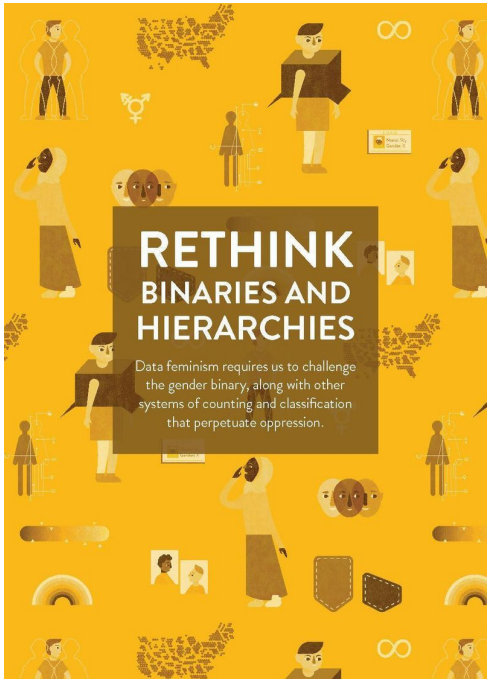
Rethinking Design-Tech Beyond Binaries

Amir H. Payberah

payberah@kth.se

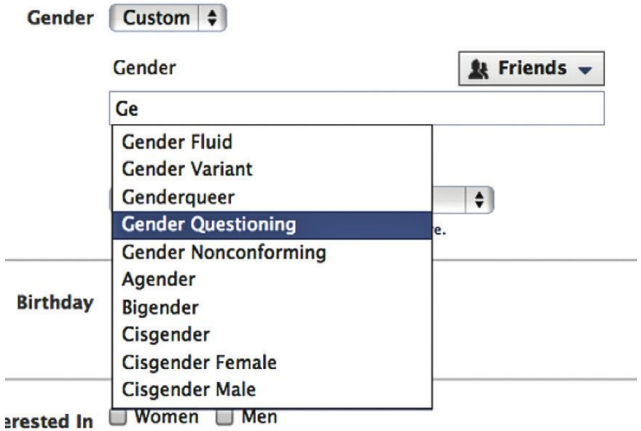
May 23, 2025

Rethink Binaries and Hierarchies



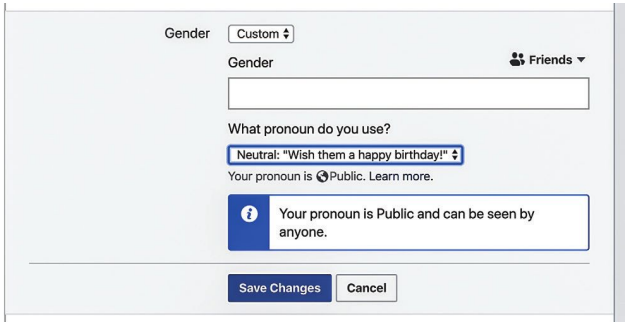
https://tinyurl.com/ypezb3hv

Facebook Gender Field



2014

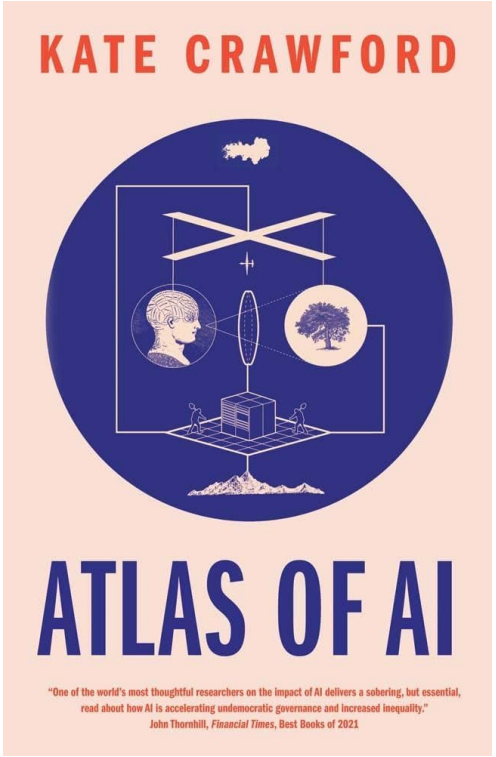
https://tinyurl.com/5eysafdd



2018

https://tinyurl.com/2p9esmjm

Atlas of AI - Kate Crawford



https://tinyurl.com/97spajab

Samuel Morton Skull Collection

- Collected over 1000 skulls in the early 1800s.
- Skull labels include racial and descriptive terms.
- Aimed to classify and rank human races objectively.

https://tinyurl.com/y7mn963k

Morton's Classification

- Used cranial measurements to argue for racial hierarchy in intelligence.

- Categorized into five races:
  - African
  - Native American
  - Caucasian
  - Malay
  - Mongolian

https://tinyurl.com/2dfv2fd3

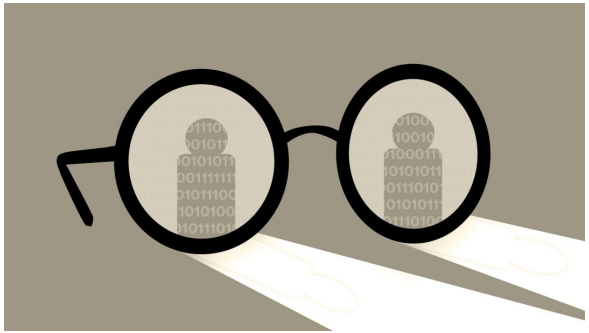
Scientific Impact and Criticism

- Considered cutting-edge science of the time.
- Cited for nearly a century to justify racial hierarchies.
- Used to legitimize slavery and racial segregation in the US.

https://tinyurl.com/z9syrybx

Reassessment of Morton's Work

- Found selective sampling, procedural omissions, and assumptions that influenced results.
- Emphasized the social context influencing Morton’s findings.
- Classification systems in AI can perpetuate biases.
- Importance of questioning underlying assumptions in data classification.

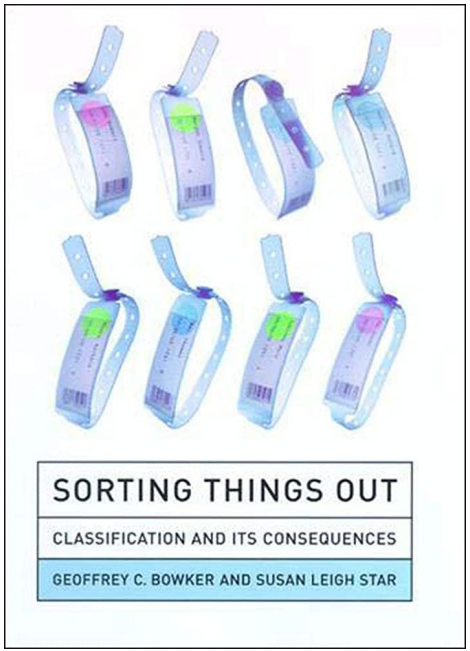


https://tinyurl.com/9nmkws7k

Questions Raised

- Who creates and changes these categories?
- What values might be encoded into them?
- What hidden hierarchies they encode?
- How do they spread and affect public policy?
- Whether these classifications should exist in the first place?

Sorting Things Out - Geoffrey Bowker & Susan Leigh Star



They explore how classification systems shape our understanding of the world and affect the people and practices within it.

Classification

- Segmentation of the world; set of boxes (metaphorical or literal) to organize items for work or knowledge production.

- Embedded in our daily life.
  - Sorting clothes for the wash.
  - Classifying medications and books.
  - Organizing folders in computers.

https://tinyurl.com/k89s9aab

Power and Ethics of Classifications

- These systems are usually invisible but crucial for societal order.
- They favor some perspectives while silencing others.
- When classifications fail, they become noticeable.
- Examples
  - Airport security systems (body scanners)
  - Automated hiring tools
  - Washed mixed color clothes

https://tinyurl.com/y653brh5

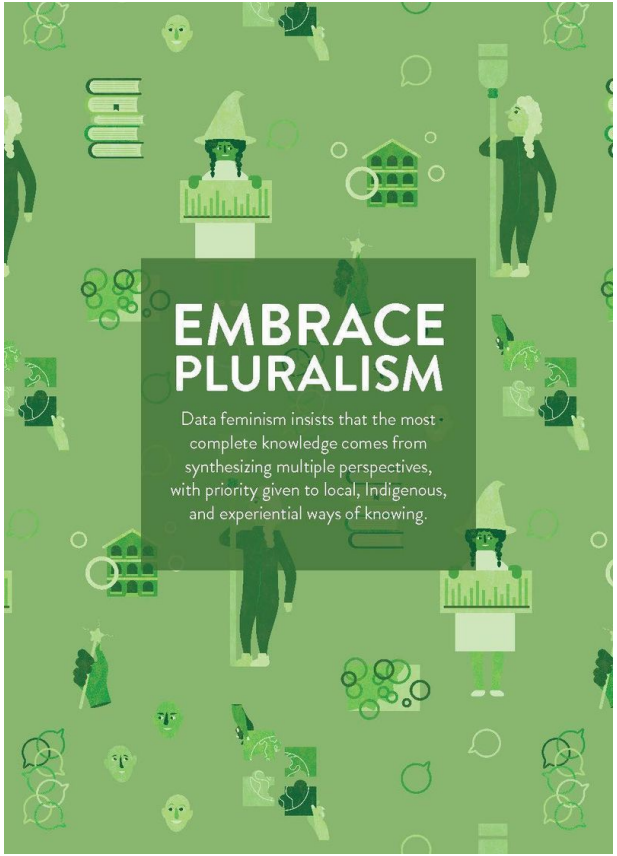
Questions for Discussion

- How do classification systems handle exceptions, anomalies, or ambiguous cases?

- What should we do amid potential classification harms?

Challenge These Practices

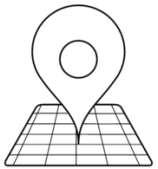
- Embrace pluralism in data science
- Value and include the marginalized communities viewpoints at all stages of the process.
- Recognizing a multiplicity of voices for a complete picture.



Anti-Eviction Mapping Project

Anti-Eviction Mapping Project (AEMP)

- AEMP: an example of how data can be used to serve the interests of marginalized communities.



Anti-Eviction Mapping Project

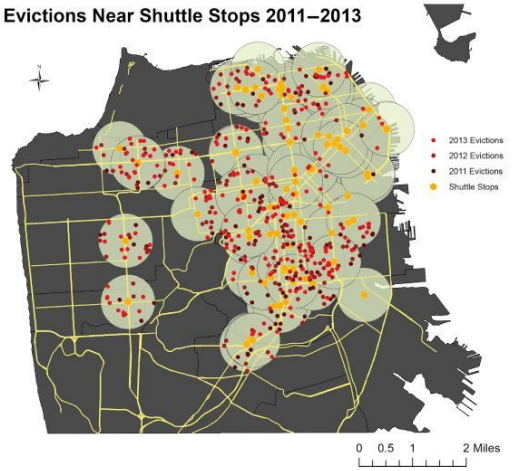
Housing Crisis and Evictions in San Francisco

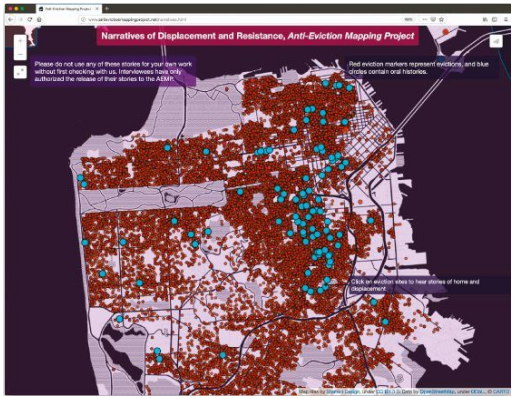
- High income gap between the richest and poorest in San Francisco.
- Wealth gap has racial and gender dimensions.
- Escalating rate of evictions since 2003.

https://tinyurl.com/ycx2fjz2

AEMP Community

- A collective of housing justice activists, researchers, artists, and historians (founded in 2013).
- With many designers and nonexperts (local residents) to map evictions.
- Committed to anti-racist, feminist, and decolonial methodologies.





Feminist Counterdata and Countervisualization

- AEMP documents the displacement of residents through gentrification and eviction, particularly in communities of color and low-income neighborhoods.

- A set of feminist counterdata collection and countervisualization strategies.
  - Where people go after they are evicted?
  - How many of those people end up homeless?
  - Which landlords are responsible for systematically evicting major blocks of the city?

- https://antievictionmap.com

https://tinyurl.com/ywnzbnxx

Embracing Pluralism in Data Science

- AEMP exemplifies the embrace pluralism principle of data feminism.
- Valuing various perspectives at all stages of the process, from data collection to cleaning to analysis to communication.
- It also means recognizing how data science methods can unintentionally silence voices for clarity, cleanliness, and control.

https://tinyurl.com/2kks39b6

Questions for Discussion

- How effective is collaborative data science (e.g., AEMP) in truly shifting power to marginalized communities, and what are the potential risks of using these methods in ways that still benefit those in power?

- Can pluralism and profit coexist in the field of data science, or are they fundamentally at odds?

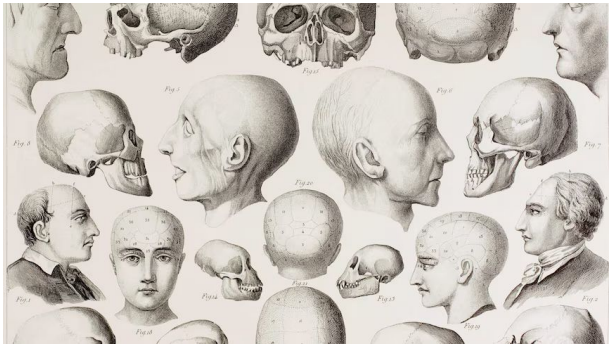
Data Cleaning

Cleaning and Preparing Data

- Data cleaning is considered a crucial part of data science.
- What might be lost in this process?
- Whose perspectives might be lost in this process?
- Whose perspectives might be additionally imposed?

Historical Roots and Ethical Considerations of Cleanliness

- Ideas of cleanliness in data have historical ties to eugenics.
  - Improve the genetic quality of human populations through selective breeding.
- Early statisticians like Pearson and Galton were also leaders in the eugenics movement.
- Influence modern beliefs in control and cleanliness in data science.

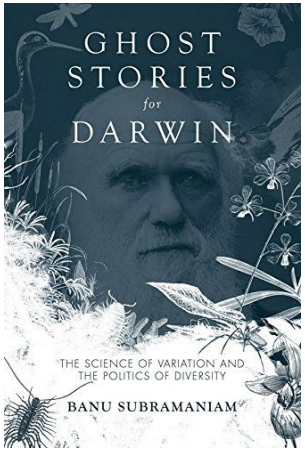


https://tinyurl.com/4b9j47yd

https://tinyurl.com/ydv28yh7

Ghost Stories for Darwin - Banu Subramaniam

- Argues that scientific practices are deeply embedded in cultural and political contexts.
- Critiques how the desire for purity and uniformity in scientific data has often mirrored broader societal efforts to control and categorize human populations.



https://tinyurl.com/ynky768r

The Cleaning Paradigm and Its Critique

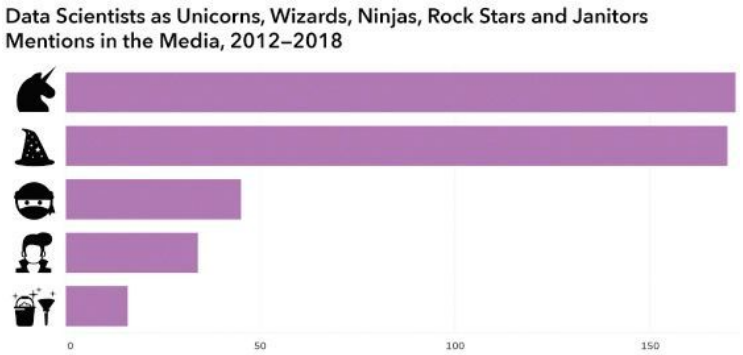
- Katie Rawson and Trevor Muñoz, in Against Cleaning argued that cleaning assumes an underlying correct order.
- Rich information can be lost during cleaning.
- Diversity-hiding trick

"This is not what I meant when I said 'we need better data cleansing!'"

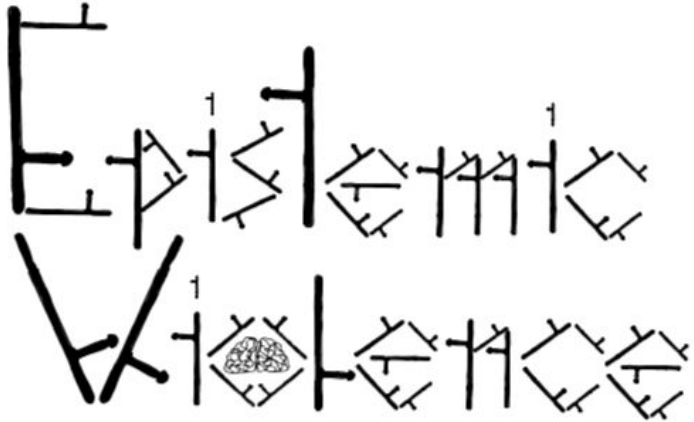
www.iwaysoftware.com/go/dataquality

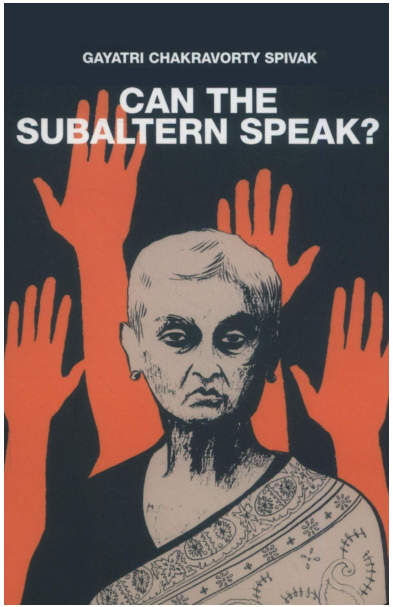
Data Scientists as Strangers

- Data do not need cleaning until there are strangers in the dataset.
  - Data scientists are often seen as rock stars, wizards, ninjas, or janitors.
- These roles imply solitary, extraordinary technical expertise.
- The solitary genius model overlooks the collaborative nature of data science.



Can The Subaltern Speak? Gayatri Spivak



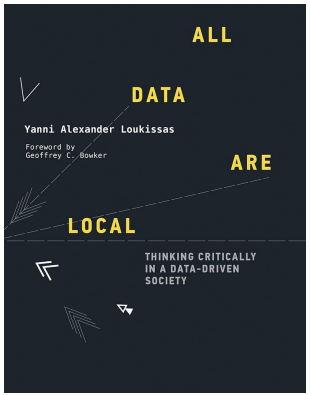


Harm caused by silencing or dismissing marginalized knowledge and perspectives.

Situated Knowledges

All Data Are Local - Yanni Loukissas

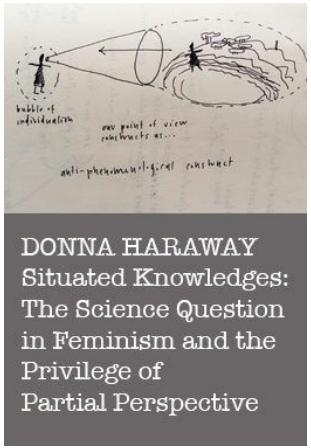
- Data is inherently local: it reflects the specific conditions, practices, and power structures of the settings in which it is created.
- Considering data settings rather than datasets.
- E.g., Clemson University Library's "upstate" term is clear locally, but confusing to outsiders.



https://lmc.gatech.edu/featured-people/yanni-loukissas

Situated Knowledges - Donna Haraway

- All knowledge is partial, no single person or group can claim an objective view of the Truth.
- People make knowledge from a particular standpoint: from a situated, embodied location in the world.

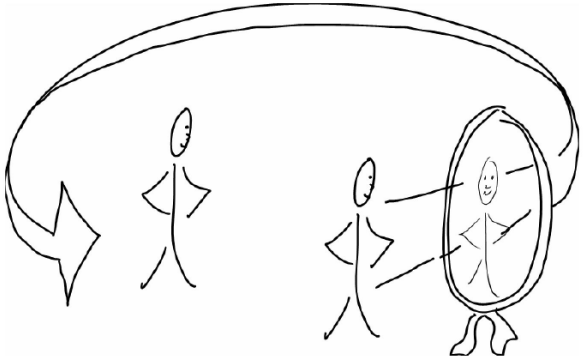


https://tinyurl.com/y6zzjhj5

- How do we start gaining a deeper understanding in data science?

Transparency and Reflexivity

- Transparency: revealing technical details.
- Reflexivity: who is doing the work and the process behind it.



https://tinyurl.com/5n99sjuj

From Data Ethics to Data Justice

Concepts That Secure Power because they locate the source of the problem in individuals or technical systems	**Concepts That Challenge Power** because they acknowledge structural power differentials and work toward dismantling them
Ethics	Justice
Bias	Oppresion
Fairness	Equity
Accountability	Co-liberation
Transparency	Reflexivity
Understanding Algorithms	Understanding history, culture, and context

Beyond Transparency and Reflexivity

- Actively and deliberately inviting other perspectives into the data analysis and storytelling process.

https://tinyurl.com/mrxy76mz

From Data for Good (Data Ethics) to Data for Co-liberation (Data Justice)

	"Data for good"	Data for co-liberation
Leadership by members of minoritized groups working in community		√
Money and resources managed by members of minoritized groups		√
Data owned and governed by the community		√
Quantitative data analysis "ground truthed" through a participatory, community-centered data analysis process		√
Data scientists are not rock stars and wizards, but rather facilitators and guides		√
Data education and knowledge transfer are part of the project design		√
Building social infrastructure—community solidarity and shared understanding—is part of the project design		√

Questions for Discussion

- What steps can we take to make data cleaning a more ethical and inclusive process?

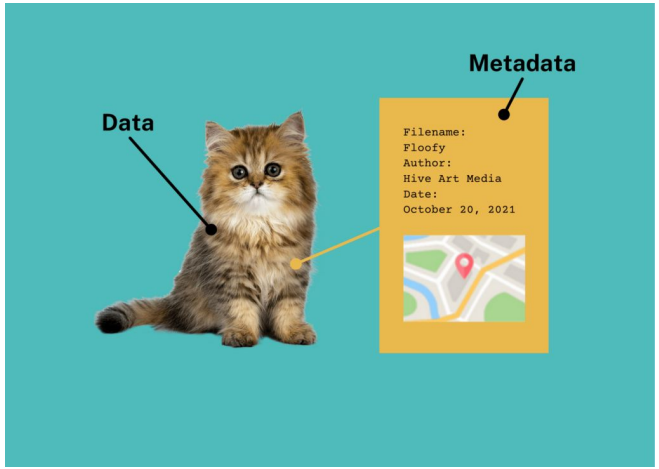
- Is it possible to achieve both technical accuracy and ethical responsibility in data cleaning, or are these goals sometimes in conflict?

Consider Context

https://tinyurl.com/3p9r2fnr

Context and Metadata in Data

- Many datasets lack context or metadata.
- Lack of context makes data exploration difficult.
- Without local knowledge, understanding power dynamics is challenging.



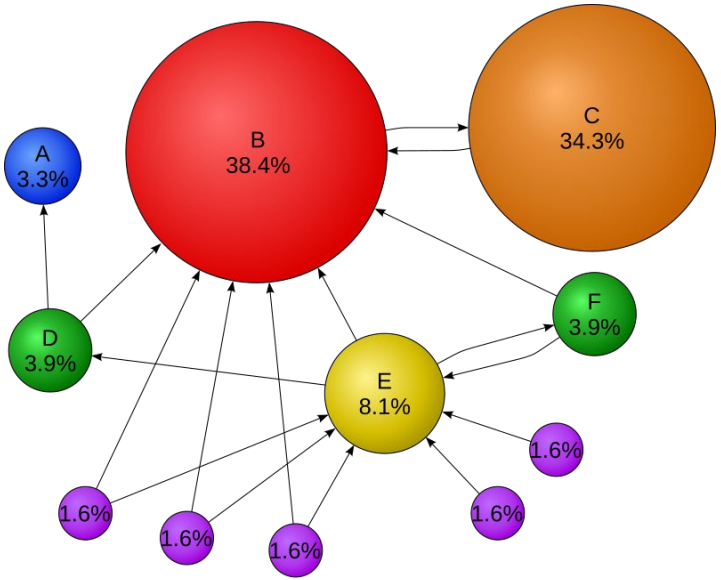
Context-Free Data Analysis

- Chris Anderson's "The End of Theory" claims data speak for themselves, eliminating the need for context.

- Statistical inference relies on sampling, but big data suggests using all data directly.

Is Correlation Enough?

- Anderson insists: correlation is enough.
- E.g., Google search
- But what happens when the number of links is also highly correlated with sexist, racist, and pornographic results?
- Correlation can reinforce societal biases.



Raw Data, Cooked Data, Cooking

- Data are not raw inputs; they are cooked through social, political, and historical contexts.
- One strategy for considering context is to consider the cooking process that produces “raw” data.
- Exploring and analyzing what is missing from a dataset.

https://tinyurl.com/yvhs3v6w

Formerly Gang-Involved Youth as Domain Experts for Analyzing Unstructured Twitter Data

- SAFELab: use Twitter data to understand and prevent gang violence.
- Deep listening and contextual analysis of social media.

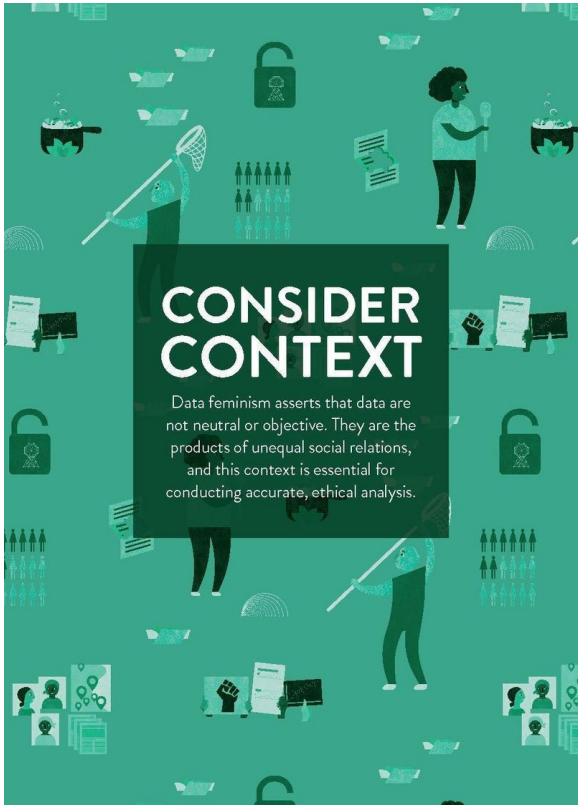
Anybody could get 👏 n dis 💩 🤷‍♀️ lackin period 🚰 💯

aint kill yo mans & ion kno ya homie

SAFE LAB

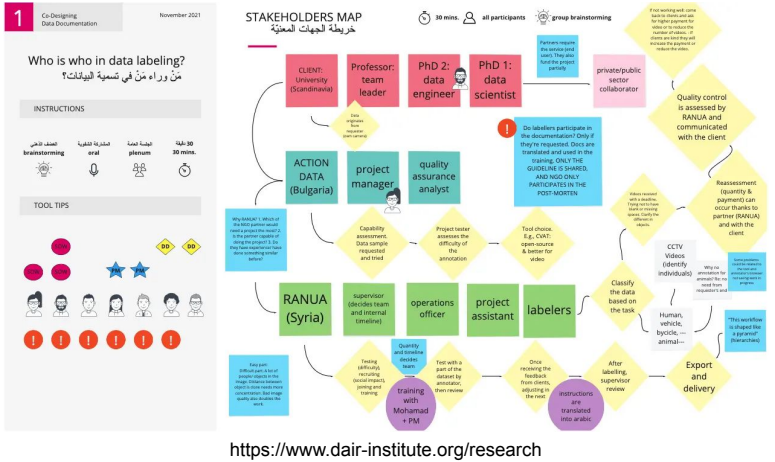
Consider Context

- Datasheets for datasets
- A participatory approach for data work



Datasheets for Datasets

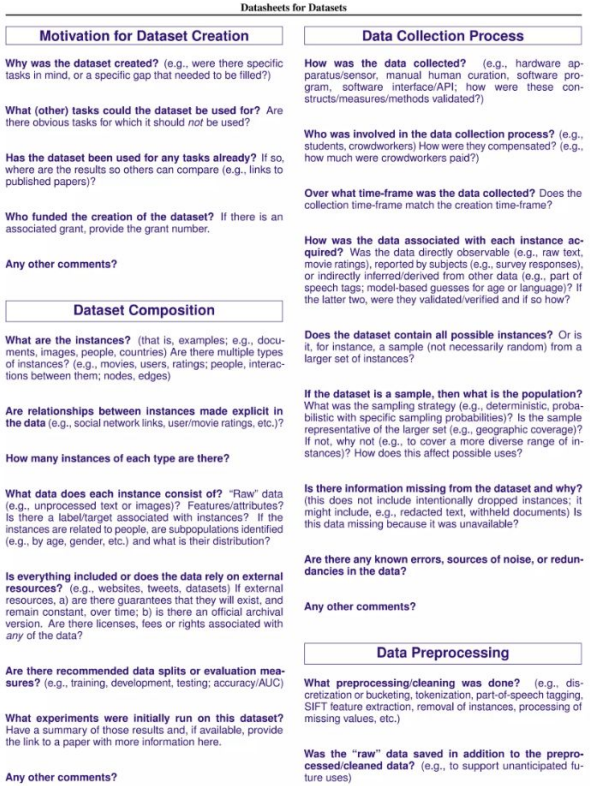
- Address the needs of two key stakeholder groups:
  - Dataset creators
  - Dataset consumers
- Objectives for dataset creators
  - Encourage careful reflection on the creation, distribution, and maintenance of a dataset, considering assumptions, risks, and potential impacts.
- Objectives for dataset consumers
  - Ensure they have the information they need to make informed decisions about using a dataset.



https://www.dair-institute.org/research

Datasheets for Datasets - Key Components

- Motivation
  - Why was the dataset created? What is its intended use?
- Composition
  - What does the dataset consist of? Who are the subjects?
- Collection process
  - How was the data collected? Any ethical concerns?
- Preprocessing
  - Any cleaning or transformations applied to the data?
- Distribution
  - How can the dataset be accessed or shared?
- Maintenance
  - Who is responsible for maintaining the dataset?



Datasheets for Datasets

Motivation for Dataset Creation

Why was the dataset created? (e.g., were there specific tasks in mind, or a specific gap that needed to be filled?)

What (other) tasks could the dataset be used for? Are there obvious tasks for which it should *not* be used?

Has the dataset been used for any tasks already? If so, where are the results so others can compare (e.g., links to published papers)?

Who funded the creation of the dataset? If there is an associated grant, provide the grant number.

Any other comments?

Dataset Composition

What are the instances? (that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges)

Are relationships between instances made explicit in the data (e.g., social network links, user/movie ratings, etc.)?

How many instances of each type are there?

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)? Features/attributes? Is there a label/target associated with instances? If the instances are related to people, are subpopulations identified (e.g., by age, gender, etc.) and what is their distribution?

Is everything included or does the data rely on external resources? (e.g., websites, tweets, datasets) If external resources, a) are there guarantees that they will exist, and remain constant, over time; b) is there an official archival version. Are there licenses, fees or rights associated with *any* of the data?

Are there recommended data splits or evaluation measures? (e.g., training, development, testing; accuracy/AUC)

What experiments were initially run on this dataset? Have a summary of those results and, if available, provide the link to a paper with more information here.

Any other comments?

Data Collection Process

How was the data collected? (e.g., hardware apparatus/sensor, manual human curation, software program, software interface/API; how were these constructs/measures/methods validated?)

Who was involved in the data collection process? (e.g., students, crowdworkers) How were they compensated? (e.g., how much were crowdworkers paid?)

Over what time-frame was the data collected? Does the collection time-frame match the creation time-frame?

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part of speech tags; model-based guesses for age or language)? If the latter two, were they validated/verified and if so how?

Does the dataset contain all possible instances? Or is it, for instance, a sample (not necessarily random) from a larger set of instances?

If the dataset is a sample, then what is the population? What was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? Is the sample representative of the larger set (e.g., geographic coverage)? If not, why not (e.g., to cover a more diverse range of instances)? How does this affect possible uses?

Is there information missing from the dataset and why? (this does not include intentionally dropped instances; it might include, e.g., redacted text, withheld documents) Is this data missing because it was unavailable?

Are there any known errors, sources of noise, or redundancies in the data?

Any other comments?

Data Preprocessing

What preprocessing/cleaning was done? (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values, etc.)

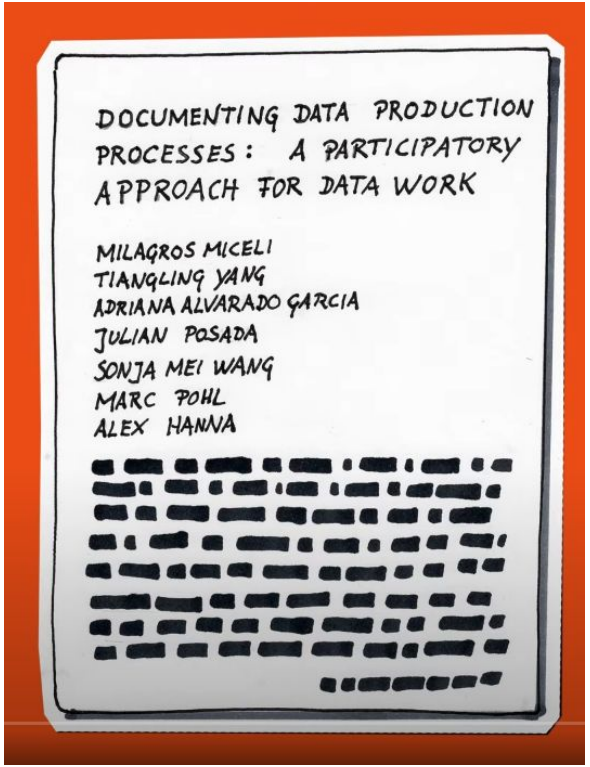
Was the "raw" data saved in addition to the preprocessed/cleaned data? (e.g., to support unanticipated future uses)

Datasheets for Datasets - Example (Health Records)

- Motivation
  - Collected to improve healthcare outcomes
- Composition
  - Contains anonymized patient records from 10 hospitals
- Collection process
  - Data collected through surveys and hospital records
- Preprocessing
  - Data anonymized and cleaned for missing entries
- Distribution
  - Available to researchers under a strict data-sharing agreement
- Maintenance
  - Updated annually by the hospital consortium

Participatory Documentation

- Involves multiple stakeholders (data creators, curators, users) in the documentation process.
- Encourages diverse perspectives and shared ownership over data production practices.



https://tinyurl.com/4ndhe9hm

Participatory Documentation - Key Steps

- Involve stakeholders early
  - Engage data collectors, curators, and users from the beginning of the project.
- Map data creation
  - Document the entire data production process, from collection to final output, noting key decisions.
- Record rationales
  - Explain why certain data collection, cleaning, or curation choices were made.
- Highlight biases and gaps
  - Identify and openly document any biases or limitations in the data.

Who Should Provide Context?

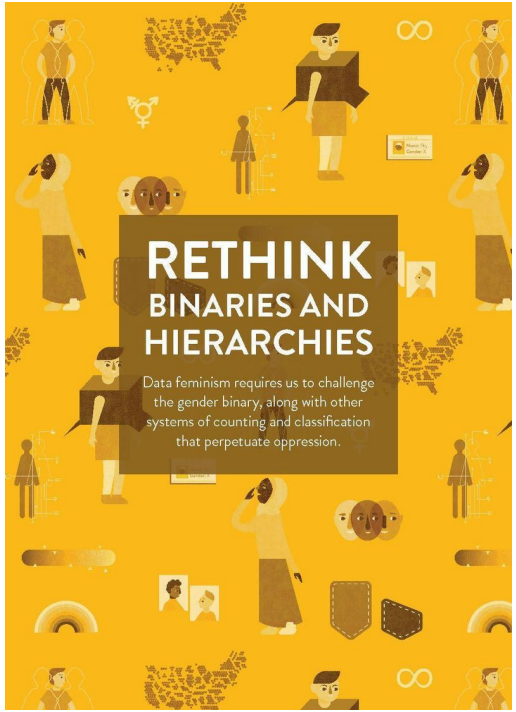
- Which actors in the data ecosystem are responsible for providing context?

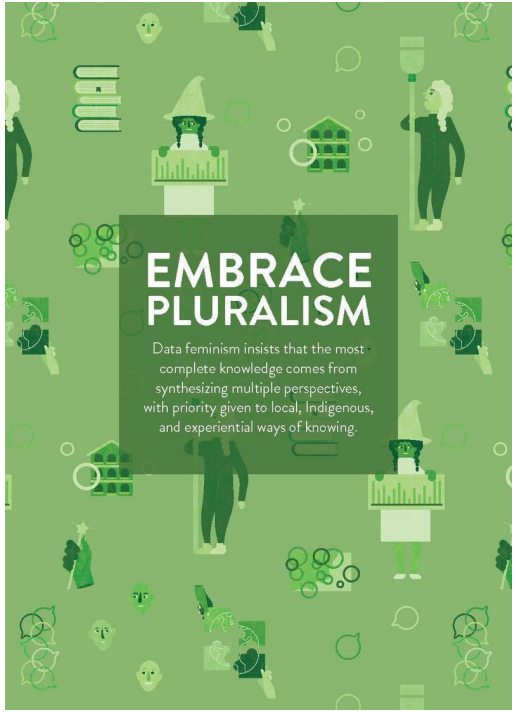
- End users?
  - Challenges of verifying data on a deadline and budget.

- Data publishers?
  - Often overstated capabilities and lack of documentation.

- Data intermediaries?
  - Librarians, journalists, nonprofits, educators can help but need funding and capacity-building.

Summary


RETHINK
BINARIES AND
HIERARCHIES
Data feminism requires us to challenge the gender binary, along with other systems of counting and classification that perpetuate oppression.


EMBRACE
PLURALISM
Data feminism insists that the most complete knowledge comes from synthesizing multiple perspectives, with priority given to local, Indigenous, and experiential ways of knowing.


CONSIDER
CONTEXT
Data feminism asserts that data are not neutral or objective. They are the products of unequal social relations, and this context is essential for conducting accurate, ethical analysis.

- Data Feminism, (ch. 3-5)
- Data colonialism: Rethinking big data's relation to the contemporary subject, Nick Couldry et al., 2019
- Against Cleaning, Katie Rawson and Trevor Muñoz [link]
- All Data Are Local: Thinking Critically in a Data-Driven Society, Yanni Loukissas (ch. introduction)
- Situated knowledges: The science question in feminism and the privilege of partial perspective, Donna Haraway, Feminist Studie, 1988
- Social media for large studies of behavior, Derek Ruths and Jürgen Pfeffer, 2014
- Artificial Intelligence and Inclusion: Formerly Gang-Involved Youth as Domain Experts for Analyzing Unstructured Twitter Data, William R. Frey et al., 2020
- Datasheets for datasets, Timnit Gebru et al., Communications of the ACM, 2021
- The Dataset Nutrition Label, Sarah Holland et al., Data Protection and Privacy, 2020
- Documenting Data Production Processes: A Participatory Approach for Data Work, Milagros Miceli et al., 2022