# MCS Portfolio: Interactive, Intelligent and Immersive Visualizations CSE 578

Joshua O'Callaghan
jdocalla@asu.edu
CIDSE, Arizona State University
Tempe, Arizona

## ABSTRACT

This course project aims to develop an interactive, intelligent, and immersive visualization for SEINEt data portal. SEINet is a collection of distributed data resources of interest to the environmental research community in Arizona and New Mexico. SEINet database is a large collection of raw data, which makes it hard to find the required information for the researchers. The project involves formulating a research problem on the data and creating a visualization to solve it. The images of certain species found in the Americas along with their elevation information is used to formulate the problem. The solution to the problem is presented in a visualization format.

## 1 INTRODUCTION

In this course, I learned about data visualization, why it is important to take specific design decisions and how to implement these design decisions. For our project we chose the SEINet Database to model because of the amount of raw data and potential we saw in the visualization we could create. With the help of Dr. Sharon Hsiao along the way, my group and I were able to create something that stood out and accomplished a goal that would not have been achievable before I had taken this course.

### 1.1 SEINet Database

To understand the scope and goal of this project, I will first explain the database where all the data values come from and the existing tools that are available to users.

The SEINet database is essentially a database filled with different types of flora with accompanying images, and descriptive text. The website can be found at the comes from the domain http://swbiodiversity.org/seinet/ where trees, flowers, cacti and other flourishing plants can be found across the vast website. The SEINet domain is one of the few interfaces which provides a distributed data resources for the environmental scientists on flora found in Arizona, New Mexico and Colorado. The data provided by SEINet includes native images on flora along with their scientific name, height, unique features, flowering season, elevation and location of the species. The portal also provides basic information like colour, nativity, length of various features of the species. The data that will be used and discussed in this paper is the SEINet cacti data.

Although the dataset that we work with deals solely with plants and flora of the Southwest region of the United States, the SEINet network stretches across many regions through other domains. The network has data on species from the Midwest, Southern Rocky Mountains, Intermountain, Mid-Atlantic, the Great Plains, Northern Mexico and Southeastern states. The data is collected from various herbariums, universities, colleges and various museums.

The portal also provides basic tools so that researches can access data that they want quickly. Collections Search, Map Search, and Exsiccati Search are the tools provided to users in the SEINet network and offer simple functionality to quickly lookup an item. The Collections Search allows a user to specify the collection of data, and offers more filters such as Family/Genus/Species. Map Search is a visualization tool which utilizes the Collections Search functionality and displays the output as triangles on a mercator projection of the world where the data items are found.The portal also provides Exsiccati Search (dried specimen of fungi) which is more of a narrowed down indexing tool rather than a search.

Although these tools attempt to solve the problem of navigating a large and complex site, there exist many problems with the tools. The Collections Search is clunky, not very intuitive to use, and slightly difficult to explore new plants/data entries. The Map Search takes a long time to execute and with the existing problems of Collections Search this tool is not very effective. Exsiccati Search is more of a filter rather than a tool itself which lies the problem. Already we can see plenty of room for improvement within the SEINet Database

### 1.2 Background

For this project, concepts of Visualization, Image Processing, and Machine Learning techniques were used.

*1.2.1 Data Visualization.* Data Visualization is the field of data science which deals with graphical representation of data and is the main focus of this course. It involves producing images that communicate relationships among the represented data to viewers of the images. Because humans learn more from visual insights than raw data Data Visualization is a very vital sector of Big Data. With proper visualizations we can convey an exact and detailed story to a user using much less time than it would to explain the data in its bare form.

*1.2.2 Image Processing and Machine Learning.* Digital image processing is the use of a digital computer to process digital images through an algorithm. It is a method to perform some operations on an image, in order to get an enhanced image or to extract some useful information from it. Many different techniques are offered in the world of Machine Learning and we will specifically be looking

at KMeans later in this paper to make conclusions from a given image. Image processing was an important part of this project since the data given to us was primarily images. The machine learning aspects of image processing and post image processing analysis was important as well. Not only were we using a KMeans algorithm to draw conclusions from the image but we also used a simple Knn algortithm for our predictions. Machine learning is defined as *the study of computer algorithms that improve automatically through experience*. Typically machine learning models will build a mathematical model or train data to then make predictions on later. While it would be ill defined to call our project a Machine Learning endeavor because of the lack , it is clear that Machine Learning techniques and methods were used throughout the project to accomplish our goal. Image processing and Machine Learning are two powerful tools that go hand in hand, and when used can provide amazing results for a project.

## 1.3 Problem Statement

This project is important because of the value it can potentially offer to researchers as well as artists/enthusiasts. The data that we are dealing with in its raw form is not very useful, as someone will not be able to make many conclusions by simply looking at the images the site provides. In its current state it offers at best a search/filter tool to be able to find a specific species/genus which is not enough if researchers want to learn more from this data quickly or if enthusiasts are looking to do something with this data without having extensive background knowledge of the flora.

The problem is that SEINet lacks the proper visual tools to display data in a meaningful format. It also could provide a lot of value to groups that it may have not considered originally. This dataset contains a vast amount of raw, unprocessed data which yields huge potential for data scientists to create something meaningful from. The goal of this project is to create a visualization that can utilize the data in a creative and useful way while providing a meaningful visualization that researchers and enthusiasts can both gain value from.

# 2 SOLUTION

In order to tackle a project of this size we first had to start with prototyping, making calculated design decisions along the way. I think it is important to point out that when we dove into this project we did not have a clear goal in mind which made the project a bit more difficult and will be discussed later in the paper. As a goal began to solidify while prototyping and working with the data, it was important to stick to that goal and create something that achieved it.

## 2.1 Prototypes

One thing my undergrad taught me through my Human and Computer Interaction course is that prototyping is the most important step in creating any sort of product. Originally the goal of this prototype were to satisfy the project requirements which was to create an interactive, intelligent, and immersive visualization. With this in mind our team broke off separately to create prototypes to try and use our data and accomplish the assignment requirements.

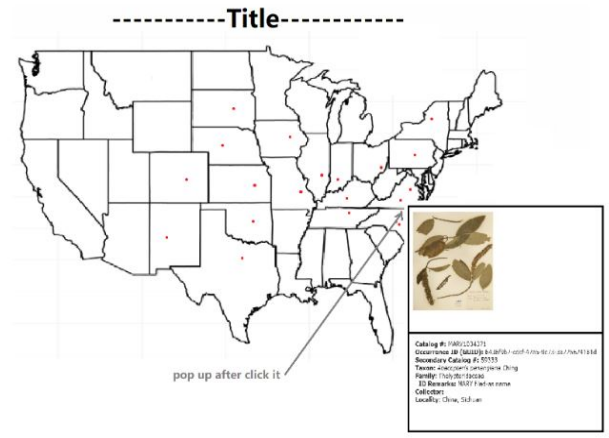


**Figure 1: Prototype 1**

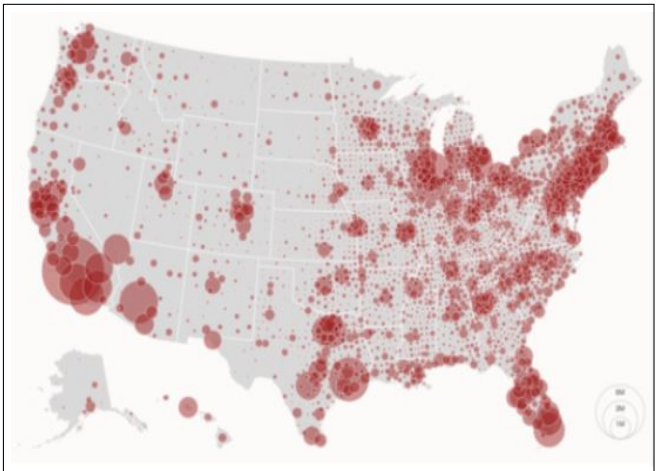

**Figure 2: Prototype 2**

*2.1.1 Prototype 1.* The first prototype (*Figure 1*) was a map graph overlaid by scatter plot type data. Each dot on the map represents a particular species found at that location. An interaction using the mouse pointer on the map would lead to a small description about that particular species. We ended up not moving forward with this design mostly because it lacked an intelligence aspect to it. It also was very similar to Map Search that SEINet provided and we wanted something more creative.

*2.1.2 Prototype 2.* The second prototype (*Figure 2*) used a map graph similar to the first prototype but instead it used a heatmap to visualize the data points.As a positive we could easily accommodate all the accumulated data in this design. The negatives of this design similarly to the first prototype was that we wanted to create something more original, and that it lacked an intelligence aspect to it.

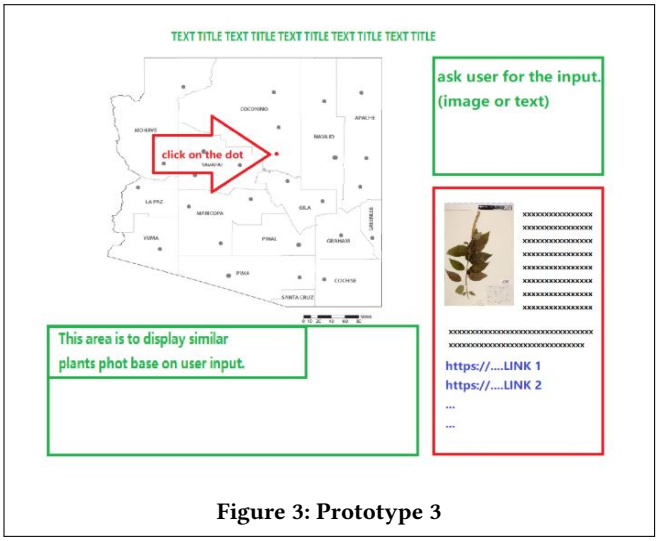


**Figure 3: Prototype 3**

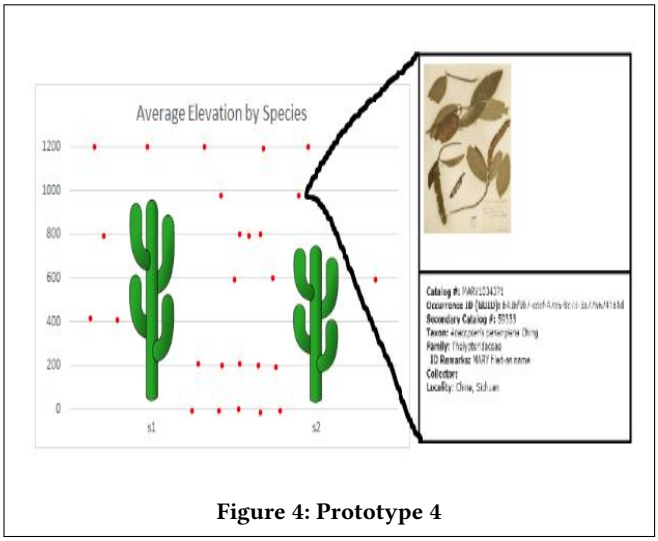


**Figure 4: Prototype 4**

*2.1.3 Prototype 3.* The third prototype (*Figure 3*) we liked more than the first two prototypes. We were able to implement some of the features we liked while improving on some of the features we didn't. We were able to keep a map graph while adding the uniqueness of using specific counties of an area we were all familiar with. This design also would ask the user for input in the form of an image or a text then returning a similar plant with appropriate information.

At this point in the prototyping phase we started to get an idea of what we wanted our goal to be with this project. As we moved closer to the problem statement defined above we realized we did not want to move forward with this design. Although this could be a valuable visualization for researchers, we wanted something that would be useful for the average cacti enthusiast as well, or even a home designer. Location data was not something we wanted to move forward with because of this.

*2.1.4 Prototype 4.* The final prototype we worked on is prototype 4 (*Figure 4*). Although this visualization lacked some of the interactivity and intelligence that the other visualizations had, we admired its simplicity. With a simple visualization we could really focus our time on a more advanced backend algorithm which is where most of our talents were.

Finally we were at the end of our prototyping phase. The design that we decided to move forward with was a modified version of prototype 4. We decided that we would add some of the input output features that prototype 3 used as well.

## 2.2 Design Decision

For our final product we decided that it was going to take images from user input and recommend the top 3 images in a subset of the SEINet database. The visualization would then compare the elevations of each of the recommended cacti as well as display their images. This design decision solved our problem because it could be useful to researchers who would want to make comparisons with other cacti and draw conclusions between color and elevation. It would also be useful to enthusiasts who wanted to create a color theme in their yard or add a small cactus to their color themed room. We were satisfied moving forward with this design because it solved our problem and was a new and creative way to visualize data that would otherwise not have significant purpose.

## 2.3 Backend Implementation

*2.3.1 Data Cleaning and Preparation.* The first portion of backend implementation was scraping the data from the SEINet portal. Because SEINet did not have an easy way to access the photos in their database, we built a script in Python to collect the images through nested loops of http requests. By using the HTML parser *Beautiful Soup* this process was painless and straight forward.

The next step in the process was to clean the data that we had collected. We decided that the most accurate recommendations would come from analyzing images in our dataset which means that we needed clear images to pull color codes from. Many of the images were photos from books so we cleaned the data manually to remove these.

*2.3.2 Data Analysis.* The next step was creating a supporting database will color data so that we could refer to it later in the matching process. Using Sklearn and Skimage, we were able to retrieve the top colors found in the images in our database. We performed this analysis on all images in our data subset then stored the data in a csv file. The file has the following format for each row: name of the cactus, filepath of the image, top 6 colors pulled from the image, and cactus elevation. The name and filepath were stored as a string, the elevation was stored as an int, and the colors were three RGB values stored in an nd-array.

The last step in the back end is a function which returns top 3 choices, based on the color. Similarly to how we built our small supporting database with colors, we used Sklearn and Skimage to evaluate color data user input. It uses the base value on RGB and calculates the distance between user input and the csv stored values using the Manhattan distance method. The distances are sorted in descending order and the names of three plants with smallest value

are returned. The top three predictions are returned to the front end and displayed.

## 2.4 Front-end/Visualization Implementation

The frontend was much less complex than the backend but still required some setup and most importantly integration with the backend. The index.html file contains all of the styling and functions that we used combining html, css, and js. For the visualization portions we use D3.js. For hosting, we created an EC2 instance on AWS running Ubuntu 18.0.4 and using Apache. In order to integrate this with the backend python code, we used the Django framework. When an image is chosen, it sends the data to the backend and calculates the necessary data using our python script. When the values are returned D3 displays the elevation of the top 3 choices on a bar chart and the input image is displayed along with a color wheel of the codes pulled from the input image. When you hover over the bars the name of the plant appears along with the image all handled with D3.

## 2.5 Technologies used

The below technologies are used to develop the visualization.

| | |
|---|---|
| **Backend:** | Python, Django, **Libraries:** Beautiful Soup, NumPy, SkLearn, SkImage |
| **Frontend:** | D3.js, HTML |
| **Hosting:** | AWS, EC2, Apache for Ubuntu |

# 3 RESULTS

Our team and I were very satisfied with the product result we accomplished. Not only did it complete the assignment, but we were able to find a problem within the dataset and solve it using a beautiful visualization (*Figure 5*). The advanced matching algorithm that we implemented allows us to see correlations of elevation. The solution we proposed is well implemented and it solves the proposed problem. There is plenty of opportunity for future work on the SEINet database. We hope that our visualization opens up some of the possibilities you can achieve with the data.

## 3.1 Reflections

I think that overall this project was a success. As mentioned previously I think that we made a mistake by starting to prototype our project without having a clear goal in mind. once we established the problem we wanted to solve, it made the work much easier and more meaningful as well. There were no known bugs with out visualization and it performed exactly as intended. If I were to do this again, the only thing I would change would be just better planning and overall group communication. Our communication improved by the end but in the beginning it made it hard to coordinate.

# 4 PERSONAL CONTRIBUTIONS

Personally, I collected and cleaned all the data for our project including writing the Python script to interface with the SEINet website. I performed all presentation items which included the Elevator Pitch, Presentation Video, and Submission Video. The videos I did alongside another group member.I also assisted in the prototyping, planning, and organizing of the project. I pitched in a little bit for the back-end implementation and front-end implementation. I also wrote the report with the assistance of two of my group members.

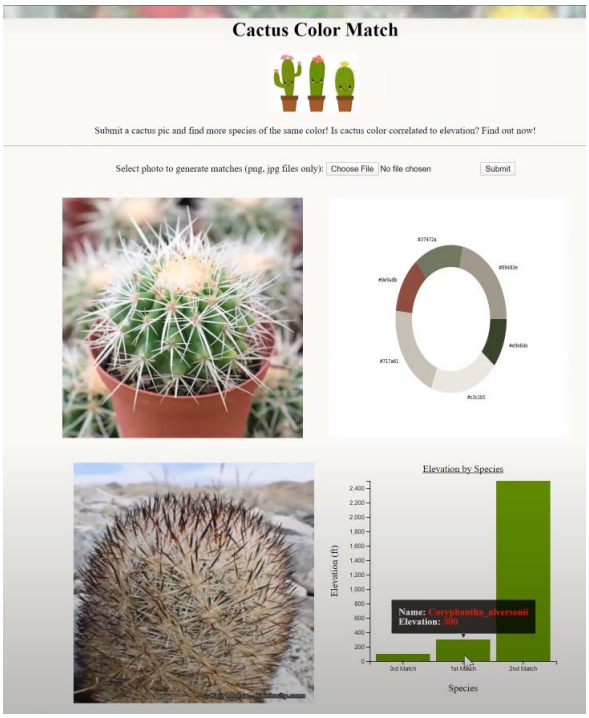


**Figure 5: Final Product Result**

# 5 TAKEAWAYS

I think the biggest takeaway I got from this group project was the value of having a motivated team. Everyone took up a specific role and performed it excellently. This project was also the first time I had written in ACM format, and the experience was well worth it because I will need it extensively for the rest of my graduate career.

## 5.1 Technical Skills

Some of the skills that I picked up were: D3, Django, Python scripting, and LaTeX. This class really helped instill the knowledge of D3.js in me and now I am able to create beautiful visualizations like the one we created for this project. Django was a new framework for me and it was very useful to learn how to integrate html, JavaScript, and Python function calls. I have been programming in Python for a few years now but I had never gotten to work with Beautiful Soup or scrape and data from a website. I am glad I got that experience because I know it will be useful in the future. I also had never done any machine learning programming in python and being introduced to libraries like numpy and sklearn is something that I value. Finally I have heard of LaTeX before and wanted to start using it to make better looking documents for my resumes. Now I finally have the experience to make that happen.

## 5.2 Team Members

My team members were: Alyssa Goldstein, Ke Fan, Haseeb Amin, and Ganesh Betha. I am thankful for such a committed group and without them this project would not have been possible.

## REFERENCES

[1] Seinet. 2020. Seinet- arizona, new mexico chapter. (2020). http://swbiodiversity.org/seinet/index.php.

[2] Karan Bhanot. 2019. Color identification in images-machine learning application. (2019). https://towardsdatascience.com/color-identification-in-images-machine-learning-application-b26e770c4c71.

[3] Cory Maklin. 2019. K nearest neighbor algorithm in python. (2019). https://towardsdatascience.com/k-nearest-neighbor-python-2fccc47d2a55.

[4] Jason Brownlee. 2020. Develop k-nearest neighbors in python from scratch. (2020). https://machinelearningmastery.com/tutorial-to-implement-k-nearest-neighbors-in-python-from-scratch/.

[5] 2019. Beautiful soup: build a web scraper with python. (2019). https://realpython.com/beautiful-soup-web-scraper-python/.

[6] [n. d.] Paletton, the color scheme designer. (). https://paletton.com/.

[7] 2019. Why image recognition is so important. (2019). https://www.netbase.com/blog/image-recognition-important/.

[8] [n. d.] Machine learning. (). https://en.wikipedia.org/wiki/Machine_learning.

[9] Adam Heitzman. 2019. Data visualization: what it is, why it's important amp; how to use it for seo. (2019). https://www.searchenginejournal.com/what-is-data-visualization-why-important-seo/288127/.

[10] Sharon Hsiao. [n. d.] Data visualization. (). http://www.public.asu.edu/~ihsiao1/slides/cse591/L1-20S.html.

[11] Sharon Hsiao. [n. d.] Principles and design. (). http://www.public.asu.edu/~ihsiao1/slides/cse591/L4-PrincipleDesign20S.html.

[12] Sharon Hsiao. [n. d.] Machine learning and visualization. (). http://www.public.asu.edu/~ihsiao1/slides/cse591/ml20S.html.

Heitzman [9]Hsiao [10]Hsiao [11]Hsiao [12]