

# NISS

## Microdata Toolkit (MDTK) Prospectus

Alan Karr

[karr@niss.org](mailto:karr@niss.org)

October 18, 2005

# Background: NISS research on SDL for numerical microdata

- Thrust
  - Risk-utility formulations for microdata
  - Utility measures tied to inference
  - Multiple candidate releases
  - Risk-utility frontiers
- Initial study
  - A. F. Karr, C. N. Kohnen, A. Oganian, J. P. Reiter, and A. P. Sanil (2005), A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality (<http://www.niss.org/dgii/TR/utilitycomp-final.pdf>)
- Continuing research
  - Additional utility measures
  - Combining SDL methods

# The Initial Study

- 8 SDL methods, each with “settable parameters”
  - Rank swapping
  - Resampling
  - Addition of noise
  - 5 forms of microaggregation
- 3 utility measures
  - KL distance between original and masked data
  - For a single regression:
    - Confidence ellipse (for estimated coefficients)
    - Confidence intervals (for estimated coefficients)
- 2 disclosure risk measures
  - % of correctly linked records
  - % of correctly linked records, weighted by distance

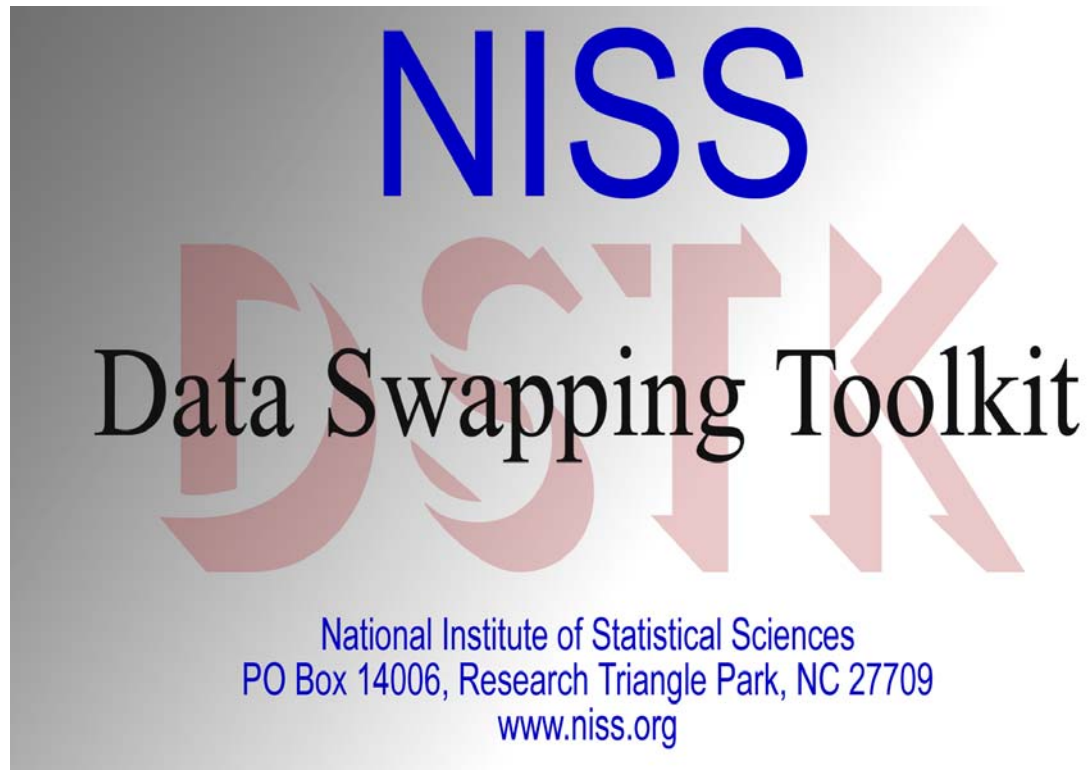
# Ongoing Research

- Utility measures that are
  - Not tied to normal data
  - Not tied to specific analyses
  - Tractable computationally
  - Example: propensity scores
- Iterated SDL methods
  - Example: microaggregation followed by addition of noise
  - Can be superior in terms of both risk and utility to each component alone

# A “Typical” Experiment

- Multiple SDL methods
  - For each, multiple parameter values
  - Possibly, multiple replications
- Multiple data utility measures
- Multiple disclosure risk measures
- Initial study
  - $8 \times 3 \times 2$  [x 5] possible releases
- Need a software tool for both research and “production” purposes: the MDTK

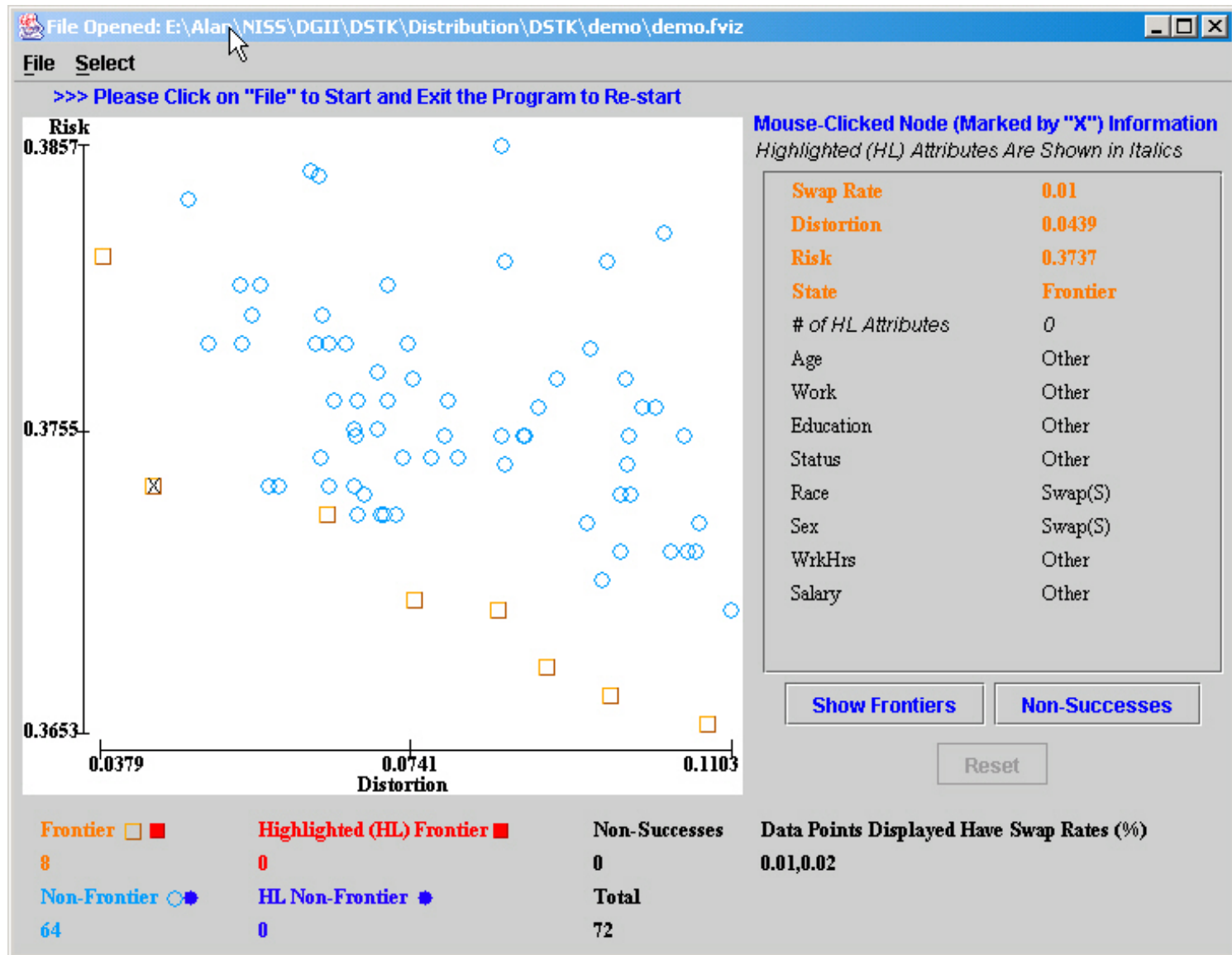
# Previous Product



# DSTK Basics

- Able to handle numerical and categorical data
- Perform experiments with multiple choices of
  - Swap attribute(s)
  - Swap rate
  - Constraints on unswapped attributes
- Three components
  - GUI for single swaps
  - Batch swap package
  - [Integrated Batch Swapper]
  - Frontier visualizer
- Written in Java
- Available at <http://ww.niss.org/software/dstk.html>

# Frontier Visualizer



# What Would the MDTK Do?

- Computational experiments similar to—but more complicated than—those done by the DSTK
  - User selects
    - Methods and one or more parameter values for each
    - Utility measure(s)
    - Risk measure(s)
  - MDTK
    - Computes risk and utility measures
    - Calculates frontier
    - Visualizes results

# MDTK Structure

- Analogous to DSTK, consisting of engines run by script (specifications) files that can be generated
  - Manually
  - Via GUI
- Modular and extensible, to allow “plug in” of
  - New SDL methods
  - New risk measures
  - New utility measures

# An MDTK Experiment

- User specifies
  - SDL methods, individually and/or in combination
  - Possibly multiple values of parameters for each
  - Data utility measure(s)
  - Disclosure risk measure(s)
- MDTK
  - Generates each masked data set
  - Calculates data utility and disclosure risk
  - Calculates and visualizes risk-utility frontier
  - Logs full results in various files

# What's in Place?

- Good code in place for
  - Individual SDL methods from initial study
  - Risk and utility methods from initial study
- Code under development for
  - New utility measures
    - KS distance
    - Cramer-von Mises statistic
    - Propensity scores
  - Iterated methods
    - Non-trivial because parameters of second method may depend on output of first

# What's Missing?

- Big issues
  - Extensibility
    - APIs
    - Data structures
    - Inter-component communication
    - ...
  - Scalability
- Approaches
  - High-end: professional programmer working 8-12 months
  - More modest
    - Professional software architect to create core
    - Postdoc or student programmers to do GUIs, viz, ...