DataWorks Responsible Computing Curriculum

v2, April 2023

Created by Annabel Rothschild, arothschild@gatech.edu

Supported by the National Science Foundation, the Public Interest Technology University Network, and the Constellations Center for Equity in Computing at the Georgia Institute of Technology

Draft: Not for public distribution

Introduction

This workshop helps a) enable participants to make sense of their work at DataWorks and how it contributes to the larger client projects we work on, and b) employ their existing knowledge and lived experience into communicating their thoughts and – should they arise – their concerns. Examples throughout the workshop are political and culturally relevant and make clear to participants that their lived experience is an essential part of the perspective they bring to data work. This also doubles as a reminder to participants that the designers of machine learning tools and other data-intensive systems can lack in those places where participants have rich lived experienced; this reiterates the importance of the perspective that participants bring to machine learning as a field.

The first version of this workshop (as described in this document) ran as an eleven-week series; however, future versions will run as a ten-week series due to a condensed version of the initial exploratory activity.

The workshop is broken up into two sections. The first (weeks 1-5) is a programming-free approach to understanding machine learning and data intensive systems. Designed to be accessible to participants without any computational background, the first section covers the underlying themes of machine learning and artificial intelligence, namely, pattern-recognition from previous examples. A decent portion of the client work at DataWorks contributes either data annotation or labeling of datasets used to train machine learning models and other data-intensive systems. This portion helps participants see themselves as part of those projects and ensure they feel they can engage with a client project's goals and motivations.

The second section (weeks 5-10/11) is a technical approach to data preparation for machine learning projects. Given the need for datasets to be well documented, standardized, and

orderly for, e.g., machine learning model consumption and training, this section supplies a practical approach to employing "good data hygiene" in the context of large-scale datasets.

Each unit is delivered as a 90-minute in-person session lead by an instructor. Some of the units include additional activities completed individually, which, in the context of DataWorks, would be included in the larger training hours portion of the workweek. Most weeks end with either a journal activity, where participants reflect individually on the unit, or in a group discussion, in which participants reflect together.

Week 1

Materials: slides, paperclips for logic chain activity

Learning goals:

1. See and copy file tree structure.
2. Practice thinking through potential uses of data.
3. Document (via logic chains) how they came to a judgement, in writing.

Provides an overview of the two sections and the workshop as a whole. Sets the tone for the workshop as both a site for learning and sharing lived experience in the context of data work. The session is focused on exploring the concept of *data as evidence*, or as a tool for demonstrating an opinion, perspective, conviction, or argument. This sets the tone for later lessons that show examples of data being employed as a sociopolitical tool.

Week 2

Materials: slides

Learning goals:

1. Understand machine learning at high level (as pattern recognition) and have a conceptual model of how data travel through an ML pipeline.
2. Understand the extrapolation process ML aims to develop; given an unfamiliar data point, it tries to find the pattern most similar to that it's seen before.
3. Describe their role in ML client projects (as labelers, transcribers, correctors, standardizers, etc.).
4. Recognize machine learning in the wild and understand that it can have both good and bad effects.

Introduces machine learning, employing CAPTCHA as a familiar example. Machine learning is made familiar as a pattern-recognition system that employs rules developed based on prior data seen. The nuances between supervised and unsupervised machine learning approaches are not explored; the level of detail we stick to is how a machine system can perform generalization. We also explore the shape of the larger machine learning and data-intensive system ecosystem, identifying and highlighting the role of participants in those systems.

Week 3

Materials: slides, prompt handouts

Learning goals:

1. Recognize how machine learning is used to extrapolate from known instances to new (unforeseen ones).
2. Develop a list of questions they plan to ask every time they start with a dataset (they will then test these out in the project in subsequent two weeks).
3. Practice evaluating a new dataset for potential applications.

This session focuses on what it means for a machine learning model to generate insight for new or as-of-yet unseen data. We do this by talking about extrapolation and how it happens. Participants go through a series of scenarios where machine learning models and systems have been employed and try to identify what kinds of data might be involved and how the system generates new insight. We close with a video from the Data For Black Lives organization as a way to demonstrate that equity and empowerment-focused applications for machine learning are out there and that, if employed correctly and thoughtfully, machine learning can have good societal impact.

Week 4

Materials: slides

This session introduces the role of data annotation platforms in machine learning and other data-intensive systems. The first time we ran this workshop, we set aside time for the participants to individually work on two data annotation platforms (Clickworker and Microworker) before working with Amazon Mechanical Turk (AMT) as a group. However, in future iterations, we can run this solely using AMT in the group session, which requires the instructor guiding participants through the interface and having more control of the session.

The format and process of performing data annotation as a worker on AMT is treated as the status quo in contemporary data work. The rest of the session is devoted to comparing that session with how things work at DataWorks and how participants experience differences in agency in those two work settings. As part of this, we engage with several performance art pieces created to explain the power differentials between workers and requesters in data annotation work. This session helps participants see and engage with the political implications of the projects we contribute to.

Week 5

Materials: slides, handout

Learning goals:

1. Synthesize what they've learned so far and communicate what parts of it matter to them.
2. Exert some ownership over the data intake process.
3. Practice communicating their area of expertise to someone with less background; SCRUM and specific project experience to activity partner.

This week is designed to wrap up the first section of the workshop. Using what they have already learned and reflected on in prior sessions, participants (re)design an intake process for new client projects that takes into mind the goals and setting of a potential client project. This is done through a data checklist, including creation of a data bibliography (answers the who, what, where, when, why of the project dataset(s)). The final activity is developing a shared intake checklist based on recommendations from all pairs.

Week 6

Materials: slides, handout

This unit is devoted to introducing the second section of the workshop series. Most importantly, it establishes the importance of practicing good data hygiene to support client projects, but also as tools for fairness and justice in data-intensive systems. To link these ideas, we talk about example organizational data lifecycles and develop a rendition of our own data lifecycle at DataWorks.

Week 7

Materials: slides, Ask a manager dataset, discussion handout, salary dataset handout

Note: this dataset should be replaced with one that is more relevant to the interests of participants in the current workshop series.

Learning goals:

1. Practice estimating data moves needed and confirm with actual work.
2. Practice documenting with the assumption that someone will immediately come behind you and need to understand what you did.
3. Practice deciding what tool to use (OpenRefine, Excel) based on the task specifications.
4. See standardization from the perspective of a client – understand why standardization matters in data work.

This unit establishes a common language around standardization at DataWorks. First, we discuss documentation, in the form of a README and a data log, as well as their different uses as documentation tools. Second, we work on estimating how long a series of data moves might and how they should be ordered. Participants then perform these data moves on the dataset, as the first half of a project that continues into the next unit.

Week 8

Additional materials: dataset and instruction pairings submitted at the end of the previous session.

Learning goals:

1. See what good documentation looks like in real time.
2. Understand how well someone else able to read your own documentation.

Participants get live insight into how well their documentation efforts can be understood by their peers. Paired with a peer, each participant hands off the data moves instruction list they wrote out, as well as a fresh version of the dataset, and see how well their peer can follow their documentation to achieve the same result. At the end of the session, paired peers provide constructive feedback to one another. As the last event in the session, we use an automated csv comparison tool to determine which pair achieved the closest mirrored data moves.

Week 9

Materials: file type cards, file type game, file type handout, reCAPTCHA lawsuit article

Learning goals:

1. Recognize difference between file types; which might be better suited to a given purpose; where they are interchangeable and not; and pointers to what kind of file you might be dealing with.
2. Revisit documentation to practice the iteration mindset.

This unit is devoted to wrapping up introducing file types and wrapping up section two of the workshop. We have a short introduction to file types and then play a quiz game. The later part of the session is devoted to a reflection on the workshop as a whole and tying together the two sections. Further, we revist the intake checklist made at the end of unit one.

A note on sensitive content

This workshop hits on two points that can be sensitive or potentially distressing to participants. First, the overwhelming nature of machine learning and artificial intelligence can distress participants and make them feel like the world is against them. Second, like any situations that pulled on lived experience, the workshop series can unintentionally trigger uncomfortable memories or recollections. On top of the opportunities for reflection built into the curriculum, the workshop instructor should keep an eye out for potentially struggling participants and take advantage of local, site-specific opportunities to engage in reflection and optimistic endeavors. For example, at Georgia Tech, this might look like making use of the outside space during conversational portions of the workshop, or scheduling guest speakers who use data for good.