Task and Data Based Considerations for Design and Recommendation of Visualizations

Aditeya Pandey
pandey.ad@northeastern.edu
Khoury College of Computer Sciences, Northeastern University

Abstract

Data visualization is a graphical representation of information and data, enabling end-users to identify meaningful knowledge from the data and make informed decisions. But the process of visualization design is intensive and prone to errors. A common mistake in visualization design is the incorrect mapping between data and task requirements and visual encoding, or in simple terms, "choice of visualization". The mapping of data and tasks to a suitable visual encoding can be challenging because of the following reasons: (1) Limited Theoretical Support: Existing visualization theory may offer limited support to identify the analytical goals of the end user, (2) Incomplete or Inconsistent Guidelines: The knowledge for designing visualizations may be incomplete, inconsistent or inaccessible, (3) Lack of Practical Resources: Shortage or lack of resources that can guide visualization practitioners in their task of visualization design. My work aims to limit these challenges and support visualization practitioners and researchers in selecting visualization design more effectively.

The overarching research goal of my thesis is: *How can we develop theory, identify visualization best practices, and build applications that enable visualization practitioners to create effective and expressive visualizations?* To answer this question, my work contributes knowledge to three visualization areas: theory, design guidelines, and recommendation systems. My thesis's theory contribution includes identifying shortcomings of existing visualization theory and recommending methods to eliminate the shortcomings. To determine the shortcomings, I conducted three novel design studies CerebroVis, Strokevis and Segmentrix+Portola. In these studies, I developed tree and network visualizations to solve novel domain problems. A posthoc analysis of the studies revealed that existing general task abstraction theory lacked specificity to describe tree visualization tasks. To improve the task abstraction specificity for trees, my research contributes a tree visualization specific extension to the established task abstraction framework Multi-Level Task Typology. My thesis also contributes visualization design guidelines. In my work, I systematically curate task-based design guidelines through a meta-analysis of empirical results of tree visualization effectiveness from a survey of over 50 papers. In addition to tree visualization, my work also contributes novel design guidelines for data glyphs and timelines. My thesis's final contribution is focused on making the design guidelines and visualization knowledge easy to access for practitioners and researchers. To do so, I contribute two novel visualization recommendation systems. Based on the theory and data gathered on trees, I developed a recommendation system to help visualization practitioners navigate the vast tree design space and choose an effective visual encoding based on their data and tasks. Besides the tree visualization recommendation system, my work also presents a knowledge-based recommendation system for genomics visualizations. Such systems would make tree and genomics visualization creation accessible to both experts and novices and improve visualization literacy.

Through a series of theoretical and practical contributions, my thesis supports mapping data and task requirements to appropriate tree and genomics visualizations. In this thesis, I also reflect on the broader applications of the contributions, specifically how my work can act as a framework to support the effective and expressive design of information visualization in general.

Contents

1 Introduction

Data visualization enables end-users to analyze the data visually. The use of visualization is pervasive and plays a critical role in revealing meaningful insights from the data in different fields. Today, machine learning and artificial intelligence methods depend on data visualization to understand the learning algorithm's underlying working. Data visualization also enables physicians to visually analyze health indicators like the patients' heart rate in clinical settings. The Nested Model of Visualization Design and Validation [32] and Design Study Methodologies [45, 49] are commonly used visualization design models and frameworks that support practitioners and researchers to create visualizations. These models and frameworks argue that the core of visualization design is mapping data and task requirements of a domain problem to suitable visualization encodings and interaction techniques. For example, an epidemiologist may want to analyze the patterns in the branching of a virus strain and compare how different strains evolve. The epidemiologist is analyzing hierarchical data and the task that they want to accomplish is related to comparing the branches of the tree. Based on the data and the task requirements of the epidemiologist the most appropriate visualization technique is the node-link tree visualization.

The process of mapping of data and task requirements is intensive and prone to errors [45]. An error in the mapping can lead to an ineffective choice of visualization [32, 29]. Ineffective visualization design can lead to the spread of misinformation by distorting the way information is perceived by people [50]. Based on my personal experience with designing visualizations for different domain problems [37, 40] and analyzing existing visualization theory [45] I found that process of mapping data and task requirements to visualization encoding can be broken down into four steps. I refer to the four steps combined as the **"Visualization Design Pipeline"** (Fig. 1). Each step in the pipeline is an action that a visualization practitioner has to perform to go from the data and task requirements to a finished visualization product. I describe the steps below and also identify the challenges they can pose for visualization practitioners:

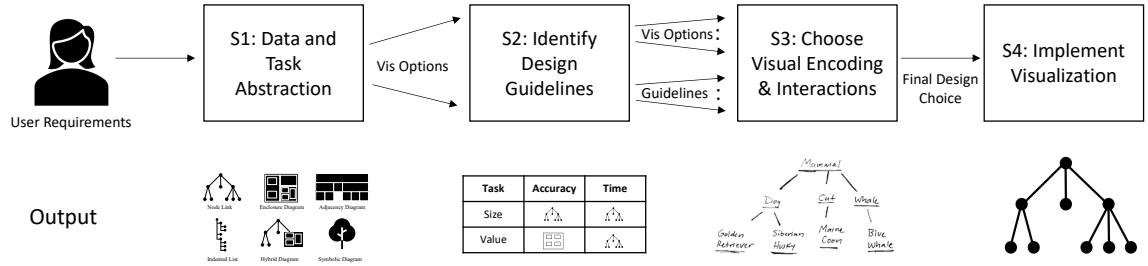

Figure 1: Visualization Design Pipeline: The top row shows a typical visualization design pipeline. Data and task abstraction (S1) identifies all the design alternatives for the problem. But among that, only a small subset of visualizations will be effective. To determine the subset, practitioners have to identify meaningful design guidelines (S2), and based on the guidelines, choose the appropriate visualization design (S3). The final step in the pipeline is implementing the visualization for the end-user (S4). The bottom row shows the output for each step in the task of designing a tree visualization.

1. **Abstract Data and Task:** The data and task abstraction phase of the visualization design pipeline maps the observed data and domain goals to generalizable abstract data and task specifications using visualization theory. For example, a biologist may be interested in results for tissue samples treated with LL-37 matching up with the ones without the peptide. A visualization designer may identify that results of the experiment are recorded as **numerical** values and translate the task to **comparison** of values between **two groups**. This transformation of data and tasks from domain-specific to abstract language is essential to enable visualization creators to effectively compare data and tasks across different domains and look for relevant techniques and strategies in different application areas.

   Challenges: Visualization researchers have proposed various data and task abstraction approaches (e.g., [33, 6, 2, 46, 27]). Adopting an appropriate abstraction approach is pivotal for visualization

design as it impacts the choice of visualization design and interaction idioms. However, selecting a proper abstraction framework requires an extensive comparison of existing literature. A practical solution to the problem of multiple options is to choose a general-purpose framework. For instance, the Multi-Level Task Typology (MLTT) framework [6] and its extended version in the Visualization Analysis and Design textbook [33] is a generic data and task abstraction framework that works well across disciplines and dataset types. However, the general-purpose frameworks sometimes lack the specificity to support task abstractions for specific dataset types such as temporal, spatial-temporal, networks, and trees.

2. **Identify Design Guidelines:** Chen et al. define "a guideline embodies wisdom advising a sound practice in creating a visualization image, designing a visual representation, or developing a visualization system." [10]. Design guidelines can be general like the "overview first and details on demand" mantra by Shneiderman [46], or specific to data type and task such as to the comparison of quantitative data is more effective with position and length visual channels (Clevland & McGill [11]). In this step, a visualization practitioner has to identify the important design guidelines for their visualization problem. This step is manual and currently requires visualization practitioners to scan and analyze information spread across scientific papers, textbooks, and technical reports.

   Challenges: The scientific community has amassed a wealth of empirical knowledge, case studies, tools, and techniques over the past decades. However, most of the knowledge is spread across scientific papers, inaccessible to the general audience of designers, and visualization practitioners [13]. The Visguides project [13], aims to resolve this problem by creating a platform where visualization researchers can share, discuss and critique design guidelines. However, the platform is still in an early phase and only has a small subset of design guidelines. Another major problem for this step is incomplete knowledge. Visualization literature is evolving, and there are many areas and aspects of visualization that have no design guidelines to support visualization practitioners. This opens up opportunities for visualization researchers to fill the gap and build visualization knowledge.

3. **Choose Visual Encoding and Interactions:** Visualization practitioners use the data and task abstraction and the design guidelines to identify appropriate visual idiom and interaction techniques. As discussed previously, the epidemiologist analyzing virus strains' evolution maps the data and task requirements to a node-link tree visualization over other forms of information visualization techniques. Visual encoding and interaction mapping allow practitioners to winnow down from a large visualization design space to find visualization technique/s that match users' goals while respecting the visualization design guidelines.

   Challenges: To select visual encoding and interaction, a visualization practitioner needs to collate and analyze large amount of information about visualization techniques and design guidelines from existing literature scattered in books, empirical studies, and survey reports. This step's manual nature, combined with large design space and unstructured availability of guidelines, can lead to a sub-optimal visualization encoding or an incorrect visual encoding to represent the data. Recently, visualization recommendation tools like Draco [31] and Voyager [55] have tried to reduce the workload from a visualization practitioner by recommending them appropriate visualization encoding based on data and task requirements. However, existing tools do not support visualization techniques like trees and networks or domain specific visualizations like visualization for genomics.

4. **Implement the Visualization:** Visualization practitioners implement the visualization encoding and interactions to a working prototype or a tool to enable users to analyze data and perform the visualization tasks. To implement a visualization design, practitioners can use a spectrum of tools ranging from low-level visualization libraries like d3.js [5] to high-level visual analysis tools like Tableau [51] and Power BI [41]. The low-level visualization libraries are flexible and enable practitioners to create customized visualization, but they can be hard to learn. Tools like Tableau and Power BI allow practitioners to develop visualization without programming, but they only support a handful of visualization techniques limiting the visualization practitioner's expressivity.

   Challenges: Given the broad spectrum of tools and the range of functionalities they offer, it is challenging for visualization practitioners to determine the solution conducive to their visualization design

problem. The tools usually have manuals that outline their functionality, but the practitioners may still need to employ a trial and error method to select the right tool. The trial and error method is time-consuming. Organizations with time and resources may afford to invest time in this part of the process. Still, it may be a disadvantage for the larger visualization practitioner audience, who may not have the time or resources to invest in this process.

The challenges in each step of the visualization design pipeline are potential research opportunities. My PhD research focuses on improving the overall visualization design pipeline by eliminating or reducing the challenges. Through a series of research projects, I demonstrate how the challenges can be solved for tree visualizations. I use tree visualization as a case study because it is beyond the scope of a PhD thesis to solve the challenges for the entire visualization design space. It is a career endeavor. The scope of my thesis is also motivated by a wide range of applications for tree visualization. Tree visualizations of hierarchical data are common in many fields such as software engineering, machine learning, geography, finance, and biology. The typical applications of tree visualizations in these fields are organization and representation of code-bases in software engineering, explainability of decision-tree models in machine learning, presentation of natural geographical phenomenon like river branching in geography, and exploration of genetic evolution data in biology [28, 33]. Therefore, my thesis contributes information on visualization creation challenges accompanied by resources and tools to help visualization practitioners design widely applicable tree visualization techniques. However, my projects are not limited to tree visualizations. Throughout this thesis, I also present visualization projects that solve the visualization design pipeline challenges for glyphs, timelines, and genomics visualizations. More broadly, my thesis also discusses a framework that can support the identification of visualization design challenges and support creating tools and resources for a broad range of visualization techniques.

Contributions: In this thesis, I present three novel visualization design studies: CerebroVis, StrokeVis, and Segmentrix+Portola. Cerbrovis is a novel visualization technique for representing cerebral arteries, and StrokeVis builds over CerebroVis to detect and diagnose stroke in patients. Segementrix+Portola presents a novel visualization system to detect anomalies in the network traffic of a data center. I applied data visualization theory to develop tree and network visualizations to solve the domain problems in all the design studies. A post hoc analysis of these studies revealed that existing Multi-Level Task Typology (MLTT) [6, 33] lacks specificity to abstract tree visualization tasks. To enable effective task abstraction for trees, I also contribute a novel extension of the Multi-Level Task Typology to include more specificity to support tree-specific tasks and a systematic procedure to conduct task abstractions for tree visualizations. With an extensive task abstraction theory for tree visualizations, I curate task-based design guidelines for tree visualizations by surveying published empirical studies. In addition to tree visualization design guidelines, this thesis also contributes task-based design guidelines for data glyphs and timelines. Based on the theory and data gathered on trees, I present a recommendation system to help visualization makers navigate the vast tree design space and choose a useful visual encoding based on their data and, importantly, their tasks. Beyond recommendation for trees, I also contribute a system to recommend genomics visualizations called Genorec. A key aspect of Genorec is its ability to guide an analyst to an existing visualization tool to help them implement the visualization.

In the remainder of this proposal, I will discuss the following topics. In Related Work, I will discuss the visualization theories, state-of-the-art in recommendation systems and provide background about tree visualizations as they are an overarching topic in this thesis. In Research Questions, I will discuss the research questions that guide my work. In Thesis Plans, I will discuss how I will answer those questions and my thesis milestones. In the remaining sections, I will discuss the thesis's accomplished work and plans for on-going and future projects.

2 Related Work

2.1 Visualization Theory

Information visualization theory provides theoretical tools like models, frameworks, and guidelines to create the visualization. In this section, I will discuss the visualization design models and frameworks. After that,

I will discuss different types of data and task abstraction frameworks. Finally, I will provide an overview of the visualization design guidelines and discuss the challenges and opportunities in this area.

2.1.1 Visualization Design Models and Frameworks

The "Nested Model of Visualization Design and Validation" was developed by Tamara Munzner to guide visualization practitioners with the creation and analysis of visualization systems [33]. The model identifies the key steps involved in visualization design, namely: domain characterization of the data and tasks, an abstraction of domain goals into operations and data types, visual design encoding and interaction techniques, and create algorithms to execute these techniques efficiently. Munzner argues that all the model steps are essential and play an indispensable role in designing effective and expressive information visualization. Sedlmair [45] et al. proposed a "Design Study Methodology Framework" (DSM) where they broke down the Nested Model by Tamara Munzner into detailed step-by-step instructions to guide practitioners on how to create a visualization design. The extended DSM framework consequently divides the visualization design into nine stages: learn, winnow, cast, discover, design, implement, deploy, reflect, and write. For each stage, authors provide real-world examples from their own research experience to help visualization practitioners understand the stages better and apply them to their problems. A pivotal step in both the Nested Model and the DSM framework is mapping the abstract data and task definitions to suitable visual encoding and interactions. In this stage, practitioners make decisions about good and bad matches of visual encoding based on their understanding of visualization design theory. While the Nested Model and DSM framework highlight the importance of the data and task to the encoding mapping stage, they do not provide sufficient knowledge to perform the mapping. The knowledge to perform data and task abstraction is available in data and task abstraction frameworks or taxonomies, and the mapping knowledge is available in the visualization design guidelines. Therefore, we will discuss the common and relevant data and task abstraction frameworks and visualization guidelines in the next two sub-sections.

2.1.2 Data and Task Abstraction Frameworks

Data and task abstraction map tasks and data from the specific domain's vocabulary into a more abstract and generic description in the vocabulary of computer science. More specifically, it is in the vocabulary of information visualization. Visualization research provides many abstraction frameworks [24, 3, 33, 44] and taxonomies [27, 1, 30, 36, 2, 46]. Out of the many abstraction frameworks, the most comprehensive is the "Three-part analysis framework" proposed by Tamara Munzner in the textbook Visualization Analysis and Design [33]. The three-part analysis framework provides information on how to abstract data and visualization problems and classify a visualization in terms of marks, channels, and interactions. The data abstraction part of the framework classifies the visualization dataset in terms of dataset type (tree, network, table, spatial data, etc.), the data types (nodes, links, attributes), and the attribute types (categorical, ordinal, and quantitative). The task abstraction part of the three-part analysis framework was adopted from the Multi-Level Task Typology (MLTT) by Brehmer & Munzner [6]. MLTT helps the user understand why a particular task is carried out and breaks down the task into high-, mid-, and low-level categorization along with the final target of the task. In the framework, each categorization consists of abstract concepts to delineate the various objectives at each stage of the task. For instance, the high-level categorizations analyze whether the visualization is used to *consume (discover, present, and enjoy)* or *produce (annotate, record, and derive)* data. The mid-level actions *(lookup, locate, browse, and explore)* describe the type of *search* carried out based on the target and location knowledge. The low-level actions *(identify, compare, and summarize)* represent the type of *query* performed on the target. Targets can be different kinds of *data (e.g., trends, or outliers)*, *attributes (e.g., extremum)*, and *topology (for network data)*.

Existing task abstraction frameworks and taxonomies are either general-purpose or dataset-specific. General-purpose frameworks and taxonomies such as the MLTT [6], Low-level components of analytic activity [2], work well across all disciplines and data-set types. However, they lack the specificity that a dataset-specific task abstraction framework or taxonomy provides. For example, Task Taxonomy for Graph Visualization [27] is a taxonomy for tasks in the field-specific to graph visualizations. This taxonomy provides a more descriptive identification of visualization goals for network visualization tasks than a generalized framework [37]. Therefore, in this thesis, I propose a method to extend a general-purpose abstraction framework (MLTT) (Sec. 6) to include dataset-specific information for trees.

2.1.3 Visualization Design Guidelines

In information visualization literature, guidelines are a set of rules or best practices that guide visualization practitioners to create visualizations. For example, in the classic study by Clevland and McGill [11], authors found out that the representation of quantitative values with position or size channel was more accurate compared to an area or angle channel. Visualization guidelines are scattered in research papers, technical reports, and surveys. Due to scattered availability of resources, visualization practitioner have a hard time to identify and find meaningful guidelines for their visualization problems [13]. The scattered guidelines also make it hard for researchers to identify the data types and tasks that have already been evaluated and the where are the opportunities for novel research [39]. Recently, there have been efforts from the visualization community to build practical resources to share and discuss visualization guidelines. Visguides [13] is a platform that enables visualization researchers to share and critique design guidelines on an open and accessible web platform. Besides Visguides, there are websites like From Data-to-Viz [16] and FlowingData [15] that educate visualization practitioners on the best visualization design practices and help them create visualization by pointing them to the right resources. Although the visualization community is making progress on communication and dissemination of design guidelines, it remains an open question on how do we systematically curate and communicate design guidelines. Therefore, there is vast research potential in this area of information visualization research. My work contributes novel design guidelines for data glyphs, timelines and trees (Sec. 7).

2.2 Visualization Recommendation Systems

Visualization recommendation systems assist practitioners to identify the most compelling visualization technique from a relatively large visualization design space [31, 19, 55, 54]. Due to the growth of design space and increased involvement of people from diverse backgrounds in visualization design, visualization recommendation tools are more important now than ever [53].

Visualization recommendation systems must take into account four considerations [53, 22]: (1) *Data Characteristics* deals with the identification of visual encoding corresponding to data type and attributes. Mackinlay [29] identified the correlation between data attributes and visual encoding. Polaris [48], the research prototype of Tableau software, further adopted the data-based recommendation system by Mackinlay to develop the "Show Me" feature. Voyager [55] is another recommendation system that automatically suggests meaningful visualizations based on the statistical characteristics of the underlying data. (2) *Task Oriented* recommendations factor in users' intentions behind visualizing data as the main criteria for recommending visualization. The current task-based recommendation system support domain-independent low-level analytical tasks like compare, summarize, distribution [42] and domain level tasks [23]. (3) *Domain Knowledge* imposes further restrictions on the results of the recommendation system as the domain expert may prefer a visualization that is more familiar or widely accepted within their domain. Specialized organizations like NASA have developed domain-based recommendations to assist in-situ visual analysis of spacecraft data [25]. (4) *User Preference* relates to factoring end users' preference in the recommendation system output. For instance, in an aesthetic evaluation of tree visualization techniques, researchers found that the sunburst chart was most preferred by the users [9]. Draco [31] has a method to factor user preference in the form of user-defined constraints.

Existing recommendation systems use the recommendation axis that suits their objective. For instance, tools like Tableau and Voyager [55] only use the *Data* axis because they want to recommend a starting point for the visualization practitioner or analyst. Tools like SeedDB [52] and VizML [19] use both *Data* and *Task* axes for recommendation but their goal is completely automated visualization recommendation. In order to achieve complete automation, SeedDB and VizML compromise on the visualization techniques and the tasks they support. The tools currently support basic one- or two-dimensional visualization techniques and low-level analytical tasks. Draco [31] is flexible and uses data, task, and user preference axes for the recommendation. However, even Draco is not capable of handling domain-specific problems or recommending specialized visualization encoding like tree visualizations.

2.3 Tree Visualizations

A *tree* is defined as a collection of nodes and links. A *node* is a data structure that can have an identifier (id) and a value. A *link* in a tree is a data structure that connects two *unique* nodes. Similar to the node, a link can also have values associated with them. A key aspect that differentiates trees from graphs or networks is the "hierarchical" relationship that exists within the nodes. Hierarchical relations categorize nodes in a tree dataset as "above" or "parent", "below" or "child", and "at the same level" or "sibling". A *tree visualization* is a graphical representation of a tree dataset.Tree visualizations of hierarchical data, common across many fields of study, are used for critical tasks ranging from the exploration of genetic data of species evolution in biology to the visual analysis of network activity in cybersecurity. As a result, a variety of tree designs are available at the disposal of designers. `Treevis.net` [43] catalogs over 300 techniques and categorizes them on geometric dimensionality ("2D", "3D","hybrid"), visual representation of hierarchy ("implicit", "explicit","hybrid") and node alignment ("axis-parallel", "radial", "free").

Data and tasks play an important role in choosing a tree visualization layout. For instance, if the practitioner wants to display the financial stock market and identify outlier stocks, they will commonly use a Treemap [21], as the market cap data can be aggregated to represent the value of a sector, and treemap representation facilitates the identification of extreme values. The tree visualization tasks are designed to acquire information about the structural and data attributes of a tree. The structural attributes provide information about the "topology" of a tree, and the data attributes provide information about the data associated with the nodes and links of the tree.

Despite the pervasiveness of tree visualizations and the importance of tree visualization tasks for effective visualization encoding, there is little formal theory in the field of data visualization to support the effective characterization of the vast design space as well as the creation of these visualizations based on the data structure and importantly the tasks of the user. While there are many general task taxonomies and ideas of how visualizations are created to support a domain task, these frameworks lack the specificity to support trees. This lack of theory makes it difficult for visualization creators to fully characterize the task space that a tree visualization could support and to design and evaluate novel tree visualizations effectively. Therefore, in my thesis I present a survey of tree visualization tasks, and create a novel task abstraction framework for tree visualization tasks (Sec. 6). I use the task theory to develop guidelines for design of tree visualizations (Sec. 7) and use the guidelines for development of a recommendation system (Sec. 7).

3 Research Questions

As I described in the introduction, a core phase of visualization design is mapping data and task requirements of a visualization problem to suitable visual encodings and interaction techniques. The mapping of data and task requirements is an intensive process and requires a visualization practitioner to make a series of decisions to create the final visualization. In Fig. 1 (visualization design pipeline), we present four intermediate steps between the data and task requirements and the final visualization design. Each stage of the pipeline is critical. However, these steps are also prone to challenges like insufficient theoretical support for design, lack of clear design guidelines, and practical tools to assist users in visualization design. An error at any stage of the pipeline can propagate downstream, affecting the final visualization choice. My thesis's primary motivation is the identification of challenges that exist in the pipeline and the creation of theory, resources, and tools that can reduce the errors at the stages.

To identify with the limitations and propose suitable solutions, my dissertation addresses the following overarching question:

"How can we develop theory, identify visualization best practices, and build applications that enable visualization practitioners to create effective and expressive visualizations?"

More specifically, this overarching research question calls for visualization research contributions that can solve the challenges at each step of the visualization design pipeline (Fig. 1). To solve the overarching research questions, I divide the problem into four more specific research questions, where each question maps to a stage in the pipeline:

1. *RQ1: What are the shortcomings of existing visualization theory that can inhibit effective and expressive visualization design?*
   To answer this question, I present three visualization design studies. In these studies, I present how we can apply existing visualization theory to solve novel visualization problems. These studies primarily contribute a visualization tool that allows users to solve critical analytical tasks more effectively in their domain. Further, a posthoc analysis of these studies identifies the shortcomings associated with the existing visualization theory.

2. *RQ2: How can we fix the shortcomings of existing visualization theory while maintaining its advantages?*
   To answer this question, I present a methodology to extend existing visualization theory and resolve their shortcomings. For this research question, we focus on the extension of the task abstraction theory for tree visualizations. In this thesis, I focus on tree visualizations because they are widely applied in many application areas, such as biology, computer science, and geography, and were common across the design studies that we discussed in RQ1. More specifically, we found that the existing task abstraction theory for tree visualizations lacked specificity. Therefore, through this contribution, I enhance the specificity of task abstraction for tree visualization in visualization theory.

3. *RQ3: How can we generate and collect visualization design guidelines and ensure that the guidelines comprehensively map the task and data configurations in visualization theory to appropriate design encodings?*
   To answer this question, I present two empirical studies that generate novel visualization design guidelines for data glyphs and timelines. I also present a methodology to curate design guidelines for visualization techniques with previously published evalution results. For tree visualizaions, I collect visualization design guidelines from a survey of published studies. The tree visualization design guidelines are build over the novel task abstraction theory discussed in RQ2. Through these projects, I contribute empirical knowledge to visualization literature.

4. *RQ4: How can we create tools and systems to help visualization practitioners and researchers access the design guidelines and improve visualization literacy?*
   To answer this question, I present two visualization recommendation systems. The recommendation systems enable practitioners and researchers to choose effective visualization encoding or design based on the data and task requirements. Through these systems we contribute a method to communicate visualization design guidelines that can significantly reduce the human-centered shortcomings of the visualization design pipeline.

By answering these questions, I contribute knowledge to the visualization literature on how to make the visualization design process robust which leads to more effective and expressive information visualization design. My work pays special attention to tree visualization because of its pervasive nature in the data analysis and visualization field. However, this thesis's findings can potentially lead to a robust visualization design pipeline for encodings beyond tree visualizations.

4 Thesis Plan

4.1 Thesis Chapters

1. **Introduction:** This chapter will provide an overarching motivation for the thesis research questions and summarize its contributions.

2. **Related Work:** The related work chapter will provide the necessary background of the visualization design theory, describe the existing research in the space of visualization design guidelines curation, and discuss state-of-the-art visualization recommendation systems. This chapter will also explain the limitations of the current research and summarize how this thesis's contributions extend the community's knowledge in these areas.

3. **Novel Visualization Design Studies:** This chapter will present two novel data visualization design studies. The projects solve real-world visualization in domains of medical diagnosis and cybersecurity. I will describe the projects as sub-chapters in this thesis, as outlined below. The projects are used as case studies to demonstrate the advantage of using the Nested Model for Visualization Design and Validation (Munzner 2019) for visualization design studies and identify the shortcomings associated with the theory. This chapter will also discuss the practical challenge of choosing appropriate visualization design idioms due to the way design guidelines are curated and communicated in the visualization literature.

    1 **CerebroVis:** This chapter will present a novel abstract representation of cerebral arteries that is more accurate than the traditional 3D representation in the task of detecting cerebrovascular abnormalities. This chapter will also present a novel framing and definition of the cerebral artery system in terms of network and tree theory and characterize neuroradiologist domain goals as abstract visualization, tree comparison and network analysis tasks.

    2 **StrokeVis:** This chapter will present a novel system to diagnose stroke in patients. This novel representation is built over CerebroVis.

    3 **Segmentrix + Portola:** This chapter will present a novel visualization tool that allows cybersecurity analysts to analyze the hierarchical organization of resources and network connections within these resources at data centers. The visualization design focuses on revealing the anomalous network connection that may lead to network attacks. This chapter will also discuss the role of visualization theory that enabled the translation of cybersecurity analysts' goals to abstract visualization tasks.

4. **Extended Task Abstraction Framework:** This chapter describes the importance of task abstraction for designing and evaluating visualizations. Task abstraction allows visualization creators to abstract the domain-specific task requirements to abstract visualization specific goals. In a survey and meta-analysis of tree visualization tasks, I found that the existing task abstraction framework for trees only offers limited specificity to describe tree visualization tasks abstractly. Therefore, in this chapter, I describe a task abstraction framework for tree visualization tasks. To supplement the task abstraction, I also contribute a tree visualization task dataset. The task dataset consists of over 200 tasks. All tasks in the dataset were abstracted with the novel abstraction framework and analyzed to better understand the state of tree visualizations. These abstracted tasks can benefit visualization researchers and practitioners as they design evaluation studies or compare their analytical tasks with ones previously studied in the literature to make informed decisions about their design.

5. **Guidelines for Visualization Design:** In this chapter, I focus on two methods for building design guidelines. First, I create design guidelines for encodings that do not have existing empirical studies. In this thesis, I present two novel empirical studies that allowed me to build design guidelines for visualizing multi-dimensional data as glyphs and presenting temporal event data as timelines. For visualization techniques with evaluations, such as tree visualization, we curate guidelines by surveying results from published studies and conducting a meta-analysis to identify the general trends and patterns in the results. In the sub chapters of the thesis, I will describe the individual projects.

    1 **Evaluating the Effect of Data Glyphs on Probabilistic Categorization Task:** In this chapter, I will present an empirical study that measures the effect of representing multidimensional data as glyphs for a probabilistic categorization task. This study contributes guidelines for the effective use of glyph designs.

    2 **Evaluating the Effect of Timeline Shape on Visualization Task Performance:** In this chapter, I will present an empirical study that measures the effect of different timeline shapes, such as linear, circular, and spiral, for representing temporal events data. This study contributes guidelines for the effective use of timeline shapes.

    3 **Collecting and Curating Task Based Design Guidelines for Tree Visualizations:** In this chapter, I will present a methodology to curate design guidelines for tree visualizations from previously published empirical studies. I will also discuss the task-based challenges associated with design guidelines curation.

6. **Visualization Recommendation Systems:** This chapter describes two novel visualization recommendation systems. I will discuss the recommendation systems in detail as sub-chapters of this thesis. I will conclude this chapter with a discussion on common design patterns that emerged from the design of recommendation systems for two very different problems.

    1 **Genorec:** Genorec recommends visualization to biologists and data analysts working with genomics data. In this chapter, I will present the methodology to design and develop a knowledge-based recommendation system for genomics.

    2 **Treevis Recommendation System:** Treevis recommendation system suggests appropriate tree visualization encoding to visualization practitioners or researchers. In this chapter, I will present in detail the methodology to develop a task based recommendation system for tree visualizations.

7. **Discussion:** In this chapter, I will reflect on the challenges faced in solving the research questions and the lessons learned in the process. In the discussion chapter, I will also discuss how we take the lessons learned from this thesis and apply it more broadly in extending visualization theory, curating design guidelines, and building visualization recommendation systems.

8. **Conclusion:** This chapter will reiterate the thesis's primary contributions and argue that I have made significant contributions to the topics of information visualization design and recommendation. I will also urge the community to build on these contributions and make the process of designing and developing visualizations accessible and effective.

4.2 Thesis Completion Timeline

I anticipate to finish my thesis in Spring 2022. Fig. 2 shows projects and the time-frame to complete the projects. Each project's publication status is available in the column "Status" of Fig. 2. For unpublished work, I present the estimated time-frame to complete the research and the tentative publication venue.

Milestone	Status	Spring 2021	Summer 1 2021	Summer 2 2021	Fall 2021	Spring 2022
Chapter 1 (Introduction)		**Write** the chapter				
Chapter 2 (Related Work)		**Write** the chapter				
Chapter 3 (Visualization Design Studies)				**Write** the chapter		
Cerebrovis	**Paper** IEEE Vis 2019, TVCG					
Strokevis			**Research**			**Submit** to IEEE Vis 2022
Segmentrix + Portola	**Poster** IEEE Vizsec 2019 Best Poster		**Submit** to IEEE Vis 2021 Short Paper			
Chapter 4 (Task Abstraction Theory)				**Write** the chapter		
Tree Visualization Task Survey	**Paper** Conditional Accept with Major Revision TVCG					
Chapter 5 (Design Guidelines)					**Write** the chapter	
Glyph Evaluation	**Poster** IEEE Vis 2019					
*Timeline Shape Evaluation**	**Paper** CHI 2020					
Design Guidelines for Tree Vis		**Research**			**Submit** to CHI 2022	
Challenges for Curating Guidelines	**Paper** BELIV Workshop 2020					
Chapter 6 (Visualization Recommendation Systems)						**Write** the chapter
Genorec	**Poster** IEEE Vis Best Poster 2020	**Submit** to IEEE Vis 2021				
Treevis Recommendation				**Research**		**Submit** to IEEE Vis 2022
Chapter 7 (Discussion)						**Write** the chapter
Chapter 8 (Conclusion)						**Write** the chapter

Figure 2: Thesis Milestones and Timeline. In the timeline shape evaluation(*) paper, I was the second author. Therefore, I will discuss the parts relevant to my thesis for the timeline evaluation paper.

5 Novel Visualization Design Studies

In this section, I summarize the design study projects of my thesis. These studies helped in identifying the challenges of visualization design pipeline. For each project, I provide an overview of the main contribution, the results and current publication status.

5.1 CerebroVis and StrokeVis

Summary: Arteries in the human brain form a network of blood flow, and a blockage or leakage in this network can lead to life-threatening cerebrovascular conditions such as a stroke or aneurysm. Conventional

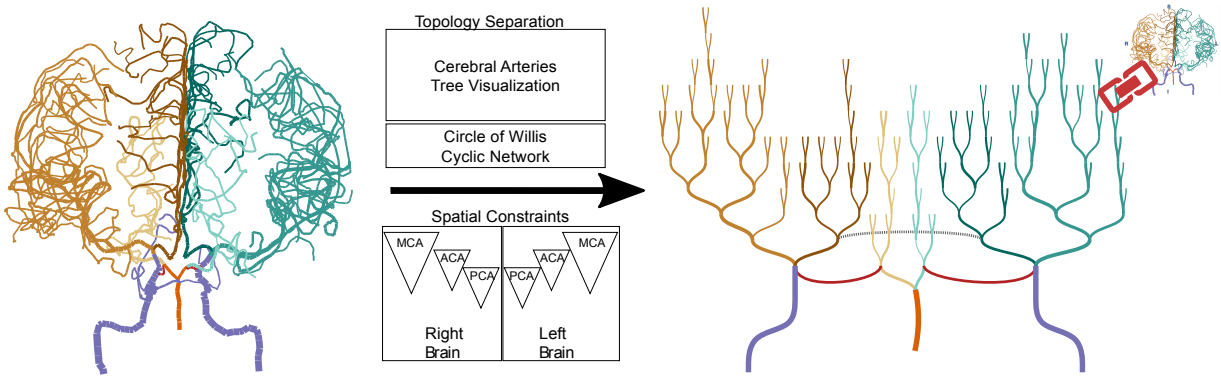

Figure 3: CerebroVis is a novel network visualization for cerebral arteries. CerebroVis uses an abstract topology-preserving visual design which is put in spatial context by enforcing constraints on the network layout. Here we show the conversion of an almost symmetrical healthy human brain cerebral artery network from a 2D isosurface visualization (left) to CerebroVis (right)

diagnostics rely on an expert neuroradiologist identifying vascular abnormalities through examination of medical images (e.g., CTA, MRA). This data is commonly rendered in 3D to assist the doctor with identification of the abnormalities. However, prior research indicates that existing representations of the 3D cerebral arteries—e.g., isosurface, volume rendering, and Maximum Intensity Projection (MIPS)—introduce visual artifacts and task performance challenges such as overplotting/occlusion [14], false impression of geometry [14], and excessive artery bends. In this design study, we present **CerebroVis** a novel 2D visualization of the cerebral artery system with spatial context to assist doctors in the identification of cerebrovascular abnormalities. The abstract visualization enables increased domain task performance over 3D geometry representations, while including spatial context helps preserve the user's mental map of the underlying geometry.

Evaluation and Results: We evaluate our new layout and the accompanying CerebroVis prototype in two ways: (1) assessing the robustness of the technique by examining 61 healthy brain scans and (2) a mixed methods study with three neuroradiologists which included semi-structured interviews and a controlled experiment simulating intracranial stenosis diagnosis. We found that our layout and implementation correctly visualizes all 61 brain scans, that neuroradiologists were more accurate at identifying stenosis with CerebroVis vs. a 3D visualization (absolute risk difference 13%), and that neuroradiologists thought CerebroVis was easy to understand and a useful addition to their diagnosis toolbox.

Status CerebroVis design study paper was accepted at IEEE Vis 2019 and published in TVCG journal in 2020. The follow-up work StrokeVis is an upcoming project. StrokeVis will build over CerebroVis layout and focus on detection of cerebral stroke.I plan to conduct the research for StrokeVis in Summer of 2021 and submit it to Vis 2022.

5.2 Segmentrix+Portola

Summary: Micro-Segmentation enables organizations to logically divide the datacenter into distinct security segments down to the individual workload level, and then define security controls for each unique segment. Tree and Network visualizations play a critical role in the development and maintenance of segmentation. In an unsegmented network, a network visualization of workload communication can help domain users assess dependencies and create segmentation policies. Whereas, in segmented networks, the visualization of traffic between individual workloads and segmented groups can be essential for monitoring security compliance. To assist cybersecurity analysts we developed Segmentrix+Portola. Segmentrix is an adjacency matrix-based tool for developing and monitoring micro-segmentation strategies. This representation is scalable, readable, and provides visibility into the entire datacenter network of large organizations. Portola is a radial layered tree visualization, it represents the organization of workloads in a datacenter and assists analysts to visualize traffic through different ports of the network.

Figure 4: Segmentrix+Portola

Evaluation and Results: Segmentrix+Portola was developed over three months at Illumio Inc., with close domain collaboration. Through expert interviews, we found that the visualization system helped find anomalies in network connection and understand the organization of workloads in a datacenter. Segmentrix and Portola were evaluated independently. In the future, we will assess the tools combined as one system.

Status Segmentrix was accepted as a Poster at Vizsec 2019. It also received the Best Poster Award for Vizsec 2019. I plan to submit a short paper to Vis 2021 which will include both Segmentrix and Portola.

6 Extended Task Abstraction Framework

From the design studies discussed in Sec. 5, I learned that the existing task abstraction theory lacks specificity to abstract tree visualization tasks. For instance, in Cerebrovis, a key task is to analyze symmetry or balance between the left and right brain arteries. However, the existing task abstraction taxonomies do not have the capability to describe the abstraction of a tree balance task. The existing task abstraction frameworks limit visualization researchers and practitioners' ability to define a domain-centric tree visualization task as a well-specified abstract tree visualization task and discover appropriate tree visualization encodings. This section presents a novel extension of the Multi-LevelTask Typology (MLTT) to accurately abstract and analyze tree visualization tasks.

Summary: In the field of information visualization, the concept of "tasks" is an essential component of theories and methodologies for how a visualization researcher or a practitioner understands what tasks a user needs to perform and how to approach the creation of a new design. In this project, I focus on the collection of tasks for tree visualizations, a common visual encoding in many domains ranging from biology to computer science to geography. In spite of their commonality, no prior efforts exist to collect and abstractly define tree visualization tasks. I present a literature review of tree visualization papers and generate a curated dataset of over 200 tasks. To enable effective task abstraction for trees, I also contribute a novel extension of the Multi-Level Task Typology to include more specificity to support tree-specific tasks as well as a systematic procedure to conduct task abstractions for tree visualizations. All tasks in the dataset were abstracted with the novel typology extension and analyzed to gain a better understanding of the state of tree visualizations. These abstracted tasks can benefit visualization researchers and practitioners as they design evaluation studies or compare their analytical tasks with ones previously studied in the literature to make informed decisions about their design. I also reflect on our novel methodology and advocate more broadly for the creation of task-based knowledge repositories for different types of visualizations.

Results: Fig. 5 presents the novel tree-specific extension to the MLTT. The tasks of tree visualization tasks broken down by *Actions* and *Targets*. The Actions use the Multi-Level Task Typology terminology to identify the types of actions users can perform in tree visualization tasks. The Targets include a novel *Nested-extension* of the existing MLTT target characterization that adds specificity for tree visualizations.

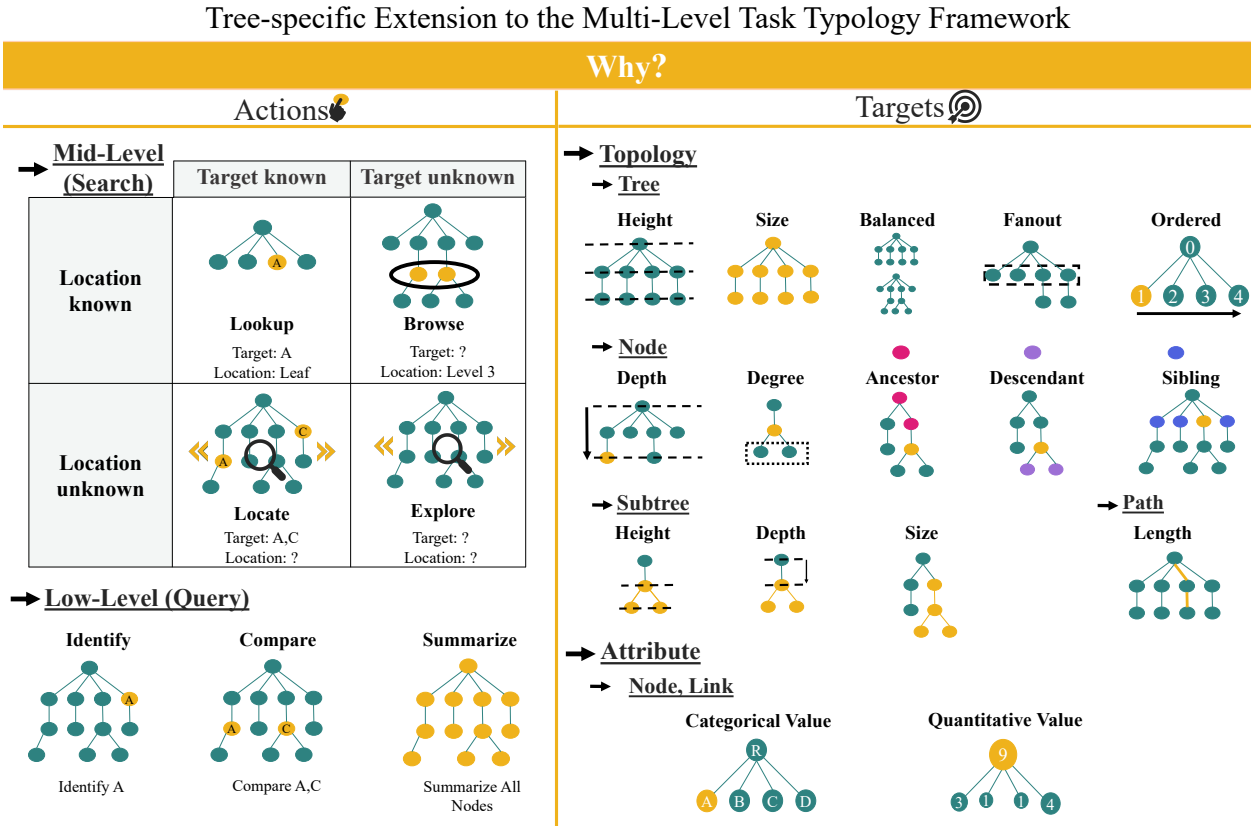

Figure 5: Task Abstraction Framework for Trees

To demonstrate the advantage of the extended task abstraction framework, let us consider the following examples: A financial analyst may want to analyze performance of stocks in different sectors of the market, particularly in the aftermath of a major global crisis to identify outliers that survived recessions and performed well despite an economic slowdown. In another instance, an epidemiologist may want to analyze the patterns in the branching of a virus strain and compare how different strains evolve over time. Applying the MLTT framework for the abstraction for the examples leads to the same abstraction result: the goal of the task is to perform "lookup" on the "topology". However, in practice the two tasks are fundamentally different. The differences become apparent when we use the extended MLTT framework (Figr. 5), and the abstraction reveals that the financial analyst is interested in "looking up" the "ancestor" nodes of outlier stocks in the stock market tree, and the epidemiologist wants to "compare" properties like "height" and "fanout" of different branches in the tree. Therefore, the novel task abstraction framework enables tree visualization practitioners to be more specific with task abstractions, leading to effective design choices.

Status: The complete report of the survey is under review for TVCG journal and is available on our survey website: `https://intervis-projects.ccs.neu.edu/Tree-Visualization-Survey/`.

7 Guidelines for Visualization Design

This section presents design guidelines for three visualization techniques: data glyphs, timelines, and trees. In my research, I found that there are two common methods to build design guidelines. The first method involves conducting an empirical study, where the researcher evaluates a visualization technique with a set of tasks to identify their effectiveness [12, 26]. In the second method, the researcher surveys previously published studies, and through a meta-analysis of the results, curates guidelines for visualization techniques [17]. I summarize two novel empirical studies for developing guidelines for data glyphs and timelines. There are many tree visualization evaluation studies in information visualization research (for e.g. [26, 47, 8]). However, the task-

based design guidelines for tree visualizations are not yet curated [39]. Therefore, for tree visualizations, I describe my plan for curating task-based design guidelines.

7.1 Data Glyphs

Summary: Categorization involves classifying objects based on their features. For example, sorting laundry before putting it in the wash by color and material. Glyphs are visualizations well suited to represent categorization data because they are used for multidimensional data in which dimensions are encoded to marks in the visual or pictorial representation. However, there has been no study or systematic evaluation of how to best encode probabilistic categorization data with glyphs to date. In this study, I evaluate whether the visual representation of the data affects categorization accuracy? And if so, how should the data be visually represented to maximize categorization accuracy? Previous research work has demonstrated that the inclusion of representations of people and human faces results in a significant improvement in memorability for natural images [20] and data visualizations [4]. Research in visualization has also shown that pictorial data encodings are recalled more accurately than simple bar charts [18] when working memory is under heavy load. Consequently, we hypothesized that a memorable glyph representation would result in a higher categorization accuracy for the second question. To test the hypothesis, we evaluated the effectiveness of anthropomorphic (human-like) glyphs as compared to abstract glyphs. In order to evaluate the effect of glyph representation on categorization accuracy, we conducted a within-subject study with 480 participants on Amazon's Mechanical Turk. Each participant completed a probabilistic categorization task with two of four different glyph designs each of which encode 3 probabilistic features. Two of the glyphs were of abstract design and two of the glyphs were human-like so that we could observe whether there was a positive benefit to the more memorable anthropomorphic glyphs, see Fig. 6 (Glyphs Evaluated).

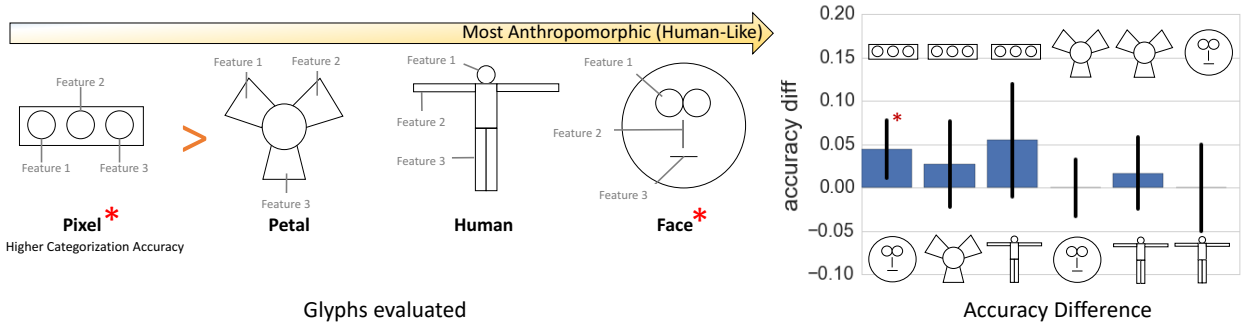

Figure 6: The left figure shows the glyphs evaluated in our study. The pixel and petal glyphs are abstract glyph designs and the human and face are anthropomorphic glyph designs. The figure on the right is a summary of the average differences in accuracy between each pair of glyphs evaluated in the study. On the y-axis, positive ratios denote that glyphs on the top of the chart had greater accuracy, and visa versa for negative. For each glyph comparison, the 99.9% C.I. is plotted, and asterisks (*) denote Bonferroni-corrected significance in accuracy of one stimulus over other.

Results: Contrary to our hypothesis, I found participants were significantly more accurate with abstract than anthropomorphic glyphs, see Fig. 6 (Accuracy Difference). The Pixel glyph visual encoding generated the most accurate categorization performance and lead to statistically significantly higher accuracy than the Face glyph. In addition, participants felt less confident with anthropomorphic glyphs in comparison to the abstract glyphs when performing the categorization task.

Status: he research is complete and the paper is written. The work received initial validation from the community as it was accepted as a poster at IEEE Vis 2019. I plan to submit the paper to a journal.

7.2 Timelines

Summary: A timeline is a visual representation of a series of events in time. Timelines have become prevalent in our daily lives as the de facto representation to show financial trends, weather details, and meeting schedules. Timelines are most commonly drawn linearly [7], where the events are organized along a straight line. In practice, however, we can find abundant examples of timelines where events are arranged in non-linear shapes like circles, spirals, grids, and other arbitrary arrangements [7]. The visualization literature provides sufficient evidence that the layout and orientation of visualizations affect user's analytical task performance [11]. However, existing work in timeline visualization evaluation has not measured the impact of timeline shape alone on user task performance for general temporal event sequence data. In this paper, I present the first study which evaluates the readability of timeline shape alone on user task performance for general temporal event sequence data. In a crowd-sourced experiment, I compare 4 timeline shapes — horizontal line, vertical line, circle, and spiral — using 3 types of temporal data — recurrent, non-recurrent, and mixed. Our study is carefully designed to evaluate timeline shapes using common everyday tasks with familiar-looking datasets. E.g., find the date associated with an historical event on a timeline or lookup your daily schedule to find what are you supposed to do tonight at 8pm. In a within-subjects study design, I measured time to complete a visualization task and the task accuracy across the 4 timeline shapes.

Results: There was evidence that task completion time is dependent on the choice of the timeline shape. Specifically, linear shapes were on an average faster to read. But, no evidence supports that users' accuracy is affected by the shape of timeline. Additionally, there was a strong preference for linear timeline among the participants.

Status: This evaluation study was accepted at CHI 2020.

7.3 Tree Visualizations

Motivation and Preliminary Research: Task-based guidelines are critical for the design of a visualization. For instance, if the user's task is to visualize the tree topology, then the preferred visualization design choice should be a node-link tree visualization as they show the topology of the tree explicitly. In information visualization theory, there are many task-based tree visualization evaluation papers (for e.g. [26, 47, 8]). However, these studies have never been collated and analyzed to create consistent task-based design guidelines for tree visualizations. Lack of task-based design guidelines inhibits visualization practitioners from navigating through the design space of tree visualizations and identifying the most effective visual encodings based on their tasks. Therefore, I plan to conduct a meta-analysis of the published empirical studies for tree visualizations and use them to generate task-based design guidelines. In my previous work, as discussed in Sec. 6, I surveyed empirical studies and also identified an exhaustive list of tree visualization tasks. Both these contributions are critical for curating the design guidelines, as explained in the research plan.

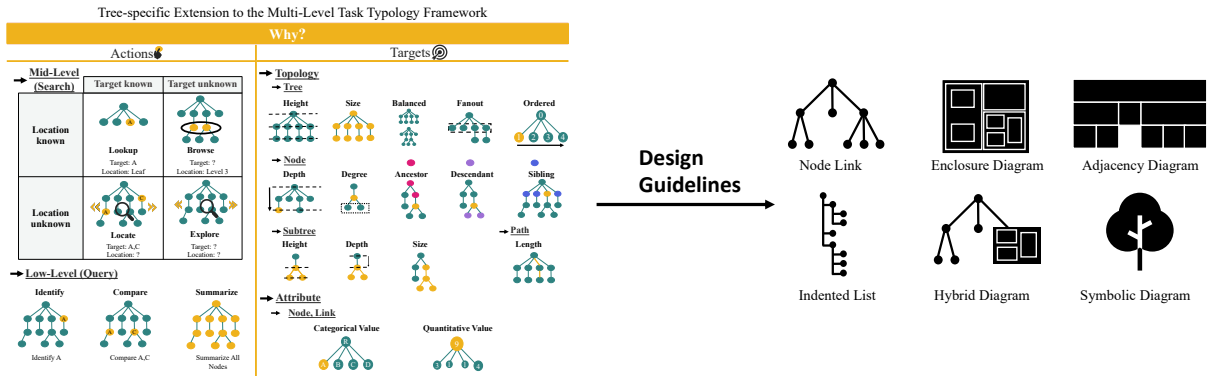

Figure 7: Task-based design guidelines help visualization practitioners to choose the most appropriate visualization encoding based on tasks of the user.

Research Plan: Task-based design guidelines assist visualization practitioners in choosing an appropriate visualization given the task requirements of the user (see Fig. 7). To build the guidelines, I will collect the tasks, the visualization encoding, and results from tree visualization empirical studies. After collecting the tasks, I will abstract the tasks using the Tree-specific task abstraction framework developed (see Fig. 5). The abstraction will enable the comparison of tasks collected from different studies. After the abstraction, I will identify the most appropriate visualization encoding for each task type based on the results of the empirical studies. To find the most appropriate visualizations, I will rank visualization techniques for each task based on consensus. The consensus in ranking will be reached if most empirical results favor a particular visualization encoding for a particular task. For tasks that do not have published results or ambiguous results, the ranking will be built through expert ranking. The expert ranking method was previously used by Nobre et al. [34] to develop a recommendation wizard for Multi-variate network visualizations. In this method, a group of experts ranks visualization encodings based on their task suitability. The experts use their research experience to perform the ranking. The expert team will include expert tree visualization researchers: Michelle Borkin and John Alexis Guerra Gomez for tree visualizations.

Expected Outcomes: This work will yield design guidelines that will help researchers, designers, and students to create tree visualizations. The guidelines will also help in the development of a tree visualization recommendation system. The guidelines will be the core of the recommendation model, which will suggest tree visualization encoding to the users. I will discuss the recommendation system in Sec. 8.

8 Visualization Recommendation Systems

In this section, I will present two visualization recommendation projects. The first project is a recommendation system for genomics visualizations. The genomics recommendation project is in the last stage of development and validation. I have already shared the initial results with the visualization community by publishing a poster. Therefore, I will discuss the completed research and provide information about how I plan to conduct the remaining work for the genomics visualization recommendation system. In this section, I will also present a recommendation system for tree visualizations. The tree visualization recommendation system project builds over my research in Sec. 6 (Task-Abstraction Theory) and Sec. 7 (Tree Visualization design guidelines). The task abstraction theory helps in guidelines research, and the guidelines will ultimately form the knowledge for the recommendation system. Since I have not started my work on the tree visualization recommendation project, I describe the research plan, expected outcome, and validation plan in detail.

8.1 Recommendation System for Genomics Visualization

Summary: Visualization tools and techniques play a significant role in the workflow of genomics researchers, and they are regularly employed in the interpretation of genomics data. However, the vast majority of genomics researchers have little or no formal training in data visualization design. Therefore, they require guidance on designing effective visualizations for a given set of data and analysis tasks. In this work, I present a knowledge-based recommendation system for genomics visualization. The system allows genomics researchers to navigate through a selection of visualization options and identify the techniques that meet their preferences and requirements. The first step in building the recommendation system was to identify the data structures, analytical tasks, and visualization designs used in genomics analysis. The required information was gathered from the survey paper by Nusrat et al. [35], where the authors contributed a data, task, and visualization taxonomy for genomics visualization. Next, I characterized the typical design workflow of a genomics visualization. As shown in Fig. 8 (B), to create a genomics visualization, a designer needs to make several design decisions like the choice of marks and channels to encode genomics data or how to layout the marks and channels. Our analysis found that design stages are sequential, meaning each step feeds into the next one. For instance, the choice of alignment depends on the choice of layout. The third step in creating a recommendation model was identifying design guidelines that inform the selection of a visualization. Design guidelines for our recommendation models are derived from general visualization graphical, and perception studies and analysis of genomics visualization literature published at visualization conferences.

Input and Output of the Recommendation System: Both data features and types of analytical tasks are inputs to the recommendation model. Data information is either provided explicitly by users or

collected automatically from the standard file formats used for genomics data. In addition to data-driven recommendations, our system supports task-based recommendations. Unlike data descriptions, tasks that users are intended to perform are difficult to infer, which requires task descriptions to be explicitly specified by users. The recommendation system's output includes a custom interactive visualization implementation and a list of existing genomics visualization tools and libraries that match the user's data and task requirements.

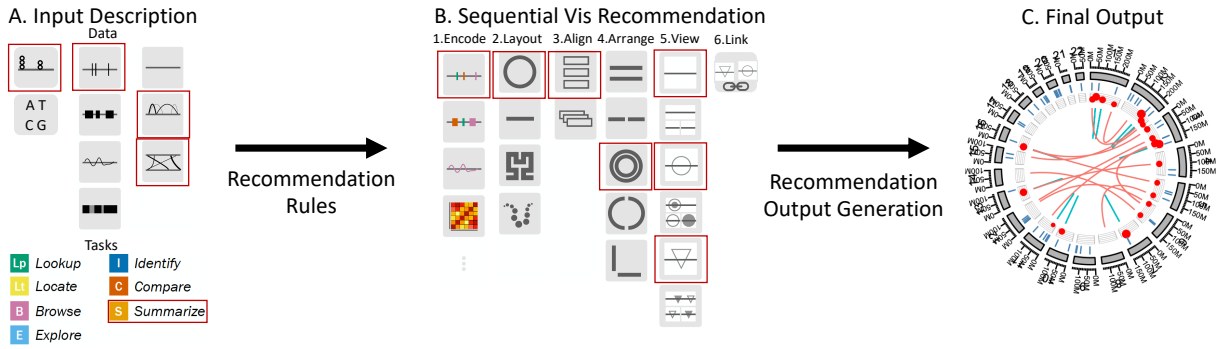

Figure 8: A schematic representation of the three stages of a visualization recommendation system for genomics data. The input visually presents the data and task specification for the system. Based on data and task requirements, the sequential recommendation model identifies suitable visual encoding at each step of the visualization grammar, and the output visualization represents the final deliverable to the user for their analysis.

Status: This work was accepted as a poster at IEEE Vis 2020 [38] and received the "Best Poster Award" at the conference. The poster gave an overview of the system design and explained in detail the recommendation model. I plan to submit this work as a full paper to IEEE Vis 2021. For the submission, I am working on developing the front-end of the recommendation system. The recommendation system's front-end will allow genomics researchers to provide their data and task specification as input to the recommendation model. The front-end will also enable the users to analyze the recommended genomics visualization. After completing the front-end development, I will evaluate the effectiveness and expressivity of the recommendation model output. To assess the effectiveness, I will carry out a case-study based evaluation with domain experts. In the assessment, domain experts will use the recommendation system and evaluate the visualization output's accuracy. I will also observe if the recommendation system helped the domain experts find an accurate but unexpected visualization output. This question will help us evaluate the serendipity and pedagogical factors of the system. To assess the expressivity, I will try creating the common genomics visualization identified in the survey paper by Nusrat et al. [35]. If the recommendation system covers many visualization techniques from the survey, it will indicate high expressivity of the recommendation system.

8.2 Recommendation System for Tree Visualization

Motivation & Preliminary Results: In Sec. 7 (Tree Visualizations), I discussed our plan for developing task-based design guidelines for tree visualizations. The design guidelines are theoretical and will be communicated to visualization creators in the 'printed on paper' form as a research paper. However, guidelines available in research form may not be easy to access or apply, see Sec. 1. Therefore, through my work, I want to make the guidelines actionable and usable by visualization creators. To make the design guidelines actionable and easily applicable, I will create a system to recommend a tree visual encoding based on a user's data and tasks. My ongoing research work on the genomics visualization recommendation system is actively helping me learn valuable system design and engineering implications for the new proposed work, including user interface designs for the recommendation tool, and techniques to decouple the recommendation engine or model from the interface to increase the portability of the recommendation engine into different programming languages such as Python and Javascript. I plan to use the skills of designing a recommendation system from genomics visualization research and combine it with the theoretical design guidelines of creating tree visualizations identified in Sec. 7 (Tree Visualizations) to develop a tree visualization recommendation system.

Research Plan: The core of recommendation system is the recommendation model. The recommendation model will contain rules to support mapping of data and task requirements to appropriate visualization encodings. The rules will be generated from design guidelines discussed in Sec. 7. The recommendation model will take in the tree data type and the user's already abstracted tasks and then generate a ranked list of supporting tree visualization encoding techniques with the help of visualization design base. For example, if node data is "aggregatable" and the task is to "identify outliers", then use a "treemap" visualization. The recommendation rules can be stored externally in a text file. Storing recommendation rules externally will allow easy update of the rules in the future. For the recommendation, I will also implement an algorithm to apply the rules and generate a ranked list of visualization recommendation. The recommendation algorithm will be implemented in Javascript to ensure seamless integration with the web-based recommendation tool. The recommendation system will be implemented as a system with a website front-end UI to allow a user to input their tree data structure and abstract tasks and get out a series of visual encoding recommendations. We will provide an interactive user interface to help users determine their data and task requirements and select them in a step-by-step guided input to the recommendation system.

Expected Outcomes: Accomplishment of this project will yield a new recommendation system built on the design guidelines derived from visualization evaluation papers, to recommend tree visual encoding techniques. The underlying code of the recommendation system, including the underlying algorithm, will be made open source and publicly available with supporting documentation. The system will be implemented with a front-end website to serve as the user interface. This system will help researchers, designers, and students alike be able to more easily navigate the vast tree visualization space, create tree visualizations, reduce implicit bias in the visualization design process, and support visualization education.

Validation Plan: I will conduct a user study with quantitative and qualitative metrics to measure confidence, trust, novelty, and serendipity with use of the new system. The new recommendation system will be evaluated and validated by visualization experts. We need visualization experts to evaluate the system because I am using visualization theory to build the recommendation model, and approval from the visualization experts will validate that the system is using appropriate visualization rules. Since the system is still under research stage, the exact plan of the evaluation will be determined closer to the study. However, the broad goal of the evaluation will be to understand if the output of the recommendation system is accurate and matches expectations of visualization creators. Furthermore, I will also like to study if the recommendation system can help visualization creators find visualization techniques they were not previously familiar with but the recommendation system helped them find it. This will be an online study without any in-person component.

9 Conclusion

In my thesis, I present theoretical knowledge and practical tools to improve visualization design for practitioners and researchers. I use the case of tree visualizations as the overarching theme to identify several major shortcomings of our theoretical knowledge about visualization design and propose solutions to overcome the existing challenges. Till this point in my research, I have solved novel visualization design problems, and in the process, identified challenges with the existing theoretical resources for designing tree visualizations. I have used these challenges to improve the theoretical support to abstract tree visualization tasks, which ultimately support the selection of accurate tree visualization encoding. Currently, I am working on using the new tree visualization theory and a survey of tree visualization studies to build tree visualization design guidelines. I plan to finish the development of guidelines by the Summer of 2021. With the new tree visualization design guidelines, I plan to develop a system for tree visualization recommendation. As the design guidelines are a prerequisite to the recommendation system, I anticipate to start the work on the tree visualization recommendation system in the Fall of 2021 and finish the work in the Spring of 2022. My work on tree visualizations has constantly drawn support from my other research projects, which also fall in the general theme of improving the visualization design pipeline. In my thesis, I present task-based visualization design guidelines for data glyphs and timelines. The design guidelines project inspired me to survey for task-based guidelines for tree visualizations. In the survey, I found that tree visualization design guidelines are not well-curated and need additional work. I also anticipate my genomics visualization recommendation system will help me design and develop the tree visualization recommendation system.

Motivated by the goal to improve the visualization design pipeline, I will make the following contributions. Through a series of practical visualization design studies, I identify the existing challenges of creating visualizations. The design studies also solve visualization design problems in the critical fields of medical diagnosis and cybersecurity. Next, I contribute theory to abstract tree visualization tasks, with a dataset for visualization creators to analyze tasks and their abstraction of over 200 tasks. I also contribute task-based design guidelines for tree visualizations, data glyphs, and timelines. Finally, to apply visualization design guidelines in a practical context, I develop visualization recommendation systems. These contributions help the visualization community understand the existing problems with the visualization design pipeline and contribute knowledge and resources to eliminate the challenges.

References

[1] Bilal Alsallakh et al. "Visualizing Sets and Set-typed Data: State-of-the-Art and Future Challenges". In: *EuroVis - STARs*. Ed. by R. Borgo, R. Maciejewski, and I. Viola. The Eurographics Association, 2014. ISBN: -. DOI: 10.2312/eurovisstar.20141170.

[2] R. Amar, J. Eagan, and J. Stasko. "Low-level components of analytic activity in information visualization". In: *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.* Oct. 2005, pp. 111–117. DOI: 10.1109/INFVIS.2005.1532136.

[3] Gennady Andrienko et al. "A conceptual framework and taxonomy of techniques for analyzing movement". In: *Journal of Visual Languages & Computing* 22.3 (2011), pp. 213–232. DOI: 10.1109/VL.1996.545307.

[4] Michelle A. Borkin et al. "What Makes a Visualization Memorable?" In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (Dec. 2013), pp. 2306–2315. ISSN: 1077-2626. DOI: 10.1109/TVCG.2013.234. URL: http://dx.doi.org/10.1109/TVCG.2013.234.

[5] M. Bostock, V. Ogievetsky, and J. Heer. "D^3 Data-Driven Documents". In: *IEEE Transactions on Visualization and Computer Graphics* 17.12 (Dec. 2011), pp. 2301–2309. ISSN: 1941-0506. DOI: 10.1109/TVCG.2011.185.

[6] M. Brehmer and T. Munzner. "A Multi-Level Typology of Abstract Visualization Tasks". In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (Dec. 2013), pp. 2376–2385. ISSN: 2160-9306. DOI: 10.1109/TVCG.2013.124.

[7] M. Brehmer et al. "Timelines Revisited: A Design Space and Considerations for Expressive Storytelling". In: *IEEE Transactions on Visualization and Computer Graphics* 23.9 (2017), pp. 2151–2164. ISSN: 1077-2626. DOI: 10.1109/TVCG.2016.2614803.

[8] M. Burch et al. "Evaluation of Traditional, Orthogonal, and Radial Tree Diagrams by an Eye Tracking Study". In: *IEEE Transactions on Visualization and Computer Graphics* 17.12 (2011), pp. 2440–2448.

[9] N. Cawthon and A. V. Moere. "The Effect of Aesthetic on the Usability of Data Visualization". In: *2007 11th International Conference Information Visualization (IV '07)*. 2007, pp. 637–648. DOI: 10.1109/IV.2007.147.

[10] M. Chen et al. "Pathways for Theoretical Advances in Visualization". In: *IEEE Computer Graphics and Applications* 37.4 (2017), pp. 103–112. ISSN: 1558-1756. DOI: 10.1109/MCG.2017.3271463.

[11] William S. Cleveland and Robert McGill. "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods". In: *Journal of the American Statistical Association* 79.387 (1984), pp. 531–554. ISSN: 01621459. URL: http://www.jstor.org/stable/2288400.

[12] Sara Di Bartolomeo et al. "Evaluating the Effect of Timeline Shape on Visualization Task Performance". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. Honolulu, HI, USA: Association for Computing Machinery, 2020, pp. 1–12. ISBN: 9781450367080. DOI: 10.1145/3313831.3376237. URL: https://doi.org/10.1145/3313831.3376237.

[13] A. Diehl et al. "Visguides: A Forum for Discussing Visualization Guidelines". In: *Proceedings of the Eurographics/IEEE VGTC Conference on Visualization: Short Papers*. EuroVis '18. Brno, Czech Republic: Eurographics Association, 2018, pp. 61–65.

[14] Elliot K Fishman et al. "Volume Rendering versus Maximum Intensity Projection in CT Angiography: What Works Best, When, and Why". In: *Radiographics* (2006). DOI: https://doi.org/10.1148/rg.263055186.

[15] *FLOWINGDATA, url = https://flowingdata.com/about/.*

[16] *From data to Viz, url = https://www.data-to-viz.com/.*

[17] J. Fuchs et al. "A Systematic Review of Experimental Studies on Data Glyphs". In: *IEEE Transactions on Visualization and Computer Graphics* 23.7 (July 2017), pp. 1863–1879. ISSN: 1941-0506. DOI: 10.1109/TVCG.2016.2549018.

[18] Steve Haroz, Robert Kosara, and Steven L. Franconeri. "ISOTYPE Visualization: Working Memory, Performance, and Engagement with Pictographs". In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. CHI '15. Seoul, Republic of Korea: Association for Computing Machinery, 2015, pp. 1191–1200. ISBN: 9781450331456. DOI: 10.1145/2702123.2702275. URL: https://doi.org/10.1145/2702123.2702275.

[19] Kevin Hu et al. "VizML: A Machine Learning Approach to Visualization Recommendation". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, pp. 1–12. ISBN: 9781450359702. DOI: 10.1145/3290605.3300358. URL: https://doi.org/10.1145/3290605.3300358.

[20] Phillip Isola et al. "Understanding the Intrinsic Memorability of Images". In: *Advances in Neural Information Processing Systems 24*. Ed. by J. Shawe-Taylor et al. Curran Associates, Inc., 2011, pp. 2429–2437. URL: http://papers.nips.cc/paper/4451-understanding-the-intrinsic-memorability-of-images.pdf.

[21] B. Johnson and B. Shneiderman. "Tree-maps: a space-filling approach to the visualization of hierarchical information structures". In: *Proceeding Visualization '91*. Oct. 1991, pp. 284–291. DOI: 10.1109/VISUAL.1991.175815.

[22] Pawandeep Kaur and Michael Owonibi. "A Review on Visualization Recommendation Strategies". In: Feb. 2017. DOI: 10.5220/0006175002660273.

[23] Stephan Kerpedjiev et al. "AutoBrief: A Multimedia Presentation System for Assisting Data Analysis". In: *Comput. Stand. Interfaces* 18.6–7 (Dec. 1997), pp. 583–593. ISSN: 0920-5489. URL: https://doi.org/10.1016/S0920-5489(97)00022-6.

[24] N. Kerracher, J. Kennedy, and K. Chalmers. "A Task Taxonomy for Temporal Graph Visualisation". In: *IEEE Transactions on Visualization and Computer Graphics* 21.10 (Oct. 2015), pp. 1160–1172. ISSN: 1941-0506. DOI: 10.1109/TVCG.2015.2424889.

[25] David M. Klumpar, Kevin Anderson, and Avangelos Simoudis. "RAVE: Rapid Visualization Environment". In: 1994.

[26] A. Kobsa. "User Experiments with Tree Visualization Systems". In: *IEEE Symposium on Information Visualization*. Oct. 2004, pp. 9–16. DOI: 10.1109/INFVIS.2004.70.

[27] Bongshin Lee et al. "Task Taxonomy for Graph Visualization". In: *Proceedings of the 2006 AVI Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization*. BELIV '06. Venice, Italy: Association for Computing Machinery, 2006, pp. 1–5. ISBN: 1595935622. DOI: 10.1145/1168149.1168168.

[28] Manuel Lima. *The book of trees: visualizing branches of knowledge*. Princeton Architectural Press, 2014.

[29] Jock Mackinlay. "Automating the Design of Graphical Presentations of Relational Information". In: *ACM Trans. Graph.* 5.2 (Apr. 1986), pp. 110–141. ISSN: 0730-0301. DOI: 10.1145/22949.22950. URL: https://doi.org/10.1145/22949.22950.

[30] M. Meyer, T. Munzner, and H. Pfister. "MizBee: A Multiscale Synteny Browser". In: *IEEE Transactions on Visualization and Computer Graphics* 15.6 (Nov. 2009), pp. 897–904. ISSN: 1941-0506. DOI: 10.1109/TVCG.2009.167.

[31] D. Moritz et al. "Formalizing Visualization Design Knowledge as Constraints: Actionable and Extensible Models in Draco". In: *IEEE Transactions on Visualization and Computer Graphics* 25.1 (Jan. 2019), pp. 438–448. ISSN: 1941-0506. DOI: 10.1109/TVCG.2018.2865240.

[32] T. Munzner. "A Nested Model for Visualization Design and Validation". In: *IEEE Transactions on Visualization and Computer Graphics* 15.6 (Nov. 2009), pp. 921–928. ISSN: 1941-0506. DOI: 10.1109/TVCG.2009.111.

[33] Tamara Munzner. *Visualization analysis and design*. CRC press, 2014.

[34] C. Nobre et al. "The State of the Art in Visualizing Multivariate Networks". In: *Computer Graphics Forum* 38.3 (2019), pp. 807–832. DOI: https://doi.org/10.1111/cgf.13728. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.13728. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13728.

[35] S. Nusrat, T. Harbig, and N. Gehlenborg. "Tasks, Techniques, and Tools for Genomic Data Visualization". In: *CGF* (). DOI: 10.1111/cgf.13727. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.13727. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13727.

[36] Sabrina Nusrat and Stephen Kobourov. "Task Taxonomy for Cartograms". In: *Eurographics Conference on Visualization (EuroVis) - Short Papers*. Ed. by E. Bertini, J. Kennedy, and E. Puppo. The Eurographics Association, 2015. DOI: 10.2312/eurovisshort.20151126.

[37] A. Pandey et al. "CerebroVis: Designing an Abstract yet Spatially Contextualized Cerebral Artery Network Visualization". In: *IEEE Transactions on Visualization and Computer Graphics* 26.1 (Jan. 2020), pp. 938–948. ISSN: 1941-0506. DOI: 10.1109/TVCG.2019.2934402.

[38] Aditeya Pandey, Sehi L'Yi, and Nils Gehlenborg. "Towards a Knowledge-Based Recommendation System for Genomics Visualization". In: ().

[39] Aditeya Pandey, Uzma Haque Syeda, and Michelle Borkin. "Towards Identification and Mitigation of Task-Based Challenges in Comparative Visualization Studies". In: (2020).

[40] Aditeya Pandey et al. "Poster: Segmentrix: A Network Visualization Tool to Develop and Monitor Micro-Segmentation Strategies". In: ().

[41] *Power BI, url = https://powerbi.microsoft.com/en-us/.*

[42] Steven F. Roth and Joe Mattis. "Data Characterization for Intelligent Graphics Presentation". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '90. Seattle, Washington, USA: Association for Computing Machinery, 1990, pp. 193–200. ISBN: 0201509326. DOI: 10.1145/97243.97273. URL: https://doi.org/10.1145/97243.97273.

[43] H. Schulz. "Treevis.net: A Tree Visualization Reference". In: *IEEE Computer Graphics and Applications* 31.6 (Nov. 2011), pp. 11–15. ISSN: 1558-1756. DOI: 10.1109/MCG.2011.103.

[44] H. Schulz et al. "A Design Space of Visualization Tasks". In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (Dec. 2013), pp. 2366–2375. ISSN: 2160-9306. DOI: 10.1109/TVCG.2013.120.

[45] M. Sedlmair, M. Meyer, and T. Munzner. "Design Study Methodology: Reflections from the Trenches and the Stacks". In: *IEEE Transactions on Visualization and Computer Graphics* 18.12 (Dec. 2012), pp. 2431–2440. ISSN: 1941-0506. DOI: 10.1109/TVCG.2012.213.

[46] B. Shneiderman. "The eyes have it: a task by data type taxonomy for information visualizations". In: *Proceedings 1996 IEEE Symposium on Visual Languages*. 1996, pp. 336–343.

[47] John Stasko et al. "An evaluation of space-filling information visualizations for depicting hierarchical structures". In: *International journal of human-computer studies* 53.5 (2000), pp. 663–694. DOI: https://doi.org/10.1006/ijhc.2000.0420.

[48] Chris Stolte, Diane Tang, and Pat Hanrahan. "Polaris: A System for Query, Analysis, and Visualization of Multidimensional Databases". In: *Commun. ACM* 51.11 (Nov. 2008), pp. 75–84. ISSN: 0001-0782. DOI: 10.1145/1400214.1400234. URL: https://doi.org/10.1145/1400214.1400234.

[49] Uzma Haque Syeda et al. "Design Study "Lite" Methodology: Expediting Design Studies and Enabling the Synergy of Visualization Pedagogy and Social Good". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. Honolulu, HI, USA: Association for Computing Machinery, 2020, pp. 1–13. ISBN: 9781450367080. DOI: 10.1145/3313831.3376829. URL: https://doi.org/10.1145/3313831.3376829.

[50] Danielle Albers Szafir. "The Good, the Bad, and the Biased: Five Ways Visualizations Can Mislead (and How to Fix Them)". In: *Interactions* 25.4 (June 2018), pp. 26–33. ISSN: 1072-5520. DOI: 10.1145/3231772. URL: https://doi.org/10.1145/3231772.

[51] *Tableau Software, url = https://www.tableau.com/*.

[52] Manasi Vartak et al. "SeeDB: Efficient Data-Driven Visualization Recommendations to Support Visual Analytics". In: *Proc. VLDB Endow.* 8.13 (Sept. 2015), pp. 2182–2193. ISSN: 2150-8097. DOI: 10.14778/2831360.2831371. URL: https://doi.org/10.14778/2831360.2831371.

[53] Manasi Vartak et al. "Towards Visualization Recommendation Systems". In: *SIGMOD Rec.* 45.4 (May 2017), pp. 34–39. ISSN: 0163-5808. DOI: 10.1145/3092931.3092937. URL: https://doi.org/10.1145/3092931.3092937.

[54] Martin Voigt, Stefan Pietschmann, and Lars Grammel. "Context-aware Recommendation of Visualization Components". In: 2012.

[55] K. Wongsuphasawat et al. "Voyager: Exploratory Analysis via Faceted Browsing of Visualization Recommendations". In: *IEEE Transactions on Visualization and Computer Graphics* 22.1 (Jan. 2016), pp. 649–658. ISSN: 1941-0506. DOI: 10.1109/TVCG.2015.2467191.