#### TL:DR;

Scientific papers are more than just words on a page—they are packed with data, figures, tables, and statistical analyses that researchers must interpret alongside the main text. Navigating these dense, data-rich papers requires more than traditional reading skills; it demands the ability to extract key insights, synthesize information across different formats, and assess the reliability of presented evidence. However, the challenge doesn’t end with the first reading. Researchers often need to revisit papers days, weeks, or even months later, only to struggle with recalling key insights, retracing their thought process, or relocating important sections. Without effective tools to capture and organize their evolving understanding, readers may find themselves repeating work, losing track of their reasoning, or missing crucial details.

Our project, *Augment Data-Intensive Reading*, seeks to address this challenge by exploring how researchers engage with data-heavy papers and identifying ways to improve their reading experience. Through user research, we found that researchers do not follow a single, linear approach to reading. Instead, their reading processes vary significantly, shaped by shifting intentions that influence how they engage with different parts of a paper. Understanding these intentions—and developing tools to help researchers navigate them—could lead to a more effective and seamless reading experience, not only in the moment but also when revisiting papers over time.

**[Figure: illustrates the complexity of a data-intensive paper]**

****

*Image Credit: https://www.wikihow.com/Remember-What-You-Read*

#### Example Story

Jack, an undergraduate student, stared at his laptop, overwhelmed by a data-intensive research paper assigned for class. Complex formulas, algorithmic descriptions, and figures filled the pages, making it difficult to follow. Lacking deep expertise, he skimmed the abstract and conclusion, jumping between sections to piece together key insights. Each unfamiliar term sent him searching online, breaking his reading flow. Despite multiple passes, he struggled to fully grasp the paper and decided to ask his mentor, Violet, for guidance next week.

Violet, a PhD student and experienced researcher, had read this paper months ago during her literature review. She had skimmed it first for relevance, then returned to specific sections for a deeper dive, annotating key points along the way. But now, when Jack asked about it, she found it hard to recall the details or retrace her exact thought process. Her notes helped, but without a clear way to track how she had engaged with the paper, reconstructing her insights was frustratingly slow.

As they sat down to discuss, their differing reading strategies and the difficulty of recalling their thought processes over time made communication difficult. This disconnect highlights a core challenge in data-intensive reading: not only is understanding a dense paper difficult, but revisiting it—recalling insights, retracing thought processes, and retrieving key information—is just as daunting.

#### User research methodology

[Briefly mention our goal of the user research - what problem are we researching]

The problem we initially set out to solve is how we can help readers understand academic papers better. From this, we hypothesized that this question can be applied in two different ways. Firstly, we identified that readers with less experience with papers (for example, undergraduate students who read papers mostly for class), might need help with understanding the paper as a whole. Secondly, readers with more experience (for example, PhD students that have to read papers regularly for their lab) might need assistance with specific sections, and augmenting the information presented.

Thus, the goal of our research was, keeping these two potential personas in mind, how do readers currently approach reading papers, and what are their pain points and struggles in the process of doing so. In addition, we also aim to understand the role of data in papers, and how readers interact with data in the reading process.

[What methods we use and why we choose these methods]

With the aforementioned goal, we decided to perform a combination of multiple methods: think-aloud study and interview/survey.

1. Think-aloud study

Through a think-aloud study, we aim to see how users are able to interact with academic papers first hand. We give them the option of two papers, and around 10-15 minutes to talk us through their process of reading the paper. We encourage them to voice their thoughts out loud as well as any questions they have throughout the reading process. At the end of their read, we ask them these three following questions that involve them summarizing what they read.

1. What are the key contributions of this paper?
2. How do the authors accomplish these contributions?
3. In what ways do these contributions improve upon prior work?

This approach will allow us to go beyond just an interview or survey, since most people are unaware with the challenges they face when reading papers, so having the opportunity to see readers interact with the paper directly will allow us to see things an interview might not tell us. We are also doing this in a digital setting, thus, we will be able to see how users interact with papers via a digital interface. Through this, we hope to see in what ways we can design our interface to make it as natural as possible for readers to adopt.

1. Survey/Interview

Before and after the think-aloud, we will be conducting both an introduction survey, as well as a post-study interview. Through questions like one’s outlined below:

1. Have you used any tools or techniques to help you read papers? Are there any tools that you wish existed that you think would help you?
2. Which sections of a research paper do you consider most important for understanding its content?
3. Can you describe how you typically navigate through a research paper?
4. Is your reading process linear, or do you follow a different pattern?
5. How do you approach understanding the “Results” or “Evaluation” sections of a paper?

we aim for readers to draw from past experiences when it comes to reading papers, as well as reflect on their experience during the think-aloud study. This provides additional room for discourse post study, since readers might have not voiced all their thoughts during the think-aloud study. In addition, using specific and directed questions, we aim to obtain more targeted responses to questions that we had. Also, we acknowledge that the paper we chose might not have had everything that could have been points of discussion, so allowing users to draw from their other experiences with other papers could give us a wider perspective on readers' approaches.

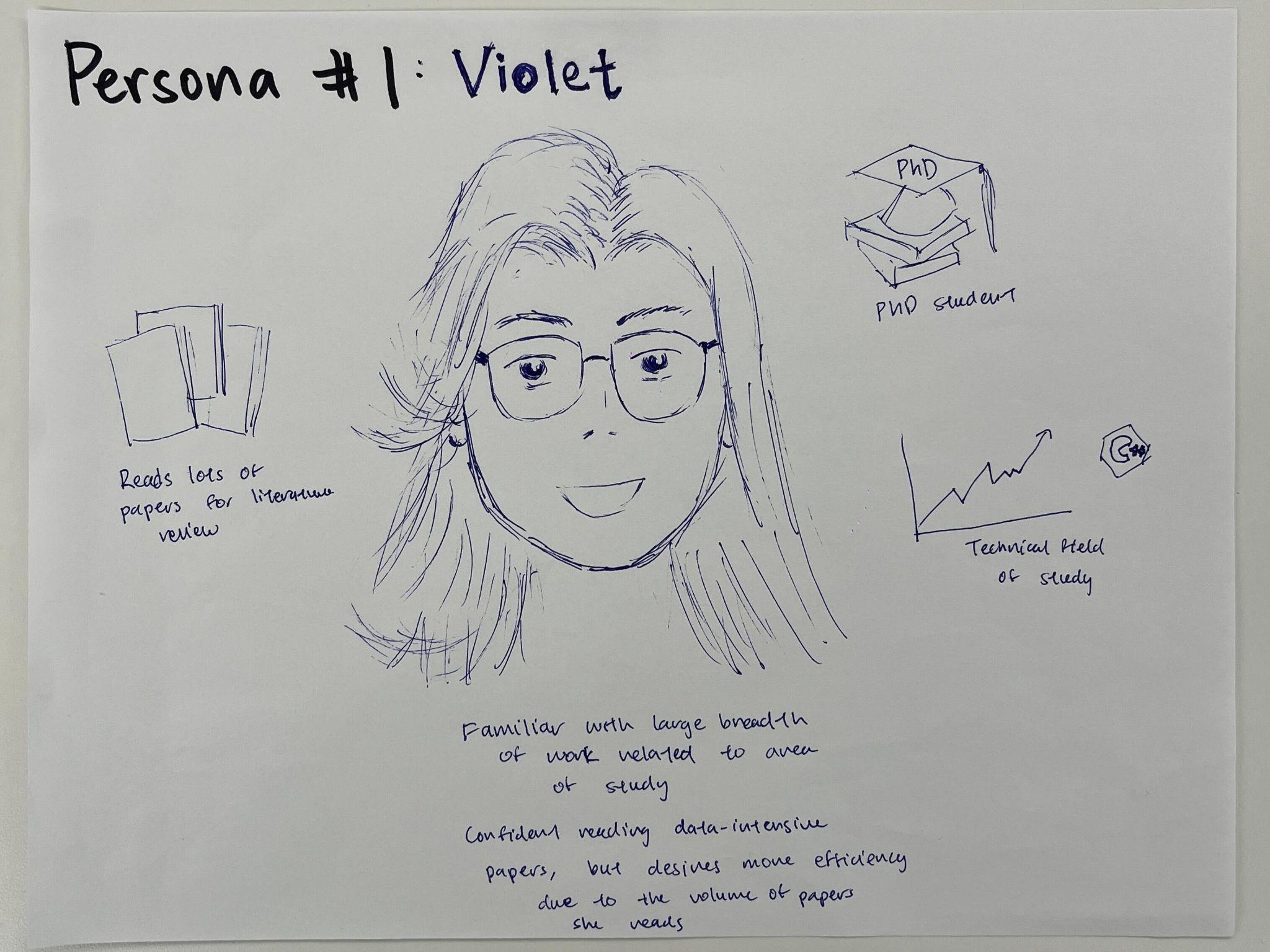
[Introduce our protocol]

We aim to perform the full interview process (survey/interview and think-aloud) with around 10 people. These interviews were done throughout the span of a week, and included a mix of both in-person and zoom interviews depending on the interviewees availability. Below we outline the whole interview process:

1. **Pre study interview/survey:** in this section, we aim to develop an initial understanding of the interviewee’s background, reading habits, and tools they currently use to help them read papers.
2. **Think-aloud study:** in this section, we present a data-intensive paper and tell the user to answer the questions mentioned in the previous section. From this, we are able to see how they interact with the paper first-hand.
3. **Post study interview:** in this section, we aim for readers to be able to reflect on their think-aloud process, and understand their perspective of data in the papers presented

#### Key findings from User research

From our user research, we determined two main user groups to target. The first is the more technical user, someone who tends to read a lot of papers and goes in great depth and breadth around a research subject. We represent this user with the persona, **Violet**:

****

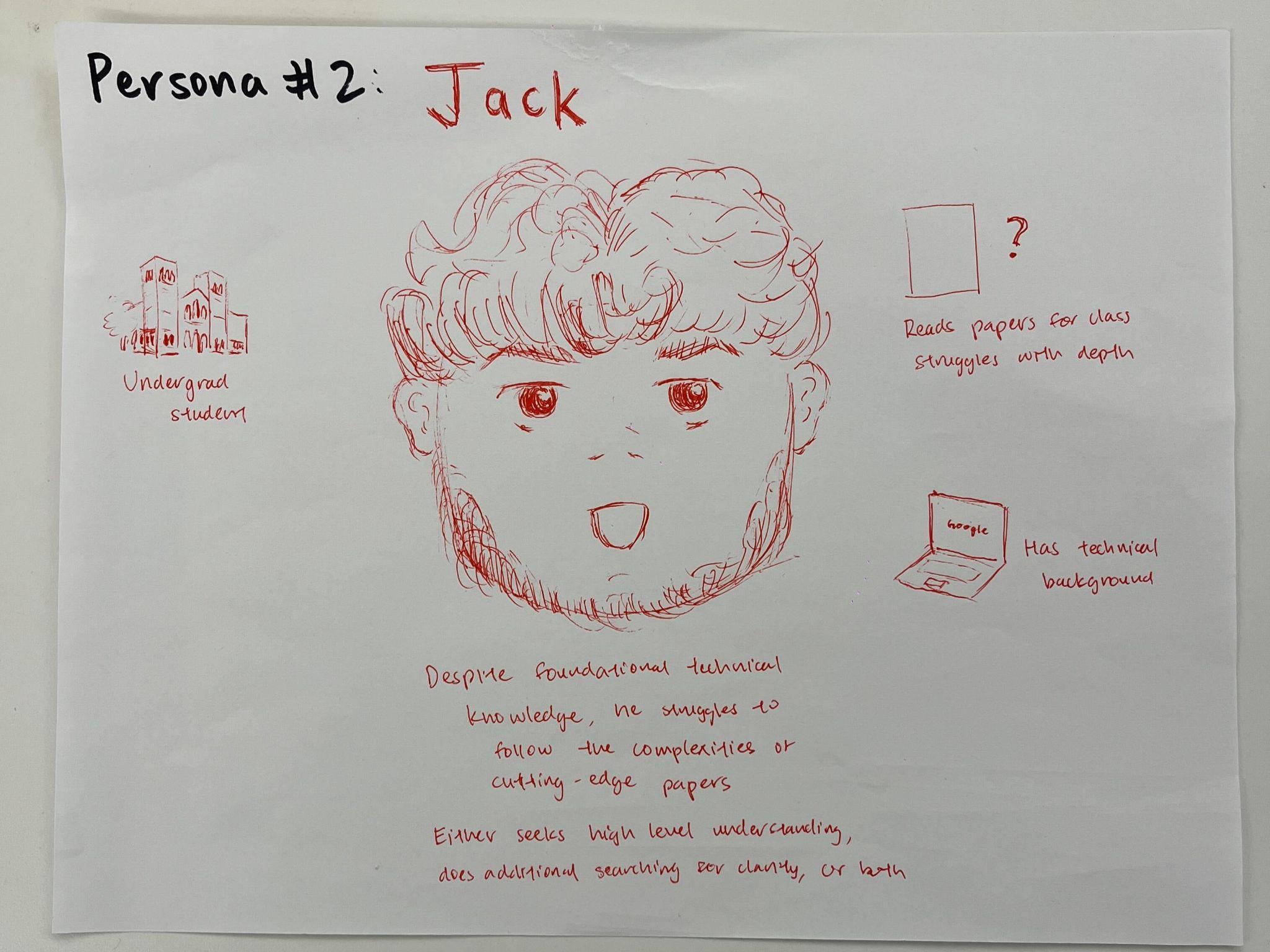
**Violet** is a PhD student in a data-intensive, technical field (think statistics, machine learning, and their applications). She regularly conducts literature review for both depth in her current focus as well as breadth to keep track of new contributions across her area of study, meaning she often reads multiple papers a day.

We determined that a regular scenario with which Violet would interact with papers would be as follows:

Violet is conducting literature review in a field related to her area of study, but not exactly the same. She is familiar with many of the common terms, but is unfamiliar with some of the formulas or models. As she potentially needs to read 20 or more papers within a week to complete her review, she needs to be able to quickly get up to speed and process information with papers she has only foundational-level knowledge in. To help, she reads on an iPad or a similar device that lets her highlight and make annotations on the paper so that she can cross-reference materials easier. With some papers, she merely skims them to understand their basic contributions, but for others she needs to dive more in depth. When skimming, she finds that she often misses info that would help understand contributions requiring a few passes through papers, and when diving deeper, she often has to manually parse referenced papers or check data analysis on her own. When diving deeper, she tends to read more linearly as she cares to be very thorough in those cases, but she wishes for a more efficient way to take in all this information without as much manual effort.

In this scenario, a tool that is able to quickly pull context from relevant papers she’s reading, or provide links between important concepts or contributions of a paper, would be greatly beneficial in improving the efficiency of her research and literature review.

Our second persona represents the less technical reader, who has interest in papers whether for class or personally and is able to understand them at a high level, but being less technical may struggle with the level of depth or breadth of a user like Violet. We represent this user with the persona, **Jack**:

****

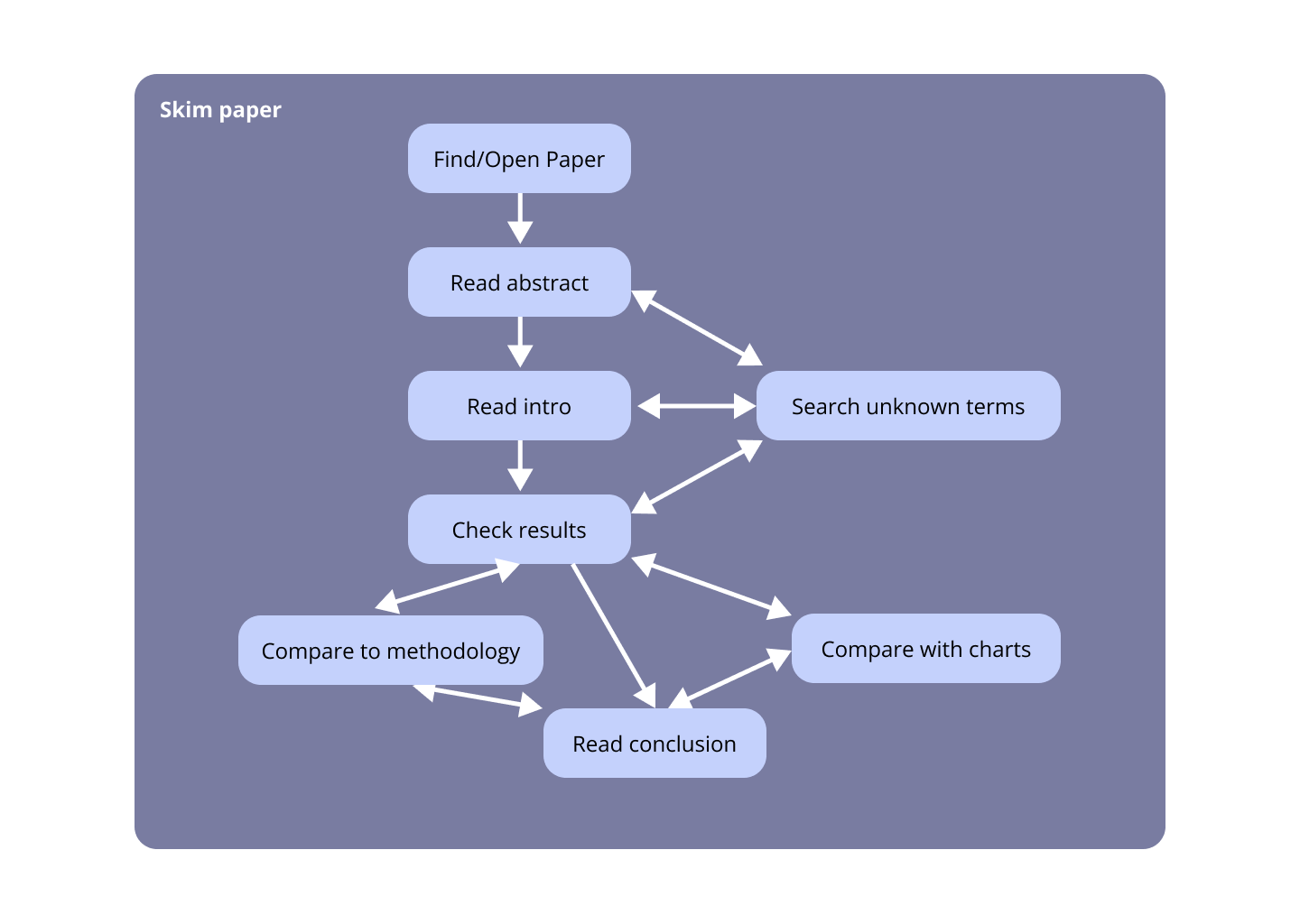
**Jack** is an undergrad student in a technical field who is required to read papers for several of his upper division classes. He has read papers intermittently in the past, and while he has some technical knowledge, he struggles to follow some papers which go into great technical depth. For the purpose of his classes, understanding them at a high level via their general contributions and results suffice.

We determined that a regular scenario with which Violet would interact with papers would be as follows:

Jack is reading a new assigned paper for his class. The paper is about some novel machine learning methods and goes into great depth in its mathematical and algorithmic foundations, as well as the data used in its experimentation. As a result, there are many complex formulas and a variety of diagrams comparing the novel model to existing baselines. Jack is familiar with some but not all the baselines, and while he has a foundational understanding of machine learning, he is unfamiliar with some of the more complex optimization techniques used in the paper. He skims most of the methodology and focuses most of his time on the results, discussions, and conclusion portion of the paper to get a sense of its contributions. As a result, he often jumps back and forth between sections of paper, reading in an erratic manner and constantly searching up terms and things he’s unfamiliar with. He wishes that he didn’t have to switch his focus from reading to searching as often as it makes it difficult to gather a clear understanding of papers.

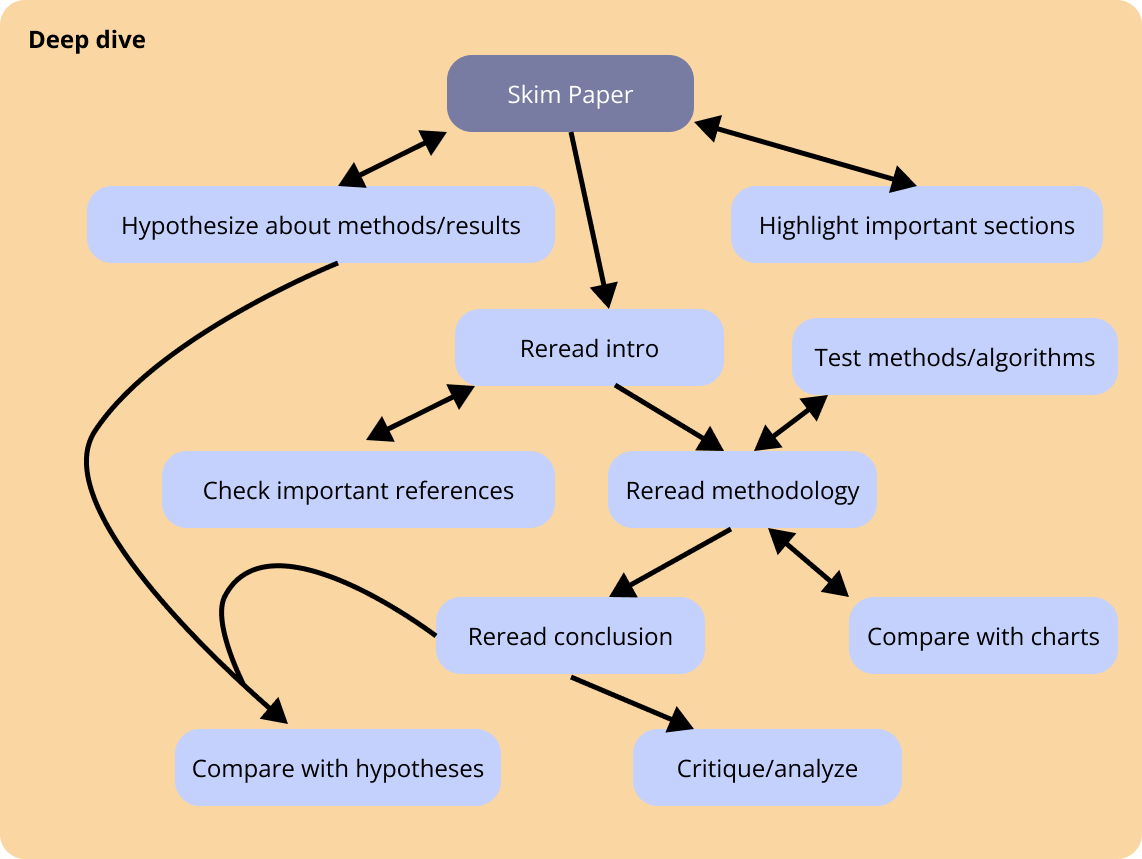
In this scenario, a tool that is able to pull in context and provide it on hover or with just singular clicks would be extremely helpful, as this would prevent the need for him to search for things on his own, interrupting his flow through the paper. Additionally, similar to Violet’s scenario, any tool which can link together significant contributions within a paper would also help Jack streamline his otherwise erratic reading.

These two scenarios provide insight into our process maps, which are inspired by the reading habits of our two personas. Our first process map describes how one skims a paper:



Violet may be checking this paper out because it’s referenced in a paper she’s doing a deep dive on, or it has a new contribution she’s never heard of, and wants to check if it is relevant to her research. Jack may be checking this paper out because it’s been assigned for class, or it was related to a paper he was assigned for class and he wants more background knowledge.

Our second process map describes a more thorough deep dive, which can involve an initial skim that involves additional steps:



Violet likely needs to do this several times over the course of a literature review and thus would like to do them as efficiently as possible. Jack may do this for a deep critique of a paper and an attempt to broaden his knowledge.

#### Problem statement

Our user research reveals that reading scientific papers is a dynamic process, driven by shifting intentions and varying levels of engagement. Readers navigate papers with different intentions—whether to gain a high-level understanding, analyze technical details, or verify findings—but the complexity of these papers makes addressing each intention time-consuming and cognitively demanding.

As readers shift between intentions, they often lose track of their progress and insights, leading to redundant effort and difficulty reconstructing their understanding when revisiting a paper. Without effective ways to capture and organize their evolving thought process, researchers struggle to efficiently retain and recall key information over time, hindering their productivity and communication.

Our project aims to address this challenge by identifying readers’ intentions and developing tools to help them track their engagement, ensuring that data-intensive reading becomes not only more efficient but also more structured and recallable.

#### Storyboard

