Kalev H. Leetaru
Yahoo! Fellow in Residence
Georgetown University

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

FROM WORDS TO INSIGHTS: RETHINKING CONTENT AND BIG DATA

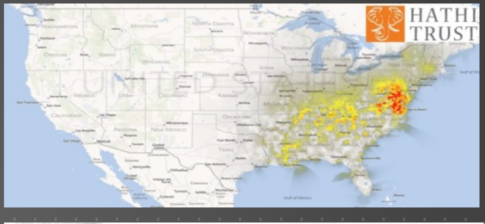
AUDIENCE QUESTION

- Have you ever provided special exports of your data to academic research projects for data mining?
  - (1) Yes
  - (2) No

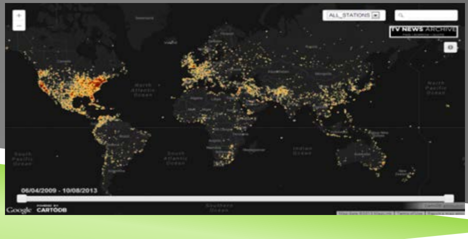
A “BIG DATA” VIEW OF SOCIETY

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

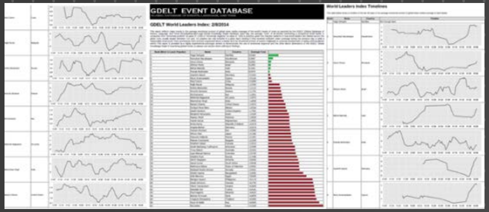


The Declassification Engine
Conference on the Computational Analysis of Official Secrecy
BRAZIL
CHILE



WHAT POWERS IT?

- Datasets: news media, social media, books, journals, video, etc…
- Computing platforms: experimental supercomputing platforms / engineering prototypes, SGI UV2 (64TB RAM + 4,000 CPUs in one machine) Google Cloud, IA Virtual Reading Room…
- 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, BigQuery, PERL Modules



GDELT EVENT DATABASE

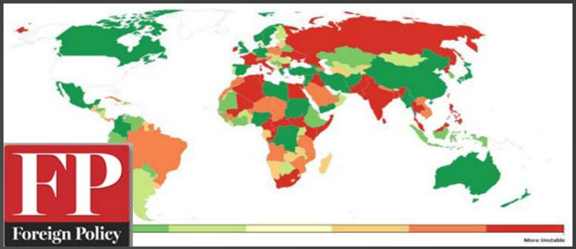
GDELT World Leaders Index: 2/8/2014

World Leaders Index Timelines

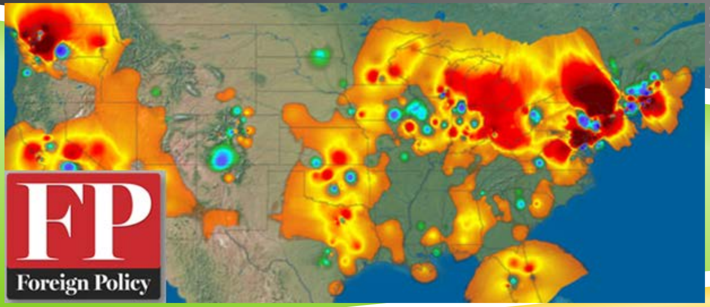
[illegible]

GDELT PROJECT

http://www.gdeltproject.org/



- Global Database of Events, Language, and Tone
- Computer-process the world’s news media to extract every mention of riots, protests, coups, peace appeals, diplomatic exchanges, etc, worldwide 1979-present, all georeferenced at the city level.
- Connect every person, organization, location, count, theme, news source, and event across the planet into a single massive network that captures what’s happening around the world, what its context is and who’s involved, and how the world is feeling about it, every single day.
- Quarter billion events 1979-present and counting…
- Daily conflict trends, world leaders index, all 100% open…
- ONLY codified numeric event records…



FP

THE GLOBAL CONVERSATION

[illegible]

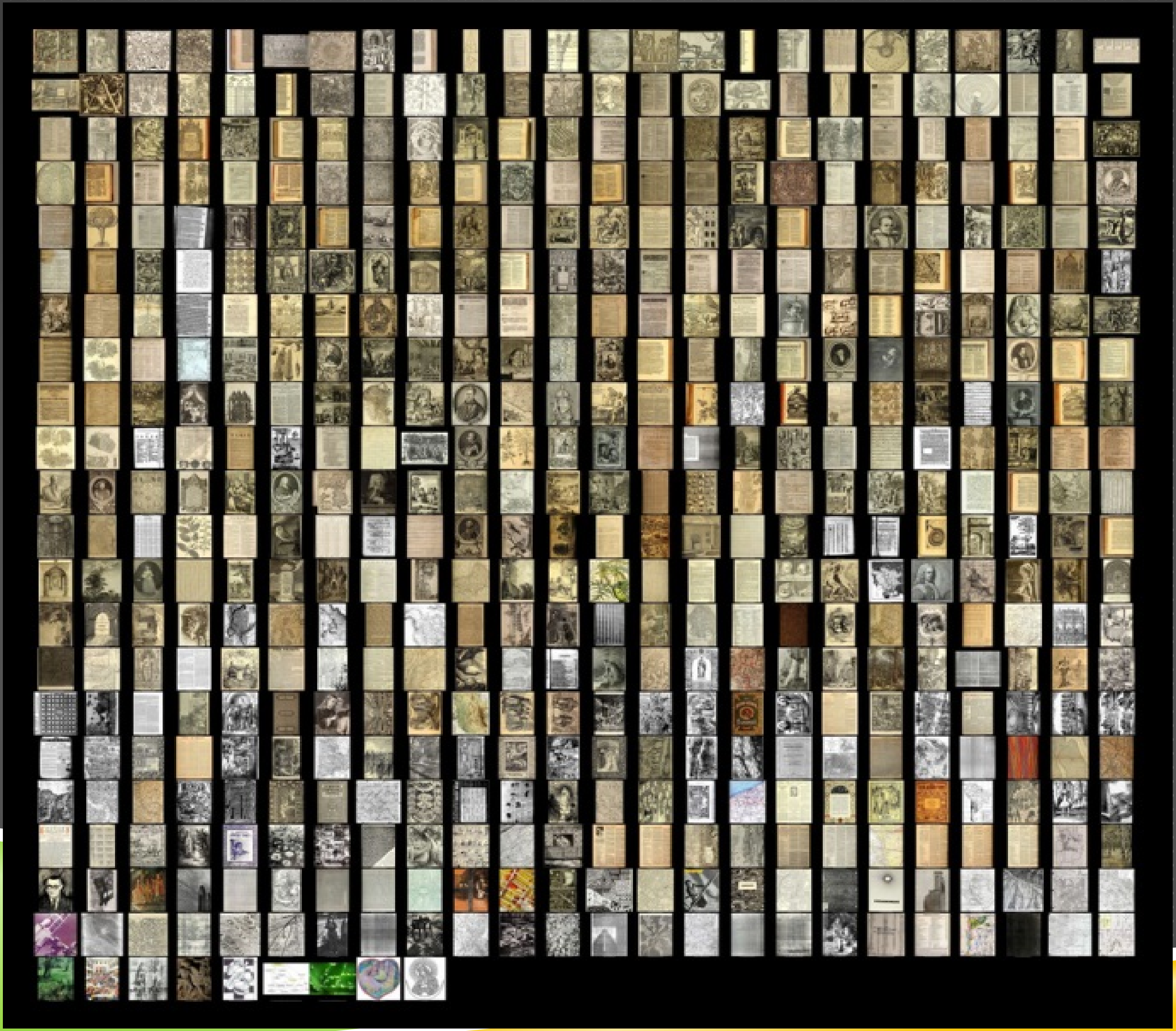
HALF A MILLENNIUM OF IMAGERY

- All current book digitization efforts have focused on recovering TEXT
- In fact, historically some have used lighting and imaging settings to enhance text for OCR at expense of fidelity of illustrations
- Images have been identified and extracted ONLY to separate them from text
- End goal is to make books fulltext searchable – books are just piles of words
- What about the images? Illustrations, drawings, charts, maps, photographs, and other visual elements…
- 300 billion words in one digital library, but how many images? No-one knew…

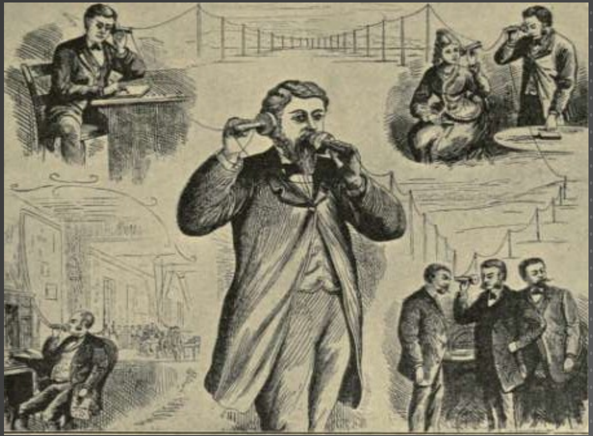
HALF A MILLENNIUM OF IMAGERY

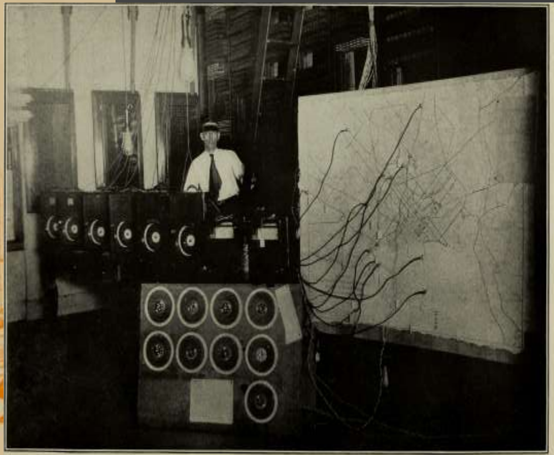
- Rise of ereaders has brought attention back to images. Image-over-text PDFs are too large for mobile devices, so need web page-like interface with ASCII text and embeddable images.
- EPUB format – ZIP file with web pages and images.
- Images are preextracted as JPEG files and embedded in text like a web page.
  - Just have to copy the images out of the ZIP file.
  - Find their reference in the web page and extract the surrounding text for context. Keyword search 500 years of images!!
  - Its literally THAT EASY!
  - No massive new technology, just a new way of thinking about content.
  - EPUB format is an industry standard, meaning this could be done on most digital libraries
- ~60M images from ~1 million books

HALF A MILLENNIUM OF IMAGERY



THE TELEPHONE



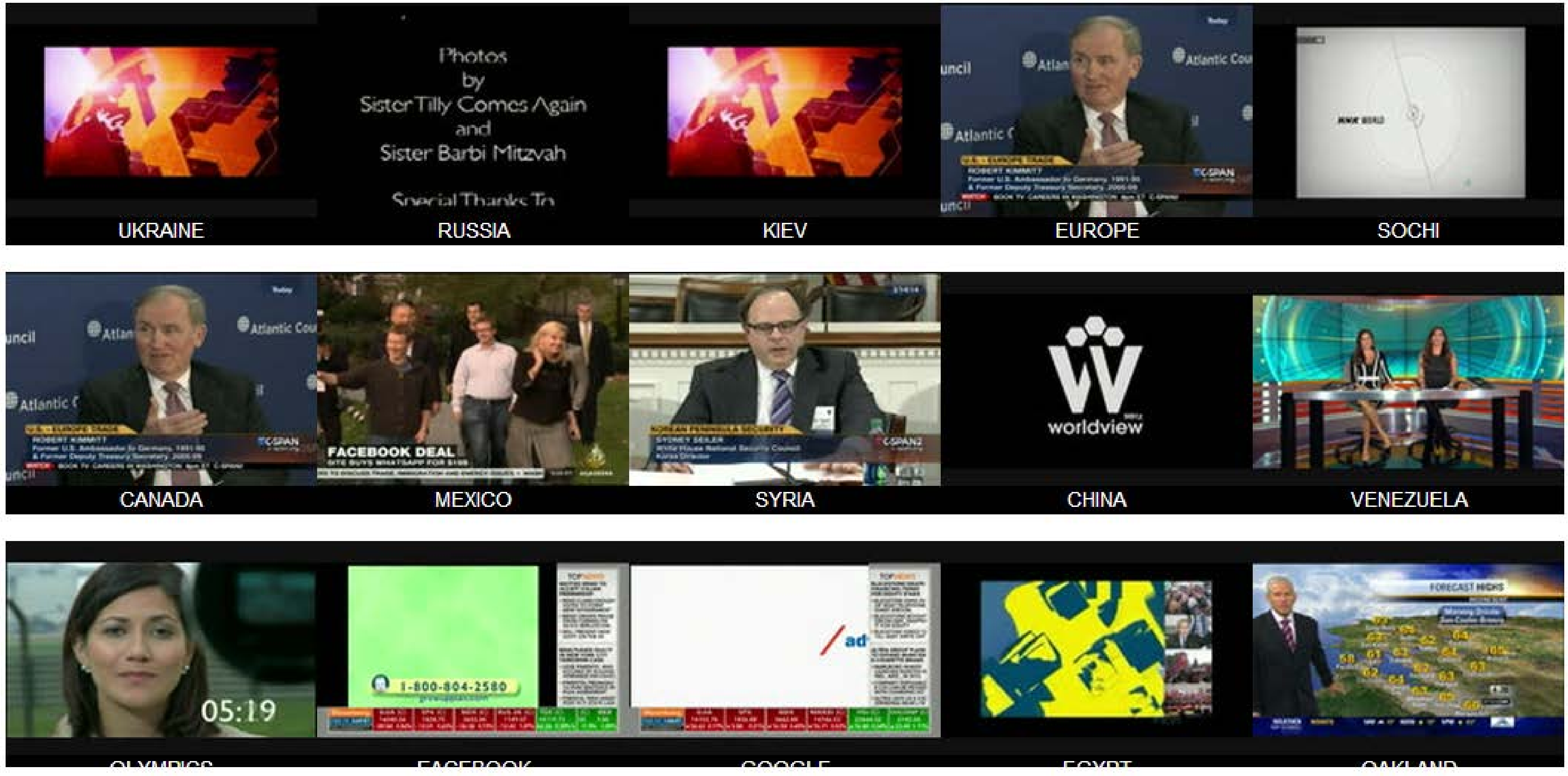


Bureau of Standards Technologic Paper No. 28

Fig. 4.—*Arrangement of connections for measuring over-all potentials*

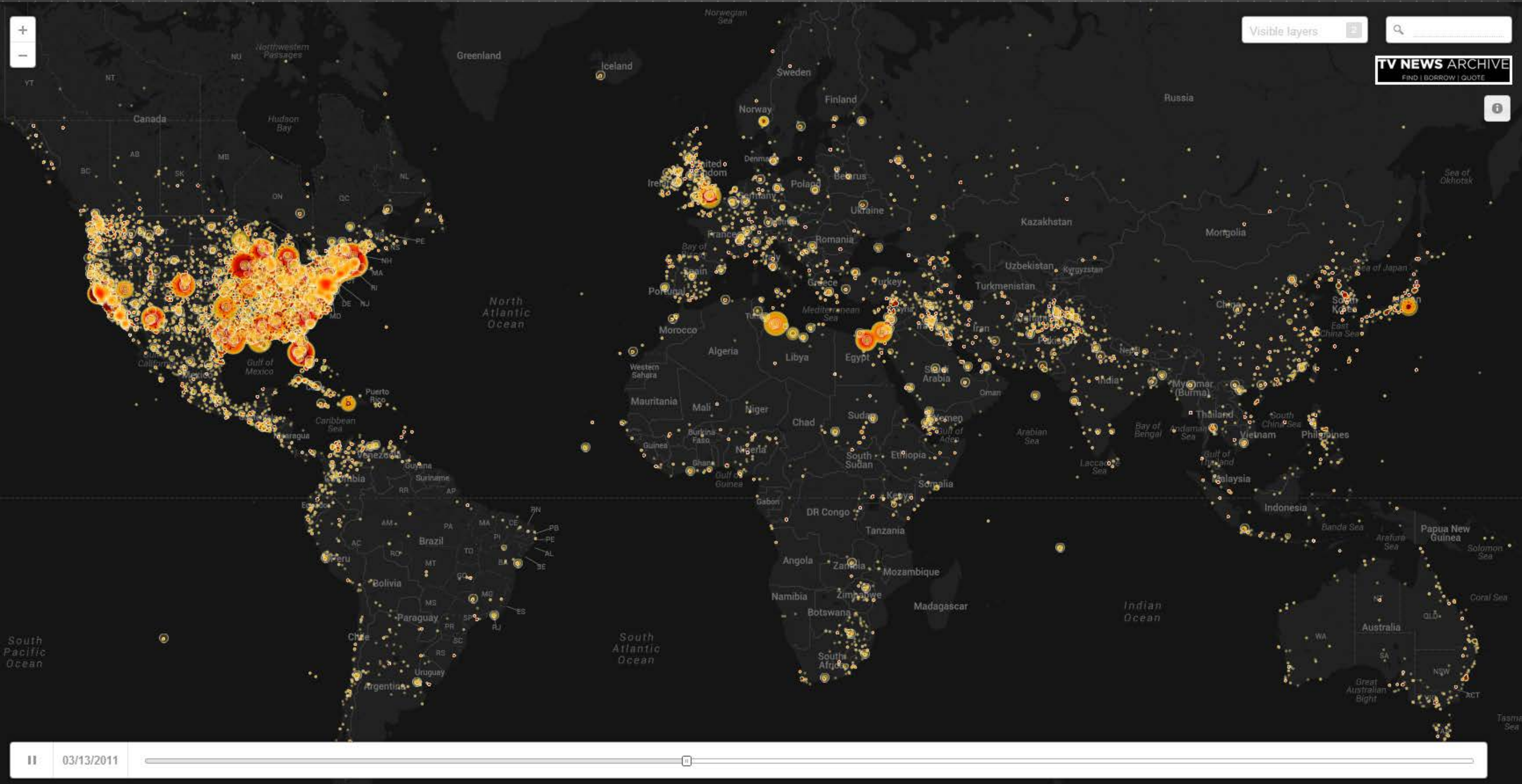
INTERNET ARCHIVE TELEVISION NEWS ARCHIVE

- Largest television news archive on earth
- 609,992 shows recorded / 488,118 shows with closed captioning streams
- 1,417 programs from 63 channels
- 2.7 billion words of closed captioning (~110 words per minute)



MAPPING THE GEOGRAPHY OF TELEVISION

- How can we creatively explore the closed captioning stream to better understand how American television news portrays the world?
- Apply "fulltext geocoding" software to the entire archive to automatically identify, disambiguate, and extract all mentions of locations in the closed captioning over all half million hours.
- All computation run on Internet Archive's servers – compute moves to data (more on this in a moment)



+
−
Visible layers
TV NEWS ARCHIVE
FIND | BORROW | QUOTE
Greenland
Iceland
Norwegian Sea
Canada
Hudson Bay
Northwestern Passages
Russia
Sweden
Finland
Norway
United Kingdom
Ireland
Poland
Belarus
Ukraine
Germany
France
Romania
Kazakhstan
Mongolia
Sea of Okhotsk
Sea of Japan
Uzbekistan
Kyrgyzstan
Turkmenistan
Greece
Turkey
Spain
Portugal
Morocco
Algeria
Libya
Egypt
Mediterranean Sea
Iran
Saudi Arabia
Oman
China
South Korea
East China Sea
India
Nepal
Myanmar (Burma)
Thailand
Vietnam
Philippines
South China Sea
Bay of Bengal
Andaman Sea
Gulf of Thailand
Malaysia
Indonesia
Banda Sea
Arafura Sea
Papua New Guinea
Solomon Sea
Coral Sea
Australia
Great Australian Bight
Tasman Sea
North Atlantic Ocean
South Atlantic Ocean
South Pacific Ocean
Indian Ocean
Arabian Sea
Laccadive Sea
Gulf of Mexico
Caribbean Sea
Puerto Rico
Western Sahara
Mauritania
Mali
Niger
Chad
Sudan
Yemen
Gulf of Aden
Burkina Faso
Guinea
Ghana
Nigeria
Gulf of Guinea
South Sudan
Ethiopia
Somalia
Kenya
Gabon
DR Congo
Tanzania
Angola
Zambia
Mozambique
Namibia
Zimbabwe
Botswana
Madagascar
South Africa
Venezuela
Colombia
Guyana
Suriname
Ecuador
Peru
Brazil
Bolivia
Paraguay
Chile
Uruguay
Argentina
Nicaragua
Mexico

03/13/2011

AUDIENCE QUESTION

- Are you currently exploring ways of making your content more available for data mining?
  - (1) Yes
  - (2) No

HOW DO WE ENABLE DATA MINING

- In talking with many of your companies, the #1 request I've heard is "how can I permit data mining while protecting my intellectual property?"
- Cheap and ready availability of data mining tools are increasing interest across disciplines and scholars are clamoring for bulk access to content to data mine.
- Academics beating down many of your doors asking for bulk exports and/or trying to work around your security tools to bulk download from your search interfaces or paying armies of undergrads to sit in computer labs and download millions of articles.
- Its not just academics! Wall Street seeing huge potential in forecasting, medical community increasingly data mining to find patterns in disease spread and treatment.
- What if there was a way to satisfy all of them AND create a new source of revenue?

HOW DO WE ENABLE DATA MINING

- Nearly all of the social media companies (Twitter, Facebook, Wordpress, Flickr, YouTube, etc) all make live APIs or streams available of their public content. Resell via commercial vendors like GNIP.
- Social media data now one of the top data sources for computer science because of its availability in data mining-friendly format.
- Yet, they own their own content, so can redistribute it – how can companies likes yours enable data mining?

THE VIRTUAL READING ROOM

- Data mining algorithms require bulk access to large portions of a collection, but not sustainable to just box up large portions of your holdings and ship to researchers (and often not permitted under license terms).
- Three most common approaches today:
  - Special extracts shipped on USB drives (great for researchers, but not sustainable or not permitted)
  - Page/Article-level APIs (most data mining tools need access to whole corpus for normalization – NYTimes has shrunk 50% over 60 years – need to know this to trace a keyword over time)
  - Custom extracts like "n grams" (most data mining tools need access to the text to see the words in context and in order, not just how many times a word was used by year)

THE VIRTUAL READING ROOM

- There's a fourth possibility:
  - Hosted "cloud" computing where researchers run their codes directly on the content holder's servers
- Think of digital libraries like ARCHIVES and not LIBRARIES. Library you check material out and read at home: content travels to analysis. Archive you physically travel to the reading room, peruse the material there, and can only take your NOTES with you – analysis travels to content.

THE VIRTUAL READING ROOM

- 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.

THE VIRTUAL READING ROOM

- For Internet Archive Television News project, we used this model to geocode the 2.7 billion words of closed captioning content.
- TV material cannot leave IA's physical premises – just like an archive.
- Instead, submitted geocoding algorithms to Internet Archive and they ran on a server physically located on IA's premises. ONLY the final list of locations (such as "Kiev, Ukraine") were returned.

THE VIRTUAL READING ROOM

- The VRR is a three-part model: technical, methodological, and legal.
- Your companies could use this same model to enable data mining of your own collections.
- Essentially a researcher would get a login to a sandboxed computer running in your data center, walled off and secured.
- Computer has access to all of the material researcher's institution subscribes to and data mining tools can run on all of the content, but NONE of the content can leave, ONLY codified research results.

- Not a cost center, instead a profit center. Make "data mining access" an optional extra on your subscription package. For an incremental additional fee, subscribing institutions get a set of VM logins for their researchers.
- Means researchers have a LEGAL and AUTHORIZED platform for all of their data mining needs, the data is even in data mining-friendly format ready to go, and you have a new revenue stream!

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 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

- Huge trend towards "infrastructure as a service". Your companies have invested massively in providing highly specialized infrastructure for storing, preserving, processing, and making searchable vast volumes of material.
- Could offer this model to companies and researchers – host their data in your infrastructure as a revenue model.

AUDIENCE QUESTION

- Are you currently exploring making more sophisticated kinds of search available, like geographic, topical, or entity search?
  - (1) Yes
  - (2) No

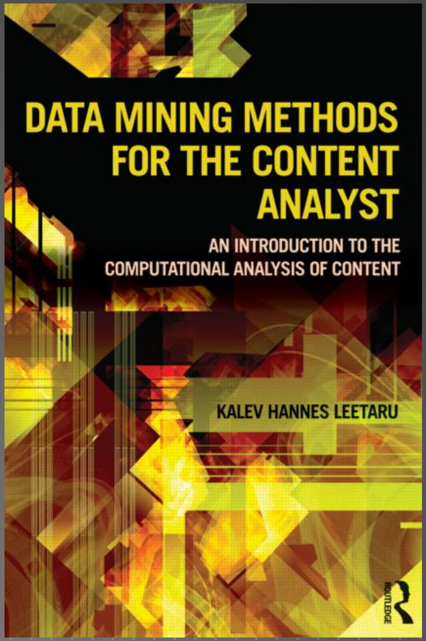
ITS NOT JUST ABOUT THE TEXT

- Huge shift in how content is used – users today think in terms of information, not its containers. They want to jump right to the piece of information they need.
- The new race is not just to provide the most content in one place, but to provide the best access to that content.
- What if I want all news coverage from Southern Ukraine? Today would require typing in a MASSIVE boolean "OR" search with hundreds of cities and landmarks. Instead, what if I could just circle that region on a map?
- What about building timelines or networks around key organizations, individuals, and locations?
- Outside of CS, academics want to study information, not mine it – if you can do all of this for them, it's a HUGE value proposition.
- Partner with key academics in special pilots to apply their tools to your data – they get publications, you get specialized results from data mining your content that you can then explore to see what tools would work best on your material. Win-Win!

THANK YOU!

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

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


DATA MINING METHODS FOR THE CONTENT ANALYST
AN INTRODUCTION TO THE COMPUTATIONAL ANALYSIS OF CONTENT
KALEV HANNES LEETARU
ROUTLEDGE