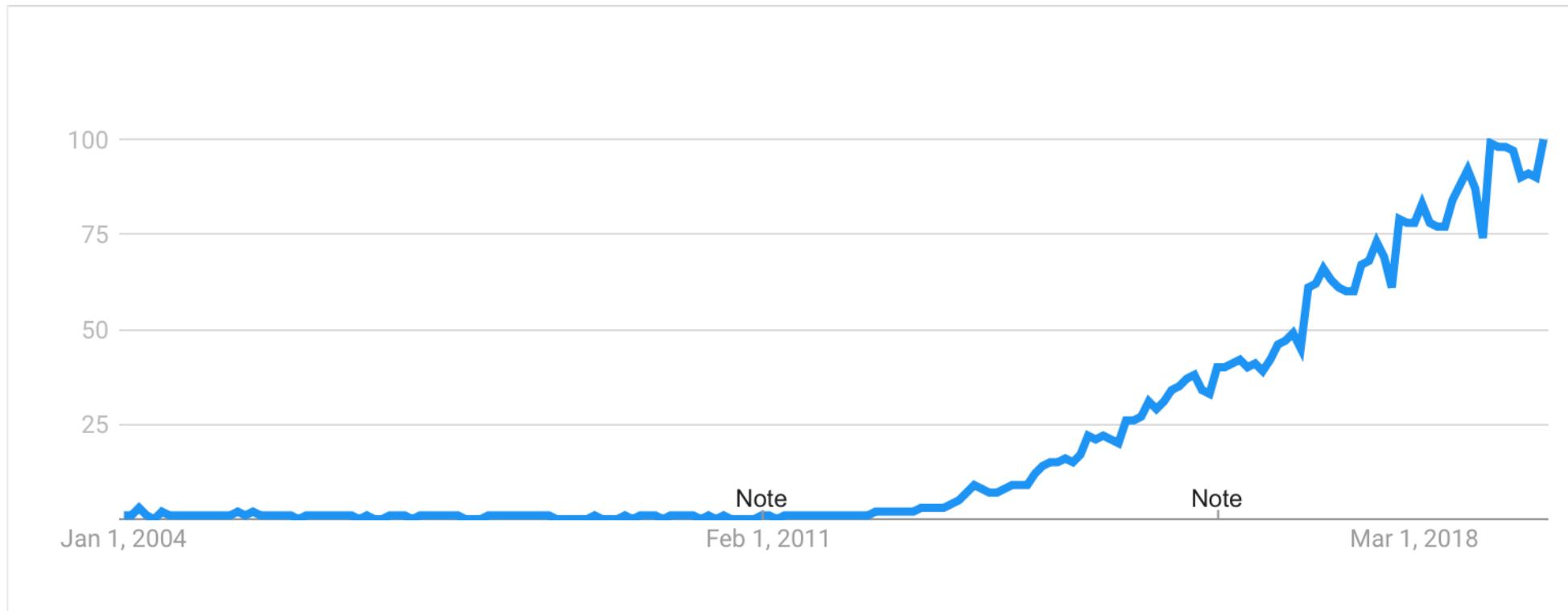


WHAT IS DATA SCIENCE?

Jeff Goldsmith, PhD

Department of Biostatistics

Data science is pretty new



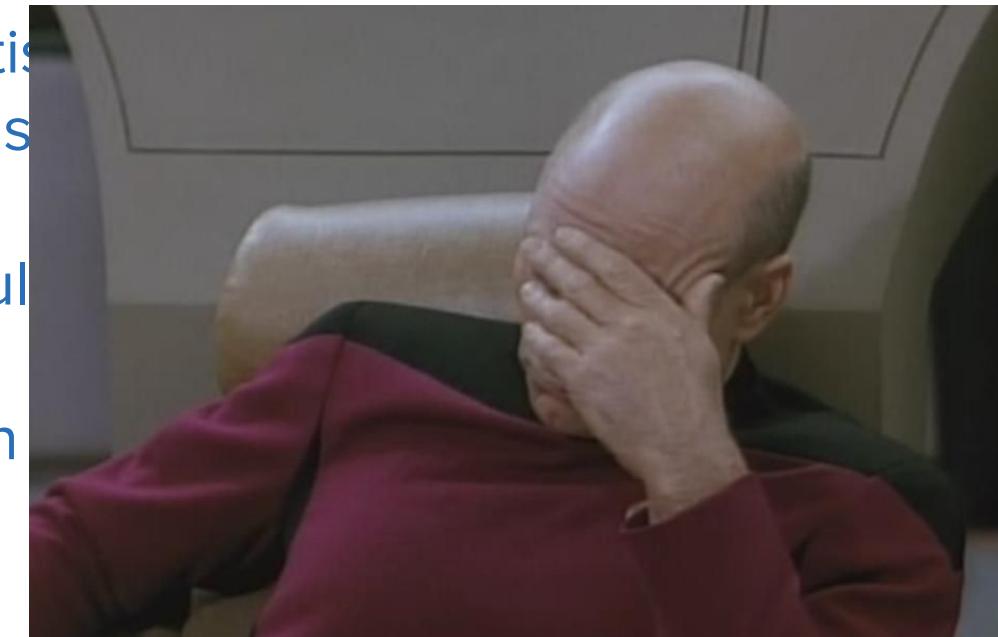
Source: Google Trends

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.”
- “A data scientist is a better computer scientist than a 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?

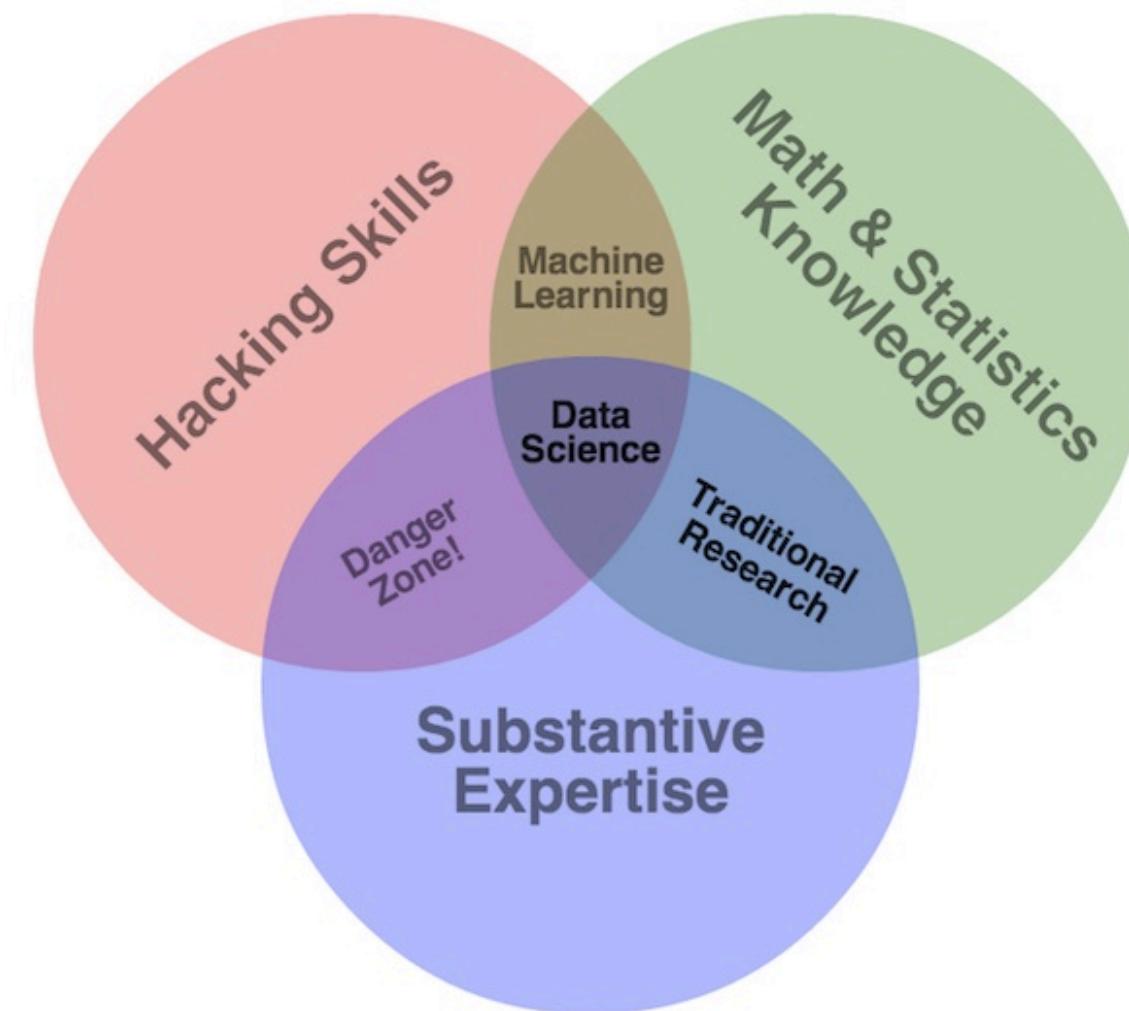
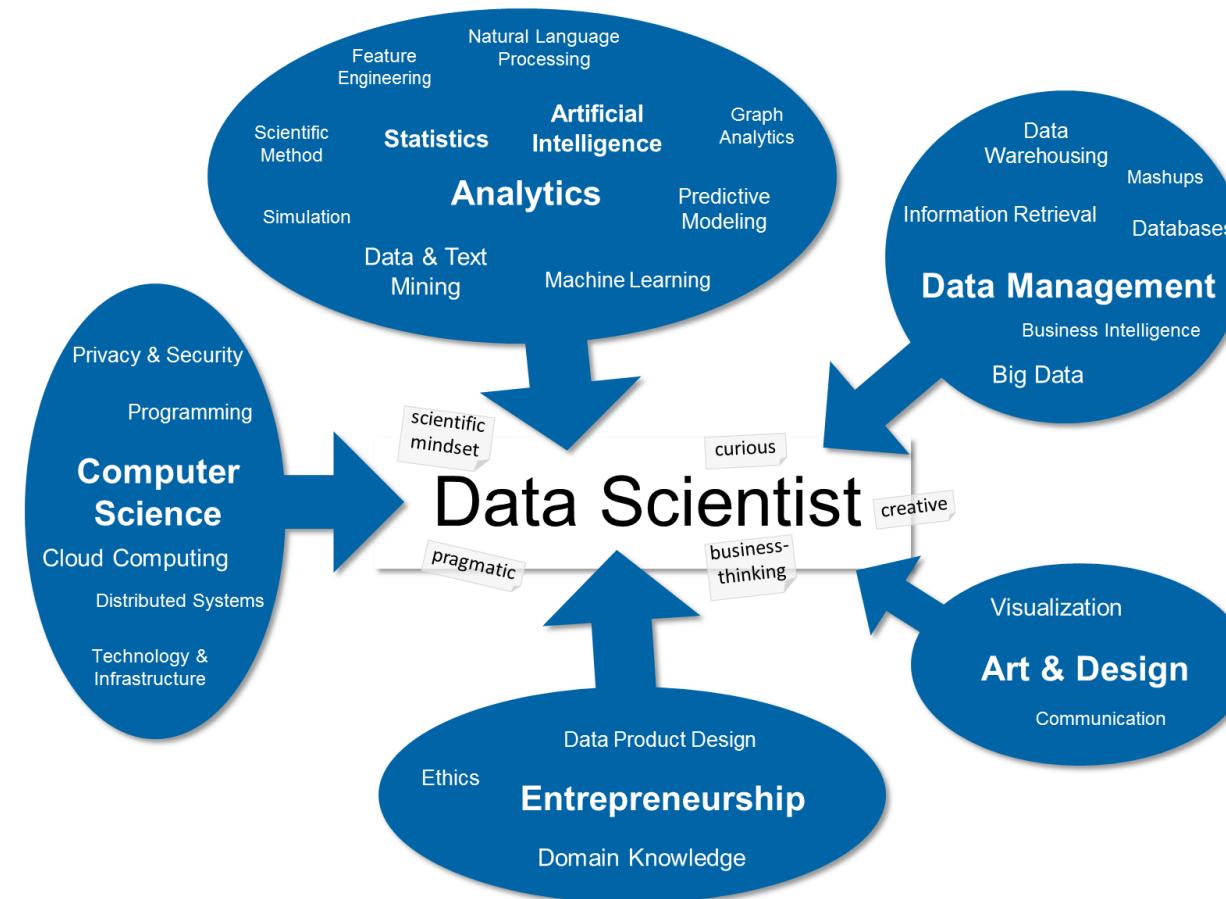


Image from Drew Conway

Maybe pictures will help?



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	PROGRAMMING & DATABASE
<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 	<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	COMMUNICATION & VISUALIZATION
<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 	<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	PROGRAMMING & DATABASE
<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 	<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	COMMUNICATION & VISUALIZATION
<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 	<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

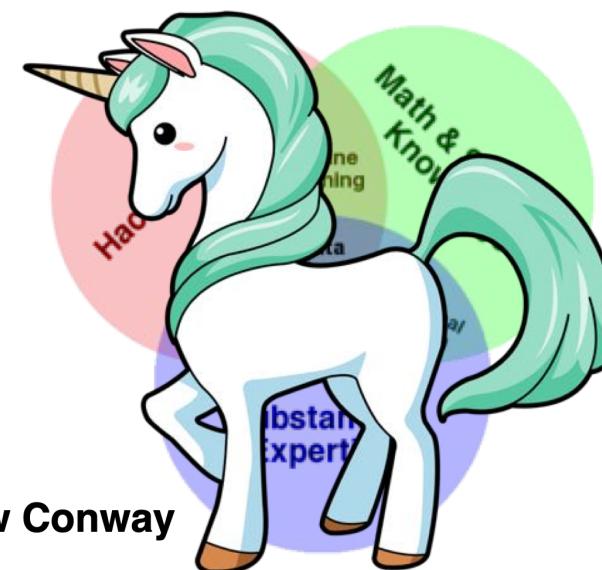
- “Data science is just ...” definitions miss the point
 - If data science is just statistics (or machine learning, or computer science, or engineering) we wouldn’t need a new term, let alone a new discipline
 - The popularity of “data science” suggests that there’s a newly recognized need
- “A data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



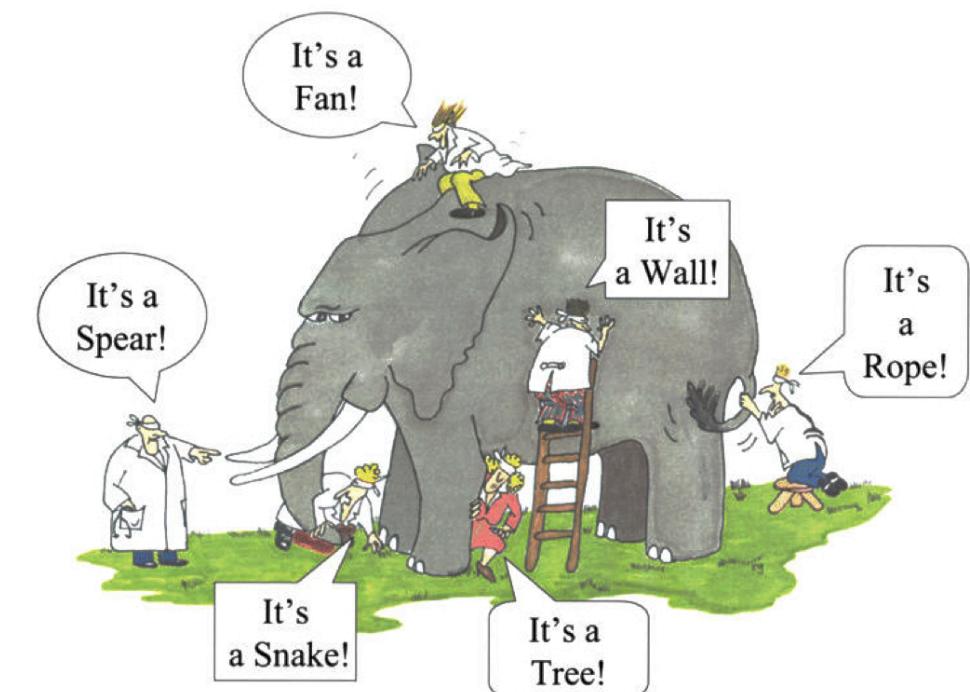
17

@angebassa

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

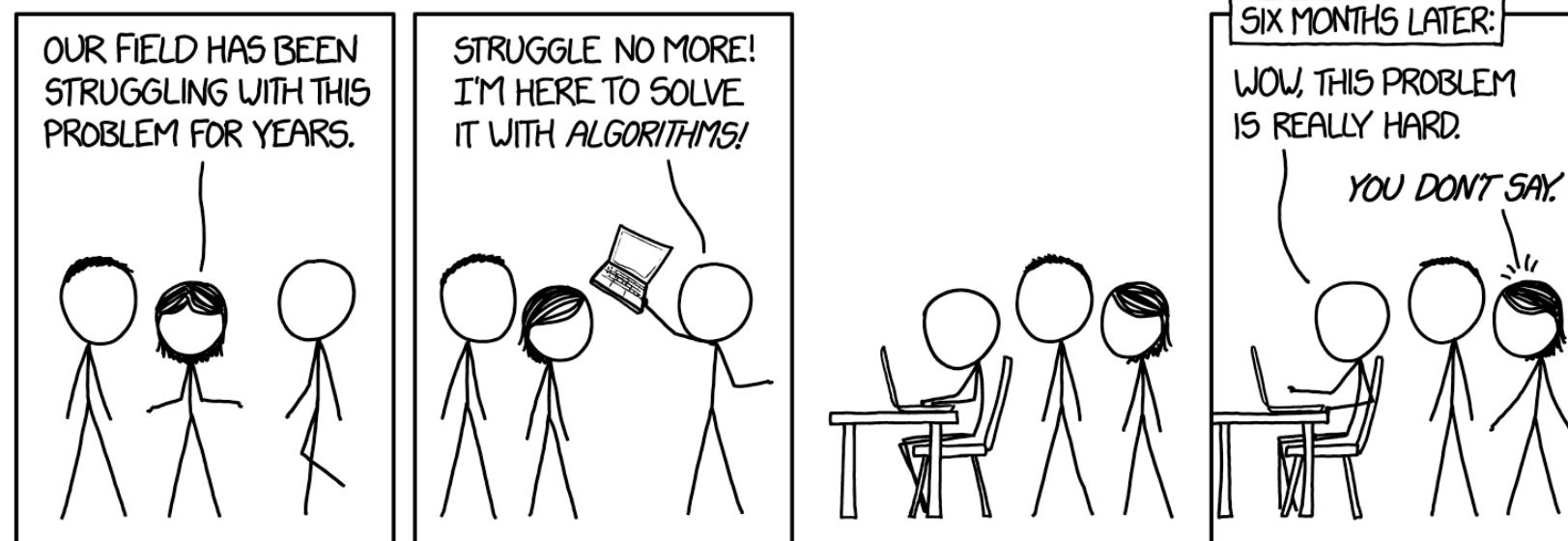
Why these definitions are good

- 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/>

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

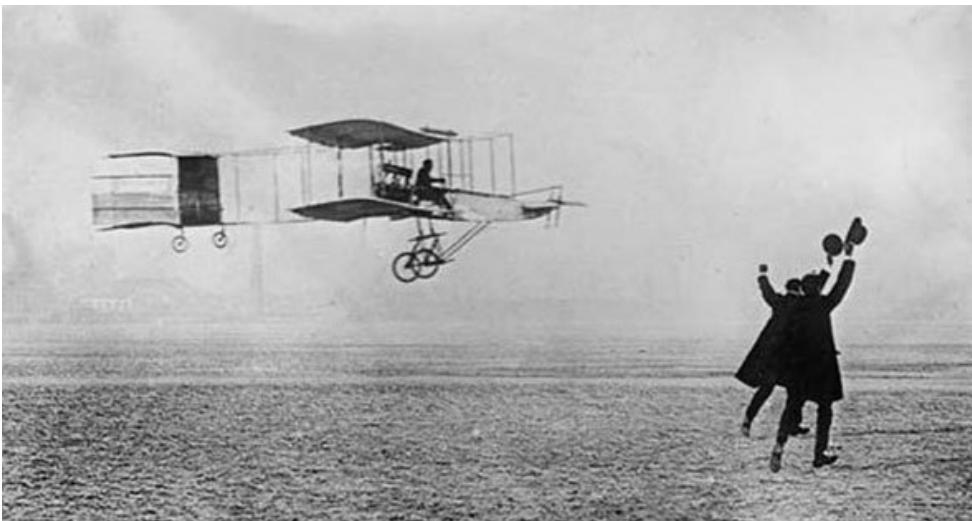
For the purpose of this class:

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

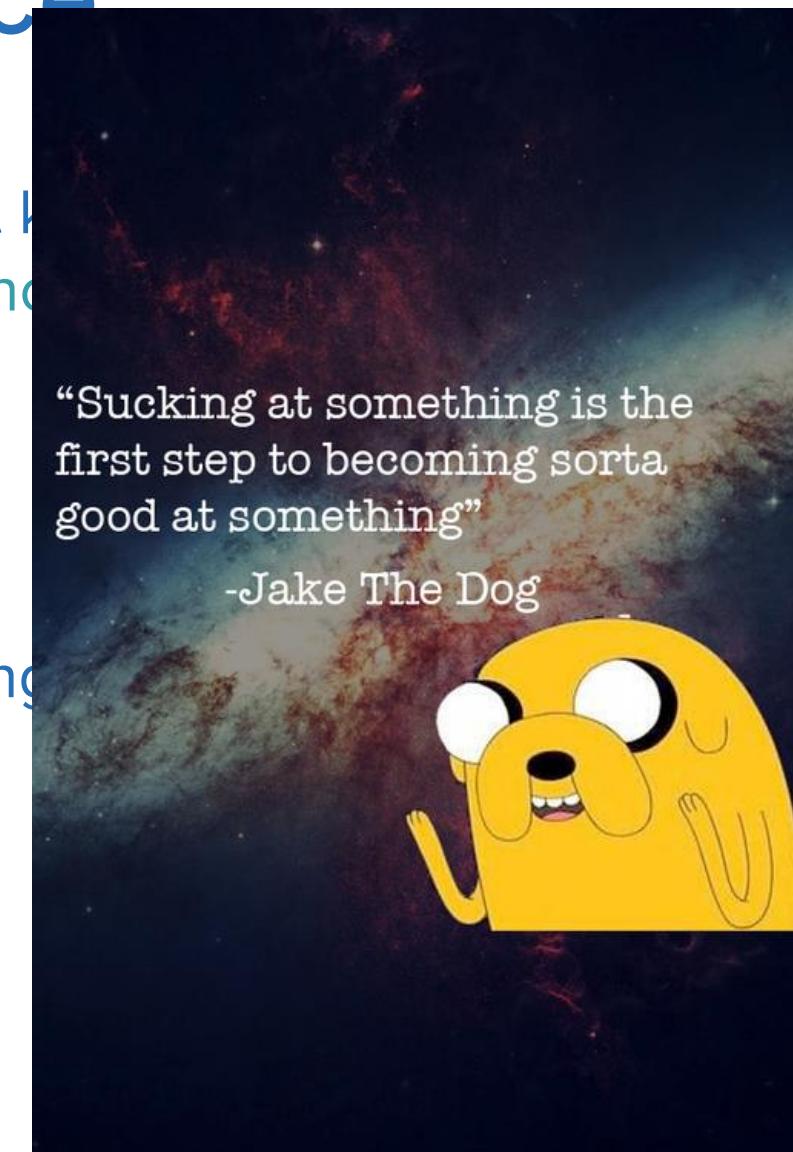
How to learn data science

- 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 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 that people 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



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

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

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
- @jtleep
- @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



David Parkins

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...
The Washington Post*

Sections ≡

The Washington Post

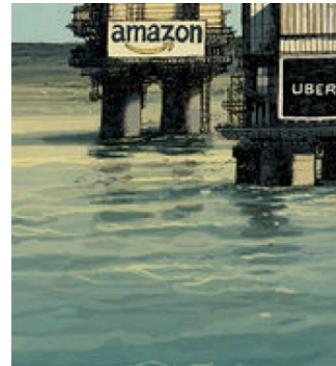


WP BrandStudio



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 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 have a revenue line for Disney. Instead, it's about giving the company the ability to know each of its fans individually, including their names, addresses, and characters they like, and how much, and where they live. This is why the annual plan is priced at \$129.99. 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

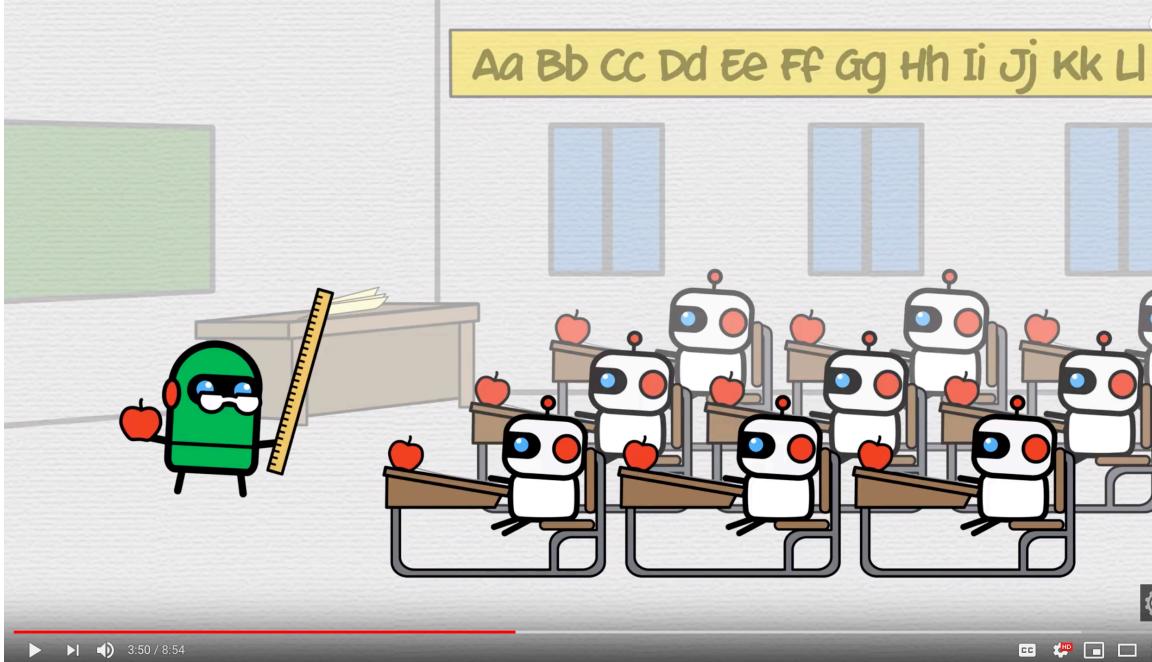
4,489 Retweets 11,130 Likes

141 4.5K 11K

AI and deep learning

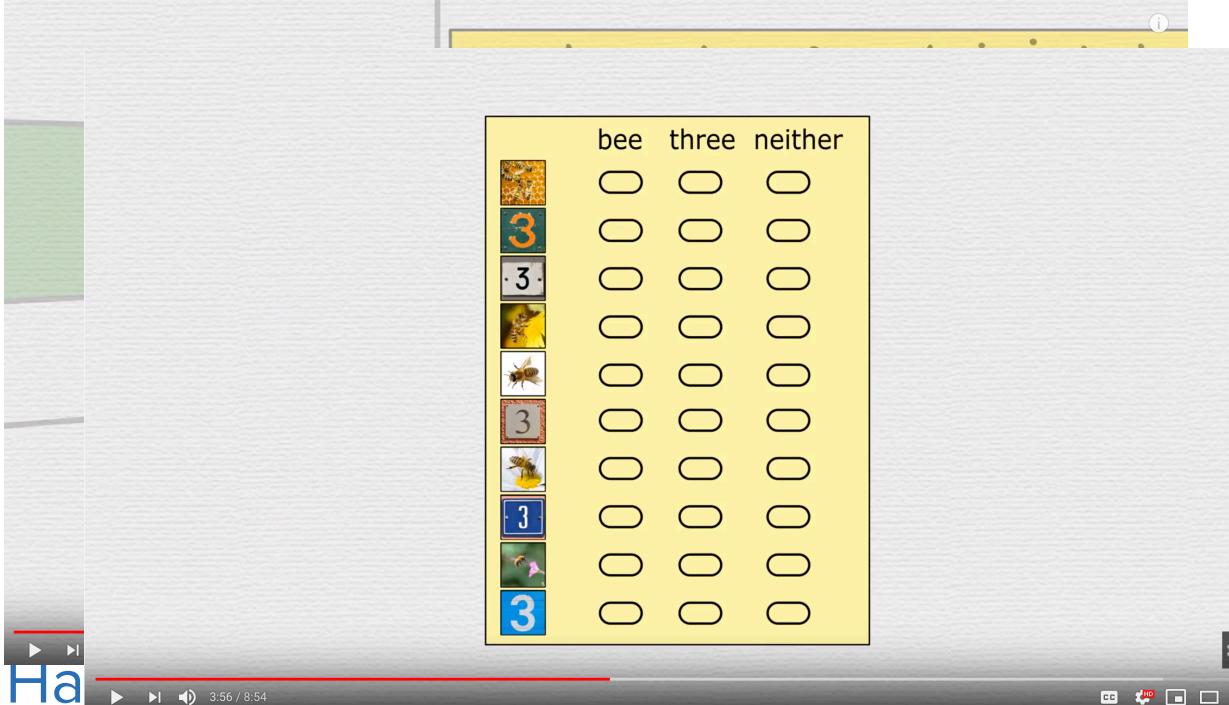
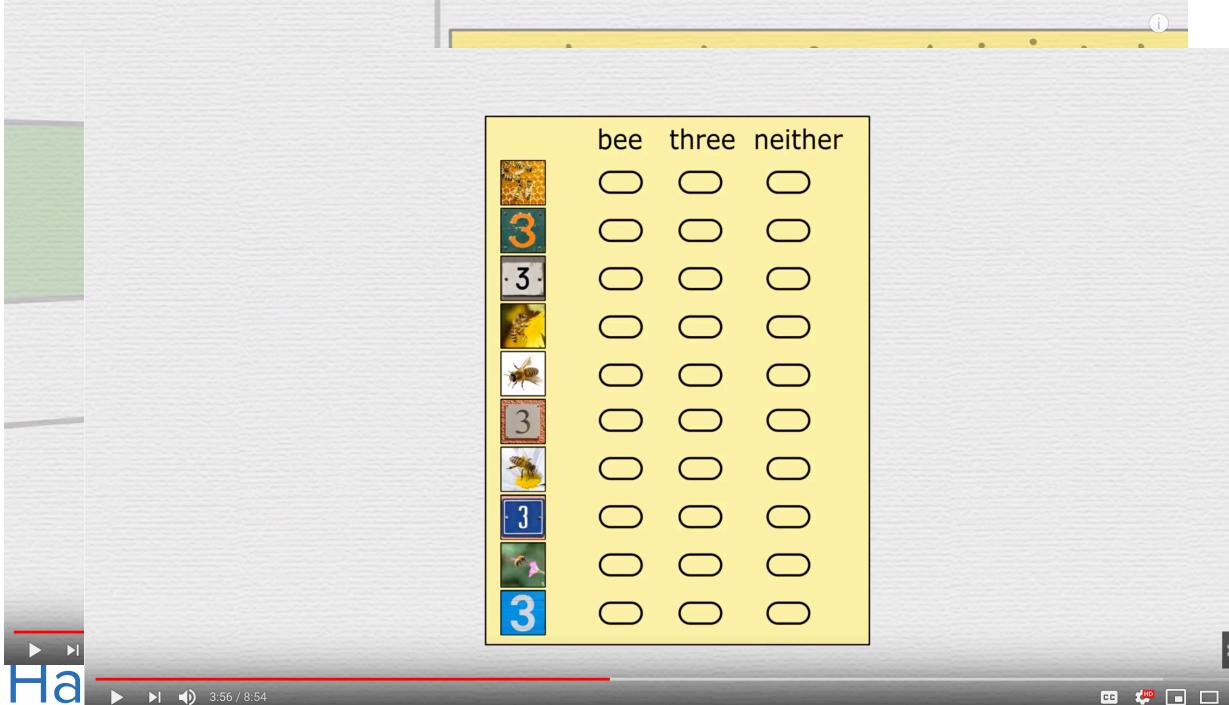
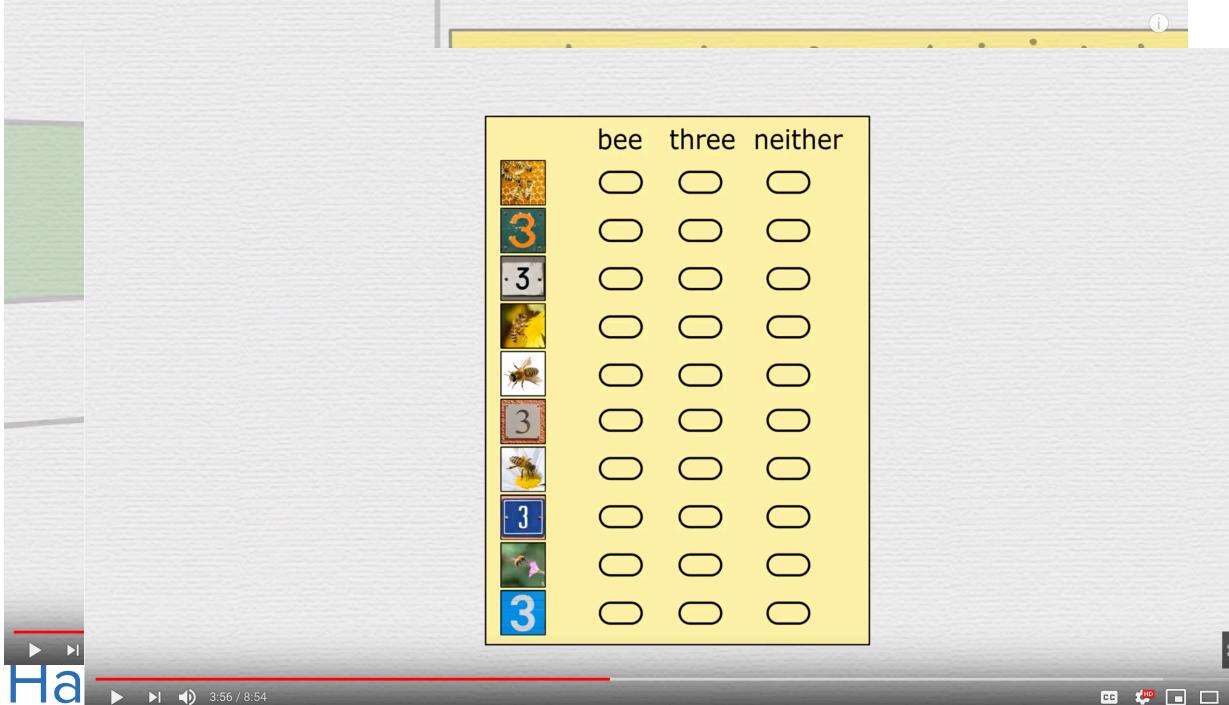
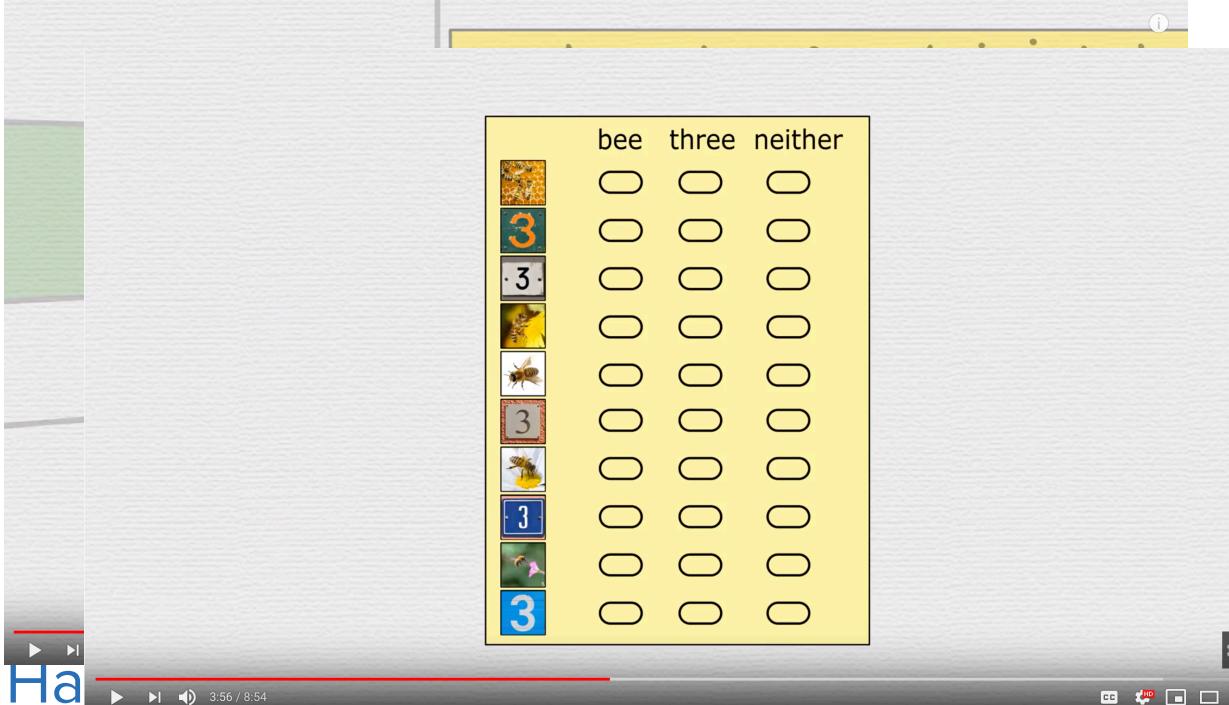
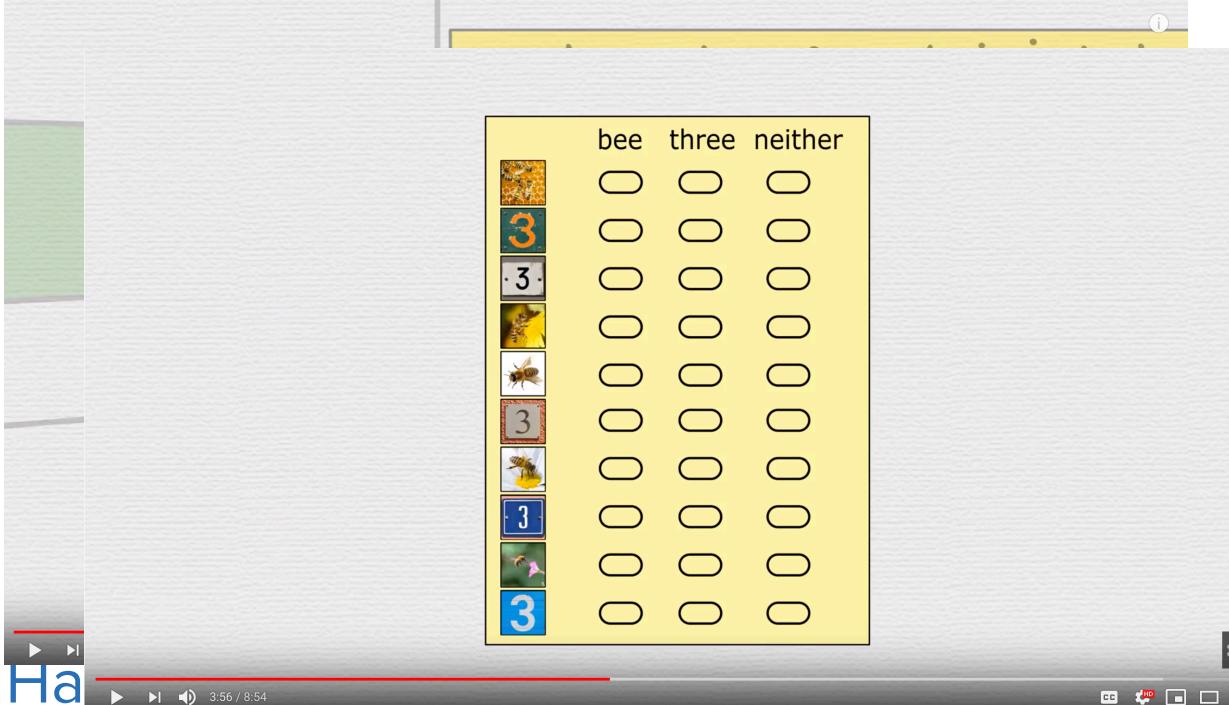
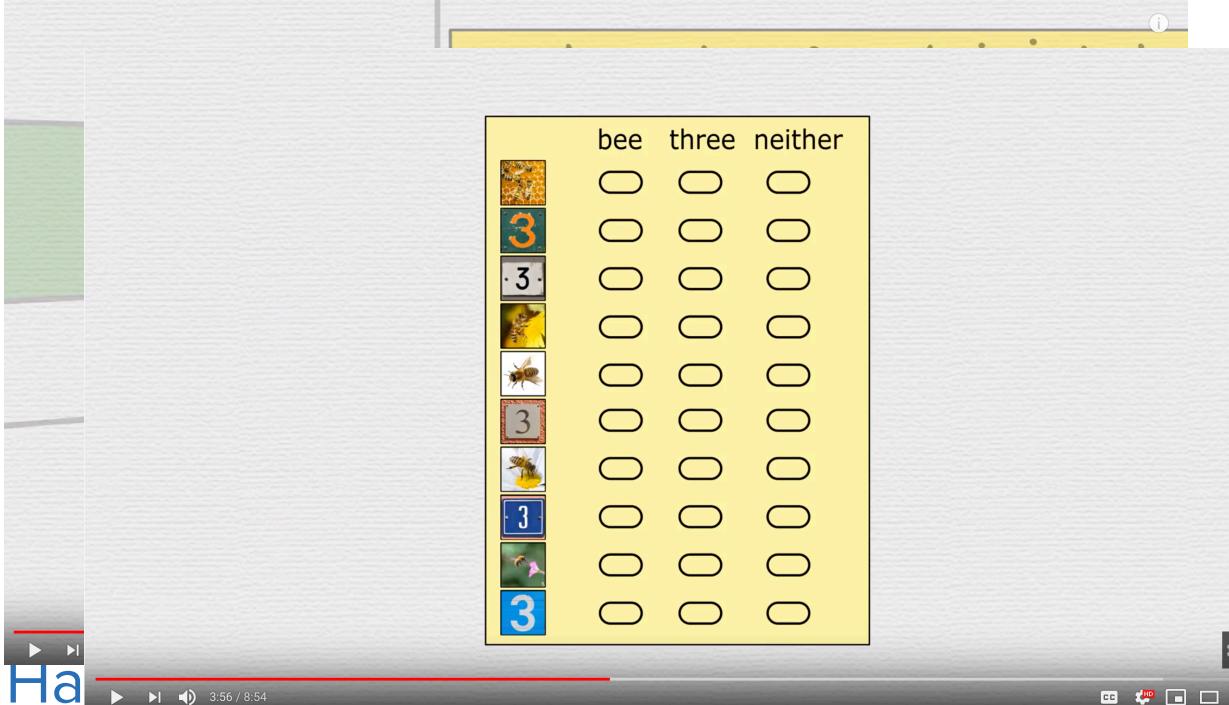
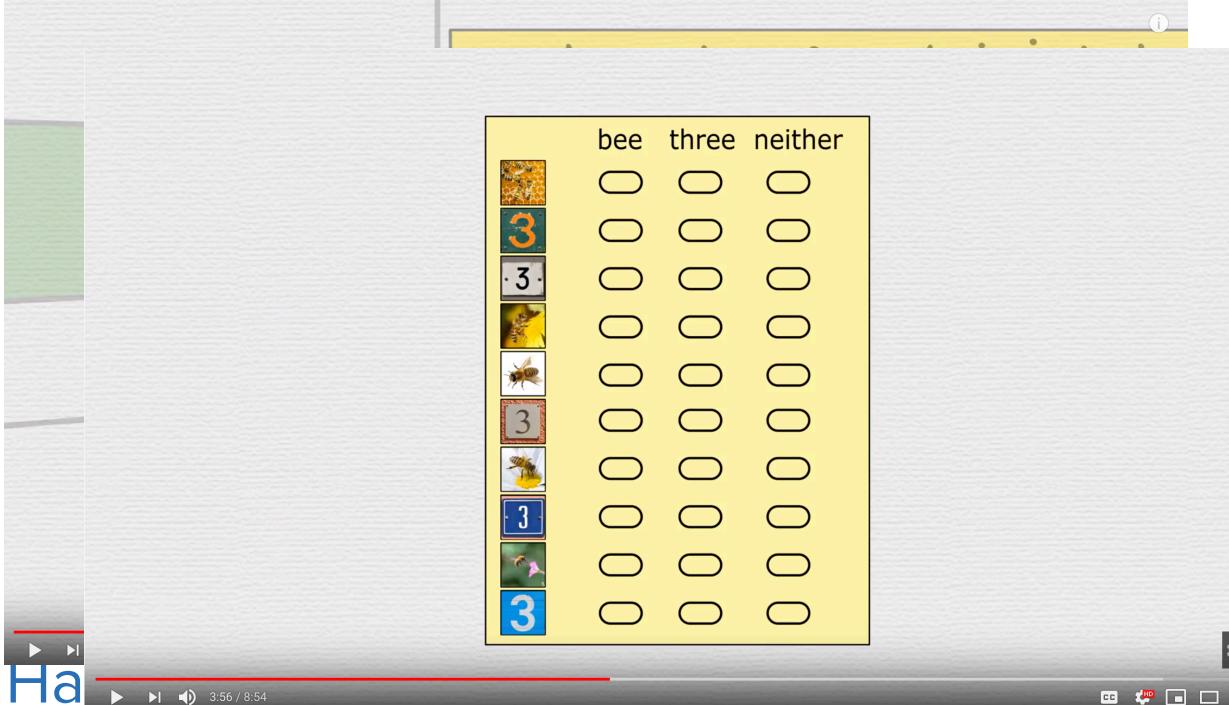
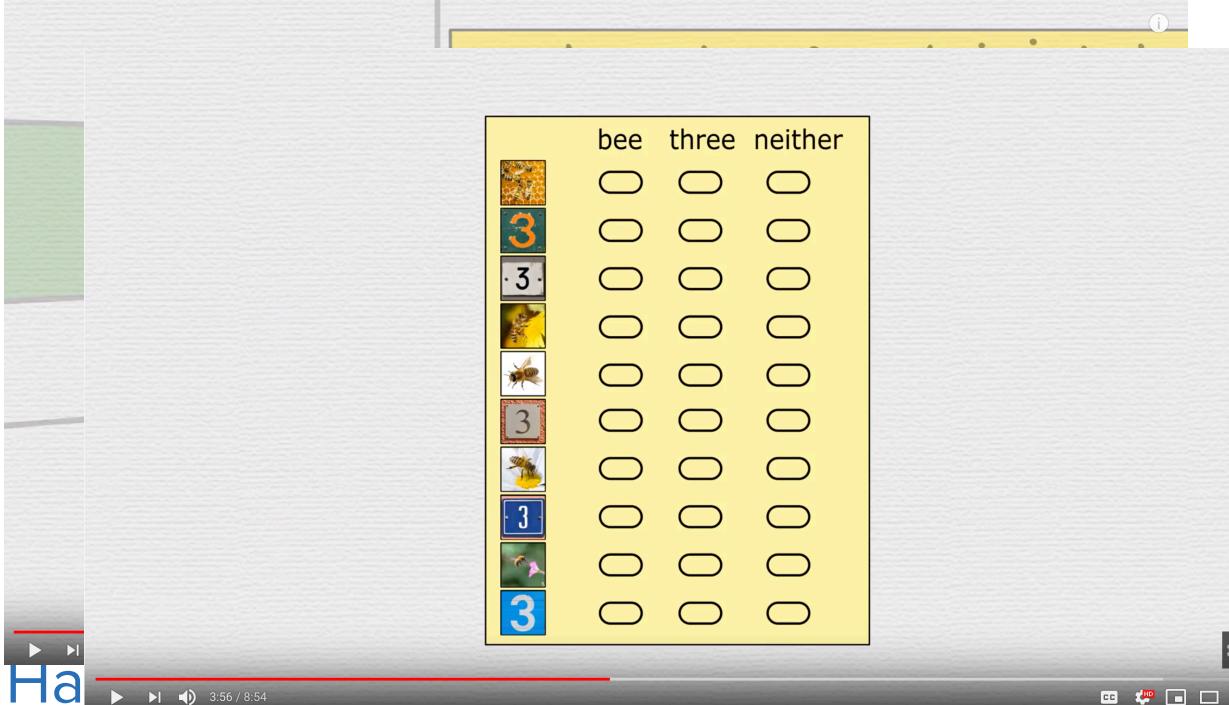
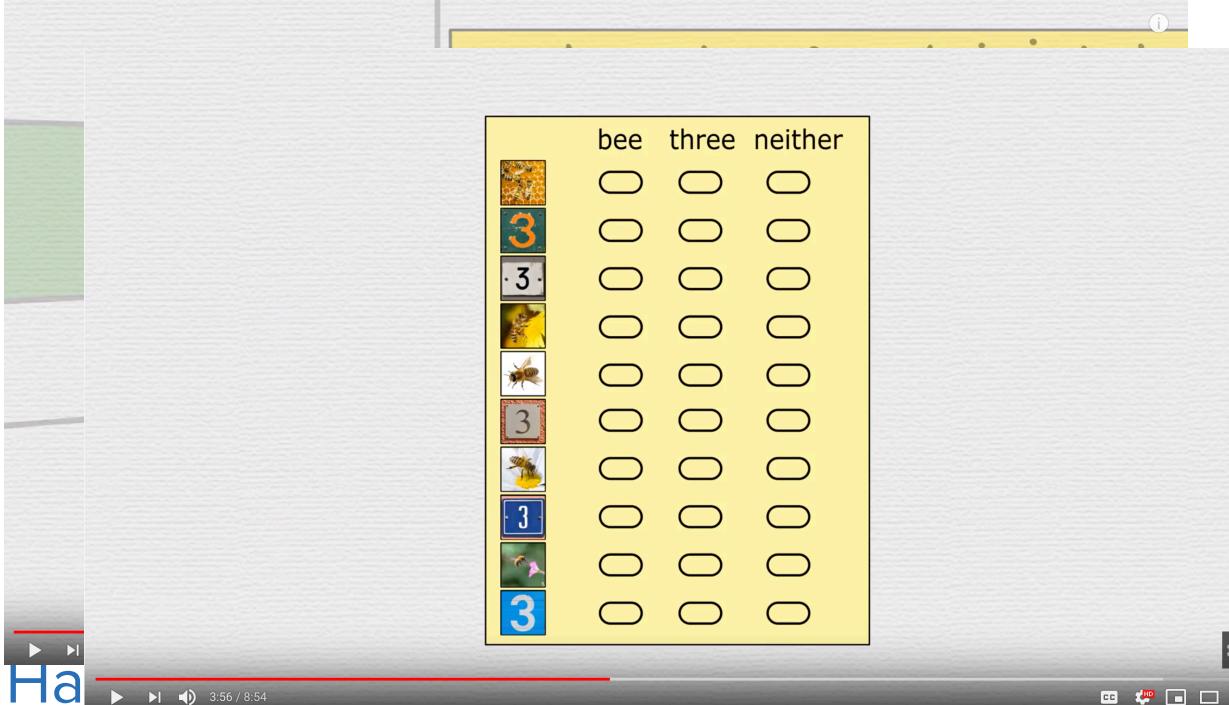
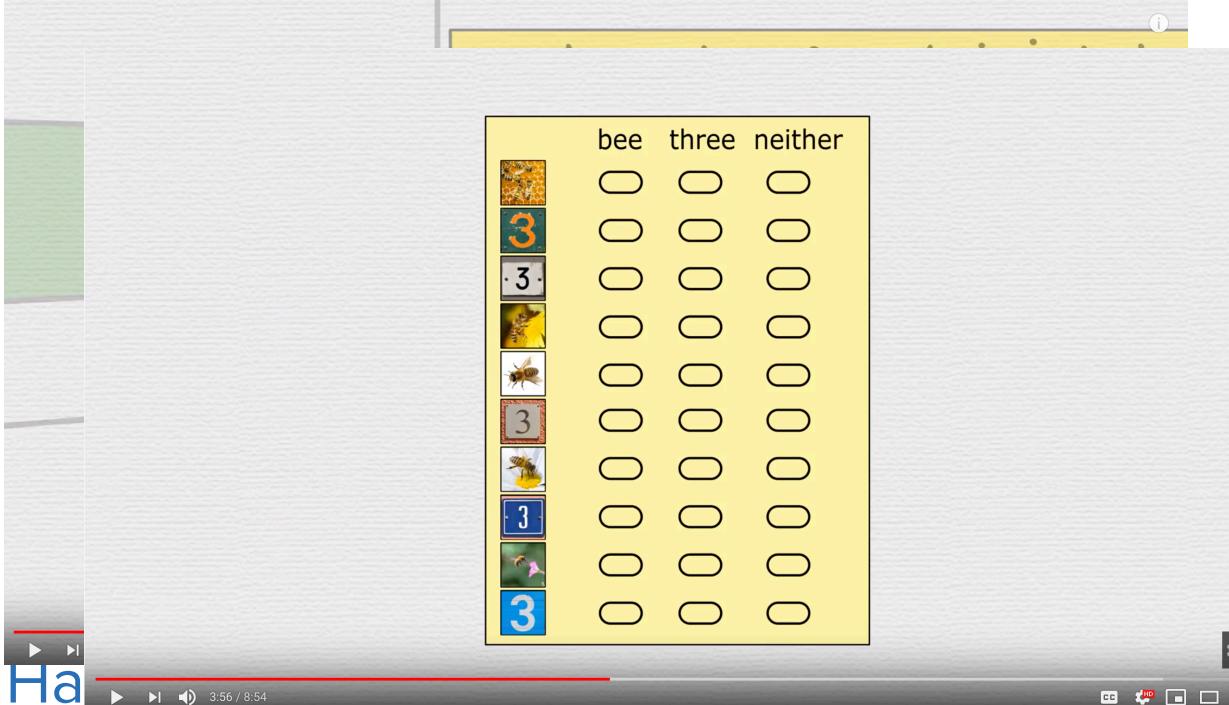
- Not a “magic bullet”
 - Predictions can be very different from the truth, even when advanced techniques are used
- Often requires massive training databases
 - Results are only as good as training data
 - Results also depend on what methods are trained to optimize
- Hard to interpret results...

AI and deep learning

- A cartoon illustration of a classroom. A green teacher character with a ruler and an apple stands on the left. Several white robot students with red eyes and an apple each sit at their desks. A yellow sign above the robots lists letters: Aa Bb Cc Dd Ee Ff Gg Hh Ii Jj Kk Ll Mm. The video player interface shows a progress bar at 3:50 / 8:54 and various control icons.
- **AI can find patterns in data that aren't there**
- **AI can learn to do things without being explicitly programmed**
- **Hard to interpret results...**

Stills from “How machines learn” by CGP Grey

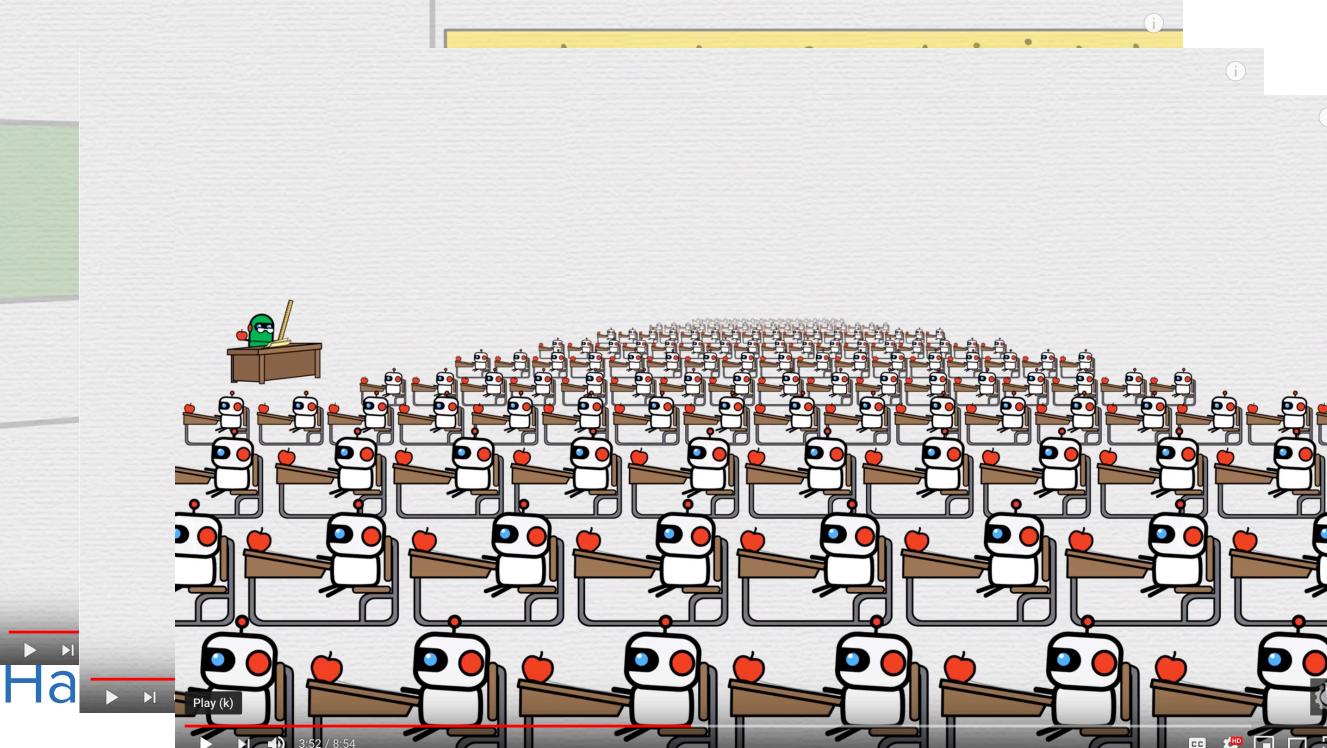
AI and deep learning

- 
- 
- 
- 
- 
- 
- 
- 
- 
- 

truth, even when advanced

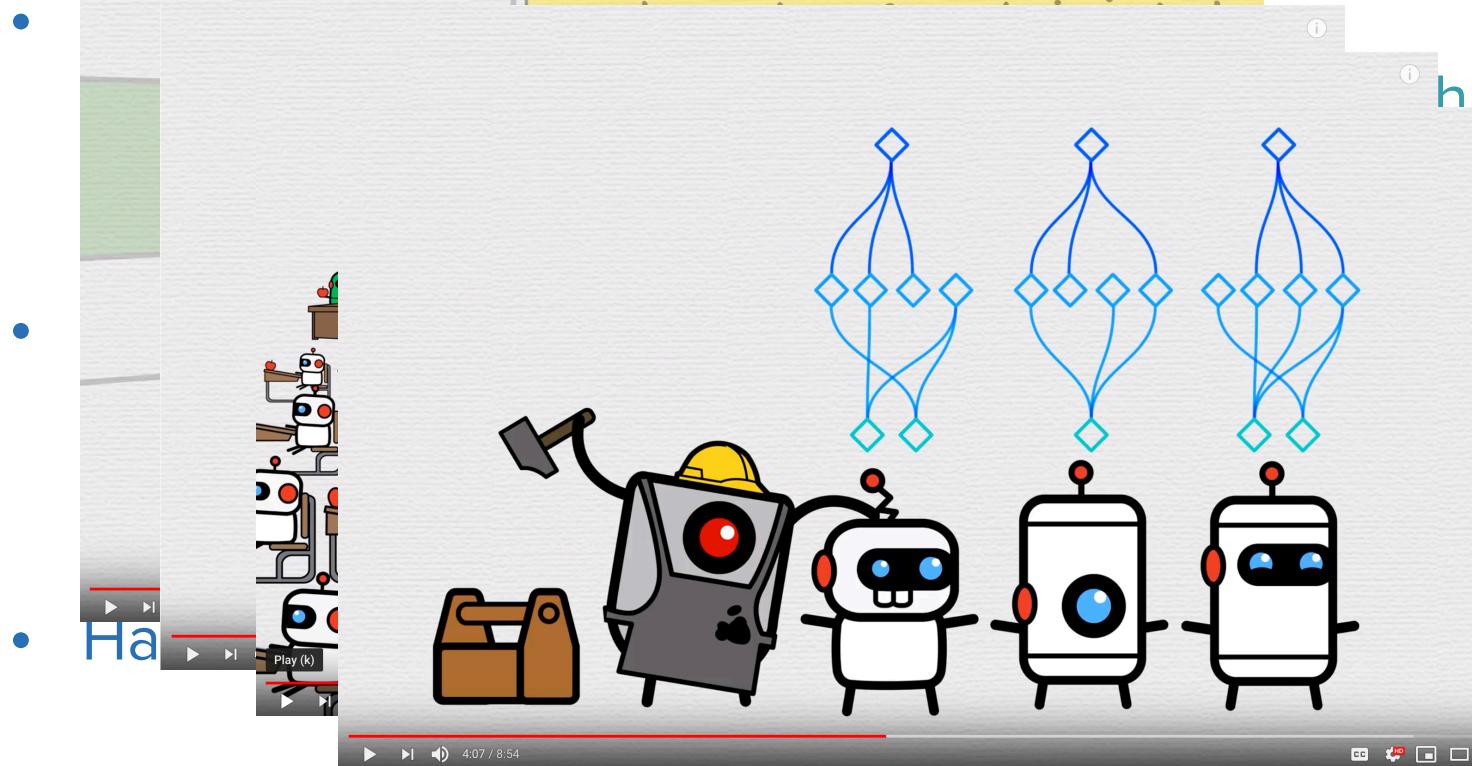
trained to optimize

AI and deep learning

-  Even simple AI can be effective, even when advanced
- Trained to optimize
- How

Stills from “How machines learn” by CGP Grey

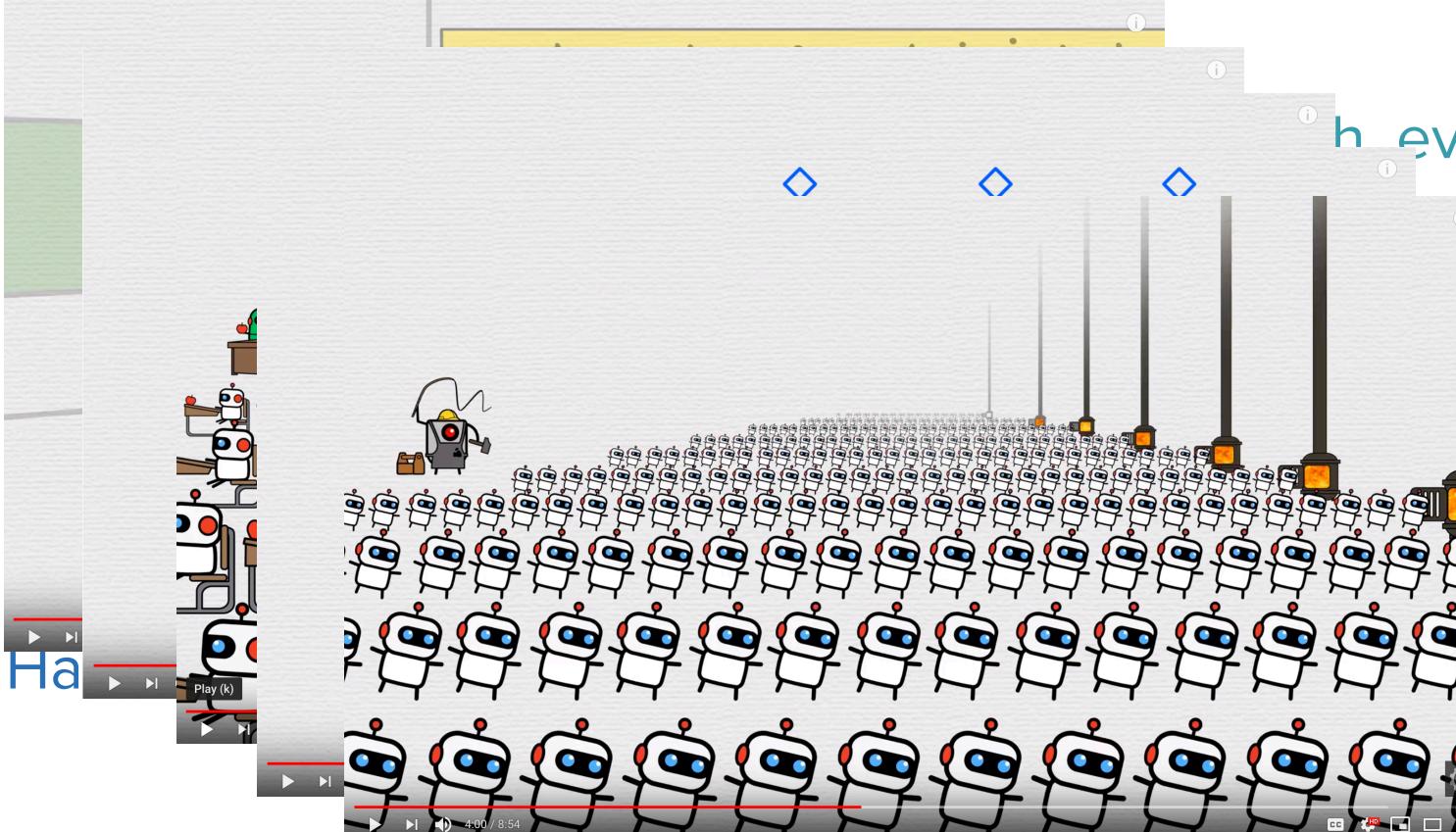
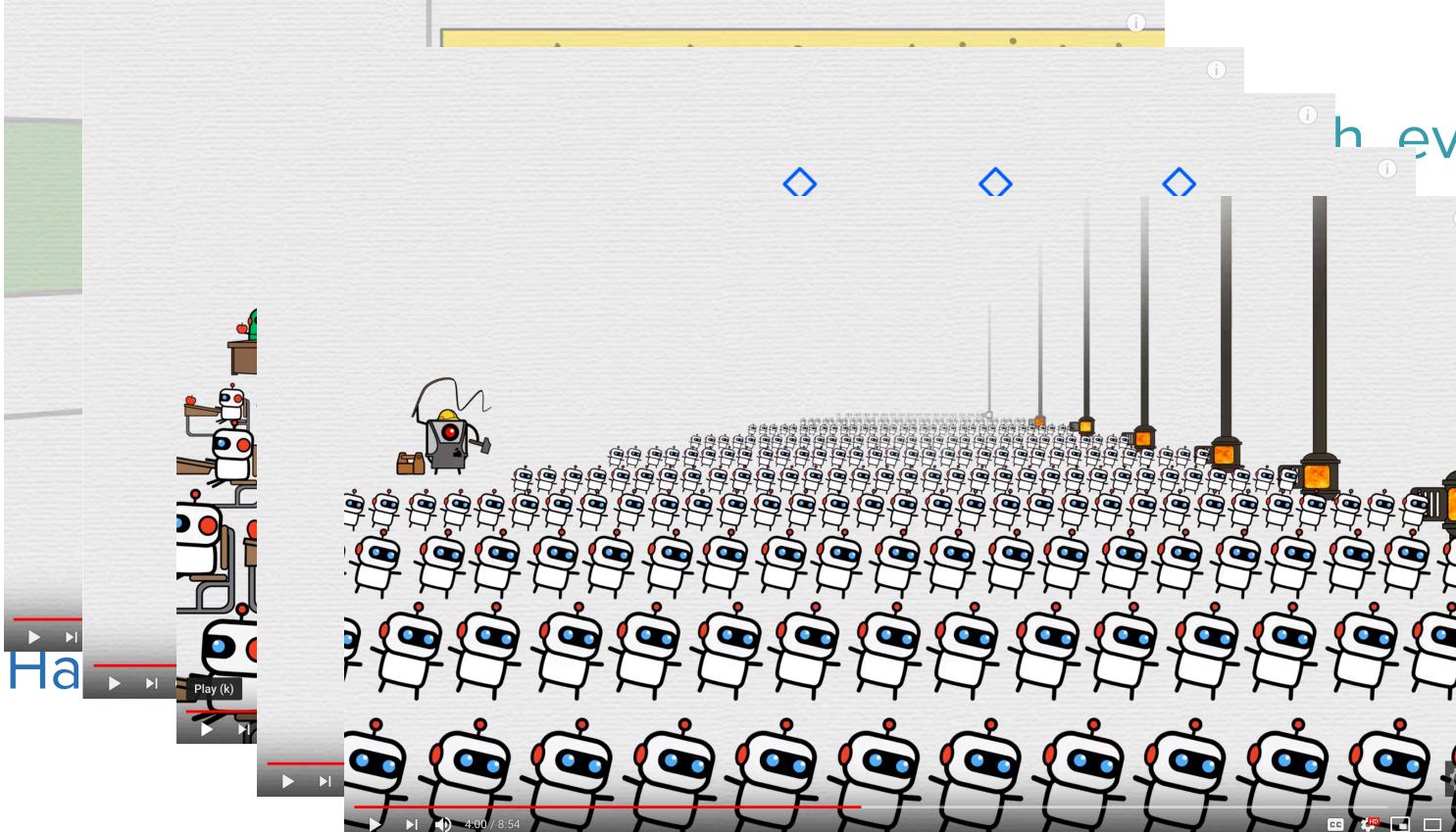
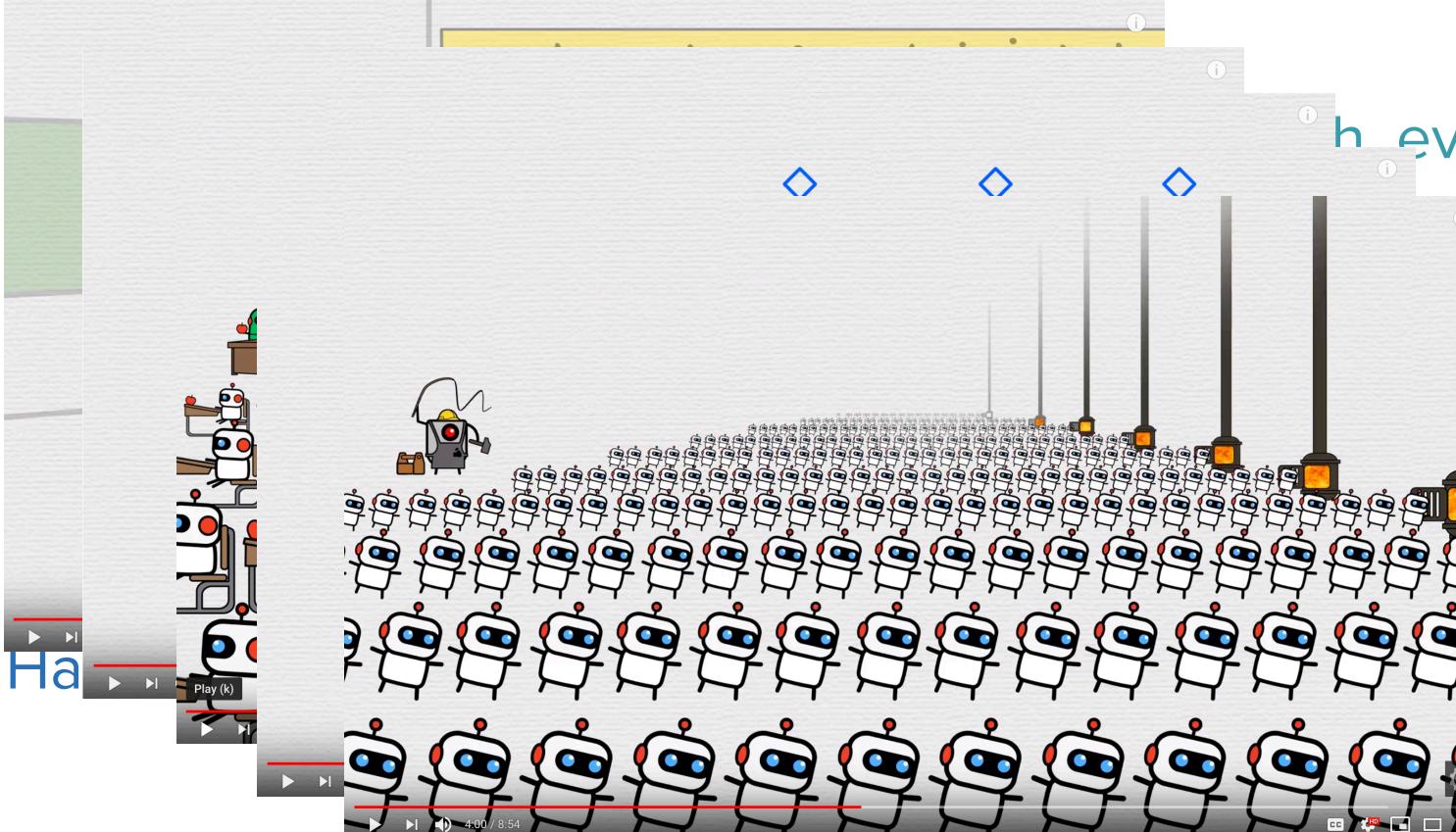
AI and deep learning



h even when advanced

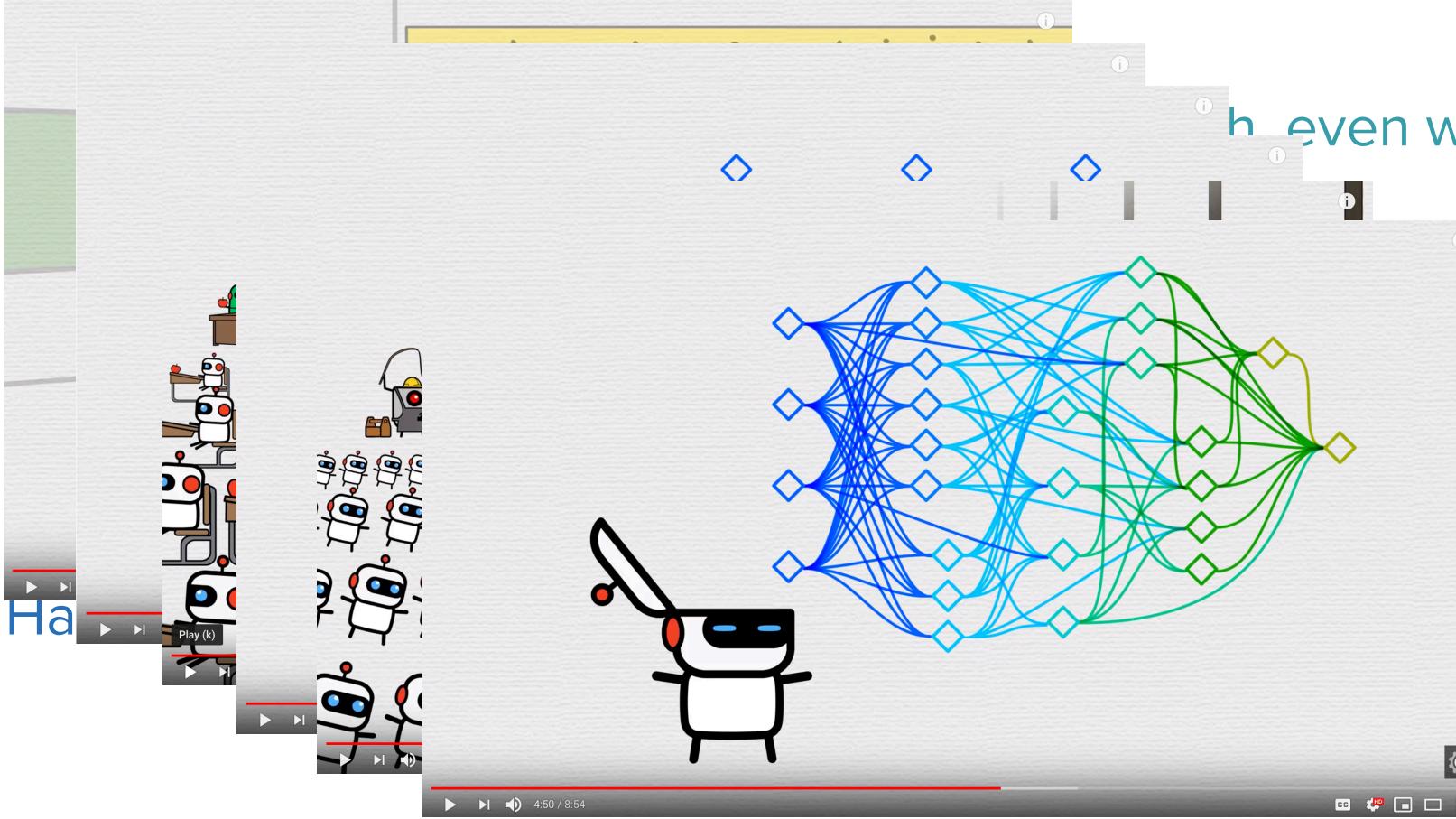
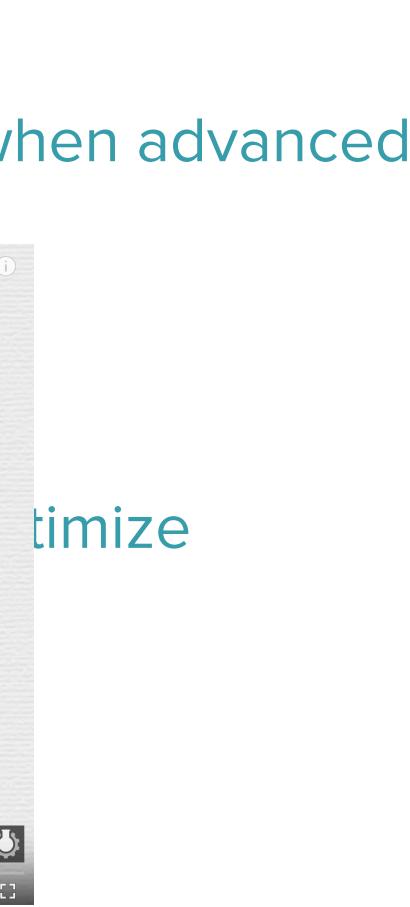
d to optimize

AI and deep learning

-  Even when advanced
-  To optimize
-  Ha

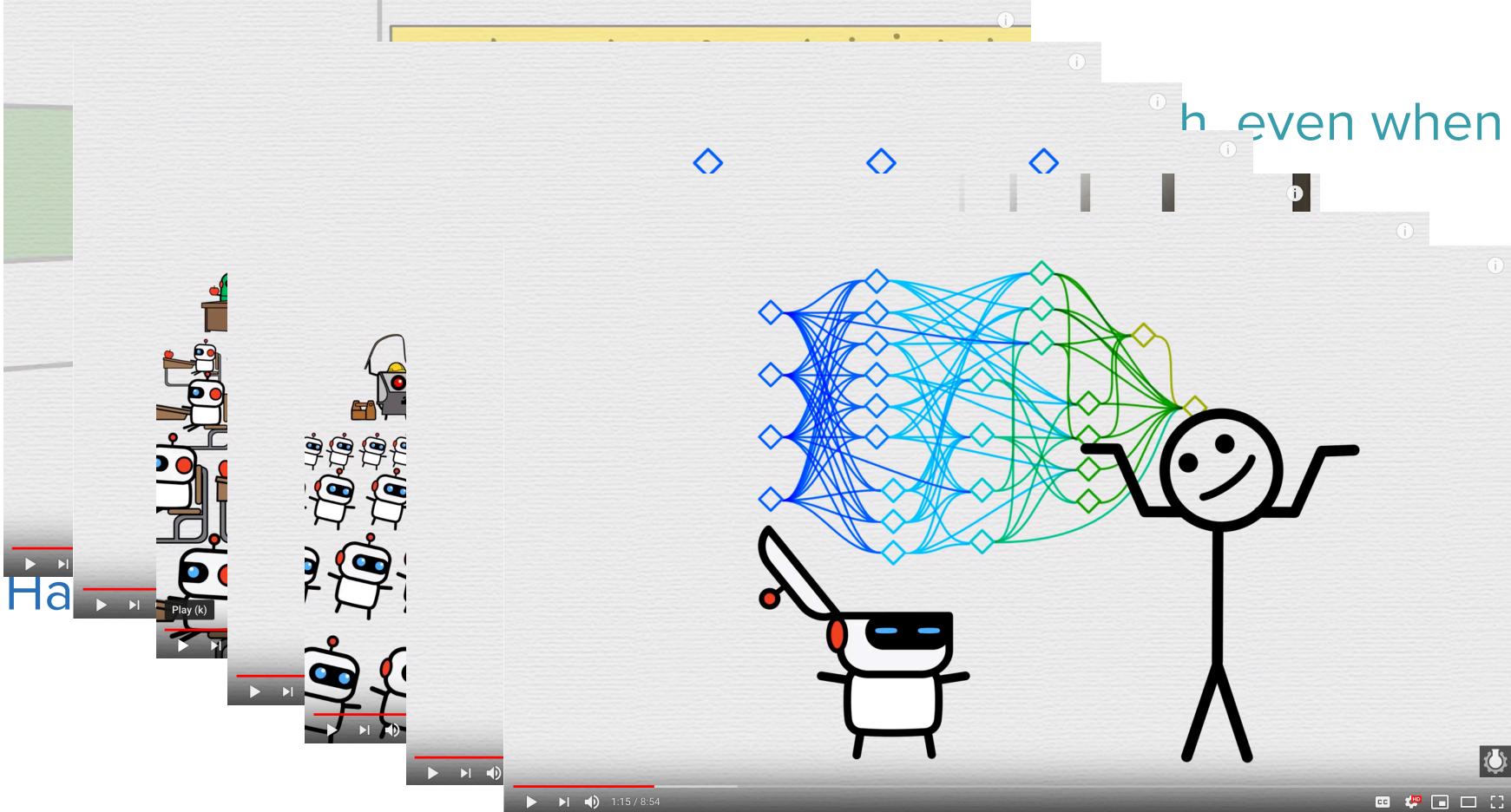
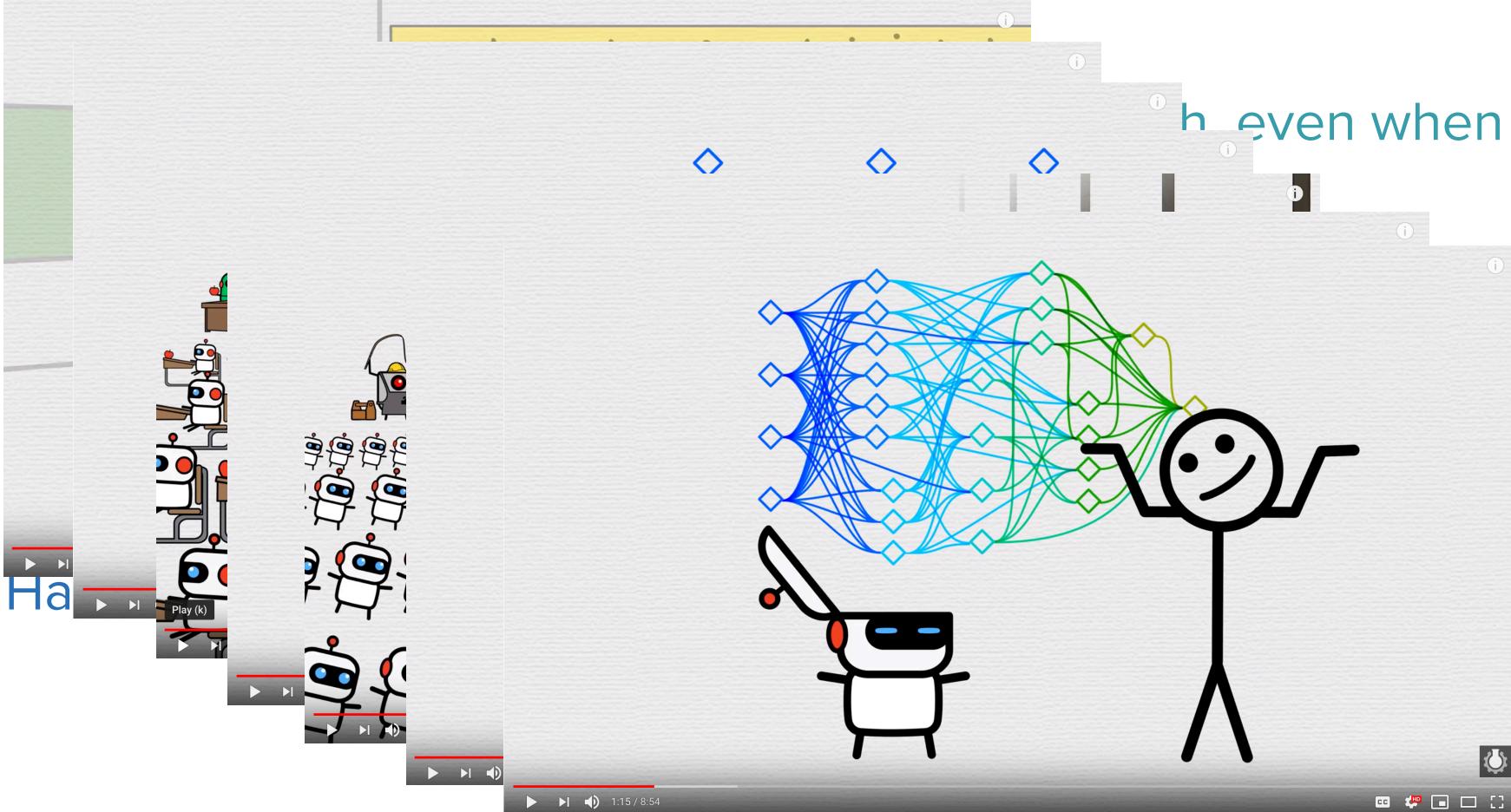
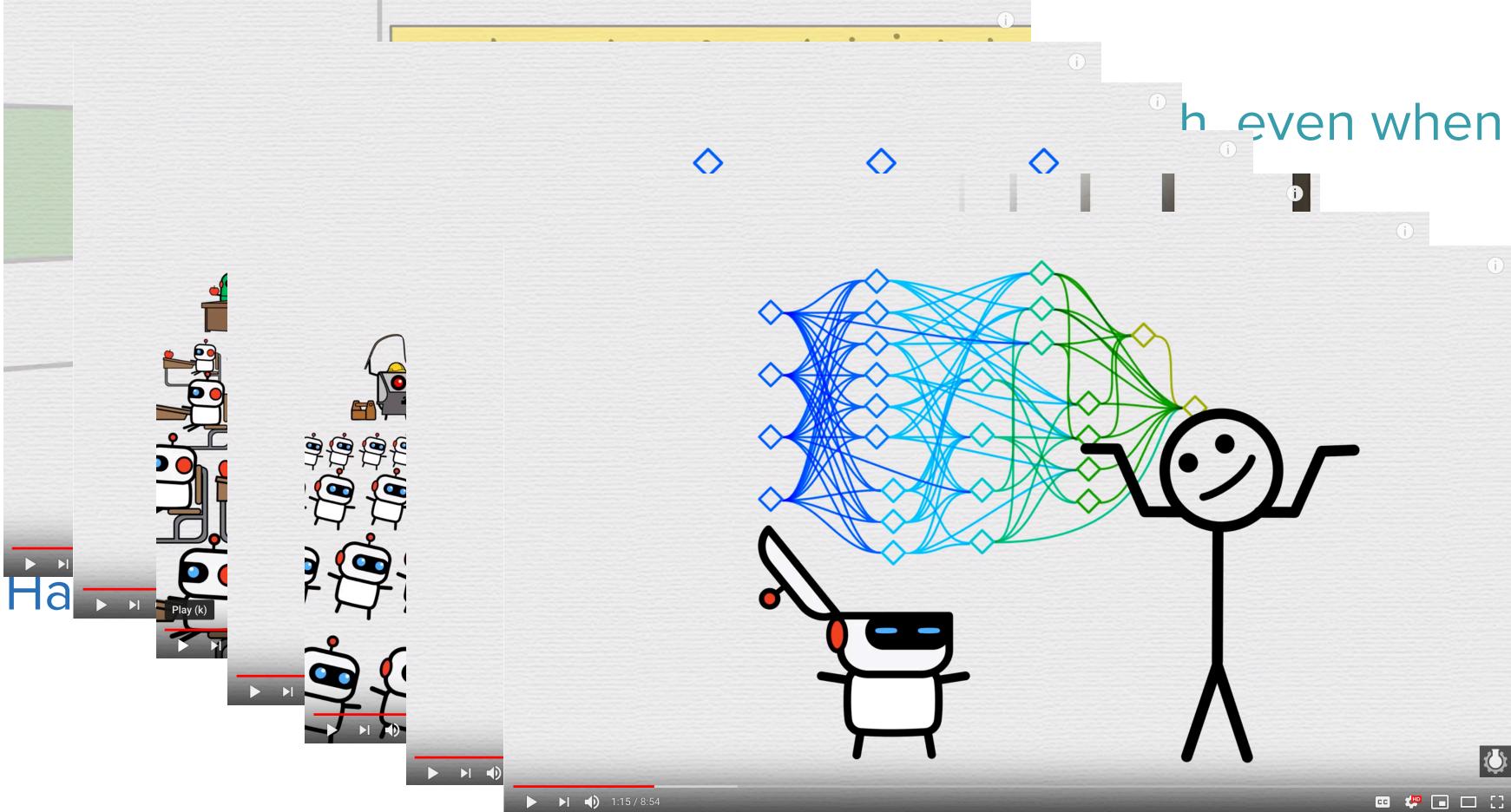
Stills from “How machines learn” by CGP Grey

AI and deep learning

-  Even when advanced
h, even when advanced
-  timize
-  Ha

Stills from “How machines learn” by CGP Grey

AI and deep learning

- A screenshot of a video player interface. The video frame shows a stick figure with a smiling face and a small robot. Above them is a complex neural network diagram with many blue and green nodes and connecting lines. In the background, there are several small robot icons. The video player has a progress bar at the bottom labeled '1:15 / 8:54'. On the left side of the video frame, there are three smaller video thumbnails. The top thumbnail shows a robot on a chair. The middle thumbnail shows a group of robots. The bottom thumbnail shows two robots. The video player has standard controls like play/pause, volume, and a settings icon.
- A screenshot of a video player interface. The video frame shows a stick figure with a smiling face and a small robot. Above them is a complex neural network diagram with many blue and green nodes and connecting lines. In the background, there are several small robot icons. The video player has a progress bar at the bottom labeled '1:15 / 8:54'. On the left side of the video frame, there are three smaller video thumbnails. The top thumbnail shows a robot on a chair. The middle thumbnail shows a group of robots. The bottom thumbnail shows two robots. The video player has standard controls like play/pause, volume, and a settings icon.
- A screenshot of a video player interface. The video frame shows a stick figure with a smiling face and a small robot. Above them is a complex neural network diagram with many blue and green nodes and connecting lines. In the background, there are several small robot icons. The video player has a progress bar at the bottom labeled '1:15 / 8:54'. On the left side of the video frame, there are three smaller video thumbnails. The top thumbnail shows a robot on a chair. The middle thumbnail shows a group of robots. The bottom thumbnail shows two robots. The video player has standard controls like play/pause, volume, and a settings icon.

Stills from “How machines learn” by CGP Grey

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 AI Race 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 Grid

How Google Translate, one of the world's most popular learning tools, can perpetuate gender bias

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
“he” with
“ng” and “she”
and other

Real talk about AI



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](#)

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

