Association for Public Data Users 2013 Annual Conference:

A Sea Change for Public Data

George Washington University
800 21st Street NW | Continental Ballroom,
3rd floor | Washington, DC
September 16-17, 2013

Timothy M. Mulcahy
Principal Research Scientist
Director, NORC Data Enclave
NORC at the University of Chicago

NORC
at the UNIVERSITY *of* CHICAGO

Overview

- Resetting Perceptions
- What is the Data Enclave?
- How do you balance data utility & confidentiality?
- How do users access the enclave?
- What functionality is available?
- How do I import and export files?
- How do I collaborate with other researchers?

The Licensing ("Trust") Model
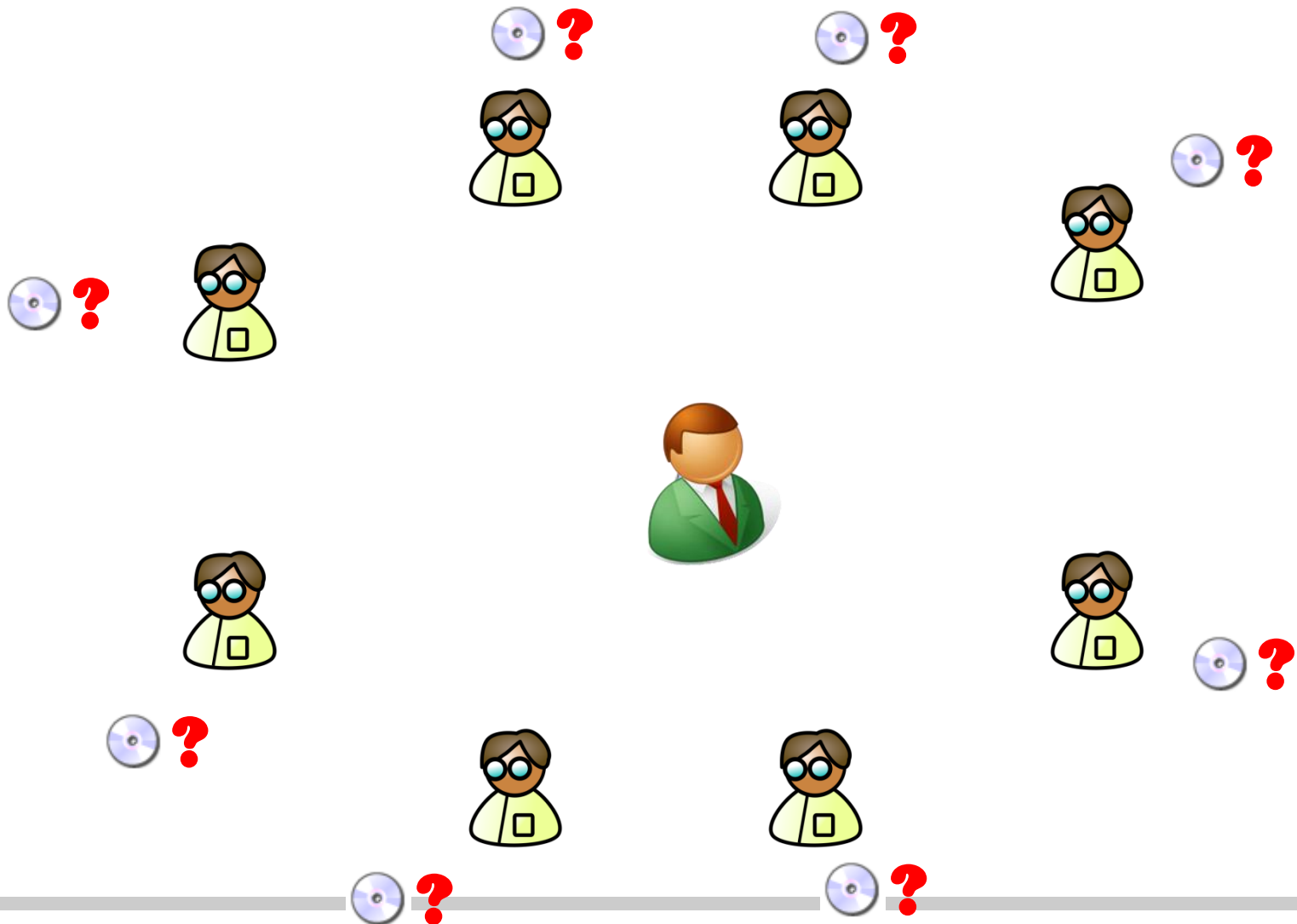
The Challenge Before Us…

- Develop data access methods that achieve the often conflicting goals of:

    - Data confidentiality

    - Protecting privacy

    - Maintaining data quality, and

    - Making data more conveniently accessible

Resetting Perceptions

- Fundamental perceptions need to be revisited and adjusted for accessing sensitive data
- Classic dissemination models need to change
- No longer pushing out sensitive data (e.g., via CDs and contracts to "trusted researchers")
- Pulling in trusted researchers through safe access nodes to secure systems
- Ensuring safe outputs / statistical disclosure control

What is the Enclave ?

The Enclave is an environment that allows for secure remote access to confidential microdata.
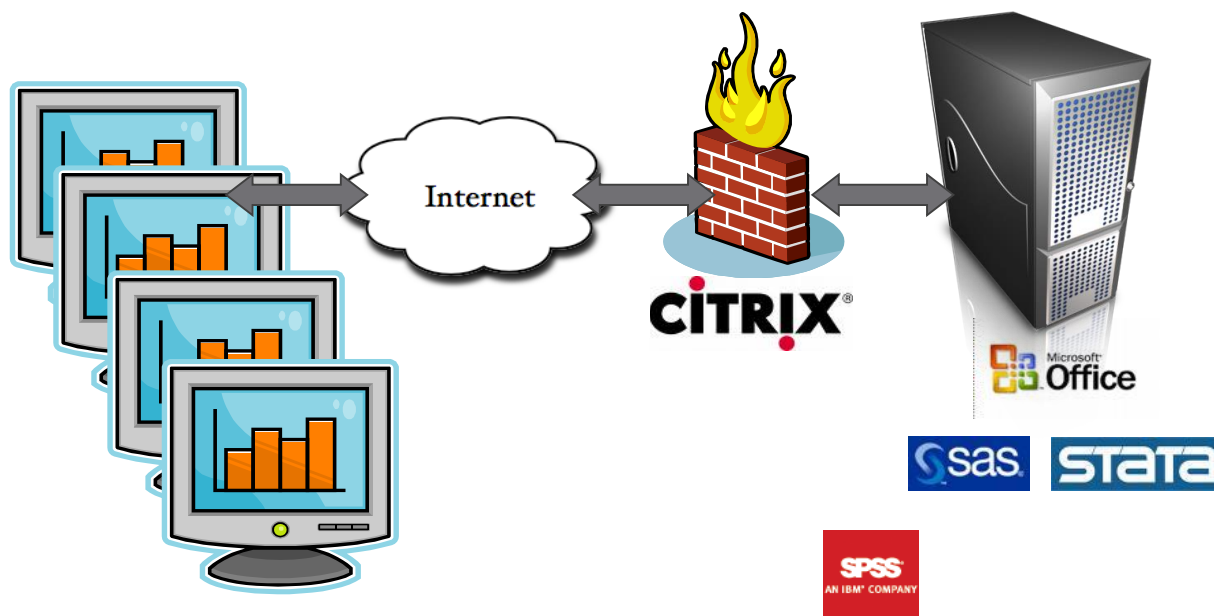
Through the use of a secure terminal VPN session, researchers analyze sensitive data in a secure, convenient and cost-effective manner; data never leave the FISMA compliant secure data center.

Security Compliance

The NORC's Data Enclave IT Security Plan is fully compliant with the Federal Information Security Management Act (FISMA), provisions of mandatory Federal Information Processing Standards (FIPS), and meets all of NIST's IT, data, system, and physical security requirements.

Per the Federal Information Security Management Act and provisions of mandatory Federal Information Processing Standards (FIPS) 199 and NIST Security Plan 800-60, and 800-53 NORC's Data Enclave system impact levels have been determined as follows:

NIST Sensitivity Element	NIST Impact Rating	Short Description of Basis for Rating
Confidentiality	Moderate	The Data Enclave contains information of a proprietary nature. If this proprietary information were disclosed, it could result in serious loss of customer trust for the Federal Statistical Agency.

How Do I Access the Enclave?

Trusted User (with secure credentials)

Trusted Token (second authentication factor)

Trusted Endpoint (Thin Client)

VPN

Virtualization Servers

VPN

Secure Data Storage

Data Access Modalities

The Data Enclave exists along a continuum of access modalities for sensitive data.

- Every data producing organization that seeks to disseminate data must determine what their goals and objectives are with respect to risk tolerance and the desired audience for data.

- In practice this means making decisions about the tradeoff between disclosure risk and analytic utility

Risk-Utility Tradeoff (cont.)

Confidentiality (y-axis)

Analytic Utility (x-axis)

- Statistical Tables and Data Cubes
- Remote Batch Processing
- Synthetic Micro-Data
- Public Use Data-File
- Physical and/or Remote Access Data Enclaves
- Licensing
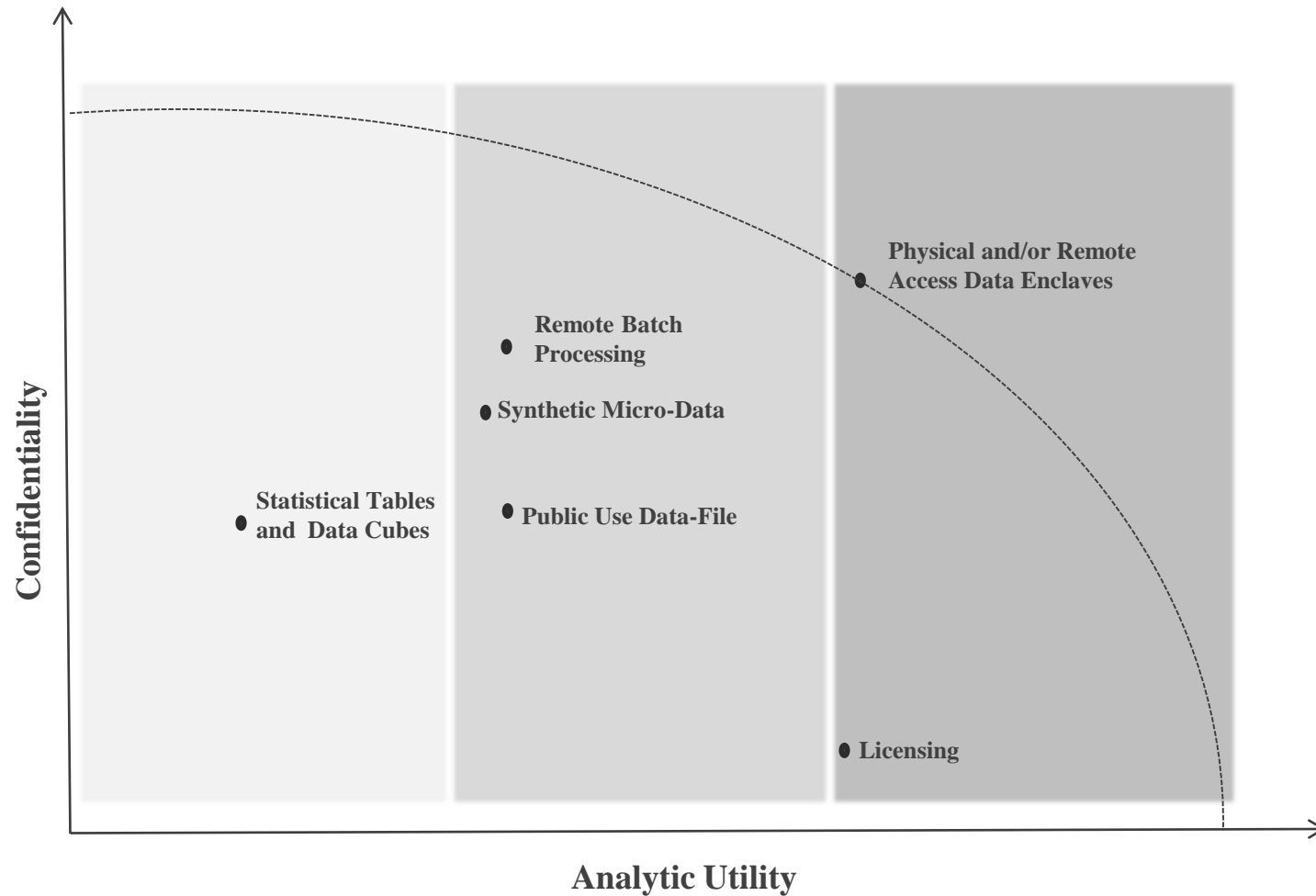
RDC vs. Remote Access

- Data security – the ability to control disclosure risk, ensure privacy, and thus maintain data confidentiality
- Both RDCs and data enclaves allow secure microdata access: similar level of data analytic utility
  - **RDC**: researchers physically access data stored at a secure physical facility
  - **Data Enclave**: researchers remotely access data stored at a file server through a secure system on a virtualized environment
- Both modalities provide high confidentiality protection: information inflow & outflow are monitored and controlled.

Data Access: Convenience

- Confidentiality and utility are not the only factors that influence the choice of data access modality
- The third factor: Convenience
- Producers' perspective:
  - How costly is it to implement an RDC or enclave?
  - How easy is it to update and document the data?
  - How easy is it to monitor researchers' work and output requests?
- Researchers' perspective:
  - How far do they need to travel to the nearest RDC?
  - How easy is it for them to conduct follow-up work?
  - How quickly does the RDC review and approve output requests?
  - How easy is it for them to seek assistance?
  - Is there any peer-to-peer researcher interaction?

Value-Add: Remote Access

The Ideal System
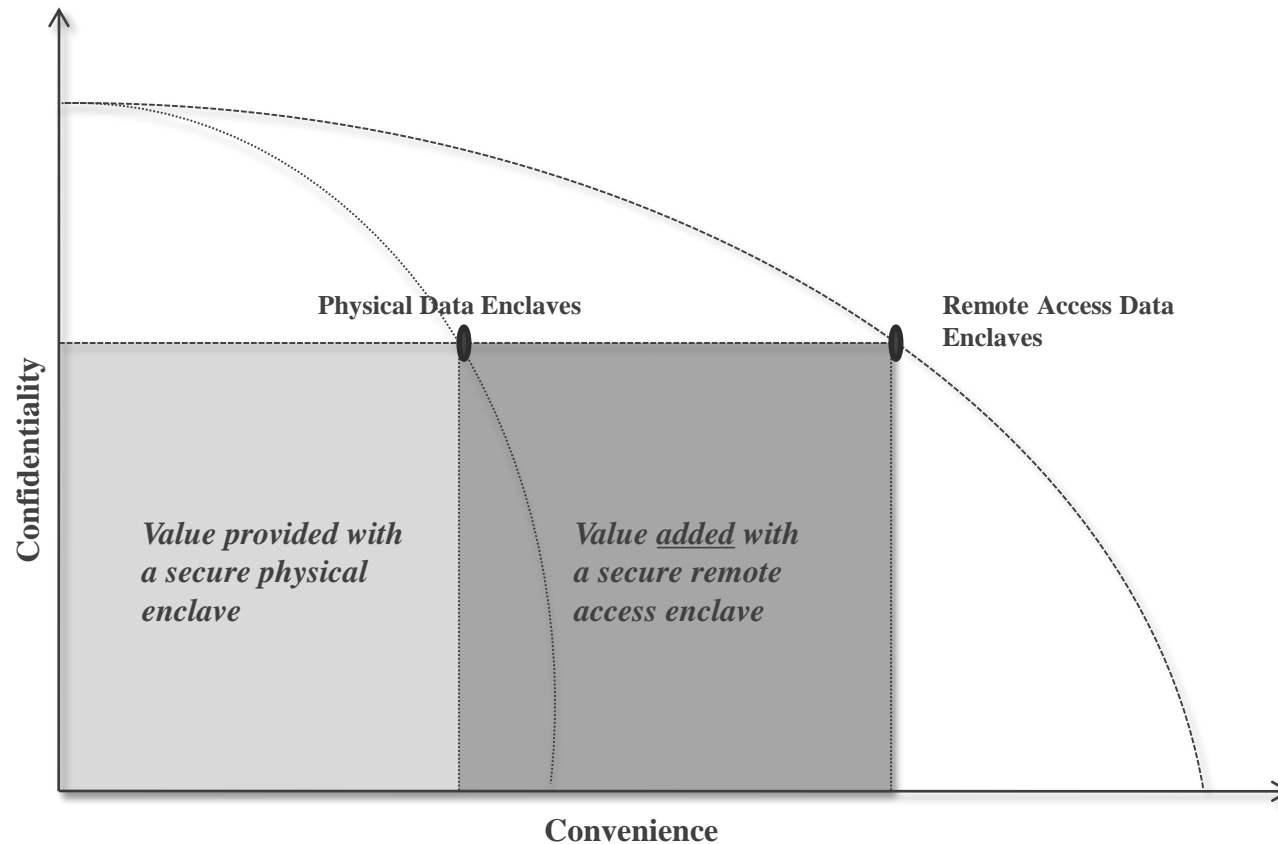
Objectives of a Modern Data Access Facility:

- Secure
- Flexible
- Low Cost
- Convenient
- Meet Replication standard
- Metadata documentation / data lifecycle

Holistic Security Protocol

+Safe Projects

Must have institutional approval and backing

+Safe People

Trained Researchers

+Safe Setting

Data and processing housed in secure network

Users access environment over secure connection

Access limited to secure machines

+Safe Outputs

Strict disclosure review of all exports

_____

= Safe Use

Statistical Applications

Statistical Applications:

‣ SAS

‣ Stata

‣ SPSS

‣ R

‣ MATLAB

‣ LISREL

‣ NLOGIT

Productivity Tools

‣ StatTransfer

‣ Microsoft Office (with Office Communicator)

‣ IHSN Microdata Management Toolkit

‣ Notepad ++

‣ Ultraedit

Streaming Applications

Data Analyses
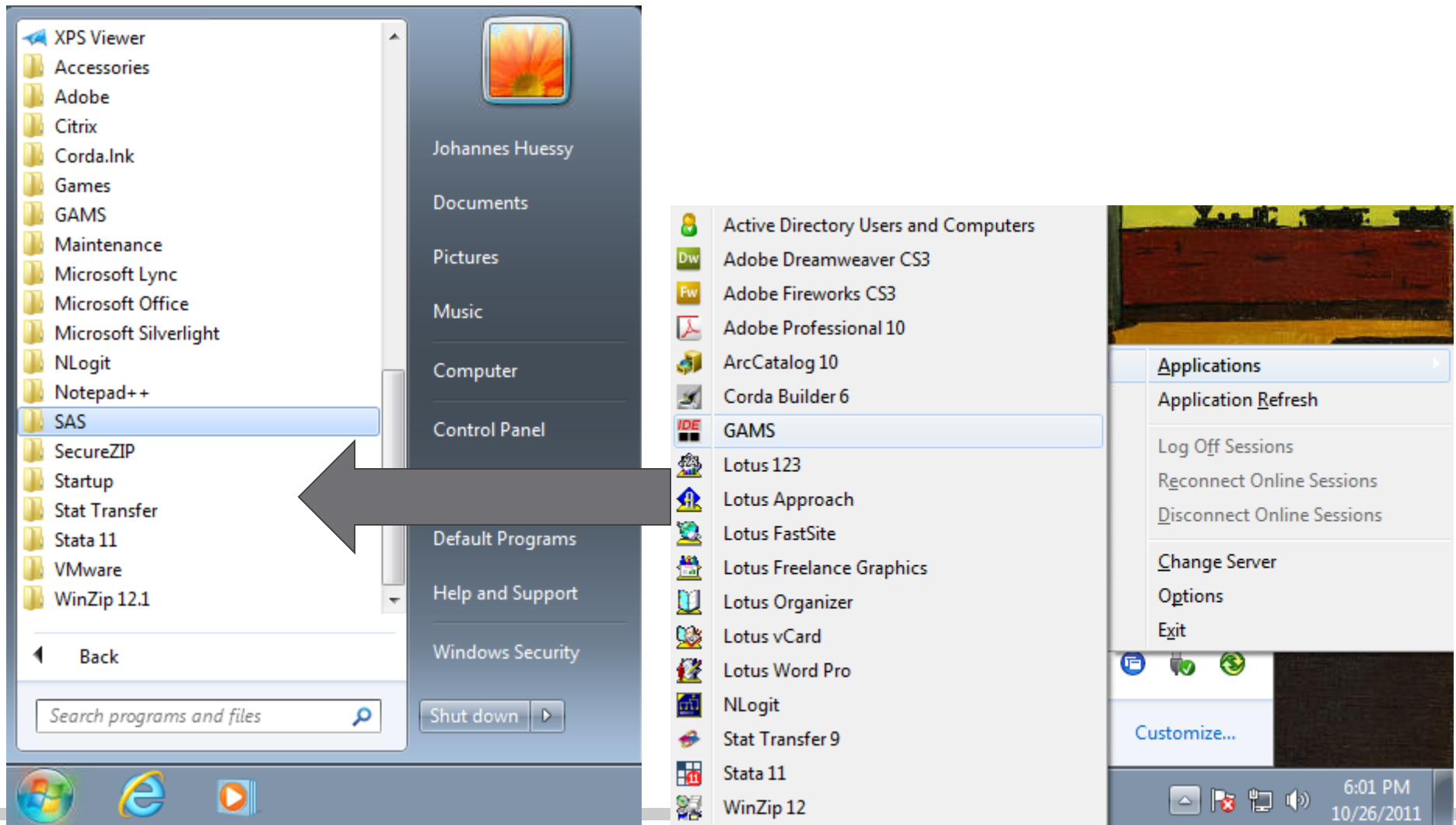
Data Queries Run on Advanced Computational Engines

- As the size and complexity of the data grows, a straightforward virtual desktop infrastructure can become inefficient. Advanced data engines are necessary to provide adequate functionality:
  - Parallel Processing
  - Advanced Databases
  - Tabulation Engines
  - Extraction Tools

Efficient Access

- Less time spent waiting for analyses to complete

- More time available for interpretation

- Increased publication quality and volume potential

Big Data Solution

Output Control

Secure Lab

Data | Work Area

Researcher Logs in

Exports/Output

Imports/Input

Disclosure Review

Online transfer site
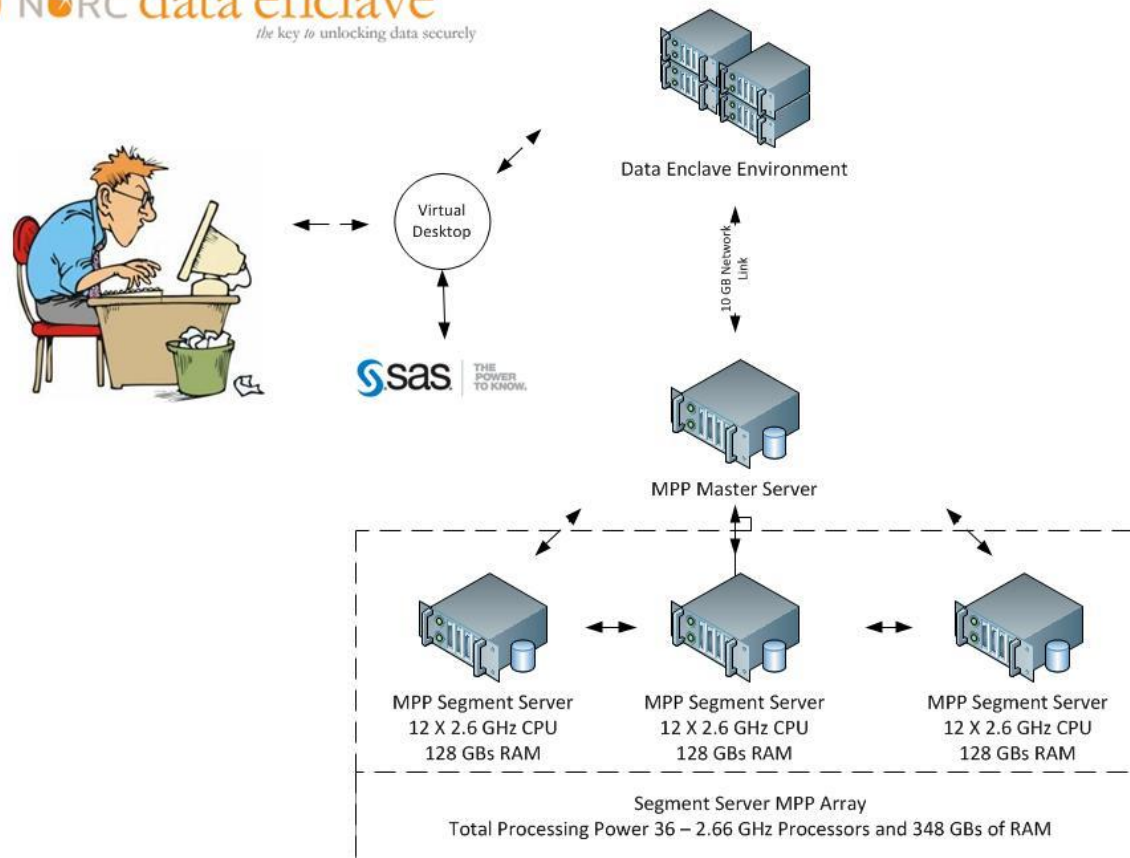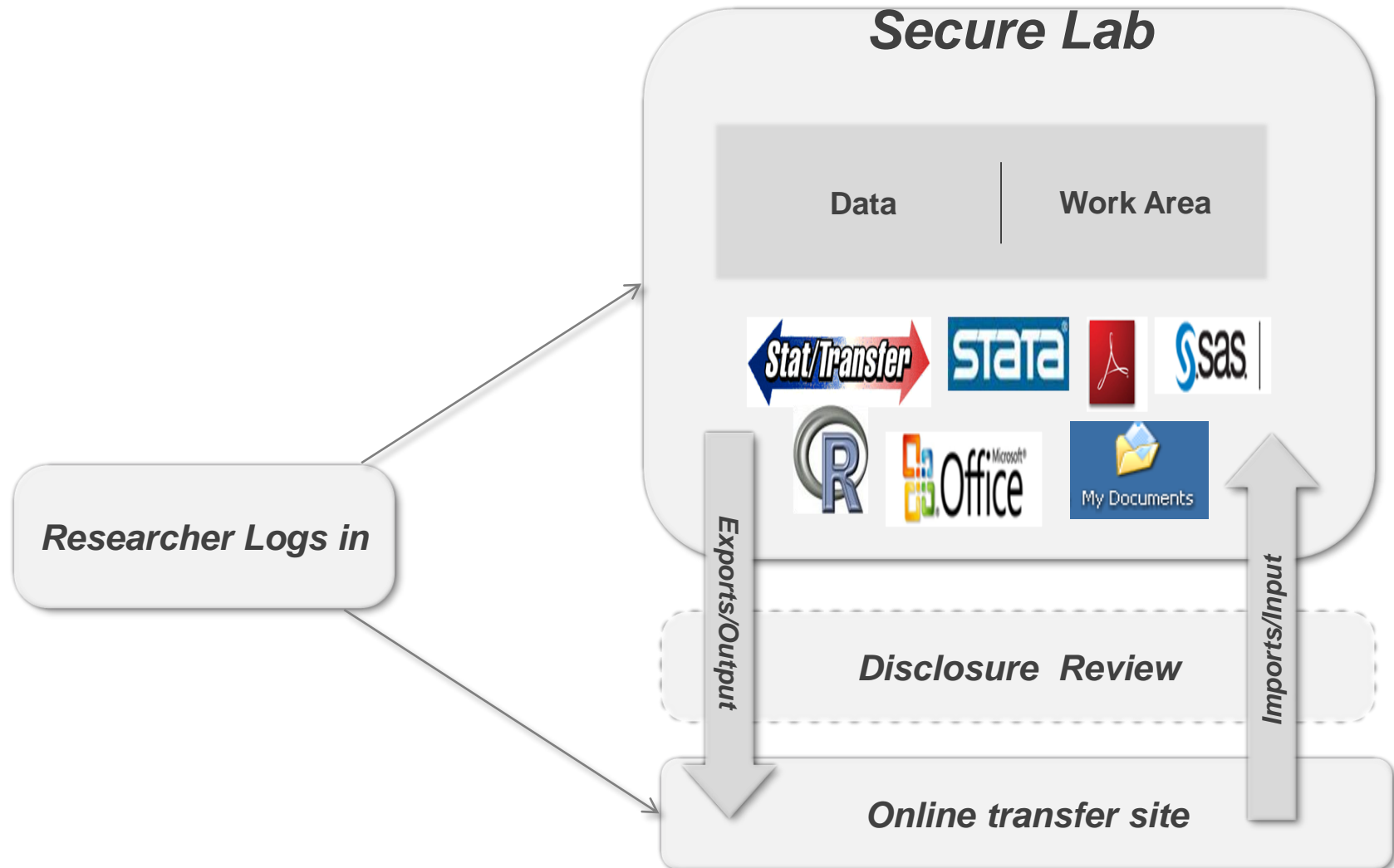
Disclosure Review

- All research results must be reviewed for disclosure risk and approved by NORC disclosure analysts before the results can be shared with approved researchers.

- NORC performs extensive disclosure analysis on all output and makes recommendation to producer
  - Primary disclosure
  - Secondary disclosure
  - Residual disclosure

Accellion File Transfer Site

For exports out of the environment, users are sent a link to approved files which can be downloaded from the Accellion transfer site. All files saved on the Accellion transfer site are encrypted at rest and in transit.
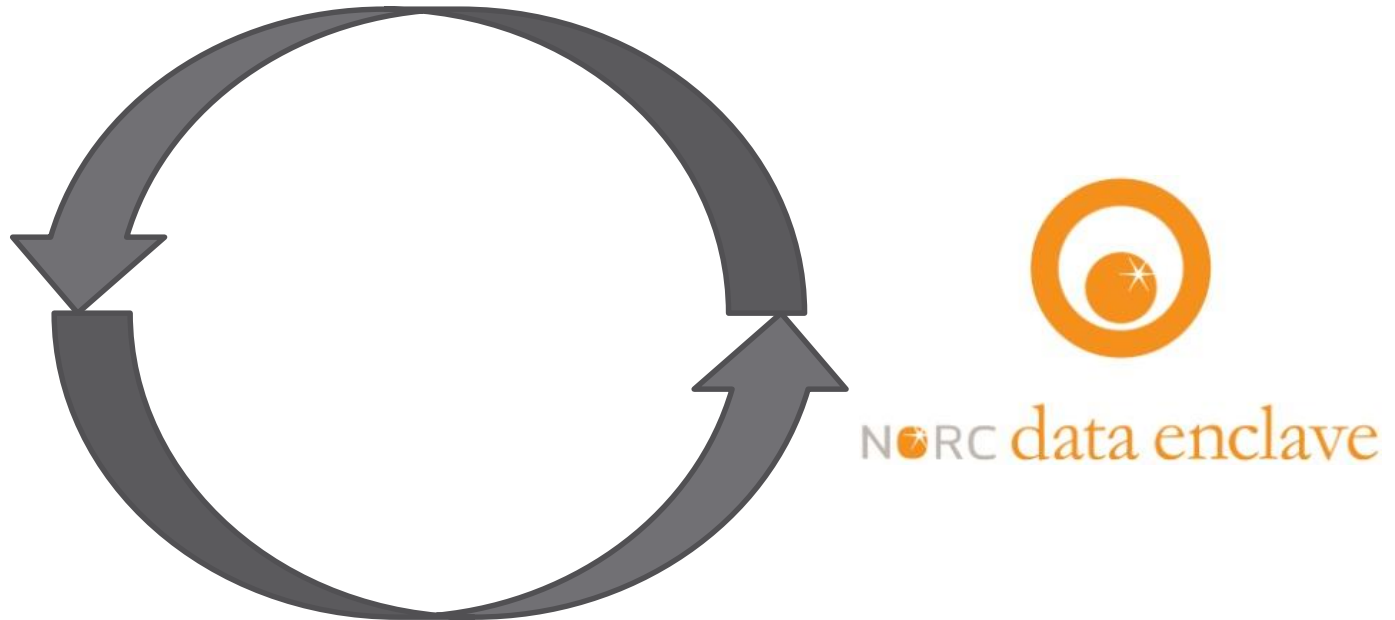
For imports into the environment, users need to request an import workspace from Enclave staff. Once this workspace has been created users can upload files and notify Enclave staff by email for transfer. Small files under 5MB can be emailed directly to dataenclavemanager@norc.org

Researcher Collaboration

PRODUCER PORTAL

GENERAL INFORMATION	KNOWLEDGE SHARING	SUPPORT

• Background info
• Announcements
• Calendar or events
• About
• Topic of the week
	• Discussion groups
	• Wiki
	• Shared libraries
	  • Metadata / Report
	  • Scripts
	  • Research papers
		• Frequently Asked Questions
		• Technical Support
		  • DE usage
		  • Data usage
		  • Quality
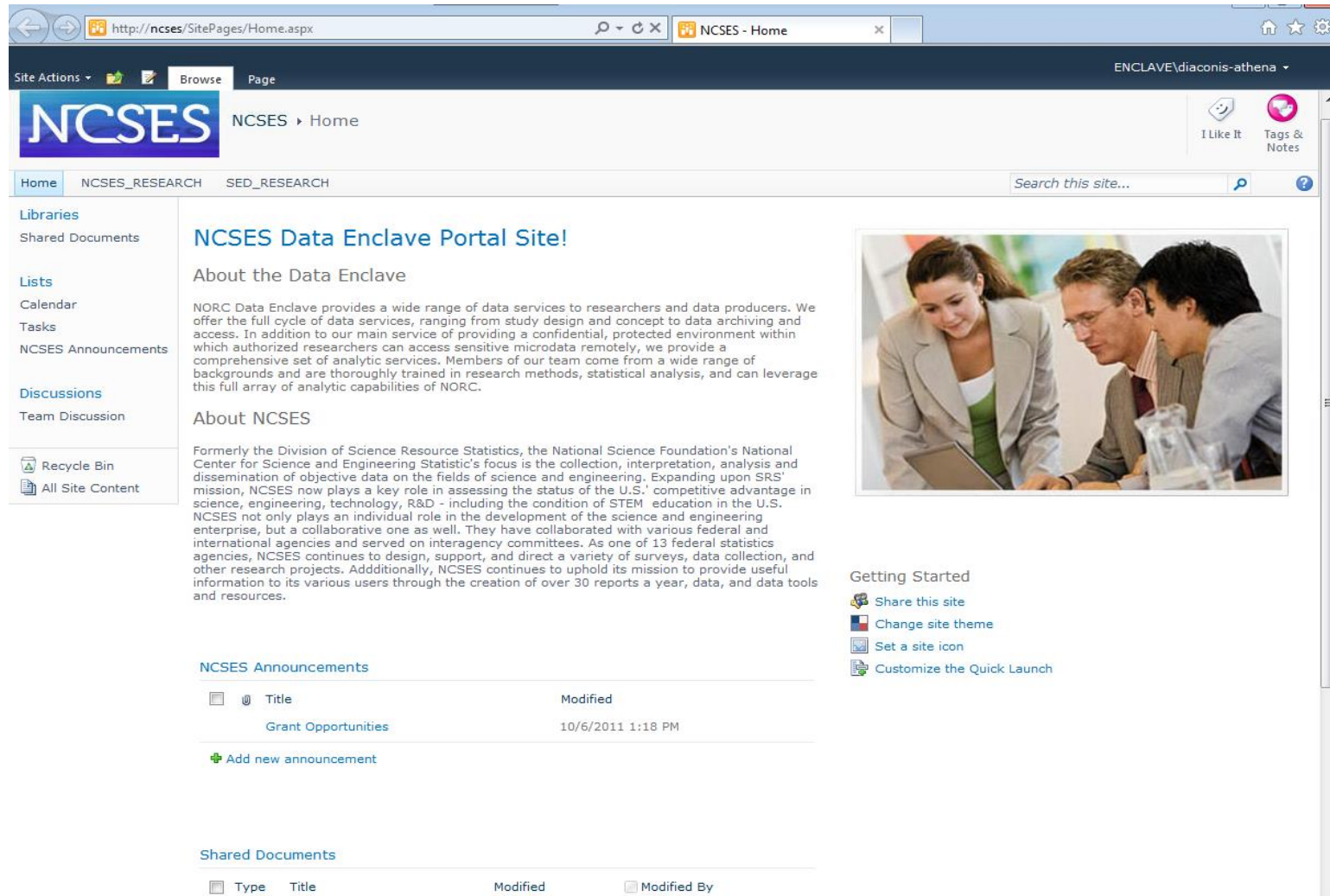
Content fully editable by producers and researchers using a simple web based interface

Private **research group portals** with similar functionalities are configured for each research project

Collaboration Tools

Instant Messaging

Thank You!
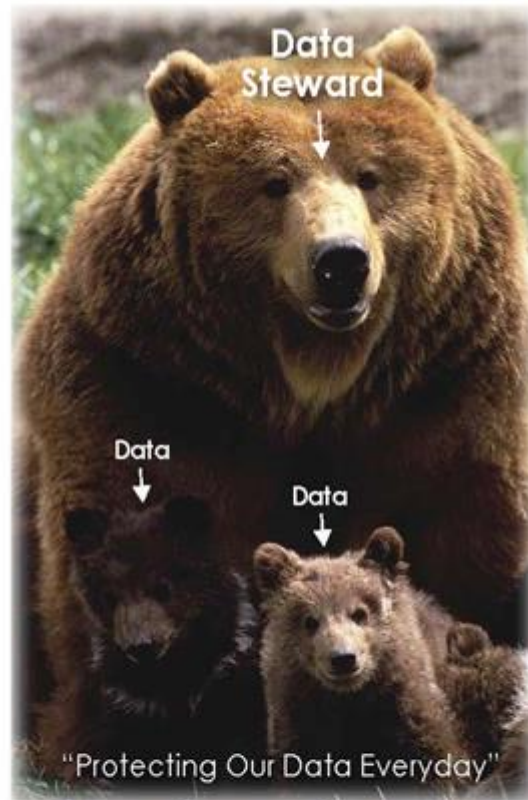
"Protecting Our Data Everyday"

Data Stewardship Week
February 11-15, 2008

USCENSUSBUREAU

NORC
at the UNIVERSITY *of* CHICAGO

insight for informed decisions™