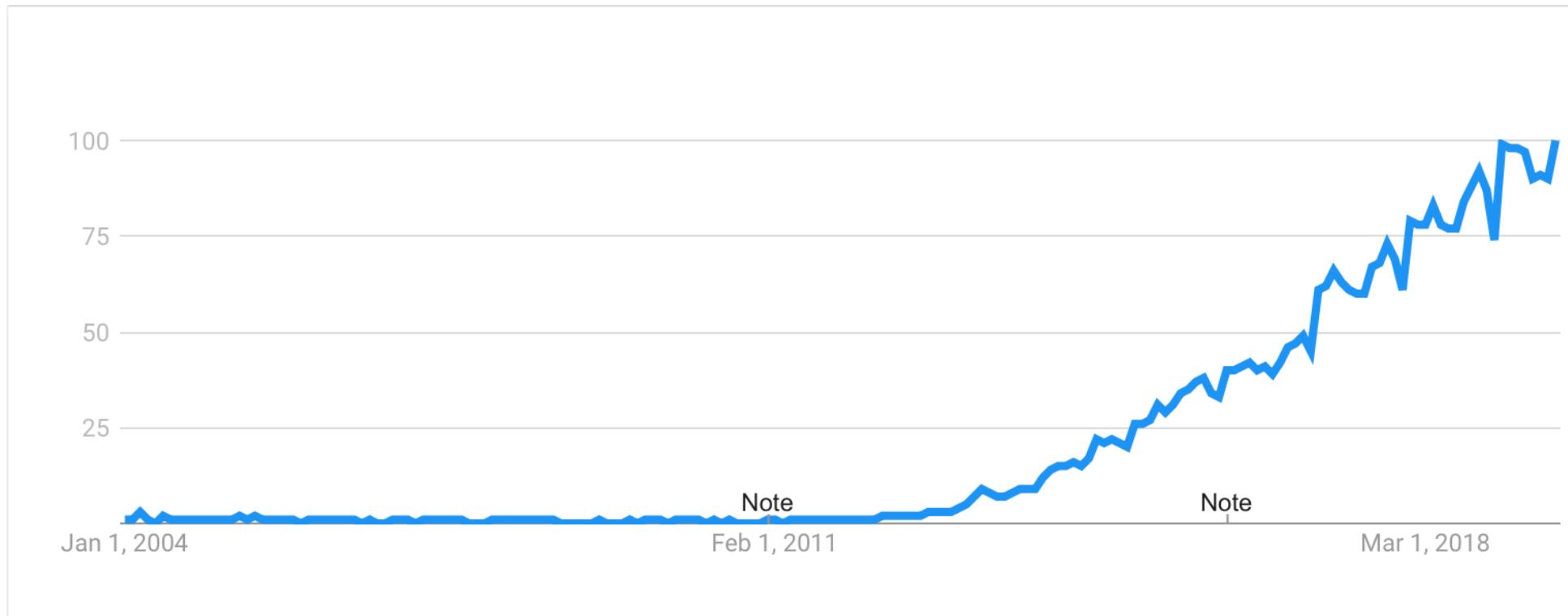


WHAT IS DATA SCIENCE?

Jeff Goldsmith, PhD

Department of Biostatistics

Data science is pretty new



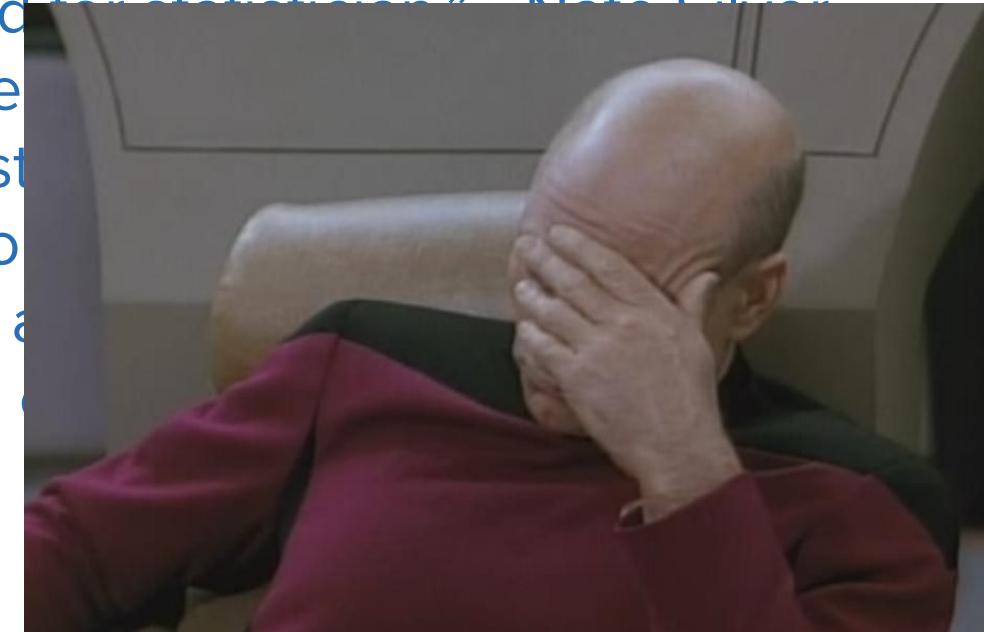
Source: Google Trends

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 statistician.” — Nate Silver
- “A data scientist is a better computer statistician than a computer scientist”
- “A data scientist is a statistician who can code”
- A data scientist is a good statistical analyst who can program
- A data scientist is a statistician who can program



Maybe pictures will help?

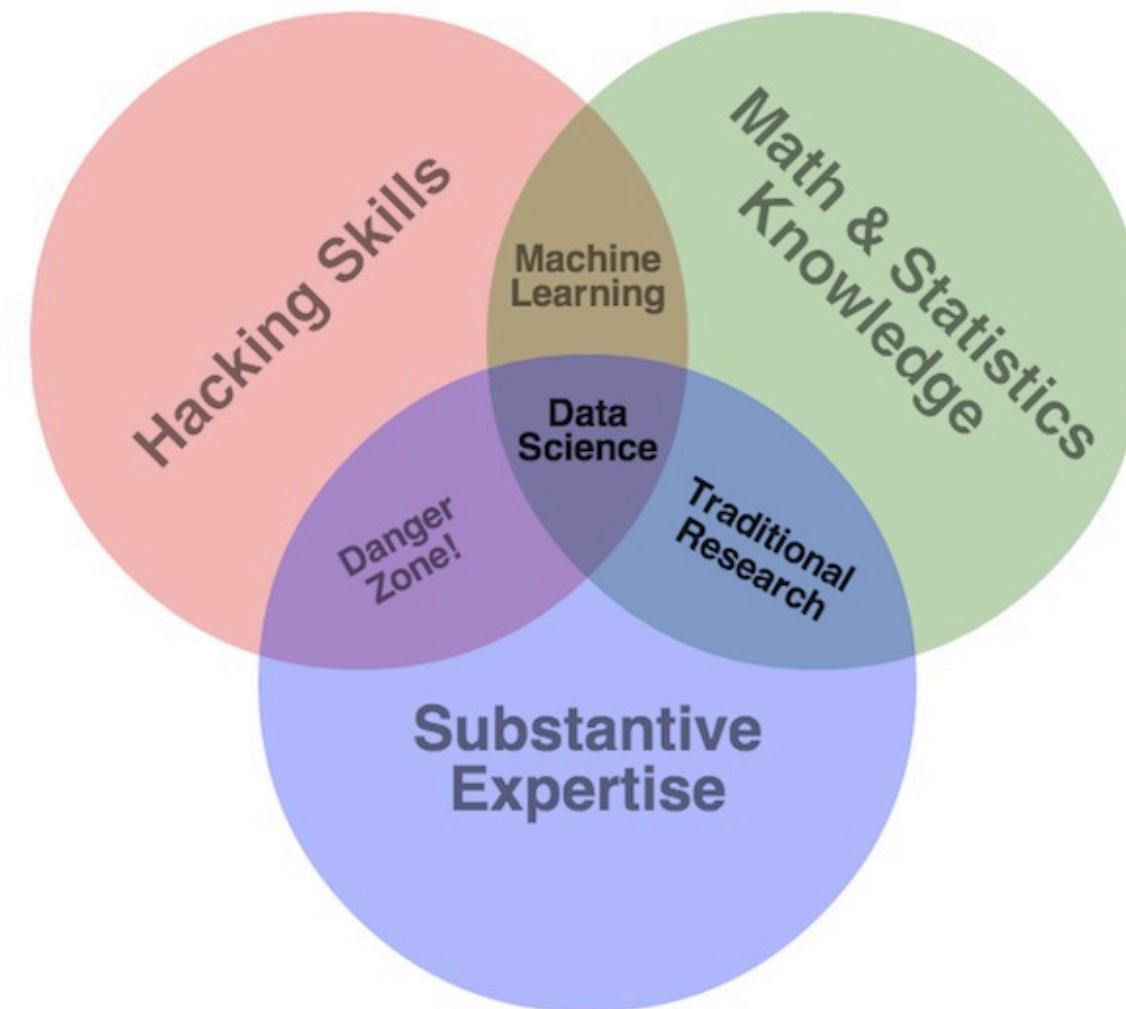
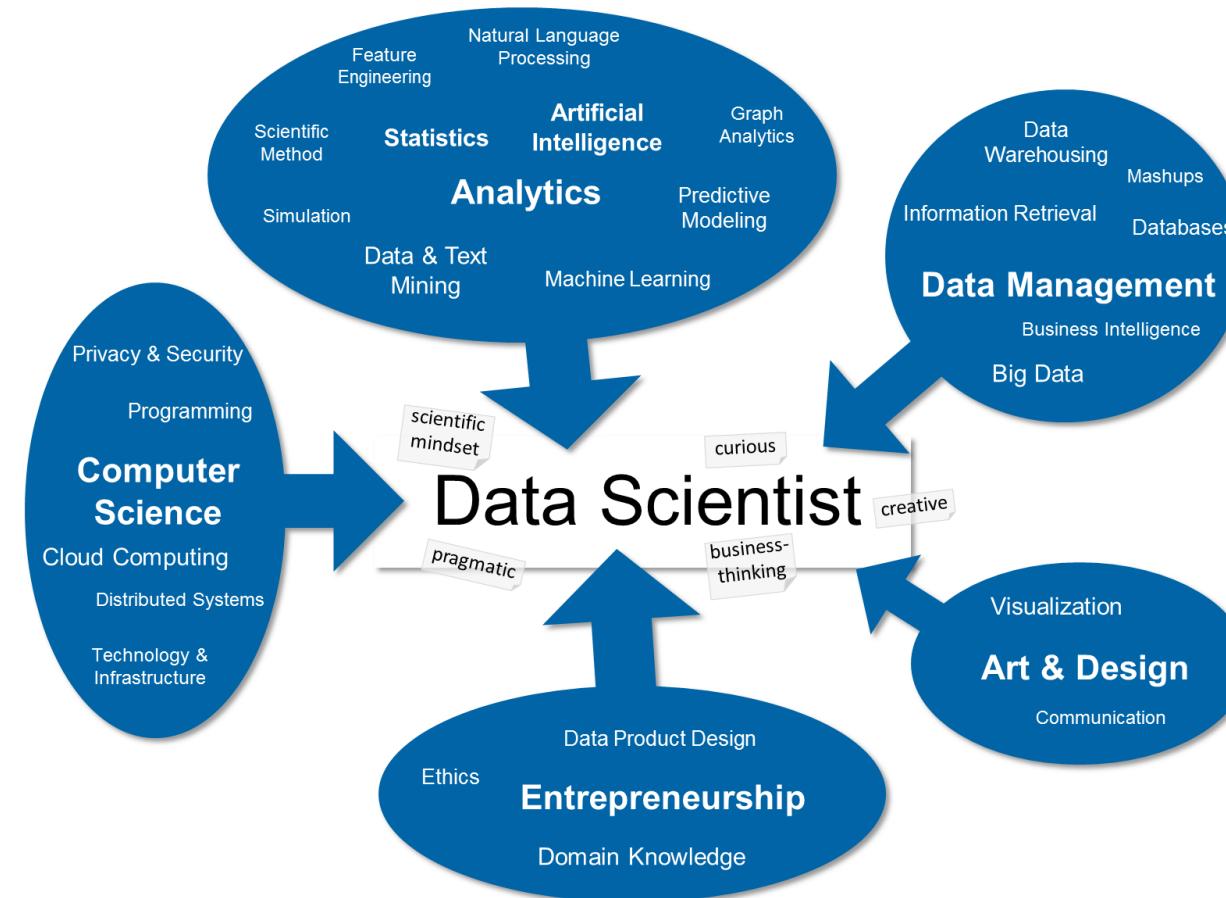


Image from Drew Conway

Maybe pictures will help?



Maybe pictures will help?

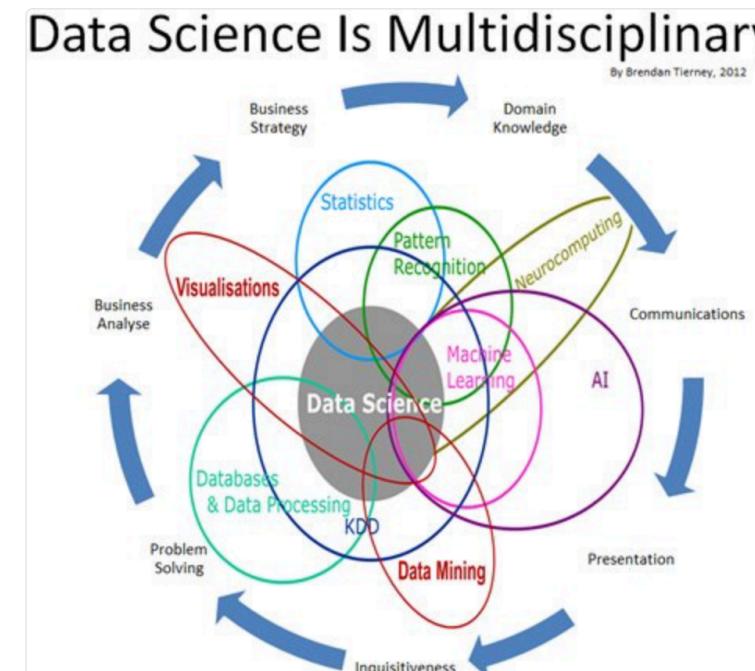


David Robinson

@drob

Follow

I'm going to blame [@drewconway](#) for this



RETWEETS LIKES
20 84



4:00 PM - 28 Apr 2017 from Manhattan, NY

From twitter

Maybe pictures will help?

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21th century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.



MATH & STATISTICS <ul style="list-style-type: none"> ★ Machine learning ★ Statistical modeling ★ Experiment design ★ Bayesian inference ★ Supervised learning: decision trees, random forests, logistic regression ★ Unsupervised learning: clustering, dimensionality reduction ★ Optimization: gradient descent and variants 	PROGRAMMING & DATABASE <ul style="list-style-type: none"> ★ Computer science fundamentals ★ Scripting language e.g. Python ★ Statistical computing packages, e.g. R ★ Databases: SQL and NoSQL ★ Relational algebra ★ Parallel databases and parallel query processing ★ MapReduce concepts ★ Hadoop and Hive/Pig ★ Custom reducers ★ Experience with xaaS like AWS
DOMAIN KNOWLEDGE & SOFT SKILLS <ul style="list-style-type: none"> ★ Passionate about the business ★ Curious about data ★ Influence without authority ★ Hacker mindset ★ Problem solver ★ Strategic, proactive, creative, innovative and collaborative 	COMMUNICATION & VISUALIZATION <ul style="list-style-type: none"> ★ Able to engage with senior management ★ Story telling skills ★ Translate data-driven insights into decisions and actions ★ Visual art design ★ R packages like ggplot or lattice ★ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization: customer tracking and on-site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Marketing DISTILLERY
(c) Krzysztof Zawadzki

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.



MATH & STATISTICS <ul style="list-style-type: none"> ★ Machine learning ★ Statistical modeling ★ Experiment design ★ Bayesian inference ★ Supervised learning: decision trees, random forests, logistic regression ★ Unsupervised learning: clustering, dimensionality reduction ★ Optimization: gradient descent and variants 	PROGRAMMING & DATABASE <ul style="list-style-type: none"> ★ Computer science fundamentals ★ Scripting language e.g. Python ★ Statistical computing package e.g. R ★ Databases SQL and NoSQL ★ Relational algebra ★ Parallel databases and parallel query processing ★ MapReduce concepts ★ Hadoop and Hive/Pig ★ Custom reducers ★ Experience with xaaS like AWS
DOMAIN KNOWLEDGE & SOFT SKILLS <ul style="list-style-type: none"> ★ Passionate about the business ★ Curious about data ★ Influence without authority ★ Hacker mindset ★ Problem solver ★ Strategic, proactive, creative, innovative and collaborative 	COMMUNICATION & VISUALIZATION <ul style="list-style-type: none"> ★ Able to engage with senior management ★ Story telling skills ★ Translate data-driven insights into decisions and actions ★ Visual art design ★ R packages like ggplot or lattice ★ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization: customer tracking and on-site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Marketing DISTILLERY
(c) Krzysztof Zawadzki

Why these definitions are bad

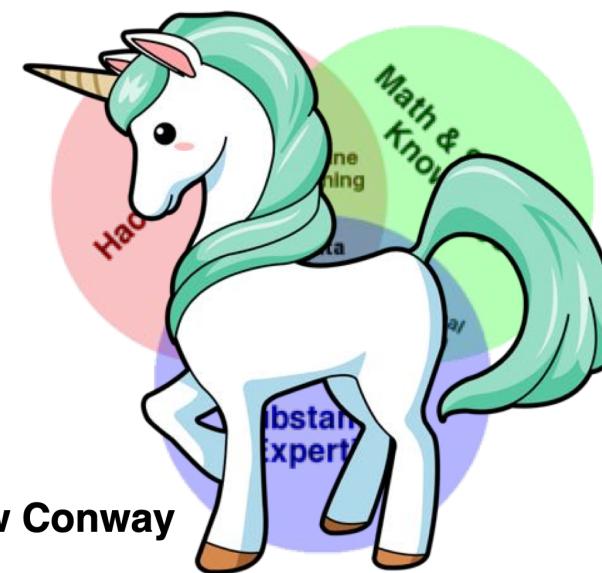
- “Data science is just ...” definitions miss the point
 - If data science is just statistics (or machine learning, or computer science, or engineering) we wouldn’t need a new term, let alone a new discipline
 - The popularity of “data science” suggests that there’s a newly recognized need
- “A data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



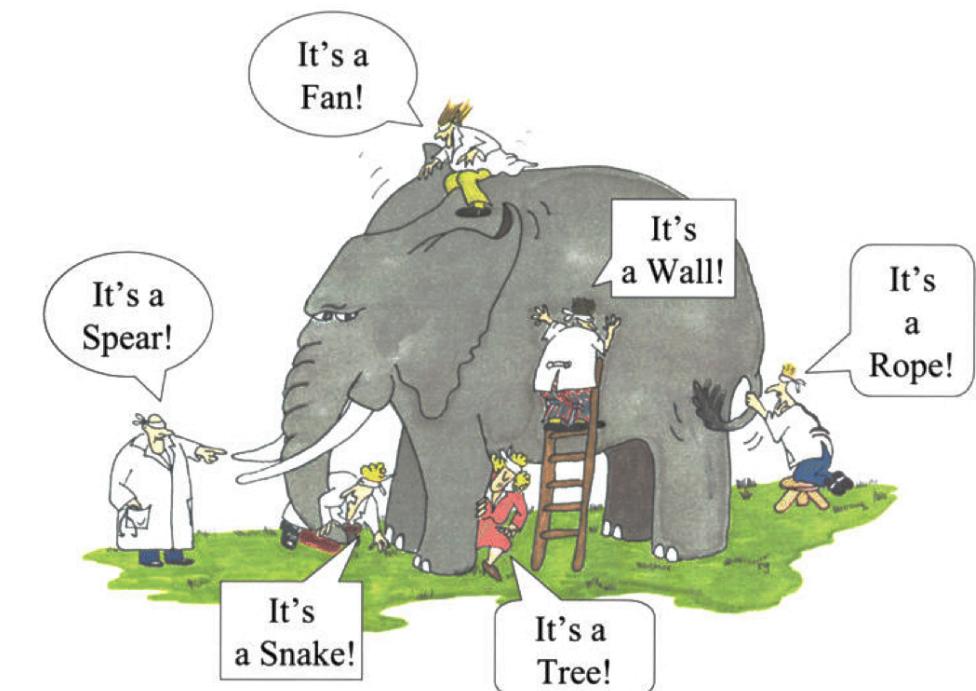
17

@angebassa

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

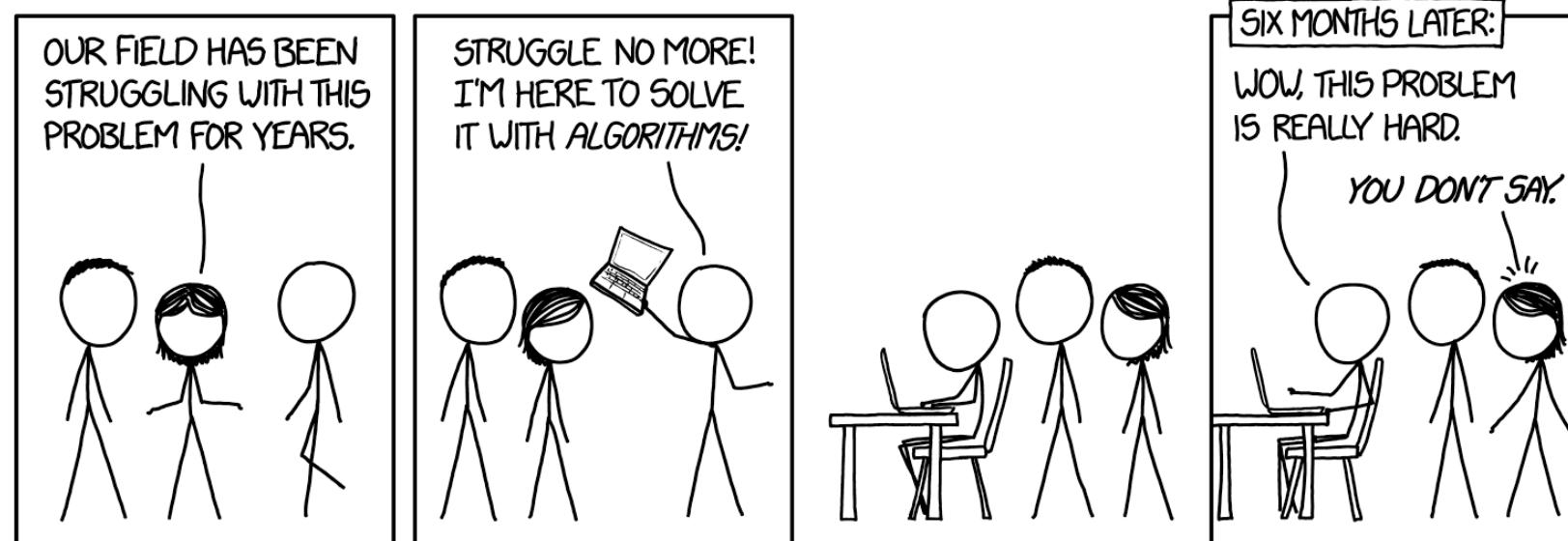
Why these definitions are good

- 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



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

Is “data science” a buzzword?

- It is used to describe a frustratingly wide collection of ideas
- It is also used in different (sometimes contradictory) ways by different people

- Nonetheless, the popularity of “data science” reflects an increasingly data-centric reality
- Dismissing “data science” ignores this reality, and blinds people to the usefulness of a new perspective across disciplines

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
 - Openness is valuable – identify errors early and fix them quickly
 - Code is an important means of communication
 - New tools encourage reproducibility, but the concept is not platform-dependent

Replication

- Reproducibility is one part of replication ...

Replication

- Reproducibility is one part of replication ...

SUBSCRIBE

SCIENTIFIC AMERICAN

English ▾ Cart 0 Sign In | Register

THE SCIENCES MIND HEALTH TECH SUSTAINABILITY EDUCATION VIDEO PODCASTS BLOGS STORE Q

POLICY & ETHICS

Is There a Reproducibility Crisis in Science?

By Nature Video on May 28, 2016

Replication

- Reproducibility is one part of replication ...

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

POLICY & ETHICS

Is There a Reproducibility Crisis in Science?

By Nature Video on May 28, 2016

Replication

- Reproducibility is one part of replication

Essay

Why Most Published Are False

John P. A. Ioannidis

Is There a Re
in

By I



Code



Estimate



Claim



Population



Question



Hypothesis



Experimental Design



Experimentor



Data



Analysis Plan



Analyst

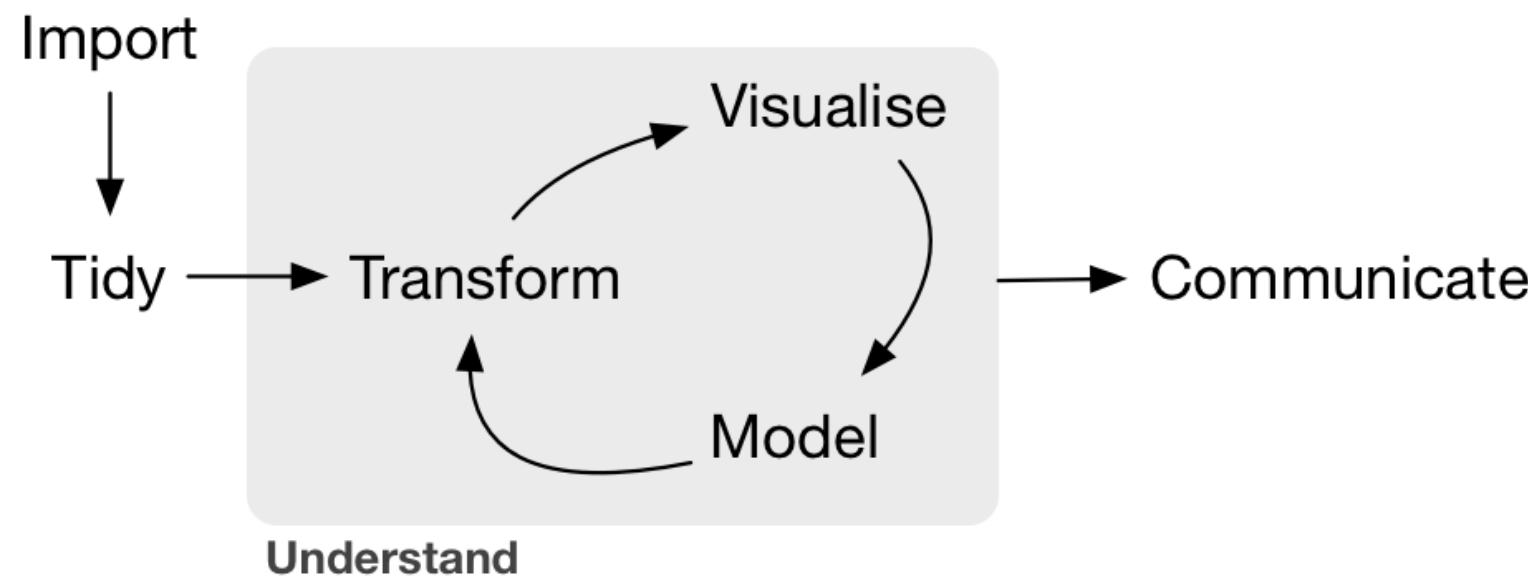
Patil, Peng, and Leek 2016

For the purpose of this class:

Data science is the use of data to formulate and rigorously answer questions in a process that emphasizes clarity, reproducibility, and collaboration, and that recognizes code as a primary means of communication.

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

A data exploration diagram



R for Data Science

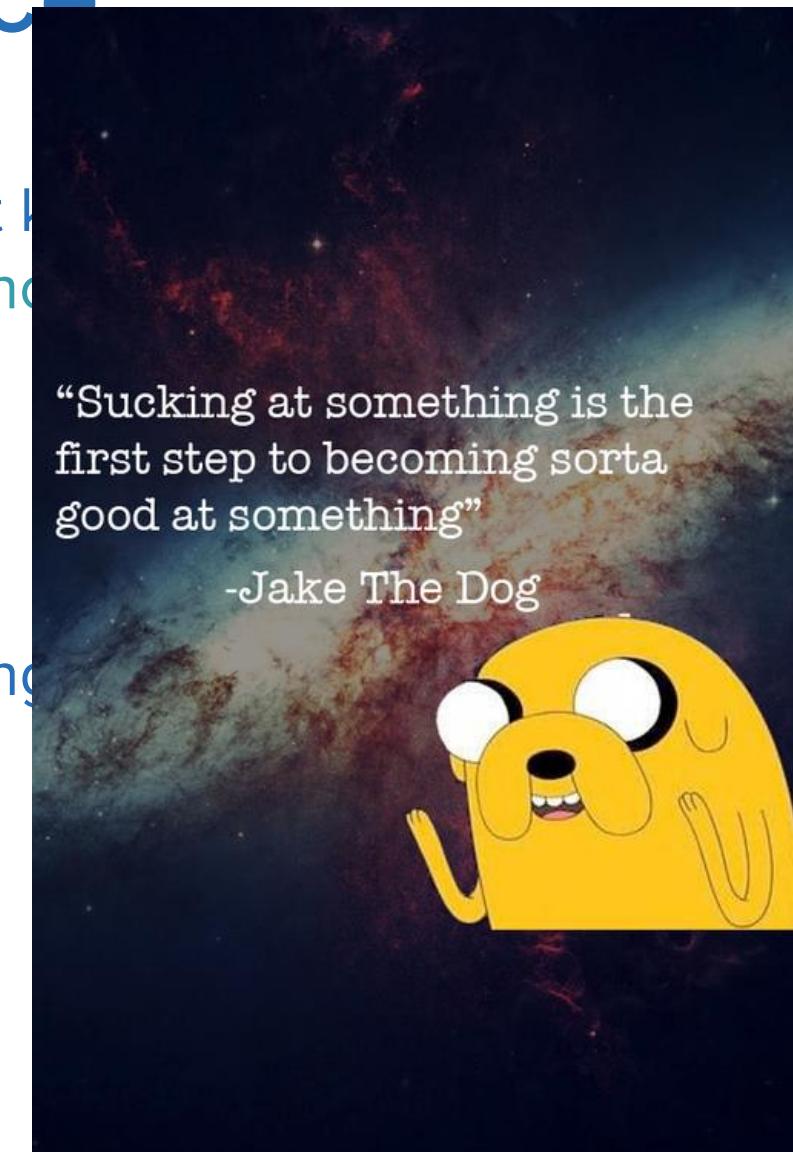
How to learn data science

- 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 know
 - Corollary: don't be a jerk to people who do know things
- Ask questions (well) and keep learning
- Pretty much the same as learning anything else, except they 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, come up with 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

2

10

66

How to learn data science

- Be on the lookout for cool stuff!



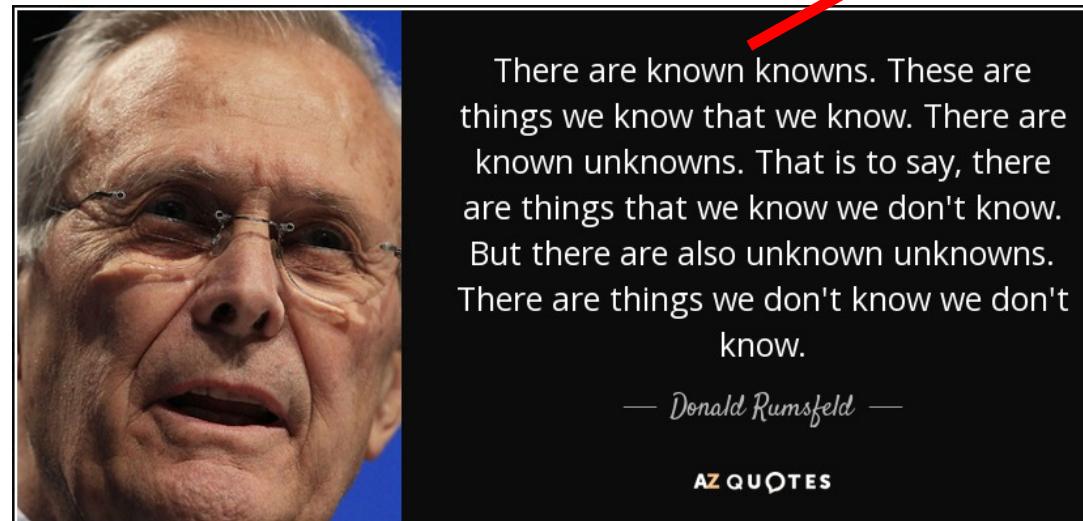
There are known knowns. These are things we know that we know. There are known unknowns. That is to say, there are things that we know we don't know. But there are also unknown unknowns. There are things we don't know we don't know.

— Donald Rumsfeld —

AZ QUOTES

How to learn data science

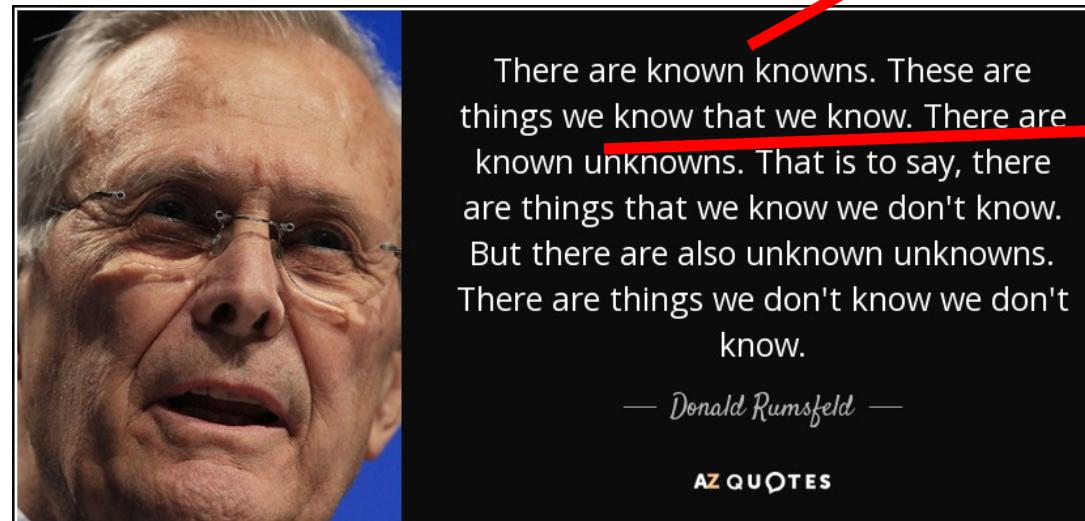
- Be on the lookout for cool stuff!



Knowledge base! :-D

How to learn data science

- Be on the lookout for cool stuff!

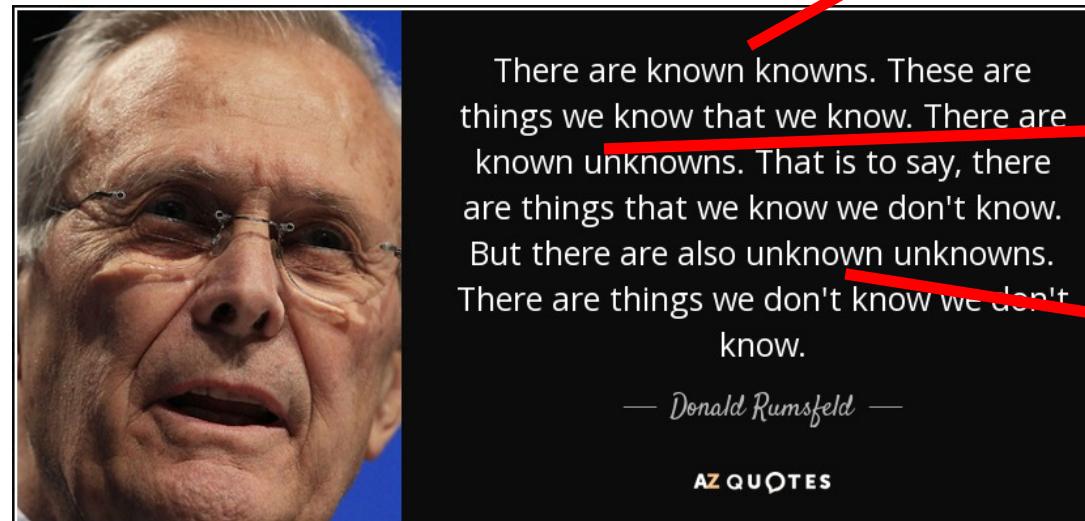


Knowledge base! :-D

Things you know
exist and can
learn how to
do :-)

How to learn data science

- Be on the lookout for cool stuff!



Knowledge base! :-D

Things you know
exist and can
learn how to
do :-)

Things you don't
know exist and
can't use :-(

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 longer oil, but data

 BrandStudio  Content from IBM Power Systems

*The data economy does not have to be...
it can be... it must be... better.*



Why big data is
"the new natural
resource"



Data as a resource

The world's most valuable resource
is no longer oil.

*The data economy is...
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 valuable asset
is no longer oil. It's data.

The data economy depends on

Sections ≡ The Wa

WP BrandStudio Content

Opinion

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

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.

escuing The



Data as a resource

The world's most valuable asset is no longer oil, it's data.

The data economy does not yet have a clear revenue line for Disney. Instead, it's about giving the company the ability to know each of its fans individually, including their names, addresses, ages, genders, and characters they like, and how much, and where they are watching. This is why the annual plan is priced at \$129.99, while individual subscription fees are trivial if Disney can use them to sell more \$5,000 cruises. The same applies for movie tickets and other products.

Opinion

Filippo Valsorda @FiloSottile Follow

Data is not the new gold, data is the new uranium.

Sometimes you can make money from it, but it can be radioactive, it's dangerous to store, has military uses, you generally don't want to concentrate it too much, and it's regulated.

Why keep uranium you don't need?

9:44 AM - 16 Aug 2019

4,489 Retweets 11,130 Likes

141 4.5K 11K

Data in health and medicine

- Data are everywhere
- Clinical trials
- Observational studies
- Genomics
- Medical imaging
- Microbiome

Data in health and medicine

- Data are everywhere
- Clinical trials
- Observational studies
- Genomics
- Medical imaging
- Microbiome
- Electronic health records?
- Mobile health technologies?
- Twitter posts?
- Search terms?
- Social networks?

A public health lens

How can we use these data to improve health?

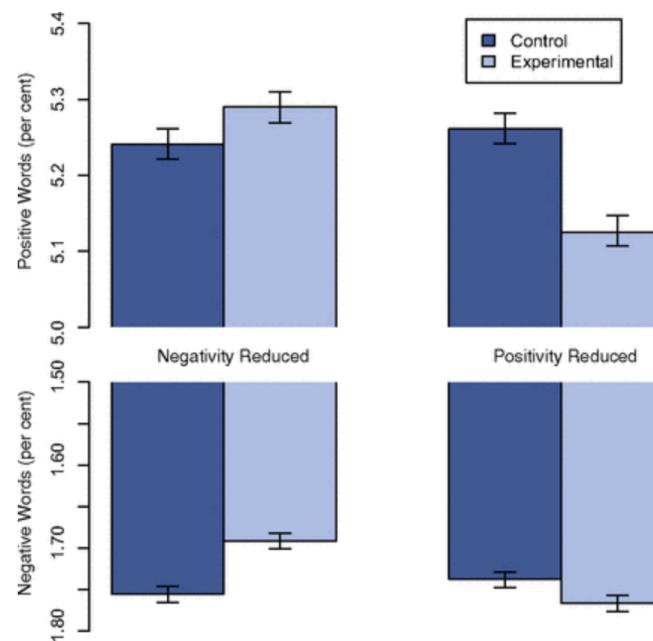
- Improve surveillance, leading to better prevention efforts?
- Better understanding of mechanisms?
- More precise and more effective outreach?

- Doing something simple and useful is better

Facebook experiments



Experimental evidence of massive-scale emotional contagion
through social networks

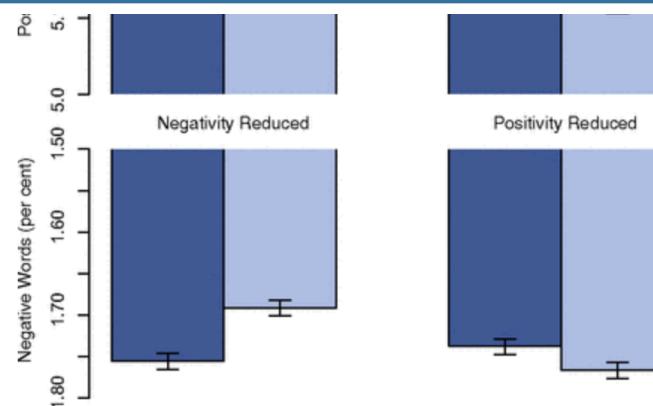


Facebook experiments

PNAS

Significance

We show, via a massive ($N = 689,003$) experiment on Facebook, that emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness. We provide experimental evidence that emotional contagion occurs without direct interaction between people (exposure to a friend expressing an emotion is sufficient), and in the complete absence of nonverbal cues.

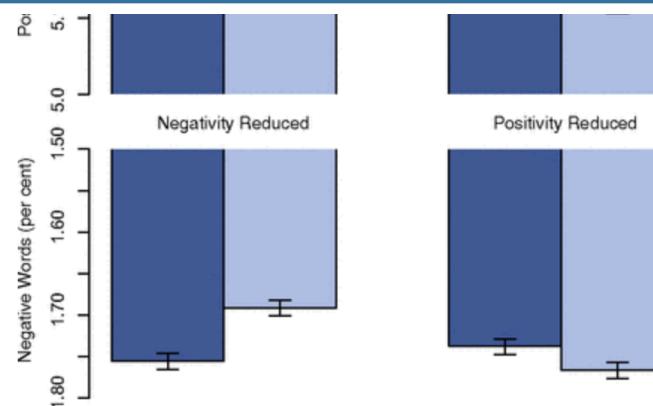


Face BBC Experiments

NEWS

Facebook admits failings over emotion manipulation study

We show, via a massive ($N = 689,003$) experiment on Facebook, that emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness. We provide experimental evidence that emotional contagion occurs without direct interaction between people (exposure to a friend expressing an emotion is sufficient), and in the complete absence of nonverbal cues.



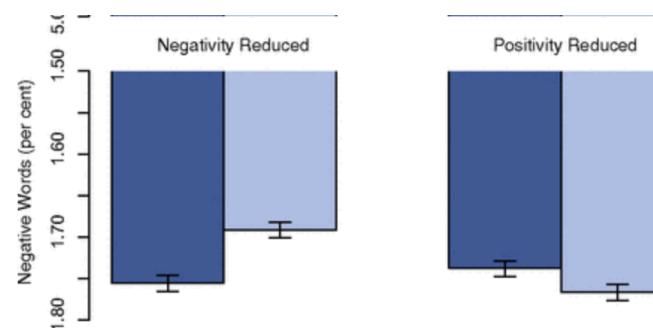
Face **BBC** Experiments

NEWS

Facebook admits failings over emotion manipulation study

the guardian

"Facebook reveals news feed experiment to control emotions



Face  experiments

NEWS 

Facebook admits failings over emotion manipulation study

the guardian

"Facebook reveals news feed experiment to control emotions

Forbes / Tech

Facebook Manipulated 689,003 Users' Emotions For Science

Face experiments

BBC NEWS

Face

ma

th

"Fa

c to

Fo

Faceb

Science



On Facebook, you may be a guinea pig and not know it.

ut

For

AI for Translation

FEATURE

The Great A.I. Awakening

How Google used artificial intelligence to transform Google Translate, one of its more popular services — and how machine learning is poised to reinvent computing itself.

AI for Translation

FEATURE

The Great Translation Experiment

Even artificial intelligence can acquire biases against race and gender

How Google used artificial intelligence to transform Google Translate, one of its more popular services — and how machine learning is poised to reinvent computing itself.

AI for Translation

The Great

How Google used
Translate, one of its
learning is

FEATURE

Even artificial intelligence can acquire biases against race and gender

Google Translate’s gender bias pairs “he” with “hardworking” and “she” with lazy, and other examples

AI for Translation

The Great

How Google Translate, one of the world's most popular learning tools, got性别偏见

FEATURE

Even artificial intelligence can acquire biases against

he is a soldier
she's a teacher
he is a doctor
she is a nurse

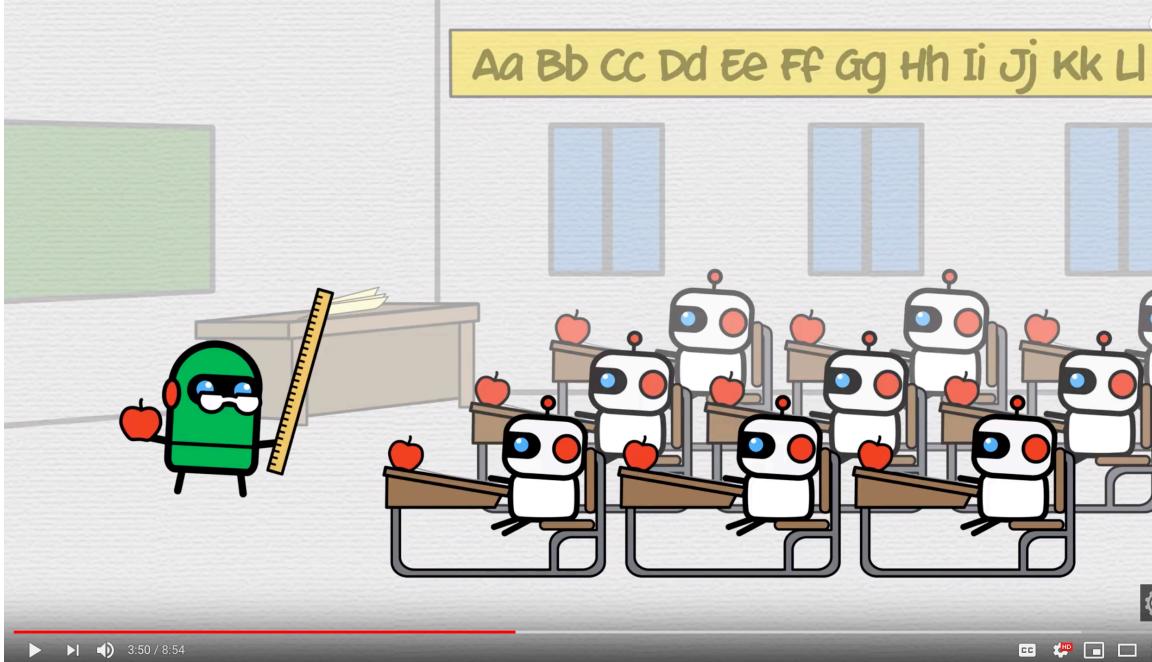
with lazy, and other
examples

slate's gender bias, replacing “he” with “she” and “he” with “she” and other examples

AI and deep learning

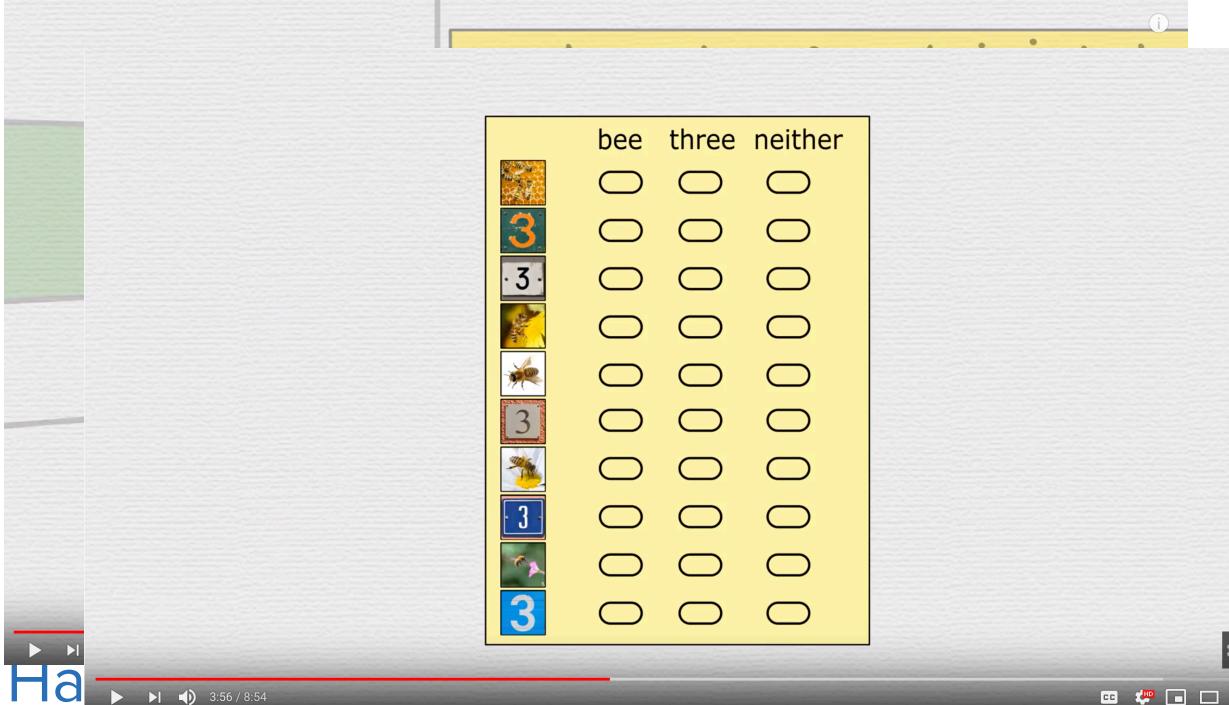
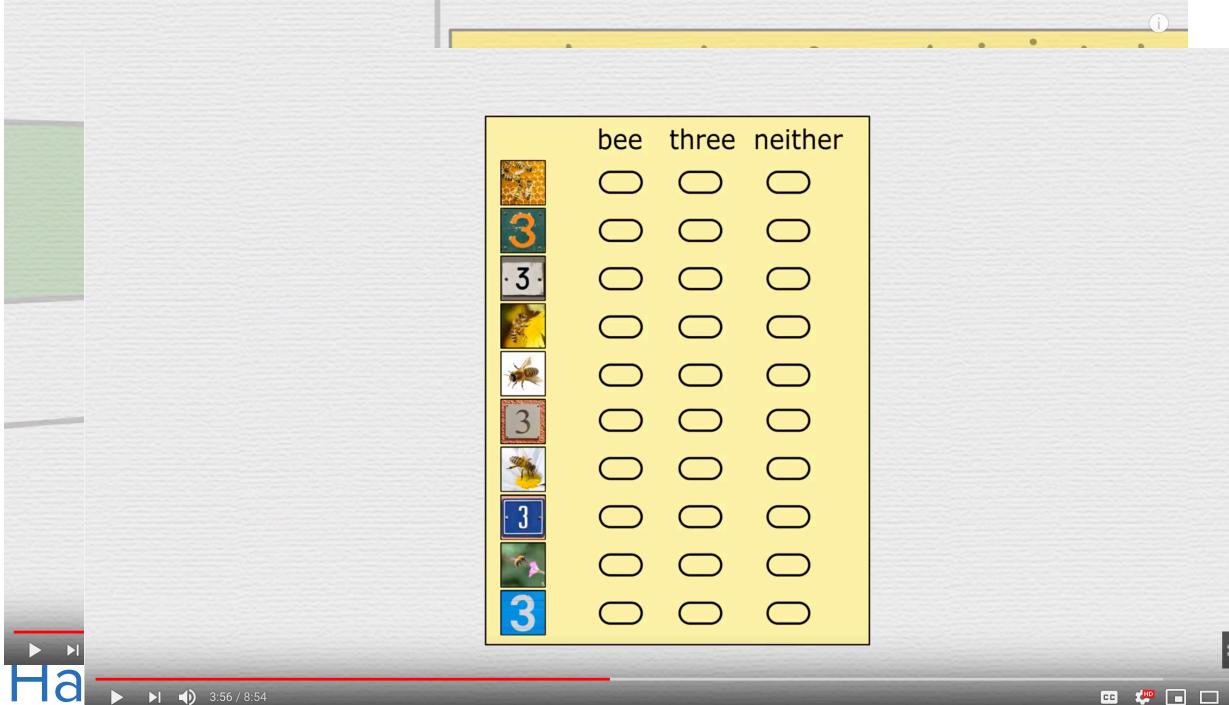
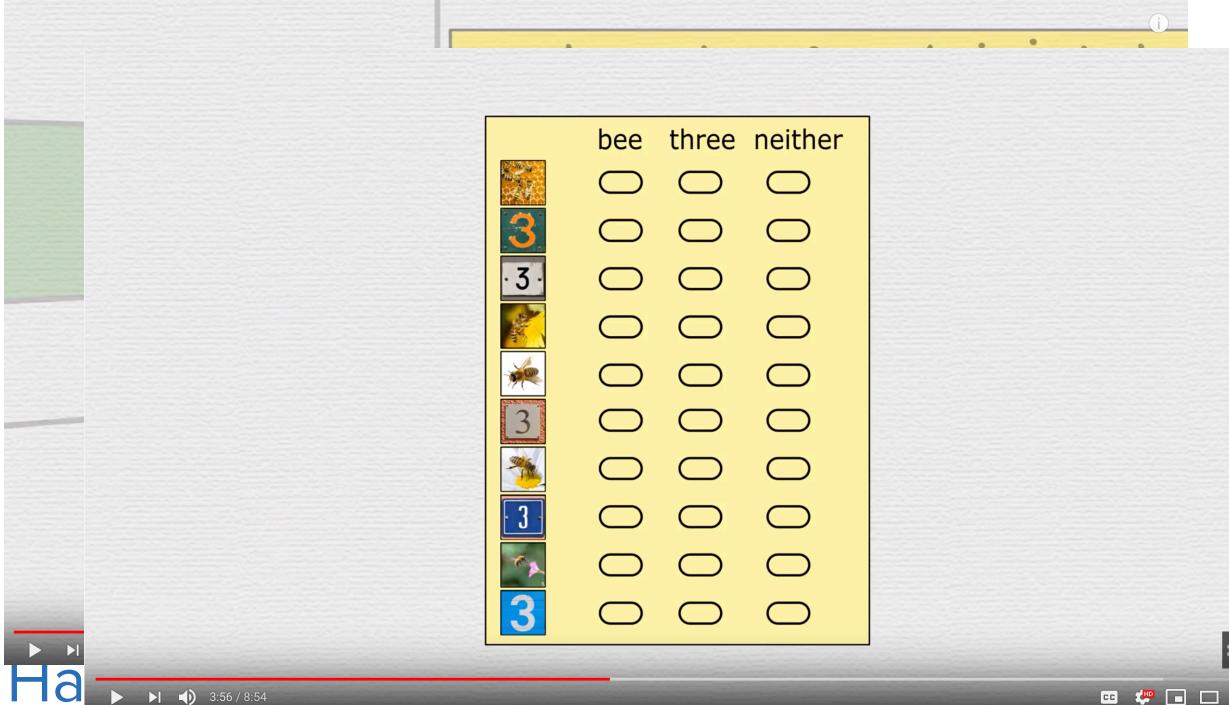
- Not a “magic bullet”
 - Predictions can be very different from the truth, even when advanced techniques are used
- Often requires massive training databases
 - Results are only as good as training data
 - Results also depend on what methods are trained to optimize
- Hard to interpret results...

AI and deep learning

-  The truth, even when advanced
machines are trained to optimize
- Hard to interpret results...

Stills from “How machines learn” by CGP Grey

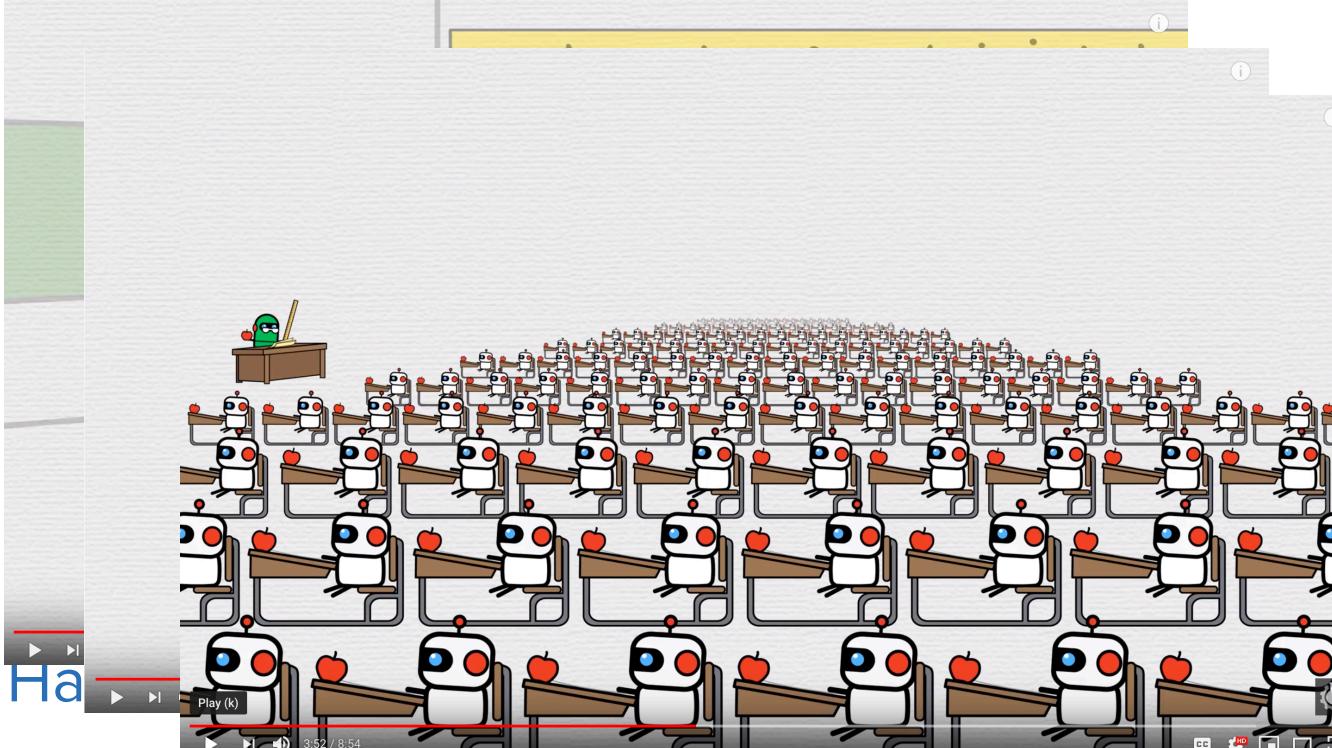
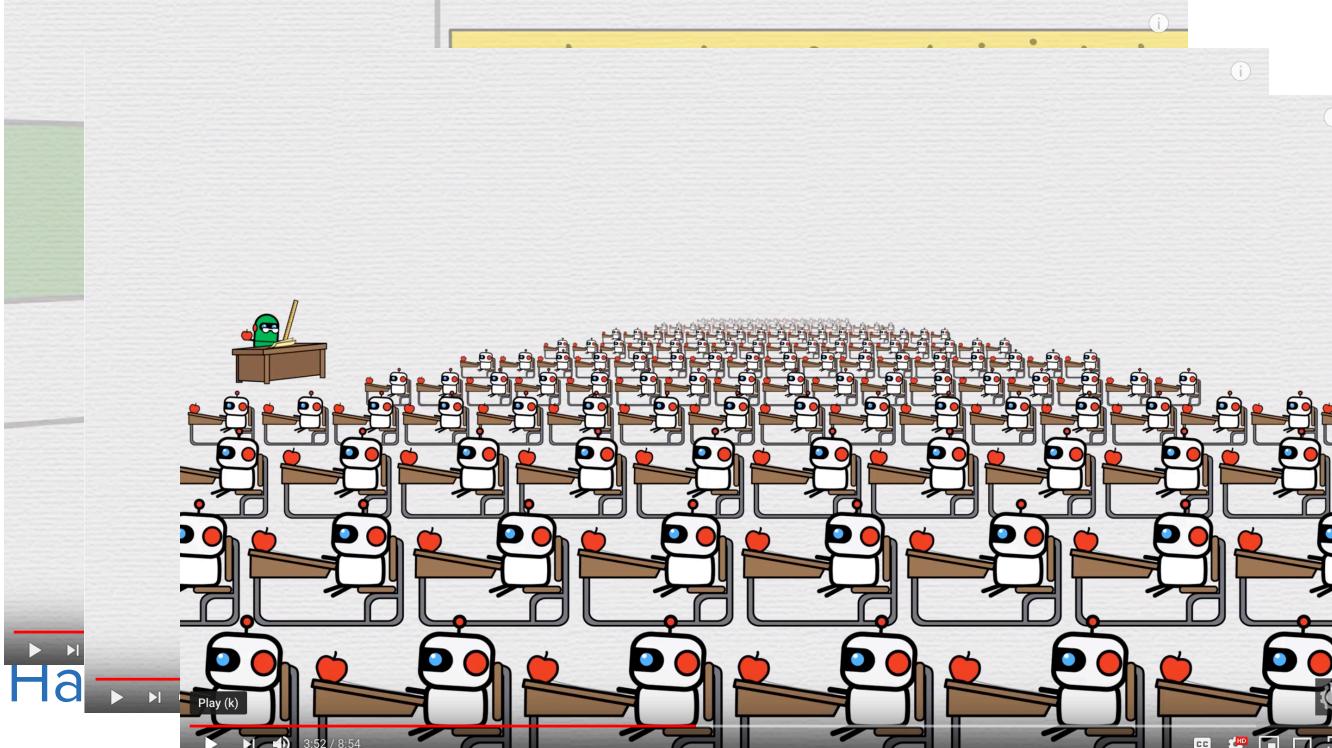
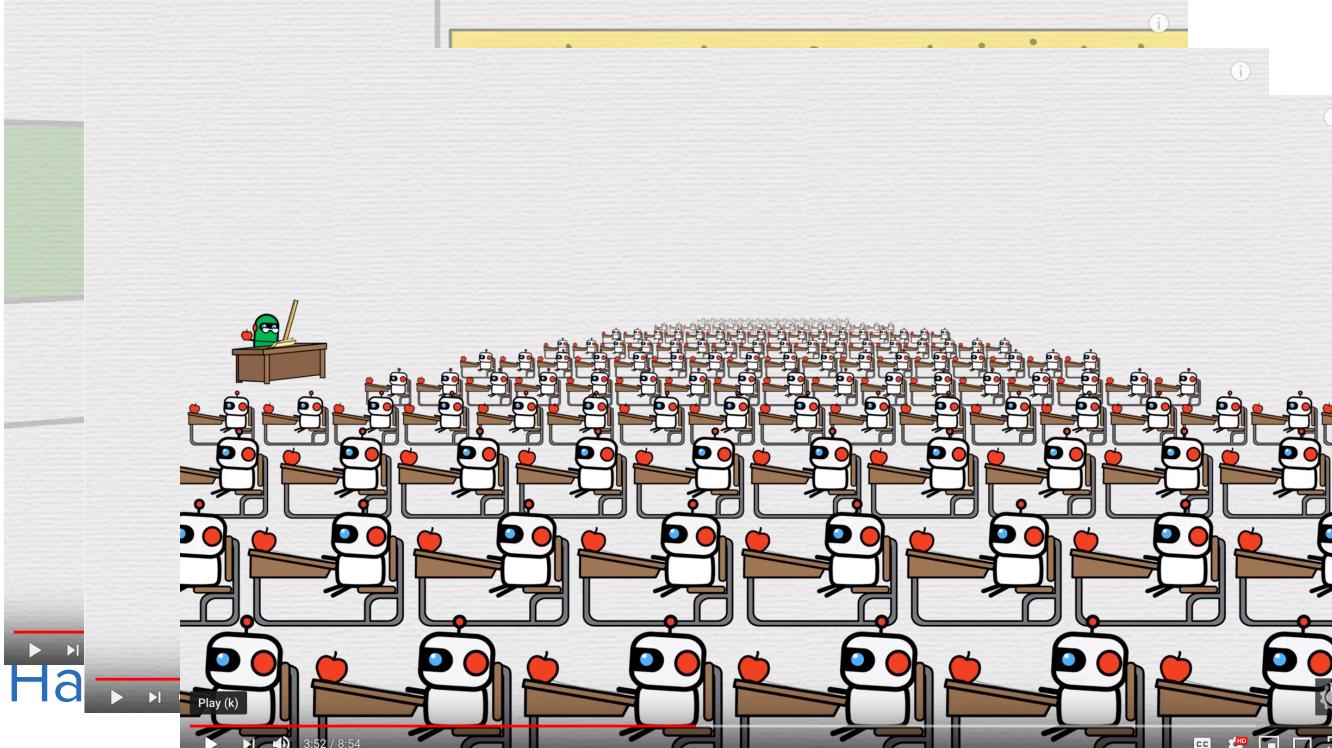
AI and deep learning

- 
- 
- 

truth, even when advanced

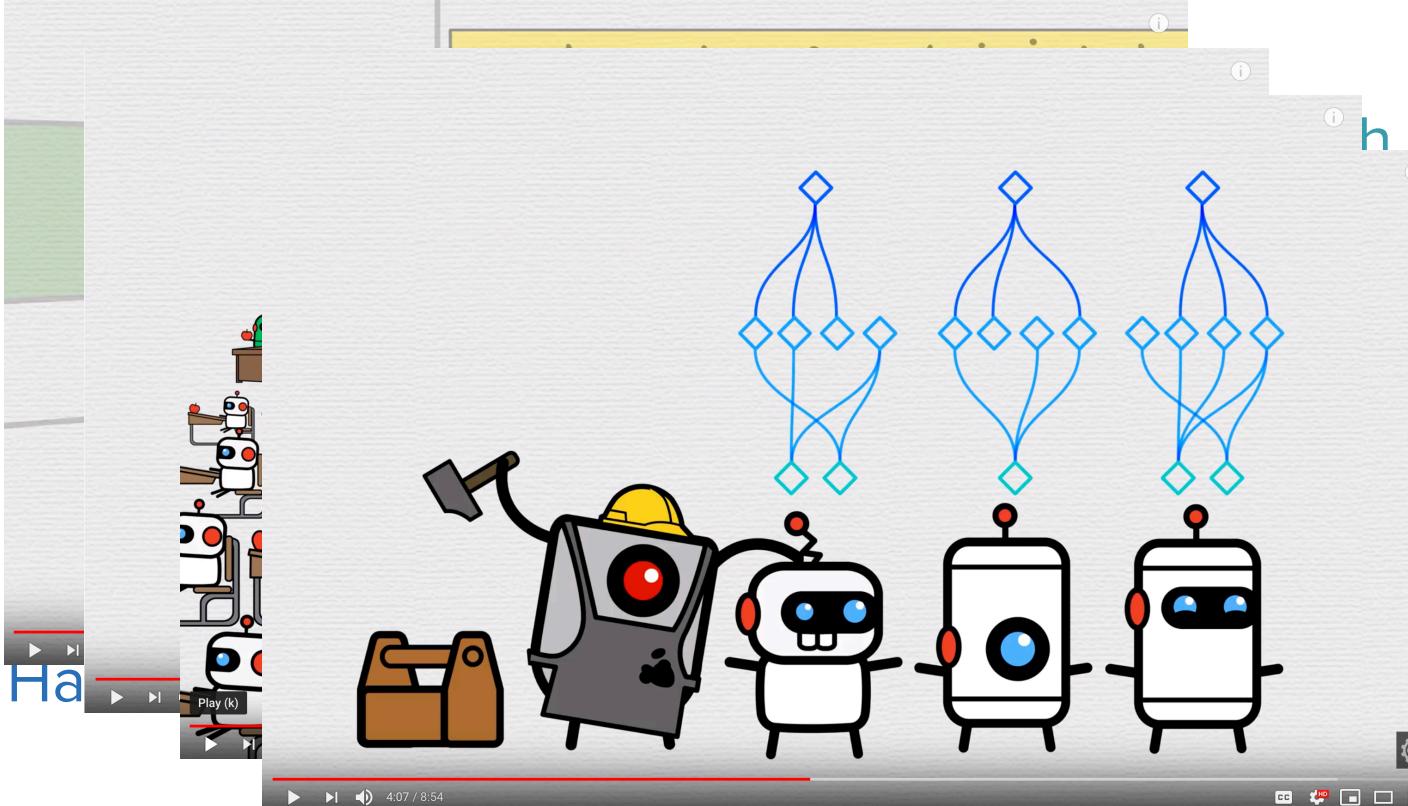
trained to optimize

AI and deep learning

-  Even when advanced, AI is still trained to optimize
-  Even when advanced, AI is still trained to optimize
-  Even when advanced, AI is still trained to optimize

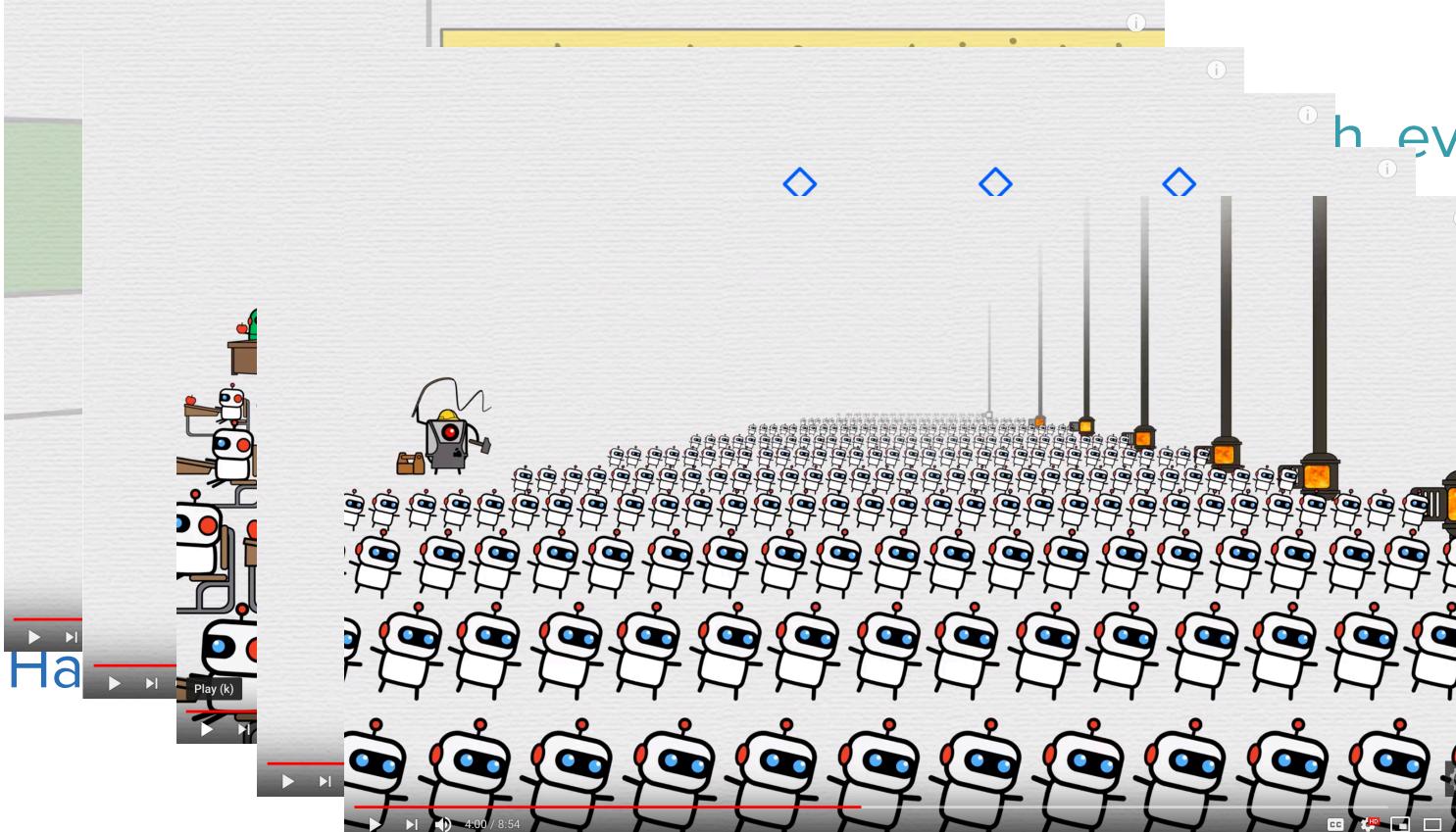
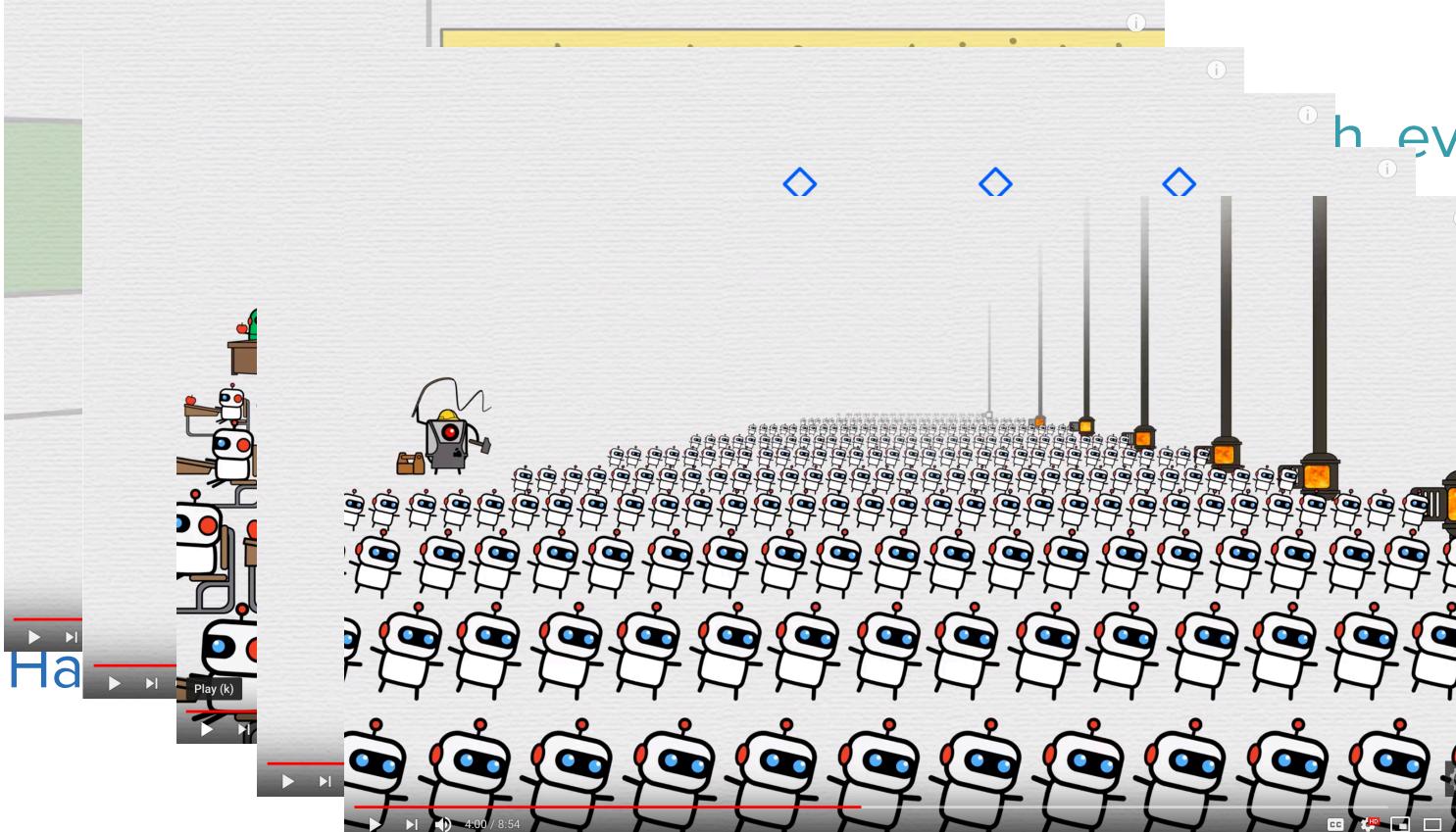
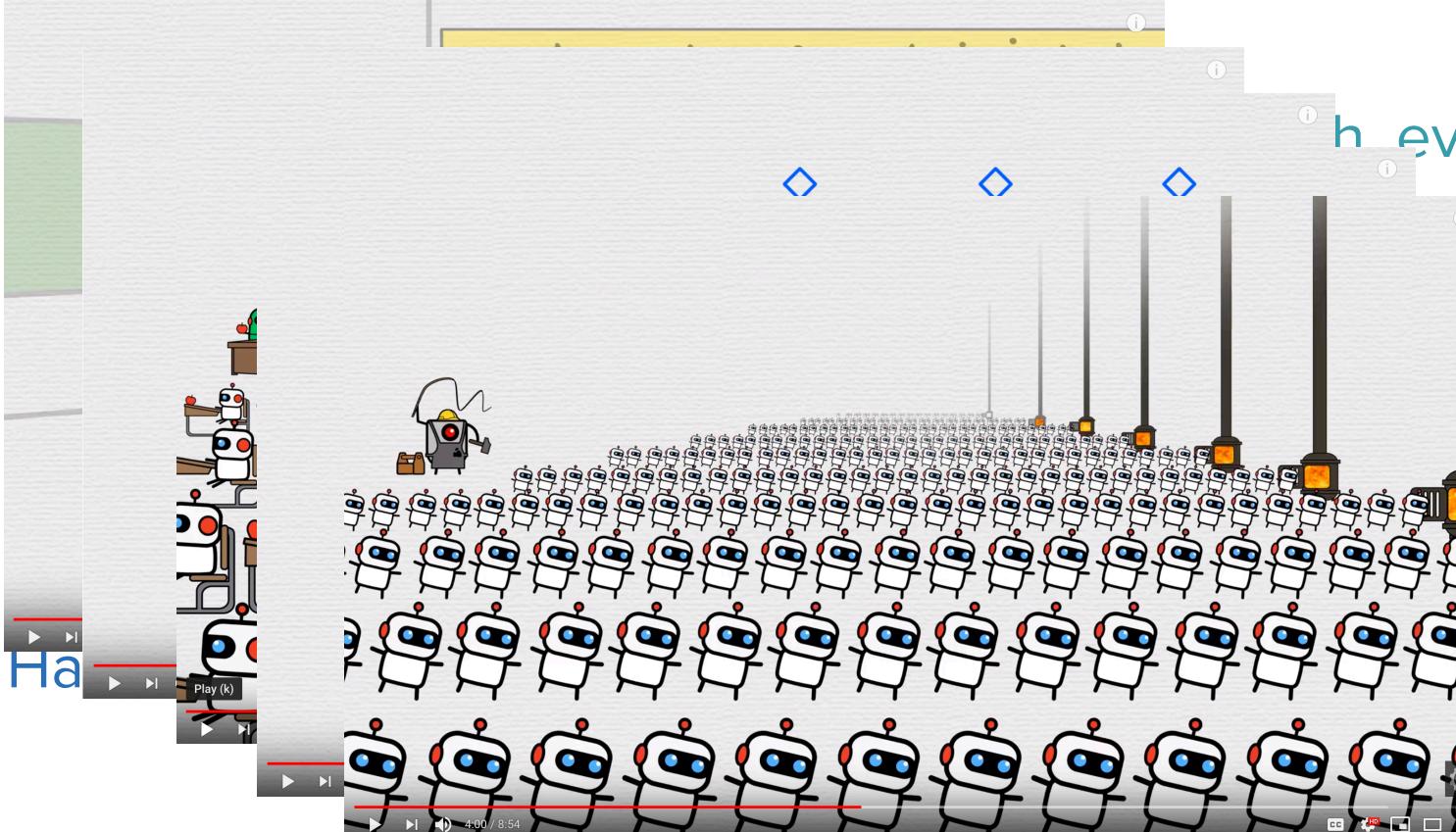
Stills from “How machines learn” by CGP Grey

AI and deep learning

- A video player interface showing a cartoon robot holding a hammer and a toolbox, standing next to three other robots. Above them are three neural network diagrams. The video player interface includes a play button, volume control, and a progress bar showing 4:07 / 8:54.
- *h even when advanced*
- *d to optimize*
- **Ha**

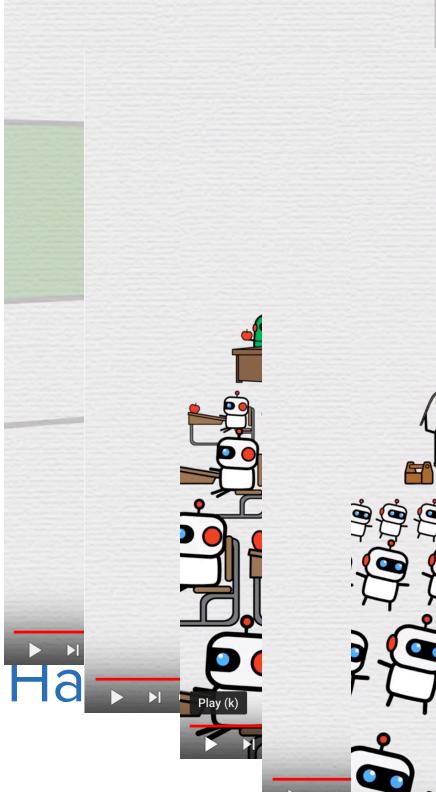
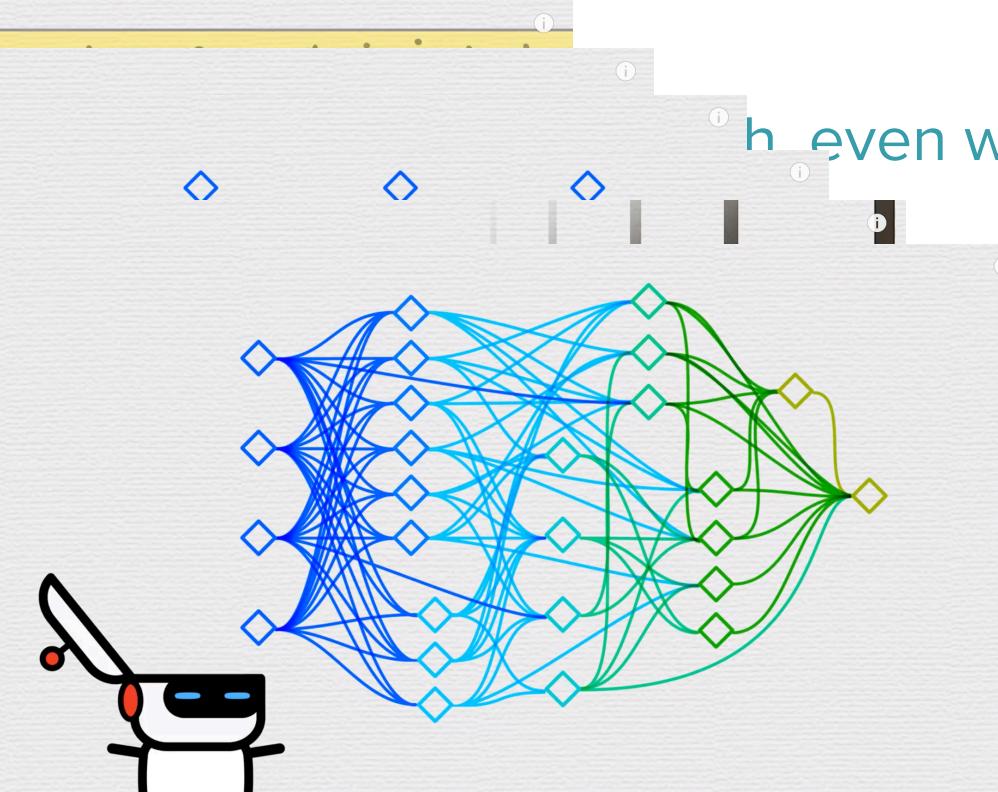
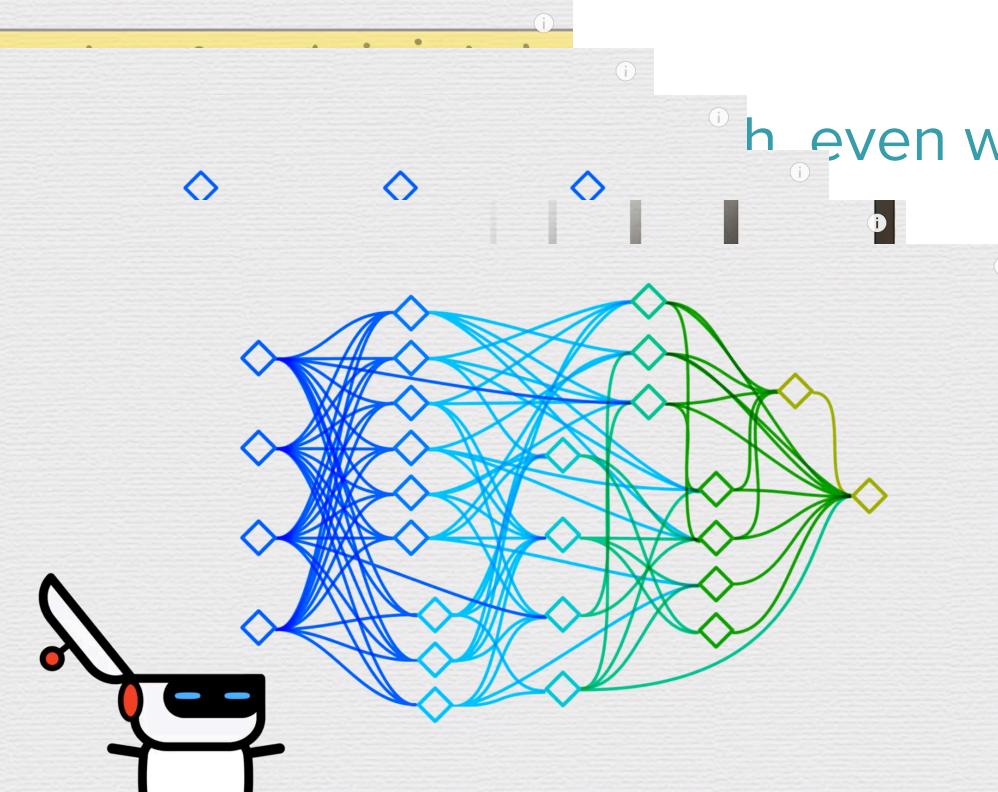
Stills from “How machines learn” by CGP Grey

AI and deep learning

-  h, even when advanced
-  o optimize
-  Ha

Stills from “How machines learn” by CGP Grey

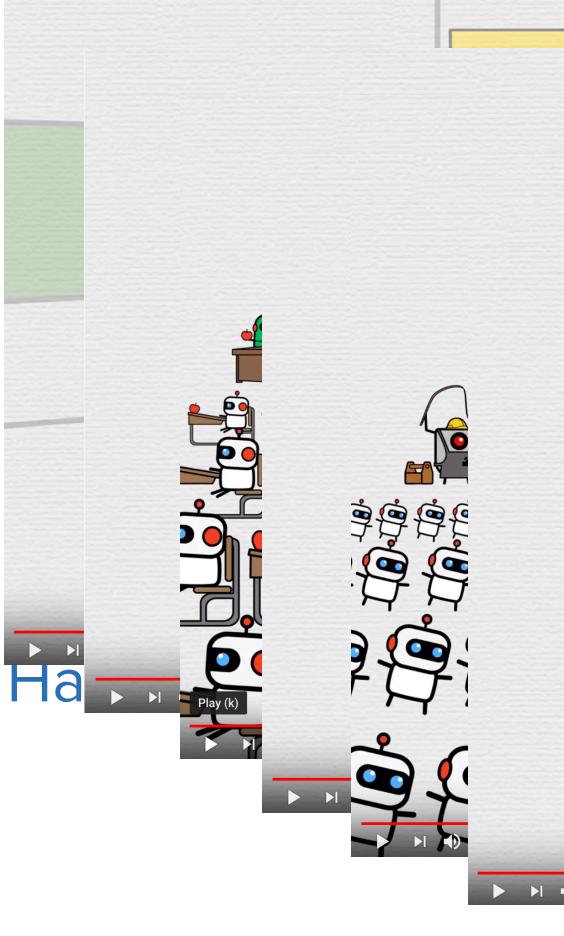
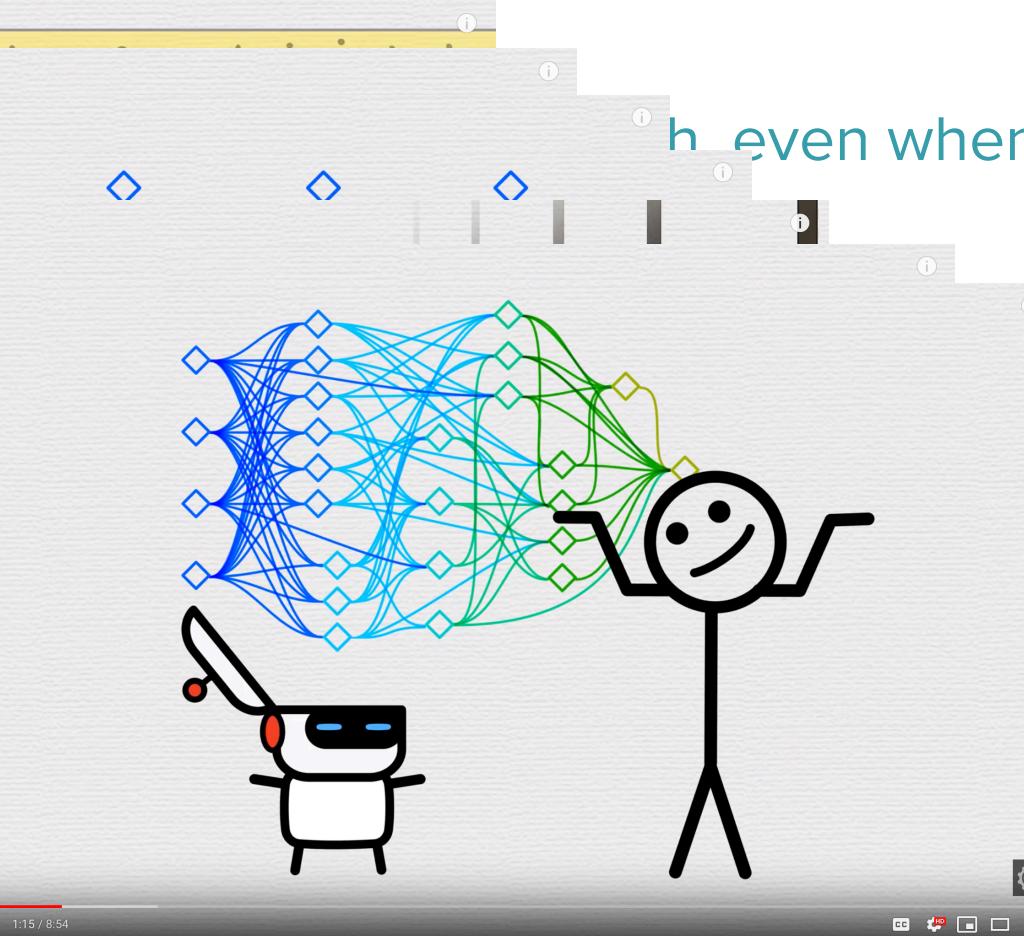
AI and deep learning

-  A video player interface showing a scene where a large white robot is interacting with several smaller, colorful robots. The video controls at the bottom include a play button, volume, and a progress bar indicating 4:50 / 8:54.
-  A video player interface showing a complex neural network diagram with many nodes and connections. The video controls at the bottom are identical to the previous frame.
-  A video player interface showing a complex neural network diagram with many nodes and connections. The video controls at the bottom are identical to the previous frame.

h, even when advanced

timize

AI and deep learning

-  A screenshot from a video showing a white robot with blue eyes and a red antenna playing a game with other robots. The interface includes a green bar at the top, a yellow progress bar, and a 'Play (k)' button.
-  A screenshot from a video showing a stick figure interacting with a complex neural network diagram. The diagram consists of many interconnected nodes (blue diamonds) and connections. A small black robot is also present near the network.
-  A screenshot from a video showing a stick figure with a speech bubble containing the text 'h, even when advanced'.

Stills from “How machines learn” by CGP Grey

Limitations of data

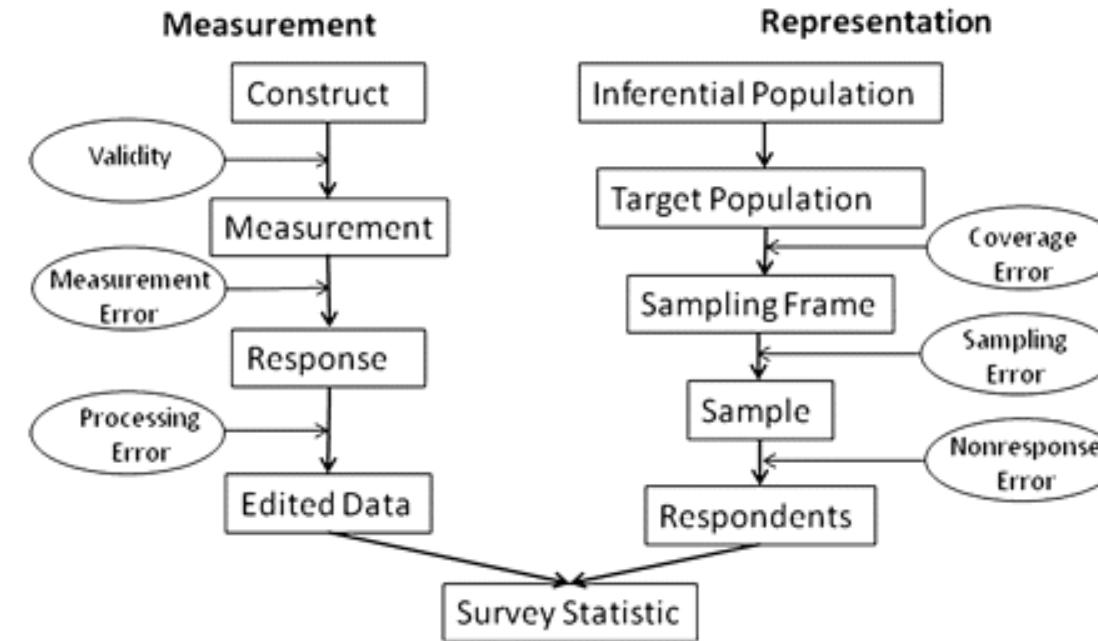
- Not trying to gang up on Google, Facebook etc
 - There's a lot to say there ...
 - But in each case, these are smart people doing interesting (maybe even important) things with cool data
- These cases point to challenges to be overcome, and are important opportunities
 - How can public health practitioners engage with non-traditional partners in a beneficial way?
 - How can tech be used or evaluated as a public health tool when it changes so rapidly?
 - How can big data overcome issues of selection bias and access?

A public health lens

How can we use these data to improve health?

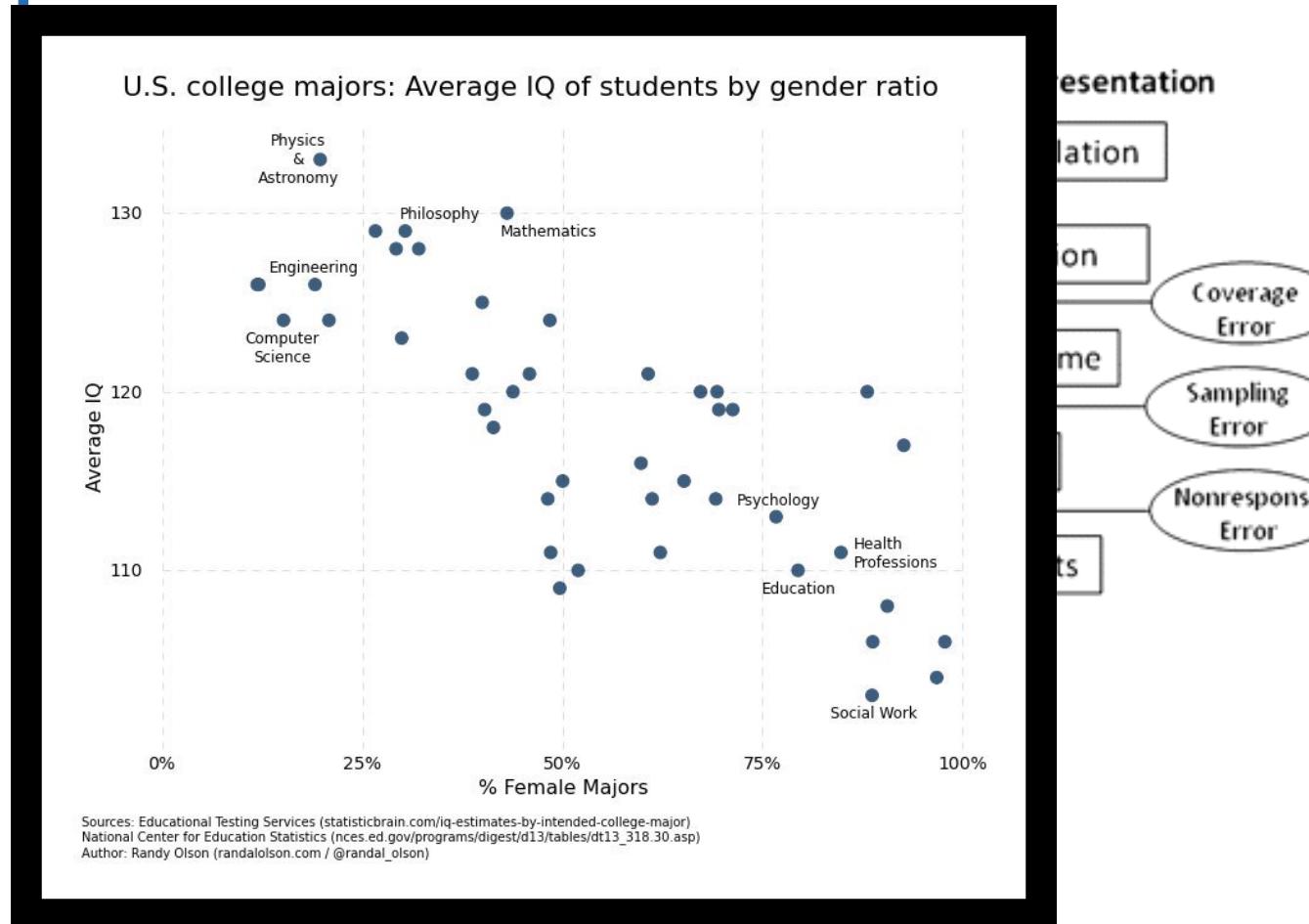
- Who is represented?
- What do measurements mean? Can they be trusted?
- Will I uncover associations? Causes?
- Can I develop new hypotheses or confirm existing ones?
- Can I design an intervention, or evaluate the success of one?

Be skeptical about data



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

Be skeptical about data



representation

lation

on

me

ts

Coverage
Error

Sampling
Error

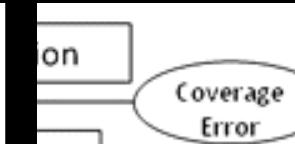
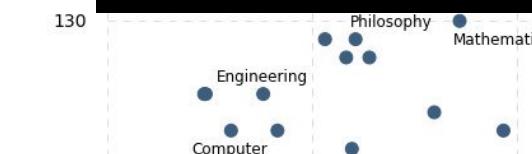
Nonresponse
Error

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

Be skeptical

Untrustworthy Data Will Lead to Poor Conclusions

Trusting all data as if it were fact is a dangerous proposition.



So, can any data ever be trusted?

The short answer is... *it depends*. Skepticism is not a free pass to disregard data you disagree with. It's a tool to ensure that the conclusions derived from data are reliable and do, in fact, reflect reality.

You also shouldn't trust data just because it "proves" a point that you're already inclined to believe. It's probably even more important to be skeptical of extraordinary claims with which your heuristics already naturally align.

Sources: Economic
National Center for
Author: Ram

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

A caveat before starting ...

People sometimes confuse fancy methods for data science.

Don't Do That.

A simple method applied to good data and clearly communicated
is **much** better than
a fancy method that no one understands applied to bad data.