HATHI TRUST RESEARCH CENTER

HathiTrust Research Center Data Capsule v1.0: An Overview of Functionality

IU Libraries' Digital Library Brownbag Series | 09.10.14

Beth Plale | Jiaan Zeng | Robert H. McDonald | Miao Chen
Data To Insight Center
Indiana University

INDIANA UNIVERSITY

ILLINOIS

Tweet us - @HathiTrust #HTRC

Many thanks …

HTRC Data Capsule@IU Team

- Beth Plale (PI)
- Jiaan Zeng
- Guangchen Ruan

Special Thanks to

- Samitha Liyanage
- Milinda Pathirage
- Zong Peng
- Earlence Fernandes
- Ajit Aluri

HTRC Data Capsule@Michigan Team

- Atul Prakash (PI)
- Alexander Crowell

Jiaan Zeng, Guangchen Ruan, Alexander Crowell, Atul Prakash, and Beth Plale. 2014. Cloud computing data capsules for non-consumptiveuse of texts. In *Proceedings of the 5th ACM workshop on Scientific cloud computing* (ScienceCloud '14). ACM, New York, NY, USA, 9-16. DOI=10.1145/2608029.2608031 http://doi.acm.org/10.1145/2608029.2608031

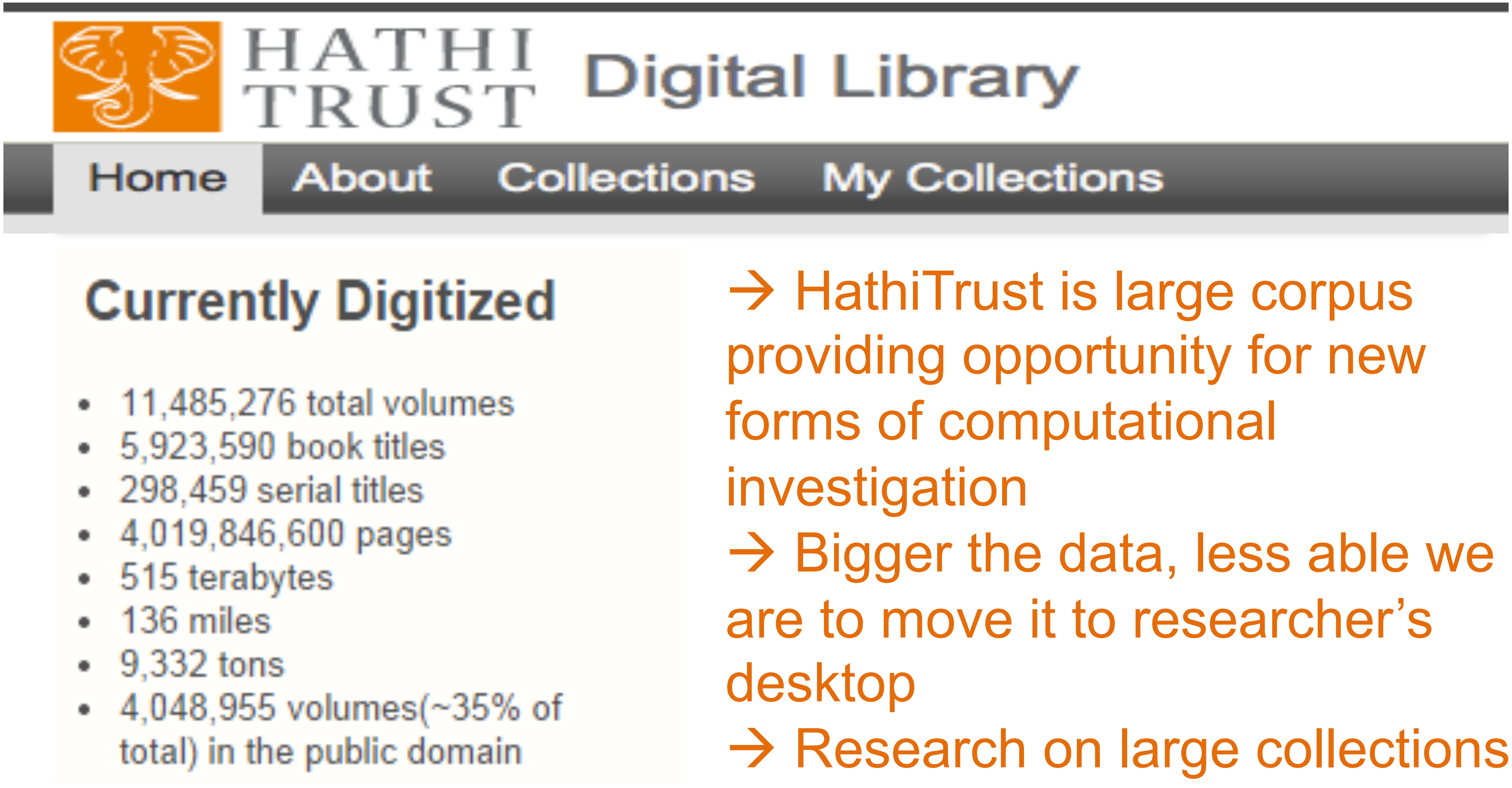
ALFRED P. SLOAN FOUNDATION
INDIANA UNIVERSITY
ILLINOIS
UNIVERSITY OF MICHIGAN
The Andrew W. Mellon Foundation

HathiTrust Digital Library

- HathiTrust is a partnership of academic & research institutions, offering a collection of millions of titles digitized from libraries around the world.
  - IU is a founding member of the HathiTrust along with University of Michigan, University of California, and the University of Virginia.

http://www.hathitrust.org

http://www.hathitrust.org/htrc



Currently Digitized

- 11,485,276 total volumes
- 5,923,590 book titles
- 298,459 serial titles
- 4,019,846,600 pages
- 515 terabytes
- 136 miles
- 9,332 tons
- 4,048,955 volumes(~35% of total) in the public domain

→ HathiTrust is large corpus providing opportunity for new forms of computational investigation

→ Bigger the data, less able we are to move it to researcher's desktop

→ Research on large collections requires

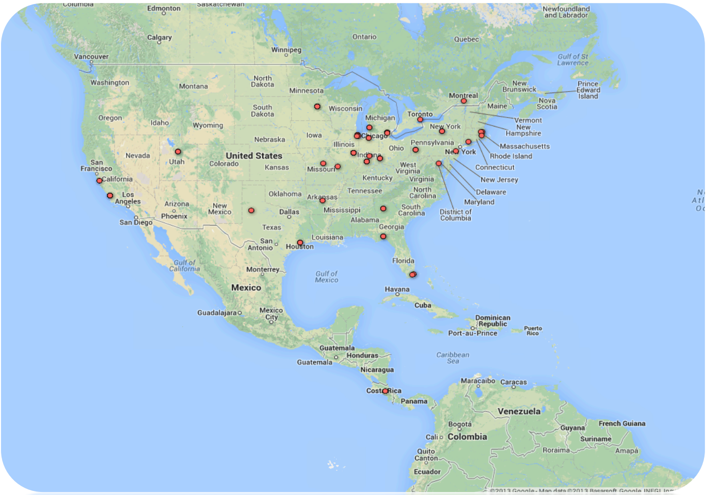
computation moves to data, not data to computation

Mission of the HT Research Center

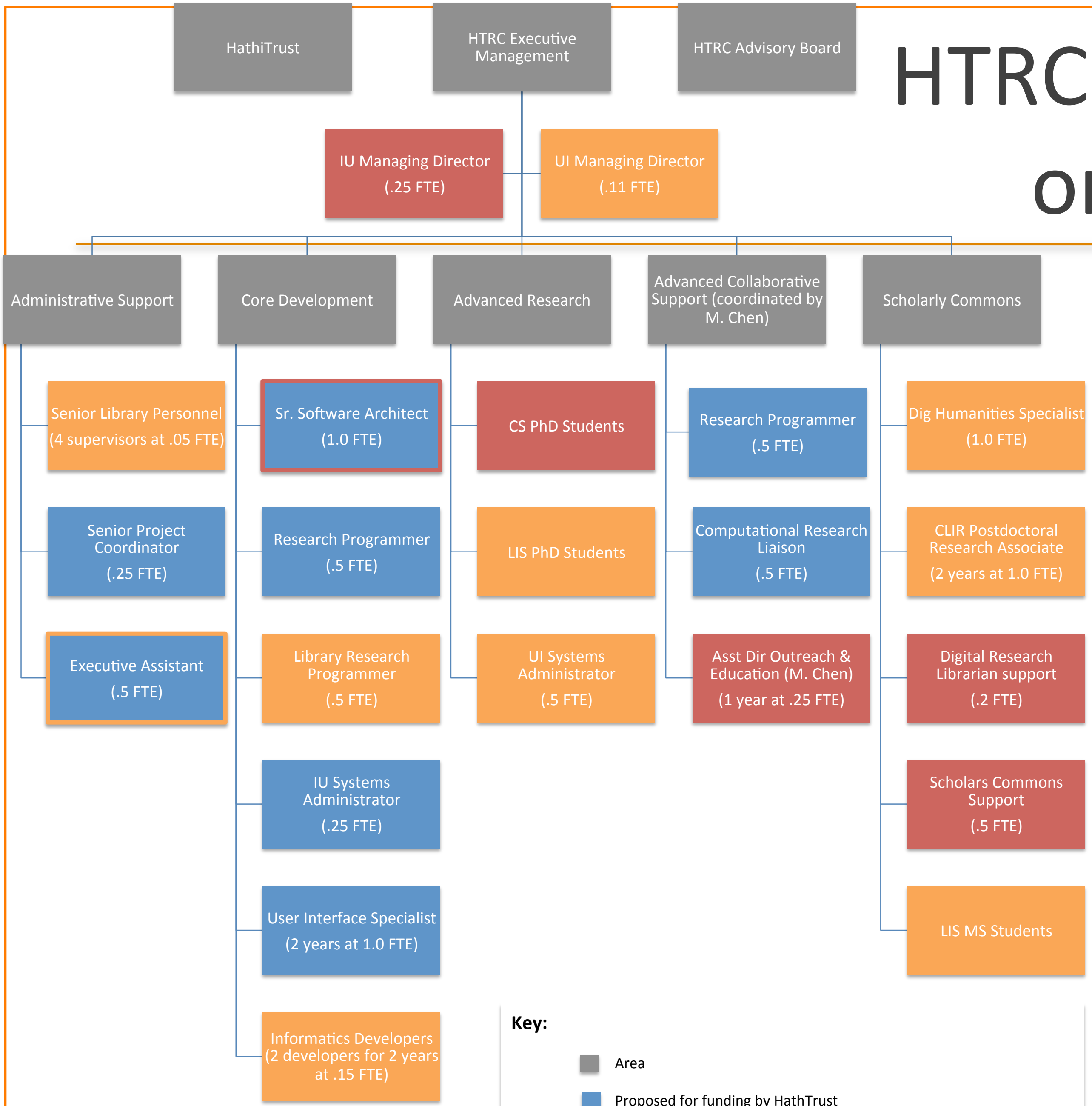
- Public research arm of HathiTrust
- Goal: enable researchers world-wide to accomplish tera-scale text data-mining and analysis
  - Develop cutting-edge software tools for processing, analyzing text
  - Develop cyberinfrastructure to enable HPC access to the HathiTrust Digital Library
- Established: July, 2011
- Collaborative center: Indiana University & University of Illinois

HTRC Timeline

- Phase I: development 01 Jul 2011 – 31 Mar 2013
  - HTRC software and services release v1.0 https://github.com/htrc
- Phase II: outreach, 01 Apr 2013 – 30 June 2014
  - 2nd HTRC UnCamp Sep ’13
- Phase III: operations, 01 July 2014 - present



HTRC 2014-2018 org chart



- ⬛ HathiTrust
- ⬛ HTRC Executive Management
- ⬛ HTRC Advisory Board
  - 🟥 IU Managing Director (.25 FTE)
  - 🟧 UI Managing Director (.11 FTE)
    - ⬛ Administrative Support
      - 🟧 Senior Library Personnel (4 supervisors at .05 FTE)
      - 🟦 Senior Project Coordinator (.25 FTE)
      - 🟦 Executive Assistant (.5 FTE)
    - ⬛ Core Development
      - 🟦 Sr. Software Architect (1.0 FTE)
      - 🟦 Research Programmer (.5 FTE)
      - 🟧 Library Research Programmer (.5 FTE)
      - 🟦 IU Systems Administrator (.25 FTE)
      - 🟦 User Interface Specialist (2 years at 1.0 FTE)
      - 🟧 Informatics Developers (2 developers for 2 years at .15 FTE)
    - ⬛ Advanced Research
      - 🟥 CS PhD Students
      - 🟧 LIS PhD Students
      - 🟧 UI Systems Administrator (.5 FTE)
    - ⬛ Advanced Collaborative Support (coordinated by M. Chen)
      - 🟦 Research Programmer (.5 FTE)
      - 🟦 Computational Research Liaison (.5 FTE)
      - 🟥 Asst Dir Outreach & Education (M. Chen) (1 year at .25 FTE)
    - ⬛ Scholarly Commons
      - 🟧 Dig Humanities Specialist (1.0 FTE)
      - 🟧 CLIR Postdoctoral Research Associate (2 years at 1.0 FTE)
      - 🟥 Digital Research Librarian support (.2 FTE)
      - 🟥 Scholars Commons Support (.5 FTE)
      - 🟧 LIS MS Students

🟦 Proposed for funding by HathTrust

🟥 Funded by Indiana University

🟧 Funded by University of Illinois

🟦 Proposed for joint funding by HathiTrust

🟦 Proposed for joint funding by HathiTrust

Key:

⬛ Area

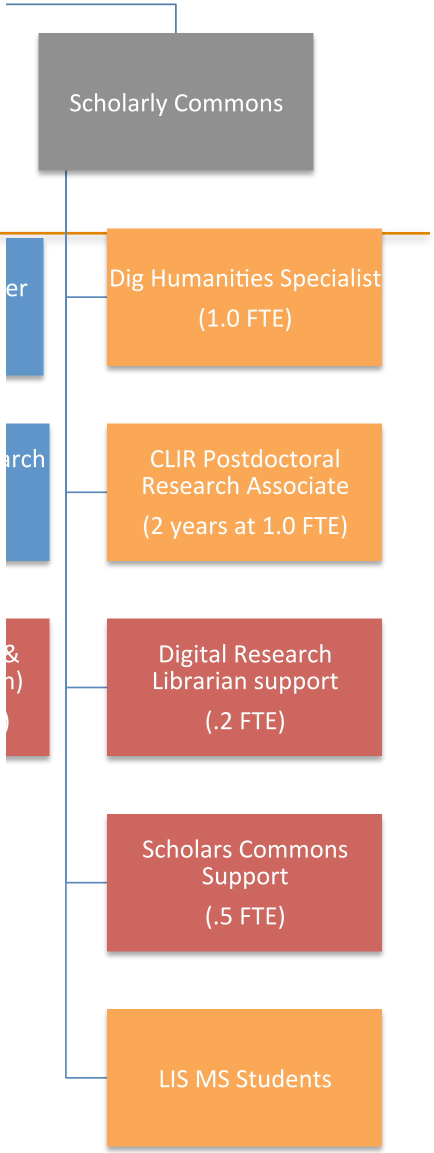
🟦 Proposed for funding by HathTrust

🟥 Funded by Indiana University

🟧 Funded by University of Illinois

Scholarly Commons User Support Service

- Develop training materials
- Educational workshops
- Tool and workset creation
- Collaborate with librarians and DH centers at HT institutions
- Assist researchers in HTRC text data mining research projects
- Led out of University of Illinois Library; smaller group at IU
- Resourced at 2.7 FTE.



Scholarly Commons

Dig Humanities Specialist
(1.0 FTE)

CLIR Postdoctoral Research Associate
(2 years at 1.0 FTE)

Digital Research Librarian support
(.2 FTE)

Scholars Commons Support
(.5 FTE)

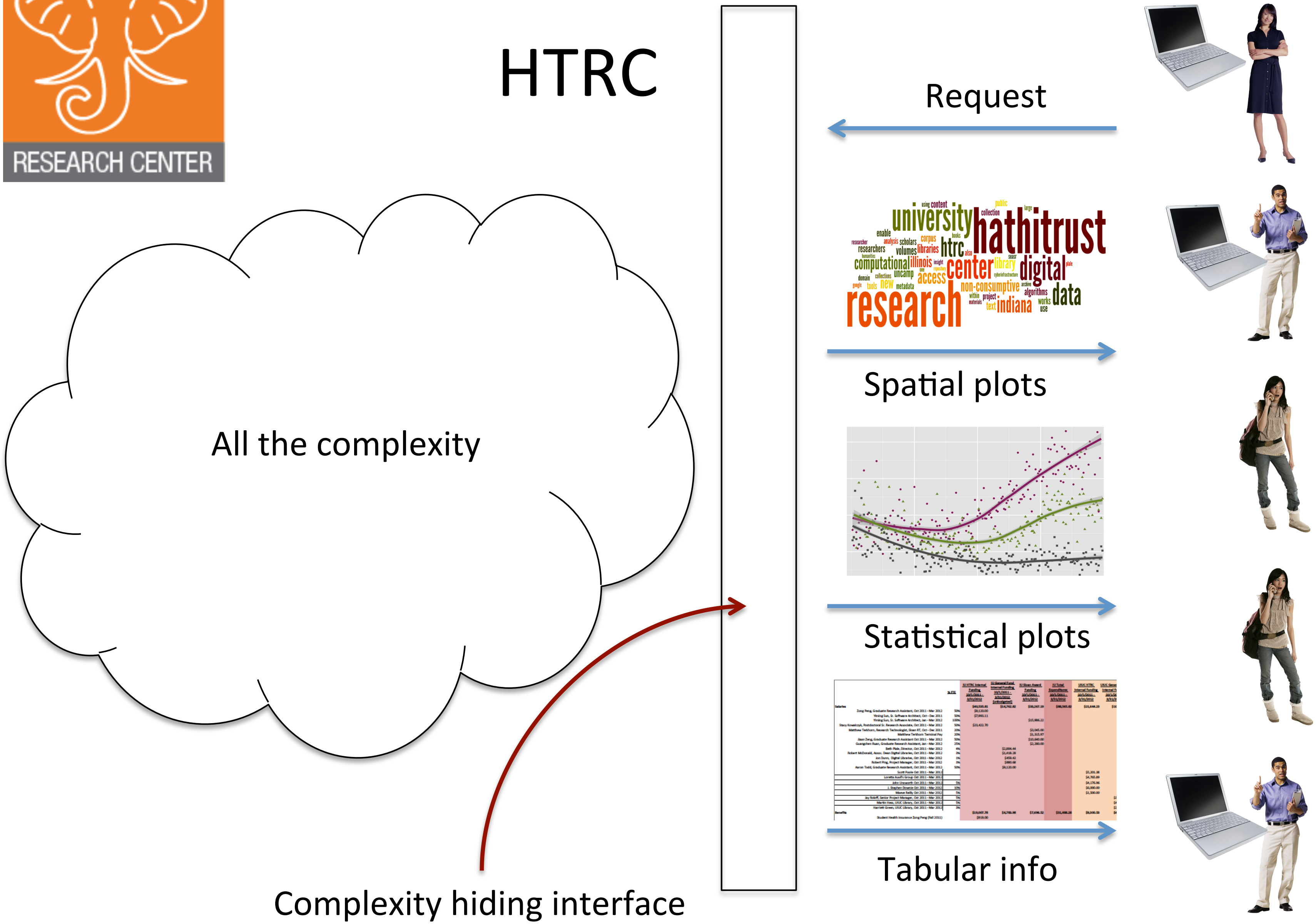
LIS MS Students

Non-Consumptive Research Paradigm

- *No action or set of actions on part of users, either acting alone or in cooperation with other users over duration of one or multiple sessions can result in sufficient information gathered from collection of copyrighted works to reassemble pages from collection.*
- Definition disallows collusion between users, or accumulation of material over time. Differentiates human researcher from proxy which is not a user. Users are human beings.

RESEARCH CENTER

HTRC



All the complexity

Request

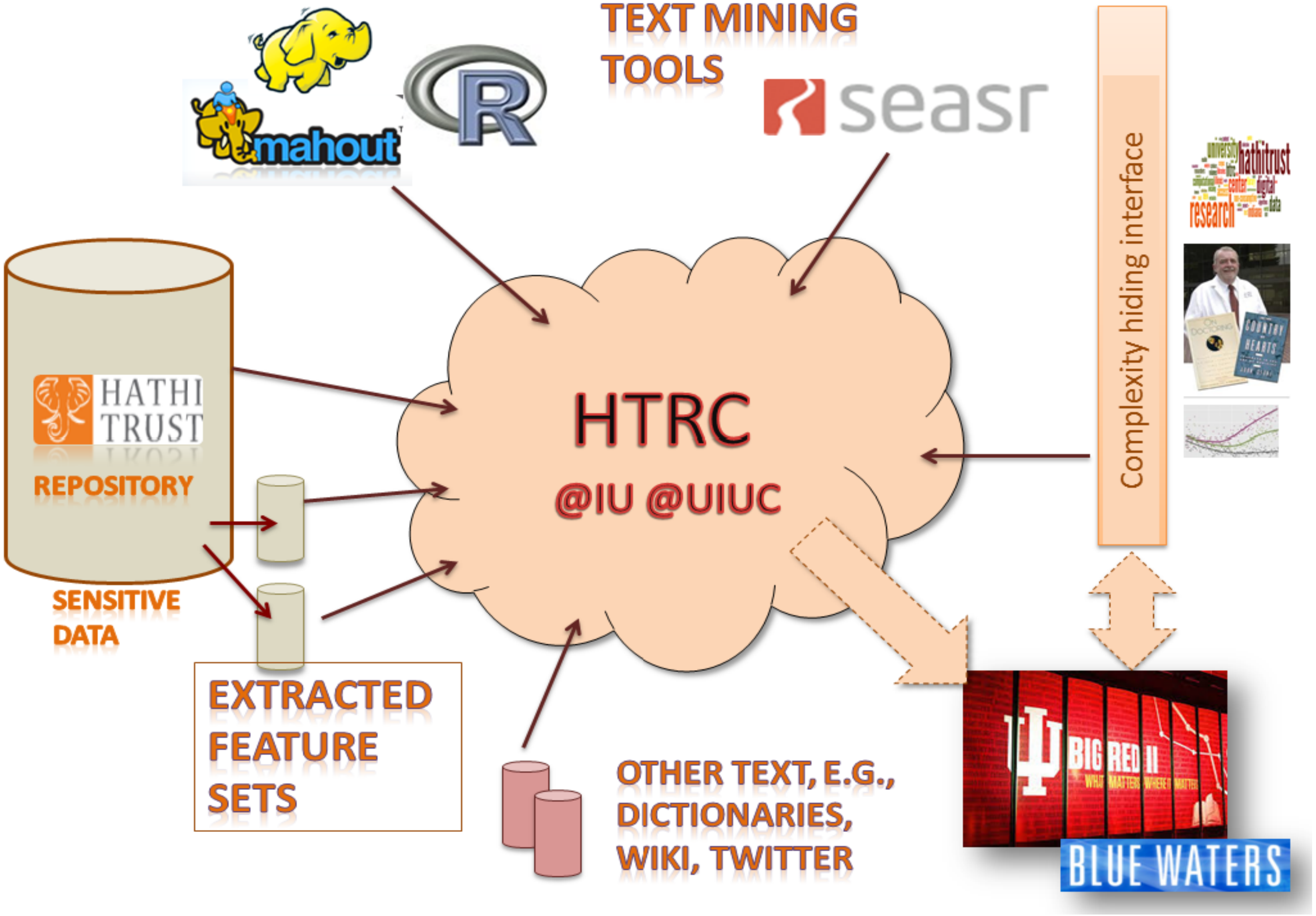
Spatial plots

Statistical plots

Tabular info

Complexity hiding interface

HTRC v2.0



TEXT MINING TOOLS
mahout
R
seasr
HATHI TRUST
REPOSITORY
SENSITIVE DATA
EXTRACTED FEATURE SETS
HTRC
@IU @UIUC
OTHER TEXT, E.G., DICTIONARIES, WIKI, TWITTER
Complexity hiding interface
BIG RED II
BLUE WATERS

HTRC v2.0 + HTRC Data Capsule

- There is a mismatch between what HTRC v2.0 provides and users' needs.
  - HTRC v2.0 provides predefined algorithms to users and runs them on users' behalf. This is to prevent copyrighted data leak.
  - However, a user usually wants to run her own algorithm and exam the results interactively.
- HTRC Data Capsule is developed to strike a balance between preventing data leak while keeping HTRC as flexible as possible to users.

Research Questions

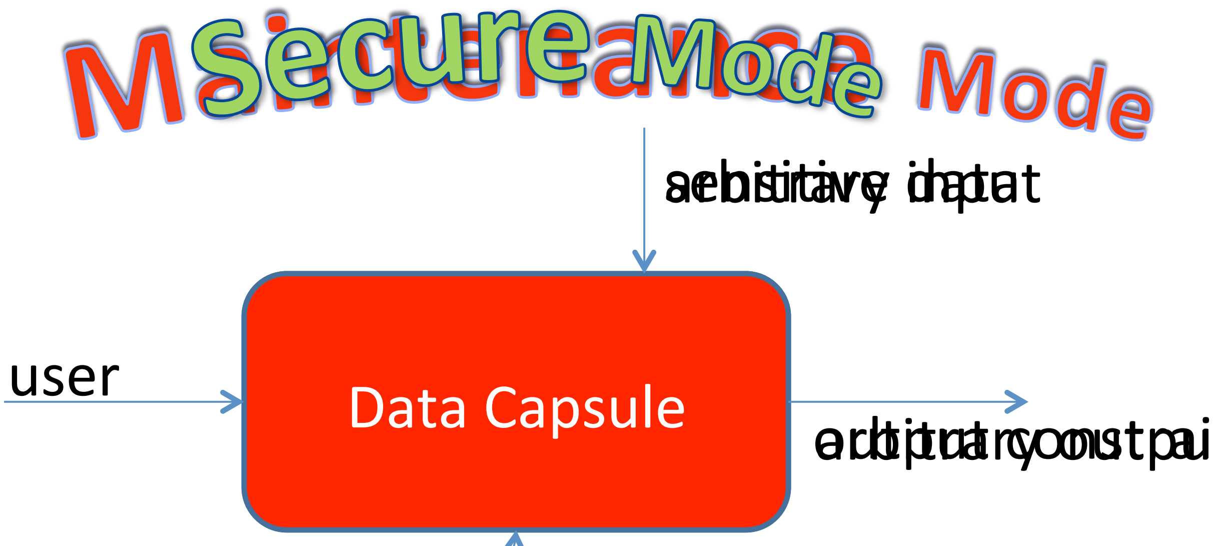
- **Non-consumptive use***: can framework provide safe handling of large amounts of protected data?
- **Openness**: can framework support user-contributed analysis without resorting to code walkthroughs prior to acceptance?
- **Large-scale** and **low cost**: can protections be extended to utilization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

HTRC Data Capsule

- Provisions virtual machines (VM) for researchers to run their algorithms over copyrighted data.
- Trusts researchers to not deliberately leak copyrighted data.
- Prevents malware acting on researcher’s behalf from leaking data.

Building Block – Data Capsule



Maintenance Mode

Secure Mode

arbitrary input

sensitive data

user

Data Capsule

arbitrary output

audit trail

Computation is carried out inside Data Capsule.

K. Borders, E. V. Weele, B. Lau, and A. Prakash.
Protecting confidential data on personal computers with storage capsules.
In 18th USENIX Security Symposium, SSYM'09, pages 367–382. USENIX Association, 2009.

Design Options

- HTRC Data Capsule extends data capsule to build a cloud environment around data capsule to serve multiple users.
  – Build the system around an existing cloud platform, e.g., OpenStack;
  – Build the system from scratch through web service and *QEMU*.

Design Options

- HTRC Data Capsule extends data capsule to build a cloud environment around data capsule to serve multiple users.
  - Build the system around an existing cloud platform, e.g., OpenStack, Eucalyptus; ✗

    (Data Capsule relies on low level control of the VM which requires a lot of customizations of existing cloud platforms. In addition, OpenStack allows a user to configure the VM network which poses threats to Data Capsule.)
  - Build the system from scratch through web service and *QEMU*.

Design Options

- HTRC Data Capsule extends data capsule to build a cloud environment around data capsule to serve multiple users.
  - Build the system around an existing cloud platform, e.g., OpenStack;
  - Build the system from scratch through web services and *QEMU.* ✓

    (This option gives us the highest degree of flexibility to implement HTRC Data Capsule.)

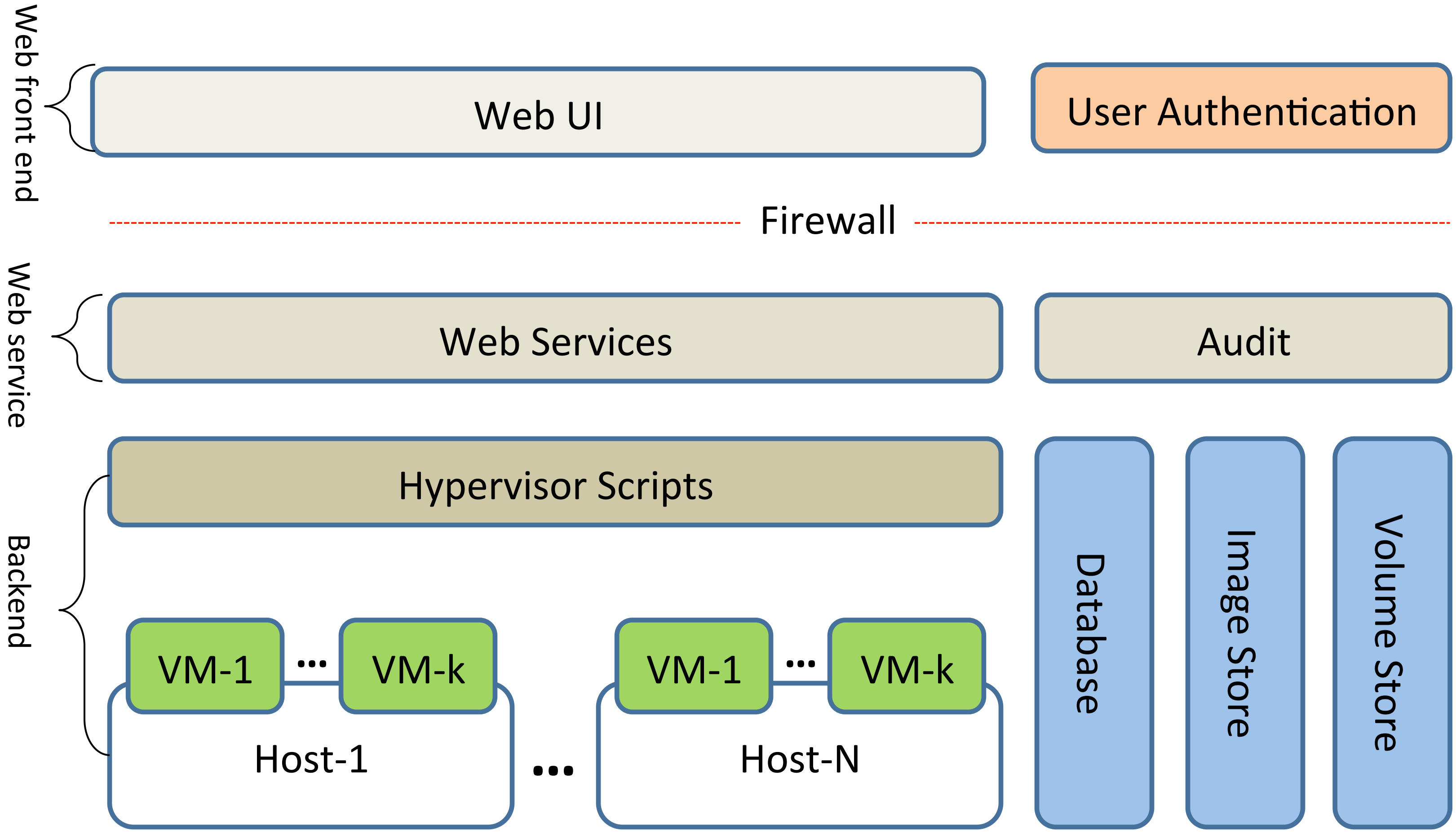
Threat Model

- The user is trustworthy.
- The virtual machine manager and the host it runs on are also trusted.
- The VM is NOT trusted. We assume the possibility of malware being installed as well as other remotely initiated attacks on the VM, which are undetectable to the user.

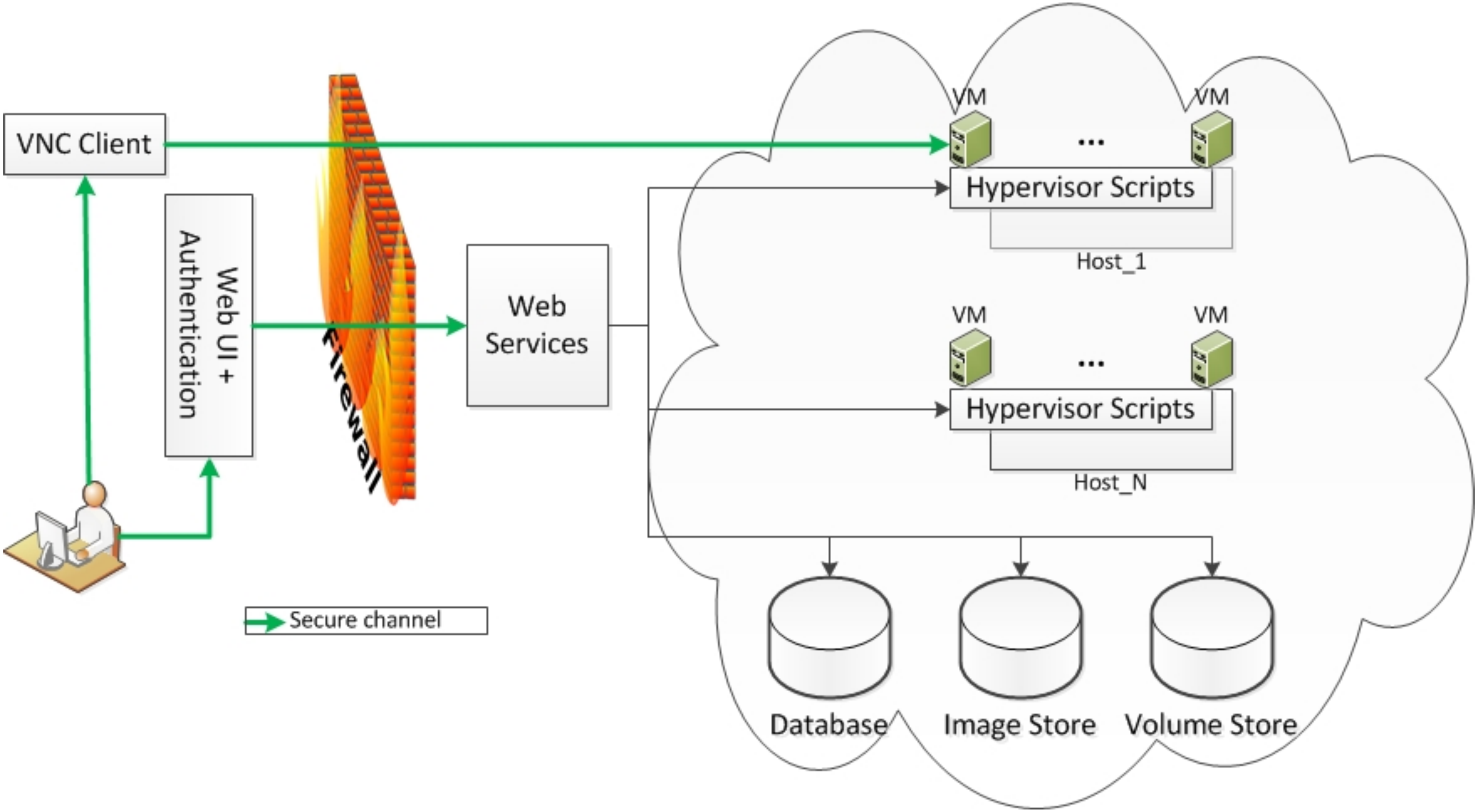
Threat Model (Cont.)

- The VNC session and final result download are two channels which data could leak from potentially.
  - For VNC session, we could encrypt the session to prevent eavesdropping.
  - For final result download, we could monitor traffic on the release channel as a means to automatically detect leakage.
- Covert channels between VMs on the same host also could leak data potentially.
  - In the future, we could run VMs on separated hosts to provide strong isolation.

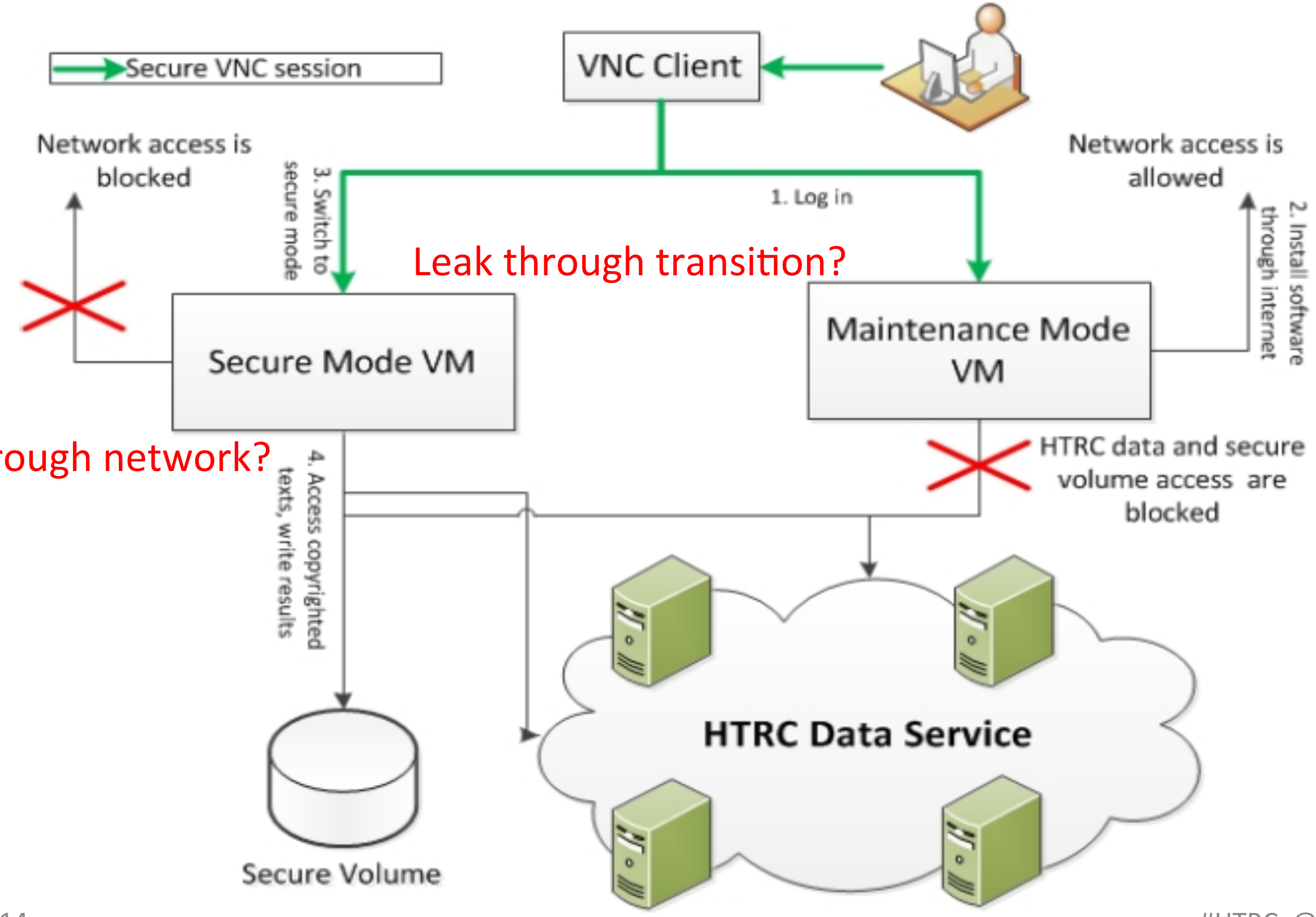
HTRC Data Capsule Architecture



HTRC Data Capsule Workflow



HTRC Data Capsule Access



Secure VNC session

VNC Client

Network access is blocked

3. Switch to secure mode

1. Log in

Network access is allowed

2. Install software through internet

Secure Mode VM

Maintenance Mode VM

HTRC data and secure volume access are blocked

4. Access copyrighted texts, write results

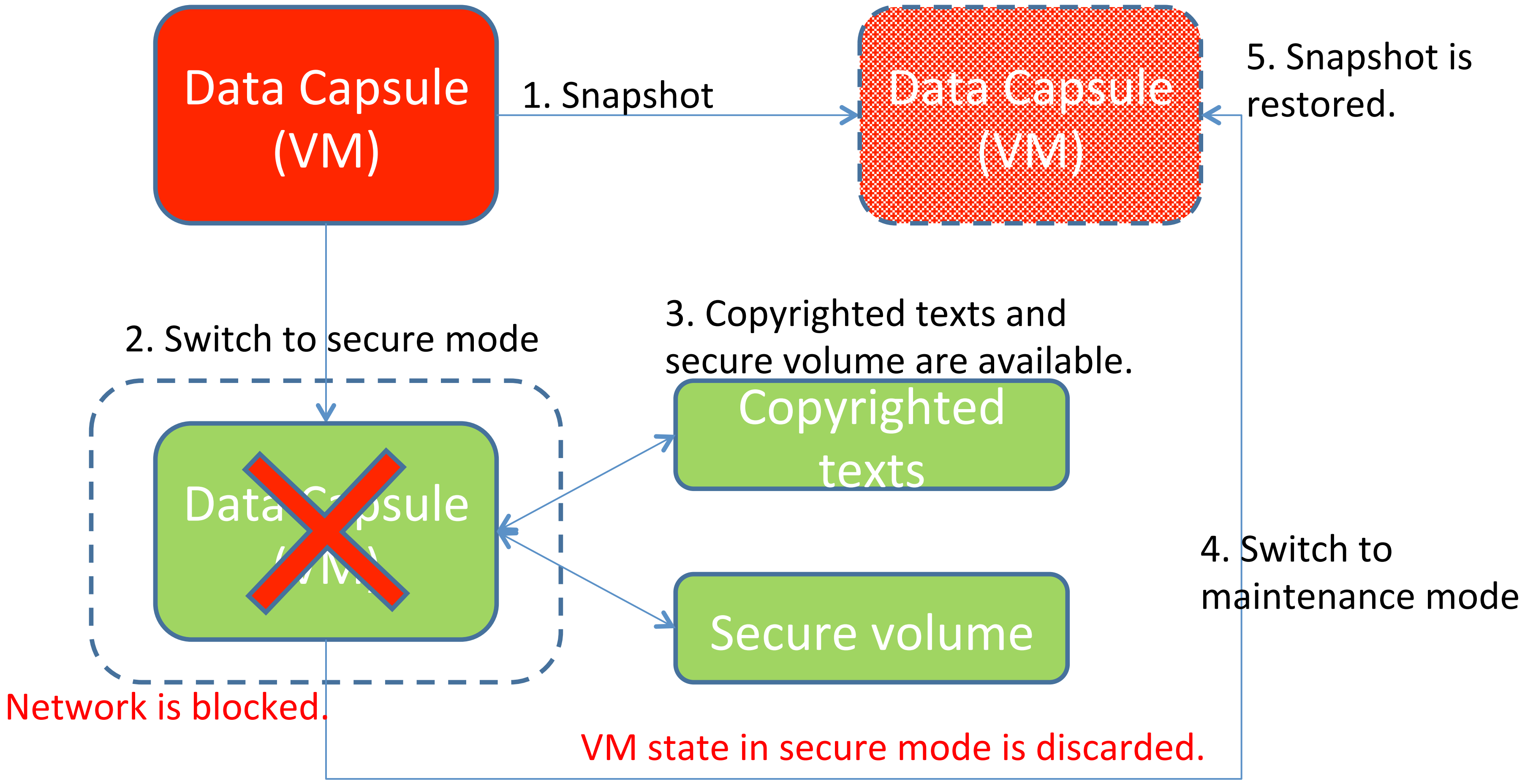
Secure Volume

HTRC Data Service

Leak through transition?

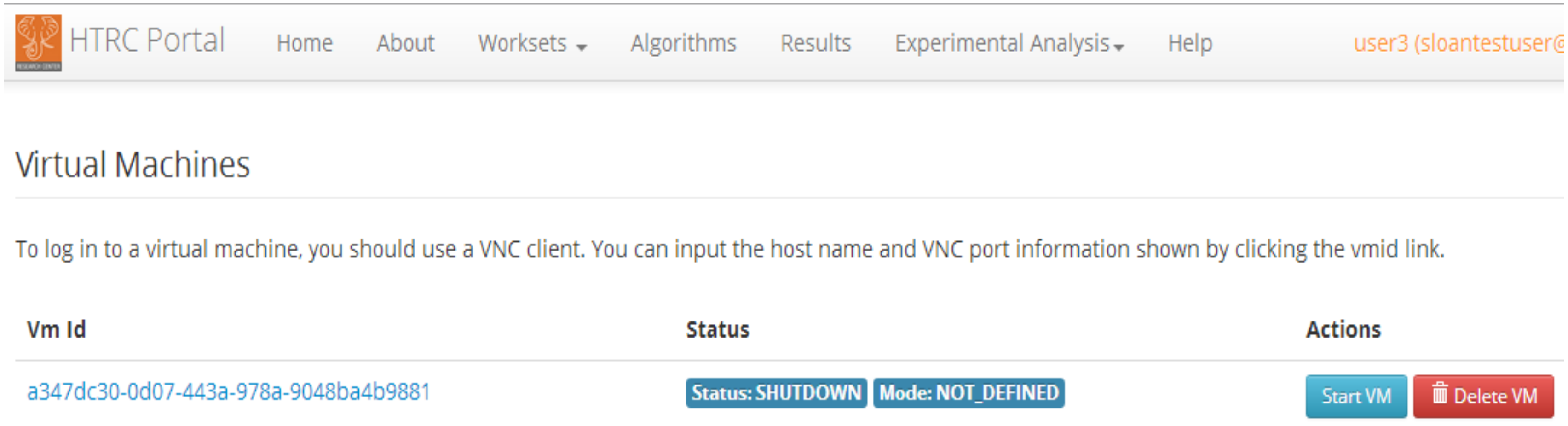
Leak through network?

Data Capsule Mode Switch



VM Operations Screenshots

VM in shutdown state.

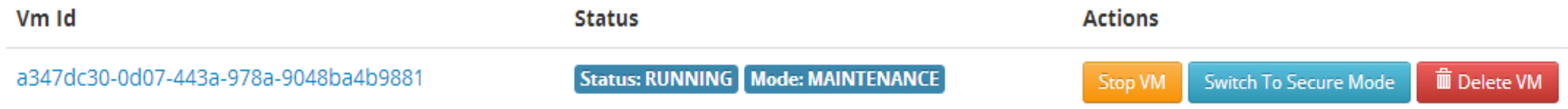


Virtual Machines

To log in to a virtual machine, you should use a VNC client. You can input the host name and VNC port information shown by clicking the vmid link.

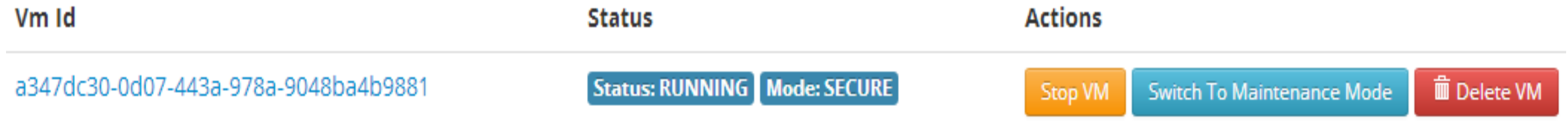
Vm Id	Status	Actions
a347dc30-0d07-443a-978a-9048ba4b9881	Status: SHUTDOWN Mode: NOT_DEFINED	Start VM Delete VM

VM in maintenance mode.



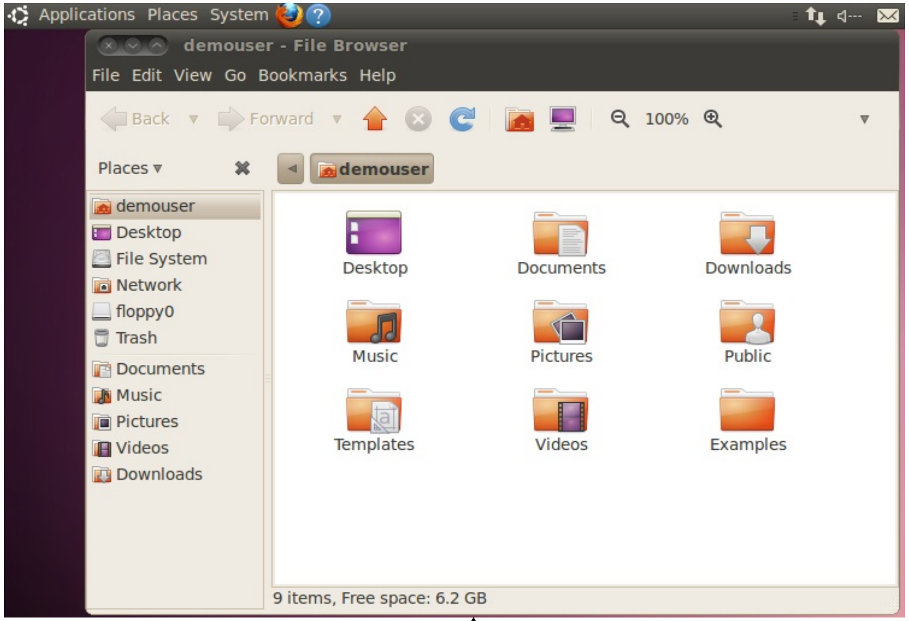
Vm Id	Status	Actions
a347dc30-0d07-443a-978a-9048ba4b9881	Status: RUNNING Mode: MAINTENANCE	Stop VM Switch To Secure Mode Delete VM

VM in secure mode.



Vm Id	Status	Actions
a347dc30-0d07-443a-978a-9048ba4b9881	Status: RUNNING Mode: SECURE	Stop VM Switch To Maintenance Mode Delete VM

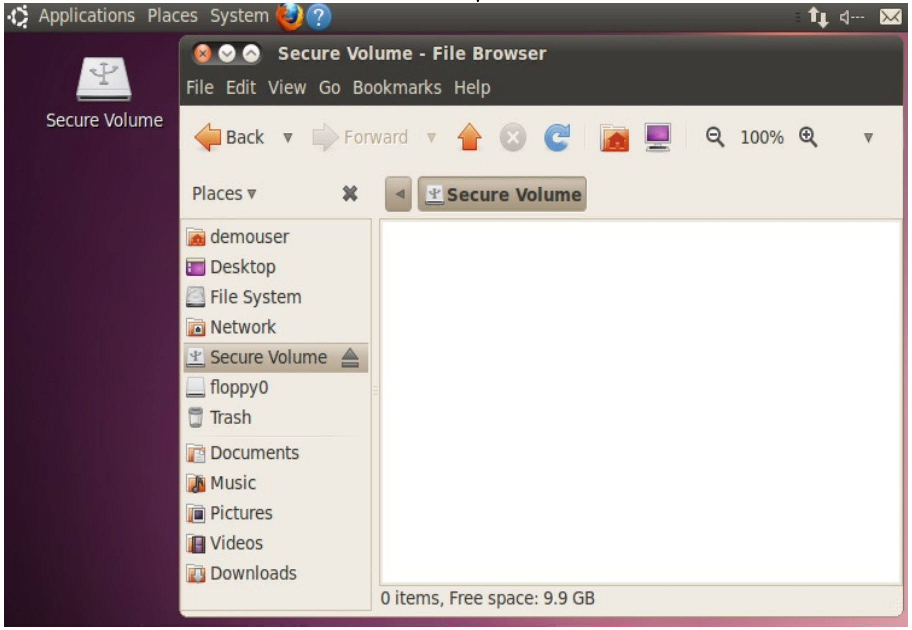
VM Access Screenshots



9 items, Free space: 6.2 GB

Maintenance Mode

Secure Mode



0 items, Free space: 9.9 GB

User Feedback

- Non-consumptive use
  - Initial users report that they can only access the internet in maintenance mode and HTRC data service in secure mode. They can neither make persistent changes to VMs in secure mode, nor access other users' VMs by SSH'ing.
- Openness and efficiency
  - Initial users report that they are able to configure the VM as needed, and run their analysis against HTRC data interactively.

Future Work

- The upcoming hands-on session!
  - Sep 15, 2-5pm
  - Wells Library E174
  - Bring your own laptop
  - One of the Scholar Commons Events
  - Register at the Scholar Commons page [1]
  - Work on text analytics tasks in the HTRC Data Capsule environment

[1] http://libraries.iub.edu/tools/workshops/workshop-listings/series-view/283/series

Future Work

- Copyrighted content in progress
- Advanced Collaborative Support
  - The award model
  - Award content is HTRC ACS staff time
  - Collaborate with scholars on addressing their research needs related to HTRC
  - E.g. prototyping, running text analysis
  - Advocate open source; encourage extending the work to a grant submission
- Scholars Commons
  - Interaction with scholars to help using HTRC tools and services
  - An interface to interact with HTRC users via the channel of scholars commons
  - Series of workshops at IU and other places
  - Weekly consulting time
  - Every Wed 2:30 – 4:30pm, IU library, Scholars Commons 157R
  - Contact: Miao Chen, Nicholae Cline

- For details http://www.hathitrust.org/htrc/faq
- General contact info
  - Beth Plale, Director HTRC, plale@indiana.edu
- Requests for capability, interest
  - Miao Chen, Asst. Director for Outreach HTRC miaochen@indiana.edu