NEWSROOM STATISTICS IN THE DIGITAL AGE

Travis Hartman

Keywords: Statistics, Investigative reporting, Data, Regression, Journalism, Outliers

Chapter One: Introduction

I came to the Missouri School of Journalism with the expressed goal of learning how to create interactive graphics. My background is in photography and photo editing. I was drawn to photography for the simple reason that I found satisfaction in creating things, but as I studied it in earnest, I found it was a craft that rewarded the process of using the camera not simply to make photos but to render visual ideas. The results of my exploration of interactive graphics have been much the same. The code/camera is not the destination but the path you walk to get there.

The first class I enrolled in for my master's program was an information graphics course, and both the strengths and weaknesses of the class have informed my studies. It was challenging and exciting to learn to think in a design and efficiently communicative fashion, but the course was centered on static design principles. Although it was good to learn the principles that govern paper, it only made me crave the interactive work more.

I've pursued much of my other coursework through independent study and weekend boot camps that pertain to emerging technologies in journalism. I also have pursued a number of classes outside the journalism school in the realm of statistical analysis in an effort to prepare myself for insightful and informed dealings with large bodies of data. My assistantship in the NICAR database library has also been a boon in this regard. I've worked through the entire process of acquiring data directly from large, sometimes contentious governmental bodies; inspecting and cleaning data to make it usable; then analyzing it via structured query language. My mentor in this regard has

been Liz Lucas, director of the database library. She has great patience and has guided me over the hurdles involved in procuring and working with data.

My eventual goal is to develop interactive graphics for a news organization. I love working with information that is current and am constantly engaged by the variety of problem-solving skills needed to be innovative and successful in a news environment. I see a number of relatively new websites (Vox, Fivethirtyeight, and Upshot to name a few) that have emerged around a much more data-centric type of journalism, so it's clear I'm not the only one who enjoys working with and consuming this type of journalism. I want to work in a place that will allow me to refine my programming skills on a daily basis and learn the logic that lies behind code. With an increased fluency, I'll be able to produce more sophisticated storytelling with greater efficiency. This fluency will also allow me to contribute meaningful work under the ever-present deadline time constraints.

The research portion of my project explores how statistics are used in the newsroom, specifically how these techniques are used in an accurate and critical fashion to help communicate the news. The information from this project will help me find the most accurate ways to present my data and give me insight into what kinds of stories are approachable within a given dataset. I have also noticed an overlap between programming and statistics. They share many core concepts of organization, hierarchy, and terminology, and this cross-pollination will only help to cement the concepts in my mind.

Chapter Two: Dispatches From The Field

Week 1 dispatch

(1.09.15 – 1.16.15)

I started working on the 12th at Graphicacy, a creative analytics firm. They do interesting work, and can already see I have much to learn. Graphicacy is a small company, with 5 members, the majority of which are ex-journalists. The interpretation of governmental data plays heavily into their work, and the presentation of trends in a clear and concise way lies at the root of their mission. Interestingly, they have their department split into designers and developers. The designers produce an illustrator file that serves as a template for the developers. The developers then match it in code, making the design interactive with the same visual assets. Seeing as how I have much to learn in both aspects, I'm sure learning to match a design in code will be beneficial, though I have a large desire to engage in the design as well. We'll see how that aspect plays out.

In terms of working this week, I've been communicating regularly with Jeff Osborn who is the Creative Director at Graphicacy about how best to begin. He knows my desire to focus on JavaScript so I'm guessing as soon as the agency wraps up it's current spate of contracts he will get me in on the ground floor of upcoming projects It would make no real sense for me to try and engage in any of the ongoing projects in the latter stages. In the meantime, I've sent out a memo of ideas that might work thematically on their blog, which is an area they would like to bolster with fresh content. Osborn and I discussed the ideas and he particularly liked my idea of a "Congressional ERA" so we

could rate the productivity of our congressmen, in a similar way that people rate sports figures based on statistics.

Since then, I've discovered a large data set from political scientists (Volden and Wiseman) and that have run statistical methods on data aggregated from Roll Call, the congressional publication that catalogues all the bills congress works on. Through this, Volden and Wiseman have established a "Legislative Effectiveness Score" based on a number of factors, but ultimately concerning how many bills a representative has initiated and how far the bill progresses through the system. The score basically boils down a series of weighted averages, tempered by many factors, including the importance of the bills passed, length of tenure of the representative, and if they are a member of the majority party. The data is free and available for download, so I grabbed it. It's a fascinating dataset and I spent the first part of the week combing through it and creating a data dictionary to categorize the information within it. Also looking into it and finding flaws in the information in order to prevent missteps down the line. After becoming quite familiar with the data, I again met with Osborn, and we plotted out a strategy for visualizing the data. or more distinctly, we plotted out that I'll explore the data and would develop a strategy based on my findings.

I've already started to visualize my findings and having not one but two developers sitting within five feet of me to help is amazing. The lead developer, Reed Spool is a patient teacher and explains things clearly, concisely, and with a large deal of wit. I have already surpassed my previous efforts and made some great mental leaps in terms of understanding how to process data with JavaScript, using D3 in particular. Spool, and the other developer, JoElle Straley, and I have also started a book club with

weekly readings and discussions. The aim is to up our coding skills, especially when it comes to creating code that is all working along the same standards. Very cool stuff.

Next week class starts, and I'm sure I'll have more to report on that front next week.

Week 2 dispatch

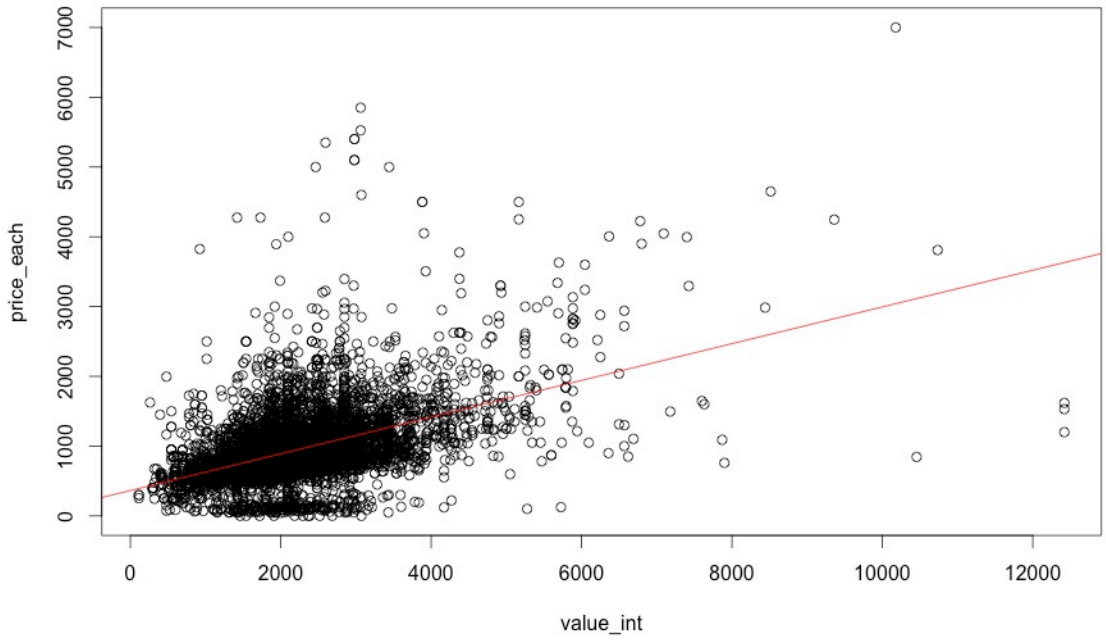
(1.17.15 – 1.23.15)

Week two was productive, despite being a short week due to MLK day on Monday. I went in to work anyhow, and managed to make some good progress on my congressional project. Tuesday was orientation at the Missouri University press club office where some classes will be held. We read through the syllabus, introduced ourselves around to each other, and got a tour of the National Press Club, which I have since become a member. Over the next couple days at Graphicacy, I was asked to start doing a little work in R, a statistical analysis software they use for data exploration. This is good, as it's a program that I used while in school, and I have a familiarity with it, though the task leans more towards the analysis side. I could certainly use a refresher course with it, and I believe working towards a goal with an interest in exploring is the best way to learn. At first I loaded up a script on an existing Graphicacy project that deals with the history of baseball. They have been working on this for quite a while and it's a beautiful poster project with an interactive version as well that's equally incredible. I see now that R is capable of doing some truly amazing things in the right hands. The last few days have

given me a radically different perspective on what R is capable of, which I truly appreciate.

But the baseball project, which has been laid aside in light of other work, is constituted of an R script that has been passed through many different sets of hands. Very few of which were looking to create code that is readable for others. It is probably far beyond the scope of my current abilities in R and while I'm sure it would be fascinating to untangle, it would take a good while for me to get up to speed on. During my weekly meeting with the creative director, we discussed that if there were a modular section of the baseball project that I could break off and work on, it might be a better way for me to contribute in a meaningful way considering my time frame and current skill level. This conversation about how I can best contribute led me to suggest that I could and would be happy to convert some of their static projects they are being commissioned for into interactive versions as practice. This was indeed very good practice for me and perhaps also a way to show clients what potential lies within their projects. This seemed amenable to him, but immediately had me start working on going through some data for an account for houseplans.com. I spent some time looking for interesting facts about their sales data and in the time I had, tried to figure out the structure, query the data with SQL, export the data, load it into R, and run a few quick lines of code on it. The business model for houseplans.com is one which the customer picks the features they like and the website shows you the blueprints available. In my analysis, I wanted to do a quick look to see if price was correlated with the total square footage in the blueprint. It would make sense that a blueprint for larger house would cost more- so loaded up the data from about

7000+ records, drew up a scatterplot and ran a linear regression with price as the dependent variable and total square footage as the independent variable.



The results will probably come as not huge surprise, but it's interesting to see the exact relationship from the data. By my best interpretation of the summary statistics, roughly 23% of the price in each blue print is explained by the amount of square footage.

In other news, I have reached out to a couple people regarding my research and have an appointment with Derek Willis at the NY Times office next Friday. He said he'd be happy to speak with me regarding my research and perhaps find some of the other Upshot staffers as well. I have also reached out to fivethirtyeight.com, but have not heard back yet. I am also trying to find a good contact at Vox to discuss my project with. An editor at Vox recently reached out to Graphicacy in reference to an annual federal budget visualization, so that may be an avenue in.

The seminar this week consisted of a round table discussion with Tom Rosenstiel and we discussed how technology is affecting the media, how the current problems with advertising have surfaced, and ways in which the industry can address audiences in a more relevant way. We then went to the Newseum and took a tour. It is a very interesting building. The museum has elegantly tackled the problem of creating a fluid environment that can respond to the evolving landscape of media and remain current in their displays.

Week 3 dispatch

(01.24.15-01.30.15)

Over the last weekend and throughout this week, I've been applying to internships and fellowships across the country. Getting into another solid working environment with excellent peers would hopefully make my progress escalate rapidly. So it's been a very busy week, crafting emails, cover letters, updating my portfolio website, and networking.

In regards to the internship I'm currently in, it was a busy week. Early on, I spent time refining and getting help programming my congressional analysis. I also took a meeting with Jody Sugrue, a former National Geographic multimedia staffer. She gave me some much needed design and user experience advice on my project. Also, in discussions about Graphicacy's annual poster work, we realized that the 2016 federal budget would be released on Feb. 2. There has been interest in this project from several avenues and producing this poster in a speedy and efficient manner has become a priority I spent a large portion of the week working to understand the existing process for producing the poster. I also spent time working through the scripts in R to assess what

raw data are needed and where to access them on the internet. It was following breadcrumbs for a long while, but after discovering a hidden cache of files from last year, everything crystalized. I'm working with the designer for the poster, Josh Korenblat, and we have been discussing ways to improve the overall concept and produce a better workflow. I have managed to implement several of his suggestions by reprogramming the automated script, and feel it will be much less laborious than in previous years.

During my weekly meeting with Jeff Osborn, I expressed my dual desire to contribute meaningful help in however I could, but also that I had a strong desire to focus on a client based interactive project that I could code in JavaScript. He understood and we discussed a few potential projects at varying stages of production. We eventually decided it might work best to work from the ground up on a project, in order to help with the ideation as well as implementation of the communication strategy for the client. Jeff has been coordinating with a health data company around the corner from us at the We Work office, and we immediately walked down the hall for an introduction. I will begin work on this next week after seeing through my responsibilities on the Death and Taxes poster.

My research on statistics in the newsroom has been coming along quite well. I've managed to conduct two interviews this week, one with Jeff Ernsthausen, a data reporter for the Atlanta Journal Constitution. Jeff is a former statistician for the federal reserve, and has most recently crafted a statistical model on past voting records to predict if a bill will pass the Georgia Senate. Named Predict-a-Bill, it's has close to 85% accuracy, and Jeff spoke about on ongoing process to refactor it in order to make it better. He discussed his process in using statistics, and his academic background will lend a unique

perspective to my eventual paper. His previous experience before working at a newspaper is a rarity among data journalists.

I also spoke with Derek Willis of the New York Times. He has recently been focused on working within The Upshot, a relatively new section of the Times that appeals to a more data literate audience. His work revolves around campaign finance and spoke about how he uses statistics on a daily basis in the research he conducts. He also talked about when he does and doesn't use statistics in the text of his work and how social science is perceived overall in the industry of journalism. His point of view is very well reasoned and full of examples that will serve as a buttress to frame the larger challenges in using statistics.

In class, we travelled to see Amber Marchand, the communications director for Senator Roy Blunt. She spoke about how best to work in a professional manner with a senatorial staff, both in terms of what to do and what not to do. I was a little surprised at how focused she was on returning calls and emails from local media. She even pointed out how Senator Blunt will personally get back to one reporter each day from Missouri, in order to remain in touch with his constituency. I also found it fascinating to hear her describe the difference in political philosophy in regards to voting for bills outright versus voting for amendments. I believe she said Blunt voted on 14 amendments during the whole of last session. Since republicans have taken back control, there has been a marked increase in the amount of amendments voted on, up to 18 in one day. She was also very eager to help us in our work, and suggested several times that we reach out to her via email should we have any further questions.

Week 4 dispatch

(1.31.15-2.6.15)

This week I dealt with the federal budget. All 1.5 Trillion dollars worth. Which in reality means I used the scripts Graphicacy had to process all the data I downloaded from the federal government that was released at 11:30 AM on Monday. The R scripts worked great, and the modifications I made helped the designer when importing the resulting PDF's of circles into illustrator. The process took a while to refine, but it's in a good place now. The last set of budget items were more laborious, as I had to compile them by hand, but were finished by the end of the week.

In conjunction with this federal budget project, one of my colleges and I discussed working on the interactive version, which seems like a grand and difficult project. Determining the organization of the user experience is a lot of work and there are many ways to imagine parsing and looking through that much data. Ideally we would be able to give a good sense of the macro and micro at the same time. But on a compute screen, it would be a challenge. I'm sure there will be more on this, as we just kicked around ideas this week.

I also made some great progress on the congressional project. Going through this process is invariably going to give me problems. But some problems, no matter how much research I do, still just don't make sense. With a developer to explain the hurdles I come up against, help me refine my code, and point me towards readings, it makes all the difference in the world. I was having a hard time updating my data once I moved my slider to choose which session of congress to display. The lead developer made a few deft

changes and explained the details of a guiding principle of D3, which is the ability to display the data in a dynamic fashion. I am now feeling good about where the project is going and how I will be able to apply these lessons to the next project.

Towards the end of the week, I managed to capitalize on meeting with the DARE group down the hall. I had a paper they wrote about how migration in, out, and within of the DC/MD/VA area will have significant effects on healthcare providers. In discussing the narrative that we might focus on for the interactive, I suggested looking at DC alone and finding contrasts. For the last 50 years the population had been falling and in the last decade it has risen 5%. While it boasts one of the youngest, wealthiest, most educated workforces, DC also has a very high HIV infection rate, and much higher than average heart disease rate. So might be very interesting to try and tease apart these stats. I'm going to pull the census data and they will be working the health data angle. The challenge I think will be to match up the geographic specifics and finding the common denominator for the most recent year available for all the data.

I interviewed Tom Meagher from the Marshall Project for my research and had a great conversation with him. Also during my visit to Bloomberg, I caught one of our hosts, Mike Dorning, in an aside and asked if he knew any reporters that use statistics in his organization. He told me to email him and e would pass along my information, which I did that day. I also discovered a good source for statistically minded journalists- in the NICAR-L Listserve, I saw that Jacob Harris has posted a skill share document online, which is to say, you type what you want to learn and what you can teach on an informal level. Several people have signed up for wanting to learn Monte Carlo simulations, which is a statistical method to compare competing statistics for small samples under realistic

data conditions. I'm guessing that these people might be interested and willing to chat with me on my research.

Week 5 dispatch

(2.7.15-2.14.15)

I am currently involved in several projects at Graphicacy. The pace at which projects are accomplished here seems slower than the editorial news world, and I'm beginning to look with more specificity at what an "abundant body of work" may entail in order for my project to be a success. I have been quite busy since moving to Washington. I've learned a lot in the last five weeks. But I do not have much material to actually show from the time I've been working. This is an issue I feel I need to correct, and with good speed.

In light of this mindset, I've redoubled my efforts with the congressional effectiveness project. I have been working to apply some of the lessons I've learned from my research interviews and decided to dig into the data to look for more storytelling elements. I started off well with this project on both the presentation and programming front. I have also developed a good bit of the mechanics for the interactive interface. To propel the next set of design decisions, I need to find the aspects of the data that are the most salient and interesting. These aspects will then dictate how the interactive is modeled from here on out. Form follows function.

To find these interesting and salient features, I have made a list of queries and will use them to usher the code as I work through the questions. I'll look for outliers, lack

of outliers, where the data is uniformly shaped and where it is not. As almost every reporter I have interviewed has noted, it is critical to turn over the data in every way possible to find the connections and disparities. I'm not sure I'll be able to answer why these permutations have occurred, but at this point, I think it's more about locating the oddities and noting they exist.

To perform this analysis, I have been using R, a statistical programming language. I have had some success, but with many starts and stops. The speed at which I'm able to progress is not great and I find my lack of knowledge with the language to be the hurdle. I do think that the time I put in now will pay off in the long run, as greater fluency in my analysis skills should shorten the overall turnaround on future projects. I'm just waiting for the day in which I don't have to Google every third piece of code in order to make trivial progress. Patience is a virtue I'm trying to cultivate. But it seems that the practice of cultivating patience requires unto itself a good deal of patience. Please note that the humor here is not lost on me.

I spent a good deal of time this week sorting through census data in preparation for my meeting with the DARE global initiative group on Monday. I'm eager to see what kind of health data they have found and at what level of geographic specificity it is. Geography and time are the points at which all this data needs to overlap, and the more granular the level, the clearer picture it will paint of the health changes over the last 40 years in the D.C. area.

Beyond the projects I'm working on this week, I interviewed Jon McClure who is a news applications developer at the Dallas Morning News. His mindset is ambitiously statistical, and cited several examples of stories that were borne from analysis directly.

McClure sees a culture of empirical investigative analysis at the Dallas Morning News. He also noted that the trust his department has built with the upper management over time has contributed to his ability to use statistics in his daily work.

I was also able to sign up for a journalism boot camp in coordination with Open Data Day DC, which is a hackathon and training event in celebration of Open Data Day. I can't say I even knew about there was an Open Data Day, but I'm a fan. The boot camp is set at the World Bank and look to be a great opportunity to work on projects with other journalists and find out about new sources of data in the federal government and elsewhere. I am very excited about this in terms of potential for story ideas. I also think it will be great to be immersed in working from an editorial mindset for a couple days. I think the boot camp will be a very valuable experience and I should have many things to report from it in my next dispatch.

Week 6 dispatch

(2.15.15-2.22.15)

The first part of week six started out with furious effort to understand how to perform and visualize exploratory statistics in the programming language R. I spent the weekend working with my congressional data in an effort to understand the range of the information contained, and to find out where there may be interesting relationships within the data. Performing the statistical exercises on the data serves a dual purpose. First to grant insight into potential narrative threads in the story I'm trying to tell. Secondly, finding the threads will determine the way the information is presented, and is the

necessary next step before making intelligent choices about the future direction of the project.

Thus, it was important that I figure out how to work in R. Over the week I managed to become slightly more adept in using it to find answers to the questions I have about the data, and have been working through my list one by one. In an effort to both track my progress on this project and eventually cross-check my methods, I'm keeping notes on everything I'm doing. This helps my learning process in terms of codifying my thoughts in terms of why I'm doing things in a particular way, as well as being able to ask another person to replicate my work as a way to verify my methods.

Beyond the R progress, I also managed to send interview requests to people at the Washington Post, Bloomberg, the New York Times, The Wall Street Journal, CIR, and fivethirtyeight.org. I heard back from Amanda Cox of the New York Times graphics department, which is really fantastic, as she has a degree in statistics and will be able to provide serious insight into the differences between statistics in journalism and academia. I also received response from Christopher Ingraham, a blogger from the Washington Post's Wonkblog who also does really great work with graphics and statistics. I hope to meet him in person for an interview. No word from any of the others thus far. My plan is to follow up with them mid-week. I also have two phone interviews for summer internships on Monday, which I'm very excited about- The Wall Street Journal graphics desk, and the Google Journalism Fellowship at Propublica.

I spent Friday and Saturday at the Open Data Day event at the World Bank, and it was an overall great learning experience. I had an opportunity to meet other journalists and listen to presentations on information that I'm mostly familiar with, but still got a lot

out of them as I'm still learn new things when going back over the basics. A new tool I learned to use is a program called open refine, which is a very powerful (and free) spreadsheet / data cleaning program. I also managed to speak at length with some Office of the Chief Technology Officer (OCTO) developers, who are heavily involved in making Washington DC's open data and maps available online. I discovered some great resources for open data and was able to talk to people with a deep knowledge of mapping, and who were happy to help answer my questions. Saturday was essentially a hackathon, where people converged in the morning to pool information, motivation, and knowledge to try and get a project proposal off the ground in a day. Very cool idea, but they had more information sessions upstairs, So I attended one on how to use API's. API stands for application program interface, and this session was probably one of the most enlightening hours that I can recall in recent memory. The instructor, a Sunlight Foundation developer, showed us how you can break a URL apart into separate pieces, then showed us how to build one back up. This rebuilt URL is how we received information from the Sunlight foundation API. Quite simple really, and fits in well with my existing schema of how to go about searching for and processing information. This knowledge opens so many doors for me in terms of programmatically finding data. I can't wait to find a use for it. The lecturer told a story of how one of his colleagues learned how to use the API in the morning, and by evening he had a twitter bot that automatically wished every senator and congressman a happy birthday via twitter, using information from the Sunlight foundation API. A somewhat trivial example, but built on a very powerful idea. I'm very excited to put this kind of knowledge to use at some point in the not too distant future.

Week 7 dispatch

(2.22.15-2.28.15)

This week was largely spent working in R doing exploring the language and working within the program to interview the congressional data. I have been refining my skills and taking notes as I go though the analysis in an effort to not only record my thought processes, but to help reinforce the underlying lessons that support them. I have started to make strides in my ability to explore the data visually, and have been making charts that act as guides in refining my lines of inspection.

The first is what's termed a set of small multiples, and shows a chart for every congressman elected between 1973 and 2010 with fever line showing how many laws they passed per session. It's quite a catalog and challenging to view on a screen, but still a good reference to be able to quickly pick out the outliers. Interestingly, Missouri had a representative named Leonore Sullivan who has the record for most laws passed per session in the timespan I have. She is a distinct outlier and I've ben trying to find out which laws she passed during that record session, in order to get a little more granularity and detail.

The second chart I made is in a similar format, but formatted for the states. It shows charts for each state and a fever line for how many laws were passed by its representatives. Previously, I ran a simple histogram that told me California has passed more laws over the 40-year stretch of my data than any other state by nearly a factor of 2. California 836 laws to its name and the next closest is New York at 480. While it is true

that California has always had a large number of delegates, many other states have had proportionately over half the number of delegates they do, but have not passed close to half the laws California has overall. It's interesting, and I decided to make a chart of the entire set of California congressman from the last 40 years just to see if there was anything of note. It appears that the representatives with the highest peaks are often short lived and the representatives with more longevity are less dramatic but steadier in their numbers over time.

If one were to try and normalize the sheer number of congressmen in California, and try to look at which states have the highest numbers of laws passed per congressman, Alaska comes out in the lead by a long shot. Alaska's lone representative, Don Young, as been an active force in legislating for over 22 years and has 74 laws to his name.

In addition, I've taken a look at the total amount of laws passed by the House over time and see that it is at it's lowest point in the last 40 years. I also see that the number of laws passed have gone through several cycles of high and low, but has continued a downward trend overall. I'm interested to see if these periods of peaks and troughs correlate to specific parties having majority, or maybe majority congress with opposition presidents. Perhaps there is a set of factors that give rise to more or less laws being passed overall.

These are some of the avenues I've been exploring this week, and am going to start using these findings to program a small narrative graphical presentation of the data in JavaScript. I'm looking forward to moving into this next phase of the project, and will be coordinating with Jeff Osborn to get help with the overall design of the piece.

In terms of research this week, I conducted an interview with Amanda Cox, a graphics editor at the New York Times. Cox has a master's degree in statistics and was very generous with her time and opinions. She made a number of fine points regarding how statistics are best used in the New York Times as well as her thoughts on some overall industry practices. It was a very enlightening conversation and her incisive comments will certainly lend gravity to my research that previously did not exist.

In class, we had a presentation from two Inspectors General representatives, Bridget Serchak from Department of Defense and Douglas Welty from the Department of State. They made a fine presentation, and as I looked at the results and recommendations of their department's investigations online, I noticed the documents were only available in PDF format. I asked them about the digital availability of results from their investigations via an API or in a database format and Serchak, responded that it was certainly a goal they were striving towards, but have not achieved it yet. She also mentioned that the Sunlight Foundation had been doing work to that end. I immediately did a search on the Internet and found the Sunlight Foundation's project. They have built a collection of scrapers that have compiled over 18,000 reports from the 65 Inspectors General, and have it available for download. This seems like an amazing resource and one that I plan on looking into, especially this summer if I manage to land an internship at a news organization.

Week 8 dispatch

(3.1.15-3.8.15)

The first part of the week I spent troubleshooting and refining data for the Death and Taxes poster. As the illustrators lay circles out and inspect the associated values, they sometimes come across peculiarities that emerge from the data. Sometimes the numbers seem odd. Sometimes the choices made by the R script seem odd. It is my job to peer into the script, decipher what it is doing to the data, and then look at the data to make sure the script is doing the correct thing. Basically, I double-check the ins and outs. And fix them if they need to be fixed. Overall the script is performing correctly which means I've managed to follow the instructions handed to me at the outset of the project and I have set the data up correctly. The few eccentricities that have emerged so far have ben resolved by researching the facts and attempting to discern the underlying meaning. For example, a large reduction in the State Department's budget from 2015 was a result of a negative budget line in the department of federal building oversight. We understood that to mean that several buildings were sold in 2015 and produced a sum that was larger than requested, resulting in a negative line item. The 2016 budget request was consequently much larger, and resulted in a rather dramatic percent change from 2015 to 2016 which was what initially threw up the red flag.

I also agreed to present my congressional project to the rest of my co-workers next week in order to get some critical feedback in terms of conception and design. I've moved from the analysis phase into the programming and design phase and have started working to bolster my knowledge of D3. I'm largely reading books and working through

tutorials in order to reinforce my tenuous understanding of the fundamentals that I may have glossed over when I was starting.

In the middle of the week, I flew to Atlanta and attended the NICAR conference. It was a marvelous event and I had the opportunity not only to learn but also to present some of my own original research this year. In December of 2014, I was awarded a Knight Prototype Grant for an idea to produce a wireless sensor array that would aggregate data for a live noise pollution map in downtown Columbia. My presentation was with two forerunners in the field of sensor journalism, John Keefe of WNYC, and Matt Waite, of University of Lincoln, Nebraska. It was a distinct pleasure to say the least.

On the first day of the conference, I was able to attend an all day python programming class that discussed refactoring and program design. The information was at the limits of my understanding, though I do believe it will inform my process going forward. The tenants of object oriented programming and modular design will be with me as I confront my next programming challenge. I'm eager to start working on it already, as if I just got a new toy for Christmas.

NICAR was also a great chance to meet potential employers, and I had several conversations that finished with them asking me to please get in touch after I graduate, which is a welcome thing to hear. It was also really nice to see old colleagues and re-connect with them both in terms of their progress in the field and to have the chance to discuss their success in depth. These lessons learned from conversations over dinner or late into the evening can be as informative as anything you see presented in a class or on a github repo.

In addition to having a great time and making new friends, I was able to secure verbal confirmations for interviews with several of the journalists I've been emailing with requests. Steven Rich of the Washington Post, Barbara Cohen of the New York Times, and Holly Hacker of the Dallas Morning News all seemed quite willing to speak with me. I'll set up the details next week and am happy, as these three in particular present a wide range of experience that should round out my set of interviews nicely.

Week 9 - Dispatch

(3.9.15 – 3.15.15)

This past week, I've been focusing on programming intently. Currently, the focus of my efforts is the congressional project I've pitched. I've been walked through the process at Graphicacy, which is to say, basically drawn up a modified client brief for the project. I wrote a script for a narrative walkthrough of the project, which centers on the legislative effectiveness score developed by two political scientists. My thought is to use a circle to show each representative, and that the diameter of the circle will indicate their effectiveness. The color of each circle will be determined by party (red/blue) and will be arranged in a half circle, to approximate their seating in the house. This will also have the added benefit of seeing an easy visual separation by party and creating a de-facto indicator gauge that shows which party has more members in the house. I have a slider that allows the user to choose the session of congress to display, and I'm also planning on small buttons that will allow sorting the circles by various data points such as state, longevity of service, and gender.

The lead developer, Reed Spool, walked me through the trigonometric calculations that were needed to plot an even number of dots in a hemi-spherical pattern, 8 rows deep. It was quite an eye opener of a math lesson. While I was moderately familiar with the basic formulas, this project is showing me why the formulas for sin, cosine and tangent are awesome. More than that, I'm starting to see how the creation of unique graphics can depend on a wide berth of mathematic knowledge. This project is a great platform to learn visual tactics beyond the pre-formed code snippets that exist and to employ other spheres of knowledge into the problem I'm working on.

Beyond that, in more of a nuts and bolts realm, I'm looking at the best ways to separate the data processing from the actual visualization part of the code. Currently I'm playing with Node.JS, which is a JavaScript wrapper for C, and it will allow me to run .js files independently of a browser. Since I'm currently using JavaScript to process my data beforehand, it makes a certain amount of sense to keep it all in the same language. Though, truth be told, I very well might get fed up with trying to learn yet another set of rules for another new language, and just use Python, which I've taken classes in and feel reasonably comfortable using.

On the research front, I've rounded out my last two scheduled interviews with Holly Hacker of the Dallas Morning News and Sarah Cohen of the New York Times. They presented me with great information and both recommended that I get in touch with Jen LaFluer, an editor at CIR. I reached out to her previously and heard nothing back, but upon the urging of these recent conversations, I'll make another attempt. I feel that a lot of the interviews are starting to echo one another, which may be a sign that I've reached a bit of a saturation point. Cohen did bring up a really interesting point of conversation that

piqued my interest, which were her recent forays into machine learning. Machine learning is basically a way to programmatically analyze unstructured data through statistical methods.

An example she noted was a presentation she was a part of at the recent NICAR conference. She presented with Janet Roberts, A Data Editor at Reuters, and Roberts shared her attempts to examine an elite group of lawyers that dominate the Supreme Court docket. She basically took all the text from thousands of court documents and tried to find similarities in them in order to give her an idea of the "shape" of all the data and where the similarities were. She had a good bit of success in this method, which she confirmed by double-checking a lot of the work by hand. But overall, the machine learning process allowed her a categorization of her data (roughly 10,000 court documents) that was easily modified to re-process and refine to get better results. I'm not sure if this is a way in which I want to extend my project at this time, but it certainly is an interesting and developing field in terms of how statistics are used to find stories.

Week 10 dispatch

(3.16.15 – 3.23.15)

In the last week, I have been focusing on my congressional visualization. After Reed introduced me to Node.JS, and the ability to work in JavaScript from the command line, I've excited to put it to use and have been running through tutorials all over the Internet. They paid off and I've finally made good headway in a couple distinct areas of the presentation.

First, I managed to break out the pre-processing of the data in Node.js, and am able to complete the current and any additional data manipulation before the project even hits a browser. This is a good practice not just for this project but for the future as well. I am now able to import, loop through the data, add a seating chart for each of the 8,000 representatives across 40 years of time, and seal it all back up in a comma separated value document ready for the script that runs in the browser.

The second portion I figured out this week is how to align the circles in a perfect radial arc no matter how many representatives there are. The total number varies a bit from year to year, and hard coding the number of reps per row was giving me problems. The static number was leaving a few data points off in some years, which was not acceptable. In the years where the total number of reps fell shy of the hard coded numbers, the circles in the last row never reached the bottom edge, and the design looked haphazard. So it was a problem that had to be fixed. On Friday, I figured out how to construct a new object and populate it with values from the total number of reps from each year. Then, divided that number by eight (one for each row) and adjusted the row number by percentages of each total to distribute the circles evenly along each row. This matrix is then applied to the trigonometry that positions each circle in the radial layout, and now it shows eight perfectly full rows despite the shifting numbers of representatives.

My next task will be to change the way the circles initially appear on the page, with the democrats to one side and republicans to the other. This not only mirrors how they actually sit in the house, but will also create a de-facto visual gauge that allows the user a quick estimate of which party has control of the house. Following that, I'll create buttons that allow the user to sort the group of representatives by various categories.

The seminar this week took us to the Supreme Court, where we toured the courtroom and took a meeting with Patricia McCabe Estrada, the deputy public information officer. She gave us an overview of how her office works and how much things have changed in the almost two and a half decades she has worked there. I was most interested in how she described the intimacy of the courtroom, noting that the lectern for lawyers arguing is incredibly close to the justices. She related one chief justice saying that if he were to reach out from where he sat and the lawyer did as well, they would just be able to touch each other.

I also spent the weekend crafting the first draft of my research paper. It's coming along well, but is not quite done yet. I have one last interview to conduct with Andrew Flowers, the Quantitative Editor for fivethirtyeight.com. I'm hoping he will be able to provide some insight into their particular model for journalism and what has created the underlying appetite for their level of analysis. This last interview should round out the set I have to draw from quite well and provide me with the finishing touches I need to complete my research paper.

Week 11 dispatch

(3.23.15 – 3.29.15)

This week I've spent the vast majority of my time working with the code of my congressional project making tweaks, modifications, and adjustments. I've been working to add tooltips to the individual circles so the information they represent is accessible. I've also been really stretching my brain out to grasp, on a very granular level, the way the data is being processed. I find it becomes much easier to manipulate the existing

program when I have a clear understanding of the abstractions. This is a pretty obvious statement, but when juggling so many variables in your head and working through the path the data takes to be processed correctly, the attitude of if "it ain't broke, dont fix it" persists. This can lead to an atrophy of the holistic program comprehension, but there are many ways to combat this withering away of details. One is to code with clear variable names. The use of variable names that detail exactly what is going into them is critical when programming, not just for others that may encounter it to help or refactor, but for ones future self. Another tactic is to indent code with purpose and rigorous consistency. Indention shows where action is taking place and what parts are subsumed by others at a glance, similar to an outline does with rough thoughts. This is simple technique that is for readability alone, as the computer does not (in JavaScript) need or care about space between statements. I've been trying to incorporate these elements into my work when I write code, and am relieved to see it when others do the same.

One of the difficulties working with talented people is that sometimes their technique far surpasses your own, and the digestion of the ideas they offer takes time. The lead developer here at Graphicacy, Reed Spool, has been helping me with my project and the code he writes is very clean and concise. As the code comes out of his mind, the narrative that goes into naming variables likely have good purchase in his imagination. I've found it takes me a good amount of time to comprehend his structures without mental hiccups. And comprehension is critical when attempting to adjust and manipulate the inner workings. It is laborious. And I've seen my coffee consumption shoot through the roof. With time and effort, I find it comes more easily, but there are no shortcuts. And I am grateful I have good examples to follow. Developing poor habits is something I've

desperately been trying to avoid in the pursuit of this internship. Learning a skill well often results in a slower uptake at the outset, but I've found it pays off and is much easier than correcting bad habits.

In other news, I reconvened with Monique, the researcher at DARE I was working with to try and find a project to collaborate on. She mentioned that a close friend of hers is in charge of an AIDS education fund (the Ryan White fund) and while giving her friend a tour of our shared workspace, she saw and loved the posters we have up in the hallway. Monique said her friend was very interested in working with Graphicacy to showcase the data their organization has accrued, and it seems like a fantastic opportunity to create some meaningful work. Very cool opportunities on the horizon.

I also finished the first draft of my research paper this week and sent it to Scott for his perusal. I think it reads well, but has room for improvement. I also had the opportunity to conduct an interview with Andrew Flower, the quantative editor at fivethirtyeight.com, and his input will dovetail very nicely with the existing structure of the paper. Looking forward to threading it in and getting that part of my work wrapped up.

Week 12 dispatch

(3.30.15- 4.6.15)

I spent this week immersed in the code for the Congressional Effectiveness project. Over the weekend I managed to get all the internal wiring done so that I have a set of buttons that allows for sorting the circles along data points. Currently the buttons represent sorting by: length of tenure, effectiveness score, and by party. I've been

parsing the code to implement the changes needed to complete the project, so even though the lead developer laid out some of the initial frameworks I was unfamiliar with in order to get me started, I've been going through the entire program with a fine toothed digital comb in order to understand the passage of information from start to finish. In expanding the features of the program, I've had to understand the new features, and have also been researching and borrowing from older projects in order to get it all working. It's always a little startling when I've been working on a problem for a long time, and finally manage to hit the right combination of functions, variables and statements. Loading it up in the browser for the millionth time, except this time, it comes to life. So, very much worth the time put in on it. But the luster fades quickly and I start to assess what needs to be done next to make the project complete. Though, in order to try and cement the learning, I find myself rapidly jotting down notes about what I've done right and what I did wrong to document my process. Inevitably there will be a next time, and I hope to have an easier time if I can learn from my past accomplishments and turn to my notes as a way to solve whatever problem I'm currently stuck on.

I also revised my graduate project article as per Scott's suggestion, which certainly is a more accessible reading experience now. Writing about complex things in a simple way is really challenging (actually one of the points I've made in my article) and tried to incorporate a lot of examples and metaphors in my work. I think it came together pretty well, and am happy with the product.

We visited the Washington Post as a class this week, and listened to a couple of the younger members on the investigative team speak about their jobs and how they work. I was particularly familiar with Steven Rich's work, as I had covered a lot of it in

preparation for my interview with him for my research. At one point, they mentioned needing to become experts on a subject within a month for whatever they were working on, and I asked how they managed to accomplish such a task, and how they knew when that had reached a level of satisfactory research. Both of them responded in terms of ordering books, reading a lot online, and trying to become knowledgeable enough to have a conversation about the topic with an expert. This sounds like a daunting task and depending on how broad a topic one might be trying to cover, seems like it might very well take much longer than a month.

Chapter Three: Self Evaluation

My progress has been steady over the past four months. Many times, I was learning when I didn't realize I was. Only upon the sudden comprehension of what yesterday was an opaque idea did the time spent focused on learning become evident. I know that I have barely dipped my toe in the water of a large ocean. But what was at first daunting and unwieldy has become a somewhat manageable project full of small problems. And that seems to be how I measure my progress now: the ability to solve small problems.

The lead developer at Graphicacy, Reed Spool, has said many times that programming is so hard at first because you're trying to overcome not only a language barrier but also cultural and logical barriers. He says often that a programmatic mindset is one that allows you to see a problem, analyze it, and break it apart into its simplest pieces. Then, the job is to attack those pieces. Those have been my small problems.

Another friend of mine has described learning to program by working on a project as "trying to review a movie in Russian… while you're learning to speak Russian." This gets at having the desire to express an idea while being governed by the stutter step of running up against your limits of knowledge in the midst of trying to express that idea. The process requires patience and optimism. And a stubborn determinism.

My work at Graphicacy has been quite fruitful in terms of my improvement. The company is bound by the clients they serve, not by a news cycle, so it only has work when contracted. By timing, I came during a trough in the natural ebb and flow of their work cycle. They were finishing up some big projects and not yet starting others. I made

efforts to engage a couple opportunities on a quid pro quo basis, but these projects stalled despite best efforts all around. In an effort to contribute and learn, I volunteered to process the data for an annual project Graphicacy has called "Death and Taxes." It is a visualization of the discretionary federal budget, and I worked with some existing scripts in the statistical package R.

Working on this project had many unexpected benefits, one of which was growing much more familiar with how out government spends money. It also led me to consider the incredibly enormous amounts that constitute millions, billions, and trillions. The sheer largeness of these numbers has given me a great deal to think about in terms of how to present financial information in the future. In addition, I was able to work in concert with the designers at Graphicacy, improve on an existing workflow, and help solve problems in a collaborative fashion. The end product is a fine poster that is informational and visually fascinating.

In the absence of direct client work, I began an independent project looking for a way to measure our elected representatives, similar to how baseball players are evaluated by a batting average or an ERA. In searching for a statistical measure to work with, I found it had already been done by two political scientists, Craig Volden of the University of Virginia and Alan E. Wiseman of Vanderbilt University. Their project worked with information compiled directly from the Library of Congress website and tracks the five stages of every bill from proposal to law by every congressman elected between 1973 and 2011. This set of information helps to compile the "Legislative Effectiveness Score" which is then weighted by their party, tenure, committee appointment, and a number of

other dimensions housed in the data. Best of all, the data was housed on the website free and available for download in excel spreadsheet format.

This data became the backbone of my Congressional Effectiveness project, which was a test piece to work on solving problems using JavaScript and the data visualization library D3. In the process of iterating over the project and going through a more extensive planning phase than I had encountered before, I found my ideas clarified in a way they had not been in my previous projects.

The process of programming and having a lead developer to answer questions as well as inspire by example was a great experience. I think it made me work harder to answer my own questions because I knew that I would be able to find the answer even if I was unable to solve it myself. Which is to say, frustration was ever present, but hope was never lost.

The Congressional Effectiveness project had many conceptual dead ends that I ran down in trying to engineer a good program. This is a good way for me to learn. I feel it's been a solid experience that has really pushed my abilities far beyond what I would have been able to do on my own in such a short amount of time. I also think my conception of how to program has been expanded, and I have a set of rigorous concepts that I will carry with me to make my code more comprehensible to others and myself. I have always heard coding was a social process, and now I'm starting to understand why.

The improvement I have made probably far outweighs the concrete work I have committed in code or in published work for Graphicacy. The majority of my effort was spent researching to gain a holistic understanding of how to code efficiently in the

absence of client work from the company. This has resulted in a different type of learning than I had anticipated, but nevertheless my skills are now far beyond what I expected.

Letter of evaluation from Jeff Osborn, Creative Director of Graphicacy

TO:
Barbara Cochran
Missouri School of Journalism
Curtis B. Hurley Chair
In Public Affairs Journalism
Washington Program Director

FROM:
Jeffrey Osborn
Graphicacy
Creative Director

RE:
Travis Hartman
Graphicacy winter 2016 intern

Dear Barbara,

Per your evaluation request, a quick recap of Travis Hartman's tenure as a Graphicacy intern:

Duties
Our goal for Travis was to find the sweet spot where, through integration with Graphicacy's work flow, he could be learning and contributing at the same time. Our understanding was that his background had afforded him a "visual eye" and a journalist's knack for storytelling –and– his current interests where to expand his capabilities in coding, specifically in ways that would help him be able to create and strategically apply interactive information graphics relevant to the emerging field of data journalism.

His duties were specifically to bring as many of his skill sets to bear on the work we were doing (where possible) while also developing his own self-directed interactive project paired with a blog post about the creative journey.

Accomplishments
Having Travis join us with thoughts/feedback on a daily basis was helpful overall. More specifically, he was able to contribute in the following ways:

1) **Death & Taxes poster** – Graphicacy annually creates a data-rich poster visualizing the U.S. discretionary budget allocations. Travis was able to help us

locate the most up to date data, do basic visualizations of those numbers in the statistical software program R, and then join us in the iterative process of final production and fact checking. He also helped in brainstorming ideas about features for our companion online interactive as well as a list of "things that we learned that were interesting" for the product launch mention we were extended by the online news source Vox.

2) **Legislative effectiveness interactive** – this was Travis' self-generated project. In tandem with the Graphicacy team he was able to work through research, data collection, concept development, and development iteration phases to create an engaging tool that would allow viewers to custom tailor an exploration of Congresspeople's productivity. This project allowed Travis to work with Graphicacy staff on both the design and development side in pursuit of good storytelling with data (both in form and function.)
3) **Center for American Progress** – Travis provided helpful feedback and support in the "post mortem" evaluation of a large motion graphic and interactives project as well as for a newly launched web design project for CAP.
4) Travis participated in a **weekly development team coding exercises** and review meeting intended to strengthen and enlarge the capabilities of this side of Graphicacy's business.

Overall Travis was an exemplary intern and a pleasure to work with – mature, intelligent, a problem solver, passionate and interested in what we do, eager to learn, a team player, someone who arrived on time and ready to go to work, of good humor, and someone willing to join and enlarge conversations about our concerns.

Travis represented himself well both for the University of Missouri and Graphicacy, and I'm confident that he will be in demand and able to contribute immediately to any future employer engaged in the creation of information graphics/data visualization/data journalism. Will be happy to offer recommendations where needed.

Best,

Jeff

Chapter Four: Abundant Physical Evidence

Death and Taxes

The Graphicacy employee that processed the Death and Taxes data last year had moved on, so I found a short set of instructions she wrote to guide me. It served as a good template, but there were many things that were left unexplained. It took a lot of time to decipher how the file structures were set up and how the more than 35 R scripts were organized in order to generate the 84 PDFs that comprise the circles in the poster.

I downloaded the 12 government spreadsheets that comprised the different aspects of the 2016 federal budget proposal and updated all the scripts with current years. I managed to improve the workflow for the designers by formatting the circle PDFs, the text labels describing the department and its budget, and removing unnecessary background elements to facilitate importing into Adobe Illustrator. I was tasked with troubleshooting the various problems that came up concerning the data throughout the length of the project, such as negative numbers in the budget, incongruities of the pre-existing script, and differences between this year's budget and last.

I amended the existing instructions with pointers on how to download and prep the data for the R scripts, noting the hurdles I overcame and the differences that occurred from last year to this year. I also made sure the folder structure I created was self-explanatory, should the person working with the data next year need to compare. Below is an excerpt of the notes I kept as well as sample PDFs from last year and this year. These reflect both the changes I made and obstacles I overcame.

NOTES:

These are notes I kept to record my progress during my participation in the Death and Taxes project. The following is primarily code from the statistical scripting language R, and details the beginning of the process from downloading, formatting, and cleaning the data. The end of the notes section provides evidence of my work with the plotting script. I changed the way the circles and accompanying text were produced in effort to aid layout of the final poster.)

The President is scheduled to release his proposed fiscal 2016 budget on Feb 2.

found instructions-
looks like a good overview-
checked all the links-
defense budget link is 404
seems to be a new place to find this info:
http://comptroller.defense.gov/budgetmaterials.aspx

- Get the following files:

- Overview Defense Budget PDF:
looks like the 2014 is here:
http://comptroller.defense.gov/Portals/45/Documents/defbudget/fy2014/FY2014_Budget_Request_Overview_Book.pdf
this table is inside the overview:
Get A-8 DoD Base Budget by Military Department and Appropriation Title (this provides the military budget change over years, as well as the military budget across four major departments, which are not included in the files we use for plotting detailed activities. See the picture at the end of this section as an example.)

- Get Detailed Budget Documents with the following names:
m1,r1,p1,o1

--looks like all R scripts are in the folder for the previous years.
r scripts output the circles into PDFs

could write a quick python script to download all the selected budget files?

- could be problematic as the locations of the files/etc can migrate, but would be nice if everything stayed same same.

---

ran script.R in non-defense folder

---

threw an error on line 59-

```
> #Make sure the row orders in "budget" are the same as those in "position"
> # if not, change "position.csv" to match them in "budget"
> budget[c("x", .... [TRUNCATED]
> budget$x <- position$x
 Show Traceback

 Rerun with Debug
 Error in `$<-.data.frame`(`*tmp*`, "x", value = c(17.2, 16.5, 17.7, 20.5,  :
 Replacement has 68 rows, data has 70
```

looking to match position to budget as per instructions-
it looks like the recovery accountability and transparency board + allowances are the differences between 2014 and 2015.
simply had to add them with their dept number and 0,0 and it worked and plotted the circles and output the PDF. looks good so far.
on to the next.

---

ran Employemnt and Training Adminstration.R in non-defense folder

---

threw an error:
Warning message:
In sqrt(budget1$X2015) : NaNs produced

but still plotted three nice circles.

---

ran IRS.R in non-defense folder

Warning messages:
1: In sqrt(budget1$X2015) : NaNs produced
2: In sqrt(agency15_1$X2015) : NaNs produced

in spot checking against last years PDF's from the folder on the sever, the circles and numbers above them appear identical.
will keep checking for accuracy, but it seems ok to ignore warnings.

_____
ran 14.R, Other Independent Agency.R in non-defense folder
---

no errors

_____
ran 21_Aviation.R in non-defense

Warning messages:
1: In sqrt(budget1$X2015) : NaNs produced
2: In sqrt(agency21_1$X2015) : NaNs produced

_____
ran 18_19_Based on Subfunc.R in non-defense folder

Warning messages:
1: In sqrt(budget1$X2015) : NaNs produced
2: In sqrt(agency19_2$X2015) : NaNs produced

_____
ran 202.R, 25.R, 26.R, 29.R, Bureau_Level_Script.R in non-defense folder
---

Warning message:
In sqrt(budget$X2015) : NaNs produced

_____
ran 26.R in non-defense folder

Error in aggregate.data.frame(lhs, mf[-1L], FUN = FUN, ...) :
no rows to aggregate
In addition: Warning message:
In sqrt(budget1$X2015) : NaNs produced

(after getting this error, I ran bureau_level_script.R and Other Independent Agency.R, then ran it again, and it worked. I'll inspect to make sure it's the same as the previous year. -- Matches up perfectly. Not sure, perhaps you need to run those listed above before 26.R)

got through all non-defense.

DEFENSE:

formatted m1, o1, p1, and r1 as per instructions. many sheets per each- assuming that it's the overall docket- removed thousands separators, deleted first and last (if needed) rows. filled in all blank spaces associated with "classified" projects.

running through the first script (Defense_wide.R), it throws an error-
line 11 of defense_wide is

> operation <- operation[c("Account.Title","Organization","FY.2014.Base.Request.with.CR.Adj.","FY.2015.Base")]

Error in `[.data.frame`(operation, c("Account.Title", "Organization", : undefined columns selected

looks like the columns are named FY.2014.Base.Request.with.CR.Adj. (C.R. for continuing resolution?) in the CSVs from the previous iteration- not sure if they changed the name or I've grabbed the wrong files somehow?

there is no header for any column named FY.2014.Base.Request.with.CR.Adj. -- looking into the produced spreadsheet to see how this was handled.

******
OK- so found You You's (the staffer who did the project last year) .xls files and had them to match against- it appears the titles change each year, I had the right files for 2015 budget, but the column titles had changed- thus the R scripts had changed as well, and when i found them in You You's folder, i had them to match to- simply had to run the new script with the files i had all along and it went swimmingly. Well, there was an asterisk after FY.2015.BASE in the xls and csv which translated to a period in the import in R. this period was throwing off the script. once i deleted the * from the csv, the script ran fine. awaay we go.
****

ran defense.procurement.R - no problems
ran defense.operation.R - no problems
ran defense.wide.R - no problems

ran defense.RDTE.R -
Warning messages:
1: Removed 1 rows containing missing values (geom_point).
2: Removed 1 rows containing missing values (geom_text).
3: Removed 1 rows containing missing values (geom_text).
4: Removed 1 rows containing missing values (geom_text).

ran defense.Personnel.R -
Warning messages:

```
1: In sqrt(army$FY.2015.Base) : NaNs produced
2: In sqrt(af$FY.2015.Base) : NaNs produced
3: In sqrt(navy$FY.2015.Base) : NaNs produced
```

on to manual completion:

should be able to write a python script to scrape from .html pages?
might be possible.
make set of depts. source that set when searching for numbers.
boom.
done.

to reformat the script to output no grid background and colored text, i have modified the last part of the script in the output as well as the spacing between the circles. the resulting PDF is larger, but easier to both grab and import into illustrator, plus easier to format with the different colors in the info.

```
# assign positions
dw[c("x","y")] <- NA
# changed the magic number from 3 to 5 for spacing.
dw$x <- seq(1,length(dw$FY.2015.Base)*5,by=5)
dw$y <- 1

# draw circles
pdf("Defense-wide.pdf",family="helvetica")
ggplot(dw) +
#commented out the xlim to allow the grade to expand as needed for the extra
spacing in the circle plotting
#xlim(-8,55)+
ylim(-8,10)+
#instered a new theme here that nullifies the formatting and drops the circles onto
a blank slate.
theme_bw()+
theme(axis.line=element_blank(),
axis.text.x=element_blank(),
axis.text.y=element_blank(),
axis.ticks=element_blank(),
axis.title.x=element_blank(),
axis.title.y=element_blank(),
legend.position="none",
panel.background=element_blank(),
panel.border=element_blank(),
panel.grid.major=element_blank(),
panel.grid.minor=element_blank(),
plot.background=element_blank())+
coord_fixed(ratio=1)+
```

```
    geom_point(shape=1,aes(size=radius,x=x,y=y))+
    scale_size_continuous(range=c(min(dw$radius)/300,max(dw$radi  us)/300))+
    #color = "colorname" is how to set the color; spacing on the  labels /
    calculations is the x and y coordinates
    geom_text(data=dw,aes(label=Organization,x=x,y=y+5),size=1,
    color="blue")+
    geom_text(data=dw,aes(label=paste(FY.2015.Base/1000000,"Bill
    ion"),x=x,y=y+4),size=1,color="green")+
    geom_text(data=dw,aes(label=f2014t2015,x=x,y=y+3),size=1,col  or="grey")
    dev.off()
```

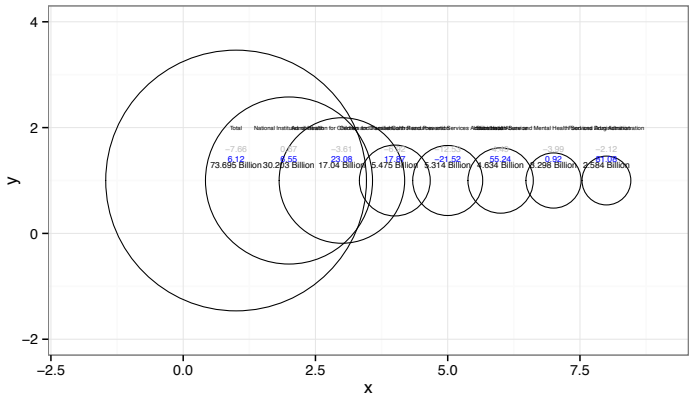
how to alternate labels to avoid overlapping?
set variable to vary 5, -5
based on odd/even in the array of

```
    # assign positions
    dw[c("x","y")] <- NA
    dw$x <- seq(1,length(dw$FY.2015.Base)*5,by=5)
    dw$y <- 1

    # draw circles
    pdf("Defense-wide.pdf",family="Helvetica")
    for (i in 1:2) {print(alternate[i])}
    ggplot(dw) +
    #xlim(-8,55)+
    ylim(-8,10)+
    _bw()+
    theme(axis.line=element_blank(),
  axis.text.x=element_blank(),
  axis.text.y=element_blank(),
  axis.ticks=element_blank(),
  axis.title.x=element_blank(),
  axis.title.y=element_blank(),
  legend.position="none",
  panel.background=element_blank(),
  panel.border=element_blank(),
  panel.grid.major=element_blank(),
  panel.grid.minor=element_blank(),
  plot.background=element_blank())+
    coord_fixed(ratio=1)+
    geom_point(shape=1,aes(size=radius,x=x,y=y))+
    scale_size_continuous(range=c(min(dw$radius)/300,max(dw$radi  us)/300))+
    #for (g in 1:2) {print (alternate[g])}+
    geom_text(data=dw,aes(label=Organization,x=x,y=(alternate[i]
    geom_text(data=dw,aes(label=paste(FY.2015.Base/1000000,"Bill
    ion"),x=x,y=y+4),size=1,color="green")+
```

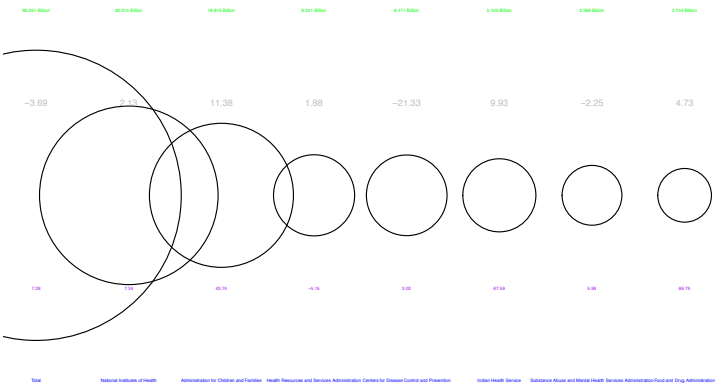
```
geom_text(data=dw,aes(label=f2014t2015,x=x,y=y+3),size=1,col   or="grey")
```

additional.csv:
kind of a nightmare, assembling it by hand, grabbing info from the pdfs.
the dept of veterans have changed the name of ambulatory care to outpatient care.
so I changed it in the spreadsheet as well.
mostly the fields were already added up but occasionally (as in the EPA) you need to add them up yourself from every occurrence in the section.

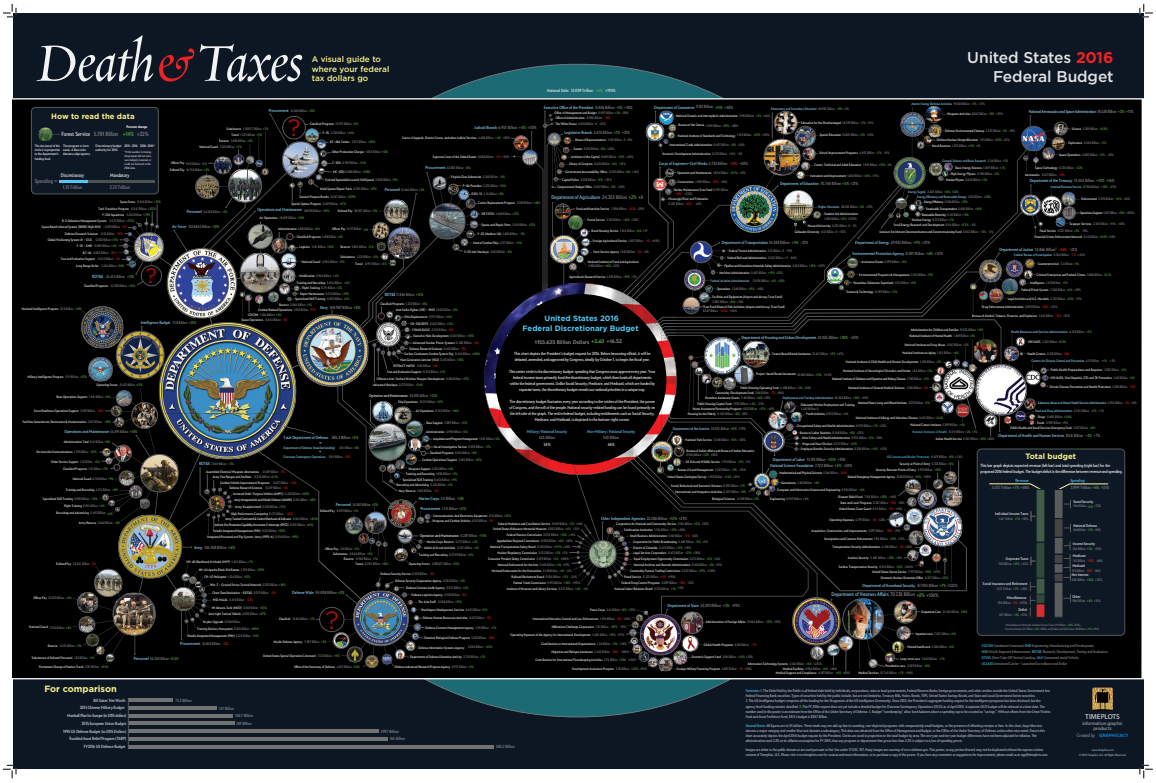
Output Improvements to the Death and Taxes production process:
This is the output from last year showing the 2015 National Institute of Health's budget.



Below is the cleaner version I created for the 2016 NIH budget.



The Final Product, Death and Taxes 2016



Congressional Effectiveness Project

This project was a main focus of my time while at Graphicacy. It provided me with a reason to use JavaScript and D3.js and work on the same kinds of problems found both in newsrooms and Graphicacy through the medium of data visualization. This project allowed me to interact with the entire staff in different facets by tapping into the design expertise of the creative director and art director, as well as the coding abilities of the lead developer. In effectively designing a project for myself, I was able to participate in all phases of the design and development through the internship. This turned out to be a very rewarding project in terms of skills advancement but also in terms of learning how

to develop and explain the design through a script and design prospectus before starting to code.

Once coding commenced, I worked more exclusively with the lead developer to architect the programming techniques to efficiently produce the graphics through legible code. I also developed my ability to decipher the many examples on the Internet from which code snippets can be lifted from. And overall, simply going through the design to development process and spending a healthy amount of time researching the concepts the lead developer laid out for me at each stage was invaluable. This process I realize made for an inefficient development cycle but in my opinion created a more resonant learning experience.

Initial pitch from an ideas memo on January 12, 2015:

-Congressional effectiveness ratings-
Is there a metric for how effective they are? Sort of like an ERA in baseball?
Turns out there is- two political scientists have devised a system to determine effectiveness based on how many bills each congressman is able to get passed into congress. They scraped the data from THOMAS, the website that houses all the day to day info on the congress- then compiled it as a basis for their metrics.
There are caveats, but it's a good start and a place to build from I think.
http://www.thelawmakers.org/#/method explains the details.
They also have data available for every congressperson since 1970 for download here:
http://www.thelawmakers.org/#/downloads
Which is great.
They do detail their media coverage, mostly which piggybacks on an earlier, related look at gender and how women wield power in the senate. I'm going to make sure they haven't missed anything and that no one else has done anything with this data in any kind of a big way

After getting a green light from Jeff Osborn, the creative director, I started looking into the data in R, cleaning the data, and building a list of questions to explore.

The following list of questions are from my Congressional Effectiveness notes. They give an insight into my process of first formulating questions and then attempting to interview the data through analysis. Below the questions, I provide some of the R script that I used to clean the data and explore relationships between the variables. Ideally, the relationships are what I would want to eventually display through the visualization. The charts following the code are an initial attempt at visualizing the data while looking for outliers and potential areas of interest.

Initial Questions

which state has passed the most laws? CA by a long shot- but why?

which congressman? - Missouri congressman

what year had the most laws passesd?

what was the percentage of rep/dem laws?

how disadvantaged is the minority party in passing laws? (see majorty party member column)

has congress been passing more or less laws overall?

do congressmen get better or worse with the longer they are in office? (seniority column)

male female divide?

does the size of your delegation from your home state affect your abilities?

member of committee? speaker? etc?

percentage vote received to enter this congress? does that affect how many laws passed?

Potential correlation to look at:

between laws passed and re-election?

between laws passed and pct vote?

between higher seniority and more laws passed?

between more bills introduced and more laws passed?

between majority party and laws passed by party members??

R Code:

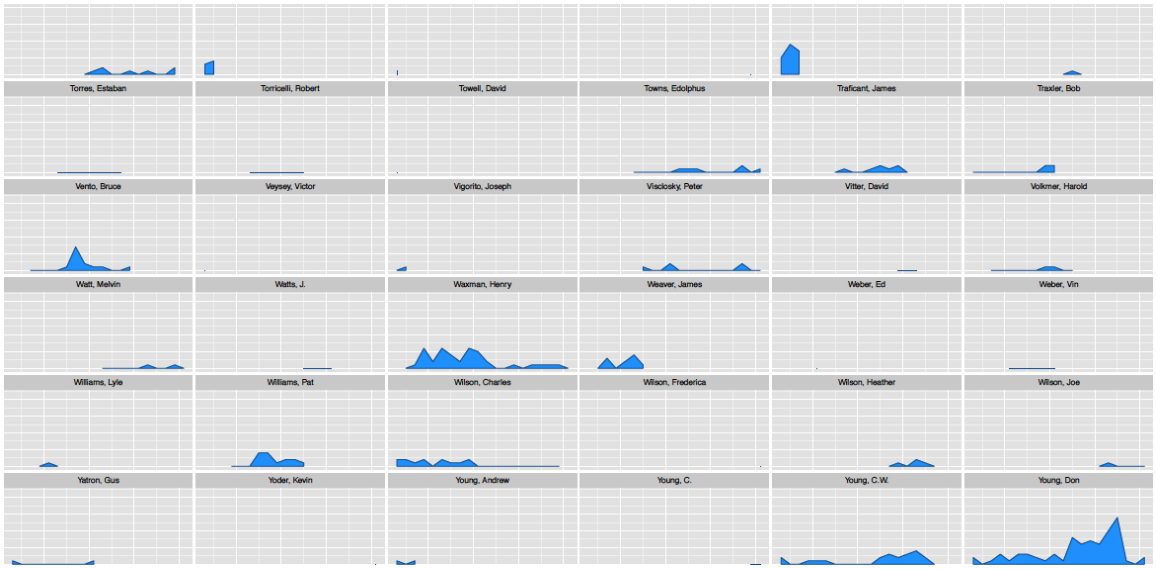
:

```
#find which st_name row has "Fl"
which(congress$st_name == "Fl")

#see the whole row
congress[5320,]

#replace the st_name data point with the correct one
congress$st_name[5320] <- "FL"


#to count NAs in congress$dem column:
        sum(is.na(df$col))
#returns 90-
#what's up with this?
#to find locaiton of NA's
which(is.na(congress$dem))
[1]    5   25  127  426  443  450  475  573  875  887  980      1023 1318 1331 1415
1459 1464
[18] 1761 1847 1904 2171 2206 2218 2293 2338 2609 2650 2662   2695 2743
2788 2807 3054 3106
[35] 3138 3231 3247 3505 3552 3582 3668 3670 3686 4001 4037   4073 4122
4136 4289 4447 4565
[52] 4742 4788 4862 4892 5017 5036 5193 5237 5304 5406 5464   5635 5685
5760 5856 5913 6078
[69] 6124 6200 6231 6296 6351 6521 6648 6678 6717 6751 6806   6975 7159
7194 7249 7261 7419
```

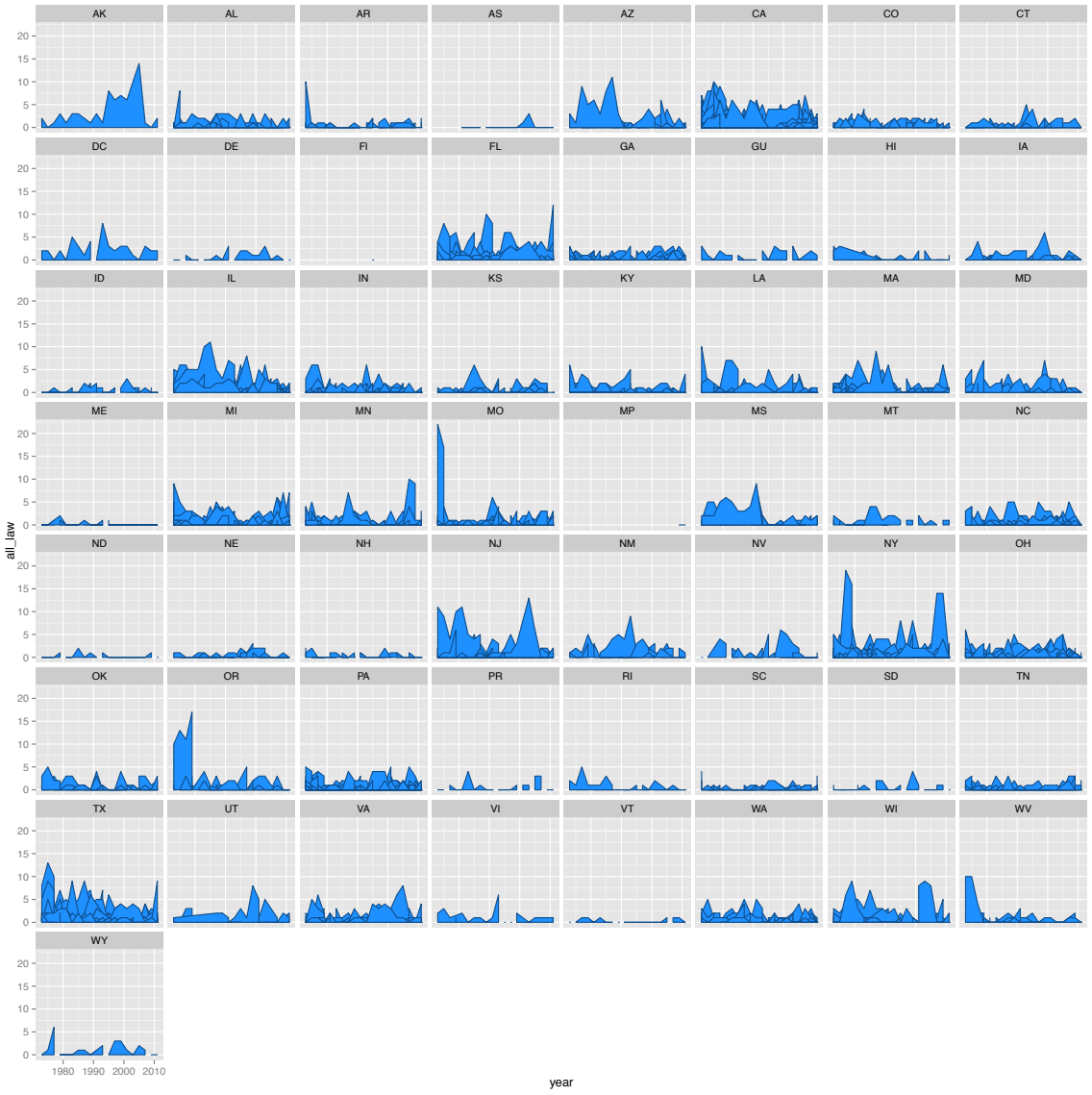
```
[86] 7602 7641 7699 7709 7871

#researched the party of each individual with an NA. replaced in excel.
#for those independents, i put 3.




############ do congressmen become more effective the longer  they are in
congress?
lm.senior.les <- lm(formula = les ~ seniority, data =            congress)
#seems to be a high significance to basically no       correlation. (9%)
summary(lm(formula = les ~ seniority, data = congress))
#see plot
plot(les ~ seniority, data = congress)
abline(lm.senior.laws, col = "red")
#also
crPlots(lm.senior.les)
#again, uptick after about 15 years. slightly better    correlation, 14% overall - if
seniorty figures into the les,   will that make the corrlation wonky?

############### which state passes the most laws??


# plot to see which states have most laws passed (MO is      highest?? NOPE- a
congressman from Missouri passed 22 laws  in one term.)
ggplot(laws.passed.in.congress, aes(x=st_name, y=all_law)) +
  geom_point(shape=1)      # Use hollow circles

#here we are:
state.laws <- (aggregate(congress$all_law,   list(state=congress$st_name), sum))

plot(state.laws)


#also: to plot small multiples for states with each      congressman having a line
  ggplot(congress, aes(year, all_law,group = thomas_name, fill=thomas_name))+
geom_ribbon(aes(ymin=0, ymax=all_law),   fill= "dodgerblue", colour =
"dodgerblue4")+facet_wrap(~st_name,ncol=8)
```

Plots from R scripts exploring Congressional Data:

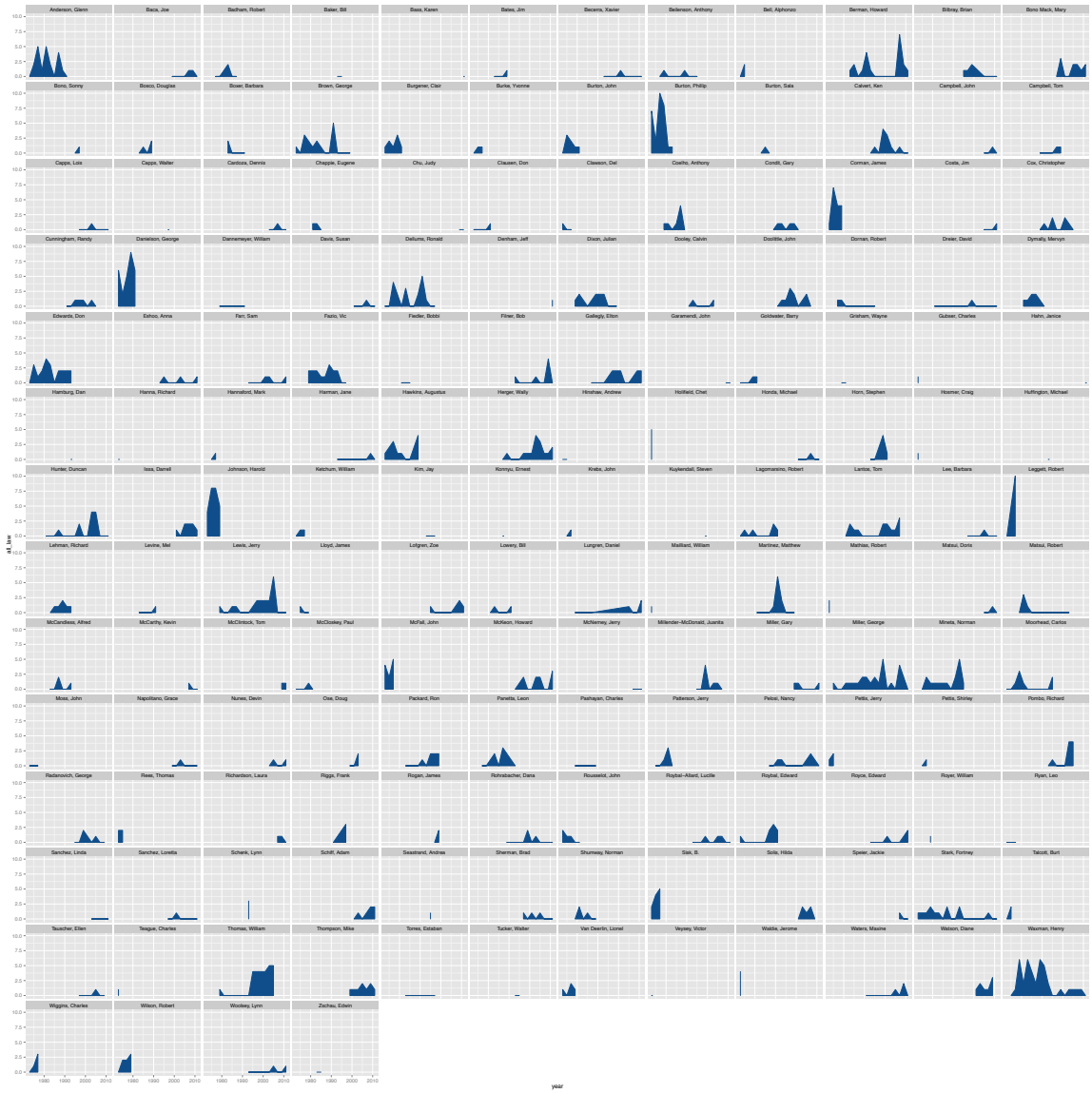
This is a detail from a visualization of every representative in the data set. Note Alaska Rep. Don Young's 23 years of service in the lower right hand portion of chart.



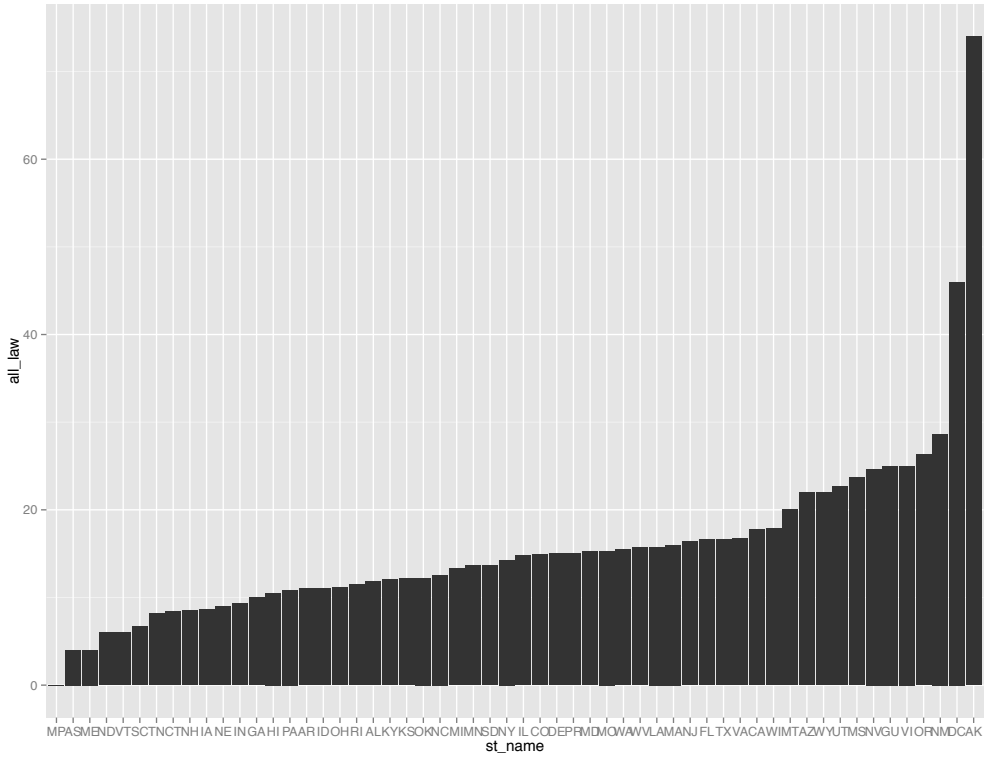
This shows the number of laws passed by state from ’73-‘11. California is at the top with over 800.



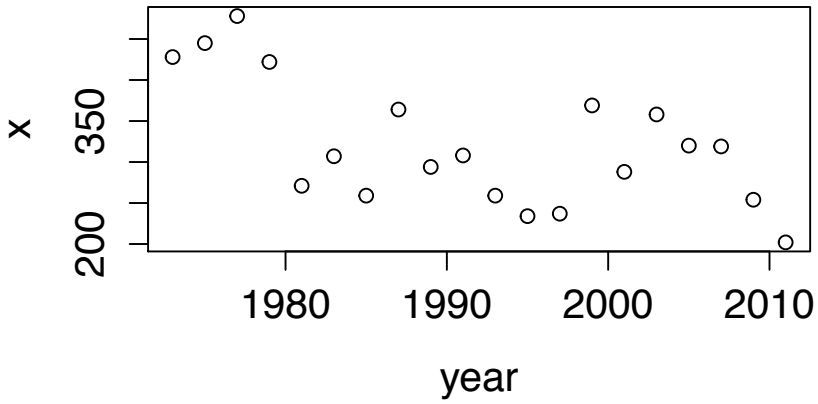
Visualization of over 800 laws passed by California representatives from 1973 to 2011.



Percent success of passing a law per delegate in each state (Alaska is the outlier, as the influential Don Young has been representing the state for 23 terms and is the only delegate from Alaska in the 38 years of data.)



Total number of laws passed per year from 1973-2011.



Design Prospectus:

After the initial analysis, the design prospectus is the next step in the Graphicacy framework for projects. This document helps to crystalize the concept of the project for the client in terms of what the agency is planning to do for the execution phase. It details what information is there, what it shows, and what story is trying to be told through both words and sketches. This document provides a basic guideline or script to refer back to in terms of guiding the project from here on out.

The initial brainstorm for the project was a four-stage presentation that visually explored the statistical underpinnings of the Legislative Effectiveness Score, and then allowed the user to explore the data themselves. After much discussion, the eventual deployment was agreed upon to start with the data exploration facet and if that proved a simple task, move on to the visualization of the statistics. The result was that creating program was much more challenging than originally anticipated, and it was a worthy project to dwell on and use to explore a well-rounded presentation of data.

Congress infographic / script:
Purpose: To visualize how a legislative effectiveness score is derived, as well as how each representative has scored across 40 years of congress.
Audience: anyone interested in the American democratic process

Setup/Intro page:
*Political scientists from Ohio- grabbed a bunch of political data from Congressional Quarterly and crunched it. Scientifically.
* Boiled data down into a single number to judge effectiveness
*They wanted to look into how certain factors might affect the lawmaking process like:
the roles of political parties,
committee leaders, or
race and gender effects

and thus, developed the Legislative Effectiveness Score (LES).

Graphic: small stat box on house of reps:
How many:
Gender breakdown:
Average length of term:
Number of people represented per rep:

Part 1:
First a look into how bills are passed:
Graphic Interface – multi-stage stepper to move through the explanatory process
Similar in execution to this:
http://elections.nytimes.com/2012/ratings/electoral-map?pagewanted=all

Text 1: Scientists looked and measured five parts of the bill passing –
Graphic – showing how bill progresses through to be law 5 stages,
And each stage a bill goes through accumulates the score. When it fails to move, the score for that bill stops being added.

Caveat: Explain that a simple average of how many bills sponsored / bills passed does not take into account how bills vary by importance

Graphic – show three circles describing the different types of bills –
All bills are not created equal. Some are more important than others-
Show the levels of commemorative, substantive and substantive and notable.

(Hopefully can carry circles through transition as page slides, to keep visual and conceptual continuity and help viewer understand we're still talking about he same things.)

Part 2:

Stepper stays – still the main navigation for the page (though scrolling works)
With out circles that represent the different types of bills,
Graphic – watch them grow in relation to their weighting in the statistical formula.
One stays same size, one grows 5x and one grows 10x to represent how they are weighted. Explain this in concise text

Now in bringing it all together, we can see how the type of bills and number of bills affect the overall LES-
Graphic: interactive circle with sliders that allow user to change number of bills and types of bills. A representative circle is shown (neutral color) and changes in response to user input. Perhaps create pre-sets that show different congressional all-stars from certain years?

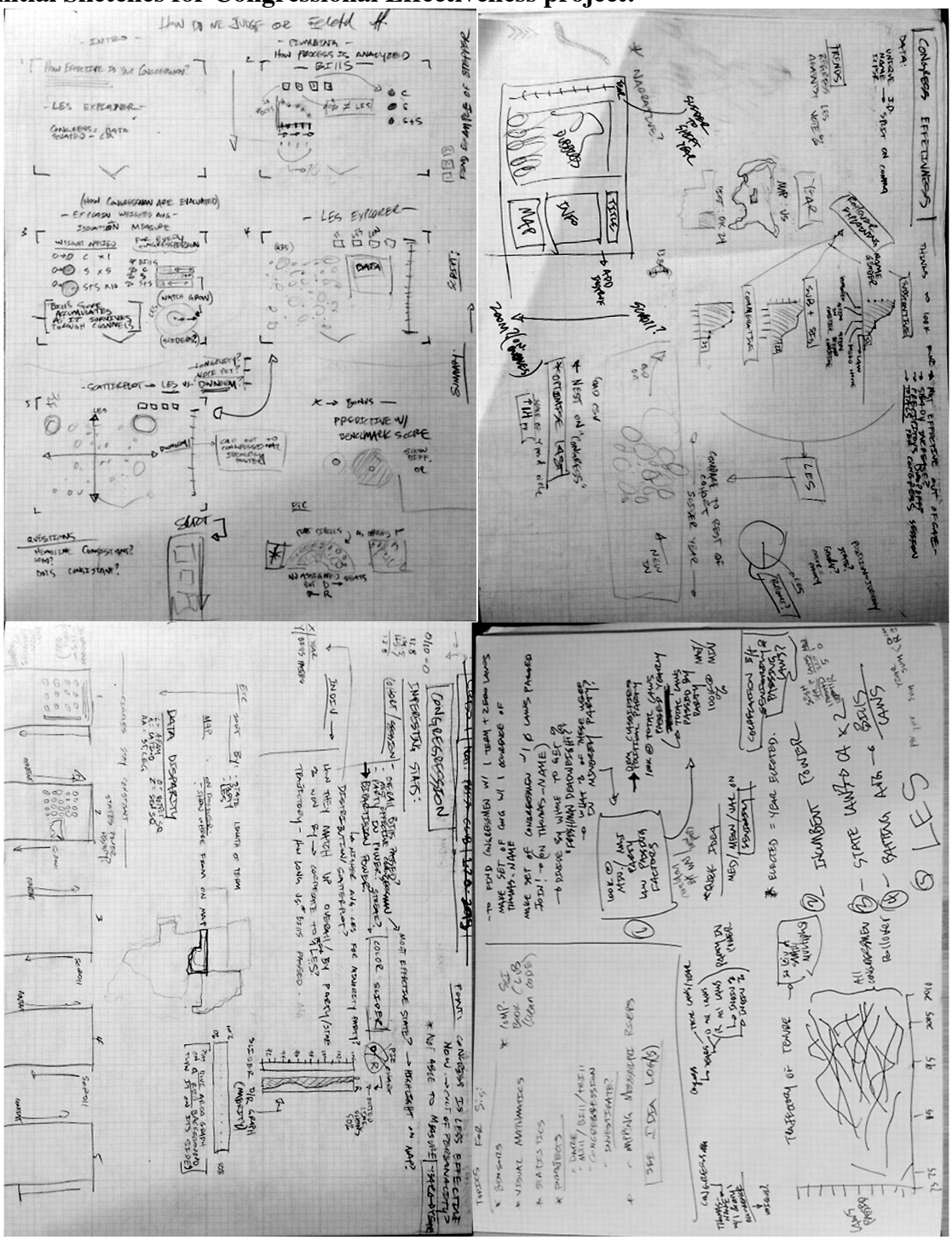
Part 3:

Text: explore the reps from 1973-2010. Size of circle is indicative of LES score. Roll over to see more detail.
Graphic -
LES interactive: - display 435 circles in arc like they sit in Rep.
Explain slider to access years of congress- roll over to see who / stats.
Sort by state:
Sort by gender:
Sort by LES
Sort by longevity:

Part 4 – scatterplot?

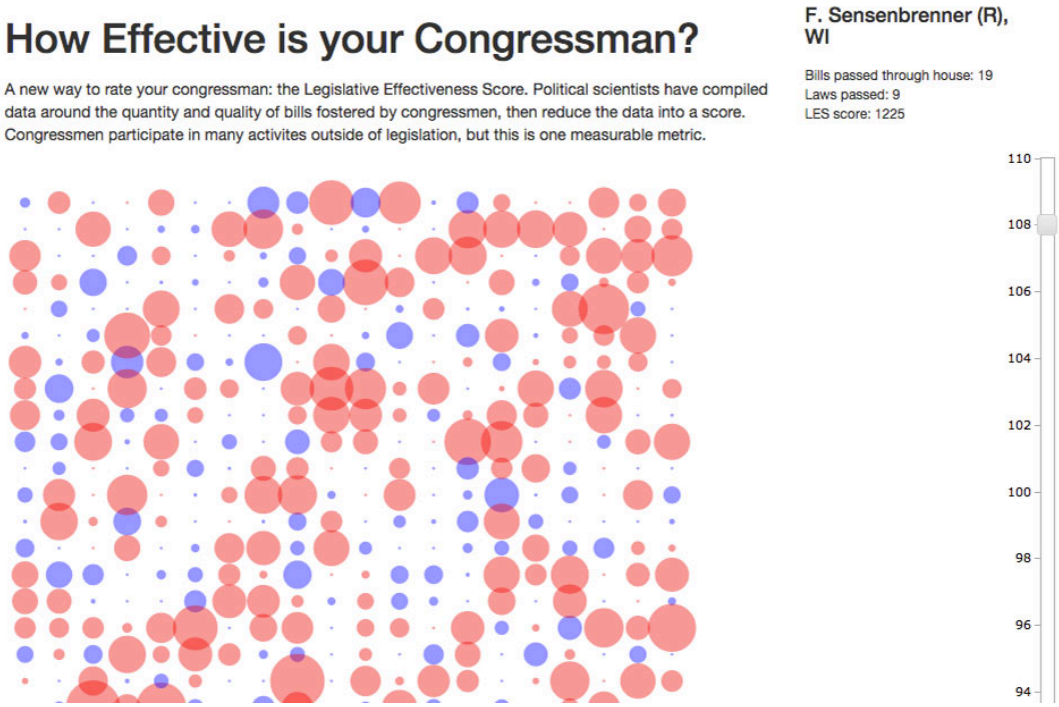
Measure LES vs:
DWNOM? -> direct tie into the congressional bias poster
Vote percentage? Women's influence?
Longevity of term? Bonus: Expected LES?

Initial Sketches for Congressional Effectiveness project:



D3 Production:

Version 1 of Congressional Effectiveness project – line of circles that transforms on click into a grid. [Click here to see it.]



Version 2 – The circles are arranged to mirror actual seating, and on click, the circle comes to front, with information displayed below. Buttons are on top left to sort by data points: Legislative Effectiveness Score, party (default) or tenure. [Click here to see it]