# Geono-Cluster: Interactive Visual Cluster Analysis for Biologists

Subhajit Das*, Bahador Saket*, Bum Chul Kwon, and Alex Endert

**Abstract**—Biologists often perform clustering analysis to derive meaningful patterns, relationships, and structures from data instances and attributes. Though clustering plays a pivotal role in biologists' data exploration, it takes non-trivial efforts for biologists to find the best grouping in their data using existing tools. Visual cluster analysis is currently performed either programmatically or through menus and dialogues in many tools, which require parameter adjustments over several steps of trial-and-error. In this paper, we introduce Geono-Cluster, a novel visual analysis tool designed to support cluster analysis for biologists who do not have formal data science training. Geono-Cluster enables biologists to apply their domain expertise into clustering results by visually demonstrating how their expected clustering outputs should look like with a small sample of data instances. The system then predicts users' intentions and generates potential clustering results. Our study follows the design study protocol to derive biologists' tasks and requirements, design the system, and evaluate the system with experts on their own dataset. Results of our study with six biologists provide initial evidence that Geono-Cluster enables biologists to create, refine, and evaluate clustering results to effectively analyze their data and gain data-driven insights. At the end, we discuss lessons learned and implications of our study.

## 1 Introduction

Clustering is the task of summarizing and aggregating complex multi-dimensional data in such a way that items in the same group are more similar to each other than those in different groups. Domain experts often want to perform clustering to find groups of data items that share common characteristics with respect to data attributes. For example, a biologist who wants to investigate genome data can cluster gene sequential data according to similarity between their expression profiles. Clustering has a widespread application in several domains [1], [2], [3].

Our paper aims to accommodate the process of interactive visual clustering for biologists. Like other domain experts, biologists also want to cluster their data and visualize the result to investigate patterns, relationships, and structures among data instances and attributes. However, not all biologists often have formal data science training. The lack of knowledge in data science often prevents users from clustering their data and from interpreting the results in the biological context using the existing tools. Based on our collaborations with a group of biologists, we found that they use tools such as SAS and/or programming languages like R to run cluster analysis on their data. These programming languages (or tools) require users to specify clustering algorithms and parameters in written scripts. The absence of interfaces designed specifically for clustering tasks frequently required by biologists may increase execution costs and impede the adoption of clustering methods.

There is a large body of visual analytic systems that employ visual clustering as a part of high dimensional data analysis (e.g., [4], [5], [6], [7], [8], [9], [10]). Some of these visual analytic systems are often complex, and require careful tuning, steering, and parameterization of the clustering models. Interaction complexity in such systems often poses fundamental usability challenges for those domain experts who may not have formal data science training [11]. Furthermore, it is challenging for domain experts to directly apply their knowledge into the clustering processes. For example, biologists exploring genome data might want to merge two clusters because of the similarity of evolutionary history of the genes located in two clusters. Alternatively, they might want to subdivide a specific cluster to estimate the disease risk of genes in different sub-clusters in a specific population. As such, current tools are ill-equipped to help biologists build and explore alternate groupings based on their domain expertise, hindering their ability to discover patterns in the data. Many such tools lack usable interactions to allow domain experts to translate domain-specific questions and hypotheses about the data into model parameters to foster the exploratory process of their tasks.

To tackle the challenges for biologists, we present **Geono-Cluster**, a visualization tool that applies the "by demonstration" [12] paradigm. Instead of requiring biologists to transform their clustering tasks into system specifications by going through layers of menus or programming it, Geono-Cluster allows biologists to directly apply their domain expertise by visually demonstrating how their expected changes should look like (e.g., dragging one cluster and dropping it over another cluster to show their interest in merging the clusters). By translating these demonstrations into numerical processes that update the underlying cluster distance functions, the system predicts biologists' intentions and generates potential clustering results (e.g., different visual clustering outputs that merged those two clusters). Thus, Geono-Cluster is not designed solely for constructing the most accurate cluster model, but instead to help users glean insights through data exploration facilitated by the the process of testing multiple clustering hypotheses realized as alternative models. While at times these two goals can be met at the same time by specific models (accurate and domain-relevant models), the exploration of alternative models may at times lead users to choose models which sacrifice overall model accuracy for the benefit of allowing them to understand a new aspect of the data. We have developed Geono-Cluster in collaboration with biologists investigating disease risks frequency across different populations. We closely followed the design study protocol [13] to derive system requirements, tasks to be supported,

*Bahador Saket and Subhajit Das contributed equally to this work.*

- *Bahador Saket, Subhajit Das, and Alex Endert are with Georgia Institute of Technology. E-mail: saket, das, endert@gatech.edu*
- *Bum Chul Kwon is with IBM Research. E-mail: bumchul.kwon@us.ibm.com.*

and design guidelines based on feedback from biologists.

We conducted a qualitative study with six expert biologists. In this evaluation, we observed how our tool helps biologists to cluster their data and identify challenges they encounter while using our tool. We also conducted a semi-structured interview to collect biologists' feedback and new ideas. Our results demonstrate that Geono-Cluster enables biologists to build, refine, and evaluate clustering outcomes with intuitive demonstration-based interaction and to interactively explore the results through multiple views.

## 2 Related Work

Different domains are seeing a surge in data collection at an alarming rate, which needs to be efficiently analyzed [14]. Clustering a dataset is a popular approach to understand the inherent structure of large datasets and is used in several critical domains [1], [2], [3]. Many tools and programming languages such as R, Matlab, SAS, and Python support cluster analysis. Despite the flexibility, using such methods often require an intermediate to advanced knowledge of programming skills. For this reason, domain experts often need to go through a steep curve of learning these programming languages. Furthermore, these tools often lack visual feedback and interactivity, which make users difficult to understand the results and to reconfigure the setting for improved results in next iterations. Thus, the lack of interactivity can increase execution costs and impede the data exploration process [15].

### 2.1 Interactive Visual Clustering Analysis

Researchers have been investigating various techniques and approaches to facilitate interaction in clustering analysis, with the goal of bringing a human in the loop. Effective user interaction is critical to the exploratory data analysis process, and thus to the success of the visual analytic systems for visual clustering. A large body of previous work designed and implemented interactive tools to support interactive visual clustering and analysis (e.g., [5], [10], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25]).

Clusterophile [17] and Clusterophile 2 [5] are both designed to enable users to explore different choices of clustering parameters and reason about clustering instances in relation to data dimensions. Datta et al. built an interactive clustering system - CommunityDiff, showing a mechanism to visualize ensemble space by using a weighted combination of various clustering algorithms to aid identifying patterns, commonalities, and differences [26]. iVis-Clustering [27] is another tool that supports document clustering based on a widely used topic modeling method called latent Dirichlet allocation (LDA). Hu et al. [21] and Guo [23] separately developed interactive tools that allow users to select features while clustering their data. ClusterSculptor [20] is another tool that aids data scientists in the derivation of classification hierarchies in cluster analysis. VisBricks [28] provides multiform visualization for the data represented by clusters (it enables users to select which visualization technique to use for which cluster). In a different project, Basu et al. [19] proposed a tool that allows users to move data items and build clusters of data items from a larger set, while the system suggests data items which can be further added to the set. ClusterVision [10] is a more recent tool that enables users to cluster data using a variety of clustering techniques and parameters and then ranks clustering results utilizing five quality metrics.

Geono-Cluster differentiates itself from the aforementioned work mainly by supporting biologists' visual clustering analysis. Many of the existing visual analytic systems often require careful tuning, steering, and parameterization of the clustering models [5], [20], [27]. In such systems, analysts need to translate their analytic goals into clustering specifications by going through layers of menus. Unlike existing tools, Geono-Cluster enhances user interaction expressivity by enabling users to interactively define clustering results by their demonstration on data items, which is more user-friendly and easy to understand for domain experts.

### 2.2 Interactive Clustering Analysis for Biologists

There exist a few visualization tools that are designed for clustering analysis of biological data. StratomeX [29] is an interactive visualization tool that enables users to explore the relationships of subtypes across multiple genomic data types. StratomeX is mainly designed to support tasks with "comparative nature" (e.g., evaluate how well two or more stratifications support each other). CComViz [30] is a different application that uses the parallel sets technique to compare clustering results. Kern et al. proposed novel methods for evaluating and comparing cluster results and implemented their methods into StratomeX [31]. XcluSim [32] is another tool for bioinformatics data helping users to compare multiple clustering results, supporting a diverse set of algorithms.

Unlike other tools that are mainly designed to support tasks with "comparative nature", Geono-Cluster is designed to cover a different category of tasks such as customizing, merging, and splitting clusters. Moreover, Geono-Cluster aims to reduce biologists' cognitive cost and enhance interaction expressivity by implementing the "by demonstration" approach [12]. Geono-Cluster enables biologists to apply their domain expertise into clustering processes by interacting with visuals representing data items and clusters (e.g., a biologist can express that a data item does not belong to a cluster by dragging it out of a cluster). As a result, the system finds the most appropriate clustering results based on the user interactions (e.g., finding clustering results where the selected data item does not belong to the specified cluster).

### 2.3 Demonstration-Based Interaction

Demonstration-based interaction has been applied to many applications. A common application of the technique in human-computer interaction is "programming by demonstration" [33]. Other domains that have successfully used the "demonstration-based" paradigm include data cleaning [34], [35], database querying [36], temporal navigation [37], visual data analysis [38], and visualization construction [12], [39]. For example, Kandel et al. [35] enables users to demonstrate desired changes to tabular dataset by making direct edits to the table elements (e.g., select and delete empty rows) [40]. In response to the given demonstrations, the system suggests potential transformations to accept to generalize the demonstrated change and update the data table. Prior works also enabled steering dimension reduction models by demonstrating relative similarity between data items (e.g., [41], [42]).

Our work enables biologists to demonstrate their desired clustering results by directly manipulating visual elements representing clusters (e.g., moving a subset of data items from one cluster to another). In response to users' demonstrations, the system computes possible clustering results and recommends them. Inspired by previous work [12], [35], each recommendation provides a visualization which gives an overview of the clustering result and a textual explanation.

## 3 Formative Assessment

Here we explain a formative assessment that we conducted to characterize users' workflow, derive tasks and requirements from it, to generate design guidelines to design the system.

### 3.1 Characterizing domain experts, data, and tasks

The motivation of this work stems from an ongoing project in which we have been collaborating with biologists at the Georgia Tech. We have been working with the biologists over the past 13 months to design and build solutions for supporting interactive visual clustering of disease risk factors.

The dataset used by the biologists is from Genome Wide Association Studies (GWAS Catalog) [43] which includes published SNPs (single-nucleotide polymorphisms, representing differences in a single DNA building block, called a nucleotide), and association studies to analyse genetic sequences. Through this dataset, biologists intend to determine "alleles" that correlate to various diseases and traits. Alleles are various forms of a gene that are formed by mutation and are found at the same place on a chromosome. Using GWAS dataset biologists analyse SNPs to find how do they vary between various genome samples.

During data analysis, biologists often focus on certain features of their dataset such as, disease/trait, SNP identification number, risk allele frequency, p-value, and odds ratio/beta. Focusing on those values, they try to answer questions like, how and why disease risk frequencies differ across populations, what are the statistical power to detect those known SNPs, and how well associations found in one population can transfer/replicate well to another population. To answer such questions, researchers cluster their data to investigate patterns and relationships of position on the genome, risk allele, and risk allele frequencies that impact diseases risk frequencies across different populations. This is an iterative process and biologists frequently create customized clusters, merge/split clusters, and investigate sub-clusters within a specific cluster to test their hypotheses based on their expertise.

To cluster and visualize their data, these biologists currently use tools/programming languages like Python, R, and SAS. They revealed that the current process of clustering and visualizing the data is rather time-consuming, cumbersome, and occasionally error-prone. Our observations as well as researchers' feedback show that they often need to write and execute scripts to accomplish their tasks. Writing scripts becomes even more challenging when they want to perform more specific tasks such as merging two or more clusters, as they have to translate these operations into the proper syntax, sequence, etc. To overcome these challenges, we designed an interactive visual clustering tool that enables visual data clustering specifically tailored for biologists.

### 3.2 Tasks and Requirements

Following a user-centered method [44], we began our iterative design process by investigating current practices, needs, and challenges. We conducted multiple group discussions with two biologists at the Georgia Tech. We started our discussions with the biologists by asking them: 1) what kinds of questions do they ask and answer while exploring their data? 2) why do they perform clustering tasks during their analysis?, and 3) how do they currently create clusters? Then, we freely continued our conversation that touched upon the tools, analytic methods, and challenges they face during the process. We took notes during all the group discussions. We then read through our notes to gain a better understanding of the requirements and challenges these biologists encounter while clustering their data. After reading the data, we identified the meaningful text segments (e.g., *"[...] here we combine these two clusters."*). We then assigned a code phrase that describes the meaning of the text segment (e.g., merging clusters).

We initially identified three commonly performed clustering operations that are currently challenging for biologists to complete using existing programming languages and tools. Here we define "clusters" or "clustering" interchangeably in two contexts: (1) Algorithmic clustering models such as K-Means,etc, and (2) Group of data items assigned to a collection based on an algorithmic cluster model represented visually as a grouped node view.

**T1: Hand-craft, Merge, and Split Clusters:** Biologists apply their domain knowledge to create customized clusters to better understand which factor(s) is causing the ascertainment bias on the dataset that are being used popularly. For example, one of the biologist stated: *"Given the identified SNPs [single-nucleotide polymorphisms] that are associated with common disease and traits, it's interesting to create a cluster of SNPs."* In addition, biologists apply their domain expertise to merge or split two or more clusters depending on how related they think the clusters are based on given feature(s). For example, one of the biologists mentioned: *"Depending on the evolutionary history of the genes, two or more clusters can be really related to each other. If ascertained they are related, we will merge them as one cluster."* Another biologist reported that *"In my new project, we are comparing Africans to non-Africans. In this case I merge Americans, East Asians, and Europeans as one cluster, and compare that to Africans data."*

**T2: Divide each cluster to sub-clusters:** Biologists often investigate sub-clusters within a specific cluster to: 1) understand which other factors can affect the cluster, 2) compare two clusters based on the member data items in each, and 3) see trends and patterns in the sub-clusters, with respect to chosen features, We noticed that the biologists found existing solutions challenging because they had to write lines of scripts to compute and visualize sub-clusters in a given cluster. Furthermore, the existing methods prohibit rapid iteration and visualization of results, which inevitably prolongs the exploratory clustering process to understand their data better.

**T3: Adjust feature contributions:** Biologists need to easily see by how much different attributes/features contribute to computing a cluster. Moreover, they often need to adjust the importance of different features used for computing a cluster. Biologists currently have to programmatically adjust the importance of features, execute the code, and visualize the outcome. They often repeat this process multiple times until they achieve a satisfactory result. They need interactive methods to view and refine feature contributions.

### 3.3 Design Guidelines

We needed to explore alternatives and make design decisions to better support the aforementioned tasks. In particular, GeonoCluster should be easy to use by experts who do not have formal data science training. We developed a set of design guidelines to inform those interested in developing visual analytic tools for domain experts (in particular biologists). These guidelines are based on existing tools designed for supporting visual data exploration for biologists [31], [32], mixed-initiative systems [45], and our experiences through several design iterations with biologists.

**G1: Shifting the burden of specification from the biologists to the systems.** The existing tools and technologies put the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2020.3002166, IEEE Transactions on Visualization and Computer Graphics



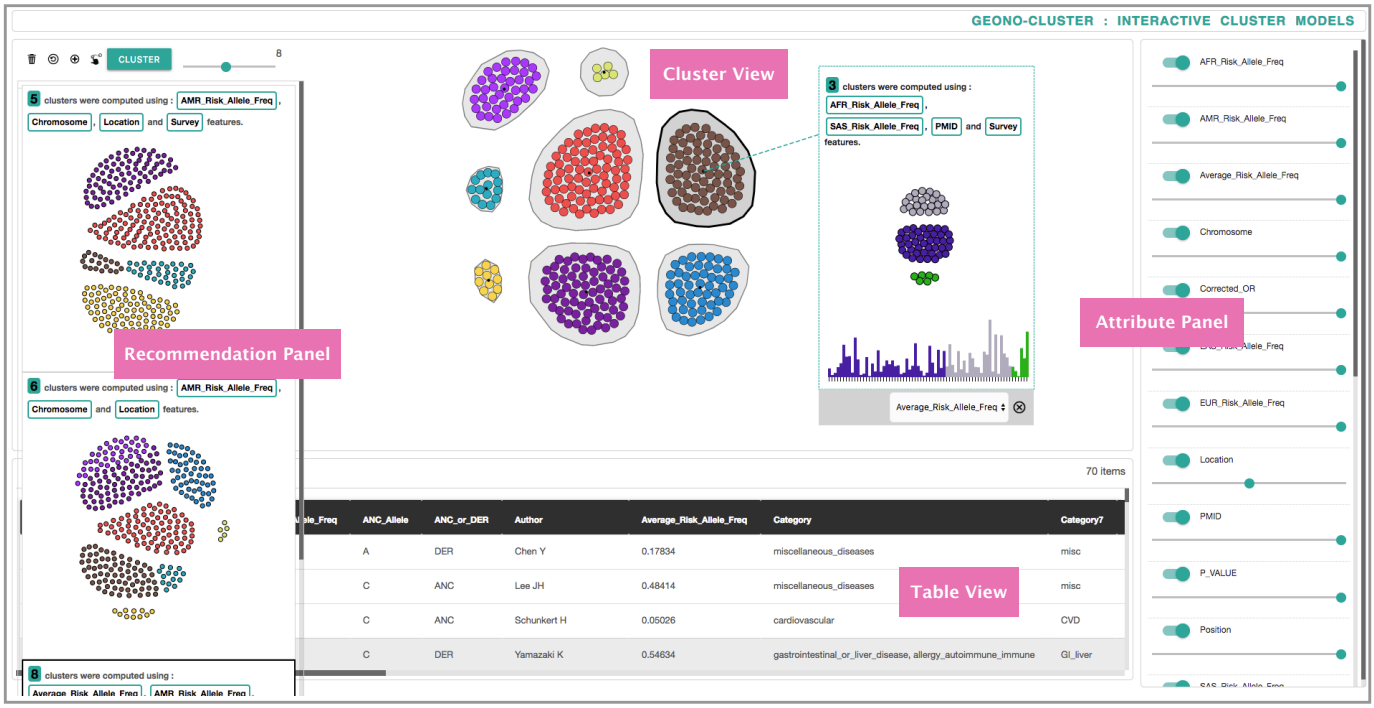


Fig. 1. The Geono-Cluster user interface consists of a Cluster View, a Recommendation Panel, a Table View, and an Attribute Panel. Cluster view visualizes the clustered data and provide a medium for users to provide visual demonstrations. Recommendation panel shows different clustering results based on the demonstrations provided by users. The Table view and Attribute Panel show the raw data and attribute weights.

burden of specification on biologists. For instance, one of the biologists noted: *"It sometimes takes time to perform specific tasks [using Python]. I have to Google and find out how to do it."* Instead of requiring biologists to specify the clustering models by programming or going through layers of menus, the tool should provide an environment that enables them to demonstrate how the expected clustering outcomes should look like [12]. By translating the given demonstrations, the system could estimate the biologist's intention and generate appropriate results. This way we could balance the responsibility between the biologist and the system – biologists provide visual demonstrations, based on this, the system infers potential clustering results and recommends them.

**G2: Enable user interaction to drive recommendations.** As analysts explore their data, their interests will evolve [46]. Our initial observations and interviews also showed that biologists need to explore various clustering models rapidly during their data analysis process. One potential approach to support such a rapid data analysis is to recommend potential cluster models that biologists should consider during their data analysis process [31], [45]. Furthermore, the clustering recommendations should be adapted for biologists' analytic goals. The recommendation engine should steer multiple clustering models based on biologist-specified expected visual outcomes. In addition, biologists can also directly adjust feature contributions to update the clustering results. In aggregate, these interactions create demonstrations which serve as the primary units by which biologists communicate their expected changes to the system.

**G3: Enhance interpretability of recommendations.** Biologists reported their interest in seeing more details about different clustering results while skimming through different recommendations. However, not all biologists might be familiar with technical terms used to describe a cluster such as silhouette value. Therefore, recommended clustering results should be presented in a transparent manner so that biologists can extract the most important and understandable information (e.g., contributing features) used for clustering results. One powerful approach to enhance transparency of the recommended clustering options is to use natural language [47] to explain them. This way biologists can learn about the recommended clustering outcomes without having to know about more technical terms describing each clustering outcome.

## 4 GEONO-CLUSTER

Based on the tasks and guidelines, we developed Geono-Cluster, a visual clustering tool for biologists. All components of the Geono-Cluster were implemented using JavaScript, D3.js, and Python.

### 4.1 Usage Scenario

In this section, we motivate the design of our system and illustrate the functionality via a usage scenario. We indicate how a domain expert can utilize Geono-Cluster to perform visual cluster analysis on the GWAS Catalog dataset [43]. This dataset includes detailed information regarding the identified single-nucleotide polymorphisms (SNPs) associated with common diseases and traits (e.g., position on the genome, risk allele frequencies, p-value, effect sizes, etc.). SNP is a region on the gene where more than one allele (A, C, G, T) is observed and each row on the dataset is a SNP [48].

Megan is a biologist who wants to compare populations from the *GWAS* dataset to understand disease risk factors related to geographical regions (e.g., if gene samples collected from "America" are more prone to cancer than gene samples collected from "Europe"). She launches Geono-Cluster to cluster the data, and to compare associated sub-populations. First, Megan skims through different features on the Table View (see Figure 1).

Megan knows that there are two types of gene samples: ANC and DER. ANC samples are the genes that are derived from either humans or monkeys. DER are the gene samples that are

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2020.3002166, IEEE Transactions on Visualization and Computer Graphics



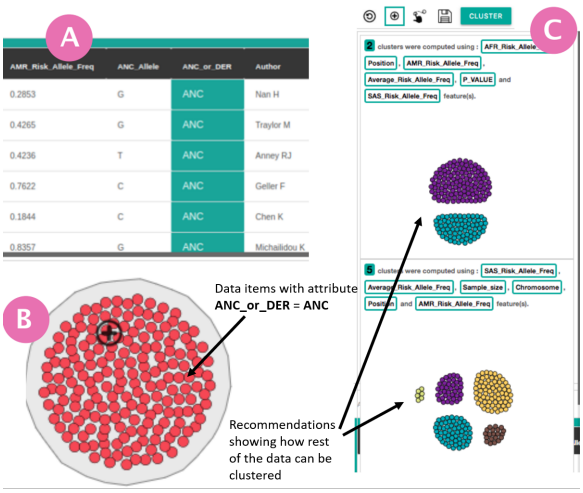


Fig. 2. **A)** Megan clicks on a cell in the column *ANC-or-DER* with the value *ANC*. The system automatically selects all data items with ancestry *ANC*. **B)** She drags the selected data items and drops them to the cluster view. The system automatically represents data items as red circles and places them in an independent cluster. **C)** The system also recommends potential clustering layouts of the non-interacted data instances based on the demonstration provided by Megan.

derived from the mixture of humans and monkeys. Megan starts her exploration by comparing the disease risk factor between the two types of genes samples based on their ancestry. Megan first skims through different features to find the *ANC-or-DER* feature on the Table View. She clicks on a cell in the column *ANC-or-DER* with the value *ANC* in the Table View to demonstrate her interest in selecting all the data items with ancestry *ANC*. In response, Geono-Cluster automatically selects all data items with ancestry *ANC* (see Figure 2-A). Megan then demonstrates her interest in clustering data items with *ANC* value by dragging them from the Table View and dropping them to the Cluster View. In response to the demonstration, the system automatically represents data items as red circles and places them in the *Red Cluster* (see Figure 2-B). At this point, the system also recommends potential clustering results based on the demonstration provided by Megan (see Figure 2-C). Even though Megan's interactions may lead to grouping the data based on the chosen categorical data attribute (*ANC or DER*), in essence, this is a start to allow a user demonstrate their intent to find a clustering model that represents agreeable clusters in the data. Their interactions are inferred as implicit intents by the system to find the most appropriate cluster model as opposed to just group the data by a set of categorical variables.

Megan opens the recommendation panel and previews other clustering options through the thumbnail previews. She finds one of the clustering results recommended by the system interesting. She clicks on this thumbnail (the first recommendation), which updates the Cluster View with the recommended cluster layout by adding the *Blue Cluster* and the *Purple Cluster* (along with the *Red Cluster*) in the Cluster View (see Figure 3-A).

Megan explores the data items within each cluster by hovering over each data item to see its details. She notices that while the *Red Cluster* contains genome samples with ancestry *ANC*, the recently added clusters (*Blue Cluster* and the *Purple Cluster*) contains all the gene samples with ancestry *DER*. She further notices that most of the items in the *Blue Cluster* have the chromosome value higher than 8, and the genome samples belong to the region *America*. Now she understands what each clusters represent.

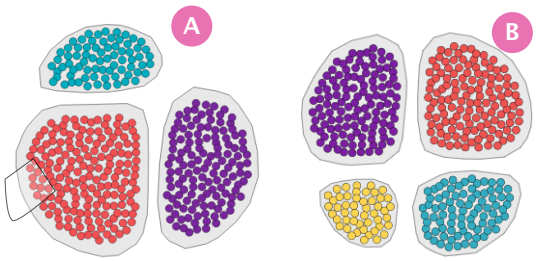


Fig. 3. **A)** Megan uses lasso tool to select a subset of data items from the red cluster and drags them out. **B** The system automatically finds other similar data items and defines the yellow cluster containing them.

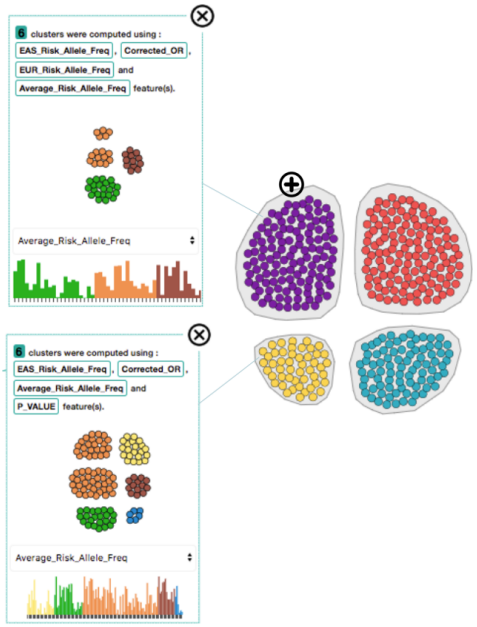


Fig. 4. Megan clicks on the "+" icon to open the sub-cluster panel for clusters purple and yellow. Bar chart views showing comparisons of the feature Average-Risk-Allele between these clusters.

Next, Megan demonstrates her interest of excluding data items with ancestry *ANC* that belong to *Africa* from the *Red Cluster*. To do so, she lasso-selects a subset of data items with ancestry *ANC* that belong to the region *Africa* from the *Red Cluster* (see Figure 3-A). Important to note that points closer to each other in a cluster are expected to be similar to one another based on the applied cluster model. Thus using the lasso selection, Megan selects similar points from a cluster for further analysis. She then drag-and-drops these points out of the cluster. In response, the system automatically finds other similar data items with ancestry *ANC* that belong to the region *Africa*, and then defines the *Yellow Cluster* containing these data items.Further, the system updates the recommendations in the recommendation panel accordingly.

Looking at the *Purple Cluster*, and the *Yellow Cluster*, Megan realizes that the items in these two cluster are with ancestry *ANC* and *DER* respectively. Megan wants to compare the distribution of the feature *Average-Risk-allele* between these two clusters to compare their disease risk factors. She clicks on the "+" icon (see Figure 4), which is shown upon hovering on a cluster, to open the sub-cluster panel for each cluster. Each sub-cluster further clusters the data items per cluster. Also, the sub-cluster panel contains a bar chart, highlighting the distribution of a chosen feature (*Average-Risk-allele*) through a drop-down selector for all data items in the

parent cluster (see Figure 4). After inspecting the distributions, Megan does not notice any significant difference in the value of *Average-Risk-Allele* between the *Purple* and the *Yellow* cluster.

Megan explores the *Blue Cluster* (with *DER* ancestory) to inspect its sub-cluster layout and distribution of the feature *Average-Risk-Allele*. When she compares the distribution of feature contributions of the *Yellow Cluster*, she discovers that the genes sampled from *Africa* with ancestory *ANC* has much higher disease risk factor than those sampled from other regions with ancestry *DER*.

Megan decides to merge the clusters *Blue* and *Purple*. To do so, she demonstrates her interest in merging the clusters by drag-drop the *Blue* cluster on the *Purple* cluster. In response, the system recommends new cluster layouts on the Recommendation Panel. She previews the thumbnails from the recommendation panel and selects the second recommendation, which results in placing 3 clusters in the Cluster View. To continue discussing the findings and implications with other colleagues, she exports a *.png* screenshot of the current cluster layout. She also saves the results as a *.csv* file to investigate them more in other programs like R and SPSS.

### 4.2 Views and User Interface

Geono-Cluster's interface consists of: a Cluster View, a Recommendation Panel, a Table View, and an Attribute Panel. See Figure 1.

**Cluster View** visualizes the clustered data as Figure 1 shows. For testing their hypotheses, biologists often perform actions at the level of data items (e.g., move data items from one cluster to another). We visually present each cluster and its members on the Cluster View. The colored circles in each group represent members of a cluster; the surrounding hull represents the cluster. Users can hover over a circle, which prompts relevant attribute details of the data. Users can specify the number of clusters using the slider shown on the top-left. Cluster View is an environment similar to a spatial workspace in which users can move data items to structure their information and provide visual demonstrations **(G1)**. For example, a biologist might notice a set of data items should not be in a specific cluster. Thus, she can demonstrate that those points belong to a different cluster by dragging them from one cluster to another. The system uses the visual demonstrations provided by the users to steer the underlying recommendation engine **(G2)**.

The Cluster View shows an overview of the clustering results and then encourages users to query additional information (e.g., tooltips, attribute distribution histograms) as they explore the data. The visual representation of the Cluster View powered by a force-directed layout algorithm shows the size and shape of each cluster, the number of clusters, and an overview of clustered data items without overwhelming users with too much information. Furthermore, the design encourages cluster-level interactions such as merging two clusters, or splitting a cluster to refine or customize a cluster. Within each cluster the position of the node (a data item) placed at the center represents a stronger cluster membership than those that are on the periphery. This allows users to understand the clustering probabilities of each data item in relation to others. However, each cluster in the layout is positioned by the force-directed layout simulation that does not capture if two clusters are similar or different. We deliberately restricted ourselves to communicate that information as we intended users to inspect differences between clusters by viewing the thumbnail previews of recommended models.

**Recommendation Panel** shows different clustering results. Based on users' demonstrations on the Cluster View, the system recommends a set of appropriate clustering outputs. To compute the recommended clustering results, the underlying recommendation engine takes into account different (1) clustering techniques/algorithms; (2) combinations of attributes/features; and (3) clustering hyperparameters (i.e., varying 'k' for k-means clustering technique). Read section 4.4 for more details.

During the design process of Geono-Cluster, we examined different ways of presenting recommended clusters. We first considered showing all the recommended clustering results as small thumbnails in the Recommendation Panel. The biologists liked the idea and the way that we recommended clustering results. However, the main challenge that biologists encountered was that they were not able to infer detailed information from the small thumbnails. Thus, they requested adding textual description of details about each clustering result in the recommendation. Currently, each thumbnail includes a textual description about the number of clusters, features used to compute the clustering recommendations, and a visualization of the clustering result **(G3)**.

Initially, we designed the recommendation module to update the view with new clustering recommendations whenever users show their demonstrations and/or adjust feature contributions. However, our users revealed that such approaches may distract their ongoing investigations on the current results. Thus, we compute cluster recommendations in the background but do not show the results immediately. Once the computation is done, a notification pops up, encouraging users to explore the results on demand by toggling the 'show recommendations' button (see Figure 1).

**Table View** shows a tabular representation of the loaded dataset where each row is a data item (see Figure 1).Biologists specifically requested adding this view since it enabled them to check the raw data. It provides standard table interactivity, with the one feature that shows similar rows to those selected. From the design study we noted that the existing workflow of the biologists involved exploring the data in MS Excel, then using "R" to run clustering models, and then export the data back to MS Excel.

**Attribute Panel** lists the attributes of the loaded data set as Figure 1 shows. Users can turn on and off a set of attributes which directly affects the clustering algorithm. Furthermore, users can also adjust attribute contributions, specifying relative importance of the selected attributes to define cluster memberships **(G2)**.

### 4.3 Interactions

In this section we discuss how Geono-Cluster supports interactive operations commonly performed by biologists.

**Merging and Splitting Clusters (T1)**: To merge two or more clusters, users first click on a cluster. They then demonstrate their interest in merging two clusters by drag-and-dropping the cluster on top of another cluster. Users can drag point(s) out of the cluster and drop into either i) another cluster or ii) a blank space (on the Cluster View). Drag-and-drop items into blank space is translated as forming a new cluster of the selected items outside the current cluster (see Figure 2). Demonstration-based cluster customization enables users to interact with the data directly and removes any mid-level instruments such as control panels or menus.

The merge interaction is derived from the previous work by Sarvghad et al. [38], in which they enabled HIV researchers to merge bars in bar charts by dragging one bar and dropping it over another bar. Biologists liked this interaction design and found it "direct and intuitive". To split clusters, we initially enabled biologists to select the data items by clicking on each circle representing

a data item. However, biologists found it cumbersome and time-consuming. So, we implemented the lasso-selection such that users can select multiple data items easily. This operation allows user to brush over a set of data samples (represented as circles) in the Cluster View. In response the system extracts those samples from the current cluster and places them in a new cluster. If data samples from multiple clusters are selected (using lasso selection), then the system makes a new cluster from these lasso picked data samples.

**Sub Clustering (T2)**: Hovering over a cluster reveals a plus button. Users can click on it to open a subcluster panel on the Cluster View, which shows subgroups of the data items within the selected cluster. In addition, a bar chart shows the distribution of a chosen attribute. Alongside, text description highlights the attributes that were used to compute the sub-clusters. Given that the users are not experts in data science, we do not present the quality metrics (e.g., silhouette scores, homogeneity score, etc.) Instead, we describe cluster models by showing thumbnail previews of clustering results with text descriptions as Figure 4 shows.

**Delete data items or Clusters (T3)**: Our discussion with biologists revealed that they sometimes need to 'exclude' data items or clusters from their analysis while testing a hypothesis. Thus, we initially implemented the 'delete' feature by enabling users to select a subset of items or clusters from the main view and click on the delete icon. However, when we showed it to the biologists, they had trouble due to inconsistencies between the button-based interaction and other demonstration-based interaction.

Currently in Geono-Cluster users cam drag-drop a selected cluster on the delete icon shown on the top-left of the interface to show their interest in moving the selected cluster out of the layout. Similarly, they can drag-drop individual data items to demonstrate their interests in removing them from the cluster assignment.

**Creating Customized Clusters**: Users can select a subset of data items by clicking on the rows shown on the Table View (each row represents a data item). After selecting a subset of rows, users can drag-and-drop them on the Cluster View to demonstrate their interest in creating a clustering, in which all the selected data items fall in the same cluster. Users can iteratively repeat the process, and each drag-and-drop operation forms a new cluster in the Cluster View. Participants liked this idea as they found the design and the workflow of this interaction consistent with other interactions.

### 4.4 Computational Techniques

This section describes the underlying computational techniques which enable Geono-Cluster to recommend cluster models by incrementally steering (multiple cluster models) them to adhere to demonstrated user preferences. Our cluster model recommendation process includes the human in the loop. On a high level, the user shows their intentions on a cluster layout. Based on the operations, Geono-Cluster models multiple cluster algorithms and finds top $k$ closest cluster models to the users' intention. Then, the user can refine the results through a series of customizations (instrumented through the interactions described above). In response, Geono-Cluster automatically finds close variants of cluster models and updates the recommendations in the Recommendation Panel. In summary, the system finds a set of cluster models with a distance function that reflects user-demonstrated cluster assignments.

**Multiple clustering models:** The clustering task begins when the user requests a new cluster layout (when they press the cluster button in the interface). In response, Geono-Cluster generates multiple clustering models $M$. Each cluster model $M_i$ in $M$ ($M_1$,$M_2$,$M_3$,$M_4$, ... $M_T$) is defined by a careful combination of a learning algorithm $\omega_i$ and a set of $p$ hyperparameters $\phi$, defined as $\phi_{i1}$, $\phi_{i2}$, $\phi_{i3}$, $\phi_{i4}$, ... $\phi_{ip}$. Applied clustering algorithms include K-Means, DBScan, Agglomerative Clustering, and Spectral Clustering. In the evaluation of the system, we used K-Means cluster model as we observed through our design-study that most of our users are familiar with packages in "R" to use K-Means clustering (with default parameterization) to cluster the genome data. However, depending on the need of the user and the data used, Geono-Cluster can be extended to use other clustering methods. Nevertheless, each algorithm has its hyperparameters. For example, K-Means is a learning algorithm with "k" and the "max-iteration" value as an input hyperparameter. Furthermore, each model $M_i$ in $M$ is assigned a metric score $S_i$ to compute $S$, which defines the quality of the clustering output (a higher $S_i$ means a better cluster definition). Geono-Cluster uses Scikit-Learn's ML package to construct and evaluate the cluster models using various quality metrics (e.g., Silhouette Coefficient, Davies-Bouldin index)

**Recommendation Technique:** Geono-Cluster ranks the models in $M$ by their scores $S$ explained below, and visualizes the best clustering layout in the Cluster View. Further, the system allows the user to inspect top $f$ best cluster models from the ranked models $M$, through the Recommendation Panel (see Figure 1-a). If a user makes any customization to the shown cluster model $M_c$ (e.g., merge or split clusters), the system automatically updates the recommendations by computing a new set of $M$ cluster models, except the model $M_c$, which is currently shown in the Cluster View. Per iteration, the system updates $S$ and the ranking of the models $M$ based on user interactions with the data. Next it visualizes the best model in $M$ in the Cluster View and shows thumbnail previews of the top $f$ models in the Recommendation Panel (**G2**).

Geono-Cluster's model recommendation finds the closest fitting cluster assignments, whenever the user customizes the current cluster layout in the Cluster View. However, there can be scenarios that no cluster recommendation matches the user's intended changes. This may occur when users seek clustering results, which are mathematically infeasible. There could be various reasons for it, such as, users may have a different understanding of the data than what the data actually contains, or the data may have noise, etc. In such cases, users may need to be educated to understand the reasons for a different clustering result, which we plan to integrate in the workflow in the future. Currently, in such cases, Geono-Cluster still responds with the nearest best clustering output, though it may not resemble the layout shown by the user. Furthermore, to ensure users can see unexpected clustering results (to ideate and explore), every few iterations the system also recommends a set of cluster models that are randomly parameterized and thus clusters the data in an unexpected way. While our approach may seem similar to active learning (AL) [49] as in both, users specify feedback to the input data that drives the generation of a model. However, in our case the data does not have class labels. Furthermore, in AL, the system "asks" users to give feedback on specific data points, while in our technique users have the freedom to interactively explore and provide feedback any time along the process.

**Clustering Metric:** Initially the system does not have cluster assignments or labels for any of the data instances. Thus to compute $S$ the initial cluster models $M$, are evaluated using the Silhouette Score metric [50]. This metric is computed using the mean intra-cluster distance, and the mean nearest-cluster distance for every

data instance. As users interact and assign clusters to a set of data instances $I$, Geono-Cluster applies two types of metrics to calculate $S$. To compute the first metric $S_1$, the system finds all the correlational features $fc_k$ and non-correlational features $fc_k^{'}$ that describes each cluster ($k = 0$ to $g$ clusters). Here, $fc_k$ and $fc_k$ defines how the user characterises each cluster. Next, when $M$ is computed the system computes the correlational features $fc_{ik}$ and $fc_{ik}^{'}$. The system compares $fc_k$ with $fc_{ik}$ (and $fc_k^{'}$ with $fc_{ik}^{'}$) for each model in $M$ to derive the clustering metric $S_1$, that describes how closely the cluster model $M_i$ adheres to the clusters defined by the user ($S_1$ is normalized between $0-1$, higher is a better model). The second metric $S_2$ is based on the labels assigned to data items by users for $I$. The system finds other data instances $J = N - I$ ($N$ is all data instances) that are similar to the user interacted data instances using cosine similarity distance metric based on their attribute values (categorical variables are one-hot encoded). We apply a similarity threshold $\beta$ to find a satisfactory number of similar data instances, the value of which is set empirically with multiple trials on the GWAS data. The current prototype does not allow users to interactively control this threshold. However, in future on users request it can be interactively specified using a slider widget. The system automatically assigns to these similar data instances ($J$) the same class assignments that the users assigned to $I$. Next using this labelled data, the system applies Homogeneity index score [50] to compute $S_2$. This metric uses true labels and assigned labels by the system (when a cluster model $M_i$ is applied to the data) to give a score to each model in $M$. The final score $S$ is defined as the weighted linear combination: $S = \lambda_1 * S_1 + \lambda_2 * S_2$. Here the weights $\lambda_1$ and $\lambda_2$ are hyperparameters that are assigned based on how well the clustering outputs satisfied the biologists' expectations.

**User driven feature selection:** A cluster model $M_i$ is driven by a set of features $F$ = $f_{i1}$, $f_{i2}$,$f_{i3}$,$f_{i4}$ ..... $f_{ik}$ as input to compute the distance function which assigns a set of data items $D$ to individual clusters $C$. In Geono-Cluster, the set of features $F$ is either computed using feature selection methods e.g., "select K Best" [51], "PCA" [52] or can be retrieved from users if they specify a set of features and their relative weights (from the Attribute Panel supporting the task **T3**). When users specify a set of $k$ features $F_u$ = $f_{i1}$, $f_{i2}$,$f_{i3}$,$f_{i4}$ ... $f_{ik}$ with respective weights for each feature ($W_u$ = $w_{i1}$, $w_{i2}$,$w_{i3}$,$w_{i4}$ ... $w_{ik}$, the system updates the distance function in the clustering algorithm. The distance function is represented as $\sum_{i=1}^{k}\sum_{j=1}^{n}\left\|x_i^{(j)} * w_i^{(j)} - c_j\right\|^2$, where $c_j$, is the $j$th cluster centroid and $w_i^{(j)}$ is the user assigned feature weight.

**Sub Clustering:** When triggered by users, the system builds a sub-cluster model $M_{si}$, for data instances $E$, member of a selected cluster $C_i$. Unlike the set of main cluster models $M$, only a single sub-cluster model is generated per cluster (**T2**). For sub-clustering we relied on the parameterization of the best-recommended cluster model for the entire data i.e, best-found parameterization of the K-Means cluster model. To avoid further compute times that may impact real-time interactions, we did not construct and test multiple cluster models for sub-clustering. However, clicking on the "add subcluster" button again for the same selected cluster $C_i$, the system recomputes the sub-cluster model $M_{si}$, by randomly choosing a new set of a learning algorithm $\omega$ and hyperparameters $\phi$; e.g., it picks a new "k" on the "K-Means" cluster model. This technique allows users to rapidly browse a large set of sub-cluster models.

**Similar item selection:** Users click on a cell ($q_j$) of a quantitative attribute on the Table View to select a value $v_j$ of the data item $d_i$. Geono-Cluster finds a set of $r$ data instances, $U$ = $d_a$,$d_b$,$d_c$ ... $d_r$, each of whose value $v_j$ falls within a threshold range, say $[+eps, -eps]$. The parameter $eps$ is set for each quantitative attribute $Q$ by heuristics and can be adjusted. This technique allows users to pick data instances which are similar, based on the selected quantitative attribute $q_j$. Further, users can select another quantitative attribute cell $q_k$. Next, from the set of selected data instances $U$, the system finds all instances $V$ which fall within a threshold range of the value selected for attribute $q_k$. Here the size of $V$ is less than that of $U$. This technique allow users to filter and select a subset of data instances $V$ from the Table View. For categorical features $X$, Geono-Cluster performs exact feature value matching instead of matching data items based on a predefined range. Users can drag-drop these $V$ data items to the Cluster View as a single cluster ($C = C_1$). They can continue selecting another set of data items, then add them to the cluster view as a new cluster ($C = C_1, C_2$). Users complete the data exploration or they can request the system to find a model $M_i$ iteratively (**T3**).

## 5 Evaluation

To evaluate Geono-Cluster, we performed a qualitative assessment with six biologists to collect subjective feedback and observational data. Our study had two main goals: (1) collect qualitative feedback on Geono-Cluster's features and design, and (2) observe how experts perform visual clustering analysis using Geono-Cluster. In particular, our study indicates how Geono-Cluster helps domain experts gain insights into data by interactively building clusters.

### 5.1 Participants and Setting

We recruited 6 biologists (2 female, 4 male), all with graduate degrees related to Biology, Bio-Statistics or Bio-Informatics. They had $1-2$ years of experiences working with *Gene* related datasets. They had not participated in our preliminary evaluation of Geono-Cluster and were also not involved in the design of Geono-Cluster. All participants were familiar with the concept of data clustering and had previous experience with data grouping with at least one data analysis tool (e.g., SAS, R, etc.). Further, as they had previously worked with *GWAS* catalogue data, they were familiar with all the data attributes in the dataset. During the entire study participants used a computer with 17-inch screen and used a mouse to interact with the system. The study took approximately 50 minutes and we rewarded each participant with a $20 gift card.

### 5.2 Procedure

**Introduction and Training:** Participants were briefed about the purpose of the study and their rights. After filling out the study consent form and a questionnaire on demographics, we asked participants to watch a tutorial video of Geono-Cluster. The video walked the participants through different features and interactions provided by the tool. After watching the video, we asked participants to work with the tool for 10 minutes. We encouraged the participants to ask as many questions as they want during this stage.

**Main Study:** The participants were asked to explore the *GWAS* Catalogue [43] data that includes published *SNPs* and association studies. In particular, we asked the participants to imagine their colleagues asked them to analyze the dataset using the visualization tool for 30 minutes and report their findings. Participants were instructed to verbalize analytical questions they have about the data, the tasks they perform to answer those questions, and their

answers to those questions in a think-aloud manner. In addition, we instructed them to come up with data-driven findings rather than making preconceived assumptions about the data. The interviewer played a role of 'active listener' during the think-aloud protocol.

**Follow-up Interview:** After each participant completed the task, the experimenters asked participants to explain major obstacles of the tool and describe what they liked or disliked about the tool.

### 5.3 Data Collection and Method of Analysis

We screen and audio-recorded the whole study. During the main study, the experimenter took notes while participants interact with the system. We also collected feedback from a semi-structured interview with open-ended questions at the end of the study. We analyzed around 300 minutes of screen-capture videos from six participants. First, one of the authors transcribed the audio recording of the study. Then, two coders (first and second authors) read the transcribed data (including the think-aloud sessions and the interview responses) to parse a set of meaningful text snippets. After reading the data, each of the coders independently assigned codes (a word or phrase) to best describe the text snippets. Finally we consolidated the codes from the two authors by focusing on the aspects of the responses which highlighted positive or negative feedback with respect to *usability of the system*, *easy of use*, *learning curve*, *future feature requests* or *strategies pertaining to exploratory data analysis using clustering models*. In the following section, we use **P1** to **P6** to respectively denote the participants one to six who participated in the evaluation.

### 5.4 Results and Feedback

Overall, all participants found Geono-Cluster easy to use and effective in performing cluster analysis tasks. Below, we categorize and discuss the findings of our qualitative study in more details.

**System usability**: All participants found Geono-Cluster's workflow easy to use, intuitive, and engaging. P2 remarked *"I can keep trying new ideas to quickly test different ways to cluster this data."* P4 said *"It's so easy to use, I can quickly iterate and learn about the data much faster, than using packages in R to cluster data."* Further, many other participants found visualization to be a very good medium to learn about the data by exploring different clustering results. P5 said *"I never knew that I can use visual methods to explore clustering result. Currently I use R to cluster my data, then export a CSV file to my team-mates."*

**Consistency with user mental model**: Participants found the design and workflow of Geono-Cluster consistent with their mental model and expectations. In particular, participants found that it is intuitive to visually demonstrate tasks such as creating, merging, and splitting clusters by demonstration. For example, P3 mentioned: *"it feels intuitive to merge clusters by dragging and dropping one cluster over another one. [...] this is what I would expect to happen."* P5 stated: *"I liked the idea of creating a cluster of items by moving the data items from this table to the empty space [dragging the data items from the Table View and dropping them on the Cluster View to create a cluster]."* Further P2 added: *"Compared to programming, using this kind of tool is more straight forward and faster."* Consistency and natural mapping between user's intent and the actions required for performing the intent is important in designing new interactions.

**Perceived control over data analysis process**: While using Geono-Cluster, P1, P4, and P5 commented on their level of control over the data analysis that resulted from their freedom in interacting with visualizations instead of going through layers of menu items. For example, P1 mentioned: *"This is great because I can construct my own cluster and tell the system how I want my clustering outcome looks like."* P4 stated: *"It is a powerful idea to enable analysts to use their knowledge about the data items to interactively create clusters [visually demonstrate their expected clustering outcome]. I specifically like how this allows merging and splitting clusters."* The level of interaction directness [53] with the visual representation contributes towards increasing the perceived control of the participants over the data analysis process.

**Difficulty in splitting a cluster**: Participants found the lasso interaction intuitive and easy to use. However, with lasso selection participants were not very exact about the data items that they wanted to select. For example, after selecting a subset of data items, P3 noted: *"It is hard to be exact with this selection. I don't want this specific point to be selected."* In such cases, participants had to either deselect the items that were selected incorrectly by clicking on them or try to lasso select again. Going forward, we envision designing advanced interaction techniques for easier selection of data items that are located in a close distance from one another.

**Interpretability of recommendations**: Although some participants liked how the recommendations were presented, two participants could not immediately understand why specific recommendations are suggested. For example, P2 mentioned: *"I understand what each cluster represents which is good, but I am not sure why these recommendations."* and P3 stated: *"I am curious how these recommendations are added."*. Going forward, we suggest systems to explore design alternatives to explain the reasoning behind recommendations. In situations when the system does not find any cluster recommendations that matches user's demonstrated changes, Geono-Cluster shows the nearest best clustering layout. In such scenarios, users may be surprised to see the abrupt or strikingly different recommendations. In the future, we are thinking of explicitly communicating this conflict in textual description. At the same time, we want to introduce a more variety of models so that the system can perform deeper search to find desirable results.

## 6 Observations

Our user study reveals that participants usually began exploring the data by framing a hypothesis, asking the questions they want to know, and then performing a set of tasks (as described in section 3.2) through Geono-Cluster's interface to find the answers. Interestingly, we observed that participants often took two different approaches to perform visual data clustering: **Top-down** and **Bottom-up**. Below, we describe each approach in more details.

### 6.1 Top-down Visual Data Clustering Approach

P1 started his data analysis process by asking *"How does the gene samples differ in disease risk factor by regions and chromosome factors?"* To that end, P1 clustered data items by selecting a set of features from the Attribute Panel and then pressed the *Cluster* button. Next, he checked the recommended cluster layouts from the Recommendation Panel to explore other clustering results based on another set of features. In response, he updated the list of features to cluster the data by and triggered Geono-Cluster to generate a new cluster layout. P4 also followed the same approach; however, he did not have any question to begin with. He initialized the process by pressing the cluster button to start with an initial clustering. Next, he hovered over data items in each cluster to familiarize himself with the data items and find similarity or dissimilarity. He

also checked the Table View to compare different data items from various clusters. If the clusters did not match his mental model, he would adjust the features from the Attribute panel. He would then preview the recommended clustering options to further explore a wide range of cluster outputs. This process continued until he was satisfied with the clusters and had a better sense of the data.

A main point here is that in the top-down approach participants mostly avoided interaction at the data item level, but instead they dealt with the full range of features from the Attribute Panel. P1 also verified this point by saying: *"I relied on cluster button to cluster the data, as I do not specifically know much about the data items, so did not use the table's drag-drop feature. Similarly, I did not customize the clusters by using lasso or drag-drop feature initially. I rather re-computed the clusters based on a new set of features that I specify."* However, P1 later confirmed that over iterations when he was more confident about the data, he started using the split and merge operations to customize shown clusters.

### 6.2 Bottom-up Visual Data Clustering Approach

Remaining participants (P2, P3, P5, and P6) followed the Bottom-up approach, in which they mainly relied on interaction at the data item level. They first created a customized cluster by dragging data items from the Table View and dropping them on the Main view as opposed to relying on the cluster button. These participants often interacted with data items to demonstrate their expected outcome.

P2 started her clustering analaysis by asking *"How does the gene samples derived from humans/monkeys (ANC) vary from gene samples derived from mixing humans and monkeys (DER) with respect to various diseases?"* To answer the question, P2 placed all the *ANC* gene samples into one cluster and a few *DER* gene samples into another cluster from the Table View. P2 remarked: *"my strategy is to select a set of data points [items] based on the gene's ancestry, then drag-drop to create a cluster"*. P2 then previewed the recommendations to explore other options to cluster the data based on his specification of clusters. In this process, P2 did merge/split clusters to test different ideas to cluster the data using the lasso-selection and the cluster drag-drop feature. P2 said: *"I also rely on the lasso tool to define other clusters from this, if the cluster appears too big"*. Using Geono-Cluster, many participants were able to customise clusters in this fashion to find interesting insights from the data, that they found needed further analytical investigations/research with their peers or mentors. For example, one participant was able to find a significant difference in *Average-risk-allele-frequency* between two sets of clusters by iteratively following this bottom-up visual clustering approach.

P6 also followed the same approach. P6: *"I want to know if the gene with chromosome factor higher than 6 sampled from America, have higher cancer risk factor? To seek an answer, I find the Table View's data item selection feature quite useful, as I can define my own clusters based on chromosome value or the region the gene was sampled from."* We noticed that when P6 explored the initial set of cluster layouts, he paid attention to the suggested features (in the Recommendation Panel) to understand how the cluster is defined. In some cases, P6 did not agree with the recommendations or the features that were used to derive the results. To provide his feedback for updated results, he customized the best-perceived cluster layout by splitting the existing clusters using the lasso tool and merging smaller clusters into one. P6 added: *I am using the lasso feature to take out all the data items which have chromosome value less than 6. Also, the smaller clusters with 5 or fewer data samples are confusing, so I merge them into one*.

## 7 Discussion

**Generalizability of the approach:** The methodology for this work stems from a design study [54] in biology. This inherently makes our contribution domain-specific, and solves a very specific problem that we discovered through working with biologists. However, the underlying interaction technique behind Geono-Cluster is generalizable and can be applied in other domains and on other tabular datasets. Our demonstration-based interaction design is a bottom-up approach where users provide demonstrations by interacting with data items. As such, this technique works whenever users can bring their knowledge by interacting with data points and provide demonstrations. For instance, another dataset that biologists use is cancer dataset to cluster patients based on the likelihood to be diagnosed with cancer. In this case, the domain experts may know patients with certain chromosome value or blood count level.

**Human bias in interactive clustering:** The human-in-the-loop nature of Geono-Cluster introduces potential user biases in visual data exploration. In fact, some amount of human bias exists in most interactive systems (e.g., control panel style interfaces). However, the key goal of Geono-Cluster is to help users explore different aspects of data while testing their hypotheses using clustering models. More importantly, the goal of our tool is not to help users build the most accurate cluster model, but rather to: (1) explore alternate models and their outputs, and (2) validate these models based on metrics that are meaningful to them (instead of relying on conventional metrics). The results of our user studies also show that biologists using Geono-Cluster successfully gleaned insights from the data and learned about their data at the end of their exploration process, and not just construct a set of clustering models.

**Extending current interactive clustering approaches:** Previous interactive clustering tools (e.g., Clusterophile [17] etc.) follow a top-down approach, where users define cluster parameters through control panels to build cluster models. To interact with these tools, users should know various cluster parameters and how to adjust them. Instead, GeonoCluster follows a bottom-up approach in which it enables users to apply their domain knowledge by interacting at the data instance level (without having to learn model parameters or metrics), and the system infers users' intent from the given demonstrations to recommend cluster models. Also, unlike other interactive clustering tools, our work solves a specific problem in a domain that we discovered through working with biologists. That being said, with this work we tried our best to provide as many affordances as possible empowering the biologists' to perform the desired set of tasks to rapidly ideate many clustering models. However, we acknowledge that there may be other useful methods and guidance that can make the current workflow easier and intuitive for users to better support their analysis process.

**Model Feedback and Interpretation:** Periodic discussion and informal inputs from the biologists clarified that model interpretation and feedback (to the model) is of critical value to them. For example, when Geono-Cluster shows a set of clustering recommendations, users may need to know how they differ from each other, or what logic was implanted to define the displayed clusters. There are many ways to explain this to the user; however, we only selected methods which do not require any technical expertise from the user. Our final design explains a cluster by using a natural language-based approach to communicate the features that were used to compute the clustering distance function. In particular, we avoid showing technical information such as silhouette coefficient or exact feature weights to provide a high-

level model explanation that does not overwhelm the users with a bag of information that might not be easy to interpret. Our qualitative feedback hints that our approach made Geono-Cluster not only easy to use but also an engaging tool to continue data exploration by rapidly testing different ideas to cluster the data.

**Cluster Model Comparison:** While representing multiple clustering results show different ways to partition the data, model comparison to understand trade-offs between these clustering options is critical. However, in our current prototype we do not support explicit cluster model comparison. For example, users cannot perform a pairwise comparison of two cluster models side by side [55], or they cannot select a few chosen cluster models to see the results in a way which facilitates direct comparison. Based on our interviews with the biologists, comparing cluster models was not posed as a requirement to us. Therefore, we deliberately did not include cluster model comparison as one of the design goals of the system. However, as visual analytics researchers, we understand that being able to compare multiple cluster models, may positively aid model selection and enhance the tools use case.

**Limited Model Explanation:** Geono-Cluster explains a cluster model by highlighting the top $k$ features that were used to compute the underlying distance function using a natural language expression. Though the simplicity of the explanation is helpful for non-experts, in certain cases this may pose as a very limited explanation of a clustering model. For example, two cluster models may be based on the same set of features, but the defined clusters are strikingly different. In this case, users may get confused to interpret the difference between these models.

**Scalability:** The current interface and the supported interactions (i.e., split and merge technique) is tested with 3000 (approximately) data items. However, we understand that as the size of the data grows, the interaction techniques such as drag-and-drop interaction and lasso-selection tool may be less responsive. In the user study, P6 noted that the lasso-selection was less effective for large clusters when the data items became too small to select or notice (often partially obscured by neighboring data items). We envision multiple ways to enhance scalability in future iterations of this tool.

## 8 Conclusion

In this paper, we introduce Geono-Cluster that is designed to help biologists visually cluster their data for exploratory analysis. The proposed technique leverage the domain knowledge of the users by allowing a demonstration based interaction methodology, which recommends multiple cluster models according to users' intent. Based on collaborative studies with biologists, we built a set of task requirements and design guidelines for our prototype. The technique shown exemplifies a model of interaction which allows non-experts in data science interactively construct clustering models by specifying their preferences. This spares them the burden of going through layers of menus and control panels to transform their expectations to outputs or to comprehend complex model parameters or metrics to find the right clustering model. Our study provides valuable lessons for researchers who design visual clustering tools for biologists.

## References


[1] R. Nugent and M. Meila, "An overview of clustering applied to molecular biology," in *Statistical methods in molecular biology*. Springer, 2010, pp. 369–404.

[2] G. M. Downs and J. M. Barnard, "Clustering methods and their uses in computational chemistry," *Reviews in computational chemistry*, vol. 18, pp. 1–40, 2002.

[3] D. J. Bartholomew, F. Steele, J. Galbraith, and I. Moustaki, *Analysis of multivariate social science data*. Chapman and Hall/CRC, 2008.

[4] M. desJardins, J. MacGlashan, and J. Ferraioli, "Interactive visual clustering," in *Proceedings of the 12th International Conference on Intelligent User Interfaces*, ser. IUI '07. New York, NY, USA: ACM, 2007, pp. 361–364.

[5] M. Cavallo and Ã. Demiralp, "Clustrophile 2: Guided visual clustering analysis," *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2018.

[6] K. Chen and L. Liu, "Vista: Validating and refining clusters via visualization," *Information Visualization*, vol. 3, no. 4, pp. 257–270, Dec. 2004.

[7] J. Seo and B. Shneiderman, "Interactively exploring hierarchical clustering results [gene identification]," *Computer*, vol. 35, no. 7, pp. 80–86, July 2002.

[8] N. Cao, D. Gotz, J. Sun, and H. Qu, "Dicon: Interactive visual analysis of multidimensional clusters," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2581–2590, Dec 2011.

[9] D. H. Jeong, C. Ziemkiewicz, B. Fisher, W. Ribarsky, and R. Chang, "ipca: An interactive system for pca-based visual analytics," in *Proceedings of the 11th Eurographics / IEEE - VGTC Conference on Visualization*, ser. EuroVis'09, Chichester, UK, 2009, pp. 767–774.

[10] B. C. Kwon, B. Eysenbach, J. Verma, K. Ng, C. D. 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, Jan 2018.

[11] A. Endert, L. Bradel, and C. North, "Beyond control panels: Direct manipulation for visual analytics," *IEEE Computer Graphics and Applications*, vol. 33, no. 4, pp. 6–13, July 2013.

[12] B. Saket, H. Kim, E. T. Brown, and A. Endert, "Visualization by demonstration: An interaction paradigm for visual data exploration," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 331–340, Jan. 2017.

[13] M. Sedlmair, M. Meyer, and T. Munzner, "Design study methodology: Reflections from the trenches and the stacks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2431–2440, 2012.

[14] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "Biogrid: a general repository for interaction datasets," *Nucleic acids research*, vol. 34, no. suppl_1, pp. D535–D539, 2006.

[15] S. K. Card, G. G. Robertson, and J. D. Mackinlay, "The information visualizer, an information workspace," in *Proceedings of the SIGCHI Conference on Human factors in computing systems*. ACM, 1991, pp. 181–186.

[16] J. Wenskovitch and C. North, "Observation-level interaction with clustering and dimension reduction algorithms," in *Proceedings of the 2Nd Workshop on Human-In-the-Loop Data Analytics*, ser. HILDA'17. New York, NY, USA: ACM, 2017, pp. 14:1–14:6.

[17] Ç. Demiralp, "Clustrophile: A tool for visual clustering analysis," *CoRR*, vol. abs/1710.02173, 2017. [Online]. Available: http://arxiv.org/abs/1710.02173

[18] D. Sacha, M. Kraus, J. Bernard, M. Behrisch, T. Schreck, Y. Asano, and D. A. Keim, "Somflow: Guided exploratory cluster analysis with self-organizing maps and analytic provenance," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 120–130, Jan. 2018.

[19] S. Basu, D. Fisher, S. M. Drucker, and H. Lu, "Assisting users with clustering tasks by combining metric learning and classification," in *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, ser. AAAI'10. AAAI Press, 2010, pp. 394–400.

[20] E. J. Nam, Y. Han, K. Mueller, A. Zelenyuk, and D. Imre, "Clustersculptor: A visual analytics tool for high-dimensional data," in *2007 IEEE Symposium on Visual Analytics Science and Technology*, Oct 2007, pp. 75–82.

[21] Y. Hu, E. E. Milios, and J. Blustein, "Interactive feature selection for document clustering," in *Proceedings of the 2011 ACM Symposium on Applied Computing*, ser. SAC '11. New York, NY, USA: ACM, 2011, pp. 1143–1150.

[22] J. Liu, E. T. Brown, R. Chang, and C. E. Brodley, "Dis-function: Learning distance functions interactively," in *Proceedings of the 2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, ser. VAST '12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 83–92.

[23] D. Guo, "Coordinating computational and visual approaches for interactive feature selection and multivariate clustering," *Information Visualization*, vol. 2, no. 4, pp. 232–246, Dec. 2003.

[24] J. Z. Self, M. Dowling, J. Wenskovitch, I. Crandell, M. Wang, L. House, S. Leman, and C. North, "Observation-level and parametric interaction for high-dimensional data analysis," *ACM Trans. Interact. Intell. Syst.*, vol. 8, no. 2, pp. 15:1–15:36, Jun. 2018.

[25] A. Dubey, I. Bhattacharya, and S. Godbole, "A cluster-level semi-supervision model for interactive clustering," in *Machine Learning and Knowledge Discovery in Databases*, J. L. Balcázar, F. Bonchi, A. Gionis, and M. Sebag, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 409–424.

[26] S. Datta and E. Adar, "Communitydiff: Visualizing community clustering algorithms," *ACM Trans. Knowl. Discov. Data*, vol. 12, no. 1, pp. 11:1–11:34, Jan. 2018. [Online]. Available: http://doi.acm.org/10.1145/3047009

[27] H. Lee, J. Kihm, J. Choo, J. Stasko, and H. Park, "iVisClustering: An Interactive Visual Document Clustering via Topic Modeling," *Computer Graphics Forum*, 2012.

[28] A. Lex, H. Schulz, M. Streit, C. Partl, and D. Schmalstieg, "Visbricks: Multiform visualization of large, inhomogeneous data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2291–2300, Dec 2011.

[29] A. Lex, M. Streit, H.-J. Schulz, C. Partl, D. Schmalstieg, P. Park, and N. Gehlenborg, "Stratomex: Visual analysis of large-scale heterogeneous genomics data for cancer subtype characterization," *Comput. Graph. Forum*, vol. 31, no. 3pt3, pp. 1175–1184, Jun. 2012.

[30] G. G. Jianping Zhou, Shawn Konecni, "Visually comparing multiple partitions of data with applications to clustering," pp. 7243 – 7243 – 12, 2009.

[31] M. Kern, A. Lex, N. Gehlenborg, and C. R. Johnson, "Interactive visual exploration and refinement of cluster assignments," *BMC Bioinformatics*, vol. 18, no. 1, p. 406, apr 2017.

[32] S. L'Yi, B. Ko, D. Shin, Y.-J. Cho, J. Lee, B. Kim, and J. Seo, "Xclusim: a visual analytics tool for interactively comparing multiple clustering results of bioinformatics data," *BMC Bioinformatics*, vol. 16, no. 11, p. S5, Aug 2015.

[33] A. Cypher and D. C. Halbert, *Watch what I do: programming by demonstration*. MIT press, 1993.

[34] J. Lin, J. Wong, J. Nichols, A. Cypher, and T. A. Lau, "End-user programming of mashups with vegemite," in *Proceedings of the 14th International Conference on Intelligent User Interfaces*, ser. IUI '09. New York, NY, USA: ACM, 2009, pp. 97–106.

[35] S. Kandel, A. Paepcke, J. Hellerstein, and J. Heer, "Wrangler: Interactive visual specification of data transformation scripts," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '11. New York, NY, USA: ACM, 2011, pp. 3363–3372.

[36] M. M. Zloof, "Query by example," in *AFIPS National Computer Conference*, 1975.

[37] B. Kondo and C. Collins, "Dimpvis: Exploring time-varying information visualizations by direct manipulation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 2003–2012, Dec 2014.

[38] A. Sarvghad, B. Saket, A. Endert, and N. Weibel, "Embedded merge & split: Visual adjustment of data grouping," *IEEE Transactions on Visualization and Computer Graphics*, 2017.

[39] B. Saket and A. Endert, "Demonstrational interaction for data visualization," *IEEE Computer Graphics and Applications*, vol. 39, no. 3, pp. 67–72, May 2019.

[40] B. Saket, S. Huron, C. Perin, and A. Endert, "Investigating direct manipulation of graphical encodings as a method for user interaction," *IEEE transactions on visualization and computer graphics*, 2019.

[41] E. T. Brown, J. Liu, C. E. Brodley, and R. Chang, "Dis-function: Learning distance functions interactively," in *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, Oct 2012, pp. 83–92.

[42] A. Endert, C. Han, D. Maiti, L. House, and C. North, "Observation-level interaction with statistical models for visual analytics," in *2011 IEEE conference on visual analytics science and technology (VAST)*. IEEE, 2011, pp. 121–130.

[43] (2018) GWAS Catalog, https://www.ebi.ac.uk/gwas/. [Online]. Available: https://www.ebi.ac.uk/gwas/

[44] D. A. Norman, *Things that make us smart: Defending human attributes in the age of the machine*. Basic Books, 1993.

[45] E. Horvitz, "Principles of mixed-initiative user interfaces," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '99. New York, NY, USA: ACM, 1999, pp. 159–166.

[46] J. W. Tukey, *Exploratory data analysis*. Reading, Mass., 1977, vol. 2.

[47] S. Zolaktaf and G. Murphy, "What to learn next: Recommending commands in a feature-rich environment," 12 2015, pp. 1038–1044.

[48] E. Reed, S. Nunez, D. Kulp, J. Qian, M. P. Reilly, and A. S. Foulkes, "A guide to genome-wide association analysis and post-analytic interrogation," *Statistics in Medicine*, vol. 34, no. 28, pp. 3769–3792, 2015. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.6605

[49] B. Settles, "Active learning literature survey," University of Wisconsin–Madison, Computer Sciences Technical Report 1648, 2009. [Online]. Available: http://axon.cs.byu.edu/~martinez/classes/778/Papers/settles.activelearning.pdf

[50] "Sk learn - clustering metrics," https://scikit-learn.org/stable/modules/classes.html, accessed: 2019-07-15.

[51] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011.

[52] A. Malhi and R. X. Gao, "Pca-based feature selection scheme for machine defect classification," *IEEE Transactions on Instrumentation and Measurement*, vol. 53, no. 6, pp. 1517–1525, Dec 2004.

[53] M. Beaudouin-Lafon, "Instrumental interaction: An interaction model for designing post-wimp user interfaces," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '00. New York, NY, USA: ACM, 2000, pp. 446–453.

[54] M. Sedlmair, M. D. Meyer, and T. Munzner, "Design study methodology: Reflections from the trenches and the stacks," *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 12, pp. 2431–2440, 2012. [Online]. Available: https://doi.org/10.1109/TVCG.2012.213

[55] M. Gleicher, "Considerations for visualizing comparison," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 413–423, Jan 2018.



**Bahador Saket** is currently a fifth year PhD student at the Georgia Institute of Technology. His research combines methods from data visualization and human-computer interaction (HCI) to develop and evaluate natural interaction techniques for visual data analysis, and apply these to interactive systems in different critical domains.

**Subhajit Das** is a Ph.D. Computer Science student at the School of Interactive Computing, Georgia Institute of Technology, USA. His research interests include interactive machine learning, model optimization/selection, and designing human-in-the-loop based visual analytic systems.

**Bum Chul Kwon** is Research Staff Member at IBM Research. His research area includes visual analytics, data visualization, human-computer interaction, healthcare, and machine learning. Prior to joing IBM Research, he worked as postdoctoral researcher at University of Konstanz, Germany. He received his M.S. and Ph.D. from Purdue University in 2010 and 2013, respectively. He received his B.S. in Systems Engineering from University of Virginia in 2008.

**Alex Endert** is an Assistant Professor in the School of Interactive Computing at Georgia Tech. He directs the Visual Analytics Lab, where him and his students explore novel user interaction techniques for visual analytics. His lab often applies these fundamental advances to domains including text analysis, intelligence analysis, cyber security, decision making, and others. He received his Ph.D. in Computer Science at Virginia Tech in 2012.