CS 573 Final Project Proposal

Basic Info

Project title: **An American Day**

Team members

- Ben McMorran (bjmcmorran@wpi.edu, benmcmorran on Github)
- Francisco Sanchez (fmsanchez@wpi.edu, friscis on Github)

Project repository: https://github.com/benmcmorran/DataVisFinal

Background and Motivation

The American Time Use Survey (ATUS) is a continuously run survey by the US Government that asks respondents aged 15 and older to recall what they did in the last 24-hour period. Over 3,000 samples are collected per month. The data is used for a wide variety of purposes such as demographic research, quantifying the amount of non-paid time that people work, and measuring time spent on elder and child care.

The reason the ATUS data was selected for this project was because this data consists of a large amount of data that can be approached and analyzed in countless different ways, which allows us to come up with our own unique approach to visualizing our analysis. Another reason, we selected this dataset was because when we searched for previous visualizations of this data most of them seemed to do a poor job at conveying the results of their analysis. As a result, we thought we could apply the skills and knowledge from this course to develop an interactive visualization that is versatile and concise. Some personal motivations for choosing this project include our own curiosity for how different demographics influence the ways people utilize their time, as well as being able to explore and see where our own time use fits best within the data.

Project Objectives

We identified three main questions that our visualization will try to answer.

1. Which activities do Americans do most commonly throughout different time periods (day, week, year, etc.)?
2. How do common activities vary by location within the US?
3. Are there significant differences in activity prevalence between different classes (age, sex, race, education, and income)?

Answering these questions should give us a better understanding of how Americans spend their time and could reveal interesting differences between time use for different classes of Americans. For example, we may find gender differences in the amount of time spent on housework. These differences can help to quantify how far we are as a society from an egalitarian ideal.

Additionally, this project is designed to appeal as a casual exploratory tool to a wide audience. Everyone has experience with time use from their daily life, and they will likely find comparing their own experiences to the averages to be an interesting and fun exercise.

Data

Original data comes from the Bureau of Labor Statistics (http://www.bls.gov/tus/) in the form of several linked tables for each year of the survey with information about activities, respondents, and households of respondents. However, this raw data is very detailed and can be difficult to work with, so instead we will be collecting our data using ATUS-X, the American Time Use Survey Extract Builder (https://www.atusdata.org/atus/). This online tool provides an easy interface for extracting relevant attributes from the large ATUS data set and summarizing activities at the daily level if desired.

Data Processing

The data obtained from ATUS-X is already mostly clean, but some processing will still be required to extract the quantities of interest. ATUS-X data is provided at either the respondent level or the individual activity level. Both of these data sets are far too fine grained for our visualization needs. A preprocessing tool will be need to aggregate the individual activity data into distributions of activities by time of day, and to aggregate by respondent attributes like age, sex, and race. Additionally, the survey does not use uniform random sampling, but instead oversamples minorities and certain locations to ensure good statistics for these populations. A weighting factor is assigned to each respondent to account for this oversampling, and this weight will need to be incorporated into our preprocessing tool to ensure that overall averages accurately reflect the American population.

We are planning to implement the preprocessing as a Python script that can be run directly on the CSV files produced by ATUS-X and output CSV files suitable for consumption by d3. Python has good CSV support built-in and none of our preprocessing is statistically intensive enough to demand the use of a dedicated statistical tool like R. Both team members are already familiar with Python, making it a natural choice.

Visualization Design

We approached the design process by first brainstorming ideas individually before coming together to discuss our different approaches. This allowed us to have a broader basis on which to implement our final design. We had various different ways to visualize comparative data, through bar charts, line graphs and choropleth, as well as time disturbed data through stacked area charts, scatter plots and stacked bar charts. Additionally, we came up with some ways to represent hierarchical data through sunburst charts and tree maps, and some ways to summarize the entire data set through matrixes and common activity timeline bars.

The three designs we came up with to represent the dataset can best be summarized as an age focused visualization, and customization focused visualization, and a comparative focused visualization. The age focused visualization utilized a stack area chart with age as an x-axis from 15 to 85 years of age, and had the option to show a sunburst chart as well as another stacked area chart based on a 24-hour period for any age you hover over. The 24-hour period stacked area chart uses percentage of respondents doing a given activity for the areas, and the 24-hour time scale for the x-axis. Additionally, by clicking an activity from the 24-hour period stacked area chart, you would bring up a choropleth of the united states, that shows the distribution of respondents doing the selected activity based by state location. This design does a good job of conveying the continuous nature of age, and has the opportunity to show the full hierarchy of activities, however it is not very good for comparing across various classes, and only allows comparisons via stacked visualizations.

The customization focused visualization utilized various drop down and additive buttons to select various different configurations for the graphs to represent the data. This visualization utilized line graphs to show how time affects various activities, stacked horizontal and vertical bar charts to show how classes affect the percentage of time spent doing each activity. The benefit of this visualization is that it allows a great deal of user interaction and customization so the user can view whatever (s)he feels is most important, and given the compact design can show all visualizations without having to scroll down. The drawbacks of this visualization is that certain insights may be lost because it relies almost entirely on user interaction.

The comparative focused visualization utilized a combination of a stacked area chart and timeline bars to show what percentage of respondents are doing each activity at each hour of the day, as well as breaking that up into each of the different classes as well as buckets of age groups. Additionally, by clicking an activity on the stack area chart, this visualization brings up a line graph of the distribution of respondents doing that activity throughout an average day. Underneath this line graph is a bar chart that shows how many hours on average each class spends doing that activity in a 24-hour period, with the ability to select any class to add it to line graph. Finally, there is a vertical hover cursor that displays the time distribution of each state on a US map. This visualization is good to make comparisons between classes because they are on a shared axis, and it reduces area for the visuals by using a stacked area chart and as a result is fairly compact. Some cons with this visualization is that common activities may be very similar across classes, the map is only for display, and that the age is a ratio but is only represented in discrete bins.

Our final design (see final sheet in appendix) attempts to incorporate the best elements of all three earlier designs. The main view uses a stacked area bar chart over a 24-hour period (by default) to quickly convey the distribution of activities over the course of a day. Unlike earlier designs, the time scale of this graph can be changed using a slider at the top to cover a week, year, or lifetime. Time has a deeply hierarchical nature, and adding this slider lets the user easily explore time at all these scales. Although the stacked nature of a stacked area graph means that activities cannot be compared along a common baseline, we felt that this tradeoff was acceptable given how well the stacked area bar chart conveys the part-to-whole relationship. The number of categories should be small enough (< 15) that easily distinguishable colors can be used for each activity type and activities will not be lost.

Below the stacked area graph is a list of demographics by age, sex, race, employment status, and income. Next to each item is a quick visual summary of the most common activity of the given demographic at a given time of day, using the same horizontal axis as the stacked area bar chart above. Clicking on one of these bars will filter to stacked area bar chart to show only results from that demographic. Clicking again will remove the filter. The horizontal summaries for each demographic, while certainly not comprehensive, provide a way to quickly compare all classes without extra interaction. More detailed comparison is available at the individual activity level.

Hovering over the stacked area graph will display a vertical cursor at a certain time of day. A sidebar displays an exact percentage breakdown for activity distribution at that time of day. If the user is not hovering over the chart, the sidebar simply displays averages for the time period.

Clicking on an area in the stacked area graph will filter the graph to display only that activity, turning the stacked areas into a simple line graph with a single line. This facilitates better comparison of activity distribution over the time period by using a common baseline. Additionally, the summary bars for each demographic at the bottom are replaced by a horizontal bar whose length represents the number of hours per day that the demographic spends on the selected activity. Again this design facilitates direct comparison between demographics in the most perceptually accurate way by using a length encoding on a common baseline.

A checkbox is displayed next to each demographic. When the box is checked a new line is added to the line graph specifically for that demographic. The color of the line matches the color of the horizontal bar, creating an implicit legend. This allows the user to compare certain demographics for a given activity across the time period on a common baseline.

Finally, hovering over the line graph displays a vertical cursor and similar sidebar information as the stacked area graph. Additionally, a US map is displayed when zero or one demographics are selected. The coloring of the map represents the distribution of the activity over the country at the given point in time, or on average if the user is not hovering over the graph.

Must-have Features

- Stacked area chart that displays the percentage of respondents doing a given activity throughout a day
- 24-hour breakdown of the activities each demographics is doing during each block of time
- Selectable demographics to compare how each demographics’ stacked area chart compares to the entire data set and other demographics
- Line graph for each activity, accessible by clicking the activity on the stacked area chart
- Average number of hours per day per demographic based one which activity line graph is showing
- Checkboxes for each demographic to add to the line graph for any activity
- US map that is gradient shaded based on which states (or counties) report the highest average number of hours

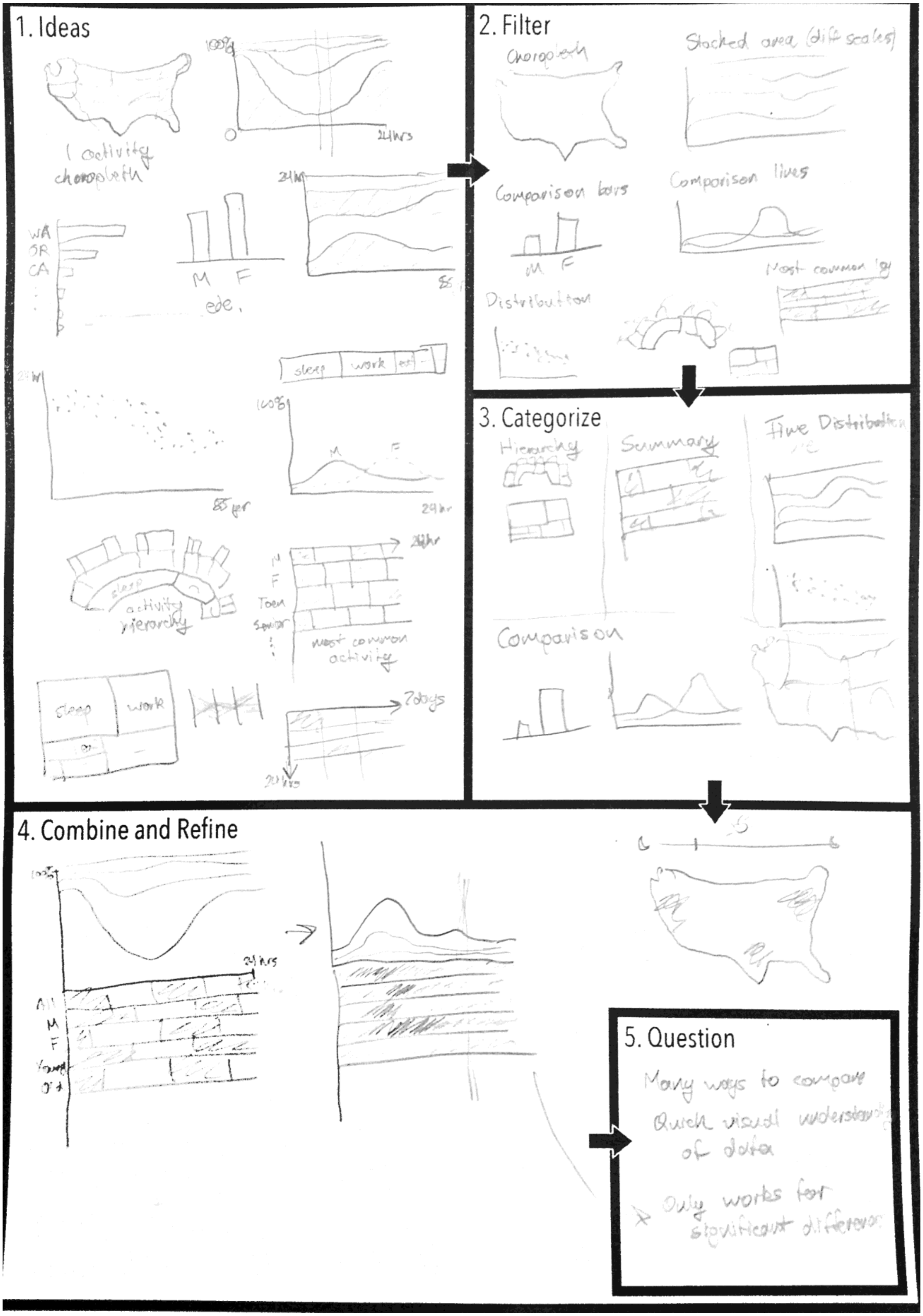
Optional Features

- Selector for different time scales for the stacked area chart (week, year, and lifetime)
- Hover vertical cursor and expanded info side panel with exact data values for stack area chart
- Hover vertical cursor for the line graph with expanded info side panel and real-time updates to the US map
- User survey to input your own data and see what demographic your data fits best into

Project Schedule

Our project schedule is broken down weekly. Dates indicate deliverables for that date.

- November 1: Project proposal completed.
- November 8: Complete preprocessing tools completed and non-interactive d3 implementation.
- November 15: Integrate real data with d3, add interactivity, tweak as needed.
- November 22: Collect feedback from friends and classmates on prototype.
- November 29: Integrate feedback, polish prototype visualization, collect more feedback in class.
- December 6: Integrate feedback, prepare final website.
- December 13: Polish process book, prepare final webcast, present in class.



1. Ideas

2. Filter

3. Categorize

4. Combine and Refine

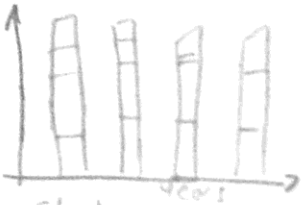
5. Question

Many ways to compare

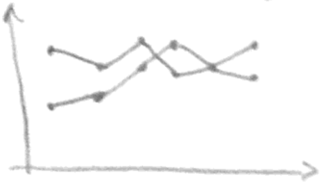
Quick visual understanding of data

Only works for significant differences

1. Ideas

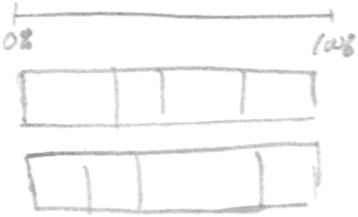


Stacked bar chart activities throughout time

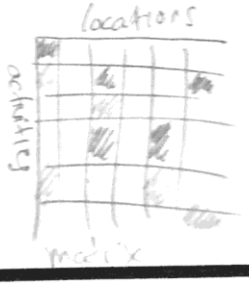


Scatter plots to compare classes

pie chart



Stacked horizontal bar charts for ~~different~~ locations



2. Filter

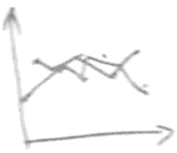
stacked bars are similar and differ only in data sets

pie chart is redundant to horizontal bar

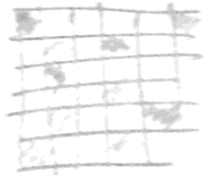
3. Categorize

percent based

comparison between classes



Comparison large data

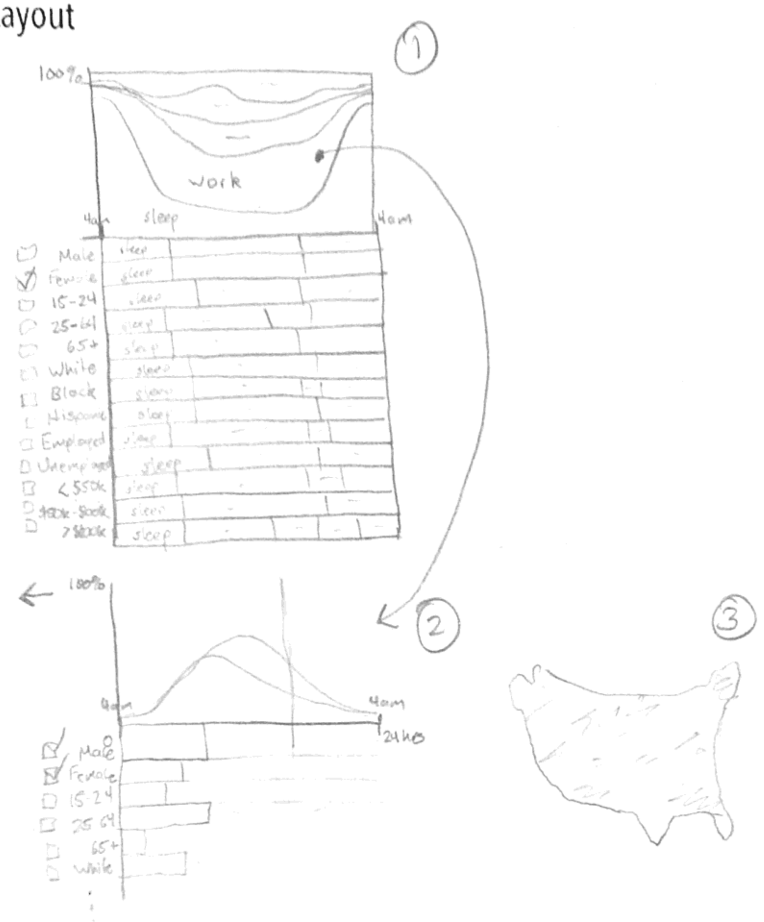


4. Combine and Refine

5. Question

yes?

Layout



Title:

Auther: Ben McMorran

Date: Oct 31 2016

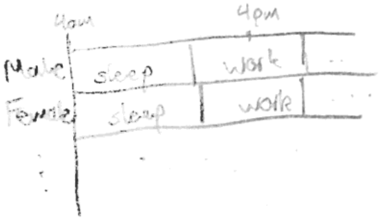
Sheet: 1

Task:

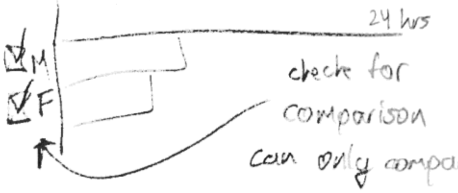
Operations

Click on activity in ①, animate to selected activity ②. Filter in ① w/ checkboxes, Compare in ② w/ checkboxes, Slide cursor over ② to animate ③ (if only 1 or zero comparisons selected), Back arrow to return to ①

Focus



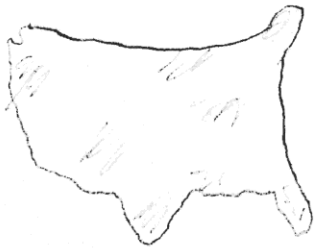
shows most common activity at time of day for class



check for comparison

shows # of hrs/day on activity

can only compare within type (age, sex, etc



shaded for avg day normally if 1 comparison checked then shading is filtered shows hour of day as cursor slides across ②

Discussion

- most common activity may be similar enough that bars under ① aren't interesting

+ comparison between classes on shared axis

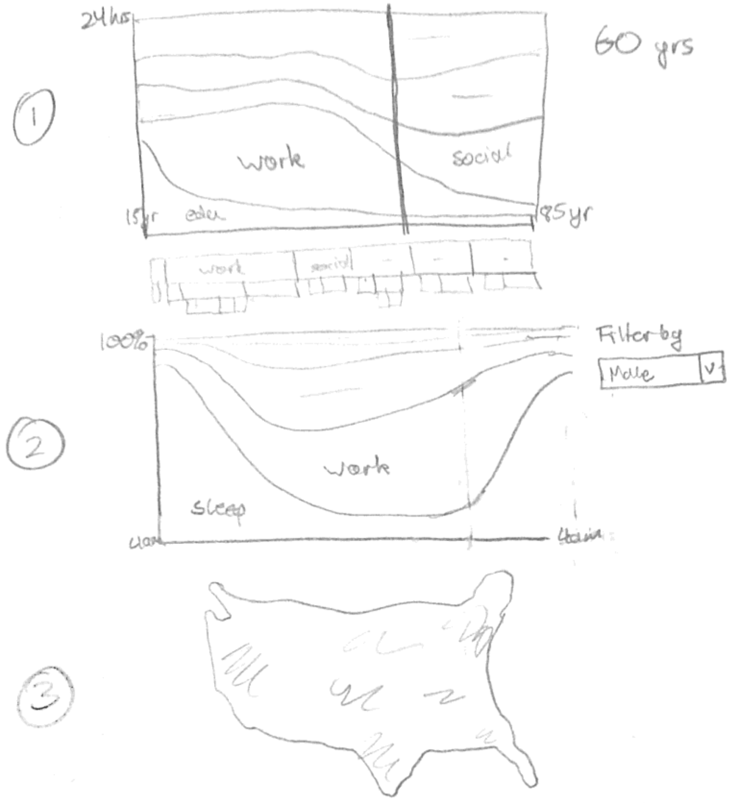
- map only for display, not interactive

+ stacked area reduces to single area for better readability

- age is ratio, but represented in discrete (large) bins

+ compact comparison of all classes

Layout



Title:

Auther: Ben McMorran

Date: 10/31/16

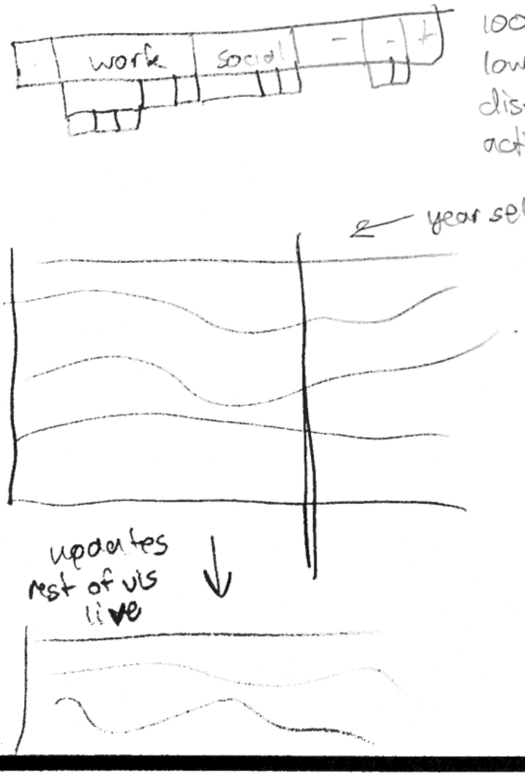
Sheet: 2

Task:

Operations

Slide across ① to select age. Hierarchy distribution updates, Dropdown in ② to filter day distribution to certain class. Click activity in ② to show distribution in ③. Slide across ② to show certain hour.

Focus



100% stacked bar lower bars show distribution within activity hierarchy

Discussion

- hard to compare across classes

+ conveys continuous nature of age

+ full hierarchy of activities

- all comparisons between activities are stacked. Unaligned worse than aligned.

Title:

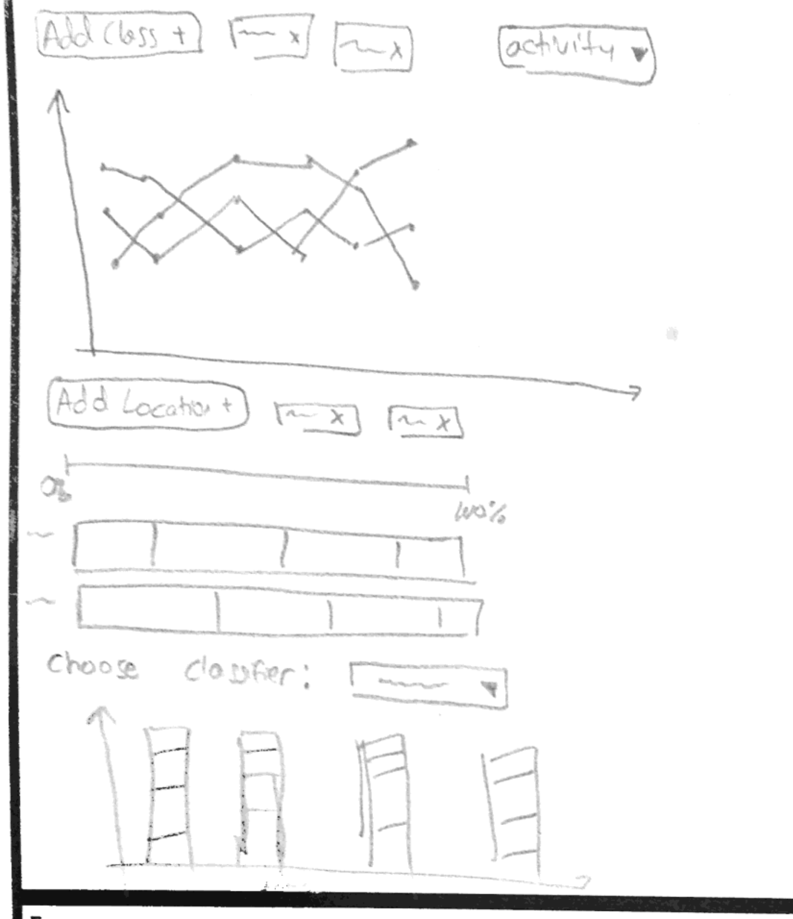
Auther: Francisco Sanchez

Date: 10/31/16

Sheet: 3

Task:

Layout



Operations

Add Class +

Button allows uses to select a class to add to the scatterplot

Activity ▾ allows user to select activity for comparison.

Add Location +

similar button to add class

classifier ▾

allows user to track a certain subgroup through the years

Focus

Each set of visualizers features on interactive set of classifers to compare classes and characteristics to one another.

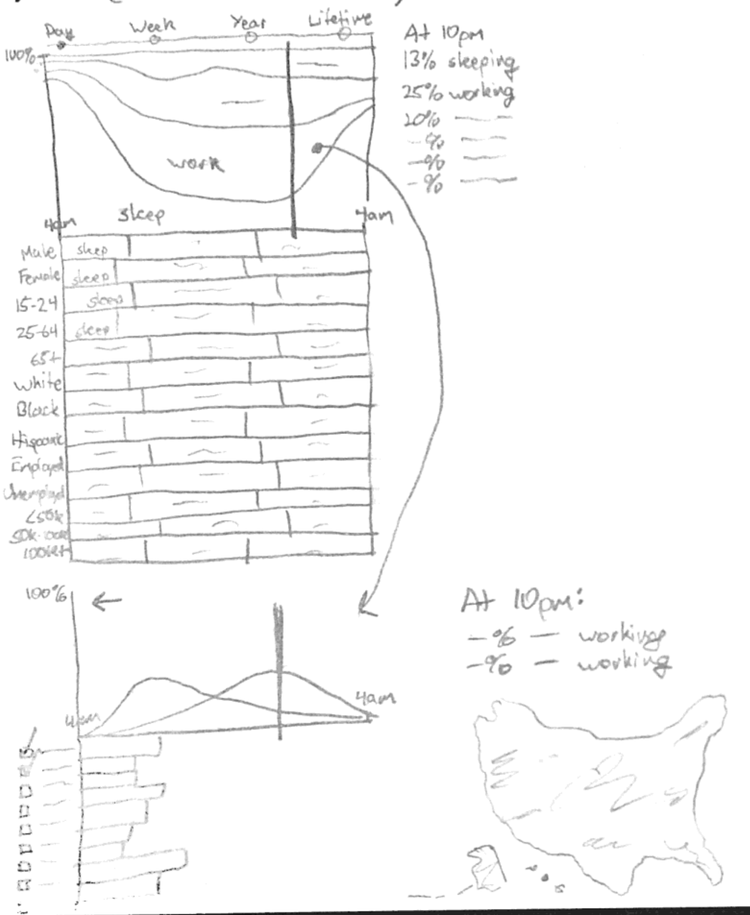
Discussion

+ allows deep user customization

+ can show many comparison on a single page

- some insights may be lost.

Layout (similar to sheet 1)



Title:

Auther: Ben & Francisco

Date: 10/31/16

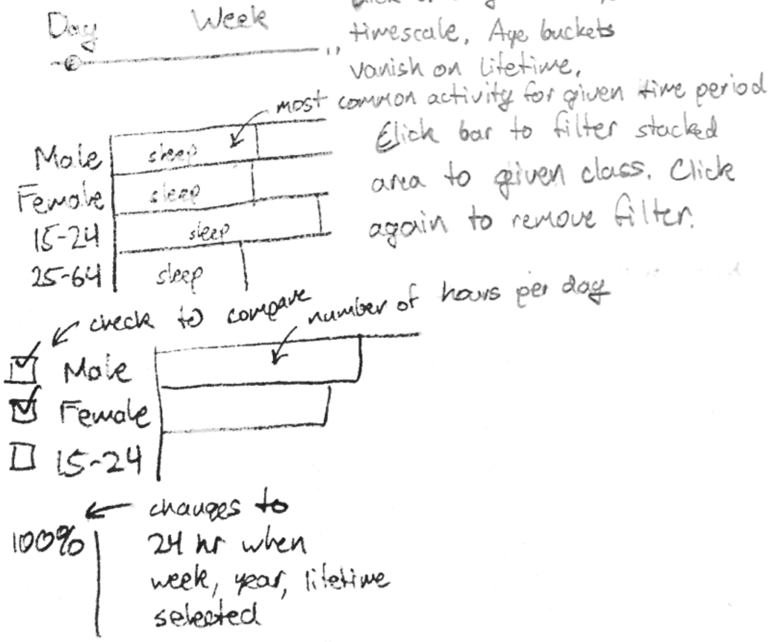
Sheet: 4

Task:

Operations

Hover over main chart to get % breakdown @ point in time.
Change timescale at top.
Click bar to filter by class.
Click activity to remove all other activities
Check classes to compare
Slide over single activity to update % breakdown and map

Focus



Click or drag to change timescale. Age buckets vanish on lifetime.

most common activity for given time period

Click bar to filter stacked area to given class. Click again to remove filter.

check to compare

number of hours per day

changes to 24 hr when week, year, lifetime selected

Detail

Class-level summaries may look silly at anything other than Day. Check w/ real data.

Need real-time interactive updates for sliders to be effective

Preprocessing to reduce respondent level data to summaries