Using Visual Communication Design
To Optimize Exploration of Large Text-Mining Datasets

by

John Joseph Montague

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Arts

Humanities Computing

University of Alberta

Abstract

How can the principles and concepts applied by visual communication designers be used to assist in exploring and understanding the massive, complex volumes of data now available to Digital Humanities researchers? One method we might employ to help us more easily comprehend the implications of these large data sets is visualization. While potentially helpful, visualizations can be more effective still if they are constructed in such a way as to allow or even encourage the viewer to interact with them, exploring the data looking for patterns that might lead to new insight. Visual communication designers try to strike a balance between "clarity" and "immediacy"; clarity, meaning the reader's ability to identify and recognize communicative elements like text and symbols, and immediacy meaning the ease with which the reader can understand the message. This same two-part formula can be applied to visualizations of Big Data relationships.

Almost by definition, immediacy (understanding) will be at a premium when displaying massive amounts of complex data. With especially complex new forms of Big Data, "message" might simply involve determining "where more specialized researchers might want to look more closely." Greater viewer engagement will invariably help promote exploration, more exploration will bring with it a greater likelihood of new insights or questions, and these insights/questions might suggest new research directions in which traditional scholars may want to focus their attentions.

This optimized engagement will principally be the result of two things; the purposeful application of visual communication and perception theory, and the creation of opportunities for dynamic interactivity between the user and the data, which will also provide the means of

exploration. Dynamic interaction affords a more likely path to successful visualization as defined by Petre & Green, that is, visualization "that makes accessible the particular information the user needs" (56). Visual communication designers are always asking themselves what the user needs and crafting their designs accordingly; data visualization design should involve this same consideration.

Table of Contents

List of Figures

Introduction

> "Data visualization" is a wide term that applies to visual representations that attempt to help viewers better understand data. These visualizations help viewers to more easily see and interpret patterns and trends in the data. Being able to see these trends allows people to more quickly articulate their ideas and thoughts.
>
> - Russel Cooke -

We live in an era of burgeoning data, a deluge made all the greater by both the automated generation of much of this data, as well as the convenience with which it can be accessed. Easily communicating with people living on the other side of planet is something we have taken for granted for some time, but social media now allows us to monitor or create the latest news about anywhere in the world while almost anywhere in the world, using our smartphones, the powerful personal computers we carry with us almost everywhere. From financial transactions and healthcare records, to emails and tweets, to the digitization of all the world's literature, the volume of digital data continues to grow at an amazing rate, and will for the foreseeable future.

The principles of composition and perception as employed by both artists and designers, can and should be used to aid in the communication of the relationships in large data sets. Visualizations of large data in the humanities can and should be studied and optimized to encourage greater reader engagement and promote more useful interactive exploration of the data sets, and to help broader audiences explore and understand complicated data.

As put forth by Drucker ("Humanities Approaches"; emphasis original), "The polemic I set forth here outlines several basic principles on which to proceed differently by suggesting that what is needed *is not a set of applications to display humanities "data" but a new approach that uses humanities principles to constitute capta and its display.*" I believe Drucker to be only

partially correct. We certainly would benefit from a new approach to the display of capta, that is, a more effective way to display the knowledge gained from discussion, debate, analysis and interpretation, such that very complex issues are made more clear. However, I argue that we would also benefit, perhaps more so, from addressing the efficacy of our approach to visualization of data for interpretation.

The explosion of the World Wide Web has certainly been impactful, but the changes it heralded have been primarily behavioural. Big Data holds the promise to impact the world more like the advent of the written word, or the creation of science; it promises to add a new chapter to the book of human understanding. "Big Data", explains Cukier et al. (28), "is about more than just communication: the idea is that we can learn from a large body of information things that we could not comprehend when we used only smaller amounts."

With so much data being generated, the sheer volume of information available for research and analysis is awe inspiring. In order to make this Big Data useful, it is important that we give due thought to how we interact with and present this information to those with an interest in using and studying it. It is insufficient simply to collect all the data the world has to offer without also developing efficient and useful ways with which we might examine, analyze and otherwise exploit this data. This thesis will examine the question of how the tools and principles used by visual communication designers might be applied to the interpretation and communication of data gleaned from large text-mining research projects, as well as why it is important that they be employed to do so. The paper is organized as follows:

Chapter One, *Evolution of Communication Media*, presents a brief history and discussion of the genesis of communication media, from the earliest cave art approximately forty thousand years

ago, to today. It highlights the role of abstraction in accommodating the growing communication needs of our expanding civilizations, and argues that the complexities we now face as a result of the speed and volume of information available to us, are such that entirely new ways of examining and analyzing information are necessary.

Chapter Two, *Big Data & The Future*, examines the challenges and opportunities offered by Big Data. We have the opportunity to learn new things when we study something at a scale at which we have never before been capable. With the sheer amount of data produced however, and the speed at which it is produced, its transmission in real-time, and the endless ways in which it may or may not be structured, our current methodologies for envisioning, exploring and analyzing data can be overwhelmed. Traditional close study of human digestible quantities of data needs to be supplemented with new inter-disciplinary paradigms that allow for the exploration of much larger volumes of data, in search of areas for closer inspection. One valuable potential inclusion in these new methods is visualization.

Chapter Three, *Visualization*, defines visualization, and outlines the benefits of including visual communication design as an integral component of integrated interdisciplinary scholarly work, and makes a case for the improvement of visual communication design specifically in Digital Humanities research work. It then outlines some of the challenges, opportunities, strengths and weaknesses of visualizing data, and explains why principles of design and theories of visual communication can and should be utilized to help overcome the challenges. The chapter ends with a presentation of the three researchers whose work forms the basis of the categorization and evaluation of contemporary text-mining visualizations in Chapter Five.

Chapter 4, *Design & Perception*, begins with a discussion of the nature of visualizations as visual metaphors representing abstract data, and how general principles of visual communication design can be applied to assist in their creation and evaluation. It goes on to explain several pertinent theories of visual communication, including Semiotics, Gestalt and Colour Theory, and suggest how those theories might be applied to assist in clarity and viewer engagement.

Chapter Five, *Contemporary Text Mining Visualizations*, reports on a survey of recent visualizations created by researchers involved in automated textual analysis. Some of these visualizations present research results, others represent the visual output of exploratory tools which are themselves research results. Each is explained, categorized by role, and evaluated via a discussion of its use of the elements and principles of visual communication design and according its adherence to the meta-principles presented in Chapter Three.

Chapter Six, *Case Studies*, presents two text-mining visualization tools developed with a team of researchers during my studies at the University of Alberta. The goal and functionality of each is described, along with the benefits conferred by the application of various principles of visual communication design.

1.0 Evolution of Communication Media

> "To know where we are today, and indeed, where we are going, we need to understand where we as a discipline came from"
>
> - Graham, Milligan & Weingart -

1.1 The Origin of Media

The earliest evidence of human mark making known exists preserved very deep in ancient caves. Cave paintings found in various sites throughout Europe, as well as more recent finds in Indonesia, have been dated as far back as approximately 40,000 years by testing the age of small clusters of mineral bumps laid down close to the the paintings (Aubert et al. 224). Since their discovery, academic debate has been vigorous surrounding the nature of the paintings. Scholars remain uncertain whether they were decorative, or spiritual, or instructive, some combination of these three, or for some other purpose altogether.

What is certain is that these cave paintings are the earliest examples of humans recording information that for whatever reason they deemed important. They therefore arguably also represent the earliest appearance of what we would term visualization. These visualizations are an abstract representation of information, although the level of abstraction at work is speculative; does the figure of an animal represent one animal of the kind depicted, or does it represent the concept hunting? This ability to conceive, record, share and understand abstract information is one of the things that, as far as we know, separates us from other animal species, and the abstraction technology that has been most responsible for allowing us to share our knowledge across generations, is writing.

Ancient cave artists were members of hunter-gatherer groups whose subsistence activities likely required little in the way of record keeping beyond perhaps to say “we were here”, or “this is what we hunt.” As humans began to settle in larger, more agrarian societies, our need to communicate and record more complex information about things like commerce, law and spirituality grew. To do that, we required a more robust means of abstraction than simply figurative depictions.

1.2 The Need for Change

The historical record suggests we have been using abstract text in one form or another to record and access meaningful data since approximately 8000 BCE, when ancient inhabitants of Mesopotamia began inscribing clay tablets with information about their agricultural and manufactured goods (Wikipedia). As the volume and complexity of our recorded data grew, we developed new technologies for storage, organization and retrieval. Clay tablets gave way to lighter, more portable papyrus or vellum scrolls, but these were physically limited, and “it was these physical limitations—the length of the papyrus roll and the number of rolls that could be stored together—that tended to define the divisions of literature” (Grout). This was overcome with the introduction of the bound codex, which led to the book as we now know it. The clay inscriptions of 8000 years ago evolved into the more complex cuneiform logo phonetic writings of the Sumerians, Mesopotamians and Assyrians. These gave way to true alphabets, with separate, fully abstract symbols for all consonants and vowels, like the Latin alphabet used in English today.

Each of these new technologies, physical or abstract representational, developed as a response to newly identified needs and wants of users, and each made our information storage and retrieval more efficient. "As models of reading shifted from contemplation to humanistic study, a new graphical organization was required. If a text was to be searchable, so that points could be called out of it for an argument, it needed navigational features" (Drucker, "Humanities Tools" 1). At points along the way, we developed rules for writing left to right and top to bottom (in English, at any rate), we embraced punctuation (for the most part), and we started using both upper and lower case letters. We invented headings, and chapters, and alphabetical order, and even page numbers, because they helped us organize, understand, and find the information we wanted.

1.3 Complex Problems & Interdisciplinarity

As the nature of our lexical requirements became more complex we required new and innovative ways to envision and communicate them. In the intervening millennia between our first visual recordings and the present, we have relied on our brainpower and often lengthy, tedious labour to analyze, organize and retrieve our information. We developed methods and technologies to assist us in our cataloguing, analysis and retrieval of information, but none of these technologies has had the impact, both in challenge and promise, that we are witness to in the current digital age, and Humanities Computing is uniquely well positioned to address the complexities inherent to interpreting the massive and growing data now available to us.

"Complex problems require the perspectives of different disciplines to solve for them" (Mendel & Yeager 1). As an interdisciplinary field of study, Digital Humanities sometimes

draws on the expertise of all manner of humanist and scientific scholars, programmers, writers, and designers. This is important, argues Repko, because "integration addresses the challenges of complexity" (Repko 4). Davidson and Goldberg suggest that the nature of the issues faced by today's world are of a complexity such that integrational interdisciplinarity is the key to overcoming these problems, stating not only that with integration "each field is richer and deeper because of the others", but also that "new interdisciplinary paradigms help us uncover whole new areas and objects of study" (Davidson & Goldberg 2). This final point is a strong argument in favour of the potential role of exploration via visualization in identifying new and important information and areas ripe for close study.

Canadian Research Chair in Digital Humanities at Brock University, Dr. John Bonnet tells us, regarding Canadian cultural, political and communications thinker Harold Adams Innis' work, "To survive, Innis believed that humans have to think consciously about their instruments and signs of communication. They have to change them, play with them, complexify them, even violate them if they are to adapt to new challenges and environments" (Bonnet 6). Developing emergent disciplinary paradigms, convergent fields, hybrid methodologies and new publication models brings complex challenges, but also the opportunity, as Bonnet deems necessary, to "change them", to "play with them", to create more and varied tools, providing researchers with more options in their tool selection and almost invariably inspiring a wider variety of research directions.

More than ever, interdisciplinarity in pursuit of new knowledge and solutions to ever more complex problems can benefit from the particular insights offered by the humanities. As we continue to search for new knowledge within large data sets of growing complexity, one of the tools we can and should use to help us unlock the insights offered by the humanities, is visualization.

“By visualizing information, we turn it into a landscape that you can explore with your eyes, a sort of information map. And when you’re lost in information, an information map is kind of useful” (McCandless).

1.4 Summary - Evolution of Media

Humans have been recording abstract representations of information for around 40,000 years. As our societies evolved, becoming more complex and generating ever more information to catalogue, organize, retrieve and communicate, so to have our technologies for doing so. We are now at a point where the volume and complexity of the data we create is so great, and changes so rapidly, that our traditional modes of study are no longer sufficient. New multi-disciplinary methods of analysis offer the promise of identifying new and exciting areas of study, and visualization is a tool that may help point us in interesting directions.

2.0 The Era of Big Data

> "The greatest hope for renewing our shared theoretical traditions in humanities research, and perhaps the only possible route, is to use massive stores of data digitally."
>
> - Benjamin M. Schmidt -

2.1 The Challenge of Big Data

The first wave of Humanities Computing work, inspired by Father Roberto Busa in the late 1940s, focused on the analysis of existing text collections via the creation of such things as concordances and indices. The second wave, from around the mid-1990s, brought with it a name change to Digital Humanities and opened the door for new multi-disciplinary paradigms, for discussion surrounding what it means to be gathering and studying digital data and how those processes affect traditional humanities study, and for the development of digital environments for the collection, organization and study of this new born-digital data. The third wave of Digital Humanities, now upon us, is the era of Big Data.

John Smith (20) said, "Humanists have always been explorers. They sail not on seas of water but on seas of color, sound, and, most especially, words." In his acceptance lecture for the inaugural award bearing his name, Robert Busa (5) quoted Plautus;

> "'Tis as you'd turn a stream upon your field; which if you do not, it will all run waste into the sea.' I use those verses as a metaphor: in informatics, the oceans include data banks, WWW, the Internet (i.e. information retrieval informatics) and, second, the informatics of writing, printing, and publishing, in which multimedia are included. The small stream is what we are doing, which I summarize as computerized text analysis, or language hermeneutics, i.e. interpretation, i.e.all our ways of questioning the whys of language."

Furthering Smith's earlier analogy, we are adrift in a sea of data, and Busa felt we were at risk of floundering if we failed to develop new methods and approaches, such as data visualizations, to effectively manage and interpret the deluge of data pouring over the horizon.

2.2 What Makes Data BIG?

> "The information produced and consumed by humankind used to vanish – that was the norm, the default. The sights, the sounds, the songs, the spoken word just melted away. Marks on stone, parchment, and paper were the special case… Now expectations have inverted. Everything may be recorded and preserved, at least potentially."
>
> - James Gleick -

There is ongoing debate surrounding what does or does not constitute Big Data, although one lay definition proposed by Graham et al. defines it simply as "information that requires computational intervention to make new sense of it." One thing all the definitions have in common is that as the name suggests, Big Data is vast, often so massive and complex as to make gaining a meaningful understanding of especially large, multidimensional data sets next to impossible using traditional close study of the information contained therein. The relationships between the variables are simply too complex and numerous to conceive and internalize. In an attempt to "define the proportional dimensions and challenges specific to Big Data", Meta Group analyst Doug Laney identifies Big Data as being comprised of "the three Vs": high volume, high velocity, and high variety ("3D Data Management").

Volume

As digital technologies become more powerful and more commonplace, data is being produced and stored at an almost unfathomable rate. Dragland reports that ninety percent of the all the world's recorded data, most of it naturally digital, unstructured, and thus in a state almost impossible to meaningfully examine without computational intervention, was generated in just the last two years (Dragland).

More traditional texts are being digitized every day, and born digital data is outstripping that by far. 2011 saw the creation of 1.8 zettabytes of digital information - 1.8 trillion gigabytes. From 2006 to 2011, the amount of data in the digital universe increased by a factor of nine. (Gantz & Reinsel 1). With such a massive proliferation of information, we need new tools to help organize, search, and understand the enormous volumes now available to us.

Velocity

Velocity is the speed at which data is created, transmitted and stored, and in many cases, analysed and subsequently visualized. In the past, it was not uncommon for databases to update only every day, or even every several days, as computing power was such that significant time was required to process what we would now consider only modest amounts of data. Data now is not only created in real-time or near real-time, the preponderance of online devices, both wireless and wired, allows machines to transmit data the moment it comes into being. "'Nowcasting,' the ability to estimate metrics such as consumer confidence, immediately, something which previously could only be done retrospectively, is becoming more extensively used, adding considerable power to

prediction. Similarly, the high frequency of data allows users to test theories in near real-time and to a level never before possible" (McGuire, Manyika, & Chui)

The speed at which data is being created is truly unprecedented. On average, every minute of every day we create and deliver around 350,000 Tweets (Krikorian), upload more than 300 hours of video to Youtube (Youtube Statistics), and register around 2,500,000 Google queries (Google Search Statistics), and these are only a few of the more well known high velocity online services. "More data cross the internet every second than were stored in the entire internet just 20 years ago" (McAfee et al.), and the challenge we face is coping with the speed at which these volumes of data are being created and used in real time. The nature of graphic images, allowing not serial but parallel interpretation, may help with this bottleneck.

Variety

> As explained by Dwayne Spradlin, CEO of the nonprofit Health Data Consortium, "We know that data is everywhere, and it's growing at an exponential rate. The problem is that this data is everywhere, but it's hidden and packed away, and it's incredibly fragmented"
>
> - Natalie Burg -

As expressed by Laney in his 2001 research paper, "no greater barrier to effective data management will exist than the variety of incompatible data formats, non-aligned data structures, and inconsistent data semantics" (2). Not only are we generating data at previously unrealized rates, but we are outputting and storing that data in a daunting myriad of different and often mutually incompatible forms. This data structure disparity must be overcome so we might more readily and reliably recognize the relationships between significantly divergent data types.

2.3 Why Big Data Matters

In their 2013 paper "The Rise of Big Data: How It's Changing the Way We Think About the World," Kenneth Cukier and Viktor Mayer-Schoenberger explain that the big deal with Big Data is that "we can learn from a large body of information things that we couldn't comprehend when we used only smaller amounts." What though, are those things?

Among the detractors of Big Data and machine reading, a belief is often expressed that the very nature of at least some of the humanities precludes the utility of digital automation in their understanding. Stephen Marche summed up the perceived problem in his 2012 article "Literature is Not Data; Against Digital Humanities" with the emphatic statement "Literature cannot meaningfully be treated as data… Literature is the opposite of data." To me this statement presumes to assign an irrevocably "untouchable" status to some areas in the humanities, and is essentially suggesting that because we don't yet understand all the things that exploring Big Data might tell us, we may as well not even bother investigating where it might lead.

Much has been written in opposition to the notion that Big Data, and machine learning in particular, can provide meaningful insight, especially in rhetorically robust arenas such as the study of literature, or philosophy. If supporters of Big Data were suggesting that algorithmic analysis of the arts should or even could replace the traditional close reading techniques that have been used in the academy for centuries, I could understand the temptation to decry the dangers of this relatively unproven instrument of exploration. This of course is not the case, and as Rob Kitchin explains, "Big Data should complement small data, not replace them" (Carrigan).

Even the most fervent proponents of the possibilities of studying Big Data recognize that

the data itself is best utilized to establish correlation, not causation; the patterns in massive data sets can show us what relationships exist, but not necessarily why they exist. As Podesta et al. tell us;

> "Identifying a pattern doesn't establish whether that pattern is significant. Correlation still doesn't equal causation. Finding a correlation with Big Data techniques may not be an appropriate basis for predicting outcomes or behavior, or rendering judgments on individuals. In Big Data, as with all data, interpretation is always important" (7).

This correlation/causation relationship guarantees that the close reading methods already firmly established across the academy will remain valid and valuable no matter the extent of Big Data utility. The promise being explored is merely an exciting addition to the toolset already in use.

2.4 The Future of Big Data - Internet of Things

Ubiquitous connectivity is the driving force behind what many refer to as the Internet of Things (IOT), or less frequently, the Internet of Everything or the Industrial Internet. "A very large contributor to the ever expanding digital universe is the Internet of Things with sensors all over the world in all devices creating data every second. The era of a trillion sensors is upon us" (Rijmenam). Ever increasing machine to machine communication, made possible with the proliferation of inexpensive sensor technology and mobile, reliable, cloud-based communication networks, promises to change the world at least as radically as did the birth of the World Wide Web, making "everything in our lives, from streetlights to seaports 'smart'" (Burrus). GIS and GPS supplied geolocative information, coupled with low energy RFID tags that indicate the identity of a particular individual or object, and a stream of data from countless sensors will change the way we interact with the world, or perhaps more accurately, the way the world interacts with us.

"In our houses, cars, and factories, we're surrounded by tiny, intelligent devices that capture data about how we live and what we do. Now they are beginning to talk to one another. Soon we'll be able to choreograph them to respond to our needs, solve our problems, even save our lives" (Wasik). Bill Wasik, senior editor at Wired, suggests that the most remarkable thing about this phenomenon, what he terms "the programmable world" is not the sensors, nor that our objects are linked; it's that once we have enough of them linked, all these disparate data gathering and analyzing items will become "a coherent system… a single giant machine" which will change how we imagine our daily lives. Dr. Hermann Kopetz, of the Vienna University of Technology agrees, stating, "The novelty of the IoT is not in the capability of a smart object ... but in the expected size of billions or even trillions of smart objects that bring about technical and societal issues that are related to size" (308). Global analysts project that by 2025 there will be one trillion networked devices worldwide (Wasik), each of them generating and/or collecting, and many of them sharing, data. The impact of all this data on human interaction based fields as diverse as sociology, architecture, history and business will be impressive, as long as we can reasonably analyze and utilize it.

Wasik outlines three stages he has identified, through which we he says must pass before we will realize the true potential of the coming revolution. The first stage involves growing the network, that is, simply increasing the number of sensors, networks and wireless communication enabling tech in the world. We are well on our way to establishing the critical mass necessary for stage one, with prolific smartphone usage providing us with a widely accepted interface for control of sensor embedded objects like home electronics systems and wearable health data monitors.

Stage two, says Wasik, involves "coordinating their [the device's] actions to carry out simple tasks without any human intervention." In its simplest form, think of this as a series of if-then statements embedded into your environment. If you approach within 500 meters of your home, then your air conditioner turns on, for instance. The possibilities are near endless, but the collected data will need to be analyzed, as always, to determine where we can create positive benefit.

The third stage involves overlaying systems programming on top of the smart objects, or as Wasik explains, "not just tying together the behavior of two or more objects—like the sprinkler and the moisture sensor—but creating complex interrelationships that also tie in outside data sources and analytics." Imagine that same sprinkler that is not only tied to a moisture sensor, but that checks current and historical weather patterns to decide when and how much water to distribute, or traffic lights that respond to congestion patterns, accidents and construction to optimize traffic flow.

In addition to radio frequency identification, wireless sensor networks, addressing schemes (to identify specific objects), and data storage and analytics, Gubbi et al. identify visualization as a critical element in the potential success of the IoT (1647), mirroring my argument that engaging visualizations (assisting with immediacy), coupled with clarity of information are necessary for the realization of the potential of Big Data interactions. "For a lay person to fully benefit from the IoT revolution, attractive and easy to understand visualization has to be created. As we move from 2D to 3D screens, more information can be provided in meaningful ways for consumers. This will also enable policy makers to convert data into knowledge, which is critical in fast decision making" (Gubbi et al. 1649).

Though there are efficiencies to be realized with such connectedness, achieving them will be a challenge. There are difficulties arriving at a concensus on standards to make data structured enough to be used this way, as well as difficulties overcoming the objections of some organizations to sharing their data, and with security in general, and besides says Bruce Sterling, the benefits offered by the IoT will mainly be realized by the technology giants who monitor and monetize the data. “Google and Facebook don’t have “users” or “customers.” Instead, they have participants under machine surveillance, whose activities are algorithmically combined within Big Data silos” (Sterling 6).

As Big Data gives way to even bigger data, we will need help sorting through it all to form hypotheses regarding where to look more closely in order to engage in the kinds of analysis that will help us answer questions and identify both problems and solutions. Well designed visualizations can offer that help, both making the data more conceivable from an analysis standpoint, and in the creation of interfaces for the monitoring and control of devices and networks.

2.5 Summary - Big Data

Front page news events like Edward Snowden’s May 2013 revelations disclosing the secret NSA (America’s National Security Agency) collection and analysis of massive amounts of ostensibly private data both domestically and internationally, are now making Big Data analytics part of the public lexicon (Risen). Major corporations, led by IBM with $1.3 billion in Big Data revenue in 2012, are scrambling to adopt practices that will allow them to capitalize on the volume of information now being generated. Global Big Data related revenues in 2012 were $11.6 billion, and are projected to break $18 billion in 2013 (Kelly et al.).

While one might shudder to think what the NSA will do with all that information, it is worth keeping in mind, as suggested by Meier, that "the national security establishment is often in the lead when it comes to initiatives that can also be applied for humanitarian purposes." Outside of the world of politics and espionage, what can researchers and academics do with the volume of data now, or at least soon to be at our disposal? Matthew Jockers asks, "How do we mine them [texts] to find something we don't already know?" (qtd. in Smith)

To utilize such massive corpora and avoid succumbing to the dilemma of what McCormick et al. refer to as "information without interpretation" (4) we need to find the most effective ways for end-users to explore, visualize and interact with Big Data such that it is both meaningful and understandable, if possible by both the trained and untrained eye. I would like to modify Dr. Jockers' question, asking instead more specifically, "How can we better use *visualizations* to find something we don't already know?" How can we use our natural abilities as pattern finding machines to aid in exploratory research?

As suggested by Mayer-Schönberger & Cukier, one of the differences between working with Big Data and working with more traditionally manageable bodies of information, is a tendency to embrace the benefits offered by inexactitude or generality rather than focusing on the benefits offered by precision. Their suggestion that "what we lose in accuracy at the micro level, we gain in insight at the macro level" (14) mirrors the ability of both visualization and abstraction to quickly communicate broad understanding with minimal time and effort.

Big Data analytics and computer-supported visualization offer ways to read collections as our cognitive abilities are stretched to their limit with the sheer volume of data available (Araya 74). A data set requiring visualization might now be so large, and contain so many variables, that

we cannot possibly examine and understand all of it at once. “Scholars increasingly [have begun to] use visualization to look not only at textual content but also spatiotemporal data, non-text artifacts and related metadata over the longue duree” (Zepel). In order to meaningfully display and make sense of the huge volumes of data now being studied and presented, we must often make decisions about what data to show and how, as well as what not to show, and those are tasks perfectly suited to a visual communication designer.

3.0 Visualization

> "Nineteenth century culture was defined by the novel, twentieth century culture was defined by the cinema, and the culture of the 21st century will be defined by the interface."
>
> - Lev Manovich -

In this chapter I argue the necessity of including visual communication design as a key component of integrated interdisciplinary scholarly work, discussing both the potential benefits and shortcomings of data visualization in Digital Humanities.

In the conclusion to his 2006 paper "Is it Now Time to Establish Visualization Science as a Scientific Discipline?", Remo Burkhard explains that, "The benefit of an established discipline [called] Visualization Science [would be] more impact, both in teaching, research, and business" (Burkhard 6). Thomas and Cook second the notion, suggesting the necessity of creating "a science of visual representations based on cognitive and perceptual principles that can be deployed through engineered, reusable components" (6), and further that, "visual representation principles must address all types of data, address scale and information complexity, enable knowledge discovery through information synthesis, and facilitate analytical reasoning" (6). While "science"may not be entirely appropriate as a description, there is certainly benefit to be realized from the careful study and controlled, purposeful application of the principles of visual communication.

3.1 Visualization & Digital Humanities

3.1.1 What is visualization?

> "Visualization offers a method for seeing the unseen. It enriches the process of scientific discovery, and fosters profound and unexpected insights."
>
> - Bruce McCormick -

Card et al. (7) define visualization as, "the use of computer-supported, interactive, visual representations of data to amplify cognition", while Erbacher (623) explains it as "the creation of graphical images for the representation of data through either abstract or physical relationships." Visualizations draw on the nature of human perception to afford abstraction. Abstraction is a tool that Liao et al. tell us, "allows us to learn general patterns with limited training data" (2) and when used effectively, can greatly assist with the rapid understanding of complex information.

While it has long been understood that a well-constructed visualization can assist in understanding complex data, formal study of the mechanisms present is only a fairly recent phenomenon, with modern psychology and research into perception having their origins only in the late nineteenth century (Carlson 18). "A picture is often cited to be worth a thousand words and, for some (but not all) tasks, it is clear that a visual presentation… is dramatically easier to use than is a textual description or a spoken report." (Shneiderman 1)

There's been a good deal of discussion surrounding the importance and place of aesthetic consideration in data visualization, both in Digital Humanities and elsewhere. Some of this literature focuses on whether or not aesthetically pleasing, if often incomprehensible data, can in fact be classified as art (Lang 9), while most focuses on establishing what benefit, if any, visualization

offers. I make no suggestion as to whether or not data visualization can or should be considered art, rather I am suggesting simply that principles of composition and perception as practiced to some extent by both artists and designers, can and should be used to aid in visual analytics, in the communication of the relationships within large data sets.

"Visual analytics is the science of analytical reasoning facilitated by interactive visual interfaces" (Thomas & Cook 4). It is a multidisciplinary field that includes, among other less relevant areas of focus:

> • Visual representations and interaction techniques that take advantage of the human eye's broad bandwidth pathway into the mind to allow users to see, explore, and understand large amounts of information at once
>
> • Data representations and transformations that convert all types of conflicting and dynamic data in ways that support visualization and analysis (4)

3.1.2 Why is visualization important to DH?

> "In today's world of networked communications the digital humanities have a special role to play in helping the humanities reach out."
>
> - Alan Liu -

In 1998, Roberto Busa noted, "...computerized speleology, to retrieve deep roots of human language, is fundamental in all disciplines. At this level, humanities are the prime source and principle for all sciences and technologies" (7). It seems as though those controlling the purse strings of post secondary funding are either not listening, or very short sighted. As evidenced by recent cuts to government funding, post-secondary programs have of late been under assault.

Recently, the Alberta government slashed $147 million dollars from the University of Alberta's funding, and as a result, using a selection method that favours currently profitable programs over those with substantial long term human benefit but less immediate profit potential, the U of A was forced to cut twenty programs in the arts. Even with the news that "Alberta's government is pouring $50-million back into its universities and colleges, a sharp mid-year reversal of course after it cut $147-million from postsecondary education in March's provincial budget" (Bradshaw), those dropped programs remained dropped.

As ongoing debates within DH continue regarding just who is involved and exactly what they are up to, we remain a growing but still relatively unknown, low profile "discipline." Instability in post-secondary funding is likely to continue most strongly impacting programs with low enrollment and thus low "profitability"; there is no petroleum industry snatching up all the DH grads, thereby fueling registration. "The digital cuts across disciplines and perspectives ... This gives a strong [theoretical] incitement for institutions to support the digital humanities" (Svensson). If we want to ensure a measure of stability for the DH program, and for the Humanities in general, we need to find ways to maintain and hopefully increase our profile and enrollment.

By addressing the quality of their visualizations, digital humanities researchers may be able to produce research output that is more widely and easily understood. Engaging, well constructed interactive visualizations can help promote exploratory research of large DH data sets by relevant academic parties not previously aware of or interested in DH. Raising awareness of the potential benefits to be realized from DH research may help digital humanities as a discipline gain more widespread acceptance and understanding among both the general public and the humanities community, and perhaps more importantly, among those in the broader academic community whose decisions will impact funding levels.

3.2 Challenges & Opportunities

> "Graphical competence demands three quite different skills: the substantive, statistical, and artistic. Yet now most graphical work is under the direction of but a single expertise—the artistic. Allowing artist-illustrators to control the design and content of statistical graphics is almost like allowing typographers to control the content, style, and editing of prose. Substantive and quantitative expertise must also participate in the design of data graphics, at least if statistical integrity and graphical sophistication are to be achieved"
>
> - Edward Tufte -

Visualization involves the presentation of data in an often abstract visual manner, such that a user can "get insight into the data, draw conclusions, and directly interact with the data" (Keim 1). Shneiderman suggests that a good mantra for using visual tools in information seeking might be, "overview first, zoom and filter, then details on demand" (1), and this is precisely what well constructed visual data interactions afford, giving the user a feel for the big picture before allowing them to filter the data as they feel most relevant, in search of insights or meaningful patterns.

Static visualizations, some digital and all those created for print, are often relatively spartan, focusing attention on the results through careful design, with all affordances removed. Visualizations generated from large data sets, however, tend to be overly dense and have to include interactive affordances to be useful. As Big Data becomes ever bigger, it becomes increasingly important that we create visualizations allowing the user to examine and interact with the data in multiple ways. Static visualizations are sufficient when their creator wants to exhibit a specific relationship in the data, but not so much when the data is very large, and the multitude of possible relationships therein are very complex.

In these instances, we can benefit greatly from creating interactive tools that enable a user to perform what amounts to exploratory research on large data sets, what McCandless calls

being a "data detective." Exploratory research is that which is conducted prior to the definition of a clear research question, or prior to the clear understanding of the scope of an issue. When so engaged, the viewer is using "high-bandwidth human perceptual and cognitive capabilities to detect patterns and draw inferences from visual form" (Vande Moere 2), potentially developing new and unintentional insights into the data, and perhaps generating a hypothesis to be tested. In short, creating compelling visualizations with excellent user engagement will help to encourage the kind of exploratory research that I believe to be one of the key components in making interactive visualizations effective at uncovering new knowledge.

"Never before in history has data been generated at such high volumes as it is today. Exploring and analyzing the vast volumes of data is becoming increasingly difficult. Information visualization and visual data mining can help to deal with the flood of information" (Keim 1). This flood of information provides us with a new and unique opportunity to develop insights into data of which we have heretofore been ignorant, but the complexity and sheer scale of the information we might now be examining can preclude summarizing it neatly for analysis and presentation. The following section examines some of the ways in which visualizations can both potentially help or hinder.

3.2.1 Strengths

"Only a picture can carry such a volume of data in such a small space."

- Edward Tufte -

Keim (1) summarizes the strength of visualization saying, "The advantage of visual data exploration is that the user is directly involved in the data mining process," allowing our

unique perception abilites to spot patterns that might otherwise remain unseen. “CAPTCHAs ... automated tests to distinguish humans from software robots in an online environment” (Chew & Tygar 1), exploit this ability of ours to more readily recognize patterns in noisy, non-homogenous or unstructured visual data (Keim 1) than can machines by presenting online visitors with a visual identification task only accurately performable by a human.

In his blog, University College Cork’s Paul O’Shea suggests some of the ways in which data visualization might be beneficial, including increased cognition (Card 1), and reduction of mental load (Cawthon & Moere 2). I would add to this list the real-time representation of fast moving data, increased opportunities for user interaction, and increased user engagement, bringing with it the potential for increased research volume and exploration.

“The visual data exploration process,” says Keim (1), “can be seen as a hypothesis generation process: The visualizations of the data allow the user to gain insight into the data and come up with new hypotheses.” Combining the storage, retrieval and display abilities of powerful computers with human perception abilities makes visual data exploration particularly well suited for exploratory research, where research goals are often initially vague and shift automatically as users gain insight into the data.

Frawley describes a promising method of exploration, knowledge discovery, involving the “nontrivial extraction of implicit, previously unknown, and potentially useful information from data” (qtd. in Healey, “Effective Visualization” 13). Healey hypothesizes that knowledge discovery can help the user filter unwanted information, thus reducing the volume of data being visualized, helping ease common display bottlenecks as described by Erbacher (1), and allowing the user to discover and explore previously unknown trends and relationships in the data.

Interactive visualizations can help offset a difficulty identified by Erbacher, arising from the display of especially large volumes of information. “It is unfeasible to assume that even future display capabilities will be able to visually display the volume of data being discussed,” he says, “especially considering the rate of growth of data collection processes in comparison with the growth rate of display technology” (1). By utilizing Shneiderman’s “overview first, zoom and filter” model (1), and displaying data using generalization, as abstract visual metaphors, and by compartmentalizing massive data, such that the viewer only sees portions of it at once, as necessary for their understanding, interactive visualizations can permit exploration of data sets even when our devices might be incapable of displaying the set in its entirety, at maximum granularity.

Keim (1) explains that, “For data mining to be effective, it is important to include the human in the data exploration process and combine the flexibility, creativity, and general knowledge of the human with the enormous storage capacity and the computational power of today’s computers.” Effective visualization tools can help researchers “focus their full cognitive and perceptual capabilities on their analytical processes, while allowing them to apply advanced computational capabilities to augment their discovery process” (Thomas & Cook; 28).

3.2.2 Weaknesses

> “Illustrators too often see their work as an exclusively artistic enterprise .... Those who get ahead are those who beautify data, never mind statistical integrity.”
>
> - Edward Tufte -

While developing a growing presence in the public consciousness, visualizations are not a one size fits all panacea for all Big Data’s analysis and communication needs. Different kinds

of visualizations have different characteristics; there are areas where they offer benefit, and there are areas where their utility is limited or non-existent. Simple representative visualizations, like histograms for instance, are insufficient to display complex interrelationships. Others, such as dendrograms, especially if you are working with huge data sets, quickly become an illegible mass of inter-connectivity. Word clouds are good at showing the relative frequency of words in a text or topic, but not at comparing several texts simultaneously. Network diagrams produce some beautiful results, but suffer from the same difficulties as dendrograms; large data sets quickly lead to illegibility.

A potential problem with using the abstract visual metaphors provided by visualization to describe data is a lack of precision in what's being displayed. There is greater ambiguity in the interpretation of visual metaphors than there is in the reading of text, for instance, leading to a wider spectrum of interpretation, potentially leading to erroneous conclusions, especially given our tendency to find meaningful patterns in meaningless noise, a phenomenon known as apophenia (Wikipedia, "Apophenia"), or as described by Shermer, "patternicity." This has to be weighed however, against the exploration utility of visualizations, especially interactive ones. Visualizations are not designed to provide proof; they generally either illustrate something already discovered, or in the case of those designed for exploration, enable insights that might provide interested parties with a direction in which to explore.

Bresciani and Epplen claim the pitfalls of visualization are "due to the fact that the meaning of symbols and colours are not universal" (quoted in O'Shea, 2012). In his blog, after explaining some of the ways in which data visualization might be beneficial, O'Shea goes on to give several valid examples of cultural perception bias, before accurately pointing out that visual literacy is to

an extent a learned skill that not everyone has. While it certainly is true there are circumstances, cultural or otherwise, that would render much visualization incomprehensible, O'Shea is lamenting visualizations' failure to achieve truly universal access to the meaning of the data they represent. Doing so, providing universally interpretable acces to complex data, would be a significantly more impressive achievement than using visualizations to suggest some sort of correlation that might be worth looking at more closely, and O'Shea's lament is a little akin to complaining that "this treatment for the cold is pretty good, but it would be better if it cured cancer too." It's undeniably true, but it's a lot to expect, and if visualization can help direct research questions, it has to be considered a worthwhile exercise.

The creation of visualizations presents us with many of the same kinds of challenges faced by early innovators working with vellum scrolls as data communication and interaction needs outgrew the existing medium, that is, how best to communicate the important parts of a growing amount of complex information and ensure as much clarity and efficiency as possible. As Healey & Enns suggest, harnessing human vision effectively for data visualization purposes … requires that we construct displays that draw attention to their important parts (12). The first step in this process is to understand what is most important, and the second step is to draw the viewer's attention to it; Principles of visual communication can help us do just that.

3.3 What makes a good visualization

Evaluating a visualization, like evaluating all design, first requires an understanding of its purpose. A visualization, however compelling, cannot be deemed effective unless it performs in

accordance with the role intended by its creator. As with most forms of visual communication, the intended function of a given visualization is likely to have a significant impact on the form it takes. The following three sections summarize the work of researchers that will be used in Chapter Five to evaluate and discuss the works included in a survey of contemporary text-mining visualizations.

3.3.1 MacEachren & Taylor on Roles

Determining the best choice of visualization for a given circumstance is a complex matter, with a number of taxonomies having been developed to help organize the possibilities. MacEachren & Taylor (3) classify cartographic visualizations in accordance with their role, which they suggest is driven by the point in the research process at which the visualization is being produced, "from initial data exploration and hypothesis formulation through to the final presentation of results" (Dibiase 14). I extend their classification of visualizations for cartographic purposes to include those developed and used for more general depiction of abstract spatial relationships in data, suggesting that *all* visualizations are meant to principally perform one of the following tasks

Presentation – These visualizations are meant to present previously discovered information in a way that is understandable to the viewer. As such, clarity/legibility is especially important.

Confirmation/Analysis – These are meant to assist in determining the meaning of a particular group of data elements. Visualizations meant primarily for analysis would benefit most from accuracy in their data representation.

Synthesis – These allow the combination of multiple data sets or streams. Synthesis visualizations are important tools in the identification of interrelationships.

Exploration – Exploration visualizations are meant to help users when they are unsure exactly what they are looking for. It is especially important that exploration based visualizations afford the user opportunities for interaction. This kind of exploratory research is often a starting point, to be followed by careful analysis. Precision in these visualizations is secondary.

It is the last kind of visualization, exploration, in which I am most interested, and on which the visualization tools presented in Chapter Six focus; those that allow the user to not simply understand a predetermined analysis of the data, but to explore it in hopes of uncovering new information.

3.3.2 Bertin on Efficiency

Jacque Bertin's work *Semiology of Graphics: Diagrams, Networks, Maps* explains his semiotic based "Image Theory", and is concerned with the optimal design of information graphics. Bertin defined efficiency as a measure of how easily the viewer of an "image" could glean the intended information at a glance; an image he defined as that part of an information graphic that can be isolated during a moment of perception. He proposed the existence of a number of variables that can be used by designers in the translation of information into visual communication. These he categorized into "planar" and "retinal" variables, with planar variables being the relative position on the X and Y axes, and retinal variables being what are now considered standard tools in a designers visual communication toolbox; colour, texture, shape, orientation, size, and value.

Bertin (149) proposed that efficient visualization is limited to three-components, two planar and one retinal, and while he is not alone in suggesting a limitation to the effective number of components we are capable of processing, no universal consensus exists regarding the specifics

of that limitation. We easily comprehend two planar components represented on spatial axes, along with a third retinal component, but our interpretive ability becomes more challenged once we introduce more components. Several workarounds have been employed to combat this problem, including “dividing and conquering”, or “linking” wherein the viewer is provided with different views of the same data, and more recently, the display of data at varying levels of granularity (Green 3). Bertin’s notion of efficiency is a significant indicator of the level of clarity present in any given visualization, and will be considered in Chapter Five, *Contemporary Text Mining Visualizations*, when evaluating each surveyed visualization’s adherence to Schlatter & Levinson’s core meta principles for working with very complex systems..

3.3.3 Schlatter & Levinson’s Meta-Principles

The challenge for designers in a Big Data display sense, is to determine how best to utilize elements and concepts of visual communication such as scale, white-space, labeling and colour (what Petre & Green would call ‘secondary notation’) to assist the viewer in understanding and gaining insight into complex, multidimensional data. “Multidimensional data visualization involves representation of multidimensional data elements in a low dimensional environment, such as a computer screen or printed media. Traditional visualization techniques are not well suited to solving this problem.” (Healey, “Effective Visualization” 1)

Schlatter & Levinson explain the tendency for complex digital applications (a description I here extend to include interactive visualizations) to either look great or be highly functional, but not both. “What makes these applications challenging to design is also what makes them

compelling to use—real-time access to data… [that] provides highly visual, interactive ways to explore" (xi). They outline three core "meta-principles" (xiv) to combat the difficulty inherent to working in such complex systems:

Consistency – visual language needs to define conventions and use them consistently in order to be first, understandable, and second, re-usable.

Hierarchy – sensible and appropriate hierarchical display of information helps the viewer perceive and interpret the relative importance of elements in a visualization.

Personality – has the potential to create engagement, and builds expectations in users regarding an application's functionality and for whom and what it is intended.

Dastani (601) proposes "a process model for effective data visualization … based on the assumption that effectiveness of visualizations depend on the effective use of the capability of the human visual system to perceive visual structures" and argues that effective data visualization requires a "structural correspondence condition", whereby the relationship between data and the perceptual structures used to represent it is preserved. With this statement, Dastani is explaining the need for the visual representation of structures and data to be appropriate, clear, non-contradictory and consistent, in order to promote optimal viewer understanding, reinforcing the importance of Schlatter & Levinson's notions of 'consistency' and 'hierarchy'. Employing Schlatter & Levinson's meta-principles will help in the creation of interactive, engaging, visualizations, that are clear and understandable despite their reliance on and need to display massive amounts of data, even in real-time.

3.4 Visualization Summary

Representing and decoding large, complex collections of information is inherently difficult. Visualization, the presentation of data as often abstract visual metaphor, allows human pattern recognition to aid in both understanding and knowledge discovery. Visualizations that are automatically generated however, tend to be too dense and/or unstructured to be useful, and compounding this, exhibiting especially large amounts of information often leads to a display bottleneck.

A well constructed, visualization is created to efficiently perform one or more particular functions; presentation, confirmation/analysis, synthesis or exploration, each of which can impact the visualization's design. Effective exploratory visualizations often afford the user interactive opportunities, and can encourage and enable exploration, potentially assisting in the formation of research hypotheses.

4.0 Design & Perception

Visualizations are employed primarily to make complex information either more understandable, more interesting, or both. To do this successfully, they must exploit aspects of how we as humans perceive and interpret visual information. Making design decisions based on an understanding of theories of visual communication can help ensure the maximum performance of any given visualization, regardless of its role.

4.1 Principles of Design

"One of the earliest stages of a traditional design process consists of preparation, investigation and analysis, in which the designer typically locates relevant data and facts to have a better overview of the design problem and to frame possible end solutions" (Vande Moere 3). As designers of Big Data visualizations, we have to become familiar with the data such that we can fathom which relationships are most important; which are most likely to illustrate interesting or enlightening insights.

As explained by Vande Moere, "Abstract data is characterized by its lack of a natural notion of position in space, and its visualization examples include financial models, textual analysis, transaction data, network traffic simulations and digital libraries" (2). When presented with exceptionally large, complex and difficult to comprehend data sets such as these, often derived from what for most readers would be unfathomable algorithms, designers must find a way to make that information legible. Due to the abstract nature of this kind of data, the designer must

determine “how to invent new visual metaphors for presenting information and developing ways to manipulate these metaphors to make sense of the information” (Eick 1).

Visual communication designers often use the partnered terms ‘legibility/readability,’ or alternatively ‘clarity/immediacy,’ to describe two of the fundamental evaluations of visual communication design. Legibility (or clarity) is literally a measure of how easily the viewer can discern the elements that must be read and interpreted when viewing the design. Historically, this refered to the literal legibility of a typeface used in a piece of visual communication, but it is also used to describe the discernability of non-text symbols and elements. Lidwell et al. define legibility as, “the visual clarity of text, generally based on the size, typeface, contrast, textblock, and spacing of the characters used” (124), while Tinker refers to legibility as “factors affecting ease and speed of reading” (4). Illegibility, in the context of a visualization, is often a result of the complexity of the information being presented, its sheer volume, or thoughtless presentation. Viewers can be overwhelmed by the task of trying to sort through a poorly designed visualization to develop even a rudimentary understanding of what is in front of them.

Readability (or immediacy) is, as defined by Lidwell et al., “the degree to which prose can be understood, based on the complexity of words and sentences” (162). Tinker explains it as a means to “measure the level of mental difficulty of reading material” (5). Broadening this explanation to include non-prose visual communication, readability is an evaluation of the extent to which a message being communicated is readily understood. Readability can be improved by employing widely practiced principles of visual communication design, leading to greater viewer engagement, more time spent interacting with complex visualizations, and ultimately, greater understanding and new insights. As explained by Healey and Enns, “an image regarded as interesting or beautiful

can encourage viewers to study it in detail" (10), and any image studied in detail is more likely to reveal its secrets to the viewer.

Functional design needs the right amount of both legibility and readability, and this requirement is one of the key separations between design and art. A lack of either or both is a fundamental weakness observed in much of the data visually presented by non-designer humanists. "Designers", as Vande Moere notes, "are required to create novel forms of information art, that appeal also to user engagement and aesthetics and provoke long and enjoyable usages," rather than to focus solely on task metrics and effectiveness (8). It seems when researchers are presenting their own large data sets the focus is too often on accuracy and completeness, rather than functionality of the visualization. The line and bar graphs, word clouds, and illegible network diagrams reflect, as stated by Liu, "the near-total imaginative poverty of the field in crafting an aesthetics of data" ("The State of the Digital Humanities" 27).

4.2 Theories of Visual Communication

In order to build an effective data visualization, the designer's knowledge of the subject matter, the purpose or goal of the visualization, and the likely users must be coupled with his or her familiarity with human perception and cognition in order to make effective use of the tools of a visual communicator. The following theories of perception and visual communication have the potential to directly impact the efficiency, utility, and communicative effectiveness of visualizations. Their understanding and implementation is important in order to make sound decisions regarding the use of Petre & Green's 'secondary notation'.

4.2.1 Pre-Attentive Processing

Pre-attentive processing, as described by Healey (“Effective Visualization” 18) refers to the ability of our visual sensory system to quickly and easily detect certain visual features, including hue, orientation, intensity and motion, when presented with an image or scene. “The front end of the visual system”, explains Wolfe, “can process vast amounts of information in parallel” (“Guidance of Visual Search” 104), suggesting that we process an incredible amount of information before we have a cognitive understanding of what we are seeing. Healey, Kellog and Enns hypothesize that preattentive processing can be exploited to create visualization tools to assist in data analysis, and that “such tools … should allow users to perform certain types of visual analysis very rapidly and accurately” (“High Speed Visual Estimation” 2).

“Reading textual information is considered a perceptually serial process as the reader must perceptually interpret each character in sequence to interpret a word and subsequently each sentence. A graphical image on the other hand can be interpreted in parallel, allowing a conceptualization of the image to be interpreted essentially instantaneously.” (Erbacher 1)

Understanding how we perceive visual stimuli before consciously interpreting them may help visualization designers more efficiently display large, complex data sets. “To deal with the still-overwhelming excess of input, the visual system has attentional mechanisms for selecting a small subset of possible stimuli for more extensive processing while relegating the rest to only limited analysis.” (Wolfe & Horowitz 1). In exploratory visualizations, this valuable characteristic of human vision helps the user detect patterns, enabling identification of areas that are potentially worth closer study.

4.2.2 Gestalt Theory of Perception

"Many theories of visual perception assume that before attention is allocated within a scene, visual information is parsed according to the Gestalt principles of organization." (Moore & Egeth 339) The Gestalt theory of perception, as refined by Max Wertheimer in 1910, states in essence that the whole is other than the sum of the parts. That is, our perception of what we see is something other than simply an evaluation of the component parts of the scene in front of us. Gestalt theory, which contains the following precepts, suggests that our brains automatically help us make sense of what we see by organizing scenes in some predictable ways;

Figure & Ground - the viewer requires adequate contrast between foreground and background to avoid confusion.

Similarity - items in a scene that are similar enough to one another, relative to other items, will tend to be experienced as a single group.

Proximity - items in a scene that are proximal enough to one another, relative to other items, will tend to be experienced as a group.

Continuation - as much as possible, the brain seeks a smooth, continuing line, and tends to skip over small distractions.

Direction - items on the page that in one way or another point in the same direction, or that have a prominent directional axis, will tend to be experienced as a group, and can focus viewer attention in the direction in which they point.

Gestalt theory holds that what one sees when experiencing a scene is modified in accordance with what one has previously seen, and by what one wants or expects to see. "Experience influences

what ‘readers’ of graphical representations look at and hence what they see, so that readership skills – both perceptual and interpretive … must be learned” (Petre & Green 1). This reinforces the importance of employing consistent visual language across a series of related visualizations or even within a single complex visualization, as called for by Schlatter & Levinson. Failure to do so risks losing the benefits offered by experience.

4.2.3 Semiotics & Sign Theory

Semiotics is the study of signs, or rather, the study of things that represent other things. Almost all signs require some degree of previous knowledge to understand, though some are more intuitive than others. “A primary technique to achieve improved visual communication is to use clear, distinct, consistent visible language. Visible language refers to all the verbal and visual signs that convey meaning to a viewer” (Marcus 2). In semiotic terms, all signs are classified into three areas:

Iconic - signs that are similar to or appear to be resemble what they represent. Male/female washroom signs are an example.

Indexical - signs that are not similar to, but have a logical connection to what they represent. A red stoplight, or a hand waving in greeting are examples.

Symbolic - having neither similarity nor logical connection to what it is they represent. The letters forming this sentence are an example.

The meaning of particular symbols may be at best unclear and at worst misleading given a viewer’s prior experience. Visualization designers need to be aware of the potential for ambiguity

and confusion in their designs, and where possible, employ logically sound symbolism for the abstract metaphors they build.

4.2.4 Social-Cognitive Theory

Social cognitive theory, as outlined by Bandura, deals with "the process by which people make judgements and decisions" ("Social Foundations" 61), suggesting that one does not passively witness an image, but one instead actively arrives at a conclusion regarding that image through a series of perception-affecting mental activities, such as those identified by Bloomer in *Principles of Visual Perception* (12-15);

Memory - new images are interpreted by recalling stored images, and often trigger unexpected connections to related items.

Projection - viewers project meaning based on among other things, their personality and mental state at the time of viewing.

Expectation - viewers tend to see what they expect to see, and often overlook details that don't fit their pre-existing mental model. This functions like "visual cognitive dissonance;" viewers embrace what fits with their existing model and discard what does not.

Selectivity - viewers filter out details that aren't relevant at the time, to avoid overload. This is particularly useful when you know what you are looking for.

Habituation - viewers tend to ignore things they see frequently. Conversely, unfamiliar things tend to stand out, particularly if an existing mental model allows for their appearance.

Salience - viewers tend to notice those things that have a particular relevance or significance to them.

Dissonance - people can only mentally process what they see at a certain speed. This is why parallel processing of information is such an important feature of pre-attentive processing.

Culture - a person's age, gender, upbringing, ethnicity, etc. all have a significant impact on their interpretation of an image.

Language - like culture, the way we use words to describe, interpret and remember ideas differs widely.

Ware tells us that "seeing is not at all the passive registration of information. Instead it is active and constructive" ("Visual Thinking" 8). It is a visualization designer's job to make decisions that highlight and clarify for the viewer the parts of a visualization that are the most relevant to him or her at the time, and to provide feedback that is visually and conceptually consistent throughout the user's experience, to best enable their active understanding.

4.2.5 Colour Theory

In order to make optimal use of colour in visualizations, we must choose colours that both support the functionality required by a visualization, and that help the viewer become engaged, though arguably engaging the viewer should always be a part of required functionality. Colour theory, as developed and practiced for centuries in the art and design worlds (Newton; von Goethe; Chevreul), "is a set of principles to guide the mixing of color for specific psychological and physiological responses" (Eliassen), and provides ample answers to the problem of comfortable engagement using colour.

Below are several questions that designers might consider in order to help themselves address optimal colour functionality in a given visualization, inspired by a list compiled by Healey ("Choosing Effective Colours" 1):

> • Is colour use allowing rapid, accurate identification and grouping of individual data elements, such that patterns can be identified, and inferences drawn?
>
> • Is colour (hue/saturation/element transparency) being used effectively to communicate what Bertin would refer to as "retinal variables" - those not being plotted using the XY(Z) plane?
>
> • Are analagous colours, those that are similar in hue, being used to represent data that is not analagous? Doing so could lead to confusion.
>
> • Does the colour pallette create a sense of comfortable harmony, or is it dissonant, creating unease in the viewer and potentially limiting their engagement.

4.3 Summary - Design & Perception

Visual communication designers strive to ensure an appropriate balance of clarity and immediacy in their work.; clarity, so the work is interpreted correctly, and immediacy so viewers will be engaged with it at length. This same consideration can be applied to visualization design where, especially in the case of large complex data sets, clarity can be a challenge. Visual communication theories offer insight into ways to harness human perception and cognition to maximize the likelihood of both clarity and immediacy, and thus, visualization effectiveness.

Pre-attentive processing is the phenomenon whereby humans perceive some aspects of what they see even before they have a chance to think about it. Understanding this phenomenon

can help designers create more efficient visualization, as defined by Bertin, and better use them as a means of exploration.

The tenets put forth in the Gestalt theory of perception try to describe the laws governing the human ability to perceive stability and order in a complex system of visual stimuli, to help determine how best to visualize especially complex and noisy data sets.

Semiotics, the study of signs and symbols, can help designers avoid ambiguity and confusion when abstractly representing information. Visualizations, whether for the purpose of exploration or analysis, should take pains to ensure that the perception created by the arrangement of elements presented to the viewer is not misleading, and consistently reinforces where possible the relationships between those elements.

Social cognitive theory suggests viewers engage in a series of mental activities when interacting with a visualization, potentially helping designers direct their attention to the visual components that are most relevant to them at a given time.

Colour theory can be used both to assist in understanding, by selecting colours that are logical and which display appropriate contrast, and engagement, by selecting a colour palette that is neither jarring nor boring. A greater level of viewer engagement with a visualization leads to more time exploring that visualization, which in turn can lead to new insights and/or the discovery of questions worthy of additional close research.

5.0 Contemporary Text-Mining Visualizations

5.1 Survey Introduction

As explained by Cukier & Mayer-Schoenberger (28), the big deal with Big Data "is that we can learn from a large body of information things that we could not comprehend when we used only smaller amounts." The trick is, how do we go about sifting through a volume of data that is larger than our limited cognitive capacity to digest? One answer to this question, at least as it pertains to textual information, is text mining. Text mining involves the extraction of new information from textual sources by deriving patterns from the unstructured data via automated statistical analysis. These analyses include such operations as text categorization, text clustering, concept extraction, sentiment analysis, document summarization, named entity recognition, and topic modeling, among others.

The goal of text mining is to turn unstructured texts into data for analysis and perhaps ultimately, hypothesis generation. While the extraction of data from texts may be automated, the interpretation of that data is not. In his work "The End of the Beginning", Moretti explains that what we are seeing when we use these kinds of text mining visualizations are actually models of the data being studied (Moretti 157), which is itself a model of reality. Being by definition incomplete, abstract representations, it is important that we study the visual models emerging from research in Digital Humanities to determine how they might be best designed to facilitate humanities exploration, traditional or otherwise, and to begin answering the question of how

we might use visualizations to find things we don't already know. As always, interpretation is a uniquely human process, and visualization is one means of assisting researchers in interpreting data produced by automated textual analysis.

5.2 Text Mining Visualization Survey

The following is a survey of contemporary text-mining visualizations employed by Digital Humanities researchers. Each is classified according to its role or roles as identified by MacEachren & Taylor (*Analysis/Confirmation; Synthesis; Presentation; Exploration*) and following a brief discussion of the presence or absence of effective design principles and use of 'secondary notation' as described by Petre & Green (56), each is evaluated according to its employment of Schlatter & Levinson's core "meta principles" (*Consistency; Hierarchy; Personality*) for combating the difficulty inherent to working in complex systems.

5.2.1 Docuburst

Visualizing Document Content Using Language Structure

Principal Role(s) - Exploration

DocuBurst was designed as a complement to existing textual visualizations, which provided content detail but no consistent view across documents (figure 1). It presents a highly engaging, highly interactive visualization of "document content based on the human-annotated IS-A noun and verb hierarchies of WordNet (Fellbaum) which can provide both uniquely and consistently-shaped glyph representations of documents, designed for cross-document comparison." (Collins et al. 1).

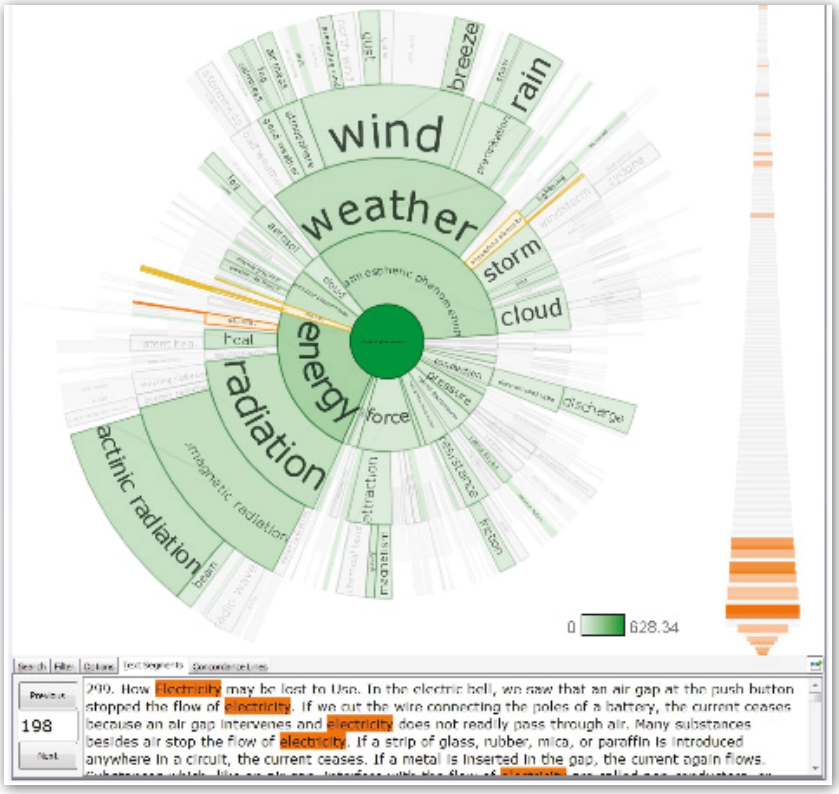


Figure 1. Docuburst. - Visualization of document content based on the human-annotated IS-A noun and verb hierarchies of WordNet (Fellbaum).

Docuburst uses a visually exciting sunburst arrangement of wedge shaped glyphs to display its high level data. Sensibly, the core circle contains the topmost level of information, with each subsequent outward division representing a “subsetting” of data. Assisting the viewer to connect any item from one layer of data to those above and below it, each radial division visually falls within the bounds of the data from which it is derived.

Two additional linked visualizations provide additional information. A colour encoded, stacked bar graph displays a visual representation of word distribution, and a text box displays selected texts.

Docuburst’s use of colour is both engaging and comfortable for the user, with little visual dissonance in the pallette. Sufficient saturation change is used to indicate selected items without

creating undue distraction. Glyph transparency is not only meaningfully attached to a selection of user controlled variables, it also creates an engaging feeling of depth.

Docuburst displays all glyph names, so long as they can be displayed within the glyph shape using a defined minimum text size. Confusion arises because the tags will be displayed using the largest type size possible, suggesting an undue importance to short names that are displayed using a large type size.

Strengths - Consistency & Personality

Weaknesses - Hierarchy

5.2.2 FacetAtlas

Multifaceted Visualization for Rich Text Corpora

Principal Role(s) - Exploration

According to Cao et al. (1172), “There is a lack of effective analysis tools that reveal the multifaceted relations of documents within or across the document clusters.” FacetAtlas displays both global and local patterns simultaneously using “a multifaceted graph visualization to visualize local relations and a density map to portray a global context” (figure 2). Auto-converting search results from a one-dimensional list into a visual graph-based representation affords the user an interactive exploration of the data’s multifaceted relationships.

We naturally cluster visual elements that are sufficiently proximal to one another. Clusters are further highlighted by FacetAtlas using semi-transparent amorphous blobs of colour. As a user zooms for more information on a cluster, additional nodes and edges that were obscured by

the semi-transparent features become more visible, implementing Shneiderman's (1996) Visual Information Seeking Mantra; "overview first, zoom and filter, then details on demand" (1).

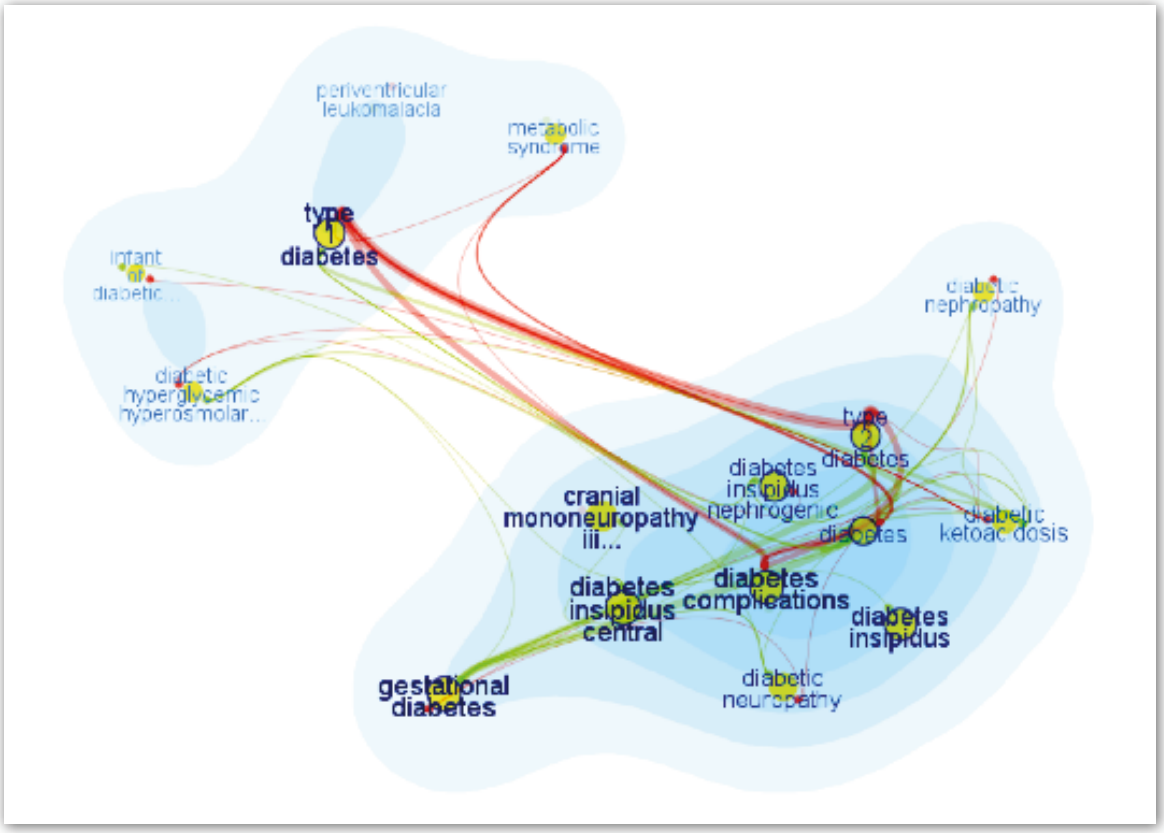


Figure 2. FacetAtlas - Displaying both local (network graph) and global data (density map) simultaneously.

Further, both node text labels and edges vary their weight and transparency based on their relative values, which makes one of the more unique features of FacetAtlas, its affordance of what the team refers to as "dynamic facet-based context switching" (Cao et al. 2) intriguing. The user alters the primary visualization layout arrangement to get a new perspective, while consistently maintaining his/her analytic focus, simplifying visual comparisons between multiple variables and the user's target variable.

Strengths - Hierarchy, Personality & Consistency

5.2.3 ThemeRiver

Visualizing Theme Changes Over Time

Principal Role(s) - Synthesis & Exploration

"ThemeRiver provides users with a macro-view of thematic changes in a corpus of documents over a serial dimension" (Havre, Hetzler, & Nowell 2). Using the metaphor of a river, ThemeRiver uses stacked area graphs to display the relative appearance of themes within a corpus, over time (figure 3).

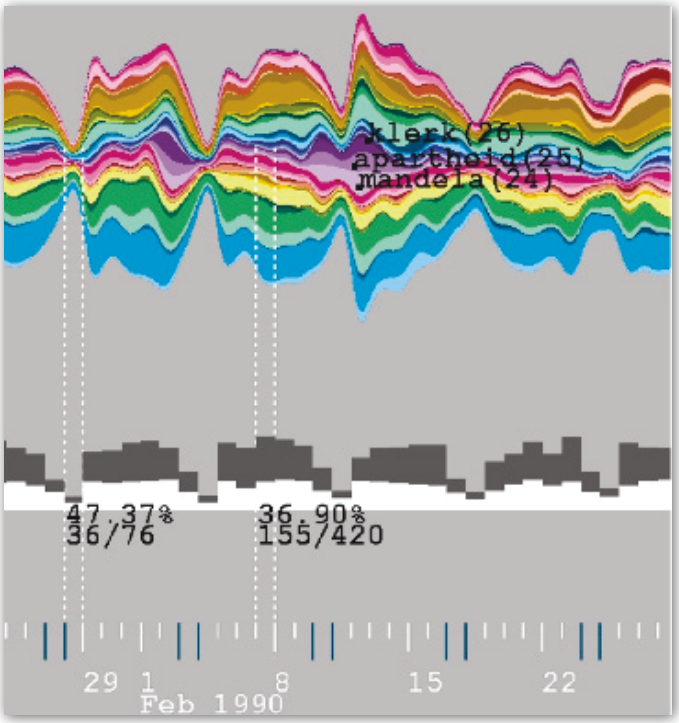


Figure 3. ThemeRiver - Stacked area graphs plotting theme changes over time are sometimes difficult to use for even reasonably accurate comparison of multiple streams. Additional diachronic information can be plotted across the bottom of the viewer.

The ebb and flow of themes is displayed alongside notable current events, and users have the option of displaying multiple streams alongside one another. As well, the user can examine

additional data in histogram form simultaneously, for insight into document totals at a given moment in the stream display. While the tool offers an interesting opportunity for hypothesis generation, ThemeRiver lacks any affordance for document level text analysis.

The multi-coloured, organically flowing shapes of the ThemeRiver display are undoubtedly eye-catching. Users will likely be interested in exploration if only to see what kinds of shapes their data creates.

The stacked area graphs do a reasonable job of identifying large scale disparities in the data but make it hard to intuitively determine the relationship between two elements with similar values. As well, the graphs are prone to misinterpretation, with the shape of inner areas having the potential to affect perception of the shape of those on the outside. Text labels are a consistent colour and scale, offering no tertiary affordances for the user to gain insight into anything beyond a stream's name.

Strengths - Personality

Weaknesses - Hierarchy & Consistency

5.2.4 MoodViews

Tools for Blog Mood Analysis

Principal Role(s) - Confirmation/Analysis

"MoodViews enables a range of applications, providing a window into aggregate states-of-mind of masses of people." (Mishne & Rijke 2) Moodviews tracks and visualizes "mood" info provided by LiveJournal users around the world via user provided tags. Visualization is

accomplished exclusively via line graphs, but the tool offers three views; Moodgrapher, Moodteller and Moodsignals (figure 4). Moodgrapher displays moods over time as provided by blogger tags,

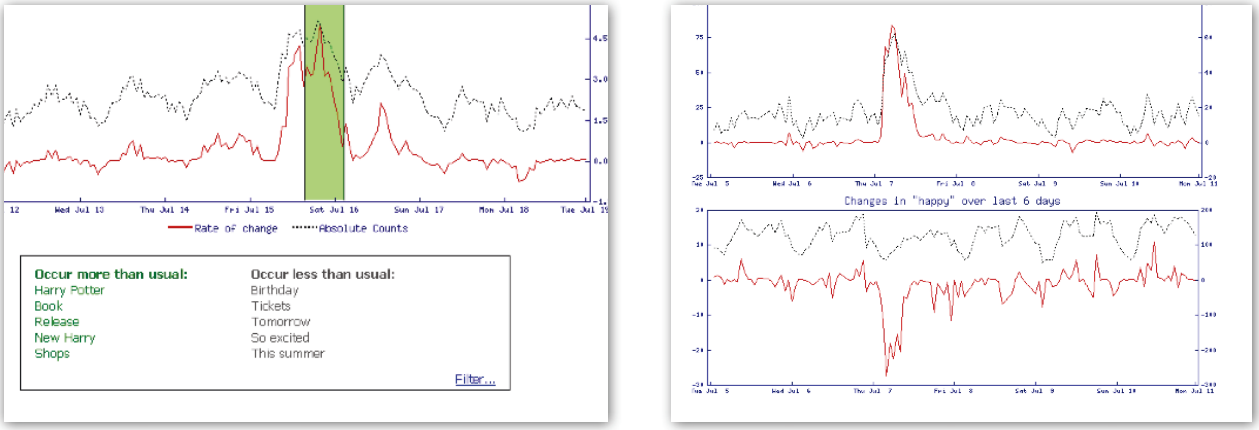


Figure 4. - Moodviews - Traditional line graphs displaying “mood” gleaned from tags provided by LiveJournal users offer little by way visual engagement.

Moodteller affords some measure of prediction using a semantic analysis of blog entries, and Moodsignals promotes insight into likely explanations for spikes in the data by displaying lists of words most associated with the posts responsible for the mood during a selected time span.

MoodView’s traditional line graphs are easy to understand, with up representing an increase in value, and right representing the passage of time. In addition, the MoodSignals tool displays terms that appear more often than normal using green, suggesting a positive correlation, and those appearing less using a more neutral black.

The traditional line graphs used by MoodViews hold little in the way of visual interest for the user. As well, no typographic hierarchy exists at all, with all text rendered using the same typeface and size, leaving the user to struggle with where to look for information.

Strengths - Consistency

Weaknesses - Hierarchy & Personality

5.2.5 Phrase Nets

Generating Visual Overviews of Unstructured text

Principal Role(s) - Confirmation/Analysis and Exploration

According to van Ham, Wattenberg, & Viégas, "[A] problem in the visual display of text involves legibility. In most visualizations, one wants to use spatial position as a meaningful variable. Yet a readable set of words obeys spatial constraints on alignment, grouping, and type size. The conflict between positioning and legibility can lead to displays that are hard to read or where spatial position is essentially random" (1). Phrase Nets (figure 5) uses "phrases" as its unit of analysis, and displays graphs, the nodes of which are words, and the edges of which link the words by user-specified requirements. The query "X of Y" for instance, will return all groups of three words with "of" in the middle. It then uses regular expressions to further focus user inquiries.

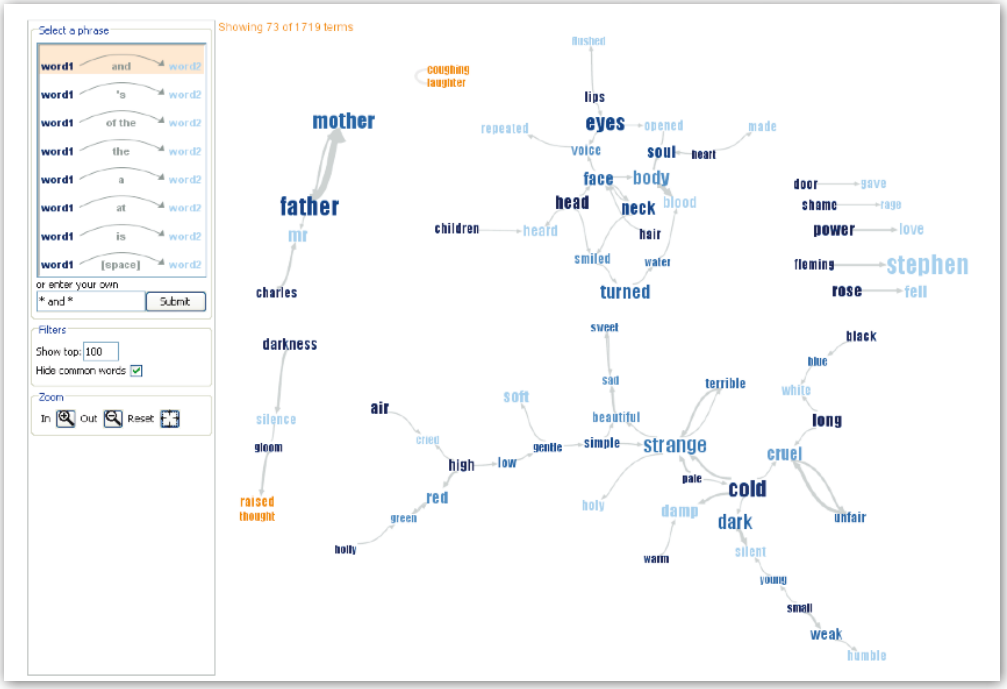


Figure 5. Phrase Nets - Phrase Net uses regular expressions to allow the user to customize searches, entering searches like "X of Y" to find all three word groups with "of" in the middle. Users can choose from a drop down menu, or create their own formulae.

Using word scaling and edge thickness to indicate relative number of word occurrences and connection strength respectively, Phrase Net uses edge compression to simplify the visualized structure and help overcome the issues identified above, arising from the conflict between positioning, scale and legibility. The variety in scale, colour saturation, and value create a feeling of three-dimensionality that is often used by visual communicators to establish and maintain viewer interest. As well, the organic shapes used by Phrase Nets to reinforce grouping will be read by users as "friendly" or "welcoming", which is likely to help maximize user engagement. Finally, we naturally perceive larger elements as well as darker elements on a white field as more visually important, making Phrase Nets' application of larger scale and darker values to words with higher data values visually justifiable.

Strengths - Personality, Hierarchy & Consistency

5.2.6 Luminoso

Multi-Dimensional Semantic Space

Principal Role(s) - Exploration

"Luminoso is an interactive application that aids a researcher in exploring […] semantic spaces in a way that is intuitive for discovering semantic patterns from the dimensionality reduced data." (Speer et al. 1) Luminoso allows the inclusion of "canonical" documents, that is, documents with "known" semantic content against which to compare documents in the corpus.

Figure 6. Luminoso - Documents in a network are displayed using colour temperature and proximity to assist the user in identifying clusters to explore more closely.

Luminoso's interface revolves around the grabbing of a data point in the display, which "simultaneously allows a user to rotate their view using that data point, view associated text and statistics, and compare it to other data points." (Speer et al. 1) Selecting a point also highlights all of the connected, semantically associated data points (figure 6).

Luminoso offers no document level view, but does allow real time addition or deletion of documents to and from the viewing area. The main viewer window plots documents in a network diagram, utilizing colour temperature and proximity to create visual clusters for further exploration, and scale of data points to indicate relative frequency of concept appearance. Underwhelming scale variety, poor use of whitespace and text with uncomfortable line lengths and margins indicate little attention has been paid to viewer engagement.

Strengths - Consistency

Weaknesses - Hierarchy & Personality

5.3 Topic Modeling Visualization Survey

This section presents a similarly evaluated survey of visualizations specific to a text-mining operation called Topic Modeling. "A topic model", says Iwata et al. (1), "is a hierarchical probabilistic model, in which a document is modeled as a mixture of topics, where a topic is modeled as a probability distribution over words," or more generally, an automated text mining technique that tries to identify groups of words with a tendency to occur together more regularly within the same documents in a corpus, defining these groups as "topics." Topic modeling algorithms, in identifying groups of words that are likely to co-occur, tend to ignore both terms that occur uniformly across a text, as well as those that occur very little within a text at all.

We can calculate the relative similarity of any number of given topics, as each word used in all the documents within the corpus has a measurable contribution to every topic, and likewise, each topic has a measurable contribution to each of the documents within the corpus. These two matrices, word contribution to topics and topic contribution to documents, are two of the most fundamental measures to be generated via topic modeling. As explained by Chaney and Blei (2), "One of the main applications of topic models is for exploratory data analysis, that is, to help browse, understand, and summarize otherwise unstructured collections," gaining insight into areas like what themes are prevalent across a corpus, the influence of various subjects on one or more author's works, or how each of these may have changed over time.

5.3.1 InPhO

The Indiana Philosophy Ontology Project

Principal Role(s) - Synthesis & Exploration

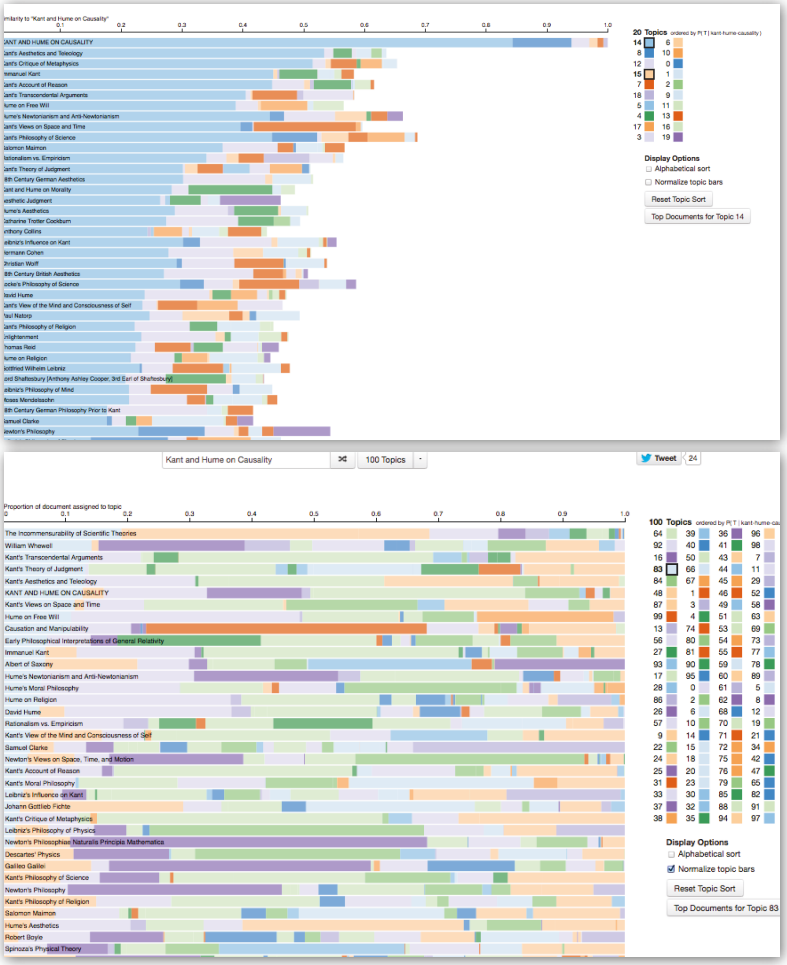


Figure 7. InPhO - InPhO displays the relative topic similarity between a corpus of documents, and a target document. Each topic is colour coded and explorable, while the main display re-orders when any topic is selected. The bottom image displays the normalization feature, stretching each document's profile to the same length, highlighting relative versus asolute values.

The InPhO ("InPhO Topic Explorer") visualization displays the statistical similarity between each document in a corpus and a target document. InPhO uses a bar graph to display relative similarity, and offers a real-time option to display 20 to 120 topics, with each topic shown as a smaller, colour-coded portion of an article's bar (figure 7). A topic's size is determined by its

contribution to the article. Any selected topic re-orders the articles according to use of that topic. on several levels. For clarity, InPhO offers a "normalization option, wherein all articles' bars are shown at the size of the target bar, for ease of ascertaining topic contributions in articles with little similarity to the target.

InPhO's use of bar graphs for topic comparison across documents is engaging, clean and easy to understand. Displaying each document as a series of colour coded topic bars, consistently ordered most to least prominent according to what topic is selected, provides the user with an intuitive and immediate understanding of each document's similarity to the collection. The normalization option, stretching each document's bar graph to the same length as the target document, offers further clarity when the user is concerned with ratios of topics in document composition, rather than absolute number of appearances per document, and an intuitive re-ordering feature makes it easy to focus on a particular topic. Typographic and visual hierarchy is subtle but comfortable in the visualization's initial state, employing ample white space to assist in navigation. Bolding and edge enhancement are used to indicate elements related to those selected by the user.

Strengths - Hierarchy, Personality & Consistency

5.3.2 D-VITA

Visual Analytics of Dynamic Topic Models

Principal Role(s) - Exploration, Synthesis & Confirmation/Analysis

"D-VITA supports end-users in understanding and exploiting topic modeling results by providing interactive visualizations of the topic evolution in document collections and by browsing documents based on keyword search and similarity of their topic distributions" (Günnemann et al. 1).

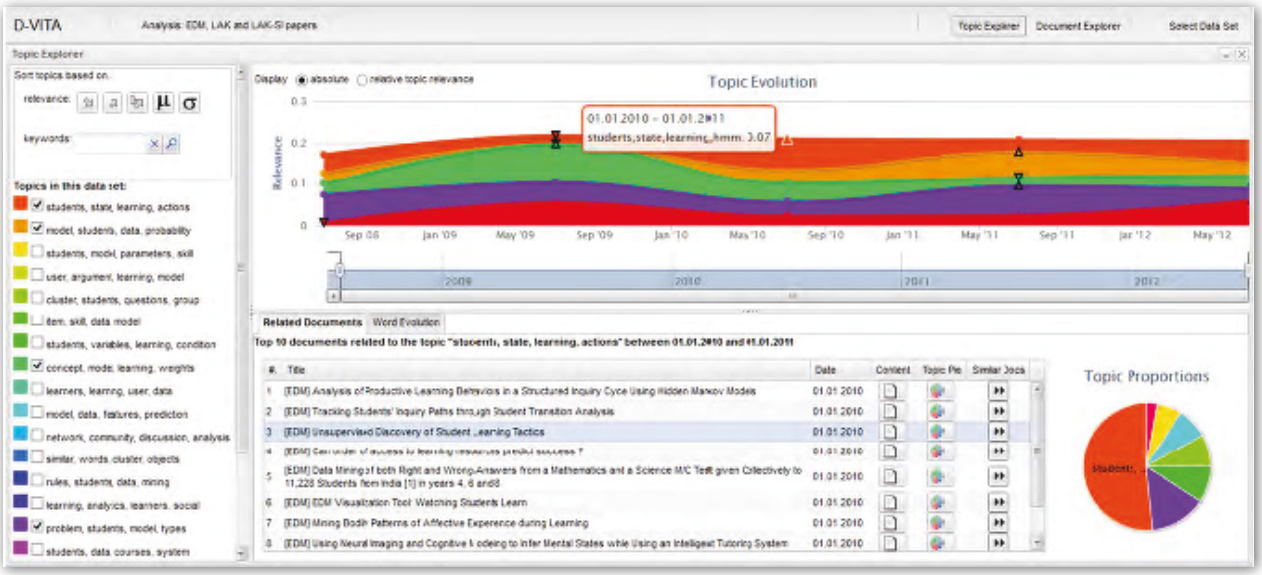


Figure 8. D-VITA - D-VITA's multiple linked views, employ little beyond colour coding of some elements to assist users in navigating the multiple panels presented, or to provide any level of visual engagement. Users may initially find the organic shapes and saturated colours compelling, but the generic nature of the rest of the visualization and the lack of navigation and interpretation affordances will likely limit the extent of a user's desire to explore.

As shown in figure 8, D-VITA provides users a multiple visualization environment for exploring topic models, initially via the organic, stacked area graph of *Themeriver*, providing users "with a macro-view of thematic changes in a corpus of documents over a serial dimension" (Havre et al. 2). Selecting a topic in the Themeriver panel brings up a list of the documents exhibiting the selected topic, and by selecting a document from this list, the distribution of topics in the document is visualized as a pie chart. Additionally, individual documents can be inspected, and users can produce a list of documents with similar topic contribution.

D-VITA provides multiple linked views, but employs little enough beyond colour coding of some elements to assist users in navigating the multiple panels presented, that it is difficult for a new user to establish conventions of operation and navigation. While users may find the organic shapes of the ThemeRiver portions of D-VITA compelling, the remainder of the display is

cluttered and almost antiseptic, dominated by hierarchically similar lists of black type on a white background. Also, as in ThemeRiver, the stacked area graphs make it hard to intuitively determine the relationship between two topics with similar values.

Weaknesses - Hierarchy, Personality & Consistency

5.3.3 MetaToMATo

Metadata and Topic Model Analysis Toolkit

Principal Role(s) - Exploration & Confirmation/Analysis

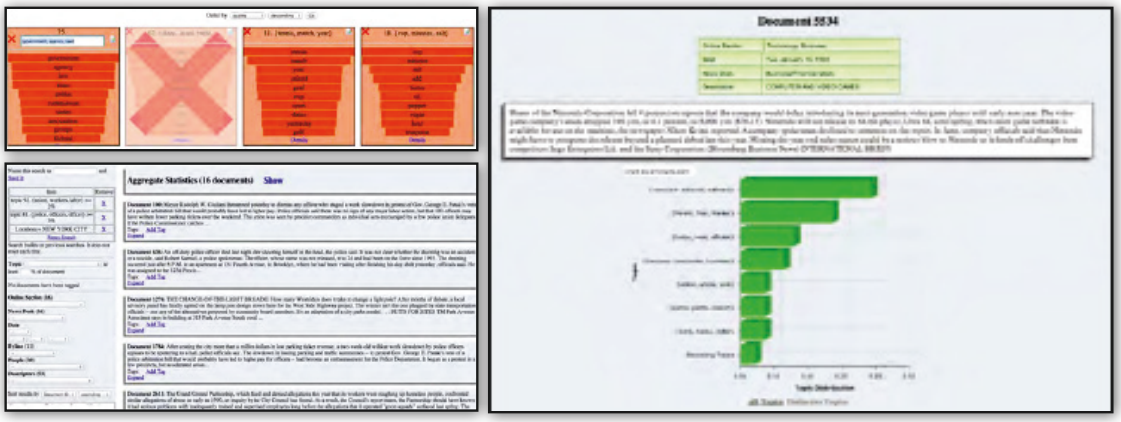


Figure 9. MetaToMATo - MetaToMATo provides multiple, linked visualizations for exploration of document collections. At top left, words in each topic are displayed hierarchically. At right, documents can be further explored as they are discovered.

MetaToMATo (figure 9) is “a visualization tool that combines both metadata and topic models in a single faceted browsing paradigm for exploration and analysis of document collections” (Snyder et al. 2). Topics are displayed as ordered lists of hierarchically important words, as indicated by type size and bolding, each of which labels a normalized bar representing the probability of that word occuring within the topic. These bars are similarly coloured to facilitate user grouping of them as a single type of information.

The Document's view in MetaToMATo is more comprehensive than most of the tools surveyed, providing a short sample of each document returned in a search, along with user selected "quick view" metadata. While typographic hierarchy is used to help the user navigate, these panels have tight margins, and their tabular form is not optimal for rapid reading. Users can expand documents, view all their metadata, view a graph of the distribution of topics in any document, and view topics particularly distinctive to this document. The expanded document view provides ample white space, but the three-dimensional bar graph feels generic, and once again, the text box has uncomfortably tight margins and long line lengths.

MetaToMATo uses colour to connect related visualizations, rather than to separate one element within a visualization from another, helping provide the user with a more understandable overview. While comprehensive in its offering of filters, metadata and tags with which to explore a corpus, MetaToMATo provides no high-level corpus wide visualization of topics or documents, and no diachronic views for temporal analysis, both of which I feel are key components of exploratory research.

Strengths - Hierarchy & Consistency

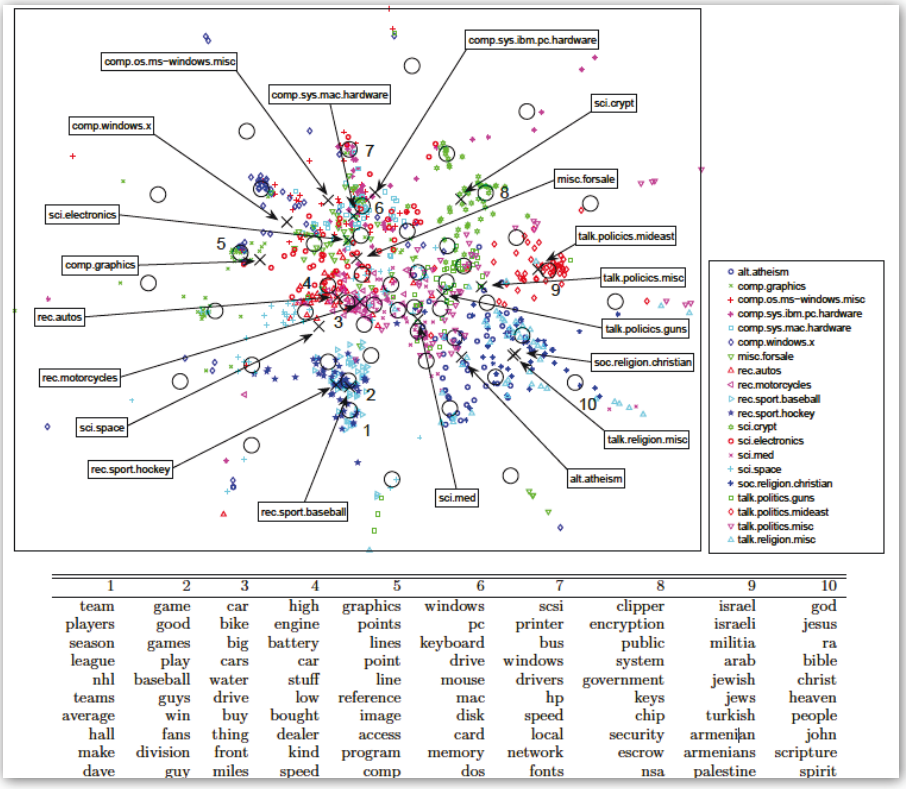
Weaknesses - Personality

5.3.4 PLSV Method

Probablistic Latent Semantic Visualization

Principal Role(s) - Presentation & Synthesis

Iwata et al.'s PLSV method utilizes the EM algorithm (expectation maximization) to fit the model to a given set of documents, and as such, documents with similar semantics will appear close together, even if they share few words, and thus, topics. Less a tool than a unique topic modeling algorithm, the PLSV method assumes that both topics and documents can be plotted in either a 2 or 3 dimensional Euclidean space, and simultaneously "estimates topics and visualizes documents in one probabilistic framework (Iwata et al. 2)."



1	2	3	4	5	6	7	8	9	10
team	game	car	high	graphics	windows	scsi	clipper	israel	god
players	good	bike	engine	points	pc	printer	encryption	israeli	jesus
season	games	big	battery	lines	keyboard	bus	public	militia	ra
league	play	cars	car	point	drive	windows	system	arab	bible
nhl	baseball	water	stuff	line	mouse	drivers	government	jewish	christ
teams	guys	drive	low	reference	mac	hp	keys	jews	heaven
average	win	buy	bought	image	disk	speed	chip	turkish	people
hall	fans	thing	dealer	access	card	local	security	armenian	john
make	division	front	kind	program	memory	network	escrow	armenians	scripture
dave	guy	miles	speed	comp	dos	fonts	nsa	palestine	spirit

Figure 10. PLSV Method - Scatter plots rendered with no use of scale or transparency to assist in interpretation, and an otherwise spartan, generic feeling are unlikely to inspire exploration. Linked panels provide additional

The closer a document is to a topic in the visualization space, the greater the chance the document will contain the topic. “The topic proportions of a document are determined by the distances between the document and the topics in the visualization space” (1).

As shown in figure 10, the visualization offered by Iwata et al. consists mainly of scatter plots displaying topic and document distribution in a two-dimensional field. Additional information is provided below, in a text list of probable words for ten topics selected within the visualization, and in a sidebar legend. The plots are difficult to interpret, making no use of scale or transparency where it might help. Additionally, the scatter plots feel generic, and suggest only an expert might usefully interpret them. The PLSV topic modeling algorithm may well offer a more refined calculation of these relationships, but the visualization itself, and lack of user affordances, offers little likelihood of user engagement or extended exploration.

Weaknesses - Personality, Hierarchy

5.3.5 Unnamed Tool - Chaney & Blei

Visualizing Topic Models

Principal Role(s) - Exploration

In their 2012 paper “Visualizing Topic Models”, Chaney and Blei present an unnamed tool for the exploration of topics and documents within a corpus (figure 11). Beginning with topics identified via Latent Dirichlet Allocation, users dig down to related topics and documents, exploring document text as desired. Each selected document is presented alongside a pie graph displaying topic contribution to that document.

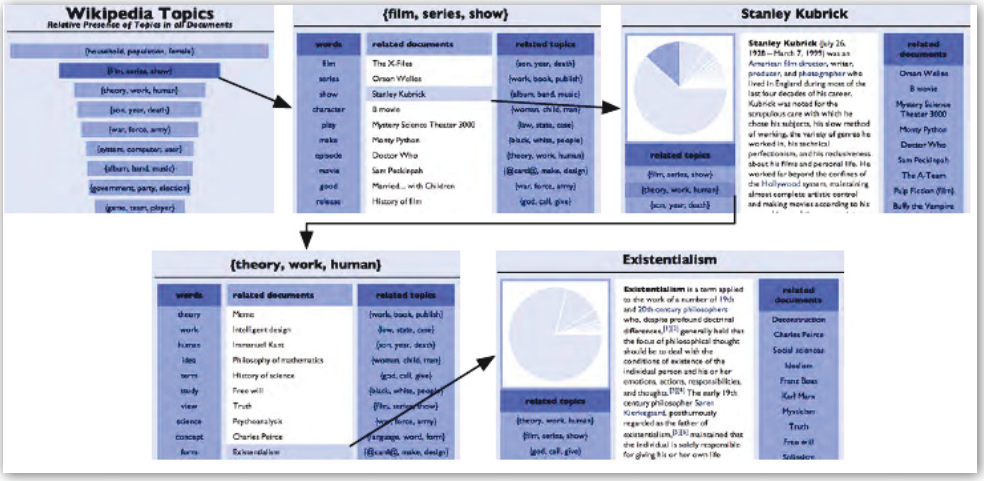


Figure 11. Unnamed Tool by Chaney & Blei - Users dig through layers of data, beginning with topics identified within a corpus. Each selection offers new lists of related data to which the user may navigate and explore.

This tool lacks a high level corpus view beyond the found topics and their prevalence, and each selection made hides the user's current data, in favour of those deemed relevant to their selection. As such, it is not possible to display multiple topics, words or documents side by side, for comparison and analysis.

Presenting the user with a reasonable number of options, usability conventions are readily understandable. Navigation is fairly linear and logical, and facilitated by panels with ample room for the content to breathe. The monochromatic colour scheme provides a unified feeling to the tool as a whole, but inconsistencies in its application can hinder navigation, as a given data type may appear in several positions on screen and with different background colours, depending on the user's position within the framework. The clear typographic hierarchy and the monochromatic colour pallette combine to suggest a user friendly environment, even for the non-expert user.

Strengths - Personality & Hierarchy

Weaknesses - Consistency

5.3.6 Serendip

Topic Model-Driven Visual Exploration of Text Corpora

Principal Role(s) - Exploration, Confirmation/Analysis

As the name suggests, Serendip is a text-analysis tool that promotes serendipitous discovery. Alexander et al. identify corpus level, document level, and word level, as three desirable levels of "zoom" in a text-analysis tool, supporting Shneiderman's (1) overview first then details on demand mantra. "To enable fluent fusion, a system must provide not only a set of views for looking at the data from multiple viewpoints, but also connections between the different types of information allowing a reader to move smoothly across scales, data types, and research questions… Users can make a more diverse set of findings if they can view the data from a variety of different angles." (Alexander et al. 3)

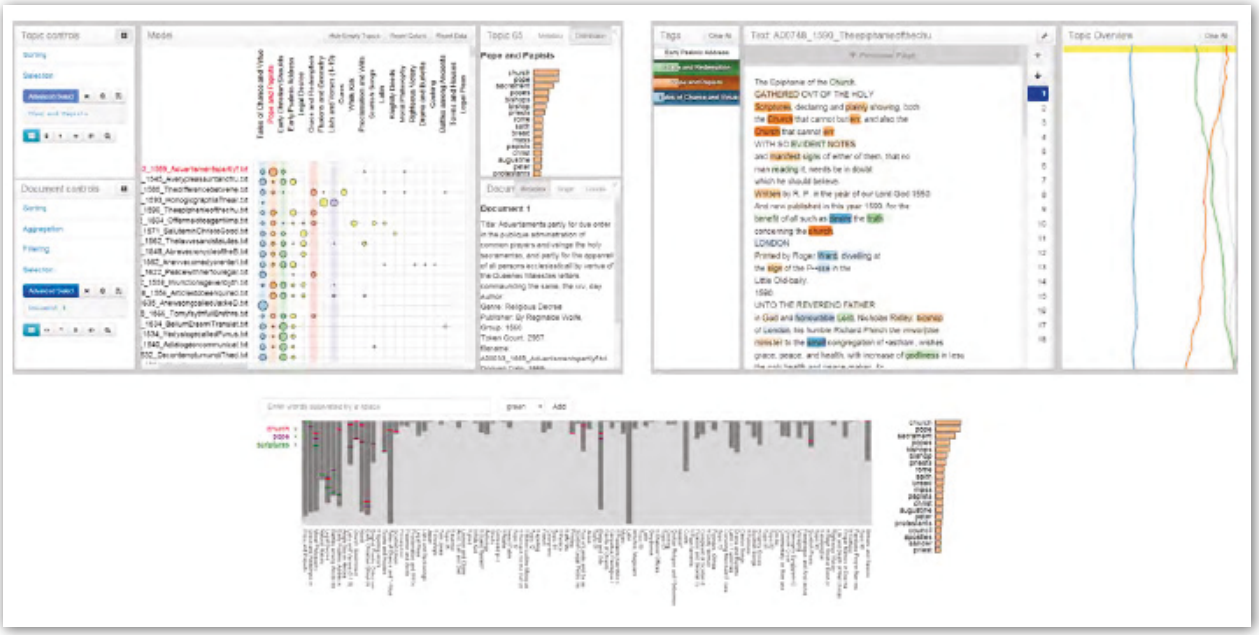


Figure 12. Serendip - Users are provided a visually connected suite of visualizations, affording them the opportunity to explore three important levels of zoom in text analysis; corpus, document and word.

Serendip offers linked visualizations, affording the user the chance to observe relationships that a single visualization often hides. A two-dimensional matrix corpus overview offers users a re-orderable matrix to help the user discover high-level patterns, then select specific documents and topics to look at more closely. As shown in figure 12, the matrix employs colour and scale to assist the viewer's interpretation. Histograms and line graphs, colour coded to match the overview, are used to signify the contribution of individual words to topics.

Serendip's strengths lie in the provision of multiple access points, each of which makes consistent use of colour density to help display value; the main topic/document overview, displaying topic appearances in documents and word contribution to selected topics, the text viewer, allowing document navigation by density of one or more topics, and the rank viewer, displaying how words rank in various topics. As well, Serendip makes effective use of whitespace, colour, scale and text direction in the construction of their layout, encouraging curiosity and exploration.

Strengths - Personality, Hierarchy, Consistency

5.3.7 Termite

Visual Analysis System for Human-Centered Iterative Topic Modeling

Principal Role(s) - Exploration & Confirmation/Analysis

Termite, like Serendip, offers a main view consisting of a two-dimensional matrix displaying word contribution to topics, using scale and colour to aid in grouping and value comprehension. Termite goes a step further, using a "seriation algorithm that both reveals clustering structure and

promotes the legibility of related terms" (Chuang, Manning & Heer 1) and maintaining natural reading order in word lists to preserve a semblance of contextual understanding in their displays.

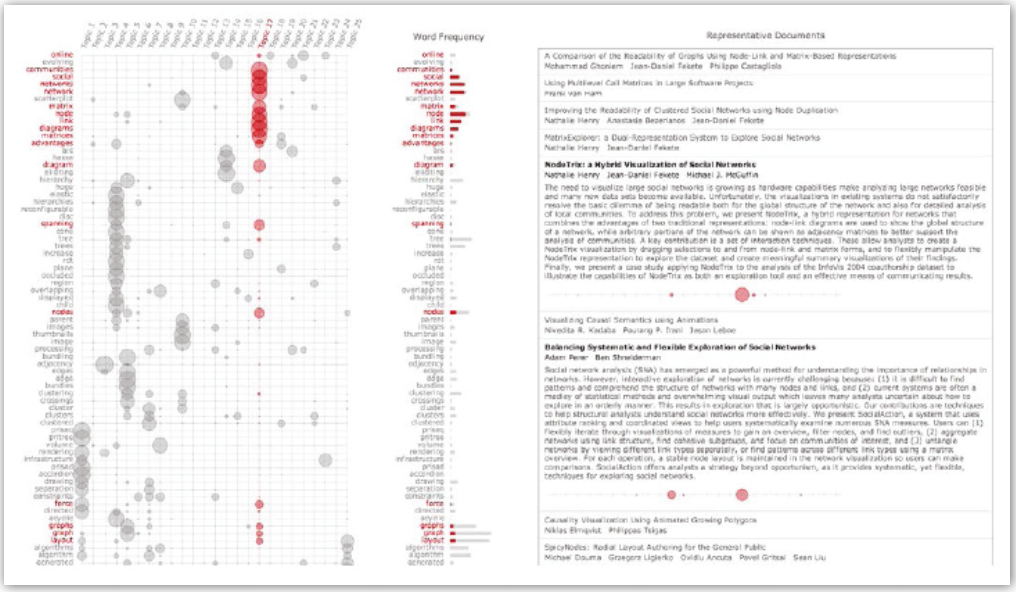


Figure 13. Termite - Consistent use of circular shapes provides an engaging, unified feeling to the visualization. Scale, transparency and and seriation help identify interesting clusters of data.

As shown in figure 13, colour coded histograms display word frequency distribution, relative to the full corpus, in a linked panel when a topic is selected in the main panel. A final panel automatically displays a selection from the most relevant documents in the corpus, along with a modified line graph illustrating topic signal strength throughout each document, using the same familiar colour coded circles from the main viewer.

Consistent use of circular shapes as targets, even on line graphs, consistent use of shape volume to indicate scale of data, and consistent use of colour to indicate related data elements makes Termite an engagingly friendly collection of visualizations to navigate. The additional optional use of seriation to aid in clustering helps the user quickly identify the words with the greatest relevance to a given topic.

Strengths - Consistency, Personality, Hierarchy

5.4 Visualization Survey Summary

Output from text-mining can assume a near infinite variety of forms. In this survey alone, principal visualizations consisted of exploded pie charts, network diagrams, diachronic stacked area graphs, line graphs, bar graphs, histograms, word lists, scatter plots, two-dimensional matrices, word clouds and dendrograms.

Each of these visualization varieties is functional, but fundamentally limited in some capacity in its ability to communicate especially complex relationships. Simple traditional visualizations, like histograms and line graphs for instance, alone have insufficient capacity to display complex multivariate relationships. Dendrograms, especially when your data sets are massive, quickly become an illegibly dense mess. Word clouds are great at showing the words in a topic, but not at easily and clearly comparing several texts or topics to one another. Network diagrams often produce engaging, suggestive results, but suffer from the same difficulties as dendrograms; large data sets quickly lead to illegibility. The variety of different visualization techniques evident across the survey reinforces the difficulty inherent to displaying complex text mining data, and suggests that multiple, linked, interactive, visualizations can help the user overcome the limitations inherent in any single visualization by providing affordances for each other – a histogram for instance might aid in interpreting a dendrogram by providing insight and clarity into variables not directly measured or shown.

Those visualizations using traditional line graphs and scatter plots, and to a lesser extent bar graphs and pie charts as a their main feature will likely generate low levels of engagement. As with other applications of visual communication design, creating sufficient unity and variety

among visual elements is a key to establishing and maintaining user interest. Traditional line graphs and scatter plots have insufficient scale and shape variety in their constituent elements to create much visual interest. Displaying additional data in those same charts, using variable line thickness or point size for instance, increases the potential for exploration, and creates additional viewer engagement.

The difference in the desire to engage with a personality deficient visualization versus a visualization with a strong personality should not be underestimated. Creating a desire to play and explore goes a long way toward encouraging the user to investigate and eventually gain insight into complex, multi-variate data.

Creating a visualization with personality is not an entirely separate consideration from establishing hierarchy and a consistent visual language. It is in fact important to consider how consistency and hierarchy can be *used* to create engagement. The thoughtful control of visual design elements like shape, scale, colour, alignment, proximity, and active white space can help not only establish and maintain a logical, hierarchical visual language, they are among the means by which a designer establishes "feel", or as Schlatter & Levinson describe it, personality.

The more complex the relationships being measured, and the greater the number of desired display variables, the more complex the required visualization solution. Single visualizations offer an easier time creating clarity, as well as fewer limitations to creating powerful engagement, or immediacy, than linked visualizations, but at some point a single visualization is insufficient to efficiently display the desired variables, and gives way to multiple linked visualizations, a change which brings with it both challenges and opportunities tied directly to the affordances it offers.

Visualizations with few user affordances are the simplest to design. Multiple linked visualizations on the other hand, provide more exploratory potential, but are also more demanding environments in which to create both engagement and consistent visual hierarchy, as they must establish and then adhere to a helpful, consistent, cross panel, visual language, in order to remain clear.

If the intent is simply presentation of established data, in a static visualization in print for instance, the designer knows in advance all the pieces with which they are working. Once they begin offering affordances for exploration, there arises the potential for a great many unknowns, depending on the data set. Visualizations designed for a specific data set can be easily massaged to display that data in an engaging fashion. Visualizations which allow user upload of data need to consider the range of potential values that *might* drive the visualization.

Multiple, linked visualizations, those with the most user affordances for modification and customization of results, are the most complex systems to design, but also offer the greatest chance for new knowledge discovery. I believe there is great potential benefit in creating the sort of interactive visualization that enables a user to custom filter unwanted information in order to perform what amounts to exploratory research on large data sets, affording them the opportunity to gain insight into data prior to forming a specific research question, and perhaps to generate testable hypotheses.

6.0 CASE STUDIES

In the following two sections, I present two visualization tools designed for the exploration of text mining data. By exploring combinations of visualizations, I hope to increase not just the amount, but also the utility of the data shown, and to do so in a way that engages the viewer, not simply allowing, but instinctively encouraging them to spend time exploring the data.

I propose that creating visualizations that promote greater user engagement will invariably increase exploration, that increased exploration will create a greater likelihood of new insights or questions, and that these insights or questions have the potential to suggest unique research directions in which more traditional scholarly experts can focus their studies.

6.1 Dendrogram Viewer

Figure 14 displays a traditional dendrogram showing the agglomerative clustering of five hundred computer journal articles over a sixty-year span. A fixed set of five hundred articles can hardly be considered “Big Data”, but note how congested and illegible the data is, even at this scale. No principles of design are assisting user interpretation, nor are there any affordances allowing the user to explore the data.

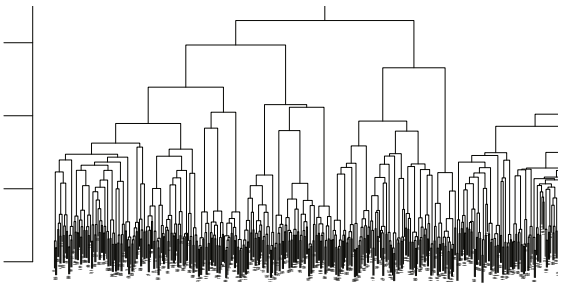


Figure 14. Traditional Dendrogram

Figure 15. Concept Sketch of Modified Dendrogram - Combination of modified dendrogram and diachronic histogram displaying a collection of five hundred philosophy papers, by type.

Figure 15 is a concept sketch for an interactive visualization tool showing that same clustering, but combining a modified dendrogram, showing journal clustering, with a diachronic histogram,showing the prevalence of journal types over time. This sketch formed the basis of a text mining visualization tool developed in conjunction with my colleagues at the University of Alberta[1]. This tool, which we refer to as the "Dendrogram Viewer" (Montague et al. "Seeing the Trees"), combines multiple visualizations to facilitate exploration of a sample set of documents, tagged with meta-data such as author gender, date of publication and document "type", as defined by the researcher. In it we make use of Shneiderman's "Visual Information Seeking Mantra; Overview first, zoom and filter, then details on demand", as presented in his 1996 paper "The Eyes Have It" (1).

Figure 16. Dendrogram Viewer -The collection overview, using colour, proximity, direction and scale to help the user easily understand the overall pattern in the data. Multiple linked views reinforce relationships in part by utilizing a shared, consistent colour scheme.

We first provide an overview using visual cues to amplify perception of relevant grouping and scale (figure 16). A controlled colour palette, using colours of similar saturation minimizes interference caused by dissonance or value differences. Scale changes in the rectangular glyphs indicate the volume of data in a given cluster element. Shape proximity reinforces algorithmic grouping of journal types. These visual conventions help the the user to sense patterns in the data and begin to ask questions like, “Why do these clusters seem to dead-end here?” (a - figure 16), or, “why is there a large clump of Orange in some years and not others?” (b - figure 16).

The viewer can “zoom and filter” by selecting either individual papers or clusters to see how they relate to the clusters both above and below in the data (figure 17). Further, the diachronic histogram at the bottom provides simultaneous insight into how the “kinds” of papers appeared over time. As well, the user can toggle what is displayed in each graphical cluster, choosing between

the “types” of publication content we have defined, and the decade of publication of the articles contained within the cluster, depending on the nature of their inquiry. In the panel at bottom-right, the viewer is able to dig deeper into “details on demand” about individual papers in the groups he or she selects, accessing citation information and any appropriate meta-data, such as author age or gender.

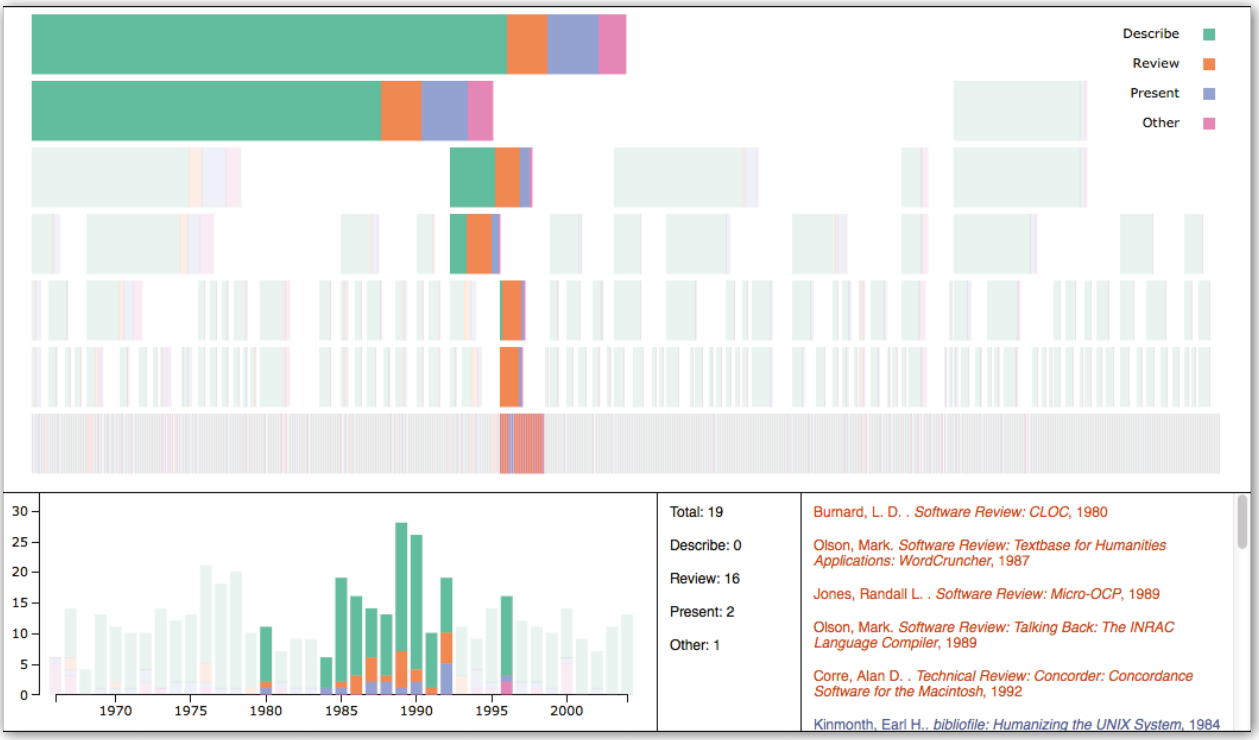


Figure 17. Dendrogram: Zooming & Filtering - The user can explore relationships by selecting either a paper, or a cluster, to view how it cluster through data. Relevant documents are identified in the panel at right, while the histogram hilights the years contributing documents to the selection.

This prototype uses a fixed, relatively small data set, but in the future, the user will have the ability to make decisions about things like how many papers to examine, over what years, as well as what meta-data to track as variables. Interacting with the visualization in this way allows the viewer to explore the data set in ways the visualization’s creators might not have imagined, opening up a myriad of possibilities for new knowledge discovery.

6.2 Topic Model Galaxy Viewer

Having surveyed existing visualization tools, I became involved in the development of a tool that provides targeted affordances for the exploration of topic modelling results from a large text corpus, requires little expertise on the part of the user in response to Alexander et al's (9) call for "model agnosticism," and allows the user to drill down and uncover more layers of data as their research or interests required (figure 18).

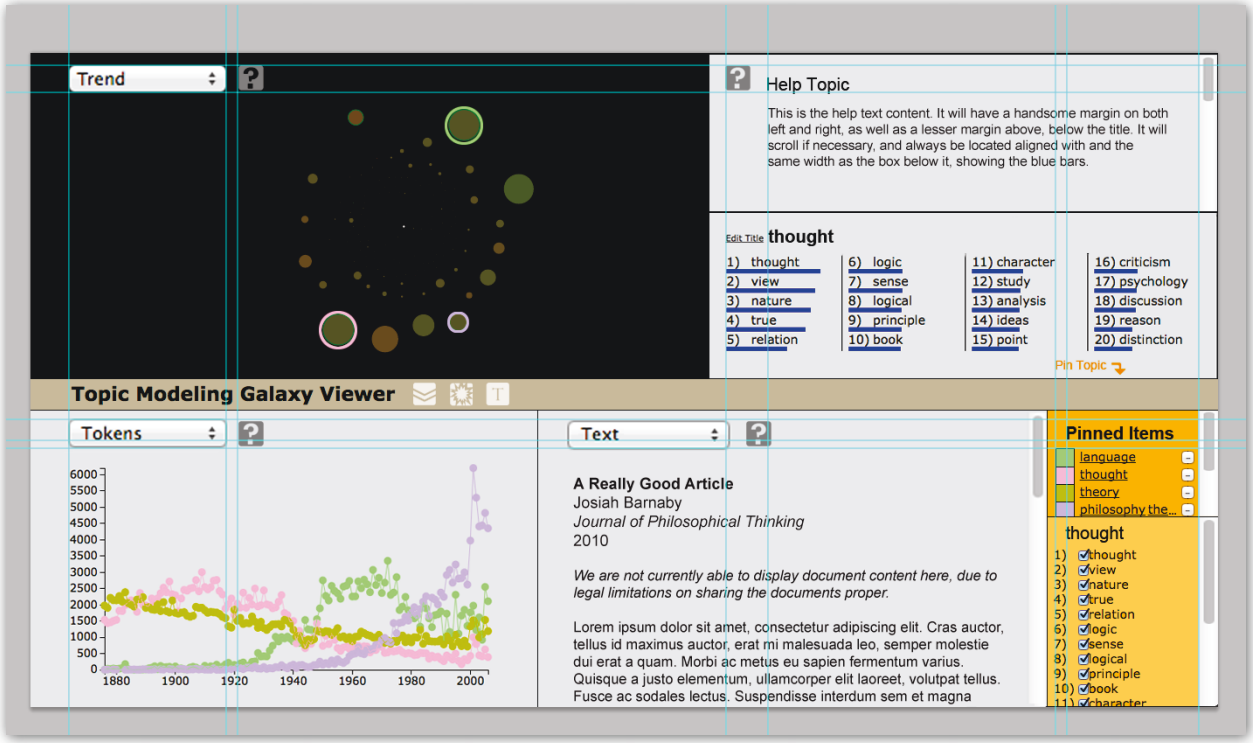


Figure 18. Topic Model Galaxy Viewer Concept Sketch

One of the greatest challenges to creating useful topic model visualizations involves enabling the user to see, explore, interact with, and meaningfully interpret topic modeling data from the highest level view of the entire corpus, down to the level of the individual documents, topics or words. This tool (figure 19), also developed in conjunction with colleagues at the University of

Alberta[2] , to which we refer as the *Topic Model Galaxy Viewer* (Montague et al. "Exploring Large Datasets"), addresses this challenge by offering users a customizable selection of visualization panels in a multi-faceted, browser based environment, which can simultaneously exhibit different aspects of the corpus being modelled, both synchronically and diachronically, as desired by the user to support their exploration. Users can interact with and explore a corpus of documents from any altitude, altering their focus from very broad to very narrow.

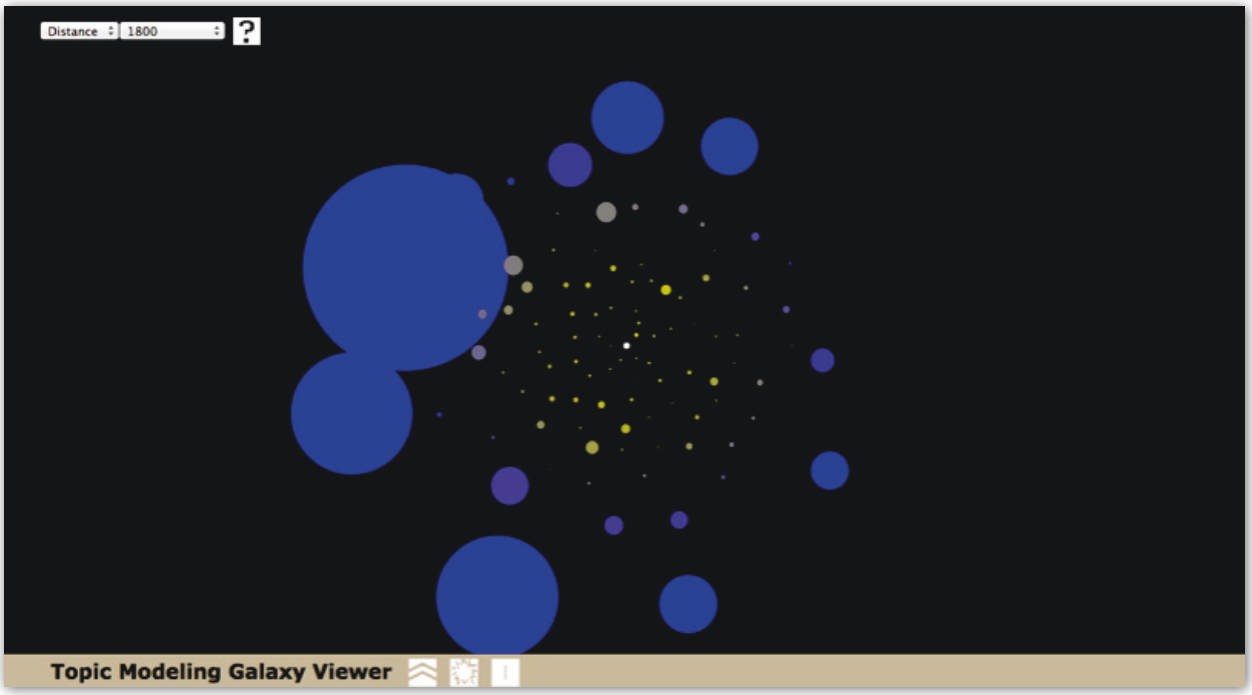


Figure 19. Topic Model Galaxy Viewer - Topic model results are plotted in a 2000 dimension space and rendered in a pseudo three-dimensional "galaxy" view.

The customizable selection of concurrently viewable panels, offers the user a choice of several visualizations, each a different facet of the data within the corpus. The main galaxy viewer provides an overview of all the topics gleaned from the corpus via the topic modeling algorithm, plotted in a zoomable, abstract, two thousand dimension, pseudo-3D galaxy. Users can slide open the secondary view panels, which by default display a list of documents in the current selection,

and a diachronic graph of token appearances in pinned topics (figure 20). When the secondary panels are open, the main viewer remains visible and accessible in the top half of the screen.

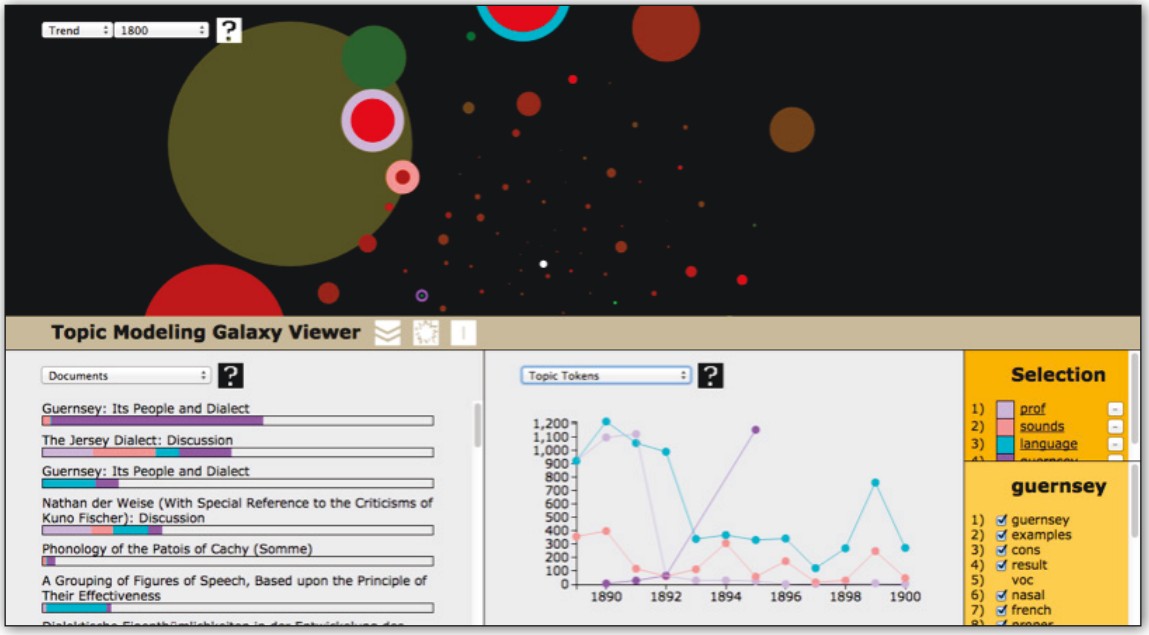


Figure 20. Secondary View Panels - These panels are customizable, allowng the viewer view the linked visualizations that are most useful to them at a given moment.

Snyder et al. (1) explain that "effectively exploring and analyzing large text corpora requires visualizations that provide a high level summary." Our main galaxy viewer provides a high level visualization of all the topics in the corpus, with galactic centre represented by the "zero topic", shown as a white circle, and is essentially the topic in which every word has an equal chance of occurring. Each plotted topic displays a number of characteristics; size, distance from galactic centre, distance from all the other topics, and finally, colour. A topic's size is the most straightforward of these traits, and is determined by the topic's prevalence across the corpus, with those topics appearing most frequently being largest.

Each topic's distance from galactic centre is a factor of what we refer to as that topic's strength. A topic's strength is representative of how widely it appears, and how often, across many documents in the corpus. Low strength topics, those closest to galactic centre, tend to exhibit only localized associations, that is, they appear in only a few documents. Stronger topics, those furthest from galactic centre, tend to appear in many documents throughout the corpus.

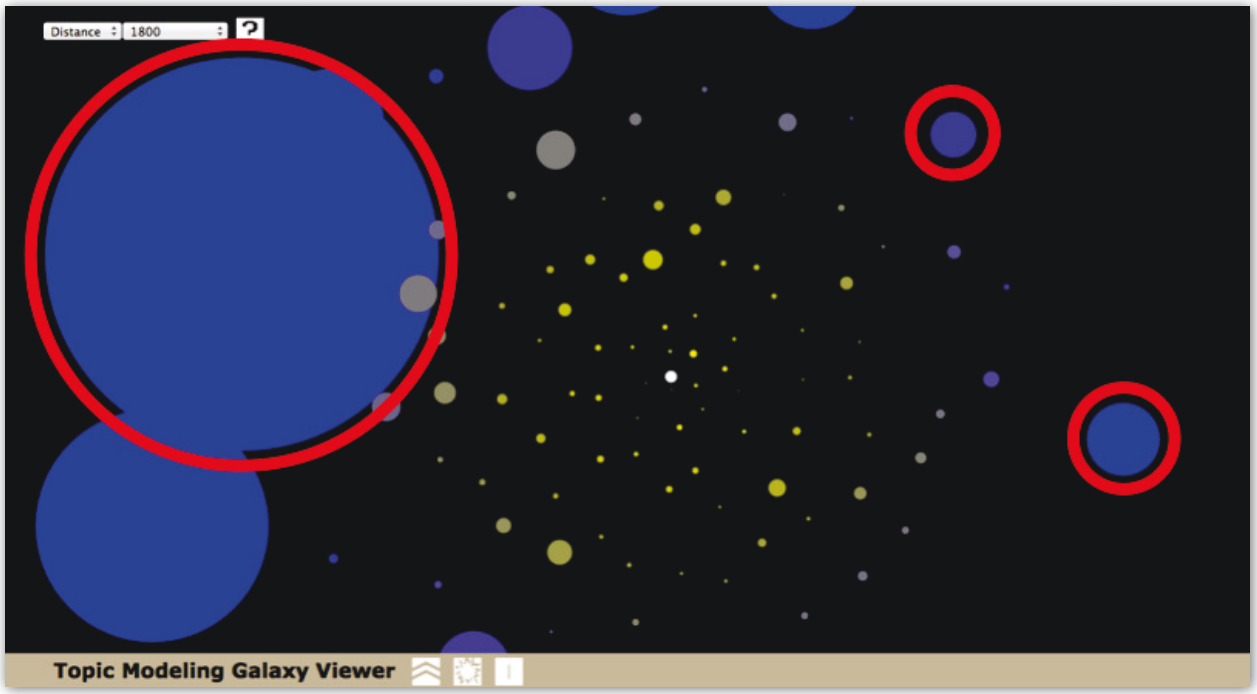


Figure 21. Topic Position - There is no absolute "direction" in the Galaxy Viewer. The circled topics will all appear in approximately the same number of documents across the corpus. The two smaller circles, being closer together, are more likely to have words in common than either is with the large topic at left.

A topic's position relative to all other topics is indicative of the chance those topics share words. There is no absolute "east/west/north/south" in the galaxy viewer, only relative positioning. A topic on the far right of the display is not more "rightish", but it does have a lower likelihood of words in common with a topic on the far left, and both topics in figure 21, given their positions far from centre, would appear in many documents across the corpus, when compared to topics

closer to centre. To reinforce the notion that absolute position is meaningless, and potentially aid in exploration, we have included an agitate function for the main viewer frame, which shakes up the main viewer by quickly re-applying the force direction algorithm responsible for calculating topic positions. Since absolute direction is meaningless, the re-plotted positions provide a new and different arrangement of data points for the user to explore, potentially prompting insights.

The user can choose what single variable they want topic colour to represent from options presented in a drop-down menu. Currently the options are limited to strength, further reinforcing topic distances from galactic centre, especially in cases where scale causes topics to overlap, and trend, which calculates each topic's overall appearance curve over the time span of the corpus. Those increasing in usage glow hot, while those decreasing in usage glow cold, with the severity of both indicated via increased colour saturation. A topic's specific trend curve can be viewed via the tokens panel, in one of the secondary windows.

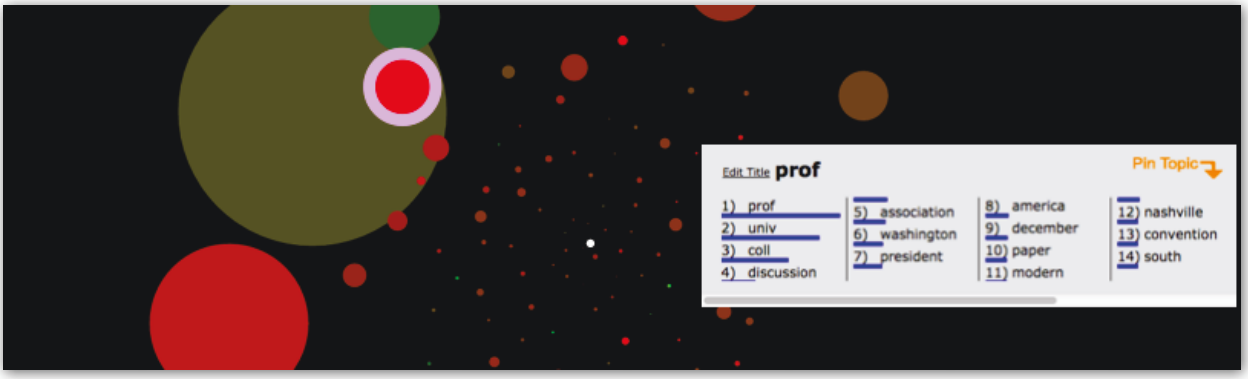


Figure 22. Topic Word Contribution - On selecting a topic, the top twenty most relevant words are displayed.

On selecting a topic in the main viewer (figure 22), a popup up window displays the topic's twenty most relevant words, with bars indicating each word's normalized relative contribution to the topic. The popup window also offers options to edit the topic's name, and to "pin it" to the list

of topics that will be displayed in secondary panels. Pinned topics are indicated via a green ring surrounding their icon in the main viewer.

The two secondary panels offer a selection of linked visualizations selected via a dropdown list located in the top right of each window. The user is able to choose which visualizations they would like to explore together, moving from the high level summary, deeper into the data, simultaneously experiencing multiple visualizations of related corpus data from unique vantage points. As espoused by Chuang et al. (2), users are able to drill down into additional layers of information as desired. Since any single visualization is fundamentally limited in its ability to communicate very complex relationships, like Chuang et al's *Termite*, and Snyder et al's *MetaToMATo*, this concept merges multiple visualizations into what Snyder et al. term "a single faceted browsing paradigm for exploration and analysis of document collections" (2).

7.0 Conclusion

Dendrogram Viewer

Agglomerative Clustering of Journal Articles

Principal Role - Exploration

The Dendrogram Viewer enables zooming and filtering by allowing the user to select clusters of data, to view the flow of clustering through the collection, or individual documents within a given cluster, to view year of publication or other pertinent meta-data. Additionally, users can toggle between using colour to display either publication date or document type, depending on their interest.

A harmonious colour pallette creates a sense of calm, offering no visual dissonance. Sufficient distinction exists between the hues used to allow immediate perception of the nature of a given document or cluster of documents. Consistent application of colour across all panels aids in navigation and exploration, easily connecting an element in one panel with other relevant panels.

Scale changes indicate the relative size of clusters of documents, and the spatial arrangement of clusters reinforces algorithmic grouping.

Typographic hierarchy is not overly well developed, though information hierarchy is supported by making the main view panel notably larger then the secondary panels.

Strengths - Consistency, Personality

Weaknesses - Hierarchy

Topic Model Galaxy Viewer

Targeted Affordances for Topic Model Exploration

Principal Role - Exploration

The Topic Model Galaxy viewer allows exploration of topic model results from the level of the entire corpus, down to the individual document level. The main viewer, initially a full screen view, displays topic modeling results in a space reminiscent of a galaxy. Additional panels offer a selection of linked visualizations displaying things like document text, most relevant words

The Galaxy Viewer makes good use of scale and colour. Topics are presented as dots of varying size and colour, with larger topics appear more often across the corpus. The user can select what variable he or she wants colour to represent from a drop down list; strength or appearance trend. The relative position of topics is indicative of their likeliehood of shared words.

A harmonious colour pallette offers no dissonance, and contrast with the black background while sufficient for colour identification, is not distracting. Consistent use of colour across all panels aids navigation and exploration. The red/blue colour scheme differentiating between topics that are trending hot/cold is sensible, though subtleties in the saturation of the topics with relatively flat appearance trends render them all very similar.

Scale differences easily distinguish between topics that appear often, and those that appear only seldom. Scale variety, and the expanse of space between the dots provides an engaging, pseudo-3D experience for the viewer.

Typographic hierarchy is only moderately developed, offering only rudimentary assistance to navigation. The full screen main viewer remains identifiably the largest and most important

panel, even when the secondary panels are opened, helping the user return to full corpus exploration having dug deeper into the secondary panels.

Strengths - Consistency, Personality, Hierarchy

The capabilities of the media with which we communicate with one another have evolved over time as our communication needs have grown and become more complex. We are now at a point where we are capable of collecting and studying previously unfathomable amounts of data, and there seems to be no growth slow-down on the horizon. The advent of the World Wide Web heralded a massive surge in data production and recording, and moving forward, as smart objects and the Internet of Things connect the world in more and varied autonomous ways, we will need new and different ways to evaluate a growing volume of data.

With huge collections of data, there exists the opportunity to learn things you could not learn from studying a smaller data set. The speed of new data production and transmission, the volume produced, and the variety of formats however, make Big Data a challenge to interpret. As massive data becomes more and more pervasive, we need entirely new ways to examine and understand it.

This paper opened with the question, "How can the principles and concepts applied by visual communication designers be used to assist in exploring and understanding the massive, complex volumes of data now available to Digital Humanities researchers? With its interdisciplinary nature, Digital Humanities is uniquely positioned to develop new paradigms of discovery, partnering the computational analysis power of the digital age with the traditional close study methods of the humanities, and as I propose, with the skills and communication principles employed by visual

communicators. Though detractors often suggest automated analysis will never be an effective way of studying the semantically rich content of the humanities, we can certainly look for patterns from a new altitude, and employ more traditional means to unlock those pattern's secrets. Interpretation, as ever, is a uniquely human activity.

One group of powerful exploration tools at our disposal, a means by which we can use design principles to help display and understand massive data, are visualizations. The presentation of data as abstract visual metaphors offers benefits including increased understanding, a reduction of mental load and faster knowledge digestion. Visualizations exhibit data in an abstract form, allowing our abilities as pattern finding machines to aid in exploring it and hopefully offer insights. A survey of contemporary text-mining visualization tools suggests much of the existing body of text-mining visualizations fail to make adequate use of visual communication theory. The purposeful application of principles and theories of design and perception, and the affordance of opportunities for interaction with the data will lead to more user time spent exploring, and therefore to more hypothesis generation, and hopefully to more knowledge discovery.

Several academics (Burkhard; Thomas & Cook) have proposed the need for creating "a science of visual representations based on cognitive and perceptual principles that can be deployed through engineered, reusable components" (Thomas and Cook; 6). Pre-attentive processing, whereby we perceive some information before having a chance to cognitively process it, the Gestalt theory of perception, which tries to explain how we organize elements in our field of vision, semiotics, the study of signs and symbols, social cognitive theory which describes the process whereby we come to understand the things we see, and colour theory; these would each be

relevant additions to such a science which, if employed intelligently, could help ensure both clarity and engagement are maximized in visualizations of large complex data.

According to MacEachren & Taylor (3), visualizations are employed for one of four purposes; presentation and communication of known and understood information, analysis and confirmation of data, synthesis of multiple data streams for the discovery of relationships, or exploration for the purpose of new knowledge discovery. Large quantities of automated text mining data can benefit from tools that allow interested parties to explore the data in order to develop research hypotheses. When the data is especially large and complex, difficulties arise in meaningfully displaying it all. One of the limiting factors is screen real estate relative to size of the data set and the number of data points being simultaneously displayed. A means of overcoming the difficulties inherent to displaying and interpreting massive amounts of data is to produce tools offering affordances for "zooming."

Zoomable tools are those that offer the user the greatest ability to take advantage of Shneiderman's "Visual Information Seeking Mantra; Overview first, zoom and filter, then details on demand" (1). These tools, which often make use of linked visualizations, best allow the user to navigate through layers of data representing the corpus at different levels of granularity. They are strong enablers of fruitful exploratory research, but because of the complexities involved in navigating and understanding such complex data, they are also the most complicated visualization tools to develop. Schlatter & Levinson propose several meta-principles to help overcome these complications (xiv); the establishment and use of a consistent visual language throughout the visualization, a sensible information hierarchy, so users understand and can locate what is most important to them, and personality, the quality of a piece of visual communication that encourages

users to engage with it. Applied principles of visual communication can significantly contribute to the success of all three of these points as we strive to find ways to display, explore and understand larger than ever collections of data.

The amount of information available to examine, now and in the future, provides us with an opportunity to develop insights into data of which we were previously ignorant. The volume is also such that limiting exploration to experts and professional researchers is likely to continue to result in a massive backlog of unexplored data. As such we would benefit from sharing with as many people as possible new and better tools and methods that help organize, search, and eventually understand the enormous volume of information now available.

Notes

1 · The team involved in the production of the *Dendrogram Viewer*;
John Montague (University of Alberta)
Ryan Chartier (University of Alberta)
Geoffrey Rockwell (University of Alberta)
Stan Ruecker (I.I.T. - Chicago)

2 ·The team involved in the production of the *Topic Model Galaxy Viewer*;
John Montague (University of Alberta)
Ryan Chartier (University of Alberta)
John Simpson (University of Alberta)
Geoffrey Rockwell (University of Alberta)
Stan Ruecker (I.I.T. - Chicago)

Works Cited

Alexander, Erin., et al. "Serendip: Topic Model-Driven Visual Exploration of Text Corpora." *IEEE Conference on Visual Analytics Science and Technology*. Marriott Rive-Gauche. Paris, France. 14 November 2014. Conference Presentation. Web. 30 May, 2015.

"Apophenia." *Wikipedia*. Wikimedia Foundation, n.d. Web. 20 Nov. 2015.

Araya, Agustin A. "The Hidden Side of Visualization." *Techné: Research in Philosophy and Technology*; 7.2 (2 November 2003) : 74-119. Print.

Aubert, M., et al. "Pleistocene Cave Art From Sulawesi, Indonesia." *Nature* [London] 514.7521 (2014): 223-227. GeoRef. Web. 20 Nov. 2015.

Bandura, Albert. *Social Foundations Of Thought And Action : A Social Cognitive Theory*. n.p.: Englewood Cliffs, N.J. : Prentice-Hall, c1986. Web. 20 Nov. 2015.

Bertin, Jacques, and William J. Berg. *Semiology Of Graphics : Diagrams, Networks, Maps*. n.p.: Redlands, Calif. : ESRI Press : Distributed by Ingram Publisher Services, 2011. Web. 20 Nov. 2015.

Bradshaw, James. "Alberta Government Makes $50-million U-turn on Postsecondary Funding Cuts." *The Globe and Mail*. The Globe and Mail, 7 Nov. 2013. Web. 20 Nov. 2015.

Bresciani, Sabrina, and Martin J. Eppler. *The Risks of Visualisation: A Classification of Disadvantages Associated with Graphic Representations of Information*. Lugano, Switzerland. University of Lugano (USI) Institute for Corporate Communication, 2008. Print

Burg, Natalie. "How Big Data Will Help Save Healthcare." *Forbes Magazine*, 10 Nov. 2014. Web. 20 Nov. 2015.

Burkhard, R. "Is It Now Time To Establish Visualization Science As A Scientific Discipline?." *Tenth International Conference On Information Visualisation* (IV'06) (2006): 189. Publisher Provided Full Text Searching File. Web. 20 Nov. 2015.

Burrus, Daniel. "The Internet of Things is Far Bigger Than Anyone Realizes." *Wired*. Web. Accessed 20 May 2015.

Busa, Roberto. "Picture a Man." *Literary and Linguistic Computing* 14(1): 5–9. Oxford University Press. Debrecen, Hungary. 6 July 1998. Web. 13 May 2015.

Carlson, Neil R. *Psychology : The Science Of Behavior*. n.p.: Boston : Allyn & Bacon, c2010. Web. 20 Nov. 2015.

Cao, Nan, et al. "Facetatlas: Multifaceted Visualization For Rich Text Corpora." *IEEE Transactions On Visualization And Computer Graphics* 16.6 (n.d.): 1172-1181. Science Citation Index. Web. 20 Nov. 2015.

Card, Stuart.K., et al. *Readings in Information Visualization: Using Vision to Think*, Morgan Kaufmann Publishers, 1999. San Francisco, CA. Print.

Carrigan, Mark. "Rob Kitchin: "Big Data Should Complement Small Data, Not Replace Them."" *Impact of Social Sciences. The London School of Economics and Political Science*, 27 June 2014. Web. 20 Nov. 2015.

Cawthon, Nick, and Andrew Vande Moere. "Qualities of Perceived Aesthetic in Data Visualisation". 2007. Citeseer. Web. 20 Nov. 2015.

Chaney, Allison and David Blei. "Visualizaing Topic Models." n.p. Association for the Advancement of Artificial Intelligence. 2012. Web. 20 Nov. 2015.

Chevreul, Michel E. *The Principles of Harmony and Contrast of Colors*. 1854. Print.

Chew, Monica and J.Tygar. "Image Recognition CAPTCHAs." *In Proceedings of the 7th International Information Security Conference (ISC 2004)*, Springer, Sep. 2004, 268-279.

Chuang, Jason., Manning, Christopher., and Heer, Jeff. "Termite: Visualization Techniques for Assessing Textual Topic Models." *AVI '12, 21-25 May, 2012. Capri Island, Italy*. Web. 9 Apr. 2015.

Collins, Christopher., Carpendale, Sheelagh., and Penn, Gerald. "Docuburst: Visualizing Document Content Using Language Structure." *Computer Graphics Forum* 28.3 (2009) : 1039–1046. Web. 16 Sep 2014.

Cooke, Russel. “The Importance of Data Visualization.” *Spin Sucks*. 17 September 2014. Web. 16 May 16 2015.

Cukier, Kenneth, and Viktor Mayer-Schoenberger. “Rise Of Big Data: How It’s Changing The Way We Think About The World, The [Notes].” *Foreign Affairs 3* (2013): 28. HeinOnline. Web. 20 Nov. 2015.

Dastani, M. “The Role Of Visual Perception In Data Visualization.” *Journal Of Visual Languages And Computing* 13.6 (n.d.): 601-622. Science Citation Index. Web. 20 Nov. 2015.

Davidson, Cathy., Goldberg, D. Theo. “A manifesto for the Humanities in a Technological Age.” *The Chronicle Review* 50.23 (2004) : B7. Web. 6 May 2015.

DiBiase, David. “Visualization in the Earth Sciences.” *Earth and Mineral Sciences Bulletin* 59:2, 13-18. (1990). Web. 22 Nov. 2015.

Dragland, Åse. “Big Data, for better or worse: 90% of world’s data generated over last two years.” *ScienceDaily. SINTEF*. 22 May 2013. Web. 17 Oct. 2014.

Drucker, Johanna. “Humanities’ ‘Tools’ in Digital Environments 2.” 2008. TS. Proposal for Humanities Approaches. Web. 16 May 2015.

- - -. “Humanities Approaches to Graphical Display.” *Digital Humanities Quarterly*. 5.1 (2011). Web. 6 May 2015.

Eades, Peter. *Handbook of Human Centric Visualization*. 1994. Print.

Eick, Stephen G. "Visualizing Online Activity." *Communications Of The ACM* 44.8 (2001): 45-50. Business Source Complete. Web. 20 Nov. 2015.

Eliassen, Meredith, MSLIS. "Color Theory." *Salem Press Encyclopedia* (2014): Research Starters. Web. 22 Nov. 2015.

Erbacher, Robert. F. "Exemplifying The Inter-Disciplinary Nature Of Visualization Research." *11Th International Conference Information Visualization (IV '07)* (2007): 623. Publisher Provided Full Text Searching File. Web. 20 Nov. 2015.

Fellbaum, Christiane. (Ed.). *WordNet: An Electronic LexicalDatabase.* MIT Press, Cambridge, USA, 1998. IEEE Explore. Web. 23 Nov. 2015.

Gantz, John, and Reinsel, David. "Extracting Value from Chaos" *IDC iView.* Jun 2011 Web. 10 May 2015.

"Google Search Statistics." - Internet Live Stats. n.d. Web. 20 Nov. 2015.

Graham, Shawn., Milligan, Ian., and Weingart, Scott. "[Big Data]" *The Historian's Macroscope: Big Digital History.* Under contract with Imperial College Press. Open Draft Version 2013, Web. May 8, 2015.

Green, Marc. "Toward a Perceptual Science of Multidimensional Data Visualization: Bertin and Beyond" 1998. Web. 16 Sep. 2014.

Grout, James. "Scroll and Codex" *Encyclopaedia Romana*. Essay. n.d. Web. 15 May 2015.

Gubbi, Jayavardhana, et al. "Internet Of Things (Iot): A Vision, Architectural Elements, And Future Directions." *Future Generation Computer Systems 29. Including Special sections: Cyber-enabled Distributed Computing for Ubiquitous Cloud and Network Services & Cloud Computing and Scientific Applications - Big Data, Scalable Analytics, and Beyond* (2013): 1645-1660. ScienceDirect. Web. 20 Nov. 2015.

Günnemann, Nikou. et al. "An Interactive System for Visual Analytics of Dynamic Topic Models." *Datenbank-Spektrum*, 13(3), 2013: 213-223. Web. 23 Nov. 2015.

Havre, Susan., Hetzler, Beth., and Nowell, Lucy. "Themeriver: Visualizing Theme Changes Over Time" *IEEE Symposium On Information Visualization*, 2000 (INFOVIS 2000) (2000): 115. Publisher Provided Full Text Searching File. Web. 20 Nov. 2015.

Healey, Christopher. "Choosing Effective Colours For Data Visualization" *Proceedings Of Seventh Annual IEEE Visualization '96* (1996): 263. Publisher Provided Full Text Searching File. Web. 20 Nov. 2015.

- - - "Effective Visualization of Large Multidimensional Datasets." *Department of Computer Science, University of British Columbia*. 1996. Thesis. Web. 9 Sep. 2014.

Healey, Christopher G., and James T. Enns. "Perception And Painting: A Search For Effective, Engaging Visualizations." *IEEE Computer Graphics And Applications 2* (2002): 10. Academic OneFile. Web. 20 Nov. 2015.

Healey, C. G., K. S. Booth, and J. T. Enns. "High-Speed Visual Estimation Using Preattentive Processing." *ACM Transactions On Computer Human Interaction 3.2* (1996): 107-135. British Library Document Supply Centre Inside Serials & Conference Proceedings. Web. 20 Nov. 2015.

Huang, Weidong. *Handbook of Human Centric Visualization.* New York, NY: Springer, 2014. Print.

"InPhO Topic Explorer." InPhO Topic Explorer. N.p., n.d. Web. 20 Nov. 2015.

Iwata, Tomoharu et al., "Probabilistic Latent Semantic Visualization: Topic Model for Visualizing Documents." *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 363-371. 24–27 Aug. 2008. Las Vegas, Nevada, USA. Web. 10 Sep. 2015

Jockers, Matthew. *Macroanalysis : Digital Methods And Literary History.* Champaign: University of Illinois Press, 2013. Project MUSE. Web. 20 Nov. 2015.

Kelly, Jeff. "Big Data Vendor Revenue And Market Forecast 2012-2017 - Wikibon." *Big Data Vendor Revenue And Market Forecast 2012-2017* - Wikibon. Wikibon, 19 Feb. 2013. Web. 20 Nov. 2015.

Keim, Daniel. "Information Visualization and Visual Data Mining." *IEEE Transactions On Visualization And Computer Graphics*, 7.1 (Jan-March 2002) : 100-107.

Kopetz, Hermann. “Real-Time Systems: Design Principles For Distributed Embedded Applications.” n.p.: New York : Springer, c2011. Web. 20 Nov. 2015.

Krikorian, Raffi. “New Tweets per Second Record, and How! | Twitter Blogs.” *Twitter*, 16 Aug. 2013. Web. 20 Nov. 2015.

Laney, Doug. “3-D Data Management: Controlling Data Volume, Velocity and Variety.” *Gartner Blog Network*. Meta Group inc. Stamford, CT. 2001. Web. 10 Jul. 2014.

Lang, Alexander. “Aesthetics in information visualization.” *Trends in Information Visualization* (2009): 8. Web.11 Aug. 2014.

Liao, Lin, et al. “Behavior recognition in assisted cognition.” *Proceedings The AAAI-04 Workshop on Supervisory Control of Learning and Adaptive Systems*. 2004. Web. 12 May 2014.

Lidwell, William, Kritina Holden, and Jill Butler. *Universal Principles of Design*. Gloucester, MA: Rockport, 2003. Print.

Liu, Alan. “The State Of Digital Humanities: A Report And A Critique.” *Arts And Humanities In Higher Education: An International Journal Of Theory, Research And Practice* 11.1-2 (2012): 8-41. Philosophers Index with Full Text. Web. 20 Nov. 2015.

- - - “Where is Cultural Criticism in the Digital Humanities?” *Debates in the Digital Humanities*. ed. Matthew K. Gold. Minneapolis: University of Minnesota Press, 2012 : 490-509.

MacEachren, Alan M., and D. R. F. Taylor. *Visualization In Modern Cartography*. n.p.: Oxford, U.K. ; New York : Pergamon, 1994., 1994. University of Alberta Library. Web. 20 Nov. 2015.

Manovich, Lev. *Trending: The Promises And The Challenges Of Big Social Data*. n.p.: University of Minnesota Press, 2012. University Press Scholarship Online. Web. 20 Nov. 2015.

Marche, Stephen. "Literature is Not Data; Against Digital Humanities." *Los Angeles Review of Books*. 28 October 2012. Web. Accessed May 16, 2015.

Marcus, Aaron. *Graphic Design For Electronic Documents And User Interfaces*. n.p.: New York, N.Y. : ACM Press, 1992, 1992. University of Alberta Library. Web. 20 Nov. 2015.

Mayer-Schönberger, Viktor, and Kenneth Cukier. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston: Houghton Mifflin Harcourt, 2013. Print.

McAfee, Andrew, and Erik Brynjolfsson. "Big Data: The Management Revolution." *Harvard Business Review*. N.p., 01 Oct. 2012. Web. 20 Nov. 2015.

McCandless, David. "David McCandless: The beauty of data visualization". *TED Talks*. 2010. Web Video. 25 October 2013.

McCormick, Bruce H. *Visualization in Scientific Computing*. New York, NY: ACM SIGGRAPH, 1987. Print.

McGuire, Tim, James Manyika, and Michael Chui. "Why Big Data Is The New Competitive Advantage." *Ivey Business Journal* 76.4 (2012): 1-4. Canadian Reference Centre. Web. 20 Nov. 2015.

Meier, Patrick. "Research Agenda for Visual Analytics." *iRevolutions: From innovation to Revolutions,* June 8 2009. Web. 14 May 2015.

Mendel, Joanne, and Yeager, Jan. "Knowledge Visualization in Design Practice: Exploring the Power of Knowledge Visualization in Problem Solving." *Parsons Journal for Information Mapping*, 2.3. Summer 2010. Web. Apr. 2015.

Mishne, Gilad., and De Rijke, Maarten. "Moodviews: Tools For Blog Mood Analysis." *Technical Report - American Association For Artificial Intelligence* Ss 06/03 (2006): 153-154. British Library Document Supply Centre Inside Serials & Conference Proceedings. Web. 20 Nov. 2015.

Montague, John,. et al. "Seeing the Trees & Understanding the Forest: Visualization Models for Text Mining." 2014. TS. University of Alberta, Edmonton. Print.

Montague, John,. et al. "Exploring Large Datasets with Topic Model Visualizations." 2015. TS. University of Alberta, Edmonton. Print.

Moore, Cathleen M., and Howard Egeth. "Perception Without Attention: Evidence Of Grouping Under Conditions Of Inattention." *Journal Of Experimental Psychology: Human Perception And Performance 2* (1997): 339. Academic OneFile. Web. 20 Nov. 2015.

Moretti, Franco. “The End of the Beginning. A reply to Christopher Prendergast.” *New Left Review*. (2006): 157. Web. 10 Dec. 2014.

Newton, Isaac. *Opticks*. London: Royal Society of London, 1704. Print.

O’Shea, Paul. “Data Visualisation - Anaesthetic or Aesthetic?” *PaulOShea.org*, 2012. Web. 1 Dec. 2013.

Petre, Marian. and Green, Thomas R. G. “Learning To Read Graphics: Some Evidence That ‘Seeing’ An Information Display Is An Acquired Skill.” *Journal Of Visual Languages And Computing 4.1* (1993): 55. British Library Document Supply Centre Inside Serials & Conference Proceedings. Web. 20 Nov. 2015.

Podesta, John, et al. “Big Data: Seizing Opportunities, Preserving Values.” *Report of the Executive Office the President*. Washington, DC. 2014. Web. 9 Sep. 2014.

Repko, Allen F. “Interdisciplinary Research : Process And Theory”. n.p.: Thousand Oaks, Calif. : SAGE Publications, c2012. Web. 20 Nov. 2015.

Rijmenam, Mark., “Why The 3V’s Are Not Sufficient To Describe Big Data.” *Datafloq*, 7 August 2014. Web. 20 May 2015.

Risen, James, and Laura Poitras. “N.S.A. Gathers Data on Social Connections of U.S. Citizens.” *The New York Times*. The New York Times, 28 Sept. 2013. Web. 20 Nov. 2015.

Schlatter, Tania, and Deborah A. Levinson. *Visual Usability: Principles And Practices For Designing Digital Applications*. n.p.: Waltham, Mass. : Morgan Kaufmann, c2013 (Norwood, Mass. : Books24x7.com [generator]), 2013. University of Alberta Library. Web. 20 Nov. 2015.

Schmidt, Benjamin. "Theory First." *Journal of Digital Humanities,* 1.1. 2011. Web. May 14, 2015.

Shermer, Michael. "Michael Shermer: The pattern behind self-deception". *TED Talks*. 2010. Web Video. 21 Nov. 2015.

Shneiderman, Ben. "The Eyes Have It. A Task By Data Type Taxonomy For Information Visualizations." *The Craft Of Information Visualization* (2003): 364-371. ScienceDirect. Web. 20 Nov. 2015.

Smith, John. *A* "New Environment For Literary Analysis." *Perspectives in Computing* 4. 2/3, (1984): 20-31.

Smith, Steve. "By text-mining the classics, UNL professor unearths new literary insights", *UNL Announce*. nd. Web. Jan. 2014.

Snyder, Justin et al. "Topic Models and Metadata for Visualizing Text Corpora." *Proceedings of the NAACL HLT 2013 Demonstration Session*. 5–9. Atlanta, Georgia. 10-12 June 2013. Web. 12 Sep. 2015.

Speer, Robert, et al. "Finding your way in a multi-dimensional semantic space with Luminoso." *Proceedings of the 15th International Conference on Intelligent User Interfaces.* 7-10 February, 2010. Hong Kong, China. 385–388. Web. 6 May 2015.

Sterling, Bruce. *The Epic Struggle of the Internet of Things*. New York: Strelka, 2014. Print.

Svensson, Patrik. "Beyond the Big Tent". *Debates In The Digital Humanities*. Ed. Matthew K. Gold. n.p.: Minneapolis : University of Minnesota Press, c2012. Web. 20 Nov. 2015.

Thomas, James J. *Illuminating the Path: the Research and Development Agenda for Visual Analytics*. Los Alamitos, CA: IEEE, 2005. Print.

Tinker, Miles. A. *Legibility of Print*. Ames/IUSA, Iowa State University Press. 1963. Print.

Tufte, Edward R. *The Visual Display Of Quantitative Information*. n.p.: Cheshire, Conn. : Graphics Press, c2001. Web. 24 May 2015.

"Statistics." YouTube. YouTube, n.d. Web. 20 Nov. 2015.

van Ham, Frank, Marting Wattenberg, and Fernanda Viegas. "Mapping Text With Phrase Nets." *IEEE Transactions On Visualization And Computer Graphics 15.6* (n.d.): 1169-1176. Science Citation Index. Web. 20 Nov. 2015.

Vande Moere, Andrew. "Form Follows Data." *Iris-Isis Publications At Okk-Editions* 9.(2005): 31-40. British Library Document Supply Centre Inside Serials & Conference Proceedings. Web. 20 Nov. 2015.

von Goethe, Johann W. *Theory of Colours*. 1840. Print.

Ware, Colin. *Information Visualization: Perception For Design*, Third Edition. n.p.: Waltham, Mass. : Morgan Kaufmann, c2013 (Norwood, Mass. : Books24x7.com [generator]), 2013. University of Alberta Library Web. 20 Nov. 2015.

Ware, Colin. *Visual Thinking For Design*. n.p.: Burlington, MA : Morgan Kaufmann, 2008. Web. 20 Nov. 2015.

Wasik, Bill. "In the Programmable World, All Our Objects Will Act as One." *Wired*. Conde Nast Digital, 14 May 13. Web. 20 Nov. 2015.

"Writing." Wikipedia. Wikimedia Foundation, n.d. Web. 20 Nov. 2015.

Wolfe, Jeremy M. "Chapter 17: Guidance Of Visual Search By Preattentive Information." *Neurobiology Of Attention* (2005): 101-104. ScienceDirect. Web. 20 Nov. 2015.

Wolfe, Jeremy M., and Todd S. Horowitz. "What Attributes Guide The Deployment Of Visual Attention And How Do They Do It?." *Nature Reviews Neuroscience 6* (2004): 495. Academic OneFile. Web. 20 Nov. 2015.

Zepel, Tara. "Visualization as a Digital Humanities _____ ?" *HASTAC*, 2 May 2013. Web. 20 Nov. 2015.