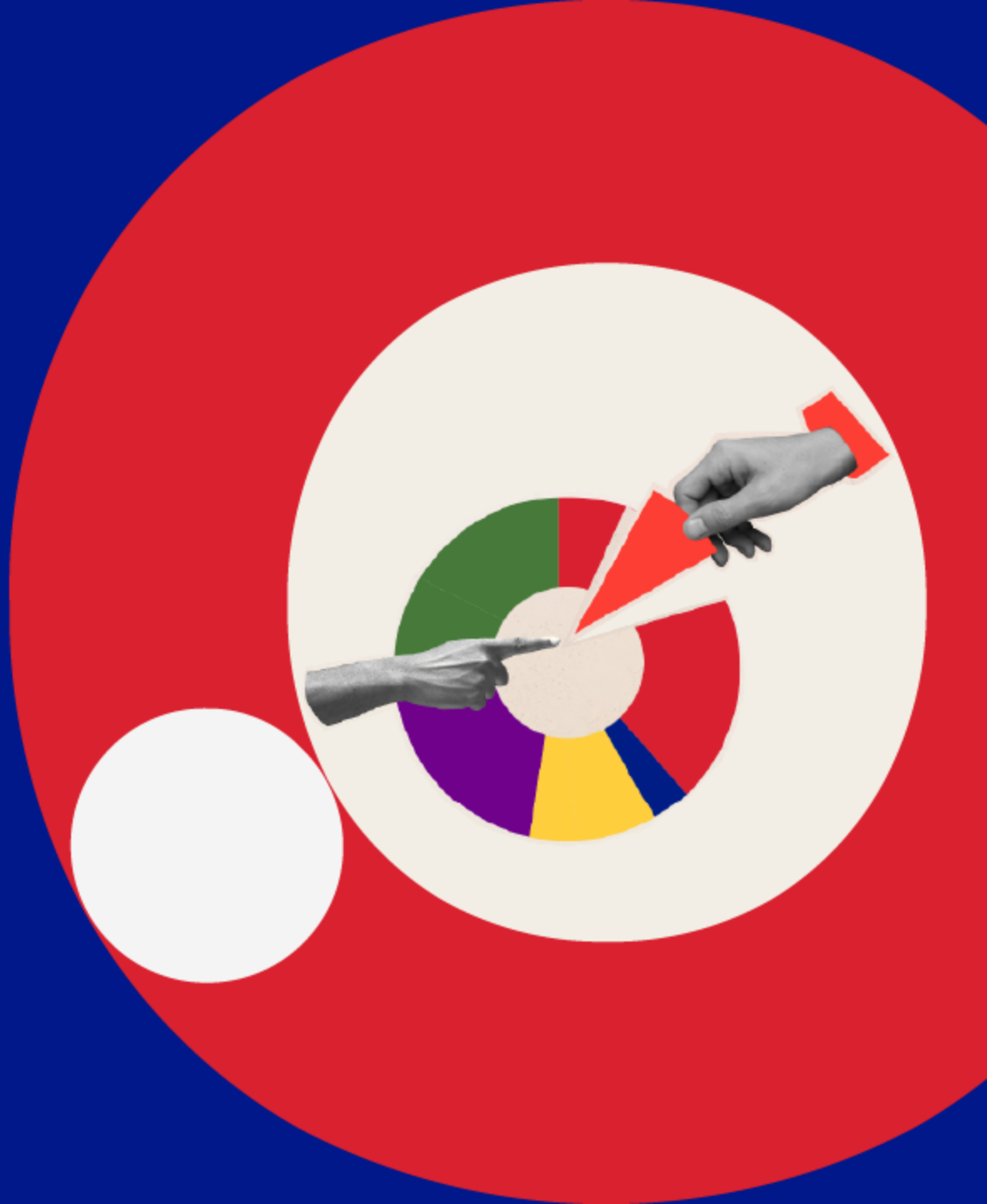
WEBINAR

Open data, academia, and ethics: responsible use of AI-generated and synthetic data in research



data.europa academy

5 June 2026

10:00 – 11:00 CEST

Rules of the game

The webinar will be recorded and published on the data.europa academy

For questions, please use the ClickMeeting chat

Please reserve 3 min after the webinar to help us improve by filling in our feedback form

data europa academy

Today's speakers

Flora Kopelou
European Data Portal, Publications Office of the EU

Dr Thomas Lampert
Professor of Computer Science, University of Strasbourg

Dr Ella Hafermalz
Associate Professor, Vrije Universiteit Amsterdam

data europa academy

Indicative agenda

10.00 – 10.05	Opening and introduction – *Flora Kopelou*
10.05 – 10.25	Introduction: AI-generated and synthetic data – *Dr. Thomas Lampert*
10.25 – 10.45	How generative AI and synthetic data reshape trust and research – *Dr. Ella Hafermalz*
10.45 – 11.00	Q&A and closing remarks – *Flora Kopelou*

data europa academy

AI-generated and synthetic data

Prof Thomas Lampert

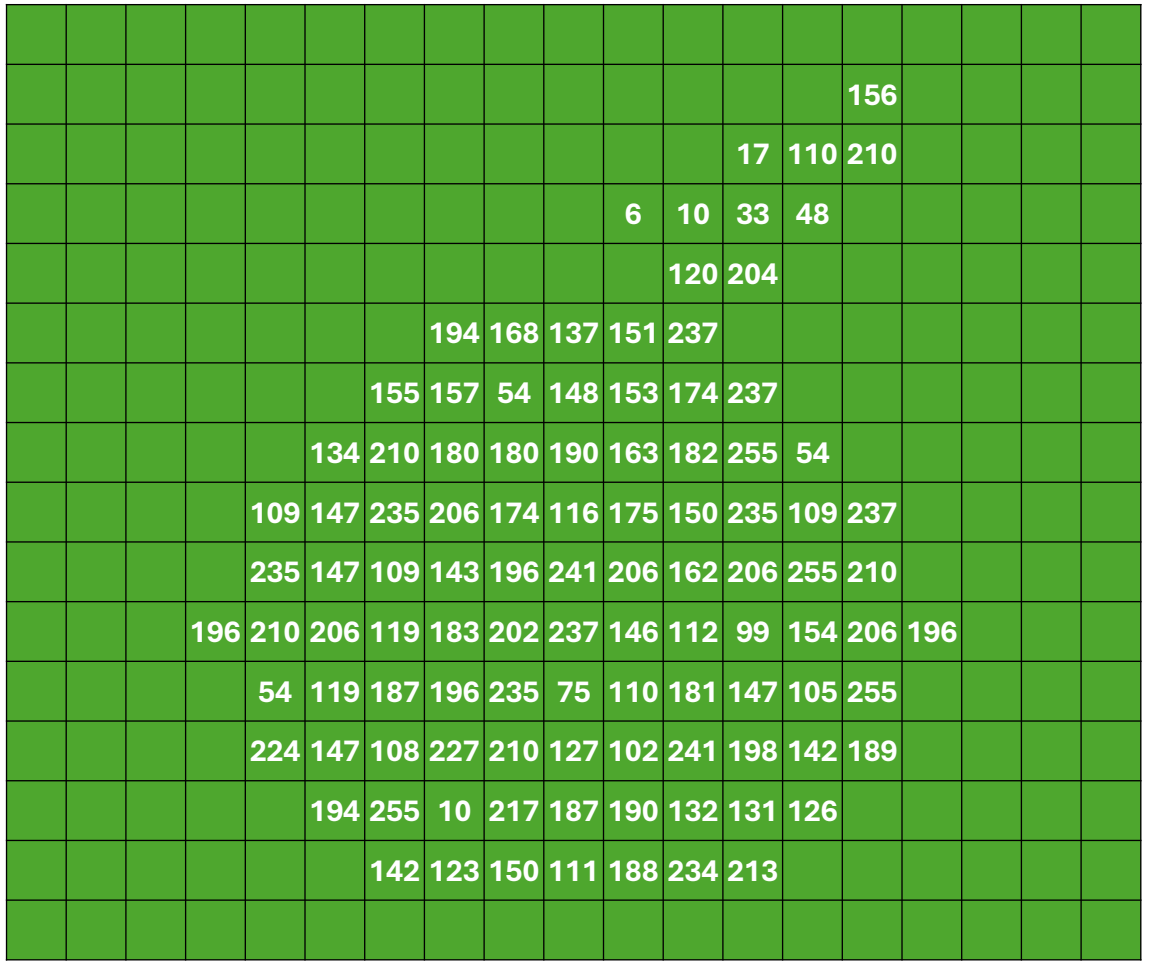
University of Strasbourg

Machine Learning

- Data
  - Images

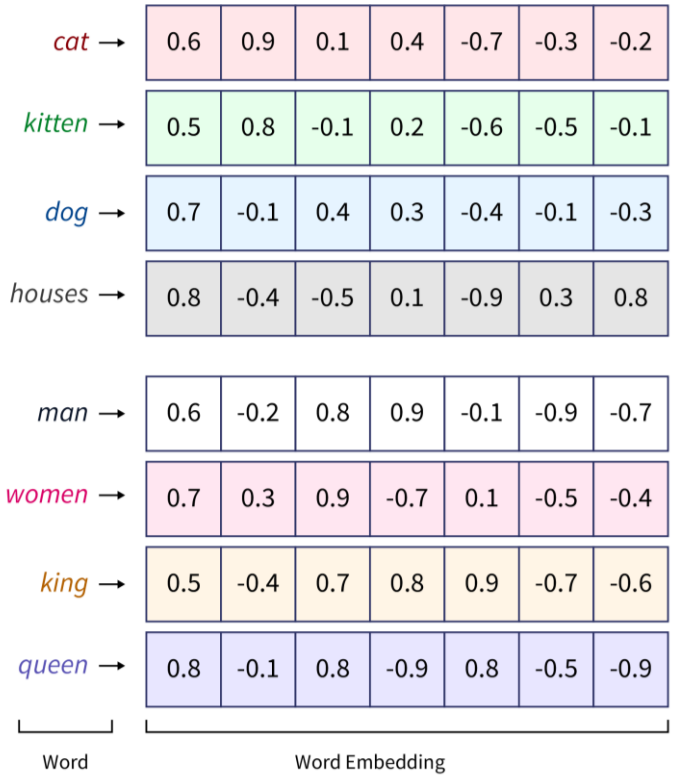
Machine Learning

- Data
  - Images



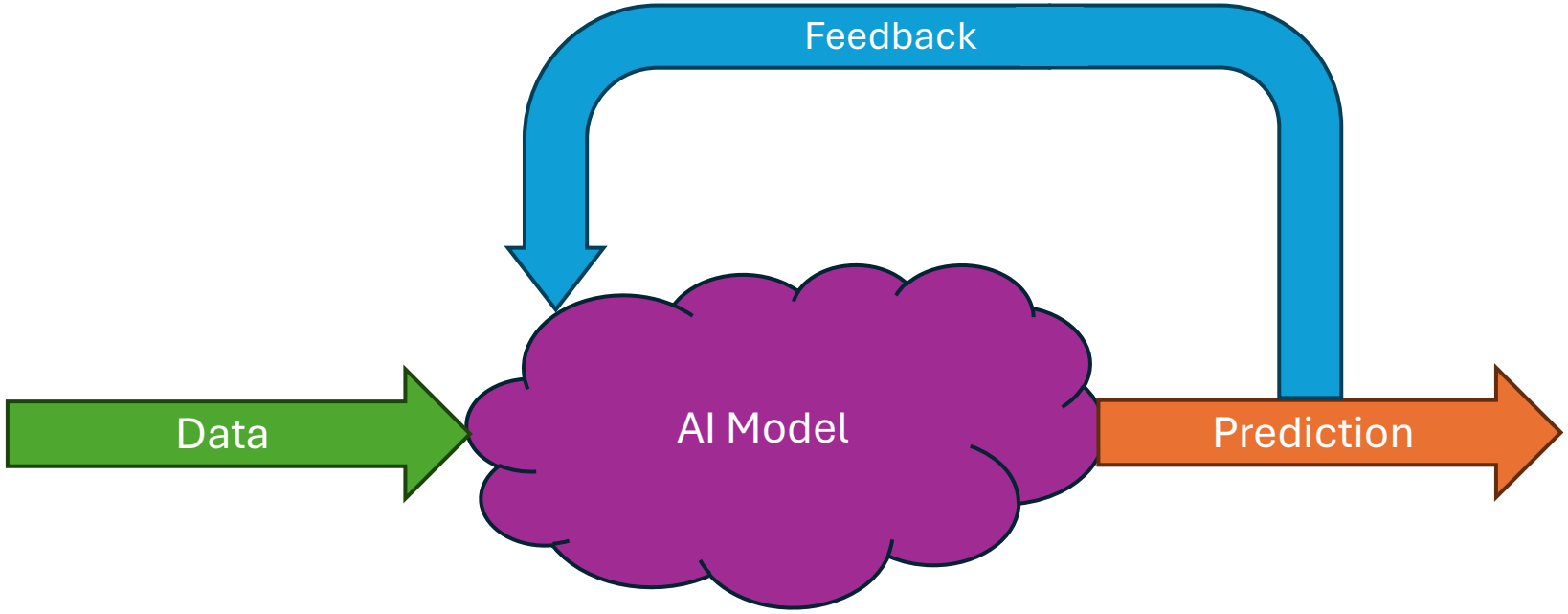
Machine Learning

- Data
  - Text



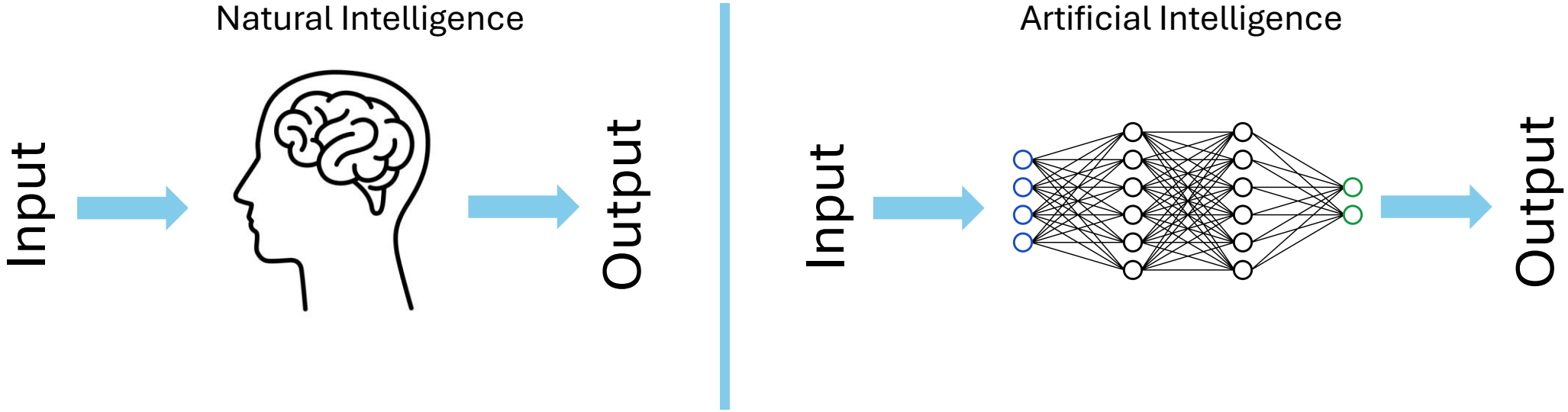
N. Malingan, Word Embeddings with TensorFlow, Scaler Topics

The Machine Learning Process



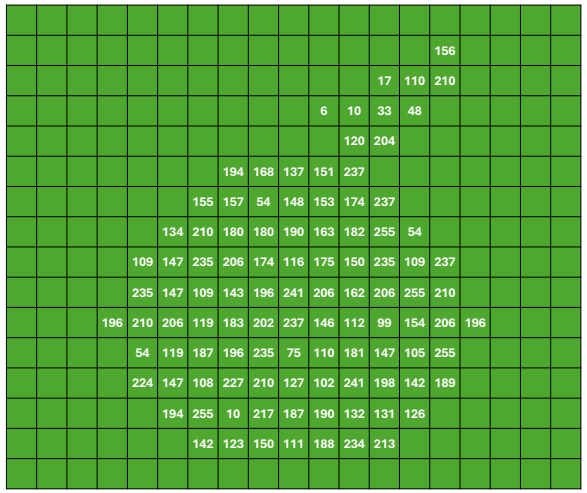
Machine Learning

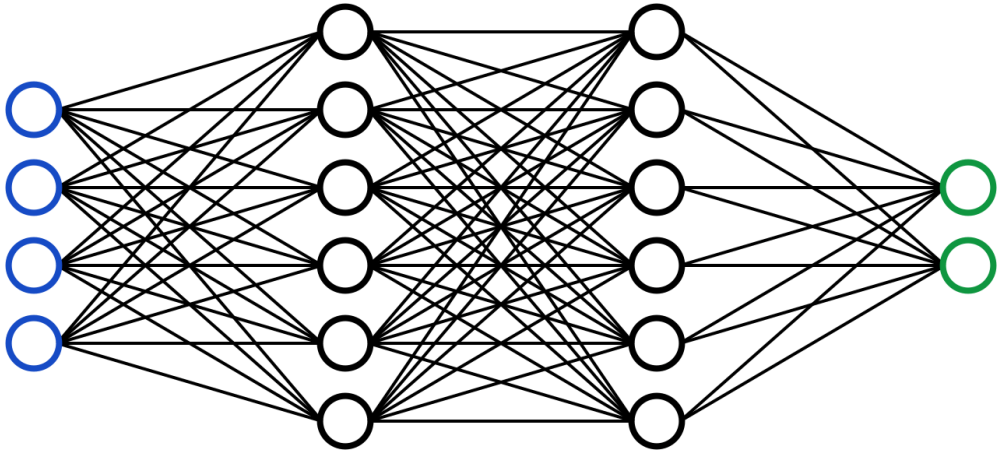
- (Current) AI reproduces the results and not the cognitive process

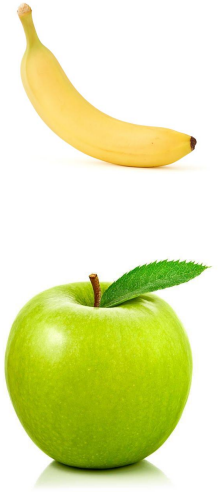


Machine Learning

- Supervised Learning







Machine Learning

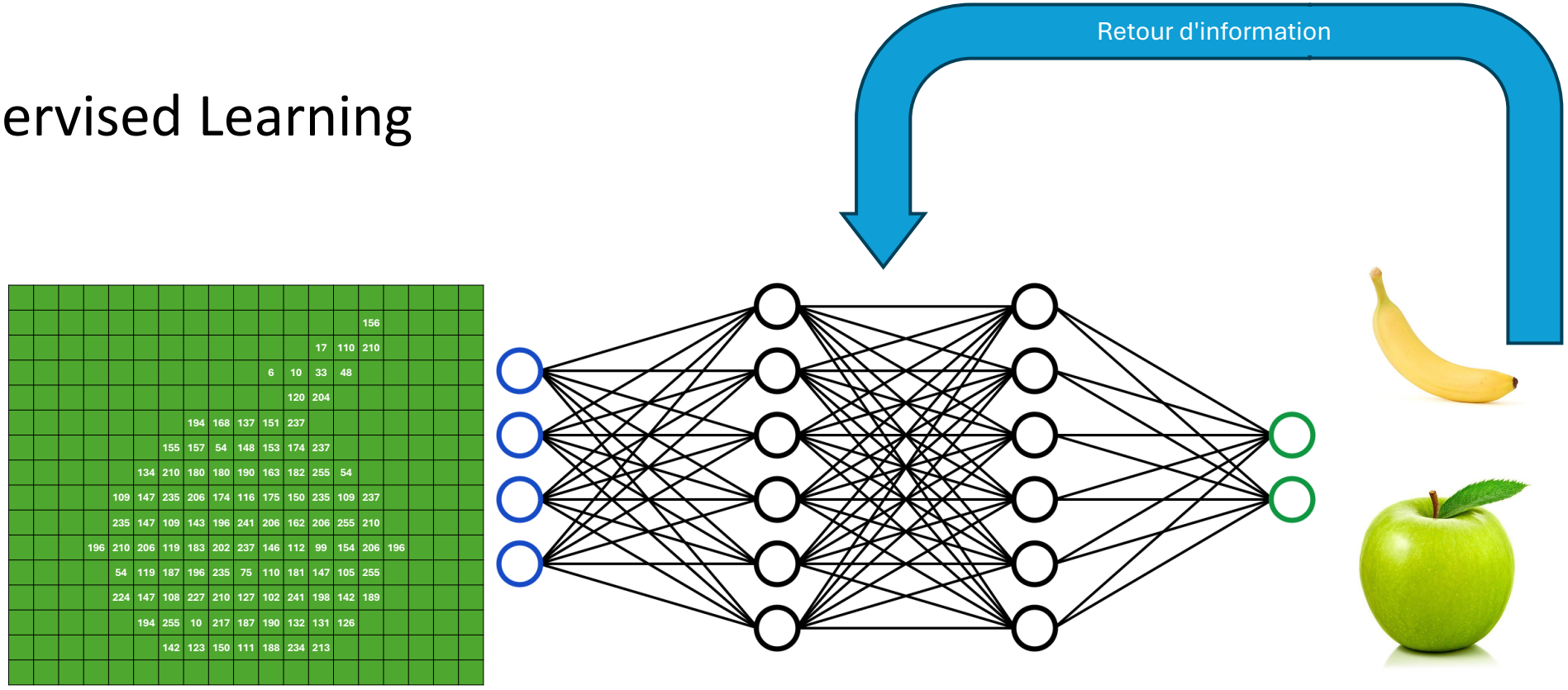
- Data

Bananas

Apples

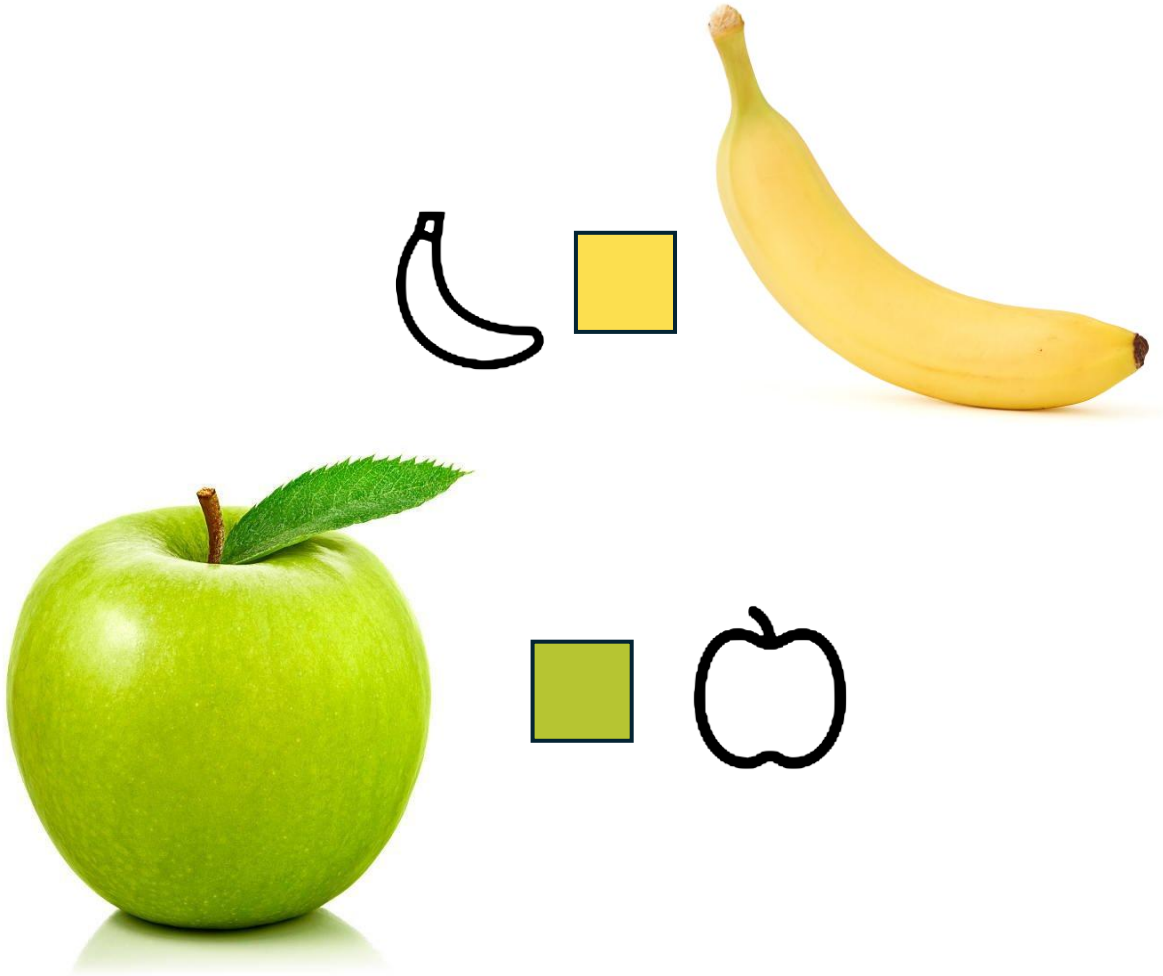
Machine Learning

- Supervised Learning



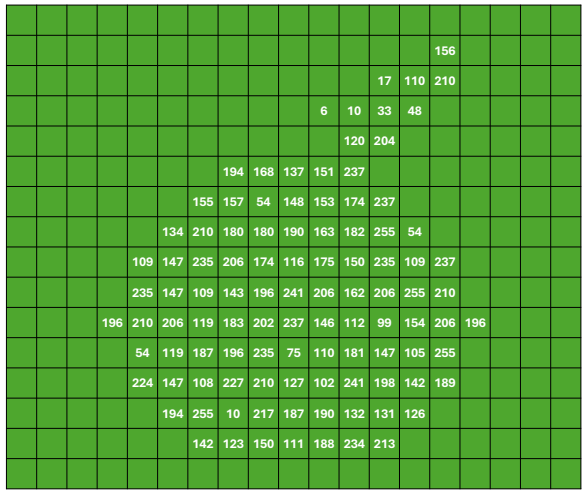
Machine Learning

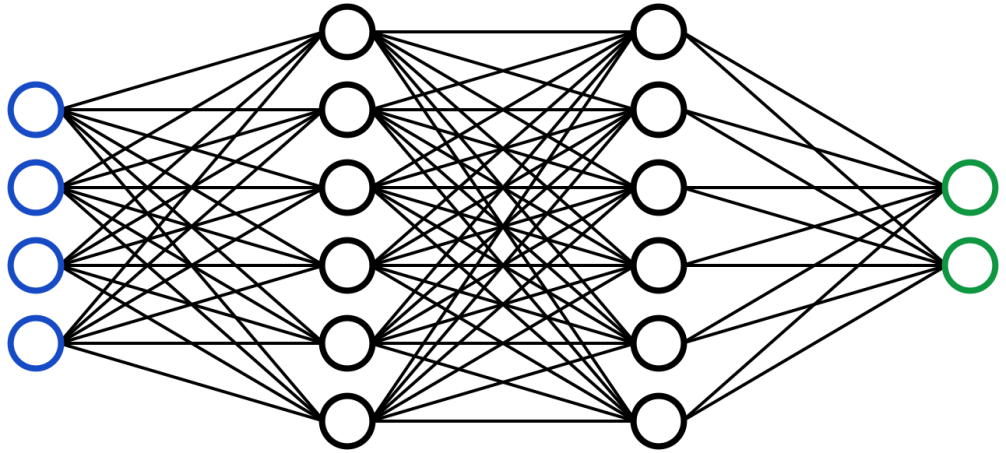
- Supervised Learning



Machine Learning

- Supervised Learning





Generative Models

- Data

Bananas

Apples

Generative Models

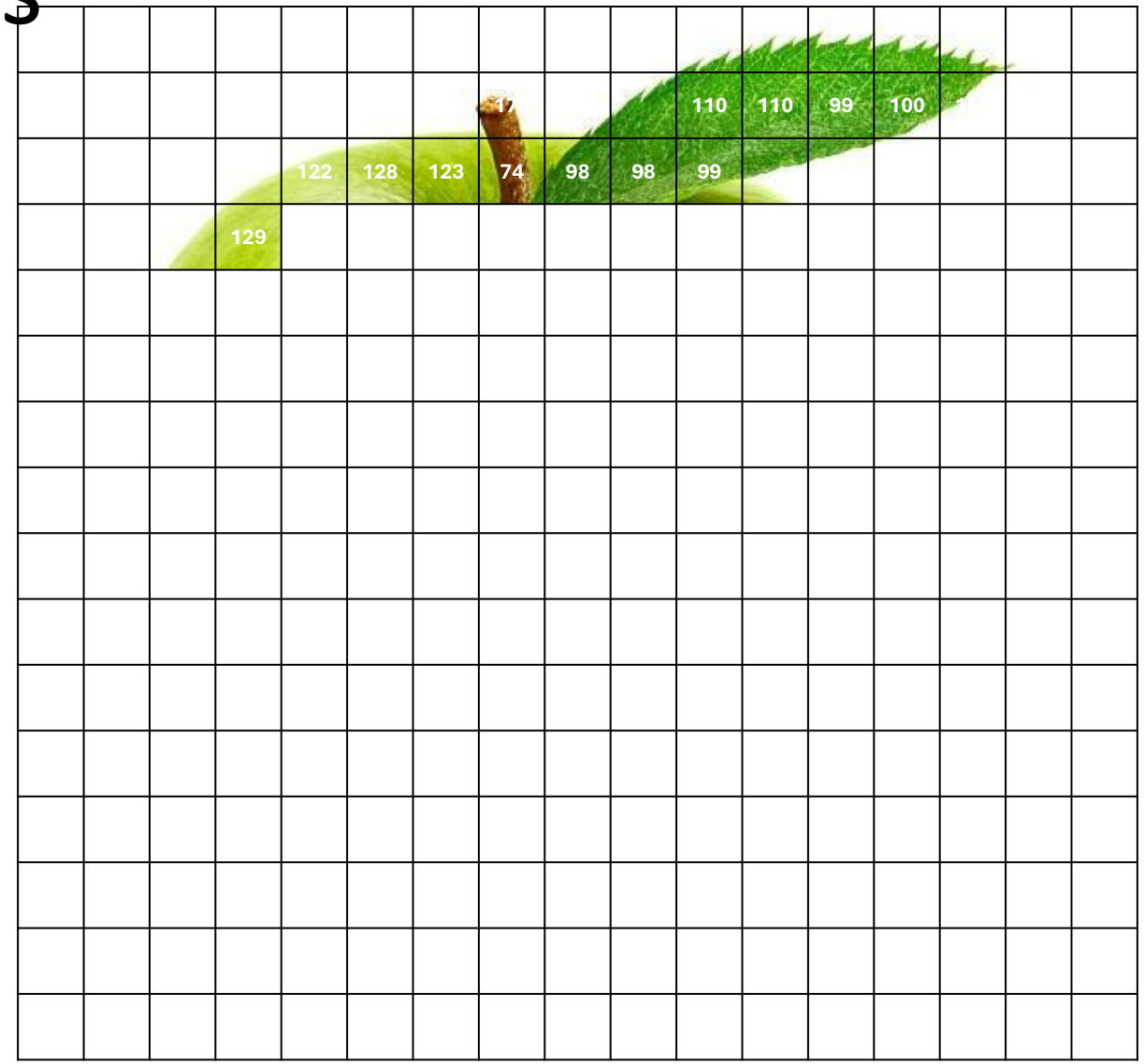
- Data

~~Bananas~~

~~Apples~~

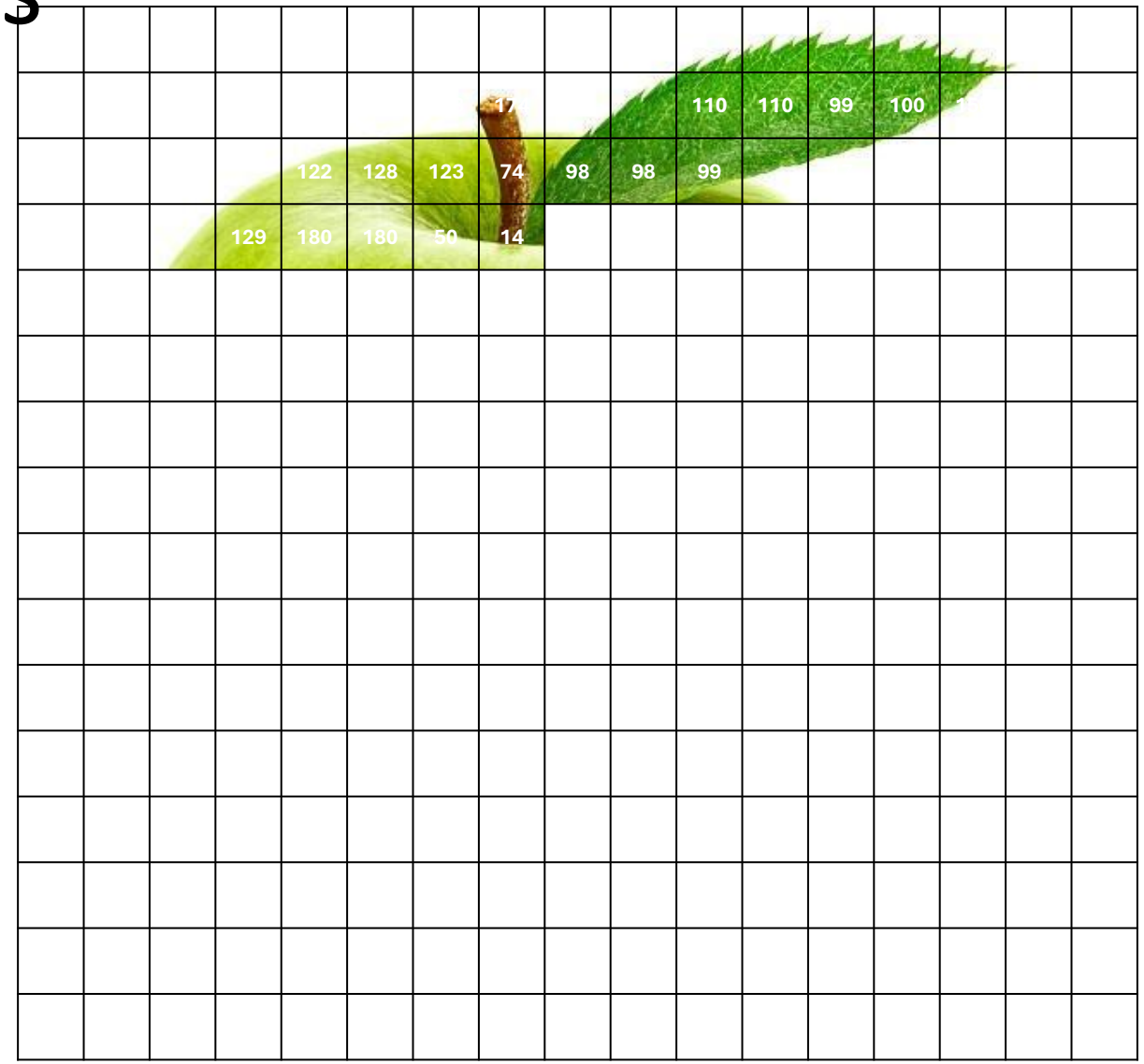
Generative Models

- Unsupervised Learning
  - Generative Models



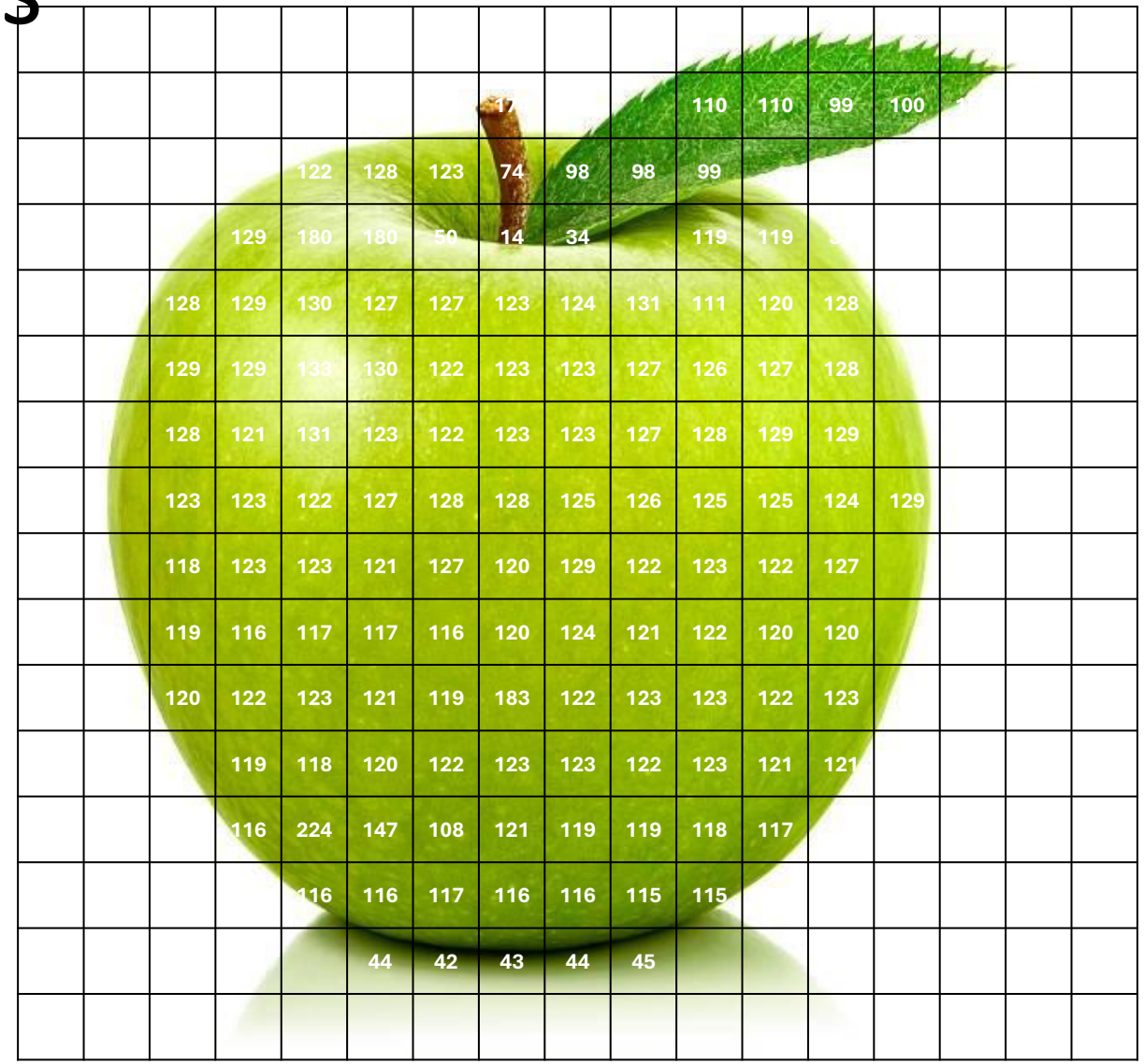
Generative Models

- Unsupervised Learning
  - Generative Models



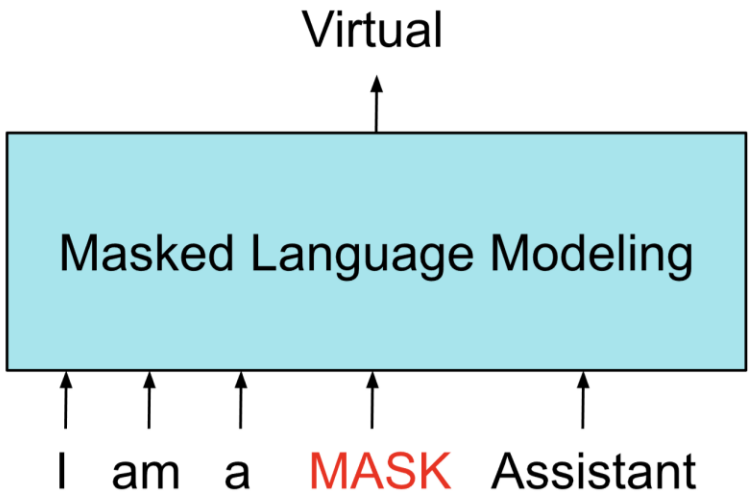
Generative Models

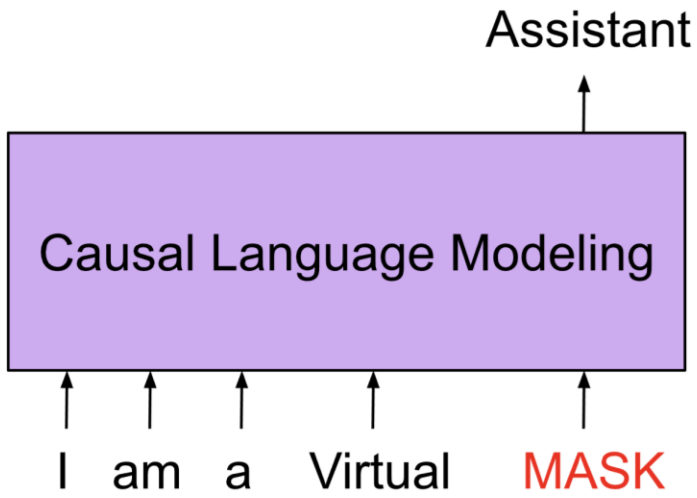
- Unsupervised Learning
  - Generative Models



Generative Models

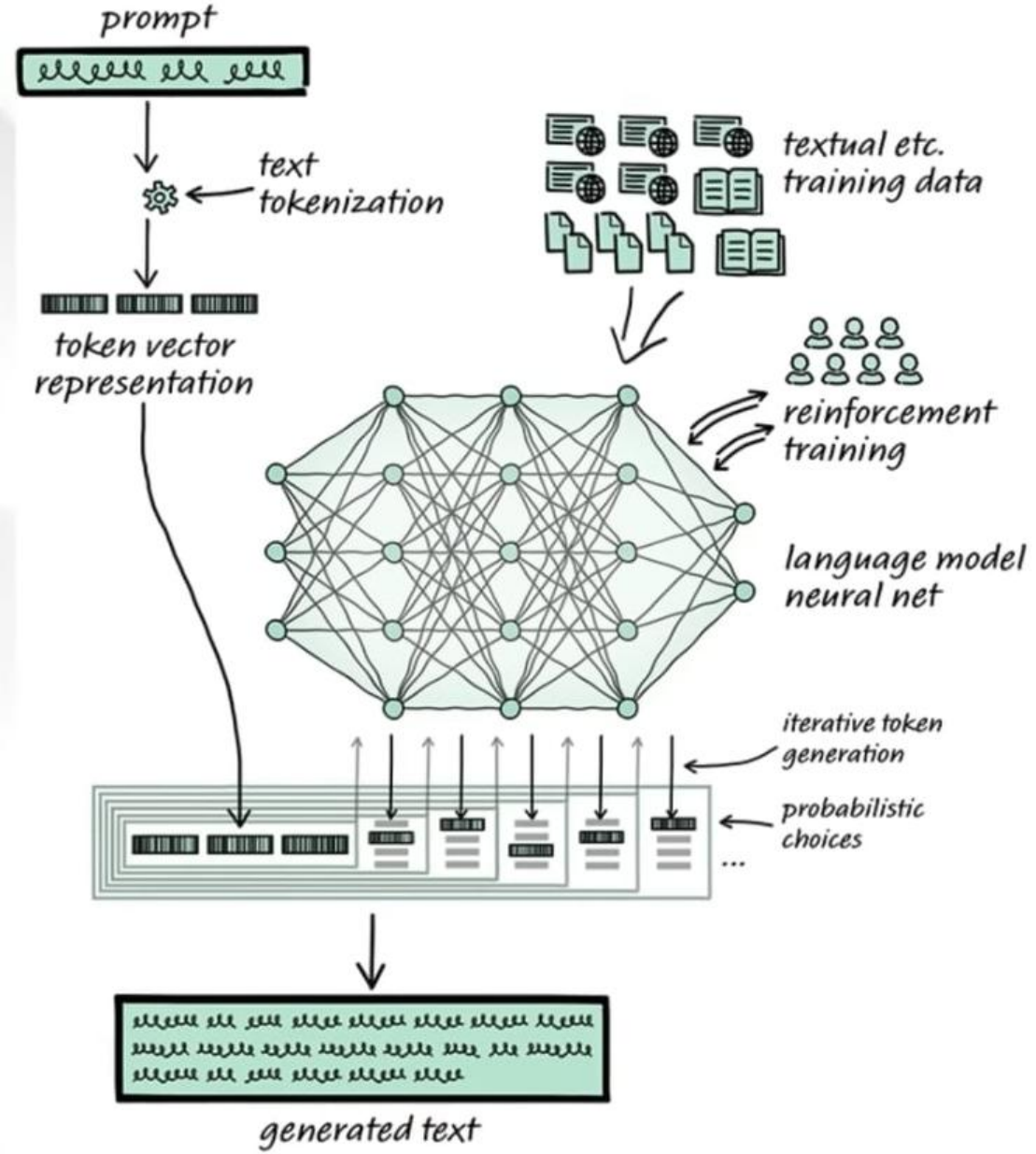
- Unsupervised Learning
  - Generative Models





Generative Models

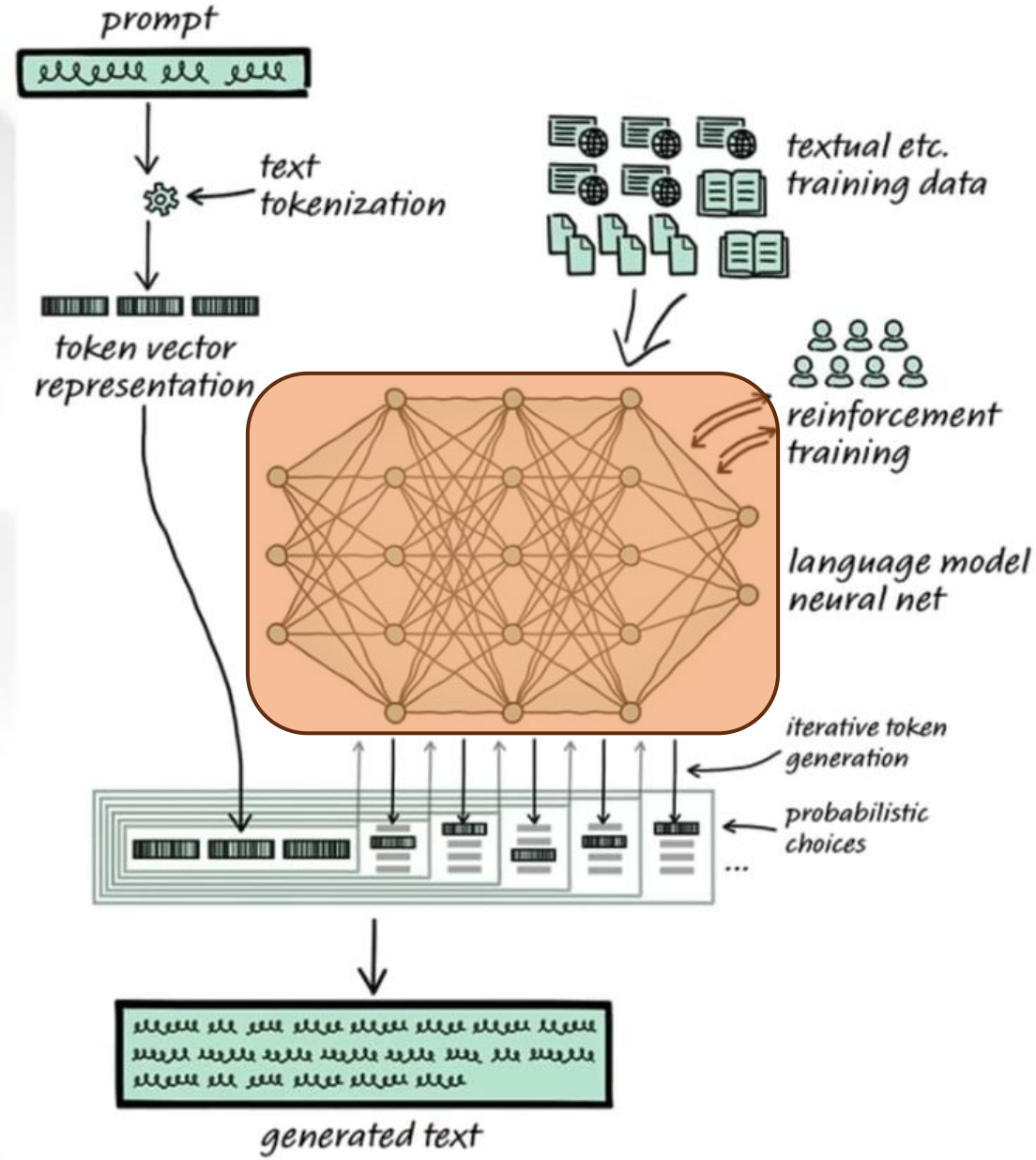
- ChatGPT
  - **G**enerative
  - **P**re-Trained
  - **T**ransformer



Wolfram: “What Is ChatGPT Doing … and Why Does It Work?”

Generative Models

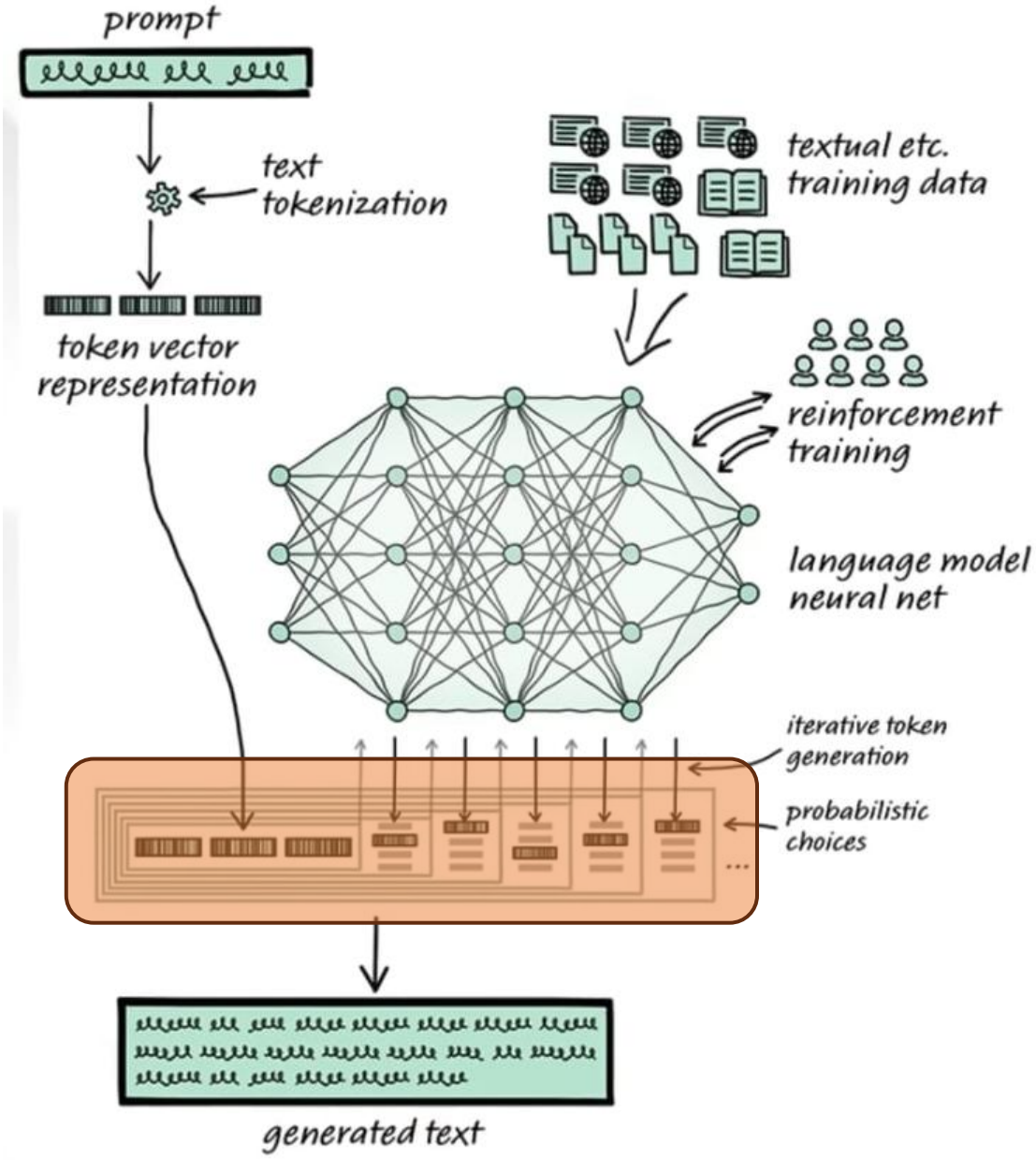
- ChatGPT
  - **G**enerative
  - **P**re-Trained
  - **T**ransformer



Wolfram: “What Is ChatGPT Doing … and Why Does It Work?”

Generative Models

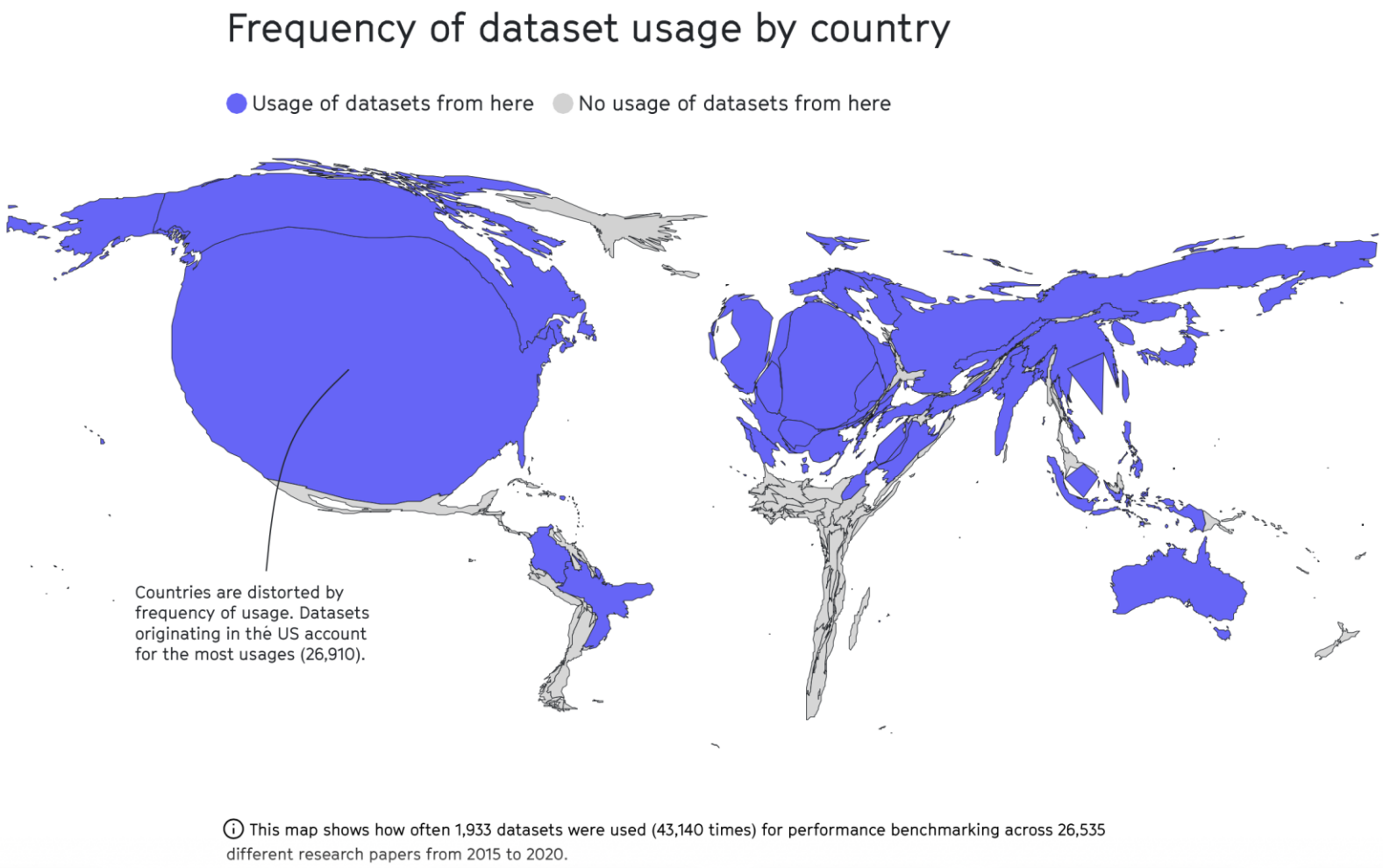
- ChatGPT
  - **G**enerative
  - **P**re-Trained
  - **T**ransformer



Wolfram: “What Is ChatGPT Doing … and Why Does It Work?”

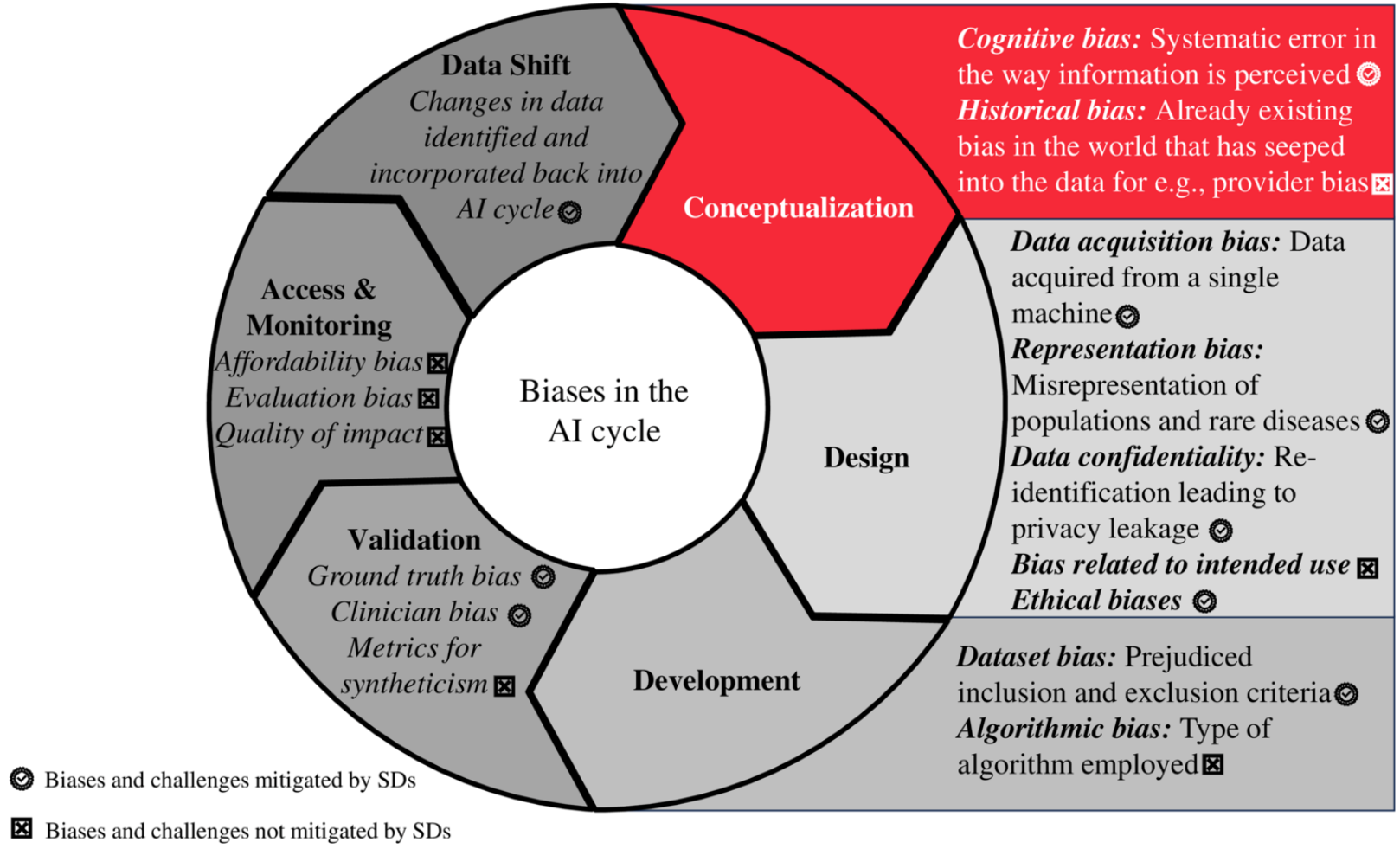
Generative Models

Bias



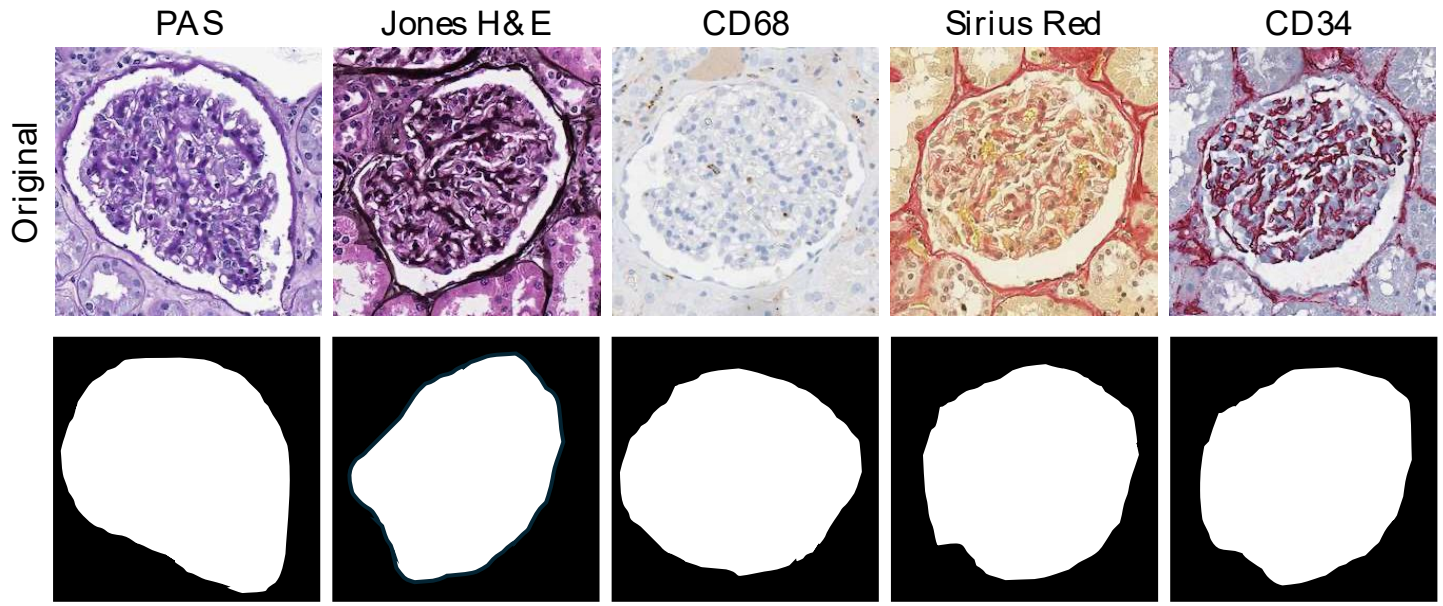
B. Koch, et al., Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research, 2021.

Gen AI in Research



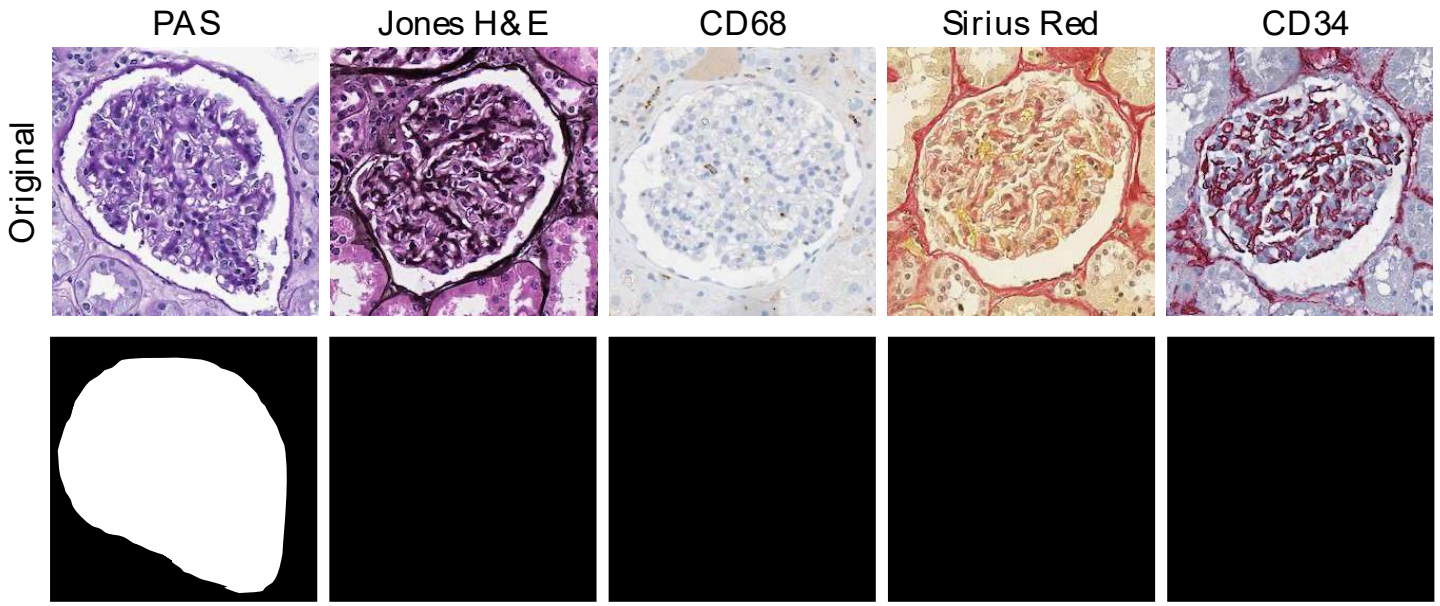
Umer and Adnan, Generative artificial intelligence: synthetic datasets in dentistry, BDJ Open, 2024

Medical Imaging



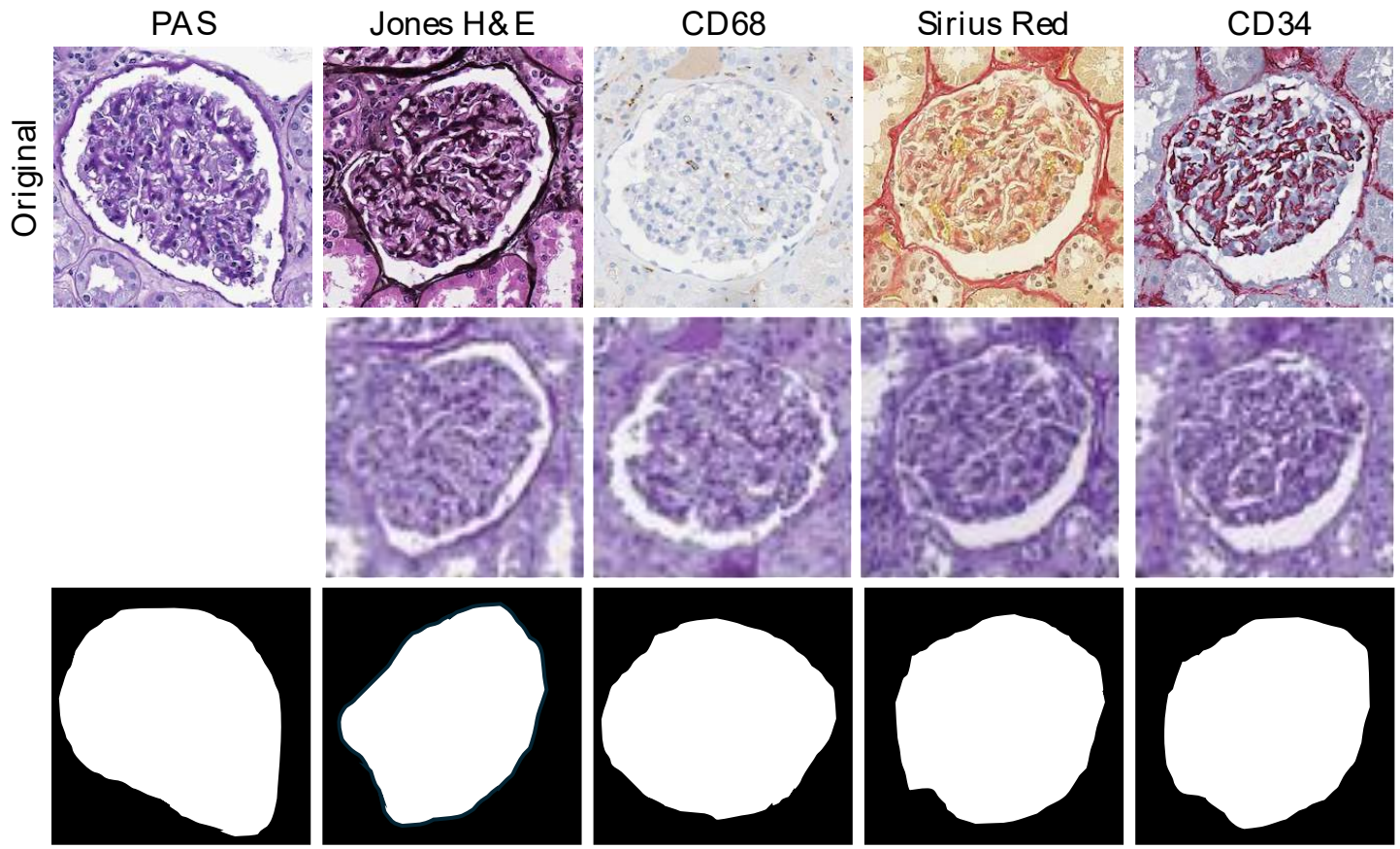
Medical Imaging

- Ground truth bias

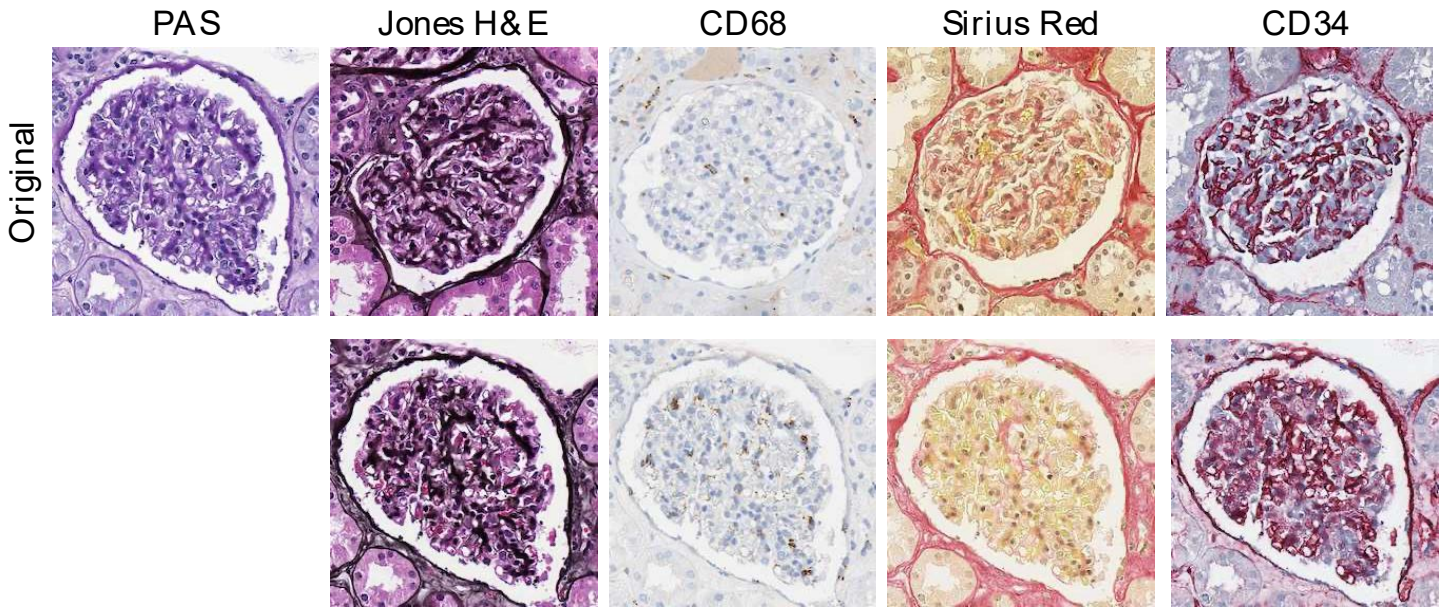


Medical Imaging

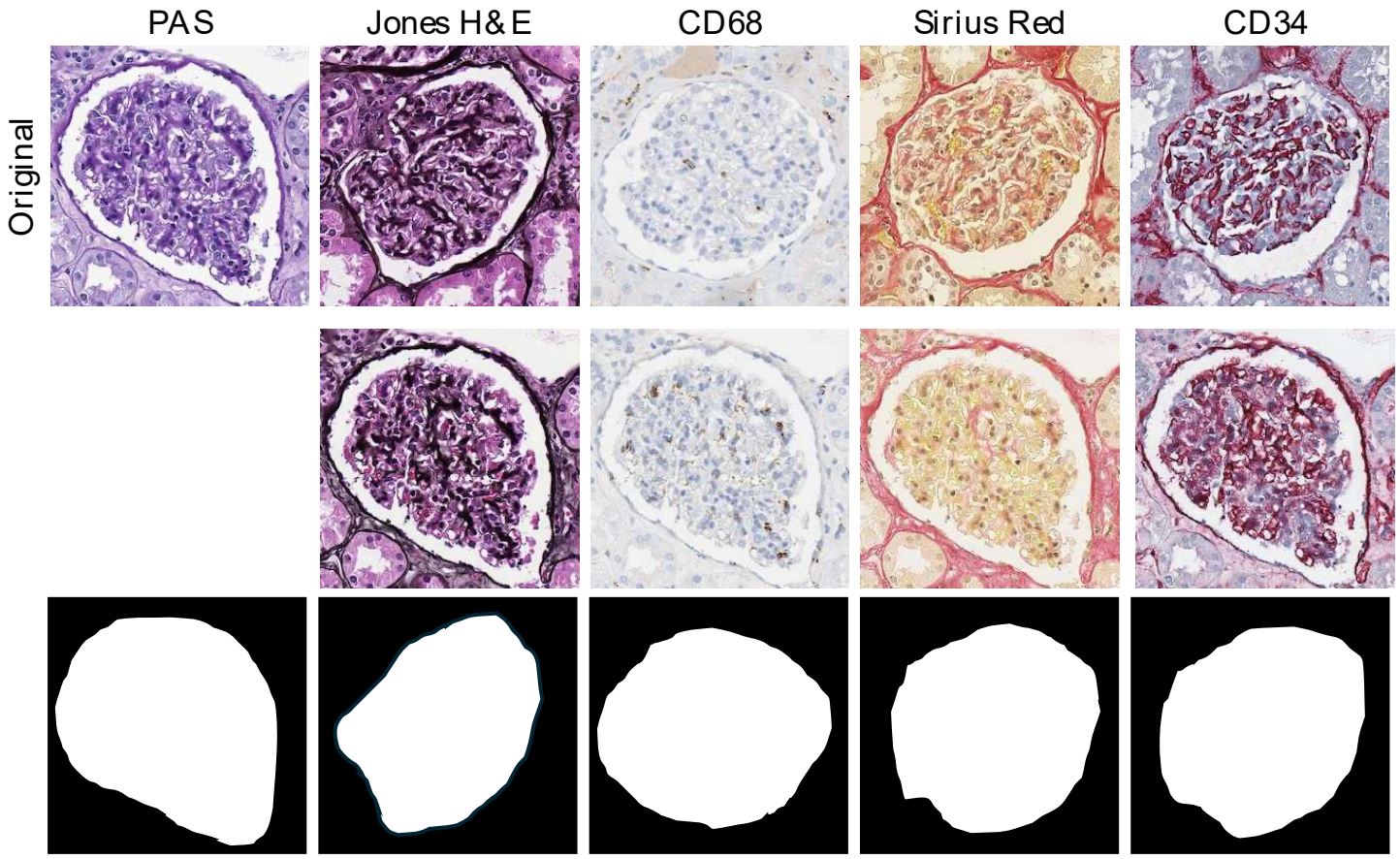
- Ground truth bias



Medical Imaging



Medical Imaging



AI-generated and synthetic data

Prof Thomas Lampert

University of Strasbourg

How Generative AI and Synthetic Data Reshape Trust and Research

Ella Hafermalz
Associate Professor – Work and Technology
Vrije Universiteit Amsterdam

VU VRIJE UNIVERSITEIT AMSTERDAM

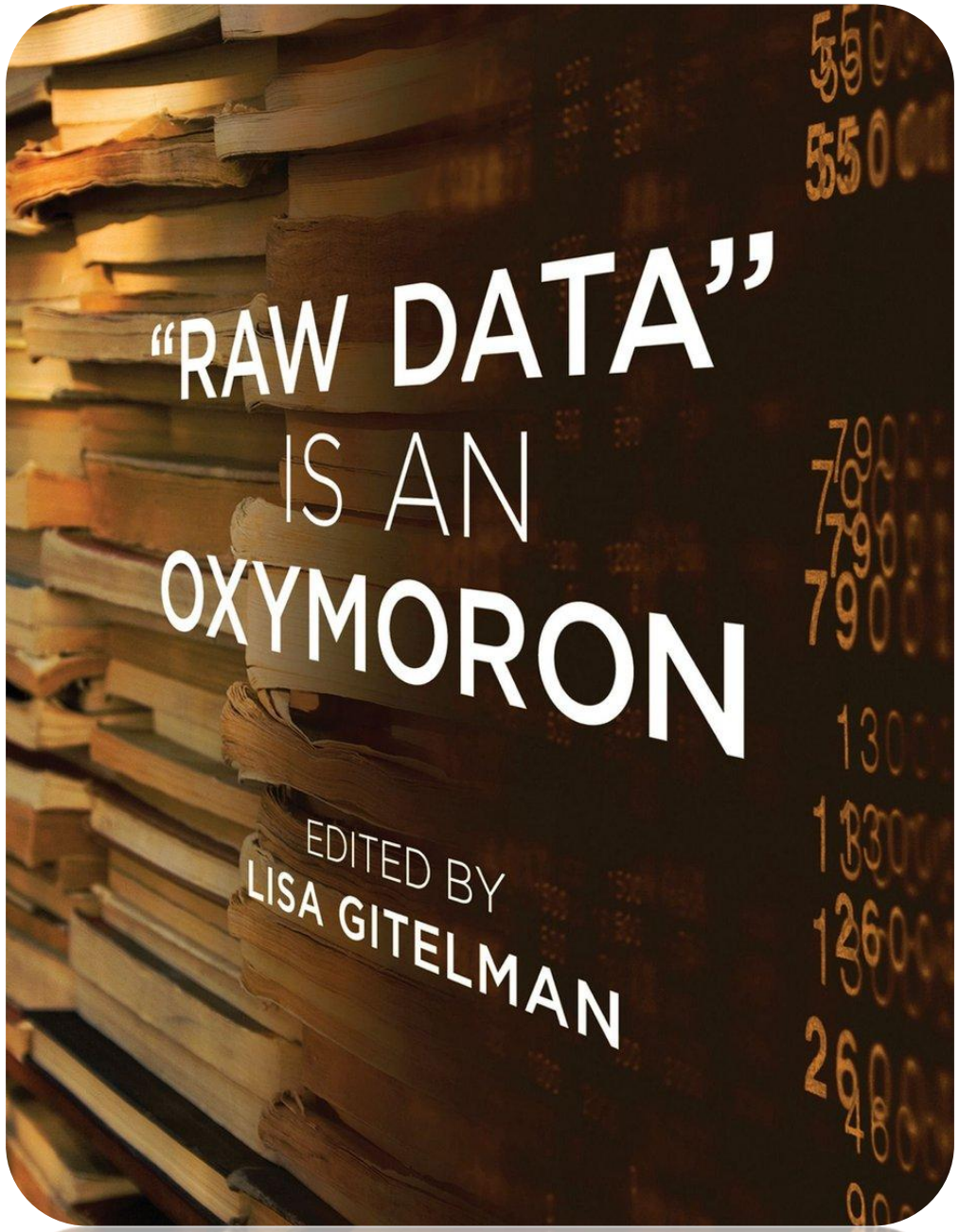
School of Business and Economics

Center for Digital Innovation

Data has always been constructed

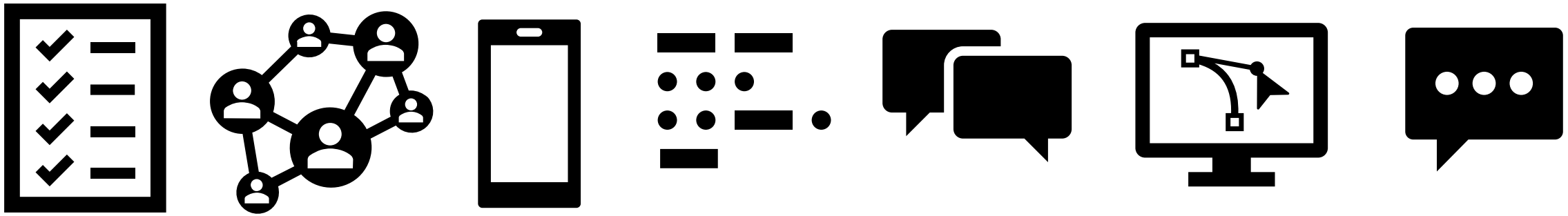
Reflecting on the "construction" of data rather than the "collection" of data - i.e. the *mushroom picking* approach to qualitative research.

Alvesson, M., & Sandberg, J. (2018). Metaphorizing the research process. *The Sage handbook of qualitative business and management research methods: Methods and challenges*, *2*, 486-505.


"RAW DATA"
IS AN
OXYMORON
EDITED BY
LISA GITELMAN

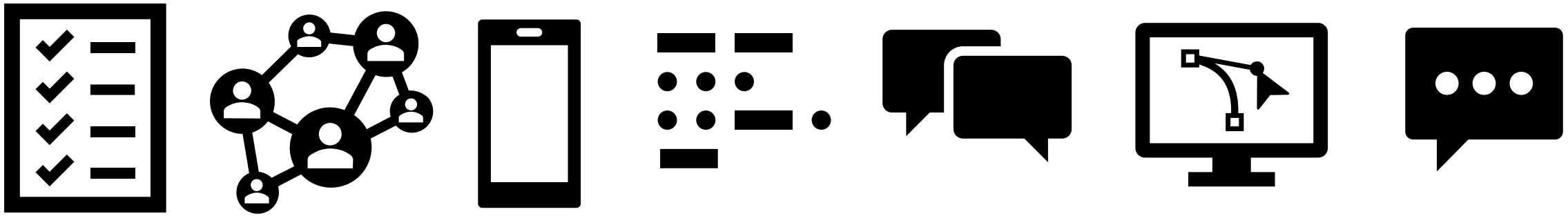
What is the synthetic data ‘threshhold’?

IS MY DATA SYNTHETIC? AT WHAT POINT?



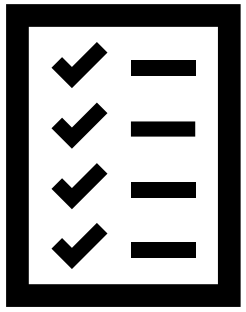
1 2 3 4 5 6 7

IS MY DATA SYNTHETIC? AT WHAT POINT?



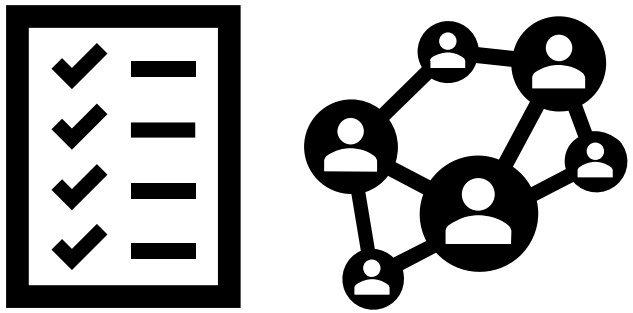
1 2 3 4 5 6 7

IS MY DATA SYNTHETIC? AT WHAT POINT?



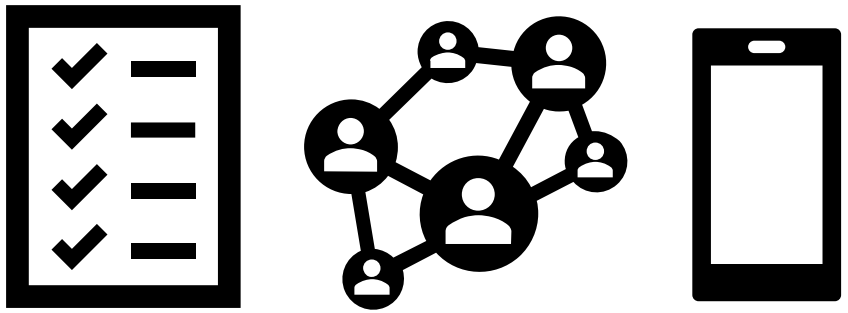
1

IS MY DATA SYNTHETIC? AT WHAT POINT?



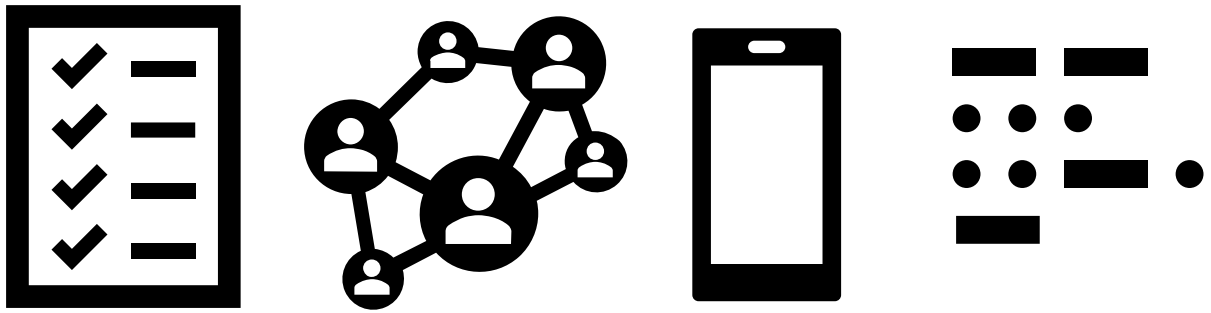
1 2

IS MY DATA SYNTHETIC? AT WHAT POINT?



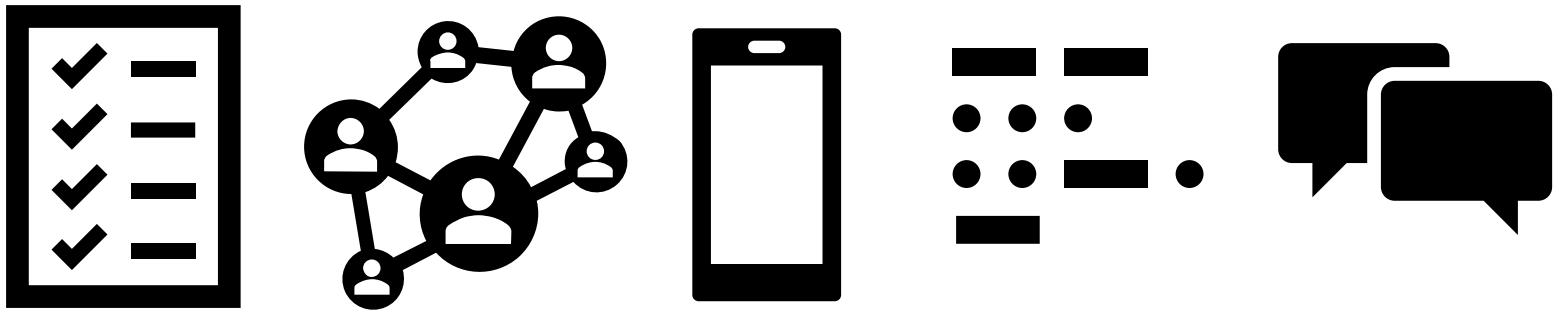
1 2 3

IS MY DATA SYNTHETIC? AT WHAT POINT?



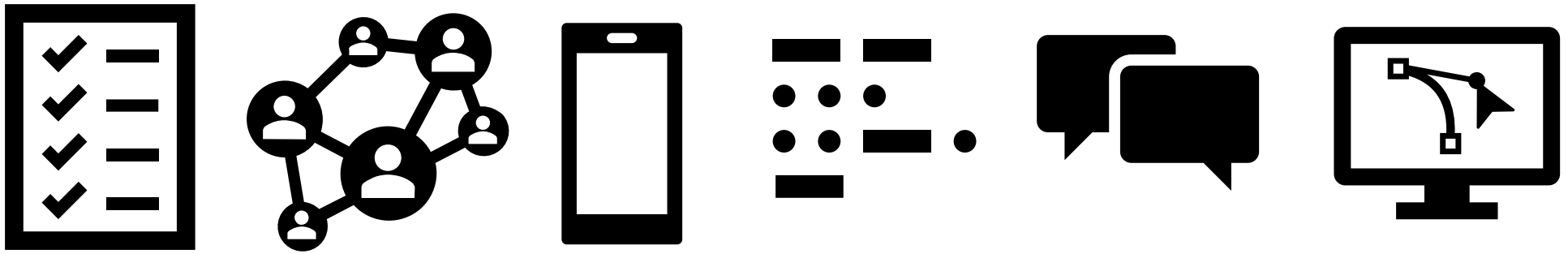
1 2 3 4

IS MY DATA SYNTHETIC? AT WHAT POINT?



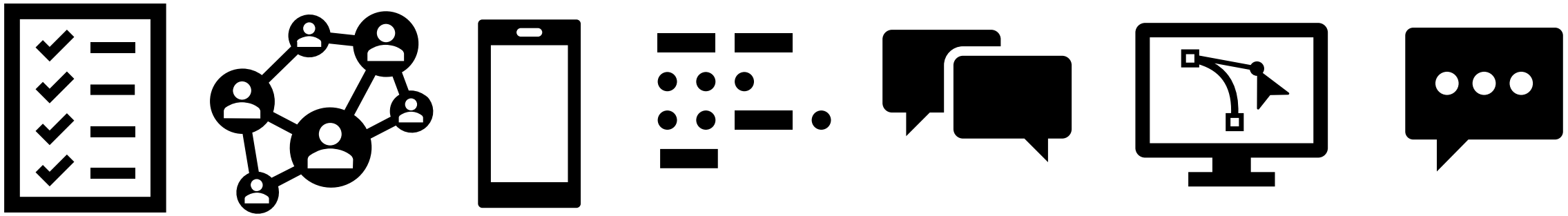
1 2 3 4 5

IS MY DATA SYNTHETIC? AT WHAT POINT?



1 2 3 4 5 6

IS MY DATA SYNTHETIC? AT WHAT POINT?



1 2 3 4 5 6 7

What’s new here?

Business Horizons
Volume 67, Issue 5, September–October 2024, Pages 511-523

Managing a ChatGPT-empowered workforce: Understanding its affordances and side effects

Jana Retkowsky, Ella Hafermalz, Marleen Huysman

Show more

+ Add to Mendeley Share Cite

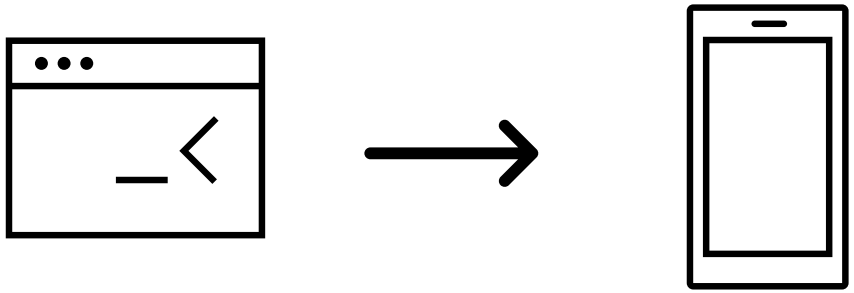
https://doi.org/10.1016/j.bushor.2024.04.009

Get rights and content

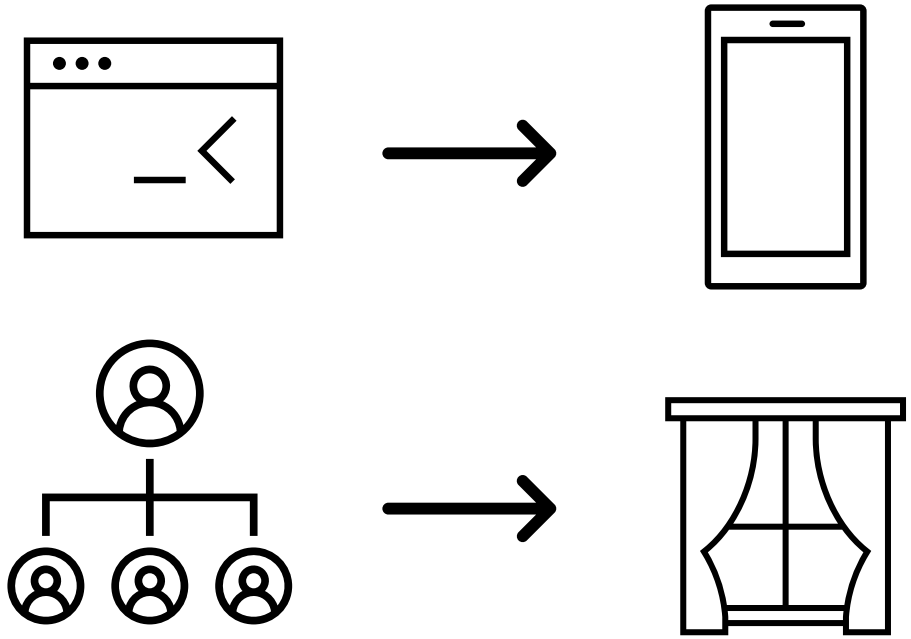
Under a Creative Commons license

Open access

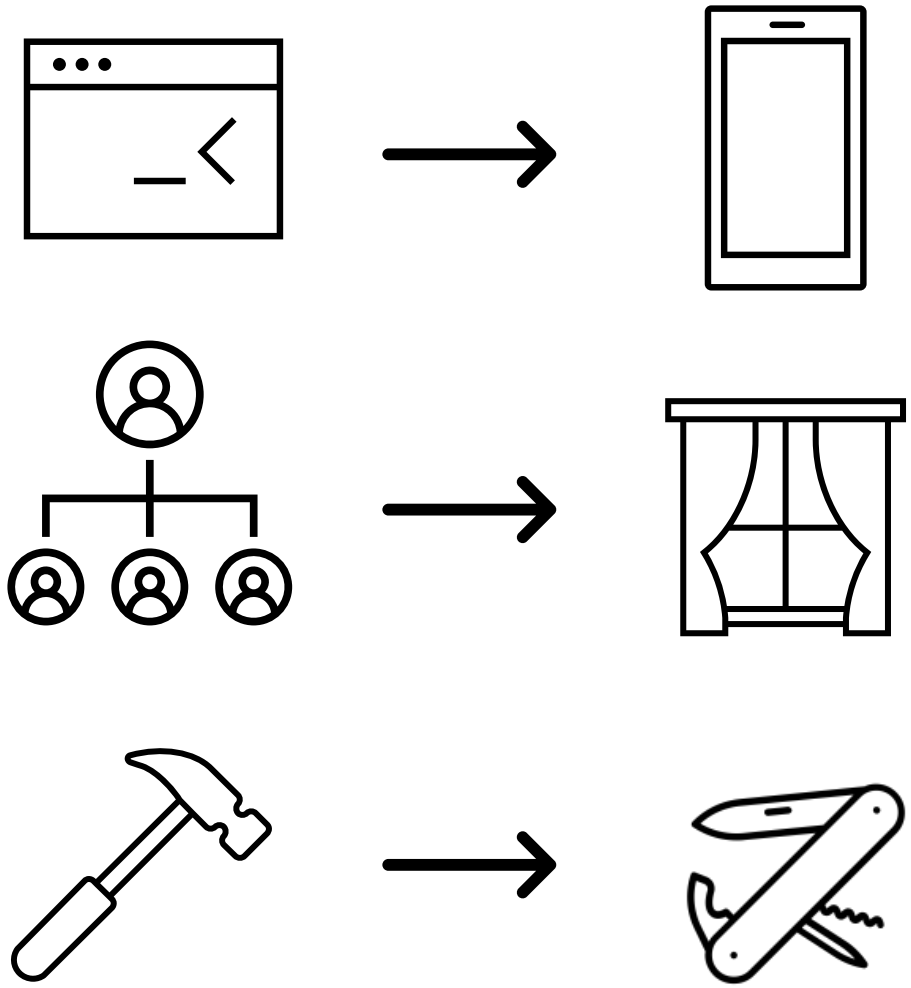
TRADITIONAL AI VS GENERATIVE AI



TRADITIONAL AI VS GENERATIVE AI



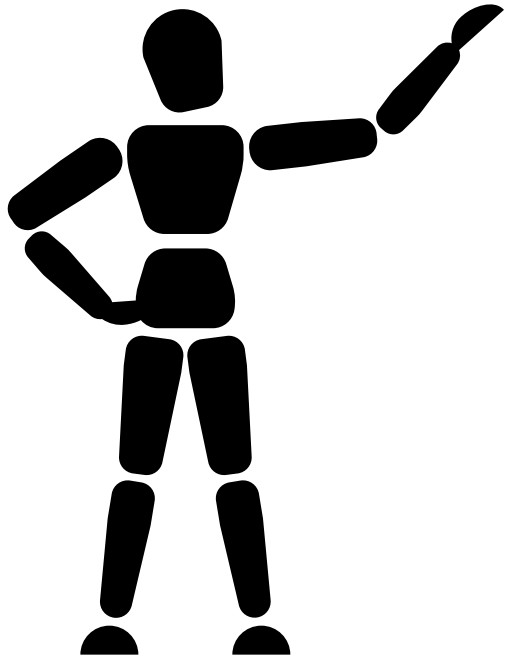
TRADITIONAL AI VS GENERATIVE AI



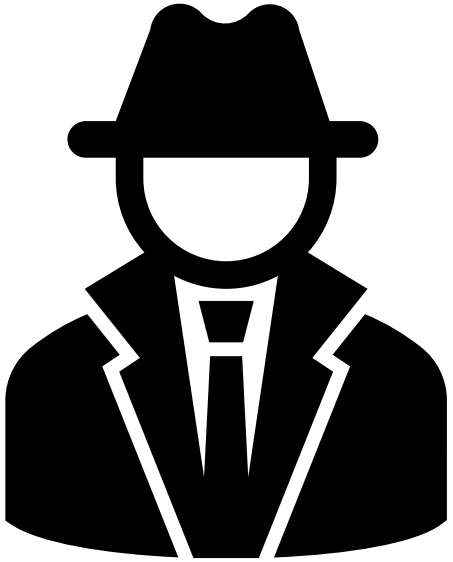
3 different forms of synthetic data in Qualitative Research

1. Legitimate and declared → e.g. illustrative vignettes
2. Fraudulent and covert → e.g. fabricated interviews, deep fakes
3. Intertwinement → merging of human and AI outputs

1. LEGITIMATE AND DECLARED

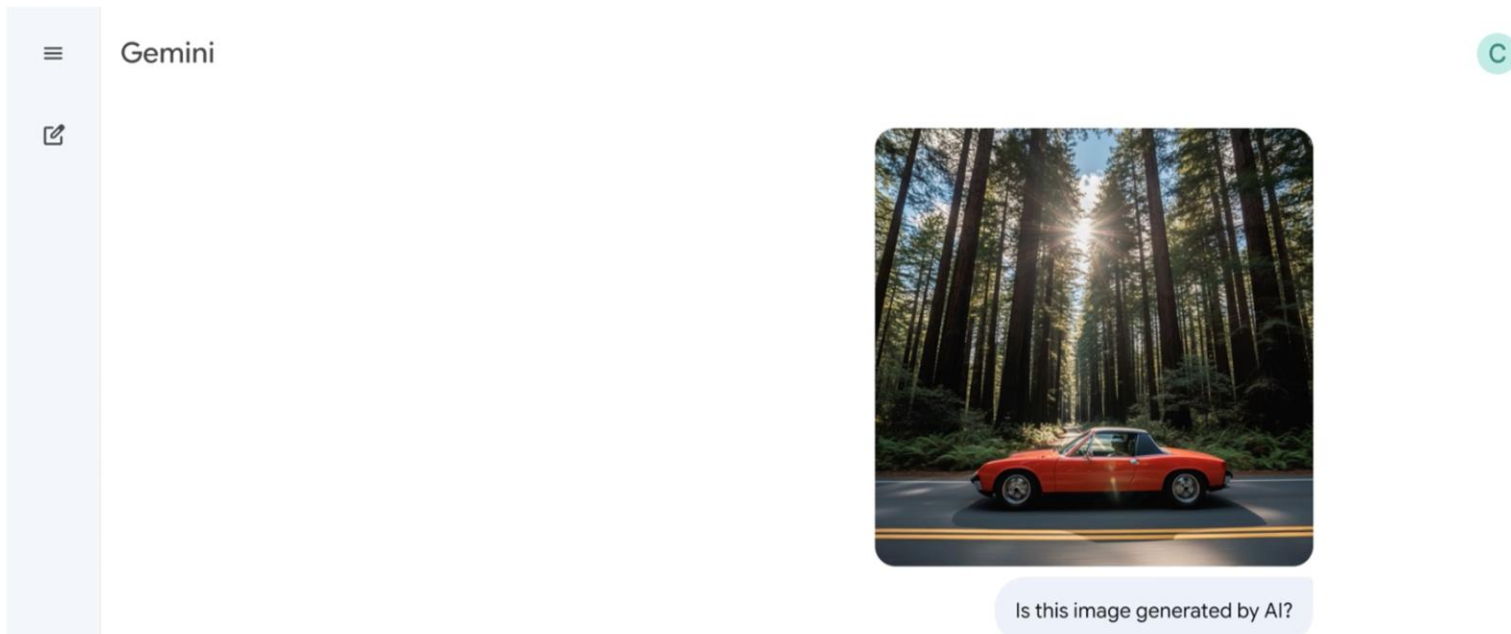


2. FRAUDULANT AND COVERT



We believe it's critical to know when an image is AI-generated. This is why all media generated by Google's tools are embedded with our imperceptible SynthID digital watermark.

Today, we are putting a powerful verification tool directly in consumers' hands: you can now upload an image into the Gemini app and simply ask if it was generated by Google AI, thanks to SynthID technology. We are starting with English language prompts for images, and will expand to more languages, audio and video soon.



how to remove google banana pro watermark

AI-overzicht

Removing the Gemini Nano Banana Pro watermark (part of Google's SynthID system) can be done using image editing tools, specialized AI removers, or by re-prompting. Effective methods include using the Magic Eraser in Google Photos, applying the open-source GeminiWatermarkTool, using free online AI removers like Pixelbin, or prompting the AI to "extend the background" to move the watermark out of frame.

Top Methods to Remove the Watermark

- **Google Photos Magic Eraser:** Open the image in the Google Photos app, select "Edit" -> "Tools" -> "Magic Eraser," and tap the watermark to remove it.
- **GeminiWatermarkTool (CLI/Drag & Drop):** Download the `GeminiWatermarkTool-Windows-x64.exe` from GitHub, then drag and drop your image onto the executable to remove the watermark.
- **Free Online AI Removers:** Use tools like NoteGPT or Pixelbin to upload the photo and automatically erase the SynthID digital watermark.
- **Prompt Engineering (Re-generation):** When generating, ask the AI to "extend the background of the image whilst keeping the current image the same," which often places the watermark in a crop-able area. Another prompt is: "Recreate the picture with 100% similarity... but without including the icon on the bottom right-corner".
- **Alternative AI Editors:** Upload the image to ChatGPT or other AI editors and ask them to remove the bottom-right watermark.

Gemini Nano Banana / Pro watermark maintenance tool

13 feb 2026 — Always back up your original images before processing. The author assumes no responsibility for any dat...

GitHub

This is how I remove the gemini watermark from images ...

21 okt 2025 — sankalp_pateriya. • 4mo ago. It's not that deep, just use Magic Eraser in Google...

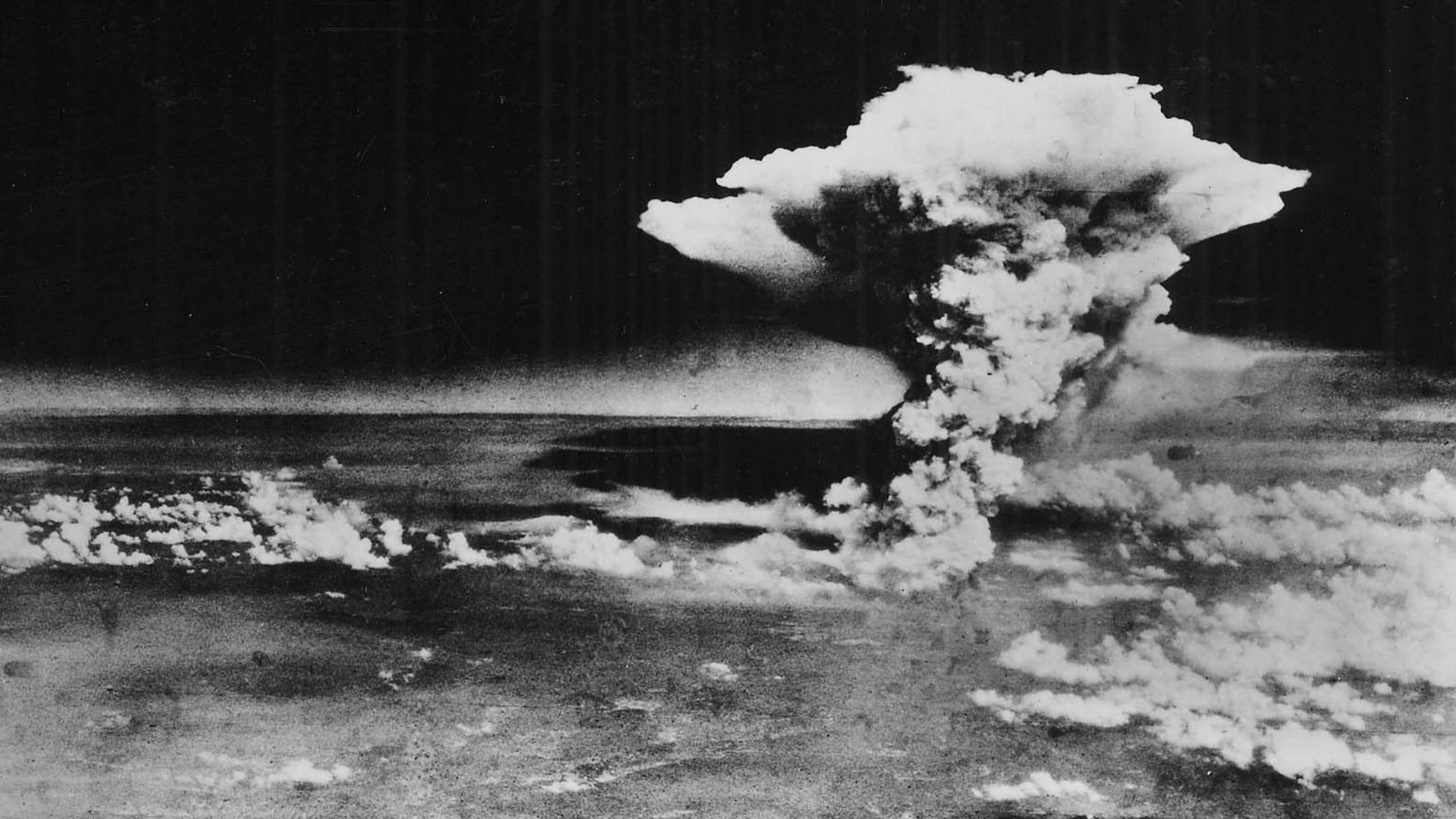
Reddit

Easy way to remove the Gemini Watermark from Your Nano Banana ...

3 dec 2025 — Comments Section * JohnyChingas. • 3mo ago. I use Google Photo...

Reddit

Alles tonen



3. INTERTWINEMENT

4.2. Reaching intertwinement: ChatGPT as a number one knowledge assistant

We heard that ChatGPT increasingly became a part of our participants' daily work. We describe this stage, where ChatGPT becomes part of normal flow of work, as *intertwinement*. The idea of intertwinement is inspired by literature that looks at how people relate to technologies that they are familiar with and reliant upon. Reading glasses and a blind person's cane are classic examples where it is not clear where the human ends and the technology begins; they operate as one closely linked, intertwined entity (Merleau-Ponty et al., 2013). Likewise, over time ChatGPT became our participants' go-to place for assistance. Some noted they use ChatGPT every day and always have a browser window open, ready to chat. It became an indispensable routine part of their work, and participants referred to it as "my personal assistant" or "it's a part of my workflow." In this stage, the tool becomes such an integral part of the work that it is no longer clear to the user what aspects of their work are which are human and which are AI generated.

(Retkowsky, Hafermalz, and Huysman, 2024)

Code of Practice on marking and labelling of AI-generated content

Page Contents

This code of practice aims to support compliance with the AI Act transparency obligations related to marking and labelling of AI-generated content.

Share

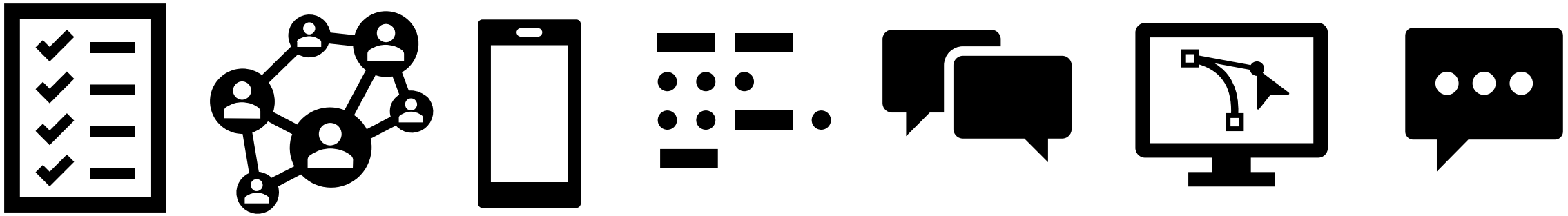
Marking and labelling of AI-generated content

The obligations under Article 50 of the AI Act (transparency obligations for providers and deployers of generative AI systems) address risks of deception and manipulation, fostering the integrity of the information ecosystem. These transparency obligations pertain to marking and detection of AI generated content and labeling of deep fakes and certain AI generated publications. They complement other rules like those for high-risk AI systems or general-purpose AI models.

To assist with compliance with these transparency obligations, the AI Office has kick started the process of drawing up a code of practice on transparency of AI-generated content. The code will be drafted by

https://digital-strategy.ec.europa.eu/en/policies/code-practice-ai-generated-content

WHAT TO DECLARE, AND HOW?



1 2 3 4 5 6 7

Conclusion

Synthetic data and the visibility of knowledge production

THE OPPORTUNITY: IMAGINATIVE RESEARCH

Synthetic and AI-generated data open new possibilities for qualitative and creative inquiry.

Create personas that don't exist

'Interview' people who aren't here

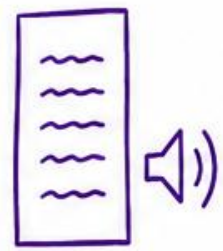
Explore scenarios, futures, alternatives

Blend art and science

More ways to ask "what if?"

THE RISK: CREDIBILITY AT RISK

The same tools that create new possibilities also make it harder than ever to prove what happened.



Transcripts, audio, even videos can be deep faked

Low cost.
Low skill.
Hard to detect.

Unmanageable burden on gatekeepers of research integrity

Risk of cynicism: everyone suspects everything, so nothing is trusted

WHAT WE NEED: A NEW INFRASTRUCTURE OF TRUST

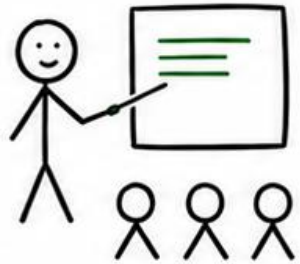
METHODS & PROCEDURES

Stronger study design, documentation and audit trails.
Make the process visible, not just the output.

TECHNOLOGIES

Secure capture, provenance tracking, watermarking, authenticity tools and standards.

EDUCATION & CULTURE



Train researchers, reviewers and students to ask better questions, read evidence critically and use tools responsibly.

PROPORTIONAL & CONTEXTUAL

Different methods, different risks, different standards.
Right level of evidence for the claim being made.

TRANSPARENCY & DEMONSTRATION

Move from "trust me" to "here is how this was produced, validated and why it is credible enough for this purpose".

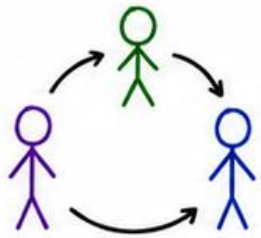
LEARN FROM THE PAST

Drawing on decades of work in qualitative and critical data studies, research integrity, and open science.

BUILD NEW PRACTICES

Combine methods, procedures, technologies and education into integrated, practical workflows.

WORK TOGETHER



THE GOAL

Research that is creative and bold, and knowledge that is credible, defensible and worth trusting.

Ella Hafermalz

Questions and comments welcome

VU VRIJE UNIVERSITEIT AMSTERDAM

School of Business and Economics

Center for Digital Innovation

Q&A Session

Flora Kopelou
European Data Portal, Publications Office of the EU

Dr Thomas Lampert
Professor of Computer Science, University of Strasbourg

Dr Ella Hafermalz
Associate Professor, Vrije Universiteit Amsterdam

data europa academy

Stay up-to-date on our 2026 activities!

data.europa academy

Continue the discussion after the webinar!

Open data, academia, and ethics: responsible use of AI-generated and synthetic data in research

Submitted by Hannah KROKER on Thu, 21/05/2026 - 16:46

Topic: Academy webinars

Have you joined our webinar on the responsible use of AI-generated and synthetic data?

The use of synthetic data offers up countless new possibilities in research but requires careful consideration of the potential risks involved. In our webinar, our two guests from academia examined how emerging technologies can be applied responsibly to produce valuable research.

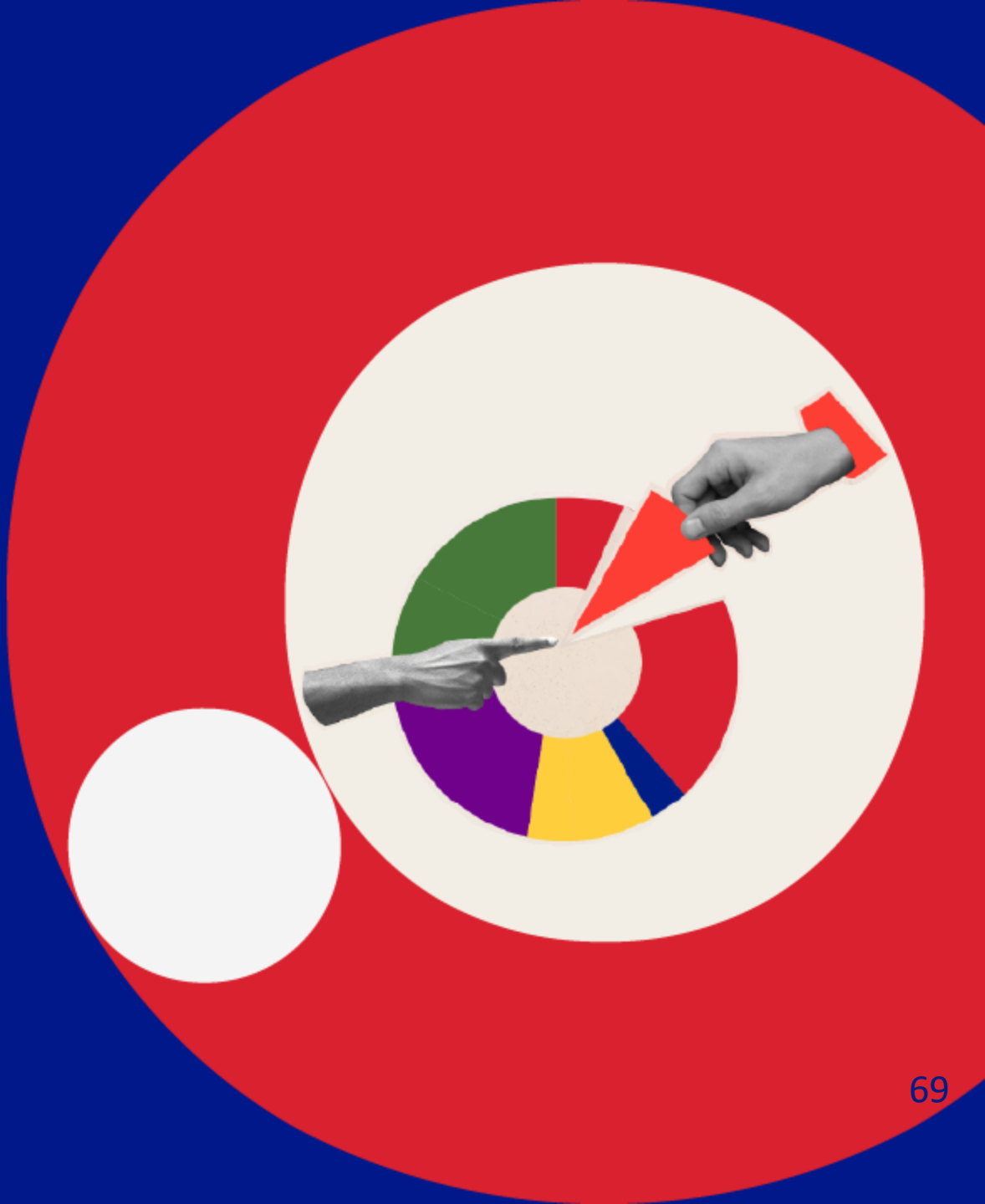
Have you previously worked with synthetic data? In your opinion, what should researchers keep in mind when using such techniques?

Let us know in the comments below!

Login using your EU login account and join our Data reusers group to share your thoughts in the comment section down below.

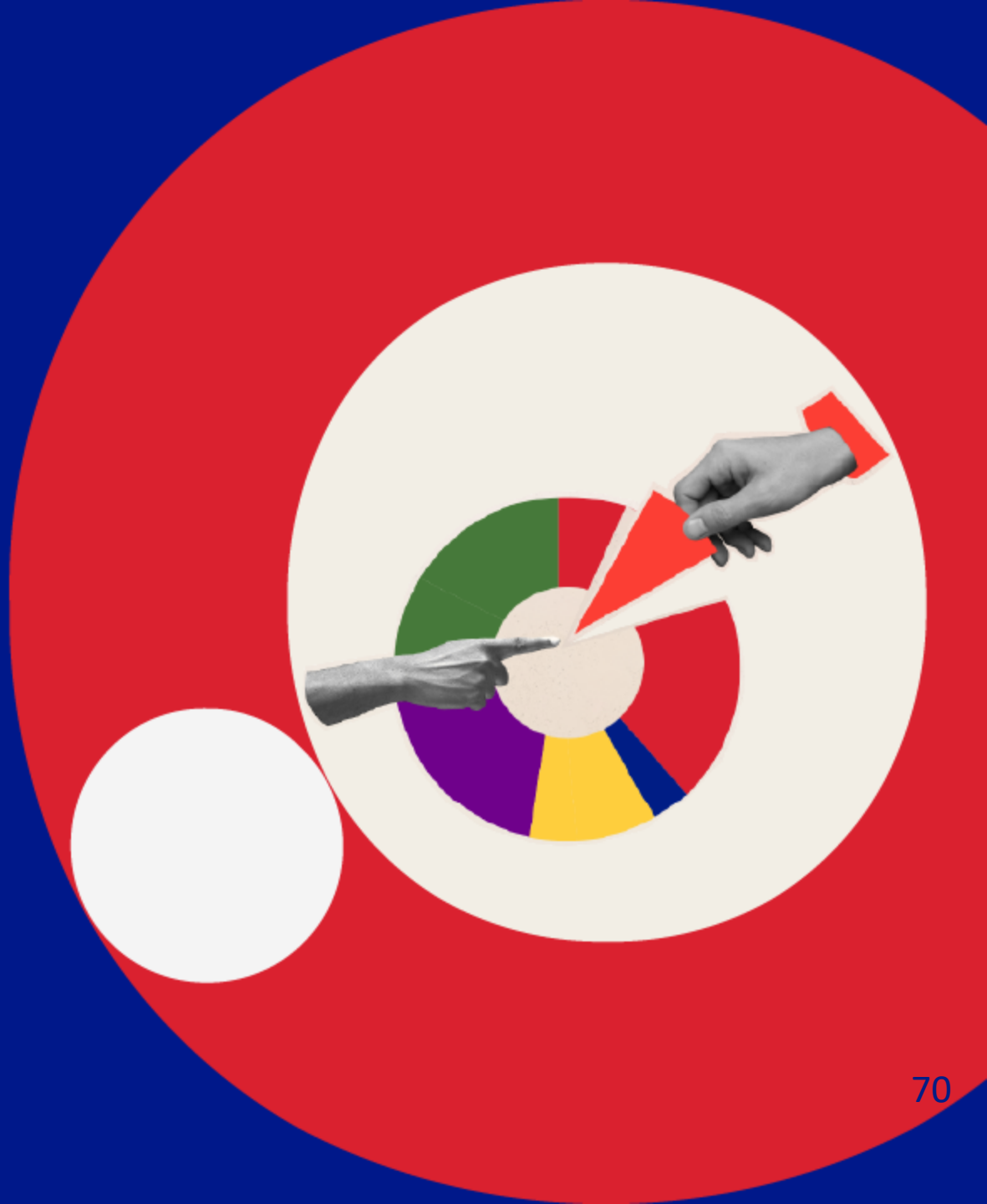
data europa academy

Your opinion is important to us



data europa academy

Thank you!



data europa academy