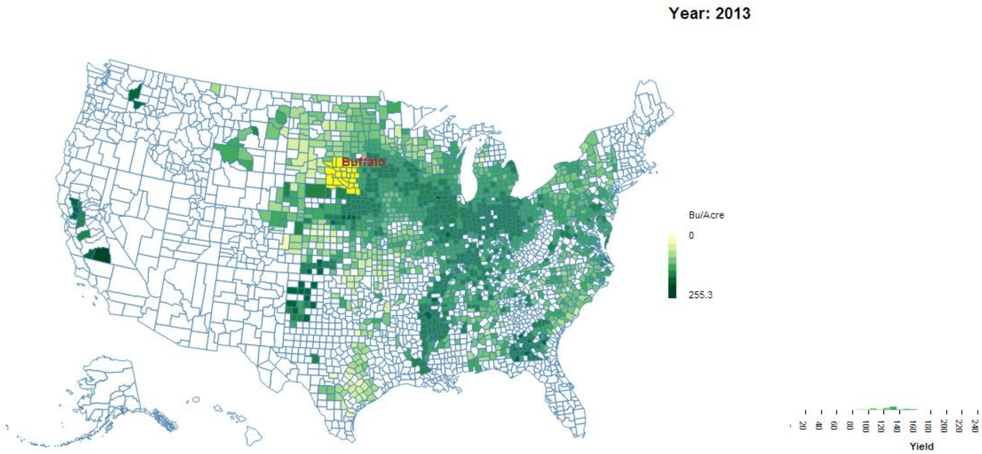
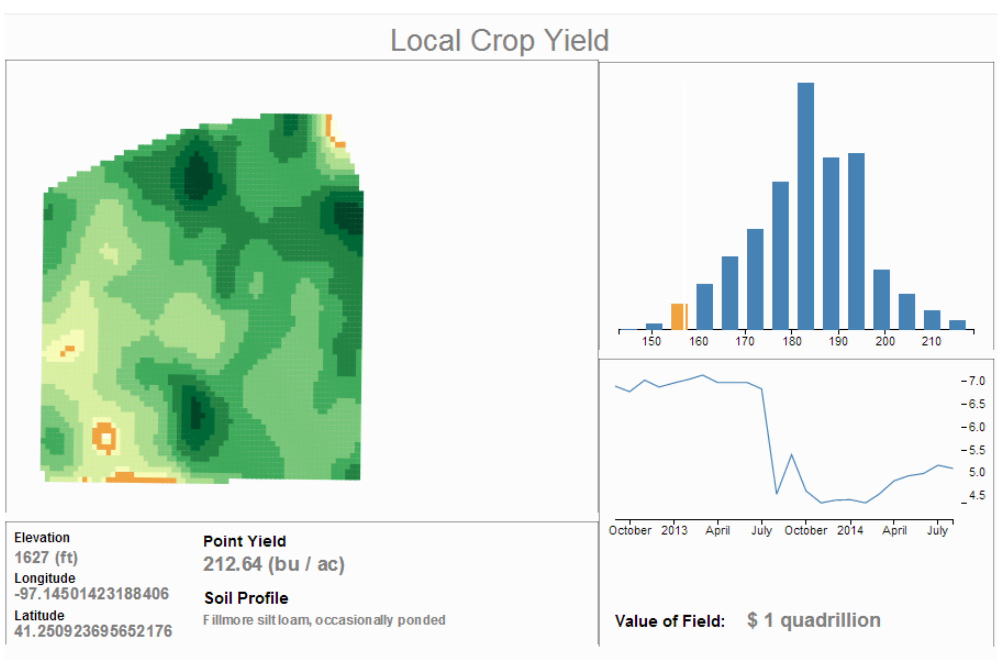
BitFarm Process Notebook

J. Benjamin Cook, Ryan King, Chad Hornbaker, Conor Myhrvold

For our milestone, we outline what we have accomplished and what we aim to do for the remainder of the project.



BitFarm Visualization Prototypes



Index

Click on the page to go to section

Overview and Motivation

Unprecedented amounts of data are available on modern farms. From yield monitors to weather sensors to infrared imaging, farmers try to keep track of every detail. However, most can't take full advantage of this data. Much is never reviewed after being collected, and what is, often remains inaccessible, trapped in complicated legacy software, agronomist reports and countless spreadsheet pages.

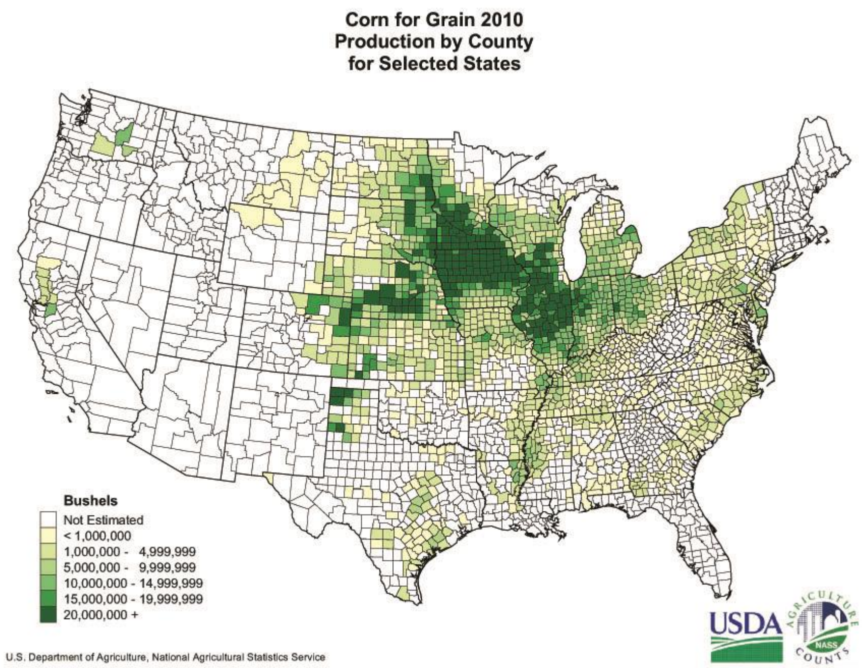
As a result, agricultural production lags far behind other industries in data-driven decisions. But increasingly this type of decision making is farming's best hope. Farming is a challenging business. Reduced water availability, and a general rise in operating costs, narrows the difference between success, and losing everything. Extreme volatility in the commodities markets contributes to uncertain cash for cash crops. New irrigation regulations, pesticide use, and fertilizer contribute to the problem: farmers must find new ways to increase yield.

Our initial project concept was for a d3.js data visualization tool for farmers and other like-minded individuals that can serve as a one-stop shop, allowing us to see a farm not through only the eyes of a 20th century agriculturalist, but also as a 21st century data scientist: the Bloomberg Terminal for the farm, merging information from the farm itself with country, state, and national trends.

As the project has taken shape, we realized there is a bigger story. With access to over a century of crop yield data, we are able to visualize the changing face of farming in America. Technology and climate both have a significant effect on the efficiency of farming. Our visualization allows a user to explore the trends in crop yield over time and also view yield data from selected individual farms to understand the macro is connected to the micro. This is a basic implementation of a concept that could easily be developed into a more robust tool in the future to help farmers make the most of their data.

Related Work

For all of the talk about the changing face of agriculture and US land use, there have been few attempts to show a concomitant picture of what this change looked like in the 20th century, and now into the 21st.



Example national corn yield by country visualization. Our county level data comes from the same source, thus our model holds a similar appearance to the figure above. We extracted the data from 1910 to 2013 (most current), and made it selectively interactive. We also chose a similar color scheme since the greens and yellows give the appearance of growth and vegetation.

We dug deeper and gathered many disparate, but related datasets. Our displays of corn production, use two main visualizations to show the big picture, the local picture, and everything in between (through interactive selection).

Questions

We are motivated by the general question, what affects a farm over space and time in the United States?

More specifically, we aim to answer, how do different farm influences fit together to show the big picture?

As we made our visualizations, we were able to specify our questions. Our visualization now directly answers:

1) What does the corn crop yield change look like over the past century in the US?
2) What are key indicators affecting the yield of a local farm?
3) How does the above compare to the context of the county, or larger area (up to national)?

Data

The data from a variety of sources. Here we list the datasets used in our visualizations, where they can be obtained, how they are presently used or how we envision using them in our final product:

1 – National Weather Data, United States Historical Climatology Network

Daily precipitation, and min/max temperatures affect yield. We show the relationship between climate (aggregated weather trends) and yield in the County Level visualization. After pulling down weather data from the last century we cleaned and transformed the data with a Map/Reduce job to obtain a manageable clean sample that we can visualize.

2 – County Yield Data, USDA National Agricultural Statistics Service (NASS)

Shows yield by county from 1910-2013. Together, it's a national picture. NASS maintains relatively updated and clean survey data from farmers around the country. To use this data we simply downloaded CSV files and removed irrelevant columns with a simple Python script.

3 – Local Yield Data came from correspondence with a Nebraska farmer.

From combines (geo-located data, sampled 6 times/sec). Proprietary to others, but we can use publically. An example of a farmer's individual yield. The county and national data place this small plot into context. Local yield data is summarized by fitting a Gaussian Process to the data so we can show the spatial variability in local yield.

4 – Soil Type Data, UC Davis Soil Resource Laboratory

Affects yield and type of crop suitable to be grown (although we focus on corn). Soil data is downloaded thorugh the UC Davis soil lab API, which allows us to access a high dimensional overview of soil by longitude and latitude.

5 – Grain Price Data, USDA Economic Research Service (ERS)

Affects the crop value; how much a farmer can get from their land. We summarize this with monthly averages of price in dollars per bushel. We use this data to show how valuable a farmer's field is at any given time.

6 – Soil Moisture Data, USDA National Water and Climate Center (NWCC)

From 1978 – 2013. Soil moisture affects yield. Scraped with Selenium. This data will be incorporated in our final visualization with the climate time series data on our county yield data page.

IPython notebooks used to clean the data:

Weather Data IPython Notebook

County Data IPython Notebook

Exploratory Data Analysis:

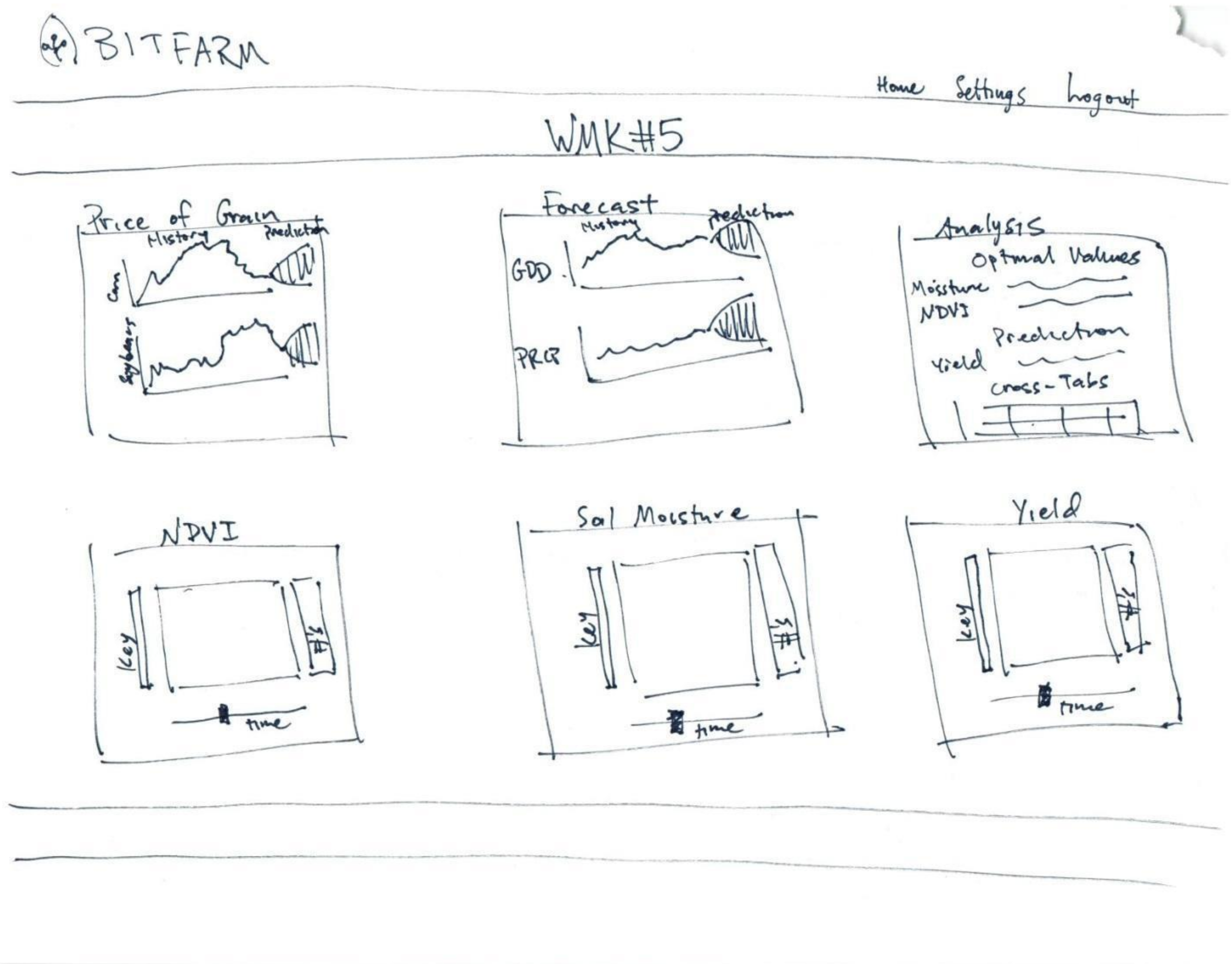
We started with separate visualizations of local yield, county-level yield and climate data to get a feel for what the data look like so we could begin to think about how the different types of data depend on one another.

We learned several important things from these exploratory visualizations that are informing how we build our final visualization. First, there is a major technology trend, and yield has dramatically increased over time overall, with variance for any individual area (i.e., some has decreased, and what's remained productive, has increased a lot). There is a high degree of spatial variation in corn yield. Yield tends to be highest in the Midwest United States; however, corn is throughout the US.

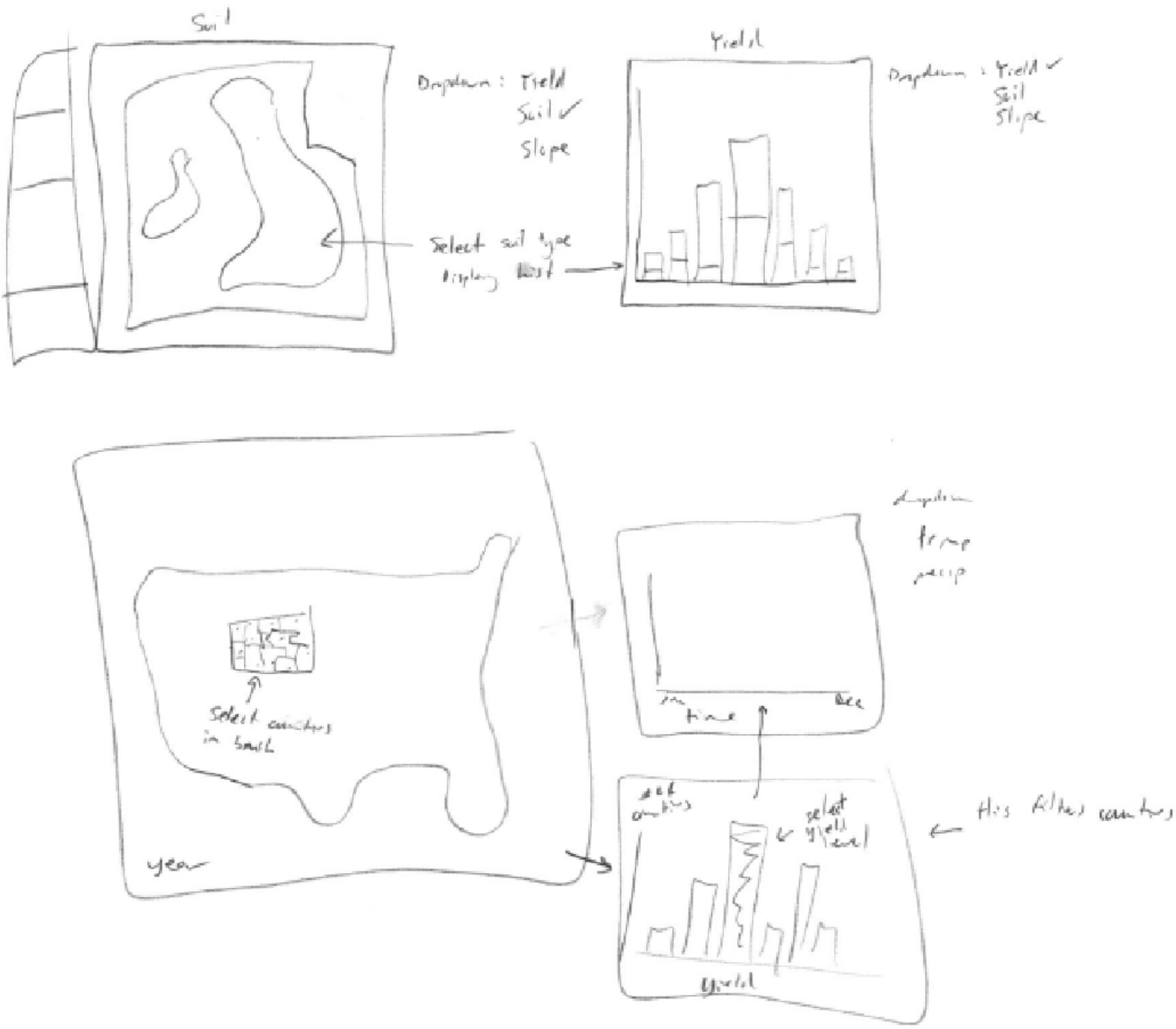
We also did some basic statistics to aggregate and summarize the data. For example, we trained a Gaussian Process on the local-level yield to show the patterns of high yield in a spatially smooth way. Because the climate data are several gigs in raw form, we also aggregated temperature by plotting average monthly values.

Design Evolution

Our initial design concept was for a dashboard with several floatable visualizations, each showing a metric affecting farming success, the crop yield. The idea was that the layout could be customized; each visualization would be expandable, draggable and snapped into place, much like an app interface on a Windows Phone or an app dashboard for a PC.

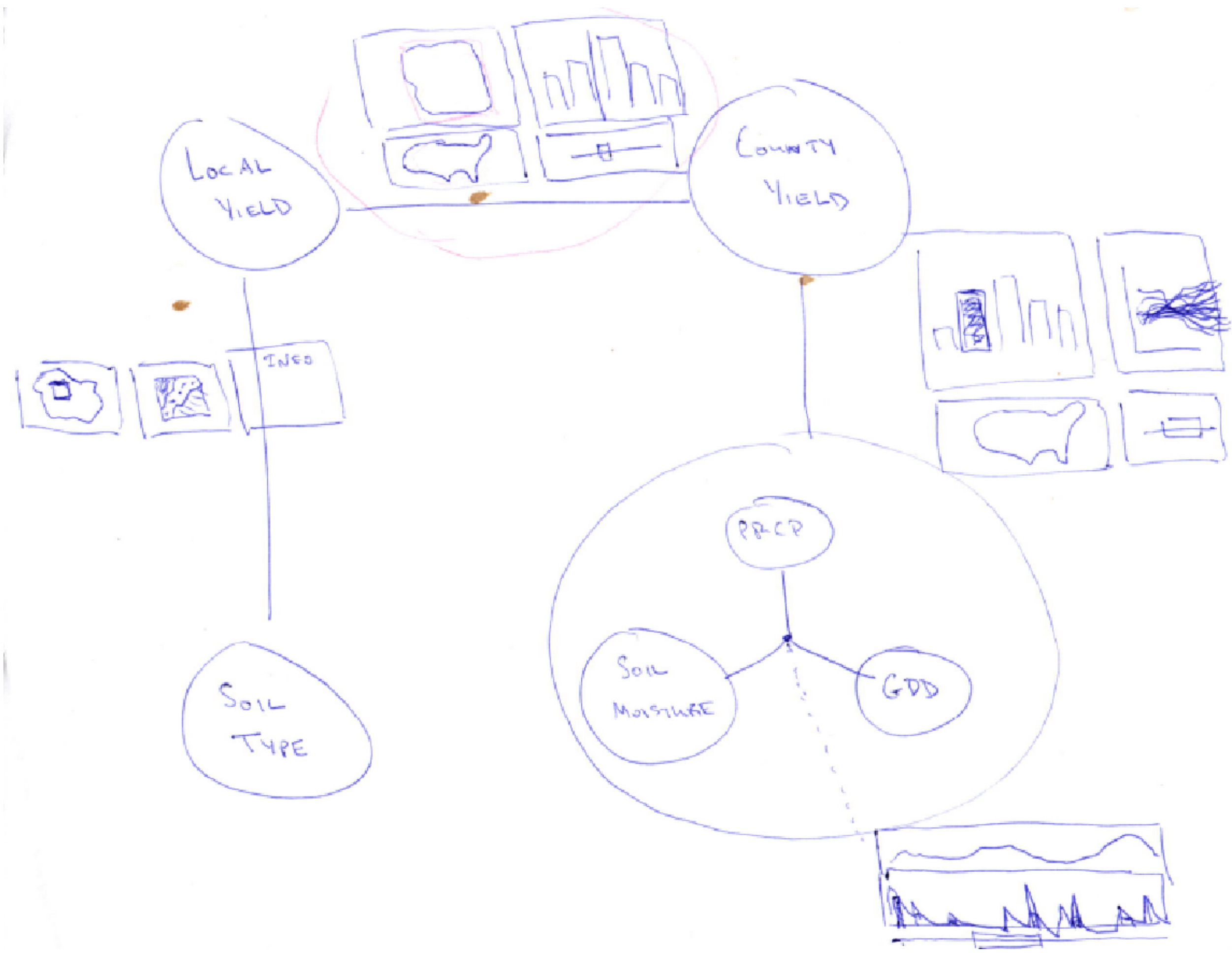


As we began to better understand the data and explore possible implementations, we gradually began to move away from the dashboard concept and focus on the visual components that communicated the best story for our data. The dashboard concept may be useful as a future implementation in an app, but we did not see it as the best way to communicate our intended story and allow a user to intuitively discover and explore the data.

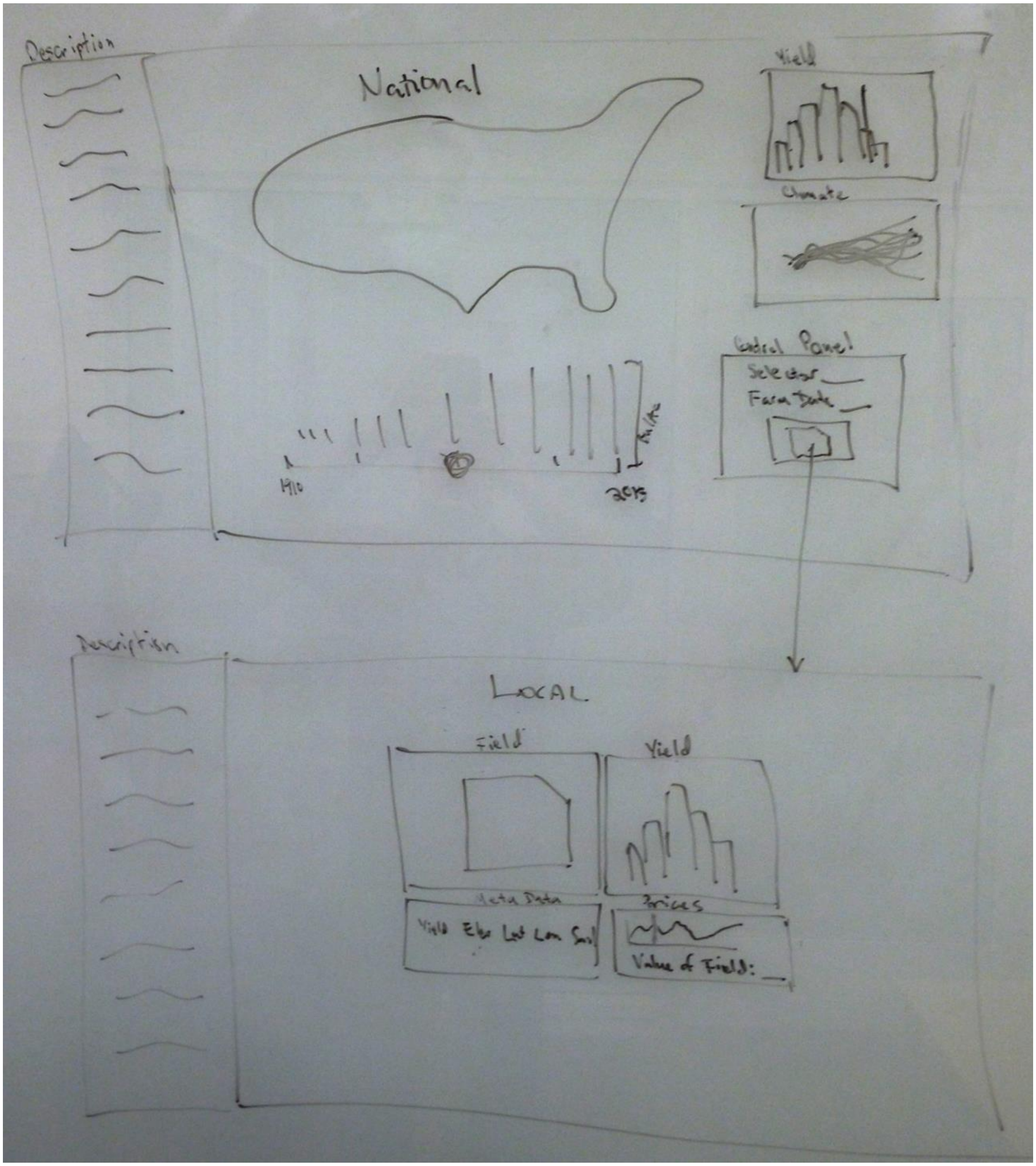


The sketch above shows our initial concept for the simplified interface. Rather than multiple disparate views, we focus on a single view at the national level *(bottom)*, and a view of a single farm *(top)*. Nearly all of our available data can be displayed in these two simple views. A user can look at the yield data at specific points in a field and see where the field is producing the most, as well as compare the yield to the current and future market price. At the national level, the user can view other counties around the country with similar output, as well as seeing the historical trends in yield in the context of weather and location.

Linking the data between the micro and macro level was a difficult conceptual task. The sketch below depicts the information flow we believed would work well for addressing our research questions. At the local level, we can see which parts of the field have a certain soil type. We thought this may be an interesting link to the point yield data. At the field level, we want to not only see the yield patterns, but have a way to compare it to how well other counties across the country are doing. This broadens our view to the county and national level, where we can see trends in weather and other factors such as soil moisture. The addition of historical data provided a more compelling view of how well modern American farms are doing in comparison to only a few decades ago.



<u>Current Design</u>



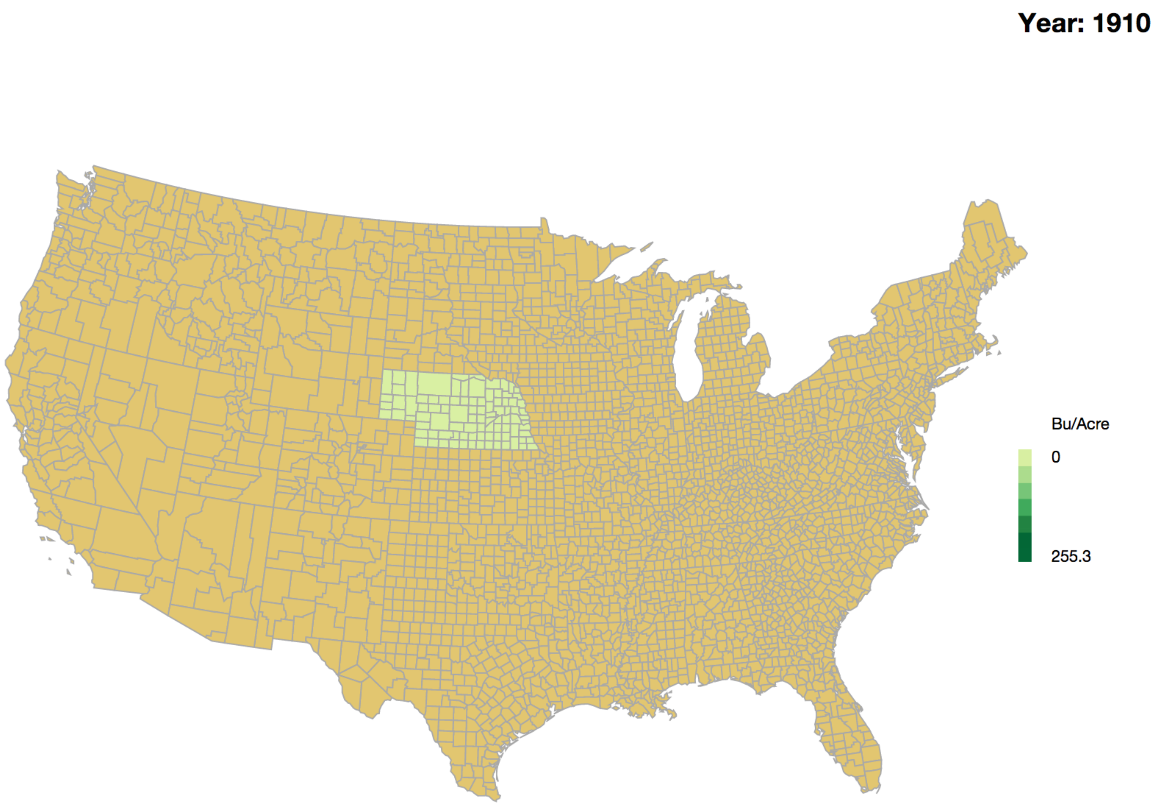
We narrowed down our ideas further to detail the specific information we want to see in the two primary views. The National view (above) shows the historical trends and has the option to view any of the individual farms in the dataset (still very limited at this time). The Local view implements our desire for a way to explore a farm's data at a very specific level. The transition and cross-information between these views will be further refined in our final product.

Implementation

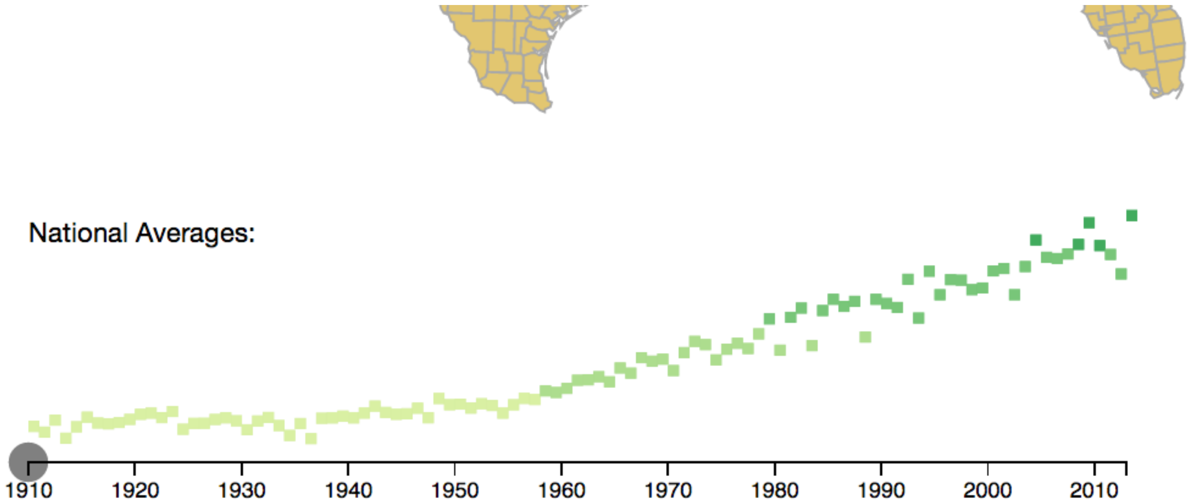
The layout consists of a primary visualization of national crop yield of corn for the past 100 years and a secondary visualization of field level data.

National Yield View

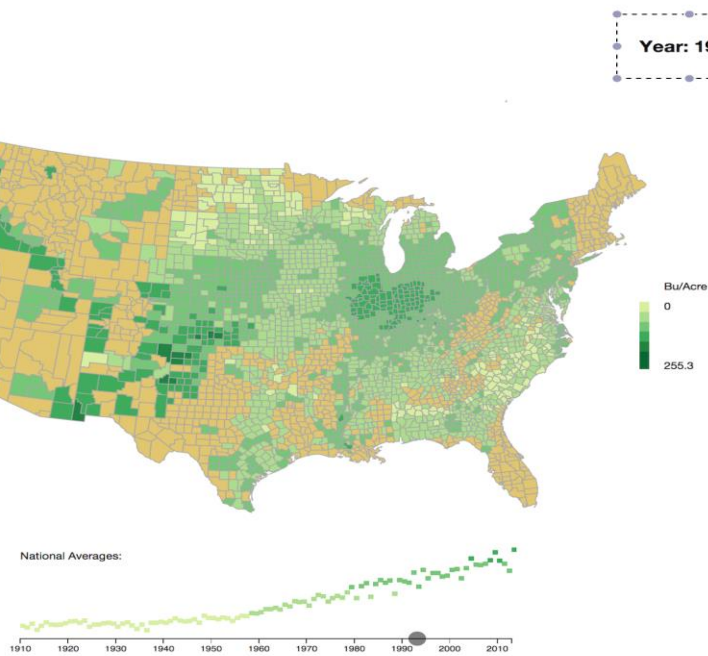
The National Yield View starts by loading the data of corn yield for the earliest year in the dataset (currently 1910). A choropleth is then generated using the county yield data. For comparative power, the color range of the cholorpeth is scaled to the min and max yield for the entire data set.



As indicated by the map's legend, darker green shades correspond to higher yields. Green was chosen because of the intuitive association with plant growth. A light brown color is used to shade in any county without yield data as an analogy to an unfarmed field's soil color. In addition to the choropleth, a scatter plot is generated with the average yield of all counties per year:



This axis of the scatter plot also serves as range slider that allows for interactive selection of which year's data is displayed on the choropleth. For aesthetic reasons, the axis's ticks were placed at ten year intervals. To keep track of what year has been selected, there is a tooltip in the top right corner of the graphic that logs the currently selected year:



In addition to the aforementioned aspects of the visualization, there are several additional features that are designed to enhance the user's ability to glean trends about crop yields:

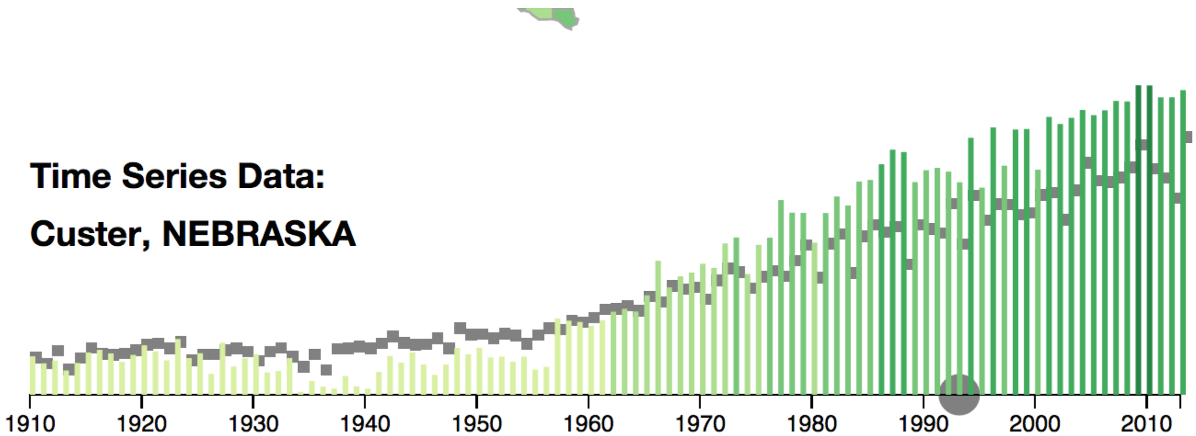
Toggle Selection Mode

Radial Buttons allow for seamless transition between the graphic's two main selection methods:

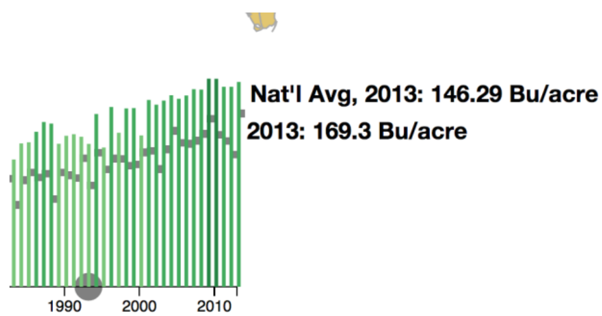
1.) Point Selection:

When the user mouse hovers over any county, the name is displayed.

Additionally, the user can click on any county to display historic yield data of that county. This county data is plotted against the national average data for comparative purposes. To simplify the graphic, the national averages are greyed:

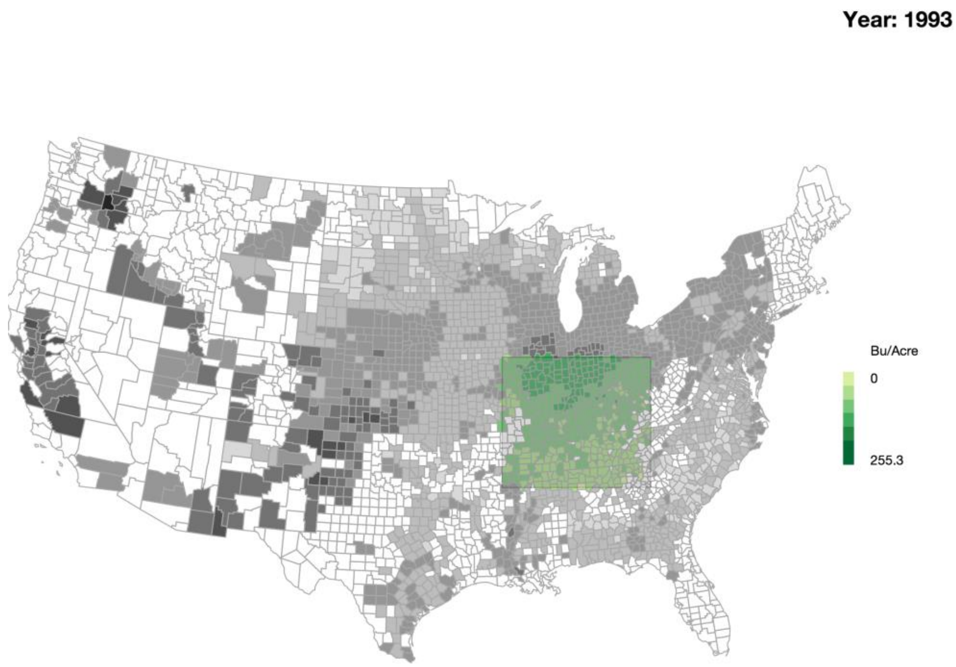


To see the actual value at any year, the user can simply hover over the appropriate bar. Additionally, the user can hover over national average scatter plot value to generate a tooltip that will display the numerical average. This feature allows the user to quantify exactly how well any county is doing compared to the national average:



2.) Two Dimensional Brushing:

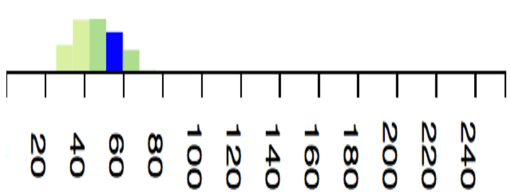
This second selection method allows for the user to choose a group of counties to focus on. When selected, the brushed counties maintain their original color scheme while the rest of map is gray scaled:

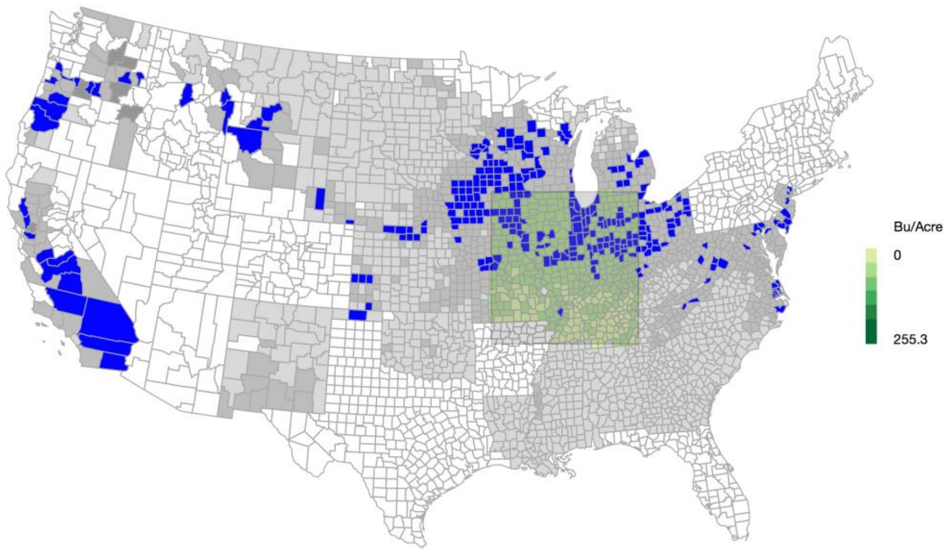


When a region is selected, the data displayed is truncated to only include the select counties. This allows a user to obtain better insight to regional trends as opposed to national averages.

Side Panel Histogram

Yields for the country (or select counties) are used to create a histogram of the distribution. A user can click on the histogram bars to select all counties with that yield range:





The idea of this graphic is identify regions which had similar yields to hopefully discover trends in what generated those similar yields.

Ongoing Programming

1.) Weather Data Visualization:

Now that the above features have been implemented, the next major feat to accomplish on the National Yield View will be to incorporate weather data. Currently, there is a placeholder panel for where the weather data visualization will go. This visualization will include year round data for the brush selected region in the form of a line chart. There will be a check box feature that allows the user to select which data they are most interested in displaying (temperature, precipitation etc.). Additionally, when the user clicks on the aforementioned histogram's bars and selects all counties with a particular yield, the weather data for all highlighted counties will be added to the line plot. The major goal of this weather visualization is to aid the user in the identification of weather trends and yield.

2.) Additional Crop Types:

There will be a drop down menu that allows the user to select from a list of crop types other than corn to remake the above visualizations. This additional feature should export the utility of our previous visualizations to analysis of many different kinds of crops.

3.) Web Link to Local Field View:

There will be a button, that when clicked, will link the user to the field data available for any selected group of counties.

Local Field View

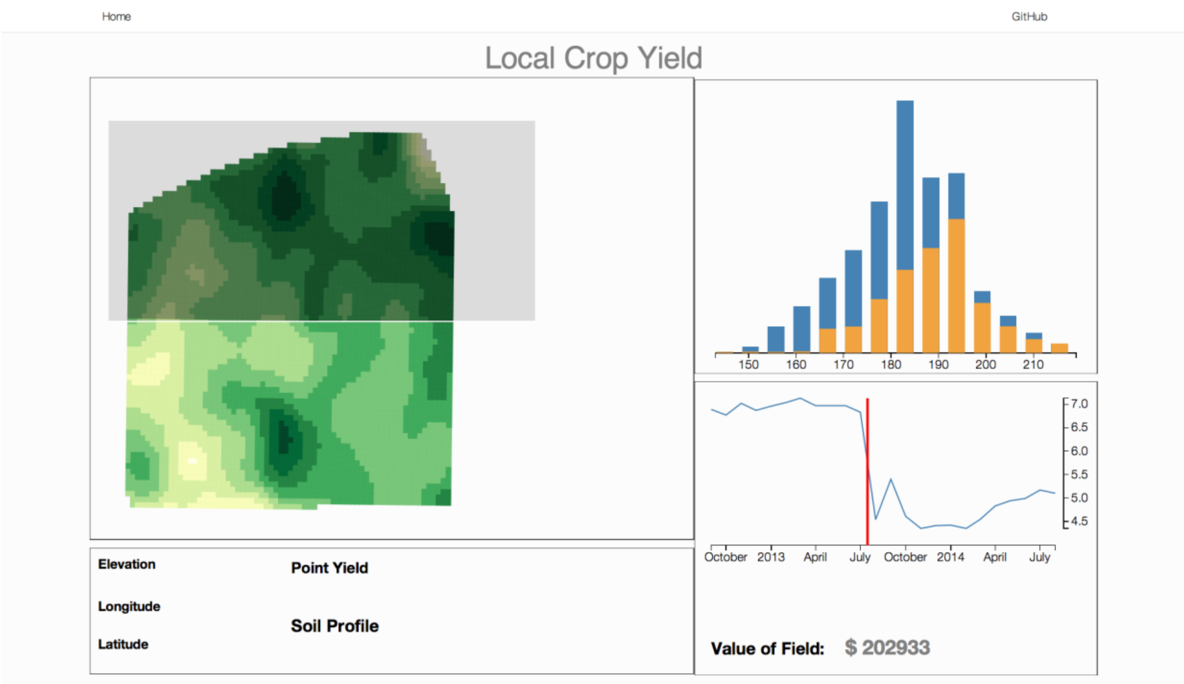
The goal of the Local Field View is to show an in-depth look at maize yield on a farm in Nebraska, including how yield is linked to the soil and how the price of crops affects the value of the field. The Local Field view starts by loading a heat map style view of the field (which is inferred from a Gaussian Process) a histogram of the yield distribution, and the market prices for that crop over a range of months.

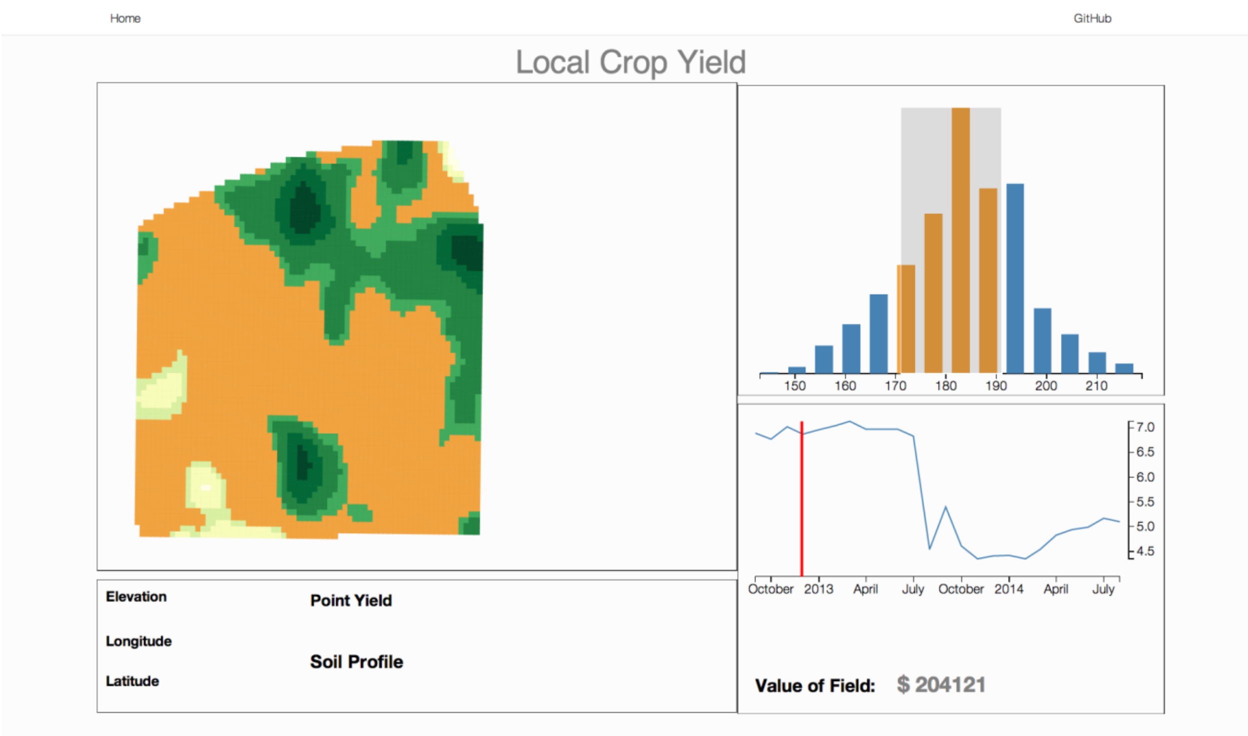
Field View: The field is shaded by bushels per acre. The user can brush over any rectangular portion of the field to view the distribution of yield for that portion in relation to the overall field distribution. The selected area on the field relates to the orange bars on the histogram at the right.

Yield Histogram: This contains both a static view of the crop distribution and the relative distribution of any selected areas of the field. The user can also brush the histogram to select a range of yield values, which will color the field map accordingly (See bottom figure).

Crop Prices: This shows the price per bushel (in dollars) over a year, projected out several months. The user can click on any portion of the chart to quickly calculate the estimated value of the field (or selected area) for that date.

Information pane: This contains basic metadata for a point on the field as the user hovers over it with the mouse.





Current Bugs and Issues: The basic components of this view are running, but there are several bugs we are still addressing.

- There is a conflict with the brush over the field that doesn't allow us to display the point yield metadata.

- The time selector is not lined up properly so the value of the field is not entirely accurate yet.

- Labels and keys are not yet implemented.

- Selected area histogram changes widths and doesn't line up well with the overall histogram.

- Google Earth map data: For the final implementation, we hope to have the field data overlaid on Google Earth imagery of the field. This will give the visualization more context and has been partially implemented in a different branch (i.e. not the master branch).

- Additional fields: Our prototype only has a view of one field. We have data for several other fields that will be available for the final implementation. The user will be able to select a specific field either from the Local Field view or the National Level view.

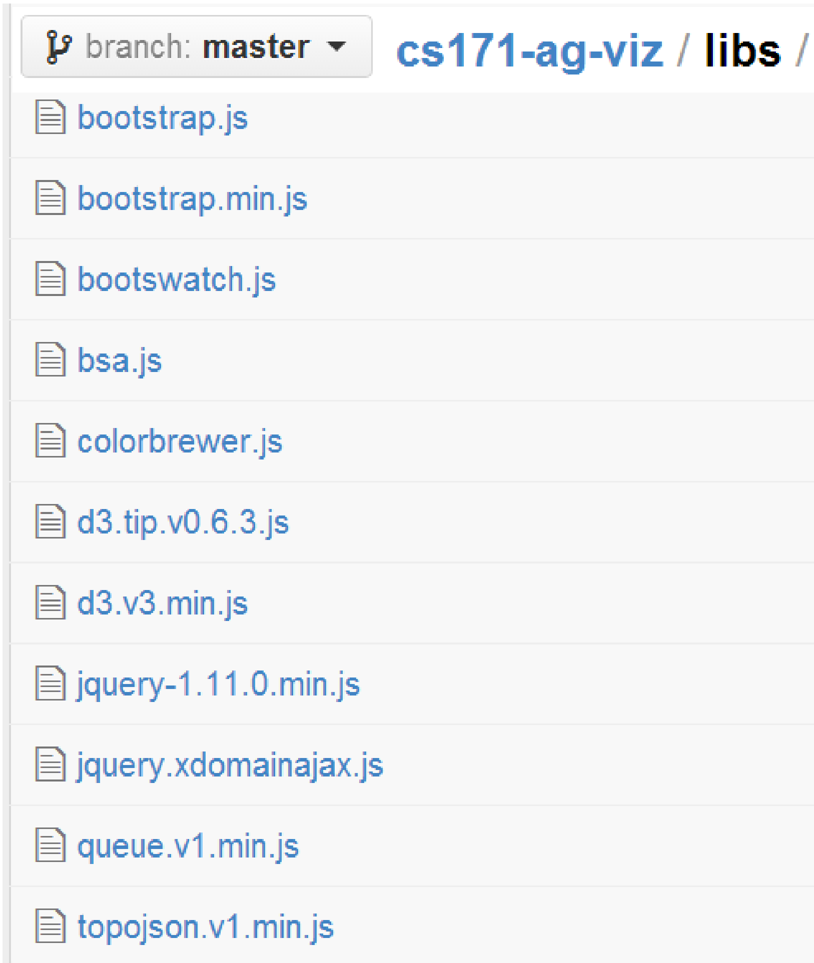
Evaluation

There were several key trends we noticed as part of the changing story of corn crop yields over time. First, if you look at the corn crop yield over time, you'll see the geographic area dramatically change throughout the 20th century. Even more recently, from 2000 to 2013, the Southwest of the US experienced record droughts. From our visualization, we see a noticeable decline in crop yield in the Southwest over this same time period, which warrants further scrutiny, and is exactly the type of observation we have in mind, to be able to see with this visualization.

We'll hold off on further evaluations until our project is nearing completion.

Libraries

We used the following JavaScript libraries in constructing our visualizations:



Additional Resources

Climate Change Will Reduce Crop Yields Sooner Than Thought

Historical Track Record – Crop Production

Treemap Interactive Tool

Visualize GPS Agronomic Data with Maps to Make Data Easier to Understand

Visualizing Historical Agricultural Data: The Current State of the Art

Visualization of a Crop Season: The Integration of Remotely Sensed Data and Survey Data