UNIVERSITY OF PITTSBURGH SCHOOL OF INFORMATION SCIENCES
Fall 2012

LIS 3970 Seminars Special Topics: Digital Scholarship
Combined: INFSCI 2965 & INFSCI 3150

Instructor: Stephen Griffin, Visiting Professor, Mellon Cyberscholar

Office Number and Telephone: SIS 620; 412 624-9315

Office Hours: By appointment or anytime by e-mail

E-mail: sgriffin@pitt.edu

Homepage: http://www.ischool.pitt.edu/people/griffin.php

Class Sessions: Tuesdays, 12-2:50 PM

Course Objectives

Course Rationale. Contemporary research and scholarship is increasingly characterized by the use large-scale datasets and computationally intensive tasks. Vast amounts of data are used by scholars to better map the cosmos, build more accurate earth system models, examine in finer detail the structures of living organisms, and gain new insights into the behaviors of societies and individuals in a complex world. Similarly, humanists are rapidly integrating newly digitized corpora, digital representation of cultural artifacts and spatial and temporal indexed data into their scholarly endeavors.

Purposes of This Course. This course will chart the development of digital scholarship from the beginning of the use of models and abstracted forms to conceptualize and represent knowledge and physical phenomena to state-of-the-art projects today that are transforming the nature of inquiry in many disciplinary domains. The course will be descriptive in nature. The goal will be to understand digital scholarship in terms of high-level methodological approaches and conceptual frameworks as well as to examine the technological, academic and social contexts that underpin successful endeavors. Case studies of exemplary state-of-the-art projects will be the vehicle for exploring the ways in which scholars, using internet-based open data, technologies and tools are dramatically expanding the problem space of domain scholarship in many areas and creating new methods for analysis of information and presentation of research results. A focus will also be on the natural role of collaboration and communication in digital scholarship. Class assignments will be tailored for each student to meet their interests and support their career goals.

Course Requirements

Form of weekly assignments: Students will be required to produce each week a 1-2 page "document" that can and should include actively formatted materials such as links, videos, images, etc. drawing on the materials presented and discussed in the class session. Students can also submit a Powerpoint presentation, Prezi, or new types of information presentation tools and visualizations. A portion of each class will be devoted to asking individual students to present their work so that the class can discuss it openly and explore various aspects of it. The "document" should elucidate as many of the following attributes as possible:

- Researcher(s)/location/disciplinary domain(s)
- Driving question/hypothesis/motivation
- data and computational resources used
- analytical methods applied
- workflow considerations
- results – basic form
- validation and testing
- results – refined presentation
- dissemination
- efforts taken to enable reuse of data/resources
- related work (what did this work build on and/or add to)
- impact and importance

Many of these issues will the focus of the class discussion of the work.

Final Project. Each member of the class is to complete a larger final project that either is an individual work of digital scholarship, or that examines in detail an outstanding example of digital scholarship. Important in this regard is how creatively and effectively the student presents their work.

Course Syllabus

Digital Scholarship – Transforming Inquiry through Data and Computation

Session 1: Data-Intensive Scholarship and Research Today
This introductory session will lay out the broad landscape of data-intensive research and scholarship today and describe a set of exemplary projects at the cutting-edge. Projects selected are meant to demonstrate the power, scope and complexity (technological and social) of data-intensive scholarship and will include work in the sciences, humanities and interdisciplinary research areas.

Projects to be discussed include:
National Virtual Observatory
http://virtualobservatory.org/whatis/history.aspx
Center for Embedded Network Sensing
http://research.cens.ucla.edu/

The Mesur Project
http://mesur.informatics.indiana.edu/
Digital Michelangelo Project*
http://graphics.stanford.edu/projects/mich/
Venetus A Manuscript of Homer's Iliad
www.vis.uky.edu
http://vis.uky.edu/vis-media/imaging-the-iliad/
Electronic Cultural Atlas Initiative
www.ecai.org
Rome Re-born: Digital Reconstruction of the Roman Forum 400 AD
http://www.romereborn.virginia.edu/
Oyez: A Multimedia Archive of the Supreme Court of the United States
http://www.oyez.org/
Mirex: Music Information Retrieval Evaluation eXchange
http://www.music-ir.org/mirex/wiki/MIREX_HOME
Integrative Biology Web Resources Portal
http://lib.berkeley.edu/BIOS/ib.html

> * Although this project was begun more than 10 years ago, it is still producing creative work from the original and derived data sets and the site contains a wealth of references and tools. It has been featured in more than 50 media publications and films. More than any other, this project demonstrated the power of an interdisciplinary group of scholars using exceptionally large and complex data to gain insight into difficult questions.

Session 2: Foundations
Mathematical description and representation of physical entities and natural processes and systems

- Computation, analysis and modeling
- Selected early models of terrestrial and extraterrestrial phenomena
- Epistemological issues associated with mathematical modeling of physical reality and change
    - Models as incomplete and non-verifiable abstractions

Resource materials:
Timelines and histories on the web
mathematics: http://www.math.wichita.edu/~richardson/timeline.html
computation: http://webspace.ship.edu/cgboer/computertimeline.html
IEEE Computer History Timeline:
http://www.computer.org/cms/Computer.org/Publications/timeline.pdf (pdf)

[Epistemology readings are very weighty and probably won't appeal, but here are a few if any students are interested.]

- Black, M. 1962. Models and Metaphors. Ithaca, NY: Cornell University Press Foucault, M. 1994. The Order of Things: An Archaeology of the Human Sciences. Vintage Books (addresses language issues)

- Foucault, M. 1972. The Archaeology of Knowledge & The Discourse on Language. New York: Pantheon Books.
- Whitehead, A.N. 1974. Science and Philosophy. New York: Philosophical Library
- Nagel, E. 1979. Teleology Revisited and Other Essays in the Philosophy and History of Science. New York: Columbia University Press
- Nagel, E. 1961. The Structure of Science. New York: Harcourt, Brace & World.
- Machlup, F.; Mansfield, U. eds. 1983. The Study of Information: Interdisciplinary Messages. New York: Wiley
- Creager, A; Lunbeck, E.; Norton, M. eds. 2007. Science Without Laws: Model Systems, Cases, Exemplary Narratives. Duke University Press, Durham
- Devlin, K. 1991. Situations as Mathematical Abstractions in Situation Theory and Its Applications, ed. Barwise, J. et. al. Stanford, CA: CSLI.
- Lakoff, G.; Johnson, M. 1980. Metaphors We Live By Chicago: The University of Chicago Press.
- Hempel, C. 1965. Aspects of Scientific Explanation. New York: Free Press
- Koyre, A. 1958. From the Closed World to the Infinite Universe. New York: Harper & Brothers.

Session 3: Setting the Stage for Computational Science as the "third paradigm" of Research

- large-scale computing (Supercomputers/early parallel architectures)
- computational models and algorithms
  - mapping applications to architectures
- Grand Challenges*
  - application/domain/algorithm/software/hardware matrices
- Origins of Scientific Visualization
- National Research and Education Networks (NREN) and the Internet
- World Wide Web, Browsers, Search Engines

Resource materials:

Whitepapers and Reports Motivating Early Federal Investment

- Report of the Panel on Large Scale Computing in Science and Engineering, 1982 (Lax Report)
- NSF Working Group on Computers for Research) A National Computing Environment for Academic Research, 1983 (Bardon-Curtis Report)
- Office of Technology Assessment, Supercomputers: Government Plans and Policies, 1982
- High Performance Computing and Communications Supplements to President's Budget to Congress. The complete library of these is available at:

  [These documents, beginning in the FY1992 budget were filled with descriptive material about the "Grand Challenges" and colorful scientific visualizations and were instrumental in persuading Congress to appropriate funds to Supercomputer Centers. In the High Performance Computing and

Communications Program 1992 Supplement to the President's Budget, nine Grand Challenge Applications were proposed that were to be successfully addressed with Teraop Computers by the year 2000. This class session will question how successful this was, both in terms of achieving sustained teraop performance and in constructing computational models of the application areas. (The areas were Climate Modeling, Fluid Turbulence, Human Genome, Ocean Circulation, Quantum Chromodynamics, Semiconductor Modeling, Superconductor Modeling, Viscous Fluid Dynamics and Vision and Cognition. These will be revisited in a later class session.).]

The full catalogue of Supplements the President's budget are archived on the Federal Networking and Information Technology Research and Development Program, the "Nation's primary source of Federally funded revolutionary breakthroughs in advanced information technologies such as computing, networking, and software." [from the web site]

http://www.itrd.gov/Publications/index.aspx

* This phrase was coined by Ken Wilson, Nobel Laureate in physics and head of the Cornell Supercomputer Center in the late 1980s. It was first used at a high performance computing conference in Hawaii in 1987 and has demonstrated remarkable staying power.

Timelines on the web:

IEEE Computer History Timeline
http://www.computer.org/computer/timeline/timeline.pdf

Semiconductors in Computers Timeline
http://www.computerhistory.org/semiconductor/timeline.html

Moore's Law – Transistor Count charts 1970-2010
[I have a number of good ones]

Kryder's Law and Hard Drive capacity over time
http://en.wikipedia.org/wiki/Kryder%27s_law - Kryder.27s_Law

Hobbes Internet timeline (long version):
http://www.zakon.org/robert/internet/timeline/

Jason Powers Internet timeline (simple example):
http://imjustcreative.posterous.com/?tag=visualisation

Roads and Crossroads of the Internet History: by Gregory Gromov
http://www.netvalley.com/cgi-bin/intval/net_history.pl?chapter=1

Internet Society
http://www.internetsociety.org/internet/internet-51/history-internet
[many policy and Federal Support resources here]

Internet History:: NSFNET
http://www.cybertelecom.org/notes/nsfnet.htm

history of computers:
http://www.hitmill.com/computers/computerhx1.html

Session 4: The Digital Libraries Initiatives
This session will begin to discuss the role of the digital libraries projects in shaping and organizing the explosive expansion of digital content to advance research and scholarship in many domains.

- new ICT technologies and tools
- interoperability at multiple levels
- digital content in many forms, data fusion and complexity
- distributed repositories and sharing resources
- expanding the problem domains of existing disciplines and creating new interdisciplinary research areas (numerous project examples here)
- new international dialog, collaboration and communities of practice

Resource materials:
Digital Libraries Initiative Program Announcements
Program Planning Workshop Materials & Reports:
IITA Workshop 1995
http://dbpubs.stanford.edu:8091/diglib/pub/reports/iita-dlw/main.html
Santa Fe Workshop 1997
www.sis.pitt.edu/~repwkshop/papers/dl1997.pdf
Chatham Workshop 2003
http://www.sis.pitt.edu/~dlwkshop/
Phoenix Workshop 2007
http://www.sis.pitt.edu/~repwkshop/index.html

International Collaborations
NSF/JISC: http://www.dlib.org/dlib/june99/06wiseman.html
NSF/DFG/EC: [Many documents relevant to the international collaborative work are available from instructor's personal digital library]

Early Significant Achievements: Project sites
Instructor has an extensive record of digital libraries research, infrastructure development and related activities from 1994 to date in the form of presentations by researchers (~1500 Powerpoint presentations…), in-depth workshop materials, study and project reports (using a variety of media), seminal events, new conference series, discourse communities discussions, etc. in his personal digital library.

The class may be asked here to do a "look ahead" and discuss how rapidly the "data deluge" came about. Particularly useful in this regard are reports such as the IDC White Papers – "The Expanding Digital Universe" produced for EMC in March 2007 and the follow-on report "A Digital Universe Decade – Are You Ready?" produced in 2010. These provide a variety of statistics and charts showing data growth. The instructor has given several talks on this topic and will draw from those as well.

A handout will be an invited article on the history of the Interagency Digital Libraries Initiative for the journal Library Hi Tech.

Session 5: Data-Centered Scholarship and Research Introduction

The following class sessions are organized about different categories of data associated with data-intensive scholarship and research. These categories are:

- data from large-scale simulations
- experimental/observational data captured by high-throughput digital instruments and recording devices in an automated system
- experimental data requires human effort at some stage of acquisition
- internet-based hybrid datasets consisting of born-digital data and data resulting from the continuing digitization of large analog collections
- complex higher-order data objects resulting from computation to digitally enhance and/or create digital models of physical entities and spaces
- digitized holdings of Library Collections

Associated with these data are research activities that in turn have been given unique labels. There is some overlap and the labels are not used within the larger research communities in a consistent way.

eScience

- data from large-scale simulations
- experimental/observational data captured by high-throughput digital instruments and recording devices in an automated system

Data-Driven Research

- experimental/observational data captured by high-throughput digital instruments and recording devices in an automated system
- experimental data requires human effort at some stage of acquisition

Data-Intensive Research

- internet-based hybrid datasets consisting of born-digital data and data resulting from the continuing digitization of large analog collections
- complex higher-order data objects resulting from computation to digitally enhance and/or create digital models of physical entities and spaces
- digitized holdings of Library Collections

In each session, case studies of projects will be presented and discussed with the goals of understanding the motivation and significance of the work, the nature and sources of data and stages of refinement, specific scholarly approaches and analytical methods used, technologies employed, contextual information and metadata, various roles and actors involved and speculation on how this work might connect with and/or contribute to a larger body of related work within and across disciplinary boundaries.

The students will be asked to visit project sites on the web , locate and use tools and datasets for knowledge discovery, analysis of datasets, problem solving, etc. Required readings will include selections from:

Borgman, C. L. 2007. **Scholarship in the Digital Age: Information, Infrastructure and the Internet.** Cambridge, MA: MIT Press

Hey, T., Tansley, S. and Tolle, K. **The Fourth Paradigm: Data-Intensive Scientific Discovery**. Redmond, WA: Microsoft Research

The major cyberinfrastructure reports: **Revolutionizing Science and Engineering through Cyberinfrastructure** by Dan Atkins and others for NSF (2003)and **"Our Cultural Commonwealth"** the American Council of Learned Societies (2006) are also important resources.

Larsen, R., Wactlar, H., **Knowledge Lost in Information**, Report of the NSF Workshop on Research Directions for Digital Libraries. Chatham, MA, June 15-17, 2003.

Arms, W., Larsen, R. **Building the Infrastructure for Cyberscholarship**. Report of a Workshop Held in Phoenix, AZ, April 17-19, 2007. Sponsored by the National Science Foundation and the Joint Information Systems Committee http://www.sis.pitt.edu/~repwkshop/index.html

There are numerous other papers that describe new disciplines and interdisciplines. These are important because they demonstrate in a very convincing fashion that data-intensive research is a transformative mode of inquiry, capable of generating altogether new forms of scholarship and not merely a way of extending established forms. Here is just one example:

Michael F. Goodchild and Donald G. Janelle. Toward Critical Spatial Thinking in the Social Sciences and Humanities. GeoJournal. 2010 February

Session 6: eScience & Data-Driven Research

The next two class sessions will introduce eScience and Data-driven Research. These research modalities involve:

- data from large-scale simulations
- experimental/observational data captured by high-throughput digital instruments and recording devices in an automated system
- experimental data requiring human effort at some stage of acquisition

"eScience" as a term has been used on different occasions to encompass various types of research activities. It will be used here to denote research that reflects the continual evolution of computational science begun in the 1980s (the "third paradigm"). Applications include *simulations* of phenomena either too large or small, fast or slow or too complex to explore in a research laboratory. Examples are better mapping the cosmos, building more accurate earth systems models and examining the basic building blocks of life on earth. Large-scale eScience simulations produce massive datasets, resulting from the solution of complex mathematical models over broad parameter spaces. Primary problem domains are in the physical, biological and geosciences.

eScience applications advance in step with the computational speed (now petaflop scale; soon exaflop scale). For optimal overall productivity, input and output data must be proximate to the computational engines as data bandwidth to the processors can be the rate-controlling step in a computation. The results of large-scale simulations are most frequently presented as scientific visualizations.

[Computation may involve solutions of partial differential equations or determination of probability distributions, etc. As with all research, goals include insight, inference, deep understanding and increased predictive capability. Variables are mathematical continuums – hence more computation produces greater resolution. In the case of probabilistic computations, more computation provides greater measures of certainty. Simulations are compared with observations to validate the models.]

"Data-Driven Research" as described here refers to efforts that involve the analysis of large amounts of *experimental data* at various stages in a larger overall research project. A distinction is made between two types of experimental data: a) that which is collected and prepared using fully automated systems; and, b) that which is collected and prepared but which requires some level of human involvement in the process. Examples of the former are where data capture is accomplished using high-throughput digital instruments and recording devices such as sophisticated astronomical instruments, particle accelerators, environmental sensors, medical diagnostic equipment and many others. The petabytes of raw output data often requires significant computation to yield the basic data for the analysis software; but the overall process is essentially automated.

Experimental data that which requires human involvement at some point in capturing or preparing data for analysis often includes that from social network analysis, bibliographic citation analysis and demographics studies.

Resource Materials:

Examples taken from:

- IEEE International Conferences on e-Science and Grid Computing [2005-2012]
- 2005: http://www.cloudbus.org/escience/index.html
- 2006: http://www.escience-meeting.org/eScience2006/
- 2007: http://www.escience2007.org/
- 2008: http://escience2008.iu.edu/
- 2009: http://www.escience2009.org/
- 2010: http://www.escience2010.org/
- 2011: http://www.escience2011.org/
- 2012: http://www.ci.uchicago.edu/escience2012/

Large-scale collaborative eScience Centers and Projects

- Pittsburgh Supercomputer Center
- National Center for Supercomputer Applications
- San Diego Supercomputer Center

- Los Alamos National Lab
- Argonne National Lab
- CERN, Switzerland

National Virtual Observatory
http://virtualobservatory.org/whatis/history.aspx
Astronomy: Infrared Processing and Analysis Center
http://www.ipac.caltech.edu/project/20
Center for Embedded Network Sensing
http://research.cens.ucla.edu/

Class Discussion:
In the High Performance Computing and Communications Program 1992 Supplement to the President's Budget, nine Grand Challenge Applications were proposed that were to be successfully addressed by the year 2000. The class will be asked to discuss whether these goals were accomplished, both in terms of achieving sustained teraop performance and in constructing computational models of the application areas. (The areas were Climate Modeling, Fluid Turbulence, Human Genome, Ocean Circulation, Quantum Chromodynamics, Semiconductor Modeling, Superconductor Modeling, Viscous Fluid Dynamics and Vision and Cognition)]

Session 7: eScience & Data-Driven Research – part 2

This session will continue the discussion begun in the prior class meeting.

Resource Materials: TBD

The class will discuss particular projects as transformative or innovative in a particularly influential way. We will be looking for projects that proved to be turning points in a line of research or those that opened an entirely new line of inquiry. The reference points will be from the material presented in Session 5 and the assignments coming out of that session.

Hands-on Assignment
The class will be asked to locate data sets of special personal interest and if appropriate upload the data to a web-based data analysis and visualization site. The data could then be visualized data in different ways and multiple interpretations might be drawn. From this, the class could come to recognize and agree upon what is the most informative visualization for particular data sets. This could be done in real-time in the class. A particularly popular sites for doing this is IBM's Many Eyes:
http://www-958.ibm.com/software/data/cognos/manyeyes/

Session 8: Data-Intensive Research and Scholarship - overview
Data-intensive research and scholarship is used here to refer to scholarly activities based on several categories of very complex data.

a) Internet-Based Hybrid Datasets Consisting Of Born-Digital Data And Data Resulting From The Continuing Digitization Of Large Analog Collections And Records

The research space made possible by hybrid datasets is exceptionally large and diverse. Disciplinary research that study change over long time periods and geographic locations is one example. This might mean using spatial and temporal indexed data to produce historical maps and timelines, combining satellite imagery with field notes to identify and reconstruct archaeological sites and discover new sites in the same vicinity.

b) Data Objects And Collections Resulting From Digital Enhancement and/or Digital Representation Of Physical Entities, Systems And Spaces

The area of interest may be significant cultural artifacts, virtual representations of larger structures and places, ancient manuscripts, illustrations, statuaries and other objects which have sustained significant deterioration, decomposition, or destruction over time for any number of reasons. Success of the research effort may require detailed knowledge of the object in its original physical condition and relevant contextual information drawn from analog collections.

The digital remediation of deteriorated objects, or creation of digital surrogates often requires large-scale computation on lower-order, raw data from sophisticated analytical instruments (laser-scanning, x-ray CT, multi-spectral illumination devices) combined with state-of-the-art computer vision and graphics techniques. The complex constructs that result can be highly accurate representations but also constitute petabytes of data. However in many cases, the digital representation will prove to be of greater scholarly value than the original physical object as the reconstructive process will provide resolutions [which can be adjusted] and detail far beyond that available from traditional examination. [And the can be duplicated and shared on demand.]

Methods originally developed for the sciences have proven to be useful in non-scientific research: multispectral analysis and x-ray computed tomography developed for medicine is being used to restore legibility to inscriptions on severely deteriorated ancient manuscripts; textual analysis for multi-version documents draws upon methods developed for biological comparison of species. The enhancement and recreation of historic documents and objects combined with other knowledge resources is transforming humanities research.

c) Digitized holdings of Library and Museum Collections

Some of the most valuable data to humanists, as well as researchers studying evolutionary processes and environmental change (paleontologists, p-biologists, p-botanists, p-climatologists, et.al.) are found in library and museum collections. Large-scale digitization of analog collections of library and museum holdings, including material culture, and open access to on-line books, scholarly journals, and other knowledge resources gives new opportunities for scholarship and research in many domains and at the same time dramatically increases the user base of individual memory institutions. In many instances converted analog resources provides critical contextual information to born-digital data.

Resource Materials:
There are many digital libraries projects that fit this category. Some are multi-national, large-scale efforts that have been ongoing for more than a decade. Just a few examples would include:

- Cuneiform Digital Library (UCLA lead)
- Perseus Project (Tufts lead)
- Electronic Cultural Atlas Initiative (University of California, Berkeley lead)
- EDUCE Project (University of Kentucky lead)
- Brown University Shape Laboratory projects
- World-Historical Dataverse (Univ of Pittsburgh)
- ECHO (European Cultural Heritage Online)
  http://echo.mpiwg-berlin.mpg.de/home
  …

Students can explore all selected project sites in order to discover scholarly activities that appeal directly to them. They would produce a document explaining their findings in the form of a paper, Powerpoint, Prezi; etc. and share this with the class.

Session 9: Data-Intensive Research: Inscriptions, Scripts & Manuscripts before Invention of the Printing Press

This session will focus on computer vision techniques and collaborative efforts between information technologists and domain scholars to recover inscriptions made on a variety of media with the goal of adding factual information to the evidence base for the humanities, while at the same time pushing information technology research into new areas.

The following projects are of interest:

- Cunieform Tablets (laser scans and 3-D rendering techniques) Ref: http://cdli.ucla.edu/
- Papyrus and Scrolls (x-ray computed tomography, laser scans, multispectral illumination) Ref: http://www.rch.uky.edu/
- Handwritten Manuscipts from Scriptoriums (multispectral illumination and laser scanning) Refs: St. Chad's Gospel
- Homer Multitext Project: http://chs.harvard.edu/chs
- Venetus A: www.**wired**.com/gadgets/miscellaneous/news/2007/06/iliad_scan
- Palimpsests (x-ray fluorescence imaging)
  Ref: http://www.archimedespalimpsest.org/imaging_experimental4.html
- Virtual Vellum Project
  http://www.shef.ac.uk/hri/projects/projectpages/virtualvellum
- Perseus Project: http://www.perseus.tufts.edu/hopper/

Session 10: Data-Intensive Research: 2D Imaging of Paintings, Objects and Structures

This session will focus of technologies used for digital imaging of works of art and significant cultural objects and four related research issues: a) creation; b) image search and retrieval; c) presentation; and, d) applications and use.

Visual imagery is rich in features including shape, texture, subsurface structure, color, forms and content aspects, media-based characteristics and other attributes visible to the human eye in some instances, but requiring technology mediation in other instances to make features visible. Multi-modal acquisition is often required to gain new insight into the origin and purposes of images. Rendering and other presentation methodologies are often required for measurements of degradation and to suggest remediation strategies.

Technologies for analysis of imagery include:

- computational photography and enhanced video
- multi-spectral illumination (visible, IR, UV)
- laser scanning
- x-ray computed tomography
- reflectance methodologies (Fresnel, IR, …)
- x-ray fluorescence

References and Examples:

1. Computational photography application examples:
   Research slides from Mark Levoy, Stanford U and Noah Snavely, U Washington
2. Reflected infrared light application example:
   http://esciencenews.com/articles/2012/06/18/reflected.infrared.light.unveils.never.seen.details.renaissance.paintings
3. Computerized analysis of van Gogh's painting brushstrokes
   Research slides from James Wang
   Article in IEEE SIGNAL PROCESSING MAGAZINE [37] JULY 2008
4. Comparative Study of Cypriot Icons
   Research slides from James Wang
5. Study of ancient Chinese maps from the National Palace Museum, Taipei, Taiwan, slides and videos from Brent Seales, U Kentucky
6. Shape analysis of early-American maps
   Slide presentations from Peter Bajcsy
7. BAILANDO Projects at U California, Berkeley
   http://bailando.sims.berkeley.edu/index.html
8. Global Memory Net Image Collection
   http://memorynet.org/
9. Rock Art (image feature matching and data mining) Ref: Eamonn Keogh
   http://www.cs.ucr.edu/~eamonn/ slides from presentation by Eamonn Keogh

Resource Material:

- Report Of The Delos-NSF Working Group On Digital Imagery For Significant Cultural And Historical Materials, Ching-chih Chen, Alberto Del Bimbo Co-Chairs

- Report of the NSF-Mellon Foundation Workshop on Digital Imagery for Works of Art, Kevin Kiernan, Charles Rhyne and Ron Spronk, Co-Chairs

Session 11: Data-Intensive Research: 3D Imaging of Objects, Artifacts and Structures

The motivation for the work described in this class session is to focus on digital restoration of degraded media. This research benefits many disciplinary domains: anthropology, archeology, classics, et.al. At the same time the domain inquiry pushes IT research in areas related to computer vision, information retrieval, computer graphics, color analysis, data mining and many others. By so doing interdisciplinary collaborations is fostered.

Digital surrogates of 3-D objects

1. Reassembly/reconstruction/virtual completion of broken artifacts. Refs:
2. Forma Urbis Roma: http://formaurbis.stanford.edu/
3. Pot Sherds from Petrus: http://www.lems.brown.edu/shape/
4. 3-D Colonial Philadelphia artifacts: http://tsp.cs.drexel.edu/pmwiki/pmwiki.php?n=INHPProject.Presentation
5. Computational photography examples showing construction of 3D objects and structures from large series of random 2D photos from open sites on the web. Research slides from Mark Levoy, Stanford U, Noah Snavely, U Washington and Svetlana Labeznik, U North Carolina http://graphics.stanford.edu/projects/Examples of archaeological site reconstruction from mixed source materials: Refs: Peter Allen, Columbia U
6. Digital Morphology Laboratory, U Texas, Austin, http://www.digimorph.org/
7. Grouping and classifying of North American arrowheads by date and culture. U California, Riverside has collection of > 1M arrowheads that are being digitized, grouped and classified. Research slides from Eamonn Keogh

In many domains, classification of objects is a primary focus. It is used to group and distinguish among classes and types of artifacts in order to create understandable units as well as to associate them with particular time periods, locations, cultures and functional utility. This allows scholars to discern similarities, differences and relationships among types of objects and obtain a clearer understanding of the larger topics of interest. Computation extends the ability for feature matching and comparison dramatically. [In one case noted above, 1,000,000 million projectile points were the focus of grouping and analysis.] Typology is the term used to describe hierarchical classifications.

List of Conferences and Professional Groups:
Computer Applications & Quantitative Methods in Archaeology
http://caaconference.org/
http://digitalhumanities.org/centernet/
[to be expanded]

Session 12: Data-Intensive Research: 2D & 3D Imaging – Part 2

This session continues the discussions begun in the prior two class meetings and gives students the opportunity to display their assignments to the rest of the class.

Session 13: Information Visualization and Virtual Spaces

This session will focus on the recent trends to visualize large sets of quantitative data and explore creative and innovative data visualizations. Visualizations of results has long been an essential tool for presenting the output of high performance computer simulations in the sciences and used extensively in analyses of vast quantities of large-scale, complex datasets. More recently humanists have focused on information visualization as a useful tool to gain insight into large text collections and more heterogeneous datasets.

Virtual reality has similarly been used in a variety of application areas in the physical and geo sciences, but also used in developing entertainment – computer games.

The class will explore some of the more common forms of information visualization as well as certain outstanding examples of cultural virtual reality projects such as Rome Reborn, a digital reconstruction of Rome as it existed at approximately A.D. 320. The project itself contains a large number of models and accompanying scholarly papers. http://www.romereborn.virginia.edu/. The project is part of a much larger endeavor, the Virtual World Heritage Laboratory. The mission statement of the laboratory offers and insightful summary as to the value that humanists place on digital representations and tools to pursue their scholarly goals:

> In recent years, humanists in many disciplines have been using 3D digital technologies to capture and model their objects of study, from humble artifacts such as vases or furniture to entire cities such as ancient Rome. 3D has become a new and powerful form of scholarly expression and communication. The mission of the Virtual World Heritage Laboratory is to apply these new tools not only as interactive illustrations but also as heuristic instruments of discovery. The scope of our interests--as implied by the phrase "World Heritage"--includes the entire human record. The focus of our investigations, as is suggested by the phrase "Virtual World"--is the metaverse and how it can make possible experiences and experiments that--short of time travel--would otherwise not be possible. Click here to read a recent article by Director Bernard Frischer setting forth the key ideas behind the creation of the VWHL.
> http://vwhl.clas.virginia.edu/mission.html

Resource Material:
There are a large number of annual conferences either devoted entirely to information visualization in various disciplinary domains, or which contain tracks on data visualization. Often, data visualization is part of large conferences and meetings focused on data mining, data analytics, data science and related research areas. Others are associated with commercial and entertainment industries. A few notable annual events are:

ACM SIGGRAPH Conferences
IEEE Scientific Visualization
IEEE Information Visualization
IEEE Analytics Science and Technology
VisWeek
Infographics World Summit
International Conference on Information Visualization Theory and Applications

Project sites for class discussion:
Center of Interdisciplinary Science for Art, Architecture and Archaeology, University of California, San Diego
http://cisa3.calit2.net/
Electronic Cultural Atlas Initiative
www.ecai.org
[to be expanded]

Session 14: Audio and Music Digital Scholarship

This session will give a broad overview of audio achives and the rapidly emerging field of music information retrieval and analysis. The audio archive to be focused on is the Oyez Project. http://www.oyez.org.

The Oyez project was originally part of the Michigan State/Northwestern University National Gallery of the Spoken Word Digital Libraries Initiative project. Oyez received separate funding in 2003. The project is a multidisciplinary effort to create a permanent and free archive comprising all audio recorded in the Supreme Court of the United States since October 3, 1955. The archive serves a multidisciplinary audience for research and a general public interested in the activities of the nation's most powerful yet remote legal body whose decisions have affected the foundations of the American Life, while the debate and legal arguments that led to the decisions have been essentially been inaccessible except to segments of the legal community. At present, the audio archive contains 110+ million words in 9000+ hours of audio synchronized to the sentence level (largest existing database of spoken language data). Electronic transcripts are synchronized to the audio files enabling quick search-and-retrieval when searching audio. For forty years individual justices who spoke at oral argument were not identified in the recordings. The project has addressed this issue through speaker-specific acoustic models that add highly accurate speaker identification. Elements of this rich database may be accessed through a custom player created to make use of the audio as easy as possible. The site draws a large audience – approximately 500,000 visits per month. [from www.oyez.org]

Music information analysis and retrieval is a newly emerging field within computer and information sciences and is generating high levels of enthusiasm. Major technology companies are developing applications based on this research at a rapid pace due to market demand. There is an International Society and an annual series of conferences devoted to music research. The web site for the Society contains a wealth of information.
http://www.ismir.net/

A central resource for music information researchers is the the Music Information Retrieval Evaluation eXchange (MIREX), an evaluation framework and infrastructure founded and managed by the International Music Information Retrieval Systems Evaluation Laboratory (IMIRSEL) [2] at the University of Illinois at Urbana-Champaign (UIUC). MIREX provides infrastructure for the scientific evaluation of the many different retrieval and analysis techniques and algorithms being developed by researchers engaged in music information research and the development of music digital libraries. http://www.music-ir.org/mirex/wiki/MIREX_HOME

Session 15: Roles of Libraries and iSchools

The final session can be devoted to a number of different areas of consideration. One might be to discuss what new roles and responsibilities (and opportunities) data-intensive scholarship offers to libraries and iSchools.

It is essential to data-intensive scholarship and research that data resources and tools be carefully managed over the entire research cycle and beyond. Research data must be managed from the *earliest* stages of the information lifecycle in order to create highly functional collections that naturally grow, connect to other resources and remain available for reuse indefinitely into the future. Funding agencies are now mandating this in many instances. Data centers can capture and preserve adequate documentation relating to the content, structure, context, experimental parameters and environmental circumstances of scientific data collections. (The Astronomy Community has been exceptionally good at this.) Computational models, algorithms, software and visualization facilities are also usually maintained by the laboratories at which eScience and data-driven research is performed. Open source software tools are proliferating on individual sites across the web. Data grids and cloud storage environments are being used increasingly. But many would argue that more centralized, explicit arrangements are necessary.

Libraries are destined to play a major role in support of data intensive research. Domain scientists and scholars often do not have the skills, nor interest in managing data for reuse and long-term preservation and archiving – they are focused on using it to accomplish their research. The role of Libraries is already evolving from that of repositories and providers of knowledge resources to being an active agent in the research process. Libraries have the internal structure and are adding staff with the skills to maintain large data sets, create collections of tools, create metadata linkages, add provenance information, address long-term preservation and archiving, and attend to all of the tasks associated with data curation. The iSchools are producing graduates with the extensive skills relevant to all of the above.