ICPSR

Inter-university Consortium for Political and Social Research

Archiving and Sharing Confidential Data in the Social Sciences

George Alter

Director, ICPSR

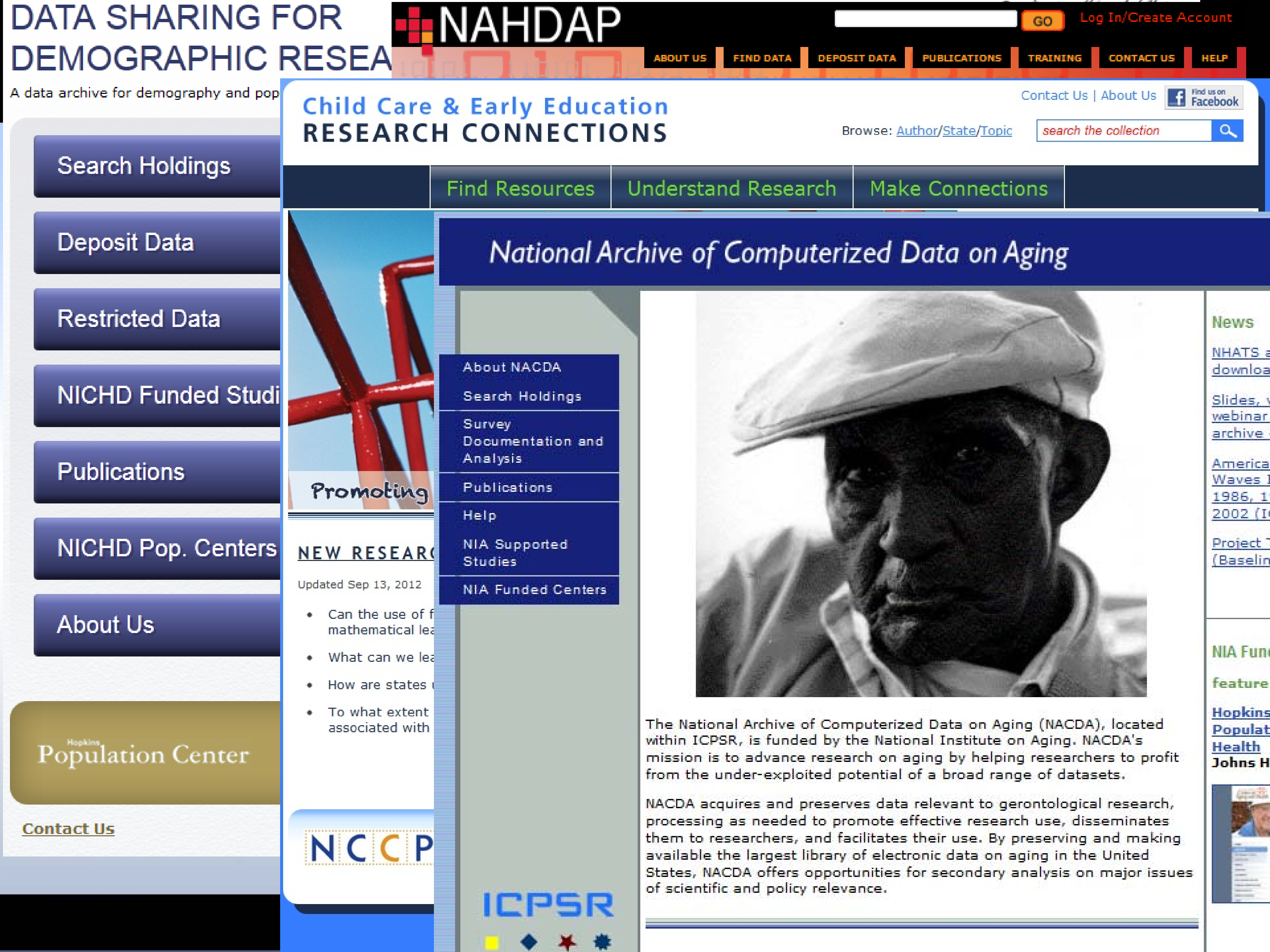
About ICPSR

- Established in 1962 to share the American National Election Studies
  - Partnership of 21 universities
- Today: More than 700 members
  - ~400 U.S. institutions
  - 46 national memberships
- 8,000 data collections
- Data available 24/7 for download and online analysis

Mission: ICPSR provides leadership and training in data access, curation, and methods of analysis for a diverse and expanding social science research community.

What we do

- Acquire and archive social science data
- Distribute data to researchers
- Preserve data for future generations
- Provide training in quantitative methods



NEW RESEARC

Updated Sep 13, 2012

- Can the use of f mathematical lea
- What can we lea
- How are states
- To what extent associated with

The National Archive of Computerized Data on Aging (NACDA), located within ICPSR, is funded by the National Institute on Aging. NACDA's mission is to advance research on aging by helping researchers to profit from the under-exploited potential of a broad range of datasets.

NACDA acquires and preserves data relevant to gerontological research, processing as needed to promote effective research use, disseminates them to researchers, and facilitates their use. By preserving and making available the largest library of electronic data on aging in the United States, NACDA offers opportunities for secondary analysis on major issues of scientific and policy relevance.

Sharing Confidential Data

- Disclosure risks in social science data
- Protecting confidential data
  - Safe data
  - Safe places
  - Safe people
  - Safe outputs

Disclosure risks in social science data

Research designs can increase disclosure risk :

- Geographically referenced data
- Longitudinal data
- Multi-level data:
  - Student, teacher, school, school district
  - Patient, clinic, community

Protecting Confidential Data

- **Safe data**: Modify the data to reduce the risk of re-identification
- **Safe places**: Physical isolation and secure technologies
- **Safe people**: Training and Data use agreements
- **Safe outputs**: Results are reviewed before being released to researchers

Safe data

Disclosure risks can be reduced by design:

- Multiple sites rather than single locations
- Keeping sampling locations secret
  - Releasing characteristics of contexts without providing locations
- Responsive sampling procedures

Safe data

Data masking

- Grouping values
- Top-coding
- Aggregating geographic areas
- Swapping values
- Suppressing unique cases
- Sampling within a larger data collection
- Adding “noise”
- Replacing real data with synthetic data

Safe places

- Data protection plans
- Remote submission and execution
- Virtual data enclave
- Physical enclave

Data Protection Plans should address risks:

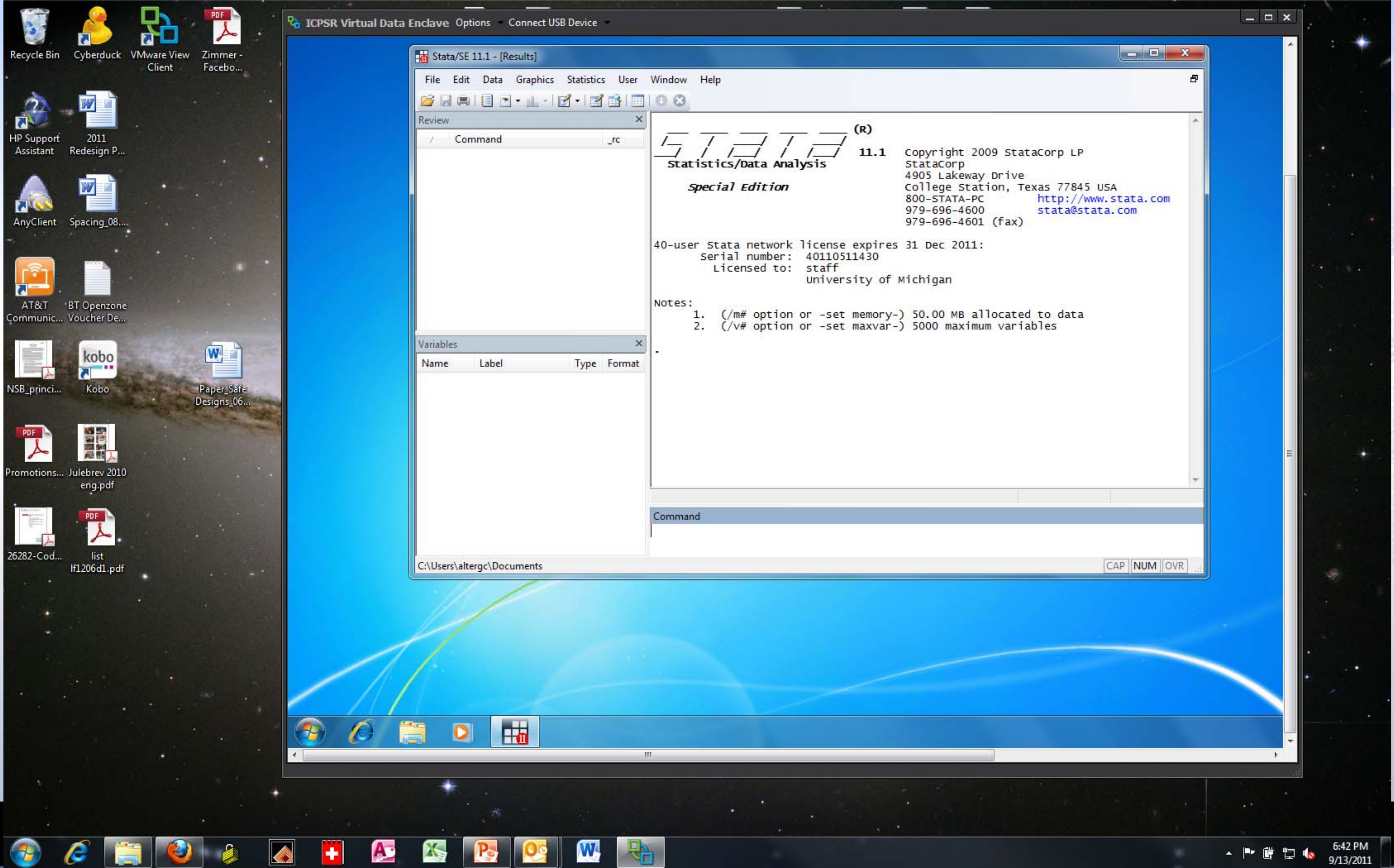
- **unauthorized use** of account on computer
- **computer break-in** by exploiting vulnerability
- **hijacking** of computer by malware or botware
- **interception** of network traffic between computers
- **loss** of computer or media
- **theft** of computer or media
- **eavesdropping** of electronic output on computer screen
- **unauthorized viewing** of paper output

We often focus too much on technology and not enough on risk.

Controlled environments allow review of outputs

- **Remote submission and execution**
  - User submits program code or scripts, which are executed in a controlled environment
- **Virtual data enclave**
  - Remote desktop technology prevents moving data to user’s local computer
  - Requires a data use agreement
- **Physical enclave**
  - Users must travel to the data

Virtual Data Enclave

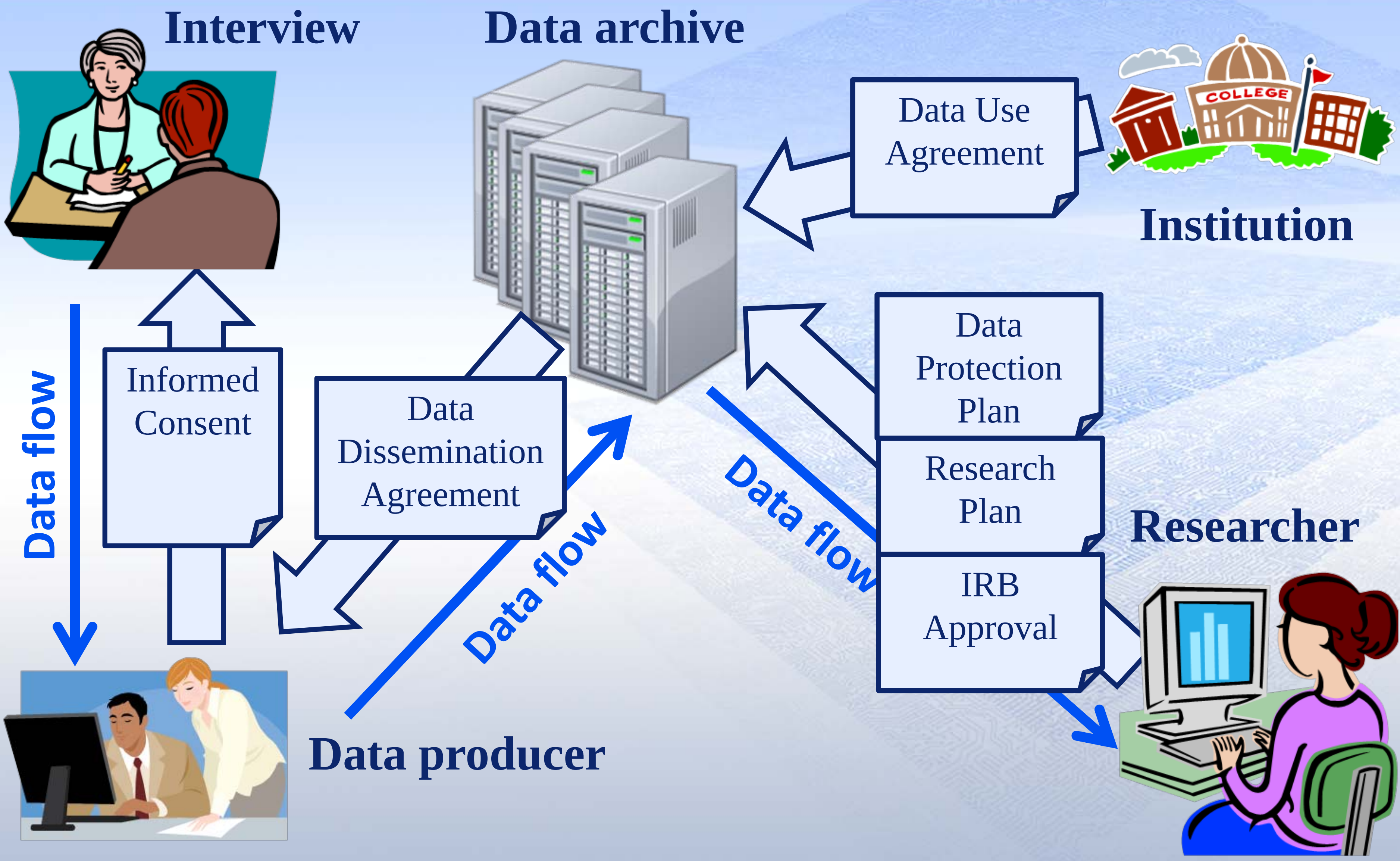


```
                                    (R)
                              11.1   Copyright 2009 StataCorp LP
  Statistics/Data Analysis           StataCorp
                                     4905 Lakeway Drive
     Special Edition                 College Station, Texas 77845 USA
                                     800-STATA-PC        http://www.stata.com
                                     979-696-4600        stata@stata.com
                                     979-696-4601 (fax)

40-user Stata network license expires 31 Dec 2011:
       Serial number:  40110511430
         Licensed to:  staff
                       University of Michigan

Notes:
      1.  (/m# option or -set memory-) 50.00 MB allocated to data
      2.  (/v# option or -set maxvar-) 5000 maximum variables

.
```

C:\Users\altergc\Documents

Safe people

- Data use agreements
- Training



Interview

Data archive

Data Use Agreement

Institution

Data flow

Informed Consent

Data Dissemination Agreement

Data flow

Data Protection Plan

Research Plan

IRB Approval

Data flow

Researcher

Data producer

Safe people

- Parts of a data use agreement at ICPSR
  - Research plan
  - IRB approval
  - Data protection plan
  - Behavior rules
  - Security pledge
  - Institutional signature

Data Use Agreement: Behavior rules

To avoid inadvertent disclosure of persons, families, households, neighborhoods, schools or health services by using the following guidelines in the release of statistics derived from the dataset.

1. In no table should all cases in any row or column be found in a single cell.
2. In no case should the total for a row or column of a cross-tabulation be fewer than ten.
3. In no case should a quantity figure be based on fewer than ten cases.
4. In no case should a quantity figure be published if one case contributes more than 60 percent of the amount.
5. In no case should data on an identifiable case, or any of the kinds of data listed in preceding items 1-3, be derivable through subtraction or other calculation from the combination of tables released.

Data Use Agreement: Institutional Commitment

The Recipient Institution will treat allegations, by NAHDAP/ICPSR or other parties, of violations of this agreement as allegations of violations of its policies and procedures on scientific integrity and misconduct. If the allegations are confirmed, the Recipient Institution will treat the violations as it would violations of the explicit terms of its policies on scientific integrity and misconduct.

What are the consequences of violating the terms of use agreement for ICPSR data?

Subjects who participate in surveys and other research instruments distributed by ICPSR expect their responses to remain confidential. The data distributed by ICPSR are for statistical analysis, and they may not be used to identify specific individuals or organizations. Although ICPSR takes steps to assure that subjects cannot be identified, users are also obligated to act responsibly and not to violate the privacy of subjects intentionally or unintentionally.

If ICPSR determines that the terms of use agreement has been violated, one or more of the steps will be taken which may include:

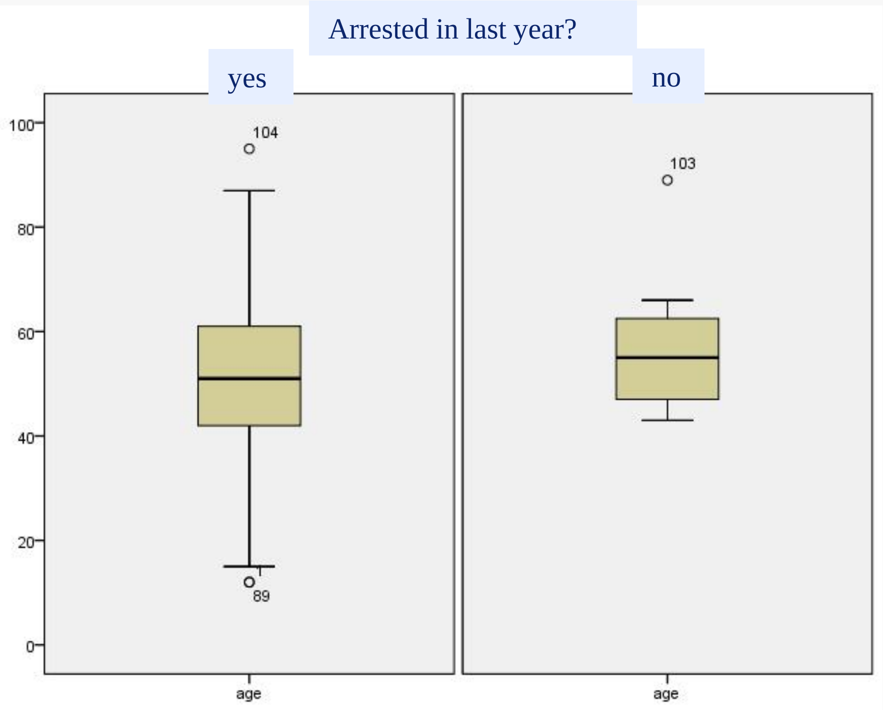
- ICPSR may revoke the existing agreement, demand the return of the data in question, and deny all future access to ICPSR data.
- The violation may be reported to the Research Integrity Officer, Institutional Review Board, or Human Subjects Review Committee of the user’s institution. A range of sanctions are available to institutions including revocation of tenure and termination.
- If the confidentiality of human subjects has been violated, the case may be reported to the Federal Office for Human Research Protections. This may result in an investigation of the user’s institution, which can result in institution-wide sanctions including the suspension of all research grants.
- A court may award the payment of damages to any individual harmed by the breach of the agreement.

Training: Disclosure risk online tutorial

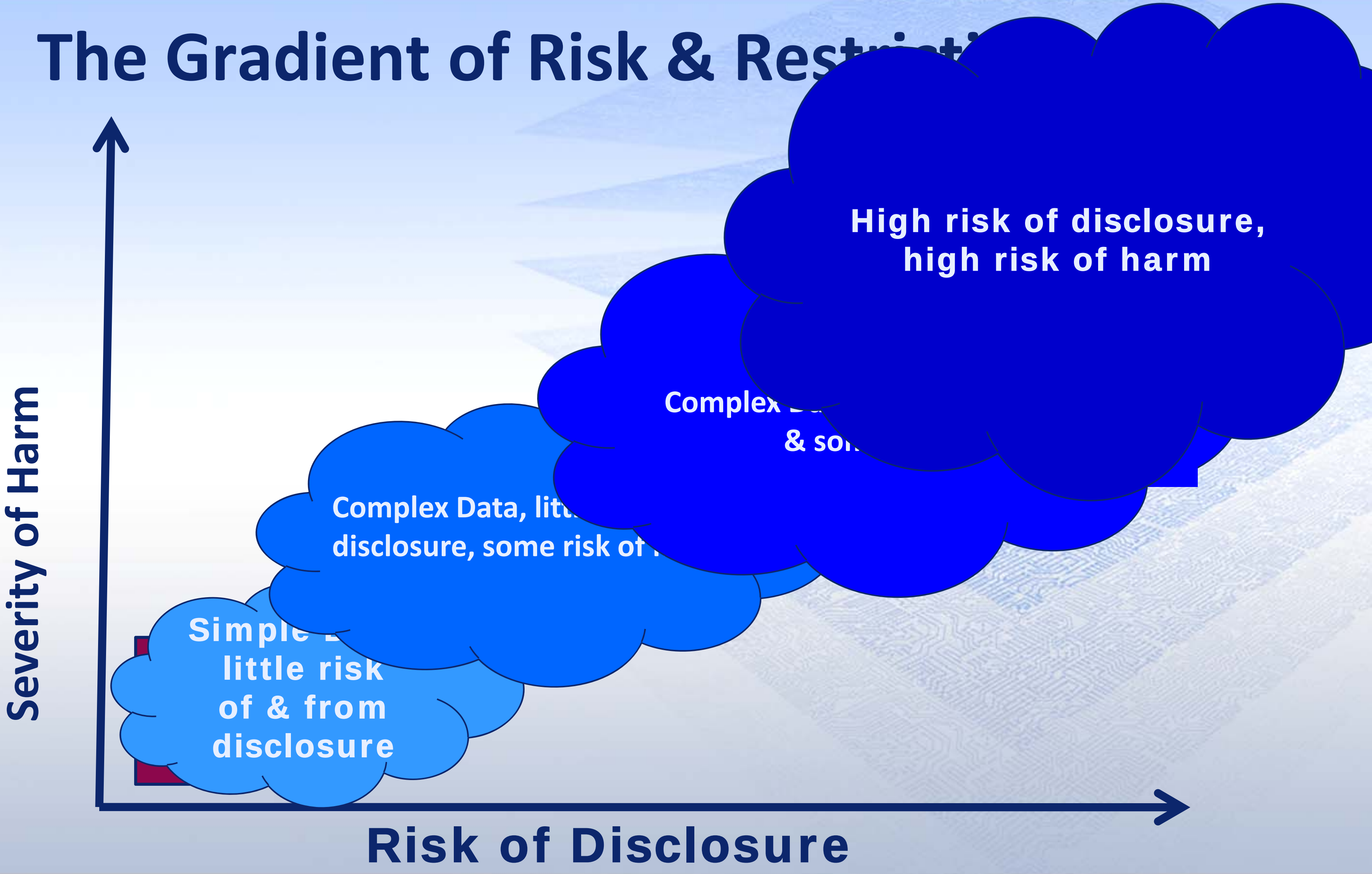
Disclosure: Graph with extreme values example

Data were collected for a sample of 104 people in a county.
Among the variables collected were age, gender, and whether the person was arrested within the last year. Box plots below show the distribution of age, one plot for those arrested and one for those who were not. The number labels are case number in the dataset.
The potential identifiability represented by outlying values is compounded here by an unusual combination that could probably be identified using public records for a county in the U.S. --someone approximately 90 years old was arrested in the sample. Including extreme values is a disclosure risk for identifiability when combined with other variables in the dataset.

N	104
min age	12
max age	95
mean age	51
std dev	15
% female	5.2
% arrested	5.8



The Gradient of Risk & Restricti



High risk of disclosure, high risk of harm

Complex D
& so

Complex Data, litt
disclosure, some risk of

Simple
little risk
of & from
disclosure

The IRB of the data producer should establish a data dissemination plan,

But

- Many IRBs lack expertise in disclosure risk
- Data persist much longer than membership of the IRB
- Centers of expertise in disclosure risk should be available to advise and succeed IRBs

Data repositories offer secure dissemination services

- Dissemination of confidential data is too burdensome for most projects
- Data security technology is changing rapidly
- Repositories are long-lived institutions

Institutions that receive data are responsible for security and behavior

- IRBs of data recipients should defer to protocols established by previous IRBs
- Institutions are responsible for compliance
- IT security is necessary but not sufficient
- Data users must understand disclosure risks and behave safely
- Data users should pay the costs of access to confidential data

Thank you!

George Alter
altergc@umich.edu
www.icpsr.org