Tutorial 1: Representing Data

Week 2

January 2026

1 Research Question & Hypothesis Building

Imagine the current state of generative AI tools (e.g., LLMs, image and video generation models, ChatGPT) and how they are shaping the world and our lives in domains such as education, work, creativity, misinformation, mental health, or inequality.

Identify one specific domain and propose an interesting research question that investigates benefits or harms of generative AI in that context.

1. Suppose you have sufficient funding to conduct a traditional experiment. Redefine your research question, formulate a clear hypothesis, and describe your experimental design. How would you divide participants into groups, and what outcome(s) would you measure?

2. Suppose you do not have sufficient funding and must rely on big data. What data sources could you use? Redefine your research question and hypothesis accordingly. What are the main limitations and ethical risks of this approach?

2 Data sources

What sources of data have you encountered so far in your studies and/or in your life? What types of data were they? What were they used for?

3 Data about you

1. Data is being collected all the time from our behaviours online. What data do you think exists about you? Where is it stored? Who has access to it? What is it used for?

2. A data scientist put together all the personally identifiable data about you that's on the internet, what might they learn about you? What would they miss or get wrong about you?