



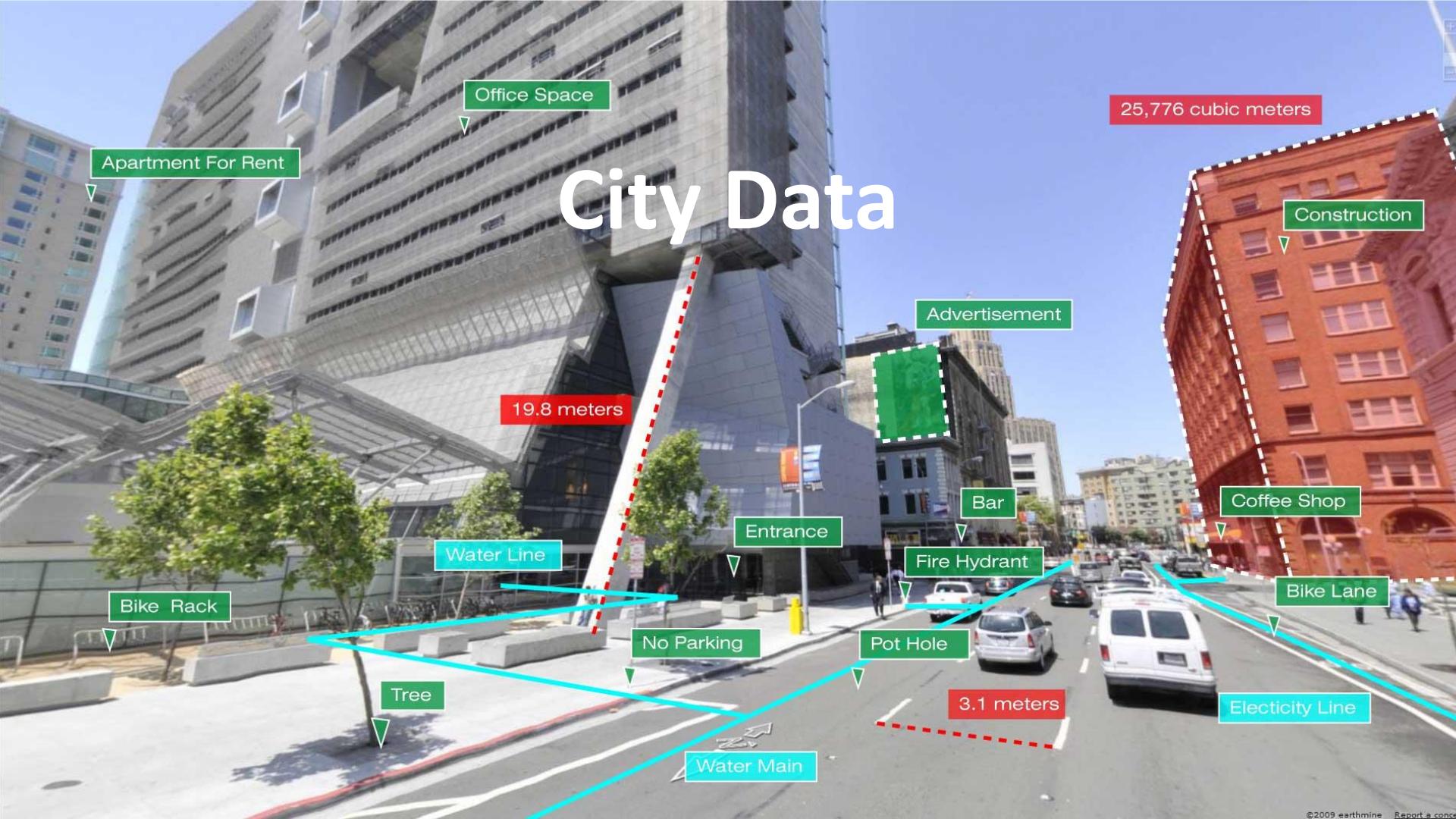
Applied Generative-AI for Digital Transformation

Democratization of Data and AI

Abel Sanchez, John R. Williams

MIT Campus Course | Cambridge, MA

DATA CHALLENGE



City Data

THE COMING FLOOD OF DATA IN AUTONOMOUS VEHICLES

RADAR
~10-100 KB
PER SECOND

SONAR
~10-100 KB
PER SECOND

GPS
~50KB
PER SECOND

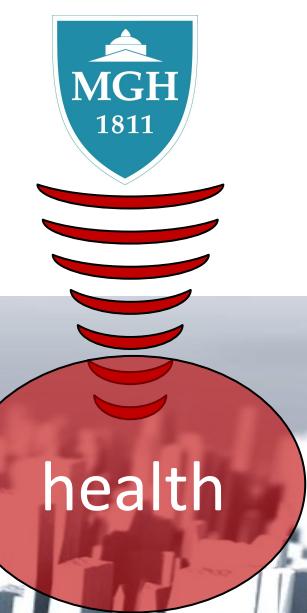
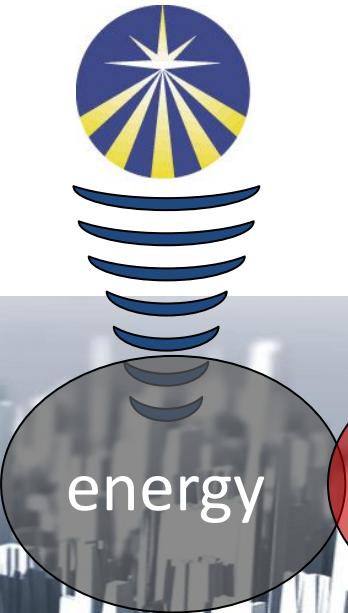
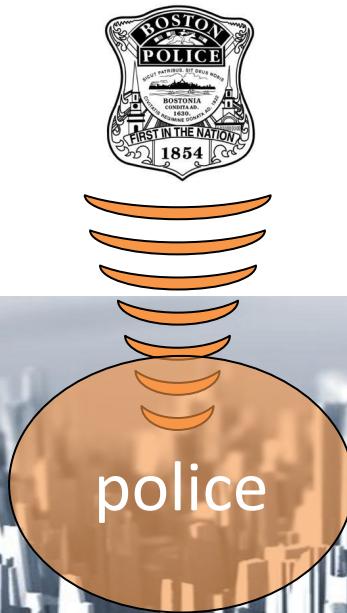
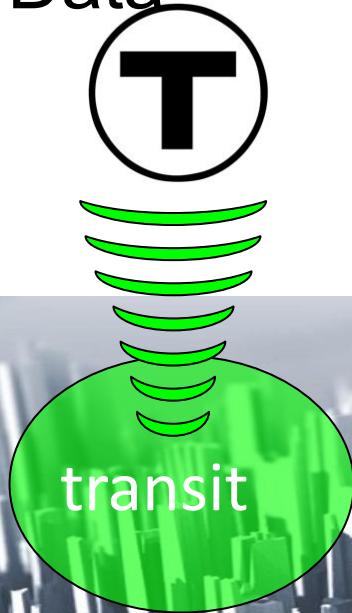
CAMERAS
~20-40 MB
PER SECOND

LIDAR
~10-70 MB
PER SECOND

AUTONOMOUS VEHICLES
4,000 GB
PER DAY... EACH DAY



City Data





DATA





DATA

REPORTS

OPEN GOVERNMENT

CONTACT



User Guide

For details on Data.gov's new home page redesign, please click [here](#).

The Home of the U.S. Government's Open Data

Here you will find data, tools, and resources to conduct research, develop web and mobile applications, design data visualizations, and more.

258,224 DATASETS AVAILABLE



Search

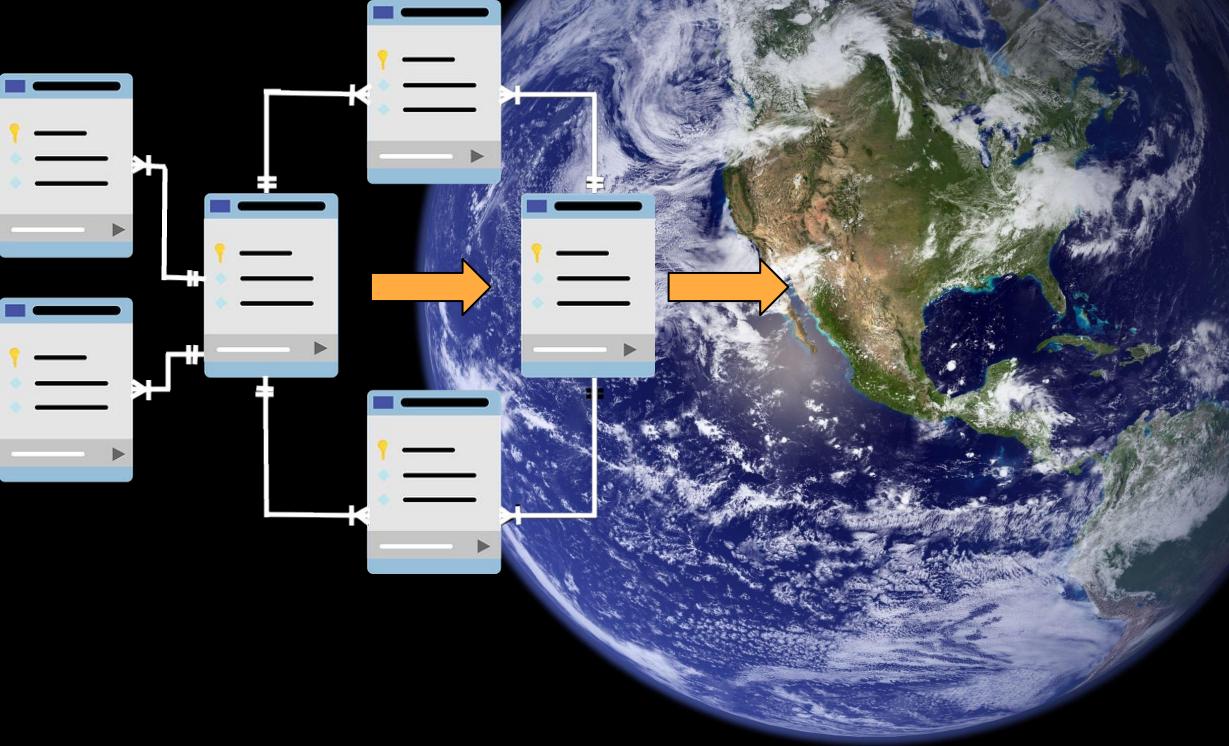


In the U.S., length of bill to congress

- 2,500 pages, The Affordable Care Act
 - 17,000 pages of regulations
- 2,300 pages, Dodd-Frank Act
 - 8,843 pages of regulations



Who is going to read it?



Forcing a
data model
onto the
world

The background of the slide features a grayscale photograph of a large stack of papers or documents, creating a textured, layered effect. The text is overlaid on this image.

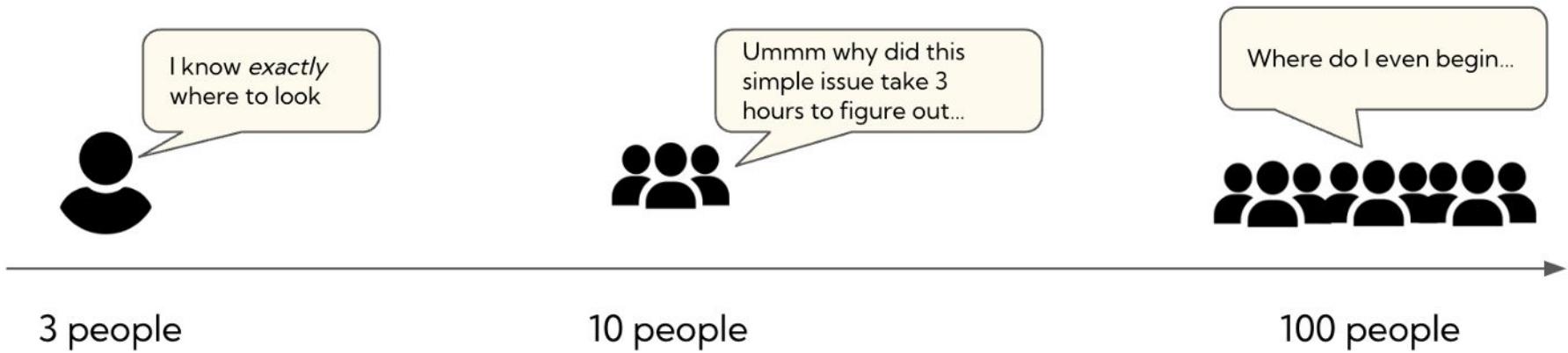
OPPORTUNITY: DATA DEMOCRATIZATION



We have failed to democratize data access. We have failed to realize the data driven corporation.



The data challenge



Common scenario



Idealism



Discovery



Procurement



Integration



Abandonment

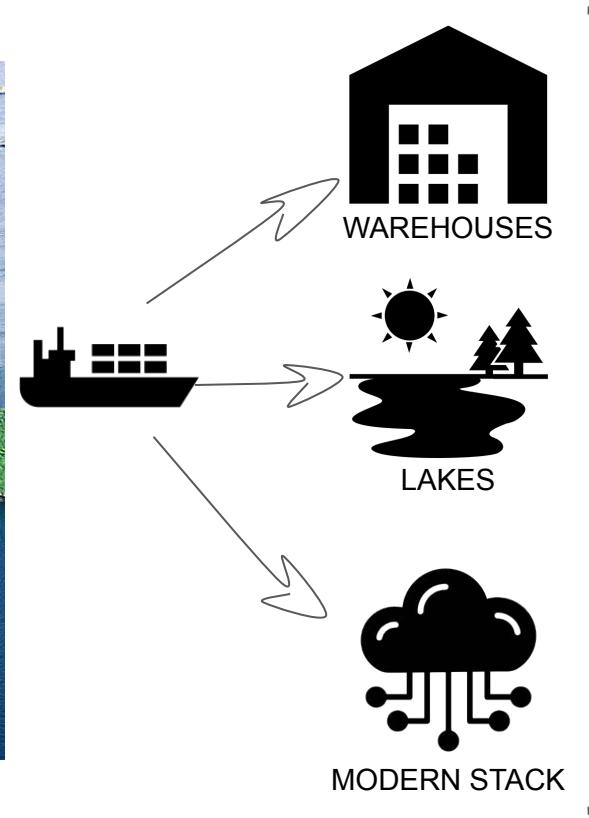
Result: we do not know what we know



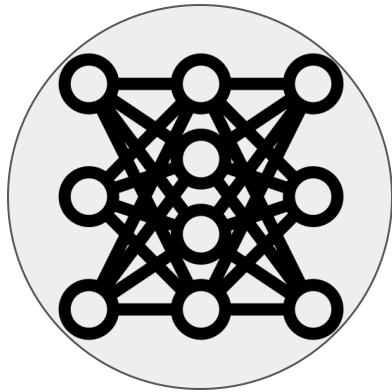
We have tried to integrate



Has not worked!



How do we reason over data at scale? We have difficulty.
So far, we have failed to democratize data.



LLMs are a new opportunity

Simple database

Students	
	StudentID
	CollegeID
	FirstName
	LastName
	BirthDate
	Email
	City
	Region
	Country

Colleges	
	CollegeID
	Name
	Students
	City
	Region
	Country

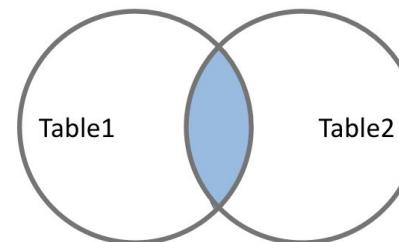
```
1 -- -----
2 -- TABLE COLLEGES
3 --
4
5 CREATE TABLE `colleges` (
6     `collegeID`      int NULL,
7     `name`           varchar (20) NULL,
8     `students`       int NULL,
9     `city`           varchar (15) NULL ,
10    `region`         varchar (15) NULL ,
11    `country`        varchar (15) NULL
12 );
13
14 --
15 -- TABLE STUDENTS
16 --
17
18 CREATE TABLE `students` (
19     `studentID`      int NULL,
20     `collegeID`      int NULL,
21     `firstName`      varchar (20) NULL ,
22     `lastName`       varchar (20) NULL ,
23     `birthDate`      date NULL ,
24     `email`          varchar (30) NULL ,
25     `city`           varchar (15) NULL ,
26     `region`         varchar (15) NULL ,
27     `country`        varchar (15) NULL
28 );
29
```

```
30 --
31 -- POPULATE COLLEGES
32 --
33
34 INSERT INTO `colleges` VALUES (1,'MIT',11,'Cambridge','MA','USA'),
35 (2,'Brown',9,'Providence','RI','USA'),
36 (3,'Dartmouth',6,'Hanover','NH','USA'),
37 (4,'Stanford',17,'Stanford','CA','USA'),
38 (5,'Yale',12,'New Haven','CT','USA'),
39 (6,'Columbia',31,'New York','NY','USA'),
40 (7,'Harvard',23,'Cambridge','MA','USA'),
41 (8,'Princeton',9,'Princeton','NJ','USA'),
42 (9,'Johns Hopkins',24,'Baltimore','MD','USA'),
43 (10,'Northwestern',21,'Evanston','IL','USA'),
44
45 --
46 -- POPULATE STUDENTS
47 --
48
49 INSERT INTO `students` VALUES (1,1,'Nancy','Davolio','1948-12-08',
50 (2,9,'Andrew','Fuller','1952-02-19','andrew@yahoo.com','Dallas','T
51 (3,8,'Janet','Leverling','1963-08-30','janet@hotmail.com','Miami','P
52 (4,3,'Margaret','Peacock','1937-09-19','maggie@outlook.com','Phoenix','M
53 (5,4,'Steven','Buchanan','1955-03-04','steve@apple.com','Denver','M
54 (6,7,'Michael','Suyama','1963-07-02','mike@icloud.com','Portland','M
55 (7,6,'Robert','King','1960-05-29','rob@gmail.com','San Francisco','M
56 (8,5,'Laura','Callahan','1958-01-09','laura@gmail.com','Memphis','M
57 (9,2,'Anne','Dodsworth','1966-01-27','anne@msn.com','Portland','M
58 (10,1,'Ivy','Johnson','1986-01-20','ivy@gmail.com','Chicago','IL')
```

Structure Query Language (SQL)

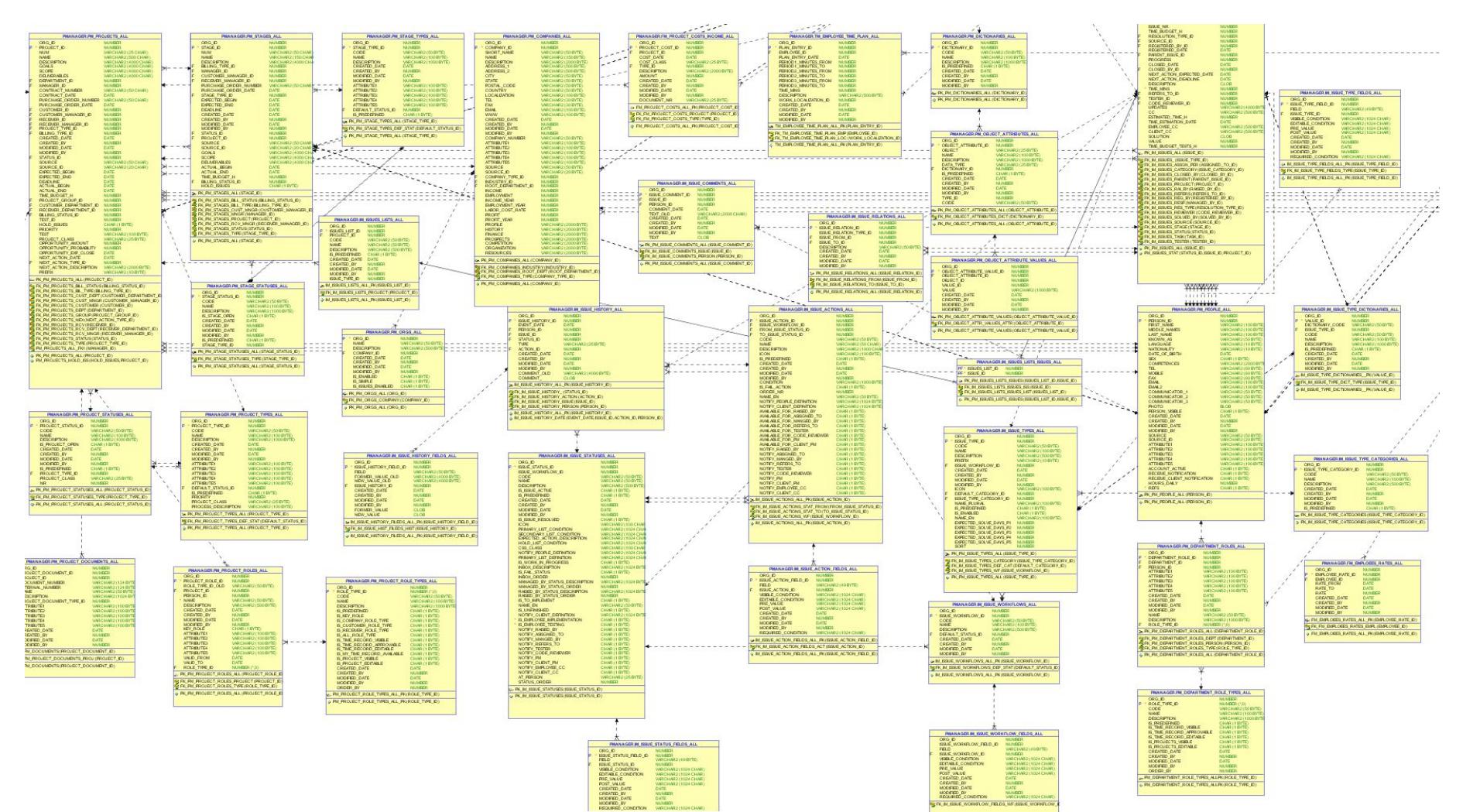
INNER JOIN

```
SELECT table1.column, table2.column  
FROM table1  
INNER JOIN table2  
ON table1.column = table2.column
```

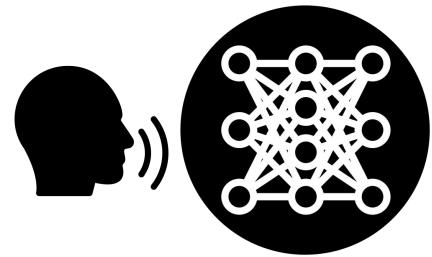


Active Learning - Talk to your data

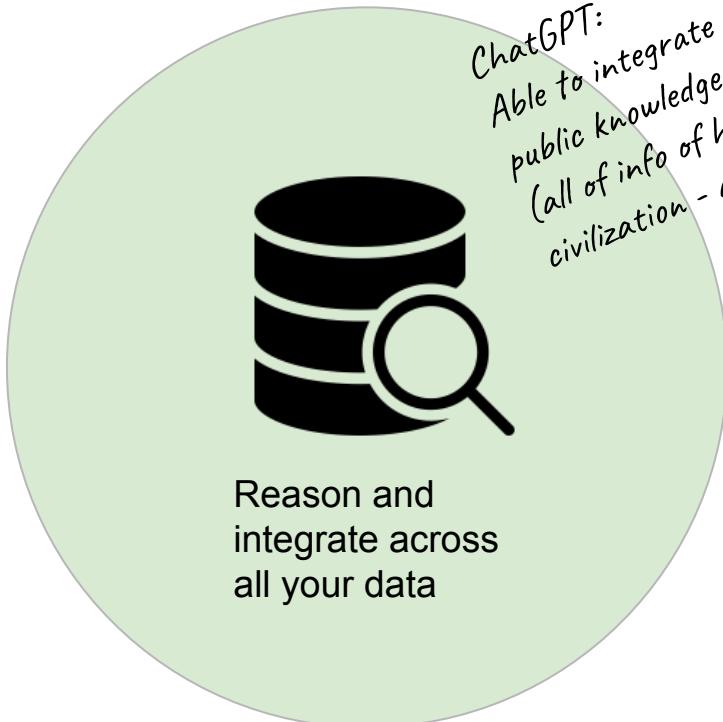




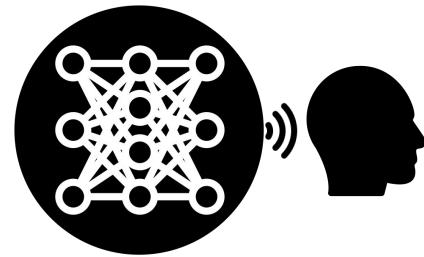
Generative-AI is a new opportunity



Ask question in
plain language



Reason and
integrate across
all your data



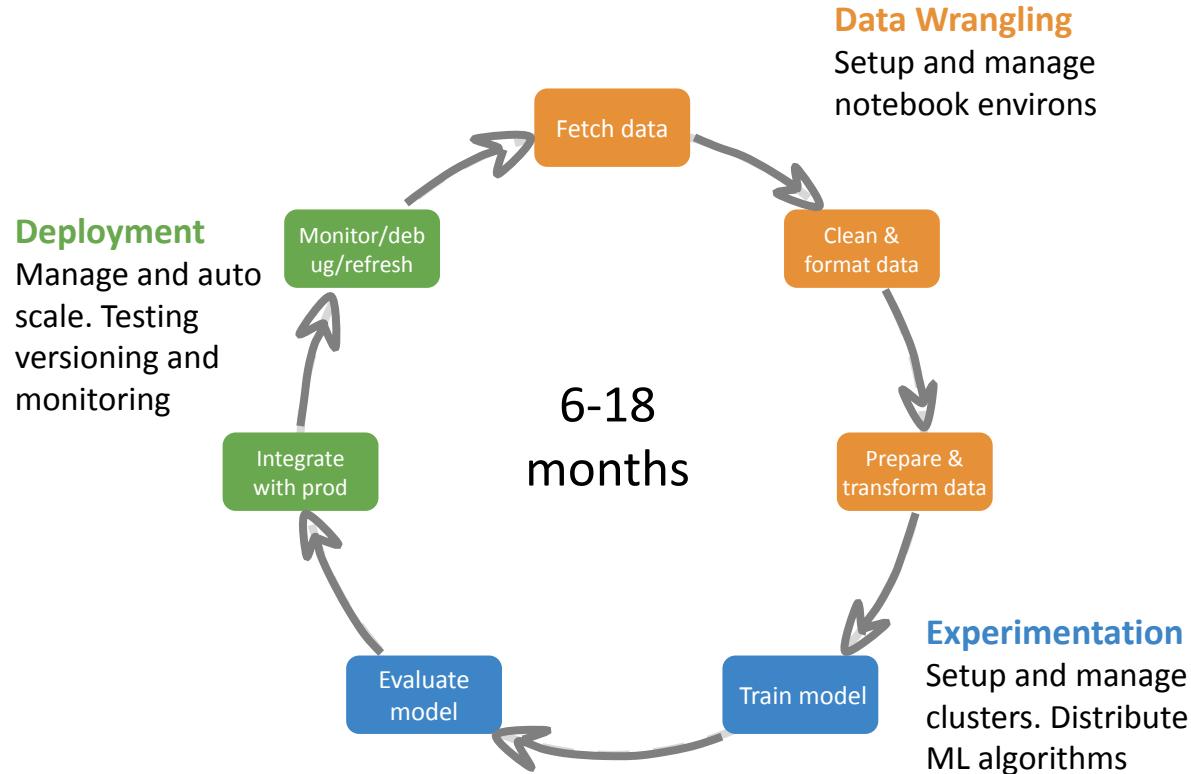
Get response in
plain language



OPPORTUNITY: AI DEMOCRATIZATION



Data Science Cycle

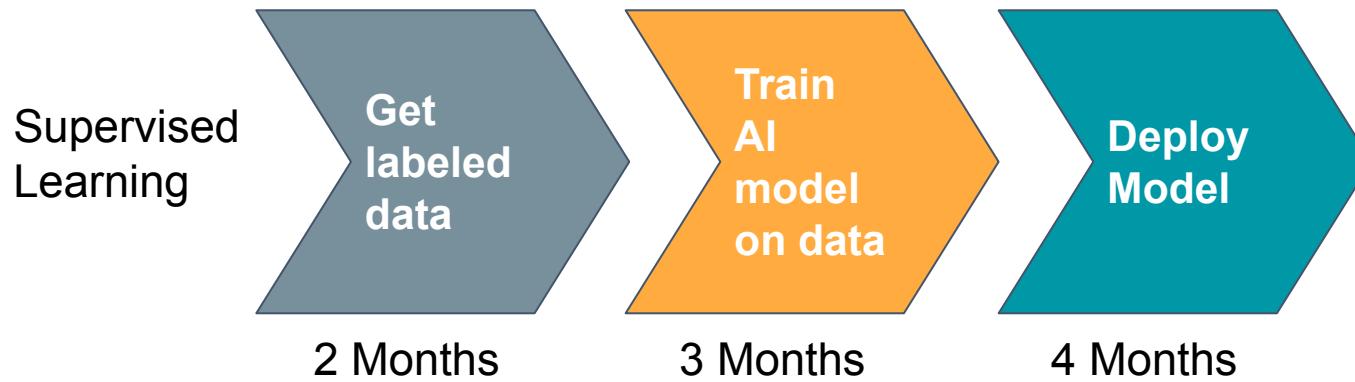


Barriers: cost/access/time

- Access: not everyone has **access** to talent, AI engineers.
- Cost: not everyone has the **budget**. Current AI engineer offers have surpassed \$900K.
- Time: not everyone has the **time**. It takes 6-to-18 months.



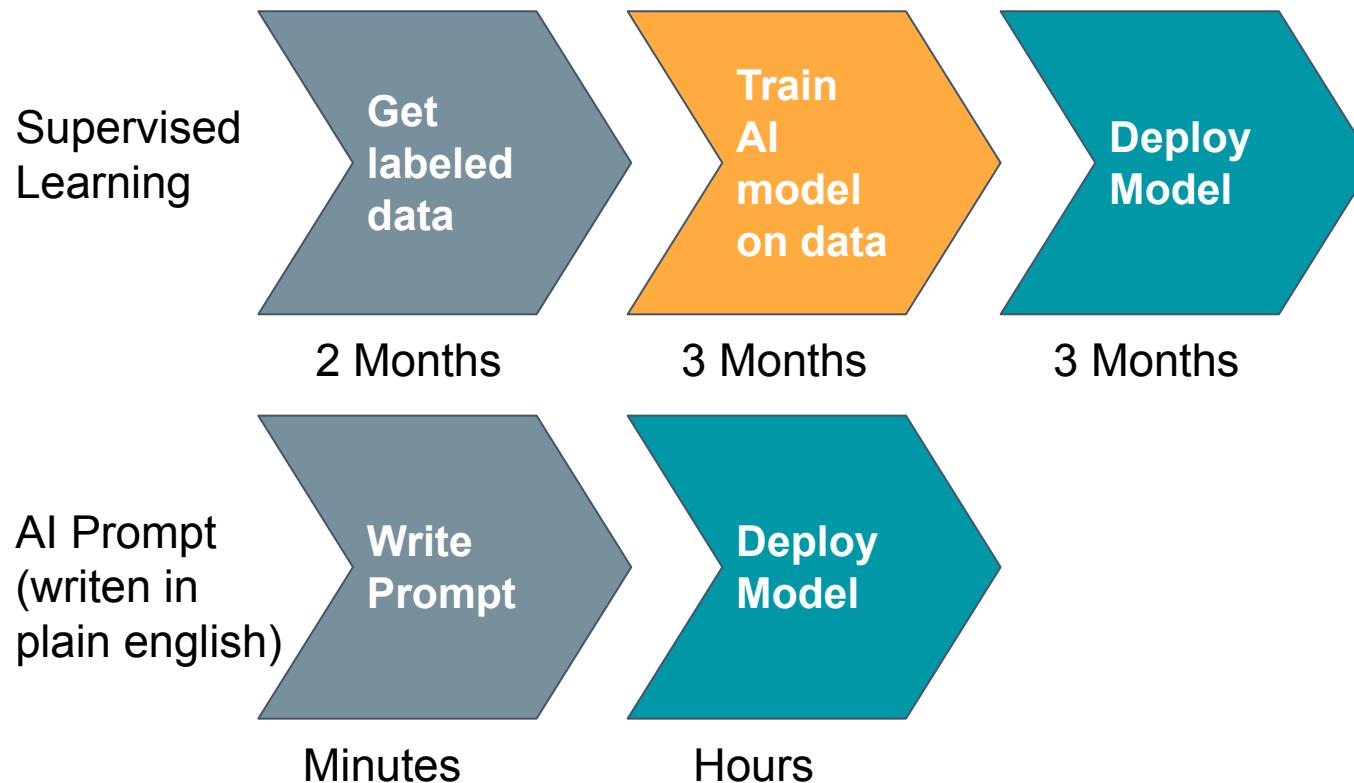
Prompting AI Applications - Example sentiment analysis



Active Learning: Prompting for development



Prompting AI Applications - Example sentiment analysis



A different way to build AI applications

 **Andrej Karpathy** 
@karpathy · [Follow](#) 

The hottest new programming language is English

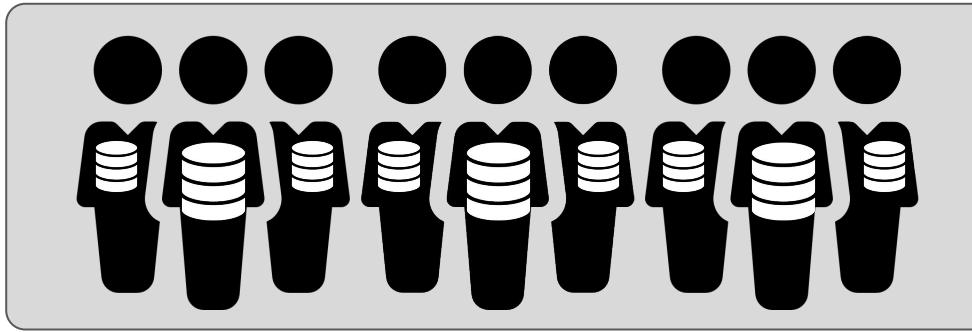
3:14 PM · Jan 24, 2023 

 [Read the full conversation on Twitter](#)

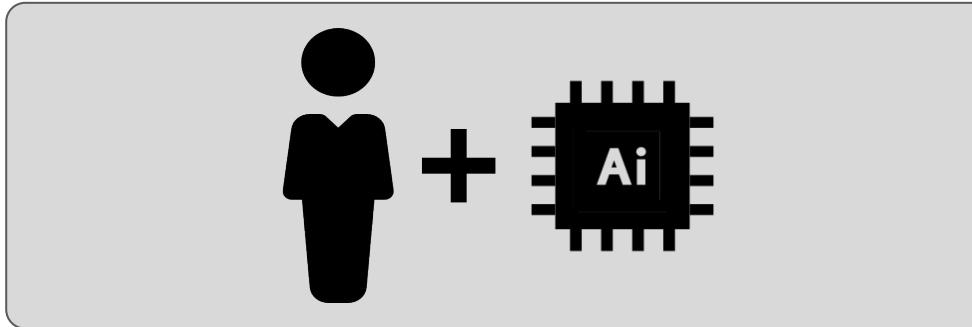
 2.4M Views  Reply  Copy link

\$10 Million AI on a \$2 Budget

The past
(the long cycle)



The opportunity
(the short cycle)

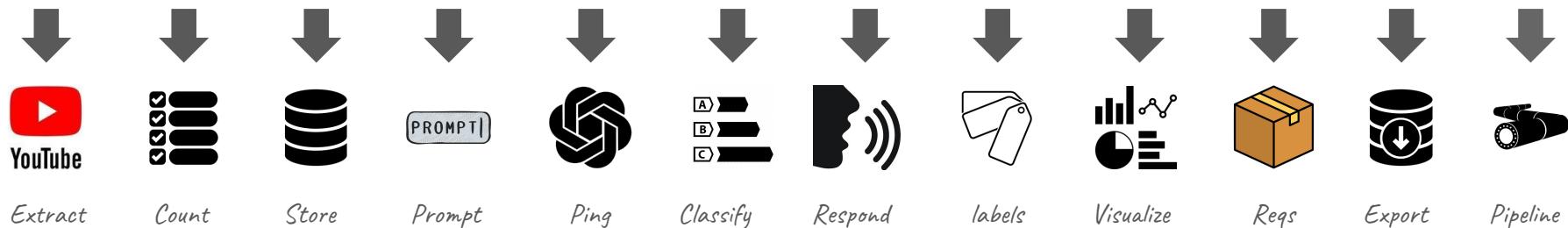


Converting an AI problem to a data engineering problem

"In-context learning", involves using LLMs off-the-shelf, without any fine-tuning, and controlling their behavior through clever prompting and conditioning on private contextual data. This approach effectively reduces an **AI problem to a data engineering problem, which most startups and big companies already know how to solve.**

- No need for a specialized team of ML engineers
- No need to host infrastructure
- Converts AI problem to a data engineering problem
- Many companies and startups understand DE

Sample moderation - LLM App



Influencer: Your latest video got 3M+ views and 10,000+ comments.

For each comment, you want to know:

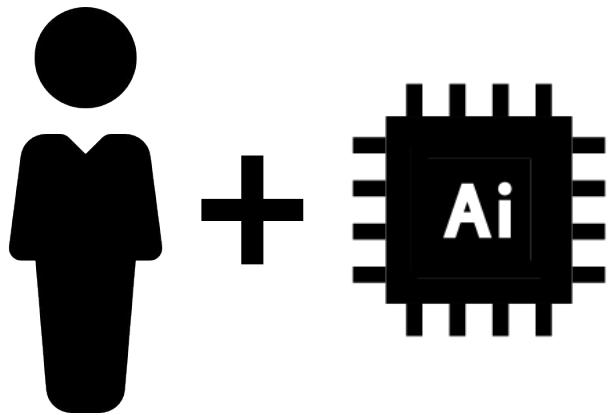
- Positive or negative
- Angry - some to worry about?
- Spam
- Requires response

Economics of Human vs Gen-AI

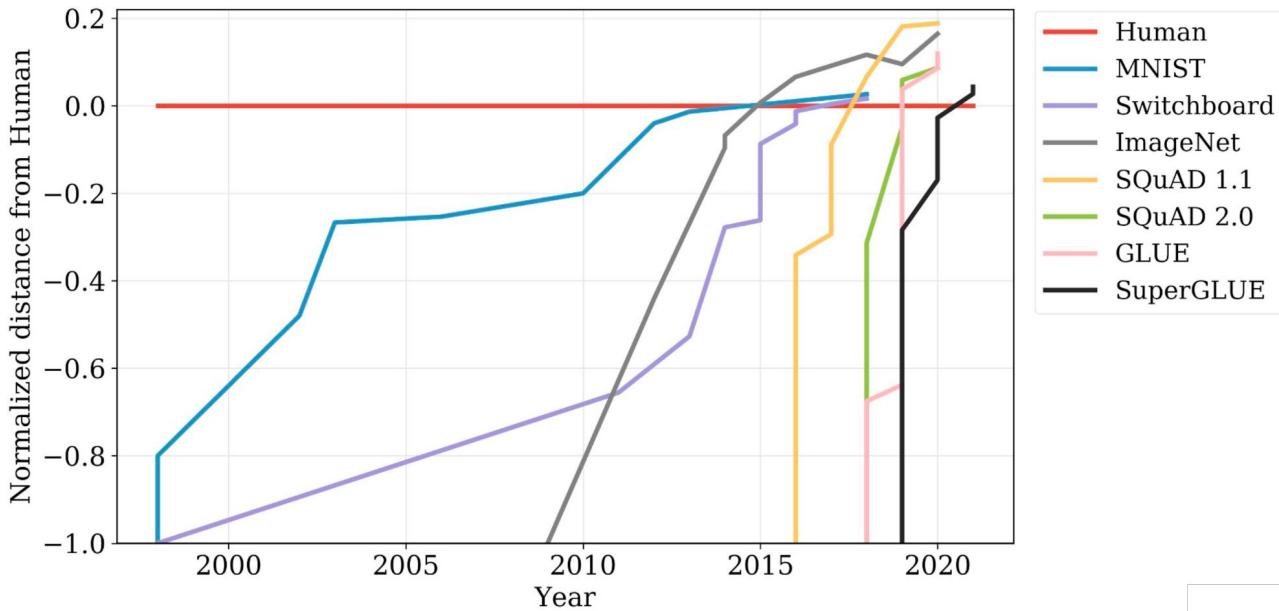
- Graphic artist image
 - Gen-AI cost: ~\$.001, takes ~1 second
 - Human cost: \$100s minimum, takes hours or days
 - 100,000 times cheaper (assuming \$100)
 - 3,600 times faster
- Lawyer, summarize and answer questions on a complex legal brief
 - Gen-AI cost: ~\$.001, takes ~10 second
 - Human cost: \$100s min, takes hours or days
- Therapist
 - Gen-AI cost: ~\$.001, takes ~10 second
 - Human cost: \$100s, takes 1 hour
 - Availability: scalable and global

The augmented human

Not human replacement,
human enhancement.
Knowledge is no longer
limited to experts.
You too can be X.



Surpassing human level performance faster



Digital Humans

The Best That Never Was

Marcus Dupree, had an explosive start to his college football career at the University of Oklahoma. His story is famously documented in ESPN's "30 for 30" series, titled "The Best That Never Was," highlighting the potential professional earnings and impact on the sport that went unfulfilled.



The Shame of College Sports

A litany of scandals in recent years have made the corruption of college sports constant front-page news. We profess outrage each time we learn that yet another student-athlete has been taking money under the table. But the real scandal is the very structure of college sports, wherein student-athletes generate billions of dollars for universities and private companies while earning nothing for themselves. Here, a leading civil-rights historian makes the case for paying college athletes—and reveals how a spate of lawsuits working their way through the courts could destroy the NCAA.



Photograph by Evan Kafka for The Atlantic

CULTURE

THE SHAME OF COLLEGE SPORTS

A leading civil-rights historian makes the case for paying college athletes—and reveals how a spate of lawsuits working their way through the courts could destroy the NCAA.

By Taylor Branch

Name, Image, and Likeness (NIL)

In July 2021, they adopted interim rules permitting student-athletes the ability to benefit from their name, image and likeness, also known as “NIL.” This was an unprecedented move by the NCAA, which had historically prohibited athletes from receiving any compensation in connection with their NIL.



Changing the game

Iowa basketball star Caitlin Clark, who's never had a full-time job, is worth over \$3 million—and she's changing the financial playbook for college athletes. The high-scoring guard led Iowa to the 2023 NCAA final. Now she's reshaping tickets, ads and events in women's basketball.

Caitlin Clark, over \$3 million — changing the financial playbook

BY DYLAN SLOAN

April 2, 2024 at 3:56 PM EDT



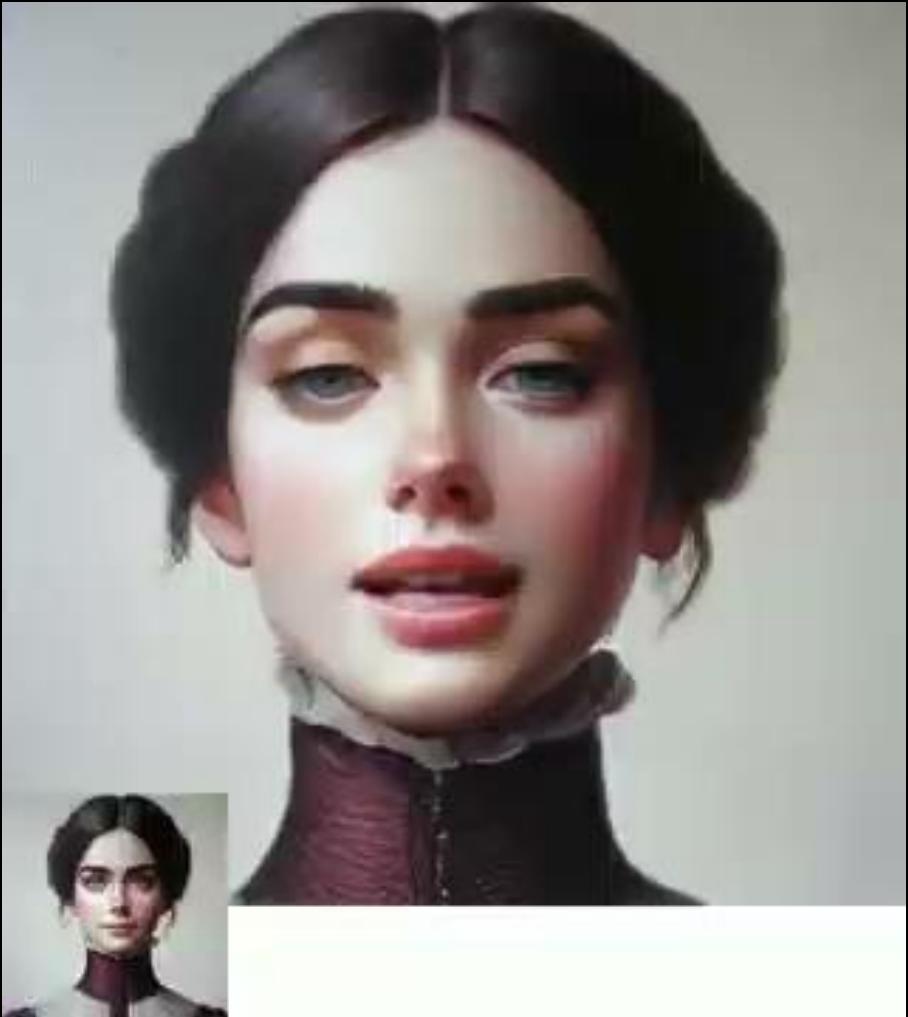
How easily can we create a digital human?

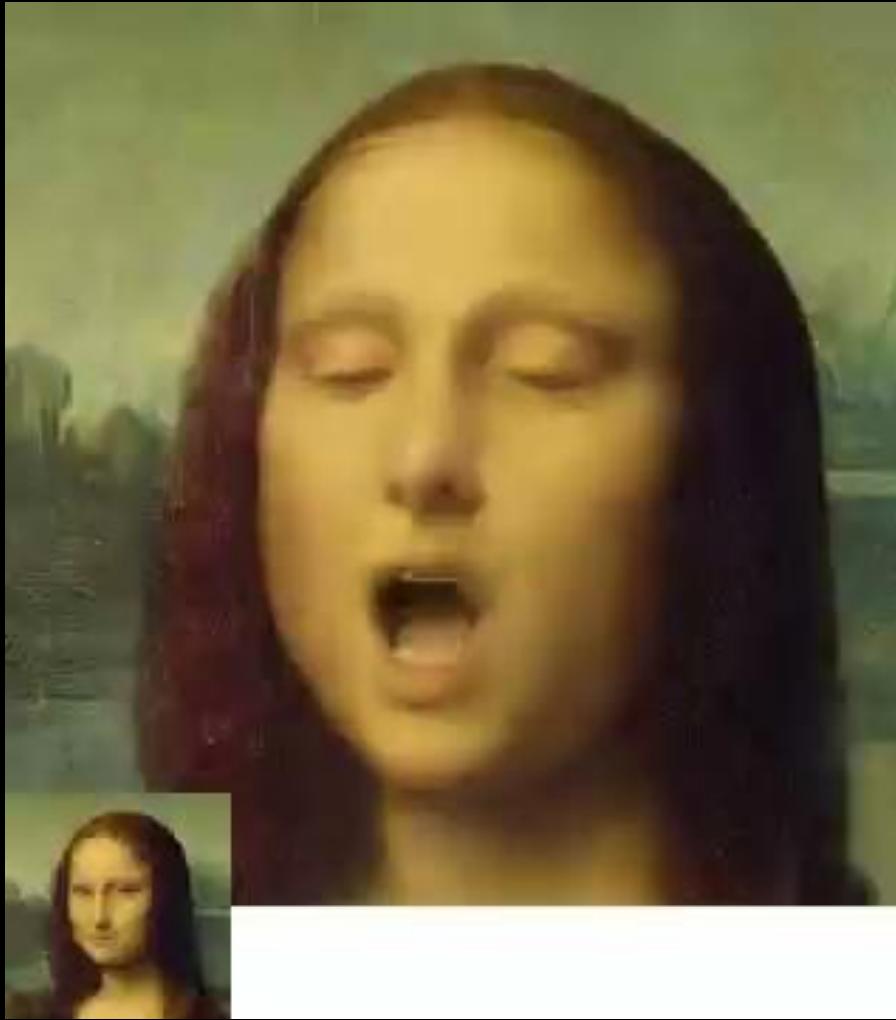
- Cost
- Access
- Time











All it takes is a single image + an audio file



Real-time translation:

The speaker only speaks English!



I'm about to speak six languages in this video



A digital Abel?

Can a digital Abel ...

- Teach a class
- Answer questions
- Record lessons
- Personalize content
- Deliver in multi-language



N

I

L

*Twitter
viral video*





Ariel M.

★ 5.0 (288)

Level 1 ♦♦♦

📍 United States 💬 English

fiverr.

About me

Hi I'm Ariel. I'm a US based content creator with over 7 years of experience. I specialize in Social Media Management, Content Creation, and video production. I am also well trained in... [Read more](#)

Skills

Adobe Photoshop expert

Promotional video creator

Product promoter

Customer service representative

Microsoft Excel expert

+4

Can you simulate a
human? The behavior?



Fashion industry

- Anne Klein is testing technology from AI Fashion that generates fashion shoots based on photos of real-life models.
- AI tool won't necessarily completely replace photoshoots, but it enables companies to build assets at scale
- Startups offer services that use AI to generate images based on a brand's clothing line. In some cases models are completely AI-generated.



Models likeness

Platforms like AI Fashion give models an opportunity to appear in paid campaigns just by providing old images of themselves and letting AI do the rest. The platform creates AI-generated images of human models for fashion campaigns and e-commerce sites.

CIO JOURNAL

These Models Gave Up Photoshoots to Sell Their AI Likenesses

Platforms like AI Fashion give models an opportunity to appear in paid campaigns just by providing old images of themselves and letting AI do the rest



Focus groups

- Focus group Q&A without the expensive monthslong process of setting up a real focus group
- Digital twins of 27,000 real individuals
- AI algorithm can generate answers to questions like whether a female in her 30s would pay a 10% increase on a streaming service subscription



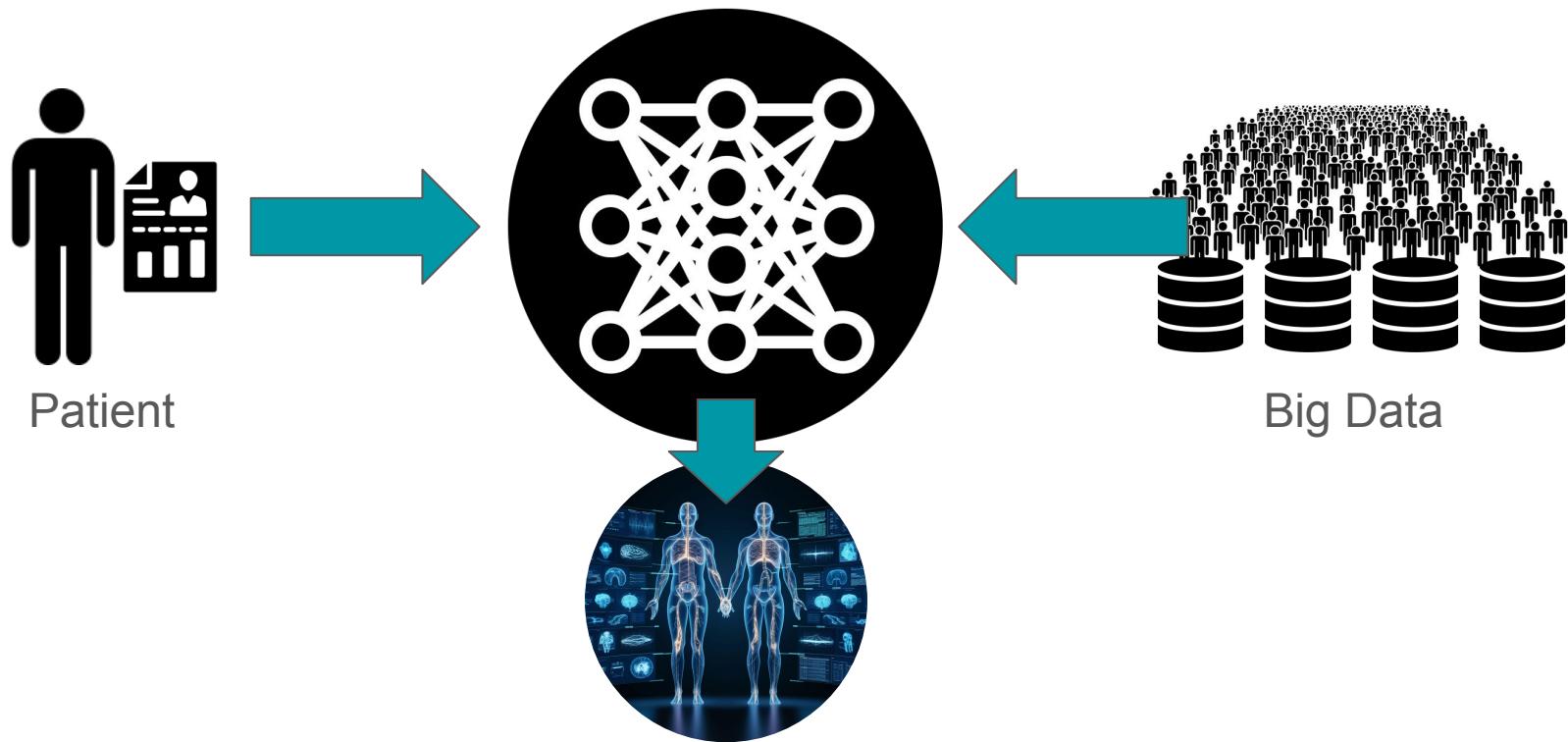
Clinical trials

- Digital twins AI models of people that predict how a particular disease might progress over time.
- Using baseline data points of a participant's health, the model runs a model on vast amounts of clinical data, and generates a digital twin for that individual that forecasts how their disease would progress if they were in the placebo group.



Time (months)	Baseline	3	6	9	12
ADAS Cancellation	0	1.3 ± 0.9	1.4 ± 1.0	1.3 ± 1.1	1.4 ± 1.2
ADAS Commands	0	0.2 ± 0.4	0.2 ± 0.5	0.4 ± 0.6	0.5 ± 0.7
ADAS Comprehension	0	0.1 ± 0.2	0.1 ± 0.3	0.1 ± 0.3	0.2 ± 0.4
ADAS Concentration	5	0.0 ± 0.0	0.0 ± 0.1	0.0 ± 0.2	0.0 ± 0.1
ADAS Construction	1	1.1 ± 0.8	1.1 ± 0.8	1.0 ± 0.8	1.0 ± 0.8
ADAS Delayed Word Recall	-	7.4 ± 1.5	7.6 ± 1.8	7.9 ± 1.6	8.2 ± 1.5
ADAS Executive	5	5.0 ± 0.0	4.8 ± 0.6	4.4 ± 0.9	3.8 ± 1.3
ADAS Ideational	0	0.1 ± 0.4	0.2 ± 0.5	0.4 ± 0.6	0.6 ± 0.8
ADAS Naming	1	1.0 ± 0.6	0.9 ± 0.7	0.9 ± 0.8	0.9 ± 0.8
ADAS Orientation	3	3.2 ± 1.5	3.3 ± 1.6	3.4 ± 1.5	3.4 ± 1.7
ADAS Remember Instructions	1	1.0 ± 0.7	0.9 ± 0.8	0.9 ± 0.9	0.8 ± 0.9
ADAS Spoken Language	0	0.1 ± 0.2	0.1 ± 0.2	0.1 ± 0.3	0.2 ± 0.4
ADAS Word Finding	0	0.1 ± 0.4	0.3 ± 0.5	0.4 ± 0.6	0.5 ± 0.7
ADAS Word Recall	3	3.5 ± 1.3	3.9 ± 1.3	4.3 ± 1.5	4.7 ± 1.7
ADAS Word Recognition	4	4.1 ± 2.6	5.1 ± 3.0	5.6 ± 3.2	6.5 ± 3.1
CDR Community	1	1.1 ± 0.3	1.1 ± 0.4	1.1 ± 0.5	1.2 ± 0.5
CDR Home Hobbies	2	1.8 ± 0.4	1.6 ± 0.5	1.4 ± 0.6	1.3 ± 0.6
CDR Judgement	1	1.1 ± 0.5	1.2 ± 0.5	1.3 ± 0.5	1.3 ± 0.6
CDR Memory	1	1.1 ± 0.4	1.1 ± 0.5	1.2 ± 0.5	1.3 ± 0.5
CDR Orientation	1	1.1 ± 0.4	1.2 ± 0.5	1.2 ± 0.5	1.3 ± 0.5
CDR Personal Care	1	1.0 ± 0.3	1.1 ± 0.5	1.1 ± 0.5	1.2 ± 0.7

Digital twins in clinical trials



A black and white photograph showing a large stack of papers or documents. The stack is composed of numerous pages, some of which are slightly visible at the top. The overall texture appears slightly grainy.

WE NEED IT

lit

The background of the slide is a photograph of the interior of a grand library, likely the Trinity College Library in Dublin. The room is vast with high ceilings and arched wooden beams. Numerous tall, dark wood bookshelves are filled with books, reaching almost to the top of the frame. A balcony level is visible on the left, featuring a railing with a blue plaque that reads "PALLISERIANA". The overall atmosphere is one of historical knowledge and academic tradition.

Average Person



750 books/life



130 million exist



0.0000058

Digital Data Universe

175 zettabytes

1.4 billion x bigger
(than number of books)

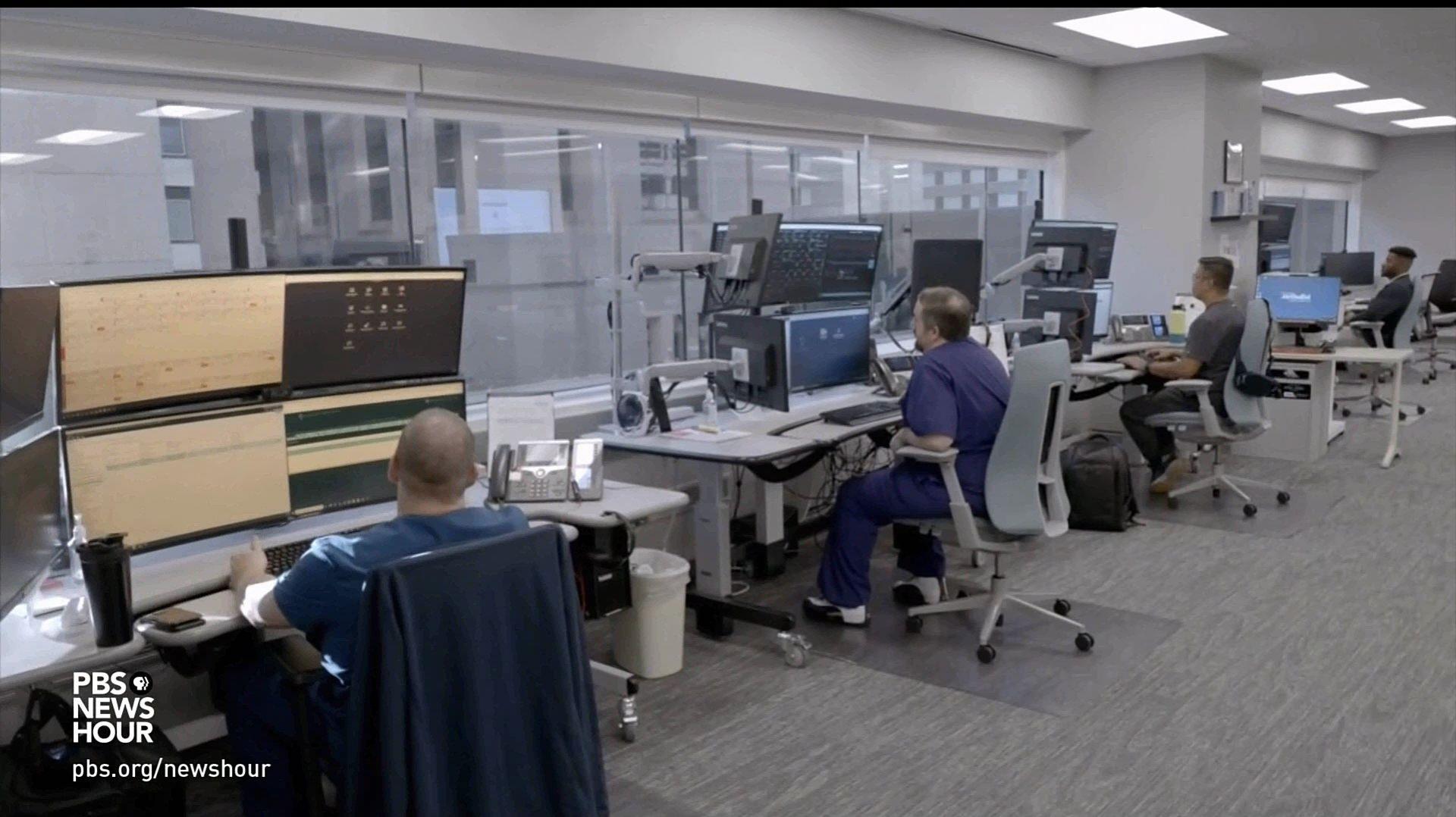
32,000 deaths/year, leading cause of death, ages 4-to-35



Potential for
80% reduction

Healthcare crisis in many nations





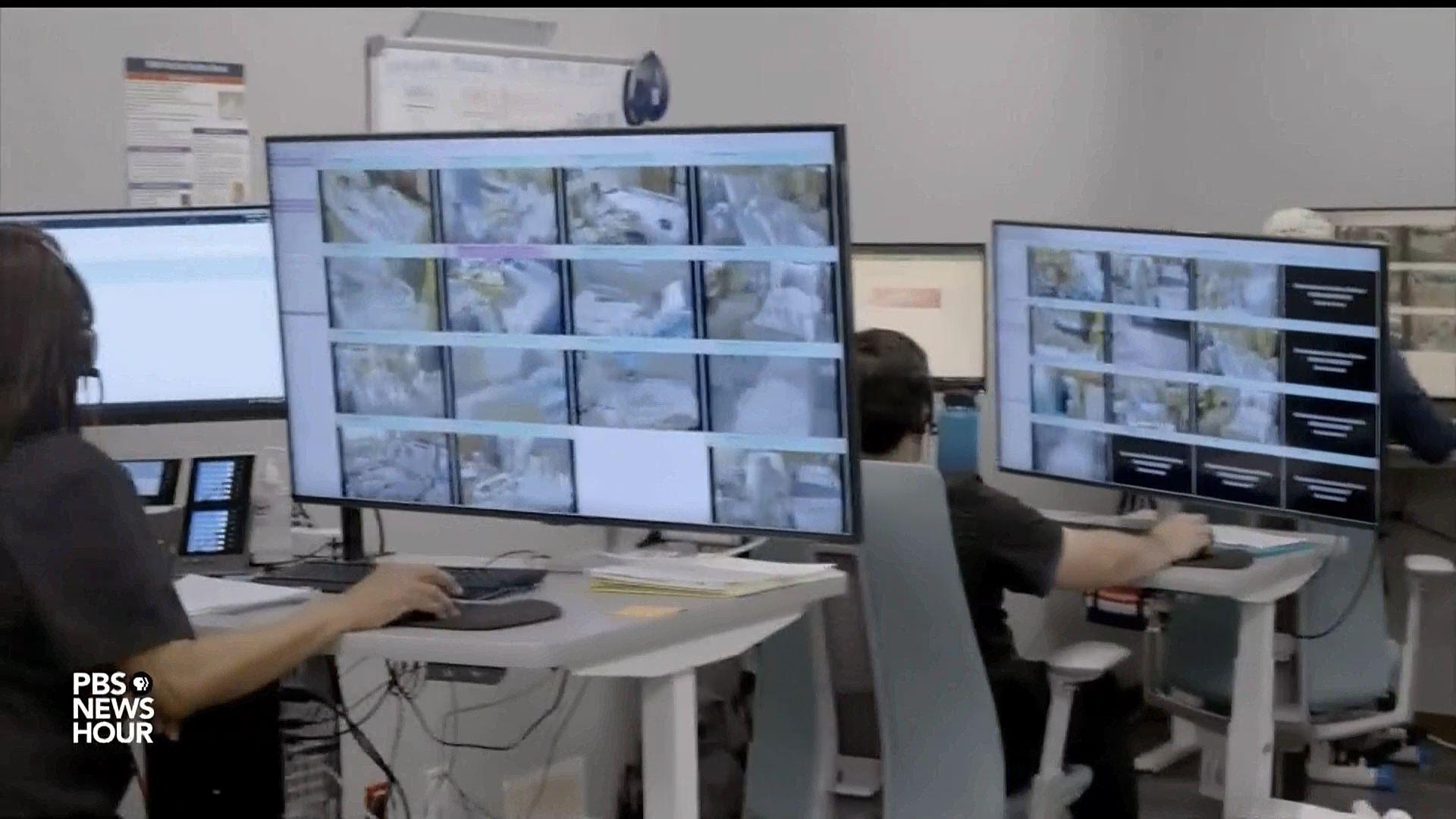
PBS
NEWS
HOUR

pbs.org/newshour



PBS
NEWS
HOUR

pbs.org/newshour



NURSE WORKFORCE GAP



200,000 - 500,000

bedside nurses

Cybersecurity

You cannot read 1M lines

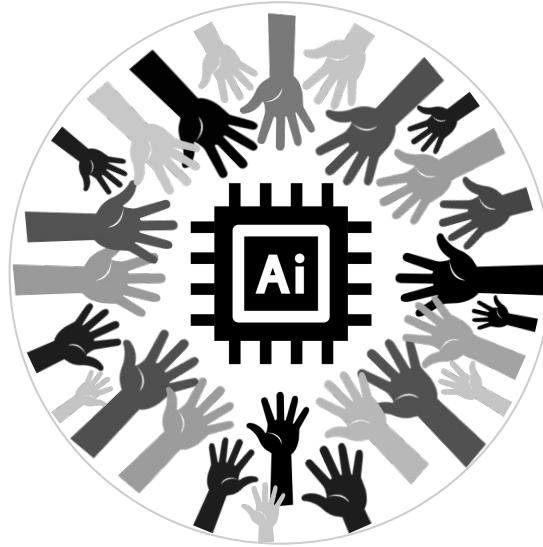
- An LLM can read it
- It will make mistakes
- It reduces the problem
- Makes it manageable

A large wall of code, likely a screenshot of a terminal or code editor showing a massive amount of text. The code is a mix of various programming languages and syntax, appearing as a dense grid of colored text (blue, green, yellow, red, etc.). It includes comments, variable declarations, loops, conditionals, and function definitions, though no specific logic can be discerned due to the sheer volume.

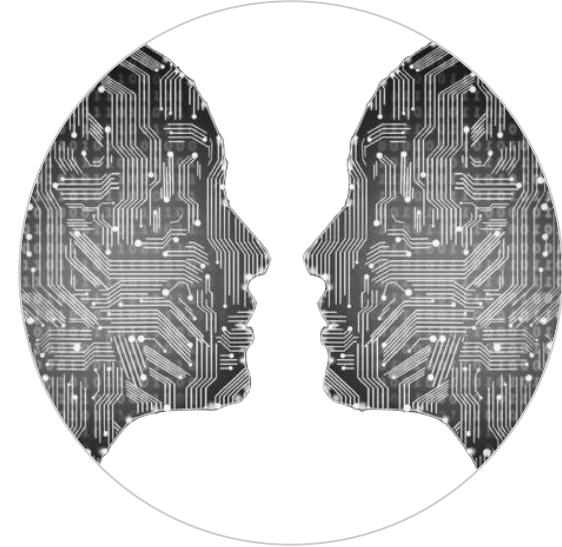
Democratized access - A new opportunity



Data



AI



Human
Simulation

Q&A

You too can be an author

Example: You too can be an author.

LLM prompt: Create a story based on the following ...

A young girl, 11 years old, working Arizona-Sonoran desert. Location is on the coast of the Sea of Cortez, in the sonoran desert. Agriculture themes, vineyards, citrus trees, the aroma smell of oranges, grapefruits, and lemons



First page ...

Isabel's laughter danced on the desert wind, swirling between emerald grape leaves like secrets whispered across generations. Her sun-kissed skin gleamed against the fiery Baja California sky, a vibrant contrast to the dusty earth clinging to her bare feet. At ten, Isabel was already a seasoned farmhand, her life woven into the tapestry of their vineyard nestled near the Sea of Cortez.

Every morning began with the rooster's crow and the vibrant fragrance of citrus groves. Isabel would race through the orchard, inhaling the sweet symphony of oranges, grapefruits, and lemons, each scent holding a memory: the tang of lemonade on a scorching summer day, the sticky sweetness of fresh marmalade, the warmth of her Mom's famous citrus cake.

Days were symphonies of labor and laughter. Isabel, alongside her siblings and weathered vaqueros, weaved through the endless rows of grapevines. Her nimble fingers plucked sun-ripened clusters, the juicy weight a familiar comfort. The air buzzed with chatter and the rhythmic snap of scissors, punctuated by bursts of song as they filled baskets to overflowing.

Create image to
fit the story?



Create image to
fit the story?

