16/12/25

Curating Data: A Portfolio

Clara Holst (CH), AU737540, 202306022@post.au.dk
5th Semester, Cognitive Science BSc, Arts Faculty, Aarhus University
Jens Chr. Skous Vej 2, 8000 Aarhus, Denmark

Critical Data Studies, Elective
Curating Data Course
Total Characters: 563.741 ~ 235 pages (text and code)

Assignment	Title	Character Count
1	Critical Report: Curating the Holst Family Coin Collection	10.728 (text)
2	Methodology Report: “Liao Huanxing” Documentation in Wikidata	13.279 (text)
3	Visualizing Consumer Psychology: A Critical Analysis of E-commerce Behavioral Data	14.399 (text) 51.000 (code)
4	Website Creation: Curating Data / Making Sense of Knowledge	429.428 (code)

Curatorial statement

This portfolio presents a curated collection of work from the Curating Data course, documenting my exploration of how data transforms into knowledge through critical curation practices. As a Cognitive Science student, I approached this journey with particular interest in how humans organize and interpret information.

My work engages with four key curation practices: collecting, categorizing, visualizing, and archiving. Each assignment demonstrates how these practices shape understanding, from transforming physical coins into structured data to making marginalized histories visible through Wikidata. Even extending this exploration to algorithmic categorization through recommender systems analysis.

Theoretical frameworks from critical data studies inform this work, particularly concepts of data assemblages and the DIKW hierarchy (Data, Information, Knowledge, and Wisdom). These perspectives reveal curation as an active, interpretive process rather than neutral organization. Through creating relationships between different data forms, I have generated new meanings and insights.

Ethical considerations form a crucial foundation throughout this portfolio. I have maintained commitments to responsibility and care in representation, adherence to GDPR and privacy standards, balancing open access with proper attribution, and continuous reflexivity about power dynamics in data work.

Critical Report: Curating the Holst Family Coin Collection

Clara Holst (CH), AU737540, 202306022@post.au.dk
5th Semester, Cognitive Science BSc, Arts Faculty, Aarhus University
Jens Chr. Skous Vej 2, 8000 Aarhus, Denmark

Critical Data Studies, Elective
Curating Data Course
Characters: 10.728 ~ 4,5 pages

1. Process of Collecting, Digitising, and Datafying

The project began with the Holst Family Coin Collection of 75 coins. Each coin was systematically photographed on both sides with a OnePlus 8T and measured (weight, diameter) using digital tools. These physical and contextual details (monarch, mint, catalog reference) were entered into Google Sheets.[1] Later on, 6 duplicates of existing coins in the collection were excluded.

The dataset was organised into 25 variables: Identification (ID, Year, Country, Monarch), Physical Attributes (Material, Weight, Size), Numismatic Details (Mint, Rarity, Reference), and Collection Management (Condition, Current Worth). This process transformed the heterogeneous objects into a structured archive. By doing so, the collection became not only preserved but also reconfigured into a tool for comparative analysis.

[1] ClaraHolst_HolstFamilyCoinCollection_A1.xlsx

2. Critical Reflection on Design Decisions

The choice of variables was shaped by a dual purpose: to serve both numismatic cataloguing and personal collection management. Therefore, standard identifiers like Year, Country, and Reign are essential for any coin dataset, enabling classification and historical placement. Additionally, variables like Material, Weight, and Diameter objectify the coins through measurable, quantitative properties, aligning with traditional numismatic practice.

The decision to include a subjective "Condition" grade creates a limit to the collector's personal perception. Furthermore, sensory experiences like the coin's "feel" or the detailed historical narrative behind a design change (e.g., the shift from silver to aluminium bronze in Danish coins post-WWI) were lost, as they resist easy quantification into spreadsheet cells. More significantly, ‘Mintage Quantity’ was excluded; a key variable for objectively assessing rarity. Instead, rarity was approximated using Numista.com’s 0 - 100 index, where 0 denotes common and 100 denotes extremely rare.

Additional limitations arise from practical constraints. Photographing with a phone may result in color distortions or inconsistent lighting, and measurement tools have finite precision. The dataset also reflects personal selection choices, so coins collected opportunistically or for sentimental reasons may overrepresent certain types or countries, limiting representativeness.

3. How Decisions Shaped the Dataset

These design decisions fundamentally shaped the resulting knowledge. The tabular, quantitative structure of the spreadsheet makes comparative analysis exceptionally efficient. One can instantly filter all coins from Margrethe II's reign or sort by weight to see material changes over time.

However, this structure also flattens the objects. The coin as a holistic historical artifact is fragmented into discrete cells. The rich, contextual stories behind the coins; why a particular motif was chosen, the economic conditions influencing a metal change, are silenced in favour of comparable data points. The dataset hereby becomes limited to the grid, as discussed by Dourish (2017).

The grid is defined as an "anticipatory" structure that dictates what information is relevant. By choosing variables like "Current Worth," the dataset is framed through a collector/market lens, elevating economic value as a key metric of importance. The omissions create a silence, limiting the dataset's utility for deep numismatic research and making it reliant on external catalog references (KM#, Schön#) for authority.

4. Connection to Course Theories and Concepts

First, the very act of datafication resonates with the concept that "raw data is an oxymoron" (Gitelman & Jackson, 2013). The coins themselves are not data. They become data only through a series of deliberate, "cooking" processes: selecting which coins to include (excluding duplicates), choosing which attributes to measure, and defining categories like "Colour" (Silver, Golden) or "Condition." Each variable represents a choice about what is worth capturing, transforming the messy reality of physical objects into a clean, structured dataset. This dataset is not a neutral reflection of the collection but a constructed representation shaped by my decisions, tools, and purposes.

Second, the design decisions exemplify the power and limitations of quantification (Wernimont, 2021). Quantification is an agential practice that makes the world manageable and comparable. By turning traits like material composition and size into numbers and labels, the coins were digitized for sorting, filtering, and analysis; the core functions of a spreadsheet. Yet Wernimont argues that quantification is not merely descriptive, but also world-making. The inability to capture a coin's historical narrative or aesthetic appeal means these aspects are excluded from the dataset's 'official' knowledge, showing how quantification empowers some forms of knowledge while marginalizing others.

In conclusion, curating the coin collection into a dataset was an exercise in knowledge creation, not mere transcription. The resulting spreadsheet is a powerful tool for specific queries, but is also a product of specific curatorial choices that highlight certain truths while silencing others.

5. Metadata as Contextual Infrastructure

Beyond the tabular dataset, I created structured metadata to situate the Holst Family Coin Collection within its wider context. The metadata records key aspects such as the total size of the archive, its temporal and geographic scope, the device used for photographing the coins, the location of image capture, and the date of the last update. It functions as an interpretive layer, documenting the conditions under which the dataset was produced (Greenberg et. al., 2023).

Details such as photographing the coins with a OnePlus 8T in Aarhus V, Denmark, and excluding five duplicate coins reveal the practical and methodological choices that shaped the collection. These elements increase transparency and allow future users to understand not only the coins but also the processes, constraints, and decisions that influenced their digital representation (Horsch, M.T. et. al., 2021).

6. Ethical Considerations in Curating the Coin Collection

Ethical considerations are central to curating the Holst Family Coin Collection. The collection's methodology aligns closely with emerging ethical guidelines for dataset curation (Mahima Pushkarna et al., 2022). Assessing a coin's condition or estimating its market value involves subjective judgment. Making these interpretive steps explicit through metadata and documentation prevents the dataset from appearing more objective than it is.

Including collectors' names, photograph locations, and technical details about image capture required careful reflection on necessity and potential exposure. While the dataset contains no sensitive personal data, a minimalist and purpose-driven approach ensures responsible practice (David Mindel et al., 2021).

The personal connection to the coins adds an ethical dimension. Many coins were collected as souvenirs or found in natural settings, and even coins with little monetary value hold significant sentimental importance. This influenced which coins were documented and how, blending personal attachment with numismatic categorization. This approach demonstrates a holistic, purpose-driven method of cultural preservation that goes beyond mere cataloging (Will Orr et al., 2024.)

The dataset also incorporates external sources, including rarity indices and catalog references from Numista and other authorities. Proper citation respects intellectual property and ensures transparency. Overall, the ethical stance emphasizes care, accuracy, and openness, acknowledging both technical and personal factors that shaped the collection.

Conclusion

Curating the Holst Family Coin Collection transformed physical coins into a structured dataset, enabling systematic preservation and comparison. Decisions such as variable selection, grading condition, and approximating rarity shaped the knowledge produced while highlighting the limits of quantification. The dataset allows efficient analysis but reduces the coins' historical and aesthetic narratives to discrete cells.

This process illustrates how datafication constructs knowledge, showing that even carefully measured attributes cannot fully capture the coins' stories or material significance. Metadata and ethical considerations provide context and transparency, acknowledging personal connections, subjective judgments, and external references without overemphasizing them.

Overall, the Holst Family Coin Collection dataset demonstrates the interplay of curatorial choices, personal meaning, and analytical utility, revealing how context and deliberate decisions shape both the representation and understanding of cultural objects.

References

- Dourish, P. (2017). Spreadsheets and Spreadsheet Events in Organizational Life. In The Stuff of Bits: An Essay on the Materialities of Information (pp. 81–104). The MIT Press.

- Gitelman, L., & Jackson, V. (2013). Introduction. In "Raw Data" Is an Oxymoron (pp. 1–12). The MIT Press.

- Greenberg, Jane., Mingfang Wu., Wei Liu., Fenghong Liu.; Metadata as Data Intelligence. Data Intelligence 2023; 5 (1): 1–5. doi: https://doi.org/10.1162/dint_e_00212

- Holst, C. (2025). *Complete dataset* [Google Sheets]. Google. X ClaraHolst_HolstFamilyCoinCollection_A1.xlsx

- Horsch, M.T., Chiacchiera, S., Cavalcanti, W.L., Schembera, B. (2021). Research Data Infrastructures and Engineering Metadata. In: Data Technology in Materials Modelling. SpringerBriefs in Applied Sciences and Technology. Springer, Cham. https://doi.org/10.1007/978-3-030-68597-3_2

- Kitchin, R. (2022). Introducing Data. In The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences (pp. 1–19). SAGE Publications Ltd.

- Mindel, D. (2021). Ethics and digital collections: a selective overview of evolving complexities. Journal of Documentation.

- Numista. (n.d.). *Numista coin catalog*. Numista. Retrieved September 24, 2025, from https://en.numista.com/

- Orr, W., & Crawford, K. (2024). Building Better Datasets: Seven Recommendations for Responsible Design from Dataset Creators. arXiv.org.

- Pushkarna, M., Zaldivar, A., & Kjartansson, O. (2022). Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. Conference on Fairness, Accountability and Transparency.

- Wernimont, J. (2021). Quantification. In N. B. Thylstrup, D. Agostinho, A. Ring, C. D'Ignazio, & K. Veel (Eds.), Uncertain Archives: Critical Keywords for Big Data (pp. 427–431). The MIT Press.

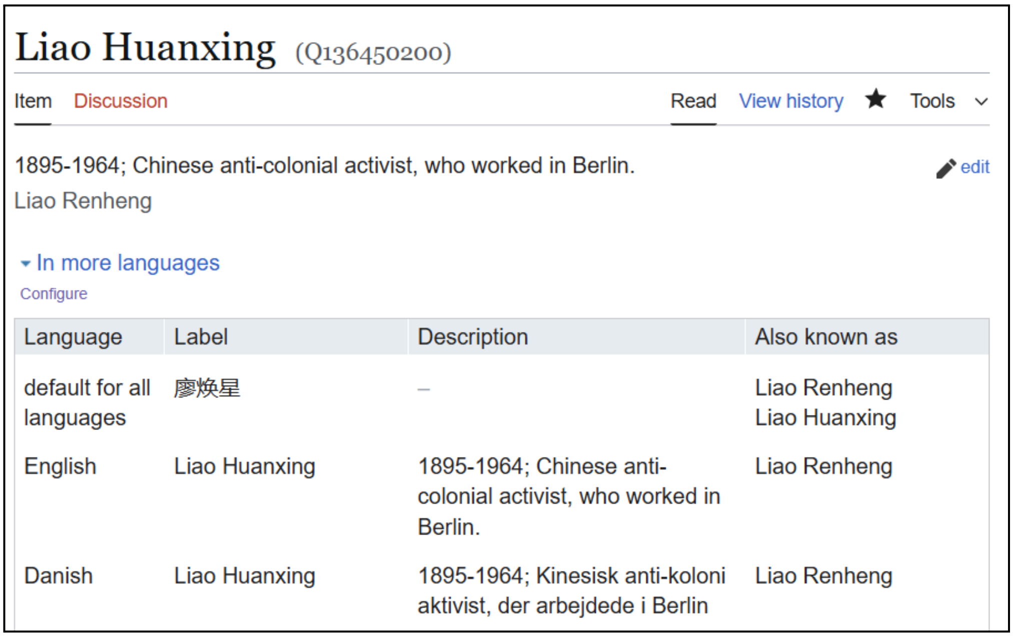
Methodology report: “Liao Huanxing” documentation in Wikidata

Clara Holst (CH), AU737540, 202306022@post.au.dk
5th Semester, Cognitive Science BSc, Arts Faculty, Aarhus University
Jens Chr. Skous Vej 2, 8000 Aarhus, Denmark

Critical Data Studies, Elective
Curating Data Course
Characters: 13.279 ~ 5,5 pages

Introduction and Context

This report details my work for Wikidata as seen in Fig. 1, which centred on the biography of Liao Huanxing (1895-1964), a Chinese anti-colonial activist whose life and work were documented by the Dekoloniale project. The primary aim was to transform the narrative, qualitative information from this marginalized history into structured, machine-readable data on Wikidata. This process served a dual purpose: to enhance the digital visibility of a figure often omitted from mainstream historical narratives and to engage practically with the principles and challenges of Linked Open Data (LOD). By contributing to an open knowledge graph, the task aimed to connect Liao Huanxing’s story to a wider web of historical data, thereby supporting the findability, accessibility, and reusability of decolonial knowledge resources.



Liao Huanxing (Q136450200)

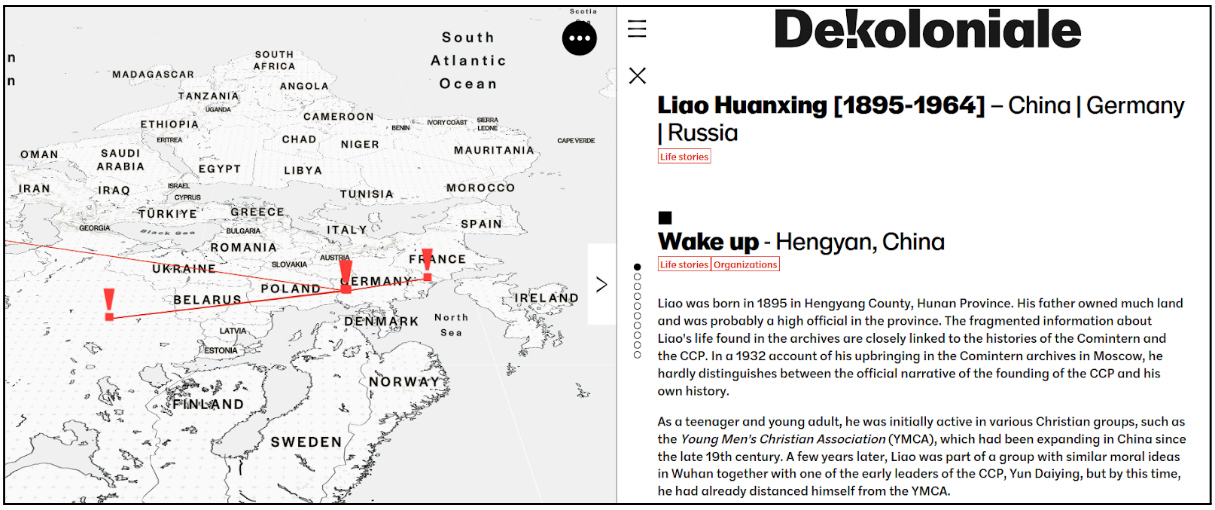
1895-1964; Chinese anti-colonial activist, who worked in Berlin.
Liao Renheng

Language	Label	Description	Also known as
default for all languages	廖焕星	–	Liao Renheng Liao Huanxing
English	Liao Huanxing	1895-1964; Chinese anti-colonial activist, who worked in Berlin.	Liao Renheng
Danish	Liao Huanxing	1895-1964; Kinesisk anti-koloni aktivist, der arbejdede i Berlin	Liao Renheng

Fig. 1: Liao Huanxing’s new page in Wikidata, Source: Wikimedia/Wikidata, CC BY-SA 4.0.

Personal Role and Responsibilities

My primary contribution was the extensive research and data enrichment for the Wikidata item for Liao Huanxing (Q136450200). Figure 1 shows the newly created item page. Initially, this item did not exist on any Wikimedia project. I was responsible for a comprehensive review of his biography on the Dekoloniale website (Fig. 2) to identify key entities and properties for structuring.



Dekoloniale

Liao Huanxing [1895-1964] – China | Germany | Russia

Life stories

Wake up - Hengyan, China

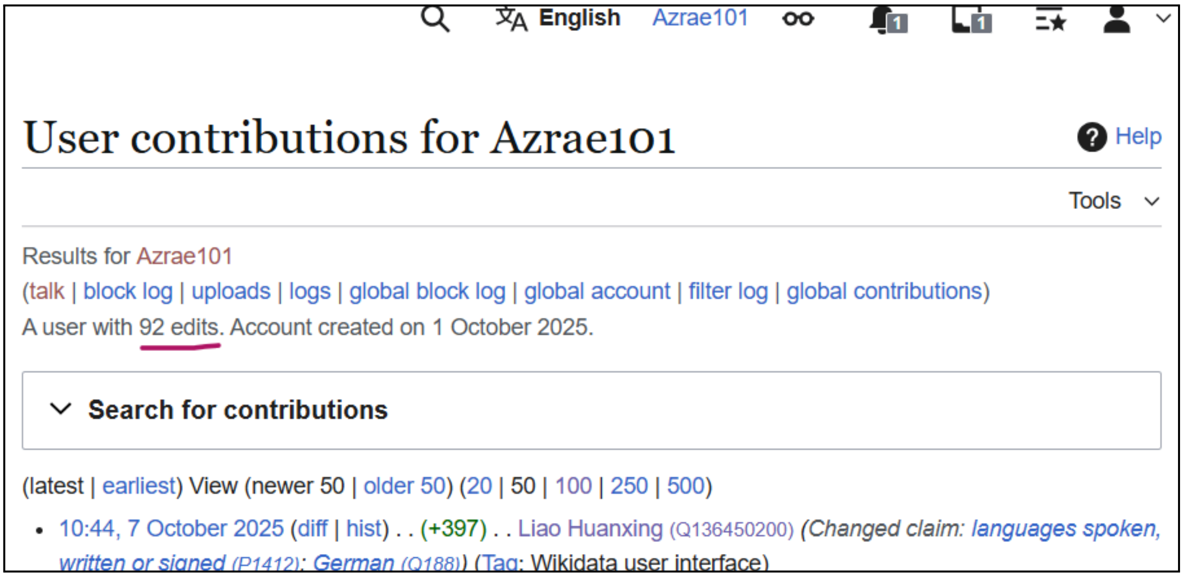
Life stories | Organizations

Liao was born in 1895 in Hengyang County, Hunan Province. His father owned much land and was probably a high official in the province. The fragmented information about Liao's life found in the archives are closely linked to the histories of the Comintern and the CCP. In a 1932 account of his upbringing in the Comintern archives in Moscow, he hardly distinguishes between the official narrative of the founding of the CCP and his own history.

As a teenager and young adult, he was initially active in various Christian groups, such as the *Young Men's Christian Association* (YMCA), which had been expanding in China since the late 19th century. A few years later, Liao was part of a group with similar moral ideas in Wuhan together with one of the early leaders of the CCP, Yun Daiying, but by this time, he had already distanced himself from the YMCA.

Fig. 2: Liao Huanxing on the Dekoloniale project website, Source: Dekoloniale.de, CC BY 4.0.

Then I executed the creation and population of numerous statements on his Wikidata item. My contributions, a sample of which is shown in Figure 3, included adding specific details such as his residences, education at Wuhan University, language skills, political affiliations, and alternative names. A critical part of my role involved meticulously adding references for each statement, linking the data back to its source to ensure verifiability and honour the Dekoloniale project as a key knowledge producer.



User contributions for Azrae101

Results for Azrae101

(talk | block log | uploads | logs | global block log | global account | filter log | global contributions)

A user with 92 edits. Account created on 1 October 2025.

Search for contributions

(latest | earliest) View (newer 50 | older 50) (20 | 50 | 100 | 250 | 500)

- 10:44, 7 October 2025 (diff | hist) . . (+397) . . Liao Huanxing (Q136450200) (Changed claim: languages spoken, written or signed (P1412): German (Q188)) (Tag: Wikidata user interface)

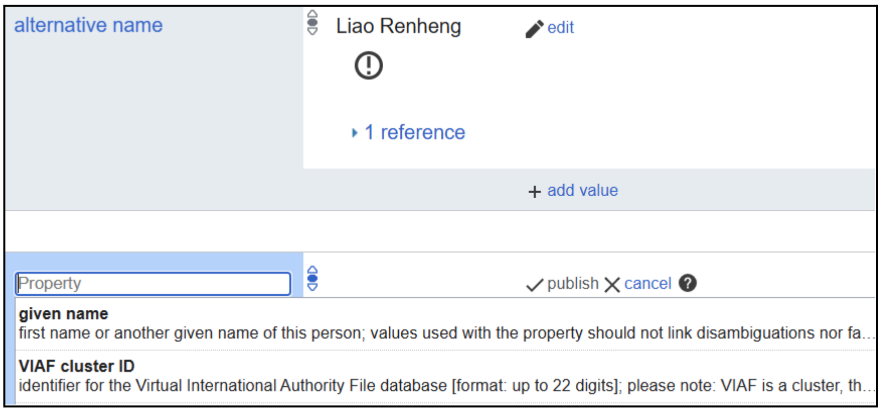
Fig. 3: Overview of my own contributions to Liao Huanxing's Wikidata item.
Source: Wikimedia/Wikidata, CC BY-SA 4.0.

Methodological Approach

My methodological approach began with a close reading of the Dekoloniale biography of Liao Huanxing to extract factual elements that could be represented in Wikidata. This step involved identifying key pieces of information such as places of residence, education, political affiliations, language skills, and significant relationships. The aim was to determine which elements could be translated into structured statements while maintaining as much historical and contextual accuracy as possible (Figure 2).

Understanding the structure of Wikidata was essential for this process. Wikidata is organized around items, each representing a unique entity such as a person, place, or organization. Each item is identified by a Q-number. These items are described using statements, which consist of a property, identified by a P-number, and a value, which can either be another item or a literal value such as a date or string of text. Each statement can also include qualifiers, which provide additional contextual information such as the start or end dates of an event, and references, which link the statement back to a source. Together, these statements form subject-predicate-object triples, the foundation of Linked Open Data. Figure 4 illustrates the process of adding a statement to Liao Huanxing's Wikidata item, showing the interface used to input properties, values, qualifiers, and references. This structured approach allows information to be interlinked across the knowledge graph and connected to other relevant items, increasing its findability, accessibility, and usability.

To maximize the interconnectivity of Liao Huanxing's data, I prioritized linking to existing items in Wikidata whenever possible. For example, rather than entering the university he attended as a free-text value, I linked it to the existing item for Wuhan University (Q461313) using the property "educated at" (P69). This approach ensured that Liao Huanxing's information became part of the broader semantic network, making it discoverable alongside other related historical figures and institutions. Temporal information, such as the dates of political party membership or residence in a specific city, was added using qualifiers to provide historical context. The careful addition of these statements is shown in Figure 4, while the resulting network of interlinked data can be visualized in Figure 5.



given name
first name or another given name of this person; values used with the property should not link disambiguations nor fa…

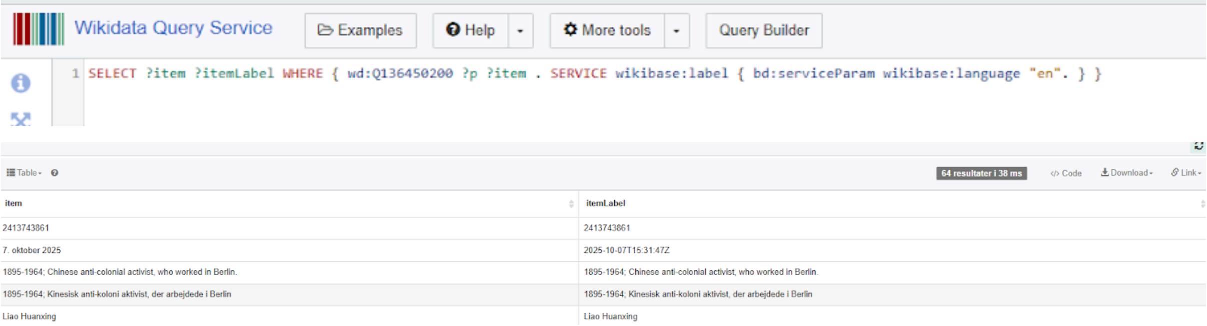
VIAF cluster ID
identifier for the Virtual International Authority File database [format: up to 22 digits]; please note: VIAF is a cluster, th…

Fig. 4: Adding a statement to Liao Huanxing's Wikidata. Source: Wikimedia/Wikidata, CC BY-SA 4.0.

To organize the extracted information efficiently, I developed a mental model that categorized data into person-centric entities. I separated statements into thematic groups such as education, occupations, relationships, and political affiliations. Each statement was added systematically, property by property, ensuring that all values were linked to the most accurate and relevant existing Wikidata items. This methodical approach allowed for careful management of complex information and ensured that the narrative from the Dekoloniale biography was preserved as much as possible in a structured format (Figure 1).

The process also involved iterative validation and refinement. As I added statements, I monitored warnings and suggestions from Wikidata's interface, which often indicated missing properties, potential duplicates, or formatting issues. Addressing these required careful interpretation of the source material and a balance between strict data modeling rules and the nuances of historical narrative. The interface feedback and validation process is visible in Figure 4, demonstrating how statements are checked for consistency and correctness.

Once statements were added, the resulting network of information could be visualized using the Wikidata Query Service. Figure 5 shows how Liao Huanxing's item is connected to other entities within the knowledge graph, highlighting the interconnectivity achieved through careful data modeling and linking. This visualization also confirms that qualifiers and references were applied consistently, ensuring both accuracy and discoverability.



```
SELECT ?item ?itemLabel WHERE { wd:Q136450200 ?p ?item . SERVICE wikibase:label { bd:serviceParam wikibase:language "en". } }
```

64 resultater i 38 ms

item	itemLabel
2413743861	2413743861
7. oktober 2025	2025-10-07T15:31:47Z
1895-1964; Chinese anti-colonial activist, who worked in Berlin.	1895-1964; Chinese anti-colonial activist, who worked in Berlin.
1895-1964; Kinesisk anti-koloni aktivist, der arbejdede i Berlin	1895-1964; Kinesisk anti-koloni aktivist, der arbejdede i Berlin
Liao Huanxing	Liao Huanxing

Fig. 5: Visualization of Liao Huanxing's data connections in the Wikidata Query Service (SPARQL query, https://w.wiki/FpKs). Source: Wikimedia/Wikidata, CC BY-SA 4.0.

Overall, this methodological approach allowed the transformation of a rich, narrative biography into structured, linked data that could be queried, analyzed, and connected to other knowledge resources. By carefully considering the properties, qualifiers, and references used, I ensured that the structured data maintained the integrity of the original biography while enhancing its discoverability and usability within the broader Wikidata ecosystem. Figures 2, 4, and 5 together illustrate the full process, from extracting information from the Dekoloniale biography to modeling it in Wikidata and visualizing its connections in the knowledge graph. This approach demonstrates the potential of Linked Open Data to make marginalized historical figures visible and connected across multiple digital platforms.

Challenges and Considerations

The process was not without significant challenges. A primary difficulty was the inherent tension between narrative nuance and the rigid structure of a database. The biography conveyed the significance of Liao's activism through story, whereas Wikidata required its reduction to discrete, pre-defined properties. As seen in Figure 5, Wikidata's interface often showed warnings when it did not recognize new information, a process that often felt reductive, stripping away the contextual meaning of his political work. Furthermore, ambiguities in the source text, such as imprecise dates, posed a problem. For example, stating a residence in Berlin without exact years required a decision on how to represent this temporally, balancing accuracy with the information available.

Ethically, I was highly conscious of the power dynamics inherent in categorization, as discussed by Bowker and Star (2000). Assigning labels like "activist" and selecting his ethnic group as "Han Chinese" are not neutral acts; they are interpretive choices that frame his identity in specific ways. Ensuring that every claim was rigorously sourced from the Dekoloniale project was my way of anchoring these categorizations in the original, community-driven narrative, thereby mitigating the risk of misrepresentation.

Critical Reflection

Reflecting on this process through the lens of our course readings reveals profound insights into curation as knowledge production. The notion that "raw data is an oxymoron" (Gitelman and Jackson, 2013, p. 2), was vividly demonstrated. The data I created on Wikidata was anything but raw; it was a product of my interpretation, the constraints of the platform's ontology, and the specific perspective of the Dekoloniale source. This aligns with Kitchin's (2022) concept of "data assemblages," where data is shaped by a complex sociotechnical system: in this case, comprising the Dekoloniale editors, the Wikidata platform, its community norms, and myself as a curator. The work of Bowker and Star (2000) on the politics of classification was also ever-present.

By fitting Liao Huanxing’s life into Wikidata’s property schema, I participated in making his history visible within a particular standardized system, one that, as Ford and Illadis (2023) note, carries its own Western and structural biases. The outcome is a form of knowledge that is highly accessible and linkable but also flattened. While the process successfully makes a marginalized history more visible to algorithms and automated systems, it simultaneously risks decontextualizing it, separating the factual "what" from the narrative "why" that gives the facts their deeper meaning and political power. Structured data should thus be understood as a supplement to, not a replacement for, rich narrative sources (Dekoloniale, Figure 2).

Conclusion

In conclusion, this exercise provided invaluable lessons about the potential and limitations of Linked Open Data for recuperating marginalized histories. I learned that making such histories visible requires more than just uploading facts; it demands a careful, ethical, and critically aware approach to data modeling. The power of LOD lies in its ability to connect disparate pieces of information, potentially placing a figure like Liao Huanxing on the same digital map as more widely known historical actors.

However, this comes at the cost of narrative richness. The key lesson is that structured data projects like this are most powerful when understood as supplements to, not replacements for, detailed narrative sources like the Dekoloniale biographies. They serve as vital indexes and entry points, guiding users to the richer, contextualized knowledge held within the original resources.

Ultimately, the work of digital curation is a continuous negotiation between the logic of the database and the complexity of human experience, a practice that is fundamentally about making conscious and responsible choices in the production of knowledge.

References

- Bowker, G. C., & Star, S. L. (2000). Why classifications matter. In Sorting things out: Classification and its consequences (pp. 319-326). MIT Press.
- boyd, d., & Crawford, K. (2012). Critical questions for Big Data. Information, Communication & Society, 15(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878
- Ford, H., & Illadis, A. (2023). Wikidata as semantic infrastructure: Knowledge representation, data labor, and truth in a more-than-technical project. *Social Media + Society, 9*(3). https://doi.org/10.1177/20563051231195552
- Gitelman, L., & Jackson, V. (2013). Introduction. In "Raw Data" Is an Oxymoron (pp. 1–12). MIT Press.
- Illadis, A., & Russo, F. (2016). Critical Data Studies: An introduction. Big Data & Society, 3(2). https://doi.org/10.1177/2053951716674238
- Kitchin, R. (2022). Critical Data Studies. In The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences (pp. 21–41). SAGE Publications Ltd.
- Kitchin, R. (2022). Introducing Data. In The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences (pp. 1–19). SAGE Publications Ltd.
- Kitchin, R. (2022). Small Data and Data Infrastructures. In The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences (pp. 44–59). SAGE Publications Ltd
- Wikidata. (2025). *Liao Huanxing (Q136450200)*. Wikimedia Foundation. https://www.wikidata.org/wiki/Q136450200 (CC BY-SA 4.0)
- Dekoloniale. (n.d.). *Liao Huanxing biography*. Dekoloniale Memory Culture in the City. https://dekoloniale.de/ (CC BY 4.0)
- Wikidata Query Service. (2025). *SPARQL visualization for Liao Huanxing (Q136450200)*. Wikimedia Foundation. https://w.wiki/FpKs (CC BY-SA 4.0)
- Wikimedia Commons. (n.d.). *Screenshots and interface elements from Wikidata*. Wikimedia Foundation. https://commons.wikimedia.org/ (CC BY-SA 4.0)

Visualizing Consumer Psychology

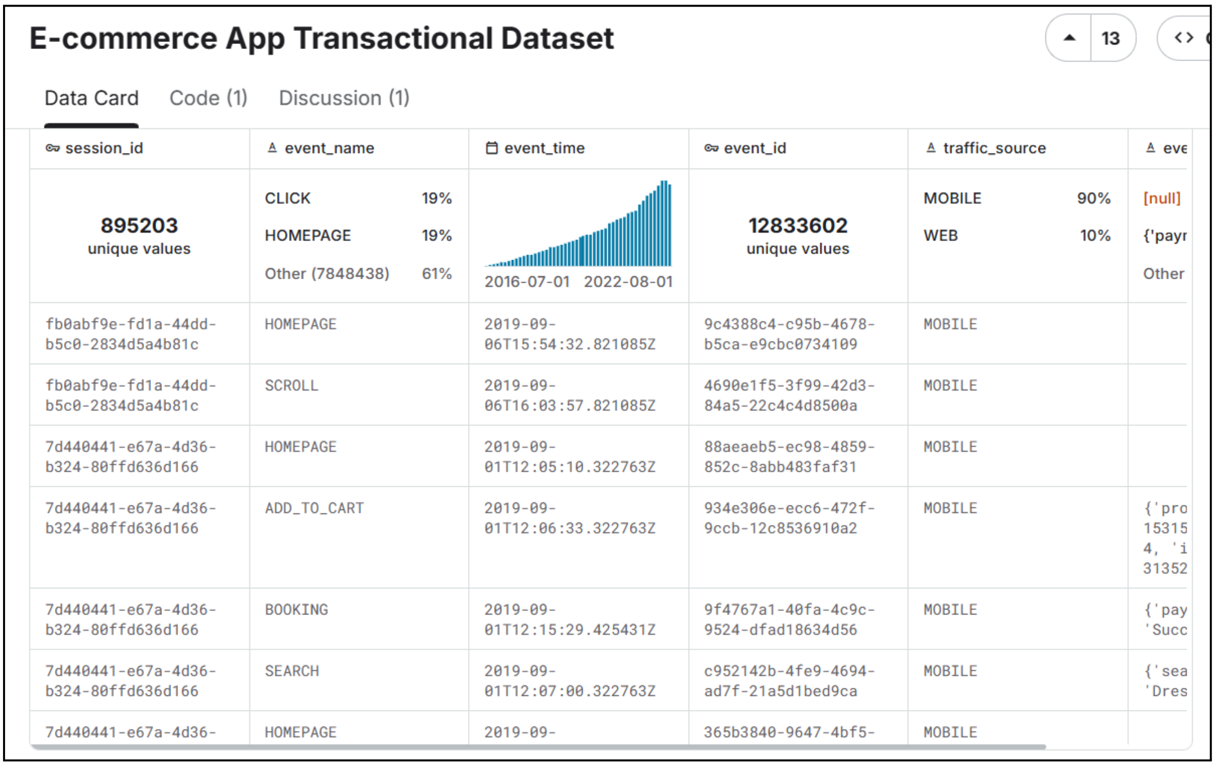
A Critical Analysis of E-commerce Behavioral Data

Clara Holst (CH), AU737540, 202306022@post.au.dk
5th Semester, Cognitive Science BSc, Arts Faculty, Aarhus University
Jens Chr. Skous Vej 2, 8000 Aarhus, Denmark
Critical Data Studies, Elective
Curating Data Course

Characters: 14.399~ 6 pages

Introduction

This report analyzes 1.25 million e-commerce transactions from a Kaggle dataset, shown in Figure 1, through two distinct data visualization approaches: a traditional visual analytics method and an interactive feminist/critical data studies perspective. The goal is to explore how these different methods reveal or suppress understandings of consumer psychology and price perception, aligning with the course objective to practice data visualization from multiple viewpoints.



session_id	event_name	event_time	event_id	traffic_source	eve
895203 unique values	CLICK 19% HOMEPAGE 19% Other (7848438) 61%	2016-07-01 2022-08-01	12833602 unique values	MOBILE 90% WEB 10%	[null] {'payr Other
fb0abf9e-fd1a-44dd-b5c0-2834d5a4b81c	HOMEPAGE	2019-09-06T15:54:32.821085Z	9c4388c4-c95b-4678-b5ca-e9cbc0734109	MOBILE	
fb0abf9e-fd1a-44dd-b5c0-2834d5a4b81c	SCROLL	2019-09-06T16:03:57.821085Z	4690e1f5-3f99-42d3-84a5-22c4c4d8500a	MOBILE	
7d440441-e67a-4d36-b324-80ffd636d166	HOMEPAGE	2019-09-01T12:05:10.322763Z	88aeaeb5-ec98-4859-852c-8abb483faf31	MOBILE	
7d440441-e67a-4d36-b324-80ffd636d166	ADD_TO_CART	2019-09-01T12:06:33.322763Z	934e306e-ecc6-472f-9ccb-12c8536910a2	MOBILE	{'pro 15315 4, 'i 31352
7d440441-e67a-4d36-b324-80ffd636d166	BOOKING	2019-09-01T12:15:29.425431Z	9f4767a1-40fa-4c9c-9524-dfad18634d56	MOBILE	{'pay 'Succ
7d440441-e67a-4d36-b324-80ffd636d166	SEARCH	2019-09-01T12:07:00.322763Z	c952142b-4fe9-4694-ad7f-21a5d1bed9ca	MOBILE	{'sea 'Dres
7d440441-e67a-4d36-	HOMEPAGE	2019-09-	365b3840-9647-4bf5-	MOBILE	

Figure 1. Screenshot of the E-commerce App Transactional Dataset on Kaggle (Pratama, 2023), showing an overview of customer, product, transaction, and click-stream tables, which capture user behavior and transaction data from a fashion e-commerce app.

Visual Analytics Approach

The first visualization is a scatter plot mapping perceived value against actual product prices, with customers divided into nine behavioral segments (e.g., Value-Optimistic, Budget-Pessimistic). This follows Cui's (2019) definition of visual analytics as "the science of analytical reasoning facilitated by interactive visual interfaces."

To clarify the analytical framework, the behavioral segments were constructed by cross-tabulating two dimensions. First, customers were categorized into three spending tiers based on their transaction history: "Budget" (lowest average spend), "Value" (mid-range), and "Premium" (highest average spend). Second, their "perceived value" rating for products was classified as "Optimistic" (perceived value > 110% of actual price), "Realistic" (90-110%), or "Pessimistic" (<90%). This creates the nine segments, such as a "Budget-Optimistic" shopper who spends little but tends to overvalue products.

The data pipeline, implemented in Python, began by merging three primary datasets: transactional records, customer demographics, and product information. A key technical challenge was extracting structured data from the JSON-like product_metadata field, which was parsed using Python's ast.literal_eval() function to isolate individual product IDs and prices for analysis. This step was crucial for accurately linking transaction amounts to specific products and their categories.

The visualization itself was generated using Plotly to create an interactive, multi-faceted scatter plot. The code employs a stratified sampling method (sample_data = merged.sample(min(3000, len(merged)), random_state=42)) to ensure visual clarity while maintaining statistical representativeness. Each subplot corresponds to a specific age-gender demographic, with the facet_col and facet_row parameters enabling direct comparison across segments. The visual encoding uses color for behavioral categories, point size for transaction amount, and includes both a perfect perception line (y=x) and ±10% confidence bands to immediately contextualize value estimation accuracy.

The statistical results also provide a quantitative backbone for these visual patterns, revealing a consumer landscape dominated by mid-range, value-conscious shoppers. The near-identical sizes of the three Value segments (Value-Optimistic: 16.8%, Value-Realistic: 16.8%, Value-Pessimistic: 16.8%) are particularly striking. This symmetrical distribution within the Value tier, which collectively represents 50.4% of all customers, suggests that while consumers cluster in the mid-price range, their psychological orientation toward value is evenly split between optimism, realism, and pessimism. This indicates that spending level and perception bias operate as independent psychological dimensions within the marketplace.

First and foremost, Figure 2 reveals interesting systematic patterns. Perceived value aligns closely with actual price at lower levels, clustering around the “perfect perception” line, but becomes more scattered as prices rise. This indicates greater uncertainty in how high-value products are perceived, possibly due to reduced price familiarity or aspirational tendencies among consumers.

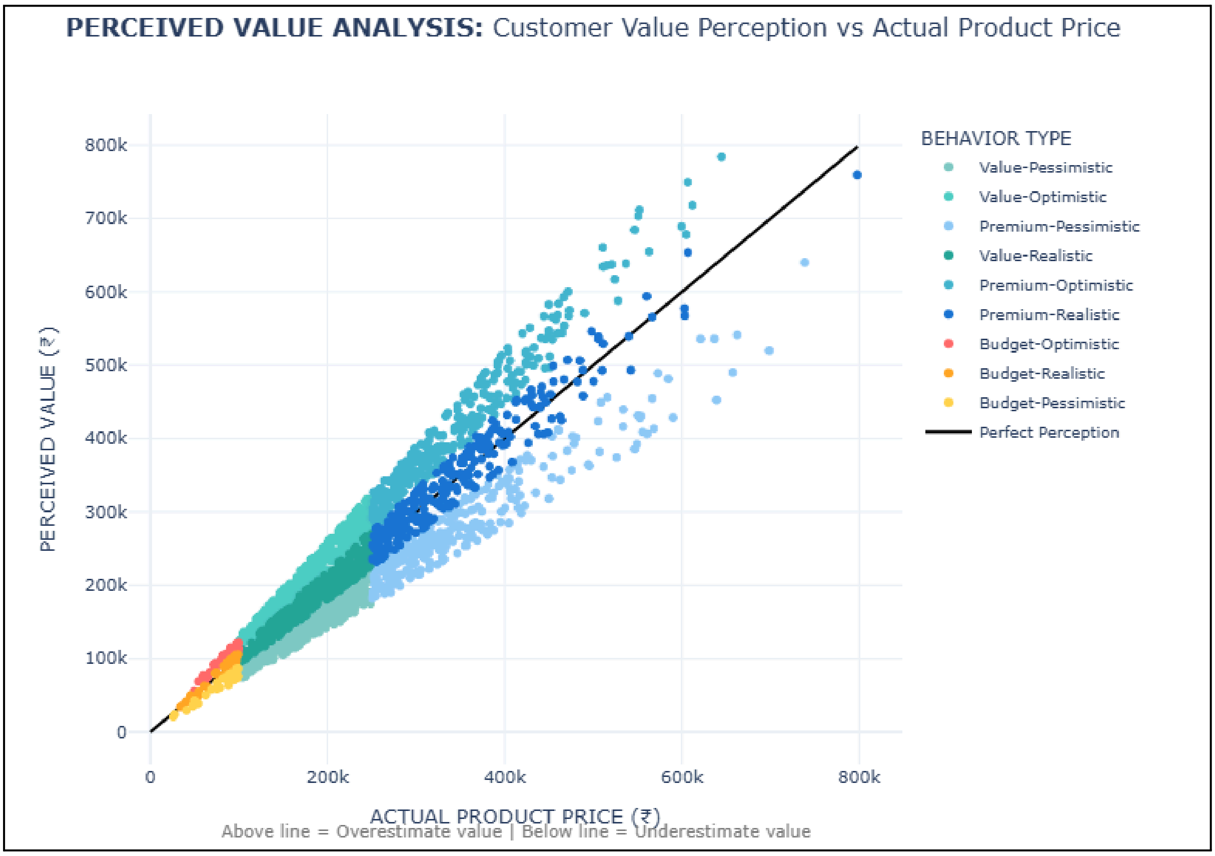


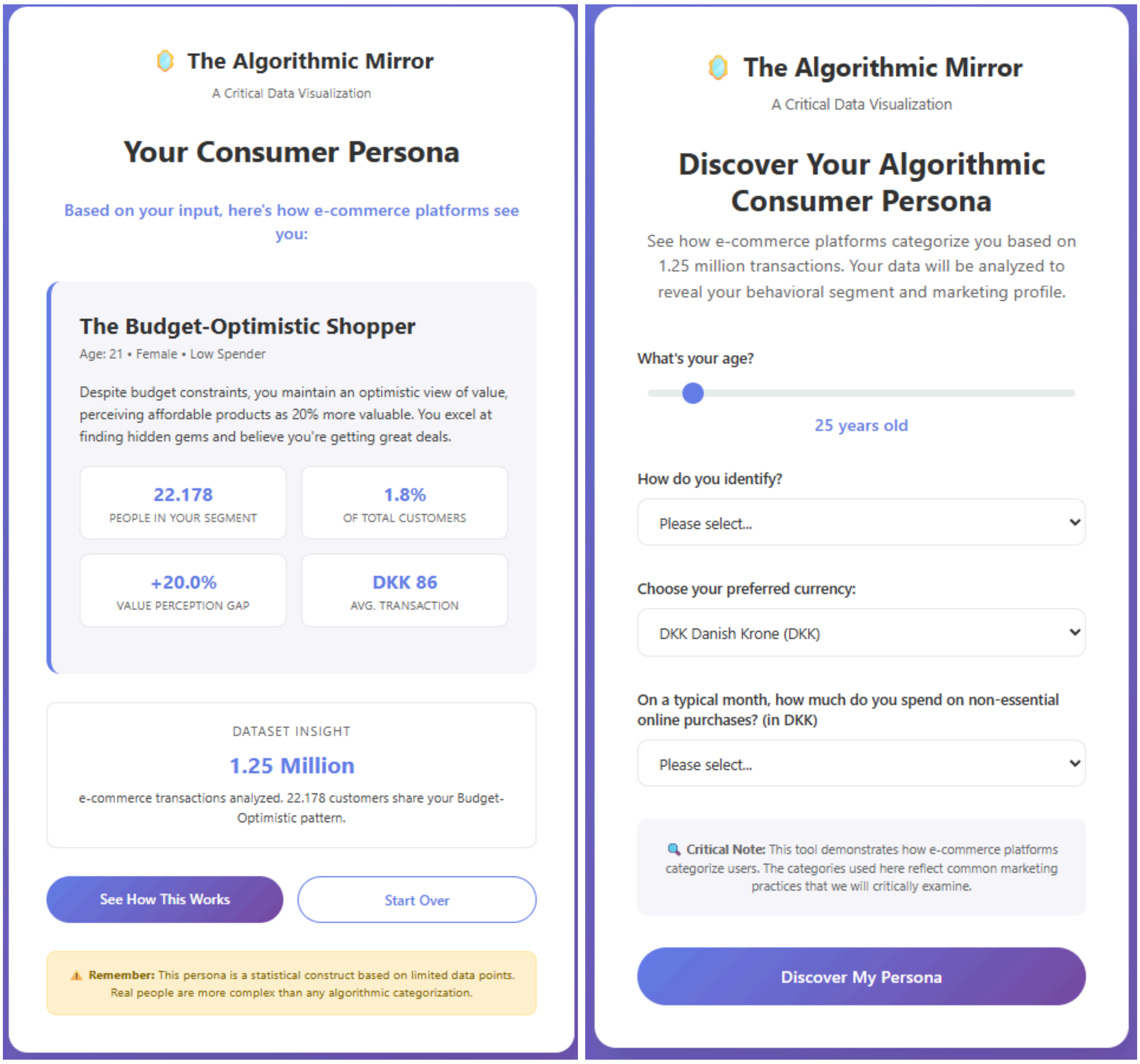
Figure 2: Scatter plot comparing actual product prices (x-axis) against perceived value (y-axis) across 1.25 million transactions. Points colored by 9 behavioral segments (Budget/Value/Premium × Optimistic/Realistic/Pessimistic) and sized by transaction amount (₹100K-₹1M range). Black diagonal indicates perfect perception; 50.4% of customers fall in Value segments, with perception gaps ranging from -20% to +20% across behaviors.

Despite its analytical clarity, this approach has key limitations. It privileges quantitative precision and abstracts complex psychological processes into data points (D’Ignazio & Klein, 2020). The “perfect perception” line imposes a normative standard that equates accuracy with value, reinforcing capitalist ideals and marginalizing alternative consumer perspectives.

Furthermore, the statistical aggregation masks individual stories and cultural contexts. The near-perfect balance between optimistic and pessimistic segments (16.8% vs 16.8% in the Value tier) might suggest market equilibrium, but this mathematical symmetry erases the lived experiences of consumers whose value perceptions may be perfectly rational within their specific economic or cultural contexts. The dataset's geographic origin in Indonesia is particularly relevant here, as the price points and perception patterns must be understood within local economic conditions rather than universal psychological laws.

Critical Data Studies Perspective

I developed "The Algorithmic Mirror" website to visceralize datafication through interactive engagement (D'Ignazio & Klein, 2020) as seen in Figure 3. The website invites users to input demographic information and spending habits, mirroring e-commerce platform data collection practices. In return, users receive a personalized consumer persona that categorizes them within the same behavioral segments analyzed in the original dataset.



The Algorithmic Mirror

A Critical Data Visualization

Your Consumer Persona

Based on your input, here's how e-commerce platforms see you:

The Budget-Optimistic Shopper

Age: 21 • Female • Low Spender

Despite budget constraints, you maintain an optimistic view of value, perceiving affordable products as 20% more valuable. You excel at finding hidden gems and believe you're getting great deals.

22.178 PEOPLE IN YOUR SEGMENT	1.8% OF TOTAL CUSTOMERS
+20.0% VALUE PERCEPTION GAP	DKK 86 AVG. TRANSACTION

DATASET INSIGHT

1.25 Million

e-commerce transactions analyzed. 22.178 customers share your Budget-Optimistic pattern.

⚠ **Remember:** This persona is a statistical construct based on limited data points. Real people are more complex than any algorithmic categorization.

The Algorithmic Mirror

A Critical Data Visualization

Discover Your Algorithmic Consumer Persona

See how e-commerce platforms categorize you based on 1.25 million transactions. Your data will be analyzed to reveal your behavioral segment and marketing profile.

What's your age?

25 years old

How do you identify?

Please select...

Choose your preferred currency:

DKK Danish Krone (DKK)

On a typical month, how much do you spend on non-essential online purchases? (in DKK)

Please select...

🔍 **Critical Note:** This tool demonstrates how e-commerce platforms categorize users. The categories used here reflect common marketing practices that we will critically examine.

Figure 3: Interactive website where users input age, gender, and spending habits to generate their algorithmic consumer persona. The interface aims to mirror e-commerce data collection practices. Access it here: https://azrae101.github.io/Assignment3_CDS/index.html

The website's architecture was designed to operationalize critical data studies principles through three key methodological choices. First, it employs the same categorization logic used in the visual analytics approach; classifying users into Budget/Value/Premium spending tiers and Optimistic/Realistic/Pessimistic perception segments, but applies this algorithm to user-inputted data rather than historical transactions.

This creates a direct bridge between the objective analysis and personal experience, allowing users to encounter the same classificatory systems that e-commerce platforms use to profile consumers.

Second, to foreground the geopolitical situatedness of data, the website incorporates a multi-currency model that contextualizes the original dataset's Indian price points (in Indian Rupees) against comparative fashion expenditure data from Denmark (DKK), the Eurozone (EUR), and the United States (USD). This design choice, drawing on Statistics Denmark (2025), NielsenIQ (2025), and Bureau of Labor Statistics (2025) data, makes visible how purchasing power and consumption norms vary dramatically across geographic and economic contexts; a dimension often erased in universalizing behavioral analyses.

Third, the interface deliberately mirrors the data collection practices of commercial platforms, requesting age, gender, and spending habits while offering only binary gender options. This intentional limitation serves as a critical design feature rather than an oversight, demonstrating what D'Ignazio & Klein (2020) term the "tyranny of categories" by making users experience firsthand how complex identities are flattened into marketable segments. The subsequent "Critical Analysis" section explicitly deconstructs this process, acknowledging the dataset's erasure of non-binary identities and the cultural specificity of its Indonesian origin.

This interactive approach demonstrates what Seaver (2022) terms the "semantic gap" between lived experience and data representation. When users encounter their algorithmic persona, they experience the dissonance between their self-perception and their data-double. The website's self-reflexive design includes critical disclaimers and an analysis section that exposes its own mechanisms, acknowledging that personas are statistical constructs based on limited data points.

The website extends beyond mere demonstration to create what D'Ignazio & Klein (2020) call "visceral engagement" with datafication. Unlike the analytical distance of the scatter plot, the moment of recognition: or more often, misrecognition, when users see their assigned persona transforms abstract data patterns into personal experience. This emotional response is pedagogically crucial: it reveals how algorithmic categorization feels from the subject position, making palpable the power dynamics embedded in commercial data systems.

The critical framework embedded in the website explicitly addresses the ethical feedback loops identified by Milano et al. (2020), where algorithmic categorization leads to targeted marketing that shapes behavior, which in turn reinforces the original categorization. By walking users through this cycle and providing comparative demographic data, "The Algorithmic Mirror" enables what Seaver (2022) might call "infrastructural recognition" - understanding not just what the algorithm says, but how it operates within broader economic and social systems that convert identity into capital.

Reflection on Design and Viewer Experience

The design choices in both visualizations directly shape viewer understanding through different epistemological frameworks. The scatter plot creates analytical distance, positioning viewers as objective observers of consumer behavior patterns. This traditional approach reveals systematic price perception trends but risks dehumanizing consumers by reducing them to data points.

Conversely, "The Algorithmic Mirror" positions users as both subject and analyst of their own datafication. This creates what D'Ignazio & Klein (2020) call "visceral engagement," transforming abstract patterns into personal experience. However, this approach carries its own risks: by employing the same categorical systems used by e-commerce platforms, it may inadvertently reinforce the reductive logic it seeks to critique.

A key difference lies in how they handle demographic data. The scatter plot, while colored by behavior, abstracts away individual demographics to reveal macro-trends. The interactive website, in contrast, explicitly asks for age and gender to generate a persona, mirroring the profiling practices of commercial platforms. This direct solicitation forces the user to confront how these attributes are used as proxies for behavior, a point critically examined in the website's "How It Works" section.

Both methods grapple with the fundamental challenge identified by Drucker (2011): data requires arrangement, and every arrangement privileges certain perspectives while suppressing others. The simulated "perceived value" metric in both approaches highlights how consumer psychology is often inferred rather than measured, demonstrating Seaver's (2022) concept of engineers attempting to "close the semantic gap" between human experience and computational representation.

The importance of geographical context is deeply embedded in the dataset's Indian origin (as detailed on the website). The price points and consumption patterns are specific to a particular socio-economic environment. For instance, the prominence of "Value" segments and the specific price ranges (in Indian Rupees) reflect the priorities and purchasing power of this specific market. A visualization for a Scandinavian or North American audience would likely reveal a different distribution of behaviors, shaped by distinct cultural and economic factors. This underscores how data, value, and consumer priorities are not universal but are profoundly shaped by their geographic and socio-economic context.

Conclusion

These contrasting visualization approaches demonstrate that data representation is never neutral. The visual analytics method reveals systematic patterns in price perception but reduces human complexity to quantifiable metrics. The interactive critical approach fosters personal engagement with datafication processes but must constantly question its own methodological foundations.

The consistent value misperception patterns across all segments suggest universal psychological tendencies that both enable and complicate demographic targeting. These visualizations reveal not only how consumers perceive prices but also how data systems reconstruct human identity as analyzable patterns, raising ethical considerations for consumer autonomy and fairness (Milano et al., 2020).

Furthermore, the website practices what the feminist data studies framework advocates: using data to challenge power by making the invisible processes of datafication visible and emotionally resonant. The self-reflexive design: which constantly questions its own categorical assumptions and methodological limitations, models an alternative approach to data visualization that rejects the "view from nowhere" in favor of situated, partial knowledge that acknowledges its own constructed nature.

Ultimately, the project shows that a single dataset can tell multiple stories. The scatter plot tells a story of macroeconomic trends and psychological biases, while the interactive website tells a story of algorithmic identity and personal reflection. The geographical and cultural specificity of the data further reminds us that these stories are situated and partial.

By employing multiple visualization methods, we can better understand both consumer behavior patterns and the power dynamics embedded in their representation. The most responsible approach to data analysis acknowledges its limitations and remembers that behind every data point lies a human story that exceeds algorithmic capture.

Website creation: Curating Data / Making Sense of Knowledge

Clara Holst (CH), AU737540, 202306022@post.au.dk
5th Semester, Cognitive Science BSc, Arts Faculty, Aarhus University
Jens Chr. Skous Vej 2, 8000 Aarhus, Denmark

Critical Data Studies, Elective
Curating Data Course
Characters: 429.428 (code and text)
Github code repository:
https://github.com/Azrae101/CuratingDataScience

Website:
https://azrae101.github.io/CuratingDataScience/index.html

Abstract

Curating Data: Making Sense of Knowledge is a digital exhibition that brings together the site's key sections, including Home, Curatorial Statement, Assignments, Workshops, Notes, Games, References, and About. Created by Cognitive Science student Clara Holst, the exhibition presents a semester-long investigation into how data becomes knowledge through practices of collecting, categorizing, visualizing, and archiving. The Curatorial Statement introduces the theoretical foundations that guide the project, including critical data studies, data assemblages, the DIKW hierarchy, and posthuman approaches to curation.

The Assignments, Workshops, and Notes pages showcase practical and analytical work that transforms physical objects, historical narratives, and large datasets into structured and interpretable forms. Interactive elements in the Games section support learning through flashcards, quizzes, and matching activities, while the References page provides the academic and data sources that anchor the exhibition.

Across all pages, the project highlights the ethical, interpretive, and cognitive dimensions of data curation and demonstrates how curatorial decisions shape meaning, structure experience, and influence the ways knowledge is created and encountered.