# A Critical Analysis of the Usage of Dimensionality Reduction in Four Domains

Dylan Cashman, Mark Keller, Hyeon Jeon, Bum Chul Kwon, Qianwen Wang

**Abstract**—Dimensionality reduction is used as an important tool for unraveling the complexities of high-dimensional datasets in many fields of science, such as cell biology, chemical informatics, and physics. Visualizations of the dimensionally-reduced data enable scientists to delve into the intrinsic structures of their datasets and align them with established hypotheses. Visualization researchers have thus proposed many dimensionality reduction methods and interactive systems designed to uncover latent structures. At the same time, different scientific domains have formulated guidelines or common workflows for using dimensionality reduction techniques and visualizations for their respective fields. In this work, we present a critical analysis of the usage of dimensionality reduction in scientific domains outside of computer science. First, we conduct a bibliometric analysis of 21,249 academic publications that use dimensionality reduction to observe differences in the frequency of techniques across fields. Next, we conduct a survey of a 71-paper sample from four fields: biology, chemistry, physics, and business. Through this survey, we uncover common workflows, processes, and usage patterns, including the mixed use of confirmatory data analysis to validate a dataset and projection method and exploratory data analysis to then generate more hypotheses. We also find that misinterpretations and inappropriate usage is common, particularly in the visual interpretation of the resulting dimensionally reduced view. Lastly, we compare our observations with recent works in the visualization community in order to match work within our community to potential areas of impact outside our community. By comparing the usage found within scientific fields to the recent research output of the visualization community, we offer both validation of the progress of visualization research into dimensionality reduction and a call for action to produce techniques that meet the needs of scientific users.

**Index Terms**—Dimensionality Reduction, Projection, Visualization, Critical Analysis

## 1 Introduction

Recent years have witnessed the ubiquitous application of dimensionality reduction (DR) techniques across various domains. By converting hundreds or thousands of dimensions into just two, DR enables intuitive visualization of high-dimensional data, which greatly aids exploratory data analysis (EDA), especially in noise filtering and pattern identification. Consequently, DR has evolved into an essential component of many data analysis workflows [1], [2]. One of the most popular DR methods, t-SNE [3], has been cited more than 45,000 times as of June 2024, according to Google Scholar. In addition, DR is a core component in many visual analytics systems [4]–[6].

At the same time, concerns regarding the use and interpretation of these techniques have emerged [7]–[10]. First, distortion and information loss are inevitably induced when applying DR can easily lead to inaccurate interpretation of the high-dimensional data [2], [8], [11]. Second, inexperienced users often interpret projections in a manner that violates their original design intention. For example, one common mistake is the assumption that the distance between two clusters in the projected 2D manifold directly reflects the similarity between them [12], [13]. There is ongoing debate regarding the usage and interpretation of dimensionality reduction in data analysis.

It is important for visualization researchers to consider if the visualization of DR results is meeting the needs of the scientific community at large. While there are tools and preprocessing pipelines recommended within domains, they may be designed explicitly for particular types and scales of data typical in those domains and thus not be more broadly applicable. Despite the ubiquitous use of DR, there does not appear to be a standard workflow for interpretation and use across domains. Different users and fields approach the analysis in unique ways. There may be a gap in the needs of the scientific community and the current capabilities of DR algorithms and visualization systems that integrate them.

To address these challenges, we present a critical analysis of the usage of DR for high dimensional data analysis *outside of computer science*. We seek to answer the following research questions.

RQ1 **What are the usage patterns of DR outside of computer science?** These usage patterns may differ by what particular techniques they are using (t-distributed Stochastic Neighbor Embedding i.e. t-SNE, Principal Component Analysis i.e. PCA, or Uniform Manifold Approximation and Projection i.e. UMAP) or by what characterizes the data that is used. Answering this question informs the algorithmic design of visualization systems that allow for the exploration of high-dimensional data.

RQ2 **How are views of dimensionally-reduced data interpreted?** The methods of interpretation may also


- *Dylan Cashman is with Brandeis University. E-mail: dylancashman@brandeis.edu*
- *Mark Keller is with Harvard Medical School. E-mail: mark_keller@hms.harvard.edu*
- *Hyeon Jeon is with Seoul National University. E-mail: hj@hcil.snu.ac.kr*
- *Bum Chul Kwon is with IBM Research. E-mail: bumchul.kwon@us.ibm.com*
- *Qianwen Wang is with University of Minnesota. E-mail: qianwen@umn.edu*

depend on the analytic goals and what tasks are used to reach those analytic goals.

RQ3 **Do the usage patterns and interpretation methods differ by field?** The variance in usage and interpretation across domains can provide insight into the generalizability and extensibility required by DR algorithms and tools.

RQ4 **Are there any gaps in the available DR algorithms or tools that are opportunities for visualization researchers to adapt existing tools or to develop new tools?** There may be low-hanging fruit to apply the learnings from design studies in visualization research to meet unmet needs within domain use cases that haven't been addressed adequately by our community.

To answer these research questions, we follow a three-step analysis of scientific literature. First, we conduct a bibliometric analysis of 21,249 academic publications within scientific domains that use DR (Section 3). The analysis determined broad trends and practices (RQ1) and how those practices differ between communities at a high level (RQ3). Next, we present a survey of 71 papers from four scientific fields (Biology, Chemistry, Physics, and Business) to uncover lower-level differences between usage patterns (RQ1) and the interpretations of dimensionally-reduced views of data offered within scientific publications (RQ2) (Section 4). Full results of our survey are presented in a table in the appendix as well as an online browser of screenshots and metadata available at https://dimension-reduction-vis.github.io/.

We summarize the results from these two steps of the investigation in a comparative analysis of the findings from all fields in Section 5. As part of our analysis of these results, we review concurrent literature within the visualization community, including STARs, surveys, and reviews. These works typically review the usage of DR within visualization systems, and we conclude that there is a mismatch between the tasks identified in visualization systems vs. those found in our sample of domain science papers. Through this comparison, we uncover gaps between the visualization needs in domain scientists and the solutions offered by visualization researchers. We describe how these gaps lead to opportunities for visualization researchers (RQ4).

We find that simple linear projections like PCA are much more frequently used than nonlinear techniques like t-SNE and UMAP, even though nonlinear techniques are more frequently found in visual analytics systems [14]. We also find that dimensionality reduction visualizations are frequently used for *both* confirmatory and exploratory data analysis, in ways that may lead to bias or spurious interpretations. We describe three common workflows across the spectrum of confirmatory to exploratory data analysis. We outline the different ways that the visualization of the dimensionally reduced data is interpreted across four fields. In section 6, we provide takeaways to domain scientists and visualization researchers, including open questions for how visualization researchers can provide simpler-to-use projection techniques that provide insights on the underlying data while mitigating bias.

## 2 Related Work

In this section, we review related surveys, state-of-the-art reports, and other publications that investigate the usage of DR within visualization, our four selected subject areas (biology, chemistry, physics, and business), and science at large. Our goal in this section is to situate our work against existing studies so that a reader might know first what novel findings are presented in this work, and second where to look for alternative viewpoints on the topic.

We emphasize that our survey is the *first* survey from a visualization researcher's viewpoint looking outwardly at the usage of DR in domain sciences. Our survey thus differentiates from existing surveys that look inwardly at the usage of DR within visualization research [1], [2] and surveys of domain scientists looking inwardly at the usage of DR within their own domains [15]–[18]. Here, we review both to contrast their methods from this critical analysis.

### 2.1 Surveys on DR from Visualization Researchers

Visualizing and interacting with DR methods have become important topics in the visualization community, sparking visualization researchers to conduct various surveys. We categorize previous literature into the target of their analysis.

**Survey on DR Techniques** The most common type of DR-related surveys is, not surprisingly, the ones about DR techniques. These surveys aim to clarify the advantages and disadvantages of DR techniques, thereby supporting practitioners in selecting proper DR techniques in their analysis [2], [14], [19], [20]. For example, Espadoto et al. [14] surveyed 44 DR techniques and quantitatively examined their performance using five quality metrics. Nonato and Aupetit [2] also compared 28 different techniques, providing guidelines to select DR techniques by the analytic task. Etemadpour et al. [20], Xia et al. [21], and Sedlmair et al. [22] also provide similar guidelines, where empirical user studies ground their guides.

**Survey on Tasks** This family of surveys taxonomizes DR-related analytic tasks. They thereby aim to gain a further understanding of how practitioners use and interact with DR projections. For example, Sacha et al. surveyed visualization papers that included interaction techniques with dimensionality reduction algorithms, revealing the procedure in which analysts interact with DR [1]. Nonato and Aupetit [2] systematically survey and taxonomize the type of analytic tasks. It is not a literature survey, but Brehmer et al. [23] also revealed task sequence using DR projections for HD data analysis by conducting interview studies with analysts.

**Survey on DR Quality metrics** Quality metrics for DR assess the extent to which DR projections suffer from distortions [24]. As different quality metrics focus on different structural characteristics (e.g., local neighborhood structure [25] or cluster structure [11]), selecting appropriate DR quality metrics that match target analytic tasks is important for reliable data analysis. Surveys regarding quality metrics, therefore, aim to guide analysts in choosing appropriate metrics. Thurn et al. [26] and Bertini et al. [27], for example, organize DR quality metrics regarding which structural characteristics they focus on. Lee and Verleysen [28] share a

similar goal but concentrate on neighborhood preservation-based methods.

Our work differs from these works by focusing on the usage patterns and interpretations of DR methods across diverse fields of science, rather than limiting our scope to the visualization community. While previous surveys potentially oversample from design studies driven by visualization researchers or collaborations including them, we aim to investigate whether the usage patterns (RQ1) and interpretations (RQ2) of DR in the wild differ from those familiar to visualization experts. Unlike prior analyses that compare and contrast various visualization works within the same field, we compare and contrast the application of DR methods across diverse scientific disciplines (RQ3). By conducting extrinsic observations, our study provides insights that have the potential to guide the development of more generally applicable tools (RQ4), especially for managing biases and distortions inherent in DR techniques.

### 2.2 Surveys from Subject Area Researchers

Within individual subject areas, surveys, meta-analyses, or guidelines, papers act as an implicit or explicit reference on how to use various types of algorithms as part of the analysis of that subject area's data. Within biology, there exist several works comparing the usage of different dimensionality reduction algorithms on various types of biological data [15]–[17], [29], [30]. Similar works can be found in physics and astronomy [31], epidemiology [32], and chemistry [18]. These works provide some review of practices within the silo of a single field, and present some recommendations on the usage of DR within a particular high-dimensional data analysis pipeline. In contrast, our work looks both within and across domains. We also present our findings from a computer science perspective, interrogating the gap between the usage of dimensionality reduction in visualization research and those being used out in the wild.

## 3 Bibliometric Analysis Across Subject Areas

In this section, we use a bibliometric analysis of all recent papers citing the publications that introduce DR methods.

### 3.1 Bibliometric Analysis Objectives and Design

This analysis aims to identify broad trends and practices across domains outside of computer science (RQ1) and how those practices differ between communities at a high level (RQ3).

First, we identify 78 publications initially introducing DR methods ("DR papers"), based on Table 1 of a recent survey of DR methods by Espadoto et al. [14]. We exclude from our analysis general machine learning methods such as "neural networks" which were considered in the original table from Espadoto et al. [14] but are used more generally for other methods than DR. We also exclude methods for which a single originating publication to reference could not be identified. We acknowledge that this may bias our analysis, but we believe that the 78 papers provide a broad enough sample to provide insights. Using the Semantic Scholar Academic Graph [33], we query for all academic papers

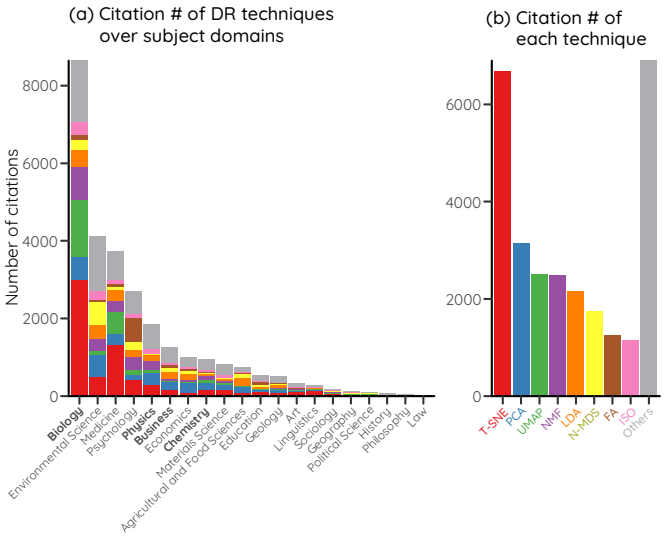



Fig. 1. Results of our bibliometric analysis on citation counts. (a) The number of citations of DR techniques over each subject domain. (b) The number of citations in which each DR technique was obtained.

(excluding preprints) published since 2013 that cite those publications. We use the Semantic Scholar classification of subject area, which is calculated using a machine learning method (see Appendix for more details). We exclude citing papers in Computer Science, Mathematics, and Engineering from our analysis based on the subject areas on record in Semantic Scholar, as we are interested in understanding how DR methods are used in other subject areas.

We present results in the form of counts and proportions. However, proportions are influenced by the introduction of new DR methods: the proportion of t-SNE usage might drop as a result of UMAP being introduced and used *in addition* to t-SNE). This could result in an observed decrease in the proportional usage of t-SNE, even if its use consistently increased over the period of analysis. As a result, we additionally compute percentile rankings of papers using the CP-EX method described by Bornmann and Williams [34], using the entire Semantic Scholar corpus to build cumulative percentages of papers with each citation count value in each subject area and year.

Acknowledging that citation counts are influenced by field size, we also consider metrics that are standardized to enable comparison across subject areas. In addition, we stratify citation counts by cited method and subject area and convert them to rankings.

### 3.2 Bibliometric Analysis Results and Discussion

We first sought to quantify the usage of the 78 DR methods within each subject area based on direct citations. We find that there are 136,956 unique papers published between 2013 and 2023 that cite at least one of the DR papers in the dataset. Of those 136,956 papers, 74,720 have been mapped to at least one subject area via one of the two subject area mapping methods. 74,357 papers (99.5% of those with at least one subject area) have been mapped to at most three subject areas using either method. Using the latter subject area assignments, there are 21,249 citing papers assigned to at least one of 20 subject areas other than Computer Science,

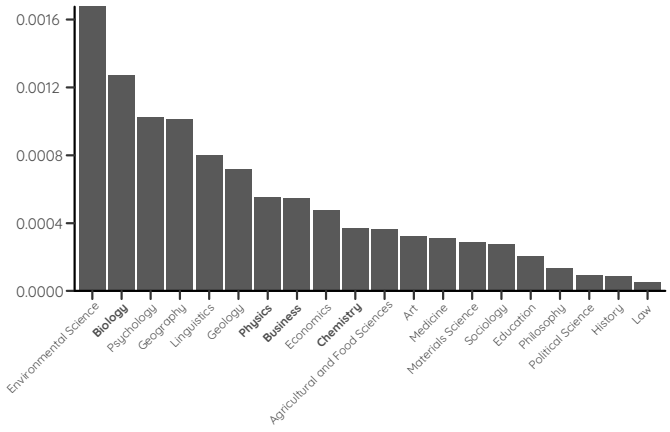



Fig. 2. Results of our bibliometric analysis. The Y axis encodes the fraction of the number of papers in each subject area that cite a DR paper over the total number of published papers in the subject area.

Mathematics, and Engineering. We use these 21,249 papers and subject areas for the results that follow in this section. We will highlight the four scientific domains that we focus on in our subsequent literature review to call attention to the variance in the domains that we have chosen.

Through summation of citing paper counts within subject areas, we observe that the top ten areas citing the 78 DR papers are Biology (first), Environmental Science, Medicine, Psychology, Physics, Business, Economics, Chemistry, Materials Science, and Agricultural and Food Sciences (tenth) (Figure 1a). Instead of summing across subject areas, we find that t-SNE, PCA, and UMAP are the three most highly cited methods (Figure 1b).

Considering the number of citing papers as a proportion of all papers in each subject area, we instead see that the top ten areas are Environmental Science (one in 596 papers cites a DR method), Biology, Psychology, Geography, Linguistics, Geology, Physics, Business, Economics, and Chemistry (one in 2,722 papers cites a DR method) (Figure 2). When we stratify citation counts by cited method and subject area and convert them to rankings, it becomes apparent that while t-SNE and PCA appear in the top three methods for 18 and 16 of the 20 subject areas, respectively, UMAP is only within the top three methods for Biology and Medicine (Figure 3). Other notable findings are that N-MDS is the top-ranked method in Environmental Science and Geography, FA is the top-ranked method in Psychology and Education, and LDA is the top-ranked method in Agricultural and Food Sciences (Figure 3).

For ease of interpretation and visualization, we next consolidate lesser-cited DR methods by mapping those that do not appear in the top three methods cited within any subject area to the category 'Other'. This results in nine DR method categories: t-SNE, PCA, UMAP, Nonnegative Matrix Factorization (NMF), Linear Discriminant Analysis (LDA), Non-metric Multidimensional Scaling (N-MDS), Factor Analysis (FA), Locally Linear Embedding (LLE), and 'Other'. Using these nine top-used DR method categories, we next considered the percent usage of each DR method within each subject area. We find that the percentage of citations mapped to the 'Other' method category is largest (compared to the eight remaining categories) in the subject areas Environmental Science, Physics, Business, Economics, Chemistry, Materials Science, Education, Geology, History, and Philosophy, which suggests that there is not a single overwhelmingly dominant DR method used in these fields. Together, the top three DR methods per field are cited more than the bottom 75 methods in Biology, Medicine, Psychology, Agricultural and Food Sciences, Linguistics, Sociology, Geography, and Law. This may suggest that in these subject areas, only a small set of DR methods are established or accepted or that data characteristics in such areas are amenable to certain DR methods.

We next consider how the 78 DR methods have been cited over time. We use percentile rankings of citation counts stratified by subject area and year to enable temporal comparison. Focusing on t-SNE, PCA, and UMAP in Biology, Physics, Business, and Chemistry, we find that percentile rankings for t-SNE, PCA, and UMAP increased between 2013 and 2022 in Biology, Physics, and Chemistry. In Business, percentile rankings for t-SNE and UMAP increased while those for PCA stayed consistent (Figure 4). These results are consistent with the publication of t-SNE in 2008 and UMAP being preprinted in 2018. The general trend between 2013 and 2022 across subject areas is that percentile rankings of DR methods have increased over time.

In summary, we find that while the use of dimensionality reduction is prevalent and generally rising across most fields (RQ1), the techniques used vary greatly by field (RQ3). In fields relating to biology and medicine, t-SNE and UMAP are popular techniques, while in fields relating to physics, chemistry, or business, PCA is popular. While there are other techniques used, t-SNE, PCA, and UMAP are the three most cited methods in recent years.

## 4 Subject Area Survey

The bibliometric analysis provided some insights into the broad trends and practices across domains outside of computer science (RQ1) but did not provide insight into the types of data used or nuances about the usage of particular techniques. Likewise, such a high-level analysis could not analyze the visual interpretations offered (RQ2). As a subsequent study, we conduct a close reading of a survey of literature in research outside of computer science.

### 4.1 Grounded Analysis

We began with a grounded analysis [35], [36] to consider a theory for analyzing and comparing works based on their DR. Our goal with this initial investigation was to ground the analysis of our subsequent literature review, bootstrapping the dimensions on which the usage and interpretation of dimensionality reduction techniques were spread among different scientific disciplines.

For this grounded analysis, we loosely gathered papers from varied domains using Google Scholar and Scopus. In both search engines, we searched for articles listing the different dimensionality reduction techniques in their title, abstract, or keywords and with subject areas of Arts and Humanities, Economics, Econometrics and Finance, Nursing, Physics and Astronomy to get a varied collection

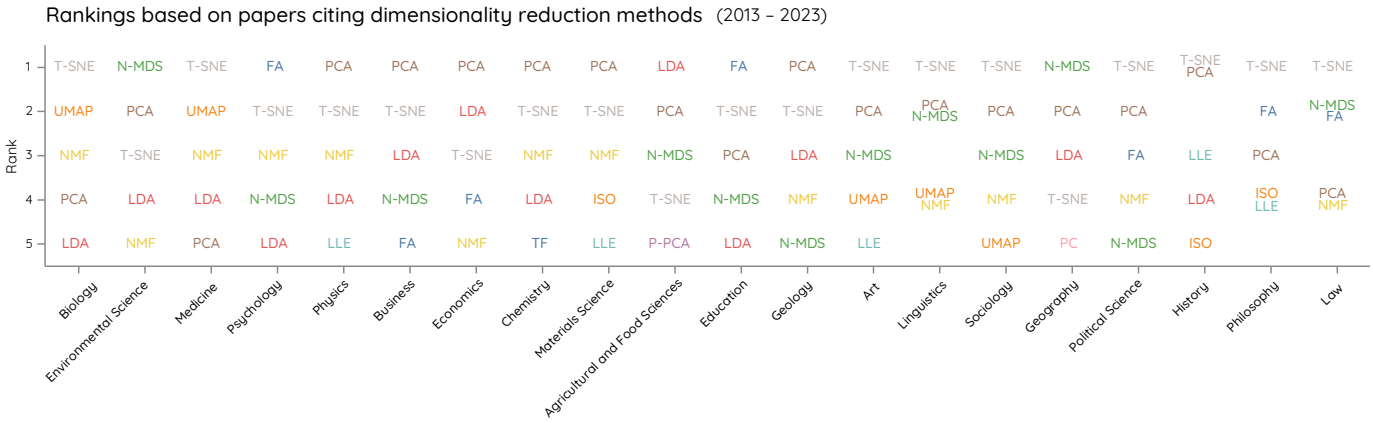



Fig. 3. Rankings based on papers citing dimensionality reduction methods (2013-2023). Methods are ranked based on the number of citing papers in each subject area. Subject areas are ordered left-to-right based on descending total count of papers citing the 78 considered dimensionality reduction methods. Ties are displayed in alphabetical order. Full method names can be found in the appendix.

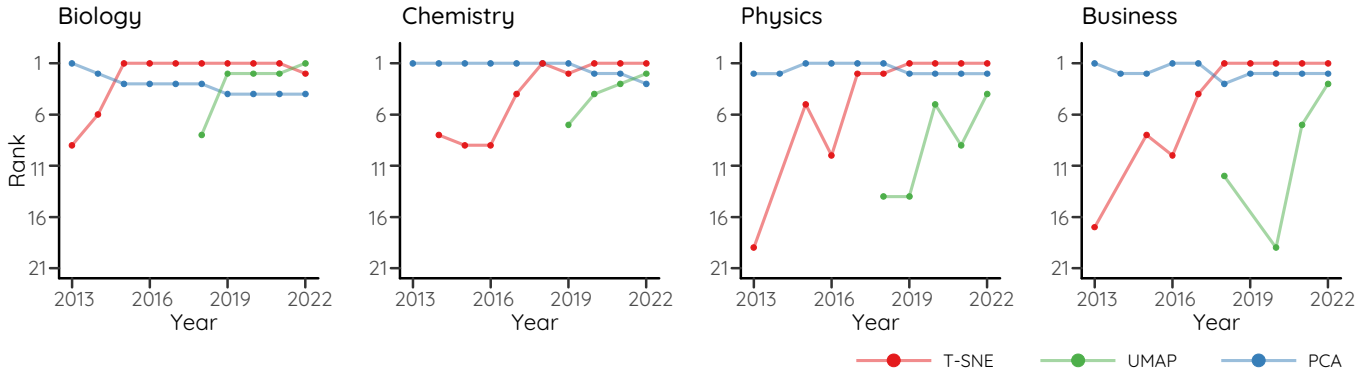



Fig. 4. Percentage rankings of three most widely used DR techniques (T-SNE, UMAP, and PCA) across years (2013 – 2022) in four domains we focused on. While PCA's percentage ranking stayed still or slightly decreased, the ranking of UMAP and t-SNE drastically increased through the past decade.

of papers. Between these two searches, we found articles in single-cell genomics [15]–[17], [29], [30], [37]–[39], business/finance [40], [41], urban studies [42], physics [31], [43], and epidemiology [32].

In our reading of these papers, we found that the usage of dimensionality reduction techniques varied in the analytical goal, the techniques used, the preprocessing and parameterization of those techniques, and the interpretation of the results of those techniques. We also found some similarities and differences between the scientific subject areas. For example, many single-cell genomics papers used a 2-D scatter plot of dimensionally reduced data as one step in a larger analytics pipeline, only gathering hypotheses to later test, while other domains drew conclusions directly from the reduced space. Similarly, the types of data, including the number of rows and columns, seemed to naturally lead to different types of interpretations. The design decisions of the visualizations used also seemed to vary, as some papers would rely on many annotations and highlights to explain how the dimensionally reduced data should be interpreted, while others provided only a few sentences in the main text. In particular, we found that the usage of DR differed primarily in the **data** shape and DR algorithm being used on that data, the **design** of the visualization of the dimensionally-reduced data, the low-level **tasks** used in the interpretation of that visualization, and the larger **workflows** that the DR was a part of.

### 4.2 Selection of papers

Drawing from our bibliometric analysis, we first decide on four subject areas to focus on based on their diversity in DR usage: Biology, Chemistry, Physics, and Business. In particular, we found that Biology was the field that cited the top three methods the most and featured a large body of research, while Chemistry, Physics, and Business were fields that featured usage of alternative DR algorithms. In addition, several of the coauthors of this work had experience working with domain scientists in these four fields. While there were other fields that were more popular or featured unique usage, such as Environmental Science, they were either very similar to the chosen fields or they were outside of the area of expertise of the co-authors. We note that this selection of four domains does limit the generalizability of our results to just those domains.

Then, we use a literature database to search for papers matching keywords relevant to dimensionality reduction. We required that any paper in our analysis i) uses a dimensionality reduction algorithm as part of its data analysis and ii) presents a visualization of the dimensionally reduced data.

We use the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guideline to conduct systematic reviews [44]. We chose Scopus[1] as our literature database, based on its reproducibility and its wide coverage of the subject areas we chose [45]. The search strategy searches for keywords ("Dimensionality Reduction", "UMAP," "t-SNE," "PCA," "Projection") within Scopus subject areas of chemistry, biology, physics, and business within the last five years. Subject areas were based on the Scopus database labels. These projection techniques were used based on the bibliometric findings that these were the most common methods. In addition, these methods have been studied previously by the visualization community.

Our initial Scopus search returned 52141 matching papers. We then filtered to only those papers with more than 5 citations, according to Scopus, to filter out papers with low impact, reducing the number of publications to 21083. Next, we conducted a stratified sample from this set down to 2000 publications, with 500 from each of our 4 subject areas. From this set of 2000 publications, we used the Zotero Reference Manager's[2] *Find Available PDF* function from our academic library's connection to identify 930 of those publications for which we could easily find the PDF for review. While the additional PDFs could eventually be found, we believe that this filtering to 930 publications should serve as a fairly uniform subsampling of the 2000 publications, which was appropriate since we ultimately sought to select just a small sample of the papers. This sampling is allowable because we were not aiming to completely survey all works using dimensionality reduction in these fields; rather, we aimed to merely sample them.

We divided these 930 publications amongst the four authors and scanned them to filter out any publications that did not have any visualization of the dimensionality reduction results, which removed 62% of the publications. Then, from the resulting set of 347 publications, each author randomly selected five publications from each of the four subject areas, as designated by Scopus. This resulted in 71 papers being reviewed: 20 Biology, 20 Chemistry, 17 Business, and 14 Physics (upon closer reading, 3 out of the 20 Business and 6 out of 20 Physics were incorrectly classified by Scopus and instead came from other fields, i.e. industrial design and engineering, and so were excluded from our analysis).

We discuss insights from this paper's winnowing process in section 6. We note that the relatively small sample of 71 papers included in our survey are not completely representative of the more than fifty thousand papers returned in our initial search in Scopus. However, we also believe that 71 papers is a large enough sample to provide valuable insights into the usage of DR outside of computer science. In addition, we investigate the author list of these 71 papers to ensure that there is not an overrepresentation of any particular authors. Across these 71 papers, there were 705 unique authors with only a single author appearing on more than one paper (two). We believe this sample represented a diversity of authors across these domains. In addition, we analyzed the listed affiliations of these papers and found that only 6 out of 705 unique authors listed an affiliation in a department of computer science or information science [46]–[51].

1. Query strings included in the appendix.
2. https://www.zotero.org/about/

## 5 Findings

In this section, we describe the findings from our literature review. We organize our findings based on our grounded analysis found in Section 4. We present findings related to the **data** being projected, the **design** of the visualization of the projected data, the **tasks** used in interpreting the projection according to the in-line text and captions, and the high-level **workflows** that the projections serve within the flow of the publications. We list the subject area when citing works from the survey to identify similarities and differences (**RQ3**) observed during our in-depth review. Classification results can also be viewed and explored in our online browser, available at https://dimension-reduction-vis.github.io/. This browser includes screenshots of visualizations from each paper that we read as part of this report.

### 5.1 Data

#### 5.1.1 Data Shape

We recorded the number of rows (i.e., points) and columns (i.e., dimensions) of the data fed into the DR method if they were reported by the authors. The data varied significantly across the 71 publications we surveyed. The size of the data ranged from as few as 7 data points [52] to more than twenty million [47]. The number of dimensions of data ranged from 4 [53], [54] to more than 13 thousand [55].

When there are many data points, points are drawn with some level of opacity to communicate the density of the projected space as in Physics Cheng et al. [46], as in Figure 6. In studies with few data points, the data points were typically categorized to understand which points were similar and which were different. For example, in Hasan et al. [56] (Physics), eight different metals were analyzed by their concentrations found in either soil samples (SS) or food samples (FS) grown in that soil. PCA is used to project those samples into a two-dimensional space, and the relative location of the two types of samples in the projected space is used to conclude that the distribution of metals was different. Similar phenomenon was observed also for Chemistry (e.g., relating 8 different states of a kinase [57]), Business (e.g., analyzing relationships between the economies of European nations [58] or Chinese corporate brands [53]), and Biology (comparing different strains of bacteria [59]).

#### 5.1.2 DR Method

We also examined the type of DR method used to process the data. We identified that higher dimensional data necessitates domain-specific and/or nonlinear dimensionality reduction techniques and that the types of data found in different subject areas drove different methods. Biology and Chemistry papers tended to employ more nonlinear dimensionality reduction methods (e.g., UMAP, t-SNE). The Biology publications that reported input dimensions reported an average of $\mu = 2186$ dimensions and Chemistry $\mu = 751$, compared to Business $\mu = 19$ and Physics $\mu = 12$. It is likely that the phenomena being

captured in high dimensional data are not often visible or might be difficult to identify in a linear projection like PCA. For example, in the Chemistry field, Mazher et al. [60] analyzed a dataset of 5,688 data points with 273 dimensions, comparing different non-linear dimension reduction methods, including UMAP, t-SNE, and a newer technique, Potential of Heat-diffusion for Affinity-based Trajectory Embedding (PHATE). The choice of the DR technique may also be the result of disparate tasks in each subject area, which we discuss shortly.

Surprisingly, in each domain, there were examples of papers that used multiple dimensionality reduction techniques because the techniques were seen as being useful for different types of data. An example is found in Physics Lee et al., in which various data extracted from a small sample of newborn blood is analyzed [48]. The blood samples were processed to read several categories of data: (1) cellular composition, (2) plasma cytokines/chemokine concentration, and (3) several other biological measures like protein composition and metabolomic data. In this case, PCA was used to demonstrate the separability of the data for categories (1) and (2). However, the authors suggested that a domain-specific technique, Data Integration Analysis for Biomarker discovery using Latent cOmponents (DIABLO), designed for biomarkers, was needed when integrating the third category of data because of *"the complexity of the data ... and the heterogeneous nature of data measured on different scales and technological platforms"*.

In 14/71 (20%) of the papers in our survey, a technique besides PCA, tSNE, or UMAP was used. This alternate technique was frequently a factor analysis, which is a statistical technique that uncovers primary factors that result in the separation of data points. In this analysis, the dimensions of the data are combined in a linear combination into two different components similar to PCA, but use a different method that is more standard within their domain, such as orthogonal projections to latent structures discriminant analysis (OPLS-DA) within Biology [61]. Single-cell Biology is unique in that some nonlinear techniques have been developed by computational biologists that are specially designed for high-dimensional biology data [47], [62].

## 5.2 Design

We categorize the design of the visualization of the DR view by its plot type, its plot style, and the annotations used.

### 5.2.1 Plot Type

We consider four plot types, all variations on the scatter plot. 3D and 2D describe scatterplots with three and two dimensions, respectively. 2D+ describes a two-dimensional scatterplot integrated with an additional plot, as in Figure 5a [66] (a), or if a three-dimensional plot is used where one of the dimensions is not a projected axis, but instead used to view correlation as in Chemistry Figure 5b [64]. Likewise, 1D+ features a one-dimensional plot of the data points, as seen in Business Figure 5c [67]. The overwhelming majority (63/71) of papers included a 2D plot, with 9 3D plots, 8 2D+ plots, and 7 1D+ plots.

### 5.2.2 Plot Style

We report whether plots have a legend, drawn axes, gridlines, and small multiples of scatterplots. Legends were less common in Business (24%) and Chemistry (50%) than Physics (57%) and Biology (85%). Axes were usually drawn (89%), while gridlines were less frequent (37%), although they were markedly more frequent in Business (65%). Small multiples were rarely used in Business (12%), Chemistry (25%), and Physics (36%), but frequently in Biology (65%). In some cases, small multiples were used to analyze more than two composite dimensions of the dimensionally-reduced data as in Business [68] or Physics [56] (Figure 7a), while in other cases, the same view of the DR space is presented multiple times with different variables encoded by color as in Biology (Figure 7b) [66].

We additionally record whether the plots are annotated using glyphs or symbols to signify another variable, textual labels directly on the plot, highlighting, and captions. Captions, or textual descriptions spatially attached to the figure, were frequently used across our survey corpus (82%). Textual labels directly annotating visual elements (51%), highlighting (37%), and glyphs or symbols (24%) were used less frequently with no clear patterns of difference between different subject areas. Annotations were commonly used to highlight clusters and identify relationships between clusters and other variables. In particular, many papers including Chemistry [69] and Biology [70] would draw enclosing circles around clusters or to signify 95% confidence ellipses, and also use color and shape to show additional attributes beyond those used in the projection. These annotations were often explicit, using text directly on the plot to label clusters. In some cases, arrows or connecting edges were drawn to show connections between data in two different known groups, as seen in Figure 7c. Axis titles often note the percentage of variance explained using text. We observe that authors also draw quadrants and annotate different sections with interpretation. There is no one common way of explaining axes.

## 5.3 Tasks

### 5.3.1 Task Descriptions

Through our grounded analysis and our close reading of our survey of subject areas, we identified seven common tasks that were commonly used in describing how the plots of DR data should be interpreted. Each task is illustrated and described below, and the percent of papers the task is found in is reported. To address RQ4, we compare our seven tasks with the tasks identified in three previous visualization works: Etemadpour et al. [20], Nonato & Aupetit [2], and Xia et al. [21]. Notably, we only consider tasks that are relevant to the interpretation of static views of dimensionally-reduced data, rather than those tasks available in interactive systems. As a result, we exclude tasks that explicitly refer to user interactions with DR, such as those from Sacha et al. [1].

**Single Point** (9/71 = 13%). In this task, a single data point is highlighted and described, whether in the main text or within a caption. Its place within the projected space

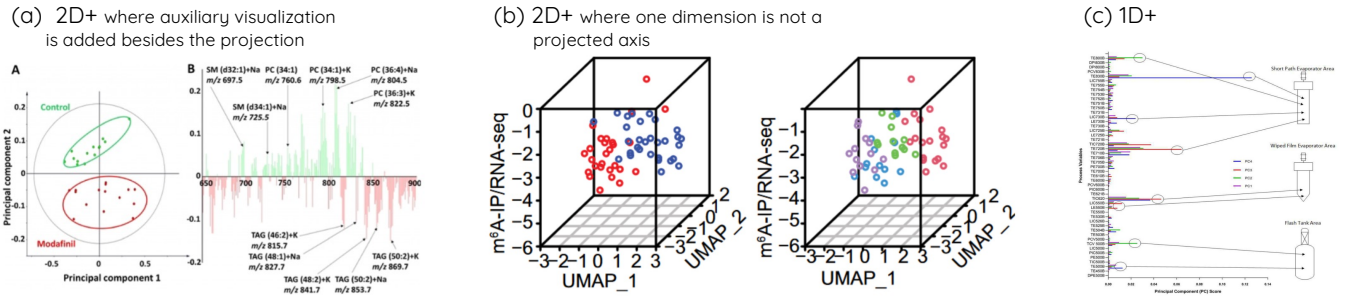



Fig. 5. The example figures for 2D+ and 1D+ plot types (Section 5.2.1). (a) 2D+ plot in which an auxiliary *plot* is added to augment the projection. (b) 2D+ plot in which an auxiliary *axes* is added to represent the dimensions that are not a projected axis. (c) 1D+ plot consists of a 1D projection and an auxiliary plot. Found in [63]–[65], respectively

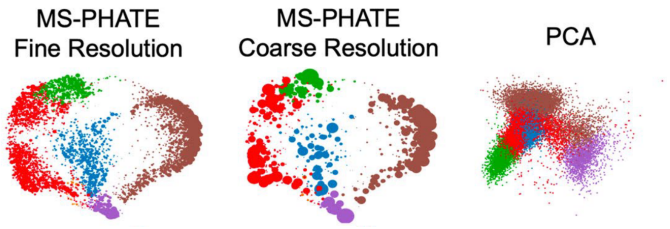
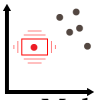
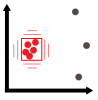



Fig. 6. Three plots of 22 million measurements of peripheral blood mononuclear cells, or PBMCs, gathered via flow cytometry [47]. When the number of elements in the projections is high, transparency is often used so that the general distribution of the data in the projected space can be seen.

is used to provide insight into the particular point, as in Physics identifying the distance of a control point from the rest of volatile compounds in the DR view [71]. The closest proposed task is from Etemadpour et al. [20] to identify the closest cluster to a given object (*fCluObj*). Nonato & Aupetit mention a similar task of *finding a seed point*, and Xia et al. do not include a task on identifying a single point. We find that the interpretation of a single point is not a common task within our survey, but when it is a task, it typically involves identifying proximity to a cluster.



**Multiple Points** (8/71 = 11%). In this task, multiple data points are described and compared. This comprises the tasks of identifying nearest neighbors from Etemadpour et al.. but can also be broader to account for expected relationships between points. For example, in Biology Cui et al., particular macrophages are identified in a PCA view as being related to pulmonary fibrosis lungs [72]. In Business Feuillet et al., ten years of French soccer club seasons are projected into a PCA plot, and multiple points corresponding to multiple seasons of the same club are analyzed to understand changes in strategy over time [73]. This example may be similar to the task of Nonato & Aupetit of *identifying a path* within dimensionally-reduced data. Again, this is not a common task, suggesting that individual points are not typically the object of analysis within our survey.



**Single Cluster** (23/71 = 32%). In this task, multiple data points are described and compared based on their visual clustering within the projected view. It is identified by both Nonato & Aupetit and Xia et al. as an exploration task to discover a cluster within a projected view. In Etemadpour, this task is separated into two subtasks: *#SClu* i.e., estimating the number of subclusters within a cluster, and *#Obj* estimating the number of objects within a cluster. However, we did not find counting to be a common action in analyzing a cluster. Instead, a cluster might be analyzed to develop an explanation for the isolation of some particular subgroup within the data. Examples include Biology Potluri et al. identifying a set of patients with a particular metastatic disease phase being visually clustered together in a PCA view of antibody profile data [55], or Chemistry Wang et al. identifying a cluster of transitional states of an enzyme as shown in Figure 7c [57].

**Multiple Clusters** (46/71 = 65%). In this task, multiple clusters are analyzed. This could be to compare different groups. This is found in Chemistry comparing treated vs. untreated samples [52], [74] and different organic sample locations [75], [76]. It is also found in Physics to compare different types of celestial bodies [77]. Alternatively, it could be to try to define the different clusters that emerge. As an example, in Business [78], Ji et al. interpreted the distance between clusters of different types of biofuel as being greater than the distance between clusters based on the temperatures at which those fuels were burned. This was the most common task found in our survey, but the comparison of clusters is not explicitly included as a task in any of the visualization works we compared to. Closest is the *distance comparison* from Xia et al., but that task specifically refers to identifying the closest cluster to a given cluster rather than interpreting the distance and relative locations of two arbitrary clusters. This may not be a focus of visualization research because it is fraught with potential misinterpretations we describe in Section 5.5.1, including the interpretation of global distances and nonlinear axes.

**Global Patterns** (34/71 = 48%). This task was the second-most-common of the seven we identified. In this task, different regions of the projected view independent of particular clusters are ascribed meaning. This typically involved in-

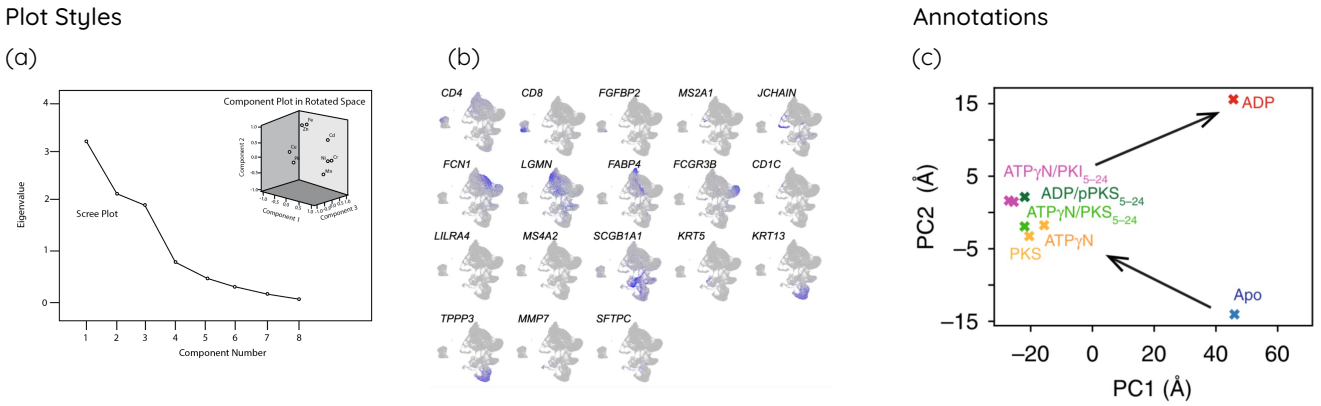
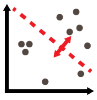



Fig. 7. The example figures for different plot styles and annotations. (a) Small multiples leveraged to analyze more than two composite dimensions [56] (figure manually reproduced by authors due to copyright claims) (b) Small multiples to compare different variables at once [66] (c) Annotations for depicting the connections between data items [57]

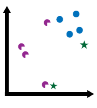
terpreting the axes or quadrants of the projected view in order to explain clusters or verify known relationships in the data. Related subtasks for interpreting global patterns are identified in Nonato & Aupetit including *naming* and *discovering relationships* between reduced dimensions and original dimensions. Examples can be found in Chemistry in interpreting different clusters of states of matter [79] or locations where samples were taken [80] and likewise in Physics to explain principal components of climate data [81].



In some cases, more than two dimensions of the dimensionally-reduced data are interpreted. Biology Jin et al. [59] analyze the first four principal components for correlations with input features using a loading plot, seen in Figure 10, which then informs the interpretation of the corresponding PCA view. Similar analyses of the principal components are split into one-dimensional plots shown in Figure 13 side-by-side with traditional two-dimensional scatterplots (not shown) in Biology [82].

The analysis is not only axis-aligned, as in Chemistry Wang et al. [57] (Figure 7c), where diagonal directions are interpreted as state transitions of an enzyme. This type of linear interpretation of a space is similar to the identified task of *discovering a path* within the projected view from Nonato & Aupetit.

We believe that the popularity of this task within our survey points to unclear guidelines on how to interpret dimensionally-reduced data. The potential distortions outlined in Nonato & Aupetit indicate that the broad analysis of the dimensionally-reduced data can result in misconceptions about the source data. We believe that there are opportunities for visualization researchers to provide more precise interpretations via the mitigations suggested in Nonato & Aupetit, which we describe in Section 6.



**Relationship to Other Variables** (29/71=41%). In this very common task, points in the projected view are additionally encoded with some additional attribute and the projected view is compared to any natural ordering provided by the additional encoding, commonly to verify that the projection separates the data by that additionally encoded data. For example, in Biology Ocasio et al. (Figure 11, far right), scatterplot points are colored by expressed transition state, and the transition of colors in the plot is annotated to demonstrate that the projected view captures the phenomenon of state transition. Often, a categorical class is used in a symbol or color encoding, which was found in Business [83], Biology [84], and many others. In other works in Biology, it was common to present complex data in small multiples with each scatterplot encoding a different variable, as seen in Figure 7b [66] and others [85]–[87]. This technique was sometimes found in Chemistry as well [88]. This very common task was not included explicitly in the visualization works we compared to, and we believe it should be an opportunity for research, as the use of glyphs, symbols, and other encodings can affect the perception of dimensionally-reduced data.

**Outliers** (5/71 = 7%). In this last task, outlier points in the projected view are described to interpret the data projection. In Business Onuferová et al., Slovakian travel agencies are analyzed both statistically and visually. Visual outliers within a PCA view are then termed as "extremes" [89]. In a similar manner, Business [90] highlight outliers in a PCA view of a data envelopment analysis of heat management companies to identify companies that are at risk of bankruptcy. The identification of outliers is a task described by Nonato & Aupetit. The interpretation of outliers is related to tasks identified by Etemadpour et al. (*fCluClu*) and Xia at el. (*membership identification* and *distance comparison*) in which a point is identified as being close to one cluster vs. another. However, in our survey we found that points far from clusters (i.e. outliers) were more frequently described rather than those close to cluster centers.

### 5.3.2 Summary of Differences from Prior Works

Compared with Etemadpour et al., we did not find that the count of the number of subclusters (#*SClu*), the number of outliers (#*Out*), the number of objects within a cluster (#*Obj*), or the relative densities of clusters (*rDens*, also identified as a

primary task in Xia et al [21]) to be commonly analyzed [20]. In contrast, in our survey we found that the analysis was generally coarse, describing a phenomenon found across a dataset, and so the count of objects within clusters (or their visual density) wasn't discussed. In addition, outliers were identified in smaller datasets in several Business works, but they were analyzed individually rather than counted.

Besides *density comparison*, Xia et al noted three additional typical tasks: *cluster identification*, *membership identification*, and *distance comparison*. The identification of clusters is found in our literature review, but the measurement of distance and the cluster membership of individual points are not commonly found.

Nonato and Aupetit present a much more complete list of 32 tasks, with one subgroup group of 8 tasks, *Explore Items in Base Layout*. These tasks include *Discover Clusters*, *Discover Paths*, and *Discover outliers*. However, it also includes interactions such as navigation and brushing, and some more local investigation into neighborhoods. In addition, out of the more than 40 papers surveyed, the tasks were only found in a maximum of 7 papers in their survey, suggesting that they were not prevalent tasks in the visualization literature.

We believe that the analyses and interpretations we observed in our literature review represented a different sample than those surveyed in the four prior works, which observed the use of dimensionality reduction in visualization research. In our observations of our literature review, counts of clusters or objects may not have been important because they are not useful in confirming prior hypotheses about the data. The authors typically used the presence of clusters or the separability of clusters to provide evidence that particular phenomena well-known in their communities (like the difference between cell types or physical sample sources) are identifiable in the data. The number of individual subclusters is not relevant to these types of hypotheses. The place of dimensionality reduction in workflows within visualization papers may be different than the place found within domain papers.

The four prior works are not a complete union of the proposed task analyses of dimensionality reduction use cases, although we note that Nonato and Aupetit do cite many design studies in their analysis. It is likely that there have been design studies, including collaborations between visualization researchers and domain scientists. We do not include individual design studies in the scope of this paper. However, future work in surveying design studies in visualization research could potentially surface novel tasks. While there are existing surveys on dimensionality reduction in visualization research, they do not include a task meta-analysis. In section 6, we further discuss the role of design studies and surveys in our opportunities for visualization researchers.

### 5.4 Workflows

Across domains, we encountered three common workflows where dimensionality reduction results were explicitly visualized in describing a data analysis process. These contrast in some ways to popular understandings of the role of visualization, such as Pirolli and Card's sensemaking loop [92] or Van Wijk's model of the value of visualization [93]. These models could suggest that the visualization of high-dimensional data in a 2D space would largely be done for exploratory data analysis, to make sense of the data, understand trends, and generate hypotheses. However, in our analysis, we found that there was often a mix of confirmatory data analysis, in which the generated projection was used as an ad-hoc statistical test [94]. In this case, the visual separation of clusters, for example, could be used as evidence that data is separable in the high-dimensional space. This separability is then extrapolated to make a judgment about the value of the data for later analysis.

**Workflow 1: Exploratory Data Analysis** One common workflow is to use a projection for exploratory data analysis to generate hypotheses, which are then verified by other statistical tests. In this workflow, a complex phenomenon is being studied to develop a greater understanding, and the goals can be developed iteratively as the data is explored. The workflow is typically seen when a rich dataset was assumed to be related to a particular phenomenon, but the causality or model connecting the data to the phenomenon was not known, as in Biology mass spectrometry data [63] and gene expression data [95], Physics remote-sensing reflectance data [96], or heuristically gathered data in Business such as forest management [97], agricultural [98] or nutritional [99] indicators, and survey data [100], [101].

For example, in Philipsen et al. [63], mass spectrometry imaging data of the fly brain is used to understand the effect of a treatment, modanifil. This generates hundreds of spectra, each one a dimension of data, and the goal of the analysis is to understand which spectra might change in response to the treatment. As seen in Figure 5a, the spectroscopy data is projected down to two dimensions with PCA, showing separability in the second principal component. Then, a loading plot shows which spectra are most discriminatory in the second principal component. This generates hypotheses that the bands of spectra that peak or valley may quantify the effect of the treatment. After identifying these spectra, the authors explain *"it is hard to interpret the precise differences based on the scores and loadings. Hence, the changes in the level of each molecular species are measured and evaluated with statistical analysis."*

Similar workflows were used in Physics literature. As described in Conterosito et al. [102], dimensionality reduction can be applied to physical data to "speed up analysis with the specific goals of assessing data quality, identifying patterns where a reaction occurs, and extracting the kinetics." This type of analysis was also used to generate hypotheses in Kobaka et al. [103] to identify differences in concrete mix designs.

**Workflow 2: Confirmatory Data Analysis** In the second type of workflow, the visualization of the dimensionality reduction results is used to draw a conclusion rather than generate a hypothesis. The hypothesis is often that the high dimensional data has sufficient information to separate the variable of interest, often a binary or categorical variable. This workflow is frequently used when there is a well-known phenomenon that is potentially expensive to measure. An alternative method may be proposed to gather high-dimensional measurements of the object to be classified and then hypothesize that in this high-dimensional measurement, the data is separable. This measurement may

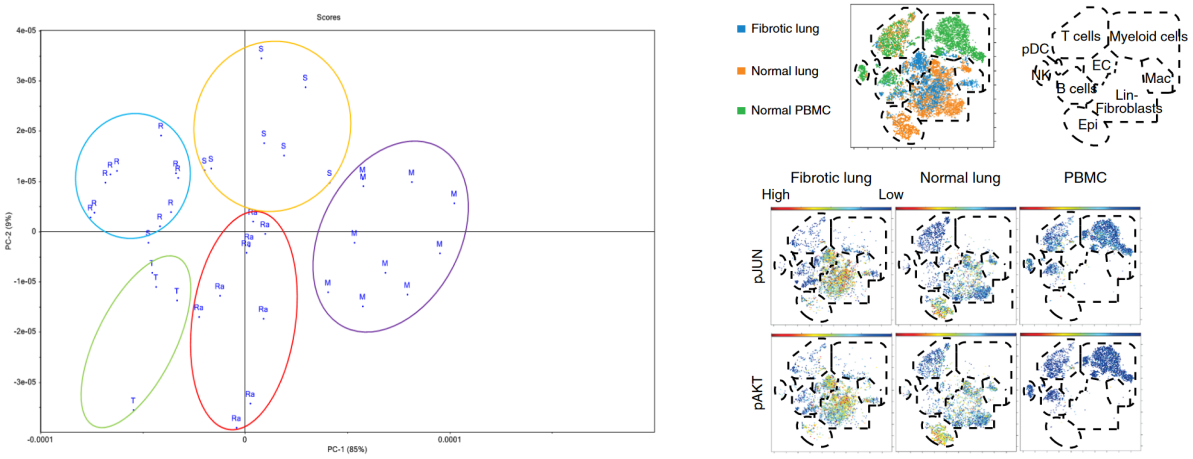



Fig. 8. We identify three common workflows for using visualization of dimensionality reduction results, two of which are shown here and one is evident in a previously highlighted figure in Figure 5a. The three workflows differ based on whether the visualization is used for exploratory data analysis as in Figure 5a [63] where the relationship between frequencies in a spectroscopy and clusters is being explored, confirmatory data analysis as in (a) of this figure [91] where known groupings in the data are verified to exist in the projected space, or a combination of both as in (b) of this figure [72]. In this combined workflow, known groupings are first used to informally validate the layout. Then, the relationship between additional variables (in this case, the activations of two transcription factors in different types of tissue) is explored in the projected space.

be a novel process (i.e. Physics electric tongue [91], Surface-enhanced Raman scattering sensors [104]), Chemistry (i.e. different types of spectrometry of liquids [105]–[107] or electric measurements of scent [49], [108]), Biology (i.e. microbiome data [81], MALDI fingerprinting [109] or gene expression data [49]), or Business (i.e. novel qualitative analyses of businesses, countries, or groups of consumers [53], [54], [58]).

Physics Pauliuc et al. [91] provides an example where PCA is used to reduce data from a measurement of honey using a process called a *Voltammetric Electronic Tongue* to determine which flower influenced its flavors. In this common type of confirmatory data analysis workflow, the separability of clusters in the visualization (see Fig. 8 (b)) is used as confirmation that the process can successfully recover the type of flower, or if the honey was not the product of a single type of flower but rather a combination of flowers, which is a less desirable type of honey. The process being validated may also be a statistical process, as in Business [110], where authors conduct a cluster analysis based on survey data analyzing internet habits of older populations in Spain. A PCA view of the data is annotated with the convex hull of the discovered clusters, showing the separability of most of the clusters.

A novel example of this workflow is found in Physics Otten et al. [67], where a deep generative model is used to generate events in a physical process. The analytical goal is to interpret the data generated from the deep generative model, evaluating it for use in further analysis. PCA is used to reduce the dimensionality of a hidden layer of the generative model to the first two principal components. However, before visualizing the data, that two-dimensional space is converted to polar coordinates, and samples are taken on a grid in polar coordinates (Fig. 12). Samples on this grid are given a visual encoding representing multiple dimensions of the underlying physical data.

Lastly, this type of workflow was used not only to confirm hypotheses but also to reject hypotheses. Chemistry Nurani et al. [111] used H-NMR spectroscopy to extract metabolite data about turmeric plants in order to classify the particular species. They use several dimensionality reduction techniques to identify clusters and evaluate separability and conclude that PCA is only able to distinguish certain species and not others: "*Chemometrics of PCA could not differentiate C. longa, C. xanthorrhiza, and C. manga clearly (data not shown). It might be caused by the large variations of the variables; therefore, the principal components (PC) were not able to represent the original variables.*" Interestingly, they do not include a plot of the results they use to make this conclusion. They go on to state that a different DR technique was able to distinguish between them: "*Observation using supervised pattern recognition, namely PLS-DA using 7 PC, could classify C. longa, C. xanthorriza, and C. manga resulting in three different classifications.*"

It is likely that this type of workflow is broadly analogous to assessing the value of a projection by its perceptual cluster separability, which is a quantitative metric describing the distance and clarity of separation that has been identified as a quality metric for projection techniques [112]. Cluster separability on projections of data with known ground truth can be used in a quantitative evaluation or comparison of different projection techniques.

**Workflow 3: Confirmatory then Exploratory Data Analysis**
In the third workflow, a projection is generated to ultimately be used for exploratory data analysis to generate hypotheses for the correlation between data features. However, in order to build a greater level of confidence in the projected view, it is first evaluated via confirmatory data analysis.

This workflow is frequently seen in domains where there is complex input data being used to study a phenomenon that is not well understood. First, the projection is inspected visually, with points colored according to some known

quantity that should be separable within the data. Visual separation confirms that there is some meaning in the layout of the points in the projected view. Then, the projection is repeated but with additional data encoded on each point to develop new hypotheses and potentially enrich the understanding of the meaning of known separable groups. The analysis then continues into additional exploration and confirmation of those hypotheses through other statistical tests.

This type of workflow was typical in Biology (eg. [55], [72], [113]–[115]) where many physiological processes interact and are typically represented by a high dimensional dataset, and there is a large space of potential hypotheses that can be narrowed through exploration of high dimensional data. As an example, in Cui et al. (see Fig. 8(c)), a tSNE projection of lung cells is first projected to confirm that abnormal tissue is separable from normal tissue, as well as cell type [72]. Then, that view is colored by the variables of interest, pJUN and pAKT, to understand if their concentration differs from abnormal tissue to normal tissue. It is identified that fibrotic lung tissue sees high readings of both variables in a particular cell type, fibroblasts, which are then further analyzed. This technique can sometimes be used to color many variables of interest, making use of small multiples (see Biology Figure 7b).

One notable example was found in Business Teng et al. [65]. A factory process is being optimized, and PCA is used to find a subset of processes that might be easily experimented on together without disrupting too much of the manufacturing process. First, PCA scores for the first four components are shown against the high dimensional features and confirmed to find meaningful clusters, and then the clusters are used to drive the design of experiments and optimizations using domain knowledge of the factory itself. This analysis is seen in Figure 5c.

### 5.5 Gaps between Research and Data Analysis

Our literature review identified gaps between the practical use of DR and the relevant research conducted by the visualization community. The findings ignite open challenges for the visualization community to reduce the gap.

#### 5.5.1 *Inappropriate Usages*

Our analysis reveals inappropriate usages of DR techniques that could lead to erroneous conclusions about the relying data, e.g., relying on assumptions that were not guaranteed by the methods being used. Please refer to Figure 11 for examples.

**Inappropriate choice of DR techniques** The visualization community has analyzed which DR techniques are best suited for various visual analytics tasks, such as cluster identification and neighborhood search. For example, Xia et al. [21] reveal that UMAP and t-SNE are the most appropriate techniques for cluster identification tasks. They found that PCA is not suitable for the task, which means that PCA shows inaccurate cluster representations. This is largely achieved by conducting benchmark studies, where the appropriateness of DR techniques is evaluated using scores from DR quality metrics [14] or human task accuracy [20], [21].

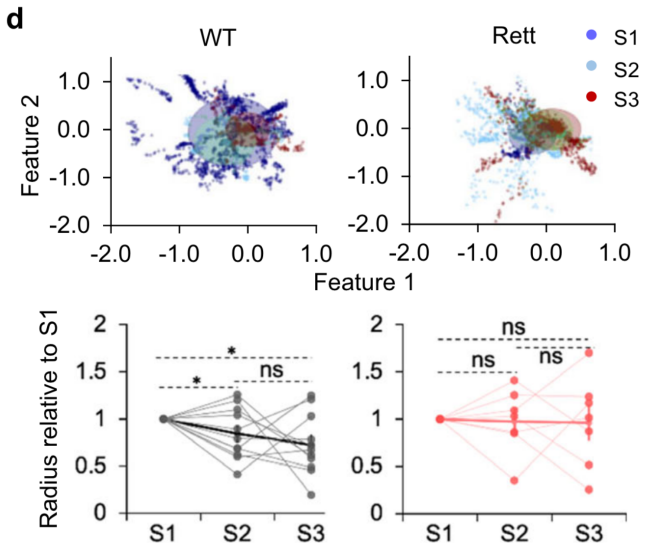



Fig. 9. In Xu et al. [50], authors use a Variational Autoencoder (VAE) to reduce the dimensionality of in vivo image frames of mouse neurons. The authors recorded brain imaging data for two types of mice doing three types of tasks, S1, S2, and S3. The linear distance from each point to the cluster center of points in task S1 is reported in the bottom plots. However, the VAE method can distort distances, making this possibly misleading information (more in Section 5.5.1).

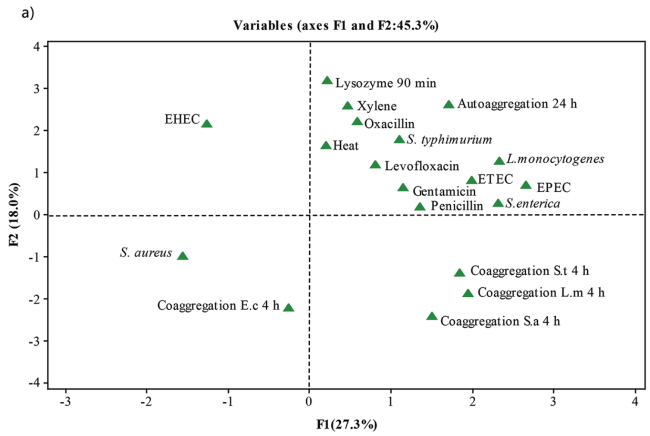



Fig. 10. In Jin et al. [59], quadrants of a PCA loading plot (in which points represent features) are used to understand how features are (anti-)correlated with the first two principal components.

However, we identify that research works in four domains often use DR techniques that do not match with their task. We especially find that PCA is widely used for cluster identification tasks (Figure 11a), although it is less suitable for the task [21]. This inappropriate usage degrades the reliability of the findings made by the research work; for example, the clusters found by the practitioners may not stay as clusters in the high-dimensional space [8], [119].

**Inappropriate plotting of DR techniques** The visualization community also informed practitioners how to plot DR projections properly. For example, Faust et al. [120] emphasized that conventional $x$ and $y$ cannot be used to interpret nonlinear DR techniques, proposing a new nonlinear axes visualization technique. Jeon et al. [121] and Aupetit et al. [122] claimed that conventional brushing that selects 2D

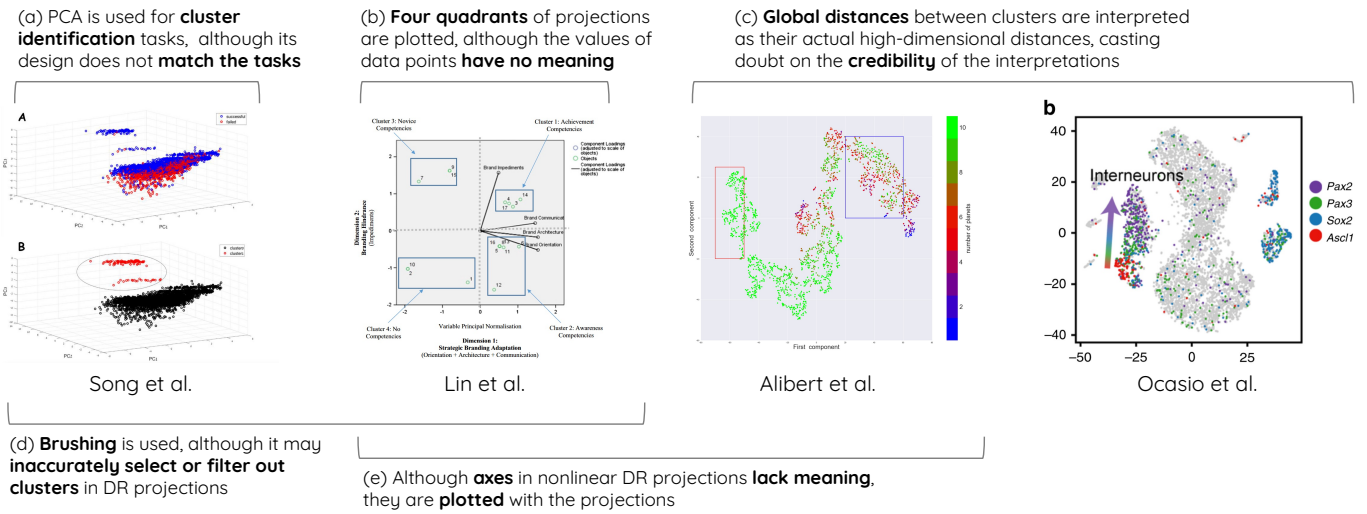



Fig. 11. The examples of our findings on inappropriate usage of DR techniques (Section 5.5.1), found in [53], [116]–[118]. These improper usages lead data analysis to have limited reliability, casting doubt on the conclusions in which research works made.

regions should not be used for DR projections, contributing new brushing techniques that locally resolve distortions.

Still, we find cases where research works in four domains do not align with such guidelines. For axes plotting, we identify papers that plot axes with titles such as UMAP1 & UMAP2 or TSNE1 & TSNE2 or with grid lines (Figure 11e). We also find the case in which four quadrants of the dimensionally reduced view are used to define four clusters (Figure 11b), suggesting that the positive and negative directions of the axes hold semantic meaning for the authors. Regardless, these annotations could mislead audiences, in particular, those unfamiliar with the properties of the underlying DR techniques used.

In terms of brushing, we find that several papers emphasize the clusters using fixed-shape brushes (e.g., rectangular, ellipse, or spheres), which violates the guidelines made by the visualization community (Figure 11d). As with inappropriate technique selection, such a violation degrades the credibility of brushed clusters.

**Inappropriate interpretation of data patterns** The visualization and machine learning community provided guidelines to interpret DR projections based on the design of the DR techniques used. For example, Wattenberg et al. [12] guided practitioners not to interpret global distances between clusters in t-SNE plot as their distances in the original high-dimensional space. The claim has been further verified by many articles, not only for t-SNE but also for other nonlinear DR techniques like UMAP [13], [62], [123]–[125].

However, our review reveals that research works in four domains often do not comply with such guidelines. For example, it was common to interpret the global distances between clusters in DR projections that do not preserve the global structure (Figure 11c), casting doubt on the credibility of the analysis results.

In another example, Physics Xu et al. [50] project frames of mouse brain imaging data into a 2D latent space using a Variational Autoencoder (VAE) (Fig. 9). The authors stratify the data points into groups based on experimental conditions. For each group of points, they compute a distribution radius defined as "*the average distance of all frames to the center in the 2D latent space.*" The authors proceed to draw conclusions by comparing the distribution radii among experimental conditions. Because VAE is a potentially nonlinear method, it is possible for distances in the 2D space to be distorted, preventing linear comparison.

**Execution of statistical test using dimension-reduced data** The biology and biostatistics community provided guidance not to conduct statistical tests using dimension-reduced data [126], [127]. This is because the data is distorted during the reduction process and then fed to the statistical test, which means that it has been "double-dipped."

However, several research works in four domains apply statistical tests like the *t*-test to the PCA results, casting doubt on the validity of the test (e.g., [116]). Even though the solutions for this problem have been widely proposed in the biology community [128]–[131], the incorrect execution of statistical tests is rampant in the field.

### 5.5.2 *Issues with Reproducibility*

We find that in most cases, the authors specified the software package (e.g., `scikit-learn` [132]) that was used to perform DR and reported how they preprocessed data. However, papers failed to mention whether optional parameters (e.g., perplexity in the case of t-SNE) were used, degrading their reproducibility. This is because most works have used the default hyperparameter settings of the library they used. Still, there are possibilities in which hyperparameters are cherry-picked to generate DR projections that best align with the papers' hypotheses. The absence of a hyperparameter report also raises concerns about whether the hyperparameter values have been properly optimized, negatively impacting not only reproducibility but also the credibility of the data analysis.

## 6 Discussion

In this section, we summarize the takeaways of our survey for both target audiences: 1) visualization and machine

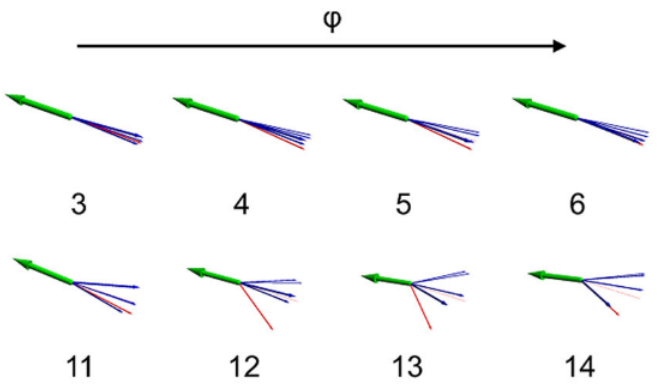



Fig. 12. In Otten et al. [67], the PCA scores are converted to polar coordinates, and points within that space are visualized as vectors, with the thickness of the green arrow and directions of the blue and red arrows encoding physically meaningful values in the original data.

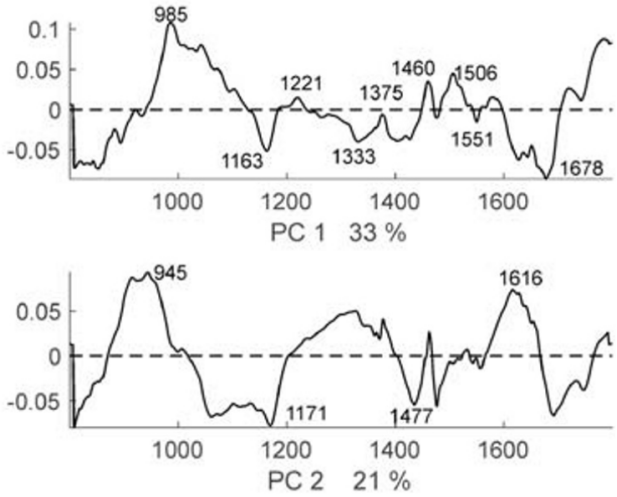



Fig. 13. In Diehn et al. [82], spectra from pollen samples comprising five species of grass were obtained using Fourier-transform infrared spectroscopy (FTIR) for chemical characterization. PCA was applied to a dataset of fifty observations (each observation representing the average of 20 pollen grains from a single plant) of the spectral range from 800 to 1,800 $cm^{-1}$. The authors plot the loadings for PCs 1 and 2 against the measured spectral range. The authors note that two groups of grass species (which appear in different regions of a PC1-PC2 scatterplot) can be distinguished from each other by taking advantage of the extreme values observed in the loading plots around 1,678 $cm^{-1}$ for PC1 and around 945 $cm^{-1}$ for PC2, which correspond to known molecular vibrations of proteins and carbohydrates.

learning researchers and 2) domain-centered data analysts and practitioners. We note that due to our selection of four domains, there are some limitations to the generalizability of our findings, but we believe that the diversity of our four domains lends weight to our findings and, thus, our takeaways and open problems.

### 6.1 Takeaways For Domain Researchers

A theme of this survey is that there is a disconnection between DR-focused research in the visualization community and DR usage in the subject areas of Biology, Business, Chemistry, and Physics. Despite a diversity in the types of findings that authors discuss when referencing DR plots, we find that most use 2D scatterplots for visualization. Certain types of such findings may benefit from alternative visual encodings or may be apparent without the usage of dimensionality reduction. We encourage researchers using DR methods to ask themselves what value the DR added to their analysis. This way, the process of creating visualizations can be centered around the communication of those findings that are intrinsic to the usage of DR.

Domain scientists and visualization researchers should recognize the ways in which DR results can be interpreted and visualized correctly. Solutions to problems discussed in Section 5 have been proposed in certain subject areas, but researchers in others may not yet be aware. For instance, den-SNE and densMAP published in Biology enable interpretation of point density that would be erroneous using the original t-SNE and UMAP algorithms [133]. PHATE, also published in Biology, is a nonlinear DR method that preserves global structure and patterns, such as branch points and trajectories [62]. Methods that account for double usage of data during statistical testing have been proposed in Biostatistics [128]–[131].

We hope this survey motivates further qualitative analyses of how dimensionality reduction techniques are used and visualized within and across fields. Domain scientists may be able to provide DR usage guidance that is tailored to the data types and analysis workflows that are commonly encountered in the field.

### 6.2 For Visualization Researchers

We organize our takeaways for visualization researchers into two categories. First, we provide takeaways into the usage of DR in visual analytics systems, and then focus on the need for developing new visual analytics techniques to address the needs of domain scientists.

#### *6.2.1 Mismatches in Usage Patterns*

There were many differences between the types of dimensionality reduction used in our observed domain literature and that found in visualization research. In a recent survey of the use of embeddings in visualization systems by Huang et al. [134], it was identified that most visual analytics systems use nonlinear dimensionality reduction algorithms, such as t-SNE or UMAP. Similarly, according to a survey by Espadoto et al., 33 out of 44 (75%) dimensionality reduction techniques used in visualization papers are nonlinear methods. In our report, we find that PCA is overwhelmingly the most common technique. Visual analytics designers could consider using linear techniques more frequently, because of their familiarity and ease of interpretation - especially if the phenomenon being studied is visible within the linear projected view or the number of data points is low. Nonlinear techniques can be superior to linear techniques at cluster separability and anecdotally in identifying clusters in image databases. However, Based on the workflows identified in section 5, linear techniques may be more helpful in confirming or generating hypotheses.

In addition, visual analytics designers should consider alternatives to scatterplots that better support their users' tasks. If the goal is to understand differences between

known clusters or categories in the data, it may be better to include other encodings of cluster distances or vector distances between data points. Alternatively, visual analytics designers can consider interpretability as an objective in their choice of linear projection [135].

Lastly, we note that out of all papers included in our report, none used the types of interactive analysis that are frequently featured in visual analytics systems for the type of cluster analysis frequently used on high-dimensional data [136], [137]. Instead, domain scientists use standard techniques that are more easily reproducible. Creating easily usable open source software and publicizing methods for interpretability through tutorials or guidelines papers in meaningful venues for domains could improve our ability to reach other communities.

#### 6.2.2 Research Opportunities for the Visualization Community

**Needs for a unified guideline that informs when to use which techniques** The visualization field provided several guidelines for selecting a DR technique that matches analytic tasks [2], [21], [22]. Also, the community has provided several empirical studies that ground these guidelines [14], [21], [138]. However, as described in the recent survey by Espadoto et al. [14], these guidelines are fragmented, and the interaction techniques developed by the visual analytics community also broaden the ambiguity in selecting appropriate DR for a given context dimensionality reduction [1]. There is a need for simple guidelines on when to use what technique and how to incorporate interactive techniques without introducing bias. This would mitigate the risk of early adopters within the domain sciences from using novel techniques. An empirical guidelines work was published more than a decade ago by Bertini et al. [27], but the landscape of dimensionality reduction has changed enough that there is a demand for more guidelines.

**Tangible and detailed guidelines beyond the selection of techniques** This survey reveals that domain researchers often use DR techniques in ways that can lead to misinterpretations (Section 5.5.1), and their methodologies often lack reproducibility Section 5.5.2. These findings clearly indicate the need for more detailed guidelines that help domain researchers . As aforementioned, the visualization field provides several guidelines for using dimensionality reduction (DR), but these guidelines mostly map analytic tasks to techniques [2], [21]. They support domain researchers in selecting good techniques but do not offer insights on interpreting and communicating the projections. More tangible guidelines that fit domain researchers with lower visualization and machine learning literacy are thus needed. For example, a detailed protocol to comprehensively report DR execution and its results in their paper may substantially benefit in enhancing the reproducibility of DR-based visual analytics. This challenge has been noted previously by dimensionality reduction researchers [120], [139], [140]; based on the usage patterns found in our surveyed domain papers, we recommend visualization researchers build toolkits and target new venues for their design studies.

**Packaging up our techniques into a toolkit** We generally found that domain scientists used DR techniques that were standard packages in common languages such as *R* or *Matlab*, or in some cases used languages developed for their particular type of data like viSNE [141] or PHATE [62]. However, the visualization field currently lacks libraries that serve various techniques developed by a community. It is clear that we can improve the usability of these techniques by packaging them in clean interfaces available in package managers, which would also improve the likelihood that domain experts will use them [24]. We also recommend that researchers make these packages more actionable. As Draco [142] did for general visualization design, we can help domain experts by building a framework that automatically recommends DR techniques and hyperparameter settings that align with the experts' task and data domain.

**Greater impact for design studies through targeted publishing** Our study indicated that in the four fields we have studied, there has not been penetration of best practices from our community. We believe there is an opportunity for greater impact if we encourage authors of design studies to publish within the venues of the application domain, in addition to visualization domains. By presenting the value of our approaches on datasets that are meaningful to domain scientists within their venues, we can make it easier for domain scientists to understand the value and risk of appropriate interpretations.

We also recommend that surveys of design studies also target applied domains. While most surveys count domain scientists among the intended audience of their work, it may be unlikely that surveys in the visualization research community are ever encountered by domain scientists. To mitigate this gap, we recommend that visualization researchers consider writing an executive summary of survey findings and share that within applied domains, potentially as letters or notes within their professional publications.

## 7 Conclusion

In this paper, we describe the state of the art in the usage and interpretation of dimensionality reduction in domain-specific data analysis across four domains: biology, chemistry, physics, and business. We conduct three iterations of analysis: 1) a bibliometric analysis of papers citing dimensionality reduction techniques, 2) a loose analysis of papers using dimensionality reduction techniques, and 3) a structured analysis of papers across four domains from the last 5 years. We classify their usage and interpretation of DR techniques and then describe qualitative findings. We believe this study provides valuable insights to both domain scientists and computer science researchers in understanding the usage of tools and the gaps that could drive further research within the visualization community.

## References

[1] D. Sacha, L. Zhang, M. Sedlmair, J. A. Lee, J. Peltonen, D. Weiskopf, S. C. North, and D. A. Keim, "Visual interaction with dimensionality reduction: A structured literature analysis," *IEEE transactions on visualization and computer graphics*, vol. 23, no. 1, pp. 241–250, 2016.

[2] L. G. Nonato and M. Aupetit, "Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 8, pp. 2650–2673, 2019.

[3] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 11, 2008.

[4] F. Cheng, M. S. Keller, H. Qu, N. Gehlenborg, and Q. Wang, "Polyphony: an interactive transfer learning framework for single-cell data analysis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 1, pp. 591–601, 2023.

[5] A. Narechania, A. Karduni, R. Wesslen, and E. Wall, "Vitality: Promoting serendipitous discovery of academic literature with transformers & visual analytics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 1, pp. 486–496, 2022.

[6] Y. Li, J. Wang, P. Aboagye, C.-C. M. Yeh, Y. Zheng, L. Wang, W. Zhang, and K.-L. Ma, "Visual analytics for efficient image exploration and user-guided image captioning," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, no. 6, pp. 2875–2887, 2024.

[7] T. Chari and L. Pachter, "The specious art of single-cell genomics," *PLoS Computational Biology*, vol. 19, no. 8, p. e1011288, 2023.

[8] M. Aupetit, "Visualizing distortions and recovering topology in continuous projection techniques," *Neurocomputing*, vol. 70, no. 7-9, pp. 1304–1330, 2007.

[9] J. Lause, P. Berens, and D. Kobak, "The art of seeing the elephant in the room: 2d embeddings of single-cell data do make sense," *PLOS Computational Biology*, vol. 20, no. 10, pp. 1–5, 10 2024. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1012403

[10] H. Jeon, H. Lee, Y.-H. Kuo, T. Yang, D. Archambault, S. Ko, T. Fujiwara, K.-L. Ma, and J. Seo, "Unveiling high-dimensional backstage: A survey for reliable visual analytics with dimensionality reduction," *arXiv preprint arXiv:2501.10168*, 2025.

[11] H. Jeon, H.-K. Ko, J. Jo, Y. Kim, and J. Seo, "Measuring and explaining the inter-cluster reliability of multidimensional projections," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 1, pp. 551–561, 2022.

[12] M. Wattenberg, F. Viégas, and I. Johnson, "How to use t-SNE effectively," *Distill*, vol. 1, no. 10, p. e2, 2016.

[13] H. Jeon, Y.-H. Kuo, M. Aupetit, K.-L. Ma, and J. Seo, "Classes are not clusters: Improving label-based evaluation of dimensionality reduction," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, no. 1, pp. 781–791, 2024.

[14] M. Espadoto, R. M. Martins, A. Kerren, N. S. T. Hirata, and A. C. Telea, "Toward a Quantitative Survey of Dimension Reduction Techniques," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 3, pp. 2153–2173, Mar. 2021. [Online]. Available: https://ieeexplore.ieee.org/document/8851280/

[15] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. W. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell, "Dimensionality reduction for visualizing single-cell data using UMAP," *Nature biotechnology*, vol. 37, no. 1, pp. 38–44, 2019.

[16] T. A. Ujas, V. Obregon-Perko, and A. M. Stowe, "A guide on analyzing flow cytometry data using clustering methods and nonlinear dimensionality reduction (tSNE or UMAP)," in *Neural Repair: Methods and Protocols*. Springer, 2023, pp. 231–249.

[17] Z. A. Clarke, T. S. Andrews, J. Atif, D. Pouyabahar, B. T. Innes, S. A. MacParland, and G. D. Bader, "Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods," *Nature protocols*, vol. 16, no. 6, pp. 2749–2764, 2021.

[18] F. Anders, C. Chiappini, B. X. Santiago, G. Matijevič, A. B. Queiroz, M. Steinmetz, and G. Guiglion, "Dissecting stellar chemical abundance space with t-SNE," *Astronomy & Astrophysics*, vol. 619, p. A125, 2018.

[19] D. Engel, L. Hüttenberger, and B. Hamann, "A survey of dimension reduction methods for high-dimensional data analysis and visualization," in *Visualization of Large and Unstructured Data Sets: Applications in Geospatial Planning, Modeling and Engineering-Proceedings of IRTG 1131 Workshop 2011*. Schloss-Dagstuhl-Leibniz Zentrum für Informatik, 2012.

[20] R. Etemadpour, R. Motta, J. G. d. S. Paiva, R. Minghim, M. C. F. de Oliveira, and L. Linsen, "Perception-based evaluation of projection methods for multidimensional data visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 21, no. 1, pp. 81–94, 2015.

[21] J. Xia, Y. Zhang, J. Song, Y. Chen, Y. Wang, and S. Liu, "Revisiting dimensionality reduction techniques for visual cluster analysis: An empirical study," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 1, pp. 529–539, 2022.

[22] M. Sedlmair, T. Munzner, and M. Tory, "Empirical guidance on scatterplot and dimension reduction technique choices," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2634–2643, 2013.

[23] M. Brehmer, M. Sedlmair, S. Ingram, and T. Munzner, "Visualizing dimensionally-reduced data: Interviews with analysts and a characterization of task sequences," in *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization*, 2014, pp. 1–8.

[24] H. Jeon, A. Cho, J. Jang, S. Lee, J. Hyun, H.-K. Ko, J. Jo, and J. Seo, "Zadu: A python library for evaluating the reliability of dimensionality reduction embeddings," in *2023 IEEE Visualization and Visual Analytics (VIS)*. IEEE, 2023, pp. 196–200.

[25] J. Venna, J. Peltonen, K. Nybo, H. Aidos, and S. Kaski, "Information retrieval perspective to nonlinear dimensionality reduction for data visualization," *J. Mach. Learn. Res.*, vol. 11, pp. 451–490, 2010.

[26] M. C. Thrun, J. Märte, and Q. Stier, "Analyzing quality measurements for dimensionality reduction," *Machine Learning and Knowledge Extraction*, vol. 5, no. 3, pp. 1076–1118, 2023.

[27] E. Bertini, A. Tatu, and D. Keim, "Quality metrics in high-dimensional data visualization: An overview and systematization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2203–2212, 2011.

[28] J. A. Lee and M. Verleysen, "Quality assessment of dimensionality reduction: Rank-based criteria," *Neurocomputing*, vol. 72, no. 7, pp. 1431–1443, 2009, advances in Machine Learning and Computational Intelligence. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231209000101

[29] A. K. Kimball, L. M. Oko, B. L. Bullock, R. A. Nemenoff, L. F. van Dyk, and E. T. Clambey, "A beginner's guide to analyzing and visualizing mass cytometry data," *The Journal of Immunology*, vol. 200, no. 1, pp. 3–22, 2018.

[30] T. Liechti, L. M. Weber, T. M. Ashhurst, N. Stanley, M. Prlic, S. Van Gassen, and F. Mair, "An updated guide for the perplexed: cytometry in the high-dimensional era," *Nature Immunology*, vol. 22, no. 10, pp. 1190–1197, 2021.

[31] G. Traven, G. Matijevič, T. Zwitter, M. Žerjal, J. Kos, M. Asplund, J. Bland-Hawthorn, A. R. Casey, G. De Silva, K. Freeman *et al.*, "The galah survey: classification and diagnostics with t-SNE reduction of spectral information," *The Astrophysical Journal Supplement Series*, vol. 228, no. 2, p. 24, 2017.

[32] S. Sakaue, J. Hirata, M. Kanai, K. Suzuki, M. Akiyama, C. Lai Too, T. Arayssi, M. Hammoudeh, S. Al Emadi, B. K. Masri *et al.*, "Dimensionality reduction reveals fine-scale structure in the japanese population with consequences for polygenic risk prediction," *Nature communications*, vol. 11, no. 1, p. 1569, 2020.

[33] R. M. Kinney, C. Anastasiades, R. Authur, I. Beltagy, J. Bragg, A. Buraczynski, I. Cachola, S. Candra, Y. Chandrasekhar, A. Cohan, M. Crawford, D. Downey, J. Dunkelberger, O. Etzioni, R. Evans, S. Feldman, J. Gorney, D. W. Graham, F. Hu, R. Huff, D. King, S. Kohlmeier, B. Kuehl, M. Langan, D. Lin, H. Liu, K. Lo, J. Lochner, K. MacMillan, T. Murray, C. Newell, S. R. Rao, S. Rohatgi, P. L. Sayre, Z. Shen, A. Singh, L. Soldaini, S. Subramanian, A. Tanaka, A. D. Wade, L. M. Wagner, L. L. Wang, C. Wilhelm, C. Wu, J. Yang, A. Zamarron, M. van Zuylen, and D. S. Weld, "The semantic scholar open data platform," *ArXiv*, vol. abs/2301.10140, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:256194545

[34] L. Bornmann and R. Williams, "An evaluation of percentile measures of citation impact, and a proposal for making them better," *Scientometrics*, vol. 124, no. 2, pp. 1457–1478, Aug. 2020. [Online]. Available: https://doi.org/10.1007/s11192-020-03512-7

[35] H. Starks and S. Brown Trinidad, "Choose your method: A comparison of phenomenology, discourse analysis, and grounded theory," *Qualitative health research*, vol. 17, no. 10, pp. 1372–1380, 2007.

[36] A. Diehl, A. Abdul-Rahman, B. Bach, M. El-Assady, M. Kraus, R. S. Laramee, D. A. Keim, and M. Chen, "Characterizing grounded theory approaches in visualization," *arXiv preprint arXiv:2203.01777*, 2022.

[37] M. C. Cieslak, A. M. Castelfranco, V. Roncalli, P. H. Lenz, and D. K. Hartline, "t-distributed stochastic neighbor embedding (t-SNE): A tool for eco-physiological transcriptomic analysis," *Marine genomics*, vol. 51, p. 100723, 2020.

[38] T. S. Andrews, V. Y. Kiselev, D. McCarthy, and M. Hemberg, "Tutorial: guidelines for the computational analysis of single-cell rna sequencing data," *Nature protocols*, vol. 16, no. 1, pp. 1–9, 2021.

[39] M. W. Dorrity, L. M. Saunders, C. Queitsch, S. Fields, and C. Trapnell, "Dimensionality reduction by UMAP to visualize physical and genetic interactions," *Nature communications*, vol. 11, no. 1, p. 1537, 2020.

[40] C. Lee, "How can we use neural network with entity embedding for product valuations? a case study for the car industry," *International Journal of Information Management Data Insights*, vol. 3, no. 2, p. 100187, 2023.

[41] M. Vriens, S. Chen, and C. Vidden, "Mapping brand similarities: Comparing consumer online comments versus survey data," *International Journal of Market Research*, vol. 61, no. 2, pp. 130–139, 2019.

[42] J. Wang and F. Biljecki, "Unsupervised machine learning in urban studies: A systematic review of applications," *Cities*, vol. 129, p. 103925, 2022.

[43] K. Ch'ng, N. Vazquez, and E. Khatami, "Unsupervised machine learning account of magnetic transitions in the hubbard model," *Physical Review E*, vol. 97, no. 1, p. 013306, 2018.

[44] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan *et al.*, "The prisma 2020 statement: an updated guideline for reporting systematic reviews," *International journal of surgery*, vol. 88, p. 105906, 2021.

[45] M. Gusenbauer and N. R. Haddaway, "Which academic search systems are suitable for systematic reviews or meta-analyses? evaluating retrieval qualities of google scholar, pubmed, and 26 other resources," *Research synthesis methods*, vol. 11, no. 2, pp. 181–217, 2020.

[46] T. Cheng, F. Harrou, Y. Sun, and T. Leiknes, "Monitoring influent measurements at water resource recovery facility using data-driven soft sensor approach," *IEEE Sensors Journal*, vol. 19, no. 1, pp. 342–352, 2018.

[47] M. Kuchroo, J. Huang, P. Wong, J.-C. Grenier, D. Shung, A. Tong, C. Lucas, J. Klein, D. B. Burkhardt, S. Gigante *et al.*, "Multiscale phate identifies multimodal signatures of covid-19," *Nature biotechnology*, vol. 40, no. 5, pp. 681–691, 2022.

[48] A. H. Lee, C. P. Shannon, N. Amenyogbe, T. B. Bennike, J. Diray-Arce, O. T. Idoko, E. E. Gill, R. Ben-Othman, W. S. Pomat, S. D. Van Haren, K.-A. L. Cao, M. Cox, A. Darboe, R. Falsafi, D. Ferrari, D. J. Harbeson, D. He, C. Bing, S. J. Hinshaw, J. Ndure, J. Njie-Jobe, M. A. Pettengill, P. C. Richmond, R. Ford, G. Saleu, G. Masiria, J. P. Matlam, W. Kirarock, E. Roberts, M. Malek, G. Sanchez-Schmitz, A. Singh, A. Angelidou, K. K. Smolen, The EPIC Consortium, D. Vo, K. Kraft, K. McEnaney, S. Vignolo, A. Marchant, R. R. Brinkman, A. Ozonoff, R. E. W. Hancock, A. H. J. Van Den Biggelaar, H. Steen, S. J. Tebbutt, B. Kampmann, O. Levy, and T. R. Kollmann, "Dynamic molecular changes during the first week of human life follow a robust developmental trajectory," *Nature Communications*, vol. 10, no. 1, p. 1092, Mar. 2019. [Online]. Available: https://www.nature.com/articles/s41467-019-08794-x

[49] J. Wei, Y. Zeng, X. Gao, and T. Liu, "A novel ferroptosis-related lncrna signature for prognosis prediction in gastric cancer," *BMC cancer*, vol. 21, no. 1, pp. 1–12, 2021.

[50] P. Xu, Y. Yue, J. Su, X. Sun, H. Du, Z. Liu, R. Simha, J. Zhou, C. Zeng, and H. Lu, "Pattern decorrelation in the mouse medial prefrontal cortex enables social preference and requires MeCP2," *Nature Communications*, vol. 13, no. 1, p. 3899, Jul. 2022. [Online]. Available: https://www.nature.com/articles/s41467-022-31578-9

[51] X. Xu, Z. Xie, Z. Yang, D. Li, and X. Xu, "A t-sne based classification approach to compositional microbiome data," *Frontiers in Genetics*, vol. 11, p. 620143, 2020.

[52] B. Babinszki, E. Jakab, Z. Sebestyén, M. Blazsó, B. Berényi, J. Kumar, B. B. Krishna, T. Bhaskar, and Z. Czégény, "Comparison of hydrothermal carbonization and torrefaction of azolla biomass: Analysis of the solid products," *Journal of analytical and applied pyrolysis*, vol. 149, p. 104844, 2020.

[53] F. Lin and W.-S. Siu, "Exploring brand management strategies in chinese manufacturing industry," *Journal of Brand Management*, pp. 1–29, 2020.

[54] I. Malafronte and J. Pereira, "Integrated thinking: measuring the unobservable," *Meditari Accountancy Research*, vol. 29, no. 4, pp. 805–822, 2021.

[55] H. K. Potluri, T. L. Ng, M. A. Newton, J. Zhang, C. A. Maher, P. S. Nelson, and D. G. McNeel, "Antibody profiling of patients with prostate cancer reveals differences in antibody signatures among disease stages," *Journal for ImmunoTherapy of Cancer*, vol. 8, no. 2, 2020.

[56] A. B. Hasan, A. H. M. S. Reza, S. Kabir, M. A. B. Siddique, M. A. Ahsan, and M. A. Akbor, "Accumulation and distribution of heavy metals in soil and food crops around the ship breaking area in southern Bangladesh and associated health risk assessment," *SN Applied Sciences*, vol. 2, no. 2, p. 155, Feb. 2020. [Online]. Available: http://link.springer.com/10.1007/s42452-019-1933-y

[57] Y. Wang, M. V.S., J. Kim, G. Li, L. G. Ahuja, P. Aoto, S. S. Taylor, and G. Veglia, "Globally correlated conformational entropy underlies positive and negative cooperativity in a kinase's enzymatic cycle," *Nature Communications*, vol. 10, no. 1, p. 799, Feb. 2019. [Online]. Available: https://www.nature.com/articles/s41467-019-08655-7

[58] I. Jonek-Kowalska *et al.*, "Assessing the energy security of european countries in the resource and economic context," *Oeconomia Copernicana*, vol. 13, no. 2, pp. 301–334, 2022.

[59] Y. Jin, B. Luo, J. Cai, B. Yang, Y. Zhang, F. Tian, and Y. Ni, "Evaluation of indigenous lactic acid bacteria of raw mare milk from pastoral areas in Xinjiang, China, for potential use in probiotic fermented dairy products," *Journal of Dairy Science*, vol. 104, no. 5, pp. 5166–5184, May 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S002203022100401X

[60] A. Mazher, "Visualization framework for high-dimensional spatio-temporal hydrological gridded datasets using machine-learning techniques," *Water*, vol. 12, no. 2, p. 590, 2020.

[61] S. Carbajo-Pescador, D. Porras, M. V. García-Mediavilla, S. Martínez-Flórez, M. Juarez-Fernández, M. J. Cuevas, J. L. Mauriz, J. González-Gallego, E. Nistal, and S. Sánchez-Campos, "Beneficial effects of exercise on gut microbiota functionality and barrier integrity, and gut-liver crosstalk in an in vivo model of early obesity and non-alcoholic fatty liver disease," *Disease models & mechanisms*, vol. 12, no. 5, p. dmm039206, 2019.

[62] K. R. Moon, D. van Dijk, Z. Wang, S. Gigante, D. B. Burkhardt, W. S. Chen, K. Yim, A. v. d. Elzen, M. J. Hirn, R. R. Coifman *et al.*, "Visualizing structure and transitions in high-dimensional biological data," *Nature biotechnology*, vol. 37, no. 12, pp. 1482–1492, 2019.

[63] M. H. Philipsen, E. Ranjbari, C. Gu, and A. G. Ewing, "Mass spectrometry imaging shows modafinil, a student study drug, changes the lipid composition of the fly brain," *Angewandte Chemie*, vol. 133, no. 32, pp. 17 518–17 522, 2021.

[64] H. Yao, C.-C. Gao, D. Zhang, J. Xu, G. Song, X. Fan, D.-B. Liang, Y.-S. Chen, Q. Li, Y. Guo, Y.-T. Cai, L. Hu, Y.-L. Zhao, Y.-P. Sun, Y. Yang, J. Han, and Y.-G. Yang, "scm6A-seq reveals single-cell landscapes of the dynamic m6A during oocyte maturation and early embryonic development," *Nature Communications*, vol. 14, no. 1, p. 315, Jan. 2023. [Online]. Available: https://www.nature.com/articles/s41467-023-35958-7

[65] S. Y. Teng, B. S. How, W. D. Leong, J. H. Teoh, A. C. S. Cheah, Z. Motavasel, and H. L. Lam, "Principal component analysis-aided statistical process optimisation (paspo) for process improvement in industrial refineries," *Journal of Cleaner Production*, vol. 225, pp. 359–375, 2019.

[66] E. Wauters, P. Van Mol, A. D. Garg, S. Jansen, Y. Van Herck, L. Vanderbeke, A. Bassez, B. Boeckx, B. Malengier-Devlies, A. Timmerman, T. Van Brussel, T. Van Buyten, R. Schepers, E. Heylen, D. Dauwe, C. Dooms, J. Gunst, G. Hermans, P. Meersseman, D. Testelmans, J. Yserbyt, S. Tejpar, W. De Wever, P. Matthys, CONTAGIOUS collaborators, M. Bosisio, M. Casaer, F. De Smet, P. De Munter, S. Humblet-Baron, A. Liston, N. Lorent, K. Martinod, P. Proost, J. Raes, K. Thevissen, R. Vos, B. Weynand, C. Wouters, J. Neyts, J. Wauters, J. Qian, and D. Lambrechts, "Discriminating mild from critical COVID-19 by innate and adaptive immune single-cell profiling of bronchoalveolar lavages," *Cell Research*, vol. 31, no. 3, pp. 272–290, Mar. 2021. [Online]. Available: https://www.nature.com/articles/s41422-020-00455-9

[67] S. Otten, S. Caron, W. de Swart, M. van Beekveld, L. Hendriks, C. van Leeuwen, D. Podareanu, R. Ruiz de Austri, and R. Verheyen, "Event generation and statistical sampling for physics with deep generative models and a density information buffer," *Nature communications*, vol. 12, no. 1, p. 2985, 2021.

[68] Y. Yang, M. He, and L. Li, "Power consumption estimation for mask image projection stereolithography additive manufacturing using machine learning based approach," *Journal of cleaner production*, vol. 251, p. 119710, 2020.

[69] B. J. Collins, S. P. Kerns, K. Aillon, G. Mueller, C. V. Rider, E. F. DeRose, R. E. London, J. M. Harnly, and S. Waidyanatha, "Comparison of phytochemical composition of ginkgo biloba extracts using a combination of non-targeted and targeted analytical approaches," *Analytical and Bioanalytical Chemistry*, vol. 412, pp. 6789–6809, 2020.

[70] S. Bernardo-Bermejo, E. Sánchez-López, M. Castro-Puyana, S. Benito, F. J. Lucio-Cazaña, and M. L. Marina, "An untargeted metabolomic strategy based on liquid chromatography-mass spectrometry to study high glucose-induced changes in hk-2 cells," *Journal of Chromatography A*, vol. 1596, pp. 124–133, 2019.

[71] E. Núñez-Carmona, M. Abbatangelo, and V. Sberveglieri, "Innovative Sensor Approach to Follow Campylobacter jejuni Development," *Biosensors*, vol. 9, no. 1, p. 8, Jan. 2019. [Online]. Available: https://www.mdpi.com/2079-6374/9/1/8

[72] L. Cui, S.-Y. Chen, T. Lerbs, J.-W. Lee, P. Domizi, S. Gordon, Y.-h. Kim, G. Nolan, P. Betancur, and G. Wernig, "Activation of jun in fibroblasts promotes pro-fibrotic programme and modulates protective immunity," *Nature Communications*, vol. 11, no. 1, p. 2795, 2020.

[73] A. Feuillet, M. Terrien, N. Scelles, and C. Durand, "Determinants of coopetition and contingency of strategic choices: The case of professional football clubs in france," *European Sport Management Quarterly*, vol. 21, no. 5, pp. 748–763, 2021.

[74] Y. Yuan, X. Wang, M. Jin, L. Jiao, P. Sun, M. B. Betancor, D. R. Tocher, and Q. Zhou, "Modification of nutritional values and flavor qualities of muscle of swimming crab (portunus trituberculatus): Application of a dietary lipid nutrition strategy," *Food Chemistry*, vol. 308, p. 125607, 2020.

[75] C. Castro, L. Luz, J. Guedes, D. Porto, M. F. Silva, G. Silva, P. Ribeiro, K. Canuto, E. Brito, D. Zampieri, C. Pessoa, and G. Zocolo, "Metabolomics-Based Discovery of Biomarkers with Cytotoxic Potential in Extracts of Myracrodruon urundeuva," *Journal of the Brazilian Chemical Society*, 2020. [Online]. Available: http://jbcs.sbq.org.br/audiencia_pdf.asp?aid2=5812&nomeArquivo=2019-0398AR.pdf

[76] Y. Duan, F. E. M. Santiago, A. R. Dos Reis, M. A. de Figueiredo, S. Zhou, T. W. Thannhauser, and L. Li, "Genotypic variation of flavonols and antioxidant capacity in broccoli," *Food Chemistry*, vol. 338, p. 127997, 2021.

[77] A. Clarke, A. Scaife, R. Greenhalgh, and V. Griguta, "Identifying galaxies, quasars, and stars with machine learning: A new catalogue of classifications for 111 million sdss sources without spectra," *Astronomy & Astrophysics*, vol. 639, p. A84, 2020.

[78] R. Ji, L. Su, H. Cheng, Y. Wang, J. Min, M. Chen, H. Li, S. Chen, S. Wang, G. Yu, L. Zhang, and J. Han, "Insights into the potential release of dissolved organic matter from different agro-forest waste-derived hydrochars: A pilot study," *Journal of Cleaner Production*, vol. 319, p. 128676, Oct. 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0959652621028766

[79] A. Kriston, A. Podias, I. Adanouj, and A. Pfrang, "Analysis of the effect of thermal runaway initiation conditions on the severity of thermal runaway—numerical simulation and machine learning study," *Journal of The Electrochemical Society*, vol. 167, no. 9, p. 090555, 2020.

[80] X.-k. Ma, X.-f. Li, J.-y. Zhang, J. Lei, W.-w. Li, and G. Wang, "Analysis of the volatile components in selaginella doederleinii by headspace solid phase microextraction-gas chromatography-mass spectrometry," *Molecules*, vol. 25, no. 1, p. 115, 2019.

[81] D. Xu, H. Lu, G. Chu, L. Liu, C. Shen, F. Li, C. Wang, and N. Wu, "Synchronous 500-year oscillations of monsoon climate and human activity in northeast asia," *Nature Communications*, vol. 10, pp. 1–10, 09 2019.

[82] S. Diehn, B. Zimmermann, V. Tafintseva, M. Bağcıoğlu, A. Kohler, M. Ohlson, S. Fjellheim, and J. Kneipp, "Discrimination of grass pollen of different species by FTIR spectroscopy of individual pollen grains," *Analytical and Bioanalytical Chemistry*, vol. 412, no. 24, pp. 6459–6474, Sep. 2020. [Online]. Available: https://link.springer.com/10.1007/s00216-020-02628-2

[83] K.-H. Nam, D. Y. Kim, H. J. Kim, I.-S. Pack, H. J. Kim, Y. S. Chung, S. Y. Kim, and C.-G. Kim, "Global metabolite profiling based on GC–MS and LC–MS/MS analyses in ABF3-overexpressing soybean with enhanced drought tolerance," *Applied Biological Chemistry*, vol. 62, no. 1, p. 15, Dec. 2019. [Online]. Available: https://applbiolchem.springeropen.com/articles/10.1186/s13765-019-0425-5

[84] H. Du, X. Zheng, Q. Zhao, Z. Hu, H. Wang, L. Zhou, and J.-F. Liu, "Analysis of structural variants reveal novel selective regions in the genome of meishan pigs by whole genome sequencing," *Frontiers in Genetics*, vol. 12, p. 550676, 2021.

[85] R. Manco, I. Averbukh, Z. Porat, K. Bahar Halpern, I. Amit, and S. Itzkovitz, "Clump sequencing exposes the spatial expression programs of intestinal secretory cells," *Nature communications*, vol. 12, no. 1, p. 3074, 2021.

[86] A. F. Aissa, A. B. Islam, M. M. Ariss, C. C. Go, A. E. Rader, R. D. Conrardy, A. M. Gajda, C. Rubio-Perez, K. Valyi-Nagy, M. Pasquinelli *et al.*, "Single-cell transcriptional changes associated with drug tolerance and response to combination therapies in cancer," *Nature communications*, vol. 12, no. 1, p. 1628, 2021.

[87] S. Sengupta, S. Das, A. C. Crespo, A. M. Cornel, A. G. Patel, N. R. Mahadevan, M. Campisi, A. K. Ali, B. Sharma, J. H. Rowe *et al.*, "Mesenchymal and adrenergic cell lineage states in neuroblastoma possess distinct immunogenic phenotypes," *Nature cancer*, vol. 3, no. 10, pp. 1228–1246, 2022.

[88] S. M. Moosavi, A. Nandy, K. M. Jablonka, D. Ongari, J. P. Janet, P. G. Boyd, Y. Lee, B. Smit, and H. J. Kulik, "Understanding the diversity of the metal-organic framework ecosystem," *Nature communications*, vol. 11, no. 1, pp. 1–10, 2020.

[89] E. Onuferová, V. Čabinová, and T. Dzurov Vargová, "Analysis of modern methods for increasing and managing the financial prosperity of businesses in the context of performance: a case study of the tourism sector in Slovakia," *Oeconomia Copernicana*, vol. 11, no. 1, pp. 95–116, Mar. 2020. [Online]. Available: https://journals.economic-research.pl/oc/article/view/1755

[90] R. Štefko, J. Horváthová, and M. Mokrišová, "The application of graphic methods and the dea in predicting the risk of bankruptcy," *Journal of Risk and Financial Management*, vol. 14, no. 5, p. 220, 2021.

[91] D. Pauliuc, F. Dranca, and M. Oroian, "Raspberry, rape, thyme, sunflower and mint honeys authentication using voltammetric tongue," *Sensors*, vol. 20, no. 9, p. 2565, 2020.

[92] P. Pirolli and S. Card, "The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis," in *Proceedings of international conference on intelligence analysis*, vol. 5. McLean, VA, USA, 2005, pp. 2–4.

[93] J. J. Van Wijk, "The value of visualization," in *VIS 05. IEEE Visualization, 2005.* IEEE, 2005, pp. 79–86.

[94] H. Wickham, D. Cook, H. Hofmann, and A. Buja, "Graphical inference for infovis," *IEEE transactions on visualization and computer graphics*, vol. 16, no. 6, pp. 973–979, 2010.

[95] J. F. Muñoz, T. Delorey, C. B. Ford, B. Y. Li, D. A. Thompson, R. P. Rao, and C. A. Cuomo, "Coordinated host-pathogen transcriptional dynamics revealed using sorted subpopulations and single macrophages infected with Candida albicans," *Nature Communications*, vol. 10, no. 1, p. 1607, Apr. 2019. [Online]. Available: https://www.nature.com/articles/s41467-019-09599-8

[96] P. K. Lange, P. J. Werdell, Z. K. Erickson, G. Dall'Olmo, R. J. Brewin, M. V. Zubkov, G. A. Tarran, H. A. Bouman, W. H. Slade, S. E. Craig *et al.*, "Radiometric approach for the detection of picophytoplankton assemblages across oceanic fronts," *Optics Express*, vol. 28, no. 18, pp. 25 682–25 705, 2020.

[97] F. Riccioli, R. Fratini, E. Marone, C. Fagarazzi, M. Calderisi, and G. Brunialti, "Indicators of sustainable forest management to evaluate the socio-economic functions of coppice in tuscany, italy," *Socio-Economic Planning Sciences*, vol. 70, p. 100732, 2020.

[98] M. Berton, S. Bovolenta, M. Corazzin, L. Gallo, S. Pinterits, M. Ramanzin, W. Ressi, C. Spigarelli, A. Zuliani, and E. Sturaro, "Environmental impacts of milk production and processing in the eastern alps: A "cradle-to-dairy gate" lca approach," *Journal of cleaner production*, vol. 303, p. 127056, 2021.

[99] M. Moreira, J. García-Díez, J. De Almeida, and C. Saraiva, "Evaluation of food labelling usefulness for consumers," *International Journal of Consumer Studies*, vol. 43, no. 4, pp. 327–334, 2019.

[100] E. Maghlaperidze, N. Kharadze, and H. Kuspliak, "Development of remote jobs as a factor to increase labor efficiency," *Journal of Eastern European and Central Asian Research (JEECAR)*, vol. 8, no. 3, pp. 337–348, 2021.

[101] G. Hetenyi, A. Lengyel, and M. Szilasi, "Quantitative analysis of qualitative data: Using voyant tools to investigate the sales-marketing interface," *Journal of Industrial Engineering and Management (JIEM)*, vol. 12, no. 3, pp. 393–404, 2019.

[102] E. Conterosito, M. Lopresti, and L. Palin, "In situ x-ray diffraction study of xe and co2 adsorption in y zeolite: Comparison between rietveld and pca-based analysis," *Crystals*, vol. 10, no. 6, p. 483, 2020.

[103] J. Kobaka, "Principal component analysis as a statistical tool for concrete mix design," *Materials*, vol. 14, no. 10, p. 2668, 2021.

[104] R. Yasukuni, R. Gillibert, M. N. Triba, R. Grinyte, V. Pavlov, and M. Lamy de la Chapelle, "Quantitative analysis of sers spectra of mnsod over fluctuated aptamer signals using multivariate statistics," *Nanophotonics*, vol. 8, no. 9, pp. 1477–1483, 2019.

[105] R. Zhou, X. Chen, Y. Xia, M. Chen, Y. Zhang, Q. Li, D. Zhen, and S. Fang, "Research on the application of liquid-liquid extraction-gas chromatography-mass spectrometry (lle-gc-ms) and headspace-gas chromatography-ion mobility spectrometry (hs-gc-ims) in distinguishing the baiyunbian aged liquors," *International Journal of Food Engineering*, vol. 17, no. 2, pp. 83–96, 2020.

[106] E. Schievano, M. Sbrizza, V. Zuccato, L. Piana, and M. Tessari, "Nmr carbohydrate profile in tracing acacia honey authenticity," *Food chemistry*, vol. 309, p. 125788, 2020.

[107] D. Suhandy and M. Yulia, "The use of uv spectroscopy and simca for the authentication of indonesian honeys according to botanical, entomological and geographical origins," *Molecules*, vol. 26, no. 4, p. 915, 2021.

[108] H. Li, X. Ming, Z. Liu, L. Xu, D. Xu, L. Hu, H. Mo, and X. Zhou, "Accelerating vinegar aging by combination of ultrasonic and magnetic field assistance," *Ultrasonics Sonochemistry*, vol. 78, p. 105708, 2021.

[109] V. Z. Petukhova, A. N. Young, J. Wang, M. Wang, A. Ladanyi, R. Kothari, J. E. Burdette, and L. M. Sanchez, "Whole cell maldi fingerprinting is a robust tool for differential profiling of two-component mammalian cell mixtures," *Journal of the American Society for Mass Spectrometry*, vol. 30, no. 2, pp. 344–354, 2018.

[110] C. Llorente-Barroso, M. Sánchez-Valle, and M. Viñarás-Abad, "The role of the internet in later life autonomy: Silver surfers in spain," *Humanities and Social Sciences Communications*, vol. 10, no. 1, pp. 1–20, 2023.

[111] L. H. Nurani, A. Rohman, A. Windarsih, A. Guntarti, F. D. O. Riswanto, E. Lukitaningsih, N. A. Fadzillah, and M. Rafi, "Metabolite fingerprinting using 1h-nmr spectroscopy and chemometrics for classification of three curcuma species from different origins," *Molecules*, vol. 26, no. 24, p. 7626, 2021.

[112] M. Sedlmair and M. Aupetit, "Data-driven evaluation of visual quality measures," in *Computer graphics forum*, vol. 34, no. 3. Wiley Online Library, 2015, pp. 201–210.

[113] M. Kosicki, F. Allen, F. Steward, K. Tomberg, Y. Pan, and A. Bradley, "Cas9-induced large deletions and small indels are controlled in a convergent fashion," *Nature communications*, vol. 13, no. 1, p. 3422, 2022.

[114] T. Higa, Y. Okita, A. Matsumoto, S. Nakayama, T. Oka, O. Sugahara, D. Koga, S. Takeishi, H. Nakatsumi, N. Hosen *et al.*, "Spatiotemporal reprogramming of differentiated cells underlies regeneration and neoplasia in the intestinal epithelium," *Nature communications*, vol. 13, no. 1, p. 1500, 2022.

[115] H. Wang, D. Wang, B. Luo, D. Wang, H. Jia, P. Peng, Q. Shang, J. Mao, C. Gao, Y. Peng *et al.*, "Decoding the annulus fibrosus cell atlas by scrna-seq to develop an inducible composite hydrogel: A novel strategy for disc reconstruction," *Bioactive Materials*, vol. 14, pp. 350–363, 2022.

[116] Y. Song, R. Berger, A. Yosipof, and B. R. Barnes, "Mining and investigating the factors influencing crowdfunding success," *Technological Forecasting and Social Change*, vol. 148, p. 119723, Nov. 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S004016251832033X

[117] Y. Alibert, "New metric to quantify the similarity between planetary systems: application to dimensionality reduction using T-SNE," *Astronomy & Astrophysics*, vol. 624, p. A45, Apr. 2019. [Online]. Available: https://www.aanda.org/10.1051/0004-6361/201834592

[118] J. K. Ocasio, B. Babcock, D. Malawsky, S. J. Weir, L. Loo, J. M. Simon, M. J. Zylka, D. Hwang, T. Dismuke, M. Sokolsky, E. P. Rosen, R. Vibhakar, J. Zhang, O. Saulnier, M. Vladoiu, I. El-Hamamy, L. D. Stein, M. D. Taylor, K. S. Smith, P. A. Northcott, A. Colaneri, K. Wilhelmsen, and T. R. Gershon, "scRNA-seq in medulloblastoma shows cellular heterogeneity and lineage expansion support resistance to SHH inhibitor therapy," *Nature Communications*, vol. 10, no. 1, p. 5829, Dec. 2019. [Online]. Available: https://www.nature.com/articles/s41467-019-13657-6

[119] M. Aupetit, "Sanity check for class-coloring-based evaluation of dimension reduction techniques," in *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization*, ser. BELIV '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 134–141. [Online]. Available: https://doi.org/10.1145/2669557.2669578

[120] R. Faust, D. Glickenstein, and C. Scheidegger, "Dimreader: Axis lines that explain non-linear projections," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 481–490, 2019.

[121] H. Jeon, M. Aupetit, S. Lee, H.-K. Ko, Y. Kim, and J. Seo, "Distortion-aware brushing for interactive cluster analysis in multidimensional projections," *arXiv preprint arXiv:2201.06379*, 2022.

[122] M. Aupetit, N. Heulot, and J.-D. Fekete, "A multidimensional brush for scatterplot data analytics," in *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2014, pp. 221–222.

[123] C. Fu, Y. Zhang, D. Cai, and X. Ren, "Atsne: Efficient and robust visualization on gpu through hierarchical optimization," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 176–186.

[124] H. Jeon, H.-K. Ko, S. Lee, J. Jo, and J. Seo, "Uniform manifold approximation with two-phase optimization," in *2022 IEEE Visualization and Visual Analytics (VIS)*, 2022, pp. 80–84.

[125] M. Moor, M. Horn, B. Rieck, and K. Borgwardt, "Topological autoencoders," in *International conference on machine learning*. PMLR, 2020, pp. 7045–7054.

[126] N. Kriegeskorte, W. K. Simmons, P. S. F. Bellgowan, and C. I. Baker, "Circular analysis in systems neuroscience: the dangers of double dipping," *Nature Neuroscience*, vol. 12, no. 5, pp. 535–540, May 2009, number: 5 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/nn.2303

[127] D. Lähnemann *et al.*, "Eleven grand challenges in single-cell data science," *Genome Biology*, vol. 21, no. 1, p. 31, 2020. [Online]. Available: https://doi.org/10.1186/s13059-020-1926-6

[128] L. L. Gao, J. Bien, and D. Witten, "Selective Inference for Hierarchical Clustering," *Journal of the American Statistical Association*, pp. 1–11, Oct. 2022. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/01621459.2022.2116331

[129] Y. T. Chen and D. Witten, "Selective inference for k-means clustering," *J. Mach. Learn. Res.*, vol. 24, pp. 152:1–152:41, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:247779201

[130] A. Neufeld, L. L. Gao, and D. Witten, "Tree-values: Selective inference for regression trees," *J. Mach. Learn. Res.*, vol. 23, pp. 305:1–305:43, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:235436113

[131] D. Song, Q. Wang, G. Yan, T. Liu, T. Sun, and J. J. Li, "scDesign3 generates realistic in silico data for multimodal single-cell and spatial omics," *Nature Biotechnology*, pp. 1–6, May 2023, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41587-023-01772-1

[132] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[133] A. Narayan, B. Berger, and H. Cho, "Assessing single-cell transcriptomic variability through density-preserving data visualization," *Nature biotechnology*, vol. 39, pp. 765 – 774, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:231641412

[134] Z. Huang, D. Witschard, K. Kucher, and A. Kerren, "Va+ embeddings star: A state-of-the-art report on the use of embeddings in visual analytics," in *COMPUTER GRAPHICS Forum*, vol. 42, no. 3, 2023.

[135] M. Gleicher, "Explainers: Expert explorations with crafted projections," *IEEE transactions on visualization and computer graphics*, vol. 19, no. 12, pp. 2042–2051, 2013.

[136] B. C. Kwon, H. Kim, E. Wall, J. Choo, H. Park, and A. Endert, "Axisketcher: Interactive nonlinear axis mapping of visualizations through user drawings," *IEEE transactions on visualization and computer graphics*, vol. 23, no. 1, pp. 221–230, 2016.

[137] B. C. Kwon, B. Eysenbach, J. Verma, K. Ng, C. De Filippi, W. F. Stewart, and A. Perer, "Clustervision: Visual supervision

of unsupervised clustering," *IEEE transactions on visualization and computer graphics*, vol. 24, no. 1, pp. 142–151, 2017.
[138] D. Atzberger, T. Cech, W. Scheibel, J. Döllner, M. Behrisch, and T. Schreck, "A large-scale sensitivity analysis on latent embeddings and dimensionality reductions for text spatializations," *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–11, 2024.
[139] A. Bibal, R. Marion, R. von Sachs, and B. Frénay, "Biot: Explaining multidimensional nonlinear mds embeddings using the best interpretable orthogonal transformation," *Neurocomputing*, vol. 453, pp. 109–118, 2021.
[140] P. Lambert, R. Marion, J. Albert, E. Jean, S. Corbugy, and C. De Bodt, "Globally local and fast explanations of t-sne-like nonlinear embeddings," in *2022 International Conference on Information and Knowledge Management Workshops, CIKM-WS 2022*. CEUR Workshop Proceedings, 2022.
[141] E.-a. D. Amir, K. L. Davis, M. D. Tadmor, E. F. Simonds, J. H. Levine, S. C. Bendall, D. K. Shenfeld, S. Krishnaswamy, G. P. Nolan, and D. Pe'er, "viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia," *Nature biotechnology*, vol. 31, no. 6, pp. 545–552, 2013.
[142] D. Moritz, C. Wang, G. L. Nelson, H. Lin, A. M. Smith, B. Howe, and J. Heer, "Formalizing visualization design knowledge as constraints: Actionable and extensible models in draco," *IEEE transactions on visualization and computer graphics*, vol. 25, no. 1, pp. 438–448, 2018.

**Dylan Cashman** is an assistant professor in the Michtom School of Computer Science at Brandeis University in Waltham, MA. His research interests include the development and evaluation of visual affordances that improve usability of artificial intelligence models and data science processes. Dylan received a Ph.D in Computer Science from Tufts University.

**Mark S. Keller** is a student in the Bioinformatics and Integrative Genomics PhD Program at Harvard Medical School in Boston, MA. He holds a Bachelor of Science in Computer Science from University of Maryland, College Park. His research interests include the development of interactive visualization tools for high-dimensional single-cell omics data, including for transcriptomics and chromatin accessibility experiments.

**Hyeon Jeon** is a Ph.D student at Seoul National University, Seoul, Korea. He is currently working on developing new visualizations and machine learning techniques that support reliable data analysis. Before starting his Ph.D. program, he received a B.S. degree in Computer Science and Engineering from the Pohang University of Science and Technology (POSTECH), Pohang, Korea.

**Bum Chul Kwon** is a researcher at IBM Research. His research area includes visual analytics, data visualization, human-computer interaction, healthcare, and machine learning. His primary research interest includes the development of interactive visualizations to enhance users' abilities to derive knowledge from biomedical data using interactive visualization systems. He received his Ph.D in Data Visualization from Purdue University.

**Qianwen Wang** is an Assistant Professor in the department of computer science and engineering at the University of Minnesota. Her research aims to enhance communication and collaboration between domain users and AI through interactive visualizations, particularly focusing on their applications in addressing biomedical challenges. She received her Ph.D in Electronic and Computer Engineering from Hong Kong University of Science and Technology.