Data Portraits: Aesthetics and Algorithms

by

Alexandru C. Dragulescu

BS, Cinema and Photography, Ithaca College (2000)
MFA, Visual Arts, University of California at San Diego (2005)

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning
in partial fulfillment of the requirements for the degree of

Master of Science in Media Arts and Sciences

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2009

Author . . . .

Alexandru C. Dragulescu
Program in Media Arts and Sciences,
School of Architecture and Planning
September, 2009

Certified by . . . .

Judith Donath
Associate Professor of Media Arts and Sciences
Thesis Supervisor

Accepted by . . . .

Deb Roy
Chair, Department Committee on Graduate Studies

Data Portraits: Aesthetics and Algorithms

by

Alexandru C. Dragulescu

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning
on September, 2009, in partial fulfillment of the
requirements for the degree of
Master of Science in Media Arts and Sciences

Abstract

While interacting online, one generates a multitude of personal data trails, both textual and behavioral. The *data portrait* is a way to collect, condense and represent these information trails, which are often time consuming and tedious to find and grasp when read linearly across web pages or domains, into an easy, legible, and compelling visualization.

This thesis presents various data portraiture techniques that generate both individual and collective portraits of users participating in online social media. The data used in generating the portraits are unstructured text and publishing timestamps of Twitter micro-blog posts, as well as aggregate RSS feeds from FriendFeed. The strategies for depicting people's online personas explored in this thesis focus on the compression, mapping and visual representation components of the visualization pipeline. The resulting portraits attempt to maintain a tight connection with the data, and be legible to viewers, but at the same time, venture to explore more expressive visual forms, and engage with the evolving aesthetics of cinematography, typography and animation.

Thesis Supervisor: Judith Donath
Title: Associate Professor of Media Arts and Sciences

Data Portraits: Aesthetics and Algorithms

by

Alexandru C. Dragulescu

The following served as a reader for this thesis:

Thesis Reader . . . .

Caroline Jones
Director, History, Theory and Criticism Section
Department of Architecture
MIT

Data Portraits: Aesthetics and Algorithms

by

Alexandru C. Dragulescu

The following served as a reader for this thesis:

Thesis Reader . . . .

Martin Wattenberg
Director, Visual Communication Lab
IBM

Acknowledgments

I am very grateful to my advisor, Judith Donath, for her inspiring classes, guidance, invaluable insight and critique.

Many thanks to my readers, Caroline Jones and Martin Wattenberg, for their thoughtful commentary and critique.

Thanks to Yannick Assogba, Luis Blackaller, Drew Harry, Kate Hollenbach, Ana Luisa Santos and Aaron Zinman, my colleagues and best friends at the lab.

Thanks to Aaron Blankstein, Don Goldin, Joe Tennent, Jenny Chan and Tuan Huynh for their dedication and contributions in the implementation stage.

Thanks to Linda Peterson for her guidance and patience. Thanks to all the students, faculty and staff for making the Media Lab such an unique research environment.

This thesis is dedicated to my soul mate, Laura, my parents, Elena and Emilian, and my sister, Liliana. Thank you all, for your love.

Contents

Chapter 1

Introduction

Prior to coming to the Media Lab, my projects have revolved around creating artworks with data and algorithms (Fig 1-1). In *Blogbot* [14], text extracted from blog posts related to the Iraq war and appropriated video game assets were rendered automatically into experimental graphic novels. Together with Tim Jaeger, I created *Respam* [15], a series of audio and visual compositions driven by the rhythms and contents of junk emails. These performances presented a dystopic information society, overloaded by advertising and plagued by consumerism, phishing and fraud. The same unwanted email messages provided the rhythms that generated virtual structures for the *Spam Architecture Series* [16].

My intentions have always been to find and give form to *data*, or to sculpt with *data*, in order to provoke the audience to think about this new material that has become so pervasive in our lives. Daily we consume and generate vast quantities of data. In online communications, data is the material that composes our digital personas, projections of our real selves in a mediated world.

In this thesis, I extend my previous work by exploring ways to represent and render our "data bodies". Our data bodies are formed by the multitude of information traces we leave behind while communicating online. Some of this data is public, while some of it is private. Some of it resides on the client side – on our mobile phones and computers, while some resides replicated on nodes in the computing cloud. We are often not aware, or we do not pay enough attention to whom else this data is visible.

This thesis is organized as follows. Chapter I, *Background and Related Work*, discusses the motivation behind the concept of *data portraits* as a means for evoking our data bodies, and describes my approach into this relatively new research area. I give a brief history of portraiture, highlighting milestone moments and shifting trends. I then discuss the functions of the traditional portrait, as well as the data portrait, and present related work from design, photography, and information visualization projects.

The main body of this work is discussed in Chapters 3 and 4. Chapter 3, *Designing Data Portraits* presents sketches and explorations that use typography and motion to condense an individual's micro-blog into a static or animated portrait. This chapter shows a series of typographic strategies and visualizations metaphors that are used in the final instantiations of the data portraits.

Chapter 4, *Collective Portraits*, presents data portraiture techniques that are re-purposed in the context of one's social network. Audiences can now contrast and compare a multitude of portraits synchronized in an unique timeline, as well as observe semantic interactions in an ecosystem of Twitter users.

Chapter 5, *Critique*, summarizes the results of a critique session with my thesis readers, and presents the outcome of informal interviews with museum visitors who experienced some of the collective data portraits. Finally, Chapter 6 concludes with directions for future work.

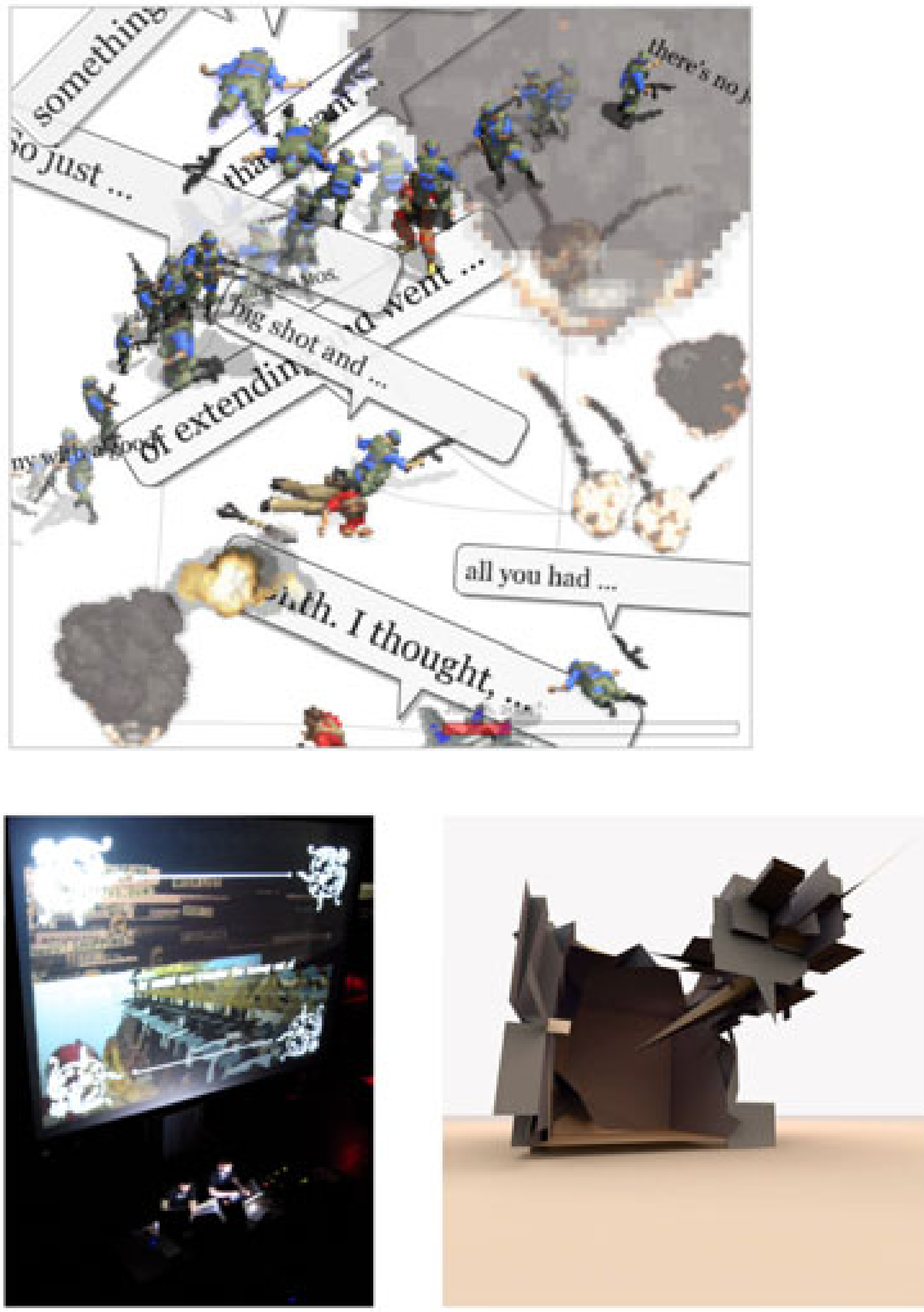

Figure 1-1: Sculpting with data, previous work: *top: Blogbot*, software for generating experimental graphic novels; *bottom left: Spam Architecture, Structure 11*, digital print; *bottom right: Respam*, audio visual performance.

Chapter 2

Background and Related Work

Increasingly we spend our lives in electronic habitats: e-mail, RSS feeds, Facebook, Twitter, online shopping and banking. Our presence in these habitats is a projection of our real selves and is manifested and perceived by the textual or behavioral data we generate while dwelling in these environments. This data might be open for read access to all other users, or it may be protected from other users, but readable and shareable by the commercial entities owning the channels of communication. Most of the time, the data streams we generate are social, they are intertwined and interconnected with other users' information, and most importantly, they reside on commercially-owned servers. Hence, the question of *Who owns the data?* is a difficult one to answer [40]. Because of the fuzzy ownership of data, and the access control concerns, the question of *How do we capture and portray our online personas?* becomes an interesting and beneficial issue to investigate.

The *data portrait* is a way to collect, condense and represent a multitude of substantial personal information trails, that are time consuming and tedious to find and grasp when read linearly across web pages or domains, into an easy, legible and compelling visualization. As such, the data portrait can increase awareness, allowing users to understand how they project themselves in online habitats, or in other words, how others perceive them, and what information is available and stands out about them. In addition, the data portrait could also serve as instrument of reflection on social relationships and interpersonal exchanges, as well as personal narratives and histories.

The design of these data portraits requires an interdisciplinary approach, which finds itself at the intersection of many domains such as information retrieval, natural language processing, design, aesthetics and visualization. The data portraiture techniques developed for this thesis have the goal of maintaining a tight connection with the data, and be legible to viewers, but at the same time, venturing to explore more expressive visual forms, and engage with the evolving aesthetics of cinematography, typography or animation. This thesis concerns itself more with the mechanics of how to represent online personas and does not attempt to exhaustively make use of all possible types of data. Furthermore, it reflects on the data portrait as an artistic genre in the light of traditional portraiture.

2.1 Related Work

The first text and visualization work that mentions and proposes the concept of *data portrait* is *PeopleGarden: creating data portraits for users* [39]. In the context of Usenet groups, message boards and chat rooms, Judith Donath and Rebecca Xiong ask the question *How can we compactly convey information about these participants to each other?* The answer is using data portraits, an *abstract representation of users' interaction history*:

> ”In real life, successful portraits show the physical likeliness of people; they also use objects in people's lives to illustrate their interests and status. This second aspect is an inspiration for our work. The on-line equivalent of one's objects is data about one's past interactions.”

PeopleGarden used a flower metaphor for the portrayal of users, and a garden metaphor for representing the whole group. Various communication patterns are highlighted through the use of colors, the number of petals or the length of the stem. PeopleGarden's goal was to facilitate forming a mental image of an individual's participation in an online interaction environment and help answer questions such as:

- Do participants here really get involved? (Post frequently or only once in a while?)
- How much interaction is there? (Do people respond to each other?)
- Do participants here welcome newcomers? (Do newcomers get many replies?)

- Who are the experts? (Who has been here for some time or posted many messages?)

Themail [33] is an email archive visualization that portrays relationships between individuals. As a data portrait, the focus is more on the relationship than the individual, however, the presence of topical patterns characterize and create impressions of people through their information history. Themail is an interactive tool meant for reflection on past conversations and events, and allows the user to gain insight on the evolution of relationships. It uses typographical devices for representation, and text analysis to filter large quantities of data, similar to the techniques used in the data portraits proposed by this thesis.

What the World Eats [26], a photo essay on the *Time* magazine's website shows the items on the dinner table of fifteen families around the globe. The photographs are extracted from Peter Menzel's book *Hungry Planet* [27]. These images are a fantastic example of data portraits. Besides the photographic likeness of the individuals in a family, there is a multitude of additional data: the quantity, types and diversity of food, the decoration in their home, the clothes they are wearing, the size of the family. There is also information about the weekly expenditure for food, as well as their favorite food or recipe. The audience is left to infer from this data impressions and characteristics of the portrayed individuals. Looking across these portraits, the audience can reflect on globalization, or health concerns. For example, there is a strong contrast between the packaged food of families from the industrialized world and the bulk foods of families from rural areas or developing nations.

A day in brands [5] is a visualization of daily services and products usage by Ana Bierzanska. This is one of the many instances of an exercise first created by an Account Executive at an advertising agency who blogged anonymously at the now defunct *http://dearjanesample.wordpress.com/*. The logos are arranged in groups and ordered by time. There is a large variety of products from toothpaste, toothbrush, clothing, drinks, software applications, phone and computer hardware, to media outlets. Actions are decoded via the type of product or service and the time of day of use. Again, data is used as a intermediary for creating an impression, a mental image about what the person behind the portrait might be like. For example, from the type of software she is using, we can infer that she is a designer or digital media maker. This self-portrait is also a vehicle for fashion statements, from clothing brands to the use of open-source applications like Firefox or

VLC.

Wish Lists [31] is a series of forty inkjet prints on paper created by Kelly Sherman. The content of the prints consists of wish lists found on the Internet. The entries are printed orderly, left-aligned and minimalistically displayed as a vertical list. At the top of the list there is the name of the author or a generic name such as *Family #13*. The items descriptions vary, from really specific products, such as *Little Tikes Cookin' Fun Interactive Kitchen Set* to very generic ones: *baby doll - any*. Through these pieces of information, the audience tries to deduct age and gender, imagine character traits or narratives, eventually reconstructing the persona of the author.

The Feltron Annual Reports [19] are a series of yearly compilations of personal statistics by designer Nicholas Felton. These highly designed and typographically elegant brochures are a form of data self-portrait. Their form is transplanted from the contemporary corporate culture, where companies give shareholders a report on financial performance and future direction. Felton keeps track of many types of data: the number and brands of beers, the number of vacation days, the average daily subway traveled miles, the weight gain, the number of burglars confronted, the countries traveled to, the ten most listened to music artists, the best wedding attended. Information is displayed using infographics devices such as line charts, pie charts and maps.

2.2 Portraiture and the Mechanical Artist

Similar to traditional portraits, the role of the data portrait is to create an evocative representation of an individual. It is therefore important to discuss the similarities and differences between the two genres and understand how the latter registers and distinguishes itself in the broader context of portraiture.

Joanna Woodall, the editor of *Portraiture, Facing the Subject* [38], a collection of essays on portraiture, delineates various paradigms in the history of portraiture in western art. Central to western art, is the idea of *likeness* and naturalistic portraiture, where physiognomic similitude represents identity. Across time and geography, this idea had various degrees of adoption. In the medieval period, identity was not represented through likeness, but through symbols of status and position, such as a blazon, or coat of arms. The realist portrait emerged at the beginning of the fifteenth century, introduced by the early Flemish painters, and perfected by Jan Van Eyck. It is characterized by ”the adoption of intensely illusionistic, closely observed facial likeness, including idiosyncrasies and imperfections”. In the sixteenth century, artists in Italy ”reconciled attention to the physiognomic peculiarities of the subject with more generalizing visual devices, such as the profile view (especially for women) or the analysis of face and body in smooth, consistently lit geometrical shapes. Such techniques were traditionally understood to attribute universal and ideal qualities to figures.”

The need for scientific likeness was materialized with the advent of photography. Furthermore, as Freud’s writings on psychoanalysis became widespread, artistic exploration shifts from depicting the external appearance, towards investigating and portraying the subconscious and the interior life of the subject. In the early twentieth century, naturalistic portraiture is further subverted by the experimentation with non-figurative painting and sculptural techniques. The dynamic between artist and sitter has also changed dramatically during this period. Whereas in the Renaissance era the power dynamic was clearly dictated by the patron who commissioned the portrait, the beginning of the twentieth century marks the act of commissioning being seen as an act of submissiveness, detrimental to artistic creativity. The portrait refers now more to the identity and aura of the artist, while the sitter becomes secondary [38].

Data portraiture follows in the tradition of modern and contemporary art, where the relationship between artist and subject becomes much more diluted. The act of commissioning is more mechanical and symbolical: in the case of Lexigraphs (see Chapter 3), Twitter users will enter their login name into a form. Knowing each individual and portraying it separately is impossible: the vast and daily expanding data needs to be synthesized by algorithm. The actual data portrait is rendered by the intermediary *mechanical artist*, a program or a collection of programs that materialize the will of the artist/programmer. The *mechanical artist* blurs the lines between art material, art instrument, art experience and art object. It also subverts the power relationship between sitter and artist: the representation of the subject’s identity is controlled now by the program. The software is one part material: the electronic canvas is the support on which the portrait manifests itself, and the interface is an integral, ”tangible” component of the data portrait that allows for exploration; and one part instrument: it contains the code that collects and generates the final art object/experience.

Unlike paint or marble, data is fluid, processable and malleable. Through code, data can be infinitely sculpted and layered into various forms and can be interactively experienced by the sitter or audience. Precisely because of these qualities of data, the artistic process of algorithmic creation in general, and data portraiture specifically, is very different from its traditional counterpart. In creating portraits, a traditional artist would use additive or subtractive processes. In data portraiture, the process of evoking ones persona is now iterative. The artist starts by looking across the data (see Section 3.1.1) and experiments with algorithms of extraction and representation of facets of one’s data body (see Chapter 3). The artist has less control on the output and result of each portrait. The subjects are evoked through broad, general algorithmic gestures as opposed to sitter-specific, carefully placed brush strokes.

The goal for the data portraits is to allow the subject to observe how facets of the online self are portrayed, and compare with the depiction of others through the same lens. As such, customizing and tweaking individual portraits (shape, speed) is not desired, so as to allow an audience to observe the same mapping across portraits. An artist’s point of view, or commentary is automatically embedded through subjective selection of various portrayal strategies, as well as the choice of metaphor, or filtering and condensing algorithms.

The algorithms in data portraiture are in a sense recipes that emulate to a certain degree the capacity for evaluating one’s persona by reading through the flat data. The tensions between portraiture and scientific visualization emerge during this process, as data portraiture is situated at the intersection of these two fields. The key decisions of the data portraiture pipeline allow for trade-offs between subjectivity and expressivity on one side, and legibility and recognizability on the other side:

- Choice of algorithm, and parameters: How do we compress the data? What do we choose to show? How do we filter out noise?

- Data mapping: How are the quantities and attributes of the data expressed? Is this a simple mapping that is immediately legible (e.g., size of type denotes frequency, opacity denotes presence or activity)? Does it require a legend or an explanation of the mapping?

- Metaphor and rendering style: How do we visually represent the data and patterns? Is it abstract (e.g. using lines and color) or do we use a metaphor (e.g. silhouette, clock) to arrange the data?

2.2.1 Metaphor

The metaphor is well known as a literary device employed by poets. However, we are often unaware of the crucial role metaphors play in our perception and day-to-day living, as exposed by George Lakoff and Mark Johnson in their book *Metaphors we live by* [23]:

> ”The concepts that govern our thought are not just matters of the intellect. They also govern our everyday functioning down to the most mundane details. Our concepts structure what we perceive, how we get around in the world, and how we relate to other people. Our conceptual system thus plays a central role in defining our everyday realities. If we are right in suggesting that our conceptual system is largely metaphorical, then the way we think, what we experience and what we do every day is very much a matter of metaphor.”

Structural metaphors allow us to understand an abstract idea through another concept that is more clearly grounded and delineated in our experience. We borrow them from one domain of experience to characterize and understand concepts in other domains. *Orientational metaphors* are based on our physical spatial interaction with our environment and navigating the real world with our bodies. Containers, paths, center-periphery, force, links, balance are central to orientational metaphors. For example, ”Happy is up; sad is down”. Our language reflects what we experience physically through our bodies and how our bodies, in this case, posture, reflect emotional state: ”That boosted my spirits. You’re in high spirits. Thinking about her always gives me a lift. I’m feeling down. I’m depressed. My spirits sank.”

The data portraits at the center of this thesis, are using both structural and orientational metaphors as a mechanism to facilitate perception, but also as a vehicle for artistic statement. Chapters 3 and 4 show explorations of how personal data can be imagined and represented through the use of various metaphors. One's collection of information traces and artifacts is imagined as a human body (Sections 3.3 and 4.1), a clock (Section 3.1.7), a cellular organism (Sections 3.2 and 4.3), or as architecture (Section 4.2).

> ”The reason we have focused so much on metaphor is that it unites reason and imagination. Reason, at the very least, involves categorization, entailment, and inference. Imagination, in one of its many aspects, involves seeing one kind of thing in terms of another kind of thing – what we called metaphorical thought. Metaphor is thus *imaginative rationality*.” [24]

Finding the right metaphor is challenging. Some metaphors will highlight certain aspects of the concept meant to be illustrated or understood, while at the same time hide others. For example, a clock is great for representing rhythmical data, but it might be too mechanical to illustrate an individual. The qualities of the visualized data need to be taken into account when employing metaphors. Can a piece of data be considered solid and perpetual, so that it can be portrayed as the building block for a physical structure or an architectural fragment? Metaphors may have rich and complex associations and interpretations, some of them experiential and some cultural, which might not reflect the traits of the data, or might help infer undesirable qualities. On the other hand, the large spectrum of social data needed to be visualized might not be entirely characterizable by the chosen metaphor. Judith Donath writes on the perils of metaphor scope or coverage, critiquing the PeopleGarden visualization [39] that used a garden of flowers as a metaphor to represent Usenet users [12]:

> ”The flower imagery is highly evocative, which is both its appeal and its limitation. Gardens are organic, we associate them with life, beauty, freshness. This imagery is fine when the visualization is of, say, a welcoming support group. It seems jarringly inappropriate for depicting dry technical announcements or vitriolic flame wars.”

2.2.2 Words

The information sources for the data portraits presented in Chapter 3 are the RSS feeds of Twitter micro-blogs. The two components of interest are the contents of the post: the raw unstructured text, together with the publishing timestamps. The goal is to extract and represent patterns in the text by filtering and compressing the original content. Because of the limited message length of 140 characters, micro-blog users are already doing a first stage of compression by trying to communicate ideas succinctly, using a small number of words.

Twitter users are supposed to answer the question *What are you doing?* but the range of messages show a large variety of uses. Some of the posts describe present or past actions, while other are observations or musings. Some posts are replies in conversation threads with one or multiple users, while others simply link to photos, or other websites.

Text analysis research [28] has shown that ”the words people use in their daily lives can reveal important aspects of their social and psychological worlds” and that ”particles, which serve as the glue that holds nouns and regular verbs together, can serve as markers of emotional state, social identity, and cognitive styles.”

By using word frequency/counts or more sophisticated filtering techniques like *tf-idf* [30], the Lexigraphs series, discussed in Chapter 3 builds a collection of relevant words for a certain user which are then employed to render the portrait. The portrait only filters, condenses and attempts to replay behavioral rhythms, while the task of judging, comparing and extracting gist is left to the audience. In *Designing Visualizations of Social Activity: Six Claims* [17], Thomas Erickson argues that visualization should allow user to interpret information and suggest rather than inform:

> ”Our response is that accurately presenting information is not the point of a social visualization; its primary role is to provide gist for inferences, and, in fact, it is less important that the inferences are correct. Our users have proved very comfortable with making best guesses from incomplete information. Thus, it is OK to distort activity, to magnify small amounts of activity, and to dampen large amounts of activity; for example, it is much more important for users to

> be able to tell whether there are 3 or 7 people present, than whether there are 103 and or 107 present. Ideally, the ambiguity of the visualization should be clear to users.”

In other text visualizations, like the *State of the Union Address in Words* [36], the words are mapped into charts and the actual words appear only in the legend. In contrast, the words are an essential part of the portraits put forth by this thesis. There are both advantages and disadvantages to this technique. When words are used in the portrait, there is the obvious advantage of not needing to go back and forth to consult the legend. However, there are some representation and perceptual challenges. As in the case of tag clouds, differentiating between word sizes could be problematic, given that frequency is mapped to the height of the font, while the total area of one word is influenced by the length of the word in letters. There are also font differences depending on what letters make up the word. Letters with ascenders or descenders look larger than letters without them, and in the case of fonts that are not fixed-width, the letter *w*, for example, will be larger than letter *i*. Despite the theoretical perceptual issues, the popularity and widespread use of tag clouds in web applications and blogs make this technique accessible and familiar [34].

2.2.3 Time

Tone and rhythm of activity or posting in online mediated communications could be considered the equivalent of real life gait. In real life, we tend to make a variety of conscious or subconscious social judgment calls based on the body language, posture, or manner of walking of other people. Artists across media and time have closely observed and rendered gait. As an example, seventeenth century sculptor Domenico Bernini quotes his father [37]:

> ”If a man stands still and immobile, he is never as much like himself as when he moves about. His movement reveals all those personal qualities which are his and his alone”.

As another example, in 2D or 3D animation a character’s walk cycle can be an expressive form to render both physical, emotional or psychological attributes [35].

Unlike regular blog posts that have no length limits, micro-blogging has another important dimension: rhythm. Blogs are much more about exposing ideas at length, paying attention to narrative structure and grammar. Micro-blog posts resemble telegraphic snippets, short bursts of ideas and actions, in a stream of consciousness way. Their temporal nature can reveal characteristics of the author. For example, if someone is posting regularly messages of self-promotion, he or she could be seen as self-centered, in the same manner in which an arrogant walk might trigger similar assumptions.

Studies of collaboration in online communications [21] [4] have shown that temporal rhythms in email archives set expectations about availability and interruptability and can be uncovered and predicted. Furthermore, communication rhythms play an important factor in characterizing the nature of relationships between individuals [29].

As noted by Anthony Aveni in his book *Empires of time: calendars, clocks, cultures* [2], rhythm is embedded in our nature, both on a biological and psychological level. Our anatomies observe and replay a multitude of parallel cycles: the .1-second oscillation of brain waves, the 1 second oscillation of our hearts pumping blood, the 6-second cycle of breathing in and out, the circadian cycle with the interweaving between being awake and being asleep, or the reproductive circa-lunar, and circ-annual cycles. Most experimental biologists believe that the timing systems in mammals are found in the brain, more exactly in a group of nerve cells in the hypothalamus called suprachiasmatic nuclei (SCN) . Destroying these cells in rats, seems to affect their sleep-wake cycle. When people are deprived of the light-dark, cycle and live for weeks in either dark caves, or rooms with constant light, first, their circadian cycle of 8-hour stretches to 16 hours and then they psychologically lose track of time. Hence, each of us has a biological clock that is constantly synchronized to the cycle of day and night. Anthony Aveni acknowledges best the connection between the larger rhythms of nature and our inner clocks[2]:

> ”We’ve got rhythm! There’s no doubt about it. And our desire to memorize, mark, and record it seems to have been bred within us right from the start, before we could even call ourselves human. Like all living organisms, we march in time to the dependable beat impressed upon us by nature’s background music.”

Chapter 3

Designing Data Portraits

This chapter describes the types of data portraits explored in this thesis as well as the iterative process involved in designing them.

3.1 Process: Portraiture with Motion and Typography

Process is an important part of the final work. I believe that the directions experimented with, both the ones that were finally pursued, as well as the ones that proved to be unfruitful, are as important as the result. In addition, showing the branching of ideas, implementations and sources of inspiration will allow a better understanding of the goals of this thesis, and serve as documentation for both successful and failed strategies.

3.1.1 Pre-visualization: Getting to Know the Data

Browsing Data One of the first steps in the visualization pipeline is to understand the range, breadth and type of data that will be processed. At the beginning of the data collection stage, I periodically reviewed the last hundred messages posted by Twitter users. This provided a very quick way of understanding the stream of messages, while keeping an eye on the distributed crawling process. I also issued SQL queries to check various characteristics of this cloud of information: daily posting volume for various users, messages containing certain keywords (a rough search), distribution of geographical location

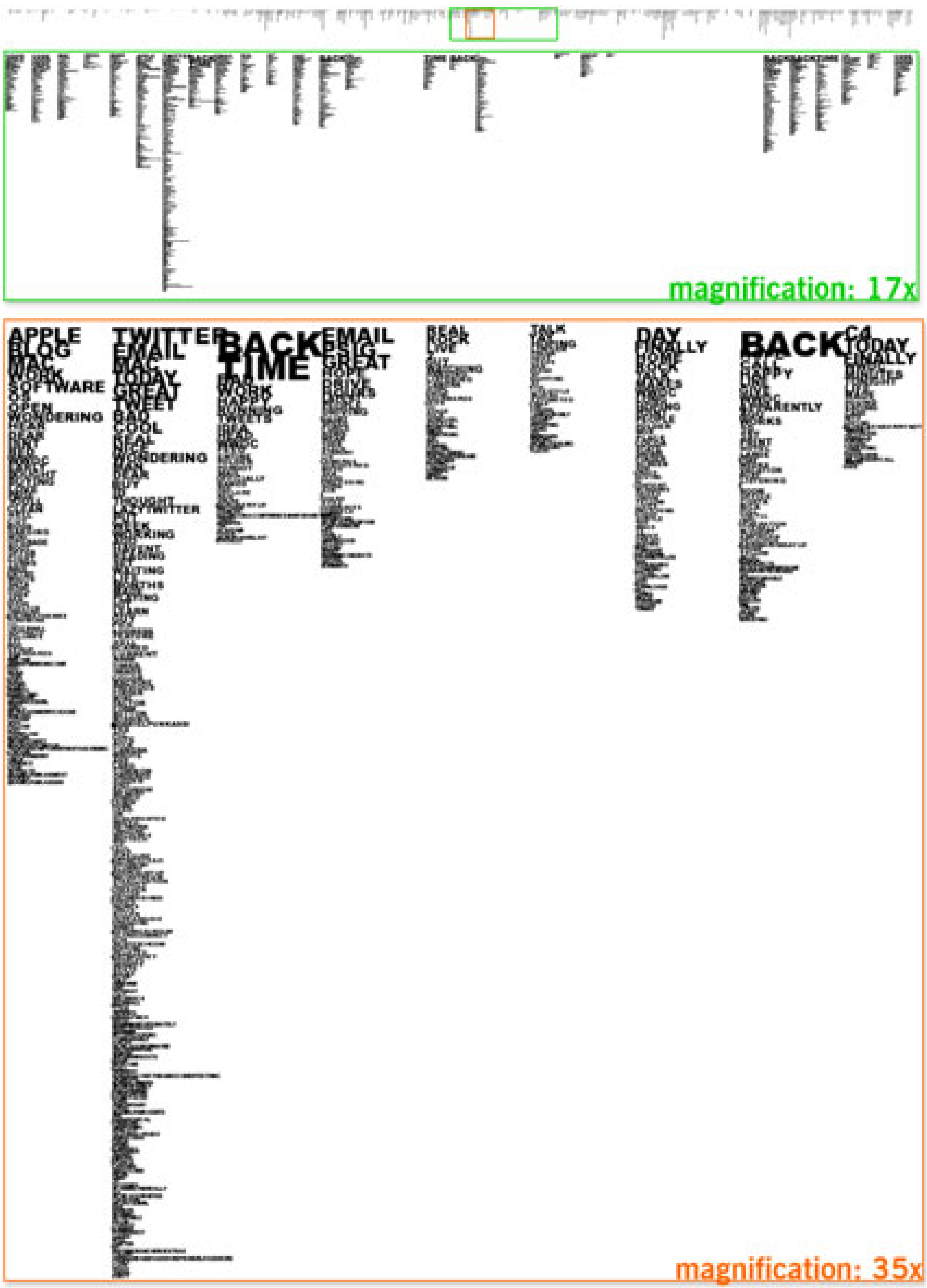

Figure 3-1: Previsualization sketch: the most used words of at Twitter micro-blog in a period of 365 days, ordered on daily columns, most frequent words at the top (Python, Python Image Library - pil).

of users, percentage of protected accounts versus public accounts, usage and distribution of clients, and rate of growth of new accounts.

As the database grew quickly overtime, it became clear that I needed to limit the dataset to messages from English-speaking users only. I have used a Python implementation of TexCat [10] [25], an n-gram based language classifier, to assign a composite score of all the posts for each micro-blog. This approach was mostly accurate, but could not assess very well micro-blogs with bilingual content. Despite a varied corpus used to train the language classifier, this method would also have poor results on posts containing the ever evolving “netspeak”. For example, Twitter users introduced the use of hashtags to tag posts pertaining to a certain theme or topic (e.g., #twitterfail, #sxsw, #earthquake), or the practice of re-tweeting or rebroadcasting somebody elses message (e.g., “RT:”). I have used the language score and the weekly posting volume to focus the data collecting process on mainly English-writing micro-blogs and high-volume users. To minimize page trips, low-volume users would also be crawled and indexed less often than high-volume users.

Filtering and Compression One of the first programmatic sketches I have created was the timeline for a period of one year for one Twitter micro-blog (Fig 3-1). This was done specifically as a means to understand the extent of data. I have filtered out English stop-words using the stop-word dictionary of NLTK [6]. Time is laid out on the horizontal axis, with each column of text representing one day of posts. The remaining words (after removing stop words) are arranged vertically and are sorted by number of occurrences (highest frequency at the top). The size of type also reflects the number of occurrences. The number of posts per day is not visible, as the words that occurred in a single day are merged and sorted in a single list. This previsualization sketch shows that significant spatial compression is needed to display legible text in a horizontal timeline.

3.1.2 Twelve Typographical Devices

The intention of the visualizations presented in this chapter is to use in the representation stage the very words found in the data set. By eliminating the noise in the communication and filtering out the meaningless tokens, I wanted to provide the viewer of the visualization with the essence of the data stream. In other words, I wanted to leverage the viewer's subjectivity, judgment and perception in evaluating the topical and temporal patterns exhibited in the data portrait. This thesis is mainly concerned with mechanisms of representation and methods for mapping and translating information using visual metaphors. Using typographical elements to show the temporal and topical patterns naturally comes to mind. For inspiration, I have looked at poster and book cover designs, a frugal format that along the years has evolved extremely creative methods. In one page, designers must catch the wandering eye of the audience and be able to communicate using type, color, shape and texture.

In the following, I enumerate some of the attributes of type and/or design devices that can be used to replay rhythm or represent numerical data. This is by no means an exhaustive list, but synthesizes representative elements conceptualized and extracted from various images that showed expressive ways of arranging and laying out text[1].

1. Scale/Color/Direction

- All these attributes are easily mapped to numerical values and can be animated.
- Color can be used to separate categories or separate entities.
- Tint or transparency can be used to indicate the intensity of a particular attribute.

Fig 3-2:a De-construct.com (Process, Poster for Blanka, 2006)

2. Repetition

- Used to express quantity.
- Used to create negative or positive space.
- Used to create shapes, figures, or to give texture to an area.

Fig 3-2:b Bronislaw Zelek (Ptaki, Poster for The Birds, by Alfred Hitchcock. 1965)

[1]These images were saved from the stream of image-bookmarking services such as *ffffound* [22] which do not record proper references to the author or date of creation. Whenever possible, I have identified the origin of the images.

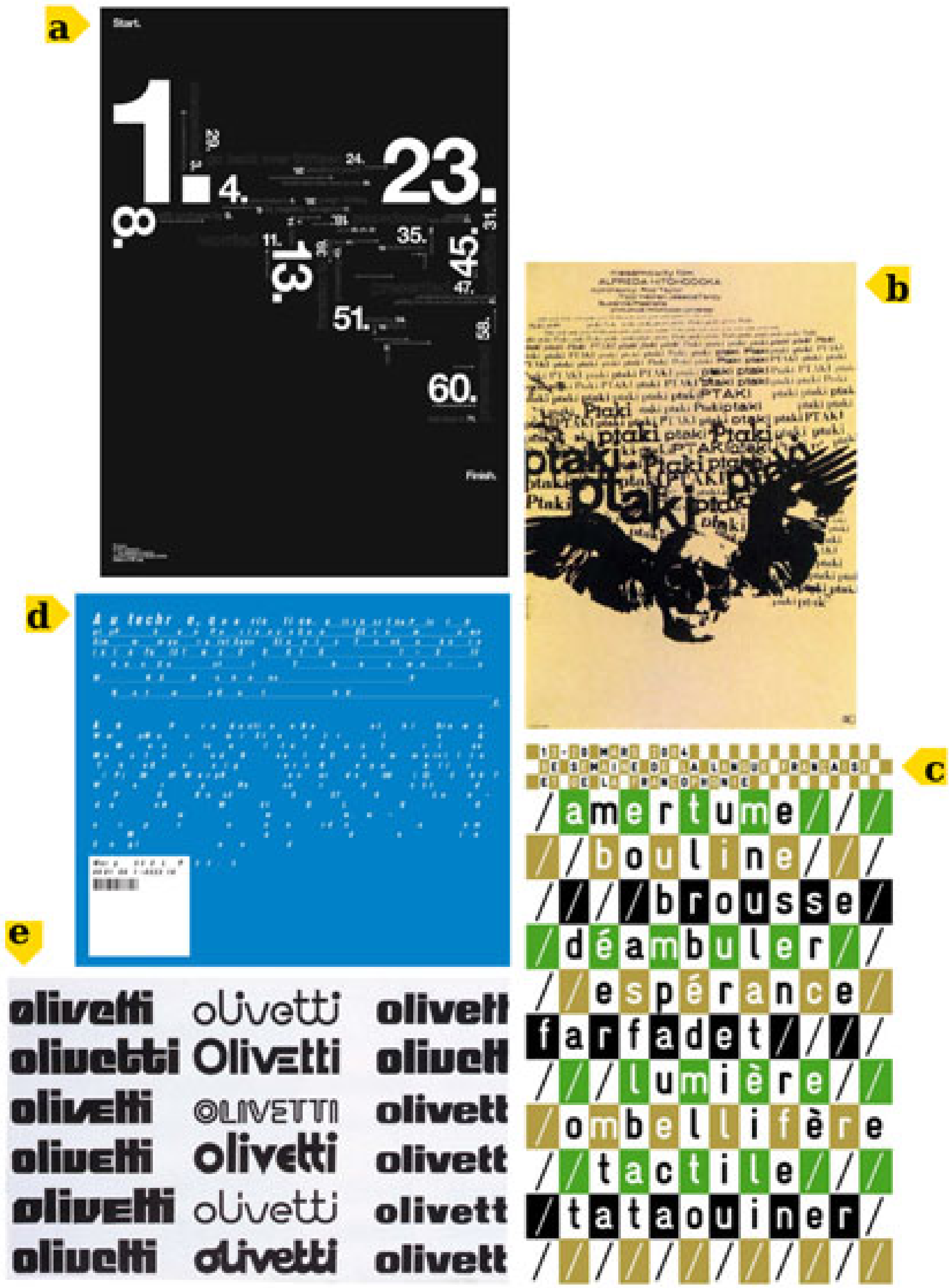

Figure 3-2: Twelve typographical devices that can be used to express data: (a) Scale/Color/Direction, (b) Repetition, (c) Foreground/Background, (d) Leading/Kerning, (e) Font Weight.

3. Foreground/background

- The tension and contrast between foreground and background can be used to create patterns.
- Can be used to focus attention on a certain event (if animated) or area (if still image).

Fig 3-2:c Philippe Apeloig (Poster for La Semaine de la Langue Franaise et de la Francophonie, 2004)

4. Leading and Kerning

- The distance between letters and words and/or rows can be used to create rhythm and patterns.

Fig 3-2:d The Designers Republic (Album Cover for Autechre's Quaristice, 2008)

5. Font weight

- Alternating light and heavy can create rhythm and patterns.
- Used to highlight events when animated.

Fig 3-2:e Walter Ballmer (Olivetti logo treatment, Milan, Italy 1960)

6. Filling space with text

- The label becomes the object. Figures or shapes made of text pack dual meaning/intent.

Fig 3-3:f John Yunker (Country Codes of the World, 2007)

7. Accents, annotations

- Annotations can be used to make additional statements or add context.
- Swirls and swooshes found in graphic novels or music notation can be used to express motion and intensity.

Fig 3-3:g John Stump (Sheet for Faerie's Aire and Death Waltz)

Figure 3-3: Twelve typographical devices that can be used to express data: (f) Filling space with text, (g) Accents, annotations, (h) Lines, timelines, grids, (i,j) Chart embedded, or around type.

8. Lines, timelines, grids

- Lines, timelines and grids can be used to anchor text, mark time, keep tempo.

Fig 3-3:h Unknown designer (Poster for the Zurich Design Museum)

9. Chart embedded or around type

- Could a graph charting the usage for a term, be embedded or wrapped around the word itself?

Fig 3-3:i Reilly Stroope (Sorry, Illustration, 2006)

Fig 3-3:j Pedro Ramirez Vazquez, Eduardo Terrazas, Lance Wyman (Poster, Olympic Games Mexico 1968)

10. Capitalization

- Letters could be capitalized and used as markers for tempo or certain occurrences.

Fig 3-4:k Tauba Auerbach (Alpha Bible, 2006)

11. Breaking the type

- Obscure certain parts of the bodies or letters to replay a pattern or rhythm. Hard to achieve without loosing legibility.

Fig 3-4:l Unknown Designer (Book Cover for The Hidden Curriculum by Benson Snyder, 1970)

12. Use the third dimension

- Layering of multiple lines of text.
- Through perspective and depth highlight or hide layers.
- Add a third dimension to the actual body of the letters. This third axis can be used to express quantity.

Figure 3-4: Twelve typographical devices that can be used to express data: (k) Capitalization, (l) Breaking the type, (m,n,o) Use the third dimension

Fig 3-4:m Brian Dettmer (Sculpture Series, Book Autopsies)

Fig 3-4:n Martina Walthert, Naima Schalcher, Kerstin Landis (Ausstellungsplakat Hans Finsler HGKZ, 2006)

Fig 3-4:o Unknown Designer (Out of Print, Zurich Museum of Design)

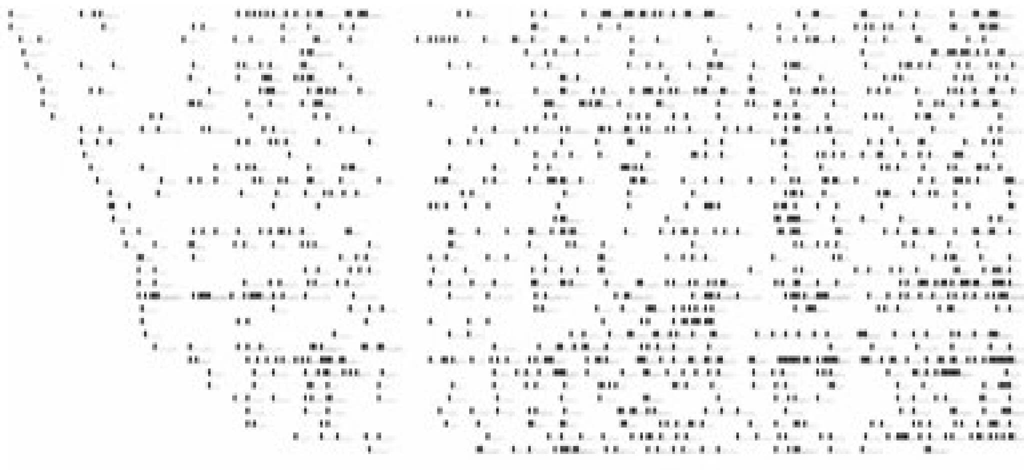

Figure 3-5: Keyframes sketch: Word occurrences are marked by a 2-pixel rectangle (Python, ActionScript)

3.1.3 Sketch: Keyframes

The previsualization sketch (Fig.3-1) showed that because of the varying length of words, a lot of screen space is wasted by trying to use the actual words to mark occurrences. Hence, compression is crucial. In an effort to further compress horizontally the timeline and fit a year of posts into one screen, I have scaled down each day to 4 pixels (Fig 3-5). In this design, the occurrences of words are marked by 2-pixel wide rectangles. The words are not sorted in a single list per day, instead each word occurs on its own distinct timeline. The most *n* frequent words are extracted, after removing stop-words. Frequency is calculated simply by counting the words. In this sketch, *n* is 35.

The vertical arrangement of the parallel word axes is dictated by the order of their appearance in the micro-blog posts. For example, ”twitter”, ”hard”, ”back” appear in the beginning of the micro-blog while ”guess” and ”iphone” appear later on. Each rectangular keyframe marker is annotated with a copy of the word in a lighter gray. When looking at the image as a whole, the repetition accentuates denser areas and allows isolated word keyframes to be identified. While this arrangement shows the exact rhythm of one word for a long period, it makes reading across a day very difficult as two words can occur in one day, but be really far apart because of the fixed placement of the word axes.

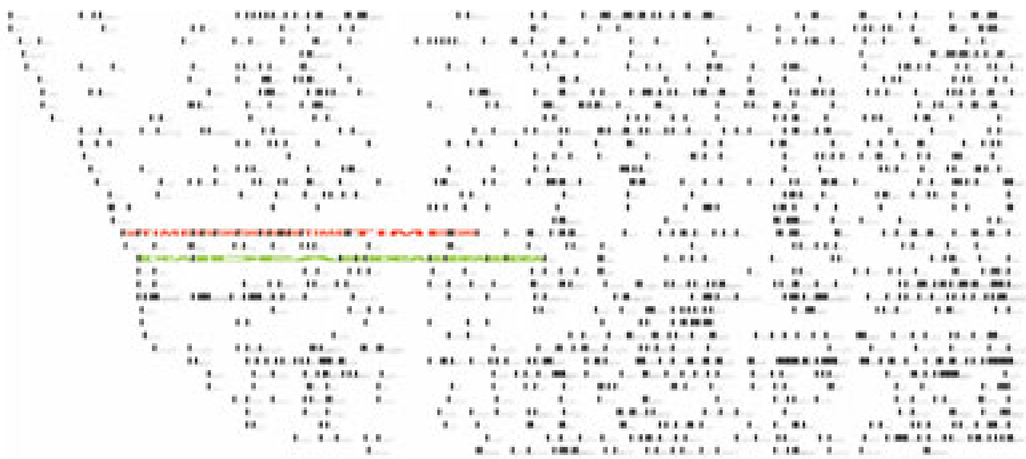

Figure 3-6: Keyframes sketch: Fitting text in between the markers (Illustrator)

Figure 3-7: Keyframes sketch: Fitting text in between the markers. Each word axis has randomly assigned color (Python, ActionScript)

A subsequent iteration on this design shown in Fig 3-6 attempted to make the individual word axes more legible. In a third attempt show in Fig Fig3-7 the same token was stretched or contracted to fit in between the keyframe markers on each word axes. In addition to the fact that type was unaesthetically stretched, the visualization was misleading, as the larger the gap between keyframes, the larger the text. Even if the transparency of type would be proportional to the size, and thus the word would appear less visible, dense areas would still be hard to read as the text is compressed.

3.1.4 Sketch: Waveforms

The design described in the previous section highlighted the fact that the intervals between occurrences are as important as the event itself, as rhythm in music, for example, is a product of both sound and its absence. Hence a natural question arises: how can one depict graphically both presence and absence? The inspiration for the waveform sketches came from sound editing and non-linear video editing application such as Audacity [1] (Fig 3-8), or Avid Media Composer [3]. Waveforms are graphic representations of audio pressure over time. Waveforms from loud sounds fluctuate further up and down from the average middle line that represents the resting atmospheric pressure.

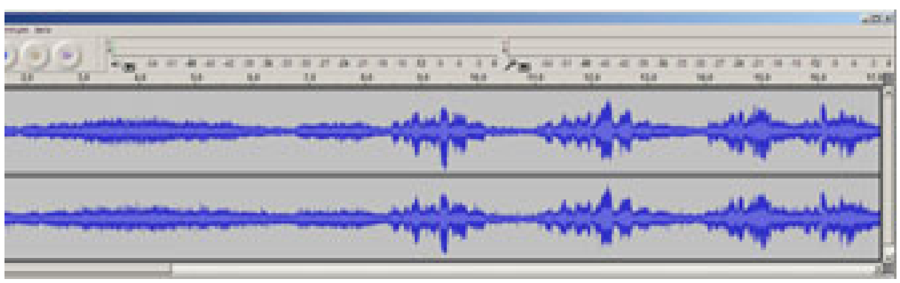

Figure 3-8: Waveform visualization in Audacity, a sound editing application

The sketch in Fig 3-9 shows a possible mapping: word occurrences are the peaks in the waveform while the silence becomes a valley. The height of the waveform also shows if a word occurs multiple times in a day (Fig 3-12). This phenomenon was not represented in the keyframe marker sketch (Fig 3-5).

To further evaluate the waveform as a representation metaphor, I have manually created a sketch where the word axis from previous sketches becomes a ribbon of repeated tokens. The ribbon is stretched vertically when that particular word occurs and compressed vertically when the word has not been used (Fig 3-10). This study includes several solutions for separating the axes through background and foreground colors. The repetition of words leads to a perception problem. Because the day length is fixed and the word

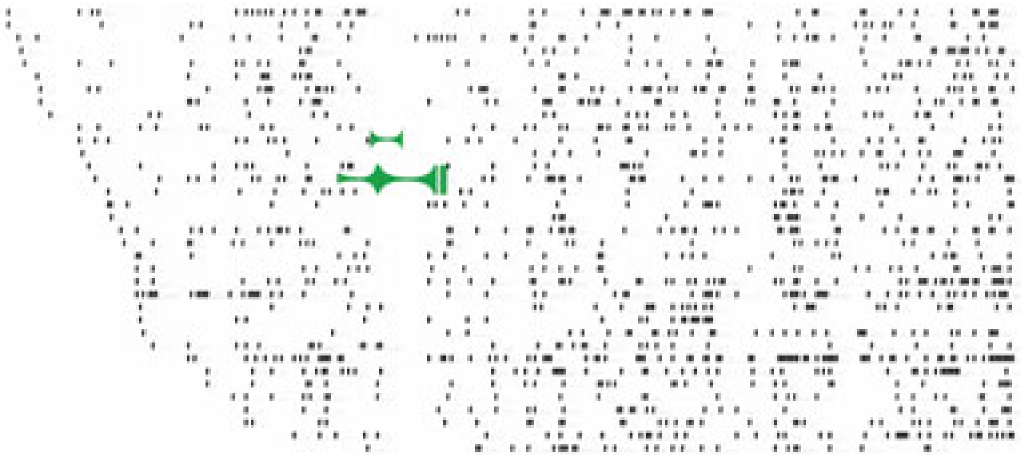

Figure 3-9: Waveform sketch: Waveforms as metaphor for visualizing the absence and presence of a word (Illustrator)

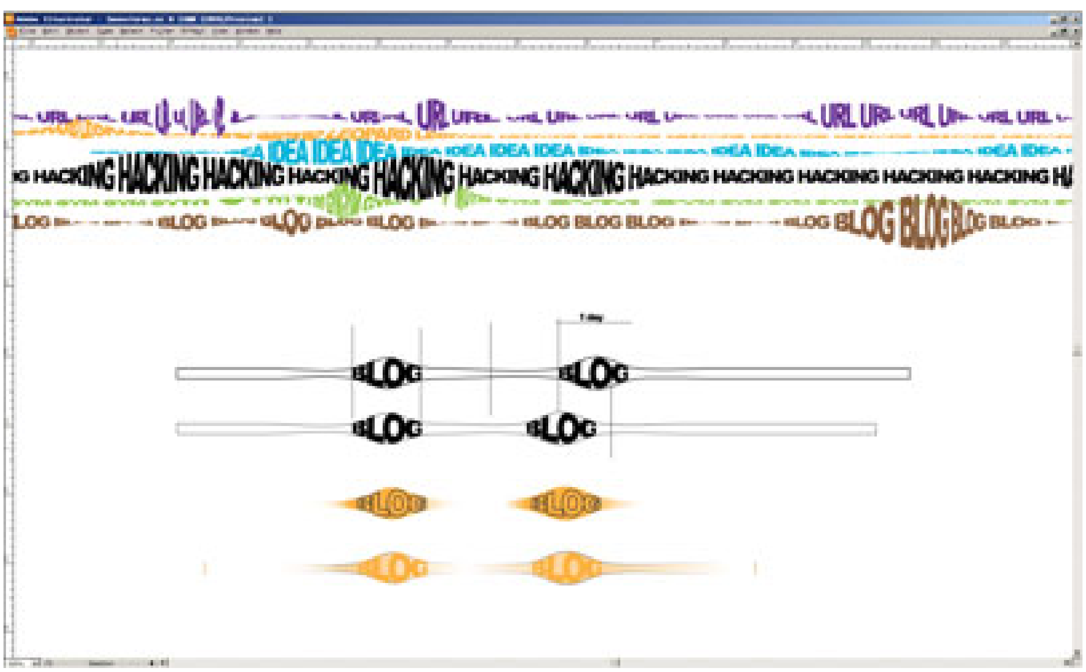

Figure 3-10: Waveforms hand-drawn study: text is warped to fit a waveform pattern (Illustrator)

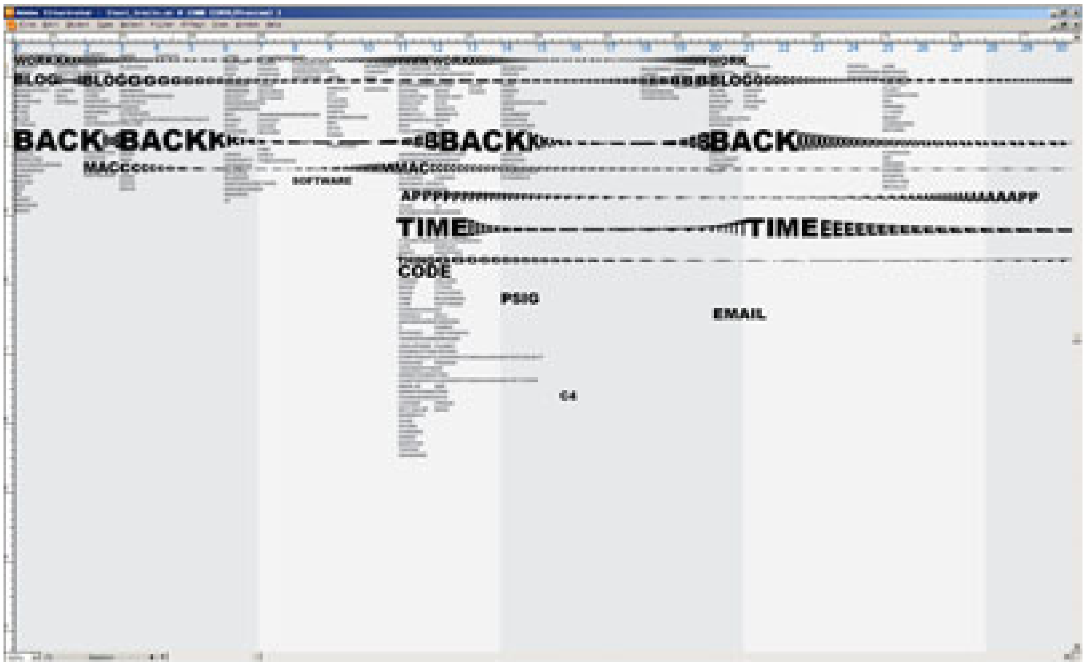

Figure 3-11: Waveforms hand-drawn study: the first and last letter of the word are used to create the ”valley” of the waveform allowing for better legibility (Illustrator)

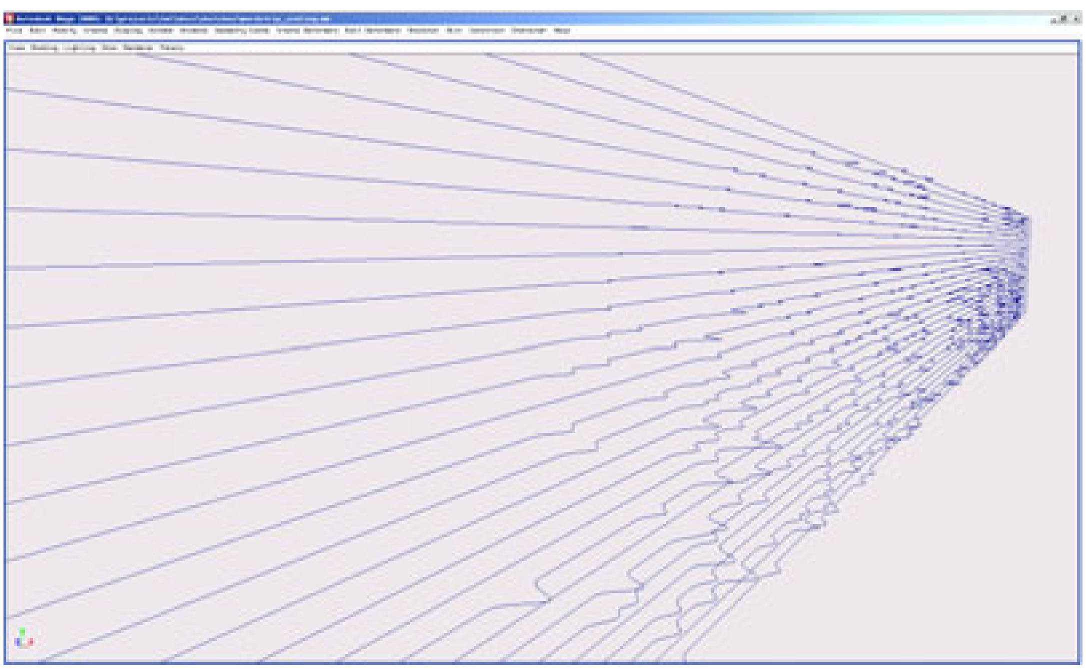

Figure 3-12: Waveforms sketch: Visualization of word occurrences on a 3D cubic spline. The intensity of peaks on the curve is directly proportional to the number of times a word occurs in a day. (Python, OpenMaya, Maya screenshot)

length varies from axis to axis, the interval of warped text does not always coincide with the full word. As an example, on the “blog” axis the ideal warp should happen on the full word: BLOG**BLOG**BLOG. This is not usually the case, and the result could be such as: BL**OGBLOGB**LOG or BL**OGBL**OG. Fig 3-11 shows a study of several solutions to address this issue. The word can be resized to the width of one day and always be stretched correctly, if the first and last letters are used to create the valley that denotes its absence.

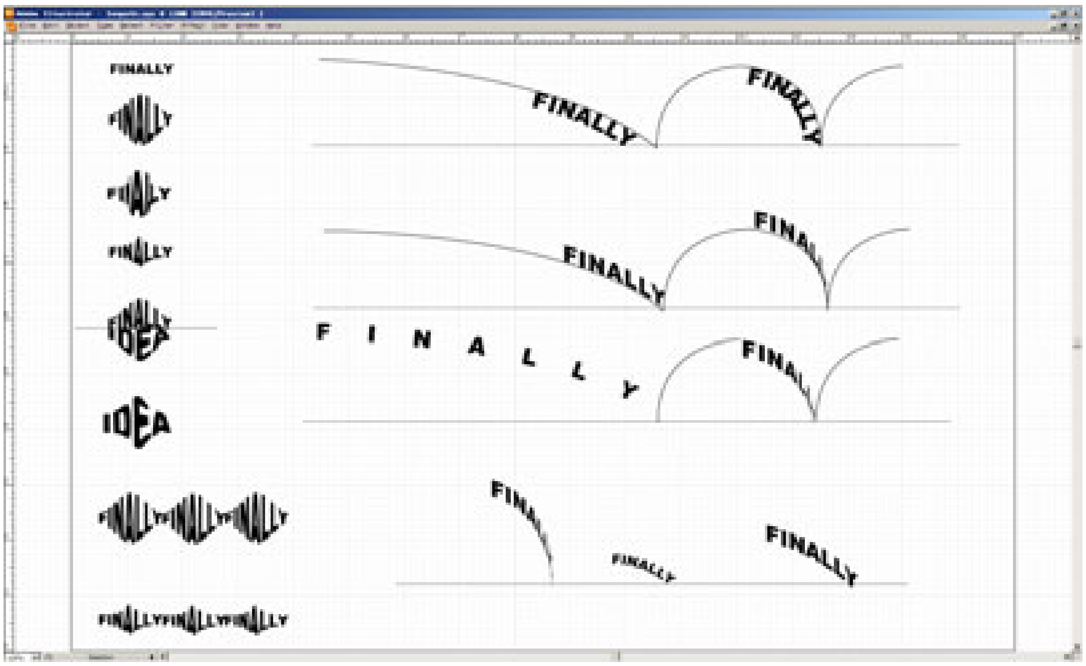

Figure 3-13: Type-on path method hand-drawn study (Illustrator)

The algorithmic sketch in Fig 3-14, shows another problem with this design. If the axes are too close to each other, keyframes could overlap and the top ribbon would be occluded (the z-order [2] increases as the word axes are added). If the axes are too far apart, a lot of wasteful negative space is introduced (Fig 3-15).

Fig3-13 shows other explorations of typographical strategies to represent flow and rhythm. Previous sketches examined what can be done by challenging the inherent rectangularity of type by warping the letters. This study employs the use of type-on- path method where letters are aligned to a cubic curve that describes the occurrence of a particular word. The length of interval of ”silence” for that particular word would be encoded in the angle

[2]Z-order refers to the order of drawing objects on a two-dimensional canvas. Objects with lower z-order are behind objects with higher z-order.

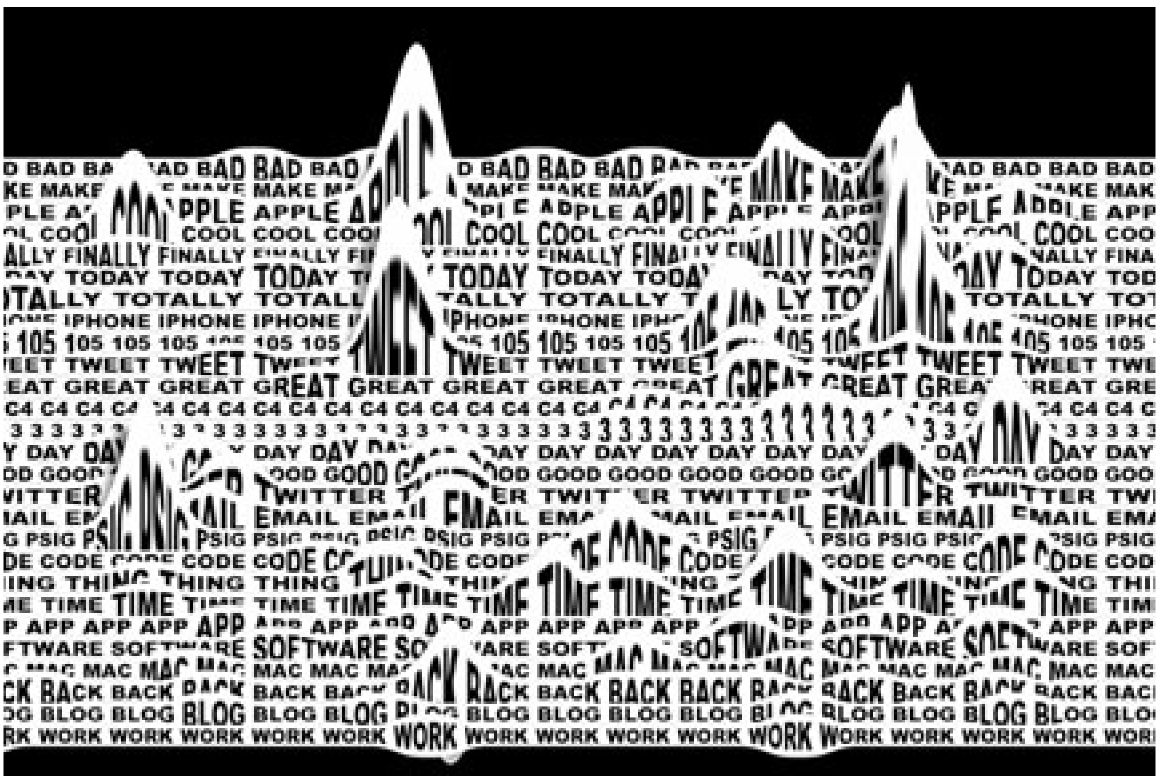

Figure 3-14: Waveforms sketch: text ribbons occlude each other when layered in 2D (Python, OpenMaya, Maya render)

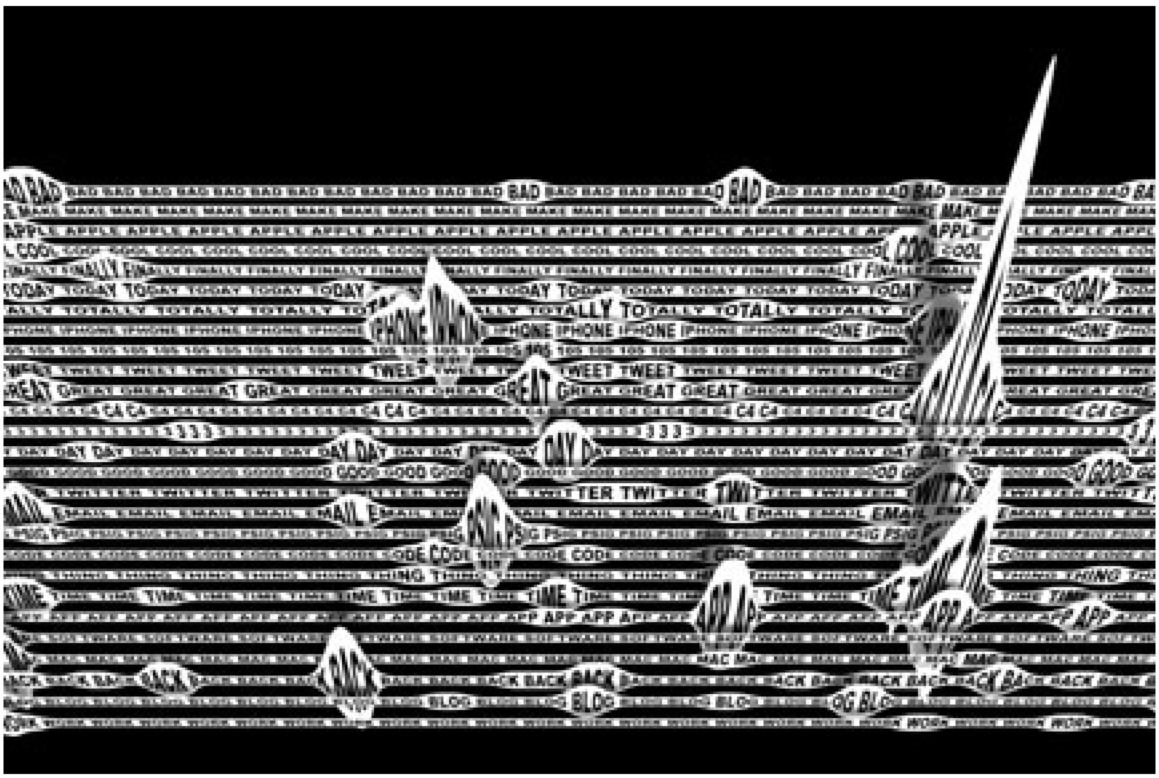

Figure 3-15: Waveform sketch: in a 3D perspective the text ribbons are more legible, yet a lot of space becomes unused (Python, OpenMaya, Maya render)

of the curve tangent at the point of occurrence. Under this schema, at acute angles that can occur when the word appears in consecutive days, the type either breaks down or it is too vertical, becoming less or completely illegible. Hence, the word would be more legible after a larger silence and less legible if reoccurring at short intervals. With audio at least, one might argue that this could be a correct perceptual representation, as a loud drum sound after a period of silence leaves a larger impression than a series of softer drum sounds.

3.1.5 Sketch: Depth of Field

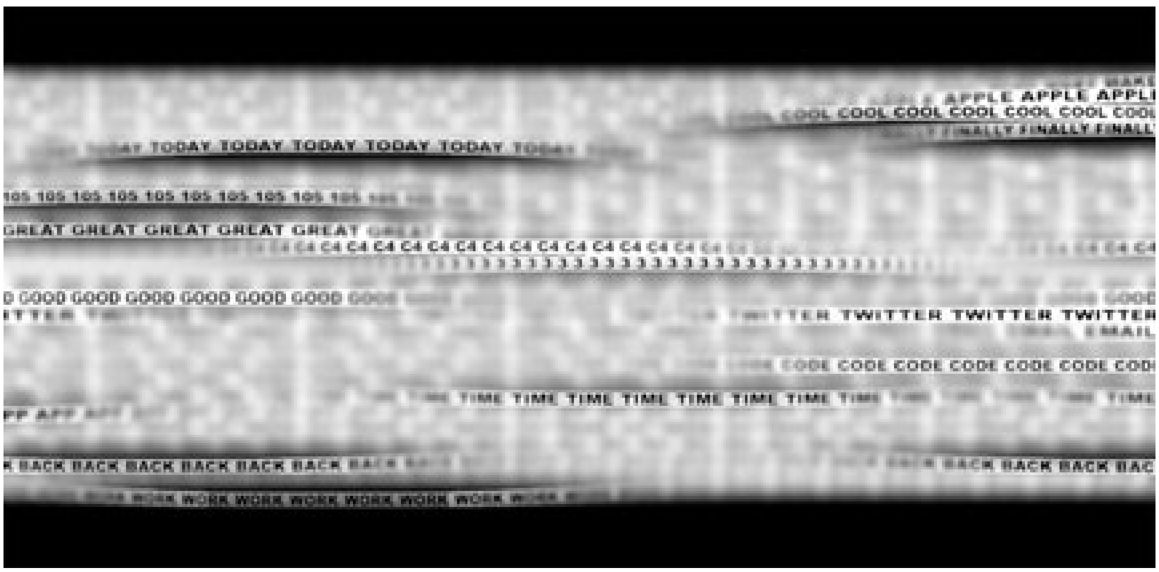

Figure 3-16: Depth of field study: words come in and out of focus as they appear (Python, OpenMaya, Maya render)

Inspired from a cinematographic technique that allows to separate and highlight certain elements in the shot, the depth of field sketch (Fig 3-16) is a variation of the waveform sketch. Instead of warping the word on the vertical Y axis, the word is pushed back on the Z axis, away from the camera's region of focus and thus blurred.

Overall the sketches presented so far led me to the realization that the nature of data does not fit the form and metaphor of representation. Hence, I have abandoned these directions of research. The words do not appear in a continuous stream, and they resemble more of a collection of *on* and *off* states. To use again an analogy from music, the data is more like a drum beat, and less like a bagpipe sound that is always present and sometimes has variations in frequency. It is not a continuous stream whose intensity varies, but separate events.

Figure 3-17: Wordtraces screenshot: daily columns of words, move from right to left (Python, Actionscript, Flash Player)

3.1.6 Wordtraces: Transition to Animation

One of the problematic issues with the previous sketches is that one tends to read the visualization horizontally, following each word axis separately, trying to perceive patterns. However, it is hard to correlate words horizontally and establish context because the days are tightly compressed. It is possible to allow this type of vertical reading either by interactive browsing (i.e., selecting one day and highlighting all the words that occurred that day) or by animating the words as they happen and playing back the entire timeline. In representing rhythms, the interactive browsing option would still have a more difficult time, as highlighting the words in one day requires the altering in the rendering of the highlighted words (either through size or color). This change would perturb the perception of the rhythm captured in the static visualization.

Figure 3-18: Wordtraces animation final time-lapse frame for user mmc (Python, Actionscript, Flash Player)

Figure 3-19: Wordtraces animation final time-lapse frame for user danielpunkass (Python, Actionscript, Flash Player)

The screenshot in Fig 3-17 shows a frame of an alternative visualization that captures both rhythm and context through animation. The layout is similar to the *Keyframes* and *Waveforms* sketches shown previously in Fig 3-5 and respectively Fig 3-11. The tokens are aligned vertically to their corresponding daily columns and they each maintain the same screen height as they appear on their predefined horizontal axis. Time flows from right to left: the current day appears at the right of the screen, while past days move off screen to the left. At the bottom of the screen, a bar chart moves, synchronized with the words and shows for each day, the number of posts.

Because there can be multiple posts in a day, tokens appear together on the screen grouped by post. Token groups from different posts are separated by a slight delay. Thus, context and meaning can be established even if sometimes words in the same post appear at different screen heights (as the tokens are locked to a fixed horizontal axis).

Using animation to fade in and move the words according to their daily occurrences, allows the viewer to understand the rhythms of posting and the gist of a user's data stream. One shortcoming of this approach is the fact that the animated playback only leaves a temporary latent image in the viewer's mind. This image is sufficient for perceiving rhythm and understanding context, but reassessing global patterns becomes time consuming, as the data stream needs to be replayed. A static image would allow the viewer to revisit and scrutinize the aggregate view of the information stream. To address this issue, the next version I have developed (Fig 3-18 and Fig 3-19) plays the data stream, and at the end of the animation it displays an averaged image of all frames, similar to the photographic multi-exposure process. At the right of the screen, these visualizations show the list of most frequently used words ordered by appearing order and scaled proportional to their number of occurrences.

3.1.7 Sketch: Clocks

For the Wordtraces sketches, the mapping for the animation speed was set to three seconds for one day of real time. Interactive controls to allow a viewer to change the speed of playback could have been added. However I was reluctant to do so, as our perception of time is already very fuzzy. Although the words flow at a constant rate on the screen, time seems to expand and contract when looking at the visualization and this is caused by different factors: volume of words per day, ratio of active vs. inactive days, the measure of interest in the data, and the frequent repetition of certain words. Allowing for interactively changing the playback speed would interfere even more with the process of reading and understanding the unit of time.

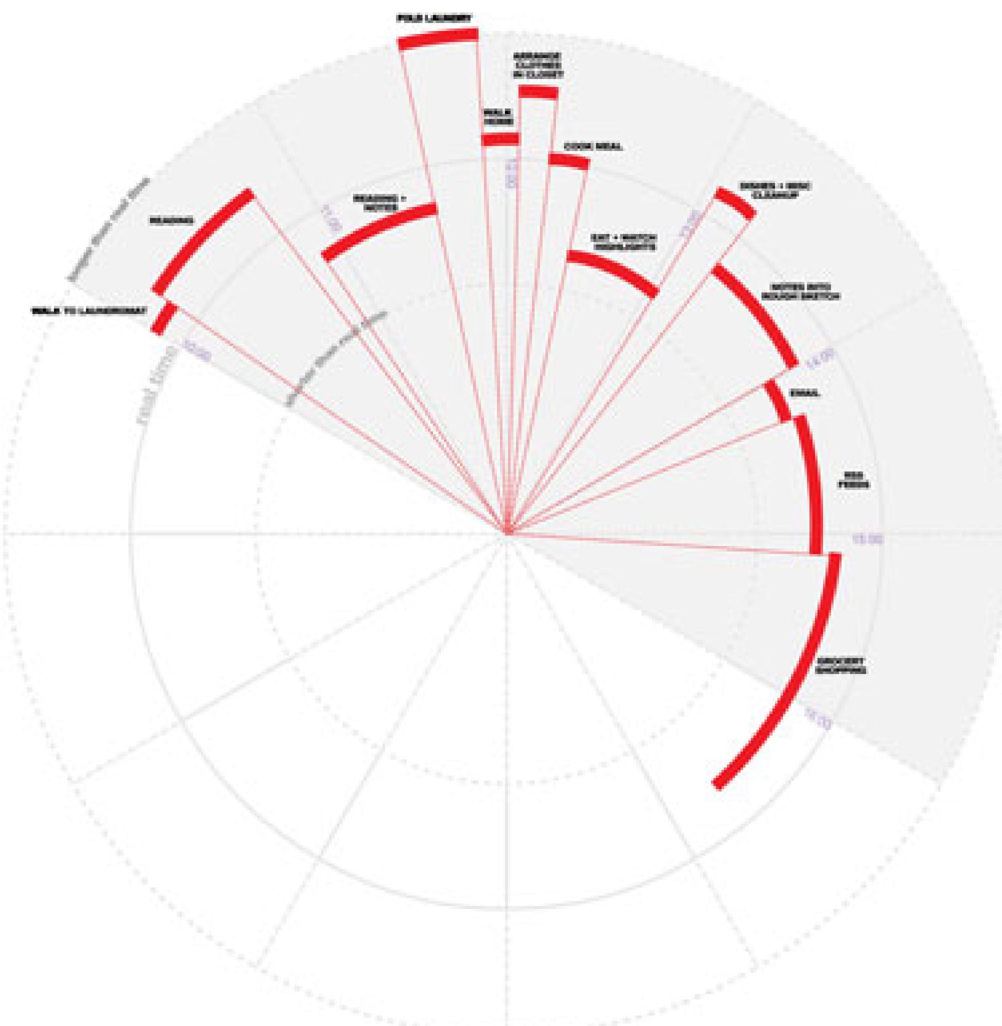

Figure 3-20: Time Perception Visualization, time tracking and logging class exercise (Illustrator)

The visualization in Fig 3-20 gives an idea of how segmented and irregular one's time perception is. This visualization is a result for a class exercise which required keeping a log of my activities for six hours, and trying to estimate the time each task took. The tasks are overlaid over the surface of a clock and span six hours starting at 10:00 am. The circle rendered with a continuous line and labeled ”real time” represents the actual measured duration of each task. The perceived duration is rendered with a thick solid red line and measured by the length of the corresponding arc. Chores such as folding laundry were

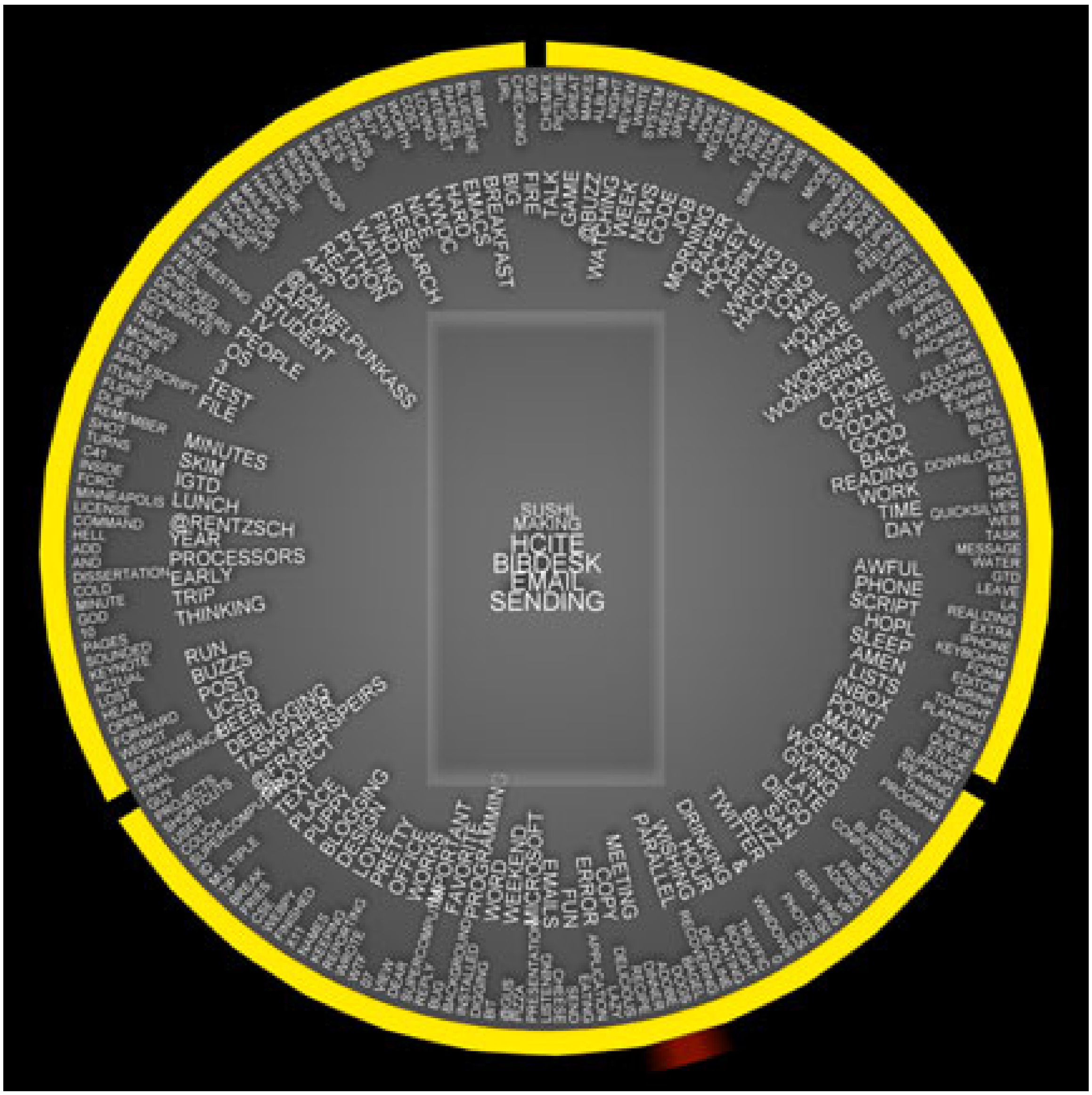

Figure 3-21: Clock Visualization Sketch for user mmc (Python, OpenMaya, Maya render)

perceived as longer than their actual duration, hence they were drawn on a circle with a larger radius. I could however be more accurate estimating duration of tasks that were shorter or spent near clocks such as cooking a meal or reading email. These latter tasks were drawn either on the same radius circle as ”real time” or on smaller radius circles to produce shorter than ”real time” arc lengths. This class exercise inspired me to use the clock as a metaphor for the information layout and as instrument to reinforce the unit of time. The resulting animated sketch is depicted in Fig 3-21. The most relevant words used by the user are arranged around in concentric circles. The hour arm is moved outside the body of the clock. Its rotation has a metronome-like function: to keep and indicate the exact tempo of days passing by. Each day is again mapped to two seconds. At the beginning of the two-second day, the words used that day move on the vertical strip in the center of the clock. Just before the arm completes a revolution, thus marking the end of the day, the words return to their original location on the concentric circles. Fig 3-22 shows experimentation with different aligning algorithms for laying out a large number of relevant keywords in concentric circles.

While this sketch reinforces the tempo of flowing data through time, some of its shortcomings is again the lack of a cumulative view at the end of playback, as well as means for highlighting patterns of word occurrences after playback. This can be partially solved by scaling the tokens based on their total number of occurrences, and also ordering them by scale on a spiral, instead of concentric circles. Establishing context for the words as well as separating posts during the day is another difficult goal to achieve for this sketch. For users with high volume postings per day, it is difficult to fit the words in the central vertical strip of the clock, while at the same time maintaining the proper cumulative scale for each token.

3.2 Amoeba

As shown in the previous section, although the clock metaphor is an effective mechanism for keeping the tempo, it breaks down when we think about the qualities of the data source. A clock of words seems an appropriate apparatus for marking time, but it is not suitable for portraying the organic and fluid qualities of the consciousness behind the words. In other words, the metaphor feels too mechanistic, and hides the fact that the source of the words displayed is an individual.

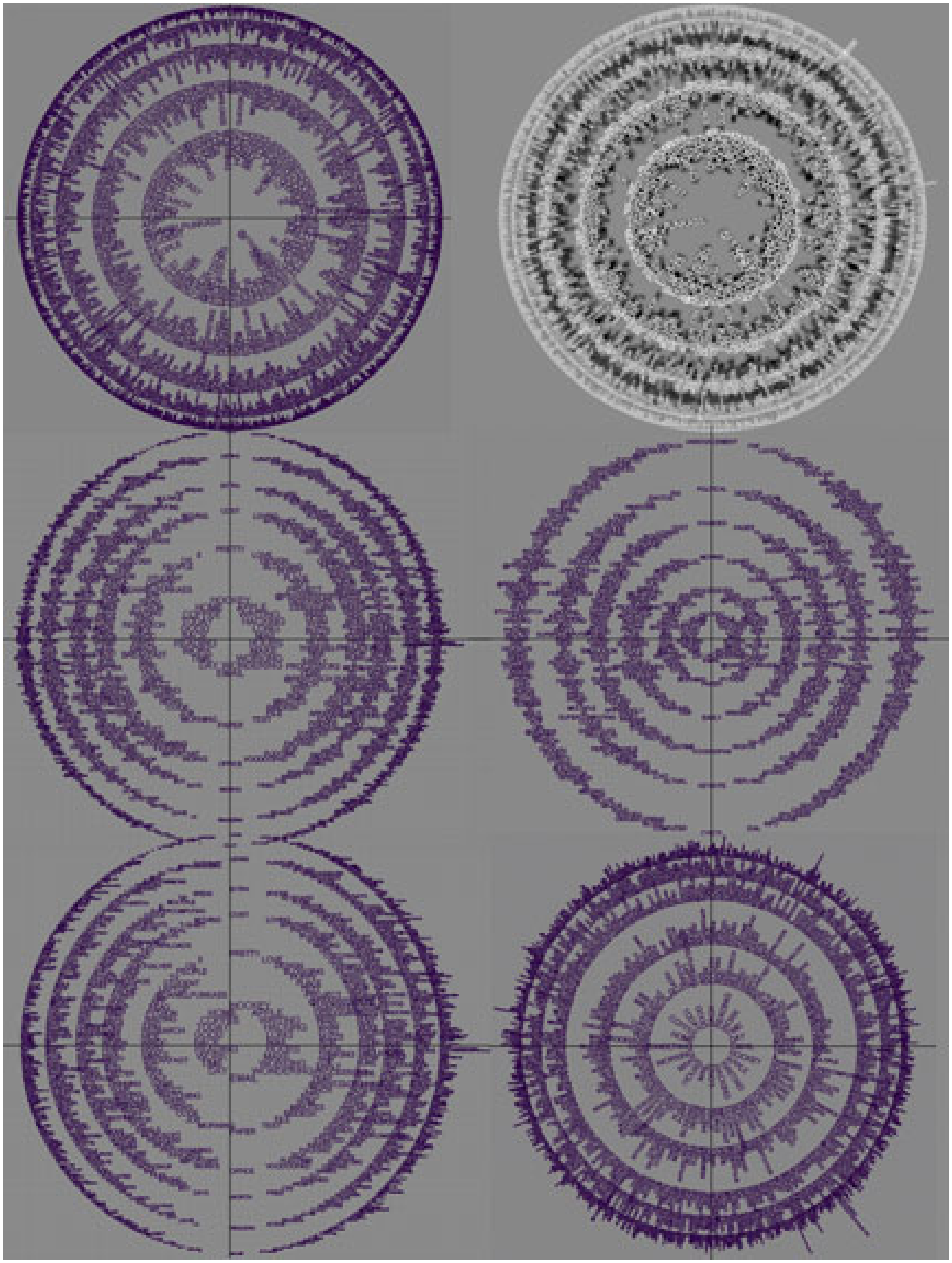

Figure 3-22: Clock sketch, various alignments *from top left, clockwise*: inner tangent to circle, inner tangent to circle rendered, horizontally centered on circle, outer tangent to circle, left side align to circle, horizontally centered on circle and scaled according to frequency (Illustrator)

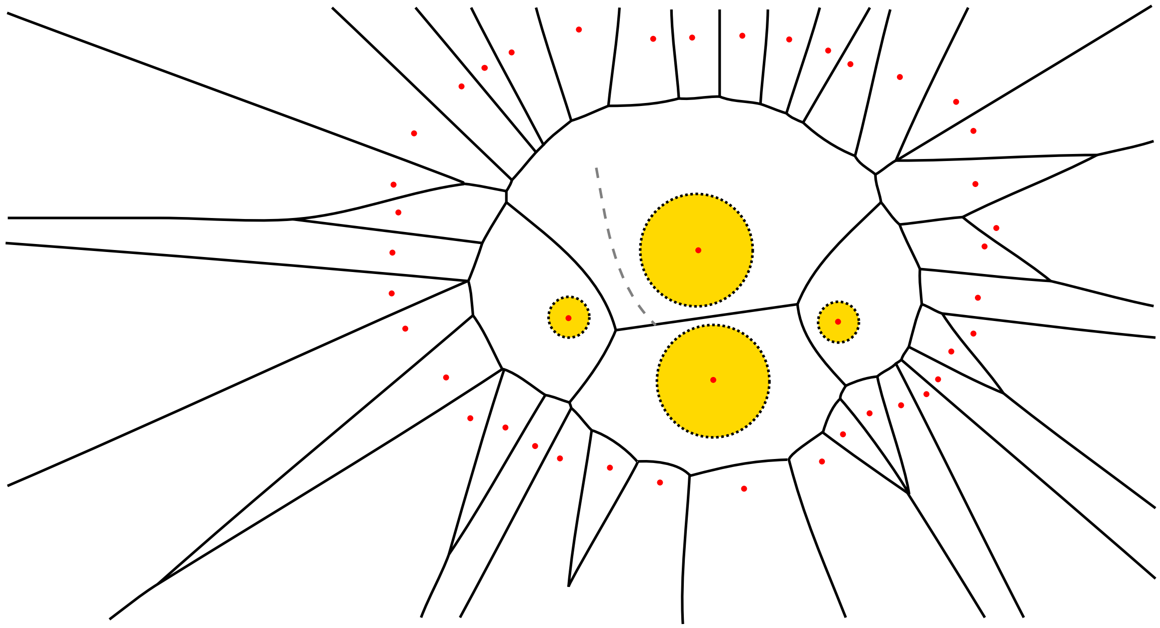

Figure 3-23: Additively weighted Voronoi diagram. The sites are drawn in red. The yellow disks represent the weight for each site. The dashed line shows the other branch of the hyperbola that is not used as a boundary. (Java)

The next iteration of sketches uses a *Voronoi* diagram [11] as the layout structure. I will show shortly how this metaphor mitigates the shortcomings of the clock metaphor outlined above. The Voronoi diagram is a popular space partitioning schema in computational geometry, and has applications in various fields. For example, my previous work as part of the *Scalable City* project [8] used Voronoi diagrams to divide an utopian suburban landscape into roads and house lots.

A Voronoi diagram is composed of generating points (also known as *sites*) and *cells*. Each site has a corresponding cell or region of influence. Each point within a cell is closer to that cell's site when compared to the sites of all other cells. The boundary between two sites always crosses through the midpoint of the line connecting them.

For this visualization, entitled *Amoeba* (an explanation of the name follows shortly), I used a particular version called an *additively weighted Voronoi diagram* (Fig 3-23) which adds a weight parameter for each site. This weight is represented as the radius of a disk centered at the same point as the site. The intersection of two Voronoi cells is now one of the branches of a hyperbola opened towards the smaller disk.

As in the previous sketches, the data used for the Amoeba visualization (Fig 3-24),

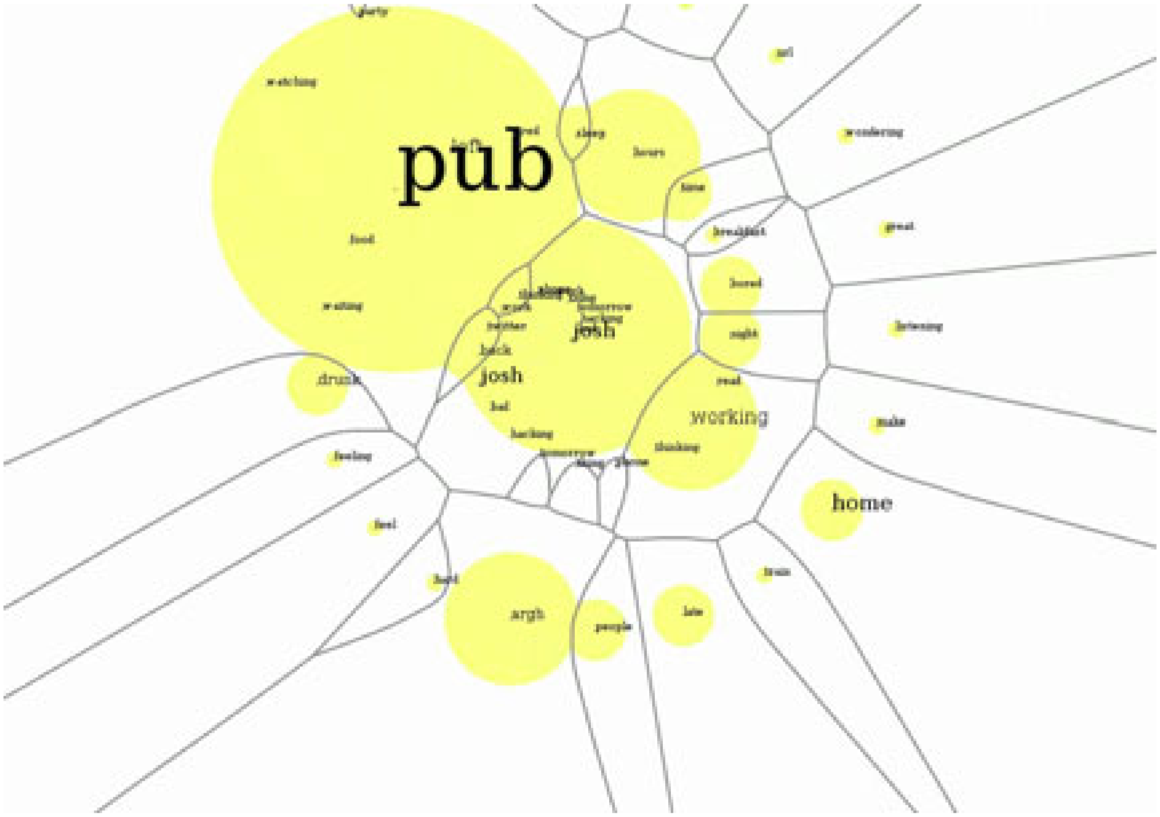

Figure 3-24: Amoeba visualization: (Java)

are the most frequently used words from the feed of a Twitter user. Each word maps to a site: the actual word is displayed at the site location while its frequency drives the weight of the site. The sites are arranged at equal arc distances on an arithmetic spiral, in the chronological order of their first appearance in the data stream. The layout of the spiral starts with at least one revolution offset, to avoid cluttering the center of the diagram.

This sketch was also animated by playing back the daily occurrences of words. The animation can be played using two modalities: cumulative and temporal. In the cumulative mode, the weight of the site and the size of the corresponding word increase with each occurrence. At the end of the animation, the diagram illustrates the number of occurrences for each word. In the temporal mode, the weight increases with each occurrence but also constantly decreases each day with a factor inversely proportional to the total number of occurrences. As such, words with higher frequency decrease in size slower, and if a user did not post for many days the diagram converges to the initial configuration in which all sites have minimal weights.

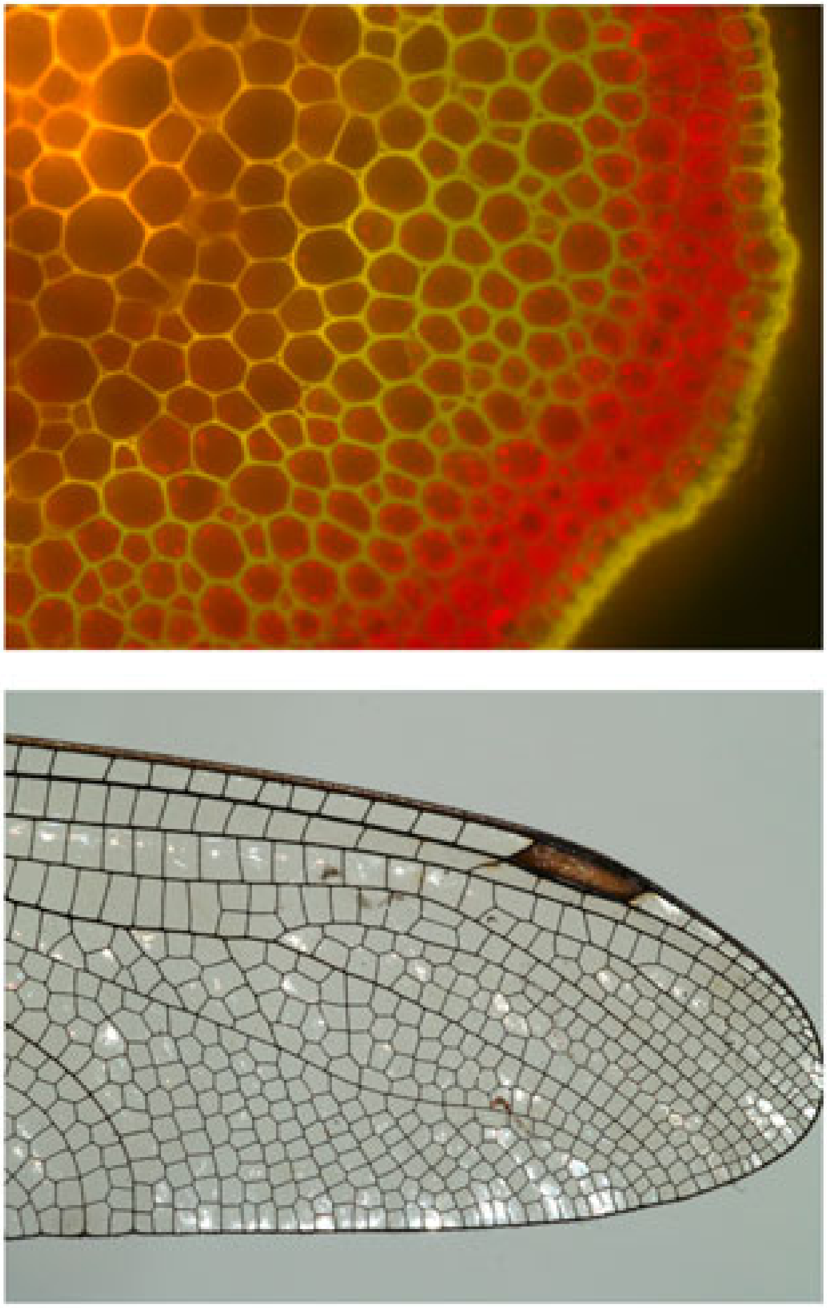

Figure 3-25: Voronoi diagrams in nature: *top: Psilotum nudum*, cross section of aerial stem, autofluorescence, reprinted with permission, Liz Kabanoff, University of Western Sydney, Australia, 2008, *bottom: Wing of a dragonfly*, reprinted with permission, Rolf Müller, www.rachaimer.de, 2007

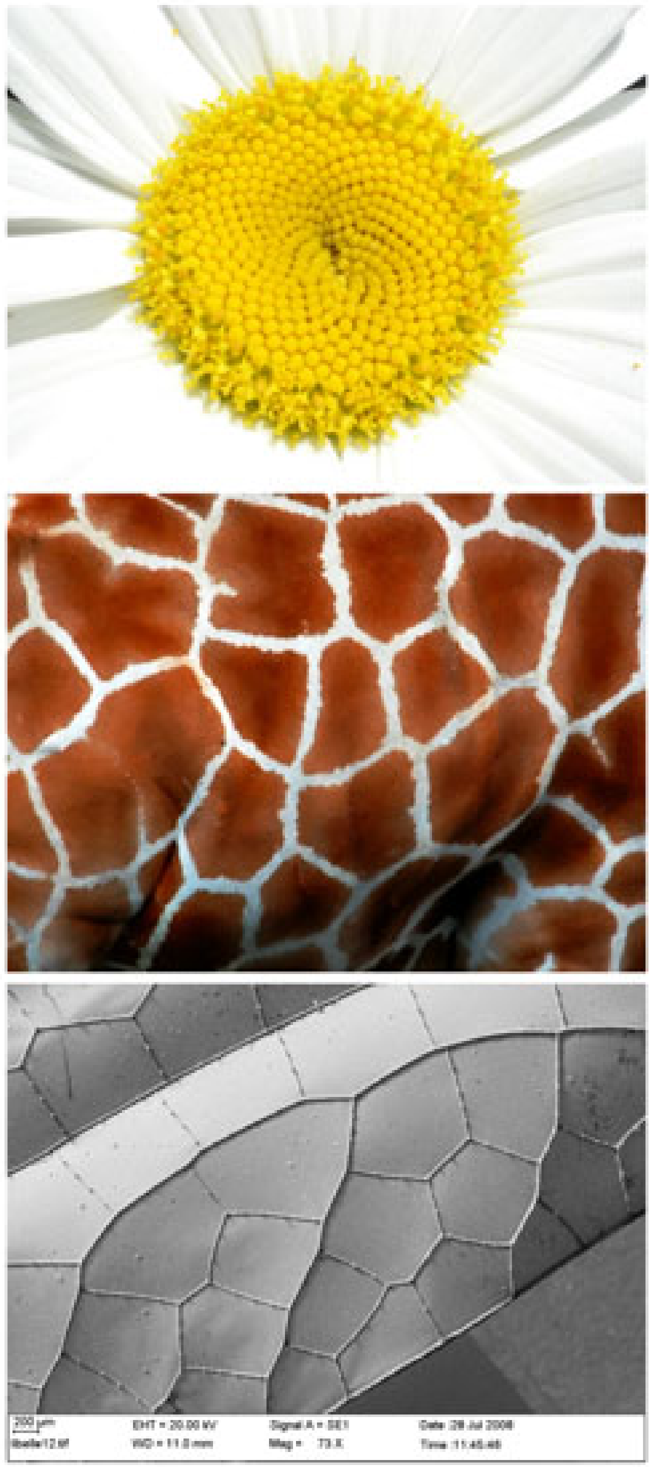

Figure 3-26: Voronoi diagrams in nature: *top: 7060103, daisy, close-up*, reprinted with permission, Allison J. Sebastian, 2009 ,*middle: Diagram of a Giraffe*, reprinted with permission, Scott Aldous, 2005 ,*bottom: Ophiogomphus cecilia*, *wing detail*, reprinted with permission, Andreas Weck-Heimann, www.seelensack.de, 2008

Both modalities described above, depict the user's posting patterns. Finally the last iteration of this sketch adopted a hybrid approach: the size of the type is driven as in the temporal modality, while the weight of the Voronoi site is driven cumulatively. Not only that this visualization captures the subject's posting patterns, but also produces, at the end of the animation, a unique print for each user visualized – given the unique method of aligning the sites and various frequencies for the corresponding words.

The weighted Voronoi diagram seems to be a perfect vehicle for the ”data organism” visualization metaphor. The diagram is the result of an exact computation that can be directly mapped to data and it constructs a form that feels organic and reminds the viewer of a multi-cellular organism, hence its title: *Amoeba*. Its appearance is familiar to us: similar Voronoi patterns are found in nature in various skin textures, tissue structures, or the formation of soap bubbles (Fig 3-25 and 3-26). We will encounter the Amoeba metaphor again in Section 4.3 where it is used to render a simulation of a group of Twitter users, in order to observe semantic similarity and social interaction.

3.3 Lexigraphs

Lexigraphs (Fig 3-27) is one of the last instantiations in the sequence of data portraits sketches and experiments. This visualization synthesizes elements from all previous sketches, borrowing devices that proved fruitful.

The salient words displayed in the Lexigraphs portraits are part of a different collection of words. Whereas previous experiments use tokens selected just by their frequency across all the posts in one Twitter micro-blog, the Lexigraphs portrait relies on three types of tokens:

- Single words as used in previous sketches;
- Bi-grams and tri-grams (groups of two and respectively three words);
- Twitter usernames (tokens that have the ”@” symbol in front of them).

The tokens with the largest *tf-idf* score [30] are selected to appear in the portrait. The *tf-idf* is a statistical measure of the relevance of a term within a collection of documents.

The corpus used for computing the tf-idf weight was constructed by randomly choosing all the message updates of around ten thousand English-speaking Twitter users. While for previous sketches stop-words needed to be explicitly separated, the *tf-idf* method implicitly eliminates stop-words by assigning a very low score to words that appear very frequently in the corpus. In addition, words such as ”twitter” or ”tweet” that are heavily used in Twitter micro-blogs are also automatically pruned because of their low *tf-idf* score.

Figure 3-27: Lexigraphs: portrait of Twitter user *ev* taken on (Actionscript, Flash Player)

The list of final tokens is duplicated and aligned to two curves that define a silhouette for the subject. The words are ordered chronologically by the date of their first occurrence, most recent words appearing at top of the curves. On the left side of the silhouette, the day-to-day rhythm is replayed through the animation of words, while on the right side, the cumulative score for each token drives the size of the type, which eventually settles towards the end of the animation. One can get a sense of the daily posting rhythms, as well as the user's interests and conversations. The terms starting with the "@" sign are usernames whom the subject of the portrait has mentioned or has replied to.

The tempo is maintained by several mechanisms. At the bottom of the portrait, a progress bar animation shows the passing of time. As before, one day maps to two seconds in the animation. The current day, month and year are overlaid in the center of the progress bar. On the left side of the silhouette, the words increase in size at the beginning of the two-second day and disappear towards the end of the day, thus creating a metronome effect – a mechanism borrowed from the *Clock* sketch. A line chart tracking the number of posts each day is animated and overlaid at the bottom of the portrait. On the cumulative right side, words exhibit a behavior employed in the *Wordtraces* sketch: when changing position or size they leave behind a copy at 1% transparency. This technique enables the highlighting of major shifts in token usage over the many months and sometimes years of existence of a micro-blog.

The inspirations for the Lexigraphs figure were the sculpted bust, the passport photo and generic user icons from various desktop or web applications. These portraits usually include the head, shoulder and chest of the subject. As such, Lexigraphs continues using the body metaphor that was employed in Amoeba. However while the representation of the body in the Amoeba visualization can be interpreted as a more abstract portrait, a chart or a partitioning of space into areas, in Lexigraphs the figure is more evident and literal and clearly signals the data body of an individual, fleshed out with their specific content and rhythms.

In the structure of the Lexigraphs data layer, the Twitter user icon image is used in the bottom most layer, immediately underneath the layer of the left curve and it is assigned a *multiply blending mode*. Hence the colors of the image get picked up in the words of the left curve providing a simple and effective way to personalize and differentiate this portrait from others, as well as further the separate the two sets of words on the left and right. Displayed as a group, these portraits feature the individuality of each subject and highlight the external events that unite them, as discussed in section 4.1.

3.4 Data Ghosts

In the Data Ghosts visualization (Figures 3-28 and 3-29), the two-dimensional representation of words and rhythms from the Lexigraphs series has been translated into three-dimensional space, resulting in a virtual sculpture of words in perpetual motion. This visualization has been part of an interactive art installation entitled *Metropath(ologies)* [13], designed by the Sociable Media Group and first installed at the MIT Museum.

When visitors are walking into the Metropath(ologies) installation space, they see themselves on a computer monitor filmed by a surveillance camera. From time to time, another presence in the scene is revealed - a ghostly figure, emitting a stream of letters and words, appearing to move into the same physical place occupied by the visitors. The digital personas and individuals in real life are sharing the same space within the installation (Fig 3-30). The avatar's motion is picked randomly from a database of motion clips. The words it emits are the most frequently used words in a randomly chosen Twitter micro-blog; each animation clip corresponds to a different Twitter user. The avatar is rendered from a virtual camera that matches the position and attributes of the live video camera, including static masks applied over the avatar video clips that simulate occlusion by the real columns that are virtually located in front of the avatar. The masked video clips are mixed in with the real-time camera video stream to create a space in which visitors co-habit in real-time with the "data ghost".

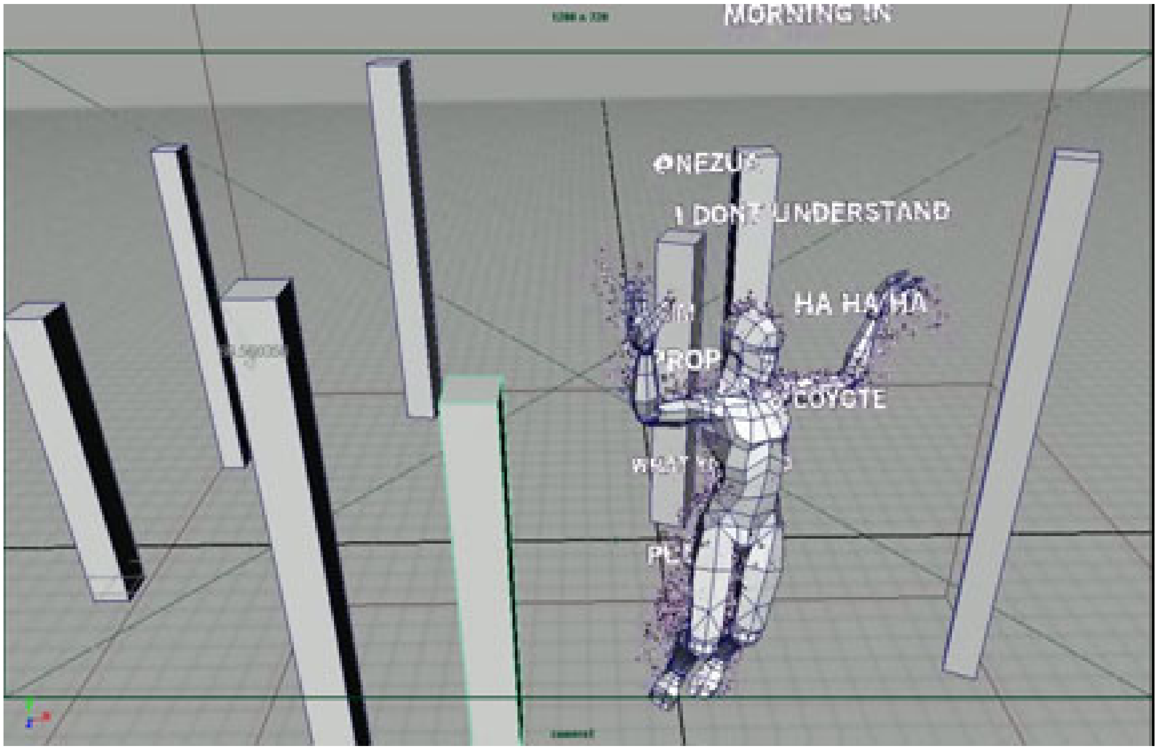

Figure 3-28: Data ghosts avatar wireframe

Figure 3-29: Data ghosts avatar rendered version

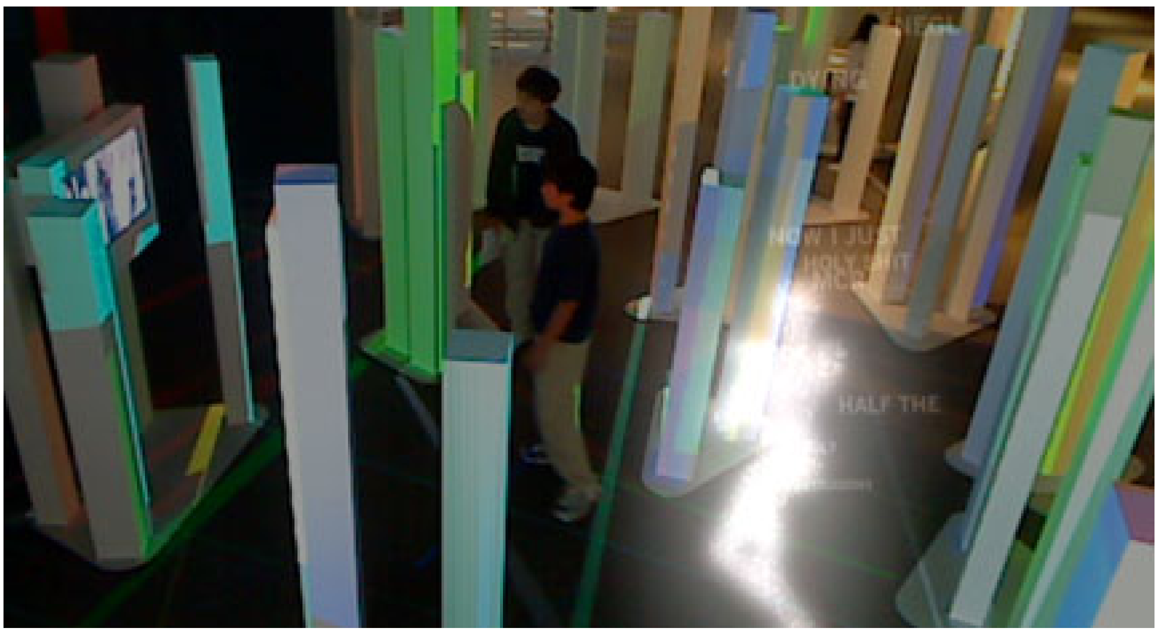

Figure 3-30: Data ghosts screen

Chapter 4

Collective Portraits

The information used to create the data portraits presented in this thesis is intrinsically social. Twitter users have conversations with each other, they mention the name of other users with whom they participate in both off-line and on-line events, or *re-tweet* (repost) each other's messages. Twitter users, in a sense, produce performances for their followers, and sometimes even for a larger public audience. The collective portraits described in this chapter intend to evoke both individual and group traits by focusing on the social facets of data.

Richard Brilliant, in his book *Portraiture* [7], highlights an important tension between the individual and the group in traditional seventeenth-century Dutch portraiture:

> ”In Dutch group pictures, the integrated ensemble may prevail over the independent individual, but in them the strong emphasis on the realistic depiction of specific individuals permitted each person's portrait to compete for close, if momentary, visual and psychological attention. The Dutch artists seemed to compel the viewer's eye to move from face to face never losing sight of the others in the contemplation of the one. Such works are peculiarly cooperative and collusive in their nature, because each person in the group contributes to, and draws from, the presentational dynamic of the whole.”

One's character or persona can be further portrayed by foregrounding the subject's behavior in a social group, or by focusing on the narratives or relationships involving other people. The group acts as a common denominator for the depictions of individuals. Both the group and the individual are inherently linked in contouring each other's identities: the group gains traits from the sum and affinities of the participants, while the individual receives further features through association with the group.

This chapter describes three types of collective data portraits. The *Lexigraphs* visualization described in Section 4.1 is closer in nature to exhibits in museums or galleries, where portraits are displayed as a series in order to invite audiences to look across subjects, artistic styles, or time periods. The *Cityscape* (Section 4.2) and *Ecosystem Simulation* (Section 4.3) are more akin to group portraits, where social activity within the group is made visible together with individual characteristics.

4.1 Lexigraphs

The Lexigraphs collective portraits (Fig 4-1) is a continuation of the single-user interactive visualization described in the previous chapter. Multiple Twitter users are portrayed together, their timelines and post streams synchronized with each other. The portraits are bin-packed onto the screen randomly. Their size is proportional with their total number of posts. Their animated silhouettes fade in on the screen in the order of their first post. In the top left corner, the date changes every two seconds, setting the tempo. The silhouettes fade out slightly when the corresponding user has not posted in more than a month. The audience can compare and distinguish each user by their various color schemes, posting rhythm, message volume and perhaps most importantly, topical rhythm and distribution. Various events, such as the presidential debates or election can be easily identified as representative keywords ripple through the silhouettes of the participating users.

4.2 Cityscape

Previous sketches of data portraits (e.g., Lexigraphs, Data Ghosts) use the body as a metaphor to signal to the audience that the trail of information from online communications constitutes our virtual body via that we are known for and judged by. In Cityscape, I experi-

Figure 4-1: Lexigraphs, collective portrait, *top:* frame from animation, 31 randomly bin-packed users, *bottom:* detail(Python, OpenMaya, Maya render)

ment with an architectural metaphor, proposing the idea that the data we generate online is the result of us living in various habitats, using various tools. We often think of data as ephemeral and intangible however in reality, the data lives on the *cloud*, is replicated across server nodes and across Internet platforms and domains (e.g., forum, blogs, file sharing and hosting services). There will always be some instance of our data stored online in the index of web crawlers, such as Google cache, or versioned at the Internet Archive. Data is hardly ever deleted due to performance and semantic reasons. First of all, deletion from a relational database systems is an extremely expensive operation, as it involves cascading to multiple relational tables and reorganizing multiple indices. Hence, "deleted" data is usually simply marked as hidden and not physically removed. Second of all, once posted, information quickly becomes interconnected hence difficult to remove without breaking conversation flows. In social web applications, for example, deleting a user account does not usually trigger the deletion of their posts without which comment threads would become incomplete. Therefore, data can be thought of as persistent, with an almost solid quality, and can be represented as such, using architectural logic.

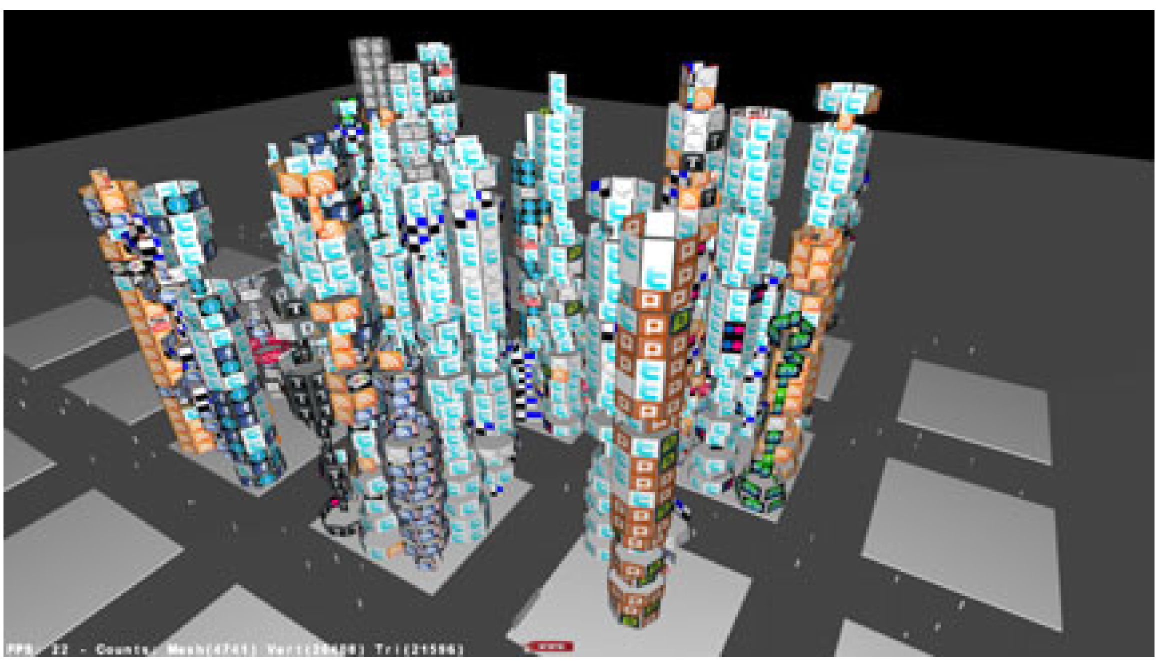

Figure 4-2: Cityscape real-time application: FriendFeed users, visualized with logos of feed channels (Java, JME, LWJGL)

Fig 4-3 shows a point of view into such an architectural rendering of the FriendFeed

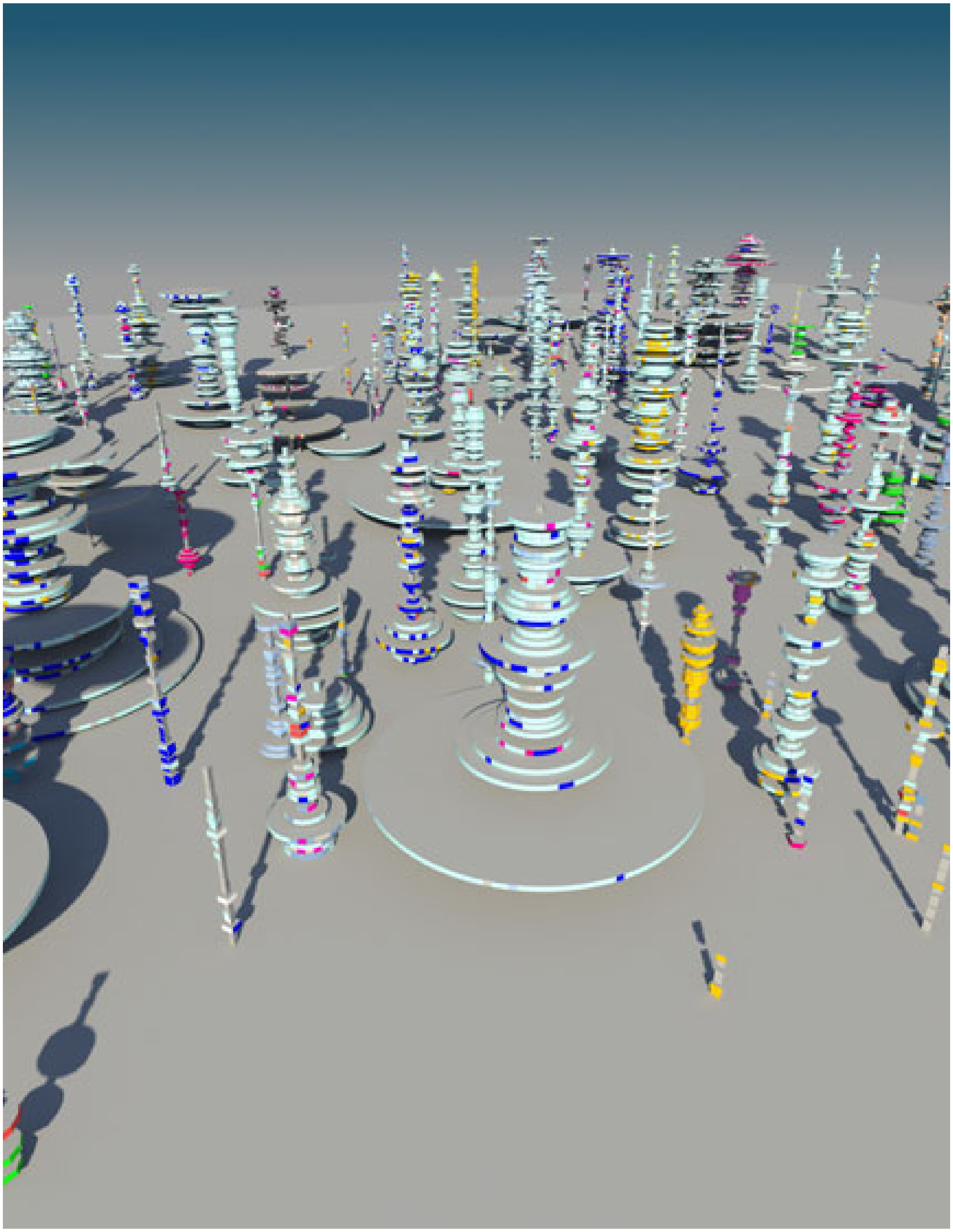

Figure 4-3: Cityscape render: 200 random FriendFeed users, color coded structures according to feed origin (Python, OpenMaya, mentalray)

habitat [9]. FriendFeed is a social web application that acts as an RSS feed aggregator, a place to collect one's activity across habitats: photos from Flickr, messages from Twitter, posts from Tumblr or other blogs, wishlists from Amazon, etc. Each tower is the architectural representation of one FriendFeed user. These towers of information have a chronological structure and are constructed of circular discs representing the activity for one day. The radius of each disc is proportional to the number of posts of the corresponding day. The number of daily posts is also expressed in the number of facets in the polygonal cylinder. All daily disc or ”floors” in these towers have the same height. Days without posting activity do not generate ”floors” instead leave slight gaps to further advance the height of the tower. The posts within the daily structure are represented using three modalities:

- By projecting onto each facet the logo of the corresponding RSS application (Fig 4-2);
- By coloring each facet into a distinct color, borrowed from the logo of the RSS application (Figures 4-3 and 4-5);
- By rendering the actual post with text or image (Figures 4-6 and 4-7).

The Cityscape visualization is in fact animated and realized two different ways: movie form and using a real-time OpenGL engine. While the former allows for large quantities of information to be rendered in one screen, the latter allows for instant, interactive viewing and navigation. The architectural forms allow users to perceive posting rhythms, volume and types. Once materialized as architectural form, the habitat can be navigated and experienced at both macro and micro level.

The social connections between users, as well as the interconnected nature of data are not very well expressed in this visualization, and need to be further explored. For example, in the current renders the towers are randomly grouped into rectangular city blocks (Fig 4-2), or bin-packed randomly based on their largest radius, on the horizontal plane (Fig 4-5). However, some social aspects of data are still represented in the scale and length of the facet representing a feed entry. An entry on FriendFeed can be commented or liked. The more comments an entry has, the larger the surface of the facet. The number of ”likes” is encoded into the length of the facet protruding away from the center of the tower (Figures 4-6 and 4-7). These data mappings allow to differentiate between users who exhibit social behavior and users that are lone aggregators.

Cityscape is an example of portraiture where each subject has an abstract architecturalized body with specific properties. The collective aspects of this visualization allow the audience to further extract meaning by comparing across the individual, social and chronological dimensions. An interesting aspect that Cityscape reveals, and invites the audience reflection upon, is the fact that our mediated communication is done almost entirely through business-owned channels, as signaled by the landscape of logos.

4.3 Ecosystem Simulation

The Amoeba portrait described in Section 3.2 was also extended into a collective portrait (Figures 4-4 and 4-9). The intention here was to augment the personal data in the portrait with social information, reproducing in more detail the interactions with other users. The metaphor of the user as a biological organism maintains its validity in group mode. The animation evokes an ecosystem of organisms interacting under a microscope. The simulation places members of a social network randomly outside the screen. All users are synchronized to a unique timeline which is being played back at the usual rate of two seconds for each day. Each user's organism unfolds based on the daily posts, as explained in Section 3.2. A force of attraction is applied between two users if they mention each other's names, or use similar words in their posts (stop-words are excluded).

We usually experience the Twitter users as separate entities, on separate pages. Conversations are visible but make sense only after a page jump. The Ecosystem simulation attempts to overcome this aspect by allowing one to follow parallel and similar content. Semantic proximity translates to spatial proximity. Various narratives among two or more users are revealed by proximity and highlighting of words in their user bodies.

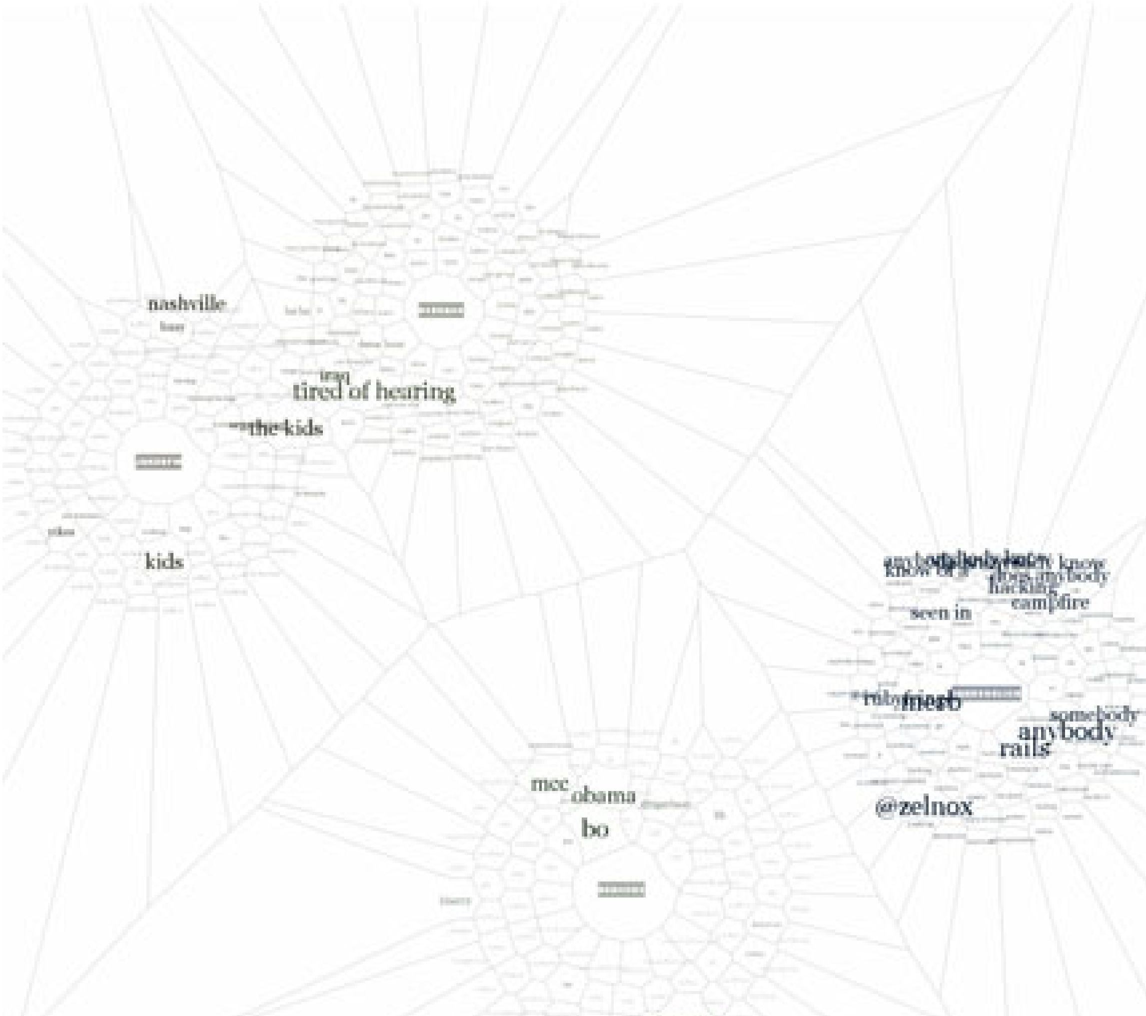

Figure 4-4: Ecosystem visualization: Four Twitter users, frame from animation, forces of attraction not visible (Java, Java2D, phys2d)

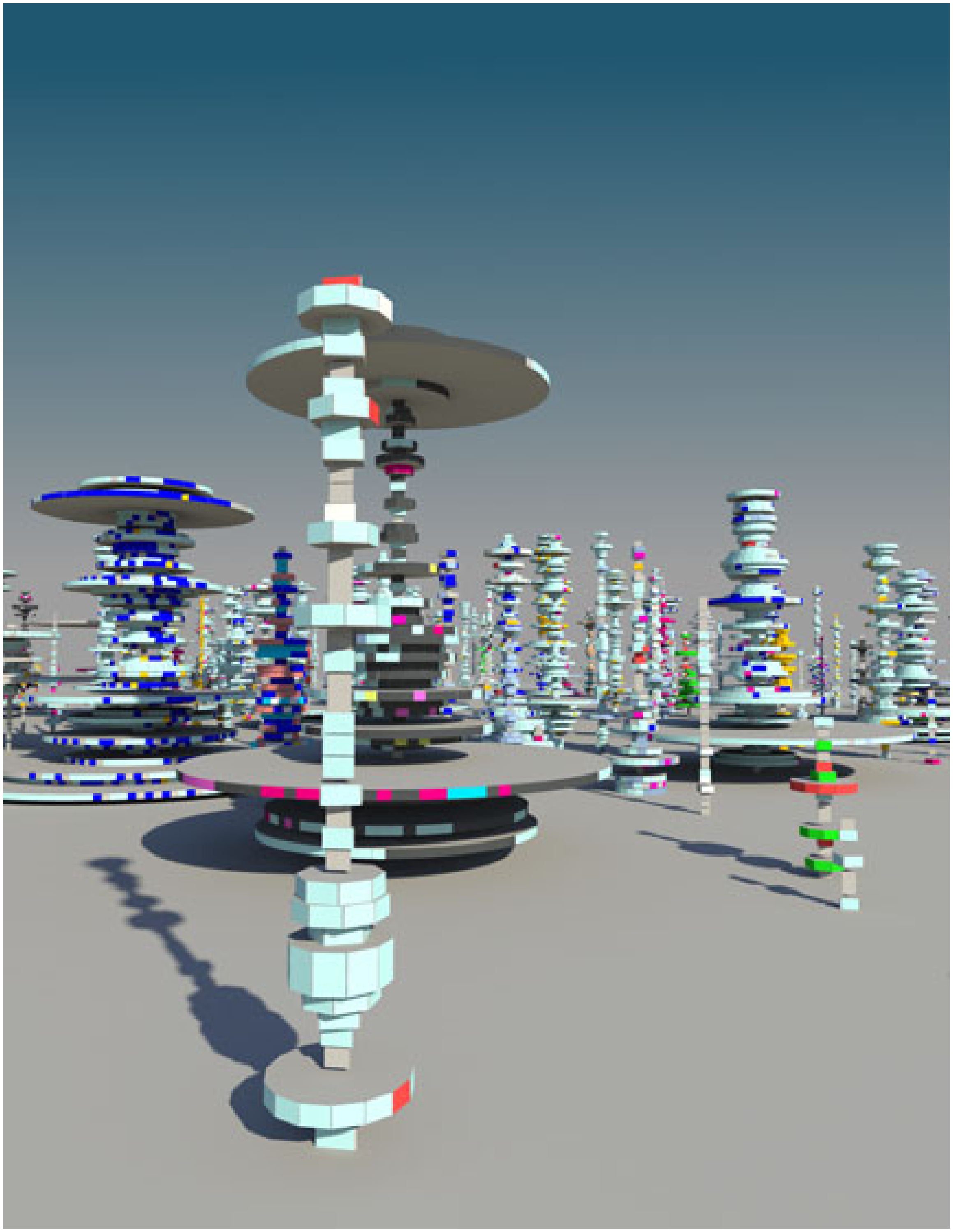

Figure 4-5: Cityscape render: 200 random FriendFeed users, alternate view, color coded structures according to feed origin (Python, OpenMaya, mentalray)

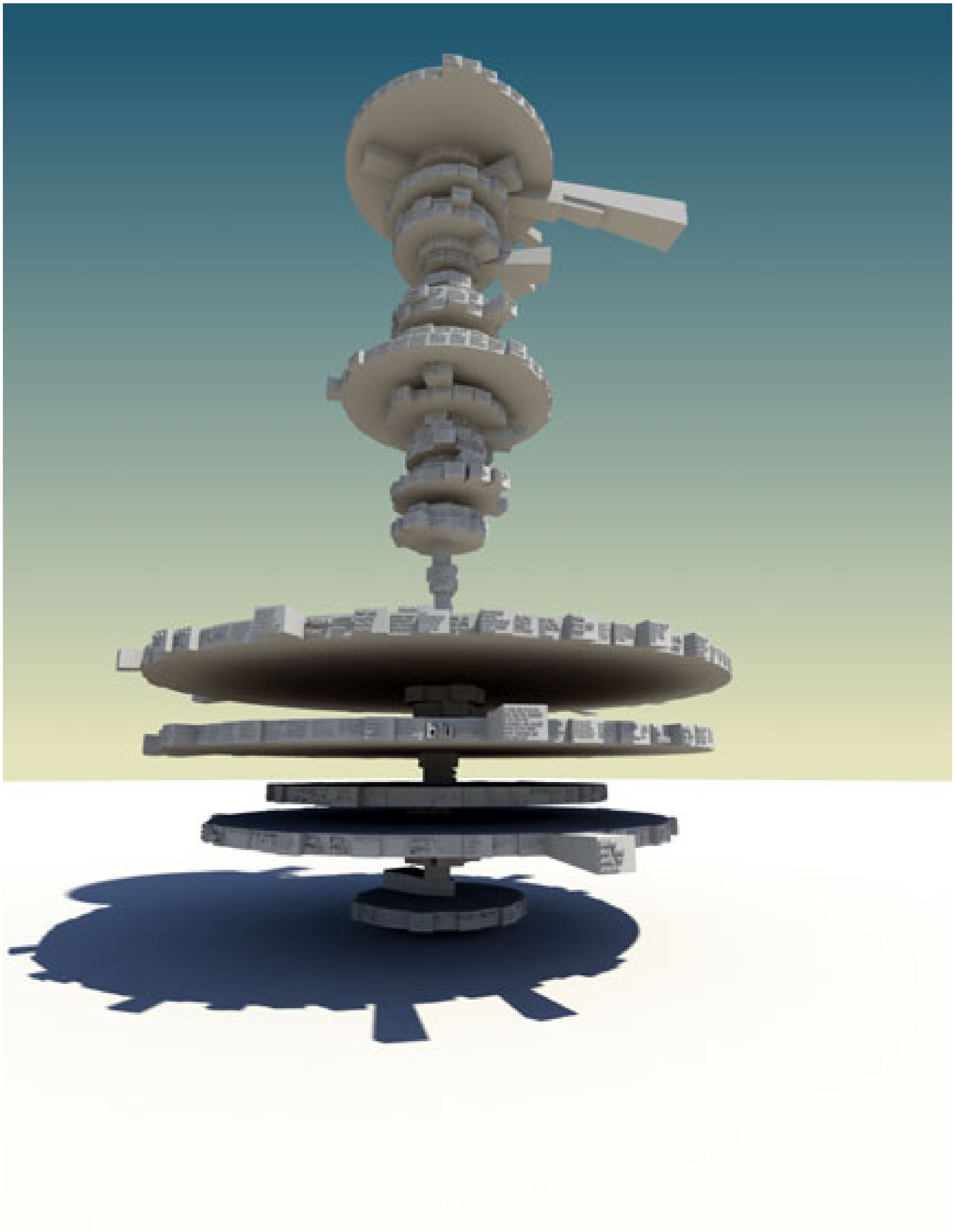

Figure 4-6: Cityscape: FriendFeed leolaporte user, posts rendered with text (Python, Open-Maya, mentalray)

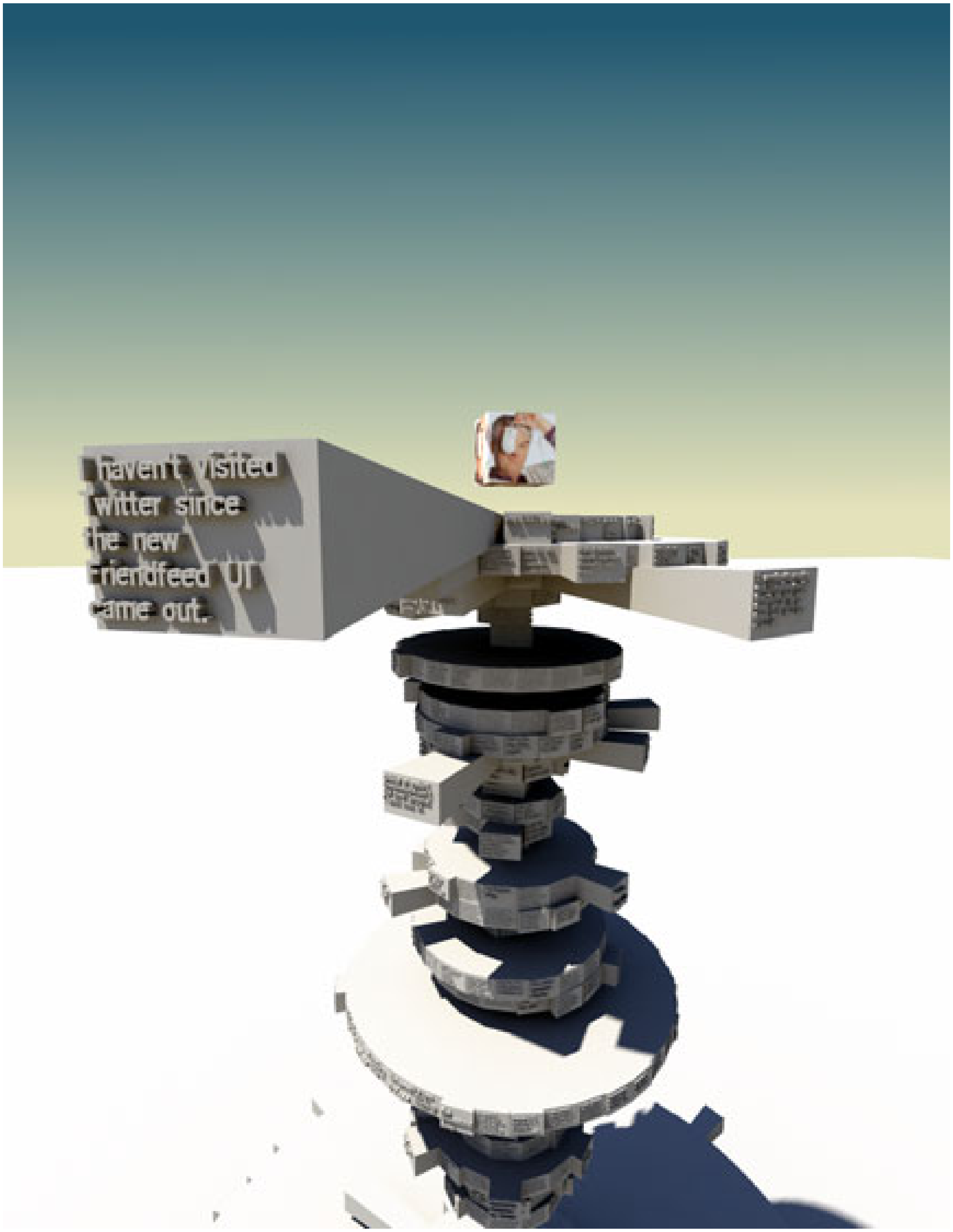

Figure 4-7: Cityscape: FriendFeed leolaporte user, detail, posts rendered with text (Python, OpenMaya, mentalray)

Figure 4-8: Cityscape, Metropath(ologies) installation view *top, bottom:* FriendFeed users, visualized with content of posts (Java, JME, LWJGL)

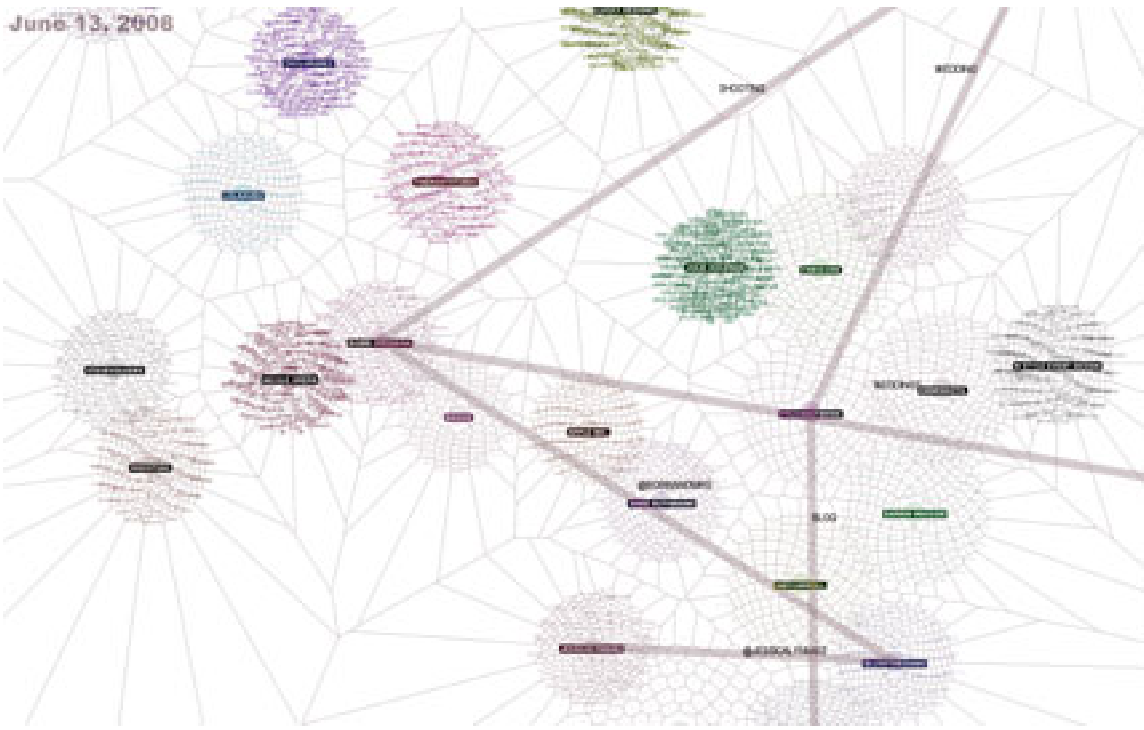

Figure 4-9: Ecosystem visualization: social network on Twitter, lines show forces of attraction based on common words or usernames from conversations, frame from animation (Java, Java2D, phys2d)

Chapter 5

Critique

The work presented in this thesis finds itself at the intersection of art, design and science - sometimes using artistic strategies to challenge the viewer, sometimes using design devices to communicate, and scientific methods to collect and summarize the data. Critique is an established form of evaluation in visual art departments, design boutiques, or animation studios. A recent paper by Saul Greenberg and Bill Buxton[20] highlights the benefits of the critique process even for the field of Human Computer Interaction, where usually the norm for evaluation is the usability study. In a critique session, the artist or designer is being challenged by a series of questions. The critique panel conveys likes, dislikes, suggestions, and brings into conversation related work. This probing and constructive criticism allows the author to gain a deeper understanding of the work being proposed.

My critique panel was composed of my thesis readers, Judith Donath, Martin Wattenberg and Caroline Jones (their bios are attached in the Appendix). I have first presented the work, and the discussion followed.

Among the first things that both Caroline and Martin commented on in context of the Lexigraphs portraits, was that their shapes were very similar and some of the content was not very interesting to look at. The Amoeba portraits were more visually attractive, as the configuration of the body was affected more acutely by the corpus of words. Martin suggested to try to inject a stronger editorial stance in these portrayals, by choosing and highlighting ”things that have an edge to them, would provoke the viewer, make the viewer uncomfortable, or make the viewer say: oh I didn’t know that”.

Caroline urged me to think about the position and the role of the artist towards these very voyeuristic representations as the people did not asked to be portrayed. The portraiture analogy comes with certain criteria attached. For example, the power dynamics and relationship between artist and sitter which is minimal in these data portraits. Judith intervened and brought up the idea that using the term *portraiture* is also a provocation to rethink and redefine the portrait in the twenty first century, in an age of public electronic personas. Judith also brought into the discussion the reappropiated images in the Andy Warhol portraits, as a way to highlight the difference between the process of private sitting and the process of portraying people via their data, something more akin to taking snapshots of public performances.

Martin remarked on the good timing of this work in adding to the discourse around Twitter as it is becoming mainstream (Oprah opened an account!) and part of the national conversation. I explained how the interactive instantiation of Lexigraphs is intended as a mirror for users to see themselves. Caroline suggested that embedded the portrait in the home page as an avatar, as a representation on one's self, might be useful.

There were some questions about the timing of the animation and I realized I could have presented more clearly how the rhythms are represented. Judith made a commentary on how ”people have no days or nights” in electronic communication: we are always plugged in, the day is one unit, no morning, afternoon or evening.

To extend this conversation, Martin added how the process of creating the portraits is similar to ”building a new type of camera and then taking pictures with this camera”. He urged me to ”turn this camera towards something you care about... do some portraits of people you have some emotional connection with”. Another interesting experiment would be, for example, to take a person I know very well and tweak the parameters and filtering algorithms to enhance and capture some aspects of that person that would be lost through the generic, universal parameters.

Caroline mentioned Etoy's *Mission Eternity* series [18] as a project that deals with our society's obsession with recording data, and in a larger sense, with our sense of memory and time. She asked me to think about my attitude towards these portraits. ”Are these portraits ironic, abstract, euphoric or techno-euphoric?”

Judith remarked that I needed to articulate what changes when I re-frame these portraits using different metaphors (architecture, body, or organism). ”What is it that these different metaphors bring out in the way we think about the data?”

The critique was very helpful in pointing conceptual issues that needed more investigation and articulation, and suggesting possible future directions for this work.

In addition to the critique session, I also conducted informal interviews with museum visitors who participated in the Metropath(ologies) exhibition and have looked at the Lexigraphs portraits. Below are some answers to my question ”What was your experience like, and what were your thoughts when looking at these portraits?”

> ”I could see in parallel different patterns of how people use Twitter... some erratic some always posting. I don’t feel guilty for not updating as often... Sometimes I ask myself after I post: am I spamming everyone, am I posting really boring stuff?”

> ”In the evenings sometimes I pass on the streets and I really like looking into other people’s apartments. This was kinda the same... looking into a window in their life.”

> ”It was interesting to see what people are talking about ... one is always heading somewhere, one is processing data, one is walking the dog and tweeting...”

> ”Sometimes you didn’t get an impression of what that person feels towards a certain things. You would just see the words ”the Lakers” but you couldn’t tell if they hated or liked them.”

> ”I wanted to see my portrait and the portraits of my friends and see if we talk about the same things.”

> ”I think I am going to lock my Twitter accounts and only let my friends see my messages.”

> ”It kinda sickens me all this attention to our selves... technology has turned us into these egocentric maniacs really no one cares about what you are doing this second.”

> ”It was funny ’cause you could distinguish certain types: the dad, the student, the sports-obsessed person ... and see how certain things dominate their lives or how they keep coming back to the same few things: the kids, Obama, their favorite TV show.”

I also found an interesting reflection on the Lexigraphs series in artist and designer Lou Susi’s livejournal entry [32]:

> ”*Data Portraits* took the unsuspecting tweets of the twitterverse to create a portrait visualization of each user ... words { the smallest common denominator allowing some balance between private and public exposure } from the user — participant’s tweetstream make up the outline of head, neck and shoulders ... words on the left are from recent tweets, words along the right of the data head silhouette are the most-used throughout your tweetstream history ... the words of each portrait pop forth from the black backdrop much like a smooth data-persona tag cloud, quite literally outlining your word stream of the moment for you
>
> what strikes me the most about these linear-textual gestural snapshots is the cocoon like and ghostly bodily presence of each figure ... there’s also a wonderful sense of swirling ... the words seem to envelope or mummify a pre-set human form ... besides certain key words that pop out { i am guessing the word size follows the same sort of rules of frequent use that most tag cloud methodologies implement }, there is little differentiation from portrait to portrait ... the shape of the head, neck and shoulders remains the same ... and the words simply outline or ’contain’ the previously human form
>
> at first i thought that the **datamummification** might be a purposeful artistic and aesthetic choice ... i don’t think i get the sense that my portrait would look that much different than anyone else as far as physical attributes are concerned

... same height, same weight, same shape, same lack of eyes, mouth, ears and hair ... you are your words in these portraits ... you are the ghostly echo-trace of your micro-bloggings ... a bit sobering ... a little scary ... and unless you are lucky enough to have micro-entered some emotionally-laden and unique words over the last year or so, you are just as unique as everyone else on Twitter ...

part of me wants to think these implications are an intentional affect of the visualization as portraiture ... and if nothing else, perhaps we can see this as a subconscious expression of the artists involved ... maybe there is no true participatory auto-magical means to create this sort of portrait ... or perhaps the effect is completely intentional ... a statement about machine-produced { app-influenced } human behavioral modification ... the media we use shapes our behavior, and now we quite literally all itch to tell it all right now ... a sort of electronic OCD ..."

Chapter 6

Conclusion and Future Work

I have presented several automatic data portraiture techniques for generating individual and group portraits of users participating in social media. These portraits use mainly unstructured text and timestamps from Twitter micro-blogs, as well as aggregate RSS feeds from FriendFeed.

I have argued that the *data portrait* is an emerging concept that deserves attention and needs to be materialized for two main reasons. It can be used as a mirror to reflect one's data body – the collection of their online personal information trails. Furthermore, it can serve as a vehicle for aesthetics discourse around data ownership and control, privacy, and surveillance in our society.

I have looked at important moments in the history of portraiture and analyzed the ever-changing dynamics between artist, sitter and audience, in order to understand how the data portrait genre is similar and different from its traditional counterpart. I have showed and discussed the conceptual, perceptual and technical challenges of the data portrait-making process centered around an algorithmic visualization pipeline. I have focused on the compression, mapping and visual representation components of this pipeline, and less on the collection and mining of all possible types and sources of information one generates when interacting online. I have described my explorations and sketches, and discussed typographical and animation devices that can be employed to visualize topical and temporal rhythms. While doing so, I have highlighted the key decisions of the data portraiture pipeline, and showed how it can allow for trade-offs between subjectivity and expressivity on one side, and legibility and recognizability on the other side.

This thesis shows how data portraits can contain parallel intentions, in varying degrees: *utilitarian* on one hand, presenting patterns and rhythms in data in the spirit of scientific information visualization, and *aesthetic* on the other hand, provoking thought and raising awareness as art pieces. This thesis also reflects on the *mechanical artist*, the collection of software that acts as an intermediary between the artist and audience, and in the same time blurs the lines between art material, art instrument, art experience and art object.

The techniques I have employed show how important the selection of metaphor is, for both cognition and as a perceptual conduit, but also as an instrument for artistic statement and commentary. For example, the more aesthetically pleasing and engaging Amoeba portraits present an utopian view of Twitter as an organic, almost lyricist environment. The Lexigraphs collective portraits show users as bodies with similar, generic shapes, and might be perceived as a more dystopian view of a media landscape populated by clones. Metaphors can be employed to provoke the viewer to think about the qualities of the data. The artist can provoke further stronger reactions from the audience if there is a mismatch between the qualities of data and the scope or coverage of the metaphor. In addition to the metaphor, the artist can embed commentary and statements in these data portraits by certain choices during the different stages of the visualization process: the means of data collection and the types of data collected, the choice of algorithms and parameters for filtering and condensing the information, the mapping between data and entities of representation and the rendering style of the visual metaphor.

Two main future work directions involve further investigating the utilitarian and the aesthetic functions of the data portraits. On the aesthetic side, I would like to explore other metaphors and employ more radical algorithmic gestures to potentially reveal unexpected facets of personal data and relationships in social media environments. One of the reoccurring phenomena in the interviews with museum visitors, is the intention to reassess one's behavior in these mediated habitats after experiencing the collective portraits, which prompts for further investigation of the utilitarian side of the data portrait as a mirror for one's data body. Towards this goal, I see several possible extensions. For example, it might be interesting to make data portraits interactive by adding multiple layers of data types and sources, and allowing the user to select the layers of interest. Furthermore, it would be interesting to allow the user to switch between different lenses, or algorithmic interpretations of the data. For example, superimposing a sentiment analysis layer over the content would enable the audience to understand how the subject of the portrait feels about the topics being discussed. In environments with both private and public profiles, it would be interesting for users to analyze and compare their public versus private personas. In the case of the collective portraits, it might be useful for the audience to interactively select and group certain users of interest, in order to observe how the relationships between them play out. I would also like to further investigate the architecture metaphor and develop a real-time navigational interface that will allow new users to quickly understand and explore such online habitats. This would be extremely useful in the context of various current social media services such as Twitter or FriendFeed, whose home pages contain explanations of what each is intended for, but lack an intuitive mechanism for introducing a potential or new user to the full capabilities of the environment.

Appendix: Critic Biographies

Judith Donath is the director of the Sociable Media research group at the MIT Media Lab and a Faculty Fellow at Harvard's Berkman Center for Internet and Society. Her work focuses on the social side of computing, and she is known internationally for pioneering research in social visualization, interface design, and computer mediated interaction. She created several of the early social applications for the web, including the first postcard service and the first interactive juried art show. Her work with the Sociable Media Group has been shown in Boston's Institute for Contemporary Art, in several New York galleries and is now the subject of a major exhibition at the MIT Museum. Her current research focuses on creating expressive visualizations of social interactions and on building experimental environments that mix real and virtual experiences. She has a book in progress about how we signal identity in both mediated and face-to-face interactions. Dr. Donath received her doctoral and master's degrees in Media Arts and Sciences from MIT, her bachelor's degree in History from Yale University, and has worked professionally as a designer and builder of educational software and experimental media.

Caroline Jones studies modern and contemporary art, with a particular focus on its technological modes of production, distribution, and reception. Trained in visual studies and art history at Harvard, she did graduate work at the Institute of Fine Arts in New York before completing her PhD at Stanford University in 1992. Previous to completing her art history degree, she worked in museum administration and exhibition curation, holding positions at The Museum of Modern Art in New York (1977-83) and the Harvard University Art Museums (1983-85), and completed two documentary films. In addition to these institutions, her exhibitions and/or films have been shown at the San Francisco Museum of Modern Art, the Hirshhorn Museum and Sculpture Garden in Washington DC, the Hara Museum Tokyo, and the Boston University Art Gallery, among other venues. She is the recipient of fellowships from the National Endowment for the Humanities and the John Simon Guggenheim Foundation (among others), and has been honored by fellowships at the Wissenschaftskolleg zu Berlin and the Max Planck Institüt (2001-02), the Institute for Advanced Studies in Princeton (1994-95), and the Stanford Humanities Center (1986-87). Her books include Machine in the Studio: Constructing the Postwar American Artist, (1996/98, winner of the Charles Eldredge Prize from the Smithsonian Institution); Bay Area Figurative Art, 1950-1965, (1990, awarded the silver medal from San Francisco's Commonwealth Club); and Modern Art at Harvard (1985). She co-edited Picturing Science, Producing Art (1998), and has published on subjects ranging from Francis Picabia to John Cage to new media art in journals such as Critical Inquiry, Res, Science in Context, caareviews online, and Cahiers du Muse national d'art moderne.

Martin Wattenberg is a computer scientist and artist. He is the founding manager of IBM's Visual Communication Lab, which researches new forms of visualization and how they can enable better collaboration. The lab's latest project is Many Eyes (http://www.many-eyes.com), an experiment in open, public data visualization and analysis. Prior to joining IBM, Wattenberg was the Director of Research and Development at SmartMoney.com, a joint venture of Dow Jones and Hearst. His work at SmartMoney included the ground-breaking Map of the Market. Wattenberg is known for his visualization-based artwork, which has been exhibited in venues such as the London Institute of Contemporary Arts, the Whitney Museum of American Art, and the New York Museum of Modern Art. Wattenberg holds a Ph.D. in mathematics from U.C. Berkeley. He lives in Winchester, Massachusetts.

Bibliography

[1] Audacity. `http://audacity.sourceforge.net`.

[2] Anthony Aveni. *Empire of Time: Calendars, Clocks, Cultures*, chapter 1. University Press of Colorado, 2002.

[3] Avid media composer. `http://www.avid.com`.

[4] James Bo Begole, John C. Tang, and Rosco Hill. Rhythm modeling, visualizations and applications. *Proceedings of the 16th annual ACM symposium on User interface software and technology*, pages 11–20, 2003.

[5] Ana Bierzanska. A day in brands. `http://www.abinarylife.com/2008/05/a-day-in-brands/`.

[6] Steven Bird, Edward Loper, and Ewan Klein at all. Nltk, natural language toolkit.

[7] Richard Brilliant. *Portraiture*, chapter 3. Reaction Books, 1981.

[8] Sheldon Brown. Scalable city. `http://scalablecity.net`.

[9] Paul Buchheit, Jim Norris, Sanjeev Singh, and Bret Taylor. Friendfeed. `http://www.friendfeed.com`.

[10] W. B. Cavnar and J. M. Trenkle. N-gram-based text categorization. *Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval*, 11:161–175, April 1994.

[11] Mark de Berg, Otfried Cheong, Marc van Kreveld, and Mark Overmars. *Computational Geometry: Algorithms and Applications*, chapter 7. Springer-Verlag, 2000.

[12] Judith Donath. A semantic approach to visualizing online conversations. *Commun. ACM*, 45(4):45–49, 2002.

[13] Judith Donath, Alex Dragulescu, Aaron Zinman, and Yannick Assogba. Metropathologies. `http://web.mit.edu/museum/exhibitions/connections/`.

[14] Alex Dragulescu. Blogbot. `http://www.sq.ro/blogbot.php`.

[15] Alex Dragulescu. Respam. `http://www.sq.ro/respam.php`.

[16] Alex Dragulescu. Spam architecture. `http://www.sq.ro/spamarchitecture.php`.

[17] Thomas Erickson. Designing visualizations of social activity: six claims. *CHI '03 extended abstracts on Human factors in computing systems*, pages 846–847, 2003.

[18] Etoy. Mission eternity. `http://missioneternity.org/cult-of-the-dead/`.

[19] Nicholas Felton. Annual reports. `http://www.feltron.com`.

[20] Saul Greenberg and Bill Buxton. Usability evaluation considered harmful (some of the time). In *CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 111–120, New York, NY, USA, 2008. ACM.

[21] James M. Hudson, Jim Christensen, Wendy A. Kellogg, and Thomas Erickson. I'd be overwhelmed, but it's just one more thing to do: availability and interruption in research management. *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 97–104, 2002.

[22] Keita Kitamura. ffffound. `http://ffffound.com`.

[23] George Lakoff and Mark Johnson. *Metaphors we live by*, chapter 1. The University of Chicago Press, 1980.

[24] George Lakoff and Mark Johnson. Metaphors we live by. In Jody O'Brien, editor, *The production of reality: essays and readings on social interaction*, part 10, page 114. Pine Forge Press, fourth edition, 2006.

[25] Thomas Mangin. Python texcat. `http://thomas.mangin.me.uk/software/python.html`.

[26] Peter Menzel. What the world eats. `http://www.time.com/time/photogallery/0,29307,1626519_1373764_last,00.html`.

[27] Peter Menzel and Faith D'Aluisio. *Hungry Planet: What the world eats*, chapter 1. Material World, 2007.

[28] James W. Pennebaker, Matthias R. Mehl, and Kate G. Niederhoffer. Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 54:547–577, February 2003.

[29] Adam Perer and Marc A Smith. Contrasting portraits of email practices: Visual approaches to reflection and analysis. *Proceedings of the working conference on Advanced visual interfaces*, pages 389–395, 2006.

[30] Gerard Salton. *Automatic Text Processing - The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, 1989.

[31] Kelly Sherman. Wish lists. http://www.kellysherman.net/work/wishlists2.htm.

[32] Lou Susi. datamummification. http://loususi.livejournal.com/45434.html.

[33] Fernanda B. Viégas, Scott Golder, and Judith Donath. Visualizing email content: portraying relationships from conversational histories. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 979–988, New York, NY, USA, 2006. ACM.

[34] Fernanda B. Vigas and Martin Wattenberg. Tag clouds and the case for vernacular visualization. *Interactions*, 15(4):49–52, August 2008.

[35] Chris Webster. *Animation: the mechanics of motion*. Elsevier, 2005.

[36] Ben Werkschul and The New York Times. State of the union in words. http://www.nytimes.com/ref/washington/20070123_STATEOFUNION.html.

[37] Rudolf Wittkower. *Bernini the Bust of Louis XIV*, page 72. Oxford University Press, 1951.

[38] Joanna Woodall. *Portraiture, Facing the Subject*, chapter 1. Manchester University Press, 1997.

[39] Rebecca Xiong and Judith Donath. Peoplegarden: Creating data portraits for users. In *Proceedings of the 12th annual ACM symposium on User interface software and technology (UIST '99). ACM Press*, pages 37–44. ACM, 1999.

[40] Mark Zuckerberg. On facebook, people own and control their information. http://blog.facebook.com/blog.php?post=54434097130.