Data analysis tools for statistical non-experts

Eunice Jun

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington
2023

Reading Committee:
Jeffrey Heer, Co-Chair
René Just, Co-Chair
Tyler H. McCormick

Program Authorized to Offer Degree:
Paul G. Allen School of Computer Science & Engineering

University of Washington

Abstract

Data analysis tools for statistical non-experts

Eunice Jun

Co-chairs of the Supervisory Committee:
Jerre D. Noe Endowed Professor Jeffrey Heer
Paul G. Allen School of Computer Science & Engineering
Associate Professor René Just
Paul G. Allen School of Computer Science & Engineering

Data analysis is critical to science, public policy, and business. Despite their importance, statistical analyses are difficult to author, especially for researchers with expertise outside of statistics. Existing statistical tools, prioritizing mathematical expressivity and computational control, are low-level while researchers' motivating questions and hypotheses are high-level. Researchers need to translate their questions and hypotheses into low-level statistical code in an error-prone process that involves grappling with their domain knowledge, statistics, and programming.

In this talk, I will introduce two tools that embody a new way of authoring analyses: Tea and Tisane. Researchers directly express their domain knowledge through higher level abstractions, and the tools will validate the data, select a statistical analysis, and implement it, all while educating analysts about why a statistical approach is valid. Tea helps analysts author statistical tests. Tea's key insight is that statistical test selection can be cast as a constraint satisfaction problem. Tisane enables analysts to author generalized linear models with or without mixed effects, which are difficult for even statistical experts to author. Using Tisane, analysts can express their conceptual models using a high-level domain specific language. Tisane translates these conceptual models into causal DAGs and engages analysts in a disambiguation process to arrive at an output statistical model. Real-world researchers have already used these tools to conduct analyses in published research that push their own disciplines forward. I will also introduce "hypothesis formalization," a series of cognitive and operational steps analysts take to translate their research questions into statistical implementations. Hypothesis formalization retrospectively explains why Tea improves statistical testing and directly inspired the design of Tisane.

Tea and Tisane serve as platforms for further research into computational support for statistical analysis. This talk also exemplifies how combining human-computer interaction with other areas in

and outside of computer science leads to software tools that impact real-world users.

Acknowledgements

Insert acknowledgments here.

DEDICATION

To all the wild women who have danced with me. I promise to never stop.

i stand
on the sacrifices
of a million women before me
thinking
what can i do
to make this mountain taller
so the women after me
can see farther
- legacy
Rupi Kaur

There is a special place in hell for women who don’t help other women.
Madeleine Albright

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter typically includes:

- a brief overview
- a challenges section
- a section about your approach
- a section (or subsection in the approach section) giving a dissertation outline (a roadmap of the rest of the thesis)

Statistical analysis plays a critical role in how people evaluate data and make decisions. Policy makers rely on models to track disease, inform health recommendations, and allocate resources. Scientists develop, evaluate, and compare theories based on statistical results. Journalists report on new findings in science, which individuals use to make decisions that impact their nutrition, finances, and other aspects of their lives. Faulty statistical models can lead to spurious estimations of disease spread, findings that do not generalize or reproduce, and a misinformed public.

Despite the prevalence of statistical analyses and their central importance to a number of disciplines, they remain challenging to author accurately. The key challenge in developing accurate statistical models lies not in a lack of access to mathematical tools, of which there are many (e.g., R **?**, Python **?**, SPSS **?**, and SAS **?**), but in accurately applying them in conjunction with domain theory, data collection, statistical knowledge, and programming ability **?**. Analysts must translate their implicit domain knowledge into statistical models that they can then implement and execute in code. However, this process—which requires disciplinary, statistical, and programming expertise—is out of reach for statistical non-experts who depend on accurate analyses, including many researchers.

Approach

This dissertation asks if separating the above concerns and incorporating automated reasoning in statistical software could benefit statistical non-experts. Towards this goal, I combine techniques from human-computer interaction, programming languages/software engineering, and statistics to (i) characterize the cognitive and operational steps to author statistical analyses and (ii) develop novel interactive systems that enable statistical non-experts to author valid analyses. As detailed below, I not only move between systems building and empirical studies but use each to deepen and enhance the other.

The work described in the dissertation demonstrates the following:

Thesis statement Domain-specific languages that provide abstractions for expressing conceptual knowledge, data collection procedures, and analysis intents instead of specific statistical modeling decisions coupled with automated reasoning to compile conceptual specifications into statistical analysis code can help statistical non-experts more readily author valid analyses.

Three challenges fall out of this thesis statement:

Challenge 1: How to make implicit domain knowledge explicit.

Designing abstractions focused on conceptual knowledge requires identifying what domain knowledge analysts want and can express and balancing these constraints with what automated reasoning approaches may require. What is easy to express and what is easy to assume for the sake of automation may be at odds, especially when analysts provide ambiguous specifications that could be compiled into multiple statistical analyses. The challenge therefore, is to design language constructs that are usable for analysts and useful for automated reasoning and support interactive program specification as necessary.

Challenge 2: Represent and reason about key statistical analysis decisions

A central idea in this thesis is that software systems should take on the responsibility of translating conceptual knowledge into statistical analyses. This is akin to a compilation process that requires representing the conceptual knowledge analysts express and reasoning over it to derive statistical analyses that respect statistical best practices and rules. A major challenge is in picking representations so that the reasoning is straightforward.

Challenge 3: Increase analysts' statistical knowledge/understanding

While automating statistical analysis can be helpful, analysts relying on data to make high-impact decisions (e.g., policy, scientific discovery) often need to understand why an analysis approach is appropriate and what the implications of the results are to their domain. Furthermore, software can inform how users approach future analyses. Therefore, educating analysts about the applicability and impact of statistical decisions and guiding their interpretation of results are important.

Summary of Contributions

To do (??) This dissertation makes systems, empirical, and methodological contributions:

- empirical findings of how authoring analyses requires integrating conceptual, data, statistical, and programming expertise, which we summarize in our **theory of hypothesis formalization**;
- an analysis of how the current statistical software ecosystem does not explicitly support and may even hinder hypothesis formalization, suggesting new **design opportunities and implications**;
- a **formal constraint-based model** to specify and select among common Null Hypothesis Statistical Tests in Tea;
- A **mixed-initiative approach** for "interactively compiling" linear models from conceptual and data relationships in Tisane;
- (Proposed) empirical findings on researchers' implicit semantics of variables, conceptual models, study design, and hypotheses;
- (Proposed) new language constructs and an updated interaction model for expressing and mapping these implicit semantics to statistical models implemented in a new version of Tisane; and
- (Proposed) evaluative studies on how usable and effective Tisane is compared to `lme4`.

In my thesis (**??**), I designed and implemented two systems, Tea **?** and Tisane **?**, that leverage **domain-specific languages** (DSLs) to capture analysts' implicit assumptions and conceptual knowledge. Users **interactively compile** these high-level specifications into low-level code. To infer valid statistical analyses, the systems **programmatically represent and reason about core statistical authoring challenges** as constraints and graphs (**??**). As a result, my systems prevent common analysis mistakes **??**. I also contribute a **new theory describing the cognitive and operational steps involved in authoring statistical analyses**.

1.1 Thesis outline

This dissertation contributes new domain-specific languages (DSLs) for authoring statistical analyses and a new theory describing the cognitive and operational steps involved in authoring statistical analyses. In the process of designing the second DSL, we also explored new methods for eliciting and integrating user feedback throughout programming language design. The content of thesis is as follows.

To do (**??**)

How to approach this dissertation

To do (**??**)

1.2 Prior Publication and Authorship

To do (**??**)

Chapter 2

Related work

2.1 Donald Campbell's Theory of Validity

Chapter 3

Tea: A Domain-Specific Language and Runtime System for Hypothesis Testing

This chapter covers material from a single project. There are typically 3-4 of these chapters total (but any number is fine as long as your advisor is on board).

3.1 Updating things if you've copied and pasted

If you are copying and pasting material from one of your papers, then remember to:

- Remove the abstract and instead add a little overview of the chapter and how it ties in to the rest of the thesis. You should also mention the original paper's source like: "This chapter includes materials originally published in \citet{myownppr}"

- Make sure the formatting still works – this is single column now!

- Consider rephrasing conference-paper-style language:

  - Find every place you mention some variation of "in this paper" and say "in this chapter" instead.

  - Remove or rephrase the parts where you talk about "our main contributions".

  - Rephrase the language describing code and data releases.

- Replace the conclusion section with a summary section. Again, you should tie this chapter back to the main themes of the thesis.

3.2 Related Work:...

3.3 Design of Tea's DSL

3.4 Tea's constraint-based runtime system

Why constraints? are they really necessary?

3.5 Evaluation:...

3.6 Key tensions

- inflated alpha - inherent tension in executing multiple statistical tests vs. sensitivity

3.7 Limitations, Ongoing work, Future directions

3.8 Summary of Contributions

Chapter 4

Hypothesis Formalization: A conceptual framework describing how analysts translate research questions into statistical analyses

4.1 Related Work

4.2 Content analysis

4.3

Chapter 5

Tisane: Authoring statistical models via formal reasoning from conceptual and data relationships

5.1 Tisane's DSL: Formalizing implicit domain assumptions as conceptual models

5.2 Deriving statistical models from conceptual models

5.3 Evaluation: Case studies with real-world researchers

5.4 Limitations

5.5 Summary of Contributions

Chapter 6

rTisane:

6.1 rTisane: Re-designing Tisane

6.2 Exploratory study

6.3 Key insights and system changes

6.4 Evaluation: Lab study assessing the impact of connecting conceptual and statistical modeling

6.5 Summary of contributions

Chapter 7

Conclusion

7.1 Summary of contributions

7.2 Note on methodology and implications

- strength of this work is moving back and forth between and integrating empirical studies with systems building - Even within empirical studies, explored and used many qualitative and quantitative approaches - Within systems building: Use diversity of technologies - Identify need for methodologies in the future for designing DSLs *with* end-users

7.3 Where all this is going / Why do we care about any of this?

- a reinterpretation of EUSE – programming tools as bicycles of the mind

7.4 Discussion: The role of programs and the act of programming as a reflective practice

Finding: interactive disambiguation not just necessary for refinement and automated reasoning but *useful* to analysts for reflection

**Not just higher levels of abstraction but appropriate abstractions that allow analysts to dig deeper into the appropriate parts

7.5 Future work

Has this dissertation lost its way? Further re-orienting towards what users *really* want: to understand their domain Push further in directions this work orients us - more support for understanding results, especially when some questions may not be answerable with the data/how it was collected - knowing how robust the results are –> why not just multiverse everything? **how do we resolve and come out from under the tyranny of false positive rates fear

7.5.1

7.5.2 What about in the face of LLMs?

Bibliography

Chapter A

Appendix One

A.1 Appendix section 1

Table A.1: Table in the Appendix