Kalev H. Leetaru
Yahoo! Fellow in Residence
Georgetown University

Kalev.leetaru5@gmail.com
http://www.kalevleetaru.com

FROM PROVIDER TO PARTNER: THE CHANGING ROLE OF LIBRARIES AND DATA MINING

A "BIG DATA" VIEW OF SOCIETY

- What does it look like to study the world through the lens of data mining?

- Mapping complete English text of Wikipedia: 80M locations and 40M dates via fulltext geocoding
- First large-scale examination of the geography of social media: Global Twitter Heartbeat
- Tracing spread of ideas through space over millions of books
- Spatial visualization of millions of declassified State Dept cables
- Compiling the world's constitutions in digital form
- First large-scale study of how social media is used in conflict
- Mapping half a million hours of American television news (2.7 billion words of closed captioning)
- First live emotional "leaderboard" for television (NBC/SyFy)
- Network diagram of the entire global news media (GDELT/Global Conversation) and 256M global events

WHAT POWERS IT?

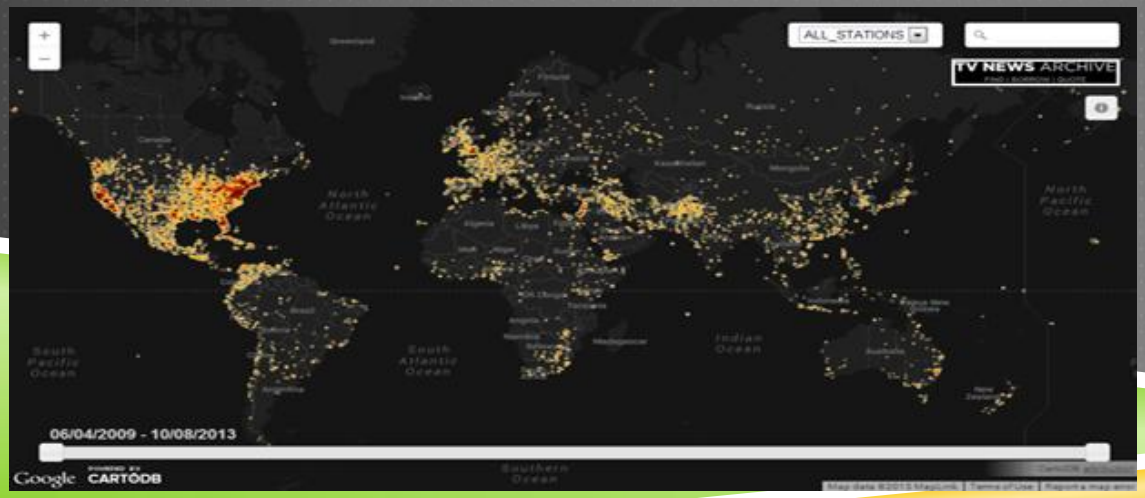
- Datasets: Wikipedia (open), Twitter (commercial), HathiTrust (~open~), Internet Archive (open), NARA (~open~), bulk digitization, global news media (commercial), JSTOR (commercial), television (VRR), web (VRR)…
- Computing platforms: experimental supercomputing platforms / engineering prototypes, SGI UV2 (64TB RAM + 4,000 CPUs in one machine) Google Cloud, IA VRR…
- Algorithms: Geocoding, Sentiment, Thematic, Topical, Network Construct, Machine Translation, OCR, Spatial Statistics, NLP, Mapping…
- Languages: PERL, R, C, C++, Java, Python…
- Tools: Gephi, Graphviz, R, ArcGIS, CartoDB, MapEngine, ImageMagick, PERL Modules

THE VIRTUAL READING ROOM

- Many of the most in-demand datasets are licensed or commercial services where data cannot be bulk downloaded, but data mining algorithms require bulk access. Example: Internet Archive's TV News Archive
- Virtual Reading Room = "virtual machine" runs on Internet Archive's physical premises. You submit your code to run on the VM where it can access all of the material, but no human can access the material, and you get back just the computed results. Removes limitations of N-Grams and other approaches. Just like a reading room in an archive, you can only take your notes with you, not any of the materials.
- Most of the major publishers have expressed interest in this model, likely to start seeing first pilot offerings in next 24 months. Will be fee-based, incremental over existing license fee. Library will be gatekeeper, handle account management and adherence to terms of use. Will place libraries squarely in a central role of enabling data mining on their campuses.

THE VIRTUAL READING ROOM

- The Virtual Reading Room provides a powerful solution to the need for bulk access for data mining, while protecting and securing intellectual property.
- Yet, also fantastic model for open collections. Assemble wide array of material in a single cloud-like environment, host on behalf of campus researchers. Customized computing environment and tools to support data mining.
- Internet Archive Virtual Reading Room used for both TV News Archive and for forthcoming "500 Years of Images" project. In latter, all books fully public and open, but VRR's unique environment vastly accelerated the development and processing cycle.

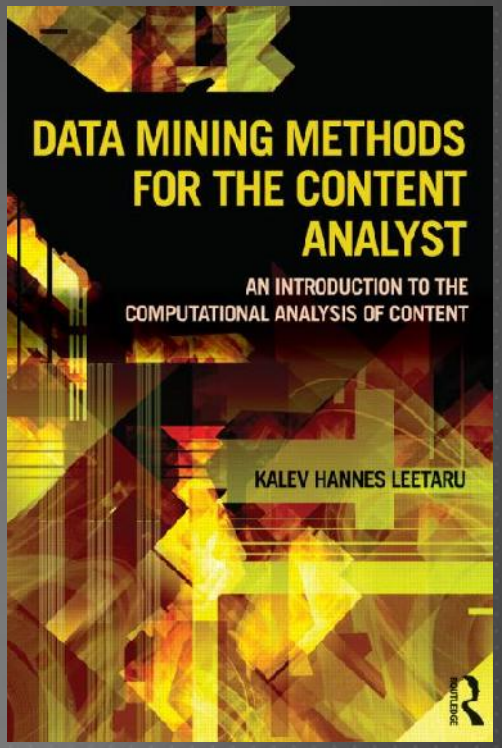


THE VIRTUAL READING ROOM

- Stable cloud environment to build common shared data mining environment for campus. Install standard tools like R, ArcMap, and Gephi. Create web interfaces and APIs to expose licensed and open source services to campus (within license agreement, such as a campus-wide OCR server API if allowed by enterprise license), or an API to an open source package.
- Cloud model makes it easy to "cloudburst" out to commercial clouds for special projects as needed, or onto NSF XSEDE resources. CyberGIS model.
- Web-based interfaces for novice users, wrap API's around tools for moderate users, and full computing environment for advanced users – all with access to the same datasets and tools.
- WARNING: not all datasets that libraries purchase permit data mining, ALWAYS check licensing agreement.

A DATA MINING SCHEMATIC

- Workflow
  - Translating a HASS (Humanities, Arts, Social Sciences) question into a computational question.
  - Securing data access.
  - Determining necessary algorithms and tools.
  - Securing computing resources.
- Lifecycle
  - What happens when the project ends?
  - Libraries as data and software repositories.



FROM PROVIDER TO PARTNER

- Libraries need to transition from being purely repositories of knowledge towards helping patrons apply that knowledge. Don't just hand a patron a book, collaborate on research.
- From PROVIDER to PARTNER.
- Columbia and Stanford's digital humanities centers are both housed in their libraries and are fantastic examples of this model. Collaborative mindset, sit down with faculty to understand their research, help them translate to a data mining approach, identify and acquire datasets and computation resources, and help execute project. Much like faculty come to library when they need a book, here they come when they need help with digital humanities.
- Help faculty understand what's possible. Purpose of my Routledge book – a "menu" that faculty can read and realize that computers can codify tone, extract topics and themes, map geography, construct networks, and visualize evolving language. Helps bridge the disciplinary gap.
- Hold regular workshops to connect faculty with potential collaborators and socialize library resources. Learn what specific data and tool needs your faculty have. Alert them to new datasets and grant programs.

FROM PROVIDER TO PARTNER

- Stanford and Columbia model of a service bureau is critical: need a standing team with diverse skills. Most HASS scholars don't have research budgets to hire students on their own.
- WARNING: Can't just hand faculty off to a CS professor working in the field. CS faculty and students only interested in technically-interesting challenges: 99.9% of HASS research doesn't cross that threshold. (Word cloud of 50 documents). Even with interesting challenge, requires translator to help disciplines talk in each other's language. Library should have standing staff and liaison with CS faculty for the largest projects.
- More CS departments require senior design courses – leverage this for no-cost skilled short-term programming support for intricate projects.
- Maintain connections with campus computing resources and fast-track cloud bursting agreements.
- DATA BROKERS.

GATEWAYS AND GATEKEEPERS

- Highest-demand datasets aren't readily available for data mining. Most publishers willing to at least have a conversation, but set very high bar.
- Tremendous damage has been done by publishers investing heavily in supporting projects that never get off the ground: many no longer willing to support academic projects without substantial cost recovery.
- Libraries can act as gatekeepers, sitting down with faculty to develop a workplan and ensure they have all of the necessary resources to complete a project before approaching a publisher and help partner faculty new to the field with more experienced ones.
- For projects with necessary resources for success, act as gateway to put them in touch with the right contacts at the data vendor and help translate their needs and act as a "guarantor".
- Some publishers willing to provide bulk exports, but only with key guarantees on data safety, security, access, and use – which libraries can monitor for them.
- Most have commercial bulk APIs, but very expensive – libraries can bulk negotiate.

INFORMATION SHARING

- Need a central mailing list and knowledge repository for announcements of new datasets, tools, programs, funding opportunities, and large-scale example projects.
- For example, Internet Archive has been looking for scholars interested in making use of its collections, such as its half-petabyte .GOV archive. How do we make the right researchers aware of these resources? (Matt Connelly at Columbia: asked him to relay .GOV archive to his colleagues, he responded with interest in scanning for FOIA reading rooms). When the image archive is released, how do we get the word out there?
- Data gift programs like the new Twitter/GNIP data access program.

LIFE CYCLE

- Residential output products of data mining projects often massive, can be tens of TB's for some projects.
- Libraries can work with faculty to identify which output products are shareable (IP and licensing considerations) and make the data available to the research community. Can require very high-bandwidth high-disk storage clouds – library can help broker or provide this.
- Increasing use of interactive web delivery of results – libraries can host specific platforms like mapping, database, and visualization platforms to provide a stable long-term environment. This is a CRITICAL area most libraries miss – faculty host in assorted cloud platforms that vanish or change a few months later and project is lost.

THANK YOU!

- Kalev H. Leetaru
- Yahoo! Fellow in Residence
- Georgetown University

- Kalev.leetaru5@gmail.com
- http://www.kalevleetaru.com