CS6604 Digital Libraries
Global Events Team Final Presentation

Presenters:

Liuqing Li, Islam Harb, Andrej Galad

{liuqing, iharb, agalad}@vt.edu

Instructor:

Dr. Edward A. Fox

Virginia Polytechnic Institute and State University

Blacksburg, VA, 24061

April 27, 2017

VirginiaTech

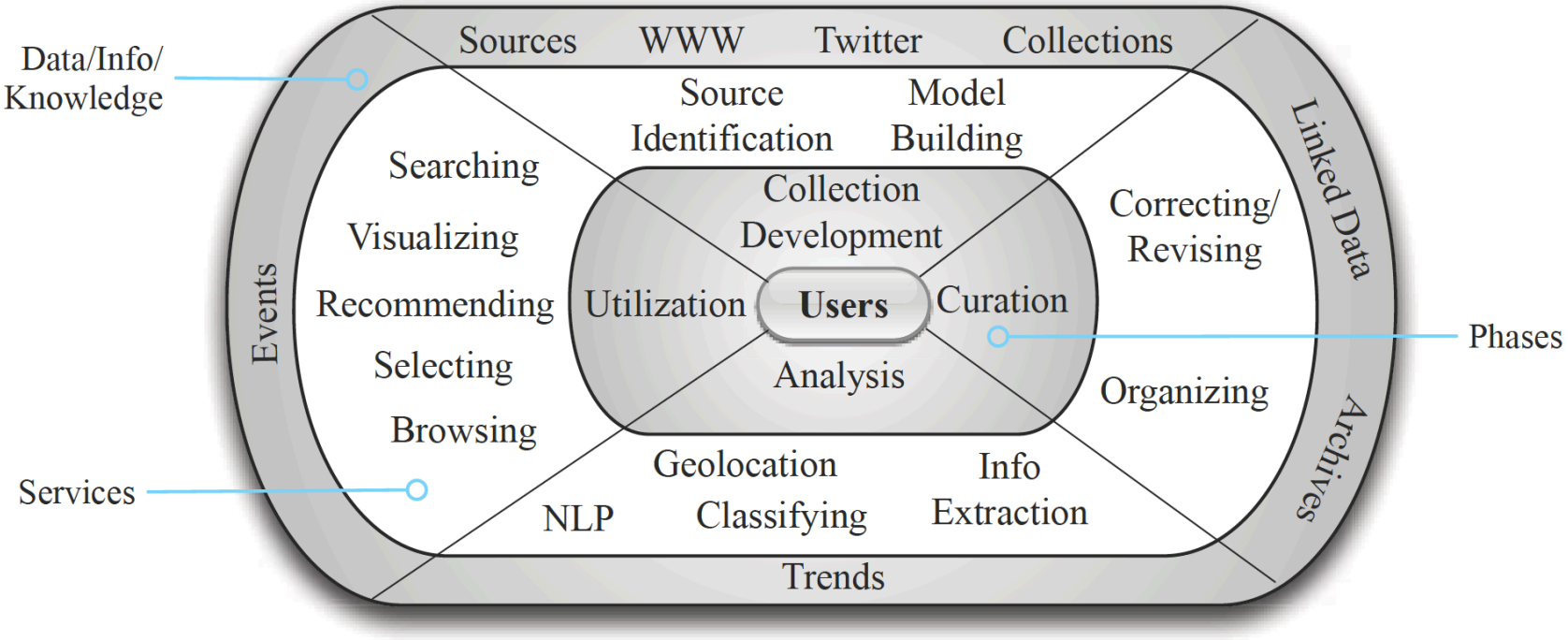
Invent the Future

Outline

- Background
- Implementation
  - Data Collection
  - Data Processing
  - Data Visualization
- Future Work
- Acknowledgement

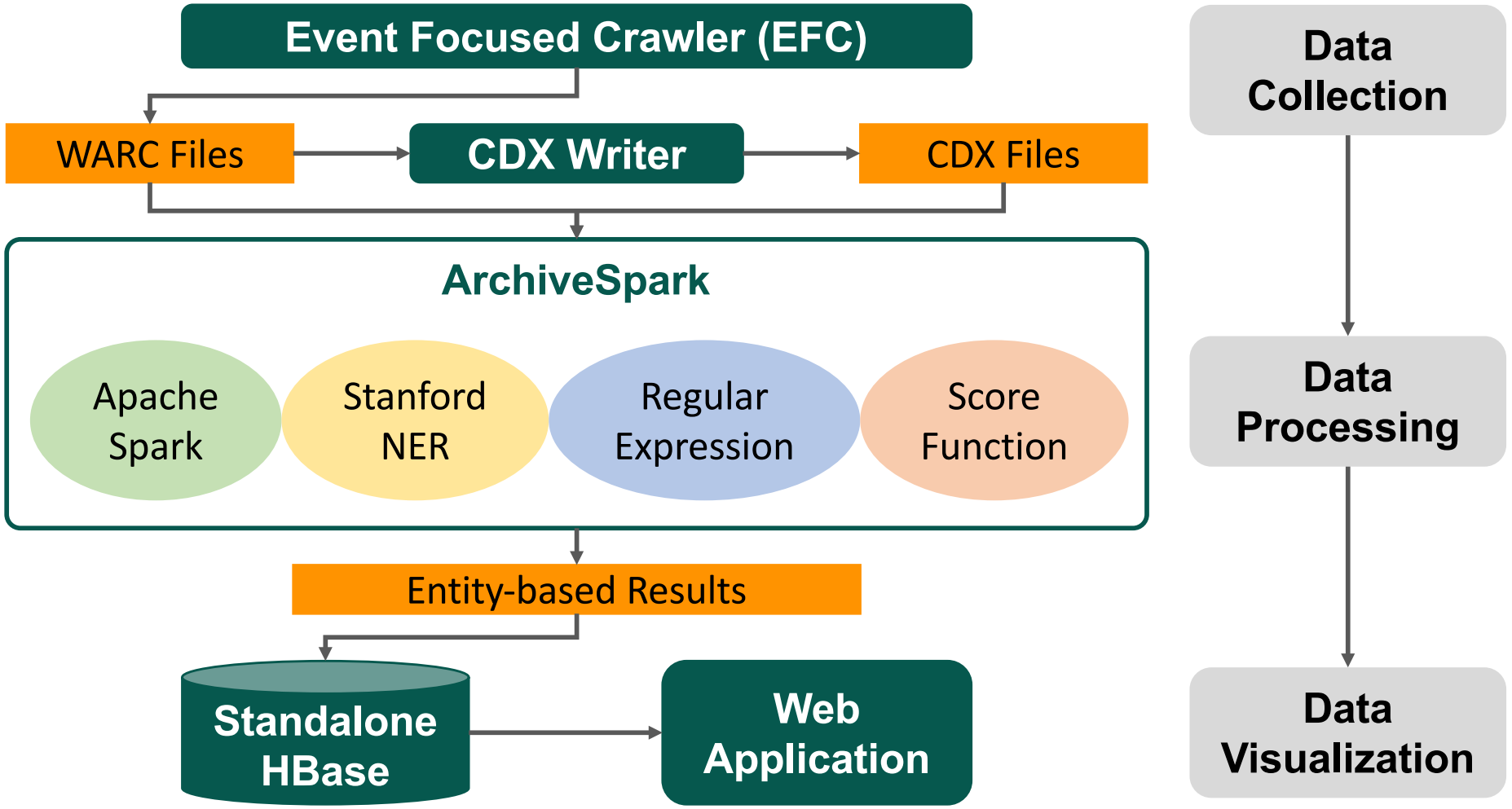
Background

- GETAR*
  - Global Event and Trend Archive Research
  - Architecture



* Edward A Fox, Donald Shoemaker, Chandan Reddy, Andrea Kavanaugh, III: Small: Collaborative Research: Global Event and Trend Archive Research (GETAR), NSF grant IIS - 1619028, 2017-2019. http://eventsarchive.org

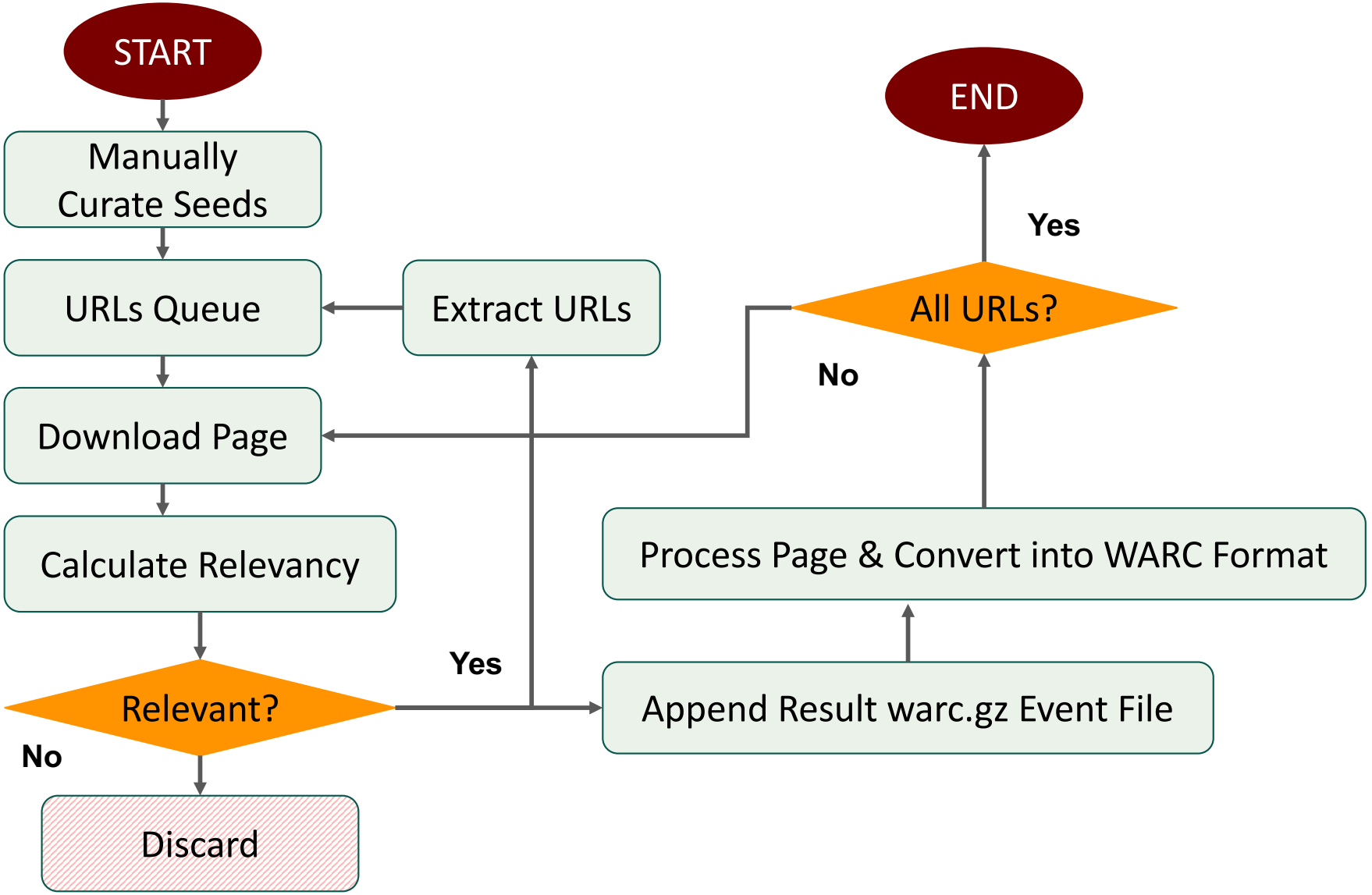
Implementation – Architecture



Events of Interest

School Shooting Events	Year
Virginia Tech Shooting	2007
Northern Illinois University Shooting	2008
Dunbar High School Shooting	2009
University of Alabama Shooting	2010
Worthing High School Shooting	2011
Sandy Hook Elementary School Shooting	2012
Sparks Middle School Shooting	2013
Reynolds High School Shooting	2014
Umpqua Community College Shooting	2015
Townville Elementary School Shooting	2016

Focused Crawler – Collecting / Archiving



WARC Libraries

- Wget (Version 1.14 or later)

```
wget \
  --mirror \
  --warc-file=YOUR_FILENAME \
  --warc-cdx \
  --page-requisites \
  --html-extension \
  --convert-links \
  --execute robots=off \
  --directory-prefix=. \
  --span-hosts \
  --domains=example.com,www.example.com,cdn.example.com \
  --user-agent=Mozilla (mailto:archiver@petekeen.net)\
  --wait=10 \
  --random-wait \
```

WARC Libraries

- Wpull

Note

When resuming downloads with `--warc-file` and `--database`, Wpull will overwrite the WARC file by default. This occurs because Wpull simply maintains a list of URLs that are fetched and not fetched. You should either rename the existing file manually or use the additional option `--warc-append` or move the files `--warc-move`.

- `--warc-append`
- `--warc-move`: Move WARC files out of the way for resuming a crashed crawl.

WARC Libraries

- WARCIO: WARC (and ARC) Streaming Library
  - Python 2.7+ and 3.3+
  - Post-Processing: Read / Write WARC format

```python
from warcio.warcwriter import WARCWriter
from warcio.statusandheaders import StatusAndHeaders

import requests

with open('example.warc.gz', 'wb') as output:
    writer = WARCWriter(output, gzip=True)

    resp = requests.get('http://example.com/',
                        headers={'Accept-Encoding': 'identity'},
                        stream=True)

    # get raw headers from urllib3
    headers_list = resp.raw.headers.items()

    http_headers = StatusAndHeaders('200 OK', headers_list, protocol='HTTP/1.0')

    record = writer.create_warc_record('http://example.com/', 'response',
                                       payload=resp.raw,
                                       http_headers=http_headers)

    writer.write_record(record)
```

Ten Events Collections

- Naming Convention
  - [location]_[year].warc.gz

```
islam 788109490 Apr 18 13:45 townville_2016.warc.gz
islam 758426598 Apr 18 13:43 umpqua_2015.warc.gz
islam 696778729 Apr 18 13:38 omaha_NE_2011.warc.gz
islam 661944603 Apr 18 13:40 Sandy_2012.warc.gz
islam 454081122 Apr 18 13:29 VT_2007.warc.gz
islam 279509497 Apr 18 13:41 Sparks_2013.warc.gz
islam  15635547 Apr 18 13:36 dunbar_chicago_2009.warc.gz
islam  14708587 Apr 18 13:41 reynolds_2014.warc.gz
islam   5705504 Apr 18 13:36 huntsville_2010.warc.gz
islam    531513 Apr 18 13:35 NIU_2008.warc.gz
```

Tools for Data Processing

- ArchiveSpark
  - Apache Spark framework for Web Archives
  - Easy data extraction
  - Input: WARC and CDX files

- CDX Writer
  - Python script to create CDX files of WARC files
  - Format: CDX N b a m s k r M S V g
    - e.g., edu,vt,cnre)/ 20170422005601 http://cnre.vt.edu text/html 200 BT3ILJXROIILHBKQPNYDUCUVZRDKG3OA - - 9478 20104749 data/Virginia-Tech-Shooting_20070416.warc.gz

Data Preprocessing

- Webpage Cleaning
  - Extract Raw Text
    - payload.string.html.body.text
  - Remove jQuery & JavaScript
    - { WPGroHo.syncProfileData(hash, id); }, …
  - Remove tags
    -
, <p>, …
  - Remove markers
    - *, |, +, …
  - Remove stopwords
    - a, about, the, …

Data Processing

- Entity Extraction
  - Basic Parsing
    - event name and date
  - Stanford NER (Integrated model)
    - entities, shooter name
  - Regular Expression
    - event date
    - shooter name and age
    - number of victims
    - weapon list
- Score Function
  - $tf * df$

HBase

- Build-in ImportTsv Utility
  - Import Data into HBase

Table Name	globalevents	
Row_Key	Event_Date + Event Hash Value	20070416217787922
Column Family	event	
Column	event: name	Virginia Tech Shooting
	event: date	20070416
	event: shooter_age	23-year-old
	event: shooting_victims	32 victims
	event: entities	Virginia;Tech;VA;University;…
	event: entities_count	146900;62415;13940;7732;…
	event: entities_url	url1,url2,url3,url4,url6;url2,url3,url4,url5;url1,url3,url4,url5,url6;…

Data Processing – Demo

- Key Stages
  - Initialization
    - Create Spark Session
    - Create NLP Core
    - Create Storage
  - Processing
    - Extract Event Name/Date/URL
    - Extract Name Entities
    - Extract Other Event Features
  - Export and Import
    - Generate TSV file
    - Import TSV file into HBase

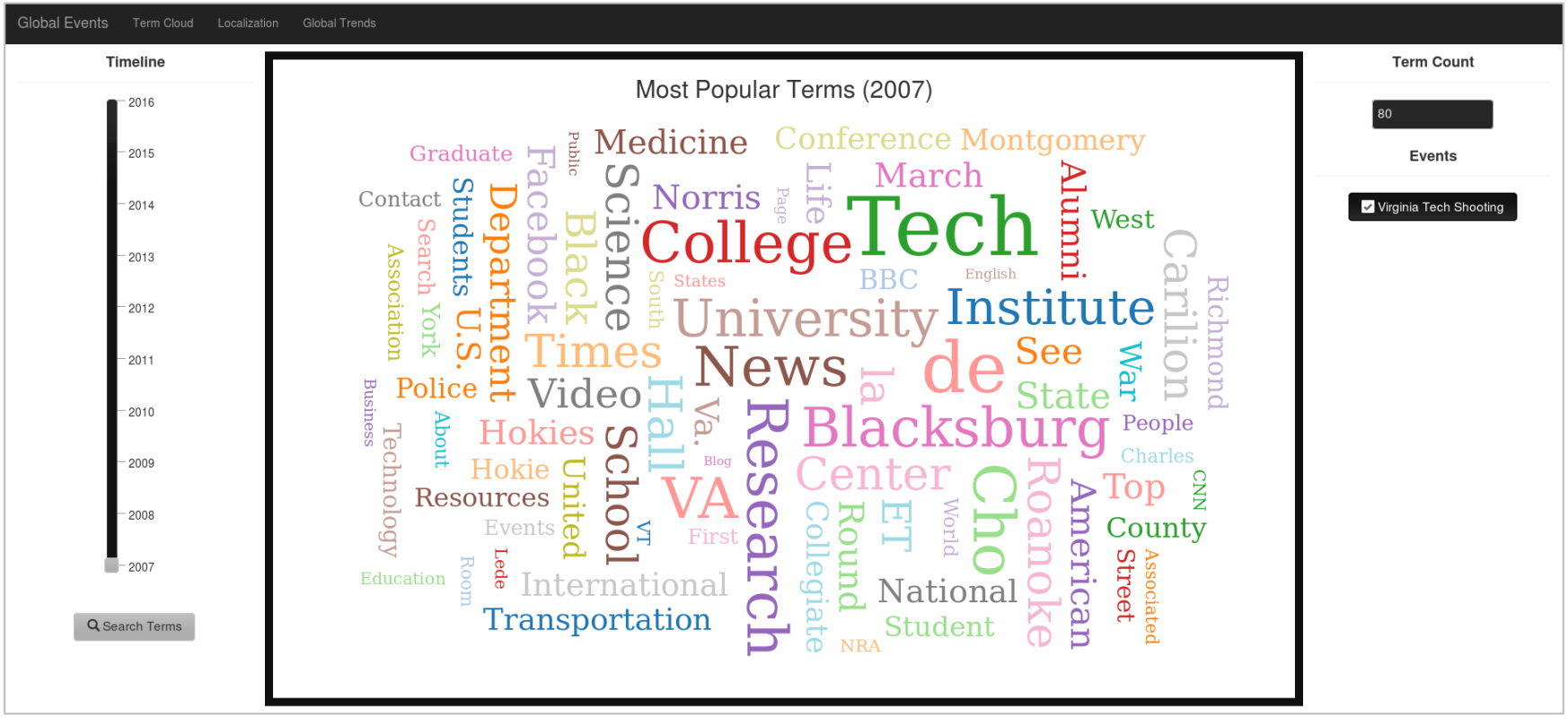
Global Events Viewer

- Efficient visualization of long-term global events
  - Show representative terms -> link to corresponding URLs
  - Visualize events' trends over time (time series)
- Java 7 Spring Boot Web application
  - Build system - Gradle
  - Embedded Tomcat Web server
  - Backend - HBase, in-memory
  - Frontend - D3.js, Bootstrap

https://github.com/dedocibula/global-events-viewer

Global Events Viewer – Demo

- Key Components
  - WordCloud, Range Selection, URL List, Trends



Problem Faced

Data Collection

Encoding problems (UTF-8, ASCII and others)

Get more relevant seeds for old events

Data Processing

Lack of documentation (ArchiveSpark)

Version conflict (CDX Writer, Kernel in Jupyter)

JVM issue (Spark)

Data Visualization

Spring boot IntelliJ setup

JQuery UI

Lessons Learned

Data Collection
WARCIO
Focused Crawler

Data Processing
ArchiveSpark
Spark & Scala (Map/Reduce Process)

Data Visualization
D3 WordCloud
D3 Dynamic Line Charts

Future Work

Data Collection

- Wayback Machine
- Automatic Routine for Focused Crawler
- Event Extension (Sources, Time, Space)

Data Processing

- Standalone Mode -> Cluster Mode
- Name Entity Recognizer
- Automatic Processing (CDX Writer and HBase)

Data Visualization

- Localization – Datamaps
- Weapons

Acknowledgement

Projects

NSF IIS - 1319578	III: Small: Integrated Digital Event Archiving and Library (IDEAL)
NSF IIS - 1619028	III: Small: Collaborative Research: Global Event and Trend Archive Research (GETAR)

Organizations

Internet Archive
L3S Research Center

Persons

Instructor	Dr. Edward A. Fox
Alumnus	Dr. Mohamed Magdy Farag
Labmates	Prashant Chandrasekar, Xuan Zhang

Thank you !

Questions?