Dimensionality Reduction in the Wild: Gaps and Guidance

Michael Sedlmair, *Member, IEEE*, Matthew Brehmer, Stephen Ingram, and Tamara Munzner, *Member, IEEE*

Abstract— Despite an abundance of technical literature on dimension reduction (DR), our understanding of how real data analysts are using DR techniques and what problems they face remains largely incomplete. In this paper, we contribute the first systematic and broad analysis of DR usage by a sample of real data analysts, along with their needs and problems. We present the results of a two-year qualitative research endeavor, in which we iteratively collected and analyzed a rich corpus of data in the spirit of grounded theory. We interviewed 24 data analysts from different domains and surveyed papers depicting applications of DR. The result is a descriptive taxonomy of DR usage, and concrete real-world usage examples summarized in terms of this taxonomy. We also identify seven gaps where user DR needs are unfulfilled by currently available techniques, and three mismatches where the users do not need offered techniques. At the heart of our taxonomy is a task classification that differentiates between abstract tasks related to point clusters and those related to dimensions. The taxonomy and usage examples are intended to provide a better descriptive understanding of real data analysts' practices and needs with regards to DR. The gaps are intended as prescriptive pointers to future research directions, with the most important gaps being a lack of support for users without expertise in the mathematics of DR, and an absence of DR techniques for comparing explicit groups of dimensions or for relating non-linear embeddings to original dimensions.

Index Terms—Dimension reduction, high-dimensional data analysis, taxonomy, tasks, usage.

1 Introduction

Dimension reduction (DR) is the process of reducing a high-dimensional dataset to a lower-dimensional representation that retains most of the important structure. It has been an active research area throughout several decades and across many domains, from its origins in psychology [41, 51] through statistics [11] to machine learning [5, 40, 45, 24] and visualization [22, 25, 50].

The DR literature is heavily focused on mathematical and algorithmic descriptions of new techniques, including complexity analysis and benchmarks. However, considering DR from the perspective of problem-driven visualization, with a focus on the tasks and data of potential users [35], leads to many open questions: how can we evaluate whether a particular technique meets the needs of users? More fundamentally, what do DR practitioners actually do? And what are their tasks, considered at the abstract level?

A close reading of the technical literature does reveal some implicit assumptions regarding the concerns of a practitioner who seeks to use DR, but these vary significantly between and even within practitioners' domains. Synthesizing a coherent picture of what DR practitioners might care about is difficult because these assumptions have never been explicitly articulated. Our interest in finding an understanding of these needs grew as we moved from reading the previous work, to developing and validating DR techniques of our own [22, 48], and finally to building a DR system around our own assumptions regarding user needs [21].

Motivated by the lack of information on DR practices and needs of data analysts, we embarked on a two-year qualitative research project to gain insight into how DR was actually being used **in the wild** by a broad sample of data analysts spanning many application domains. We borrow the phrase "in the wild" from human-computer interaction (HCI), where it differentiates investigation of users and tasks in real-world settings from studying user behavior in artificial laboratory settings; the goal of in-the-wild studies is maximizing the realism of findings [31]. Our work has two primary goals: First, we sought a systematic understanding of how, when, and why analysts use DR. The scope of our investigation also included circumstances where analysts have high-dimensional data but opt not to use DR. Second, we wanted to identify gaps and mismatches between analysts' practices and the capabilities provided in currently available DR techniques, with the hope of directing future technique research.

To achieve these goals, we collected information from 24 data analysts, primarily via semi-structured interviews. We also surveyed existing literature in order to collect more examples of DR usage, and selected five papers with interesting descriptions of how DR algorithms were used to solve a domain problem. From this rich corpus of data, we extracted and distilled usage examples depicting DR as it is used among data analysts. Around these usage examples we created a descriptive DR-usage taxonomy characterizing real analysts and their processes. This taxonomy allows for a discussion to occur around which data analysis tasks are well served by current DR techniques, as well as those that are not. We used the taxonomy to compare our findings of analysts' practices and needs to the capabilities of the current state of the art in DR techniques. This comparison was possible given our own experience in this area [48, 22, 21, 38] and our familiarity with the DR technique literature. In so doing, we identified discrepancies between the needs of real users and the capabilities of current DR techniques. We classified these into two groups: *gaps* reflecting user DR needs that were unfulfilled by the suite of currently available techniques, and *mismatches* reflecting the opposite situation, where offered techniques are in excess of actual user needs.

During our survey of the existing literature, we also sought to identify implicit assumptions about user tasks and data, especially when we noted differences between domains. The most notable split was between the visualization and machine learning literatures, particularly in terms of benchmark dataset characteristics. Figure 1 shows two canonical datasets: the *Swiss roll* dataset heavily used in the machine learning literature [5, 40], and a cluster dataset of the sort heavily used in visualization [38]. The synthetic Swiss roll dataset shows off the capabilities of the *manifold following* non-linear DR techniques, such as Isomap [40] and Laplacian Eigenmaps [5], that assume that the high-dimensional data lies on a densely sampled manifold that should be "unrolled" to a lower-dimensional representation. The cluster dataset appears to completely violate the manifold assumption. We were skeptical that methods optimized for one would work well on the other, and wondered which dataset were more accurately reflected in analysts' practices and needs. Our DR usage taxonomy shed light on this question: both dataset types represent real-world scenarios. Some users are focused on dimensions, and others on clusters. For those concerned with dimensions, some focus on working with the existing dimensions, and others on synthesizing new ones. For those concerned with clusters of points, some focus on identifying implicit groups from DR layouts, and others on matching between these and explicit groups that are provided with the dataset as classes.

This paper addresses two groups of target readers. The first group

• *Michael Sedlmair, Matthew Brehmer, Stephen Ingram, and Tamara Munzner are with the University of British Columbia, E-mail: [msedl,brehmer,sfingram,tmm]@cs.ubc.ca.*

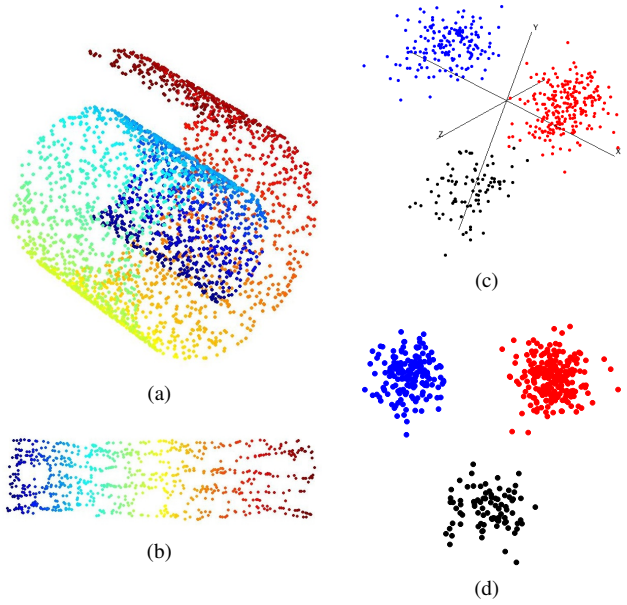

Fig. 1. **(a/b)** The Swiss Roll dataset used to illustrate and justify manifold DR techniques: a) Original 3D dataset; b) Isomap "unrolls" the manifold in 2D [40]. **(c/d)** A synthetic dataset with visually separable clusters that does not correspond to a uniform and dense sampling of a high-dimensional manifold: c) Original 3D dataset; d) PCA reduction to 2D.

is that of general visualization researchers with little expertise in DR, because it describes DR in a very accessible way from a usage perspective rather than from the usual highly technical perspective of previous work. The second group is that of researchers actively engaged in developing DR techniques, to provide a systematic lens on what users of their techniques are doing. In particular, we offer the summary of gaps and mismatches in Section 6 to channel future research directions.

In short, the contributions of this paper are a taxonomy of DR usage, a set of usage examples summarized according to it, and a set of gaps and mismatches between real usage and available techniques.

2 Related Work

We review related work on studies of high-dimensional data analysis in the wild, and on the use of taxonomies for characterizing DR and related analysis tasks.

High-dimensional Data Analysis in the Wild: A small body of work exists that studies high-dimensional data analysis as conducted by real data analysts in the wild. Most of these examples appeared in the format of visualization design studies [37] providing a detailed description of real users and tasks, albeit focusing on a single domain problem. One example is a recent design study by Pretorius et al. [34], which describes DR for parameter optimization in the domain of image analysis. Another is the Vismon design study [8], which supported sensitivity analysis of high-dimensional data stemming from fish population models; it was informed by preliminary results of the study presented here.

Others have raised the issue of user guidance in DR: In our own previous work we presented DimStiller [21], a system for providing user guidance via DR workflows. It addresses the conjectured needs of "middle-ground users": visualization generalists and application domain experts who do not have a deep understanding of DR mathematics. Cunningham [13] also offers guidance for DR usage, with a particular focus on real-world situations where the number of dimensions in a dataset dwarfs the number of points.

No previous work has offered a cross-domain understanding of tasks, needs, and problems of high-dimensional data analysts.

Taxonomies: There is a great deal of previous work in classifying DR algorithms based on different distinctions, including feature selection and feature extraction [13, 24, 49], linear and non-linear [24], globally and locally operating techniques [17], or convexity and spectrality [46]. The corresponding DR algorithm component of our taxonomy is simpler than these technically-driven taxonomies, because we focus on concisely classifying DR from a usage point of view.

The use of the taxonomy as a comprehensive representation of theory has a rich tradition in information visualization. For high-dimensional data analysis, Bertini et al. conducted a systematic literature review of high-dimensional quality measures [7]. Our own recent work [38] provides a taxonomy of the characteristics of visual separation factors of high-dimensional datasets in scatterplot visualizations. That effort was also inspired by the preliminary results of this work.

All existing taxonomies in high-dimensional data analysis are largely a product of DR technique- or data-driven enquiry; in contrast, we aim to complement this body of work by providing a perspective that is primarily usage- and task-driven. In information visualization, there are several general examples of task taxonomies. Some are based on the authors' own experience in conjunction with a thorough consideration of the current state of the art [2, 39], and others on observations of user behavior in controlled settings [1].

Our work is the first descriptive taxonomy of high-dimensional data analysis and DR usage that is based on interviewing domain expert users about their tasks conducted in the wild.

3 Process

We followed an iterative data gathering and analysis approach in the spirit of *grounded theory* [12]. Grounded theory is a common approach in the HCI community; it has been used to identify best practices [19], to inform design [16], and to characterize user behavior [33]. These methods are recently beginning to gain ground in visualization as well [23, 42, 27]. A methodology in the spirit of grounded theory allowed us to better understand the practices, tasks, and context of a broad collection of high-dimensional data analysts across different domains. The key aspects that we embraced were the alternation between data collection and analysis, the identification and refinement of conceptual relationships grounded in the data, and the subsequent generation of theory and critical reflection.

Our process was further informed by *survey methodology* [43]. Survey methodology refers to studying a representative sample of individuals, not to individual data collection methods such as questionnaires or online survey tools. We selected this approach to address our initial hypotheses regarding a potential mismatch between the characteristics of benchmark datasets and those encountered by data analysts in the wild.

Since high-dimensional data analysis spans many different domains, we decided for a *cross-domain* approach, which distinguishes us from the afore mentioned examples that focus on an in-depth investigation of a single culture, domain, or work context. Cross-domain studies in the wild are still rare, and to the best of our knowledge, our work provides the first in visualization. Examples in HCI include the work of Dourish et al. on strategies and attitudes surrounding system security [15].

3.1 Participants

Overall, we interviewed 24 data analysts from a range of different domain backgrounds. We distinguish between 19 *primary* interviewees with practical DR usage examples and 5 *secondary* interviewees, colleagues or supervisors of the primary interviewees who worked on the same problem and additionally attended group interviews. Interviewees were chosen via convenience and snowball sampling [14]. All had experience conducting analysis tasks with high-dimensional data and were highly trained in their domain.

3.2 Data Collection

Our primary data collection method for studying high-dimensional data analysts was semi-structured interviews. We used a set of closed- and open-ended questions relating to high-dimensional data and DR, which evolved over the 2-year span of data collection and analysis.

We conducted 1-2 interviews per data analyst, 22 in total, ranging in duration from 30 minutes to 4 hours. Most interviews were individual interviews (18) while others were group interviews with multiple data analysts (4). Some interviews were conducted remotely via telephone or Skype (8: 5 voice only, 3 voice and video). One to three interviewers guided the sessions and took detailed written notes for the purpose of later analysis. In the case where only one interviewer was present, audio recording was used (with interviewee consent). In addition to interview notes and recordings, we also gathered documents from analysts, which included their published papers, unpublished manuscripts, theses, datasets, screenshots of visualizations of their data, and their email correspondence with us.

To broaden the corpus of data, we surveyed many papers involving DR, including both those proposing DR techniques and those featuring the application of DR for solving a particular domain problem.

3.3 Analysis

Using a grounded approach [12], we concurrently performed data collection and analysis, culminating in the representation of our findings in the DR usage taxonomy described in Section 4.

Coding: Interview notes and other materials collected from interviewees were subject to *open coding*, a process for identifying themes and concepts. This was followed by *axial coding*, the grouping of codes with conceptual relationships, guided by our background knowledge and insights accumulated from previous data analysis.

Usage Examples: This stage involved the creation of concise *summary reports* and the extraction of one or more *usage examples* for each primary interviewee. While most of the interviews resulted in one usage example, we also had several interviews in which data analysts described multiple yet fundamentally different analysis tasks or situations in which they used or wanted to use DR; these became multiple usage examples, allowing for their concise description in our taxonomy. In the case of one interviewee, we did not have enough information to derive a concise usage example, so their data was excluded from subsequent analysis. This process resulted in 22 usage examples from 18 primary interviewees. The structure of the summary reports and their contained usage examples reflected the results of our axial coding. These codes reflected the interviewee's domain background and expertise, high-level research goals, characteristics and assumptions regarding their data, characteristics of their data analysis tasks, the means by which they perform these tasks, and the problems they encounter while performing them. Both codes and reports were iteratively revised and critiqued as data was collected.

From our literature survey, we derived 5 usage examples with interesting descriptions of high-dimensional data analysis and dimensionality reduction usage [10, 11, 30, 36, 40]. Many others were excluded as their DR usage examples were either too similar to those previously identified, or they did not provide sufficient detail regarding DR usage.

Overall, this process led us to 27 usage examples: 22 from interviews, 5 from literature.

Taxonomy: By the end of the 2-year data collection and analysis period, it became apparent that we were reaching data saturation; no new axial codes were emerging. In other words, we could now describe each of our identified usage examples using the set of axial codes. The descriptive taxonomy is the result of *selective coding*, a process of arranging our axial codes, based on their importance and representativeness, into hierarchical relationships.

Judgement of Usage Examples: Finally, relying on our background in this area, we assessed each usage example with regards to whether the analyst was ultimately successful in achieving their analysis goals (*succeed*), successful but having experienced considerable difficulty or uncertainty (*struggle*), or unsuccessful (*fail*). The *struggle* and *fail* cases contributed to our identification of research gaps and mismatches discussed in Section 6.

4 Taxonomy

The taxonomy that we derived from our systematic data analysis is descriptive rather than prescriptive, in the same spirit as the taxonomy of Lam et al. on visualization evaluation [28]. It can be used to describe DR usage at the level of specific DR techniques and tasks, and also at a more general level, pertaining to high-level analysis goals. The usage examples in Section 5 are discussed at both of these levels.

Figure 2 presents the taxonomy. The root node is *high-dimensional data*, which splits into three top-level branches of *User*, *DR*, and *Task*. The hierarchical structure of the taxonomy is meant to reflect the structure of our data; it does not represent a decision tree, as the branches are not mutually exclusive. A single usage example can be represented by several nodes: an analyst may have multiple task interests, or use different DR algorithms to approach a single problem.

The major contribution of the taxonomy is the task branch, which abstracts data analysts' interests into a set of low-level tasks. The user and the DR branches also help in providing an understanding of our usage examples, but are less novel and thus are minor contributions.

4.1 User

We describe *domain* and *DR expertise* as important *user* factors. The domains captured by our usage examples include bioinformatics, machine learning, mathematics, computational chemistry, structural chemistry, computer vision, fisheries sciences, journalism, life sciences, marine and ocean sciences, HCI, search engine optimization, and statistics.

We differentiate between a *DR-Math expert*, one with a deep mathematical understanding of DR techniques, and one who is*DR-naïve*, without this technical understanding. DR-naïve individuals treat DR algorithms as black-box tools for acquiring insight into their data. Certainly, DR-naïve individuals are experts within their own domains of data analysis; some have strong mathematical backgrounds, but not in the specific mathematical niche that is DR. These users seek to apply DR algorithms to their high-dimensional data, but struggle with selecting the right techniques, and cannot always be sure of whether they should trust what they see in low-dimensional visualizations.

The distinction between expert and naïve in our taxonomy is binary, in order to be concise and provide a simple lens for considering expertise. However we acknowledge that in reality expert and naïve are points along an axis traversed by analysts as they learn about and use DR. We classified 13 interviewees as being closer to DR-naïve, 8 as being DR-Math experts, and 3 as having aspects of both, sitting somewhere between an expert and naïve. The high number of individuals classified as being DR-naïve is interesting considering that very little research focuses on these users; DimStiller is a notable exception [21].

4.2 DR: Dimension Reduction

At the top level of the DR component, we differentiate between usage examples that involve DR (*yes*), and those that do not (*no*). A key point is that only some high-dimensional data analysis tasks require DR. There are many visualization and analysis techniques, such as the use of parallel coordinates or scatterplot matrices, that can be used without reducing the number of dimensions in a dataset. Despite this broad set of options that do not require DR, some interviewees had attempted to use DR even in situations where it was perhaps not the most appropriate choice; one such example is covered in Section 5.

Purpose of DR: In cases where DR is used (*yes*), the first of two boxes is the *purpose of using DR*. We differentiate usage of DR for *data analysis*, versus for *algorithmic input*. For the data analysis case, we found that analysts often used DR techniques to reduce the dimensionality of the data to be able to visualize it, with 2D scatterplots (20 of 27 usage examples), 3D scatterplots (3 of 27) or scatterplot matrices (2 of 27). Some analysts used DR techniques for data analysis directly, rather than as a precursor to visualization. For instance, the MUSIC usage example is an example where the analyst used PCA to identify and name the principal components of her data, and to understand how the original dimensions contributed to these principal components.

A different purpose of DR is algorithmic input, a common case in machine learning. In this case, the goal is to reduce the dataset's dimensions in order to improve the performance of downstream algorithmic processing. This practice can improve an algorithm's computational efficiency by avoiding the curse of dimensionality [6], or

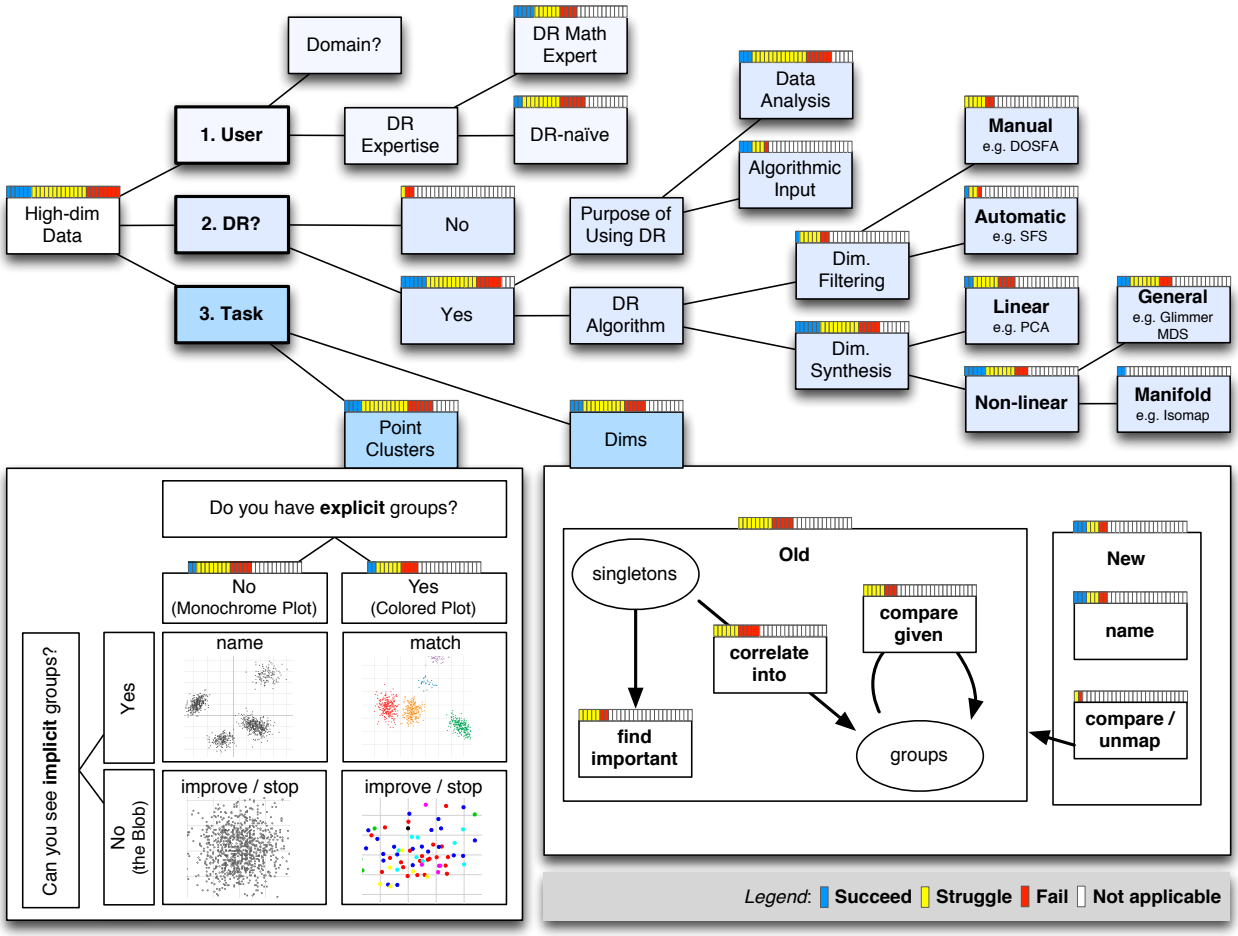

Fig. 2. DR usage taxonomy with three components (1) User, (2) DR, and (3) Task. The taxonomy is a classification rather than a decision tree allowing usage examples to constitute multiple paths through it. The color-coded bars atop many of the boxes indicate the overall number of usage examples in the box, colored according to our judgement of the DR usage outcome: blue for *success*, yellow for *struggle*, and red for *fail*.

make a predictive model more reliable, robust and accurate. We identified 7 usage examples where DR was used algorithmic input. DR for algorithmic input can be done without any further data analysis at all, which happened in 2 out of 7 usage examples. Examples of data analysis in the other 5 cases included inspecting the original high-dimensional data and/or the reduced low-dimensional data that is intended for algorithm input.

DR algorithm: The next DR usage box accounts for the choice of *DR algorithm*. Following the vocabulary of visualization community [32, 50], we divide algorithms into *dimensional filtering*, in which less interesting dimensions are filtered out, and *dimensional synthesis*, in which old dimensions are combined into new synthetic dimensions. In the machine learning literature, these categories are known as feature selection and feature extraction, respectively [49].

For dimensional filtering, we further differentiate between *manual* and *automatic* filtering techniques. Examples of manual filtering techniques user-defined quality metrics for finding interesting dimensions [25], or Yang et al.'s DOSFA approach of dimensional filtering, ordering and spacing [50]. An example of an automatic filtering technique is Sequential Forward Selection (SFS) [24] in which the "best" dimension is selected and others are added iteratively until no further improvement is made, relative to a threshold selected a priori.

In terms of dimensional synthesis, we further differentiate between *linear* and *non-linear* techniques. Linear techniques such as PCA [26] or classical MDS [41, 51] produce new dimensions that are a linear combination of the original old dimensions. However, many datasets have structure that can only be adequately revealed by non-linear techniques. These are further divided into *manifold* and *general* techniques. We define manifold techniques as the subset of non-linear techniques with the underlying assumption that the data lies on a densely sampled manifold, such as Isomap [40] and Laplacian Eigenmaps [5]. Although the machine learning literature often uses manifold synonymously with non-linear, we distinguish manifold techniques from general non-linear techniques; Section 6.4 discusses our rationale. An example of the general techniques is the distance scaling MDS approach of Glimmer [22].

4.3 Task

At the top level of the task component, we differentiate between two interests a high-dimensional data analyst might have: interests in *point clusters* and interests in *dimensions*. We are aware that from a purely mathematical perspective, points and dimensions could be transposed. However, most of our interviewees clearly considered these to have fundamentally different semantic meanings, so the taxonomy preserves this distinction.

Point Clusters: These tasks pertain the identification and verification of point clusters in the data (21 of 27 usage examples). In terms of visual data analysis, clustering is strongly tied to visualizing dimensionally-reduced data in scatterplots: 19 of the 21 usage examples associated with an interest in clusters involved the use of scatterplots.

In our taxonomy we frame cluster analysis in terms of the intersection between two questions: do *explicit* groups exist?, and are *implicit* groups visible? An explicit group means that the dataset has an associated class structure; the classes are typically shown by color coding the points in a scatterplot. These classes might come directly with the

data, or be assigned using a clustering algorithm run by the analyst, or may even be the result of manual labeling. If explicit groups do not exist, then there are no class labels for the points and corresponding scatterplots are displayed in monochrome. By visible implicit groups, we refer to proximity relationships in the lower-dimensional layout of points, forming a distinguishable structure; that is, there are multiple separated blobs in the scatterplot. If no implicit groups are visible, then the entire scatterplot is a single undifferentiated blob. We found that the fundamental reason that people look at scatterplots of dimensionally reduced data is to see if the implicit groups match their mental model of the dataset.

These two questions can be answered with *yes* or *no*, resulting in the 2x2 outcome matrix (Figure 4, lower left). The four cells are:

Yes explicit, yes implicit (top-right): If explicit groups are available for color-coding the points and if visually separable implicit groups are visible, a typical task is to evaluate the match between explicit and implicit groups: that is, checking whether the colors match with the spatial structure in the layout. When they match, an analyst will often declare victory, as the spatial layout indicates a *true positive*: the classes in the explicit grouping are in fact trustworthy. When they do not match, an analyst may attempt to improve the explicit grouping. For instance, this could involve adjusting the parameterization of a clustering algorithm.

No explicit, yes implicit (top-left): As we discuss in previous work [38], the visual separability of monochrome groups differs perceptually from the separability of color-coded groups. When no explicit groups are available, but implicit groups are visible, an analyst could assume that these groups represent meaningful clusters; the spatial layout is once again indicating a *true positive*. These implicit groups may then be subsequently named and labeled, resulting in explicit groups; the analyst has thus crossed into the previous cell.

Yes explicit, no implicit (bottom-right): In this cell, the color-coded points of the explicit groups are mixed together in the spatial layout, with no visually distinguishable class separation. An analyst may consider this result to be a *true negative*, meaning that the proposed class structure of the explicit group does not truly reflect the structure of the dataset. She may decide to discontinue this part of the analysis process at this point. Alternatively, she may also conjecture this visual result to be indicative of a *false negative*, meaning that clusters are just not visible with the current set of choices and parameters for clustering algorithm, DR technique, and visualization. In these cases, an analyst may iteratively select alternative choices and parameters until either separable class structure is visible (thus crossing into the upper right cell), or when they are confident that the results are indicating a *true negative*, that no structure is indeed the reality.

No explicit, no implicit (bottom-left): In this case, there is no visible structure in the monochrome spatial layout: an analyst sees a single large blob. This situation is similar to the previous one, in that an analyst may either infer a *true negative* and stop, or infer a *false negative* and continue trying other choices and parameter changes for DR technique and visual encoding. If these changes do not reveal any apparent structure, data analysts may try crossing into another cell by applying a clustering algorithm to create explicit groups.

These four situations should be interpreted as end points of axes that aid thinking about point clusters, rather than fixed states. It is common that data analysts start with few or no visible implicit groups, and then move upwards towards visible implicit groups by incrementally improving their clustering algorithm and/or DR parameter settings. It is also common to move from left to right, a sign that clusters are being iteratively detected and labeled.

Dimensions: These tasks pertain to the analysis of the dimensions of the dataset (18 of 27 usage examples), rather than clusters of points. We differentiate between interests in the original high-dimensional, or *old*, dimensions, and interests in the *new* dimensions that are the result of dimensional synthesis DR algorithms. Once again, the hierarchy is not a decision tree: a data analyst can be interested in both old and new dimensions within the same usage example.

We identified three common tasks within the old category. Some involve single dimensions, and others groups of dimensions. The *find important* task is about finding which among the old dimensions are important, according to some particular metric of interest. A common metric is the variance that a single dimension contributes to the overall variance (7 of 18 usage examples).

Analysts may also be interested in groups of old dimensions (13 of 18 usage examples). A task is to consolidate singleton dimensions into groups of *correlated* dimensions. If explicit groups of dimensions exist within the dataset, a task may be to *compare explicit groups*. A common case is when a dataset was produced by a predictive simulation model, where there exists a group of input dimensions and a group of output dimensions. This often calls for sensitivity analysis, which assesses whether small changes among the input dimensions yield small or large changes among the output dimensions [8]. While sensitivity analysis is typically conducted without DR, if there are a large number of dimensions in either group, then DR would be helpful; we discuss the challenges of this intersection in Section 6.2. In usage examples involving explicit groups of dimensions, the most common case was two groups of dimensions, either having an input/output or cause/effect relationship (5 of 8). The remaining 3 examples involved the more general case of *n* groups of dimensions.

Tasks involving an interest in new dimensions (8 of 18 usage examples) necessarily imply the usage of dimension synthesis techniques. One common task is to *name new* dimensions; that is, the analyst attempts to ascertain the semantic meaning of the proposed new dimensions. A common way to do this is to inspect the original high-dimensional data plotted within the context of the new low-dimensional layout in a scatterplot, wherein the analyst may be able to discern an interesting semantic relationship along the low-dimensional axes. For instance, image databases are easy to show as thumbnail images next to their corresponding points in a scatterplot: inspecting thumbnails of an image database of faces reduced to 3 dimensions via Isomap reveals three axes: up-down pose, left-right pose, and illumination [40]. Similarly, using MDS to lay out the similarities between Morse codes in a 2d scatterplot reveals that one axis is the number of clicks per letter, while the other axis is the number of long clicks relative to the number of short ones [11].

Another task associated with new dimensions is *comparing*, or more specifically, *unmapping* new dimensions to corresponding old dimensions. This task only occurs when an analyst has an interest in both old and new dimensions (2 of 8 usage examples associated with new dimensions) This task can be conducted in a straightforward fashion with linear techniques, by inspecting the extent to which any particular old dimension contributed to the synthesis of a new one. In PCA the relation between old and new dimensions is often referred to as the "loading" of the (new) principal components [26]. Most nonlinear techniques do not support unmapping, since the mapping that occurs between old and new dimensions is hard to interpret in a meaningful way.

Note that while determining correlation between old dimensions is a task in the taxonomy, we have not found any examples of analysts being interested in correlation between new dimensions. This is not surprising, as many DR techniques, such as PCA, produce new synthetic dimensions that have as little correlation between them as possible. In such cases, searching for correlation between new synthetic dimensions is unlikely to yield meaningful results.

5 Usage Examples

We now describe 8 concrete usage examples. 7 of which are based on interviews with 5 analysts; MoCap and Music each describe two usage examples. The remaining usage example was extracted from a DR usage paper [30]. We selected these so as to have good coverage of the taxonomy described in the preceding section. Table 1 summarizes all 27 usage examples, beginning with the 8 described in this section.

DR for Algorithmic Input (MoCap A & B)**:** The MoCap usage example pertains to a researcher in the domain of machine learning who is interested in building predictive motion capture models [44]. Using 45 carefully calibrated accelerometers, gyroscopes, and magnetometers, each attached to a part of the body of human subjects, he captures

fail / strug / happy	Usage Scenario	Paper/ Interview	1. User			2. DR							3. Task				Gaps						
			Domain	DR expertise		Purpose of Using DR		DR Algorithm					Point Clusters		Dims		User Gaps: DR-naïve			Task Gaps: Dimensions		Data Gaps: Assumptions	
				DR math expert	DR-naïve	Algo. Input	Data Analysis	Dim. Filtering: Manual	Dim. Filtering: Auto	Dim. Synthesis: Lin	Dim. Synthesis: Non-lin: Gen.	Dim. Synthesis: Non-lin Man.	Verify Explicit	Identify Implicit	Old	New	Concept.	Interpret.	Guidance	Groups	Unmapping	Categorical	Scalability
happy	MoCap A	Interview	machine learning		[X]	[X]			[X]	[X]			[X]						[X]				
fail	MoCap B						[X]	[X]	[X]	[X]			[X]					[X]	[X]				
fail	Music A	Interview	HCI		[X]		[X]			[X]			[X]	[X]	[X]		[X]	[X]	[X]			[X]	
strug	Music B						[X]			[X]					[X]	[X]	[X]	[X]	[X]		[X]	[X]	
strug	Concept	Interview	life sciences		[X]		[X]	[X]		[X]	[X]			[X]		[X]	[X]	[X]	[X]				[X]
fail	NPAlgo	Interview	machine learning	[X]		[X]		[X]							[X]	[X]				[X]			[X]
happy	BRDF	Paper [30]	graphics	[X]			[X]			[X]		[X]				[X]							[X]
fail	FishPop	Interview	fisheries sciences		[X]										[X]				[X]	[X]			
fail	SeqAln A						[X]			[X]	[X]		[X]	[X]				[X]	[X]				
happy	SeqAln B	Interview	bioinformatics	[X]	[X]	[X]					[X]			[X]					[X]				
fail	SeqAln C						[X]				[X]			[X]					[X]				[X]
fail	GamMdl	Interview	machine learning	[X]										[X]	[X]					[X]			[X]
fail	Search	Interview	search optimization		[X]		[X]			[X]	[X]		[X]	[X]	[X]	[X]	[X]		[X]		[X]	[X]	
strug	ProstCan	Interview	bioinformatics		[X]	[X]	[X]	[X]	[X]	[X]	[X]		[X]		[X]		[X]	[X]	[X]	[X]			[X]
strug	EpiGen	Interview	bioinformatics		[X]		[X]			[X]	[X]			[X]	[X]		[X]	[X]	[X]	[X]			
strug	StrucGen	Interview	bioinformatics		[X]								[X]	[X]	[X]		[X]		[X]	[X]		[X]	[X]
strug	FlockSim	Interview	mathematics	[X]			[X]	[X]						[X]	[X]			[X]		[X]			
strug	CompBio	Interview	comp. biology	[X]	[X]	[X]	[X]	[X]	[X]				[X]	[X]				[X]	[X]				[X]
strug	ChemRel	Interview	comp. chemistry	[X]			[X]				[X]		[X]		[X]			[X]		[X]		[X]	
strug	MedImg	Interview	comp. vision	[X]			[X]	[X]						[X]	[X]			[X]		[X]			
strug	TxtDocs	Interview	journalism	[X]	[X]		[X]				[X]		[X]	[X]			[X]	[X]	[X]				[X]
strug	BoatAct	Interview	marine & ocean sci.		[X]		[X]			[X]	[X]		[X]		[X]		[X]	[X]	[X]			[X]	
strug	Polymers	Interview	structural chemistry		[X]		[X]			[X]	[X]			[X]			[X]	[X]	[X]				
strug	Quadrup	Paper [36]	graphics	[X]	[X]	[X]	[X]			[X]						[X]		[X]					
happy	ArtShp	Paper [10]	graphics	[X]		[X]					[X]		[X]										
happy	Faces	Paper [40]	graphics	[X]			[X]					[X]				[X]							
happy	MorseCd	Paper [11]	statistics	[X]			[X]				[X]			[X]		[X]							
Total count				13	14	7	20	7	4	13	13	2	12	15	13	8	10	14	17	9	2	6	9
				57%	61%	26%	74%	26%	15%	48%	48%	7%	44%	56%	48%	30%	37%	52%	63%	33%	7%	22%	33%
			of 18 interviewees and 5 papers			of 27 usage examples																	

Table 1. 27 Usage examples we extracted from interviews and papers. Each line reflects a usage example summarized in terms of the taxonomy described in Section 4 (left side), and the gaps described in Section 6 (right side). The usage example at the top are further described in Section 5.

motions while walking, standing up, sitting down, lying down, kneeling, etc. From these sensors, a large number of time-varying derived variables (~25 per sensor) are recorded, resulting in datasets of ~10K points and ~1K dimensions. These datasets are subsequently used to train a motion classifier.

DR is used for two different purposes, *algorithmic input* (usage example MoCap A) and data analysis (MoCap B). In the former, *linear* PCA or SFS *automatic filtering* [24] are used to reduce the number of dimensions from ~1K to roughly 30, a number chosen by manually inspecting scree plots. The reasons for applying DR, typical for machine learning domain problems, are data compression, algorithmic efficiency, and avoiding the curse of dimensionality [6]. For the latter, the analyst is interested in verifying that the *clusters of explicit groups*, those corresponding to the ground truth of his motion capture recordings (the type of movement), do indeed exist as visibly distinct blobs within the layout of points in a scatterplot. To this end, he further reduces the data using the same DR techniques, PCA or SFS, to either 2 or 3 dimensions, then visualizes the result with color-coded 2D or 3D scatterplots. However, he often finds himself in the cell of *yes explicit, no implicit*: the color coding of the explicit groups does not unambiguously match the implicit groups. Despite this failed attempt to verify his explicit groups, he continues to use the low-dimensional reduced data for training the classifier and usually attains satisfying classification results. We conjecture that his result was a false negative.

PCA of Listening Histories (Music A & B)**:** The Music data analyst is an HCI graduate student interested in the listening behavior of digital music consumers. For this purpose, she gathered listening history data and demographic information from ~300 users of the last.fm music streaming service. The resulting dataset had 48 dimensions, which included continuous dimensions, such as the number of tracks streamed per day or per login session, as well as categorical dimensions, which included the user's gender and geographical location.

Her initial interest was to cluster users into groups with similar listening behavior (usage example Music A). She hypothesized several groups a priori, such as users who listen to the same music repetitively vs. users who listen to new music, or users who listen during the day vs. those who listen at night. She used PCA and scatterplots in hopes of both verifying these hypotheses and generating new ones, with both monochrome scatterplots as well as the application of k-means clustering to generate colored scatterplots. However, this approach did not result in meaningful results. Problems included uncertainty about what k to use for the number of clusters, and whether the undifferentiated blob from *no implicit groups* represented a true or false negative. We thus classify this case as *fail*, since her needs were not met.

Giving up on her original goal of identifying clusters, the data analyst changed her focus to *finding important* dimensions and consolidating dimensions by *correlate* them into groups (Music B). To do so, she again used PCA and examined the first 13 principal components, which accounted for 75% of the variance. For each of these 13 principal components, she related them back to the original *old* dimensions; by closely analyzing this *unmapping*, she identified proper *names* and meanings for the *new* synthetic 13 dimensions. These new dimensions were meant for guiding other researchers designing music recommender systems. In terms of DR, the goals were therefore to account for as much of the variance of the original data as possible with a small number of new dimensions, while maintaining an understandable semantic mapping between old and new dimensions. No visualization was involved in this second usage example [4].

MDS for Clustering Research Concepts (Concept)**:** The Concept data analyst is a Computer Science graduate student with a research focus on the visualization of social networks. In one of her projects, she is interested in visualizing the expertise of researchers in the life sciences domain. Her goal is to create a concept relationship visualization, to be used by researchers in the life science, providing them with an overview of the dataset via higher-level concept clusters, and allowing them to identify others working in areas related to their own. The dataset stems from a database of all researchers in life science domain in which each researcher is represented by a set of ranked research concepts. Overall there are 20,000 concepts, including terms such as "DNA", "cancer", or "North Carolina".

Her data analysis task was to identify point clusters of concepts. To do so, she computed a distance matrix of concepts based on their co-

occurrence in the researcher database. This matrix was then used as input to classical MDS to be displayed as a 2D scatterplot. As classical MDS did not scale, she manually filtered the dataset from ~20K to 400 concepts and used the 400 x 400 distance matrix as input. However, the scatterplot did not reveal any meaningful cluster structure, only an undifferentiated blob of points (*no explicit, no implicit groups*). She also tried Glimmer MDS, but that did not reveal visible cluster structure either. Eventually, she remained uncertain as to whether she was seeing a true negative or false negative; we thus classify her case as *struggle*.

Reduce Parameters of Algorithms (NPALGO)**:** This usage example involves a computer science professor with a strong mathematical DR background. He uses empirical methods to study algorithms. One particular project addresses the question of how to construct good prediction models for NP-hard algorithms, such as the traveling salesman problem [29]. He measures the time required to run an algorithm for a NP-hard problem across many different parameter settings, where the settings are regularly sampled from the available range.

This process results in a dataset with between ~100K and one million points. There are two *groups* of dimensions: ~100 features in the group of input dimensions, and the measured runtime as a single output dimension. The dataset is then used to train a predictive model for the algorithm at hand. Although the prediction model that is constructed in this fashion works well, humans have a hard time understanding this 100D feature space. The data analyst therefore would prefer to reduce the dataset to a lower dimensional set of 5-10 features, with a resulting model that retains nearly the same predictive power as the one that uses the entire feature space. Abstractly, this goal can be described as getting insight into *groups* of dimensions and to synthesize *new* dimensions without diminishing predictive power.

The data analyst knows that the input dimensions are highly correlated to each other, so he would like to use a dimensional synthesis technique. However, he has not succeeded in finding a technique that meets his needs, because the critical relationship of the input dimensions to the output dimension is not taken into account in the DR process. For example, a specific input dimension that contributes very little to the overall variance may nevertheless have a huge impact on the output run time; the converse is also possible. It is not useful to apply a DR technique solely on the input dimensions, because the crucial information about the intrinsic relation of the input dimensions to the output dimensions is not considered in the computation. Neither can he run synthetic DR on all 101 dimensions at once, because the distinction between input and output groups of dimensions is not maintained by any existing algorithm. That is, the importance of original input variables with regard to their influence on the output variable is not acknowledges by common DR metrics, such as maintaining variance as in PCA, or reducing stress as in MDS. He thus has no choice but to use *manual filtering* instead of synthetic DR techniques.

A Data-driven Reflectance Model (BRDF)**:** This *successful* usage example is extracted from Matusik et al.'s paper "A Data-Driven Reflectance Model" [30]. The researchers constructed a generative reflectance model for image synthesis, based on a large database of photographs of 104 physical objects made from different materials. Successive pictures of each material were taken as the light source was systematically moved with respect to the camera, resulting in a densely sampled set of measurements. Extensive resampling and post-processing led to a dataset with ~4M derived dimensions and 104 points representing the way that light reflects off the object. The goal was to drastically reduce this data to a set of *new* meaningful dimensions. This lower-dimensional representation could be used to construct a simple model that would allow for the generation of new materials with characteristics that were a blend of other physical materials photographed. After using PCA to reduce the data to 45 dimensions, Matusik et al. showed that the result model of this linear DR technique generated physically implausible reflectances, whereas the non-linear DR method of charting [9] resulted in a correct model with 15 dimensions to which they could assign meaningful names, such as "redness", "metallic-like" or "dustiness".

Sensitivity Analysis, not DR (FISHPOP)**:** Finally, we describe an example from a biologist with a strong statistics background who studies fish population models. Using these, she gives recommendations about balancing the risks of overfishing with commercial and private fishing interests. She compares and evaluates several different mathematical models that simulate the behavior of fish populations. All of these take a set of input parameters, such as carrying capacity and productivity, typically generated via regular sampling in the space of possible parameter configurations. This sampling pattern is clearly visible in the *fisheries* dataset example in our visual cluster separation taxonomy [38]. The output of these models is an indication of the probability that a fish population will die out [20].

This dataset is a canonical example of *explicit groups of old dimensions*. Her main concern is sensitivity analysis: checking whether small changes in input dimensions lead to small or large changes in output dimensions. Here, there is no need for DR. She was not using DR at the time of the initial interview. However, she did apply several DR techniques over the following months, but ultimately resolved that no off-the-shelf DR technique was appropriate for her dataset, thereby continuing with sensitivity analysis.

We include FISHPOP as an example of a situation where DR is not required despite a high dimensional dataset. In this case, sensitivity analysis was sufficient and there was no need to reduce dimensionality.

6 DISCREPANCIES: GAPS AND MISMATCHES

We derive seven gaps and three mismatches from the analysis of 27 usage examples. The gaps describe problems of users that are not sufficiently addressed by current DR literature. Table 1 illustrates the usage examples which fell into these gaps. The mismatches are situations where the capabilities of techniques proposed in the literature were not required to solve the particular problems of these users. Gaps and mismatches are also summarized in Table 2. We consider our usage examples to be an existence proof of these gaps; we do not make quantitative claims about their prevalence. We discuss them here given that we identified each of them at least in one usage example, and because their description may be useful for directing future research.

6.1 User: DR-naïve Gaps

More than half of our interviewees can be considered as being DR-naïve, suggesting that this type of user is not uncommon, exists in many domains, and should not be neglected by the research community. DR-naïve users have high-dimensional data, a potential need for DR, and extensive domain knowledge. However, they have markedly less technical knowledge about DR, as discussed in Section 4.1. We identified three DR-naïve gaps:

Conceptual: *What is DR doing?* — Most of the current technical DR literature requires a sophisticated grasp of the underlying mathematics in order for techniques to be used effectively. It explains DR solutions from an *implementation model* point of view; that is, with a strong focus on *how* the technical DR solution works. End users, especially those who we call DR-naïve, would benefit from simplified yet faithful *conceptual models* of these techniques, models which can easily be incorporated into a user's *mental model*. An analogous example from the HCI domain refers to withdrawing money from an ATM machine: the conceptual model presented to the users is rather simple (use debit card and PIN to get money), hiding the full complexity of the implementation model, which involves time-critical transaction processing. In DR, PCA is easy to understand in terms of finding principal components, and even MDS has an intuitive analogy of springs and forces between points. These are simplified yet faithful conceptual models which hide the full mathematical complexity. As a result, PCA and MDS are often the first choice of DR-naïve analysts. The conceptual gap is a hurdle that prohibits many potential users from selecting and applying DR, and can lead to misapplications of DR techniques.

Interpretation: *What do the results mean?* — Many of our DR-naïve interviewees struggled with gauging the effectiveness of a DR technique, such as when inspecting the visual layout of a reduced dataset

in a scatterplot. These users know enough about DR to select a particular technique, but not enough to fully interpret its results. Their uncertainty can be cast as concerns about potential false negatives and positives. One frequently recurring example among many of our interviewees, including some who we considered to be *DR-Math experts*, was how to interpret a scatterplot of a reduced dataset that does not contain visually separable clusters (*no implicit groups*). Such an instance could either indicate a true negative, that there is no cluster structure in the dataset, or as a false negative, an artifact of inappropriate choices of DR techniques and/or parameters. Some interviewees stated this concerns explicitly. However, a worse case is a user who unquestioningly assumes that no visible implicit groups are always indicative of a true negative.

Guidance: *What algorithm/parameterization to use?* — This gap follows from the previous one, and points to a lack of support for the user's questions regarding how to proceed. In the undifferentiated blob example, the problem of *true negative vs. false negative* leads them to ask: when should attempts to use other techniques be made? At what point should analysis stop? For DR-naïve users, these questions are hard to answer without guidance, often resulting in trial-and-error approaches. Guidance in these situations may be provided by matching data and task characteristics with the assumptions of the DR algorithms. The concise description of abstract tasks offered in this paper is a step towards such systematic guidance. Our previous work with DimStiller [21] was another step in this direction, however, there is much left to do.

6.2 Task: Dimension Gaps

We describe two gaps of unfulfilled DR needs arising from *dimension* tasks, as characterized in our taxonomy (Section 4.3).

Dependent groups: Many of our usage examples involved *comparing explicit groups of dimensions*. Some analysts, such as in the FISHPOP usage example, perform this comparison and have no need for DR at all, and sensitivity analysis alone may represent an adequate solution for them. However, many other analysts have an additional need to reduce the dimensions of their data. This combination does not inherently result in a problem. For instance, the FLOCKSIM analyst was able to meet her needs with DR in the form of filtering; the STRUCGEN analyst was satisfied by reducing each group of dimensions independently from one another, then comparing the reduced groups. However, in other cases, such as in the NPALGO example, the groups possessed an inherent and important dependency which is not accounted for by off-the-shelf DR algorithms. NPALGO is a specific instance of this gap, where there is a need for DR combined with a need for sensitivity analysis between input and output groups. Reducing the input (or output) dimensions without taking into consideration their inherent relation leads to meaningless results. Reducing both groups together also results in the loss of the critical distinction between the dimensional groups. In general, this problem may occur in situations with n groups of dimensions and an arbitrary arrangement of relations between these groups. No existing DR technique (to our knowledge, or that of our interviewees) handled dependent dimensional groups appropriately.

Non-linear unmapping: In Section 4.3, we mentioned the task of relating the new synthetic dimensions back to old ones, or *unmapping the new dimensions*. When interviewees performed this task, they used PCA because there was no support for unmapping with any non-linear technique. However, in many cases their data was not linear, leaving their analysis needs poorly met. For example, the MUSIC B data analyst needed intuition about the mapping between old and new, and would have benefited from a non-linear synthesis technique. The SEARCH analyst faced the same problem, where linear methods fell short when synthesizing and unmapping new dimensions.

While we are well aware that unmapping non-linear combinations is a difficult undertaking, even a partial solution would improve the state of the art. Interactive visualization may well be a fruitful avenue to pursue, for helping users explore a complex non-linear dimensional mapping space.

6.3 Data: Assumption Gaps

Finally, assumptions about data characteristics that make certain DR synthesis techniques unusable were apparent in several usage examples. Although there were many instances of noisy, sparse, or incomplete data, we do not belabor these well-known issues here. We identified two further stumbling blocks with respect to data characteristics:

Categorical dimensions: Categorical data is poorly handled by most existing DR algorithms, which are designed to work efficiently with continuous, numerical data. Workarounds, such as mapping categorical data with k values into k new binary dimensions, are stopgap measures. Correspondence analysis is designed for categorical data, but does not solve analysts' problems, as it lays out the dimensions, not the individual points, which is particularly problematic if analysts look for *pointclusters* [18].

Scalability: Many real-world datasets are far too large for commonly used DR algorithms, causing inconvenient breakdowns for many analysts. A large number of points or dimensions challenges most algorithms, however the scalability problem is particularly severe for distance matrices, which grow quadratically with the number of points. Extreme ratios between dimensions and points, when there are many more dimensions than points, also pose scalability challenges [13, 47].

6.4 Mismatches

We identified three *mismatches* that should evoke caution on the part of both DR users and DR technique developers. As opposed to *gaps* where users have unmet needs, these mismatches represent the opposite situation, where offered techniques are in excess of user needs.

DR: The first mismatch to note is that some users with high-dimensional data have no need for DR. While this statement might sound obvious to some, DR-interested researchers might be quick to assume a need for DR where it does not in fact exist; we have certainly noticed this tendency in our prior research. The *DR→No* box in the taxonomy serves as an explicit reminder of this fact. The FISHPOP analyst's goals were met with sensitivity analysis, gauging whether small changes in the input dimensions result in small or large changes in the output dimensions; there was no need for DR per se.

Vis: The second mismatch is that some users do not need visualization. The *DR Purpose* box relates to this question: when the purpose of using DR is solely for algorithmic input, there is no need for visualization. More subtly, even if the user has a goal of data analysis, there still may not be a need for visualization. NPALGO and MUSIC B usage examples illustrate non-visual data analysis.

Manifold DR Techniques: Our in-the-wild investigation led us to a better understanding of our original hypothesis and we indeed identified mismatches between user needs and the assumptions of manifold following techniques. We note that none of the interviewees chose to apply manifold techniques; both usage examples where these techniques were applied were found in papers.

Densely sampled: Manifold techniques assume densely sampled data, however, we encountered many situations in which it was not clear to data analysts, in particular DR-naïves, whether their datasets meet the assumptions of a densely sampled manifold or not. We have identified two rules of thumb to understand when a dataset is likely to qualify as a densely sampled manifold. First, all dimensions should be numerical and continuous, which negates any datasets containing categorical dimensions. Second, the dataset should be generated by a process that has the characteristics of continuous sampling. For example, in BRDF pictures were taken of materials where the light source moved in small and regular intervals and in MOCAP the sensors measure body part motion over small time intervals. Real-world measurements are a common case for manifolds, but manifolds may exist elsewhere: with NPALG, the regularly changing values reflected algorithm parameters. We note that these aspects seem not to be met by many real-world datasets. Consider, for instance, the survey data in the MUSIC example, or the research CONCEPTS dataset.

Single vs. multiple manifolds: In addition to densely sampled data, many of the well-known manifold techniques such as Isomap [40] and

User Gaps	**Conceptual**	Missing conceptual models for many DR algorithms.
	Interpretation	Difficulties understanding and trusting the visual layout of reduced datasets, in terms of true/false positives/negatives.
	Guidance	Difficulties in selecting which algorithms to use, and realizing when one has reached a stopping condition.
Task Gaps	**Dependent groups**	Need for DR that takes into account dependencies between explicit groups of dimensions.
	Unmapping	Need for non-linear DR that supports relating synthetic new dimensions to original / old dimensions.
Data Gaps	**Categorical**	Assumption that dimensions are continuous or ordered, rather than categorical.
	Scalability	Assumption that the number of points, or dimensions, or ratio between them is constrained.
Mismatches	**DR**	User has high-dimensional data but does not need DR.
	Vis	User has no data analysis needs, or can perform analysis without visualization.
	Manifold	User does not need to name new synthetic dimensions, or the dataset was not generated via dense sampling.

Table 2. Summary of gaps and mismatches. *Gaps* are user needs that are not sufficiently addressed by current algorithms and tools. *Mismatches* describe situations where offered algorithms/techniques were not needed by the analysts.

Laplacian Eigenmaps [5] have a further assumption, that the data resides on a **single** manifold. Formally, the data distribution along all continuous dimensions needs to be homogeneous rather than heterogeneous or clumpy; this third rule of thumb indicates single manifolds. The Swiss Roll dataset is the canonical example for a single manifold, with the goal being to carefully unroll the single manifold. This assumption matches only with the needs of users who care about the *name new dimensions* task. In particular, such techniques are less likely to meet the users' need in cases where they care about *point clusters*. Point clusters would be due to either a non-uniform distribution of samples on a single manifold, or that the clusters represent samples taken from **multiple** different manifolds. MoCap is an example from our study where a multiple manifold structure is likely, with one densely sampled manifold per movement type. The machine learning community has noted the instability of these older single manifold algorithms [3, 46] and newer techniques such as t-SNE [45] have been proposed for following multiple manifolds and better supporting clustering tasks. However, the question of whether a dataset reflects the result of dense sampling along continuous dimensions remains.

7 Discussion

Our taxonomy provides the first systematic analysis and description of DR usage *in the wild*, adding a new usage-based perspective to the large body of technique-driven DR literature. Following the nature of descriptive taxonomies [1, 28], our main contribution is not a radically new perspective on a problem; rather, our taxonomy provides a structured description, a holistic lens, and a concise vocabulary for talking and thinking about high-dimensional data analysis with respect to DR. In particular, we introduced a differentiation between tasks involving *point clusters* and those involving *dimensions*, the idea of *explicit* and *implicit* groups of points as a way to think about visual clustering, and of *old* and *new dimensions* for framing dimension-related tasks. We also emphasize *true/false positives/negatives* as a way to consider whether the the visual representation of a dimensionally reduced dataset is faithful to its true structure. We envision several ways in which the descriptive understanding we offer could prove helpful to researchers and practitioners of visualization and DR. Those conducting design studies [37] involving high-dimensional data can use the taxonomy to guide the categorization and abstraction of user problems and tasks, including the decision of whether the application of DR is an appropriate choice.

Researchers presenting new DR techniques can use the taxonomy to concisely state assumptions about which tasks are supported, rather than leaving this description implicit in a way that places a burden on the reader. We also hope to stimulate and inspire future research directions in DR techniques through the seven gaps and three mismatches identified here. The most critical open areas are the DR-naïve gaps, and dimension gaps of group-aware DR and support for non-linear unmapping of new dimensions. The classification of 27 usage examples in Figure 2 reveals that while nearly half of DR usage for algorithmic input led to successful outcomes (3 out of 7), most users who used DR for data analysis struggled or failed in their attempt (of 22, 13 struggled and 6 failed). These numbers clearly underline the need for further usage-centered DR development and research.

Limitations and Future Work: This work is not without its limitations. Our own background is in visualization, and our particular interests for this project were in high-dimensional data analysis with a strong focus on, yet not an all-encompassing knowledge of, DR. As is inevitable with qualitative research [12], this lens influenced which participants we invited to our study, how we questioned them, which papers we selected and read, how we coded the data, and eventually also what we decided to report in the paper. This is, we do not claim a complete coverage of all possible and relevant DR and high-dimensional usage patterns. High-dimensional data analysis is a diverse and rich area full of many other interesting real-world aspects. While our data gathering and analysis process has focused on the usage of DR, we naturally encountered other aspects including such as data cleaning, outlier detection, model optimization, and subspace clustering. We see our work as a first step towards a better and more systematic understanding of data analysis "in the wild" and hope that others will build upon our work and methodological approach, extend the taxonomy with new findings, and broaden our understanding with new and different perspectives.

Lessons Learned: Interviewing a broad set of users across many domains proved to be a difficult adjustment for us, compared to the familiar design study approach where the goal is to attain a deep understanding of a single domain problem. Many of the data analysts we talked to worked on difficult questions in complex domains, such as computational chemistry. To understand data analysts' needs and problems a certain amount of domain understanding is imperative. On the other hand, attaining an in-depth understanding for all 19 domain problems would have resulted in years or even decades of work. Finding the balance between those two extremes was a challenging process. Our approach was to engage in task and data abstraction on the fly during the interviews, in order to get immediate feedback without iterative refinement through multiple interviews.

8 Conclusions

In this paper, we presented a descriptive taxonomy of DR usage *in the wild*, grounded in the analysis of usage examples from many different domains. The taxonomy provides an abstract understanding of what practices data analysts with high-dimensional data and DR needs are engaging in. Together with the usage examples we described, it serves as a new usage-centered lens on DR, complementing the rich corpus of technical DR literature. We also specifically outlined research gaps and mismatches between real usage and available techniques which became apparent to us during the course of our study. We hope that this usage-centered approach to high-dimensional data analysis and dimensionality reduction encourages others to continue in this methodological spirit.

Acknowledgements

We thank the data analysts who participated in this study for their time and energy. We thank NSERC for funding this project, Jessica Dawson, Miriah Meyer, Torsten Möller, Hamidreza Younesy, and Steven Bergner, for feedback and/or assisting with the interviews.

REFERENCES

[1] R. Amar, J. Eagan, and J. T. Stasko. Low-level components of analytic activity in information visualization. In *Proc. IEEE Symp. Information Visualization (InfoVis)*, pages 111–117, 2005.

[2] R. Amar and J. T. Stasko. A knowledge task-based framework for design and evaluation of information visualizations. In *Proc. IEEE Symp. Information Visualization (InfoVis)*, pages 143–150, 2004.

[3] M. Balasubramanian and E. Schwartz. The Isomap algorithm and topological stability. *Science*, 295:7, 2002.

[4] D. Baur, J. Büttgen, and A. Butz. Listening factors: A large-scale principal components analysis of long-term music listening histories. In *Proc. Conf. Human Factors in Computing Systems (CHI)*. ACM, in press.

[5] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pages 585–591. MIT Press, 2001.

[6] R. Bellman. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, 1961.

[7] E. Bertini, A. Tatu, and D. A. Keim. Quality metrics in high-dimensional data visualization: an overview and systematization. *IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis)*, 17(12):2203–2212, 2011.

[8] M. Booshehrian, T. Möller, R. M. Peterman, and T. Munzner. Vismon: Facilitating analysis of trade-offs, uncertainty, and sensitivity in fisheries management decision making. *Computer Graphics Forum (Proc. EuroVis)*, in press.

[9] M. Brand. Charting a manifold. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pages 961–968. MIT Press, 2003.

[10] A. Bronstein, M. Bronstein, A. Bruckstein, and R. Kimmel. Matching two-dimensional articulated shapes using generalized multidimensional scaling. In *Proc. Conf. Articulated Motion and Deformable Objects (AMDO)*, pages 48–57. Springer, 2006.

[11] A. Buja and D. Swayne. Visualization methodology for multidimensional scaling. *Journ. Classification*, 19:7–43, 2002.

[12] K. Charmaz. *Constructing Grounded Theory. A Practical Guide Through Qualitative Analysis*. Sage Publications, 2006.

[13] P. Cunningham. Dimension reduction. In *Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval*, pages 91–112. Springer, 2008.

[14] A. Dix, J. Finlay, G. D. Abowd, and R. Beale. *Human-Computer Interaction*. Pearson Education Limited, 3rd edition, 2004.

[15] P. Dourish, R. E. Grinter, J. D. de la Flor, and M. Joseph. Security in the wild: user strategies for managing security as an everyday, practical problem. *Personal and Ubiquitous Computing*, 8(6):391–401, 2004.

[16] L. Findlater, J. McGrenere, and D. Modjeska. Evaluation of a role-based approach for customizing a complex development environment. In *Proc. Conf. Human Factors in Computing Systems (CHI)*, pages 1267–1270. ACM, 2008.

[17] S. France and J. Carroll. Two-way multidimensional scaling: A review. *IEEE Trans. Systems, Man, and Cybernetics*, 41(5):644 –661, 2011.

[18] M. Friendly. *Visualizing Categorical Data*. SAS Press, 2000.

[19] D. Furniss and A. Blandford. Usability work in professional website design: Insights from practitioners' perspectives. In *Maturing Usability: Quality in Software, Interaction and Value*, pages 144–167. Springer-London, 2008.

[20] C. Holt and M. Bradford. Evaluating benchmarks of population status for Pacific salmon. *N. American Journ. Fisheries Management*, 31(2):363–378, 2011.

[21] S. Ingram, T. Munzner, V. Irvine, M. Tory, S. Bergner, and T. Möller. DimStiller: Workflows for dimensional analysis and reduction. In *Proc. IEEE Symp. Visual Analytics Science and Technology (VAST)*, pages 3–10, 2010.

[22] S. Ingram, T. Munzner, and M. Olano. Glimmer: Multilevel MDS on the GPU. *IEEE Trans. Visualization and Computer Graphics*, 15(2):249–261, 2009.

[23] P. Isenberg, T. Zuk, C. Collins, and S. Carpendale. Grounded evaluation of information visualizations. In *Proc. CHI Workshop BEyond time and errors: novel evaLuation methods for Information Visualization (BELIV)*. ACM, 2008.

[24] A. Jain, R. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(1):4–37, 2000.

[25] S. Johansson and J. Johansson. Interactive dimensionality reduction through user-defined combinations of quality metrics. *IEEE Trans. Visualization and Computer Graphics*, 15(6):993–1000, 2009.

[26] I. T. Jolliffe. *Principal Component Analysis*. Springer, 2nd edition, 2002.

[27] Y. A. Kang and J. T. Stasko. Characterizing the intelligence analysis process: Informing visual analytics design through a longitudinal field study. In *Proc. IEEE Symp. Visual Analytics Science and Technology (VAST)*, pages 19–28, 2011.

[28] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. Empirical studies in information visualization: Seven scenarios. *IEEE Trans. Visualization and Computer Graphics*, in press.

[29] K. Leyton-Brown, E. Nudelman, and Y. Shoham. Empirical hardness models: Methodology and a case study on combinatorial auctions. *Journ. ACM*, 56(4):22:1–22:52, 2009.

[30] W. Matusik, H. Pfister, M. Brand, and L. McMillan. A data-driven reflectance model. *ACM Trans. Graphics*, 22(3):759–769, 2003.

[31] J. E. McGrath. Methodology matters: Doing research in the behavioral and social sciences. In R. M. Baecker, J. Grudin, W. Buxton, and S. Greenberg, editors, *Readings in Human-Computer Interaction: Towards the Year 2000*, pages 152–169. Morgan Kaufmann, 2nd edition, 1995.

[32] T. Munzner. Visualization (Chapter 27). In *Fundamentals of Graphics*, pages 675–707. AK Peters, 3rd edition, 2009.

[33] B. A. Nardi, S. Whittaker, and E. Bradner. Interaction and outeraction: Instant messaging in action. In *Proc. Conf. Computer Supported Cooperative Work (CSCW)*, pages 79–88. ACM, 2000.

[34] A. J. Pretorius, M. A. P. Bray, and A. E. Carpenter. Visualization of parameter space for image analysis. *IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis)*, 17(12):2402 –2411, 2011.

[35] A. J. Pretorius and J. J. Van Wijk. What does the user want to see? what do the data want to be? *Information Visualization*, 8(3):153–166, 2009.

[36] L. Reveret, L. Favreau, C. Depraz, and M.-P. Cani. Morphable model of quadrupeds skeletons for animating 3D animals. In *Proc. Eurographics Symp. on Computer Animation (SCA)*, pages 135–142. ACM, 2005.

[37] M. Sedlmair, M. Meyer, and T. Munzner. Design study methodology: reflections from the trenches and the stacks. Submitted to InfoVis 2012.

[38] M. Sedlmair, A. Tatu, M. Tory, and T. Munzner. A taxonomy of cluster separation factors. *Computer Graphics Forum (Proc. EuroVis)*, in press.

[39] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Proc. IEEE Symp. on Visual Languages*, pages 336–343, 1996.

[40] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[41] W. S. Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 17:401–419, 1952.

[42] M. Tory and S. Staub-French. Qualitative analysis of visualization: A building design field study. In *Proc. Conf. BEyond time and errors: novel evaLuation methods for Information Visualization (BELIV)*, pages 7:1–7:8. ACM, 2008.

[43] W. Trochim. *The Research Methods Knowledge Base*. Cornell Custom Publishing, 2nd edition, 1999.

[44] O. Tunçel, K. Altun, and B. Barshan. Classifying human leg motions with iniaxial piezoelectric gyroscopes. *Sensors*, 9(11):8508–8546, 2009.

[45] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journ. Machine Learning Research*, 9:2579–2605, 2008.

[46] L. van der Maaten, E. Postma, and H. van den Herik. Dimensionality reduction: A comparative review. Technical report, Tilburg University, TiCC-TR 2009-005, 2009.

[47] M. West. Bayesian factor regression models in the "Large p, Small n" paradigm. In *Bayesian Statistics*, pages 723–732. Oxford University Press, 2003.

[48] M. Williams and T. Munzner. Steerable, progressive multidimensional scaling. In *Proc. IEEE Symp. Information Visualization (InfoVis)*, pages 57–64, 2004.

[49] I. Witten, E. Frank, and M. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2011.

[50] J. Yang, W. Peng, M. O. Ward, and E. A. Rundensteiner. Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets. In *Proc. IEEE Symp. Information Visualization (InfoVis)*, pages 105–112, 2003.

[51] G. Young and A. Householder. Discussion of a set of points in terms of their mutual distances. *Psychometrika*, 3:19–22, 1938.