Implementing Criteria for Rigor in an Evolutionary Biology Design Study

Jen Rogers, Austin Patton, Luke Harmon, Alexander Lex, Miriah Meyer
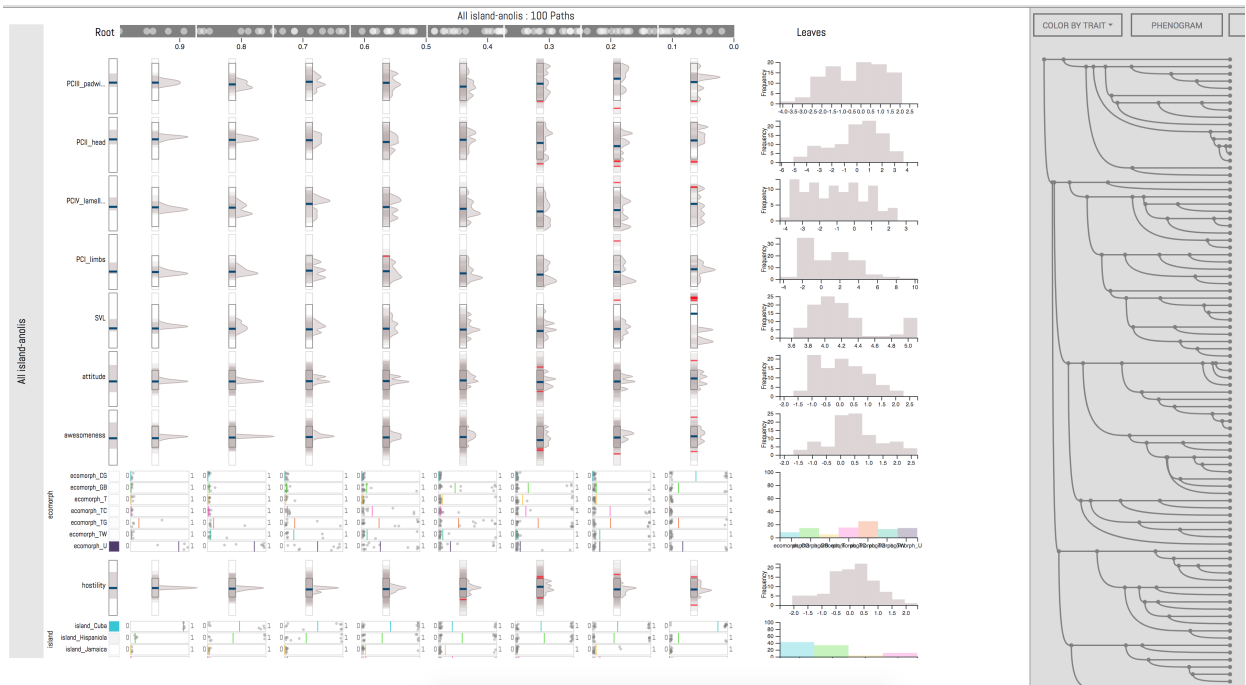

Fig. 1. Screenshot of Trevo

Abstract— A fundamental goal in evolutionary biology is to explain the diversity of life on earth. To address this, researchers reconstruct ancestral histories of living species and their characteristics to better understand when and how these characteristics evolved. These histories are commonly represented as phylogenetic trees showing possible relationships of living species. In these trees, the leafs are extant and typically well-described species, while the ancestral species encoded in the internal nodes are commonly reconstructed through statistical models. These models also infer likely traits of the ancestral species, with a varying degree of certainty. To analyze evolutionary patterns, scientists analyze several traits of the extant and reconstructed species simultaneously.
Visualizing multiple attributes under uncertainty in the context of the tree topology is a challenging problem, which we address in this paper. We developed Trevo, a tool to analyze evolutionary patterns in trait development in a design study with a team of biologist. Our contributions include (1) a characterization of domain goals and tasks; (2) two novel visualization techniques for visualizing a summary of attributes over time in a sub-tree, and (b) a ranking and exploration interface for evolutionary patterns. We also contribute the insight gained from our experiments in rigor, with actionable recommendations for how to apply these criteria in a design study.

Index Terms—Radiosity, global illumination, constant time

---◆---

1 INTRODUCTION

Design studies are am important type of vis inquiry. Many ds have been reported on, spanning a broad range of application areas. Also work on theoretical underpinnings, include process models [9-stage,DAF,ADR], decision models [NM,NBGM], and educational approaches [DSLM]. Despite this maturing of ds, open questions remain about what we can learn from ds and how to conduct them with rigor.

Recent work by Meyer and Dykes propose a re-framing of ds

- *Jen Rogers, Alexander Lex, Miriah Meyer are with the University of Utah*
- *Austin Patton and Luke Harmon are with the University of Idaho.*

that stresses a broad set of knowledge contributions that are acquired through the process of design and reflection. This reframing is grounded in interpretivist epstilomolgy, and proposes a set of six criteria for rigor in ds. Work lays out considerations for rigor, but does not provide specific examples of how criteria are considered in practice, or what types of learning happens when they are considered.

In this work we experiment with considerations for three rigor criteria: ABUNDANCE, REFLEXIVITY, and TRANSPARENCY. Our experimentations took place within the context of a design study with evolutionary biologists, and consisted of a variety of methods throughout the study.

More specifically, we conducted a 1-year design study with a lab of biologists studying.... Over the course of the ds we employed techniques such as an immersive, 3-month field study; structured and systematic self-reflectiion; and careful curation of documents and other design artifacts. Through collaborative reflection we identified several

methodological insights that emerge from our experiments.

The contributions of this work are a complementary set of visualization design innovations from our collaboration with biologists, as well as methodological results from our criteria experiments. The contributions include: Two new techniques for supporting visual analysis of multivariate trees Three methodological insights and recommendations for rigor considerations Furthermore, this work serves as an example of how some criteria can be considered in practice, as well as the diverse types of knowledge contributions possible through their consideration.

2 DOMAIN BACKGROUND AND CHARACTERIZATION

2.1 Background

The driving force in evolutionary biology is to explain "the diversity of form and function" of the living world [?]. Evolution, as Darwin concisely described it, is descent with modification. To understand the evolutionary process, it is important to determine when a given characteristic evolved and whether it was a result of common ancestry or other forces. This is important in understanding not only how animals adapted to their environments, but the rate and mutation of virus strains, such as COVID-19.

How do evolutionary biologists begin to answer these questions? They study a population of organisms to establish hypotheses about the underlying forces of evolutionary change that can generalize to other populations. For example, Anole lizards are heavily studied to infer how environment influences evolution.

First, they reconstruct the ancestry for the population of species to a single common ancestor. These hypothesized evolutionary relationships are represented as a tree. Each splitting of a branch represents a speciation event, characterized by enough change in genetic differences between two organisms where they are no longer able to reproduce with viable off-spring. The tree provides the relational context to understand when and why an organism evolved the way it did. For example in our lizards, when did the dewlap emerge in the lizard history and did it emerge multiple times in the tree? Determining if this physical characteristic emerged independently in separate parts of the tree can indicate that this was a result of the lizards adapting to their environments. To answer if the dewlap emerged in separate parts of the tree, analysts map characteristics for the species and ancestors on to the phylogenetic tree. Characteristics of the species are referred to as traits. Traits can be genetic sequences or morphological measures, for the scope of this paper, traits refer to measured morphological characteristics. They lack the actual ancestral traits, so they reconstruct these traits using traits of existing species and if available, fossil records. The statistical evolutionary model they use to reconstruct traits is referred to as ancestral reconstruction(or character mapping).

The concept of one all-encompassing tree of life is a myth. There is a significant amount of uncertainty in both the tree topology and the reconstructed ancestral traits. A given phylogenetic tree is constructed from a large group of simulated trees that vary slightly in topology. One consensus tree is calculated from this group of trees and is determined to have the most likely topology sourced from the collection. Often these trees have ranges for possible branch times.

Because analysts estimate ancestral trait values back up the tree, there is significant uncertainty in these values that increases as you go back in time. In addition to the challenge in trait uncertainty, analysts often need to account for multiple traits at once. To do this, they compare different traits mapped on to the nodes of mirrored trees side-by-side. This is difficult with just two traits, but often they are considering up to 10 traits for a given tree. As expressed by a participant in a work shop we conducted early in the study, "if you have one continuous trait you can do things. If you have two - ok. If you have three or four or five, there is nothing really sufficient, " (CITE AUDIT TRAIL?).

Our collaborators study these evolutionary relationships, often comparing multiple trees, to get an understanding of how traits change. The challenge of multiple traits and significant uncertainty in the topology of the tree and ancestral trait values, compounds to a very difficult problem for analysis. This was a main motivation in our collaboration with the University of Idaho, in which we conducted a design study to explore how visualization could help with some of these challenges.

3 RELATED WORK

3.1 Conducting Design Studies

Extensive groundwork has been done on methodologies for conducting applied visualization research [15, 22, 24]. **WILL ADD MORE HERE**

Design study methodology is fundamental to research of this nature [22]. Design studies are characterized by a collaboration between visualization researchers and domain experts motivated by a problem that is "wicked" in nature [16]. Wicked problems are characterized by the inability to define the problem without the solution [6]. The process of tackling this type of problem is iterative and non-linear. This also means that proposed solutions for wicked problems, and the process that drove their development are also hard to evaluate. So how do we begin to tackle this? **AFTER TALKING DURING CLASS ABOUT WICKED PROBLEMS MAYBE I SHOULD REFRAME THIS?**

Though there is a lot of reporting on design studies, what kind of knowledge we can gain from them and how can we conduct them with rigor remain open questions.

Case studies are the most common validation for design studies [22], in which the tool is evaluated by real users, using real data to answer questions they have. Many reports on case study evaluations are characterized as insight-based [10], where the evaluation is embedded in the descriptions of insight the domain analyst gains from using the tool. Detailed descriptions are provided for an outside party, who reads the case study description, gets context for the tools use and its value in answering questions within the domain space. Case studies provide ecological validity.

Unfortunately, case studies are often boring. It leads one to question whether something that many people skim over in their reading of a design study paper, is worth including in its traditional form as evident validation. The traditional structure of a design study paper leaves little to no room for the rich, interesting, insight that emerges from experimentation in the problem space.

Do we need case studies? As a conceptual framework, yes. But we argue for their use in combination of other methodologies for establishing trustworthiness and evaluating usefulness.

In qualitative work, rigor is a way to establish trustworthiness or credibility in the research [9]. Meyer and Dykes defined a set of six criteria for rigor in a design study [16]. The six criteria are: INFORMED, REFLEXIVE, ABUNDANT, PLAUSIBLE, RESONANT, and TRANSPARENT. We specifically investigated reflexive, abundant, and transparent in the context of the design study with evolutionary biologists.

3.2 Phylogentic Visualization

It is understood by the evolutionary biology community, that a given tree has a level of topological uncertainty. There are often several tree topologies that are equally likely to be true. Branch times are not definite, but have estimated ranges. The uncertainty in branch times are commonly represented as explicit ranges for the branch points of a tree. Though this works for communicating branch time ranges, the uncertainty in equally probable topological variation is not accounted for. Research within the VIS community has been done to address this variation by designing tools to compare topology of a collection of trees [11, 12, 17].

A common way to communicate topological uncertainty in the domain space is with a DensitTree plot, in which a collection of dendrograms varying in topology are overlaid with some transparency to see where there is more consensus in the topology and where it is more ambiguous [5]. However, it can be hard to understand the topological variation in these. As expressed by a participant early in the study process when discussing the difficulty of interpreting these representations, "some of that has to do with uncertainty, some of that I think also has to do with the inadequacy of the approach to communicate that uncertainty" CITE THE AUDIT TRAIL.

As mentioned, analysts need to reconstruct evolutionary histories by mapping estimated trait values on to the topology of a tree. More often then not, there are multiple traits that need to be accounted for at once, making a very challenging multivariate visualization problem. Work has been done in the visualization community for addressing the difficult task of visualizing multiple attributes in the context of topology by creating taxonomies and detailing recommended use [1]. This provides a foundation for understanding best practices for representing multivariate networks based on analysis tasks, but does not account for uncertainty in the multivariate data.

Linearization of paths in a tree is a technique used to address the challenge of viewing multiple attributes and the topology at once [18–20]. By linearizing the topology, attributes can be shown below or beside for comparison. This however, is used for mainly single paths and not a distribution of them.

4 PROCESS

4.1 Core Phases

This design study was a year-long collaboration with Evolutionary biologists at the University of Idaho. The process began with email correspondence with the lab and background work into literature in the domain space.

4.1.1 Field Study

We conducted a three month immersive field study near the beginning of the collaboration. Two more visits were conducted after the initial visit. During the field study, the first author was fully immersed in the lab of our collaborators. She was given a desk in the lab space, where she spent the majority of her time. She gave a presentation to introduce the lab to design study concepts and outline what she would be doing during her visit. She interviewed members of the lab and conducted a creative visualization opportunity workshop with he members. She took regular reflective memos and notes during the study.

4.1.2 Design Phase

The beginning stages of sketching and prototyping began during the field study. However, the bulk of development happened upon the first authors return to her home lab. Design was an iterative process. The main collaborators were very involved in providing feedback on design iterations. Much of this correspondence happened through email and visits to our collaborators lab. Artefacts generated during this process included sketches, illustrator mock-ups, and prototype iterations.

4.1.3 Reflection Phase

Regular reflective memos were taken during the field study, however, the reflection phase began with the construction of the audit trail, in which authors revisited and reflected on the body of collected artefacts. Through this process, insight emerged on the design of the tool, as well as the design study process. With repeated discussions with the vis team, this insight was articulated into lessons learned and actionable recommendations for methodology moving forward.

4.2 Domain Collaborators

Our primary collaborators were the head of the lab at the University of Idaho, as well as a graduate researcher at Washington State University currently collaborating with the Harmon lab. We worked closely with them throughout the study and continued regular correspondence with them after the field study. Secondary collaborators included the graduate students in the lab. They participated in 1-3 interviews, a half day workshop conducted at the University of Idaho, and were active members in the lab during the field study.

4.3 Experimentation with Criteria for Rigor

Meyer and Dykes defined six criteria to establish rigor in the subjective, often hard to evaluate design study process, in which they "leave it to design study researchers to decide, and argue for how best to achieve these criteria - given the specific people, data, and context involved in the study," [16]. We have experimented with these criteria, focusing

on three of them: Reflexive, Abundant, Transparent. In the following sections we will detail how these were integrated into the methodology and the evidence of our findings drawn from this experimentation.

4.3.1 Reflexive

Reflexivity involves "explicit and thoughtful self-awareness of a researchers own role in a study," [16].
    // NEED TO ADD MORE ON WHAT

- Reflexivity through memoing

Reflexivity is a way to gather data [21]. A common method of doing this in qualitative work is through memoing. "Memos can help to clarify thinking on a research topic, provide a mechanism for the articulation of assumptions and subjective perspectives about the area of research, and facilitate the development of the study design" [4].

Reflexivity has been a major consideration in qualitative research, especially in the field of nursing [13]. It is argued that qualitative work is fundamentally reflexive, and "the very nature of qualitative research requires the researcher to assume a reflexive stance in relation to the research situation, participants and data under study" [4]. Researcher bias and perspective is an inherent part of qualitative research, eliminating them from the research process is impossible [13]. The way to mitigate bias is to be explicit on the thoughts, feelings, and perspective of the researcher.

*what are we doing in the research community? Seems to be a void of design study papers that mention reflexivity?

Though reflexivity has been reported on in the HCI community [2, 21, 23], it has been slow to adopt reflexive practices in research due to scrutiny on subjectivity in the review process. This comes as a cost to the research [21], in which the interesting space for speculation and exploration can be stifled by pragmatism and usability testing [7]. There remains a void of papers reporting on reflexivity in VIS research. This was a motivating factor for including this criteria in our work.

As our first author embedded herself into the domain space, she took regular reflective notes throughout the study in the form of regular memos. These reflections were reflexive in nature and included: how feelings on position in the lab, insecurities that were limiting the research, social dynamics, friendships, and how those dynamics affected the research.

4.3.2 Transparent

"Transparent reporting should be self-critical and include errors, failures, analytical dead ends – the joys and mistakes," [16].

There has been significant attention to transparency or lack thereof in our research community. Transparency is considered important for establishing rigor [16, 25] and validation of our often qualitative and ethnographic work, yet we have yet to achieve a standard threshold for the level of transparency in our reporting [25].

Research that is more transparent may even be subject to greater scrutiny in the review process [25]. Another main concern for releasing data and other material is protecting the privacy of participants [26].

Access to data and source code has been a common way to address transparency. There has been recent experimentation in releasing a temporally organized "audit trail" as supplementary material to the research [?, 14]. Other work has focused on making data not only accessible, but understandable and interpretable through visualization [3]. How and to what extent we should be transparent remains an open question. This was a motivation to include transparency in our experimentation for rigor in this study.

In an attempt to transparently communicate the design study process, we created an audit trail website two thirds into the study. This contains annotated records including: notes, email correspondence, Sketchbook scans, photos of collaborator sketches, links to papers, and notes from audio. We will be releasing this website.

4.3.3 Abundant

"A design study with abundance has rich details; many voices, data sets, contexts, and designs; and significant time in the field," [?].

Abundance is something not often explicitly focused on in research reporting, but is still present. Recent work in immersive design studies advocate for immersion to "facilitate collaboration and accelerate project development by building trust and deepening the dialogue between collaborators" [8].

- Immersive field study

We conducted a three month immersive field study in the lab of our collaborators in which the first author sat at a desk among the other graduate researchers. In several respects, she became a peer. She attended lab research meetings, bounced ideas off of the group as she would in her own research environment.

- Reached out to other labs

In an attempt to include multiple voices and data sets, we reached out to other labs outside of the core research lab with which we were collaborating. A primary collaborator gave recommendations for other labs that were currently conducting research that would be applicable. We gathered a list of researchers at the University of Utah that have recent research involving phylogenetic analysis. We sent out emails to these labs which we have included in the audit trail website (CAN POINT TO THE EMAIL). Though this did not work out, we have insight on how to conduct this moving forward.

- An Abundance of Sketching

We used sketching for brainstorming features, understanding the domain space, and communication with domain collaborators. These "sketched" artifacts also include any sketching done by collaborators. As a result, we have a rich collection of artefacts that provides a record of the speculative process of learning important domain concepts and the design development. These are discoverable in the audit trail website.
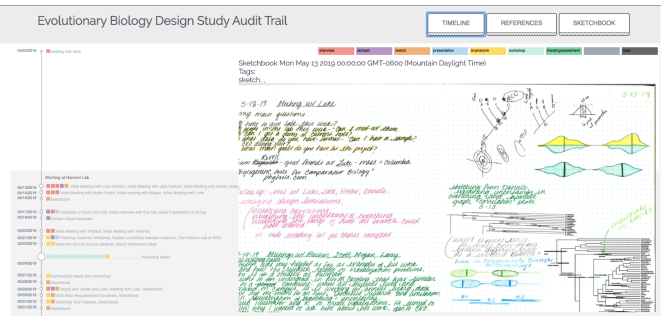

Fig. 2. description for audit trail

5 DOMAIN CHARACTERIZATION

Associated trait data for both internal and leaf nodes were also loaded using Multinet's API. There is a distinction between leaf nodes representing the species, and the internal nodes representing common ancestors. Leaf nodes have associated traits that are measured and definite. Internal nodes have associated traits that are estimated and contain significant uncertainty. Traits can be continuous (such as tail length, head length) or discrete (such as whether they lay eggs, island of origin). They then map these reconstructed traits to the hypothesized topology of the tree. The trait values for internal nodes are estimated from the species trait values. The reconstructed continuous traits include an estimated value and a 95 percent confidence interval. The discrete traits include the probabilities that the internal node belongs to each of the possible states.

The tree diagrams how species are related up the tree. Each node has several associated attributes. This creates a rich, multivariate tree.

5.1 Domain Requirements and Tasks

The fundamental goal of our collaborators' research is to explain why the living world evolved the way it did. To address this, their analysis is focused on understanding when and how traits evolved in a population. To answer this, they needed the ability to view trait values for multiple attributes in the context of the topology. To address this question, we have characterized five tasks:

T1: Understand the uncertainty in reconstructed ancestral traits

As mentioned earlier, there is significant uncertainty in the reconstructed traits for internal nodes. Our collaborators lack adequate visual representations of trait values that include the uncertainty. 95 percent confidence intervals for continuous traits are considered in comparisons of trait change through the history to determine what trait changes are actually significant. High uncertainty for discrete traits is indicated by equal or very similar probabilities for all possible states for an internal node.

T2: Define sub-trees by topology or trait values

Comparisons of trait histories are more often made between sub-trees. Our collaborators need the ability to define sub-trees by topology and trait values. For example, an analyst might want to compare trait values between island groups of a population of salamanders. They need to be able to define groups by the species island category.

T3: Identify evolutionary outliers

It is important for our collaborators to be able to identify paths that are significantly higher or lower in trait values than the rest of the sub-tree. For example, they want to be able to easily identify paths that have a with significantly larger body mass than the rest of the sub-tree they are exploring. To assist with this, these outliers should be explicit in the visualization.

T4: Identify trait patterns that indicate known evolutionary events

An important task in our collaborators analysis is identifying path trajectories that fit a pattern associated with an evolutionary event. There is currently no way to do this visually with existing tools. These patterns occur between pairs of paths and require the comparison of trait trajectories of multiple species in a tree. An example of this would be identification of situations of convergence, where an analyst would identify paths that are separate n the topology of the tree, branching early in the tree history from one another. These paths would diverge in trait value, and then nearing the species in the recent history, they would converge to very similar trait values. This known pattern of convergence is an indication of adaption in a population, where species grew similar in physical characteristics because they inhabit similar environments.

T5: Compare path histories for multiple attributes

Comparisons are important in identifying how traits are changing. These comparisons are between different species' paths for a given trait or between different traits of the same species or sub-tree. Often a significant pair is determined from identifying a known pattern in multiple traits, not just a single one. Considering this, it is essential to be able to identify significant trait changes in multiple traits at once. Viewing multiple traits at once has been a difficulty for our collaborators, traditionally relied on comparisons of reconstructed traits on separate trees.

6 VISUALIZATION DESIGN - CAN WE INCORPORATE LIZARD STORY HERE FOR ADDED CONTEXT TO THE VISUAL DESCRIPTION??

- Explain that the tool has views we are not claiming as a contribution?.

- reiterate the 2 novel contributions.

There are two technical contributions that emerged from this design study. A trait view that visualizes a distribution of trait values in the context of the tree topology, and a pattern view that identifies patterns within path pairs.

6.1 Trait View

The trait view characterizes trait trends independent, but within the context of the topology of the tree. This allows comparison of trends between sub-trees composed of paths that are not in a single monophyletic group, but scattered in the tree topology. For example, a user wants to determine when a significant divergence in body mass occurs in the tree between lizards on the island of Hispaniola vs Cuba. Species that are found on both island are scattered across the tree and isolating these paths allows for more direct comparison unobscured by tree topology.

6.1.1 Projecting the tree topology to a single line

The internal nodes of the tree is binned by time and the topology of the tree is projected to a single line 3. This allows for comparison on trait distributions unhindered by the topology of the tree. Sub-trees can be defined by range value of a continuous trait or category of a discrete trait. Species nodes are left out of this projection.
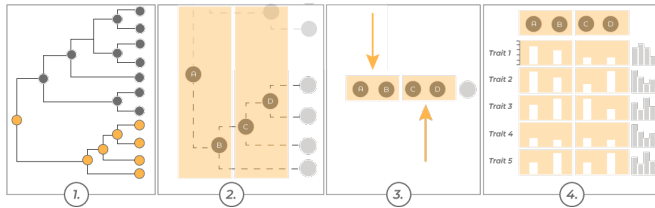

Fig. 3. (1)Topology of the entire tree with subtree highlighted (2) Zoomed in subtree binned by time (3) Binned internal nodes of Subtree projected to single line (4) Rows for each trait show distribution of values for internal nodes in a given bin.

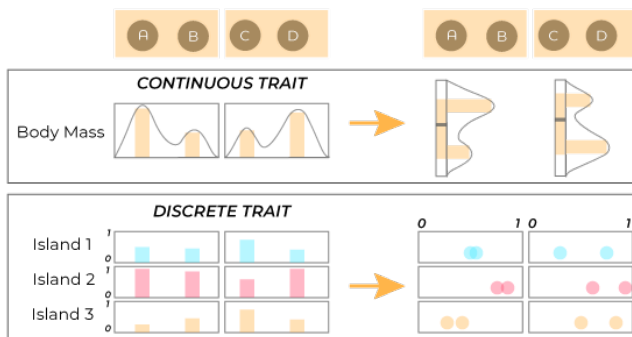### 6.1.2 Continuous and Discrete Trait Encoding

Fig. 4. description for trait breakdown

Because the internal nodes are binned, each trait bin contains a visual representation for a sample of trait values. As mentioned earlier, there is significant uncertainty in the internal node trait values.

- Continuous

For continuous traits, this is represented by 95 percent confidence intervals for the estimated values in the paths of the pattern view 6 and the binned range rectangles of the trait view 1. The distributions of the continuous trait values are represented as a Kernel Density Estimate plot.

- Discrete

Probabilities for discrete traits are represented in the trait view as separate one-dimensional plots for each state 4. The axis for these plots range from probability of 0 to probability of 1. All probabilities on internal nodes for each state are plotted in their respective plots. To keep from obscuring the nodes when they share very similar probabilities, jitter was added to their position in the plots. The average for each state probability is plotted as a line in the plots, colored by the state category. The discrete view design was iterated on several times before coming to its current iteration (CITE AUDIT TRAIL HERE?).

6.1.3 Compare Sub-trees

To better identify differences in distributions across sub-trees, the user can define sub-trees from possible states or value ranges for species in the sidebar, or brushing internal node ranges in the trait distribution.

- Sub-tree by Brush

The user identifies a distribution of high values for tail length in a later time bin. They want to isolate those paths into a new subgroup to see how that trait changed across the history of the tree for that subgroup and whether other traits changed in a similar way. For example, did body size decrease in these paths that evolved a longer tail?

- Sub-tree by sidebar

As mentioned earlier, it is important for an analyst to isolate subgroups from the entire tree. They can do this by brushing distributions of interest, or by specifying a range for leaf nodes. This is done in the sidebar and can be done for discrete or continuous traits. For continuous traits, the analyst can specify a range on a slider. For discrete traits, they can specify a subgroup by a given state in a discrete category. For example, an an analyst wants to determine whether there is a significant difference in snout length for mainland lizards vs island lizards. They define subgroups for each of these from the sidebar then overlay these groups to see when snout length values diverge for the two groups and by how much, and whether other traits follow a similar trajectory.

6.2 Pattern View

The pattern view allows analysts to identify path pairs that follow a known pattern that infers an evolutionary event. For example, convergence is a pattern that an analyst can use to score ranked pairs that indicates species' adaption to respective environments. This pattern is defined by pairs that have branched early in the tree history, have diverged significantly in trait value, and then have converged to similar trait values for the existing species at the tips of the tree. This is studied extensively using data on the Anole lizard taxa, which is composed of a large, diverse set of species and have known subgroups that share similar characteristics. These groups of similar species inhabit islands of similar environments, such as, a group of Anole species defined as "twig" Anoles have evolved independent of one another, on separate island but are all small, with long tails that allow them to occupy the small branches of trees on their respective islands.

6.2.1 Known Patterns

There are 6 preset patterns that an analyst can choose from to score path pairs 5. These pattern characterizations emerged from repeated deliberation with our collaborators. Though there are known patterns in the domain space that indicate potential evolutionary events, some of these patterns did not have a definitive name. For example, there was debate on what the pattern currently referred to as "anti-convergence" should be defined as. What characteristics about this pattern are most significant? What exactly does this indicate? Is it character shift or displacement? This lead to an interesting deliberation of what exactly the relationships between path value trajectories would look like as visual representations, something they had not yet explored. Patterns are specified by adjusting 3 metrics. There is a drop down to select these patterns that adjusts the metrics accordingly.
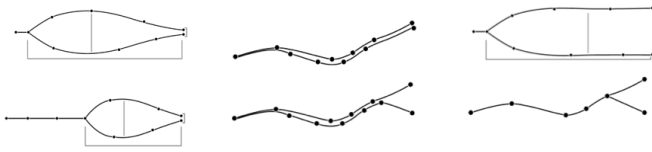
Fig. 5. There are 6 present patterns that can be identified in the tool.

Pattern	Distance	Delta	Closeness
Convergence (Ancient)	1.0	1.0	1.0
Convergence (Shallow)	-1.0	1.0	1.0
Divergence (Ancient)	1.0	1.0	-1.0
Divergence (Shallow)	-1.0	1.0	-1.0
Conservativism	1.0	-1.0	1.0
Anti-Convergence	1.0	-1.0	-1.0

Table 1. metric weights for each defined pattern

6.2.2 Metric to define a pattern

Distance, Delta, and Closeness are three metrics used in combination to identify a given pattern. Distance refers to the distance up the tree where the paths diverged from a common ancestor 6. Delta is the maximum difference in trait value after path divergence. Closeness is the difference in trait value at the tips of the tree. These metrics are adjusted based on the pattern chosen. The metric weights for each pattern are defined in table 1. Weight for each metric can either be 1, 0, or -1. If all of the weights are 1, paths with the maximum distance from branch point, maximum difference in value once branched, and closest value at the leaf nodes are scored the highest.

6.2.3 Pair Plots

Every possible pairing of paths in the tree are calculated. Then an initial trait is used to calculate the ranking of each pair. Each pairs' path trajectories for a trait are scored based on a set of three metrics that are adjusted according to the pattern that is chosen. The pairs are then sorted by combined score and the top one percent of all pair combinations are taken and visualized in the pair plots 6.

6.2.4 Heat map of Other Traits

This component emerged from a meeting with collaborators reviewing the path view, after which they asked if it was possible to have an indication of whether these paths were ranked high in other paths (CAN CITE AUDIT TRAIL?). These patterns of trait change are less significant if present in only one trait. Considering this, the pattern view calculates the top pairs for a given trait then for those top pairs, calculates the score for every other trait to determine whether they rank high for other traits. Pairs ranked high in multiple traits include a heat map where each square represents the other traits 6. Squares with darker saturation have a higher ranking. The trait name is shown on hover. Pair plots can be sorted by number of other top ranked traits. For example, an analyst is exploring top ranked pairs for convergence our Anole lizards. She begins with the trait Snout Vent Length, a common metric for lizard body size. She also wants to see how tail length and toe pad size rank for convergence so she sorts the ranked pairs by frequency in top ranked additional traits.

7 INSIGHTS AND LESSONS LEARNED

In this section, we describe three lessons learned that emerged from reflection-on-action of artifacts from the design study. These were refined and articulated further after meetings with our fellow visualization researchers.

7.1 Lesson 1: Explicit, systematic reflection was demonstrably productive

There were two types of reflection that were productive in the process: Reflection-in-Action and Reflection-on-Action.

- cite schon here??

- Reflection-in-Action

Consistent reflection was adopted through the design study process. This is defined as Reflection-in-Action, or reflecting during the activity. Reflection-on-action helped propel the process forward - especially during the early phases of the study where the domain space was not well known. Reflective memoing was the main medium for this type of reflection. This was done before and after meetings with collaborators. In pre-meeting reflection, notes from previous meetings would be reviewed to inform what content we had to build from. Points to hit during the meeting were defined before a meeting to establish a loose agenda. Post-meeting memos included main points that were discussed as well as what talking points were met and what were still open.

- cite reflective memo paper here??

Regular reflective memos were also reflexive. They included feelings on presence, insecurities, and dynamics in the lab. By making these explicit, we were able to identify limitations in the research that could be acted upon and changed methodology during the process because of this. For example, an author felt uncomfortable audio recording. She was explicit about this in her notes. Reviewing these at a later point and discussing with another author, she realized the detail in her current note-taking could be improved, and the benefits of recording outweighed her hesitation about it. The change to incorporate audio recording allowed for more abundant reflection-on-action. (CITE WHERE THIS HAPPENED IN THE AUDIT TRAIL?)

- Reflection-on-Action

The second type of reflection was "Reflection-on-Action". Within this, there were two examples of reflection-on-action that happened during the process. The first was reflection via reflective transcription.

- Reflection via reflective transcription

After we started to record meetings, we collected audio recordings for every meeting moving forward. Reflective transcription would happen within the same day or the next day that the audio was recorded. Reflective transcription did not involve transcribing the audio word for word, but used as a way to take notes and reflect on the meeting. During the first reflective transcriptions, it became apparent that the recordings afforded deeper context and detail than note taking during meetings could. It not only capture what was said, but how it was said. Recordings were revisited at a later time to get added information and context for something that had been written down in the notes.

- Reflection via Artifact Curation

Is there room for Tavory and Tammerman's importance of revisiting notes here? [?]. Two thirds into the study process, we began to curate all of the artifacts we had collected, with the intention of creating an audit trail website that provided a transparent documentation of our process. We did not anticipate that the act of curating and organizing artifacts facilitated reflection-on-action that played a vital role in design development. For example, the discrete trait representation in the trait view was re-designed after a more informed domain understanding and identification of re-emerging concepts after curation and categorization of previous notes and sketches. (CITE WHERE THIS HAPPENS IN THE AUDIT TRAIL?)

- When should we reflect?

When reflection should occur remains an open question in applied visualization research [?]. The time and regularity of reflection during the study process depends on the type of reflection. Reflection-in-action that was systematic and consistent allowed for identification of limitations that affected the research. As well, regular reflection-in-action
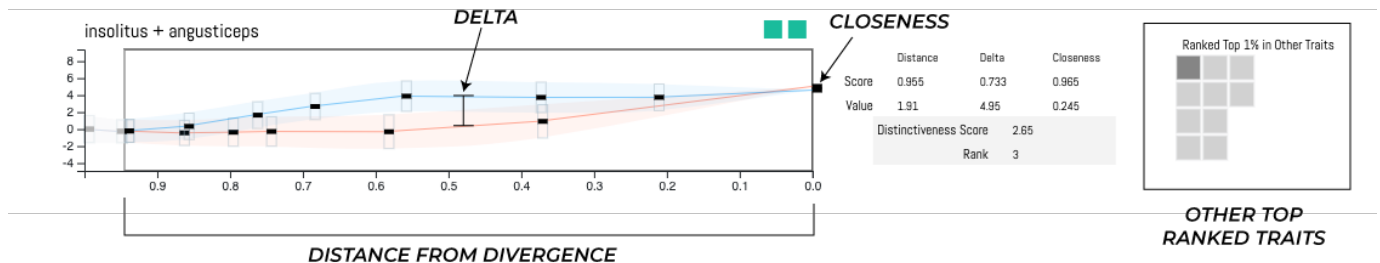
Fig. 6. description for pair plot

facilitated reflection-on-action by providing an abundance of material to speculate and draw from. Though we do not have rigid recommendations for when reflection-on-action should occur, it should be done during the process of design and development, as it was a demonstrable catalyst for changes in the design and methodology during the design study process.

7.1.1 Recommendations

- Reflect before and after meetings with collaborators

- Record audio

- Curation of artifacts is a process that facilitates reflection. Record insight emergent from this.

7.2 Lesson 2: When artifacts are discoverable, they support transparency and reflection

7.2.1 The Audit Trail

Through the design study process, artefacts were organized and curated with the intention of releasing a publicly available audit trail for transparency in communication of our process. In creating the audit trail website however, we realized that the curation of artefacts allowed for significant reflection. Ultimately, the audit trail served two roles; transparent communication and reflection-on-action.

Initial organization was done using a spread sheet, where each artefact was given a row, including the date it was created and associated meta data. This became the backbone for the artifact timeline. The temporality of the timeline organization showed the evolution of ideas, and allowed identification of what emerged when. Each artefact row has columns for initial categorization of artifact type. Then it can get categorized and tagged. Tagging allowed the artefact to be found by searching for a keyword.

7.2.2 Recommendations

- Make your records discoverable (by you and others).

- Create a system for organization and curation as soon as you begin collecting

- Categorization and tagging makes artefacts more discoverable and allows recurring concepts to float to the top.

7.3 Lesson 3: Changes in collaboration patterns are evidence for criteria

An author conducted an immersive field study in our collaborators lab, in which she developed relationships with lab members. This resulted in a long-term deep engagement. During reflection, we noticed how communication moved from more formal, less-frequent, longer correspondence to informal, frequent, shorter text messages as the author became more immersed in the lab. This shift in correspondence from formal, long, descriptive correspondence to frequent, short, informal text messages is evidence for ABUNDANT as well as INFORMED criteria.

- ABUNDANT

As a long term engagement was established, so was a mutual trust and collegiality. The author would currently consider most of the lab members friends. This evolution aligned with the change in social dynamic.

- INFORMED

As the author's understanding of the domain space deepened, her conversation on domain concepts with collaborators accelerated. They would not have to take time to explain background information to her, and she did not have to take extra time to look up concepts. Speed of communication improved and required less effort.

7.3.1 Recommendations

- Look for evidence of immersion, such as:

  - Shifting modes of communication
  - Ways in which interaction with collaborators can be indicators/evidence of criteria.

8 DISCUSSION

technique the whole process what worked and what did not
  * abundance in reaching out to collaborators *

9 CONCLUSION

But what does it mean.

REFERENCES

[1] The State of the Art in Visualizing Multivariate Networks — Request PDF. Library Catalog: www.researchgate.net.

[2] S. Avle and S. Lindtner. Design(ing) 'Here' and 'There': Tech Entrepreneurs, Global Markets, and Reflexivity in Design Processes. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pp. 2233–2245. Association for Computing Machinery. doi: 10.1145/2858036.2858509

[3] R. Barcellos, J. Viterbo, L. Miranda, F. Bernardini, C. Maciel, and D. Trevisan. Transparency in practice: Using visualization to enhance the interpretability of open data. In *Proceedings of the 18th Annual International Conference on Digital Government Research*, Dg.o '17, pp. 139–148. Association for Computing Machinery. doi: 10.1145/3085228.3085294

[4] M. Birks, Y. Chapman, and K. Francis. Memoing in qualitative research: Probing data and processes. 13(1):68–75. doi: 10.1177/1744987107081254

[5] R. R. Bouckaert. DensiTree: Making sense of sets of phylogenetic trees. 26(10):1372–1373. doi: 10.1093/bioinformatics/btq110

[6] R. Buchanan. Wicked Problems in Design Thinking. 8(2):5–21. doi: 10.2307/1511637

[7] S. Greenberg and B. Buxton. Usability evaluation considered harmful (some of the time). In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pp. 111–120. Association for Computing Machinery. doi: 10.1145/1357054.1357074

[8] K. W. Hall, A. J. Bradley, U. Hinrichs, S. Huron, J. Wood, C. Collins, and S. Carpendale. Design by Immersion: A Transdisciplinary Approach to Problem-Driven Visualizations. pp. 1–1. doi: 10.1109/TVCG.2019. 2934790

[9] L. Krefting. Rigor in Qualitative Research: The Assessment of Trustworthiness. 45(3):214–222. doi: 10.5014/ajot.45.3.214

[10] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. Empirical Studies in Information Visualization: Seven Scenarios. 18(9):1520–1536. doi: 10.1109/TVCG.2011.279

[11] B. Lee, G. Robertson, M. Czerwinski, and C. S. Parr. CandidTree: Visualizing Structural Uncertainty in Similar Hierarchies.

[12] Z. Liu, S. H. Zhan, and T. Munzner. Aggregated Dendrograms for Visual Comparison Between Many Phylogenetic Trees. pp. 1–1. Conference Name: IEEE Transactions on Visualization and Computer Graphics. doi: 10.1109/TVCG.2019.2898186

[13] S. Mantzoukas. The inclusion of bias in reflective and reflexive research: A necessary prerequisite for securing validity. 10(3):279–295. doi: 10. 1177/174498710501000305

[14] N. Mccurdy, J. Gerdes, and M. Meyer. A Framework for Externalizing Implicit Error Using Visualization. 25(1):925–935. Conference Name: IEEE Transactions on Visualization and Computer Graphics. doi: 10. 1109/TVCG.2018.2864913

[15] S. McKenna, D. Mazur, J. Agutter, and M. Meyer. Design Activity Framework for Visualization Design. 20(12):2191–2200. doi: 10.1109/ TVCG.2014.2346331

[16] M. Meyer and J. Dykes. Criteria for Rigor in Visualization Design Study. pp. 1–1. doi: 10.1109/TVCG.2019.2934539

[17] T. Munzner, F. Guimbretière, S. Tasiran, L. Zhang, and Y. Zhou. TreeJuxtaposer: Scalable Tree Comparison Using Focus+Context with Guaranteed Visibility. In *Proceedings of the ACM Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '03)*, pp. 453–462. ACM. doi: 10 .1145/1201775.882291

[18] C. Nobre, N. Gehlenborg, H. Coon, and A. Lex. Lineage: Visualizing Multivariate Clinical Data in Genealogy Graphs. 25(3):1543–1558. doi: 10.1109/TVCG.2018.2811488

[19] C. Nobre, M. Streit, and A. Lex. Juniper: A Tree+Table Approach to Multivariate Graph Visualization. 25(1):544–554. doi: 10.1109/TVCG. 2018.2865149

[20] C. Partl, S. Gratzl, M. Streit, H. Pfister, D. Schmalstieg, and A. Lex. Pathfinder: Visual Analysis of Paths in Heterogeneous Graphs. In *Poster Compendium of the IEEE Sympoisum on Visualization in Data Science (VDS '15)*. IEEE.

[21] J. A. Rode. Reflexivity in digital anthropology. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pp. 123–132. Association for Computing Machinery. doi: 10.1145/ 1978942.1978961

[22] M. Sedlmair, M. Meyer, and T. Munzner. Design Study Methodology: Reflections from the Trenches and the Stacks. 18(12):2431–2440. doi: 10. 1109/TVCG.2012.213

[23] P. Sengers, K. Boehner, S. David, and J. J. Kaye. Reflective design. In *Proceedings of the 4th Decennial Conference on Critical Computing: Between Sense and Sensibility*, CC '05, pp. 49–58. Association for Computing Machinery. doi: 10.1145/1094562.1094569

[24] U. H. Syeda, P. Murali, L. Roe, B. Berkey, and M. Borkin. Design Study "Lite" Methodology: Expediting Design Studies and Enabling the Synergy of Visualization Pedagogy and Social Good. doi: 10.31219/osf.io/mghj3

[25] P. TalkadSukumar, I. Avellino, C. Remy, M. DeVito, T. Dillahunt, J. McGrenere, and M. Wilson. Transparency in Qualitative Research: Increasing Fairness in the CHI Review Process. doi: 10.31219/osf.io/auhpd

[26] C. Wacharamanotham, L. Eisenring, S. Haroz, and F. Echtler. Transparency of CHI Research Artifacts: Results of a Self-Reported Survey. doi: 10.31219/osf.io/3bu6t