Process book

Bogdan Kulynych, Cosmin Rusu, Jean Bejjani

Watershed flows

December 13, 2018

Introduction

This document will guide the reader through the process of creating the visualizations for the project for Data Visualization (CS-480) EPFL course. It will lay out the motivation, scientific background, and pointers to using and understanding the visualization.

The project is made with Natural Capital Project (https://naturalcapitalproject.stanford.edu) and P. James Dennedy-Frank, a post-doctoral researcher at Stanford University.

Team

1. **Cosmin Rusu:** Data Science Master's @ EPFL
2. **Bogdan Kulynych:** Ph.D. student @ EPFL, working on Privacy, Security, and Machine Learning
3. **Jean Bejjani:** Masters in MicroEngineering and Robotics, Minor in Computer Science @EPFL, and Research Assistant @ LIS EPFL.

Motivation

Source-water protection and nature-based infrastructure have become major investment drivers for water security around the world. In many cases, these programs have project goals that include increased water during the dry season and decreased flooding.

However, such programs have often failed to seriously investigate the evidence about the likely water quantity changes that would occur in such programs.

There is a need for a good visualization to show managers and other stakeholders the evidence from existing studies of the changes in river flow after the controlled land-cover change.

This will help them base their decisions on the best evidence available.

The managers will want to know:

1. What is the river flow response to potential interventions?
2. What is the river flow response to potential interventions in sites like theirs?
3. Where are the gaps, in which we don't know what the likely changes in river flow may be?

Target audience

Our target audience are scientists and stakeholders concerned by how to manage and maintain water distribution. Scientists are interested in studying the impact of different environmental treatments on natural watersheds. Climate change is expected to cause an increase in the occurrences of floods and droughts. Our target audience contains also the people with the will and power to use the results of this project to effectively implement the right actions in the right places in specific watersheds.

Data description

Watersheds, or catchments, are pieces of land where precipitation collects and drains off into a river, bay, or other body of water. It is important to be able to manage watersheds for ecological reasons-for maintaining natural habitats and climate-and geopolitical reasons-watersheds are often used as political boundaries.

Watersheds can be controlled through constructions like dams, or using nature-based solutions, like changing cutting out the forest around the catchment. One of the ongoing tasks of Natural Capital at Stanford is understanding the performance of various nature-based solutions in different conditions.

Our goal for this project is to summarize data from studies that performed experiments with watersheds in easy-to-comprehend graphics. The users should be able to dig in the details when needed. Finally, we should show and take extra care about the statistical significance of the studies.

Dataset description

The dataset contains extracted key information from various studies that performed experiments with watersheds around the world in the last several decades. It is the result of ongoing research; currently, it contains 34 rows describing different study results.

The data will come in an SQLite relational. The main table contains data about particular watersheds where experiments have been made. The rows contain data about the watersheds information (lat/long, rainfall, elevation, etc), the results of the experiments performed (experiment type, amount & type of flow change, and the normalized version of the results). This table is linked to a watershed that serves as the control and to a table that has relevant references.

This database contains information from a number of studies on the effects of different types of land-cover change on streamflow. **catchmentExperiments** is made up of 4 primary tables: **catchmentSites**, **experimentalCatchments**, **controlCatchments**, and **referencePapers**.

Catchment Sites

Each site is made up of multiple catchments, each of which may serve as a unique study. **catchmentSites** provides basic information about the sites where experiments have taken place, while **experimentalCatchment** provides detailed information about the specific catchment in which a study was done and the results of the study.

Control and experimental catchments

controlCatchments are the catchments which serve as controls for those in the experimental catchments, while **referencePapers** provide information about the source of the results. The results of 4 different hydrologic outcomes of interest are included: annual water yield (the total flow out of a watershed in a year), as well as measures of low-flow, peak flow, and groundwater recharge that can vary between sites-these, should be comparable across sites on a relative basis despite the difference in the actual measures. Both **controlCatchments** and **experimentalCatchments** have foreign keys to **catchmentSites** to link the catchments to the

sites. **experimentalCatchments** also has a foreign key to **controlCatchments**. Finally, both **catchmentSites** and **experimentalCatchments** have lists of citations that have foreign keys to **referencePapers** so that each catchment is linked with the source of the information for reliability. Further detail on the fields in each table will be provided later in the text below.

The dataset is getting updated

Natcap team is working on filling the database as we are writing this, so it changes and gets updates all the time. You can download the latest database that we have from our Github repo.

Implementation

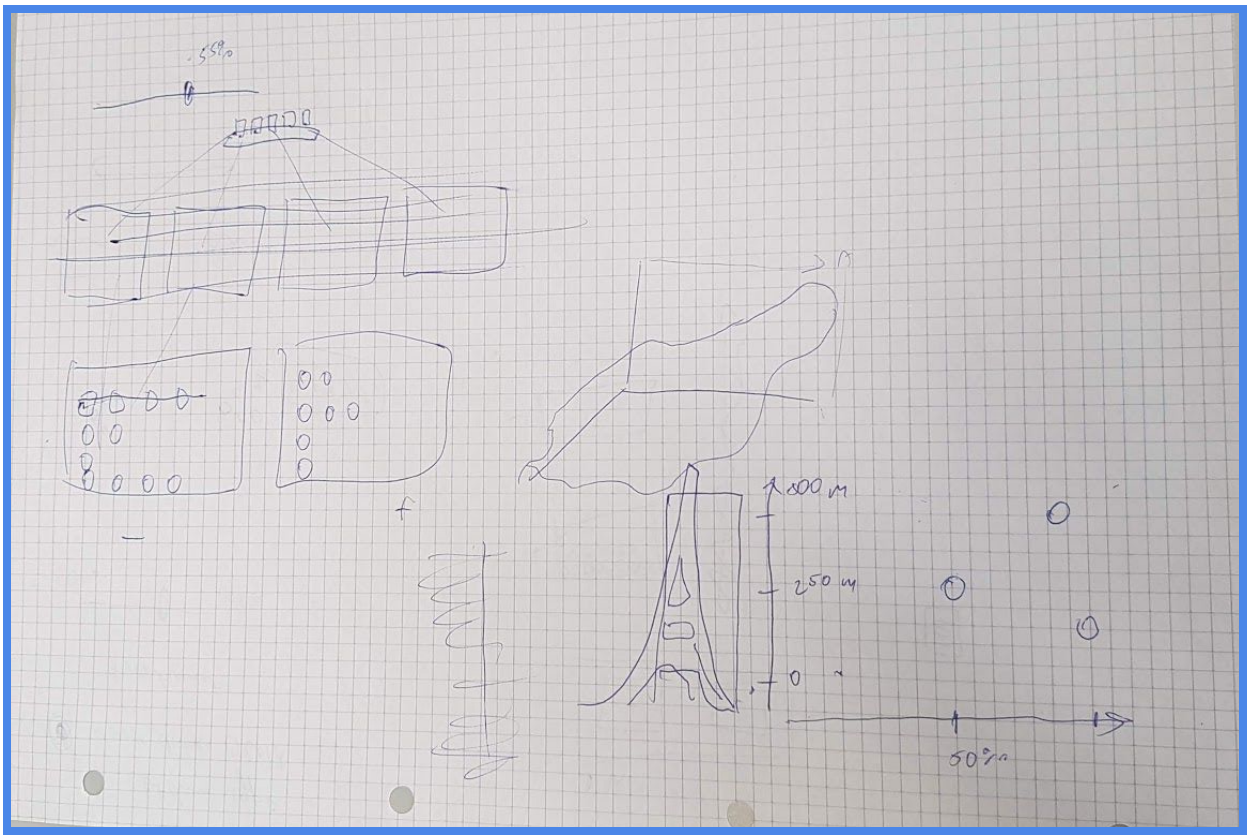
We started by understanding what the project is all about. We scheduled a video conference with James Dennedy-Frank to understand exactly what he would like to be able to visualize as well as to ask the questions we initially had. After this initial call, we have split our project in 4 milestones. In the first milestone, we brainstormed and crafted initial wireframes that we then presented to James in another video conference. In the second milestone, we crafted a prototype of the kind of visualizations we want in Pandas. In the third milestone, we implemented the visualizations we thought to capture the most information and are the most useful to the project, as interactive visualizations using D3.js. In the last milestone, we simply put everything together and we wrote this book.

Milestone 1

In this milestone, we clearly defined our goals and what answers we want to ask. We coordinated with the data owners to understand their needs and we crafted our ideas around the project. We brainstormed and tried to split the problems into smaller problems that we can tackle. We have also set up a Trello board for us to be able to distributively work together towards the same goal.

Milestone 2

The second milestone was mainly about trying different things and getting feedback from the NatCap team. We would craft some visualizations ideas and even prototyped some in Pandas.



Milestone 3

In the thirds milestone, we already had some satisfying ideas and visualizations prototypes, so we started crafting the interactive data visualizations in D3.js, and started writing the story that the webpage will tell. To this extent, we asked the NatCap team exactly what their needs were, where they wanted to go next, and what is the real motivation behind the research that has been done and is being done.

Milestone 4

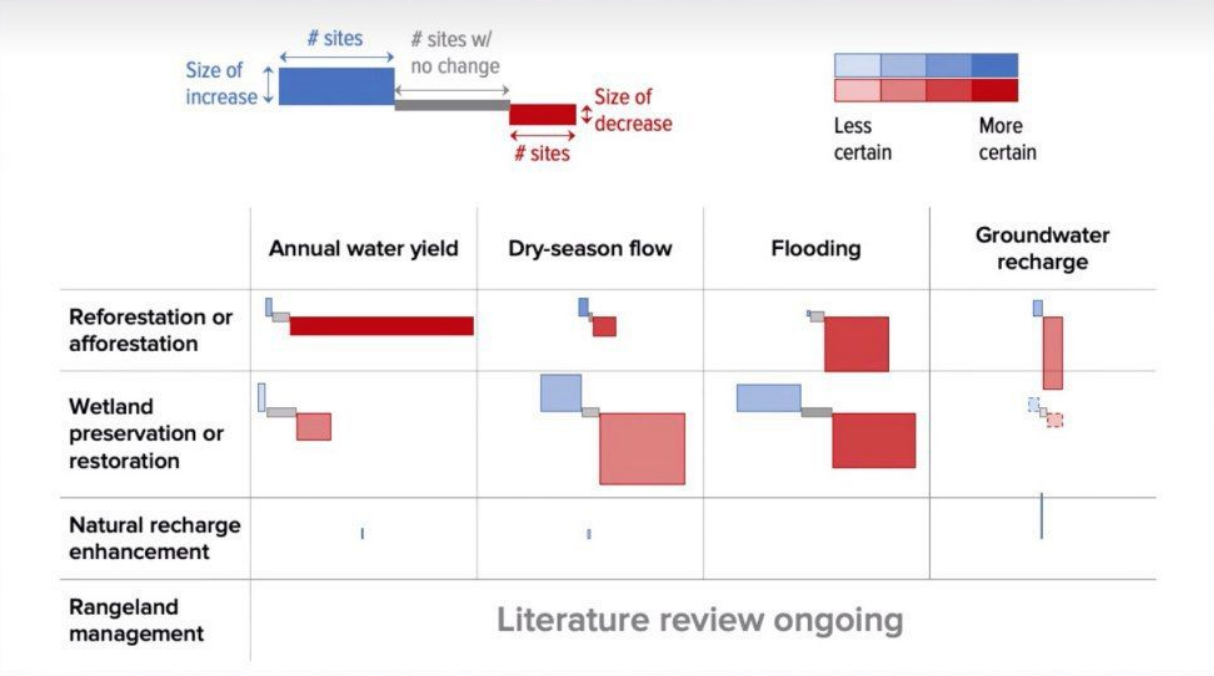
The last milestone was about putting every piece together and writing down this report book.

Concept

We started by meeting with James to discuss what exactly are the requirements. After we understood their end goal and we have clarified most of the questions we had with regards to the project, we agreed on meeting another date so that we have time to brainstorm and came up with ideas. That way, James would not bias our thinking, and we could have full freedom and creativity.

Brainstorming

After coming up with many ideas, most of them very optimistic (because although the ideas were great, the data was either not available or it was in a small amount). After presenting our ideas to James, he showed us what he was thinking about. In particular, the most important thing he said is that they need more things, more dimensions, to be seen in our visualizations. He then showed us the following prototype:



There are a lot of problems with the above visualizations:

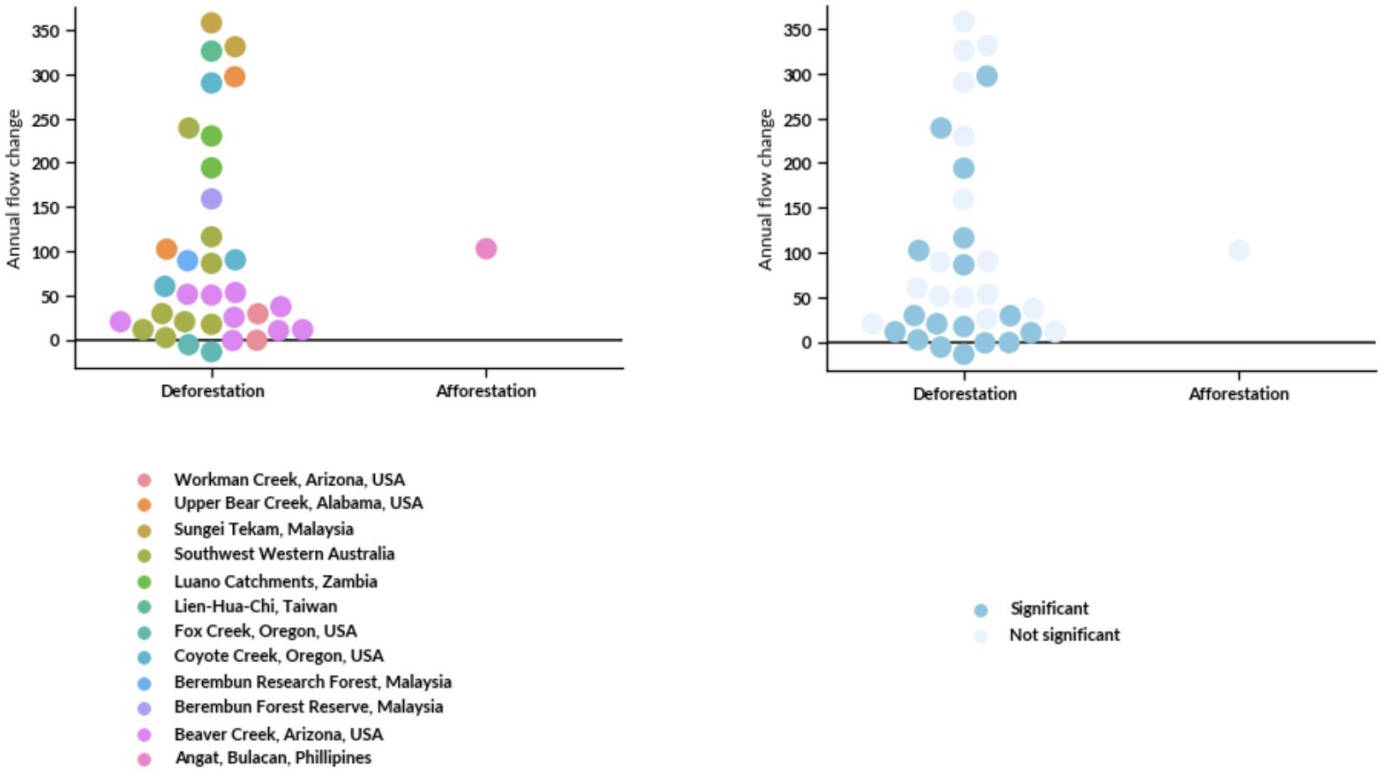
- The area does not clearly give information;
- Some rectangles are too extreme (i.e. one dimension is much higher than the other one);
- The colors reflect the aggregation, not individuals;
- The rectangle is interpreted as area (one quantity), but it contains two dimensions of the data;
- No fine-grained breakdown;
- It is unparseable.

Inspiration

Our inspiration came from brainstorming a lot and trying out different prototypes. We have started doing data analysis on the small data we had, and the swarm plots were the most natural visualization we have thought of. Moreover, for an aggregate perspective on the data, we have chosen to create histograms. This was needed because sometimes the people were interested more on the whole overview of the experiments, not on individual experiments.

Initial prototype

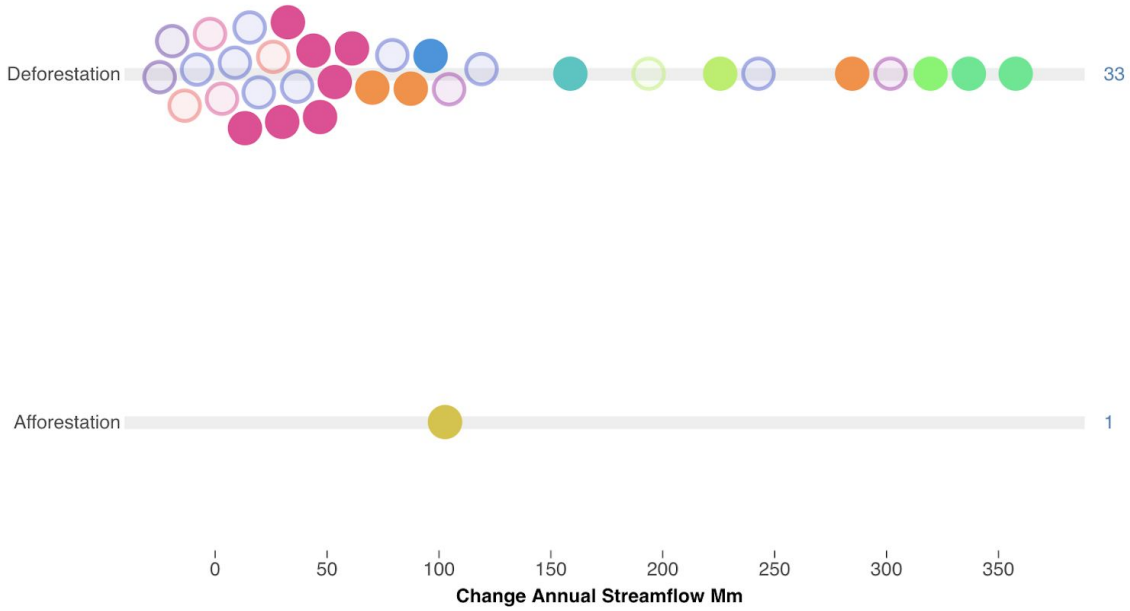
We came up with the following prototypes:



In the first **swarm plot**, we have successfully been able to give the viewer a much broader set of information. First, the colors represent the **sites** of each experiment. The fact that our points are represented as circles given the user the whole overview of the experiments across all experiments. On the X-axis, we show the available **Treatment types**: in this case, deforestation (cutting trees) and afforestation (plant trees). The Y-axis represents the **measurement** conducted after the treatment has been performed. In this case **Annual Flow Change**. Other measurements are: **Change Peak Flow**, **Change Low Flow** or **Change Annual Streamflow.**

The second **swarm plot** (the one on the right) removes the information about the sites of each experiment and instead uses color opacity to display significance or non-significance accuracy of each experiment.

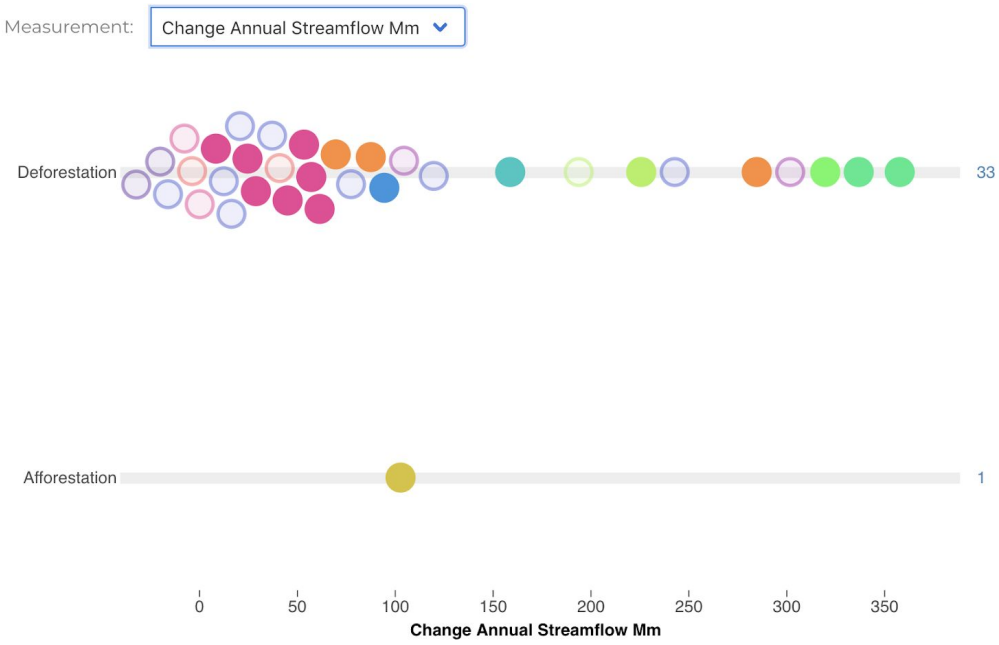
In our final plot, we have combined both **ideas** and created a **swarm plot** where we have used different colors for different pallets and opacity to encode the significance or non-significance of the experiments.



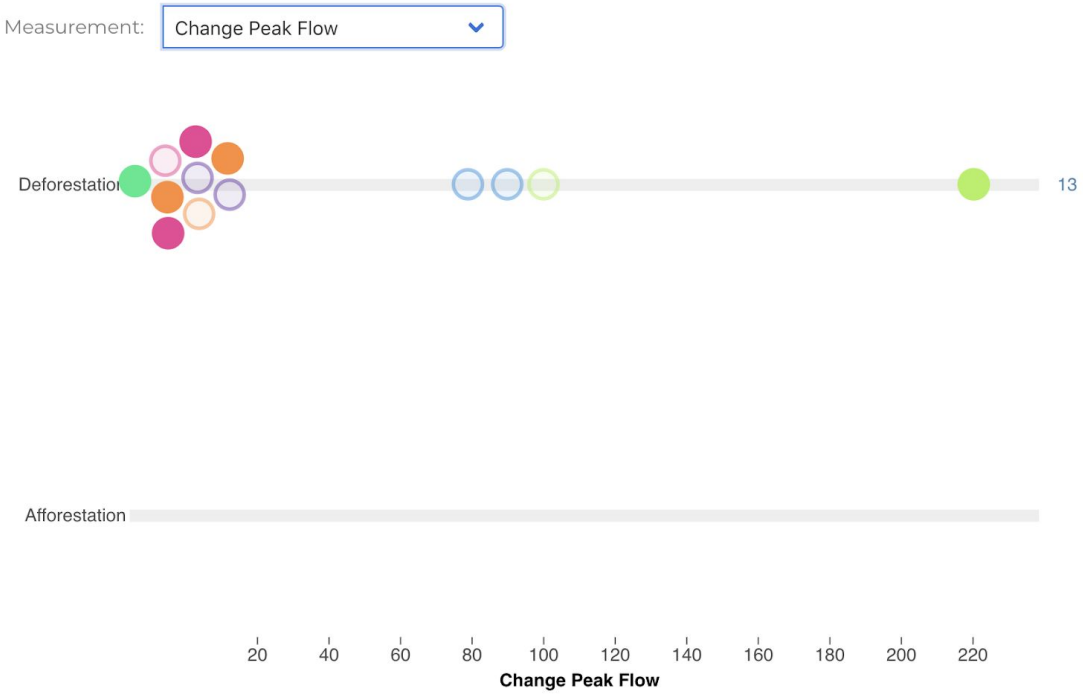
Interaction

Clearly, since our initial prototype was in Pandas, we did not have any interaction at that point. After implementing the visualizations in D3, we had to introduce interaction. The interaction was simple and most important it **satisfied the requirements** that we got from James.

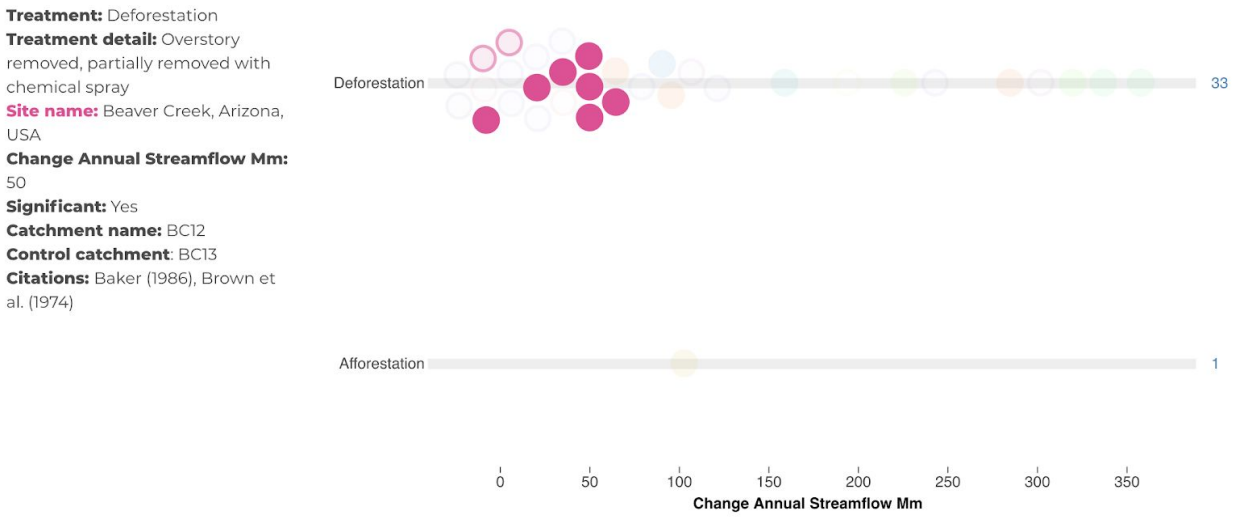
The first and probably most important interaction that the researchers needed was the ability to see results for different measurements. To this extent, we have added a dropdown select next to the visualization.



Each time the user selected another option, the visualization adopts accordingly.

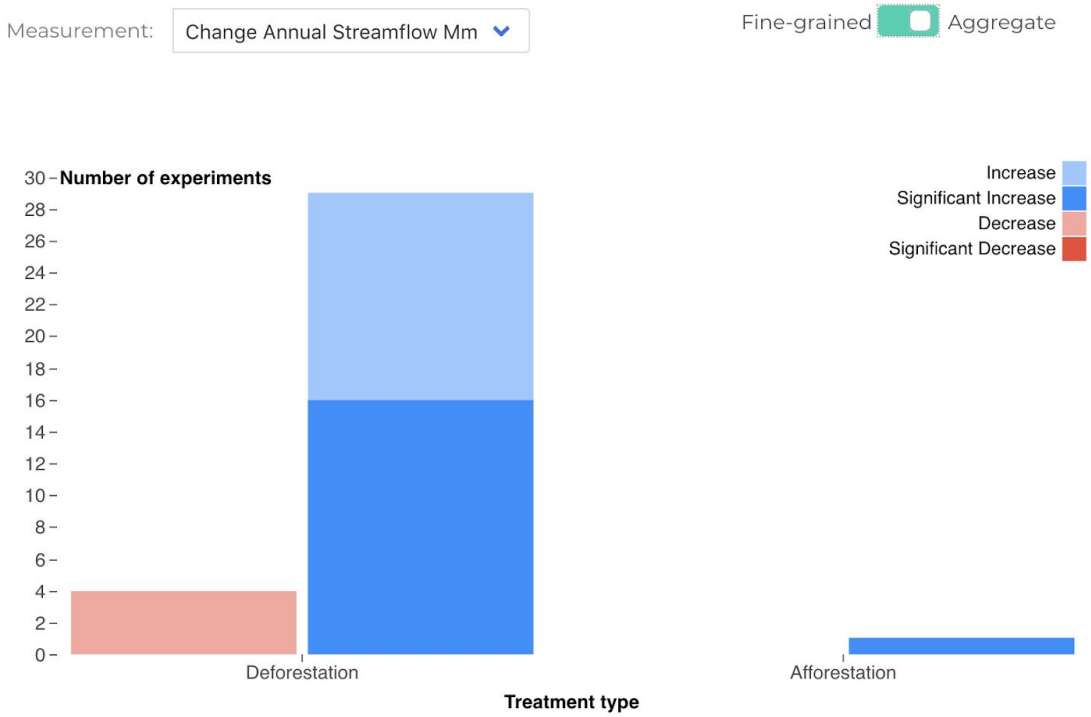


Another useful interaction, for the researchers, is the ability to hover over every experiment and see more details about each of them. Clearly, our current visualization covers only a subset of the data dimensions we have. With that said, we have introduced a detail box that shows up every time the users hovers a specific experiment. We have considered initially a tooltip, but that turned out to cover other experiments (circles), and that was not **desired**.



The hover feature also highlights the experiments performed on the same site. In the above case, **Beaver Creek, Arizona, USA**. Non-significant experiments are represented as hollow circles. This helps researchers really understand if the results are consistent or not.

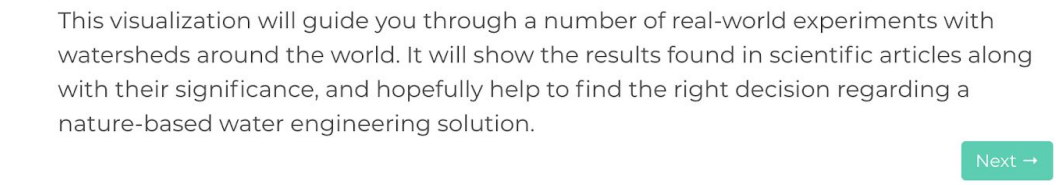
Another useful interaction is the ability to quickly switch from a fine-grained view (all the experiments) to an **aggregate** perspective of the data. This is done via a toggle.



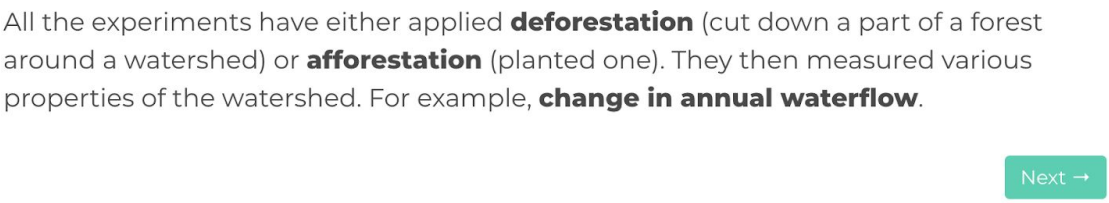
Story

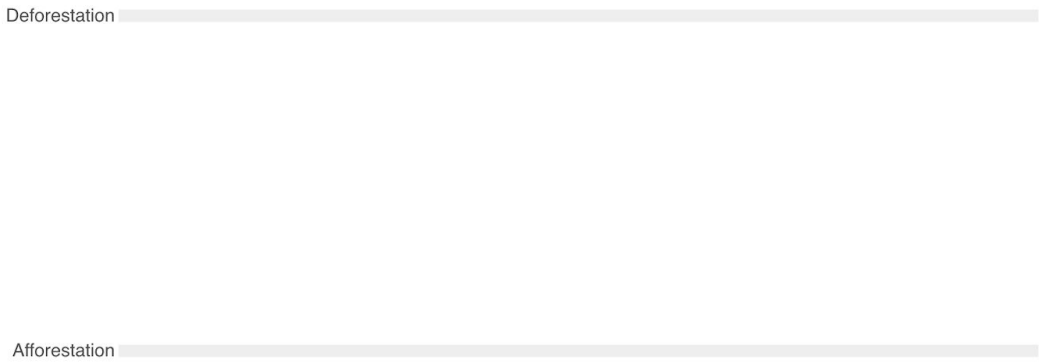
Our data is rather complex and it contains a lot of verbiages that is rather confusing. That's why our story was the **most challenging** part of the project. Every time we showed the visualization to someone that had no previous experience with this domain, they were getting confused. We came up with the solution to use the Medium-articles style, where we would have different copies of the same plot and at the end have interaction. Indeed, this might work for the audience not familiar with this domain, but it is **not practical** for our target audience - that is, researchers and stakeholders in charge of water distribution. That is why we have decided to tell our story by using a **guided tour approach**. Concretely, when the user lands on the visualization page, it is guided by our tour and he can explore the data while learning about how to use all of the available tools at the same time. For convenience, the left arrow keyboard is tied to clicking the **Nex**t button. At the end, the user is invited to explore the data himself, unlocking all the features.

In the first step, the user is informed about the tour.

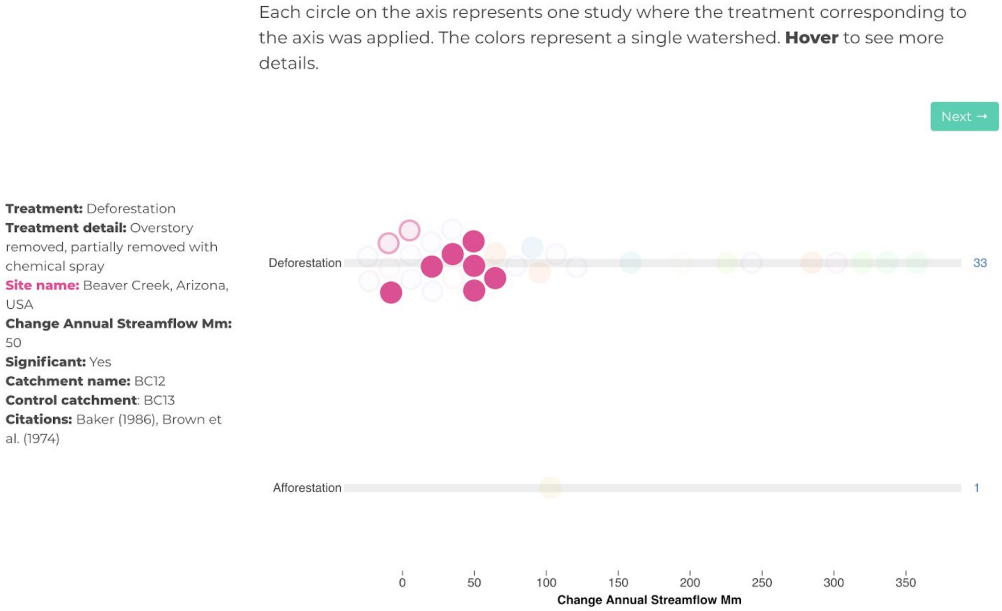


Then, he learns about the different **treatments** tied to each experiment.



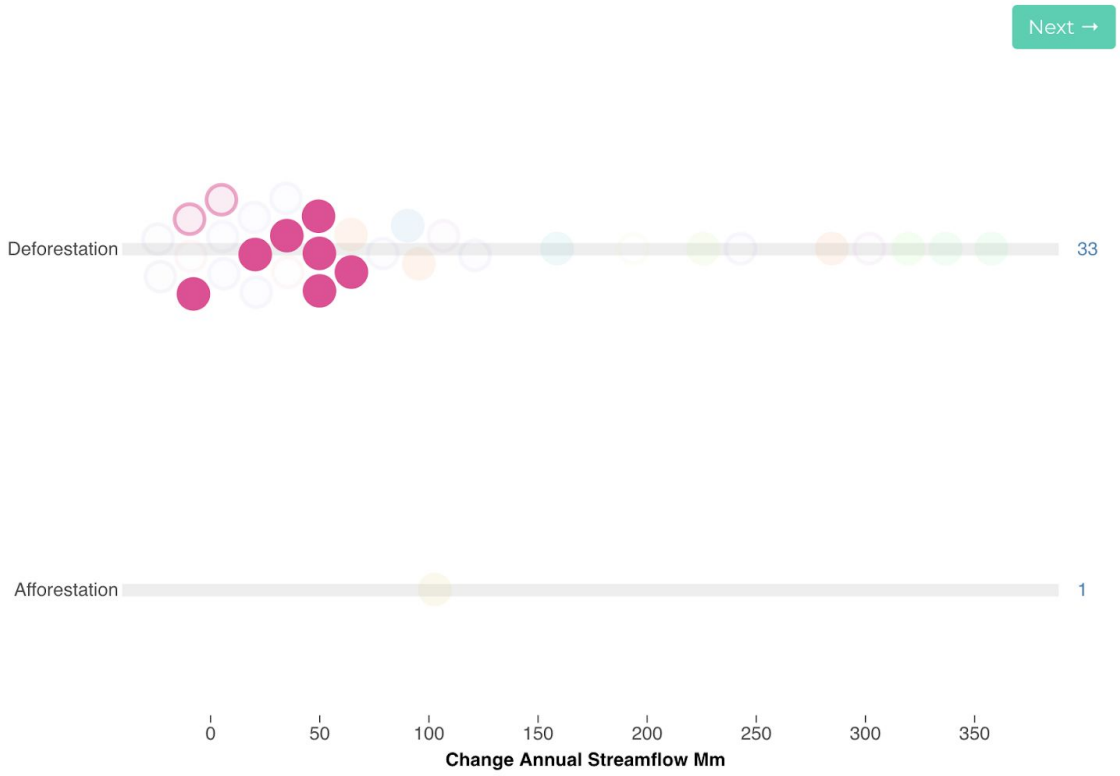


He is then invited to **hover** over the experiments to learn more about each of them.



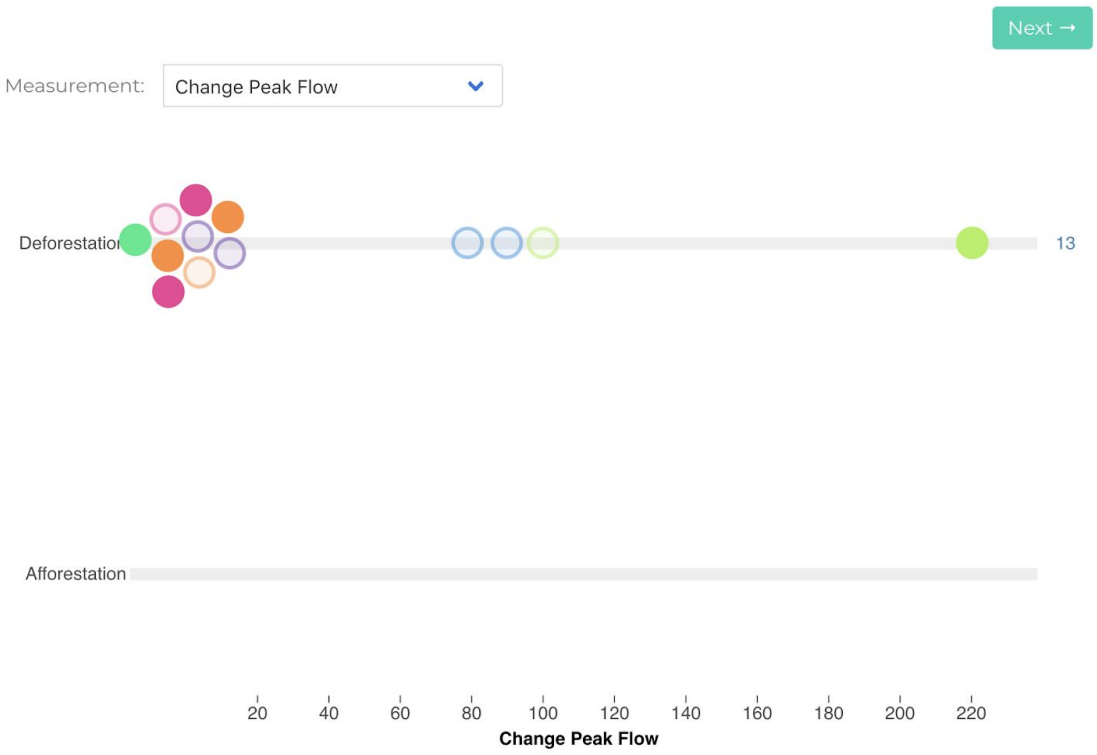
Next, the user understands now the difference between significant and non-significant experiments and how they are represented on the graph (encoded in the opacity of the circles).

Most of them discovered that deforestation leads to an increase in streamflow. Two of them, however, are **not significant** (hollow circles). It is important to see which experiments are not significant.



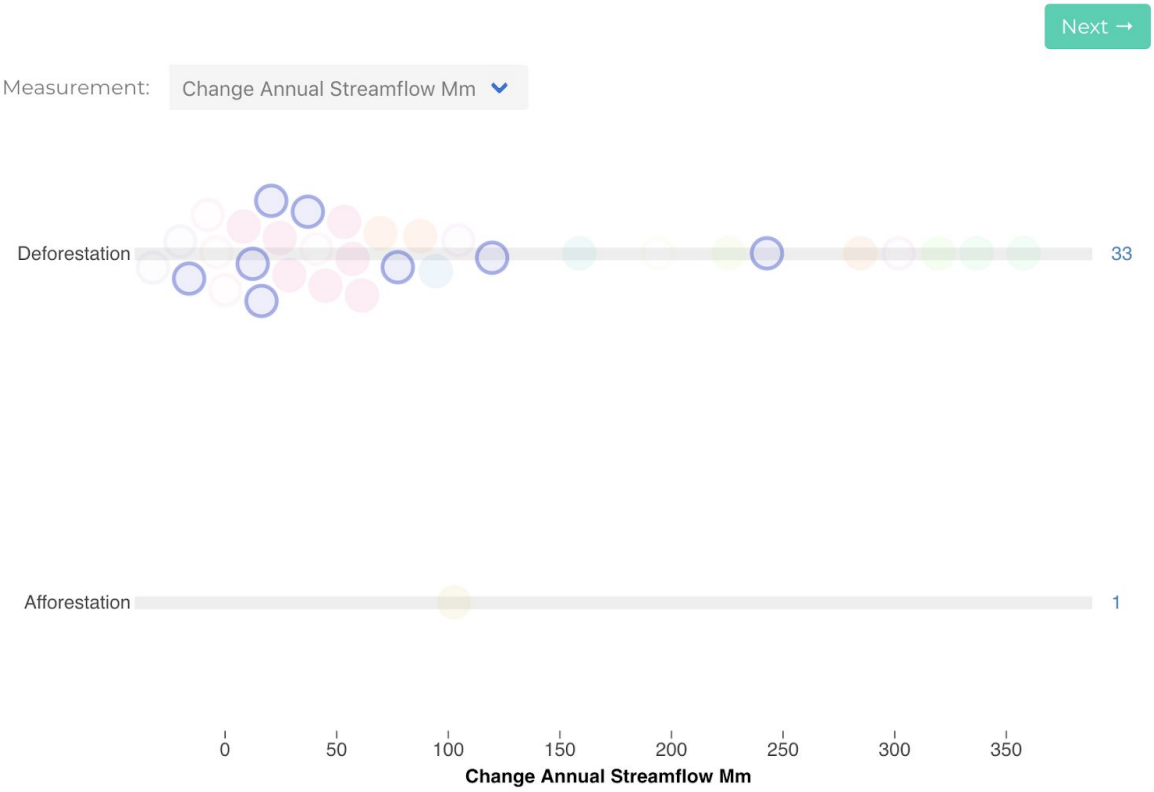
The user is invited to check the different kinds of measurements.

There are **other measurements** reported in the studies. This, for example, is **change in peak water flow**. It is indicative of floods. You can check different kinds of measurement now.



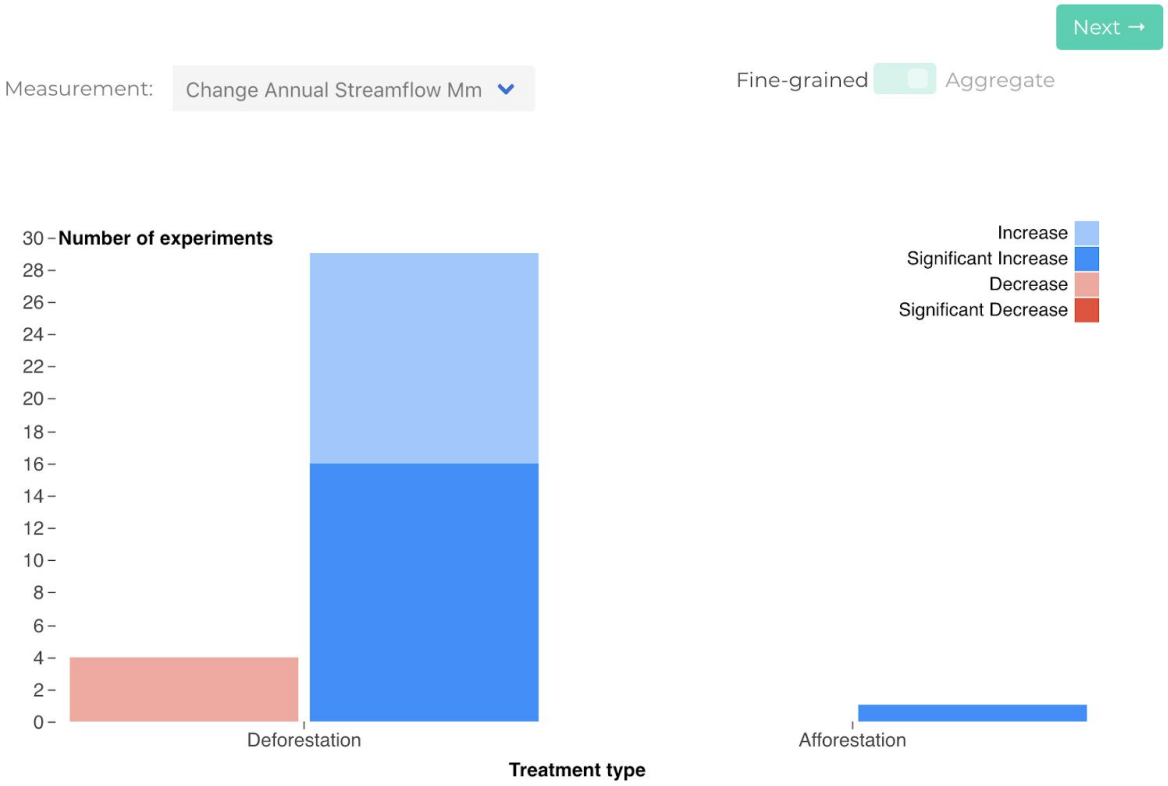
While in the tour, the user also sees different observations about the data. This is useful because he can **learn** what kind of **insights** he can get from the data.

All experiments in this watershed in **Southwest Australia** detected increase in annual streamflow, but all of them are also not significant. It is not clear without expert help if the results are conclusive.

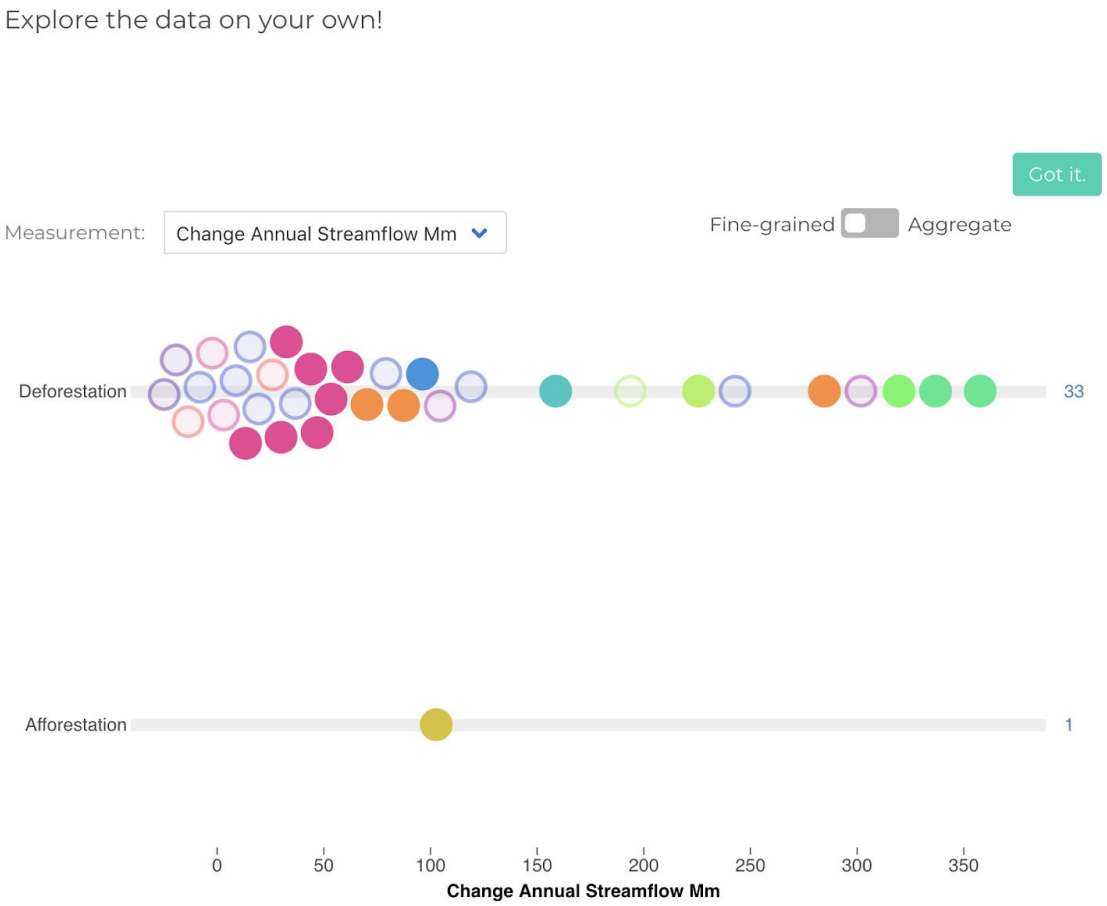


Next, the user is guided to the **aggregate** view of the data.

To check the big picture, you can take a look at the **aggregate plot**. It does not show particular studies, but only the general trend — whether the treatment increases the measurement or not.



Finally, the user is invited to explore the data on its own, with the knowledge just acquired.



Implementation

Technical Details

We started off by spinning a repository and a static website on **Netlify** that would update every time we pushed to the master (continuous deployment). Netlify is a static website hosting platform that has more features than GitHub Pages. For example, it has a limit of 3 builds per minute, compared to GiHub's 10 builds per hour, it has 1 click rollbacks and most important for us, asset optimizations. Here is a list of all the differences.

We have used jQuery for DOM manipulation because of its simplicity and easy integration with D3. We have taken into consideration React, but we decided to keep things simple, at least in the beginning.

As a CSS framework, we have used Bulma over Bootstrap mainly because we find it more beautiful and it is **Flexbox**-based. FontAwesome was used for quickly integrating common icons and because of its easy integration with Bulma. The fonts for the website were gathered from Google Fonts. We have used two Bulma extensions, namely Bulma-Timeline and Bulma-Switch.

For the visualization part, we have used Pandas and Seaborn for quickly prototyping and then D3.js for the final implementation.

Since our data was stored in an SQLite format, we have quickly come across an open source library for parsing and querying SQLite databases directly in the browser, SQL.js. We wanted to use such a library because we knew that our database will change as a result of acquiring more data, and so we wanted to be as flexible as possible. Compare this to the other solution where we would need to have a separate pipeline that converts the data from **SQLite** format to **JSON** or **CSV** for example. The disadvantages of this method are obvious.

Other libraries that we used are WickedCSS for CSS easy CSS animations and Skrollr for scroll animations and events based on the scroll of the user.

Overview and Functionality

The website is available at epfl-dataviz.netlify.com and the source code can be seen at github.com/rusucosmin/dv. The main contains information about the dataset, a timeline of our journey into developing this project and the team members.

The most important page, the visualization page, launched the story and it lets user explore the data in a fun and exciting way. We believe our story is interesting because not only it lets the user explore the data himself, but he learned something new from another domain.

Evaluation

In this section we will discuss the results we have obtained, the data insights we have achieved, pitfalls to look for as well as the prospect for future improvements and work.

Data insights

Clearly, more studies need to be conducted with other kinds of treatments, and more data is needed to draw strong conclusions.

From the data, we already have, in most cases, deforestation results in significantly increased flows, but not always.

If we take a look at the Beaver Creek watershed in Arizona, USA, there have been several experiments applying deforestation to this watershed. Most of them discovered that deforestation leads to an increase in streamflow. A few of them, however, are not significant. For this watershed, one can be rather sure that deforestation in similar conditions will lead to an increase in annual water flows. Not all watersheds boast the level of certainty like Beaver Creek does.

All experiments in a watershed from Southwest Australia detected an increase in annual streamflow, but all of them are also not significant. It is not clear without expert help if the results are conclusive.

Work distribution

We were doing weekly meetings and all sharing ideas and exploring possibilities . Most of the Ideas came during these brainstorming sessions and everyone gave contributions that were implemented in our final design. Jean Marc worked on the pre-processing and analysis of the data on the structure of the visualization. Bogdan worked on prototyping and partly on the visualization. Cosmin worked on the development. This repartition let join forces on work in parallel to maximize efficiency.

Future work

As a future work, we think it would be interesting to explore the visualization of different dimensions of the data, as a result of the requirements being changes. If more than was available, more insightful visualizations can be drawn. This experiments are very time-consuming and take a lot of time to do, hence the necessity to explore the results is obvious.