ICPSR

Inter-university Consortium for Political and Social Research

Sharing Confidential Data

George Alter

Director, ICPSR

Disclosure: Risk & Harm

- What do we promise when we conduct research about people?
  - That benefits (usually to society) outweigh risk of harm (usually to individual)
  - That we will protect confidentiality
- Why is confidentiality so important?
  - Because people may reveal information to us that could cause them harm if revealed.
  - Examples: criminal activity, antisocial activity, medical conditions...

What are We Afraid of...

- Direct Identifiers
  - Inadvertent release of unnecessary information (Name, phone number, SSN…)
  - Direct identifiers required for analysis (location, genetic characteristics,…)
- Indirect Identifiers
  - Characteristics that identify a subject when combined (sex, race, age, education, occupation)

Who are We Afraid of?

- Parents trying to find out if their child had an abortion or uses drugs
- Spouse seeking hidden income or infidelity in a divorce
- Insurance companies seeking to eliminate risky individuals
- Other criminals and nuisances
- NSA, CIA, FBI, KGB, SABOT, SBL, SMERSH, KAOS, etc...

Deductive Disclosure

- A combination of characteristics could allow an intruder to re-identify an individual in a survey "deductively," even if direct identifiers are removed.
- Dependent on
  - Knowing someone in the survey
  - Matching cases to a database

Current Survey Designs Increase the Risks of Disclosing Subjects' Identities

- 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:

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

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 analysis
- Virtual data enclave
- Physical enclave

Safe places

Data Protection Plans should address risks:

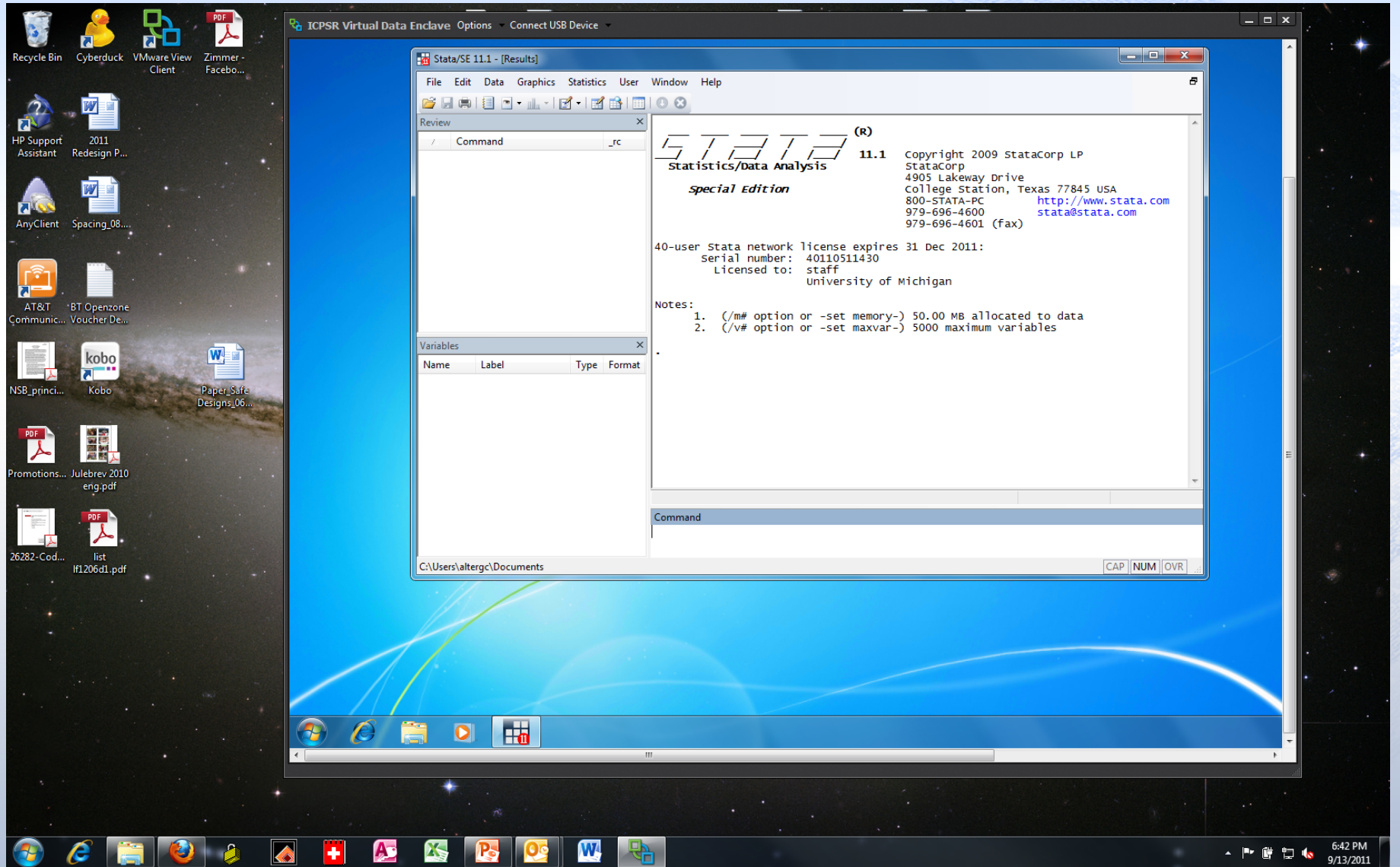
- **unauthorized use** of account on computer
- **computer break-in** by exploiting vulnerability
- **hijacking** of computer by malware or botware
- **interception** of network traffic between computers
- **loss** of computer or media
- **theft** of computer or media
- **eavesdropping** of electronic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

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

Safe places

- **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

The Virtual Data Enclave (VDE) provides remote access to quantitative data in a secure environment.



Safe people

- Data use agreements
- Training

Safe people

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

ICPSR

Inter-university Consortium for Political and Social Research

Interview

Data archive

Data Use Agreement

Institution

Data flow

Informed Consent

Data Dissemination Agreement

Data flow

Data flow

Data Protection Plan

Research Plan

IRB Approval

Researcher

Data producer

ICPSR

Inter-university Consortium for Political and Social Research

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.

ICPSR

Inter-university Consortium for Political and Social Research

Data Use Agreement

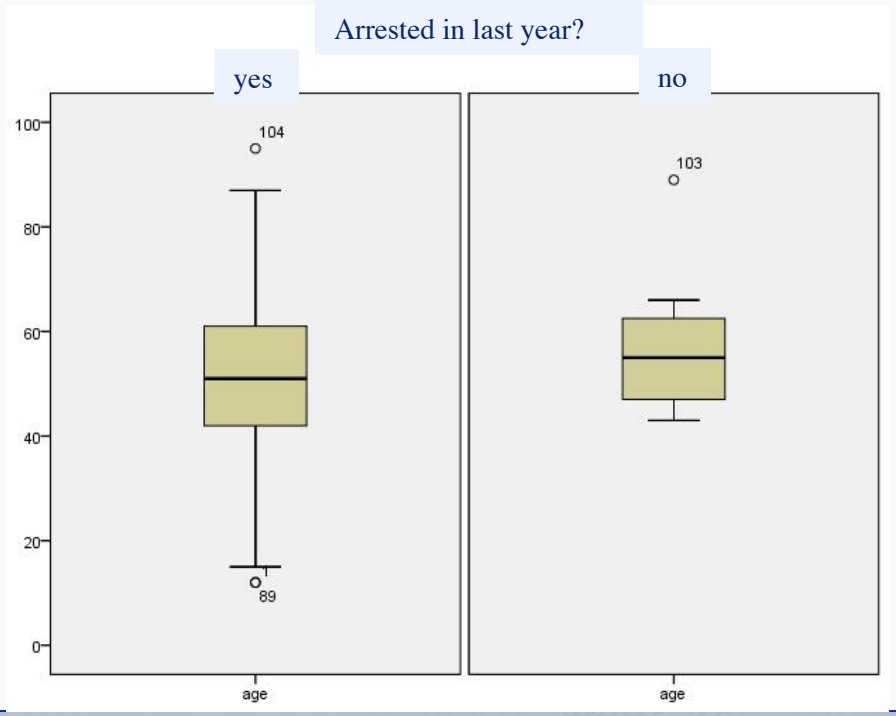
The Recipient Institution will treat allegations, by NAHDAP/ICPSR or other parties, of violations of this agreement as allegations of violations of its policies and procedures on scientific integrity and misconduct. If the allegations are confirmed, the Recipient Institution will treat the violations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

Safe People: 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



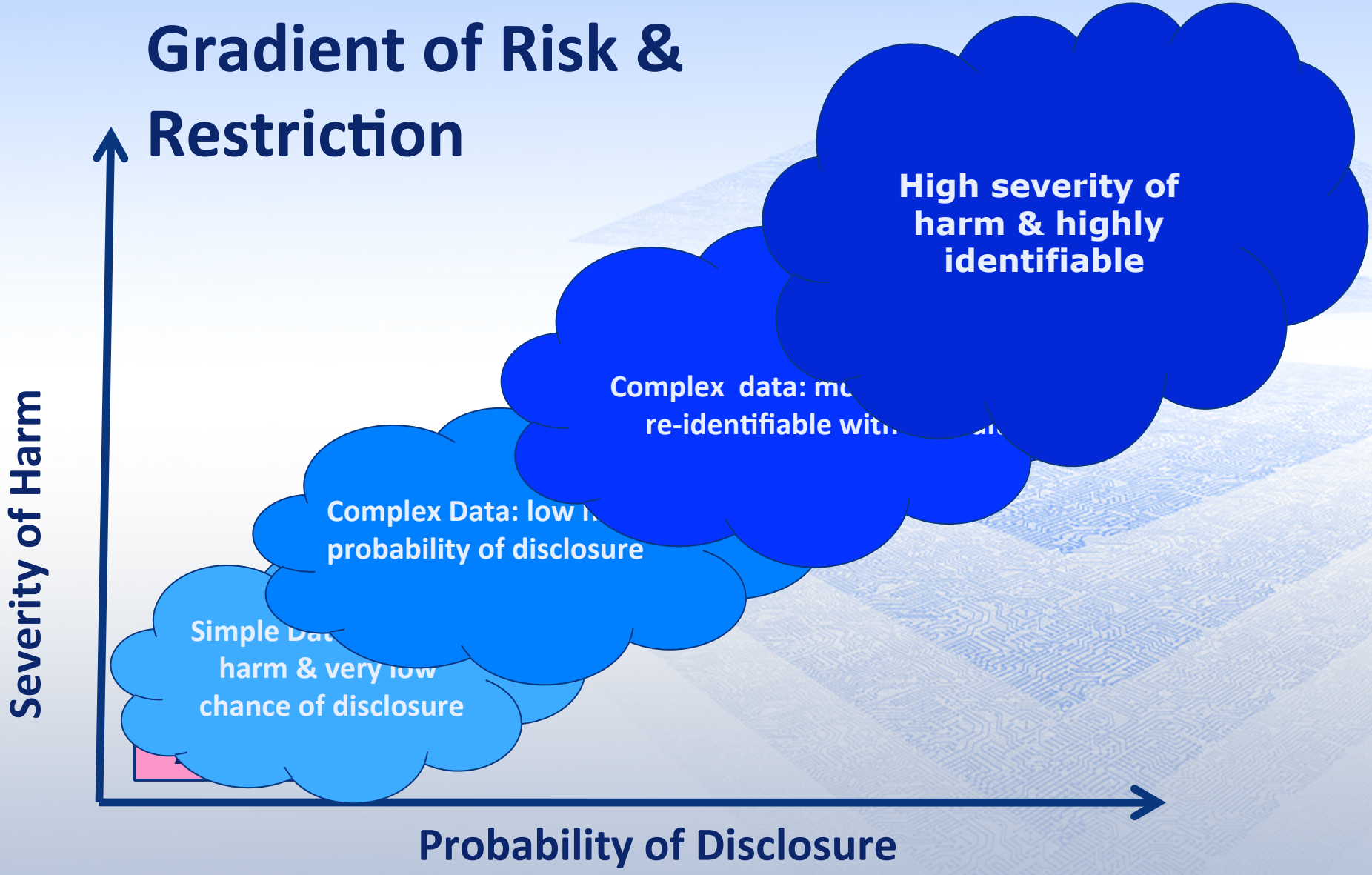
Safe outputs

- Controlled environments allow review of outputs
  - Remote execution systems, Virtual data enclaves, Physical enclaves
- Disclosure checks may be automated, but manual review is usually necessary

ICPSR

Inter-university Consortium for Political and Social Research

Gradient of Risk & Restriction



ICPSR

Inter-university Consortium for Political and Social Research

Thank you