Teleoscope: Exploring Themes in Large Document Sets By Example

PAUL BUCCI, University of British Columbia, Computer Science, Canada
LEO FOORD-KELCEY, University of British Columbia, Computer Science, Canada
PATRICK YUNG KANG LEE, University of British Columbia, Computer Science, Canada
ALAMJEET SINGH, University of British Columbia, Computer Science, Canada
IVAN BESCHASTNIKH, University of British Columbia, Computer Science, Canada

Qualitative thematic exploration of data by hand does not scale and researchers create and update a personalized point of view as they explore data. As a result, machine learning (ML) approaches that might help with exploration are challenging to apply. We developed Teleoscope, a web-based system that supports interactive exploration of large corpora (100K-1M) of short documents (1-3 paragraphs). Teleoscope provides visual programming workflows that have semantic and computational meaning; helping researchers to retrace, share, and recompute their sense-making process. Attempting to create qualitative "themes" rather than "topics," our NLP approach tunes an ML model to "think like you" without significant retraining. Here, we present our two-year design process and validation of Teleoscope, including a multi-week study with qualitative researchers ($N = 5$), a six-month field deployment with a qualitative research group, and an on-going public release.

CCS Concepts: • **Human-centered computing** → **User interface programming**; **Field studies**.

Additional Key Words and Phrases: datasets, machine learning, visualization

ACM Reference Format:
Paul Bucci, Leo Foord-Kelcey, Patrick Yung Kang Lee, Alamjeet Singh, and Ivan Beschastnikh. 2023. Teleoscope: Exploring Themes in Large Document Sets By Example. 1, 1 (February 2023), 28 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Exploring data at scale within a large corpus of documents is difficult, particularly if you want to interpret your data by telling a story or explaining a phenomenon. Qualitative research often focuses on interpretation, which we can think of as enriching data with context and meaning [21, 25, 45]. For large corpora, summative statistical approaches are often used at the cost of interpretation. Qualitative research often focuses on interpretation by hand, which involves extensive reading to look for themes within data and connect empirical observations to theory or outside knowledge. But, working by hand is slow and limits the scale of what can be analyzed.

Machine learning (ML) is a statistical approach to data analysis; with the advent of large language models (LLMs), modeling context and meaning is becoming possible. A common summative ML approach is topic modelling: take an

Authors' addresses: Paul Bucci, pbucci@cs.ubc.ca, University of British Columbia, Computer Science, 2366 Main Mall, Vancouver, British Columbia, Canada, V6T 1Z4; Leo Foord-Kelcey, University of British Columbia, Computer Science, 2366 Main Mall, Vancouver, British Columbia, Canada, V6T 1Z4; Patrick Yung Kang Lee, University of British Columbia, Computer Science, 2366 Main Mall, Vancouver, British Columbia, Canada, V6T 1Z4; Alamjeet Singh, University of British Columbia, Computer Science, 2366 Main Mall, Vancouver, British Columbia, Canada, V6T 1Z4; Ivan Beschastnikh, bestchai@cs.ubc.ca, University of British Columbia, Computer Science, 2366 Main Mall, Vancouver, British Columbia, Canada, V6T 1Z4.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2023 Association for Computing Machinery.
Manuscript submitted to ACM

entire dataset and try to extract large-scale topics that label and partition document sets within a corpus. In contrast to topics, qualitative researchers aim to produce cross-cutting *themes* by enhancing existing, but perhaps non-obvious, connections between documents with explanation and theory.

LLMs use document similarity metrics for common semantic understandings [1, 10, 15, 40, 49]. To help qualitative researchers, a trained LLM would need to have the same understanding of document similarity as the researcher. However, training a LLM to think like one particular researcher would be impractical. Instead, in our work, we develop techniques and tools to *refine* LLMs through iteratively defining similarity metrics according to user interactions that are already part of the qualitative research workflow.

Good qualitative interpretation requires *interaction* with data to make sense of it. The "way we think" when making sense of data can be called a *schema*. The concept of a *schema* comes from psychology and refers to a cognitive framework that helps to organize and interpret particular events, behaviour and information. We might have many schemas that serve as frameworks to help interpret different events. These frameworks are difficult to introspect (i.e., know entirely through internal thoughts), but underlie our reactions to events in the world. For our purposes in this paper, the intuition, curiosity, and ad hoc hypotheses that researchers have about their data can be understood as expressions of their underlying schemas. An important part of the qualitative research process can be thought of as identifying, expressing and externalizing the researcher's schema through labeling, arranging, modifying, and enriching data [7]. We can think of a schema as the internal cognitive structures and processes that produce our intuition and understanding that helps make sense of data. During research, as a person interacts with their dataset, they update and re-externalize portions of their implicit schemas and therefore updating their understanding. By externalizing, they are updating their explicit schemas. These processes feed back on each other.

To support thematic exploration of data [27], it would be ideal if we could train an ML model that matches the user's schema. But, training such an ML model is impractical with existing approaches: the schema is hidden, unique, and mutates every time the user interacts with data. Training and re-training such an ML model would therefore take a long time and require the user to generate many examples. Further, qualitative researchers often do not work alone; it is necessary to inspect, retrace, and share your thinking processes with other people on a research team.

To target this space between human and machine sense-making, we built Teleoscope, a system to support multiple users interactively exploring a large corpus based on LLM's document similarity. With our system, users can iterate, collaborate, and group documents based on their own understandings of document similarities. Our system will then produce 'fast enough' and 'good enough' models by dimensionally reducing the LLM based on user-defined groups so that users can inspect and update their cognitive schemas alongside the ML models.

Problem: Thematic exploration of data by hand does not scale. But ML models cannot be easily used because they are not specific enough to capture a particular researcher's schemas. Further, researchers need to update their schemas as they explore data, especially as they share findings with others.

We present *Teleoscope*, an interactive web-based system for exploring large datasets (100K-1M) of short documents such as social media posts. Teleoscope was designed to assist qualitative researchers during the discovery, data collection, and organization phases of a research study. Using Teleoscope, researchers can quickly search for themes across a document corpus, organize documents, and test ad hoc hypotheses before committing to a deeper analysis. Teleoscope's interface is a visual data workflow editor that controls a cloud-based backend of machine learning (ML) workflows which calculate document similarity and clustering based on a user's document grouping and workflow organization on the interactive workspace (see Fig. 1).

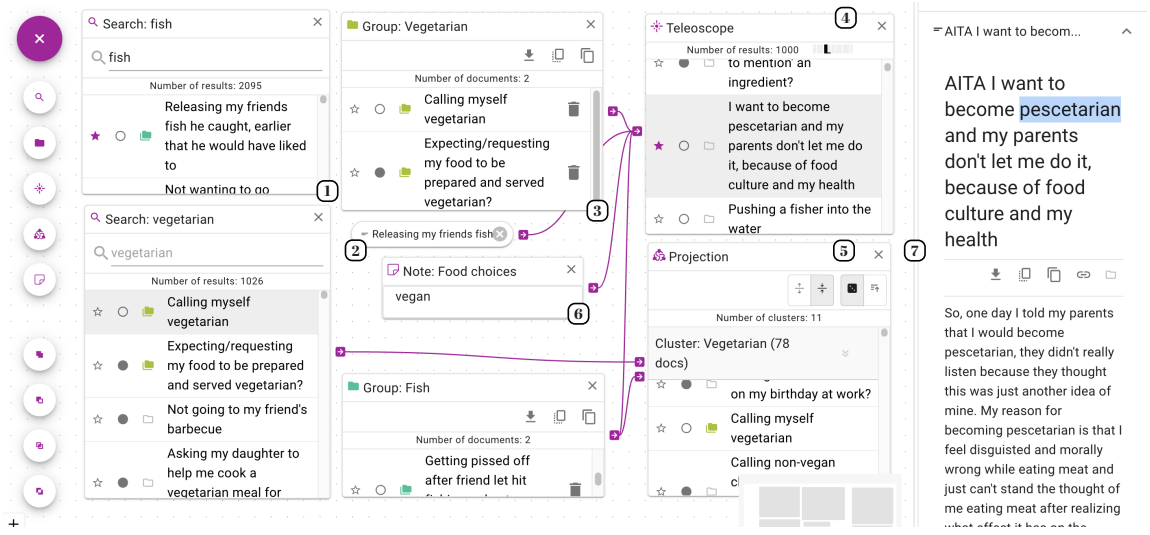

Fig. 1. (1) Users start by performing a **keyword search** to explore documents; (2) **Documents** are dragged onto the workspace; (3) Documents can be put into **groups** for organization; (4) **Teleoscope** nodes can use documents, notes or groups as control inputs; (5) **Projections** create clusters using groups as control input; (6) **Notes** can contain arbitrary text which is also vectorized and can be used as a control input to a Teleoscope; (7) the **sidebar** has a quick viewer for documents, saved items, bookmarks, and settings. Keyboard navigation is included for quick exploration and group creation.

A key difficulty in using ML to support exploratory qualitative research is to define the right document similarity metric. This is because document similarity from a machine perspective can be quite different from a user's understanding of document similarity (see Fig. 4). Teleoscopes's natural language processing (NLP) system is a special case of semi-supervised topic modelling based on document similarity. However, the machine-generated outputs of topic modelling are often disappointingly general and miss the nuance that a researcher would prefer to have about their document groups [38, 44]. A researcher may decide that two documents that seem very different to the machine are part of the same theme. Or a machine may decide that two documents are very similar that a researcher would not think belong together. Teleoscope uses researcher's understanding of similarity as expressed through grouping and mixing documents to define and express a distance metric that is used for dimensionality reduction and clustering.

Teleoscope is best thought of as a tool for *thematic exploration* rather than topic modelling. Our key insight is to continually update document representations based on user interaction that are *good enough* and *fast enough* to support iterative sense-making. We use the natural interactions of a qualitative researcher when interpreting documents such as *arranging*, *annotating*, and *grouping*. Teleoscope translates these interactions into inputs for a custom distance metric for ranking, dimensionality reduction and clustering. Put another way, the user starts from a large model and proceeds to refine it by (1) exploring documents in terms of keyword search; (2) discovering document similarities; (3) grouping documents together to indicate their personal understanding of similarity; (4) "mixing" them by connecting them to a *teleoscope*[1] operation to discover what the machine "thinks" is similar; (5) building up a workflow graph (used as input for dimensionality reduction); (6) iterating and refining input (see Fig 1).

[1]Our on-screen operation was unfortunately named for our tool early in the design process and has not been revised yet. In this paper, we use lower-case and italics to refer to the operation, and capitalize the system. Our apologies to the reader.

We want to use the power of ML on large corpora, but human cognitive schemas are typically implicit. ML systems need data to be explicit, but making too many definitive choices too early can cut off important interpretive directions for a qualitative researcher [12]. Iteratively making schemas explicit allows us to slowly approximate and develop a researcher's schema by customizing an ML model to their research context.

We designed Teleoscope to make cognitive schema expression as explicit as possible by creating computable visual traces so that researchers could retrace, inspect, and modify their schemas at the same time as refining machine data representations. Effectively, we wanted to make a computable mind map that corresponds directly to an ML model state (an ML schema to match a cognitive schema). The word "Teleoscope" is a portmanteau of telescope and the Greek word *telos* (as in, teleology), which refers to explaining the *purpose* of something. Teleoscope is a way for researchers to intuitively explore and explain meaning and purpose for their datasets.

As such, our design goals targeted the creation of a system that allows continuous live interaction, as opposed to many topic modelling packages that run as python programs in the terminal [16, 17, 41, 42]. Similarly, we wanted to a visualization system that captured the *process* of developing a thematic understanding of the data so that research collaborators could share their thought processes with each other. This is as opposed to topic modelling visualization systems that focus on high-level statistics, topic hierarchies, word counts, and labels [17, 23, 48]. Finally, we believe that the embodied experience of arranging documents is an important part of creating the sense of understanding. To support our target users—who are not computer experts—we wanted to make a web platform that allowed for fast enough interaction that users could feel as if they were having an embodied, creative, improvisational experience with the data. Our design goals are summarized here:

- **DG1. Create an interactive system where researchers directly manipulate data to produce ML models which then feed back into their understanding of the data.** Imbue common qualitative interaction metaphors with computation; specifically those that correspond to actions that qualitative researchers are familiar with from data interpretation and exploration, such as iterative *arranging*, *annotating*, and *grouping* of documents. Qualitative researchers often use emergent interpretive processes such as physically re-arranging and annotating documents until patterns and meanings emerge.
- **DG2. Help researchers externalize their thought processes (cognitive schemas) as visual traces that can be *retraced, modified,* and *inspected* by collaborators.** When interpreting data, capturing the steps of a thought process is useful when inspecting and changing your conclusions, and then later sharing them with others. If the visual representation traces your thought process but also has computational meaning, then changing the representation gives feedback to your schemas and vice-versa. Then, you can modify the representations when collaborating with other team members or reviewers to inspect, share and modify workflows and understandings.
- **DG3. Facilitate an on-going feeling of improvisational, creative, in-the-moment data exploration for researchers.** The feeling of continual interactivity was important to us to facilitate through our system design choices. We wanted users to be able to continually interact with the system while inevitably large computations completed. We also wanted to limit the scope of computations to produce results that are *good enough to inspect, but not perfect* since we expect them to be quickly iterated on.

To support our design goals and evaluate our system, we ran three field deployments. In the first deployment (N=5), we asked qualitative researchers to use Teleoscope to explore a shared topic over the course of a few weeks, then discuss their experience and findings at a focus group. For the second deployment, we recruited a three-person qualitative research team to use Teleoscope as part of their data collection and analysis process and responded to their design

requests. For the third deployment, we released a live public version of Teleoscope which is currently running and available. This helped us to gain insight into our research goals.

At a high level, we were interested in seeing whether we had created a tool that would be effective in supporting qualitative researchers in the data collection phase, particularly in externalizing their cognitive schemas. We wanted to iteratively develop interaction and visual metaphors that capture these schemas and ensure that they are understandable to non-expert users. Once we developed the tool to a deployable level, we wanted to study how Teleoscope was used in a real collaborative research study.

Although our main design inspiration and target demographic are qualitative researchers, our studies demonstrate that Teleoscope supports people in general-purpose text data exploration and discovery. We performed a software design process over two and a half years which was heavily informed by our own use of Teleoscope for qualitative analysis.

Our design goals, design studies, and software design process have produced the following contributions to the CHI community:

- **C1. Collaborative, interactive cognitive schema data exploration and sense-making.** We contribute a system that concretely demonstrates a method for expressing and updating cognitive schemas alongside ML models as a form of collaborative data exploration in large corpora. We also provide our simple-yet-hard-to-develop interaction metaphors, assembly of ML approaches, and systems engineering approaches that facilitate the real-time interactive feeling of continuous improvisation required to make cognitive schema development possible alongside ML model interaction. To our knowledge, no other system has attempted to demonstrate a design for aligning machine and cognitive schemas, particularly at scale with a real-time system.
- **C2. Design studies of cognitive schema data exploration.** We further contribute studies of our novel approach that detail both development and long-term field deployments of the system. Our studies are instructive for the community in that they detail a long a difficult design process for a simple outcome, and culminate in an unique long-term field deployment that involved a full real in-situ research use case. They further shed light on how qualitative methodologies can be understood with collaborative ML systems.
- **C3. An open-source, usable, and adaptable product that can be collaboratively used by multiple team members.** Teleoscope is provided as a research product that is both publicly deployed at [anonymized for review after consultation with CHI chairs] and is an open-source software for others to adapt and deploy themselves. Multiple users are able to collaborate in exploring over 500K documents on the same workspace; the deployment features user authorization, is cloud-native, works in the browser, and is robust enough to run continuously.

In this paper, we will address *Related Work* in Section 2 where we will explain the theoretical motivation for Teleoscope for qualitative methodologies, differentiate Teleoscope from other visualization and topic modelling tools, and explain the NLP approaches we are drawing from. We then explain the design of our system and motivate it with respect to qualitative researcher needs in *Teleoscope Design* in Section 3. After, we outline our participant-focused *Design Studies* in Section 4, wherein we performed an iterative low-cost design process, and three long-term field deployments to test and refine the viability of our approach. We discuss the results of our studies in light of our design goals and research questions in Section 5, and finish with a short *Conclusion* in Section 6.

2 RELATED WORK

In this section, we outline the theoretical positioning of Teleoscope within the literature. We explain the relevant technologies, outline the theoretical connection to qualitative research, and motivate our design decisions relative to common approaches for creating interfaces for exploring large document sets and topic modelling. Teleoscope draws from three categories of related work: data visualization, the NLP field of topic modeling, and computer-supported sense-making as articulated through qualitative methodological theory and approaches.

As an interactive system, Teleoscope primarily takes inspiration from topic modelling visualization and qualitative analysis software. Similar to common topic modelling visualizations addressed below, Teleoscope structures and annotates a vast document space with labels and groups, supports creative and interactive exploration of documents along with document structures, and uses human-in-the-loop machine clustering of documents. We differentiate ourselves by taking a provenance-based approach in that we make the questions of "what did we find?" and "how did we get here?" the primary focus of manipulating objects in the interface (**DG1**). As a result, we further differentiate ourselves in terms of facilitating (1) collaborative sense-making (**DG2**); (2) at a large scale while remaining improvisational (**DG3**); and (3) targeting themes rather than topics.

2.1 What qualitative researchers need

Teleoscope's design is largely inspired by the qualitative practice of thematic analysis [9] and the methodological considerations therewith. This is well captured in the following quote by Nowell et. al:

> *To be accepted as trustworthy, qualitative researchers must demonstrate that data analysis has been conducted in a precise, consistent, and exhaustive manner through recording, systematizing, and disclosing the methods of analysis with enough detail to enable the reader to determine whether the process is credible. [39]*

Teleoscope takes the same value propositions from the *analysis* phase and applies it to the *data collection* phase. Due to the human-in-the-loop ML modelling approach and the large amount of data, the collection methodology must also satisfy the standards of being precise, consistent, and exhaustive. However, at scale, there may be hundreds or thousands of documents that are thematically identical for a particular study. Researchers need to determine which documents are most similar within a corpus, structure the documents into groups, then find exemplar documents to construct a sample for more in-depth analysis. Teleoscope supports this data curation process by providing potential theoretical rigour that would apply standard thematic analysis by hand by focusing on visualizing the *provenance* of the ML model state (see below for a discussion on provenance).

How then do we decide which documents are worth looking at? The Teleoscope approach prioritizes the sense-making *process* which is best understood by the concept of making tacit *schemas* explicit, as articulated by Berret and Munzner [7]. In quantitative methodologies, the standard approach is that scientists formulate their research as a series of falsifiable claims which are supported or rejected due to empirical evidence. In empirical qualitative methodologies, a researcher's job is to *interpret* empirical evidence so that it makes sense for both the researcher and the reader. Berret and Munzner explain the act of sense-making as moving between tacit schemas (i.e., gut feelings, unconscious ideas) and explicit schemas (i.e., drawings, writings, and other data representations). By turning a tacit schema into one or more explicit schemas a researcher inspects their tacit schemas and makes them available for critique by others.

Importantly, sense-making is an iterative process that starts from ambiguity. Chen et al. [12] argue that ML support for qualitative analysis in social science research needs to "*[shift] the focus to ambiguity.*" To paraphrase, computer scientists often create supposedly definitive and accurate ML models of textual data too early: before there has been a

period of ambiguity. This is antithetical to qualitative definitions of academic rigour where it is considered important to preserve ambiguity for as long as possible. This allows for more time to make connections between data points and enrich theoretical insight while developing themes about the data. Good themes are *rich* (more interconnected) rather than *accurate* (categorically definitive) [33].

To balance the tension between making schemas explicit and retaining ambiguity, we designed Teleoscope to show users results that were "good enough" and "fast enough." This motivated our choice to make an asynchronous processing system: users can continue to explore and arrange documents while waiting for a longer background ML calculation to complete. Since the act of arranging documents takes place within a node-based visual workflow editor, the arranging and linking of elements also has explicit computational meaning which is continually updated as the user "updates" their implicit cognitive schema.

Teleoscope contributes to qualitative methodologies by importing standards of analysis into standards of document collection at scale (**DG3**). It also demonstrates a method for making the process of tacit sense-making explicit by asking the user to primarily manipulate workflows which represent their sense-making process (**DG2**). Unlike other systems, Teleoscope makes a schema expression an explicit goal, and provides a direct correspondence between the schema as expressed through visual workflow and ML model (**DG1**).

2.2 Visualization Approach

2.2.1 Provenance. Our choice to focus on visualizing provenance comes from a desire to make an ML system that can be used by qualitative researchers who are not computer experts, yet still need to explain their use of ML system to stakeholders such as collaborators, supervisors, and clients [3] (**DG1**, **DG2**). Many topic model visualizations choose to directly visualize the underlying document clusters, that is, the *result* of the clustering process [38, 44]. In contrast, Teleoscope focuses on capturing and displaying the exploratory *process*. In the systems and visualization communities[50, 51], this is referred to as *provenance* tracking. For making sense of data, the importance is placed on creating a visual trace of a history of user's interactions with the data to answer not just "What are the results?" but also "How did we get here?" [20]. A standard provenance system that we are all familiar with is *undo/redo*; a visualization of undo/redo could be as simple as a list of previous actions performed on the interface.

Beyond simply keeping an undo/redo history, researchers are very interested in creating reproducible and explainable results [46, 50]. This helps during the process of analysis, publication, and review. For Teleoscope, keeping track of the provenance of an ML result means tracking the inputs and user-selected data processing elements in a chained workflow. The user is engaged in a process of data wrangling for sense-making [5, 8], i.e., *exploration*, *annotation* and *curation* of data [34]. In *Trrack* [14], Cutler et al. demonstrate a library for tracking branching histories for actions on web-based visualizations. This is a maximal approach for provenance, where every action is tracked and is reproducible. Histories can be linear, branching, or networked, allowing for both high-scale overviews and detail-on-demand approaches [51].

Although the Teleoscope system does record every user action in a comprehensive history system, we hide most of that history from the user. Instead, we rely on the user creating small, comprehensible workflows that they commit to via grouping and connecting in a workflow graph. In Xu et. al's taxonomy of topic model approaches [50], we are modeling coupled user and application state via an entity graph using semantic interactive visual analysis.

Making document sources the primary manipulation objects in a workflow was not our first design choice nor is it often what visualization systems choose. Similar to other systems [13, 19, 24, 26, 48], we first tried keywords and topic lists as the primary manipulation objects using a dashboard metaphor. However, we arrived at design insights that

(1) the process of iteratively arranging documents itself is critical to sense-making; and (2) the spatial arrangement of documents often captures their relationships, which are important for remembering and inspecting thought processes.

Teleoscope uses concepts and design ideas from existing provenance systems, but differs by applying these ideas to qualitative understanding of corpora and topic modelling. Rather than capturing a full history of user actions, we chose to make provenance traces our primary manipulation objects. Therefore, researchers are directly creating the story of "how we got there" in a visual form that can be shared with collaborators, inspected, and updated (**DG1**, **DG2**).

2.2.2 Topic Modelling and Themes. With Teleoscope we depart from standard topic modelling because we are not interested in creating *categorical* topics, but instead are interested in supporting researchers in the first stage of developing rich, specific *themes*. The difference between categories and themes is subtle, but it is important to qualitative researchers. Rather than creating deductive categories, qualitative researchers are interested in inductive interpretation.

The difference is illustrated well with *Cody*, Rietz et al.'s excellent qualitative coding support tool [43]. *Cody* addresses the problem of annotating specific sentences with qualitative codes (similar to tagging) by connecting keywords with set operations to create inclusion/exclusion rules. These rules are then applied across a corpus. This supports a style of coding that is fundamentally categorical, that is, the main action of analysis is to organize the different sentences into code categories. Although this is very useful from an organizational standpoint, the tendency is to create code hierarchies that capture more and more documents; by contrast, a good theme from a thematic analysis standpoint is very specific and involves interpretation. Teleoscope supports an interpretive approach since we are *not* attempting to support users in *labeling* documents; instead we focus on document *relationships*.

Many topic modelling visualizations take the approach of acting directly on topics. For example, *Serendip* [2] and *TopicSifter* [23] both approach the problem of topic exploration and discovery by allowing users to filter and/or drill down into topics. For these tools, the topics are the primary objects of interaction, and the goal is to explore the topics themselves. Both are a top-down approach. Teleoscope is most conceptually similar to *Scholastic* by Hong et al. [22] because we also take a human-centred approach and provide a hierarchical clustering algorithm as a "machine-in-the-loop" approach. Their elegant approach of visualizing the document space differs greatly from our choice of visualizing the process of discovery. Also, as with *Cody*, they focus on document coding, which we do not.

Teleoscope differs from the above tools in that it does not attempt to visualize, categorize, or make claims about the entire corpus. Instead, we take a bottom-up inductive approach which allows for orders of magnitude larger exploration, but with orders of magnitude less to display. In this way, although we use semi-supervised topic-modelling NLP methods (explained below), our conceptual approach does not manipulate topics directly; instead the user interacts with the ML model implicitly while interacting with documents and groups similar to how they would typically with qualitative analysis and thematic exploration. Put another way, we are making mini-topic models from of a restricted subset of documents based on customized distance metrics that we are calling *themes*.

Teleoscope differs from topic modelling tools by focusing on (1) discovering relevant documents rather than making claims about a whole corpus; (2) supporting collaboration, reproducibility, and schema expression; (3) co-developing themes with the computer through creative data exploration; and (4) allowing users to interact in real time with a larger corpus than other tools (**DG1**, **DG2**, **DG3**).

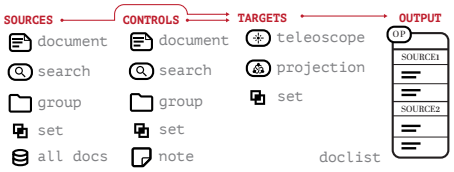

Fig. 2. Above are the different types of nodes and operations in the workflow graph. Document **sources** can be a single document, a group of documents, the results of a keyword search, or all documents in the corpus. **Controls** for the **target** Teleoscope ranking operation or the Projection clustering operation can be any subset of the corpus or a note (which includes arbitrary text). Operations **output** a list of document lists which are partitioned per document source.

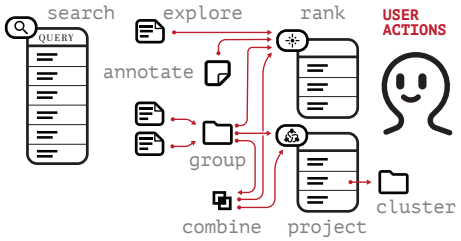

Fig. 3. Users start a new workflow using a keyword **search**. Then, they can **explore** documents by dragging them onto the workspace, skimming titles and reading. When they find documents they like, they can **annotate** and **group** the documents; they can further **combine** sources with set operations. Any source can be **ranked** according to controls. Teleoscope can also **cluster** documents for them into groups.

2.3 NLP Approach

To facilitate the concept of *themes* over *topics*, we take the approach of *modelling by example* to allow users to explore and then structure the document space. Lissanddri et al. [28] describe example-based search approaches as having a resurgence in popularity. A variety of example-based query systems have been introduced that attempt to synthesize the query from examples [18, 28, 29]. Teleoscope follows these works by using semantic similarity to drive exploration (see Fig. 4 and Section 3.1 for details). Teleoscope differs from these works by not attempting to construct deterministic queries or models of the entire corpus. Instead it relies on the interaction process and the user's own sense of saturation to determine the extent of the exploration.

When a Teleoscope user has finished a phase of exploring via semantic similarity, they can switch to structuring the document space via semi-supervised dimensionality reduction. Variations on this approach are used in recent human-centred ML and visualization systems [4, 17, 32, 47]. The premise is to take a large language model and reduce dimensions along which a similarity metric is defined. Teleoscope uses the Universal Sentence Encoder (USE) for the base exogenous model, which encodes all documents as 512-vectors [11]. In Teleoscope, we use grouped documents as control inputs to define the similarity metric (i.e., this is the "supervision" part of our semi-supervised topic model approach). We use Universal Manifold Approximation and Projection (UMAP) for reduction to five dimensions [31]. Clustering uses Hierarchical Density-Based clustering (HDBSCAN*) [30].

Teleoscope uses a typical assembly of already-existing NLP tools, but uniquely capitalizes on a conceptual and practical correspondence between semantic example-based search, dimensionality reduction, and cognitive schemas (**DG1**).

3 TELEOSCOPE DESIGN

In this section, we outline our design decisions for Teleoscope. Our design process took over two years and involved many iterations. As we used Teleoscope for our own data exploration, demonstrated Teleoscope to others, and performed initial pilot studies, our engineering and UX requirements expanded.

Our design takes the most direct inspiration from applying processes used in thematic analysis to data collection. Importantly, Teleoscope does not facilitate thematic analysis, but instead applies ideas from thematic analysis to data

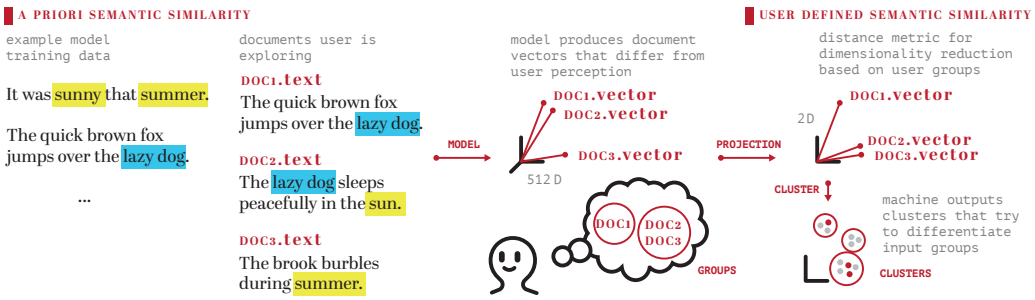

Fig. 4. ML similarity scores are based on vectorizing text and measuring the distance between vectors. In the above example, a model would assess *doc1* and *doc2* to be most similar to each other because they share the phrase *lazy dog*. A sophisticated model, with enough training data, may be able to establish semantic similarity between *sun* and *summer*.
The similarity score between *doc2* and *doc3* would still be correctly measured as lower than *doc1* and *doc2*. However, as a qualitative researcher who is performing interpretive research, you would want the ML model match your personal interpretive schema. In this example, you would want the model to emphasize the similarity between *sun* and *summer*.

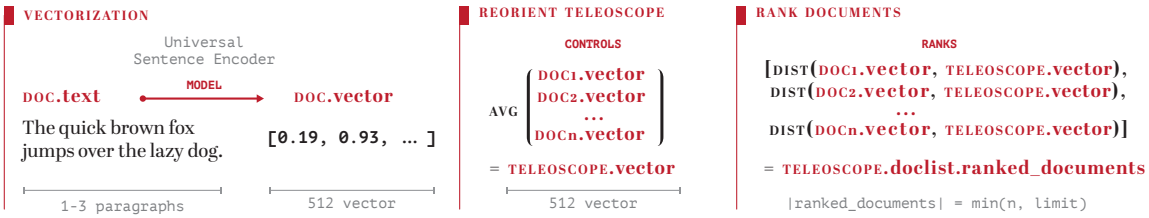

Fig. 5. Teleoscope encodes all documents using the same model. We chose to use the Universal Sentence Encoder (USE) model, but any model that takes text and outputs vectors could be substituted. Teleoscopes are oriented by creating a Teleoscope vector, which is an average over all vectors from nodes connected to the control input, which may include any source, i.e., documents, groups, searches, or notes. Each document in sources is ranked by distance to the Teleoscope vector. If no sources are indicated, all documents in the database are ranked. For frontend display speed, we limit the number of stored ranks.

collection to increase theoretical and practical rigour. For thematic analysis, Braun and Clarke emphasize that during the data familiarization phase, there needs to be *active* engagement with the data:

> *Use whatever format works for you (e.g., annotating transcripts, writing comments in a notebook or electronic file, underling portions of data) to highlight items potentially of interest. Note-making helps you start to read the data as data. Reading data as data means not simply absorbing the surface meaning of the words on the page, as you might read a novel or magazine, but reading the words actively, analytically, and critically, and starting to think about what the data mean. [9]*

We contend that active engagement via annotation, direct manipulation, and exploration is vitally important to exploring an opinionated data collection at scale. Arranging and re-arranging data allows patterns to emerge; in Teleoscope, data workflows can be arranged and rearranged as well as individual documents. Further phases of thematic analysis are also reflected in the Teleoscope workflow, which are referenced below.

Creative, engaged interaction with data required systems engineering choices that made real-time feedback possible, motivating our development of a drag-n-drop interface with a distributed backend. We initially designed Teleoscope as

a simple dashboard interface for manipulating keywords with a bag-of-words model, however, found that manipulation was an important part of sense-making.

After about a year of our own design iterations, we moved into a second phase of user-driven design by recruiting social scientists for an in-depth usability study, analyzed the study results with a visualization research group, and then performed a field deployment with a dedicated qualitative research group. We also designed the system to be robust enough to be used by the general public and released Teleoscope openly on the internet. Our studies and data collection are detailed in the next section. This section describes the outcome of the design iterations so that the reader is able to understand the current Teleoscope system design, which is the third major redesign and includes design up to and including the public release.

3.1 Teleoscope interface concepts

Teleoscope consists of an infinite whiteboard *workspace* where users connect data *sources* to computational operation *targets* to create a computational graph. There is also a sidebar for quick navigation and reading. A chain of sources are used to control the ML operations and form the basis of our workflows. From an interaction perspective, the workflows become curated digital artifacts that leave a trace of the researcher's exploration of the data. From a computation perspective, the workflows are user-curated provenance graphs.

3.1.1 Workspace. We developed the Teleoscope workspace to reflect the process of arranging data on a table, due to conversations with qualitative researchers about needing to arrange data to make sense of it (**DG2**). We wanted Teleoscope to reflect a researcher's personal meaning of their data arrangements with computational meaning in the ML model (**DG1**). Therefore, the Teleoscope workspace is a drag-n-drop visual workflow editor which produces computational graphs. We have two types of nodes in the workflow graph: document *sources* and *target* operations. We chose a direct manipulation metaphor for positive transfer from whiteboard apps that qualitative researchers are familiar with (such as Miro) and to mimic the process of arranging and rearranging documents on a physical desktop such as in affinity diagramming. Based on user feedback, we settled on a whiteboard workspace metaphor with a visual workflow editor. We also implemented a quick viewer in the right hand sidebar and keyboard navigation for skimming document titles and body text. See Fig. 6 for an annotated depiction of the following sections.

3.1.2 Sources. A *document* is the primary interaction object in Teleoscope. Documents display a title, text and metadata information within the node or in the quick viewer when selected. Best performance for documents is 1-3 paragraphs, such as social media posts. Users can review documents to decide whether they are topically relevant by grouping, bookmarking, or using them as input to operations. A *search* is a fuzzy keyword search across the full document set; queries are in plain text and support set operations like *and*, *or*, and *not*. Keyword search was provided since this is a common search interaction style that qualitative researchers have often been explicitly trained in and are used to thinking about. Documents can be sorted into labelled *groups*, either by dragging and dropping into a group node window, or by selecting the group in a drop down. A *note* can contain arbitrary text, which is then vectorized live as the user types. Users can use notes to create annotations, or they can use the note to drive *teleoscope* operations.

Computationally, all sources are lists of one or more vectors. Interactionally, sources are meant to be treated like (1) document collections; and (2) examples that will drive the ML system to find relevant documents according to what the user arranges to be together.

3.1.3 Targets. Targets are the ML operations placed at the end of a workflow and produce collections of document collections as outputs. Each *source* will create a new document collection, whereas each *control* will manipulate distance metrics and similarity scores. Targets are meant to reflect the user's mental model as they explore the data so that the user can update their mental model along with the machine model.

The *teleoscope* operation ranks source documents relative to the average vector of all control documents (see Fig. 4). If no sources are connected to a *teleoscope*, it will rank all documents in the corpus; otherwise it will rank each source subset independently.

The *Projection* operation runs semi-supervised dimensionality reduction and *clustering*.

Control inputs to a projection define a distance metric: if two documents are in the same group, their distance is set to a minimum; otherwise, their distance is the cosine distance between their vector encodings. Users can choose to include an additional rule in the distance metric that sets the distance between groups to a maximum. Doing so forces the system to separate groups when clustering; otherwise, groups may be clustered together.

In the projection operation, *source* inputs can be used to define the domain of the documents being clustered. If no *source* inputs are provided, users can decide between a random subset of the corpus, or a selection of documents ranked relative to the average vector of *control* inputs.

3.1.4 Source and target operation: Set Operations. Set operations are provided so that users may non-destructively combine sources and inspect results. They include standard set operations of union, difference, intersection and exclusion (the complement of intersection). They are the only operations which are both sources and targets; they can be combined to create chains of operations (see Figure 6).

3.2 Workflow Design Patterns

3.2.1 Rank. The Teleoscope ranking operation can be used to explore documents from the whole corpus, but if a source is connected, the document list can be ordered relative to the control. The controls can be any source; in Fig 6, we can see a search source being ranked by a note control.

3.2.2 Chaining Set Operations. Set operations can be used to join and filter documents as well as intermediary connective operations. In Fig 6, we can see that set operations are chained to get a specific document source which is then ranked by the Teleoscope operation. Set operations can be used for non-destructive editing of groups, construct larger sources out of smaller sources, or to narrow down large groups. When combined with machine-generated groups after clustering, they can also give insight to the way in which the machine clusters overlap with the user-defined groups, either providing confidence in a user's group choice or indicating an area for further exploration.

3.3 Technical systems design

Reflecting our goal of real-time interaction (**DG3**), Teleoscope's backend is engineered to support continuous iteration and interaction. This was a non-trivial task which is not always supported by other systems due to the high computational workload. Through internal testing, we found a distributed backend was necessary to run Teleoscope computations without blocking user interaction. We chose to precompute, cache, and distribute as much as possible to make on-the-fly calculations seem quick. For example, a Teleoscope rank of cached data is nearly instantaneous despite sorting hundreds of thousands of documents (average 1601 ms for update on UI including network latency). For operations that are not instant, the interface allows continuous interaction with other elements. We believe that this user experience goal is *vitally* important to Teleoscope being capable of facilitating the flow state needed for sense-making.

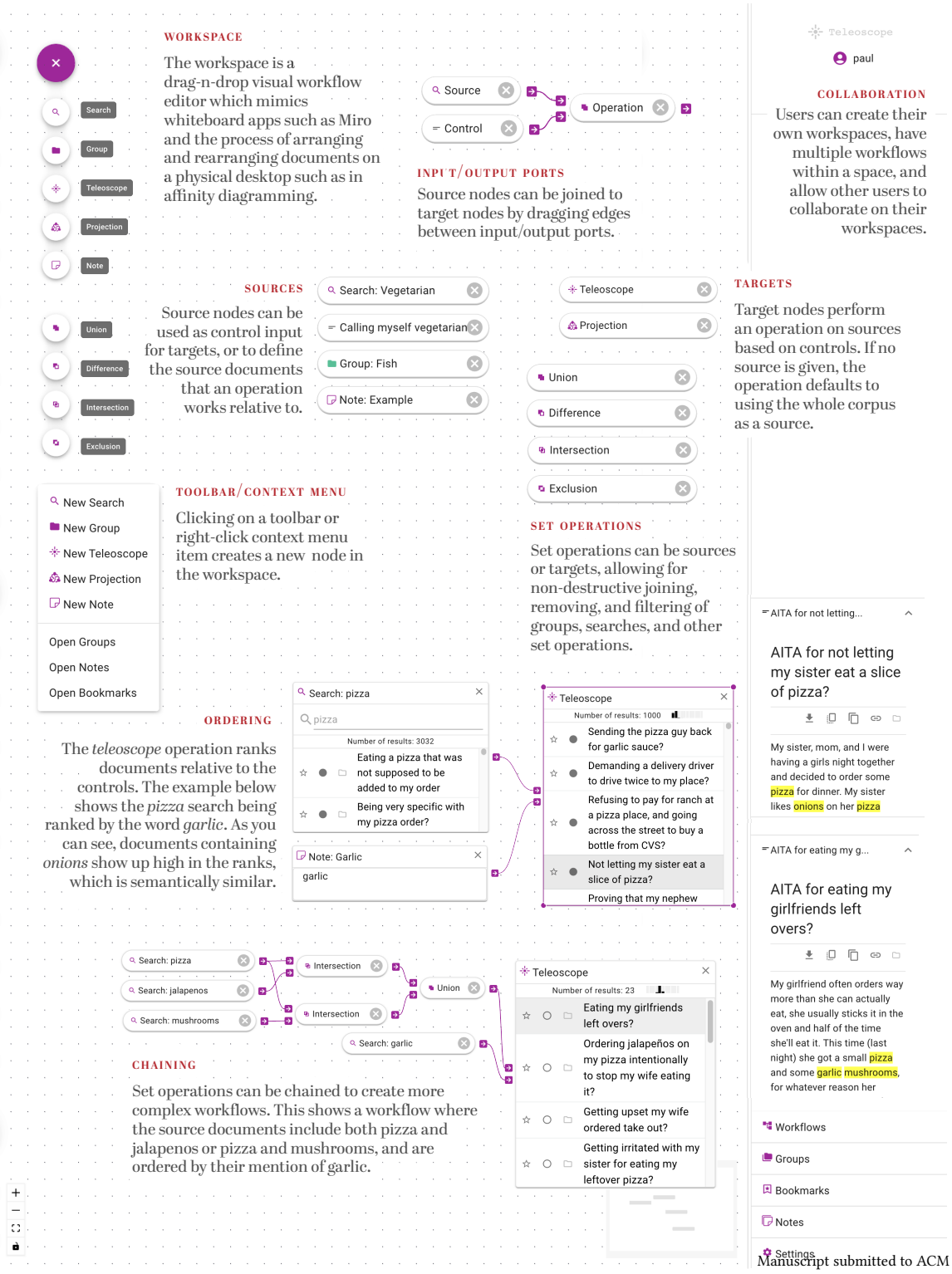

Fig. 6. The above is an annotated illustration of the Teleoscope workspace, explaining workspace operations and giving example workflow patterns. Annotations are in serif fonts whereas the workspace items are sans-serif. See Figs. 1 and 2 for detailed window depictions and conceptual overviews of operations.

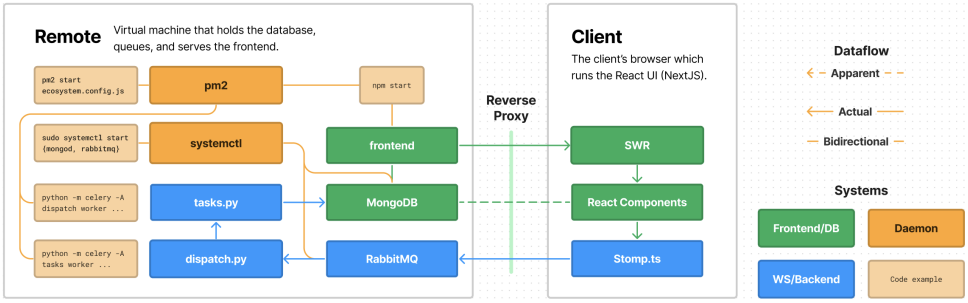

Fig. 7. We designed a strict dataflow for Teleoscope to aid with concurrency and to mask remote processing latency. The NextJS client uses a polling strategy with the Stale-While-Revalidate (SWR) library to pick up state changes from MongoDB. The client sends API requests to RabbitMQ and concurrent Celery workers read from RabbitMQ and write to MongoDB. Dispatch and task modules are split and can run on different virtual machines for load balancing.

From a community perspective, we wanted Teleoscope to be available for qualitative researchers who are non-computer experts, which necessitated a robust enough system to survive a production-level environment on the open internet, including almost-one-click deployment, user accounts, security, and backup systems. Teleoscope is continuously available live at [anonymized for review after consultation with CHI chairs] for use by the general public. We have an active user community on Discord where we take bug reports and design requests.

3.3.1 Frontend. Our frontend is built using the NextJS ecosystem to manage React development and deployment, ReactFlow for the graph drag-n-drop workflow system, and Material UI for the design elements, as well as a variety of smaller libraries and custom components. NextJS was chosen because of its large user community and full stack support, including data fetching and user authorization libraries. We chose ReactFlow after experimenting with a number of whiteboarding, windowing, and drag-n-drop libraries; it is also a mature and actively maintained freemium product. Material UI implements Google's material design in React.

3.3.2 Backend. Teleoscope uses a distributed backend with RabbitMQ for messaging and Celery to execute tasks. Our database is MongoDB. To ensure continual service, our system is daemonized with native linux *systemctl* as well as the *pm2* library for node and python applications. Both perform process monitoring and memory management.

3.3.3 Dataflow and History management. React uses a virtual document-object model (DOM) to ensure a strict dataflow model for user actions and system state. Similarly, we designed a dataflow policy such that the frontend (almost) entirely makes requests to the backend to manage system state on the server. This means that actions that mutate database state are strictly sent via a secure websocket connection to RabbitMQ through a well-defined API. With the exception of user registration, there is no direct database mutation by the frontend. Similarly, backend state updates are managed by the Stale-While-Revalidate (SWR) data caching and fetching system in NextJS.

Keeping this strict policy has benefits and drawbacks, mostly having to do with interaction availability. Any large-scale calculations and mutations are offloaded to the backend while the frontend waits for data to be marked as stale to refresh the local client cache. The trade-off is that some state changes that require a backend response may be impacted by network latency.

4 DESIGN STUDIES

This section outlines our design study processes. We evaluated Teleoscope through (1) informal piloting and internal analysis using low-cost evaluation methods; (2) a multi-week study and focus group with qualitative researchers (N=5) including a post-hoc expert review of our interface and study data with a visualization group; (3) a multi-month field deployment with a qualitative research group; and (4) an on-going public release. We used data from Reddit as archived by PushShift [6] up to their latest published data in January 2023. Except for initial in-lab piloting, all research was conducted under approval of our institution's research ethics board and all participants signed initial consent forms and were reminded of their right to halt participation in the studies if desired.

We focused on one subreddit, the *r/AmItheAsshole* advice forum since we were ourselves interested in data on social norms; our field deployment collaborators also wanted access to *r/nursing*. *r/AmItheAsshole* has roughly 650K documents and *r/nursing* has roughly 100K documents. Data included only posts, not comments. The last post date was February 2023, since Reddit significantly restricted their data API access in response to ChatGPT.

4.1 Design Study 1: Initial piloting and internal analysis

Teleoscope was initially developed for our own use as a research tool for searching through large corpora of text and went through a series of very quick design iterations with a single designer. As we developed it into a more robust and large-scale system, we began to incorporate more human-centred design techniques into our design process as our research and development team grew. Since this was early stage design, the interface design changed significantly as we determined design requirements. We are including only light details on our informal methods so as to faithfully report on our process. During this year-long phase, we also ran a series of tests on different NLP approaches. These were incorporated into our system during our informal user testing.

4.1.1 Participants. For our informal methods, the users we refer to are members of our design team and the larger lab members who were not involved in Teleoscope development. In terms of expertise, all users were trained competent-to-expert computer scientists; some members of our team are trained competent-to-expert UX and qualitative research practitioners.

4.1.2 Methods. We used a variety of informal low-cost UX evaluation methods to motivate our early design choices, including cognitive walkthroughs, heuristic evaluations, and informal observations with both people from our design team and from our larger research lab. The cognitive walkthroughs and heuristic evaluations were performed with standard methods with heuristics taken from the Nielsen Norman group [35–37]. Informal observations were performed on low-level interactions such as menu clicking and basic keyword searches to discover and amend heuristic violations.

4.1.3 Results. The results of our initial evaluations were a set of guiding backend and frontend design requirements that aligned with standard UX heuristics. We summarize the most relevant heuristics here to explain our early design directions:

Visibility of System Status. Our original design used a dashboard metaphor where each module displayed system state such as included/excluded keywords, document similarity statistics, and topics. Our initial corpus visualization attempts repeatedly pointed towards common visualization solutions of weighted adjacency matrices of keywords and documents, but with corpus sizes of a million documents, pixel overlap became a problem very quickly. Further, the connection between modules in a dashboard is hidden. Therefore, we decided to move towards a windowing system, eventually creating window modules that had visible input/output areas.

Recognition over recall. By moving commands and system state out of menus/collapsible dashboard modules, we opted for a design with minimum display of information. This is contrary to an approach that attempts to display the full system state through statistics or summary graphics. We deemed this unnecessary for our core interaction goal; opting instead for an Overview/Details-on-Demand design pattern where the *process* was visualized rather than the full system state.

Error recovery. We created a robust history system where every system action is logged. After many discussions about how much history to display to the user, we opted for an algebraic workflow metaphor. This way, the user can directly manipulate the "history" of their actions.

4.1.4 Design Study 1: Conclusion. During this study, we quickly worked through interaction and visualization approaches that are common in topic modelling. After informal user feedback and using the system ourselves, we started to develop interaction methods that focused on directly manipulating data provenance.

4.2 Design Study 2: Multi-week user study and Focus Group

Once our front and backend designs had stabilized, we released our first version for a multi-week study with representative target users. The premise was to simulate a research team working on the same research question within a provided dataset. We were interested in the following research questions:

- **E2.RQ1**. In what ways did participants understand and use features such as collaboration, search and Teleoscope ranking?
- **E2.RQ2**. How did participants incorporate Teleoscope into their understanding of qualitative research processes?

We were also interested to see the extent to which Teleoscope could hold up in a simulated production environment and welcomed ongoing bug reports. Therefore, the interface was developed to be robust enough for participants to use on their own devices, outside of a lab environment.

4.2.1 Participants. We recruited participants who had competence with qualitative methods: at least an upper-level undergraduate course and/or equivalent research experience. Eight participants were recruited; three dropped out (N=5 final count). Of participants who remained, one was a PI at a university who leads qualitative research in Nursing, one was a senior PhD student in Sociology, and two were upper-level undergraduates in Sociology attending a directed studies in qualitative methods course, and one was an upper-level undergraduate in Psychology. Participants were reimbursed at a rate of twice local minimum wage due to their status as expert users.

4.2.2 Methods. Participants were introduced to Teleoscope and each other during a one-hour training session where we brainstormed a shared research topic. Then, participants were instructed to use Teleoscope for at least 10 hours before the focus group, scheduled three weeks later. For each session that they used Teleoscope, they wrote in a diary, detailing (1) the *theme* that they explored; (2) the *process* by which they explored the theme; and (3) any *collaboration* features they used; and (4) bugs or features requests. No design changes were made during these weeks, but minor bugs were fixed. System logs were kept throughout this period.

Diary entries were analyzed using affinity diagramming before a focus group with participants, which took three hours including lunch. Focus group involved: (1) diary discussion; (2) brainstorm on problems encountered; (3) brainstorm on feature requests and design solutions to problems; and (4) explanation of ML concepts and a brainstorm on better alignment between visual and interaction metaphors. We video and audio recorded the focus group and used large printouts of the Teleoscope interface to draw and annotate problems and design ideas (see Fig 8).

Fig. 8. Large printouts of the Teleoscope interface were used to draw and annotate problems and design ideas with both the focus group (left), and the visualization group during a post-hoc analysis (right).

4.2.3 Post-hoc Analysis by Visualization Group. After we had analyzed and summarized participant results, we presented Teleoscope in two multi-hour analysis sessions with a Visualization research group. Our results reflect the analysis of that group along with our own analysis and solution brainstorming.

4.2.4 Diary Results. The topic that was chosen by the group of study participants for investigation was *Critical and end of life care* within the *r/AmItheAsshole* dataset. The group brainstormed starter keyword search ideas of *Medical Assistance in Dying (MAID)*, *End of life care*, *Palliative*, *ICU*, *Failures*, *Emergency rooms*, *Emergency care*, *Lack of beds*, *Overcrowding*.

Teleoscoping differs from keyword searches. Participants found that the Teleoscope ranking system differed from a normal keyword search. For many participants, it took some time to (1) differentiate results of a keyword search from results of a Teleoscope rank operation; and (2) differentiate valid results that did not meet their expectations due to the documents that existed in the corpus from invalid results due to bugs or mental model inconsistencies. There were negative transfer effects from being used to keyword searches which took multiple sessions to unlearn.

For example, P1 searched for "palliative" and was "*...surprised by how many posts were about animals at end of life, which does not fit our defined research topic.*" P1 then wished "*...there was a feature that would take everything I had*

already put within one group and give me 'more like this'", which was exactly what the Teleoscope rank operation was designed to do. Multiple participants corroborated this sentiment in their diaries (P5, P7, P8). This indicated a problem in the participants' mental model, likely due to (1) how we were representing the Teleoscope operation on the workspace; (2) our training, documentation and support materials; and (3) not enough time to learn the tool.

However, P1 reported for their third session that they spent a long time looking through documentation and support videos to understand the possibilities of Teleoscope:

> *Today I also spent time trying to go down the rabbit hole of different searches to try to gain a true appreciation for how this machine learning approach to data collection differs from just keyword searching within the Reddit search. This was really evident to me when I found a post where the OP had a palliative/terminal illness, and I wanted to find others where this was the case. I made a new folder for this category, then used the Teleoscope feature, and immediately found one other post where the OP has cancer and was asking a friend to not mourn their death. It would be extremely difficult to keyword search for this type of topic, but it's a very interesting and important area to capture (OPs with terminal illnesses). This was a great exploration! (P1)*

This indicated that it was possible to learn the difference between keyword search and Teleoscope rank, but that the learning curve was steep enough to require multiple hours of usage and documentation review.

Teleoscope can support quick, iterative workflows. P7 articulated a very clear document review strategy and seemed to understand the tool very quickly. Ignoring the group topic, they searched for documents related to their own research program by skimming titles:

> *I was interested in [AITA] posts about gay marriage, which is a topic tangentially related to my own research. I populated the [gay wedding] group with results whose titles caught my eye. I should note that I very rarely read the actual documents. If the title was vague, I occasionally skimmed the first few lines. (P7)*

They further suggested adding a document quick viewer to aid in skimming. They then organized documents into groups, relabelling and changing the groups as they developed their understanding of the corpus and the tool. Then, they switched between using the *teleoscope* operation and the group feature to find relevant documents:

> *After adding about half a dozed documents to the group 'gay stuff', I noticed that many of the documents are about gay panic. That is, the fear of being (wrongly or correctly) as gay, the dislike of anything perceived as gay, and an aversion from being around gay people. I changed the group name to 'gay panic' to reflect this...once I had about 13–14 results, I opened the Teleoscopes window for the group. Looking at the first two pages of results, none of them that weren't already in the group seemed very relevant, mostly judging by their titles and occasionally by the first couple of sentences in the document. I refined the search...[with a] couple of documents that I thought particularly demonstrated gay panic. (P7)*

Positive and negative transfer effects from other qualitative research software. Participants' prior extensive experience with qualitative research software allowed them to have a much more clear mental model of the tool without extensive training, which indicates the possibility of positive transfer effects. Many desired features were given as examples from tools that they had experience with, such as Google docs, MaxQDA, and NVivo. Unsurprisingly, they mostly expected Teleoscope to work like other interfaces they had previously used. To our surprise, none of the participants used any of the collaboration features.

P7 noticed many problems with the interface and made many suggestions for design changes that we brought to the focus group, including a lack of annotation and coding features, document export features, and overall corpus visualization features: "*As this point of my exploration of this theme, being able to play around with how [documents] connect to one another would I believe might have helped refine my thinking.*"

The above results were summarized and presented at the focus group, motivating our central discussion points.

4.2.5 System Log Results. During the user study, the system maintained a log of user actions as they interacted with the system. Across the 5 participants, 7 sessions with the *teleoscope* operations were tracked (one participant created 3 separate sessions for themselves). The mean number of actions tracked per session was approximately 310 (median = 286, minimum = 97, maximum = 684). Actions included such things as session initialization, creation/movement/deletion of windows, keyword searches, instantiation of *teleoscope* ranking and results, in short, any conditions where the state of the user workspace was altered.

The actions were logged and then visualized to better understand user interactions. Generally, users made use of an iterative process to find documents of interest, alternating between putting documents into groups and instantiating new *teleoscopes* to find new documents relevant to their queries and then sorting them into their groups. The number of *teleoscopes* created across the study were relatively small however (between 2 and 4).

4.2.6 Focus Group Results. The focus group provided insight into the needs of qualitative researchers with different levels of expertise with computer supported analysis. We used thematic analysis to review the results. We report here on the most prominent needs that emerged.

Mental model need: *teleoscope* state needs to be inspectable. In our tested design, *teleoscope* state was coupled directly to a single group. When the group was updated, the *teleoscope* rank changed. Individual documents within the group could be selected to weight the *teleoscope* search vector closer to that document. This confused participants about the state of the *teleoscope* rank. We decided that the *teleoscope* inputs and outputs needed a more explicit visual representation and decoupled from a single group, allowing for multiple input sources.

Mental model need: Use direct manipulation for all features, including workspace interactions. Participants expected more features to use direct manipulation such as drag-n-drop, infinite canvas, and organizing documents.

Mental model need: Clarify the Teleoscope exploration metaphors onscreen. For example, participants wondered whether they were being "*dropped off in the landscape*" of documents and "*going down different paths*" (P7). If so, they wanted a record of the paths and some way to compare paths directly onscreen. Participants agreed that "*seeing it*" and "*understanding how close documents are to each other in space [is important].*" (P1)

Feature need: Support set operations and filtering. Participants were most familiar and received direct training in keyword manipulation. Therefore, they were familiar with set operators and wanted to use them to gain confidence that they were being thorough enough in their search.

Conceptual need: Confidence in path saturation. Participants agreed that they did not need to show total path *exhaustion*, rather, they needed a sense of *saturation*. Their imagination of the use case for Teleoscope was for data collection, which did not mean finding every piece of relevant data but instead finding enough *representative* samples of data.

Conceptual need: Explaining the methodology to reviewers. Participants were concerned about how to explain the Teleoscope process to reviewers at a high level, but were not concerned with the details of the Teleoscope process. As long as they had a clear metric that they could point reviewers to (e.g., a paper that describes the metric), they did not particularly care about which metric to use. This was a defensive publication strategy: keyword searches are reported on directly to show paper saturation but there seems to be an understanding among current reviewers that the immense size of possible data sets meant that any reasonable metric could be argued for and used.

Conceptual need: differentiating data collection from analysis. Participants were unsure whether using Teleoscope constituted a violation of rigour such that it mixed data collection and analysis research phases. Particularly

if text-level analysis was to be supported, they felt that it might not be appropriate to allow for both in the interface at once. We found that participants used Teleoscope as we intended: they explored data in a manner that extended beyond keywords searches. For example, participants reported the following:

> *I was excited when the tool came up with things about the topic, but not including keywords that I used. (P1)*
> *I use it to find papers that I wouldn't have found. (P7)*

This allowed us to believe that Teleoscope could be used for a longer and more in-depth research project.

4.2.7 Post-Hoc Analysis and Recommendations from Visualization Group. We presented our results and current interface to a visualization group at our university. After two multi-hour analysis sessions, the visualization group recommended the following:

Visualize workspace relations. Up until this point, our interface did not include any visual graph concepts since our original abandoning of adjacency matrices. The recommendation was to clarify system state further by treating windows as nodes which could be used as input sources for operations.

Clarify data type representation. Since the results of workspace operations were effectively variations on ordered lists, the recommendation was to create visual homogeneity where data types were the same, and visual differentiation where they were different. From this, we developed our current paradigm of source, control, and target data types, and a unified output datatype of a list of document lists.

Create a quick-viewer to allow for minimized windows. A sidebar was recommended to allow documents to remain as minimized pills while users reminded themselves of document contents.

Allow for multiple workflows. Users wanted to simultaneously pursue multiple explorations with quickly accessible workflows. They further recommended that we create chainable workflows to target both provenance and reusability.

Our main takeaways from the visualization group recommendations was to move our window metaphor into a workflow metaphor with simplified workspace objects. This required a major redesign for both our front and backend systems to support graph operations. It also introduced questions of concurrency and graph directionality. For example, our first imagination of this design introduced cycles into the graph; as such, we made the decision to make sources strictly "left-handed" and targets "right handed" with the exception of set operations. This had further impacts on our state management system, which had to be redesigned to work with a graph-based structure.

4.2.8 Design Study 2: Conclusion. In this study, we investigated how real qualitative researchers would use Teleoscope in a simulated research environment. A summary of our findings corresponding to the research questions is:

- **E2.RQ1**. Participants understood the overall concept of the Teleoscope ranking operation and workspace as a whole, likening it to walking down different paths in an unknown region. Even though this metaphor came from the participants, the operation inputs and system state were unclear; therefore we decided that the path metaphor should be visualized as directly as possible. Participants did not use any collaboration features of Teleoscope, which may be due to a lack of impetus, our study environment, or because of our design. In the next study, we addressed these possibilities to ensure collaboration features were used.
- **E2.RQ2**. Participants interpreted Teleoscope primarily as a data collection tool. Further, they were unsure about whether it infringes on analysis and had concerns about academic rigour if it does.

4.3 Design Study 3: Multi-Month Field Deployment

After completing a redesign after Study 2, we were interested in a long-term focused field deployment with qualitative research groups who could bring their own research needs to us. We also wanted to move Teleoscope out of a simulated environment and into a production environment where real-world motivations and difficulties would be encountered.

Study 2 had many study contrivances, such as a research topic that none of the participants were using for their own research. We felt that it was important to see how Teleoscope would perform when participants were subjected to all the benefits, consequences and costs of real research. We also committed to ongoing feature development (and bug fixes) using a quasi co-design methodology.

Specifically, we were interested in the following research questions:

- **E3.RQ1**. When put in a real-world environment, how did researchers incorporate Teleoscope into their own research practice? Which features, processes, and workflows within Teleoscope were commonly used and in what way?
- **E3.RQ2**. Using Teleoscope, to what extent were researchers able to feel confident that they were able to retrieve data using criteria that are important to qualitative researchers, i.e., richness, saturation, etc.?

During this period, we also moved Teleoscope out of a test environment into a publicly-available production release. This involved adding many security features and backup systems which were exposed to the threat of arbitrary internet attacks. We had at least one successful security breach which was dealt with and mitigated; the number of unsuccessful attacks are unknown.

4.3.1 Participants. Three PIs and research groups were recruited; only one research group was able to commit to long-term use of Teleoscope. The final team was comprised the PI from Study 2 and a research team of two graduate RAs that were recruited specifically to use Teleoscope. Participants were not reimbursed for their time by us since they were being supported extensively by our design team for their own research project; our understanding is that RAs were compensated as normal by the PI.

4.3.2 Methods. Teleoscope was deployed in a standard beta release manner where participants were given a private link to Teleoscope until we transitioned to a full public release. Participants were invited to participate in a Discord server to make bug reports and feature requests. Depending on the request, occasional emails and video interviews were conducted. Logs were kept of system use to compare with study results. For the purposes of this paper, we finished data collection after six months, but use is ongoing by the research team.

4.3.3 Results. The research team used Teleoscope for data collection for their research project on nurses and structural inequality as articulated in Reddit's *r/nursing* forum, which is where working nurses post about their day-to-day problems. Of their own initiative and in alignment with their own qualitative methodological approach, the team's main conceptualization of data collection was via an external google document that contained a set of keywords to be used in searches (see Figure 9). After exploring the data using a variety of features in the Teleoscope workspace (described below), they would update their keyword list with successful and unsuccessful searches. Here are the ways in which the research team used the Teleoscope workspace features:

Teleoscope helped discover unknown terminology. Even though the research team was composed of people experienced in nursing culture, they were not always aware of the terms that were used by on-the-ground nurses.

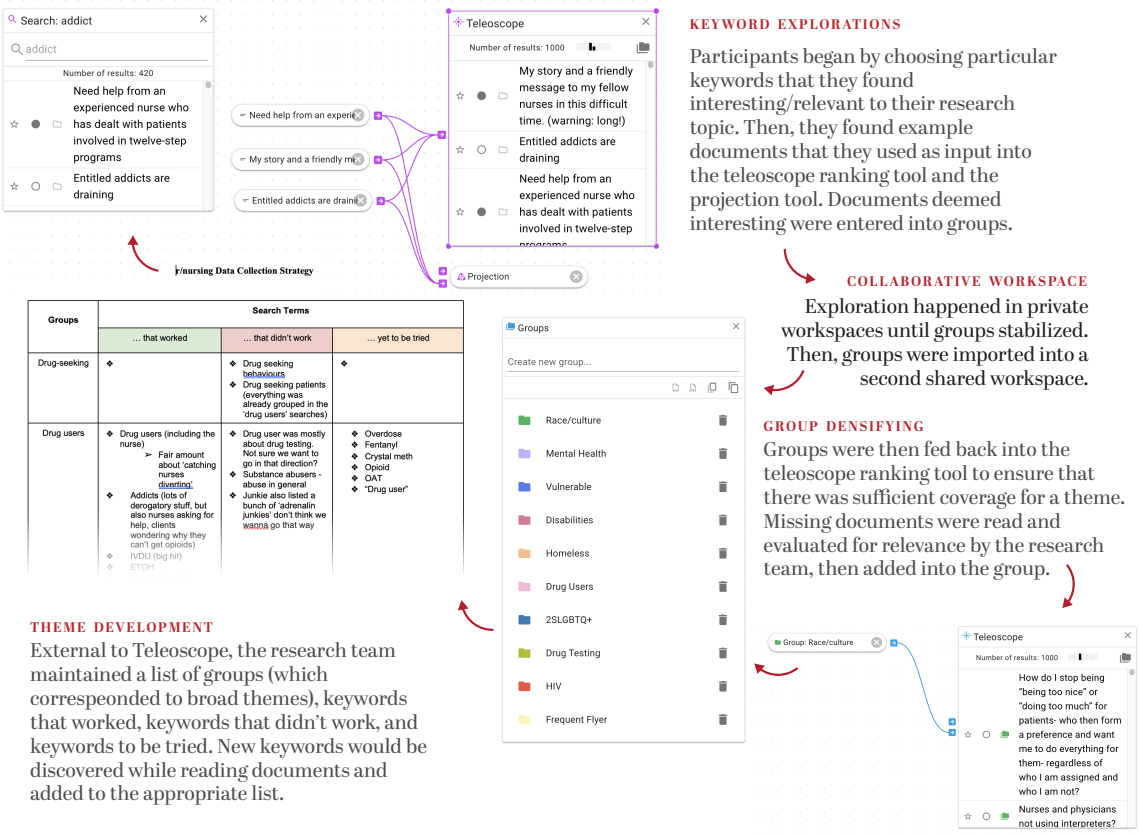

Fig. 9. An example of a workflow from *Design Study 3* our long-term deployment. Pictures are actual screenshots of research artifacts from our participant research team's data-gathering phase. Participants worked both individually and collaboratively on the Teleoscope interface, and collaboratively on Google Docs, Zoom, and in-person. Due to the existing qualitative research culture in their research group, keyword searches were the focus of their data collection approach. ML features were used to discover new keywords, find thematically similar documents that did not have specific keywords, and to saturate groups with relevant documents.

> *We search based on these weird, predefined words keywords that we think relate to structural inequities. But we also have to guess in advance what language people might be using...We didn't think that people would delicately [post using the term] 'people who experience structural inequities'...But we did try a lot of words we wouldn't use, you know, like addict, junkie...We were trying to use the system in a way that would get us further than those keywords alone. (P1)*

The Teleoscope system helped to populate their keyword search document with search terms that they would not have predicted a priori.

Teleoscope helped with search saturation. The researchers reported (1) making groups from documents from their keyword searches; (2) piping the groups into the Teleoscope ranking operation as controls; (3) determining which documents they had not yet read and (4) adding those documents to the groups. This helped to see the parts of the document space that they had not yet captured with a keyword search.

> *Putting each one of our single groups like indigenous, vulnerable disabilities, [an RA] put them into the teleoscope and then basically went through to see like which ones we hadn't read...We were just trying to expand our data set and be exhaustive. (P1)*

Working iteratively between keyword searches and ML functionality was important for exploring and structuring the research topic. The researchers reported that they used the system to iterate and structure their ideas about their research topic.

> *It was actually really helpful to start with keyword searches. And to be able to build out this groups structure, and then Teleoscope from there. Whereas maybe if we'd had a more like drilled in topic, we could have just gone from there.(P1)*

The researchers used Teleoscope to develop ad hoc themes out of their original categorical approach as their understandings of the target data grew:

> *There are a bunch of different like origin categories we needed to go off of because of the way our topic is. We couldn't even [try to search for categorical terms such as] 'Oh, this is about emergency departments...vulnerable [people] or inequities'...because when people are dragging on someone who uses drugs, who comes to the emergency department every week. That's not the words they're going to use. They're going to use super stigmatizing language, probably like what we've seen in a lot of cases and be like, 'I'm so frustrated. This junkie comes into work all the time. He's just drug seeking. He's like plugging up a bed for everyone else who needs it.' (P1)*

Projections have potential to allay methodological concerns. In the focus group, it was brought up as a concern that using Teleoscope might be too close to analysis and confuse methodological rigour (in rigorous qualitative research, data collection and analysis stages are kept distinct so as not to predetermine results). However, the projection operation, which was added after the focus group, seemed to have potential to populate the interface with unexpected results:

> *One of the things that I had really worried about methodologically was that with Teleoscope, I almost felt like you were kind of like deciding what your findings might be...I feel like [the projection operation] is really addressing some of that for me, because I feel like it's bringing you all these like adjacent topics to what you're looking for. And I feel like it really broadens your idea so much further in the potential data [since] you're exposed to so much more of the subreddit than you would be through just keyword searching. And I feel like that's really methodologically sound. (P1)*

The speed and ease of use helped with the sense of completion as well:

> *When it comes to big data sets, it takes way too long to get a sense of what's going on. [The projection operation] is really nice, quick [and has] potential to expose that kind of stuff so that your findings aren't [close to the] single keyword that you put out...With qualitative research that's a really interesting and powerful thing to be able to do. (P1)*

Multiple workflows/workspaces allowed for reproducible exploration and collaboration. The research team approached collaboration by exploring on their own in independent workspaces, discussing their results, and then collating the results into a single shared workspace. Outside collaboration tools were also used, such as email, Zoom, and Google Docs. The independent workspaces served as drafting areas, where individual researchers could explore many ideas without committing to the larger team's conception of the dataset, then came together with refined groups

and keywords. After they collated their results, they performed further data exploration as a team on a single workspace by systematically using our ML operations to ensure completeness.

One of Teleoscope's key design features is to create reproducible workflows that are able to be inspected by collaborators. The above method of exploring in separate workspaces and combining in a shared workspace was enabled by the guarantee of maintaining reproducible results.

Along with the above findings, researchers submitted a variety feature requests and bug reports. Set operations were the last feature to be developed after some discussion and redesign of our backend graph processing system.

4.3.4 Design Study 3: Conclusion. In this study, we performed a customized field deployment for a real qualitative research team and took ongoing bug reports and design feedback. The summary of our findings for our research questions are:

- **E3.RQ1**. Researchers incorporated Teleoscope into their research practice as a data collection tool, working between an external Google doc for keywords and the interface itself. Teleoscope was used to explore parts of the document space where keywords would not be easy or obvious to find. The most common features used were the keyword search, document reading and grouping. Teleoscopes were used after grouping. Projections were used after Teleoscoping and grouping.
- **E3.RQ2**. Teleoscope helped to provide confidence that a corpus was being more fully and rigorously explored by providing both ranked and randomized example documents.

4.4 Design Study 4: On-going Public Release

We are hosting an ongoing public beta release of Teleoscope (publicly deployed at [anonymized for review after consultation with CHI chairs]). This is not a formal user study; instead it is an ongoing test and demonstration of our system robustness in terms of performance, security, and availability. By committing to a live release on a cloud platform, we were forced to develop the following security and availability measures:

- **(Nearly) one-click deployment**. Teleoscope can be deployed on a new AWS virtual machine using Ansible playbooks in nearly one click. This was developed after a sprinkler accident destroyed our original non-public servers and motivated our move to the cloud. We needed to re-deploy Teleoscope often enough to spend time developing an automatic deployment system.
- **Robust backup system**. We expected a catastrophic security breach at some point and developed a backup system. When our database was indeed hacked and erased, we redoubled our backup system to two small-scale hourly offsite backups as well as a daily backup.
- **User roles/API limiting/Reverse proxy/SSL/TLS**. Earlier this year, Reddit restricted usage of the data API, which interfered with our data collection strategy. It also introduced the threat of large-scale data scraping from Teleoscope. As such, we restricted Teleoscope to registered users (open to anyone to register), creating a robust internal/external user role scheme for MongoDB and RabbitMQ, put in place data security measures such as limiting our API throughput, and set up a SSL/TLS reverse proxy to encrypt messages between client and server.

Teleoscope remains online as of the writing of this paper and continues to gain users.

5 DISCUSSION, LIMITATIONS AND FUTURE WORK

In this section, we discuss the results of our studies with regards to our design goals, research questions, and provide directions for future work.

Through our studies and design work, we settled on a process-focused graph-based workflow metaphor rather than a dashboard metaphor (**DG2**). This was due to our focus on the process of arranging and connecting documents as being the most important interaction focus (**DG1**, **DG3**); it also increased visibility of system state and allowed for direct inspection of computational elements. Researchers found that they were able to find confidence that they had searched "enough" of the document space through using the Teleoscope ranking operation, and found potential in the Projection clustering operation in supporting rigorous document space exploration (**E2.RQ2**, **E3.RQ2**). After we iterated on our design to allow for collaboration and make it obvious to our participants how to use collaboration features, they were able to satisfy their need for theoretical saturation through collaborative iterative searches using Teleoscope (**E2.RQ1**, **DG2**).

A limitation of a workflow metaphor is that it does not visually model the unknown document space very well. In part, this helps manage the cognitive overload since users are only shown what they have explicitly requested and built. However, having some dashboard elements such as summary statistics of corpus exploration relative to selected groups would be an obvious next design step. In our estimation, this would require a more sophisticated graph processing approach.

Due to our users' use of an external keyword Google doc, an open question is how and whether to incorporate keywords back into Teleoscope as a primary interaction element (**E3.RQ1**). Perhaps with more experience with Teleoscope, set operations would be used to manage keywords more directly. However, it may not be necessary to treat keywords as a primary interaction element within Teleoscope. We already imagine Teleoscope as part of a qualitative research tool ecosystem: current research shows that Teleoscope can work within existing analysis workflows between tools such as NVivo or PowerBI. We found it interesting and useful that researchers were able to creatively include Teleoscope in their research process using external tools that they were already familiar with, and to iterate between them and Teleoscope. For this reason, we support exporting to common document formats such as XLSX, DocX and JSON. However, having interface standards that make Teleoscope output interoperate with other qualitative software is important.

Since this is a research product, Teleoscope does not yet incorporate standard enterprise product features such as arbitrary data import. We imagine that an enterprise version of Teleoscope would need to support swapping/encoding with different exogenous models directly in the user interface, particularly as large language models become more sophisticated.

In our estimation, a large part of Teleoscope's success is demonstrated by its on-going public release and use by the research team that we recruited for field deployment in Study 3. We are continuing to recruit more research teams and hope to see Teleoscope develop into a more mature product as they use it.

Teleoscope's impact on methodology. To what extent is it desirable for an ML model to "think like you" when you are a qualitative researcher? One of the most interesting and relevant questions for Teleoscope is the way in which it can impact qualitative research methodologies. This was brought up by multiple participants, and is an important question for both adoption of Teleoscope and for shedding light on other burgeoning ML-supported methodologies that deal with data interpretation.

Current qualitative methodologies place a high level of importance on disciplined and ethical data collection, analysis, and reporting. Since interpretation necessarily comes from the perspective of the researcher, researchers attempt to maintain distance from analysis until data collection is concluded. Teleoscope could be seen as presupposing a thematic structure before detailed analysis has begun, which would violate the principle of letting the themes emerge from the data.

However, our participants articulated this mostly as fear of cognitive bias about the data, not that it fundamentally interfered with having an opinionated approach to data collection. It is impossible to do data collection from a truly unbiased perspective, since researchers still have editorial discretion in choosing their research interests, topics, and data sources. Our participants were satisfied with Teleoscope bringing up *unexpected* results along with *expected* results as a way to counteract bias. So, the answer to the question at the start of this section is that you might want the ML model to "think like you" when you are searching for expected results, and then to not "think like you" when you are challenging your expectation. Teleoscope can do both.

We believe that Teleoscope provides an opportunity to directly inspect and compare biases by inspecting and comparing the workflows that trace how a document collection came to be. Returning to the Nowell quote from Section 2, we believe that Teleoscope can demonstrate that data *collection* has been done in a precise, consistent, and exhaustive manner through *tracing the data collection process* with enough detail to enable the *research team* to determine whether the process is credible (**DG1**, **DG2**).

6 CONCLUSION

In this paper, we presented Teleoscope, a web-based system that supports interactive exploration of large corpora (100K-1M) of short documents. We developed it in response to the need of qualitative researchers to explore large corpora in meaning-based ways using natural interaction techniques. Targeting qualitative "themes" rather than "topics," we attempted to create a system that makes an ML model "think like you" without significant retraining. Teleoscope provides ML-based workflows that have semantic and computational meaning. These workflows help researchers to retrace, share, and recompute their sense-making process. We reported on the design, engineering, evaluation, and deployment of our system. Our public deployment of Teleoscope is ongoing. We plan to continue improving on Teleoscope and to maintain it for use by the broad community of qualitative researchers.

REFERENCES

[1] [n. d.]. https://openai.com/blog/chatgpt

[2] Eric Alexander, Joe Kohlmann, Robin Valenza, Michael Witmore, and Michael Gleicher. 2014. Serendip: Topic model-driven visual exploration of text corpora. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*. 173–182. https://doi.org/10.1109/VAST.2014.7042493

[3] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* 58 (2020), 82–115.

[4] Deepak Suresh Asudani, Naresh Kumar Nagwani, and Pradeep Singh. 2023. Impact of word embedding models on text analytics in deep learning environment: a review. *Artificial Intelligence Review* 56 (2023), 1–81.

[5] Rajiv Badi, Soonil Bae, J Michael Moore, Konstantinos Meintanis, Anna Zacchi, Haowei Hsieh, Frank Shipman, and Catherine C Marshall. 2006. Recognizing user interest and document value from reading and organizing activities in document triage. In *Proceedings of the 11th international conference on Intelligent user interfaces*. 218–225.

[6] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, Vol. 14. 830–839.

[7] Charles Berret and Tamara Munzner. 2022. Iceberg Sensemaking: A Process Model for Critical Data Analysis and Visualization. *arxiv.org* (4 2022).

[8] Christian Bors, Theresia Gschwandtner, and Silvia Miksch. 2019. Capturing and visualizing provenance from data wrangling. *IEEE computer graphics and applications* 39, 6 (2019), 61–75.

[9] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association.

[10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[11] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175* (2018).

[12] Nan-Chen Chen, Margaret Drouhard, Rafal Kocielnik, Jina Suh, and Cecilia R Aragon. 2018. Using Machine Learning to Support Qualitative Coding in Social Science. *ACM Transactions on Interactive Intelligent Systems* 8 (6 2018), 1–20. Issue 2. https://doi.org/10.1145/3185515

[13] Jaegul Choo, Changhyun Lee, Chandan K Reddy, and Haesun Park. 2013. Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. *IEEE transactions on visualization and computer graphics* 19, 12 (2013), 1992–2001.

[14] Zach Cutler, Kiran Gadhave, and Alexander Lex. 2020. Trrack: A Library for Provenance-Tracking in Web-Based Visualizations, In IEEE Visualization Conference (VIS). 116–120. https://doi.org/10.1109/VIS47514.2020.00030

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL]

[17] Mennatallah El-Assady, Rebecca Kehlbeck, Christopher Collins, Daniel Keim, and Oliver Deussen. 2019. Semantic concept spaces: Guided topic model refinement using word-embedding projections. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 1001–1011.

[18] Anna Fariha and Alexandra Meliou. 2019. Example-driven query intent discovery: Abductive reasoning using semantic similarity. *arXiv preprint arXiv:1906.10322* (2019).

[19] Samah Gad, Waqas Javed, Sohaib Ghani, Niklas Elmqvist, Tom Ewing, Keith N Hampton, and Naren Ramakrishnan. 2015. ThemeDelta: Dynamic segmentations over temporal topic models. *IEEE transactions on visualization and computer graphics* 21, 5 (2015), 672–685.

[20] Marti A Hearst and Duane Degler. 2013. Sewing the seams of sensemaking: A practical interface for tagging and organizing saved search results. In *Proceedings of the symposium on human-computer interaction and information retrieval*. 1–10.

[21] Monique M Hennink, Bonnie N Kaiser, and Vincent C Marconi. 2017. Code saturation versus meaning saturation: how many interviews are enough? *Qualitative health research* 27, 4 (2017), 591–608.

[22] Matt-Heun Hong, Lauren A Marsh, Jessica L Feuston, Janet Ruppert, Jed R Brubaker, and Danielle Albers Szafir. 2022. Scholastic: Graphical Human-AI Collaboration for Inductive and Interpretive Text Analysis. *The 35th Annual ACM Symposium on User Interface Software and Technology*. https://doi.org/10.1145/3526113.3545681

[23] Hannah Kim, Dongjin Choi, Barry Drake, Alex Endert, and Haesun Park. 2019. TopicSifter: Interactive search space reduction through targeted topic modeling. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, IEEE, Vancouver, Canada, 35–45.

[24] Hannah Kim, Barry Drake, Alex Endert, and Haesun Park. 2020. Architext: Interactive hierarchical topic modeling. *IEEE transactions on visualization and computer graphics* 27, 9 (2020), 3644–3655.

[25] Kori A LaDonna, Anthony R Artino Jr, and Dorene F Balmer. 2021. Beyond the guise of saturation: rigor and qualitative interview data. , 607–611 pages.

[26] Ching-Hung Lee, Chien-Liang Liu, Amy JC Trappey, John PT Mo, and Kevin C Desouza. 2021. Understanding digital transformation in advanced manufacturing and engineering: A bibliometric analysis, topic modeling and research trend discovery. *Advanced Engineering Informatics* 50 (2021), 101428.

[27] Yuan Li, Anita Crescenzi, Austin R Ward, and Rob Capra. 2023. Thinking inside the box: An evaluation of a novel search-assisting tool for supporting (meta) cognition during exploratory search. *Journal of the Association for Information Science and Technology* (2023).

[28] Matteo Lissandrini, Davide Mottin, Themis Palpanas, Yannis Velegrakis, and HV Jagadish. 2019. *Data Exploration Using Example-Based Methods*. Springer.

[29] Denis Mayr Lima Martins. 2019. Reverse engineering database queries from examples: State-of-the-art, challenges, and research opportunities. *Information Systems* 83 (2019), 89–100.

[30] Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.* 2, 11 (2017), 205.

[31] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. 2018. UMAP: Uniform Manifold Approximation and Projection. *The Journal of Open Source Software* 3, 29 (2018), 861.

[32] Christofer Meinecke, David Joseph Wrisley, and Stefan Jänicke. 2021. Explaining semi-supervised text alignment through visualization. *IEEE Transactions on Visualization and Computer Graphics* 28, 12 (2021), 4797–4809.

[33] Albine Moser and Irene Korstjens. 2017. Series: Practical guidance to qualitative research. Part 1: Introduction. *European Journal of General Practice* 23 (10 2017), 271–273. Issue 1. https://doi.org/10.1080/13814788.2017.1375093

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

[35] Jakob Neilson. [n. d.]. 10 usability heuristics for user interface design. https://www.nngroup.com/articles/ten-usability-heuristics/

[36] Jakob Nielsen. 1992. Finding Usability Problems through Heuristic Evaluation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 373–380. https://doi.org/10.1145/142750.142834

[37] Jakob Nielsen and Rolf Molich. 1990. Heuristic Evaluation of User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 249–256. https://doi.org/10.1145/97243.97281

[38] Sergey I Nikolenko, Sergei Koltcov, and Olessia Koltsova. 2017. Topic modelling for qualitative studies. *Journal of Information Science* 43, 1 (2017), 88–102.

[39] Lorelli S Nowell, Jill M Norris, Deborah E White, and Nancy J Moules. 2017. Thematic analysis: Striving to meet the trustworthiness criteria. *International journal of qualitative methods* 16, 1 (2017), 1609406917733847.

[40] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[41] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12

(2011), 2825–2830.

[42] Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic* 3, 2 (2011).

[43] Tim Rietz and Alexander Maedche. 2021. Cody: An AI-Based System to Semi-Automate Coding for Qualitative Research. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3411764.3445591

[44] Matthias Rüdiger, David Antons, Amol M Joshi, and Torsten-Oliver Salge. 2022. Topic modeling revisited: New evidence on algorithm performance and quality metrics. *Plos one* 17, 4 (2022), e0266325.

[45] Favourate Y Sebele-Mpofu. 2020. Saturation controversy in qualitative research: Complexities and underlying assumptions. A literature review. *Cogent Social Sciences* 6, 1 (2020), 1838706.

[46] Claudio T Silva, Juliana Freire, and Steven P Callahan. 2007. Provenance for visualizations: Reproducibility and beyond. *Computing in Science & Engineering* 9, 5 (2007), 82–89.

[47] Fabian Sperrle, Mennatallah El-Assady, Grace Guo, Rita Borgo, D Horng Chau, Alex Endert, and Daniel Keim. 2021. A Survey of Human-Centered Evaluations in Human-Centered Machine Learning. In *Computer Graphics Forum*, Vol. 40.3. Wiley Online Library, 543–568.

[48] Silvia Terragni, Elisabetta Fersini, Bruno Giovanni Galuzzi, Pietro Tropeano, and Antonio Candelieri. 2021. OCTIS: Comparing and Optimizing Topic models is Simple!. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Online, 263–270. https://doi.org/10.18653/v1/2021.eacl-demos.31

[49] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[50] Kai Xu, Alvitta Ottley, Conny Walchshofer, Marc Streit, Remco Chang, and John Wenskovitch. 2020. Survey on the analysis of user interactions and visualization provenance. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, 757–783.

[51] Jun Yuan, Changjian Chen, Weikai Yang, Mengchen Liu, Jiazhi Xia, and Shixia Liu. 2021. A survey of visual analytics techniques for machine learning. *Computational Visual Media* 7 (2021), 3–36.