VIDIA: a HUBzero Gateway for Data Analytics Education

Jeanette Sperhac and Steven M. Gallo
Center for Computational Research, University at Buffalo

11th Gateway Computing Environments Conference
San Diego Supercomputer Center
San Diego, CA
2 November 2016

CENTER FOR **COMPUTATIONAL RESEARCH**

University at Buffalo
The State University of New York

SCIENCE GATEWAYS .ORG

Outline

1. The collaboration
2. The VIDIA instance
3. Supporting teaching
    - Tools
    - Coursework
4. Enabling research
5. Demonstration

The VIDIA Collaboration

- Academic supercomputing facility
- Buffalo, New York
- 10,000+ cores, 170+ Tflops compute capacity

- Four-year liberal arts college
- Oneonta, New York
- Enrollment 5,900 students

CENTER FOR **COMPUTATIONAL RESEARCH**

University at Buffalo
The State University of New York

SUNY ONEONTA

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

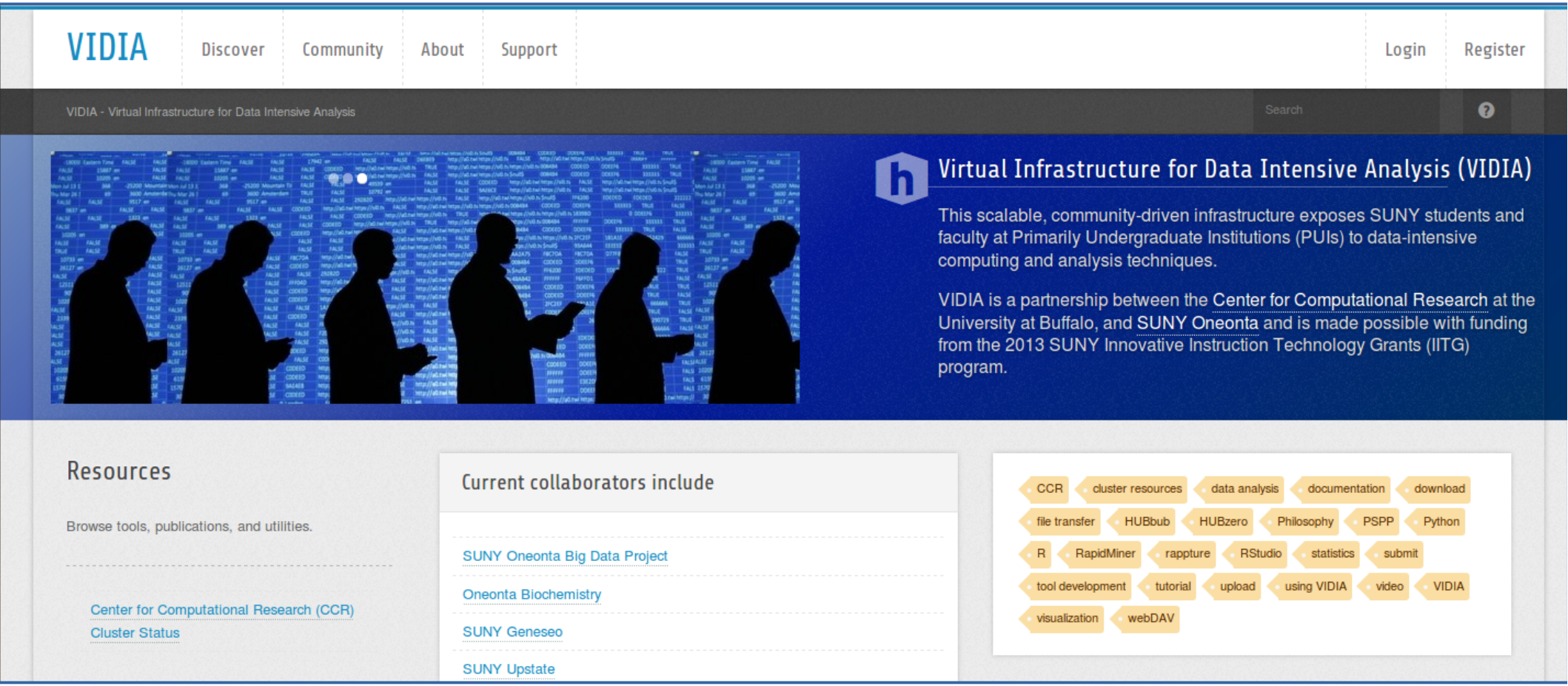
University at Buffalo
The State University of New York

SUNY ONEONTA

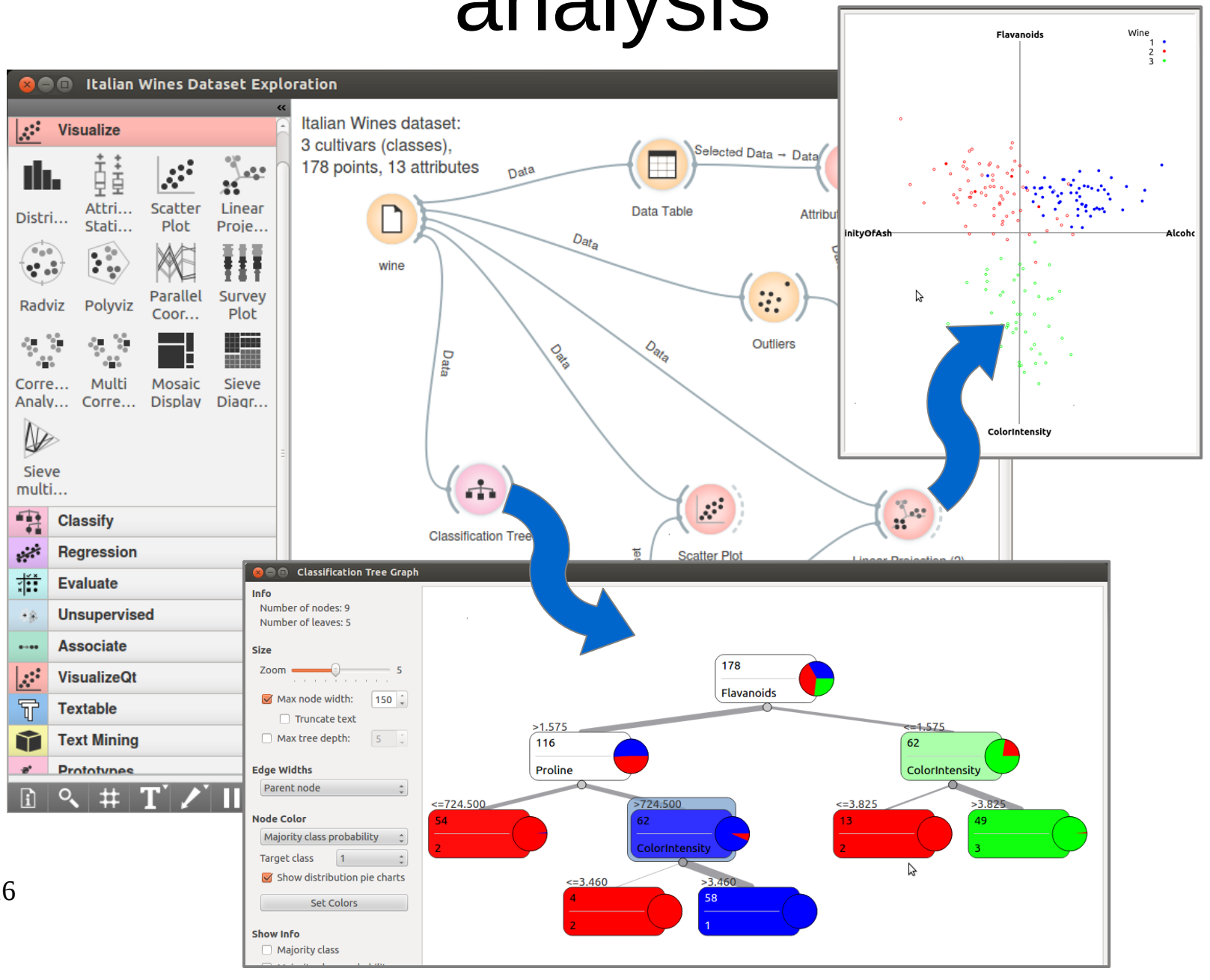
How is this platform different?

- Developed for scientific collaboration
- Access via web browser
- Computational horsepower
  - Deployed tools run in virtual containers
  - HPC access via *submit*

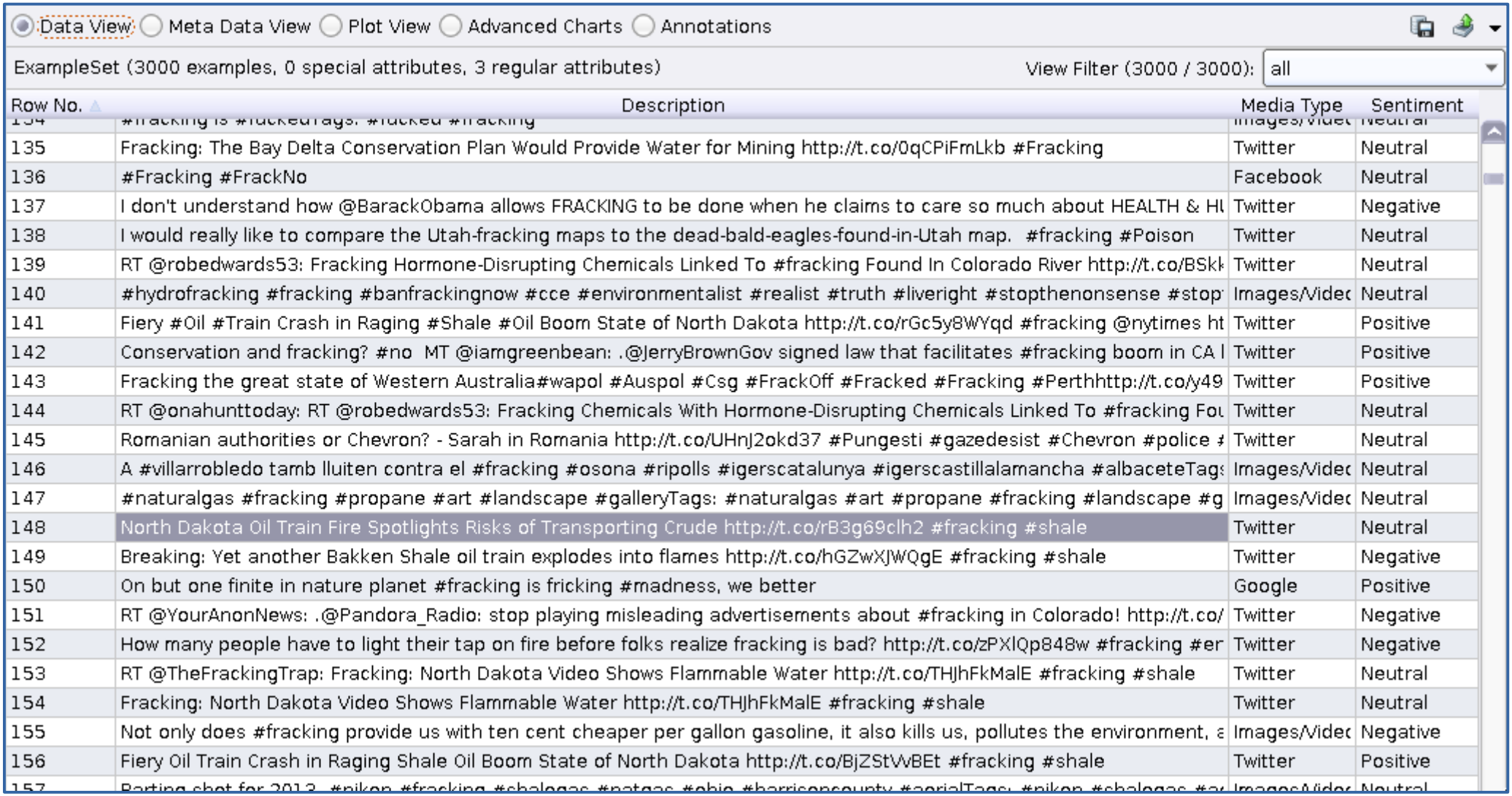
vidia.ccr.buffalo.edu



Provide workflow tools for data analysis



Curate datasets of social scientific interest



Data View | Meta Data View | Plot View | Advanced Charts | Annotations

ExampleSet (3000 examples, 0 special attributes, 3 regular attributes) View Filter (3000 / 3000): all

Row No.	Description	Media Type	Sentiment
134	#fracking is #fuckedtags: #fucked #fracking	Images/Videc	Neutral
135	Fracking: The Bay Delta Conservation Plan Would Provide Water for Mining http://t.co/0qCPiFmLkb #Fracking	Twitter	Neutral
136	#Fracking #FrackNo	Facebook	Neutral
137	I don't understand how @BarackObama allows FRACKING to be done when he claims to care so much about HEALTH & HU	Twitter	Negative
138	I would really like to compare the Utah-fracking maps to the dead-bald-eagles-found-in-Utah map. #fracking #Poison	Twitter	Neutral
139	RT @robedwards53: Fracking Hormone-Disrupting Chemicals Linked To #fracking Found In Colorado River http://t.co/BSkk	Twitter	Neutral
140	#hydrofracking #fracking #banfrackingnow #cce #environmentalist #realist #truth #liveright #stopthenonsense #stop	Images/Videc	Neutral
141	Fiery #Oil #Train Crash in Raging #Shale #Oil Boom State of North Dakota http://t.co/rGc5y8WYqd #fracking @nytimes ht	Twitter	Positive
142	Conservation and fracking? #no MT @iamgreenbean: .@JerryBrownGov signed law that facilitates #fracking boom in CA l	Twitter	Positive
143	Fracking the great state of Western Australia#wapol #Auspol #Csg #FrackOff #Fracked #Fracking #Perthhttp://t.co/y49	Twitter	Positive
144	RT @onahunttoday: RT @robedwards53: Fracking Chemicals With Hormone-Disrupting Chemicals Linked To #fracking Fou	Twitter	Neutral
145	Romanian authorities or Chevron? - Sarah in Romania http://t.co/UHnJ2okd37 #Pungesti #gazedesist #Chevron #police #	Twitter	Neutral
146	A #villarrobledo tamb lluiten contra el #fracking #osona #ripolls #igerscatalunya #igerscastillalamancha #albaceteTags	Images/Videc	Neutral
147	#naturalgas #fracking #propane #art #landscape #galleryTags: #naturalgas #art #propane #fracking #landscape #g	Images/Videc	Neutral
148	North Dakota Oil Train Fire Spotlights Risks of Transporting Crude http://t.co/rB3g69clh2 #fracking #shale	Twitter	Neutral
149	Breaking: Yet another Bakken Shale oil train explodes into flames http://t.co/hGZwXJWQgE #fracking #shale	Twitter	Negative
150	On but one finite in nature planet #fracking is fricking #madness, we better	Google	Positive
151	RT @YourAnonNews: .@Pandora_Radio: stop playing misleading advertisements about #fracking in Colorado! http://t.co/	Twitter	Negative
152	How many people have to light their tap on fire before folks realize fracking is bad? http://t.co/zPXlQp848w #fracking #er	Twitter	Negative
153	RT @TheFrackingTrap: Fracking: North Dakota Video Shows Flammable Water http://t.co/THJhFkMalE #fracking #shale	Twitter	Neutral
154	Fracking: North Dakota Video Shows Flammable Water http://t.co/THJhFkMalE #fracking #shale	Twitter	Neutral
155	Not only does #fracking provide us with ten cent cheaper per gallon gasoline, it also kills us, pollutes the environment, a	Images/Videc	Negative
156	Fiery Oil Train Crash in Raging Shale Oil Boom State of North Dakota http://t.co/BjZStVvBEt #fracking #shale	Twitter	Positive
157	Parting shot for 2013 #nikon #fracking #shalegas #natgas #ohio #harrisoncounty #aerialTags: #nikon #shalegas #a	Images/Videc	Neutral

Enable access to HPC resources

VIDIA Hardware

HUBzero instance: Dell PowerEdge R720xd

- 2x 6-core Intel Xeon E5-2630

  2.30 GHz, 15M cache

- 48 TB raw SATA disk space

  ~36 TB usable

- 128 GB memory

  16x8GB - 1333MHz DIMMS

Teaching on HUBzero

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

Initial Teaching Toolset

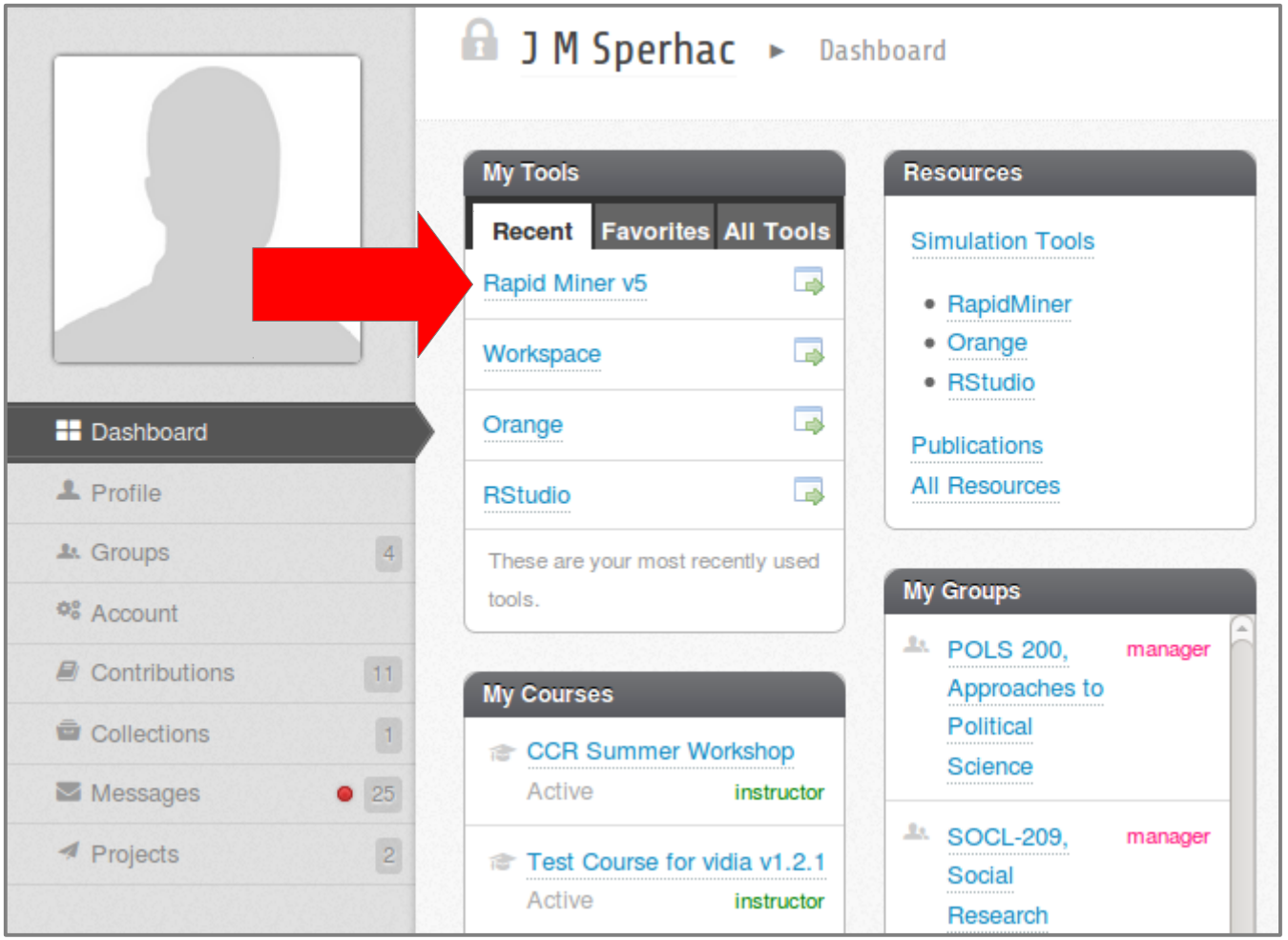
Features wanted:

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

Tools deployed:

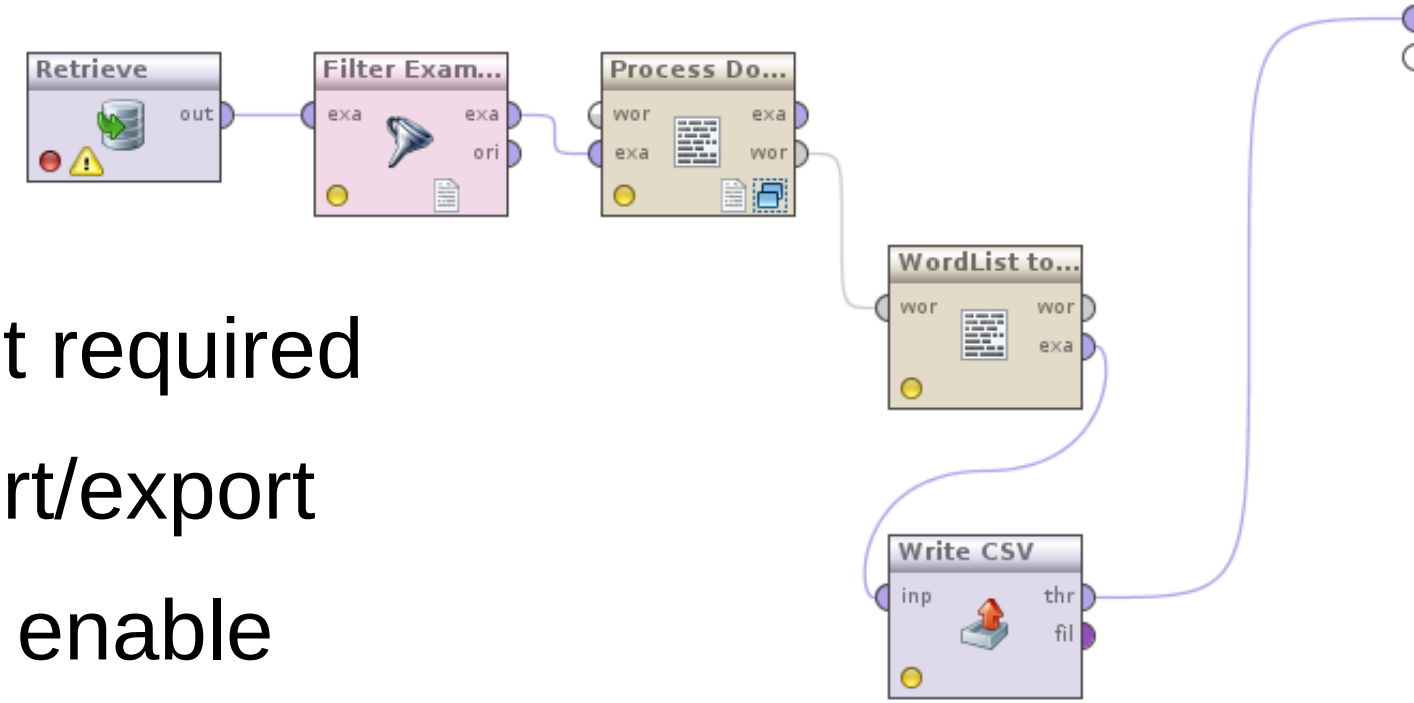
- RapidMiner
- Orange
- RStudio
- Python/Octave IDEs
- Gnu PSPP

Deployed tool on VIDIA

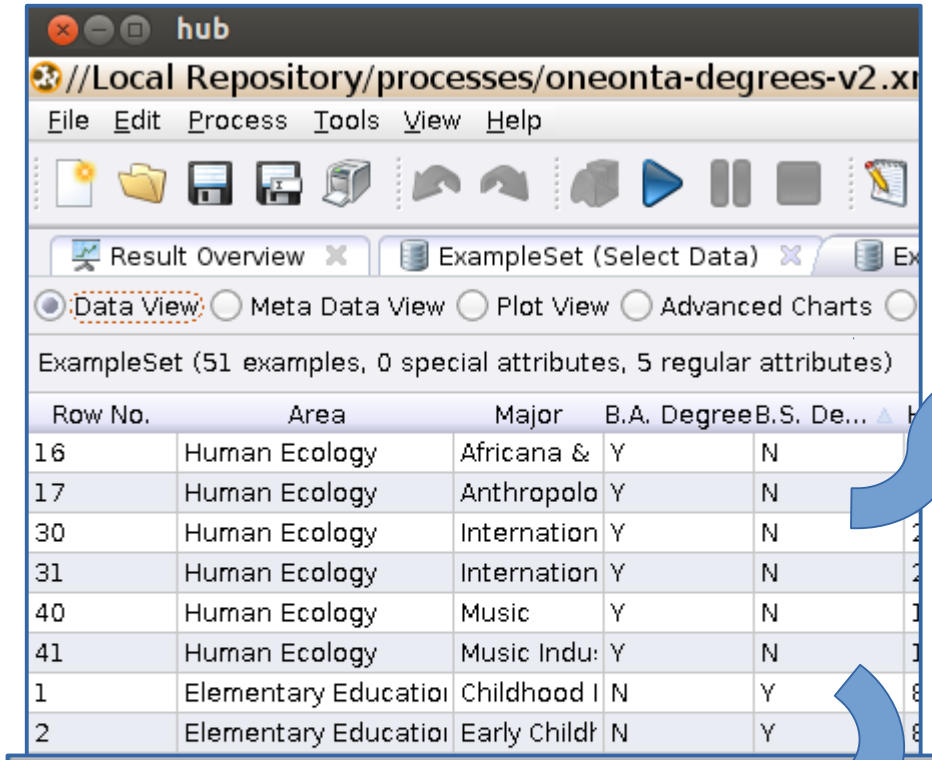


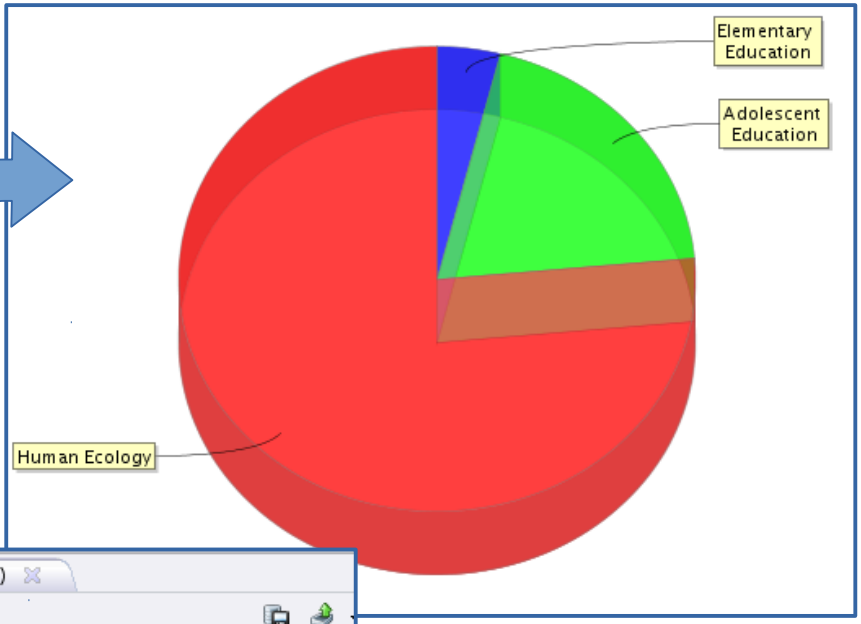
RapidMiner Data Mining Tool

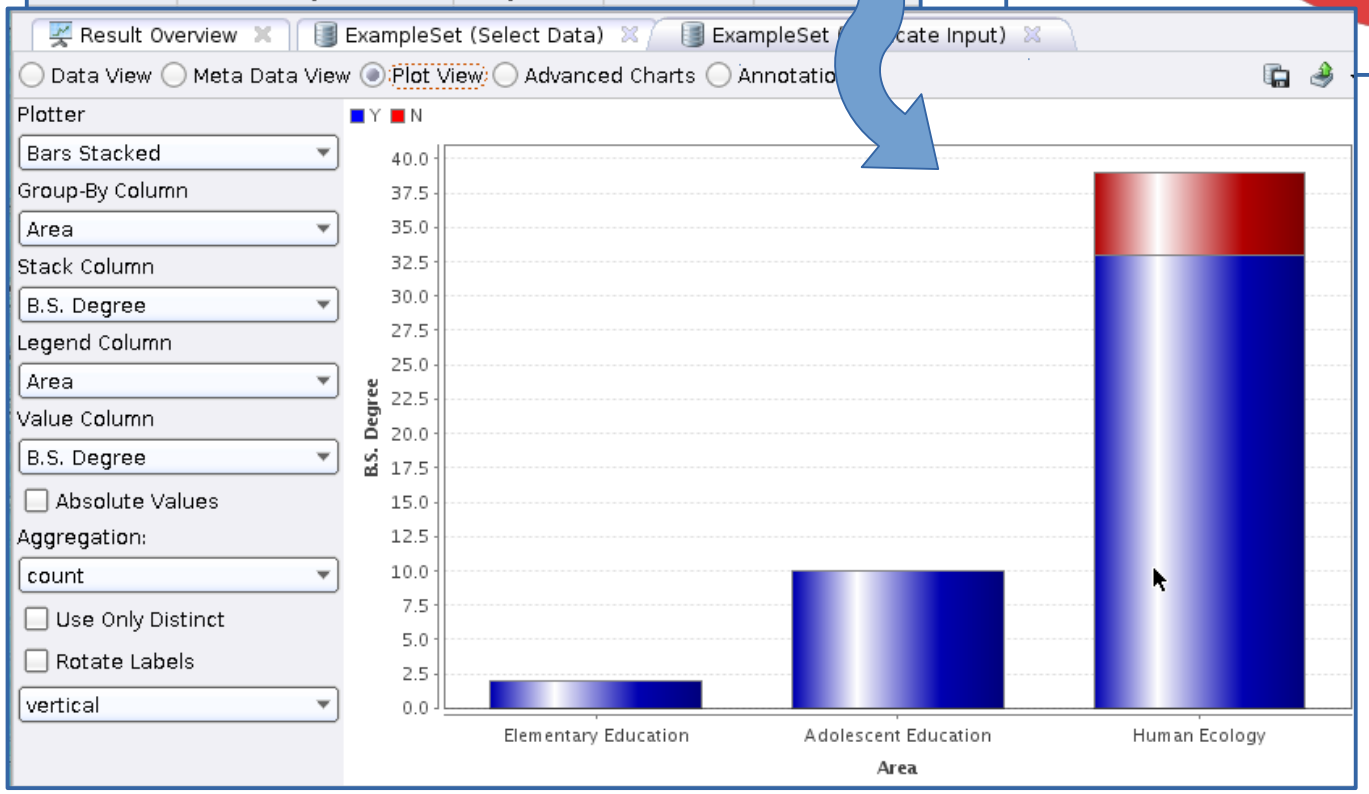
RapidMiner Tool



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


hub
//Local Repository/processes/oneonta-degrees-v2.xr
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 Educatio Childhood I N Y
2 Elementary Educatio Early Childh N Y


Elementary Education
Adolescent Education
Human Ecology


Result Overview
ExampleSet (Select Data)
Data View
Meta Data View
Plot View
Advanced Charts
Plotter
Bars Stacked
Group-By Column
Area
Stack Column
B.S. Degree
Legend Column
Area
Value Column
B.S. Degree
Absolute Values
Aggregation:
count
Use Only Distinct
Rotate Labels
vertical
Y N
B.S. Degree
40.0
37.5
35.0
32.5
30.0
27.5
25.0
22.5
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Elementary Education
Adolescent Education
Human Ecology
Area

rapidminer

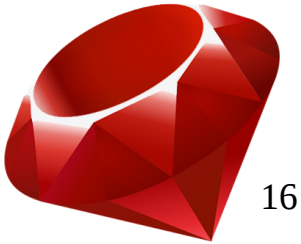
Development Environment
(write your own tools)

Features:

- Linux container
- Command line
- HPC access
- Source code control
- GUI builder
- Compilers

Supported languages:

- Fortran, C, C++
- Python, Perl, Ruby
- Octave
- Java
- R
- ...



Coursework on VIDIA

- POLS-284: U.S. Foreign Policy
- PHIL-231: Media Ethics
- SOCL-209: Social Research Methods
- SOCL-294: Animals and Society
- SOCL-244: Environmental Sociology

CENTER FOR COMPUTATIONAL RESEARCH

University at Buffalo
The State University of New York

SUNY ONEONTA

Case Study: Animals and Society

- Sociology majors, 200 level
- No programming experience
- Comparative, social scientific analyses

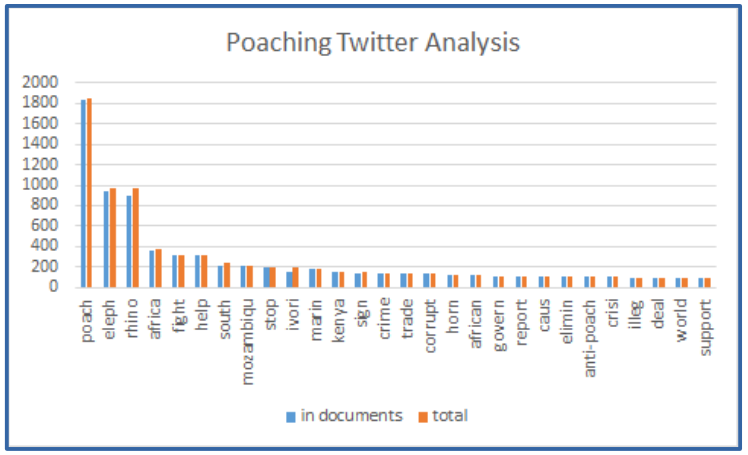
Gather, organize, and interpret social media data

CENTER FOR **COMPUTATIONAL RESEARCH**

University at Buffalo
The State University of New York

SUNY ONEONTA

Case Study: Animals and Society



Chose search terms from journal articles on animal poaching.

Collected data set of *1831 tweets* on these search terms.

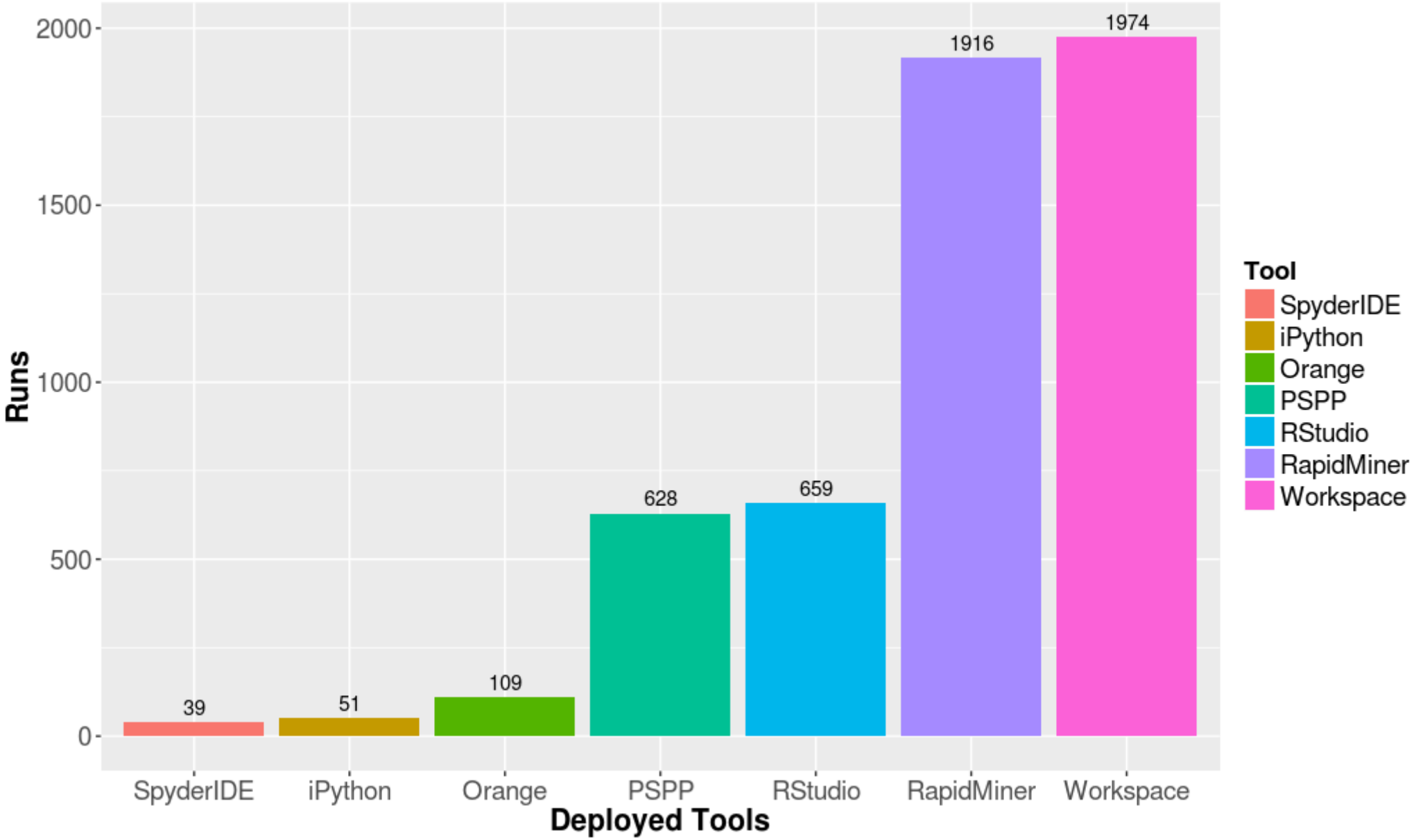
Generated list of commonly occurring terms using RapidMiner.

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

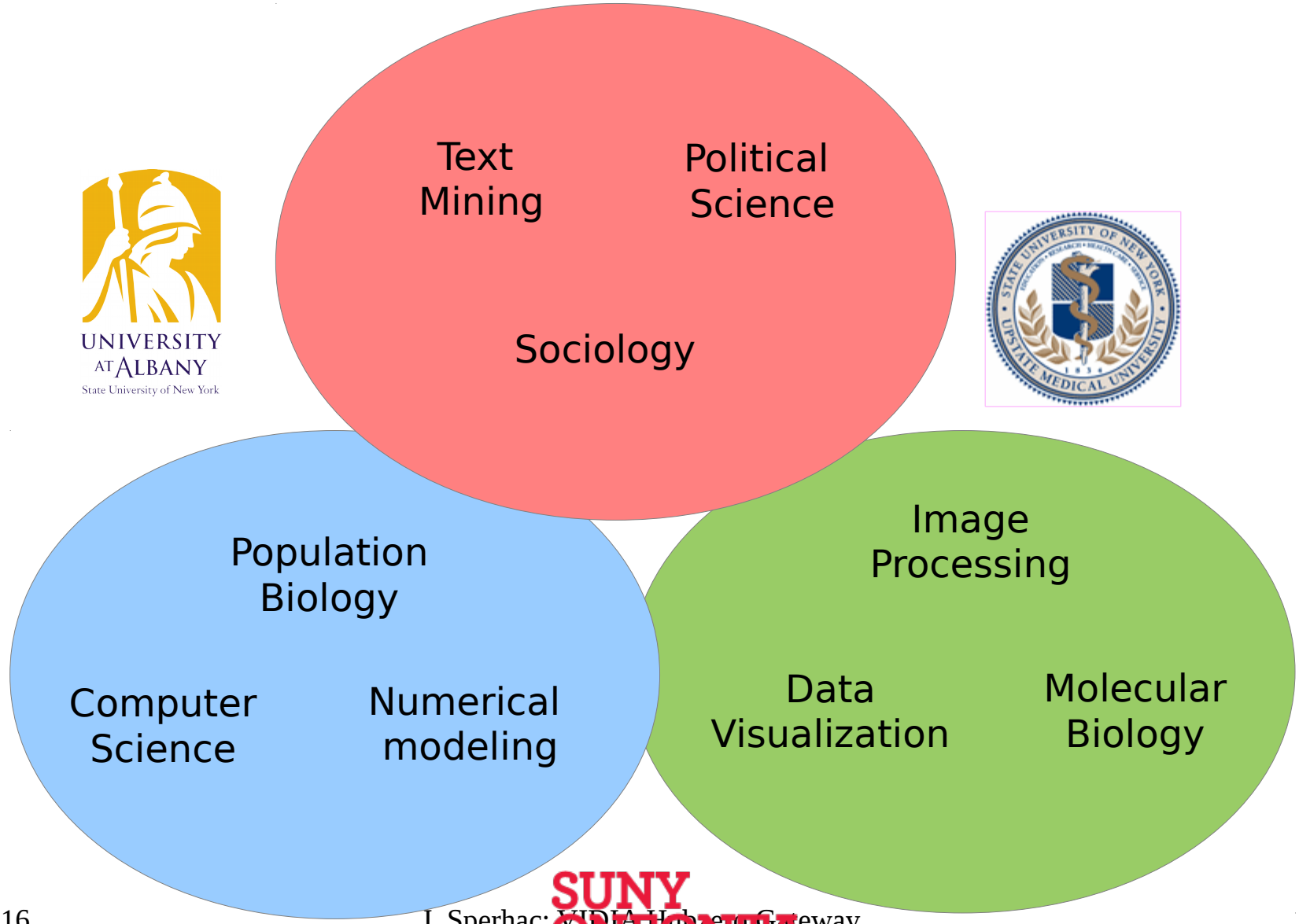
- Elephant (946 tweets/976 total) and Rhino (896/976)
- Africa (363/376) and Fight (313/317)
- Help (311/312) and Stop (192/198)

A SUNY Oneonta student, quoted in W. Wilkerson, "Using 'Big Data' in a Political Science Research Methods Course"

VIDIA Tool Sessions, to 2016-09



VIDIA supports research...

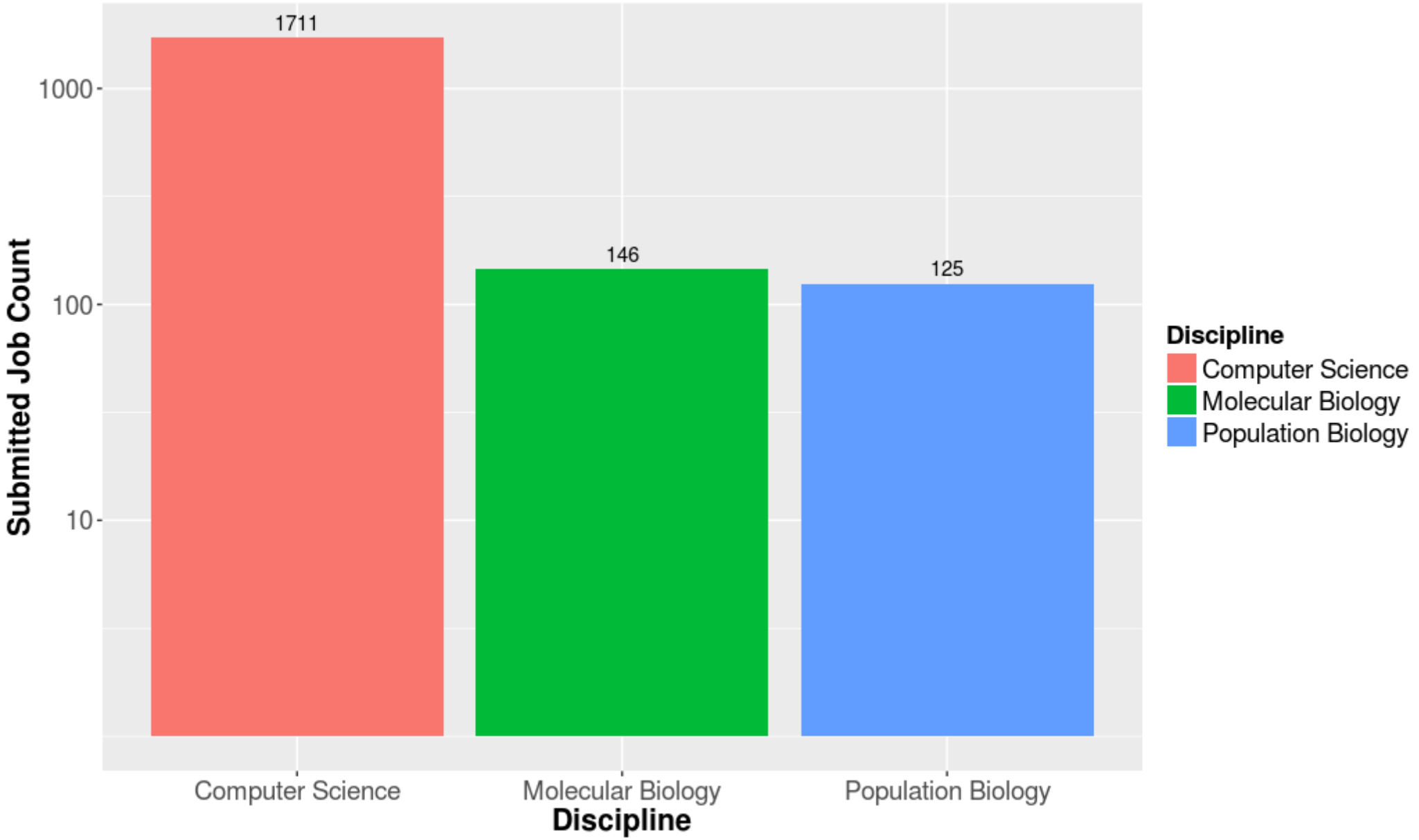


Enabling Research

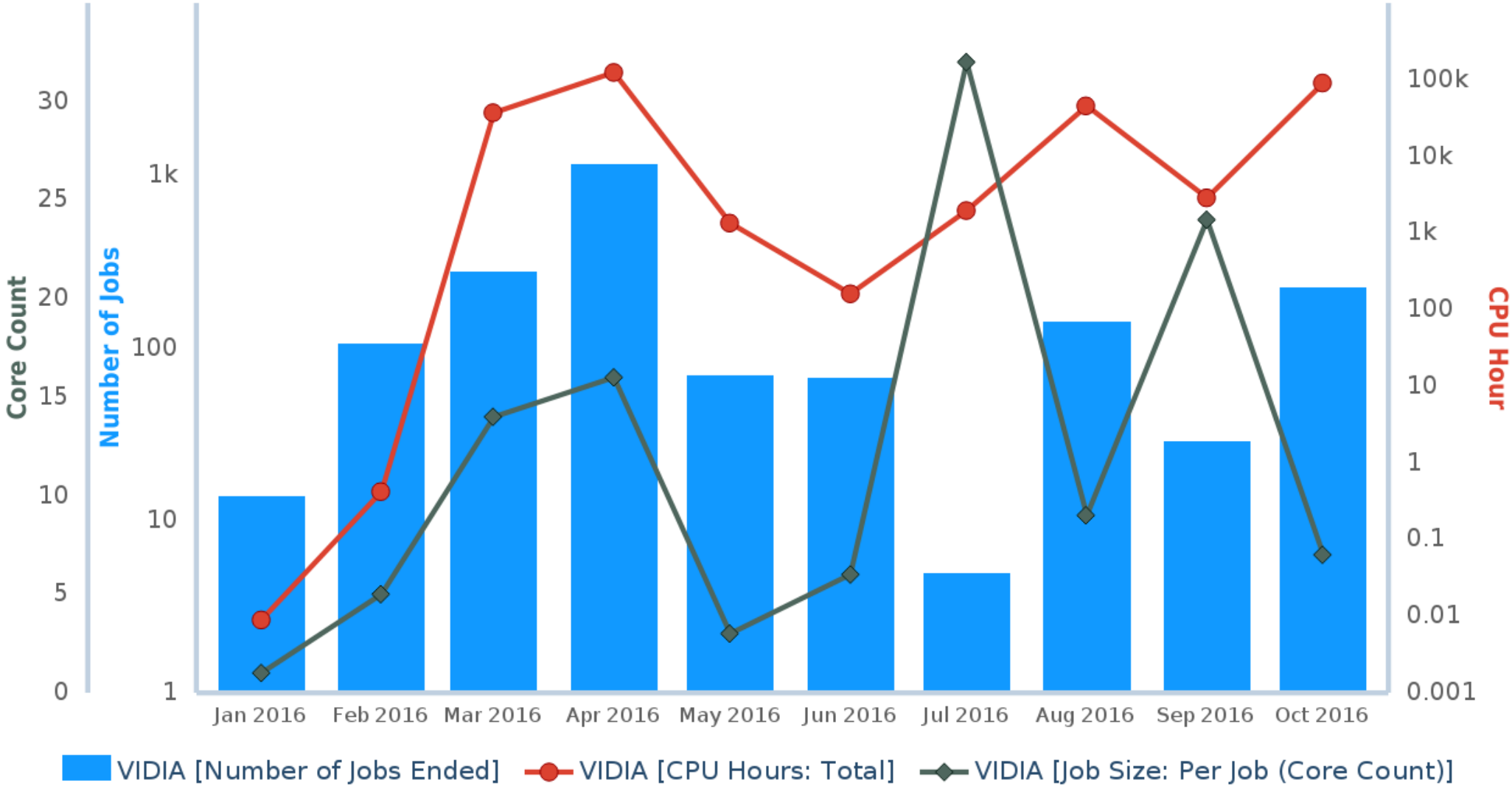
Research	Tool	Discipline
Electron Microscopy Image Analysis	HPC	Molecular Biology
Supreme Court in Public Opinion	R	Political Science
Controversies in Animal Rights	RapidMiner	Sociology
Software Anomaly Detection	HPC	Computer Science
American Shad Migration Modeling	R, HPC	Population Biology

VIDIA HPC Cluster Usage

2016-01 to 2016-10

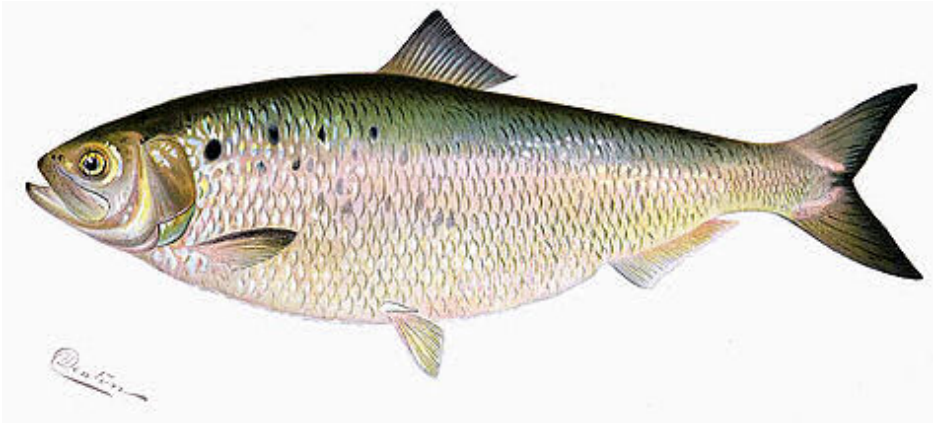


VIDIA HPC Cluster Usage 2016



XDMoD
METRICS ON DEMAND

American Shad Migration Model



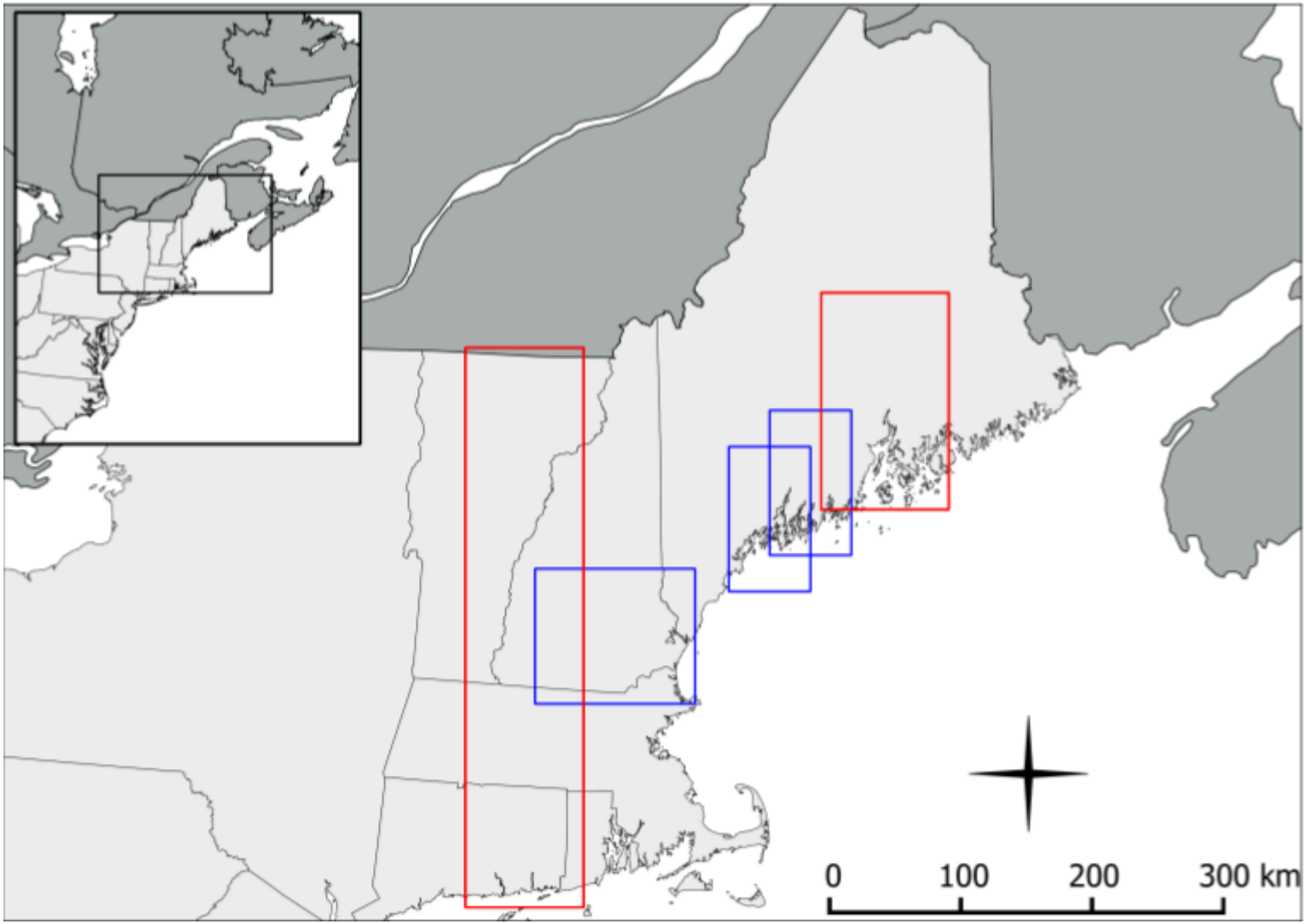
- Daniel Stich, Oneonta
- Data from literature, surveys
- Statistical models
- Advice for dam managers, fisheries

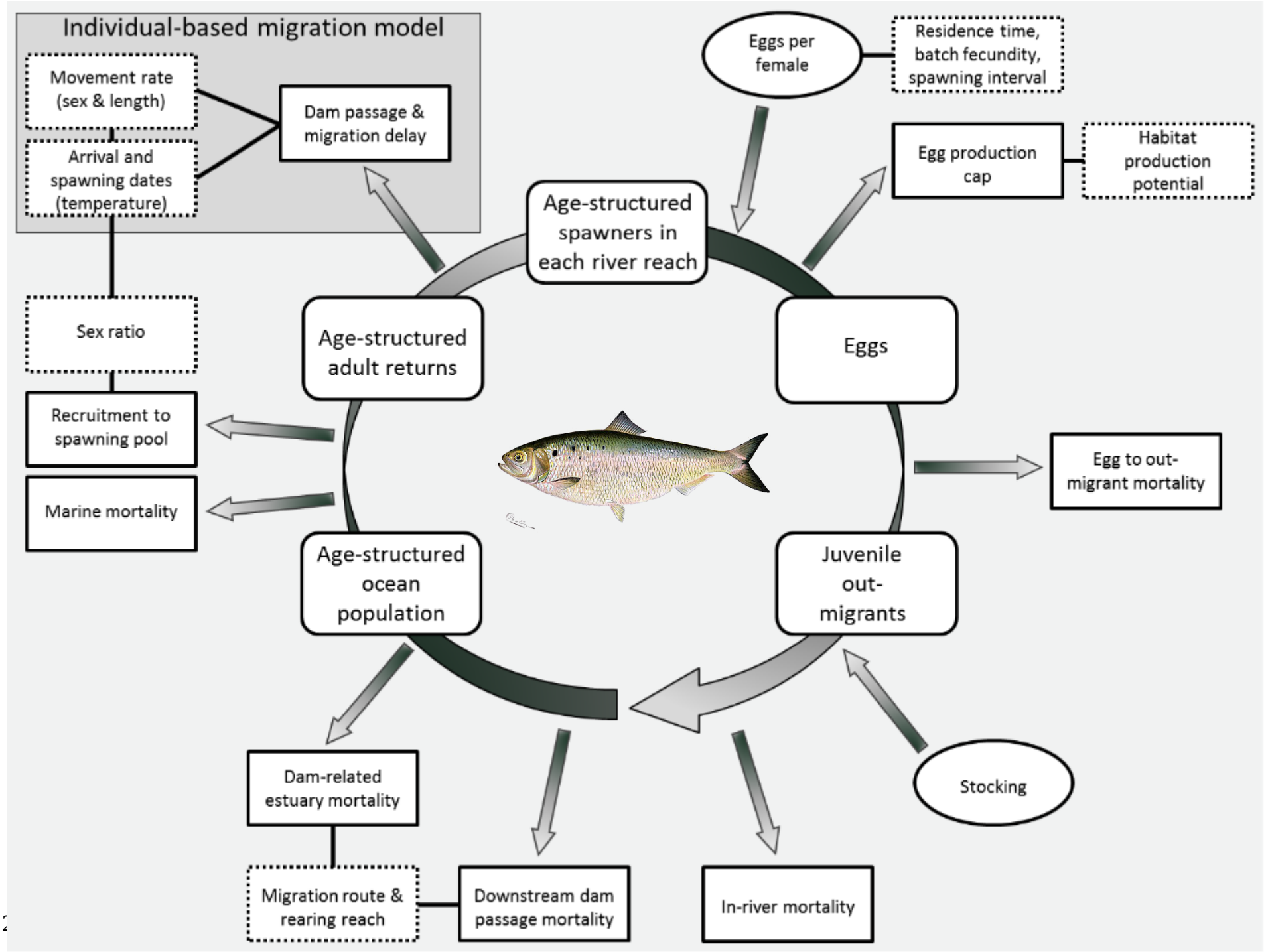
CENTER FOR COMPUTATIONAL RESEARCH

University at Buffalo
The State University of New York

SUNY ONEONTA

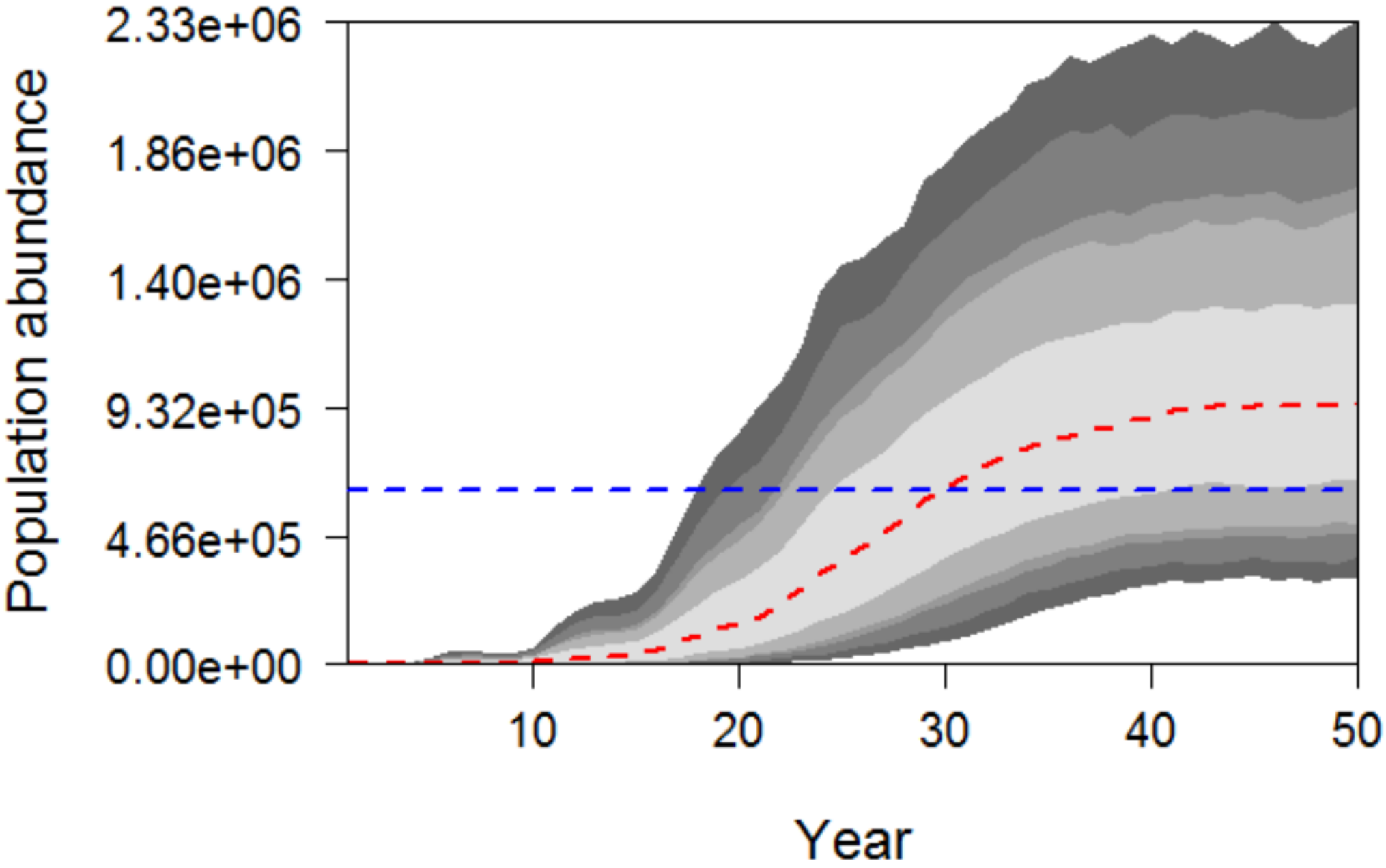
Shad migration: river systems



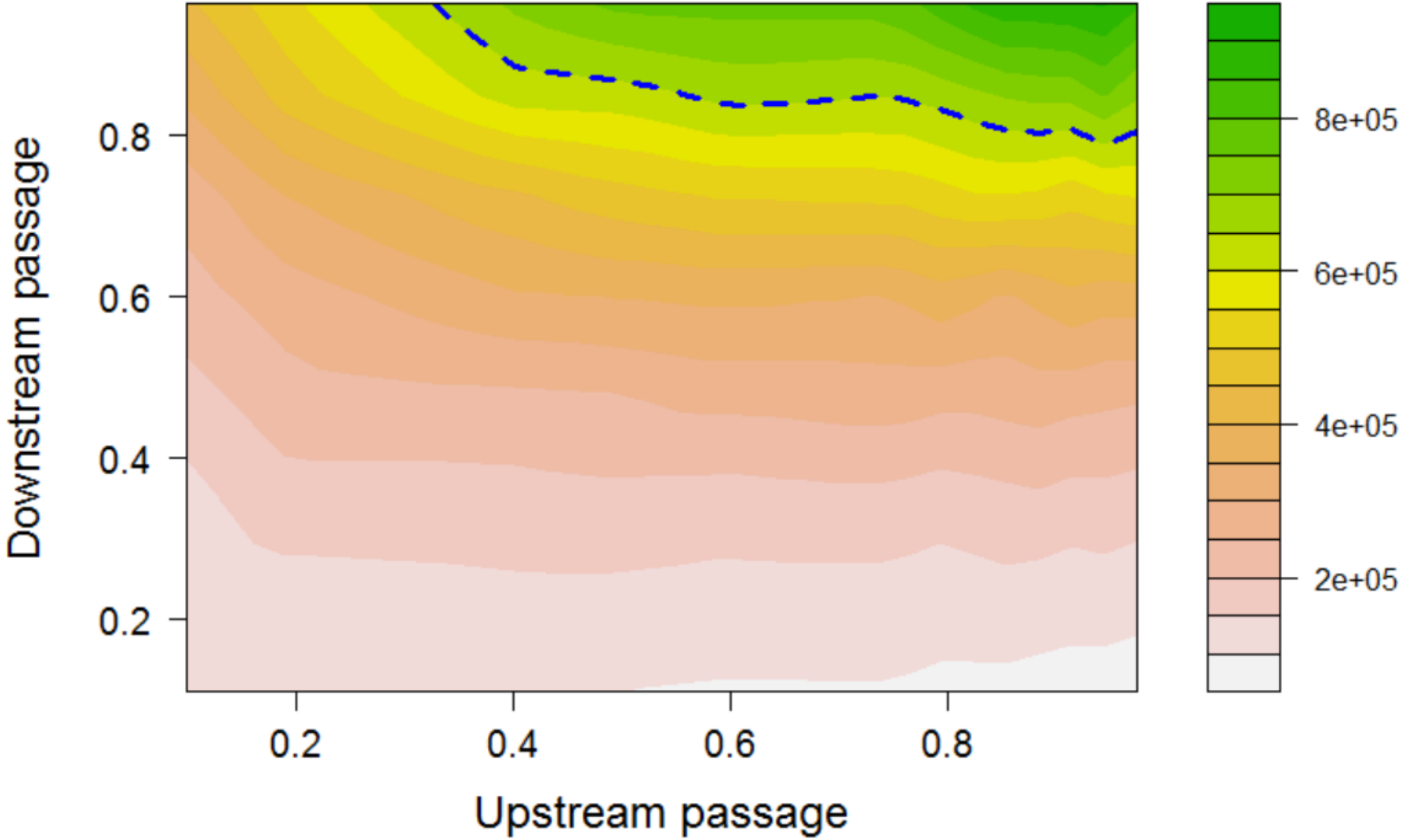

Individual-based migration model
Movement rate (sex & length)
Arrival and spawning dates (temperature)
Dam passage & migration delay
Sex ratio
Recruitment to spawning pool
Marine mortality
Eggs per female
Residence time, batch fecundity, spawning interval
Egg production cap
Habitat production potential
Age-structured spawners in each river reach
Eggs
Egg to out-migrant mortality
Juvenile out-migrants
Stocking
In-river mortality
Downstream dam passage mortality
Migration route & rearing reach
Dam-related estuary mortality
Age-structured ocean population
Age-structured adult returns

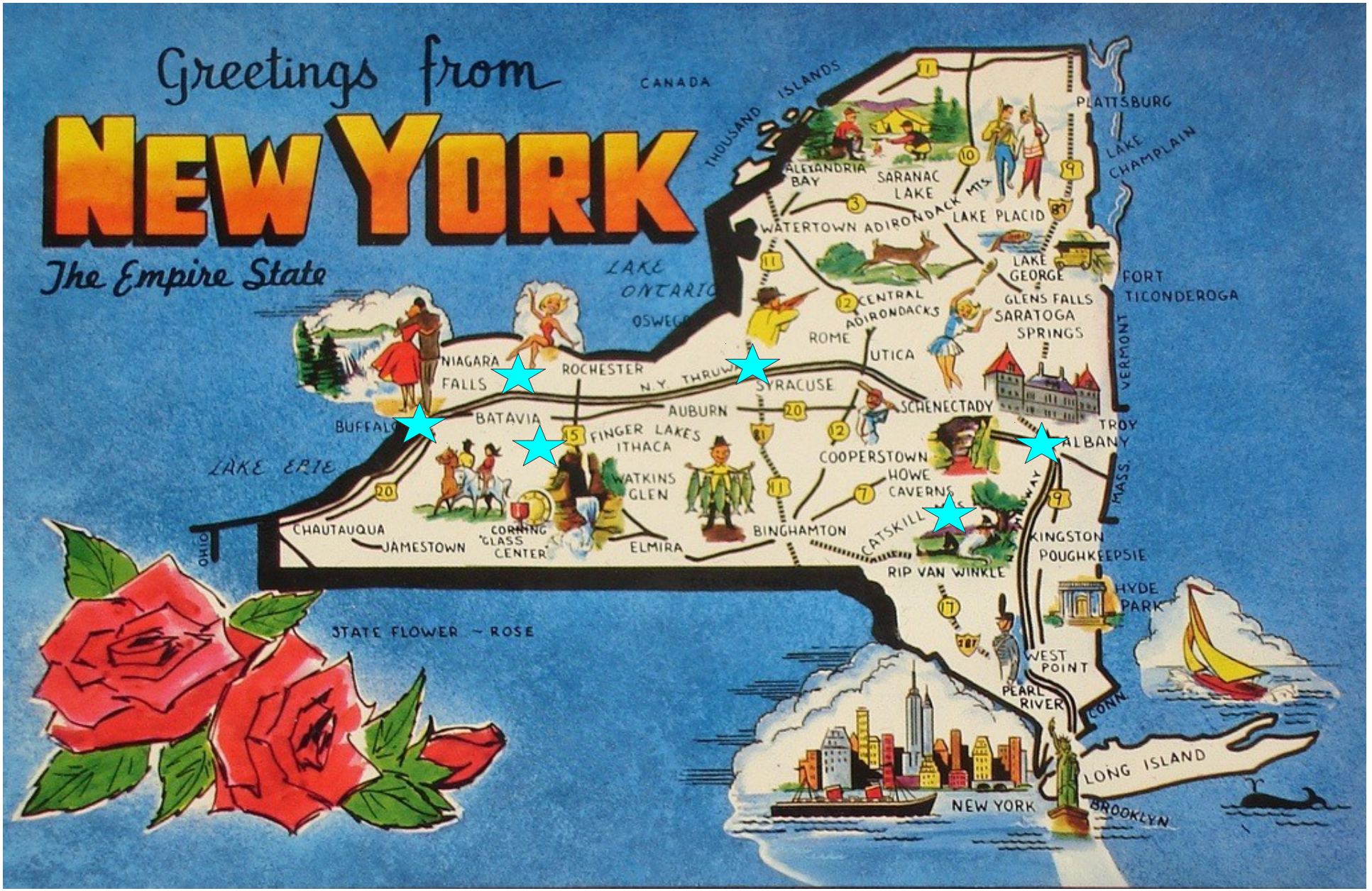
Preliminary shad model output

Penobscot River, no dams



Shad model: spawner abundance




Greetings from
NEW YORK
The Empire State
CANADA
THOUSAND ISLANDS
PLATTSBURG
LAKE CHAMPLAIN
ALEXANDRIA BAY
SARANAC LAKE
ADIRONDACK MTS.
WATERTOWN
LAKE PLACID
LAKE ONTARIO
LAKE GEORGE
FORT TICONDEROGA
CENTRAL ADIRONDACKS
GLENS FALLS
SARATOGA SPRINGS
ROME
UTICA
VERMONT
NIAGARA FALLS
ROCHESTER
N.Y. THRUWAY
SYRACUSE
SCHENECTADY
TROY
ALBANY
BUFFALO
BATAVIA
AUBURN
FINGER LAKES
ITHACA
COOPERSTOWN
HOWE CAVERNS
LAKE ERIE
WATKINS GLEN
MASS.
CHAUTAUQUA
JAMESTOWN
CORNING "GLASS CENTER"
ELMIRA
BINGHAMTON
CATSKILL
KINGSTON
POUGHKEEPSIE
RIP VAN WINKLE
HYDE PARK
OHIO
STATE FLOWER ~ ROSE
WEST POINT
PEARL RIVER
CONN.
LONG ISLAND
NEW YORK
BROOKLYN

Virtual Infrastructure for Data Intensive Analysis (VIDIA)

SUNY Buffalo/CCR
- Steve Gallo
- Tom Furlani

SUNY Geneseo
- Kirk M. Anne

SUNY Albany
- Ming Ying
- Yizhen Chen

SUNY Oneonta
- James Greenberg
- Gregory Fulkerson
- Brett Heindl
- Achim Koeddermann
- Brian M. Lowe
- Bill Wilkerson
- Dan Stich

SUNY Upstate Medical
- Nicholas Stam