Guide Dogs Visualization

VISUALIZATION FINAL ASSIGNMENT

Shlomo Shmouely shlomo.shmouely@gmail.com 318296779
Itay Segev adcitay@gmail.com 209146067

Contents

Question 1: What?
The data used in this study

Our dataset is the extended Guide Dog examination dataset, a table-based dataset given to us by Dr. Anna Zamansky as part of her collaboration with the Israel Guide Dog Center for the Blind. Note that this is not the simplified dataset that was given through the Moodle website of this course, but the full version of it.

The original dataset is stored in a large, complex and static table, where each row contains a dog's personal information, its test scores across 50 different aspects (subtests), and whether or not a specific subtest score disqualifies the entire test.

Since the original dataset was disorganized, inconsistent, missing information and spanning across three different files, we converted it into a new schema that is better structured, cleaner and is more suitable for visualization. The new schema consists of 4 tables, as described in the next section.

It's worth noting that we don't list the type for the ID properties because they are used only for the database structure and have no meaning outside of it. Also, properties listed as "Unique" are properties that cannot be classified as ordered or categorical data (such as names, IDs and references to IDs)

The dataset tables
The "SubTestDescriptions" Table
In the original dataset, each subtest result has two columns, one for the absolute rating and one for flagging disqualification of the entire test. The title of the first column describes the subtest type.

Every row in the original dataset is converted into a corresponding entry in the "Tests" table and into about 50 entries (for simplicity, we ignore 3 of the tests that don't have numerical score) in the "SubTestResults" table, one entry for every subtest.

This table contains the textual and logical description of each of these two columns, and is used to identify the type of the subtest in each SubTestResults entry.

Name	Type	Details
ID	-	The ID of the subtest
Description	Unique	The description of the subtest, in Hebrew.
Column ID	-	Column ID in the original table used for holding the subset score (This field only exists for the conversion process)

The "Tests" Table
This table contains the information of each test. It is used for grouping results of subtests and for identifying the dog and the corresponding subtests it took (used for filtering).

Name	Type	Details
TestID	-	The ID of the test
DogID	Unique	The ID of the dog
Date	Ordered Sequential	The birth date of the dog (Sequential because we know which date comes before another)
tchecklist	Unique	A unique identifier for the entire test

3

The "Dogs" Table

This table contains a dog's personal information. The unique properties for each dog are their names, their IDs and their parents' IDs. The other properties can be described either as categorical or ordered.

Name	Type	Details
ID	-	The ID of the dog
Name	Unique	The name of the dog, in Hebrew (In the original dataset, it included the dog's ID)
Name (English)	Unique	The name of the dog, in English
Birthday	Ordered Sequential	The birth date of the dog
Age at Training	Ordered Ordinal (also: categorical)	The age of the dog (in weeks) when the training started (between 9 to 14 weeks)
Status	Categorical	The current status of the dog (e.g. guiding, training, not suitable)
Secondary Status	Categorical	The current stage in which the dog is in (e.g. training, dead)
Father ID	-	The dog's father ID (from original dataset)
Mother ID	-	The dog's mother ID (from original dataset)
Gender	Categorical	The dog's gender (Male/Female)
Breed & Color Code	Categorical	The dog's breed code
Passed (Derived from data)	Categorical	The test results (pass or fail) ("true" if the status is guiding, false otherwise)

Score (Derived from data)	Ordered Sequential	A metric calculated based on the dog's children success in the tests. Dog without children is scored 0. (Used in the family relations "heatmap" view) $$score = \frac{\#children\_passed\_the\_test}{\#children}$$

The "SubTestResults" Table

This table contains all the results for the subtests. Each ~50 entries are linked to one test, using the "Test ID" field.

Name	Type	Details
SubTest ID	-	The ID of the subtest result entry
Test ID	Unique	The ID of the entire test
SubTestKind	Categorical	The type of the subtest (This is the ID in SubTestDescriptions)
Score	Ordered Quantitative Sequential	The score of the subtest (typically, in the range of 1 to 5, or 1 to 10, but may vary)
Disqualified	Categorical	Specifies if this subtest disqualifies the entire test

Derived Data

For the "Test Results" view in the visualization, we calculate two other structures, based on aggregations of the entries of "SubTestResults".

All the entries are initially filtered by the dog properties using "TestID" in "SubTestResults" and "Tests", and the attributes "DogID" and "ID" in "Tests" and "Dogs" respectively.

To accumulate the results we use D3's internal "nest" filter - `d3.nest().` We use different settings for each of the two structures.

We then attach the subtest metadata from the SubTestDescriptions table.

Note: Some attributes (marked with an asterisk) are calculated only for the data display

The "subtests summed scores" array

For every subtest type, there is an object in the array with the following structure:

Name	Type	Details
SubTest Kind ID (Logical representation: key)	Categorical	The ID of the subtest type (in "SubTestDescriptions")
SubTests Values (Logical representation: values)	Ordered Quantitative Sequential	An array of values objects (see below)
SubTest Metadata (Logical representation: info)	-	The subtest description information (Taken directly from SubTestDescriptions)

Each entry in the "values" represents one possible answer for the "" field and the number of entries in "SubTestResults" that contain this answer. The entries are represented with objects with the following structure:

Name	Type	Details
Answer Code (Logical representation: key)	Categorical	The answer code (Most of the time on a scale of 1-5 or 1-10)
# Answers (Logical representation: values)	Ordered Quantitative Sequential	# of votes for this answer
Cumulated Sum of Answers* (Logical representation: info)	-	The cumulated sum of the "# Answers" field across they array. Used for both the visualization itself and for the textual data).

The "summed disqualifications" array

For each different kind of subtest, there is an object in the array with the following structure:

6

Name	Type	Details
SubTest Kind ID (Logical representation: key)	Categorical	The ID of the subtest kind (in "SubTestDescriptions").
SubTests Values (Logical representation: values)	Ordered Quantitative Sequential	An array of values objects (see below).
SubTest Metadata (Logical representation: info)	-	The subtest description information. (Taken directly from SubTestDescriptions)

Each entry in the "values" represents one possible answer for the "Disqualified" field and the number of entries in "SubTestResults" that contain this answer. The entries are represented with objects with the following structure:

Name	Type	Details
Answer Code (Logical representation: key)	Categorical	The answer code (True/False)
# Answers (Logical representation: values)	Ordered Quantitative Sequential	# of votes for this answer
Cumulated Sum of Answers* (Logical representation: info)	-	The cumulated sum of the "# Answers" field across they array. Used for both the visualization itself and for textual data).

Other attributes:

Name	Type	Details
Total Tests*	-	Contains the total number of tests.

Question 2: Why?
The reasons for visualization

For the scope of this assignment, we create 3 different visualizations. Each visualization presents different aspects of the data and introduces new insights.

The "Dog Data Overview" visualization

The "Dog Data Overview" visualization help us in various user-tasks:

First User Task: Enable users to **discover relations between different dog traits**. For example – gender and training age - we can see that female dogs typically start training at an older age (mostly at 12 weeks or more) than male dogs.
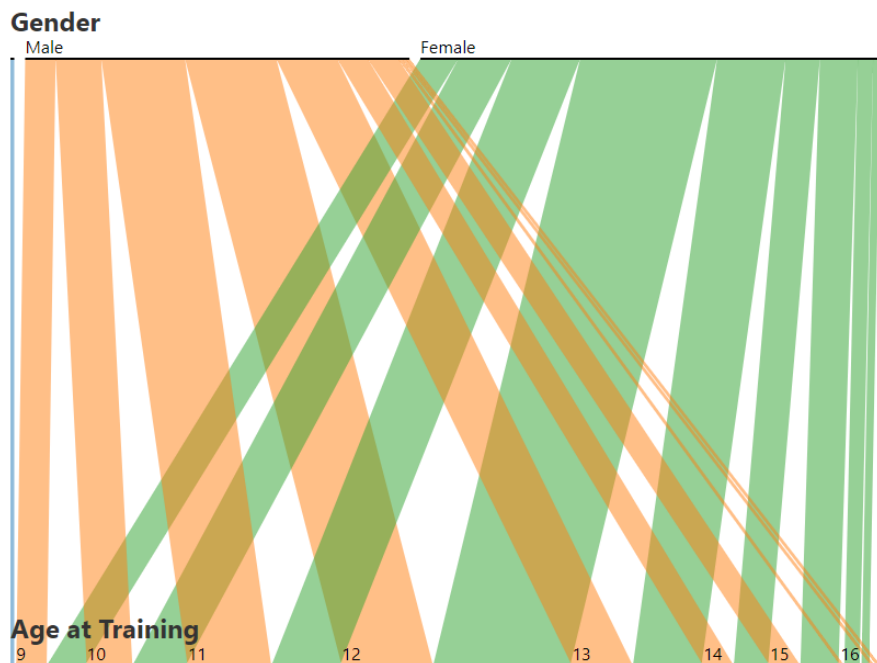
Second User Task: Enables users to **compare how different values are distributed across each category**. For example, we can see that there are more female dogs than male dogs in the dataset (55% females vs. 45% males).
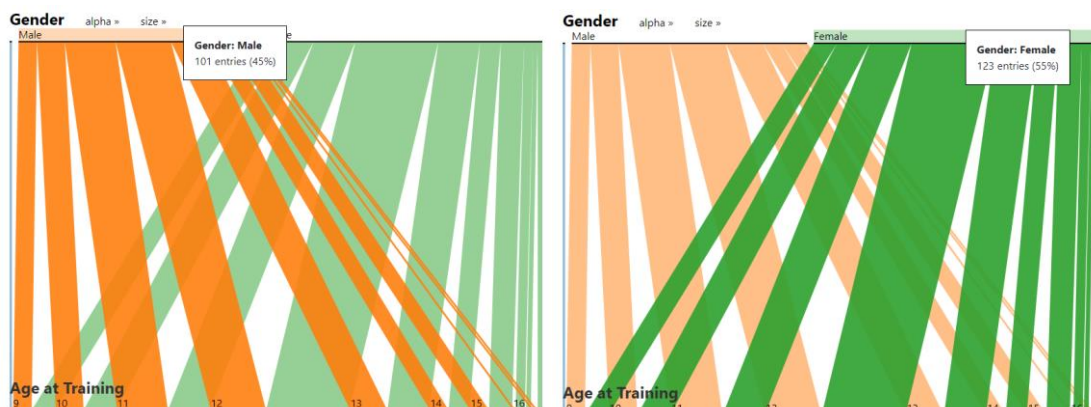
Also, we can see that the majority of the dogs in the dataset are Labradors (marked by the LAB).
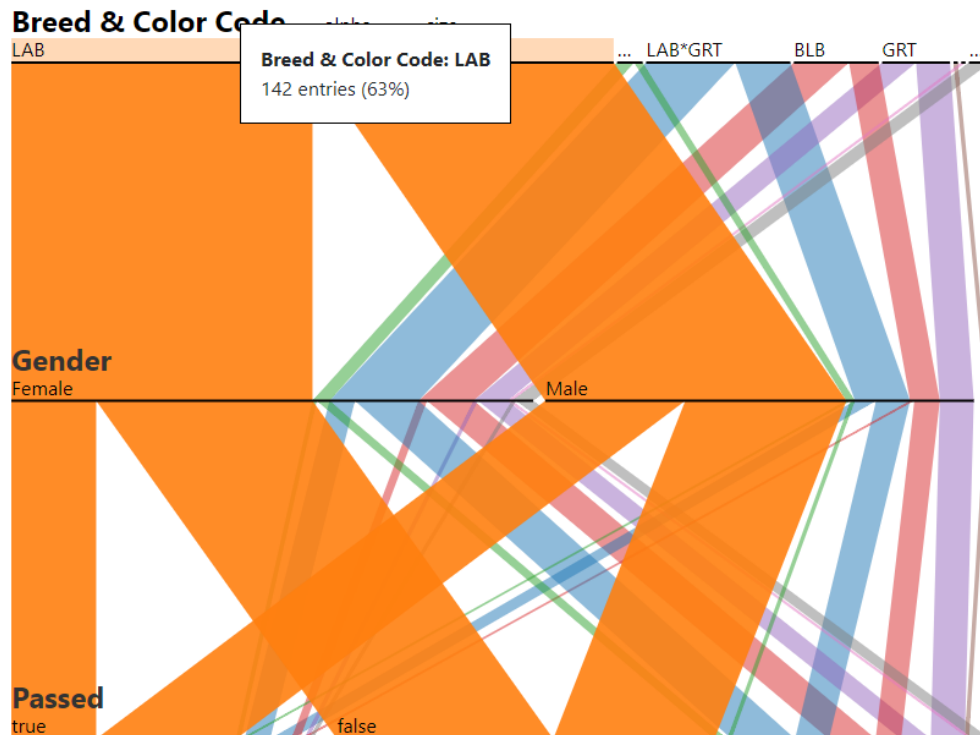
Third User Task: Enable users to **identify traits that influence the entire test result**. For example, a higher percentage of the Labradors passes the test compared to other breeds. Also, 40% of male dogs pass the test compared with 25% of female dogs.
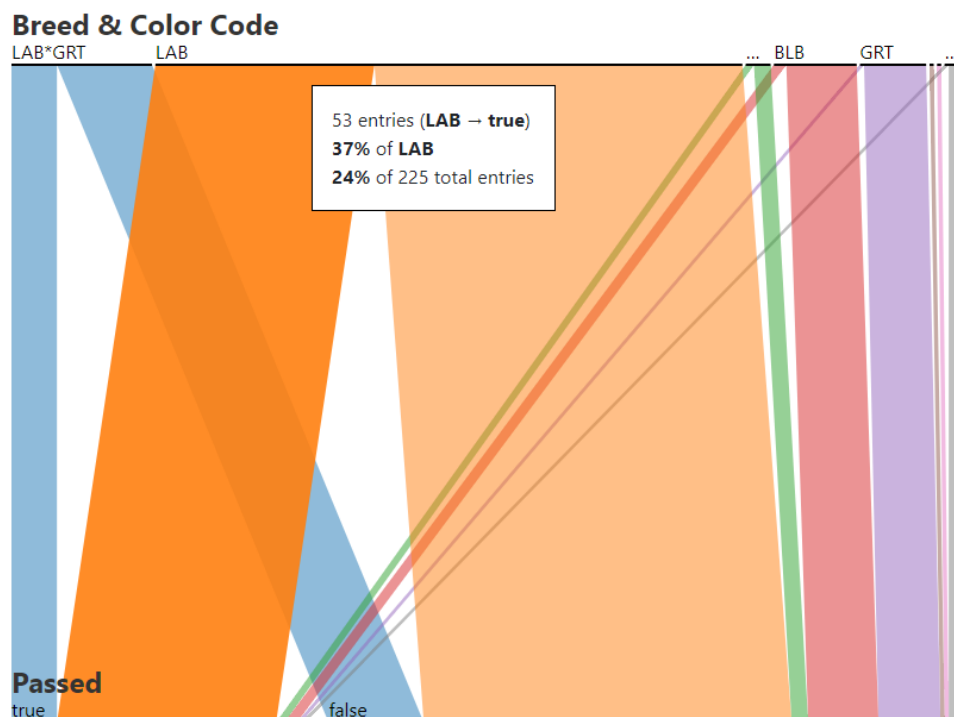
53 entries (**LAB → true**) **37%** of **LAB** **24%** of 225 total entries	3 entries (**BLB → true**) **15%** of **BLB** **1%** of 225 total entries	1 entries (**GRT → true**) **6%** of **GRT** **0%** of 225 total entries

Essentially, discovering the traits and trends that makes a dog more probably to become a guide dog could help breeders and trainers to be more effective when it comes to choosing the dogs they work with.

The "Family Relations" Visualization

First user task: Enable users to **discover the graph of family relations**. For example, you can easily find the parents of a specific dog in the dataset.
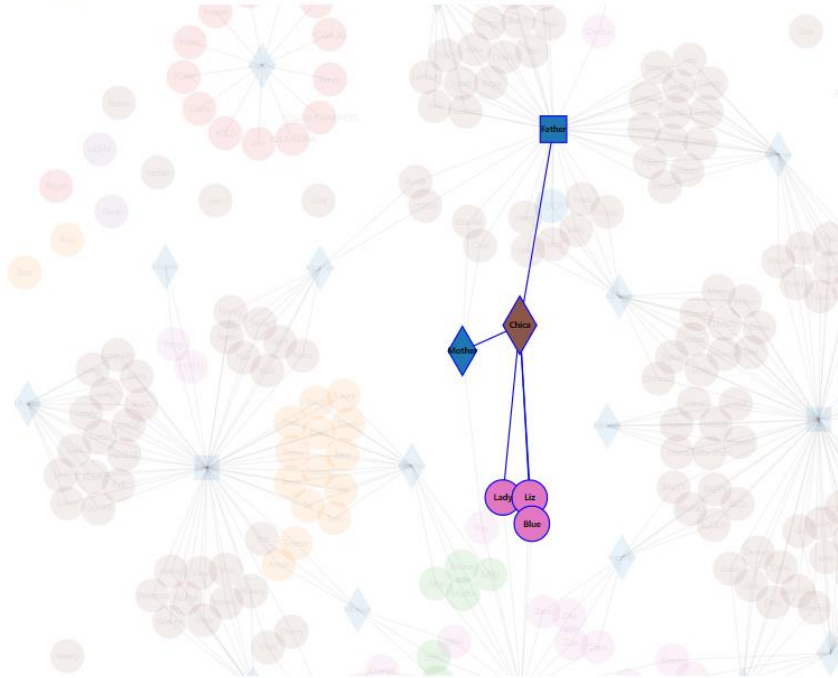
Second user task: Enables users to **discover distribution of dogs by "Breed & Color" or Gender.** For example, brown nodes represent Labradors.
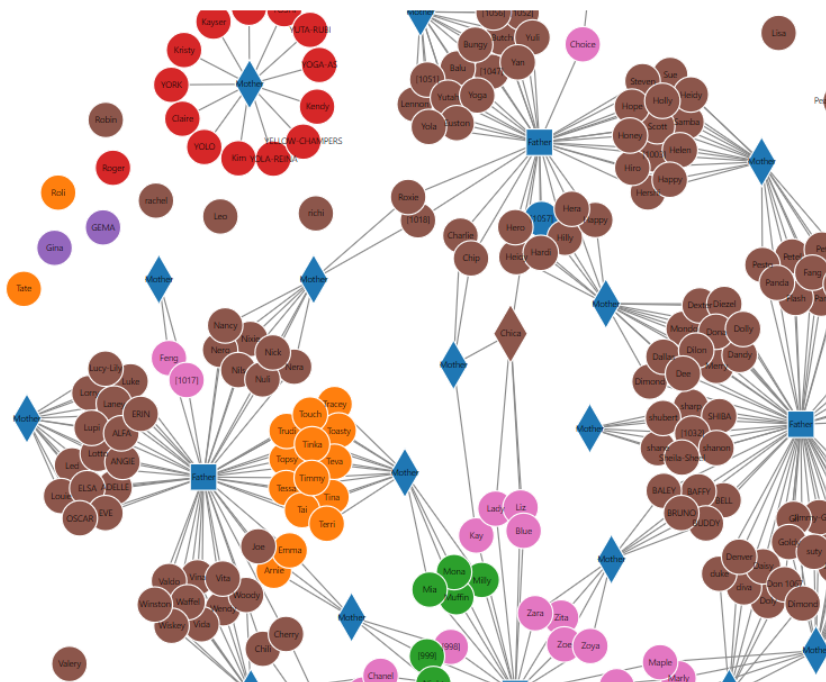
Third User Task: Enables users to **locate dogs who passed or didn't passed**. In this case, blue nodes indicate passes, whereas red nodes indicate failed.
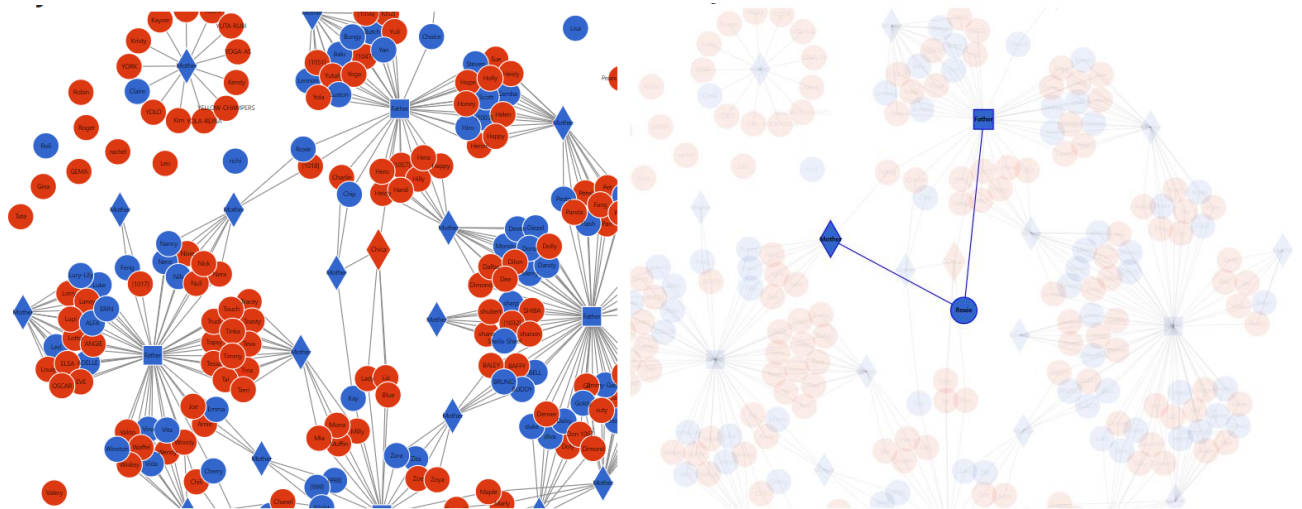
Fourth user task: Enables users to **discover the most suitable parents for breeding**. In this visualization, the darker the node, the more suitable is the parent.
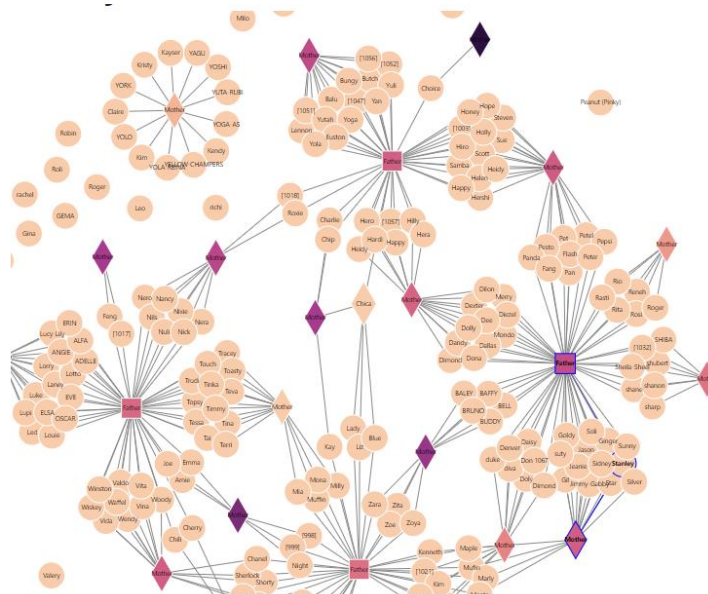
This visualization helps us identify the best parents for breeding, so that breeders could efficiently breed dogs with a higher genetic based chance of success to become a guide dog.

The "Test Results" visualization

<u>First User Task:</u> Enable users to **discover the distribution of scores for each subtest**. For example, the distribution for "המנעות ועכבות כתוצאה מלחץ" is different from the distribution of "תוקפנות רכושנית כלפי כלבים".

<u>Second User Task:</u> Enable users to **identify similar subtests**. For example, many of the subtests concerning aggressiveness ("תוקפנות") are addressing similar topics and therefor results with similar distributions

<u>Third user task:</u> Enable users to **identify high-scoring subtests**. In the example above, the majority of the dogs scores the highest grade (marked by the light blue color)

<u>Fourth User Task:</u> Enable users to **compare similar subtests**. For example, there are a few tests that are related to fears ("פחדים"). However, their score distributions are very different.

13

Fifth User Task: Enable users to **identify key subtests where dogs may fail more often**. For example, many dogs fail on "ויסות-עצמי ירוד" subtest, compared to other subtests in the dataset.

פעילות יתר כתוצאה מלחץ

נרגשות

ויסות-עצמי ירוד

חוסר שקט כשמאמן אינו פעיל

פחד גבהים

נבחנות ללא הפסקה

רמת אנרגיה גבוהה

חוסר יכולת מיקוד / ריכוז

התרגשות מתנועה

רדיפה אחר חיות
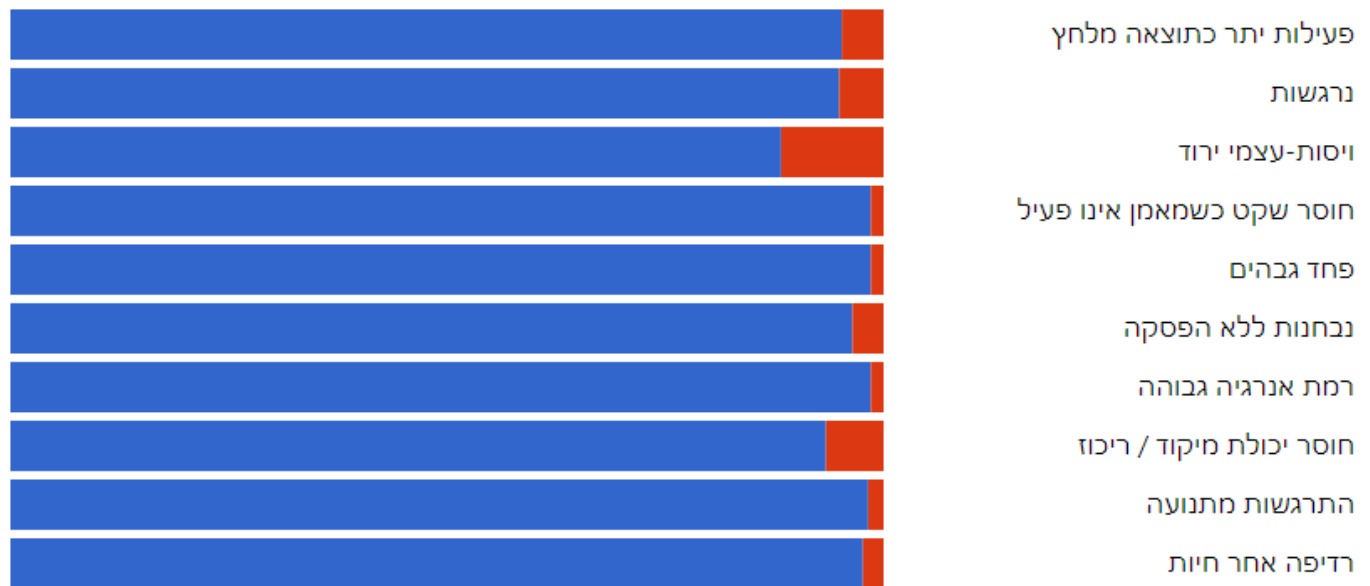
Sixth User Task: Enable users to **explore trends based on different dog traits**. For example, 1.3% of male dogs failed the "נרגשות" subtest (upper bar below) compared to 8.7% of the female dogs (lower bar below).

נרגשות

נרגשות

Identifying these subtests and trends would help the trainers know what things they need to emphasize or strengthen for different kinds of dogs, to increase the rate of dogs that become guide dogs.

Question 3: How? + Evaluation Implementation details

The Dog Data Overview

We use a visualization called "Parallel Sets", a general visualization for multidimensional categorical data.
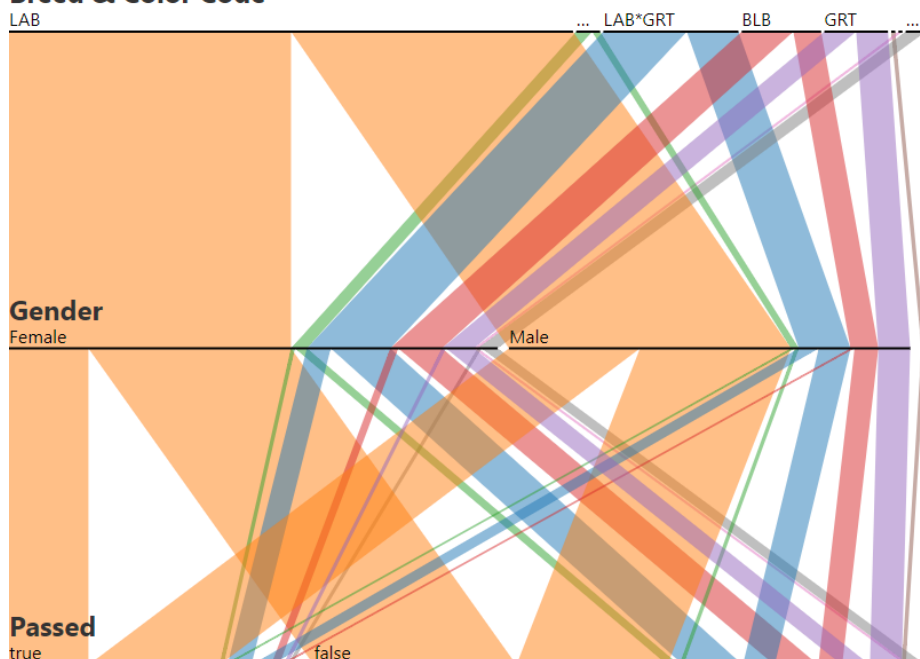
Each selected attribute from the list on the right ("Breed & Color Code", "Gender", "Passed", etc.) is denoted with a set of horizontal bars with each of its possible values. The width of each bar denotes the absolute number of matches for that value.

Starting with the first selected attribute ("Breed & Color Code"), each of its possible values is connected to a number of possible values of the next attribute, showing how this value is subdivided. This subdividing is repeated recursively, producing a tree of "ribbons".

For example, the LAB breed has a similar amount of male and female dogs, but the LAB female has less chances to pass than the LAB male.

The colors of the ribbons are selected arbitrarily using D3's default categorical color scale.

Hovering on these ribbons brings up a tooltip that contains the "path" (e.g. "LAB → Male → true") - the subdivisions of the data up to the next (bottom) attribute. The tooltip also presents the

number of entries in the ribbon, the path to the previous (upper) attribute, the percentage relative to the previous attribute and the total amount of entries in the visualization.
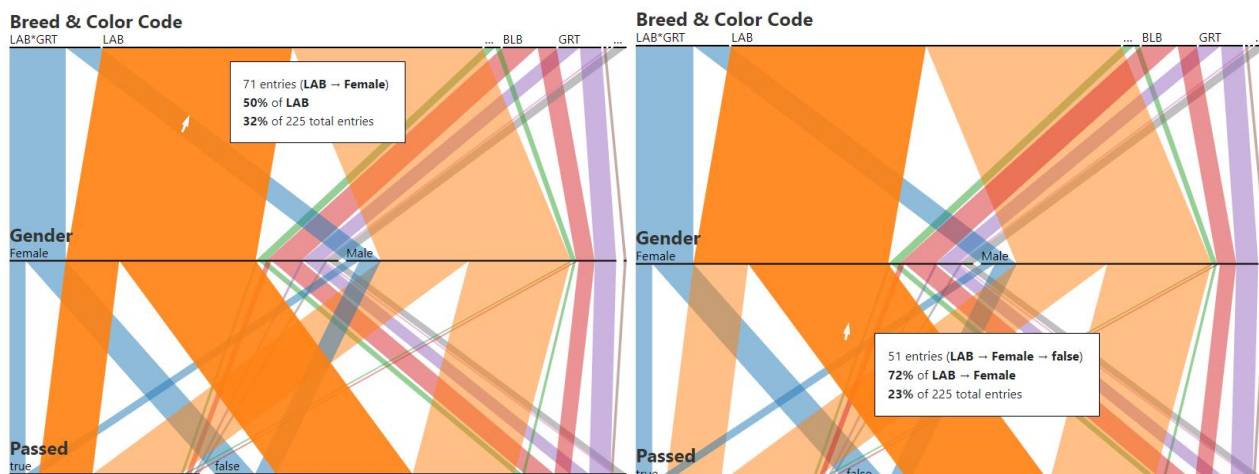
In this visualization, the attributes are the categorical attributes each dog has, as denoted in the data table (a minimum of two attributes must be selected). We also added a sidebar (filters) that allows the user to choose what categorical attributes they want to display.
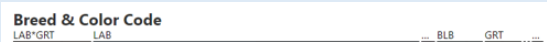
The user can reorder the attributes and possible values by dragging them. They can reorder the values by name or frequency, by clicking on "alpha" and "size" accordingly.
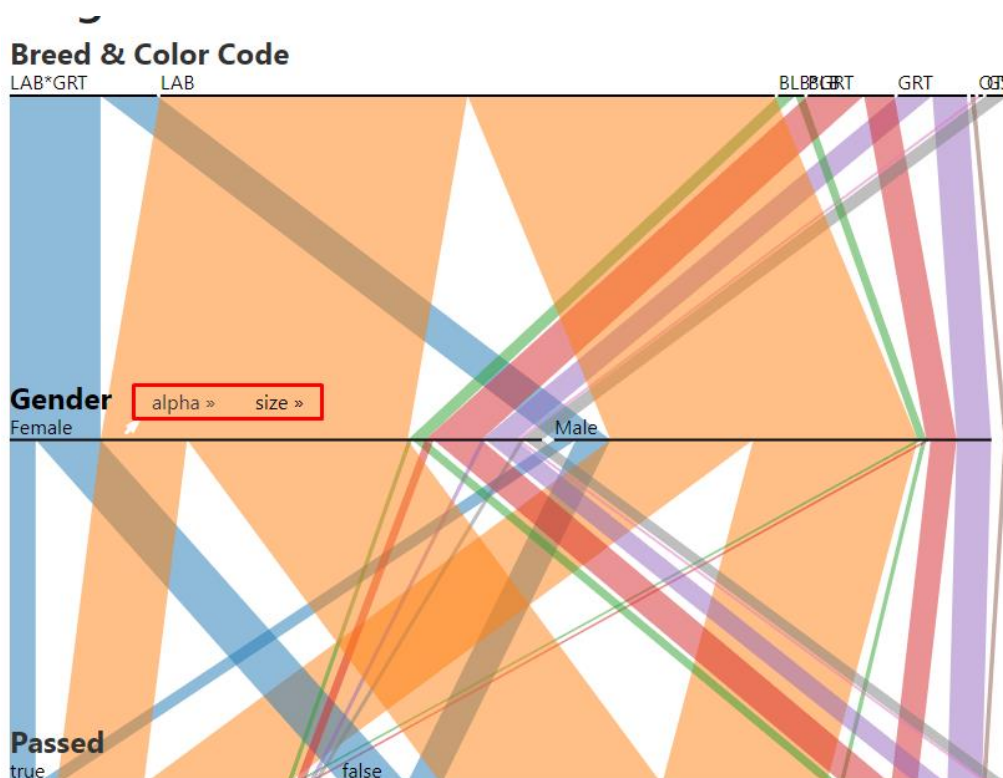
Changing the top attribute changes the color scale accordingly. By hovering, the user can see the information of that particular ribbon.

Breed & Color Code

LAB*GRT   LAB   BLBBGRT   GRT   OGS

Gender
Female   Male

> 51 entries (**LAB → Female → false**)
> **72%** of **LAB → Female**
> **23%** of 225 total entries

Passed
true   false
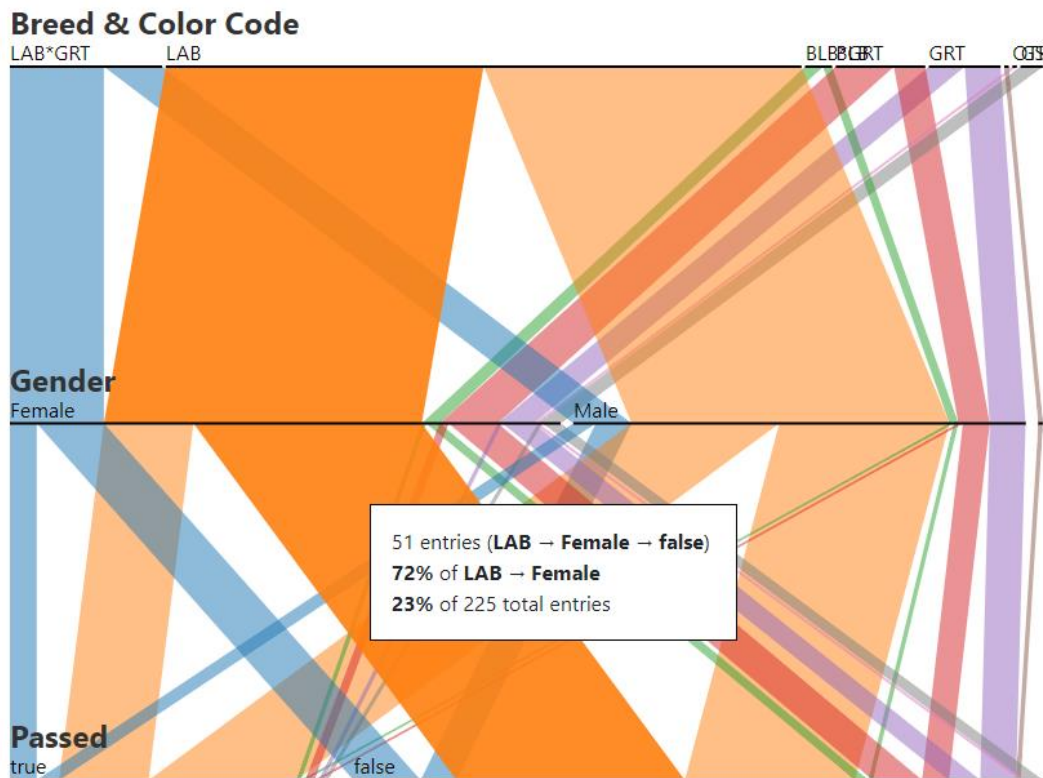
Visual Mapping

Data Attribute	Visual Property
The dogs' attributes	Horizontal "dimensions"
Attribute values	Horizontal lines in each "dimension"
Main attribute (First attribute)	Top row in the visualization Graph colors adjusted according to different values
# entries per attribute	Width of each horizontal line in the "dimensions" The Tooltip: 71 entries (**LAB → Female**) **50%** of **LAB** **32%** of 225 total entries

17

Subdivisions of the data	Subdivisions of the ribbons
	The Tooltip: 71 entries (**LAB → Female**) **50%** of **LAB** **32%** of 225 total entries
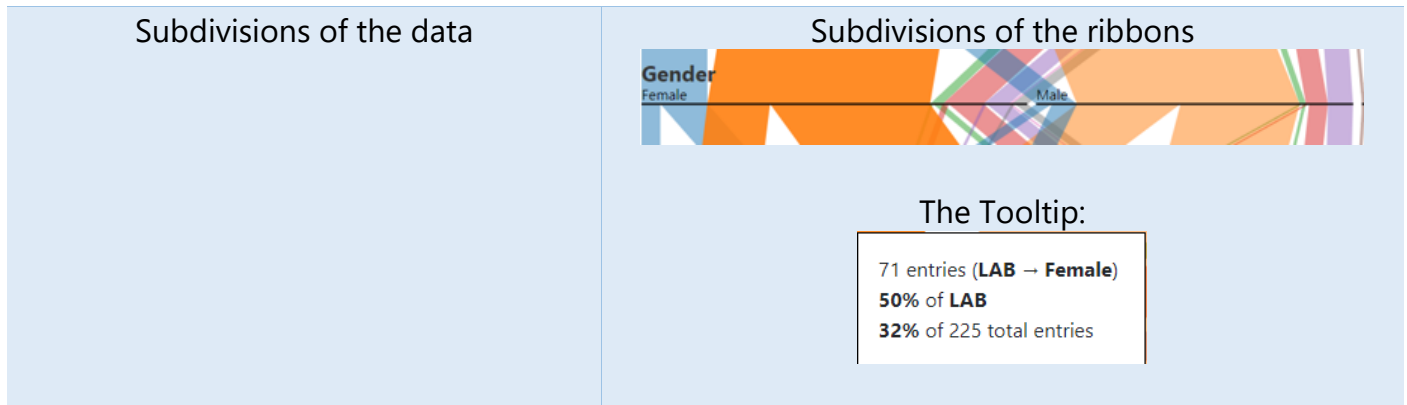
User Interaction

To support user interaction we added a sidebar that allows the user to choose what attributes to display.

Initially, we decided to work with 3 major attributes without giving the user a choice: "Breed & Color", "Gender" and "Passed" (a derived attribute), thinking these were enough for the sake of this visualization.
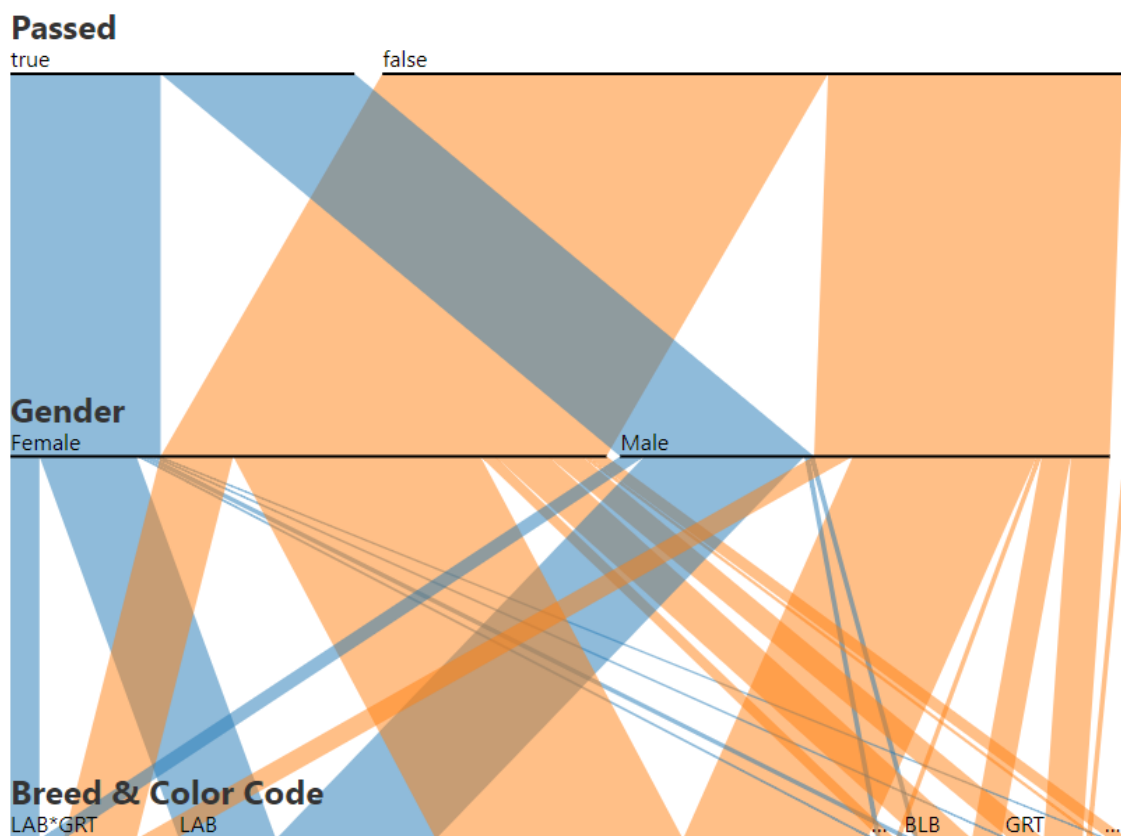
However, we saw that we have more categorical attributes, and adding them to the visualization seemed to provide more insight, we added a set of toggles for the attributes that were currently there, so the user can enable/disable them by choice.
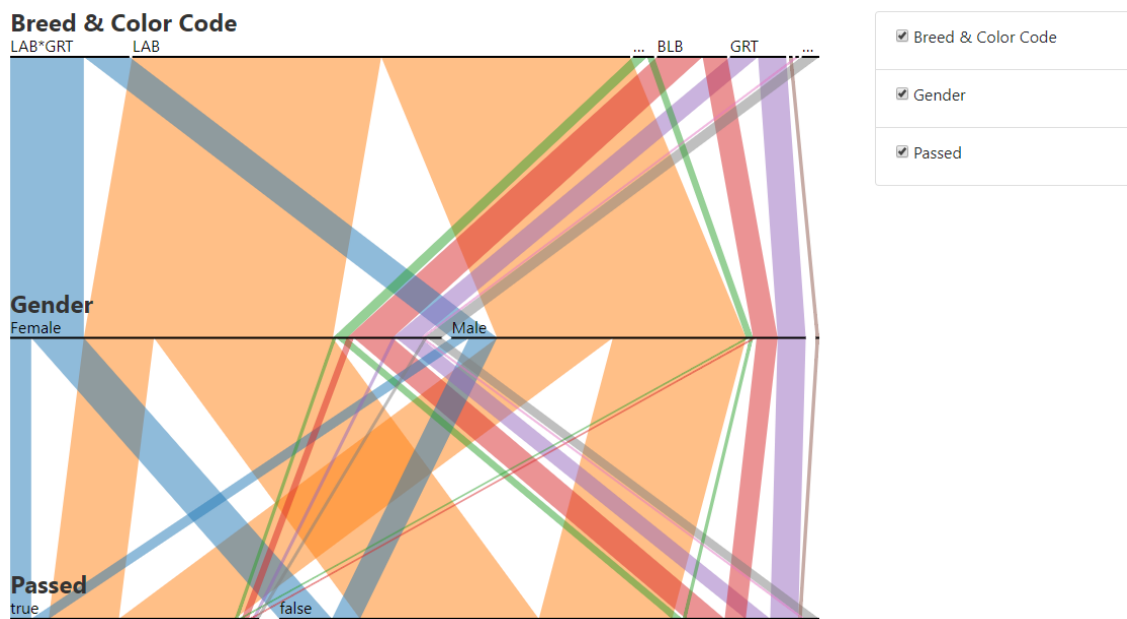
And then added more attributes:
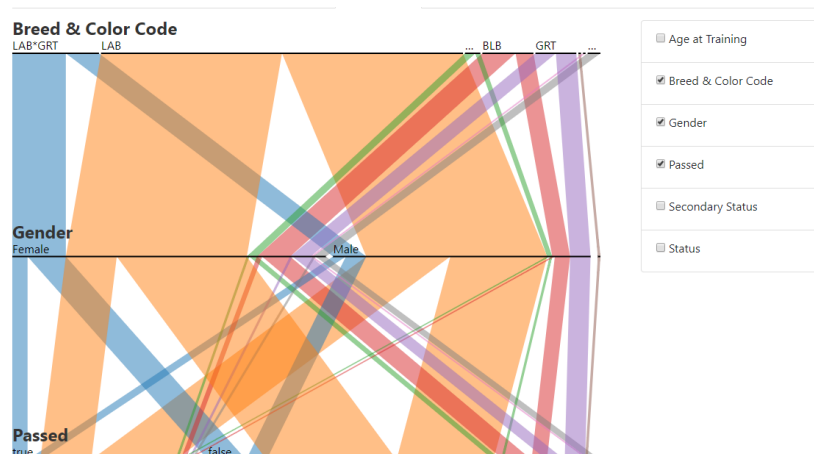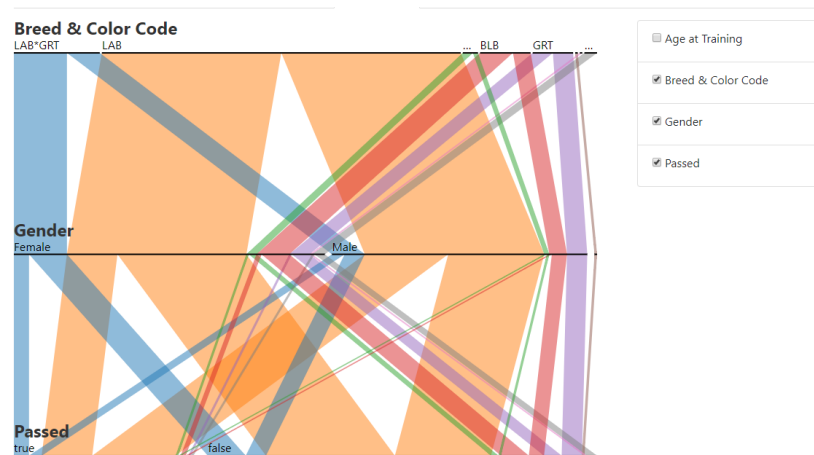
Evaluation

This visualization enables user to quickly group and compare the dogs by their different traits.

In addition, it enables a quick way to search and find correlations between traits in general, and traits that may be more important for guide dogs in particular. (Likelihood to pass the test).

This visualization clearly shows how many dogs passed and failed the test relatively to the rest of their attributes.

The Family Relations View

This visualization uses an interactive, undirected mathematical graph. It shows the 3 different aspects of family relations, with an auto-updating legend.

In all of the visualizations, circles represent children, diamonds represent mothers, and squares represent fathers. Dogs that are related are connected on this chart. Children not connected to anyone are considered Orphans (circles not connected to anything).

The first one shows the dogs that passed (blue) or did not pass (red), and their family relations.
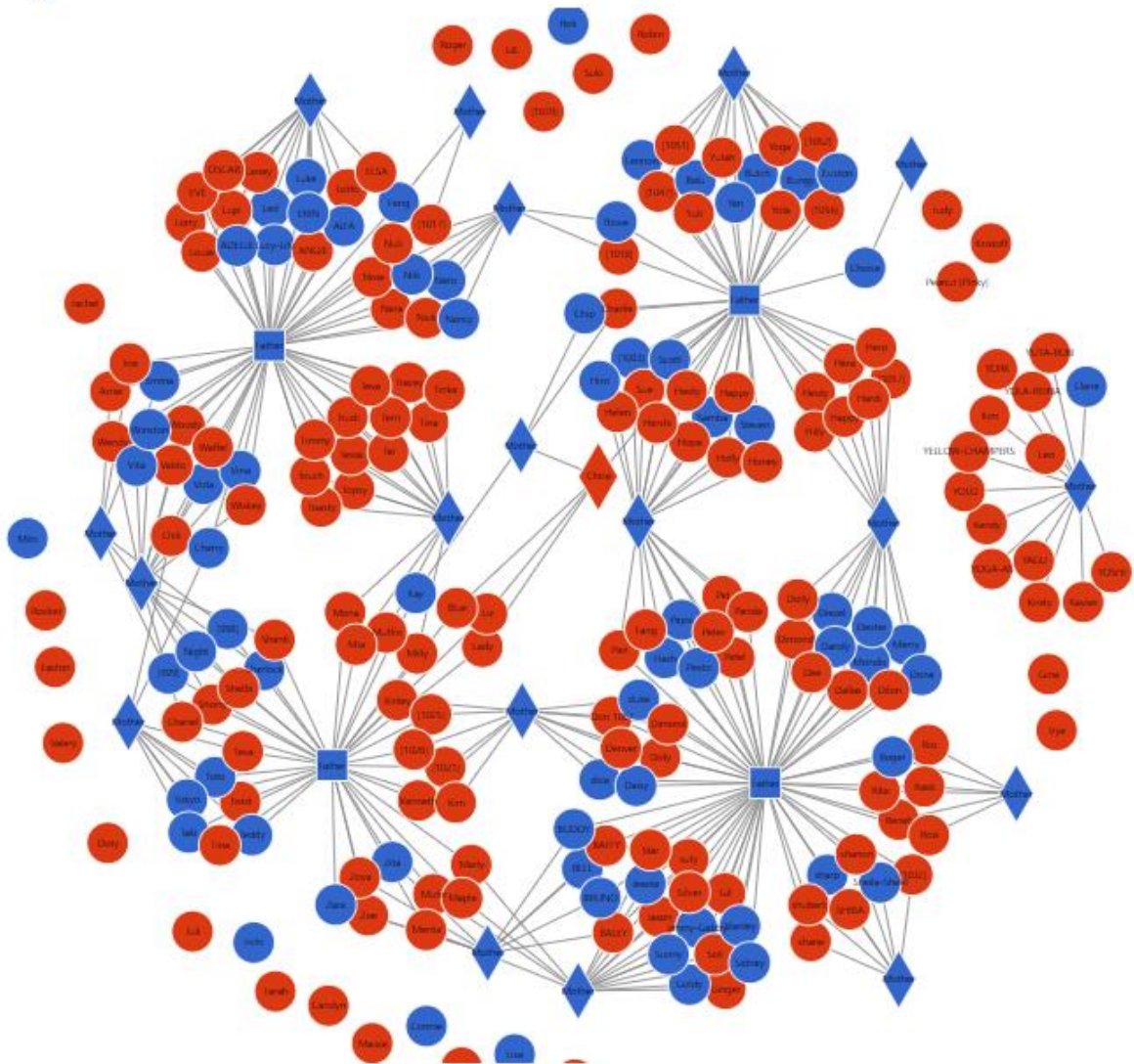
The second one shows the how common is each breed. For example, you can see that the LAB (brown) is the most common one in the graph.
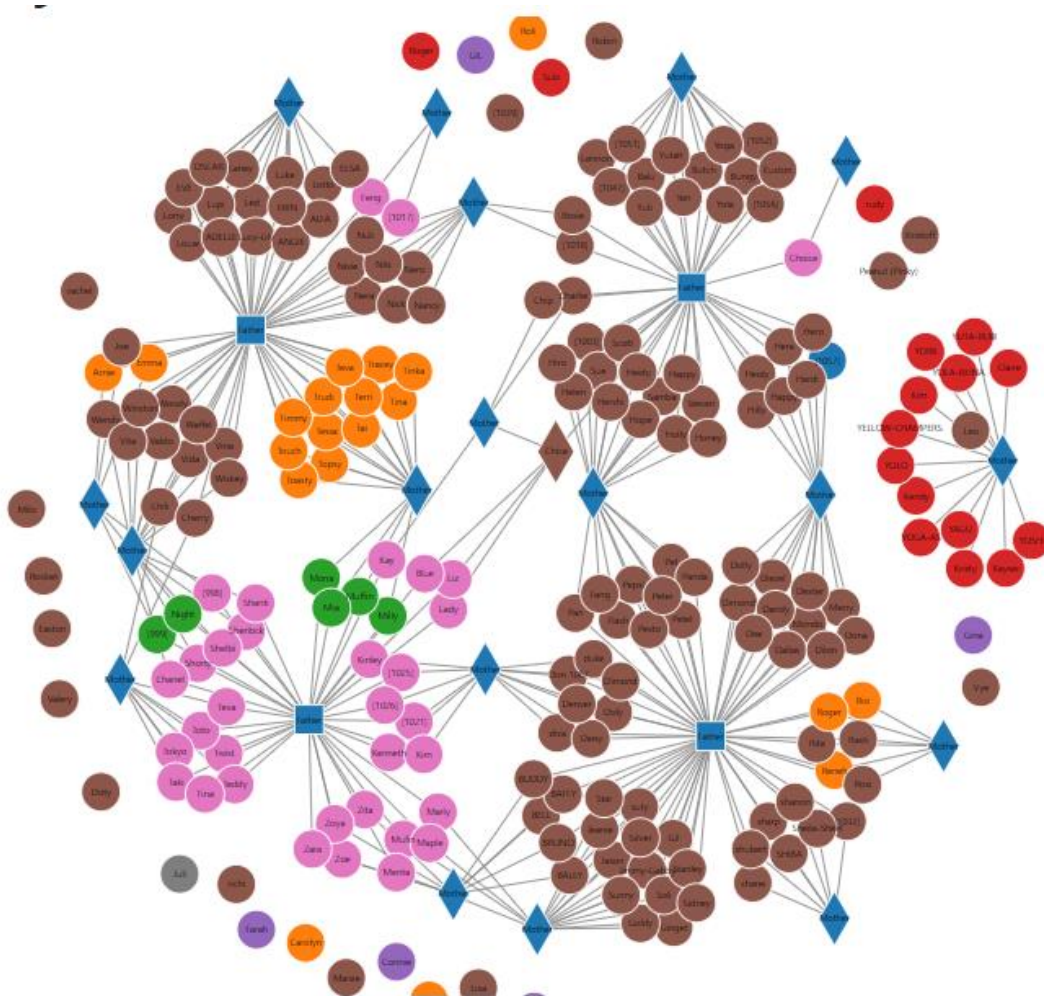
The color scale used for this mode is google's 10c scale. We chose this scale because the colors are more vibrant and the order is more intuitive than d3's category10.

The third one shows the likelihood of a specific parent to have a child that passes the test. The darker it is, the higher is the likelihood. In this chart, the black node on the top right has the best chances.
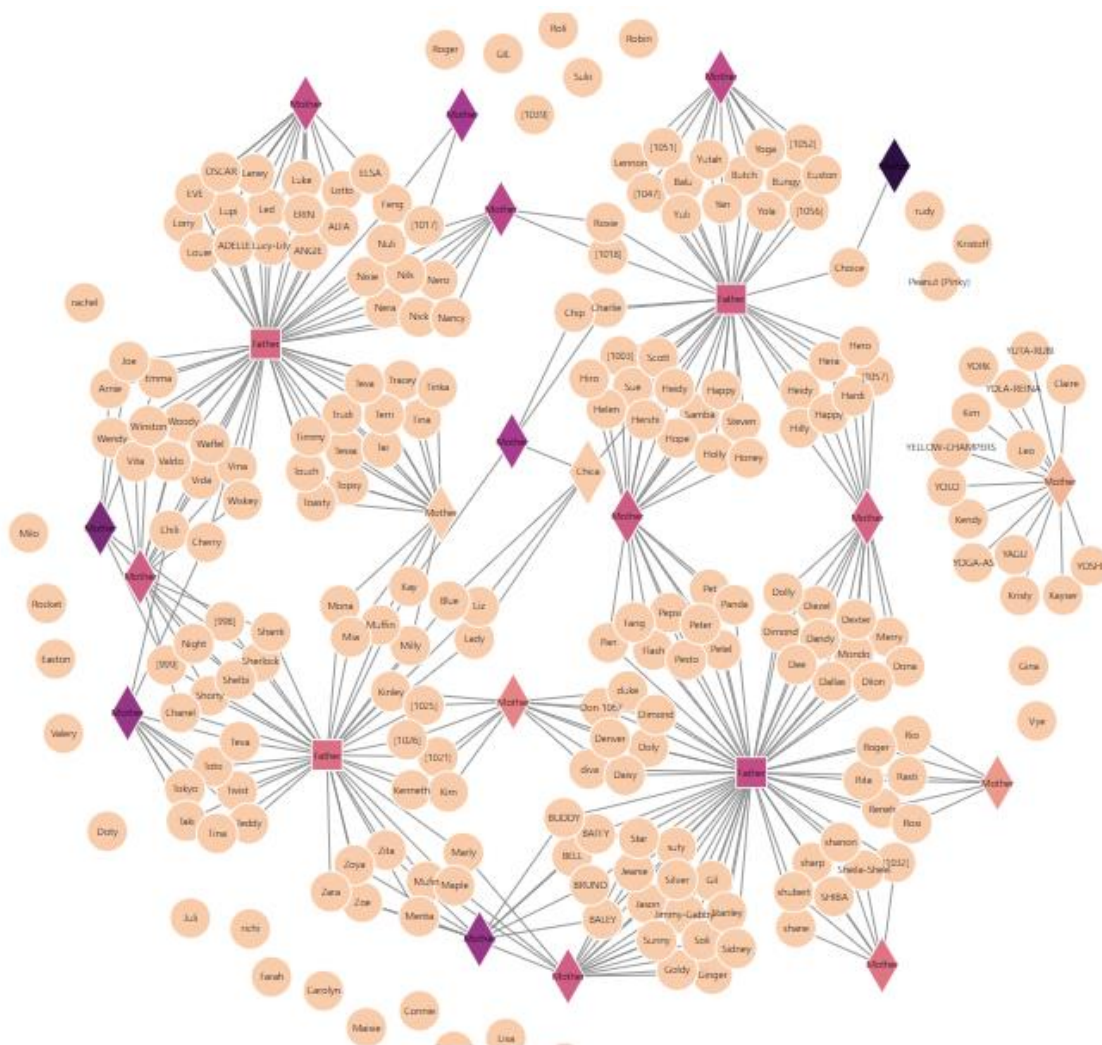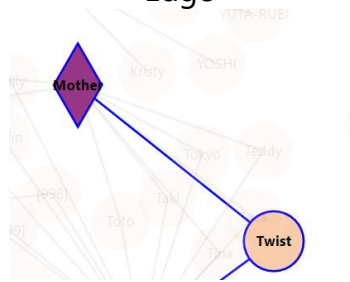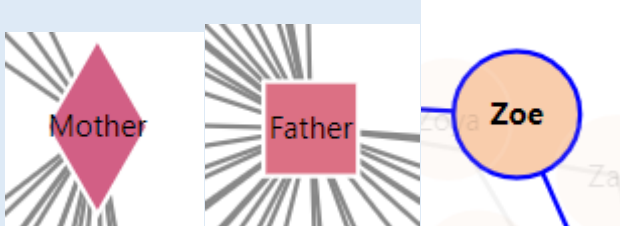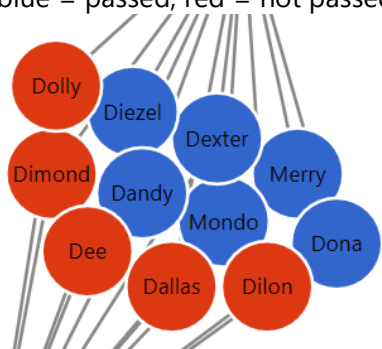
The color scale used here is a custom color scale I found that is colorblind-friendly.

Visual Mapping

Data Attribute	Visual Property
Each node	A shape on the graph.
Relation between two dogs	Edge
Node kind	Shape: Diamond for mother Square for father Circle for a child
Passed/Not Passed	Color in "Passed/Not Passed" view. (blue = passed, red = not passed)

Breed & Color Code	Color in "Breed & Color Code" view. (using google 10c color scale)
Score	Color in "Breed & Color Code" view. (using a custom scale)
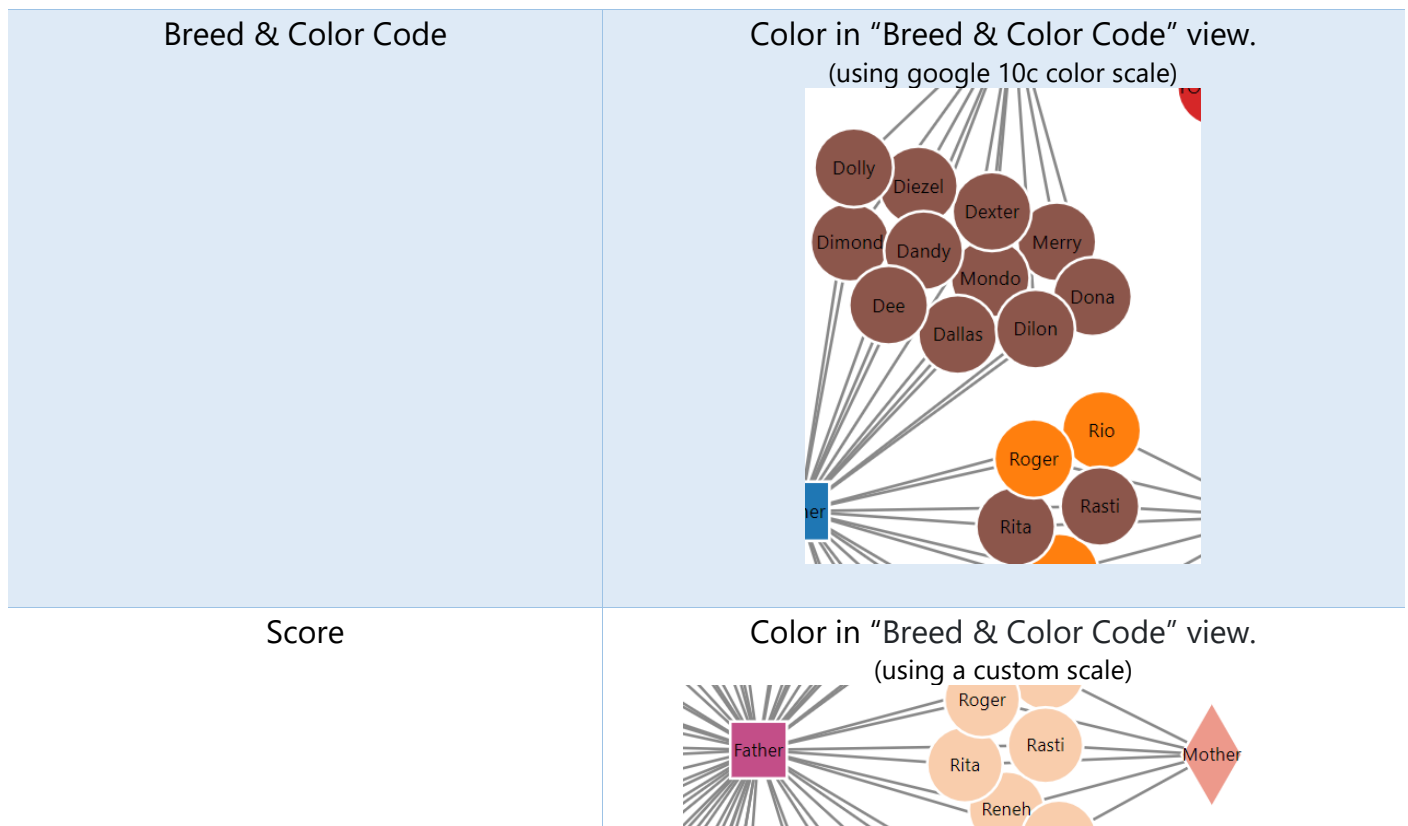
User Interaction

Users can interact with the visualization in multiple ways. They can use the display filter to choose the between the three charts available. They can also use the show selectors to hide/highlight specific elements on the chart.

Users can zoom in and out the graph itself using the mouse wheel.
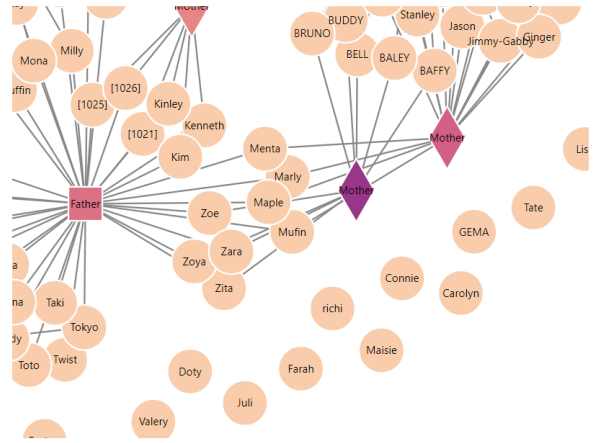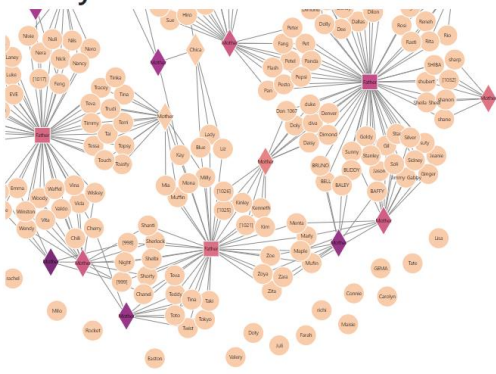
When they hover a node, this node and the nodes directly connected to it are highlighted.
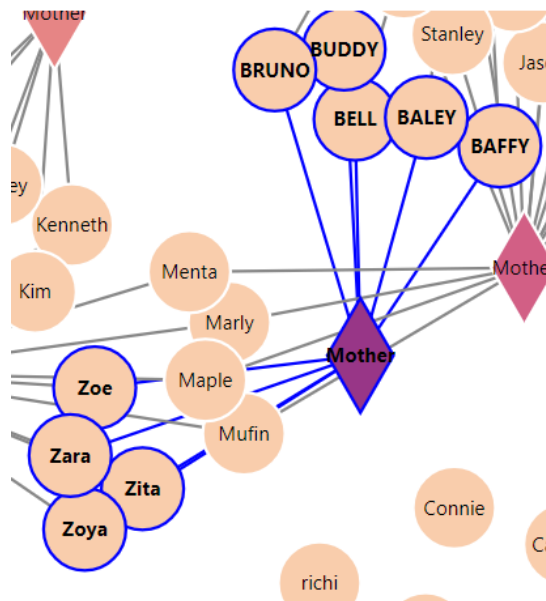
If the user also left-clicks, all the other nodes are hidden.
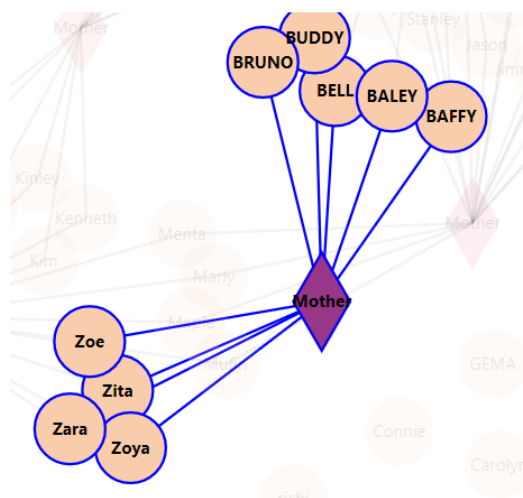
Evaluation

User can immediately see the dominant breed, how the breeds are distributed, and identify siblings and other family connections. Users can also see which dogs passed and which didn't, and they can use the heatmap to determine which dogs are best for future breeding.

Users can zoom out and see the overall picture. They can also see the places where data is missing. Unfortunately, we didn't get a better dataset from Dr. Zamansky until the deadline, so we couldn't improve our data.

Tests results visualization

This visualization is a list of stacked bar charts. It shows the 3 different aspects of family relations, with an auto-updating legend.

The first one is a categorical distribution of the grades. Each bar displays the score distribution of the specific subject (Color codes for the score is in the legend).
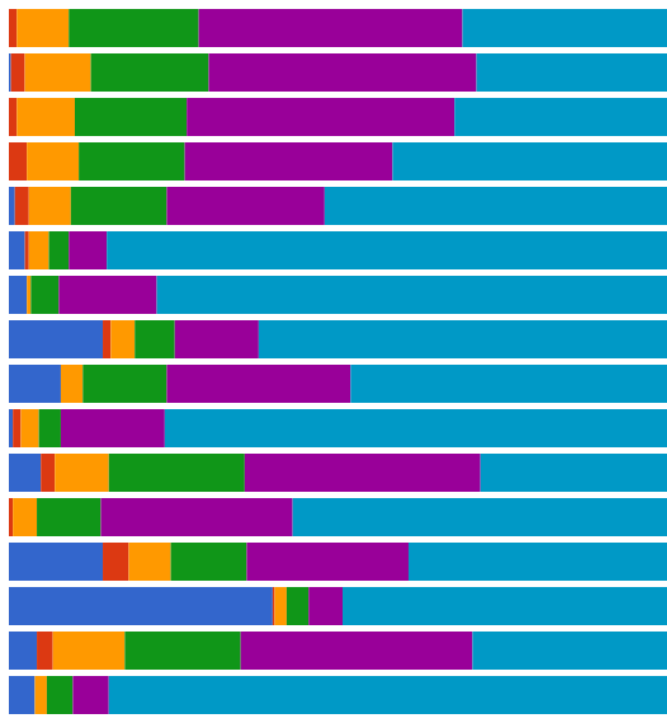
The second one highlights the probability of achieving different grades. Here, the darker the color of the bar, the more likely this would be the grade for that subtest. For instance, when we look at "פחד ממדרגות", we can see that it's very likely that most dogs will get good scores in this subtest.
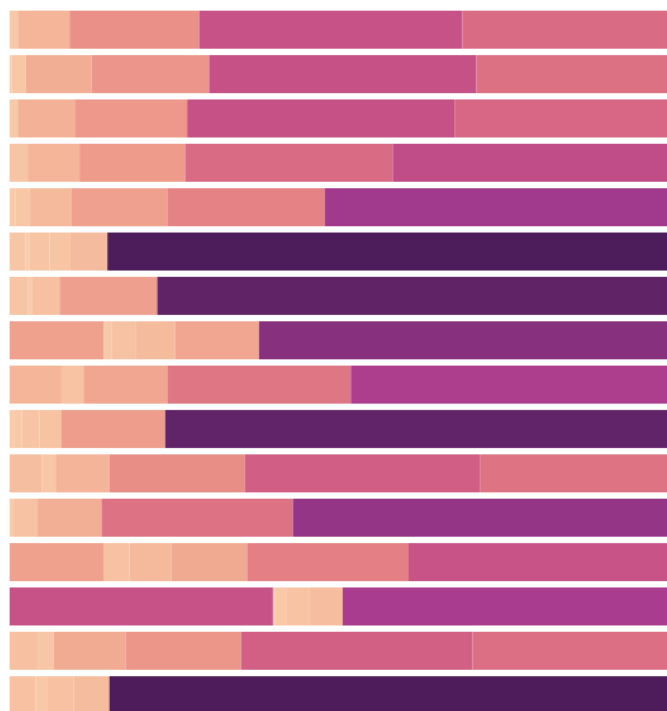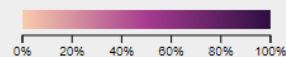
The third visualization highlights the subtests that are more likely to fail the entire tests. The blue bars represent the % subtest that weren't disqualified, and the red bars represents the % of the tests that were disqualified.

For example, "ויסות-עצמי ירוד" subtest was a leading reason for disqualification.

Visual Mapping

Data Attribute	Visual Property
Subtest	**A stacked bar chart** (Row in this visualization) חרדה במקומות לא מוכרים
Number of filtered subtests	**Width of the stacked bar charts** (Default vs. Female dogs only) חרדה במקומות לא מוכרים חרדה במקומות לא מוכרים
Different options	**Different bars** **Colors in option 1**
Number of entries	**Width of each stacked bar** Tooltip: **Option:** "4" **Tests:** 132 (39.88%) **Culmulated:** 227 (68.58%)
Probability of options	**Colors in option 2**
Passed/Not Passed	**Colors in option 3**

User interaction

Users can switch displays using the radio buttons.

Users can show or hide the labels using the checkbox

Users can search for a particular subtest using the search bar
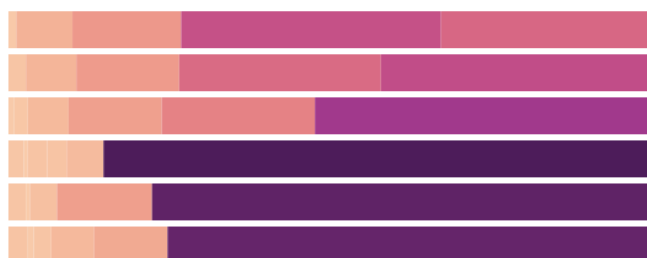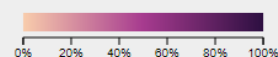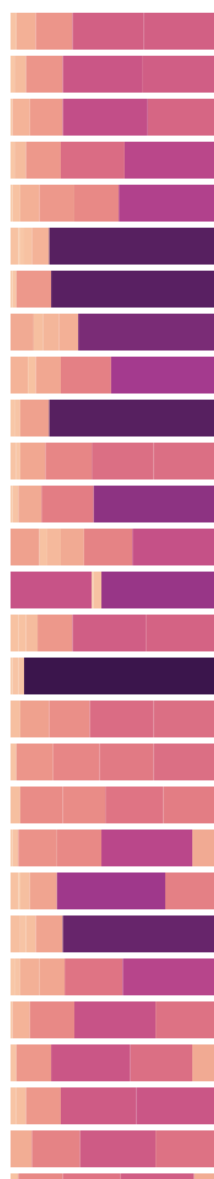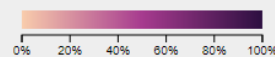
Users can filter the results based on gender and breed.For example, we're filtering anything out but female Labrador dogs. Note that the width of the stacked bars is changing, based on the number of results in the selection.

Hovering on a bar shows the full details of that option.

Evaluation

This visualization provides quick insight to distributions of the subtests results.

In particular, it enables users to see the most common grade in every subtests and other statistics (high/lower values, and high/low probabilities).

This visualization also allows us to quickly identify key subtests when dogs may disqualify the entire test more often. This can be done across the entire population, or for more specific groups.

Evaluation

References

Code Resources Used:

- D3 library - https://d3js.org
- D3.parsets – Parallel Sets for D3 - https://github.com/jasondavies/d3-parsets
- D3 queue - https://github.com/d3/d3-queue
- Bootstrap 4 beta – https://getbootstrap.com