

# Teaching Big Data analysis in the Social Sciences using a HUBzero-based platform

Jeanette Sperhac, SUNY Buffalo, Jim Greenberg, SUNY  
Oneonta

HUBbub, Indianapolis, IN  
29 September 2014

CENTER FOR COMPUTATIONAL RESEARCH

---



**SUNY  
ONEONTA**



# Outline

1. The Collaboration
2. Goals
3. Hardware, Deployed Tools, and Datasets
4. Accomplishments
5. Challenges
6. Plans

CENTER FOR COMPUTATIONAL RESEARCH

---

# The Collaboration



- Academic supercomputing facility
- Buffalo, New York
- 8000+ cores, 100+ Tflops compute capacity
- Four-year liberal arts college
- Oneonta, New York
- Enrollment 5,900 students

CENTER FOR COMPUTATIONAL RESEARCH



**SUNY  
ONEONTA**

# Adopting Social Media Analysis at Oneonta

1. Social Sciences sought a data analysis environment
2. Oneonta could not offer the needed resources on campus
3. UB CCR and Oneonta collaborated to stand up a HUBzero instance

CENTER FOR COMPUTATIONAL RESEARCH

---



**SUNY  
ONEONTA**

# Social Sciences Course Goals

- Encouraging critical thinking
- Deploying ideas from texts in new directions
- Applying theoretical perspectives and concepts

***Achieving student engagement through data-driven research***

CENTER FOR COMPUTATIONAL RESEARCH

---



**SUNY  
ONEONTA**

# Case Study: Society and Animals

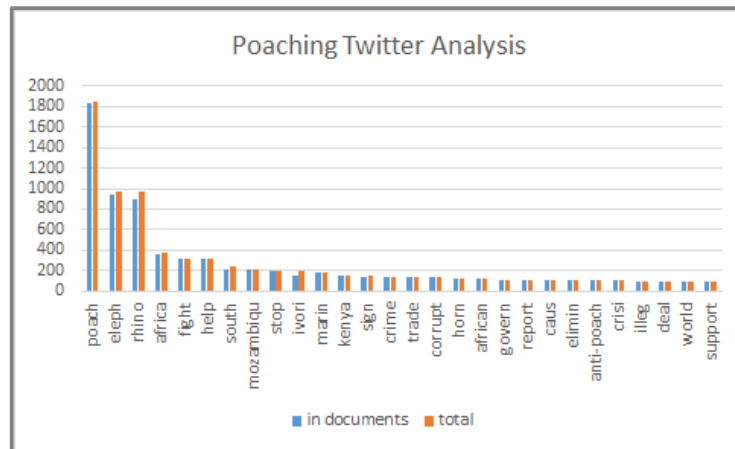
- Sociology course, 200 level
- Social science majors without programming experience
- Comparative/historical, social scientific analyses

***Gather, organize, and interpret social media data***

CENTER FOR COMPUTATIONAL RESEARCH

---

# Case Study: Society and Animals



A data set of 1831 tweets was gathered from Trackur.com, using the terms Poaching, Elephant, Rhino, Lion, Tiger, India, China, and Africa.

Search terms were chosen based on literature review of journal articles pertaining to poaching. A wordlist of commonly occurring terms was generated using RapidMiner V 5.3 software.

The keyword poach occurred 1857 times. Top *co-occurring terms*:

- Elephant (946 tweets/976 total) and Rhino (896/976)
- Africa (363/376) and Fight (313/317)
- Help (311/312) and Stop (192/198)

# Collaboration Goals

- Create a social sciences big data discovery environment
- Support social science teaching *and* research
- Leverage High Performance Computing (HPC) resources

CENTER FOR COMPUTATIONAL RESEARCH

---



**SUNY  
ONEONTA**



vidia.ccr.buffalo.edu

*Virtual Infrastructure  
for Data Intensive Analysis*

# VIDIA

Using the HUBzero platform:

- Provide workflow tools for data analysis
- Curate large datasets of social scientific interest
- Enable access to HPC resources

***Deployed October 2013***





# Why HUBzero?

- Provide platform for tool and HPC access
- Easy on campus IT staff
- Access anytime, anywhere
- Resources can be selectively secured
- Students may access resources after course conclusion

# VIDIA Hardware

## **HUBzero and webserver:** Dell PowerEdge R720xd

- 2x 6-core Intel Xeon E5-2630 (2.30 GHz, 15M cache)
- 48 TB raw (~36 TB usable) SATA disk space
- 128 GB memory (16x8GB - 1333MHz DIMMS)

## **Analysis:** 4x Dell PowerEdge R520

- 6-core Intel Xeon E5-2430 (2.20 GHz, 15M cache)
- 4.8 TB raw (~4 TB usable) SAS disk space
- 96 GB memory (6x16GB - 1600MHz DIMMS)

CENTER FOR COMPUTATIONAL RESEARCH

---

# VIDIA HUBzero Instance

- Open-source HUBzero v1.2.2
- Debian 7
- Housed in CCR machine room
- Workspace access restricted through Group
- Container template and sessions on separate disk from root

CENTER FOR COMPUTATIONAL RESEARCH

---

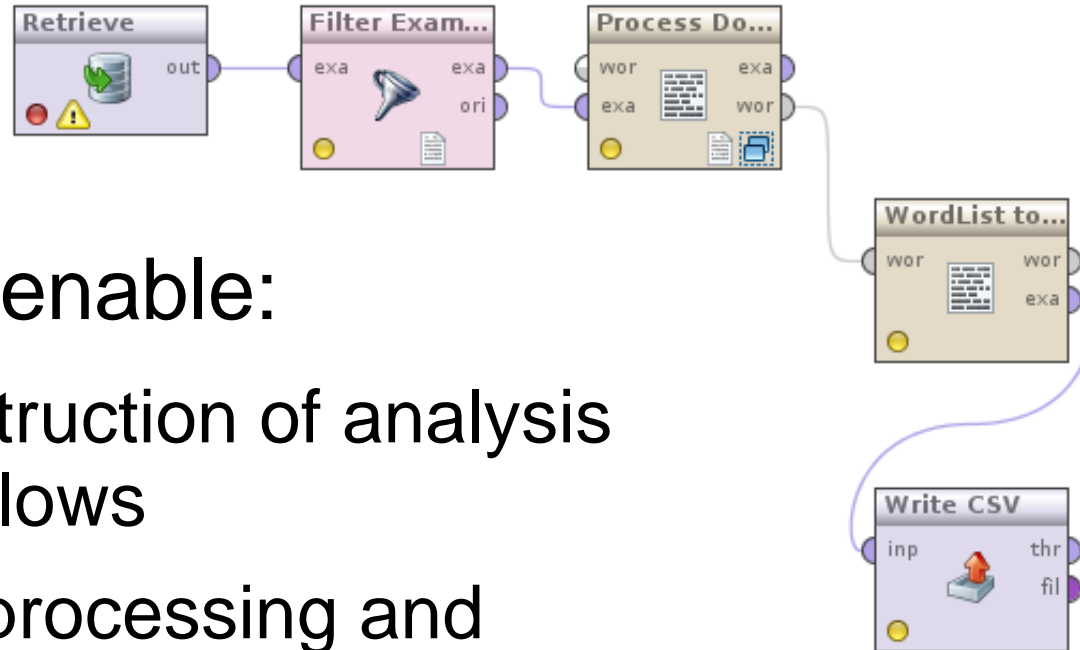


# Data Mining Workflow Tools

- Graphical User Interface
- Powerful, easy to use
- Open source, extensible



# Why RapidMiner?



- Operators enable:
  - Construction of analysis workflows
  - Text processing and analysis
  - Plotting and visualization
- Coding not required
- Data import/export

hub

Local Repository/processes/oneonta-degrees-v2.xml RapidMiner 5.3.013 @ localhost

File Edit Process Tools View Help

Process XML

Write input CSV

Process

Retrieve de... Duplicate l... Filter Data Select Data Write Outp...

Operators

- Process Control (37)
- Utility (54)
- Repository Access (6)
- Import (27)
- Export (18)
- Data Transformation (1)
- Modeling (118)
- Evaluation (29)
- Text Processing (48)
- Reporting (6)

Repositories

- Samples (none)
- DB
- Local Repository (jmsper)

Context Parameters

Filter Data (Filter Examples)

condition class attribute\_v...

parameter st... A. Degree=| & &

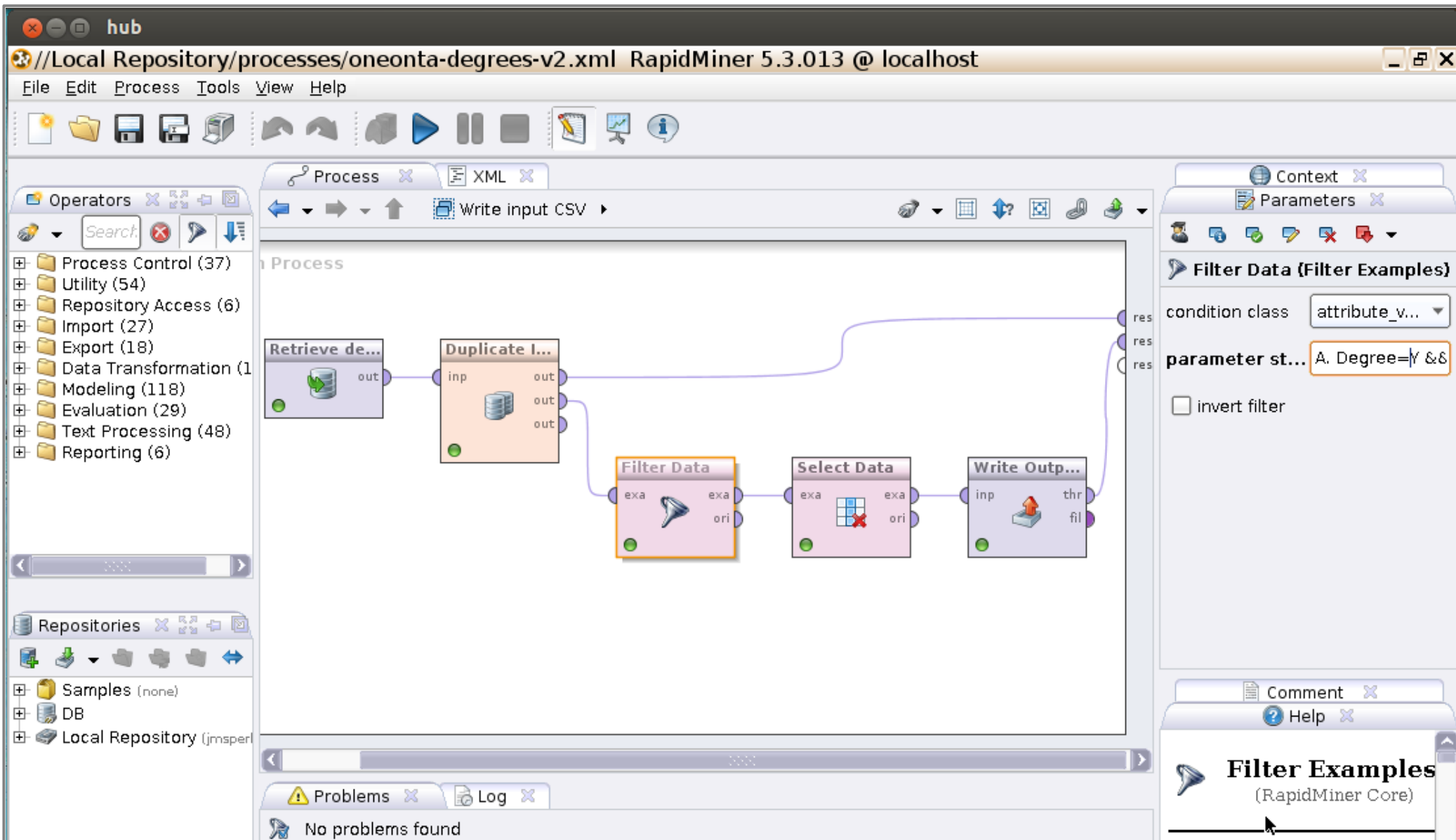
invert filter

Comment Help

Filter Examples (RapidMiner Core)

Problems Log

No problems found





hub

Local Repository/processes/oneonta-degrees-v2.xl

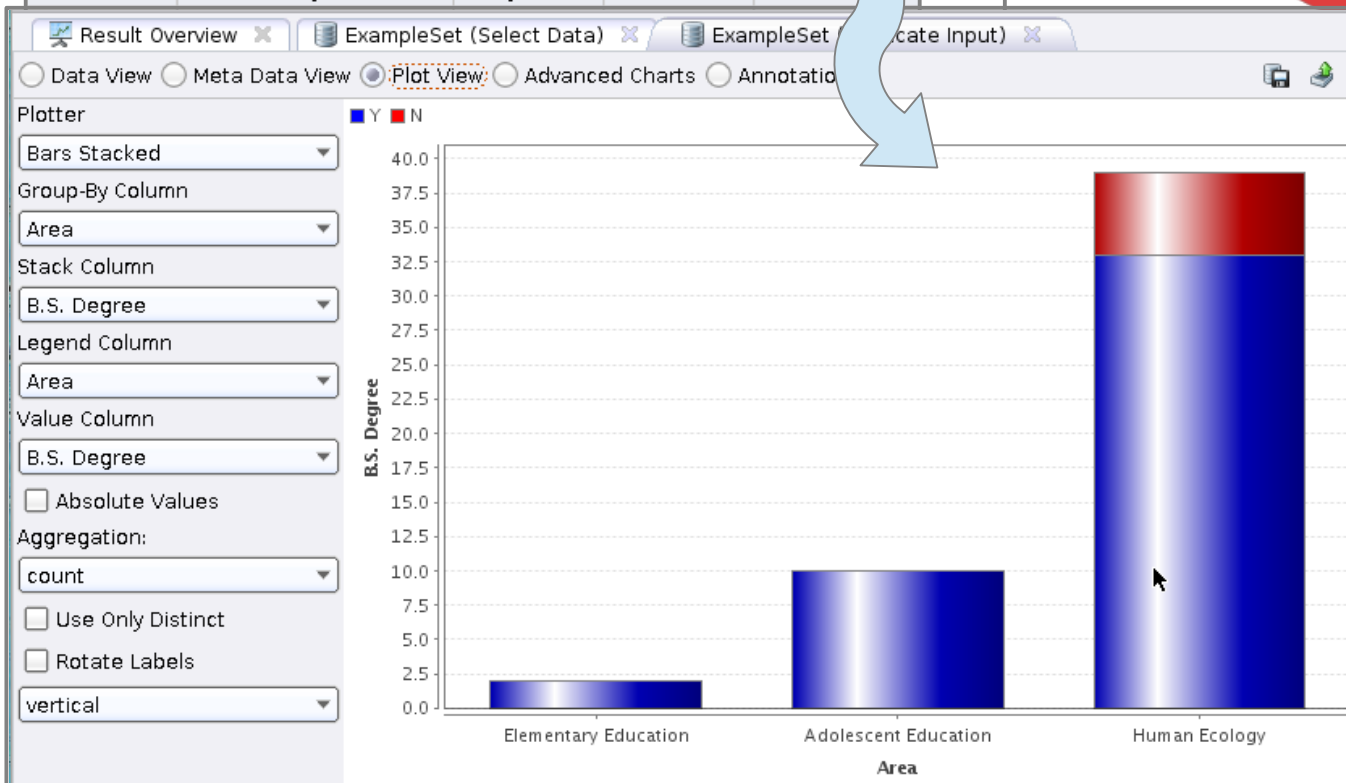
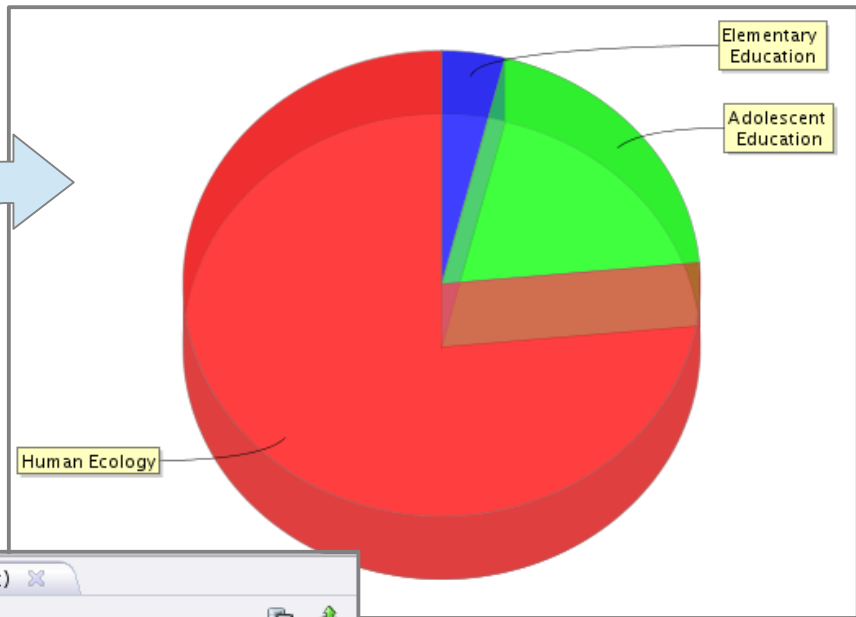
File Edit Process Tools View Help

Result Overview ExampleSet (Select Data)

Data View Meta Data View Plot View Advanced Charts

ExampleSet (51 examples, 0 special attributes, 5 regular attributes)

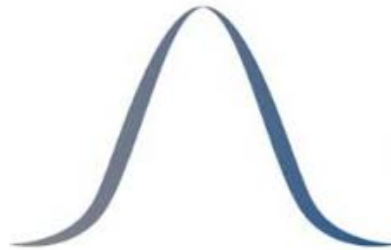
Row No.	Area	Major	B.A. Degree	B.S. De...
16	Human Ecology	Africana &	Y	N
17	Human Ecology	Anthropolo	Y	N
30	Human Ecology	Internation	Y	N
31	Human Ecology	Internation	Y	N
40	Human Ecology	Music	Y	N
41	Human Ecology	Music Indu:	Y	N
1	Elementary Education	Childhood I	N	Y
2	Elementary Education	Early Childf	N	Y



# Curating Datasets

Collect tweets for social science analysis:

- Partner with social dataset providers
- Enable students to capture own datasets



# How many tweets?

- Hydraulic fracturing: 70,000
- Animal welfare (orca): 370,000
- U.S. Supreme Court: 310,000
- Israel and Palestine: 2,000,000

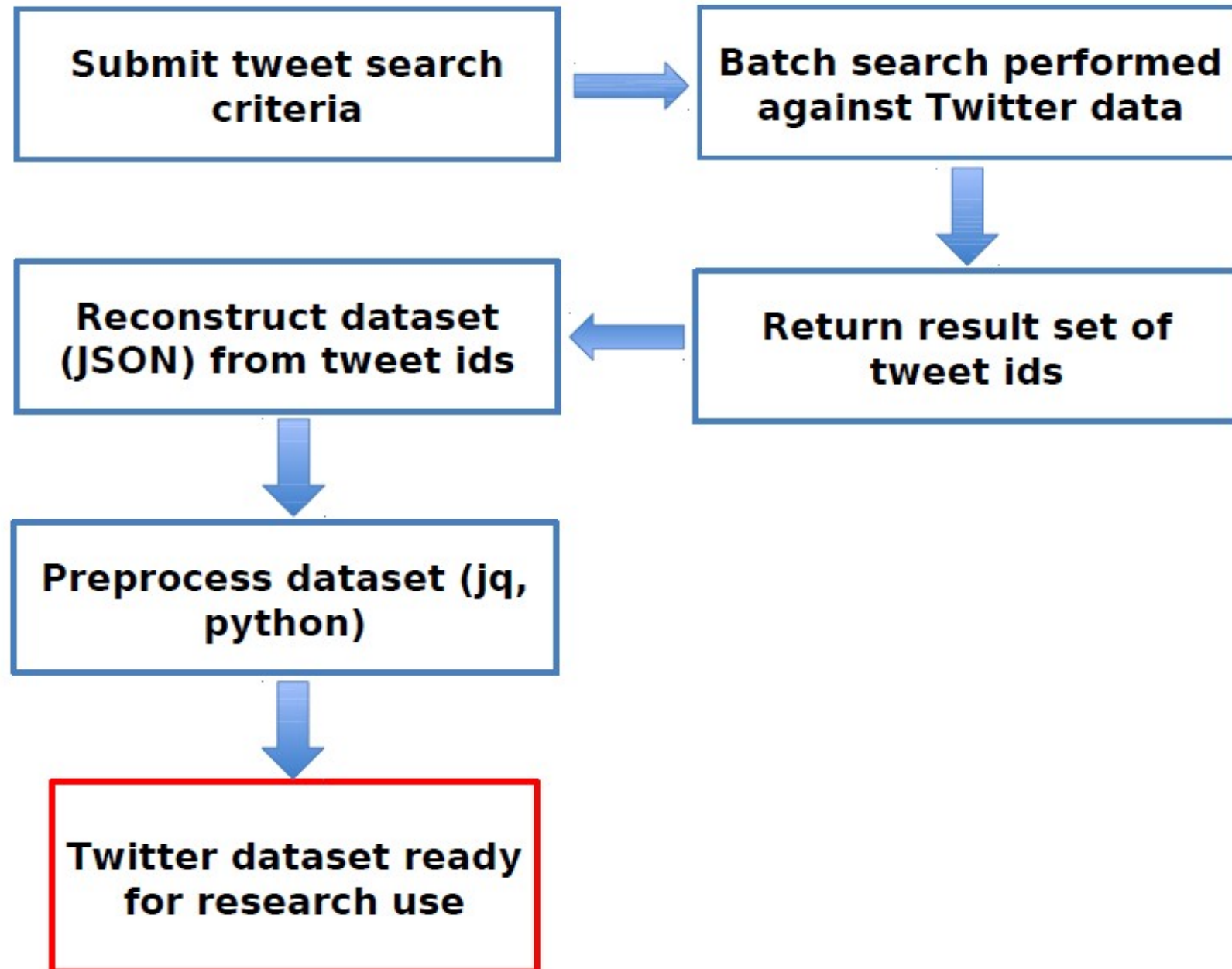


# Pitfalls: Purchasing Twitter Data

- Subscription service
  - Data reproducibility (no sampling)
  - Historical searches
- Licensing
  - Only tweet ids can be redistributed
  - Twitter/SUNY agreement
- Dataset reconstruction
  - Rate limited: 24,000 to 72,000 tweets/hour



# Twitter Data Acquisition



# VIDIA: Spring 2014

- Supported three Oneonta courses
- Deployed three data analysis tools
- Registered 76 student users
- Assigned student analyses:
  - k-Means Clustering
  - Word Co-Occurrences
- Enabled 25+ simultaneous tool sessions

CENTER FOR COMPUTATIONAL RESEARCH

---

# RapidMiner Sessions

on VIDIA

Month	Registered Tool Users	Tool Sessions Run	Total Tool Walltime (days)	Total Tool CPU Time (hours)
April 2014	69	460	16.8 days	3.36 hours
May 2014	40	186	12.0 days	2.52 hours
Cumulative Use: April 2014 to 18 Sep 2014	86	992	107 days	99 hours

CENTER FOR COMPUTATIONAL RESEARCH



**SUNY  
ONEONTA**

# Challenges

- User training: learning the platform and tools
- Technical performance details (Container creation and management)
- Scheduling HUBzero updates
- Browser compatibility (Java plugins)
- Dataset acquisition

CENTER FOR COMPUTATIONAL RESEARCH

---



# VIDIA: Fall 2014

- Supporting five Oneonta courses in the Social and Political Sciences
- Running HUBzero version 1.2.2
- Deploying additional tools (Gnu PSPP)
- Anticipating ~150 new student users
- Incorporating curated datasets

CENTER FOR COMPUTATIONAL RESEARCH

---



**SUNY  
ONEONTA**

# Plans

Plug HUBzero as an extensible teaching platform

- Include other SUNY colleges on VIDIA:
  - Brockport
  - New Paltz
- Spread the word:
  - National Park Service
  - EDUCAUSE 2015
  - SALT 2014

CENTER FOR COMPUTATIONAL RESEARCH

---

# SUNY



## The VIDIA Team



Gregory Fulkerson, Ph.D.  
Assistant Professor of Sociology



James Greenberg  
Director, TLTC



Brett Heindl, Ph.D.  
Assistant Professor of  
Political Science



Achim Koeddermann, Ph.D.  
Associate Professor of  
Philosophy and Env. Sciences



Brian M. Lowe, Ph.D.  
Associate Professor  
of Sociology



Diana Moseman  
Instructional Designer and  
Programmer, TLTC



Harry Pence, Ph.D.  
Distinguished Professor of  
Chemistry



Tim Ploss  
Instructional Designer, TLTC



Bill Wilkerson, Ph.D.  
Associate Professor of Political  
Science



Steven M. Gallo  
Lead Software Engineer  
CCR, University at Buffalo



Jeanette Sperhac  
Scientific Programmer  
CCR, University at Buffalo

CENTER FOR COMPUTATIONAL RESEARCH

