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

Insert introduction here.

This chapter typically includes:

- a brief overview
- a background section
- 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)

Alternatively: you could choose to remove the background section and make a separate background chapter directly following this introduction chapter.

Chapter 2

Related work

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 Future work

7.4.1

7.4.2 What about in the face of LLMs?

Bibliography

Chapter A

Appendix One

A.1 Appendix section 1

Table A.1: Table in the Appendix