Proposal - **Vis-a-vis**

(keep to a page)

1) Team members + team name
   Kushank Raghav
   Kritika Versha
   Yuan Chen
   Ayshwarya Balasubramanian
   Shu Wang

2) Briefly describe your proposed idea (paragraph or two). What’s the statistical/mathematical concept you’ll be explaining?

3) Where the data is coming from

Timeline
Keep an updated (and detailed) timeline here (date,person,task)

Executive Summary

(keep this to under two pages, 1 is better)

1) **Provide a short summary of your project (include a description of who is this for).**

Our project is to visualize UMSI graduates' salaries by geographic location and by industry type. This visualization could be a valuable tool for faculty, staff, current students, and prospective student. Faculty and staff members could use the visualization to see where most students are headed in terms of location or industry. They could also use the tool to find outliers in terms of industry or location. For example, does the library industry pay too little or does the North East offer a much higher salary than other areas? For students, this could help them determine which industry they want to enter or avoid, as well as figure out where alumni tend to be location wise. Geographic locations are divided into regions i.e the West, Midwest, South-West, South-East, Mid-Atlantic and International. Each regional visual can be elaborated to states present in the region when needed by the user.

2) **What are the objectives for your end-user? (provide a bulleted list). Think of these as the domain level blocks of your block model diagram)**

- Compare salaries by geographic locations
- Compare salaries by industry
- Determine lowest and highest paid locations and industries
- Understand the distribution in salaries by state and regions
- Understand distribution in salaries by industry
- Compare number of students entering different industries or regions

3) **What kind of data did you find or generate to support this project?**

The UMSI Career Development Office provided us with a CSV file that contained three columns: Salary, Industry, and Geographic Location for five years. The CSV file was divided into different sections based on the year from which the data is sourced. To create our visualization, we cleaned the geographic location data to only include the state. We also added a corresponding region for each state which was added as a separate column. To ease creation of the visualizations, we also used Python to generate the minimum, median, and maximum salary for each industry and state/region.

4) **What are the abstract tasks (the "data comparisons") necessary?**

- Compare counts across different categories (state, region, industry)
- Compare median and ranges across different categories (state, region, industry)

5) **Provide a high level description of your proposed solution (one paragraph). You can refer to images on other pages of this document**

Our solution allows users to interact with the visualizations to find the information they are seeking. The top visualization allows users to view successive “deeper” levels to find the information about a subset of industries or geographies. At the top-most level, a user will see salary ranges and medians across all US regions and all industries. As they click on the visualization, they can view a subset of the data by region or by industry. At the bottom-most level, a user can see data related to just one state. This visualization allows users to find information most important to them. For example, if they know that they want to work in a particular state, they can find the range and median of salaries in that state.

6) **What are the visual comparisons that are featured as part of your implemented solution and how do they support the data comparisons above (try to refer back to the data comparisons to indicate which visual comparison supports which data comparison try to indicate succinctly why these work on the expressiveness/effectiveness scale).**

Data comparison 1: Compare counts across states, regions, or industries
A user would visually compare the height of the line between different categories. The y-axis has tick marks that would help a user estimate the count. This is expressive because a user can determine what the count is. It is mostly effective because a user can compare heights relatively easily. It would have been even more effective if the heights were right next to each other, but that would not have worked due to the need to compare median and ranges. In addition, the user can view the actual counts for regions on the bottom visualization.

Data comparison 2: Compare median and ranges across states, regions, or industries
A user would visually compare the beginning position, ending position, and line length to compare ranges of salaries across states, regions, or industries. He/she would visually compare the x-position of the dots to compare medians. This is expressive because a user can find out the values of the minimum, median, and maximums by using the tick marks on the x-axis. It is effective because users can compare line length and position of the circles easily.

Screenshots of current “working” solution

(use as many pages as you need)

Provide screenshots (with short captions or annotations describing what’s going on) here. **For the sketch** this can either be a working prototype or high-quality wireframes (Illustrator or Photoshop generated--no napkin sketches). These should be your current “best” guess as you go and then should be replaced with the screenshots of your final solution. For your sketch deadline--please be ambitious--I know you won’t necessarily be able to implement everything but I want to see more than a pie chart with a drop down box. **For your final report, replace these with final screenshots of your working system. Put in 3-4 “configurations” of your visualization that illustrate how it is to be used and briefly point out why the configuration is interesting. Provide a link to the Web page if it exists.**

Questionnaire

(2-3 pages)

Fill out the questions below and **keep them updated** as you evolve your solution. These can be quick answers, but do spend them time doing this.

Problem definition:

1) **What are the domain tasks? (Keep a running list here, you can filter down to the main ones you want later)**

    - Compare starting salaries by state or region
    - Compare starting salaries by industry
    - Compare salaries of different geographic locations or industries.
    - Find a subset of salary information based on geography or industry (For example, a user might only want to find median salaries in California, in the Library industry, or in California and in Library)
    - Compare number of graduates entering different states or regions
    - Compare number of graduates entering different industries

2) **What are your user's biases/limitations? What are they coming in with?**

We assume that the users are current and potential UMSI students and faculty and staff members. Some of these users might have no idea where students typically place. Others might have false assumptions about starting salaries or where students place.
In addition, these users might not have much experience with interactive data visualizations or with abstractions of data. For this reason, we decided to make our salary encoding simpler, moving from a boxplot to a simple range and median. We believe this makes the salary information easier to understand.

3) **What do you *not* know about them? (things that you can't get from the interviews/meetings/analysis)**

We do not know how the interpretations of the visualization would influence their decisions in picking the school, what reviews the current student would hold about their school and how motivating or demotivating the application would be for its audience.

Data:

1) **Describe the data you have to work with. Provide a list of different variables and their type. Keep this up to date with the data you are keep/tossing/adding**

We are working with a data set which has five variables: Salary, Industry, State, Region, and Year. We've also computed three variables for various combinations of industries, states, and regions: Minimum, Median, and Maximum Salaries

Salary - Integer
Industry - String
State - String
Region - String
~~Year - Integer~~
Minimum - Integer
Median - Integer
Maximum - Integer

Data comparisons:

1) **What data comparisons do they need to do as part of their job? Try to identify the kinds of comparisons they need to make and "statistical" or "analytical" operations (e.g., finding outliers, detecting trends, etc.) Add and remove from this list as you develop your visualization, but have a starting list before you do anything else. Keep this list numbered.**

We would want to give our users the ability to filter by region, state, or industry. By this kind of visualization the user will be able to statistically analyse which regions or states in the US pay the highest salary jobs and the density of the population that go into those regions will help the users analyze their chances in making into those jobs themselves. The users will also be able to understand the difference in salary ranges and would be able to spot the outliers on the map or the box model, if any.

Encoding:

1) **How will you encode the variables above? Be specific and make sure you're covering the comparisons above**

Bipartite Visualization

Since the focus is to track annual salary of the industry types by US geographic location, interactive map in D3.js can be chosen as the most feasible option.

A. **Elaboration: Region -> State**
   The user has the ability to visualize and compare the salaries across different categories of industries according to the **regions** i.e West, Midwest, South-West, South-East,Mid-Atlantic and International.By hovering and clicking over the region

bars , the user can visualize a more detailed version of salaries and UMSI graduate distribution in each region as seen in the screenshot visualization.

B. **Filter: Industry Type**
   The second comparison is by **industry** comparison across various regions which enables the user to track the annual salaries according to the geographic regions.
C. **UMSI graduate count distribution**
   The third bipartite graph gives an overview about which locations and industries are UMSI graduates heading .

2) **What interaction techniques are you using (refer back to the list of 7) and how are you using them.**

Abstraction/Elaboration: Average annual salary on maps by cities/states/regions.Salary details can either be observed by hovering over regions or it can be elaborated from regions to states using clicks over region.
Encoding: Each regions are encoded with a different colors.Similarly color encoding is provided for industries.Text encoding is done to give additional details.Highlighting is used to enhance user attentiveness
Connect: The edges in the graduate distribution bipartite graph is also used to increase the expressiveness and the effectiveness of the visualization.

3) **For every data comparison, provide the visual comparison supported by your visualization (refer back to the answers above). Example: if they need to compare the means, how do they do this visually?**

Comparing median of salaries across different industries/regions - visually by comparing position of circles.
Comparing number of graduates entering a region/industry - comparing length of bars.
Comparing different regions/industries - comparing color hue of bars.

4) **How does your encoding expressive given the data comparisons needs (argue for both the encoding and interaction)?**

Color hue encoding expresses all regions and industries.
Length of bars expresses number of people entering a region or a industry.
Circle position expresses median of salary for a particular industry or region. Also, position encoding is used to express minimum and maximum salaries for a particular industry or region.

5) **Why is your solution effective given the data comparison needs (argue for both the encoding and interaction)?**

Region and Industry is encoded in Color Hue. ‘Ranking of Perceptual tasks’ states that this is an effective choice. Since this is a bipartite graph, Position encoding was not appropriate. The count of people in a particular region and industry is encoded in ‘Length’. Count of people is a Quantitative variable and ‘Ranking of Perceptual tasks’ tells us that this is an effective choice. Median is encoded in position and since salary is a quantitative variable, position is the most effective choice. Gestalt principles state that figure and ground should be clearly distinguishable. Hovering over a bar highlights the links between region and industry and this design satisfies this criteria. Also, Gestalt principle of connectedness states that elements that are connected group together. The connection between region and industries supports this grouping principle. It is effective because we want to group region and relevant industries (and vice versa) together.

1) **Given the domain/abstract/encoding, describe an experiment (or experiments) you would run.**

It would be fascinating for us to know which regions are the most filtered and which regions are the most sought after by the target audience. It would also be intriguing to analyze who the logged in user would be and for what he/she draws from the visualization before quitting the application. Another aspect that would be interesting to analyze would be that, how many people would visit this visualization, if the data is constantly updated every year.

2) **What are the different risks based on the nested block model (upstream and downstream risks, again, think domain, abstraction, encoding, and algorithm)**

Given the domain, our biggest challenge would be if the user chooses to select all the regions and compare all the values at once then, the visualization should be clear and should continue to make sense to the user without seeming to be like information overload and the data should support techniques such as brushing, filtering etc so that the user would find the information that he /she is looking for, easily and fast.

3) **Did you try your solution on anyone? What did they have to say? If you did an evaluation, describe it in detail here.**

We have not tried out solution yet on anyone, but we intend to do so on our peers in SI so as to get feedback on what we could improve and how easily or how complex they think of the solution to be.

Reflections:

- What worked?

Using a bipartite graph for visualization. Hovering over a particular category such as region highlighted relevant connections between region and industry and vice versa.

- What didn’t?

The map visualization presented in the draft version was insufficient to provide a distribution of UMSI graduates across different regions and simultaneous comparison of the salaries across regions and states.

- What tradeoffs did you make? (what comparisons *aren't* you supporting?)

Our initial sketches of filtering the data using the year didn't work as the dataset provided to us was sparse. As a result, the data wasn't spread uniformly over all the industries, regions and states.In our visualization, we are confined to compare the salaries across different states within a region. It does not support comparison across states present in different geographic regions.

Working todo list:

1) Keep a list of your "implementation" tasks here. This should be the list of things you need to do to get from your wireframe/prototype to the finished product? As you finish, remove them, but leave the ones you didn't get to (remember, you should be ambitious in your sketch).

1. A static graph with region-industry distribution of number of people.
2. Make the graph interactive using BiPartite, hover over either one region or one industry to show break-down in other parameter.
3. Add salary range (lines and dots) next to each break-down item
   (Didn't implement: didn't find a way to position range lines and dots properly on the graph we have.)
4. Make salary range distribution a separate graph

Notes from 1st peer round robin:

- Other ways for geographic data (other than map)
- Frequency in a city (hot cities)
- Questions make sense to students (**what people care about for full time job**)
    - Group 7: zoom in to a region, see all salaries ::: median may not make sense
    - How to incorporate box plot in map
    - Develop domain problems (e.g. trend)
    - Normalize data, consider living expenses
    -
- Ways to avoid bias: we don't have the whole data (not every student reported)
    - Avoid treating as if this is the whole population
-
- Some ideals: comparing with other iSchools
    - predicting

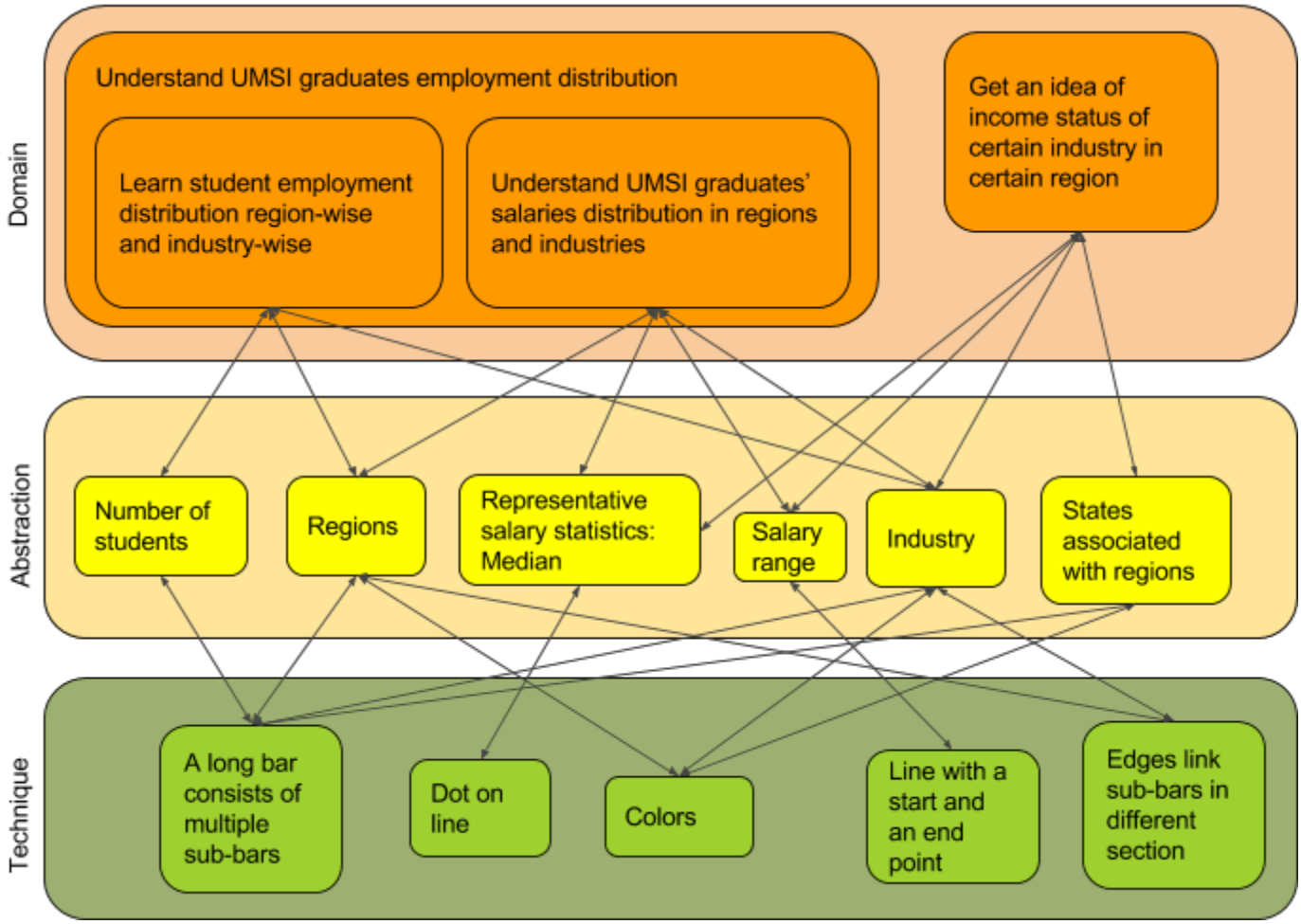
Notes from 2nd Peer Round Robin:

- We would want to know the range.
- Have a separate and similar salary structure and map it in a way we have similar to region and industry. Code salary as it is coded currently.
- Is the median most important statistic? What about Percentiles?
- If you mouse over a selection, a pop up shows salary summary statistics.
- Select by clicking multiple selections in each category such as industry, region.
- Median salaries by region and industry?
- Why Median?
- Let users choose statistics such as Mean or Median.
- If we select a label such as a region and another label such as industry and show salary statistics for these selections in a bar graph - so salary statistics for computing in midwest region. so if you select only one selection such as midwest, we will have a line showing minimum, median, maximum corresponding to each industry will be shown next to it and viceversa.
- Explore graph by selecting a salary percentile. So which industry has the top most paying job?

Nested Block Model

(as many pages as you need)

Draw the Nested Block and Guideline Model (http://www.cs.ubc.ca/labs/imager/tr/2013/NBGM/) for your solution. This can be a simplified version (you don't need the complex nesting or arrows within each level), similar to the ones see in class. However, each domain task should have a box which should be mapped to abstract task(s) and then to encoding blocks, and so on. There should be a connection that supports design decisions.

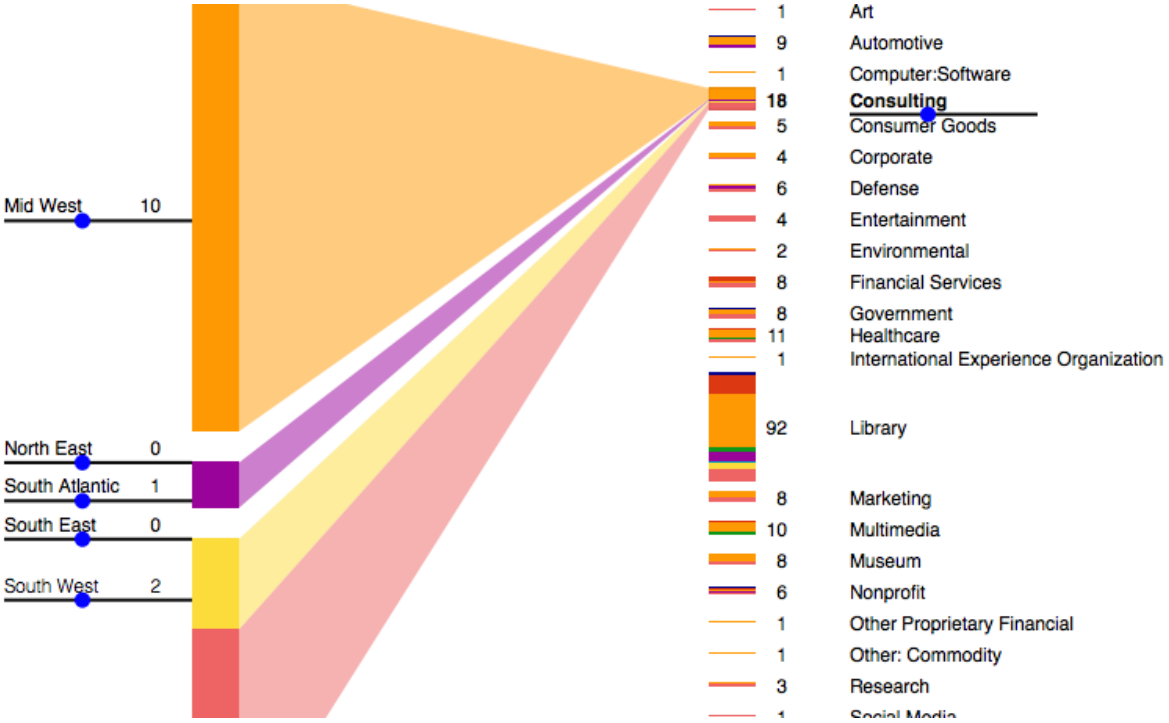


Archive of Experiments (sketches)

(as many pages as you need)

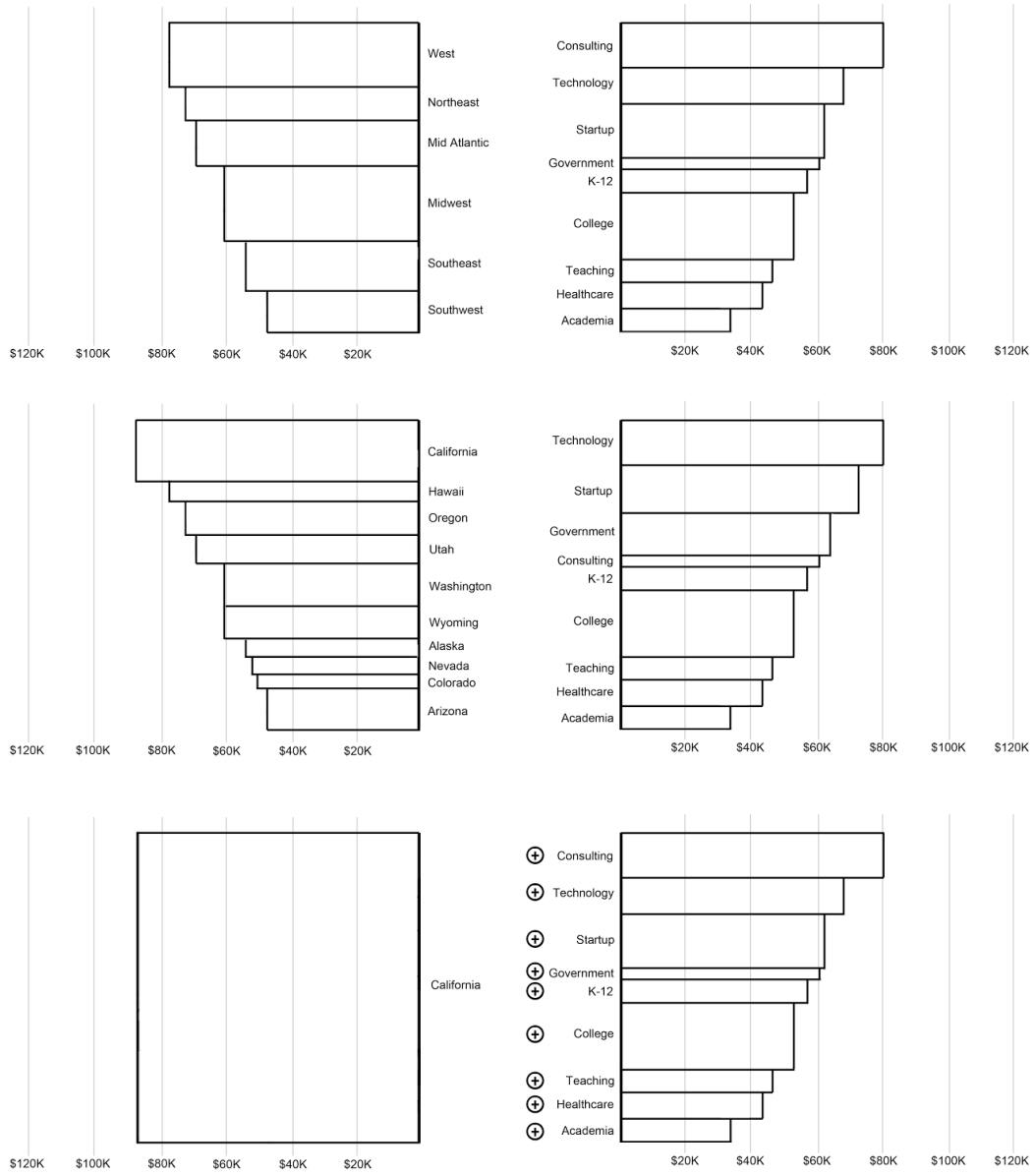
As you try various experiments keep a record of them here. These can be napkin sketches, whiteboard scans, Illustrator documents, prototype images, etc. Anything goes. Try to keep track of why you like them or decided to toss them. Do not skip this!

April 14



We wanted to display salary range and median next to each data rectangle, but it didn't work well with position.

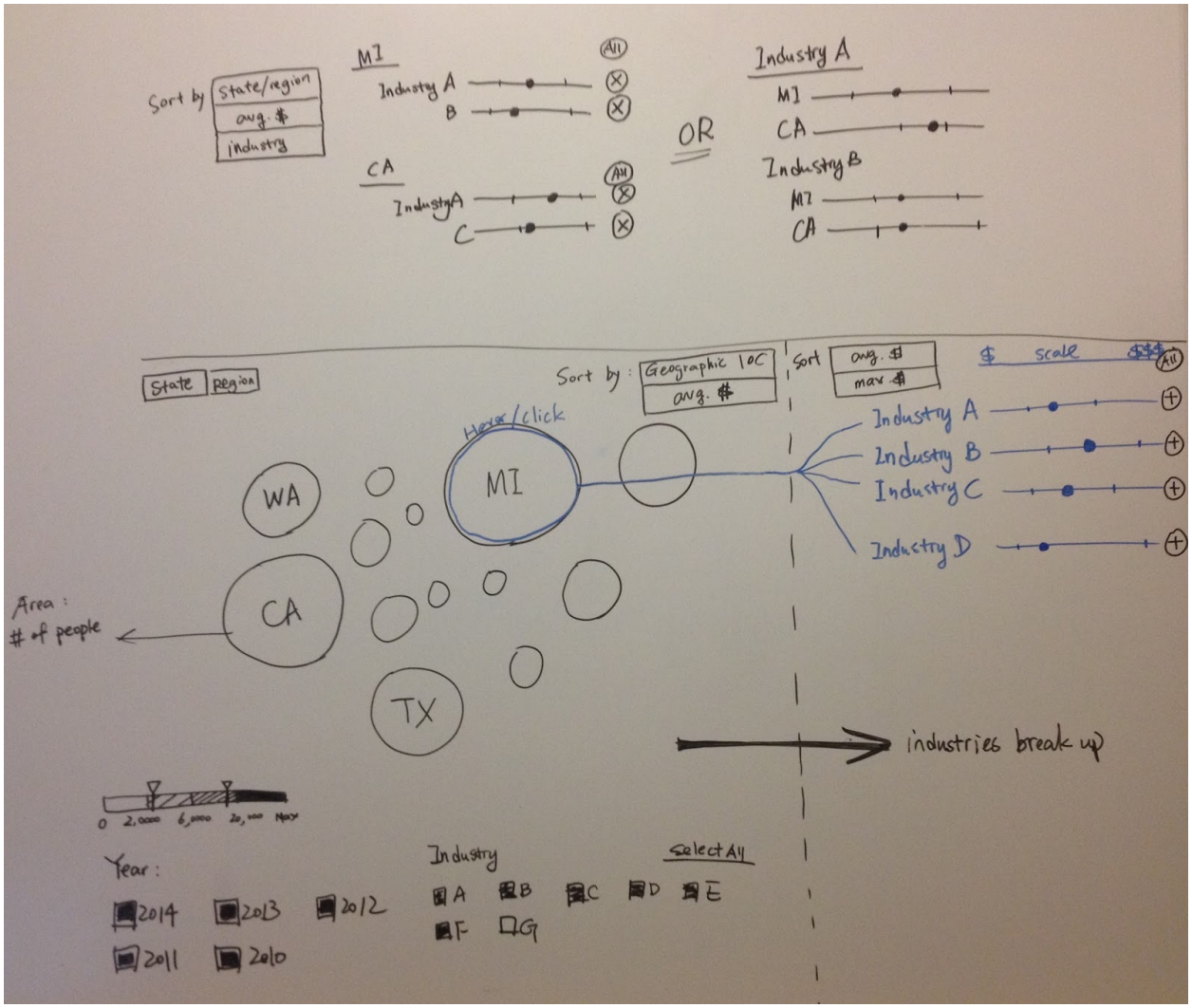
April 10



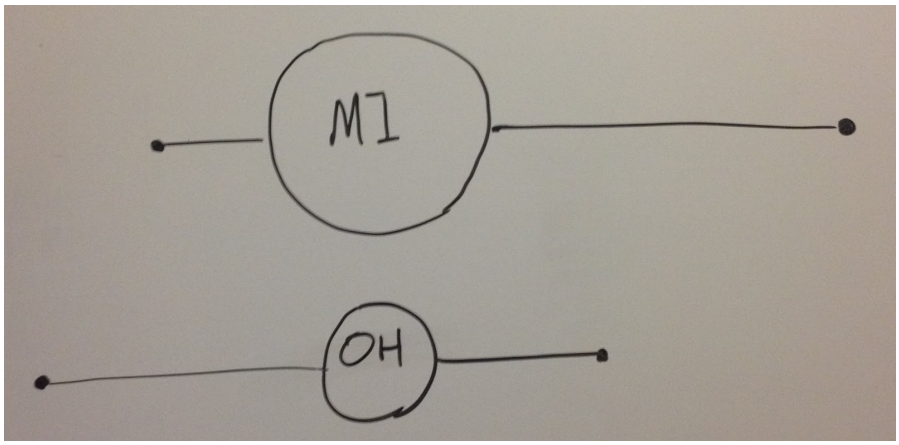
In this visualization we encoded the median as the width of each rectangle and the count for each category as the height of each rectangle. We decided not to encode the counts and medians as a rectangle because a people perceive a rectangle holistically instead of perceiving height and width. Thus it would be less effective to encode two unrelated variables as a rectangle.

We did decide to keep the back to back between the region side and industry side from this visualization. We thought it was an effective way for users to understand that they were viewing the same data from two different categorical groupings.

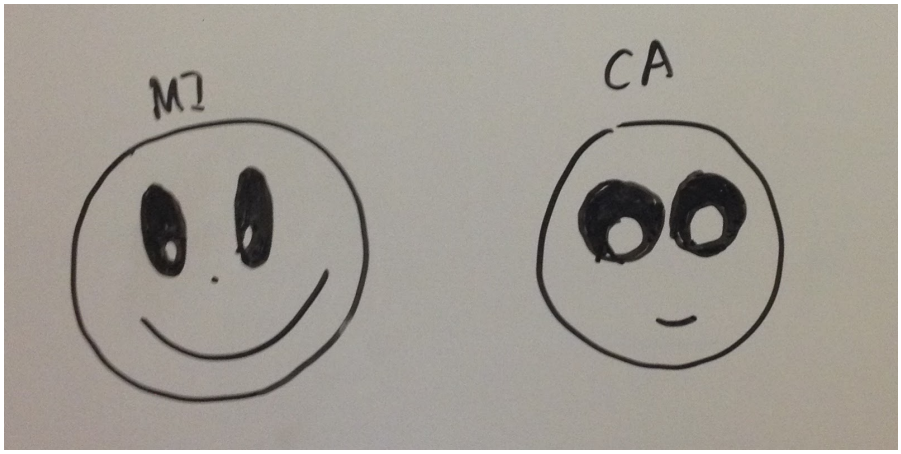
April 2

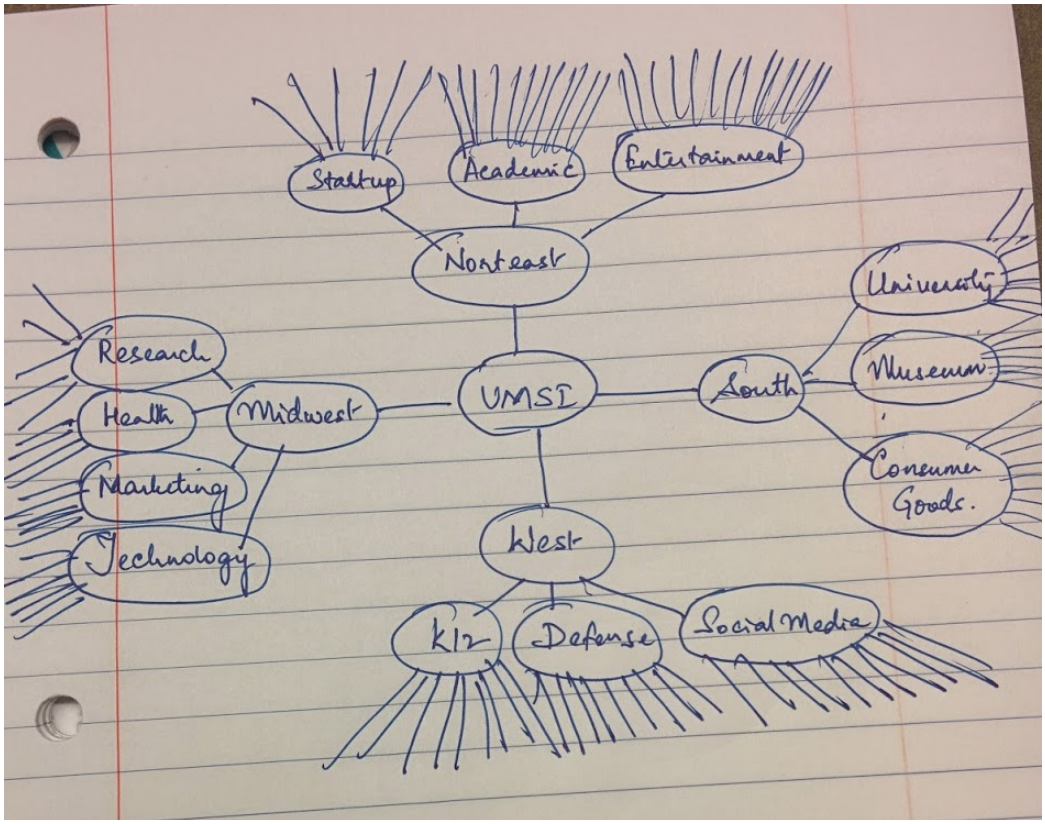


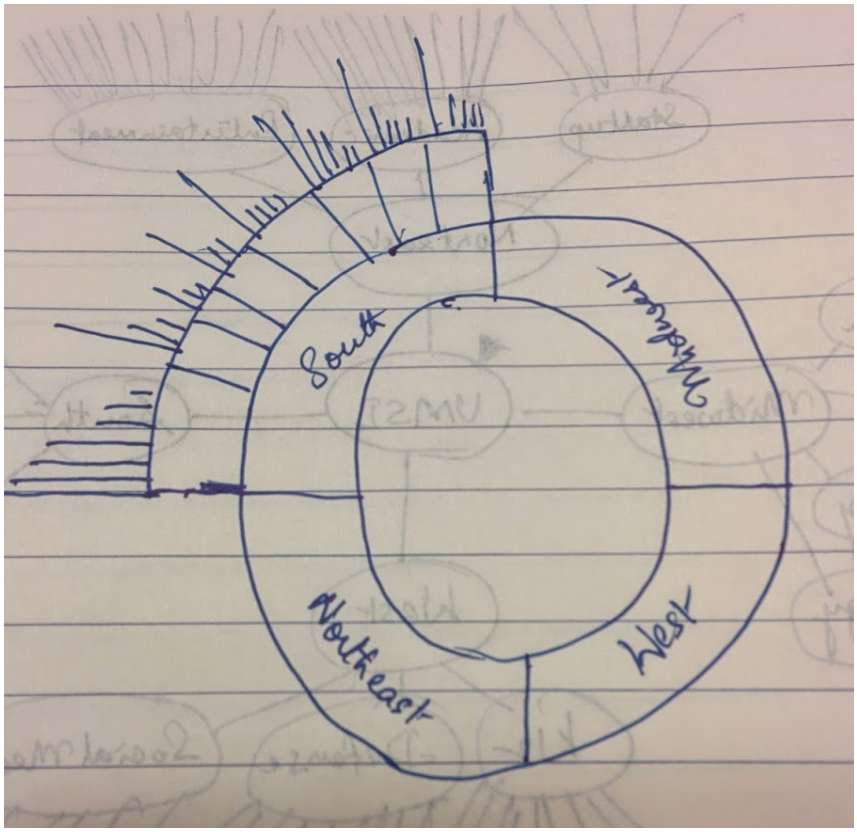
One alternative:



This is kidding:

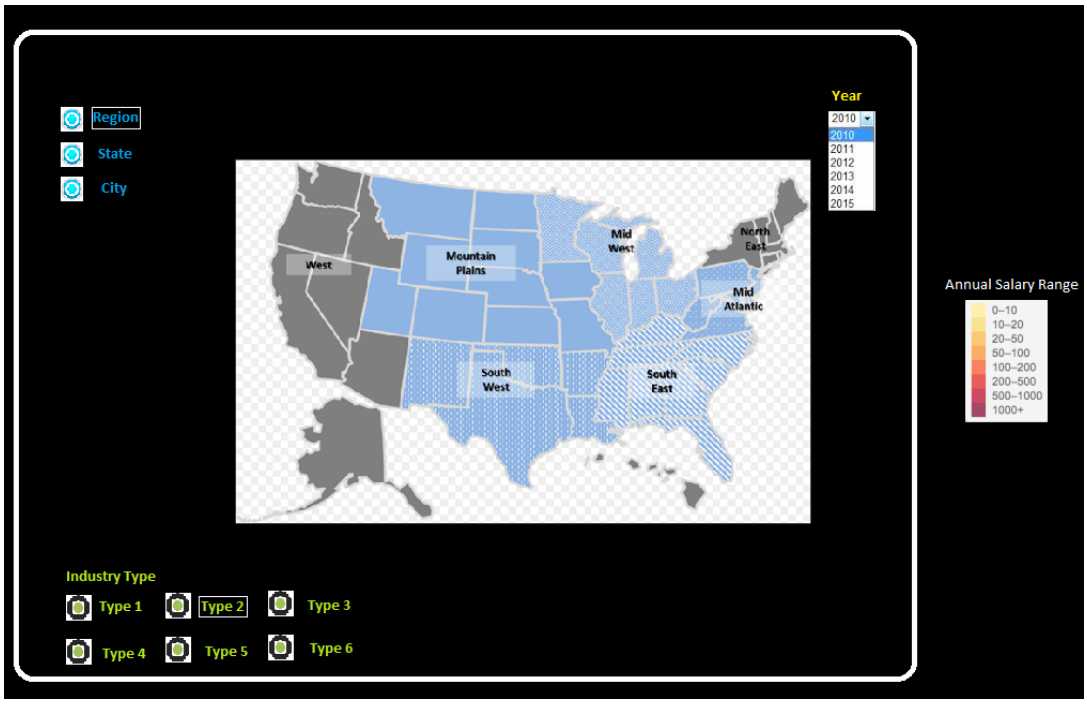



Startup
Academic
Entertainment
Northeast
University
Research
Health
Midwest
UMSI
South
Museum
Marketing
Technology
Consumer Goods.
West
K12
Defense
Social Media

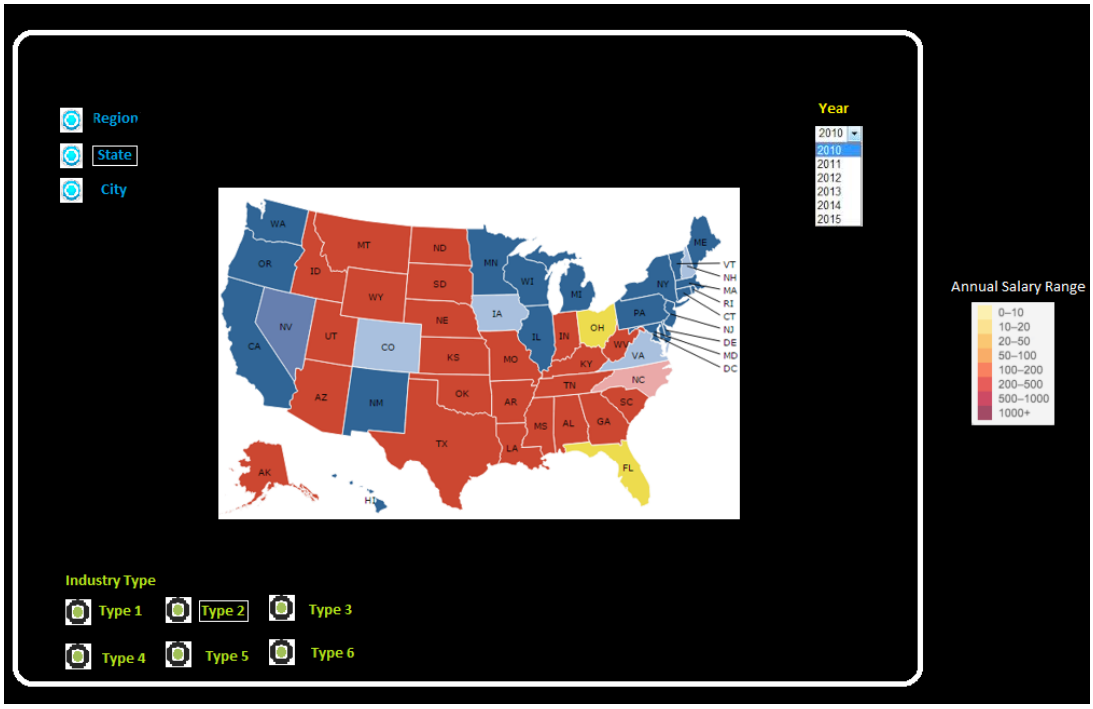

South
Midwest
Northeast
West

Mar 29

Map Visualizations:
Visualization of salary and industry by region:

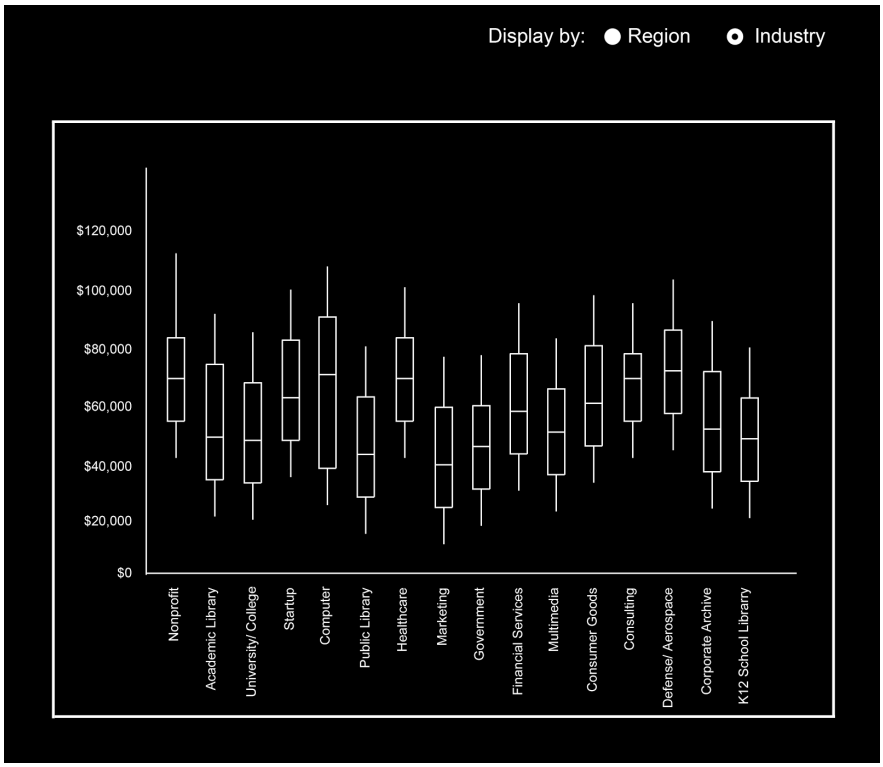


Visualization of salary and industry by states:

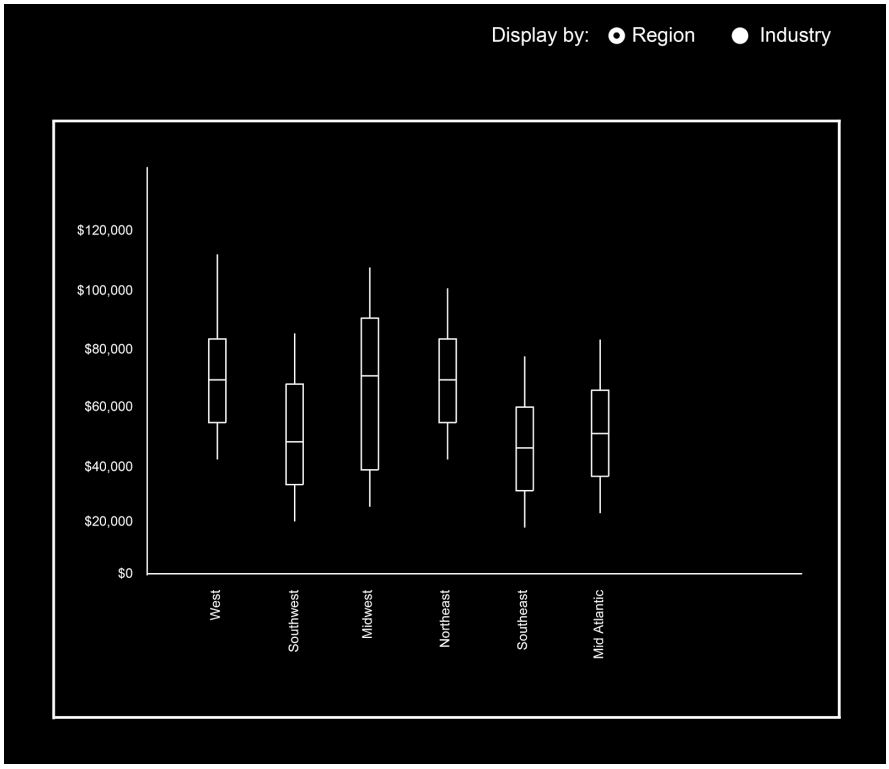


Visualization of salary and industry by cities

Boxplot visualization by industry



Boxplot visualization by region



We decided to move away from doing a map visualization because the location did not actually matter and thus the position on a map did not encode any information. In addition, we found that it was difficult to encode more than one variable in the map. For example, we could use color to encode number of people going into that region or state, but then we could not think of a way to encode the median salary or range.

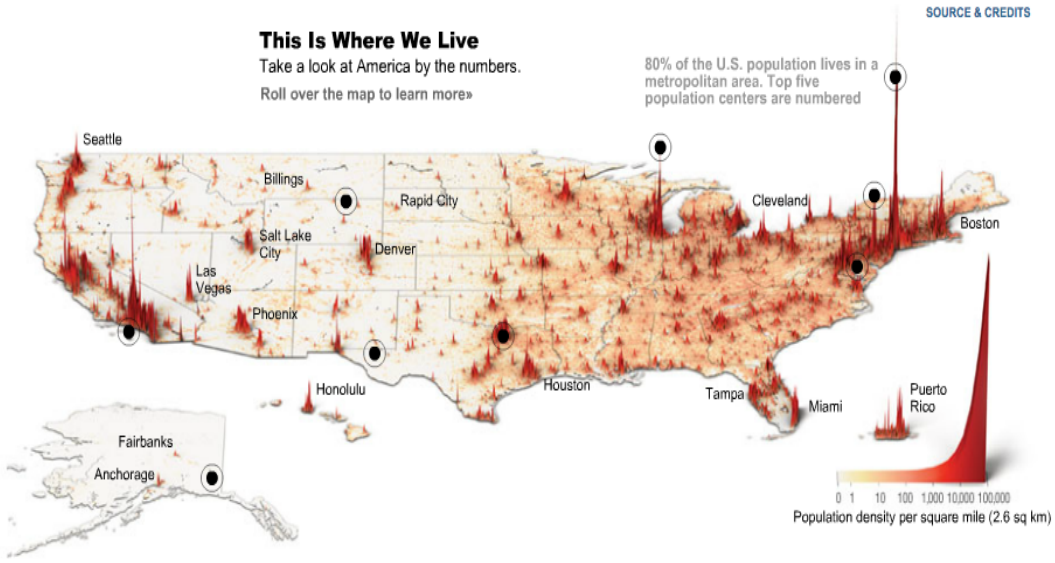
We also decided to move away from boxplots because many people unfamiliar with data visualizations do not know how to read them. We felt that the 25% and 75% percentiles were not as important as the minimum, median, and maximum salaries. Thus, in later iterations we decided to only use the minimum, median, and maximums.

Inspiration

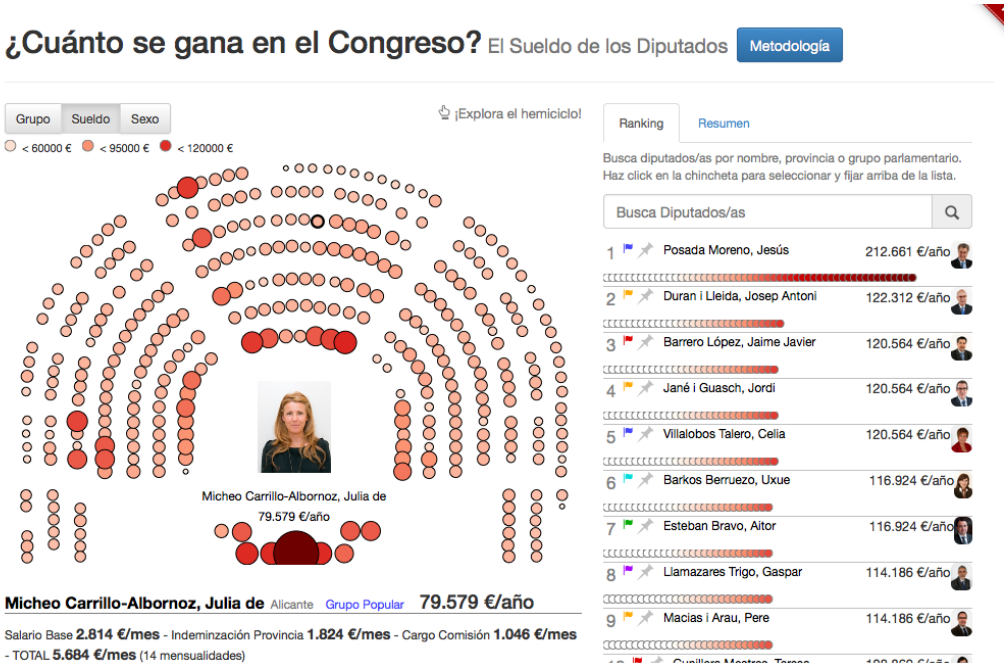
(as many pages as you need)

Keep screenshots (and references) to other solutions that you're using as inspiration for your solution (things you found on the Web, in books, in papers, etc.). Keep notes about what works/doesn't. Keep adding to this as you work. Do not skip this!

Where American lives
http://content.time.com/time/interactive/0,31813,1549966,00.html
Inspiration: hot spots

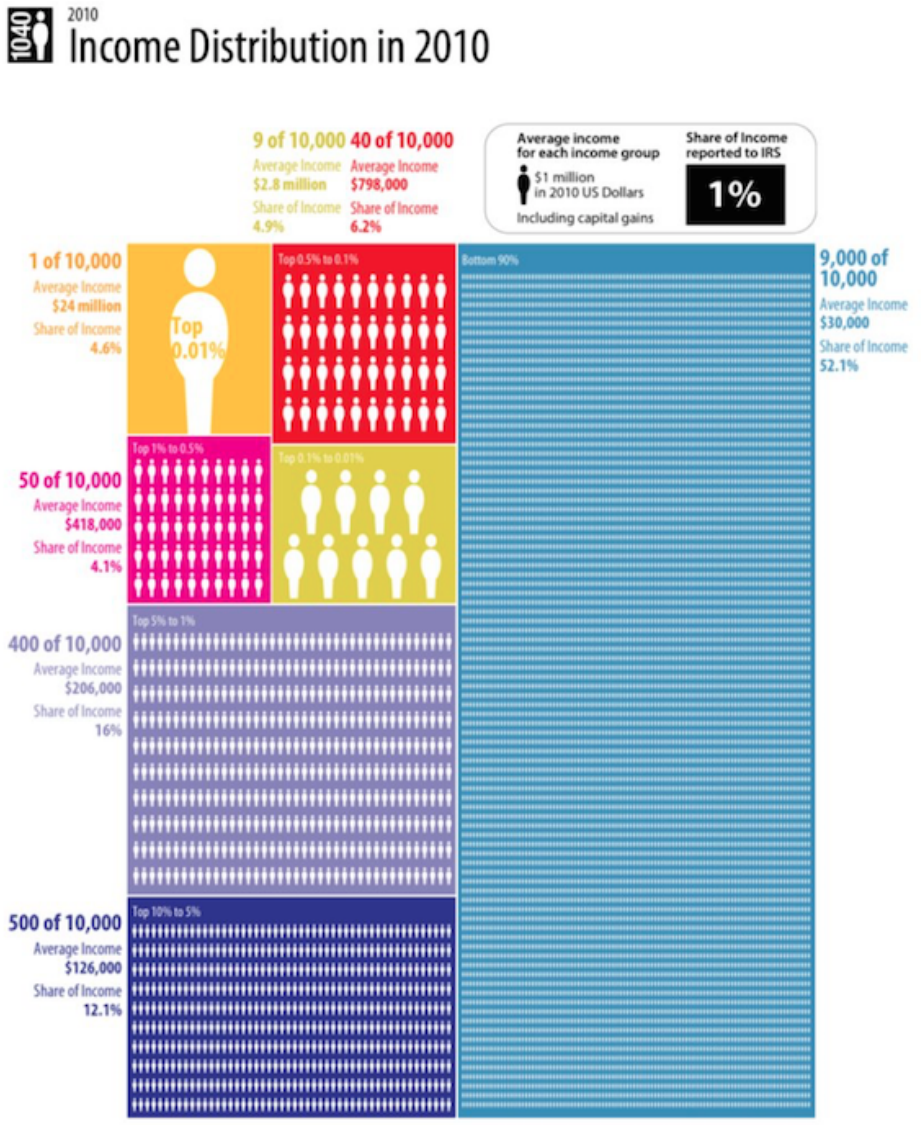


Spanish Parliament – MP's Salaries D3js Visualization (in Spanish)
http://sueldosdiputados.herokuapp.com/
Inspiration: multiple filters + connection between major graph and others

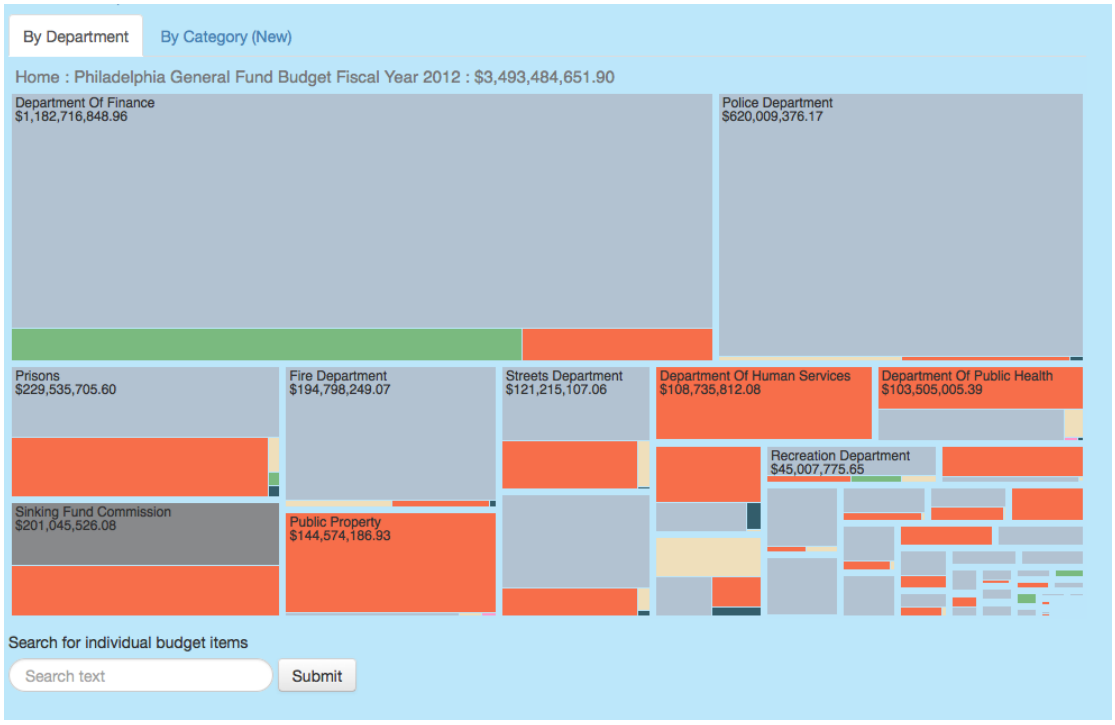


Income distribution

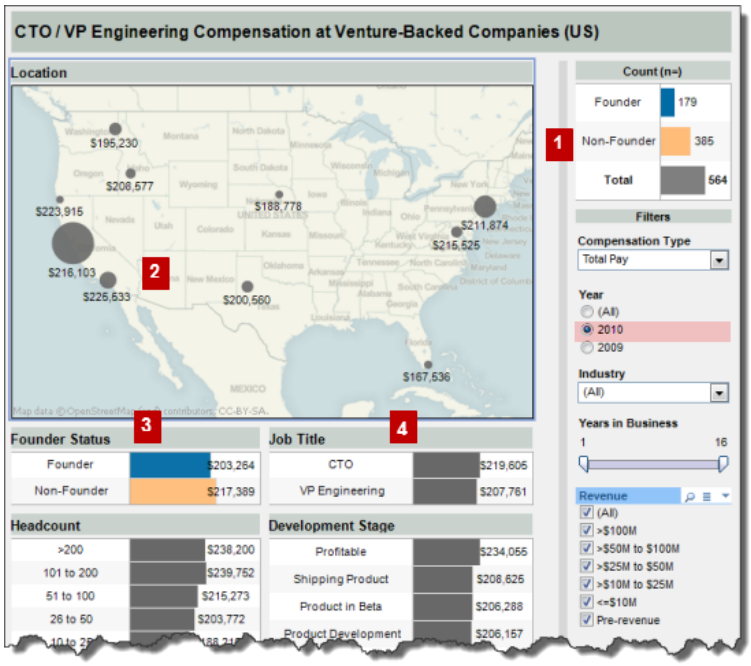
Inspiration: person shape or whatever glyph + geometric shape and area



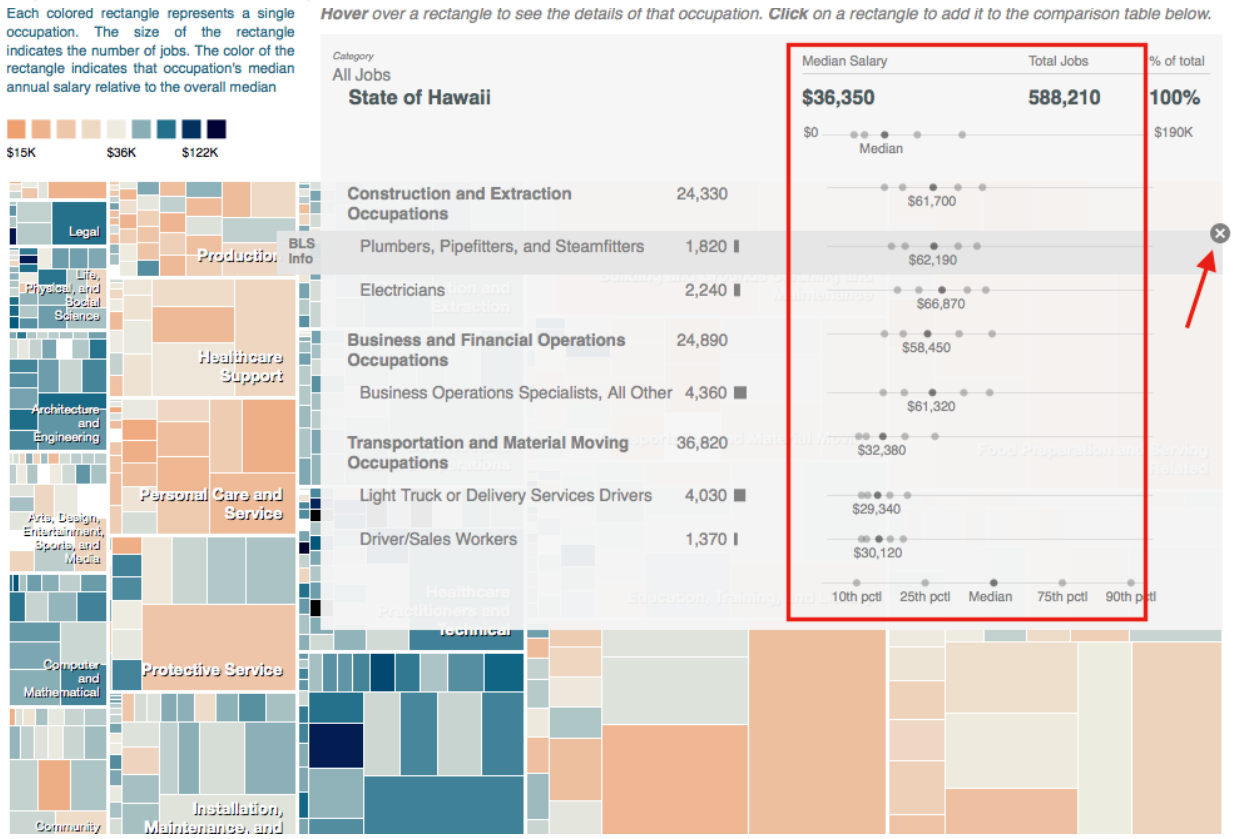
Mandel for Controller Bulldog Budget
http://budget.brettmandel.com/
Inspiration: exploration: drill down and expand



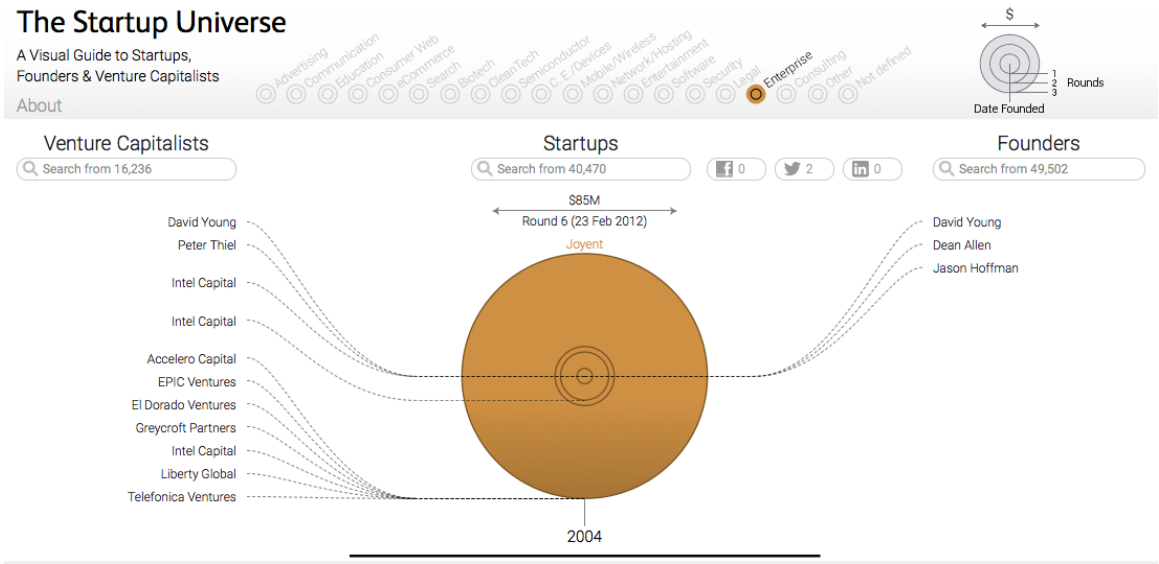
Visualization of Startup CTO Equity and Salary Data
http://www.socalcto.com/2011/05/visualization-of-startup-cto-equity-and.html#sthash.L2pyVdW2.dpuf
interactive: http://www.datarevelations.com/startup-cto-salary-qnd-equity-data-us
Inspiration: similar to what we proposed + note this one's encoding is terrible

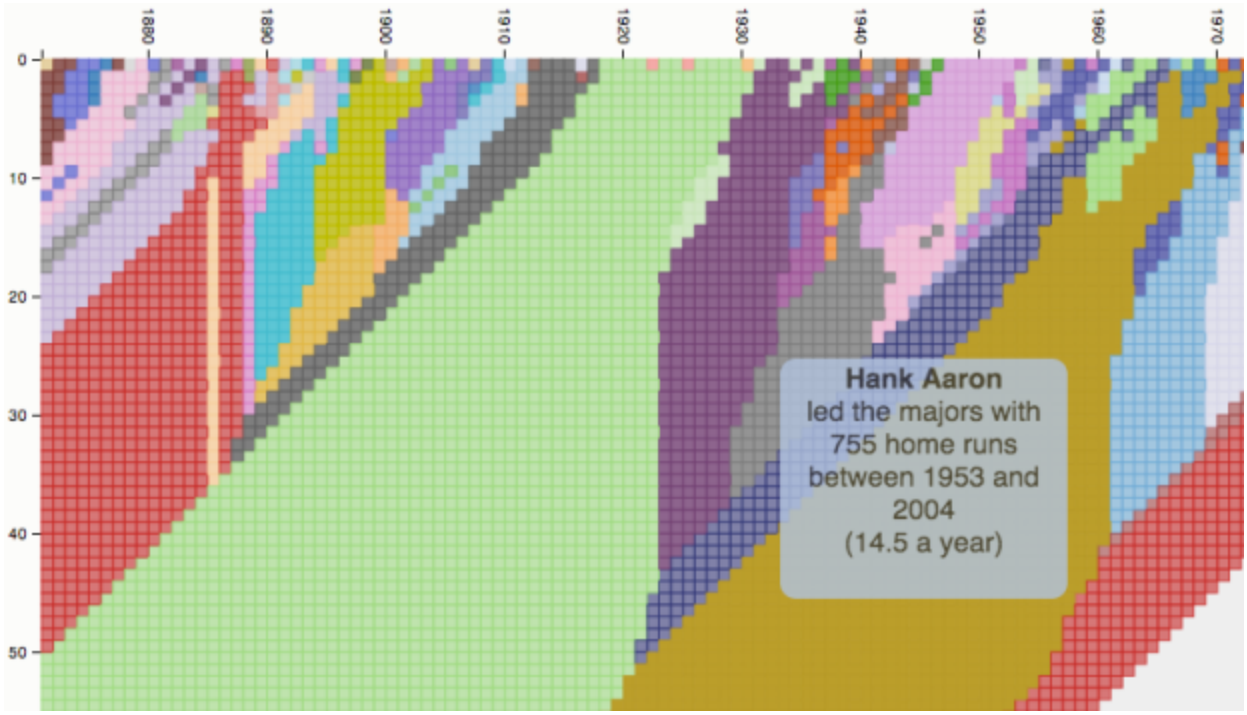


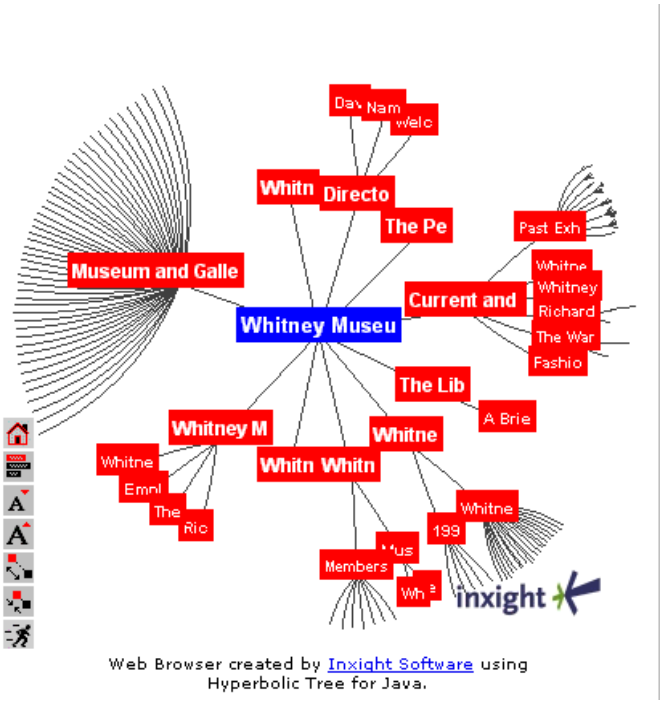
Another tree map for salaries by occupations
http://www.uhero.hawaii.edu/static/dashboard/jobs.html
Inspiration: add for comparison

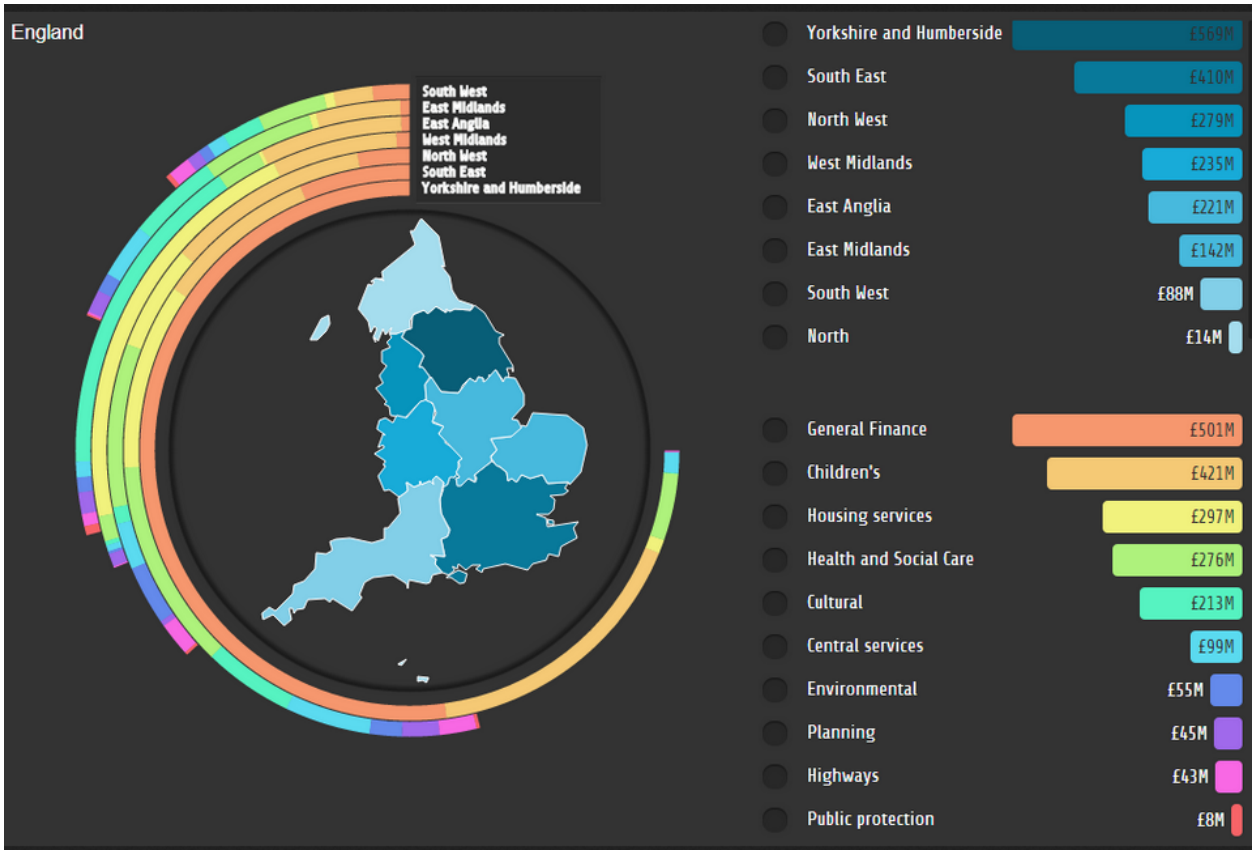


The Startup Universe
http://visual.ly/vizbox/startup-universe/
Inspiration: explore interaction


The Startup Universe
A Visual Guide to Startups,
Founders & Venture Capitalists
About
Advertising
Communication
Education
Consumer Web
eCommerce
Search
Biotech
CleanTech
Semiconductor
C. E. Devices
Mobile/Wireless
Network/Hosting
Entertainment
Software
Security
Legal
Enterprise
Consulting
Other
Not defined
$
Rounds
Date Founded
Venture Capitalists
Search from 16,236
Startups
Search from 40,470
0
2
0
Search from 49,502
Founders
$85M
Round 6 (23 Feb 2012)
Joyent
David Young
Peter Thiel
Intel Capital
Intel Capital
Accelero Capital
EPIC Ventures
El Dorado Ventures
Greycroft Partners
Intel Capital
Liberty Global
Telefonica Ventures
David Young
Dean Allen
Jason Hoffman
2004


1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
0
10
20
30
40
50
Hank Aaron
led the majors with
755 home runs
between 1953 and
2004
(14.5 a year)


Museum and Galle
Whitney Museu
Whitn
Directo
The Pe
Past Exh
Current and
Whitney
Richard
The War
Fashio
The Lib
A Brie
Whitney M
Whitne
Whitn Whitn
Whitne
Emnl
The
Ric
Whitne
199
Members
inxight
Web Browser created by Inxight Software using
Hyperbolic Tree for Java.


England
South West
East Midlands
East Anglia
West Midlands
North West
South East
Yorkshire and Humberside
Yorkshire and Humberside £569M
South East £410M
North West £279M
West Midlands £235M
East Anglia £221M
East Midlands £142M
South West £88M
North £14M
General Finance £501M
Children's £421M
Housing services £297M
Health and Social Care £276M
Cultural £213M
Central services £99M
Environmental £55M
Planning £45M
Highways £43M
Public protection £8M