

In Appendix A, we first describe additional details about the participatory design process, as well as domain-specific artifacts collected from contextual inquiry. Next, in Appendix B, we articulate the space of problems amenable to VQSs and describe how the sensemaking processes (introduced in Section 6) fit into different parts of the problem space. Finally, in Appendix C, we provide supplementary information regarding our analysis methods and results for the evaluation study.

A ARTIFACTS FROM PARTICIPATORY DESIGN

Information about each participants can be found in Table 3.

|                  | ID | Dataset        | Participated in Design | Position     | Years of Experience | Dataset Familiarity |
|------------------|----|----------------|------------------------|--------------|---------------------|---------------------|
| Astronomy        | A1 | DES            | ✓                      | Researcher   | 10+                 | 3                   |
|                  | A2 | Kepler         |                        | Postdoc      | 8                   | 5                   |
|                  | A3 | Kepler         |                        | Postdoc      | 8                   | 5                   |
| Genetics         | G1 | Mouse          | ✓                      | Grad Student | 4                   | 4                   |
|                  | G2 | Cancer         |                        | Grad Student | 2                   | 2                   |
|                  | G3 | Mouse          | ✓                      | Professor    | 10+                 | 2                   |
| Material Science | M1 | Solvent (8k)   | ✓                      | Postdoc      | 4                   | 5                   |
|                  | M2 | Solvent (Full) | ✓                      | Professor    | 10+                 | 5                   |
|                  | M3 | Solvent (Full) | ✓                      | Grad Student | 3                   | 5                   |

Table 3: Participant information. The Likert scale used for dataset familiarity ranges from 1 (not familiar) to 5 (extremely familiar).

During the contextual inquiry, participants demonstrated the use of domain-specific tools for conducting analysis in their existing workflow, including:

- Image Cutout Service (Astronomy)
- Short Time-series Expression Miner (Genetics)
- Solubility Database (Material Science)



Fig. 5: Screenshots from contextual inquiry. Left: A1 performs data smoothing to clean the data and then examines a light curve manually using a Jupyter notebook. Right: G2 uses a domain-specific software to perform clustering and visualize the outputs.

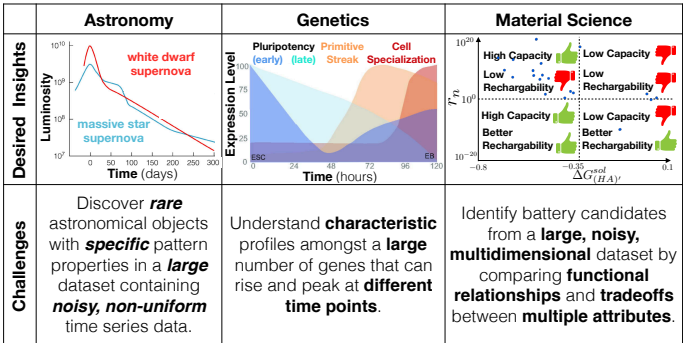


Fig. 6: Desired insights, problem and dataset challenges for each of the three application domains in our study.

Our collaboration with participants is illustrated in Figure 7, where we began with an existing VQS (Zenvisage, as illustrated in Figure 8) and incrementally incorporated features, such as dynamic class creation (Figure 9), throughout the PD process.

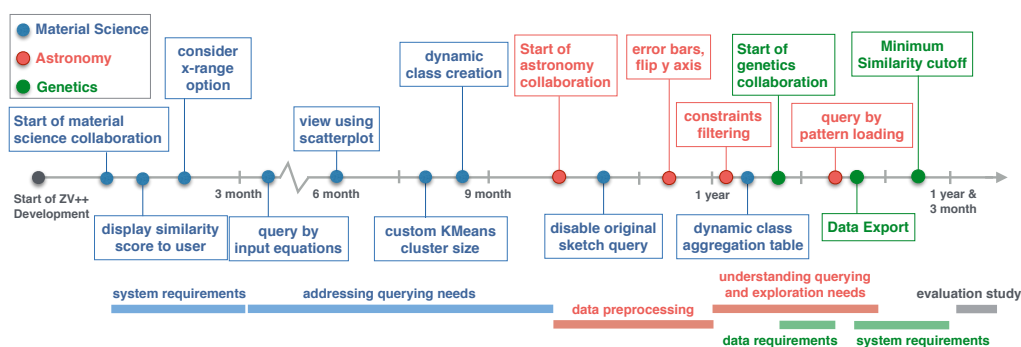


Fig. 7: Timeline for progress in participatory design studies.

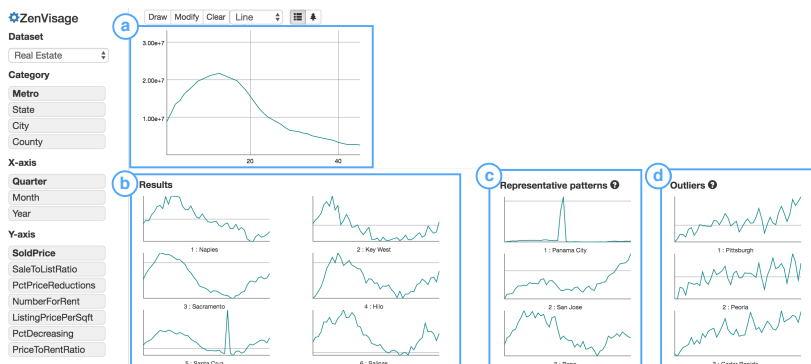


Fig. 8: The existing Zenvisage prototype allowed users to sketch a pattern in (a), which would then return (b) results that had the closest Euclidean distance from the sketched pattern. The system also displays (c) representative patterns obtained through K-Means clustering and (d) outlier patterns to help the users gain an overview of the dataset.

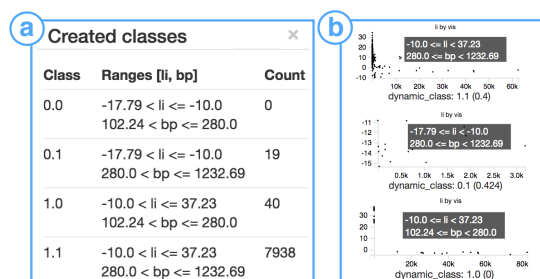


Fig. 9: Example of dynamic classes. (a) Four different classes with different Lithium solvation energies (li) and boiling point (bp) attributes based on user-defined data ranges. (b) Users can hover over the visualizations for each dynamic class to see the corresponding attribute ranges for each class. The visualizations of dynamic classes are aggregate across all the visualizations that lie in that class based on the user-selected aggregation method.

## B CHARACTERIZING THE PROBLEM SPACE FOR VQSs

We now characterize the space of problems addressable by VQSs and describe how each sensemaking process fits into different problem areas that VQSs are aimed to solve. Visual querying often consists of searching for a desired pattern instance ( $Z$ ) across a visualization collection specified by some given attributes ( $X, Y$ ). Correspondingly, we introduce two axes depicting the amount of information known about the visualized attribute and pattern instance as shown in Figure 10.

Along the **pattern instance** axis, the visualization that contains the desired pattern may already be **known** to the analyst, exist as a pattern **in-the-head** of the analyst, or be completely **unknown** to the analyst. In the **known** pattern instance region (Figure 10 grey cell), systems such as Tableau, where analysts manually create and examine each visualization one at a time, is more well-suited than VQSs, since analysts can directly work with the selected instance without having to search for which visualization exhibits the desired pattern. We define *top-down pattern search* as the process where analysts query a fixed collection of visualizations based on their in-the-head pattern (Figure 10 blue). On the other hand, *bottom-up data-driven inquiries* (Figure 10 green) are driven by recommendations or queries that originate from the data (or equivalently, the visualization), since the pattern of interest is unknown and external to the user.

The second axis, **visualized attributes**, depicts how much the analyst knows about which  $X$  and  $Y$  axes they are interested in visualizing. In both the astronomy and genetics use cases, as well as past work in this space, the attribute to be visualized is **known**, as data was in the form of a time series. In the case of our material science participants, they wanted to explore relationships between different  $X$  and  $Y$  variables. In this realm of **unknown** attributes, context creation (Figure 10 yellow) is essential for allowing users to pivot across different visualization collections.

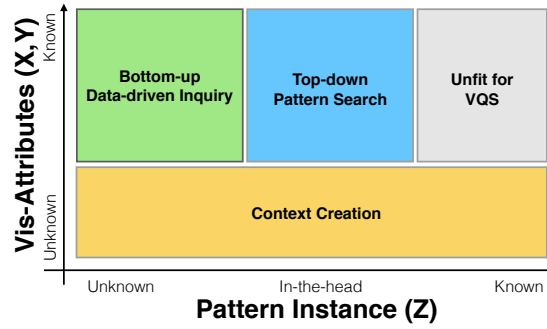


Fig. 10: The problem space for VQSs is characterized by how much the analyst knows about the visualized attributes and the pattern instance. Colored areas highlight the three sensemaking processes in VQSs for addressing these characteristic problems. While prior work has focused solely on use cases in the blue region, we envision opportunities for VQSs beyond this to a larger space of use cases covered by the yellow and green regions.

## C EVALUATION STUDY PROTOCOL

Here, we detail the procedures that were conducted during the evaluation study. At the beginning of the study, participants were asked a set of pre-study survey questions to collect basic information about participant's dataset, scientific questions, and existing workflows. While this information was similar to the ones collected through participatory design and contextual inquiry (Section 4), the pre-study survey ensured that we have background information even for the "blank-slate" participants (who were not part of the earlier design study).

- What is your current role as a scientist? What are some examples of recent questions you have researched?
- Describe the workflow that you currently use to analyze and make sense of this type of data.
- Can you describe an interesting finding you found with your current workflow and the process you took to obtain this insight?

After the tutorial and overview of the system, participant's selected dataset was loaded in. Participants were asked about their familiarity with the dataset and their analytical goals for the session.

- How familiar are you with this dataset? If you have worked with this dataset before, is there any insight that you already know from this dataset?
- What is your goal for this dataset? How long have you been working with this dataset? What are you hoping to accomplish with this dataset?

During the main experiment, participants engaged in talk-aloud exercises as they explored their data. These two semi-structured interview questions were often posed when participants begin a new line of analytical inquiry.

- What is your current goal in this phase of the exploration? What type of insights are you hoping to obtain?
- What actions are you planning to perform? How are you operationalize to achieve those goals?

In addition, we occasionally remind participants that they ask for help on something they want to accomplish on *zenvisage++*, but were not sure about the sequence of interactions. They were also encouraged to use other tools in their existing workflow alongside *zenvisage++* to perform their analysis.

At the end of the study, we interviewed participants with a set of open-ended questions regarding their experience with *zenvisage++*, as listed below.

- How was *zenvisage++* different from your existing workflow?
- Can you describe how you would use *zenvisage++* in your current workflow?
- On a scale of 1-10, how interested would you be in adopting this tool for your day-to-day workflow?
- What were some insights that you have gained from this session?
- Given the insights that you have obtained from *zenvisage++*, what additional analysis will you run downstream before you publish these results?
- What are the pros/cons for using *zenvisage++*?
- Were there any queries that you were unable to address with *zenvisage++* during today's session?
- What are additional features in *zenvisage++* that would help with your scientific workflow or serve your scientific need?

## D EVALUATION STUDY ANALYSIS DETAILS

We analyzed the transcriptions of the evaluation study recordings through open-coding and categorized every event in the user study using the following coding labels:

- Insight (Science) **[IS]**: Insight that connected back to the science (e.g. “This cluster resembles a repressed gene.”)
- Insight (Data) **[ID]**: Data-related insights (e.g. “A bug in my data cleaning code generated this peak artifact.”)
- Provoke (Science) **[PS]**: Interactions or observations that provoked a scientific hypothesis to be generated.
- Provoke (Data) **[PD]**: Interactions or observations that provoked further data actions to continue the investigation.
- Confusion **[C]**: Participants were confused during this part of the analysis.
- Want **[W]**: Additional features that participant wants, which is not currently available on the system.
- External Tool **[E]**: The use of external tools outside of *zenvisage++* to complement the analysis process.
- Feature Usage **[F]**: One of the features in *zenvisage++* was used.
- Session Break **[BR]**: Transition to a new line of inquiry.

| Domain   | IS | ID | PS | PD | C | W  | E  | BR | F  |
|----------|----|----|----|----|---|----|----|----|----|
| astro    | 4  | 12 | 13 | 57 | 2 | 18 | 20 | 22 | 67 |
| genetics | 8  | 12 | 7  | 35 | 4 | 13 | 1  | 21 | 52 |
| mat sci  | 14 | 8  | 7  | 44 | 8 | 11 | 3  | 12 | 48 |

Table 4: Count summary of thematic event code across all participants of the same domain.

In addition, based on the usage of each feature during the user study, we categorized the features into one of the three usage types:

- Practical **[P]**: Features used in a sensible and meaningful way.
- Envisioned usage **[E]**: Features which could be used practically if the envisioned data was available or if they conducted downstream analysis, but was not performed due to the limited time during the user study.
- Not useful **[N]**: Features that are not useful or do not make sense for the participant’s research question and dataset.

The feature usage labels for each user is summarized in Figure 11. A feature is regarded as *useful* if it has a **P** or **E** code label. Using the matrix from Figure 11, we compute the percentage of useful features for each sensemaking process as:  $\frac{\# \text{ of useful features in process}}{\text{total \# of features in process} \times \text{total \# of users}}$ .

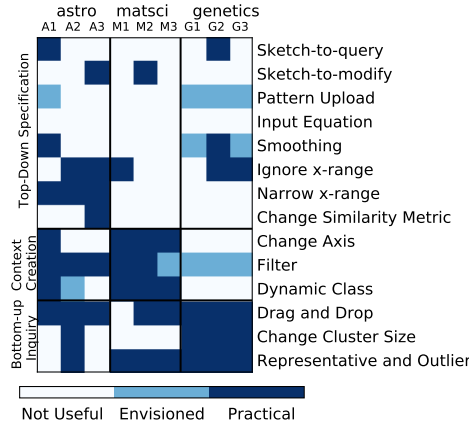


Fig. 11: Heatmap of features categorized as practical usage (P), envisioned usage (E), and not useful (N). Columns are arranged in the order of subject areas and the features are arranged in the order of the three foraging acts. Participants preferred to query using bottom-up methods such as drag-and-drop over top-down approaches such as sketching or input equations. Participants found that context creation via filter constraints and dynamic class creation were powerful ways to compare between subgroups or filtered subsets.

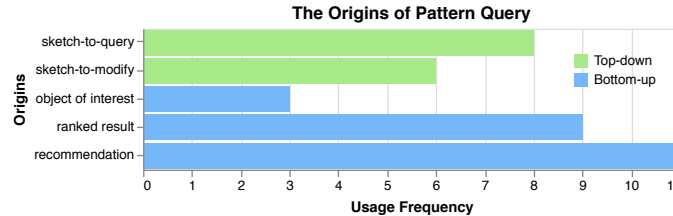


Fig. 12: The number of times a pattern query originates from one of the workflows. Pattern queries are far more commonly generated via bottom-up than top-down processes.

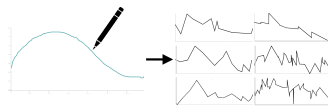
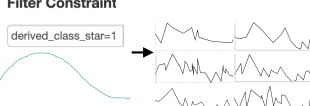
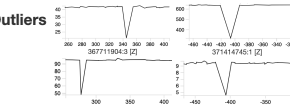
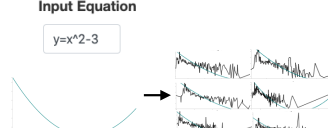
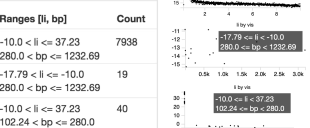
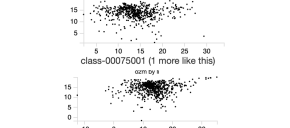
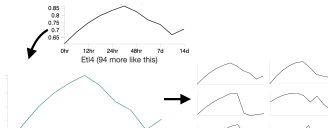
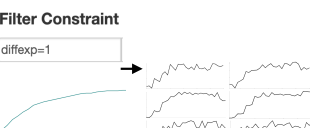
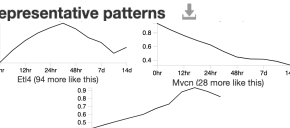
| Sensemaking Process   |  |  |   |       |                     |      |                       |  |                      |    |                       |  |                     |    |                      |  |  |
|---|--|--|---|-------|---------------------|------|-----------------------|--|----------------------|----|-----------------------|--|---------------------|----|----------------------|--|--|
|   | Top-Down   | Context Creation   | Bottom-Up   |       |                     |      |                       |  |                      |    |                       |  |                     |    |                      |  |  |
| Domain  | <b>Astronomy</b><br><b>Goal:</b> Discover potential supernovae candidates that exhibits peak-then-decay pattern<br>   | <b>Astronomy</b><br><b>Support:</b> Examine data regions that are more likely to have supernovae candidates<br><b>Filter Constraint</b><br>  | <b>Astronomy</b><br><b>Support:</b> Identify and eliminate sources of data anomalies to improve match accuracy for finding candidates<br><b>Outliers</b><br> |       |                     |      |                       |  |                      |    |                       |  |                     |    |                      |  |  |
|   | <b>Material Science</b><br><b>Support:</b> Find data classes that follows desired functional pattern to understand which solvent types exhibit certain tradeoffs and relationships<br><b>Input Equation</b><br> | <b>Material Science</b><br><b>Goal:</b> Compare characteristics from different data classes to find a solvent that satisfies desirable properties<br><b>Created classes</b><br><table><thead><tr><th>Ranges [li, bp]</th><th>Count</th></tr></thead><tbody><tr><td>-10.0 &lt; li &lt;= 37.23</td><td>7938</td></tr><tr><td>280.0 &lt; bp &lt;= 1232.69</td><td></td></tr><tr><td>-17.79 &lt; li &lt;= -10.0</td><td>19</td></tr><tr><td>280.0 &lt; bp &lt;= 1232.69</td><td></td></tr><tr><td>-10.0 &lt; li &lt;= 37.23</td><td>40</td></tr><tr><td>102.24 &lt; bp &lt;= 280.0</td><td></td></tr></tbody></table>  | Ranges [li, bp]   | Count | -10.0 < li <= 37.23 | 7938 | 280.0 < bp <= 1232.69 |  | -17.79 < li <= -10.0 | 19 | 280.0 < bp <= 1232.69 |  | -10.0 < li <= 37.23 | 40 | 102.24 < bp <= 280.0 |  | <b>Material Science</b><br><b>Support:</b> Understand the overall tradeoffs and relationships between data attributes<br><b>Representative patterns</b><br> |
|   | Ranges [li, bp]  | Count  |   |       |                     |      |                       |  |                      |    |                       |  |                     |    |                      |  |  |
| -10.0 < li <= 37.23   | 7938   |  |   |       |                     |      |                       |  |                      |    |                       |  |                     |    |                      |  |  |
| 280.0 < bp <= 1232.69   |  |  |   |       |                     |      |                       |  |                      |    |                       |  |                     |    |                      |  |  |
| -17.79 < li <= -10.0  | 19   |  |   |       |                     |      |                       |  |                      |    |                       |  |                     |    |                      |  |  |
| 280.0 < bp <= 1232.69   |  |  |   |       |                     |      |                       |  |                      |    |                       |  |                     |    |                      |  |  |
| -10.0 < li <= 37.23   | 40   |  |   |       |                     |      |                       |  |                      |    |                       |  |                     |    |                      |  |  |
| 102.24 < bp <= 280.0  |  |  |   |       |                     |      |                       |  |                      |    |                       |  |                     |    |                      |  |  |
| <b>Genetics</b><br><b>Support:</b> Search and browse for genes belonging to the same cluster<br> | <b>Genetics</b><br><b>Support:</b> Compare known properties of genes belonging to different clusters<br><b>Filter Constraint</b><br>   | <b>Genetics</b><br><b>Goal:</b> Understand characteristic pattern profiles in the dataset<br><b>Representative patterns</b><br>   |   |       |                     |      |                       |  |                      |    |                       |  |                     |    |                      |  |  |

Table 5: Table of example usage scenarios from each domain for each sensemaking process. We find that our participants typically have one focused goal expressible through a single sensemaking process, but since their desired insights may not always be achievable with a single class of operation, they make use of the two other sensemaking processes to support them in accomplishing their main goal.