April 10, 2014

Getting Started With High Performance Computing for Humanities, Arts, and Social Science

XSEDE

Extreme Science and Engineering Discovery Environment

Alan B. Craig, PhD
acraig@ncsa.uiuc.edu

Who am I?

- HASS Specialist for XSEDE (50%)
- Senior Associate Director for I-CHASS
- Research Scientist at NCSA (27 years)

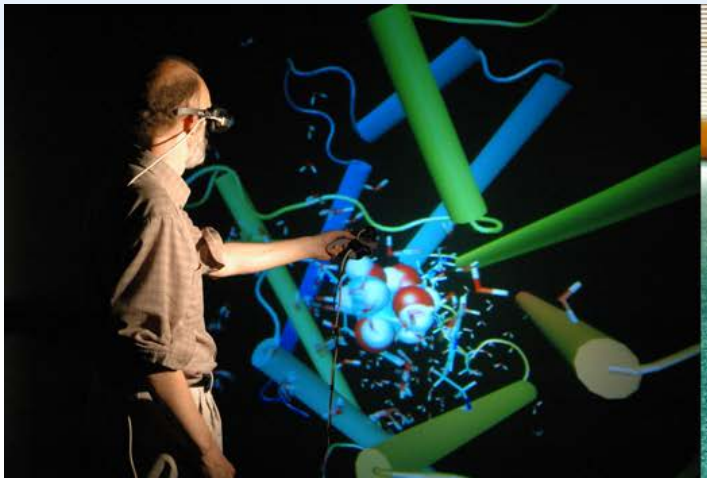
- I am a point of contact for you to help engage with appropriate experts

NCSA

I-CHASS


THE
VALLEY
OF THE
SHADOW
SEASR
SEASR: Engineering Knowledge for the Humanities
Virtual Vellum

What are my interests outside of XSEDE?



Virtual Reality

Augmented Reality

Personal Fabrication

Visualization
Representation of Information
Human-Computer Interaction

The Order of this Presentation

- Today will be *extremely informal.* Let's have a dialog, not a one way presentation

- Concepts
- Application Areas
- Mechanics
- Application Examples throughout

There may be *some* duplication with yesterday

First – Some Definitions

- HASS
- HPC
- Supercomputer
- Data
- Model
- XSEDE
- ECSS

Why would anyone in HASS want to use HPC?

- Humans are good at certain things
  - Reasoning
  - Interpretation
  - And more
- Computers are good at certain things
  - Repetitive Tasks
  - Identifying potential anomalies
  - Some reasoning, interpretation
  - Identifying relationships
  - And more

Why would anyone in HASS want to use HPC?

- It's an issue of *scale*
- It's an issue of *scale*
- It's an issue of *scale*

- *Scale of amount of data*
- *Scale of amount of computation*
- *Scale of amount of storage*
- *Scale of problem scope*

Scale Example

It Takes a Team

- Large projects often require interdisciplinary team
- Content Expertise
- Computational Expertise
- Visualization Expertise
- Project Management Expertise

What is Data?

- Data is information
- In our case it is information in electronic form
- Data can be:
  - Numeric
  - Text
  - Visual Image
  - Audio
  - Video
  - Etc. (any kind of signals)

Representation of Information

- Data can be manipulated in useful ways
- Computer can help with this
- We need to choose good representations for different purposes, and have the computer do the work
- We can transform from and to different representations

What Are Some Things People Do?

- Text Analysis
- Image Analysis
- Video Analysis
- Audio Analysis
- Network Analysis
- GIS
- Simulation
- Visualization
- Display and Interaction

Text Analysis

- Statistics (word counts, co-occurrences, etc.)
- Entity recognition
- Clustering / categorization
- Etc.

- Genre identification
- Topic Modeling
- Sentiment Analysis
- Etc.

Text Analysis

- Find things in a collection of text
- What is in this collection of text?
- Machine Learning applications

Text Analysis

- Interesting tools:
  - Mallet
    - MAchine Learning for LanguagE Toolkit
    - Topic Modeling
    - Installing on Blacklight at PSC
    - http://mallet.cs.umass.edu
- We can discuss appropriate tools for your project. Different strengths and weaknesses
- Some scale better, different capabilities…

Image Analysis

- What is in this image?
- Find images that…
- OCR
- Who painted this?
- Who wrote this manuscript?
- Again, machine learning often used

Image Analysis – Authorship Example

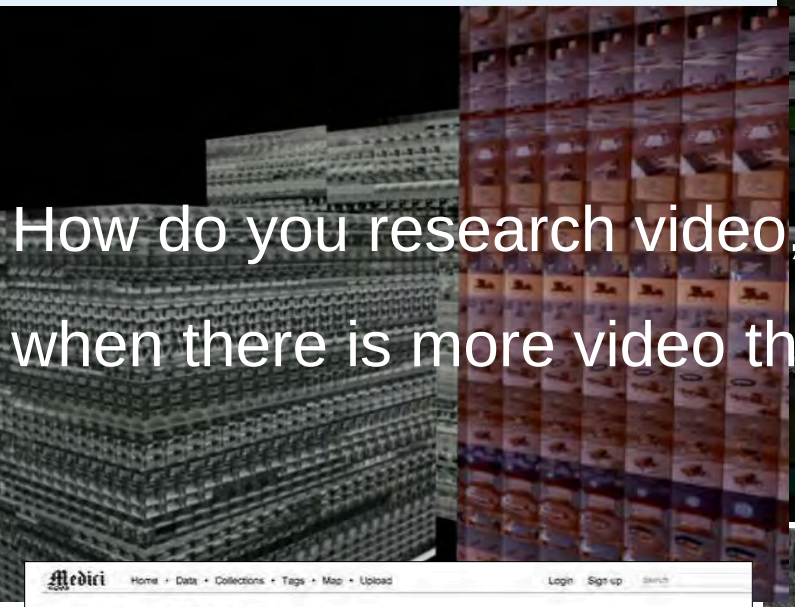
15th-century manuscripts, 17th and 18th-century maps, and 19th and 20th-century quilts

ISDA – DID - McHenry

Video Analysis

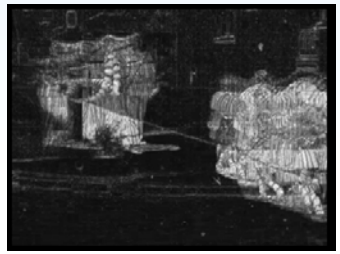
- Video is being created far faster than we can view it – EG – YouTube, Surveillance, etc.
- It is more than just the case of multiple images
- Scene Identification
- Contents – EG Phones
- Cinematographic Elements – Camera moves, lighting, etc.
- Visualization of movies

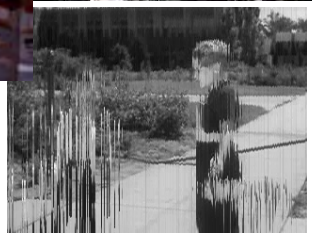
Large Scale Video Analytics: On-Demand, iterative inquiry for moving image research

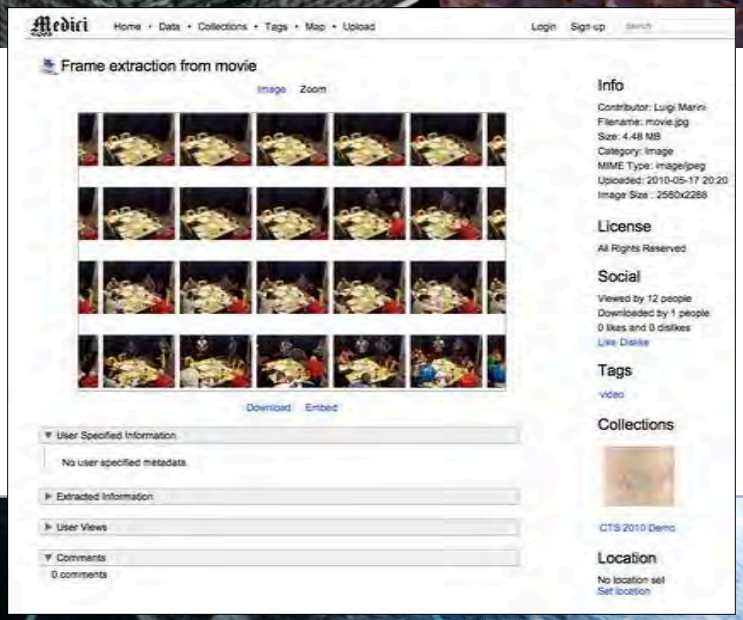


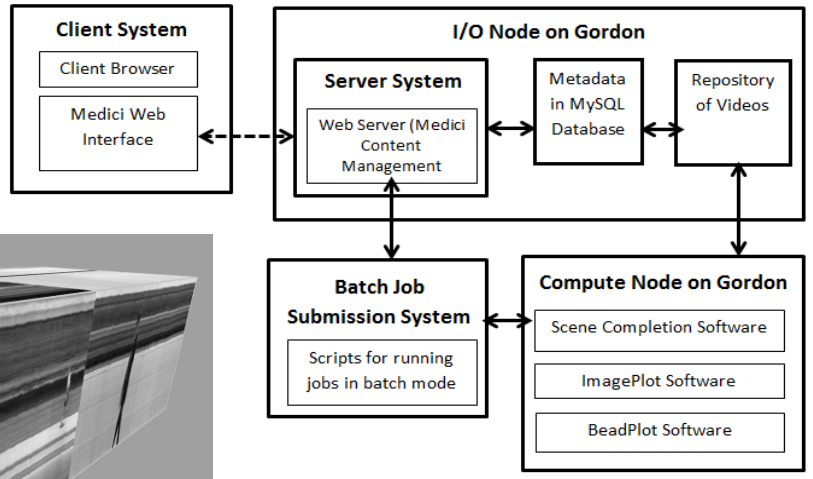
How do you research video,

when there is more video than can ever be watched?









XSEDE

Audio Analysis

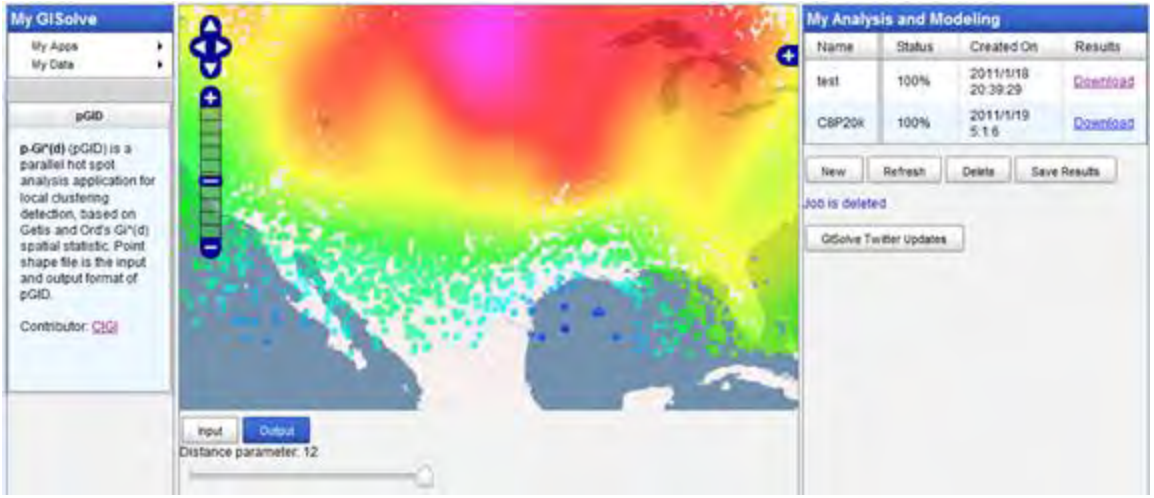
- Audio to text
- Audio Search / Retrieval
- Inflection Analysis
- Classification
- Audio with Video
- Audio Feature Extraction

Network Analysis

- Looks at relationships
- Who is connected to who?
- Who is connected to where?
- Etc.
- Think about facebook… think about twitter… think about other social media….
- Think about world events and news items…

GIS

- Map based information
- Spatial Studies
- Often combined with others listed here
- See Yan's presentation this afternoon.

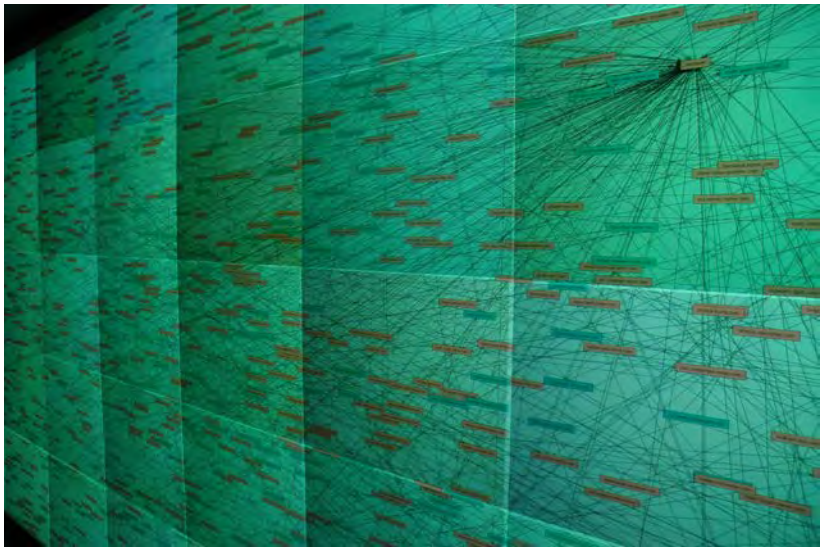


Simulation

- One of traditional uses of supercomputers for scientists
- Can simulate events to see how they might play out
- Think about propagation of ideas as diffusion…
- Historical counterfactuals
- Other ideas?

Visualization

- Represent data visually (and other senses)
- Can show relationships
- Can show the unseen



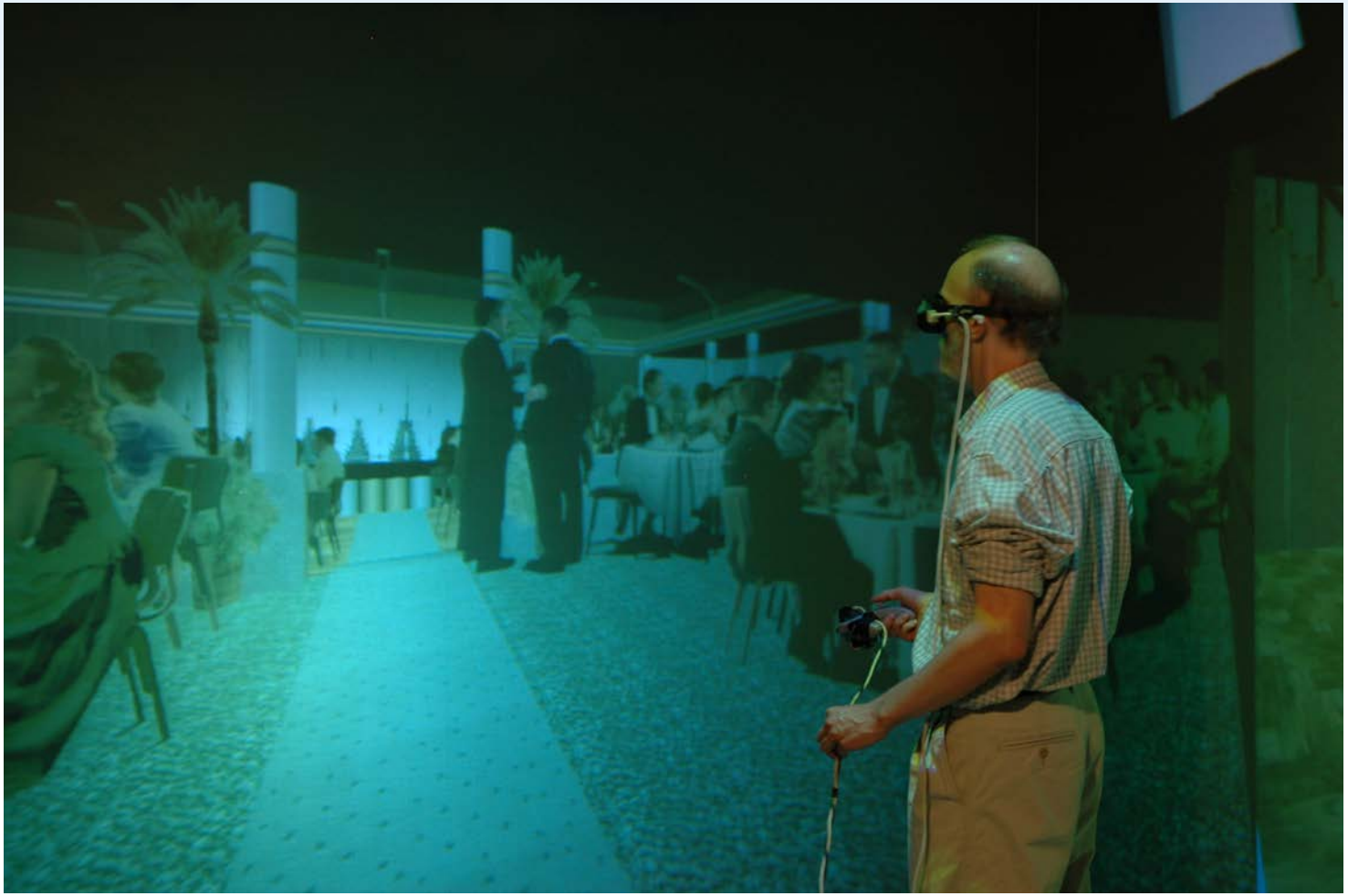
Display and Interaction

- Virtual Reality
  - Purely synthetic environment
  - Bodily engaged
- Augmented Reality
  - Real world combined with digital
  - Bodily engaged
- Note that Facebook just bought Oculus for $2B
- Microsoft bought $150M in VR and AR patents

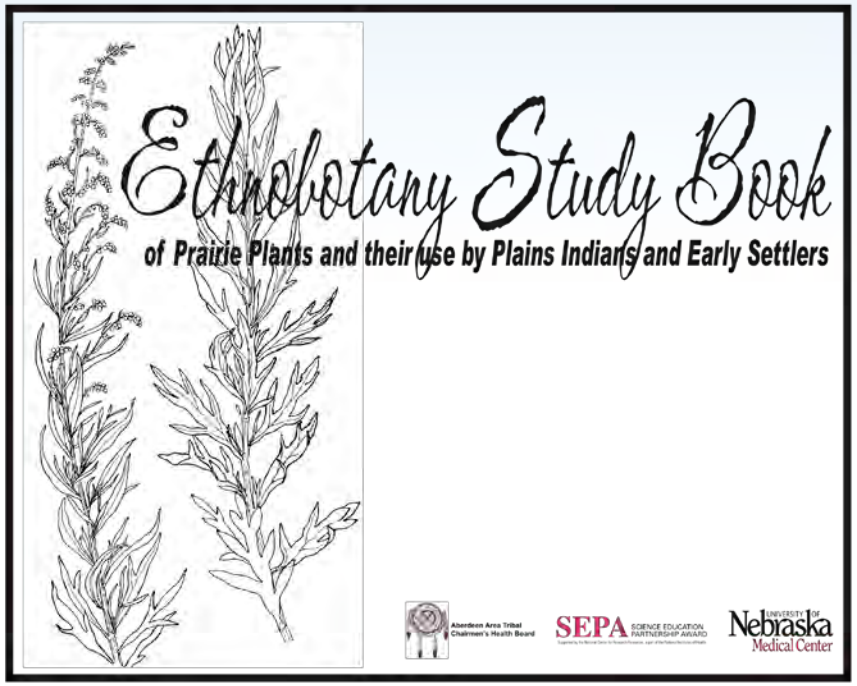
VR Example – Harlem in 1920s

Bryan Carter et al.
Central Missouri State University, U. of Missouri, UIC, U. of Arizona
African American Studies, English Literature, Pedagogy, Communications, and Computer Science.

VR Example – Harlem in 1920s



Augmented Reality Examples / Demos



Presentation and Interaction

3D Information in Books

Time Varying Information in Books

Augmented Reality Examples / Demos

AR in Archeology

- NSF ARC-1025298
- Field Experience
- Physical Component

Gateways

- Strategy for building communities, lowering barrier to entry
- GIS gateway already exists
- Building Video Analysis Gateway
- Others to follow – Text, Image, etc.

What is HPC and What is XSEDE?

What is XSEDE?

- XSEDE is a network of national resources (that *you can use*) that act as a coherent system for High Performance Computing including:
  - Processing
  - Storage
  - Networking
  - Visualization
  - Expertise
  - Etc.

Current state of HASS and HPC

- There are HASS projects using XSEDE successfully
- Many others are interested, but haven't taken the leap yet
- Some HASS problems don't fit the XSEDE mold…. We are changing the mold!
- XSEDE / HPC Communities can benefit from input from HASS Communities

Four categories of researchers

- Have code, have expertise
- 3rd party codes, may or may not be on HPC
- No code, but great idea
- No idea what to do but interested

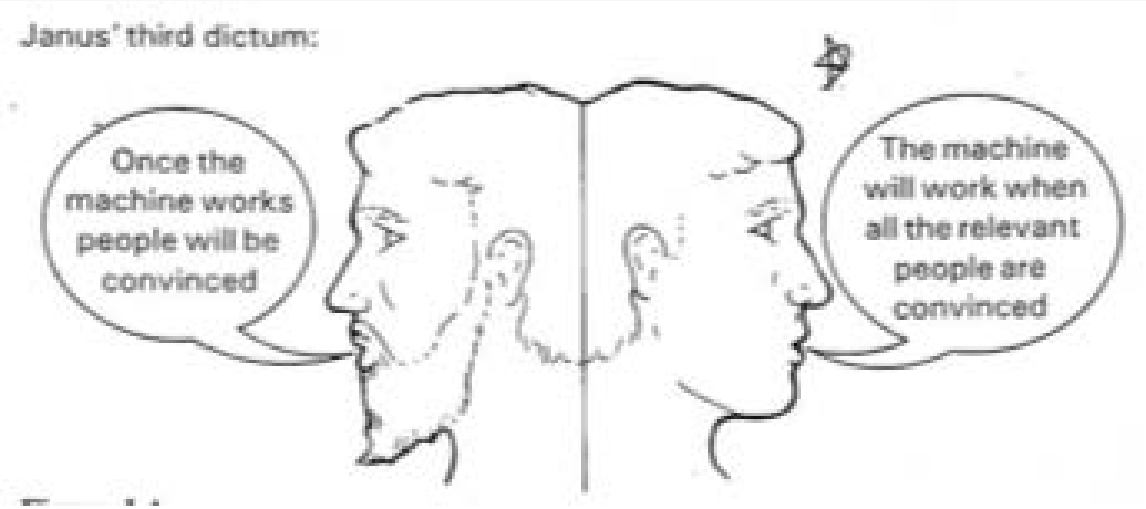
- We can work with all of these folks

Language issues between HASS and HPC

- We speak different languages… words overlap
  - Model
  - Data
  - Simulation
  - Big
- We have different expectations
  - Black and white vs. grey answers
  - Researcher fit to the computer vs. the computer fit to the researcher

Working with HPC and HASS researchers

- A different way of thinking……



Bruno Latour – Science in Action

The workflow issue

- Batch vs. real time / on demand interactive

Technical hurdles in HASS computing

- Data / Digitization
- Application development

- There is help available!

Getting Started With XSEDE

- Contact me!
- Resources are free! (by proposal)
- Technical assistance available

Getting Started With XSEDE

- Eligibility
  - U.S. Researchers
  - International collaborations with a U.S. PI
  - NSF Fellows

Getting Started with XSEDE

- Startup Allocations
  - For exploring, testing, timing, getting started
  - Low overhead application (XSEDE Portal, demographics, short abstract)
- ECSS
  - Assistance from technical experts
- XRAC Allocations
  - Peer reviewed (after experience with startup)

Survey of some of the current and proposed HASS projects

Virtual Worlds Exploratory (VWE)

- Game log analysis
- Massive networks
- Predict behavior
- Compare with other scenarios

Model Networks in Public Health

- Spread of disease
- Analysis of behavior

Testing Multiple Specifications of Theories of Decision Making

- Michel Regenwetter
- Comparing different models of decision making

Search Engine Results Analysis

- Investigate search engine results with respect to perception / portrayal of groups
- Different SERP algorithms, etc.

Digital Humanities Text Analysis and Mining at Large Scale

- Beth Plale
- Indiana University
- Stewards of Hahti Trust
- Exploring the types of things they can do with the large document corpus
- Also have educational allocation:
  - UnCamp for digital humanities with HathiTrust corpus

“Bandits and Browsing: Data Mining and Network Analysis for Library Collections”

- Harriet E. Green
- “This project will build a scalable system for library collection analysis and recommender system development. Based on the data analyses resulting from this project, the team would begin development for an enhanced recommender system for library catalogs and digital libraries that retrieves richer search results from a library collection search based on network analysis of subject relevancy, circulation data of items, and usage data for items that share interrelated subjects. In order to build this test bed for algorithm and functionalities in the recommender system, the project will utilize the advanced computing resources of XSEDE to develop self-optimizing search algorithms and network analyses that would run against the bibliographic and catalog data in library catalogs and digital library indexes.”

An Implementation of Topic Modeling that Addresses Humanists' Interest in Historical Change

- Ted Underwood
- 500,000 texts from Hahti Trust
- Genre Classification (Machine Learning)
- Topic Modeling (Various Types)

Computationally Exploring the Underpinnings of the Civil War and Views on the South Using a Billion-Page Digitized Book Archive

- Vernon Burton
- “We have assembled nearly two billion pages of digitized materials from the nineteenth and twentieth century to perform the most extensive analysis ever performed of nineteenth century views on the Civil War and the South. Using Clemson's Palemetto supercluster and XSEDE's Blacklight systems we are performing a wide array of emotional, thematic, and geographic analyses of this collection. Given the size of the collection, it is simply intractable for a human to ever consume even a minute fraction of the material and so computational analysis is critical. XSEDE's Blacklight system will be used for the final analysis portions of the project that require a large number of cores in a very large shared memory footprint for the final geographic and network analysis.”

Abraham Lincoln Correspondence - Proposed

- Lincoln Library
- Storage Allocation
- Letters to / from Lincoln
- ~60 TB Hi-resolution scans
- Also have processing needs for automated cropping and analysis
- Visualization

Simulating the Cultural Evolution of Literary Genres - Proposed

- Graham Sack
- Columbia University
- NetLogo
- Inspired by Pandora's Music Genome Project
- "Efforts thus far have been descriptive. Can we build a model to explore potential generative mechanisms?"

Data Harvesting

- Public Databases
- Web Crawling
- Social Media Feeds

- Terms of Use
- Technical Constraints
- Analysis (e.g. Radian6)

What We Don't Know

- How do we know what we don't know?
- Individual, group, society, world
- Identify conceptual gaps
- Machine vs. human

Evaluation

- Please fill in survey at:

  http://bit.ly/CSUXSEDE

April 10, 2014

Thanks for listening!

Alan B. Craig, Ph.D.

Associate Director, Human-Computer Interaction

Institute for Computing in Humanities, Arts, and Social Science

Research Scientist

National Center for Supercomputing Applications

University of Illinois

acraig@ncsa.uiuc.edu

XSEDE

Extreme Science and Engineering Discovery Environment