Capturing User Intent when Brushing in Scatterplots
Supplementary Material

Kiran Gadhave, Jochen Görtler, Zach Cutler, Carolina Nobre,
Oliver Deussen, Miriah Meyer, Jeff Phillips, and Alexander Lex

1 Overview

We provide detailed supplementary material with this manuscript. A snapshot of this data is also archived in the osf.io project with DOI 0.17605/OSF.IO/8VPE6.

Note that we are not tracking any personal information, such as IP address, on any of these websites.

Results

- Screenshots of example results in Section 2

- Live-demo on our website: http://vdl.sci.utah.edu/predicting-intent/

- The attached supplementary video.

Study

- The live study interface http://vdl.sci.utah.edu/predicting-intent-study/.

- The detailed study results in Section 3.

- An overview of the stimuli in Section 3.

- The code to generate the stimulus data https://github.com/visdesignlab/intent-data-generation.

- The analysis code and the original data https://github.com/visdesignlab/intent-study-analysis/

- The provenance data exploration website https://vdl.sci.utah.edu/intent-study-analysis/

Formative Interviews

- Data about code frequency for the two coders is shown in Section 4.

- Pictures of the datasets and annotations generated ruing the formative interviews in Section 4.

Source Code

- The source code of our system http://vdl.sci.utah.edu/predicting-intent/.

- The source code of our provenace library https://github.com/visdesignlab/provenance-lib-core.

2   Results: Prediction and Intents

Here we demonstrate a series of selections and the corresponding predictions using various datasets. These brushes and interactions can be recreated in the live version of our prototype at `http://vdl.sci.utah.edu/predicting-intent/`.
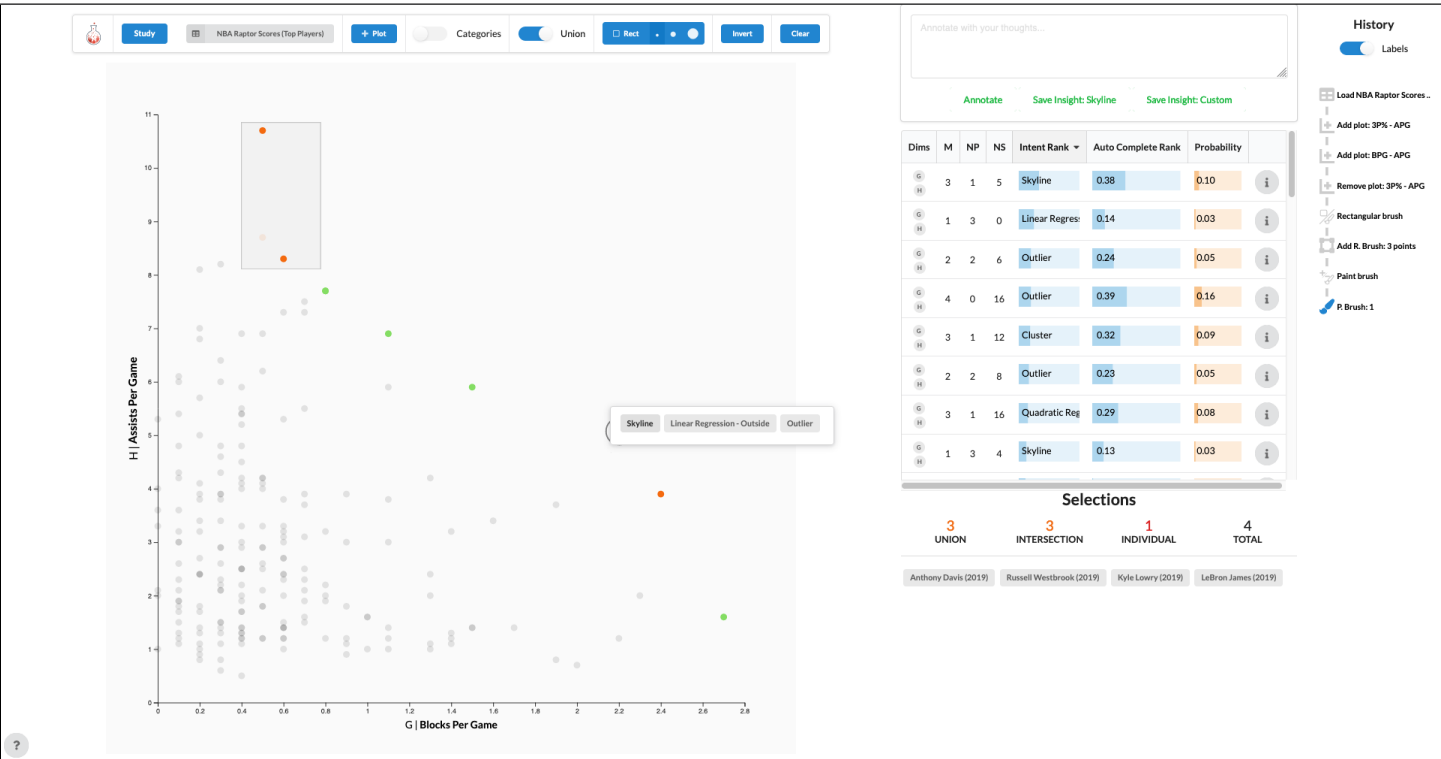

Figure S1: Visualization of assists per game versus blocks per game for the top NBA players in 2019. The selection was performed using a combination of freeform brushes and small rectangular brushes. There are four additional points that are part of the skyline but were not brushed initially. By clicking on the "Skyline" prediction, these additional points can be selected as well and submitted to the provenance graph.
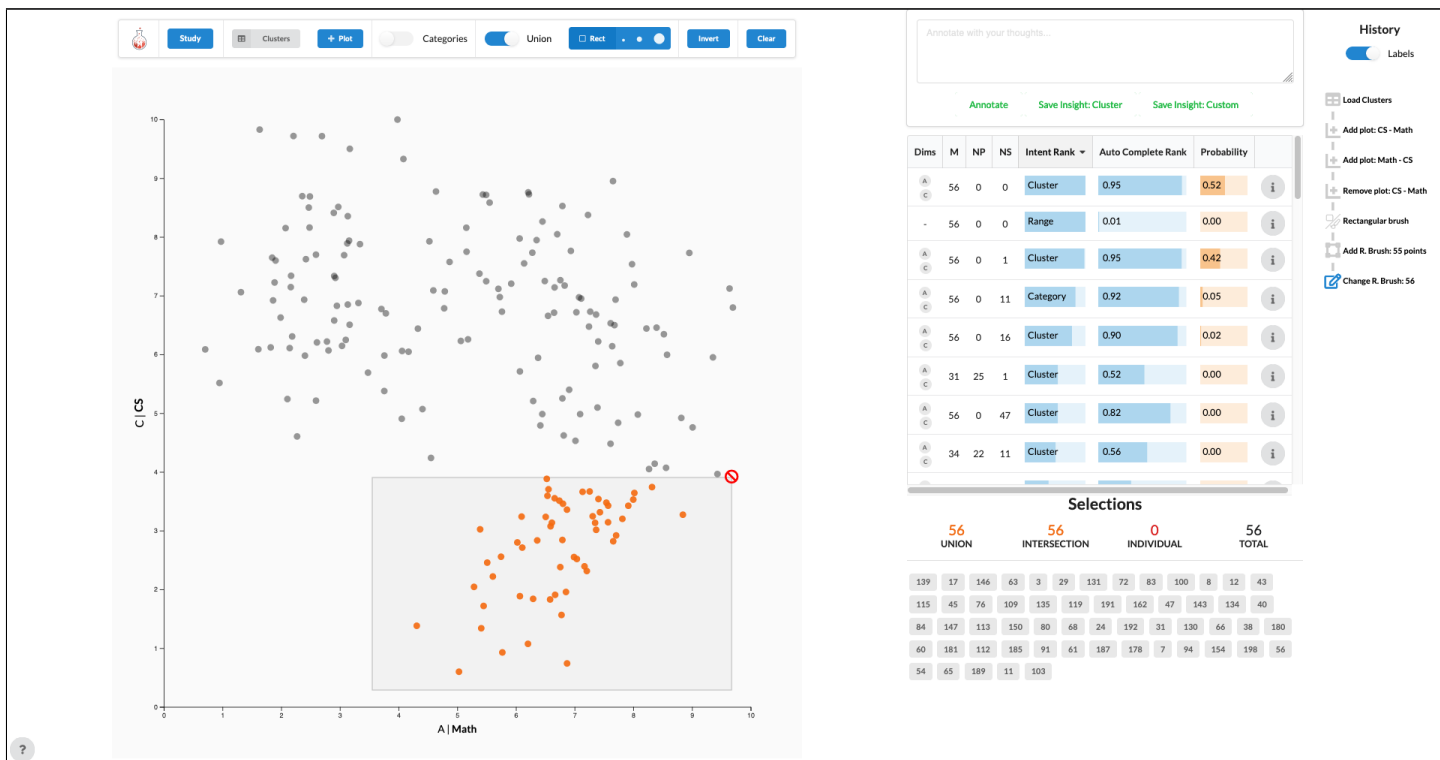
Figure S2: Selection of a cluster pattern in a synthetic dataset. All of the points belonging to the cluster are selected, as indicated by the "M", "NP", and "NS" columns in the prediction view. A second cluster was ranked on 3rd place, because one point was not selected ("NS").

Figure S3: Items that are homes in Salt Lake City that are outliers with regards to list price over taxes according to the local outlier factor algorithm. The three dark grey points were not brushed by the user but are part of this particular set of outliers. Interestingly, the point marked in green is not part of the outlier set, as indicated by the "NP" column.
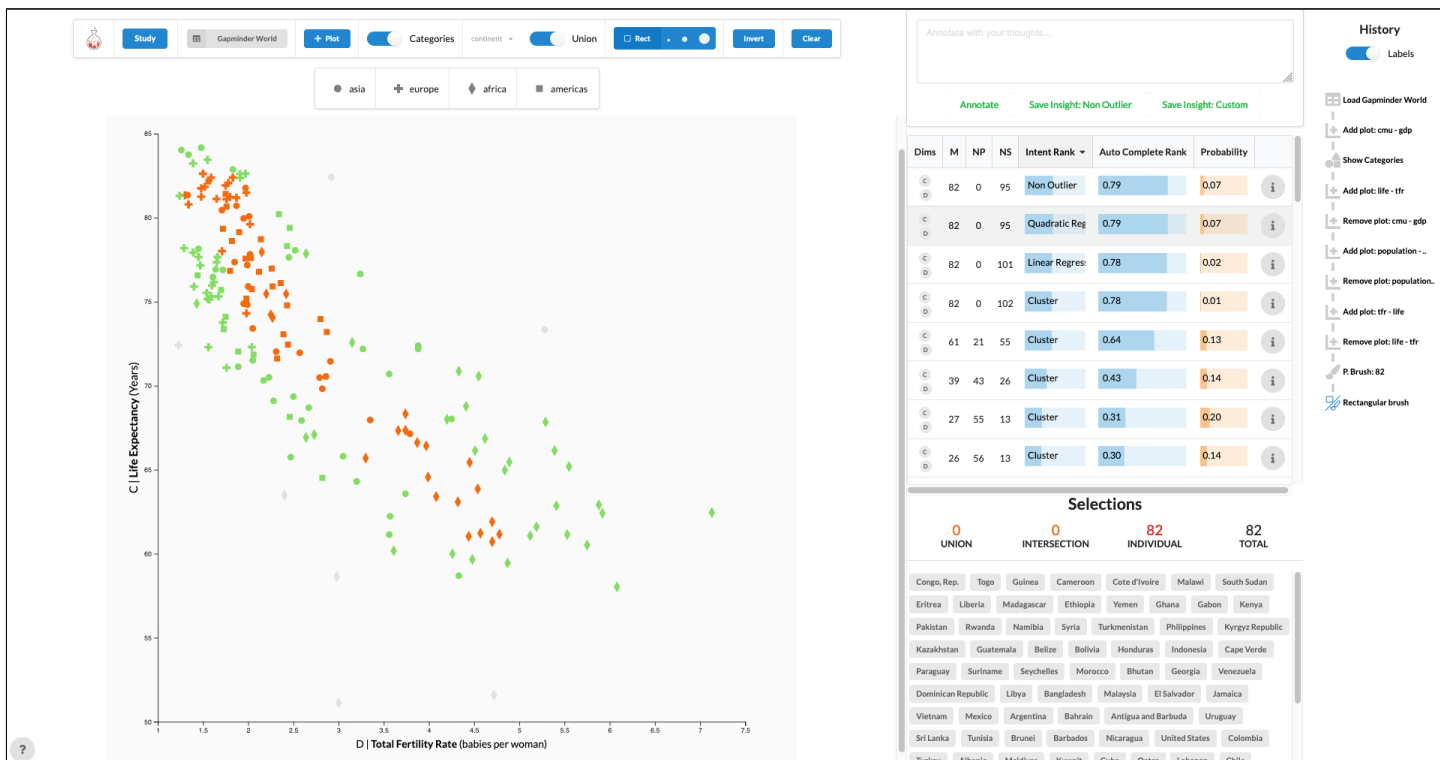
Figure S4: The total fertility rate correlates strongly with the life expectancy from the Gapminder dataset. By selecting a portion of such correlated points, linear regression and quadratic regression are both top suggestions.
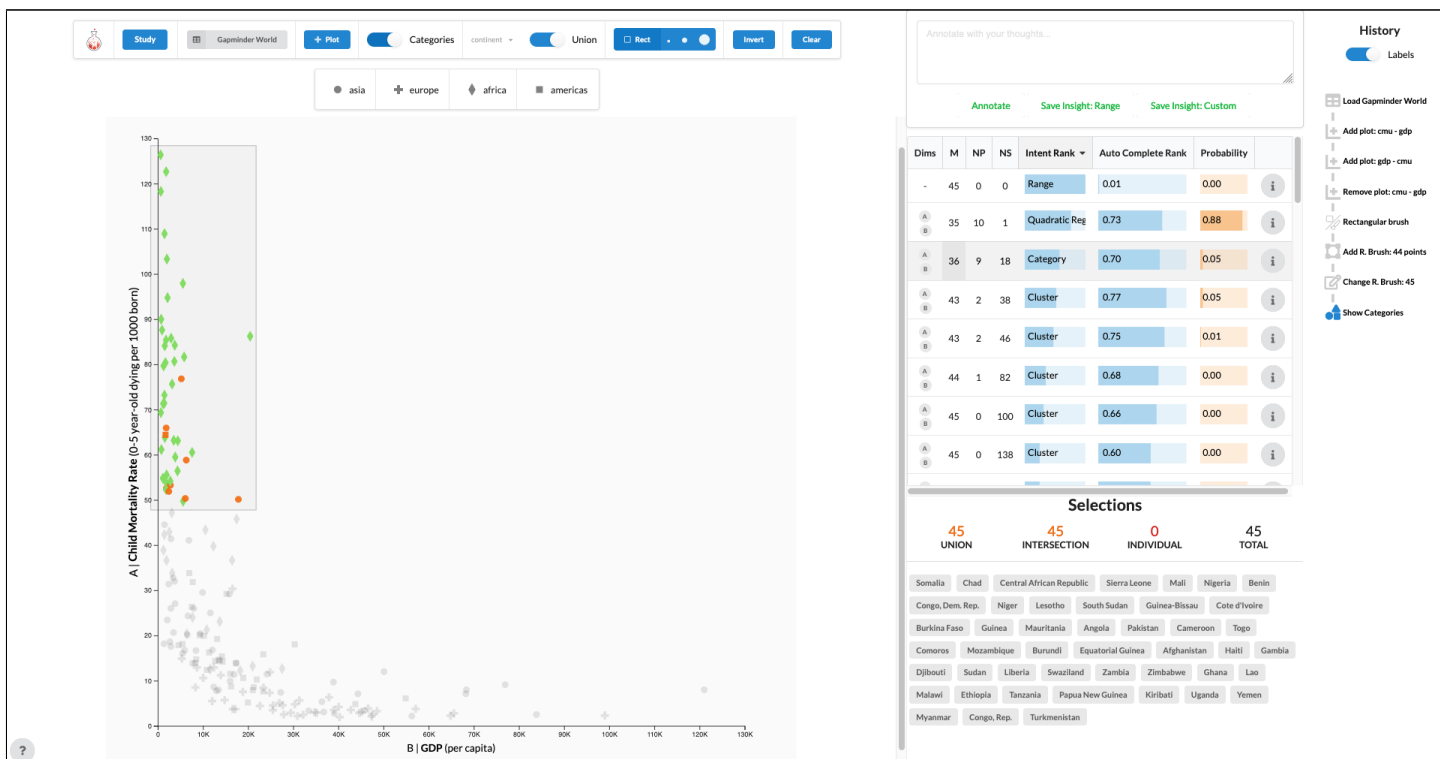

Figure S5: Visualization of Child Mortality Rate over GDP from the Gapminder dataset. Here a brush was made for child mortality larger than 50 per 1000 children. Our system ranks Range query the highest because the data can be linearly separated by this query. The third highest rank shows a dire fact. Most of these countries with high child mortality ("M", shown in green) are in Africa, computed by the Category pattern.

Figure S6: Visualization of list price versus property taxes in a real estate dataset. This example demonstrates the ability to use our approach to auto-complete a selection. An analyst has selected just seven data points (orange). Our method predicts various patterns, among them a skyline, outliers, and points that do not fit a linear regression. The intent of the analyst was to select the points that deviate from the correlation. To do that, they just have to select this prediction and apply it as their selection.
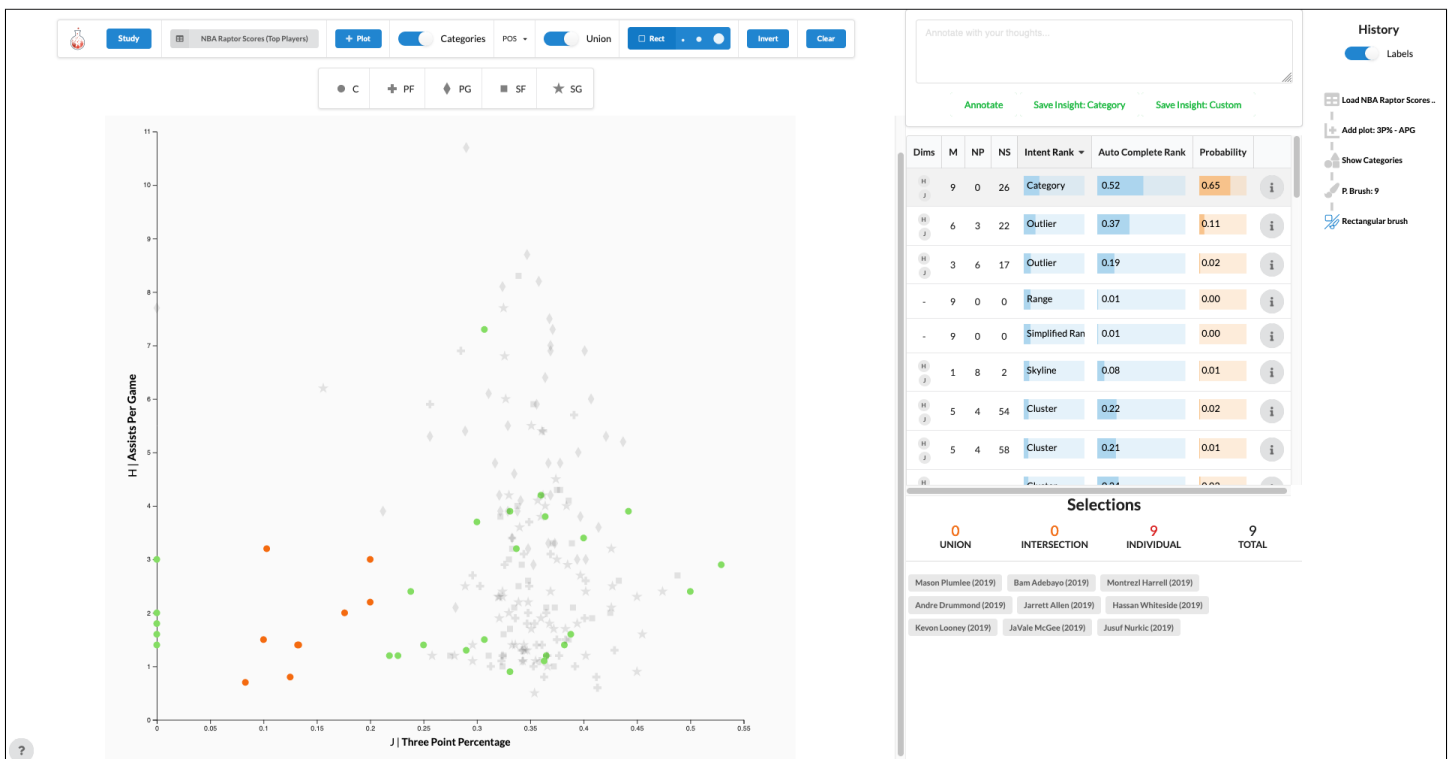

Figure S7: Visualization of three point percentage versus assists per game for top NBA players in 2019. A brush selection was made to select a small group of players who play Center (orange). Our top prediction selects all Centers, highlighting points in green that would otherwise be difficult to select. By clicking on the prediction, you can easily select all Centers.
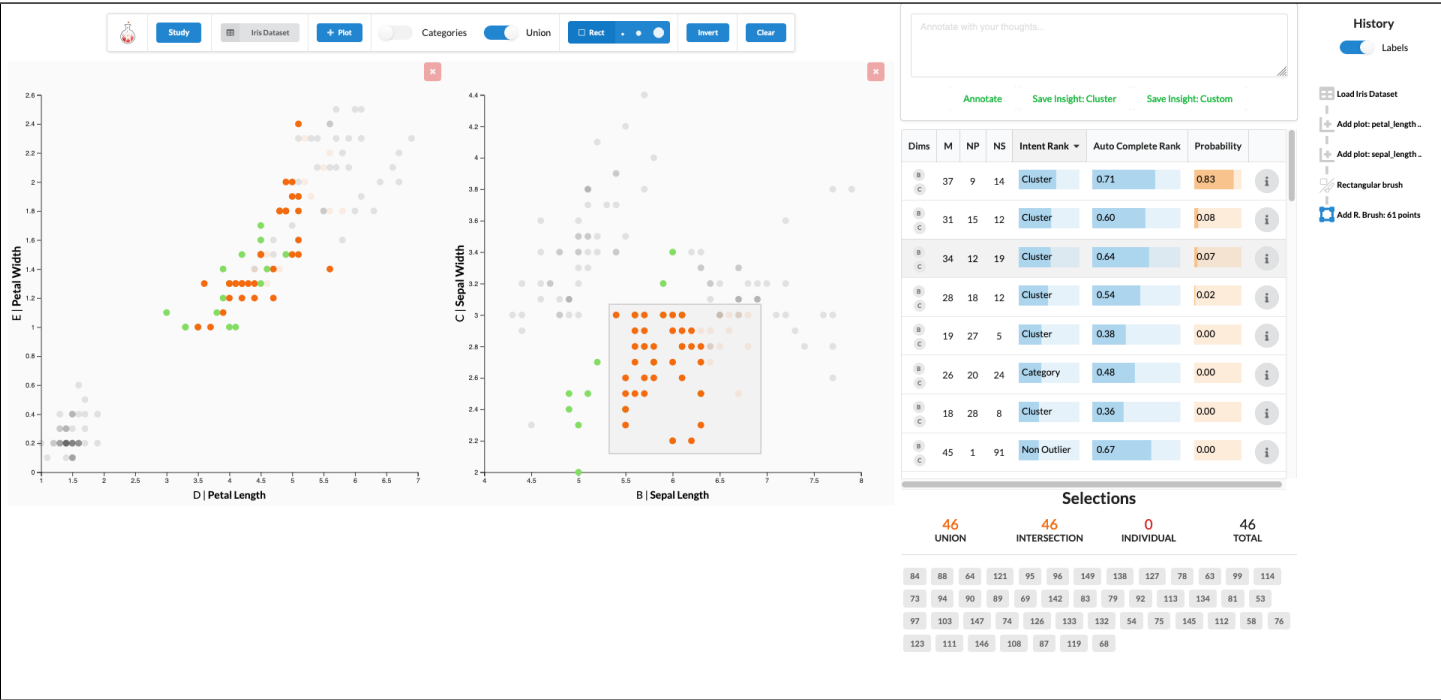
Figure S8: Visualization of the Iris dataset. The left plot shows petal width versus petal length, while the right plot show sepal width versus sepal length. This example demonstrates how our system functions with higher dimensional data.
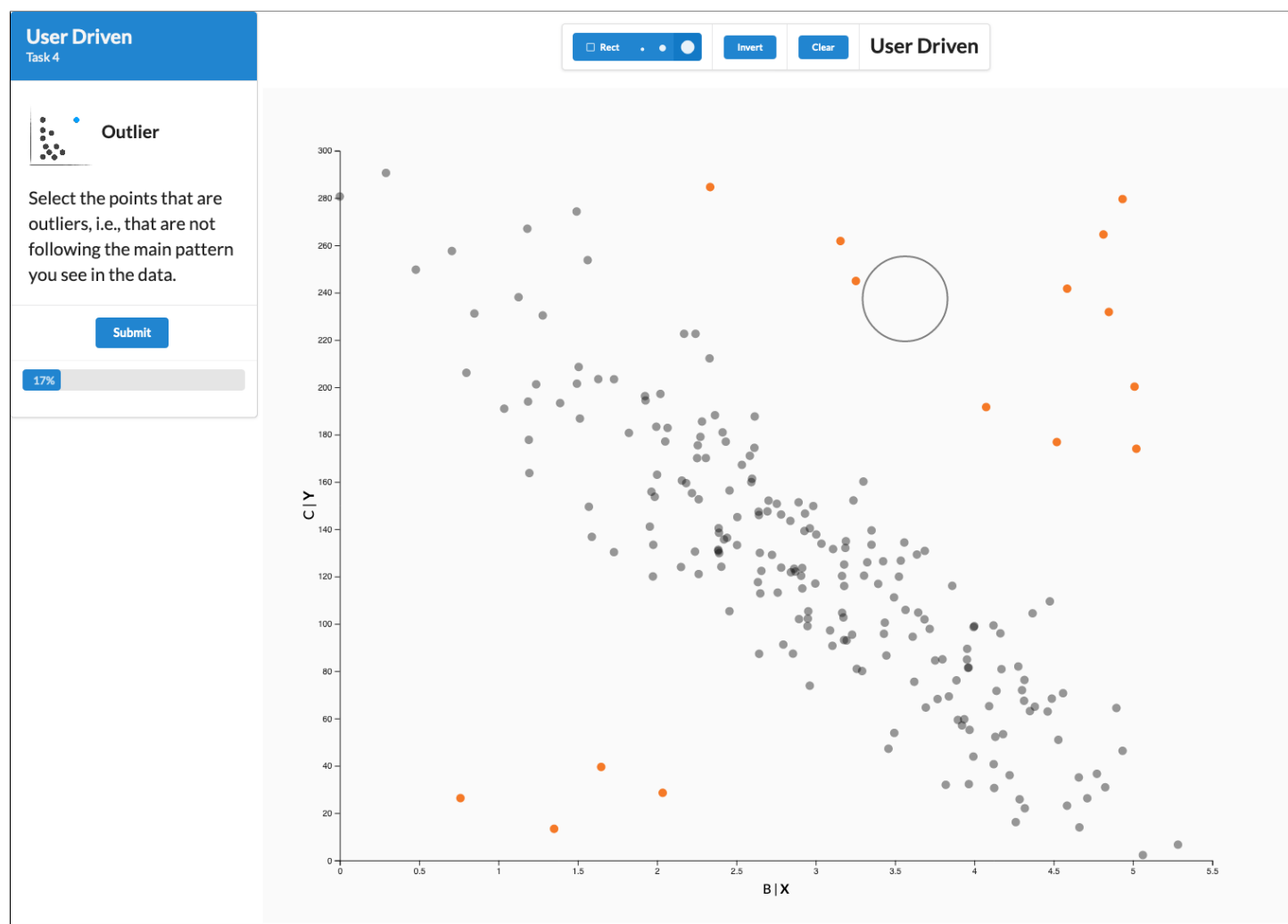
3 Quantitative Study

3.1 Stimuli

Figure S9: The user interface as used in the user-driven condition. Task instructions are shown on the right.
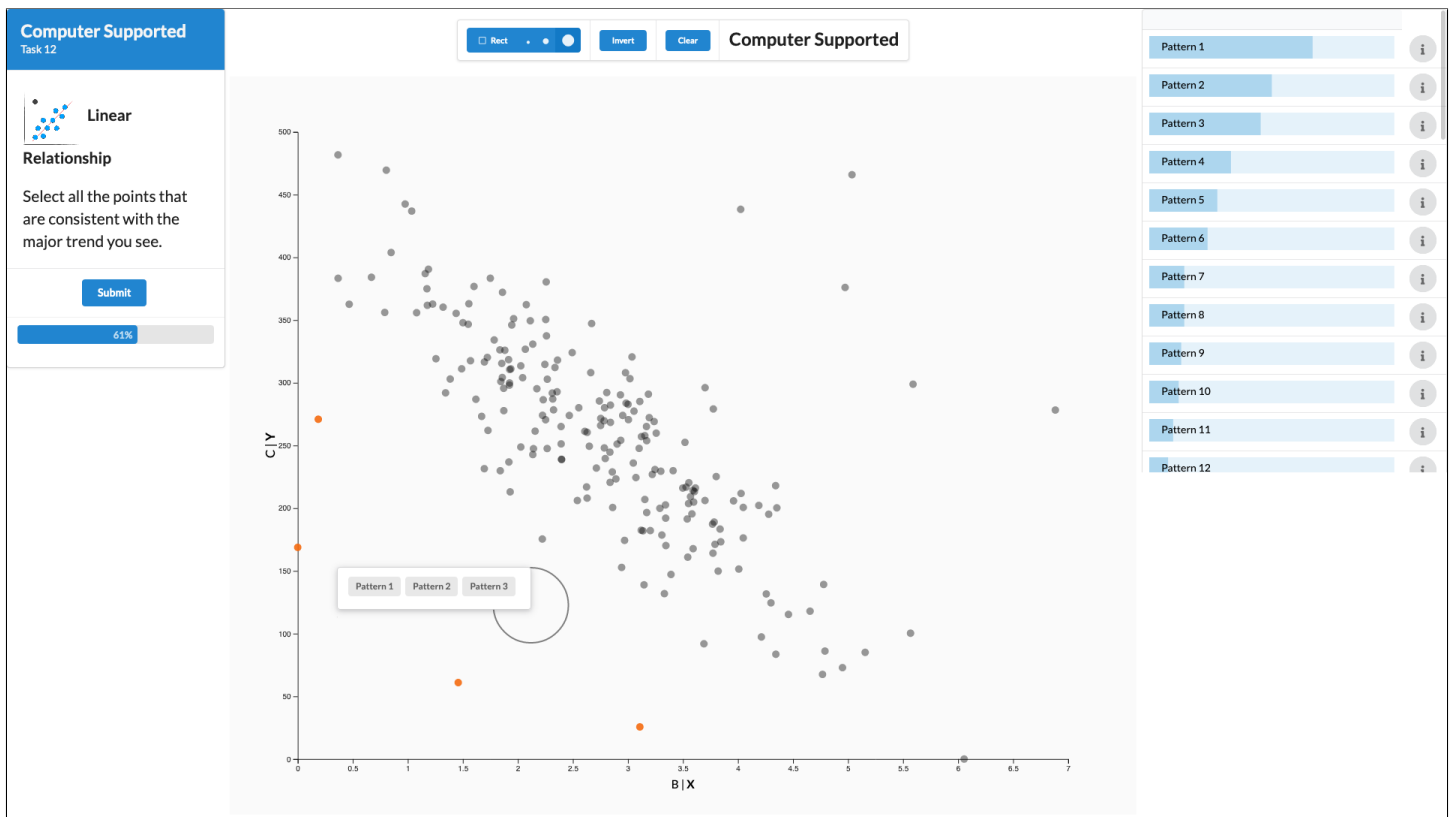
Figure S10: The user interface as used in the computer-supported condition. Task instructions are shown on the right, predictions are shown in a pop-up and on the right.
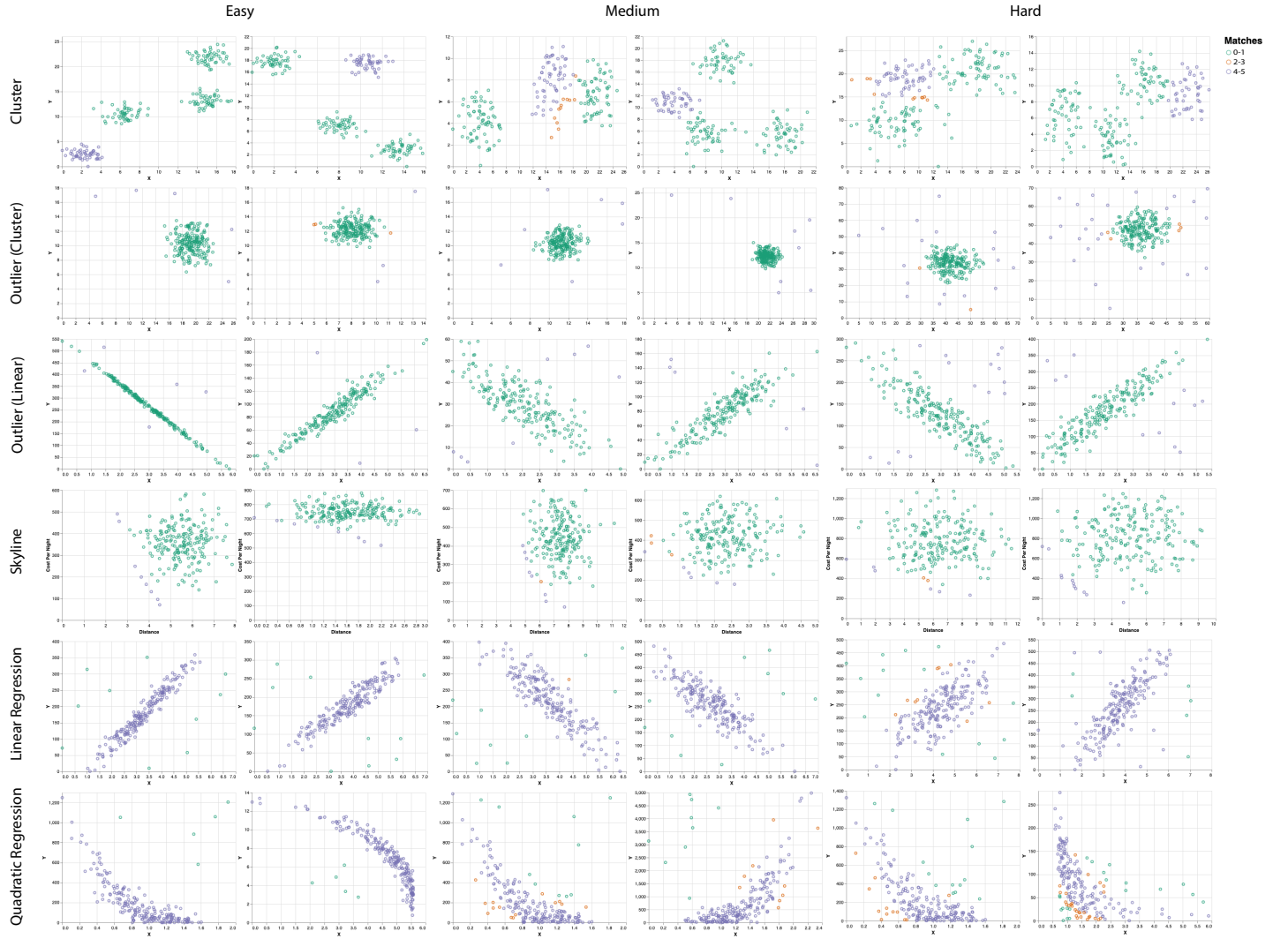
Figure S11: The datasets used as stimuli. Correct points are shown in violet, ambiguous points in orange. Other points in green.
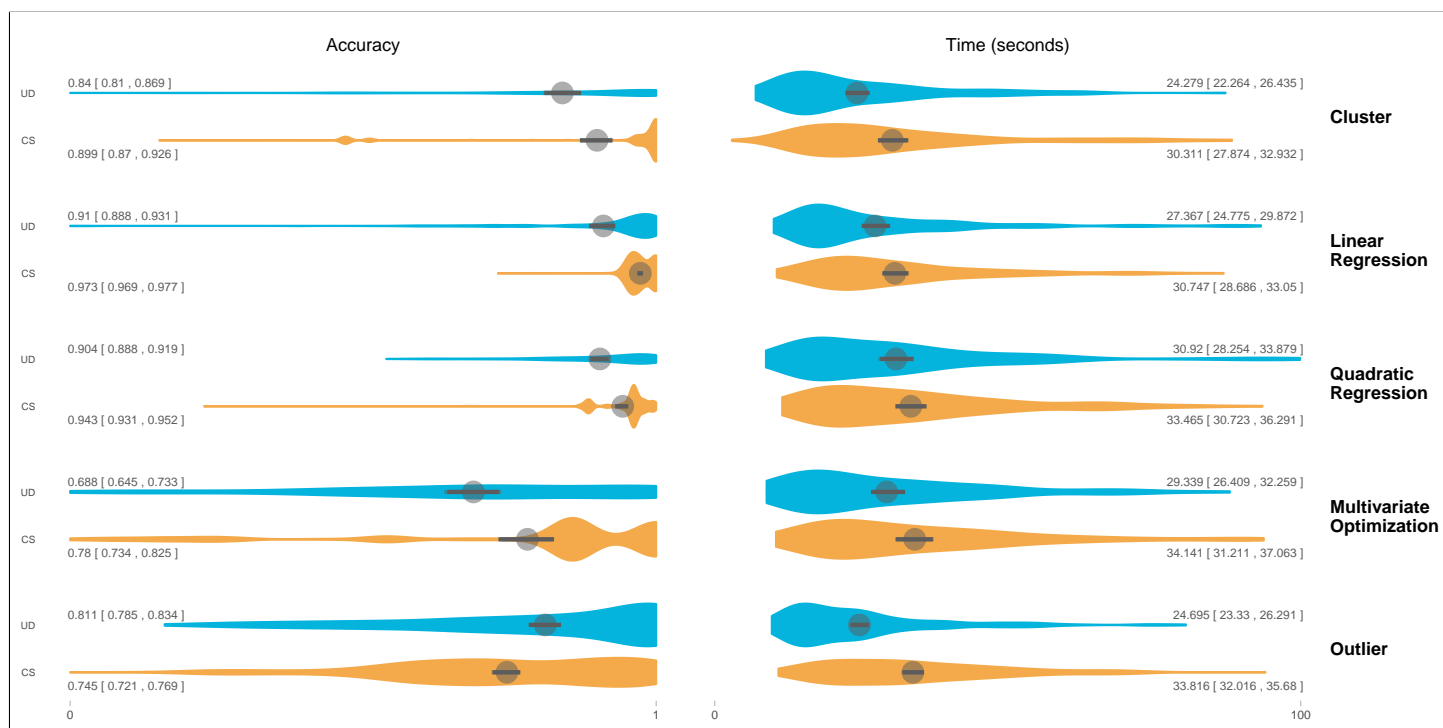
3.2 Results

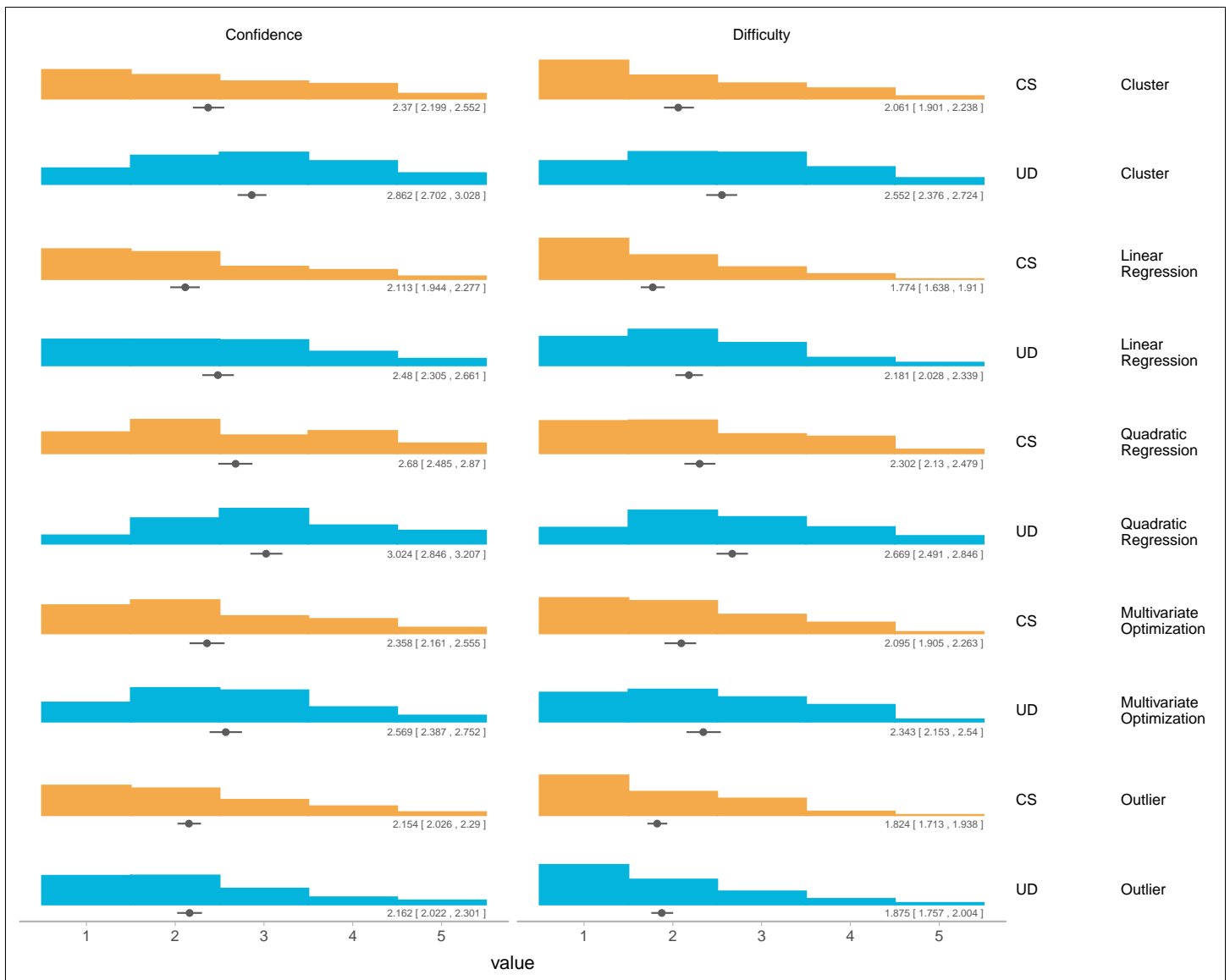**Figure S12:** Accuracy and time to completion for medium/hard task.

Figure S13: Perceived difficulty and confidence for medium/hard task.
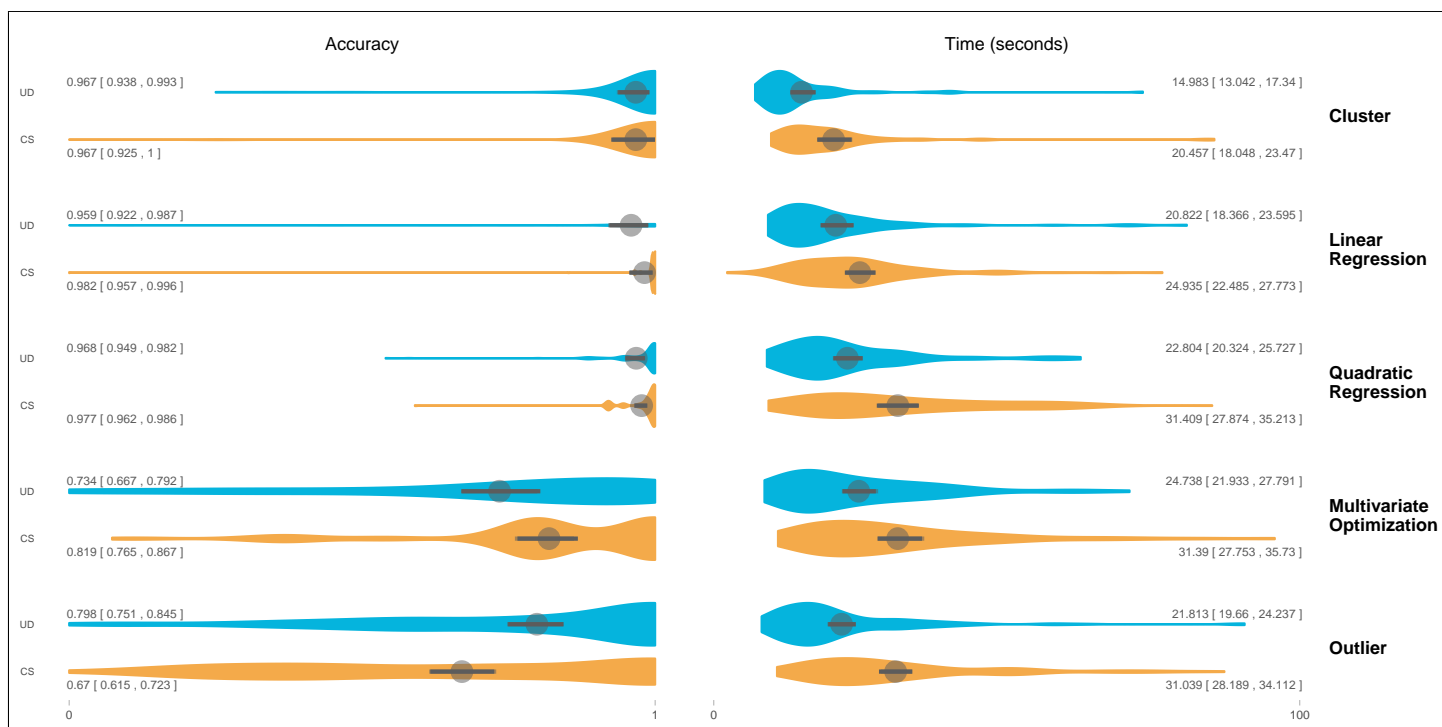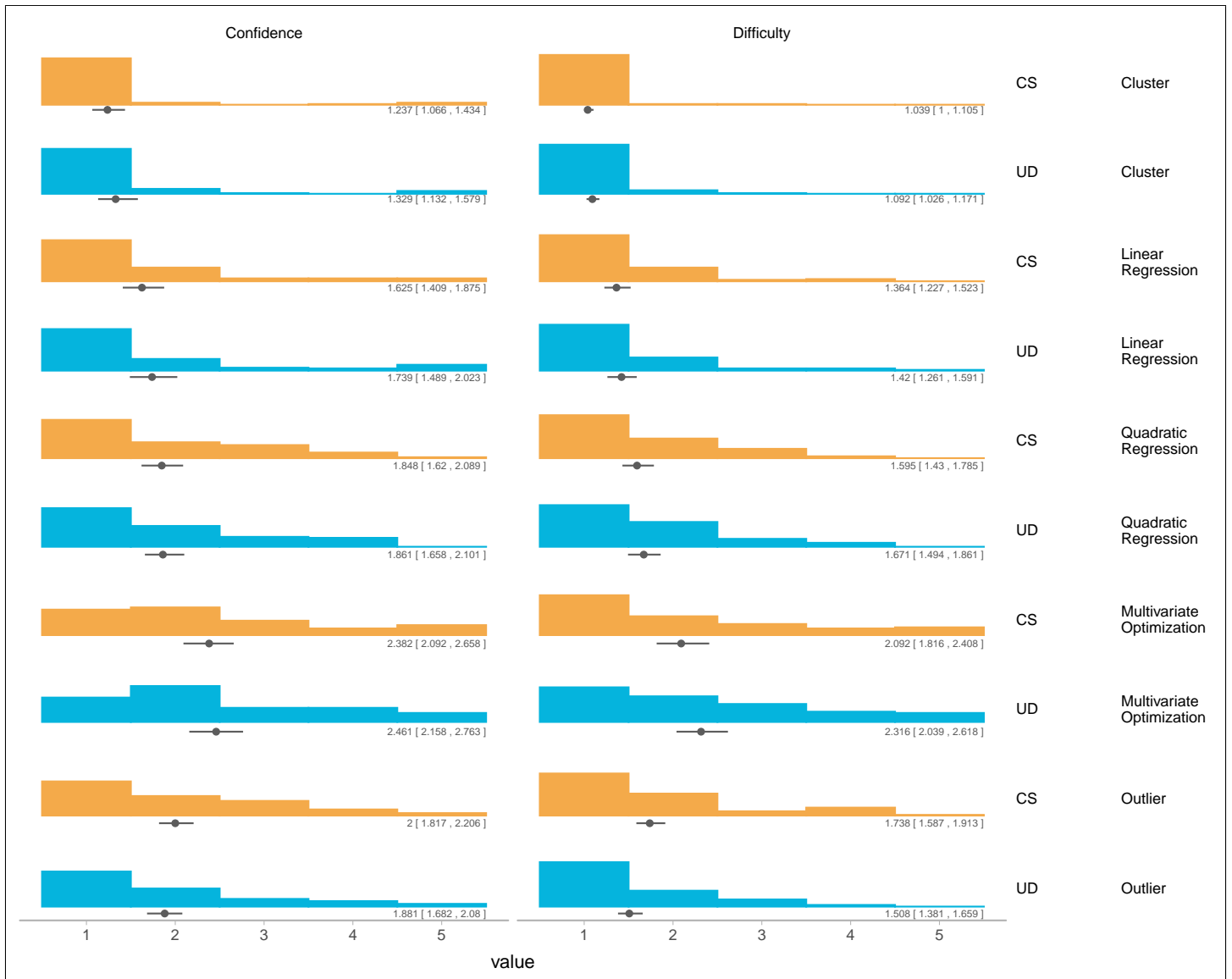
Figure S14: Accuracy and time to completion for easy tasks.

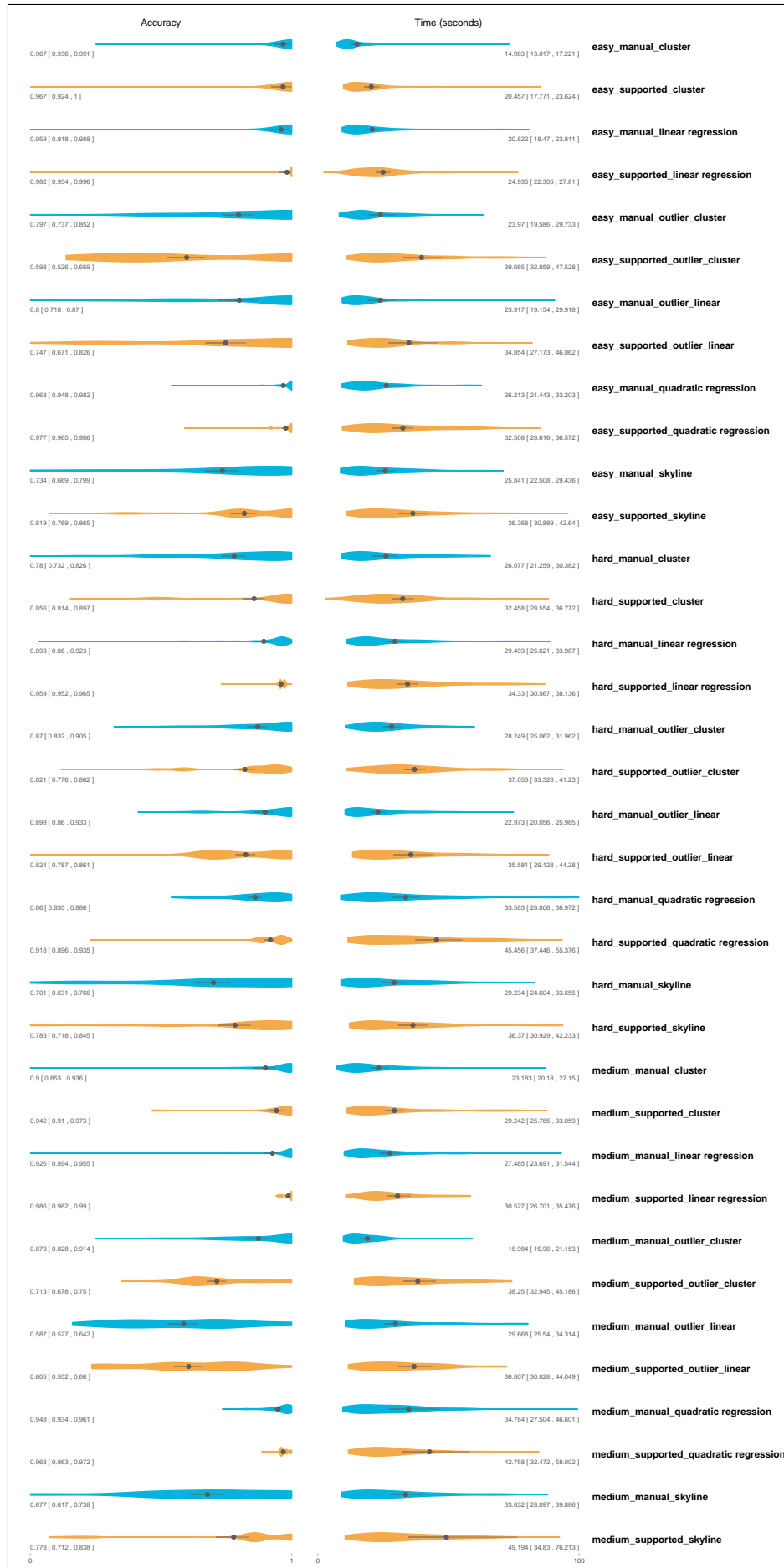**Figure S15:** Perceived difficulty and confidence for all easy tasks.

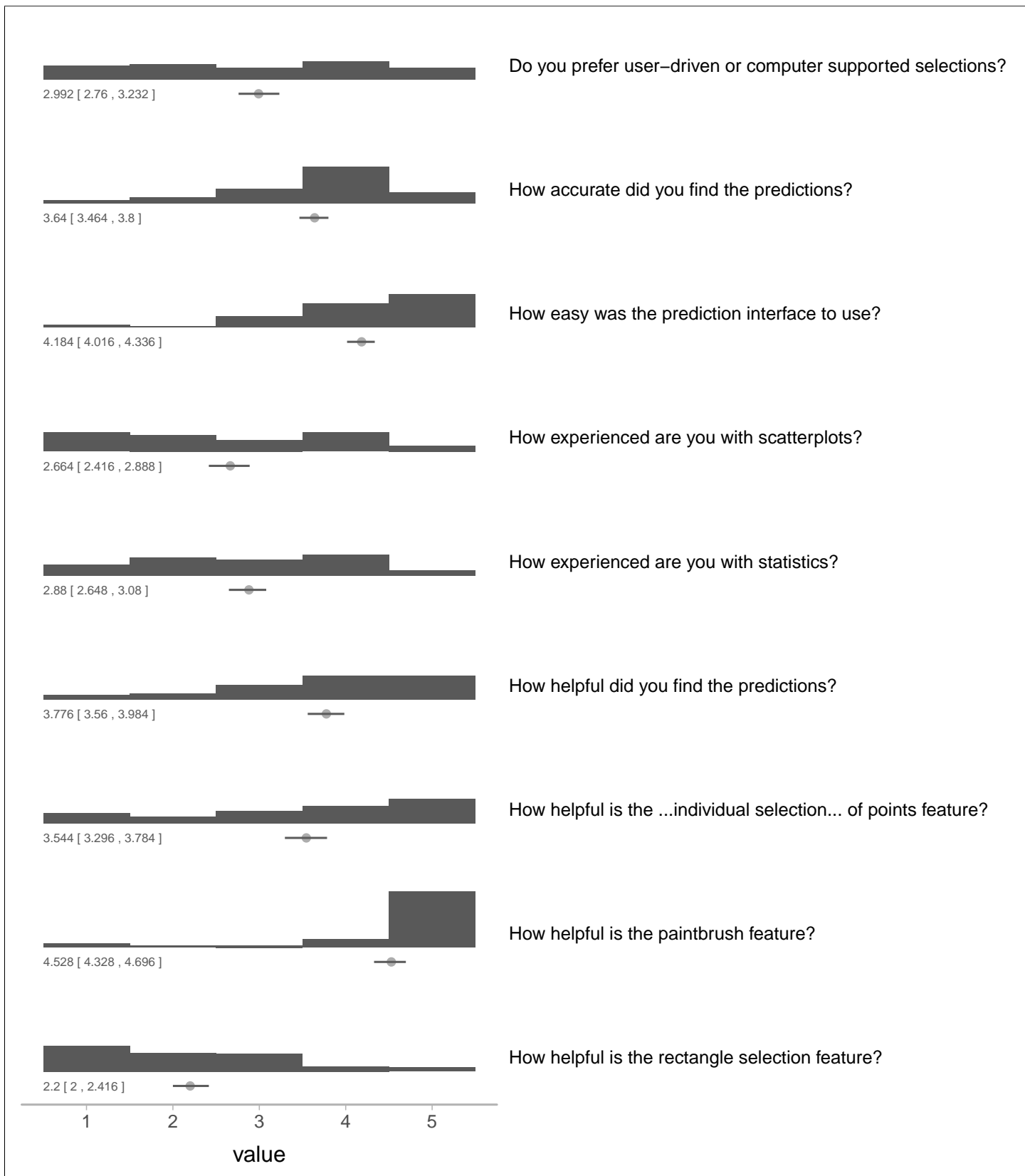**Figure S16:** Accuracy and time to completion for all tasks individually.

2.992 [2.76 , 3.232]   Do you prefer user−driven or computer supported selections?

3.64 [3.464 , 3.8]   How accurate did you find the predictions?

4.184 [4.016 , 4.336]   How easy was the prediction interface to use?

2.664 [2.416 , 2.888]   How experienced are you with scatterplots?

2.88 [2.648 , 3.08]   How experienced are you with statistics?

3.776 [3.56 , 3.984]   How helpful did you find the predictions?

3.544 [3.296 , 3.784]   How helpful is the ...individual selection... of points feature?

4.528 [4.328 , 4.696]   How helpful is the paintbrush feature?

2.2 [2 , 2.416]   How helpful is the rectangle selection feature?

value

Figure S17: Survey responses.

	Pattern asked in tasks				
Pattern predicted	Cluster	Linear Regression	Quadratic Regression	Multivariate Optimization	Outlier
Cluster	256 (223)	6 (6)	82 (78)	13 (2)	43 (25)
Linear Regression - Within	0(0)	136 (136)	54 (54)	1 (0)	4 (1)
Quadratic Regression - Within	0(0)	82 (82)	84 (84)	0(0)	2 (1)
Multivariate Optimization	0(0)	0(0)	0(0)	175 (163)	21 (10)
Linear Regression - Outside	0(0)	1 (0)	0(0)	0(0)	0(0)
Quadratic Regression - Outside	0(0)	0(0)	0(0)	2 (0)	0(0)
Non-Outlier	1 (0)	40 (39)	28 (28)	0(0)	1 (1)
Outlier	0(0)	0(0)	0(0)	22 (2)	327 (174)

Figure S18: Analyzing the matches between patterns participants were instructed to select (columns) and patterns they chose from our set of predictions. Cells show the number of times participants selected a prediction for task type and how many of those selections resulted in a correct (accuracy > 0.75) solution.

4   Formative Study

In this section we show the results of our coding of our formative interviews in Table 1. We also show images of the scatterplots marked up by domain scientists to highlight their intents.

	Multivariate Optimization	Category	Regression	Range	Cluster	Outlier
Interview 1	0\1	1\1	4\5	4\2	1\1	3\1
Interview 2	0\0	2\2	8\6	1\0	3\0	6\6
Interview 3	0\0	0\0	6\2	0\2	3\0	1\1
Interview 4	0\0	4\8	8\11	0\0	0\3	4\7
Interview 5	0\1	0\1	6\5	4\2	1\0	1\1
Interview 6	0\0	0\0	0\0	2\5	1\4	1\1
Total	0\2	7\12	32\29	11\11	9\8	16\17

Table 1: Codes for different patterns across interviews for two different coders. The first number in each cell refers to frequencies of the code in a interview found by coder 1 and the second number refers to same information by coder 2.
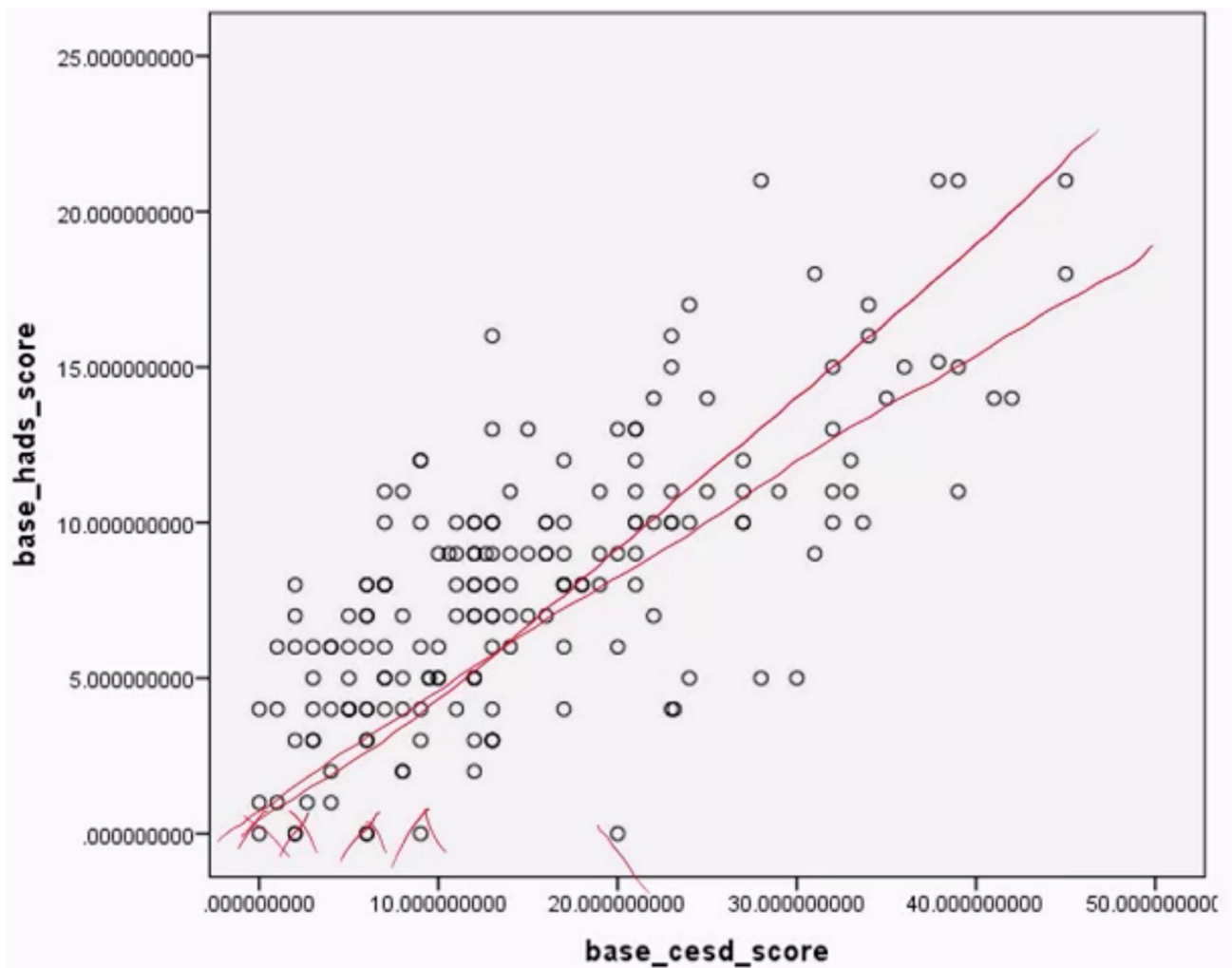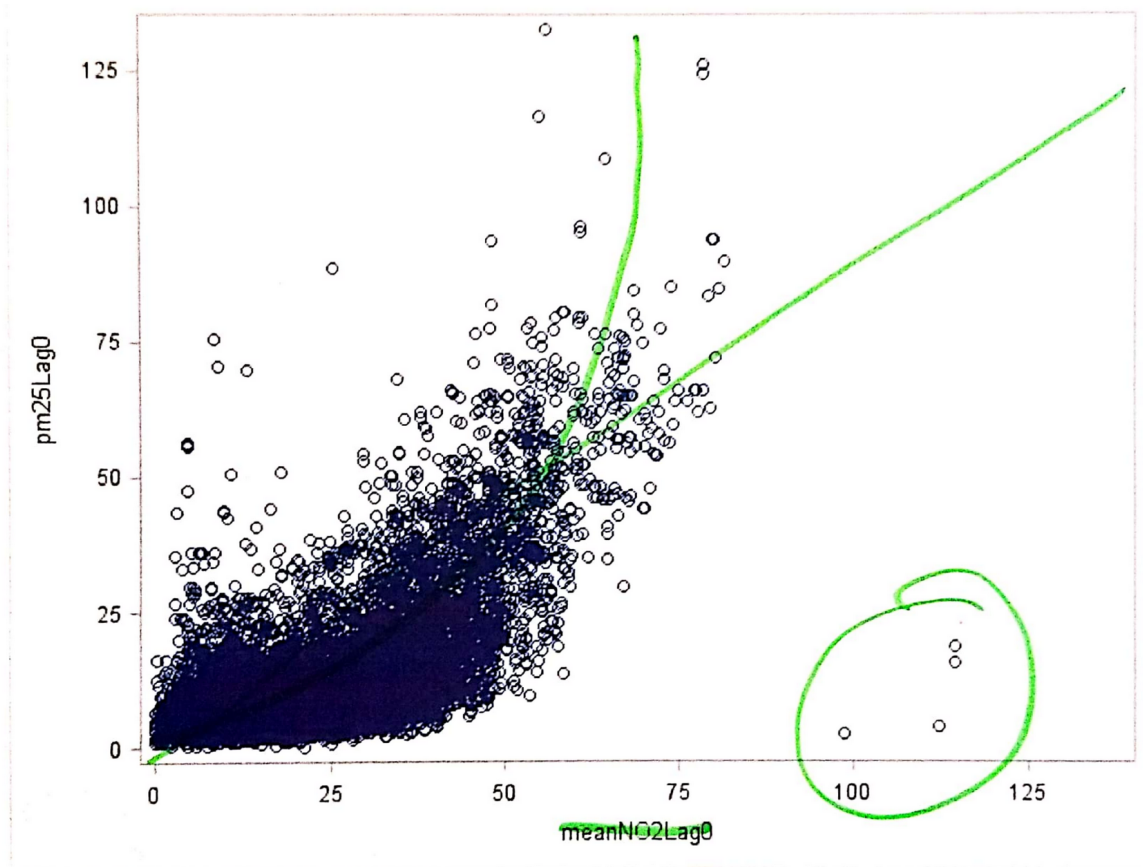
Figure S19: This plot was provided by faculty in Nursing. They look at the linear correlation of the points and the outliers at the bottom.

Figure S20: This plot was provided by faculty from Psychiatry. They are interested in the trend of the data, outliers and also color code the points with some categories like season, sex or age.
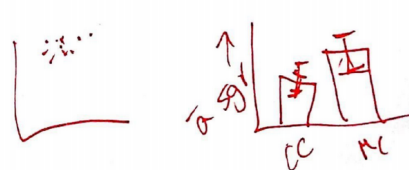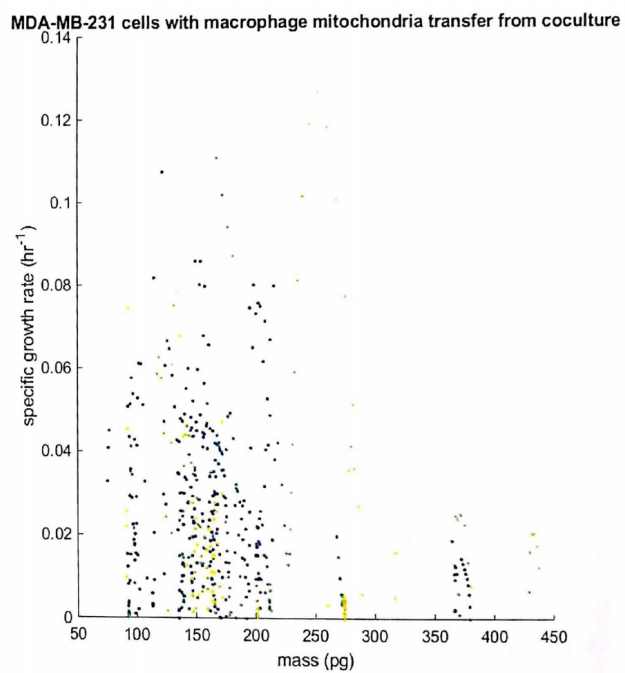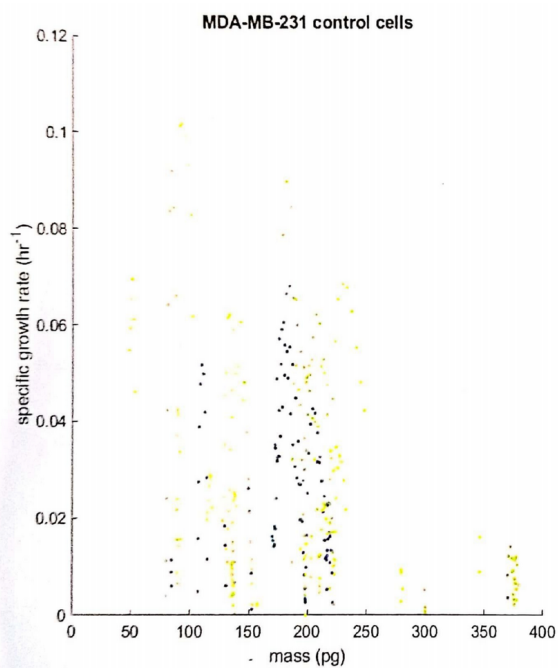
Figure S21: This plot was provided by a graduate student in Chemical Engineering. They are interested in looking at clusters of points around a range obtained from domain literature.
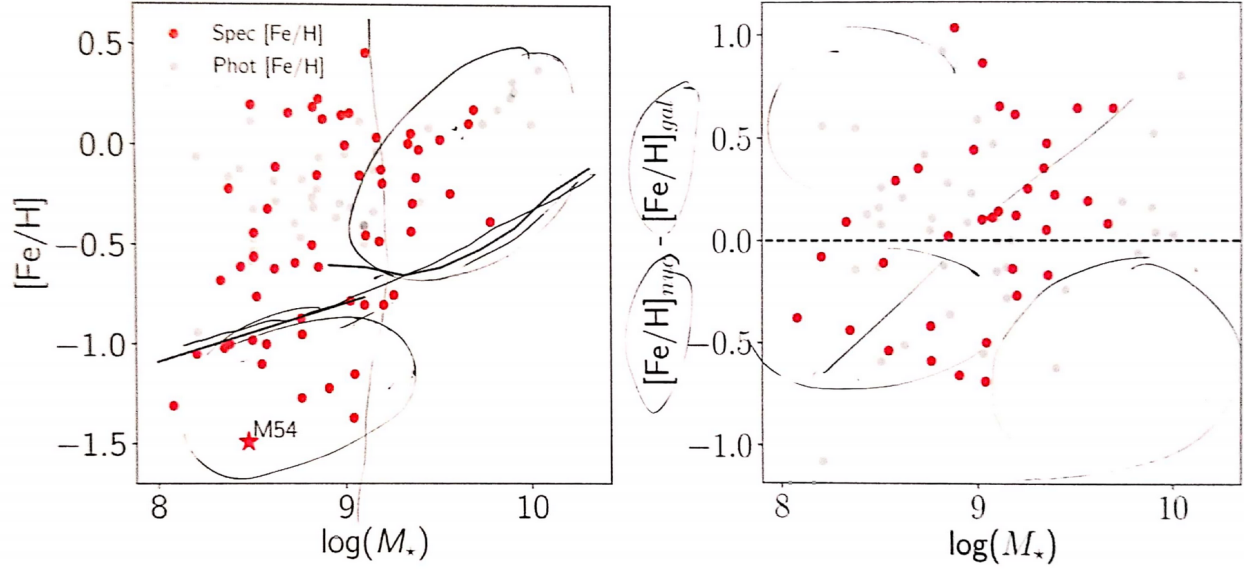
Fig. 9 Metallicity estimates for NSCs in early-type galaxies. *Left:* NSCs in higher-mass galaxies are all metal-rich, while a range of metallicities exist at lower host masses. The spectroscopic data (red symbols) include results from four studies (Koleva et al., 2009; Paudel et al., 2011; Spengler et al., 2017; Kacharov et al., 2018). The photometric metallicity estimates (grey symbols) are from Spengler et al. (2017). The estimate for the nucleus of the Sgr dwarf (labeled M 54) is that of its dominant metal-poor population (e.g. Mucciarelli et al., 2017). M 54 is the only NSC plotted here for which individual stars can be resolved and their abundances determined. The two solid lines are the mass-metallicity relations of Kirby et al. (2013) at the low mass end, and Gallazzi et al. (2005) at the high mass end. *Right:* the difference in metallicities between the NSC and their host galaxies. Data are from Koleva et al. (2009); Paudel et al. (2011); Kacharov et al. (2018). Difference values above zero indicate that the NSC is more metal-rich than its host, which predominantly occurs in high-mass galaxies.

Figure S22: This plot was provided by faculty from Astronomy. They look for trends in the data and also color code the points based on a methods of data collection. They also try to find outliers.
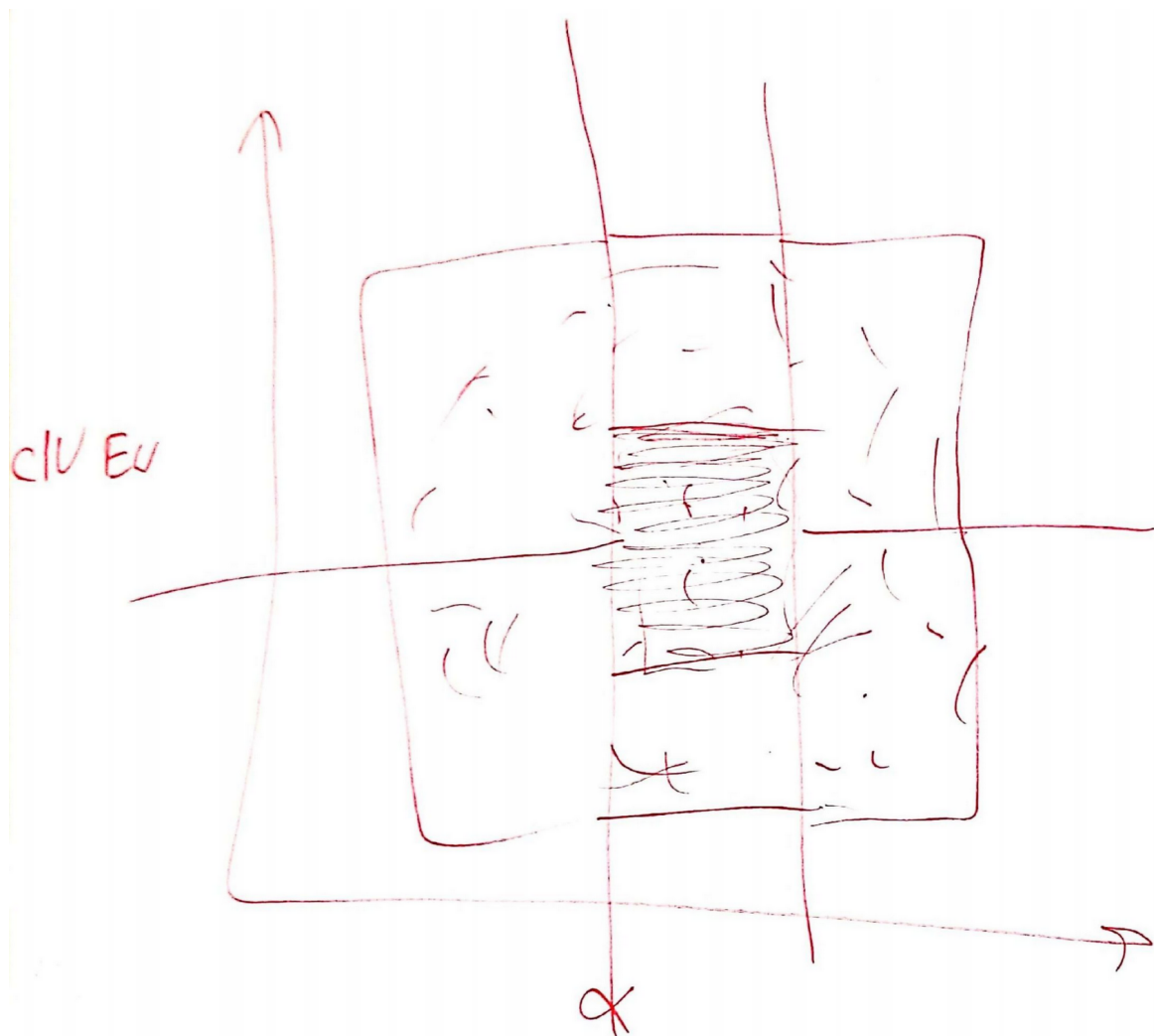
Figure S23: This plot was provided by faculty from Astronomy. They look for ranges in a scatterplot and then subset the data on that range.
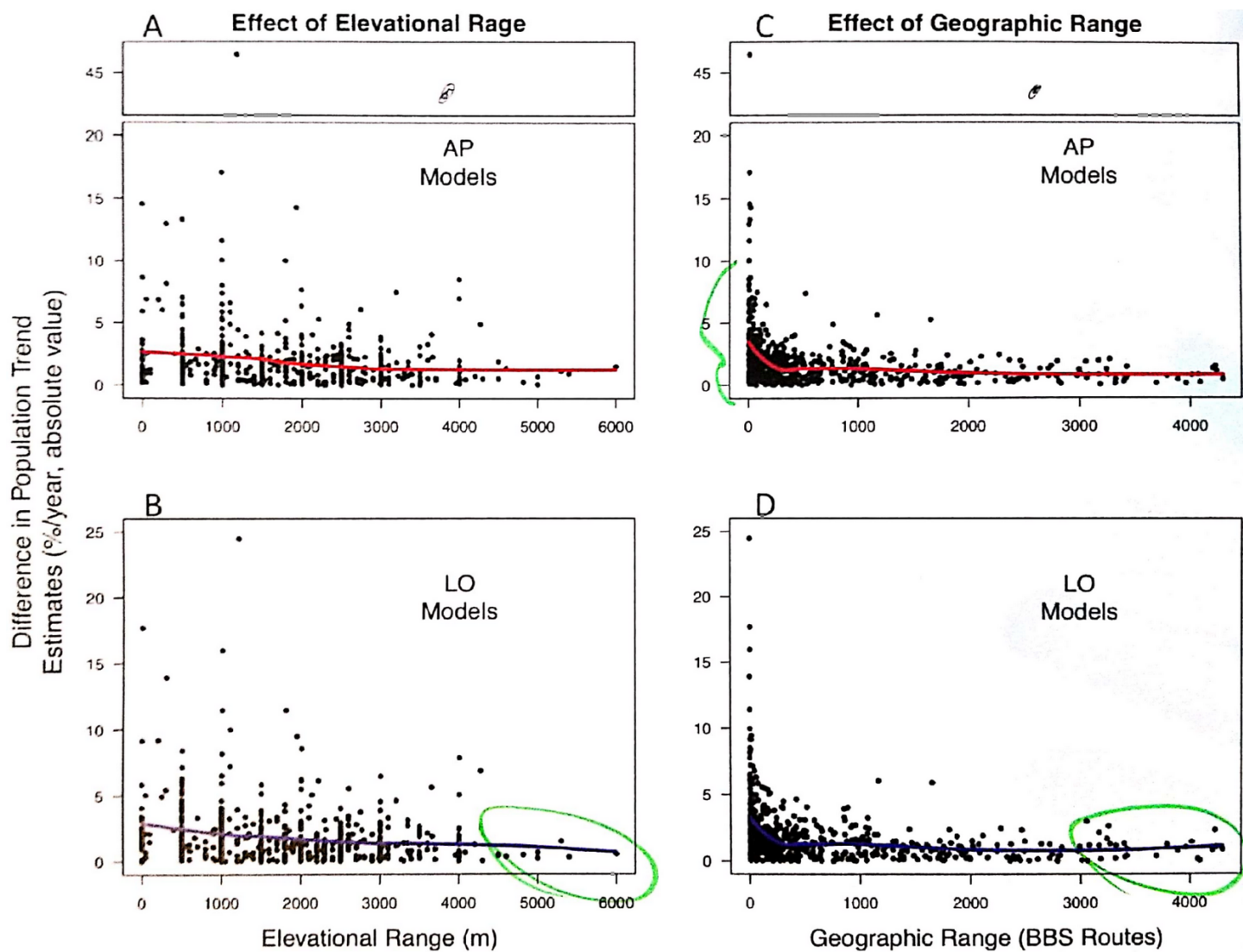
Figure S24: This plot was supplied by a research scientist from Surgery. The look for optimizations across elevation and geographic ranges.