Data Thinking Bootcamp: Tech Freedom Schools Edition

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

Cierra Robson
Ida B. Wells Just Data Lab
Department of Sociology
Harvard University

PROJECT ABSTRACT. 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: DESCRIPTIVISM AND PRESCRIPTIVISM IN LANGUAGE, HEALTH, PSYCHOLOGY, & CULTURE

The first week sets the stakes and exposes learners to difficult questions, and begins shaking the foundations for concepts that might traditionally be attended with psychological rigidity.

To start helping learners build psychological flexibility in approaching core data thinking skills, an experiential approach is necessary for exposure to difficult concepts and communication about these in semi-public spaces such as the learner team chat or GitHub issues where homework, reading, watching & listening exercises are assigned and experience reports recorded.

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.

Subsequent to language, we turn to numbers. Who trusts numbers? Who uses numbers to make decisions? We use open-ended questions (Gibb and Altosaar Li 2023) to elicit questions from learners and guide their cognitive processes toward several such focal points of discussion. Readings such as (Ogle 2019, Odell 2017, Zarya 2023, Hoang 2022) further help "pull the rug out" from beneath concepts learners' might be anchored to given their diverse backgrounds and upbringings, such as time, money, capitalism, and other social constructs that relevant to the key concepts woven throughout this course such as immigration, the criminal legal system, education, health care, climate and the environment.

To ensure learners are given ample space to explore their potential for building psychological flexibility, informed consent and mental health are key, which can be reinforced through readings such as (Fiorella 2022).

Topics. We illustrate the differences 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, storytelling, and open source intelligence.** (Chipchase 2017, Fiorella 2022, Holmes 2013, Cooperrider 2023, Villarreal 2022, Effler 2010, Gould 2015)
- **Health.** (Mikkelsen and Abramoff 2023, Zarya 2023, Bridges 2011).
- **Journalism.** (Angwin 2023).

Skills. The first week serves to begin exposing you to stimuli of your choice (such as embarking on this course! or choosing what to read).

- **Speaking on Zoom.** This helps socialize you to the common fear of public speaking, notice which other learners or instructors also share this fear, and expose ourselves collectively. Such exposure during the first week helps begin to dull the emotional response to the stimulus of public speaking and associated anxieties or fears, if they exist.
- **Signing up for new tools and paying attention to friction, pain points, cognitive processes.**
- **Writing.** Experience report.
- **Asking for help.**
- **Creating a visualization of 10 million 3-1-1 calls in New York City.**
- **Asking large language models for help reading.**

WEEK 2: MATHEMATICS, STATISTICS, AND THINKING LIKE A LIMITED PARTNER BUYING OUT

This week we help learners try on a new "mask", (Johnstone 1987), and learn mental models to help them understand the incentives that belie the data they work with. We will help learners find categories of permeable experience: vision, species classification, collective behavior, machine learning models, languages/dialects, sectors of the economy, color (Twomey et al. 2021), justice, borders, skin tone, race, ethnicity, disease, toxics ("J&J Knew for Decades That Asbestos Lurked in Its Baby Powder", n.d.), and so on. When something becomes categorized, and a group of humans agree on this category (or a single human with power or status agrees), then this category or delineation can be exploited by the market, and the invisible hand of the market can step in and helping society delineate further borders and create efficiencies.

Such factors and data journeys are essential to practicing the art and science of data thinking.

In this week, learners will be guided through a brief history of statistics, starting from disputes at the foundations of this field (Lenhard 2006), and from how things like the Normal distribution were first used to segregate Black people and white people by cognitive ability.

Such mathematical and theoretical foundations, combined with perspectives from sociology, anthropology, and ethnography of data and numbers will help learners draw their own conclusions (or lack thereof!) from data.

Following such foundation material, learners will be able to use large language models to engineer prompts to learn prescriptive and descriptive ways of describing a simple machine learning model.

They will then apply this model to data of their choice. In parallel, they will be tasked with developing a research question, extending their first week's experience report, and critiquing fellow learners' research questions.

The data thinking framework of "feel, ask, do, think" serves to anchor learners in a process that helps them decide on stimulus-response-stimulus patterns to guide their exploration of concepts in the class. By this second week, learners are expected to become familiar with core data thinking skills of asking for help (from fellow learners, instructors, large language models), articulating in writing descriptions of experience, basic programming in python and the command line (in GitHub codespaces), and generating simple visualizations from data of their choosing.

WEEK 3: DATA CLEANING, DUCKDB, AND DATABASES

WEEK 4: VISUALIZATIONS AND METADATA

(Heer and Moritz 2023) will be used to visualize all of the team chat to date.

WEEK 5: ACTIVE LEARNING WITH A HUMAN IN THE LOOP

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

REFERENCES

Angwin, J. (2023, February 4). *Journalistic Lessons for the Algorithmic Age – The Markup*. (Retrieved June 2, 2023, from https://themarkup.org/hello-world/2023/02/04/journalistic-lessons-for-the-algorithmic-age)

Bridges, K. M. (2011). *Reproducing Race: An Ethnography of Pregnancy as a Site of Racialization* (1st ed.). University of California Press. https://www.jstor.org/stable/10.1525/j.ctt1ppjpz

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/

Cooperrider, K. (2023, February 23). *The allure of stories - Many Minds podcast*. DISI. (Retrieved June 2, 2023, from https://disi.org/the-allure-of-stories/)

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

Effler, E. S. (2010). *Laughing Saints and Righteous Heroes: Emotional Rhythms in Social Movement Groups*. University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/L/bo8367574.html

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

Gibb, K., & Altosaar Li, J. (2023). *Once Upon*. Once Upon. (Retrieved June 2, 2023, from https://help.onefact.org/once-upon)

Gould, D. (2015). When your data make you cry. In *Methods of Exploring Emotions*. Routledge.

Heer, J., & Moritz, D. (2023). Mosaic: An Architecture for Scalable & Interoperable Data Views. *Private communication and early access provided to one fact foundation - do not re-cite*.

Hoang, K. K. (2022, September 6). Spiderweb Capitalism: How Global Elites Exploit Frontier Markets. In *Spiderweb Capitalism*. Princeton University Press. https://doi.org/10.1515/9780691229102

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

Johnstone, K. (1987). *Impro: Improvisation and the Theatre*. Routledge. https://doi.org/10.4324/9780203446294

Lenhard, J. (2006). Models and Statistical Inference: The Controversy between Fisher and Neyman-Pearson. *The british journal for the philosophy of science*, 57(1), 69–91. https://www.jstor.org/stable/3541653

Mikkelsen, M. E., & Abramoff, B. (2023). *COVID-19: Evaluation and management of adults with persistent symptoms following acute illness ("Long COVID") - UpToDate*. (Retrieved June 2, 2023, from https://www.uptodate.com/contents/covid-19-evaluation-and-management-of-adults-with-persistent-symptoms-following-acute-illness-long-covid) (: Mark E Mikkelsen, MD, MSCE Benjamin Abramoff, MD, MS)

Odell, J. (2017). *No such thing as a free watch*. (Retrieved June 2, 2023, from https://www.jennyodell.com/free-watch.html)

Ogle, V. (2019, May 1). Time, Temporality and the History of Capitalism. *Past & present*, 243(1), 312–327. https://doi.org/10.1093/pastj/gtz014

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/

Twomey, C. R., Roberts, G., Brainard, D. H., & Plotkin, J. B. (2021, September 28). What we talk about when we talk about colors. *Proceedings of the national academy of sciences*, 118(39). https://doi.org/10.1073/pnas.2109237118

Villarreal, A. (2022, November 1). The logistics of fear: violence and the stratifying power of emotion. *Emotions and society*, 4(3), 290–306. https://doi.org/10.1332/263169021X16518516966303

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

Zarya, M. (2023). *Nourishment*. Nourishment. (Retrieved June 2, 2023, from https://www.thenourishmentproject.com/)

J&J knew for decades that asbestos lurked in its Baby Powder. (n. d.). (Retrieved June 2, 2023, from https://www.reuters.com/investigates/special-report/johnsonandjohnson-cancer/)