DATA VISUALIZATION IN THE FIRST PERSON

Philip DeCamp

BS Electrical Engineering and Computer Science
Massachusetts Institute of Technology, 2004
MS Media Arts and Sciences
Massachusetts Institute of Technology, 2008

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning, in partial fulfillment of the requirements for the degree of Doctor of Philosophy at the Massachusetts Institute of Technology
February 2013

AUTHOR: Philip DeCamp
Program in Media Arts and Sciences
October 4, 2012

CERTIFIED BY: Deb Roy
Associate Professor of Media Arts and Sciences
Thesis Supervisor

ACCEPTED BY: Professor Patricia Maes
Associate Academic Head
Program in Media Arts and Sciences

DATA VISUALIZATION IN THE FIRST PERSON

Philip DeCamp

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning, on October 4, 2012, in partial fulfillment of the requirements for the degree of Doctor of Philosophy at the Massachusetts Institute of Technology

Abstract

This dissertation will examine what a first person viewpoint means in the context of data visualization and how it can be used for navigating and presenting large datasets. Recent years have seen rapid growth in Big Data methodologies throughout scientific research, business analytics, and online services. The datasets used in these areas are not only growing exponentially larger, but also more complex, incorporating heterogeneous data from many sources that might include digital sensors, websites, mass media, and others. The scale and complexity of these datasets pose significant challenges in the design of effective tools for navigation and analysis.

This work will explore methods of representing large datasets as physical, navigable environments. Much of the related research on first person interfaces and 3D visualization has focused on producing tools for expert users and scientific analysis. Due to the complexities of navigation and perception introduced by 3D interfaces, work in this area has had mixed results. In particular, considerable efforts to develop 3D systems for more abstract data, like file systems and social networks, have had difficulty surpassing the efficiency of 2D approaches. However, 3D may offer advantages that have been less explored in this context. In particular, data visualization can be a valuable tool for disseminating scientific results, sharing insights, and explaining methodology. In these applications, clear communication of concepts and narratives are often more essential than efficient navigation.

This dissertation will present novel visualization systems designed for large datasets that include audio-video recordings, social media, and others. Discussion will focus on designing visuals that use the first person perspective to give a physical and intuitive form to abstract data, to combine multiple sources of data within a shared space, to construct narratives, and to engage the viewer at a more visceral and emotional level.

THESIS SUPERVISOR: Deb Roy
Associate Professor of Media Arts and Sciences
MIT Media Arts and Sciences

DATA VISUALIZATION IN THE FIRST PERSON

Philip DeCamp

THESIS SUPERVISOR: Deb Roy
Associate Professor of Media Arts and Sciences
MIT Media Arts and Sciences

DATA VISUALIZATION IN THE FIRST PERSON

Philip DeCamp

THESIS READER: Martin Wattenberg
Computer Scientist and Artist
Google, Inc

DATA VISUALIZATION IN THE FIRST PERSON

Philip DeCamp

THESIS READER: Isabel Meirelles
Associate Professor of Graphic Design
Northeastern University

Table of Contents

1 Introduction

As this dissertation was being written, researchers at CERN announced the confirmation of a subatomic particle likely to be the Higgs boson. Confirming the particle's existence to a significance of 4.9 sigmas involved the analysis of about 10^{15} proton-proton collisions [Overbye, 2012] using sensors that record over one petabyte of data each month [CERN, 2008]. When the Large Synoptic Survey Telescope begins operation in 2016, it is expected to record image data at a rate of over one petabyte per year [Stephens, 2010]. Increasingly, scientific research is turning to massive datasets that no one person could hope to view in a lifetime, and that require dedicated data centers and processing farms just to access, let alone analyze.

Two years ago, the term "Big Data" entered our lexicon to refer to the growing trend of data analysis at very large scales, a trend that extends also to areas far beyond the hard sciences. Advances throughout information technologies have made it practical to collect and analyze data at scale in many areas where raw data was previously limited or prohibitively expensive. In particular, the explosion of online populations and communication devices, as well as digital sensors that are inexpensive enough to stick on everything, have made it possible to collect data from vastly distributed sources at little cost. The result has been a surge of interest in addressing a diverse range of problems, new and old, by applying massive amounts of computing to massive amounts of data.

The Santa Cruz Police Department has recently begun using crime pattern analysis tools to plan daily patron routes for officers [Olson, 2012]. Tools have been built to analyze large corpora of legal documents in order to predict the outcome of patent litigation and to aid in case planning [Harbert, 2012]. Several companies are developing commercial tools to optimize retail spaces

using data collected from in-store cameras, point-of-sale data, RFID tags, and others.

The most visible practitioners are the Internet companies: Google, Facebook, Amazon and others. These companies collect click-stream data, online transactions, communications, user generated content, and anything else that might be used to drive services for advertising, retailing, social networking, and general information retrieval. Nearly every person with a computer or phone is both a frequent contributor and consumer of information services that fall under the umbrella of Big Data.

Facebook alone counts about one-sixth of the world's population as its active users, who upload 300 million photographs every day [Sengupta, 2012]. Users of YouTube upload over ten years of video every day [YouTube, 2012]. These social networking sites are now a significant part of our global culture, and offer some of the most extensive records of human behavior ever created. One of the most fascinating examples of data mining comes from the online data site, OkCupid, which has a corpus of the dating habits of around seven million individuals. Using this corpus, they have published findings on ethnic dating preferences, the interests that most strongly differentiate between heterosexuals and homosexuals, and the seemingly random questions that best predict if a person might consider sex on a first date ("In a certain light, wouldn't nuclear war be exciting?") [Rudder, 2011]. The growing corpora of personal data offer new ways to examine ourselves.

And so the motivations for Big Data analysis are many, from scientific research, to mining business intelligence, to human curiosity. In turn, there are also many motivations to communicate effectively about Big Data, to explain what all this data is, disseminate scientific results, share insights, and explain methodology. These are all motivations behind the work described in this document, which will examine approaches to data visualization that make the analysis and communication of complex datasets clear and engaging.

1.1 Challenges

The datasets that will be examined in this document, like many of the datasets just described, pose several challenges to visualization.

First, they are far too large to view completely. They generally require ways to view and navigate the data at multiple scales.

Second, they are heterogeneous, comprised of multiple kinds of data collected from many sources, where sources might be defined at multiple levels as people, websites, sensors, physical sites, television feeds, etc. Drawing out the relationships between multiple sources of data often requires finding effective methods of synthesis.

Third, they are usually unique in structure. The more complex the dataset, the less likely it is to resemble another dataset collected in any other way. This places greater need to develop specialized visualization tools that work with a particular dataset.

There are many ways of distilling a dataset, and for very large datasets, any visualization will involve significant compression. The structure of the database might be viewed diagrammatically. Large portions of data can be reduced to statistical summaries, indexes, or otherwise downsampled. Fragments can be shown in detail. Different sources or relationships can be viewed in isolation. But looking at only one such view can only show a small part or single aspect. Forming an understanding of the whole must be done piece-by-piece, and through the exploration of broad summaries, details, components, relationships, and patterns.

1.2 Approach

The approach this document takes towards visualization is to represent large datasets as physical environments that provide a concrete form to abstract and complex data, that can be explored and seen from multiple viewpoints, and that bring multiple sources of data into a shared space.

The goal is not just to show such an environment, but to place the viewer inside of it and present data from a first person perspective. The intent is to tap into the viewers' physical common sense. When confronted with a physical scene, we have powerful abilities to perceive spatial structures and information that 2D or schematic representations do not exploit. We can reason about such scenes intuitively and draw many inferences about physical relationships pre-attentively, with little or no conscious effort. Our ability to remember and recall information is also influenced, and often enhanced, by spatial context. Last, a first person perspective can provide a more vivid sense of *being somewhere* that can help to create more engaging graphics.

This dissertation will:

Define what a first person viewpoint means in the context of data visualization.

Present a body of visualization work that demonstrates techniques for establishing a first person viewpoint, and how those techniques can be put into practice.

Examine the response received from exhibition of the work and provide critique.

1.3 Applications

The bulk of this dissertation is comprised of visualizations that use first person to address challenges encountered in real applications. Much of the work began with developing tools for retrieval and analysis, created for use in my own research in areas of computer vision and cognitive science, or by other members of the Cognitive Machines research group. Much of the work has also been created or adapted for use in presentations to communicate research methods and results.

One of the most widely seen exhibitions of the work occurred at the TED 2011 conference. Deb Roy gave a 20-minute talk on research from the Media Lab and from Bluefin Labs, a data analytics company of which Roy is cofounder. The majority of the visual content consisted of data visualizations, created primarily by myself and Roy, that were intended to explain the research to

a general audience. A video of this event was made publicly available shortly after the talk, has been seen by millions of viewers, and has generated discussions on numerous high-traffic websites. Some of the critique received from these discussions will be examined in Section 5 in evaluation of the work.

Other work has been created for use on US national broadcast television, which will be discussed in Section 4.2.

1.4 Terminology

The terms *data*, *data visualization*, *information*, *information visualization,* and *scientific visualization* are not always used consistently. This document adopts several working definitions to avoid potential confusion.

Most important is the distinction between *data* and *information*. For the purposes of this document, data is like the text in a book, and information is what is communicated through the text. A person who cannot read can still look at the text and perceive the data, but does not derive the information. Similarly, a bar chart maps quantities, data, to the size of bars. What the bars represent and the inferences drawn from the chart are information.

Unfortunately, this distinction between data and information has little to do with extant definitions of *data visualization* and *information visualization*. [Card et al., 1998] offer a definition of *visualization* as "the use of computer-supported, interactive, visual representations of data to amplify cognition." They further distinguish between *scientific* and *information visualizations* based on the type of data. *Scientific visualization* refers to representations of spatial data, such as wind flow measurements. *Information visualization* refers to representations of *abstract data*, data that is non-spatial or non-numerical, and that requires the designer to choose a spatial mapping. However, these definitions are ambiguous when working with heterogeneous data.

[Post et al., 2003] define *data visualization* as including both scientific and information visualization. For simplicity, this document uses *data visualization* exclusively to refer to any visual representation of data.

2 The First Person

What does a *first person viewpoint* mean in the context of data visualization? For software interfaces, a first person viewpoint implies a navigation scheme in which the user moves through a virtual environment as if walking or flying. And while we refer to such systems as *first person interfaces*, our categorization of viewpoint also include many elements beyond 3D navigation. Furthermore, a data visualization might not be interactive at all, but an image or animation. The concept of first person extends to all of these mediums, as it does to cinema, painting, video games, and literature.

Defined broadly, the first person depicts a world from the eyes of a character that inhabits and participates in that world. The third person depicts a world from the viewpoint of a non-participant, a disembodied observer. To extend the terminology, the term zeroth person will refer to a representation that establishes no sense of a world or characters at all, as in an abstract painting or instruction manual. Most data visualizations, like bar charts, also fall into this category[1].

The distinction between viewpoints is not always clear. Whether a representation presents a world and characters, and whether the viewpoint represents that of an inhabitant or of no one, may all be ambiguous. Furthermore, the criteria used to make such judgments depend on the properties and conventions of the medium.

In video games, the distinction between first and third-person shooters is based on a slight shift in camera position. Figure 1 shows a first-person shooter, where the player views the world

1 Second person is conspicuously omitted here due its infrequent use. The second person is looking at yourself through someone else's eyes. This is simple to accomplish linguistically with the word *you*. Representing the viewer visually is more difficult, but might include looking at a photograph or video recording of yourself, or the rare video game in which the player controls a character while looking through the eyes of an uncontrolled character, as seen in the first boss fight of the NES game *Battletoads*.

Figure 1. A first person shooter.

Figure 2. A third person shooter.

Figure 3. Diego Velázquez. *Las Meninas*. 1656.

from the eyes of his character. Figure 2 is from a third-person shooter, where the camera is placed a few feet behind the character. In either case, the player identifies with the character in the game and views the environment from a perspective very close to that of the character. Our categorization of viewpoints is not something that is defined absolutely, but relative to the norms of the medium. In the medium of 3D-shooter video games, Figures 1 and 2 represent the narrow range of viewpoints normally found, and so we call the one that is slightly closer from the character's perspective *first person*.

Categorization of viewpoint may be more ambiguous for images, which provide less obvious cues as to whether or not the image represents the viewpoint of some character. An interesting example of viewpoint in painting is provided by Michel Foucault in *The Order of Things* [Foucault 1970]. In the first chapter, Foucault meticulously examines Diego Velázquez's Las Meninas and the different ways it relates to the viewer. At first glance, the viewer might see the 5-year old princess standing in the center of the room and the entourage surrounding her. These characters occupy the center of the space and initially appear to be the focus of attention in the painting, providing a typical third person view in which the viewer, outside the painting, views a subject of interest within the painting.

On closer inspection, many of the characters are looking out of the painting fairly intently, including a painter, Velázquez himself, who appears to be painting the viewer. A mirror in the back of the room also reveals the royal couple standing in the position of the viewer. These elements give the viewer the role within the scene, as a person being painted, possible the king or queen. The center of focus is not the princess, but rather, the princess and entourage are there to watch and perhaps entertain the royal couple as they pose for a portrait. The focus is on the viewer. A first person view.

Foucault also describes the dark man hovering at the door in the back. Compositionally, he mirrors the royal couple, but stands behind the space of the room while the royal couple stand in

front of it. The historical identity of this character is known, but Foucault suggests that it might double as a representation of the viewer, as someone who has happened into a scene and pauses to look in. A second person view.

Intruding into the left of the painting and occupying nearly the entire height is a canvas on which the represented artist is painting. To give so much prominence to the back of a canvas is unusual, and Foucault theorizes it may be intended to guide the viewer's thoughts away from the representation and towards the physical canvas that he is looking at in reality, which, too, has nothing behind it. The painting is not a scene, but just a canvas. A zeroth person view.

What we consider to be a first person viewpoint is not defined by any single element, and as discussed in the Velázquez example, different elements within a representation can support contrasting interpretations. What the elements of a first person viewpoint have in common is that they establish some form of egocentric relationship with the viewer, where the viewer does not perceive the representation as a configuration of light and symbols, but as a physical world that includes them, that the viewer might interact with, or where the world might affect the viewer in some way. Viewpoint is associated most strongly with visual perception, where first person is the viewpoint that most strongly creates a sense of *perceived immersion*, of the viewer perceiving a scene as surrounding himself. However, the purpose of the Velázquez example is to show that viewpoint also occurs at a cognitive level. The different viewpoints presented in the painting all come from the same visual stimuli, but differ in how that stimuli is interpreted.

The work shown in this thesis will not attempt to manipulate viewpoint as subtly as this painting, but will approach viewpoint as something that extends beyond perception and that includes this kind of psychological engagement. Colin Ware authored *Information Visualization* [Ware 2004], which focuses on the perception of visualizations. In the book, Ware includes a brief discussion on the topic of *presence*:

> "One of the most nebulous and ill-defined tasks related to 3D

> space perception is achieving a sense of *presence*. What is it that makes a virtual object or a whole environment seem vividly three-dimensional? What is it that makes us feel that we are actually present in an environment?
>
> Much of presence has to do with a sense of engagement, and not necessarily with visual information. A reader of a powerfully descriptive novel may *visualize* (to use the word in its original cognitive sense) himself or herself in a world of the author's imagination—for example, watching Ahab on the back of the great white whale, Moby-Dick."

A viewer is more likely to feel immersed and engaged in a representation that "feels" like a physical environment, and so the concept of presence is at the core of what a first person viewpoint means within this document. As Ware notes, the concept is not well defined, and within the book, he does not attempt to delve much deeper into the subject. There is still much to explore in what defines presence, how to establish it, and how it might be applied to data visualization.

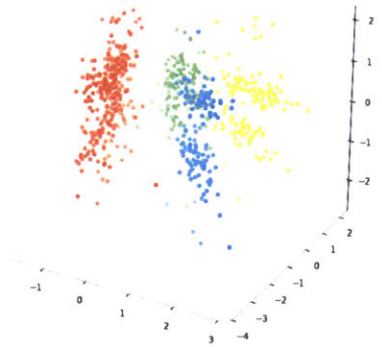


Figure 4. A typical scatter plot may be navigated in the first person, but does not provide a sense of physical engagement.

Establishing a sense of presence involves more than just representing a 3D space. Any 3D scatter plot can easily be explored using a first person navigation scheme, but even so, the representation may not provide a sense of being in a physical environment. A representation of flying through a nearly empty space, populated sparsely by intangible floating dots, is perceptually unlike any view of the real world we are likely to encounter, and more to the point, unlikely to evoke a similar experience.

Our minds learn to recognize particular patterns of visual stimuli and, through experience, associate them with patterns of thought and reasoning, described by Mark Johnson as *image schemata* [Johnson, 1990]. When looking at the small objects on top of a desk, we are likely to perceive the support structures of stacked objects, how we might sift through the objects to find a paper, or how a coffee mug will feel in our hand. When we look around from the entrance of building, we are likely to draw inferences about the layout of the building, where we might find an elevator, and how we can navigate towards it. One way to define presence is to say that representations with presence more strongly evoke

the image schemata we associate with physical environments, and lead to similar patterns of thought and engagement.

There are many individual techniques that might be used in visualization design to establish presence to varying degrees: creating a sense of depth and space, representing data in the form of a familiar object or structure, emulating physical properties like gravitational acceleration and collisions, emulating physical navigation and interaction, rendering naturalistic details and textures, or providing the viewer with a clear sense of position and scale. The rest of this document will provide more concrete examples of these approaches, and will examine how to establish presence and first person engagement for both photorealistic and non-photorealistic environments.

3 The Human Speechome Project

In 1980, Noam Chomsky proposed that a developing child could not receive and process enough stimulus from his environment to account for learning a complex, natural language. The theory followed that, if true, then part of language must be accounted for by biology, and aspects of language are hard-wired in the brain [Chomsky, 1980]. This argument is widely known in linguistics as the poverty of stimulus, and through several decades and into the present day, a central challenge in this field has been to identify the aspects of language that are innate, the aspects that are learned, and the relationship between the two.

Language might be viewed as the product of two sets of input, genetics and environment. Of the two, genetics is the simpler to quantify. The human genetic code is about 700 megabytes, and several specimens are available for download. But the environment includes all of the stimulus the child receives throughout development, including everything the child sees and hears. One of the difficulties in responding to the poverty of stimulus argument is that it is difficult to produce an accurate figure for the amount of environment data a child actually receives or how much might be useful. But the number is certainly greater than 700 megabytes, and likely lies far in the realm of Big Data.

Capturing the input of a child is a difficult and messy task. Standard approaches include *in vitro recording*, in which the child is brought into a laboratory for observation. *In vivo* recording is usually performed by sending scientists into the home environment for observation, or with diary studies in which a caregiver records notable events throughout the a child's development. In vivo methods provide more naturalistic data collected that comes from the child's typical environment, with in vitro methods only observe the child's atypical behavior in an unfamiliar laboratory.

However, all of these approaches suffer from incompleteness and capture only a tiny fraction of the child's input. As a result, each time the child is observed, he is likely to have developed new abilities during the time between observation, making it difficult or impossible to determine how those abilities were acquired.

Other researchers have lamented that the lack of high-quality, longitudinal data in this area is largely to blame for our poor understanding of the fine-grained effects of language on acquisition [Tomasello and Stahl, 2004]. A more complete record might answer numerous questions about what fine-grained vocabulary actually looks like, the influence of different environmental factors on development, and the patterns of interaction between children and caregivers that facilitate learning.

The poverty of environmental data was one of the motivations behind the Human Speechome Project (HSP). *Speechome* is a portmanteau of speech and home, meant also as a reference to the Human Genome Project. Where the Human Genome Project created a complete record of a human's genetic code, HSP intended to capture the experience of a developing child as completely as possible with dense, longitudinal audio-video recordings. Recent advances in digital sensors and storage costs offered an alternative solution to the problem of observing child development: to install cameras and microphones throughout the home of a child and simply record everything. Of course, recording the data is the comparatively easy part. The difficult task that HSP set out to address was how to develop methodologies and technologies to effectively analyze data of that magnitude.

3.1 Setup

I began working on HSP shortly after its conception in 2005. The family to be observed was that of my advisor, Deb Roy, and his wife, Rupal Patel, a professor of language-speech pathology at Northeastern University. Roy and Patel were expecting their first child, and initiated the project several months into the pregnancy. This provided enough time to instrument the house, develop a recording systems, and construct a storage facility before the

Figure 5. The HSP recording site.

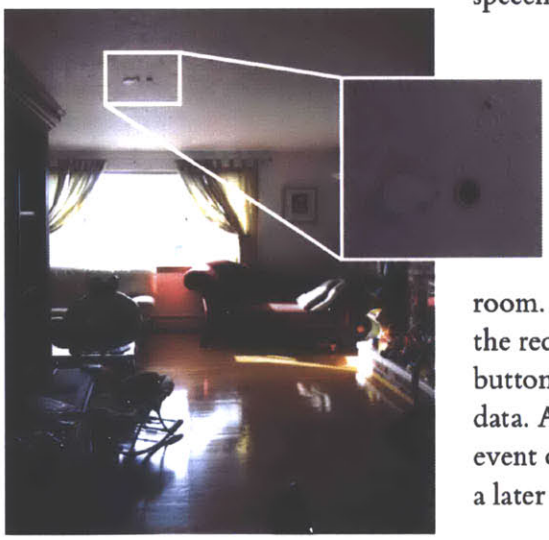


Figure 6. A camera and microphone mounted in ceiling. The microphone is the small silver button near the top.

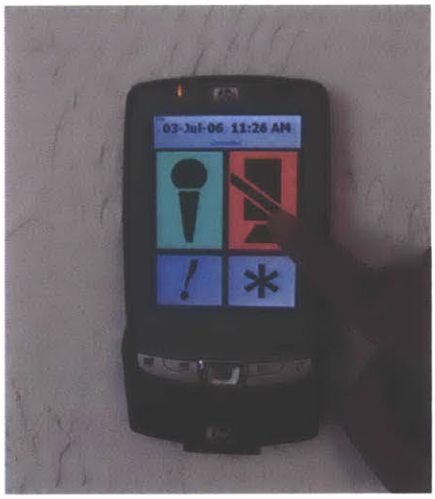


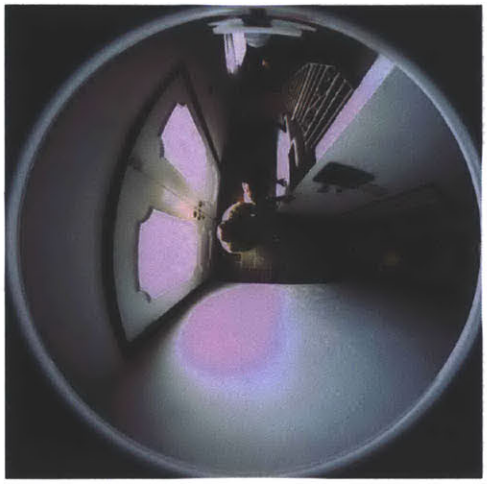
Figure 7. One of the recording control panels mounted in each room.

child arrived. Eleven cameras and fourteen microphones were installed in ceilings throughout most rooms of the house. Video cameras were placed near the center of each ceiling looking down, and were equipped with fisheye lenses that provided an angle-of-view of 185 degrees, enabling each camera to capture an entire room from floor to ceiling. The audio sensors were boundary layer microphones, which sense audio vibrations from the surface in which they are embedded and use the entire ceiling surface as a pickup. These sensors could record whispered speech intelligibly from any location in the house.

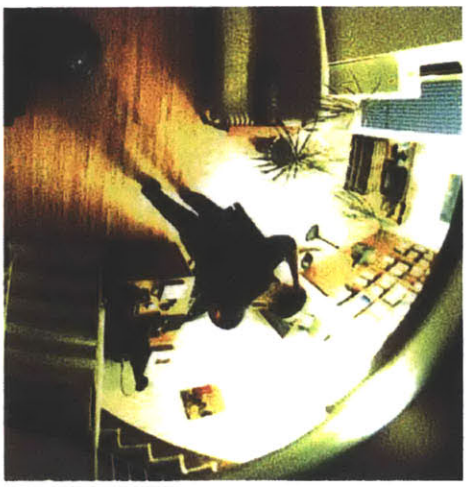
The goal of recording everything was not entirely possible, and over the course of three years, the participants would require moments of privacy. Participants in the home could control recording using PDAs—an older type of mobile device that resembles a smart phone without the telephony—that were mounted in each room. Each panel had a button that could be pressed to toggle the recording of audio or video. Another button, the "oops" button, could be pressed to delete a portion of recently recorded data. And last, an "ooh" button, could be pressed to mark an event of interest so that the event could be located and viewed at a later time.

Recording began the day the child first came home from the hospital, and completed after the child was three years old and speaking in multi-word utterances. The corpus from this project includes 80,000 hours of video, 120,000 hours of audio, and comprises about 400 terabytes of data. This data is estimated to capture roughly 80% of the child's waking experience within the home, and represents the most complete record of a child's development by several orders of magnitude. A more detailed account of the recording methodology and system can be find in [DeCamp, 2007].

Compared to what the child actually experienced, this record is certainly not complete. It does not contain recordings of smell, touch, taste or temperature. It is limited to audio-video from a set of fixed perspectives, and does not show things in the same way the child saw them, or with the same resolution. Yet, nearly



2005/07/29 12:13 PM
The Child Arrives



2005/07/29 05:03 AM
Myself, exasperated, trying to get the recording system to work hours before the arrival.

every aspect of the child's experience is represented, in part, within the data. What the child said and heard, his interactions with others, his patterns of sleep and play, what he liked or disliked, are all forms of information contained in the audio-video record. But the analysis of any such information is predicated on the ability to extract it.

3.2 The Data

The audio-video recordings are referred to as the *raw data*. Most analysis requires extracting more concise forms of data from the audio-video, like transcripts of speech, person tracks, prosody, and others, which are referred to as metadata. Extracting useful metadata from audio-video at this scale can be difficult. Automatic approaches that rely on machine perception are cheapest, but available technologies limit the kinds of information can be extracted automatically and the accuracy. Manual approaches that require humans to view and annotate multiple years of data can be extremely expensive, even for relatively simple annotation tasks. And in between are human-machine collaborative approaches, in which humans perform just the tasks that cannot be performed automatically.

At the inception of HSP, it was unclear what information could

be extracted using available tools, what new tools could be developed, and what information would be economically feasible in the end. So the project did not begin with a specific set of questions to answer, but rather a range of inquiry about language and behavior. An exploratory approach was taken towards choosing paths of research that balanced the relevance of potential results against the expected cost of mining the required data. Although the ultimate goal of the project was to develop a model of language acquisition grounded in empirical data, many of the significant contributions came from the methodologies researchers developed to extract relevant behavioral information from raw data.

Linguistic analysis required transcripts of the recorded speech. A key goal of the project was thus to transcribe all speech that occurred in the child's presence during his 9th to 24th months, representing the period just before he began to produce words, and ending after he was communicating in sentences and multiword utterances. Current speech recognition technologies were unable to transcript the speech with any reasonable accuracy. The audio recordings contain unconstrained, natural speech, including significant background noise, overlapping speakers, and the baby babble of a child learning to talk. Furthermore, although the audio quality was relatively high, recordings made with far-field microphones still pose problems for the acoustical models used in speech recognition. Brandon Roy led efforts to develop an efficient speech transcription system that uses a human-machine collaborative approach. Roy's system locates all audio clips containing speech and identifies the speaker automatically, then organizes the audio clips into an interface for human transcription [Roy and Roy, 2009]. As of this writing, approximately 80% of the speech from the 9 to 24 month period has been transcribed, resulting in a corpus of approximately 12 million words.

Person tracking, or identifying the locations of the participants within the video, was required to analyze interactions, spatial context, and often as a starting point for more detailed video analysis. Person tracking in a home environment requires following people moving between rooms, severe lighting contrasts between indoor lights and the natural light entering windows,

and attempting to track a child that was frequently carried by a caregiver. George Shaw developed an automatic, multi-camera tracking system used to extract much of the track data that will be shown in this document [Shaw, 2011].

Many other forms of data have been extracted to varying degrees of completeness. Most of these will play a smaller role in the following discussion, but may be of interest to those developing methods of analyzing human behavior from audio-video recordings. A few of these include:

Prosody: The intonation of speech, including pitch, duration, and intensity for individual syllables. Many aspects of caregiver prosody have turned out to be significant predictors of vocabulary developing in the child [Vosoughi, 2010].

Where-Is-Child Annotations (WIC): Annotations describing the room in which the child was at any given point in the recorded data, and whether the child was awake or sleeping. This metadata was largely used to quickly locate the child within the data, both for data navigation tasks, and to reduce unnecessary processing of data irrelevant to the child's development.

Head Orientation: Head orientation is a useful indicator of gaze direction and attention, what the participants are looking at, if the child is looking at a care giver directly, or if the child and caregiver share joint-attention within an interaction [DeCamp, 2007].

Affect Classification: The emotional state of the child during different activities [Yuditskaya, 2010].

Sentiment Classification: The attitude or emotional polarity of a given utterance. For example, "Awesome!" has a positive sentiment and, "Yuck!" a negative sentiment.

Taking the raw data together with the metadata, the HSP corpus is large, contains multiple forms of interrelated data, and is unique.

3.3 TotalRecall

After the recording process began, the immediate question became how to look through the data, verify its integrity, and find information of interest. Skimming through just a few hours of multi-track audio-video data can be time consuming, let alone finding specific events or interactions. This led to the development of TotalRecall, a software system designed for retrieval and annotation of the HSP corpus. This interface did not use any 3D graphics or address issues of viewpoint, but serves here as a baseline for a more conventional approach.

The TotalRecall interface provides two windows. A video window that displays the raw video, with one stream at full resolution and the other streams displayed at thumbnail sizes on the side. The timeline window provides visual summaries of the audio-video recordings. The horizontal axis represents time, and can be navigated by panning or zooming in and out of different time scales. Each horizontal strip represents one stream of audio or video.

The audio data is represented with spectrograms, a standard visualization of the audio spectrum over time. Users can skim spectrograms to find areas of activity within the audio. With some practice, users can learn to quickly separate different types of audio. Human speech contains formant structures that generate zebra stripe patterns. Doors and banging objects, like dishes, produce broad spectrum energy bursts that appear as sharp verti-

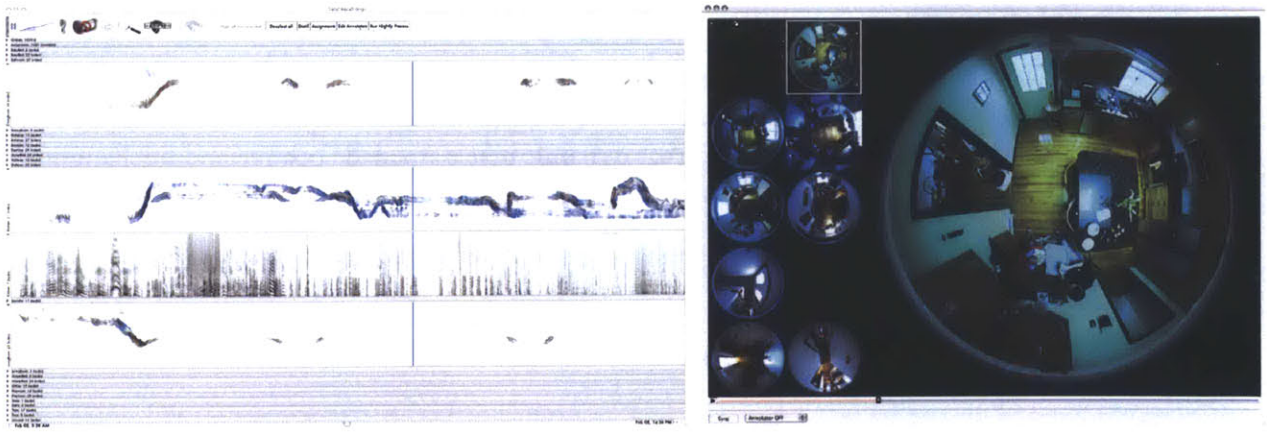


Figure 8. The TotalRecall interface used for browsing the HSP data.

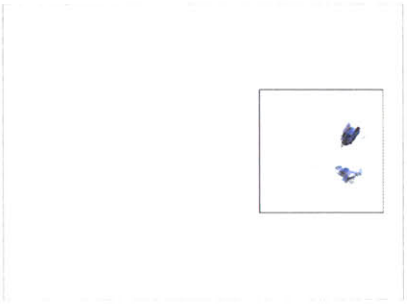
cal lines. Running water and air conditioners produce sections of nearly uniform noise.

Summarization of the video was more challenging. The standard method used in most video editing and retrieval interfaces is to show individual frames, often selecting frames scene boundaries or points of significant change. This approach works poorly for the HSP video, which contains no scene changes or camera motion. Most rooms are unoccupied, and where there is activity, it comprises only a small portion of the image. Consequently, identifying the differences between video frames requires close attention and more effort relative to edited video.

However, the consistency of the video offers other advantages. Most of the content of a given stream is already known. The living room camera will always show the same of the living room, and the portions of greatest interest are changes in the foreground. Rather than try to show the whole contents of the video frames, an image stack process was used to transform each stream of video into a video volume, a continuous strip that depicts only the motion within the video.

The process begins with a stream of raw video.

The per-pixel distance between adjacent frames. The distance map generated for each frame is then used to modulate the alpha channel, such that dynamic pixels are made opaque and unchanging pixels are made transparent.



These images are then composited onto a horizontal image strip, with each subsequent frame of the video shifted a few more pixels to the right. This maps the vertical position of motion onto the vertical axis of the image, and maps both time and horizontal position onto the horizontal axis.

The result transforms moving objects into space-time worms, where each segment of the worm represents a slice of time. More generally, the process converts continuous video into a continuous image. Similar to spectrograms, users can view a set of video volumes for all the streams of video and, with minimal training, quickly identify where and when there was activity in the home. By itself, this was of great value in searching through hours or months of 11-track video. With additional experience, viewers may quickly learn to identify more specific patterns as well. From the number of worms, users can identify the number of people in a room, and from the size, differentiate between child and caregivers. The level of intensity indicates the amount of motion, with the limitation that people at complete rest may nearly disappear for periods of time. The coloration provides information about lighting conditions, and can be used to follow some brightly colored objects, including articles of clothing and certain toys. Some activities also produce noticeable patterns, including instances when the child was in a bounce chair or playing chase with a caregiver.

A similar *image stack* process for visualizing video was described previously in [Daniel, 2003]. In this work, Daniel et al. render the video as an actual 3D volume. Our application for video volumes was different in that it we needed to view longitudinal, multi-track video. Consequently, we adapted the approach by flattening the image stack into a flat, straight rectangular strip in order to make it more suitable for display on a multi-track timeline.

Rony Kubat developed the main window of TotalRecall, with Brandon Roy, Stefanie Tellex, and myself. I developed the video window, along with all audio-video playback code. This video volume technique was developed by Brian Kardon, Deb Roy, and myself. A more detailed account of the system can be found in [Kubat, 2007].

3.4 Partitioned Views

One of the choices made in the design of TotalRecall was to present different source of data separately, each in its own partitioned view. An advantage of this approach is that it presents each source accurately and simply, and makes explicit the underlying structure of the corpus. However, this partitioning obscures the relationships between sources of data. The representation of the data is partitioned rather than composed.

In particular, there is a strong spatial relationship between all the sensors in the house that has been largely omitted. Consequently, viewing a person moving between rooms, or viewing a caregiver speaking to the child in the dining room, can require some effort to follow. In these cases, the user must watch an event multiple times from multiple views, repeatedly finding the desired track of audio or video out of the many presented, and to mentally compose that information to gain a complete picture of the activity. Similarly, the interface does not provide a clear overview of the whole environment, the spatial layout and the participants present at a point in time.

Figure 9. An interface for audio, video and motion data created by Ivanov et al.

A partial solution may have been to include a map view that presents the space as a whole. For example, Yuri Ivanov et al. developed an interface similar to TotalRecall for a dataset containing multi-camera video, person tracks, and motion sensor data. As seen in Figure 9, one view displays the video, another the timeline with annotations, and another the map of the space with overlaid motion data.

The addition of the map view is useful in understanding the spatial arrangement of the environment and interpreting motion data. However, it does little to combine the different types of data, and the spatial data is still separate from the video. As with TotalRecall, gaining an understanding of the environment from the visualization is not a simple perceptual task, but requires the user to cross-reference a spatial view, temporal view, and video view.

Both the Ivanov interface and TotalRecall present data in the as a set of multiple, mostly abstract views. Again, this approach was likely suitable for their respective purposes as browsing interfaces for expert users. However, even for us "experts," comprehending and navigating video across cameras was difficult. And for an untrained user, a first glance at TotalRecall does not reveal much about what the data represents. When using the system to explain the Human Speechome Project, it required around 10 minutes to explain how to interpret the different visual elements, much as it was described here, and what they reveal about activities within the home. In the end, the audience may still only have a partial picture of what the data contains as a whole.

In motivating HSP, a narrative frequently told in demonstrations was that we had captured an ultra-dense experiential record of a child's life, which could be used to study how experience affected development and behavior. While many found this idea compelling, skeptical listeners would sometimes argue that while a great amount of data about the child had been recorded, it did not capture much of what the child experienced. It was easy to understand the skeptics because they were presented with a disjoint set of data that bore little resemblance to their own experiences of the world.

While the data is far from a complete experiential record, part of the issue is literally how one looks at the data. In the next example, the same set of data will be presented in the first person as a way that more clearly evokes the subjective experiences of the participants.

3.5 Surveillance Video

The raw video of the HSP corpus is surveillance video, which, taken by itself, is not always the most engaging or cinematic view of an environment. The video does not focus in on any particular area of interest, and any activity is usually limited to a small region of the total image. This emphasizes the setting and deemphasizes the people within it. Furthermore, it provides a third person viewpoint where the overhead angle forces the viewer to

look down into the scene from above, rather than a more typical eye-level shot as if the viewer were actually within the scene.

Figure 10. A man under surveillance in *The Conversation*.

In cinema, shots of surveillance or CCTV footage are usually diagetic, indicating that a character is being recorded within the narrative. This device has been notably used in films like *The Conversation*, *Rear Window*, and *The Truman Show*. These shots often have ominous or lurid undertones, and tap into a cultural uneasiness surrounding the proliferation of surveillance and loss of privacy [Levin, 2006]. And indeed, although the HSP participants were aware of being recorded and in control of the system, this discomfort with the idea of constant surveillance surfaced frequently in discussion of the project, with terms like "Big Brother" voiced more than occasionally. Although this does not detract from any information in the video, it can give the viewer of the system the sense of being an eavesdropper. And in presentations, this can be a distraction in the scientific intent of the project.

One HSP researcher, Kleovoulos Tsourides, performed a clever experiment by first tracking a person within a clip of video, then using the track data to reprocess the video, zooming into the region containing the person and rotating each frame to maintain a consistent orientation. This virtual cameraman system made the video appear as if shot by a cameraman following the person using a normal-angled lens. This system was not completely developed and had few opportunities to be demonstrated, and it may be that for people unfamiliar with the data, the effect may not have seemed markedly different. But for those of us working on the project that had been viewing the surveillance video for several years, the transformation was remarkable. It replaced the impression of surveillance video with the impression of cinematic video, and gave the impression that the video contained substantially more information and detail. Of course, the process only removed information, but by removing what was irrelevant made the relevant information that much greater.

3.6 HouseFly

In addressing some of the limitations of TotalRecall, I created a new interface for browsing the HSP data called HouseFly. Rather than partition the sources of data, HouseFly synthesizes the data into a 3D simulation of the home. The user can navigate this environment in the first person and watch events with a vivid sense of immersion. Because of the density of the HSP data, the system can render the entire home in photographic detail, while also providing rapid temporal navigation throughout the three year recording period. HouseFly also serves as a platform for the visualization of other spatio-temporal metadata of the HPS corpus, combining multiple types of data within a shared space for direct comparison. As a tool for communication, the system makes the data immediately accessible and engages the viewers in the recorded events by bringing them into the home.

Figure 12. Raw video used to construct the 3D model below.

Figure 11. The HouseFly system showing an overhead view of the recorded home.

Figure 13. Google Earth 3D.

Figure 14. Google StreetView.

Figure 15. Video Flashlights.

3.7 Related Work

The virtual reconstruction of physical locations is a general one, with broad applications in the visualization of spatial environments. One of the most visible examples is Google Maps, which includes a StreetView feature that provides a street-level, first-person view of many cities. For general path planning, conventional maps may be more efficient, but the first-person view provides additional information on what a location will look like when the traveler is present, and can be used to identify visible landmarks for guidance or find specific locations based on appearance [Anguelov et al., 2010]. Google Maps and similar services provide coverage of very large areas, but primarily as snapshots in time, with limited capabilities for viewing events or for temporal navigation. As a web interface, spatial navigation is also highly constrained, and the user must navigate rather slowly between predefined locations.

Sawhney et al. developed the Video Flashlights system for conventional video surveillance tasks, which projects multi-camera video onto a 3D model of the environment. This system does not rely on static cameras, and uses a dynamic image registration to automatically map the video data to the environment. It can also present track data within the environment [Sawhney, 2002]. The flashlight metaphor of Video Flashlights is one of using video data to illuminate small regions of the model. It places the video in a spatial context and combines connects recordings to each other and to the environment, but the range of exploration is limited to localized areas.

HouseFly builds on these technologies and uses more recent graphics capabilities to to perform non-linear texture mapping, allowing for the use of wide-angle lenses that offer much more coverage. But the most significant advantages of HouseFly are provided by the data. The HSP corpus provides recordings of a complete environment, in detail, and over long periods of time. This led to the design of an interface that provides freer exploration of the environment and through time.

3.8 Constructing the Environment

The first step in developing HouseFly was the creation of a spatial environment from the data. The environment has three components: a 3D model of the house that provides the geometry, video data used as textures, and a spatial mapping from the *geometry* to the textures.

The model of the house is a triangular mesh, created in Google SketchUp. The model is coarse, containing the walls and floors, doorways, and using simple boxes to represent fixtures and large furniture. This model was partitioned manually into *zones*, where the geometry within each zone is mapped to a single stream of video. Generally, each zone corresponds to a room of the house.

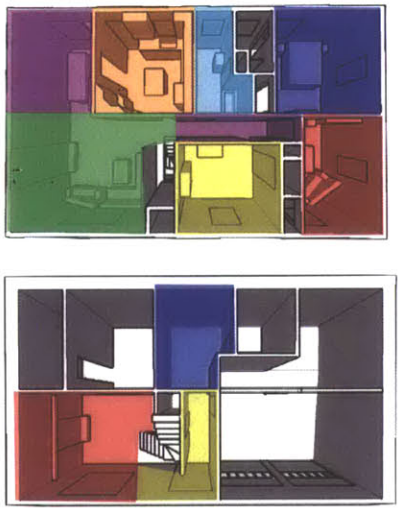


Figure 16. Partitioning of the environment geometry into zones, each of which is textured by a single camera.

Creating a spatial mapping for each stream of video requires a mathematical model of the camera optics and a set of parameters that fit the model to each camera. The extrinsic parameters of the camera consist of the camera's position and orientation. The intrinsic parameters describe the characteristics of the lens and imager.

Given the use of a fisheye lens, it is simpler to ignore the lens itself and instead model the imager surface as a sphere. The zenith axis of the sphere, Z, exits the front of the camera through the lens. The azimuth axis, X, exits the right side of the camera. $Z \times X$ is designated Y and exits the bottom of the camera. The center of the sphere is C.

To map a point in world space, P, to an image coordinate, U, P is first mapped onto the axes of the camera:

$$\tilde{P} = [XYZ](P - C) \tag{1}$$

$\tilde{P}$ is then projected onto the sensor sphere:

$$\theta = \cos^{-1} \frac{\tilde{P}_z}{|\tilde{P}|} \tag{2}$$

$$\phi = \tan^{-1} \frac{\tilde{P}_y}{\tilde{P}_x} \tag{3}$$

where θ is the *inclination*, and ϕ is the *azimuth*. Last, (θ, ϕ) is mapped into image coordinates:

$$U = \begin{bmatrix} S_x \theta \cos \phi + T_x \\ S_y \theta \sin \phi + T_y \end{bmatrix} \tag{4}$$

where S_x and S_y are scaling parameters, and T_x and T_y are translation parameters.

Thus, Equation 4 contains four *scalar* parameters, while Equations 1-3 require six scalar parameters: three to define the center of the sensor, C, and three to define the orientation as a set of Euler angles: yaw, pitch, and roll. Together, these equations define a mapping function between world coordinates and image coordinates, $f(P : \Theta) \rightarrow U$, where Θ represents the ten camera parameters.

Camera Calibration

Finding the ten parameters for each camera comprises the calibration process. This is performed by first finding a set of *correspondence points*, which have position defined in both image and world coordinates. For each correspondence point, the image coordinates are specified by clicking directly on the video frame. World coordinates are extracted from the 3D model in Sketchup and entered manually into the calibration interface. Given a sufficient number of correspondence points, a Levenberg-Marquardt nonlinear solver was used to fit the parameters.

Figure 17.The interface used for camera calibration. Rony Kubat developed the interface, and I developed the camera model and parameter solver.

Texture Mapping

Figure 20 shows the simplified geometry for one zone of the partitioned environment model, and below shows the texture used for that region. Normally, with a rectilinear lens, the texture can be mapped to the geometry at the vertex level. That is, the camera model defines a function that maps a world coordinate to a texture coordinate, and that function is applied once for each vertex of the model geometry. With a rectilinear lens, the texture for each point of the triangle can be computed accurately by linearly interpolating the texture coordinates of its vertices.

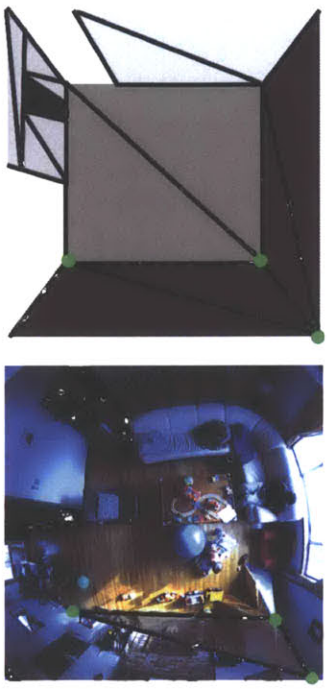


Figure 20. Partitioning of the environment geometry into zones, each of which is textured by a single camera.

However, the fisheye lenses are not modeled well with linear functions. Note that in Figure 20, although the geometry above is rendered from an angle to approximately align with the texture below, the match is not very accurate. The edges between the floor and walls are straight on the geometry, but appear curved in the texture, leading to distortion. This distortion grows greater towards the edges of the texture as it becomes more warped and non-linear. Figure 18 shows the result of using a piece-wise linear, per-vertex mapping, where distortion becomes increasingly severe as the model approaches the edges of the texture.

Subdividing the geometry produces a finer mapping and reduces distortion, but requires far more video memory and processing. Beyond a certain threshold, when the vertices of the subdivided geometry no longer fit in memory, the geometry must be loaded dynamically according to the user's current position, greatly increasing the complexity of the renderer.

Figure 18. Per-vertex mapping.

Figure 19. Per-fragment mapping.

Fortunately, modern GPUs feature programmable shaders that can perform non-linear texture projection. Instead of mapping each vertex to the texture, the renderer loads the camera model onto the graphics card, which then computes the correct texture coordinate for each pixel at every render pass. Unless the graphics card is being taxed by other operations, per-fragment incurs no detectable performance penalty while eliminating completely the non-linear distortions, as in Figure 19.

Rendering

Given the spatial model, textures, and mapping, the house can be rendered in full. Each zone is rendered separately. For each zone, the associated texture object is *bound*, the camera parameters are loaded into the fragment shader, and the geometry is sent to the graphics card for rasterization and texturing.

A benefit of this approach is that any frame of video may be loaded into the texture object and will be projected onto the environment model without any preprocessing. As a result, animation of the environment is just a matter of decoding the video streams and sending the decoded images directly to the texture objects. In this document, all the video is prerecorded, but in future applications, live video streams may be viewed just as easily.

Controls

HouseFly provides fluid navigation of both time and space. The user may go to any point in time in the corpus and view the environment from any angle.

Temporal navigation is very similar to conventional interfaces for multi-track video playback. A collapsible timeline widget displays the user's current position in time, and may be clicked to change that position. Using either a keyboard or a jog-shuttle controller, the user may control his peed along the time dimension. The data can be traversed at arbitrary speeds, and the user may watch events

in real-time, thousands of times faster, backwards, or frame-by-frame.

HouseFly was designed to create a similarly first-person viewpoint into the data. It supports two primary schemes for spatial navigation, both of which follow a metaphor of moving through the environment rather than moving the environment.

First, navigation can be performed with a keyboard and mouse using the same controls as a first-person shooter. The *WASD* keys are pressed to move forward, left, backward, and right, and the mouse is used to rotate. The *Q* and *E* keys are pressed to increase or decrease elevation. The drawback to this scheme is that it requires both hands, making it difficult to simultaneously control time.

Second, navigation can be performed using a SpaceNavigator input device. This device consists of a puck mounted flexibly on a heavy base, where the puck can be pushed-pulled along three axes, and rotated about three axes, providing six degrees-of-freedom. Navigation with the SpaceNavigator provides full control over orientation and position using only a single hand. The drawback is that this device requires significant practice to use effectively.

Figure 21. SpaceNavigator and jog-shuttle controller used for time-space navigation.

Audio

For audio, one stream is played at a time. The system dynamically selects the stream of audio recorded nearest the user's location.

While this simple approach does not capture the acoustic variation within a room, it does capture the variations between rooms. When the user is close to people in the model, their voices are clear. When the user moves behind a closed door, the voices from the next room become accurately muffled. When the user moves downstairs, he can hear the footsteps of the people overhead. Such effects may not draw much attention, but add greatly to the immersiveness of the interface.

When playing data forward at non-real-time speeds, SOLA-FS is used to correct the pitch of the audio [Hejna, 1991], which improves comprehension of speech [Foulke 1969].

Implementation

HouseFly was developed in the Java programming language. 3D graphics were produced with OpenGL using the Java bindings provided by the JOGL library. Shader programming was done in GLSL and Cg. Functionality that required significant optimization, like video decoding, was written in C. C was also necessary for interfacing with input devices, including the SpaceNavigator and jog-shuttle controller.

The graphics engine developed for HouseFly is similar to a 3D game engine, with similar techniques used to manage assets, script events, and handle user input. This engine was used for most of the original visualizations in this document.

Roughly, the hardware requirements of HouseFly are below that of most modern first-person shooter games. The system runs smoothly, between 30 and 60 frames-per-second, on personal computers and newer laptops. More specific figures depend greatly on how the software is configured and the data being accessed.

The most significant bottleneck of the system is video decoding. First, due to the size of the corpus, the video must be pulled from the server room via Ethernet. Network latency is largely mitigated through aggressive, predictive pre-caching. Alternatively, if not all the data is required, a subset may be stored locally for better performance.

Second, video decoding is currently performed on the CPU. The HSP video is compressed using a variant of motion JPEG with a resolution of 960 by 960 pixels at 15 frames per second. On a laptop with a two-core 2.8 GHz processor, the system can replay four streams of video without dropping frames. On an eight-core 2.8 GHz processor, frame dropping becomes noticeable at around eight streams of video.

However, it is not always necessary to decode many streams of video. For most viewpoints within the house, only three or four rooms are potentially visible. To improve performance, HouseFly dynamically enables or disables playback of video streams based on the location of the user.

Baseline Summary

The tools provided by HouseFly encourage the exploration of the data as a whole environment rather than as individual streams. The user can navigate to any point in time within the corpus and view it as a rich 3D environment, filled with objects and people that move, speak, and interact with the environment. The user can move closer to areas of interest, pull out to look at multiple rooms or the entire house, and follow events fluidly from one room into the next. Rather than looking down into the scene, the user can look from within, gain a clear sense of the spatial context that connects both data and events, and receives a much closer approximation of what the participants saw, heard, and experienced.

The system still has significant limitations in its ability to reconstruct the environment. The people and objects do not actually stand up and occupy space, but are projected flatly onto the surfaces of the environment. There is no blending between zones, so when a person walks to the edge of one zone towards the next, the person is chopped into two texture pieces recorded from two

Figure 22. Examples of the home environment as rendered in HouseFly.

different viewpoints. Also, many parts of a given room are not visible from the camera, and there is no data available for texturing areas under tables and behind objects. Just filling these blind areas with black or gray made caused them to stick out conspicuously next to the textured areas. Instead, the video data is projected onto these areas regardless of occlusions, and the area underneath a table is given the same texture as the top of the table.

Surprisingly, though, many viewers tend not to notice these issues. In demonstrations, listeners frequently asked how the people are rendered into the environment, and only realized that the people were not 3D models but flat projections after the camera was moved to floor.

Most methods of acquiring geometry from video are far from tractable, and even if applicable to the HSP video, would produce numerous artifacts and draw more attention to the limitations of the representation. So while the model HouseFly provides is coarse, there is enough detail within the video to provide a vivid depiction of a naturalistic, 3D environment.

3.9 Visualizing Metadata

In addition to the audio-video, the HSP corpus contains many forms of metadata. HouseFly provides several ways to incorporate metadata into the scene, combining it with other data and placing it in context. In turn, the metadata also provides methods for searching and navigating the corpus, greatly improving the accessibility of the data.

Much of the metadata used for temporal navigation is placed on the timeline widget, shown in Figure 23. The timeline displays the user's place in time, and can be expanded to show an index of the audio-video recordings organized by room. The green bars represent audio recordings and the blue bars represent video recordings. Clicking on a room within the timeline moves the user temporally to that place in time and also moves the user spatially to an overhead view of the selected room.

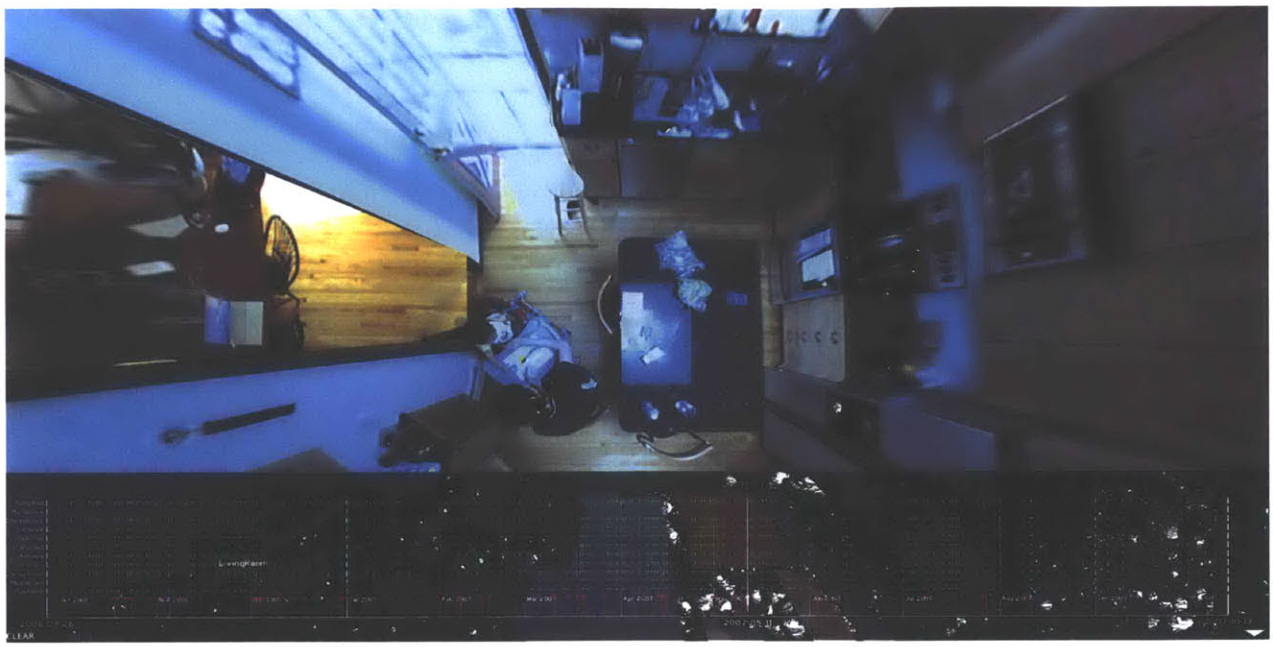


Figure 23. The timeline in HouseFly.

The orange and yellow bars are the Where-Is-Child annotations, showing the room in which the child is located. The bar is orange if the audio for that period has been transcribed, and yellow otherwise, where transcripts will be discussed later. The viewer can then browse through the timeline to quickly determine the location of the child at a given point in time, and view the child by clicking on the bar.

The small flags at the bottom of the timeline are bookmarks. The user can create a bookmark associated with a point of time, and may optionally associate the bookmark with a room of the house or a specific camera position.

As described earlier, the PDA devices mounted in each room as control panels also contained an "ooh" button that could be pressed to mark significant moments. These events were incorporated into the system as bookmarks, represented here by the pink flags. The user may browse these bookmarks to view events like the child's first steps and many of his first words. One ooh event was made after describing the recording system to the child, explaining that years later, the child would be able to go back and find that moment in the data.

Transcripts

Transcripts of the speech are used in several ways. Most directly, the transcripts can be rendered as subtitles, similar to closed-captions. This is helpful in revealing verbal interactions when skimming the data, and also aids comprehension of the language, where the child speech in particular can be difficult to understand.

The transcripts are fully indexed and searchable. The user can query transcripts by typing in a word or phrase. All the matching instances will be selected, and are placed on the timeline, where the user can browse through the selected instances one by one. The color of the transcript as it appears on the timeline indicates the speaker, where green represent caregiver speech, and red represents child speech.

Any given word can be used as a lens to explore patterns of language development. By typing in any given word, the user can quickly find the child's first use of that word. By browsing subsequent instances, the viewer can hear how that word developed over time. By also viewing the context in which the word was used, the user can determine if the child was using the word accurately, if he was requesting an object or merely identifying it, or if the child over or under-generalized the meaning of the word.

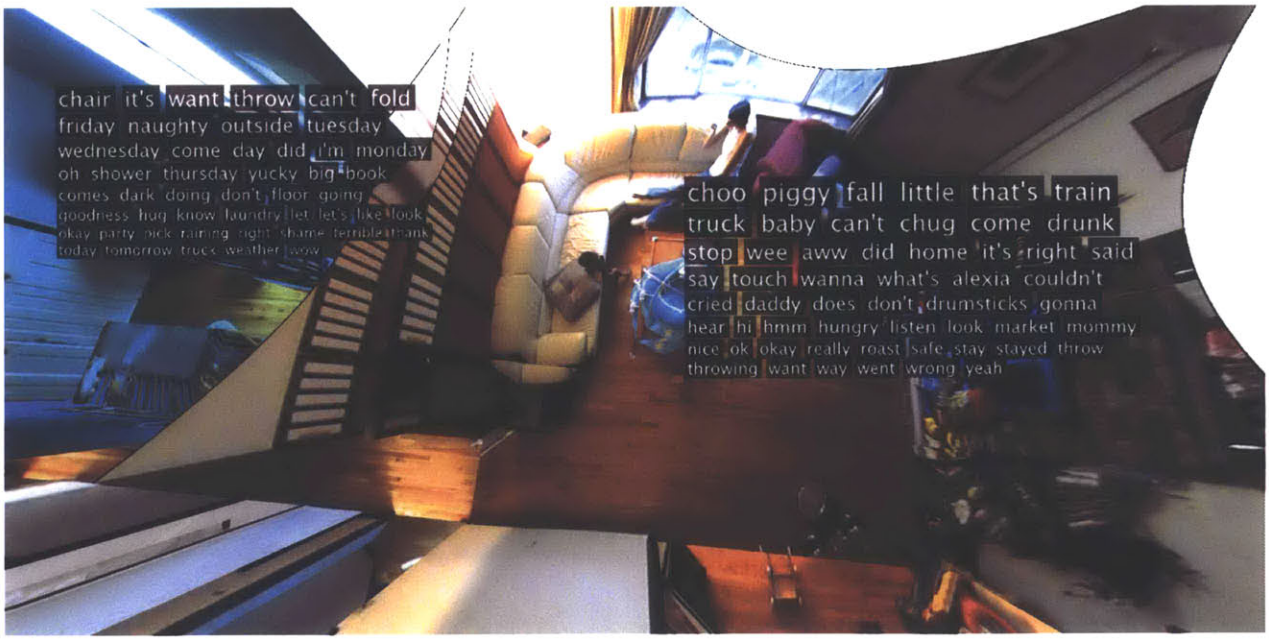


Figure 24. Summarization of transcripts as tag clouds.

HouseFly also summarizes the contents of transcripts, displaying a distribution over word types in each room as a tag cloud. By default, turning on tag clouds shows the word type distribution from the previous 30 minutes of data. If any data has been selected and placed on the timeline, the word clouds will display a summary of all selected data. For example, after performing a transcript search for the word "fish," the timeline will contain the set of the thousands of transcribed utterances containing that word, and the tag clouds will display the distribution of words that co-occurred with "fish." This enables the user to rapidly view the linguistic context of that word. The user may compare how the word was used in the child's bedroom, which contained fish magnets and a fish mobile, next to how the word was used in the kitchen, where fish was something to eat. Arbitrary segments of data may also be selected from the timeline and similarly summarized.

Tracks

Person tracks were generated by identifying and following blobs of motion or color through each video stream. The resulting track data was mapped into the coordinate space of the environment using the same camera models that HouseFly uses for texture mapping. Extracting accurate 3D coordinates from 2D video is a difficult problem and is not addressed in this work. Instead, the mapped track data assumes a fixed elevation of one meter from the floor for all objects. Objects are frequently visible from multiple cameras simultaneously, particularly around doorways and the edges of rooms, which produces multiple, overlapping tracklets, or partial tracks. After the tracklets were mapped into a unified coordinate system, the tracklets that overlapped and were thought by the tracking system to correspond to the same object were merged, resulting in a set of full tracks that extend across multiple rooms. This is a quick overview of what is a complicated and messy process, which is described in greater detail in [Shaw, 2011].

An advantage of the spatially consistent view provided by HouseFly is that the multi-camera tracks can be rendered directly in the same environment model. When track rendering is enabled,

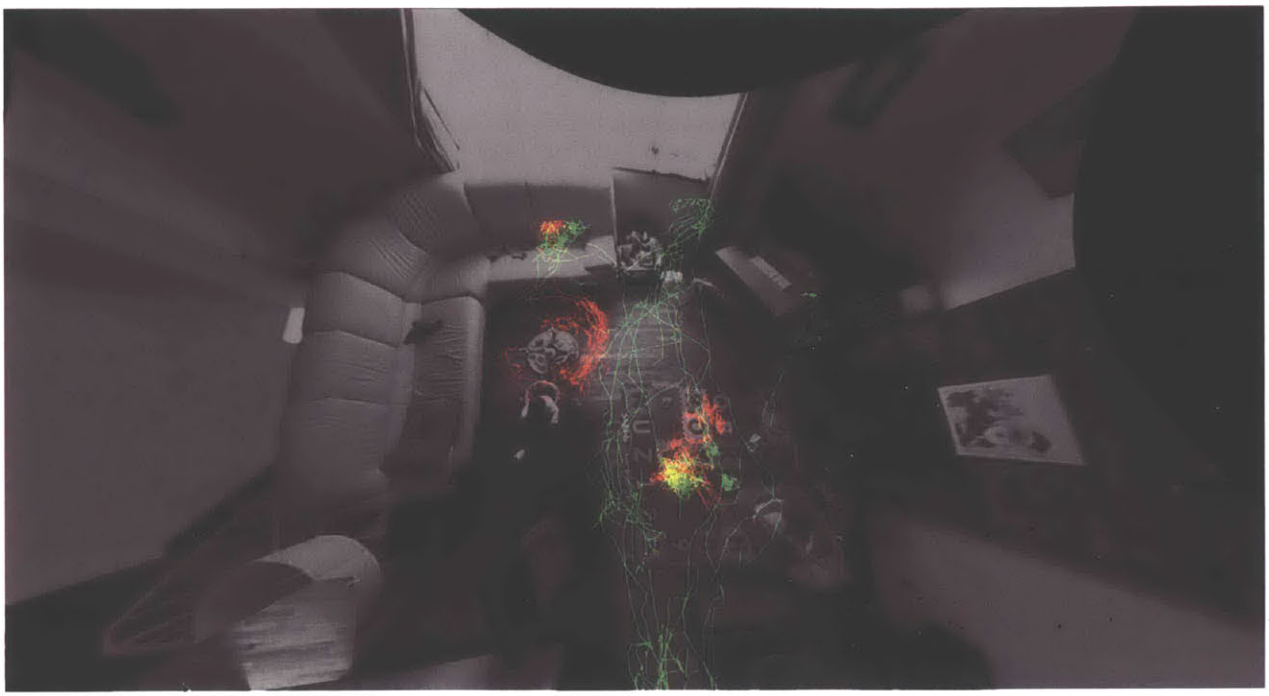


Figure 25. 30 minutes of track data rendered into the environment. The green represents the caregiver, the red the adult.

HouseFly renders the video in grayscale to make the tracks more visible, while still enabling the tracks to be viewed in context. Figure 25 shows 30 minutes of track data. The red tracks indicate the child, and the ring structure in the upper-right surrounds the child's walking toy intended for the child to walk around. The green tracks represent the caregiver, who made several trips into the kitchen via the dining room on the left, as well as one trip to a computer in the lower-right hand corner. The yellow spot in the center of the room resulted from both child and caregiver occupying the same location.

HouseFly can also render tracks by mapping time to elevation, such that the tracks begin on the floor and move upward as time progresses. For small amounts of track data, this can better reveal the sequence of events, as illustrated in Figure 26. However, the 3D structure can be difficult to perceive from a 2D image. While the structure is made more evident in the interface through motion parallax, improving the legibility of the 3D tracks remains a challenge for future work. This technique has previously been explored in [Kapler and Wright, 2004].

The selection of what track data to render uses the same selection mechanism as the tag clouds. When track rendering is enabled, whatever time intervals have been selected in the timeline de-

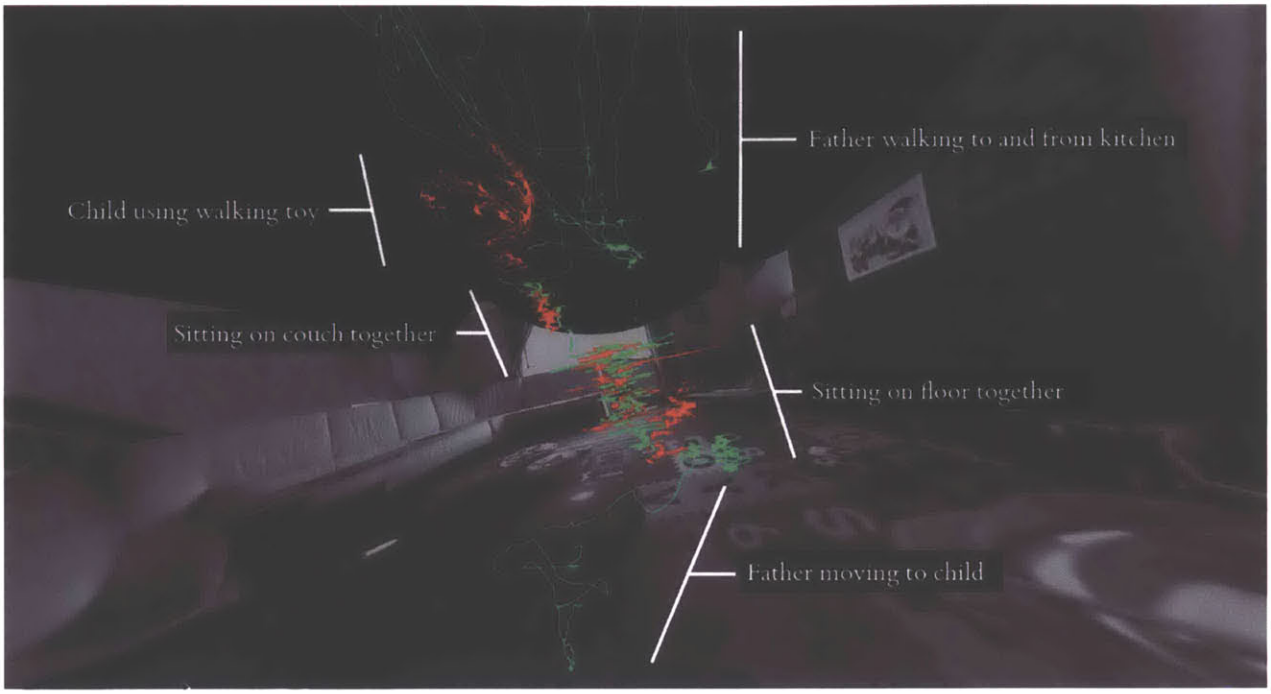


Figure 26. The time of track points mapped to vertical height to show sequence of events.

termines the track data that is shown. When track rendering is enabled, any data that the user traverses is automatically selected, so that the user may simply skim or jump through the video and the corresponding tracks will appear. If the user searches the transcripts for all instances of *car*, then those instances will be selected, and all available track data that co-occurred with that word can be viewed.

The tracks also provide a way to make spatial queries of the data. The user may click on any area of the environment, drag out a sphere, and the system will locate all tracks that intersect with that sphere and select the corresponding intervals of time, as in Figure 27. The user can then browse all data that contains activity in a particular region of the house, or enabled tag clouds to get a summary of what people said in different locales.

Figure 27. A spatial query performed by selecting a region of the environment, shown as a sphere.

Queries

The different kinds of metadata are all linked together using a shared selection mechanism, which provides the user with great flexibility in how to query the data. If the user is interested in the spatial distribution of a given word or phrase, he can query the

transcripts to retrieve the corresponding track data. If the user is interested in the patterns of speech that occur by the kitchen sink, he can begin by selecting that spatial region and view a summary of the words produced there. If he wants to view the activities of a given day, he can select that region of time and view the tracks and transcripts from that period. In any of these queries, the user can easily retrieve the raw audio-video data and view specific events in detail. Each form of metadata provides an index over the entire corpus.

3.10 Presenting HSP to an Audience

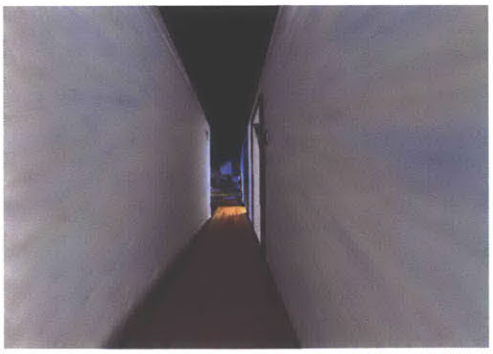
One benefit of HouseFly is that it shows many streams of data in a way that is immediately recognizable. Even for those completely unfamiliar with the project, viewing HouseFly clearly depicts the observed environment and the scope of the recorded data. This has made HouseFly useful as a communication tool for presenting the HSP project to new audiences.

For the TED presentation, HouseFly was used to give an overview of the home environment, the data recorded, and the implications for behavioral analysis.

The visualization begins with an overhead shot of 3D model constructed from the recorded video.

The camera swoops into the child's bedroom, revealing that the model is an explorable environment. By the standards of current computer graphics, this may not seem to particularly impressive. Yet, in the majority of demonstrations, including TED, this moment draws an audible response of excitement from the audience. This might be due to first presenting an graphic that appears to be an image and then defying those expectations, or perhaps the novelty of exploring actual video recordings in the same way as a video game.

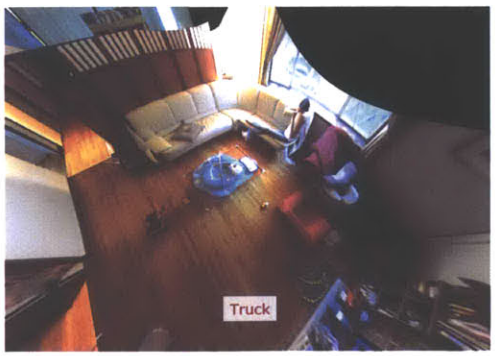
The guest bedroom is shown from the inside.



The camera then goes through the door and flies down the hallway.

The camera moves a short ways down the stairs and peeks into the first floor of the house, showing that the lower level is there as well.

The camera then goes into kitchen and performs a full turn, showing the completeness and detail of the model.

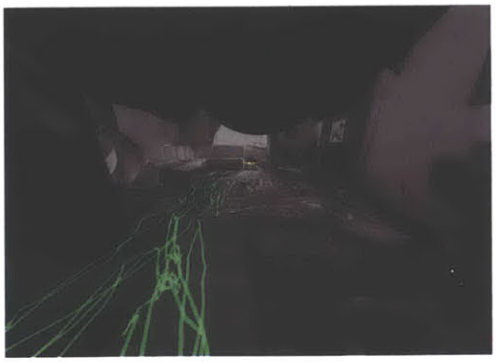


The next shot presents an example of a typical caregiver-child interaction. The camera moves into the living room, where the child sits on the floor and the nanny on the couch. The audio and video begin to play, bringing the model into motion. The nanny asks the child to find a fire truck, and the child walks to the shelf to find it, and selects an ambulance instead. The camera follows the child, showing the ability to look closely at areas of interest. Subtitles of the speech are shown at the bottom of the screen to improve speech comprehension.

The next shot is to explain person tracking. The camera moves to an overhead view of the living room, and the time of the video shifts to a point at which the child and father are sitting on the floor.

The video transitions to grayscale, and as the father and child move about, their paths are rendered on screen, the child in red and father in green. This is the data generated by the video tracking system.

Time speeds up, indicated by both the speed of the video and a clock in the upper-right corner, until approximately half an hour of video has been played. As the camera pulls up, the tracks can be seen to extend to other rooms, and that the caregiver has made several trips through the dining room and into the kitchen. The child's tracks circle around a point in the floor, where the underlying video shows the child's walking toy.



To make the sequence of events more evident, the time of the track data is mapped to the elevation of the tracks, such that the earliest tracks begin at the floor and rise into the air as they move forward in time. For this, the camera moves nearer to the floor, almost level with the tracks, before the tracks spread vertically into this 3D structure.

Viewing the tracks again shows that the father and son began in the center of the room, moved to the couch for a while, and then split up, with the father going to and from the kitchen, while the child walked around his toy.

3.11 Wordscapes

One of the key goals of HSP was to study language within context. The context of language contains innumerable factors and can be difficult to model in great detail, but one salient factor that can be extracted efficiently from video is the locations of the participants within the home. Different locations are correlated with different types of activities, and thus different patterns of speech. Speech in the kitchen often involves words about eating and cooking, while speech in the living room contains more words about toys and books. The HSP corpus contains many thousands of such spatial-linguistic correlations, some predictable, and some not. By analyzing these correlations, we hoped to identify the roles that different activities play in language development.

Linguists make a distinction between word *types* and *tokens*. Every instance of the word *green* in this document is a distinct token, but all the tokens belong to the same *green* type. The starting point of the analysis was to construct a spatial distribution of each word type learned by the child that described how likely a word type was to be produced at any location in the house. For each type, all tokens were extracted from the transcripts. For each token, the locations of all participants in the home were extracted by applying a person tracking system to a 20 second window of video centered on that token.

The result of this process was a set of 2D points for each word. Figure 28 shows the set of 8685 points found for *water*. The difficulty when plotting so many points directly is that there is a large amount of overlap between them, making it hard to accurately gauge density. The points can be made smaller to reduce overlap, but this makes them more difficult to see at all in sparsely covered regions.

Figure 28. 8685 utterances of *water* spoken throughout the home.

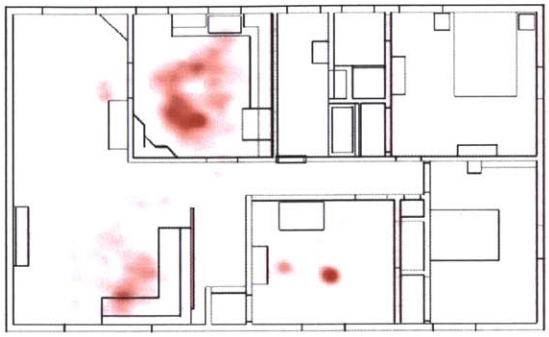


Figure 29. Heatmap of estimated distribution.

To provide a more consistent view of the density, the points were converted into a continuous distribution function using a kernel density estimation process described in [Botev, 2010]. Figure 29 shows a heatmap of this distribution. This avoids issues of overlap, although, as is a problem with all heat maps, our quantitative judgment of color is not very accurate.

In this case, where the domain of the function is small relative to the inflections of the function, a 3D surface plot might be more accurate. But in this application, a disadvantage of using either a heatmap or a surface plot is that they obscure the underlying samples and no longer appear as an aggregation. And none of these methods provide a clear sense of the physical space being examined or the scale at which these patterns of activity occur. This motivated me to develop a new kind of plot that would better communicate what this data represented and the amount of processing required to produce such a glimpse into the use of a single word.

Figure 30 presents a different view of the data as a *Wordscape*. Instead of representing the samples as dots or representing the distribution as a smooth contour, each sample is rendered as 20 seconds of track data. The representation uses a physical metaphor of rendering the tracks as ribbons that, as they are added to the scene, lay on top of one another to form a topographic distribution.

In creating this visualization, each track is first modeled as a finely segmented, planar line strip. As each vertex of the line strip is added to the scene, a height table over the discretized space is checked to determine how many segments have already been

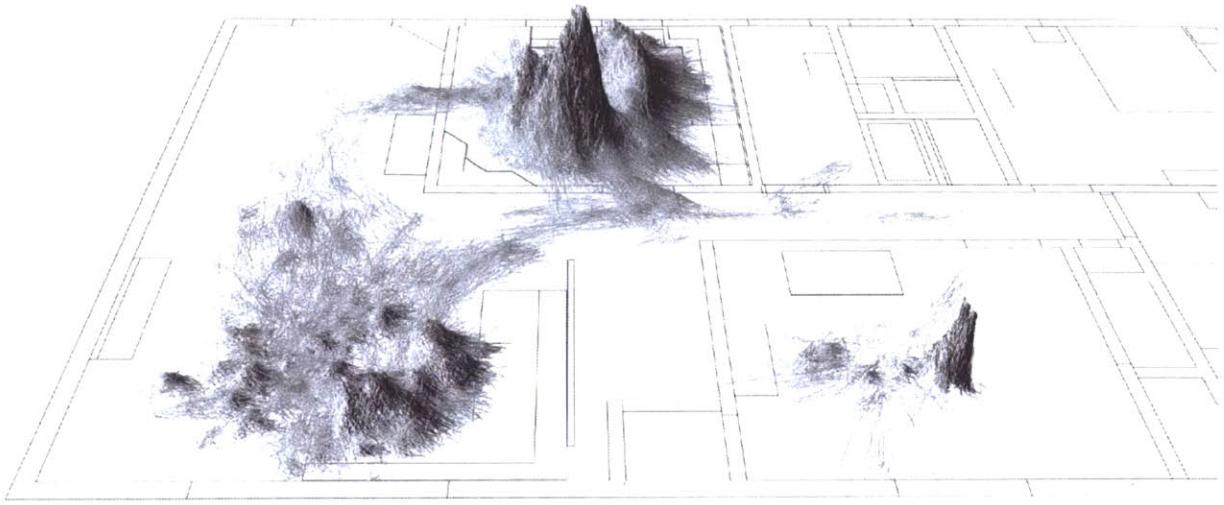


Figure 30. Thousands of person tracks combined to reveal a spatial distribution of the word *water*.

placed in that location, which is used to set the z-coordinate of the new vertex. After all the segments have been placed, the line strip is smoothed with a box filter to remove both noise present in the raw track data and aliasing effects generated by the use of a discrete height table.

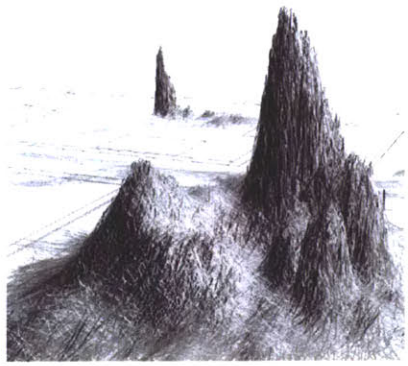


Figure 31. Detail of peak in kitchen.

The line strip is then converted to triangles in the form of a ribbon of constant width. This ribbon is bent along its length to form an extruded *V* shape so that it does not disappear when viewed from the side, but retains at least 35% of its perceived width from any angle. The geometry is rendered with a Lambertian shader using a single directed light source and smoothed normals. Although the ribbon geometry is bent, the normals are computed as if the ribbons laid flat so that the crease down the center remains invisible.

This method of rendering track data is believed to be novel, and offers several advantages. It shows the distribution of word production, as the heatmap does, and likely provides a better sense of the quantitative density through the use of height rather than color. At the same time, it shows the individual samples, providing a look at the form of the underlying data and a rough sense of the size of the dataset. And it connects the two modalities, the individual samples and the distribution, through a physical metaphor of a pile of ribbons that is easily recognizable and more intuitive than a representation of a Gaussian convolution.

The interface built for this visualization uses the same rendering engine as HouseFly and enables the user to move through the plot in the same manner. The similar use of first person navigation provides a greater sense of the physical space than the 2D maps of the previous page, although without the rich detail provided by the video, the scale and nature of the environment is still not as apparent. However, there may be several ways to combine these two views of the home. The following animation presents one approach, and was produced to explain the data at the TED presentation.

The video opens on a scene in the living room rendered in 3D in the manner of HouseFly. This orients the viewer with a representation that immediately recognizable.

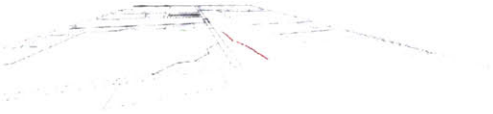
Here, the nanny is standing near the wall at the end of the couch. The child is camouflaged in this still image, but is standing nearby between the couch and coffee table.

The audio and video begin playing immediately and present the viewer with an example of a single sample point, a short interaction containing a particular word.

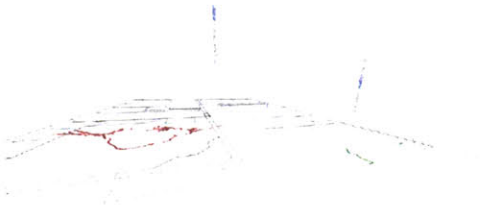
As the two people move through the room, a thick colored ribbon is drawn behind them to mark their paths, directly relating the track data with the movements of the inhabitants. Only the most recent seconds of track data is highlighted with color, red for the child and green for the caregiver, and fades to gray as it grows longer.

The nanny asks, "Would you like some water?" and extends a glass to the child, and the child replies, "No!" and turns away. I wanted to visually connect each track to be shown with the occurrence of *water*. So, just as the caregiver says, "Would you like some water?" a caption of this utterance rises vertically from the ground from her position, with the word *water* highlighted in blue. The caption continues to rise until it moves outside of view.

At this point, the video also begins to fade out in order to shift focus from the event to the track and transcript data.



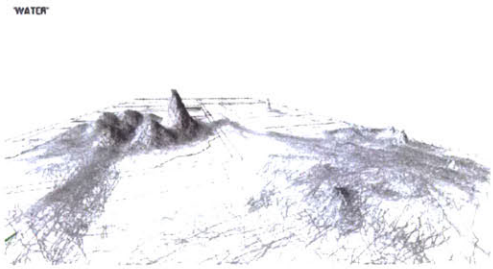
The video fades out completely and the camera moves to the side to setup the next beat. The tracks of the child and nanny continue to extend.



The aggregation of the tracks begins slowly. A few more tracks begin to appear in throughout the home in the same manner as the first, each generating a text caption.

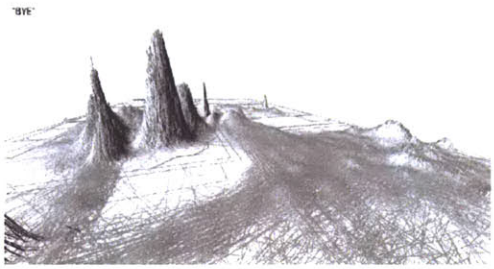
At its climax, the video builds to a frenzy of text and tracks.

Unlike a heatmap or surface plot, the viewer sees the individual samples that establish the distribution. This provides a rough, qualitative sense of the significance, which can otherwise be a difficult concept to explain to audiences unfamiliar with statistics.



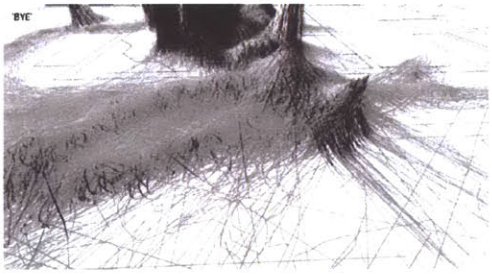
As the last of the captions leaves the screen, the distribution emerges. Utterances of water are shown to be highly concentrated within the kitchen area of the home.

For the next beat, we wanted to draw a comparison between the spatial distributions of different word types. In an early draft, the wordscape for *water* would build-out by sinking into the floor, followed quickly by the build-in of the *bye* wordscape emerging from the floor. The problem is that at first glance, the distributions of many words appear similar and have significant peaks in the kitchen, the living room, and the child's bedroom. For the *bye* wordscape, shown right, the most salient visual difference is that larger peaks of the kitchen, which is somewhat misleading because *bye* was a more frequent word and simply generated more samples.

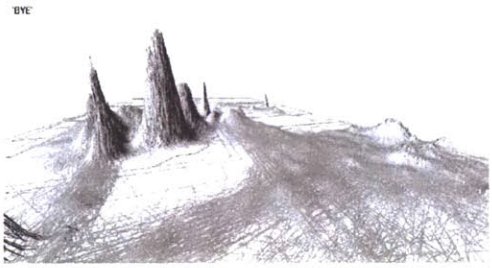


Much of the importance of a given word type's spatial distribution is in how it differs from the spatial distribution all speech. For analysis, it is useful to plot the difference between distributions or a normalized distribution. But for presentation, this introduces another level of abstraction, and would break the metaphor of the pile of tracks.

Instead, before the transition to *bye*, a change is made to the camera position to focus on an area of significant divergence from the mean distribution. Specifically, the camera moves to an area just outside the kitchen door and next to the stairs that lead to the entrance of the house.



When the wordscape transitions to *bye*, the viewer sees significant growth in this region and the formation of a small mound. This particular mound represents a specific type of event: of people saying "bye" to those in the nearby rooms before leaving the home.



The camera then moves back to show more of the distribution.

More than showing the spatial distributions of word types, this animation also shows what the data represents. The opening gives the viewer a concrete image of how each line represents a small event in the child's life. The use of animation is used to link the single event to the distribution to show how quantitative patterns of behavior emerge from many such events. The animation shows the process with visual excitement, and results in a representation that has more presence and physicality than a typical surface plot.

This video does not address the linguistic analysis performed

by researchers, or present the quantitative results. More detailed information on the spatio-linguistic analyses can be found in [Roy, B. 2012] and [Miller, 2011], with additional publications pending.

In the animation shown, track data was provided by George Shaw, and transcript data by Brandon Roy. Deb Roy and I conceived of the visualization, and I created the software system and animation to produce it.

3.12 First Steps: Making Data Personal

The rise of personal data offers opportunities to look at people in new ways, and thus new ways to tell dramatic stories. For the recorded participants of the Human Speechome Project, the collected data is extraordinarily personal. Deb, the father of the home, has referred to the data as the largest collection of home videos ever collected. For the child, the data is a unique record of his own early development.

Tangential to its scientific goals, HSP has brought to surface implications for how information technologies may eventually change the way we record events from our lives. Today, to augment and share our memories, we have access to collections of photographs and videos taken from the sparse set of events we think will be notable and merit documentation. As it becomes increasingly feasible to collect data anywhere at all times, we can create vastly more comprehensive records of our past that might capture unanticipated events or events that might not seem notable until long after they have passed.

Beyond the collection of data, there are implications for how such data might be accessed. HouseFly offers a unique contribution to the HSP participants as a way to review memories that is far more evocative of reliving those events than photographs or conventional video, one in which the users can travel through the scene, be immersed in it, and view details of events from new perspectives. Indeed, the participants have used HouseFly for this purpose, using it in their home to browse through personal copies made of a portion of the corpus.

We wanted to share this aspect of the project at the TED conference, and to create a video that would connect to the audience on an emotional level to show the implications for how the research might one day impact everyday life. So for the last clip of the presentation, I made a video of the child's first steps, an exciting milestone to which most parents can relate.

The video begins with an establishing shot, a familiar overhead view of the home.

The next action is to bring the viewer into the home and establish the scene and the atmosphere with greater detail.

The camera swoops into the living room, through the dining room, and into the kitchen.

The camera briefly pauses at the grandmother making dinner in the kitchen, the only other person in the home, before continuing out the door on the right and into the hallway.

The camera enters the hallway just as the father and child arrive. The child stands up, and the father beckons the child to walk towards him. "Can you do it?"

The child takes a several slow steps towards the father. He shows his excitement by whispering, "Wow," which is repeated by the father.

After a few steps, the child ends is walk and falls back down to a crawling position.

The dénouement.

The video freezes. The camera pulls out from the home, continuing until it vanishes. This was the last clip of the presentation, so this transition also served to close out discussion of the HSP project.

In this video, the ability to navigate through the home is used to show details of the scene to establish atmosphere and draw the viewer into the event. Aesthetically, it turns what might otherwise look like typical surveillance video into something less sterile and more intimate, increasing the drama of the event. The video is not the strongest example of a *data visualization*, as it shows a fairly literal depiction of a single event. But the means to access such moments is been provided by the tools developed to organize and retrieve the recorded data. The video does not just represent a parent that had a camcorder at a lucky moment, but the ability to recall any such moment that may have occurred years ago in any part of the environment. It suggests the possibility of a future in which there is little need to hold and operate a camera.

This video was the last played at the TED presentation, and was one of the videos most frequently mentioned by viewers in online discussions. The content of this discussion suggests that it successfully connected with the audience, and many viewers commenting that the clip was touching or "had me almost crying." It also successfully promoted interest in the personal implications of the research, and several viewers expressed a wish for a similar record of their own life. A more detailed account of audience feedback is provided in Section 5.

3.13 Additional Applications

HouseFly was initially developed for the HSP data, but can be used to browse other datasets that include suitable, multi-camera video recordings. Such video is commonly collected by surveillance systems used in many businesses and other facilities, which offers possibilities for the analysis of human behavior in other environments. Indeed, several HSP researchers have explored this topic, and employed the same methodologies as HSP to analyze how people utilize retail spaces, and how store layout and customer-employee interactions impact sales. Through partnerships with several companies, data was collected from multiple locations, including several banks and an electronics store. The data collected includes video, but due to the more public nature of the environments, does not include audio.

The construction of the camera and environment models required by HouseFly, as described in 3.8, requires only a few hours of effort, and was easily performed to bring data from three different retail environments into the system. A notable advantage of HouseFly is that it greatly simplifies tasks that involve following unfamiliar individuals through large and crowded environments. Even more so, tracking groups of people that enter the store together, or interactions between a customer and employee, where the participants may at times separate and occupy different areas of the store. Providing a coherent overview of the space enables the user to view the entire space without switching between cameras, and to follow complex activities at whatever distance is most convenient.

Figure 35 shows three hours of track data extracted from bank video. Here, the green tracks indicate customers and the red tracks employees. The customer-employee classification is performed automatically using a system developed by George Shaw. The classifier uses both appearance and motion features of the persons tracked. Although the employees do not wear uniforms, they adhere to a standard of dress that can be modeled as a color histogram and classified with accuracy significantly greater than chance. The paths of the employees are also distinct from those of the customers, where employees occupy certain seats more often, stand in the teller area behind the counter, and enter doors to back rooms. Using this information, Shaw's system could separate employees from customers with approximately 90% accuracy.

Browsing this set of tracks quickly reveals which areas of the stores were used more than others. Many customers used the ATM, a few used a computer console installed in the lower-left, and none perused the pamphlets to the left of the entrance shown in the lower-middle. One experiment being performed by the bank was the installation of two Microsoft Touch Tables in the lower right. Within this interface, the user can select the track data around these tables and retrieve all recordings of people in that area. In this set of data, customers sat at the tables only slightly more than the employees, and interacted with the table interface just as frequently as they used the table as a writing sur-

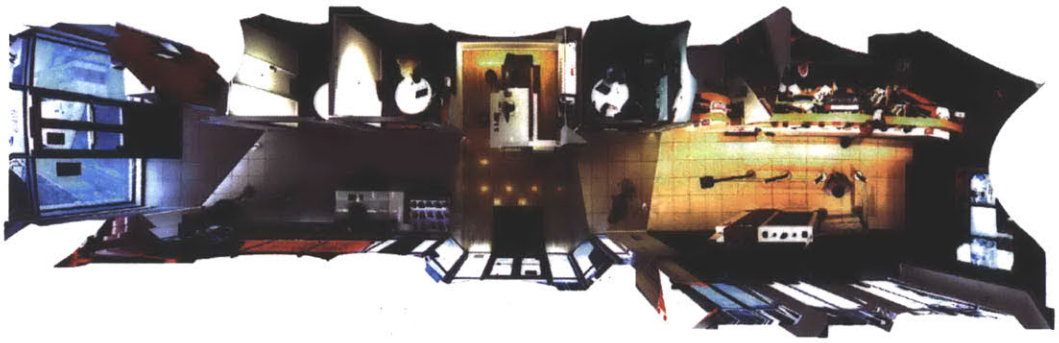


Figure 32. Bank in North Carolina recorded with 11 cameras.

Figure 33. Bank in Manhattan recorded with 20 cameras.

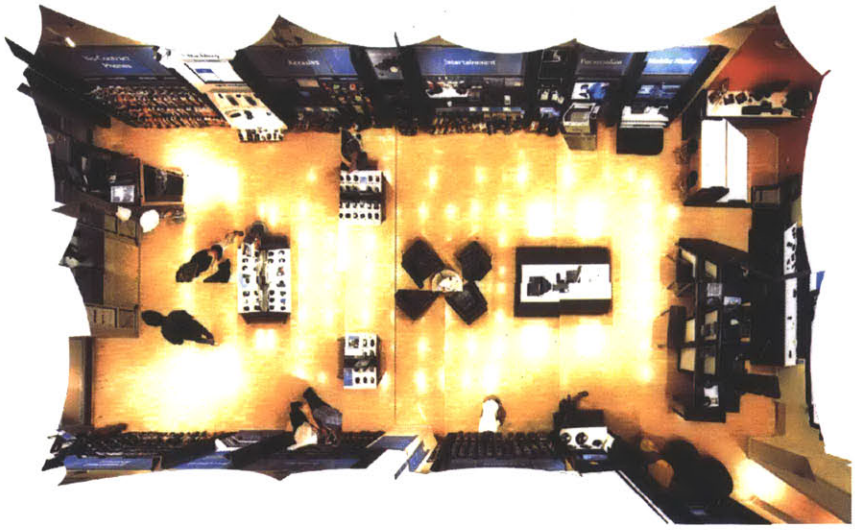


Figure 34. Electronics store recorded with 8 cameras.

Figure 35. Three hours of track data in a bank. Red lines indicate customer tracks, green indicates employees.

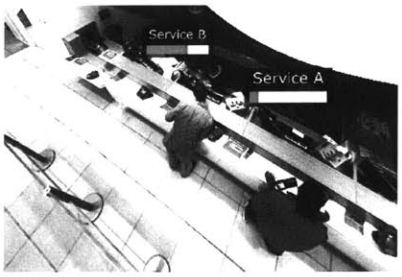


Figure 36. Customer transactions.

face or as a place to set their coffee.

Retail stores also generate electronic records of customer transactions. In a bank environment, transactions include deposits, withdrawals, and other financial events. In the electronics store, transactions consisted of items purchased by each customer, or point-of-sales data. Figure 36 shows two anonymized transactions in the bank environment, which are rendered in HouseFly as progress bars.

Much of the analysis performed involved the combined analysis of the track data with the transactions. Automatic analysis of the track data can be used to determine how long each customer waited in queue, which can then be used to model how waiting in queue affects transaction rates. One phenomenon discovered was that when customers conducted transactions over $1000, they would interact with the teller three times longer before initiating the transaction. While the process was identical regardless of the amount of money, the social interaction was substantially different.

In recent years, an industry has begun to grow around the use of video analytics for similar purposes. Several companies now offer services to count the number of customers entering a store, how long they remain, and which areas generate the most traffic. In the past, such studies were performed manually, with observers in the store recording this information on clipboards. Analysis of space utilization was performed with spaghetti plots, in which the analysts would physically lay colored string throughout the environment in order to explore traffic patterns. The rapid decline of recording costs and improvements of computer vision will likely play a pivotal role in how retailers approach store layout and design. This line of inquiry was more extensively pursued and described by my colleagues in [Rony, 2012; Shaw, 2011].

4 Social Media

At the time of this writing, the social network site Facebook has over 900 million active users. In other words, one-seventh of the world's entire population has logged into Facebook within the past month [Sengupta, 2012]. Other services like Google+, Twitter and LinkedIn also have memberships in the tens to hundreds of millions. There has been great interest in the analysis of these networks. Some of that interest is financially motivated, where the vast size and personal nature of social networks may hold lucrative new opportunities for personalized advertisement, tracking personal interests and identifying consumer trends. Other motivations include finding ways to use the networks as effective tools for political organization, disaster response, and other applications that call for rapid, mass communication. Other motivations are in the social sciences, where the vast amounts of data from these networks may reveal much about human social behavior.

The work presented so far has focused on recreating real places as simulations as a way to view data. The spatial layout of the environments and physical appearance of many objects was naturally defined by the data itself. Attempting to place the viewer inside data that is non-spatial or abstract, like that collected from a social network, presents several challenges: defining a coherent 3D space to hold the data, visually communicating what abstract data represents when it has no naturally recognizable form, and giving non-physical data a sense of presence.

What is the point of making an abstract dataset look physical? Several possibilities will be explored in this section, but I will provide one general argument here. Providing a physical representation to abstract data is the same as using a physical metaphor to explain an abstract concept. It provides the viewer with a con-

crete image of something that might otherwise be communicated only symbolically, and can thus be a powerful tool for helping to conceptualize what the data represents, facilitate reasoning through physical common sense, and to make the data familiar, relatable, and engaging.

4.1 Connecting Social Media to Mass Media

This section describes work I performed in collaboration with Bluefin Labs. Bluefin is a media analytics company founded by Deb Roy, my academic advisor, and Michael Fleischman, a former member of my research group.

Bluefin aims to analyze the relationships between mass media and social media. One of the primary objectives has been to measure audience response to television programming. Methods of audience measurement often involve soliciting viewers to participate in focus groups, to keep diaries of their viewing habits, or to use electronic devices that automatically record and send this data to the analysts. Bluefin's approach is instead to measure the *unsolicited* response of the audience by collecting and analyzing the public comments individuals post online to blogs and social network sites.

This analysis involves the construction of two very large data structures: a *mass media graph*, and a *social media graph*. For the mass media graph, dozens of television channels are continuously recorded and processed. Numerous types of data are extracted from the television content, but relevant to this discussion is the identification of every show (e.g. *Seinfeld*) and commercial (e.g. "Coca-Cola Polar Bears, Winter 2011"). Each show and commercial constitutes a node in the mass media graph, and the edges of the graph connect the shows to all commercials that played within it.

For the social media graph, public comments are collected from Twitter and Facebook. Each identified author becomes one node in the graph, and the edges of the graph represent lines of communication and connect authors that send or receive messages to

one another.

These two graphs are constructed from different sources, and the challenge remains in finding the connections between the two. Television is a popular conversation topic and generates millions of comments on social media sites every day. Each time an author writes about a piece of televised content, he constructs a referential link between himself and the content. These links provide a connective web between the social and mass media graphs such that the authors are not only connected to those they communicate with, but also to some of the things they communicate about.

For humans, such links are easy to find. We are adept at dereferencing natural language and linking speech to objects. But finding these links at scale is a large endeavor that requires parsing billions of comments and linking them to millions of audio-video events, requiring computer systems that can both parse natural language and identify television content. The payoff is that the resulting synthesis of the two graphs can reveal a wealth of unsolicited feedback about TV programming, the effectiveness of advertisement campaigns, the television viewing habits of individuals, and the dynamics of shared conversation topics across social groups.

I worked with Bluefin to create a visualization as a way to efficiently explain the approach of this analysis and to illustrate this relationship between social and mass media. This visualization was intended for a general audience, and was presented at the TED conference amongst other venues. Mass media and social media are both abstract and nebulous networks of information, and one of the goals was to provide an image of the two networks that was concrete and easy to conceptualize. A second, editorial-minded goal was to make the visualization evocative of the scale and complexity of the data.

Figure 37 shows an early iteration of a data browsing interface that uses a standard network diagram representation of dots and lines. This image shows a small subset of the data pertaining to a single television show, *Supernatural*, and its audience. The show is represented by the red dot in the center. All the authors that

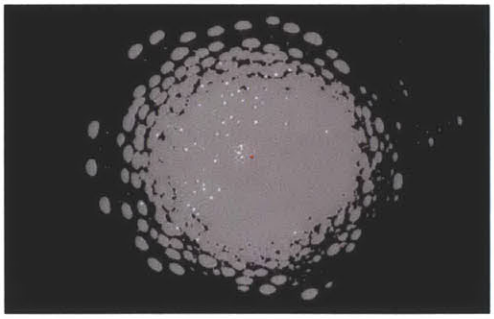


Figure 37. Graph of people commenting on the television show *Supernatural* on Twitter.

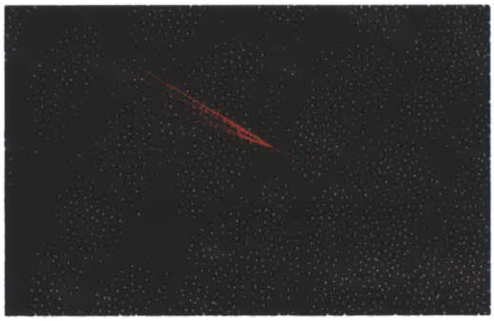


Figure 38. Detail of the graph, highlighting one author and connected followers.

Figure 39. Pieter Bruegel. *The Tower of Babel.* c. 1563.

have written about the show are drawn as white dots, and all the authors that have not written about the show but follow someone who does are drawn in gray. In total, there are about 51,000 nodes and 81,000 edges to this graph. The nodes have been organized using the implementation of the *sfpd* algorithm provided by GraphViz [Ellson, 2003].

The representation of a graph as a dots and lines on a plane is a conventional approach. It can be highly functional and useful for analysis, but is not always the most exciting. Here, the records of over 50,000 individuals have been reduced to a mathematical abstraction that says little about the scale or nature of the data itself. There are enough dots to saturate the image, yet it does not provide a visual impression of being anything massive or impressive, or of being anything at all.

Like most network diagrams, this graphic provides no sense of scale. The issue is not in communicating a quantitative scale, of which there is none, but to provide the qualitative feeling of scale that one gets when entering a cathedral, or even looking at a vividly rendered naturalistic image of a large space, as in Figure 39. This is an issue of presence. Creating a representation that looks and feels like an actual place must give the viewer a sense of absolute scale and communicate how the viewer relates to the environment physically.

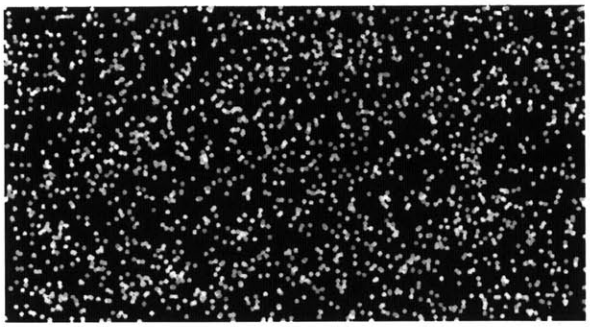
Painters have long known techniques to make paintings look large, but it is an interesting exercise to revisit those techniques to identify the minimum number of details that must be added to make abstract data appear large. Our perception of size and depth is well studied in terms of the perception of different depth cues and the interpretation of those cues to build a mental model of a spatial environment. The categorization of depth cues is not consistent across literature, but the table on the following page provides descriptions of 12 established cues, collected from a survey, [Cutting, 1997], and several additional sources.

This document will not discuss all of these in detail, but the table is provided to define terms as I present an example of using these cues to establish a sense of space.

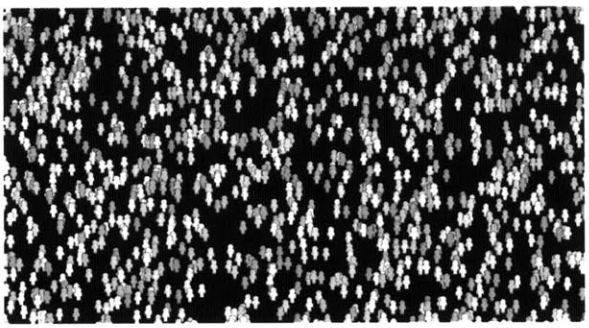
Cues for Depth and Size Perception

Source	Description	Information Provided (Physically possible, although not necessary perceived by humans!)
	Relative Sources	
Occlusion or Interposition	Closer objects block more distant objects from sight, indicating the order of depth.	Ordinal
Relative Size	The difference in apparent size between similar objects indicates ratio of distance from viewer.	Ratio
Relative Density or Texture Gradient	The scaling of textures or object groupings according to distance.	Ratio
	Absolute Sources	
Familiar Size	If the absolute size of an object is known, its apparent size indicates absolute distance.	Absolute if size is known.
Motion Perspective	Motion perspective includes *motion parallax* and *radial outflow*. Motion parallax is the apparent speed of an object in motion relative to the viewer or another object. Radial outflow is the areal scaling of an object is its distance to the viewer changes [Gomer, 2009].	Absolute if velocity is known. Otherwise, a ratio.
Elevation	If the viewer is positioned over a ground plane, the base of objects resting on that plane indicates distance, and closer objects will have a lower position in the visual field.	Absolute if the eye height is known. Otherwise, a ratio.
Aerial Perspective	Viewing objects through mediums that are not completely transparent, including air, causes distant objects to appear desaturated.	Absolute if opacity of medium is known.
Defocus Blur	For optical lens systems, the amount by which an object is blurred indicates its distance from the focal plane. Further, depth of field is shallower for near focus [Mather, 1996].	Absolute if lens characteristics and aperture are known.
Object Dynamics	Knowledge of dynamical systems can provide depth information from the apparent speed, acceleration, or other motion of objects. For example, the apparent acceleration of an object in free fall can indicate absolute distance [Hecht, 1996]. These cues are not yet well defined and less researched than the others.	Absolute if dynamical model is known.
	Not produced by 2D displays	
Stereopsis	The displacement between apparent positions of an object as seen by the left and right eye.	Absolute, but requires stereoscopic display.
Convergence	Fixating both eyes on a near object causes them to point inward, where the angle provides depth information.	Absolute, but requires stereoscopic display.
Accommodation	The ocularmotor flexing of the eye's lens to bring an object into focus provides a sense of depth.	Absolute, but requires holographic display.

Information from [Cutting, 1997], except where cited otherwise.

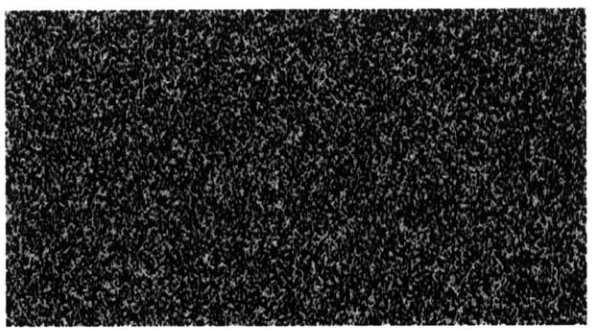


Beginning again with a set of dots, each represents a single author, and have been arranged and colored randomly. The only depth cue is occlusion. The scale of the image might be anything, from microns to light-years.



I first tried to replace the dots with icons of people, objects of familiar size. The viewer might now guess a rough scale (my office mate estimated it to be the size of a soccer field), although the representation is still quite flat.

The icons could be made more realistic or changed to photos, but even real physical objects do not evoke a strong sense of scale when removed from other depth cues. Experiments on this subject have involved showing objects like playing cards, to participants under restricted viewing conditions that removed other depth cues and asking the participants to judge the distance and size of those objects. The results revealed that when shown a normal-sized card five meters away, participants were more likely to report seeing an unusually small card at around two meters away [Predebon, 1992; Gogol, 1987]. In the absence of other information, our visual perception will readily disregard our knowledge of object size.

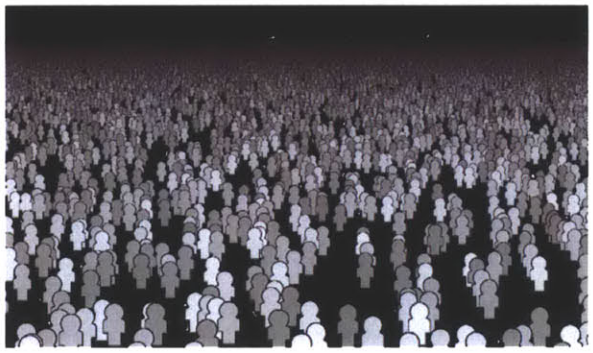


A further problem of using a 2D layout is that regardless of the representation chosen for the authors, attempting to show more than a few thousand will result in an indiscernible texture.

Positioning the icons within a volume appear more spatial, but provides only a relative depth through contrast in size and texture and occlusions.

Navigating within this 3D scene provides motion perspective information, but without a clear sense of the viewer's own velocity or size, the absolute scale remains unresolved.

Aerial perspective, or fog, might indicate absolute depth if the density is known, but even so, we are poor at using it to judge depth. In most cases, including the image on the left, it primarily provides an ordinal measure of depth. In this image, however, it does help in improving the visual contrast between and far and near objects and reinforcing the sense of volumetric space.



In this image, the volumetric layout has been abandoned and the icons have been return to a plane, but now viewed from a lower angle. The result is an image that provides a much more vivid sense of a large space containing a vast number of people.

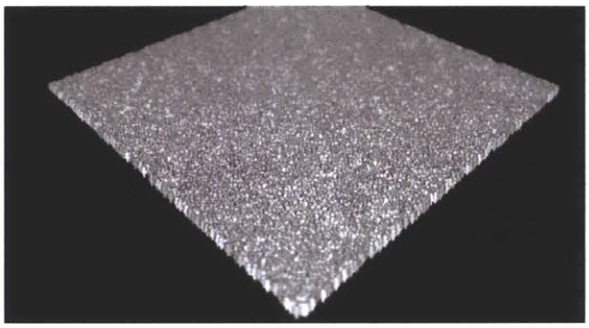
The planar configuration accomplishes two things. First, it establishes a linear perspective. Linear perspective is not, in itself, a source of depth information, but a system of interpretation. It combines multiple sources of information and resolves the constraints and ambiguities between them to produce a spatial model of a scene. In particular, linear perspective largely relies on the heuristic of interpreting elements that are apparently collinear in the retinotopic image as also being collinear in physical space. Here, increasing the colinearity of the icons greatly reduces the perceived ambiguity of the scale and depth of each.

Second, the icons now create an implicit ground surface, which provides an organizing structure that helps to define the space. With the lower camera angle, the ground surface establishes a horizon line and elevation cues, and also creates a strong texture gradient that extends from foreground to

background. This information further reinforces the linear perspective.

This is a complex way of saying that to create a sense of absolute space, it helps to have a ground surface. The different techniques of providing depth information are important, but any cue is likely to be ambiguous without an organizing structure.

The use of a ground surface is not limited to large spaces, and might be instrumental in defining a space of any scale. At left is an image of the same icons, but grouped more closely together, without aerial perspective, and viewed from a higher vantage point. The image does not look tiny, but substantially less expansive than the previous image.

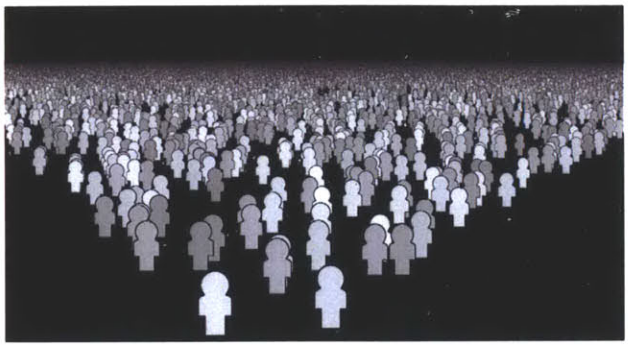


Defocus blurring can emulate the depth-of-field created by optical lens systems, including the human eye. Shallow DOFs emulate the focus on near objects, and when added to an aerial photograph or other large scene using a tilt-shift camera or digital manipulation, can sometimes produce a striking miniaturization effect. Here, without the richer scale cues provided by naturalistic imagery, the effect is still present, but less pronounced.

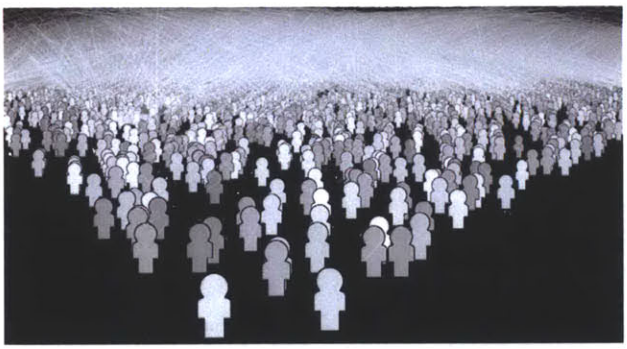
For interactive visualizations, defocus blurring is best avoided as it predetermines what the viewer may focus on. In passive mediums, like photography and cinema, we are more accepting of having our focus guided, and even appreciate the bokeh of a photograph or the way a narrowly-focused movie scene lifts the actors out of the background. In interactive mediums, the user is more likely to want to control what he sees and to explore different parts of a scene, where the inability to adjust his focus may be an annoyance. This is demonstrated by video games, where the emulation of DOF has recently become an easily achievable effect and is now a feature of many popular engines [Hillaire, 2008]. While it is still too early to judge its impact, initial opinion appears predominately negative.

4.2 Social Network Visualization

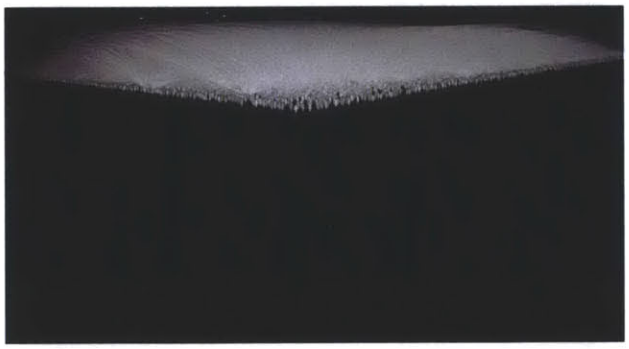
After achieving the desired effect of scale, the system was then used to produce the following visualization.



The visualization opens with a shot of the authors, arranged as discussed.



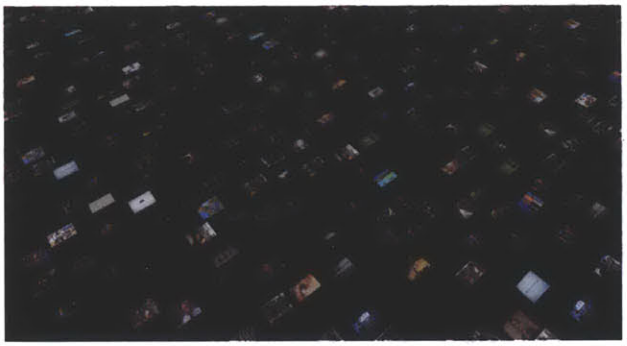
Lines are added that connect authors that communicate to each other, revealing an intricate social graph.



The camera pulls back to reveal more of the network.

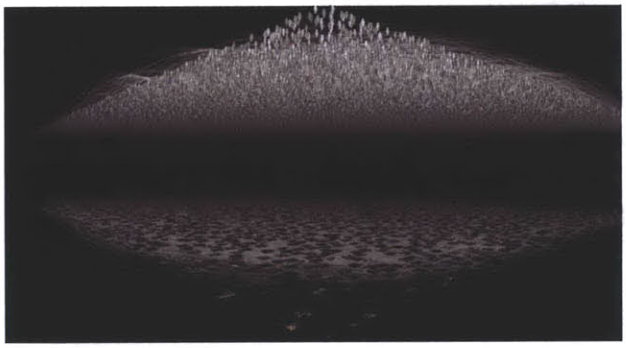
The motion of the camera includes significant lead-in and lead-out acceleration to emulate physical inertia. This helps significantly to maintain a sense of a physical space and navigation.

Nodes representing television shows and commercials are then added to the scene, organized on a new plane below the authors. Each node is shown as video panel, which provides a large amount of visual activity and excitement to the scene.



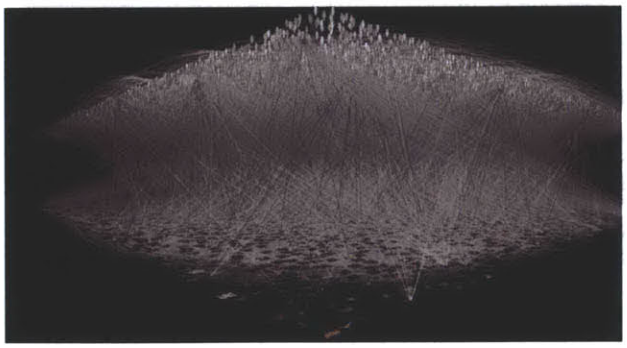
The camera moves closer to the mass media nodes to set up the next beat.

Edges are added that connect commercials to the television shows in which they aired, revealing the mass media graph.



The camera pulls out, showing the two media graphs.

The two graphs are shown as separate, but parallel, planes, preparing the audience for a third dimension.



The two graphs are then connected. Each line connects an author to a show or commercial that the author has written about.

This image represents one of the main points of the visualization: to provide a conceptual bridge between mass media and social media that invites new inferences.

All of the lines are removed. Now that the data-set has been explained at a distance, the next few beats present examples of types of patterns found in the connected graphs.

The first pattern begins with a single author that has written about a show, illustrated by a bright line that extends from the show to the author.

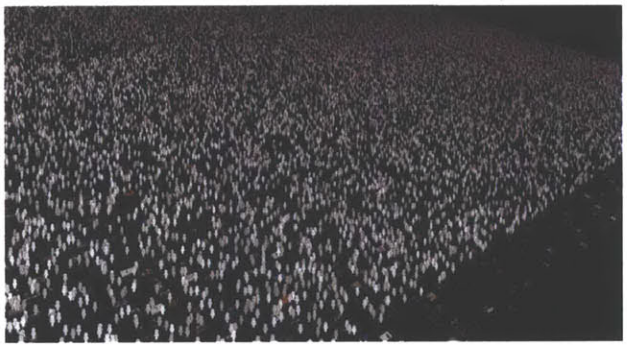
Lines then extend from the one author to the other authors that received those comments.

More lines extend between authors, showing a small network of people that communicate with one another.

These authors also write about the same show, and form a *co-viewing clique* that communicates about a shared interest.

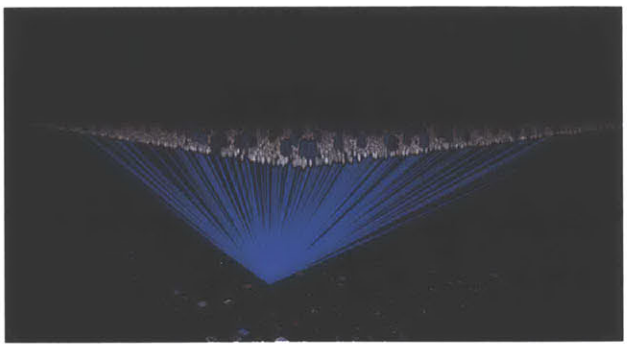
The lines are removed, and the visualization moves to the second pattern. Here, many lines shoot out from a single author creating a firework effect, showing an amateur critic that comments on many shows and is read by many people.

In one version, we attempted to make the animation of the lines less sudden, slowing it down and using multiple build ins. However, test audiences responded well to the more dramatic explosion of lines, and the effect was retained.

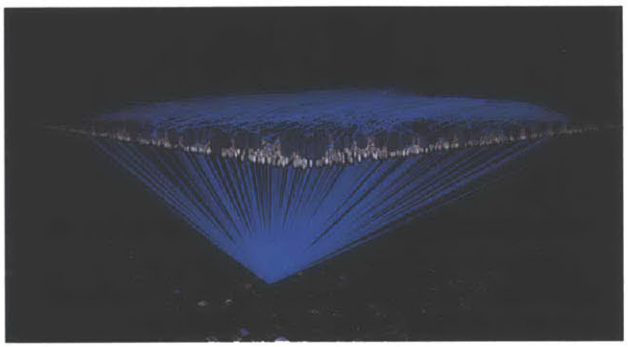


The third pattern looks at a specific television show. The camera moves forward and dives through the crowd.

The camera stops on a video showing a recent State of the Union address, and holds for a few seconds. This is an event that generates a huge number of comments.



As the camera returns to the side, lines shoot from the State of the Union to thousands of authors.



And then all the authors that have received comments about the State of Union are highlighted.

The visualization closes by showing some of the most salient phrases used in this discussion. Although the video does not delve any deeper into the content analysis, this shot is meant to raise the topic for further discussion.

This visualization shows two very different datasets, explains what each dataset represents, how they are connected, and a few of the inferences that might be drawn from those connections. The visualization is far from naturalistic, but provides sufficient cues to establish a strong sense of scale and space, resulting in a visualization that is more engaging and evocative than the earlier 2D version shown in Figure 37. It also provides a more cinematic approach to animation, which is used to connect different viewpoints of the scene, from distant shots that show the large portions of the graphs, to very close shots that show only a single node.

4.3 News Television

For many news events, public reaction is an essential part of the story, and news television networks are increasingly turning to social media as a way to gauge this reaction. The relationship between news and social media is still being defined, and studios continue to develop effective practices for using social media for journalistic purposes. This section discusses the development of a visualization for news television intended to report on public reaction within social media, and the design considerations involved when creating visualizations for a medium like television.

In late 2011, one of the largest media events in America was coverage of the GOP presidential primaries, wherein seven of the candidates running for the Republican Party nomination participated in a series of televised debates. As coverage of an election process, public reaction to the debates was a primary focus, and ABC News wanted to air a segment analyzing social media response to a debate being held on December 10th. Producers from ABC had seen the social network visualization created for TED, discussed in Section 4.2, and thought that it might adapt well to television. And so in collaboration with Bluefin Labs and Isabel Meirelles, I extended the visualization and produced a two-minute segment to be played the morning after the debate on *This Week With Christiane Amanpour.* The data involved would be just a single piece of televised content, the recording of the debate itself, and all the Twitter comments it generated.

Most data visualization literature focuses on design for print, projector, or computer display, and draws from the established design practices of those fields. Little is mentioned of visualization design for television, and few visualizations are shown on television beyond basic charts that report political polls, stock prices, or one-dimensional product comparisons. There are several reasons why television is not an ideal medium for data visualization, but also compelling possibilities for using television to reach large audiences and strengthen the use of empirical analysis in popular discourse.

One of the primary limitations of television is the picture qual-

ity. Most television sets have a low native resolution, vary greatly in size and aspect ratio, and are usually viewed from a distance. Furthermore, they are better optimized to display naturalistic images, like film and photographs, and poorly optimized to displaying high-contrast edges, as with text, and fine lines. Designing an effective graphics for television involves reducing text where possible, giving a large amount of space to every element that the viewer must see clearly, and avoiding complicated layouts that divide the space of the screen. This is antithetical to visualization design for print, where it is possible to present intricate information within a single view, and to allow the viewer to look over it from up close.

A second limitation is that television does not allow the viewer to examine things at his own pace, or control the flow of the presentation. For many programs, viewers are not expected to even look at the television much at all. The segment I created was targeted to play on a Sunday morning just after the debate, a weekend when many viewers might actually sit down to watch the morning news. The proposal of creating a segment for a weekday was considered impractical, because the audience was expected to be preparing for work and might only glance at the television occasionally, largely undermining the purpose of airing a visualization.

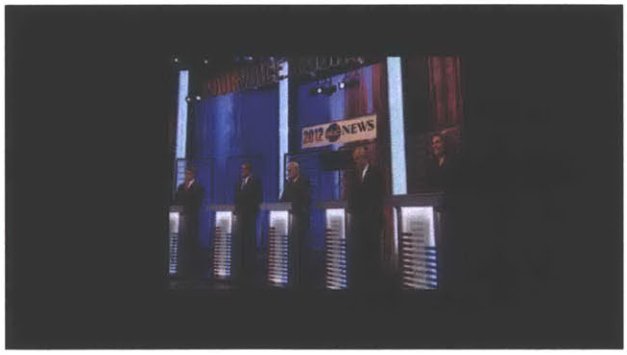
These issues are not unique to television. Image quality and resolution are significant issues whenever showing graphics on a distant screen, as when presenting with a projector. The issues of pacing and passive communication are present whenever communicating to many people at once. Edward Tufte, one of the standard bearers of information design, has also described this as problematic, and has argued that providing a paper handouts prior to presentations can give the audience "...one mode of information that allows *them* to control the order and pace of learning" [Tufte, 2007]. But handouts are not always practical.

When trying to visually communicate complex ideas in these circumstances, it may be necessary to accept the limitations of what the medium can show within a single view, and compensate by taking advantage of what can be shown in sequence at lower

resolution. The use of 3D animation provides several ways to do this. Camera movements can be used to briskly bring the viewer from one view to the next. Animation can communicate causal and process information that would otherwise call for a textual explanation. The third reason is that animation provides additional visual information through motion perspective that can significantly help perception when pushing against the limits of a low resolution display. Objects that appear as a small smudge in a static image are sometimes easy to identify in motion.

4.4 Visualization of the GOP Debate

In the visualization that follows, video and social media data was provided by Bluefin Labs and Twitter. Topic analysis of the Twitter data was performed by Mathew Miller. Editorial focus, caption writing, and transcript authoring was performed by Russell Stevens and Tom Thai. Deb Roy and Isabel Mierelles provided design input. I was the primary designer and producer of the visualization itself. The debate ended at 10pm on December 10, 2011. The complete video was sent to the ABC News team, who trimmed the video down, revised the transcript, and approximated 12 hours after the debate ended, aired the segment.



Establish subject

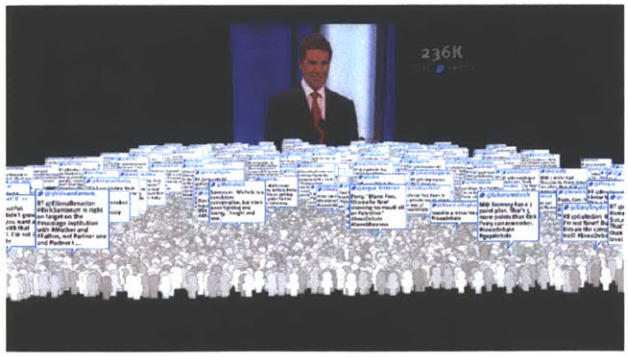
The opening shot establishes the subject, the GOP Debate, shown as video footage provided by ABC News.

Introduce a single point of data

The next beat introduces a single comment made about the debate, using the same author icons as the social graph visualization.

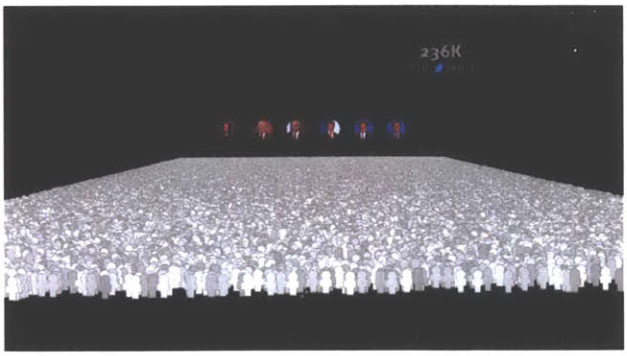
Introduce rest of data

The camera pulls back as the rest of the comments are added to the screen. 236,000 comments were identified, however, only 70,000 are shown in the scene so that the icons would not become overly small. The analysis to be shown is accurate for the entire dataset.



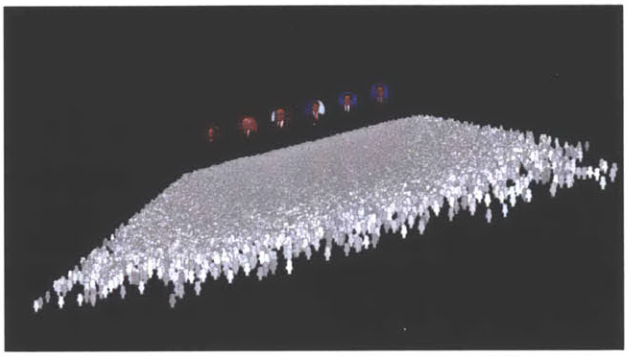
Summarize setting

The camera holds in this position for a moment, creating an image of tens of thousands of people watching the debate and offering their comments on it. This provides a literal, visual explanation of what the data represents.



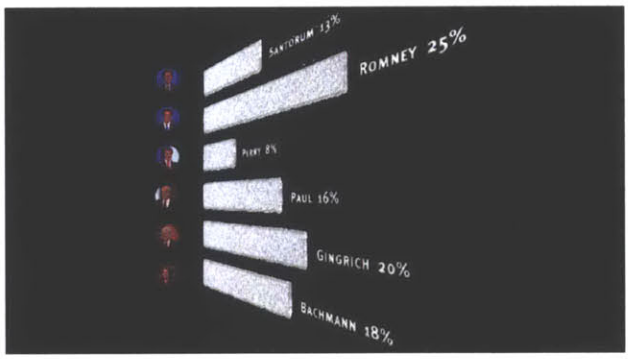
Introduce candidates

The giant screen builds out, and icons of each of the candidates in the debate builds in.



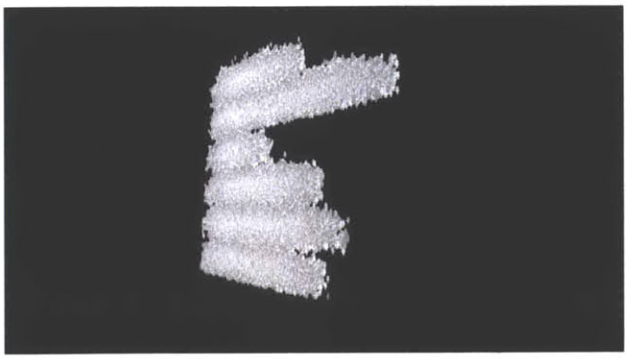
Transition to next view

The authors begin moving, creating a moment of intense visual activity. Instead of just cutting to the next view, the transition is animated to show that this new view uses the same data and to maintain continuity.

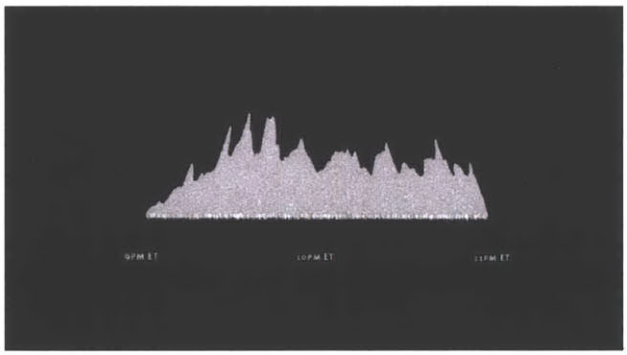


Show the volume of comments about each candidate

The comments organize themselves into a bar chart showing the amount of discussion about each candidate. Mitt Romney generated the most comments and Rick Perry the least.

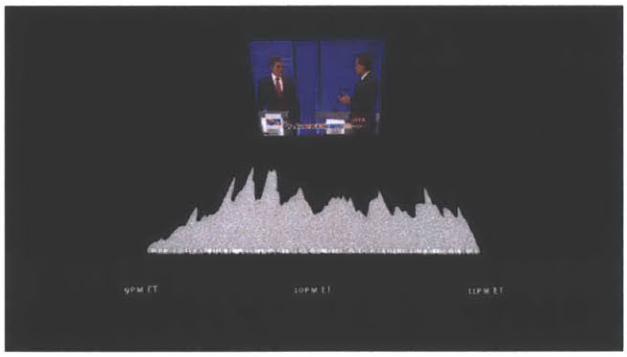


Transition to next view



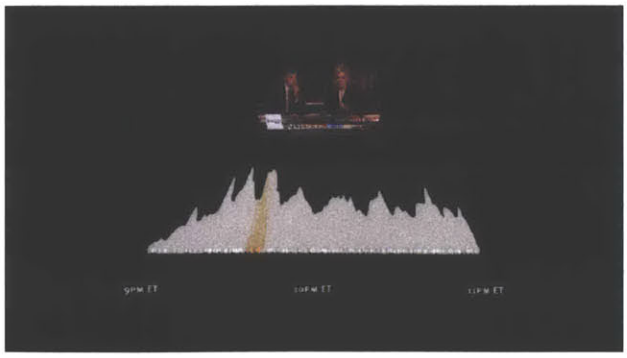
Show volume of comments over time

The comments reorganize themselves into an area plot that shows the volume of comments received throughout the debate. The peaks of this graph indicate notable events that drove discussion, with the largest peaks generated between a half-hour and an hour into the debate.



Introduce event

The screen re-emerges, showing the single event of the debate that generated the greatest number of comments. After Perry criticizes statements made by Romney in a book, Romney extends his arm and offers to bet Perry $10,000 that those statements were never made.

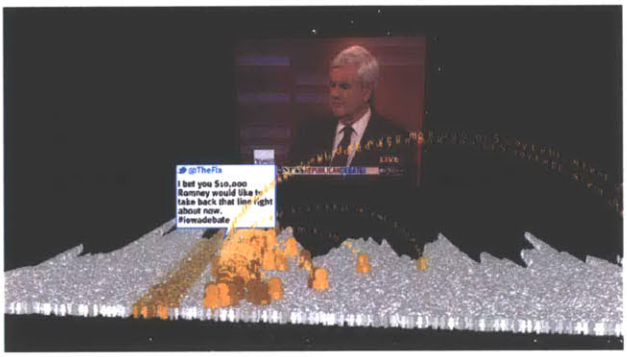


Show response to event

A number of comments are highlighted, showing the social media response to this event.

Show detail

A specific comment about the event is shown. This comment was representative of the overall opinion, where most regarded Romney's bet as a gaffe that made him appear as a rich, frivolous with his money, and disconnected from the working class.



Show propagation of comment

The comment shown was *retweeted*, or posted again by other authors, many times throughout the event. The text of the comment streams out from the original comment to each of the retweeted versions, showing its pattern of propagation.

A primary contribution of this visualization is the way it connects different views of the data using a persistent representation. The comments made about the debate are introduced once, explained through a metaphor of a large crowd of people watching the debate. The comments are then rearranged to bring the viewer through different aspects of the public response, including overall volume, response to individual candidates, the response over time, and response to a single event within the debate.

While the visualization shows the volume of comments about a single event, it does not go into further detail about who responded or what was actually written. The software that was developed provides several views that may have addressed this and offered a more detailed analysis, but were ultimately left out of the final video due to editorial decisions.

Figure 40 is taken from an earlier draft of the visualization, using data from a previous debate. In this debate, Perry accused Romney of hiring illegal immigrants to work on his property. The people icons at the bottom of the image represent the comments

Figure 40. Sentiment breakdown of comments about Perry's accusation that Romney hired illegal immigrants to work on his property.

about this specific event. The icons are colored red, white, or green to indicate if the coment expressed a negative, neutral, or positive sentiment. In this instance, the image shows that the general sentiment towards this event was slightly more negative than positive. Another unused view organized the comments by demographic groups, showing the breakdown of comments according to author gender, age, and interests.

5 Critique

With the exception of the GOP Debate visualization discussed in Section 4.4, all of the videos described in this document were used in a 20-minute presentation at TED 2011 delivered by Deb Roy. This video was posted online shortly after the presentation, and at the time of this writing, has been viewed over 1.5 million times on the TED website. The video is also been viewed on YouTube, and has been used as in-flight entertainment by Virgin Airlines. Subsequently, the presentation has generated thousands of comments online from unsolicited viewers. While Roy was responsible for constructing and delivering the presentation, I was the primary creator of most of the graphics, including all of the video content, and much of the feedback referred explicitly to the visualizations.

This kind of feedback is often more useful in qualitative assessment than quantitative, but to provide a rough sense of audience response, a portion of the comments were coded for sentiment and tallied. 299 comments were collected from the TED website, not including one comment in a foreign language and two comments made involved researchers. Of these, 62 comments expressed an opinion specifically about the visualizations and graphics. I believe this is a relatively high fraction considering that many comments did not containing any specific details on the talk, and that the talk was not about data visualization and mentioned the design of the visualizations used only briefly. Each comment was hand coded as expressive a positive, negative, or neutral sentiment. Of these comments, 5 were negative, 57 were positive, and 0 were neutral.

Informativeness is one of the key attributes for which most visualizations strive. As Edward Tufte describes, "Excellence in statistical graphics consists of complex ideas communicated with

clarity, precision, and efficiency" [Tufte, 1986]. Several commenters thought that the visualizations failed in this regard:

> "Actually, I found many of the visualizations more distracting than clarifying, especially the social media ones. Lots of little TV screens in a grid, flying through space... uh..."

Conversely:

> "The data visualizations in this presentation are very impressive. They manage to provide overwhelmingly complex ideas and data in an easily interpretable format."

> "I'm a software engineer. I was staggered by the level, detail and complexity of the information and analyses that he has displayed without batting an eyelid. Why are 'fancy graphs' important? Because it helps people like you understand complex information :)"

Several viewers were specifically impressed by the sense of immersion created by the visualizations:

> "While designed to monitor his son's development, his computer system ended up giving him an unparalleled glimpse into his own life and that of his family. He can literally search through footage using spoken words and behaviors. Using multiple angles and simulation software, he can virtually live through his past experiences in the first person!"

> "Never come across something so powerful that almost gets us back in time... fantastic stuff."

Comments on the aesthetics and production of the visualizations were almost uniformly positive. However, a few of the viewers thought that the visualizations were designed to hype research that would otherwise be poor or uninteresting:

> "Data visualizations are most useful when they help peo-

ple understand complex information. When they are used to make pretty standard observations look like 'expensive' research, they become dangerous."

The comments collected did not provide critiques much more detailed than that shown. However, in general, opinion was very positive and often enthusiastic. Viewers found the visualizations informative, immersive, and technically and visually impressive.

Many comments did not discuss the visualization work explicitly, but imply that they may have achieved their goals in making the research interesting and relatable:

> "As I said on Twitter last week, Deb Roy's talk at this year's TED was among my favorites ever. Its mixture of science, data, visualization, and personal story touched all my hot buttons, and touched me personally."

> "There's no doubt that Mr. Roy's approach to researching the development of his son's language is, at first glance, a bit creepy. Document every waking hour of your family's life using an array of ceiling-mounted cameras all over your house? Yep, creepy. ... But as Mr. Roy and MIT's work is demonstrating, the ability to record everything, archive it, analyze it and share it with others can have the most wonderful, human and un-creepiest results."

The most frequent criticism on the presentation, appearing in 40 comments on the TED website, was that the results were disappointing or too obvious:

> "This was so disappointing. A year of recording audio and video, significant time analyzing, tons of money and technology – and all we learn is that 'water' is mostly spoken in the kitchen and a few other obvious tidbits?"

I cannot take responsibility for much of the presentation, but will offer a response. In a short 20-minute presentation, it can be difficult to describe the results of several years of linguistics analysis in great detail. The visualizations that were created were focused

more on explaining the data, the methodologies developed, and the potential impact of the research, which we felt would have broad relevance to a general audience.

HSP has produced a number of scientific findings that we were not able to fully disseminated at TED. For example, the three caregivers of the household – the father, wife, and nanny – continuously adjusted the complexity of their utterances in the presence of the child in a way that seems designed to help him learn language to a surprisingly and previously unobserved degree. Given the difficulty of tracking exactly which words the child does and does not know at a given moment and taking that knowledge into account each time they spoke to the child, a reasonable interpretation is that caregivers subconsciously tracked the child's receptive vocabulary and predictively tuned their language to serve as linguistic scaffolding.

A second point is that even very obvious things require empirical observation to model scientifically. To use the example of the critic, it is quite expected that *water* would be spoken most frequently in the kitchen. However, measuring the precise frequency empirically, and being able to compare that to the frequency of other word types, is surprisingly difficult. What the wordscape visualization (Section 3.11) shows is that we have developed a way to collect such data, and have verified that this data conforms to what we might expect. This is, in itself, an important step in building scientific models of linguistic development.

In working with this data further, my colleagues have discovered a surprisingly strong influence of non-linguistic social and physical context – what is happening, where, and when – in predicting the order in which the child learned his first words. By combining linguistic factors, such as the frequency or prosody of words heard by the child, with non-linguistic context, they have been able to create the most precise predictive model of word learning ever created for a given child [Miller, 2011].

6 Conclusions

This dissertation has presented a body of work that has utilized the first person and 3D graphics to address challenges of viewing, navigating, analyzing, and communicating information embedded in big, heterogeneous data sets. The datasets used in this work included a variety of data collected for real world applications. And while the design of each visualization was tailored to the specifics of each dataset, each relied on the same generalizable approach of placing the viewer inside the data. Many aspects of the first person viewpoint and its implications have not been deeply explored previously, and this document makes several contributions to this area:

An approach to the first person viewpoint that encompasses the notion of presence and of creating a sense of physical engagement through visual perception. Where previous work has focused more on navigation schemas and immersive display technologies, this dissertation has extended the idea that many aspects of the first person can also play a significant role in visualizations presented on 2D displays, or that may not even be interactive. Not all the works in this thesis produced as strong a sense of first person engagement as video games, virtual reality rigs, or actual physical environments, but our sense of immersion does not need to be overwhelming to have an impact.

Methods of visualizing complex datasets as simulated environments to facilitate intuitive, spatio-temporal perception and navigation. The HouseFly system incorporates many previously established techniques, but as a whole, no simulation of a real environment has been created previously with a similar level of spatial detail and temporal depth. HouseFly presents multiple sources of data in a way that immediately reveals the environment as a whole and enables users to identify and follow activities seamlessly across

multiple sensors, levels of detail, and time. It also demonstrates a unique approach to retrieval that combines spatial, temporal, text, and annotation based queries.

An approach to data storytelling that leverages the use of 3D graphics to compose and sequence shots in a more cinematic manner, including similar techniques for establishing context and subject matter, focusing viewer attention, and explaining relationships between different views of the data. The use of cinematic techniques in data visualization has been discussed previously, e.g. [Gershon, 2001], but clear examples of the approach are still uncommon, with substantial room left for exploration.

An approach to creating more engaging visualizations by placing the viewer inside the data. This dissertation has examined how the first person can provide a vivid sense of being in a physical scene, provide novel perspectives and visual excitement, and, as discussed in the visualization of a child's first steps in Section 3.12, even help to establish a more personal connection with the data.

The evaluation of the thesis work has focused on the navigation of complex datasets, clarity of communication, and ability to present data in a way that provides meaning to the data and promotes engagement. These are significant goals in both research and communication. However, with regards to applying first person interfaces for analysis tasks, this work is still in an exploratory stage. The research of both HSP and Bluefin Labs involves the application of novel methodologies and technologies at very large and challenging scales. Much of the effort behind this dissertation has been focused on developing methods of collecting data and attempting to uncover the new forms of analysis this data makes possible. As applications for these datasets become more clear, future work will be required to identify specific analysis tasks that call for optimization, and to evaluate the performance of the techniques discussed quantitatively.

Further effort will also be required to reduce the labor and expertise required to create such graphics, and to develop better software tools. Creating 3D interfaces for data visualization currently requires significant ability in software development, as

well as specialized knowledge of graphics hardware, algorithms, and software libraries. For designers without such experience, the approach to visualization discussed here may be difficult or prohibitively expensive to replicate. Many of the visualization frameworks available are limited to producing standard plots and 2D graphics. The *Processing* language, developed by Ben Fry and Casey Reas, is a notable exception, and provides a simplified interface to OpenGL and other libraries that enable novice developers to more easily produce 3D graphics, typography, and multimedia content [Fry, 2004]. However, Processing is still closer to a language than a visualization engine, a simplified dialect of Java with additional libraries, and does not include many of the higher level tools required for highly functional 3D interfaces. A short list of desirable tools might include a system for asset and scene-graph management, scripting and animation, unified geometry collision and picking, a flexible renderer that facilitates procedurally generated graphics, and a GUI library that integrates both 2D and 3D interface components. Still, any software framework that integrates these tools would only mitigate the effort of software development. The result would resemble a 3D video game engine, which still require significant expertise and learning to use effectively. Making 3D visualization truly accessible to non-programmers will require more radical developments in tool design.

I do not claim that the approach to visualization argued for in this document is appropriate for all applications, or even most. Creating a 3D interface to visualize sales figures at a financial review meeting would be unlikely to illuminate the data any better than a simple line plot, and an extravagant waste of effort. But as we encounter new and increasingly massive datasets, there is greater need to understand these datasets as complex systems and to view them from many perspectives. This dissertation has shown how placing the viewer inside the data may achieve this goal, and in the process, to produce graphics that show something new, insightful, and beautiful.

6 Citations

Anguelov, Dragomir, Carole Dulong, Daniel Filip, Christian Frueh, Stéphane Lafon, Richard Lyon, Abhijit Ogale, Luc Vincent, and Josh Weaver. (2010) "Google Street View: Capturing the World at Street Level." Computer. 43:6, pp 32–38.

Botev, Z.I., J.F. Grotowski and D. P. Kroese. (2010) "Kernel Density Estimation Via Diffusion." *Annals of Statistics*, 38:5, pp 2916–2957.

Card, Stuart K., Jock D. Mackinlay and Ben Shneiderman. (1999) *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann Publishers.

"CERN: Let the Number-Crunching Begin: The Worldwide LHC Computing Grid Celebrates First Data." Web. Retrieved September 20, 2012. <interactions.org>

Chomsky, Noam. (1980) *Rules and Representations*. Columbia University Press.

Daniel, G. and M. Chen. (2003) "Video Visualization." in Robert Moorhead, Greg Turk, Jarke J. van Wijk (eds.) *IEEE Visualization 2003*. pp 409–416. IEEE Press. Seattle, Washington.

DeCamp, Philip. (2007) "HeadLock: Wide-Range Head Pose Estimation for Low Resolution Video." M.Sc in Media Arts and Sciences Thesis. MIT.

Ellson, J., E.R. Gansner, E. Koutsofios, S.C. Gansner, E.R. Koutsofios, S.C. North and G. Woodhull. (2003) "Graphviz and Dynagraph – Static and Dynamic Graph Drawing Tools." in M. Junger and P. Mutzel (eds.) *Graph Drawing Software*. pp 127–148. Springer-Verlag.

Foulke, Emerson and Thomas G. Sticht. "Review of Research on the Intelligibility and Comprehension of Accelerated Speech." *Psychological Bulletin*, 72:1, pp 50–62.

Fry, Ben. (2004) "Computational Information Design." PhD in Media Arts and Sciences Thesis. MIT.

Gogel, W. C. and J. A. Da Silva. (1987) "Familiar Size and the Theory of Off-sized Perceptions." *Perception & Psychophysics*. 41, pp 318–328.

Gomer, Joshua A., Coleman H. Dash, Kristin S. Moore and Christopher C. Pagano. (2009) "Using Radial Outflow to Provide Depth Information During Teleoperation." Presence. 18:4, pp 304–320.

Harbert, Tam. (2012) "Can Computers Predict Trial Outcomes from Big Data?" *Law Technology News*. July 3, 2012. Web. Retrieved September 20, 2012.

Hecht, Heiko, Mary K. Kaiser, and Martin S. Banks. (1996) "Gravitational Acceleration as a Cue for Absolute Size and Distance?" *Perception & Psychophysics*. 58:7, pp 1066–1075.

Hejna, Don and Bruce R. Musicus. (1991) "The SOLAFS Time-Scale Modification Algorithm." BBN Technical Report.

Hillaire, Sébastien, Anatole Lécuyer, Rémi Cozot and Géry Casiez. (2008) "Depth-of-Field Blur Effects for First-Person Navigation in Virtual Environments." *IEEE Computer Graphics and Applications*. 28:6, pp 47–55.

Johnson, Mark. (1990) *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason*. University of Chicago Press.

Kapler, Thomas and William Wright. (2004) "GeoTime Information Visualization." *INFOVIS '04: Proceedings of the IEEE Symposium on Information Visualization*. pp 25–32.

Kubat, Rony, Philip DeCamp, Brandon Roy and Deb Roy. (2007) "TotalRecall: Visualization and Semi-Automatic Annotation of a Very Large Audio-Visual Corpora." *Ninth International Conference on Multimodal Interfaces* (ICMI 2007).

Kubat, Rony. (2012) "Will They Buy?" PhD in Media Arts and Sciences Thesis. MIT.

Levin, Thomas Y. (2006) "Rhetoric of the Temporal Index: Surveillant Narration and the Cinema of 'Real Time.'" In Thomas Levin, Ursula Frohne, and Peter Weibel. (eds.) *CTRL Space: Rhetorics of Surveillance from Bentham to Big Brother*. pp 578–593.

Miller, Matthew. (2011) *Semantic Spaces: Behavior, Language and Word Learning in the Human Speechome Corpus*. M. Sc. in Media Arts and Sciences Thesis. MIT.

Olson, Mike. (2012) "Guns, Drugs and Oil: Attacking Big Problems with Big Data." Strata Conference. Santa Clara, California. February 29, 2012. Keynote Address.

Overbye, Dennis. (2012) "Physicists Find Elusive Particle Seen as Key to Universe." *The New York Times*. [New York] July 4, 2012.

Post, Frits H., Gregory M. Nielson and Georges-Pierre Bonneau. (2002) *Data Visualization: The State of the Art*. Research paper. Delft University of Technology.

Predebon, John. (1992) "The Role of Instructions and Familiar Size in Absolute Judgments of Size and Distance." *Perception & Psychophysics*. 51:4, pp 344–354.

Pylyshyn, Zenon W. (2003) *Seeing and Visualizing*. MIT Press.

Roy, Brandon C. and Deb Roy. (2009) "Fast Transcription of Unstructured Audio Recordings." *Proceedings of Interspeech 2009*. Brighton, England.

Roy, Brandon C., Michael C. Frank and Deb Roy. (2012) "Relating Activity Contexts to Early Word Learning in Dense Longitu-

dinal Data." *Proceedings of the 34th Annual Meeting of the Cognitive Science Society*. Sapporo, Japan.

Rudder, Christian. (2011) "The Best Questions for a First Date." *oktrends*. Web. Retrieved September 25, 2012 <blog.okcupid.com>.

Sawhney, H. S., A Arpa, R. Kumar, S. Samarasekera, M. Aggarwal, S. Hsu, D. Nister and K. Hanna. (2002) "Video Flashlights – Real Time Rendering of Multiple Videos for Immersive Model Visualization." *EGRW '02 Proceedings of the 13th Eurographics Workshop on Rendering*. pp 157–168.

Sengupta, Somini. (2012) "Facebook's Prospects May Rest on Trove of Data." *The New York Times*. May 14, 2012.

Shaw, George. (2011) *A Taxonomy of Situated Language in Natural Contexts*. M.Sc. in Media Arts and Sciences Thesis. MIT.

Stephens, Matt. (2010) "Petabyte-Chomping Big Sky Telescope Sucks Down Baby Code." *The Register*. Web. Retrieved September 20, 2012.

Tomasello, Michael and Daniel Stahl. (2004) "Sampling Children's Spontaneous Speech: How Much is Enough?" Journal of Child Language. 31:01, pp 101–121.

Tufte, Edward. (1986) *The Visual Display of Quantitative Information*. Graphics Press LLC. Cheshire, Connecticut.

Tufte, Edward. (2006) *Beautiful Evidence*. Graphics Press LLC. Cheshire, Connecticut.

Vosoughi, Soroush. (2010) *Interactions of Caregiver Speech and Early Word Learning in the Speechome Corpus: Computational Explorations*. M.Sc. in Media Arts and Sciences Thesis. MIT.

"Youtube: Press Statistics." (2012) Web. Retrieved September 25, 2012 <www.youtube.com/press_statistics>.

Yuditskaya, Sophia. (2010) *Automatic Vocal Recognition of a Child's Perceived Emotional State within the Speechome Corpus*. M.Sc. in Media Arts and Sciences Thesis. MIT.

Zhe, Jiang, Fei Xiong, Dongzhen Piao, Yun Liu and Ying Zhang. (2011) "Statistically Modeling the Effectiveness of Disaster Information in Social Media." *Proceedings of IEEE Global Humanitarian Technology Conference* (GHTC). Seattle, Washington.