What is your name and affiliation?
London Bielicke, cs phd student

What is your field of research or study?
Computer science

Read the introduction to rTisane, making highlights + notes + connections as you go. Explain your process.

- Highlight seems to follow the storytelling flow of the paper
    - E.g. problem, example, motivation, solution, …
- Write note but want to do a summary in the note, not just one-to-one
- Highlight content as Question
- Label as a purpose of highlight

Go back and reread the introduction, trying to answer the question: What is the methodology of the paper? Use the add read feature to make additional highlights + notes + connections with another color. Explain your process during this second read.

- Focus on methodology part
- Connect new highlight content with previous highlight (e.g. both are solution to problem)

Do you think using this tool helped deepen your understanding of the text? How so?

- It will be more useful connecting section (e.g. Link between sections)
- Make me think more intensely about what i highlighted comparing to highlight without purpose (because of label and note)
- It is cool to see the “mind map”

Did you encounter any difficulties using our interface?

- Pretty straight forward
- Hard to drag
- When Add node, have to reposition the graph

Rate the tool’s usefulness and what other features you would want to see.

- Usefulness 7/10 - would be better if integrated into current tools like an extension?
- Learnability 1.5/10 (1 means easiest to learn, 10 means hardest to learn)

Additional features you’d like to see in a more fleshed out version?

- Jump to highlight when click node
- Auto-zoom

Please take a screenshot of your highlights and graph to send to us.

## 1 INTRODUCTION

In order to answer research questions and test hypotheses, analysts must translate their research questions and hypotheses into statistical models. To do so accurately, analysts need to reflect on their implicit understanding of the domain and consider how to represent this conceptual knowledge in a statistical model. For example, consider a health policy researcher interested in accurately estimating the influence of insurance coverage on health outcomes. To formulate a statistical model, they consider prior work on how insurance coverage, race, education, and health outcomes relate to each other and other constructs. Then, they go to formulate a statistical model including or excluding covariates to account for confounding in these relationships [7].

A researcher who skips this process may overlook relevant conceptual relationships or implicit assumptions, resulting in statistical models (and conclusions) that are faulty or meaningless as answers to their motivating research question.

Key to this explanatory modeling process is analysts' domain knowledge, captured in *process models* [20] or *conceptual models* [13]. Conceptual models include variables and their relationships that are important to a domain. Figure 1 shows an example conceptual model from our exploratory study (Section 3). A number of software tools exist for building conceptual models. For example, Tisane [15], an open-source library for authoring generalized linear models with or without mixed effects, enables analysts to explicate their conceptual models and derives valid statistical models from them. Tisane has helped HCI researchers catch and fix analysis bugs prior to publication [4]. Other tools such as Dagitty [28] and DoWhy [25] also support analysts in externalizing conceptual models as causal graphs to reason through statistical modeling choices. These software tools support (i) conceptual model specification and (ii) statistical model formulation based on expressed conceptual models.

To benefit from these tools, analysts must be able to accurately externalize their implicit conceptual models (goal (i)). This goal presents two usability challenges. First, tools should make it easy for analysts to express their conceptual models. At the very least, tools should not hinder specification. Second, analysts need guidance on which implicit assumptions are important to externalize. Addressing both challenges is particularly important for making these analysis tools usable for domain experts who have statistical experience but limited expertise (i.e., many researchers).

**Figure 1: Visual overview of paper.**
Through an exploratory study, we investigate how to better support statistical non-experts in specifying their conceptual models (Section 3). Based on findings, we develop rTisane, a system for specifying and refining conceptual models in order to derive statistical models (Section 5). We compare rTisane to a scaffolded workflow in a within-subjects controlled lab study (Section 6). We find that using rTisane to externalize conceptual models deepened consideration of implicit assumptions and helped maintain analysis intent. We also find that rTisane enabled a few analysts to author statistical models they were not able to author on their own. For others, rTisane's output statistical models fit the data better or facilitated iteration.

After analysts externalize conceptual models, tools must formulate statistical models (goal (ii)) in order to obtain high-quality statistical inferences. To ensure quality, there are two challenges to statistical model formulation: fidelity of the statistical model to the conceptual model and good statistical model fit to data. These criteria provide checks on one another. For instance, for any data set, an overfit statistical model can be found that satisfies the model fit criterion as well as possible without accurately representing the analyst's implicit conceptual model. As another example, a statistical model representing an unreasonable conceptual model may not fit real-world data well. We prioritize correspondence of conceptual models to statistical models and then, given this correspondence, consider statistical model fit.

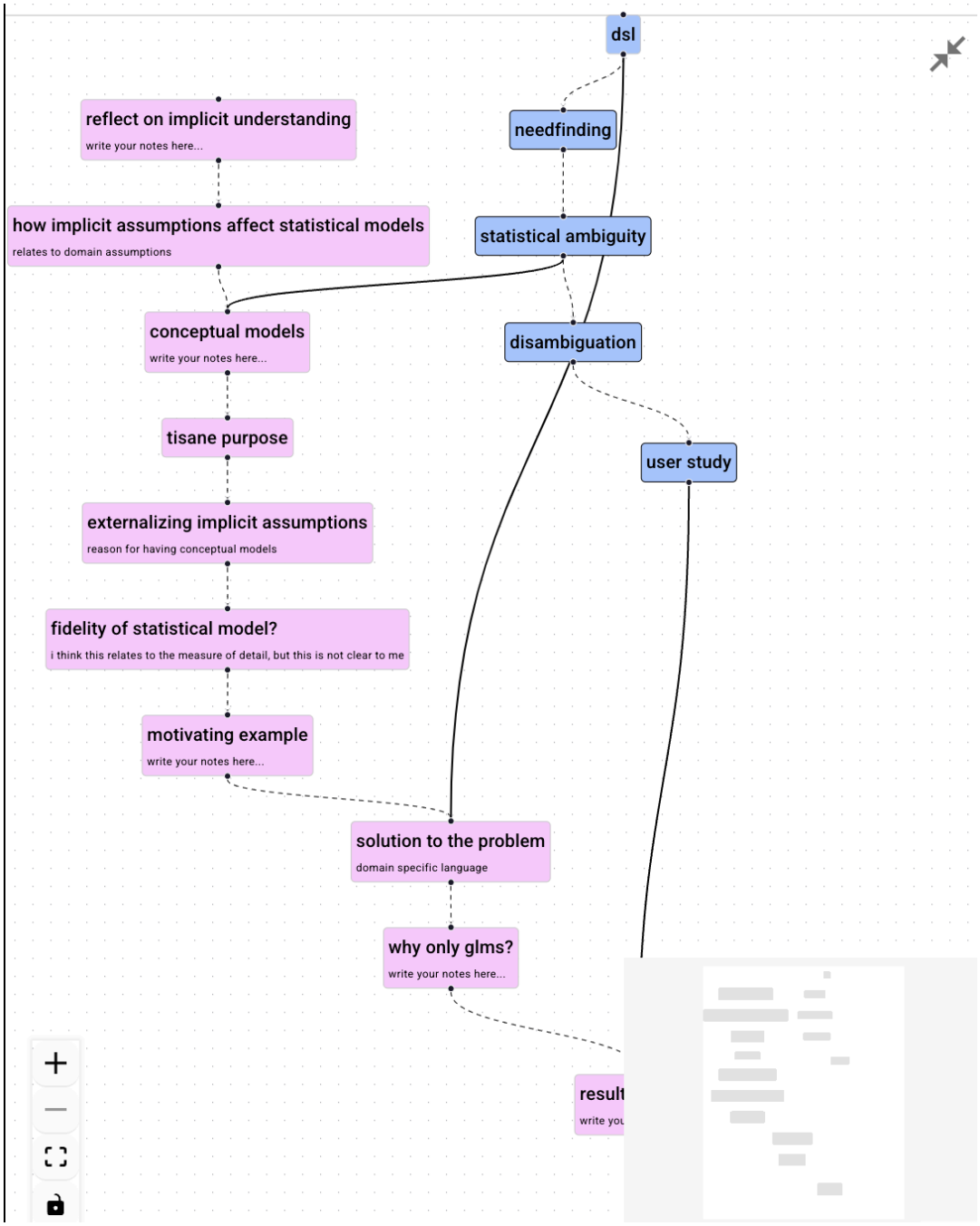
This paper investigates how to support both accurate conceptual model specification and quality statistical model formulation. We focus on the design and implementation of a domain-specific language (DSL) for expressing conceptual models and using conceptual models to author statistical models. We focus on DSL design since end-users and graphical systems alike can benefit from DSLs. Our users are analysts who have domain expertise, experience with generalized linear modeling, and experience programming in R, but are not statistical experts. We refer to these end-users as *statistical non-experts*.

We start with an exploratory study to identify challenges statistical non-experts face when expressing their conceptual models

findings, we develop rTisane, a system for externalizing conceptual models to author generalized linear models (GLMs). rTisane consists of (i) a DSL for expressing conceptual models and (ii) a two-phase interactive disambiguation process for refining conceptual models and then deriving statistical models. rTisane leverages an informative graphical user interface (GUI) for disambiguation. The result of this entire process is a script for fitting a statistical model that is guaranteed to reflect the expressed-then-refined conceptual model. To assess the impact of rTisane on conceptual model specification and statistical model formulation, we compare rTisane to a scaffolded workflow without tool support in a within-subjects lab study. We find that rTisane's DSL makes it easy for analysts to specify conceptual models and guides them to think more critically about their implicit assumptions. Furthermore, rTisane helps analysts focus on their analysis intents, and analysts are not surprised by rTisane's output statistical models. Of 13 analysts, three were only able to author a statistical model by using rTisane. Another six analysts were able to author statistical models that fit the data just as well, if not better, than statistical models they author without tool support. Figure 1 visually shows the three parts of this paper.

In summary, we contribute

- A study identifying how statistical non-experts want and are capable of expressing their implicit domain assumptions,
- The open-source rTisane system[1], which provides new language constructs for expressing conceptual models and a two-phase interactive disambiguation process for resolving ambiguity in conceptual models and deriving statistical models, and
- Evidence from a controlled lab study about how tool sup-


dsl
reflect on implicit understanding
write your notes here...
needfinding
how implicit assumptions affect statistical models
relates to domain assumptions
statistical ambiguity
conceptual models
write your notes here...
disambiguation
tisane purpose
user study
externalizing implicit assumptions
reason for having conceptual models
fidelity of statistical model?
i think this relates to the measure of detail, but this is not clear to me
motivating example
write your notes here...
solution to the problem
domain specific language
why only glms?
write your notes here...
result
write you