

Ethical & Responsible AI

Sanjaya Edirisinghe
Sanjaya@bu.edu



DEPARTMENT OF ELECTRICAL AND INFORMATION ENGINEERING
UNIVERSITY OF RUHUNA

Bio

- Director of IT
- University of Peradeniya, BSc Eng
- Boston University – MSc (AI & ML)
- Massachusetts Institute Technology (MIT) – PgDip
- SanjayaE@bu.edu

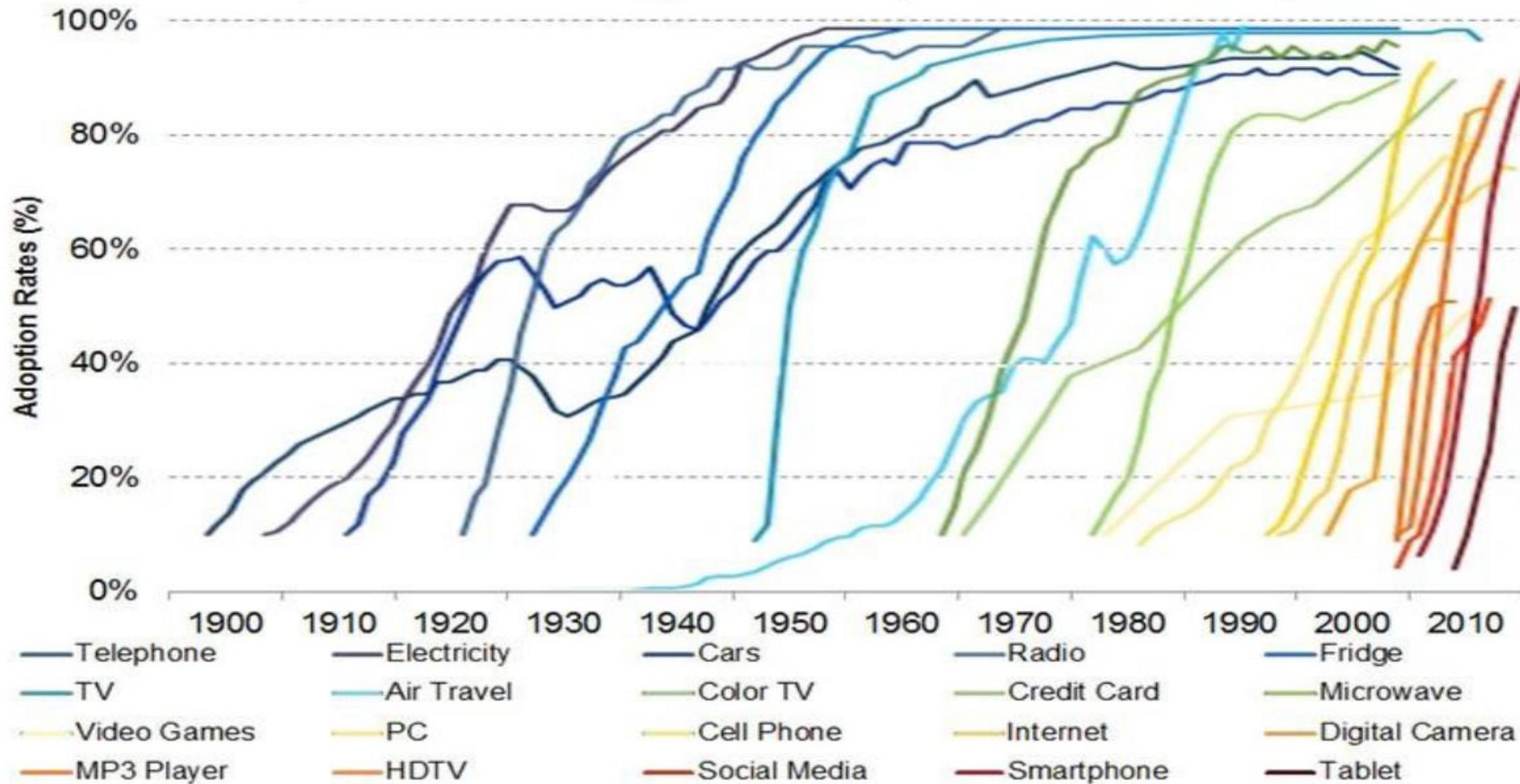


Agenda:

- Why AI
- Impact of AI
- AI Ethics
- AI Ethics framework
- Long term Social Impact of AI

Why AI ?

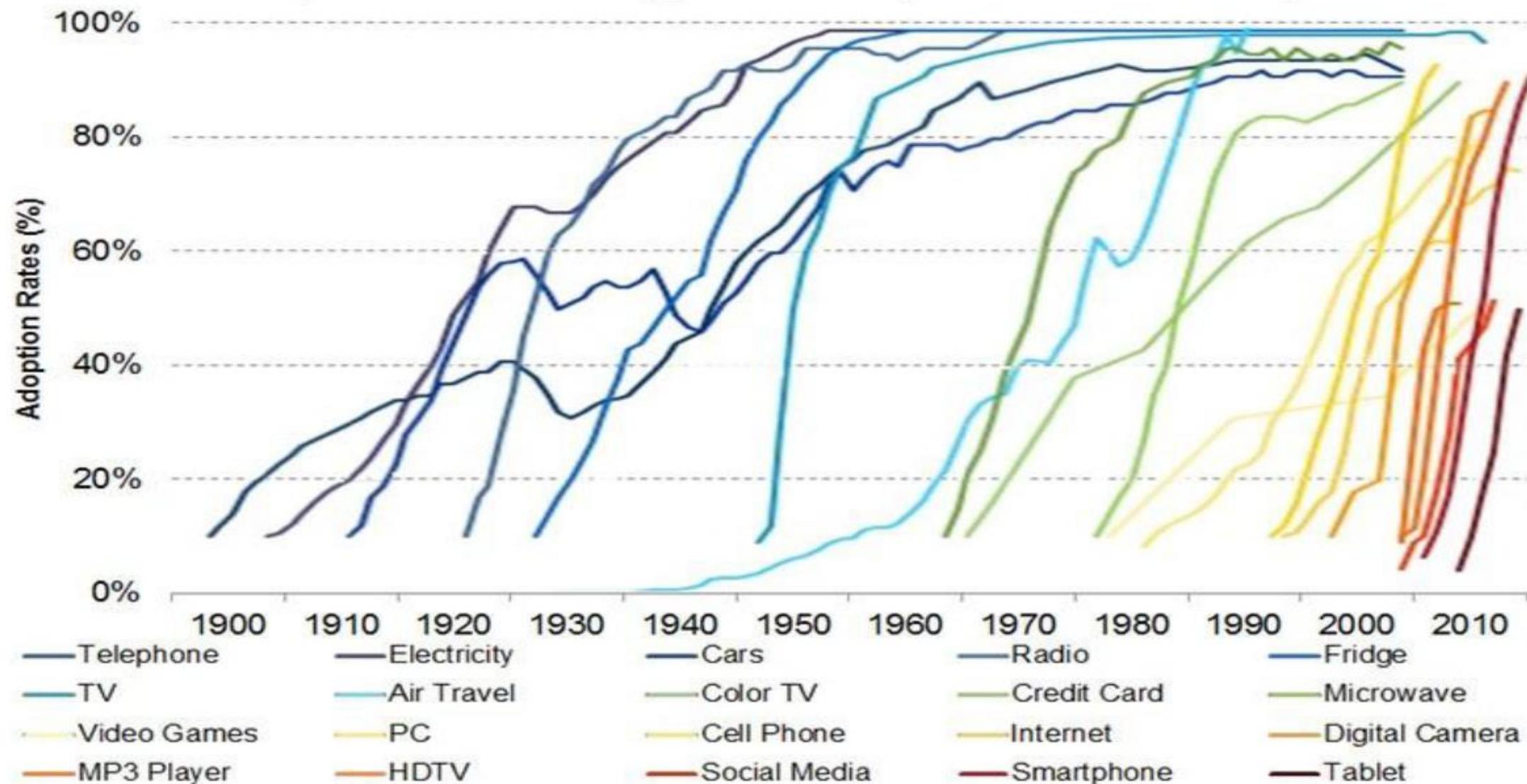
Industrial & consumer trends for last ~125 years



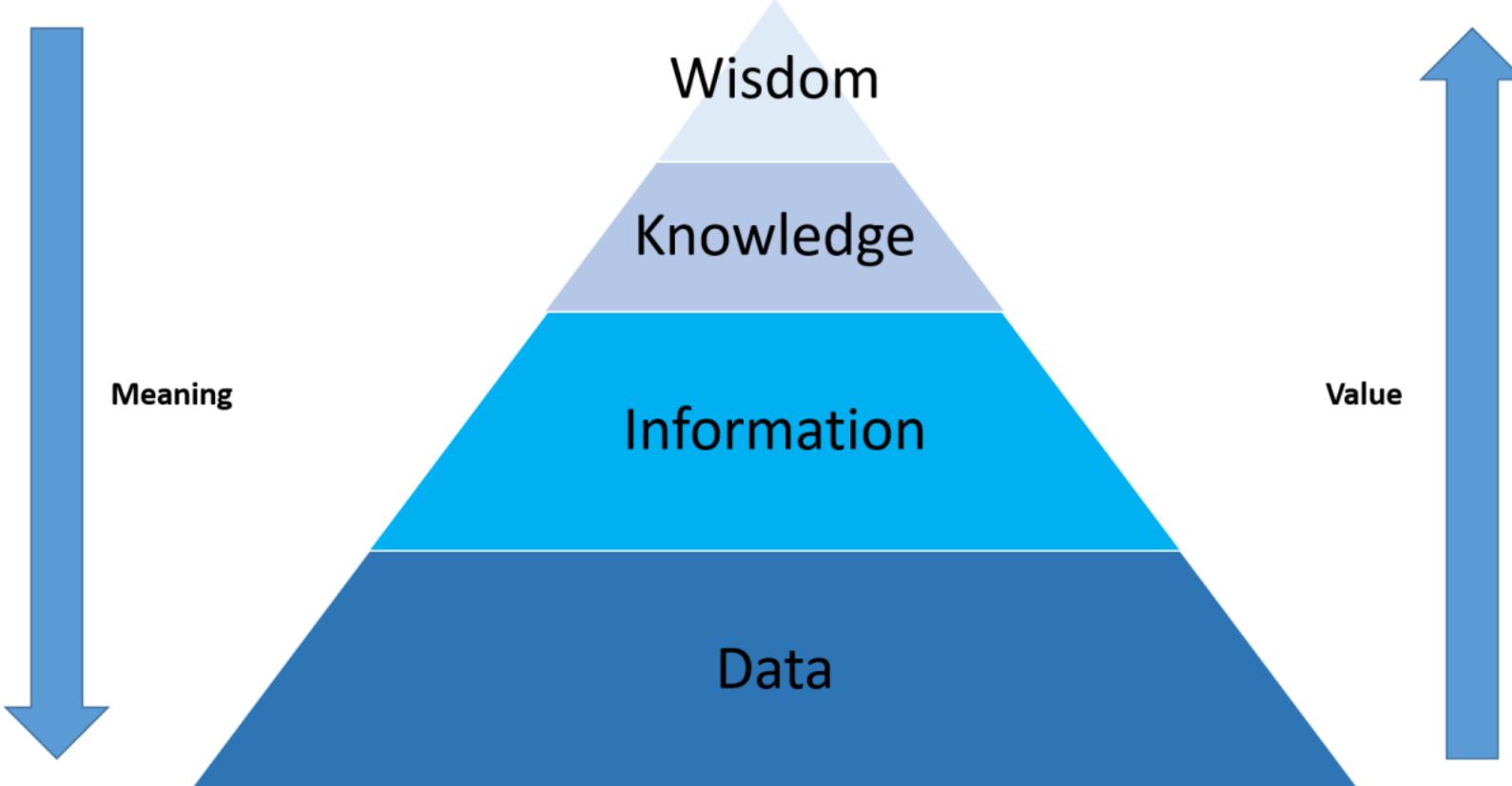
Mechanical + Electrical

Electronic

Data



DIKW Pyramid of Data (Science)

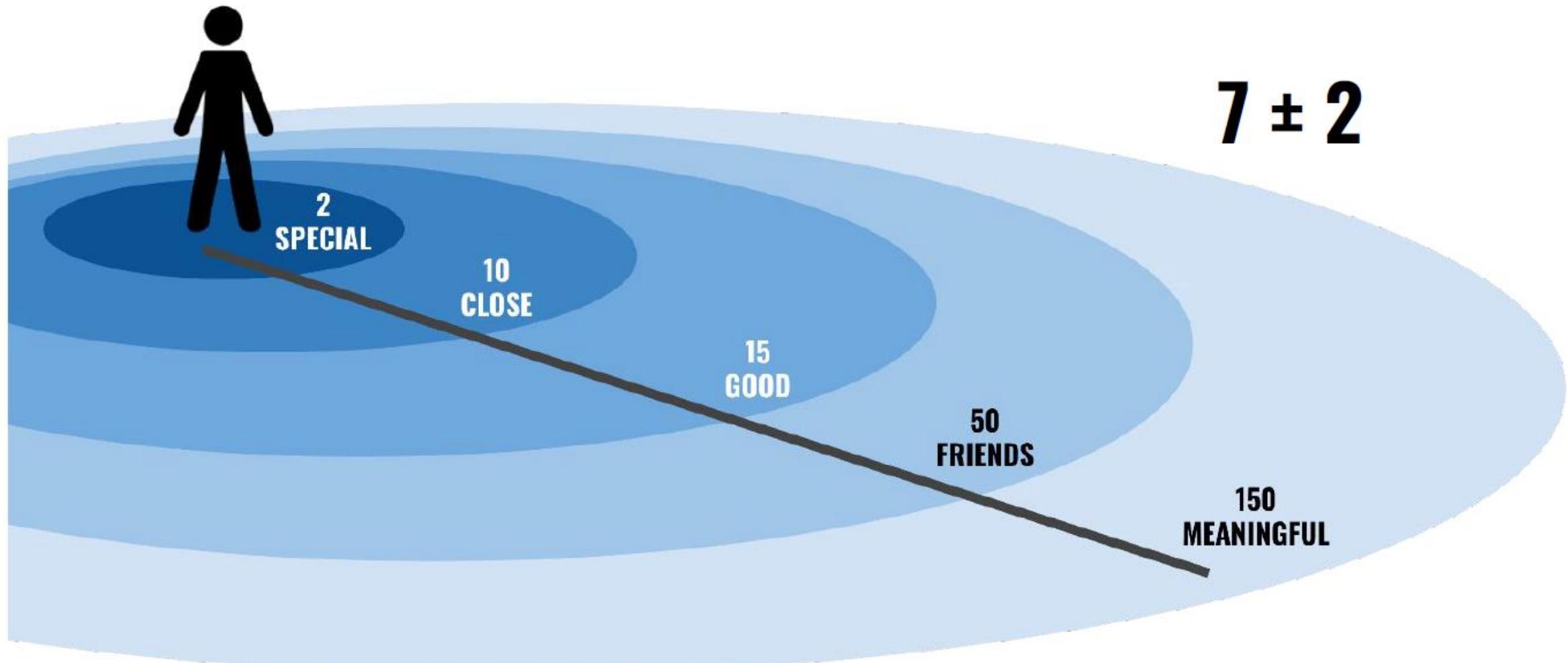


Data >> Value

Time (Speed)

AI (Accelerator)

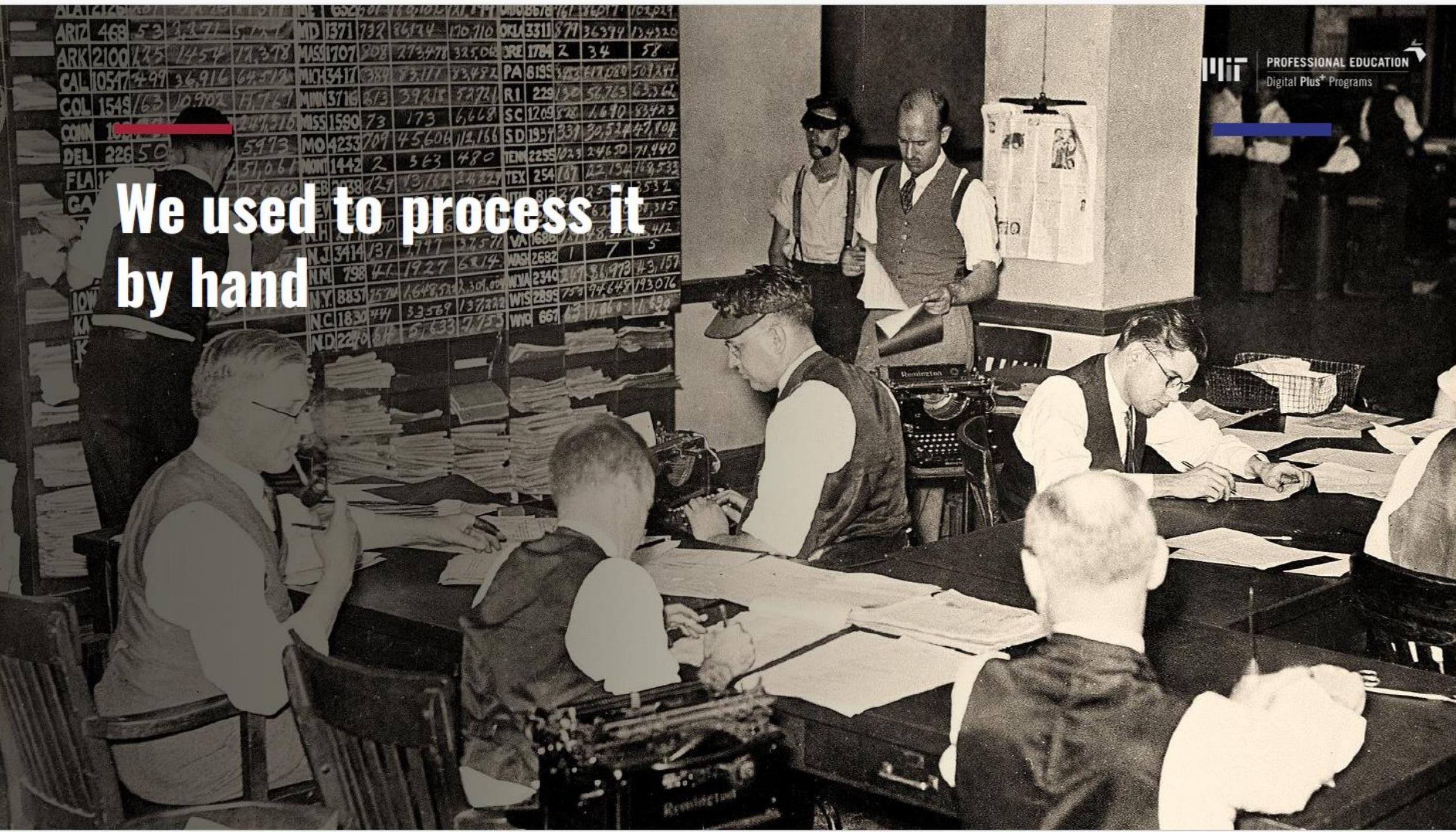
Human Limitations

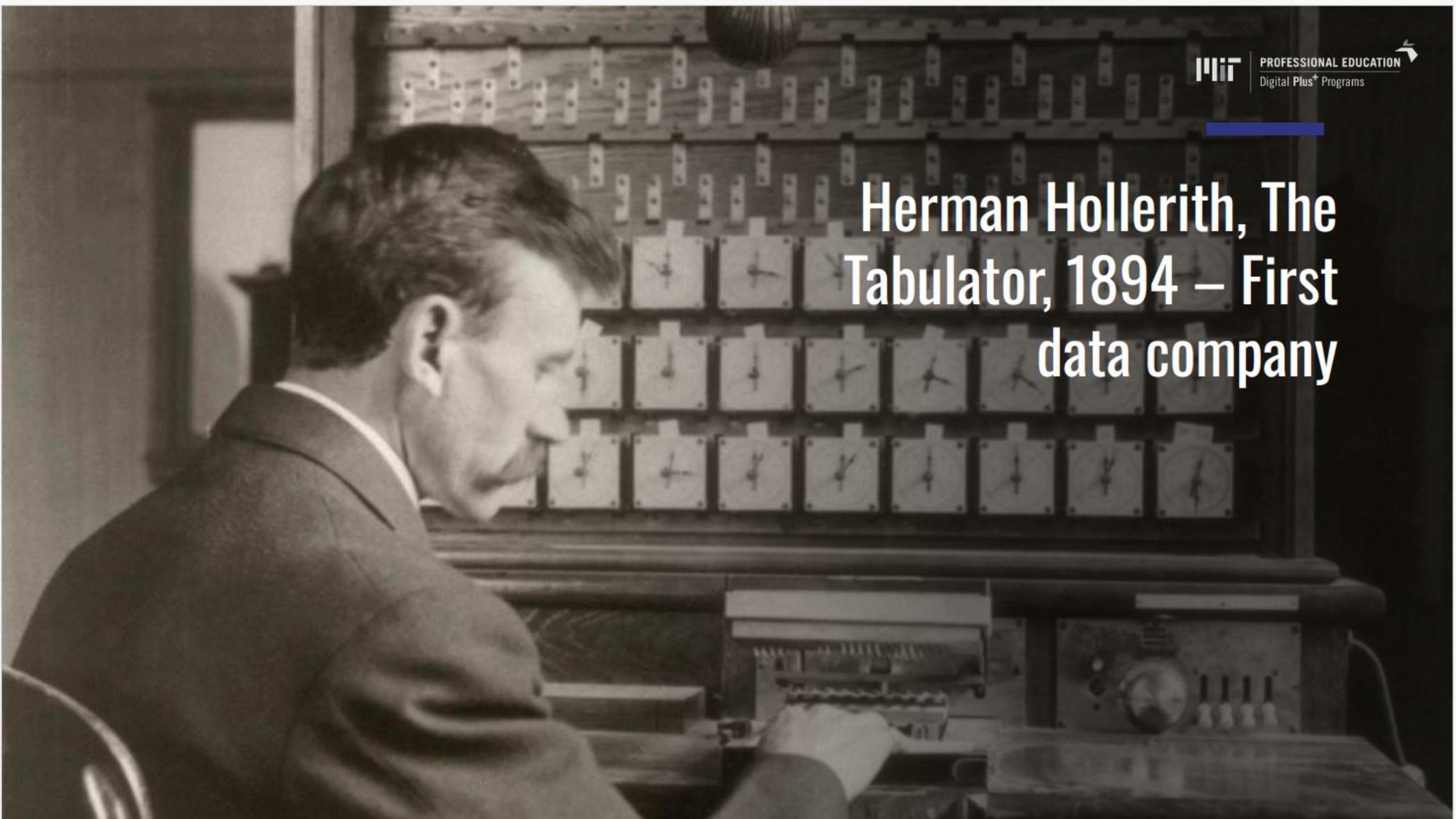


We used to collect
data by hand

The image shows a vintage computer system from the mid-20th century. The main component is a large, light-colored console with a prominent vacuum-tube indicator panel at the top. Below the panel is a large, rectangular CRT monitor displaying a grid of data. The data consists of two columns of names and numbers, likely representing state abbreviations and their corresponding values or codes. The entire scene is set against a dark, textured background that appears to be a wall or a piece of furniture. Superimposed over the bottom left portion of the image is a large, bold, white text message that reads "We used to process it by hand".

We used to process it by hand





Herman Hollerith, The Tabulator, 1894 – First data company

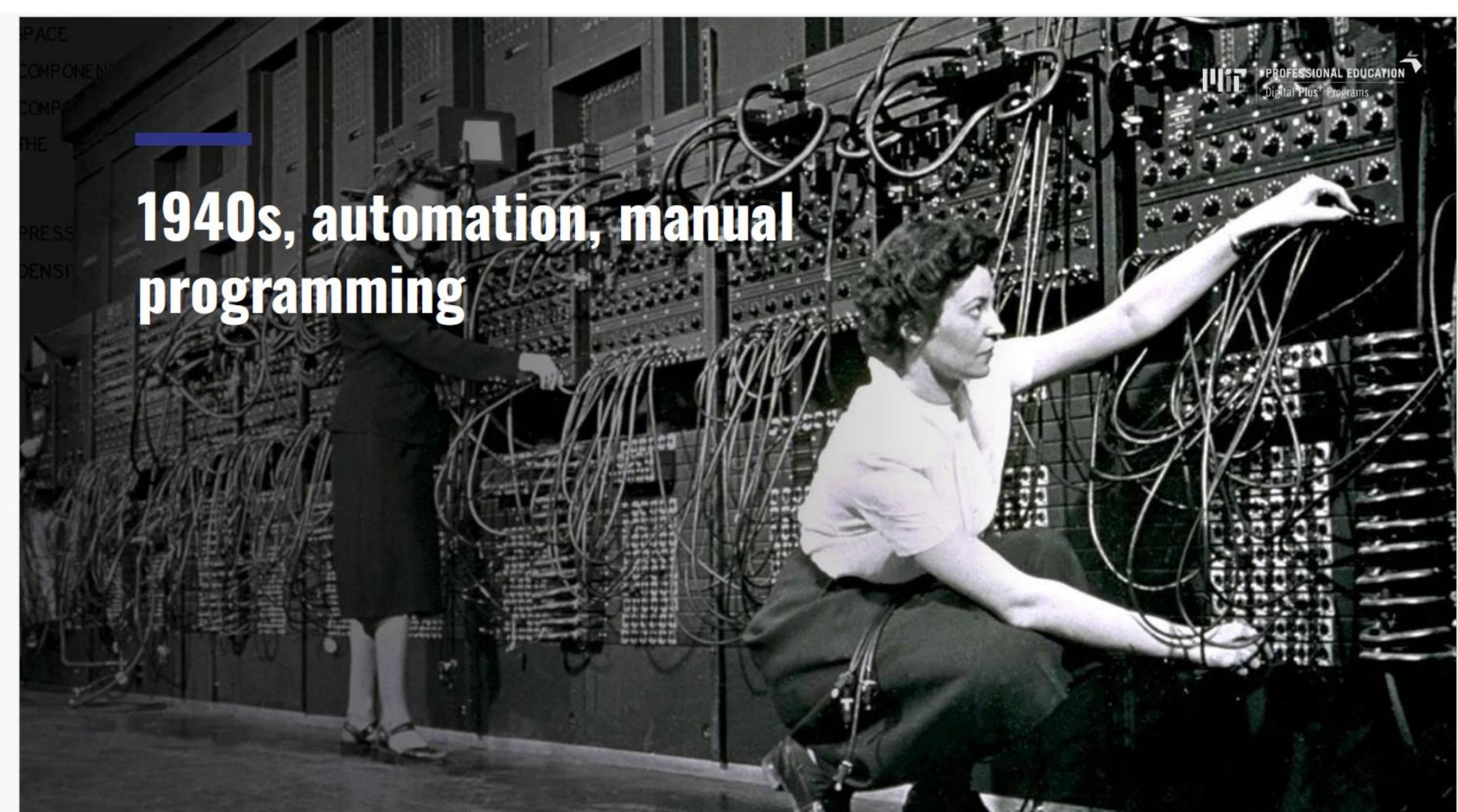


1920s, human layers
of integration

PACE
COMPONENT
COMP
THE

PRESS
DENSIT

1940s, automation, manual programming



1960s, capturing
instructions

1982



Starting MS-DOS...

Really easy,
right?

HIMEM is testing extended memory...done.

C:\>C:\DOS\SMARTDRV.EXE /X

MODE prepare code page function completed

MODE select code page function completed
C:\>dir

Volume in drive C is MS-DOS_6

Volume Serial Number is 40B4-7F23

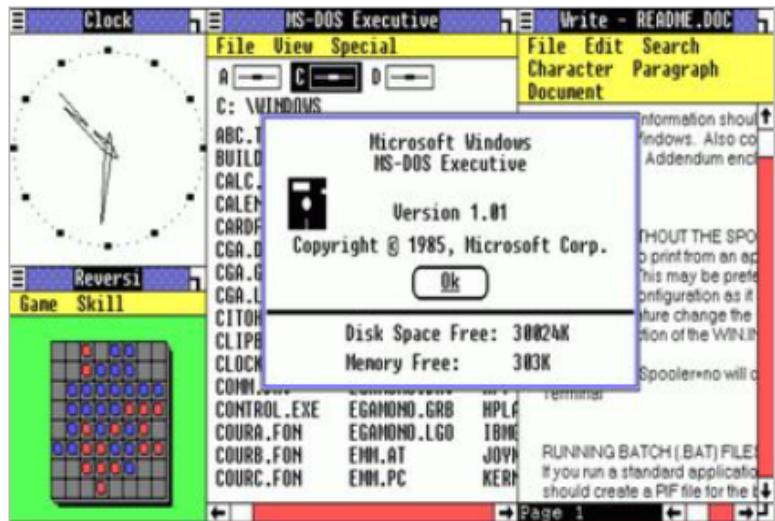
Directory of C:\

DOS	<DIR>	12.05.20	15:57
COMMAND	COM	54 645	94.05.31
WINA20	386	9 349	94.05.31
CONFIG	SYS	144	12.05.20
AUTOEXEC	BAT	188	12.05.20
		64 326 bytes	15:57
		24 760 320 bytes free	

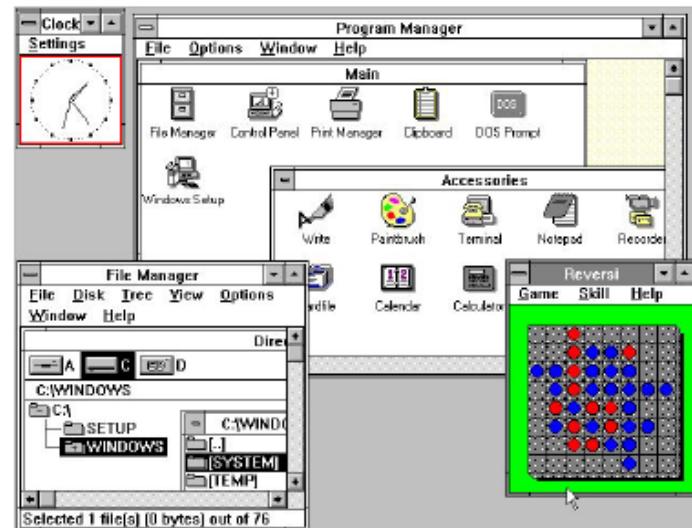
C:\>_

Graphical User Interface (GUI)

Windows 1.0



Windows 3.0



Windows 95



80s/90s, GUIs

We have arrived at making requests in plain language

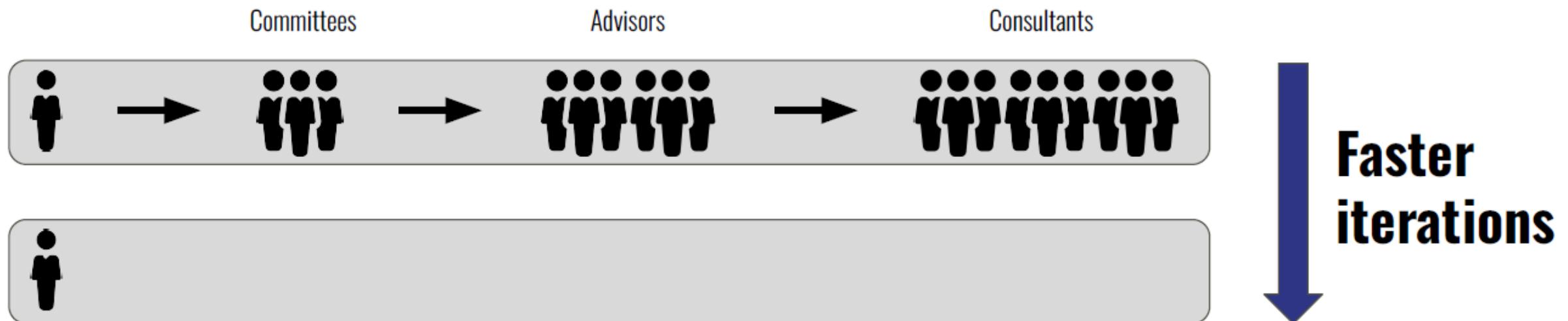
(not like Alexa)



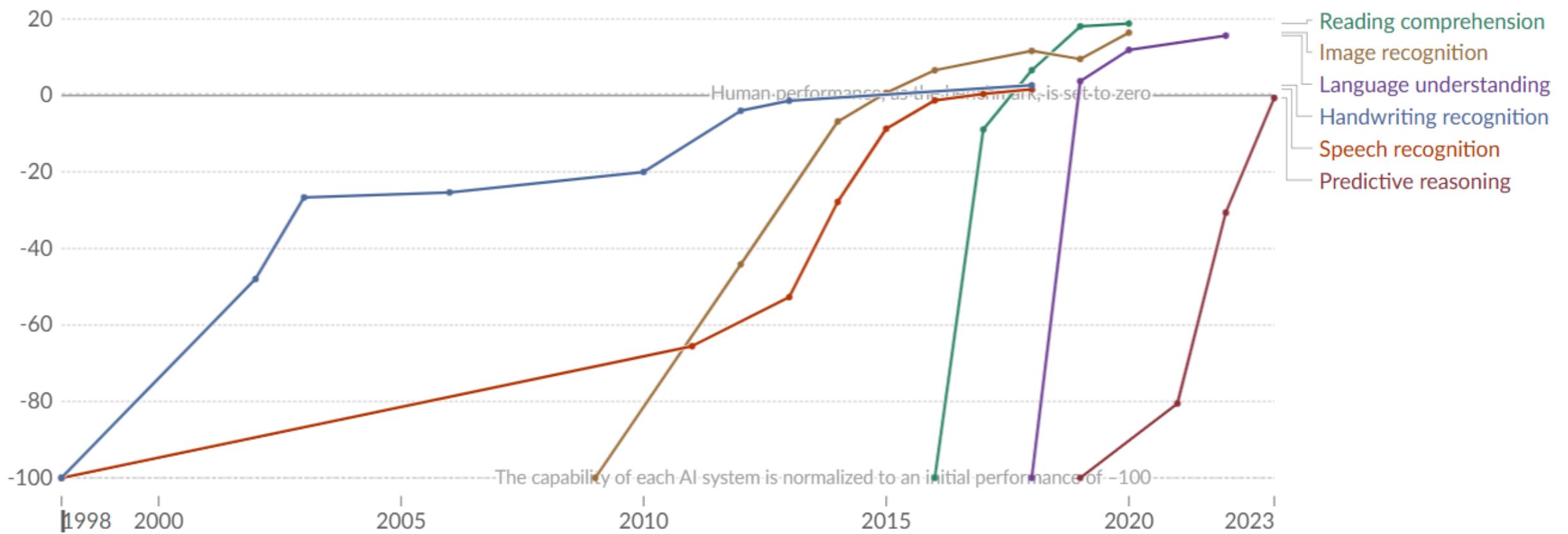
Generative AI



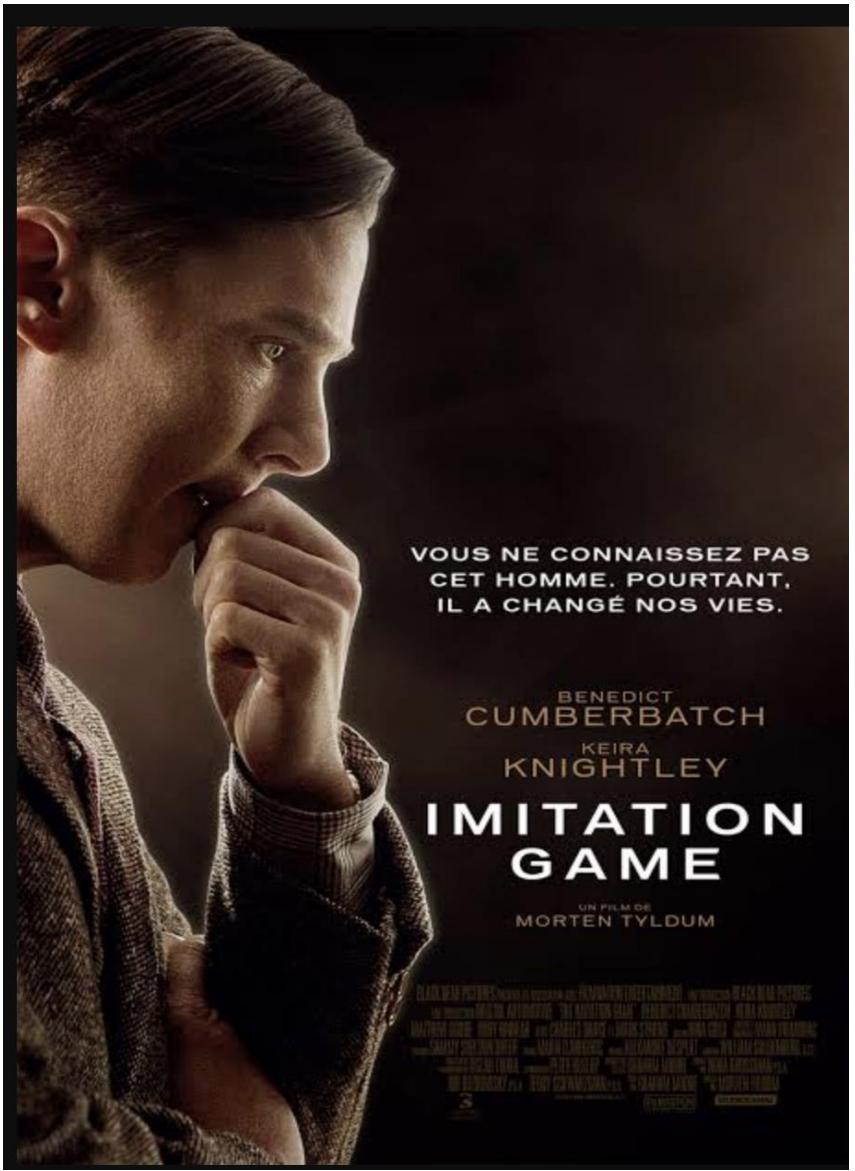
Post-ChatGPT, Moving at the speed of thought



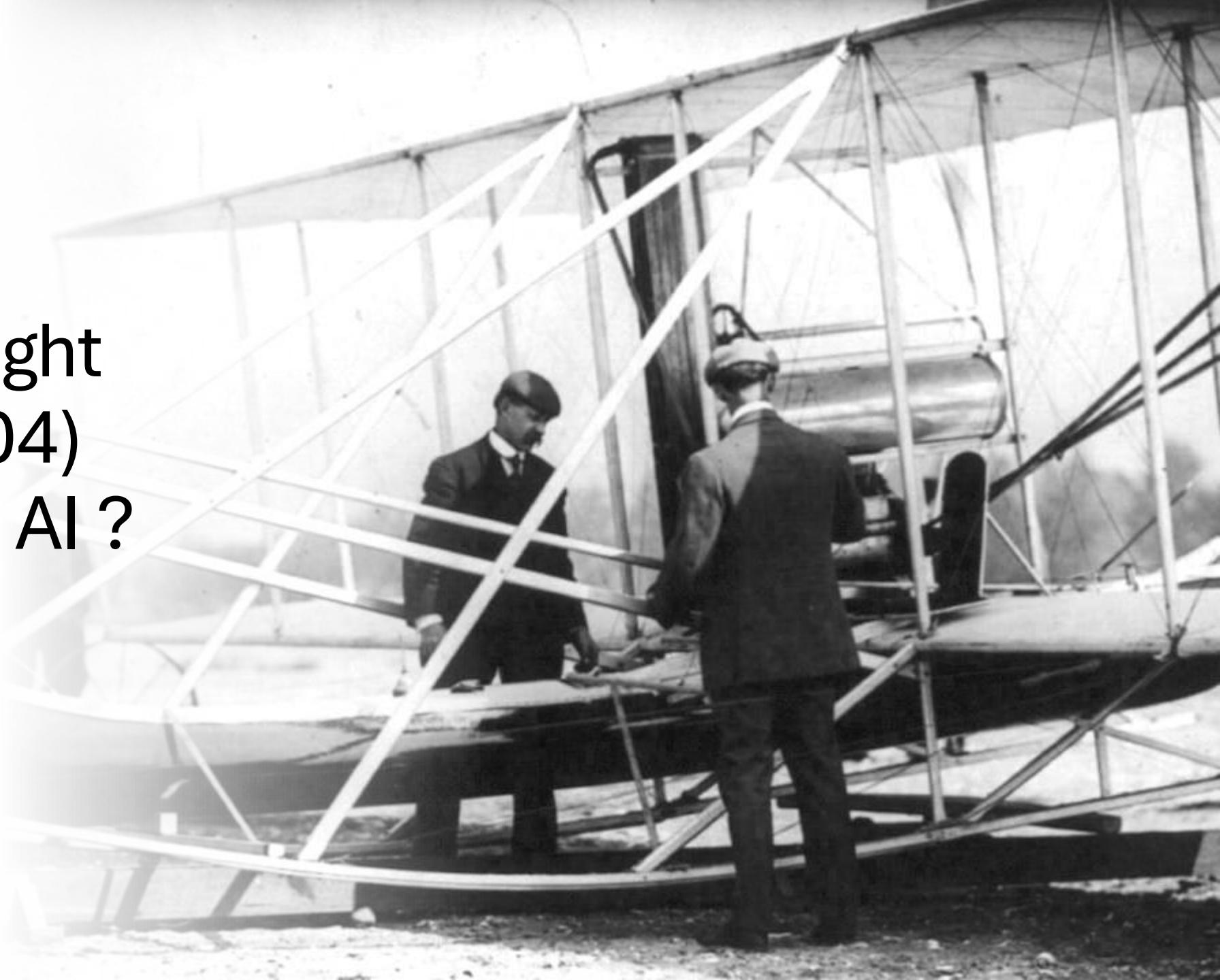
Test scores of AI systems on various capabilities relative to human performance



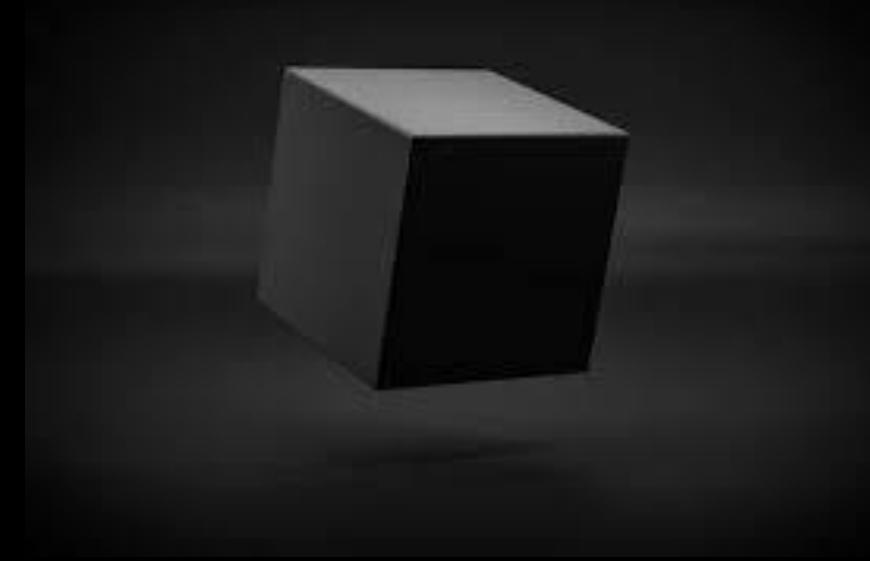
AI Ethics



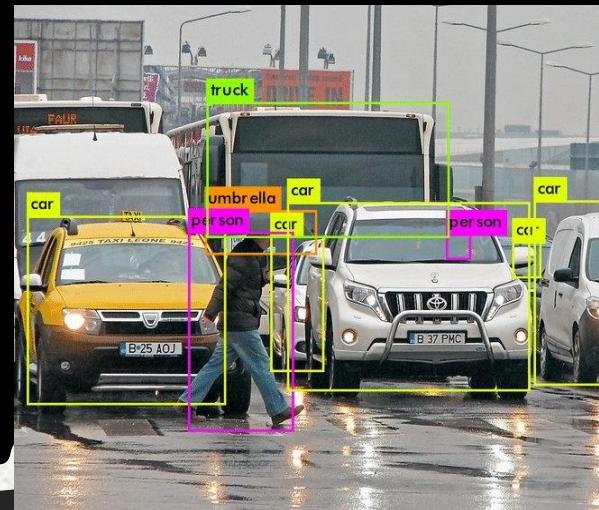
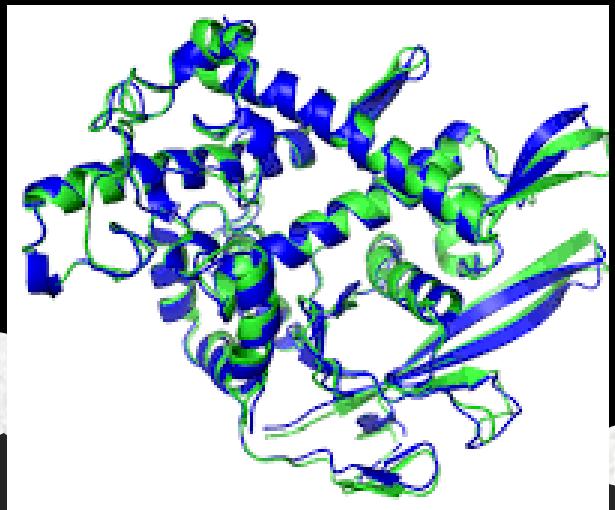
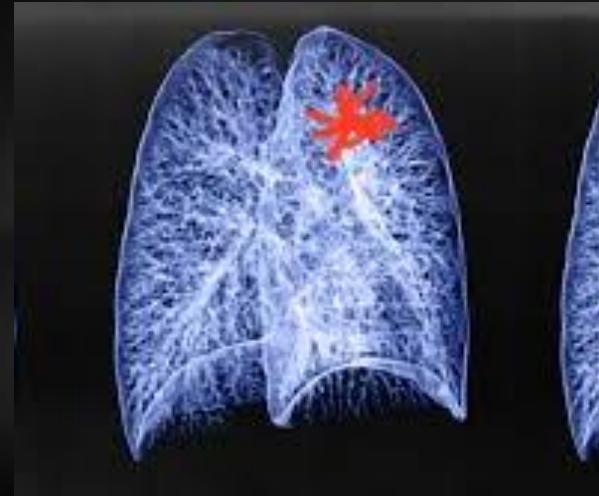
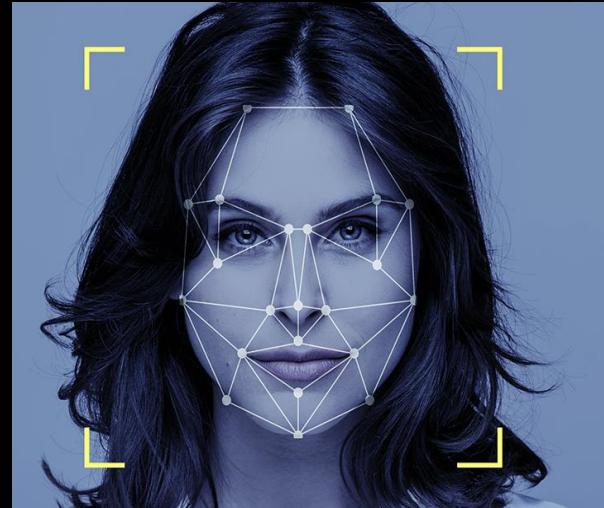
Are we at Wright
Brothers (1904)
Moment with AI ?



Missing the fundamentals
and can't explain current
AI capabilities work

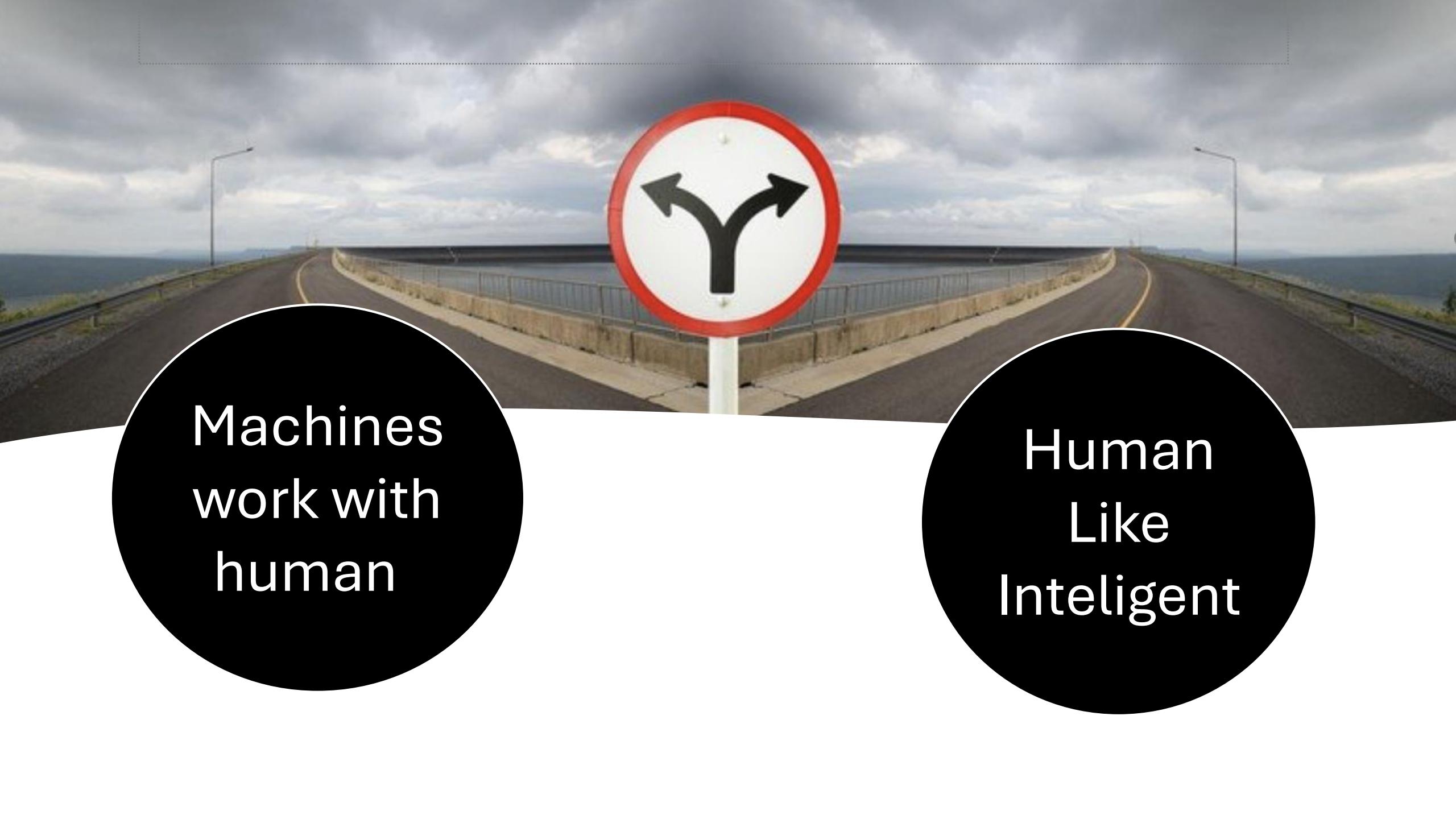


AI Success Stories ..



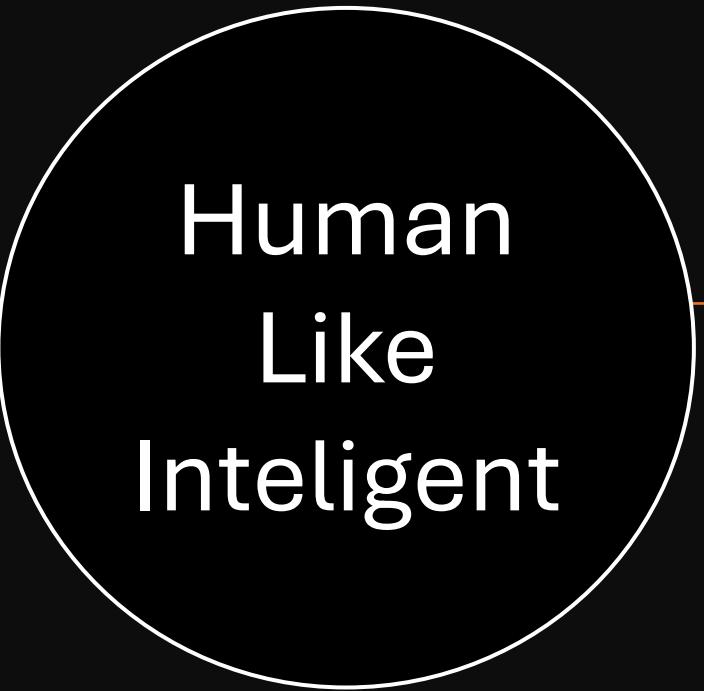
AI failed stories..





Machines
work with
human

Human
Like
Intelligent



Human
Like
Intelligent

**What are AI ethics ?
who sets them?**

AI ethics focus areas



Fairness &
Bias
Mitigation



Transparency &
Explainability



Accountability &
Responsibility



Privacy & Data
Protection



Safety & Security



Sustainability

AI Regularization



In general, all industry and consumer protection & technology regulations are applicable for AI as well.

- EU's General Data Protection Regulation (GDPR) – 2018
- California Consumer Privacy Act (CCPA)
- Sri Lanka - Personal Data Protection Act (PDPA), No. 9 (2022)

AI Specific Regulations:

- European Union's Artificial Intelligence Act (EU AI Act)
- Generative AI Accountability Act - SB 896 (California)

AI Guidelines/Framework (Not regulations)

- NIST AI 600-1, Artificial Intelligence Risk Management Framework



Fairness & Bias Mitigation



Data, Algorithm, Design Deployment ,Maintenance

- E.g. Facial recognition AI trained on Western datasets underperforms for darker skin tones.
- Mitigation strategies: Balanced datasets, fairness constraints in ML models.

Protected Classes

- Race,
- Color,
- Religion,
- Sex (including pregnancy, sexual orientation, or gender identity),
- National Origin
- family medical history

Other biased, but not protected classes.

- Culture
- Social -economic Status
- Income, employment status

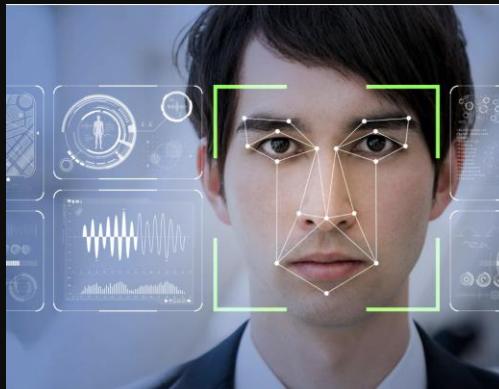


Transparency & Explainability

- Transparent on purpose, design, development, and deployment processes.
- Explain and document everything (if possible)
- Continues user feedback.
- 3rd party audits, work with regulators etc.
- User techniques like SHAP, LIME, interpretable model



Privacy & Data Protection



Issues:

- Deepfake fraud
- Profiling/Impersonation
- Unethical People ranking

Regulations:

- EU's GDPR
- (CCPA)
- Sri Lanka - Personal Data Protection Act (PDPA), No. 9 (2022)

Solutions

- Data Encryptions
- MFAS



Safety & Security

- AI systems controlling critical infrastructure must be fail-safe.
- Risks of AI in cybersecurity: Deepfake fraud, automated phishing attacks.
- Securing AI systems against adversarial attacks.



Accountability & Responsibility



Who is responsible for Boeing 737 MAX crashes and the role of automation failures ?

- Who is responsible for AI-caused harm? (Developers, businesses, governments?)
- Need for AI governance frameworks in organizations.





Sustainability



- high Energy Consumption
- Carbon footprint/impact on global warming
- High Resource Utilization

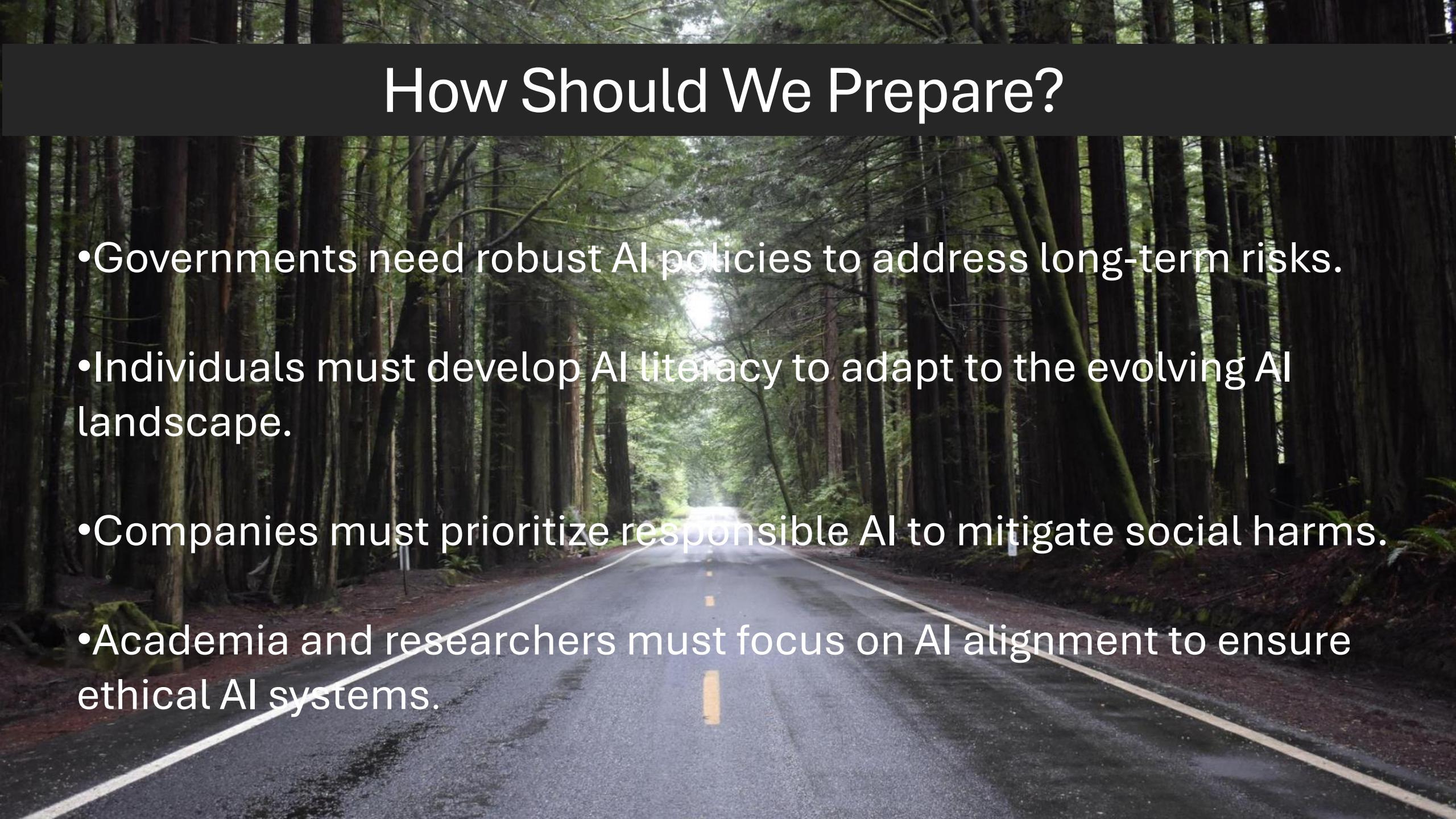


Social Impact of AI

- Human Longevity
- Personalized medication & medication
- Workforce Transformation & Job Displacement
 - Data is the new asset
- **AI taxation/Universal Basic Income (UBI)**
- **AI-augmented human**



How Should We Prepare?

A photograph of a two-lane asphalt road curving through a dense forest. The road is flanked by tall, dark evergreen trees with mossy branches hanging over the sides. Sunlight filters through the canopy, creating bright patches on the road surface. The perspective leads the eye down the center of the road towards a bright opening in the distance.

- Governments need robust AI policies to address long-term risks.
- Individuals must develop AI literacy to adapt to the evolving AI landscape.
- Companies must prioritize responsible AI to mitigate social harms.
- Academia and researchers must focus on AI alignment to ensure ethical AI systems.