

Web Archiving and Digital Library Projects and Technologies

Edward A. Fox (fox@vt.edu) and Zhiwu Xie (zhiwuxie@vt.edu)

With slides and through collaboration with:

Co-PIs: Jefferson Bailey (IA), Andrea Kavanaugh, Donald Shoemaker,
Steve Sheetz

GRAs: Prashant Chandrasekar, Liuqing Li, Ziqian Song

Former GRAs: Mohamad Farag (Alexandria), Sunshin Lee (Radford),
Islam Harb (on leave), Seungwon Yang (LSU), . . .

Presentation for Linux/Unix Users Group @ VT (VTLUUG)

15 March 2018

<http://fox.cs.vt.edu/talks/2018/20180315WADL-VTLUUG.pdf>

Outline

- Context
- GETAR proposal
- IDEAL results – Sunshin Lee
- IDEAL results – Mohamed Farag
- Selected GETAR projects
- Welcoming collaboration

Context - 1

- Understanding the world by collecting, archiving, analyzing
 - Providing access to information: digital libraries
 - From theory: 5S (Societies, Scenarios, Spaces, Structures, Streams)
 - To algorithms, applications, systems, collections, user studies
-
- Library related
 - Any type of information: multimedia as well as text
 - Applications: archaeology, autism, civil engineering, education, epidemiology, events, fingerprinting, fisheries, global change, hurricanes, national archiving (Qatar), neuroscience, news, physics, school shootings, sociology, trails, Web, . . .

Context - 2

- Digital Library Research Laboratory, 2030 Torgersen Hall
 - Director, Edward A. Fox, <http://fox.cs.vt.edu>
- University Libraries
 - Center for Digital Research & Scholarship <http://scholar.lib.vt.edu/staff/zxie/>
 - Zhiwu Xie, Director, Digital Library Development
 - Digital Libraries & Repositories <https://lib.vt.edu/collections/digital-library.html>
- WADL Workshops (2013, 2015-2018):
<http://fox.cs.vt.edu/wadl2018.html>
- VTechWorks sites for DLRL and related courses
 - <https://vtechworks.lib.vt.edu/handle/10919/18732>, 47780, 19081, 18655, 50956

Part 1

- Extracts from the GETAR proposal (NSF IIS-1619028 and 1619371)
- Virginia Tech and Internet Archive, 2016 – 2020
- **Global Event and Trend Archive Research**
 - <http://eventsarchive.org>

Problems / Questions

- How can *K-12 students, the general public, and interdisciplinary teams* study and research the **important global events and trends** that relate to *worldwide grand challenges*?
- How can information systems support those needs in an integrated fashion, empowering users through **interaction with content** across the broad *information life cycle*?
- How can the growing collections of **Internet archives** be integrated with both the *constantly changing current version of the WWW* and stream-oriented communications like *tweets*?
- How can those involved in *planning, policy making, innovation, economics, engineering, and the social sciences* engage in focused as well as **longitudinal studies (from the End of Millennium to the present: 1997-2020)**, in an interdisciplinary context, through those systems and collections?

Goals

- To aid *interdisciplinary research and education*, regarding **important global events and trends**, which can benefit from *DL access to Internet content*, starting in *1997, up through 2020*.
- To develop next generation interactive and integrated information systems, **connecting DLs and archives**, connecting sources and documents, and connecting *webpages and tweets* (and other user submitted content).
- To advance the state-of-the-art in *DL and NLP* with regard to handling *archives, analyzing documents, adding value* to metadata and collections, and **expanding the scope of interaction** across the information life cycle.

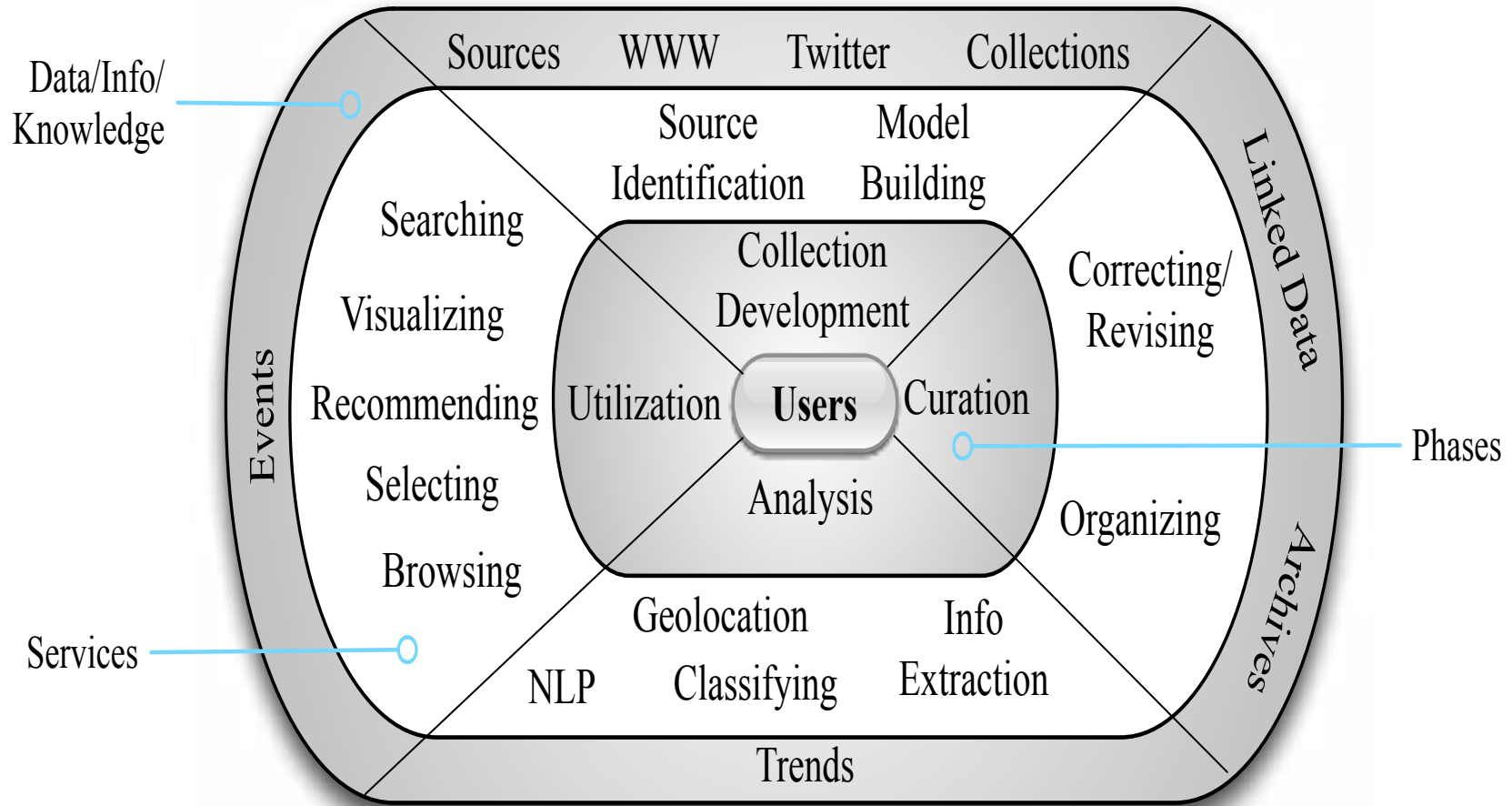
Objectives - 1

- To develop methods, deploy them *at the Internet Archive (IA)*, and
 - **get high quality collections from IA's** archives (of WARC files),
 - aiming to *find all relevant webpages*
 - (including forms like Usenet posts that have similar functionality to tweets)
 - *for important events we identify over the period 1997-2000.*
- To devise *interactive* techniques resulting in *rich models for events and trends*,
 - that will lead to **enhanced focused crawling**, accurate *classifiers*, and
 - helpful *information visualization*.
- To devise interface development methods for DLs, that lead to
 - generic solutions where possible, but also facilitate **tailoring interfaces**
 - to the needs of particular disciplines.
- To **aid stakeholders**, through interactive interfaces, to engage in
 - **development and curation** of collections of tweets and webpages
 - about *events and/or trends*, with high quality, that will support the
 - *needs in their discipline*, as well as assist in interdisciplinary research studies.

Objectives - 2

- To **aid stakeholders**, through interactive interfaces (with NLP), to *deal with errors, spam*,
 - variations in doc. length / structure / focus, multiple languages /sublanguages, and their
 - varying needs for analyzing and representing/describing/summarizing interesting content.
- To **aid stakeholders**, through interactive interfaces, to *analyze, visualize, and access*
 - those collections (with maps, timelines, social networks, and faceted browse, search, and
 - exploration of content), in ways appropriate for their needs and disciplines, and to
 - integrate the interaction with the collection development and curation activities.
- To **integrate DL and archive methods** so collections can both be preserved
 - for the long term, and easily accessed through highly interactive interfaces.
- To **aid stakeholders** to advance *research and education concerning important global events*
 - and trends, including climate change, development, disasters, energy, globalization,
 - innovation, policies, resilience, social movements, and violence.

GETAR Architecture

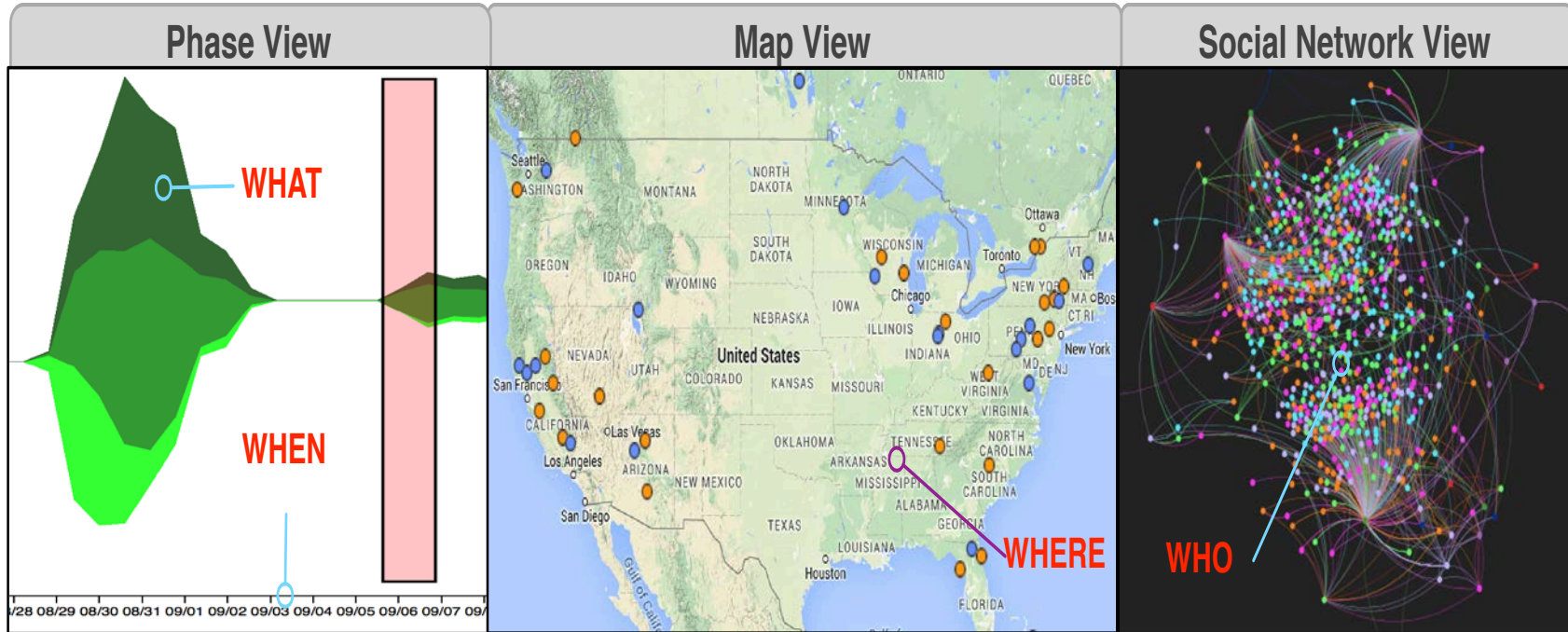


Selected Events from 54 Identified at
End of Millennium (EoM, 1997-2000)

Year	Event
1997	El Nino warms ocean wildlife (thru '99)
	Toyota Prius makes a debut in Japan
	Kyoto Global Warming Conf/Protocol
	'bird flu'; Hong Kong kills 1M chickens
	Hong Kong (transfer of sovereignty)
	Massacres in Algeria
	Iraq expelled US weapons inspectors
1998	Microsoft releases Windows 98
	Smoking ban: CA restaurants & public
	Al Qaeda bombs US embassies
	Ethiopian-Eritrean War: >10K dead,
	Hurricane Mitch in Honduras
	Congo/Africa's World War: kills >2.5M
	Google, Inc. is founded

Year	Event
1999	UN announces the 6 billionth baby born
	Napster (music download) debuts
	Kosovo War: NATO air strikes
	Turkey: Richter scale 7.4 earthquake
	Columbine High School shooting
	Two viruses afflict computers worldwide
	Kargil War between India and Pakistan
2000	Y2K (Year 2000 problem)
	Toyota released Prius worldwide
	Cyclone Eline Mozambique
	Al-Qaeda attack on USS Cole in Yemen
	Mad cow disease alarms Europe
	W. Nile Virus: Israel, France, Jordan, US
	US Bush v. Gore election; no FL recount

PhaseViz



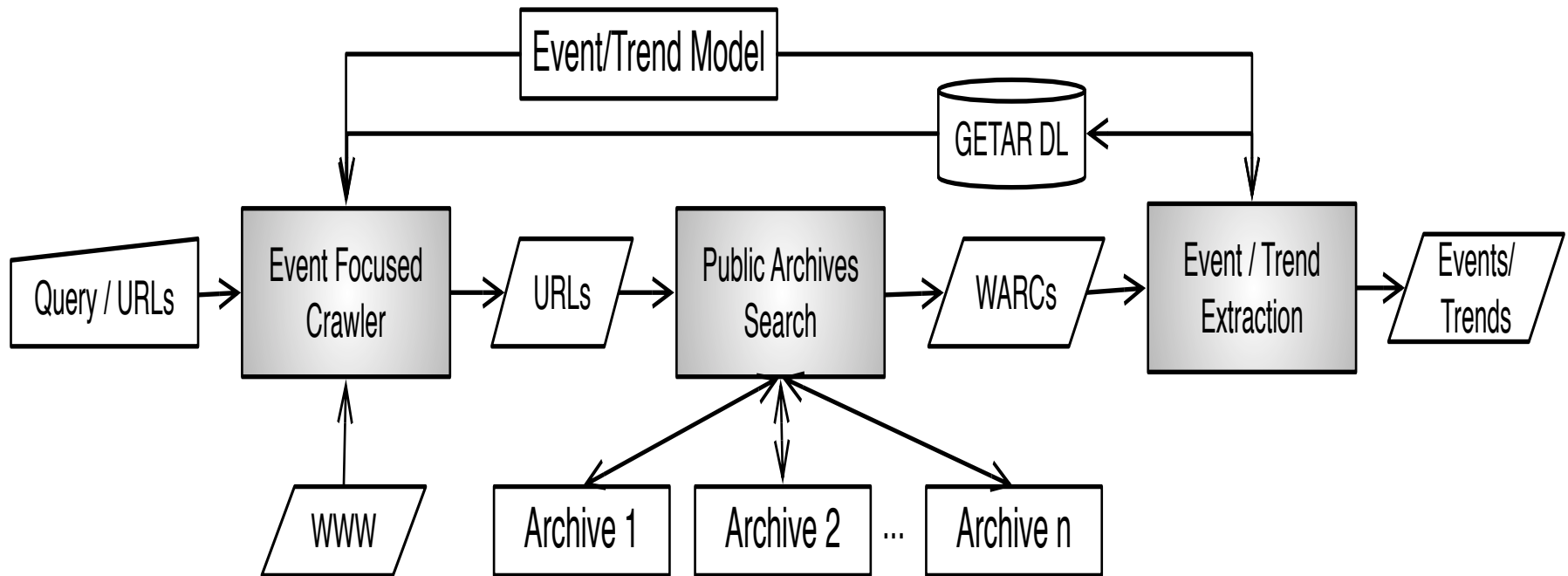
Tweets in Selected Region (Classes: 1-Response, 2-Recovery, 3-Mitigation, 4-Preparedness)

ID	is_R	Text	WHAT	Class	Date	Timestamp
33836	0	More #RedCross volunteers prepared to drive straight into #Isaac path #Fox59http://trib.al/dQAORO		4	Wed, 29 Aug 2012	1346209617
33829	1	RT @CraigatFEMA: Hurricane #Isaac, dangerous storm surge, heavy rainfall to be follow by additional flooding http		4	Wed, 29 Aug 2012	1346209620

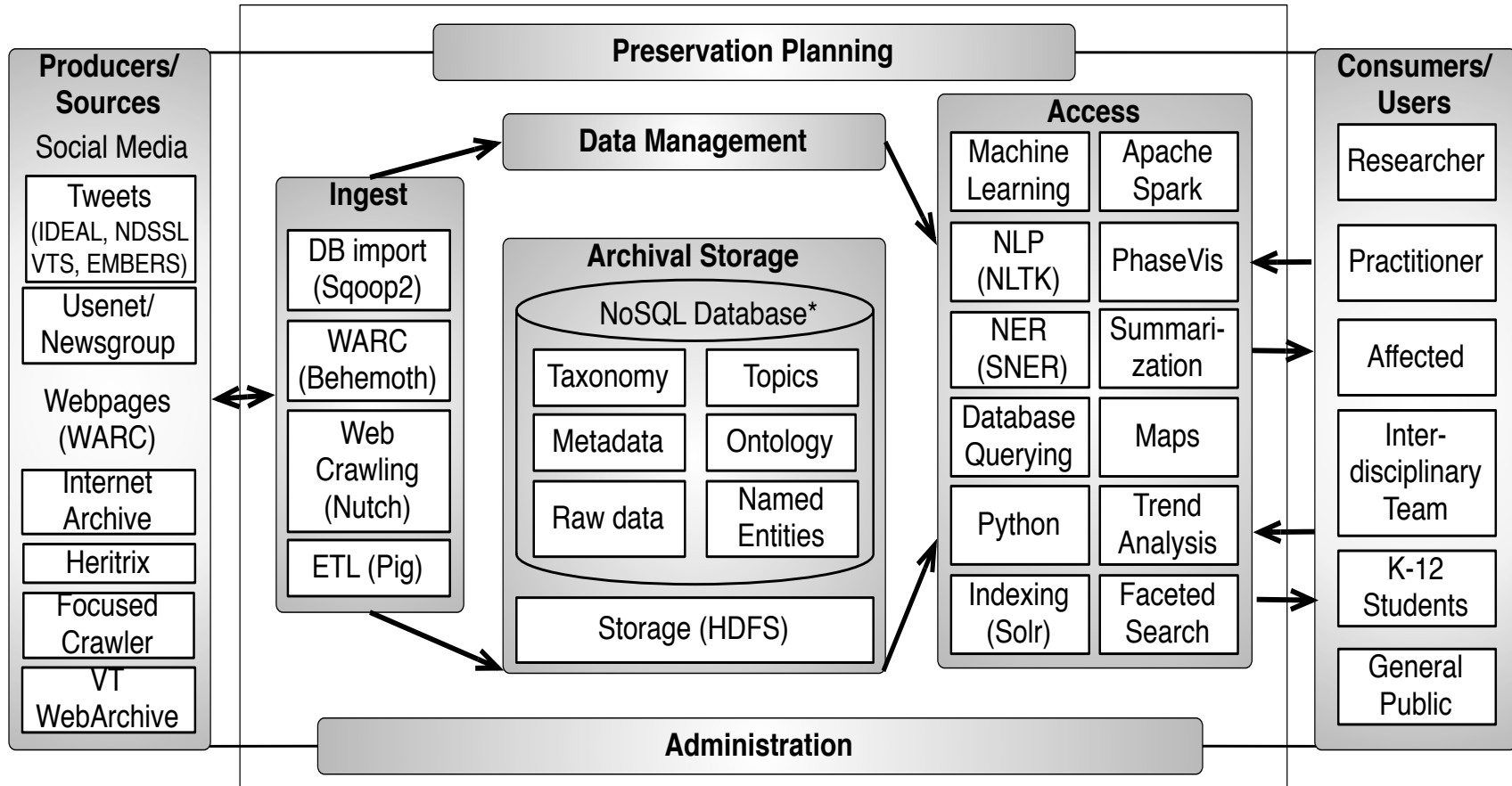
**IDEAL event webpage collections:
CS4984 (Computational Linguistics)**

Category	Collection	Event Year(s)	Location	# of Webpages
Disease	Ebola	2014	World	15,000
Earthquake	Virginia earthquake	2011	Virginia, USA	8,765
Fire	Brazil club fire	2013	Brazil	690,281
Flood	Pakistan flood	2011	Pakistan	20,416
Hurricane	Hurricane Sandy	2012	East Coast, USA	75,929
Shooting	Tucson shooting	2011	Tucson AZ, USA	37,829
Community	Blacksburg events	2011-2012	Blacksburg VA, USA	16,024

Archives, the GETAR DL, and WWW,
used to extend IA extraction results



GETAR DL architecture



Area	Description	Investigators	Courses
Core Research			
Analysis, access	Databases, HCI, information visualization, machine learning, ontologies, statistics	Fox, Franck, Huang, North, Sheetz	BIT 4524, 4544, 4614; CMDA/CS/ STAT3654; CS5764
Library, information, data	Archives, big data, curation, data management, decision support, exploring, knowledge engineering, searching	Fox, French, Nicholls, Speer, Thomas, Zobel	CS4624, 5604, 6604; FOR3604; GRAD5134
NLP	Arabic, document analysis, errors, information extraction, summarization, topic identification	Eubank, Fox, Rozovskaya	CS4624, 4984, 5984, 6804
Applied Research Across Disciplines			
Geospatial	Car crashes, data structures, GIS, maps, queries, traffic, tweets, weather and crops	Baird, Lu, Sforza	GEOG1115, 1116
Simulation	NDSLL: epidemiology, diffusion in networks, planning response	Eubank, Lewis, Swarup	GBCB5874, 7994
Climate change	Adaptation, biodiversity, conservation, ecology, ecosystems, effects on plants& animals, environment, sea-level rise	Bukvic, Jelesko, Kalkstein, Quinn	GEOG2994, 4974, 4994; PPWS4994
Economics	Development, families, game theory, Middle East, smart cities, social networks	Ball, Korkmaz, Salehi-Isfahani	ECON3004

Area	Description	Investigators	Courses
Applied Research Across Disciplines			
Energy	Green engineering, nuclear policies	Avey, McGinnis	ENGR3124
History	Globalization, Soviet history	Ewing	HIST1214, 1215, 2124, 3394, 3554
Innovation	CIE: entrepreneurship, impact of resources, industry collaboration, social and technology based ventures	Junkunc	MGT3064, 3074, 4064, 4094
Resilience	Concentrations, dependency, disasters, evacuations, planning, policy, relocation, supply chains, urban and regional, vulnerability	Bohland, Bukvic, Lawrence, Murray-Tuite, Zobel	CEE5620, 5660; GRAD5134
Sociology	Crises, global issues, social inequality, social movements, social participation, violence, social networks, communication behavior and effects (incl. in Maasai society)	Baird, Kavanaugh, Shoemaker, Wimberley	SOC2034, 2044, 3304, 3504, 3854, 4354, 4424, 4444, 4764, 5424
Political Science	National security, world politics, nuclear policies	Avey, Nicholls	PSCI1004, 1024, 1034, 2034, 2054, 2064, 3114, 3514-6, 3524, 3544, 3564, 3515-6, 3624, 3634, 3684-5, 3794, 4734, 5254, 5264, 5284, 5384, 5424, 5444, 5464, 5474, 5514, 5524, 5584, 5624, 5634, 5644

Part 2

- Slides from Dr. Sunshin Lee
- IDEAL: NSF IIS-1319578 – 2013-2017

• **Integrated Digital Event and Archive Library**

Collecting Webpages

- Started 2007
- Used Internet Archive (IA)
 - 66 collections
 - ~11TB
- Shooting, earthquake, bombing, hurricane
- Problem: very low precision

Collection Name	Last Crawl	Data (all time) ▾	Docs (all time)
Texas fertilizer plant explosion (Ap	#98765: Feb 2, 2014	1.5 TB	7,398,544
Hurricane Sandy (October 2012)	#130896: Oct 9, 2014	921.5 GB	14,085,550
Global Food Crisis	#57243: Oct 24, 2012	716.6 GB	6,151,325
Boston Marathon Bombing: Twitte	#71632: May 24, 2013	545 GB	7,103,524
Guatemala Earthquake	#59578: Dec 2, 2012	521.9 GB	2,719,020
April 16 Archive	#5008: Apr 28, 2008	287.9 GB	4,742,265
CTRnet - Emergency Preparedness	#131089: Oct 11, 2014	176.3 GB	2,333,228
Indonesian Volcanic Eruption, Tsur	#23669: Nov 2, 2010	147.8 GB	2,528,739
Brazil NightClub Fire	#63372: Feb 2, 2013	94.9 GB	2,118,616
Virginia Tech Shootings (Decembe	#41480: Jan 3, 2012	57.9 GB	1,505,721
Northern Illinois University Shootir	#4916: Apr 16, 2008	45.5 GB	631,082

Collecting tweets

- Over 1375 collections for multiple projects

- <https://docs.google.com/spreadsheets/d/13wUfD-BI49Wkloq8ZezfqTuuwQV0PKIIS9umm0RoM80/edit?ts=59b98e32#gid=0>

Project	Collection name	Total # of tweet	Started at	Collection tool	Analysis service
IDEAL	<u>Archive DB</u>	1,657,541,394	2012	yTK ¹⁾	<u>Analysis using Hadoop</u>
IDEAL	<u>1% sampling</u>	Maintenance	2015	DMI-TCAT ²⁾	<u>Analysis</u>
IDEAL	<u>User following</u>	10,407,631	2015	DMI-TCAT ²⁾	<u>Analysis</u>
IDEAL	<u>Keyword tracking</u>	20,984,747	2015	DMI-TCAT ²⁾	<u>Analysis</u>
GETAR	<u>Collection</u>	127,148,171	2015	yTK ¹⁾	<u>Analysis using Hadoop</u>
GETAR	<u>Collection</u>	230,995,656	2016.9	SFM ³⁾	<u>Analysis</u>
NIH	<u>Keyword tracking</u>	622,692	2015	DMI-TCAT ²⁾	<u>Analysis</u>
Total		2,047,700,291			

Collection Example: GETAR Prototype

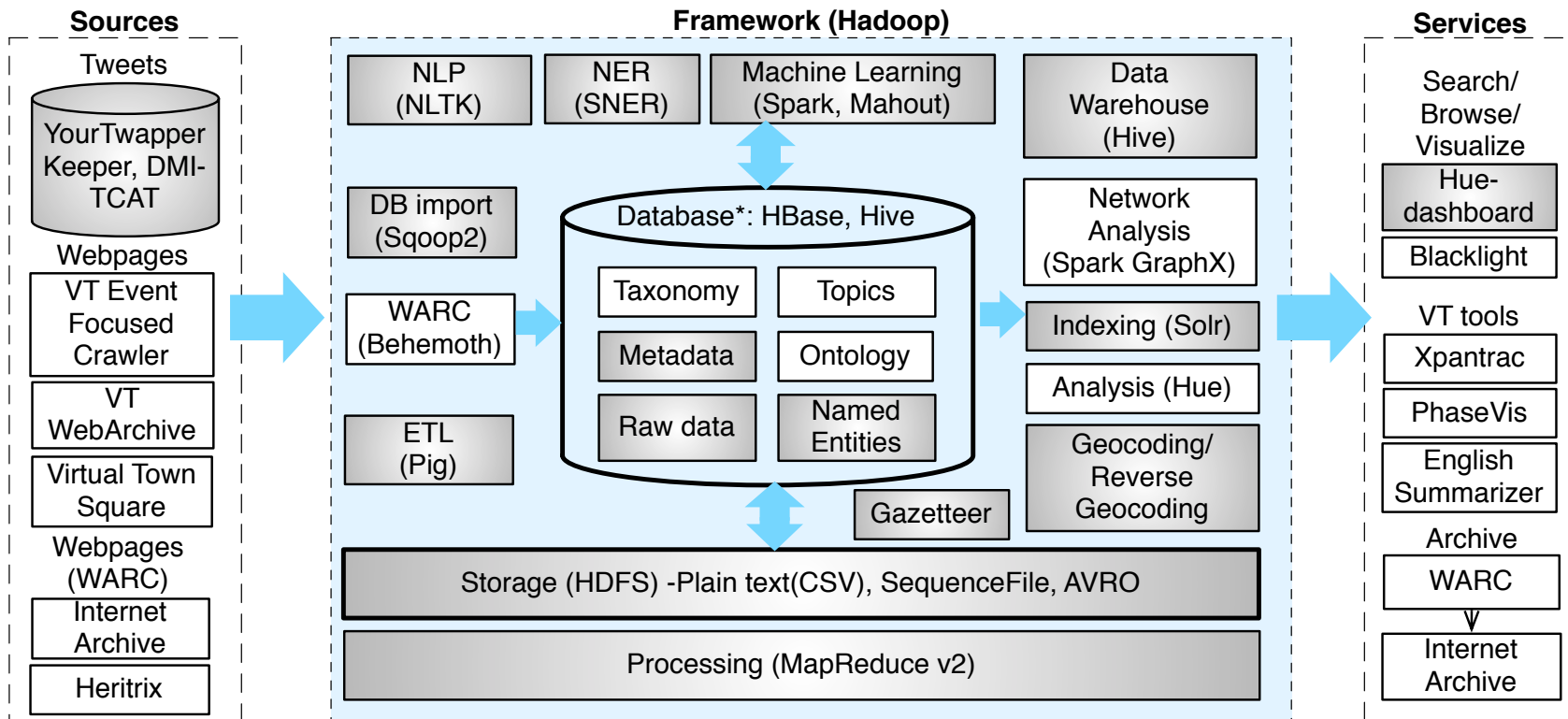
- Research key global challenges, e.g., climate change (as well as opportunities), innovation, and resilience
- Initial Collection Effort
 - Started 10/8/2015
 - 315 collections
 - 127,148,395 tweets (as of 3/15/2018)
 - Including global warming,
 - Internet of things,
 - population,
 - and the environment

Archiving and Analyzing using Bigdata Hadoop cluster

- Using Desktop PCs
 - # of Nodes: 20 + 1 (Solr)
 - CPU: Intel i5 Haswell Quad core 3.3Ghz * 20, + Xeon 8C
 - RAM: 704 GB (20 * 32 + 64)
 - HDD: 149 TB (20 * 7 + 9)
 - Backup: 32TB, 8.3TB NAS
- Servers
 - Tweet collecting
 - Web crawling
 - Geocoding
 - Search (Solr)



IDEAL System Architecture



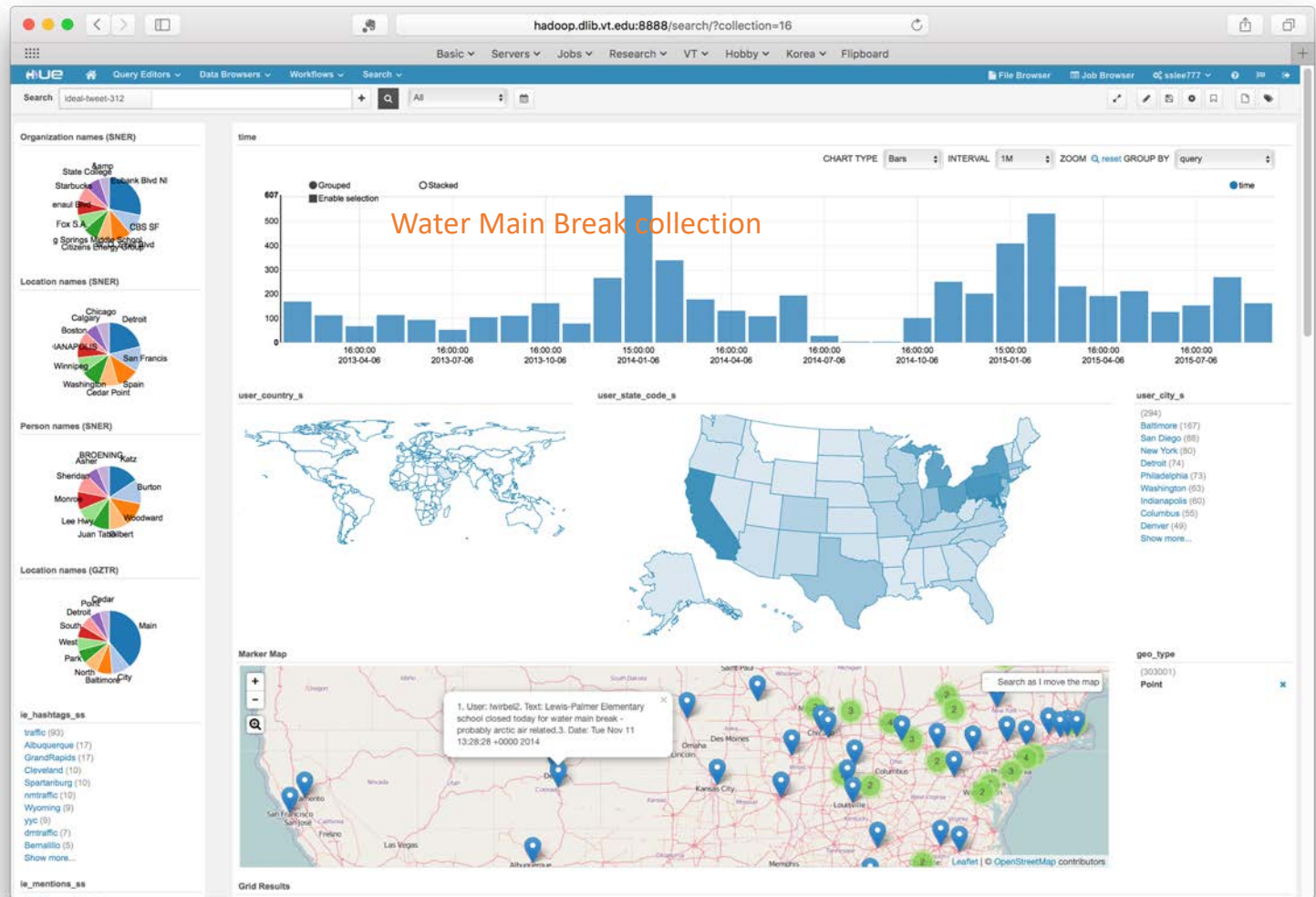
* IDEAL project: Integrated Digital Event Archiving and Library

* Highlighted (as grey) are related to this research

External Tools

- Spark and Mahout:
 - Classification, clustering
 - Topic analysis (LDA), Frequent Pattern Mining
- Solr/Lucene and (Geo)Blacklight: Search/(Faceted) Browse
- Natural Language Processing and Named Entity Recognition
 - NLTK (Python)
 - SNER (Stanford NER)
- . . .

Analysis and Visualization Example



What causes Water Main Breaks?

for **New York, NY**

Home Forecast Radar & Maps **News & Video** Severe Weather Social Fishing Your Day Star

Extreme Temperatures, Water Main Breaks Go Together

AccuWeather.com

By Alex Sosnowski, Expert Senior Meteorologist
Jan 26, 2011; 7:48 AM ET

People in St. Louis, Philadelphia, New York, Nashville and most recently Washington, D.C. have been getting quite a winter as far as problems with water main breaks.

Of course the problem is no stranger to other cities and smaller towns.

Water running through the pipes does not freeze, it is the ground around the pipes that causes many of the problems.

Meteorologist and geology buff Jim Andrews points out, "Soil is always moving very slowly down hill."



AccuWeather.com for **New York, NY**

Home Forecast Radar & Maps **News & Video** Severe Weather Social

News Video Blogs Personalities

Heat Causes 700 Water Main Breaks Daily in Houston

By Vickie Frantz, AccuWeather.com Staff Writer
Aug 21, 2011; 5:10 AM ET

Days of hot weather and aging water pipes have resulted in daily water main breaks throughout Texas.

Houston's Mayor Annise Parker told MSNBC that the city is experiencing over 700 breaks a day along 7,000 miles of pipes. The city normally averages about 200 breaks a day in the summer.

Arlington is averaging about six breaks a day, Fort Worth is dealing with about nine or 10 a day, and Ennis has gone from about one a week to about one a day, according to Cbslocal.com.



Pipe repair image courtesy of Photos.com

ACTION NEWS 6abc.com

Local/State **Montco water main break possibly a result of quake**
Tuesday, August 23, 2011

WPXI-TV PHILADELPHIA, PA

HOME ACTION NEWS 6ATV.COM MOST POPULAR LOCAL/STATE NEWS NATION/WORLD NEWS RIZZABRE NEWS BUSINESS/FINANCE NEWS CONSUMER NEWS ENTERTAINMENT NEWS



ADVERTISE WITH US

- CURRENT OFFERS
- BUILD & PRICE
- GET & QUOTE
- FIND A DEALER

TOYOTA

Home > Breaking News - MassLive.com > Palmer

Palmer water main break may be result of earthquake, says town official

Published: Tuesday, August 23, 2011, 9:07 PM Updated: Tuesday, August 23, 2011, 9:32 PM

By Patrick Johnson, The Republican

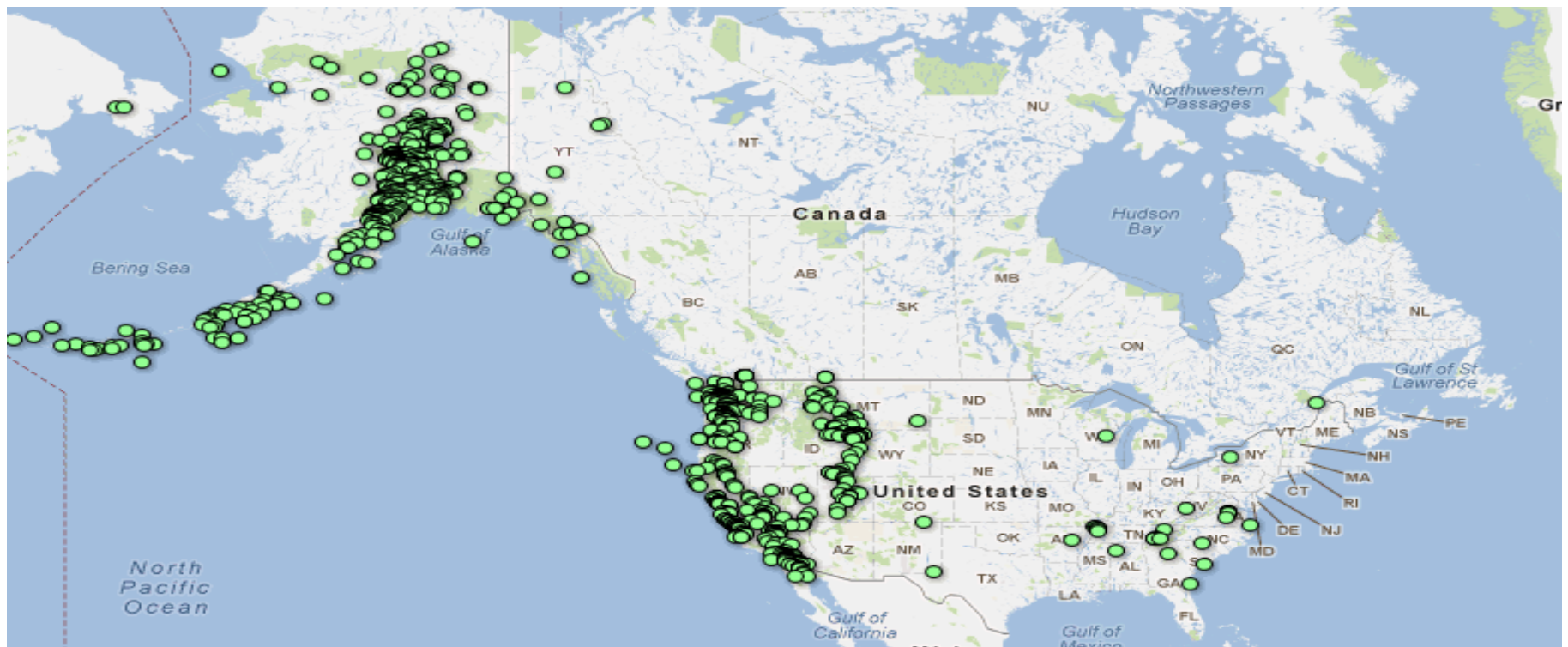
CAUSES OF WATER MAIN BREAKS

- CORROSION WEAKENS PIPES
- EXTREME HEAT OR COLD
- SHIFTING OF GROUND

© 2011 WWW.ACCUWEATHER.COM

What causes Water Main Break? => Earthquakes (USGS)

Mar. 1 – Apr. 5, 2012



Who is involved in a WMB ?

- Fix water pipe
 - Water utility
 - City/town utility
- Traffic
 - Police
- Affected
 - Citizen
- Who else ?



Lakewood, NJ, June. 2014



West Philadelphia, PA, June. 2015

Datasets for Geo-location Research

- 5 types of road side small disasters
 - Includes specific location info.
 - 3 small and 2 large collections
 - Collected 2/1/2013 to 6/30/2014
 - 17 months, start and end days varying

Size	Dataset	Number of tweets
	Total	6,039,888
Mid-size	Water main break	155,657
	Sinkhole	231,579
	Pothole	324,849
Big-size	Car crash	2,510,317
	Car accident	2,817,486

Features, combinations of features

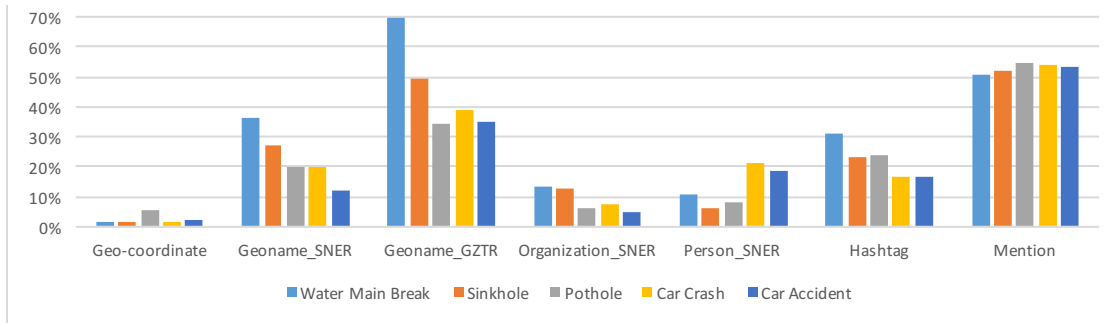


Figure 11. Percentages of tweets that have particular features

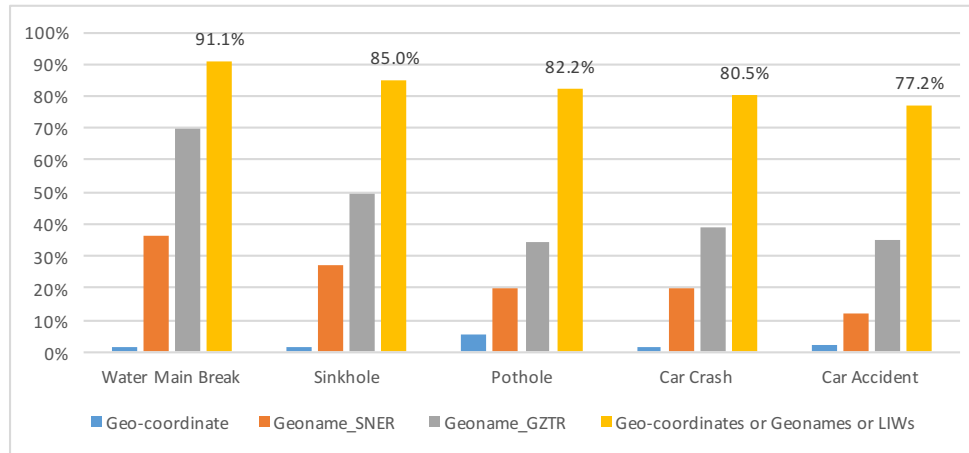


Figure 12. Percentages of tweets with features or union of features

Up to 91.1%

LIWs may help to geo-locate tweets

Part 3

- Slides from Dr. Mohamed Magdy Farag
- IDEAL: NSF IIS-1319578 – 2013-2017

• **Integrated Digital Event and Archive Library**

Archive-It Collection Quality

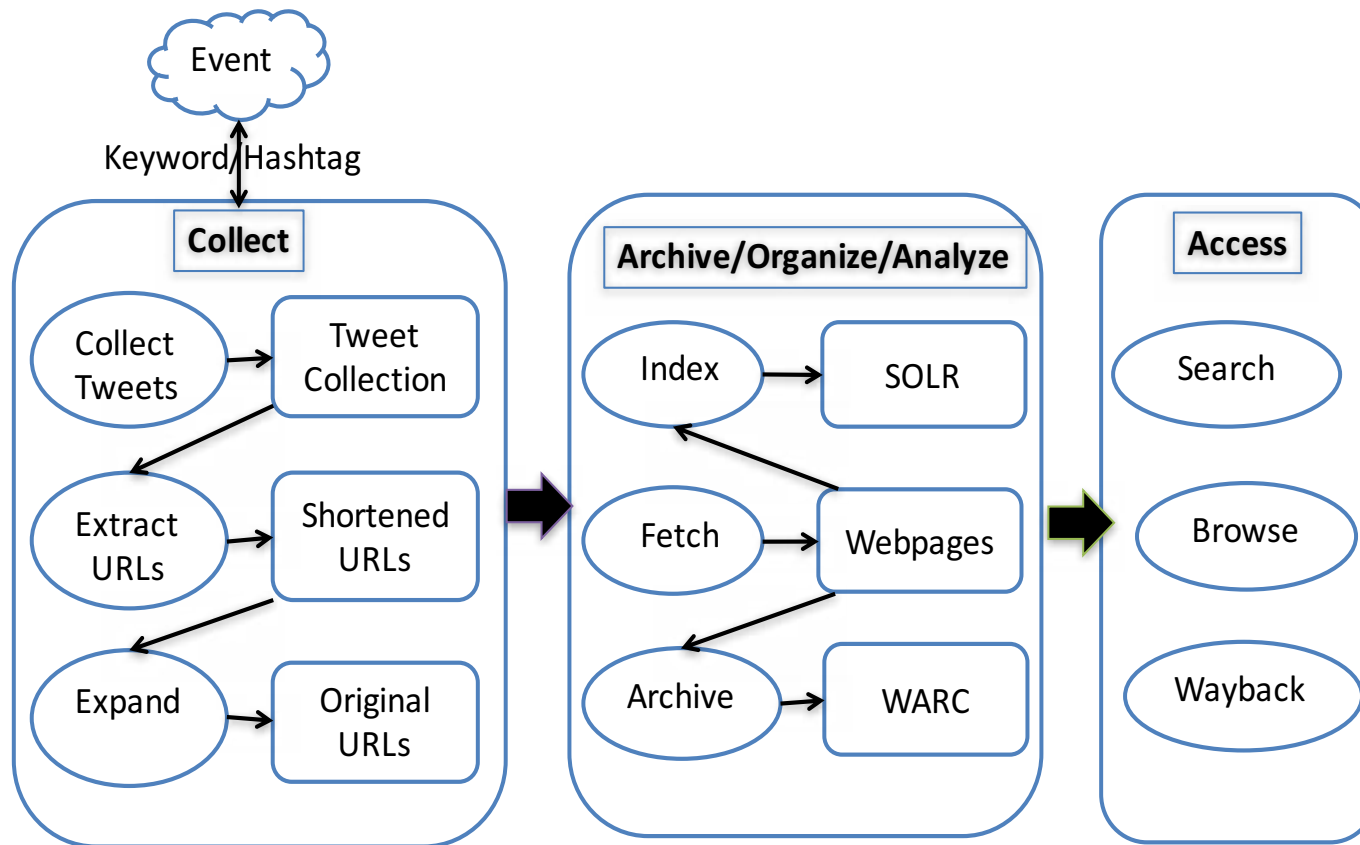
Table 4. Classification Results

Collection	Rel. (%)	Non-rel. (%)	# HTML Pages
Alabama University Shooting	1.4	98.6	6470
Brazilian School Shooting	8.8	91.2	1120
Connecticut School Shooting	17.5	82.5	3238
Northern Illinois University Shooting	26.7	73.3	15385
Norway Shooting	13.5	86.5	7419
Youngstown Shooting	40.0	60.0	3427

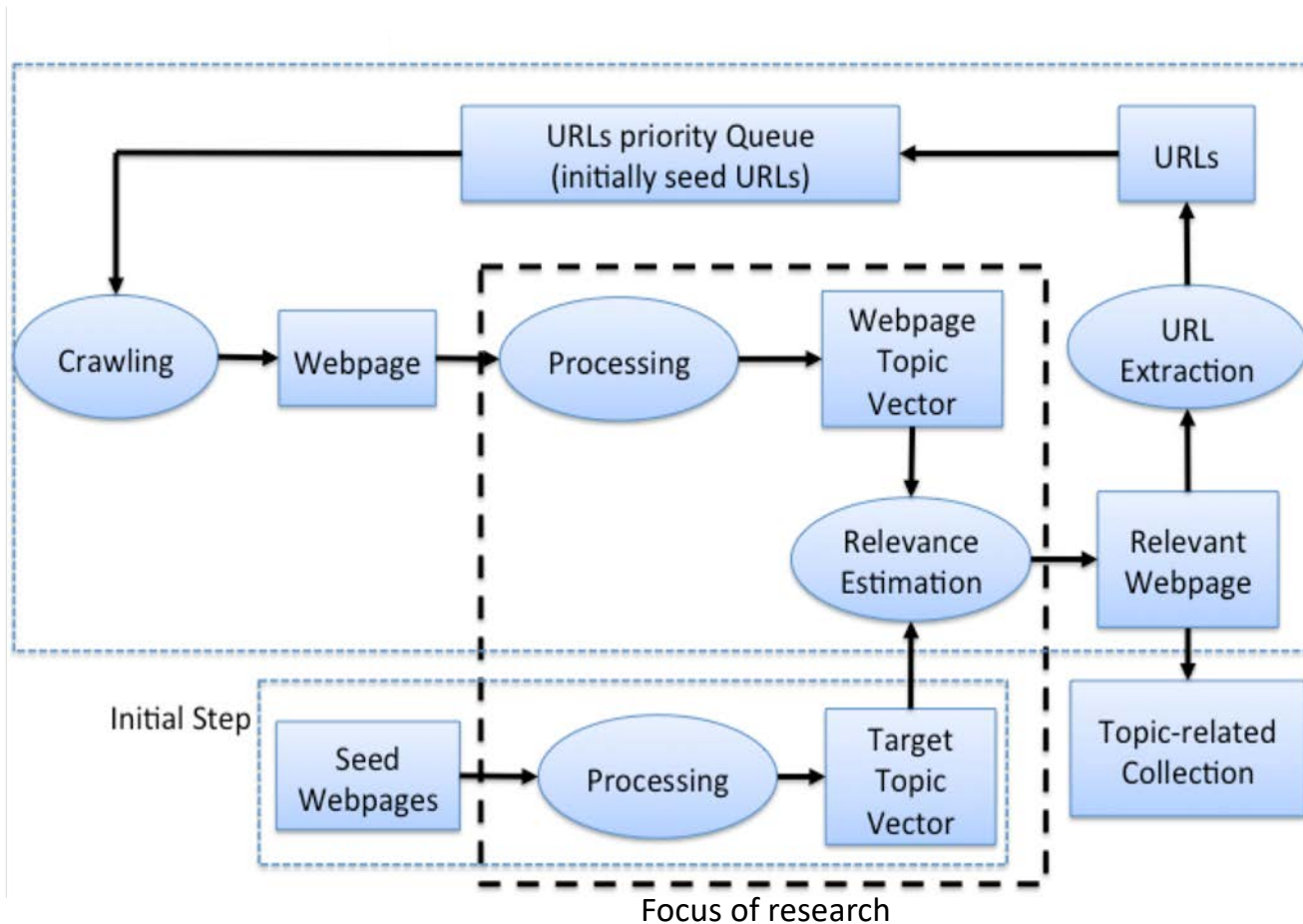
Representative Event Types

Event Types	Examples
Shooting	Oregon, California, Orlando
Earthquake	Ecuador, Japan
Flood	Texas Floods
Fire	California wild fire
Bombing	Boston bombing
Plane Crash	Egyptair, germanwings
Building Collapse	East Harlem
Protests/Riots	Egyptian revolution
Political Issue/Conflict	Brexit, Turkey coup, Greece Bailout referendum
Hurricane	Joaquin, Sandy, Katrina
Terror Attack	Paris, Brussels, Nice
Train Derailment	Amtrak188
Scandal	Panama Papers, Sepp Blatter
Community	Lovewins

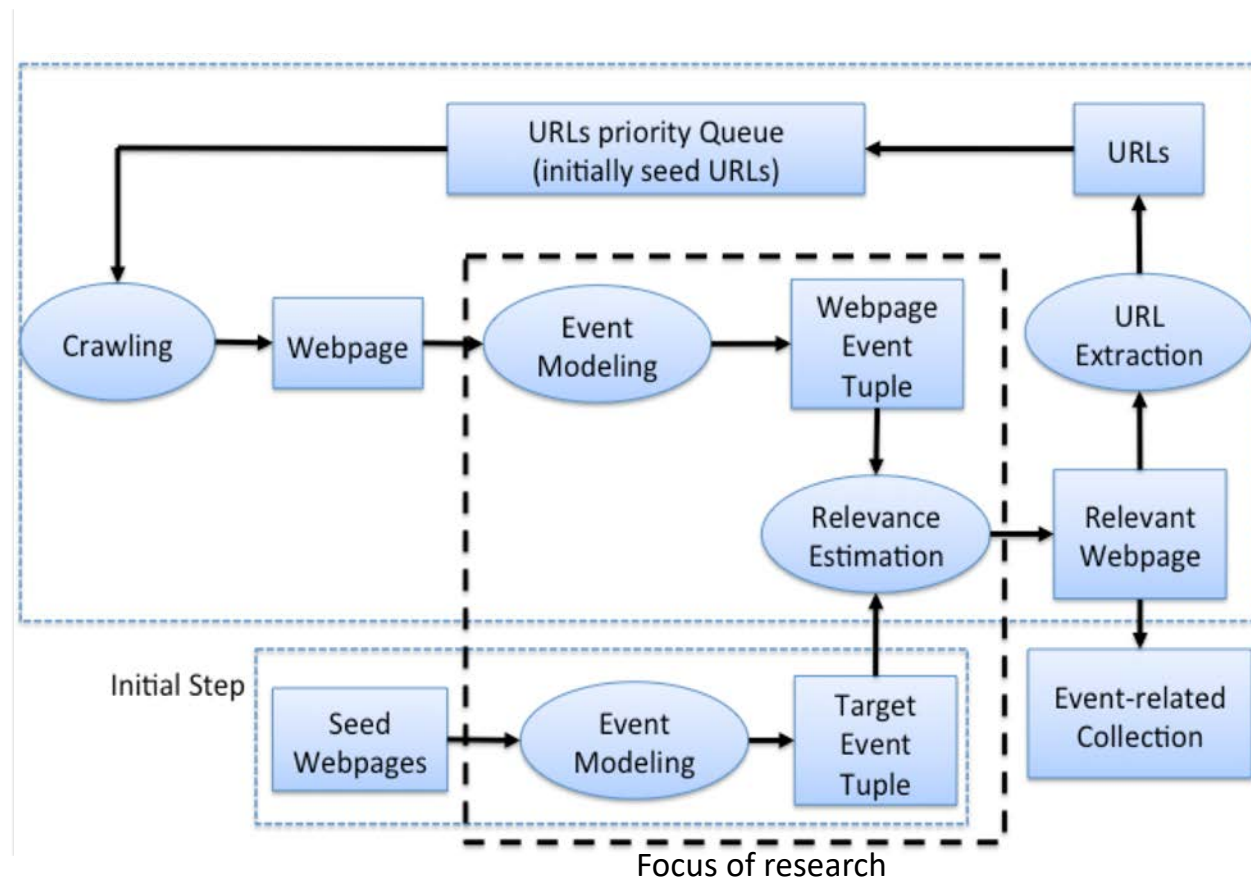
Data Flow



Baseline Focused Crawler



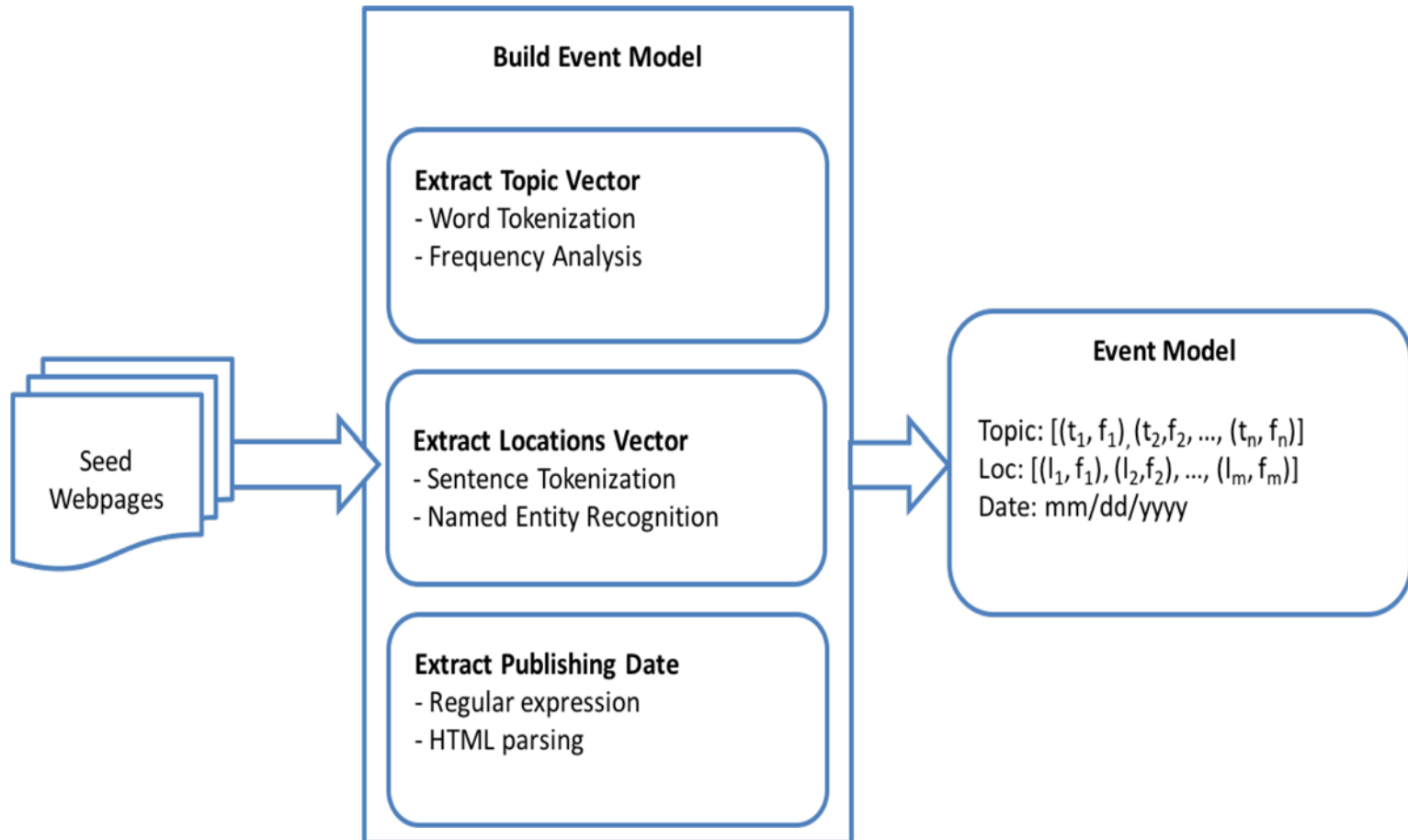
Event Focused Crawler



Event Modeling and Representation

- We define an event as
 - something (e.g., a disaster),
 - which happened in a certain place, and
 - at a certain time.
- Event E is a tuple $\langle T, L, D \rangle$.
- T = topic of event, L = location, D = date
- i.e.: what, where, and when.

Building Event Model



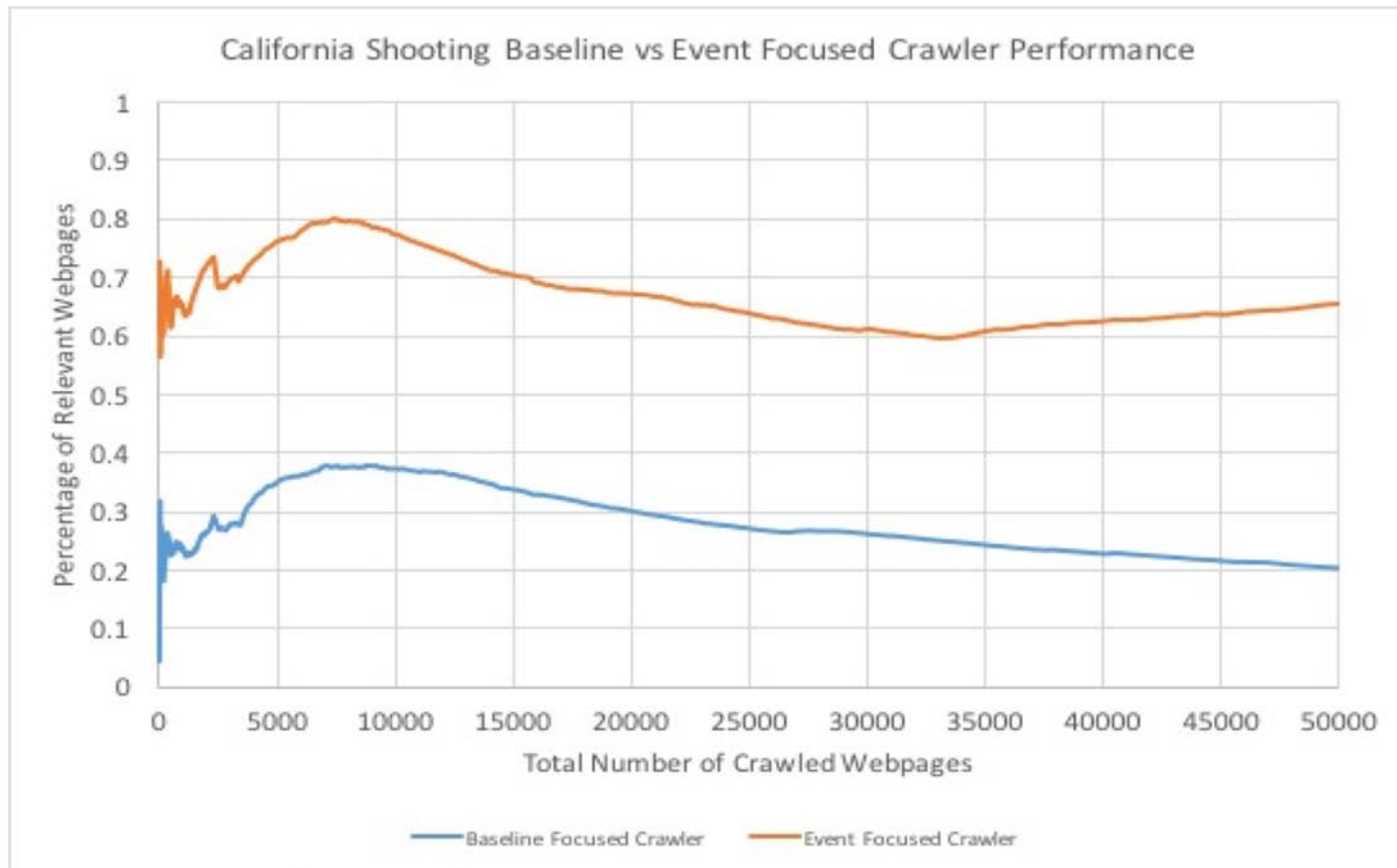
California Shooting Model

Topic	shoot	0.93
	san	0.513
	bernardino	0.465
	said	0.357
	wa	0.323
	2015	0.321
	peopl	0.31
	california	0.305
	polic	0.258
	suspect	0.177
Location	San Bernardino	1
	California	0.51
	Calif.	0.44
Date	2015-12-02	

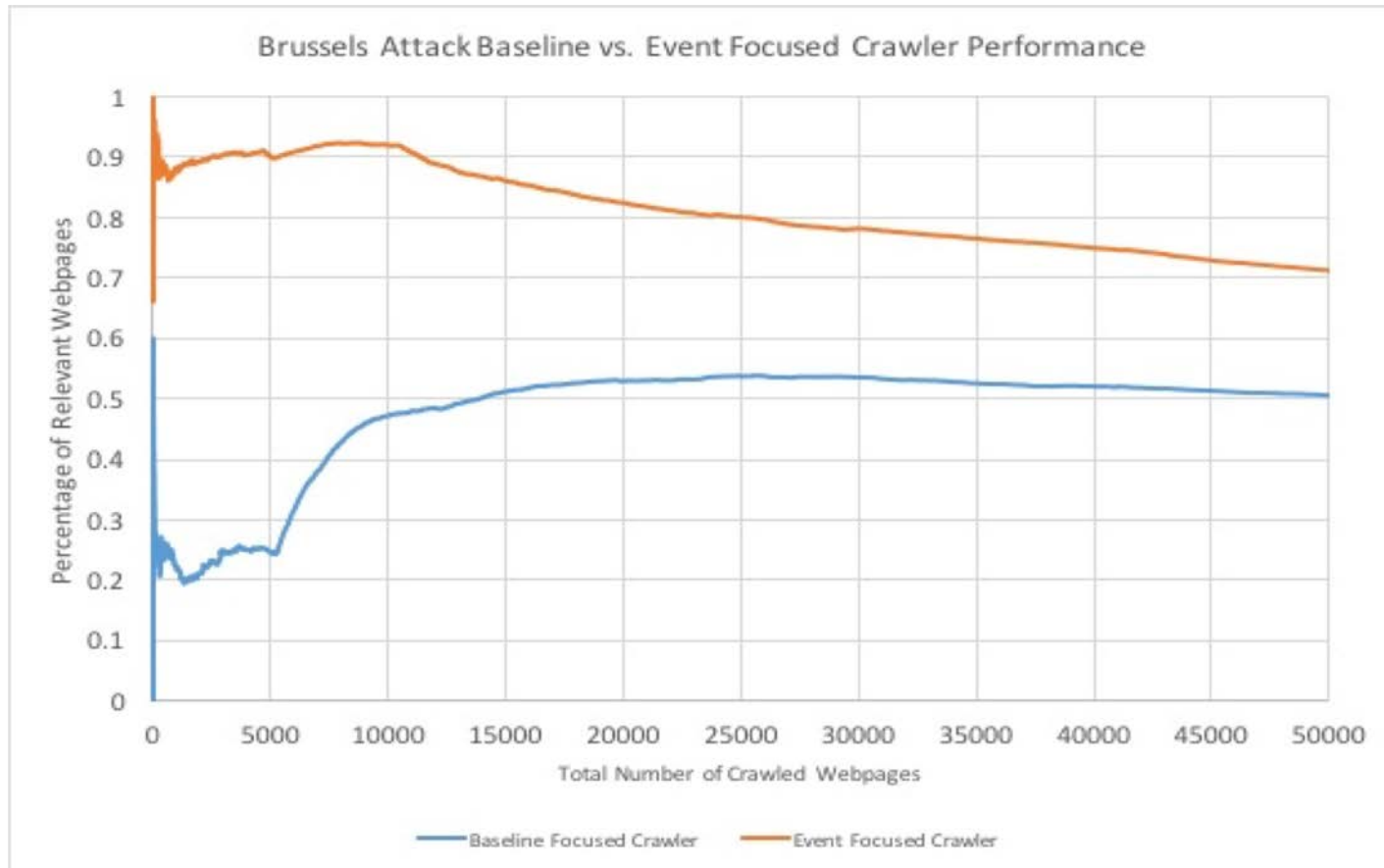
Datasets

Event	Type	Location	Date	# of Seed URLs	# of desired webpages
California Shooting	Shooting	San Bernardino, California, USA	12/2/2015	4,161	50,000
Brussels Attack	Terrorist Attack	Brussels, Belgium	3/22/2016	4,691	50,000
Oregon Shooting	Shooting	Roseburg, Oregon, USA	10/1/2015	22,354	100,000
Egyptair Plane Crash	Plane Crash	Mediterranean Sea, Alexandria, Egypt	5/19/2016	1,211	10,000
Panama Papers Leak	Document Leak	Panama	4/3/2016	18,260	100,000
Orlando Shooting	Shooting	Orlando, Florida, USA	6/12/2016	1,988	50,000
Paris Attack	Terrorist Attack	Paris, France	11/13/2015	88,835	500,000
Ecuador Earthquake	Earthquake	Ecuador	4/16/2016	11,348	100,000

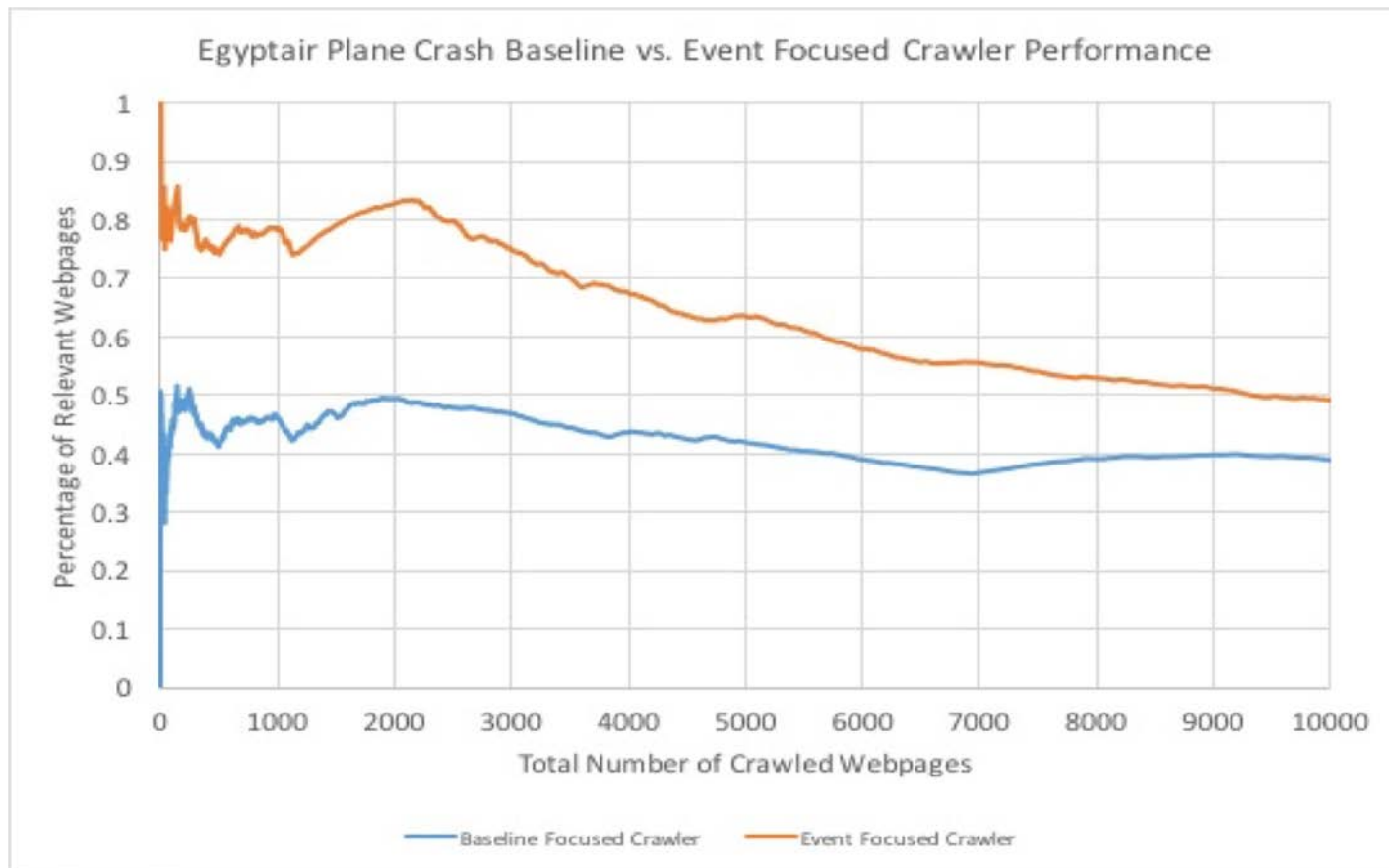
Large Scale: California Shooting



Large Scale: Brussels Attack



Large Scale: Egyptair Plane Crash



Twider: A Hybrid Model for Role-related User Classification on Twitter

Presenter: Liuqing Li

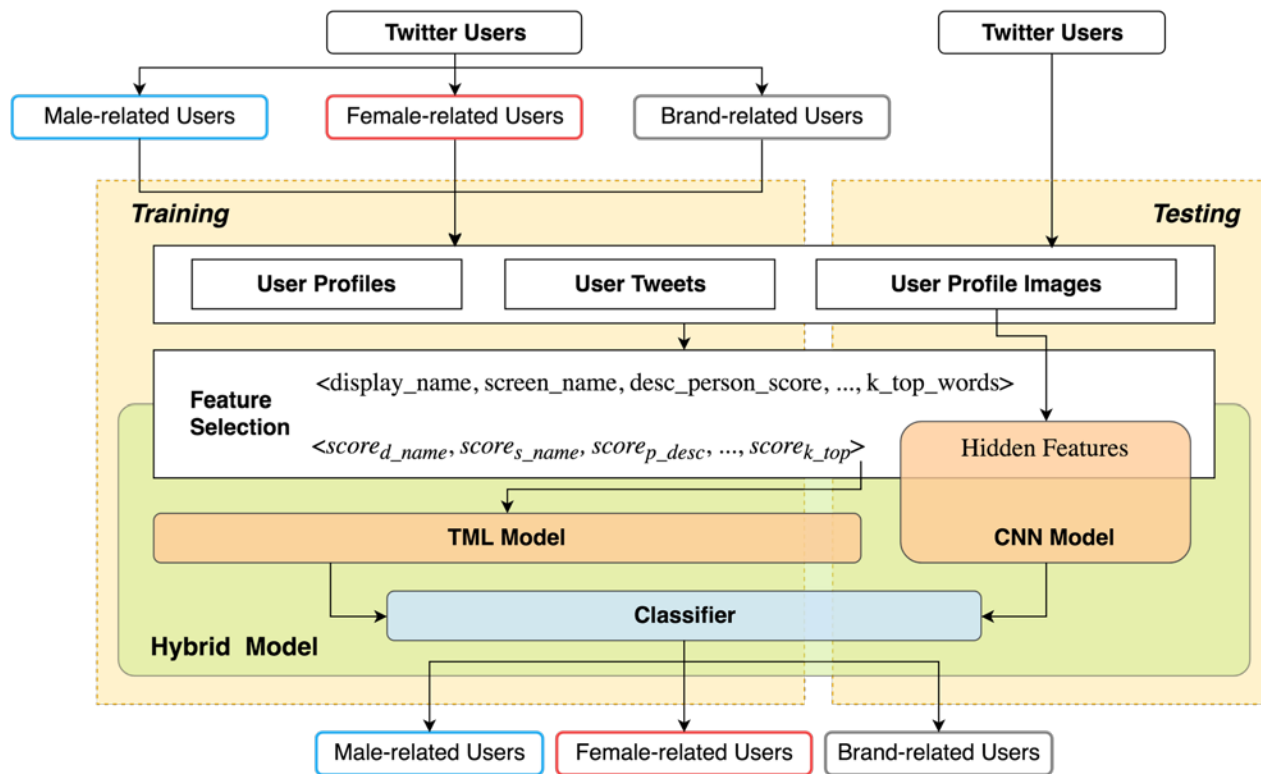
Digital Library Research Laboratory

Virginia Polytechnic Institute and State University

Blacksburg, VA, 24061

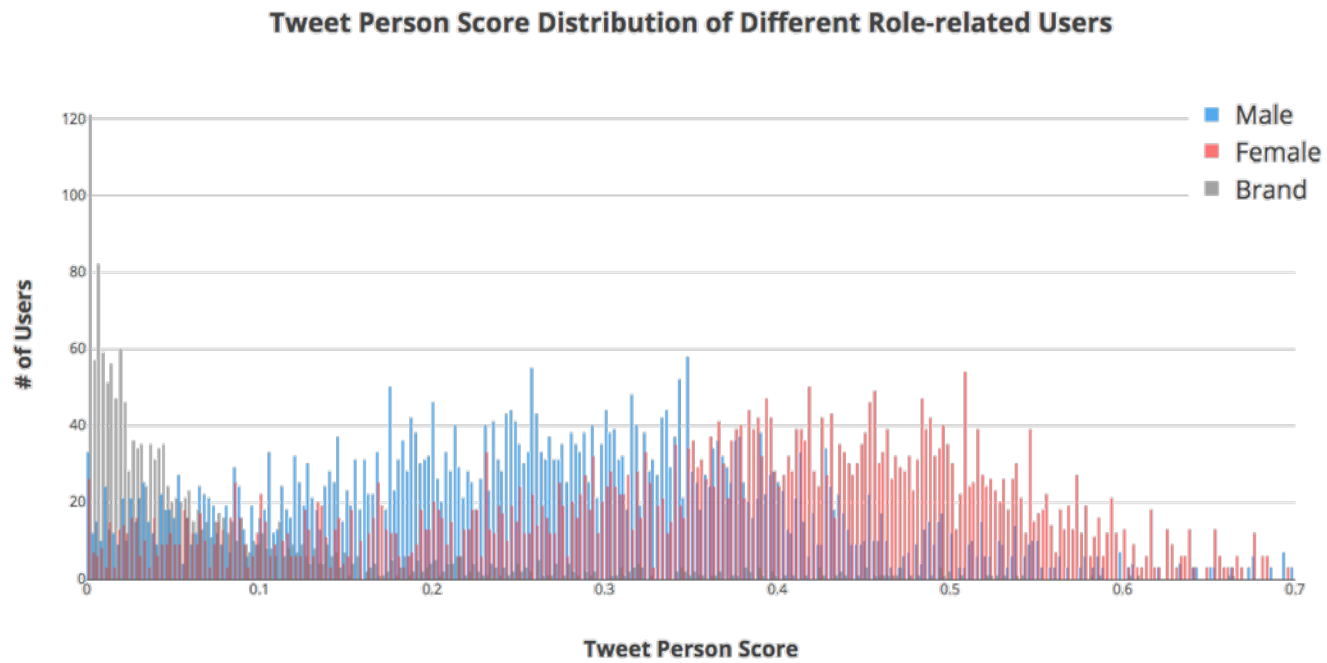
February 20, 2018

Framework



Discussion

- Features



Historical Tweet URL Analysis

Presenter: Liuqing Li

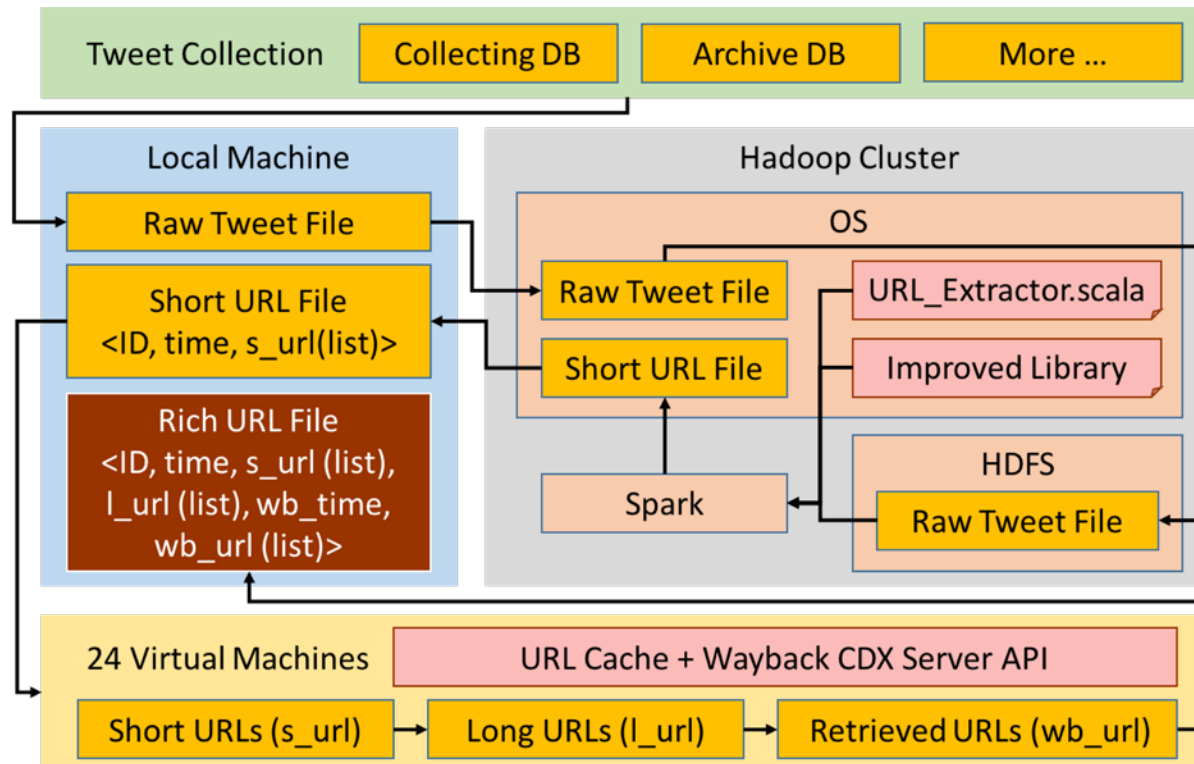
Digital Library Research Laboratory

Virginia Polytechnic Institute and State University

Blacksburg, VA, 24061

February 20, 2018

Framework

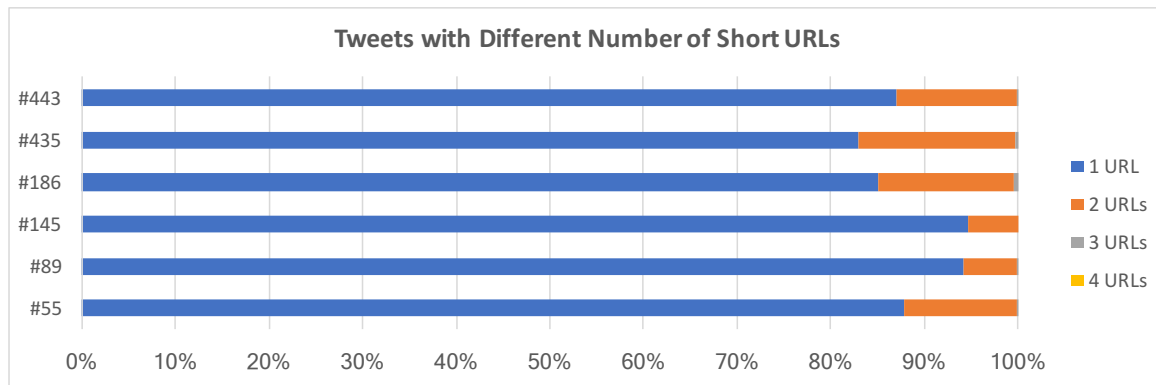


Preliminary Results

- Data Collections

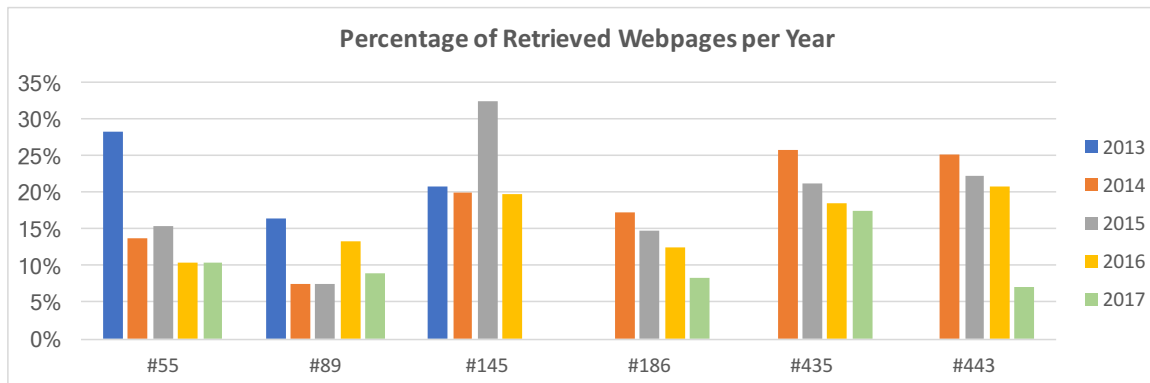
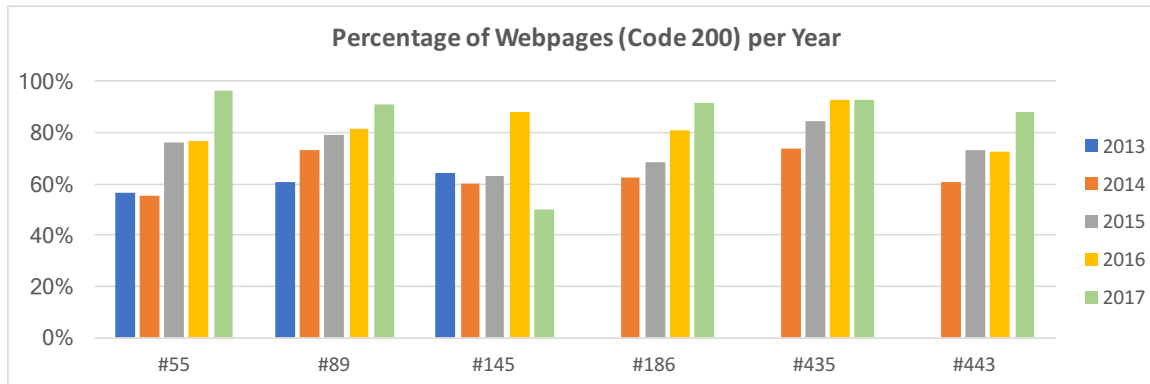
- #55 – Quantico shooting 2013/03/22
- #89 – santa monica shooting 2013/06/07
- #145 – nevada school shooting 2013/10/21
- #186 – shooting california 2014/05/25
- #435 – Ottawa Shooting 2014/10/22
- #443 – Marysville shooting 2014/10/24

Preliminary Results



ID	# of URLs	# of Unique URLs	Percentage
#55	36,704	1,377	3.8%
#89	32,089	3,092	9.6%
#145	21,497	2,602	12.1%
#186	665,895	62,096	9.3%
#435	186,600	15,686	8.4%
#443	59,554	6,338	10.6%

Preliminary Results



School Violence - GETAR

Presenter: Jason Callahan & Dr. Shoemaker

Digital Library Research Laboratory

Virginia Polytechnic Institute and State University

Blacksburg, VA, 24061

February 20, 2018

Themes to Evaluate/Refine

- Victim (gender, race, age, student vs non-student)
- Suspect (gender, race, age, student vs non-student)
- Type of weapon used (firearm, blade, etc.)
- Geographical location/region (population size)
- Suspect killed vs. suspect survives

Visualizations

- Themes of tweets/URL
- Locations of tweets (potential geotags)
- Media response tweets (news vs. non-news sources)
- Emotional response tones/themes (measured by volume/frequency)
- Clustering of terms/related incidents (hashtags of events/suspects/victims consolidated)
- Word clouds
- Pie/bar charts to illustrate the refined themes
- Time sequence tracking of refined themes
- Maps of twitter data if geotags are available

Technology on Trail Study

Presenter: Abigail Bartolome
Digital Library Research Laboratory
Virginia Polytechnic Institute and State University
Blacksburg, VA, 24061
February 20, 2018

Trail Study

“Each one is different; each has a soul”- Triple Crown veteran, Karen Berger, on which trail is her favorite.

	Appalachian Trail Topics	Pacific Crest Trail Topics	Continental Divide Trail Topics
Topic 1	#indigenous, #tairp, #amerianindian8, day, knob, mcafee, trailva	California, #pct2017, 2, story, tips, resupply, #pics	help, #bravethecdt, #hikecdt, today, @cdnst1, vote, great
Topic 2	va, catawba, sunrise, halfway, #backpacking, just, oc4444x2400	@pctassociation, like, today, #lwcf, win, did,great	#bravethecdt, @cdnst1, #hikecdt, help, great, support, today
Topic 3	days, amp, long, mountain, complete, miles, week	mount, adams, goat, rocks, @hogansog, washington, view	#bravethecdt, help, today, @cdnst1, support, vote, great
Topic 4	#travel, #bestseller, black, 1, awol, books, 2	#orshow, booth, gear, come, free, #pct2017, @danner	#bravethecdt, help, today, great, @cdnst1, support, #hikecdt
Topic 5	new, going, woman, 80yearold, solo, sisters, twin	wild, lost, #travel, #bestseller, oprahs, #7, #8	#bravethecdt, #hikecdt, @cdnst, support, help, great, today
Topic 6	hiker, #at2017, @thetrek, 5, update, thruhiker, thruhikers	taking, #backpacking, months, job, better, 4, day	#hikecdt, #bravethecdt, @cdnst1, great, help, support, today
Topic 7	#at2017, @thetrek, #trail, gear, list, things, #photography	#pctdays, new, instagram, year, weeks, posts, bitesized	help, #bravethecdt, today, #votecdt, vote, 25k, @cdnst1

From January-May, topic analysis reflected Appalachian Trail valued experiences, while Pacific Crest Trail focused more on the logistics of planning a hike. Are these part of hiking culture? Was this influenced by geographical (and schedule) differences?

Trail Cultures:

- Avid Hiking
- Conservation practices

What can we learn about these cultures? What can we learn about their language?

Surprising Trail Countercultures:

- Nude Hiking
- Actively denying conservation practices

Why do these hikers behave in this way? What are their motivations? Who is attracted to these countercultures and can they be infiltrated?

GETAR Collection + GeoBlacklight

Presenter: Ziqian Song

Digital Library Research Laboratory

Virginia Polytechnic Institute and State University

Blacksburg, VA, 24061

February 20, 2018

Homepage

- <http://mule.dlib.vt.edu:3033/>

The screenshot shows the GETAR homepage with a dark blue background. At the top, there is a navigation bar with various icons and links. The main header features the GETAR logo and the tagline 'Global Event and Trend Archive Research'. Below this, a large section titled 'Explore and discover...' contains a search bar and the text 'Web archive of urgent global challenge events and initiatives'. The main content area is divided into three columns: 'Collection', 'Subject', and 'Hashtags', each with a 'More' link. Below these columns is a 'Find by location' section featuring a world map with a search bar. The footer contains copyright information and links to 'About GETAR', 'DLRL', and 'EOLIC 7/27/2019'.

GETAR
Global Event and Trend Archive Research

Explore and discover...
Web archive of urgent global challenge events and initiatives

Search

Collection	More
diamondring	1303615
wattheclipse	693565
nonstop	300781
exploratoraudi	195255
totalclipse	95501
safeclipse	79003
ipmaphanddownsation	742
globalnewsmapdata	720
foralmapdata	663

Subject	More
experience	684012
safety	471589
see the power of music	343257
photos pictures	303332
forecast	287447
satellite experience	269366
spanish	179841
eclipse	128882

Hashtags	More
science2017	2398125
science2017	1315388
eclipse	1133341
science	554965
totalclipse	452746
thepowerofmusic	383174
see	299092
weathering	142336
mondymotivatothitjett.com@tspi	71743

Find by location

Search here

© 2017 Digital Library Research Laboratory at Virginia Tech | About GETAR | DLRL | EOLIC 7/27/2019

TweetBank

Presenter: Shou Niu

Digital Library Research Laboratory

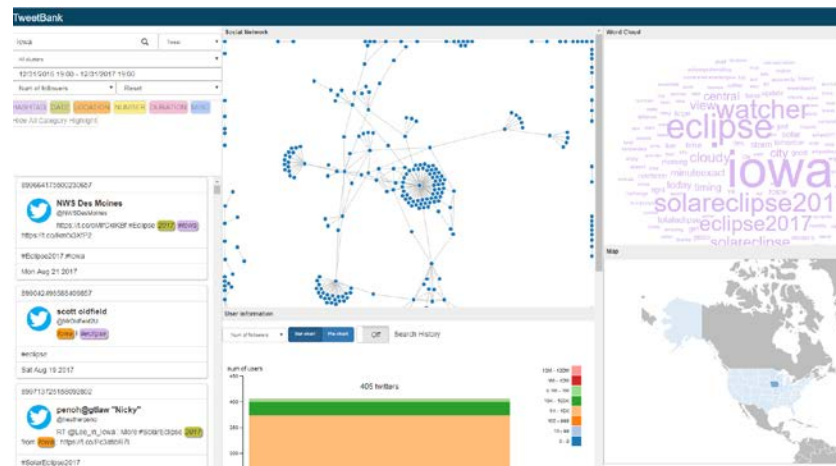
Virginia Polytechnic Institute and State University

Blacksburg, VA, 24061

February 20, 2018

TweetBank

- A web portal to explore GETAR Twitter collection.
- Developed in Fall17
- Functions:
 - Searching
 - Tweet viewing
 - Social network
 - User information
 - Time-line
 - Keywords
 - Geo-locations



http://mule.dlib.vt.edu/cs5604f17_fe/TweetBank/src/

Planned Activities – Welcoming Involvement

- Collaboration with Internet Archive to aid research community
- Aid some 30 local stakeholders
- Variety of interfaces across information life cycle
- Collect, Add Value, Archive, Analyze, Search/Browse, Visualize

- Displays outside 2030 Torgersen Hall (DLRL)

- Many volunteers: CS4624, CS5604, CS6604, Theses, Independent Studies, and others at all levels

Summary

- Context
- GETAR proposal
- IDEAL results – Sunshin Lee
- IDEAL results – Mohamed Farag
- Selected GETAR projects
- Welcoming collaboration