

# CULTURAL ANALYTICS:

## ANALYSIS AND VISUALIZATION OF LARGE CULTURAL DATA SETS

A proposal from [SOFTWARE STUDIES INITIATIVE](#) @ CALIT2

Prepared by Dr. Lev Manovich  
Director, [Software Studies Initiative @ CALIT2/CRCA](#)  
Professor, [Visual Arts Department @ UCSD](#)  
email: [manovich@ucsd.edu](mailto:manovich@ucsd.edu)  
URL: [www.manovich.net](http://www.manovich.net)

Version: September 30, 2007

### SUMMARY

Can we create **quantitative measures** of cultural innovation? Can we have a **real-time detailed map** of global cultural production and consumption? Can we **visualize flows** of cultural ideas, images, and trends? Can we visually represent how cultural and lifestyle preferences – whether for music, forms, designs, or products – **gradually change over time**?

Today sciences, business, governments and other agencies rely on computer-based analysis and visualization of large data sets and data flows. They employ statistical data analysis, data mining, information visualization, scientific visualization, visual analytics, and simulation. We believe that it is time that we start applying these techniques to cultural data. The large data sets are already here – the result of the digitization efforts by museums, libraries, and companies over the last ten years (think of book scanning by Google and Amazon) and the explosive growth of newly available cultural content on the web. (For instance, as of June 2007, Flickr has 560 million images – together with tags created by users and other metadata automatically logged by Flickr servers).

The envisioned highly detailed interactive visualizations of cultural flows, patterns, and relationships will be based on the analysis of large sets of data comparable in size to data sets used in sciences (i.e. a few or even dozens of terabytes). The data sets will come from two main sources. The first source is media content - games / visual design / music / videos / photos / art / blogs / web pages, etc. In visualizing this content, we will use not only already existing metadata (such as image tags created by the users) but also new metadata which we will generate by analysing the media content (for instance, image processing of Flickr images). The second source is digital traces left when people discuss, create, publish, consume, share, edit, and remix these media.

Visualizations should provide rich information presented in different formats, i.e., graphics, text, numbers, time graphs, etc. This information will be

placed in larger contexts – for instance, geo maps overlaid with economical, sociological and/or historical data.

Visualizations should be designed to take full advantage of the largest gigapixel wall-size displays available today – that are being constructed at CALIT2. Scaled-down versions of these visualizations running on the web or user's PC or smart phone should also be available.



Optional: data is harvested from the web in real time.

Future direction: a user can perform analysis of the data herself using visualization as a starting point (like in GIS); as more processing power becomes available, such analysis eventually would be done in or close to real-time.

## Introduction

This proposal draws on the ideas in the recent [NSF "Cyberinfrastructure Vision"](#) report (2006), [ACLS "Cyberinfrastructure for Humanities and Social Sciences"](#) report (2006), as well as current work in [data mining and visualizing the web and blogosphere](#), web and business analytics, information visualization, scientific visualization, and software art.

The NSF report emphasizes high performance (petascale) computing, data analysis and visualization. The [2006 summer workshop on Cyberinfrastructure for the Humanities, Arts and Social Sciences](#) at the San Diego Supercomputer Center has similarly encouraged the participants to think about research projects on that scale. However, so far I don't know of any project in humanities that has begun to address this challenge using **contemporary cultural data sets** (as opposed to historical ones).

In the present decade our ability to capture, store and analyze data is increasing exponentially, and this growth has already affected many areas of science, media industries, and the patterns of cultural consumption. Think, for instance, of how search has become the interface to global culture, while at the

same time recommendation systems have emerged to help consumers navigate the ever-increasing range of products.

We feel that the ground has been set to start thinking of **culture as data** (including media content and people’s creative and social activities around this content) **that can be mined and visualized**. In other words, if data analysis, data mining, and visualization have been adopted by scientists, businesses, and government agencies as a new way to generate knowledge, let us apply the same approach to understanding culture.



Traffic monitoring, Florida Department of Transportation

Imagine a real-time traffic display (now available in car navigation systems) – except that the display is wall-size, the resolution is thousands of times greater, and the traffic shown is not cars on highways, but real-time cultural flows around the world. Imagine the same wall-sized display divided into multiple frames, each showing different data about cultural, social, and economic news and trends – thus providing a **situational awareness for cultural analysts**.

Imagine the same wall-sized display playing an animation of what looks like an earthquake **simulation** produced on a super-computer – except in this case the “earthquake” is the release of a new version of a popular software or an important consumer electronics product (such as iPhone), the release of a new major movie, or the announcement of an important architectural project. What we are seeing are the effects of such “cultural earthquake” release over time and space. Imagine a wall-sized computer graphic showing **the long tail** of cultural production that allows you to zoom to see each individual product together with rich data about it (à la real estate map on [zillow.com](http://zillow.com)) – while the graph is constantly updated in real-time by pulling data from the web. Imagine a visualization that shows how other people around the world remix new videos created in a fan community, or how a new design software gradually affects the



kinds of forms being imagined today (the way Alias and Maya led to “blobs” in architecture).



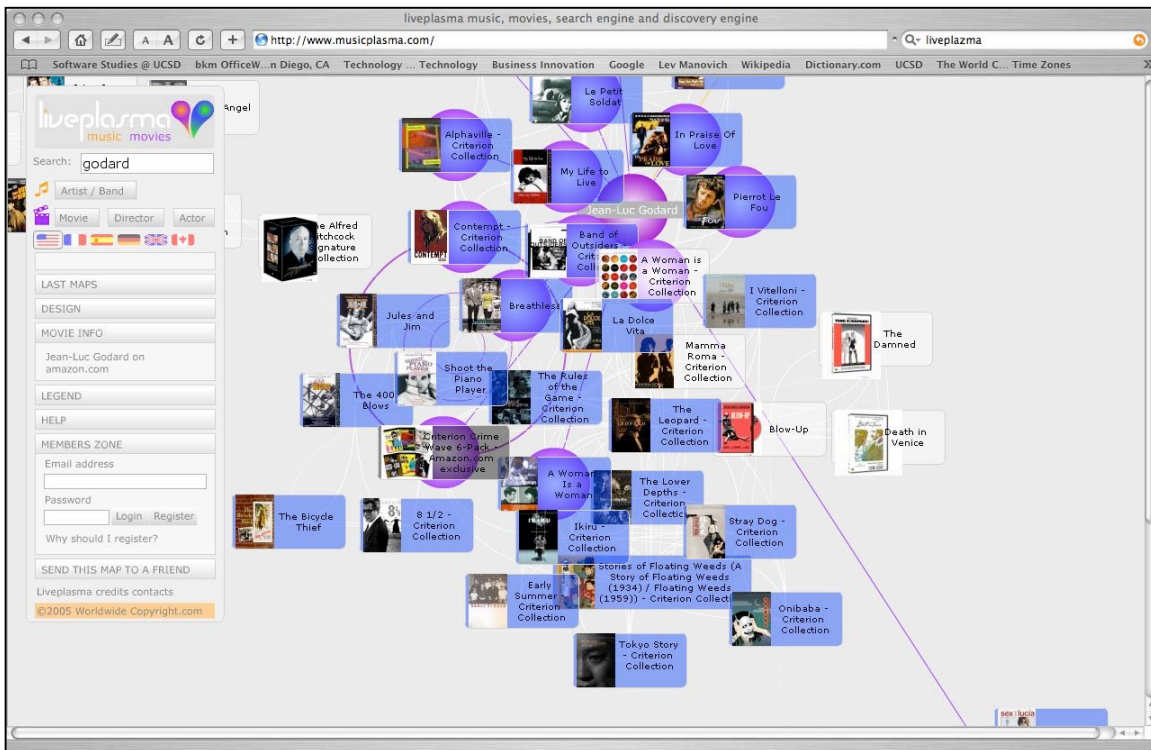
AT&T Global Network Operations Center





Our proposal builds on this work while hoping to overcome what we see as its limitations:

- (A) In comparison to the amounts of data used today in science or being processed by Google, Amazon, and other big Web companies, such projects use very small amount of data;
- (B) As data sources, they usually use media and metadata available on the web (often using API's of Amazon, Google Maps, etc. – for instance, see [liveplasma](#)). In other words, the projects are driven by what data is easily available, rather than by more challenging theoretical questions and research agendas in art history, film and media studies, communication, urban anthropology, sociology, etc., or by currently important cultural issues discussed in media.

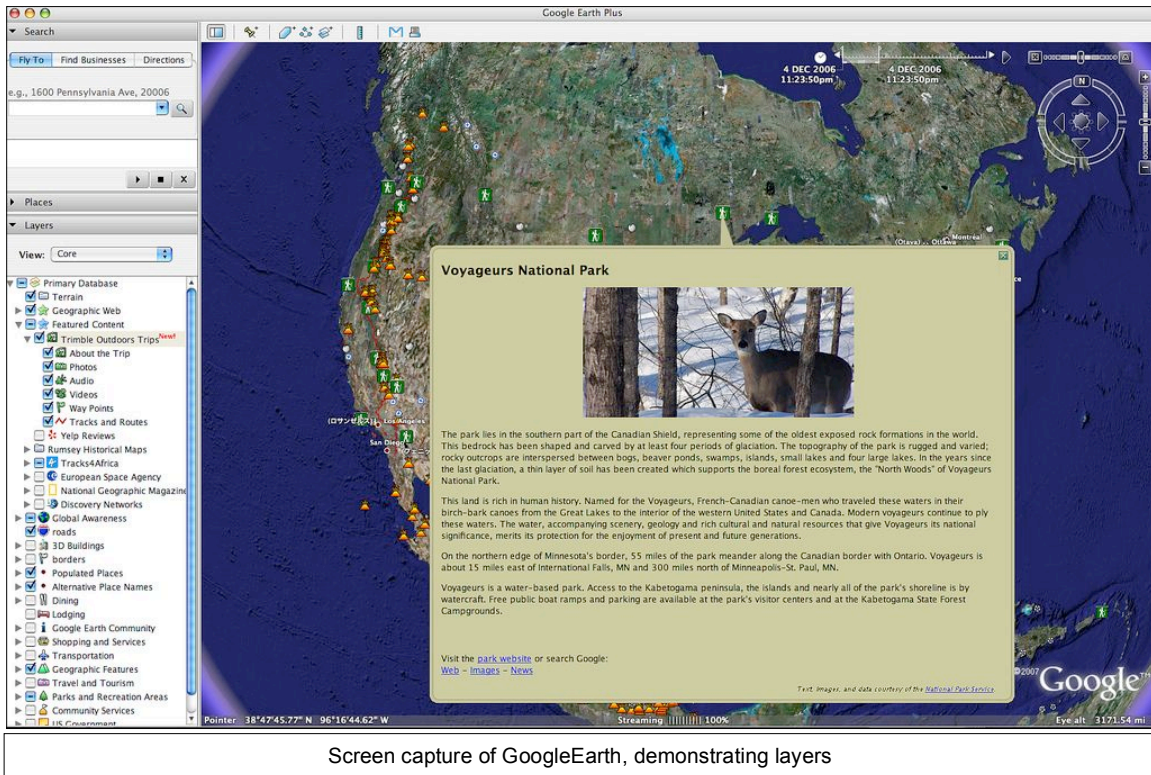


Screen capture of a liveplasma visualization

- (C) They typically rely on existing meta-data, and do not look “inside the data,” i.e. they don’t perform data analysis. For instance, existing visualizations of Flickr data use existing metadata and don’t do any image processing on the images themselves. Similarly, Amazon’s recommendation system uses available metadata about the books and users’ activities on amazon.com; as far as I know, it is not connected to

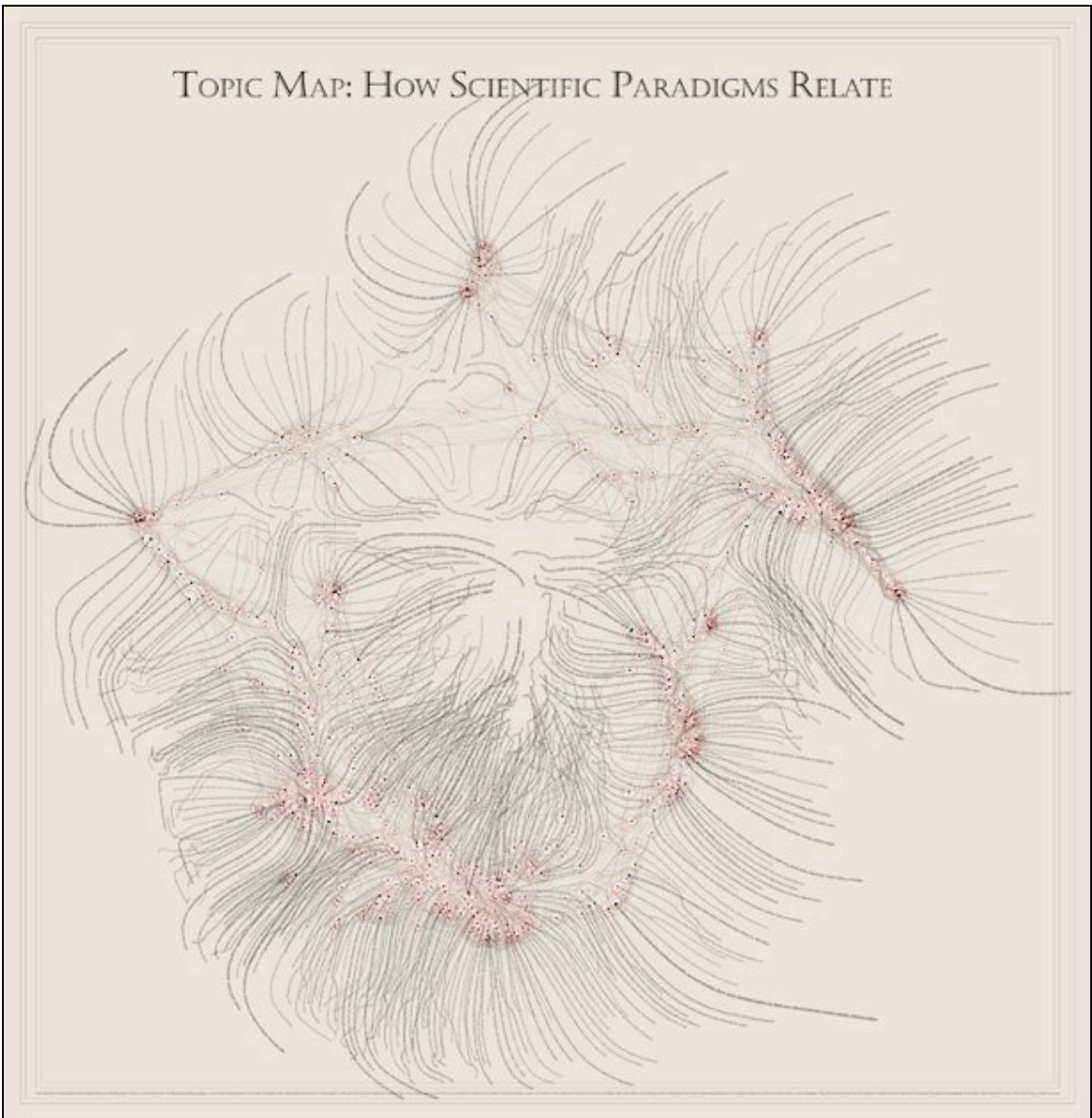
Amazon's software that does analysis of book's contents ("Statistically Improbable Phrases")

- (D) They do not layer many types of data together (as is done in GIS applications - including GIS-type applications for general public such as Google Earth)

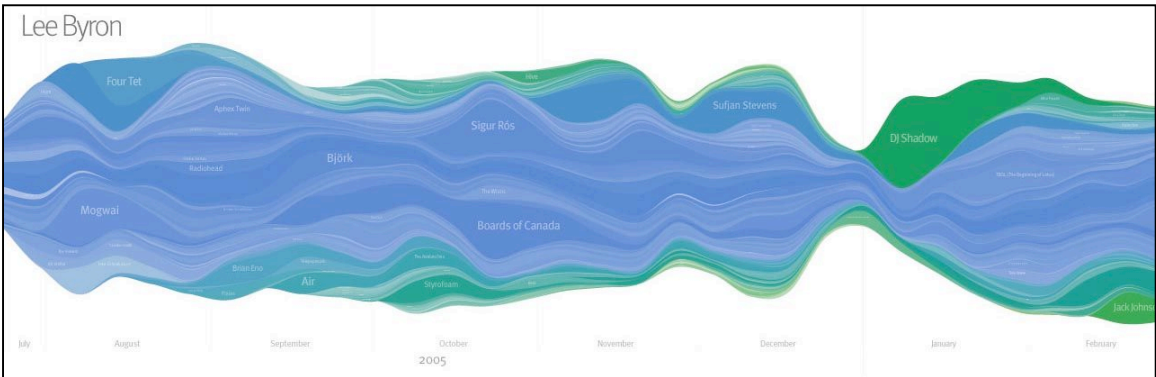


Screen capture of GoogleEarth, demonstrating layers

- (E) Cultural visualizations that show lots of detail often are static (for instance, [Listening History](#), Bradford Paley's [map of science](#)). Visualizations which are dynamic usually have less detail, they don't let the user to drill down into the data, and they rarely use multiple kinds of data. As an example of a functionality and interface we would like to see in a cultural visualization, look at real estate web site [Zillow](#). It allows the user to zoom into the street level view of the US to see estimated housing prices for every block and get detailed information about houses on the market, including time data of how a house's price changed over a 10 year period.



Bradford Paley's "Map of Science"



Lee Byron's "Listening History"



(See [www.visualcomplexity.com](http://www.visualcomplexity.com), [infoaesthetics.com](http://infoaesthetics.com), [Viz4All](http://Viz4All), [mapping science](http://mapping science) for examples of current work on visualization of cultural data and social networks.)

## information aesthetics

FORM FOLLOWS DATA - DATA VISUALIZATION & VISUAL COMMUNICATION

**clusterball wikipedia categories** 19 june 2007

a set of circular graphs of 3 levels of Wikipedia category pages & their interconnections. the parent node is centered in the graph. pages that are linked from this parent node are rendered inside the ball. pages that are linked to the latter (secondary) nodes are rendered on the outer ring. the color of the edges between them represent their depth from the parent node. highly connected groups of pages are clustered clump together, hereby forming topical groups.

Screen capture, infoaesthetics.com

**Filter by:**

- Art (31)
- Biology (36)
- Business Networks (20)
- Computer Systems (22)
- Food Webs (7)
- Internet (26)
- Knowledge Networks (92)
- Multi-Domain Representation (40)
- Music (17)
- Others (32)
- Pattern Recognition (16)
- Social Networks (56)
- Transportation Networks (30)
- World Wide Web (49)

Screen capture, visualcomplexity.com

In [Digital Humanities](#), people are doing statistical analysis of historical texts – but these texts almost always come from the past the results are not been turned into compelling visualizations. Currently, there is work at NCSA to create [platforms to enable mining and visualization of large sets of humanities data](#) (MONK, NORA) – but once again, they are designed to work on texts rather than than visual media. For example, “The NORA project, funded by the Andrew W. Mellon Foundation, aims to produce software for discovering, visualizing, and exploring significant patterns across large collections of full-text humanities resources in existing digital libraries.” Thus **we have a window of opportunity to pioneer the idea of analysis and visualization of large sets of visual and media data – in particular, contemporary data available on the web.**

Since digital humanities work so far has focused on the analysis of canonical historical texts that belong to “high culture” and it has not taken advantage of the wealth of contemporary cultural data available on the web, let us consider another body of work by non-academic actors that focuses on current cultural and social data. The examples are software such as Google Earth and Microsoft’s [Photosynth](#), GIS applications, recommendation systems, and projects such as the [Music Genome Project](#) and [NewsMap](#). These software and projects have their own limitations. The projects which do visualize large amounts of data - for instance, [mapofscience](#) and similar work, [maps of cyberspace](#), Google Earth, GIS applications – do not explicitly focus on culture (by which we mean media content and related activities of people who consume, create, share, and re-mix this content). Other types of applications are designed to run on the web in real time and be accessible via a browser. This limits the amount of data being analyzed and the visual presentation of these data. In other words, these software are not designed to take advantage of the components of next-gen cyberinfrastructure developed at CALIT2 and similar places: display walls with the resolution in hundreds of megapixels, data storage measured in petabytes, petascale computing, new scientific visualization techniques, etc.

The same applies to recently emergent web-based visualization software efforts which aim to democratize visualization by letting users upload their own data sets and then visualize them using provided tools (IBM’s [Many Eyes](#) and [Swivel](#)). (As the public interest in these software projects demonstrates, the idea of visualizations social and cultural data either designed for the general public or created by the public is gaining momentum. Similarly, the numbers of cultural visualization projects - even if they are done on a small scale for now - have been growing exponentially over last few years as demonstrated by [stats at visualcomplexity.com](#). It is also relevant to mention here the [Netflix prize](#) that invites the public to work on data mining algorithms by using a dataset provided by the company.)

While Google, Amazon, Netflix, and other large web companies capture and analyze very large data sets (users’ searches, product purchases, movie rentals, etc.), they have not focused their efforts on visualizing this data or analyzing it to ask larger theoretical questions about how culture works and how

it changes today (at least, according to publicly available information about what these companies are doing.)

Other projects have focused on using digital traces to create detailed records of social activities (MIT Media Lab's [Reality Mining](#); Microsoft's [MyLifeBits](#)). However, they are also not concerned with creating rich visualizations of these data or using the captured data to explore larger cultural and social issues relevant today. The same goes for the recently emerged paradigm of "live blogging" or "life streaming" where people frequently (potentially continuously) send photos, video, text (and potentially other media) to a private or a public site/blog. (Of course, such life blogs are a rich source of data for Cultural Analytics).

### **The New Paradigm: Approaching Culture as Big Data**

We propose a new paradigm for cultural analysis and visualization which, in parallel to web and business analytics, can be called **Cultural Analytics**. (Other terms that can be also used are **Cultural Datamining**, **Culture as Data**, or **Big Humanities**). The projects in this paradigm will share certain features that will make them different from existing work as summarized above:

- (A) **We will focus on visual and media data** (as opposed to text as it is done in digital humanities so far). This type of data requires more storage and computation and expertise in computer graphics and visualization – thus making CALIT2 an ideal place for this work;
- (B) **We will use very large data sets comparable in size to data sets used in most data-intensive science projects** – thus participating in the vision set by NSF;
- (C) While we also want to data-mine cultural data from the past, we will focus on analyzing and mapping **contemporary global cultural** production, consumption, preferences, and patterns. Consequently, we will use the web as a data resource (photos, video, music, books bought, games, etc.) in addition to other data sources. To put this in a different way, while our proposal is informed by the work in digital preservation, digital museum communities, our emphasis is different. In contrast to digital humanities, we want to **focus on current cultural data, with historical data acting as support for the present**. Similarly, in contrast to digital preservation paradigm, our emphasis is not on archiving the past for its own sake but on understanding the present and thinking about the future. In other words, the goal of our work is not preservation for its own sake but analysis – whether human analysts looking at visualizations or software doing analysis of the data.



- (D) In contrast to existing work **we will combine existing metadata with the original analysis of actual data** (feature extraction, segmentation, clustering, data-mining, etc.);
- (E) Taking a clue from visual information systems used today in a variety of industries, we want to construct not only single visualizations but also **suites of visualizations, graphs and maps which show different kinds of information next to each other** - thus taking advantage of the latest large resolution display systems. (For an example of layouts used in existing commercial applications, see Barco's [iCommand](#).) We believe that such displays are necessary if we want to really make sense of rapidly changing global cultural patterns and trends today. In other words, we want to provide **“situational awareness” for “culture analysts”** - ranging from media critics to trend forecasters to decision makers at companies that produce media and lifestyle products worldwide.
- (F) The projects should use the largest displays systems in the world (resolution-wise) – i.e., **the displays currently built at CALIT2's [VIS](#)**



Barco's iCommand

[Lab](#) and [IVL](#). (Of course, they also should be available as desktop and/or a web application in scaled version for wider dissemination). This will allow us to present different kinds of information next to each other in a way that has not been done before. (In 2007, the UCSD division of CALIT2 will complete the largest display system in the world – a wall made from one hundred twenty eight 30 inch monitors resulting in a combined resolution of 500 megapixels. Each separate monitor is driven by its own PC. Therefore, the result is not simply in a large passive

monitor but rather in a very large “visual computer” which can calculate and display at the same time.)

- (G) In contrast to existing analysis and visualizations done by individual designers and commercial companies (such as data-intensive recommendation systems created by Amazon, Netflix, Tivo and other media companies), we will explicitly position our work in relation to humanities, cultural criticism, contemporary cultural debates, as well as NSF’s vision for next-generation cyberinfrastructure, grid computing, and other cultural and scientific agendas;
- (H) Our proposed paradigm of Cultural Analytics can be also connected to the recently established paradigm of [visual analytics](#). We believe that the vision of visual analytics - combining data analysis and data visualization to enable “discovery of the unexpected within massive, dynamically changing information spaces” – is perfectly applicable to cultural data sets.
- (I) The most important difference between our proposal and the existing work is conceptual. We want to use analysis and visualization of large cultural data sets to ask **more challenging questions** than the current generation of visualization projects is able to do. For example: given that U.S. government has recently focused on creating [a better set of metrics for innovation initiatives](#), can we create quantitative measures of cultural innovation around the world (using analysis and visualization of cultural data)? Can we track and visualize the flow of cultural ideas, images, and influences between countries in the last decade – thus providing the first ever data-driven detailed map of how cultural globalization actually works? If we feel that the availability of information on the web makes ideas, forms, images, and other cultural “atoms” travel faster than before, can we track this quantitatively, visualizing how the development of the Web speeded up cultural communications over the last decade?

## How to Do Cultural Analytics

First, get lots of cultural data – preferably, well-structured and with some metadata. Here are some examples: thousands of twentieth century movies already available in digital format (P2P networks and commercial providers); 500,000 art history images collected by [Artstor](#) (funded by Mellon Foundation); millions of hours of video in the [BBC motion gallery](#), 500+ millions of photos on Flickr; records of activity in Massively Multiplayer Online Role-Playing Games (MMORPG) and Second Life; over 500,000 hours (~60 years) of continuous data on daily human behavior captured by MIT’s [Reality Mining](#) project. Ideally, we will get data from one of the leading social media companies such as Google, Flickr, Amazon, Sony, MySpace, etc. since these companies today have more well-

structured data on current cultural interactions (along with the media itself) than any public resource. Alternatively, we can harvest the data from the web ourselves and/or obtain the data from the public collections such as Artstor. And if we want to start with smaller data sets, well-structured media collections are easy to find (for instance, the project [6 Billion Others](#) will soon make available 450 hours of video interviews with appr. 6,000 people from 65 different countries; see [infosthetics.com](#) for more examples of publically available data sets).

The data set(s) can either be directly visualized; alternatively, they can be first statistically analyzed and then visualized. (While at present this analysis may have to be done in advance for very large data sets, in the future increasing computation speed should enable both analysis and visualization to be performed in real time at the user's request.) The visualizations should allow the user to zoom between an overview of all data and the individual data. (For instance, the user should be able to view individual photographs or video clips, read blog entries, etc.) Visualization should have resolution and detail compatible to the most demanding scientific visualization done today. At the same time, we want to use more abstract visual language developed in information visualization and software art fields (for examples, see gallery at [processing.org](#)). Depending on the size of the data set, visualization can be static or animated or interactive. Of course, interactive is best! Ideally, the user will be given controls which allow her/him to change what the dimensions are, filter which data are shown, change graphical style, etc.

The following are some possible scenarios:

Harvest data from the web or get an existing data set from a third party -> visualize using existing metadata

Harvest data from the web or get an existing data set from a third party -> perform data analysis such as such as image processing to generate new metadata -> visualize

Get an existing data set from a third party or harvest data from the web -> perform data analysis such as such as image processing to generate new metadata -> analyze metadata using data mining, other statistical analysis, etc. -> combine with other data sets -> visualize

## **Potential Applications**

Cultural Analytics has the potential to generate new approaches for studying cultural history and contemporary culture. (If slides made possible art history, and if the movie projector and video recorder enabled film studies, what new cultural disciplines may emerge out of the use of visualization and data analysis?)



At the same time, Cultural Analytics can provide a new application area for research in large-scale visualization, data storage, data analysis, and petascale computing as outlined in the previously cited NSF report.

Possible commercial applications: Cultural Analytics should be of interest to providers of media metrics, trend forecasters, social media companies and, in principle, all big media companies and publishers.

Other applications: humanists, social scientists, cultural critics, museums, digital heritage projects, as well as cultural consumers at large. In fact, everybody involved in culture today – from the individual members of a global “cultural class” to governments around the world which are competing in knowledge production and innovation – would be potentially interested in what we want to do – measures of cultural innovations, detailed pictures of global cultural changes, real-time views of global cultural consumption and production.

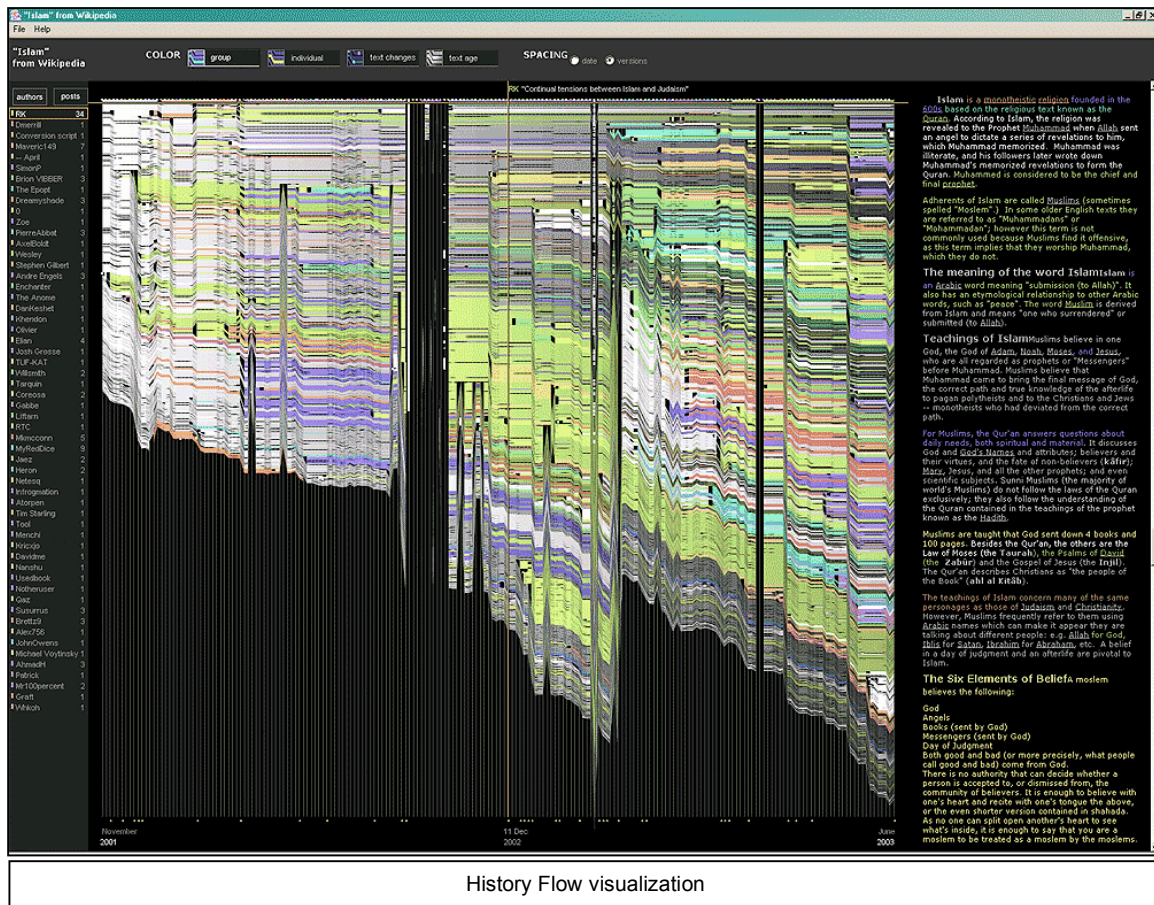
Cultural Analytics projects will also complement the work done at CALIT2's [Center of Interdisciplinary Science for Art, Architecture and Archaeology](#).

### **An Example of Cultural Analytics: Visualizing Cultural Change**

Humanities disciplines, cultural critics, museums, and other cultural institutions usually present culture in terms of self-contained cultural periods. Modern theories of history such as work by Kahn (“scientific paradigms”) and Foucault (“epistemes”) also focus on stable periods rather than transitions. In fact, very little intellectual energy has been spent in the modern period on thinking about how cultural change happens. Perhaps this was appropriate given that until recently the cultural changes of all kinds very usually slow. However, since the beginnings of globalization in the 1990s, not only have these changes accelerated worldwide, but the emphasis on change rather than stability became the key to global business and institutional thinking (expressed in the popularity of terms such as “innovation” and “disruptive change.”)

Recently, software such as [Google Trends](#) and Nelson's [BlogPulse](#), as well as individual projects such as [History Flow](#) and [visualizations of Flickr's activity](#) started to diagram how cultural preferences or particular cultural activities change over time. However, so far these projects / services either use small data sets or they generate very simple graphs.

In order to create the first rich and detailed visualizations of cultural change, we will use data sets representing not one but two or more “cultural periods.” As an example, consider a visualization that shows how Renaissance changed into Baroque in Italian paintings of the 15<sup>th</sup>-17<sup>th</sup> centuries. Such a visualization can solely rely on already available metadata about the data sets: for instance, paintings' titles, dates, sizes, available sociological data about the areas of Italy



History Flow visualization

during the given period, etc. Alternatively, it can combine these metadata with an analysis of the dataset itself – in this case, we would use image processing to do image segmentation, feature extraction, and object classification. This new metadata can also be subjected to statistical analysis.

Other examples of interesting and culturally and economically important data sets:

- (J) The changes in patterns of cultural consumption around the world as a result of globalization;
- (K) The proliferation of styles in popular and dance music over the last few decades;
- (L) Changes in the characteristics of video games, movies, blogs, or any other currently popular cultural form over periods of time.
- (M) Visualizing the changes and development in terms using Wikipedia – comparing how frequently different articles get updated over time, tracking development of new articles.

In the case of visualizing cultural change in the last few decades, the visualizations can also use economic and sociological data about the people/places (many such data sets are now available for free on the web.)

## **An Example of Cultural Analytics: Globalization Viewer**

Real-time (or near real-time) information displays are used in a variety of professional settings: business information dashboards; financial analysis and trading; situational awareness; command and control; visual analytics applications; GIS applications; monitoring of large systems such as factories, telecommunication systems, or space missions; interfaces of vehicles, etc. Yet such displays have never been deployed to present the detailed views of cultural processes. What currently exist are simple applications that can show one kind of cultural activity without any context: photos being uploaded on Flickr in real-time, most discussed themes in blogosphere (technorati.com), popular search terms (the display at Googleplex entrance).

Using the map of earth as the interface, our Globalization Viewer would show real-time (or near real-time) information about the patterns of cultural consumption and production around the world. The information is rich, detailed, and it covers many categories. The user can zoom into particular geographical areas to see what movies are playing there, popular books, music, and other products, the proportions between local and foreign brands, new architectural projects, exhibitions and festivals, topics discussed in local local press and in local blogs, etc. Optional: a spatial interface can be combined with a timeline interface that shows how cultural patterns change over time in different areas (see timeline visualizations of social and economic data produced by [Gapminder](#) - recently bought by Google - for interface examples).

Although such a real-time viewer can be connected to many current cultural discussions and commercial interests, we feel that thinking about it as a visual tool that reveals in detail a how local cultures change due to globalization would be particularly relevant today. By taking advantage of the resolution of CALIT2 wall-size displays, culture around the world for the first time can be visualized as a dynamic process in sufficient detail. The web provides enough data sources to drive such visualizations, therefore is simply a matter of collecting these data from various web sites and displaying them together in the compelling manner.

We believe that such a display will be of interest to a variety of players – from cultural institutions to potentially any company including social media companies operating on the Web.

## **An Example of Cultural Analytics: Virtual World GeoViewer**

(Prepared by Noah Wardrip-Fruin, Assistant Professor, Communication Department, UCSD; Associate Director, Software Studies Initiative [nwf@ucsd.edu](mailto:nwf@ucsd.edu))

Most of the research on virtual worlds (such as massively multiplayer games and worlds like Second Life) falls into two camps: (a) qualitative work with small numbers of users and (b) quantitative work on topics such as world population, activities of users, and so



on. Almost none of the research engages with the fact that these are \*worlds\* which are moved through geographically. It also does not combine them with the kinds of geographic visualizations that have proven powerful for understanding our own world (from popular web services such as Zillow to high end ocean atmosphere science). This project will bring a spatial approach to studying virtual worlds, providing a powerful research tool as well as a compelling visualization for users and the interested public. It has the potential to significantly shift how we see this important emerging cultural phenomenon.

As we imagine the results, users will be able to see a virtual world not just as a geography, but also as software, updated in real time. This allows for the visual traversal and data querying present in applications such as those built with GeoFusion, but we can see much more: What are the users doing and what is the world software doing? This might make it possible, for example, to see how common patterns of leveling behavior emerge in an MMO. The pathways of the fastest-leveling players could be traced and examined, especially in relation to software operations such as instance creation, joining and leaving guilds, item acquisition, etc. This would be of interest to researchers, to the game companies that design and publish MMOs, and also to players eager to discover new ways of engaging the world. Conversely, a researcher doing qualitative work in an virtual world might (with permission) trace the journeys of their informants, with markers for significant events (the first purchase of land in Second Life, the early leveling events in an MMO) looking for patterns of relation with the world not apparent through interviews with the subjects.

We believe the most powerful approach to this sort of work would involve the types of high-resolution tiled displays being created at Calit2. At the same time, if a smaller-scale web version could be made available this might spark broad community interest and conversation. Possible partners include Sony Online Entertainment (who has expressed an interest in working with us) and Linden Labs (who seem to be interested in alternative visualizations of Second Life).