

“Garbage In, Garbage Out” Revisited: Labeling and Dataification Practices Across Disciplines

R. Stuart Geiger, Ph.D
Assistant Professor

University of California, San Diego
Department of Communication
Halicioğlu Data Science Institute

[**stuart@stuartgeiger.com**](mailto:stuart@stuartgeiger.com) | see papers at [**stuartgeiger.com**](https://stuartgeiger.com)
CIC 2022, Virtual | slides at [**tinyurl.com/GIGOCIC**](https://tinyurl.com/GIGOCIC)

Who am I? An interdisciplinary nomad concerned with data and its discontents, in many contexts

- **Three degrees from weird “make your own degree” programs:**
 - B.A. Humanities (Univ of Texas at Austin)
 - M.A. Communication, Culture, and Technology (Georgetown Univ)
 - Ph.D Information (UC-Berkeley School of Information)
- **Core training: history and philosophy of science and technology; anthropology; information science; natural language processing**
- **Ph.D on the social-technical construction of knowledge in Wikipedia**
- **Four years as “staff ethnographer” at UC Berkeley Institute for Data Science, studying various “tribes” of data scientists and developers**
- **Now joint faculty in Communication and Data Science at UC San Diego, teaching required graduate “Data Ethics and Society” class**

A universal concern: dataification

- A concept from Science & Technology Studies (STS), an interdisciplinary field that takes sci & tech as its object of research (see Cukier and Mayer-Schonberger 2013)
- How is the world, in all its complexity & richness, reduced to data?
- Measurement methods and data labeling are crucial, but are seen as 'boring' and not worth discussing --- just give us results/AUC!
- Students love data science as a degree because they can work across many domains, disciplines, contexts, application areas, etc.
 - But they mostly re-use data collected by others, with little understanding of how data was collected and its limitations

Data: raw, cleaned, and cooked



Dataification: the map is not the territory

"All models are wrong, but some are useful" - George Box

"On Inexactitude in Science" by Jorge Luis Borges:

"...In that Empire, the Art of Cartography attained such Perfection that the map of a single Province occupied the entirety of a City, and the map of the Empire, the entirety of a Province. In time, those Unconscionable Maps no longer satisfied, and the Cartographers Guild struck a Map of the Empire whose size was that of the Empire, and which coincided point for point [...]

In the Deserts of the West, still today, there are Tattered Ruins of that Map, inhabited by Animals and Beggars; in all the Land there is no other Relic of the Disciplines of Geography"

The world is changing, but how we dataify the world is changing faster --- and with little documentation

High-stakes example: Predictive policing, using historical crime data to model future crime: where, when, and (even more controversially) by whom

Ethical and methodological issues in using past generations' outcomes to predict the current generation's futures, especially in predicting recidivism for bail decisions

"Crime data" is not a full or representative sample of all crime; many biases:

- Institutional biases: Where has the police department sent officers to patrol the most? Are certain kinds of kinds of crimes or activities deemed to be a higher priority than others? Ex: drugs vs auto theft
- Individual biases: When officers are on patrol in a neighborhood, which officers focus on certain kinds of "suspicious" people or activities more? Which residents report crime, and which reports are investigated?
- In San Diego, *Shotspotter* microphone-based gunshot detection system is only in 4 low-income minority neighborhoods; >20% false positive rate; 'activations' send officers racing to a neighborhood

Our Garbage In, Garbage Out Revisited Project

“On two occasions I have been asked, "...Mr. Babbage, if you put into the machine wrong figures, will the right answers come out?" ... I am not able rightly to apprehend the kind of confusion of ideas that could provoke such a question.”

— Charles Babbage (*1864*)

Our Garbage In, Garbage Out Revisited Project

“On two occasions I have been asked, "...Mr. Babbage, if you put into the machine wrong figures, will the right answers come out?" ... I am not able rightly to apprehend the kind of confusion of ideas that could provoke such a question.”

— Charles Babbage (1864)

Causes of death in London, 1660:

- People used to die of grief, lunacy, planet
- These were the most empirical and data-driven proto-scientists of their day
- How will future generations look at us?

1660.

A General BILL for this present Year,

Ending the 11th Day of December 1660.

According to the Report made to the King's most excellent Majesty,
By the Company of Parish Clerks of LONDON, &c.

DISEASES and CASUALTIES.

| | | | | | |
|---------------------------------|------|--|------|---|-----|
| A Bortive and Stillborn | 421 | Flox and Small Pox | 1523 | Palfy | 17 |
| Aged | 909 | Found dead in the Streets, Fields, &c. | 2 | Plague | 36 |
| Ague and Fever | 2303 | French Pox | 51 | Plurify | 12 |
| Apoplexy and Suddenly | 91 | Gout | 4 | Quinly and fore Throat | 21 |
| Blasted and Planet | 3 | Grief | 13 | Rickets | 441 |
| Bleeding and bloody Issue | 7 | Gripping in the Guts | 253 | Rifing of the Lights | 210 |
| Bloody Flux, Scowring, and Flux | 346 | Hanged and made away themselves | 11 | Rupture | 12 |
| Burnt and Scalded | 6 | Head-ach and Headmouldshot | 35 | Scurvy | 82 |
| Cancer, Gangrene and Fistula | 63 | Jaundies | 102 | Shot | 7 |
| Canker, fore Mouth and Thruh | 73 | Impofthume | 105 | Shingles | 1 |
| Childbed | 226 | Killed by feveral Accidents | 55 | Sores, Ulcers, broken and bruifed Limbs | 61 |
| Chrifomes and Infants | 858 | King's Evil | 28 | Spleen | 7 |
| Cold, Cough and Hiccough | 33 | Lethargy | 6 | Spotted Fever and Purples | 368 |
| Colick and Wind | 116 | Livergrown | 8 | Starved | 7 |
| Consumption and Tiffick | 2982 | Lunatick and Frenzy | 14 | Strangury | 22 |
| Convulſion | 742 | Megrims | 5 | Stopping of the Stomach | 186 |
| Cut of the Stone and Stone | 46 | Meafles | 6 | Surfeit | 202 |
| Droppy and Tympany | 646 | Mother | 1 | Swine Pox | 2 |
| Drowned | 57 | Murthered | 7 | Teeth and Worms | 839 |
| Executed | 7 | Overlaid and Starved at Nurfe | 46 | Vomiting | 8 |
| Falling Sicknefs | 4 | | | Wen | 1 |

Our Garbage In, Garbage Out Revisited Project

Two meta-research studies about how applied ML papers discuss their training data and data labeling practices (if at all), including supplemental materials:

Study of NLP-based ML application papers using data from Twitter:

Geiger, R. S., Yu, K., Yang, Y., Dai, M., Qiu, J., Tang, R., & Huang, J. (2020). “Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from?” In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 325-336).

Study of ML application papers across disciplines:

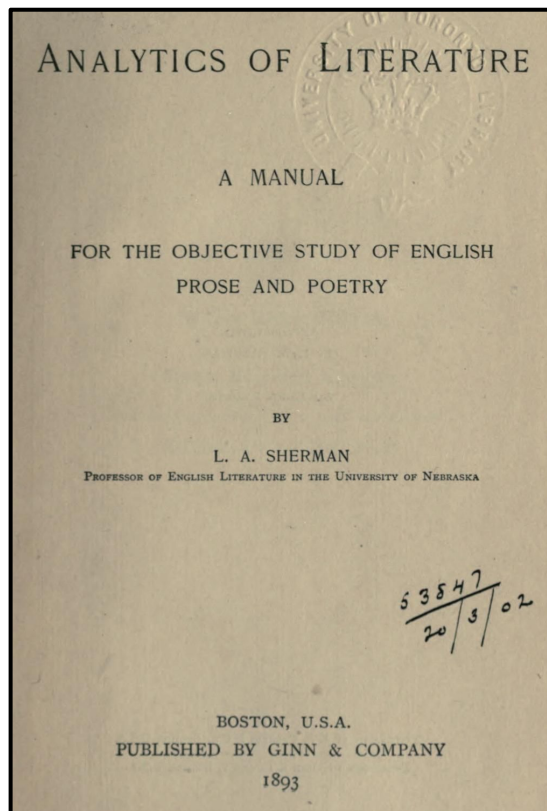
Geiger, R. S., Cope, D., Ip, J., Lotosh, M., Shah, A., Weng, J., & Tang, R. (2021). “‘Garbage in, garbage out’ revisited: What do machine learning application papers report about human-labeled training data?” *Quantitative Science Studies*, 2(3), 795-827.

The GIGO papers in four bullet points:

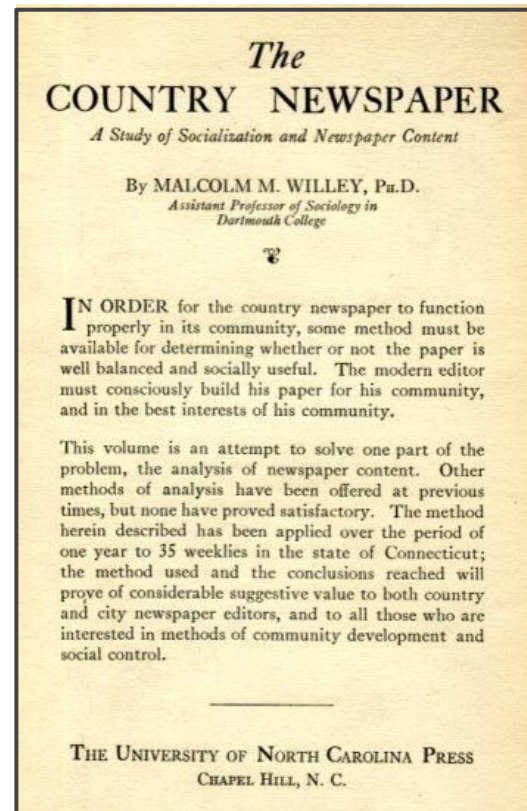
- Many of the ethical issues that arise in machine learning applications can be traced back to the quality of training data.
- The way training data is labeled by humans is often a form of structured content analysis, which has established best practices in the social sciences and humanities (also in life & ecosciences)
- RQ: How many applied ML papers report following best practices?
- A: Few, and varies substantially, showing need for more focus on data labeling practices in ML education, evaluation, and regulation.

Data labeling: structured content analysis

An established quantitative method in the humanities and social sciences, used by generations of researchers.



Sherman (1893)



Willey (1923)

Structured content analysis best practices:

“a systematic and replicable method” (Riff, Lacy, and Frederick 2013)

1. Define a “coding* scheme” with procedures, definitions, and examples.
2. Recruit and train multiple “coders” (or “annotators”, “labelers”, or “reviewers”) with the coding scheme.
3. Have coders independently code at least a portion of the same items, then calculate “inter-annotator agreement” or “inter-rater reliability.”
4. Define and follow a process of “reconciliation” for disagreements, e.g. majority rule, talk to consensus, expert/leader decides.
5. Modify coding scheme, training, and/or reconciliation as needed.

* We’ve been using “coding” to describe this work since before punchcards existed!

Our data labeling/annotation process (study 1)

Labelers: Five undergraduate students working for course credit were trained, then independently reviewed each paper

Reconciliation: Disagreement reconciled by talking to consensus, facilitated by the team leader, who made the final decision.

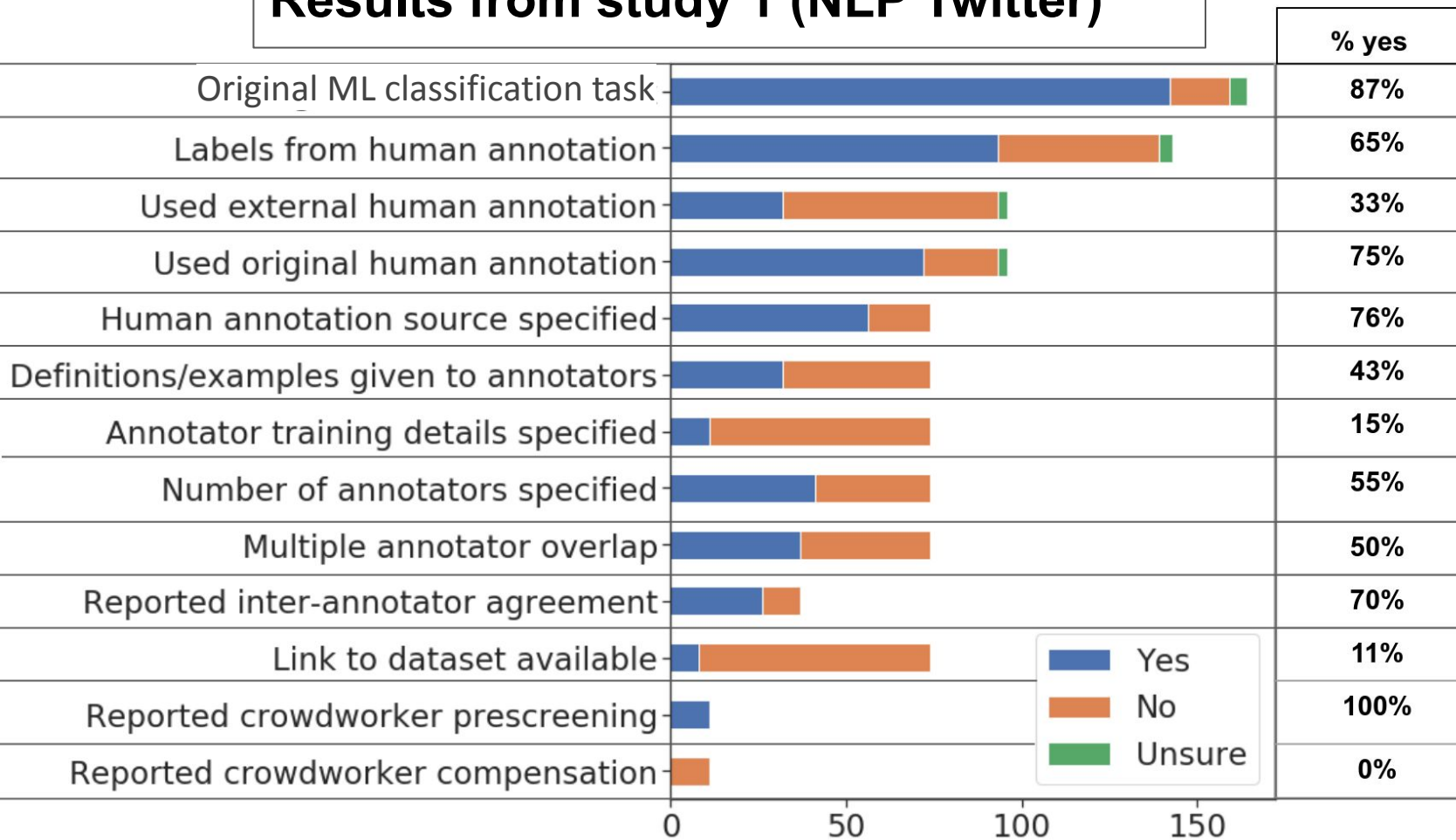
Iteration: Two rounds of labeling, after low agreement rates were found in round 1. Schema and instructions were revised and validated

Agreement: mean total agreement across all questions was 84.4%.

Questions we asked:

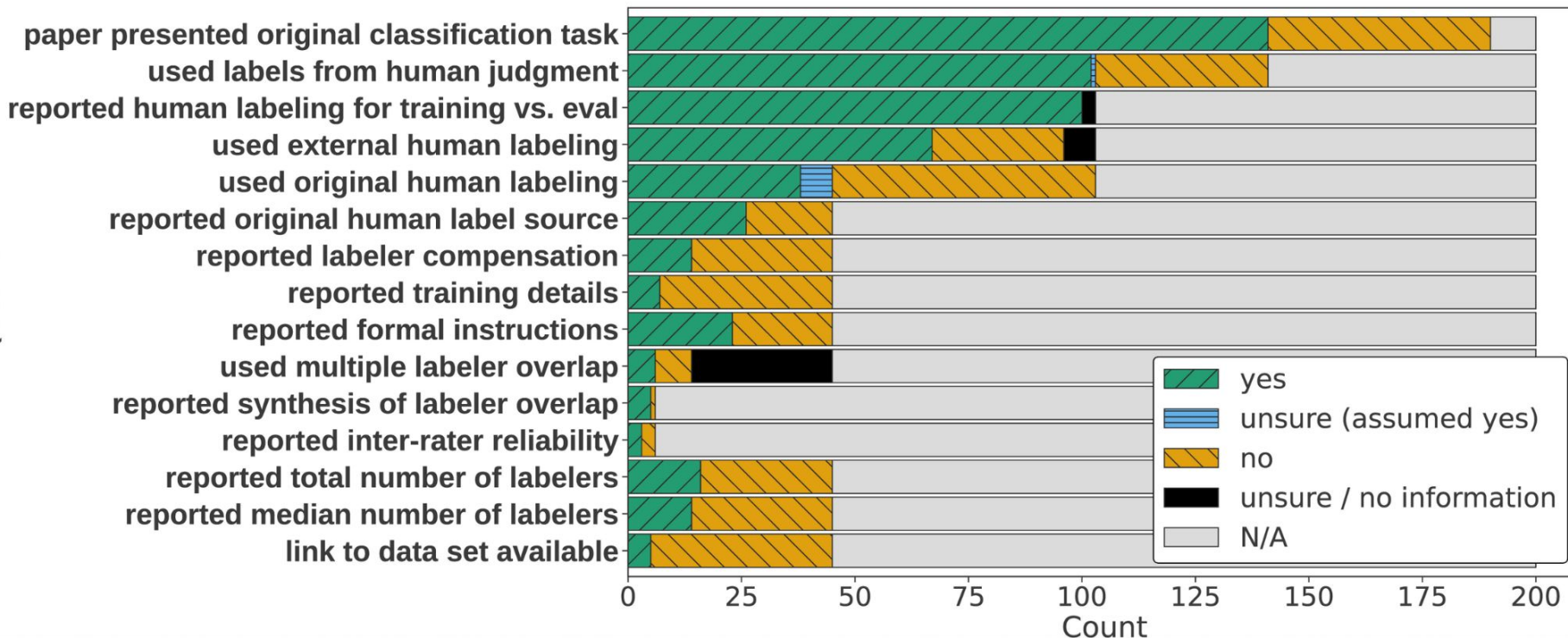
1. Is the paper presenting an original ML classification task?
2. Are the training data labels from human annotation?
3. Were the human labels from original labeling, an external dataset, or both?
4. Who labeled the dataset? (e.g. authors, turkers, experts)
5. Were the number of human annotators specified? (either total or per item)
6. Were instructions, formal definitions, or examples given to annotators?
7. Did annotators receive interactive training (beyond instructions/schema)?
8. For projects using crowdworkers, were annotators pre-screened?
9. Did multiple humans independently annotate every item (or some items)?
10. If so, were inter-annotator agreement metrics reported?
11. For projects using crowdworkers, was compensation reported?
12. Is there a link to the dataset available in the paper?

Results from study 1 (NLP Twitter)



Results from study 2 (across disciplines)

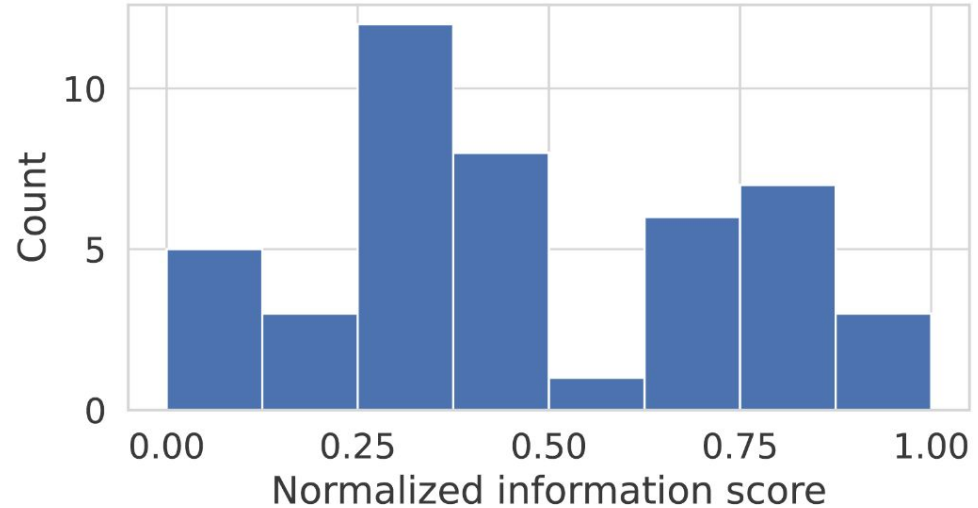
Summary results for all questions, recoded for presence of key information



Distribution of information scores

What proportion of information needed to reproduce the study was reported?

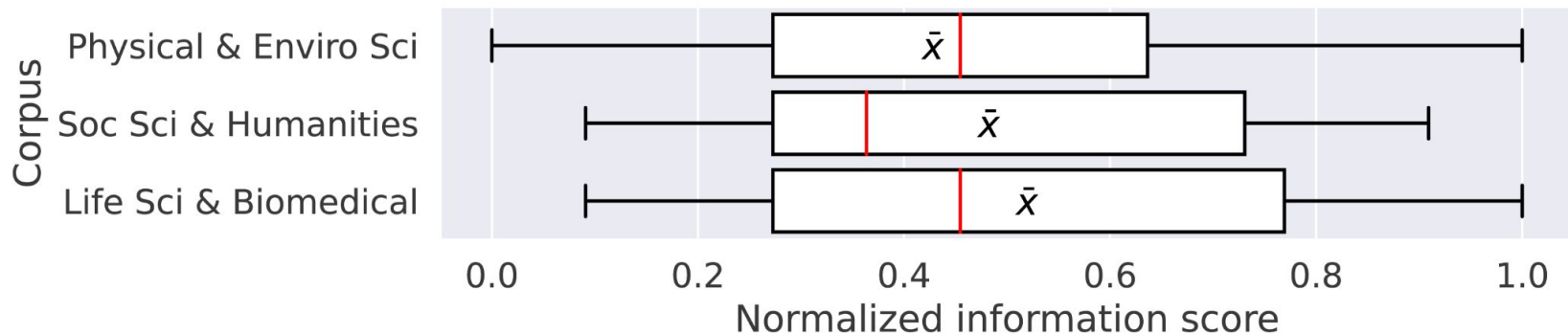
A roughly bi-modal distribution suggests there are two populations of papers/studies.



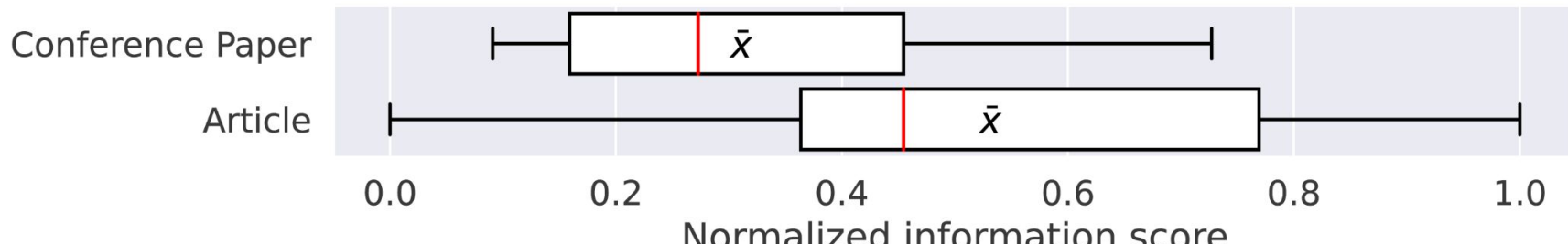
Distribution of information scores

What proportion of information needed to produce the study was reported?

Boxplot of normalized information scores by corpus



Boxplot of normalized information scores by document type



“Garbage In, Garbage Out” Revisited: Labeling and Dataification Practices Across Disciplines

R. Stuart Geiger, Ph.D
Assistant Professor

University of California, San Diego
Department of Communication
Halicioğlu Data Science Institute

stuart@stuartgeiger.com | see papers at stuartgeiger.com
CIC 2022, Virtual | slides at tinyurl.com/GIGOCIC