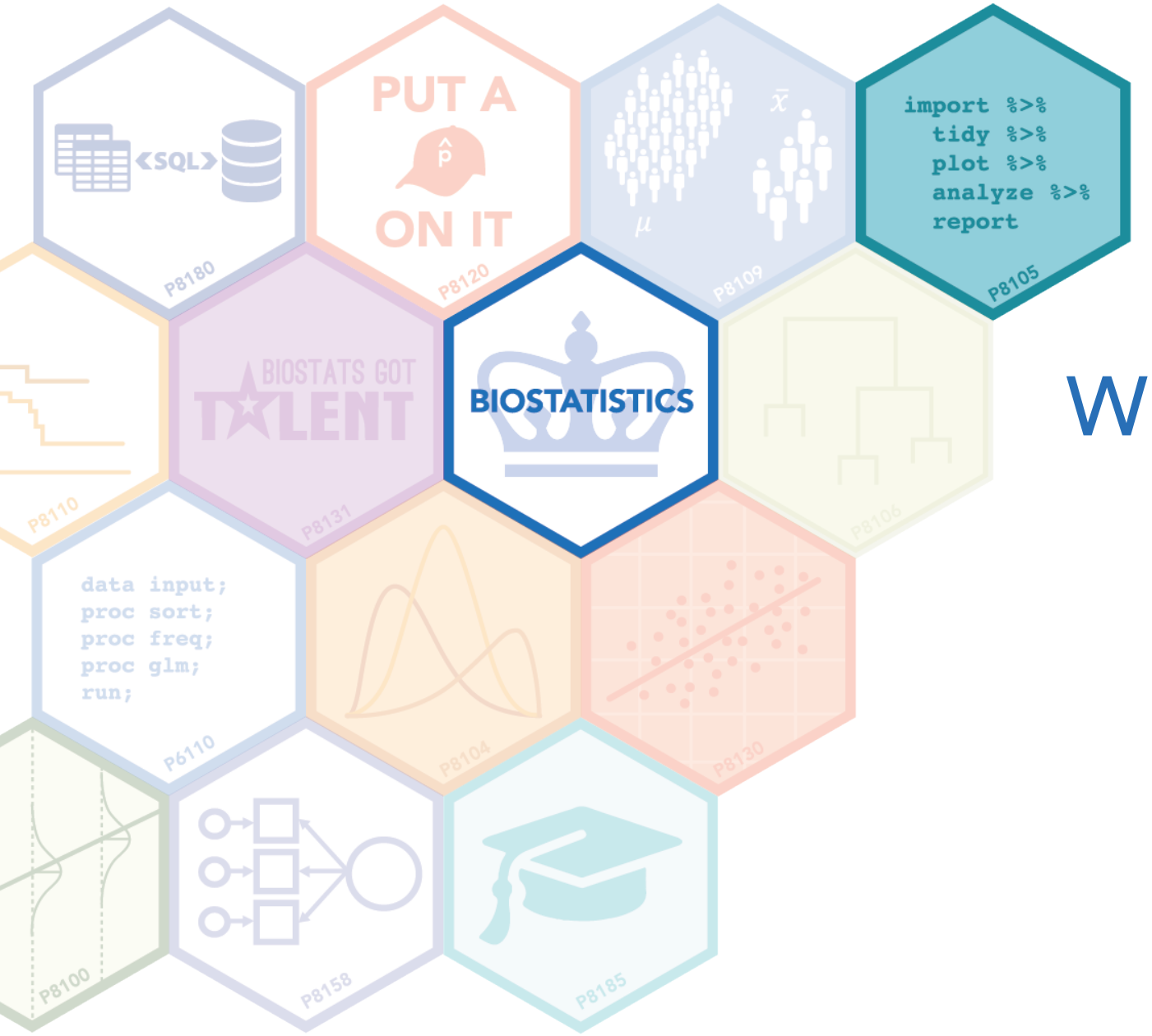


WHAT IS DATA SCIENCE?

Jeff Goldsmith, PhD

Department of Biostatistics

Data science is pretty new

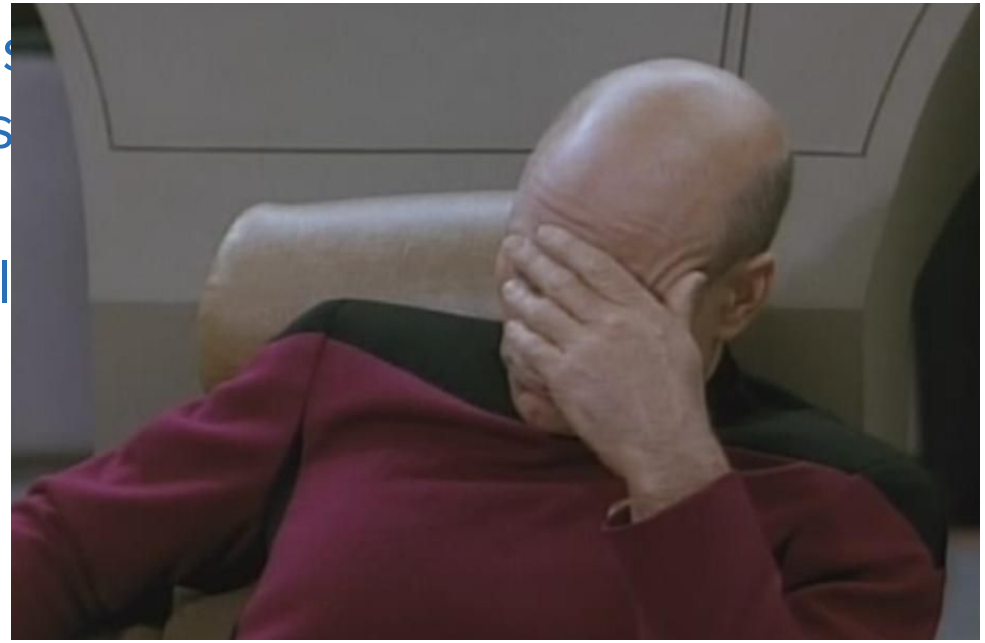
Source: Google Trends

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 statis
- “A data scientist is a better computer scientis statistician than a computer scientist.”
- “A data scientist is a statistician who is useful
- A data scientist is a good statistical analyst
- A data scientist is a statistician who codes in



Maybe pictures will help?

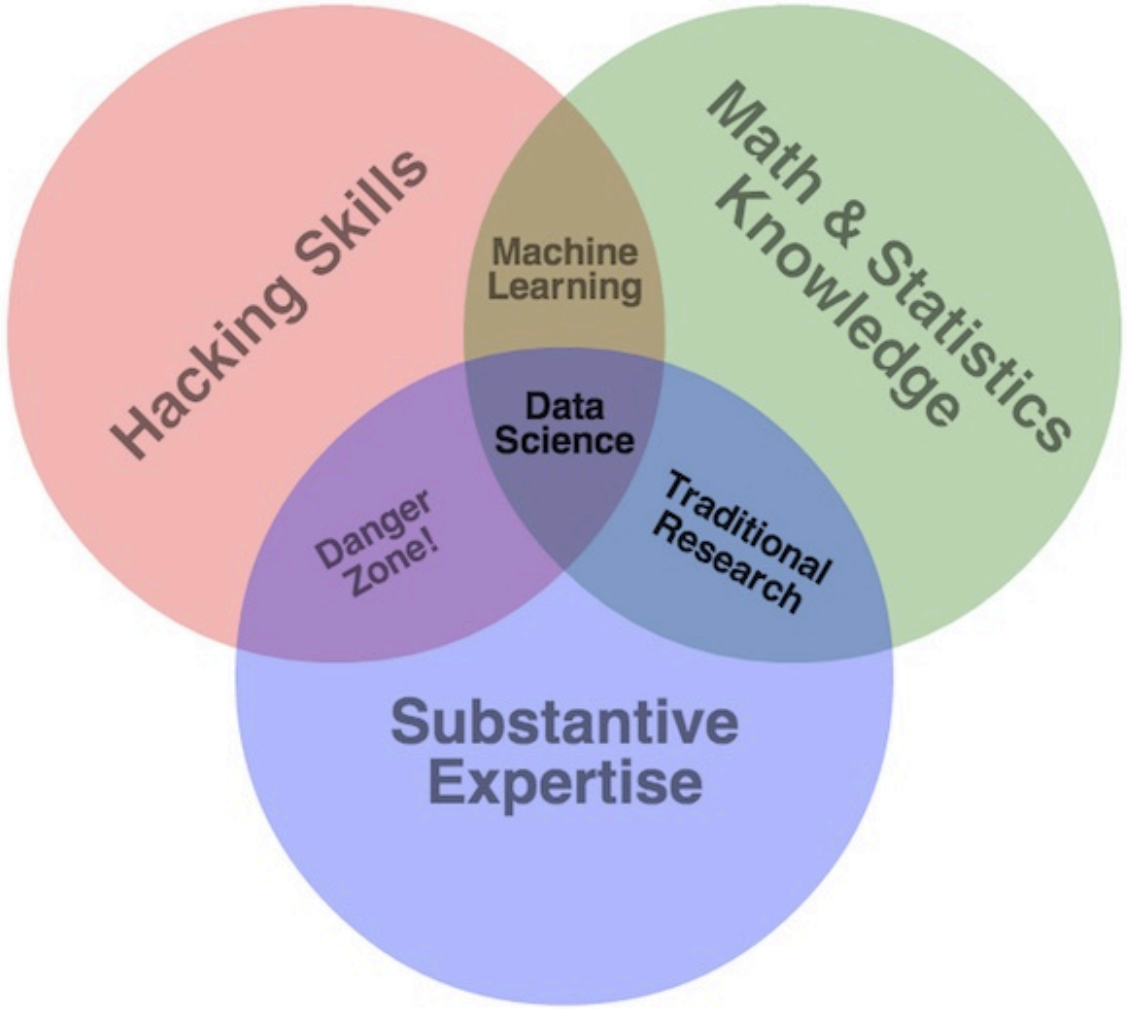
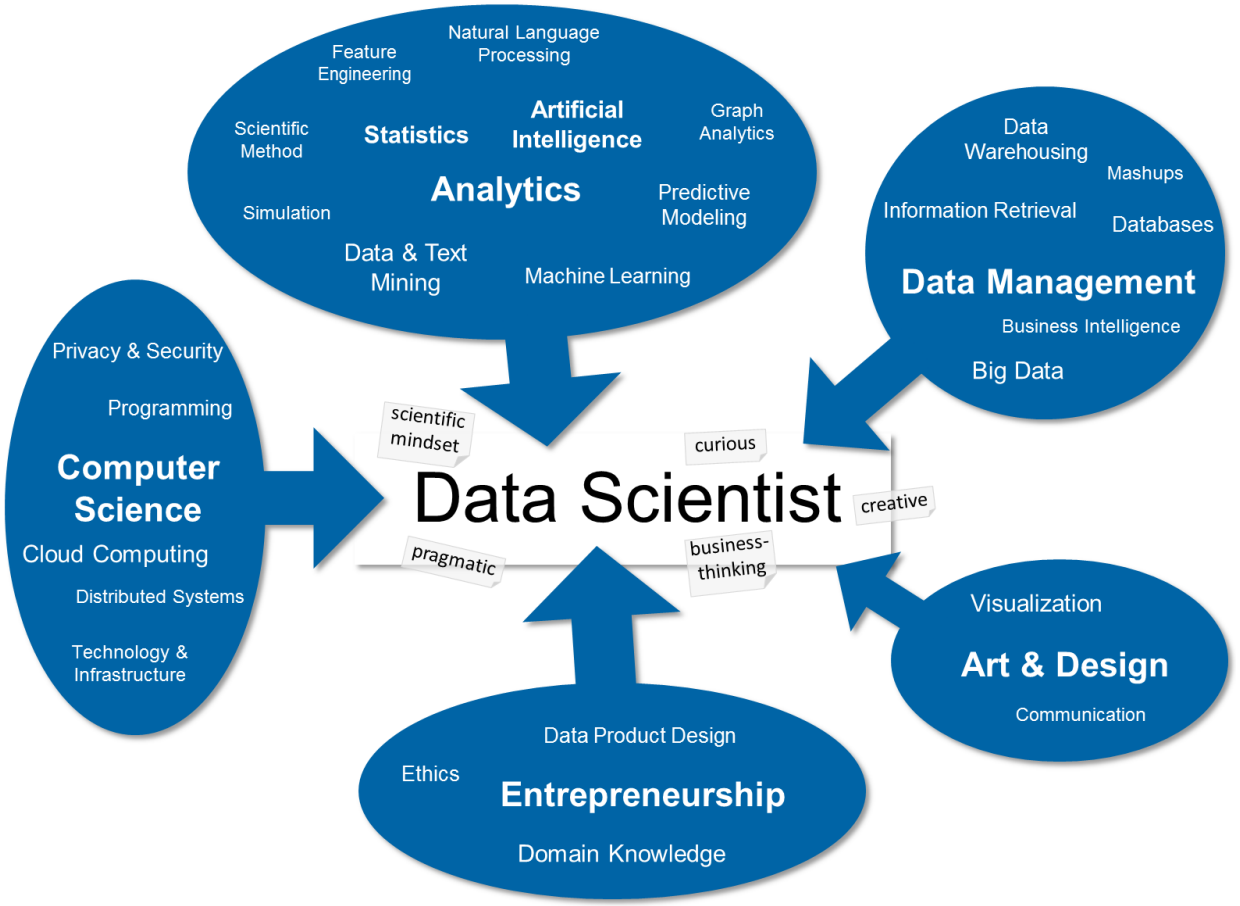


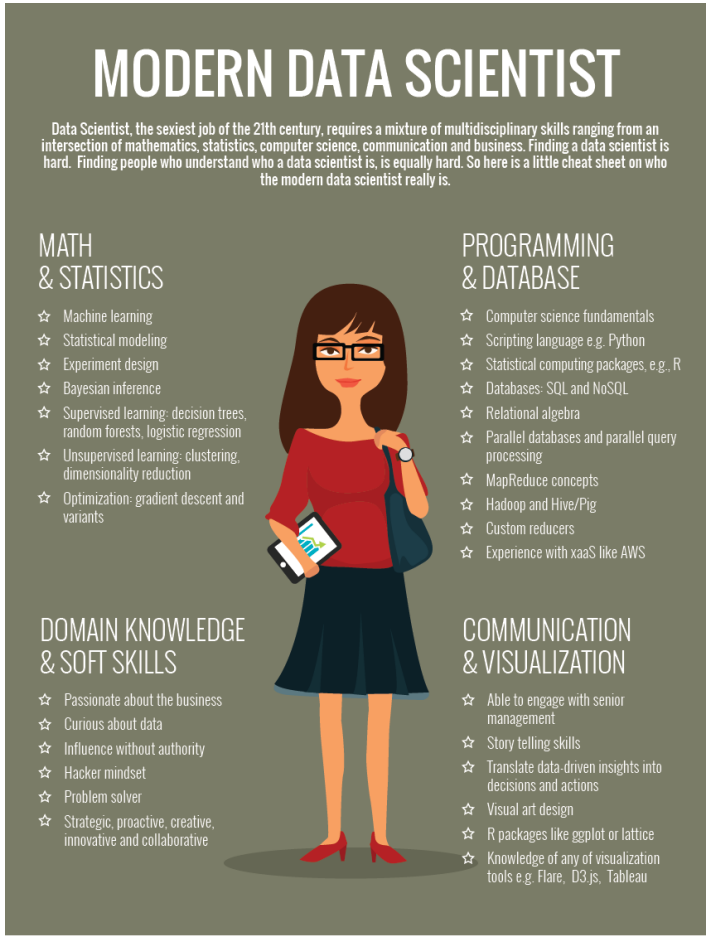
Image from Drew Conway

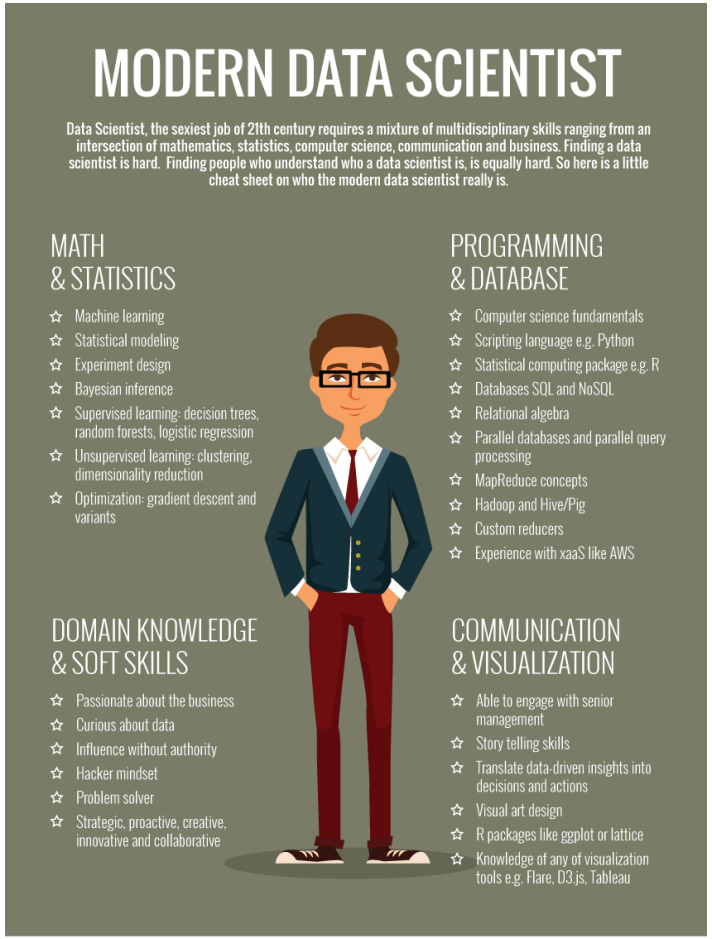
Maybe pictures will help?



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

Maybe pictures will help?





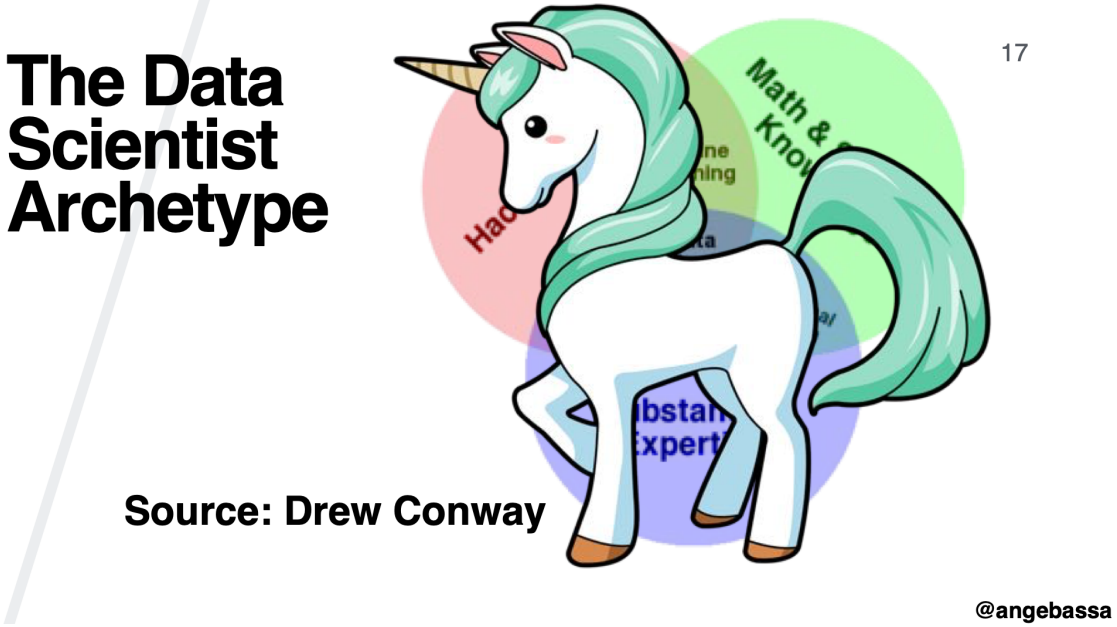
Why these definitions are bad

- "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

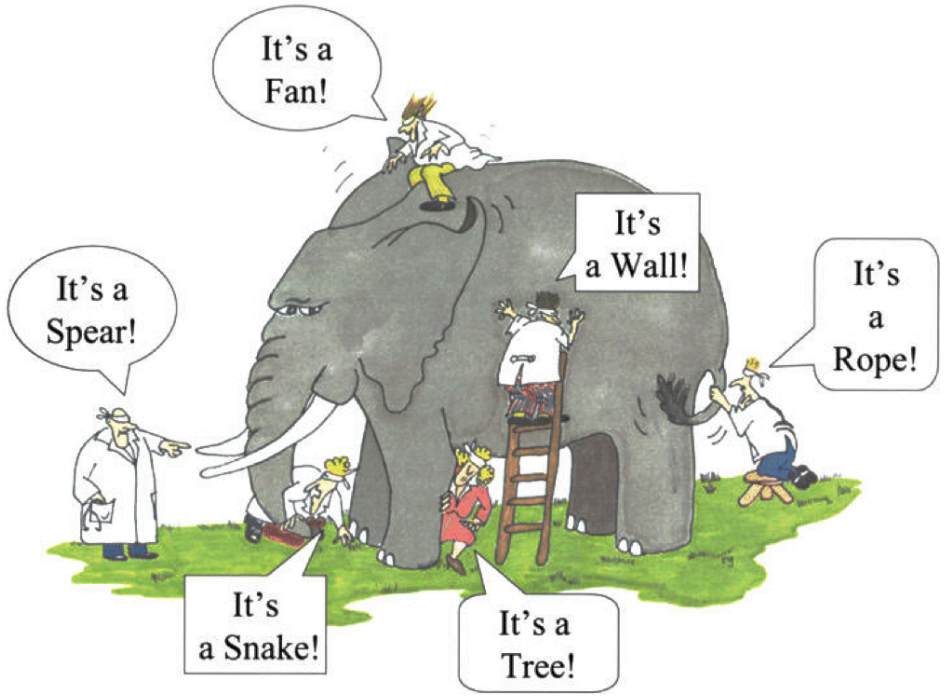
- "Data science is the combination of these 40 skills …" are unrealistic



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

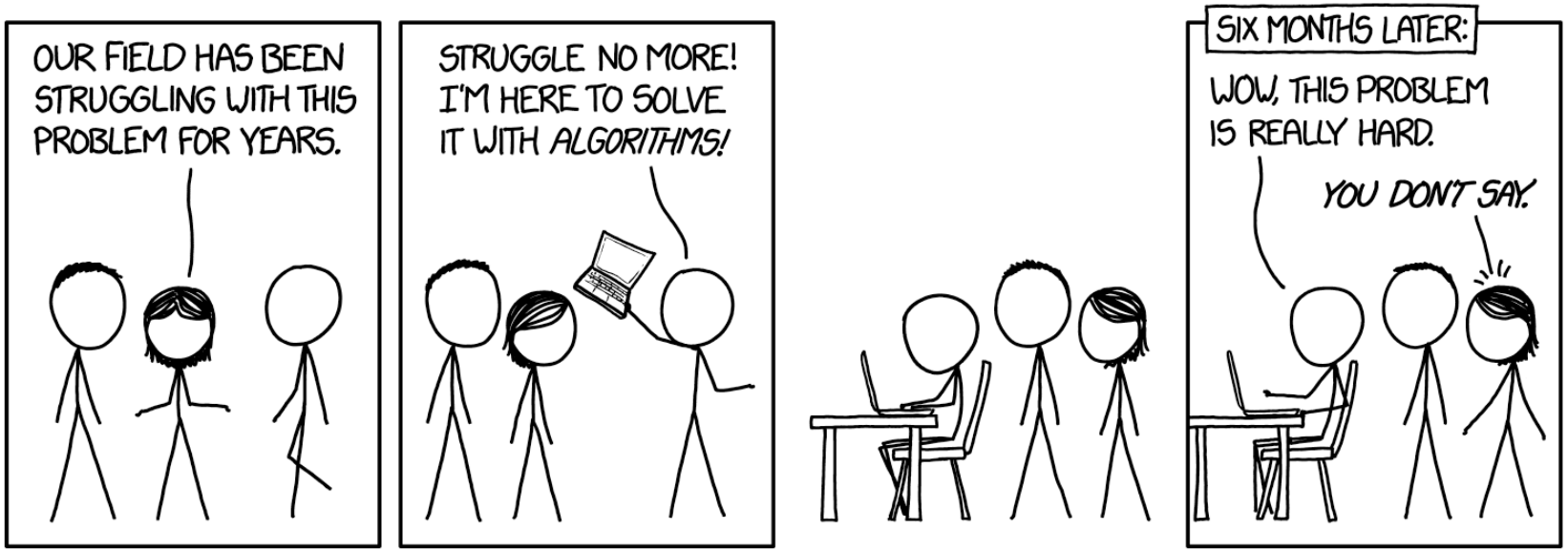
Why these definitions are good

- 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



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

What made “data science” happen

- Data science emerged in parallel to four broad trends:
  - Big data
  - Emphasis on prediction
  - Reproducibility crisis
  - Interdisciplinary

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

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, and effective communication.

- We'll focus mostly on process; how to formulate and answer questions through analyses are the focus of other courses

A data science analogy

- In 1903, the Wright Brothers had their first flight
- In 1908, the Model T rolled off the assembly line
- Both were huge advances in transportation, but in very different ways

First flight vs Model T

- Or: super fancy deep neural net vs ggplot2

- Both are data science!!
- ggplot2 may have more of an impact than AI

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

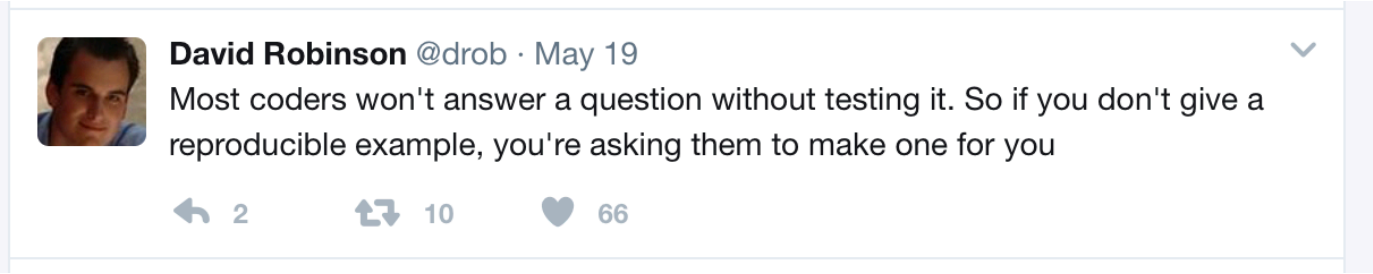
How to learn data science

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

- Pretty much the same as learning anything
  don't like to show their code

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



How to learn data science

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

DS twitter starter pack

- Follow these people to add some “knowns” to your repertoire

- @AmeliaMN
- @dataandme
- @drewconway
- @drob
- @hadleywickham
- @hmason
- @hspter
- @_inundata
- @jennybryan
- @johnmyleswhite
- @juliasilge
- @jtleek
- @kara_woo
- @kwbroman
- @rdpeng
- @robinson_es
- @seanjtaylor
- @sgrifter
- @statpumpkin
- @xieyihui

- #rstats
- #tidytuesday

Data as a resource

The world's most valuable resource is no longer oil, but data

The data economy demands a new approach to antitrust rules

Data as a resource

The world's most valuable resource is no longe

The data economy d

Sections ☰ The Washington Post

wp BrandStudio ⓘ Content from **IBM Power Systems**

Why big data is "the new natural resource"

Data as a resource

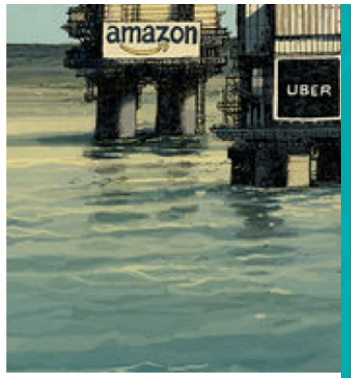
The world's most valuable resource
is no longe

The data economy d

Sections ☰ The Washington Post

wp BrandStudio ⓘ Content from **IBM Power Systems**

Is Data The New Oil? How One Startup Is Rescuing The World's Most Valuable Asset



"the new natural resource"

Data as a resource

The world's most valu
is no longe

Sections ☰ The Wa

wp BrandStudio ⓘ Conten

The data economy d

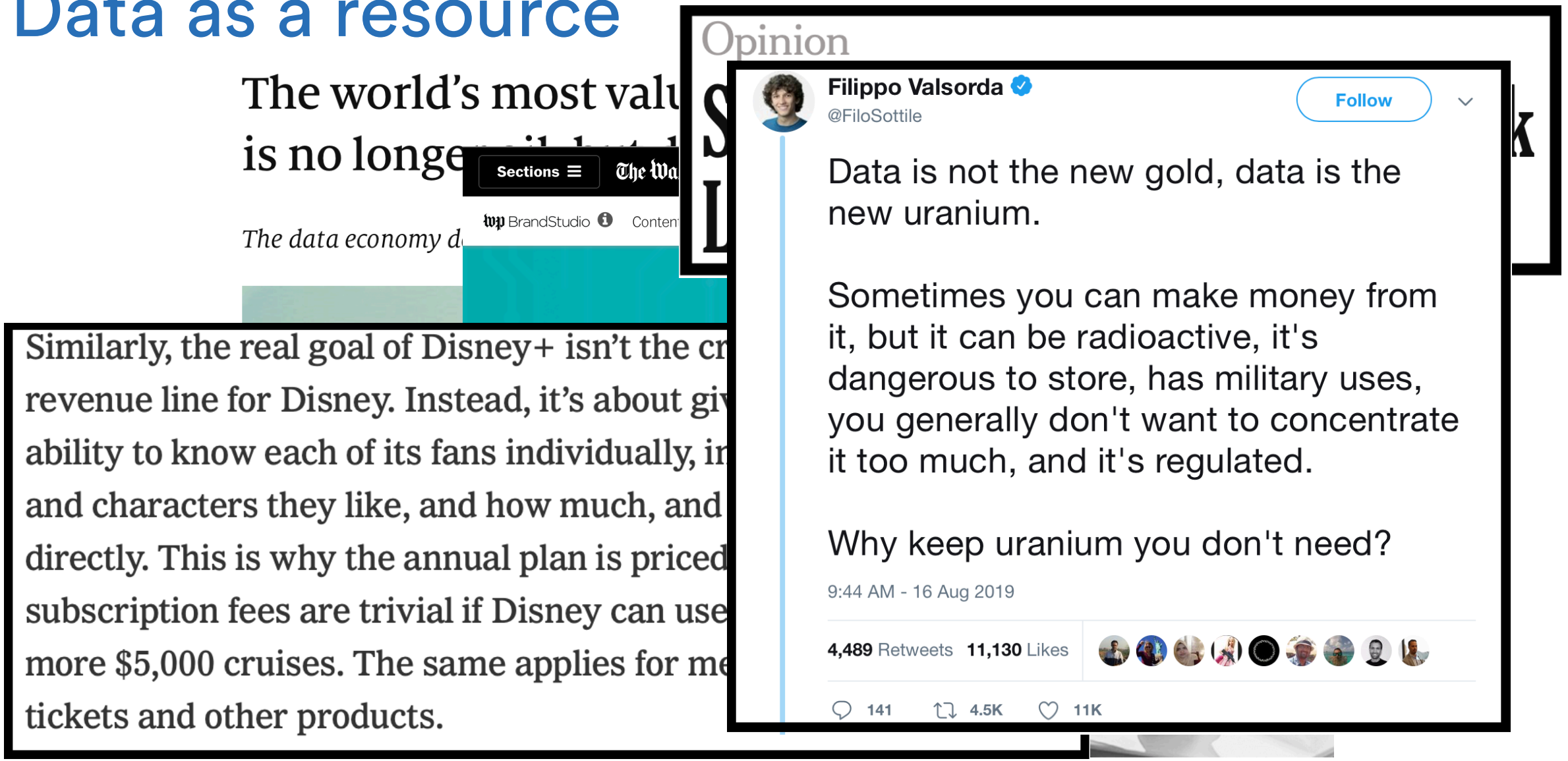
Opinion

Streaming Video Will Soon Look Like the Bad Old Days of TV

escuing The

Similarly, the real goal of Disney+ isn't the creation of a new revenue line for Disney. Instead, it's about giving the company the ability to know each of its fans individually, including what content and characters they like, and how much, and to sell to them directly. This is why the annual plan is priced at only $70. Monthly subscription fees are trivial if Disney can use the service to sell more $5,000 cruises. The same applies for merchandise, movie tickets and other products.

Data as a resource

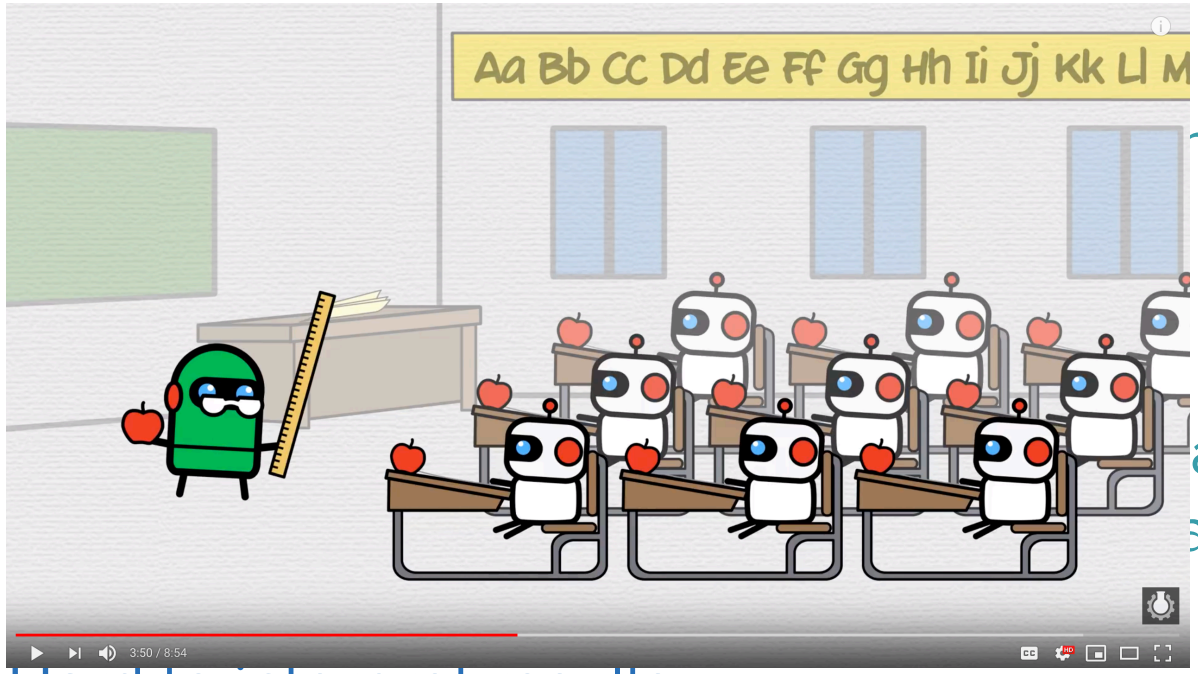


AI and deep learning

- Not a “magic bullet”
  - Predictions can be very different from the truth, even when advanced techniques are used
- Often requires massive training databases
  - Results are only as good as training data
  - Results also depend on what methods are trained to optimize
- Hard to interpret results…

Stills from “How machines learn” by CGP Grey

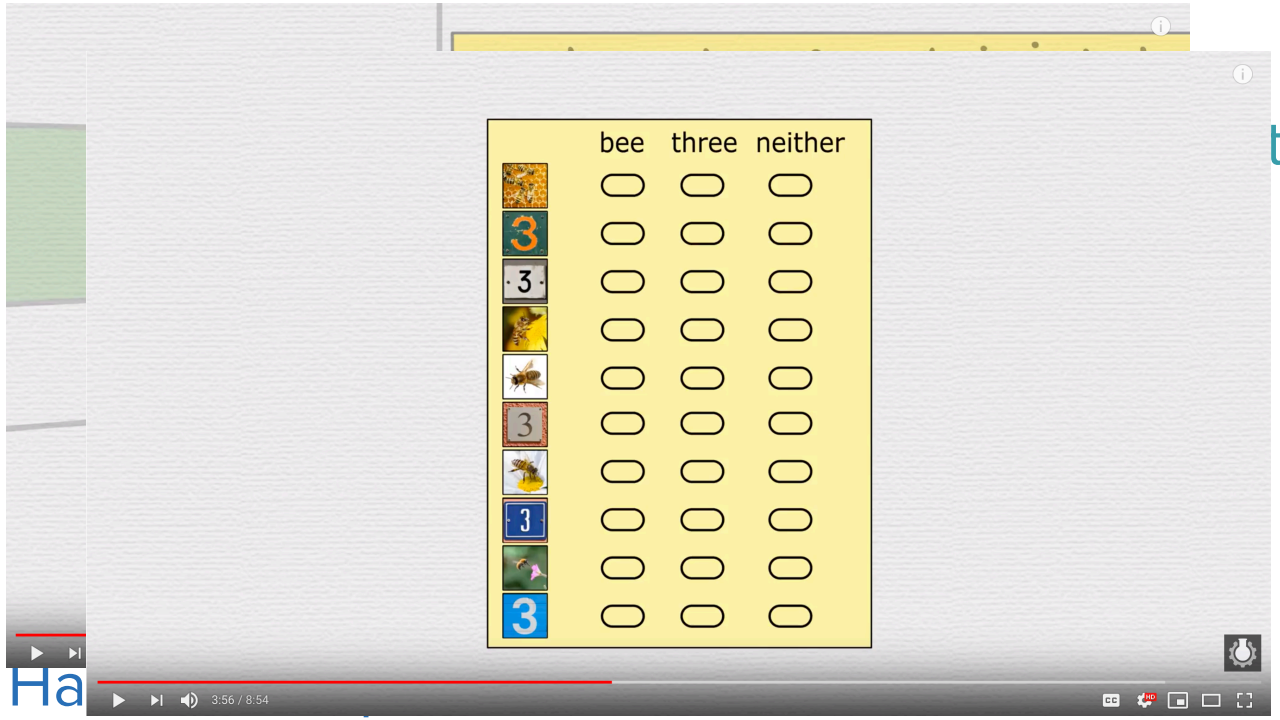
AI and deep learning

- 

ne truth, even when advanced
- a
are trained to optimize
- Hard to interpret results…

Stills from "How machines learn" by CGP Grey

AI and deep learning

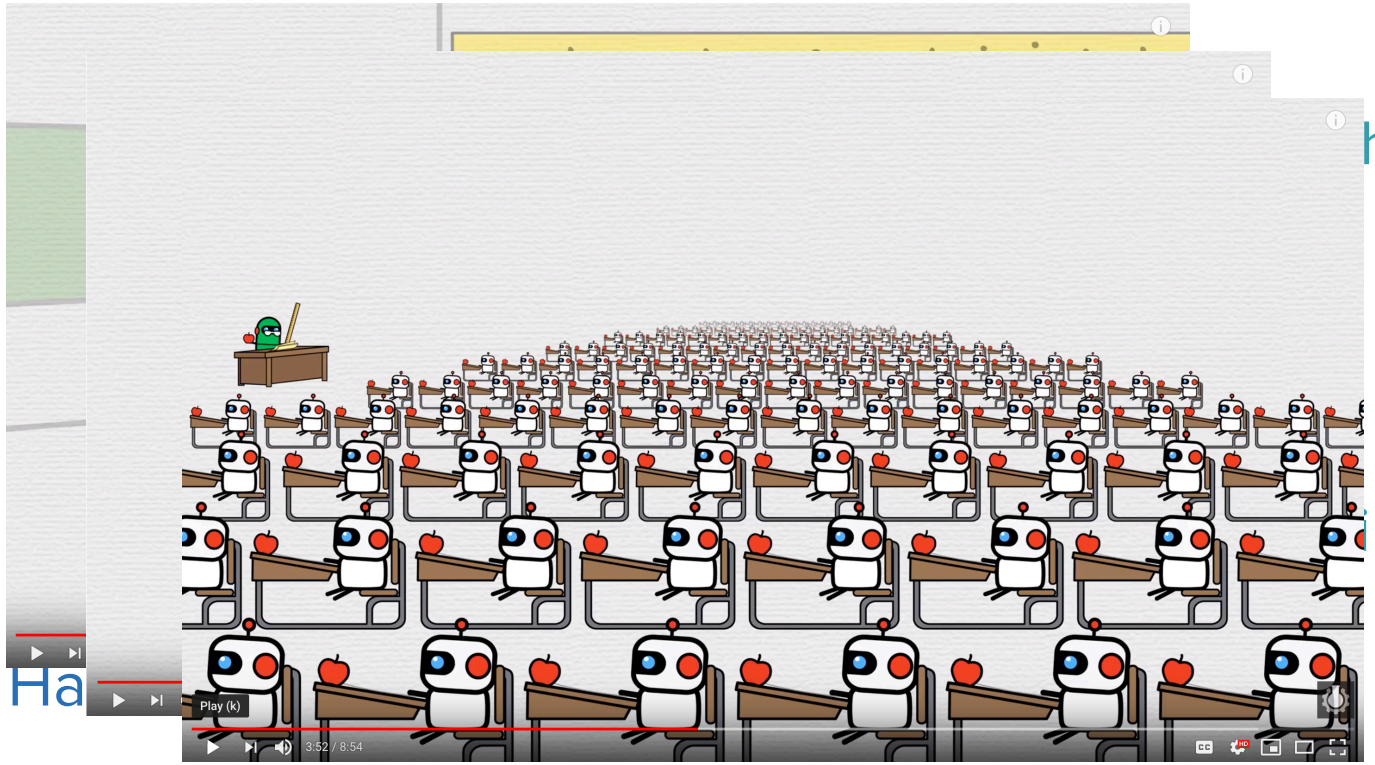
- 

truth, even when advanced
- trained to optimize
- Ha

Stills from "How machines learn" by CGP Grey

AI and deep learning

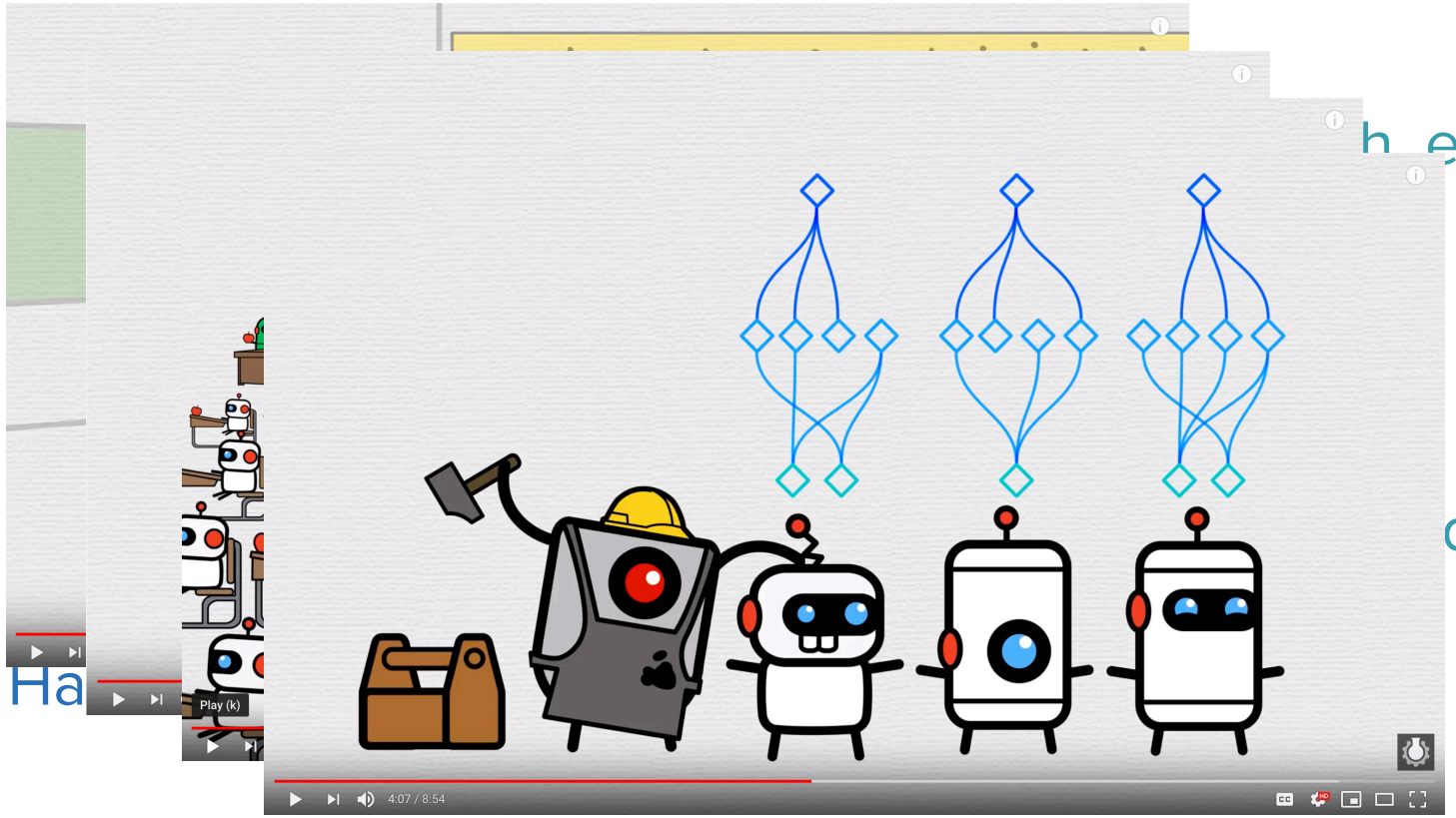
- h, even when advanced
- ined to optimize
- Ha



Stills from "How machines learn" by CGP Grey

AI and deep learning

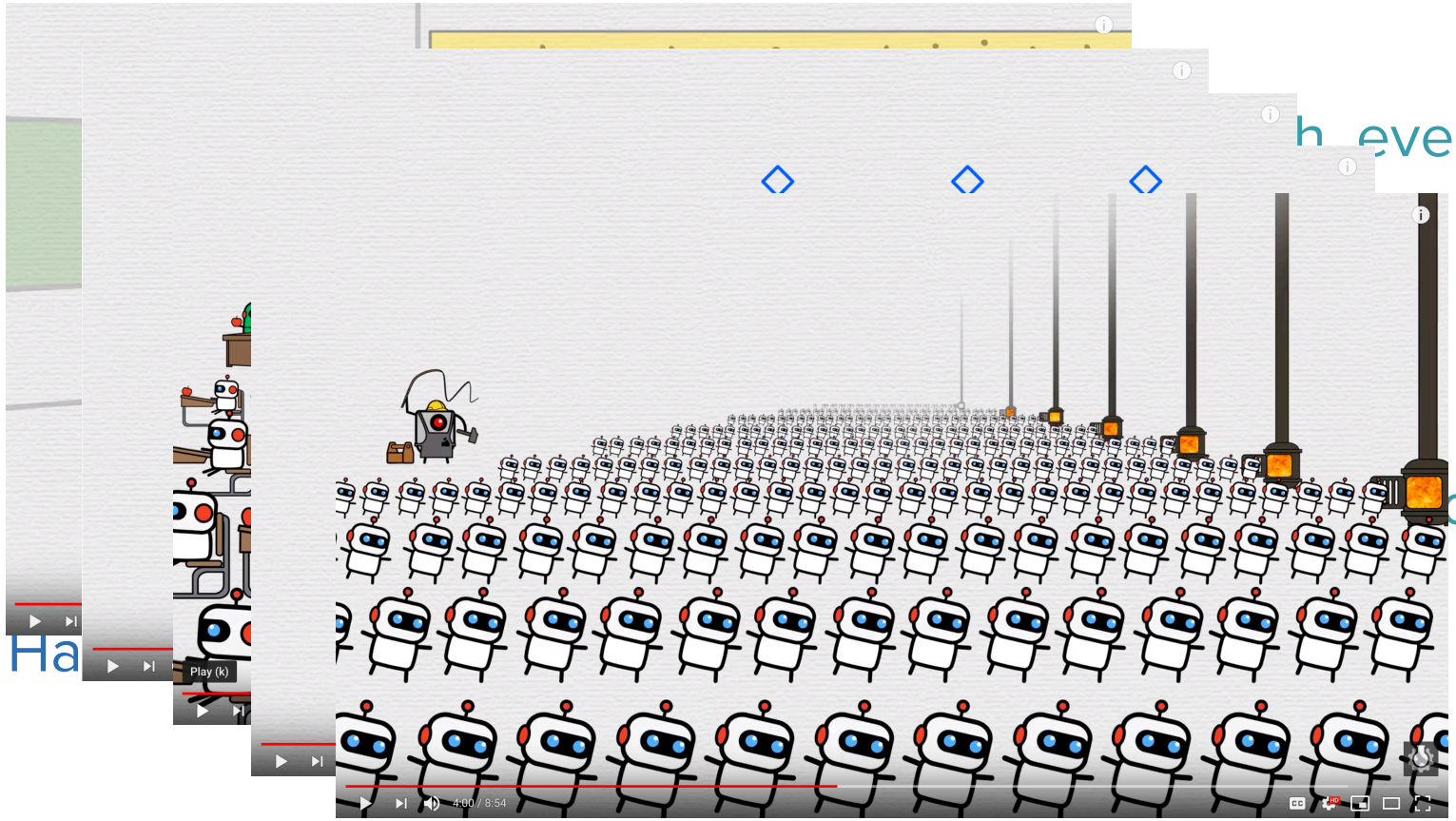
- h even when advanced
- d to optimize
- Ha



Stills from "How machines learn" by CGP Grey

AI and deep learning

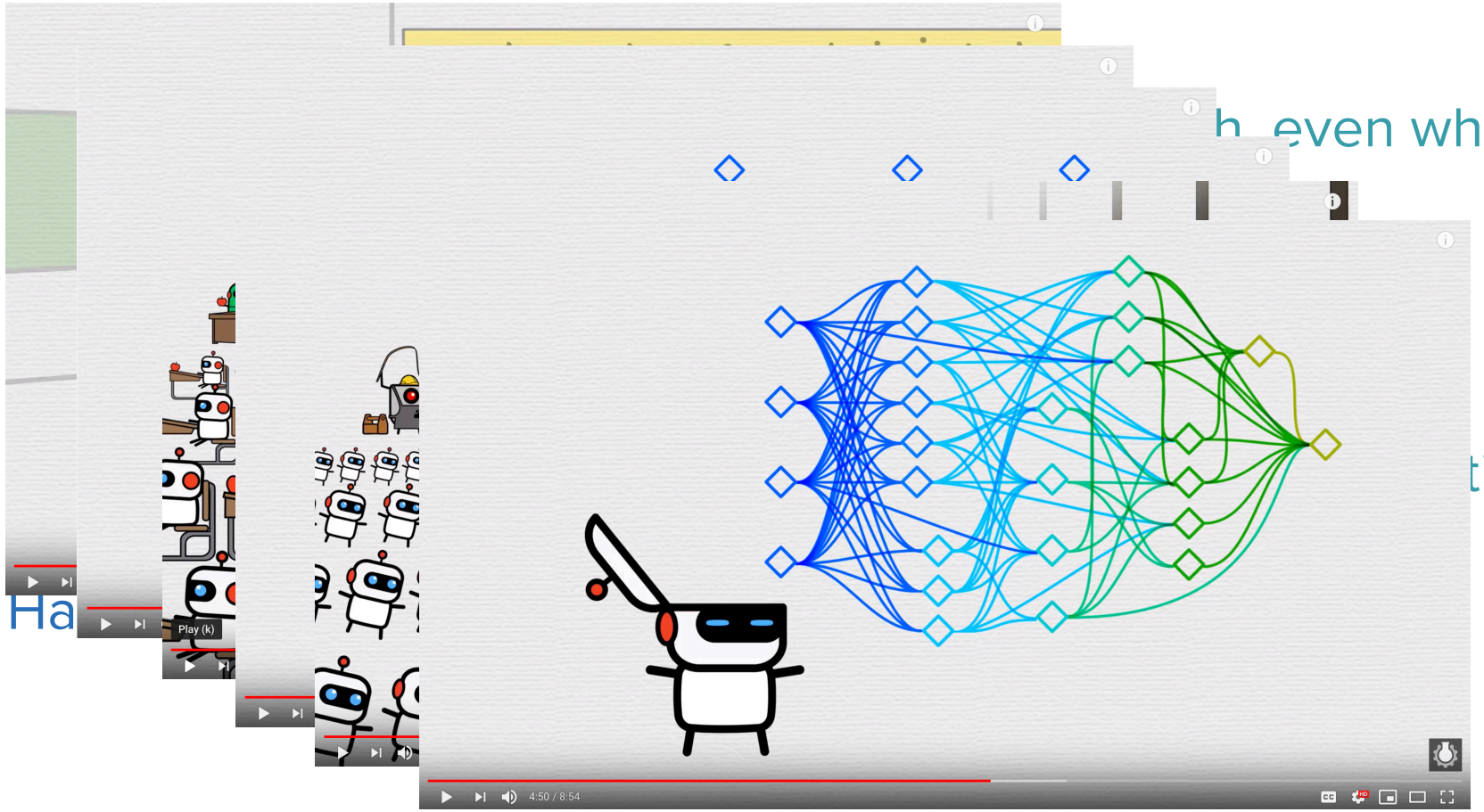
- h even when advanced
- o optimize
- Ha



Stills from "How machines learn" by CGP Grey

AI and deep learning

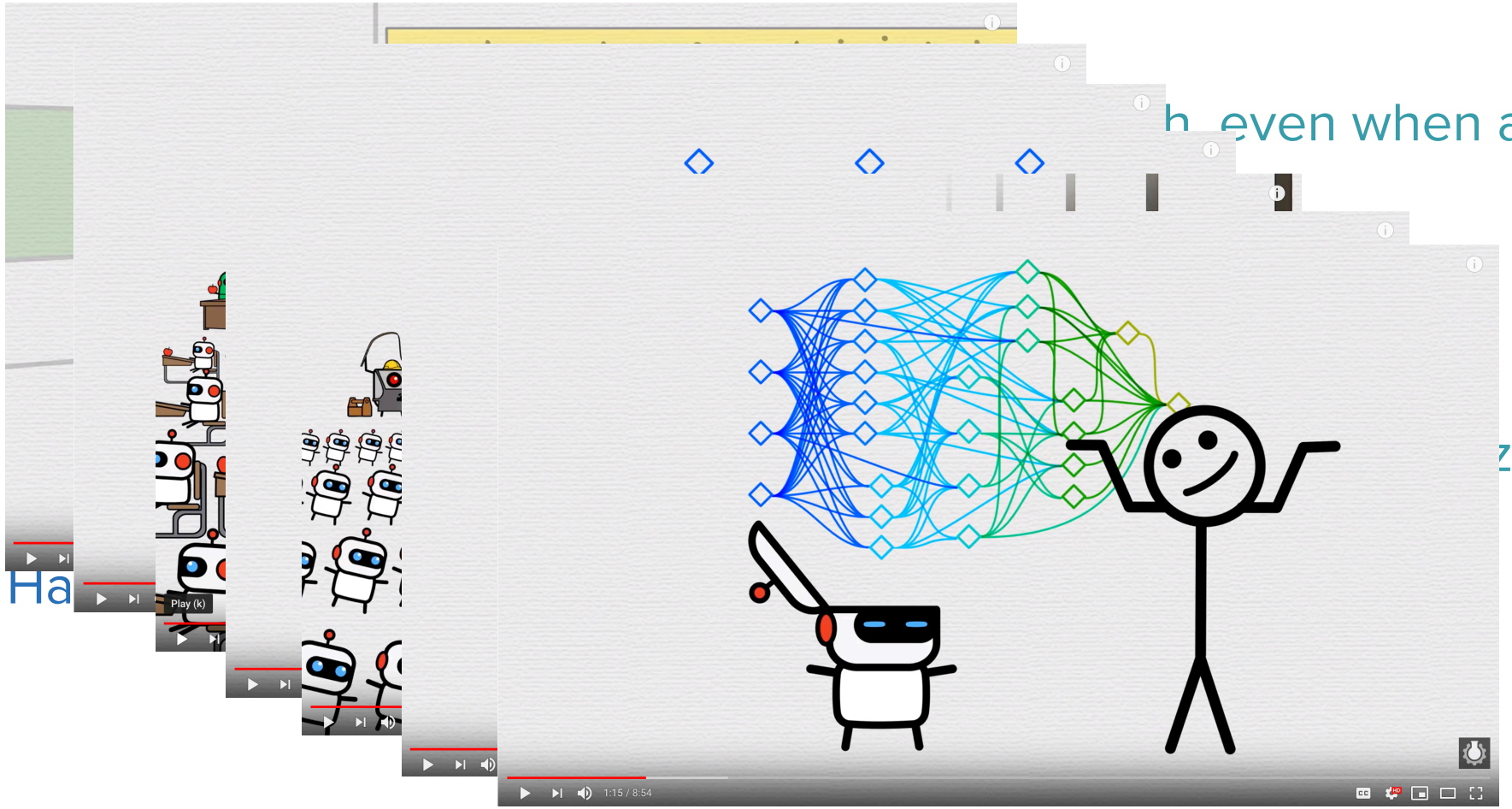
- h even when advanced
- timize
- Ha



Stills from "How machines learn" by CGP Grey

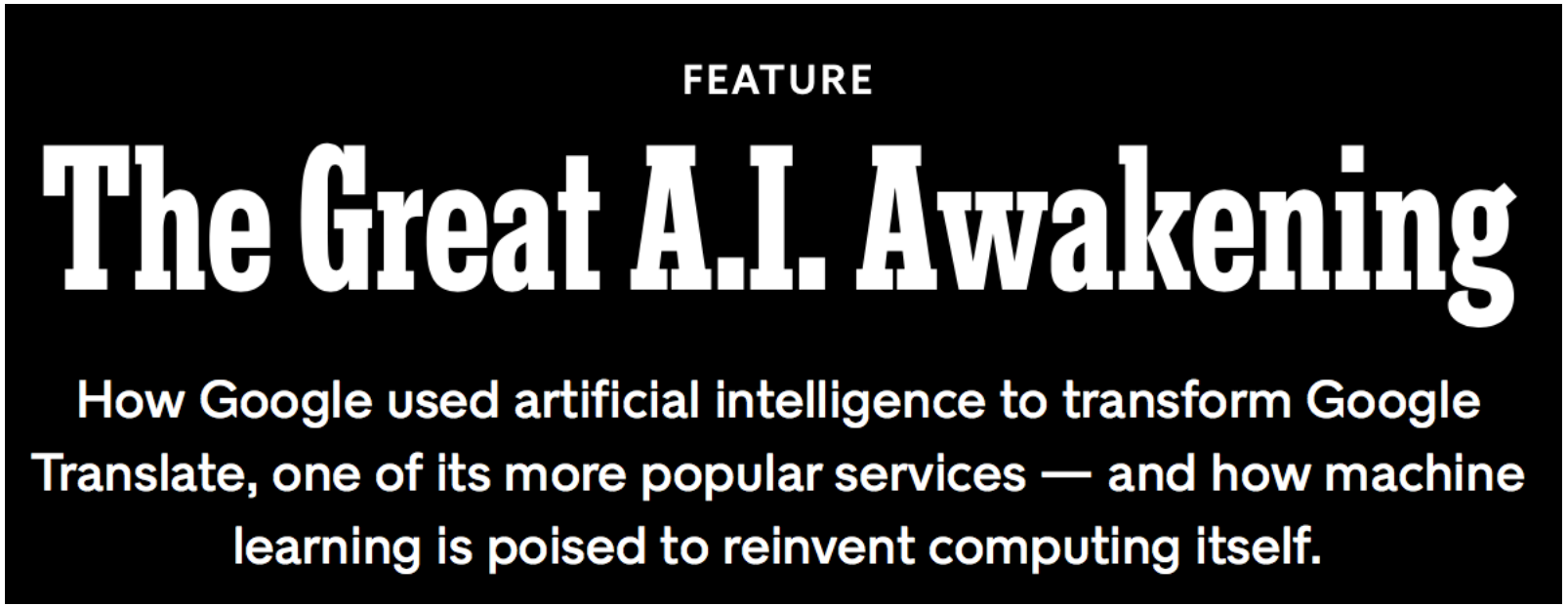
AI and deep learning

- h even when advanced
- ze
- Ha

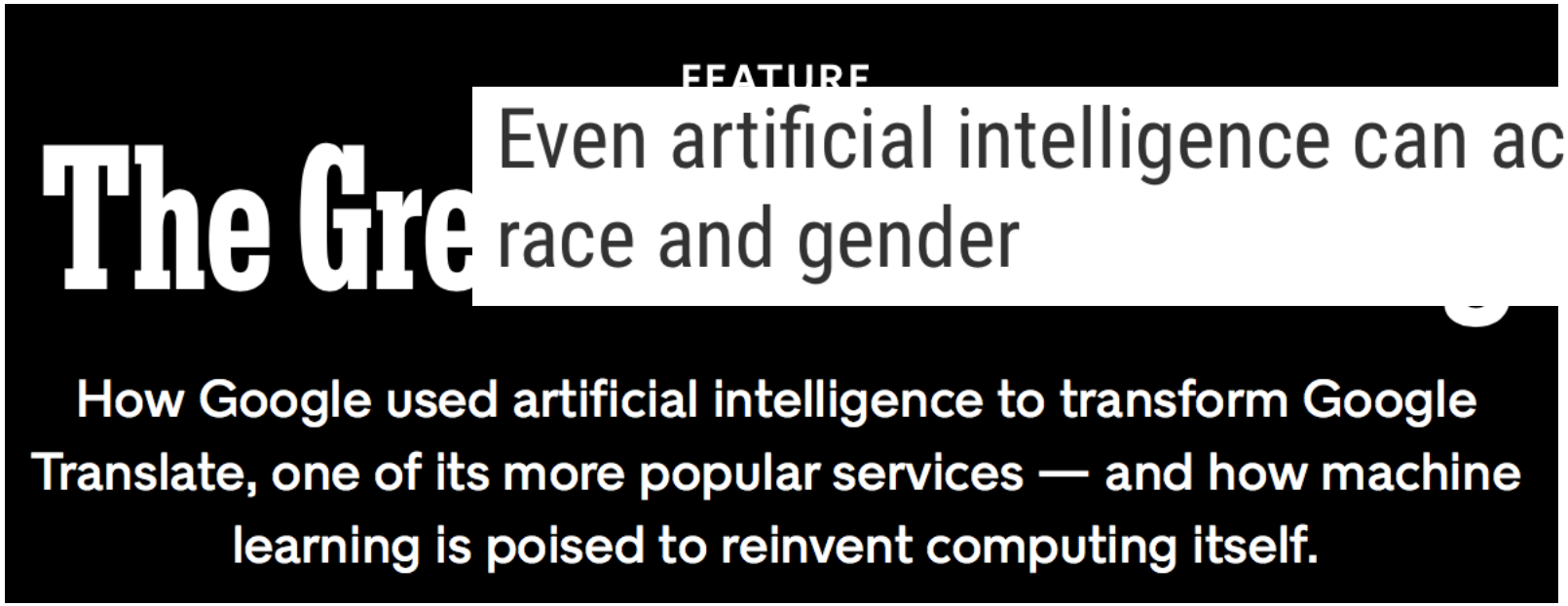


Stills from "How machines learn" by CGP Grey

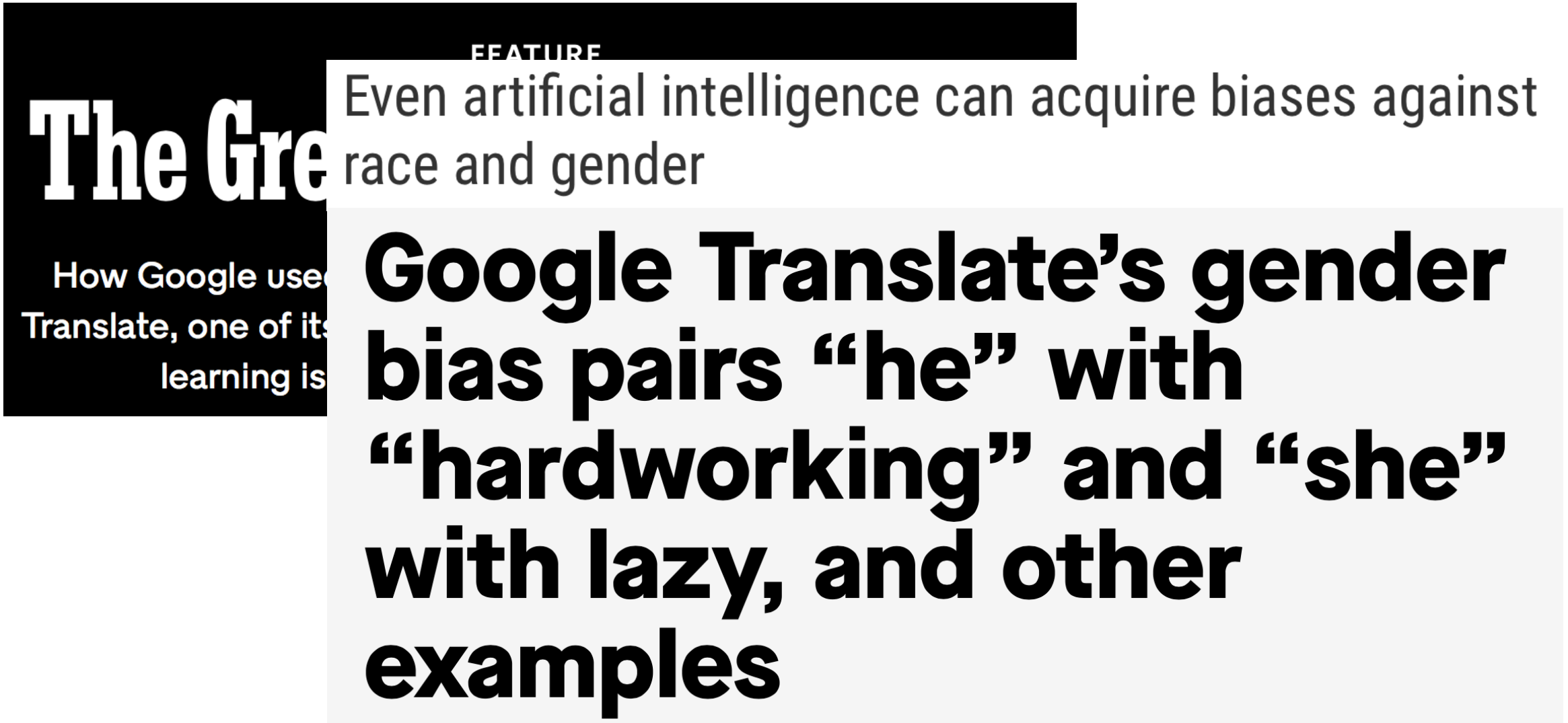
AI for Translation



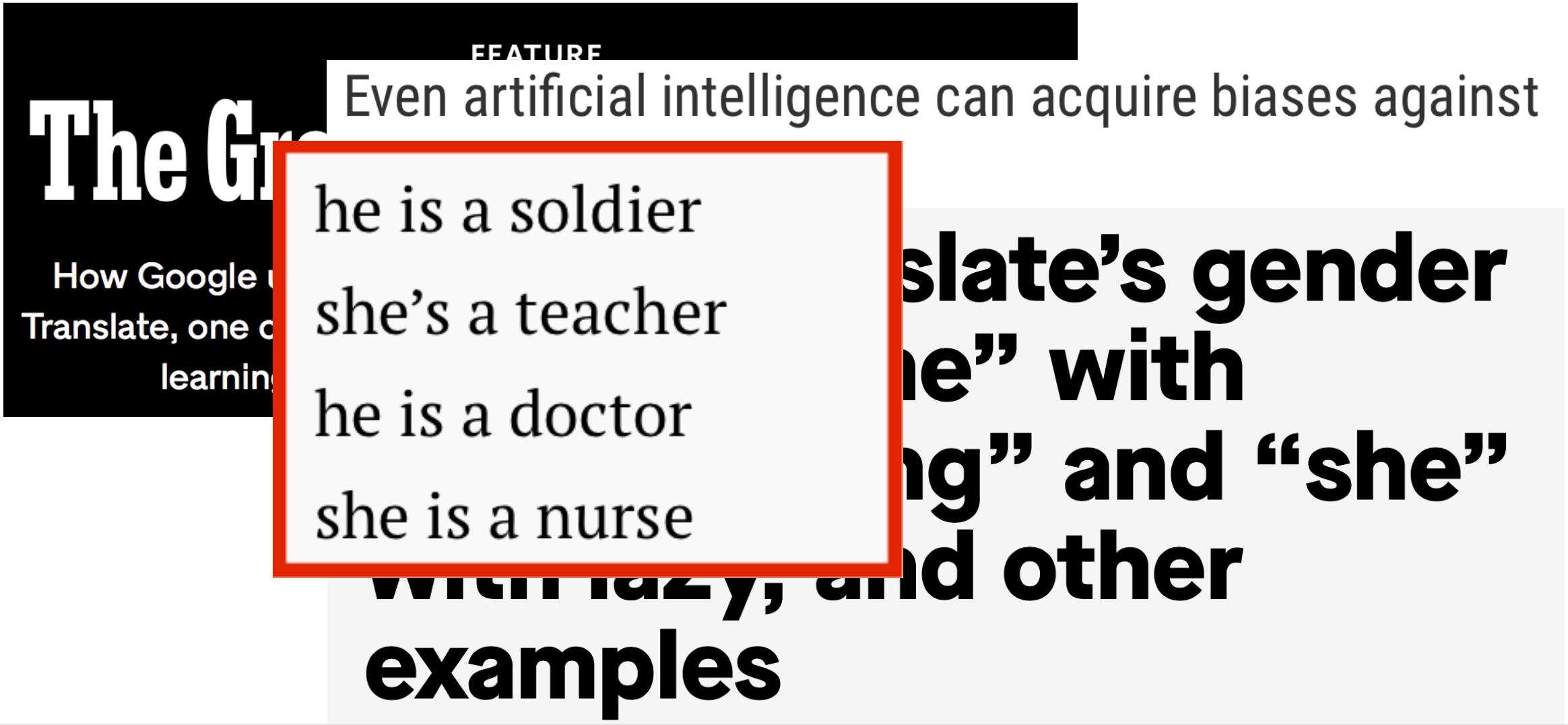
AI for Translation



AI for Translation



AI for Translation



Real talk about AI

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

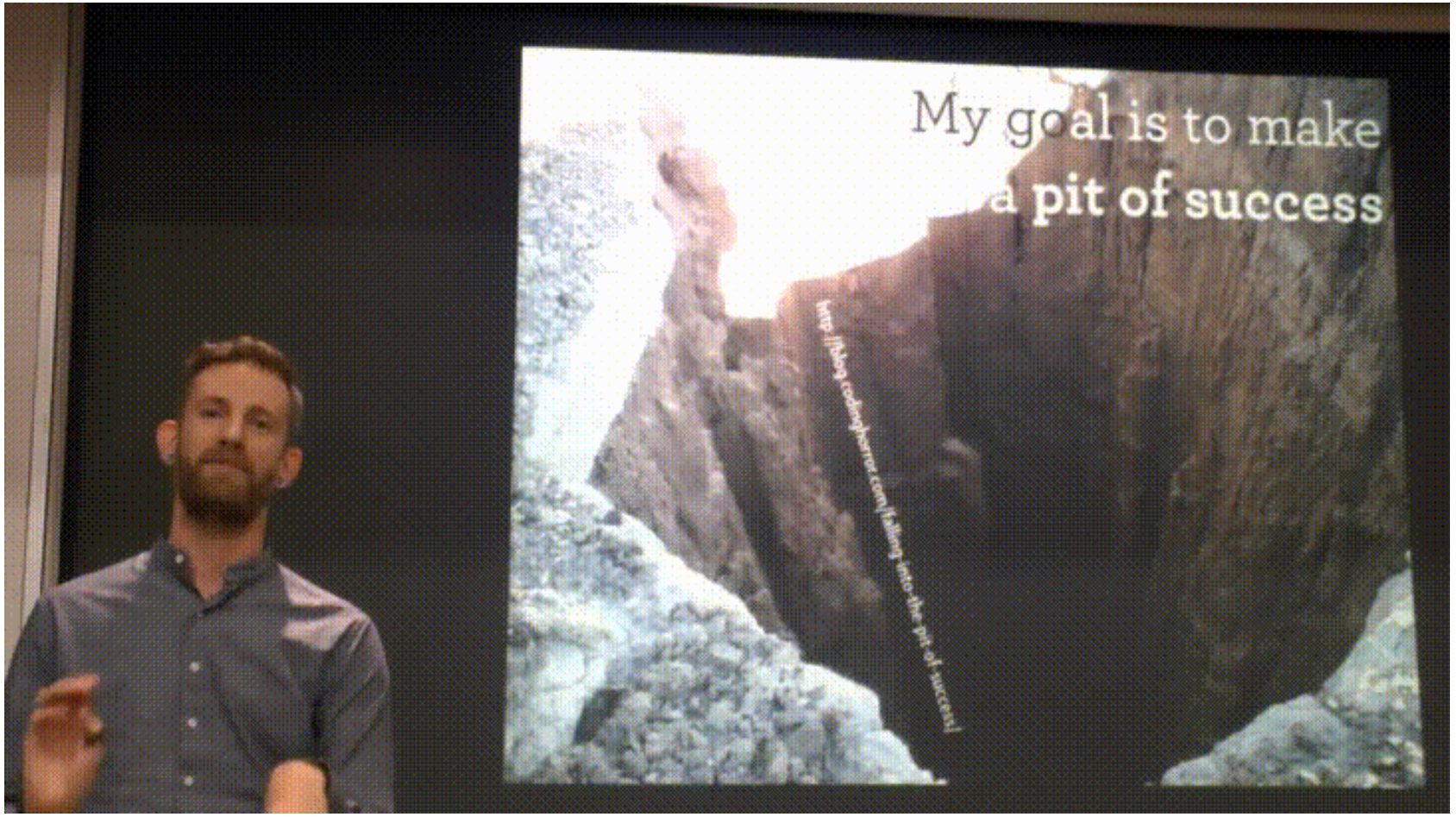
Sharing code

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

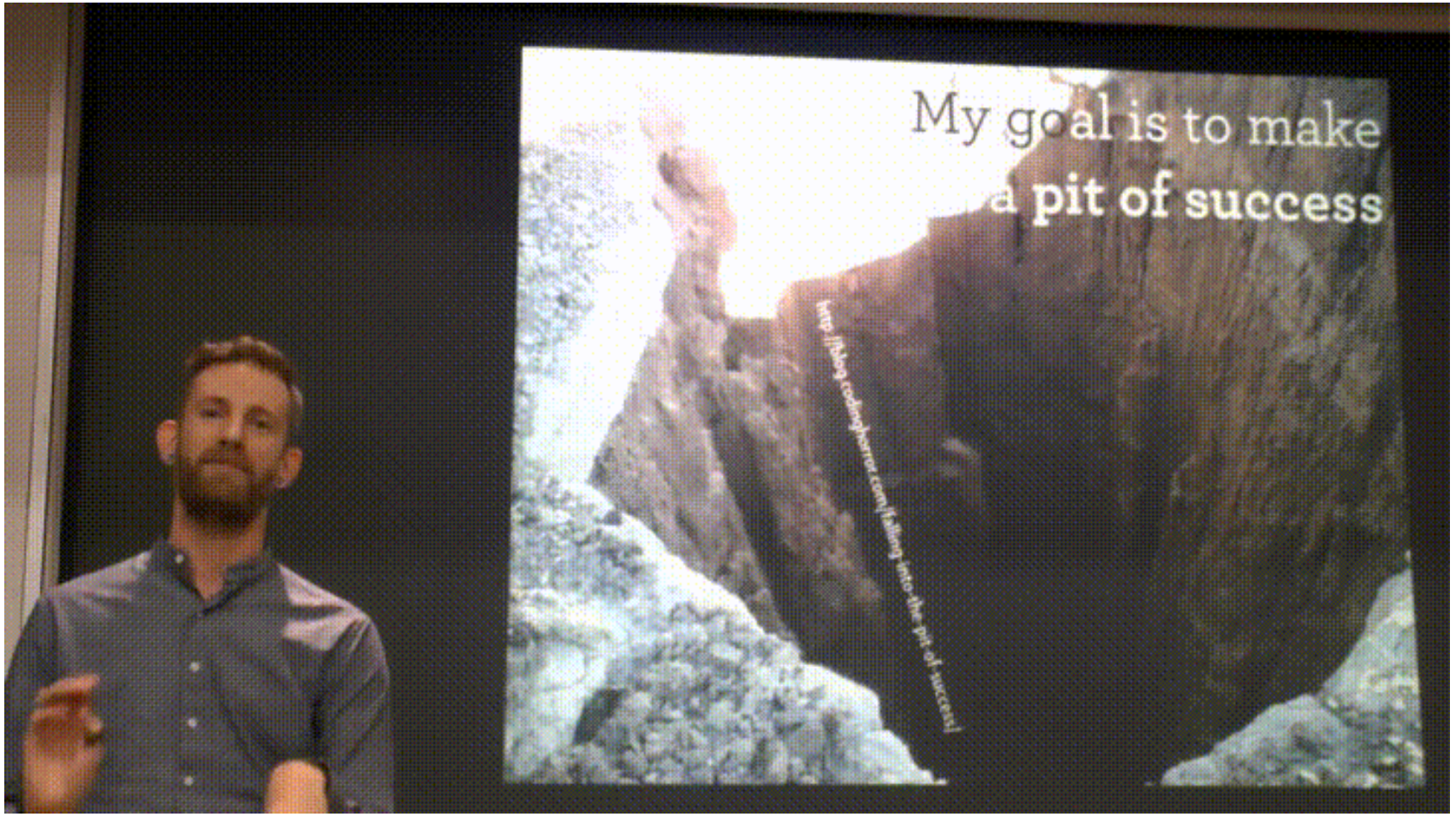
- Try to think of sharing code as a gesture of confidence and humility
  - You've done your best, and you should feel good about that
  - Everyone makes mistakes sometimes; when you do, that's fine – fix it and move on

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

Choosing data science tools



Choosing data science tools



A public health lens

How can we use these data to improve health?

- Improve surveillance, leading to better prevention efforts?
- Better understanding of mechanisms?
- More precise and more effective outreach?

- Doing something simple and useful is better

Public health data science

- Public health is great training for data science
  - Study design / sampling
  - Confounding
  - Importance of effects
  - Association vs causation
  - Ethical considerations

- Add some tools to that, and you're set

Public health data science

- Public health is great training for data science
  - Study design / sampling
  - Confounding
  - Importance of effects
  - Association vs causation
  - Ethical considerations

- Add some tools to that, and you're set