AvI - Avalanche Investigator

Exploring Characteristics of Utah's Reported Avalanches Over Time

Process Book

Jadie Adams and Max Marno

Table of Contents

Project Repository: https://github.com/jmaxmarno/avy-obs

Overview and Motivation

As winter recreationists we have spent time pouring over avalanche forecasts and observations to better understand the environment around us. Making informed decisions is a critical component for winter backcountry travelers and while published advisories are the first step in assessing risk, past observations are also relevant. The Utah Avalanche Center collects and makes available a dataset of avalanche observations from trained forecasters, and the public. These observations represent a sample of the real occurrences, but can be useful for exploratory analysis. The dataset contains observations from as far back as 60 years ago, but starting around 2009 there is a more consistent number of observations.

Our visualization illustrates how the distribution of these observations has changed over the last decade. Our project enables users to explore how reported rates of avalanches with different characteristics have changed, potentially drawing attention to the increase in backcountry traffic. As we are aware of the presence of reporting bias, our design focuses on faceting the observations into attribute sets, and displaying the normalized distributions of these attributes. Previously, interested users could only export this data as CSV or explore it in a table format. We make the data more accessible and useful by implementing mechanisms for exploration, while reinforcing the notion that observations are not a representative sample of all Utah avalanche phenomenon and absolute inference should not be made about the population distribution.

Our visualization has both user-driven interactive components as well as guided story-telling aspects to highlight points of interest. It allows the user to explore the following characteristics of avalanche observations:

- Trigger / cause of avalanche
- Aspect or cardinal facing of avalanche slope
- Discretized elevation of avalanche
- Discretized size of avalanche
- Overall distribution of observations

Our project represents this data in multiple view system and allows for granular exploration while maintaining an overall view for context.

Related Work

- The brushable heat map / time grid was inspired by a previous project for this class, “Arctic Explorer” by Dylan Wootton and Ethan Ransom: http://www.dylanwootton.com/Arctic-Explorer/. Since our project is about exploring observations over time we thought the time grid would work well for our data.

- The stacked bar plot of the main plot was implemented using D3 series, we used this example by Robert Pettersson as a reference: https://bl.ocks.org/LemoNode/5a64865728c6059ed89388b5f83d6b67
- The star plot was inspired by http://bl.ocks.org/nbremer/raw/21746a9668ffdf6d8242/. We liked the way distributions could be overlayed and decided it would be a good way to provide context to our visualization as well as a legend.

Questions

We wanted to give users the opportunity to explore a rich dataset of crowdsourced avalanche observations that is otherwise difficult to access. Considering our dataset and the presence of reporting bias, we chose to visualize the attributes as proportions of the whole. By normalizing the attribute sets in this way it is less likely that someone may incorrectly infer a trend that does not exist. Ex; 'Look, the number of avalanches triggered by skiers went way up!'.

We emphasize the fact that these observations only represent a sample of avalanches, and provide as much insight as to where people are traveling, as much as where avalanches are occurring.

However, the visualization allows one to answer questions such as:

- How does the distribution of avalanche observations change over aspect throughout the winter?
- Were observed low elevation avalanches more prevalent in one year compared to another?
- Has the proportion of observed skier triggered avalanches changed throughout the year?
- Does the width of reported slides have any correlation to the number of slides reported in a given month?

Our goal was to design a visualization that could both answer questions such as this as well as inspire the user to ask new questions.

In implementing our visualization and playing with the final design we found some of our initial questions were not as interesting as we thought, so we decided to approach the question asking a little differently. We explored the data using or visualization and noted down time points or ranges that were unique, or interesting, or helped illustrate an overall trend and incorporated them into our storytelling. We ended up doing three of these points for each of the four characteristics. In this way we were able to show the user how one might learn from the visualization and answer a dozen questions in the process.

Data

Source

Data of recorded avalanche observations is provided on the Utah Avalanche Center website: https://utahavalanchecenter.org/avalanches They have made the data available to download in a csv file.

The data recorded for each observation includes:

Date, Region, Place, Trigger, Depth, Width, Vertical, Aspect, Elevation, Coordinates, Weak Layer, Caught, Carried, Buried - Partly, Buried - Fully, Injured, Killed, Accident and Rescue Summary, Terrain Summary, Weather Conditions and History, and Comments.

Often some of these fields are left blank but they all include at least a date and nearly all have the first half of the characteristic accounted for.
For our stacked bar visualization we needed to bin/discretize the continuous variables; width and elevation. For width we used the following scale:

- Large (width > 600)
- Medium (300<width<600
- Small (300>width)

For elevation we followed a common tier commonly used in avalanche forecasting

- Above 9,500ft
- 8,000ft - 9,500ft
- Below 8,000ft

Processing

Our dataset was already relatively ‘clean’ however we expected to deal with some erroneous values as the dataset incorporates crowdsourced observations. We also needed to decide how to handle missing values for the characteristics of interest, as these were somewhat sparse we decided to mark these values as “Unknown” and include them as a category. The data we were interested in was categorical attribute sets and binned quantitative attributes. We pre-processed our data using Python to minimize the overhead needed for our visualization. This process is modular so that future incorporation with an updated dataset will be simple.

For the attributes that already had categorical labels, we simply used all of the labels present in the data. For the quantitative attributes we chose a way to bin and assign labels accordingly:

- Elevation - for elevation we used the three categories used by the Utah Avalanche Center and many other avalanche centers: below 8,000ft, 8,000ft-9,500ft and above 9,500ft.
- Width - The slide reports can include width, vertical, and depth. In exploring the data we found width was most often reported and so we decided to use width as our size measure. We binned by width using the categories: under 100 ft. across, between 100 ft. and 200 ft. across, and over 200 ft. across.

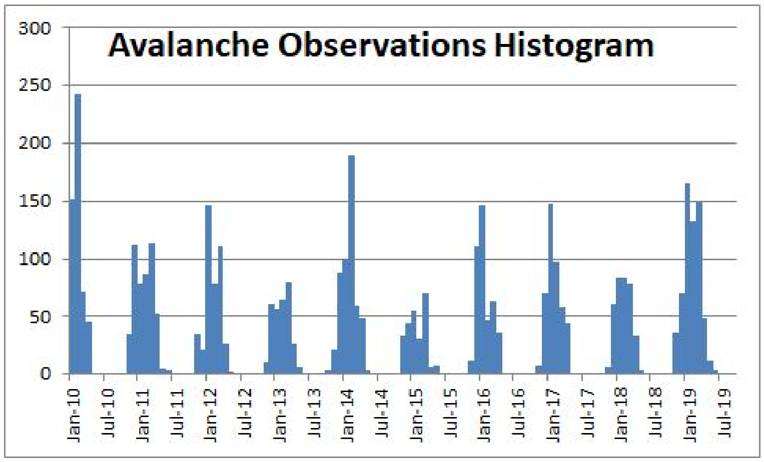
Our pre-processing script developed in Python takes the raw .csv file and converts it to a ".json" file with the appropriate nesting.

The structure looks like this:

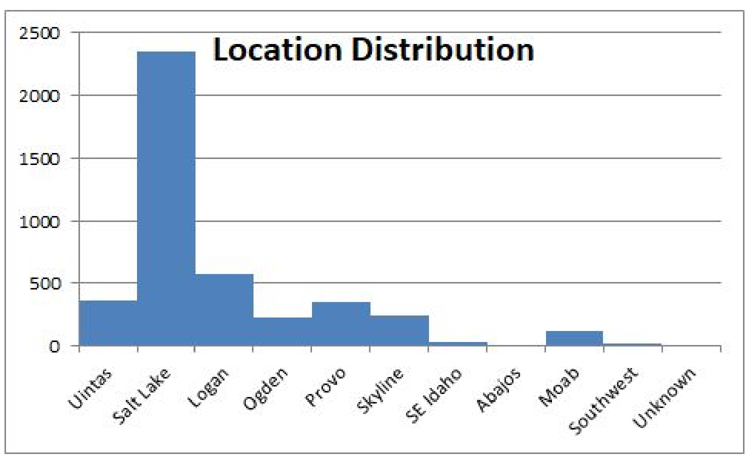
```
'''
JSON structure:
{
    year{
        month{
            total_count
            region {}
            trigger {}
            weak_layer {}
            size {}
            aspect {}
            human_stats{
                caught
                carried
                burried_partly
                buried_fully
                injured
                killed
                }
            }
        month ...
        }
      ...
    year ...
    }
    ....
}
```

Exploratory Data Analysis

Initially we made a histogram in Excel to get an overall idea of the observation distribution. Below is the histogram from 01/01/2010 to now, it is binned by month. It can be seen that in the winter months there are around 100 observations on average and in the summer months there are none.



One thing we considered was using the distribution of slide locations in our visualization. We created a histogram of the distribution of observations among the different regions as seen below.



In looking at this histogram we decided not to include any cartographic visualization in our design. As you can see the vast majority of observations are made in the Salt Lake region. This is because this is where most of the population that frequents the backcountry are. We ultimately decided viewing the proportion of reports in different regions would not be interesting because the Salt Lake proportion is so large and these proportions are less likely to change over time.

Design Evolution

Proposal

As our visualization aims to allow users to explore the change over time of the avalanche characteristics, we wanted to provide the ability to select a desired attribute set (aspect, trigger, elevation, size, etc...) and visualize the distribution of observations within this selected category over time. This steers the user away from making inferences about the real distribution of all avalanches, and provides the benefit of describing how the characteristics of the selected avalanche observations has changed (over the selected time frame). We proposed a design in which the user will specify:

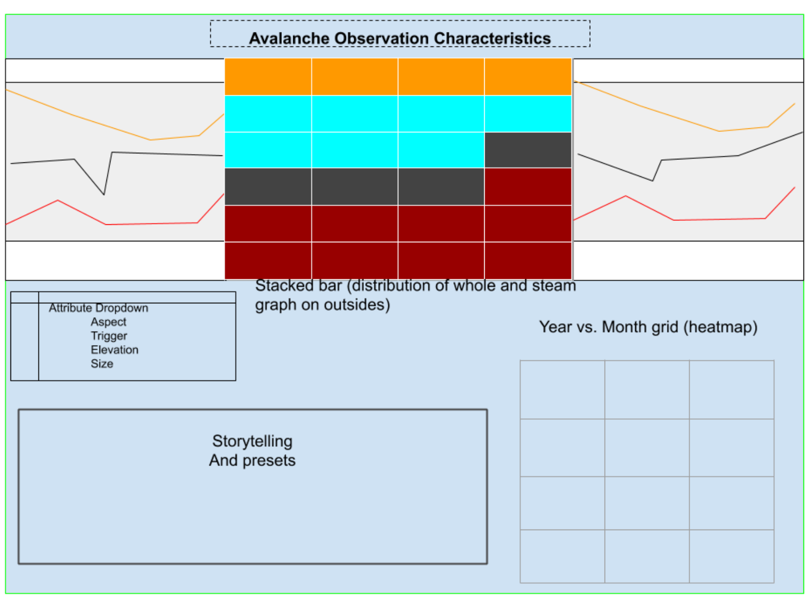
- Time frame (based on month and year) selected via brushing grid
- Category selection (aspect, trigger, size, elevation and maybe some sub-categories)

The must have features we defined in the proposal were:

- Stream/Area chart
- Attribute drop down to select what to display
- Heatmap or gid month/year selector
- Preset ‘stories’

The optional features we defined were:

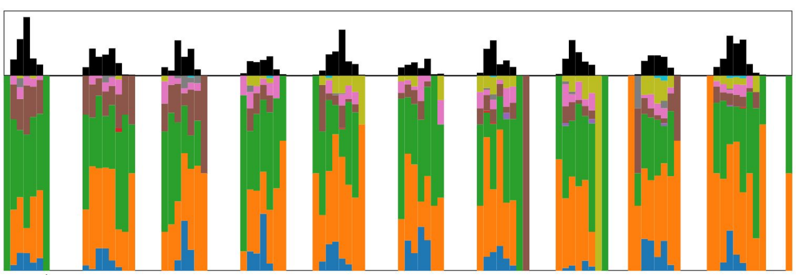
- Year clock chart
- Scatter or graph view of how attribute frequency has changed over time
- Derived attributes like weather and forecast danger (from archives)



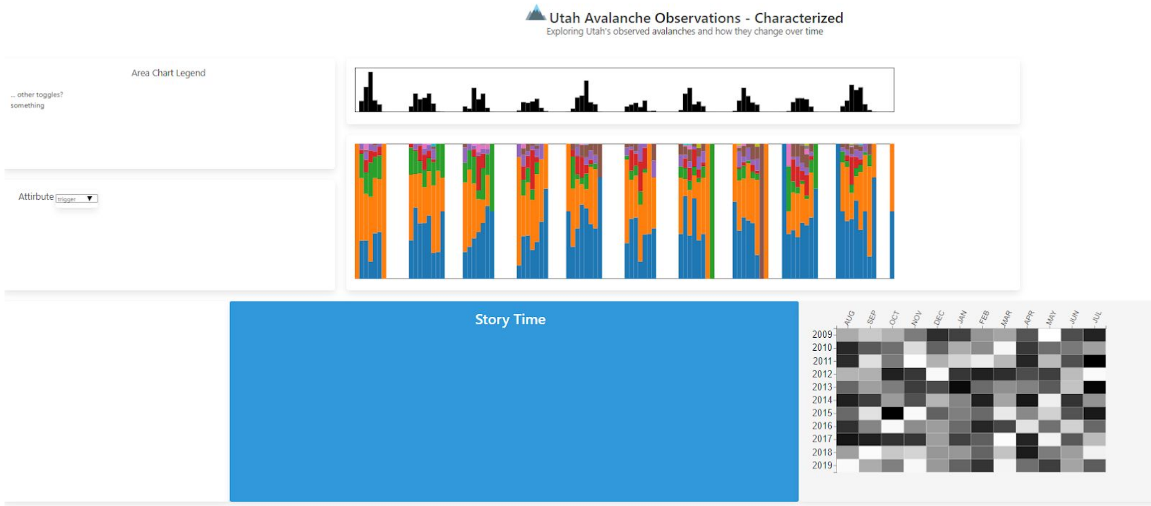
Our proposal design combined an area chart with the heat map view. The idea was user will be able to select which attributes they want to visualize using the drop down and select the year and month displayed using the heatmap, which will also indicate the number of encode observations within each cell by saturation. We decided that an area chart/stacked bar chart is an appropriate form of visualization because we are interested in showing parts of a whole and it enables the user to see which categories are most often reported and how those distributions change over time. The heat map is also an appropriate way of encoding the total number of observations because it will allow the user to visualize the cyclical properties of the data and get an idea of how these counts have changed throughout the years. We also chose to include a story and provenance section with a data-driven written description of what attributes are being displayed and how the data is filtered, as well as an overview of interesting points and storytelling.

Milestone

One of the first design changes we decided to make in implementing the milestone was to add a histogram of observations above our area chart as seen below.



We decided that the histogram gave necessary context for the normalized area chart about the total observation count. It helps illustrate, for example, why there are blank bands throughout the chart (summer months with no observations). It can also help a user understand that a proportion which seems extremely high may be an artifact of a total low observation count. We also decided that the stacked bar design was a better choice than a stream graph for the zoomed out data because it shows finer detail without looking too crowded. Our milestone design look pretty close to our proposed visualization in the end.

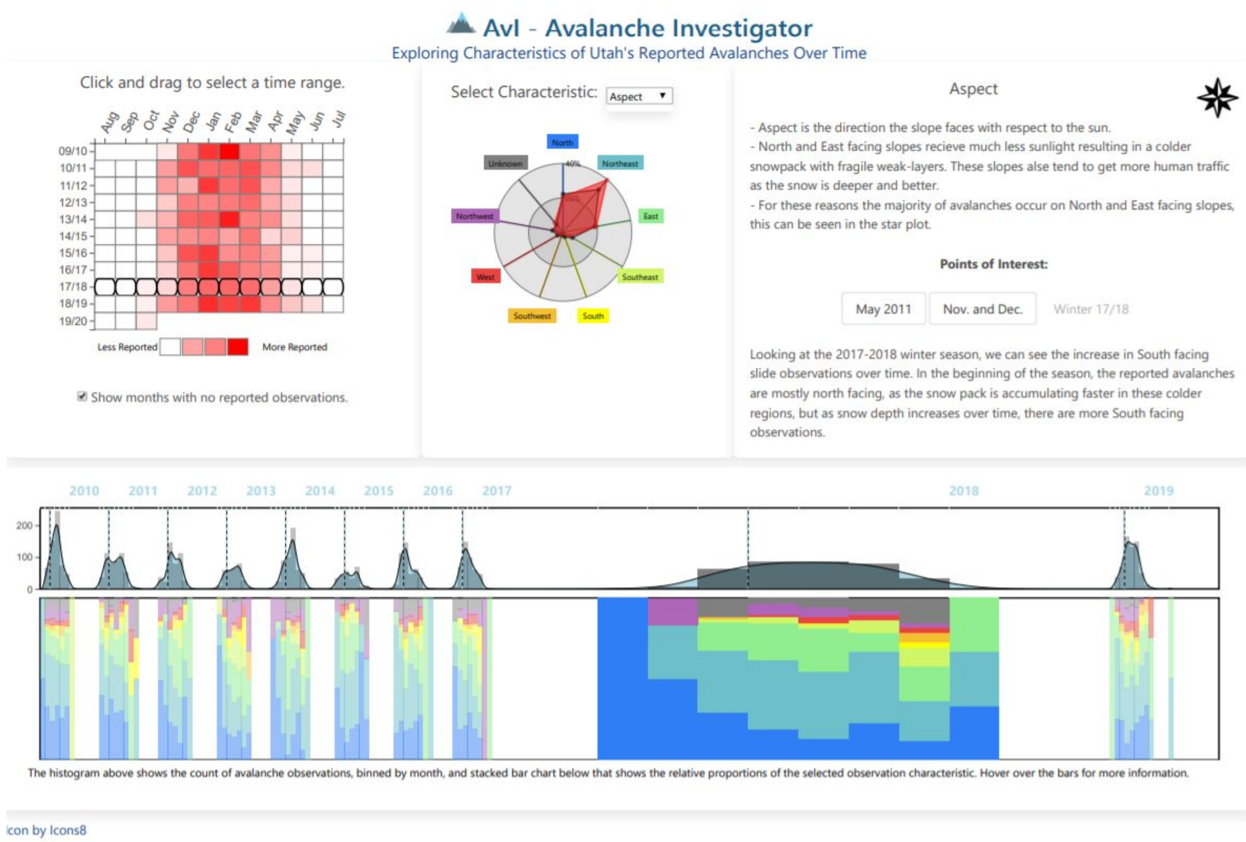


Final Design

In our final design we kept the elements of our milestone design but added a fair amount. Here are the main additions and changes we implemented:

- **Star plot / legend** - We added a star plot which shows the overall distribution of the selected characteristic and also doubles as a legend for the stacked bar chart, effectively conserving space. In considering our project at the milestone, we realized that while glancing at the entire stacked bar plot kind of gave an idea of the overall distribution, having a visualization of the averages like in the star plot would be helpful. We also use it to overlay the distribution of a selected region so you can see how that compares to the average.
- **Interaction** - We added a lot of interaction after the Milestone. We added a tooltip on hover where data is bound so the user can easily see labels and values. We also made the heat map time grid brushable to select a time region of interest. All of this allows the user to drill down on what they are interested in and learn more.
- **Storytelling** - We added a storytelling section to our visualization. We decided the four characteristics could use further explanation, especially for those who are less familiar with the language used in avalanche forecasting and reporting. We added this to the storytelling as well as points of interest which highlight interesting distributions. We decided this may also help the user understand how to use and navigate the visualization.
- **Layout -** After the milestone we made layout more compact. We chose to move the histogram and stacked bar chart below everything else so it could take the full width of the screen. We also made the visualization scale based on screen width so that it would look better on different display modes.

We also updated what we had for the milestone in a few ways. For example, we improved our color schemes, added interactivity, and added a kernel density plot to the histogram. After all of these changes this was our final design:

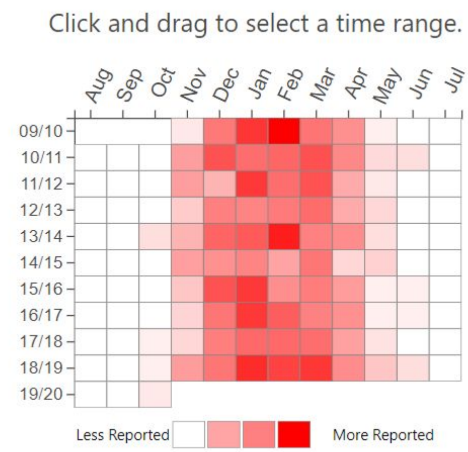


Implementation

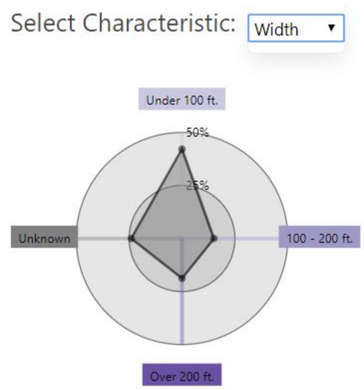
Views

Our design had four main components: the time grid, the star plot the story box, and the main plot which included a histogram and stacked bar plot. We give an overview of each of these elements below and then go on to describe the interactivity.

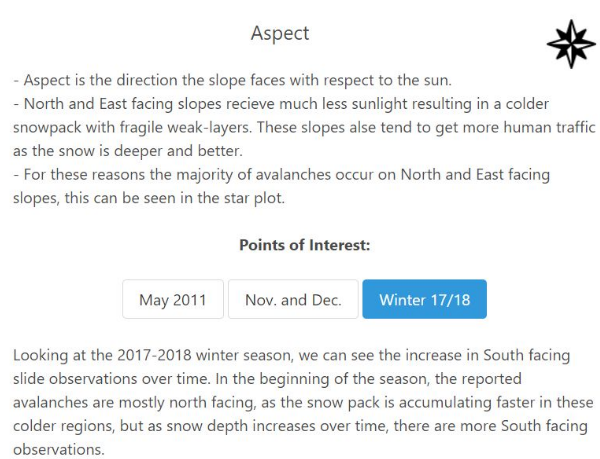
- **Time Grid -** The time grid shows the full time frame of our dataset. Each cell represents one month in this time frame where column is the month and the row is the year. We chose to start with August rather than the traditional choice of January because our data is cyclical and avalanche season usually begins in November. The number of observations in each month is encoded by the intensity of the color, where lighter red means less observations and darker red means more observations.



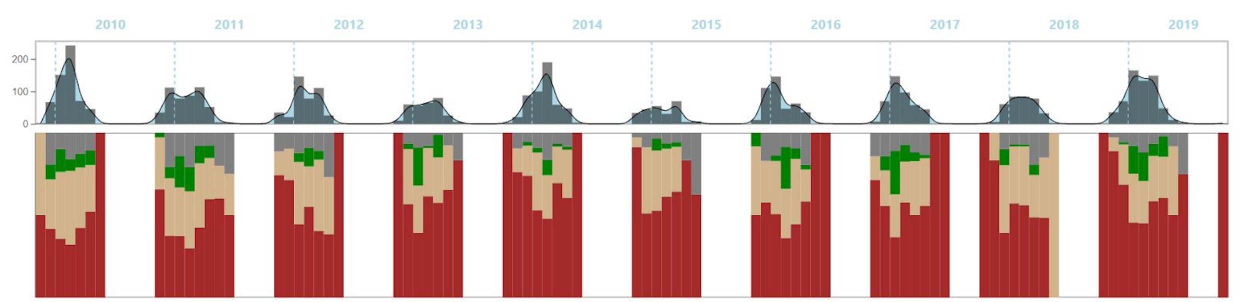
- **Star Plot -** The star plot has a few functions. It shows the categories for each of the characteristics one can select in the dropdown as well as the color encoding of these categories which is also used in the stacked bar plot. It also shows the average proportions of these categories, where the distance from the center encodes the proportion. This gives the user an overall view.



- **Story Box -** The story box updates with the characteristic drop down as well. It has a description of the selected characteristic at the top and some selectable points of interest for that characteristic beneath that.



- **Main plot -** The main plot in our visualization has two components. The first is a **histogram** which shows the distribution of total observation in each month over time. This gives necessary context for the second component, the **stacked bar plot**. The histogram encodes the number of observations by the height of each bar. The overall distribution is also encoded by height in an overlaid kernel density plot. There are dashed lines and labels for marking the boundaries of years and the month labels are available on hover. The stacked bar plot encodes the proportion of each category in the characteristic of interest during that month. This is also encoded by height and the percent is available on hover.

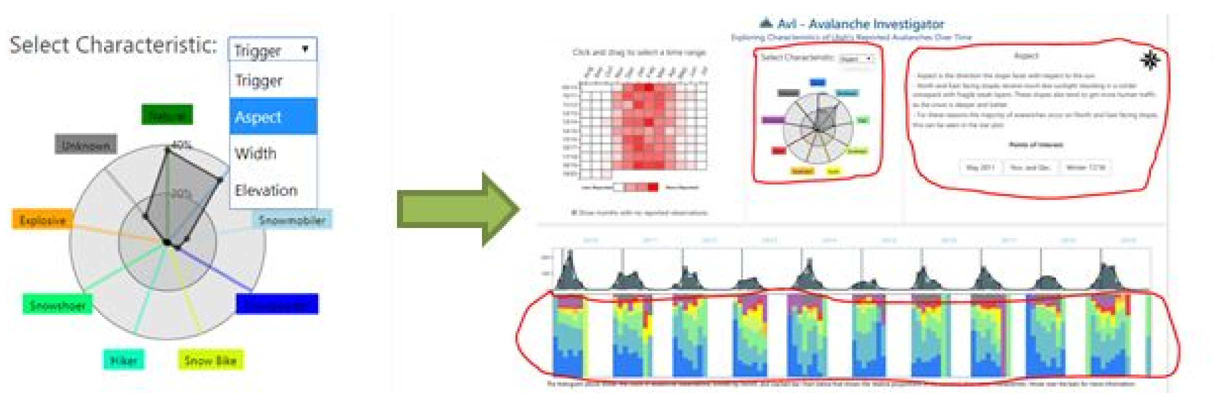


The histogram above shows the count of avalanche observations, binned by month, and stacked bar chart below that shows the relative proportions of the selected observation characteristic. Hover over the bars for more information.

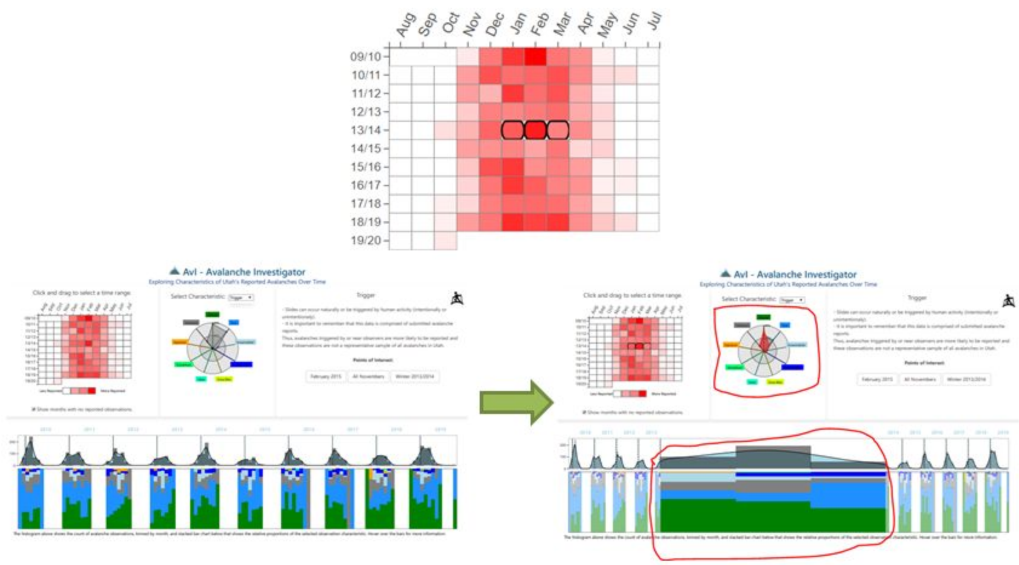
Interaction

Below is a list of the main interactive components and how they work. After each description there is an image of the interaction and what it changes in red.

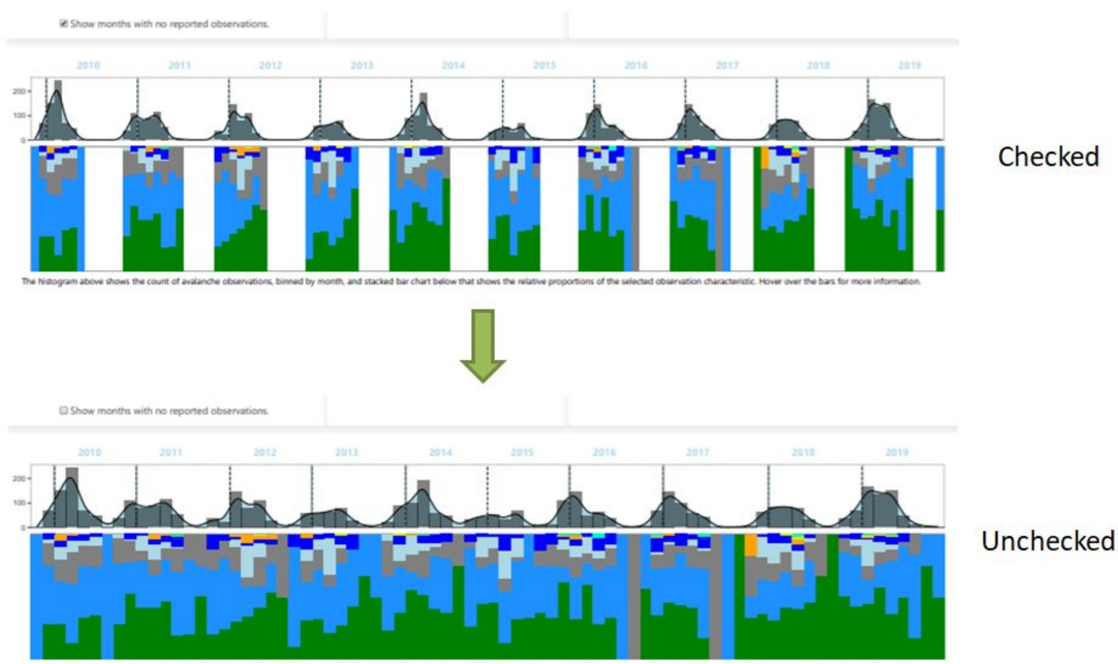
- Select Characteristic - This drop down allows you to choose which of the four slide characteristics you want to view. When you select one, the following views are updated:
    - The star plot / legend
    - The story box
    - The main plot



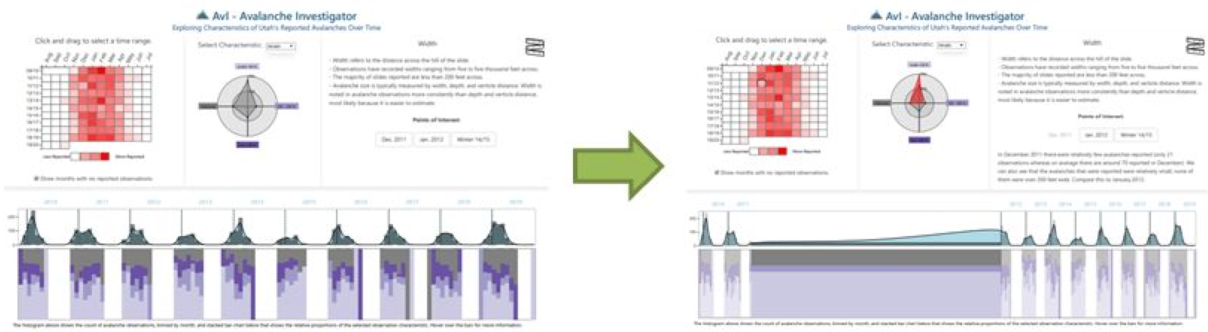
- **Brushing the Time Grid-** Brushing the time grid does a couple of things. It overlays the distribution of just selected time from on the star plot in red. This allows the user to compare the selected time frame to the average distribution. Brushing also zooms in on the corresponding bars in the main plot and grays out the rest. This allows the user to drill down on the time region of interest.



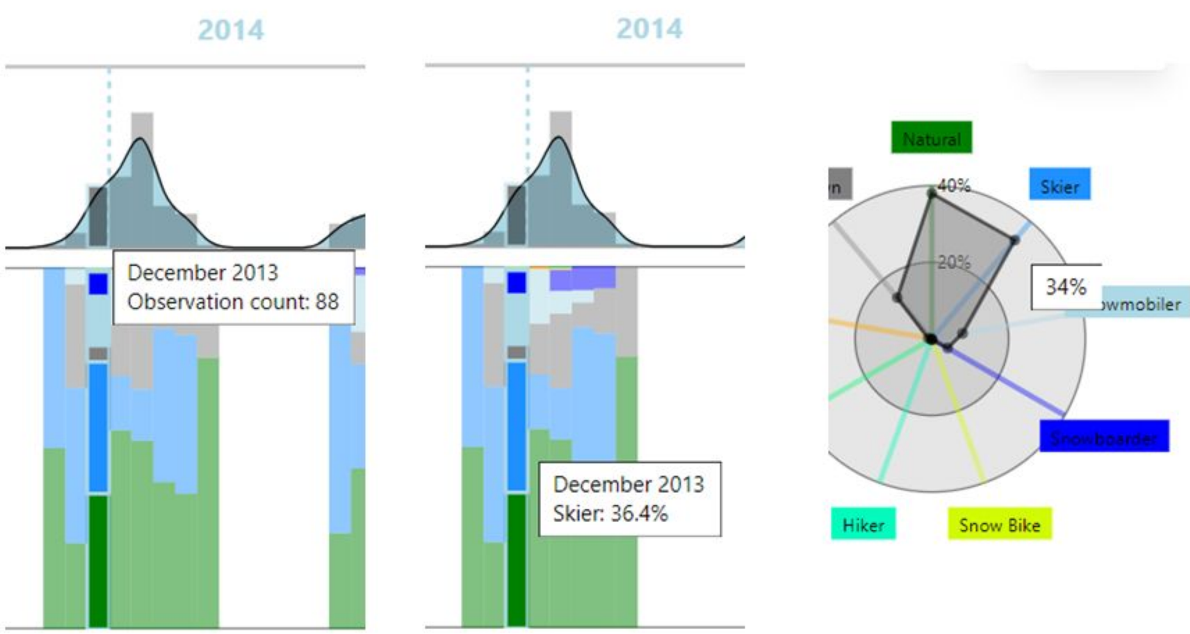
- **Show "No Observations" Checkbox -** When this box is unchecked, only months with reported observations are displayed in the main plot. When it is checked, there are blank bars for the months with no observations. This helps the user customize the view for what they are interested in. If they only care about the observations they can filter these months out, but if they want to see a full timeline they can include these months. They may, for example, want to compare how long the slide free time was in different years.



- **Story Points of Interest -** The story points of interest are like preset time grid brushes. They brush the right region in the time grid, add the overlay to the star plot, and zoom in on the appropriate bars in the main plot.



- **On Hover -** Tooltip appear on hover in a few places. When a bar in the main plot is hovered over a tool tip is added and the other bars are grayed out to make it more clear.
    - When a histogram bar is hovered on, it shows the month and year and the total count.
    - When a stacked bar is hovered over it show the month, year, category, and percent.
    - When a point on the star plot is hovered over it shows the exact percent.



Evaluation

We learned many things about the data through our visualization, most of which are added as part of our story telling. Our visualization works very well to achieve our goal of providing an effective, interactive way of exploring past avalanche observations and their characteristics. The multiple views allow for the user to drill down and get details while keeping the overall context in mind. The storytelling works to effectively give supporting information, introduce new information, and help the user see ways in which the visualization can be used and interacted with. Our project was successful in providing a better way to explore this data while reinforcing the notion that these reported observations are not necessarily representative of all slides in Utah.

The main way we could improve our visualization is by adding more characteristics and information. We did not include all of the information available about the observations for practical regions. One thing that would be interesting to add is the statistics around human involvement and how many slides involve injuries and fatalities. Our current visualization could not easily encode this but perhaps it could be improved by overlaying some timeline with the incidents.