MOVING ON WITH THE MANAGEMENT OF BIG DATA

Line Pouchard, PhD

Purdue University Libraries

Heidi Imker, PhD

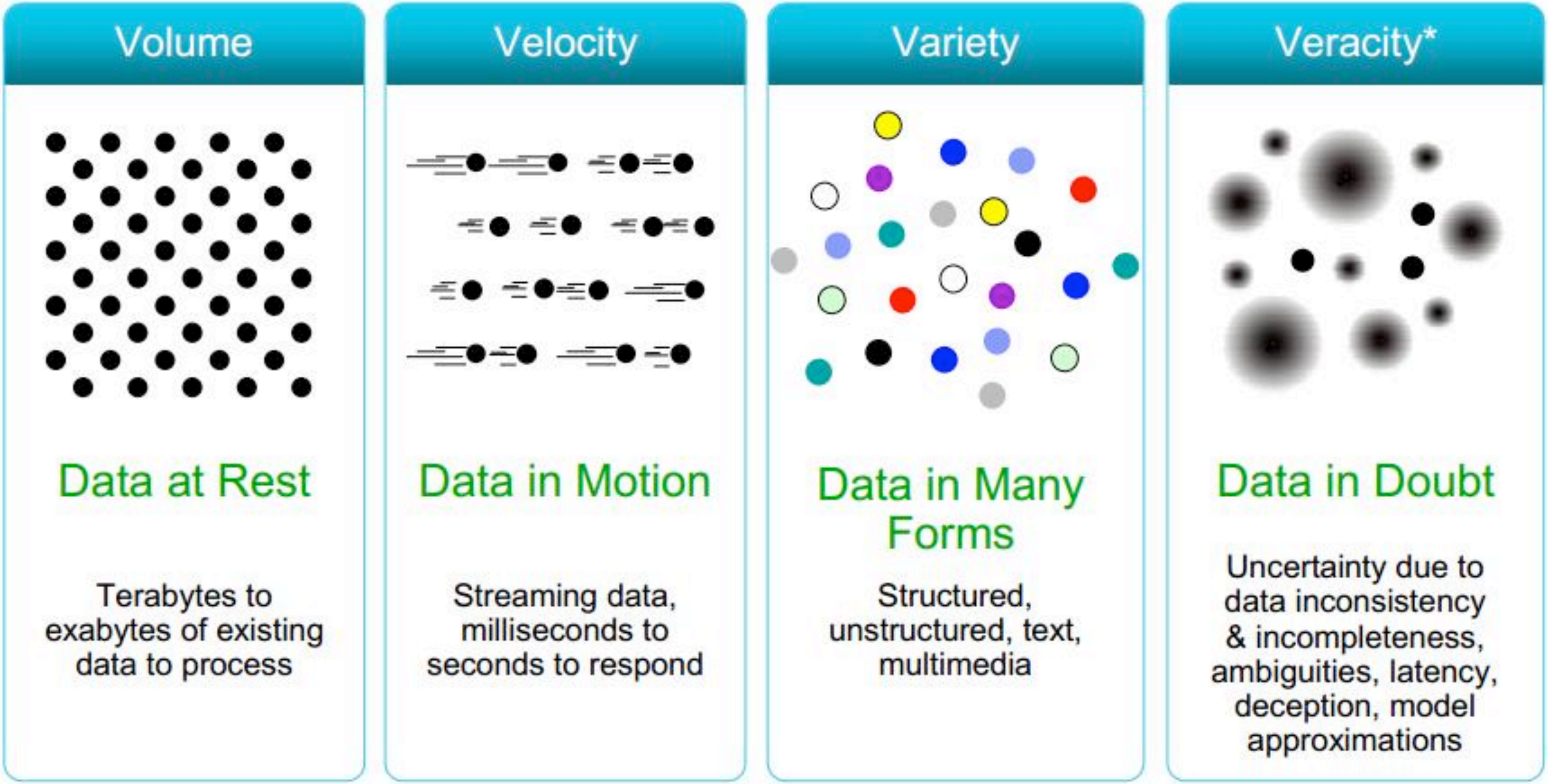
University Library, University of Illinois at Urbana-Champaign

Great Plains Network
DataFOUR Series
April 15, 2016

Outline

- The Vs of Big Data
- Big Data at UIUC
- Example of Big Data at Purdue: CAM2
  - Data Management Issues in CAM2
- Issues to consider in the choice of storage solutions
  - Storage at Purdue: Data Depot
  - Storage at UIUC
- Importance of collaborations for Big Data management
  - Collaboration with the IT and SC office at UIUC
  - Collaboration with ITAP at Purdue
- Preservation of Big Data

The Vs of Big Data



Volume

Data at Rest

Terabytes to exabytes of existing data to process

Velocity

Data in Motion

Streaming data, milliseconds to seconds to respond

Variety

Data in Many Forms

Structured, unstructured, text, multimedia

Veracity*

Data in Doubt

Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

http://www.datasciencecentral.com/profiles/blogs/data-veracity

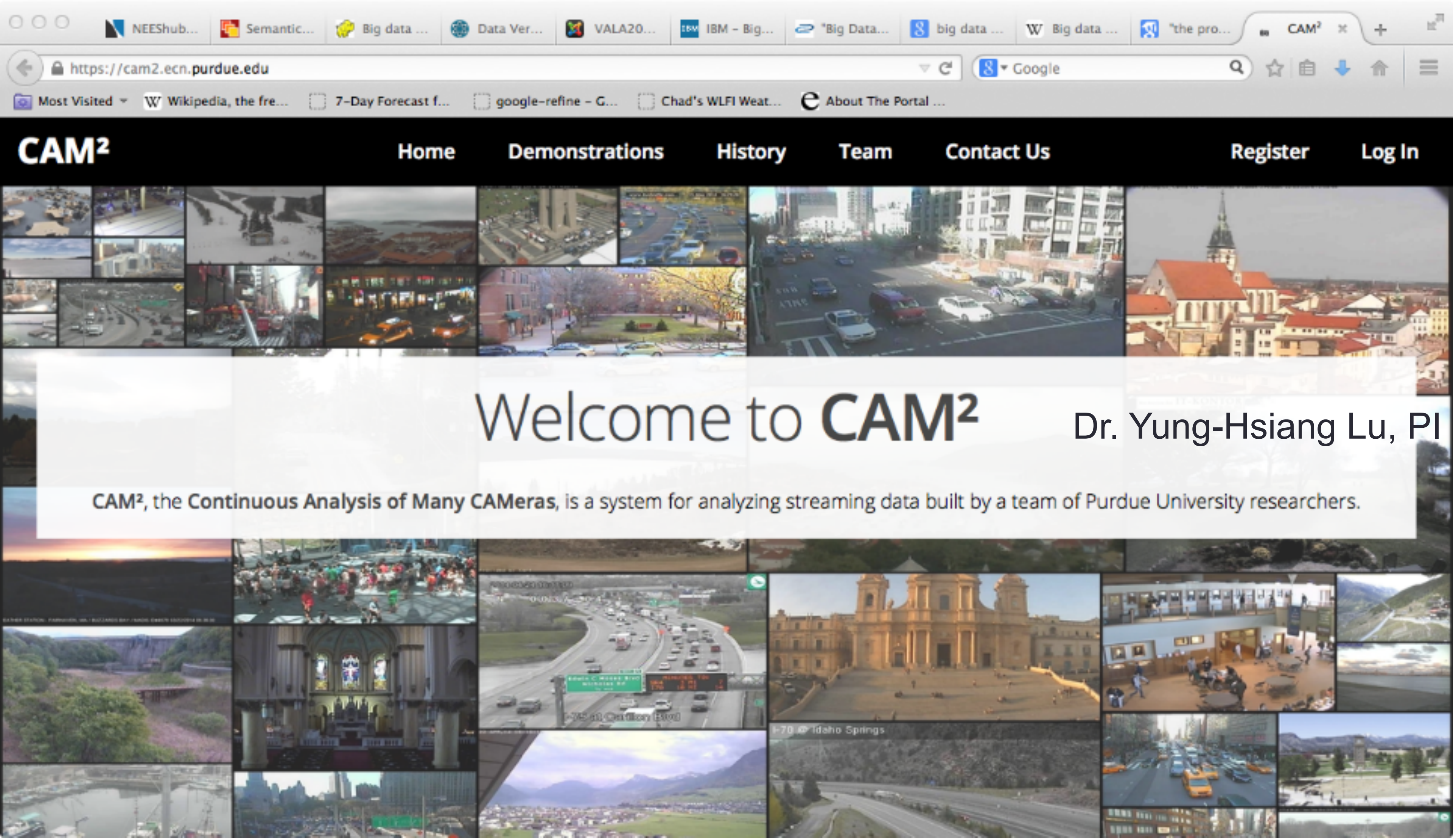
Illinois “Big Data”

Start a Research Data Service and people find you to try to give you their…

- 300 TB from a single Physics researcher
- 45 TB from a single Computational Biology study
- 3 TB from a single Animal Science Researcher
- 300 GB form a single Genomics paper

…. one Illinois imaging center has 800 TB of storage

(and our supercomputing center has >400 PB of storage)



Welcome to **CAM^2**

Dr. Yung-Hsiang Lu, PI

CAM^2, the **Continuous Analysis of Many CAMeras**, is a system for analyzing streaming data built by a team of Purdue University researchers.

Comparing policies for open data from publicly accessible international sources.
Line C. Pouchard, Megan Sapp Nelson, and Yung-Hsaign Lu
http://sites.lib.purdue.edu/linepouchard/publis/publis2014-15/pouchard04.pdf

Data management issues in CAM2

- Data access and re-use
  - policies of video streams and CCTV
  - Few policies available
- Who owns the data?
- Data storage
- Data organization
  - naming scheme
  - metadata database
- Protect metadata storage – where the intellectual property lies
- Data information literacy skills for graduate and undergraduate students in CAM2 lab
  - Learning modules and training customized to the lab

Storage issues

- Comparisons of available resources on campus are highly in demand – *and controversial*
- Different user populations have different needs
- Access to and permission levels
- Storage quotas and cost
- Data ownership and privacy issues varies
- To what extent does storage cover preservation?

Purdue Research Data Depot

- Approximately 2.25 PB of IBM GPFS
- 5x Dell R620 servers in each datacenter
- 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
- Accessible to labs/groups inside and outside Purdue
- All Purdue researchers eligible for 100GB free of charge
- Then, $150 per TB per year
- Data transfer: Globus Online

Illinois Storage Options

- Box, Google Drive, OneNote
  - Access is browser based
  - Super easy (& independent) user experience and access control
  - Roughly ≈ 1 TB or less
- File Servers
  - Access is network based (faster, but restricted)
  - Higher maintenance for IT people
  - Variable based on departmental resources
- Storage Services
  - Disk, Tape
  - Even higher maintenance
  - Cost recovery (so need $$$)

1 PB … ?

- Cheapest Storage @ Illinois = **$96/TB/year for single copy**
  - 1 PB = $96,000/1 year
  - Plus more if we want replication & a disaster recovery copy
  - Storage only (no access interface or preservation functions)

- Digital Preservation Network* = **$275/TB/year for multi-copy**
  - 1 TB = $5500/20 years
  - 1 PB = $5.5M/20 years
  - Storage and bit level preservation (but no access)

- Our conclusion is that we can't store (let alone preserve) literally **all** of it … so…

* http://dpn.org/wp-content/uploads/2015/06/DPN_FAQ.6.2.15.pdf

Tools

- Globus:
  - Secure file transfer from your machine to RDD
  - good for large and very large files
  - Web-based
  - Need Globus-Connect app for transfer to and from your personal machine
  - https://www.rcac.purdue.edu/storage/depot/guide/#storage_transfer_globus

- GitHub: www.github.com
  - Version control system
  - Mostly for code, sometimes used for data
  - Download Git and configure
  - Create a Repository – a project folder
    - Most basic unit
    - Can be public or private
    - Version control: contains all history of a file
    - Allows branching out and merging of code
  - Can be used for publication with a DOI
    - Transfer to Zenodo (EU)

Multi-Directional Collaborations

- Researchers
- The University
- Office of Research
- Central IT
- Office of Scholarly Publishing
- Sponsors
- Centers

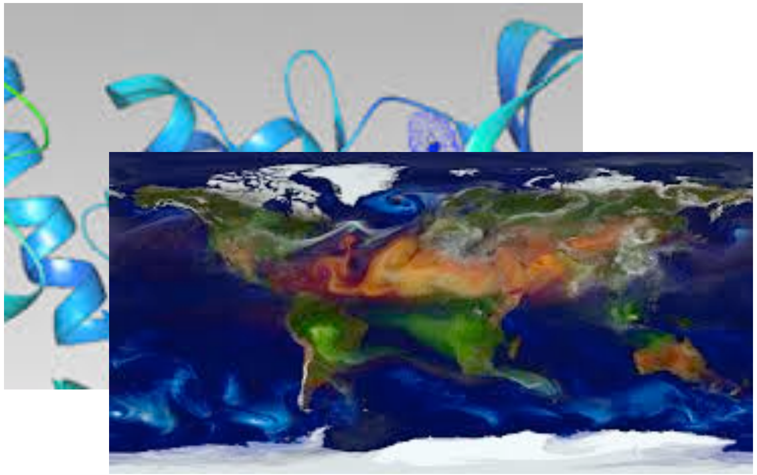
- Libraries

Illinois Collaborations w/ IT & SC Centers

- Work to provide better storage options to offer to campus
- Dataset “registration”
- Publishing sub-sets of data
- Provenance preservation

Collaborations with the IT office

- Develop common documentation about repositories, scratch space, storage, preservation platforms
- Take a common approach to advertising these resources
- Promote cross-training on data management, curation and sharing
- Work with Central IT to increase the discoverability of data



Illinois Preservation Options

- University Archives
- Records and Information Management Service
  - Administrative Records
  - Est 2006 (ish)
- IDEALS (ideals.illinois.edu)
  - Research and Scholarship
  - Est 2009(ish) (*note:* > 70K items and 19,000,000 downloads!)
  - 2 GB file size max
- Illinois Data Bank (databank.illinois.edu)
  - Research Data
  - Est May 2016
  - 15 GB max file size, but limit to 2 TB/faculty PI
  - Commit to 5 years minimum but our purview to assess long-term viability

12 Principles of Digital Preservation

British Library: http://bit.ly/1ZqCF7W

- ✓create and implement preservation plans
- ✓preserve metadata
- ✓preserve provenance
- ✓record any modifications
- ✓apply and document application of metadata standards
- ✓implement comprehensive end-to-end workflows
- ✓integrate curatorial assessments

12 Principles of Digital Preservation

? maintain file-level integrity

? monitor for emergent preservation risks

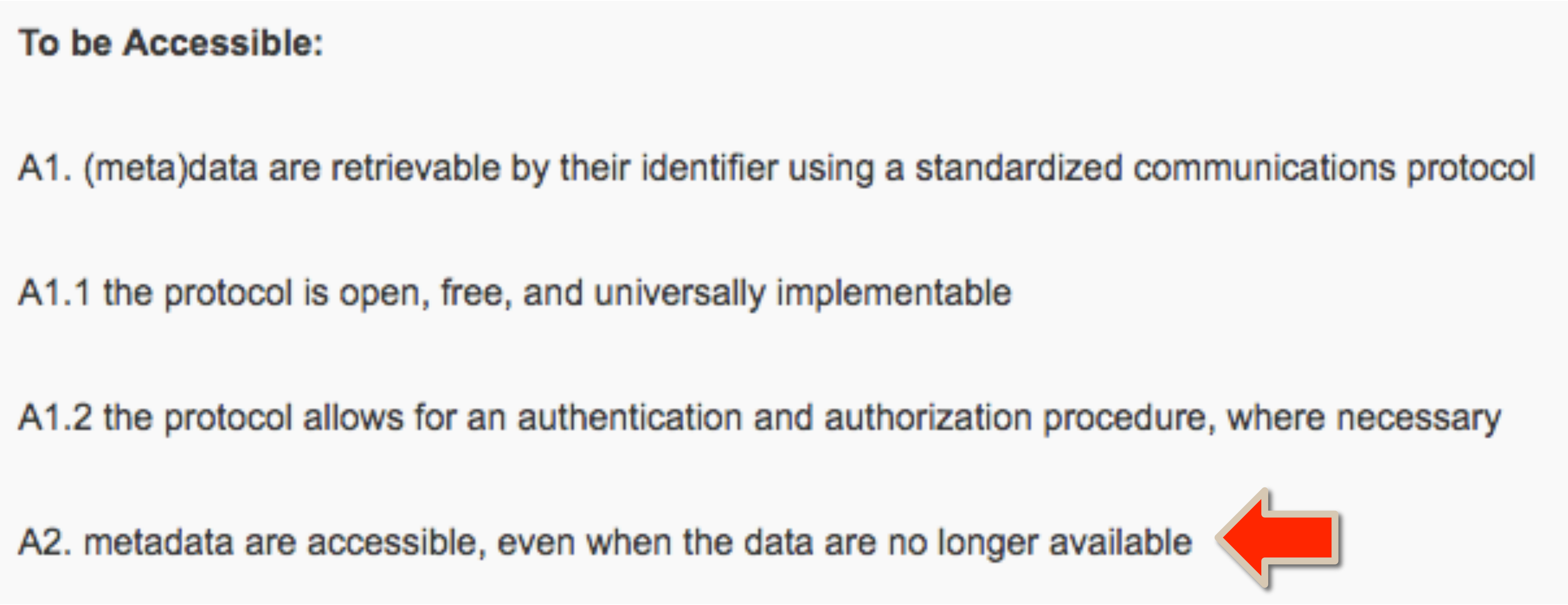
? integrate quality assurance checks

? preserve original files in our long term repository

? maintain Preservation Master copies

- Are the "can dos" meaningful if the actual files cannot be maintained "long-term"?
- If yes, how do we provide even “short-term” access ?

The FAIR Guiding Principles



To be Accessible:

A1. (meta)data are retrievable by their identifier using a standardized communications protocol

A1.1 the protocol is open, free, and universally implementable

A1.2 the protocol allows for an authentication and authorization procedure, where necessary

A2. metadata are accessible, even when the data are no longer available

http://www.nature.com/articles/sdata201618

Conclusions

- Big Data and Small Data management look similar at a high level
  - very different in the details
  - large variety of projects requires lots of customization
- Planning for management and preservation from the beginning of a project is crucial
- Less room for ad hoc solutions and grey areas
- Collaborations are more important
- Facilitating access is where efforts need to focus, not storing the data
- Demonstration need to access will help drive preservation