Lev Manovich

How to Follow Global Digital Cultures, or Cultural Analytics for Beginners

From "New Media" to "More Media"

Only fifteen years ago we typically interacted with relatively small bodies of information that were tightly organized in directories, lists and a priori assigned categories. Today we interact with a gigantic, global, not well organized, constantly expanding and changing information cloud in a very different way: we Google it.

The raise of search as the new dominant way for encountering information is one manifestation of the fundamental change in human's information environment.[1] We are living through an exponential explosion in the amounts of data we are generating, capturing, analyzing, visualizing, and storing – including cultural content. On August 25, 2008, Google's software engineers announced on googleblog.blogspot.com that the index of web pages, which Google is computing several times daily, has reached 1 trillion unique URLs.[2] During the same month, YouTube.com reported that users were uploaded 13 hours of new video to the

[1] This article draws on white paper Cultural Analytics that I wrote in May 2007. I am periodically updating this paper. For the latest version, visit http://lab.softwarestudies.com/2008/09/cultural-analytics.html .
[2] http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html .

site every minute.[3] And in November 2008, the number of images housed on Flickr reached 3 billions.[4]

The "information bomb" already described by Paul Virilio in 1998 has not only exploded.[5] It also led to a chain of new explosions that together produced cumulative effects larger than anybody could have anticipated. In 2008 International Data Corporation (IDC) forecasted that by 2011, the digital universe would be 10 times the size it was in 2006. This corresponds to a compound annual growth rate of %60.[6] (Of course, it is possible that the global economic crisis which begun in 2008 may slow this growth – but probably not too much.)

User-generated content is one of the fastest growing parts of this expanding information universe. According to IDC 2008 study, "Approximately 70% of the digital universe is created by individuals."[7] In other words, the size of media created by users competes well with the amounts of data collected and created by computer systems (surveillance systems, sensor-based applications, datacenters supporting "cloud computing," etc.) So if Friedrich Kittler - writing well before the phenomena is "social media" – noted that in a computer universe "literature" (i.e. texts of any kind) consists mostly of computer-generated files, the humans are now catching up.

The exponential growth of a number of both non-professional media producers in 2000s has led to a fundamentally new cultural situation and a challenge to our normal ways of tracking and studying culture. Hundreds of millions of people are routinely creating and sharing cultural content - blogs, photos, videos, map layers,

[3] http://en.wikipedia.org/wiki/YouTube.
[4] http://blog.flickr.net/en/2008/11/03/3-billion/
[5] Paul Virilio. *Information Bomb*. (Original French edition: 1988.) Verso, 2006.
[6] IDC (International Data Corporation). *The Diverse and Exploding Information Universe*. 2008. (2008 research data is available at http://www.emc.com/digital_universe.)
[7] Ibid.

software code, etc. The same hundreds of millions of people engage in online discussions, leave comments and participate in other forms on online social communication. As the number of mobile phones with rich media capabilities is projected to keep growing, this number is only going to increase. In early 2008, there were 2.2 mobile phones in the world; it was projected that this number will become 4 billion by 2010, with main growth coming from China, India, and Africa.

Think about this: the number of images uploaded to Flickr every week today is probably larger than all objects contained in all art museums in the world.

The exponential increase in the numbers of non-professional producers of cultural content has been paralleled by another development that has not been widely discussed. And yet this development is equally important in understanding what culture is today. The rapid growth of professional educational and cultural institutions in many newly globalize countries since the end of the 1990s - along with the instant availability of cultural news over the web and ubiquity of media and design software - has also dramatically increased the number of culture professionals who participate in global cultural production and discussions. Hundreds of thousands of students, artists, designers, musicians have now access to the same ideas, information and tools. As a result, often it is no longer possible to talk about centers and provinces. (In fact, based on my own experiences, I believe the students, culture professionals, and governments in newly globalized countries are often more ready to embrace latest ideas than their equivalents in "old centers" of world culture.)

If you want to see the effects of these dimensions of cultural and digital globalization in action, visit the popular web sites where the professionals and the students working in different areas of media and design upload their portfolios and samples of their work – and note the range of countries from which the authors come from. Here are examples of these sites: xplsv.tv (motion graphics,

animation), coroflot.com (design portfolios from around the world), archinect.com (architecture students projects), infosthetics.com (information visualization projects). For example, when I checked on December 24, 2008, the first three projects in the "artists" list on xplsv.tv came from Cuba, Hungary, and Norway.[8] Similarly, on the same day, the set of entries on the first page of coroflot.com (the site where designers from around the world upload their portfolios; it contained 120,000+ portfolios by the beginning of 2009) revealed a similar global cultural geography. Next to the predictable 20th century Western cultural capitals - New York and Milan – I also found portfolios from Shanghai, Waterloo (Belgium), Bratislava (Slovakia), and Seoul (South Korea).[9]

The companies which manage these sites for professional content usually do not publish detailed statistics about their visitors – but here is another example based on the quantitative data which I do have access to. In the spring of 2008 we have created a web site for our research lab at University of California, San Diego: softwarestudies.com. The web site content follows the genre of "research lab site" so we did not expect many visitors; we also have not done any mass email promotions or other marketing. However, when I examined Google Analytics stats for softwarestudies.com at the end of 2008, I discovered that we had visitors from 100 countries. Every month people from 1000+ cities worldwide check out site.[10] Even more interestingly are the statistics for these cities. During a typical month, no American cities made it into "top ten list" (I am not counting La Jolla which is the location of UCSD where our lab is located). For example, in November 2008, New York occupied 13th place, San Francisco was at 27th place, and Los Angeles was at 42nd place. The "top ten" cities were from Western Europe (Amsterdam, Berlin, Porto), Eastern Europe (Budapest), and South America (Sao Paulo). What

[8] http://xplsv.tv/artists/1/ , accessed December 24, 2008.
[9] coroflot.com, visited December 24, 2008. The number of design portfolios submitted by users to coroflot.com grew from 90, 657 on May 7, 2008 to 120,659 on December 24, 2008.
[10] See http://lab.softwarestudies.com/2008/11/softbook.html .

is equally interesting is the list of visitors per city followed a classical “long tail” curve. There was no sharp break anymore between “old world” and “new world,” or between “centers” and “provinces.” (See softwarestudies.com/softbook for more complete statistics.)

All these explosions which took place since the late 1990s – non-professionals creating and sharing online cultural content, culture professionals in newly globalized countries, students in Eastern Europe, Asia and South America who can follow and participate in global cultural processes via the web and free communication tools (email, Skype, etc) – redefined what culture is.

Before, cultural theorists and historians could generate theories and histories based on small data sets (for instance, "classical Hollywood cinema," "Italian Renaissance," etc.) But how can we track "global digital cultures" with their billions of cultural objects, and hundreds of millions of contributors? Before you could write about culture by following what was going on in a small number of world capitals and schools. But how can we follow the developments in tens of thousands of cities and educational institutions?

Introducing Cultural Analytics

The ubiquity of computers, digital media software, consumer electronics, and computer networks led to the exponential rise in the numbers of cultural producers worldwide and the media they create – making it very difficult, if not impossible, to understand global cultural developments and dynamics in any substantial details using 20th century theoretical tools and methods. But what if we can we use the

same developments – computers, software, and availability of massive amounts of "born digital" cultural content – to track global cultural processes in ways impossible with traditional tools?

To investigate these questions – as well as to understand how the ubiquity of software tools for culture creation and sharing changes what "culture" is theoretically and practically – in 2007 we established Software Studies Initiative (softwarestudies.com). Our lab is located at the campus of University of California, San Diego (UCSD) and it housed inside one of the largest IT research centers in the U.S. - California Institute for Telecommunications and Information (www.calit2.net). Together with the researchers and students working in our lab, we have been developing a new paradigm for the study, teaching and public presentation of cultural artifacts, dynamics, and flows. We call this paradigm **Cultural Analytics**.

Today sciences, business, governments and other agencies rely on computer-based quantitative analysis and interactive visualization of large data sets and data flows. They employ statistical data analysis, data mining, information visualization, scientific visualization, visual analytics, simulation and other computer-based techniques. Our goal is start systematically applying these techniques to the analysis of contemporary 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.

We believe that a systematic use of large-scale computational analysis and interactive visualization of cultural patterns will become a major trend in cultural criticism and culture industries in the coming decades. What will happen when humanists start using interactive visualizations as a standard tool in their work, the

way many scientists do already? If slides made possible art history, and if a movie projector and video recorder enabled film studies, what new cultural disciplines may emerge out of the use of interactive visualization and data analysis of large cultural data sets?

From Culture (few) to Cultural Data (many)

In April 2008, exactly one year later we founded Software Studies Initiative, NEH (National Endowment for Humanities, the main federal agency in the U.S. which provides grants for humanities research) announced a new "Humanities High-Performance Computing" (HHPC) initiative that is based on the similar insight:

> Just as the sciences have, over time, begun to tap the enormous potential of High-Performance Computing, the humanities are beginning to as well. Humanities scholars often deal with large sets of unstructured data. This might take the form of historical newspapers, books, election data, archaeological fragments, audio or video contents, or a host of others. HHPC offers the humanist opportunities to sort through, mine, and better understand and visualize this data."[11]

In describing the rationale for Humanities High-Performance Computing program, the officers at NEH start with the **availability of high-performance computers** that are already common in the sciences and industry. In January 2009, NEH

11 http://www.neh.gov/ODH/ResourceLibrary/HumanitiesHighPerformanceComputing/tabid/62/Default.aspx .

together with NSF (National Science Foundation) has annonced another program Digging Into Data which has articulated their vision in more detail. This time the program statement put more emphasis on the **wide availability of cultural content (**both contemporary and historical) **in digital form** as the reason for begin applying data analysis and visualization to “cultural data.”:

> With books, newspapers, journals, films, artworks, and sound recordings being digitized on a massive scale, it is possible to apply data analysis techniques to large collections of diverse cultural heritage resources as well as scientific data. How might these techniques help scholars use these materials to ask new questions about and gain new insights into our world?

We fully share the vision put forward by NEH Digtal Humanities. Massive amounts of cultural content and high-speed computers go well together – without the latter, it would be very time consuming to analyze petabytes of data. However, as we discovered in our lab, even with small cultural data sets consisting from hundreds, dozens or even only a few objects it is already viable to do Cultural Analytics: that is, to quantitatively analyze the structure of these objects and visualize the results revealing the patterns which lie below the unaided capacities of human perception and cognition.

Since Cultural Analytics aims to take advantage of the exponential increase in the amounts of digital content since the middle of the 1990s, it will be useful to establish taxonomy for the different types of this content. Such taxonomy may guide design of research studies as well as be used to group these studies once they start multiply.

To begin with, we have vast amounts of **media content** in digital form – games, visual design, music, video, photos, visual art, blogs, web pages. This content can be further broken down into a few categories. Currently, the proportion of "**born digital**" media is increasing; however, people also continue to create analog media (for instance, when they shoot on film), which is later digitized.

We can further differentiate between different types of "born digital" media. Some of this media is explicitly made for the web: for example, blogs, web sites, layers created by users for Google Earth an Googe maps. But we also now find online massive amounts of "born digital" content (photography, video, music) which until the advent of "social media" was not intended to be seen by people worldwide – but which now ends up online at social media sites (Flickr, YouTube, etc.) To differentiate between these two types, we may refer to the first category as "**web native**," or "web intended." The second category can be then called "digital media proper."

As I already noted, YouTube, Flickr, and other social media sites aimed at average people are paralled by more **specialized sites which serve professional and semi-professional users**: xplsv.tv, coroflot.com, archinect.com, modelmayhem.com, deviantart.com, etc.[12] Housing projects and portfolios by hundreds of thousands of artists, media designers, and other cultural professionals, these web sites provide a live shapshot of contemporary global cultural production and sensibility - thus offering a promise of being able to analyze the global cultural trends with the level of detail unthinkable previously. For instance, as of August 20008, deviantart.com has eight million members, 62+

[12] The web sites aimed at non-professionals such as Flickr.com, YouTube.com and Vimeo.com also contain large amounts of media created media professionals and students: photography portfolio, independent films, illustrations and design, etc. Often the professionals create their own groups – which makes it easier for us to find their work on these general-purpose sites. However, the sites specifically aimed at the professionals also often feature CVs, descriptions of projects, and other information not available on general social media sites.

million submissions, and was receiving 80,000 submissions per day.[13] Importantly, in addition to the standard "professional" and "pro-ams" categories, these sites also house the content of people who are just starting out and/or are currently "pro-ams" but who aspire to be full-time professionals. I think that the portfolios (or "ports" as they are sometimes called today) of these "**aspirational non-professionals**" are particularly significant if we want to study contemporary cultural stereotypes and conventions since, in aiming to create "professional" projects and portfolios, people often inadvertently expose the codes and the templates used in the industry in a very clear way.

Another important source of contemporary cultural content – and at the same time, a window into yet another cultural world different from non-professional users and aspiring professionals - are the **web sites and wikis created by faculty** teaching in creative disciplines to post and discuss their class assignments. (Although I don't have direct statistics on how many sites and wikis for classes are out there, here is one indication: a popular wiki creation software pbwiki.com has been used by 250,000 educators.[14]) These sites often contain **student projects** – which provides yet another interesting source of content.

Finally, beyond class web sites, the sites for professionals, aspiring professionals, and non-professionals, and other centralized content repositories, we have **millions of web sites and blogs by individual cultural creators and creative industry companies**. Regardless of the industry category and the type of content people and companies produce, it is now taken for granted that you need to have a web presence with your demo reel and/or portfolio, descriptions of particular projects, a CV, and so on. All this information can be potentially used to do something that previously was un-imaginable: to create dynamic (i.e. changing in

[13] http://en.wikipedia.org/wiki/DeviantArt .
[14] http://pbwiki.com/academic.wiki, accessed December 26, 2008.

time) maps of global cultural developments that reflect activities, aspirations, and cultural preferences of millions of creators.

A significant part of the available media content in digital form was originally created in electronic or physical media and has been digitized since the middle of the 1990s. We can call such content "**born analog**." But it is crucial to remember that what has been digitized in many cases are only the canonical works, i.e. a tiny part of culture deemed to be significant by our cultural institutions. What remains outside of the digital universe is the rest: provincial nineteen century newspapers sitting in some small library somewhere; millions of paintings in tens of thousands of small museums in small cities around the world; millions of thousands of specialized magazines in all kinds of fields and areas which no longer even exist; millions of home moves…

This creates a problem for Cultural Analytics, which has a potential to map everything that remains outside the canon – to begin generating "art history without great names." We want to understand not only the exceptional but also the typical; not only the few "cultural sentences spoken by a few "great man" but the patterns in all cultural sentences spoken by everybody else; in short, what is outside a few great museums rather than what is inside and what has been already extensively discussed too many times. To do this, we will need as much of previous culture in digital form as possible. However, what is digitally available is surprisingly little.

Here is an example from our research. We were interested in the following question: what did people actually painted around the world in 1930 – outside of a few "isms" and a few dozen artists who entered the Western art historical canon?

We did a search on artstor.org which at the time of this writing contains close to one million images of art, architecture and design which come from many important US museum and collections, as well as 200,000+ slide library of University of California, San Diego where our lab is located. (This set which at present is the largest single collection in artstor is interesting in that it reflects the biases of art history as it was taught over a few decades when color slides were the main media for teaching and studying art.) To collect the images of artworks that are outside of the usual Western art historical canon, we excluded from the search Western Europe and North America. This left the rest of the world: Eastern Europe, South-East Asia, East Asia, West Asia, Oceania, Central America, South America, etc. When we searched for paintings done in these parts of the world in 1930, we only found a few dozen images. This highly uneven distribution of cultural samples is not due to Artstor since it does not digitize images itself – it only makes available images submitted to its by museums and other cultural institutions. So what the results of our search reflect is what museums collect and what they think should be digitized first. In other words, a number of major US collections and a slide library of a major research university (which now has a large proportion of Asian students) together contain only a few dozen paintings done outside the West in 1930 which got digitized. In contrast, searching for Picasso returned around 700 images. If this example if any indication, digital depositories may be amplifying the already existed biases and filters of modern cultural canons. Instead of transforming the "top forty" into "the long tail," digitization can be producing the opposite effect.

Media content in digital form is not the only type of data that we can analyze quantitatively to potentially reveal new cultural patterns. Computers also allow us to capture and subsequently analyze many dimensions of human cultural activities that could not be recorded before. Any cultural activity – surfing the web,

playing a game, etc. - which passes through a computer or a computer-based media device leaves traces: keystroke presses, cursor movements and other screen activity, controller positions (think of We controller), and so on. Combined with camera, a microphone, and other capture technologies, computers can also capture other dimensions of human behavior such as body and eye movements and speech. And web servers log yet other types of information: which pages the users visited, how much time they spend on each page, which files they downloaded, and so on. (In this respect, Google Analytics that processes and organizes this information provided a direct inspiration for the idea of Cultural Analytics.

Of course, in addition to all this information which can be captured automatically, the rise of social media since 2005 created a new social environment where people voluntarily reveal their cultural choices and preferences: rating books, movies, blog posts, software, voting for their favorites, etc. Even importantly, people discuss and debate their cultural preferences, ideas and perceptions online. They comment on Flickr photographs, post their opinions about books on amazon.com, critique movies on rottentomatoes.com, review products on epinions.com, and enthusiastically debate, argue, agree and disagree with each other on numerous social media sites, fan sites, forums, groups, and mailing lists. All these conversations, discussions and reflections which before were either invisible or simply could not take place on the same scale are now taking place in public.

To summarize this discussion: because of digitization efforts since the middle of the 1990s, and because the significant (and constantly growing) percentage of all cultural and social activities passes through, or takes place on the web or networked media devices (mobile phones, game platforms, etc.), we now have access unprecedented amounts of both “cultural data” (cultural artifacts

themselves), and “data about culture.” All this data can be grouped into three broad conceptual categories:

- Cultural artifacts (“born digital” or digitized).
- Data about people’ interactions with digital media (automatically captured by computers or computer-based media devices)
- Online discourse around (or accompanying) cultural activities, cultural objects, and creation process voluntarily created by people.

There are other ways to divide this recently emerged cultural data universe. For example, we can also make a distinction between “cultural data” and “cultural information”:

- **Cultural data**: photos, art, music, design, architecture, films, motion graphics, games, web sites - i.e., actual cultural artifacts which are either born digital, or are represented through digital media (for examples, photos of architecture).
- **Cultural information**: cultural news and reviews published on the web (web sites, blogs) – i.e., a kind of “extended metadata” about these artifacts.

Another important distinction, which is useful to establish, has to do with the relationships between the original cultural artifact/activity and its digital representation:

- “Born digital” artifacts: representation = original.
- Digitized artifacts that originated in other media - therefore, their representation in digital form may not contain all the original

information. For example, digital images of paintings available in online repositories and museum databases normally do not fully show their 3D texture. (This information can be captured with 3D scanning technologies – but this is not commonly done at this moment.).

- Cultural experiences (experiencing theatre, dance, performance, architecture and space design; interacting with products; playing video games; interacting with locative media applications on a GPS enabled mobile device) where the properties of material/media objects that we can record and analyze is only one part of an experience. For example, in the case of spatial experiences, architectural plans will only tell us a part of a story; we may also want to use video and motion capture of people interacting with the spaces, and other information.

The rapid explosion of "born digital" data has not passed unnoticed. In fact, the web companies themselves have played an important role in making it happen so they can benefit from it economically. Not surprisingly, out of the different categories of cultural data, born digital data is already been exploited most aggressively (because it is the easiest to access and collect), followed by digitized content. Google and other search engines analyze billions of web pages and the links between them to make their search algorithms run. Nielsen Blogpulse mines 100+ million blogs to detect trends in what people are saying about particular brands, products and other topics its clients are interested in.[15] Amazon.com analyzes the contents of the books it sells to calculate "Statistically Improbable Phrases" used to identify unique parts of the books.[16]

[15] "BlogPulse Reaches 100 Million Mark" < http://blog.blogpulse.com/archives/000796.html> .
[16] http://en.wikipedia.org/wiki/Statistically_Improbable_Phrases .

In terms of media types, today text receives most attention - because language is discrete and because the theoretical paradigms to describe it (linguistics, computational linguistics, discourse analysis, etc.) have already been fully developed before the explosion of "web native" text universe. Another type of cultural media, which is also starting to be systematically subjected to computer analysis in large quantities, is music. (This is also made possible by the fact that Western music used formal notation systems for a very long time.) A number of online music search engines and Internet radio stations use computation analysis to find particular songs. (Examples: Musipedia, Shazam, and other applications which use acoustic fingerprinting.[17]) In comparison, other types of media and content receive much less attention.

If we are interested in analyzing cultural patterns in other media besides text and sound, and also in asking larger theoretical questions about cultures (as opposed to more narrow pragmatic questions asked in professional fields such as web mining or quantitative marketing research – for instance, identifying how consumers perceive different brands in a particular market segment[18]), we need to adopt a broader perspective. Firstly, we need to develop techniques to analyze and visualize the patterns in different forms of cultural media - movies, cartoons, motion graphics, photography, video games, web sites, product and graphic design, architecture, etc. Second, while we can certainly take advantage of the "web native" cultural content, we should also work with other categories that I listed above ("digitized artifacts which originated in other media"; "cultural experiences.") Thirdly, we should be self-reflective. We need to think about the consequences of thinking of culture as data and of computers as the analytical tools: what is left outside, what types of analysis and questions get privileged, and

[17] http://en.wikipedia.org/wiki/Acoustic_fingerprint
[18] http://en.wikipedia.org/wiki/Perceptual_mapping .

so on. This self-reflection should be part of any Cultural Analytics study. These three points guide our Cultural Analytics research.

Cultural Image Processing

Cultural Analytics is thinkable and possible because of three developments: digitization of cultural assets and the rise of web and social media; work in computer science; and the rise of a number of fields which use computers to create new ways of representing and interacting with data. The two related fields of computer science - image processing and computer vision - provide us with the variety of techniques to automatically analyze visual media. The fields of science visualization, information visualization, media design, and digital art provide us with the techniques to visually represent patterns in data and interactively explore this data.

While people in digital humanities have been using statistical techniques to explore patterns in literary text for a long time, I believe that we are the first lab to start systematically using image processing and computer vision for automatic analysis of visual media in the humanities contest. This is what separates us from 20th century humanities disciplines that focus on visual media (art history, film studies, cultural studies) and also 20th century paradigms for quantitative media research developed within social sciences such as quantitative communication studies and certain works in sociology of culture. Similarly, while artists, designers and computer scientists have already created a number of projects to visualize cultural media, the existing projects that I am aware of rely on existing metadata such as Flickr community-contributed tags[19]. In other words, they use information

[19] These projects can be found at visualcomplexity.org and infosthetics.com.

about visual media – creation date, author name, tags, favorites, etc. – and do not analyze the media itself.

In contrast, Cultural Analytics uses image processing and computer vision techniques to automatically analyze large sets of visual cultural objects to generate numerical descriptions of their structure and content. These numerical descriptions can be then graphed and also analyzed statistically.

While digital media authoring programs such as Photoshop and After Effects incorporate certain image processing techniques such as blur, sharpen, and edge detecting filters, motion tracking, and so on, there are hundreds of other features that can be automatically extracted from still and moving images. Most importantly, while Photoshop and other media applications internally measure properties of images and video in order to change them - blurring, sharpening, changing contrast and colors, etc. – at this time they do not make available to users the results of these measurements. So while we can use Photoshop to highlight some dimensions of image structure (for instance, reducing an image to its edge), we can't perform more systematic analysis.

To do this, we need to turn to more specialized image processing software such as open source imageJ which has been developed for live sciences applications and which we have been using and extending in our lab. MATLAB, popular software for numerical analysis, provides many image processing applications. There are also specialized software libraries of image processing functions such as openCV. A number of high-language programming languages created by artists and designers in 2000s such as Processing and openFrameworks also provide some image processing functions. Finally, many more techniques are described in computer science publications.

While certain common techniques can be used without the knowledge of computer programming and statistics, many others require knowledge of C or Java programming. Which of the algorithms can be particularly useful for cultural analysis and visualization? Can we create (relatively) easy-to-use tools which will allow non-technical users to perform automatic analysis of visual media?

These are the questions we are currently investigating. As we are gradually discover, in spite of the fact that the fields of image processing and computer vision have existed now for approximately five decades, the analysis of cultural media often requires development of new techniques that do not yet exist.

To summarize: the key idea of Cultural Analytics is the use of computers to **automatically analyze cultural artifacts in visual media extracting large numbers of features which characterize their structure and content**. For example, in the case of a visual image, we can analyze its grayscale and color characteristics, orientations of lines, texture, composition, and so on. Therefore, we can also use another term to refer to our research method – **Quantitative Cultural Analysis** (QCA).

While we are interested in both content and structure of cultural artifacts, at present automatic analysis of structure is much further developed than the analysis of content. For example, we can ask computers to automatically measure gray tone values of each frame in a feature film, to detect shot boundaries, to analyze motion in every shot, to calculate how color palette changes throughout the film, and so on. However, if we want to annotate film's content – writing down what kind of space we see in each shot, what kinds of interactions between characters are taking place, the topics of their conversations, etc., the automatic techniques to do this are more complex (i.e., they are not available in software such as MAT LAB and imageJ) and less reliable. For many types of content

analysis, at present the best way to is annotate media manually – which is obviously quite time consuming for large data sets. In the time it will take one person to produce such annotations for the content of one movie, we can use computers to automatically analyze the structure of many thousands of movies. Therefore, we started developing Cultural Analytics by developing techniques for the analysis and visualization of structures of individual cultural artifacts and large sets of such artifacts - with the idea that once we develop these techniques we will gradually move into automatic analysis of content.

Deep Search

In November 2008 we received a grant that gives us 300,000 hr of computing time on US Department of Energy supercomputers. This is enough to analyze millions of still images and video – art, design, street fashion, feature films, anime series, etc. This scale of data is matched by the size of visual displays that we are using in our work. As I already mentioned, we are located inside one of the leading IT research centers in the U.S. - California Institute for Telecommunication and Information Technology (Calit2). This allows us to take advantage of the next-generation visual technologies - such as HIperSpace, currently one of the highest resolution displays for scientific visualization and visual analytics applications in the world. (Resolution: 35,640 by 8,000 pixels. Size: 9.7m x 2.3m.)

One of the directions we are planning to pursue in the future is the development of visual systems that would allow us to follow global cultural dynamics in real-time. Imagine a real-time traffic display (a la 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 windows, each showing different real-time and historical 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 popular software, the announcement of an important architectural project, or any other important cultural event. What we are seeing are the effects of such "cultural earthquake" 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) – 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 word 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 a new language in architecture). These are the kinds of tools we want to create to enable new type of cultural criticism and analysis appropriate for the era of cultural globalization and user-generated media: three hundred digital art departments in China alone; approximately 10,000 new users uploading their professional design portfolios on coroflort.com every month; billions of blogs, user-generated photographs and videos; and other cultural expressions which are similarly now created at a scale unthinkable only ten years ago.

To conclude, I would like to come back to my opening point – the rise of search as a new dominant mode for interacting with information. As I mentioned, this development is just one of many consequence of the dramatic and rapid in the scale of information and content being produced which we experienced since the middle of the 1990s. To serve the users search results, Google, Yahoo, and other search engine analyze many different types of data – including both metadata of particular web pages (so-called "meta elements") and their content. (According to

Google, its search engine algorithm uses more than 200 input types.[20]) However, just as Photoshop and other commercial content creating software do not expose to users the features of images or videos they are internally measuring, Google and Yahoo do not reveal the measurements of web pages they analyze – they only serve their conclusions (which sites best fit the search string) which their propriety algorithms generate by combining these measures. In contrast, the goal of cultural Analytics is to enable what we may call "deep cultural search" – give users the open-source tools so they themselves can analyze any type of cultural content in detail and use the results of this analysis in new ways.

[March 2009]

[20] http://www.google.com/corporate/tech.html