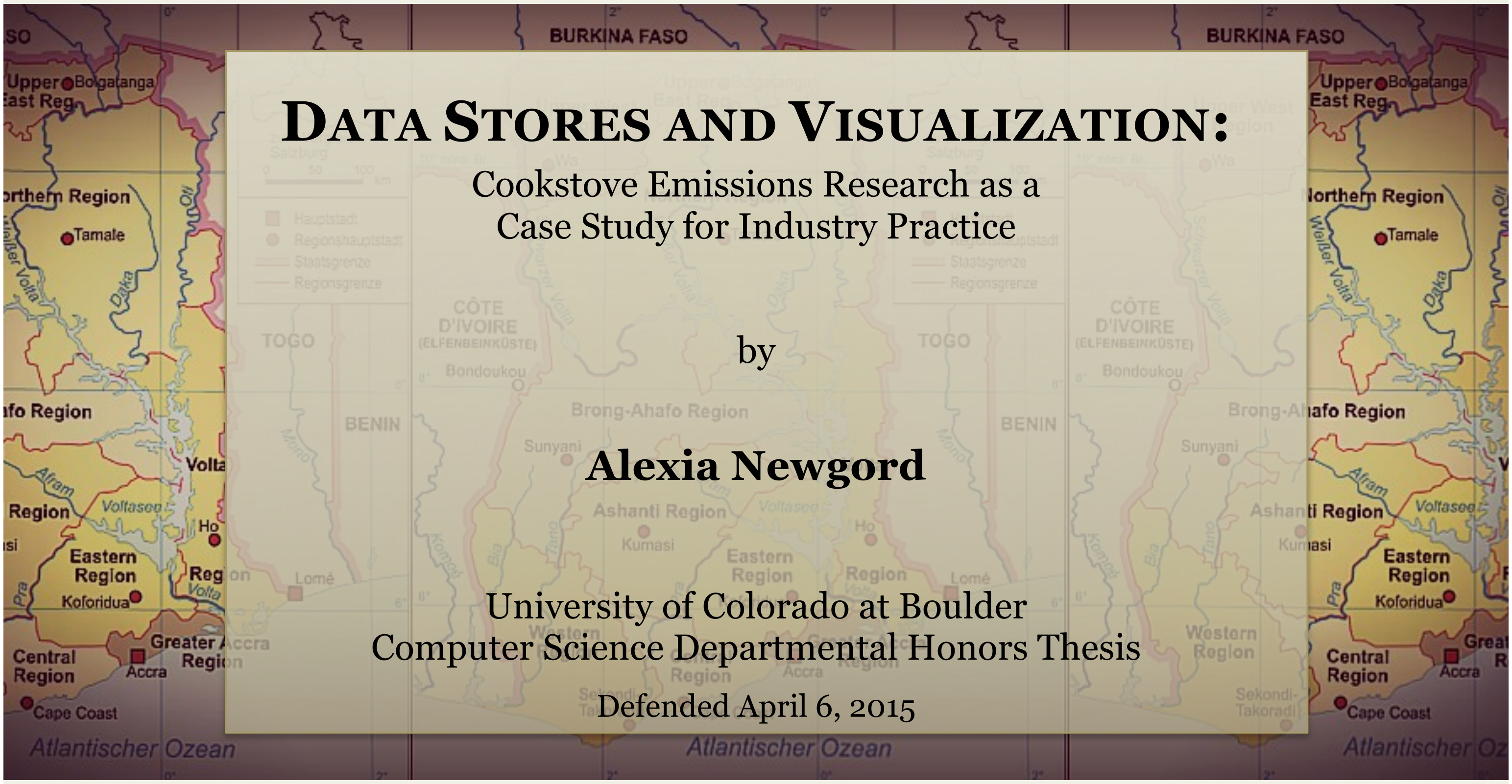


Data Stores and Visualization:

Cookstove Emissions Research as a Case Study for Industry Practice

by

Alexia Newgord

University of Colorado at Boulder

Computer Science Departmental Honors Thesis

Defended April 6, 2015

Advisor:

Kenneth Anderson, Associate Professor and Associate Chair, Computer Science

Thesis Committee:

Kenneth Anderson, Associate Professor and Associate Chair, Computer Science

Clayton Lewis, Professor, Computer Science, Honors Council Representative

Rolf Norgaard, Interim Associate Director, Program in Writing and Rhetoric

Table of Contents

ABSTRACT

Research of Emissions, Air Quality, Climate, and Cooking Technologies in Northern Ghana (REACCTING) is a group composed primarily of researchers from the National Center for Atmospheric Research, the University of Colorado at Boulder, and the Nvrango Health Research Center. Recent reports indicate that "nearly three billion people in the developing world cook food and heat their homes with open fires or cookstoves that are fueled by solid biofuels." The smoke exposure from these methods contributes heavily to global emissions and is estimated to cause four million premature deaths annually. REACCTING has been collecting vast amounts of data in search of effective alternatives for these activities and requested student assistance in managing and visualizing the data. The processing, cleaning, storage, and visualization of the REACCTING data presented technical and human challenges that are common in industry practices.

This paper primarily discusses the technical strategies and design considerations for data visualization, within the context of the data engineering lifecycle. The process of working with the REACCTING data was dynamic and the end product was an explanatory web application that presents a dataset in the form of an interactive animated map. Although the REACCTING team was already using both free and proprietary high-level software for advanced statistical, geographical, and graphical analysis, this mapping "quick-look" tool served the explanatory purpose of creating an overview of the subject's location and how that relates to exposure. The technical outcome can be used for a more comprehensive understanding of the data and, from a developer's perspective, can be applied in concept to other data problems.

Keywords

Visualization, Data Systems, Data Store, Data Engineering, Web Application, D3, Leaflet, Mapping, Open Source, Cookstove, Google Maps

I. INTRODUCTION

The visualization of information has supported human survival by enabling us to share critical data that cannot be as effectively communicated through other means. Cartography, body language, and line drawings are some of the most rudimentary examples of human dependence on visual cues. Graphic forms of information have been embedded in the threads of human communication as early as the Upper Paleolithic [1], and some even argue that up to one quarter of a single human brain is involved in visual processing, more than any other sense [2].

Since the mid-1900s computer technology has facilitated the collection, storage, and analyses of data quantities far beyond the grasp of a single human brain. Coined "big data," these data stores can lead to novel insights and increased exploratory power. However, the primary challenge that is presented by such large and diverse datasets is the ability to compress or abstract the information into human-understandable representations, without losing the true meaning of the data itself. The end goal is to be able to accurately interpret big data as if it were small data (human readable).

With this objective in mind, we can dynamically construct data systems in a way that limits the expensive overhead of "useless" data measurements and utilizes advancements in technologies and algorithms. Rather than approaching visualization and analysis as the end product for a measured action, an entire system must be carefully established for efficient and scalable data analysis.

Research in Emissions, Air Quality, Climate, and Cooking Technologies in Northern Ghana (REACCTING) is one example of an organization that uses the collection and analysis of vast measurements as a means to verify or refute a hypothesis. Composed of a team of University of Colorado/National Corporation of Atmospheric Research students and professionals, the REACCTING project investigates the substantial impact of cookstove-related emissions on human health and the environment. By compiling a large quantity of diverse data over time, such as demographic info, pollution measurements, GPS coordinates, and cookstove status, the group expects to identify novel correlations between the data sets and propose both practical and effective solutions for what is a well-known problem in this region and around the world [3].

Given this data, how can we develop a data management system that can lead to meaningful analysis and visualization? REACCTING can be interpreted as a case study for related challenges presented in industry. Therefore, by seeking to address potential issues throughout all stages of REACCTING's data engineering lifecycle, we may identify potential issues and solutions that can be applied in similar situations.

II. Data Lifecycle

By evaluating data system design, we can identify successes, failures, and innovations throughout the various stages of development. Iterations of the data engineering lifecycle enable the human mind to better navigate and interpret the realities that are abstracted, as shown in Figure 1.

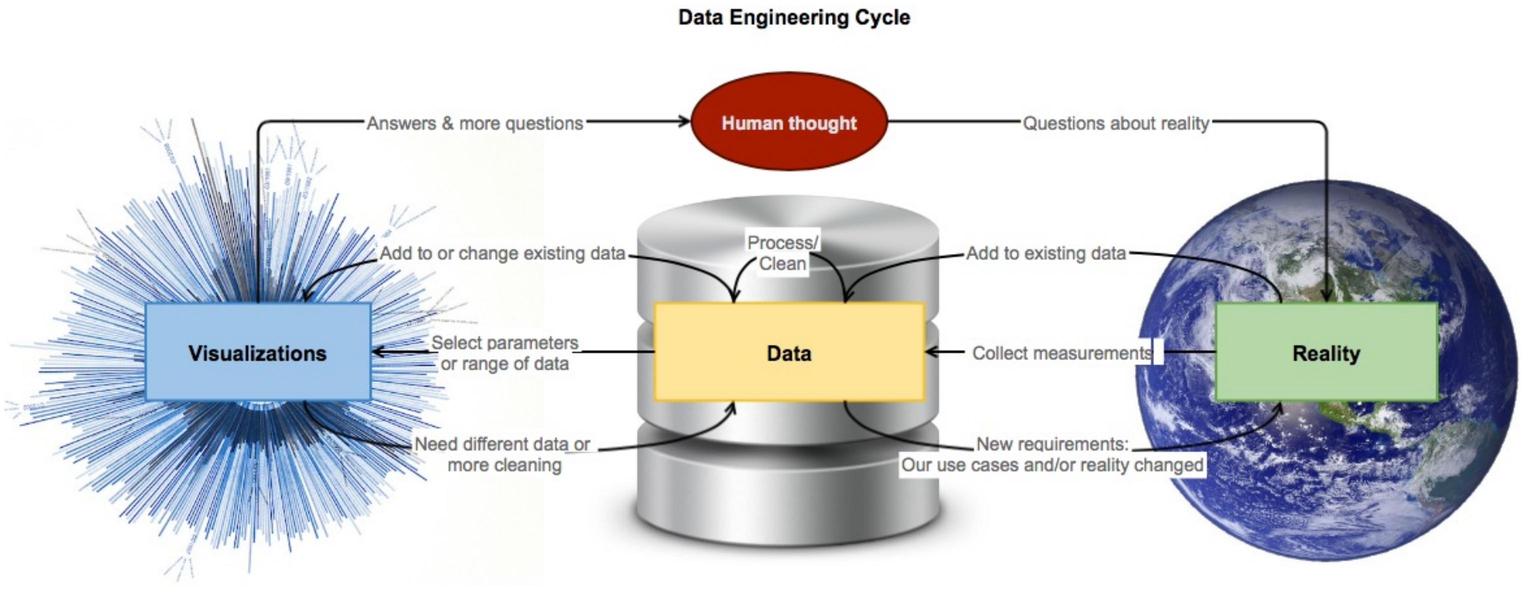


Figure 1. Diagram of the data engineering cycle.

Data is collected in response to some sort of question about our understanding of reality. Once the appropriate resources are accumulated, data that is hypothesized to provide insight to this topic is collected and stored as a new dataset or aggregated with existing related data. The method of storage can range from SQL/NoSQL databases to text files, but depends on the quantity, format, relationships, and objectives of the data collected. Highly complex data stores will likely be distributed over “clusters,” but this concept goes beyond the scope of this project.

Once the data is collected, it needs to be processed. Data processing can be loosely defined as the “collection and manipulation of items of data to produce meaningful information” [4]. More specifically, this has to do with activities that structure or clean the data, such as sorting, documentation, classification, and aggregation.

Cleaning the data comes into play when there is some sort of issue, such as corruption or “noise”, occurs during the data collection stage. Missing, duplicate, or inaccurate data are common concerns that need to be handled in a way that is appropriate for the intentions and integrity of the data. The data processing stage is also the ideal time to apply various algorithms or statistical analysis to the data. This may be as simple as formatting a timestamp or as complex as applying advanced machine learning techniques.

In many systems, the datasets will be processed and stored at various “levels.” The processing that occurred between each level is documented and, if a method or algorithm is later proven to be invalid, we can “undo” it by reprocessing the previous level.

Once the data is processed, it can be visualized. Visualizing the data may be the end goal for some systems and is a method for communicating correlations or patterns in the data. It also provides insight on further issues with the data, such that the cycle often returns to the processing or even collection stages.

Ultimately, the number of iterations depends on a variety of factors, such as the level of success at each stage and the degree of quality expected from the outcome. The process is also dependent on the project’s measureable resources, such as funding, human resources, and timeline.

A data system can be more than simply a machine that pumps out discrete visualizations based on a given input. Rather, the concept of data engineering defines a cyclic relationship between data abstraction, processing, and our understanding of reality. By embracing this cycle through methods like Agile, the longevity and value of data can be increased.

After the lifecycle is considered complete, there are additional considerations regarding data curation and sharing to be made. If others wish to access your data, either for their own research or peer review yours, is there a means for them to do that? It is increasingly common for organizations or individuals to sell/share data, although it is important that it is cited.

III. Visualization Design

The visualization of data is a means of translating quantitative information into a seeable and human-understandable format. Visualizations have been used throughout most of human history in support of improved interpersonal communication and cognitive information processing.

The earliest known visual art appeared around the time of the development of anatomically modern humans, circa 40,000 BCE, in the form of large ceremonial painting installments or portable sculptural objects. One of the oldest extant works is the painted cave, the "Great Hall for the Bulls," in modern-day Lascaux, France. Using strong outlines and soft colors, the painted images portray a landscape with various animals and a single human. Although these images were arguably used for ceremonial purposes, portions are

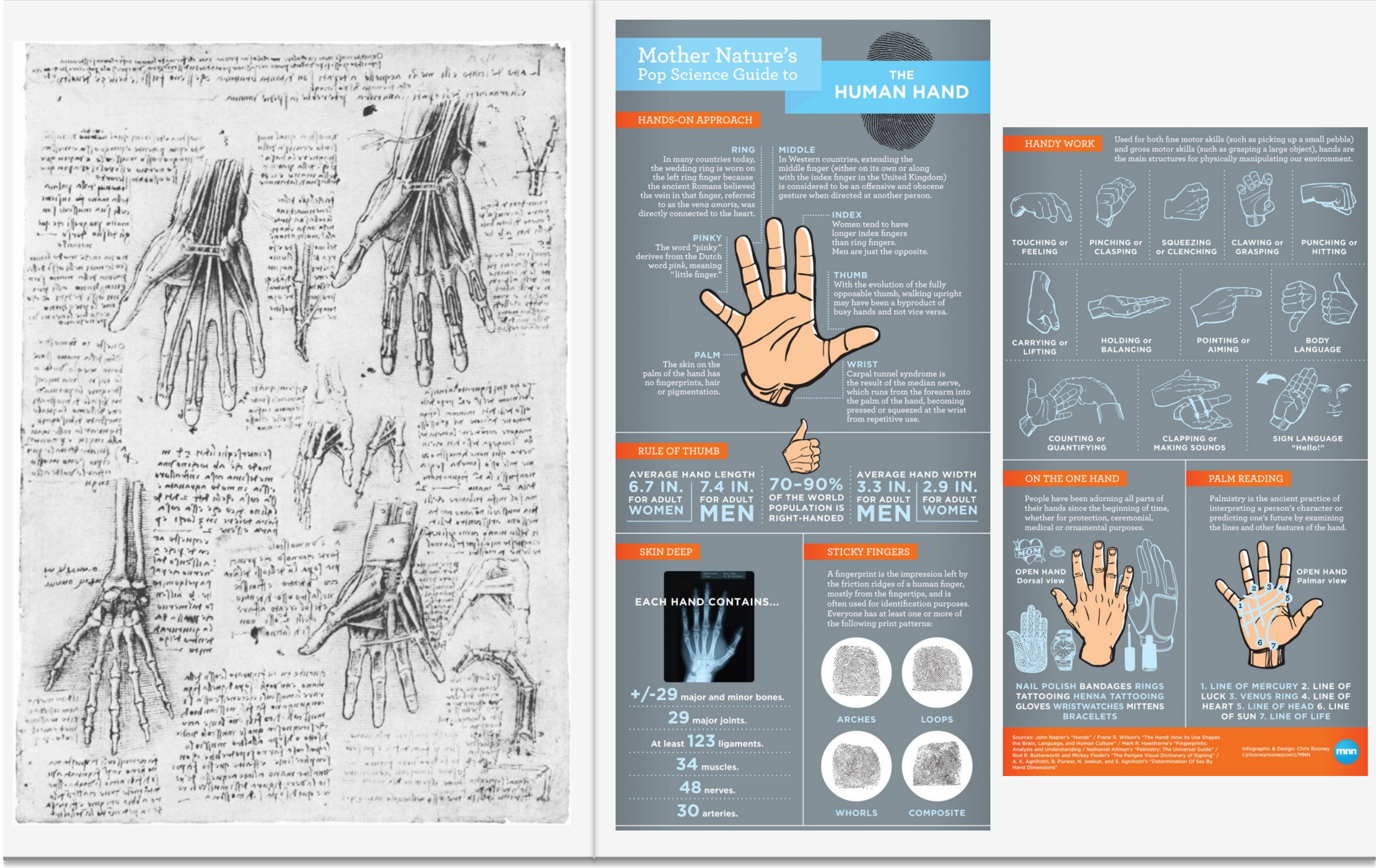
Figure 2. "The Unicorn," Great Hall of the Bulls in Lascaux cave, ca. 15,000 BCE [5].

superimposed over earlier depictions, suggesting that the artist was focused on “the act of portraying the animals rather than in the artistic effect of the final composition” [5].

Despite the unknowns surrounding this work, the composition clearly provided information to humans in ways that certainly could not be replicated with written or oral communication, especially since verbal and written communication had not yet evolved into the same level of complexity that we see today.

In addition to the enrichment of interpersonal communication, visualizations may enhance human information processing and memory retention by reducing the “cognitive load” [6]. The verbal-visual visualization format particularly demonstrates this objective. For example, Leonardo Da Vinci employed a combination of detailed drawings and encrypted words in his anatomical journals to track and record his ground-breaking observations of human anatomy. Da Vinci’s notebooks share remarkable similarities to popular modern infographics, testifying for this style’s effectiveness over time (Figure 3).

The verbal-visual format creates an accessible narrative and caters to simultaneous processing as opposed to sequential processing. This is particularly useful for modern web-browsing, where users typically navigate rapidly through pages; studies suggest that approximately one half of page visitations last for twelve seconds or less [7].


Mother Nature's Pop Science Guide to THE HUMAN HAND

HANDS-ON APPROACH

RING In many countries today, the wedding ring is worn on the left ring finger because the ancient Romans believed the vein in that finger, referred to as the *vena amoris*, was directly connected to the heart.

MIDDLE In Western countries, extending the middle finger (either on its own or along with the index finger in the United Kingdom) is considered to be an offensive and obscene gesture when directed at another person.

INDEX Women tend to have longer index fingers than ring fingers. Men are just the opposite.

PINKY The word "pinky" derives from the Dutch word *pink*, meaning "little finger."

THUMB With the evolution of the fully opposable thumb, walking upright may have been a byproduct of busy hands and not vice versa.

PALM The skin on the palm of the hand has no fingerprints, hair or pigmentation.

WRIST Carpal tunnel syndrome is the result of the median nerve, which runs from the forearm into the palm of the hand, becoming pressed or squeezed at the wrist from repetitive use.

RULE OF THUMB

AVERAGE HAND LENGTH 6.7 IN. FOR ADULT WOMEN | 7.4 IN. FOR ADULT MEN

70-90% OF THE WORLD POPULATION IS RIGHT-HANDED

AVERAGE HAND WIDTH 3.3 IN. FOR ADULT MEN | 2.9 IN. FOR ADULT WOMEN

SKIN DEEP

EACH HAND CONTAINS...

+/-29 major and minor bones.

29 major joints.

At least 123 ligaments.

34 muscles.

48 nerves.

30 arteries.

STICKY FINGERS

A fingerprint is the impression left by the friction ridges of a human finger, mostly from the fingertips, and is often used for identification purposes. Everyone has at least one or more of the following print patterns:

HANDY WORK

Used for both fine motor skills (such as picking up a small pebble) and gross motor skills (such as grasping a large object), hands are the main structures for physically manipulating our environment.

ON THE ONE HAND

People have been adorning all parts of their hands since the beginning of time, whether for protection, ceremonial, medical or ornamental purposes.

NAIL POLISH BANDAGES RINGS TATTOOING HENNA TATTOOING GLOVES WRISTWATCHES MITTENS BRACELETS

PALM READING

Palmistry is the ancient practice of interpreting a person's character or predicting one's future by examining the lines and other features of the hand.

1. LINE OF MERCURY 2. LINE OF LUCK 3. VENUS RING 4. LINE OF HEART 5. LINE OF HEAD 6. LINE OF SUN 7. LINE OF LIFE

Sources: John Napier's "Hands" / Frank R. Wilson's "The Hand: How its Use Shapes the Brain, Language, and Human Culture" / Mark R. Hawthorne's "Fingerprints: Analysis and Understanding / Nathaniel Altman's "Palmistry: The Universal Guide" / Rod R. Butterworth and Mickey Flodin's "The Perigee Visual Dictionary of Signing" / A. K. Agnihotri, B. Purwar, N. Jeebun, and S. Agnihotri's "Determination Of Sex By Hand Dimensions"

Infographic & Design: Chris Rooney (@looneyrooneycom)/MNN

Figure 3. Leonardo da Vinci [8] (pg 118), early 1500's CE and "The Human Hand" [9]

A. Principles of Analytical Design

In the organization of such effective visualizations, the revered American statistician and artist, Edward Tufte, proposes six principles of analytical design: comparisons, structure/explanation, integration of evidence, multivariate analysis, documentation, and content [10]. In this section, I will briefly analyze these six principles using my own observations and then propose additional considerations.

Principle 1: Comparisons

There is little function in the analysis of data that does not change. It is only by identifying patterns in differences that we can validate or develop hypotheses.

For example, in 1608, the telescope was discovered in Holland by the optician Jean Lippersheim. Individuals quickly observed the existence of sunspots and began formulating and modifying theories on their formation and the sun's axial rotation. In 1610, Galileo began creating daily sketches of the sun, recording his personal observations of sunspot movements over time. Eighteen months later, Galileo's findings were published in the *Istoria e Dimostrazioni Intorno Alle Macchie Solari e Loro Accidenti Rome* (History and Demonstrations Concerning Sunspots and their Properties) [11].

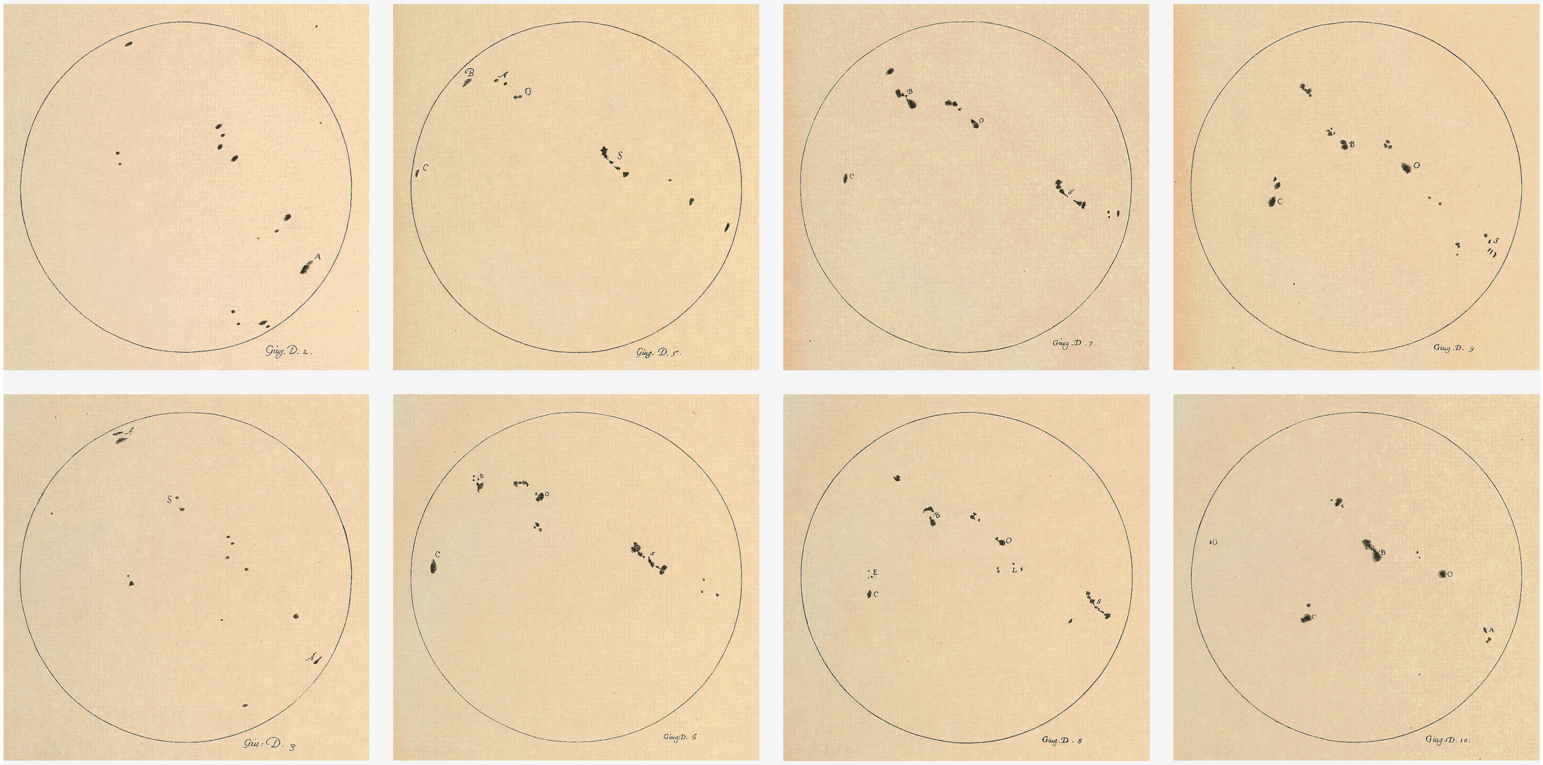


Figure 4. Six days of sunspot observations, Galileo, 1613 C.E [11].

Because Galileo made his observations at approximately the same time every day, the motions of the spots are very apparent. In a “flip-book” style animation where one image quickly and sequentially replaces the previous, the sunspots fluidly shift across the plane, giving the impression that the sun is, in fact, rotating [12].

Animations, or visual data represented over time, can identify relationships between the lapse of time and other variables. Galileo’s persistent and accurate drawings presented a definite correlation between the geographic location of the sunspots and the moment in time. This insight allowed him to correct others’ erroneous theories on the sun’s rotation [11].

With innovations like slideshows, film, and computers, animations are common and relatively easy to represent. A modern example of this is Cameron Beccario’s visualization of global wind, weather, and ocean conditions (Figure 5). The variables are updated dynamically, ranging from every three hours for the weather conditions, to every five days for the ocean surface current estimates [13]. The animation uses the JavaScript D3 library to dynamically represent wind patterns in the form of moving line patterns across the Earth. Wind patterns that are more intense will shift to the right on a

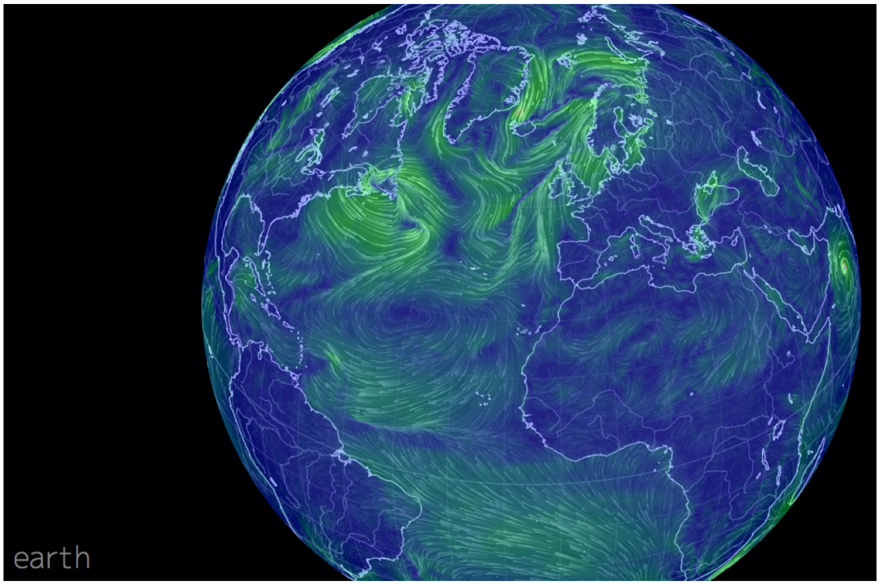


Figure 5. Screen capture of wind animation on (10-27-2014, 12:50pm MST) [13].

green to red spectrum and the user is able to click and drag to see different views of the Earth. Much like Galileo's sunspots, this visualization ties together the elements of geography and time. Viewers can easily compare the movement of weather patterns and identify regions on the Earth that are experiencing notably tumultuous weather.

Principle 2: Causality, Mechanism, Structure, Explanation

A customized visualization serves as a significant explanatory mechanism. In the process of constructing it, we can consider who the audience is and what the extent of their existing knowledge is about the dataset. What do they already assume? The focus of the correlations should attempt to identify or highlight some point or causality.

This requires knowing something about your user, which may not be an easy task. Organizations frequently collect IP addresses, cookie info, and survey data to understand more about the client, but this too may create a complex picture. For example, a study performed by ComScore in 2009 suggested that for political campaigns, up to eighty percent of advertisers' impressions were "delivered either to the wrong segments in the U.S., or to consumers outside the U.S." [14]. A visualization that is designed for a particular audience but ends up in the hands of someone with a different background or perspective may result in miscommunication.

Additionally, the user will probably make some assumptions about the visualization itself. For example, most viewers of a pie chart would assume that the sections add up to a single whole, or 100% in the case of percentages. The chart in Figure 6 is confusing because it gives the appearance that each of the three candidates for the 2012 Republican primaries has about a third of the support, which is untrue [15].

Perhaps a bar graph would have been a more appropriate graphing mode. However, even with basic line and bar graphs, comparisons are often exaggerated or made misleading by truncating or flipping the axes.

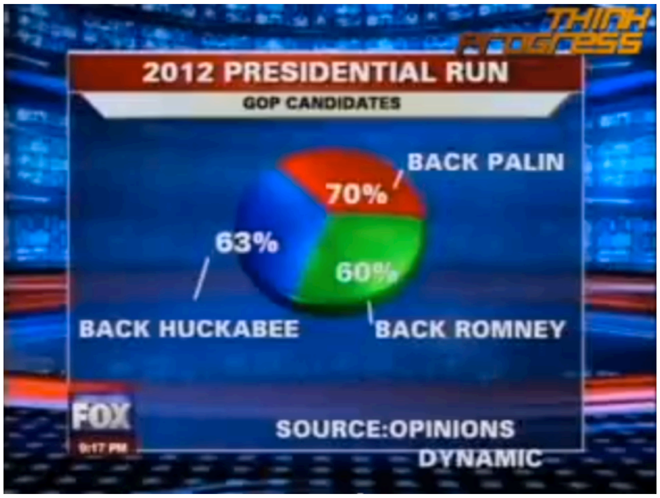


Figure 6. Pie chart that Fox Chicago aired during the 2012 primaries [15]

Principle 3: Multivariate Analysis

The most elementary graph shows the relationship between just two axes or variables, but through careful design decisions, visualizations may represent correlations between multiple dimensions and variables. Charles Minard’s classic statistical map of the “successive losses in men of the French Army in the Russian Campaign 1812-1813”, visually documents the many varying factors, like temperature, time, and distance, that contributed to the deaths of 412,000 of the 422,000 men of Napoleon’s army (Figure 7).

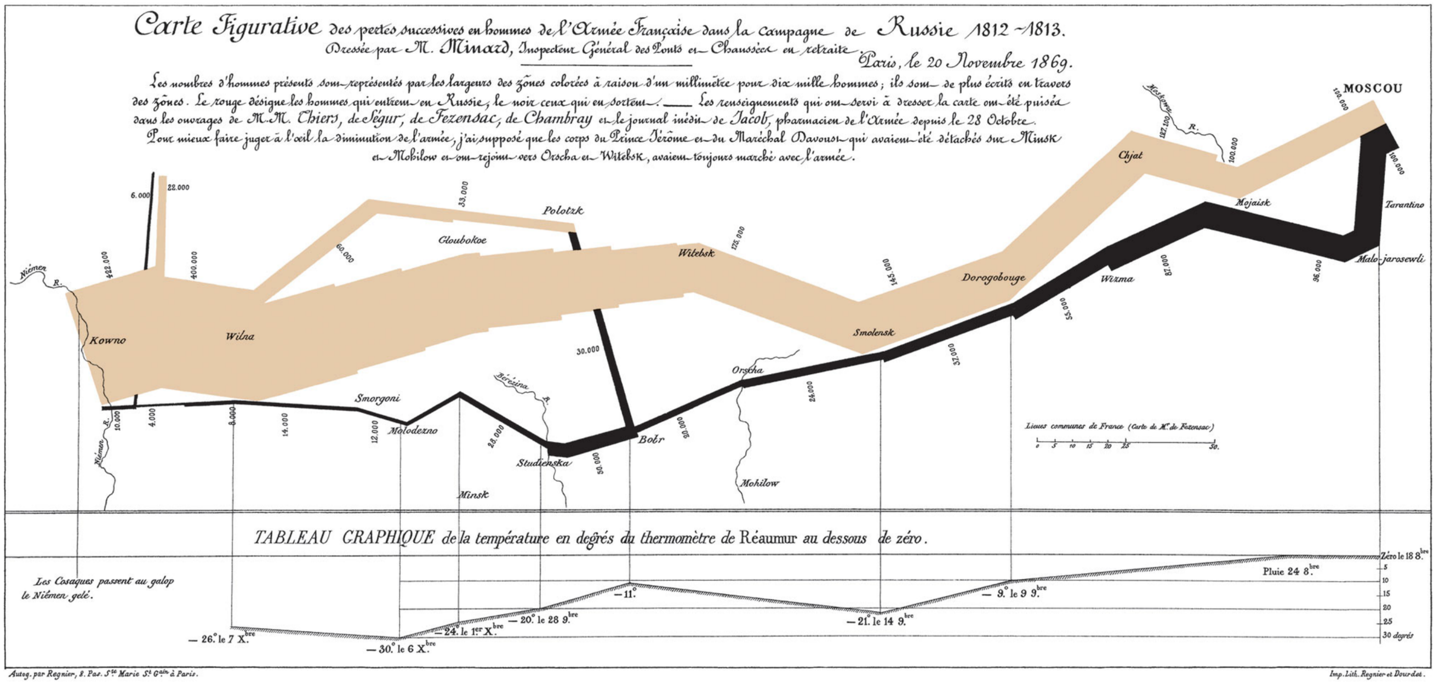

Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
Paris, le 20 Novembre 1869.
Les nombres d'hommes présents sont représentés par les largeurs des zônes colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zônes. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. — Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.
Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mobilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.
MOSCOU
Niemen R.
Kowno
Wilna
Gloubokoe
Polotzk
Witebsk
Smolensk
Dorogobouge
Chjat
Mojaisk
Wizma
Tarantino
Malo-jarosewli
Moskowa R.
Smorgoni
Molodezno
Minsk
Studienska
Bobr
Orscha
Mohilow
Berezina R.
422.000
400.000
22.000
6.000
60.000
33.000
175.000
145.000
127.100
100.000
100.000
100.000
96.000
87.000
55.000
37.000
24.000
50.000
30.000
28.000
20.000
12.000
14.000
8.000
4.000
10.000
Lieues communes de France (Carte de M. de Fezensac)
TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.
Les Cosaques passent au galop le Niémen gelé.
Zéro le 18 8.bre
Pluie 24 8.bre
— 9.° le 9 9.bre
— 21. le 14 9.bre
— 11.°
— 20.° le 28 9.bre
— 24. le 1.er X.bre
— 30.° le 6 X.bre
— 26.° le 7 X.bre
5
10
15
20
25
30 degrés
Autog. par Regnier, 8. Pas. St. Marie St. Gain à Paris.
Imp. Lith. Regnier et Dourdet.

Figure 7. Charles Minard's map of Napoleon's disastrous Russian campaign of 1812 [16].

The challenge for multivariate analysis is selecting relevant variables and parameters and limiting the scope of these parameters appropriately.

Principle 4: Integration of Evidence

Integrating evidence together into a single visualization in an aesthetic and understandable way forms the basis of visual design. The key is to balance clutter and content and use the principles of art design as tools in doing so.

Humans have long used hierarchical imagery to indicate the value of certain elements over others. For instance, the Egyptians regularly employed this technique to draw focus to a particular component in visual compositions. An example of this is the *Palette of Narmer* (Figure 8), which contains narrative imagery of the king, animalistic creatures, and slaves. The subject of interest, King Narmer, is clearly represented as the largest figure on the palette, indicating his power over the others [17].

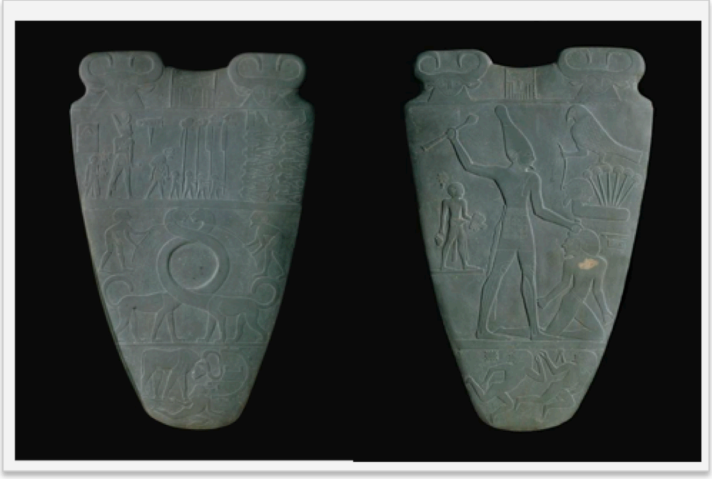


Figure 8. The Palette of Narmer (front and back), 3000 BCE, Slate [17].

Principle 5: Documentation

The principle of documentation emphasizes using credible and honest data sources. Too frequently, visualizations are based upon data that is biased, outdated, or poorly collected. Today, there is a widespread fear that vaccinations among children can cause autism. This belief has been tracked to a 1997 study published by British surgeon, Andrew Wakefield. The article suggested a link between the MMR mumps, measles, rubella (MMR) with autism among British children [18]. However, the article has since been since discredited due to, "serious procedural errors, undisclosed financial conflicts of interest, and ethical violations" [19].

Figure 9 is taken from Wakefield's article and is a prime example of how poor data can lead to a powerfully misleading visualization. Even to someone with little understanding of the negative effects of methylmalonic-acid creatinine, it is clear that the patients vaccinated with MMR supposedly have higher acid values. This was the only graphical figure in the article and Wakefield references it in a single descriptive sentence. This graph is obviously flawed since the data that it is bated on was later proven corrupt, but additionally, the lack of explanation does little to temper any assumptions that the relationship is causal. Regardless of whether Wakefield had ulterior motives or was simply careless, what Wakefield selected to reveal and not reveal ultimately triggered a panic in parents, contributing to decreased vaccinations internationally and, consequentially, an influx in vaccine-preventable diseases [19].

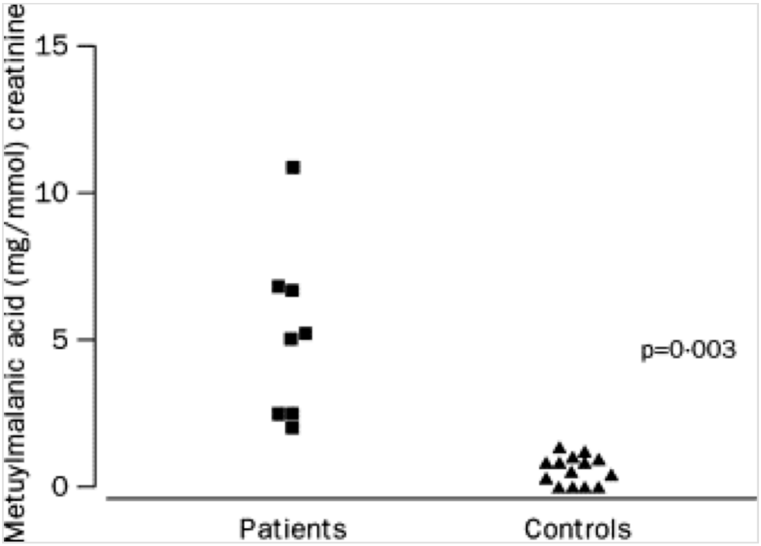


Figure 9. Graph from Wakefield's retracted article that graphs data that has since been deemed inaccurate [18]

Furthermore, the nature and source of borrowed data should be noted in order to avoid similar issues. Data is increasingly recognized as both an owned and sharable entity; citing other individuals' datasets is crucial in order to give credit where it is due and allow thorough peer reviews of these sources.

Principle 6: Content Counts Most of All

The purpose behind the data is what counts. Humans are only interesting in analyzing relevant and novel data.

An example of this is di Giorgio's sketch of a cathedral (Figure 10) [20]. Di Giorgio added value to the visualization by basing it off of the proportions of the "ideal" human being.

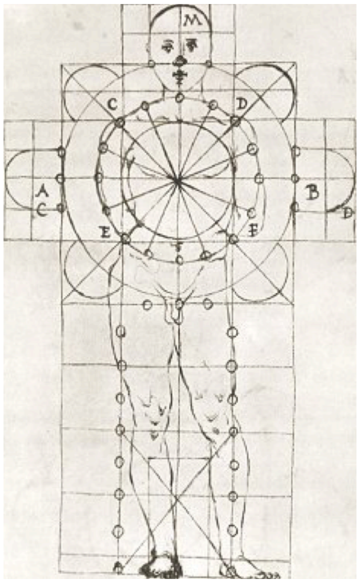


Figure 10. di Giorgio [20].

B. Visualizations as Art

In addition to Tufte's Principles of Analytical Design, it is crucial to consider the well-established basic principles of compositional design: center of interest, balance, harmony, contrast, directional movement, and rhythm [21]. How these elements are approached in the organization of a composition changes the meaning and impact of a piece.

Some designers elevate the importance of the basic principles of art beyond any other consideration. For example, even when the intricacy of the data seems to extend well beyond the scope of human understanding, "data artists" like Jer Thorp use these complex relationships as the inspiration for beautiful abstractions. In Thorp's work, "365/360," it is not clear to the viewer what "facts" or conclusions should be taken from the dataset; rather, the complexity of the data is, in itself, to be appreciated (Figure 11).

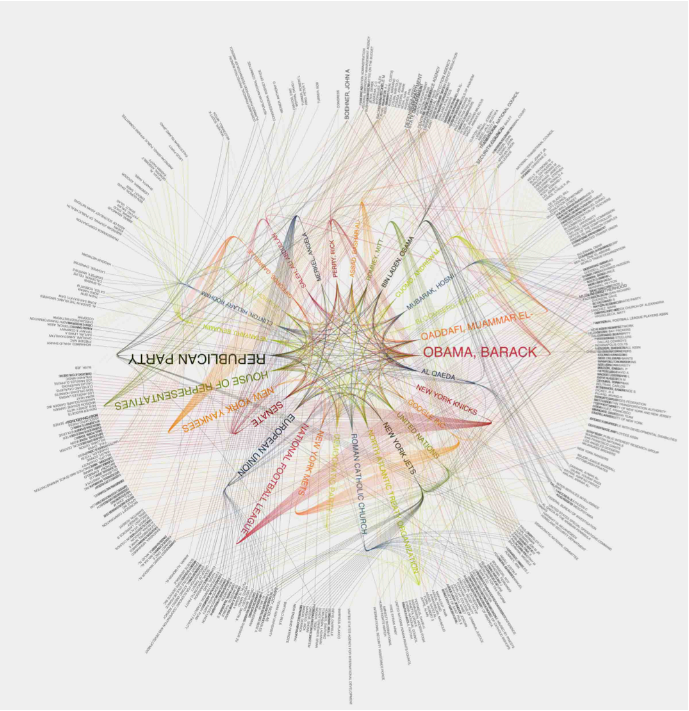


Figure 11. Jer Thorp, *NY Times 365/360*, 2011, Digital Print, 45x44 in. [22].

The complex line patterns and colors represent the seemingly organic relationships between modules, but come together to create a remarkably balanced mandala. Thorp's objective to create a narrative of huge datasets is satisfied by his ability to aesthetically "humanize" the unknown [22].

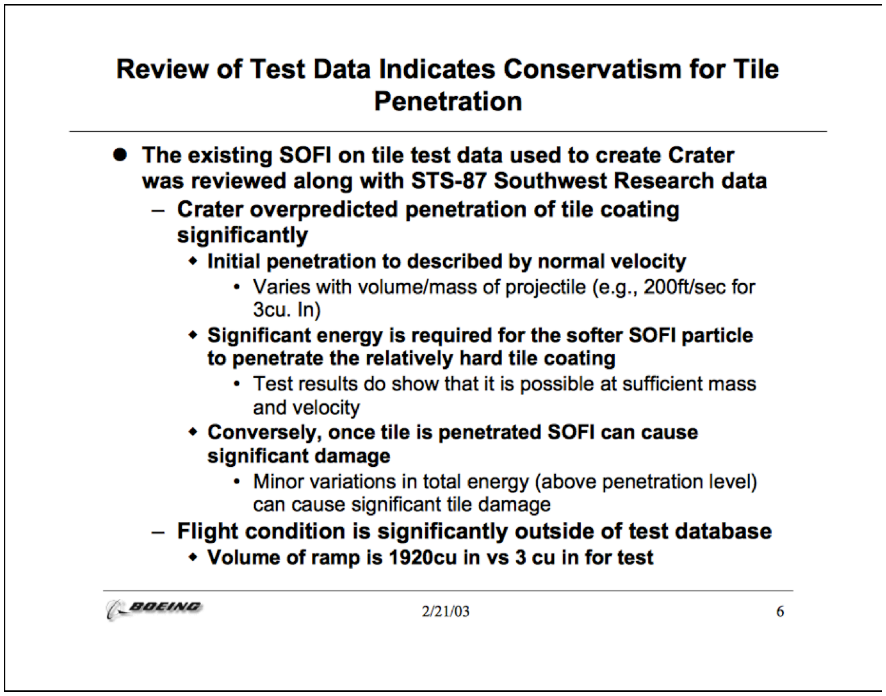
C. When Visualizations Fail

When these principles are not followed, whether intentionally or unintentionally, visualizations may be boring, outdated, or, most importantly, false. Infographics may be used to tell *part* of the story in an attempt to persuade others to buy a certain product or vote for a particular candidate or issue. Numbers give the illusion of objectivity and credibility, even though the means in which they are represented may be actually manipulate a person's interpretation of reality.

Additionally, designers face numerous tangible barriers; the effectiveness of visualizations may be limited by 2-dimensionality, pixel count, screen size, and varying quality human eyesight or cognitive perception. Visualizations mask their own construction in order to cater to varying levels of expertise and intelligence, but misreading or manipulating the given context and demographics of an audience creates ample opportunity for misuse.

In September, Google's Public Sector CTO and "Innovation Evangelist," Michele Weslander-Quaid spoke at the University of Colorado at Boulder about "Creating a Culture of Innovation." In this talk, she mentioned what is coined as the “three barriers to innovation”: culture, policy, and technology [23]. With respect to technology, data visualization should support the flow of information throughout groups and organizations. The failure of visualizations to communicate an accurate and understandable reflection of some reality can lead to poor interpersonal or inter-departmental communication, uninformed decisions, and ultimately more bad technology.

An example of a controversial and potentially misrepresentative visualization is the PowerPoint presentation that was created quickly after the *Columbia* launch (Figure 12) [24]. Since its initial release in 1990, PowerPoint has become a standard software tool used in support of oral presentations. Microsoft estimates that at least thirty million PowerPoint presentations are created daily [25]. “Slides” are presented in



Review of Test Data Indicates Conservatism for Tile Penetration

- **The existing SOFI on tile test data used to create Crater was reviewed along with STS-87 Southwest Research data**
  - **Crater overpredicted penetration of tile coating significantly**
    - **Initial penetration to described by normal velocity**
      - Varies with volume/mass of projectile (e.g., 200ft/sec for 3cu. In)
    - **Significant energy is required for the softer SOFI particle to penetrate the relatively hard tile coating**
      - Test results do show that it is possible at sufficient mass and velocity
    - **Conversely, once tile is penetrated SOFI can cause significant damage**
      - Minor variations in total energy (above penetration level) can cause significant tile damage
  - **Flight condition is significantly outside of test database**
    - **Volume of ramp is 1920cu in vs 3 cu in for test**

2/21/03 6

Figure 12. Slide by Debris Assessment Team in response to debri that struck the *Columbia* during launch [24].

a linear fashion, while the content on the slides is generally displayed in a hierarchical format. PowerPoint provides presenters a relatively easy method for sharing graphics, charts, and other visualizations for a particular lecture. It can be used as a blueprint for the direction of the discussion and can even be used as a reference after the fact.

However, the practice of using PowerPoint to drive technical conversations like that of the post *Columbia*-launch meeting may be questionable. Regardless of the "quality" of slides, PowerPoint does not natively foster a conversation between parties. Namely, "Instead of human contact, we are given human display" [25]. The meeting should have been a forum for discussion and debate as opposed to presentation. Additionally, rather than displaying the detailed quantitative data from an objective standpoint, the bullet points simply summarize the results from a high level and deceptively emphasize the optimistic perspective of the situation. Had the data presentation method been more in-depth and quantitative, there may have been an opportunity for feedback from others.

Although PowerPoint has evolved from a long history of presentation tools and methods, it should only be used when carefully determined to be the best tool for communicating an idea, rather than out of convention. The manner alone in which data is presented can completely alter a person's perception on the data itself. Clearly, if a visualization causes distorted interpretation of the data, then it can be considered a failure.

IV. APPROACH TO COOKSTOVE DATA MANAGEMENT

Naturally, the value of data visualization is positively correlated with the quality of the data on which it is based, and may face an artificial ceiling if poor data collection, cleaning, or storage occurs in earlier stages of the data engineering lifecycle. With these considerations, we can use REACCTING measurements as a case study data system design in industry practice.

Cookstove research has been present in the academic field for some time and is an area of interest for politicians, humanitarians, and environmentalists. According to the Global Alliance for Clean Cookstoves, four million premature deaths occur every year due to smoke exposure from cooking with open fires or traditional cookstoves [26].

REACCTING expands upon existing research by exploring a notably wider variety of measurements in effort to develop a more comprehensive understanding of cooking practices and impacts [27]. Additionally, the product of this research may help identify effective and feasible alternatives to current cooking methods in Northern Ghana and elsewhere. For this reason, REACCTING has a wide variety of measurements pertaining to:

- Cooking behavior
- Cooking emissions measurements
- Personal exposure and microenvironment testing
- Health measurements
- Regional Air Quality Monitoring
- GPS Coordinates and Beacon Proximity

In consideration of data system design, the varying file formats, size of memory, relationships across data, and end use cases must be closely examined. To familiarize myself with the study and data, I participated in weekly meetings, browsed through the multiple cloud hosts of the data, and read overviews of REACCTING and related research.

By successfully responding to the posed question, the productivity of the researchers involved may increase and we can identify patterns that may have been otherwise unrecognized. Furthermore, the American and Ghanaian public may also be given the opportunity to find understanding and personal meaning in the data. More importantly, the takeaways from this study may lead to valuable insight for other new and developing data systems.

V. Implementation: Data Organization

A. Data Organization and Migration

At the start of the project, the majority of data types were stored in .csv and .txt file formats. The files would then be converted to a .mat format in support of MatLab processing, with results stored in a .pdf. Finally, the hand-filled survey data was stored in a .pdf format. All of the data resided in four different cloud locations, with limited and costly storage space and unintentional access restriction by team members. The researchers ensured that the data was regularly backed up by hand on a 2.2TB Ubuntu server provided by the University of Colorado.

The initial migration of the REACCTING data into a database began with the Stove Use Monitor (SUMs) data files. The measurements were generated by a digit that resided within a close proximity to the stove (note the red device in Figure 13) [3].

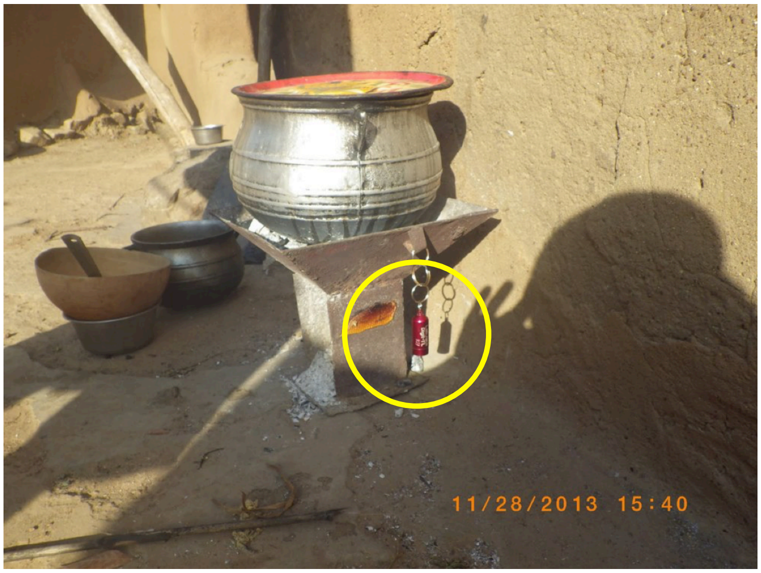


Figure 13. Cookstove with red stove use monitor [1].

The SUMs data included measurements of temperature, light, and humidity every minute and stored the calculations in the following .csv format, where the file name consisted of the device name and the data download date (e.g. “R18_09222014.csv):

Time	Temperature	Light	Humidity
3/24/2014 9:27:24 AM	39.250	0.0	0.00
3/24/2014 9:28:24 AM	38.125	0.0	0.00
3/24/2014 9:29:24 AM	37.250	0.0	0.00
3/24/2014 9:30:24 AM	36.562	0.0	0.00

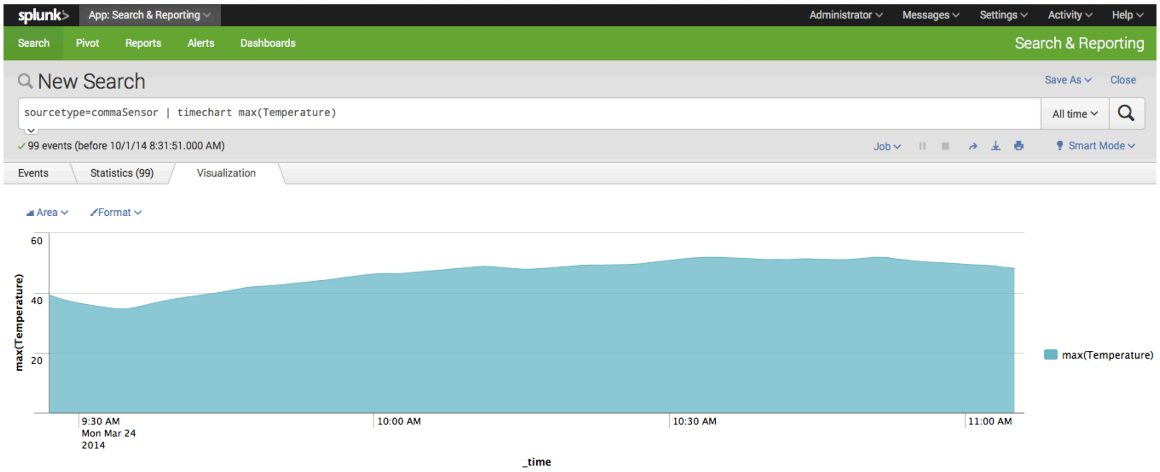
Table 1. Example of SUMs data format.

One major challenge that came with this data set was the presence of noise and corrupted data. Since the stove monitor was not in the stove itself, variables like distance to stove, location of the stove, and weather patterns directly affected the temperature readings. For this reason, there was a risk of “false positives” that indicate when the stove is in use. Furthermore, the SUMs device was also limited by the amount of heat it could withstand. For example, temperatures over 85° C would cause the data to level off, and when surpassing 125° C, the device would restart completely.

Human downloading and uploading of the .csv files also allowed for some corruption and duplication in the file set. In the case of duplicates, one file would typically contain a subset of records from the other. Deleting the smaller of the two data sets by hand generally resolved this issue.

In the process of cleaning and calibrating the data, multiple levels of processes were established, such that each level was more calibrated or cleaned than the previous. The initial goal was to store a unique collection for each level of processing. However, a single level showed to consume over 21GB of space, so it was determined that only the highest levels of processing would be imported.

Since the data is so tightly oriented around the timestamp, a database tool that was found to be particularly useful was Splunk (a database/visualization tool that is described in detail under the **Splunk** section). A preliminary chart of 100 data points was able to display the following timechart with a single SQL request (Figure 14).



```
sourcetype=commaSensor | timechart max(Temperature)
```

✓ 99 events (before 10/1/14 8:31:51.000 AM)

Figure 14. Splunk output of single SUMs .csv file

A close examination of the SUMs data helped create a framework for interpreting the organization and processing of the other data sets. The need to convert .mat files into .csv files became quickly apparent and the lack of hierarchy in the file system showed an obvious barrier to automated tasks. A long-term goal for the SUMs data is shown below in Figure 15 and can easily be applied to any of the datasets.

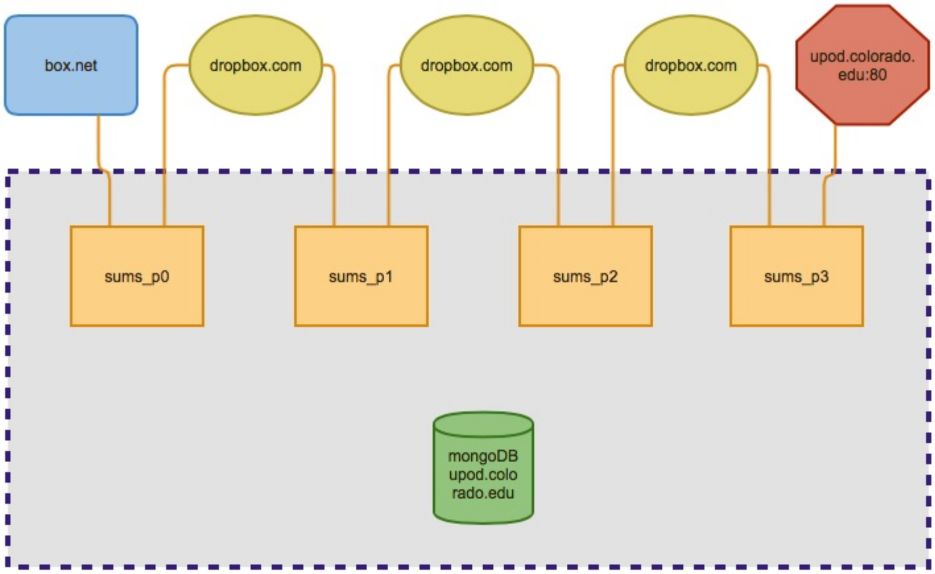


Figure 15. Initial "gameplan" for SUMs data management system.

Furthermore, Figure 16 shows diagram of the planned database architecture helped team members who were unfamiliar with databases understand how a structure file system could represent the relationships between the different data sets and various level of processing.

Another challenge that presented itself in the restructuring of the data was setting the permissions for user access. Generally, a single username and password was shared across team members for the various boxes, giving everyone with this information read/write/execute privileges.

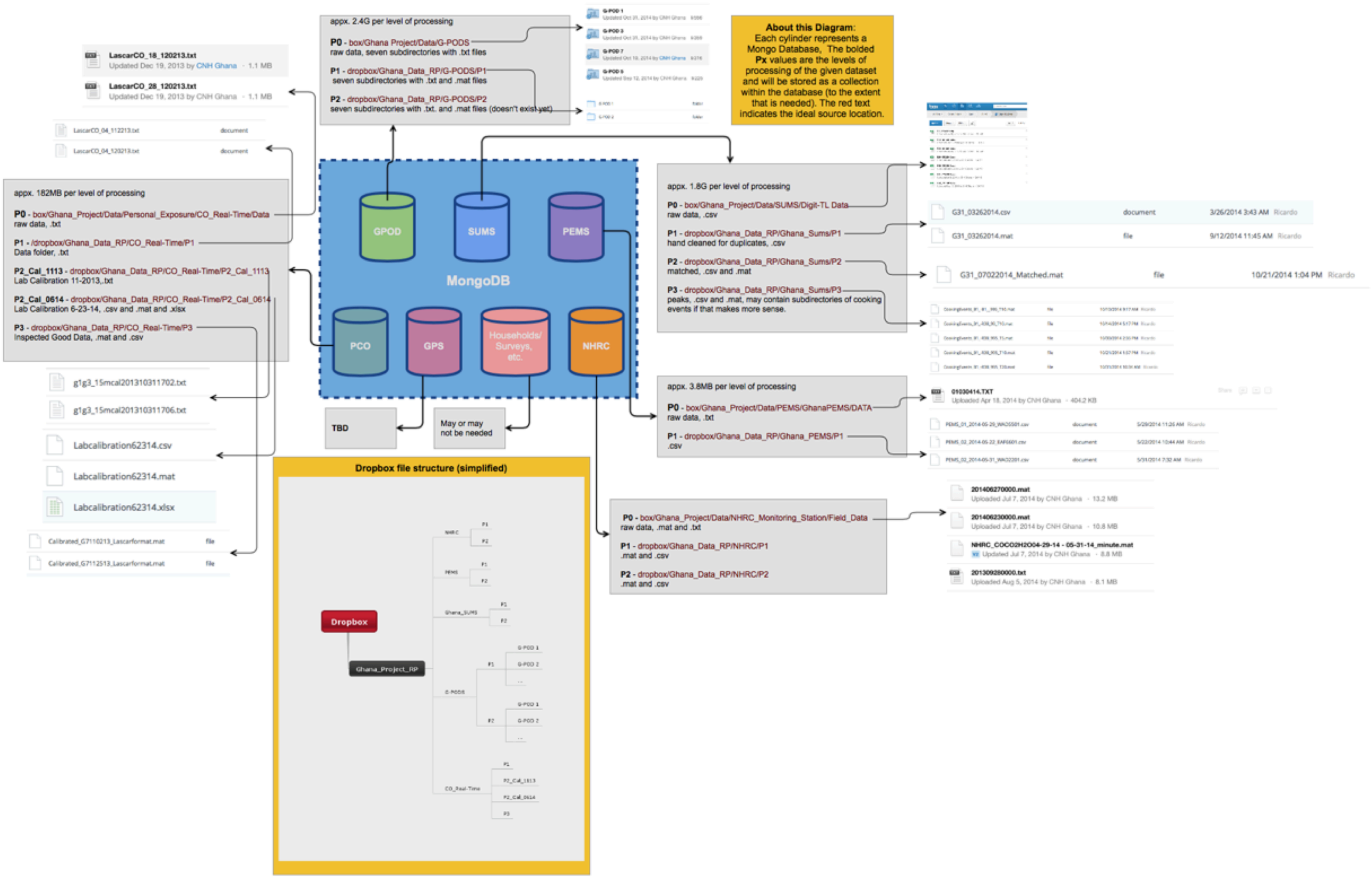

About this Diagram:
Each cylinder represents a Mongo Database. The bolded Px values are the levels of processing of the given dataset and will be stored as a collection within the database (to the extent that is needed). The red text indicates the ideal source location.
MongoDB
GPOD
SUMS
PEMS
PCO
GPS
Households/ Surveys, etc.
NHRC
TBD
May or may not be needed
appx. 2.4G per level of processing
P0 - box/Ghana Project/Data/G-PODS
raw data, seven subdirectories with .txt files
P1 - dropbox/Ghana_Data_RP/G-PODS/P1
seven subdirectories with .txt and .mat files
P2 - dropbox/Ghana_Data_RP/G-PODS/P2
seven subdirectories with .txt. and .mat files (doesn't exist yet)
G-POD 1
G-POD 3
G-POD 7
G-POD 9
appx. 182MB per level of processing
P0 - box/Ghana_Project/Data/Personal_Exposure/CO_Real-Time/Data
raw data, .txt
P1 - /dropbox/Ghana_Data_RP/CO_Real-Time/P1
Data folder, .txt
P2_Cal_1113 - dropbox/Ghana_Data_RP/CO_Real-Time/P2_Cal_1113
Lab Calibration 11-2013,.txt
P2_Cal_0614 - dropbox/Ghana_Data_RP/CO_Real-Time/P2_Cal_0614
Lab Calibration 6-23-14, .csv and .mat and .xlsx
P3 - dropbox/Ghana_Data_RP/CO_Real-Time/P3
Inspected Good Data, .mat and .csv
LascarCO_18_120213.txt
Updated Dec 19, 2013 by CNH Ghana · 1.1 MB
LascarCO_28_120213.txt
Updated Dec 19, 2013 by CNH Ghana · 1.1 MB
LascarCO_04_112213.txt document
LascarCO_04_120213.txt document
g1g3_15mcal201310311702.txt
g1g3_15mcal201310311706.txt
Labcalibration62314.csv
Labcalibration62314.mat
Labcalibration62314.xlsx
Calibrated_G7110213_Lascarformat.mat file
Calibrated_G7112513_Lascarformat.mat file
appx. 1.8G per level of processing
P0 - box/Ghana_Project/Data/SUMS/Digit-TL Data
raw data, .csv
P1 - dropbox/Ghana_Data_RP/Ghana_Sums/P1
hand cleaned for duplicates, .csv
P2 - dropbox/Ghana_Data_RP/Ghana_Sums/P2
matched, .csv and .mat
P3 - dropbox/Ghana_Data_RP/Ghana_Sums/P3
peaks, .csv and .mat, may contain subdirectories of cooking events if that makes more sense.
G31_03262014.csv document 3/26/2014 3:43 AM Ricardo
G31_03262014.mat file 9/12/2014 11:45 AM Ricardo
G31_07022014_Matched.mat file 10/21/2014 1:04 PM Ricardo
appx. 3.8MB per level of processing
P0 - box/Ghana_Project/Data/PEMS/GhanaPEMS/DATA
raw data, .txt
P1 - dropbox/Ghana_Data_RP/Ghana_PEMS/P1
.csv
01030414.TXT
Uploaded Apr 18, 2014 by CNH Ghana · 404.2 KB
P0 - box/Ghana_Project/Data/NHRC_Monitoring_Station/Field_Data
raw data, .mat and .txt
P1 - dropbox/Ghana_Data_RP/NHRC/P1
.mat and .csv
P2 - dropbox/Ghana_Data_RP/NHRC/P2
.mat and .csv
201406270000.mat
Uploaded Jul 7, 2014 by CNH Ghana · 13.2 MB
201406230000.mat
Uploaded Jul 7, 2014 by CNH Ghana · 10.8 MB
NHRC_COCO2H2O04-29-14 - 05-31-14_minute.mat
Updated Jul 7, 2014 by CNH Ghana · 8.8 MB
201309280000.txt
Uploaded Aug 5, 2014 by CNH Ghana · 8.1 MB
Dropbox file structure (simplified)
Dropbox
Ghana_Project_RP
NHRC
P1
P2
PEMS
P1
P2
Ghana_SUMS
P1
P2
G-PODS
P1
G-POD 1
G-POD 2
...
P2
G-POD 1
G-POD 2
...
CO_Real-Time
P1
P2_Cal_1113
P2_Cal_0614
P3

Figure 16. Outline for original MongoDB Structure.

The team had access to a headless linux server at upod.colorado.edu with approximately 2 TB of space (mostly unused). I was able to connect the “upod” server to the Box and Dropbox APIs for automatic syncing and weekly backups by implementing WebDav and the Dropbox Python command line tool from Crontab. However, before fully implementing this, I restricted my own write access to the data to avoid accidental overwriting or deletion of data. Permissions that were set on the Unix side would be dropped during syncs, so we had to manipulate the user account settings directly on the cloud websites to eliminate the possibility of corruption from the upod server. On Box, it is very easy to add additional read-only users to the file system. On the other hand, Dropbox counts the size of shared data against each user that has read and/or write access to it. Since the size of the shared data exceeded the allotted amount that I had for my free account, I was required to register a paid account to safely access the REACCTING data.

The backups proved to be convenient for the team and, since they were compressed into zip files, this process was more space-efficient. At one point, a team member identified that the backups were no longer working for Box. I determined that the Webdav had stopped working because of a recent power outage and resolved the issue. These complications show the importance of human oversight when using automated scripts.

B. Querying, Data Access

In order to visualize the data, I evaluated different options for accessing it. The varying data sets were stored both as zipped backups on the upod server and on the Box or Dropbox cloud. One of the more complex but robust options for data storage format was a database. A comparison between SQL and document-based database formats showed that a document-based database would be a better solution for our data because of the flexibility of storage formats, scalability, and compatibility with various tools like D3.

Using Ruby, I experimented with importing the SUMs data into a MongoDB collection. In the Ruby script, I also established indexes over all of the fields (since there was plenty of storage space). The following is pseudocode for the resulting Ruby code:

```
Import ‘csv’, ‘mongo’, ‘json’ gems

Check that argv has expected num of params.

From the filename, get the digit name and upload date.

Connect to Mongo ‘sums’ database with the specified collection name.

Walk through the CSV rows and import each as a document.  Include the
digit name and upload date as fields.  Add another field that has the
time converted into the Mongo-friendly isoDatetime format.

Create indexes on each field name
```

I created a shell script that would call the Ruby code on each of the SUMs CSV files that were in the given directory. If the dataset was more dynamic, this script could be potentially run as a cron task on a regular basis to keep the database up-to-date.

VI. Implementation: Visualization

A. Experimentation and Prototyping

In the process of determining an appropriate visualization method(s), I experimented prototyping data sets with multiple tools and software.

Splunk

Founded in 2003, Splunk is an American multinational corporation that produces software for searching, monitoring, and analyzing machine-generated data. Splunk's CEO, George Sullivan, says that it has extended beyond big data analytics into "operational intelligence," such that the analysis can keep up with organizational functions and can be used to identify problems or changes [28]. Users can browse and visualize large dynamic datasets or log files via an aesthetic web-style interface (see figure 17, [29]). Splunk is an ideal tool for time-related tasks and "quicklooks" for basic graphing or statistical analysis.

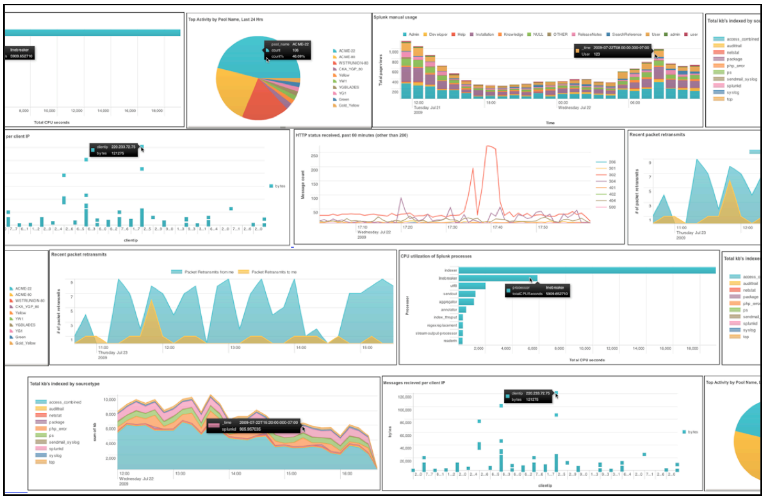


Figure 17. Examples of Splunk Graphics [29]

For the REACCTING data, I experimented using Splunk to map the SUMs data. The SUMs dataset proved to be an adequate choice because it was based on changes over time. Once the data was in the Splunk system, it was very easy to visualize with multiple graphing methods (see Figure 14). However, the format of the data presented the following challenges for importing:

1.) The SUMs data resides in many different files. The highest level of processing (P2) alone contains nearly 200 .mat or .csv files.

2.) The files have important identifying information in the filename itself. For example, "B4_09222014_Matched.csv" indicates that the data was uploaded on September 22, 2014 for the digit B4.

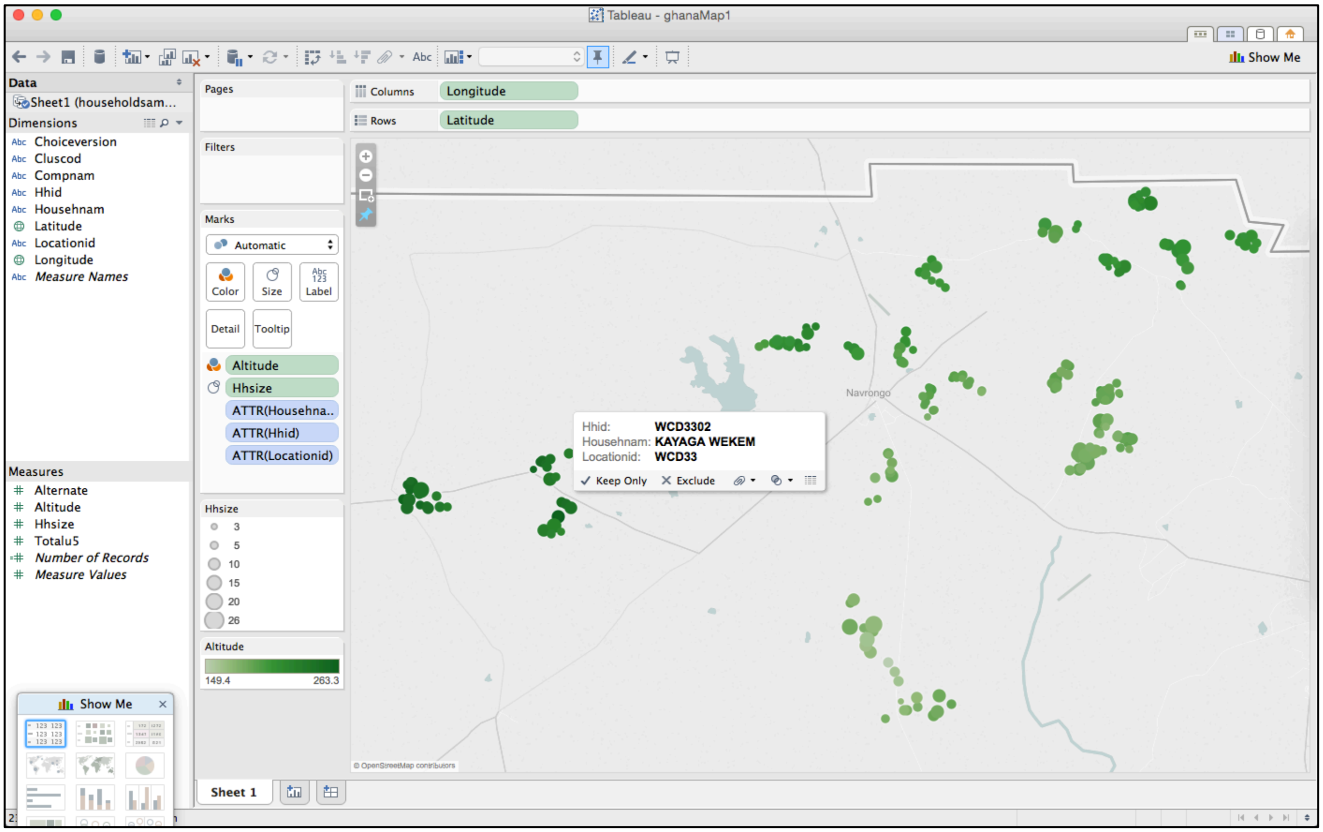
Both of these predicaments could be resolved with some scripting, but they were important to consider during these initial stages.

Splunk provides a forum to create fast and dynamic visualizations. However, since the REACCTING datasets were static and the team was already using Matlab for graphing functions, in this context, Splunk did not appear to offer enough in return for the time invested in learning the tool.

Tableau

Also founded in 2003, Tableau promotes itself as a "business intelligence software that allows anyone to easily connect to data." The data can be visualized in the form of interactive and sharable dashboards [30]. Individuals with little to no programming or SQL experience can use the Tableau software to create visualizations that dynamically reflect business changes. If a data set has coordinates, Tableau is simple to use for mapping. It is commonly used for graphing but, unlike Splunk, isn't as time-centric.

I found that Tableau offered essentially the same features as Splunk for graphing REACCTING data and experimented with mapping the households. The results were visually appealing and the markers were sized and colored based on the elevation and household size (Figure 18). However, for REACCTING's purposes, Tableau has similar disadvantages as Splunk, in that the features overlap with the existing functions offered by Matlab or ArcGis.



Columns: Longitude

Rows: Latitude

Hhid: WCD3302
Househnam: KAYAGA WEKEM
Locationid: WCD33

Figure 18. Tableau representation of household location.

Data-Driven Documents

Data-Driven Documents (D3) is a JavaScript library for creating powerful data visualizations. D3 connects data to web-based HTML or SVG documents that can be rendered by a web browser and interacted with by the client. D3's primary author is Mike Bostock and the code is entirely open source and released under a BSD license. In his book, *Interactive Data Design*, code artist Scott Murray describes D3's generation of web documents in the following steps [31]:

1.) *Loading* the data into browser memory

2.) *Binding* data to elements within the document

3.) *Transforming* these elements by setting visual properties based on the data

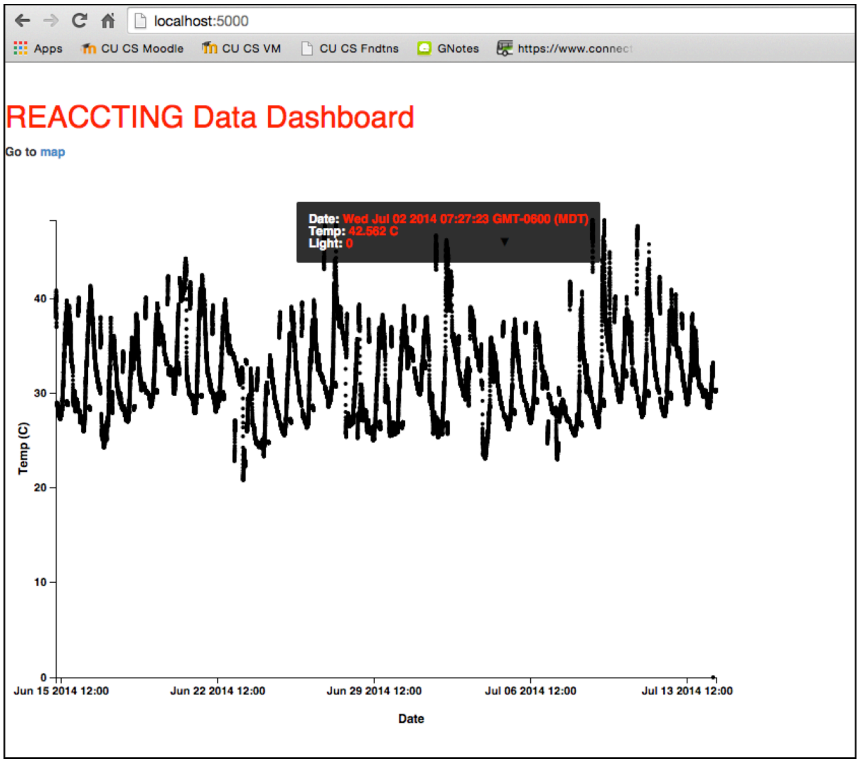
4.) *Transitioning* elements between states in response to user input

D3 less of a “boxed” software compared to Tableau and Splunk. Rather than serving as an *exploratory* tool that quickly generates predefined visualizations of datasets, D3 is best served to create customized *explanatory* presentations of the data. This way, the visualizations can be catered to a particular audience or adjusted to highlight known discoveries in the data. Murray states that, although exploratory views are more “constrained and limited,” they are more effective for focusing on and communicating important points [31].

In addition to it’s customizability, D3 also offers the benefit of being scalable, both in the sense that it can be designed to easily allow extensions in features and the vector images can be resized without a loss in quality. D3 is fast and can be made interactive through the use of tooltips, zooming, etc.

Murray notes the important distinction that D3 does not “hide” the original data from the client browser. This characteristic is important to consider when working with data that could potentially disclose identifying information about subjects in the REACCTING study.

Despite these features, D3 is a relatively advanced tool to learn, requiring programming knowledge and a basic understand



REACCTING Data Dashboard

Go to map

Date: Wed Jul 02 2014 07:27:23 GMT-0600 (MDT)
Temp: 42.562 C
Light: 0

Figure 19. D3 Visualization of SUMS data

of web application construction. Fortunately, there is a strong community of D3 developers and descriptive tutorials available online.

Since I had already imported the SUMs data into a MongoDB collection, I decided to prototype a graphed visualization of one of the digit's temperature readings over a span of several months (see Figure 17). This basic visualization required that I establish a frontend client and backend server, a process described in the *Piecing Together a Concept* portion of this paper. I exported the SUMS database to my personal computer to have access to a display and, to get started, referred to the code provided by Moujahid on Github that had a similar structure and purpose [32].

For the graph itself, I followed several D3 tutorials made available online by Scott Murray and adjusted the code to fit the format of the SUMS data [33]. The only interactive feature I made available during these stages was the tooltip, such that when users hover their mouse over a data point, it would "pop up" pertinent information about that data point.

It took the server approximately forty seconds to query the Mongo database for 42,351 records and deliver them to the D3 client in a parsable JSON format. The database contained multiple indexes to improve retrieval time, but no other optimizations were added. If the server overhead was unacceptable, perhaps the D3 code could be customized to make multiple GET requests for segments of the full time range and gradually render the SVG as the data is received.

Although the REACCTING team found the visualization interesting and fun to interact with, it provided little value to the academic research objectives since, as before, the functions overlapped with powerful existing statistical tools like MatLab and R.

Mapbox

Additionally, I experimented with the user-friendly mapping tool called Mapbox. Using geojson.io, I identified the approximate coordinates for the region in Ghana where the research is taking place and then uploaded a CSV file of the household coordinates and details (Figure 20). Mapbox easily parses CSV files and even uses the non-location information in the tooltips for the markers. As a standalone exploratory tool, Mapbox didn't offer much more functionality than basic mapping to the team (ArcGIS was already being used for this). However, I would end up using this map as a tile layer during the initial stages of the web application construction.

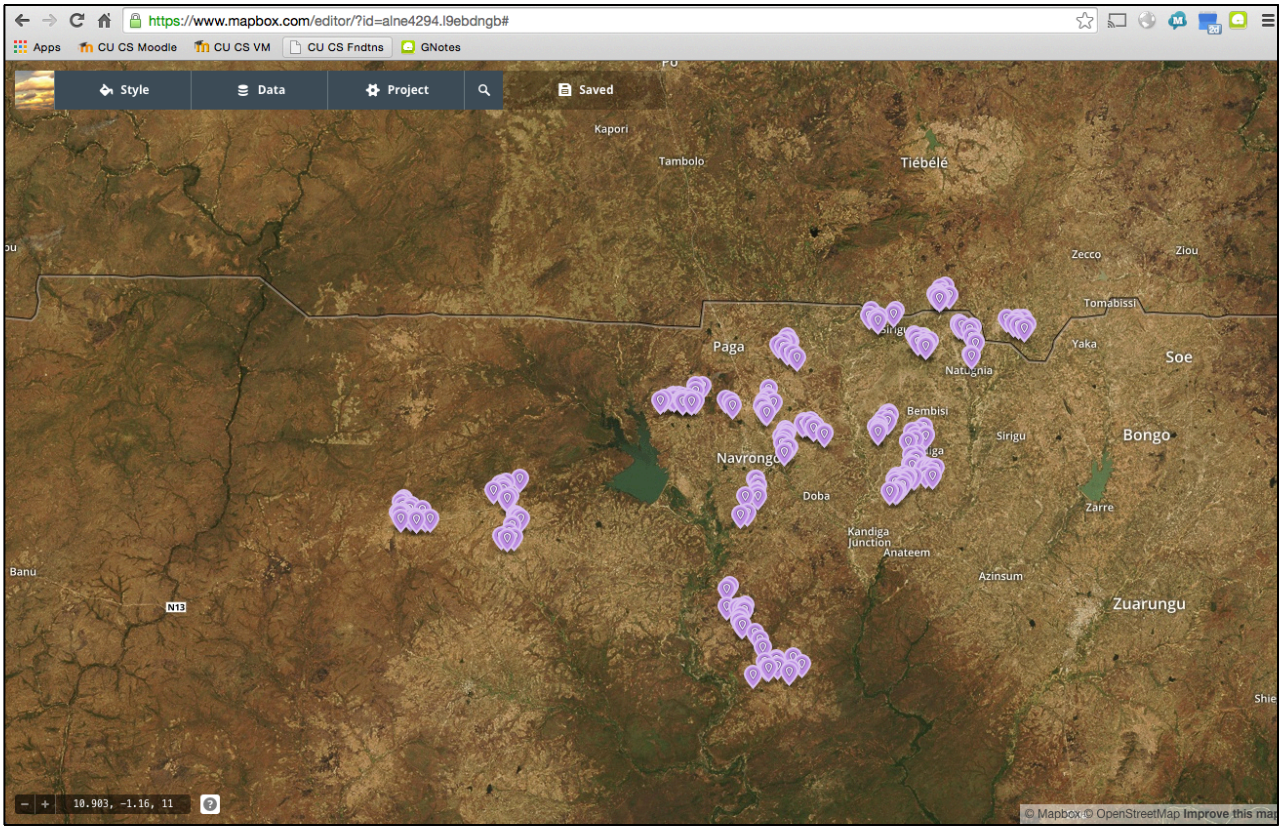


Figure 20. Mapbox map of household locations

ArcGIS

ArcGIS promotes beautiful geographical data visualization. However, I chose not to explore this option in depth because I was more interested in programmatic/open-source alternatives and a team-member was already using ArcGIS for his research.

B. Piecing Together a Concept

The REACCTING team was already using both free and proprietary high-level software for advanced statistical, geographical, and graphical analysis, such as R, MatLab, and ArcGIS. Naturally, the results from these tools were very static and non-interactive, which were appropriate for an academic papers and research. After prototyping the various visualization techniques, I presented my results to the team in February and proposed an end-goal for my visualization project.

In November, 2014, the team deployed five Android smart phones to Ghana that would track a subject's location and proximity to the firepit beacons. The relatively fresh GPS data had not yet been analyzed alone or in the context of the other datasets. A team member proposed that it would be helpful if I could provide some insight to this data through some sort of mapping tool. The end goal could create a sort of "quick-look" into the GPS data and ideally even connect it with the Personal Emissions Monitor data (PEMs), which primarily tracks an individual's exposure to C02. Although this would not be expected to provide any "groundbreaking" insights into the data, it would serve the *explanatory* purpose of creating an overview of the subject's location and how that relates to exposure. Additionally, this animation could be understood by individuals with full or limited knowledge of the complexities of the research.

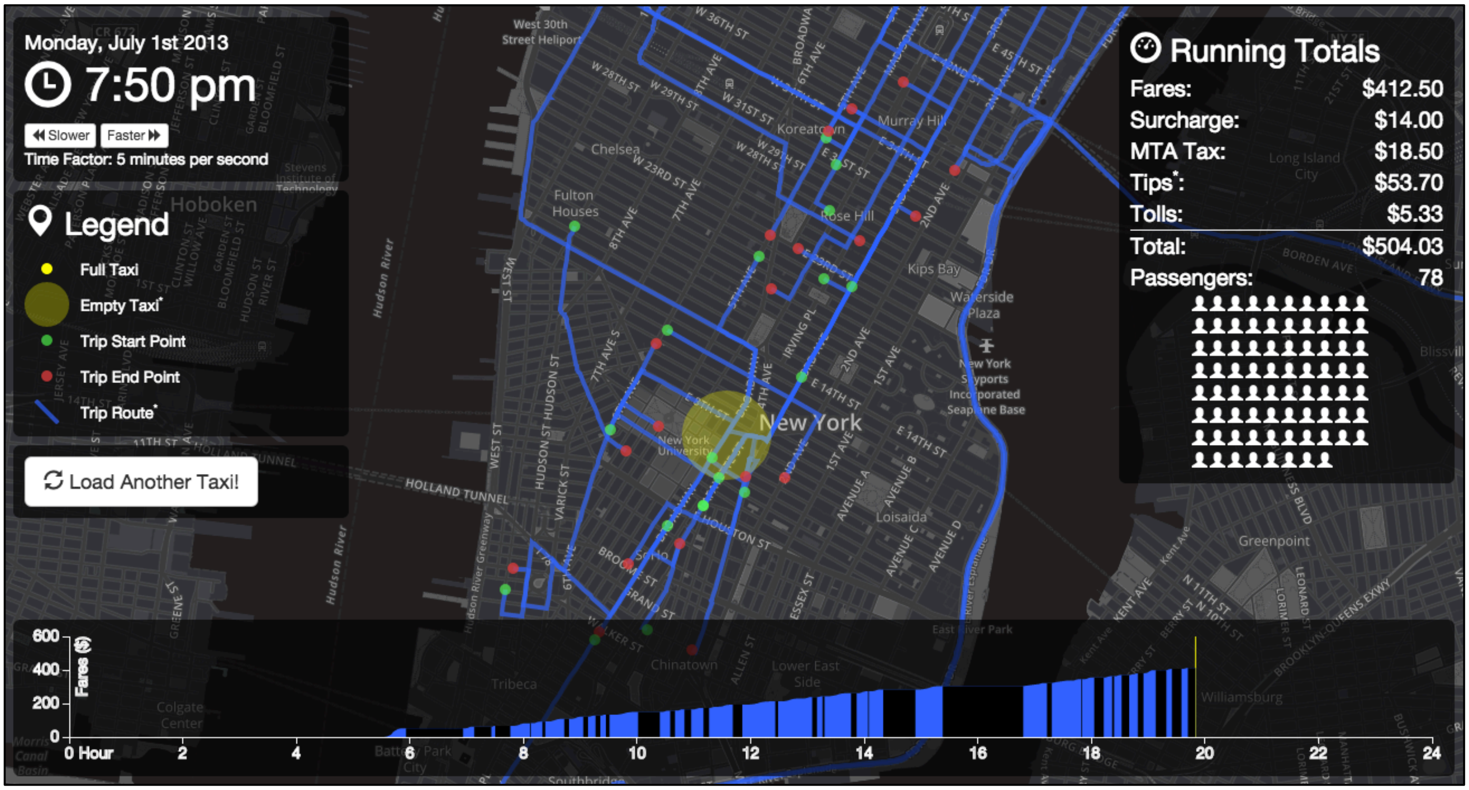
With this in mind, I began researching existing interactive maps that used edgy open-source tools like D3 and Node. The following two examples were highly influential in my final project.

Taxi Tracker

In his blog, Chris Whong describes the decision-making process he went through to create an animation of the daily movements, revenue, and number of passengers associated with a random taxi driver in New York City in 2013 (Figure 21). He even made the full source code for his project available on GitHub [34]. The animation was the winner for the "Best Motion Infographic" in 2014 [35].

To access the data, Whong used BigQuery to get a driver's trips associated with a random day. He then used node to build API calls to connect to BetaNYC's released Citibike Trip Data. The data would be processed and moved to an SQLite database before being converted to geoJSON and sent to the browser. For the frontend visualization, Whong worked with Leaflet and D3. Having access to multiple different taxis and days allows the user to compare related datasets and identify patterns of behavior. This is in line with Tufte's first principle of *Comparisons*.

The end result is an aesthetic and minimalist animation that shows the progress made by a taxi driver over a span of time. The user can interact with the visualization by adjusting the speed of the animation and requesting a new random taxi. Users can gain real insight on what it is like to be a taxi driver, a perspective that wouldn't typically be considered by most people.



Running Totals	
Fares:	$412.50
Surcharge:	$14.00
MTA Tax:	$18.50
Tips*:	$53.70
Tolls:	$5.33
Total:	$504.03
Passengers:	78

Figure 21. Screen shot from Whong's animation, "NYC: A Day in the Life" [35]

Donorschoose Dashboard

The second visualization that I found inspiring was Donorschoose Dashboard by Adil Moujahid. Like Whong, Moujahid describes on his blog what the decision-making process looked liked to create this web application. He also releases his code on GitHub [36]. Users would be able to click on a state and be shown its associated number of donations, poverty level, and frequency of donation type. Although his blog only contained a gif-style demo of the end result, it was evident how such a multi-faceted and interactive visualization could effectively reflect the relationships between different datasets [32]. The use of multiple variables and datasets aligns with Tufte's second principle of *Multivariate Analysis*.

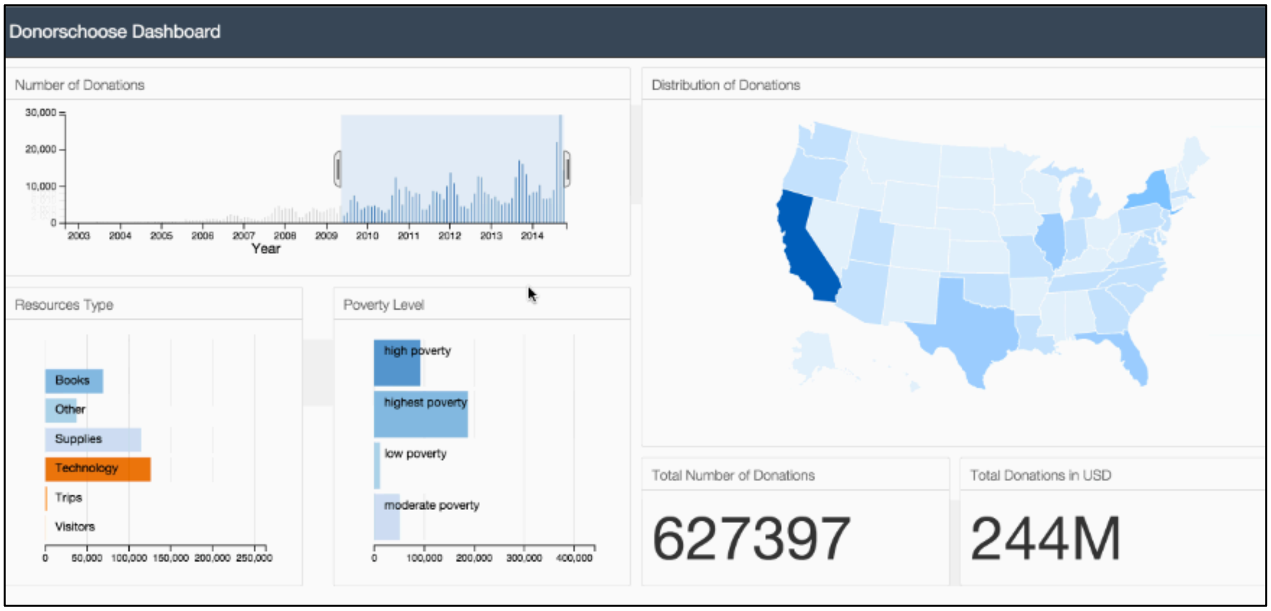


Figure 22. Screenshot of GIF that demonstrates the animation of Moujahid's Donorschoose Dashboard [32]

The Donorschoose Dashboard and Taxi Tracker shared the characteristics of connecting location data to quantitative data and allow the user to interact and control the animated elements. By displaying multiple datasets in a single visualization, this falls in line with Tufte's belief that, "Most techniques for displaying evidence are inheriently multimodal" [10]. Technically, both visualizations implement D3 with a Python Flask backend and had some sort of mapping component. The modularity of the code would potentially allow for future extensions and customizations. The visualizations are unique and clearly add value to the data that they are presenting. Since the authors both chose to describe their process in a blog and made their code publicly available, the visualizations can contribute to the community of developers.

For these reasons, I decided that a similar interactive and animated mapping application could prove to be a unique and valuable tool for explaining the REACCTING data. I had limited experience with web application development through my studies at the Unviersity of Colorado, so I anticipated a steep learning curve. Additionally, I had some concern that the animation would overlap with the capabilities of one of the teammembers who was working with ArcGIS, but I maintained communication with him to ensure that this wasn't the case.

C. Ensuring Relevance

During the presentation that I gave to the REACCTING team in February, I proposed the following goals:

- Tie together datasets (ideally, the GPS coordinates and PEMS data)
- Provide an interactive "Quicklook"
- Maintain relevance to objectives of group
- Provide insight to the GPS/Beacon dataset, which had been minimally analyzed at that time
- Potentially make the data publicly accessible (with restricted user access, due to the sensitivity of the data)

A personal goal was to align with Tufte's Principles of Design that were described earlier.

Although we agreed on the outlined objectives, it would be important to maintain consistent communication with team members throughout the development process to ensure relevance. Since our weekly meetings were cancelled periodically for various reasons, I found that email was the most effective way to ask questions.

D. Technical Implementation

The following section describes the decision-making process that ensued while creating the web application. In general, I based my selections on the following considerations:

1. What I am already familiar with
2. How lightweight and flexible each choice is; lightweight generally means less overhead for a smaller application like this
3. Which option has a stronger community or documentation supporting it

Language and Framework

After evaluating different programming language and framework options, I decided to use Python Flask for the backend server framework and Javascript for the frontend language.

Python is a popular dynamically-typed high-level language that is known for its readable yet concise syntax. Python would also prove to be a good choice because of the extensive libraries that are available. I was able to use the pymongo and datetime packages to import data from MongoDB and convert the Matlab date formatting to the standard Gregorian format (the method I used for converting Matlab time formats was described in a blog by the "Sociograph" [37]). Pip was my package installation manager of choice. I chose to implement Flask because I have past experience with this microframework and felt that a more heavyweight framework like Django or Pyramid was unnecessary at this time.

I programmed in JavaScript for the client-side browser, primarily because I was interested in implementing the D3 and Leaflet JavaScript libraries. For managing such packages, I used NPM and Bower. JavaScript is an interpretive language that is loosely typed. It features asynchronous and a strong community of developers.

JavaScript is a unique language and frequently misunderstood, partly because its name incorrectly suggests that: a.) JavaScript is related to Java, and b.) JavaScript is simply a scripting language. Rather, like Python. JavaScript can be used for advanced object-oriented design and is a full-featured language at its core [38].

I ended up implementing the following notable JavaScript packages:

- *jQuery* – Implemented to update most of the "listeners." In other words, when an action was performed by the user, jQuery initiated a behavior change in the animation.
- *Scriptaculous* – Used to build the timeline slider
    - Note: there were naming conflicts between jQuery and Scriptaculous, since both were dependent on the user of the "$" symbol [39]. This was resolved by using jQuery's noConflict() feature. Additionally, Scriptaculous had to be called in header for it to perform as expected [40].
- *Fontawesome* – Used by Leaflet Awesome.Markers plugin
- *D3* – Implemented to create GET requests and graph data
- *Leaflet* (discussed under **Mapping**)

Version Control

GitHub was originally selected for version control, but I gradually shifted to making manual backups on Dropbox due to the sensitivity of the data that I was working with (GitHub typically offers free private repositories for academic purposes, but I am still waiting for my proposal to be approved). This method added a small amount of overhead, but it wasn't a substantial hassle since I was working independently.

Data Cleaning and Storage

The data format for the GPS datasets was comma-separated. Fortunately, a teammate was able to parse and combine the hundreds of imported CSV data into one CSV file for each phone. I used convertcsv.com to transform each CSV file into a JSON file of string elements. When the animation page called a GET for the phone data, the Python backend then read in each row from the JSON file, converted each date into a readable Gregorian date and time, removed unnecessary fields, and converted the resulting dictionary back into a JSON format to submit over the network to the client. Although stepping through each row in the JSON file added to the computer's memory and performed in O(n) time, it was a necessary step to ensure that the data was clean and no larger than necessary before transmitting to the client side.

I quickly observed multiple problems with the data:

- Missing column data for a given record - some of these columns were critical for mapping
- Duplicate data – most notably, some data was repeated every hour
- Irrelevant data – for example, the records contained data of when the phones were still located in Boulder, Colorado
- Missing data – chunks of time simply weren't there
- Incorrect type assumptions from when the file type was converted from CSV to JSON – some numbers were inconsistently interpreted as strings

I took a couple of different approaches to resolve these resolve these issues. First I "hand cleaned" any records that were obviously invalid or corrupted. These rows frequently contained no data or "0" for all of the values and would appear in large chunks, perhaps due to technical issues with the phone. From the client code, I also added multiple checks like below to ignore undefined or partially-undefined records that had made it to that point (without this check, the system halts when at an undefined record):

```
if(records[j] && records[j].gpsacc){
   //We have a record with gpsacc data, so do some action with it
}
```

I also hand cleaned and added a check for coordinates substantially outside of the expected region. Many rows that recorded GPS in Boulder, Colorado were deleted, and I added the following iterator on the client side to exclude corrupt data that indicated a misreading of the GPS:

```
for (var i = 0; i < numRecords; i++){
  var record = dataObjs[i];
  if (i<5){
    console.log("RECORD");
    console.log(record);
  }
  var lat = Number(record.lat);
  var lon = Number(record.lon);

  //Check that our coordinates are nums and within the acceptable area
  if(!isNaN(lat) && lat<11.109684643110414 && lat>10.478359299402134){
    if(!isNaN(lon) && lon<-0.693511962890625 && lon>-1.41998291015625){

      //Get running total of lat and lon so we can get avg to know where
      to center the map
      latTot += lat;
      lonTot += lon;

      lats.push(lat);
      longs.push(lon);
      records.push(record);
    }
  }
}
```

Although this was another linear running-time process, it was an acceptable place to add the check since the average coordinate had to be calculated and the longitude and latitude were being added to an additional array anyway. By using additional arrays for longitude and latitude, smoothing of the animation could be added later. I had discovered that without adding these arrays, the smoothing mechanism would add enough "smoothing" objects to presumably overflow the browser memory and crash the application altogether (in our case, increasing the number of record objects from 61,000 to 247,000 caused this behavior).

For missing data, I knew that I could: a.) Ignore the issue in the code, b.) Create artificial or smooth the animation to give the appearance that the problem didn't exist, or c.) Create some visual indicator in the animation that, for the length of time that the data is missing, clearly indicates to the user that there is a gap in information. I selected option a, since the absence of data didn't appear to uncomfortably disrupt the fluidity of the animation and I felt that not enough information was yet known about the dataset to create artificial or "filler" data.

This process of cleaning, visualizing, and then cleaning the data again reflects the cyclic nature of the data life cycle. Throughout the time that I worked with this dataset, I maintained communication via email with the primary team member that had deployed it.

Mapping

Initially developed in 2011, Leaflet is an increasingly popular lightweight JavaScript library that is used for creating mobile-friendly interactive maps. Leaflet supports map layers like tile layers, markers, popups, and vector layers, and can even be interacted with through zooming or panning. Leaflet advertises that their seamless functions can be also be expanded through the use of third-party plugins [41].

Since I already had the Mapbox map that I created earlier, I plugged that in as a Leaflet tile with the following code:

```
//add a MapBox tile layer
L.mapbox.accessToken = 'pk.eyJ1IjoiYWxuZTQyOTQiLCJhIjoiZERWYV8yZyJ9.9Dp7x2OGFrLO6ZIunDSHyQ';
var map = L.mapbox.map('map', 'alne4294.l9ebdngb').setView([10.9100667, -1.0007569], 12);
```

Although it was informative to have the households displayed, I quickly discovered that Mapbox offered an insufficient resolution for this region in Ghana. Figure 23 shows how increasing the zoom resulted in a poor quality satellite image or, at a certain point, none at all.

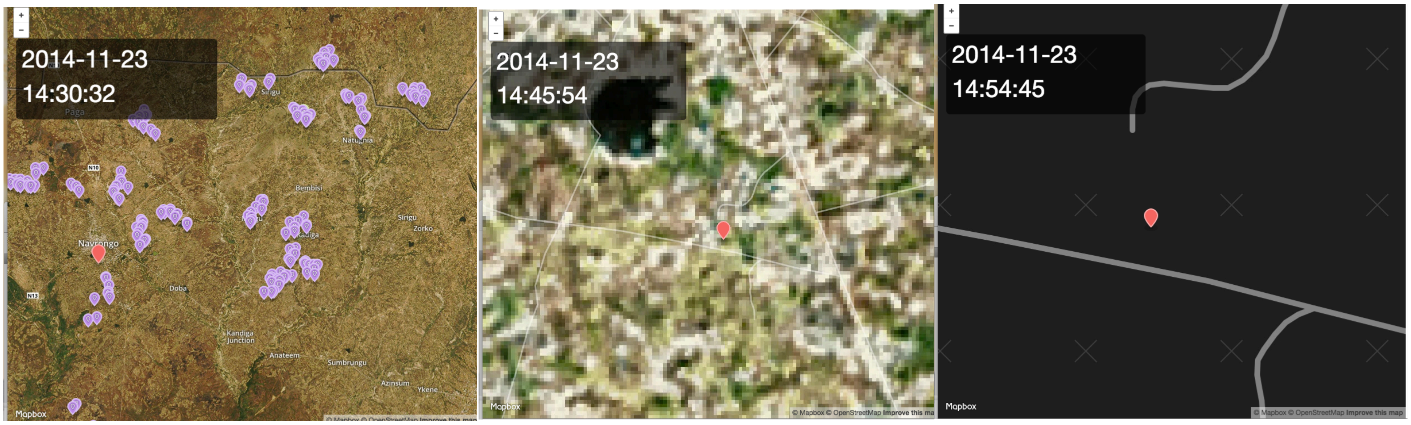


Figure 23 Mapbox resolutions of Northern Ghana at increasing zoom levels

Due to the poor resolution, I explored alternative mapping options. The most natural choice for use with Leaflet would be OpenMaps, but this service only offers graphical mapping views of an area (as opposed to satellite imagery). According to Leaflet, using Google or Bing maps would require more loading/coding overhead than OpenMaps or Mapbox, but it was common to do so. I was able to find multiple plugins on GitHub for both mapping services, which simply required that I download and import the packages to enable the Google or Bing map calls. I opted to use Shramov's Leaflet plugin library, which supported the providers Bing, Google, or Yandex .

The resolution and coloration of Bing maps seemed to be of high quality, and it had names on the roads for this region. However, calling the Bing maps required an API key and would eventually cost money after the trial period.

Google maps also offered a high-resolution picture of this region. Interestingly, the coloration of satellite imagery was noticeably cooler in temperature (see Figure 24).

This may have been due to variations in seasons/date/time or satellite sources. I was partial to the crisp coloration in Bing, but the quality of rendering appeared comparable overall.

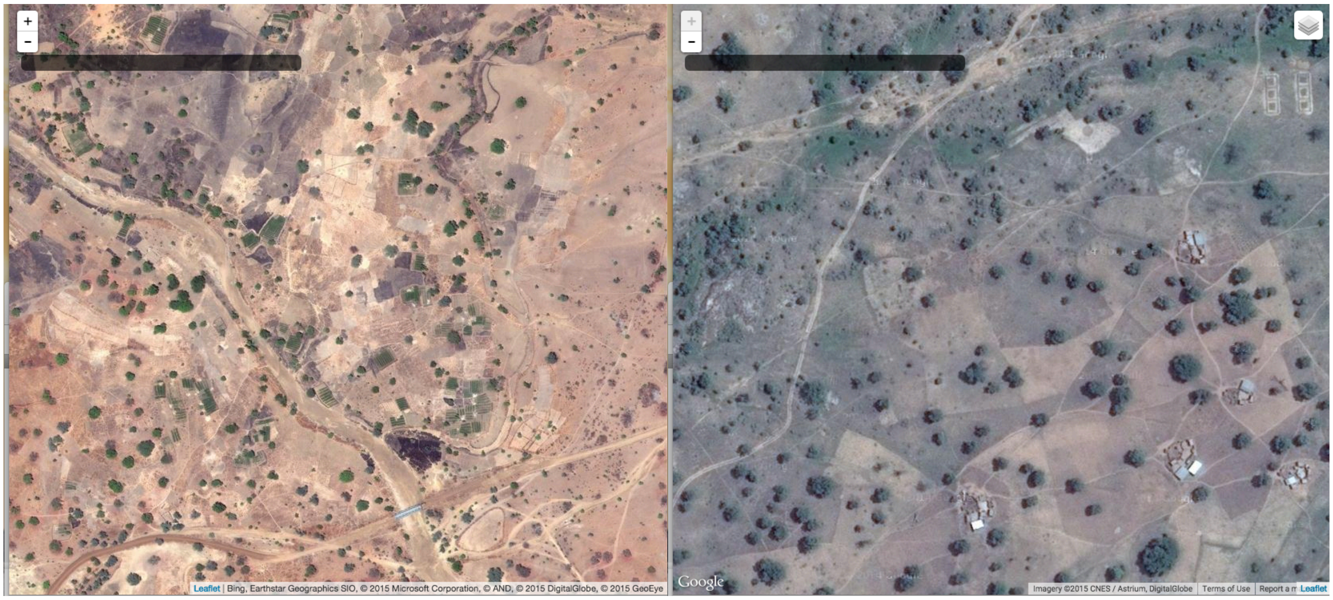

+
-
Leaflet | Bing, Earthstar Geographics SIO, © 2015 Microsoft Corporation, © AND, © 2015 DigitalGlobe, © 2015 GeoEye
+
-
Google
Imagery ©2015 CNES / Astrium, DigitalGlobe | Terms of Use | Report a m Leaflet

Figure 24. Bing Maps (left) vs. Google Maps (right) in the Navrongo region of Ghana

Google Maps also offered multiple different views of the same region at the user's discretion. Figure 25 demonstrates how the perspective adjusts as the viewing mode is selected from the top right corner of the map.

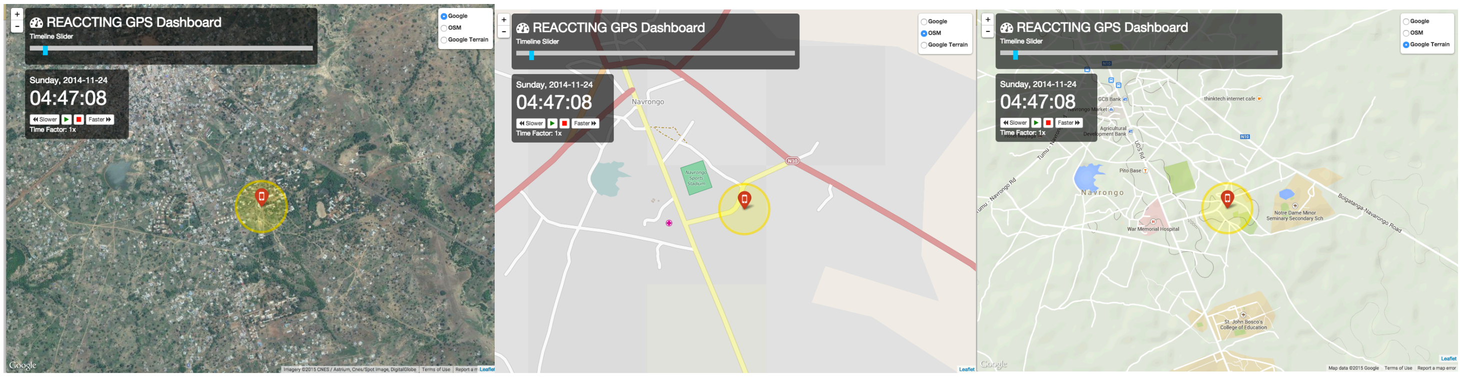
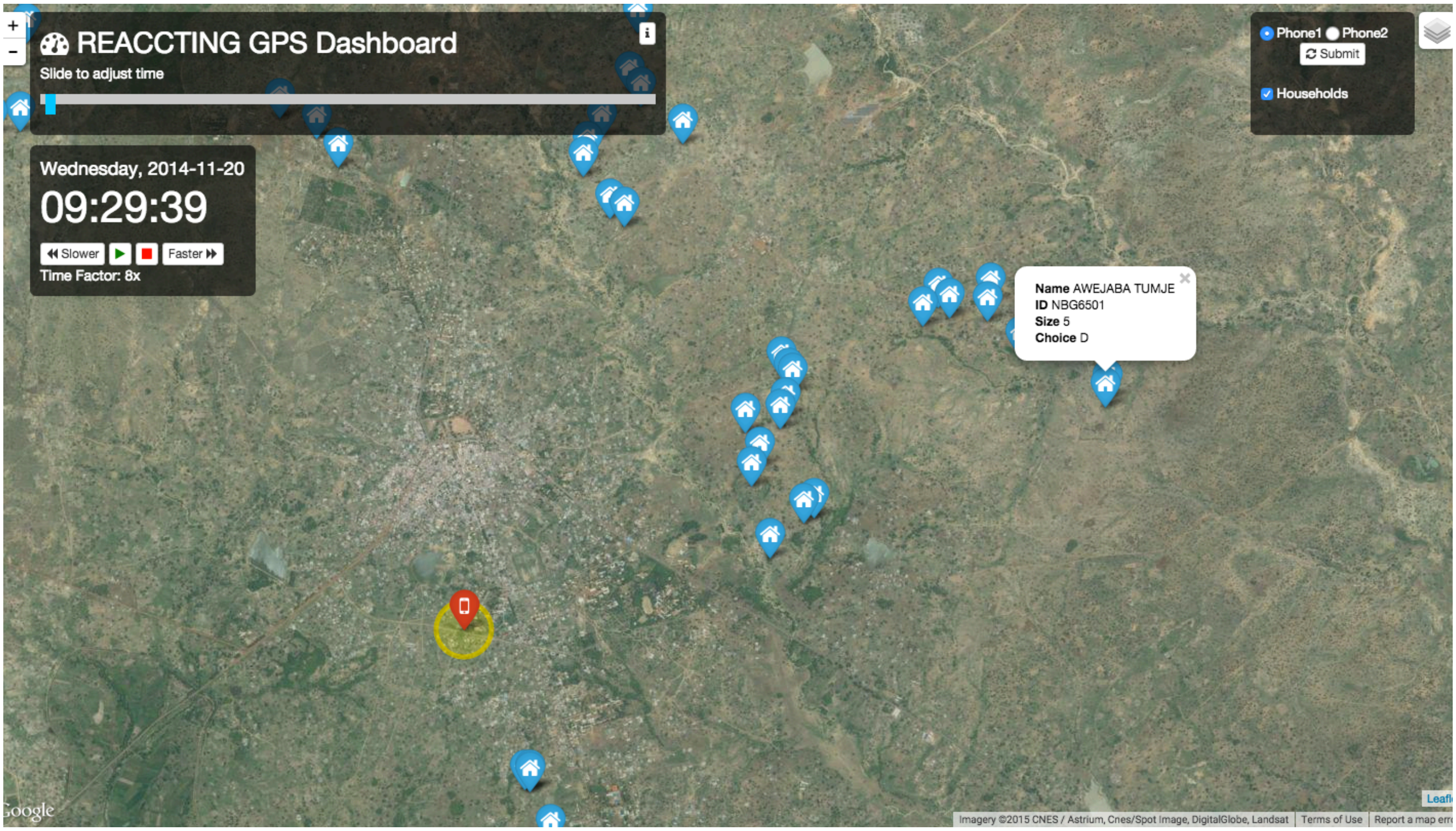


Figure 25. Google Maps as shown from the *Google*, *OSM*, and *Google Terrain* perspectives, respectively

VII. Results

Overall, the web application provided value to the REACCTING research that cannot easily be substituted with other boxed visualization or analysis tools. The content is understandable by users with varying levels of familiarity with the study and met the objectives defined with the group. Figure 26 is a screenshot of the end result.



REACCTING GPS Dashboard
Slide to adjust time
◀◀ Slower ▶ ■ Faster ▶▶
● Phone1 ○ Phone2
Submit
Google
Leafl
Imagery ©2015 CNES / Astrium, Cnes/Spot Image, DigitalGlobe, Landsat | Terms of Use | Report a map err

Wednesday, 2014-11-20

09:29:39

Time Factor: 8x

☑ Households

Name AWEJABA TUMJE
ID NBG6501
Size 5
Choice D

Figure 26. Screenshot of web application

I was able to meet the visualization goals that were outlined during the February REACCTING meeting, with future opportunity for stretch objectives. As intended, the result is a "quick-look" of the GPS data that can be navigated interactively through the use of the time slider and buttons.

By considering Tufte's Principles of Design, the visualization was also developed to be both engaging and explanatory. In the early stages of development, I referred frequently to other successful modern applications like the Taxi Tracker and Donorschoose Dashboard, which helped with effective decision-making. Similar to these visualizations, my end product meets Tufte's principles in the following ways:

- It is multi-faceted in that you can view parallel GPS datasets (Principle 1, *Comparisons*).
- It is considerate of assumptions made by the end user by providing labeling and an "information" button (Principle 2, *Causality, Mechanism, Structure, Explanation).*
- It is based on multiple datasets with significant relationships; These datasets include the Nvrango region map, household locations, and GPS coordinates of the subject (Principle 3, *Multivariate Analysis).*
- It is aesthetic and concise; this is partly through the use of CSS customizations and packages like Bootstrap and Font Awesome (Principle 4, *Integration of Evidence*)
- It is based on honest data (Principle 5, *Documentation*).
- It represents information that is significant to human understanding (Principle 6, *Content Counts Most of All*).

Due to my earlier research, I expected the cyclic nature of processing, visualizing, and cleaning the data, but was surprised by how much time was spent prototyping code or visualizations that would not actually be used in the end product. However, the data engineering process in itself added value to the research. For example, it created a forum for discussing and improving the organization and cleaning of data with the team. Furthermore, I was able to make educated decisions based on my earlier prototypes and even reuse some of my earlier work.

The final stage of this process is publicly deploying the web application. I requested to open a port on the upod server since it has plenty of space and would be free to use and the CU's OIT office approved the proposal. Now it is simply a process of matching the upod Python/JavaScript environment to the one that is set up on my personal computer. This process should be made simpler by the fact that MongoDB is already functioning on that machine, and the Python application has a requirements text file that can be used to install the dependencies.

VIII. Conclusions and Future Work

The modularity of the application would easily allow features to be added in the future, given the appropriate resources. Access to such resources (notably human resources and funding) directly contributes to the degree of success in maintaining or "curating" datasets and data tools. I feel confident in the fact that, if another student or professional were to get involved, they could certainly expand upon the existing foundation.

Such expansions or improvements could include:

- Decrease the loading time of the JSON. Options include exploring different data storage/retrieval methods, performing additional cleaning before running the application, and streaming the data in segments.
- Add a loading symbol during the GET request so the user understands what is happening.
- Improve the error handling and testing (in particular, ensure that a larger file size won't cause the browser to run out of memory).
- Add Google analytics to track traffic.
- Incorporate more datasets, especially the PEMs and SUMs data

As an ongoing study, the needs of the group are likely to change and, given the appropriate resources, the web application could evolve to support those efforts.

In consideration of the open-source visualization community, I would like to provide a tutorial on how to create such an application. Although the individual tools that I implemented are hardly novel, the way in which I pieced them together as a collective whole is unique. I cannot release the full codebase because of sensitive data, but a blog could be a sufficient platform to share this information for the benefit of developers.

The visualization of information is a creative means for humans to explore and explain real concepts. The process of working with the REACCTING project's data is an example of this powerful and unique means of communication and provides value to both the cookstove research and data visualization communities.

References

[1] G. M. Lewis, *The History of Cartography: The Origins of Cartography*, vol. 1. Chicago: University of Chicago Press, 1987.

[2] “10 Limits to Human Perception ... and How They Shape Your World,” *io9*. [Online]. Available: http://io9.com/5926643/10-fundamental-limits-to-human-perception----and-how-they-shape-your-world. [Accessed: 03-Nov-2014].

[3] “REACCTING,” *REACCTING*. [Online]. Available: http://www.reaccting.com/. [Accessed: 29-Sep-2014].

[4] C. French, *Data Processing and Information Technology*. Cengage Learning EMEA, 1996.

[5] L. A. Tedesco, “Lascaux (ca. 15,000 B.C.) | Thematic Essay | Heilbrunn Timeline of Art History | The Metropolitan Museum of Art.” [Online]. Available: http://www.metmuseum.org/toah/hd/lasc/hd_lasc.htm. [Accessed: 17-Nov-2014].

[6] The Institute for the Advancement of Research in Education (IARE) at AEL, “Graphic Organizers: A Review of Scientifically Based Research.” Inspiration Software, Inc, Jul-2003.

[7] H. Weinreich, H. Obendorf, E. Herder, and M. Mayer, “Not Quite the Average: An Empirical Study of Web Use,” *ACM Trans Web*, vol. 2, no. 1, pp. 5:1–5:31, Mar. 2008.

[8] “Leonardo da Vinci : anatomical drawings.” [Online]. Available: http://www.academia.edu/4033683/Leonardo_da_Vinci_anatomical_drawings. [Accessed: 27-Oct-2014].

[9] Jasmin, “The Human Hand [infographic] | Daily Infographic,” 26-Dec-2012. .

[10] E. R. Tufte, *Beautiful Evidence*, 1St Edition edition. Cheshire, Conn: Graphics Pr, 2006.

[11] F. L. Hawks, *Appleton’s Cyclopedia of Biography*. Appleton, 1872.

[12] Al Van Helden, “Galileo’s Sunspot Drawings,” *The Galileo Project*, 1995. [Online]. Available: http://galileo.rice.edu/sci/observations/sunspot_drawings.html. [Accessed: 17-Nov-2014].

[13] C. Beccario, “earth.” [Online]. Available: http://earth.nullschool.net. [Accessed: 17-Nov-2014].

[14] A. T. // Friday, S. 25th, and 2009-6:01 Am, “10 Reasons Why Advertising Campaigns Reach The Wrong Audience,” *AdExchanger: News and Views on Data-Driven Digital Advertising*. [Online]. Available: http://adexchanger.com/data-driven-thinking/10-reasons-why-advertising-campaigns-reach-the-wrong-audience/. [Accessed: 01-Apr-2015].

[15] “How to Lie with Data Visualization,” *Heap Data Blog*. [Online]. Available: http://data.heapanalytics.com/how-to-lie-with-data-visualization/. [Accessed: 01-Apr-2015].

[16] J. Corbett, “Charles Joseph Minard: Mapping Napoleon’s March, 1861.” 2012.

[17] *The Palette of Narmer (front and back)*. .

[18] A. Wakefield, S. Murch, A. Anthony, J. Linnell, D. Casson, M. Malik, M. Berelowitz, A. Dhillon, M. Thomson, P. Harvey, A. Valentine, S. Davies, and J. Walker-Smith, “RETRACTED: Ileal-lymphoid-nodular hyperplasia, non-specific

colitis, and pervasive developmental disorder in children," *The Lancet*, vol. 351, no. 9103, pp. 637–641, Feb. 1998.

[19] "Vaccine Myths Debunked," *PublicHealth.org*. [Online]. Available: http://www.publichealth.org/public-awareness/understanding-vaccines/vaccine-myths-debunked/. [Accessed: 01-Apr-2015].

[20] H. Millon, "The Architectural Theory of Francesco di Giorgio," *Art Bull.*, vol. 40, no. 3, pp. 257–261, Sep. 1958.

[21] "Visual Arts: Elements and Principles of Design." [Online]. Available: http://www.incredibleart.org/files/elements2.htm. [Accessed: 16-Dec-2014].

[22] "Jer Thorp | Speaker | TED.com." [Online]. Available: http://www.ted.com/speakers/jer_thorp. [Accessed: 17-Nov-2014].

[23] M. Weslander-Quaid, "Google: Creating a Culture of Innovation," presented at the The ATLAS Speaker Series | ATLAS Institute, ATLAS 100, University of Colorado Boulder, 22-Sep-2014.

[24] E. Tufte, "PowerPoint Does Rocket Science–and Better Techniques for Technical Reports." 2005.

[25] I. Parker, "Absolute Powerpoint," *The New Yorker*, 21-May-2001. [Online]. Available: http://www.newyorker.com/magazine/2001/05/28/absolute-powerpoint. [Accessed: 06-Oct-2014].

[26] "Global Alliance for Clean Cookstoves," *Clean Cookstoves*. [Online]. Available: http://www.cleancookstoves.org/our-work/. [Accessed: 03-Nov-2014].

[27] "REACCTING Report Protocol and Initial Results." .

[28] D. Harris, "How Splunk Is Riding IT Search Toward an IPO," 17-Dec-2010. .

[29] "Splunk," *Key Performance*. [Online]. Available: http://key-performance.eu/web_new/index.php/partners/technology-partners/splunk. [Accessed: 16-Feb-2015].

[30] "Tableau Business Intelligence," *Tableau Software*. [Online]. Available: http://www.tableau.com/business-intelligence. [Accessed: 30-Mar-2015].

[31] S. Murray, *Interactive Data Visualization for the Web*, 1 edition. Sebastopol, CA: O'Reilly Media, 2013.

[32] A. Moujahid, "Interactive Data Visualization with D3.js, DC.js, Python, and MongoDB," *Adil Moujahid*, 28-Jan-2015. .

[33] "alignedleft/d3-book," *GitHub*. [Online]. Available: https://github.com/alignedleft/d3-book. [Accessed: 31-Mar-2015].

[34] "chriswhong/taxitracker," *GitHub*. [Online]. Available: https://github.com/chriswhong/taxitracker. [Accessed: 16-Mar-2015].

[35] C. Whong, "Taxi TechBlog 1: Data Prep and Backend," *Chris Whong*, 20-Jul-2014.

[36] "adilmoujahid/DonorsChoose_Visualization," *GitHub*. [Online]. Available: https://github.com/adilmoujahid/DonorsChoose_Visualization. [Accessed: 01-Apr-2015].

[37] "Converting MATLAB's datenum to Python's datetime | The Sociograph." .

[38] D. Flanagan, *JavaScript: The Definitive Guide*, Fifth Edition edition. Sebastopol, CA: O'Reilly Media, 2006.

[39] imightbewrongbutidontthinkso's, "[jQuery] Using jQuery.noConflict() with Prototype/Scriptaculous," *jQuery*. 2009.

[40] T. Fuchs, “Slider,” *Scriptaculous*. [Online]. Available: http://madrobby.github.io/scriptaculous/slider/. [Accessed: 25-Mar-2015].

[41] V. Agafonkin, “Leaflet Features,” *Leaflet*. [Online]. Available: http://leafletjs.com/features.html. [Accessed: 01-Apr-2015].

APPENDIX

Big Data is a term for any collection of data sets so large and complex that it becomes difficult to process using tradition data processing applications.

Bower works by fetching and installing packages from all over, taking care of hunting, finding, downloading, and saving the stuff you're looking for.

Cron is driven by a crontab (cron table) file, a configuration file that specifies shell commands to run periodically on a given schedule.

CSS, or Cascading Style Sheets, is a style sheet language used for describing the look and formatting of a document written in a markup language.

D3 or Data-Driven Documents is a JavaScript library that uses digital data to drive the creation and control of dynamic and interactive graphical forms which run in web browsers.

Data Engineering can be described as an approach to designing and developing information systems. It can also be considered as the generation, distribution, analysis and use of information in systems.

Flask is a microframework (or light framework) for Python based on Werkzeug, Jinja 2 and “good intentions”. BSD licensed.

Git is a distributed revision control system with an emphasis on speed, data integrity, and support for distributed, non-linear workflows.

Github is a web-based Git repository hosting service, which offers all of the distributed revision control and source code management (SCM) functionality of Git as well as adding its own features.

Google Maps APIs let you embed and customize the robust functionality and everyday usefulness of Google Maps into your own website

HTML, or Hypertext Markup Language, is the standard markup language used to create web pages. It is written in the form of HTML elements consisting of tags enclosed in angle brackets (like <html>).

Javascript is a dynamic computer programming language. It is most commonly used as part of Web browsers whose implementations allow client-side scripts to interact with the user, control the browser, communicate asynchronously, and alter the document content that is displayed.

JSON, or Javascript Object Notation, is a lightweight data-interchange format that is easy for humans to read and write and easy for machines to parse and generate.

Leaflet is a modern open-source JavaScript library for mobile-friendly interactive maps.

MongoDB (from "humongous") is an open-source document database, and the leading NoSQL database.

Node.js is a platform built on Chrome's JavaScript runtime for easily building fast, scalable network applications.

NPM, or Node Package Manager, installs, publishes and manages node programs.

PEMs refers to personal emissions monitors. These monitors are worn by an individual (via a fanny pack or backpack) and measure the amount of pollution at a given time.

Python is a widely used general-purpose, high-level programming language that emphasizes code readability and generally allows programmers to express concepts in fewer lines of code than would be possible in languages such as C++ or Java.

Small Data is data that has small enough size for human comprehension.

SUMs refers to the data provided by the stove use monitors, which are digits located near the cookstoves that track variables like temperature and light.

Visualization is information that has been abstracted in some schematic form.