Adam Rule **Research Statement**

Individuals and organizations analyze data in domains as diverse as healthcare, engineering, journalism, public policy, and research to generate insights and make decisions. Analyses are often messy, spanning multiple files of diverse types which can be difficult to piece together into the exact sequence of steps used to produce a result. Analysts also rarely annotate their work with decisions or reasoning, even as small changes to how data are collected, cleaned, and modeled can lead to vastly different results. If insights from data are to be trusted, replicated, or simply reviewed, the process used to generate them must be tracked and communicated with clarity. To meet this need, ***I study and develop interactive documents that help data analysts track, perform, and share their work.***

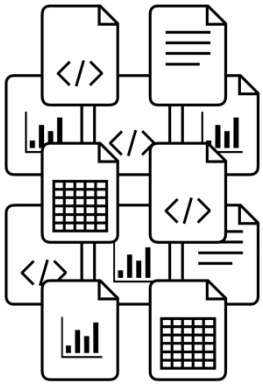

Fig 1. Data analysis routinely involves manipulating numerous files of diverse types, leading to confusion over the exact process used to produce an insight.

I am a human-computer interaction researcher (HCI) with expertise in cognitive science and medical informatics. My research aims to improve the *clarity and reproducibility* of data analyses by *developing interactive systems that encourage analytic and collaborative best practice*. I am first and foremost a keen observer, mixing computational and qualitative methods to observe how hundreds of thousands of analysts use millions of documents to track, perform, and share their work. I also build systems to test how interactive documents might better support analyses, especially collaborative ones. To date, my research has focused on two domains: use of computational notebooks (i.e., Jupyter Notebook) in academic research and use of electronic health records in healthcare. *My work has resulted in eleven major papers, won two paper awards at top HCI conferences, influenced the product roadmaps of industry leading data analysis software, and provided a corpus of over 1 million documents for research which has been downloaded over 200 times.*

SCAFFOLDING NARRATIVE & COLLABORATION IN COMPUTATIONAL NOTEBOOKS

One fundamental challenge of data analysis is tracking analyses in ways that both humans and computers understand. While computers are good at producing and consuming data, humans understand the world through narrative. This difference introduces a tension between tracking analyses through mediums which computers understand (e.g., raw data files, analysis scripts) and those that humans more easily work with (e.g., textual accounts, visualizations). Moreover, as analysts try different versions of an analysis it can be difficult, without tedious record-keeping, to track which script produced which result and why that analytical variant was even tried in the first place. In the last decade, data analysts have started using computational notebooks to address these issues, which enable analysts to mix executable code and visualizations with explanatory text in a single document. Notebooks aim to help analysts write *computational narratives* supporting collaborative, lucid, and reproducible analysis [1]. They have seen wide adoption and are used by millions of people in diverse domains. *But are notebooks being used to share compelling narratives, or simply to explore data? Do analysts find it easier to collaborate and explain their work in notebooks?*

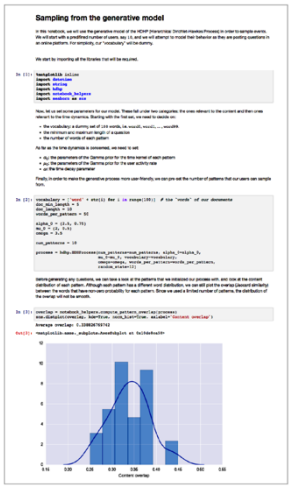

Fig 2. My research has explored how analysts use computational notebooks like Jupyter Notebook to track and share data analyses.

In my PhD dissertation I analyzed use of computational notebooks at three different scales in work that won a best paper honorable mention at CHI, the top conference in HCI [2].

First, I developed a method to scrape and analyze all 1.25 million Jupyter Notebooks hosted publicly on GitHub. Second, I hand-coded features of ~150 notebooks shared as supplementary material to academic publications. Finally, I interviewed 15 academic data analysts who used Jupyter Notebook. Together these three methods revealed a lack of explanation in computational notebooks (1 in 4 had no explanatory text) driven by data explorations which tended to produce "messy" notebooks which analysts were hesitant to clean and share. Notebooks were being used more for the way they supported iterative exploratory analysis than how they supported clear tracking and communication of process and results. Consequently, most notebooks were loose and unannotated collections of scripts that even their original authors had difficulty understanding and re-running.

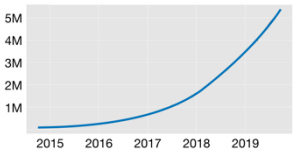

Fig 3. Publicly shared Jupyter Notebooks on GitHub over time. In 2017 I scraped and curated a dataset of all 1.25M notebooks on GitHub. This corpus has supported multiple studies by other researchers and my original analysis was recently replicated and extended by researchers at Amazon with 4M notebooks. https://github.com/jupyter-resources/notebook-research

To help reduce the *tension between data exploration and process explanation*, I developed and tested an extension to Jupyter Notebook in research that was published at CSCW, the leading HCI conference on collaborative work. [3] Through both a controlled lab study with undergraduate data science students and field deployment with academic researchers I demonstrated how simply enabling analysts to label and fold sections of their notebook aided both replication of the analysis by others and presentation of results in lab meetings. This work yielded the insight that *interfaces enabling active reading* (e.g., flipping, folding, marking up of analyses) might better support the collaborative process of data analysis.

This work also highlighted the need for *best practices*. One interviewee noted that she had received formal training on how to track experiments in paper notebooks, but lacked similar standards and training for computational notebooks. Working with leading educators and researchers, including one of the co-founders of Jupyter, I helped consolidate a set of best practices for conducting and sharing analyses in Jupyter Notebooks which was published in one of the top bioinformatics journals [4]. These practices encompassed both analytical best practices (e.g., how to modularize code and perform version control) and communicative ones (e.g., annotating process and not just results). *Aimed at practitioners, this article was viewed more than 25,000 times in the first month after its release.*

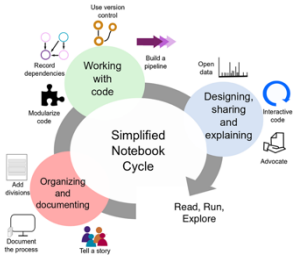

Fig 4. I helped develop best practices for the use of Jupyter Notebooks in collaboration with leading researchers and educators, including Project Jupyter co-founder Fernando Perez.

MIXING DATA, NARRATIVE, AND ACTION IN ELECTRONIC HEALTH RECORDS

While millions of analysts use computational notebooks, tens of millions of "end-user analysts" work with data without writing a line of code using systems such as spreadsheets. *How might interactive systems help these analysts work with, interpret, and communicate about data? In particular, how might notebook-like systems mixing analysis and commentary help them work more effectively?* One domain where I have begun to explore these questions is healthcare. Healthcare routinely involves the collection, analysis, and interpretation of large amounts of patient data. Moreover, team members with diverse skills and expertise need to communicate clearly about this data and its interpretation to provide and coordinate care. One of the primary tools used to do so is electronic health records (EHRs), complex software systems that support a range of healthcare activities including ordering, billing, documentation, and coordination of care. I have focused on how providers document and communicate about care using two EHR-supported formats. The first is *structured data*

fields capturing patient information such as vital signs, problem lists, and family histories. The second is *textual notes* which providers write to summarize and interpret data, justify care plans, and coordinate care. As in other domains, healthcare workers have struggled to track and communicate how they work with data, in large part due to the recurring challenge of mixing mediums that are human and machine readable. My work has provided evidence about how providers mix structured data and narrative and how tying them more closely together might better support healthcare workflows.

My recent work has explored how providers and their staff use templates to write clinical notes. These templates enable providers to create specifications for full notes or parts of notes that mix boilerplate text and dynamically retrieved patient data. Providers can invoke these templates by typing keywords such as *oneliner* into their notes. While providers have used templates for decades, few studies have examined how they do so, especially modern templates which are highly customizable and composable. My work has begun to quantify how often providers use templates and their impact on clinical documentation. For example, in one study to appear at CHI 2020, [5] I found providers use templates to document the vast majority of patient visits (95%), that most of the information imported by these templates was large data tables (such as medication lists) rather than explanatory text, and that providers primarily used personal templates rather than sharing them with other providers. In a second study, I found that one consequence of this reliance on templates is highly redundant notes, with 75% of text in notes for subsequent visits by the same patient with the same provider being exactly the same [6]. This redundancy can make it difficult for providers to quickly scan a note to see what has changed in a patient's care.

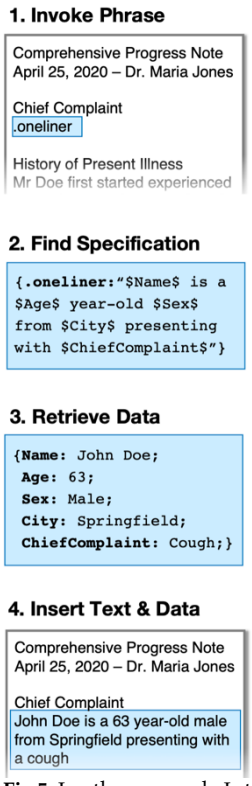

Fig 5. In other research, I study how physicians write clinical notes using custom templates mixing boilerplate text and dynamically retrieved data. This mixing of text and data in documents by invoking commands represents one form of end-user analysis.

Providers struggle to mix text and data effectively in their notes but have shown a willingness to invoke textual commands to construct documentation. This observation raises the question of how notebook-like interfaces might better support clinical workflows by more tightly integrating text and data. In one study with the Veterans Medical Research Foundation I explored how providers might simultaneously place medication orders and document having placed those orders by parsing note-text in real time [7]. This system enabled providers to write free text in their note, and intersperse that text with medication orders constructed using an inline search interface. Rather than have a separate page to place orders, and then need to reference those structured orders when describing care plans in their note, providers could just place the order from their note while writing their plan. In a usability study, providers expressed how this paradigm could save documentation time, improve patient safety by reducing discrepancies between their notes and structured orders, and how they would like to be able to both retrieve information (e.g., find most recent colonoscopy) and place other orders (e.g., x-ray of left leg) by simply typing into their notes. These results suggest untapped potential for notebooks in healthcare.

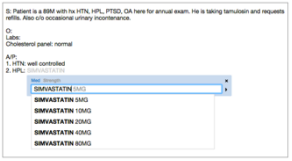

Fig 6. The ActiveNotes system enabled physicians to place medication orders inline with other note text using a domain-specific search interface, bringing the notebook paradigm to clinical notes.

FUTURE RESEARCH AGENDA

In the coming years I aim to better characterize the diverse and messy process of data analysis and develop tools that help analysts robustly work with and communicate about data.

How do people work with data? Inspecting notebooks and clinical notes only reveals part of the messy process of data analysis. I aim to develop tools and methods to observe entire analytic workflows. This is difficult because analyses often leverage multiple complex applications that may not track user history and even extend beyond computing environments. Early in my PhD I developed software to track cross-application computer use for weeks at a time without researcher intervention [8]. I have also been working with a national network of clinicians and researchers to develop standards for using EHR access logs to study clinical activities at scale [9]. I aim to build on this foundation by developing tools to track and visualize analytic workflows spanning applications. Better understanding the shape of data analysis in diverse domains will help us develop better tools to support it.

Fig 7. I developed the Traces program to enable long-term naturalistic studies of computer mediated work. Critically, Traces supports experience sampling and enables screen recording of participant's computers for weeks at a time without researcher intervention. These recordings can be used to guide follow-up interviews when visualized in a tool such as ChronoViz.

How might we support end-user analysis in diverse domains? Most prior work has focused on supporting students, academic researchers, and professional data scientists. As I have begun to do in healthcare, I aim to investigate how interactive documents might better support working with data in domains as diverse as engineering, journalism, government, and the arts without necessarily requiring users to write code. This will involve better understanding how people work with data in these domains and developing domain specific languages and interaction paradigms to support unique needs, workflows, and ways of knowing.

How might we scaffold the learning and application of data science best practices? Building on prior work I would aim to better articulate data science best practices (both analytic and collaborative) as well as develop tools that help students learn and analysts apply them in practice. While some programming best practices apply to data analysis, data analysis has different goals and outputs. Moreover it is unclear which interventions would be most effective in providing guidance (e.g., tutorials, templates, developer tools). *My ultimate aim is to help people conduct analyses that are lucid, reproducible, and sound.*

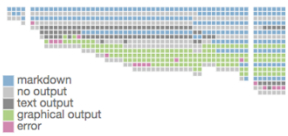

Fig 8. During my PhD, I began to develop tools to track and visualize data analyses that occurred in Jupyter Notebooks. In the future, I aim to develop tools to track and visualize analyses that span applications and extend beyond computing environments.

REFERENCES

[1] Perez F, Granger BE. Project Jupyter: Computational narratives as the engine of collaborative data science. 2015.

[2] **Rule A**, Tabard A, Hollan JD. Exploration and explanation in computational notebooks. *In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018. **(Best Paper Honorable Mention)**

[3] **Rule A**, Drosos I, Tabard A, Hollan JD. Aiding collaborative reuse of computational notebooks with annotated cell folding. *Proceedings of the ACM on Human-Computer Interaction, CSCW*, 2018

[4] **Rule A**, Birmingham A, Zuniga C, et al. Ten simple rules for writing and sharing computational analyses in Jupyter Notebooks. *PLoS Computational Biology*. 2019.

[5] **Rule A**, Hribar MR, Chiang MF. Clinical Documentation as End-User Programming. *In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020. (To Appear)

[6] Hribar MR, **Rule A**, Huang AE, Dusek H, Goldstein IH, Henriksen B, Lin WC, Igelman A, Chiang MF. Redundancy of Progress Notes for Serial Office Visits. *Ophthalmology*, 2019.

[7] **Rule A**, Rick S, Chiu M, Rios P, Ashfaq S, Calvitti A, Chan W, Weibel N, Agha Z. Validating free-text order entry for a note-centric EHR. *In American Medical Informatics Association annual symposium proceedings*, 2015

[8] **Rule A**, Tabard A, and Hollan J. Traces: A Flexible, Open-Source Activity Tracker for Workplace Studies. *CSCW Workshop on the Quantified Workplace*. 2016.

[9] **Rule A**, Hribar M, Chaing M. Using Electronic Health Record Audit Logs to Study Clinical Activity: A Systematic Review of Aims, Measures, and Methods. Journal of the American Medical Informatics Association. (To Appear)