DATA TO INSIGHT CENTER
INDIANA UNIVERSITY
Pervasive Technology Institute

HathiTrust Data Capsules : Text Mining 4.6 Billion Pages

Beth Plale
Professor of Informatics and Computing
Director, Data To Insight Center, Indiana University
@bplale

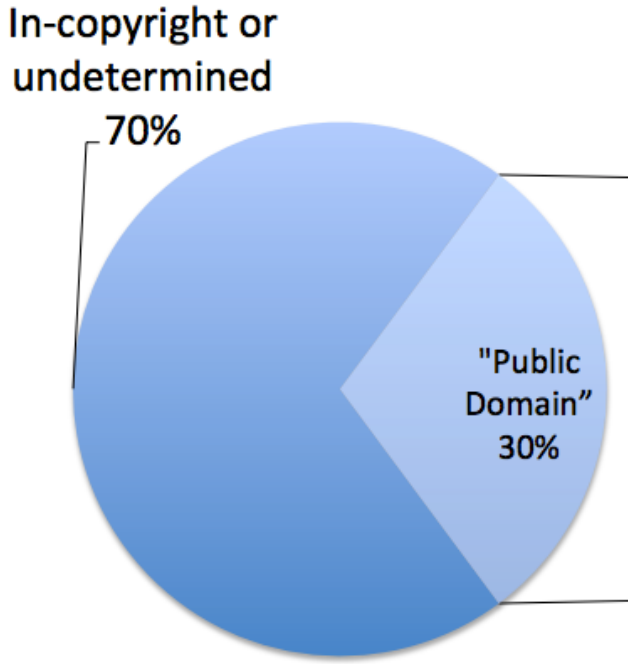
RDA RESEARCH DATA ALLIANCE

SEAD Sustainable Environment Actionable Data

RESEARCH CENTER

4.6 Billion Pages

- HathiTrust digital repository
  - 13.3 million books and growing
- Digitization sources
  - Google (96.8%)
  - Internet Archive (2.9%)
  - Local (0.3%)



HathiTrust Research Center

- Purpose: Enable researchers world-wide to carry out text mining analysis of the HathiTrust repository
- Since 70% of the digitized content under copyright, compute has to be co-located at the data. Data are “pinned” to location.
- Software services called: Secure HathiTrust Analysis Research Commons (SHARC)

Starting principles

Digital nature of content and restrictions of copyright precludes a model of checking out a set of (digital) volumes/books, taking them home, and promising to return them, or destroy them, later so

computation moves to the data (not vice versa)

How support access and computational analysis a way that doesn't constrain the research (in a non-work-limiting way)?

Research Questions

- **Non-consumptive use***: can framework provide safe handling of large amounts of protected data?
- **Openness**: can framework support user-contributed analysis without resorting to code walkthroughs prior to acceptance?
- **Large-scale** and **low cost**: can protections be extended to utilization of large-scale national (public) computational resources?

*Non-consumptive use is defined as *computational analysis of the copyrighted content that is carried out in such a way that human consumption of texts is prohibited.*

HathiTrust Data Capsule

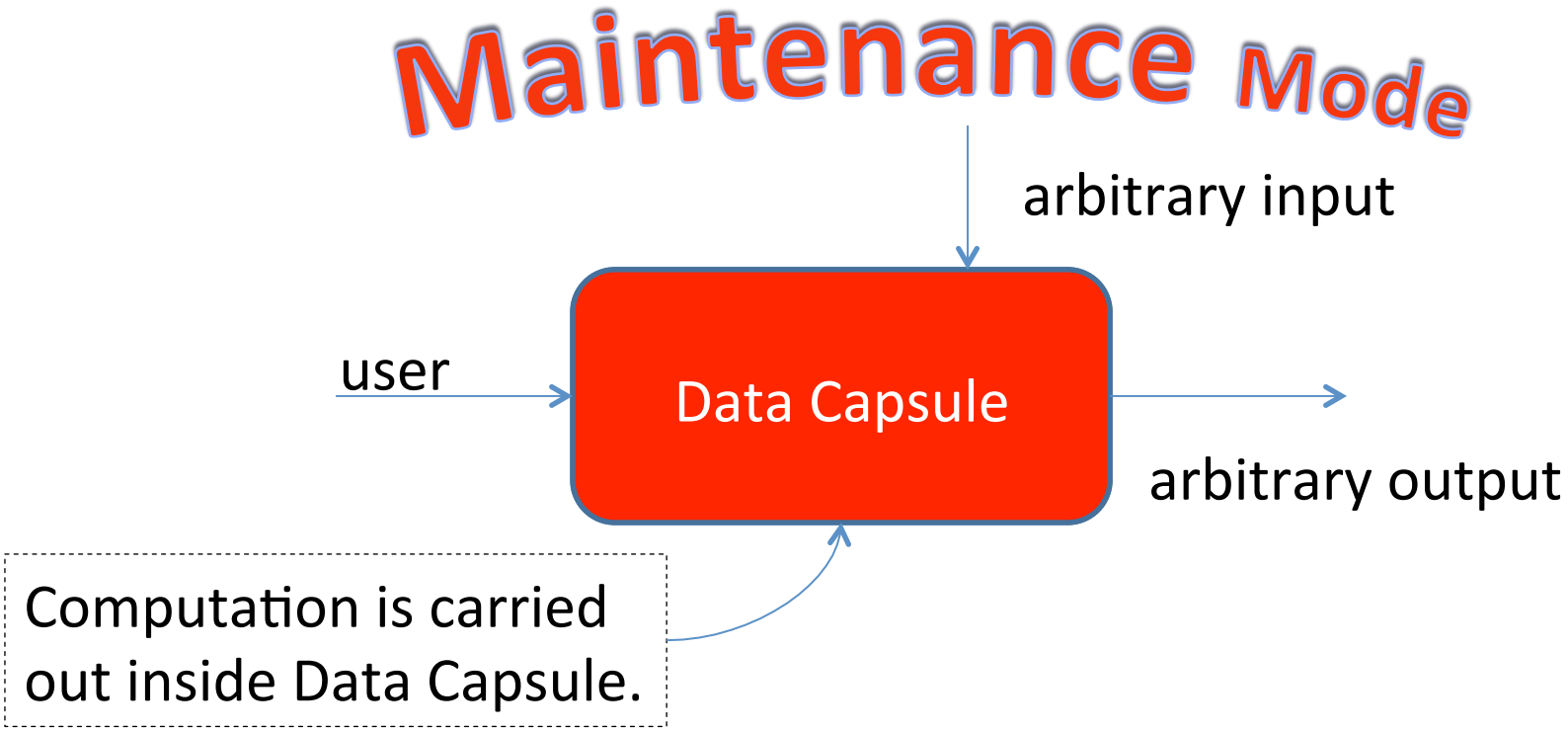
- User “checks out” a virtual machine (VM)
- VM runs in, and only in, the HTRC Services Environment*
- User owns their VM through weeks/months of analysis
- Getting stuff into VM is easy, but there is a controlled and audited process for getting results out of the VM

* A “local cloud”

Assumptions

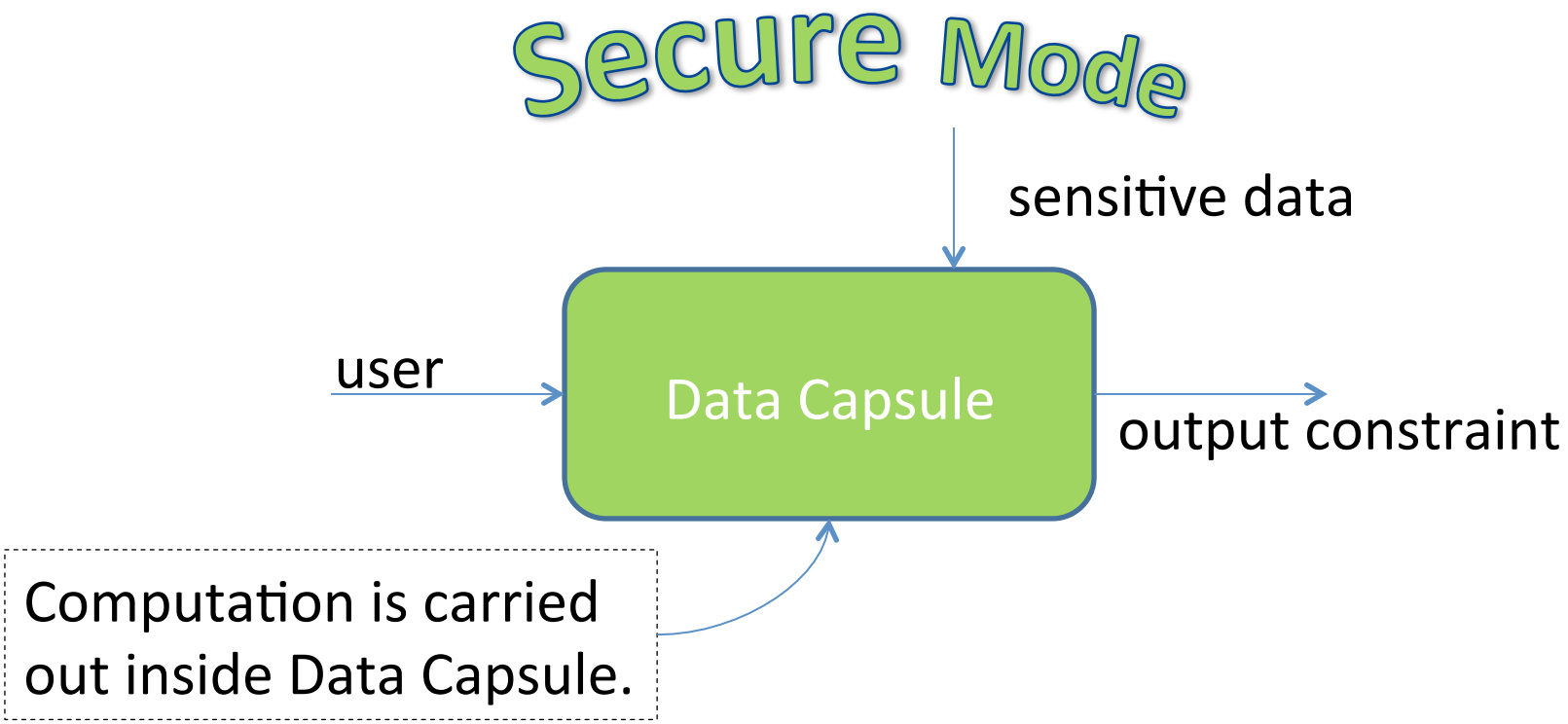
- Phase I (current phase) assumptions
  - User is trusted to not be malicious (trust users)
  - User's programs are malicious (don't trust programs)
  - A user cannot share their VM with someone else
  - A user can bring any alternate data into the VM to enrich analysis on the HT corpus, but the focus is the HT corpus (HT corpus as center of universe assumption)

HathiTrust Data Capsule Details

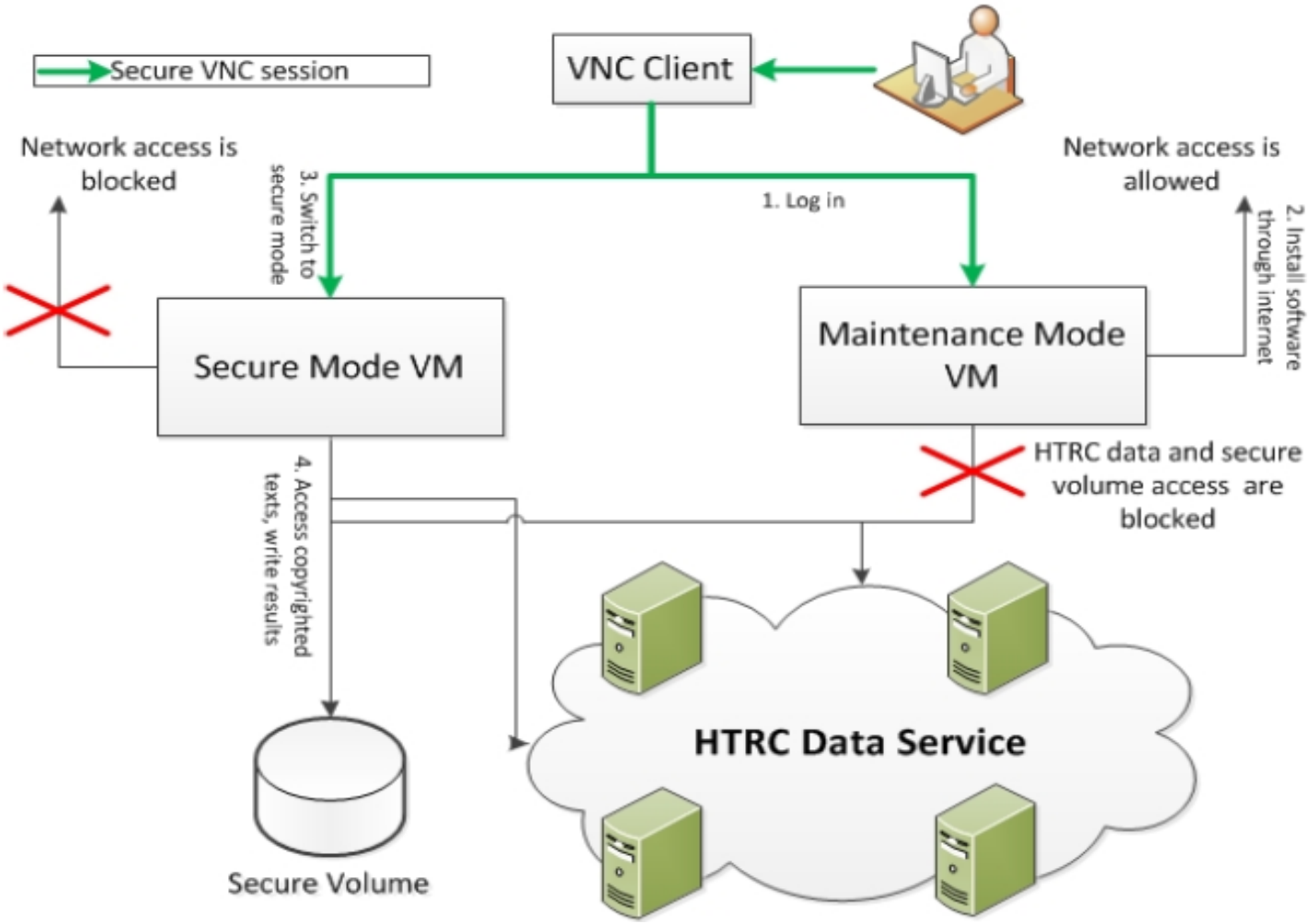


Foundations of HT Data Capsule:
K. Borders, E. V. Weele, B. Lau, and A. Prakash. Protecting confidential data on personal computers with storage capsules. *18th USENIX Security Symposium*, pp 367–382. USENIX Association, 2009.

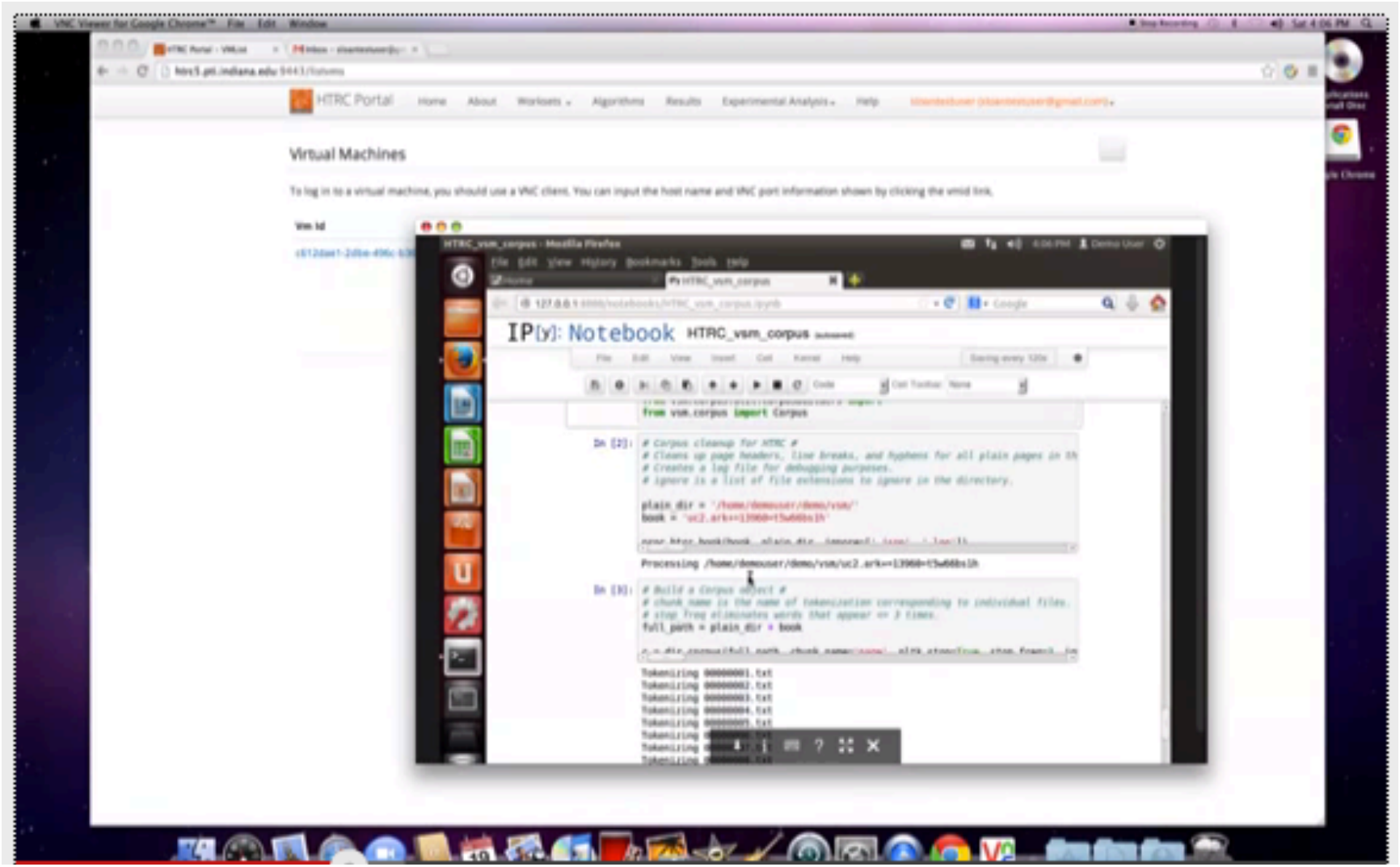
HathiTrust Data Capsule Details



HTRC Data Capsule Access



HT Data Capsule: view from researcher's couch



https://herstory.iu.edu/plale-beth.html

INDIANA UNIVERSITY

Beth Plale

Share Pin This Tweet Email

Informatics professor is at the forefront of innovations in data research

At technology conferences, Beth Plale first counts all of the women in the room (an admittedly quick task). Then, she introduces herself to every one of them.

Why tell this story and why now?

- Full professor at Indiana University, major research university in US
- Be inspiration to students who may be facing hurdles of their own to achieving their dreams