Cloud Computing Data Capsules for Non-Consumptive Use of Texts

Jiaan Zeng[1], Guangchen Ruan[1], Alexander Crowell[2], Atul Prakash[2], Beth Plale[1]

[1] School of Informatics and Computing, Indiana University

[2] Computer Science and Engineering Division, University of Michigan

HathiTrust Research Center

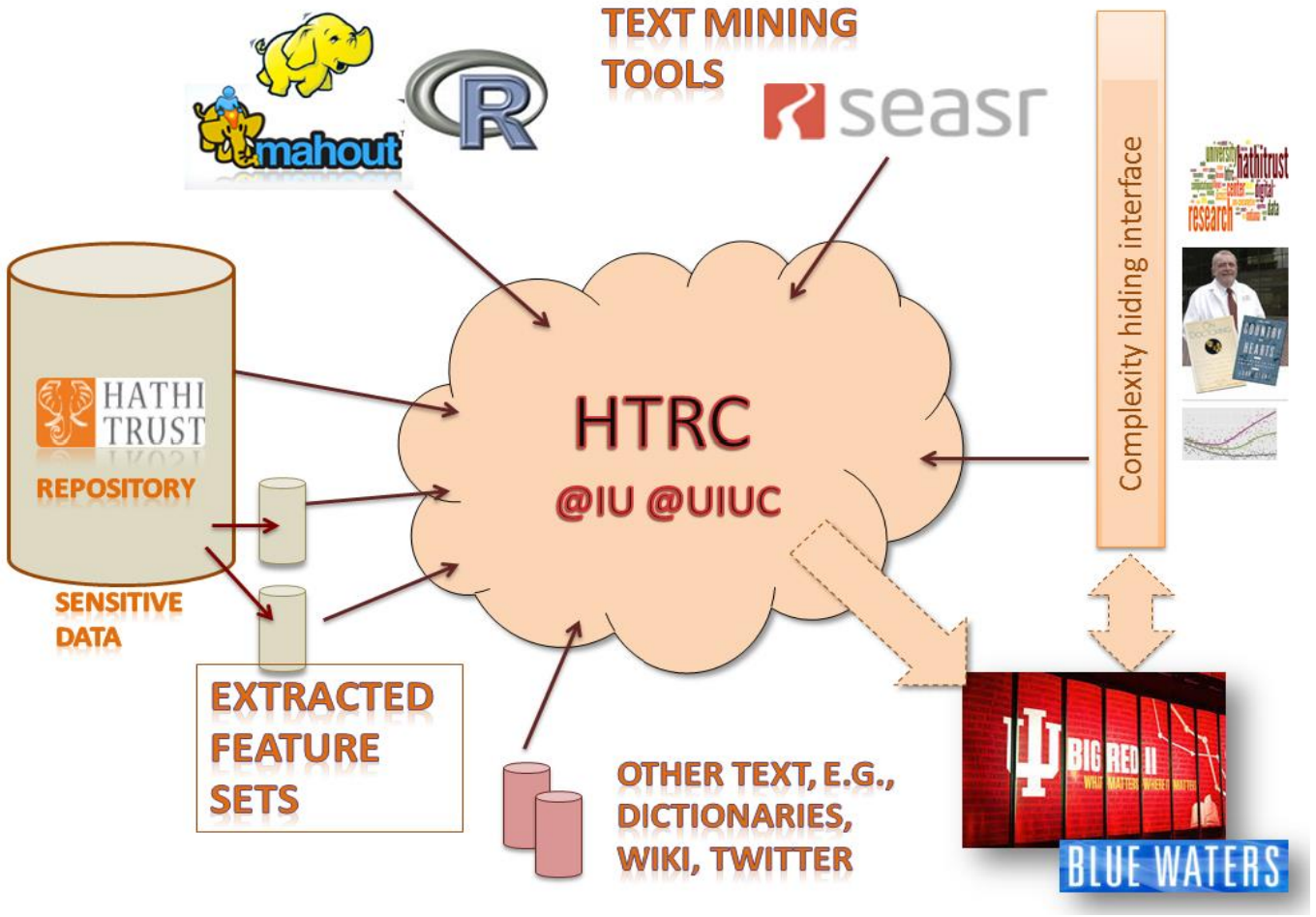
- The HathiTrust Research Center (HTRC) was established in 2011 to enable computational research across the texts and images of the HathiTrust digital repository which has 12 million digitized books.

Motivations for HTRC

- It is about BIG data.
  - Statistics of currently digitized books * :
    11,158,214 books; 3,905,374,900 pages;
    500 terabytes;
  - It needs an advanced infrastructure for text mining in such massive scale.
- Most of its data is copyrighted.
  - 66 % of total is copyrighted;
  - It suggests need for new forms of access that preserves intimate nature of interaction with texts *while* at same time *honoring restrictions on access.*

* http://www.hathitrust.org/statistics_info

HTRC v2.0


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

HTRC v2.0 (Cont.)

- 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

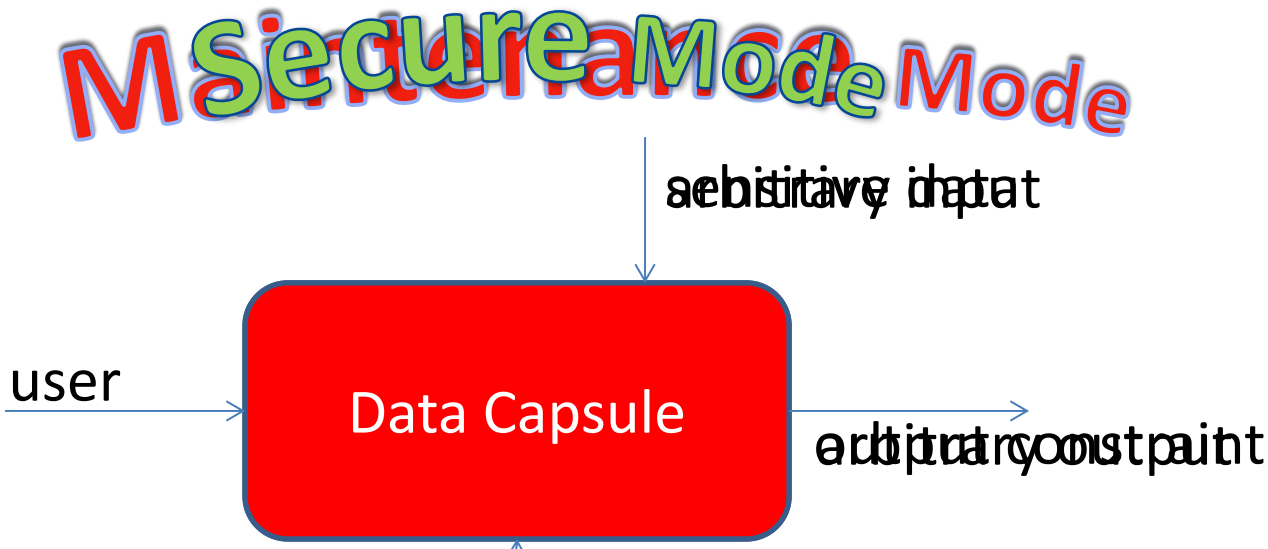
- **Non-consumptive use***: can framework provide safe handling of large amounts of protected data?
- **Openness**: can framework support user-contributed analysis without resorting to code walkthroughs prior to acceptance?
- **Large-scale** and **low cost**: can protections be extended to utilization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



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.)

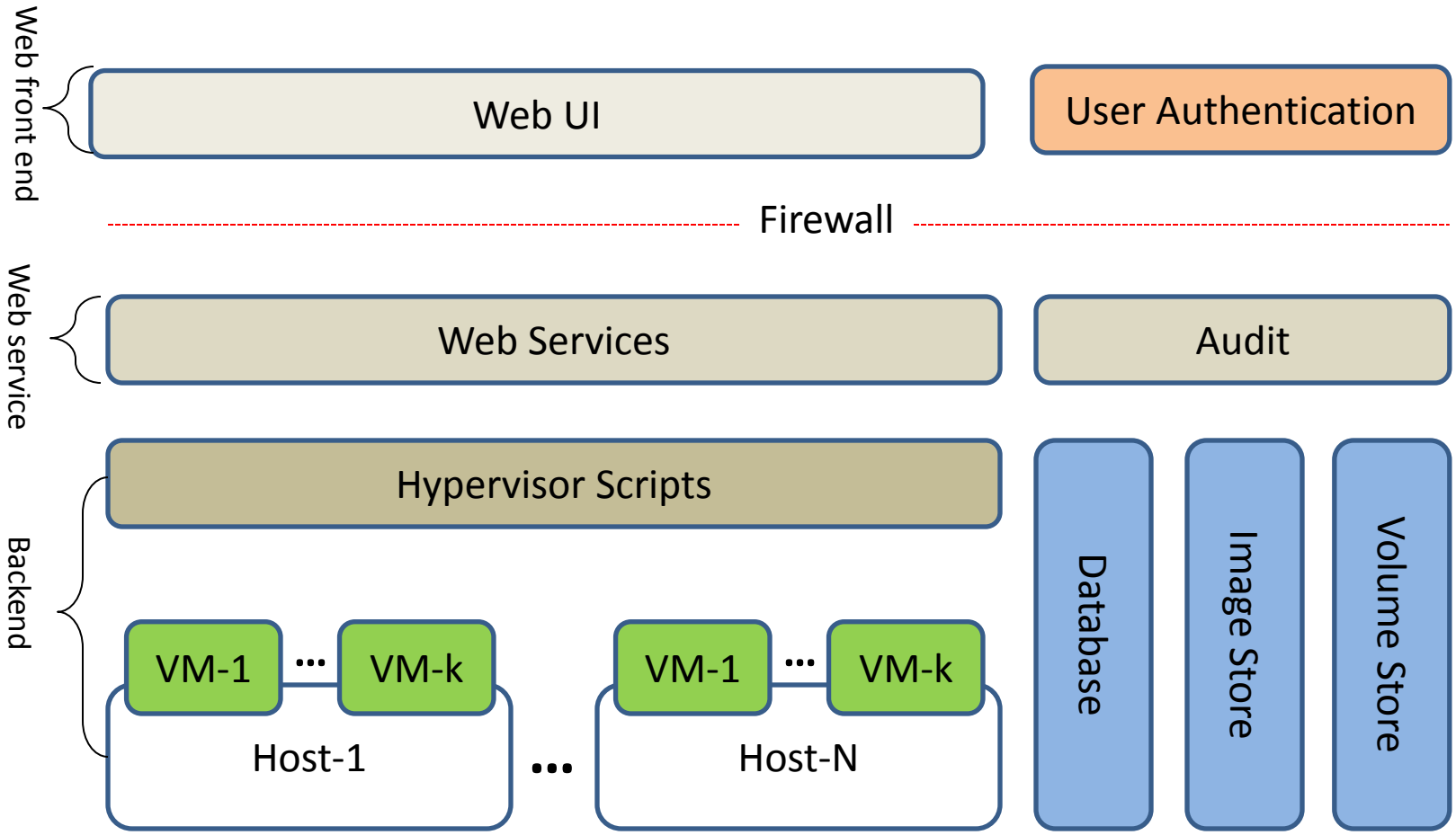
Threat Model

- 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.

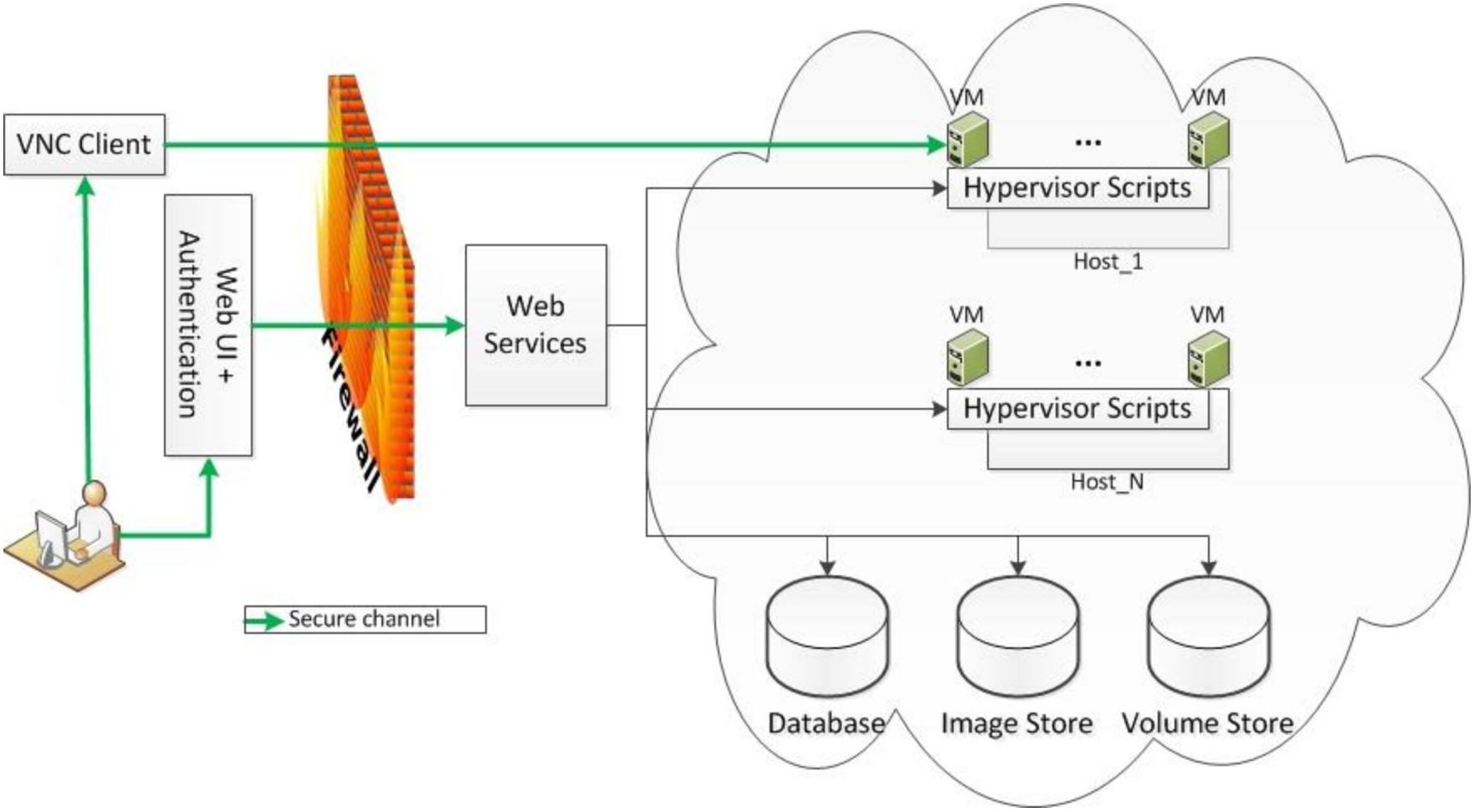
Threat Model (Cont.)

- 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.

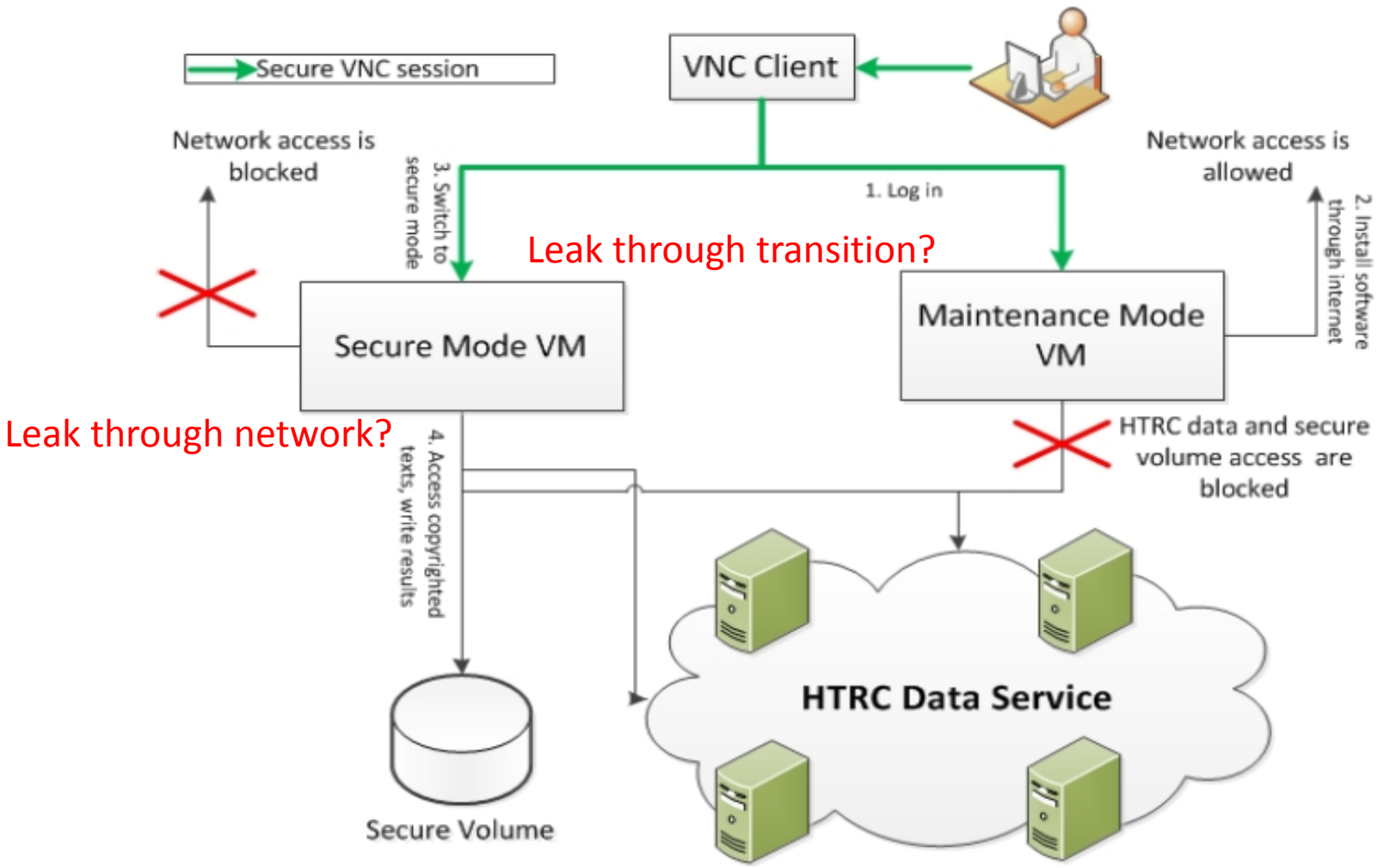
HTRC Data Capsule Architecture



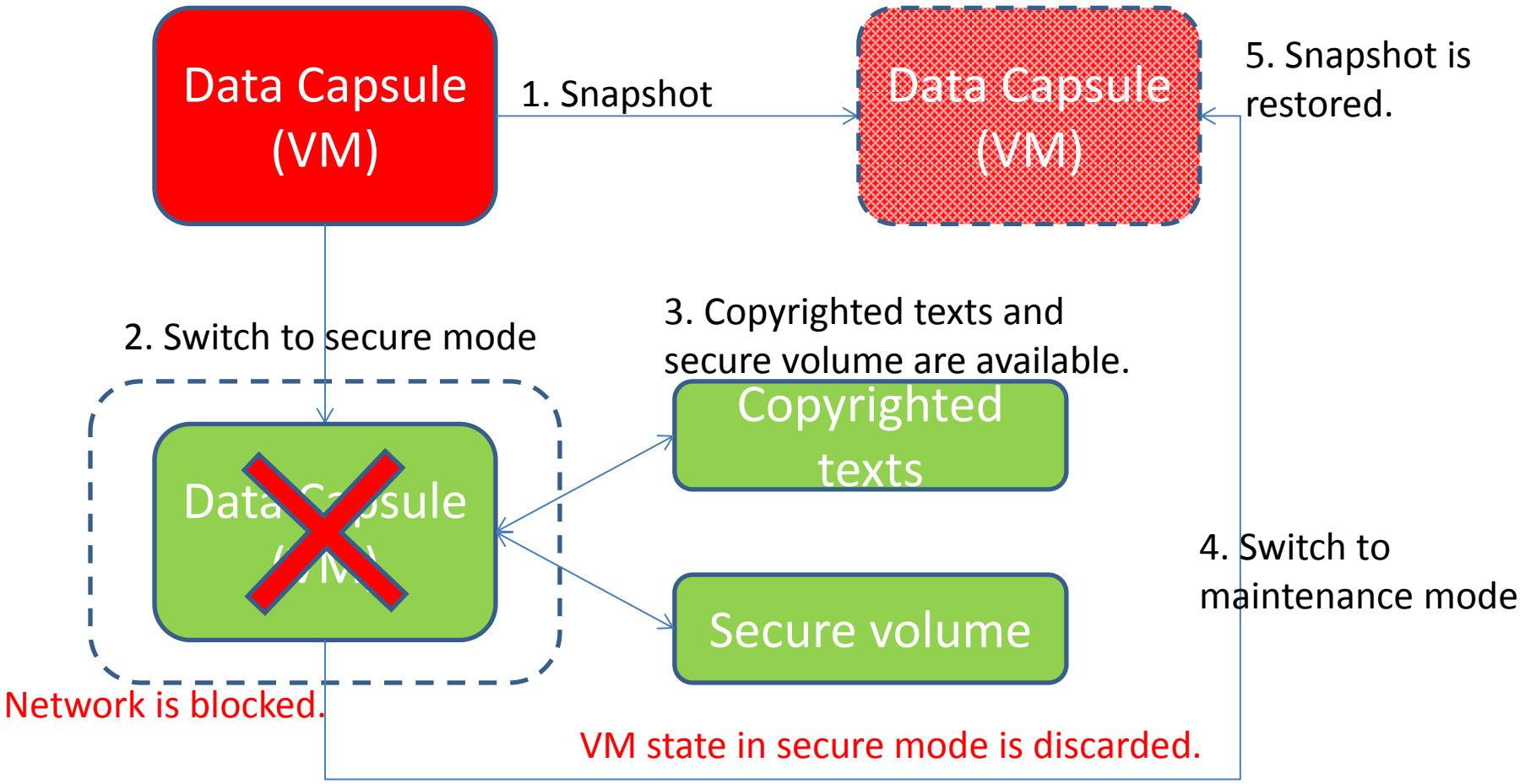
HTRC Data Capsule Workflow



HTRC Data Capsule Access

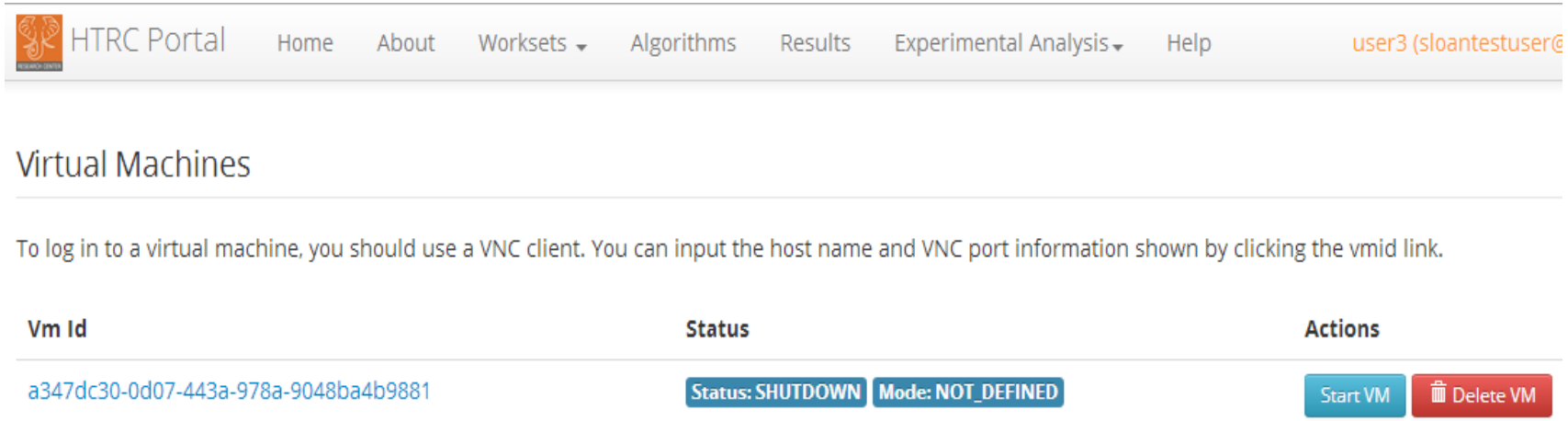


Data Capsule Mode Switch



VM Operations Screenshots

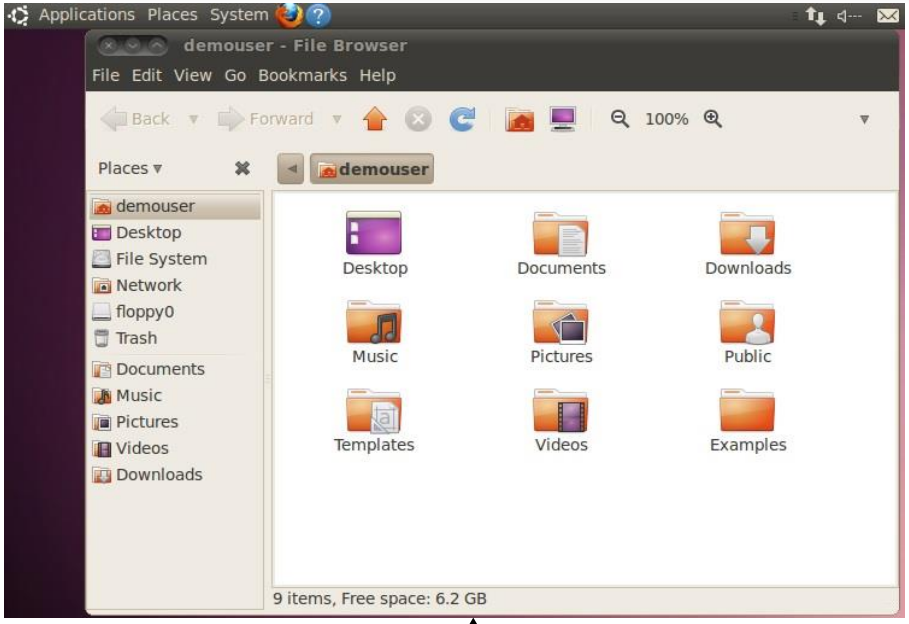
VM in shutdown state.



VM in maintenance mode.

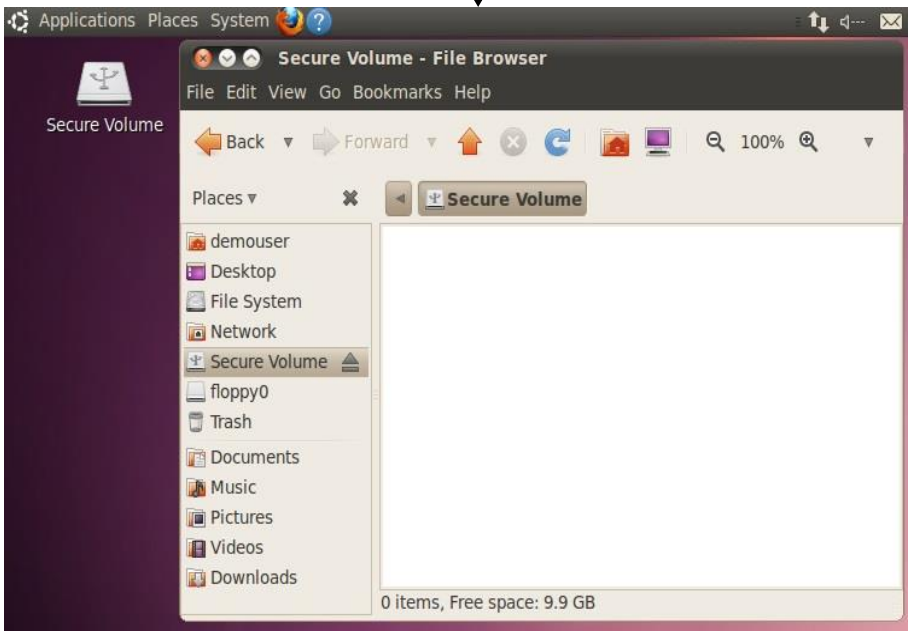
VM in secure mode.

VM Access Screenshots



Maintenance Mode

Secure Mode



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 user is not trustworthy.
  - A user may leak data through the VNC channel and encode data in the final result. A solution might be to analyze the traffic on both channels.
  - A user may use the covert channel among VMs to leak data. A solution might be to place VMs with different modes on different hosts.
- Run the data capsule in a distributed environment.
  - Run the data capsule on a cluster instead of a single VM;
  - Ship part of the computation of data capsule to public cloud resources;
  - Integrate external data sources into data capsule.

Acknowledgements

- This research is funded through a grant from the Alfred P. Sloan Foundation,
- This research is also based in part on work supported by the National Science Foundation and by Samsung.
- Special thanks to Samitha Liyanage, Milinda Pathirage, and Zong Peng at Indiana University, and Earlence Fernandes and Ajit Aluri at the University of Michigan for discussions contributing to this work.

Thanks!

Questions?