much the same as learning anything, but hard because people don’t like to show their code

# How to learn data science

- Build a broad knowledge base
- Don’t be embarrassed by what you don’t
  - Corollary: don’t be a jerk to people wh v
- Ask questions (well) and keep learning

- Pretty much the same as learning anythin e don’t like to show their code

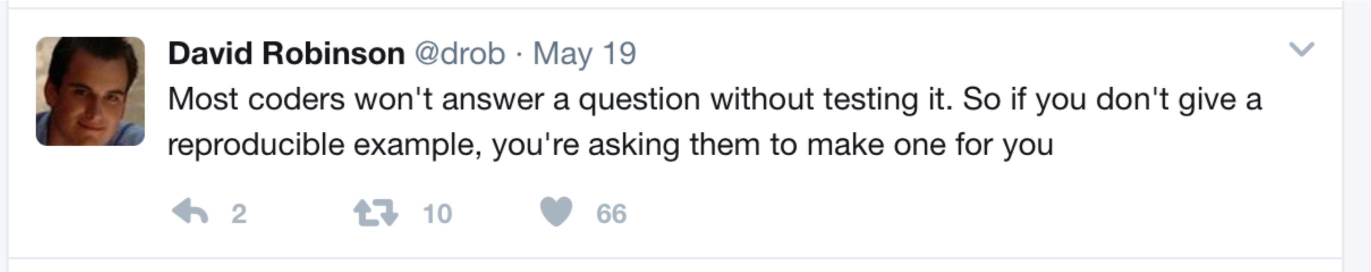


“Sucking at something is the first step to becoming sorta good at something”

-Jake The Dog

# How to learn data science

- All questions are good questions, but sometimes good questions aren't asked well
- Think through what you're trying to ask
- If your code is broken, create a simple example that illustrates what's broken



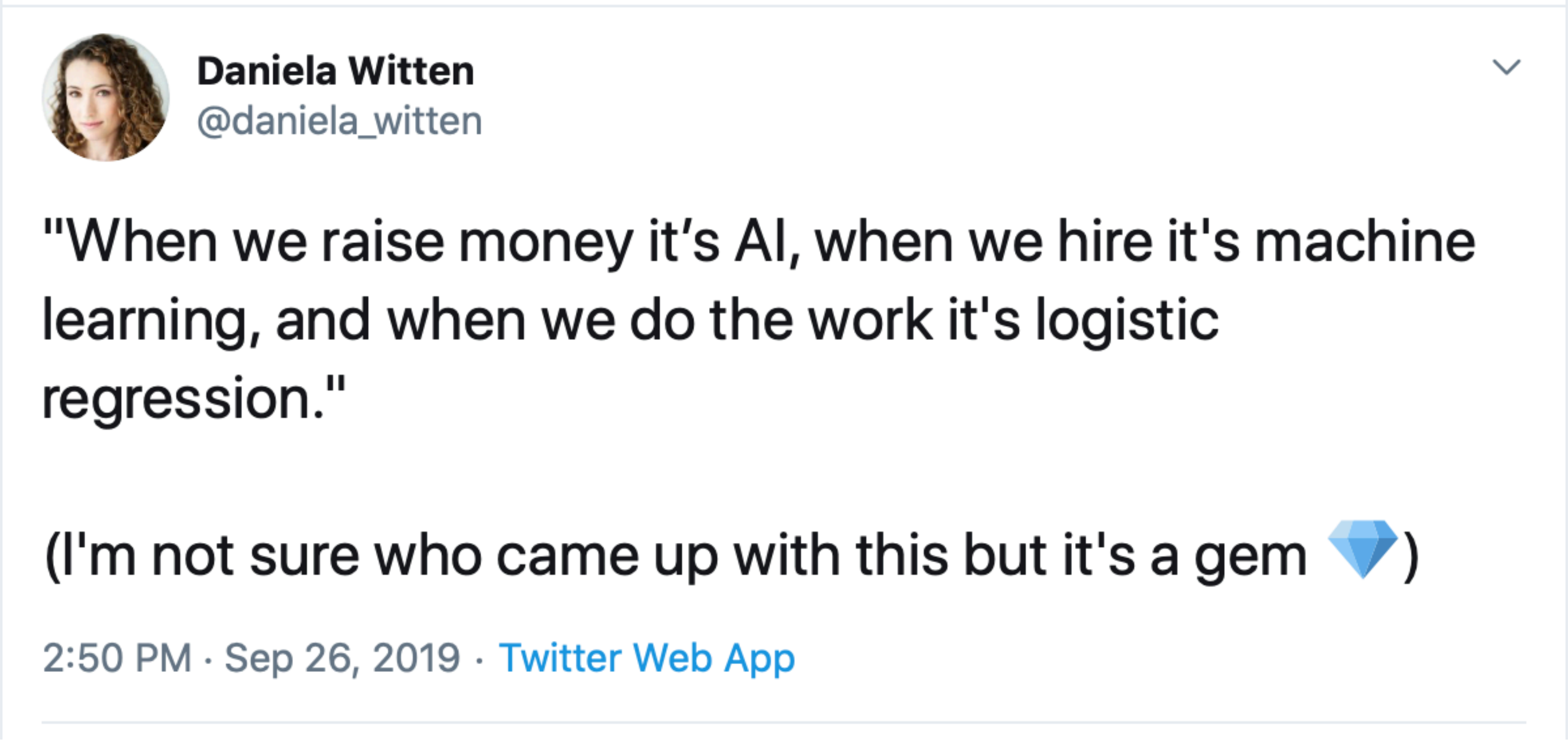
David Robinson @drob · May 19

Most coders won't answer a question without testing it. So if you don't give a reproducible example, you're asking them to make one for you

# How to learn data science

- Build up you “known knowns”
- Recognize your “known unknowns”
- Avoid “unknown unknows”

# Real talk about AI (as part of data science)



Daniela Witten
@daniela_witten

"When we raise money it’s AI, when we hire it's machine learning, and when we do the work it's logistic regression."

(I'm not sure who came up with this but it's a gem 💎)

2:50 PM · Sep 26, 2019 · Twitter Web App

# A data science analogy

1910s

# A data science analogy

1910s

1969 / 1970

# Reproducibility

- One concrete emphasis of data science is reproducibility
- Given the same data and the same code, anyone should be able to produce the same results
  - Code is an important means of communication
  - New tools encourage reproducibility, but the concept is not platform-dependent

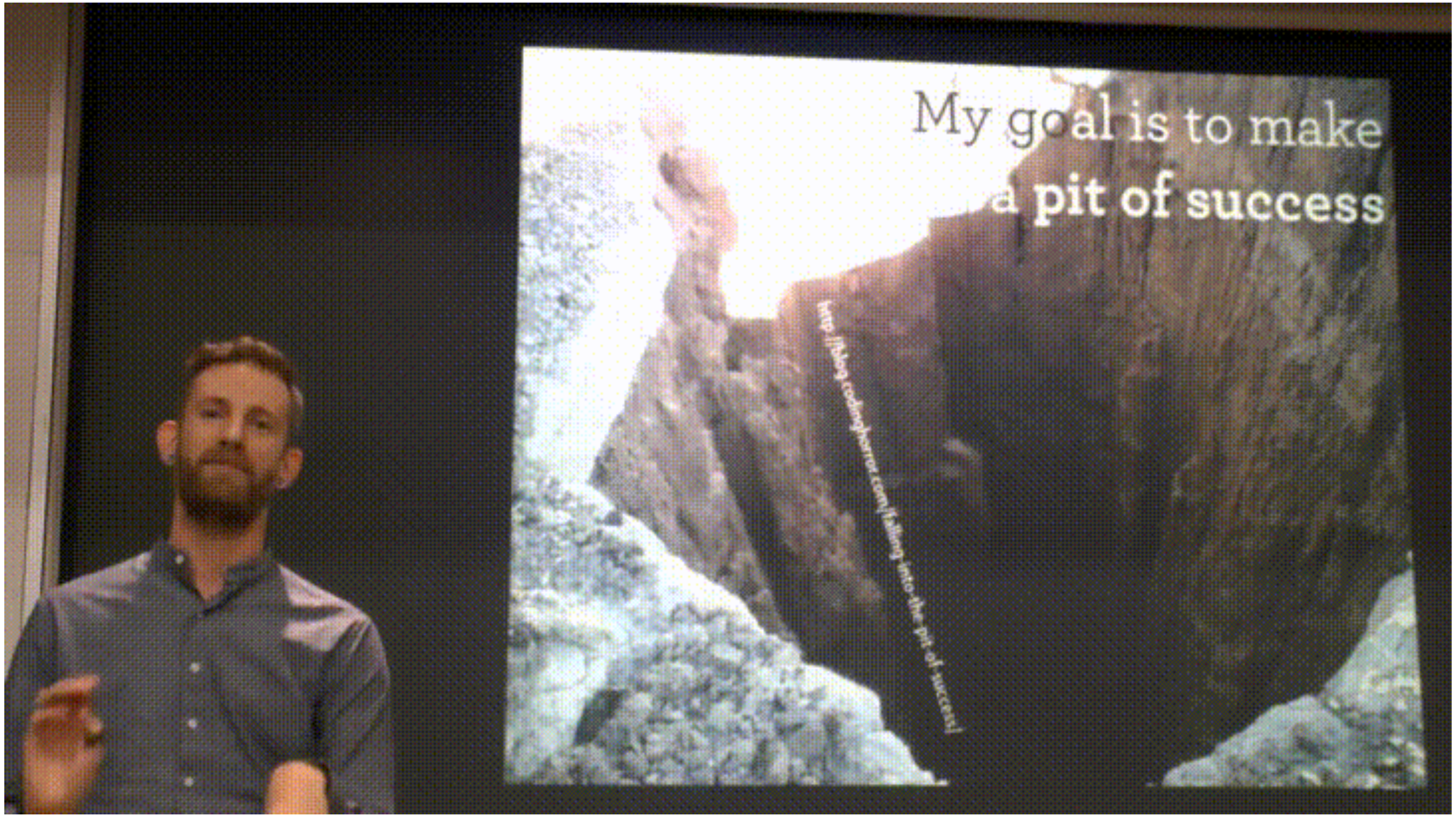
# Sharing code

- Openness is valuable – identify errors early and fix them quickly

- Try to think of sharing code as a gesture of confidence and humility
  - You've done your best, and you should feel good about that
  - Everyone makes mistakes sometimes; when you do, that's fine – fix it and move on

- Lack of transparency can reflect a lot of things
- Of these, arrogance is the most dangerous

# Choosing data science tools


My goal is to make
a pit of success
http://blog.codinghorror.com/falling-into-the-pit-of-success/

# Time to code!!