ICPSR

Researcher Passport & Virtual Data Enclave

Maggie Levenstein

ICPSR Director

Statistics Canada

August 21, 2019

ICPSR

Founded in 1962 by 22 universities, now consortium of ~800 institutions world-wide

Focus on social and behavioral science data

Current holdings

- 11,000 studies, quarter million files
- 1500 are *restricted studies*, almost always to protect confidentiality
- Bibliography of Data-related Literature with 80,000 citations

Approximately 60,000 active MyData ("shopping cart") accounts

Thematic collections of data about addiction and HIV, aging, arts and culture, child care and early education, criminal justice, demography, health and medical care, and minorities

Differentiated access to restricted data

- Goal: Protect confidentiality
  - Of individuals, groups, organizations
- Implementation can rely on safe data, safe places, safe people
  - But you don't need all three
- Differentiated access leverages
  - Different data
  - Different researchers
  - Different computing environments

Alternative access modalities

Encrypted download

- ✓ Approved, secure local computing environment
- ✓ Researcher conducts disclosure review
- ✓ Restricted data use agreement with sanctions

Virtual Data Enclave (VDE)

- ✓ Data stays on secure server
- ✓ Access from approved, local computing environment
- ✓ Third party disclosure review
- ✓ Restricted data use agreement with sanctions

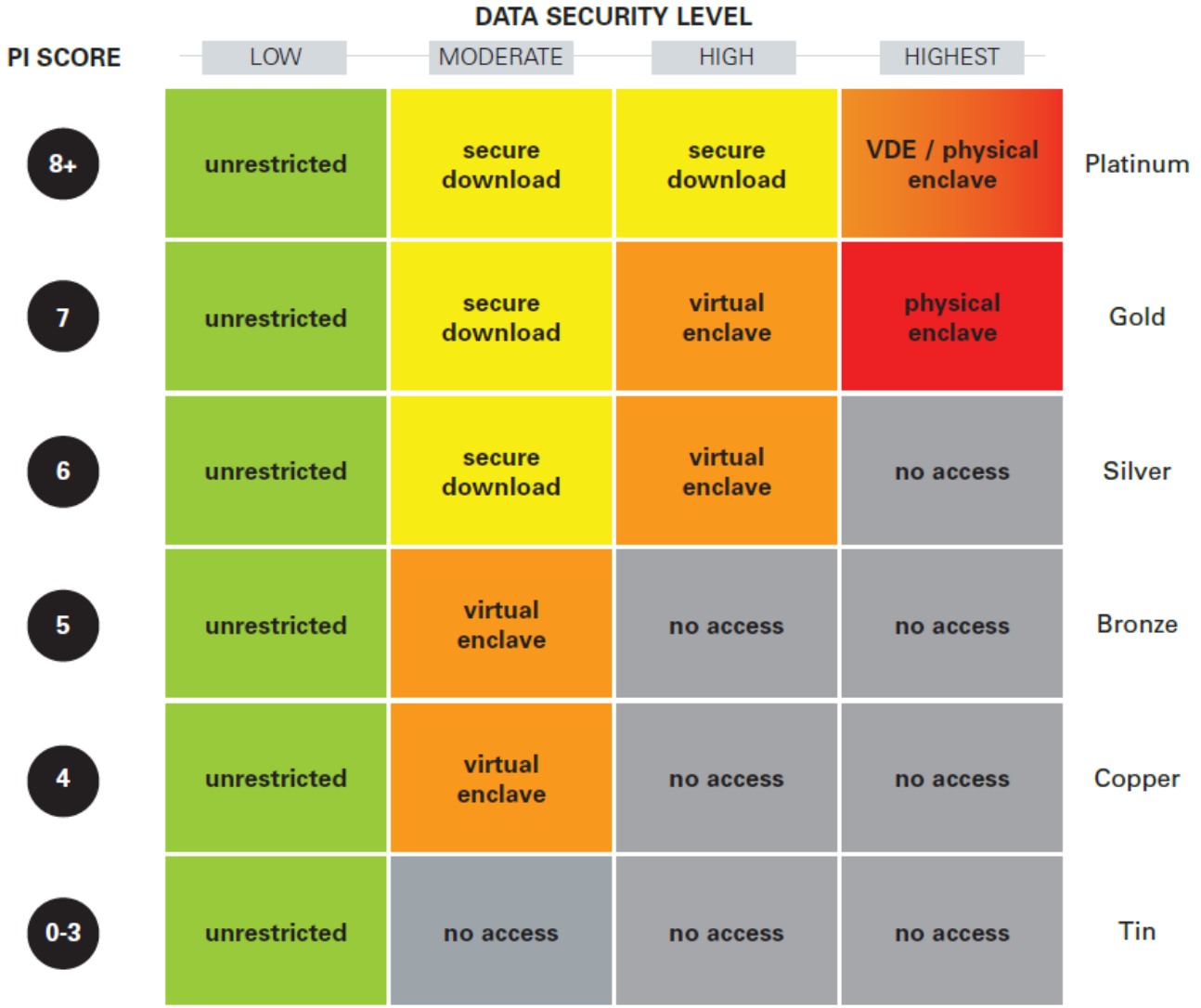
Physical Data Enclave (PDE)

- ✓ Access from approved, secure computing environment
- ✓ Third party disclosure review
- ✓ Restricted data use agreement with sanctions

Secure online analysis

- ✓ Data stays on secure server
- ✓ Automatic disclosure review

Integrating access modalities with researcher credentials



Virtual Data Enclave

Data stays on server in secure data center

- Behind secure firewall

Computation and analysis conducted on virtual machine

- Behind secure firewall
- Analytical software must be on virtual machine
- No internet access
- No access to local machine

Restricted data use agreement

- Institutional and researcher signature
- Explicit sanctions

Virtual Data Enclave

Third party review of all output to avoid disclosure

- Before output sent to researcher outside VDE environment

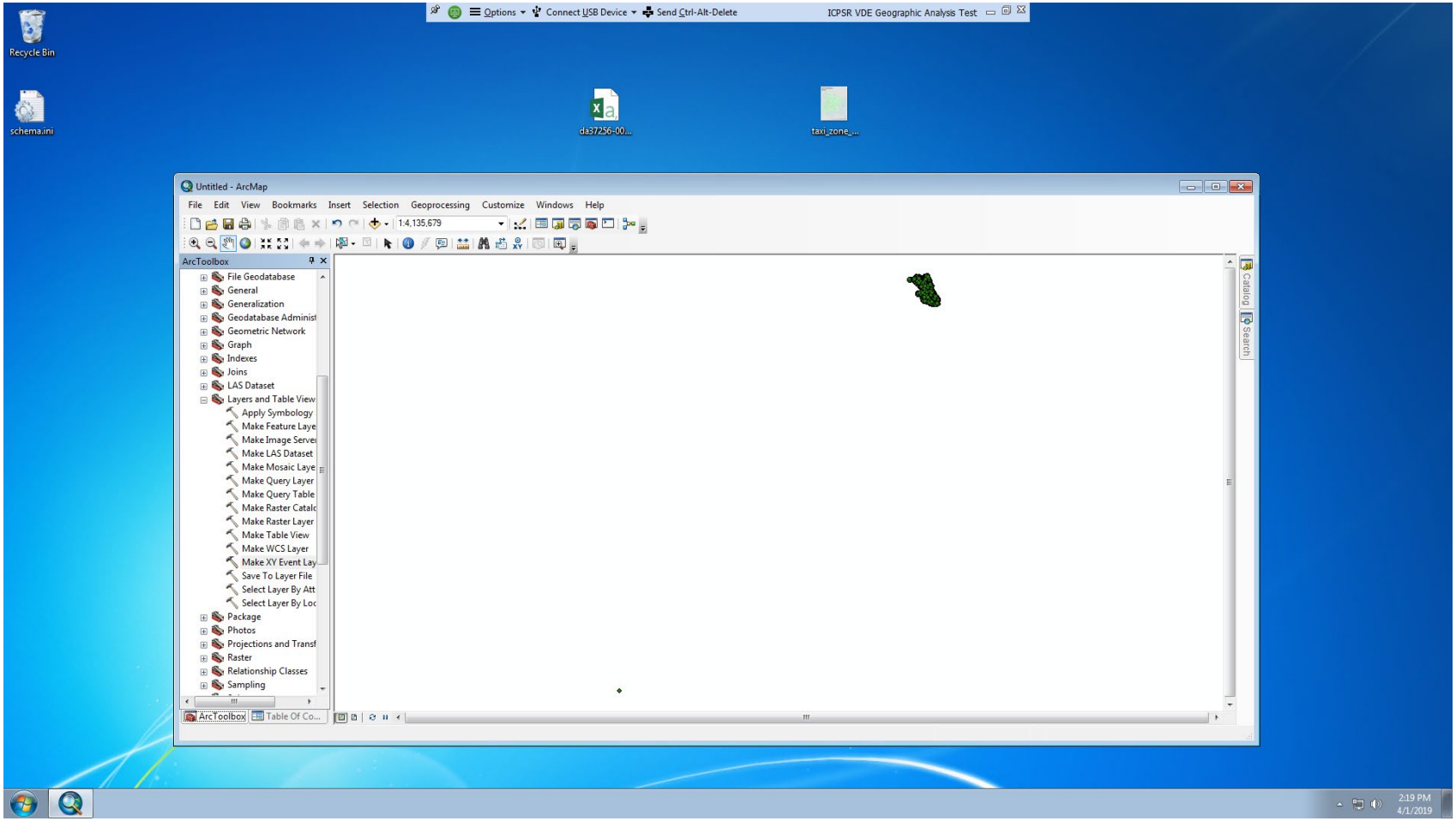
Requires training and quiz

- Establish culture of confidentiality

Research goal: automate disclosure

Mask output, not underlying data

ICPSR VDE windows environment



FSRDC VDE Pilot

Michigan Federal Statistical Research Data Center

- Longstanding physical enclave collaborating with US Census Bureau
- Will use VDI technology to allow researchers with approved projects to connect to FSRDC servers at US Census Bureau from approved locations outside a physical RDC
  - Relatively low risk datasets from US Census Bureau
  - Other statistical agencies invited to participate
- Will launch this fall

Researcher Passport

Researcher Passport: Improving Data Access and Confidentiality Protection

- ICPSR's Strategy for a Community-normed System of Digital Identities of Access
- https://deepblue.lib.umich.edu/handle/2027.42/143808
- Identifies inconsistent language and policies that impede access

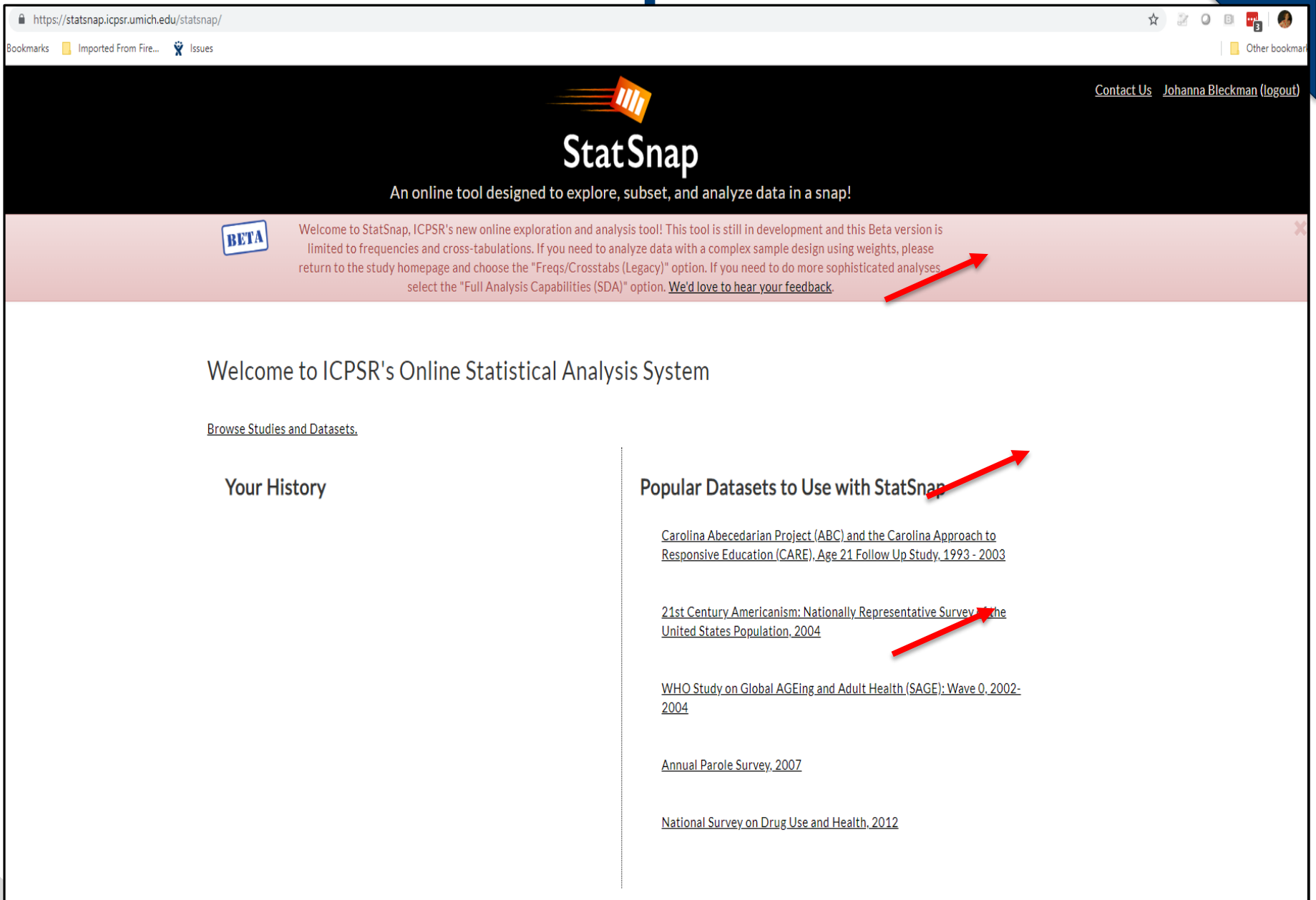
Passports for safe people

- Verified identities, institutional affiliation
- Training
- Experience (good and bad)

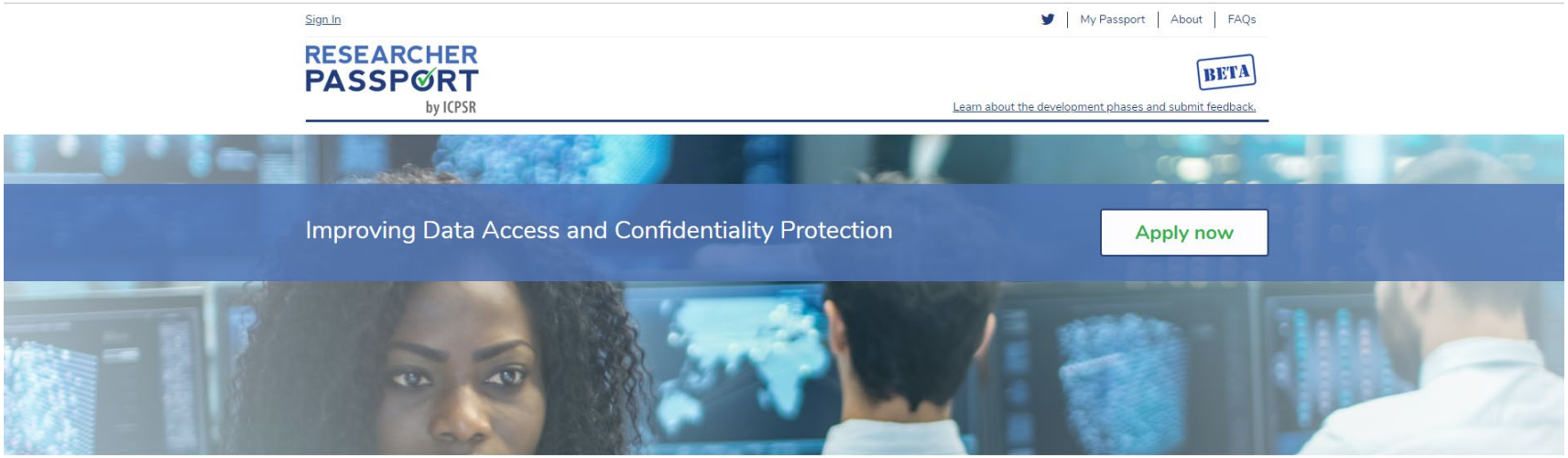
Visas to control access

- Permission to "enter" (access) specific data specifying
  - Passport holder
    - Project, Place, Period

Researcher Passport and Radius



Researcher Passport



Establishing shared understanding across repositories of what it means to be a trusted researcher

Researcher passport: next steps

- Digital identifiers associated with ICPSR MyData accounts
  - Open badges
  - Integrated into VDE application process
- Repositories issue visas to access restricted data
  - Visa controls access to secure download or VDE
    - Visa issued after signed DUA
- Establishing content of training requirements
- Governance system to provide due process for allegations of misconduct

Maintaining datasets to support the data linkage community

maggiel@umich.edu
LINKAGE LIBRARY
Upload Data
Browse Data
About
Enter search term(s)
SEARCH
LINKAGE LIBRARY
Maintaining datasets to support the data linkage community
What is Linkage Library
Comment
Comment on projects
Projects
Some have data, Some have code, Some have both
GitHub
Interface with GitHub

LINKAGE LIBRARY

Encourage researchers to share linked (or linkable) data and linkage strategies

- Algorithms, code, documents

Compare approaches across projects, datasets, disciplines

- Improve linkage practices
- Improve transparency
- Commenting – to improve linkage & build community

LinkageLibrary.org

ICPSR

This is DataJeff. He loves data.