Data Thinking: Tech Freedom Schools for the 21st Century

Dr. Jaan Altosaar Li
One Fact Foundation
University of Pennsylvania
Tartu University
jaan@onefact.org

Dr. Ruha Benjamin
Ida B. Wells Just Data Lab
African American Studies
Princeton University
ruha@princeton.edu

Cierra Robson
Ida B. Wells Just Data Lab
Department of Sociology
Harvard University
crobson@g.harvard.edu

BOOTCAMP DESCRIPTION. The material consequences of data and the artificial intelligence algorithms trained on data belie global hierarchies of power. For example, the field of algorithmic fairness may of emerged because of the economic advantages of providing advertising systems free of regulatory problems. Data-driven advertising practices, financial instrument design, credit scoring, tax planning, biomedical science, pharamceutical research and development, housing, corporate real estate, and public infrastructure are all driven by data and the humans who interpret algorithms trained on this data and make decisions based on their interpretations that affect us all across these sectors of society. Becoming fluent in the abstractions necessary to represent the incentives and market dynamics of artificial intelligence is vital for learners who seek autonomy, and community organizations that seek to assess the impact they aim to achieve. Understanding and amplifying these goals of increased autonomy and impact for under-represented learners and community organizations they might work with requires understanding the human scale prior to building mental models of algorithms that scale. Human involvement is necessary across the discipline of data thinking, from data collection, curation, standardization, analysis, visualization, communication, and advertising, alongside other core data thinking skills. Decisions about capital allocation and human resources allocation must be made at each stage a data-to-decision journey. Many people prefer to live and work within countries and systems that prioritize the worst-off among a population. In this data thinking bootcamp for Tech Freedom Schools, we take this stance, and center the creation and delivery of educational materials on the emotional journey of a learner. Besides enabling autonomy and impact, this can help bridge educational gaps between under-represented community organizations and the PhD holders who build AI (both groups bear the consequences of algorithmic decision-making systems at scale). Through this collaboration between the One Fact Foundation and instructors the Ida B. Wells Just Data Lab in the Department of African American Studies and the Department of Sociology at Harvard University, we can help learners give informed consent from sociological, anthropological, and ethnographic lenses that are vital to understand the emotions, thoughts, and behaviors of people in power who deploy artificial intelligence---and truly give every learner a chance to decide if and when to subject themselves to algorithmic decision-making that may run the risk of ignoring some of these lenses.

WEEK 1: THE STAKES - DESCRIPTIVISM AND PRESCRIPTIVISM IN LANGUAGE, HEALTH, MATHEMATICS, & CULTURE

In English style and usage, there is no right answer. This is because practitioners of data operate on the principle of parsimony: shorter descriptions of things are easier to communicate about. For if a phenomenon *must* be communicated about a certain way (a 'prescriptive' stance), then the working memory of the practitioner would suffer from increased load: two things must be kept in mind instead of one, the first being a description of a phenomenon, and the second being the rules the phenomenon *must* be described with. But cognitive load prevents proper analysis of data and subsequent decision-making, and runs the risk of a practitioner omitting unobserved confounders from analysis and making false conclusions.

To illustrate these principles to learners, we ground our initial discussion on language, as a gentle introduction to the feeling of the rug being swept from under your feet as you realize that most things are a social construct subject to the forces of cultural evolution, collective behavior, history and so on: from race to gender, to the country you live in and the verbal events of thoughts, emotions, and feelings you use to communicate each day.

For example, we illustrate the stark difference between prescriptivist and descriptivist stances toward the English language using the following readings:

- **Arts and culture**. (Deis 2015) describes how the subversive elements of hip hop can be viewed as amplifying their political impact by enabling listeners to also consider breaking "rules", in opposition to prescriptivist censorship or criticism of this music due to its nonconformism.
- **English language usage and style**. (Orwell 1946, Wallace 2001, Garner 2022, Butterfield 2015, Stahl 2023, Somers 2014) all provide descriptors of varying stances toward modern English and usage, and some such as (Garner 2022) even include weighty discussion of social justice issues that intersect with written and spoken language. (Tutuola 1994), written as oral traditions in the Yoruba culture were being succeeded by English serve to "break the fourth wall" and highlight a learner's own experience and responses to "standard English" versus how they might define "non-standard English" for themselves, alongside monitoring their implicit biases and rules throughout their data thinking journey.
- **Ethnography, anthropology, and open source intelligence.** (Chipchase 2017, Fiorella 2022, Holmes 2013)

FIGURE 1. The phenotype workflow can help practice data thinking to best validate definitions of health and disease, intellectula property, the market dynamics, etc.

LEGAL RESTRICTIONS ON NON-PROFIT RELEASE OF OPEN SOURCE ARTIFICIAL INTELLIGENCE

Care algorithms deployed across a community organizations across the United States must comply with a their jurisdictions' privacy laws, anti-discrimination laws, and antitrust laws, in addition to conforming to the rules, laws, and regulations governing data journeys and evidence-based decision-making.

Deployed open source care algorithms have merits: they can be publicly inspected for conformance with the law, compliance with clinical guidelines, and interpretability to patients and providers. Open source care algorithms can enhance clinical care, patient experience, and the efficiency of health systems – and build trust by enabling the public to assess and monitor the performance of algorithms and their safety over time. But similar to federal laws for the protection of patients and the advancement of open access biomedical research, the Internal Revenue Service must

remunerate its stakeholder, the government, by planning tax rates and receiving monies from business activities.

This creates a direct conflict with the incentive that patients and providers might have to prefer open source, non-profit backed care algorithms: the Internal Revenue Service now prohibits corporations with 501(c)(3) designation from releasing open source software. For example, one rationale for this decision is to prevent for-profit corporations from funneling the development of lucrative artificial intelligence into open source "shell" non-profit corporations.

Open source software, by definition, can be used by any entity for generating revenue, which must be taxed to remunerate the Internal Revenue Service. For example, the majority of educational institutions and hospitals in the United States have a tax-exempt 501(c)(3) designation from the Internal Revenue Service, which prohibits them from developing and using open source care algorithms unless they work with for-profit entities to release this software. Some universities even have technology transfer offices that refuse to work with open source inventions.

The One Fact Foundation, a corporation with 501(c)(3) designation from the Internal Revenue Service, was created to solve this problem: we release Data Thinking and the GPT models we build such as (Huang, Altosaar, and Ranganath 2020) through a limited liability company that, in turn, is owned by the parent 501(c)(3) and has been created solely for this purpose using a legal team with prior expertise ranging from the Linux Foundation to the Software Freedom Law Center. The One Fact Foundation's open source, nonprofit-backed engine powers care learner- and community organization-facing machine learning and artificial intelligence algorithms; our tooling scales to clinical data repositories with petabytes of data, both structured and unstructured fields of repositories and data lakes, financial transactions, real estate, health, climate, and other areas where timeseries databases, and visualization engines for business intelligence powering consumer- and client-facing product are able to scale as much as we are.

REFERENCES

Butterfield, J. (Ed.). (2015, June 18). Fowler's Dictionary of Modern English Usage. In *Fowler's Dictionary of Modern English Usage*. Oxford University Press. https://www.oxfordreference.com/display/10.1093/acref/9780199661350.001.0001/acref-9780199661350

Chipchase, J. (2017). *The Field Study Handbook*. https://www.thefieldstudyhandbook.com/

Deis, C. (2015). Hip-hop and politics. In *The Cambridge Companion to Hip-Hop*. Cambridge University Press. https://doi.org/10.1017/CCO9781139775298.017

Fiorella, G. (2022, November 23). *How to Maintain Mental Hygiene as an Open Source Researcher*. bellingcat. (Retrieved May 25, 2023, from https://www.bellingcat.com/resources/2022/11/23/how-to-maintain-mental-hygiene-as-an-open-source-researcher/)

Garner, B. A. (2022, December 22). Garner's Modern English Usage. In *Garner's Modern English Usage*. Oxford University Press. https://www.oxfordreference.com/display/10.1093/acref/9780197599020.001.0001/acref-9780197599020

Holmes, S. M. (2013, June 19). Fresh Fruit, Broken Bodies: Migrant Farmworkers in the United States. In *Fresh Fruit, Broken Bodies*. University of California Press. https://doi.org/10.1525/9780520954793

Huang, K., Altosaar, J., & Ranganath, R. (2020, November 28). *ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission*. https://doi.org/10.48550/arXiv.1904.05342

Orwell, G. (1946). *Politics and the English Language | The Orwell Foundation*. (Retrieved May 25, 2023, from https://www.orwellfoundation.com/the-orwell-foundation/orwell/essays-and-other-works/politics-and-the-english-language/)

Somers, J. (2014). *You're probably using the wrong dictionary « the jsomers.net blog*. (Retrieved May 25, 2023, from https://jsomers.net/blog/dictionary)

Stahl, D. (2023, January 3). *English in the Real World*. The Millions. (Retrieved May 25, 2023, from https://themillions.com/2023/01/english-in-the-real-world.html)

Tutuola, A. (1994). *The Palm-Wine Drinkard and My Life in the Bush of Ghosts*. Grove Paperback Page Count. https://groveatlantic.com/book/the-palm-wine-drinkard-and-my-life-in-the-bush-of-ghosts/

Wallace, D. F. (2001). Democracy, English, and the Wars over Usage. *Harper's magazine*.