Big Data infrastructure and tools in libraries

PURDUE UNIVERSITY

Line Pouchard, PhD

Purdue University Libraries

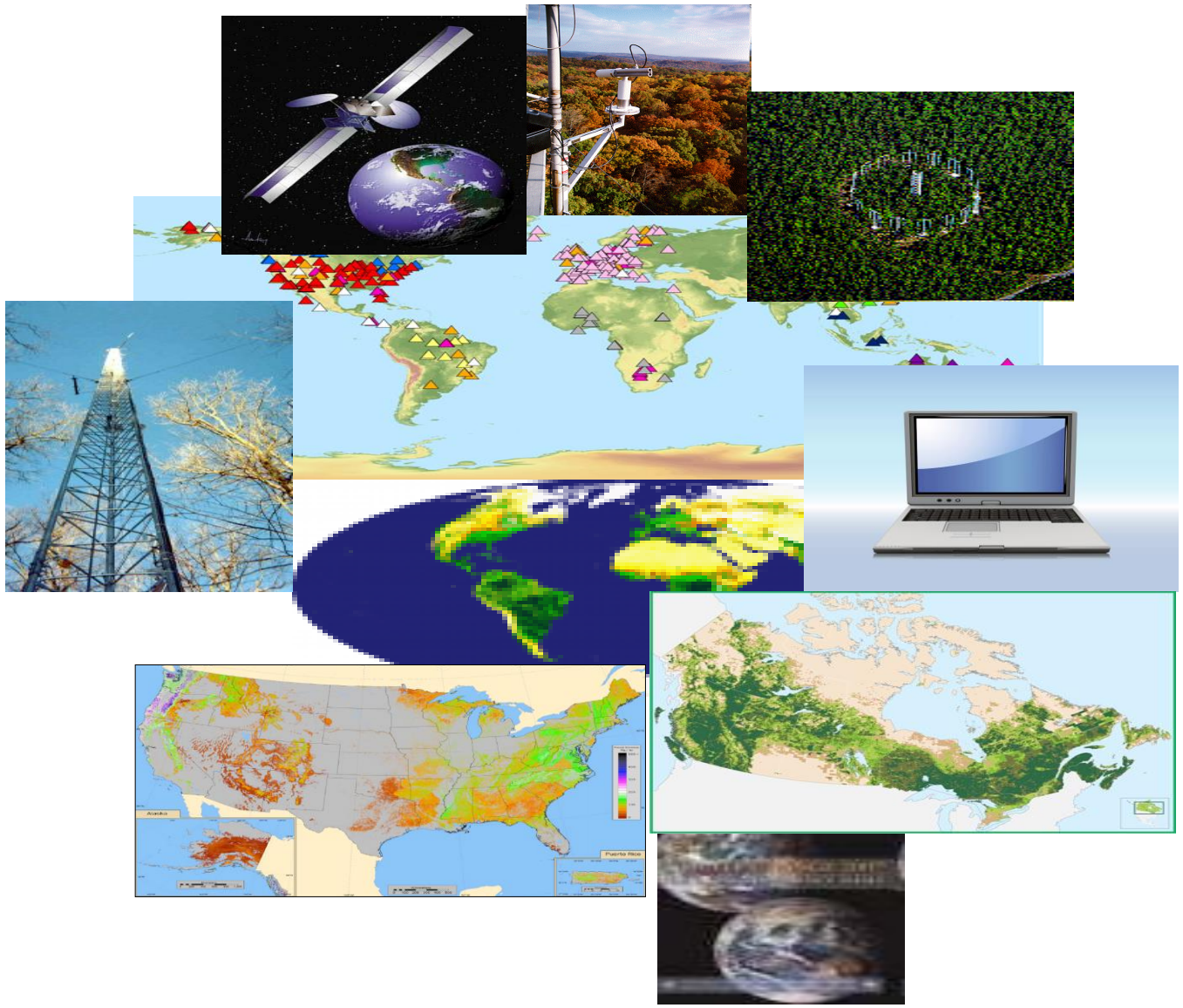
Research Data Group

08/10/2016

DATA IN LIBRARIES: THE BIG PICTURE

IFLA/
UNIVERSITY OF CHICAGO

BIG DATA: A VERY DIVERSE DATA AND METADATA ECOSYSTEM



OUTLINE

- The long tail of science
  - Needs and issues around storage options
  - Data Depot at Purdue
- The Big Data life cycle:
  - A tool for working with Big Data
- Use case: Studying policies in the CAM2 project

BIG DATA EXPERIENCES AND IT PROJECTS IN LIBRARIES

- Big Data cannot be managed, preserved, curated by libraries alone
  - A common strategy is needed
- A continuous collaboration with IT departments is required
  - Often difficult, conflicts, turf wars
  - Combination of soft skills and hardware is needed
- Need to identify the right person in both IT and libraries
- A possible division of roles could be:
  - System management by IT
  - Services provided by libraries
  - Education provided by joint teams formed of IT and Libraries staff
- A common communication strategy also helps
- A commitment of campus administration is crucial for success

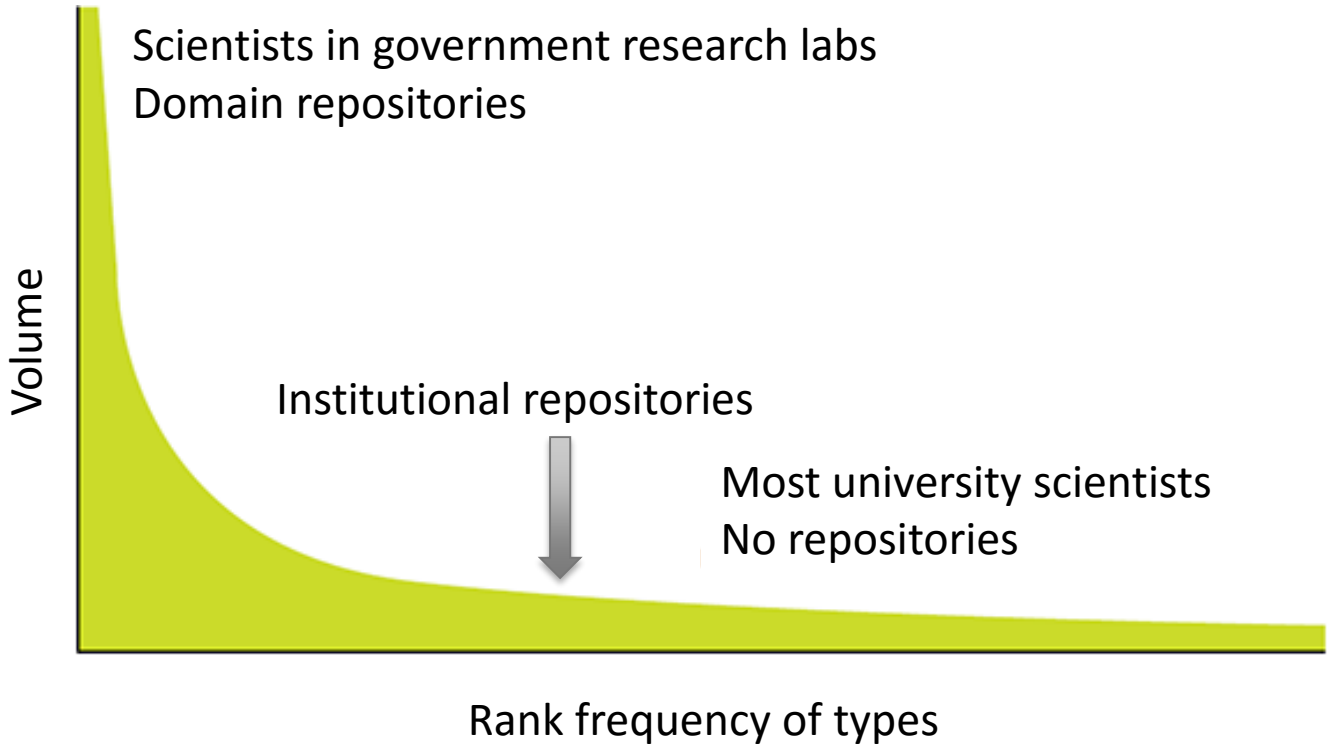
THE LONG TAIL OF SCIENCE

Head

- Big science
- Big data
- Large collaborations
- Agency-sponsored data collection
- Long-term perspective
- Common standards
- Well preserved and curated
- Expensive

Tail

- Small Data
- Small collaborations
- Individual labs
- In-labs collection
- Poorly curated and preserved

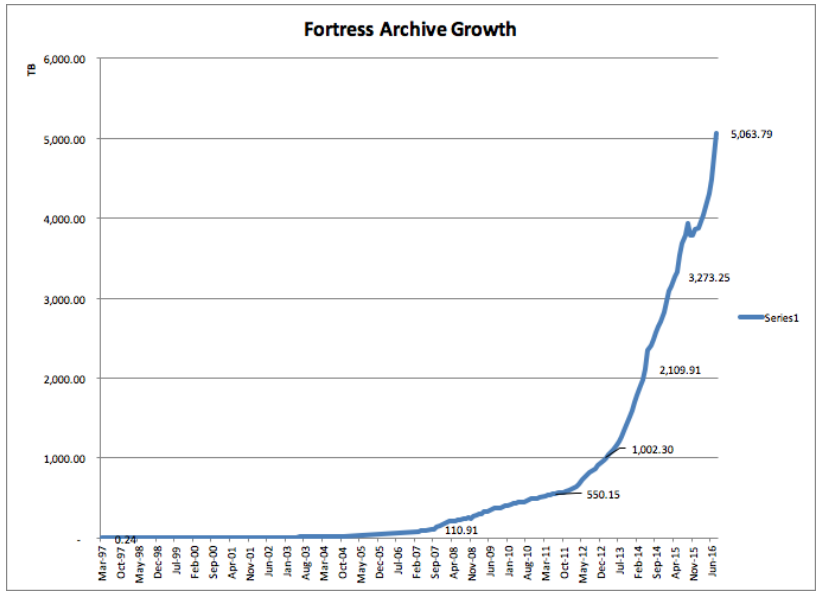


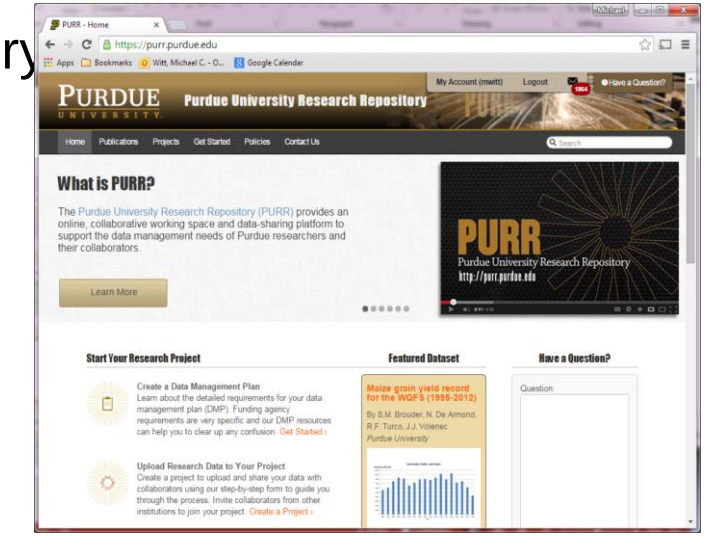
- Poor access – and visibility
- Short-term projects

Graphic by Bryan Heidorn, 2008, Shedding light on the dark data in the long tail of science.

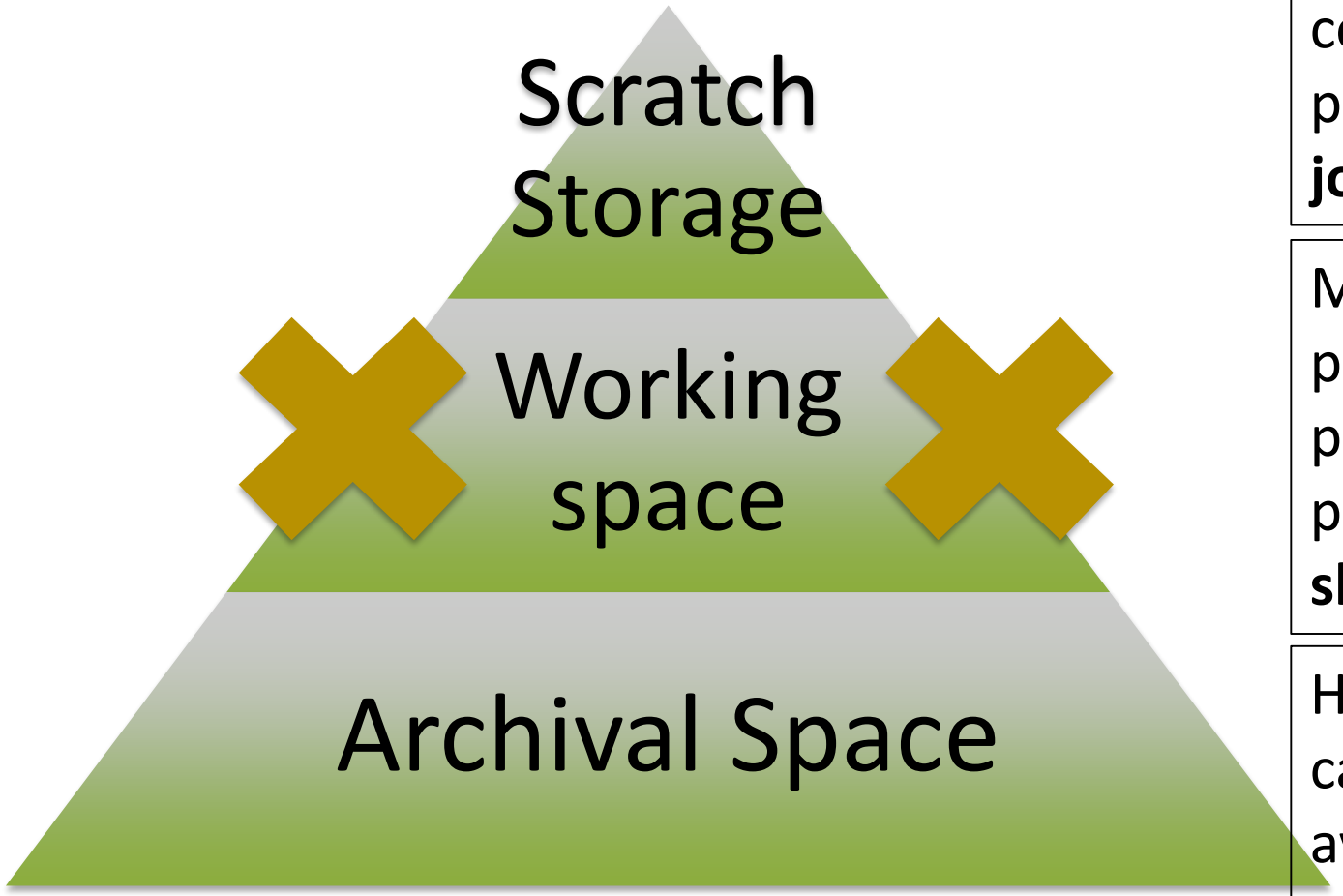
USER NEEDS AND ISSUES

- Fortress growth
  - Tape archive
  - Optimized for files >1GB
  - FTP access
  - Duplicated
- Scratch space
  - Temporary
  - Optimized for HPC
  - Had to install apps
  - Cannot be shared easily
- Purdue University Research Repository: an institutional repository
  - 100 GB per project with a grant
  - Optimized for data publication and preservation
  - Not appropriate for Big Data
- The long tail of science increasingly means Big Data
  - Very heterogeneous data (the Variety V of Big Data)
  - New problems increasingly require HPC resources
  - Data volumes also increase





BIG DATA: TIERS OF STORAGE



Fast, large, purged, coupled with clusters, per-user – **for running jobs**

Medium speed, large, persistent, data protected, purchased, per research lab – **for shared data and apps**

High speed, high capacity, well protected, available to all researchers – **for permanent storage**

A WELL-RECEIVED SOLUTION AT PURDUE: DATA DEPOT

- Approximately 2.25 PB of usable capacity
- Hardware provided by a pair of Data Direct Networks SFA12k arrays, one in each of MATH and FREH datacenters
- 160 Gb/sec to each datacenter
- 5x Dell R620 servers in each datacenter (replicated)

- **In just over a year, 280 research groups are participating**

  Many are not HPC users
- **0.75 PB used since 2014**
- **A research group purchasing space has purchased on average 8.6 TB**

RESOURCE: INVENTORY OF STORAGE OPTIONS

	Data Depot	PURR
Price	100 GB free	10 GB free, 100 GB free with grant
Available storage	No upper limit	Not available
Primary use	Storage and services, including data transfer, file structure, and tools ; group oriented	Project work space; Data publication; preservation; group oriented
Back-ups	Replicated across campus. Nightly snapshots to protect against accidental deletion	Nightly; 30 daily images
Access after you leave Purdue	Lose access. Project manager needs to be Purdue-affiliated	Lose access. Project manager needs to be Purdue-affiliated
Accessible from HPC	Directly mounted on HPC nodes Globus and other protocols to transfer data	Uses Globus to transfer data to HPC systems

Currently 7 options and 23 criteria

ROLES AROUND DATA

- Data reference questions (where to find standards)
- Reviewing/revising DMPs (providing input/suggestions)
- Data management planning (identifying metadata along lifecycle)
- Data consultation (may lead to collaborations/grants)
- Using repository (local, disciplinary)
- Promoting data DOIs
- Data information literacy (graduate students/labs)
- Finding and using data (e.g., using r3data.org
- Developing tools (e.g., Data Curation Profiles)
- Developing data resources (LibGuides, tutorials)
- Developing local data collections
- Promoting open access

A TOOL FOR WORKING WITH BIG DATA

Description of Data

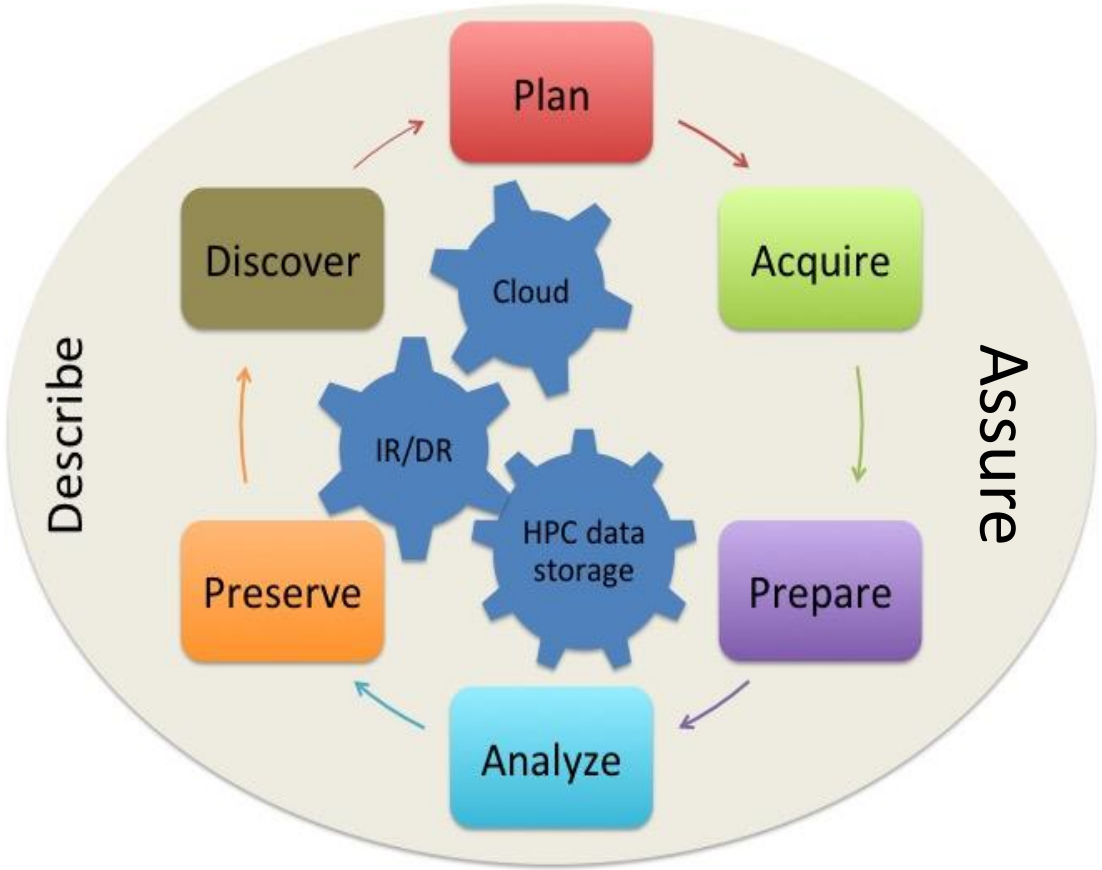
Data Formats

Documentation of Methodology

Description for Discovery

Metadata for Organization

Standards for Interoperability



Provenance for Preservation

Provenance for Reproducibility

Attribution/ citation

Intellectual Property Rights

Sharing & Access Policies

Software

Line Pouchard, 2015, "Revisiting the Data life cycle for Big Data curation," International Journal of Data Curation 10(2). doi:10.2218/ijdc.v10i2.342

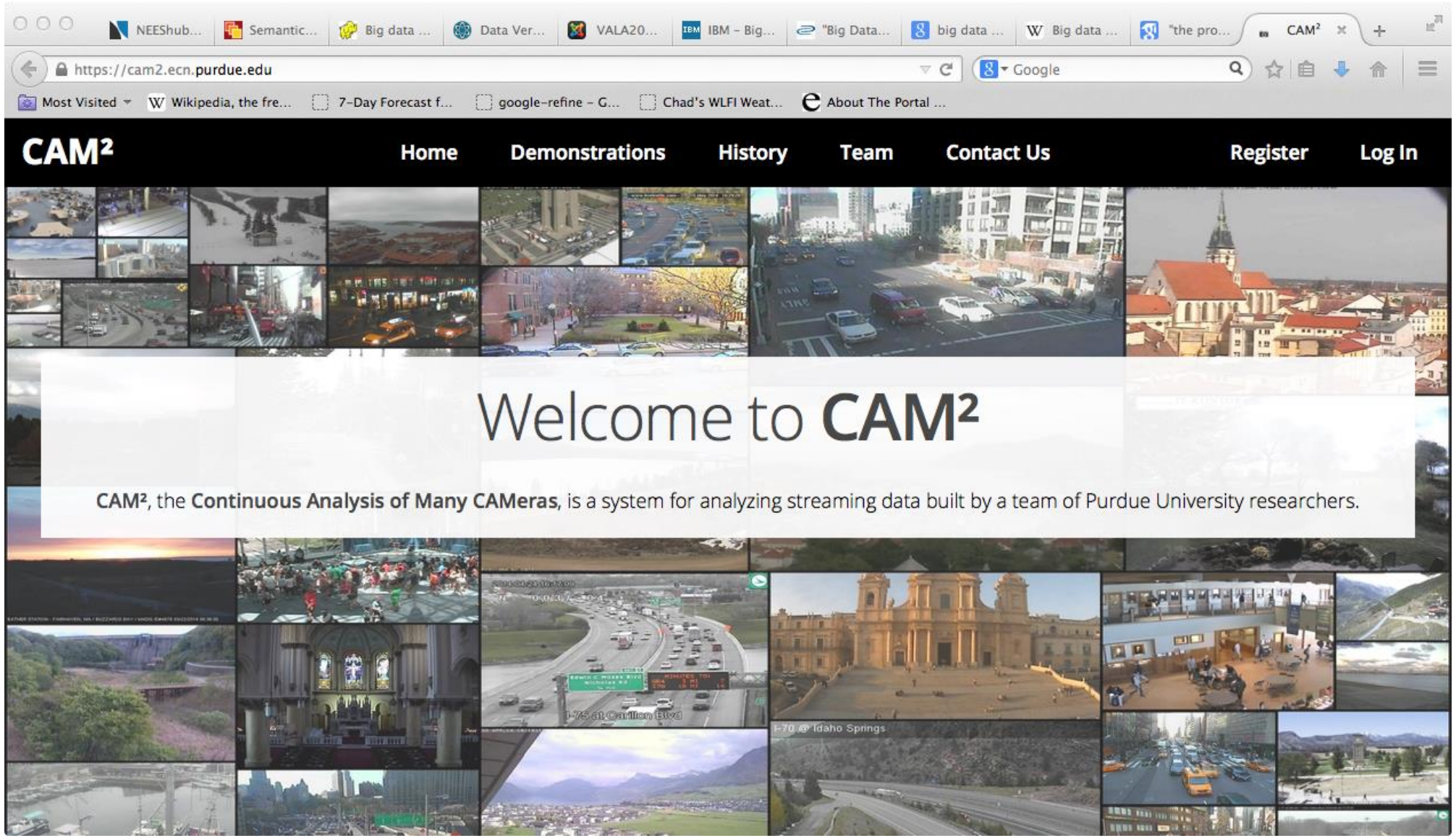
QUESTIONS INFORMING CURATION ACTIVITIES

	Plan	Acquire	Prepare
Volume	What is an estimate of volume & growth rate?	What is the most suited storage (databases, NoSQL, cloud)?	How do we prepare datasets for analysis? (remove blanks, duplicates, splitting columns, adding/removing headers)?
Variety	Are the data sensitive? What provisions are made to accommodate sensitive data?	What are the data formats and steps needed to integrate them?	What transformations are needed to aggregate data? Do we need to create a pipeline?
Velocity	Is bandwidth sufficient to accommodate input rates?	Will datasets be aggregated into series? Will metadata apply to individual datasets or to series?	What type of naming convention is needed to keep track of incoming and derived datasets?
Veracity	What are the data sources? What allows us to trust them?	Who collects the data? Do they have the tools and skills to ensure continuity?	Are the wrangling steps sufficiently documented to foster trust in the analysis?

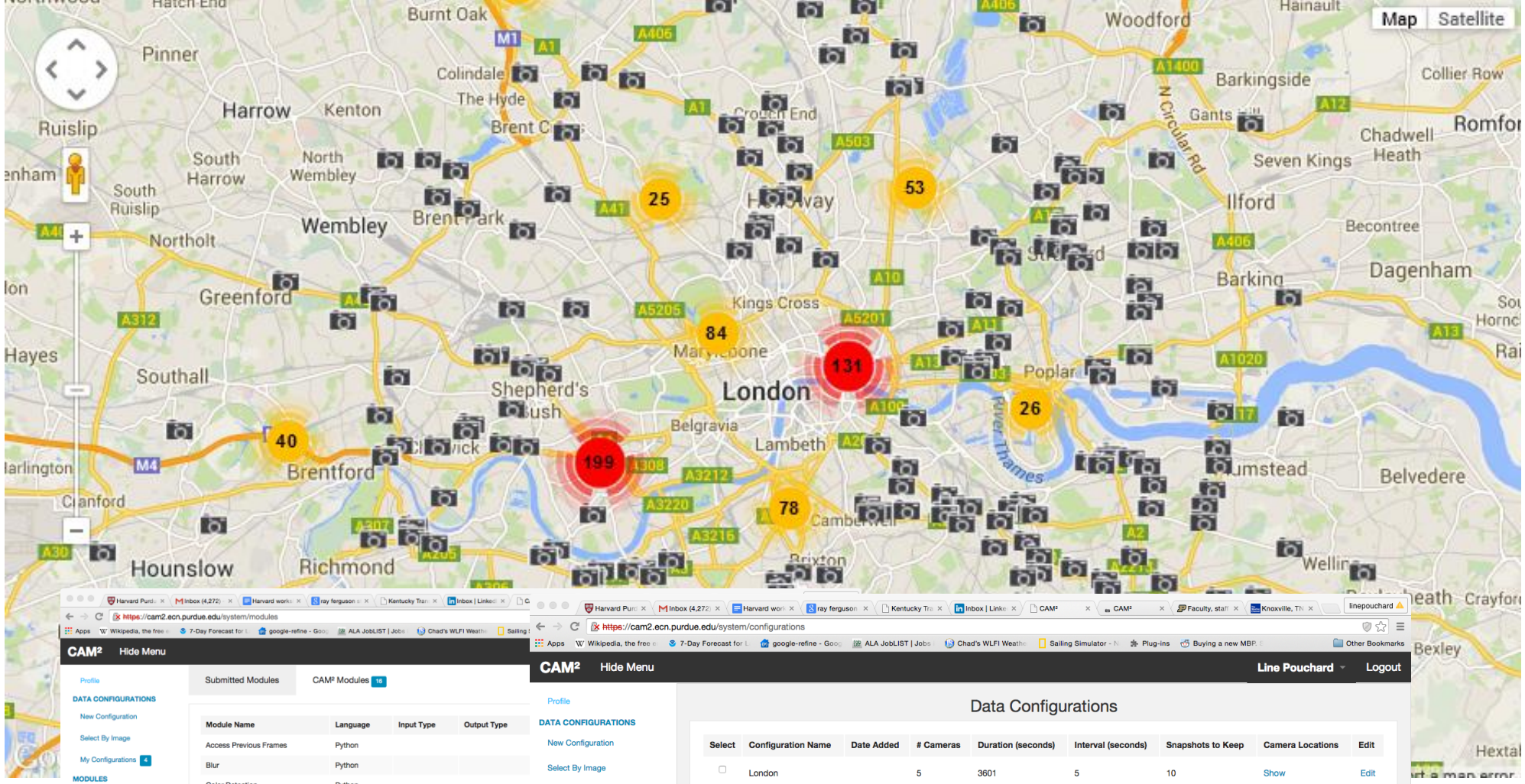
QUESTIONS INFORMING CURATION ACTIVITIES

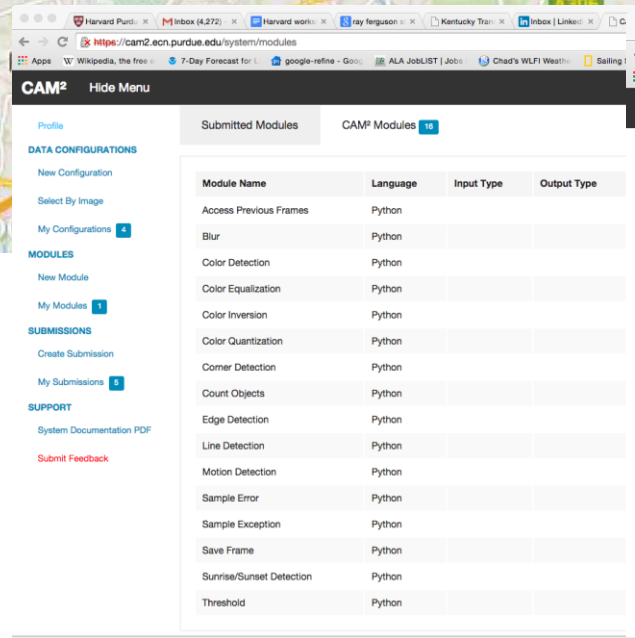
	Analyse	Preserve	Discover
Volume	Are adequate compute power and analysis methods available?	Should raw data be preserved? What storage space is needed in the long-term?	What part of the data (derived, raw, software code) will be made accessible to searches?
Variety	Are the various analytical methods compatible with the different datasets?	Are there different legal considerations for each data source? Are there conflicts with privacy and confidentiality?	What search methods best suit this data – keyword-based, geo-spatial searches, metadata-based, semantic searches?
Velocity	At what time point does the analytical feedback need to inform decisions?	When does data become obsolete?	What degree of search latency is tolerable?
Veracity	What kind of access to scripts, software, and procedures is needed to ensure transparency and reproducibility?	What are the trade-offs if only derived products and no raw data are preserved?	Providing well-documented data in open access allows scrutiny. How is veracity supported with sensitive and private data?

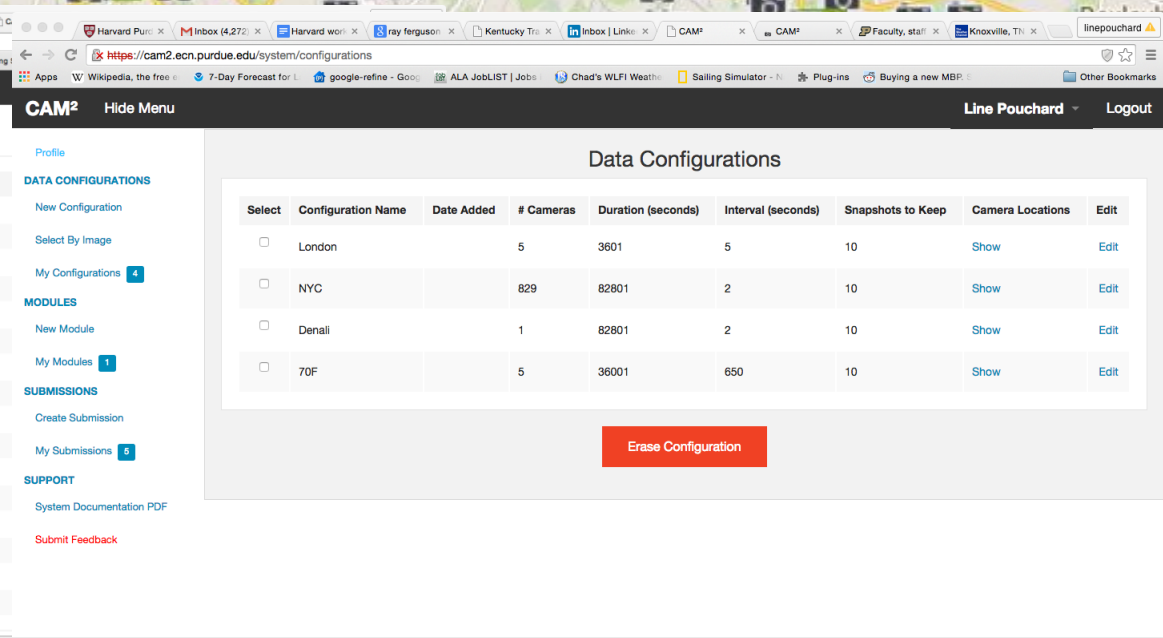
CAM2: A BIG DATA PROJECT AT PURDUE



With Dr. Yung-Hsiang Lu, PI, and Megan Sapp Nelson, Libraries


Map
Satellite
Ruislip
Pinner
Harrow
Kenton
Burnt Oak
Colindale
The Hyde
Brent Cross
Croydon End
Woodford
Barkingside
Gants Hill
Collier Row
Chadwell Heath
Seven Kings
Ilford
Romford
Becontree
Dagenham
Barking
Poplar
South Harrow
North Wembley
South Ruislip
Wembley
Brent Park
Northolt
Greenford
Hayes
Southall
Shepherd's Bush
Kings Cross
Marylebone
London
Belgravia
Lambeth
Brentford
Chiswick
Hounslow
Richmond
Camberwell
Brixton
Belvedere
Bexley
Crayford
Hextable
25
53
84
131
26
40
199
78
M1
M4
A1
A406
A503
A41
A10
A5205
A5201
A3212
A3220
A3216
A1400
A12
A13
A1020
A2
A312
A308
A307
A205
A100


https://cam2.ecn.purdue.edu/system/modules
CAM² Hide Menu
Profile
DATA CONFIGURATIONS
New Configuration
Select By Image
My Configurations 4
MODULES
New Module
My Modules 1
SUBMISSIONS
Create Submission
My Submissions 5
SUPPORT
System Documentation PDF
Submit Feedback
Submitted Modules
CAM² Modules 16
Module Name
Language
Input Type
Output Type
Access Previous Frames
Blur
Color Detection
Color Equalization
Color Inversion
Color Quantization
Corner Detection
Count Objects
Edge Detection
Line Detection
Motion Detection
Sample Error
Sample Exception
Save Frame
Sunrise/Sunset Detection
Threshold
Python


https://cam2.ecn.purdue.edu/system/configurations
CAM² Hide Menu
Line Pouchard
Logout
Profile
DATA CONFIGURATIONS
New Configuration
Select By Image
My Configurations 4
MODULES
New Module
My Modules 1
SUBMISSIONS
Create Submission
My Submissions 5
SUPPORT
System Documentation PDF
Submit Feedback
Data Configurations
Select
Configuration Name
Date Added
Cameras
Duration (seconds)
Interval (seconds)
Snapshots to Keep
Camera Locations
Edit
London 5 3601 5 10 Show Edit
NYC 829 82801 2 10 Show Edit
Denali 1 82801 2 10 Show Edit
70F 5 36001 650 10 Show Edit
Erase Configuration

THE US REGULATORY LANDSCAPE

- We were looking for sharing and re-use within the existing regulatory framework, and found nothing, so we looked at privacy
- Traditionally more concerned with protecting citizens from the government than regulating industry
- No overall data protection framework at the Federal level,
- Fair Information Practice principles (FTC) streamlined for online privacy

VARIOUS US PRIVACY ACTS

The Federal Trade Commission enforces	Fair Credit Reporting Act	Consumer Reporting Agencies must maintain accurate records and can forward records to anyone with a legitimate interest.	1970
Department of Justice	Privacy Act	Regulates the use of data by government agencies.	1974
The Federal Trade Commission enforces	Financial Modernization Act	Financial institutions must have and share a privacy policy by which customers can decline sharing their personal information with third parties.	1999
The Federal Communication Commission enforces	Cable Communications Policy Act	Cable companies are not allowed to collect or share personal information without individuals' consent.	1984
The Federal Communication Commission enforces	Video Privacy Protection Act	Video stores cannot disclose their customers' rental history.	1988
The Department of Health and Human Services	Health Insurance Portability and Accountability Act (HIPAA)	Protects patients' health information from being released to potential employers.	1996
State of California	Online Privacy Protection Act	One of the most comprehensive laws. Websites' privacy policies must be highly visible and customers must be informed of third party use of their data.	2003

THE INTERSECTION OF BIG DATA AND REGULATIONS

- Existing regulations were mostly written before Big Data came upon the scene
  - Regulations that exist may place unrealistic expectations
    - ✧For example, how do you apply the principle of notice and consent when data is re-used and aggregated?
    - ✧Our analysis will demonstrate some of the ways these policies are always not suited to BD
- Additional difficulties to enforce privacy with Big Data exist:
  - Due to buying, selling and aggregating data, enforcing privacy may be virtually impossible
  - The lack of a comprehensive framework makes it very difficult to address privacy and re-use with heterogeneous sources
  - BD has implications for how the policies are written

REF: Lane, J., Stodden, V., Bender, S., & Nissenbaum, H. (2014). *Privacy, Big Data, and the Public Good: Frameworks for Engagement*: Cambridge University Press.

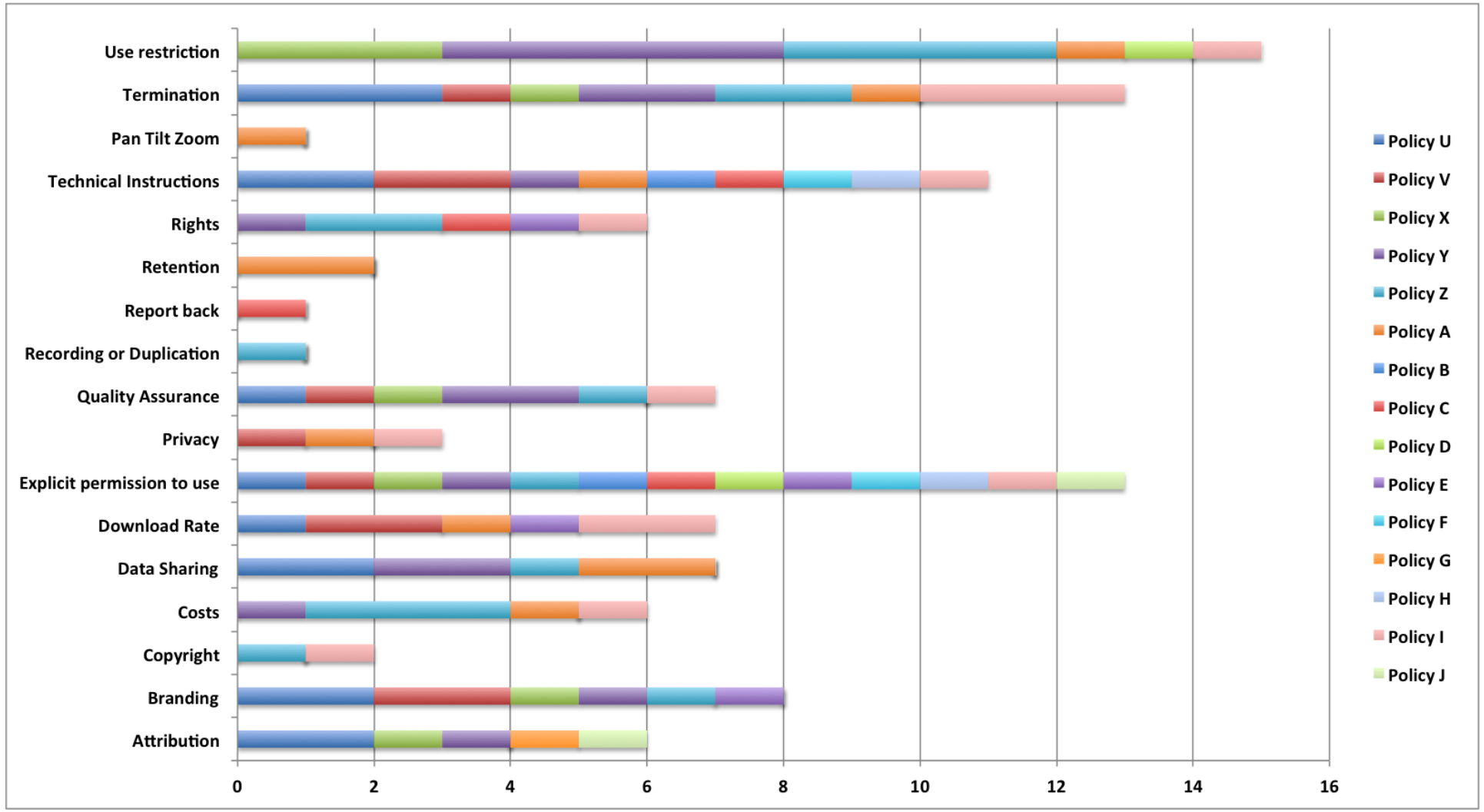
COMPARING POLICIES FOR RE-USE

- In video stream applications, data arrive at very high frequency. These applications exemplify the volume and velocity characteristics of Big Data
- Each data owner sets its own policies for using, sharing and re-using their data – the policies are different and there are different set of restrictions
- We analyze the terms that data owners use to articulate their policies and restrictions
- These terms have implications on re-use of the data for scientific research
- We also analyze the gaps that have implications for re-use

Line Pouchard, Megan Sapp Nelson, Yung-Hsiang Lu, Comparing policies for open data from publicly accessible international sources. IASSIST Quarterly, 29(4), 2015.

ANALYSIS OF POLICIES WITH NVIVO

Here is what the policies are talking about (10 ad hoc, 5 formal)



RESTRICTIONS ON SIZE & FRAME RATE

Examples of time limit or file size limit
A picture will not be captured more than once every two minutes
Allowed one picture per hour per camera
Allowed one 320 x 240 jpeg per second
No camera will be accessed more than once every five minutes
No more than a cumulative 24 hours of images that are no more than one week old.

http://mediacollege.com

A TEMPLATE FOR SHARING VIDEO CONTENT

- Data provider identification
- Download rate & file size
- Statement of re-use that allows for general scientific investigation
- A statement governing appropriate use of the data set regarding individual's privacy
- Quality Control
- Attribution
- Retention and preservation
- Accountability and report back

http://amnet.com.au

TAKE AWAY: BIG DATA INFRASTRUCTURE AND TOOLS

- The long tail of science increasingly associates with big data
- Curating Big Data cannot be done in the library alone
- We gave the example of a middle-tier storage capacity that serves both HPC and non-HPC users
- Characterizing Big Data with the 4 Vs (volume, variety, velocity, veracity) although high level helps determining potential issues for activities in the data life cycle.
- Policies are complex, confusing, contradictory, difficult to ascertain, and there is no existing, comprehensive regulatory framework in the US to provide guidance for data sharing

THANK YOU
QUESTIONS?