Title
An Online Tool for Personal Data Collection and Exploration

Author
Yau, Nathan Chun-Yin

Publication Date
2013

Peer reviewed|Thesis/dissertation

University of California
Los Angeles

An Online Tool for Personal Data Collection and Exploration

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Statistics

by

Nathan Chun-Yin Yau

2013

Abstract of the Dissertation

An Online Tool for Personal Data Collection and Exploration

by

Nathan Chun-Yin Yau

Doctor of Philosophy in Statistics

University of California, Los Angeles, 2013

Professor Mark Hansen, Chair

Advancements in technology and the changes in how people interact with data in recent years have given rise to online applications that allow people to collect data about themselves. For most applications, such as Twitter and Facebook, the collection is indirect. The primary purpose of the services is to share information with others. However, this regularly-updating online culture also provides a medium for personal data collection where people actively log data about themselves and their surroundings. This dissertation describes the development of *your.flowingdata* (YFD), an application that allows people to collect data via Twitter and to explore their data with a set of online visualization tools. Usage of the collection mechanism and visualizations is then described. Whereas most related work describes usage over a period of a week or less for in-lab users, YFD is a publicly available application and usage was studied over several months and for thousands of users. This provides a wider view into how general users, who are not necessarily "data professionals," collect and interact with their data. Study of YFD usage also provides insights for presentation of data to a wide audience and how to help them understand data.

The dissertation of Nathan Chun-Yin Yau is approved.

Deborah Estrin

Jan de Leeuw

David L. Rigby

Mark Hansen, Committee Chair

University of California, Los Angeles

2013

To Mom and Dad

Table of Contents

List of Figures

List of Tables

Acknowledgments

Thank you to my adviser Professor Mark Hansen for helping me reach this point in my graduate career. He provided many opportunities for me to explore possibilities and work on a wide variety of projects. The writing of this dissertation was not a straightforward process, and he pointed me in the right direction when I started to feel lost, along with a healthy dose of perspective. I would also like to thank the past and present members on my committee—Professors Katherine Hayles, Theodore Porter, David Rigby, Jan de Leeuw, and Deborah Estrin—who provided me feedback and support, which helped me raise the quality of my work. Thank you also to CENS, which provided an outlet to experiment.

Working from a distance, I often ran into challenges, but I had friends and family to help me stay on course. I almost certainly would not have finished without them as a support system. Thank you to my classmates who were there to talk. Thank you to my in-laws for the frequent words of encouragement. Thank you to my parents who always instilled confidence and motivation and throughout my life, have encouraged me to pursue what makes me happy. Finally, thank you to Bea, who is always there to listen and help, and for giving me more than I ever hoped for.

Vita

2004	Research Assistant, Statistics Department, UC Berkeley
2004	B.S. Electrical Engineering and Computer Sciences and Minor Statistics, UC Berkeley
2005	Teaching Assistant, Statistics Department, UCLA
2005–2006	Graduate Student Researcher, Statistics Department, UCLA
2006–2010	Graduate Student Researcher, CENS, UCLA
2007	M.S. Statistics, UCLA
2007	Graphics Editor Intern, *The New York Times*
2007–present	Editor, FlowingData

Publications

Yau N., "Seeing Your Life in Data", *Beautiful Data*. O'Reilly Media, Inc., 2009.

Yau N. *Visualize This: The FlowingData Guide to Design, Visualization, and Statistics*. Wiley, 2011.

Yau N. *Data Points: Visualization that Means Something*. Wiley, 2013.

CHAPTER 1

Introduction

Facebook has over a billion users and Twitter sees 400 million tweets per day. This always-developing, regularly-updating online culture provides a medium for personal data collection, and it has given rise to conferences, applications, and a general excitement about documenting one's life. This dissertation describes an application, *your.flowingdata* (YFD), that allows people from a general audience—from non-professionals to those professionally trained in statistics and analysis—to collect data about themselves and their surroundings within their daily routines. Its focus is on how users explore their data through a variety of custom tools and what this usage implies for visualization of personal data, as well as for general visualization made for a non-professional audience. Whereas most related work studied usage over short periods of time and a relatively small user base, YFD logged usage for about 5,000 users who individually collected half a million data points, which offers a more granular view of how people use and explore personal data.

Chapter 2 is an overview of the applications of personal data, frameworks in which to design a system that allows for the variety of applications, and general YFD architecture. While personal data can be useful for many things, from the individual to groups to anything that requires people to collect data about a topic, five main categories are described: journaling, personal informatics, identity, crowdsourcing, and citizen science. YFD was made mostly for the first two applications, which keeps data private and under the control of the user; however,

it was seen how usage can change and intertwine. For example, immediately after I made YFD available to the public, users requested a way to visually share their data with others, like the extension of a personal profile.

Chapter 3 describes how users log data on YFD. I created a flexible message syntax intending to use text messaging as the main collection mechanism, but eventually switched to Twitter, which was more efficient and cost-effective. Because YFD collection is partially via Twitter, the reader should understand how Twitter works. Twitter offers two message types, of which only the second type is used with YFD. The first is a public status update that is broadcast to a list of people. These public updates are called *tweets*. The second type of message on Twitter is a private one that is directed to another Twitter user. This is called a *direct message*. Direct messages can only be seen by the recipient. YFD users send direct messages to a Twitter account made specifically to receive data. The account username is "yfd." It is common to begin usernames with an at sign (@) to indicate one is referring to a Twitter account, so in the rest of this document I use "@yfd" when I refer to the Twitter account and "YFD" when referring to the application. Users send direct messages to @yfd with a syntax made to fit in how people already use the microblogging service.

Chapter 4 discusses the user interface and visualization tools made for YFD. The goal was to make users' data interactive, so that they could explore their data from different angles, hopefully leading to better understanding. Traditional visualization, such as bar charts and stacked area charts, were used, along with more browsable views such as a calendar heat map and standard lists. Some visualizations were application-specific, such as a tool to show the duration in between events. The tools were linked by common user interface elements and aesthetics, and whereas visualization research typically focuses on an overview first and details-on-demand approach, YFD presents details first and then overview second. It was also important that users be able to iterate over varying granularity.

Chapter 5 evaluates collection and how users explored data via visualization on YFD and provides design suggestions based on results. YFD users collected data out of personal interest and were offered no incentive and little guidance on what or when to log or look at their data. This posed challenges such as dealing with users who only logged data briefly to test the system, different start and end times, and varying expectations from people who happen to find the site. However, because usage occurred in a natural setting, the usage logs show a more accurate picture of how people collect and interact with personal data. Usage varied from short one-page visits, up to longer sessions of twenty interactions or more. A voluntary survey was also administered that asked things such as how often people used Twitter for purposes other than YFD or whether or not they grew more aware as they logged more data.

Chapter 6 provides a summary of the work in this dissertation and offers future directions for the application, as well as a broader view of presenting data to a wider audience.

CHAPTER 2

Personal Data and Applications

As technology and the Web advance, interaction with personal data grows more commonplace. People tweet on Twitter, update statuses on Facebook, and share location via foursquare. The data is stored online and accessing that data gets easier. With users accustomed to logging events, many people are eager to collect other types of data, which gave rise to topic-specific applications. There are many, but this chapter describes the applications in five main categories.

Journaling. People take photographs of memorable events and often return to collections to reminisce and reflect. Personal data collection can be applied in the same way, as a way to reflect, remember, and increase awareness of one's actions.

Personal Informatics. Data collected about the self, health-related in particular, can be used to estimate change, modify behavior, and "optimize performance."

Identity. When data and visualization are framed as a form of expression, the results can provide an image of who or what an individual or group is.

Crowdsourcing. Personal data that is useful to individuals can also be useful in aggregate, in areas such as urban planning and health.

Citizen Science The audience for personal data collection is often not trained in statistical analysis or data management ("non-professionals"), but when

experts are introduced into the loop, the extra guidance can lead to more scientific results.

The contextual nature of personal data and the variety of applications leads to different design approaches than that of traditional visualization, as well as different insight. An audience of non-professionals will view and use their data differently than those who spend hours at a time on analysis. These differences are described later in the chapter, along with an overview of YFD architecture.

2.1 Applications

Personal data collection can be useful across a wide range of applications from health to participatory sensing (Burke et al., 2006), across fields such as computer science, design, and statistics, with roots in both research and practice. It can be useful to the individuals who collect data about themselves in the way one would write in a diary and those who keep track of their habits to improve health. In aggregate, personal data can help small and large groups of people form a self-identity or achieve a larger goal.

The consistent component in these applications is the participants. Data collection is often a formal process; however, personal data collection is often outside a lab and done voluntarily out of personal interest. This poses different questions concerning visualization and interface design, which is the primary focus of this document, but before discussing design, the context and use of the data should be understood, as it affects how one should present the information and inspired some of the choices made when I implemented YFD.

2.1.1 Journaling

Long before the modern computer, Bush (1945) imagined a device called a Memex that stored an individual's paper records for quick and easy retrieval. Showing the time of the idea, he imagined a mechanism where all of these personal records were put on microfilm. In a more modern version, Bell (2001) developed CyberAll as a way to store everything in one's life digitally on a hard drive. Five years later, Gemmell et al. (2006) described the evolution of CyberAll into a more technologically advanced MyLifeBits, noting the increases in amount of data and ease of collection.

The motivation behind these projects was not to make an immediate improvement in one's self or to change behaviors based on quantitative evidence. Rather, their purpose was to augment one's memory, so that if a piece of information, such as a receipt or an old article draft, were needed, it would be available for retrieval.

At the same time, memories and experiences hold sentimental value to the owner. Gemmell et al. (2006) described MyLifeBits as a "surrogate memory" and expressed an "emotional blow" when a terabyte of data was lost due to a malfunctioning hard drive. The sentiment is similar to how we value photographs and diaries. We store artifacts of memory on hard drives and photo-sharing sites like Flickr and document events on blogs and platforms like Xanga and Wordpress. We come back to them months or years later to reflect. Sometimes the retrieval is on purpose, and other times we stumble upon old memories while looking for something else.

I initially created YFD with photo collections in mind. Emotions are often attached to photographs, which might be irrelevant to onlookers but meaningful to the individual. It is not just a picture of two people walking around. It is the picture of a couple on their wedding day about to make a lifelong commitment

Figure 2.1: Noah K. Everyday

to each other. It is not just a picture of a baby. It is the memory of holding one's newborn son for the first time. Similarly, individual data points, while quantitative, can carry the same qualitative weight because they are personal to the individual who logged them.

We usually take pictures during significant events that are atypical of a regular day; however, Kalina (2010) takes a picture of himself every day (Figure 2.1). At the time of this writing, Kalina has done this for twelve years. At the six-year mark and again at the twelve-year mark, he pieced each photograph together to make a time-lapse video of his aging self. As a whole, the collection of pictures represents a large part of Kalina's life, despite being only a small snapshot from the entirety of each day. Kalina's face changes over the years as well as the background when he moves to different apartments. While mostly entertainment for outsiders, the evolution in the photos shows something more meaningful to the individual. This is perhaps best demonstrated by the number of similar videos that have been created since. Although mostly only meaningful to Kalina, his project motivates others to document their own lives, which in turn is meaningful

to them. For example, some record the progress of their pregnancies (YouTube, 2007), whereas others track changes with weight loss and muscle gain (YouTube, 2008).

This can be extended to email that is archived more often than deleted, books and documents that are more commonly digital, and status updates on social networks that grow more frequent and data-rich. What happens when personal data is collected at higher granularity? Wolfram (2012a) analyzed a third of a million emails he sent between 1989 and 2012 and in a time series plot, he notes the years he stayed up late at night to write a book, when he sleeps, and sat down to eat dinner with his family. Similar insights are seen through his keyboard keystrokes and phone calls. This is in addition to the online scrapbook that Wolfram keeps to document significant events in his life since birth (Woflram, 2012). Kleinberg (2003) also analyzed his own email to model "burstiness" and structure in document streams.

Felton (2011) collects data on his personal habits, activities, and behaviors, such as books read, restaurants eaten at, places traveled to, and his mood, and then designs a graphical summary based on the data. The Annual Feltron Report, which as the name suggests, comes out each year. As shown in Figure 2.2, it looks like a business report—but for an individual—with maps and time series charts. Each year, Felton sells thousands of copies of the report to people who mostly only know him by his online persona. The collection of charts resonates with readers in some way, as if they were to read a stranger's personal recollection of a year or view a self-portrait hanging in an art gallery.

Popularity of these individual design works has given rise to applications that make it easier for others to do the same. For example, the online application Daytum (Case and Felton, 2010) launched after the creators saw a desire from fans of Felton's annual reports. Users can track personal metrics and build dashboards with standard statistical graphics, such as bar graphs and pie charts. Similarly,

Figure 2.2: 2008 Feltron Annual Report

DailyBooth (Pokorny et al., 2010) lets users take self-portraits every day and archive it online like Kalina. As of 2011, 13 million photos have been uploaded to the site (Tartakoff, 2011).

Relatively speaking, Daytum and DailyBooth serve a niche audience, but the idea of storing your life online has found its way into more widely used applications. For example, Facebook, which has over a billion users at the time of this writing, lets people create profiles and connect with friends and family; however, in 2011, the social network shifted focus of a person's profile from recent updates to a timeline that encapsulates a person's life from birth to present. Pictures, status updates, and major life events are shown in chronological order. Facebook hired Felton earlier that year. Similarly, Google promoted two of their services, Google+, their overarching social layer (Google, 2012c), and Gmail, their online email application (Google, 2011a), as ways to collect memories. However, the Facebook timeline and Google products arguably lack the same outsider attrac-

tion as Felton's reports, even though Felton was a designer for the former, which seems to suggest that his reports are interesting because of the high granularity of data coupled with graphic design.

YFD can also be used as a journal. For example, two parents used the application much like one would use a scrapbook for a newborn. They kept track of things like weight, feeding times, and sleeping times, so they could see trends and patterns in behavior. However, this was less about modifying behavior and more about documenting the life of a child.

2.1.2 Personal Informatics

Although data journaling focuses more on the long-term than the immediate future, the frequent collection of data by individuals naturally leads to more statistical usage. Personal informatics, often referred to as the quantified self (Kelly, 2012), self-experimentation (Roberts and Neuringer, 1998), or self-surveillance (Yau and Schneider, 2009), embraces personal data collection as a way to measure and improve health, well-being, and physical performance. This was clearly the goal of many YFD users who noted that they lost weight, started to drink more water, and found the time of day they were most productive.

Roberts and Neuringer (1998) describe self-experimentation in which an individual conducts studies on one's self. Like Felton (2011), Roberts (2004) collected data about his sleep, mood, weight and other health-related aspects of his life, but instead of an end-of-year reflection, Roberts actively tried to change his behavior and distinguish cause-effect relationships. By modifying his diet and trying various exercise and sleeping routines, he found several.

Roberts collected most data manually, but current technology allows a more automated process. This is highly visible in athletics, where companies have created wearable devices that record and upload data to a computer or a server.

Figure 2.3: Nike+ Fuelband

For example, athletics company Nike sells a wristband that tracks movement and a sensor worn in the shoe that records location and steps, as shown in Figure 2.3. The data is linked to an online server, which can be viewed on a website or via mobile applications (Nike, 2012). Jawbone (2012) and Fitbit (2012) provide similar devices.

Automated data collection lowers the barrier to entry and can provide detailed information over time. Online applications such as personal finance site Mint (Patzer, 2010), shown in Figure 2.4, and time-tracking site RescueTime (Hruska et al., 2010) use this to their advantage. The former keeps track of credit card charges, bank statements, investments, and loans, depending on how much information you choose to provide. In turn, users get an aggregated view of their finances in one place, and Mint can also provide suggestions on how to save money, based on spending habits. With RescueTime, users install a plugin on their computers, and it records applications used and websites visited. The typical goal is to manage one's time wisely, but some use the application purely for curiosity.

As services collect data on a user's behalf, privacy policies and ownership

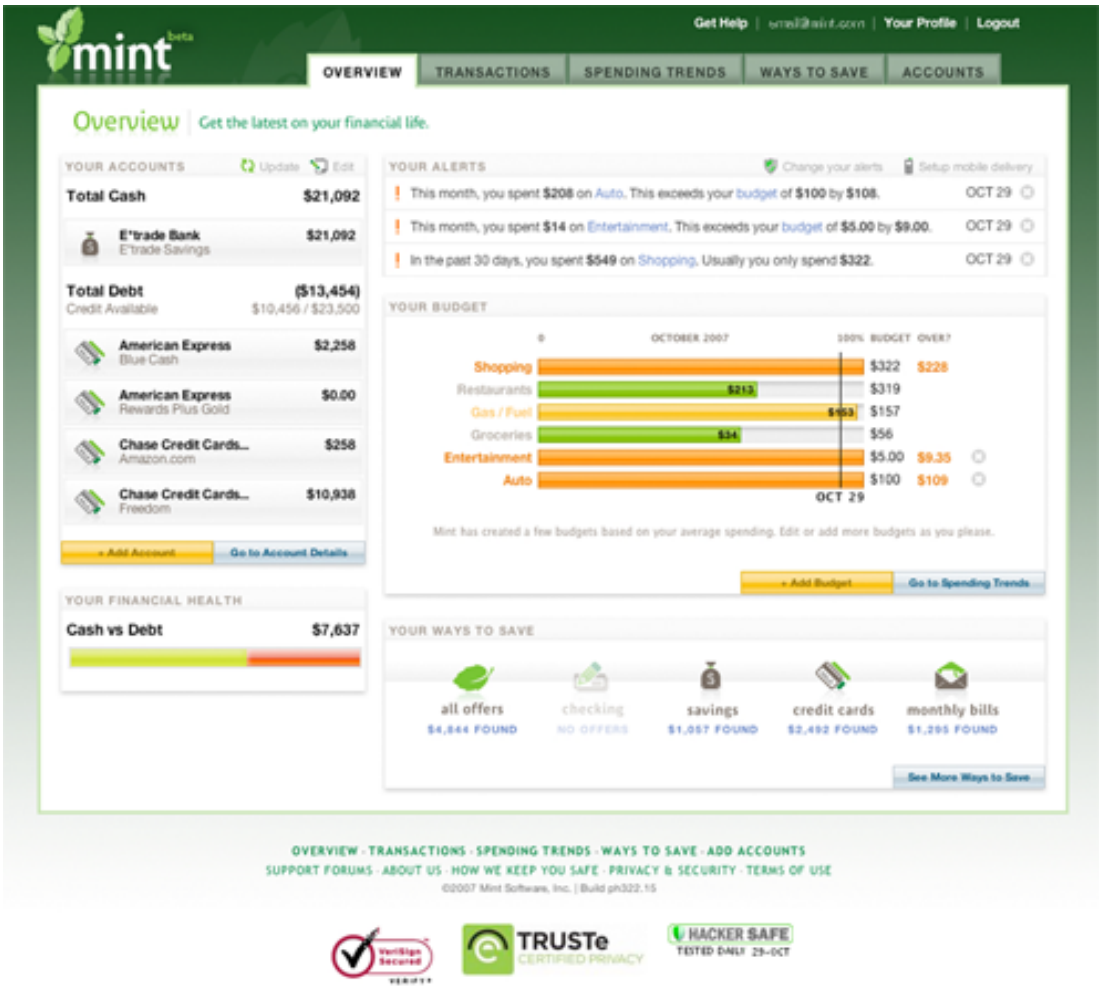


Figure 2.4: Mint Personal Finance

vary. For example, Mint and RescueTime do not sell or distribute identifying information, and your data can be downloaded or deleted at any time. However, they reserve the right to redistribute anonymized aggregates to third parties, and whereas users can download all transactions on Mint for free, RescueTime users have to pay a monthly fee to download their data. Similarly, data exports on Fitbit are only available to those who pay for the premium plan. YFD users can download their data or delete their accounts at any time.

2.1.3 Identity

Most online services frame an individual's account as a profile, which is a reflection of the user. The data that is shared affects what a profile looks like, and so the data affects that public image. Twitter and Facebook are the most prominent services that do this, where status updates and pictures make up identities and likes and retweets serve as a form of social validation.

Because others' profiles are also visible, a user can compare against and interact with friends and followers. Wolfram (2012b) took this opportunity to turn one's Facebook profile into analytics, so when an account is linked, the software reports friends' birthdays, who comments on posts, what percentage of friends are in a relationship, and which ones have the most in common with the user. Similar services such as LinkedIn (2011) and re.vu (2011) use data from LinkedIn, a social network for professional connections, turn job and education data into visual networks and resumes. Mycrocosm by Assogba and Donath (2009) lets users do this manually but frames data and charts as a way to communicate or a medium for expression. Instead of status updates, a profile on the site shows a series of graphs that represent things like what clothes a person wears, time spent napping, or the charts might just tell jokes (Figure 2.5).

One of the most common YFD requests was the ability to share data visually

Figure 2.5: Profile from mycrocosm

with others. I originally thought that people would want to keep all of their data private, so I was surprised that many wanted to make most of their data public via a dashboard-like view. I implemented a public-facing view that users could customize by moving around modules that represented facets of their data. Several shared books and movies they thought were entertaining and some published eating and weight as a form of motivation. One used the public view as a way to update his girlfriend on when he went to bed, because he was trying to sleep at an earlier hour. This of course, relied on the honor system.

2.1.4 Crowdsourcing

Personal data collection can also lead to an identity for groups of people, small and large. Between 1936 and 1945, hundreds of untrained volunteers mailed observations of the everyday in Britain to Mass-Observation. They reported things like the behavior of people at war memorials, female taboos about eating, and shouts and gestures of motorists. Mass-Observation, founded by anthropologist Tom Harrison, poet Charles Madge, and film-maker Humphrey Jennings, used these observations to describe everyday life (Holt, 2005). Although the reports were likely biased due to the self-selecting nature of the surveys, they provide a narrative that hints at what life was like at the time.

That said, because the voluntary contributions to Mass-Observation were often unstructured with non-specific instructions, it is hard to say with any certainty how accurate the reports are. They are more like anecdotes than they are comprehensive data, commonly thought to be more qualitative than quantitative (Hubble, 2010). Reports were compiled by editors rather than statisticians or analysts.

The nature of the survey goes back to the beginnings of the project. In 1936, displeased with how newspapers covered the abdication of King Edward VIII to

marry Wallis Simpson, who was twice-divorced, the Mass-Observation founders called on volunteers to write about the event. The result was a collection of anecdotes about what people looked like and what people said, which were compiled into a book. In contrast, around the same time, George Gallup, founder of the Gallup Organization, was working in the United States to judge public opinion objectively and predict presidencies (Gallup, 2012).

Everyblock (2012) is driven by similar goals as Mass-Observation, but is more focused on news and events than on public opinion. The site was originally designed to deliver local news through publicly available data, such as crime reports and restaurant inspections, but in 2011, Everyblock redesigned with a community focus (Holovaty, 2011). People can contribute to feeds about their neighborhoods, which would otherwise go unreported by news outlets, and this is accompanied by automated data feeds. So anecdotes and objective data reinforce each other; however, the shift in focus from time series charts and maps to a list-like news feed is perhaps an indicator for how the general public relates to and consumes data. Figure 2.6 shows the initial chart-centric version and the current event-based version.

Services that have gained widespread usage, in the magnitude of millions of users, tend to focus on individual data points logged recently more than on visualizing overall trends over wide time spans. For example, Foodspotting (2012) is an application that encourages users to take pictures of food dishes, and share the photographs with location attached. Individuals get to share pictures, and onlookers can see what food is in a geographic area. The Foodspotting mobile application has been downloaded three million times and people have uploaded two million photos (Ha, 2012).

Health-related applications, such as CureTogether (Carmichael and Reda, 2008) and PatientsLikeMe (Heywood et al., 2004), help users keep track of conditions and share treatments. To individual users, the service is valuable in keeping track

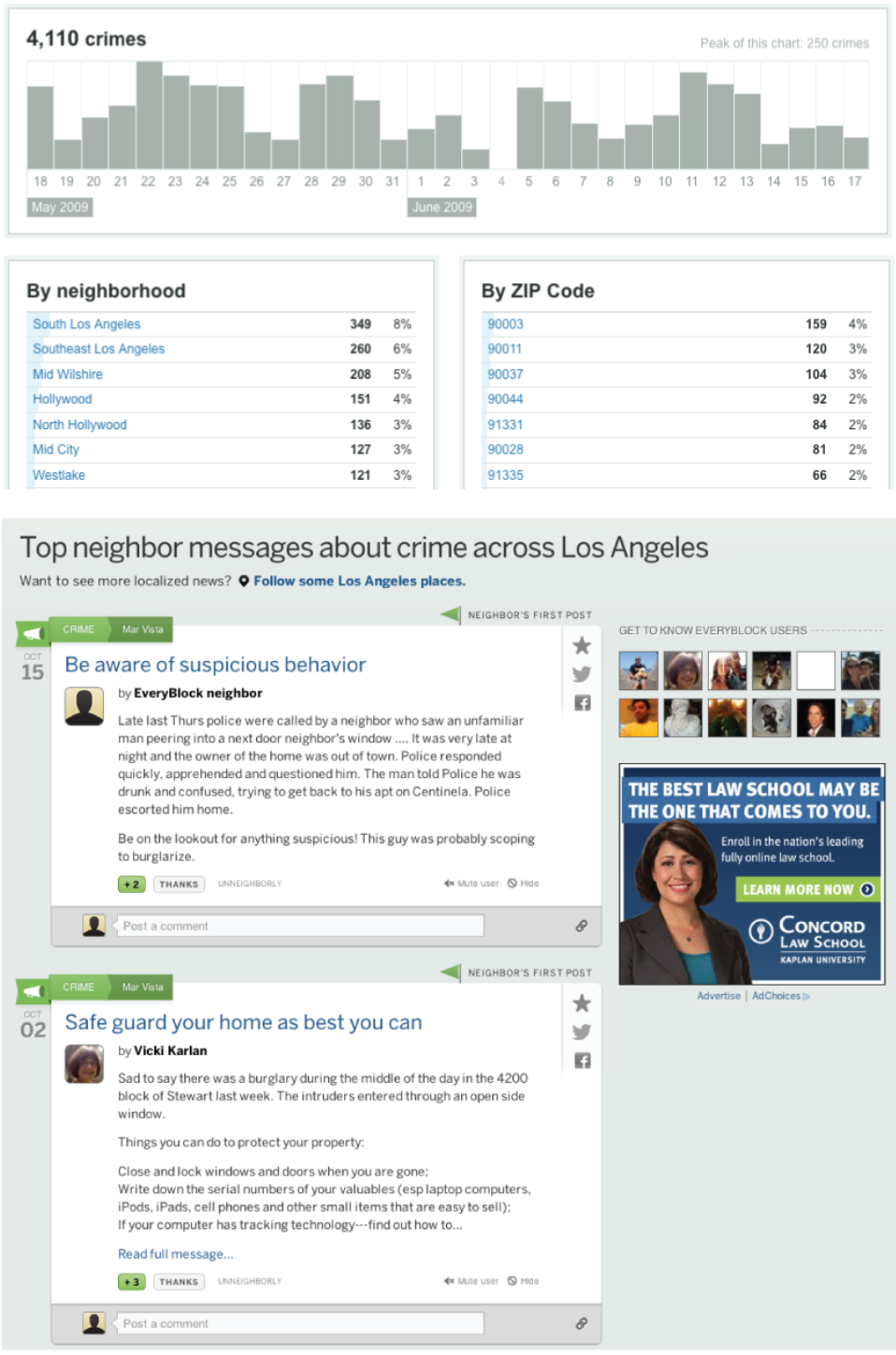


Figure 2.6: Everyblock, original (top) and current (bottom) versions

of chronic conditions, and to groups of people with the same symptoms or medical conditions, the service helps users see how others cope and what has worked and what has not. MyFitnessPal (2012) is a more focused online and mobile service that lets people keep track of weight, exercise, and eating. Users can look up caloric values for food items, or if an item is unlisted, users can contribute to the food database, which others can use. Users can also enter what type of diet they are on, which MyFitnessPal uses to report in aggregate on how well diets work, based on current users weights.

Such online communities have grown to be commonplace. Twitter was built on a similar premise: to share individual updates to a group of people who might be interested in where someone is what he or she is doing. However, as more people used the service to update status and to share news, Twitter aggregated tweet topics and estimated immediate trends to show what users currently talk about (Dorsey, 2008). Trending topics for cities was offered later and was eventually refined to show topics specific to the users one follows on the service. The aggregates give people a way to follow the news and major events, as well as find sites to pass time.

However, because millions of people use Twitter, marketers create commercials in hopes of getting people to talk about a company or a product, and spammers try to game the system for financial gain, either by injecting irrelevant links about trending topics or changing the topics themselves. Some create fake accounts. As spam can make the service unusable and irrelevant, Twitter had to develop systems to identify such tweets. Twitter also filed lawsuits against developers who created tools that made spamming the service easier (Twitter, 2012). Such challenges pose questions for research in aggregating personal data, such as with data privacy and authenticity.

2.1.5 Citizen Science

As with Mass-Observation, lack of direction and unstructured data can lead to reports that are difficult to interpret (Hubble, 2010). Participants were for the most part not "data professionals." Of course, it is not practical to require that every observer have heavy experience with data. However, it is possible for someone who knows about data collection to direct volunteers towards a more scientific method. This is the idea behind citizen science, which crowdsources data gathering. Novices can browse the data, but topic experts can also analyze the data for deeper insight. The result resembles the previous applications, but an expert is introduced into the loop.

For example, the Audubon Christmas Bird Count (Audubon, 2012) is a citizen science project that aims to create an annual census of birds in the western hemisphere, which provides population trends. Thousands of volunteer birdwatchers gather at 2,000 counting circles across the country during the winter months and count all the birds they see within a 7.5-mile radius. There are counting rules, such as participants are not allowed to count birds while retracing steps on a trail. This prevents double counts. The annual data is available to participants as well as researchers and conservation biologists.

OpenPaths (2012) is an application to raise awareness of the data that people generate by carrying phones and using the Web. The OpenPaths mobile applications let users record location and upload anonymously to a server. The data can be downloaded or viewed via the tools on the site. In addition, data can be contributed anonymously to researchers who are interested in say, looking at how people move around in an urban setting. Those who are interested create a project and request data from OpenPaths participants, and those participants can grant or deny access by issuing an encryption key that can be revoked at any time. In contrast to other services that take at least partial ownership of user

data and analyze aggregates or anonymized records, OpenPaths lets users control who sees and uses their data.

Advancing technology has also allowed people to report and record events as they happen, which has developed into a useful tool to inform the public and to make quicker decisions. For example, Did You Feel It?, maintained by the United States Geological Survey (USGS), lets people share information about an earthquake they felt, which can contribute to quick assessment of the severity of an earthquake emergency as well as earthquake research (Wald et al., 2006).

The USGS site provides a simple map interface to show where earthquakes have occurred, but the significance of citizen mapping efforts was perhaps felt most after the Haiti earthquake in 2010. WikiProject Haiti by OpenStreetMap (2010) led a collaborative effort to produce an authoritative and current map of Haiti showing roads, damaged buildings, and camps of displaced people based on satellite imagery (ITO, 2010). OpenStreetMap (2011) led a similar mapping collaboration for the earthquake and tsunami in Japan in 2011. Additionally, Google (2011b) allowed people to upload and location data to help friends and family find each other after the tsunami. This real-time reporting by individuals has naturally lent itself to journalism, such as iReport by CNN (2012), which like Everyblock, allows people to share events. However, editors filter and decide which reports become a part of official CNN news coverage.

From the art side of the spectrum, Koblin (2006) created The Sheep Market, as shown in Figure 2.7. A collection of 10,000 sheep were drawn by "workers" on Amazon's Mechanical Turk via Koblin's interface. The project is both a demonstration of the then new technology and a reflection of what motivates people to do things. Koblin has since collaborated in projects under the same motivation including Bicycle Built for Two Thousand (Koblin and Massey, 2009) and The Exquisite Forest (Koblin et al., 2012).

Placing users in the role of analyst, Heer et al. (2007) developed sense.us,

Figure 2.7: The Sheep Market

a prototype web application for "social data analysis". A suite of visualization tools were provided that allowed users to explore 150 years of data from the United States Census Bureau. Users could explore the data via visualization and comment on findings asynchronously. Many Eyes by Viegas et al. (2007) is like the next iteration of sense.us in that it allows users to upload their own datasets. Again, users can collaborate and comment on each other's visualizations, but a browsing of the site shows discussion is limited and the authors only discuss preliminary usage. Perhaps because random users are not familiar with others' data, there is limited discussion.

2.2 Building Insight

Again, it is worth emphasizing that although the applications of personal data collection can vary, the audiences are similar. Participation is often voluntary and users are typically not data professionals, but this also means that even though users might not be working statisticians, they collect data out of personal interest and are eager to learn about or reflect on what they collect.

Visualization that helps users relate to their data, but also allows them to explore deeper, played the main role with YFD. Visualization can be thought of as an interface to data whose design can change how trends and patterns are perceived, the types of inferences that users make, and provide a better understanding of

how day-to-day decisions intertwine with each other. Interactive and exploratory graphics can also help users form new hypotheses and modify preconceived ones. For example, although the goals of personal informatics might differ from those of journaling, the data generated by someone interested in self-improvement can be equally as useful in a data journal (or the same), and vice versa. Purchases can be a signal for life events, such as marriage or a new home, or an increase in time spent on baby name sites can be a signal that a baby is on the way. On the one hand, users try to save money or time, and on the other, users reflect on the past. At the same time, these data sources on spending, computer usage, and health can all affect the others.

Pousman et al. (2007) describe four types of insight in the context of a proposed subdomain of visualization, casual information visualization. The first is more quantitative, whereas the other three are more qualitative and harder to assess.

Analytic Insight. This is the traditional type of insight that comes from statistical models, testing, and analysis. Results are typically quantified.

Awareness Insight. This comes from remaining aware of data streams such as the weather, news, or stock fluctuations. By staying in view, the data becomes part of a person's everyday.

Social Insight. Through involvement in social networks or spending time in groups, people gain insight on how they fit in and how to interact with others.

Reflective Insight. When users take a step back from the data or view it in a way they are not used to, they can reflect on their lives, often from an emotional standpoint.

Casual information visualization is designed around the last three insights more

than analytic insight. Whereas more traditional visualization research is centered around increased analytical functionality for data professionals, who spend hours at time with a dataset, casual information visualization focuses on viewing data sporadically or keeping it in the background.

Ishii and Ullmer (1997) approach this challenge in the context of transferring digital bits to physical objects and ambient displays. They imagine speaking with a colleague in the office, but at the same time we are aware of the weather outside, cars driving by, and other conversations around us. If something odd happens in the background, we can easily shift our attention. We are aware of surroundings but not actively thinking about them. Similarly, Hallnas and Redstrom (2001) describe a design philosophy for "slow technology" where technology is designed for reflection rather than to increase efficiency or to optimize activities.

For personal data, we want to design a system that allows for these various types of insight. As discussed in Chapter 5 on usage, people spend the majority of their time casually browsing their data and only occasionally do they look deeper using traditional visualization. However, this is not to say that personal visualization should exclusively be casual. Instead, different views should be combined so that when something interesting is seen while browsing, users can easily switch their direct attention to that area. This leads to a system that is flexible enough for those interested in both data journaling and personal informatics, without requiring users to switch applications.

As described by Shneiderman (2003), a common visualization workflow starts with an overview of the data, lets users zoom and filter, and then details are available on demand. However, when a user logs in to YFD, the details are presented immediately in list form, and the user can move to overviews and filters after. This was especially important for new users, because with only a few data points logged, views such as stacked area charts or treemaps are not useful. The casual views are a way to see if data collection worked and a way to show

immediate change, even with just one additional data point. People need to see potential benefit, which is possibly why Felton's annual reports, Noah Kalina's photo project, or my own sharing of personal data on YFD generate interest.

The ability to easily switch between views also support a link between individual data points and aggregates of many data points. This helps non-professionals make a connection between the points that they log and the visualization that abstracts the numbers to show trends or hierarchy. From statistics education, Bright and Friel (1998) found that students often described the heights of bars in a histogram as the magnitude of an individual or observation. For example, when shown a histogram that provided the distribution of people in a group, students thought that each bar represented an individual's height rather than a cluster of people in a given height range. We observed the same misread in early versions of the visual interface for the Personal Environmental Impact Report (PEIR, Mun et al. 2009). Histograms were used to show distributions of an individual's carbon impact, but users were confused because they thought the horizontal axis represented a segment of time rather than a set of values.

Li et al. (2010) propose a stage-based model for personal informatics that consists of five stages: (1) preparation, (2) collection, (3) integration, (4) reflection, and (5) action. During the first stage, users identify what data they want to collect and how they want to collect it. Users log data during the next stage, which can be with paper and pencil or with a dedicated application such as YFD. Integration frames the data in a way that users can interpret it, such as transferring tick marks written in a notepad to a spreadsheet or in the case of YFD, saving direct messages from Twitter to a database. Once the data is easily accessible, users can reflect, analyze, and explore and perhaps take action based on findings. Users can iterate within and in between the stages. For example, reflection can lead to further preparation and collection of new data types before a user takes action to change behavior.

Although Li et al. (2010) frame the stage-based model in the context of personal informatics, where there is an implied goal to change behavior (e.g. diet), the model seems apt for systems with other applications, too. Users with other interests simply spend different amounts of time in each stage. For example, someone who journals will spend more time in the collection through reflection stages and less time in the action stage than a self-experimenter. The nature of reflection can also vary. It can be emotional and qualitative, or it can be a more quantitative and objective thought process. The goal for YFD was to support these different types of reflection and insight, and thus supporting multiple applications and motivations. (See Chapter 5 for a detailed description on usage and user insight.)

Consolvo et al. (2009) suggest eight design strategies to support behavior change in everyday life, which includes representing data in other ways than raw numbers and tables; unobtrusive collection and exploration in a user's everyday; and controllable data manipulation. Again, although these strategies are framed in the context of behavior change, we can apply them to personal data collection more generally, regardless of application. I used Twitter because people who regularly use the service are accustomed to regular updates on what they do or what goes around them. YFD grows more useful when people log more data, so when users log in, they see how their current collection volumes compare to thirty days ago. Users also spent a lot of time with a standard list view, which allows them to edit and delete data points. Finally, the collection and exploration systems were made flexible enough to support a wide variety of data types.

2.3 YFD Architecture

The architecture of the YFD application can be divided into three categories: storage, collection, and exploration and visualization. Users do not directly inter-

act with the storage component, but they are a main design consideration with the application as a whole. This section describes the technical implementation of each category, how the parts fit together, and the process behind building the application. I explain collection and visualization in more detail in Chapters 3 and 4, respectively. The goal is to build a personal data application that fits the following requirements, based on experience and usage, discussed in Chapter 5:

- Collection should be flexible enough to let users collect the data they are interested in and to change data types as that interest changes, but to still make connections between old and new data.
- Let users quickly see most recently logged data, but make older data easy to access.
- Provide visualization that lets users casually browse their data, alongside more traditional exploration tools.
- Link multiple views with similar interaction and aesthetics and data commonalities, such as time.
- Help users make a connection between individual data points and aggregates through the visual interface, which leads to inferences.

2.3.1 Storage

User data is stored in a MySQL database (Oracle, 1995), which allows straightforward queries over several tables, as shown in Figure 2.8. User-logged actions are stored in a single table, which includes columns for keywords, values, and timestamps. Metadata for each unique action is also stored per user. For example, a user might log data for food consumption with "ate" as a keyword. A description for that keyword can also be entered, which is stored in a table in the database. If YFD were used by more people—on the scale of millions—with more frequent

writes and accesses to the database, the table that stores user-logged data would need to be distributed like we did with Sensorbase (Chang et al., 2006). However, the YFD user base is smaller, so scalability was not a concern.

The user tables store username and Twitter-related information, such as time zone and whether or not a user is followed, so that he or she can send direct messages to @yfd. The actions tables store user-logged data, meta data, which is also entered by users, and tags entered as hashtags using YFD syntax. The cache tables store regularly updated information to help the site run faster and to automate data processing from Twitter to the database.

To study usage, I also stored interaction data in the database, such as pages loaded and items clicked, as well as survey questions and answers. Having the usage studies as part of the application, made implementation more straightforward and easier to link survey answers to interaction on the site.

2.3.2 Collection

Twitter, known for an online culture of frequent updates, provides a flexible API that allows other applications to make use of social service's functionality, so YFD users can collect data via SMS, desktop applications, mobile applications, and other Web services. With Twitter in the YFD data pipeline, collection is relatively straightforward using a YFD-specific syntax. Here are the main steps that are taken when a user wants to collect data with YFD:

Interest. The user takes interest in a behavior or action.

Action. The activity or behavior occurs.

Collection. A message or tweet is sent to Twitter to log a data point.

Storage. The tweet is parsed by YFD and stored in the database.

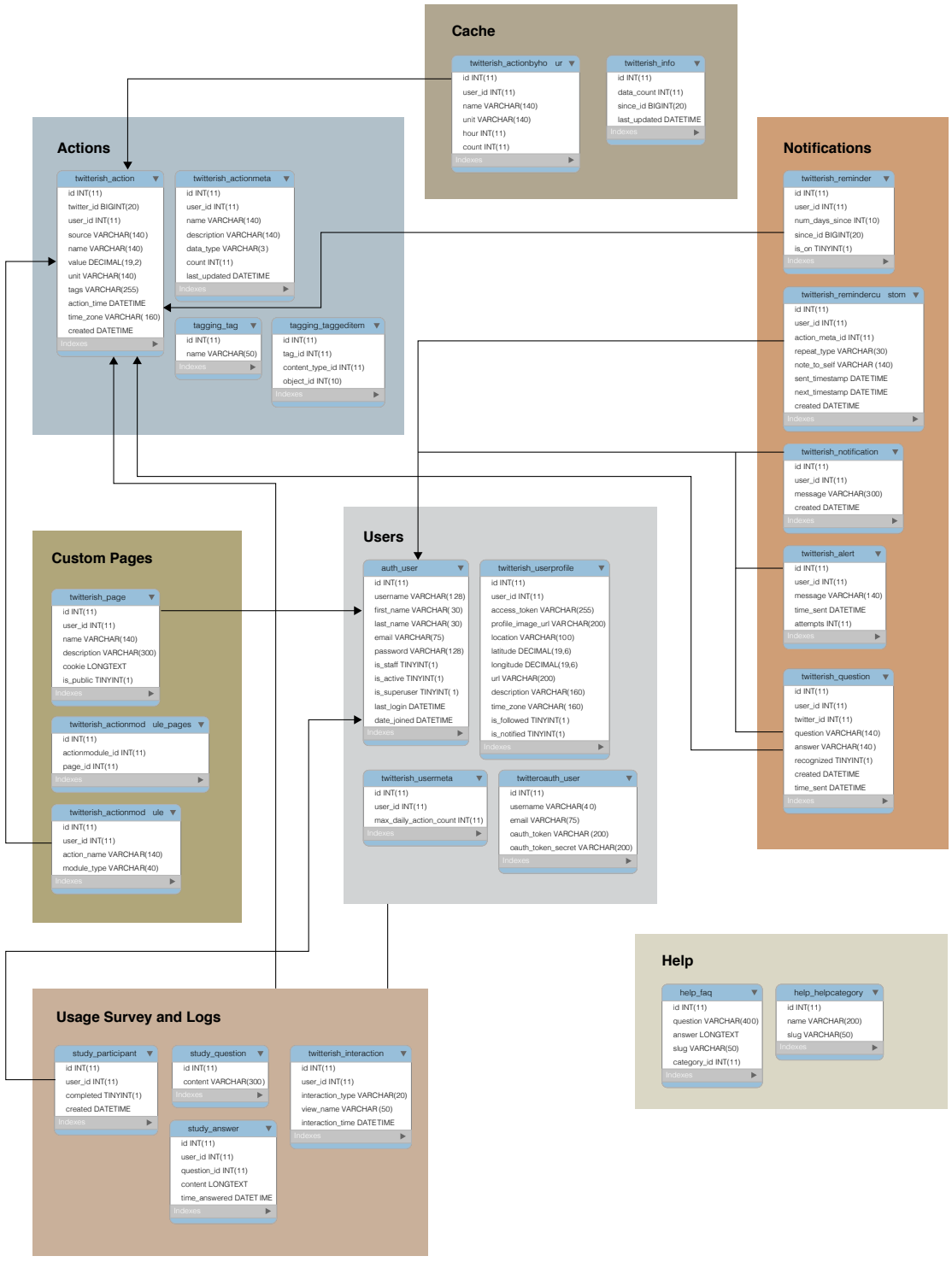


Figure 2.8: YFD Database Schema

Users must take an interest in some aspect of their lives that they can collect data about. Usually people have something in mind such as weight or food intake. Some might be interested in logging the books they read and the movies they watch. Others might want to keep track of mileage on their car. It is up to the user. Although I do suggest that new users start with a single metric to familiarize themselves with the YFD syntax and work data collection into their routine. Once they grow more accustomed to YFD, it is easier to track more.

To log a data point to YFD, the user sends a direct message on Twitter to @yfd with the syntax described in Chapter 3. When a message is received, YFD parses it and stores the data in a database. The data is visible on the site at this point and can be explored via visualization tools or downloaded as plain text. Users can also log data with the same syntax via a text field on YFD. This was common feature request and was added later, so users did not have to open a Twitter client to log data while already on YFD.

The main motivation for using Twitter was to lower the barrier to entry, since users log in to YFD with a Twitter account via OAuth, and to widen the opportunity for users to work personal data collection into their everyday routines. Those who use Twitter already tend to publicly update followers on what they are doing (e.g. eating or traveling), so it is less of a stretch to log data about personal activities in a private framework than for someone who never enters status updates on services such as Twitter or Facebook.

2.3.3 Exploration and Visualization

The YFD frontend was developed using Django (Holovaty, 2010), a Python Web framework, which uses the Model-View-Controller (MVC) design pattern (Krasner et al., 1988). It allowed for automatic URL generation, separation of data and views, and because it was in Python (van Rossum, 1991), it was more straightfor-

ward to implement data processing scripts on the server, which were also written in Python. Django handled most data processing and web page generation, however, visualization tools were implemented in a combination of Flash and Actionscript (Systems, 2012), JavaScript (Eich, 1995), and HTML and CSS (Consortium, 1995).

As described earlier in this chapter, the YFD visualization tools were made to support different applications, such as journaling and self-experimentation. So there are basic views such as an actions log which simply lists data in reverse chronological order. Users can edit and delete data points in this view. A calendar heat map visualizes data in the familiar grid format, with weekends on the ends and weekdays in between. A tag cloud, which sizes words based on volume of use, provides a standard web view into a user's data, whereas a treemap and stacked area chart provide more statistical views.

The central goal in making these tools was to allow users to interact with their data and to let them see their data from different angles. Noted by Bakker and Hoffmann (2005), students were able to understand the concept of aggregates and distributions, in addition to individual data points, when they saw data through multiple views. It seems fair that this finding also applies to non-professionals interested in data about themselves.

For example, across the different views available on YFD, users can select the segment of time they want to look at. There are preset timeframes, such as the past week, past month, or past year, but users can also select their own start and end dates. As users switch from say, the stacked area chart to the treemap, the timeframe stays consistent. This is similar to the brushing interaction described by Becker and Cleveland (1987) and later refined by Theus (2003) and Swayne et al. (2003), but across multiple views not necessarily on the same screen. With individual visualizations, users are able to search their data, which makes non-relevant data fade from the screen, but it is straightforward to return to the

original view. Finally, users can always return to the homepage where data is presented in list format, so it is easy to cycle through the views. This also lets users choose at what depth they want to view their data.

2.4 Summary

Personal data collection is often thought of as a way to quantify one's life, with the purpose of improving one's self or changing a behavior such as dieting or smoking. This is a great way to make use of personal data, but there are a variety of other applications such as journaling, forming an identity of one's self or a community, or it can serve as a source towards a greater cause in citizen science.

The tools built to browse or explore such data depends on the audience and application. If users spend most time casually browsing data or tend to leave it in the background, it makes sense to spend more time designing tools that fit in with a non-professional's everyday. Tools built for more traditional analyses can still be widely useful, but they can also be overkill, or they might abstract the data so much that users are not able to identify with the trends and aggregates. A complete personal data system allows users to quickly iterate between stages of collection and reflection, as well as provide multiple views which encourages such iteration.

CHAPTER 3

Collection

YFD enables data collection via Twitter, however, the first version of YFD did not use the service. Instead, YFD was built around email. Users sent short messages to a designated email address, the message was parsed, and the data was stored in a MySQL database. Data collection via email introduced a few challenges. First, when one sends an email, the message is usually longer than what time one woke up in the morning or what was eaten for breakfast, so users had to change how they use email. It was also important that people could log data when something happened, rather than try to remember what happened and log data when a desktop computer was available. While developing YFD, mobile email was too roundabout for the data collection process to fit into one's routine. Finally, technical issues with the email server simply led to a search for an alternative.

SMS seemed to be a natural progression, but Twitter's popularity was growing and the service offers a flexible API for developers to build applications on top of the basic message service. At the time of this writing, Twitter has over 500 million registered users around the world who send on average of 400 million tweets per day. Twitter users also share many of the same habits ideal for personal data collection on YFD, such as frequent and near-realtime updates to a network. At launch, Twitter asked, "What are you doing?" with a text box to fill in an answer such as, "Enjoying a cup of coffee" but they later updated the question to "What's happening?" evolving into a place not just for personal status updates but also a

"new kind of information network" (Stone, 2009) where people share links and information.

Twitter's growing popularity was reflected by a wide array of applications that were built using the Twitter API. Many of these third-party applications let people informally collect data about themselves, but they were basic in that the end result was usually a list of items that did not try to help users understand patterns or behaviors. The focus was on what was current or recent. Tweet What You Eat (Ressi, 2010) and FoodFeed (Lourenco, 2010) let users keep track of what they eat. Overheard.it (Snook et al., 2010) advertises itself as "eavesdropping on Twitter" and keeps a live feed of tweets, usually of amusing quotes that people hear others say in passing. Kvetch (Powazek et al., 2010) is similar to Overheard.it but aggregates complaints as well as lets people vent annoyances anonymously. Graffiter (Li et al., 2009) is a more general tool that lets you log numeric data through Twitter and other applications, such as instant messenger and bookmarking site Delicious. Individual tweets can provide context to a story, but there is value in seeing more long-term trends and finding patterns in the collection of tweets and data points. YFD makes use of Twitter's popularity, ease of use, flexibility, and online culture and applies it to personal data collection, but unlike existing applications, emphasizes patterns over time just as much as single observations.

3.1 Syntax

To log data to YFD via Twitter, users follow a defined syntax designed to fit in with typical Twitter usage. This section describes the evolution of the syntax into what it is now and provides examples of how people can use it to log various types of data.

3.1.1 Basic Syntax

In the first version of YFD after switching from email to Twitter, users were only able to log five metrics: eating, drinking, sleep, bathroom habits, and mood. Whenever users ate something, they would send a direct message to @yfd that read "ate" followed by whatever they ate. For example, if users ate pepperoni pizza, they would send a message like the following:

```
ate pepperoni pizza
```

Users entered similar messages when they drank something, using "drank" followed by what was consumed:

```
drank Dr.  Pepper
```

The keyword for mood was "feeling." For example, a happy mood could be entered with the following message to @yfd:

```
feeling happy
```

The general pattern was a keyword (i.e. ate, drank, or feeling) followed by a word or phrase, which were like categories. Tracking for sleep and bathroom habits did not require values after the keyword. Instead, users sent lone keywords, because only the time of the message was relevant. For example, when a user woke up, the following message was sent:

```
gmorning
```

As one might expect, a similar message was sent went going to sleep:

```
gnight
```

Likewise, when users went to the bathroom, they could enter:

```
peed or pooped
```

YFD was online for two weeks and had about 100 test users. Immediately, users requested support for more keywords. Smoking for those trying to quit and entertainment for those who wanted to log the books they read and the movies they watched were added, using the keywords "smoked", "read", and "watched", respectively. Whereas previous keywords where followed by a category or nothing at all, smoking required that users be able to log more than one cigarette at a time. A user could send a "smoked" message twice for two cigarettes, but this can be a chore when typing messages on a mobile phone. Instead, users were able to specify counts by following keywords with an integer, such as in the following:

```
smoked 2
```

The syntax supported this small set of keywords for about a month, but as more people were invited to use YFD, there were more requests to support new data types. Some requests included support for logging drug intake, glucose levels, exercise, and Internet usage. In its original design, YFD was built to have trackers made specifically to fit a data type. Smoking had a dedicated view, as did eating, drinking, and mood. I originally thought that customized views for a specific data type would lend to more useful visualization and a better understanding of the data. However, as users requested more, it was clear a tool that allowed users to log the data they wanted could be more useful, instead of forcing them to wait for the introduction of new data types, one-by-one. So I generalized the syntax, as shown in Figure 3.1.

The current syntax is close to the original. There are still keywords, referred to as *actions*. Actions are optionally followed by *values* and *units*. A value is numeric and units can be a word or phrase. Like the original, narrower syntax, the

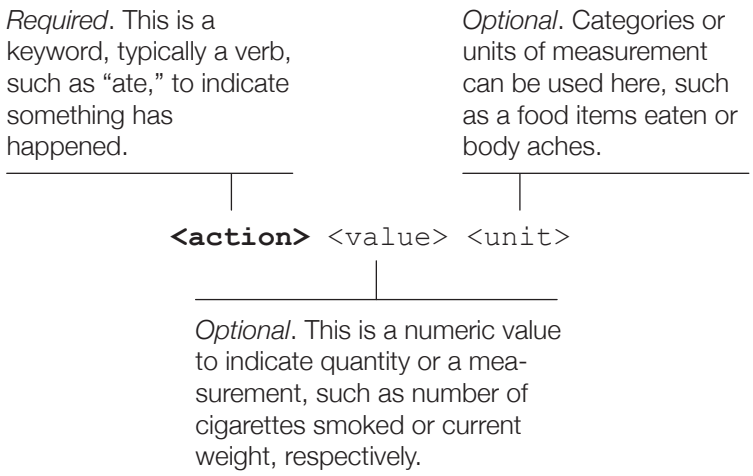


Figure 3.1: YFD General Collection Syntax

generalized syntax fits into regular Twitter activity and encourages data logging as something happens or is completed. For example, the message to log eating pepperoni pizza, can be the same message as before:

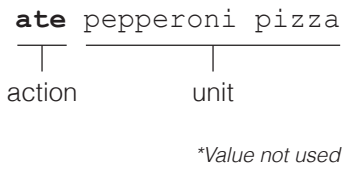


Figure 3.2: Example using general syntax, action and unit

In this case, the action is "ate" and the unit is "pepperoni pizza." No value is specified.

The inclusion of values in the syntax lets users track numeric values. Some users, for example, who keep track of how many cigarettes they smoke can send a message as shown in Figure 3.3, when they smoke two cigarettes.

Again, this is the same as before, but in terms of the generalized syntax, the action is "smoked" and the value is 2. Similarly, a user tracking drug intake could send a message like the following:

```
took 2
```

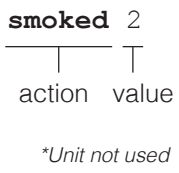


Figure 3.3: Example using general syntax, action and value

The action is "took" and the value is 2. No unit is included. If a user wanted to keep track of different types of medication used, one could include a unit, as shown in Figure 3.4.

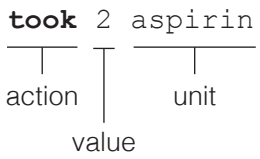


Figure 3.4: Example using general syntax, action, value, and unit

The user could change the unit if a different medication was used:

```
took 2 ibuprofen
```

The action is still "took" and the value is still 2, but in these two examples, the units are "aspirin" and "ibupofen." With this flexible syntax, the user has more control over what is tracked and how to enter it and does not have to wait for a developer to implement a feature. The parser is naive in that a message such as "took ibuprofen 2" would be interpreted incorrectly—because the numeric value does not immediately follow the action—but I chose to avoid ambiguity in the syntax.

3.1.2 Data Types

The generalized syntax lets users log four types of data: count, categorical, measurement, and event, summarized in Table 3.1. Types are user-specified via the

YFD site, discussed in Chapter 4, and the type dictates specific message syntax.

Data Type	**Syntax**	**Examples**
Count	`<action> <value>`	smoked 2, read 50
Categorical	`<action> <value> <unit>`	ate breakfast, took 2 aspirin
Measurement	`<action> <value>`	weigh 170, ran 10
Event	`<action>`	gnight, exercised

Table 3.1: Data types using general YFD syntax

The *count data type* is for actions where cumulative counts are important to the user. Smoking is one example. As before, a user might be interested in tracking the number of cigarettes smoked per day. In this case, only the value matters, and the unit is left out, because all values represent cigarettes. There are no categories. Another example is a user tracking the number of book pages read. One might send a message to @yfd like the following, after reading 50 pages:

```
read 50
```

Again, the units do not matter, because the only thing the user is tracking is number of pages read, which can be specified in a description field via the site.

The *categorical data type* lets users keep track of different units. Returning to the drug intake example, the user can keep track of the kind and amount of medication consumed by including both a value and a unit after the action "took." In some cases, the value is not needed and only the unit is used with an action. For example, if a user does not care about the number of pills taken, but rather only that medication was taken, the user could send a message like the following:

```
took aspirin
```

There is no value in this message, and the unit ("aspirin") is incremented by one.

Measurement data types are for instances when the user does not care about the aggregate over time, or a sum does not make sense. When users weigh themselves, for example, there is no need to sum the weights over a week. Instead the trends over time are the point of interest. In this case the user sends a message with the same syntax as the previous reading example, which is an action followed by a value with no unit:

```
weigh 140
```

This is an example of how the same syntax can be used to record different types of data, which is why it is important for users to specify data types on the site. The units of "weigh" (the action) are implied or again, can be explained in a description field available on the site. The user could include "pounds" as units, but that would inefficient when typing on a mobile phone.

Finally, actions are designated the *event data type* when only time of day or date are the focus. Oftentimes, users only care when something happened or the last time an event happened. For example, with sleep, only the times or the timespan in between when the user goes to sleep and wakes up matter. A user might keep track of drug intake only to remember if a prescription is taken regularly. The syntax would be the same as in the previous "took" example, but a change in data type via the site shifts focus.

3.1.3 Timestamps

YFD syntax was originally intended for in-the-moment data entry. That is, as an action was completed, users would send a message to @yfd, and the time an action occurred was assumed to be the timestamp attached to the direct message via the Twitter API. However, users requested the ability to manually timestamp their data. Sometimes users forgot to log data as something happened, or it was not convenient to log data at the time. Many users did not use Twitter via their

mobile phones, so they had to wait for access to a computer before they could log data. PEIR (Mun et al., 2009) posed a similar challenge, where users required a way to log data asynchronously. Because of the demand, the basic YFD syntax was extended so that users could include what time an action occurred.

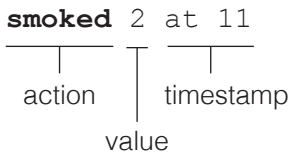


Figure 3.5: Timestamp syntax

It was important to keep YFD use consistent with use on Twitter, so the syntax was extended in a way that felt natural. Users can add timestamps manually by appending "at" and the time to their message. For example, if a user smoked two cigarettes at 11 o'clock in the morning and logged the data point in the afternoon, the user could send a message similar to Figure 3.5. The "at" indicates that the text that follows is the time that should be used instead of the time of the message. The above example is for 11 in the morning. If it were a timestamp for 4 o'clock in the afternoon, the user would also include "pm" at the end of his message:

```
smoked 2 at 4pm
```

Without an "am" or "pm" indicator, YFD assumes the time is for pre-afternoon. The only exception is when the military time format is used. If the hour is after 12 sans "am" or "pm," then military time format is assumed. For example:

```
smoked 2 at 16
```

The same syntax still applies if a user also wants to include minutes:

```
smoked 2 at 4:30pm
```

When a user includes a timestamp, YFD assumes it is the most recent occurrence of that time. For example, if someone sent the above message at 5:30pm on a Tuesday, YFD would assume that the user smoked at 4:30pm on Tuesday. If, however, the user sent the same message at 3:30pm on Tuesday, YFD would assume the user smoked at 4:30pm on Monday, because it is not yet 4:30pm or later on Tuesday.

Currently, YFD only allows users to include time of day and not an actual date. This is to encourage users to enter data as it happens and to keep in line with the in-the-moment usage of Twitter. If date entry were allowed, it should also feel natural like time does. One could allow users to enter a date in "YYYY-MM-DD" format, but that is not how people use Twitter. In future iterations, date entry could be enabled using natural language, such as "five days ago" or "last month."

3.1.4 Hashtags

Tagging has grown in popularity over the past few years (Viegas et al., 2009). It lets people append additional data or place data points into a category in an informal way. Popular bookmarking application, Delicious (Yahoo, 2010), was one of the first to popularize the concept. With Delicious, there is a tag field to enter keywords for any given bookmark, and when users go back to their bookmarks, they can filter by tags.

Twitter did not always provide tagging functionality; however, the Twitter community made up their own way to create tags, or add additional information to their tweets. Twitter then formally integrated the functionality into their framework (Support, 2010). Hashtags on Twitter work in much the same way as tags on Delicious or other applications. The difference is that hashtags always start with a hash or pound symbol (#), and they can be included anywhere in a tweet.

Like the Delicious user who filters bookmarks, a Twitter user can filter tweets by hashtag, either by selecting a hashtag on the Twitter website or via Twitter Search. For example, a Twitter user might tweet that she is "Drinking coffee at my favorite cafe #morning." Later on, that user could easily find tweets that were tagged with "#morning" and the tweet would also be included in Twitter search results for "#morning" if the user's timeline is public.

Similarly, because hashtags are common on Twitter, the functionality was also included in YFD syntax. Users can include hashtags in their messages to @yfd to categorize their data. Hashtags work with YFD the same way that they work on Twitter. For example, a user could keep track of eating for meals, in addition to food items, as shown in Figure 3.6.

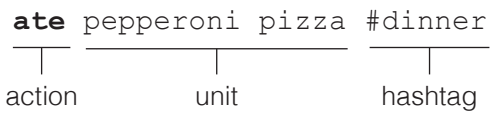


Figure 3.6: Hashtag syntax

The user can also use more than one hashtag to categorize data even further. If the user wanted to track eating habits for different meals of the day and when a meal was at a restaurant a message like the following could be sent:

```
ate pepperoni pizza #dinner #restaurant
```

Figure 3.7 shows how a message might look like with hashtags and a manual timestamp.

It does not matter in what order the hashtags are placed in the message, as long as they are not placed first. The first spot is reserved for the action. So the following would work the same as the message above. Notice the hashtags appear at the very end, and the "at 6:30pm" appears in the middle.

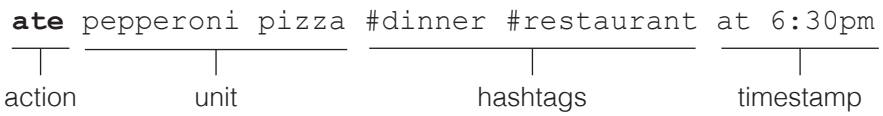


Figure 3.7: Example using hashtags and timestamp

```
ate pepperoni pizza at 6:30pm #dinner #restaurant
```

3.2 Storage

Direct messages to @yfd on Twitter are parsed with Python, and then stored in a MySQL database, which is backed up nightly. Storage in a database allows for quick and straightforward subsetting and privacy control in the user interface. Figure 3.8 shows how a message sent via Twitter is parsed and stored.

YFD was developed using Django (Holovaty, 2010), a Python web framework, and follows a model-view-control design paradigm. The model defines how data is stored, the view is what users see in their browsers, and control handles processing and exchange of data. The main model of YFD is the action. It has a name or keyword and optional unit, value or tags. It also has a timestamp. The database schema reflects this model. We can then conveniently do operations on any of the fields, provided the tables in the database do not get too large. Unlike previous project SensorBase (Chang et al., 2006), this has not been an issue with the application, and most likely will not happen for a while since most users collect data manually. Django also abstracts the actual database, which makes saving, retrieving, and deleting trivial once the models are set up.

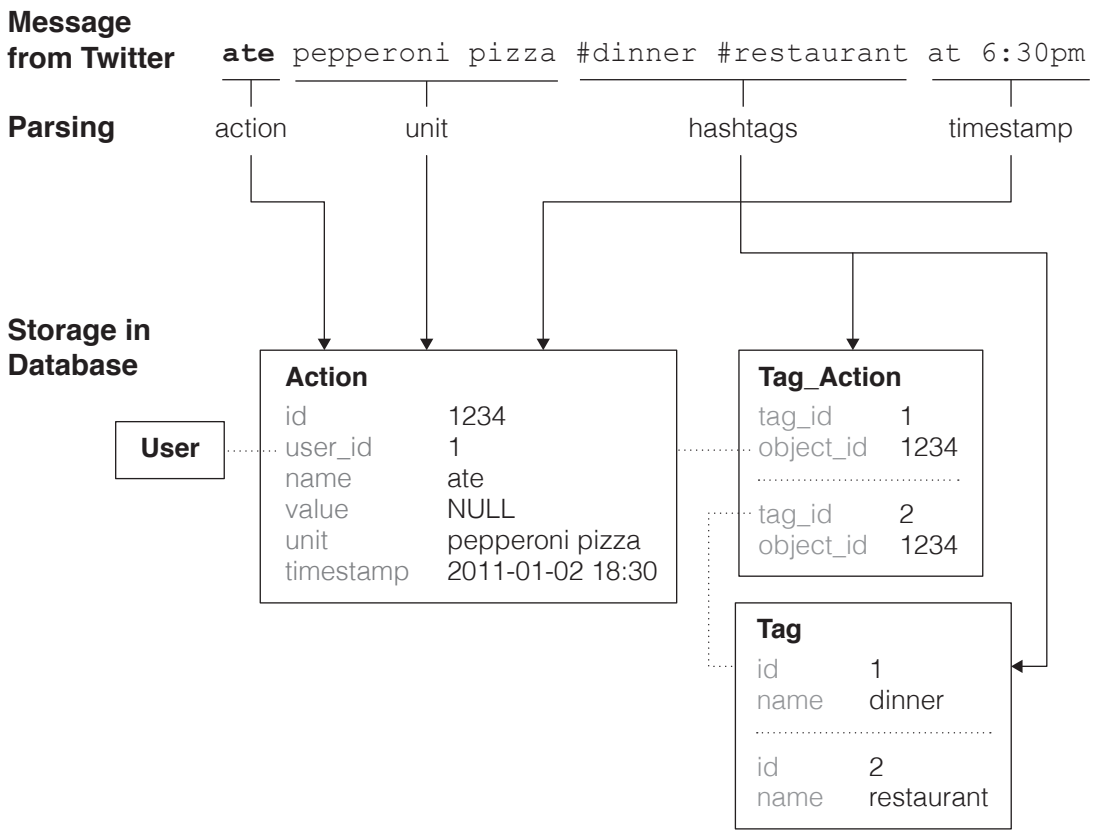


Figure 3.8: Parsing a message and storage

3.3 Reminders

One of the challenges of personal data collection is users forget to log data (Rodgers et al., 2005), so YFD provides a way to send reminders. Previous studies with SMS reminders support the addition of the feature. Leong et al. (2006) and Downer et al. (2005) discuss improvements in patient attendance for health care using reminders with SMS. In each study, attendance was at least fifty percent better for those who received mobile reminders. Gaglani et al. (2001) showed that a computerized reminder strategy was effective in increasing influenza immunization rate from five percent to over thirty.

Additionally, reminders was a common feature request among early YFD users. Some reported that they wanted to collect data about themselves, but could not

remember because it was not part of their daily routine. On the other hand, several users who had logged data over a few months reported that collection became habit, so reminders seemed like a good way to get new users accustomed to logging actions.

There are two types of reminders available on YFD: basic and custom. The *basic reminder* is based on an absence of data. If a user has not logged data in a specified number of days, YFD sends a direct message that tells the users data has not been logged data in the selected time span. For example, the following message would be sent if a user had not logged data for more than four days:

> "Psst. Just a friendly reminder to update your.flowingdata. It's been more than four days since your last update."

Because the direct message is via Twitter, it can then be forwarded to the user's email, desktop Twitter client, mobile client, or SMS message, depending on the user's settings on Twitter. These multiple pathways increase the chances that the user will see the reminder and respond in the same way they help users collect data in the moment. Users can turn this basic reminder on and off whenever they like.

YFD also provides a way for users to define their own action- and time-specific *custom reminders*. A user can select an action from a drop-down menu and select the start date and time and whether it repeats. Reminders can repeat daily, only on weekdays, every other day, weekly, monthly, or yearly. For example, a user might want to track weight, so he or she could set a weekly reminder to record that number every Friday at 7 in the morning. Each Friday the user would receive a reminder via direct message on Twitter like the following:

> "A reminder to enter data for (weight)"

Such a reminder could help with more regular updates and more reliable data since something like weight can fluctuate during different parts of the day. Mood is another example. Users who track their emotions might focus more on the negative feelings, and might only log during these times instead of during both happy and non-happy times. A reminder can help users focus on everyday emotions with a message to log mood at a fixed time during the day. In my own use, I found this to be helpful in placing more focus on positive feelings.

Like the basic reminder, custom reminders can also be edited and deleted. Users can also create as many reminders as they like. This gives users full control over what YFD sends to them over Twitter. In future iterations, one might imagine *smart reminders* that are not user-defined. Instead they might be based on patterns in previous usage.

3.4 Queries

As of this writing, queries are still in-development. Users can query or ask questions about their data by sending direct messages to @yfd on Twitter. The goal is to allow users to interact with their data in a simple way on their mobile phone, in addition to exploration on the site.

Users can either ask for a summary of their recent data or ask for the most recent entries. For the former, users send the following message to @yfd:

```
summary?
```

A message with the keyword "summary" ending with a question mark will return the number of data points logged that day and some summary statistics. The message sent by YFD to a user would look like the following:

> "8 points logged today, 10 percent more than yesterday. Most recently: ate pepperoni pizza."

Users can also ask for a summary about specific actions, and the response varies by data type. If a count data type was logged, then the sum for the day would be returned via a direct message from @yfd. The mean, minimum, and maximum is returned if an action is a measurement data type. Categorical data types will return counts for each category. For example, if a user wanted a summary about eating, the user might send a message like the following, assuming "ate" was the action word:

```
summary ate?
```

This is the same as the previous message except "summary" is followed with the action "ate." The message still ends with a question mark to indicate it is a query and not a data point to log. The response would look like the following:

> "You ate pepperoni pizza 8 times, chicken wings 4 times, and breadsticks 2 times. Most recent: pepperoni pizza."

Similarly, users can also ask for the most recent data entries with a similar message format:

```
recent?
```

The most recent data point logged and how long ago it was is sent to the user:

> "Most recent: sore finger again (1 min ago)"

Users can also focus on a specific action, as they can with summary:

```
recent ate?
```

A response similar to the following would be sent:

"Most recent: ate pepperoni pizza (2 hours ago)"

This question might be useful for someone who keeps track of the medications taken. A "recent" query ask for the most recent medication taken and at what time it was taken. Based on this information a user could decide whether or not it is time to take another dose.

Although this feedback mechanism is basic, it is easy to see how the concept could be expanded with more complex queries, such as subsetting over time or discovering patterns. Such computation could also be tied in with reminders to offer users advice based on their data logging practices. At the same time though, the messages should still be kept simple as they are only text and appear on a mobile phone with a smaller screen than a desktop computer monitor. So rather than complex and analytical insight, such messages could motivate users to explore more deeply via the exploration and visualization interfaces as well as invoke awareness and reflection.

3.5 Summary

YFD syntax started as a limited set of keywords, but was generalized based on what users wanted to track and how they were tracking it. Although simple, the syntax provides flexibility so that users can log the data they want in a way that is comfortable. They can log different types of data, add meta data through hash-tags, and modify timestamps in case they forget to log data or it is inconvenient in the moment.

Communication can also move in the opposite direction, where YFD sends reminders to users based on user-defined rules. This can help with data regularity and accuracy. In addition, users can form simple queries and receive results via direct message on Twitter, so that interaction with data is not just on the site, but also on the phone, albeit in a simple form.

Ultimately, the main motivation behind the short message format was to make data collection straightforward in a way that fit into a user's everyday. If there are barriers, users are less inclined to collect data, and the site is irrelevant. However, the more data that users logged, the better the gauge for how users explore their data with visualization tools and basic views.

CHAPTER 4

Exploration and Visualization

4.1 Introduction

YFD was made to support various applications of personal data collection, from data journaling to self-experimentation. As described in Chapter 3, the data collected by individuals is similar across a wide range of uses, but data presentation can change how the data is interpreted and what it can be used for. This chapter describes the visualization tools YFD offers and common visual elements in the application that help users interpret their data across multiple views.

Basic views that show individual points let users casually browse their data as they would a Twitter feed, whereas more exploratory and interactive tools, such as a calendar heat map and stacked area chart, show users' data over time, at different granularities, such as daily or weekly. Data can also be viewed in aggregate with tools such as a word cloud, which does not provide the most visually accurate view, but does provide users with familiarity. YFD's range of tools and interaction let users choose how deeply they want to explore their data, whether it is a quick visit to the site to see most recent data or to examine their data more closely with visualization methods designed to show more data at once.

4.2 Browsing

YFD is like an extension of Twitter. People regularly update their status on the microblogging site, and YFD provides a way to log updates privately and see those updates in aggregate over time. So the initial views of YFD were designed to provide familiarity to those who already used Twitter. Usage, discussed in Chapter 5, also showed that people spent most of their time browsing and less time with long sessions in more exploratory views.

4.2.1 User Homepage

When users log in to YFD, they see a homepage, as shown in Figure 4.1. The newest logged data point is shown in a large font, followed by a list of the most recent on the bottom in a smaller font. Small charts below the list show relative changes: percent change in number of points logged versus thirty days before, number of data points logged over time, and a histogram that shows distribution of points logged during the hours of the day. In the sidebar is more thirty-day summary on how much a user logged and what actions were logged.

The homepage serves two main purposes. The first is to quickly show users that data sent to @yfd via Twitter was saved to the database. This is the integration stage in the model proposed by Li et al. (2010) and is done automatically without user involvement, so it was important to show that the logging mechanism worked. It was common for users to log data and then briefly visit YFD.

Secondly, the nature of the web is to show what is most recent and to list items—whether they be status updates, tweets, photographs, or the news—in reverse chronological order. So it made sense to show the most recently logged data and a quick view of the past thirty days. Many of the projects referenced in Chapter 2, such as Daytum, Mint, and Many Eyes, also follow this order. We want to guide people through their data but also want to pay attention to user

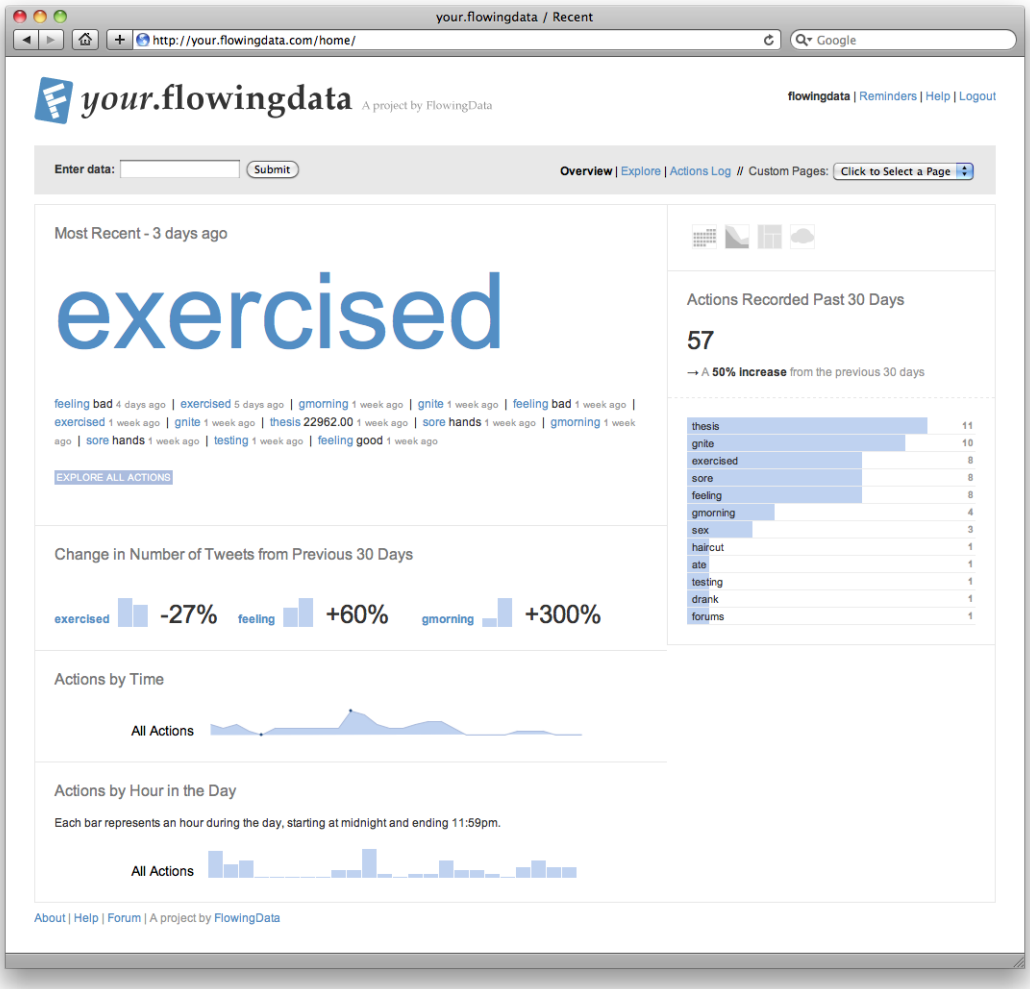


Figure 4.1: User homepage

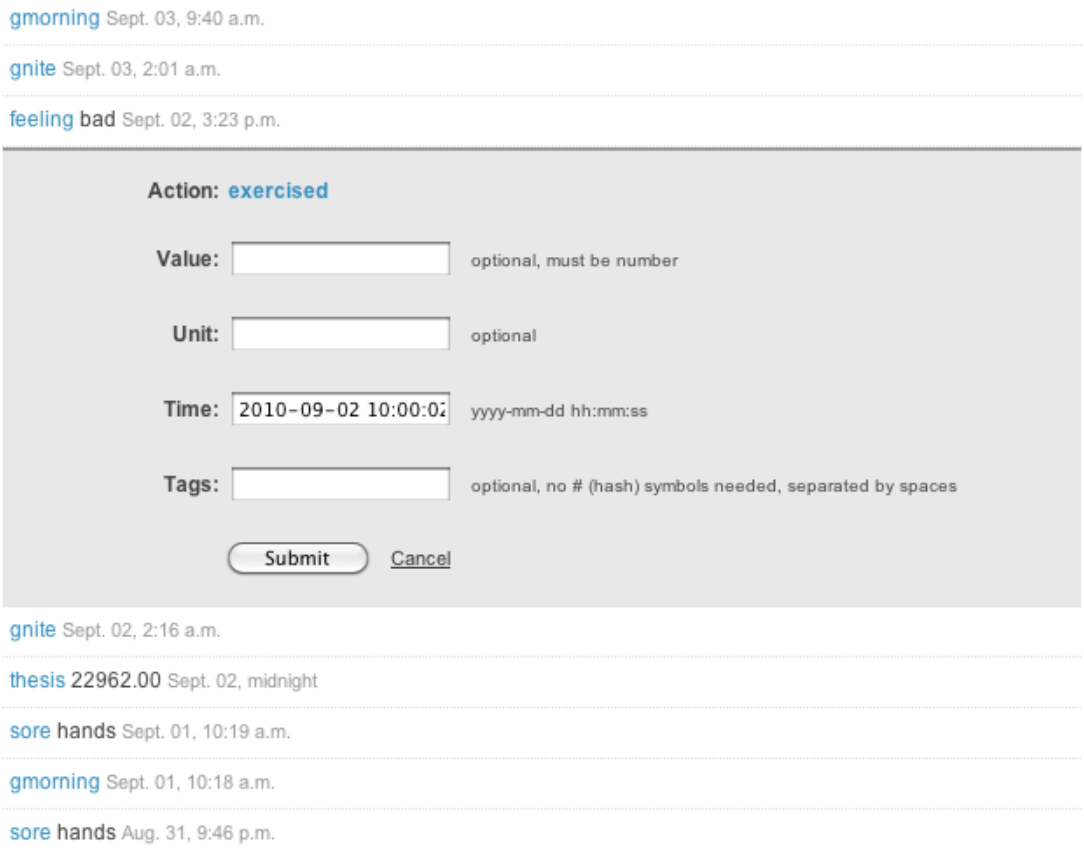


Figure 4.2: Actions log

expectations.

4.2.2 Actions Log

As shown in Figure 4.2, the actions log is simply a list of a user's data in reverse chronological order. Users can also download a tab-delimited version. Although it is a basic and straightforward view, the actions log was visited more than any of the visualization tools.

Users can see the data they logged, but more importantly, they can edit and delete it. One of the drawbacks of data collection with Twitter or SMS is that typographic errors often occur on one's mobile phone, because the physical or touch keyboards are smaller than those used with a personal computer. So YFD

provides a simple way to edit and delete actions via the actions log. As the user mouses over rows of actions, edit and delete links appear. The user can either delete the data point or edit the unit, value, or timestamp.

4.3 Single Action Views

Beyond the homepage and actions log, YFD provides visualization tools to look at multiple actions at once and to see trends for a single action. This section describes the latter. As discussed in Chapter 3, users can classify their data as categorical, event, count, or a measurement, along with a description of what the action is. The first three data types are for aggregates over time, and the measurement data type is used for non-aggregates, such as weight or the time it takes to run five kilometers. Before users see their data through the single action views they have to specify the data type, which forces them to think about how they should collect their data. The classification also allows for customization based on the type of data.

Figure 4.3 shows the view for a measurement data type. This example uses weight as the action. A dot plot is used to chart the data, with time on the horizontal axis and measurements on the vertical axis. Dots are not connected, because it is rare that users remember to or want to log data every day and in equal intervals. They might log data every few days or every few months. There might be several days in a row when data is logged and then nothing for the next chunk of time. So instead, an AB line is shown to to give a rough idea of trend. The line is dark blue while the dots are light green so that users focus more on patterns over time than they do on individual points.

Keeping with the focus of the homepage, the most recent value is shown on the bottom, but the average, maximum, and minimum are also shown. In some cases, a histogram to show distribution could be helpful, but because data logged

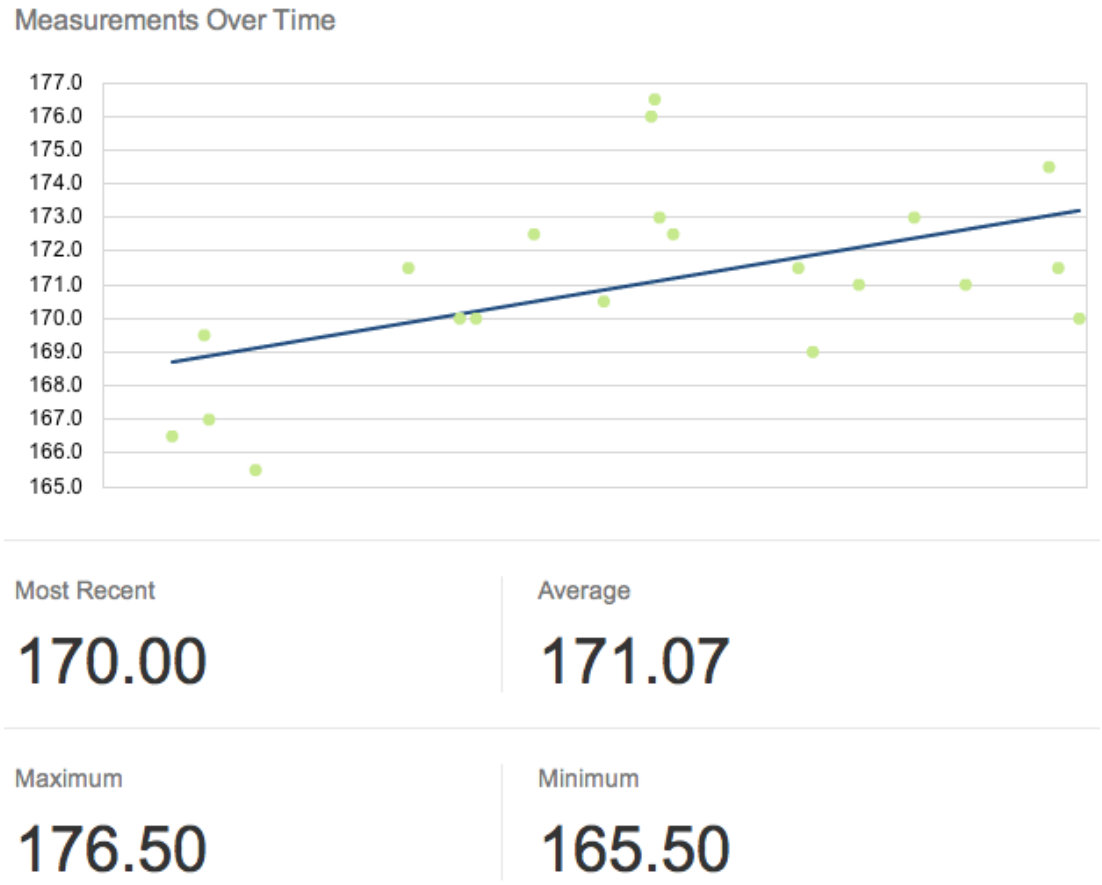


Figure 4.3: View for measurement data type

on YFD tends to revolve around time and from experience, histograms are often misread by non-professionals, no distribution charts are included in this view.

The views for categorical, event, and count data types look similar to Figure 4.4, which is for the action "drank" classified as categorical. The page starts with a one-sentence statement about the data, such as "Most commonly drank tea." The goal was to tell users something basic about their data to encourage them to look closer and perhaps think about what aggregates mean in the context of their data. For the categorical data type, the statement tells users the most common category; for the event data type, the statement tells users the most common time an event occurred; and for the count data type, the statement tells users how many times an action occurred during the most frequent hour.

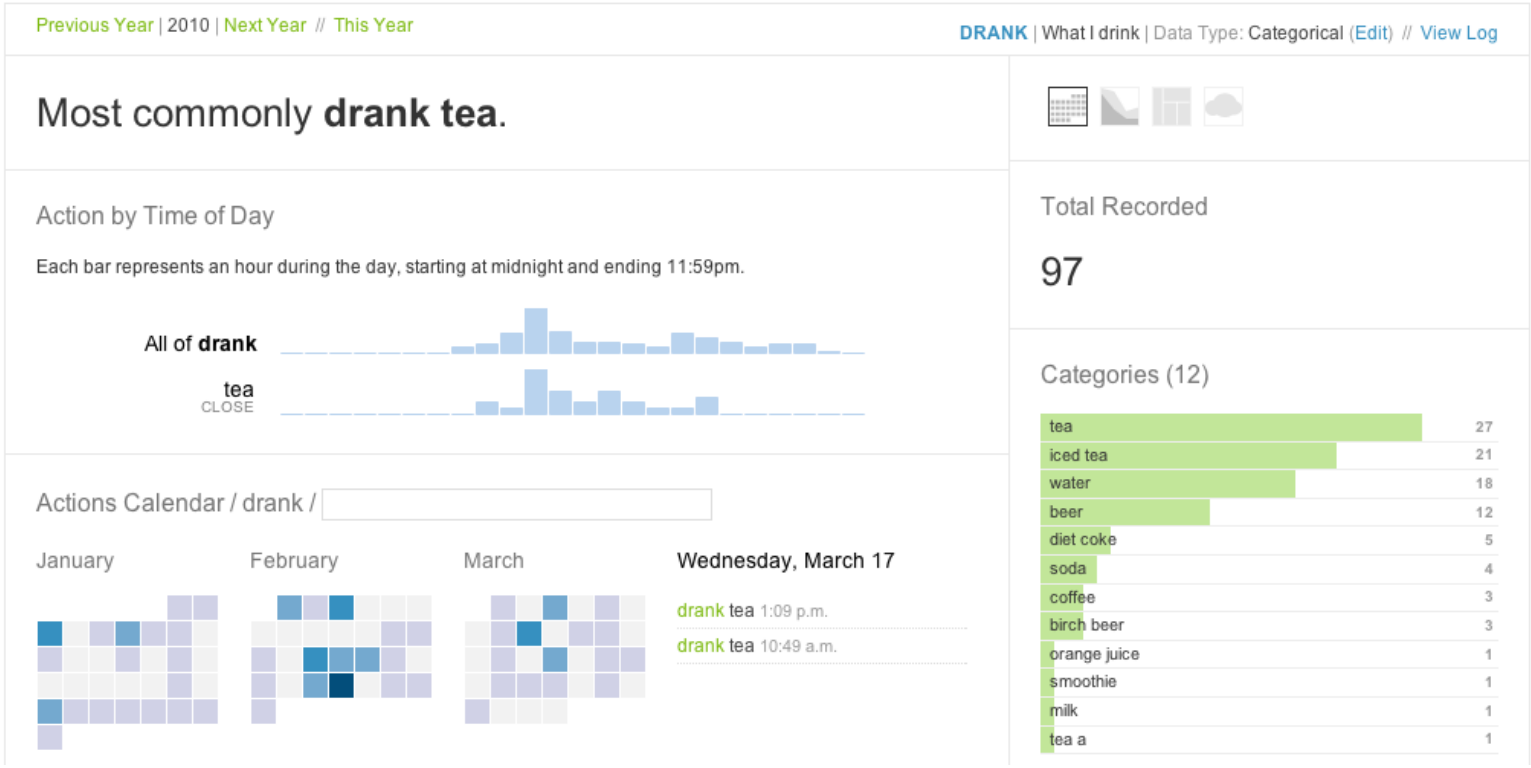


Figure 4.4: View for categorical data type

Notice that unlike the measurement data type, a histogram is used to show hourly distributions for the action. In this case, the chart seemed more intuitive, because time is on the horizontal axis, however it is hard to say if users read it correctly as no one commented on the chart type in the usage survey.

Like the homepage, a bar graph is shown in the sidebar, except units for the current action are shown instead of all actions. For example, a homepage might list ate, drank, and sleep, but in the single action view in this example shows beverages that were drank, such as tea, water, and beer.

Finally, a calendar heat map, described in more detail in the next section, is shown which colors days by number of data points logged on any date. For the categorical and count data types values are summed to determine the color scale, whereas with the event data type, each logged point is counted as a single occurrence. More on this view, which can also be used to look at all actions at once, is described in more detail in the next section.

4.4 Analysis

Although users spend most days on YFD quickly logging data and in the casual views, they also explore their data in less frequent, but longer sessions on the site. Again, the goal was to build tools that are flexible enough to display various data types and to provide consistency in interaction and look and feel, so that users do not have to learn how to use each tool. Time is inherent in the data, and the focus of most of the tools is on change and relationships over time; however, aggregate views are also available for those interested in count and volume aspects of their data. Most users have a mix of interests and data types, so YFD lets them quickly switch between views and focus.

4.4.1 Temporal Views

Once users see their most recent data, they can look back to see how the recent compares to the past. They can look at data that is years old and see how their habits have changed, and usage reflects this greater interest in change over aggregates during a given time period. Although time is present throughout the YFD interface, there are three main views that let users explore time: the calendar heat map, stacked area chart, and a custom durations tool.

4.4.1.1 Calendar Heat Map

Users can also view their data in a calendar interface. Calendars are typically used to display events such as meetings or birthdays, but the same layout can be used to display a YFD user's data. The calendars in YFD provide familiarity in both format and interaction.

Online calendars such as Google Calendar (Google, 2010), that are designed around agendas and schedules, typically do not zoom out any further than the month view; however, YFD shows a year at a time so that users can easily find the days of high, medium, and low data volumes. Cells, which each represent a day of the year, are organized in a standard calendar grid format with groupings for each month. Each month has seven columns, each representing a day of the week. The left most column represents Sundays while the right most column represents Saturdays, as shown in Figure 4.5.

Cell color depends on data type, but generally speaking, calendars follow a blue color scheme and vary in saturation. Darker shades of blue indicate higher volumes of data logging or higher counts of a given action. Conversely, lighter shades of blue indicate lower volumes or lower counts of a specified action. Days when no actions were recorded are colored a light gray. With this layout, the base calendar serves as a background, and days when something were recorded are

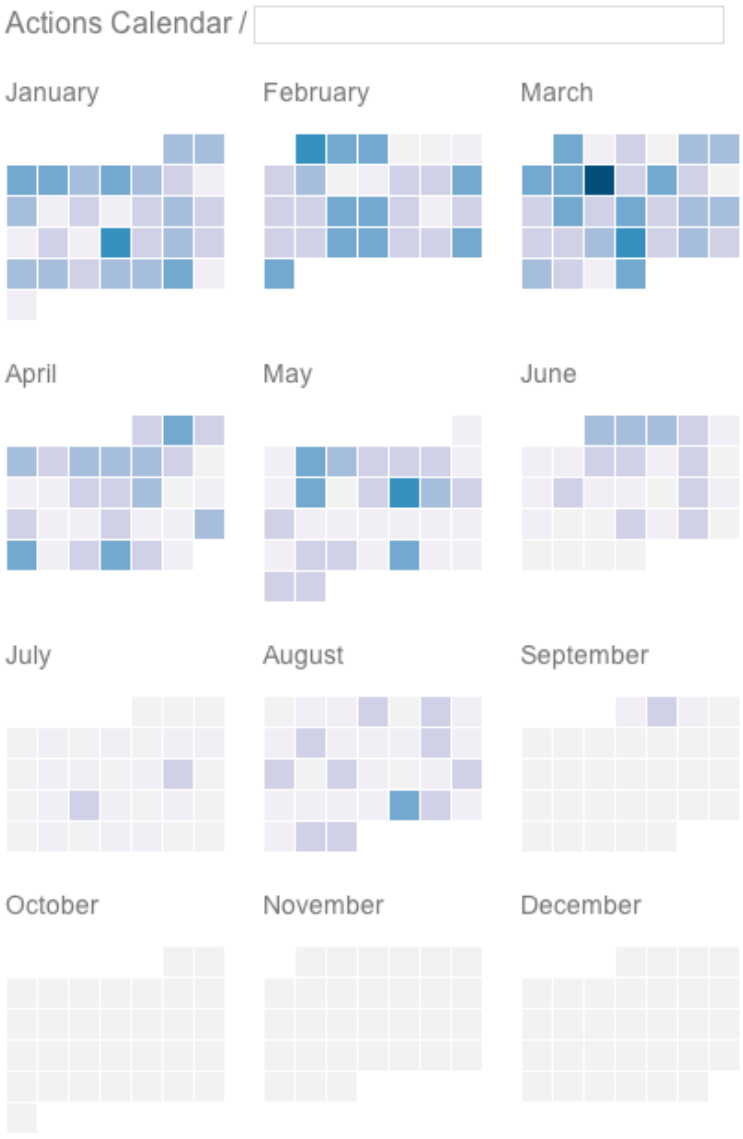


Figure 4.5: Calendar heat map

highlighted. As a whole, calendars are like a heat map (Friendly, 1994) organized by time.

When a user clicks a day on the calendar, the actions for that day appear as a list, along with the time they occurred. A time series sparkline (Tufte and Graves-Morris, 1983) is also displayed for an overview of the time of day the actions were logged. From here, the users can click on an action in the list to explore an action individually, or they can select a different day. YFD also provides a search tool to find actions of interest. As the user types a query in the search field, the calendar updates to show actions that match the query so far. For example, if the user types "a" the calendar reflects all actions that start with "a." The user can continue typing "ate," and the calendar will show color intensities for the user's food log.

The same functionality applies when the calendar view is filtered for specific actions; however the coloring scheme varies by the data type of a given action. Categorical data types are colored by number of data points logged on each day. Count data types, on the other hand, are colored by the sum of counts for the day. For example, a user might track number of cigarettes smoked with a message like this:

```
smoked 2
```

As a count data type, the days in the calendar are colored by the number of cigarettes smoked rather than the number of smoking sessions in a day.

If the data type is categorical, where the user makes use of units in the YFD syntax, the user can search by these, as opposed to the action names in the overall calendar. For example, a user might keep track of eating:

```
ate pepperoni pizza
```

The action keyword is"ate" and "pepperoni pizza" is the unit. As a categorical data type, eating could be categorized by its units, and the user could search the "ate" calendar for days "pepperoni pizza" was eaten.

4.4.1.2 Stacked Area Chart

YFD provides an interactive and searchable stacked area chart (Friendly, 1995) that, like the calendar heat map, can be used to visualize all actions or units of an action. The interaction and design is similar to the Baby Name Wizard by Wattenberg (2005). The Baby Name Wizard provides a view of baby names in the United States over time, which allows people to see trends on their own name or the names of friend and family. It is also meant to help parents choose a name for their own soon-to-be-born children. Leskovec et al. (2009) used a similar mechanism to track phrases during the 2008 United States presidential campaign, and Byron and Wattenberg (2008) used a variation to show movie box office and music listening trends.

The stacked area chart on YFD, as shown in Figure 4.6, lets users visualize and explore their interactions in the same way as users can browse names with the Baby Name Wizard. On initial load a stacked area chart is generated to show all of a user's activity during the selected time frame. If no time frame is previously selected, it shows the activity for the past year.

Height is decided by number of times an action was logged during any given time slice, so peaks appear when the user logged a lot of data relative to the amount typically collected, and valleys show when a user is inactive. Stacks are colored by volume of data for the respective action. The darker the shade of blue a stack is, the more times that action has been logged. The lighter the shade of blue, the less times an action has been logged. Finally, areas of high volume are labeled by the actions, and the labels are also sized by how much the action was

logged in total. This provides a sense of flow without having to examine closely. For those more interested in relative counts than absolute ones, YFD also provides a normalized view, such as in Figure 4.7.

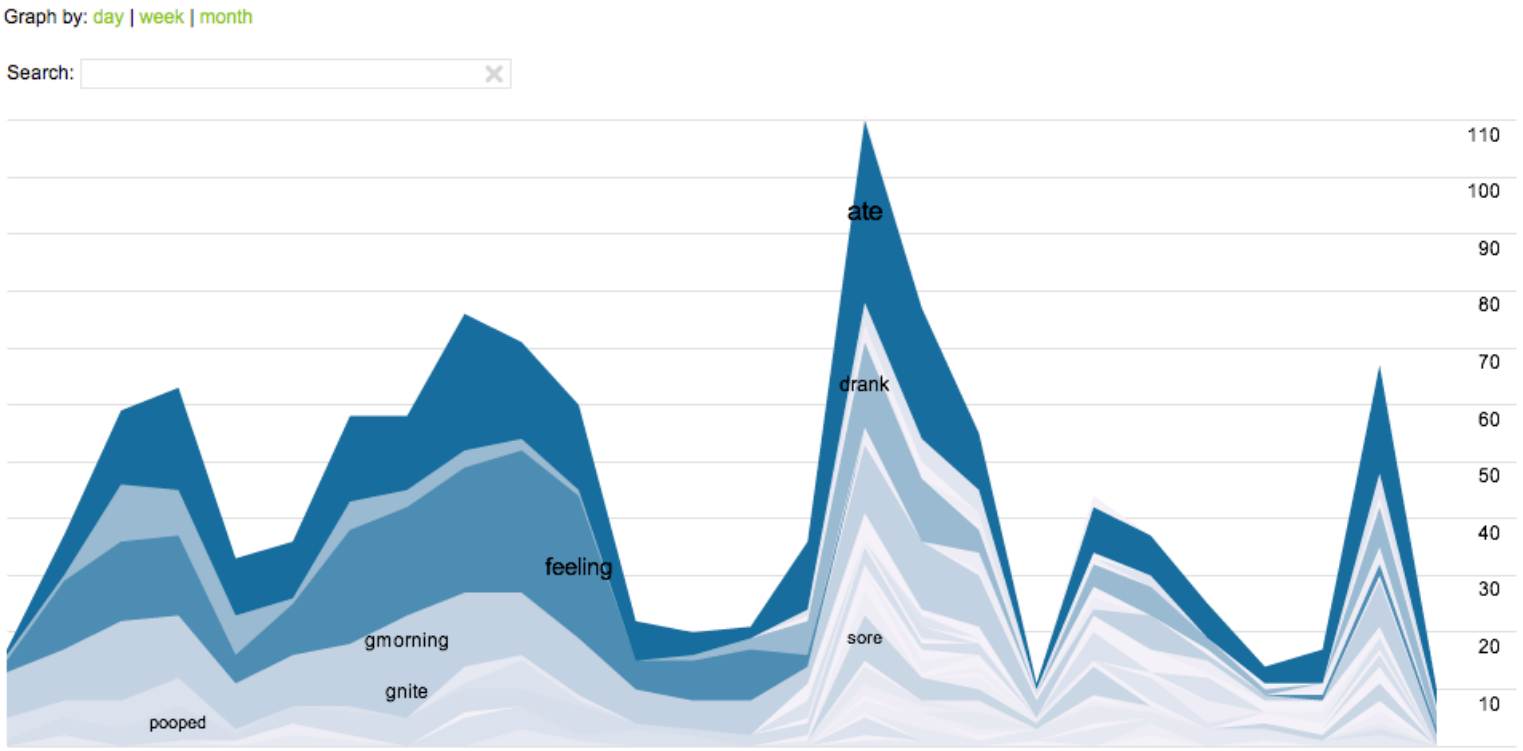


Figure 4.6: Stacked area chart

Time slices, or bins, can be changed to per day, week, or month. When evaluating data over a year, it is often more useful to see volumes on a monthly basis because if we try to fit too much data into a single view, it appears noisy and grows more difficult to read. However, if the user selects a small time frame, such as a month or two, it is perhaps better to see the data on a weekly or daily basis. The action keyword or unit name, along with the corresponding proportion or count appear as the user mouses over points on the graph.

In the overview stacked area charts, layers represent specific actions. When users double click on an action, they are taken to the action's page. Users can also visualize that specific action with the same view. For example, a user can keep track of mood with a message similar to this:

```
feeling happy
```

Whenever there is a mood change, the user can log a data point. This is unrealistic though since it is practically impossible to log every single mood one feels during the day, over a long period of time. The data would also most likely show a bias towards a particular mood such as happy or sad. Instead, it is more practical to set a reminder to log mood at a particular time of day. From experience, it was found that simplifying mood to either good or bad seemed to make mood easier to track, as shown in Figure 4.7.

Figure 4.7: Stacked area chart, normalized

Other examples include users trying to increase water intake or eat out less often. These users could keep track of liquids consumed or times eating out and at home. The options to see relative and absolute, zoom in on time frames, and change bucket size make the stacked area chart tool flexible enough for various data types.

4.4.1.3 Durations Tool

The YFD calendar and stacked area chart lets people keep track of when things happen, but in some cases the duration between events is of greater interest. For example, users interested in tracking sleep, might be more interested in how many hours they typically sleep than they are the endpoints. The durations tool is a custom YFD tool that shows durations between two actions, such as "gmorning" and "gnight." Drawing from the work of MacNeill (2010), which helps new parents keep track of when their child sleeps, the durations tool is a generalized tool that shows the time in between actions, as shown in Figure 4.8.

Figure 4.8: Durations tool

YFD users select actions from two drop-down menus, where one is for the starting action and the second is the stopping action. In the case of sleeping, "gnight" would indicate starting and "gmorning" would indicate stopping. Conversely, the user could switch the order to explore duration for time spent awake. Once a user selects start and stop actions, a visualization is generated for the current timeframe. Each row represents a day when the actions were logged. Times in between actions, or durations, with a cutoff point of 48 hours, are shown blue. Scroll over a blue area and a tooltip appears that indicates date and duration. A line indicator follows the mouse pointer and shows time of day. For example, if one were to place the mouse on the far left, it would show midnight. Move the mouse to the far right, and the indicator shows 11:59 at night. Average, maximum, and minimum duration are also shown in the sidebar on the left as summary statistics.

Through experience and informal user feedback, it seemed to be fairly common for users to forget to log a start or stop point, such as logging wake time, but forgetting to log sleep time. There is no way to find duration when a pair is incomplete. In these cases, the tool shows red tick marks for stop times and green for start. Although these marks do not indicate durations, they lend to the context of the paired data.

An alternative to pairing two actions would be to classify an action as a count or measure data type and only record durations instead of start and stop times. For example, a user could log the hours of sleep each night instead of both the time going to bed and waking up. A user's messages to log this might look like the following, to log eight hours of sleep:

```
slept 8
```

However, from experience using Twitter, it seemed more natural to post good morning and goodnight to others, which makes the transition to personal data collection easier. Additionally, separated events makes it easier to log, because

users do not have to remember start time. Instead, they do not have to remember time at all, and they can simply log a point as it happens, or soon after.

4.4.1.4 Cross-correlation Tool

Whereas the durations tool shows the time in between two actions, the cross-correlation tool, as shown in Figure 4.9, shows correlation between actions and units, across time offsets.

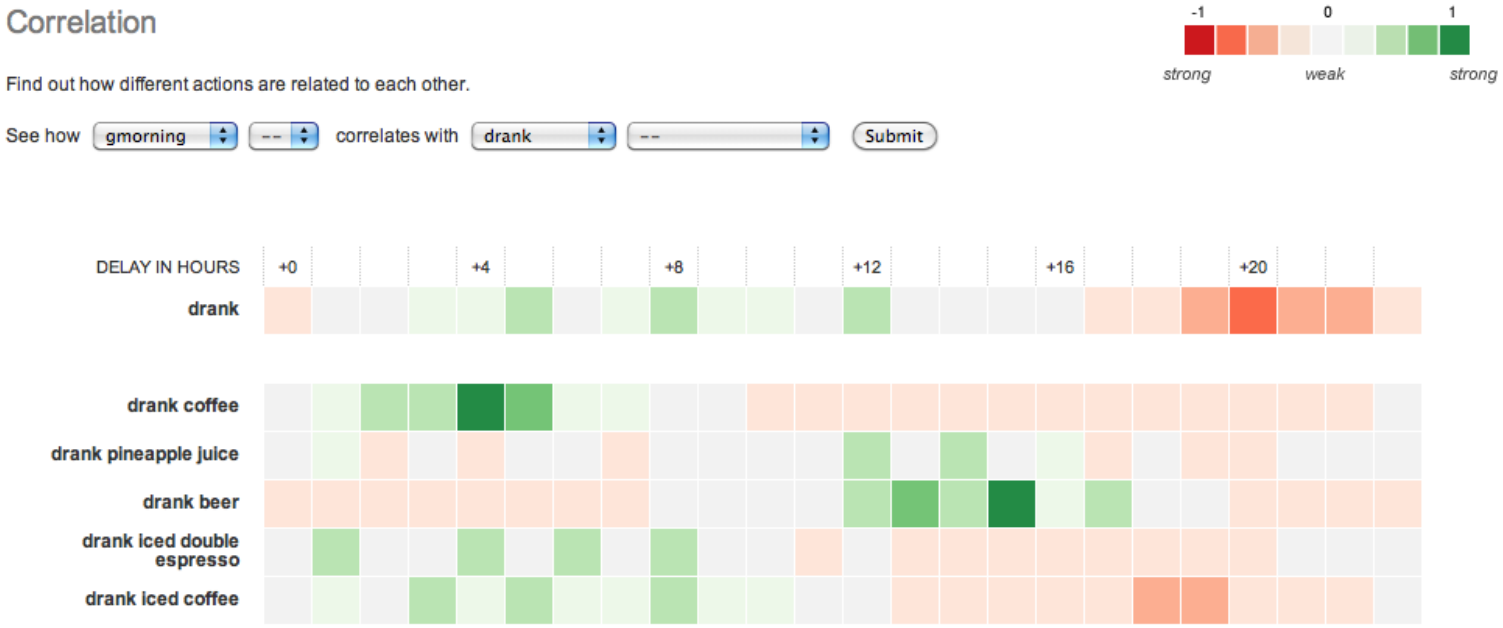


Figure 4.9: Cross-correlation tool

Users select an action from a drop-down menu and optionally a corresponding unit. The user can then see what actions correlate to the selected, in a grid visualization. Each row represents an action. They are sorted by level of cross-correlation. Columns represent hour offsets ranging from zero to twenty-three. Darker green squares indicate higher positive correlation between the corresponding row's action and the selected action with the matching hour offset. Negative correlations are shown with red squares. The darker the shade of red, the higher the absolute value of the negative correlation.

For example, a user might track when mood is good or bad and could set a reminder to ask for mood once a day in the afternoon. The user could then send a message that she was "feeling good" or "feeling bad." The action keyword would be "feeling" and the unit would be "good" or "bad." Interacting with the correlation interface, the user could select "feeling good" in the drop down menus and then be able to see what actions positively and negatively correlate to her good moods. She could also see how long before or how long after the actions occurred before she was in a good mood. If there are actions that have a high positive correlation, she could try to do what makes her happy more often, such as hang out with friends or exercise in the morning. Similarly, she could select "feeling bad" from the drown down menu and then be able to find actions that she might want to avoid.

Another application could be a search for actions that usually happen during a certain time of day. One user noted that he usually drank coffee three hours after he woke up and drank beer about ten hours after he woke up. Although this might not motivate change in behavior, it did provide a view for reflection and awareness.

4.4.2 Aggregates

When users are more interested in proportions and relative counts than they are changes over time, they can use the interactive word cloud and treemap. The former is an imperfect but familiar interface, and the latter provides an overview in a compact space. Most YFD views have a bar chart in the sidebar that shows aggregates, but only the most common actions or units. On the other hand, the full aggregate views show everything a user has logged.

4.4.2.1 Word Cloud

Word clouds have become commonplace with applications, such as bookmarking site del.icio.us (Yahoo, 2010) or photo sharing site Flickr, that deal with tags or categories. Words are typically sized by usage. The more a tag is used or the more items in a category, the larger the font for that respective tag or category.

As a traditional information visualization tool or a tool for textual analysis, word clouds are inaccurate, and no evidence has shown that the method aids analysis or improves navigation on websites (Viegas et al., 2009). Scaling of words is difficult, because there is white space inside and around letters, so the view is never as accurate as geometric shapes such as bars on a bar graph or dots on dot plot. However, word clouds continue to be ubiquitous online, and continue to be used a navigation tool.

For example, Wordle by Feinberg (2010), is a tool that is available for use on social data visualization site, Many Eyes (Viegas et al., 2007). It lets people copy and paste text into a textfield and create stylized word clouds. Users can change colors, organization, and font. All of these options do nothing for accuracy, but the tool has resonated with many users. As of this writing, users have created and saved over four million word clouds with Wordle. In Viegas et al. (2009), results from a survey found that when using Wordle, people felt creative, felt

an emotional reaction, and learned something new about the text. So although analytic insight is lacking, users seem to gain reflective insight by seeing a body of text in a new format of keywords and various colors. Feinberg (2010) refers to this as a creation of "communicative artifacts." People feel they are creating something meaningful, but this comes more from intuition than from knowing or seeing word frequencies. Feinberg notes that many people do not even realize that words are sized by frequency of use, and he concludes that a "beautiful visualization gives pleasure as it reveals something essential."

Similarly, YFD lets users visualize their actions as word clouds, as shown in Figure 4.10. Action keywords or units are displayed at once for the selected time frame, and areas are roughly sized by frequency of use. Words are also colored by the same color scheme as the other visualizations where darker shades of blue indicate greater levels of use and lighter shades of blue indicate lower levels. Words are sorted by use in descending order, so the most used action appears first and the least used action is listed last.

Figure 4.10: Word cloud, unfiltered on left and filtered on right

As described earlier, the sizing is not completely accurate because letter size is not uniform and spacing of words is not exactly the same between browsers. Text are not stylized or rotated like the words in a Wordle-generated word cloud, but one might argue that the clouds in YFD still provide the same type of emotional insight, especially since the words are personal and contextual for each user. That said, the popularity of word clouds appears to be declining in favor of more traditional graphs like bubble charts and bar graphs, and in some data-specific cases, matrix diagrams and treemaps. Zeldman (2005) referred to tag clouds as the "mullets" of the Internet, and Harris (2011) criticized the method for leaving out important context about the data that clouds represent. This perhaps is further supported by lower usage of the world cloud, discussed in Chapter 5. Nevertheless, the word cloud is provided to YFD users as a supplement visualization rather than a primary view. Perhaps phrase nets by Van Ham et al. (2009), which focuses on word relationships, or the exploratory work by Thiel (2010) which focuses on narrative and relationships, might serve as useful starting points for future tools.

Like other YFD tools, users can change time frames to compare views of different clouds. If a user views a cloud for action keywords only, he or she can click on an action to see the views for that specific action. Finally, in keeping the interaction consistent across all visualizations, word clouds can be searched via a search field location on top. As users type their queries, the cloud updates dynamically. Words that match become more prominent and those that do not match fade.

4.4.2.2 Treemap

The treemap tool in YFD is similar to the word cloud in that it shows aggregate counts for actions and units; however, it is visually more accurate because it uses the area of rectangles to indicate counts rather than words, which can have jagged edges and space in between letters. A squarified treemap (Bruls et al., 2000) is

used, as opposed to the original algorithm developed by Shneiderman (1992).

Although treemaps were originally developed to display hierarchical data, YFD treemaps currently do not make use of the ability to show nested structures. Nevertheless, the view gives users another view which can help users gain deeper analytic insight (Bakker and Hoffmann, 2005). In addition to rectangle areas, color is also used as a redundant visual cue. Rectangles are sorted from least to greatest, starting from top left and ending in bottom right. Labels are provided for more frequently used actions, and counts are displayed on mouse over.

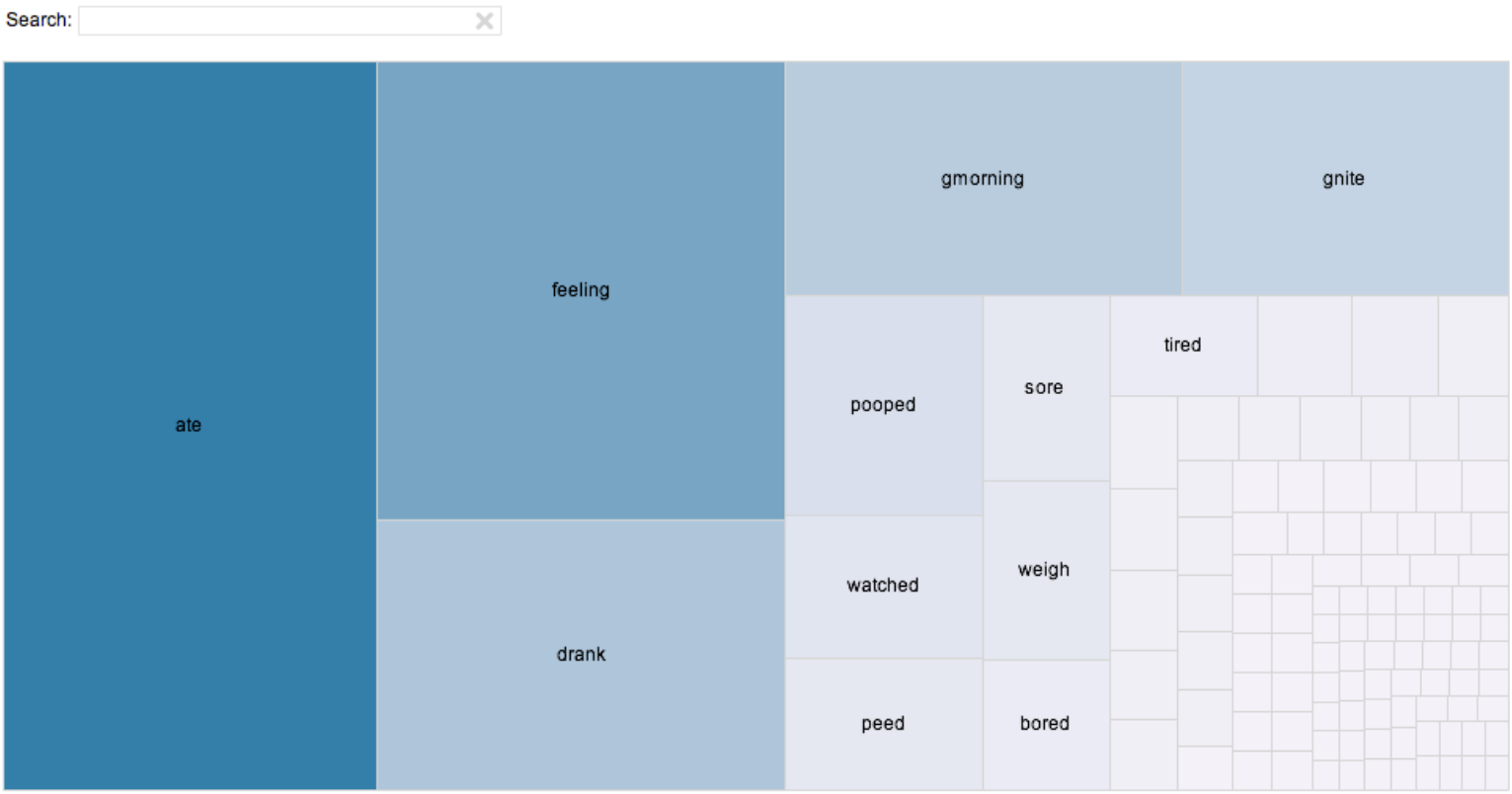


Figure 4.11: Treemap

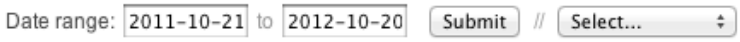


Figure 4.12: Date Range Navigation

Again, as with the other views, a search field is provided so that users can find specific actions or units, and time frame can be selected, which provides an opportunity for comparison between phases. Like the word cloud, matching actions and units are highlighted as a user searches, and non-matching actions fade to the background. A click on a specific action on the treemap takes the user to the action's specific view, as expected.

4.5 Filters and Search

With multiple views of the same data, YFD could feel like a collection of separate visualizations that are unrelated; however, by using common graphical elements and interactions in all the tools, users can visualize their data in different ways and maintain a connection between aggregates and separated aggregates over time. For example, as mentioned several times throughout this chapter, navigation on top of each page (Figure 4.12) lets users select date ranges for the data they want to see. The time range stays consistent as users switch between views so that users can make connections from temporal to aggregate to individual data points.

The ability to focus on specific actions or units is a universal theme across YFD's user interface and graphical elements. Users can search for specific actions and units in all the views either by text entry or via drop down menu. As a new action is selected, the view changes, so that the user can focus on the action of interest. The initial view for these visualizations shows all the actions at once to provide an overview, and users can see their data in greater detail by clicking on the areas they are interested in. For example, the stacked area chart shows flows of different activity over time and the calendar shows intensity for everything

before the user filters down to specific actions or units. The hope in providing such overviews is that users sense that their choices and behaviors are not separate entities that function independently of one another. Rather, the actions they log are related in some way.

That said, views of the details or the individual actions can also be interesting and lead to a deeper understanding of the bigger picture. Besides search in the visualization tools, users can also filter by hashtags, as were discussed in Chapter 3. Hashtags can be used as broad categories for users' data. For example, a user might want to separate food items by meal (e.g. breakfast, lunch, and dinner). If hashtags are used, the subset of data can be explored via the YFD visualization tools. This provides a way for users to separate their data.

Embedded visualizations are another common UI element that let users filter as well as add another view to the data that a user has logged. Willett et al. (2007) refer to these as "scented widgets" where menu items or links are enhanced with visualization. In their preliminary studies, they found that participants made more unique discoveries in unfamiliar data when menu items and navigational elements were enhanced. However, they also found that findings for the two groups equalized as users grew more familiar with the data set.

In the case of YFD, most menus that list action keywords and units are enhanced with embedded visualizations, as shown in Figure 4.13. As a whole, these lists look much like a horizontal bar graph. In overview pages, the navigation elements link to action-specific views, and elements on action-specific pages help users explore their data deeper. For example, views for categorical data types, list units in descending order of usage. When a user clicks on a category, a corresponding histogram over time appears, and the calendar updates to show activity for the selection. This interaction lets users analyze their more specific aspects of their data as well as helps users make conceptual links between various aspects of the visualization. The user can associate full aggregates over a time frame

Figure 4.13: Filters and Embedded Visualization

(e.g. number of times a user was in a happy mood over the past three months) to aggregates over time (e.g. time of day a user was in a happy mood or days of the year a user was in a happy mood).

4.6 Data Sharing

YFD was originally designed to keep all data private, but once YFD launched, users requested a way to share their data. They did not want to share all of it, but they wanted to make aggregates publicly visible. I did not implement a way to share data immediately, as I still thought it was better to keep all data private, so instead, people took snapshots of their visualizations. They shared their findings on Twitter and wrote about their experiences on their blogs. It seems to suggest that a perceived benefit of personal data collection is not just self-reflection and change, but also to share changes and selectively publish findings where others can see. This perhaps relates to YFD as an extension of Twitter, where most people tweet to a public audience.

About one month after launching YFD, I announced a way to share data. People can create custom pages made up of modules. Each module has two options. The first is the action that the user wants to share. The second is what view the user wants to show for that action. There are twelve modules users can choose from, as shown in Table 4.1, which show recent activity from the past thirty days. As shown in Figure 4.14, the combination of modules forms a view similar to a status page or dashboard.

View	Type	Description
Most recent	Single	Most recent entry for selected action.
Most recent time	Single	Time of most recent entry for selected action.
Sum of values	Aggregate	Total count.
Total logged	Aggregate	Total number of data points logged.
Average value	Aggregate	Mean value for selected action.
Values over time	Graphic	Time series of values over time.
Sum of values over time	Graphic	Daily sums of values over time.
Total logged over time	Graphic	Daily number of data points logged.
Sum of values by time of day	Graphic	Hourly distribution of summed values.
Total logged by time of day	Graphic	Hourly distribution of total data points logged.
Categories	Graphic	Distribution of units for selected action.
Recent actions	List	List of most recently logged actions.

Table 4.1: Modules Available for Custom Page

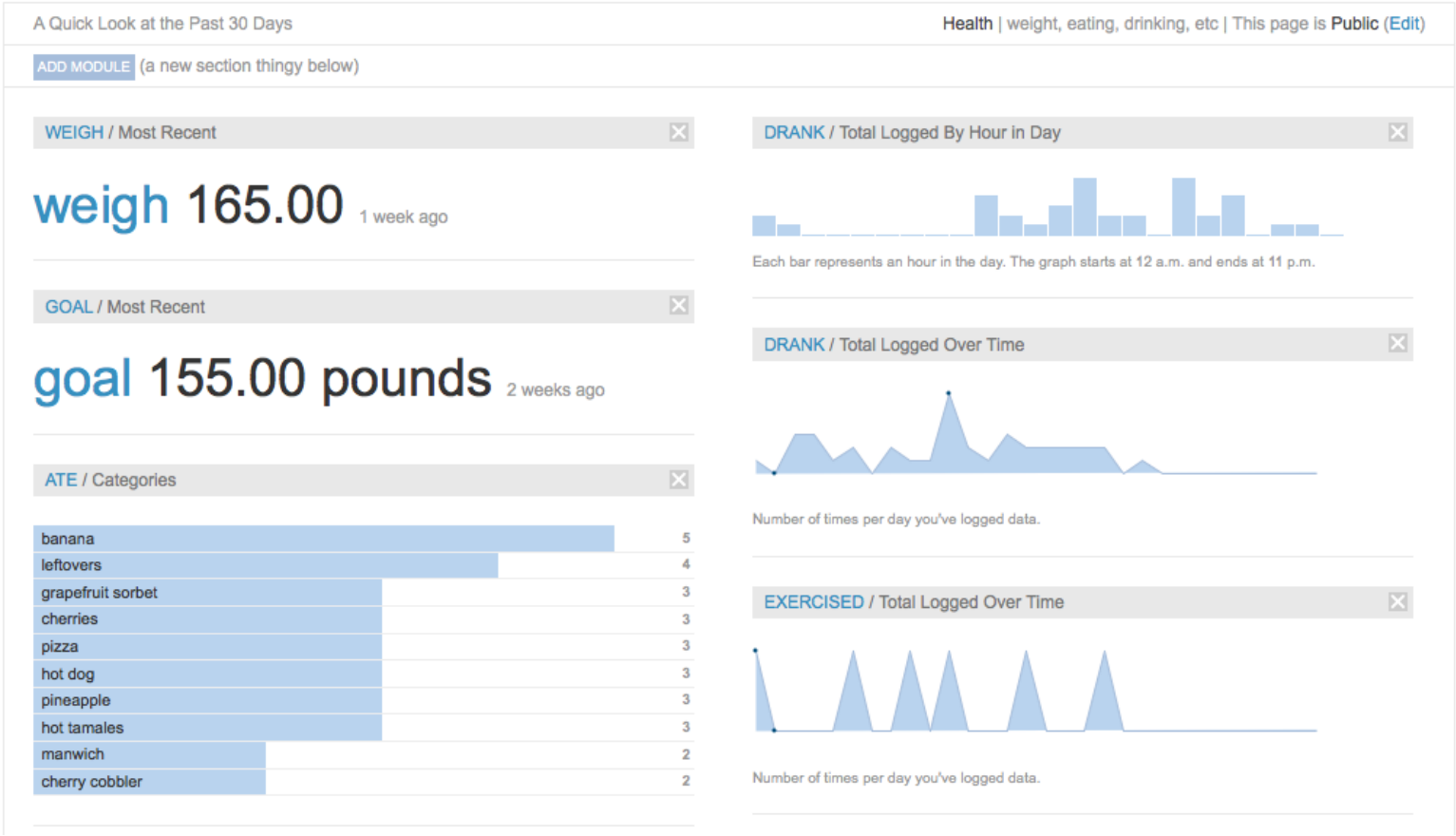


Figure 4.14: Custom Page

Users can create as many pages as they want, and there is no set limit for number of modules that can be placed on a page. Module positions can also be modified to fit a user's liking. Although a single page can contain a variety of views and actions, each page typically has a theme such as health or entertainment. An entertainment page, for example, might have the most recent movie watched, the most recent book read, and a categorical view that showed the genre of movies watched. A heath page might have information about sleep, eating, and exercise. Some users like to share their progress towards reaching a weight goal, and some like to share fitness milestones. One user created a public page that showed his most recent sleep time, so that his girlfriend could see that he was not going to sleep too late.

The flexibility of modules lets users present a variety of aspects in their data as well as control what they share and what they keep private. Many pages are public, but by default, custom pages are still private. Out of about 1,500 pages, only about a quarter of them are currently public. As a private view, the users can create pages that contain the information they are most interested in and use it as a dashboard.

4.7 Summary

YFD provides a set of visualization tools that lets users explore various aspects of their data. The user homepage and actions log focuses on what is most recent and is a quick view into a user's data. These views are also used for verification for the mechanism that translates data logged via Twitter to parsed data stored in the database. The more interactive views help users explore their data in depth, over time and in aggregate. Familiar interfaces and consistent UI elements help users stay connected with their data, and multiple views aid in understanding of how actions are related and compare to each other.

These tools at varied depths of analysis allow users to keep data journaling or personal informatics in the everyday, which involves different types of insight, as discussed in Chapter 2. Personal data collection is an exploratory process with user-specific goals, and these goals change frequently with interests.

In the next chapter, the exploratory process and usage of YFD visualization tools are discussed, as well as how the site is used as a whole and what this suggests for future design of personal data applications.

CHAPTER 5

Usage

5.1 Introduction

This chapter discusses the use of the YFD site, its visualization tools, and data collection patterns by users who come from varied areas of study, have different concepts of personal data, and started collection on their own accord. Because of the mixed user base, usage patterns also varied depending on what each user was using YFD for. Some were interested in self-experimentation whereas others used data collection as a journal.

I first developed YFD as a personal project to collect data about my own life (e.g. weight and mood) rather than a general application that others could use. However, after collecting data myself for a few months, I realized that others had an interest in doing the same, so I expanded the existing YFD application and invited about a hundred people to try it out. Based on informal feedback via email and Twitter from this small group, I developed a third, more general iteration of YFD that let anyone with a Twitter account use YFD.

This development process makes this usage study different from most others on personal data collection. The YFD application has been publicly available for most of its lifespan, and with the exception of the informal invitations sent in the early iteration, people did not receive guidelines on when to collect and were allowed to collect whatever data they wanted. Some users wrote scripts to automate data collection via Twitter. One user shared a JavaScript bookmarklet

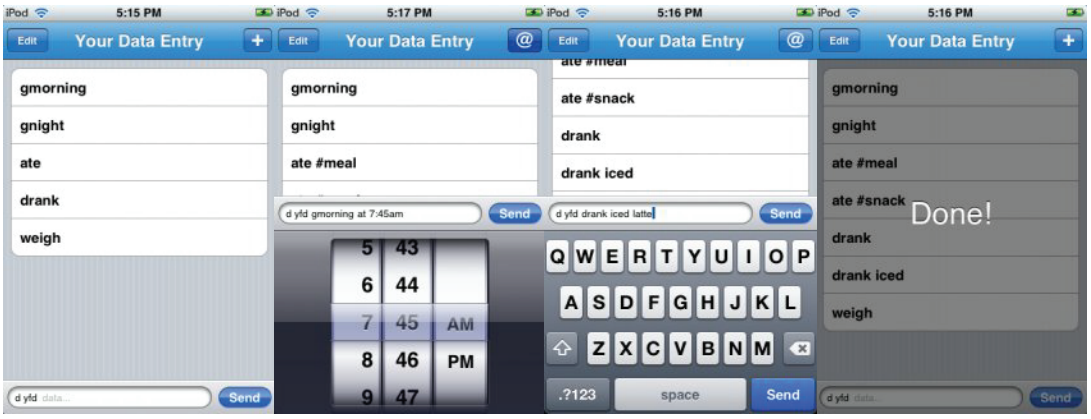


Figure 5.1: YFD iPhone app

to make data logging easily accessible in the browser, and another person wrote a browser plugin. Castillo (2009) developed an iPhone app, shown in Figure 5.1.

In the sections that follow, I describe usage during the multiple-user iterations of YFD, starting with the most recent, followed by the first version. For the current iteration, I describe site usage and interaction with visualizations in detail, as more tools were created and logging mechanisms were written to keep track of user activity on a fine-grained level. An opt-in survey was also administered. For the first iteration, I mainly look at data collection patterns since no code was written to log interactions on the site during this phase.

Figure 5.2 shows a timeline of active user counts and when each study began and ended. While the usage data collected during each iteration is different, there are general questions worth examining for each:

- Do users, regardless of their background or experience with data, gain any insights from using YFD?
- Are there visualizations or aspects of certain tools that users find more useful than others for personal data?
- Does interaction with the site motivate users to track more aspects of their

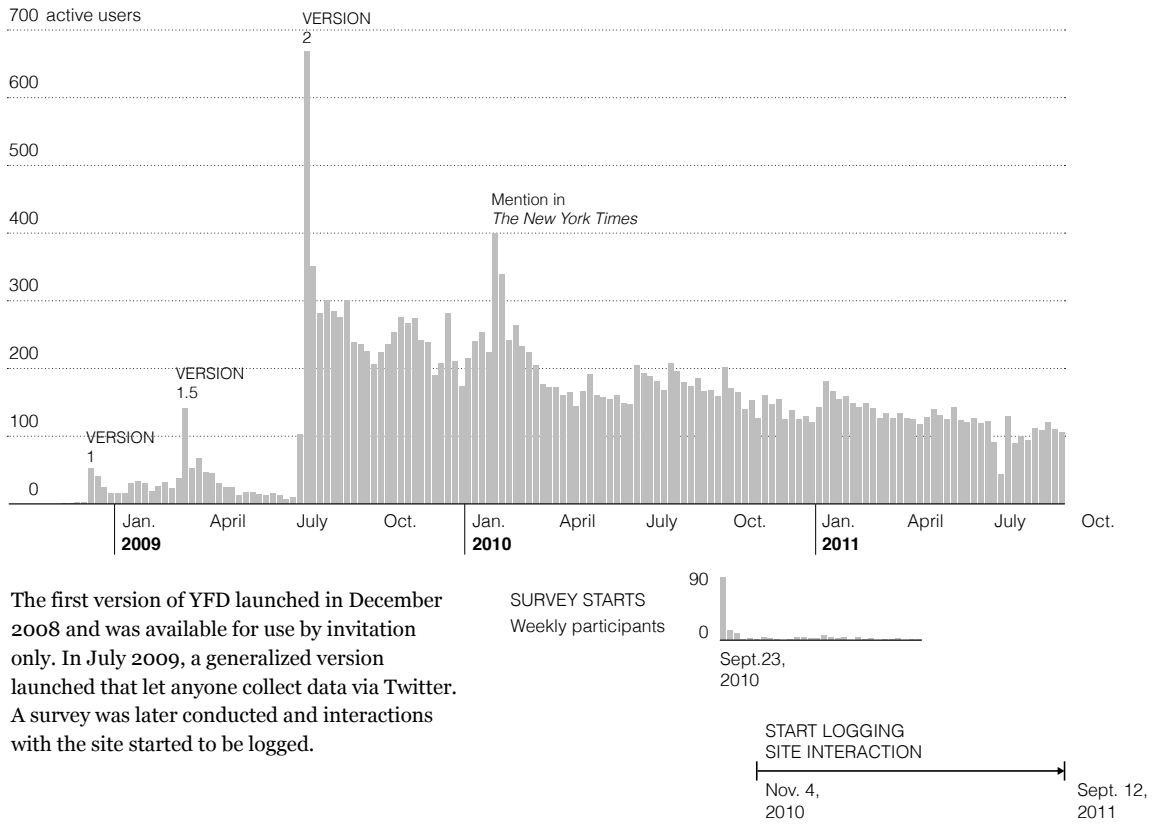


Figure 5.2: Timeline for interaction logging

lives or does data collection bring users to the site?

By exploring these questions, I make recommendations for visualization designed for personal data and the everyday and discuss how designs can differ from traditional statistical visualization.

5.2 YFD 2.0

As described in Chapter 4, YFD offers several tools that let users browse and explore their data. The actions log is a simple listing that shows data collected by time, from most recent to oldest. The bar chart appears on the sidebar of many pages and shows categories and counts. These two views are meant more for browsing than exploration. The calendar, stacked area graph, tag cloud, and treemap are more exploratory and interactive. Finally, the correlation and durations views were specifically designed for YFD and show specific aspects of users' data.

- Actions Log - Simple list of logged actions
- Bar Chart - Shows aggregates for selected time range
- Calendar - Colored according to number of actions logged
- Correlation - Grid showing cross-correlation between actions
- Durations - Time in between a pair of actions
- Stacked Area Graph - Aggregates over time
- Tag Cloud - Words sized by use
- Treemap - Rectangular regions sized by use

In addition to these visualizations, users can also create custom pages and are presented with the YFD homepage upon log in (Figure 4.1).

5.3 Setup

Interaction and engagement with the YFD online application was measured in a few different ways and at different times. General Web traffic to the site was monitored with Google Analytics (Google, 2012a) since the launch of YFD, which allowed tracking of visitors, visits, pageviews, and unique pageviews. Here are the definitions of the four metrics, as defined by Google (2012b):

- *Visits* represent the number of individual sessions initiated by visitors to site. Inactivity of 30 minutes or more are attributed to new sessions, and users who leave and come back to the site within 30 minutes will remain in the same session.
- *Visitors* are those those who come to the site and initiate a session.
- *Pageviews* are counted when a visitor loads a page. Refreshing a page counts a pageview.
- *Unique Pageviews* are the number of sessions a user loads a page one or more times.

These metrics can be viewed over time via the Google interface or downloaded to use with other analysis software, however Google Analytics only provides high-level aggregates on a per-day basis at the most detailed and does not provide data for individual sessions.

Interaction with the site on a per-user and per-session basis was logged onto the YFD server with custom code between November 14, 2010 and September 12, 2011. Each pageview was logged with an anonymized user id and a timestamp. Data collected by individual users through YFD, which was anonymized and encrypted, also had timestamps attached. Note that to keep user data private, the

content of users' logged data was also encrypted, so I could not, for example, look for changes in weight or increased water consumption.

Clicks on the calendar, editing of existing data via the Actions Log, and data deletion were also logged. This usage data allowed more specific insights on how individual users collected and interacted with their data and how similar groups of individuals engaged with the site.

In addition to the automatically logged usage data, some users opted in to take a survey that asked about their professional background, Twitter usage, original purpose for using YFD, and whether or not they found YFD useful.

The survey was taken between September 23, 2010 and May 30, 2011, with the bulk of participation during the first week. Data was also stored on the YFD server. Responses were linked with the previously mentioned interaction data. All questions were multiple choice, with the exception of the one on awareness and the last on tools, which were open-ended and let participants enter what they wanted.

See Figure 5.3 for an overview of what usage data was collected for YFD 2.0. As described, there were four main sources: Google Analytics, personal data collection, interaction with the site, and the opt-in survey.

5.4 General Usage

Since the launch of YFD 2.0, about half a million data points have been logged by over 5,000 users. As of this writing, the site receives 5,000 visits and 12,000 logged data points per month. In total, there have been 600,000 pageviews and 160,000 visits, where 18 percent of those visits are 5 pageviews or more.

Of all visits, 65 percent are by returning visitors; 38 percent of visits are the fifteenth visit or more. Most users are on modern browsers such as Firefox (38

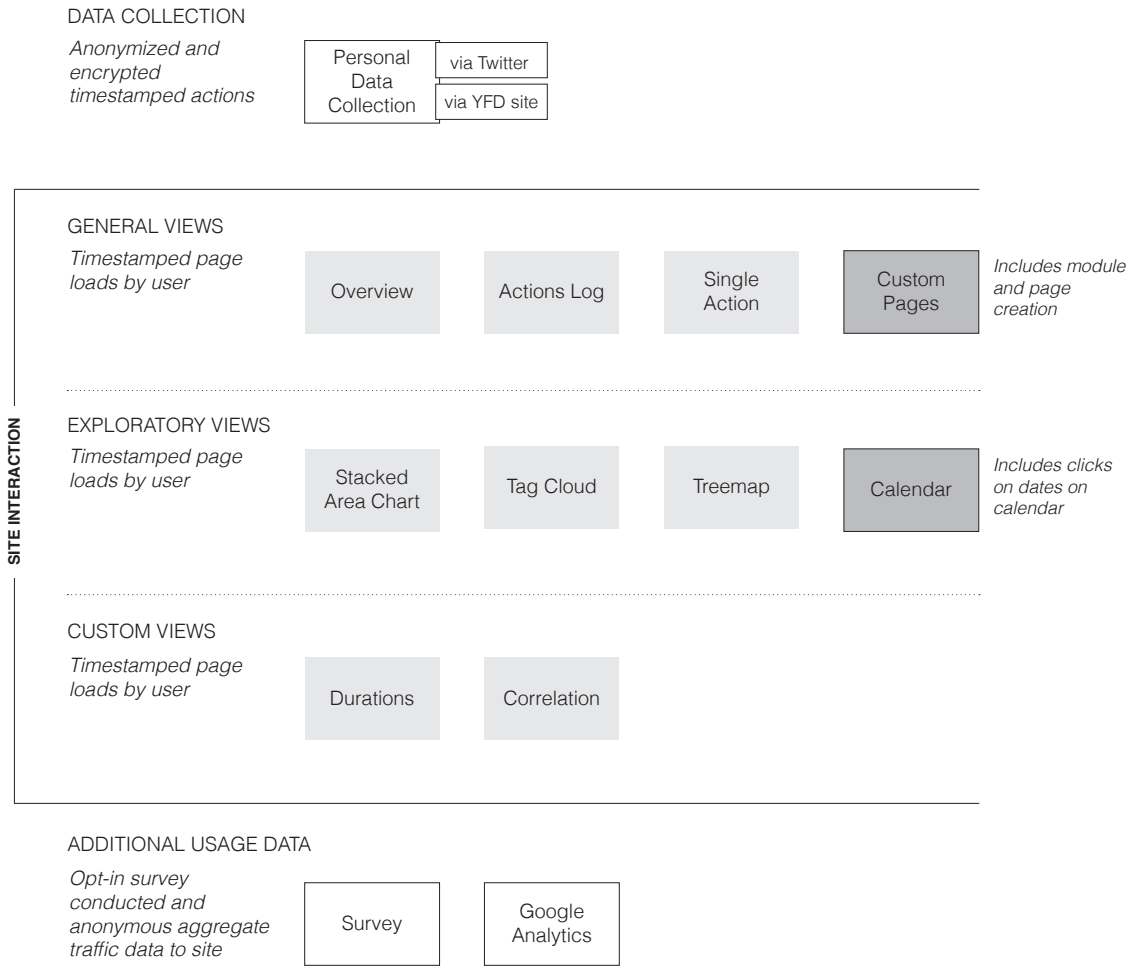


Figure 5.3: Where usage data came from

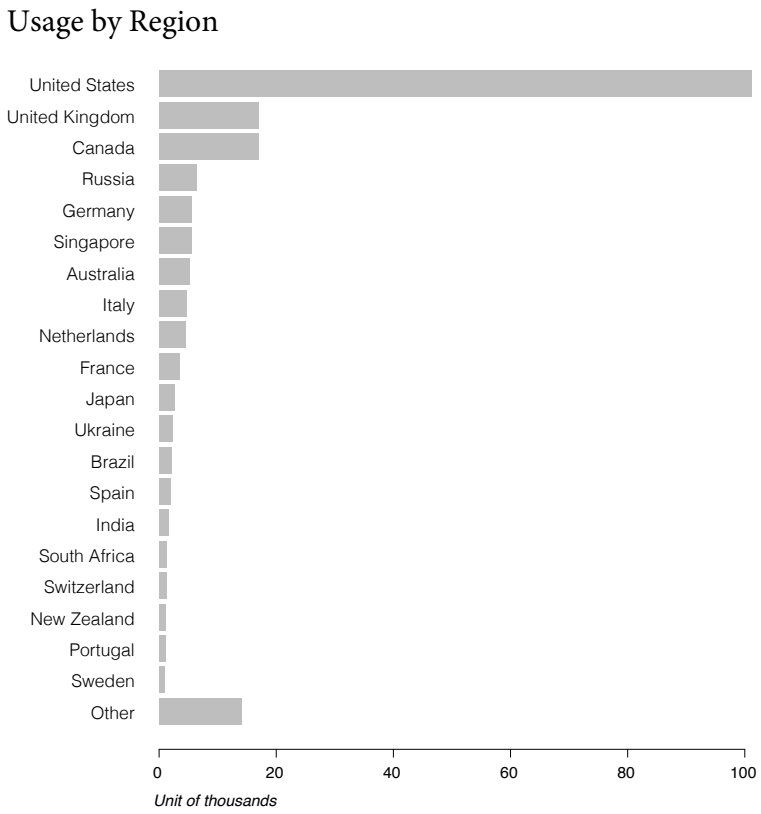


Figure 5.4: Usage by region, as reported by Google Analytics

percent), Safari (31 percent), or Chrome (22 percent). Only 6 percent use Internet Explorer, and nearly all visitors have a DSL or Cable Internet connection or faster.

Just over 50 percent of visits are from the United States, 9 percent are from the United Kingdom, 7 percent are from Canada, and the remainder come from other countries totaling 154 countries or territories around the world. More broadly speaking, as shown in Figure 5.4, 62 percent of visits are from the Americas and 28 percent are from Europe, with the remaining 10 percent from Asia, Oceania, and Africa.

5.5 Survey

Study participants were asked the following eight questions:

- "What was your original motivation in using YFD?"
- "How many days per week do you look at or use graphs in your work or studies?"
- "What is your background?"
- "How often on average do you use Twitter for purposes other than YFD?"
- "Do you use Twitter with a mobile phone?"
- "Has use of YFD made you more aware of your habits? If so, please explain."
- "What visualization tool, if any, did you find useful on YFD?"
- "What visualization and/or analysis tools, if any, would you like to see added to YFD?"

5.5.1 Results

Figure 5.5 shows the response breakdown from 165 participants. Self-experimentation, with 42 percent, was the leading original motivation for using YFD. Journaling followed with 31 percent, and general interest trailed at 25 percent. There were two non-responses.

The background and expertise of YFD users appeared to vary. When asked about graph-viewing frequency, responses were evenly distributed, with the exception of slightly higher responses for full weeks and work weeks and lower responses for the frequencies that were one day less. As for background, 36 percent answered data, which was second to the other category at 42 percent. Design and physical science backgrounds were less prominent.

Although there were people who had never used Twitter before using YFD, most (70 percent) indicated they used Twitter several times a week, and 78 percent of participants said they used Twitter with a mobile phone. Anecdotally speaking,

Survey Responses

Results from an opt-in survey that asked 165 users about their background and how they used YFD.

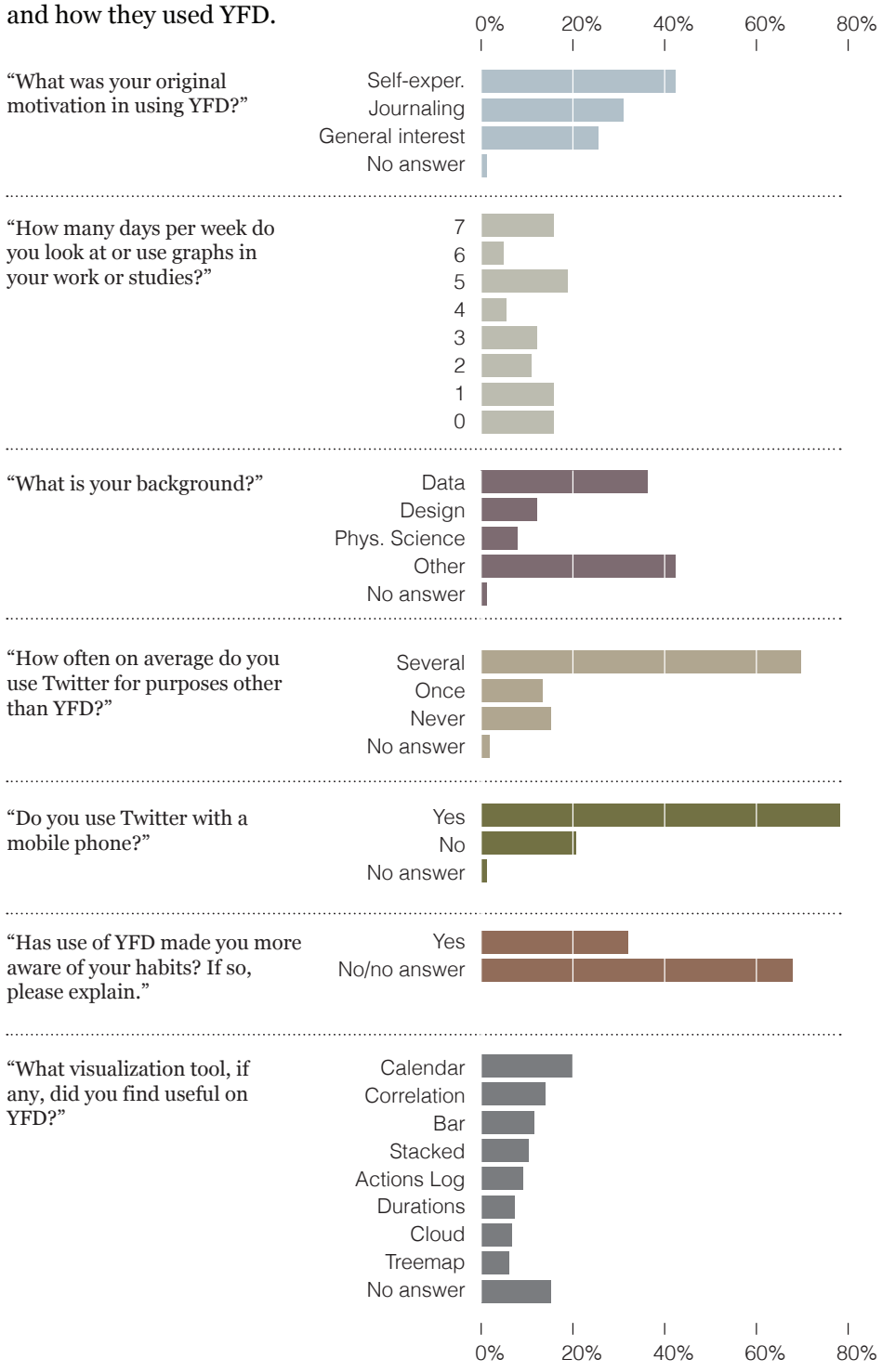


Figure 5.5: Results from opt-in survey

it seemed I had to provide more setup and starting advice to those not familiar with tweeting and sending direct messages. For example, some users not familiar with the direct message syntax on Twitter would mistakenly use the reply syntax.

Survey participants were asked, "What visualization tool, if any, did you find useful on YFD?" They were allowed to pick one or none of the tools listed above. About 15 percent of participants, who were for the most part relatively new to YFD, answered none. The calendar view was selected the most often with 20 percent; correlation had 14 percent; bar chart had 12 percent; the stacked area chart had 10 percent; and log, durations, cloud, and treemap, each had under 10 percent.

Figure 5.6 shows marginal responses to what visualization tool was useful, by motivation to try YFD. There were some differences between those interested in self-experimentation and those interested in journaling. It appears that self-experimenters found aggregation-based visualizations, such as bar charts and the treemap, more useful, whereas journalers found event-based views, such as the actions log and durations, more useful.

Finally, participants were asked, "Has use of YFD made you more aware of your habits? If so, please explain." Many of the answers were more detailed than I expected. They are perhaps the most interesting part of the survey and offer anecdotal evidence of the usefulness of YFD.

Some users drew simple awareness insights that helped them realize that they performed an action more or less than they thought. For example, one user said:

> "I didn't realise I played video games quite as much."

Another kept track of food intake:

> "Yes. I tracked my food and water intake and viewing the logs made me change a few things about what I ate and drank. I used tags to

Preferred Views by Purpose of Use

In the opt-in survey, participants were asked what YFD view they found most useful. Below is the percentage of people who chose a view, separated by their initial purpose for collecting data with YFD: self-experimentation, journaling, or general interest.

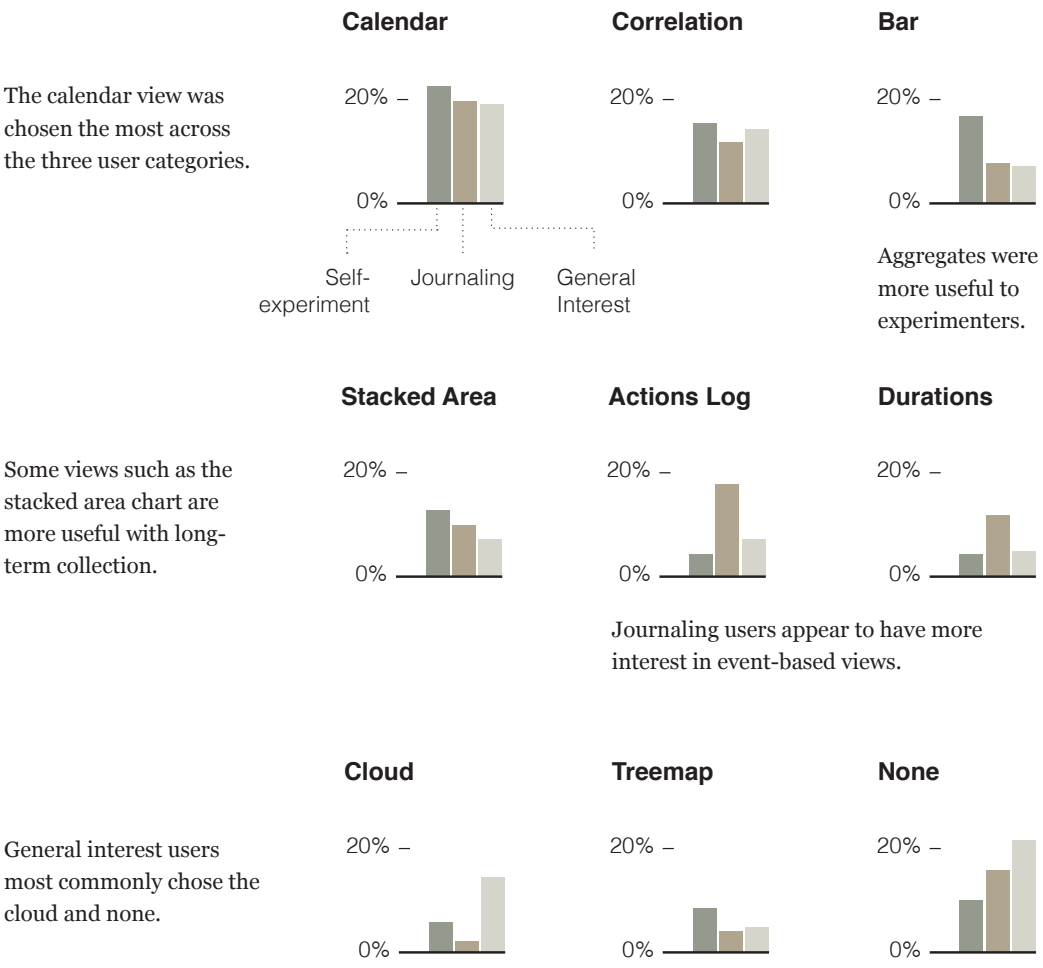


Figure 5.6: Preferred Views by Purpose of Use

> add additional information, like #caffeine, which helped me notice how much I drink caffeine products. I also used it to track the hours I slept and spending habits, which made me more aware of irregularities and areas I need to cut back on, respectively."

Some used YFD with a specific goal in mind, such as losing weight. One user noted:

> "I am getting thin. :-)"

Answers such as these and others described later reflect the flexibility of data collection on YFD and how insights can change depending on the context of the data and how and what data is entered. The answers also strongly suggest the importance of making data collection easy. Many of those who did not feel a heightened sense of awareness while using YFD shared a similar sentiment:

> "I'm new to YFD, so as net, no."

> "Not really. I wasn't very consistent about logging data."

I first encoded these responses as yes, the participant did become more aware, and no, the user did not, for further comparisons. My main interest was in what tools the awareness group found most useful as shown in Figure 5.7. The calendar was selected most often among both groups, and as should be expected, those who did not feel more aware selected none of the tools much more frequently than the awareness group.

The aware and non-aware groups can be divided more specifically though. After all, those who selected none of the tools are not necessarily the same as those who said they did not become aware. A non-response might represent users who used their own software or that all tools were found equally useful. So not aware and non-responses were separated.

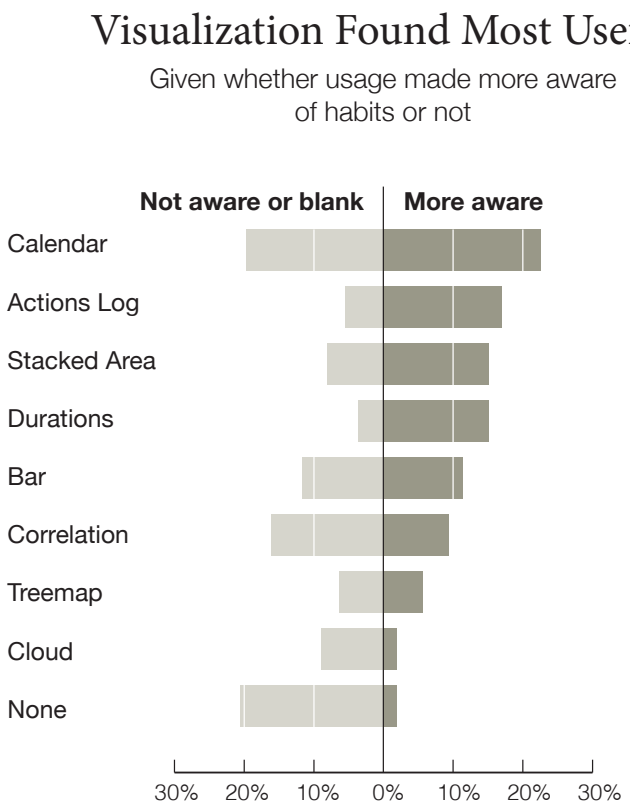


Figure 5.7: Visualization found most useful

Additionally, there are also various degrees of awareness. Responses that were classified as aware were reclassified into three groups: elementary, intermediate, and overall. This grouping is based on Friel et al. (2001), which describes three levels of statistical questions answered by reading graphs. An elementary question is one answered by a direct reading of the data such as, "What was my weight on May 20?" An intermediate question is one that asks about relationships such as, "By how many pounds did my weight change over the past year?" An overall question moves beyond a single dataset and considers the context of all datasets: "How has my weight changed over the past year and how does it relate to other factors like exercise and eating?" Although Friel et al. (2001) was written in the context of education, the concept of graph sense seemed fitting. However, it was unclear what some responses belonged to, because the survey question was about awareness and not directly about sensemaking. Such responses were classified conservatively. For example, the previously mentioned "I am getting thin" was classified as elementary, but it is possible the user saw an overall trend or associated weight data to other habits.

Using these more detailed classifications, Figure 5.8 is an elaboration of Figure 5.7. The top three visualization choices for each group are shown. The difference between non-aware and no answer is notable as are those in between the three awareness groups. The most common among those who were not more aware was a blank response, whereas the cross-correlation matrix was most chosen among the no answer group. The calendar is the only view selected most by all groups, although there was high variability within a small n of 10 for the overall. The equal selection for the Actions Log in the Elementary group is also worth noting as it ties in with the type of question answered by those users.

Figure 5.9 shows the full results for tools, purpose, and awareness. Again, notice the high response rate for the actions log and calendar among journalers and the calendar and bar chart for experimenters. In the former group, awareness lev-

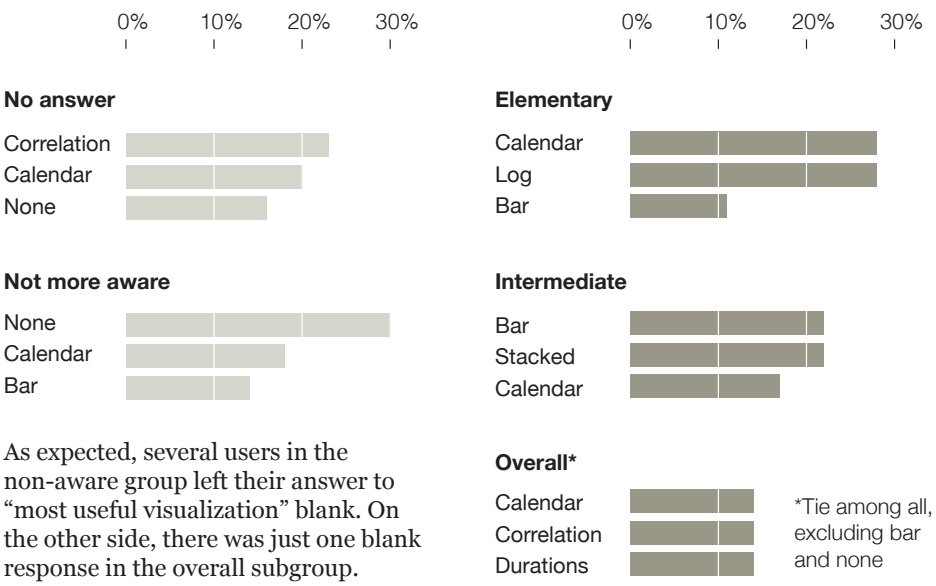


Figure 5.8: Most picked visualization

els are highest for the actions log, whereas they are much lower for experimenters.

5.5.1.1 More Responses on Awareness

There were many other responses to the awareness question that provide a view into how YFD was used and what people learned from personal data collection.

Not everyone was after specific insights. Some logged data as a journal, which drew a general feeling of awareness:

> "Very much so. I use it to track the essentials of my day-to-day, such as sleep, food intake, mood, and exercise. Noticing trends makes me aware of shifts in my homeostasis, giving me the opportunity to make different choices and interactively explore the consequences over time."

> "I track personal metrics in bursts, which shed light on the meta habit of sporadic self surveillance. The metrics are mostly behavioral: smoking, drinking, spending, running, cycling, dining, and romancing. The first three I do too much, the second pair, not enough, and I dine alone more than I romance."

Others used YFD to keep track of performance and improvement, such as in writing a book or running further and faster. In these cases, the insights are specific and premeditated.

> "Definitely. I use it to track my weight and running. Can easily see how many miles I run ea month, and how my weight is changing over time."

> "Yes, in many ways it has. I am a writer and it helps me keep track of my writing habits and moods. And, has allowed me to discover patterns in my themes and content."

Survey Responses and Awareness

By classifying answers to the awareness question of the survey as elementary, intermediate, and overall, the choice of visualization tool and level of insight can be seen.

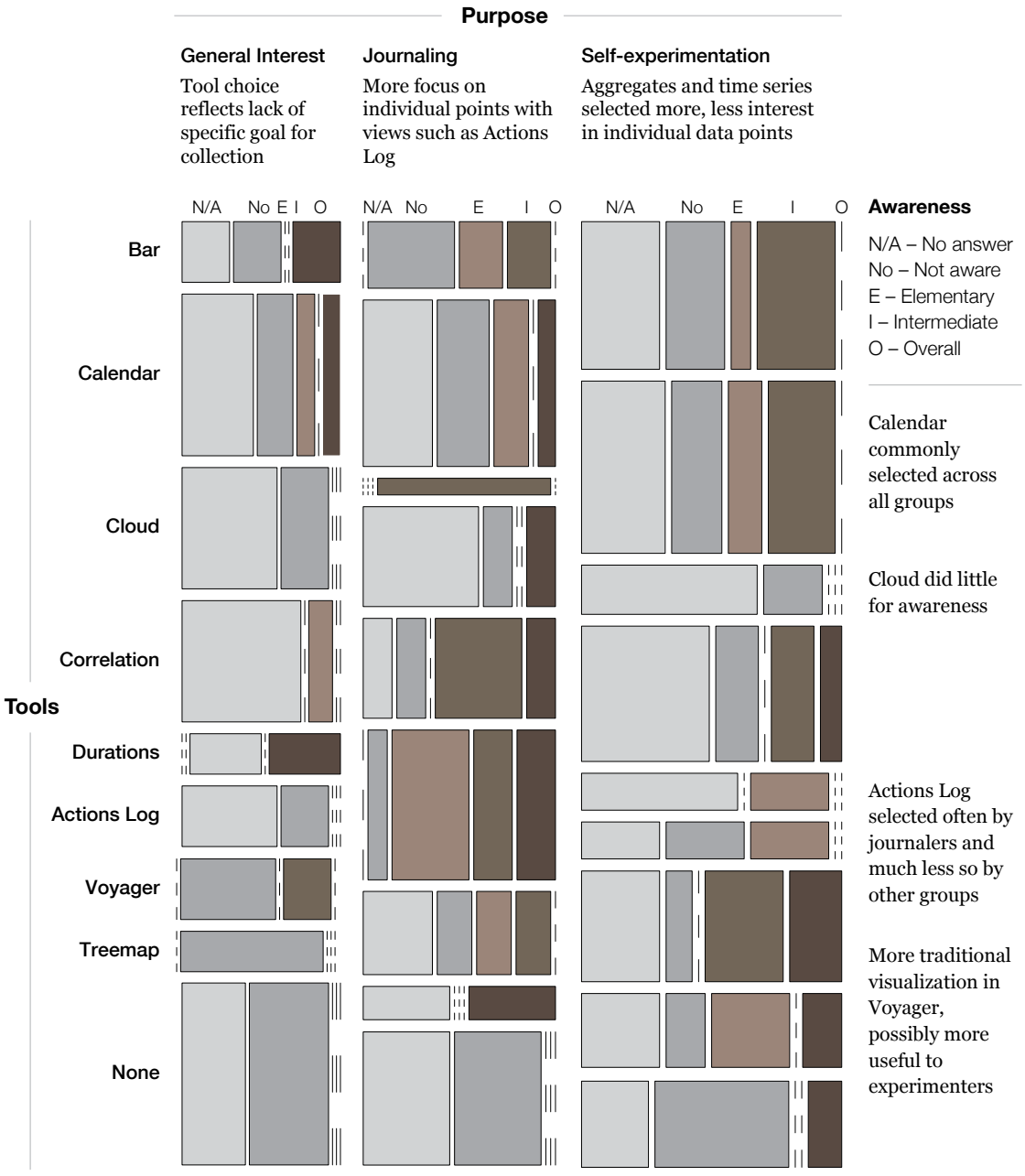


Figure 5.9: Survey responses and awareness

Sometimes such personal data collection led to unexpected results:

> "I haven't used it as much lately, but I gathered a great deal of insight from months of heavy usage. One particular takeaway was that I tend to be most productive at work around 10:00am-11:00am. I simply logged myself as being productive whenever it came up, and months later I could see that it centered around the same time. I now hold that time as precious for getting things done. I also tracked what I drank for a while and found that I was drinking less water than beer!"

> "Yes, in many ways. I could see basic stuff like how much sleep I was really getting, but also found more complex and interesting discoveries. For example, last year I was being bullied at school and I found a correlation between what color shirt I was wearing and how I was treated by my peers."

> "Yes; the cross-correlation in particular showed me how certain food and drink correlated with my waking and sleeping patterns in ways I honestly did not expect."

One user noted a pattern not with the context of the data, but the method of collection:

> "At first, it excerpts the tendency of going to bed late when I am not really efficient, and thus, persevering in inefficiency the dayafter. Then, I notice that I need to be more specific on how I measure things: efficiency is a kind word but it does help at classification. It is fuzzy. So, YFD helped me realize that I do not specify measurable targets in my work, except that of a fuzzy feeling of accomplishment."

Finally, while personal data collection is inwards facing, the awareness is not always about the self. One user discovered something new about neighbors while tracking his or her own sleep patterns.

> "Some things I thought I did frequently were actually rarer and other things the inverse. It's also made me more aware of other people's habits! I monitored when I was getting woken at night and realised there was a pattern that looked like shiftwork. I don't work shifts so I talked to my neighbour and discovered he was waking me up!"

Others were able to consistently log data, but could only keep track of a handful of metrics at a time. Actions that happened more rarely and were not part of a daily routine often were not logged:

> "Yes, I use it most consistently to track my work productivity. It has become second nature to log the completion of a task in YFD as soon as I complete it. For other less routine tasks, I have not kept up with it as much as I would have liked."

This limited amount of data led to limited insights, such as one user who regularly logged weight, but not much else:

> "Only slightly. I use YFD to track my weight and various activities (gym workouts, eating out, long bicycle rides, etc.) that I think might correlate with changes and trends in my weight. YFD has not revealed any correlations or trends that I had not previously suspected. But it has given me a longer time series of my weight than I have ever previously kept."

5.5.2 Discussion

Answers to the multiple-choice questions and the unexpectedly more detailed answers to the open-ended awareness question provide some insights about personal data collection.

- Tool preference varies with usage and purpose.
- People gain different insights depending on what data they collect and how they collect it.
- Non-experts can interact with data and gain insight.

Among self-experimenters, journalers, and those with a general interest in personal data collection, the last group – the one without a specific goal – found YFD least useful. For the most part, they did not become more aware and did not find any of the visualization tools useful. Those who did select a visualization chose either the calendar or cloud most often. There was a consensus around the calendar across all groups, but the prominent selection of the cloud was specific to the general interest group, suggesting the cloud visualization might be useful to provide familiarity to a wider audience. That said, no one in the general interest or self-experimentation groups who chose the cloud said they became more aware and only a few in the journaling group said they became more aware. This suggests that the cloud, if used at all, should only be a gateway to other tools – not an endpoint.

This is an ongoing theme in YFD users' visualization choices. With the exception of the actions log, which is not really a visualization, there were no tools in any group where participants noted awareness in the elementary, intermediate, and overall levels. For example, journalers and experimenters who selected the durations tool noted awareness as intermediate or overall. This makes sense, because the durations tool takes two action names as arguments. Users are forced

to think about more than one action at a time. The results were similar for the correlation tool, which also places actions in a linked context. On the other hand, the calendar, which was the most commonly selected tool overall, had a low proportion of people note overall awareness. Most users who selected the calendar described elementary and intermediate awareness.

So instead of trying to design a monolithic tool that lets users see all the angles of their data, it is perhaps best to build multiple tools that let users focus on the facets they are interested in at any given time. In turn, users might also learn more about their data, or about themselves rather in the case of YFD, as suggested by Bakker and Hoffmann (2005).

The variety of tools an application provides then depends on the purpose and audience. An application for a general audience requires flexible visualization that provides familiarity and guidance towards more advanced findings. However, if users have no interest in more advanced findings and are only interested in answering elementary questions, then it is not worth spending time building more exploratory tools. On the other hand, an application for journaling should first provide views of specific data points or snapshots of small timeframes and then let users look deeper. For experimenters, aggregate views are more important. Although, this is not to say that experimenters do not need basic views.

In summary, based on survey responses, it seems more efficient to design personal data applications with specific purposes than to try to provide as much flexibility as possible. A specific collection purpose also means more strictly guided design while providing more opportunity for goal-specific features and tools.

In the next section, I describe usage of the YFD site via interaction logs and how it corresponds to survey responses.

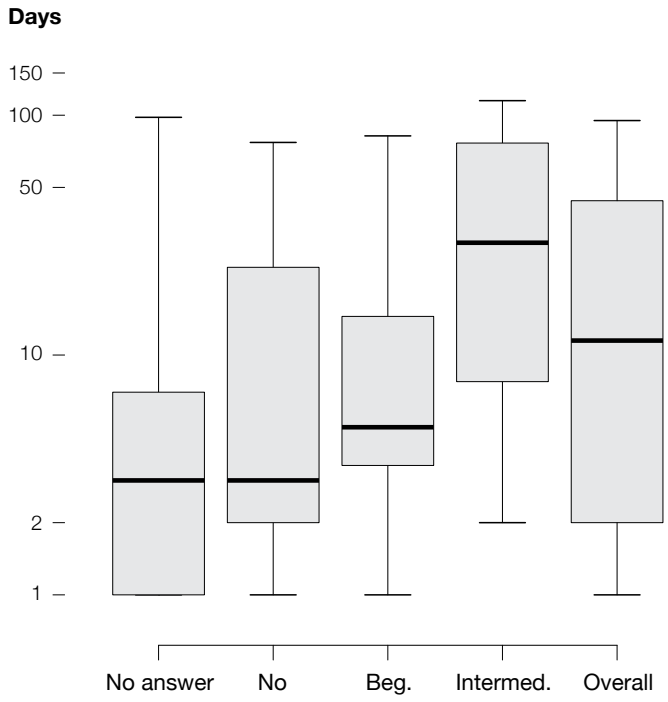


Figure 5.10: Site usage days

5.6 Interaction and Visualization

5.6.1 Survey Participants

In addition to the survey itself, I looked at the interaction logs associated with each survey participant. There were 85 survey participants who also used the site after interaction logging was implemented. Figure 5.10 shows the number of days users interacted with the site for each awareness group. As might be expected, those who became more aware with usage tended to spend more days with the tools.

Furthermore, Figure 5.11 shows a similar breakdown of awareness groups; however, it shows usage for each tool. The actions log and calendar views were the most used across all groups, and the voyager, cloud, and treemap showed decreased usage among those with more advanced findings. The aggregate views showed higher usage among those who said they did not become more aware,

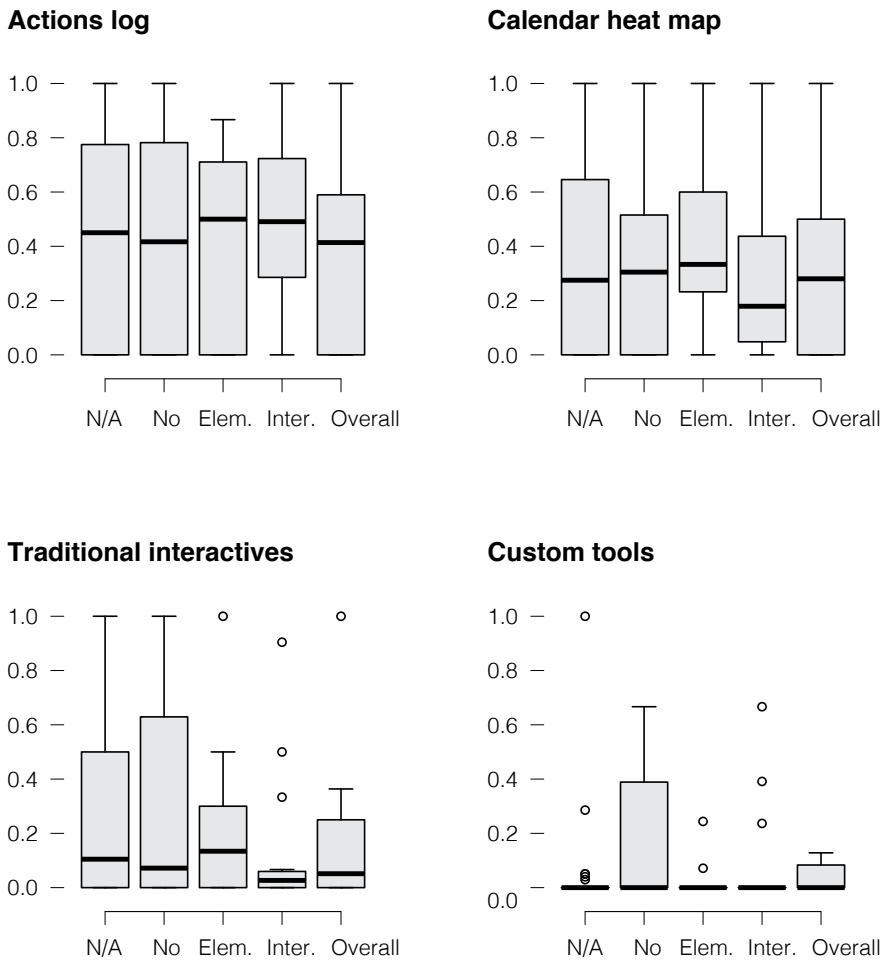


Figure 5.11: Tool usage by awareness group

which goes against intuition, but again, many of the users who answered 'no' also said they just started using YFD or did not collect enough data for it to be useful. These higher proportions for new users is discussed later.

From another angle, Figure 5.12 provides a view of tool usage based on what visualization participants selected as most useful on the site. For example, those who selected the calendar as most useful, visited the actions log and calendar most days on the site, the voyager occasionally, and rarely looked at other views. My expectation was that those who selected a given tool would use that tool

a higher proportion of days relative to others. This was not the case though, with the exception of the durations view. Many users spent noticeable time with other visualizations other than the ones they chose, which further supports an application design that incorporates multiple views.

As shown previously in Figure 5.6, tool preference differed by purpose of use. Self-experimenters preferred aggregates such as the voyager, whereas journalers gravitated towards event-based views such as the actions log. This difference in preference is also seen in the daily usage, as shown in Figure 5.13.

Similarly, a difference is seen when looking at interaction and collections rates per day. Figure 5.14 (right) shows the distribution of points logged per day for journalers versus self-experimenters. The median for self-experimenters was about one more data point logged per day, on average. Interaction-wise (left), self-experimenters used the site more per day than journalers. Here interaction is defined as any page load on the site. The median for the former was 9.79 interactions per day, whereas journalers had a median of 8.33. Both groups had higher means, which was due to two or three users who used YFD for a short amount of time and had a handful of high activity days.

Compared to the overall user population, survey participants were more active in number of days they collected data and in number of days that they used the site. The differing distributions are shown in Figure 5.16. For both groups, the large number of users who were only curious about the site but not interested in personal data collection can be seen in the left skews.

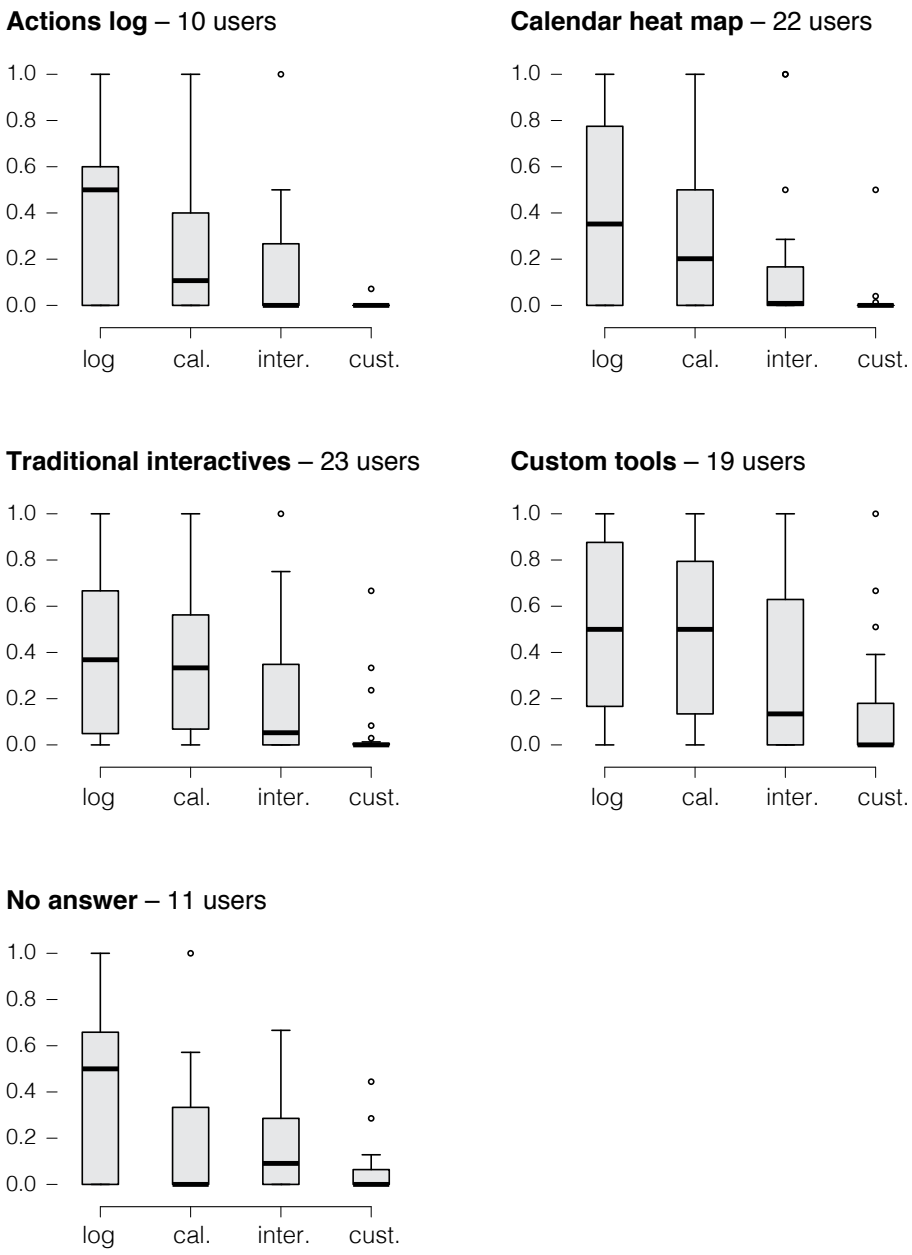
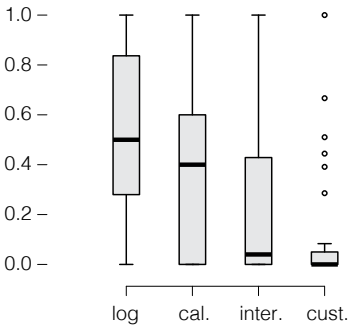


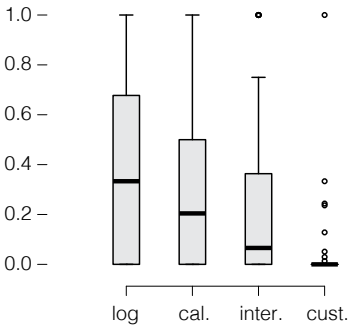
Figure 5.12: Tool usage by what survey users found most useful

Journaling



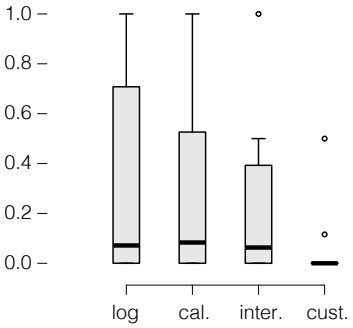
Journalers, who preferred the actions log and calendars over other tools, used the former more often compared to the self-experimenters and general interest group.

Self-experimentation



Self-experimenters showed lower usage for all tools; however, this is possible because that there were users in this group who had spent more days with YFD.

General Interest



Those in this group spent the least time with the site and often did not collect much data.

Figure 5.13: Tool usage by purpose of use

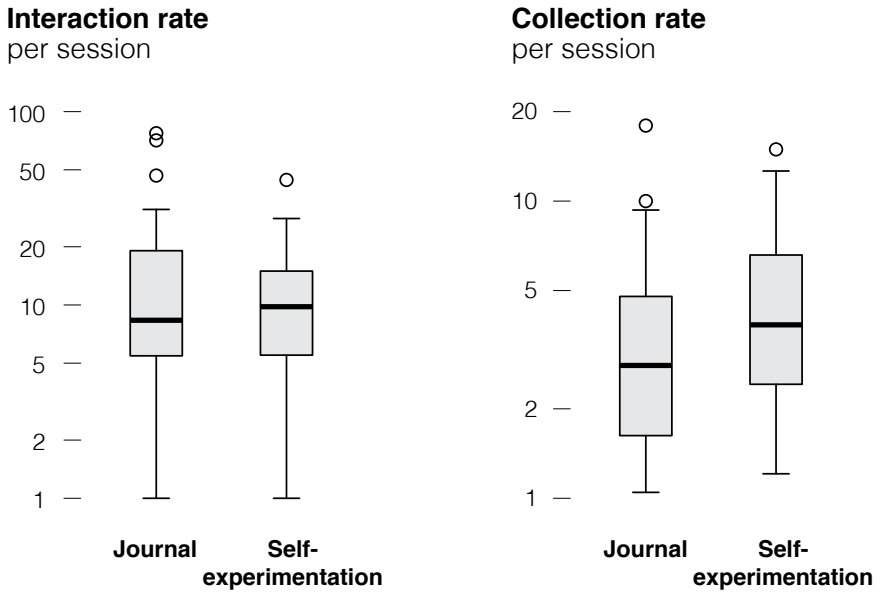


Figure 5.14: Interaction and collection rates for journalers and self-experimenters

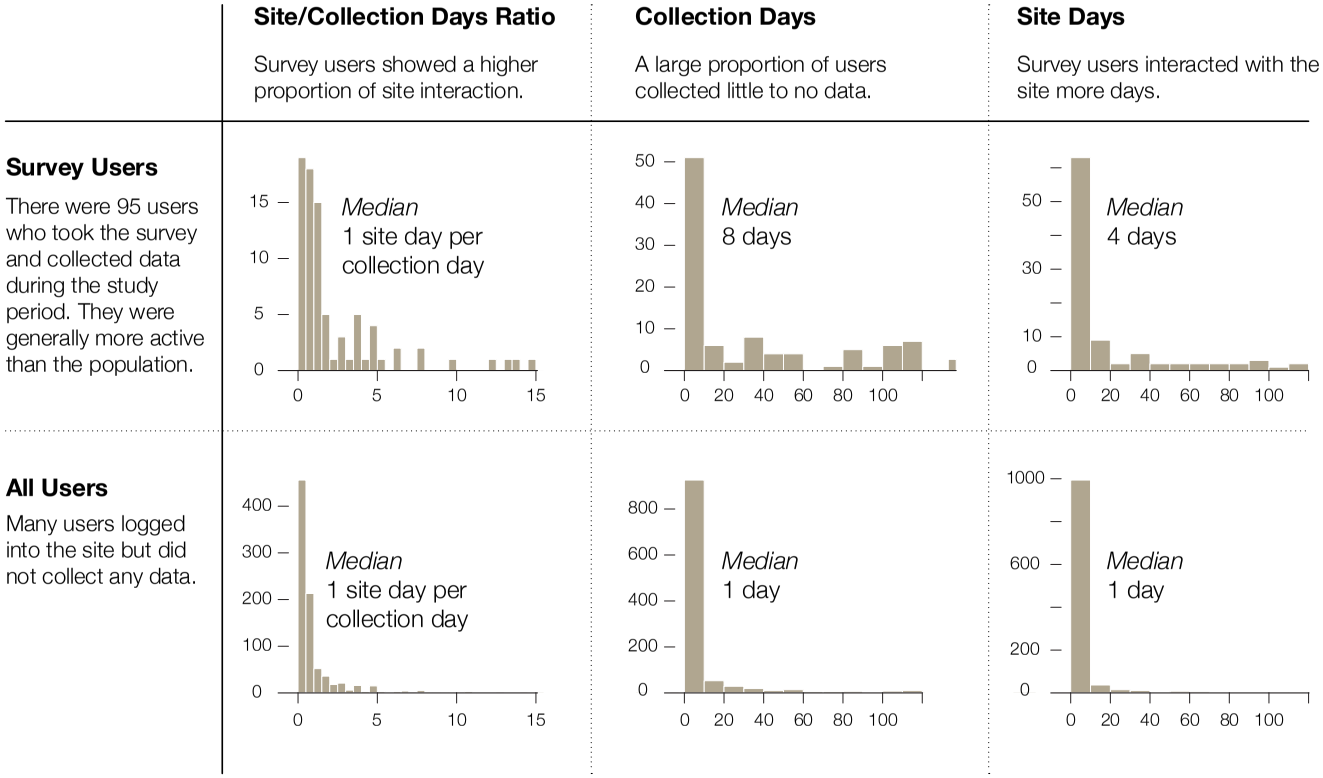


Figure 5.15: Survey users versus user population, collection and site days

In the distribution for collection days of survey users, there are small bumps in the thirty and 100-day range. The former group is most likely people who discovered YFD after I posted a survey recruitment letter about one month before interaction logging began. The latter group is likely composed mostly of those who had been using YFD before the survey began and were more inclined to answer the survey because they invested more time with the application.

Figure **??** shows tool usage distributions for the user population and survey participants. Because survey participants used the site more than all users as a whole, a subgroup of users who had at least three site days is also shown. The survey group appears to match this subgroup more closely. Table 5.1 shows a more detailed summary of tool usage among the groups.

5.6.2 All Users

In this section, visualization usage for all users is examined. Again, usage was recorded for a set amount of time, but users were able to start and stop collecting data when they wanted. Some users started before interaction logging began, whereas others started towards the end. Some users visited the site once out of curiosity, and others visited daily. These variable usage patterns were considered in the analysis.

When looking at tool usage for the overall population, it was important to account for the number of days each user visited the site. As shown in Figure 5.17, there was a varied interest between new users and more consistent users. The longer that users stayed with the site to collect data, the more usage tended to be dominated by the actions log and calendars. Those who only used the site for two days evenly visited each tool, most likely to explore the site and see what visualizations were offered. In the second quartile – users who visited the site three or four days – the change in tool usage versus the first quartile is obvious,

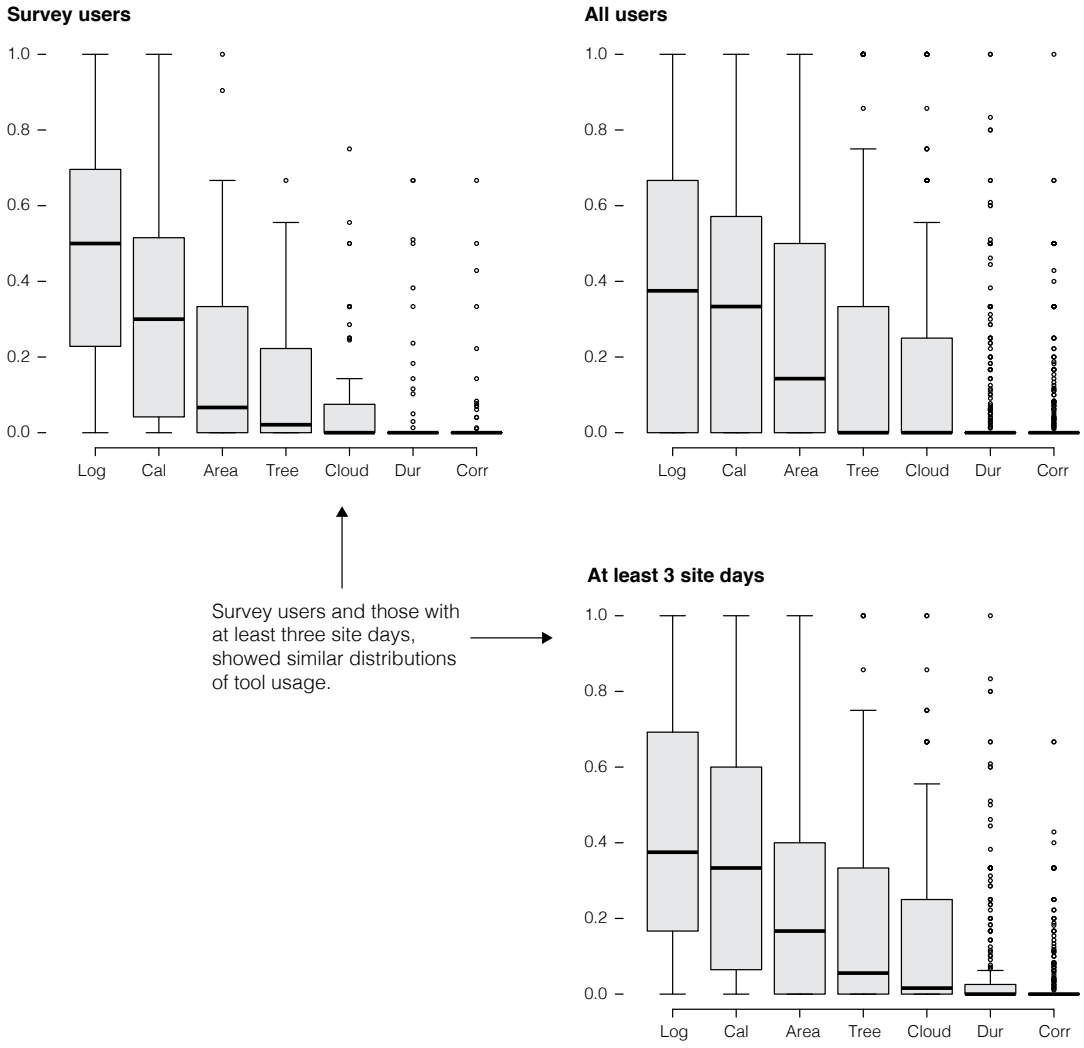


Figure 5.16: Survey users versus user population, tool usage

Group	Tool	Med	Mean	Var	Max
Survey users	Actions log	0.50	0.46	0.11	1.00
All users	Actions log	0.38	0.40	0.11	1.00
At least 3 days	Actions log	0.38	0.44	0.10	1.00
Survey users	Calendar	0.3	0.33	0.09	1.00
All users	Calendar	0.33	0.36	0.10	1.00
At least 3 days	Calendar	0.33	0.36	0.09	1.00
Survey users	Stacked area	0.07	0.18	0.05	0.90
All users	Stacked area	0.14	0.26	0.09	1.00
At least 3 days	Stacked area	0.17	0.25	0.07	1.00
Survey users	Treemap	0.02	0.11	0.02	0.67
All users	Treemap	0.00	0.18	0.07	1.00
At least 3 days	Treemap	0.06	0.17	0.05	1.00
Survey users	Cloud	0.00	0.09	0.03	0.75
All users	Cloud	0.00	0.14	0.06	1.00
At least 3 days	Cloud	0.02	0.14	0.04	1.00
Survey users	Custom	0.00	0.07	0.02	0.67
All users	Custom	0.00	0.08	0.03	1.00
At least 3 days	Custom	0.00	0.09	0.03	1.00

Table 5.1: Tool usage summaries as proportion of site days for survey users, all users, and those with at least three site days

and the proportions are less evenly distributed. In the third quartile, the actions log, calendar, and the voyager are used most often, in that order, and finally, users who visited the site more than twelve days during the study period tended to check the actions log between 30 and 80 percent of the time and the calendar between 15 to 50 percent of the time. The other visualization tools were used relatively less.

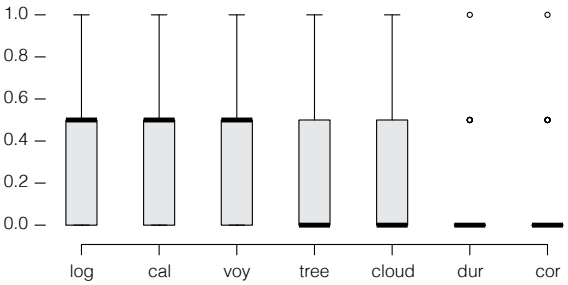
In addition to the proportion of days a view was visited, it is also useful to know how much time is spent with a tool when it is used. Average time spent on each view can be seen in Figure 5.18. The actions log had the highest average close to a minute and a half, which makes sense because people use the view to edit and delete data points, and it takes time to enter values in fields. The calendar, durations visualization, and voyager were the top three after the actions log. The treemap and cloud had the lowest average, which corresponds to the survey response for what visualization people found most useful.

5.6.2.1 Daily Usage

Figure 5.19 shows usage at a more granular day-by-day level for a sample of users and is provided for a sense of the varying amount of interaction with the YFD site. Each grid represents usage for one person between November 2010 and March 2011, and the days from Sunday to Saturday run left to right. Interaction ranges from usage every day or weekdays only, down to a visit per month. Some users, towards the top, interacted with the site often every day, indicated by dark squares, and others, towards the bottom, visited much less regularly.

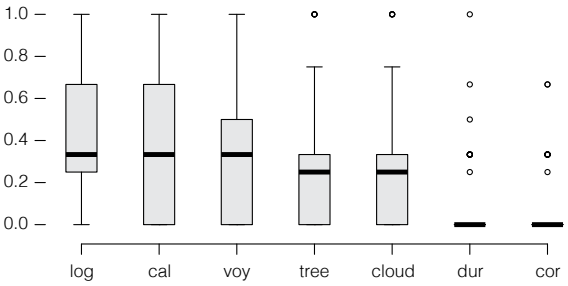
Data collection activity for the same users are shown in Figure 5.20. Like interaction levels, collection patterns also vary by user. Generally speaking, the collection levels look like an exaggerated version of the interaction levels, which seems to suggest an order of actions on YFD. However, while higher levels of

First Quartile, 2 to 3 days



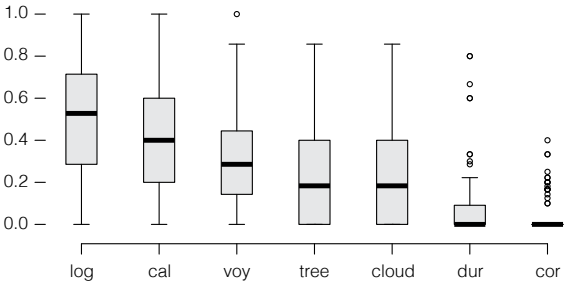
Users spend the first couple of days familiarizing themselves with the site.

Second Quartile, 3 to 5 days



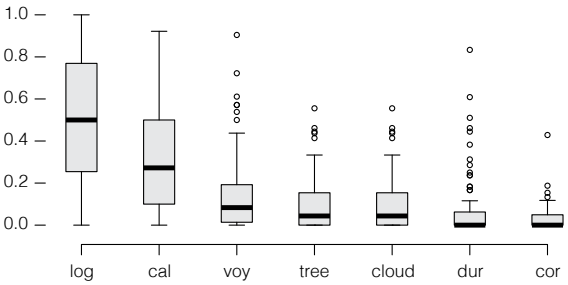
Data collection begins to become the focus.

Third Quartile, 5 to 12 days



More data has been collected at this point and some of the tools such as the voyager and durations explorer become more useful.

Fourth Quartile, More than 12 days



Tool use converges on more causual views such as the actions log and calendar.

Figure 5.17: Tool usage by site days

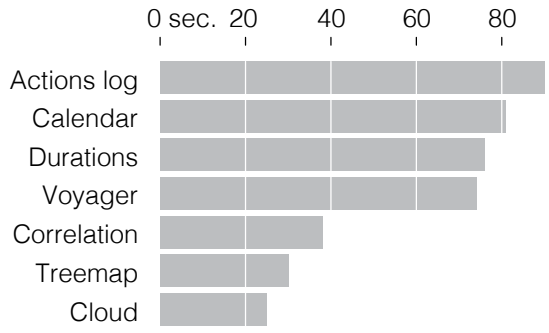


Figure 5.18: Average time spent using tools

interaction typically indicate higher levels of data collection (see Figure 5.22), it does not always go this way vice versa. It was common for people to interact with their data on the site rarely or at a low level, but collect often at a high level. For example, the first user in first row Figure 5.19 and Figure 5.20 showed lower levels of interaction (although almost every day), but collected relatively higher volumes of data.

One obvious reoccurrence is a drop in activity during the December holiday season. For example, the first user in the second row logged data almost every day in November but did not log data at all for more than three weeks around the end of December and beginning of January. The user logged regularly again after that. On the other hand, the first user in the third row stopped using the site and collecting data around the same time after high regular usage and never regained momentum. Some users, such as the last in the second row, showed the opposite with increased usage on start of the new year. Anecdotally, users with these patterns seemed to exhibit a new year's resolution burst.

Low interaction levels did not necessarily indicate low levels of data collection and vice versa, nor did high levels of interaction mean high levels of collection. For example, when a user only shows two days of logging and very little interaction with the site, one might conclude the user is not engaged. However, User D in Figure 5.21 collected data in low volumes but was consistent and showed higher

Figure 5.19: Interaction levels over time

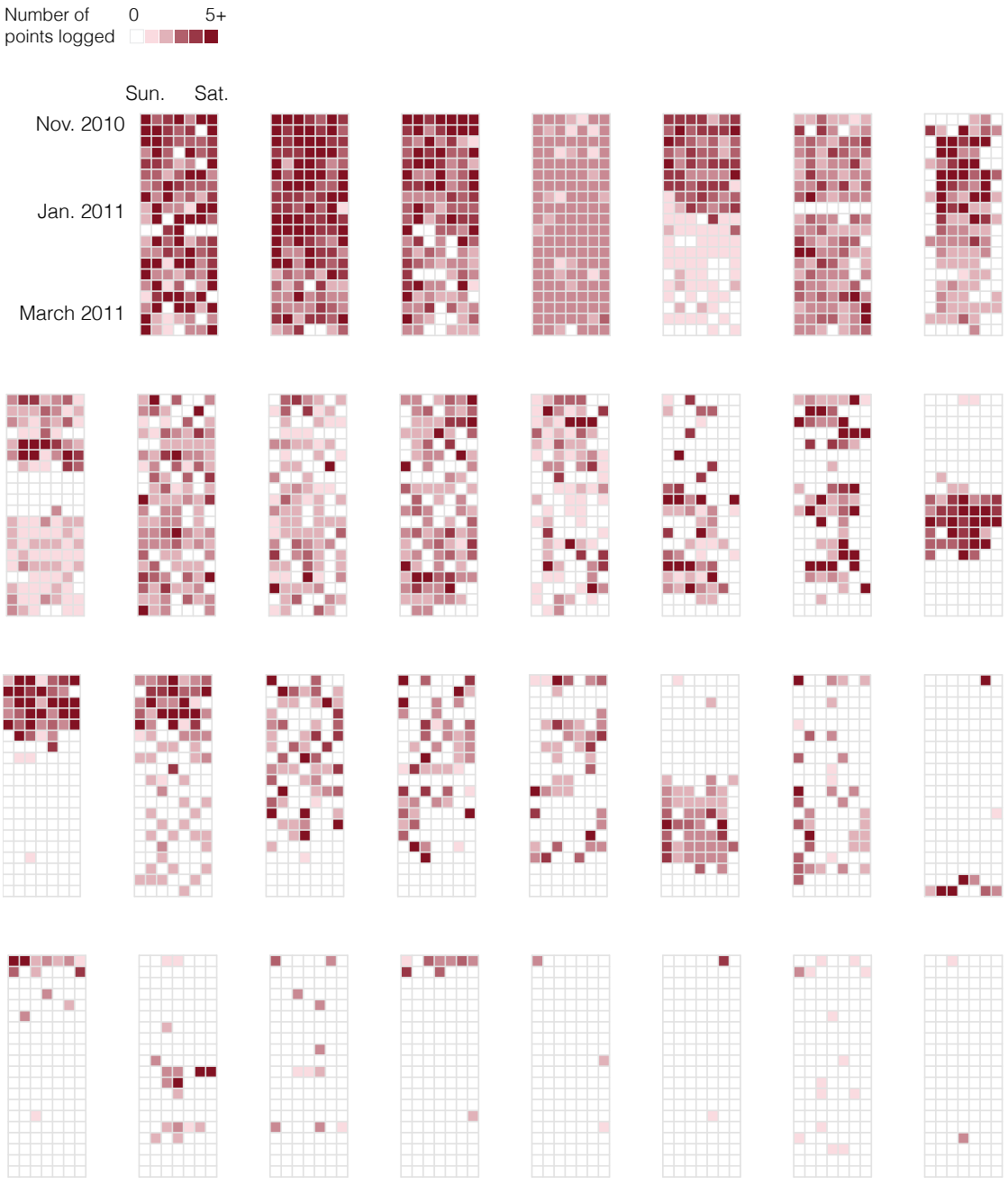


Figure 5.20: Collection levels over time

levels of interaction after collecting more data. User C collected data almost every day and interacted on the site more heavily on weekends, while User F collected data similarly, but interacted more often. The takeaway perhaps is that visualizations should be useful for both sparse and frequent data collection, and needs will change depending on the user. The needs of an individual can also change over time.

The ratio of interaction to collection levels can be seen more clearly in Figure 5.22, which plots days of logging data versus days visiting the site. Interaction tended to go up with collection. A lot of users spent equal days with interaction and collection, but there are also many who spent more days collecting than with the site. There are three users shown that have zero days of logging but a high number of days visiting the site, which does not make sense because if they did not log data, there is not much to see on the site. Most likely these users were actually bots spidering the site. It is also possible that they collected a lot of data before the interaction logging began, and they only came to the site to download their data or explore it with tools. However, this does not seem probable due to the high number of site days.

There appeared to be a gap in the 0.6 site days per logging day area, for users who collected data for more than a few days. This is more obvious in Figure 5.23, which shows the distribution of site days per logging days for users who collected data during at least 15 days (the area on Figure 5.22 where a split seems to start). The distribution (not shown) for users who used the site at least five days, the top top half from the previous Figure 5.17, showed a similar split but fewer users in the low end of site days per logging days. Users seem to either gravitate towards equal number of site and data logging days, or they favor data logging over exploration. Somewhat surprisingly, even though there was this split, the distribution of tools used, by both number of days and by average number of loads per usage day, was not significantly different between the two groups nor was the amount of data

Figure 5.21: Collection and interaction examples

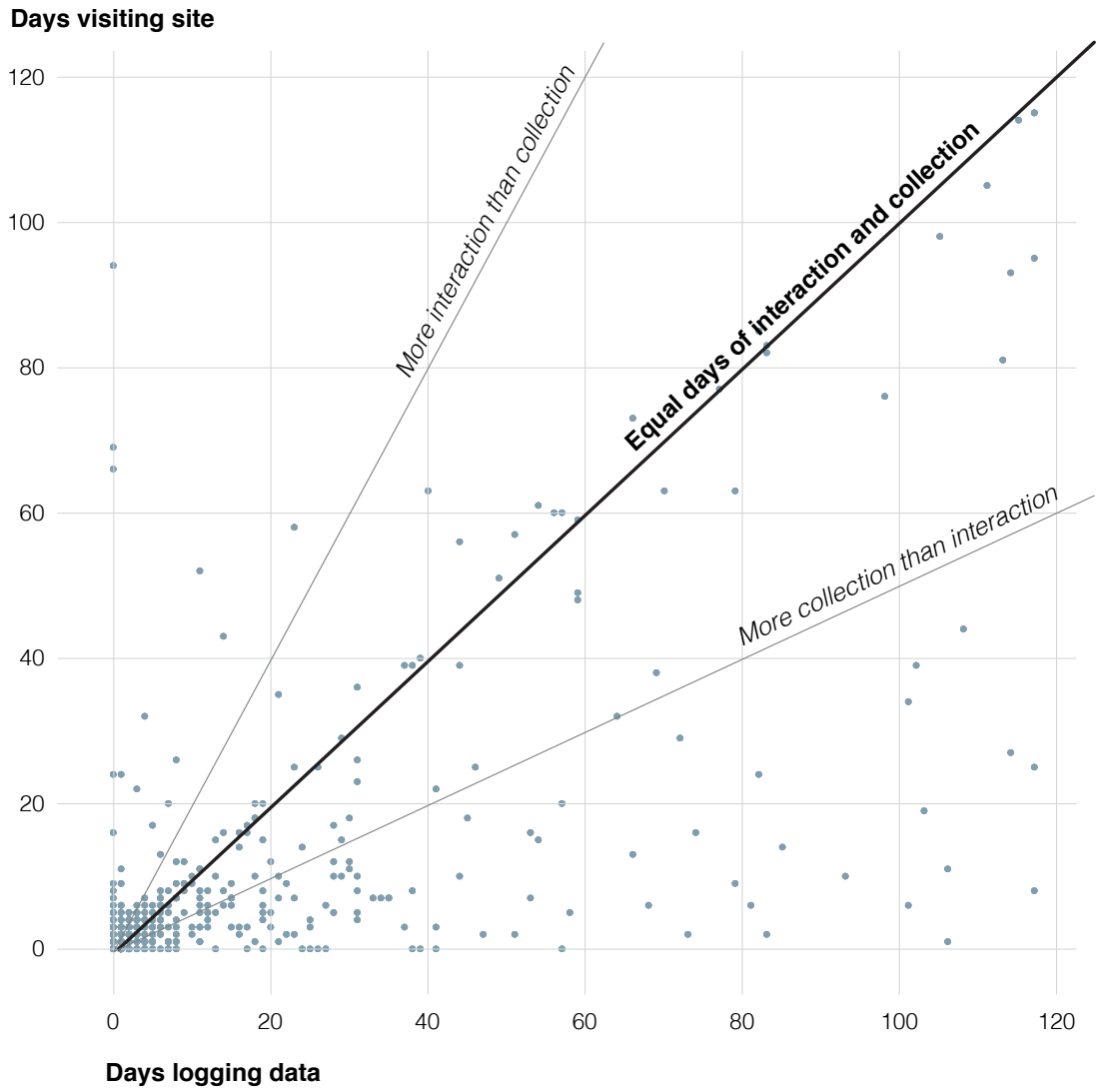


Figure 5.22: Usage patterns for data collection and interaction

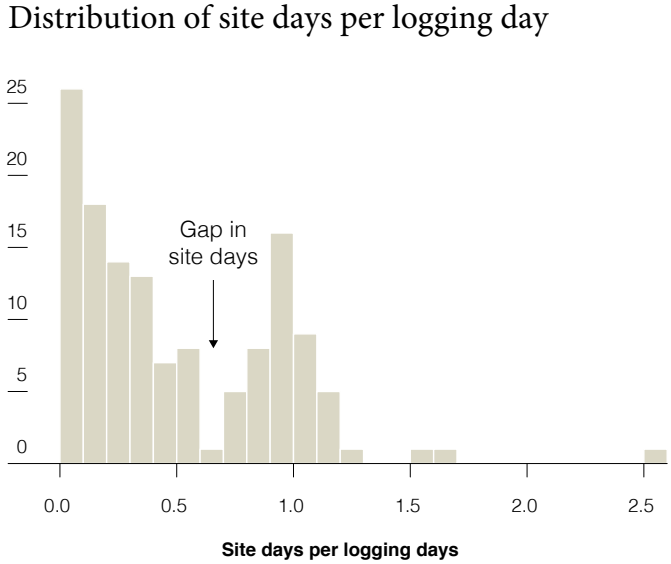


Figure 5.23: Distribution of site days per logging day

logged. The time span of usage for both groups was also similar. One group just logs in to the site more often than the other, but the depth of usage does not seem to be affected. From a user engagement perspective, this rate difference might be useful for reminders, although more testing would be required to verify this.

5.6.2.2 Usage Cycle

As mentioned earlier, users tended to move through different states of data collection, casual browsing on the site, and more in depth exploration with the interactive tools. The interaction logs are granular enough to look at this back-and-forth process. To get a sense of usage order, I looked at what views users visited before and after collecting data, during a single session. A session was defined as a series of page loads and data collection without a gap of more than 30 minutes in between any two actions. This cutoff was defined by reporting in Google Analytics that 98 percent of site visits lasted 30 minutes or less.

As shown in Figure 5.24, before collecting data, users were most often on

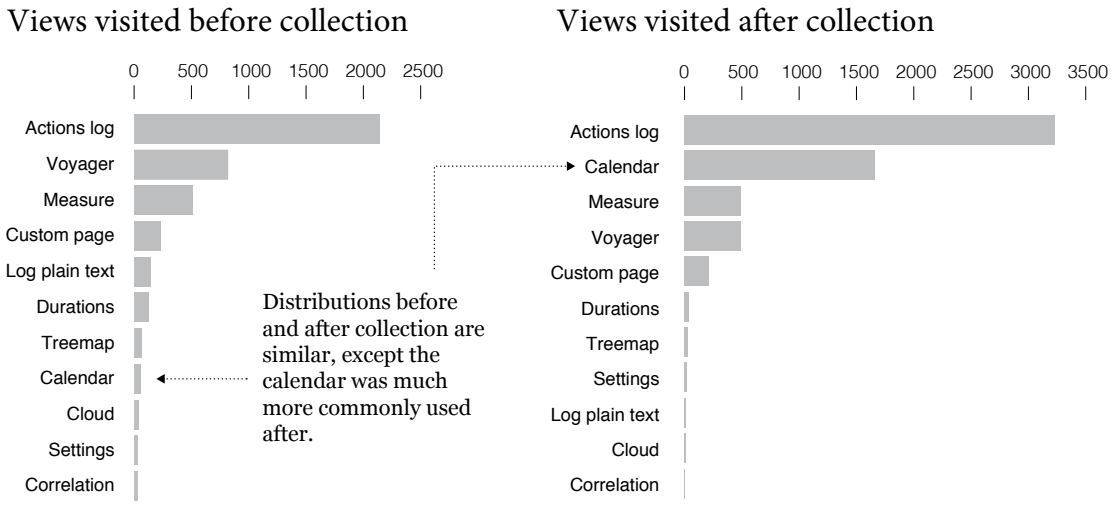


Figure 5.24: Views visited before and after logging data

the homepage or using the actions log to view and edit individual data points. The voyager was the most common exploratory tool used before logging more data. The individual action view for a measurement data type followed and then custom page view. Post-collection, the homepage and actions log were also the most common views. However, in contrast to pre-collection, the calendar and individual action views were more commonly visited post-collection, ahead of the voyager. Pre-collection the calendar views were much farther down the list. This seems to suggest that users were inclined to collect data after deeper exploration and then moved to more casual views to look for any small changes after new data was logged.

Figure 5.25 shows the collection and browsing cycle as a whole for all users. As might be expected, most sessions were for logging data, followed by casual browsing, and then deeper exploration. Casual browsing includes the homepage and actions log, whereas the analysis view group includes the visualization tools, which are more interactive. It was rare that a user went straight to analysis or made a jump from data collection to analysis.

Using model-based clustering, as described by Cadez et al. (2003), usage habits

Data Logging and Site Usage Flow

Below shows how users went back and forth between their data and exploration. Casual views include the personal homepage and the actions log, and the analysis views include the interactive tools.

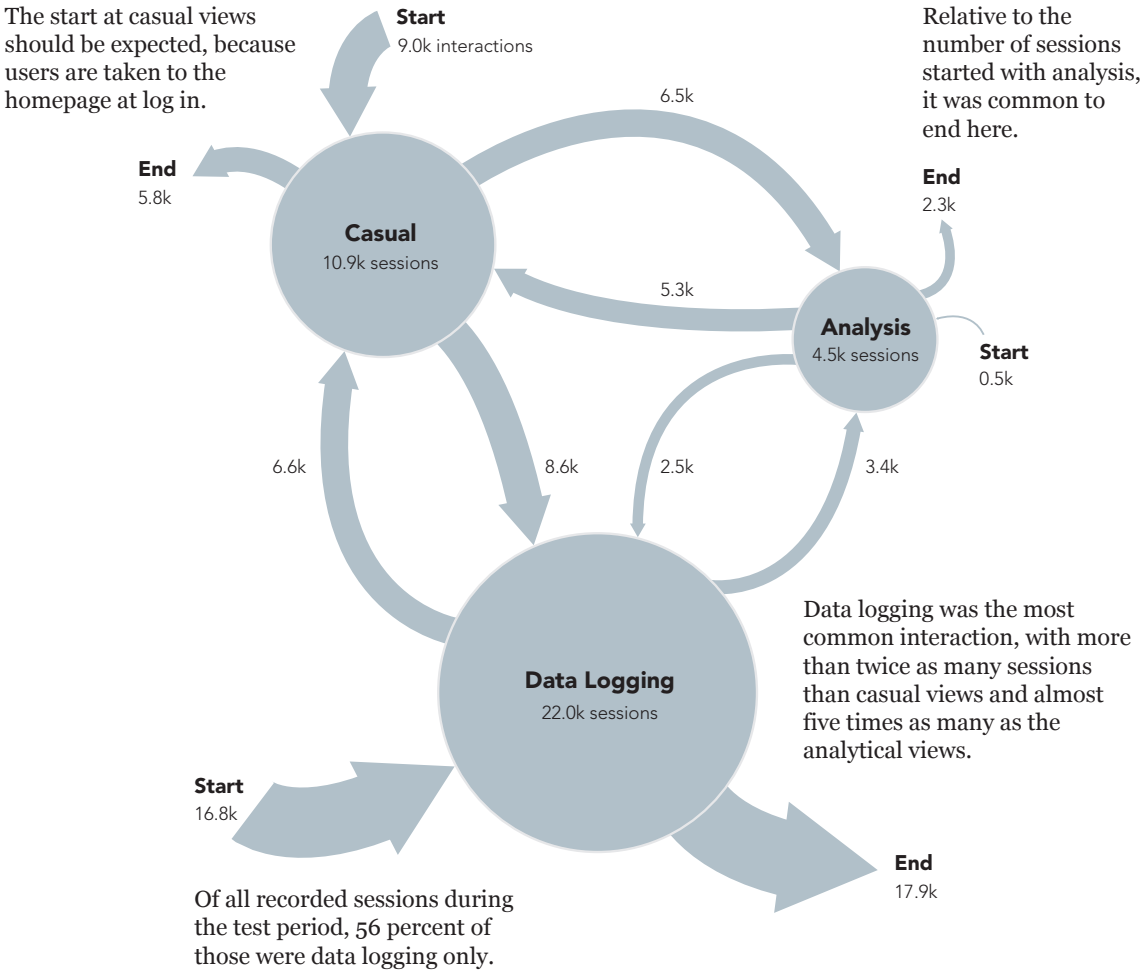


Figure 5.25: Data logging and site usage flow

can be seen in greater detail, per-interaction. Interactions on YFD were categorized as data logging via Twitter and the site, browsing (user homepage, actions log), single actions views, and analysis, such as the calendar heat map and stacked area chart. Sessions are still defined as any series of interactions without a gap of greater than 30 minutes. If a session of length L interactions is denoted as $\mathbf{x} = (x_1, ..., x_L)$ and x_i is one of the four interaction categories, $\mathbf{x}$ can be generated as a mixture of first-order Markov models:

$$\begin{aligned} p(\mathbf{x}) &= \textstyle\sum_{k=1}^{K} p(\mathbf{x} \mid c_k) p(c_k) \\ p(\mathbf{x} \mid c_k) &= p(x_1 \mid c_k) \textstyle\prod_{i=2}^{L} p(x_i \mid x_{i-1}, c_k) \end{aligned} \tag{5.1}$$

Each session is assigned to one of K clusters, c_k, using the EM algorithm. See Appendix A for code.

Figures 5.26 and 5.27 shows clustered sessions with a K of 20. Each cell represents a cluster, where each row of small squares represents a session, and each square in a session represents an action such as logging data or visiting a single action view. Each cell contains a random sample from the sessions in the respective cluster. Unlike Figure 5.25, the single action view is given its own category rather than aggregated into the analysis category.

Data logging in short bursts was the most common interaction, with longer data logging sessions occasionally. This was followed by short visits to the site, intertwined with data collection. Longer sessions were more rare, but typically involved deeper interaction with the data, either via the visualization tools, or the actions log while editing and deleting data.

Estimated transition matrices are also shown below each random sample. A fourth transition state was included to indicate the end of a session. Again, it was relatively rare to transition from data logging to analysis and more likely that users transitioned to analysis from interaction with other parts of the site. However, once users did land on an exploratory visualization, they tended to interact with the site more often than when just browsing, and it was more common to go back

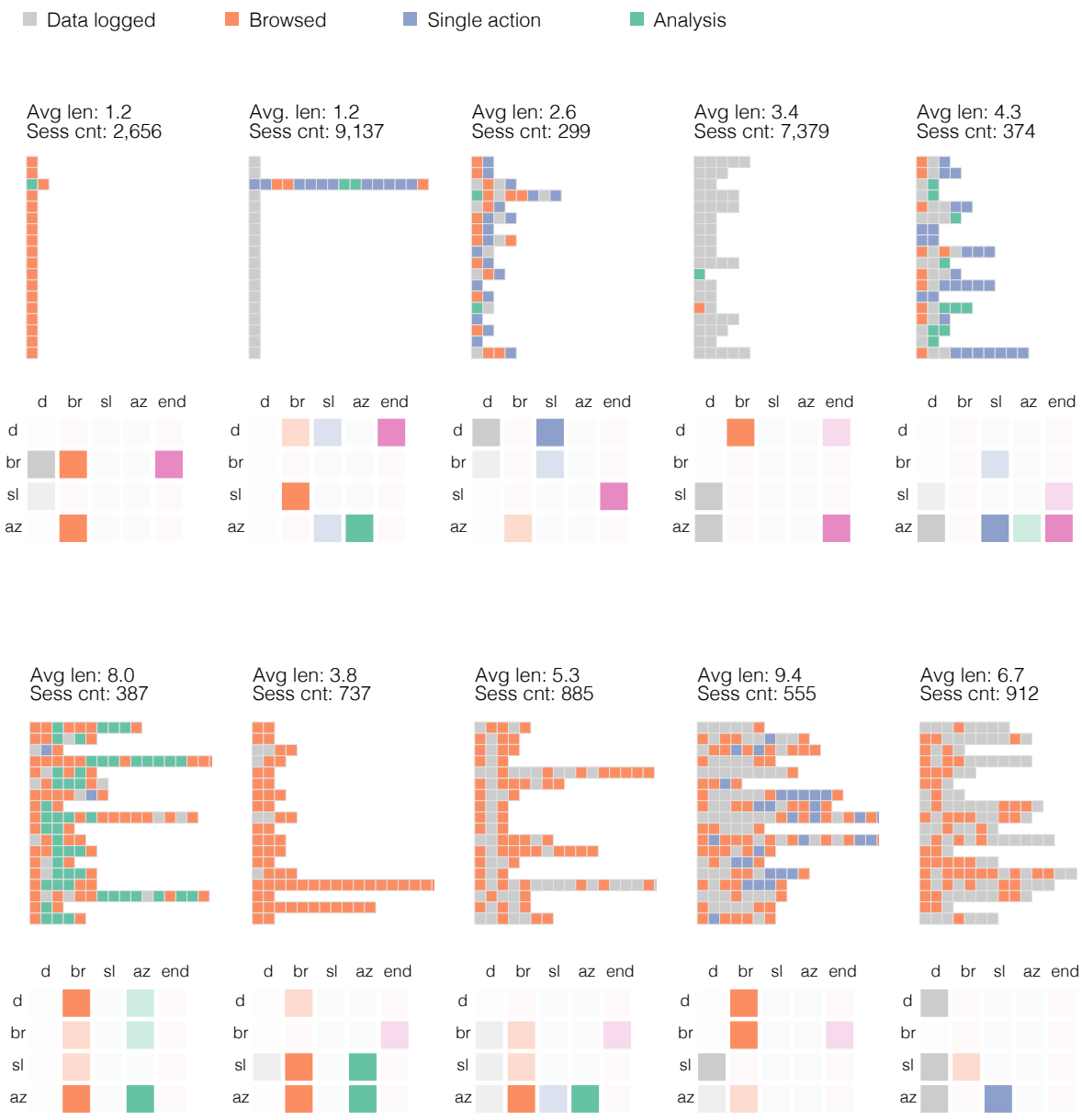


Figure 5.26: Sessions clustered using EM algorithm, 1 through 10

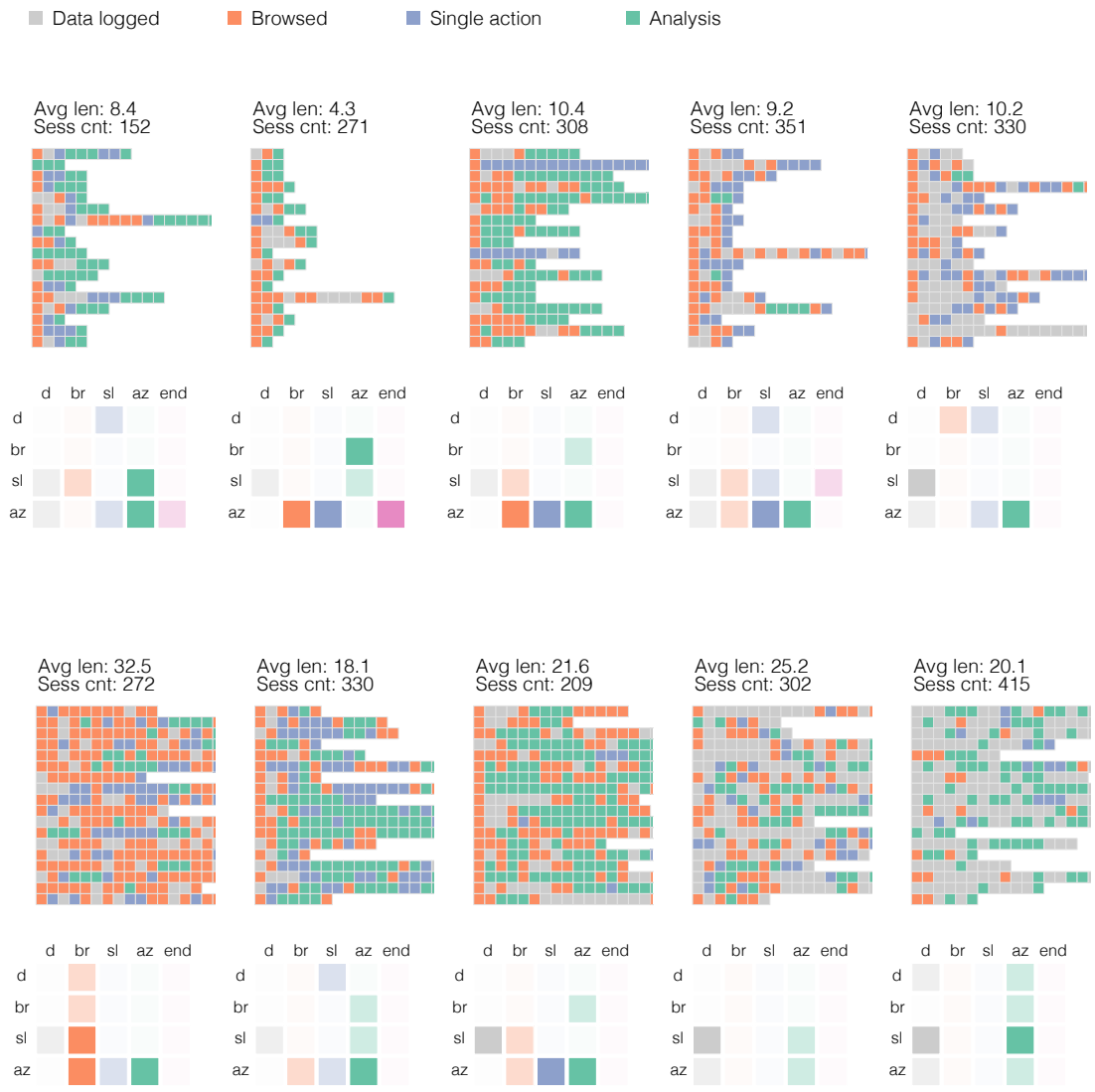


Figure 5.27: Sessions clustered using EM algorithm, 11 through 20

to one of the other interaction categories than it was to end in browsing. So it seems activity in each category encourages interaction in the others rather than users staying with one or the other. See Appendix A for simulated sessions of each cluster.

5.6.3 Discussion

The amount of time among data logging, browsing, and analysis suggests that it is worthwhile to spend more time designing visualization for casual viewing, as described in Pousman et al. (2007). Users spent the majority of their time with simpler tools. Most statistical visualization research focuses on generalized tools to explore complex patterns that data experts know to look for. Work on specialized tools that provide simple and quick insights might not be as complex, but can lead to more complex findings as users remain engaged and are encouraged to look deeper.

The order in which visualization and collection occurred is also important. In this case, the voyager was the most common view, other than the actions log and the homepage, before users collected data. The individual action view for the measurement data type was also popular. This view is less interactive, with just the ability to change time frames, but it is different from the other single action views in that it shows a fitted line on a scatter plot rather than a calendar. The overall calendar was farther down the list for views used before logging data; however, it was the most popular interactive view used after logging data.

One possible reason for the varied calendar popularity is that users found this view helpful in seeing if their logged data was entered correctly, as found by Lee and Dey (2011). In a two-person test group, users immediately looked for mistakes and anomalies when presented with a visualization of their data. This is most likely also why the actions log and homepage were most popular, more so

than they were before users logged data. Another possible reason is that logging data was similar to entering events on a calendar.

In any case, users seemed to go from exploration to collection and then to casual, and it could be useful to consider this flow in design of personal data applications. This is similar to the stage-based model proposed by Li et al. (2010), in which five stages are defined: preparation, collection, integration, reflection, and action. Due to the nature of YFD usage data, it is difficult to assess action, however, changes in sleep habit or weight loss described by survey participants suggested potential in this area. With YFD, the amount of time spent collecting and interacting with data, as well as level of reflection, varied by user, but the stages align, and there is clearly iteration between all stages.

Usage patterns between data logging, browsing, and analysis also seem similar to the the insight levels defined in the survey section (5.5.1) of this chapter. Those who took the opt-in survey described mostly elementary and intermediate insights, and a smaller proportion showed overall insight. Similarly, users spent more of their time logging data and browsing casually rather than with deep examinations. However, while this seems relationship between insight and interaction seems reasonable, I would need to administer a survey with more participants and a direct question about insight to make a better judgement.

Comparing YFD usage to that of other applications could provide more context to what has been discussed so far; however, the challenge of comparison to other studies, such as Van Kleek and Karger (2009), Elsmore et al. (2010), and Gemmell et al. (2006), is that most evaluated collection and interaction based on qualitative surveys or had relatively small user populations that were within a lab. YFD was publicly available almost from the beginning and I offered no incentives, other than the chance to use the application. Many studies also did not last very long, so there is limited opportunity to compare long-term patterns. As shown in Figure 5.17, usage and tool preference can change significantly just after two

weeks. Comparisons to commercial applications such as Mint by Patzer (2010) and RescueTime by Hruska et al. (2010), which let users record data automatically (finance and computer usage, respectively), would have been informative, but they understandably do not make user interaction data public.

Nevertheless, some comparisons are possible. For example, a similar chart to Figure 5.22 was also made to show PEIR usage in Mun et al. (2009). The PEIR chart only shows 17 users and about half of them had less than ten days of usage, but most users uploaded more days than they visited the UI. In contrast, while there were many YFD users who were collection dominant, there were also many who spent equal days with collection and interaction. PEIR was for the most part a single map view that let users interact with their data. There was also a network page, but it was relatively basic, made up of lists. It is interesting to note though that PEIR, which allowed automatic data upload, appeared to have users less engaged with the UI than that of YFD. This was most likely related to the design of the visualization I made for PEIR. The map was more on the level of the voyager on YFD, interaction-wise and less casual than say, the YFD homepage or single action views. Again this comes with the caveat that PEIR was more of a proof of concept than an application for a wider audience.

Mycrocosm by Assogba and Donath (2009) is somewhat comparable to YFD in that it lets users manually enter data; however, the underlying goal is more about an experiment in communication with basic graphs than it is about self-reflection and exploration. So more charts are made in their entirety in a single sitting rather than maintaining a continuous flow of data. Assogba and Donath (2009) reported usage for the first ten days after the announcement of the Mycrocosm application. There were 2,980 data points entered by 235 users, which is about 13 per user. During the first ten days after the announcement of YFD 2.0, there were 6,181 data points logged by 689 users, which is about nine per user. Again, the higher rate for Mycrocosm during the first ten days is most likely a factor of purpose. A

single chart is produced to communicate a single idea, so Mycrocosm users might have felt the need to complete what they started. This need to "fill in holes" was seen with YFD 1.0, which is discussed in the next section. Unfortunately, there was no long-term usage data available for Mycrocosm to compare beyond ten days, nor any detailed information on interaction with the plots.

5.7 YFD 1.0

The first multiple-user version of YFD was limited in functionality relative to 2.0. The visualizations were not interactive, and users could only log a few data types, specified by design. However, by offering users access to this limited version, I was able to gain a better understanding of how to expand the application.

5.7.1 Setup

This version was made available to about 100 users, and was online for seven months, from December 2008 through June 2009. In the middle of that time frame, I updated the application to allow a few other data types. Although the data types that could be collected was still limited. See Chapter **??** for more details.

Interaction logging was not implemented during this time, as the application began as a side project, so the only usage data available is the data that users logged and Web traffic data from Google Analytics. The latter is not especially useful, because much of the traffic was visitors that did not have YFD accounts, so for this brief usage study, I only looked at data collection.

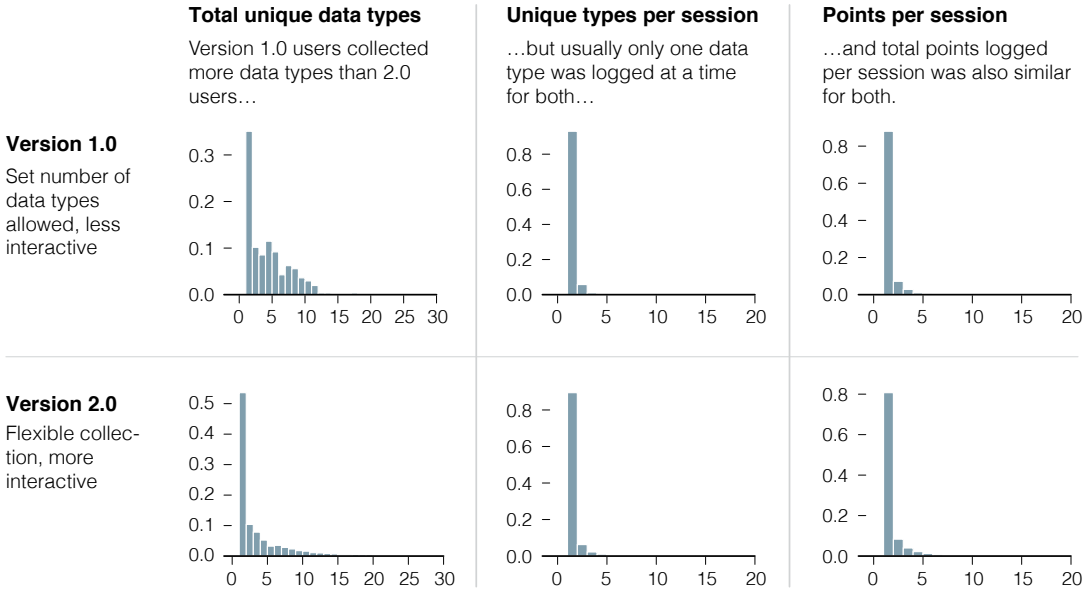


Figure 5.28: Unique data types, version 1.0 versus 2.0

5.7.2 Results

As discussed in Section 5.6.3 and shown in Figure 5.28, those who collected data with YFD 1.0 seemed to be more inclined to log the offered data types, such as eating, drinking, and sleeping. When users logged in to the site, they saw a dashboard with panels, which showed the most recent values of the respective data type.

If no data was logged for a data type yet, an empty spot or a dash was displayed. Users could reorganize the panels, but there was no way to remove them from the dashboard, which is why I believe users felt inclined to "fill in the holes" to make the view more complete. However, although YFD 1.0 users collected more unique data types as a whole, the distributions of data types and total data points logged *per session* look similar for both versions, with a slight edge to YFD 2.0 users. So after 1.0 users filled the gaps, they only continued to collect data for the metrics they were interested in.

5.8 Conclusions

The YFD user base of several thousand people provided a unique view of how people collect and interact with personal data. Unlike previous studies, YFD was publicly available, its users ranged in data skill level, and there was no incentive or recruitment to use the application.

Due to the private nature of the data that users collected, I was not able to see unencrypted data types, but the survey provided anecdotal evidence of what data people collected and their reactions. Because of the flexibility of data collection, people collected a wide range of data about themselves and their surroundings. The relatively longer-term usage of the application also provided users with insights beyond just becoming more aware of their actions. Data collection provided them with unexpected insights about their own behavior and of those around them.

Usage changed depending on whether a user was interested in self-experimentation or data journaling. The former used aggregate and pattern-finding visualization more often, whereas the latter seemed more focused on individual data points. Those with only a general interest found YFD much less useful, suggesting application-specific visualization and guidance on why one might collect personal data.

This was more obvious in the interaction logs, which showed users interacting with all the tools in the beginning (most likely to see what was available), but then settling in on more casual views, such as the actions log and individual actions views. Most interactions were short and quick, such as a single data point logged or a quick check-in on the site. The more interactive views, such as the explorer and correlations tool, were used less often, but during longer sessions.

Traditional statistical visualization, especially interactive ones, tend to follow a design of overview first, which highlights aggregates and patterns, and individual

points on demand, based on what is found in the aggregate. However, the YFD usage cycle tended to flow in the opposite direction, with most time spent with individual points and then a transition to analysis. This is not to say that it is not worth developing advanced tools for personal data, however. Rather there should be more efforts in researching how quick insight tools can be combined with deeper exploration tools, so that those with personal data can maintain an anecdotal connection but still approach their data analytically. Each provides context to the other, which potentially leads to better understanding.

CHAPTER 6

Conclusion and Future Work

This dissertation describes YFD, an application for personal data collection. Although there are many uses for personal data collection, there is still much to learn about how non-professionals interact with and analyze their data. Most related studies were short-term with a limited user base, which limits research scope. However, YFD is a publicly available application with a diverse user base, which allowed for detailed analysis of usage and opportunities to improve current and future applications.

YFD development began with a mechanism that allows users to collect data via Twitter. YFD syntax, described in Chapter 3, was designed to mimic how existing Twitter users update followers on what they are doing or what is happening around them. With an existing online culture for self-updates, YFD extends that usage to more detailed, personal data. Data collection syntax was originally restricted to specific data types, but it was later generalized so that users could collect the data they wanted to. Such flexibility allowed users to change their collection process and shift data types as their interests changed. Future work might employ natural language processing or user-defined syntax for even greater flexibility.

Once users collected data with YFD, it was clear that the application had to show immediate change for verification that the connection with Twitter worked and to provide a sense of progress. The first public version of YFD only updated once every thirty minutes (in attempt to place focus on long-term collection over individual points), and new users often emailed bug reports to notify me that

something was broken, when the site just had not updated yet. I received similar reports when the site was in maintenance mode for a code update. So the current version of YFD, although not in real-time, updates every three minutes. This is also useful for those who often edit their data in the actions log, as discussed in Chapter 5.

In Chapter 4, multiple views are described, which lets users focus on different dimensions of their data and can help in understanding aggregates and overall trends, in addition to point-to-point variations. The most recent data is shown on the user homepage, but a click on an action or an exploration option takes users to aggregate views, such as the calendar heat map or durations tool. Users can quickly switch between these views. Likewise, aggregate views link to single action views, which users can casually browse or explore in depth.

The link between browsing and analysis is important, because it can help users make inferences in their long-term and aggregate data that they would not be able to from the homepage or actions log. The casual views can produce awareness insight, whereas the aggregate views can help with reflection or analytical insight. As described in Chapter 5, users switch between the two in single sessions. It is more common for a browsing session to turn into an analysis than for a user to go straight into analysis. This is useful for those who design and develop personal data applications. The casual views can be helpful on their own, but the more straightforward interfaces can also inform people's exploration while using analytical views.

6.1 Data for a Wider Audience

As of this writing, most work with visualization in statistics assumes that users are data professionals, which is useful within the statistical community, but the audience for data has grown (and continues to), especially in areas outside of

statistics, such as computer science, design, and journalism. Visualization has evolved into more than a tool for analysis. Statistics should play a bigger role here — making data accessible to a wider, more general audience — but work in this area is often dismissed as "just making things pretty" because it does not appeal to the needs of data professionals. However, as discussed in Chapter 2, there are various types of insight other than analytic, and visualization can be and often is used for applications outside analysis. YFD is an obvious example.

The complementing insights might be more difficult to measure than accuracy and speed in graphical perception (Cleveland and McGill, 1984), but they should carry as much value, especially when trying to help non-professionals understand data. For example, skills carried over from various branches of design—such as graphic, interaction, and information—can help people relate to a dataset or better understand the context behind the numbers. It might be more difficult to measure how one relates to a dataset, but a connection with the data can encourage people to explore more critically and consider such things as what data represents, where it is from, uncertainty, and accuracy.

With personal data, context often comes attached for the individuals who collect the data, because by definition the data is about them and their behaviors. For example, those who keep track of what they eat might also remember how they felt before, during, and after a meal without actually tracking their emotions, or a spike in coffee consumption might be associated with a pending deadline. Visualization becomes a memory cue, and the quantified informs the unquantified. How do we incorporate that level of context and meta-information with more general types of data? In future work, I will explore this further.

Since 2007, I have run FlowingData (Yau, 2013b), a blog on visualization and statistics, which has provided perspective on how a wide audience reads data through various types of visualization. This past year, there were 7 million views by 4 million readers, with over a hundred thousand subscribers and followers.

Like YFD, the audience is mixed but most likely more so, coming from various backgrounds and with different levels of data literacy and expectations of what visualization is for. Many readers do not regularly work with data (nor do they plan to), whereas others are data professionals interested in learning visualization for both presentation and visual analysis. Some do not even know what visualization is. Rather, they see "pictures" and "graphics" based on numbers; however, based on comments, sharing, and experience, people are able to discern patterns and discussion typically revolves around the subject of a dataset rather than a visualization method. The most viewed and commented on blog posts are perhaps the best indicator for what readers are interested in and what I write about. Posts range from comical charts and graphics to straightforward maps to tutorials on how visualize data.

A common reaction within the statistical community to visualization that appears to be more "art" than tool is to dismiss it as inferior. Even when a graphic is shared and appreciated by hundreds of thousands of people, the inclination is to focus on everything that is "wrong" with the work. This is not to say that all work that gains mainstream attention is good, but we should examine why a visualization is so popular. For example, each year on FlowingData I choose my favorite visualization projects for that year based on use of data, aesthetics, and overall appeal, and in response to my picks from 2008, Gelman (2009) wrote that they "suck." Gelman and Unwin (2011) then expanded on the blog post and followed up on that in Gelman and Unwin (2012) and noted a misunderstanding of what information visualization research is and how it relates to mainstream graphics. The mistake was to judge the best-of picks as analysis tools, which require a certain level of efficiency and conciseness, rather than data presentations to a wide audience, which includes those who are not data professionals.

Visualization as a medium allows for a variety of applications in the same way a movie can be a documentary, action and adventure, drama, or comedy.

Although there are commonalities across movie genres, such as storytelling and cinematography, a romance drama is typically not judged in the same way one might judge a slapstick comedy. Similarly, visualization used in a comic should not be judged by the same criteria as one used for everyday analysis. Hall (2011) proposes a multidisciplinary approach to visualization critique, as work can span from scientific to artistic. Although visualization has roots dating back to the 17th century and has seen many milestones since the first statistical graphics (Friendly and Denis, 2001), this shift in visualization as a medium to not only analyze, but to browse, explore, and present in a variety of forms, has been more obvious in past years.

Technology improves and data is more ubiquitous, which provides greater opportunities to allow a general audience to interact with data and to understand their lives and surroundings from a new perspective. Usage of YFD was voluntary and the application was publicly available, which suggests an interest among individuals for data in the everyday. However, if there is any doubt, we can look to the adoption of wearable devices with fitness and wellness applications, such as Nike (2012) and Fitbit (2012), which are estimated to grow from 16.2 million in 2011 to 93 million in 2017 (Wang, 2012).

That said, as data grows more ubiquitous and people interact with it in their everyday, how should we design systems that allow people to make the most out of their data? The usage studies in this dissertation offer guidance, but it is only a start, and there is still much to improve on. For example, in 2011, there were issues with data privacy for Fitbit users. The online application allows users to log physical activity that is not measured by the device so that they can keep track of a more complete summary of calories burned each day. However, some unknowingly shared sexual activity entries, because Fitbit makes profiles public by default (Hill, 2011) and users were not aware that their data could be viewed. Had permissions been more obvious via the user interface, perhaps this would not

have been an issue.

Users of personal applications must also take ownership of the data they send to a service. If Fitbit users wanted to leave the service after such an incident, they could not download all of their data before closing their account. Only those with premium accounts can download their data, and even then they can only retrieve daily aggregates, as opposed to raw logs. It is a similar situation with Nike+. Users can retrieve data via an API, but most people will not go through the trouble. As statisticians, we know the value of data and what can be done with it, but most people do not, so the need to own one's data might be low in priority. On the other hand, there are types of data that people use every day or are more visual, such as online bookmarks or shared photographs, that might cause the same "emotional blow" that Gemmell et al. (2006) described.

If people can see the value of their data through visualization, then perhaps it will be easier for them to place more value on their data. This is what I see on FlowingData. People see a visualization project featured on the blog and then note how beautiful it is. The perception of beauty appears to transfer to the data the visualization represents, as people often remark, "Data *is* beautiful!" It is not just the appearance that readers refer to but how the data relates to the non-data world. The excitement leads to more questions, discussion, and exploration.

6.2 Future Work

While studying personal data collection and exploration through YFD, other projects were developed (refer to Appendix B), which influenced design of the application. Likewise, what I learned about how a general audience interacts with and understands data through YFD helped guide other projects. In future work, I will continue to explore these ideas. I believe that the best and perhaps only way to fully understand how a general audience interacts with personal data, or

any data for that matter, is to allow a general audience to do so and to evaluate reactions. FlowingData and consulting work with larger publications and organizations provide a medium for further exploration.

Expanding the YFD visualization toolset will also provide additional insights. Although I intend to develop additional tools and the application itself, as it is far from perfect, I plan to open source the application—the mechanisms for collection, visualization, and account management—so that others can tailor the application to their needs. As of this writing, user data is stored on the YFD server and is downloadable as delimited text files, however, I hope to develop an extendable, self-hosted application that gives people full ownership of their data. A sharing model for YFD similar to the one that of OpenPaths (2012) might also prove to be a valuable resource for researchers.

Finally, it is important to help those who develop applications, as well as those who use them, to further their understanding of data in a way that is useful, effective, and ethical. Other than developers of personal data applications, there are many people from related and new hybrid fields of study—data science (Loukides, 2010) and data journalism (Bradshaw, 2013), for example—interested in using data to make decisions and to inform. To this end, I wrote *Visualize This* (Yau, 2011), a book that provides practical examples on how to visualize and communicate with data, and *Data Points* (Yau, 2013a), a complementary book that focuses on exploring data visually. The former is used in courses and by practitioners and is available internationally in seven languages. The latter will be published a few months after this writing. Additionally, tutorials on FlowingData provide another resource to learn how to make use of data. These have been viewed over a million times.

6.3 Final Words

This dissertation describes the development and usage of YFD, an application for personal data collection. Although we should develop tools that help statisticians further understand various types of data, it is also our responsibility to help others understand data and make their own educated findings as data grows more commonplace. This requires interfaces designed for this new and growing audience, which might lead to an interest in and use of more advanced statistical tools and ultimately, a general public with greater data literacy.

APPENDIX A

Assigning Clusters

Clusters and transition probabilities were estimated using model-based clustering described by Cadez et al. (2003), using the EM algorithm. I used the following R code to cluster YFD usage data.

```
# Load data
series <- read.table('series.tsv', sep="\t", header=TRUE, as.is=TRUE)

# Randomize series for initial clusters
trainingSize <- round(length(series[,1]) / 2)
samp <- sample(1:length(series[,1]), trainingSize)
series.train <- series[samp,]
series.test <- series[-samp,]

allTransitions <- list()
clusters <- list()
clustCnt <- 20
startSize <- floor( length(series.train[,1]) / clustCnt )

# Intitialize transition matrices
for (clust in 1:clustCnt) {
    startIndex <- (clust - 1) * startSize + 1
    endIndex <- clust * startSize
```

```
    clusters[[clust]] <- startIndex:endIndex
    allTransitions[[clust]] <-
        findTransitions(series.train[startIndex:endIndex,])
}

# Run the EM algorithm
numIterations <- 0          # DEBUGGING
currLogLike <- findLogLike(allTransitions, clusters, series.train)
currDiff <- 10000   # Some big number
# for (i in 1:1) {
while (currDiff > 0.0001) {
    clusters <- vector("list", clustCnt)

    # Classify each session, based on transition matrices.
    for (j in 1:length(series.train[,1])) {
        currP <- 0
        for (clust in 1:clustCnt) {
            p <- findProb(series.train[j,2], allTransitions[[clust]])
            if (p > currP) {
                currClust <- clust
                currP <- p
            }
        }
        if (is.null(clusters[[currClust]])) {
            clusters[[currClust]] <- j
        } else {
            clusters[[currClust]] <- c(clusters[[currClust]], j)
        }
```

```
    }

    # Transition matrices, based on new clusters
    for (clust in 1:clustCnt) {
        if (length(clusters[[clust]]) > 0) {
            allTransitions[[clust]] <-
                findTransitions(series.train[ clusters[[clust]], ])
        }
    }
    newLogLike <- findLogLike(allTransitions, clusters, series.train)
    currDiff <- (currLogLike - newLogLike) / currLogLike
    currLogLike <- newLogLike
    numIterations <- numIterations + 1
}
```

The above uses the following helper functions to find log likelihoods and calculate transition matrices:

```
# Find log likelihood of series given clusters and transitions
findLogLike <- function(allTransitions, clusters, series) {
    total <- 0
    for (clust in 1:length(allTransitions)) {
        tran <- allTransitions[[clust]]
        currClusters <- clusters[[clust]]
        if (length(currClusters) > 0) {
            currSessions <- series[currClusters,2]
            f <- function(s) { findProb(s,tran) }
            currProbs <- sapply(currSessions, f)
            total <- total + sum(log2(currProbs))
```

```
        }
    }
    return(total)
}


# Output transition matrix
findTransitions <- function(series) {
    transitions <- matrix(rep(0, 20), 4, 5)
    colnames(transitions) <- c("analysis", "browse", "single",
        "datalog", "end")
    rownames(transitions) <- c("analysis", "browse", "single", "datalog")
    for (i in 1:length(series[,1])) {
        cats <- strsplit(series[i,]$session, ",")[[1]]
        if (length(cats) == 1) {
            transitions[cats[1], "end"] <- transitions[cats[1], "end"] + 1
        } else {
            for (j in 1:length(cats)) {
                if (j == length(cats)) {
                    transitions[cats[j], "end"] <-
                        transitions[cats[j], "end"] + 1
                } else {
                    transitions[cats[j], cats[j+1]] <-
                        transitions[cats[j], cats[j+1]] + 1
                }
            }
        } # @end else
    } # @end for
```

```
    # Convert counts to probability
    for (i in 1:length(transitions[,1])) {
        if (sum(transitions[i,]) >  0) {
            transitions[i,] <- transitions[i,] / sum(transitions[i,])
        }
    }
    return(transitions)
}


# Find probability of session, given transition prob matrix
findProb <- function(session, transition) {
    cats <- strsplit(session, ",")[[1]]
    prob <- 1
    if (length(cats) == 1) {
        prob <- transition[cats[1], "end"]
    } else {
        for (i in 1:length(cats)) {
            if (i == length(cats)) {
                prob <- prob * transition[cats[i], "end"]
            } else {
                prob <- prob * transition[cats[i], cats[i+1]]
            }
        }
    }
    return(prob)
}
```

Based on the estimated transition matrices in Section 5.6.2.2 and shown in Figures 5.26 and 5.27, simulated sessions are shown in Figures A.1 and A.2.

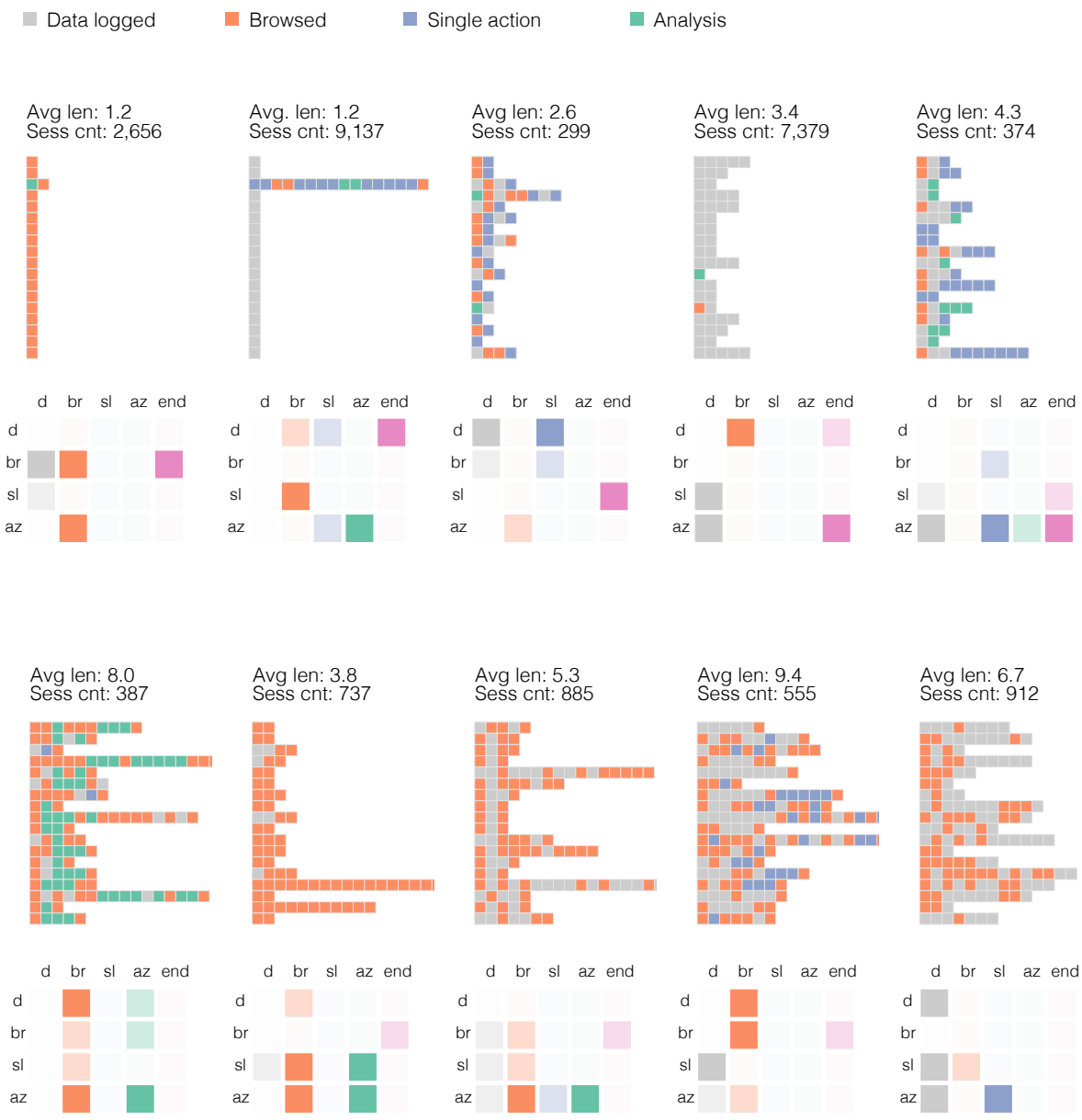


Figure A.1: Simulated sessions, 1 through 10

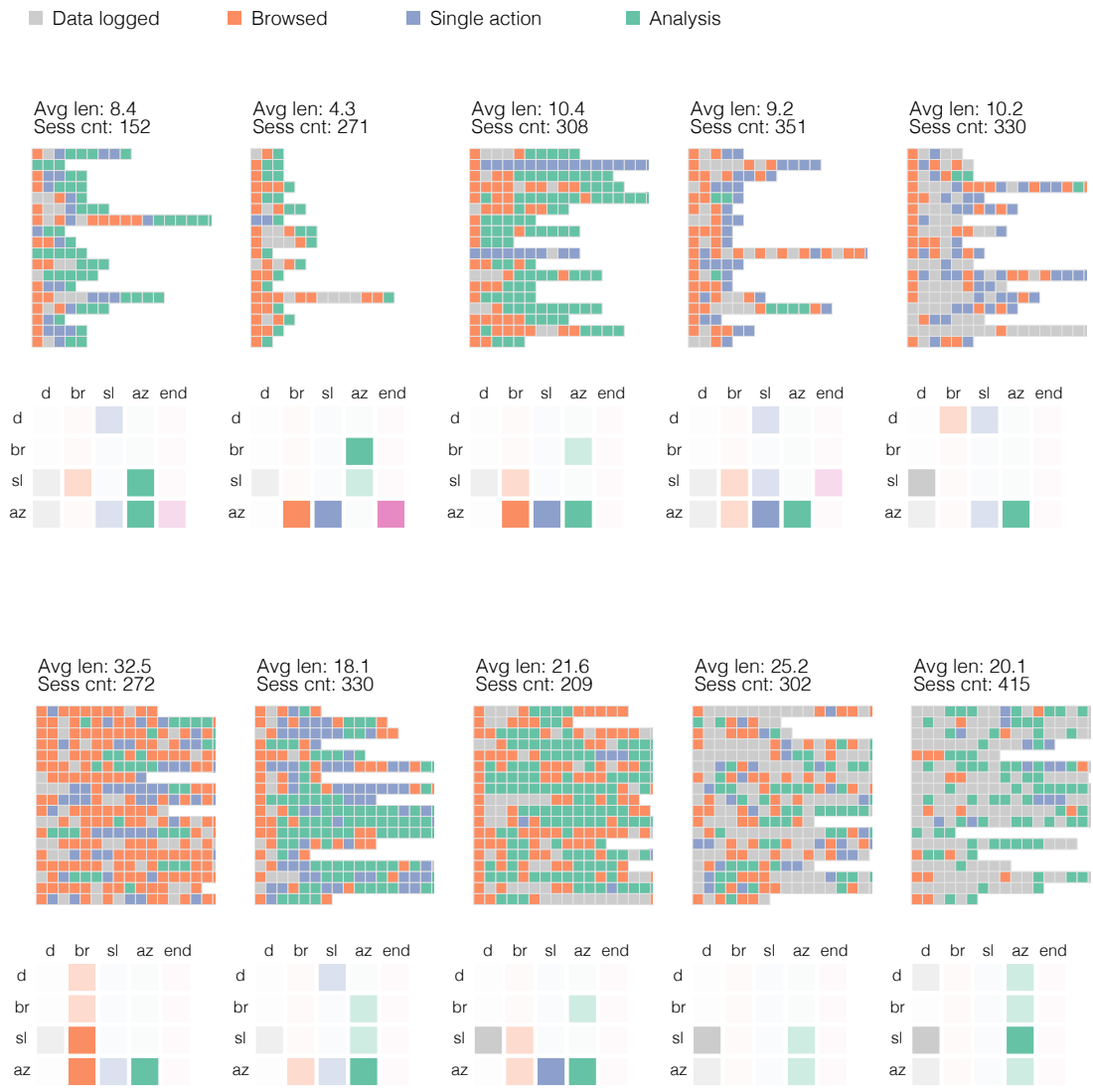


Figure A.2: Simulated sessions, 11 through 20

APPENDIX B

Reference Work

This is an appendix of projects that were developed before, during, and after YFD. The work in this appendix influenced work with YFD and experience with YFD influenced work on future projects. The work is split into two categories: personal data applications and general presentation.

B.1 Personal Data Applications

In some ways, YFD is the evolution of previous personal data projects. The work that follows represents what led to development of the Twitter-based application.

B.1.1 SensorBase

SensorBase (Chen et al., 2007) was a centralized repository to log sensor network data. The homepage is shown in Figure B.1 At the time of development, the Center for Embedded Network Sensing needed a place to store, retrieve, and share data from a variety of projects, such as those described in Burke et al. (2006). Data was originally sparse that existed mostly as flat text files in hundreds of folders and only accessible by the individual researchers who ran the projects.

We coined the term *slog*, which was a combination of "sensor" and "log" to reflect the spirit of information sharing represented by blogs. The SensorBase user interface was modeled after the user-friendly interfaces offered by popular blogging software such was Wordpress, which lets users publish, delete, and set

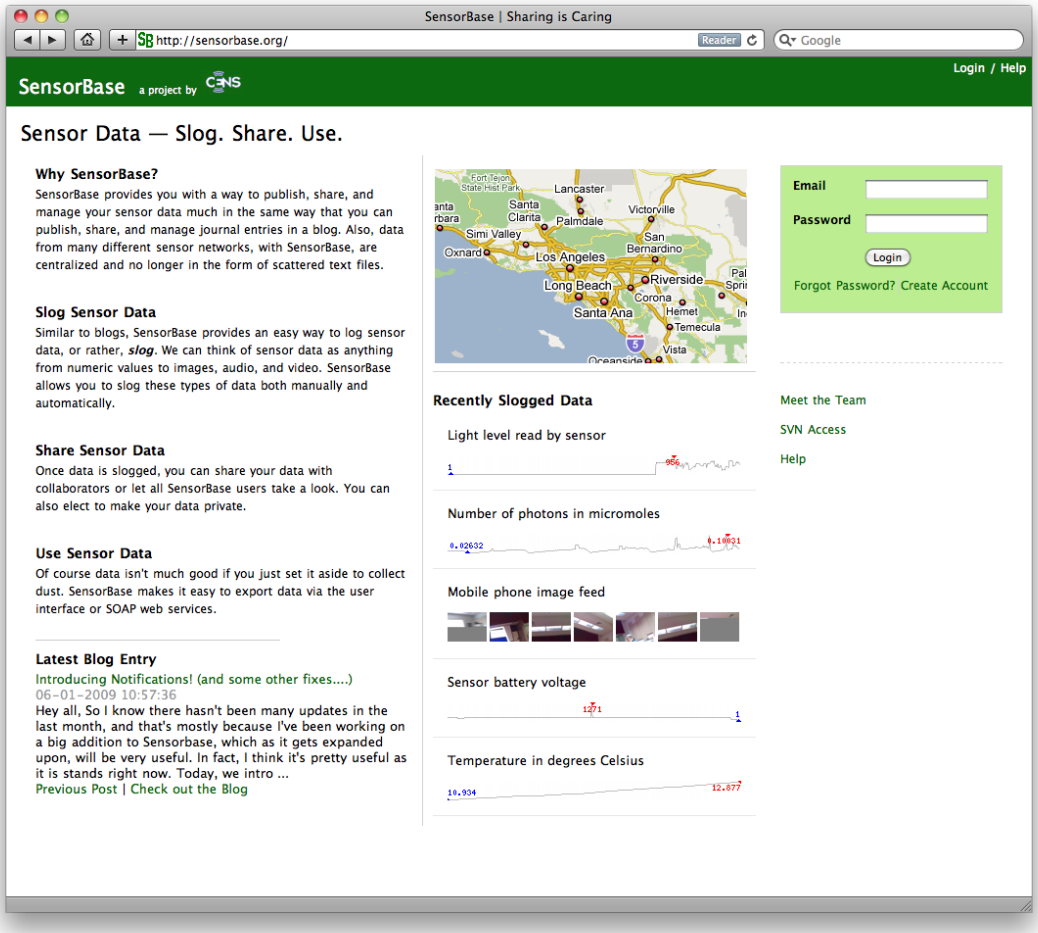


Figure B.1: Homepage of SensorBase as seen at http://sensorbase.org

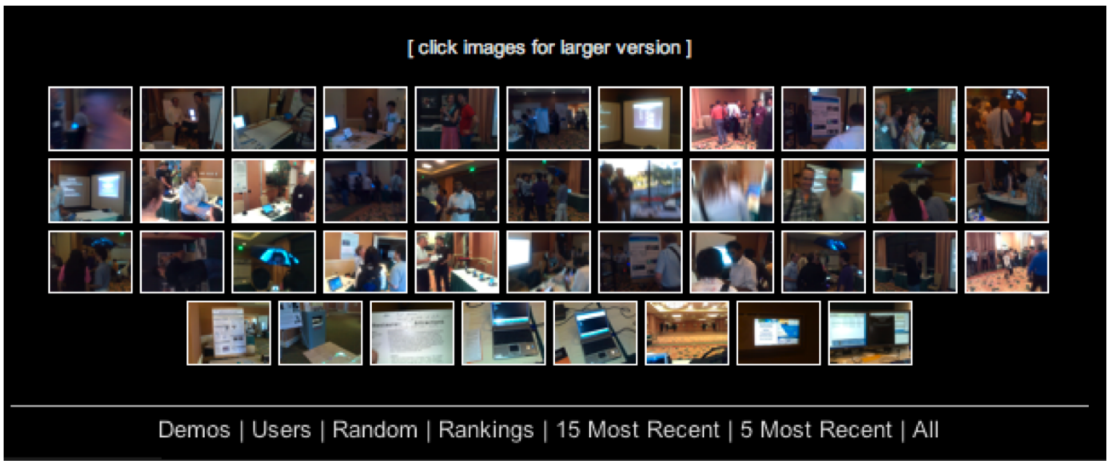


Figure B.2: SensorBase Application Demo

permissions on entries with little effort. Blogs also have RSS generated on the fly for easy notification and syndication.

The challenge was that data was heterogenous from a variety of sources, such as sensors embedded in the environment and repurposed mobile phones. Projects and applications also varied widely, so there were different demands for SensorBase. This led to a simplified design with an API that allowed others to retrieve data and build applications with SensorBase as the backend. Figure B.2 shows a demo application, which displays images from a mobile phone taken in increments automatically.

Visualization on SensorBase itself was minimal because of the variety of applications; however, small charts were included in some views to show the current status of a project. Like YFD, people often logged onto the system to see if the upload mechanism between devices and SensorBase worked.

B.1.2 Personal Environmental Impact Report

The Personal Environmental Impact Report (Mun et al., 2009), or PEIR, allowed users to log their location via mobile phone and estimate how their daily travel

choices affected the environment, in terms of carbon impact. They were also able to see estimates of how long they were exposed to high levels of particulate matter, which has been associated with increased rates of lung cancer, asthma development, and cardiovascular-related hospitalizations and mortality (Kim et al., 2004).

My goal was to visualize data so that users could explore their location and environmental estimates, as shown in Figure B.3. An interactive map on the bottom showed users their location traces colored by level of impact or exposure. Brighter green traces represent higher levels of impact or exposure, typically on highways. White dots represent idle time when the user was not traveling. The bars of color on the top represent "trips" that were estimated based on travel times and segmentation. Color corresponded to those on the map.

In concept work, I drew up a photo stream similar to that of the SensorBase demo but planned for an image browser that complemented the PEIR interactive map. The hope was to incorporate more context into a user's travels to more easily recall activity.

I also thought the use of a calendar heatmap for impact data might also be useful, as shown in Figure B.5. This was eventually used one of the main views for YFD.

B.1.3 Flowcal

Flowcal was a prototype I developed after working on YFD, as shown in Figure B.6. Taking the calendar format further, I was curious if it would be helpful to combine data from YFD, which is manually entered via Twitter, and data from existing data streams, such as Google Calendar and photo-sharing site Flickr. The interface was modeled after online calendars. Events displayed over the familiar grid layout, and pictures taken on corresponding days were also shown. Data

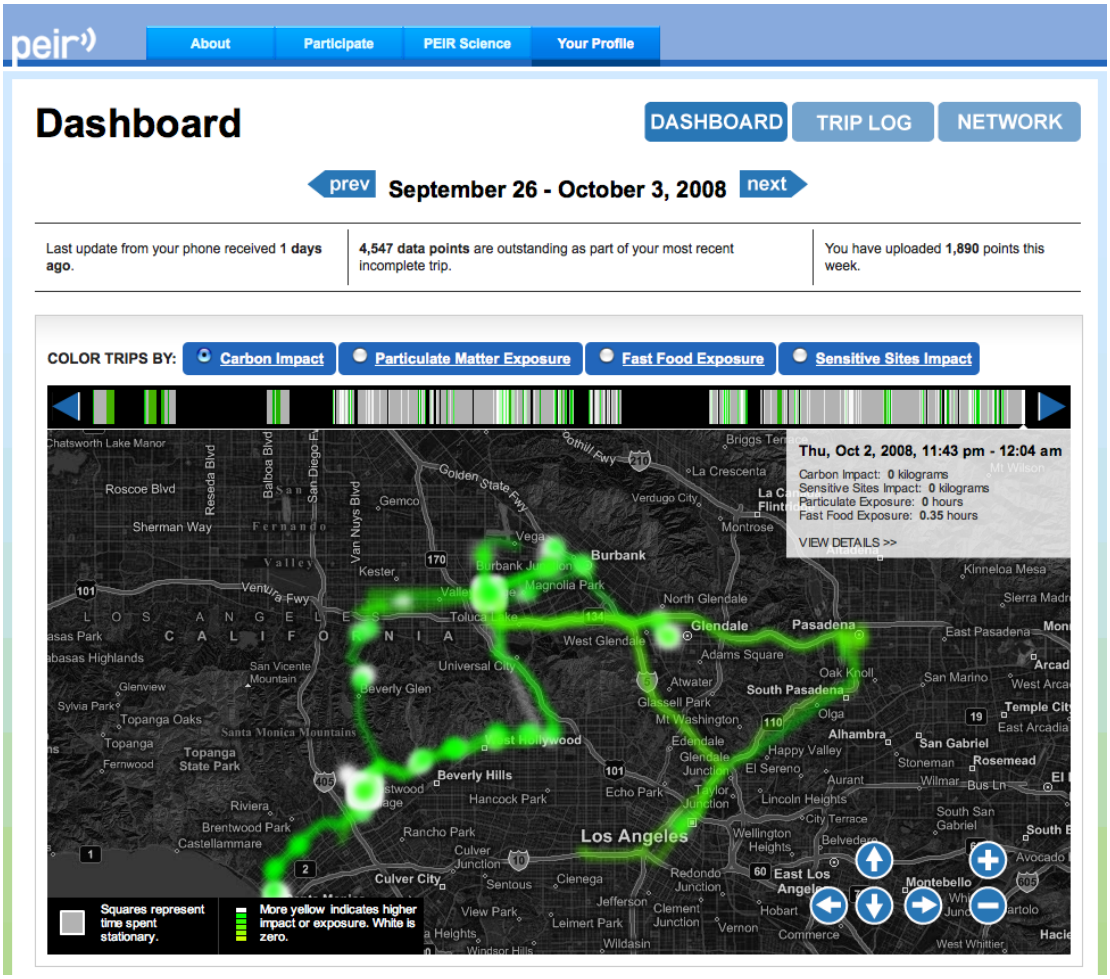


Figure B.3: PEIR Map Dashboard

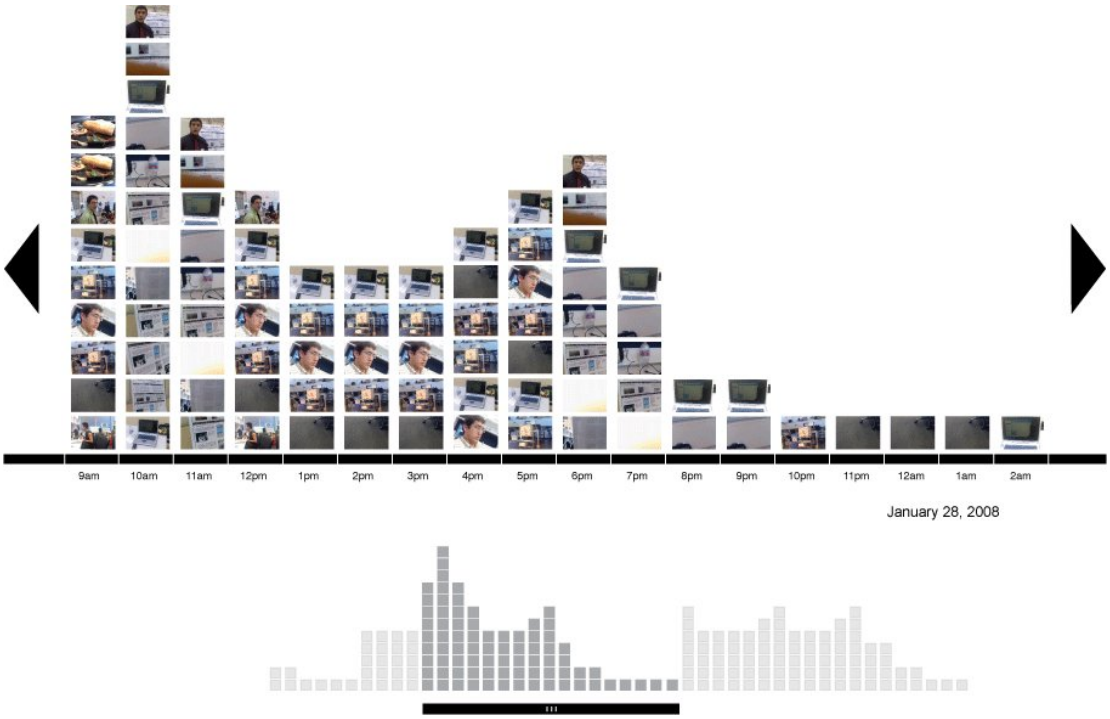


Figure B.4: PEIR Photo Timeline

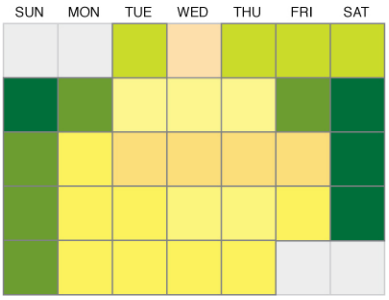


Figure B.5: PEIR Calendar Heat Map

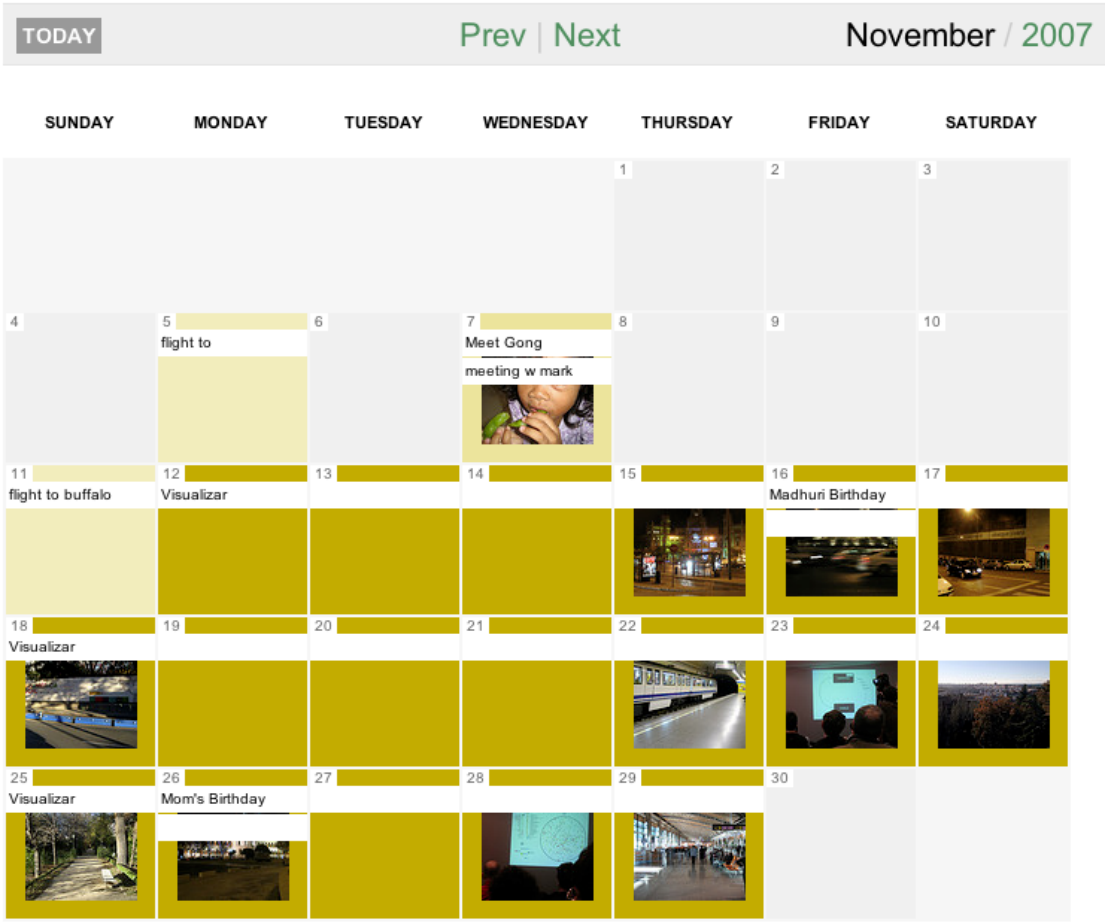


Figure B.6: Calendar Heat Map and Photos

logged on YFD and tweets on Twitter were also visible when a user selected a day.

As shown in Figure B.7, data could also be browsed by year. From personal experience, the interface made the data feel more journal-like and less analytical. I hope to extend this idea in future work.

B.2 General Presentation

Throughout development of YFD and personal data collection, I worked with magazines and newspapers to make data graphics for a wide audience via an

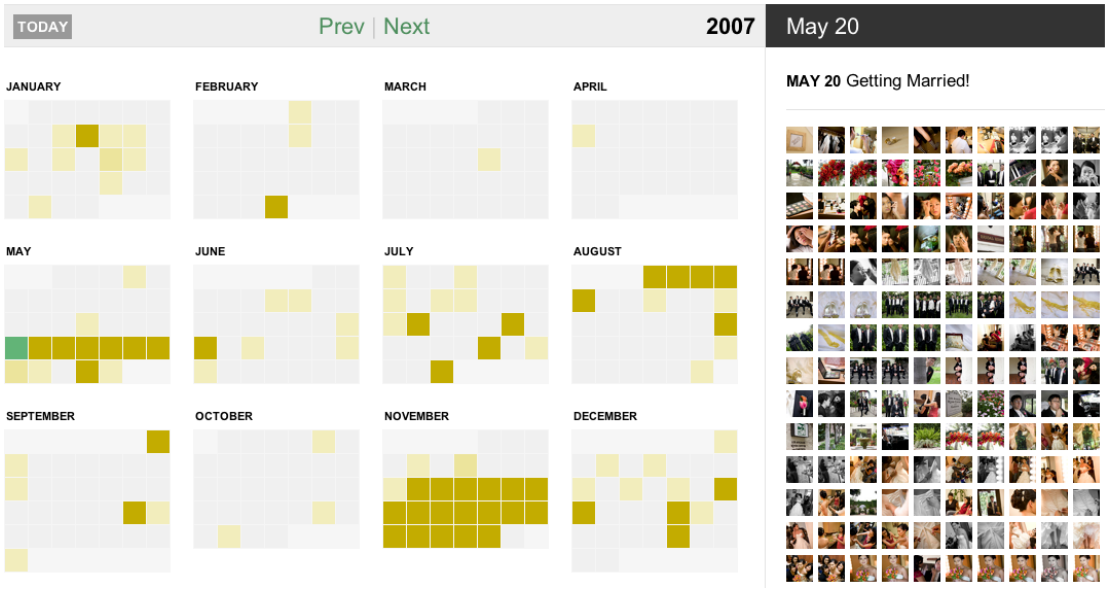


Figure B.7: Flowcal Year View

internship and as a consultant. I also did this on my own site, FlowingData. These are selected examples that allowed my studies to evolve to where I am now.

B.2.1 The New York Times

In an internship with The New York Times graphics department for a summer, graphics were designed for online and in print. The experience was significant, because before the internship, I only made statistical charts for analysis and occasionally copied them to reports. The audience was always smaller, and I paid little attention to layout, design, and journalism; however, at The New York Times graphics are used as a presentation tool and as a way to tell stories, when they are published to the newspaper. I made about 20 graphics during my time in New York.

For example, in 2007, Judge Michael B. Mukasey was appointed Attorney General of the United States. Leading up to the appointment, the graphic shown in Figure B.8 was created to compare Mukasey's rulings to other judges in the

New York Southern District. A series of three histograms was used to show distributions of rulings in different types of cases. Annotation on the graphic explains to readers how to read the distributions. Arrows point left and right from the median line indicating less and more strict, respectively. Mukasey's median sentence time was highlighted on each for easy comparison.

B.2.2 Humanflows

Humanflows, shown in Figure B.9, mapped worldwide migrations via an interactive map (Cabanzo et al., 2007). The goal behind the project was less an analytical exercise and more of a reflective one to put migration into perspective. We used migration data from the Migration Policy Institute with demographics, such as Gross Domestic Product and unemployment rates, in an effort to link the two.

The project was created during the two-week Visualizar workshop in 2007. It was a collaboration between myself and three graphic designers and was an exercise in learning how to work with people who do not speak the language of statistics. It was a challenge at first, but after the first week, we were more able to understand each other. There is conflict between people from different fields of study, but it often seems to stem from a difference in vocabulary more than a difference in ideals.

B.2.3 Animated Growth Maps

I have made various data graphics over the years (Yau, 2013c), but the most popular one, viewed over a million times and featured on the front page of a major news site, is the Wal-Mart growth map (Yau, 2010b). Shown in Figure B.10, the animated map shows store openings across the United States over time. Users can zoom in on a location of interest as the animation plays.

The growth appears organic as the rate of store openings is slow and cen-

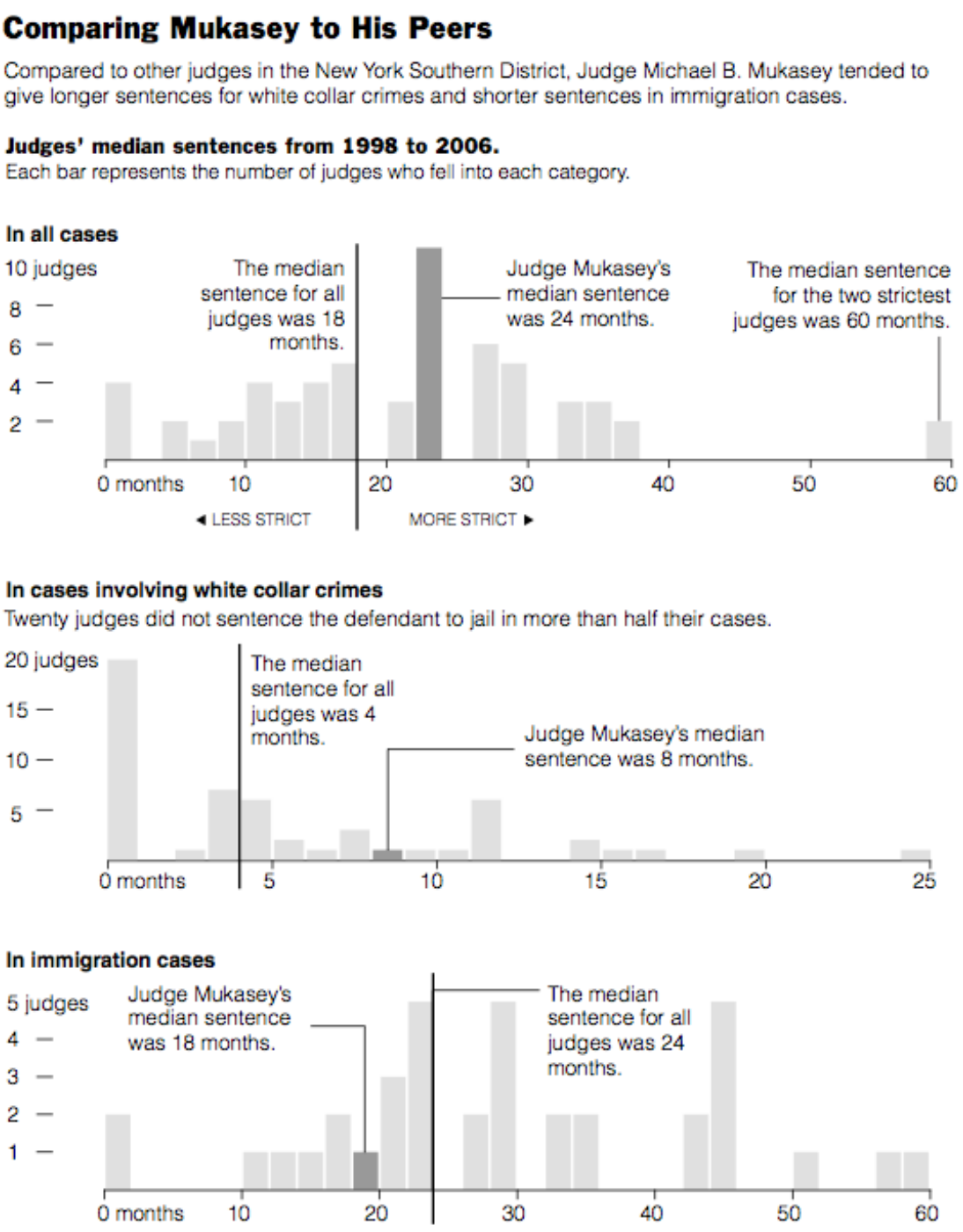


Figure B.8: Comparing Mukasey to His Peers

Figure B.9: Humanflows

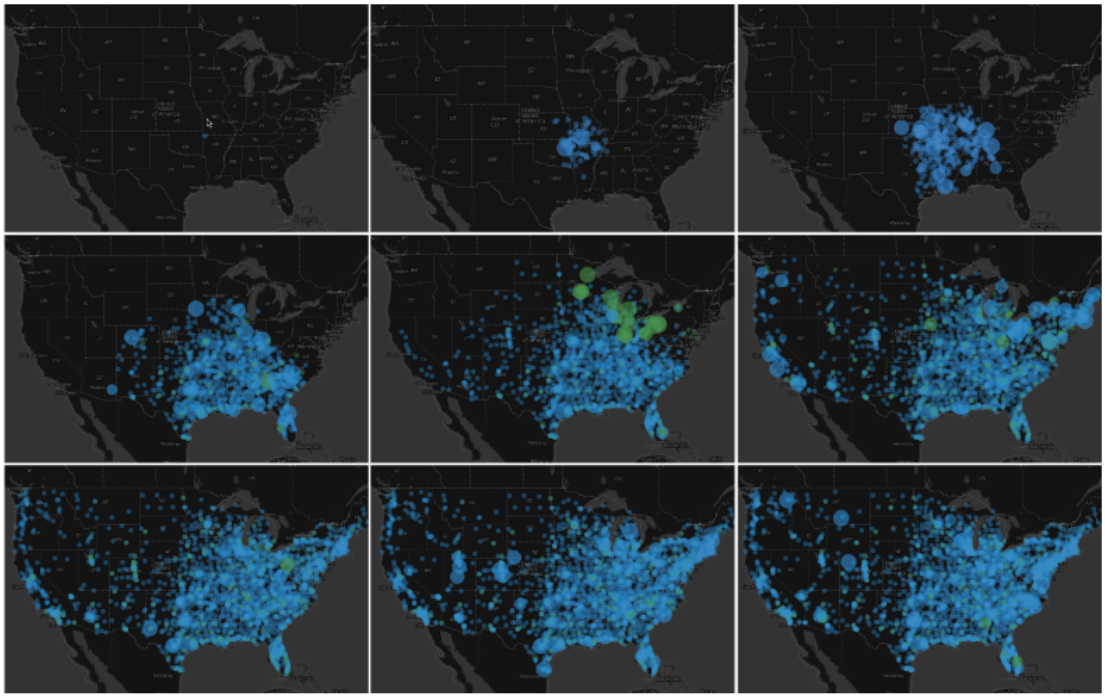


Figure B.10: Walmart Growth Map

tered in an area in the early years, but it quickly spreads outwards towards the coasts. Some readers noted that it looked like a "virus" or a "zombie apocalypse" spreading across the country. Others zoomed in to where they live and verified that a Wal-Mart did in fact open in a certain location, whereas some noted slight inaccuracies in the data. The data was originally downloaded from a personal research site, but data from an official Wal-Mart analyst now powers the map.

Because the map is available to view online, I do not know its exact uses, but I have received several permission requests from Wal-Mart corporate employees to use the map in presentations. Versions for retailers Target and Ross were also made at the request of analysts from the respective companies. I also open-sourced the code on FlowingData so that others can make their own maps.

B.2.4 World Progress Report

After much excitement over the release of world demographic data by the United Nations, about a year passed and not much had been done with the release. Progress: A Graphical Report on the State of the World is a series of graphics that was designed to provide insight into the United Nations data (Yau, 2009). This was a contrast to existing reports based on the data in that it was mostly visual. Existing reports at the time were mostly text accompanied by long and detailed spreadsheets. Although such a text-based layout can be useful for people who want to look up numbers for their specific country, it does not provide a very good overview of the rest of the world. Progress provides context to show how a country relates to others. The report focuses on mortality, population, energy, and environment.

This project was later developed as a print titled The World Progress Report, as shown in Figure B.11 (Yau, 2010a). All proceeds went to UNICEF towards earthquake relief efforts in Haiti.

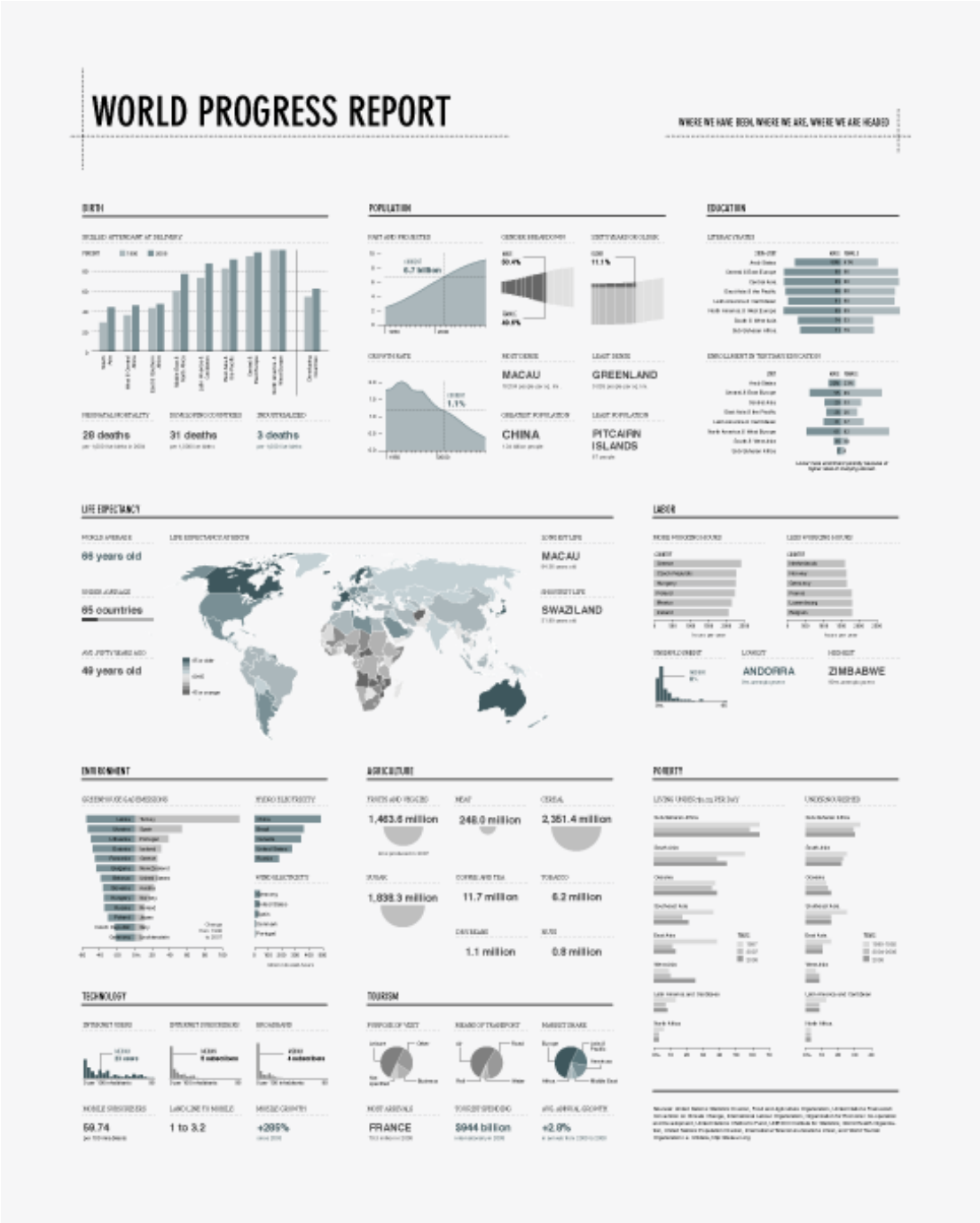


Figure B.11: World Progress Report

B.2.5 Data Underload

Data Underload is an exploration of using charts to tell stories. As visualization further develops into a medium, charts and graphs have become a way to tell jokes and communicate non-data concepts. Data Underload followed this concept as a bi-monthly web comic on FlowingData. Some graphics relate to personal curiosity while others are simply observations of the everyday, as shown in Figure B.12.

As described in this dissertation, much work has been done on how people read charts under the assumption that the only purpose is analytical insight, but not much has been done on how visualization is understood in this new form. This could be an interesting direction for future work.

Figure B.12: Data Underload #6 – Bed Head

Bibliography

Assogba Y and Donath J (2009). Mycrocosm: Visual Microblogging. In *hicss*, pages 1–10. IEEE Computer Society.

Audubon (2012). Christmas bird count. `http://birds.audubon.org`.

Bakker A and Hoffmann M (2005). Diagrammatic reasoning as the basis for developing concepts: A semiotic analysis of students' learning about statistical distribution. *Educational Studies in Mathematics*, 60(3):333–358.

Becker R and Cleveland W (1987). Brushing scatterplots. *Technometrics*, 29(2):127–142.

Bell G (2001). A personal digital store. *Communications of the ACM*, 44(1):91.

Bradshaw P (2013). What is data journalism? `http://datajournalismhandbook.org/1.0/en/introduction_0.html`.

Bright G and Friel S (1998). Graphical representations: Helping students interpret data. *Reflections on statistics: Learning, teaching, and assessment in grades K–12*, pages 63–88.

Bruls M, Huizing K, and Van Wijk J (2000). Squarified treemaps. In *Proceedings of the joint Eurographics and IEEE TCVG Symposium on Visualization*, pages 33–42. Citeseer.

Burke J, Estrin D, Hansen M, Parker A, Ramanathan N, Reddy S, and Srivastava M (2006). Participatory sensing.

Bush V (1945). As We May Think. *Atlantic Monthly*.

Byron L and Wattenberg M (2008). Stacked graphs–geometry & aesthetics. *Visualization and Computer Graphics, IEEE Transactions on*, 14(6):1245–1252.

Cabanzo M, Yau N, Moradi I, and Sanchez M (2007). humanflows. http://projects.flowingdata.com/humanflows.

Cadez I, Heckerman D, Meek C, Smyth P, and White S (2003). Model-based clustering and visualization of navigation patterns on a web site. *Data Mining and Knowledge Discovery*, 7(4):399–424.

Carmichael A and Reda D (2008). Curetogether. http://curetogether.com.

Case R and Felton N (2010). Daytum. http://daytum.com.

Castillo J (2009). yfduploader. http://www.joeycastillo.com/apps/yfduploader.

Chang K, Yau N, Hansen M, and Estrin D (2006). Sensorbase.org - a centralized repository to slog sensor network data.

Chen G, Yau N, Hansen M, and Estrin D (2007). Sharing sensor network data. *Info:*.

Cleveland W and McGill R (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):531–554.

CNN (2012). ireport. http://ireport.cnn.com/.

Consolvo S, McDonald D, and Landay J (2009). Theory-driven design strategies for technologies that support behavior change in everyday life. In *Proceedings of the 27th international conference on Human factors in computing systems*, pages 405–414. ACM.

Consortium WWW (1995). Hypertext markup lanuauge. http://wwww.w3.org.

Dorsey J (2008). Twitter trends and a tip. http://blog.twitter.com/2008/09/twitter-trends-tip.html.

Downer SR, Meara JG, and Costa ACD (2005). Use of sms text messaging to improve outpatient attendance.

Eich B (1995). Javascript.

Elsmore C, Wilson M, Jones M, and Eslambolchilar P (2010). Neighbourhood wattch-community based energy visualisation for the home. In *Nundge & Influence Through Mobile Devices workshop (NIMD)*.

Everyblock (2012). Everyblock. `http://everyblock.org`.

Feinberg J (2010). Wordle. In *Beautiful Visualization: Looking at Data through the Eyes of Experts*, pages 37–58. Oreilly & Associates Inc.

Felton N (2011). Annual feltron report 2011. `http://feltron.com/index.php?/content/2008_annual_report/`.

Fitbit (2012). Fitbit. `http://fitbit.com`.

Foodspotting (2012). Foodspotting. `http://foodspotting.com`.

Friel SN, Curcio FR, and Bright GW (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education*, 32(2):124–158.

Friendly M (1994). Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association*, 89(425):190–200.

Friendly M (1995). Conceptual and visual models for categorical data. *The American Statistician*, 49(2):153–160.

Friendly M and Denis D (2001). Milestones in the history of thematic cartography, statistical graphics, and data visualization. *Accessed: March*, 18:2010.

Gaglani M, Riggs M, Kamenicky C, and Glezen WP (2001). A computerized reminder strategy is effective for annual influenza immunization of children with asthma or reactive airway disease.

Gallup (2012). Gallup. "http://www.gallup.com/".

Gelman A (2009). Better late than never. http://andrewgelman.com/2009/04/better_late_tha/.

Gelman A and Unwin A (2011). Infovis and statistical graphics: Different goals, different looks.

Gelman A and Unwin A (2012). Tradeoffs in information graphics.

Gemmell J, Bell G, and Lueder R (2006). MyLifeBits: a personal database for everything. *Communications of the ACM*, 49(1):95.

Google (2010). Google calendar. http://calendar.google.com/.

Google (2011a). Google chrome: Dear sophie. http://www.youtube.com/watch?v=R4vkVHijdQk.

Google (2011b). Google person finder. http://google.org/personfinder/global/home.html.

Google (2012a). Google analytics. http://google.com/analytics.

Google (2012b). Google analytics support. http://support.google.com/analytics/.

Google (2012c). Google+: New dad. http://www.youtube.com/watch?v=8aCYZ3gXfy8.

Ha A (2012). Thanks to grubhub integration, foodspotting gets food ordering. http://techcrunch.com/2012/07/19/foodspotting-grubhub/.

Hall P (2011). Bubbles, lines and string: How visualization shapes society.

Hallnas L and Redstrom J (2001). Slow technology–designing for reflection. *Personal and Ubiquitous Computing*, 5(3):201–212.

Harris J (2011). Word clouds considered harmful. `http://www.niemanlab.org/2011/10/word-clouds-considered-harmful/`.

Heer J, Viégas F, and Wattenberg M (2007). Voyagers and voyeurs: supporting asynchronous collaborative information visualization. In *Conference on Human Factors in Computing Systems: Proceedings of the SIGCHI conference on Human factors in computing systems*, volume 28, pages 1029–1038.

Heywood B, Heywood J, and Cole J (2004). Patientslikeme. `http://patientslikeme.com`.

Hill K (2011). Fitbit moves quickly after users' sex stats exposed. `http://www.forbes.com/sites/kashmirhill/2011/07/05/fitbit-moves-quickly-after-users-sex-stats-exposed/`.

Holovaty A (2010). Django project. `http://djangoproject.org`.

Holovaty A (2011). Everyblock's first major redesign. `http://blog.everyblock.com/2011/mar/21/redesign/`.

Holt J (2005). Measure for Measure. *The New Yorker*.

Hruska J, Ficoa B, and Sacca C (2010). Rescuetime. `http://rescuetime.com`.

Hubble N (2010). *Mass Observation and Everyday Life*. Palgrave Macmillan, New York, NY.

Ishii H and Ullmer B (1997). Tangible bits: towards seamless interfaces between people, bits and atoms. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 234–241. ACM.

ITO (2010). Ito world at ted 2010 - project haiti. http://itoworld.blogspot.com/2010/02/ito-world-at-ted-2010-project-haiti.html.

Jawbone (2012). Jawbone. http://jawbone.com.

Kalina N (2010). Noah k. everyday. http://everyday.noahkalina.com/.

Kelly K (2012). About the quantified self. http://quantifiedself.com/about/.

Kim J et al. (2004). Ambient air pollution: health hazards to children. *Pediatrics*, 114(6):1699.

Kleinberg J (2003). Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397.

Koblin A (2006). The sheep market. http://www.thesheepmarket.com/.

Koblin A and Massey D (2009). Bicyle built for two thousand. http://www.bicyclebuiltfortwothousand.com/.

Koblin A, Milk C, Modern T, and Lab GC (2012). The exquisite forest. http://exquisiteforest.com/.

Krasner G, Pope S, et al. (1988). A description of the model-view-controller user interface paradigm in the smalltalk-80 system. *Journal of object oriented programming*, 1(3):26–49.

Lee M and Dey A (2011). Reflecting on pills and phone use: supporting awareness of functional abilities for older adults. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, pages 2095–2104. ACM.

Leong K, Chen W, Leong K, Mastura I, Mimi O, Sheikh M, Zailinawati A, Ng C, Phua K, and Teng C (2006). The use of text messaging to improve attendance in primary care: a randomized controlled trial. *Family practice*, 23(6):699.

Leskovec J, Backstrom L, and Kleinberg J (2009). Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 497–506. ACM.

Li I, Dey A, and Forlizzi J (2009). Grafitter: leveraging social media for self reflection. *Crossroads*, 16(2):12–13.

Li I, Dey A, and Forlizzi J (2010). A stage-based model of personal informatics systems. In *Proceedings of the 28th international conference on Human factors in computing systems*, pages 557–566. ACM.

LinkedIn (2011). Inmaps. `http://inmaps.linkedinlabs.com/`.

Loukides M (2010). What is data science? `http://radar.oreilly.com/2010/06/what-is-data-science.html`.

Lourenco V (2010). Foodfeed. `http://foodfeed.us/`.

MacNeill B (2010). Trixie tracker software. `http://trixietracker.com/`.

Mun M, Reddy S, Shilton K, Yau N, Burke J, Estrin D, Hansen M, Howard E, West R, and Boda P (2009). PEIR, the personal environmental impact report, as a platform for participatory sensing systems research. In *Proceedings of the 7th international conference on Mobile systems, applications, and services*, pages 55–68. ACM.

MyFitnessPal L (2012). Myfitnesspal. `http://myfitnesspal.com`.

Nike (2012). Nike plus. `http://nikeplus.nike.com`.

OpenPaths (2012). Openpaths. `https://openpaths.cc`.

OpenStreetMap (2010). Wikiproject haiti. `http://wiki.openstreetmap.org/wiki/WikiProject_Haiti`.

OpenStreetMap (2011). 2011 sendai earthquake and tsunami. http://wiki.openstreetmap.org/wiki/2011_Sendai_earthquake_and_tsunami.

Oracle (1995). Mysql. http://mysql.com.

Patzer A (2010). Mint. http://mint.com.

Pokorny B, Wheatley J, Amos R, and Myers D (2010). Daily booth. http://dailybooth.com.

Pousman Z, Stasko J, and Mateas M (2007). Casual information visualization: Depictions of data in everyday life. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1145–1152.

Powazek D, Bryant J, Snook J, and Boudreaux TJ (2010). Kvetch! http://www.kvetch.com/.

Ressi A (2010). Tweet what you eat. http://tweetwhatyoueat.com/.

revu (2011). re.vu. http://re.vu/.

Roberts S (2004). Self-experimentation as a source of new ideas: Ten examples about sleep, mood, health, and weight. *Behavioral and Brain Sciences*, 27(2):227–262.

Roberts S and Neuringer A (1998). Self-experimentation. *Handbook of research methods in human operant behavior*, pages 619–655.

Rodgers A, Corbett T, Bramley D, Riddell T, Wills M, Lin R, and Jones M (2005). Do u smoke after txt? Results of a randomised trial of smoking cessation using mobile phone text messaging. *Tobacco Control*, 14(4):255.

Shneiderman B (1992). Tree visualization with tree-maps: 2-d space-filling approach. *ACM Transactions on graphics (TOG)*, 11(1):99.

Shneiderman B (2003). The eyes have it: A task by data type taxonomy for information visualizations. *The craft of information visualization: readings and reflections*, pages 364–371.

Snook J, Rubin D, Veloso B, and Smith S (2010). Overheard.it. `http://overheard.it/`.

Stone B (2009). What's happening? `http://blog.twitter.com/2009/11/whats-happening.html`.

Support T (2010). Hashtags: a twitter community creation. `http://t.co/4pzgKHd`.

Swayne D, Lang D, Buja A, and Cook D (2003). Ggobi: Evolving from xgobi into an extensible framework for interactive data visualization. *Computational Statistics & Data Analysis*, 43(4):423–444.

Systems A (2012). Adobe flash. `http://adobe.com`.

Tartakoff J (2011). Photo social network dailybooth raises $6 million. `http://datafl.ws/283`.

Theus M (2003). Interactive data visualization using mondrian. *Journal of Statistical Software*, 7(11):1–9.

Thiel S (2010). Understanding shakespeare. `http://www.understanding-shakespeare.com/`.

Tufte E and Graves-Morris P (1983). *The visual display of quantitative information*, volume 31. Graphics press Cheshire, CT.

Twitter (2012). Shutting down spammers. `http://blog.twitter.com/2012/04/shutting-down-spammers.html`.

Van Ham F, Wattenberg M, and Viégas F (2009). Mapping text with phrase nets. *Visualization and Computer Graphics, IEEE Transactions on*, 15(6):1169–1176.

Van Kleek M and Karger D (2009). Watching through the web: Building personal activity and context-aware interfaces using web activity streams. *Understanding the User-Logging and Interpreting User Interactions in Information Search and Retrieval (UIIR-2009)*, page 36.

van Rossum G (1991). Python. `http://python.org`.

Viegas F, Wattenberg M, Van Ham F, Kriss J, and McKeon M (2007). Many Eyes: a Site for Visualization at Internet Scale. *IEEE Transactions on Visualization and Computer Graphics*, pages 1121–1128.

Viegas FB, Wattenberg M, and Feinberg J (2009). Participatory visualization with wordle. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1137–1144.

Wald D, Quitoriano V, and Dewey J (2006). Usgs "did you feel it?" community internet intensity maps: macroseismic data collection via the internet. In *First European Conference on Earthquake Engineering and Seismology. Geneva, Switzerland*.

Wang J (2012). How fitbit is cashing in on the high-tech fitness trend. `http://www.entrepreneur.com/article/223780`.

Wattenberg M (2005). Baby names, visualization, and social data analysis. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005*, pages 1–7.

Willett W, Heer J, and Agrawala M (2007). Scented widgets: Improving navigation cues with embedded visualizations. *IEEE Transactions on Visualization and Computer Graphics*, pages 1129–1136.

Woflram S (2012). The life and times of stephen wolfram: A scrapbook. http://www.stephenwolfram.com/scrapbook/.

Wolfram S (2012a). The personal analytics of my life. http://blog.stephenwolfram.com/2012/03/the-personal-analytics-of-my-life/.

Wolfram S (2012b). Wolfram—alpha personal analytics for facebook. http://www.wolframalpha.com/facebook/.

Yahoo (2010). Delicious. http://delicious.com/.

Yau N (2009). Progress: A graphical report on the state of the world. http://projects.flowingdata.com/state-of-the-world/.

Yau N (2010a). Flowingprints. http://flowingprints.com/.

Yau N (2010b). Growth of walmart. http://projects.flowingdata.com/walmart.

Yau N (2011). *Visualize This: The FlowingData Guide to Design, Visualization, and Statistics*. Wiley.

Yau N (2013a). *Data Points: Visualization that Means Something*. Wiley.

Yau N (2013b). Flowingdata. http://flowingdata.com/.

Yau N (2013c). Flowingdata projects. http://flowingdata.com/category/projects/.

Yau N and Schneider J (2009). Self-surveillance. *Bulletin of the American Society for Information Science and Technology*, 35(5):24–30.

YouTube (2007). 9 months of gestation in 30 seconds. http://www.youtube.com/watch?v=fJHz3rfPulA. Username: demiraydemir; Accessed: 2012/12/01.

YouTube (2008). My 1 year muscle gain and weight gain body transformation before and after. http://www.youtube.com/watch?v=BoY_fLbEXlE. Username: drummerboy0; Accessed: 2012/12/01.

Zeldman J (2005). Tag clouds are the new mullets. http://www.zeldman.com/daily/0405d.shtml.