Ethical & Responsible AI

Sanjaya Edirisinghe
Sanjaya@bu.edu

DEPARTMENT OF ELECTRICAL AND INFORMATION ENGINEERING
UNIVERSITY OF RUHUNA

Bio

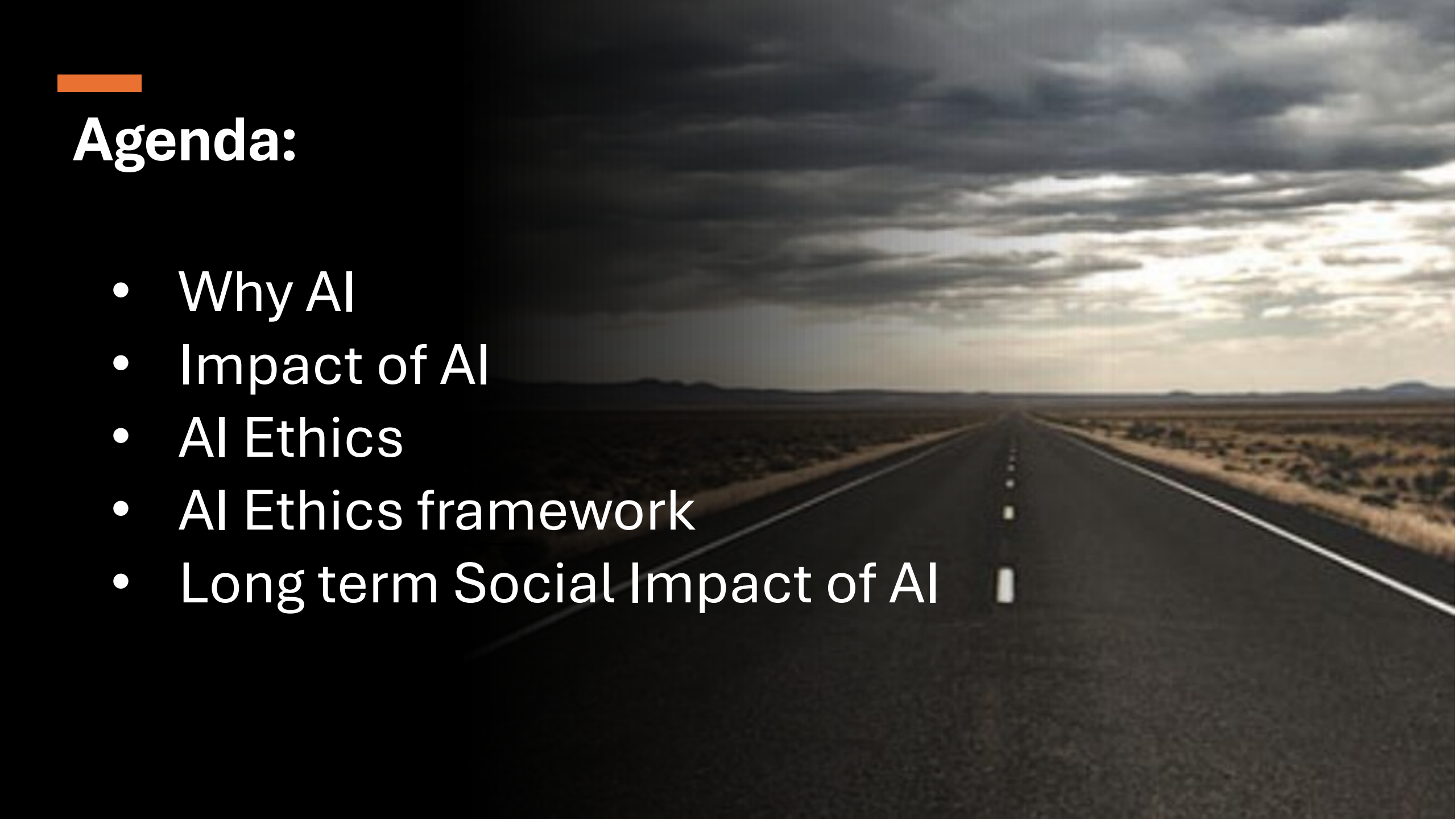
- Director of IT

  - University of Peradeniya, BSc Eng
  - Boston University – MSc (AI & ML)
  - Massachusetts Institute Technology (MIT) – PgDip

  - SanjayaE@bu.edu

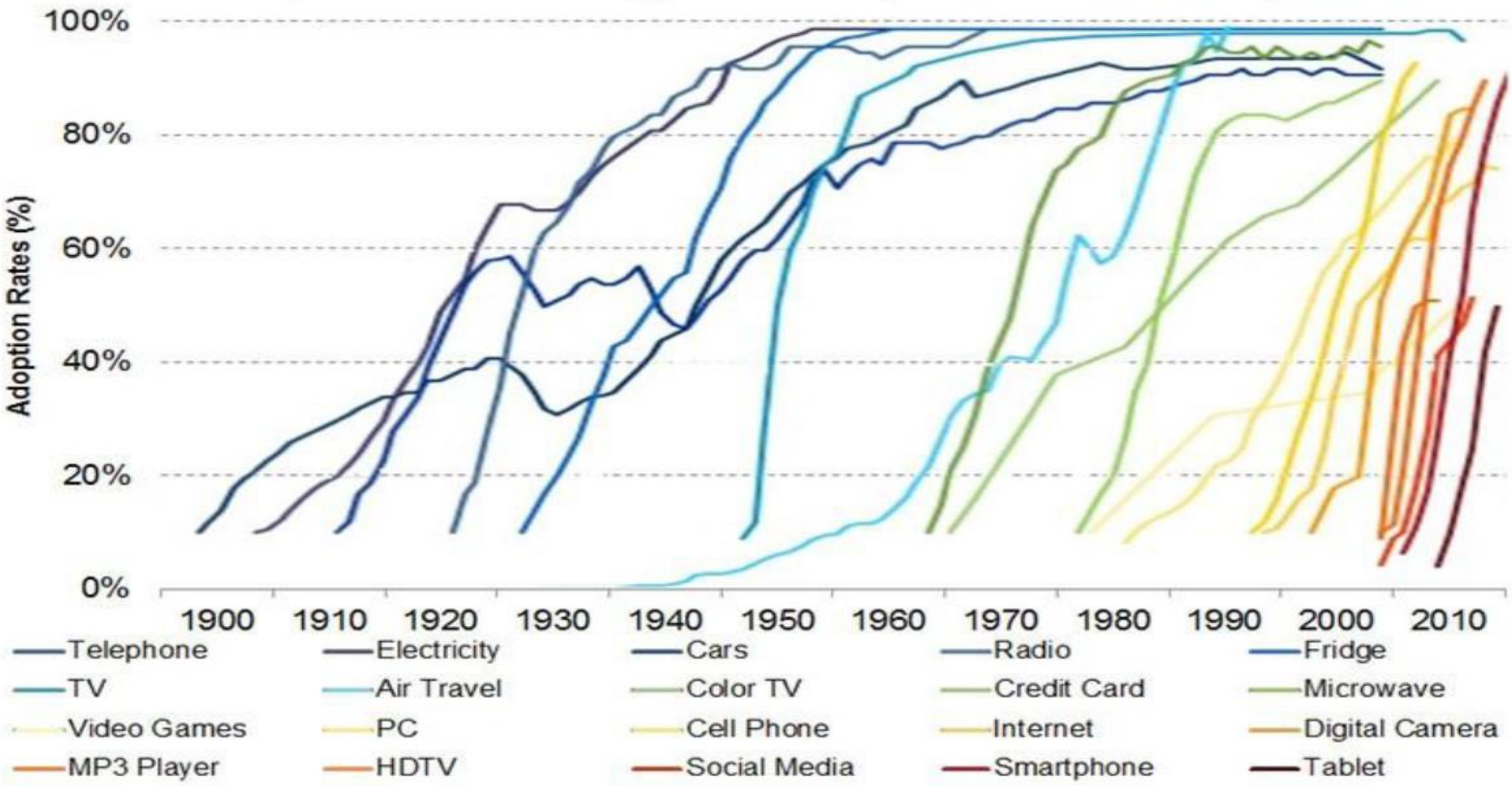
Agenda:

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



Why AI ?

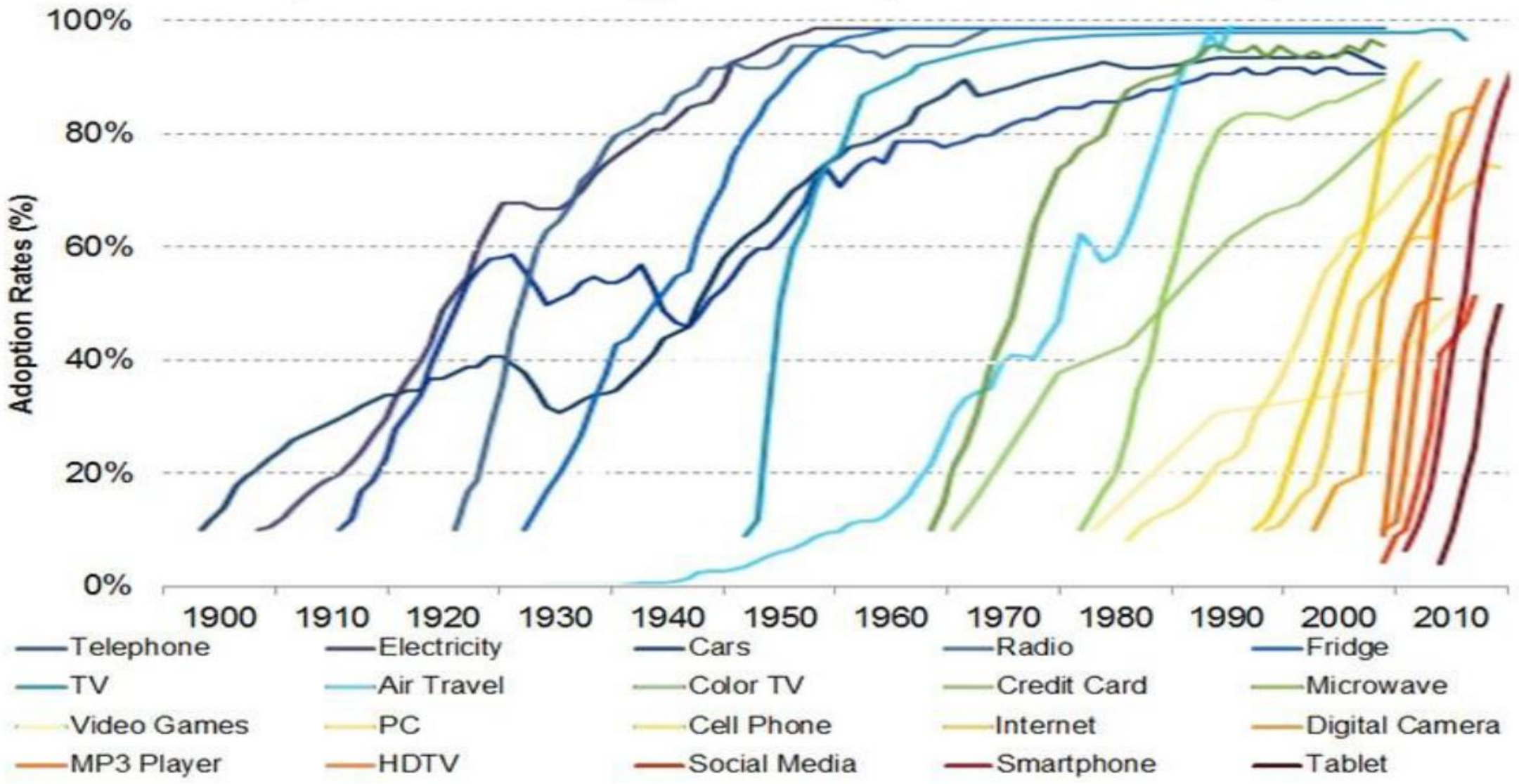
Industrial & consumer trends for last ~125 years



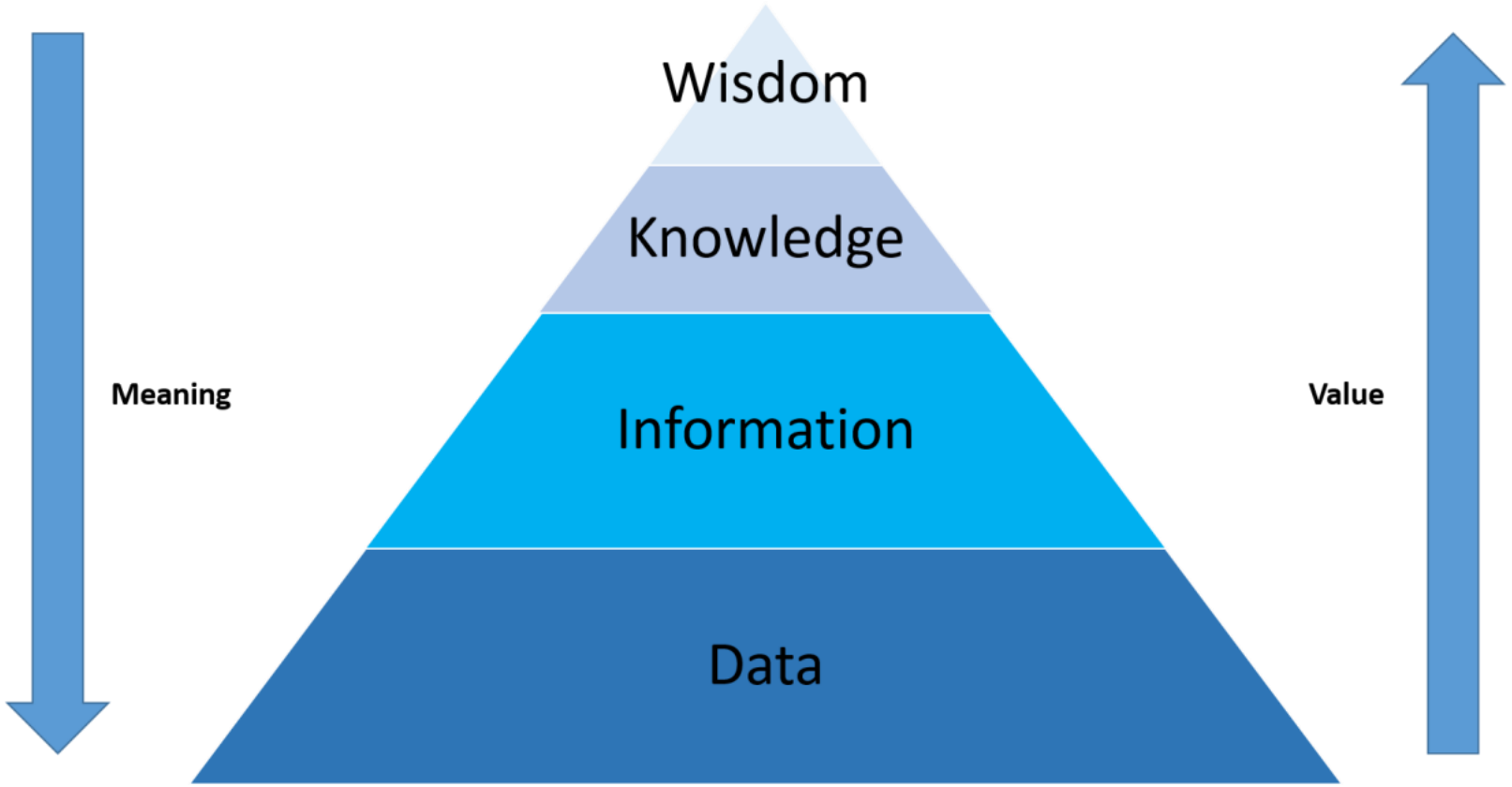
Mechanical + Electrical

Electronic

Data



DIKW Pyramid of Data (Science)

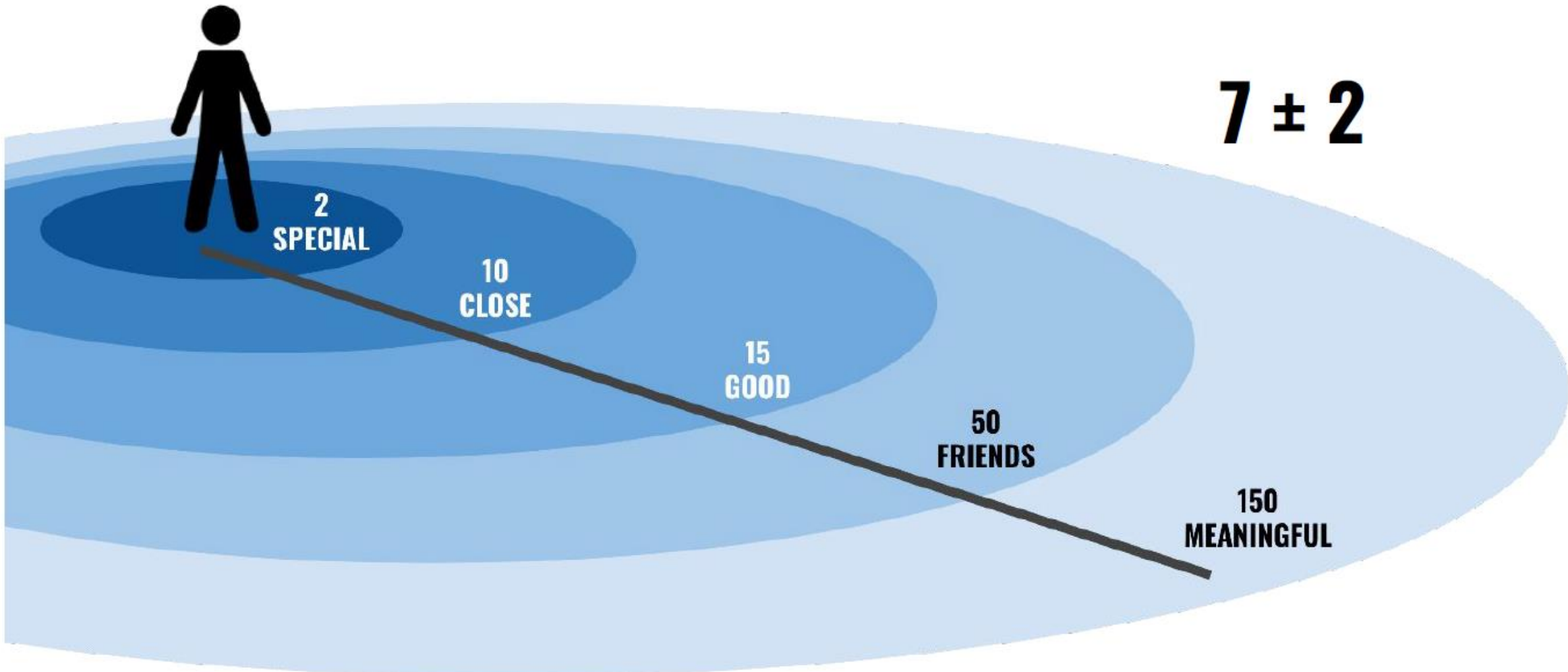


Data >> Value

Time (Speed)

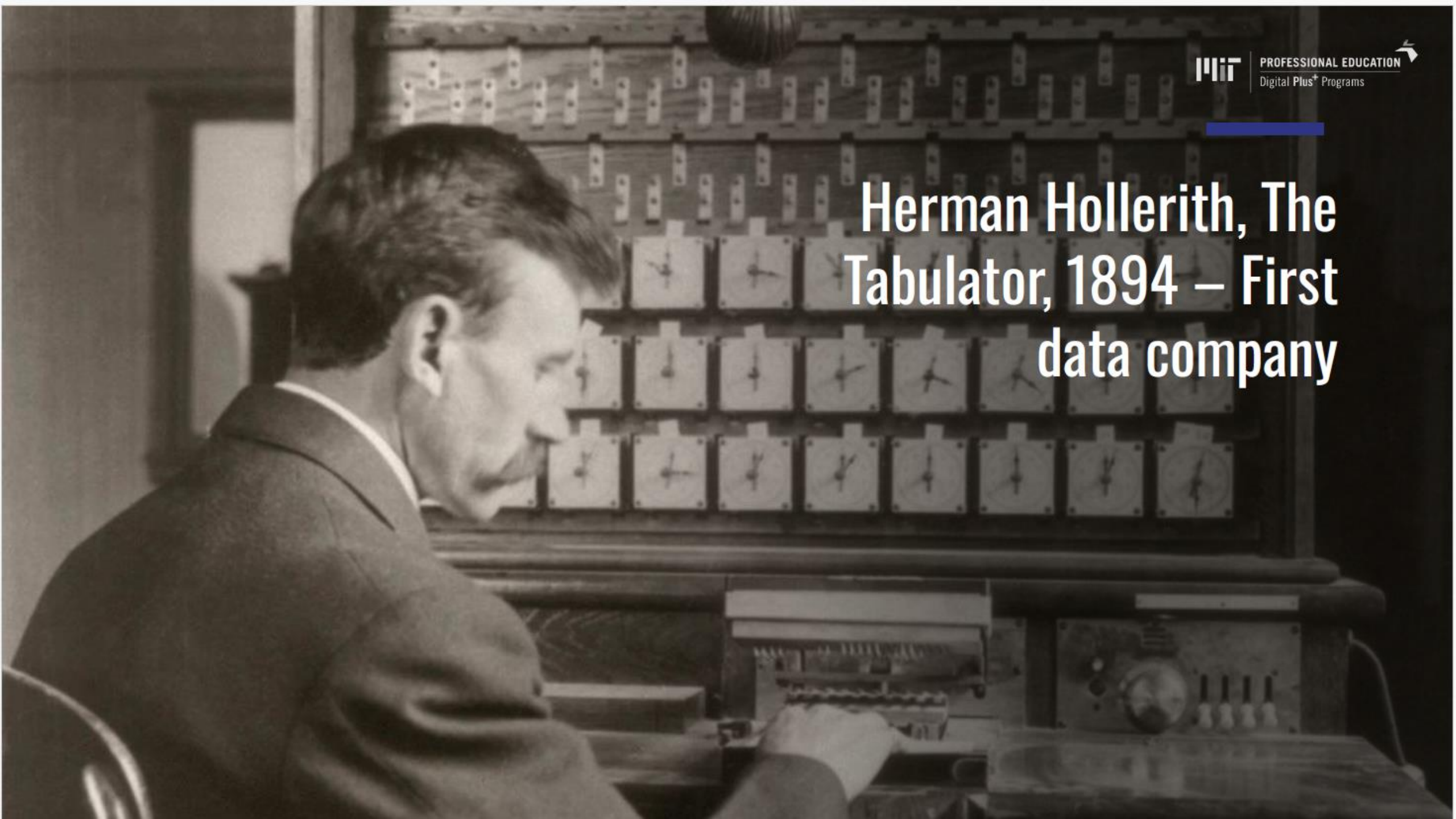
AI (Accelerator)

Human Limitations




PROFESSIONAL EDUCATION
Digital Plus* Programs
We used to collect
data by hand


We used to process it by hand
PROFESSIONAL EDUCATION
Digital Plus+ Programs

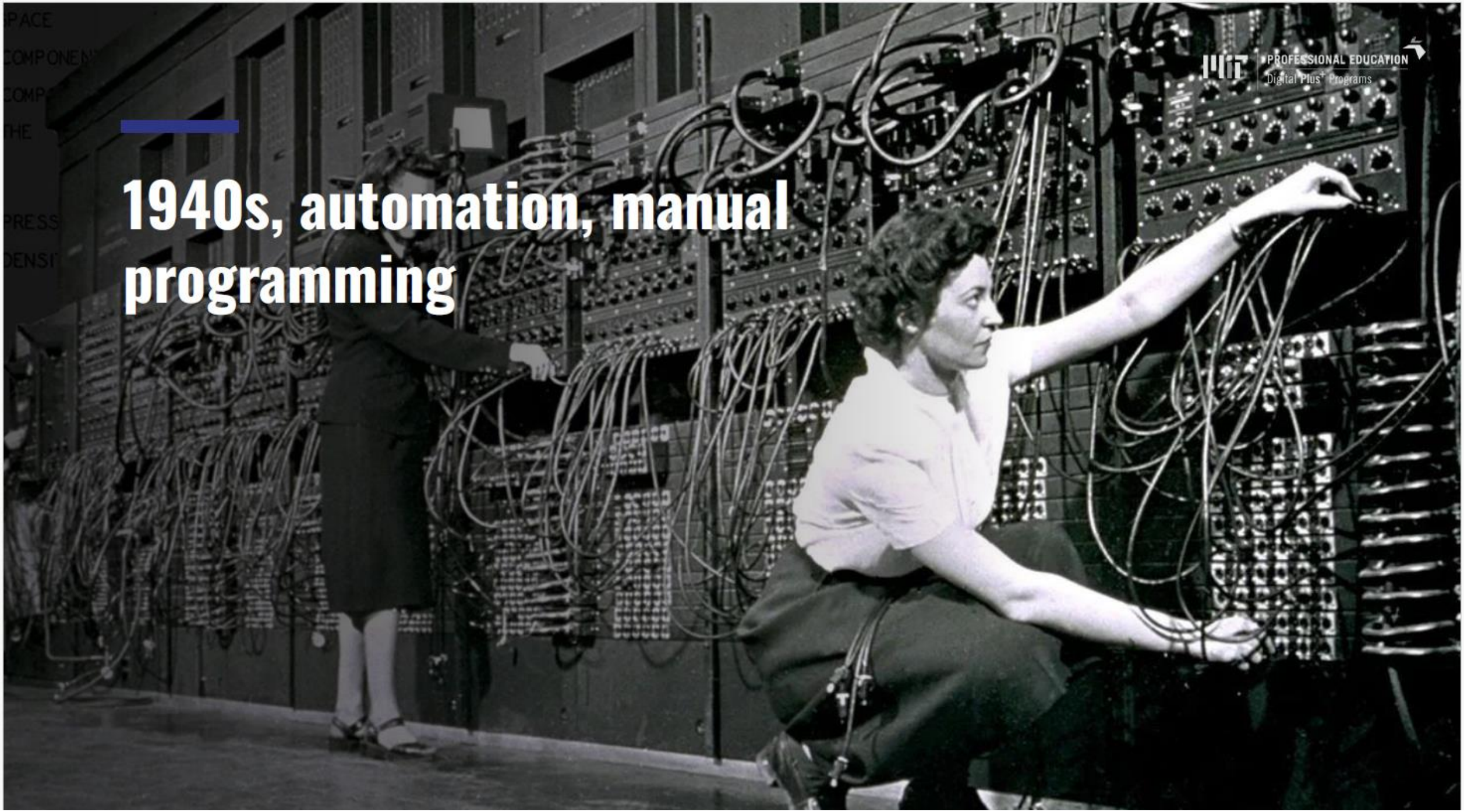


PROFESSIONAL EDUCATION
Digital Plus+ Programs

Herman Hollerith, The Tabulator, 1894 – First data company

PROFESSIONAL EDUCATION
Digital Plus+ Programs

1920s, human layers of integration



1940s, automation, manual programming

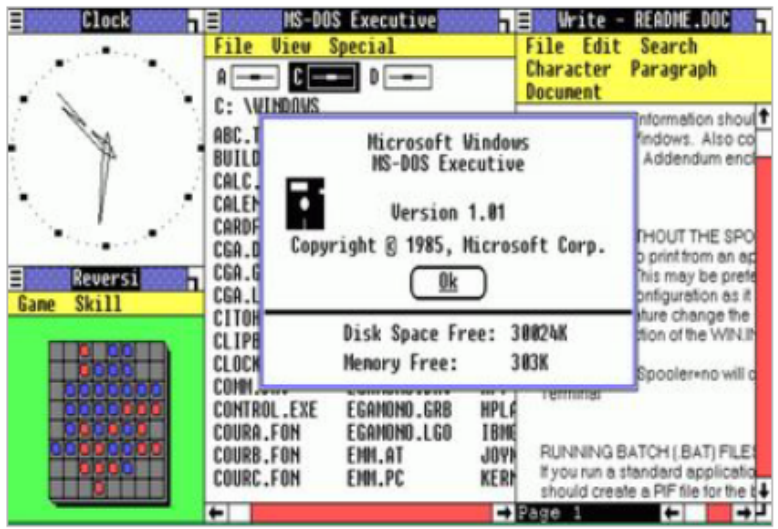
1960s, capturing instructions

1982

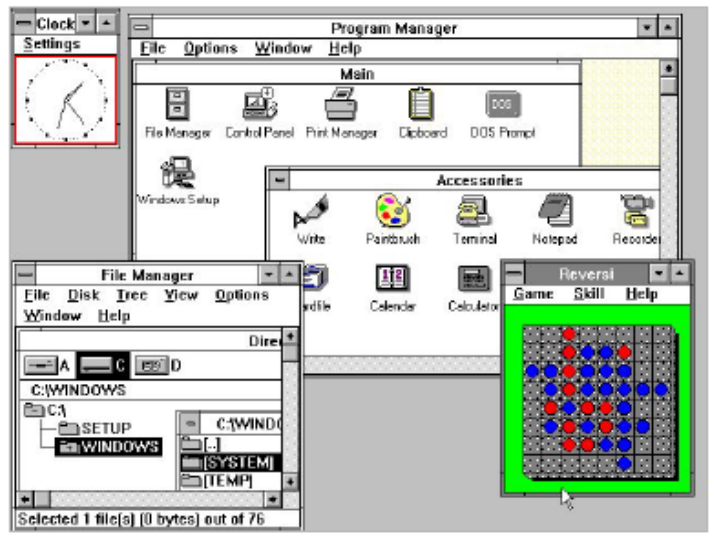
Really easy, right?

Graphical User Interface (GUI)

Windows 1.0



Windows 3.0



Windows 95

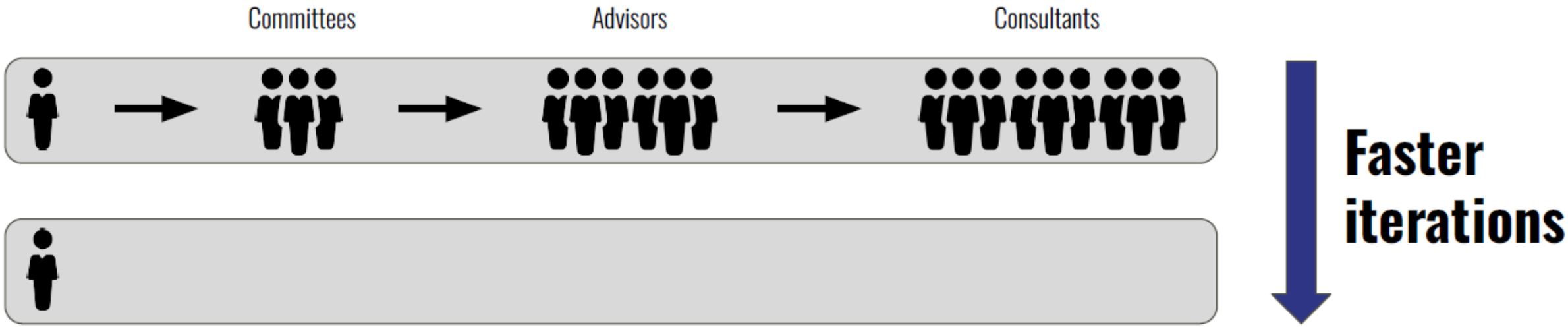
80s/90s, GUIs

We have have arrived at making requests in plain language

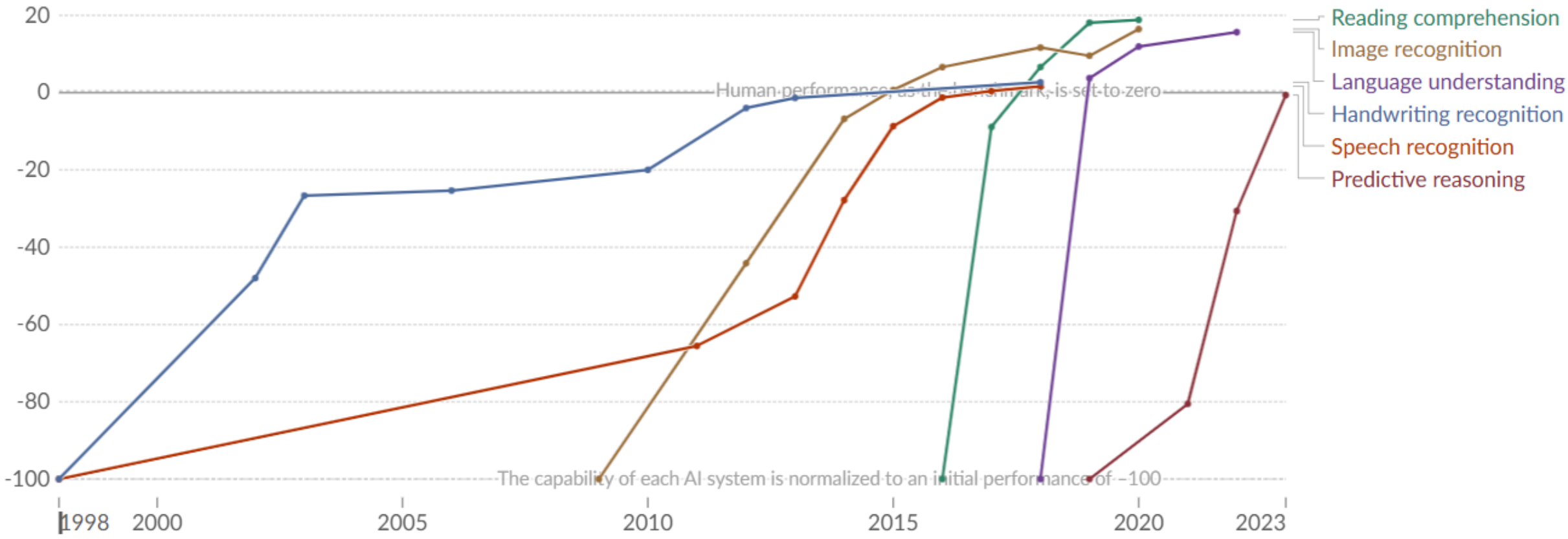
(not like Alexa)

Generative AI

Post-ChatGPT, Moving at the speed of thought

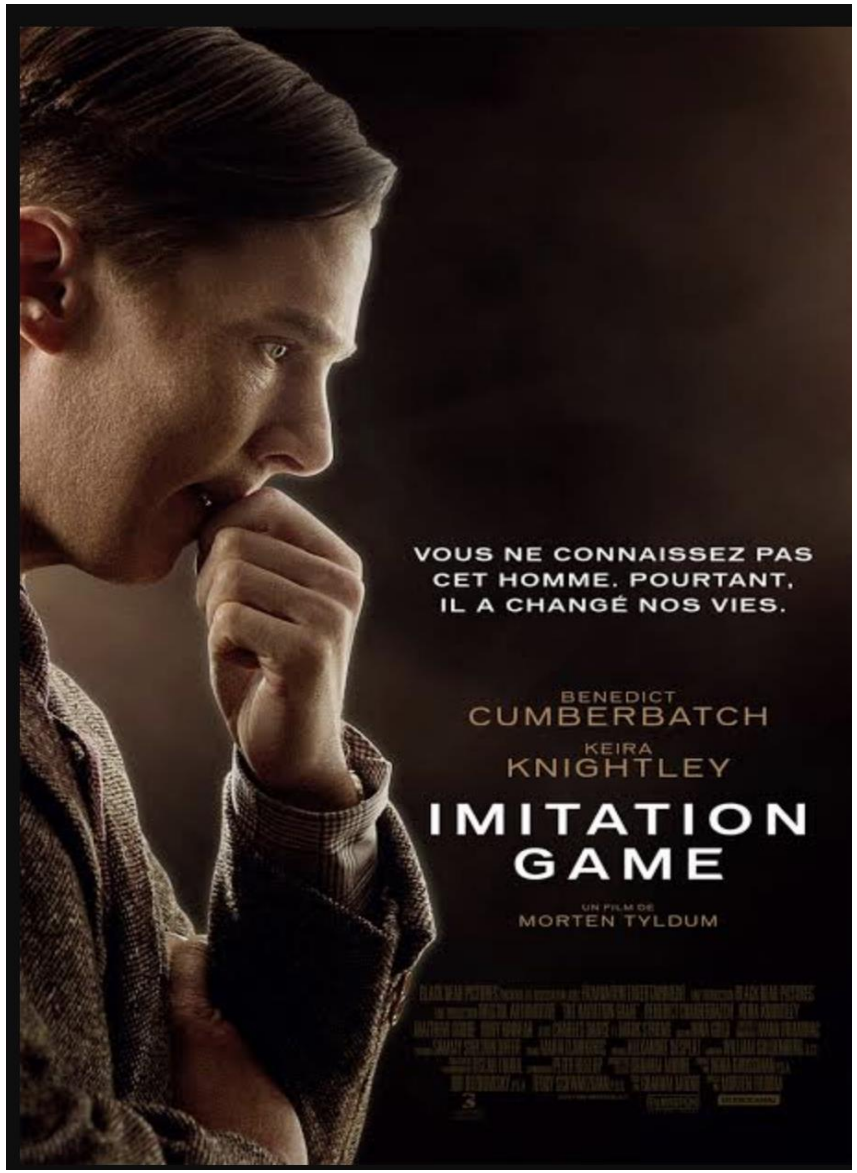


Test scores of AI systems on various capabilities relative to human performance

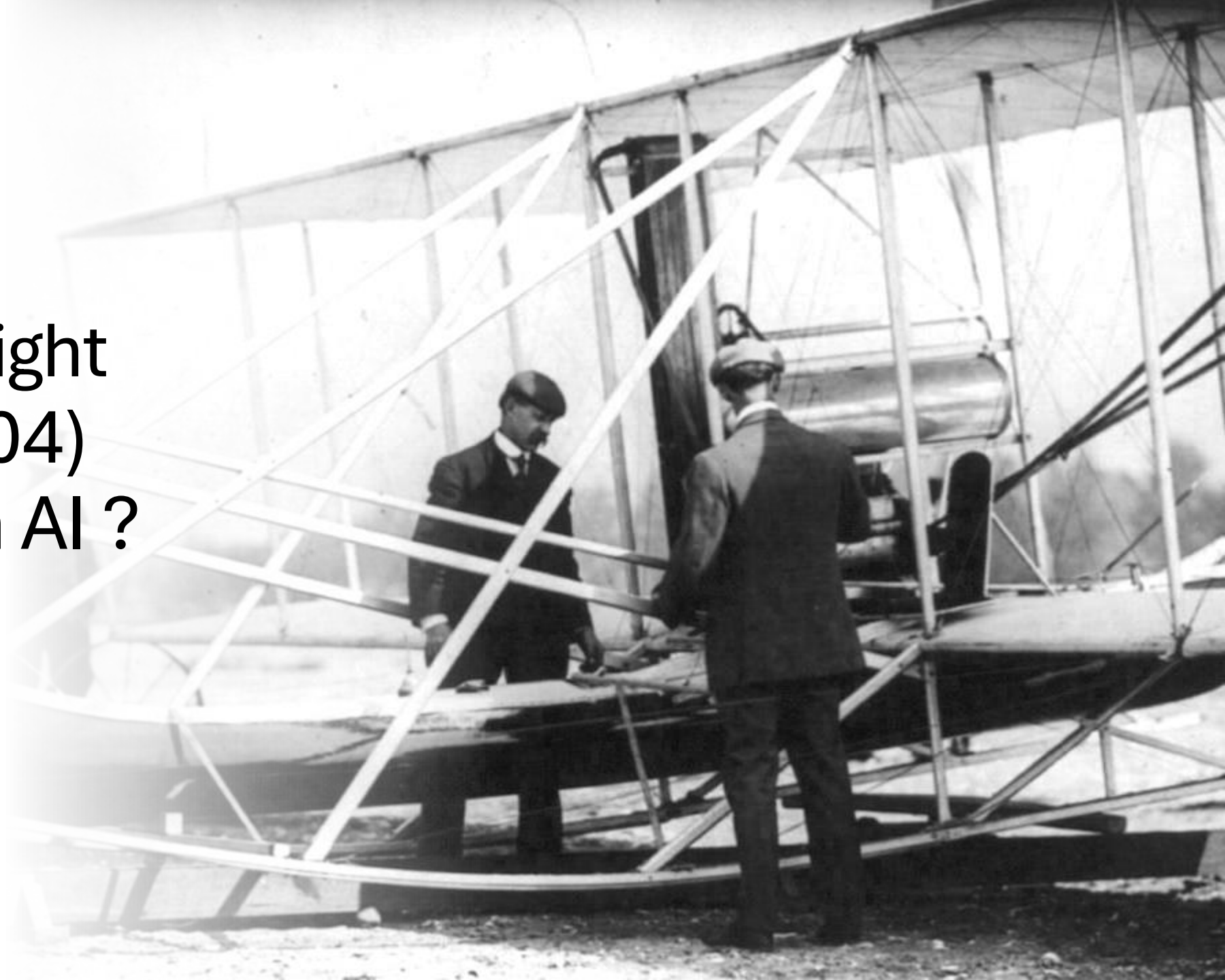


https://ourworldindata.org/grapher/test-scores-ai-capabilities-relative-human-performance

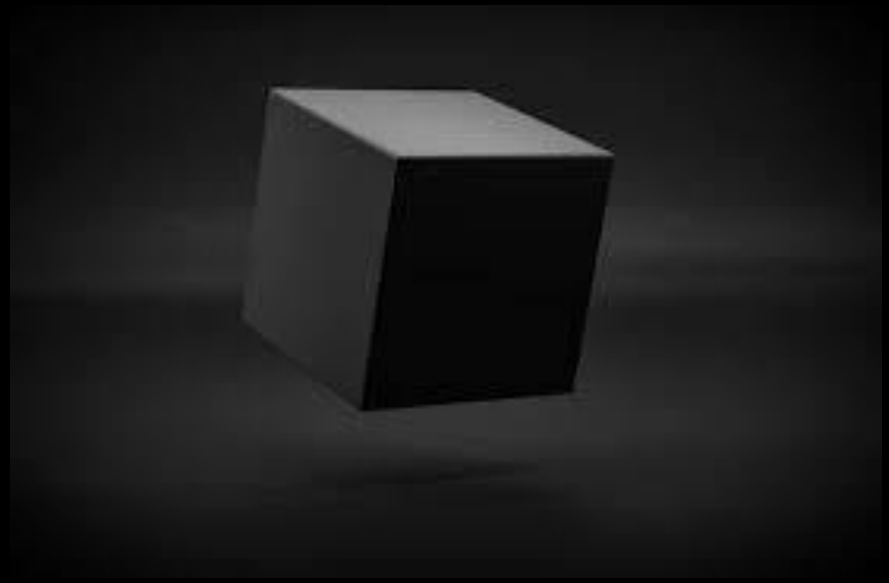
AI Ethics


VOUS NE CONNAISSEZ PAS
CET HOMME. POURTANT,
IL A CHANGÉ NOS VIES.
BENEDICT
CUMBERBATCH
KEIRA
KNIGHTLEY
IMITATION
GAME
UN FILM DE
MORTEN TYLDUM

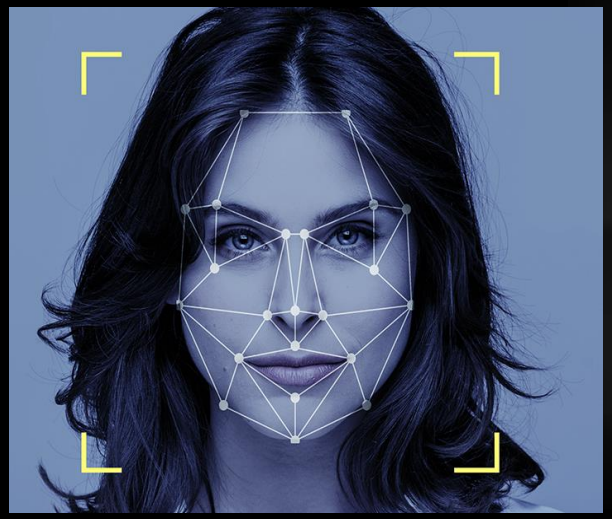
Are we at Wright Brothers (1904) Moment with AI ?

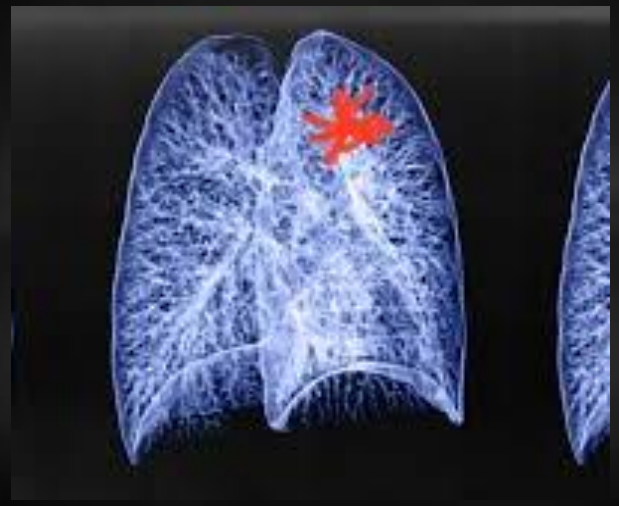


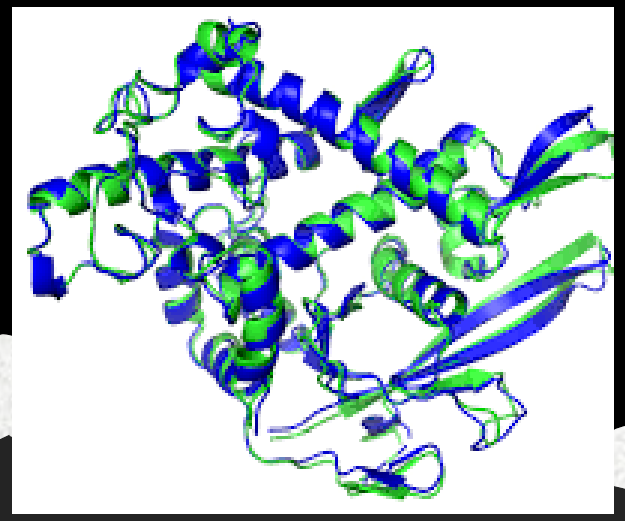
Missing the fundamentals and can't explain current AI capabilities work

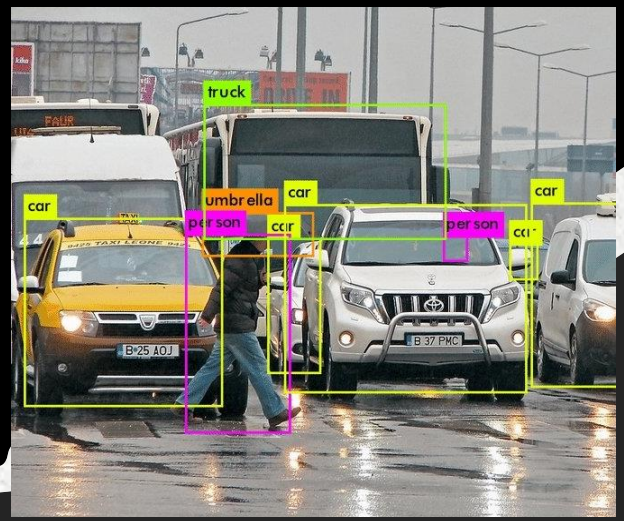


AI Success Stories ..

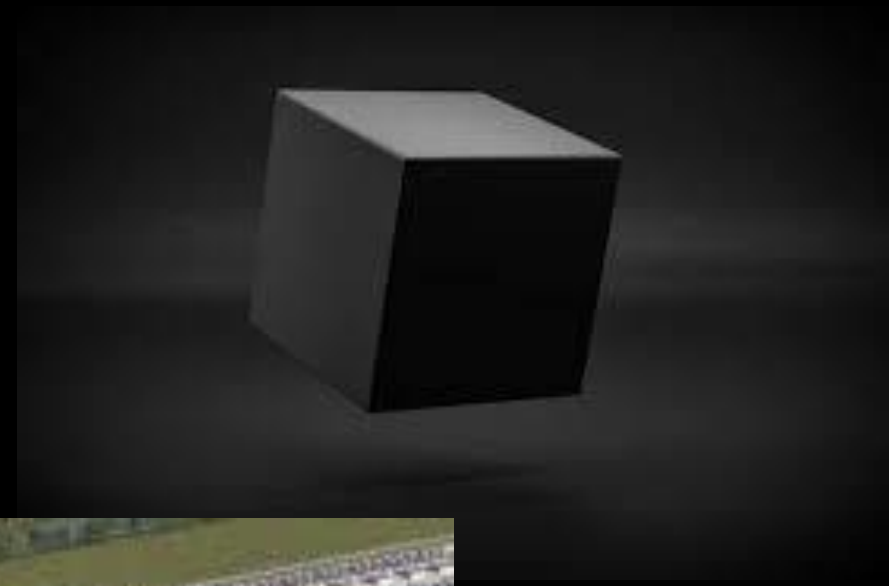






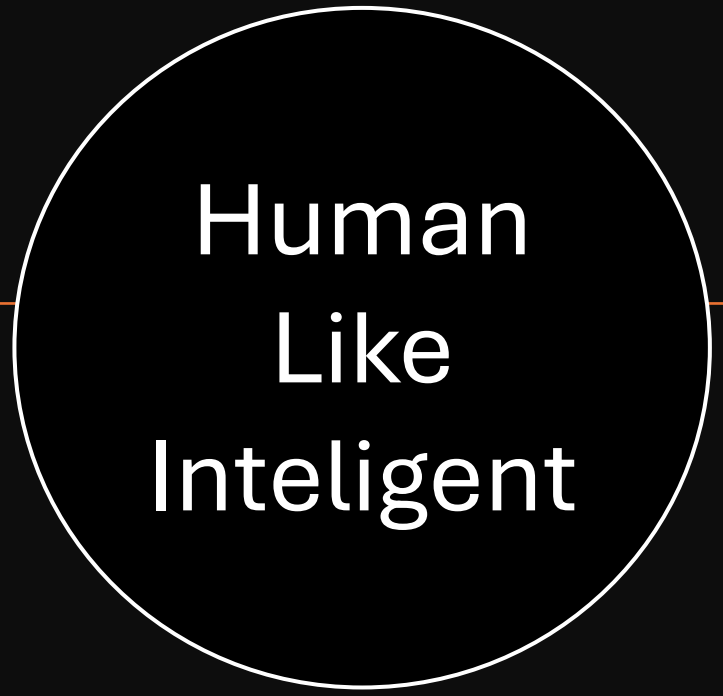


AI failed stories..



Machines work with human

Human Like Inteligent



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

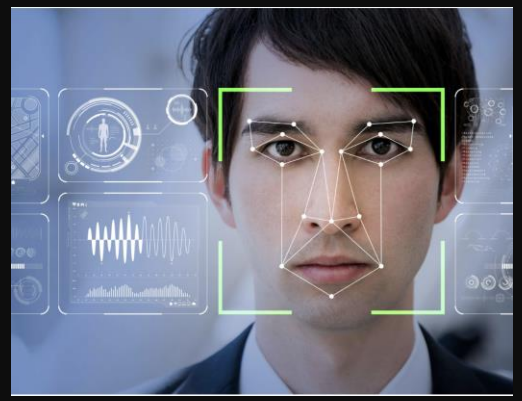
Other biased, but not protected classes.

- 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.

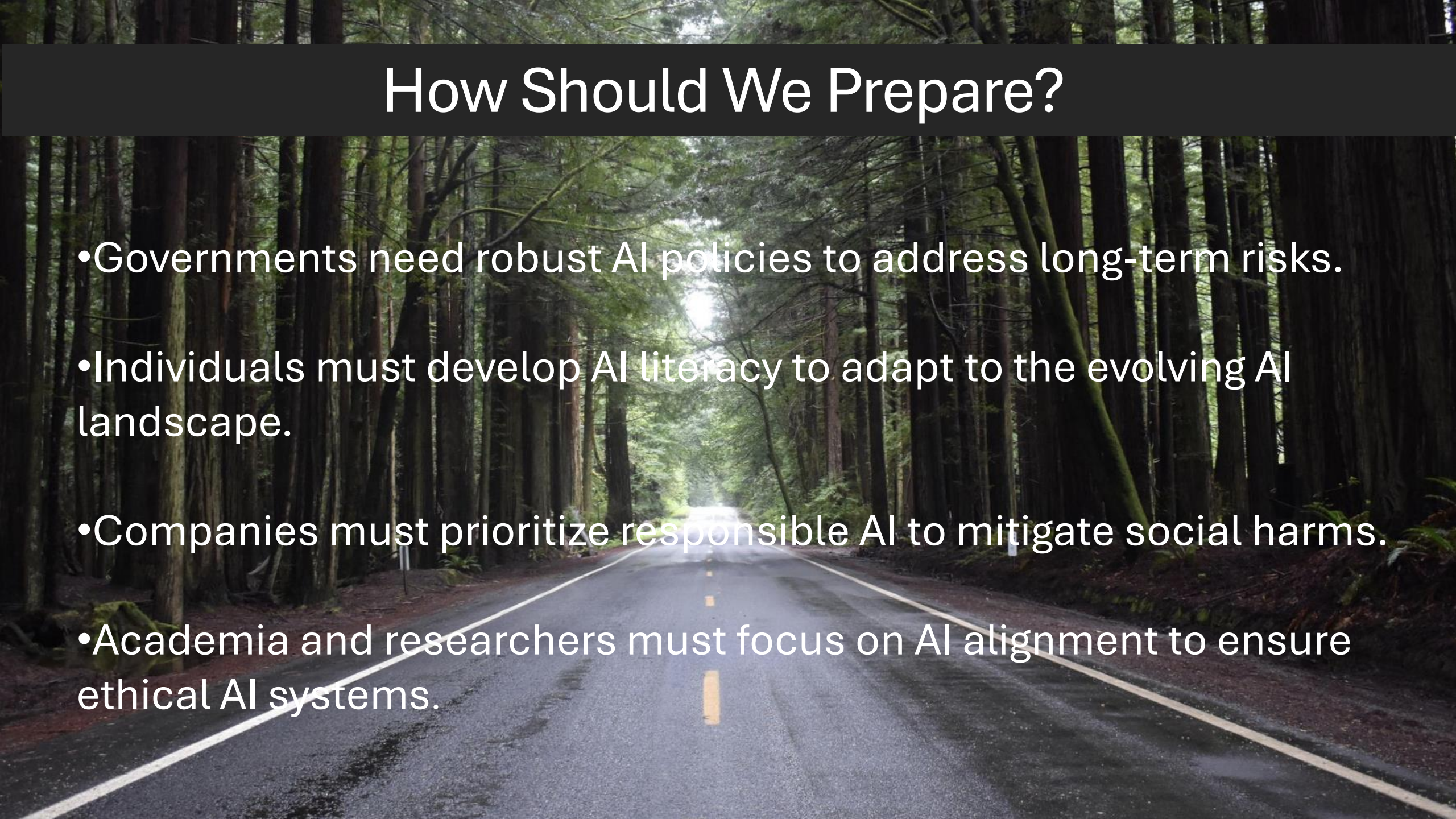
Sustainability

- 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 assert
- **AI taxation**/Universal Basic Income (UBI)
- **AI-augmented human**

How Should We Prepare?



- 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.