Community Data Portraiture: Perceiving Events, People, & Ideas within a Research Community

Doug Fritz
B.F.A Fine Art, Carnegie Mellon University (2007)
B.S. Computer Science, Carnegie Mellon University (2008)

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
in partial fulfillment of the requirements for the degree of
Master of Science in the Media Arts and Sciences
at the Massachusetts Institute of Technology

Author **Doug Fritz**
Program in Media Arts and Sciences
August 6, 2010

Certified by **Pattie Maes**
Associate Professor, Media Arts and Sciences
Thesis Supervisor

Accepted by **Pattie Maes**
Chair, Departmental Committee on Graduate Studies
Program in Media Arts and Sciences

Community Data Portraiture: Perceiving Events, People, & Ideas within a Research Community

Doug Fritz

Abstract

As a research community grows, it is becoming increasingly difficult to understand its dynamics, its history, and the varying perspectives with which that history is interpreted and remembered. This thesis focuses on three major components of research communities: events, people, and ideas. Within each of those components exploring how to construct and answer questions to improve connectivity and elucidate relationships for community members. Assuming the artifacts of a community (its publications, projects, etc) model a representation of its nature, we apply a variety of visualization and natural language processing techniques to those artifacts to produce a community data portrait. The goal of said portrait is to provide a compressed representation viable for consumption by a new researcher to learn about the community they are entering, or for a current member to reflect on the community's behavior and help construct future goals. Rather than evaluating a general technique, the tools and methods were developed specifically for the MIT Media Lab community, general principles can then be abstracted from this initial practical application.

Community Data Portraiture:
Perceiving Events, People, & Ideas within a Research Community

Doug Fritz

Thesis Reader **Judith Donath**
Founder, Sociable Media Group

Community Data Portraiture: Perceiving Events, People, & Ideas within a Research Community

Doug Fritz

Thesis Reader **Benjamin Fry**
Principal, Fathom Information Design

Acknowledgements

Thank you Pattie, for guiding my through this process and keeping me anchored when I drift.

Thank you to my readers Judith and Ben, for their help, insight, and encouragement.

Thank you to my groupmates, for reflection, support, and aid.

Thank you to my labmates, for the environment: the events, people, and ideas of the Media Lab that make this place great. To everyone at the Media Lab, thank you, but I want to especially thank Richard, Ben, Kate, Aaron, Sajid, Henry, Necsys, and the others who helped me with gathering the information for this project.

Thank you to my housemates, for the community outside of lab, for laughter, friendship, joy, and foosball.

And finally, thank you to my family, without whom I would have nothing. No matter what I do, where I move, or what events transpire, you are my eternal rock. Thank you for encouraging me to imagine, to tinker, to build, and to think.

Chapter 1: Introduction

The core of every research community is a set of ever changing and interacting ideas; these ideas evolve with time and are influenced by the relationships between the people producing them. The recent digitization of the records for these communities exposes new opportunities to make sense of history in a method unlike any previously available. The hypothesis is that communities now leave a plethora of digital traces through which computers can mine, discover, and represent the seminal events, people, and ideas. In a small community with limited lifespans, limited spatial constraints, and limited knowledge, there is a self-regulation that maintains the scale of the community to be within the bounds of what can be easily perceived. As a research community grows unfettered by time as a result of large, stable institutions; space as a result of remote collaboration; and memory as a result of Petabytes of easily available storage, it becomes increasingly difficult to perceive the dynamics of the community. A community data portrait helps us view, reflect on, and evaluate our collective actions. Much like viewing a family photo album, but one computationally derived for an entire lineage, memories and patterns are extracted from the events, people, and ideas that form a collective history. The tools and techniques outlined in this thesis were constructed to form a community data portrait of the MIT Media Lab, but their subsequent application is extendable to other communities.

Thesis Roadmap

The thesis begins by defining the higher level problem in more detail and the motivations behind why reflecting on and properly perceiving one's own community is a problem that needs addressing. Additionally, it will make an argument that there are currently no viable tools for this type of community reflection in place. Having defined the motivations, benefits, and need; it will situate itself among the various other related work from many intersecting domains, including but not exclusive to scientometrics, history, journalism, art, and business intelligence.

Next, there is an explanation of the components of the system and the reasoning behind the decisions for their construction. Once the components have been defined and explained, the core of the thesis is constructed as a series of 'lenses' which are each composed of a question about the community and the visualization of

that question through the data. The 'lenses' are broken down into the three main themes: events, people, and ideas. Within each question and representation, or 'lens', there is discussion of what worked, what did not, and why; as well as responses from interviews discussing the representation with a cross-section of the research community from the various roles of membership, ranging from new comers to alumni. All of this research, discussion, and evaluation was done specifically with the MIT Media Lab community. This decision of focus on a single community allows for the ability to act locally with real people producing immediate benefit. From the experience gathered embedded in and working with a live community, it is possible to extract principles and techniques useful to a wider audience.

Finally, it will conclude by discussing future directions and the implications of mixing and combining various 'lenses' and the methodologies that need to be in place to best perceive these research communities who now exist and grow at scales beyond our natural conception. As well as, future work making these methods available online for testing in other communities and research institutions for comparison.

Context and Definitions

Human Scale Perception - The scale at which our perceptions are evolved and optimized to perceive the universe specifically though our sense of time, relationships, and ideas. Evolutionary anthropologists postulate that our brains evolved to deal with essentially ecological problem-solving tasks.[1] Our development within the context of a world manipulatable at only a certain scale has subsequent repercussions on our perception of reality. Our world exists between the microscopic and the macroscopic universe and our inability to perceive the very large and very small is a byproduct of our development within it. This concept is extended to events, people, and ideas; where large communities, long histories, and/or too many ideas are outside the scope of what we can intuitively comprehend. This idea is additionally modified to include the nonlinear scales that are a distortion product of human perception such as the human mind's nonlinear memory of past and future events. Finally, the idea is in many ways intertwined with the common usage of the word 'intuitive', positing that interactions experienced at the human perception scale can be referred to as intuitive.

Data Portrait - "Data portraits depict their subjects' accumulated data rather than their faces. They can be visualizations of discussion contributions, browsing

histories, social networks, travel patterns, etc. They are subjective renderings that mediate between the artist's vision, the subject's self-presentation and the audience's interest" [2].

Data Lens - Is a method of transforming information to rescale, skew, or restructure its perceived nature. This could be zooming in, highlighting a specific feature, distorting, etc. Generally data lenses are visualizations and as such follow the visualization fundamentals outlined later and available in appendix A. They differ from a general visualization in their focus on the transformation of data into the 'human scale' (defined above) and by nature are tools of inquiry to expose a new way of seeing something rather than a distortion to tell a story or make a judgement. Though a story or judgement could be extrapolated later their purpose is simply to expose the nature of some dataset from a different perspective. Throughout this thesis they will be referred to simply as lens or lenses.

Community Data Portrait - Is an extension of the idea of data portraiture from Donath et al. (see above) combined with the idea of using multiple data lenses to view an entire community. A community data portrait is a series of compressions and/or distortions of a difficult to perceive entity like a community into a human scale representation similar to a portrait. As with a data portrait the artifacts of the entity become features of the representation. In this case the entity is a research community and its artifacts are the people, their documents, their relationships, and recorded actions as a collective whole.

Chapter 2: Motivations

Problem Addressed

The problem addressed in this thesis is the perception of communities as they grow beyond something of which we can maintain a representation. Specifically, the problem can be broken down into several aspects:

1. **Lilliputianism** - It is difficult to perceive what is happening at the big picture level, when our scale of perception is constrained. While hindsight is twenty-twenty, understanding where events are heading can help preemptively steer the course of action. Often, failure to see things outside of our scale is a result of missing obvious connections between the small pieces which fit into a larger whole, colloquially this is referred to as "Can't see the forest for the trees": meaning one can get so wrapped up in the details they fail to see what is happening at a larger scale.

2. **Turnover** - Knowledge is often lost as researchers come and go. While perception is not really going to solve an issue of missing data, it can elucidate the artifacts of such a researcher and expose the holes left by their absence by comparing data over time. By correctly perceiving the loss of knowledge one may take action to prevent it.

3. **Reinvention** - Previously completed projects are often "reinvented" because of a lack of awareness. If a community is too large or too old, searching for what has already been done is difficult. Web search has helped alleviate this problem, but for many communities their internal archives are not easily queryable. Additionally, search fails with many multimedia sources and conceptually related topics but textually non-overlapping data.

4. **Overload** - The number of artifacts a community produces increases in proportion with the growth of the community. However, the bandwidth of the individual remains constant. While search is obviously good at finding something in a large body of information, there is a need for extrapolating the gist of a community to understand what things might be searchable.

5. **Invisible Structure** - The explicit structure hierarchy of a community -- who is in charge, who works with who, etc -- is often not the only organizational structure. Often times there is collaboration across groups or non-equal power behind a supposed flat community. In short, there is often an unseen more fluid

structure of how people work together or make decisions. By inferring relationships between the digital traces of individuals rather than explicit titles or roles, it becomes possible to construct a representation of this unseen organization and glimpse different non-reported ways a community is organized.

6. **Ruts** - Humans are creatures of habit, without reflective exposure of our histories it is difficult to be aware of and change our actions.

7. **Drift** - It is possible for patterns of events to mildly shift at a time scale difficult to notice. This is another instance of imperceptibility based on scale of change. Just as a frog will sit calmly and boil in water with a slowly rising temperature, humans will just not notice behavior that changes gradually. An organization may, through a series of small, seemingly unrelated, changes be shifting an underlying policy. While each small change may seem innocuous, the cumulative change between the original intent and current practice would be alarming.

Human Scale Representations

> ... our brains have evolved to help us survive within the orders of magnitude of size and speed which our bodies operate at. We never evolved to navigate in the world of atoms. If we had, our brains probably would perceive rocks as full of empty space. Rocks feel hard and impenetrable to our hands precisely because objects like rocks and hands cannot penetrate each other. It's therefore useful for our brains to construct notions like "solidity" and "impenetrability," because such notions help us to navigate our bodies through the middle-sized world in which we have to navigate [3].

Human perception of the physical world is constrained by our evolution in the middle-sized world of our perceived universe. We extend this idea beyond the perception of the physical world into the perception of the social world. Though this is referencing a fairly recent quote to help define the idea, the concept of human perception relative to our normal human scale is expressed in everything from Swift's Lilliputians in *Gulliver's Travels* from 1726 to the ancient Greeks with Plato's *Allegory of the Cave* (if one agues scale to be a term relative to the constraints of one's perspective). The idea's persistence in literature over the ages is testament to it being a fundamental human concept. Within this overarching idea of misperception based on the scale at which we develop our abilities to perceive, this thesis focuses

specifically on one's perception of their community, and within that specifically research communities and their self-perception.

The goal then of augmenting the weaknesses of human perception to perceive events and patterns outside of 'human scale' is accomplished by creating various views of a community. This repackaging is achieved by viewing the data of the community -- and by proxy, the community itself -- through a series of 'lenses'. The term 'lens' is used loosely to encapsulate representations of the data artifacts a community produces which distort and/or compress said artifacts into something more easily interpretable. This thesis exclusively uses visual representations but the theory of remapping our perception of a community is not constrained to the visual domain.

The function of the community data portrait produced by these 'lenses' is to provide a compressed representation viable for consumption by a new researcher to learn about the community they are entering and situate themselves, or for a current member to reflect on the community and help construct future goals.

Reflective Views

If a feature of community lenses is their reflective nature, there first needs to be justification of the benefits of reflecting on one's community.

> Adhering to the assumption that social science or art is phronesis, whereas natural science is episteme, in the classical Greek meaning of the terms; phronesis is well suited for the reflexive analysis and discussion of values and interests, which any society needs to thrive, whereas episteme is good for the development of predictive theory, and; a well-functioning society needs both phronesis and episteme in balance, and one cannot substitute for the other [4].

Following this, the four actionable goals of social science toward phronesis were concluded as (1) Where are we going? (2) Who gains and who loses, by which mechanisms of power? (3) Is this development desirable? (4) What should we do about it? [4]. This thesis focuses almost entirely on the first question, where are we going, with an additional precondition of trying to explain who *we* are. Some attention is paid to mechanisms of power, but its subsequent evaluation or courses of action are left for further research.

To make a prediction of where we are going, one must first have a strong and valid model of the entity and its history as well as surrounding data. Thus, the course

of action should be to expose disjunction in the internal model of what one believes to be true about a community with what is actually taking place. This exposure of disjunction is inherent in viewing any representation of something already known. In this case, it is accomplished by simply displaying a representation of the community through a given lens and the viewer interpreting the representation in comparison to what they already believe they know of the community and what they think they should be seeing.

Reinforcing Community

If this thesis is for community reflection, then there is an unstated assumption that the community and a sense of community are important. To justify this, what are the factors defining a sense of community? McMillian & Chavis [5] lay out the main four elements of a sense of community as follows:

1. Membership

   Includes Membership includes five attributes:
   - boundaries
   - emotional safety
   - a sense of belonging and identification
   - personal investment
   - a common symbol system

2. Influence

   Influence works both ways: members need to feel that they have some influence in the group, and some influence by the group on its members is needed for group cohesion.

3. Integration and fulfillment of needs

   Members feel rewarded in some way for their participation in the community.

4. Shared emotional connection

   The "definitive element for true community" it includes shared history and shared participation (or at least identification with the history).

Contrasting the elements of a sense of community with the proposed plans for reflective lenses, the main element that the reflective lenses would be aiding is a shared emotional connection. By displaying items from history or creating a shared experience we tighten the identification with the emotional connection to the collective history.

Contributions

There are four main contributions of this thesis. First is the extension of data portraiture and the introduction of the idea of a community data portrait as a product of a communities collective digital artifacts. A second contribution is determining what questions are applicable to represent a community data portrait for communities of the size and scale of the MIT Media Lab. A third contribution is the underlying methods for extracting what is important from those digital artifacts, including the idea of focusing on thin slices of actual data, such as questions, or new algorithms for the discovery and classification of trendsetting ideas in a research community. Finally, the fourth contribution is evaluation of visualization techniques applicable for display and interaction with the processed information.

An interesting attribute about the Media Lab as the testbed for this research, is the scale of the community and number of artifacts. The Media Lab is reaching it's 25th anniversary this year, which is both a good time for reflection and indicates the scale of data produced. What is interesting about that scale is the betweenness of its features, the community is small enough to know many of the individuals, while too large to know all the research. It is too small for inter-citation network analysis, and too large to look at all the citations individually. New approaches and mechanisms for analysis have to be considered.

Finally, the sort of reflective community view proposed here has no strong parallel. If one is to look at comparable other services, there are corporate brochures handed to new or potential members which are all very biased and high level. On the other hand, something like an audit of an institution is very detail oriented, but misses the people and relationships. It is difficult to parse, read, or get an immediate sense of what is happening at a high level from glancing at an audit. What this thesis hopes to achieve is to provide a small slice into a community that can help summarize a more fuzzy interpretation of the whole community.

Limitations

It is important to note that there are many inherent limitations with a system that focuses on the artifacts of a research community to describe it. While it is assumed that the artifacts produced provide a stand in for the nature of the community, they are more accurately shadows cast on a wall. While we may make accurate predictions at times of the original object casting the shadow, it can never be completely accurate. For example there are many researchers that due to the nature of

their work do not publish academically, but are equally influential within their own domain. Though their shadow is great, it is cast on a different wall.

The 'lenses' talked about above and throughout this paper act as distorters to focus and filter the light shining on the object and producing the shadows. While the work focuses on how to best shine the light and interpret the shadows. It is important to note that the intended audience is other researchers, because of this, the object casting the shadow -- the community -- is something they have some memory and model of. That memory acts as an internal model to compare to the shadow. Having a memory of the original object, no matter how distorted, makes interpreting the shape of the object the shadow implies much easier. Lastly, every projection is limited in its information compared to the higher dimensional original object. By projecting the shadow into the 'human scale' data is lost, and the only hope of extruding accurate information of the original object is to cast shadows from many different directions.

Finally, it should also be noted that there is a definite bias inherent in all of these representations, and while the data portrait of a community presented here is intended to reflect the community as honestly as possible, it is subject to the choices of those designing the algorithms and presenting the information. As is the case for any portrait, the result is a negotiation between three elements: the subject's self-representation (the artifacts the community chooses to produce), the artist's bias (the programmer's goals and choices), and the viewer's interpretation (how other members of the community interpret the results).

Chapter 3: Underlying Principles

Representing a Relative View of History

This thesis presents a relative view of a community's history. The term 'relative' in this case references a postmodern notion of truth. Due to the disagreeing definitions of postmodernism across various domains and its transition into a buzz word rather than a well defined concept, there is a necessity to define it first for the scope of this thesis before its use.

> "Post modern truth is not a single thing 'out there' to be discovered. Instead truth must be assembled or constructed. Sometimes, it's constructed visibly, from many different components (i.e. scientists gathering results of multiple studies). Other times it happens invisibly by society, or by cultural mechanisms and other processes that can't be easily seen by the individual." [6]

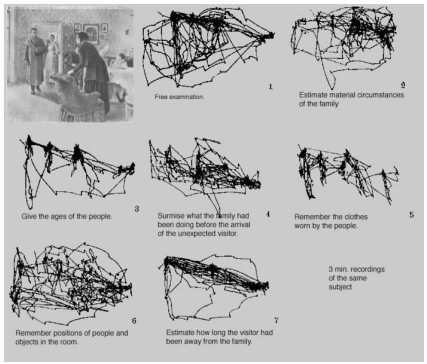
If every truth is a construction, then the tool of postmodernist understanding is deconstruction (in reference to Heidegger and Derrida). Through deconstruction one breaks down notions of truth into subcomponents for analysis. This is in part the inspiration behind the visualization framework later in this thesis where data is deconstructed and reconstructed in many different ways to derive a sense of relative truth. Finally, if relative truth is a deconstructable notion based on experiences, one must then reject the notion of an accurate summarization, or 'global narrative', instead focusing on the 'mini-narrative'. A mini-narrative being where multiple representations of small examples of local events are interpreted as more truthful to the global pattern than a summarization in a single view. Because of this assumption, the representations used to express a community, should expose small vignettes of similar experiences, rather than trying to say in a single statement what has happened in the data. An example of this idea in practice is the notion of presenting all the questions that have been asked in each Media Lab thesis. Questions mark a state of mind and are the fundamental construction of the scientific method, by presenting theses mini-narratives in multiple, these multiple perspectives form a more 'truthful' representation of the state of mind of the community.

Visual Gist

Part of the goal for a community data portrait should be to represent a visual gist of a community. In general, a gist can be thought of as the rapid recognition of patterns of patterns [7]. Much like the merging of thousands of images of the same object, spatial and temporal patterns begin to emerge from the images. In the design and art domain the same simple technique has been used to represent a gist of a scene. For example, Jason Salavon's work, seen here, is the merging of all playboy centerfolds for each of the last four decades. Our brain is able to both pick up on the major visual features, and compare those features across the instances. By grouping the data across each decade one is able to very easily compare trends in time and the major feature shifts in the data. Finally, by making use of the data as the representation or data as interface, there is no confusion over the mapping of what we are seeing or how to interpret the information, the picture is created from the pictures which are themselves still partially visible. This process becomes more difficult when one is trying to translate non-visual information into the visual domain, because there can be no reliance on the salient recognition of underlying features.

Preconditioning

The structure of each lens used to construct community views includes a question and a representation. The question is a method of parsing the information and the representation is a visual interpretation of said parsing. However, when presenting this view to the user there is a preconditioning aspect to displaying the question with representation that needs to be taken into account. Taking evidence from eye tracking work by Yarbus in 1967 and displayed here [8]. Different patterns of eye movement and scene deconstruction take place when given different tasks. From this we can conclude that the cognitive state preconditioned on the question physically affects how we see an object. Thus, although the question was meant to be represented with the visual representation,



the visual representation is in some ways constructed from the question. The implications are to be conscious of the effect of the text on preconditioning the interpretation. As such, it would be useful to try various related questions to see how this affects the efficacy of a visualization to produce the desired interpretation.

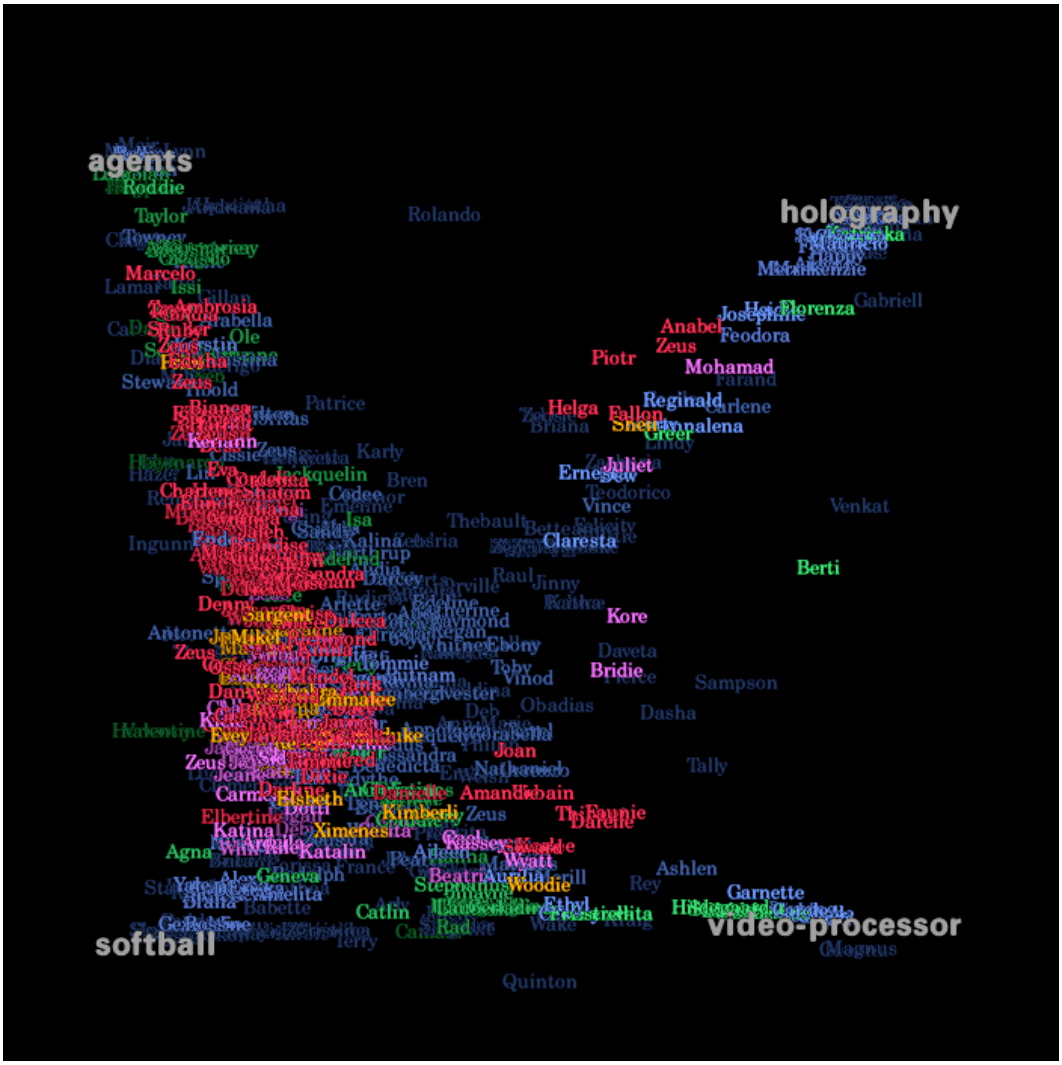
A Self-Extending Thesis

An indirect benefit of writing a thesis on theses, was the production of a command-line interface to search similar or related topics based on text input into the system. Thus, by inputing the text of this thesis during its construction, many of the references made below in the background chapter were discovered. While this is only able to search other Media Lab projects and theses it was useful to insert a given paragraph or the entire document and pivot to discover additional relevant sources. One could then add the resulting information and recurse interactively. This idea of related discovery begins to address an indirect contribution this thesis and the system it creates. It becomes possible to create just-in-time research tools, where a high-level compressed representation of the history of a community can be dynamically constructed based on the profile and interests of the viewer. If one is writing a new paper on a specific topic, the view they see of the lab's history can be constructed in a such a way as to make relevant sources and methods most prominent, decreasing reinvention, discovering key players to contact, and allowing for the most fluid extension of past ideas.

Chapter 4: Related Work

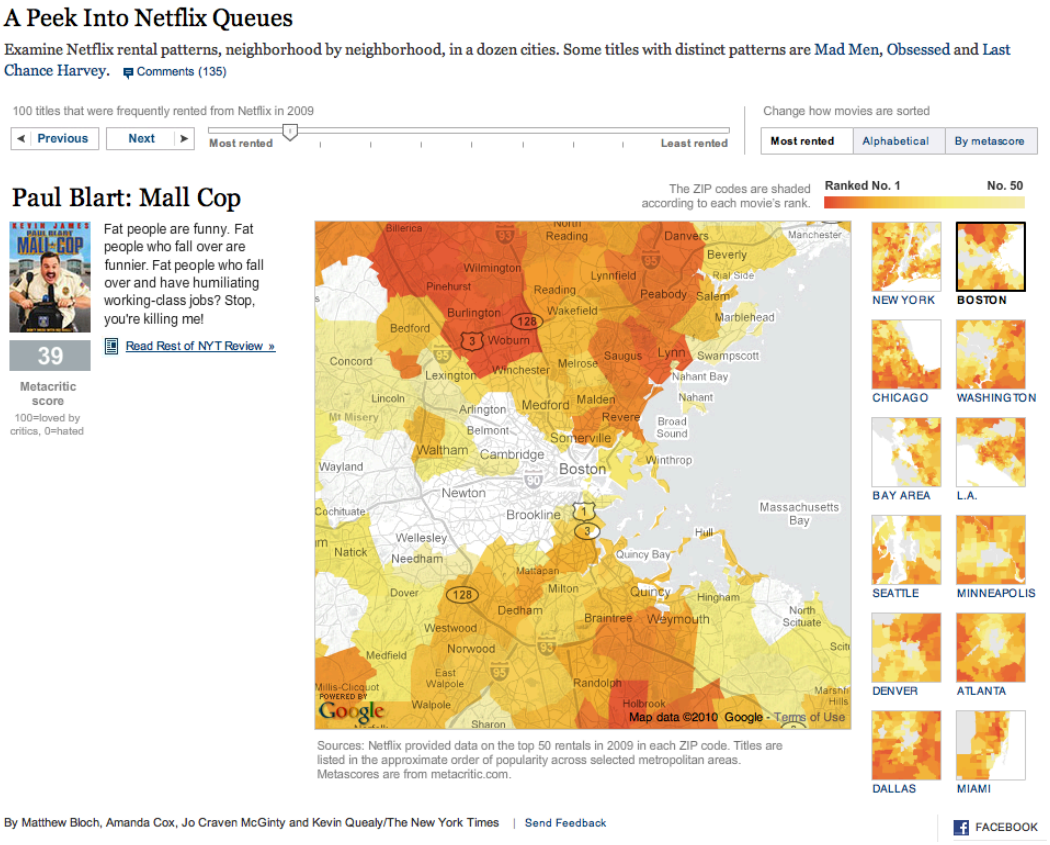
Community Visualizations

There are several other projects that have dealt with methods of visualizing



community. One early example is *Visual Who* [9] which was a project to visualize an electronic community by means of their subscription to various mailing-lists. Participants could interactively change the anchor points for the spring-based visualization and as a result explore the underlying structure of the interests of community members. This project is an early example of using the behavior of individuals to describe the nature of the community.

Additional examples from which this thesis developed are visualization



projects like the NY Times' *A Peek into the Netflix Queues* [10]. This visualization was an exploration of several major cities in the United States, with heat map visualizations of the renting frequency of top movies by zip code. There are three interesting takeway lessons from this project. First, the ego-centric nature of the piece -- the ability to look up one's own zipcode is the first thing everyone does with the this project, and that action is immediately engaging. Second, the project is an example of using the artifacts in our lives to describe community, in this case the movies rented, a premise on which much of this thesis is based. Finally, this project exposes the ability to compare and contrast models of what one thinks about a community with what one thinks about a complex cultural proxy object (like a movie). People may have a biased caricature of the characteristics of a person who rents a certain movie, as well as some rough caricature of people that live in certain areas they have experienced. By comparing how accurate the visualization maps to our internal models it is possible to judge the accuracy of the portrayal. Once a baseline for the portrayal's accuracy has been determined, one can understand the

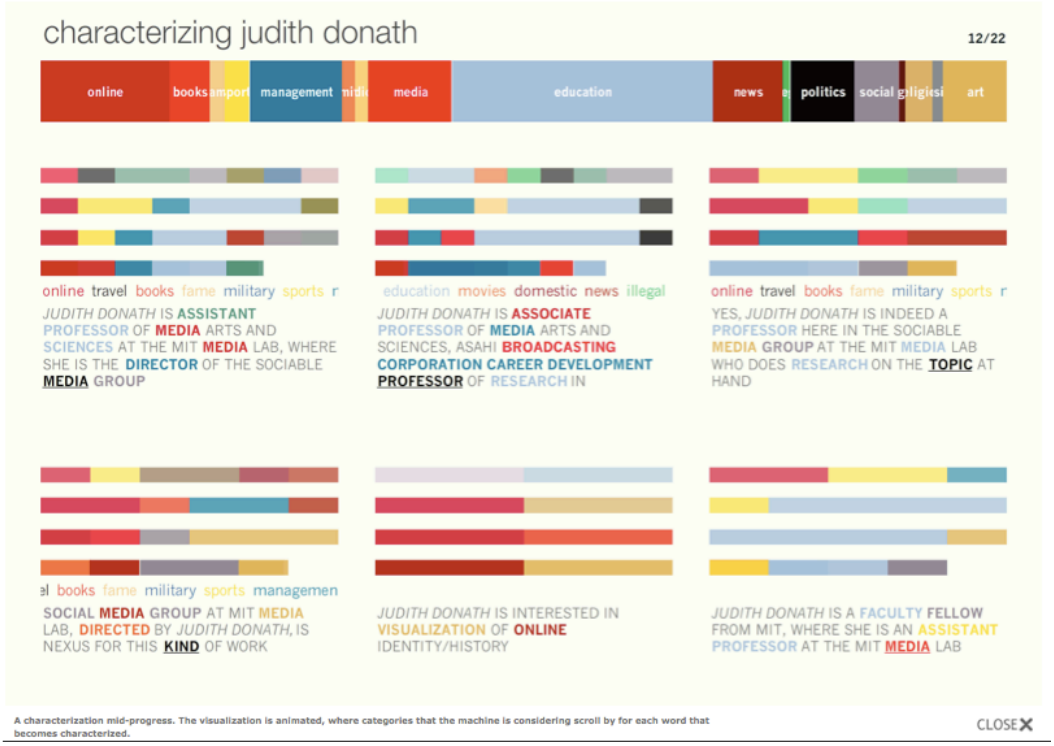
mapping of visual representation to our internal models. This mapping forms a visual language to characterize and express new models in comparison to our internal ones.

Data Portraiture

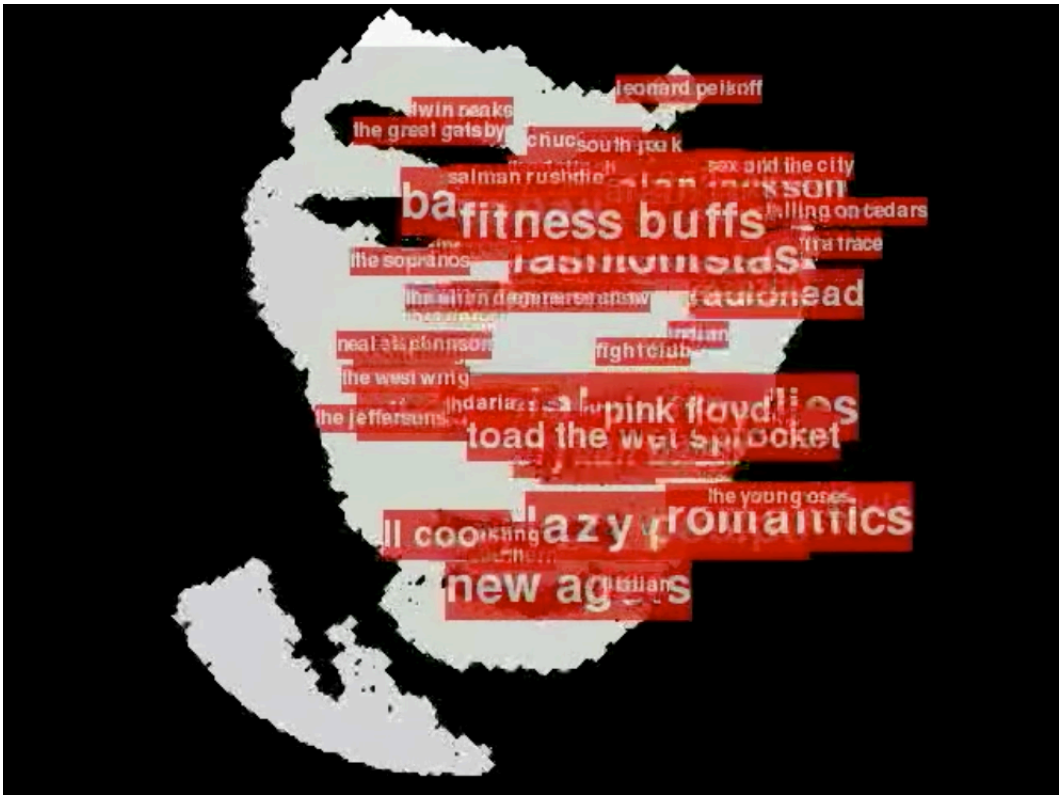
Data portraiture defined above and laid out in detail by Donath et al [2] is the subjective rendering of people by way of their data. In the majority of data portrait representations a key component is the definition of self relative to the

community. In *Lexigraphs* [11] individual data portraits are constructed out of the tf-idf (term frequency - inverse document frequency) [12] weighted words of individual twitter users and formed into portraits to represent their bodies in the bodiless existence of the internet. The relationship between the individual and the community is made visible by displaying many individual portraits next to each other. Side by side presentation of the multiple individuals helps exemplify what is important in the larger community and what makes a given portrait unique.

Another approach to relating individual to community in a data portrait is to construct the portrait by fitting an individual into a globally constructed model, such



as the case for projects, *Personas* [13] or *IdentityMirror* [14]. Personas is a project to visualize the collective representation of an individual as described through the insights and more often mis-insights of the computer. The authoritative presentation of *Personas* is contrasted with the underlying fallibility of gathering information by name alone on the Internet. This contrast hints at our over-reliance and dangers of trusting the representation. Regardless of the underlying subtext, the methodology employed is to map a given individual into a model of what computer searches return

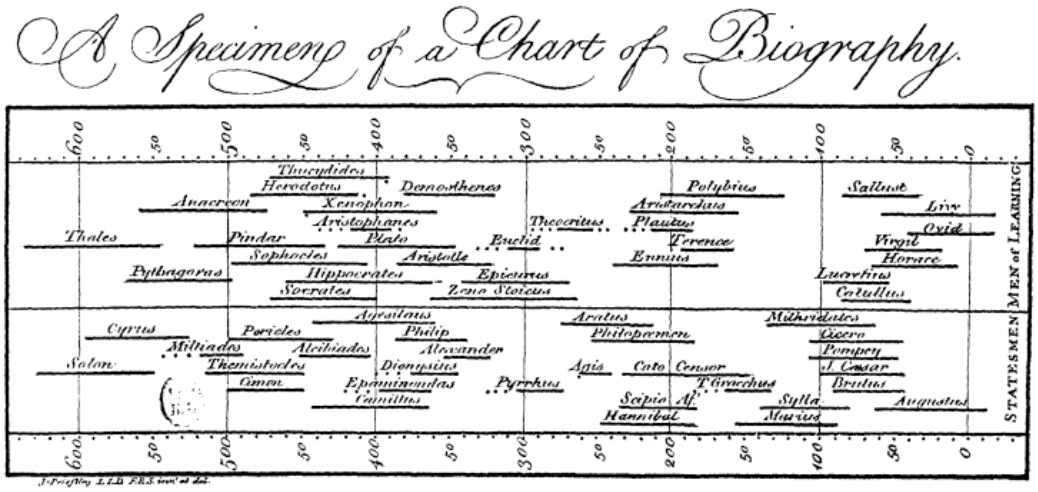


for a given name. Similarly, *IdentityMirror* is about mapping a person's online identity into a global taste fabric, where a representation of a group of users constructs a global fabric within which a given individual's representation is placed.

Finally, it should also be noted that even the use of tf-idf for the relative importance of a words in *Lexigraphs* marks a simultaneous representation of the community and the individual. Since each word is displayed as important against the background noise of the community as a whole a view of the community is produced by looking at and comparing the individual data portraits. In short, the definition of self is relative to our surrounding community, and the reverse holds, namely that a measure of community can be represented by multiple representations and comparisons of individuals.

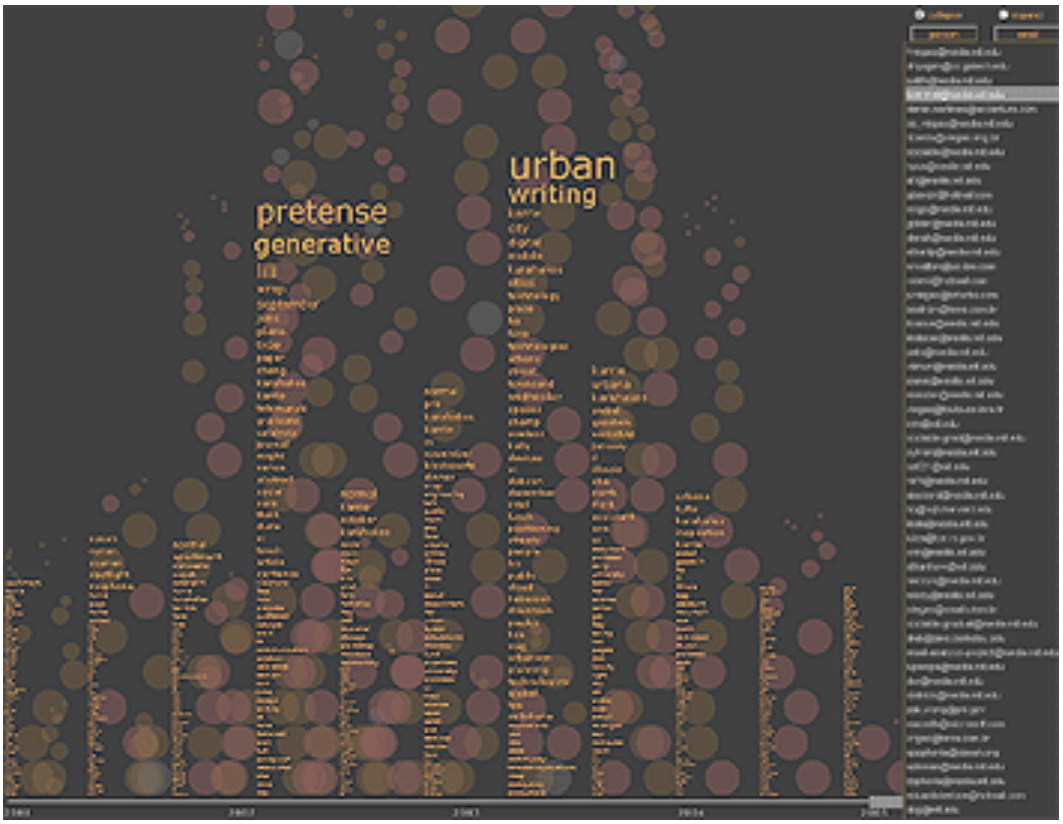
Mapping History

In any representation of events it is important to look at the long tradition of timelines and mappings of history. Within this topic subtopic there have been several example representations. To begin, one of the earliest and most well known presentations of a timeline of ideas was eighteenth-century British polymath Joseph



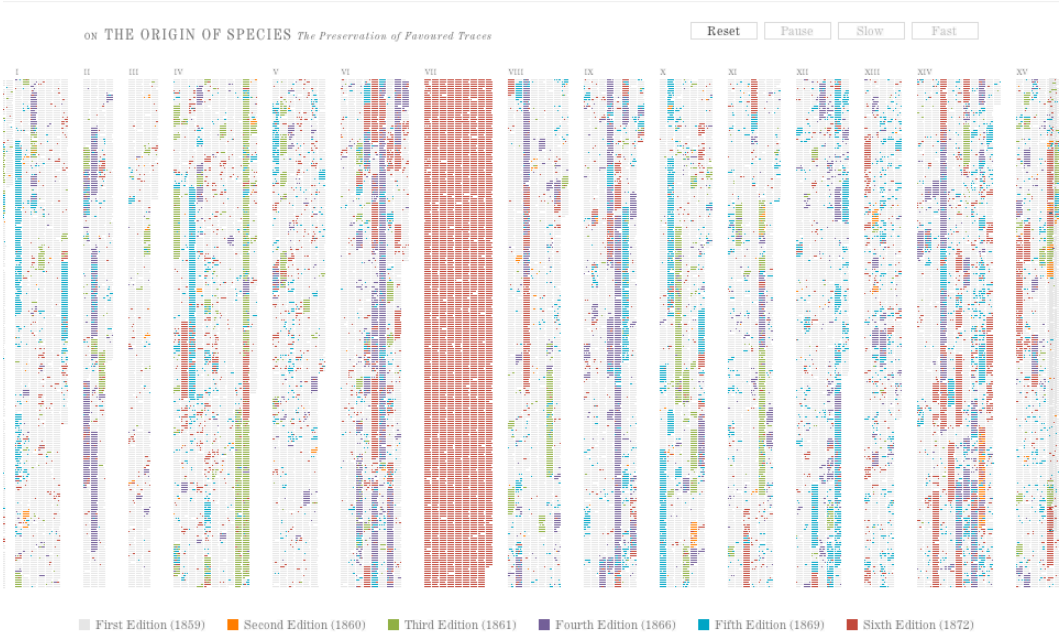
Priestley's A Chart of Biography [15]. Priestley's goal was to map an ordered structure to the history of ideas so as to make understanding, remembering, and learning the material much more salient. Unfortunately, the representation's strict linearity makes it difficult to represent conflict, divergence, and merging of ideas; as well as multiple perspectives. However, this focus on the history of ideas and the representations of content flows into later representations of scientometrics described below. It also inspires the design goals for this thesis, namely to represent trends of ideas and the history of major works or people within a community.

Obviously visualization of history is not constricted to mapping scientific

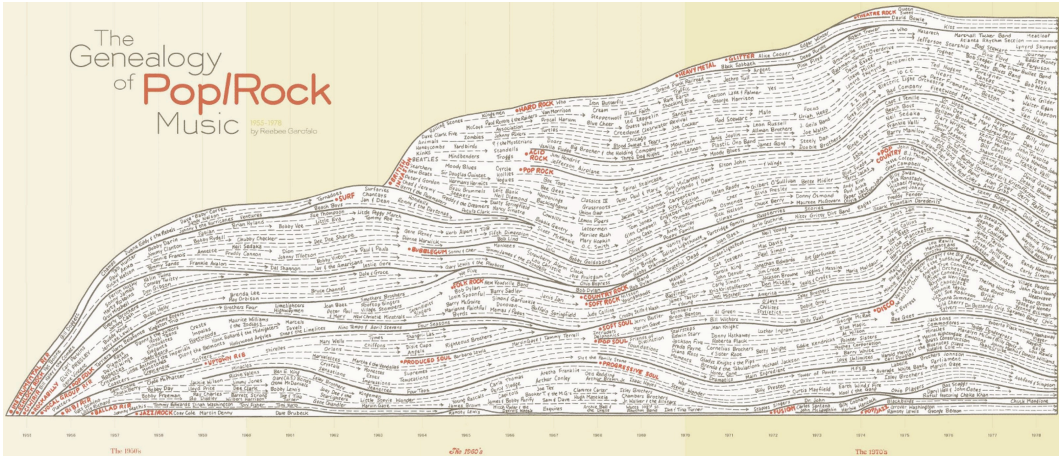


ideas. Some other representative projects include *History Flow* [16] and *TheMail* [17], which deal with the representation of content flows over time. *History Flow* exposes wikipedia edits over time, while *TheMail* allows for the browsing and reflection on one's email archives. From these projects we gather the importance of pivoting on themes, and compressing information to determine different patterns at different time scales.

Since, part of this thesis is trying to display temporal information, it is important to look at methods of displaying time. The above examples all use space to display a timeline. However, time itself can be used to represent a timeline, and the speed up of time can help reveal changes that may otherwise be missed. An example



of such a system is *On the Origin of Species: The Preservation of Favoured Traces* [18], where edits of Darwin's famous book are revealed in full over time. There are two major takeaways from this work. First, the unraveling of history through animation, simply playing back time at a speed fast enough to notice the difference and changes through the book's long history provides a new way of looking at the timeline. Second, the use of representing a one-to-one mapping of the full text is compelling. Every line correlates proportionately to a line of text in the book so there is a one to one mapping of information which affords an honest and direct interpretation of the visual representation.



Finally, *The Genealogy of Pop/Rock Music* [19] is an example of understanding genealogies and relationships as temporal in nature and mappable into some larger

timeline. What this has that this thesis has yet to fully adopt is the concept and display of a pedigree for ideas. In this case, the project maps how one style of music and performer flows into later works over time. By doing so it becomes possible to trace back the roots of each idea.

Scientometrics

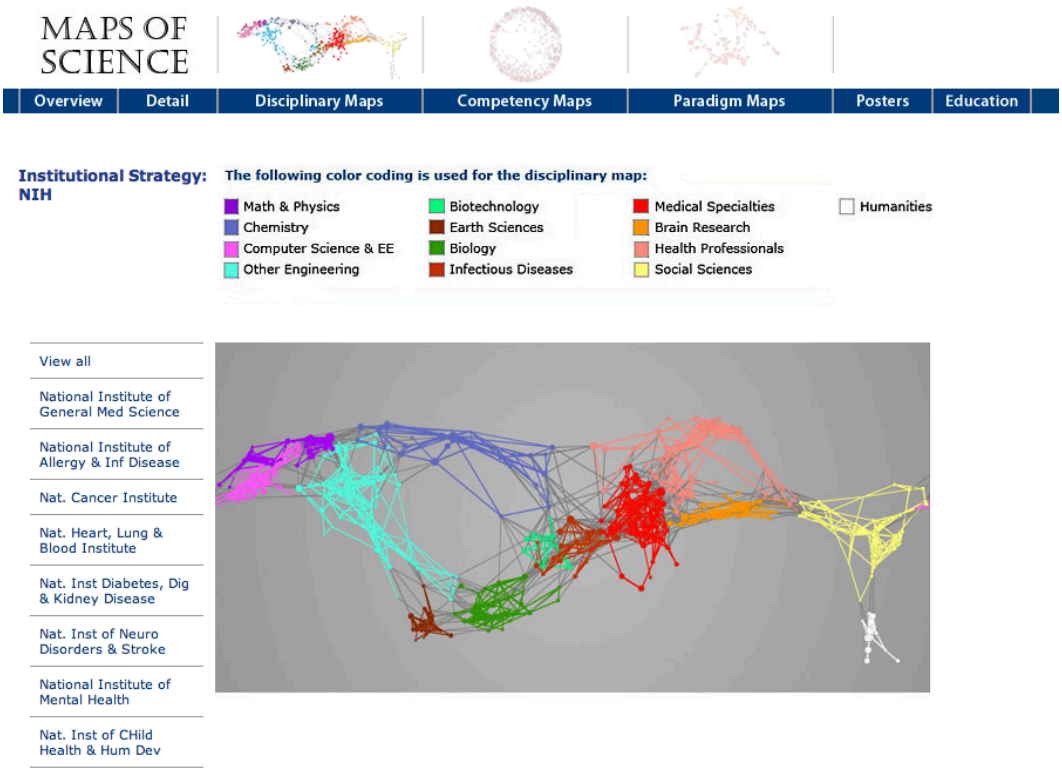
Scientometrics is the science of measuring and analyzing science. This process is practically accomplished by way of citation analysis over collections of publications. For each paper, measuring the network of citations can reveal answers to questions like: "Who are the important leaders?" or "What are the underlying structures of ideas?".

Some early work in the area include the development of the h-index [20] and its derivatives. An h-index is defined as follows --

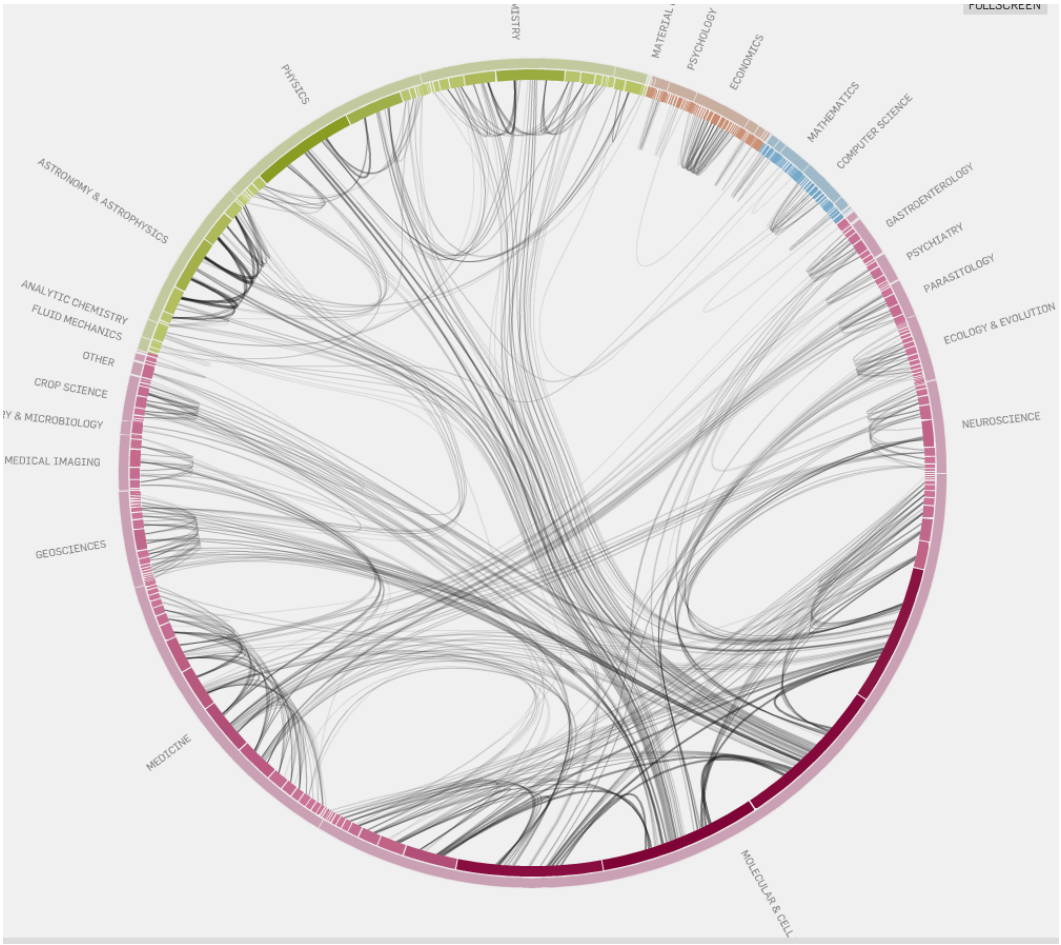
> A scientist has index h if h of [his/her] N_p papers have at least h citations each, and the other (N_p – h) papers have at most h citations each [20].

While this metric is of some value in determining the level of productivity and impact, it is also determined to be extremely misleading. This is especially true in a multidisciplinary environments like the Media Lab, where not only are some individuals publishing more or less often because of the nature of their respective fields, but who they publish with, and the frequency of being cited are dependent on their respective fields as well. As a result something like the h-index within the Media Lab is much less useful than comparing the h-index of each Media Lab professor with their comparable peers in outside departments.

Other scientometic projects, like the *Map of Science* [21], are also using citation network analysis, but are used instead to get a picture of the overall trends



across scientific research. The *Map of Science* project produced three main maps: a disciplinary map, a competency map, and a paradigm map. A discipline is defined as a cluster of journals. They grouped over 16,000 journals into 554 disciplines using similarities in their lists of references and key terms. A competency map is a display of disciplines placed relative to paradigm to measure the overlap. A paradigm is the smallest possible cluster of related scientific documents. For example, 84,000 paradigms are needed to describe the micro-structure of research in 2007 [21]. This map gives an overview of the scientific field as a whole, but it also feels unconnected to the actual researchers, the people that are producing the ideas being discussed. The graphs tell a viewer nothing about who the actual field leading individuals are.



Extending the network analysis angle of scientific visualization, projects like *EigenFactor* [22] try to measure the influence of the journal though which an article is published by means of the number of citations an article receives. In this method, a journal is considered to be influential if it is cited often by other influential journals. The importance of a citation is then measured by the influence of the citing journal divided by the total number of citations appearing in that journal [22]. The approach here is a compelling grouping and analysis of higher level structures (journals) by means of the lower level artifacts (articles). This approach could be applied to research groups, measuring the influence of a research group by the citing of other influential research groups.

Metrics on citation analysis have become the standard for evaluating a scientific community, but these methodologies become difficult when the number of citations is limited or non-existent. For example, not all projects within the Media Lab have associated papers. Additionally, many of the researchers do not publish on the same scale as one another due to the nature of their fields. Finally, although the citation analysis works well when enough articles are in place, the value of such an

approach follows a sigmoid function over the number of nodes in the network for analysis. Without enough nodes and relationships in the network the analysis techniques fails to reach critical mass.

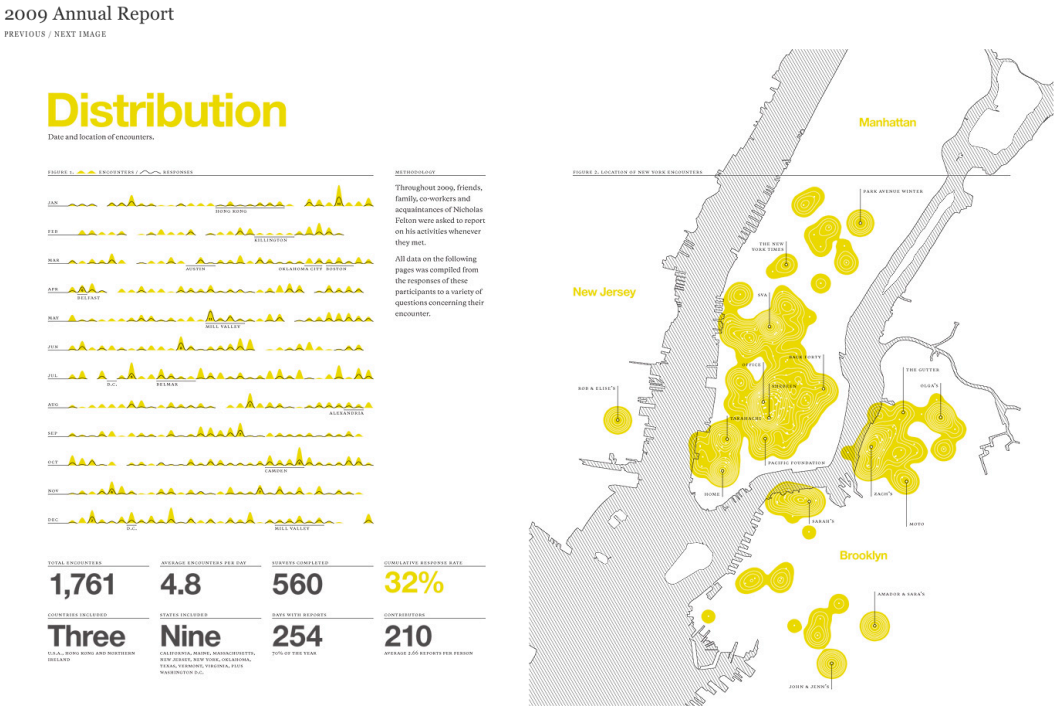
Business Intelligence (BI)

Business intelligence tools are by nature designed to explore data, extract a story, and use said story to construct an argument for why a certain decision should be made. In general, business intelligence is looking at the numerical statistics of a business' output to help optimize the goals of the business. The quality of a BI tool is measured by the ease of use to construct the story of what is happening from the data. This thesis differentiates itself from such business intelligence solutions by having a greater focus on less quantitative and hopefully less directed representations, as well as a focus of the internal people and ideas. Finally, the techniques and approaches here are in no way intended to produce business solutions, in fact more questions are probably produced than answered. However, those questions are deeper questions about the nature of one's community and with that serve the higher goal of community reflection.

Personal Informatics

Personal informatics has been a quickly rising theme over the last 7 years. Commercial sites now range to help one personally track and control topics ranging across fitness (Nike+), electricity (WattVision), diabetes (SugarStats), health (PatientsLikeMe), and mood (MoodMill), just to name a few. The common theme of these sites is the ability to track one's own behavior and by making tracked behavior more visible, affect change. The same principle holds for tracking behavior in a community. By tracking and displaying one's own contribution of ideas and relationships in a research community, one can better reflect on behavior and the community as a whole.

Another example of self representation or self reporting is the *Feltron Reports*



[23]. Here the designer offers yearly releases of all of his personal information collected throughout the year. Through complete exposure and compressed representation of the banal details of our lives we can expose interesting patterns. While the usefulness of such systems is questionable relative to the cost of self recording, *datum.com* is a site extending and automating the process to more people. Other related projects like microcosms [24] are examples of personal informatics were participants are encouraged to expose and visualize the minutia of their day. In each of these projects it is through the collective view of many different representations of banal personal statistics that a global picture of an individual begins to form.

Much like the Feltron reports, but smaller, automated, and real time, new projects like Poyozo [25], itself an extension of a project called Eyebrowse [26], are attempts at a sort of automated diary of all of one's online information. That information is stored and available for reflection and visualization. This begins the difficult process of aggregating the artifacts of an individual across various sources, and attempts to maintain privacy by storing all information local to the machine. If the future includes the continual uptake of such systems wherever personal history is constructed, it becomes easy to then construct a community portrait as a collective representation across these individual histories.

In all of these approaches the personal information of the individuals becomes aggregated and presented back to the user for re-consumption later. This is a process by which we come to understand ourselves by the reflection we form in the use patten's of our surrounding artifacts.

Knowledge Management (KM)

The most important element of research communities are ideas, thus knowledge management, who's primary function is to map and expose the sharing of knowledge and expertise across an institution is of primary concern. There is an obvious need for such systems. As an example, one survey found that 74% of respondents thought their organization's best knowledge was inaccessible and 68% thought mistakes were reproduced several times [27].

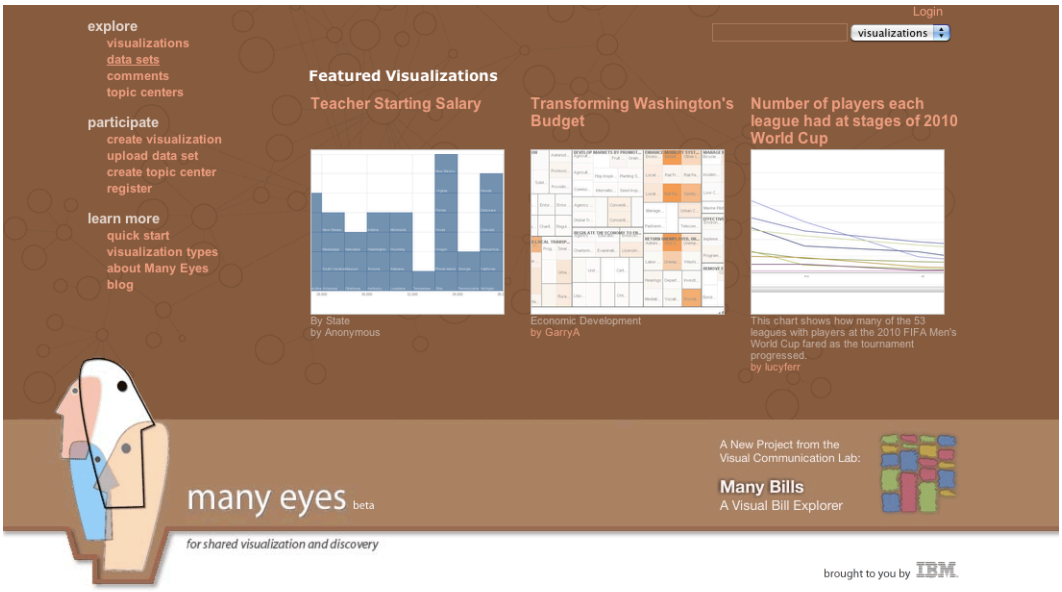
To accomplish the goal of sharing and organizing knowledge, knowledge management systems fall into a several categories. The earliest systems were expert locators constructed from directories of skilled individuals. Later e-learning tools arrived with the goal of helping share the knowledge. These centrally organized structures were later replaced by bookmarks, blogs, and wikis as members began to take a more active roll in producing knowledge that was constantly changing and needed updating beyond the control of a central bottle neck. Finally, the latest trend has been to use semantic web technologies on top of original knowledge managements systems. The trend has been toward inferring structure on top of unstructured information.

It should be noted that advances in knowledge management are often accompanied by the many research studies of how people -- especially in research organizations -- work. Projects like Lucy Suchman's *Making Work Visible* [28] are key insights into the social dynamics of communities and the process of information flow. There is a large body of work related to both knowledge management and this thesis along these lines of work-flow modeling which ultimately must come to understand the dynamics of social relationships in the transference of knowledge.

Knowledge management is about exchanging actual pieces of useful knowledge, tips, best practices, etc. In contrast, this thesis is focused on the overall gist or portrait of the community, not trying to exchange actual information, but generating a sort of community map useful for discovering that specific piece of knowledge. Improved knowledge management means externalization and storage of tacit knowledge in the system. In doing so it would improve a system that is

dependent on representing people as a product of their artifacts. More artifacts means more complete representations. So if community data portraits are useful tools, they would live symbiotically with knowledge management systems. Community data portraits reflect the state the knowledge system, and in doing so expose potential improvements for knowledge management and a high level overview of what is available. In conjunction, better knowledge management means truer and better representations of the community. There is some obvious overlap between the tools developed here and knowledge management systems. In a perfect world, a complete knowledge management system would be able to serve up all the necessary artifacts to describe a community, but as it stands, additional gathering and processing is needed to extract theses artifacts.

Collaborative Visualization



Collaborative visualization systems like *Many-Eyes.com* [29] represent an important ability to let communities discover and clarify the information they find important. This sense of agency in a wiki derived structure of finding, displaying, and remixing information becomes a key aspect of a community that wants to explore its own behavior. The difficulty often with something like *Many-Eyes* is its lack of direction in the data and representations. Because the community has no unified goal, each person judges what is important relative to their own varied internal metric, leading to a lack of consensus or resources applied to shaping the information. Because this added layer of direction is often key to wiki derivative data exploration, such systems

may be better suited for internal use within a specific close-knit community where metrics of evaluation are more unified. BI solutions like Microsoft PowerPivot which share data and visualizations in the cloud are already trying to adopt such solutions.

For community data portraits, wiki-like or collaborative feedback is an obvious extension of the representations produced. The community portraits produced could be collaboratively discussed or annotated by members of the community [30].

Artistic Visualization

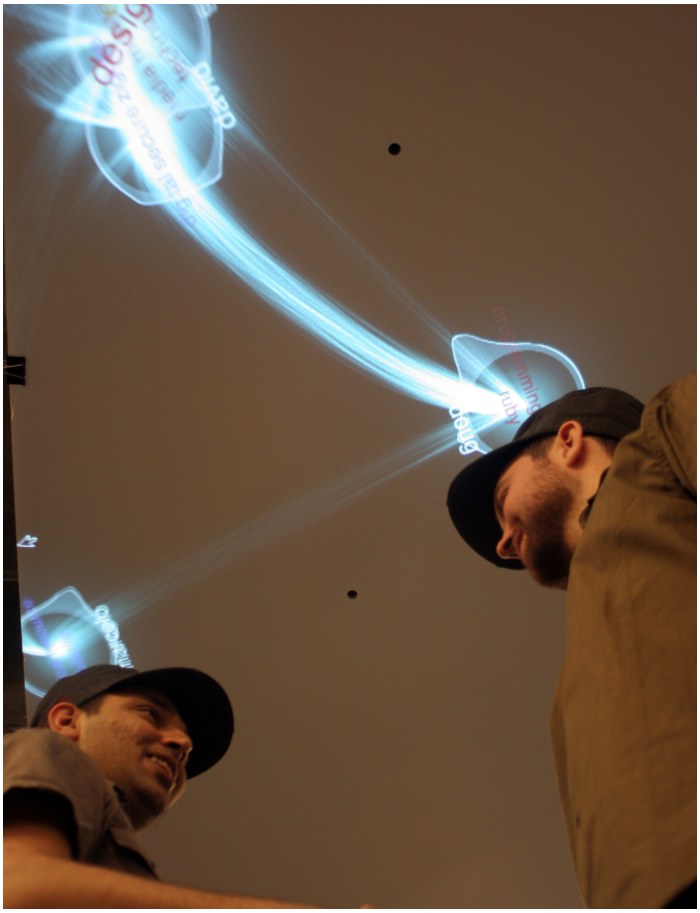
Ultimately this project is about a more emotional and artistic approach rather than a factual data analysis approach. Accordingly, there is much to take from the

artistic practice of others. For example *We Feel Fine* [31], is a project were roughly 15,000 to 20,000 sentences containing the phrase “I feel” or “I am feeling” extracted from blogs are combined with user profiles and used to generate a map of feelings for a community. This project takes inspiration from an even earlier work by Rubin and

Hansen called the *Listening Post* [32] which creates a sort of collective voice for the internet, by statically analyzing the text from chat rooms and reading it back out in various orchestrated modes. Both of these projects used small snippets of text or phrases statistically extracted to produce the community representation, and by doing so both were able to emotionally tap into representations of the people that construct the community.

Prior Work by the Author

ConnectUs

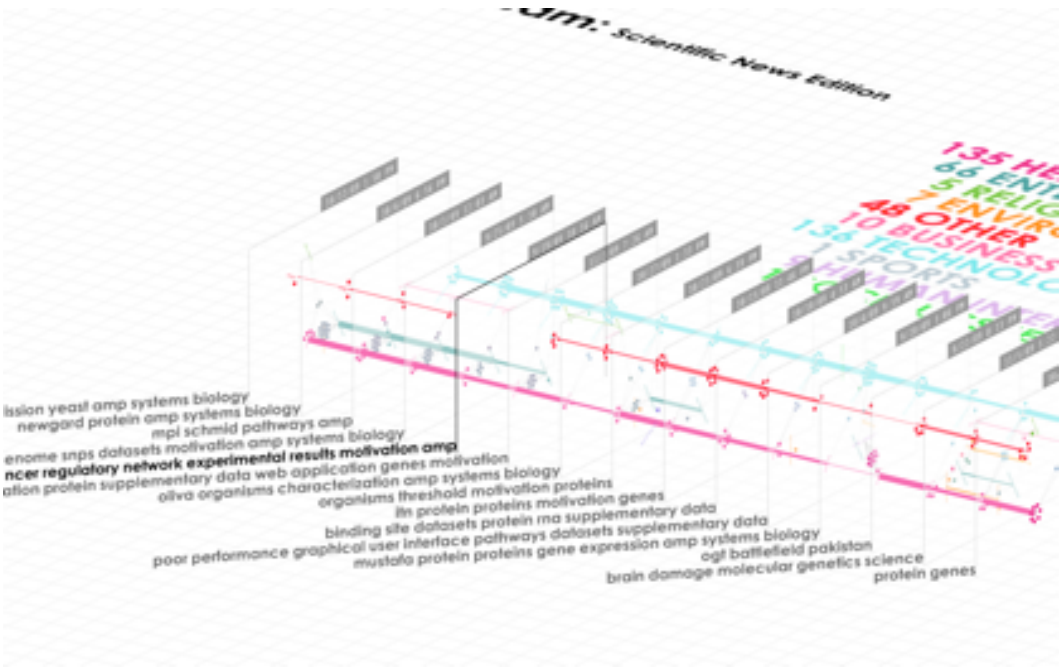


ConnectUs extracts interests of users by aggregating and clustering online life-stream data, which in turn is combined to generate maps of connectivity projected on the ceiling of a social space. The idea was to act as a compass for social navigation, giving subtle hints of connectivity between people in a room based on their interests. The system takes in delicious links and emails and measures similarities between one person and another. It then projects a form above each individual. This form acts as a sort of social compass, deforming in the direction of similarity to others in the room and displaying major tags or n-gram phrases that personify the individual's interests. When multiple people group physically it displays topics that are shared between the clustered group. Those topics are weighted by their relative uniqueness to the given group against other groups in the room.

The *ConnectUs* project marks the beginning of the idea to use the artifacts of individuals to help form connections within a community. The life-stream of traces left by an individual form's something that can be used to computationally measure the distance between individuals. For community events or conferences, where people want to begin conversations, finding some small, possibly unrelated topic to connect on can be instrumental in creating a connection within the community. This same life-stream data could, if viewed as a whole, be used to construct a community data portrait like the experiments below.

ThemeStream

ThemeStream shows streams of information such as RSS Feeds, publications, or other time-based textual documents at different scales and with different compressed views that animate smoothly between each other. Topic modeling and semantic parsing of the text are used to display themes over time in the provided dataset. It provides reflection of information streams at different conceptual levels of



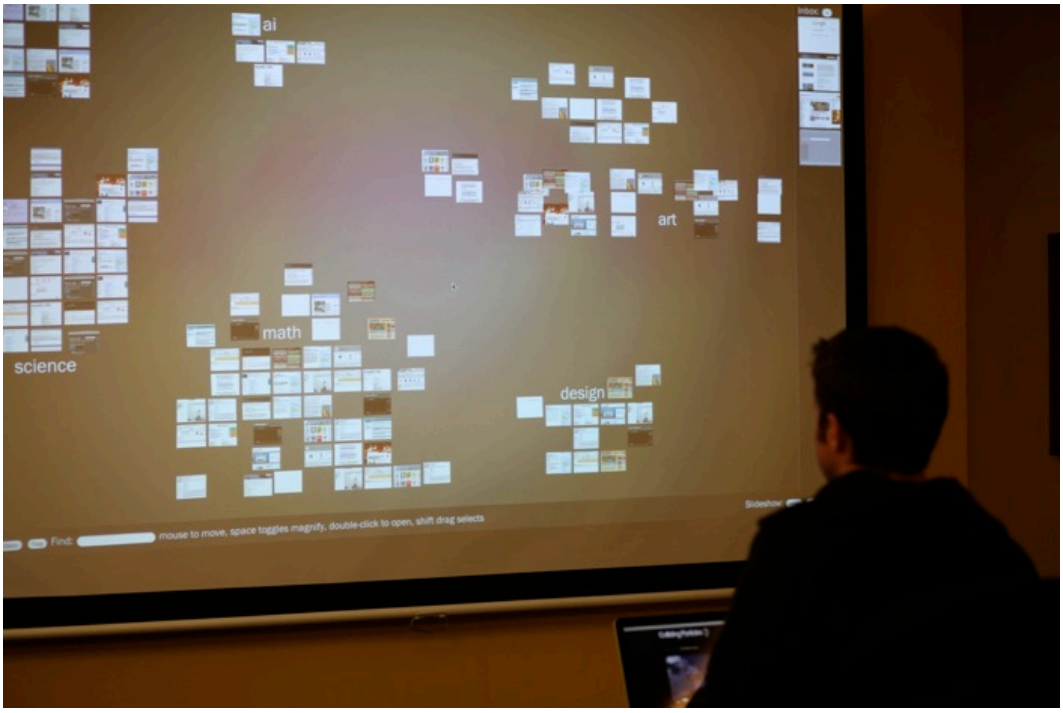
detail.

A key component here was the concept of visualizing themes shifting over time at different scales of resolution. Some ideas shift quickly while others are slow to move. Analyzing the text for patterns at different scales of resolution allows for the extraction of these many overlapping forces that affect trends. On top of this analysis of trends through text analysis, the concept of smooth animations between views of different algorithms and their results is pivotal to maintaining conceptual cohesion between various scales of analysis.

SpaceMarks

SpaceMarks is a tool for spatially organizing bookmarks and emails. It is a zoomable space for organizing one's personal information. One can intuitively group and cluster and move items, and then also bind meta data or tags to the spatially

organized objects. The backend uses reinforcement learning to create a model of why objects are placed where they are in space. The most interesting element here then



becomes comparing one's own externalized mental map to someone else's.

Related in many ways to an ideal knowledge management system, a spatial bookmarking engine like this allows for the organization of any artifact on a computer or on a website. Additionally, because all representations of relationships are organized into spatial arrangement, there is an externalization of fuzzy organization between objects that a computer can understand. This allows a computer to compare a video, to a word document, to a music file, which due to their incompatible features would not otherwise be possible. If such a system were adopted at large, the artifacts it produced would be a partial representation of the organization of ideas and thought processes for each individual. For constructing community data portraits, such data would give insightful information into how people are organizing the information they think about and what resources are related to what other topics across the organization as a whole.

What Was the Media Lab Thinking About in The Year _ _ _ _?

This project was an experiment in embodied data visualization in collaboration with Richard The. The goal was to foster understanding of abstract

information that was spatially or temporarily detached from us. The hope to achieve a more personalized, evocative perception of information that would be hard to grasp otherwise. The project took the form of an art installation in the new Media Lab building that augments the people wandering through it with statistical information. The data used was extracted from the student and faculty publications over the course

of the Media Lab's rich history.

For this project the data overlaid on individuals were three word representations of topics extracted from topic modeling (using Latent Dirichlet Allocation [33]) of the Media Lab theses. In the experiments described in this thesis, this same technique of topic modeling theses is combined with a very different representation as one of the lenses to view the Media Lab community. The forced binding of abstract concepts to actual researchers walking around was the most powerful aspect of this piece, and though the mapping was arbitrary it did help make the data more humanistic, but fogged its interpretation.

Chapter 5: Components of the System

A community data portrait requires aggregating a large amount of information, processing and binding together the resultant data, and then visualizing the results through a series of reflective lenses. The purpose of the research is to explore what questions are interesting ones to ask, and what those questions help expose about the community.

Gathering - is a deconstruction of a data source, ideally it is taking each data source and recursively deconstructing it until it is at its lowest useable feature space. In this case words and word frequencies. If possible, it is best to maintain a traversable representation of the pedigree of the decomposed features.

Processing - processing is a set of actions over the deconstructed features, it is the reconstruction of the feature space which exposes something new about the data. The processed model exists as a superset over the deconstructed feature space of the original information.

Visualizing - is what helps us see, understand, and interact with the processed model.

The system fit a very real need at the Media Lab and as such it is now intended to be taken up as the heart of the Media Lab archive in the future, to be extended and used to fuel future projects across the lab.

Gathering

> In our current society, anything which is not able to be translated into a form recognizable and storable by a computer--i.e. anything that's not digitizable--will cease to be knowledge. In this paradigm, the opposite of "knowledge" is not "ignorance," as it is the modern/humanist paradigm, but rather "noise." Anything that doesn't qualify as a kind of knowledge is "noise," is something that is not recognizable as anything within this system [34].

To gather data, it was first necessary to identify the various potential sources of the data. Data about the Media Lab is sprawled across a variety of sources and needed to be aggregated and organized. The ideal starting data would in many ways be a daily diary of every researcher's ideas and goals and thoughts organized by what they their goals are and their interpretation of the goals of others. But since such data

does not exists representations of individuals must be approximated by the other data they already produce.

Data	Getting Mechanism	Notes
Projects List Database (PLDB) - A sql database maintained by the media lab of current groups, people, and projects at the lab.	SQL interface	Many data inconsistencies and only current data. Included on going and some past projects, research groups, current people and their roles.
ML Publications Site - A publications section to the current Media Lab website.	Custom Ruby Scraper	Not well maintained, metadata inaccuracies.
NY Times	NY Times API + Ruby	Great API, still needed to write custom scraper for full text
Google Scholar	Custom Ruby Scraper	Still no API and does heavy throttling of automated requests.
MIT DSpace - MIT's institutional repository built to save, share, and search MIT's digital research.	OAI Harvester	Cumbersome and slow without an expressive interfaces, but on the plus side it is standardized
Theses files from private FTP - Henry Holtzman had a personal archive of theses.	FTP	Still had no meta data attached other than name and year.
Google News	Custom Ruby Scraper	Only small snippets and unreliably related.
CiteseerX - a scientific literature digital library and search engine	rsynced the MySQL Database	Used to fill in holes of google scholar but not comprehensive enough for the specific tasks of MIT Media Lab's publications.
Stand alone Excel Spreadsheets - examples include a list of thesis titles and students for all theses.	Exported and imported as CSV	Extremely useful, but rare, undocumented, and unpublished.
Sponsor Visits - internal database	Script to grab and parse an internal XML feed.	No way to access it directly, nor to see what is available.
Logos for all Sponsors	Ruby script to grab logos from wikipedia	Though usually organized, not all pages were structured in the same manner.

Gathering is a process of taking everything and breaking it down and storing it into its smallest meaningful components, while if possible maintaining the pedigree of each data point. So, after the data was retrieved in its raw form from the source it

needed to be broken down and stored locally. The raw form of most representations was stored in YAML [35] format for ease of reading and parsing across languages.

Initially, the database to integrate all of this information was intended to be Neo4j [36], a graph database. The reasoning behind this was that a graph representation would allow for more flexible deep queries and a schema-less integration, thus no need to worry about migrations or converting old formats as future information was added. While, the graph structure was indeed flexible, the need for processing every object in a transaction and the requirement to lock access to the database when accessing the graph made quick interactive queries and manipulations difficult. Combined with scalability concerns, and difficulty interfacing the database across languages, the decision was made to move the data to MongoDB [37] with a Ruby interface, which offered scalability and flexibility as a document based store. The document based storage mechanism fit better into a structure that was mostly concerned with the organization of documents and parts of documents connected by individuals and groups. Additionally, MongoDB has strong support for binary data for storing all the raw pdfs and images. This system could also be extended to storing video files in the same manner at a later date.

For quick retrieval both in the processing phase and in the visualization phase all documents were also indexed using Solr, a fast open source enterprise search platform from the Apache Lucene project [38]. The schemas were defined in the models and linked as dynamic fields based on type. This allows for tokenization and full search.

Once the data storage mechanism was finished each document needed to be broken down into its smallest components, in this case words and word counts, and all of the meta data needed to be extracted and merged across various sources.

The words were tokenized using a standard Indo-European tokenizer and word counts and frequencies were stored for each document. Additionally a sentence tokenizer was employed to tokenize each sentence before it was broken down into each word. The merging of metadata was a more difficult problem.

First, fuzzy based matching techniques were tried using the Solr index to help match fields and merge the results. However, the number of edge cases made it continually difficult to produce meaningful merges without destroying some other aspect accidentally. The best tool for the job ended up being *FreeBase's Gridworks Tool* [39]. The data was exported into a large CSV format and imported into Gridworks. From Gridworks, fields like names could easily be faceted and merged by iteratively fuzzy clustering the attributes. For more difficult tasks custom filters and transforms

could be written in the Gridworks DSL, Python, or Clojure. After significant cleaning the data was imported into the MongoDB structure, clean, organized, and linked.

Processing

Processing is taking the fully broken down data points in gathering, and recombining them into structured forms.

The flexibility of the Ruby interface to MongoDB and the rest of the stored data became a key feature when tasked with processing the information. Often there were issues with gathering where it was desirable to quickly pull out an element and test an idea. By having everything in an easily query-able form, running dynamic tests from the command line became a part of the work flow. Interacting with the data from the command line's interactive shell became integral to the process of thinking with the data rather than pre-planning and coding previously necessary for sketching each possible idea. An example of this sort of flexibility now includes queries like the following:

```
Text.search("visualization").map(&:creators).flatten.uniq
```

This simple one-line query, returns a weighted ordering of all people at the Media lab writing about visualization. `Text.search("visualization")` is a full-text search in all *Text* artifacts, which searches both *Thesis* and *Paper*, using Solr the returned results are in weighted order of relevance. `map(&:creators)` gets the creators for each of those documents. Finally, `flatten.uniq` makes sure the list returned has no duplicates. Though this is a play example, it does point to the extreme flexibly and simplicity of interface.

Another example:

```
WordCount.all(:word=>'art').sort_by(&:count).map(&:text)
```

```
WordCount.all(:word=>'art').map{|w|w.sentences.map(&:next)}
```

The first *WordCount* example returns all text artifacts which contain the word 'art' in sorted order by the number of times 'art' appears in the text. This is functionally similar to Text.search("art") but has more control and flexibility. For example, the second *WordCount* example returns every sentence directly following a sentence containing the word 'art' in all Media Lab documents -- a difficult to perform query in any other system.

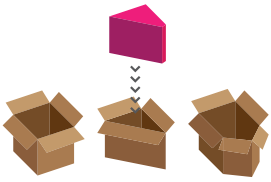
Processing is the section were questions are constructed and answered. For each of the questions in the experiments chapter below, processing included writing a Ruby script to manipulate and gather information into a form that reflects an answer to that question. This could range form using topic modeling, to natural language processing, to simple regex searches. The key to processing something so exploratory was flexibility and quickness to test an idea or theory. For larger tasks like parsing sentence structure of every phrase used by the Media Lab, scripts were run in parallel on the Condor computing cluster managed by the Harvard MIT Data Center.

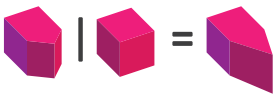
A lesson learned from processing this information in various ways was that there really are no general lessons. Since the things one needs to write are as varied as the questions one can ask -- flexibility is king. The mechanism for asking questions needs to be quick, sometimes the absurd musings elucidate rare connections or artifacts in the data.

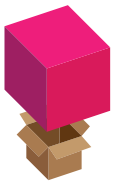
Visualizing

Finally once the information has been processed into its various forms, we need methods of returning those results in a format that people can make sense of. This is a process that is both extremely dependent on the data and on the audience. To be able to create fast visuals to display the information the Processing Library [40] was instrumental in producing results.

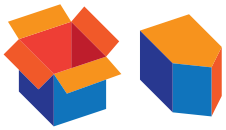
Some of the high level visualization fundamentals employed are laid out below.

1. **Shape Fitting**  How does the visualization create a group portrait to reveal summarized or related information? Something akin to a taste fabric or a related space, or connection to others. It models the behavior of the group into some sort of structured ontology. What is the best fitting box to put a group into?

2. **Caricature**  How well does it exaggerate the data of an individual against a background model of an ideal? How different is this box versus the average box?

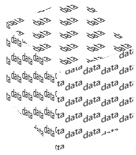


3. **Compression** How much information does it compress? This is a trade off between ease of quickly summarizing the information and having enough variance to discover something interesting. What is the smallest box we can stuff everything into while maintaining an interesting view?



4. **Engaging** How interesting is the data and its subsequent presentation? How long will a viewer spend exploring it? How pretty is the box?

5. **Intimate** Intimate data is the the level of emotional or personal connection one feels with the information. Is the box from Ikea or is it a gift from your little sister filled with family photos? It is purposefully confounding emotional response to focus on the ego, and what is important to an individual. This ego-focused approach is very helpful as an additional driving factor toward feature four, engagement.



6. **Data as Interface** The interface to information is best served if its form is itself an exposure of the underlying data. Even if that display is purely for visual texture, using the raw data add a richness to the presentation and can begin to help move toward representations that can explain why something is the way it is. This is only possible if the compressed form is unwrappable into the original source. By doing so, the inclusion of the raw form adds credibility to the representation.

These high level visualization fundamentals were the driving methods of thinking about how to represent various aspects of the data. The success of a visualization in conjunction with its design and mapping of visual features to data features, is its ability to perform the above fundamental tasks well.

Chapter 6: Experiments

> Kevin Lynch, in his pioneering book, The Image of the City, wrote: The city is a construction in space, but one of vast scale, a thing perceived only in the course of long spans of time. City design is therefore a temporal art, but... on different occasions and for different people, the sequences are reversed, interrupted, abandoned, cut across.
>
> The role of the designer is not to make a single, perfect path, but to create a space that, in Lynch's words, is "legible", one that is easily organizable into a coherent pattern. The image of the city - or of a community - is not single frame, but a series of impressions, an image in the round built from a series of successive views [9].

The representation of a community is fullest only in the combination of our projections and slices across it. Through varied lenses, we view a world we are already a part of, but whose memory is distorted by our nature. To reflect honestly on our community, there is a need to augment our perception by rescaling and reaffirming our internal models. The following designs are attempts to produce those varied lenses, whose aggregated representations form a community data portrait.

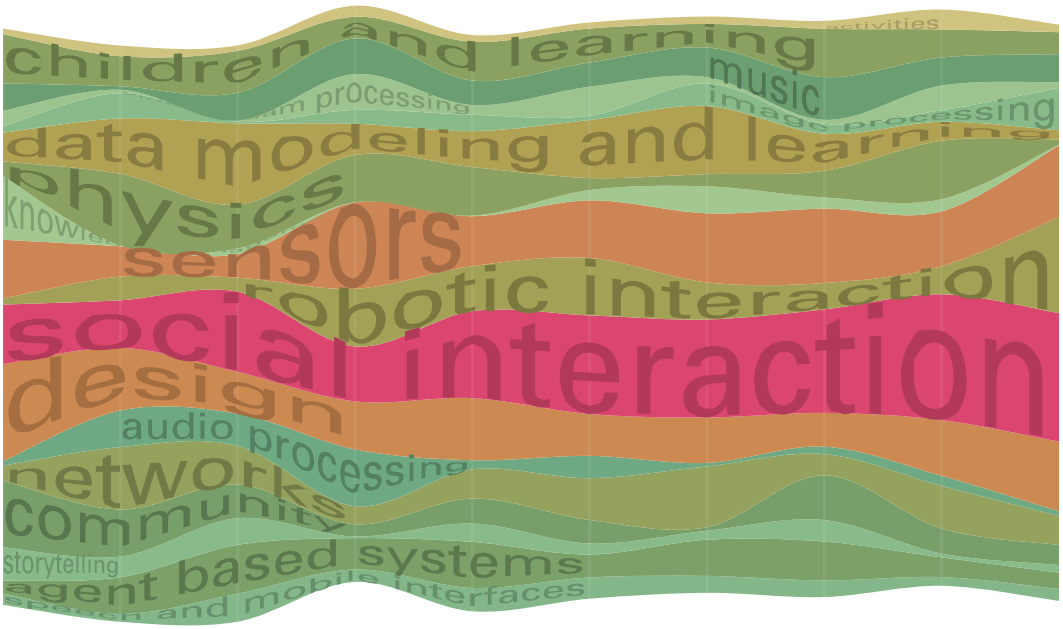
This chapter breaks down the experiments into the three core themes of community: events, people, and ideas. For each of those themes, a series of experiments are discussed. For each experiment there is a design, discussion of the design, and feedback from alumni and students in the Media Lab community.

Events

The first section of experiments are focused on the theme of events. Events are the temporal memory of a community. While this can include single events, like the opening of the new Media Lab building, the more difficult temporal element to extract is the trend of events. If each publication is taken as an event, the goal of this section would be extracting the gist of those events to help discover underlying trends.

How have topics changed over time?

Topics were extracted from Media Lab theses text using Latent Dirichlet Allocation (LDA) by way of the *Mallet Toolkit* [41], the result was then visualized in a stacked graph manner [42]. In this case, it is the top twenty topics derived from the Media Lab theses between the years 2000 to 2009. The topic labels were hand generated by looking at the frequency of the top words associated with each topic.



Using this technique, the visualization is able to display some shifts in major trends over the years at the Media Lab. Because of the Media Lab's heavy emphasis on research directions directed primarily by the Principle Investigators of each group, the most significant shifts often occur with the changing of faculty. Such as the increase of robotic interaction with the hiring of Cynthia Breazeal or image processing with the hiring of Ramesh Raskar.

Some responses to the representations included:

> "Interesting that social interaction has increased as much as it has. Not surprising, given the industry, but it doesn't feel reflective of the lab."

> "What I like about these views is that it allows me to think about the lab's impact in different areas over time, in that how i normally think about things is I think about certain groups, and i have a mental model of what groups are more prolific than

other, and this is a more objective view of what themes the lab is interested in over time."

"Where has all the music gone."

While this sort of visualization is good at producing a very high level representation and compressing the themes, the interpretation of what a theme means is often quite difficult. Adding significant papers that could be clicked on and explored for each segment of the visualization would go a long way to improving how the understanding of the representation, and adding trust to the model it represents. An additional downside to the approach includes the requirement for human labeling. Human labeling is both time consuming and like any label capable of being misinterpreted due to the biased perspective of the original labeler. Human labeling could be aided by multiple people labeling the same topic and better machine generated primary guesses for what the label should be from the underlying features. Finally, the topics have no hierarchy of concepts, the last downside could be addressed with the application of hLDA [43]. Using a method like hLDA, topics like music could have subtopics like music performance and sound processing.

Additionally, there is a concept of support visualizations. Support visualizations are works intended to further enhance or elucidate context or other features of a visualization. As an example, with certain timeline visualizations presented later in the experiments chapter, a support visualization would be one that displays events in the world next to what is happening at the lab, or a presentation of sponsor visits next to trending topics. Support visualizations act as points of comparison and cross reference of information to situate the primary representation.

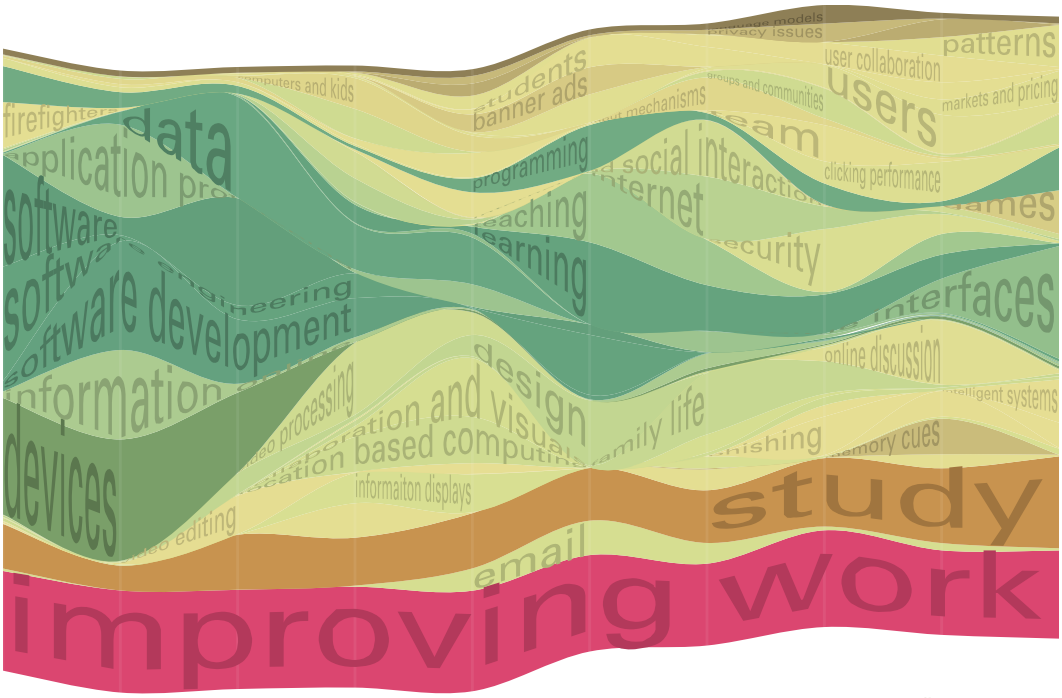
To add context to such a visualization, support visualizations were constructed in parallel. The examples above include a visual map of major events over time, and a map of sponsor visits over time. One criticism of these visualizations is that the major events displayed were too far removed from the events in the Media Lab visualization to have real connective meaning. While at times some of the visuals could help to trigger associated memories of what was happening at the lab, overall this support visualization would have been better served with a more Media Lab specific contextual history. The best approach in this case would be to have users annotate an interactive timeline on top of the visualization in a wiki oriented manner or annotate the trends directly with their thoughts of what had happened. With

regards to the sponsor visit timeline, there is obvious fault in using data points of various visual density to represent the same value, however the iconic and emotional nature of using the company logo is compelling. It would be best designed as a frequency plot of dots who's color is derived from the average of the logo, and then display the logo off to the side once for each sponsor. In general, if the visualizations were more interactive these sorts of contextualizing clues could be overlaid transparently or organized together.

How do these themes compare to other similar places?

Below is another similar visualization using topic modeling and a stream layout, but in this case the data sets are the publications from the MIT Media Lab and the publication from Carnegie Mellon's HCII department. The same time span of 2000 to 2009 was used, but twenty-five topics were chosen instead of ten.

Media Lab Publications (2000 - 2009)



CMU HCII Publications (2000 - 2009)

In the above representation we see the comparison between the MIT Media Lab and CMU HCII. There are some shared themes like improving work or efficiency of users, but in general they are quite different. Some noticeable other trends include CMU HCII's heavy emphasis on user studies as seen in the study topic which is not present in the Media Lab representation. Additionally, the Media Lab representation is much more stable in the topic areas it tends to focus on.

Here are some of the reactions received from students and alumni:

"The Media Lab is so much less bursty."

"What does improving work mean?"

"I guess that makes sense."

"I wish I could see why."

"It is what you would expect, but to see it validated in the visuals is both gratifying and unsettling."

The ability to compare one community against another is a good approach, but the example has no real grounding of what the representation means or why as a viewer it should be trusted. This again goes back to representing more of the data itself in the interface. Each node should be interactive, when selecting one aspect of

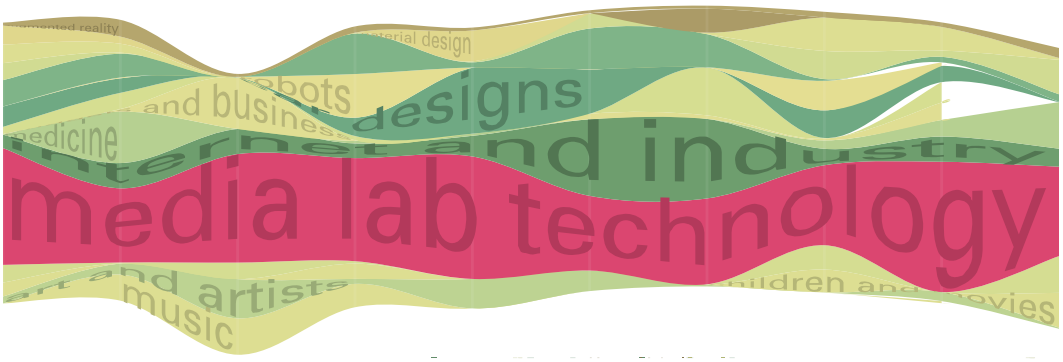
the visualization, a theme in a given year, that action should produce what the associated documents are for the that point in both visualizations.

It was only after a participant commented on the bursty nature of the CMU representation, that it occurred what feature this representation best discovered. The stability of the Media Lab is a product of its stable group-led agendas. The publications presented by CMU HCII seemed more in accordance with either trend chasing, high turn over within the department, or students setting much of the agenda. The opinion of someone who had been in both departments added that as the visualization reflects, the Media Lab does feel more diverse, but more stable in its diversity; as compared to the more unified global themes yet more frequent minor shifts in topics found at CMU. This example, expresses the ability of such reflective visualizations to evoke more questions, rather than give a definitive answer, participants were curious intrigued by the representation and most curious about why various shapes took the forms they did.

In response to the question, "What does improving work mean?" it was a label given to represent efficiency and optimization at ones given task. It is a major theme in both the Media Lab and CMU, but the misinterpretation of what it means just goes to show the fault of human labeling in these types of representations.

How does research compare to its press coverage?

Another topic modeling and stacked graph example was used to describe the themes latent in New York Times articles about the Media Lab. This data was taken from NYT articles over the same time course as the other two experiments, from 2000 to 2009. However, this example is the least revealing of the three experiments using the above mechanism. The data is not rich enough to require the compression



into themes over time. While it is possible to pull out partial events from the themes,

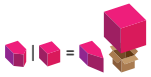
such as the section on medicine in the early 2000's correlating to the entrance of Frank Moss as the new Media Lab head and the start of the New Media Medicine group, the low number of data points for each of the themes makes the representation difficult to derive meaningful global patterns from.

This attempt could be more useful if done over all blog posts on the internet, which mention the Media Lab, or more simply the archive of internal press releases in comparison to actual research themes. If the idea it to expose some sort of disjunction between how the internal representation and the external representation of the community, then better more comparable data. Unfortunately, such information does not currently exist. Finally, the movement into yet another topic space makes comparison across previous representations meaningless. What this means is that the topics used here are different topics than the ones used to describe the internal representation, so it becomes difficult to compare topics since they are in completely different spaces. However, this use of a different topic space is unavoidable because there are so few documents in the New York Times compared to the other corpora that developing a unified topic model that would be accurate at describing both corpora equally would be impossible.

"This doesn't really tell me much."

People

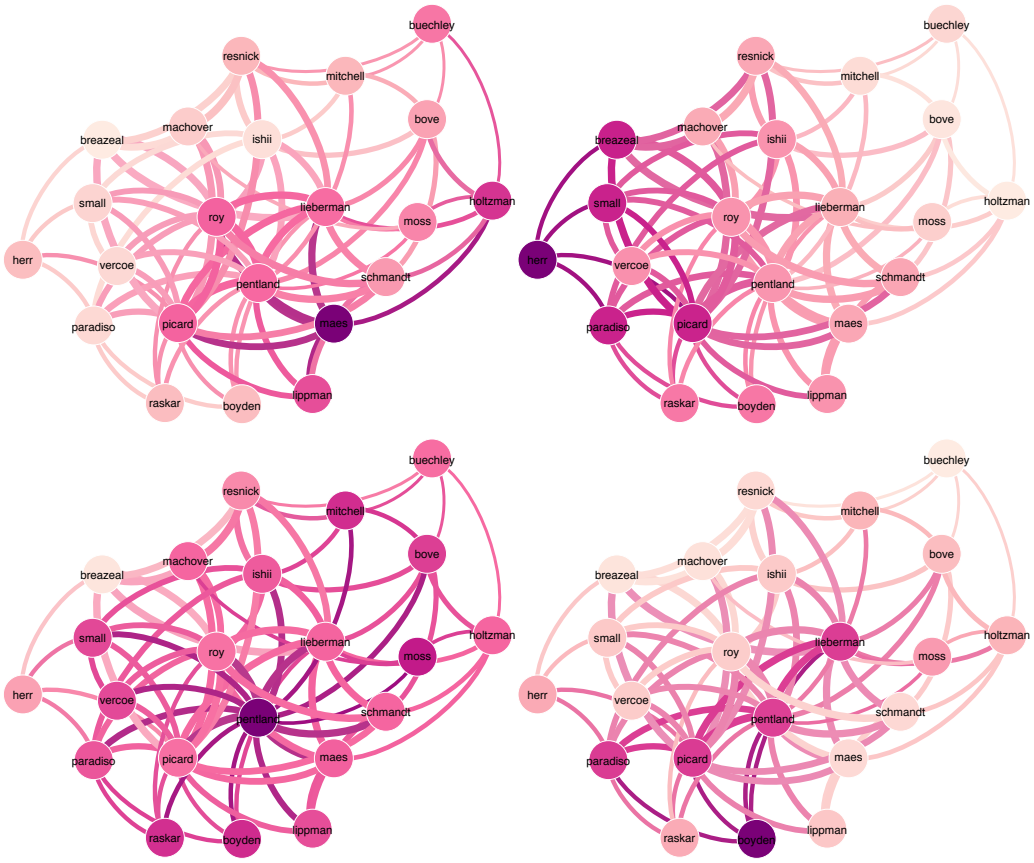
The second theme of community explored is people. Because of the difficulty to quantify a person, people are often ignored in the majority of quantitative analytics. However, community is the product of people interacting and is therefore a key theme in representing a community. Our biases, our relationships, our ways of working together are all a product of this sort of information.



How do people group together?

For each current Professor at the Media Lab, a distance measure was calculated to the three nearest neighbors. Distance was calculated as the tf-idf [12] distance between the collective works of each professor. From theses linkages, edges are drawn weighted by the strength of the linkage. Because of this rule for adding edges, each node will have exactly 3 outgoing edges and between 0 and n-1 incoming edges each with a varying strength. The graph is then rendered in a force directed layout [44], meaning that like a physical simulation of springs holding together the

nodes based on the strengths of the edges they will self-organize to reduce energy across the graph. The items are presented as a small multiple where each node is colored in a desaturating fashion decaying outward from a central node. In each version of the multiple, a different start node was chosen from which to expand outward. From this representation, one may gather similarity of language used between the professors, and how similar one is to the rest of the themes at the lab. By coloring the nodes, it makes visible clusterings of related professors by the words they use.



From these graphs it is possible to see similarity of word choices across different Media Lab Professors. From their choice of words, we can also see clusterings of connected individuals. Finally, there is an issue of interpreting what it means to be highly connected or not to the community. This could be due to a many factors, not having as many documents to represent an individual, working in a very different field, or changing fields or focus over time. This representation does not include how these connections would change and shift with time.

"That is really, really, really interesting." [In the context of realizing the implications of what high similarity of word choice shows about the interaction between Professors at the lab.]

"This has a lot of power, but it requires more explanation."

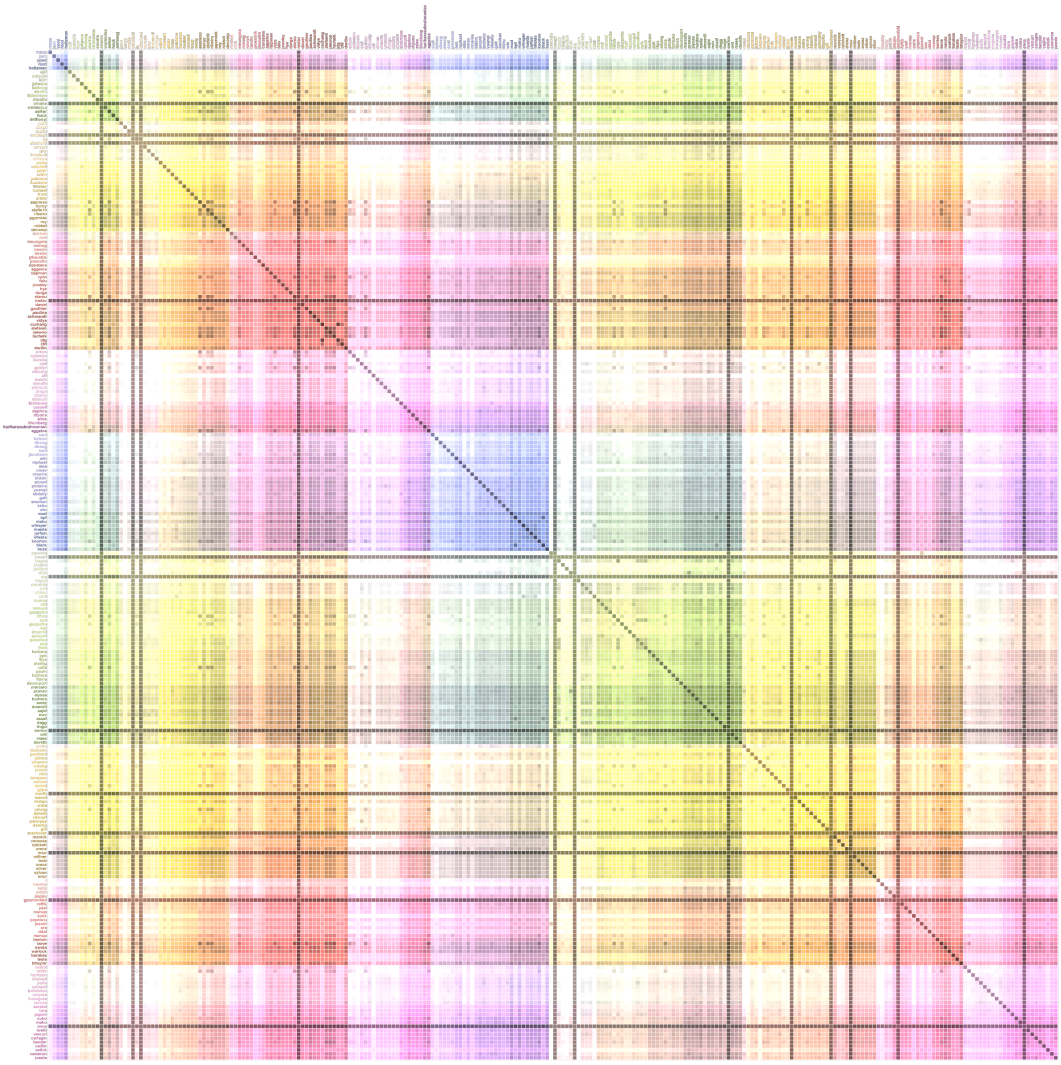
"There does seem to be a grouping of people that have been at the lab for a longer period of time, which makes me curious about what the lab's common vocabulary is."

The power of this visualization comes from the mechanism of generating edges, combined with the presentation as a small multiple. The small multiple presentation could however, be better represented as an interactive force directed graph where the center point shifts in accordance to which node is the focal point. This method did do a good job of helping to expose the invisible structure, pointed to as one of the main problems to address in this thesis. It is possible from this to see one view of how groups are related that would not be understandable from a Media Lab brochure about the various groups.

What are the unexpected similarities?

The idea in this visualization was to present what the unexpected similarities are between individuals at the lab. To do so the form was chosen so that expected similarity would form a predicable and viewable pattern and someone who was very similar, but did not fit the overall structure would be easily detected as an outlier to the pattern.

The idea in the representation below is to better be able to discover outliers in the global patterns of connections. There is an expected pattern of similarity, where members of the same research group are similar to each other, etc. However, what is interesting are the instances when what one expects turns out to be false. One wishes to be able to discover outliers in the information.



The representation is simply an N by N distance matrix between each individual to each other individual. Tf-idf text similarity across each individuals project descriptions, theses, and publications over all available years was used as the distance function between individuals. However, any other distance function based on those resources could be substituted. An example alternative measure of person to person distance could be the distance between the underlying topic models for each individual. This would most likely be a more reliable representation. The structure sorts the names by group and then by year. Each group then has its own color, and the intersecting color is the hue merge of the two, where the brightness is the similarity and darker means more similar. So there should in theory be high similarity along the center line in squares that represent each group's similarity to itself, and outliers i.e. dark positions outside of the centerline would be individuals that did not fit the semi-rigid hierarchy of the lab. These would be people that had collaborated or

are doing similar work across groups. It is also simultaneously a measure of the overall similarity of a group to the rest of the lab. These sorts of multiple representations and meanings within one graph are the root of its problem as a representation. First the number of items shown is too large. If the intention is to show similarity to others, an interactive solution where one pivots on each individual would be more apt. An N by N matrix has too much information. Second by confounding looking for outliers across groups with displaying similarity between groups, neither is discernible. The correct choice would be to normalize the representation by the similarity between the two groups, where the similarity between groups is the average between the collective members in a group.

> "I can't look at this and understand what this means as a whole."

> "I want this to be interactive."

> "Least successful."

> "Really pretty."

The response from participants mirrors the disjunction in the representation, but there is still value in displaying outliers, it is just that this representation does not make the outliers easily visible.

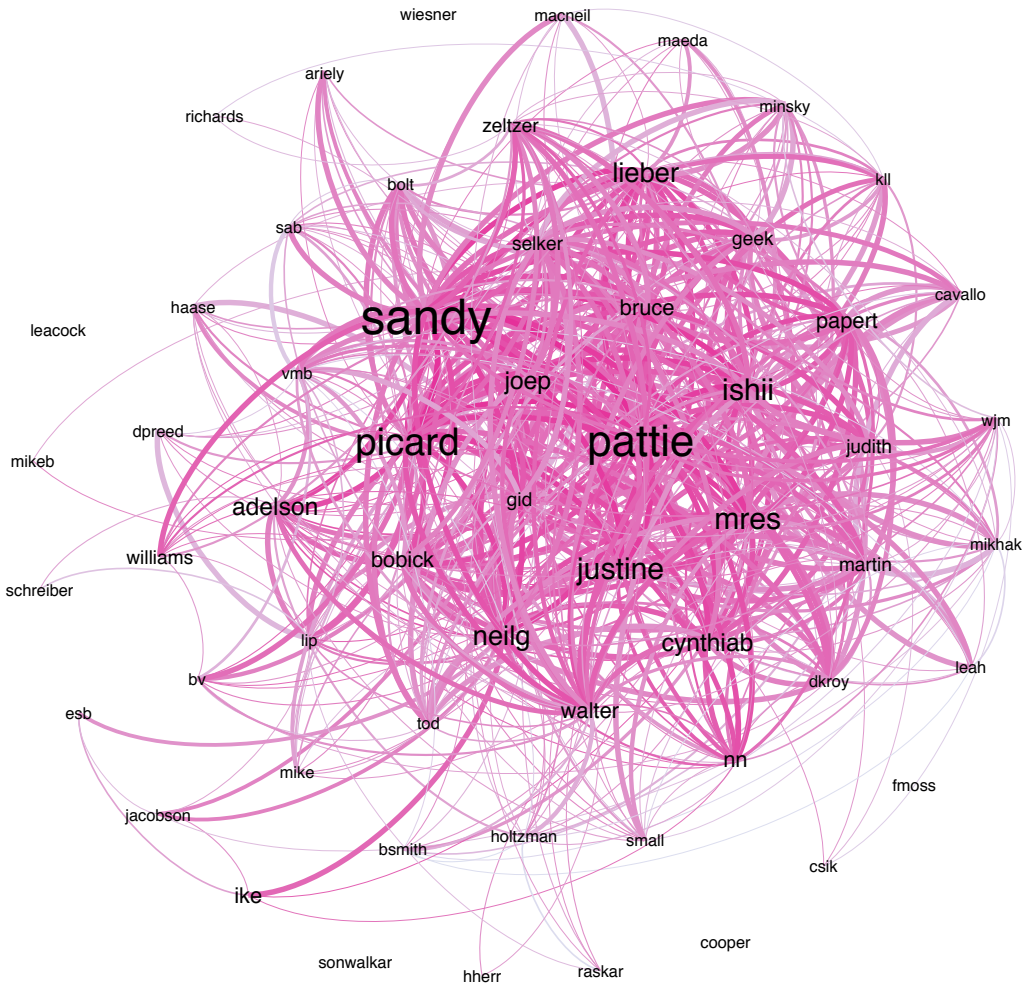
What is the terrain of the publications?

Below is a force directed graph layout [44] where the size of the name is proportional to the number of times an individual has been cited and linkages are created by co-citations between authors in some third party document.

> "Very misleading. The comparison point is two separate dimensions."

> "I think the co-citation measure is a valuable metric, but I need to understand that relative to the person as a whole."

> "Things that are obvious are the people on the periphery and the people in the middle, but the spacing and the connection between items is difficult to understand."



Because of confusion over the links, this view may be more successful if presented as a small multiple where each node has its links highlighted while all others are receded. Again this is a conflagration of representations, combined with not enough information. The total citation count and the co-citation don't mean the same thing and it is difficult to comprehend or associate the two when presented simultaneously as different features of the same graph.

What is their web-index?

Can we use someone's web search popularity to accurately plot popularity of researchers like John Maeda or Tod Machover who have poor publication indices, but are very well known in their respective fields. The measure derived below as a researcher's web-index was the normalized value of the number of results returned for each search engine where the professor's name and "mit Media Lab" were required to

be present. This search was run and averaged across Google, Bing, Yahoo, Wikipedia, Baidu, Google Images, Bing Images, and Yahoo Images.

Below are a few of the top results, see the appendix for the full set.

Name	Web-Index
Nicholas Negroponte	1
John Maeda	0.325235743480197
Pattie Maes	0.25379391323571
Seymour Papert	0.182123712324354
Hiroshi Ishii	0.173305833105389
Mitchel Resnick	0.169452867250559
Walter Bender	0.1554063611363
Tod Machover	0.141663842999596
Marvin Minsky	0.128693522078208

> "I like that it validates John's opinion of not writing papers and just getting your ideas out there"

> "It is well known that the number of google results should be treated more as random than real, it is inaccurate at best."

A measure like this is fundamentally flawed if interpreted as relevance or importance. This is a measure of search engine results, nothing more, and even that is apparently faulty. Having said that, how popular one is on the web does seem to correctly identify major figures at the Media Lab who were otherwise interpreted as minor by traditional citation metrics. However, again this is in no way an accurate measure of importance. Every web result, be it a tweet, or a NYTimes articles gets equal weight. It is therefore highly skewed to what is in the public opinion right now as well as in the opinion of people who are willing to write about such topics on the Internet. Additionally, because the query required 'mit media lab' many of the individuals how have extremely prominent careers in other institutions either before or after the lab are not represented. This requirement thus skews the results toward individuals who's identity is in some ways tied to the lab.

Another approach extending this method would be to show the citation count visualization with the same structure as the web result visualization, emphasizing those with the greatest discrepancy between the number of web-results and their number of citations.

Ideas

Finally, ideas are the most essential element of a research community. Ideas are in many ways the product of the above two themes, events and people. However, ideas are in their own way a unique method of understanding community.

What where they thinking?

In the Appendix, there is a list of the longest questions asked in all the theses from 2002 to 2009. Why would we want this? The foundation of the scientific process is about asking questions, what better representation of a scientific community than the questions they are posing to the world. Questions also represent a good metric for state of mind or framing of an idea at a given time. A few examples are highlighted below:

2002

"Why is it that when a person is shown a picture of a flower girl dressed in white, she assumes it is at a wedding, and she wonders who is getting married?" --hugo-meng
"How perfect would it be if one could throw a ball over a net, and have it come out somewhere else with the same characteristics, i.e. spin, speed and direction, all in real-time?" --florian-ms

2003

"What will the future look like?" --pangaro-ms
"What defines the unique sound of a singer?" --moo-phd

2004

"What is the soundscape of your kitchen at breakfast time?" --hugosg-ms
"What is the role of imagination in online dating?" --atf-ms

2005

"When can augmented reality and ambient interfaces improve the usability of a physical environment?" --jackylee-ms
"How are these conceptions of 'public' and 'private' space evolving today, as we use new communication technologies to weave our private social practices within public spaces?" --lilys-ms

2006

"How effectively can memory problems be addressed via information-retrieval techniques applied to a personal-data archive?" --vemuri-phd
"How may the people simultaneously known as women and as scientists - an oxymoronic social subject only beginning to break down - intervene in the construction of the potent natural-technical objects of knowledge called females?" --gemma-ms

2007

"How can new media contribute to the role of architecture, and extend its meaning, so that we can conceive new kinds of public media that contribute to the social, cultural and political meaning of places?" --orkan-ms
"What happens when we wish to collaborate, to pioneer new ideas and movements, and to mutually capitalize on one another's strengths, outside the boundaries of real space and real time, and beyond the scope of any single nation's law?" --bpf-ms

2008

"How much energy per bit is required for inter-particle communications?" --ara-ms

"How can a tangible interface retain the immediacy and emotional engagement of "record and play" and incorporate a mechanism for real time and direct modulation of behavior during program execution?" --hayes-phd

2009

"Which of our mental capacities and human experiences will we choose to amplify and enhance?" --ericr-ms

"Why are tangible user interfaces still predominantly confined to the lab, even after 20 years of compelling research?" --kumpf-ms

Some reactions received:

"I like comparing everyone's questions to each other, because it lets me get some sense of their work without it being an abstract. It has more teeth to it."

"Looking for the longest question is arbitrary. There are other metrics to pull out for their work."

"This is cool, I like this one."

"Its curious to see that some of the questions are more representative of the project and some are more representative of the person."

"This is funny, because it is my year so I really want to see this."

"I really like the notion of slicing into people's work in this way and I want to see it explored more thoroughly."

"Language, writing, style, it is a compelling format to have just a snippet. And then you can compare a lot of different people."

"It is a sort of reminiscing. Does it seem like it fits that person? I want to see the people I like, the people I don't respect, etc."

"This part connects to the people in a more personal way."

This set of representations seemed to be universally the most engaging. Because every participant was from the lab each had some sort of personal connection to someone on the list. Immediately, they would look for themselves or their friends. On top of that many found the slices to be very engaging. However, a few also commented that many were too short or too out of context to be meaningful. An interesting algorithm would be how to expand outward from a given phrase or sentence to make sure a complete thought is encapsulated. It may be possible to do so by making sure a proper sentence structure is matched. First one could tokenize the sentence, then find the pivot point using a regular expression, next

the surrounding sentence could be parsed for structure using OpenNLP to parse the tree, making sure there is a complete sentence of subject, verb, etc, and no things like demonstrative nouns pointing to concepts in other sentences. If not true adding previous or next sentences to the phrase until it satisfies the above constraints could help construct more meaningful slices more often.

Another direction this could be taken in is to analyze the questions, by looking for small repeating verbal patterns and zooming out until all that is visible are the visualization designs representing the patterns. Showing all the uses of action verbs, all the uses of the same most common word, the proportion of 'what' phrases versus 'who' phrases, etc. This comes back to a need to better cluster or represent the list of questions into something less intimidating to explore.

What were previous questions about visualization?

As an additional variation on the above experiment and a single topic was picked and then all the questions were found relating to that specific topic. Given that this thesis is related to visualization, below are the previous questions about visualizations in all Media Lab Theses from the years 2002 till 2009.

2002

"What if we started to reveal all or some of the participants' behaviors to visualize a newsgroup?" --hyun-ms
"Did thinking change as students visualized concrete spaces and actions at discrete time points?" --nbreyer-ms

2003

"How do you learn to filter information, make decisions, and visualize alternatives?" --jbeaudin-ms
"How is it that the interests of program organization and visualization are aligned?" --ch-phd

2004

"What information about authors can and should we visualize?" --ethanlp-ms
"How easy or hard was it to interpret the information in the visualization?" --ethanlp-ms
"How appealing or unappealing did you find the visualization?" --ethanlp-ms
"Is there information about the authors that is not currently shown in the visualization but which you think should be?" --ethanlp-ms
"How can the author pre-visualize the audience's behaviors?" --ppk-phd
"How does the computer monitor and visualize various uncertainties?" --ppk-phd
"Why should we be interested in visualization?" --fry-phd
"What advantages does a graphical or screen based environment offer in terms of data visualization or gestural representation that a physical system cannot?" --amanda-ms

2005

"What is being visualized?" --fviegas-phd
"What good is a visualization tool then?" --fviegas-phd
"What is the goal in visualizing archives with which the user is already familiar?" --fviegas-phd

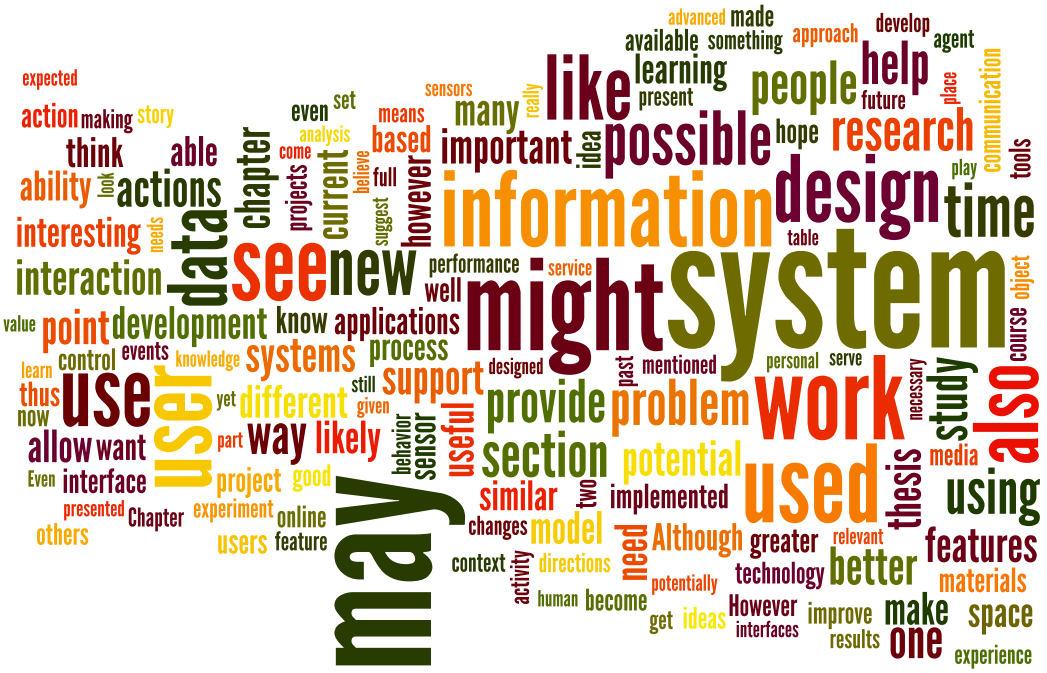
2009

"What rewards does the visualization of attribution provide to the original contributor?" --yannick-ms

These questions are interesting in the same manner as the last experiment, however now there is a more coherence theme that can be extracted. This sort of coherence does seem to allow for better comparison across the questions, but at the cost of not representing the diversity of approaches and people across the lab. It does then point to clustering or dynamic filtering being a necessary part of presenting the full set of questions to a larger audience.

What was the future like in the past?

Research is often focused on what things will be like in the future. In fact, the phrase "in the future" is common in the tested corpus. Pulling out all sentences with that phrase gives another representative of frame of mind or mini-vignette, specifically, the forward looking frame of mind. Below is just a word cloud example



of what things are said in conjunction with "in the future". See the appendix for the actual phrases.

> "It is kinda bizarre to see the hesitation in the future statements."

> "I like finding different phrases, such as in the future, that by themselves seem interesting. Then going and looking for other instances for comparison."

"I like that this isn't trying to represent the entire idea, but instead let's a random slice stand in for the whole."

"I much prefer reading the actual phrases rather than the tag cloud, but I don't mind them in conjunction."

The word cloud was not as engaging as the questions themselves, this is in many ways due to not having anything to compare the frequency of these words against. The "in the future" word cloud should be compared to an "in the past" word cloud for example.

Surprisingly, the "in the future" phrases were much less interesting and had fewer complete thoughts than the questions. Many of the "in the future" phrases would need additional supporting context.

Lastly, although the lack of removal of stop-words from the word cloud was at first considered a bug, presentation revealed an interesting and surprising trend toward much shakier and less definitive statements being used when talking about the future at the Media Lab than originally expected. This behavior is seen expressed in the extensive use of words like "may", "might", and "possible" when talking about the future. Although it makes sense that any academic writing about the future would use qualifiers, it is interesting that the model of how one perceives the lab is of such a forceful assurance in the future that is not represented in the formal writing.

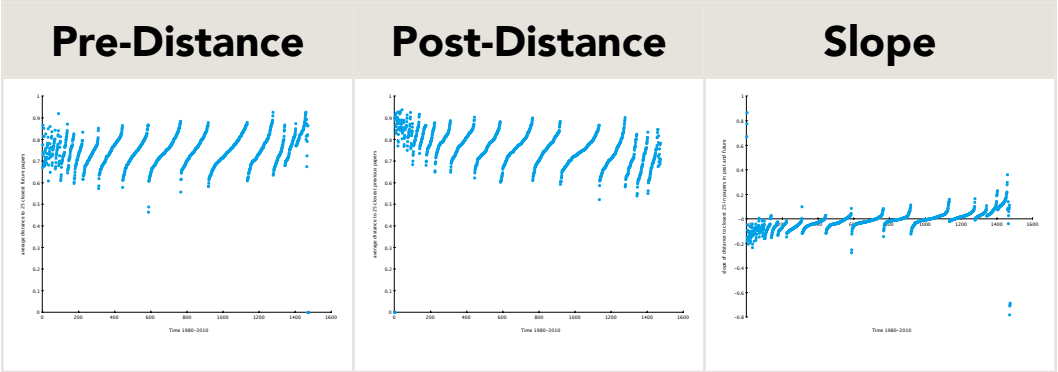
How do the ideas relate to one another?

This experiment offers a new approach for calculating papers within a community along the axes of predictive vs reflective and outlying vs mainstream. The method is related to k-means clustering [45] but focuses on partitioning based on the time position for each document.

The algorithm used is:

1. Calculate the complete distance matrix for every document to every other document.
2. For a given integer k and document x, calculate the average distance to the k nearest neighbors in the past and k nearest neighbors in the future relative to x.
3. Let m_x be the slope between the average past and future distances to document x.
4. Do this for all x.

This algorithm was applied to the Media Lab publications dataset. Using this simple metric we create a methodology for measuring, predictive, reflective, outlying, and mainstream ideas within the Media Lab relative to itself.



\ - **Predictive** - Down slope is a negative number and is closer to future papers. Think of a ball falling down the slope to the future. The steeper the slope the stronger the connection to shifting the field.

/ - **Reflective** - Up slope is a positive number and closer to past papers. This slope could be because there was just a lot of work done in the past already on this topic, or it could be that there was an event that caused this type of work to stop being produced at the lab later.

— - **Outlying** - A high value with a low slope is an outlier. Something unlike things in the past or future.

_ - **Mainstream** - A low value with a low slope is mainstream. This is everything else and generally boring or expected.

There is a global trend of increasing slope which is to be expected in the data. As one moves closer to the edges either in the past or in the future, one has many more papers to be similar to at the opposite end of the time scale. Because of this what is interesting is not really the absolute value for each data point, but the relative values within each year. The slopes of all papers in early half of the dataset may be negative, but some are more so than others, and their relative differences are revealing. It should be noted however that the comparison to the past and the future is relative only to other papers at the lab, so outside influences, like a trend that may have been started here but developed in the world at large are not being represented.

This data should not be interpreted as a chart of who is forward thinking, and who is a late-comer/boring. This is a measure of paper similarity to the things after it and the things before it. It has no notion of the quality of the papers, something that

a citation network is a much better predictor of. This method has no notion of qualities such as if a given paper was a definitive document that solved a major long standing issue in the field. Such a paper could be classified as looking backward, since it references so many things before it and being definitive would close out debate and future papers on the same topic. Conversely, a document that talks of many of an upcoming problem, but produces poor or no results to solve those issues, could be classified as a trendsetter simply because it came first, regardless of not adding any significant advancement to solving the given issue. The method is however, useful as simply another way of looking at the information. There is value here, but the interpretation of any results should be taken with caution and with the constant correlation of any insights with data from a different vantage point.

> "The problem is this is relative to the lab rather than to the world."
>
> "I think the concept is innovative, but I would like to see it run on something more global."
>
> "It is actually really interesting to see what were some of the big leaders, though I don't think the latecomers are as useful."
>
> "I don't understand how it works, so I am skeptical of what I am looking at."

The method is an interesting new metric, and in looking at the result of this algorithm (available in the appendix) it does often produce results in agreement with the internal notions one has of the given papers.

Part of the beauty of this approach is also that it does not need to look at citation networks to analyze trend setters, which is the current methodology for extracting who is setting a trend. Citation networks are good metrics, but only if the citation graph is large enough to support meaningful analysis. In this case for internal use at a single lab, with little cross group citation, such methods would not prove useful. Citations can be very useful in helping to weigh the influence of the work though. For example something like *AgoraPhone* only has four citations on google scholar, but given this data set it was highly correlated to later projects and very uncorrelated to past work it was categorized as a predictive or a trendsetter. Values here could be combined with their citation count to help express influence.

The way to really improve this process would be to do the same method, but calculate the distances and run the trendsetter algorithm for every paper, including those outside the lab. While this not currently possible for all publications it could be

done for all publications in a single conference, like CHI for example. Then the results could be broken down by resulting university or party to see who were leaders in a given field or journal. Those results could be compared against the same results derived by citation analysis or combined with them.

Chapter 7: Conclusions

Other Communities

The algorithms and methods are intended for release to others to test with their own communities. By making these same methods available it creates opportunities for better analysis across communities. How does research at university X compare to what is happening at university Y, or even within a university how does department A compare to department B? The possibilities exist to test the same approaches in domains outside of research communities, community data portraits of businesses are obvious extensions. However, any collection of people working together in the modern world are generally producing digital artifacts, and as such are candidates for a community data portrait. Processing and reflection on those artifacts should follow the same principles outlined in this thesis. These portraits are compressed views into our group actions that help use see our collective history from a new perspective. That new perspective affords use a point of view to test our awareness of how we understand our own communities -- to view with new eyes and from a new angle something we thought we knew, but was too large for us to know in full.

Future Designs

Given the knowledge gained from these experiments, and the data now available, how should the community portrait for the Media Lab be constructed in the future?

For each representation, there needs to be the ability to expose what is making the connection in the underlying data and recursively explore the components that make up the larger representations.

For the first theme of *events*, the timeline representations should all have an additional layer on top of the topic modeling trends which is editable like a wiki. This additional layer should be available so that stories can be added, annotated, amended, and the underlying papers for each of the topics that explains how they were constructed can be accessed interactively in-place.

For the second theme of *people*, to portray the community there needs to be more integration of the people as first degree objects. All of the representations presented above were abstract forms and names. Simple traditional portraits of people

at the lab by year, which are averaged and merged into a single portrait is one way of doing exactly that. Traditional portraits of individuals hit a deep emotional response, they are intimate, by averaging them together they are a compression, and by using the data itself as interface the representation is self explaining. When topics are being filtered or only a subset of the community is being shown in a certain visualization, this collective portrait could filter out the images of non-visible members so it is a dynamic average portrait of all the people who's data is being viewed at this moment.

In the exploration of relationships experiment, for example the graphs of related professors or students, there needs to more ability to expose why a connection exists. The best method to include, would be to allow a relationship to be selectable, and upon selection would reveal the actual documents or phrases that make the connection. This again follows the rule of recursively exposing the underlying construction of the portrait.

For the last theme of *ideas*, the mini-vignettes produced from interesting phrases and questions worked well, but it should be extended to group and cluster the representations into more structured and browsable form. For example, which questions were most similar, or what were all the 'who' questions? Also, more phrases should be tried, such as all of the action verbs, or all the adjectives used to describe the lab. With regards to the trendsetters experiment, the algorithm works as expected, but there is need for more explanation as to why an outcome exists. The final presentation should take the form of an interactive timeline, but on selecting each classified, trendsetter, outlier, etc.; one should be able to see what it connects to in the past and in the future. The explanation should be accomplished by visually showing the documents it is similar to, how similar they are, and by what phrases or topics they form their similarity.

Reflections

The most revealing experiments about the lab were the micro-vignettes into the lives of the individuals. The summarized representations were at times found faulty because of the way the information was grouped, or the definitiveness of the answer. This is in alignment with the original principles that a complete truth is not made up of a single grand narrative, but is the construction of many small narratives. All the attempts to expose a grand narrative at the lab were confronted with trepidation, distrust, and misguided conclusions. That is not to say they were without value, the highest level pictures were helpful, but they needed to be constructed in a way that affords their interactive deconstruction upon request. For all higher level

representations or summaries, if there is no method to drill down to expose the raw information it is difficult to trust and interpret the representation. The more abstract the representation becomes the more beautiful and effective it is at drawing people in, but often the less useful it becomes at revealing meaningful insight. Substance is found in the words and ideas.

While the original intention was to do static representation because it lowered the number of variables in evaluating the techniques and questions. Almost every approach, would have been easier and more intuitive if made interactive. Interactivity affords the key function of drilling down to expose why something is the way it is in the representation. This is a key point for community data portraiture, because by definition it is a constructed representation from all the small artifacts and individuals grouped together in various forms. Since the observers are familiar with the underlying pieces, but have often not thought of how to combine them into a collective representation, elucidating that construction is pivotal.

As such all visualizations should be designed with the expectation of having to explain themselves. When one sees a connection in the data or an artifact in the representation, interaction should allow the exposure of the underlying data at that point. The purpose of doing so is either to dig deeper or expose mistakes in the construction. Such a requirement both increases engagement and trust in the conclusions being inferred by the visual form.

The visualizations that worked best to expose and represent the community were the ones that followed the visualization principles of being intimate and data-as-interface. In many ways this follows logically from the goals, because it is trying to improve the emotional connection to the community, while simultaneously allowing for reflection and seeing the past at a different scale, there needs to be a level of trust in the representation, and a level of nostalgia or emotional binding from this data to the community. Trust is created by exposing raw data and text, and emotional binding is formed by seeing snippets of real individuals reflected in their own words.

A community data portrait is a bottom-up construction of a community from its artifacts (the digital side-effects of its members). By reflecting together on our behavior from a new shared perspective -- seeing ourselves though the eyes of a community data portrait artist -- a community can form a new vocabulary with which to discuss, relate, and steer its collective actions going forward.

Bibliography

1. Dunbar, R., The social brain hypothesis. Evolutionary Anthropology: Issues, News, and Reviews, 1998. 6(5): p. 178-190.
2. Donath, J., et al., Data portraits, in Leonardo. 2010.
3. Dawkins, R. Queerer than we can suppose. 2005 [cited 2010; Available from: http://www.ted.com/talks/lang/eng/richard_dawkins_on_our_queer_universe.html.
4. Flyvbjerg, B., Making social science matter: Why social inquiry fails and how it can succeed again, in books.google.com. 2001.
5. McMillan, D. and D. Chavis, Sense of community: A definition and theory. Journal of Community Psychology, 1986. 14(1): p. 6-23.
6. McCandless, D., The Visual Miscellaneum. 2009: Collins Design.
7. Oliva, A. and A. Torralba, Building the gist of a scene: The role of global image features in recognition. Progress in brain research, 2006. 155: p. 23-36.
8. Yarbus, A., Eye movements during perception of complex objects. Eye movements and vision, 1967. 7: p. 171-196.
9. Donath, J. Visual Who: Animating the affinities and activities of an electronic community. 1995: ACM.
10. Matthew Bloch, A.C., Jo Craven McGinty and Kevin Quealy. A Peek Into Netflix Queues. 2010 [cited 2010; Available from: http://www.nytimes.com/interactive/2010/01/10/nyregion/20100110-netflix-map.html.
11. Donath, J. and A. Dragulescu, Data portraits: aesthetics and algorithms. 2009.
12. Salton, G. and C. Buckley, Term-weighting approaches in automatic text retrieval* 1. Information processing & management, 1988. 24(5): p. 513-523.
13. Zinman, A. Personas. 2009; Available from: http://personas.media.mit.edu/.
14. Liu, H., P. Maes, and G. Davenport, Unraveling the taste fabric of social networks. International Journal on Semantic Web and Information Systems, 2006. 2(1): p. 42-71.
15. Priestley, J., A chart of biography. London. BL, 1765. 611.
16. Viégas, F., M. Wattenberg, and K. Dave. Studying cooperation and conflict between authors with history flow visualizations. 2004: ACM.
17. Vi égas, F., S. Golder, and J. Donath. Visualizing email content: portraying relationships from conversational histories. 2006: ACM.
18. Fry, B., On the Origin of Species: The Preservation of Favoured Traces. 2009.
19. Garofalo, R. The Genealogy of Pop/Rock Music. Available from: http://reebee.net/chart/poster/.
20. Hirsch, J., An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences, 2005. 102(46): p. 16569.
21. Leydesdorff, L. and I. Rafols, A global map of science based on the ISI subject categories. Journal of the American Society for Information Science and Technology, 2009. 60(2): p. 348-362.
22. Bergstrom, C., Eigenfactor. College & Research Libraries News, 2007. 68(5): p. 314.
23. Feltron, N. 2009 Annual Report. 2009; Available from: http://feltron.com.
24. Assogba, Y. and J. Donath. Mycrocosm: Visual Microblogging. 2008: IEEE Computer Society.

25. Brennan Moore, M.V.K., Christina Xu. Poyozo. Available from: http://mypoyozo.com/.
26. Van Kleek, M., et al. Eyebrowse: real-time web activity sharing and visualization: ACM.
27. Alavi, M. and D. Leidner, Review: Knowledge management and knowledge management systems: Conceptual foundations and research issues. MIS quarterly, 2001: p. 107-136.
28. Suchman, L., Making work visible. Communications of the ACM, 1995. 38(9).
29. Viegas, F., et al., Manyeyes: a site for visualization at internet scale. IEEE Transactions on Visualization and Computer Graphics, 2007: p. 1121-1128.
30. Heer, J., F. ViÈgas, and M. Wattenberg, Voyagers and voyeurs: Supporting asynchronous collaborative visualization. Communications of the ACM, 2009. 52(1): p. 87-97.
31. Harris, J. We Feel Fine. Available from: http://wefeelfine.org/.
32. Hansen, M. and B. Rubin. Listening post: Giving voice to online communication. 2002.
33. Blei, D., A. Ng, and M. Jordan, Latent dirichlet allocation. The Journal of Machine Learning Research, 2003. 3: p. 993-1022.
34. Klages, M., Literary theory: a guide for the perplexed. 2006. p. 184.
35. Ben-Kiki, O., C. Evans, and B. Ingerson, YAML Ain't Markup Language (YAMLô) Version 1.1. yaml. org, Tech. Rep, 2005.
36. Neo4j. Available from: http://neo4j.org/.
37. mongoDB. Available from: http://www.mongodb.org/.
38. Apache. Solr. Available from: http://lucene.apache.org/solr/.
39. Freebase Gridworks. Available from: http://code.google.com/p/freebase-gridworks/.
40. Fry, B. and C. Reas, Processing. Online document], 2007, Available HTTP: http://www. processing. org.
41. McCallum, A., Mallet: A machine learning for language toolkit. 2002.
42. Byron, L. and M. Wattenberg, Stacked graphs - geometry & aesthetics. IEEE Transactions on Visualization and Computer Graphics, 2008: p. 1245-1252.
43. Blei, D., et al., Hierarchical topic models and the nested Chinese restaurant process. Advances in neural information processing systems, 2004. 16: p. 106.
44. Fruchterman, T. and E. Reingold, Graph drawing by force-directed placement. Software: Practice and Experience, 1991. 21(11): p. 1129-1164.
45. Hartigan, J. and M. Wong, A k-means clustering algorithm. JR Stat. Soc. Ser. C-Appl. Stat, 1979. 28: p. 100ñ108.

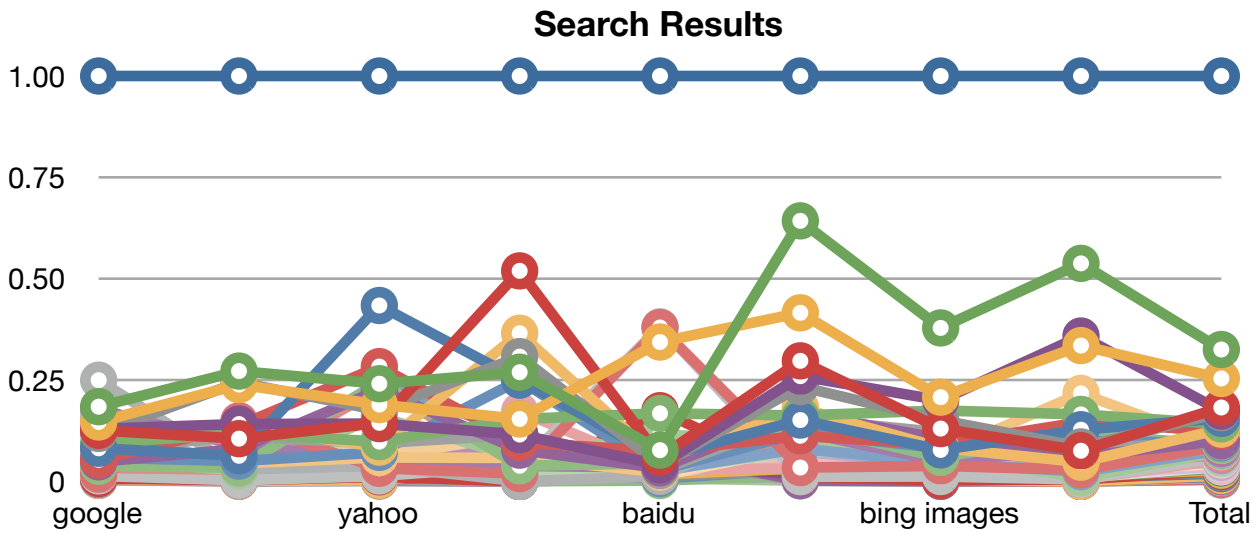
Appendix

Visualization Fundamentals

Web Popularity Results

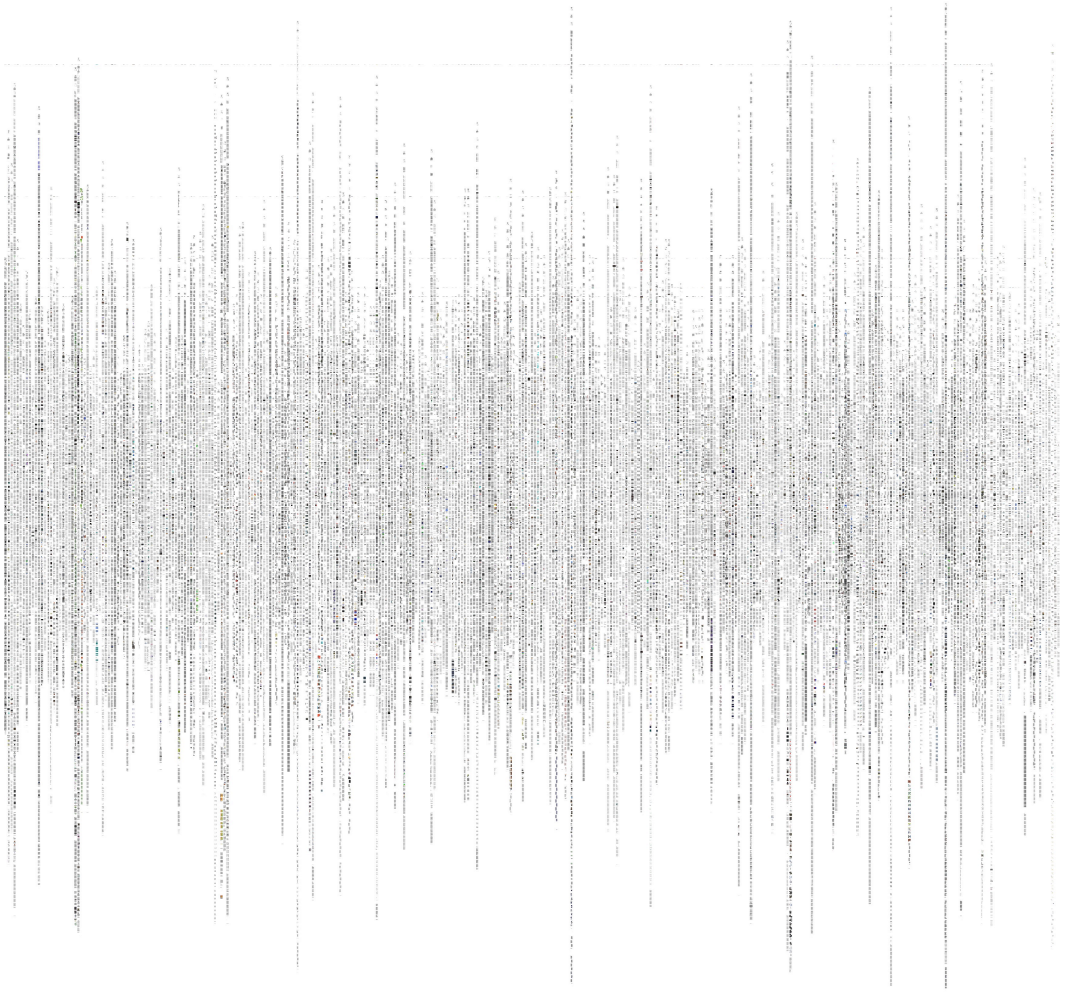
Name	Web-Index
Nicholas Negroponte	1
John Maeda	0.325235743480197
Pattie Maes	0.25379391323571
Seymour Papert	0.182123712324354
Hiroshi Ishii	0.173305833105389
Mitchel Resnick	0.169452867250559
Walter Bender	0.1554063611363
Tod Machover	0.1416638429995 96
Marvin Minsky	0.128693522078208
Frank Moss	0.124161128947195
Hugh Herr	0.0999852470227655
Cynthia Breazeal	0.0978682452322007
Judith Donath	0.0947065299125633
Deb Roy	0.0890151408400933
Ramesh Raskar	0.088723755962696
David Small	0.0872810563354772
Ed Boyden	0.0808202535303567
Leah Buechley	0.0779225139221703
Rosalind Picard	0.0770025978118295
William Mitchell	0.0739149508832005
Neil Gershenfeld	0.0733098796533513
Henry Lieberman	0.068052597978088
Dan Ariely	0.0637875169546258
Michael Best	0.0595658291127442
Michael Bove	0.0592354395959537
Chris Csikszentmihalyi	0.055214950478279
Glorianna Davenport	0.0531349054035535
Jerome Wiesner	0.053042234191178
Alex Pentland	0.0468486648269203
Joe Paradiso	0.0454707605205915
Henry Holtzman	0.0449740377406909
Ted Selker	0.0446729594414179
Muriel Cooper	0.0442952972332638
David Cavallo	0.0405581918947668
Barry Vercoe	0.0363614938031418
Joseph Jacobson	0.0355100264617719
Michael Hawley	0.0342349845454786

Name	Web-Index
David Reed	0.0324283127499 87
Justine Cassell	0.0275179008812615
Brian Smith	0.0267373582608269
Kent Larson	0.0239388114999188
Chris Schmandt	0.022281317693775
Bruce Blumberg	0.0192633910246884
Andy Lippman	0.01923670781648
Richard Leacock	0.0144469104985471
Stephen Benton	0.0140421021155983
Richard Bolt	0.0125806525647992
Ron MacNeil	0.0117419454356123
Aaron Bobick	0.0111895495704832
Bakhtiar Mikhak	0.010619450430835
Whitman Richards	0.00989068176301753
Ken Haase	0.00578460096610394
Edward Adelson	0.00485808311041919
William Schreiber	0.00484715667166277
Ike Chuang	0.00401977681932285
David Zeltzer	0.00286676857855782

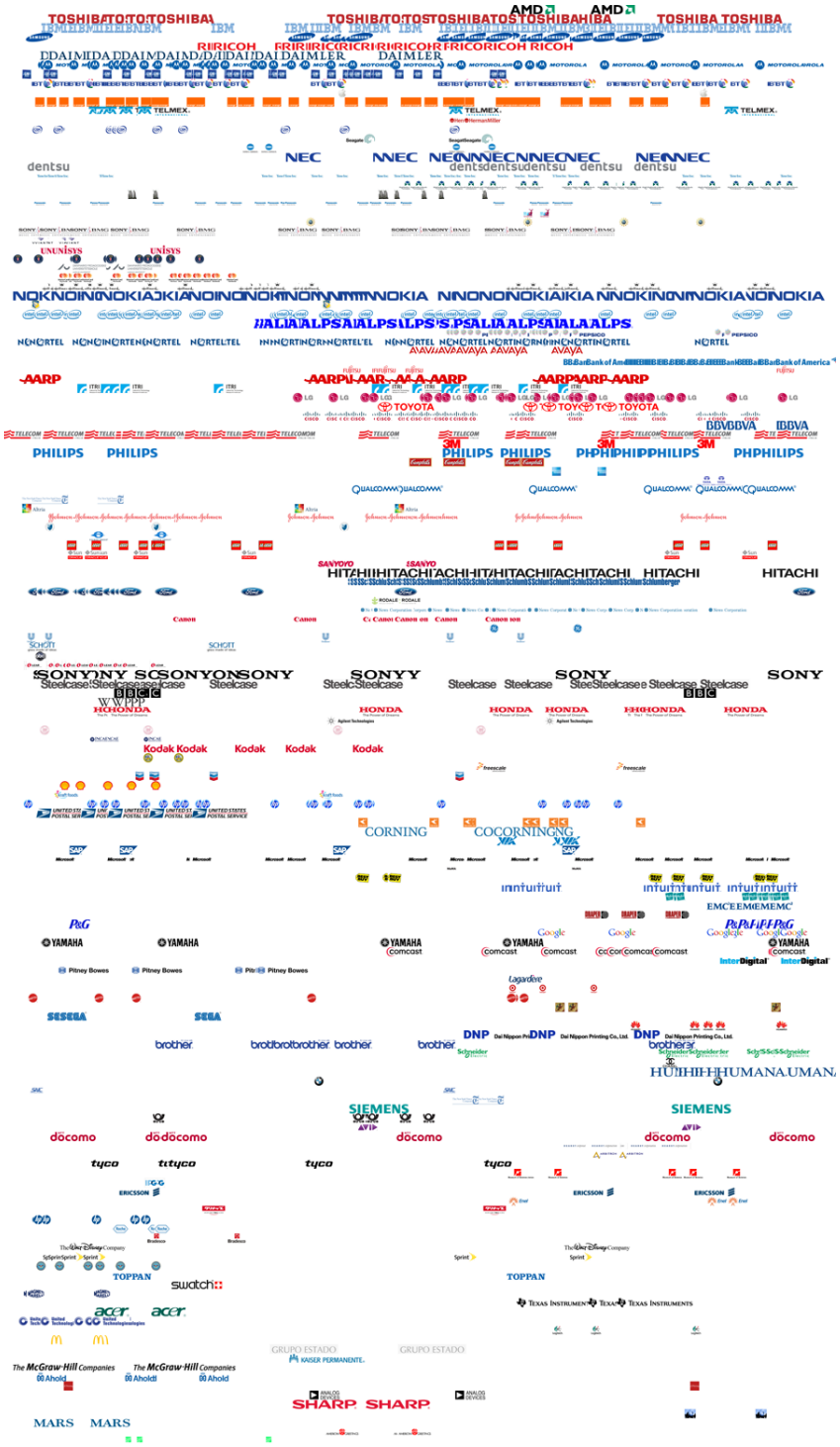


- Nicholas Negroponte
- Pattie Maes
- Hiroshi Ishii
- Walter Bender
- Marvin Minsky
- Hugh Herr
- Judith Donath
- Ramesh Raskar
- Ed Boyden
- Rosalind Picard
- Neil Gershenfeld
- Dan Ariely
- Michael Bove
- Glorianna Davenport
- Alex Pentland
- Henry Holtzman
- Muriel Cooper
- Barry Vercoe
- Michael Hawley
- Justine Cassell
- Kent Larson
- Bruce Blumberg
- Richard Leacock
- Richard Bolt
- Aaron Bobick
- Whitman Richards
- Edward Adelson
- Ike Chuang
- John Maeda
- Seymour Papert
- Mitchel Resnick
- Tod Machover
- Frank Moss
- Cynthia Breazeal
- Deb Roy
- David Small
- Leah Buechley
- William Mitchell
- Henry Lieberman
- Michael Best
- Chris Csikszentmihalyi
- Jerome Wiesner
- Joe Paradiso
- Ted Selker
- David Cavallo
- Joseph Jacobson
- David Reed
- Brian Smith
- Chris Schmandt
- Andy Lippman
- Stephen Benton
- Ron MacNeil
- Bakhtiar Mikhak
- Ken Haase
- William Schreiber
- David Zeltzer

What do all the theses look like?



Sponsors as context



Prediction/Reflection Slope

Year	Slope	Authors	Title
1992			
	-0.11554859	Barry Arons	A Review of The Cocktail Party Effect
	-0.00539726	J. Paradiso	Reaction Wheel Energy Storage
1993			
	-0.14068996	Judith S. Donath	Identity and deception in the virtual community
	-0.04474251	J. Paradiso	Synchronous Proximity Detection for Stretched Wire
1994			
	-0.16014421	Justine Cassell	Modeling the interaction between speech and
	0.0281118891	J. Moy	Please refer to the current edition of the Internet
1995			
	-0.12977521	Tinsley A.	Multi-Level Direction of Autonomous Creatures for
	-0.02437431	Lisa J. Stifelman	Feedback Generation in Audio Interfaces
1996			
	-0.133379186	Justine Cassell	A FRAMEWORK FOR GESTURE GENERATION and
	-0.01484	Shawn E. Burke,	HIGH-RESOLUTION PIEZOPOLYMER ACOUSTIC
1997			
	-0.14610427	Kimiko Ryokai,	Fantasy Play and Storytelling
	0.09936436	Joseph A.	New Technologies for Monitoring the Precision
1998			
	-0.11799843	L. Campbell, K.	Requirements for an Architecture for Embodied
	0.007060671	Bernd Schoner	Phoneme Discrimination from MEG Data
1999			
	-0.117363613	Push Singh	The Public Acquisition of Commonsense
	0.045104646	Alex Pentland	Attentional Objects for Visual Context
2000			
	-0.27573726	Kely Dobson	AgoraPhone
	0.084217247	Eric D. Scheirer	Synthetic and SNHC Audio in MPEG-4
2001			
	-0.143861191	Robert	It's about Time: Temporal Representations for
	0.11359579	J. Cassell, T.	More than just a pretty face: conversational
2002			
	-0.12322447	Axel Kilian	Challenge: SWING FOR TRANSPORT OF MATERIALS
	0.1378075171	Nitin Sawhney	Thesis Study: Collaborative Design and Learning in
2003			
	-0.02119799	Ivan Sergeyevich	Spatial Aspects of Mobile Ad Hoc Collaboration

Year	Slope	Authors	Title
	0.134990439	Nitin Sawhney	Cooperative Innovation in the Commons:
2004			
	-0.00921288	Mathew	Parasitic Mobility in Dynamically Distributed Sensor
	0.108113647	Push Singh,	OMCSNet: A Practical Commonsense Reasoning
2005			
	0.024865431	L. Bonanni, Chia-	Smart Sinks: Real World Opportunities for Context-
	0.229993451	David Gatenby	Galatea: Personalized Interaction with Augmented
2006			
	0.073988957	H. Chung, Chia-	Lover's Cups: Drinking Interfaces as New
	0.360287738	Scott G. Vercoe	Moodtrack: Practical Methods for Assembling
2007			
	0.0754814	Cati Vaucelle,	Touch Sensitive Apparel
	0.139900246	Amber Frid-	Leave Any Noise At the Signal: Participation Art
2008			
	-0.03882517	Robert Speer,	AnalogySpace: Reducing the Dimensionality of
	0.111826219	Mariana Cristina	Television meets Facebook: Social Networking

What is the Longest Question in Each Thesis?

2002

"What are the primitives (the native commands that the program is built out of, such as repeat)?" --csmith-ms

"How can we tailor content to the visitor in a museum, during his/her visit, to enrich both the learning and entertaining experience ?" --flavia-phd

"What are the creative implications of connecting instruments in a network?" --roberto-ms

"How can these views of integration versus self-determination being applied from a technology assimilation perspective?" --mmonroy-ms

"Is the speech interface active?" --yli-ms

"What limitations exist, then, for creating a commercial remote touch communication device?" --anjchang-ms

"What do you want to do next?" --earroyo-ms

"Is it necessary to perform such extensive offline analysis to produce efficient implementations?" --deva-ms

"How can we show that the unitary correction can preserve quantum states?" --brecht-ms

"Is the postural behavior that occurs in a face-to-face conversation similar to that occurring in a human computer interaction?" --atenea-ms

"Is there a compelling context for exploring emergent systems and, if so, what will be the theoretical framework defining our mode of exploration?" --lifton-ms

"Why is it that when a person is shown a picture of a flower girl dressed in white, she assumes it is at a wedding, and she wonders who is getting married?" --hugo-meng

"What kinds of social resources are exchanged between the people in the user's personal social network?" --alockerd-ms

"How often do you pick up ideas for new listening or media in general (films, radio, events etc) from the people in the above list?" --surj-ms

"What if she can get access to this information by realizing of these similarities in research areas through a visual representation?" --pinto-ms

"However, the previous studies do not answer some basic questions; how do these signals interact and coordinate with each other in grounding?" --yukiko-ms

"So how many nuclei will be polarized by the magnet?" --jasont-ms

"However, the main problem with these algorithms has always been the tuning of the dynamic programming parameters – what should the costs be for the individual pitch candidates and the transitions?" --sbasu-phd

"What were the conditions and procedures of the evaluation?" --stouffs-ms

"Is testimonial more effective as a mechanism of storytelling than stories that use their visual quality to show the other people, places, and events of the community?" --ramesh-ms

"Which interactions are best suited to tangible user interfaces; and which are better served by graphical user interfaces or other approaches?" --ullmer-phd

"When voice communication modality should change, under what network conditions should this occur, and under what circumstances does the user prefer to use each of the communication modes?" --marcoe-ms

"Why was one interaction pattern deemed more efficient or intuitive?" --rahulb-ms

"How do you begin to address the volumetric types of representations that vegetation starts to address?" --benpiper-ms

"So why don't we just make a cheaper tag?" --fletcher-phd

"Who's that standing over there?" --tang-meng

"What characteristics of the machinic embodiment of a remote person connecting to a certain public sphere will maximize feelings of comfort in the person communicating, and simultaneously maximize tendencies toward respect and empathy for this person by the people communicating from the public site of this physical tele-presence sculpture?" --monster-ms

"How would the workshop have been different if, instead of working in the open all of the time, the children could have had privacy to work on their poetry or expressions until they were ready to share their work?" --anindita-ms

"How about the buildings at this college? … any that were unusual or of special significance because of something that happened in or close to that building?" --mhadis-ms

"What will rise out of the noise and have real bearing on my baseline?" --hirzel-ms

"How often does the individual maintain distinct relationships between groups of people?" --danah-ms

"How then is it possible to acquire certain literacy skills through interaction with media other than text?" --cati-ms

"What if there are many dots in the same space that has the same property compilations, would you still have a similar response to how it is moving in relation to others?" --hyun-ms

"What changed in students' understanding, as they developed their story representation from a verb, to a written text, to a storyboard to an edited video?" --nbreyer-ms

"What kinds of internal and external representations are necessary for computational entities to form social relationships like those formed by animals?" --badger-phd

"What are they saying?" --raffik-meng

"How perfect would it be if one could throw a ball over a net, and have it come out somewhere else with the same characteristics, i.e. spin, speed and direction, all in real-time?" --florian-ms
"Why should the artisan take advice of a person that does not create handcraft?" --dkor-ms
"How then, can the site present information that is useful to every possible visitor?" --shanec-meng
"How could the user specify a motive, if the computer was allowed to ignore it?" --egon-ms
"Why not profit from the experience that the students have in living in abandoned areas of the city to discuss, for example, the pollution of the rivers and the low quality of life of the population, and how the big trash deposits endanger the health of people?" --paulo-ms

2003

"So, how can the polyketide synthase accomplish the incredible feat of assembling complex structures with extreme speed and accuracy?" --bchow-ms
"What does this error message mean?", and other questions that you might reasonably ask a knowledgeable human assistant?" --ewagner-ms
"How do we ensure that there is sufficient awareness and dialogue among stakeholders, domain experts, researchers and fieldorganizations to make the design process participatory, the emerging concepts open to peerreview, and the outcomes sustainable and accessible to all?" --nitin-phd
"When one participant reports doing 30 actions in one day and another participant reports doing 5, is this a real difference in level of activity or does it reflect different conceptualizations of what is reportable?" --jbeaudin-ms
"So these buttons, the grey ones, what's the difference do you think between these and the ones that you made?" --andrew_s-ms
"How well does the searching algorithm match torque fields to torque functions of arbitrary input motions?" --mkg-ms
"What are the compelling reasons someone would want to go to the trouble of actually carrying out the work outlined in this document?" --mankins-ms
"How can personal mobile tools augment local interaction and promote spontaneous collaboration between users in proximity?" --chardin-ms
"What is the minimum number of reflections necessary to explain a given observed data set?" --matt-phd
"Is it the fact that one is physically present, while the other seems to be remote and shown only on a screen?" --coryk-ms
"Where do we go from here?" --niloy-ms
"So, having answered the question "can we simulate?" in the affirmative, the question thus became "can we do better?" --geva-ms
"Is equally important in identifying other arrhythmic episodes?" --du-ms
"So how can we leverage an animator's knowledge of expressive animation that is implicitly contained in animation examples?" --aries-phd
"How are such systems and methods capable of expanding a user's bank of knowledge as well as broadening their sense of power and understanding?" --megan-ms
"However, again, prediction is impossible unless we have a notion of the similarity between the scenes we are attempting to predict amongst because otherwise every event looks new and unique. > 100 yrs. ?" --clarkson-phd
"What does learners' engagement in real-time programming begin to reveal about the intellectual substance of real-time programming, and of programming in general?" --ch-phd
"What we can do is take all the information that's there and let end users solve the problem of, 'how to do we make this relevant?" --ryantxu-ms
"Is it possible that children are faced with new kinds of misconceptions that are not related to the way things work in the "real world" but to the potential that each of the objects around them has?" --margo-ms
"What will the future look like?" --pangaro-ms
"How can algorithms that work for one home occupant be extended to multi-user households with intertwined activities?" --emunguia-ms
"How much of their hesitation and looking to me for authority grew from a recognition that the project represented something foreign to them but familiar to me?" --hlubinka-ms
"When is switching an effective means of taking a break or getting relevant information and when is it a distraction?" --sylvan-ms
"Is its priority high enough compared to the setup knobs configuration to pass the feedback threshold?" --taly-ms
"How then, can we use the variables that we can sense to constitute a more complex understanding?" --robotnik-ms
"What is the dominant focus of attention of the person in the instant they will use the appliance, and how will this condition their capacity to interact with it?" --mr4-ms
"How do you make sure the building is fully sprinklered, as an example, because the sprinkler system will change based upon how you divide the space?" --tmcleish-ms
"What's the difference?" --georgina-ms
"How then do we dismabiguate the correct referent object(s) from all others present in the visual scene?" --sheel-ms
"How do you make it smaller?" --osc-ms

"Why is this different from giving each product a passive tag and letting an active reader look up information in a database?" --simong-ms
"However, what if we could introduce some other interactions that can remove this degeneracy?" --murali-ms
"Is the application capable of holding the user's attention for at least 12 minutes to complete the exercise?" --vadim-phd
"What defines the unique sound of a singer?" --moo-phd
"How can you design experiences where the outcome may not be predictable, that maintains a high level of direction?" --strickon-phd
"How can a large group of anonymous individuals be given appropriate feedback, such that each individual has a sense of close control over the central interaction, while ensuring adequate structure so that all participants find the interaction pleasing?" --geppetto-ms
"When a turn is about to finish, the current speaker can explicitly hand it over to the next speaker by asking a direct question, ending with a tag question (such as "right?" --hannes-phd
"What" – what are the musical parameters and interdependent algorithms that can be utilized in the network, filling the architectural form with musical content?" --gili-phd
"What does it mean in the long term to apply lessons from animal development to the design of synthetic creatures?" --mattb-ms

2004

"How does it impact understanding?" --orenz-ms
"How far should we take this new power we are developing to mould other creatures - not to mention ourselves to suit our plans or whims?" --saoirse-ms
"How appropriate did you feel that the connection was between the types of manipulations possible with the instrument and the effects that could be controlled?" --dmerrill-ms
"Is this correlation an artifact of the specific community that we were looking at, or is it a more general effect which holds for different communities?" --tanzeem-phd
"How appropriate do you think it is analyze authors' past messages to estimate the average emotional tone of the messages?" --ethanlp-ms
"What kinds of communication between the mobile devices and the participant are effective, so that the participant is able to 1) know critical story content and context and 2) still have enough curiosity and motivation to continue a mobile cinema story?" --ppk-phd
"Who are the people you would be willing to share this type of information with?" --nmarmas-phd
"So how does this instinctive behavior pattern come to be so reliably encoded in the duckling, to the extent that it is ready to be perfectly expressed virtually from the moment of hatching?" --dlyons-ms
"What is the soundscape of your kitchen at breakfast time?" --hugosg-ms
"How does a computer know which other computer it is talking to?" --raffik-ms
"How does he take this continuous image sequence and correctly divide it into two gestures (rather than one or four or ten)?" --daphna-ms
"How do you remain aware of someone's activities and availability without encroaching on their privacy and personal space?" --vidya-ms
"What would have been the effect in the chosen final strategy if subjects were told to reduce body oscillations instead of asking them to change the placement of their feet?" --lwelti-mas
"However, what happens if a child has recorded a motion that lasts longer than the original recording?" --hayes-ms
"How do we know a neural network is doing what it's supposed to?" --fry-phd
"What is the role of imagination in online dating?" --atf-ms
"How stressed overall in comparison to other days (in general) did you feel today?" --kkliu-ms
"Is there another, more controlled way to observe the generative process, and more specifically through what methods might we best gain deeper insight into how an artist perceives structure in the artifacts he generates?" --nyssim-phd
"How does a tool remain open-ended in an effort to support exploration yet provide the necessary components for the development of specific creative works?" --msw-ms
"How would you describe this image so that you can retrieve it sometime in the future?" --whisper-ms
"Is there an appropriate compromise in physical form which maximizes the flexibility of the surface, while remaining consistent with the method of creating an on-screen parametric model?" --amanda-ms
"What mix of video and audio is useful for small groups doing remote design work?" --kkarahal-phd
"What is the minimum amount of information, and hence the number of unique part types, and the amount of state, or information within each of those parts, to specify a desired structure or type of structural behaviour?" --saul-phd
"Is it orthoscopic?" --shill-ms
"How do search engines work and how do topic maps allow for better searching?" --ouko-ms
"Which of these variables do you want to take as your independent design variables, and how do you want to propagate dependencies up or down through the tree?" --rchin-ms
"What areas does it focus on: education, politics, culture, sports, community?" --carlagm-ms
"What would have been the effect in the chosen final strategy if subjects were told to reduce body oscillations instead of asking them to change the placement of their feet?" --lwelti-ms

2005

"What can we accomplish with a given number of switches?" --manup-ms
"When can augmented reality and ambient interfaces improve the usability of a physical environment?" --jackylee-ms
"Some of this trace manifests itself in the questions commonly asked of published systems — does the system scale? is it robust? ese questions might be taken to be "can a person reasonably add to the system without it collapsing?", "is it possible to debug?", "how does it fail?" --marcd-phd
"How can we use these collections of devices for public good and personal reward?" --bcd-ms
"What more accurate representation of their learning could the students have produced?" --savalai-phd
"What can we learn about the relationship between language and power through the process of spriting—would dialect and issues of power be addressed more often in schools?" --tara-phd
"When you have a problem configuring your home appliances, what strategies do you use to solve the problem?" --jhe-ms
"Is the intelligence in the feature or the pattern recognition?" --bwhitman-phd
"What does it mean to describe a sound mathematically or programmatically in a computer?" --harrison-ms
"How are these conceptions of 'public' and 'private' space evolving today, as we use new communication technologies to weave our private social practices within public spaces?" --lilys-ms
"Where is the research being done on implanting electrodes into monkey brains and training them to control a robot arm?" --dagg-ms
"How would you begin building a way to control car animations on a computer screen?" --millner-ms
"What model were participants using to come up with a list of people with whom they felt they were exchanging the highest number of messages?" --fviegas-phd
"When do you eat dinner?" --pkaushik-ms
"What factors do you think are important in choosing a topic?" --anmol-ms
"What are the benefits of using a distributed architecture rather than a centralized one?" --arnaud-ms
"How do we organize our digital libraries?" --nfields-ms
"Did the annotator mean for the annotation to be interesting to others, or only to himself?" --golder-ms
"How long you have been using the service?" --ashwani-ms
"How much of this diffusion happens by contagion, and how much of it is a product of external forces (such as the mass media)?" --cameron-phd
"When online dating, if you had 10 minutes free were you more likely to call a friend or check mail from potential dates?" --frost-phd
"Did manufacturing and patent agreements between companies violate antitrust laws?" --jcooley-ms
"Why are nonverbal cues are better than just using digitized voice, like with voice chips (cheap, ubiquitous speech synthesis chips integrated into products to give them autonomous voice)?" --stefan-phd
"What process were these groups using to share information in a first task that either worked or didn't work, what happened when they watched the replay and strategized for a second task, and then what happened during the second task?" --joanie-phd
"What is the synchronization usually implemented?" --sunx-ms
"What architectures will allow the development of applications in this larger environment?" --ca.rocha-ms
"Did you make it into the next room?" --pgorniak-phd
"What happens when such services are not available (for example in indoor environments)?" --aggelos-phd
"So it's useful to have your bag keep track of things you carry with you everyday, but what about things you only need from time to time?" --nanda-ms
"How do we define ourselves and our legacy?" --maku-ms
"Did such experience of looking around their classroom influence the way they looked at the colors, objects, and materials in their environment?" --kimiko-phd
"How do the results of those who were in the control condition (which was constructed to reduce opportunities to cheat) compare with those who reported the highest scores from the other conditions?" --carsonr-phd
"How does a suggestion help the videographer assemble the pieces of her narrative puzzle and collect new pieces that increase the narrative possibility of her collection?" --barbara-phd
"How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?" --assaf-ms
"How would you group these four images?" --tristan-phd
"What are the repetitive properties of music at different levels, which are organized in a hierarchical way?" --chaiwei-phd
"What if the lights could be designed to work in typical interior spaces like offices and living rooms by having 'softer' transitions, better sensing and behaviors that can be useful?" --amerigo-ms

"How do people appropriate a media table into this existing physical environment and how does this platform support natural social interactions within a shared space?" --mazalek-phd
"Is it possible to predict outcome based on a small sampling at the beginning of a conversation?" --rcaneel-ms
"Why lug around excessive storage, seating and accessories, when they are only used a small percentage commuting?" --wlark-ms
"How effectively can memory problems be addressed via information-retrieval techniques applied to a personal-data archive?" --vemuri-phd
"What hint could help you know that you're making good decisions towards reaching that goal?" --erikb-phd
"Is this the extent that touch should exist in robotics?" --wdstiehl-ms
"Why focus on girls, especially when girls already show some abilities that are better than boys when it comes to certain skills such as recognizing emotions?" --sbdaily-ms
"How is it possible that these excerpts were rated with comparable similarity values to those generated with some of our methods, particularly methods 1 and 3 with components modeled jointly?" --vadan-ms
"What conditions promote the emergence of a centralized pattern of ties?" --nathan-phd

2006

"What was this person thinking about that person when these data were gen- erated?" --neptune-ms
"How important is it that you feel the presence or attention of the others during a conversation (e.g., eye contact)?" --francis-ms
"Which then would be the appropriate combination of code distribution and management technique for our network?" --jerry_bp-ms
"What role does the relationship between the movements of each individual and those of the rest of the group play in determining the character of a performance?" --aylward-ms
"How do we design collective systems that are always alive and can intensify to accomplish a goal and de-intensify to digest and relax and have loose activities until another intensification?" --arikan-ms
"However, such a scenario forces us to consider the reverse situation– what if a passageway to the resource node was available, but network communication between two active transducers at each end of the passageway was blocked?" --constans-phd
"What application of technology could address the need for rural energy?" --asun-ms
"How plausibly does this artwork communicate the thoughts|cultural notion|imagery|free intuition|feelings you had of this text?" --hugo-phd
"Is it natural for people to perform the gesture?" --enrico-ms
"How can we make mobile devices aware of the face-to-face interactions that continuously happen all around us?" --jgips-ms
"What if the application wants to draw one big, long tail, showing everywhere that the cursor has been?" --dhirsh-ms
"How much and in what ways?" --xercyn-ms
"Is it advantageous to combine multiple modalities?" --ash-phd
"What can be done to manage attention to task in a group environment with less behavioral facilitation than traditional education and activity groups for this population?" --adamb-ms
"What structural aspects of music determine or influence the acceptability of performances?" --mary-phd
"How many advisors would do this?" --jaewoo-ms
"How much are we changing the original networking stack?" --hyz-ms
"How is that assembling a large collection of changing components into a system results in something that is an altogether stable collection of parts?" --ribeiro-ms
"Is it at this moment that the child lives in a separate play room, a closet where she can store, for example, her books separately from the parents' books?" --susanne-ms
"Is this just the original clip?" --scottyv-ms
"What is the role of the machine and what is it that designers have to do to create a successful learning system?" --alockerd-phd
"Why then are we so insistent that computers do it for us?" --knorton-ms
"How can we find the shared common sense?" --ence-ms
"How many minutes would you say this activity took from the time you first moved a disk until now?" --win-phd
"When the monkey made sounds, did the sounds clarify what it was expressing? i.e. were they more helpful or just disturbing?" --rachelk-ms
"How long and how often do you think you would listen to something like this?" --nvawter-ms
"What is the optimal strategy for managing congestion on the internet while maintaining user satisfaction [63]?" --brecht-phd
"Sowa is right, and being precise, correct and unambiguous is not the fundamental answer, then what is?" --eslick-ms
"Is there a computer connecting the image on the screen and my own expressions?" --orit-ms
"How can we model the situation that two persons agree in one thing but disagree in another, and what computational performance does the extension incur?" --wdong-ms

"How can a set of distributed digital manipulatives create interfaces with new interactive and material properties?" --vleclerc-ms

"Is it possible to create maps that preserve this sense of place and successfully communicate it to the user?" --hock-ms

"However, because accessible data is free and plentiful in our networked culture, what is to prevent users from downloading or adopting hundreds of images at a time to their bag for the day?" --cml-ms

"How many times did you encounter the red screen telling you that your had watched for the number of minutes specified?" --nawyn-ms

"How do we reform the currently prevalent ideas about human interaction and reinvent browsing, advancing past the limitations present in current interactive systems?" --pauln-phd

"Did we lose all of our other guys?" --stefie10-ms

"What specifically did you dislike with the user interface, and why did you find it confusing?" --faaborg-ms

"Where do you wish to travel to?" --ayah-ms

"The answer is bound by the goals of the user: does the user want a glimpse of all response chains to a message, or are they more interested in skimming across entire threads? " --azinman-ms

"When the code is dispersed, does the user know in which order to read the code?" --sajid-ms

"Social networking since it satisfies part of his curiosities while driving, such as what other friends have traversed this same road?" --pliang-ms

"How may the people simultaneously known as women and as scientists - an oxymoronic social subject only beginning to break down - intervene in the construction of the potent natural-technical objects of knowledge called females?" --gemma-ms

"What happens if in order to solve the problem at at least one node, the global group membership requirements need to be changed?" --mallett-phd

2007

"How much access to they have and how familiar they are with modern technologies such as computers, cell phones, cameras, etc.?" --leob-phd

"Whence comes it by that vast store which the busy and boundless fancy of man has painted on it with an almost endless variety?" --orenz-phd

"How useful do you think this description would be if it were available to anyone/everyone in your phonebook (at your discretion)?" --matta-ms

"Is it a parallel system that 'copies over' values from simulation to perception, is it additory, does it compete with sensory activation?" --guy-phd

"How should these differences be compared and combined in a global similarity metric?" --grindlay-ms

"How can a system find story segments whose textual descriptions are related to the input and that are potentially interesting to the users, based on 1) common sense knowledge that is available to the system, 2) the videos' textual descriptions, and 3) the user-contributed information about the interrelations between the videos?" --edward-ms

"Is there any situation where you might remove the comments?" --andresmh-ms

"Where do actions come from?" --eepness-phd

"Is it fair that someone hosts a program while its reputation is being built and then someone else is able to come and use the program and host it with the reputation it has acquired?" --durga-ms

"Did a shock occur?" --mattxmal-meng

"What types of plans should we consider: from the ambitious and vague (i.e., "entertain guests") to common minutiae (i.e., "pick up the cellphone")?" --dustin-ms

"How does the physical form of the connectibles determine how they can be physically arranged, and how does this inform what meaning the arrangements might acquire?" --jeevan-ms

"What shall we teach our pants?" --aisen-ms

"Which of the applications supported the creation of your favorite stories?" --aisling-phd

"So considering you have been constructing story in a time-lapse form, over the past three days, do you think that if you had more time, and if you actually had experienced twenty-four hours, would you think that your time-lapse would be just about six minutes to eight minutes or do you think it would be much longer?" --hyun-phd

"How are you today?" --jorkin-ms

"How can new media contribute to the role of architecture, and extend its meaning, so that we can conceive new kinds of public media that contribute to the social, cultural and political meaning of places?" --orkan-ms

"What social cues are present?" --alea-ms

"Is any network of advice useful for this kind of prediction, or is it sensible to certain network properties, transivity for example?" --barahona-phd

"What object attributes determine canonical views?" --robbel-ms

"How much time is added to the transcription task to handle each of these error types?" --bcroy-ms

"Why should we build musical instruments when so much beautiful music can be made by the instruments we already have?" --revrev-ms

"What makes for good application-led research in ubiquitous computing?" --dskim116-meng

"However to at first select actions to be added to the scene, the problem is much more difficult: how can the human teacher, armed with incomplete knowledge of the robot's internal

representations and abilities, refer to a specific action among all that the robot might know how to perform?" --zoz-phd

"What other information would a visitor need when traveling to a new place?" --calla-phd

"What is your musical experience (trained musician, avid listener, background listener, etc.)?" --meyers-ms

"How can two beings that might look at the world through different sensors, use different ways to "break down" the world, and have different vocabularies / perceptual categories, be able to negotiate meaning among them, and effectively communicate?" --nmav-phd

"Why have everyone document how to drill a hole repetitively?" --wormulus-ms

"What do you think about offering a little training in spotting bad construction to -- sixth graders?" --jteng-ms

"How can we compose a synthesized space suggestive of a physical space that we cannot see and that never existed?" --amber-ms

"How do the main hypotheses that drove this work fare with regards to evaluation results?" --mihir-ms

"How complicated can the loss be before the laser doesn't work?" --jasont-phd

"How does construction of reflective artifacts enhance the ways people reflect on their learning processes?" --rnc-phd

"What would you prefer for yourself when going out for an evening: to feel that you have watched an interesting person, or that you have met an interesting person with whom you have had a good time?" --raphael-ms

"Is it alive?" --davidb-ms

"What is the dominant state of mind, what does the facial expression, and posture convey, fearfulness, surprise, exploration, playfulness, amazement, amusement?" --pluto-ms

"Why do you think that, in your own impression or point of view, why do you think using non-speech sounds might enable someone to be closer to, or experience some emotional aspect of their lives in away that they would be able to with speech?" --nknouf-ms

"What do you think are the merits and the drawbacks of the approach to social awareness outlined in the previous discussion?" --isha-ms

"How could a wearable computing system maintain and re-enforce social networks in these hostile environments?" --awhiton-ms

"What happens when you stop damping?" --roberto-phd

"What would it sound like if every one of your songs, on every one of your albums was playing all the time?" --jdg-ms

"How will people access, search, and sort the copious real-time and stored data available from wireless sensor networks?" --lifton-phd

"Is the design successful as a new kind of musical instrument?" --pliam-ms

"How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?" --earroyo-phd

"Is there a parallel between the way people think of balancing actions as discrete sets of isolated rules and the way people also often perceive education as a discrete process that can be segregated and compartmentalized and move learners from one state to the next in a linear fashion?" --arnans-phd

"What happens when we wish to collaborate, to pioneer new ideas and movements, and to mutually capitalize on one another's strengths, outside the boundaries of real space and real time, and beyond the scope of any single nation's law?" --bpf-ms

2008

"How much energy per bit is required for inter-particle communications?" --ara-ms

"Did you find any of the information or overviews provided by the phone labeling tool particularly confusing or lacking in information?" --rockinso-ms

"What could be gained by expanding the scale of the system from handheld objects to whole body sized arrays?" --bttaylor-ms

"Is paperless really more?" --pranav-ms

"How can we go about providing the users of these tools, as well 24

as their educators, with an easily digestible summary of the creative landscape made possible through their use?" --buza-ms

"However, once we do have a finite set of representative pitch classes, how do we determine which pitch class in a melody belongs to which one of these three major levels in the pitchstability hierarchy?" --huangcza-ms

"Which methods, if any, have you tried/are you using in order to reconcile travel-related problems (both regarding technological and social aspects), e.g. story telling before trip, video calls, maps, remote story telling over the phone, gifts?" --paulina-ms

"However observers feel about the actual work, their actions raise an interesting question – who owns public spaces?" --mud-ms

"So why are holograms so simple and holovideo displays so complex (and expensive)?" --tesla-ms

"What happens when this computer becomes something that is experienced as a social, lifelike entity that shares their physical space?" --coryk-phd

"How was the artifact perceived by the social movements in which it participated, and by the broader culture?" --tad-phd

"What should be done about users who simply re-upload other people's work with limited or no changes?" --sylvan-phd
"How useful are the studies above to explain the gender gap in the technology workplace?" --rusti-ms
"How to communicate the fact that reading a document influences it and at the same time protect the privacy of its users?" --dietmar-ms
"Is the value of a medical product established solely through clinical trial results, or do people in the real world assign additional value to it as a function of heuristics of prices, brands, and other informational signals?" --rwaber-ms
"How can a tangible interface retain the immediacy and emotional engagement of "record and play" and incorporate a mechanism for real time and direct modulation of behavior during program execution?" --hayes-phd
"When – and how – can a cellular automaton be rewritten as a lattice gas?" --ddal-ms
"So in such a natural language programming environment, the question arises—what kind of expectation should the interface give to the user?" --moin-ms
"Is this something you consider true?" --jalonso-ms
"Why is it that, free of the natural laws of the physical world, so much of virtual world design is concerned with recreating familiar physical spaces?" --dharry-ms
"Is it consistent enough across all players to allow for generalization to a single generic player?" --mtl-ms
"What technological innovations exist at the nexus of chemistry, physics, materials science, biology, and engineering?" --bchow-phd
"How does the performance of relatively simple classifiers amenable for real-time performance compare to more complex state-of-the-art classification algorithms?" --emunguia-phd
"When the child says "cup," does he use it as a request for something to drink, or as a statement that he sees a cup?" --decamp-ms
"What is the importance of designing new technologies within a clearly defined framework or constraint system if technology just progresses along it's own path, one design being replaced by another that does the job better according to simple metrics?" --zig-ms
"Why ?" --manas-ms
"How might this under-reporting of good deeds also contribute to the mean world syndrome?" --alyssa-ms
"What would the internet look like today, if suddenly our internet connection went back down to 1200 baud1 ?" --ypod-ms
"What if someone had a tool in their hands that met most of the design criteria and implemented the ideas represented in the inputs/outputz graph?" --silver-ms
"Is there any redundant information that could be done away with in the app?" --maiki-ms
"How are we seasonably aware of what is happening around us?" --starsu-ms
"However, a critical question remains: can a robot, using a simple learning algorithm and paying attention to the visual perspective of the teacher, exhibit the differences in rule choice observed in human learners?" --mattb-phd
"Did you try to reuse musical bricks 7 out of 10 created by others in your composition?" --wuhsi-ms
"How much of this process is lost when only the virtual domain is observed?" --black-ms

2009

"How important would it be to be able to review the meeting notes generated at the table on your laptop or phone after the meeting?" --hunters-ms
"How satisfied are you with the financial information provided?" --khkim-ms
"Which of our mental capacities and human experiences will we choose to amplify and enhance?" --ericr-ms
"How can we orchestrate configuration and express a task and interaction – in a meaningful way?" --gauthier-ms
"What system parameters can be adjusted to meet the privacy requests for users working/living in a ubiquitous sensor network environment?" --nanwei-ms
"What if your mobile phone could automatically detect what's around you, memorize what objects you have used throughout the day, and even warn you if your keys are left behind?" --snlee-ms
"How are you feeling today?" --joon-ms
"What is the relationship between the pleasure we get out of (or think we will get out of) a purchase and the decision to make the purchase itself?" --cfd-ms
"How to select stimuli for environmental sound research and where to find them?" --rmorris-ms
"How comfortable did the other person seem to be feeling (p during this activity?" --siggi-ms
"How the knobs, or gains, change as a result?" --mtf-ms
"What parameters of a multi-tier musical structure can and must vary to create a piece of music that conveys a certain quality of feeling?" --patorpey-ms
"Is there a way which designing and creating kinetic interactions could be as simple and direct as drawing with a pencil and paper or sculpting static objects in clay?" --amanda-phd
"How many times in this week that you expected follow-up interaction with others after you shared things with them?" --dorilin-ms
"Soliciting the patient's agenda: have we improved?" --jom-ms

"How do changes in a robot's appearance and behavior alter its persuasiveness and how it may be perceived?" --mikeys-ms
"When should two mobile devices be allowed to communicate?" --dmerrill-phd
"How can we deal with the even looser associations
between developers in different projects than those present in typical open source development?" --yannick-ms
"Why are tangible user interfaces still predominantly confined to the lab, even after 20 years of compelling research?" --kumpf-ms
"Which factor do you think is more important on positioning the views: the geometric position between images or the remote situation you are trying to follow?" --alsantos-ms