

# Efficient Tracking, Logging, and Blocking of Accesses to Digital Objects

## Cyber Security Division 2012 Principal Investigators' Meeting

October 11, 2012

Fabian Monroe  
University of North Carolina at Chapel Hill  
[fabian@cs.unc.edu](mailto:fabian@cs.unc.edu)  
919-962-1763

Michael Bailey, University of Michigan  
Charles Schmitt, Renaissance Computing Institute



renci

RESEARCH \ ENGAGEMENT \ INNOVATION

# Team

- **Fabian Monrose** is a Professor of Computer Science at **University of North Carolina at Chapel Hill**. Prior to joining UNC, he was an Associate Professor at John Hopkins University, and a founding member of their Information Security Institute. From 1999-2002, he was a member of technical staff at Bell Labs. He holds a Ph.D. in Computer Science from New York University. He has published over 70 papers in computer security, and served on numerous technical program committees and government panels.
- **Michael Bailey** is a Research Professor at **University of Michigan**. He currently directs and contributes to research on the security and availability of complex distributed systems. Prior to working at the university, he was **Director of Engineering at Arbor Networks**, and a programmer at both Amoco Corporation and Andersen Consulting. As Director of Engineering, he coordinated the actions of engineering managers, architects, engineers, and release engineering for all of Arbor's products.
- **Charles Schmitt** provides technical leadership and management for **RENCI** biological and medical science related projects. Prior to joining RENCI, he was the senior computer scientist at BD Technologies, where he assisted in software development and bioinformatics support for programs in medical diagnostics and genomics. He also served as the **primary architect** and developer of the MPM software informatics platform. He holds a B.S. degree in physics and a Ph.D. degree in CS from UNC-Chapel Hill.



# Need

- Researchers and practitioners routinely require access to large corpora of **sensitive** data for a wide host of scientific activities
- Of late, several **virtual data enclaves** have emerged as solutions for hosting such data



# Data Enclaves

- With this approach, an investigator works in a room dedicated to accessing the data
- Only **approved** researchers are allowed in the enclave.
  - Computers in the room are **not** connected to the Internet or to other external resources
  - Researchers **cannot** take individual-level data from the enclave, and all results they obtain are checked by the ***data disseminator*** for potential confidentiality breaches ***before*** they can be taken out of the enclave
- **Disadvantage:** they are inconvenient, and as a result, the amount of analysis carried out in the enclave is limited

# Virtual Data Enclaves

- The data are housed in a system owned by the data disseminator.
- Approved secondary researchers access the data remotely.
- Users of a virtual data enclave cannot store the data on their computers, and certain functions (e.g., printing) are disabled.
  
- **Advantages:**
  - Researchers access the data without traveling to a secure site
  - They avoid some of the disclosure risks posed by researchers storing data on their own machines, such as the accidental loss of CD-ROMs or sharing of data with unapproved investigators
- **Confidentiality protection** depends on researchers not **violating** terms of the data use agreement.

# Need

- Researchers and practitioners routinely require access to large corpora of **sensitive** data for a wide host of scientific activities
- Of late, several **virtual data enclaves** have emerged as potential solutions for hosting such data
- Yet, few solutions provide a secure environment for **reducing the risks** of unauthorized access to, and loss of, information hosted in these enclaves

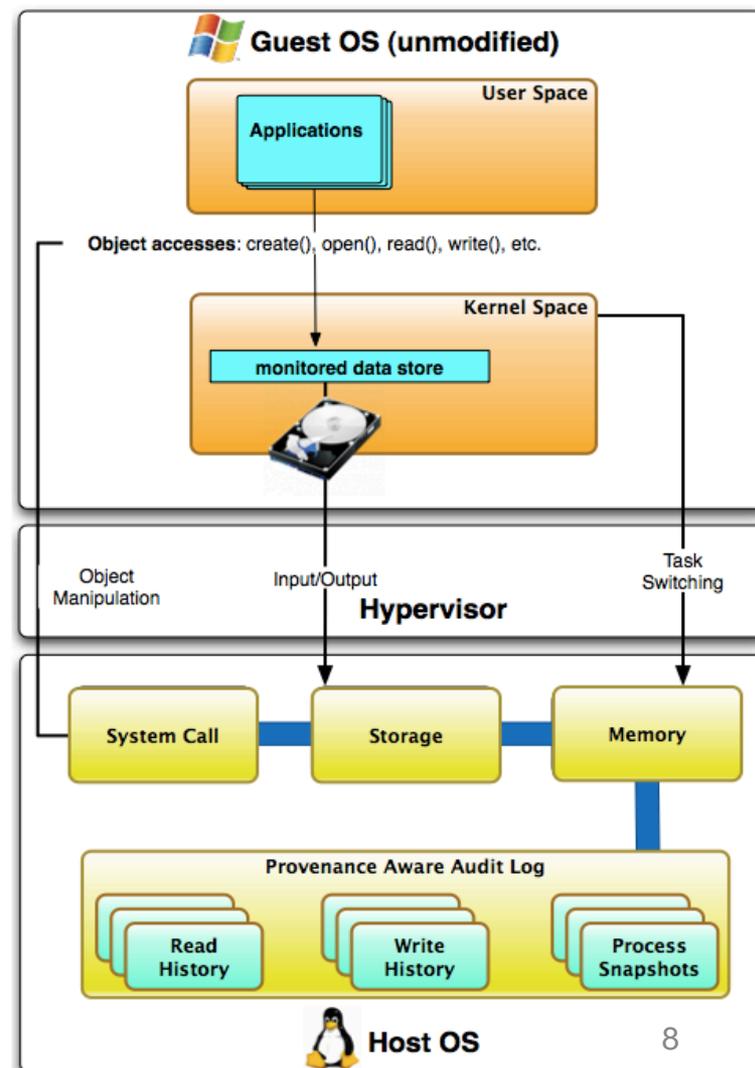


# Goals

- To design and implement techniques for tracking the chain of custody of sensitive data in a virtualized environment
- Deliverables:
  - An object tracking platform that will enable DHS and its customers to (a) **identify** and authenticate access to digital objects that originate from disk (b) **track accesses** to these objects on disks and in memory and (c) **track changes** to these objects via a provenance-aware audit trail
- Deployment and evaluation within the **Secure Research Workspace** at the Renaissance Computing Institute (RENCI)

# Approach

- Monitoring framework implemented within a Hypervisor
  - extends **TrailOfBytes** prototype
- Spans three layers: **storage, memory, and system-call** modules
  - *key idea is in monitoring access to physical memory when data is first loaded from a datastore*
- Semantic linkages captured in a **provenance-aware** filesystem
  - *provides histories via successive versioning*



# Challenges

- Efficiently bridging the “**semantic gap**” problem
- Process **attestation**:
  - follow approach taken in **Patagonix** to provide a lightweight identification facility (based on fingerprints of codepages belonging to an application)
- **Dynamic provisioning**:
  - *on-the-fly* monitoring of data spanning multiple disks
- **Selective Blocking (year 2)**:
  - Data leakage protection via a network monitoring module that blocks (or buffers) packets containing information from the monitored datastore

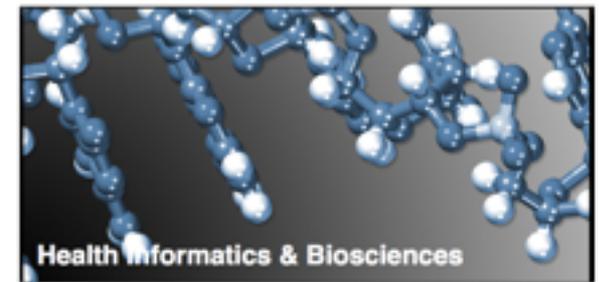
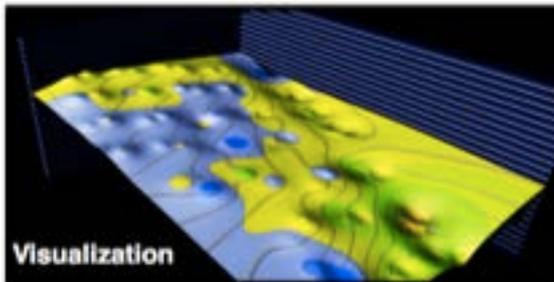
# Benefits

## ■ Enhance state of the art in digital provenance in virtual data enclaves

- ☑ **Improved data provenance representation:** A rich interface for managing and **mining the recorded information**, thereby providing deeper insights into *how* objects were manipulated
- ☑ **Limiting breaches:** Capabilities to not only record, but also to deny, unauthorized accesses or transfer of data from datastores for which provenance tracking has been enabled

# Why RENCI?

- The Renaissance Computing Institute is:
  - A provider of technology leadership for the state of NC, its citizens and its institutions
  - A pipeline that brings university expertise and cutting-edge technologies into the “real” world in order to find solutions to important state problems
  - A **multi-institutional** center that unites campus and local communities in projects that bring federal research dollars to NC



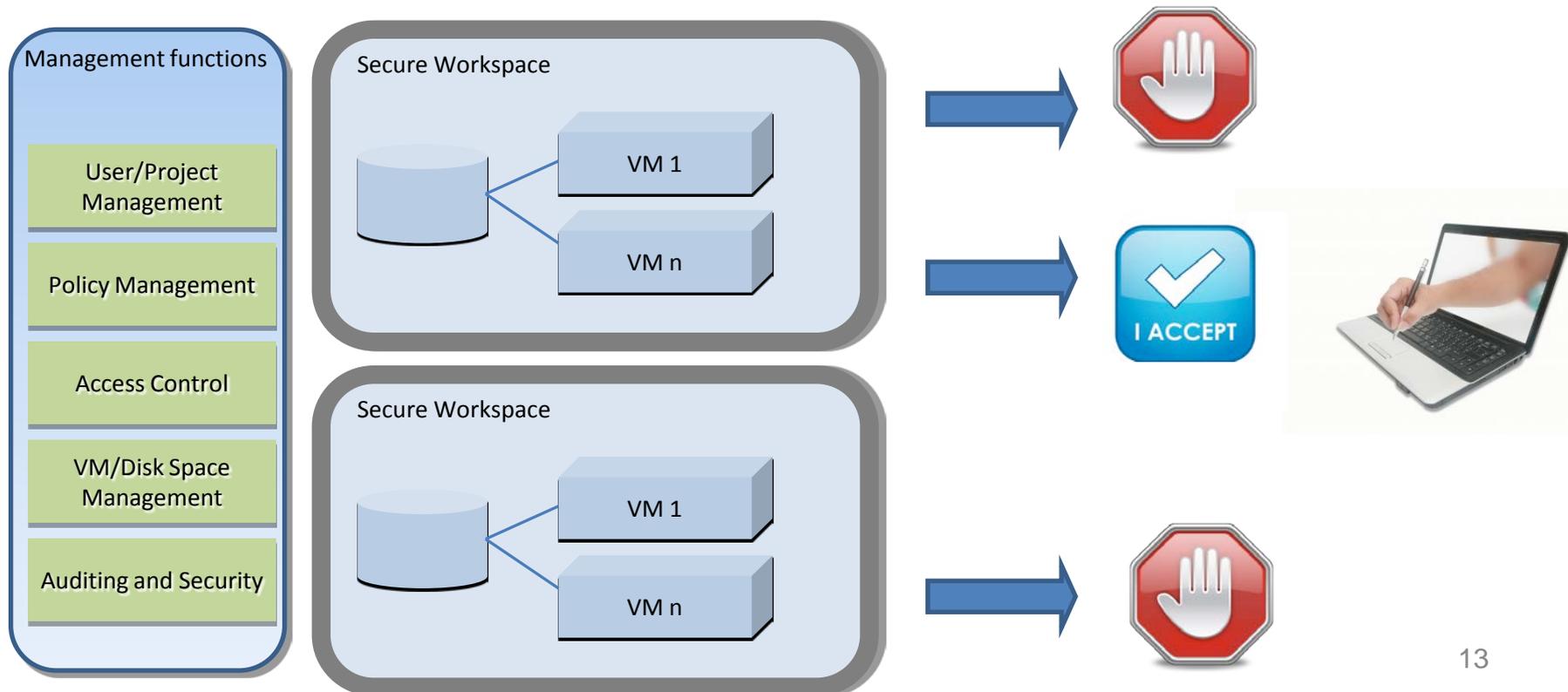
# Deployment & Evaluation

- Integration within the **Secure Research Workspace** Project
- Provide a secure environment aimed at reducing the risks of unauthorized access to, and loss of, **sensitive health** information



# Deployment & Evaluation

- Integration within the **Secure Research Workspace** Project
- Provide a secure environment aimed at reducing the risks of unauthorized access to, and loss of, **sensitive health** information



# Deployment & Evaluation

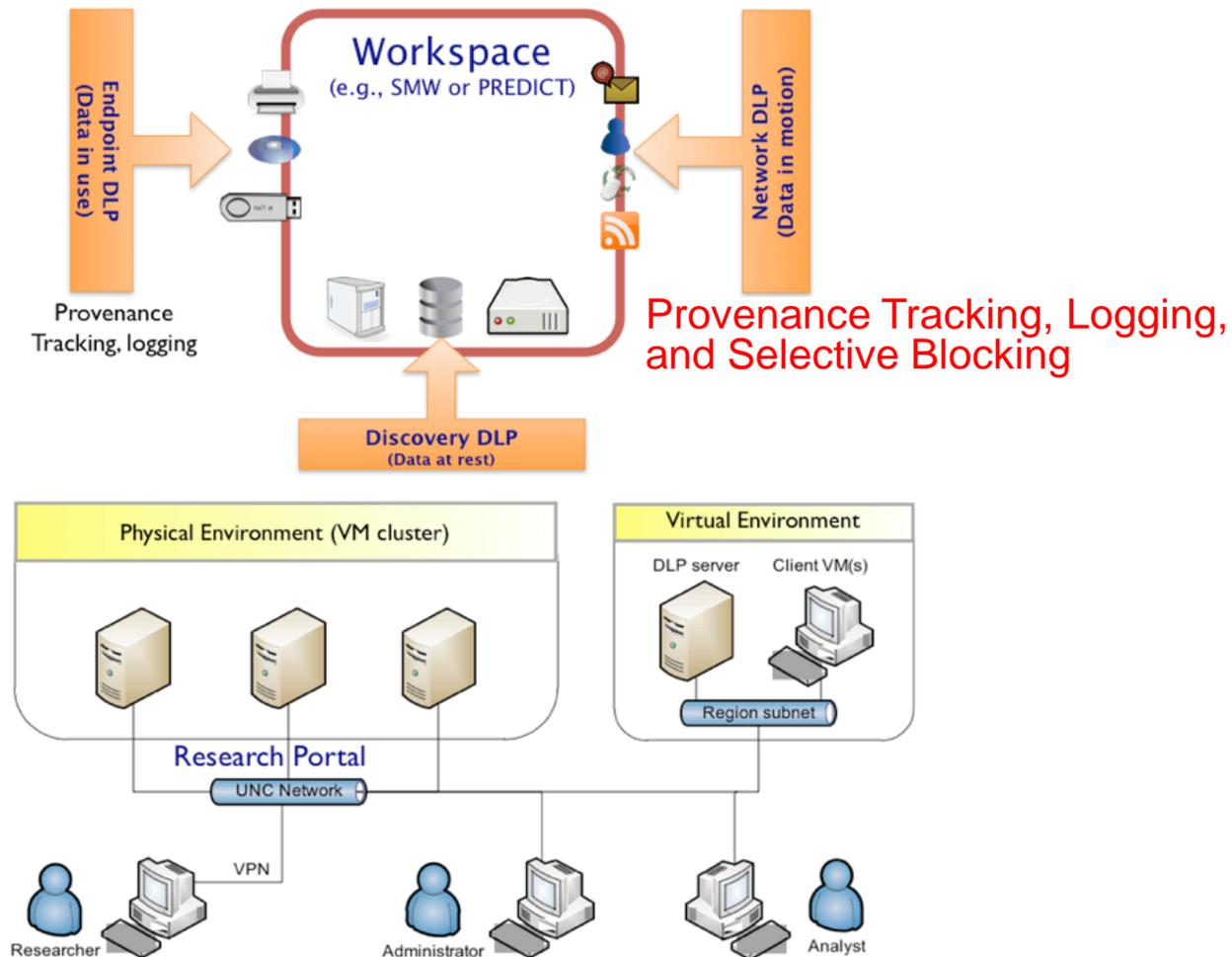
renci

RESEARCH \ ENGAGEMENT \ INNOVATION



# Deployment & Evaluation

- **Solution:** Integrate our framework within the SRW data warehouse



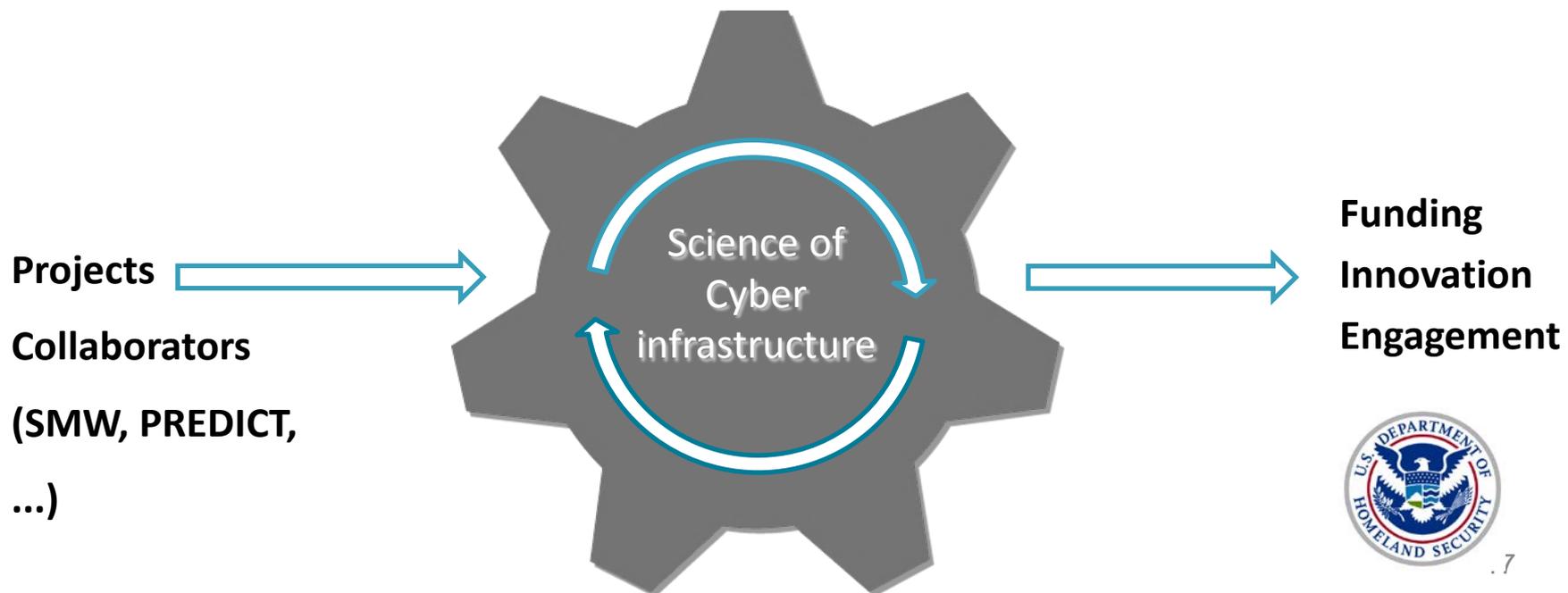
# Milestones

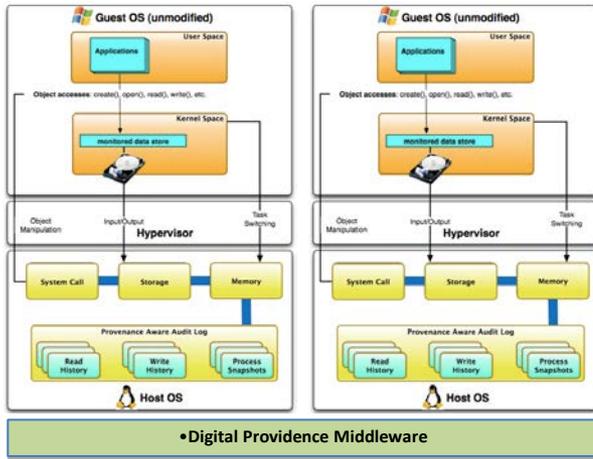
- Immediate: We are **Hiring!** See me or point candidates to <https://unc.peopleadmin.com/postings/7654> for more info.

Task Description	Month				
	1	6	12	18	24
<b>Design and Implementation</b>					
▽ Ontological Provenance Model (Task 4.1)	•	•			
▽ Development					
▷ KVM Import (Task 4.2)	•	•			
▷ Dynamic Provisioning (Task 4.3)	•	•	•	•	•
▷ Provenance Tracking (Task 4.4)		•	•	•	
▷ Capturing Semantic Linkages (Task 4.5)			•	•	
▷ Tamper Resistant Audit Trail (Task 4.6)				•	•
▷ Origin Attestation (Task 4.7)			•	•	
▷ Lightweight Process Snapshots (Task 4.8)			•	•	
▷ Multi-host Tracking (Task 4.9)				•	•
▷ Selective Blocking (Task 4.10)				•	•
<b>Testing, Deployment &amp; Outreach</b>					
▽ Deployment and Case Study (§4.4.1)			•	•	
▽ Software Release					•
▽ Publications & Documentation			•		•

# Technology Transfer Plan

- Develop and deploy within SRW project
- Make technology available to commercial partners under fair and reasonable contracts
- Possibly transition into PREDICT?





**Operational Capability:**

1. Definition of data provenance representations and development of an ontological model
- A new object tracking platform and accompanying tools that will enable customers to: (a) identify and authenticate access to digital objects, (b) track accesses to these objects on disk, in memory, and across the network, and (c) track edits to these objects via a provenance-aware audit trail
  - A rich interface for managing and mining the captured information.
  - We provide capabilities to not only record, but to also deny unauthorized accesses or transfer of data from objects

**Proposed Technical Approach:**

1. We seek to provide a system for following the chain of custody of data in a virtualized environments
- Our approach is designed to track accesses to objects that originate from disk, and capture subsequent accesses to these objects in memory, and on the network.
  - Our system provides accurate, and efficient, tracking and reconstruction tool for collating and storing events collected at different levels of abstraction, as well as a rich management and presentation interface on this data.
  - A key aspect of this approach is a robust, iterative design and evaluation process targeting the Secure Medical Workspace (SMW) Project at RENCi. The goal of the SMW project is to enable secure access to patient records collected by the Carolina Data Warehouse for Health, which houses clinical health records for all of UNC. Hospitals' patients.

**Schedule & Deliverables:**

1. Our current plan calls for these technologies to be designed, constructed, and tested over a 2-year period
2. The support for this Type I proposal will be used to fund:

Deliverable	Date
Work with customers to identify required features	Q1
Representation definition	Q1-Q2
Basic host and virtualization object tracking	Q2-Q6
Provenance middleware	Q4-Q6
Alerting and provenance querying	Q5-Q6
Deployment and evaluation	Q5-Q8

**Corporate Information:**

University of North Carolina at Chapel Hill, Fabian Monrose, 3175 Sitterson Hall, UNC-Chapel Hill, NC, 27599-3175, (919) 962-1763, fabian@cs.unc.edu