Humanities and Social Sciences Faculty Brown Bag

Visualization and High Performance Computing (HPC)
12:00 - 1:30, April 22nd, MCK 375

Have you ever wanted to mine big data? What do you do with big data once you have it? What techniques are available in exploratory data mining for quantitative research? Can we help you find appropriate big data sets?

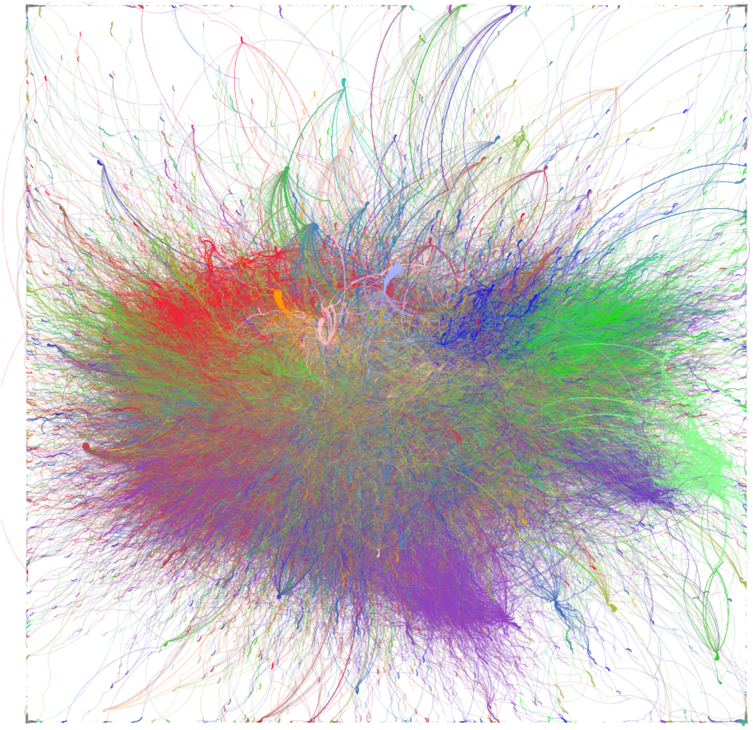
To determine the interest in holding a series of summer workshops for CAS faculty in the Humanities and the Social Sciences about using big data, visualization, and high performance computing, CASIT will hold a brown bag lunch to discuss such topics as:

- Peer research examples using big data and high performance computing (HPC)
- Available data sources as examples of what is out there
- The basics of visualization techniques
- Leveraging big data and computation processing both on and off campus
- Incorporating visualization and HPC in teaching and research materials
- Tentative summer workshop schedule
- Summer participation compensation incentives

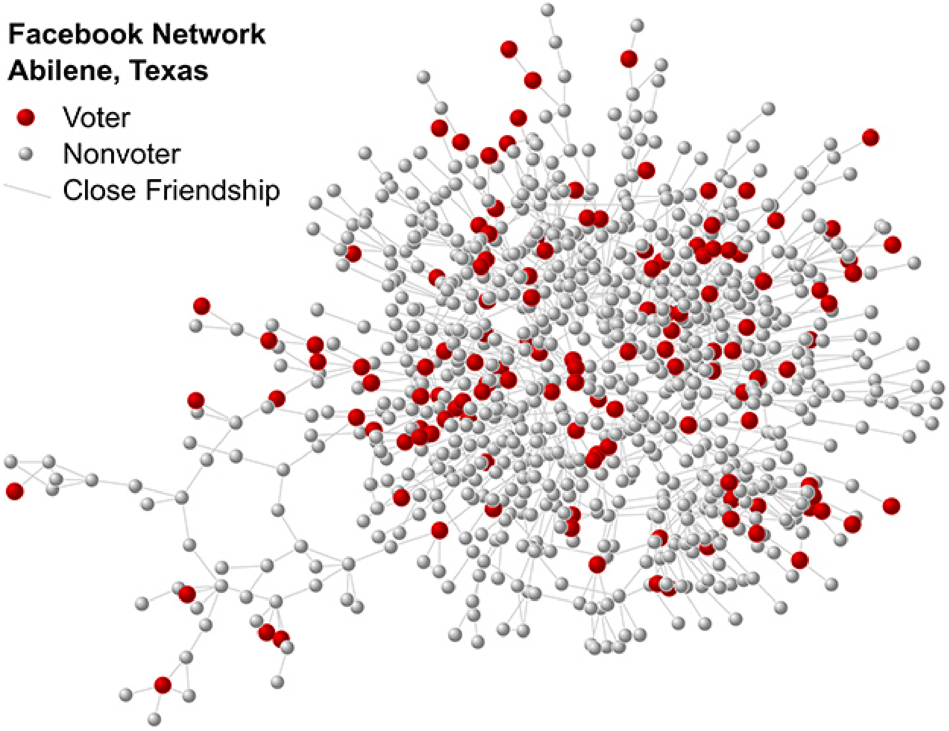
Most importantly, we want to hear from you! What do you want to explore? Our goal is to explore and develop information technology as a tool for scholarly research. We hope by demonstrating what others are doing with big data sets, high performance computing and visualization, and showing examples of available big data sets, we can stimulate ideas and further partnerships with our faculty in the Social Sciences and the Humanities.

Examples of research with big data and high performance computing:

- Text visualization
- How the meanings of words in Latin and Greek have changed over their lifetimes
- Reconstruct ancient artifacts and architecture
- Examining how Facebook boosts voter turnout
- Data analytics and forensic studies to trace President Lincoln's correspondence in time and space
- Explore the lives of slaves across the Western Hemisphere
- Examining recordings of Mopan language and how the language challenges current linguistic theory
- Studying the aural character of buildings
- Comparative search of the textural and graphic tradition connected to Leonardo's Treatise on Painting
- Produce online visualisations that capture and help interpret the complex spatial dynamics of cities
- Virality prediction and community structure in social networks



Network analysis of 500,000 tweets on #Syria by 60,000 users over 23 days in November 2011. Color key: 59% Arabic (Green), 30% English (Blue), 2.25% Hindi (Gold), 1.6% French (Red), .7% Urdu (Purple) and .6% Finnish. Everything else (Farsi, Spanish, Italian, Portuguese) is less than .5%. Conducted by R-Shief, Inc.



About one third of a million more people showed up at the ballot box in the United States in 2010 because of a single Facebook message on Election Day. Illustrative map of part of the social network of “close friends” from Abilene, Texas, who logged in on Election Day in 2010. Study led by the University of California, San Diego.

Data Source Examples:

USC Shoah Foundation

USC Shoah Foundation Videos
The USC Shoah Foundation Institute for Visual History and Education has archived a library of 52,000 video testimonies of Holocaust survivors and witnesses.
http://sfi.usc.edu/teach_and_learn/for_educators/course_development

USC Shoah Foundation - Courses
A total of 474 courses at 59 universities have drawn upon the Institute's testimonies in more than 25 academic disciplines. The courses demonstrate the range of topics and themes that the Institute's archive of video testimonies can help illuminate. Use the table below to see what the trends are for testimony-enhanced courses taught at colleges and universities around the world. conduct searches in the Institute's Visual History Archive. Log on to our Visual History Archive Online and search the entire catalog of nearly 52,000 witnesses so far catalogued and indexed of the over 52,000 testimonies housed at the Institute. Watch a sample of 1,228 video testimonies. See more at: http://sfi.usc.edu/teach_and_learn/for_educators/course_development#sthash.qW0GajhI.dpuf

Google

Google Books https://www.youtube.com/watch?v=yyrHFXbeMu8
As of April 2013, the number of scanned books was over 30 million, but the scanning process has slowed down in USA academic libraries. Google estimated in 2010 that there were about 130 million unique books in the world, and stated that it intended to scan all of them by the end of the decade.
Google Books Ngrams
http://storage.googleapis.com/books/ngrams/books/datasetsv2.html
Search and analyze the full text of any of the millions of books digitised as part of the Google Books project.

Google Correlate
http://googleblog.blogspot.com/2011/05/mining-patterns-in-search-data-with.html

Google Trends
http://www.google.com/trends/#

Google Finance
https://www.google.com/finance
40 years' worth of stock market data, updated in real time.

Yelp http://www.yelp.com/dataset_challenge

Yelp Dataset Challenge is doubling up: Now 10 cities across 4 countries! Two years, four highly competitive rounds, over $35,000 in cash prizes awarded and several hundred peer-reviewed papers later: the Yelp Dataset Challenge is doubling up. We are proud to announce our latest dataset that includes information about local businesses, reviews and users in 10 cities across 4 countries. The Yelp Challenge dataset is much larger and richer than the Academic Dataset. This treasure trove of local business data is waiting to be mined and we can't wait to see you push the frontiers of data science research with our data.

Data.gov http://data.gov

The US Government pledged last year to make all government data available freely online. This site is the first stage and acts as a portal to all sorts of amazing information on everything from climate to crime.

US Census Bureau http://www.census.gov/data.html

A wealth of information on the lives of US citizens covering population data, geographic data and education.

European Union Open Data Portal http://open-data.europa.eu/en/data/

As the above, but based on data from European Union institutions.

Data.gov.uk http://data.gov.uk/

Data from the UK Government, including the British National Bibliography – metadata on all UK books and publications since 1950.

The CIA World Factbook https://www.cia.gov/library/publications/the-world-factbook/

Information on history, population, economy, government, infrastructure and military of 267 countries.

Healthdata.gov https://www.healthdata.gov/

125 years of US health care data including claim-level Medicare data, epidemiology and population statistics.

NHS Health and Social Care Information Centre http://www.hscic.gov.uk/home

Health data sets from the UK National Health Service.

Amazon Web Services public datasets http://aws.amazon.com/datasets

Huge resource of public data, including the 1000 Genome Project, an attempt to build the most comprehensive database of human genetic information and NASA's database of satellite imagery of Earth.

Facebook Graph https://developers.facebook.com/docs/graph-api

Although much of the information on users' Facebook profile is private, a lot isn't – Facebook provide the Graph API as a way of querying the huge amount of information that its users are happy to share with the world (or can't hide because they haven't worked out how the privacy settings work).

Sites of Interest

Institute for Advanced Technology in the Humanities - explore and develop information technology as a tool for scholarly humanities research
http://www.iath.virginia.edu/projects.html

The Institute for Computing in Humanities, Arts, and Social Sciences (I-CHASS) at the University of Illinois at Urbana-Champaign
http://chass.illinois.edu/index.php/ichass-projects/

Visualizing Theatrical Text: From Watching the Script to the Simulated Environment for Theatre (SET)
http://www.digitalhumanities.org/dhq/vol/7/3/000166/000166.html

Network Visualization: The "Bush Team" in Reuters News Ticker 9/11-11/15/01
http://www.cmu.edu/joss/content/articles/volume5/JohnsonKrempel/

London School of Economics and Political Science - Luminocity Project
http://blogs.lse.ac.uk/impactofsocialsciences/2013/11/15/luminocity-project-urban-cartography
http://luminocitymap.org/#population_density_2011/7/52.600/-2.500

The CASA Blog Network
http://blogs.casa.ucl.ac.uk/

The Racial Dot Map - One Dot Per Person for the Entire United States, Demographics Research Group, University of Virginia
http://www.coopercenter.org/demographics/Racial-Dot-Map

Big Data Challenges in Social Sciences & Humanities Research
http://www.datanami.com/2014/09/08/big-data-challenges-social-sciences-humanities-research/

Sponsored by CASIT's Research Support Services (CASIT RSS), one of CASIT's outstanding services.