The Challenge & Paradox of Data Quality

Irina Shklovski

Professor of Communication & Computing

Department of Computer Science - UCPH

TEMA GENUS - Linköping University

UNIVERSITY OF COPENHAGEN

Ostrageologias by Daniel Canogar

When we talk about data and AI what are we talking about?

30 August 2021

How bird's species made new Professor Serge Belongie world-famous within Computer Vision

ARTIFICIAL INTELLIGENCE | TECHNOLOGY

NEW PROFESSOR New Professor at the Department of Computer Science and coming Director of Denmark's new Pioneer Center for artificial intelligence, Serge Belongie, allows himself and his students to think big. He recently moved from New York to Copenhagen to take Danish AI research to new heights.

Evolution of debates: privacy then data then ethics…

- Began initially with a focus on privacy concerns
  - More precise and granular data will lead to better knowledge and commercial gain
  - Creation of new types of data always creates new types of visibility
  - Who gets to control data collection?

- Shifted to discussions about purposes of data use
  - Concerns with data minimization and stated purpose of use

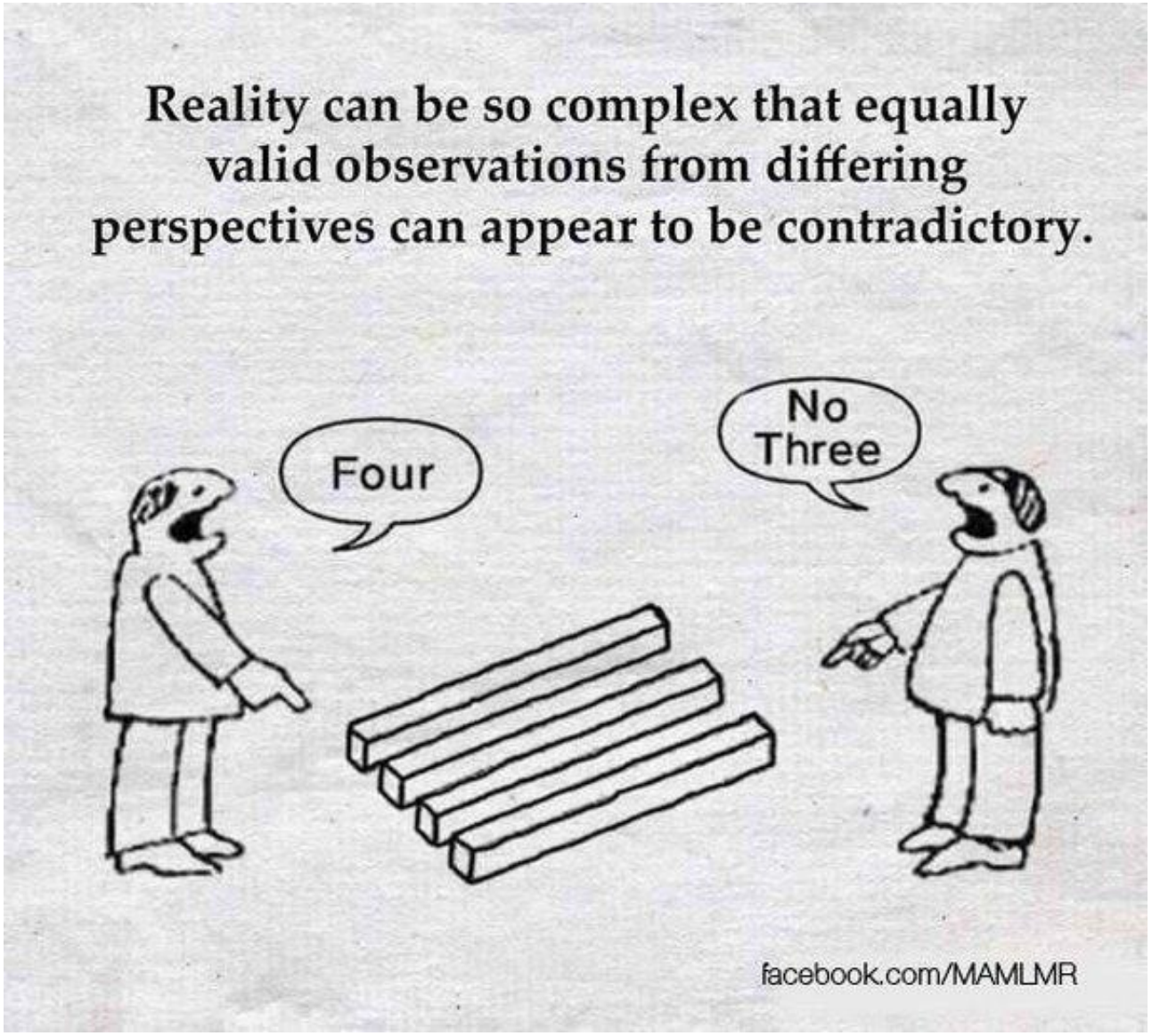
- Considered uneven and problematic outcomes of data-driven systems
  - Shift from concerns about data and privacy to ethical principles

What is/are data?

ChatGPT Image

Microsoft CoPilot

All data are socially constructed



All data are socially constructed/biased (there are no exceptions)

- How we conceive of data, how we measure them, what we intend to do with them actively frames the nature of data

- Misconceptions of data – there is no n=all:
  - All data are local/situated in contexts
  - Data are framed and shaped by the conditions of their production

- Representative datasets are a nice but impossible ideal (for the most part)
  - It is unknown what a truly representative dataset might look like in many domains
  - All data are missing something or someone – usually systematically and deliberately
  - It is difficult to know when/how the missingness might matter

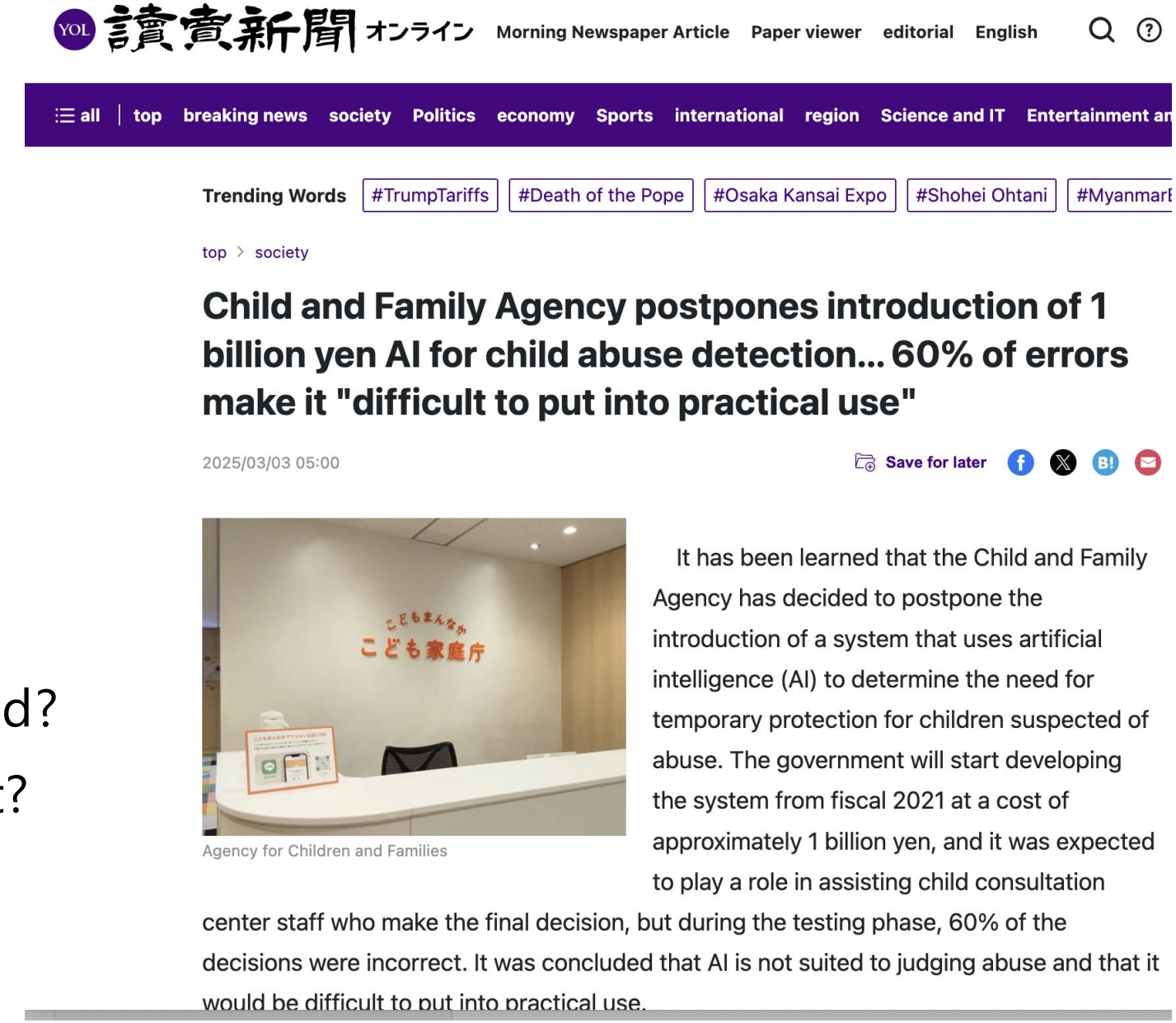
The nature of data

Mostly in ML and NLP

- A lot of attention to models
- Little attention to **data**

Is it important that

- All data are socially constructed?
- All data are local & contingent?
- Most data are of poor quality?



YOL 讀賣新聞 オンライン Morning Newspaper Article Paper viewer editorial English

all | top breaking news society Politics economy Sports international region Science and IT Entertainment ar

Trending Words #TrumpTariffs #Death of the Pope #Osaka Kansai Expo #Shohei Ohtani #MyanmarE

top > society

Child and Family Agency postpones introduction of 1 billion yen AI for child abuse detection... 60% of errors make it "difficult to put into practical use"

2025/03/03 05:00 Save for later

Agency for Children and Families

It has been learned that the Child and Family Agency has decided to postpone the introduction of a system that uses artificial intelligence (AI) to determine the need for temporary protection for children suspected of abuse. The government will start developing the system from fiscal 2021 at a cost of approximately 1 billion yen, and it was expected to play a role in assisting child consultation center staff who make the final decision, but during the testing phase, 60% of the decisions were incorrect. It was concluded that AI is not suited to judging abuse and that it would be difficult to put into practical use.

What is “data quality”?

Copilot

Data quality refers to the condition or fitness of data to serve its intended purpose in a given context. High-quality data is essential for making accurate decisions, conducting reliable analyses, and ensuring smooth operations in any data-driven environment.

Copilot

Quality data is data that is **fit for its intended use** in operations, decision-making, and planning. It meets specific standards and characteristics that make it reliable, accurate, and useful.

The problem of data quality

The Five Facets of Data Quality Assessment

Sedir Mohammed[1], Lisa Ehrlinger[1], Hazar Harmouch[2], Felix Naumann[1], Divesh Srivastava[3]
[1]Hasso Plattner Institute, University of Potsdam, Germany
[2]University of Amsterdam, Netherlands
[3]AT&T Chief Data Office, USA
sedir.mohammed@hpi.de, lisa.ehrlinger@hpi.de
h.harmouch@uva.nl, felix.naumann@hpi.de, divesh@research.att.com

ABSTRACT

Data-oriented applications, their users, and even the law require data of high quality. Research has divided the rather vague notion of data quality into various dimensions, such as accuracy, consistency, and reputation. To achieve the goal of high data quality, many tools and techniques exist to clean and otherwise improve data. Yet, systematic research on actually assessing data quality in its dimensions is largely absent, and with it, the ability to gauge the success of any data cleaning effort.

We propose five facets as ingredients to assess data quality: *data*, *source*, *system*, *task*, and *human*. Tapping each facet for data quality assessment poses its own challenges. We show how overcoming these challenges helps data quality assessment for those data quality dimensions mentioned in Europe's AI Act. Our work concludes with a proposal for a comprehensive data quality assessment framework.

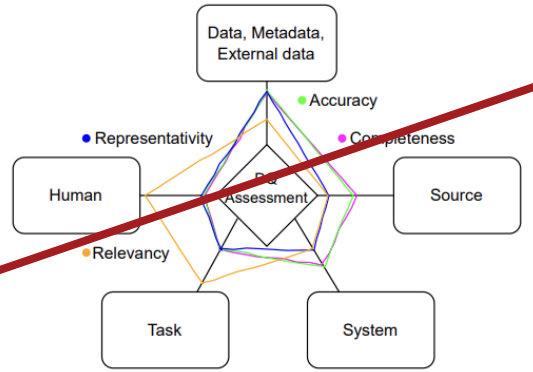


Figure 1: The five *facets* of DQ assessment and exemplary characteristics for DQ dimensions.

the *Health Insurance Portability and Accountability Act* (HIPAA), which focuses on privacy but promotes DQ dimensions, such as accuracy and

systematic research on actually assessing data quality in its dimensions is largely absent, and with it, the ability to gauge the success of any data cleaning effort.

UNIVERSITY OF COPENHAGEN

Does data quality matter?

Tech

New Japanese AI System Tracks Down Lost Items

THE SANKEI SHIMBUN APRIL 23, 2025

> AI is helping travelers find lost items faster, even with vague information, boosting return rates from under 10% to 30% across Japan's transit systems.

Technology »

AI Tool Aims to Help Conserve Japan's Cherry Trees by Assessing Many Photos

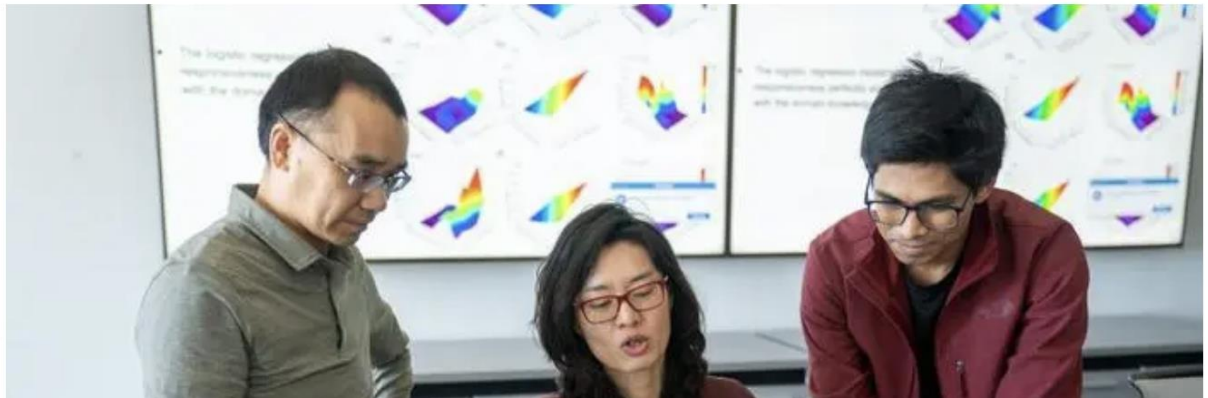
Post Share

AFP-Jiji

12:45 JST, April 23, 2025

Does data quality matter?

Virginia Tech Study Reveals Machine Learning Models Struggle to Identify Critical Health Declines

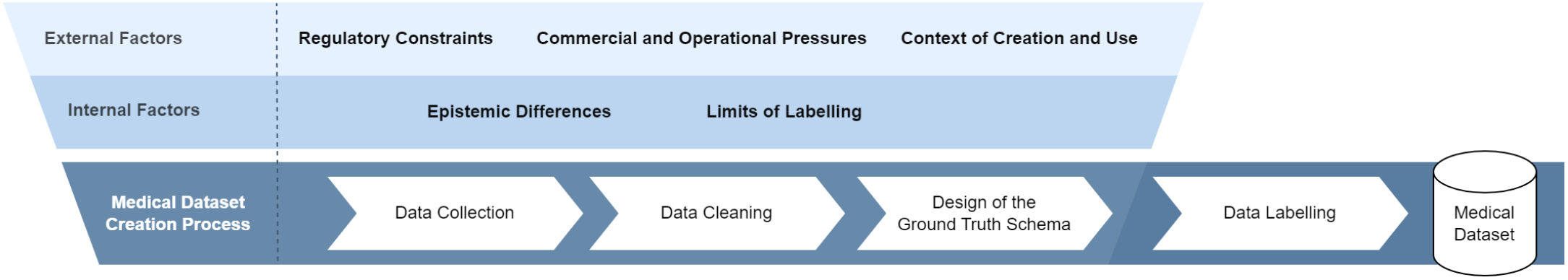


Interrogating how data are made

DCODE NETWORK
RETHINK DESIGN – CHANGE OUR FUTURE

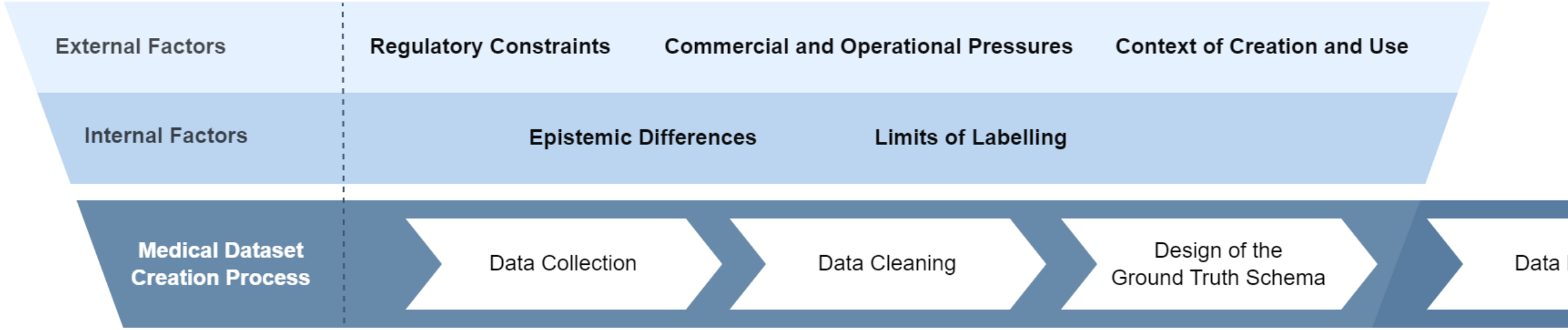
Case: creation, implementation and use of medical AI (Natalia Avlona)

→ How do experts in the health tech sector create medical datasets for the design of algorithmic systems?



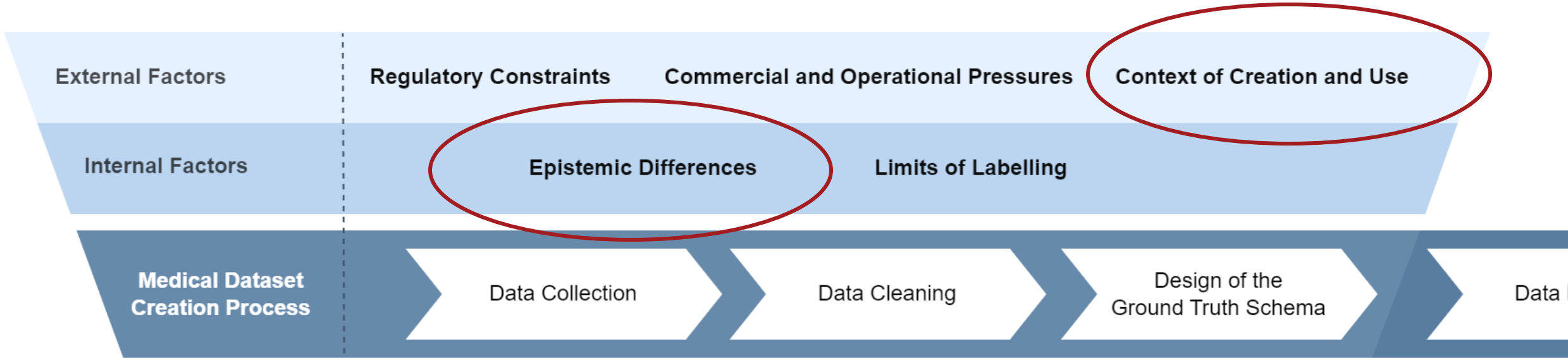
Zając, H., Avlona, N., Andersen, T., Kensing, F. & Shklovski, I. (2023) Ground Truth Or Dare: Factors Affecting The Creation Of Medical Datasets For Training AI. In *Proceedings of AIES 2023,* Montreal, Canada

Designing "ground truth"



Zając, H., Avlona, N., Andersen, T., Kensing, F. & Shklovski, I. (2023) Ground Truth Or Dare: Factors Affecting The Creation Of Medical Datasets For Training AI. In *Proceedings of AIES 2023,* Montreal, Canada

Designing "ground truth"



Zając, H., Avlona, N., Andersen, T., Kensing, F. & Shklovski, I. (2023) Ground Truth Or Dare: Factors Affecting The Creation Of Medical Datasets For Training AI. In *Proceedings of AIES 2023,* Montreal, Canada

Context of Creation and Use – a matter of geography?

DCODE NETWORK
RETHINK DESIGN – CHANGE OUR FUTURE

Data is collected in one country and model is intended to be applied elsewhere…

"So if you wanted that in the hierarchy, it could be there." (P1)

Is it aortic unfolding? Because I clearly remember this sentence from [the East African country] reports, "aortic unfolding due to chronic hypertension" (P2).

Zając, H., Avlona, N., Andersen, T., Kensing, F. & Shklovski, I. (2023) Ground Truth Or Dare: Factors Affecting The Creation Of Medical Datasets For Training AI. In *Proceedings of AIES 2023,* Montreal, Canada

Epistemic Differences – what is bias (who decides)?

DCODE NETWORK
RETHINK DESIGN – CHANGE OUR FUTURE

- Medical professionals had to explain to data scientists that to detect some types of cancer it is necessary to look at more than just the organ in question

- Providing no "extraneous or potentially biasing information"
  - "Asking a radiologist to categorise something on a picture only without getting any information on the patient. Is like asking a surgeon to look at the scars on a patient and having him tell you what kind of surgery that patient had".

Zając, H., Avlona, N., Andersen, T., Kensing, F. & Shklovski, I. (2023) Ground Truth Or Dare: Factors Affecting The Creation Of Medical Datasets For Training AI. In *Proceedings of AIES 2023,* Montreal, Canada

The problem of data quality – when is it "good enough"?

- People "torque" into data because nobody really fits all the categories well when filling out a form
  - How do you know which answer actually matters?
- High-quality data is an unattainable ideal to strive for
  - Achieving data quality is a battleground of contestation and compromise
  - There are no representative datasets in the medical domain (or anywhere)
- Within the EU regulatory context compliance becomes a dimension of data quality that orders and constrains other dimensions
  - Compliance becomes as a mode of ordering constraining what is possible in terms of accuracy, structure, and timeliness of produced data-sets.

Avlona, N. R., & Shklovski, I. (2024). Torquing patients into data: enactments of care about, for and through medical data in algorithmic systems. Information, Communication & Society, 27(4), 735-757.

Current work in data management – quality assessments

- An enormous variety of data quality dimensions is a challenge
- Recognition that quality can not be assessed WITHOUT engaging with context
- Five facets of data/context assessment
  - Data itself (meta-data, data structure, missing data)
  - Data source (perceived quality/reputation) of data source, data provenance, traceability
  - Data systems/infrastructures (recoverability, portability, compliance, audiability)
  - Data purpose/task (what is needed and is the data sufficient)
  - Data users (who is using the data, who is affected by its use)

- Mohammed et al. (2024) propose **29 dimensions of data quality**

But there is a lot of data...

CHRIS ANDERSON SCIENCE 06.23.08 12:00 PM

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

Peter Norvig, Google's research director, offered an update to George Box's maxim: "All models are wrong, and increasingly you can succeed without them."

AI Programs in Japan Are Forcing Workers to Smile More

September 12, 2024 | Kate Dedenbach, Joshua D. Nadreau, Karen L. Odash, and Nan Sato © Fisher Phillips

A Japanese supermarket chain is getting attention for implementing an AI tool called “Mr. Smile” that monitors workers for the quality and quantity of their smiles when interacting with customers, raising questions around the globe about how far to allow artificial intelligence into the workplace. Mr. Smile, introduced at eight Aeon locations earlier this year, initially monitored over 3,000 employees with artificial intelligence technology, using more than 450 elements to assess facial expressions, the length and sincerity of smiles, and the volume and tone of voice. Deeming the trial a success, Aeon just announced it will roll out the system to all 240 of its stores and monitor tens of thousands of workers across Japan “to standardize staff members’ smiles and satisfy customers to the maximum.” Could companies in the U.S. implement AI-driven emotional monitoring? Here’s what employers in Japan and the U.S. should consider when looking into AI technology that mandates specific emotions from its workers.

GOODHART'S LAW

WHEN A MEASURE BECOMES A TARGET,
IT CEASES TO BE A GOOD MEASURE

IF YOU MEASURE PEOPLE ON...	NUMBER OF NAILS MADE	WEIGHT OF NAILS MADE
THEN YOU MIGHT GET	1000's OF TINY NAILS	A FEW GIANT, HEAVY NAILS

Assumptions in the design of computational systems

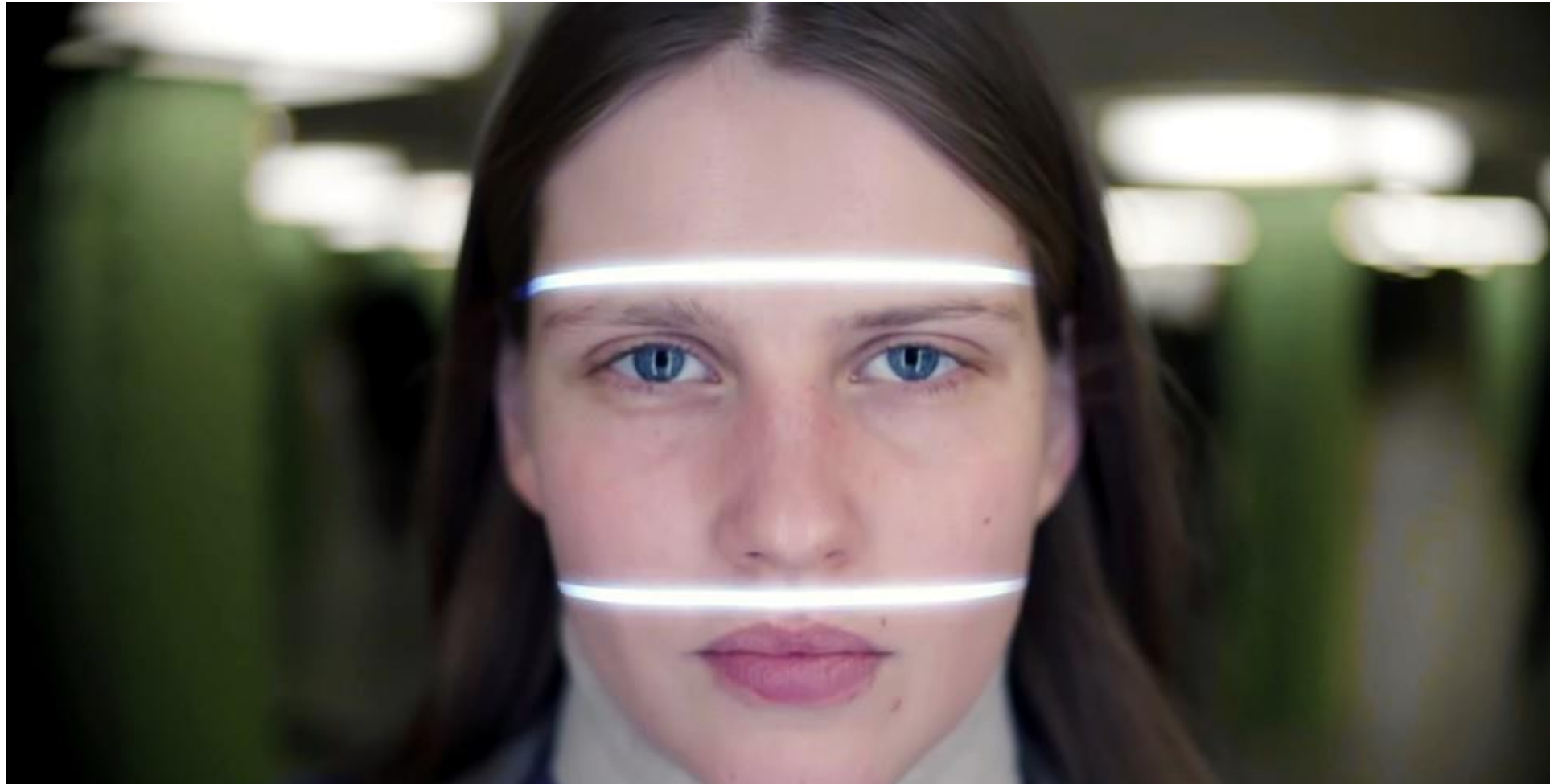
Many data-driven technology ideas resemble utopian social projects that can be naïve and misguided without much reflection on the world because engineering reasoning can be too instrumental, too focused on the individual

Phil Agre, 1997, Computing as Social Practice

The world always changes, machine learning models assume it is static

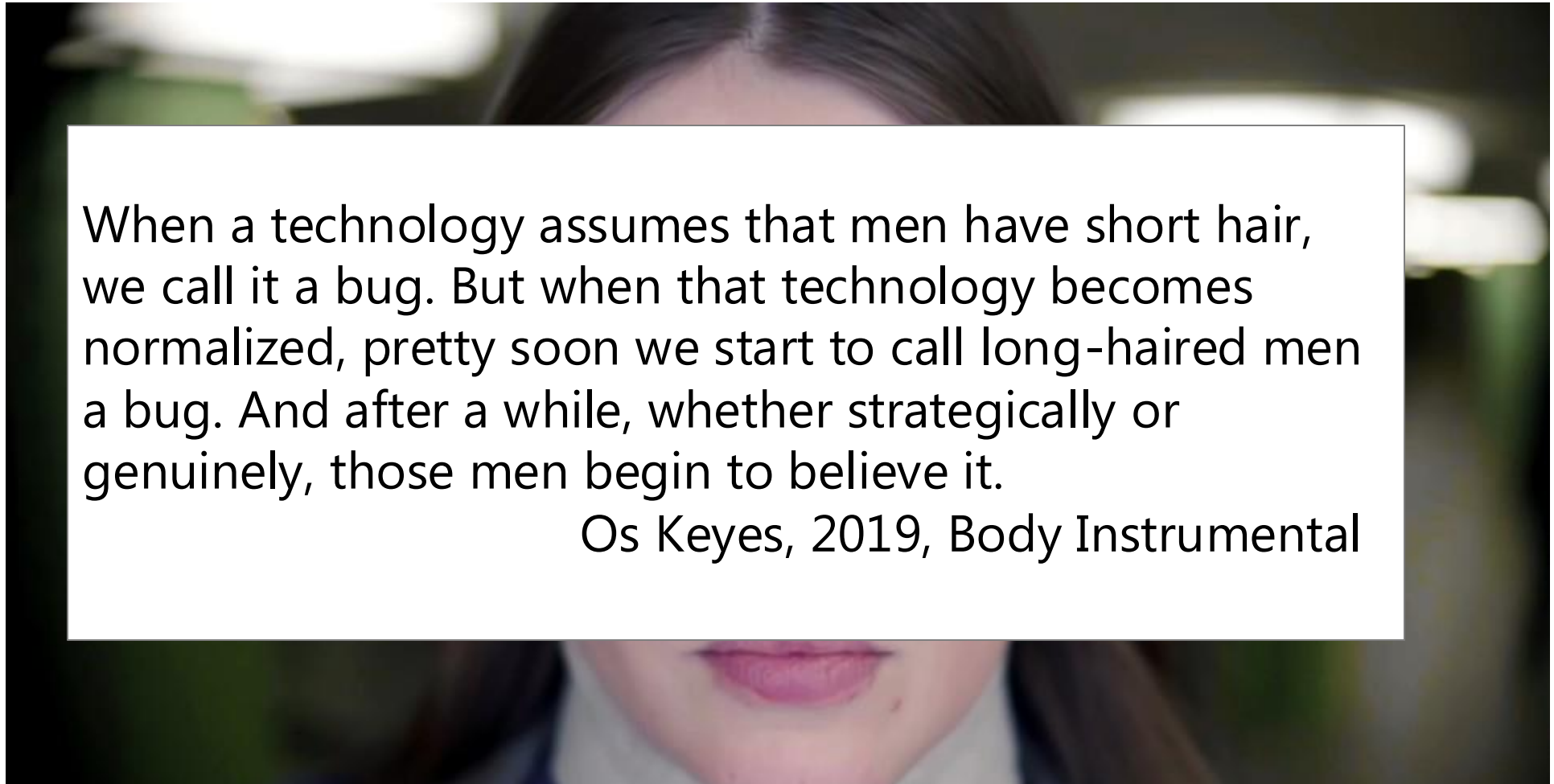
One reason for changing how we think about data

When systems we deploy change systems of incentives, they change worlds…



The reason for debates about ethics and computing

When systems we deploy change systems of incentives, they change worlds…



When a technology assumes that men have short hair, we call it a bug. But when that technology becomes normalized, pretty soon we start to call long-haired men a bug. And after a while, whether strategically or genuinely, those men begin to believe it.

Os Keyes, 2019, Body Instrumental

So what?

- Data quality is a **socio-technical concept**
  - Impossible to address by purely technical means (this is why so little progress)
  - Requires re-thinking current ways of conceptualizing data in CS

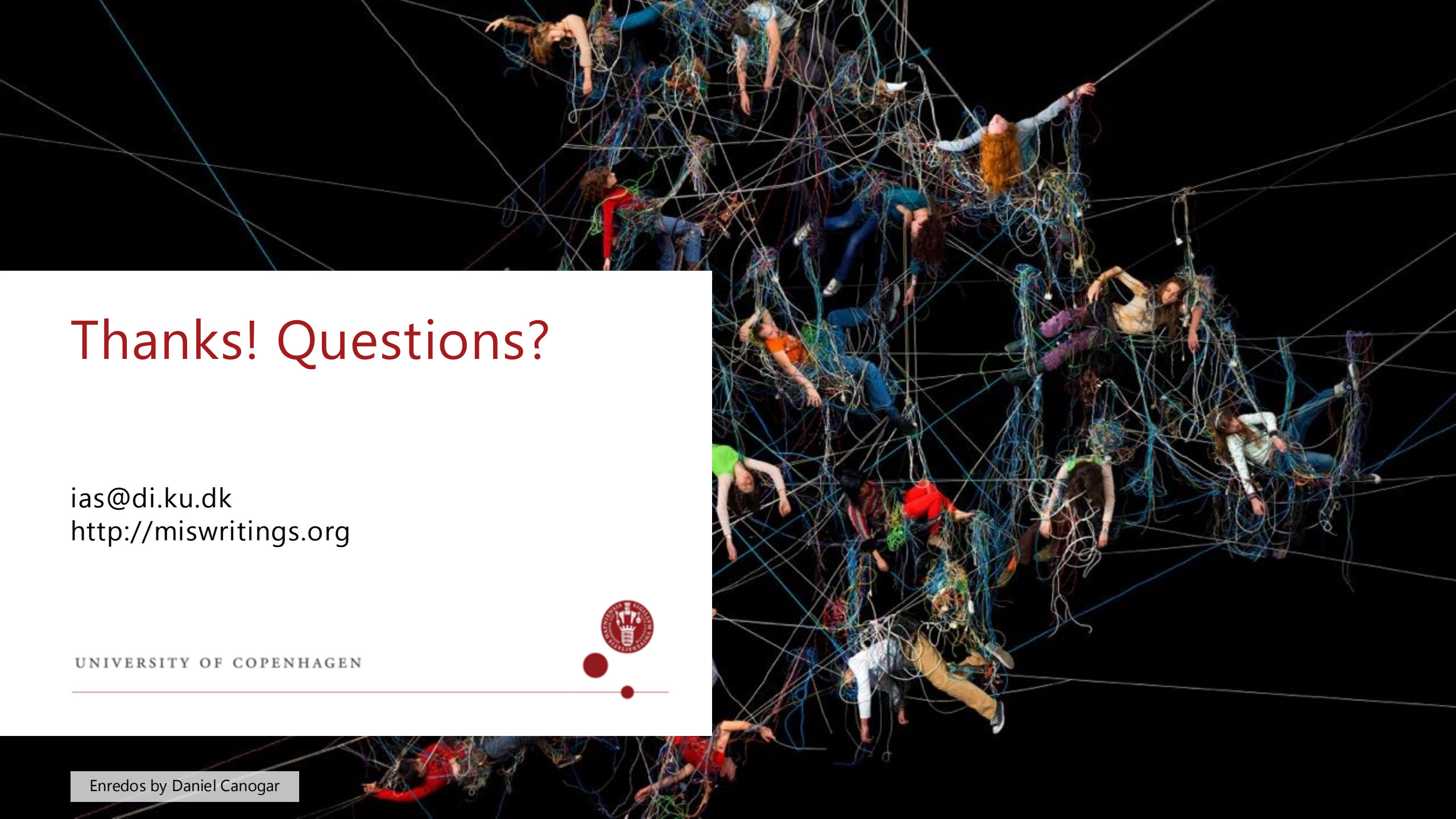
- Computing requires something to compute
  - All data are socially constructed – imperfect representations of measurable reality
  - Data are expressions of relations and politics – veneer of objectivity

- Data quality is a continuous challenge
  - Always think about the limits of your dataset (wonder how it was created)
  - Most datasets are like looking through a distorted fun-mirror (keep this in mind)

Thanks! Questions?

ias@di.ku.dk
http://miswritings.org

UNIVERSITY OF COPENHAGEN



Enredos by Daniel Canogar