
The Growth of a Nation

CS171 Data Visualization Final Project

Lauren Wood - 5 May, 2015



Introduction

The century from 1800 to 1900 marked one of incredible growth for the United States. It experienced a population explosion of over 1500%, its land population increased by over 400% and by 1900 it had developed 215,000 miles in a rail network that was unrivaled throughout the world. My goal in this data visualization is to show the growth in these areas throughout the 19th century.

Background Motivation

During a vacation in early 2015, I experienced multiple forms of transportation including air, ship and bus travel. As I marveled my ability to travel from New Hampshire to Florida in 3 hours, I thought back to the history of transportation and how the early growth of the railroads in the country led to a revolution that has made world travel a trivial event. Upon returning from vacation, out of curiosity I looked up travel time in the early 1800's and was reminded that prior to the railroads, my trip from New Hampshire to Florida would have taken three to four weeks of what was undoubtedly uncomfortable travel. During this vacation, I was already thinking about what my topic would be for this course and after returning and starting to do some research and leaving with unanswered questions, I thought this would be the perfect option for a data visualization.

Project Objectives and Questions

The primary goal of this data visualization is to show a visual progression across North America and show the amount of growth sustained during that century in state, land, population and railroads. I primarily want to see the population growth as it related to the railroads. We know that before the railroads, the only way to get from east to west coast was via covered wagon which took months and had very high risk. How much did the new train lines influence population increases? How long did it take? Were people still primarily in the cities or did they move outward? How far beyond the train lines did people move? Did populations increase because the trains were coming or because they were already there? These are some of the questions that I want to answer with this visualization.

Data and Processing

Sources

US Outlines json: [Mike Bostock's US Atlas](#) + [QGIS](#)

State Outlines json: [Eric Celeste's US GeoJSON files](#) + [QGIS](#)

US Railroads json: [Mike Bostock's US Atlas](#) + [QGIS](#)

US Population data: [University of Virginia Historical Census Browser](#)

US Railroad data: [1900 Railmap at Perry-Casteneda Map Collection at Texas](#)

US Railroad data: [American Railroads: Their Growth and Development at Central Pacific](#)

[Railroad Photographic History Museum](#)

US Railroad data: [Library of Congress Railroad Collection](#)

Latitudes and Longitudes data: [findlatitudelongitude.com](#)

Data Processing

As of today, 29 March, this project has involved a considerable amount of data manipulation before even getting the project off the ground. By starting to think about this project in January, I was able to start data crunching long before learning the d3 skills to develop the project. The biggest part of this was manipulating the railroad data. Due to the sheer amount of rail lines (215,000 miles by the year 1900), I opted to focus on primary trunk lines and left out smaller lines that were local and didn't go far. Doing this though, I still had to start with the existing shapefiles available for railroads today and using the software program QGIS, I was able to delete extraneous lines and shrink the data set down to trunk lines. I then manipulated the lines to be broken down according to decade built. After the final work in QGIS, I used the software tool ogr2ogr to convert the files to GeoJSON format suitable for working with in d3.

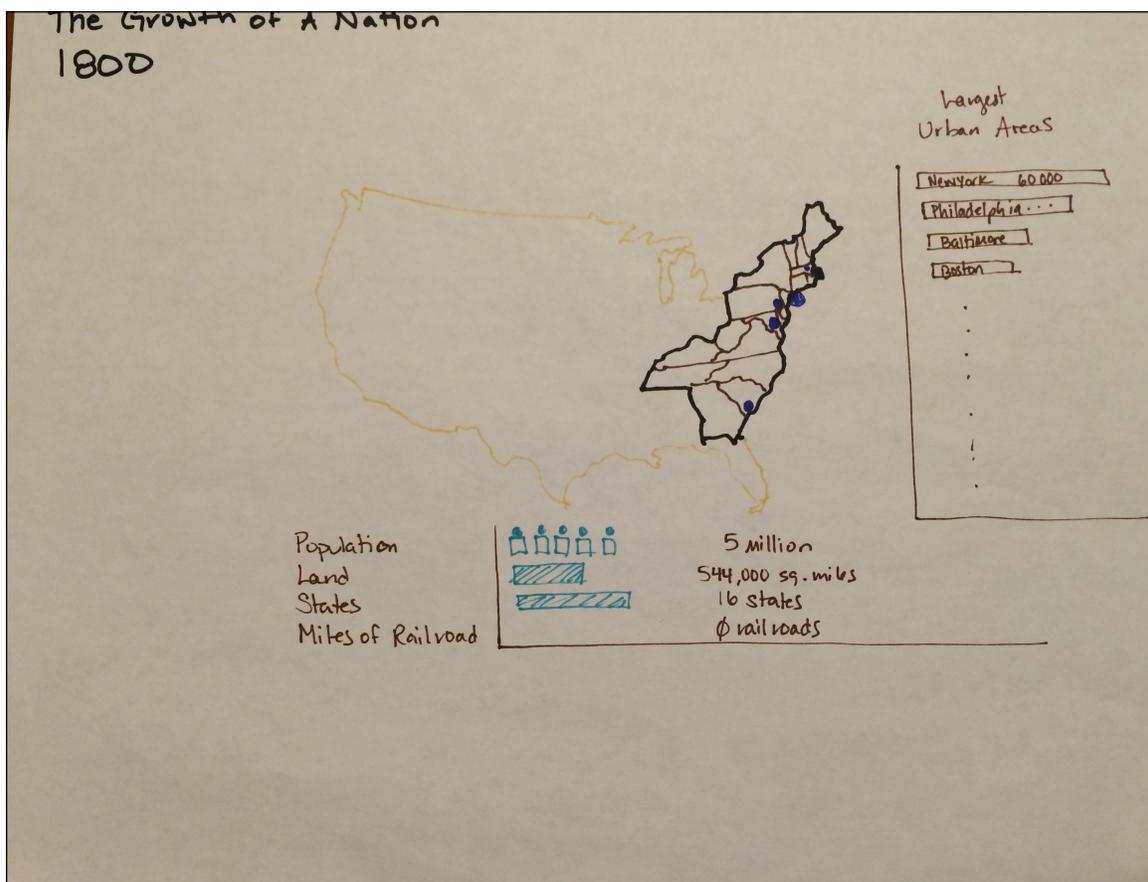
In addition to working with the rail data, I also had to collect and manipulate city data and then add latitude and longitude data to it. The geojson county files from Eric Celeste were already formatted with ID's and names which will allow me to map county census data to the county graphical objects. I also had to go through and create various US outlines as the currently available files for the outline of the US are obviously for all 50 states + territories which were obviously incorrect for the years being observed.

Visualization Design

Design Areas

Primary Map Area

The primary area for viewing will be the central map area. This will be the area seen in the images below with the map in it. It will include current US outlines, states and counties for the country at the date selected. In the images, the US outlines can be seen in black, with states in brown. Also in this area will be the railroads which are visible in purple. Also visible are the cities in blue which will be bubbles that vary in size based on current population. Cities shown will be in the top 100 largest urban areas based on census data.

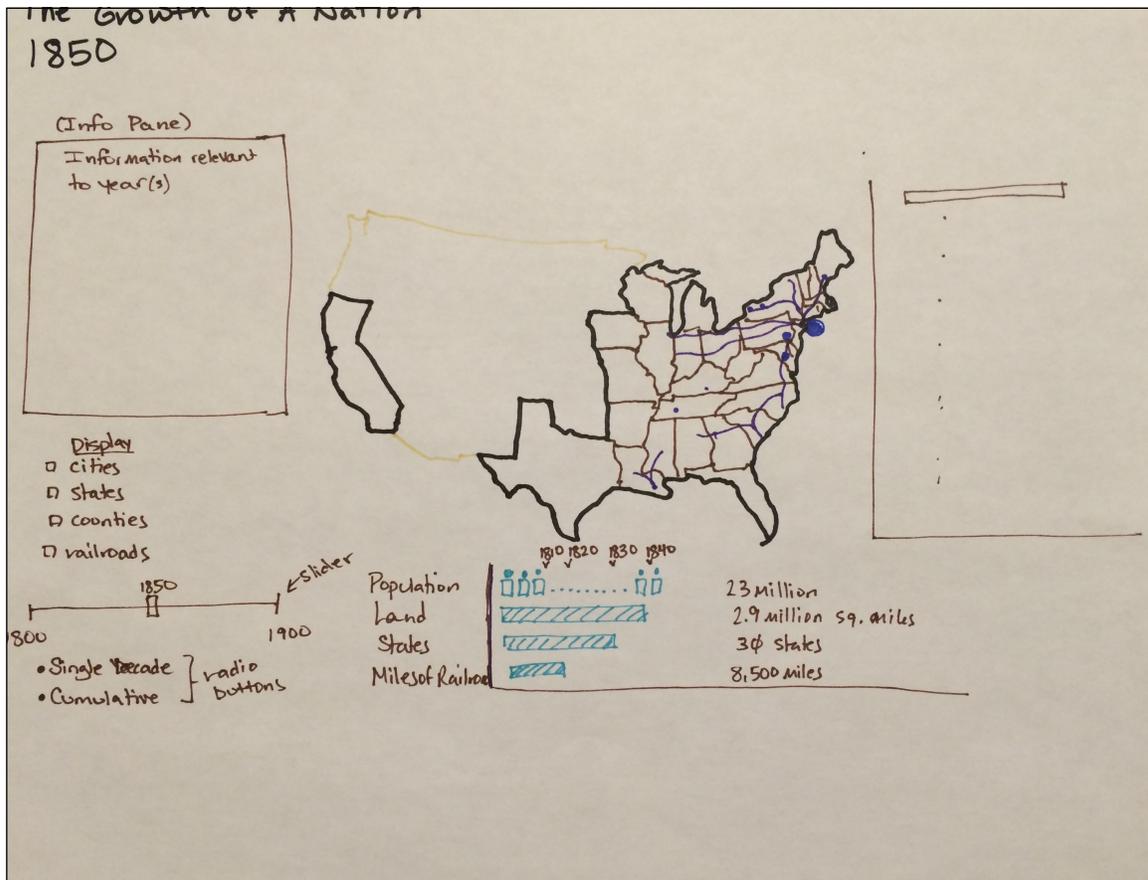


Growth of a Nation Sketch for 1800, with map, city and summary data areas

City/Urban Areas

On the right hand side of the screen will be a listing of largest cities in the US as of that year. This will use a sideways bar chart to show relative population sizes for these cities. I may also expand this to include state populations, population densities or other quantitative data

that is relative to this experience. In this field, when a bar is selected, it will make the city in the primary map area “pop” momentarily to give the user an idea of where it is located rather than trying to guess. This is particularly beneficial to users who aren’t familiar with American geography.



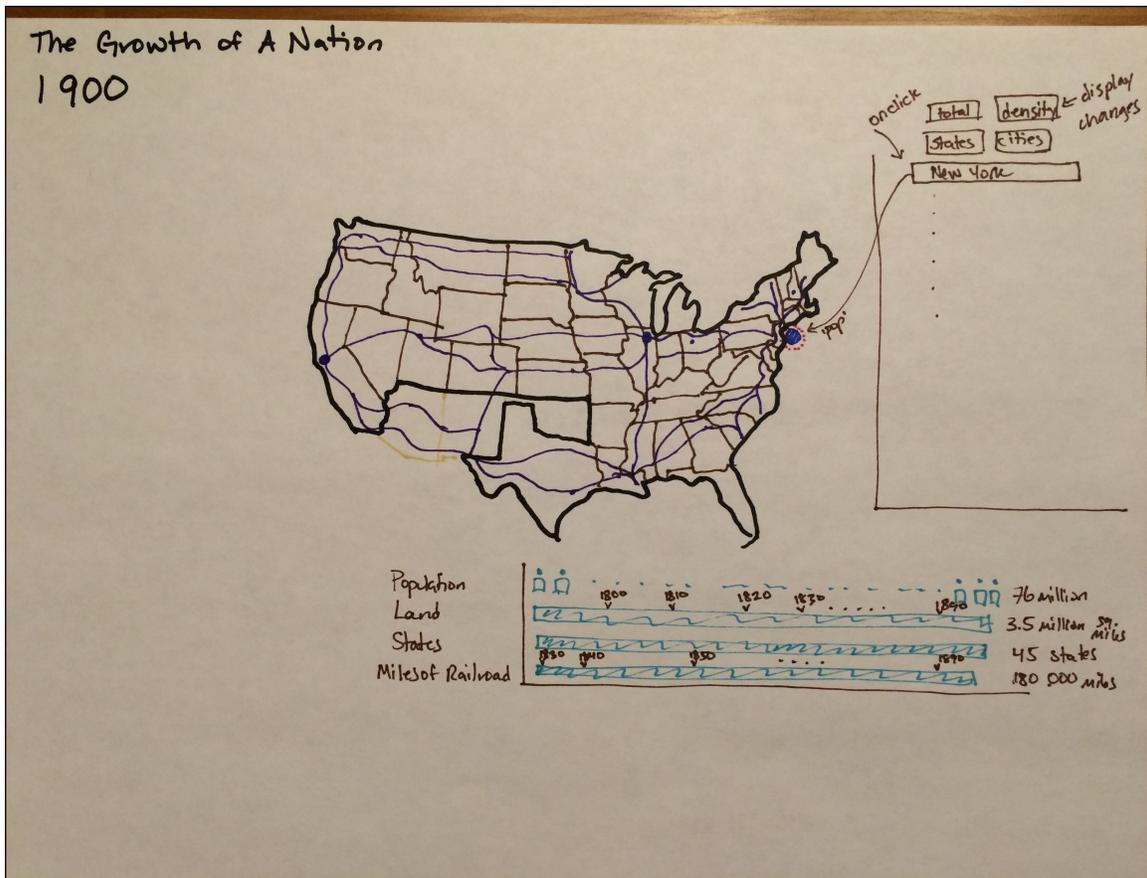
Growth of a Nation Sketch for 1850, with info, map, city manipulation and summary data areas

Growth over Time Area

At the bottom of the vis, there will be a summary bar that shows the growth over time with relative values. This will include things like population, land area, number of states and miles of railroad. These values will show not only current day values, but will have markers showing where they were in each previous decade. An example is visible in the 1850 sketch where the year markers are visible for the population entry.

Info Pane Area

On the left-hand side of the vis will be an info pane area that displays information relevant to the currently selected year. This may be information about states acquired, railroad expansion or population growth.



Growth of a Nation Sketch for 1900, with info, map, and city areas

Data Manipulation Area

In the bottom left area of the vis will be all of the data manipulation tools. This will include things such as the time bar to move from one year to the next, and also visual filters to give better insight into the data. Users will be able to view all options or they could view just city population data or just county data. This solves the problem of city bubbles becoming too large to show nearby county areas.

Design Evolution

My project changed very little between the time that I started it and the time that I completed it. I started to map out my ideas in my head in January, when the project was initially announced. This allowed me to work out the details in depth and finalize ideas long before I started to write the code. This process allowed me to make good design decisions and have clear goals when I did start development.

Features

Features Critical for Success

US Map

The biggest feature that must be implemented for a successful project is the map area in the center of the page. This map must support a time progression showing the growth of the population by county during the years from 1800 to 1900 in 10 year increments. This will be color coded to show the growth over time.

City/State Data

This is the area to the right of the screen. My initial thought is to have this display city data but it could also be used to show state data or county level data. I'm leaning toward city but this may change based on the end result. There needs to be something in that column that drills into local data, rather than the population as a whole and it needs to use the bar charts to display relative values and be click-able to show users the specific geographic regions referred to.

General Information Pane

This is the area at the bottom of the screen. In addition to the map area which will show growth over time, this area also needs to show growth over time with country statistics. Numbers/lines/people will increase over time to show the relative growth. These will be linear.

Features Nice to Have

Detailed Information Pane

The information pane would be really nice if it had a large set of information pieces to add to the vis. It's interesting to see the progression over time but it would be even better if it provided a background. For example, in one decade, 35,000 more people were leaving California each year than entering because of the end of the gold rush. I want to be able to

explain why the population decreases during that time. I also want to include information about how the railroads impacted immigration. Including information that explains WHY things happened would be fantastic.

Initial Info Screen

I'd like to add an informational intro screen that displays while all of the data is being loaded. This would include a short summary of what the visualization is about and also hide the time it takes to load all of the data.

Zoom Ability

I would like to add functionality that allows the user to zoom into 6 sections of country to see things closer up. This zoom would be for northeast, southeast, north mid central, south mid central, northwest and southwest.

Proposed Schedule

As of today, the proposed schedule is as follows. This may evolve over time as things change.

Date	
3 April	Proposal due (this document), Repository set up.
10 April	Initial framework design complete and setup. All javascript classes and basic structure completed. Must have a loadable web page with spaces for the svg elements.
17 April	Milestone 1 - Data acquisition mostly complete, data structure in place, working visualization. This is an assignment to be turned in including process book (this document) and existing code. Must have a basic structure that can progress through the years and show city data that changes in the right pane population statistics updating in the map pane and overall data visible in bottom pane.
24 April	Must be mostly working except for final finishing touches.
1 May	Finishing touches, final process book, video completed, website completed.
5 May (Tuesday)	Final due date.

Project Location Information

Repository Location: <http://www.github.com/laurenwood/laurenwood.github.io>

Website Location: <http://laurenwood.github.io>

Project Structure

index.html	The initial page
\libs	
d3.js	Used for svg elements
jquery.js	Required for bootstrap
bootstrap.js	Library for page layout
queue.js	Used for loading the (many) data files
\css	
fonts.googleapis.com	Nice free fonts
bootstrap.min.css	Again, more bootstrap stuff to bring it together
myStyle.css	My personal style sheet to control style
\js	
mapVis.js	svg object that will display the map
infoVis.js	svg object that displays the left info pane
controlVis.js	svg object that displays the control panel
chartVis.js	svg object that displays the city chart on right
summaryVis.js	svg object that displays the overview chart

Data Structure

SVG Paths

country outlines	A set of paths that show US country outline based on year
state outlines	All of the state outlines
county outlines	All US county outlines
railroad lines	The railroad trunk lines by year

Data Objects

data_by_state*	
state	
ID	Unique ID (key)
postcode	Two letter state code (AL, AK, AZ)
city (array)	
ID	Unique ID (key)
name	Name of the city
latitude	Latitude of the city
longitude	Longitude of the city
years (array)	Population of the city by year
1800	
.	
.	
1900	
county (array)	
ID	Unique ID(key)
name	Name of the county
years (array)	Population of the county by year
1800	

.
.
1900

* Note: This structure may change over time as development progresses.

Implementation

My actual implementation was as defined above. I used the previous homework assignments as inspiration for structure for the object. I had each svg object managed by a separate class that were all linked and updated by event handlers. This kept code management at a very easy level and allowed me to work on one section at a time without overlap.

Sources

Photo from front page: **“First Locomotive on the New Haven, Derby and Ansonia Railroad.**

The engineer standing by the engine was Mr. Whitlock, afterward Master Mechanic of the road. The locomotive was facing West, a short distance from the union point of the Derby and Hartford and New Haven Railroads. This track and the brick buildings have been removed. Photographed about 1871.” <http://www.tylercitystation.info/track-3---nhd-extra.html>

Project Progress

3 - 10 April

Issues

Data loading needs to speed up. Currently , it takes up to 3 seconds to load all of the data. This is too long. Timing is as follows. I may need to do some file optimization. Perhaps simplify the json files or investigate the use of TopoJSON.

Files	File Load Time	Path Display Time
Country Outline JSON Files	250ms	500 ms
Railroad JSON files	400ms	1000 ms
City Population file	30 ms	N/A
County Population file	50 ms	N/A
County JSON Outline file	1000 ms	18000 / 3000 / 5000 ms
State JSON Outline file	1000 ms	3000 / 1000 / 2000 ms

Progress

I have my first semi-working version of the project available. The configuration involves several svg objects that all work together and are each driven by their own classes. These are all in place and can create their initial layouts. There is no communication between the svg objects, but there is a map available, charts for populations, text fields for information, a control panel and summary pane at the bottom. The layout may change but I'm ready to meet with Daniel next week for initial review.

11 - 17 April

Highlights

I met with Daniel this week to review my project. I met all of the Milestone requirements for next week and had a working prototype. He gave excellent feedback, primarily relating to performance improvements. I used his information to start determining how to improve performance in my code and speed up load times. I got the event handling implemented by the end of the week and had a control panel that could control switching between years and viewing options.

Issues

The biggest issue continues to be performance. The load time is highly variable based on computer (higher processing ones load faster) and also source. When loading using a localhost server, the load time is very fast: loading it directly from the github.io server slows it down significantly. This is something that I have no control over.

Progress

The control panel is functional now with checkboxes, radio buttons and a slider that functions. The map updates with city and population colors and the chart is created with a default sorting by population.

18 - 24 April

Issues

The biggest issue this week remained with performance. I analyzed the times for each class in the visualization to determine where the slow timing was coming from. What became increasingly obvious was that the timing was in 2 places: the file loading and the map creation.

Progress

After determining the source of the processing time, I started to update things. I used mapshaper.org to shrink the size of the json map files. This improved things in two ways. First, the time to load the files into the program decreased as file size decreased and also with the lower resolution of the maps that I was drawing, the calculations took less time. I found that when I decreased the resolution by 75%, I got the best value. The lines were still curvy enough to accurately reflect their paths, but the file sizes were small enough to significantly improve file load times.

I also combed through all of the classes and did some data restructuring during this time. Prior to this week, I had a huge number of for(... in ...) statements which, thanks to Daniel, I realized were a big source of latency. I redesigned some of my data structures so that I could easier loop through them to pull out the data that I needed for my vis. While this did involve some significant rewrites of various parts of the vis, it proved beneficial in the long run and sped up timing considerably.

During this time, I also rewrote some of my mapping code so that fewer objects were rewritten each time. Previously, each time the year updated, I redrew the state and county lines, or in the case of the city view, deleted them. I re-wrote this to instead make the outlines invisible and didn't redraw them all. This sped up the processing as well.

25 April - 3 May

Progress

This week was spent on the final stages of the project. I did some layout updates to provide better viewing and finalized the information pane and the legends. I wrote up each of the info paragraphs for each decade and modified some layouts. The legends were finalized after I decided to switch from absolute county population to population density. I did this because there were some incredibly large counties out west that look like they had huge populations but it was only due to land area. Once it was formatted for density per square mile, the colors came through much better and made more sense.

This week I also finished the video which I found rather difficult. Having never created one before (I have generally just written documentation), I had a difficult time getting used to the idea of talking to an invisible microphone and describing the workings of the project. I did finally get it down however and feel that it was fairly successful.

During this time, I also opted to get some external feedback from friends and family. I got some great ideas about colors, layouts and functionality that resulted in a few changes.

Evaluation

I feel that my final visualization was very successful. My vis shows a great relationship between the trains as they grew across the country and also the population growth. The information pane provided sufficient detail as well to explain some of what was going on. I think that my favorite part of the vis though was the summary pane at the bottom. I like that it allowed a comparison of miles of track, population and land growth so that it could be seen that not all decades had equal growth.

I think that while it was a very successful visualization, there are some things that could be added to make it better. The option to zoom into areas would be nice as well as the ability to change the encoding for population. If a user had the ability to define their own population scale, they could better focus in on detailed areas that they were interested in.

Additional Sources

In addition to the sources listed above, I also spent a considerable amount of time reading history books to better understand the history of the railroads. These books are below:

EuDaly, Kevin, et al. *The Complete Book of North American Railroading*. Minneapolis: Voyageur Press, 2009. Print.

Grant, H. Roger. *Railroads and the American People*. Bloomington: Indiana University Press, 2012. Print

Hubbard, Freeman. *Encyclopedia of North American Railroading: 150 Years of Railroading in the United States and Canada*. New York: McGraw-Hill Book Company, 1981. Print

Historical Guide to North American Railroads. 3rd ed. Ed. Jeff Wilson, Randy Rehberg. Waukesha: Kalmbach Books, 2014. Print.

Law, Bill. *Fifty Railroads that Changed the Course of History*. New York: Firefly Books, 2013. Print.

Rails Across America: A History of Railroads in America. Ed. William L. Withuhn. New York: SMITHMARK Publishers, 1993. Print.

Wolmar, Christian. *The Great Railroad Revolution: The History of Trains in America.* New York: PublicAffairs, Perseus Books, 2012. Print