NISS Data Swapping Toolkit

Ashish Sanil

National Institute of Statistical Sciences

ashish@niss.org

November 20, 2003

NISS DSTK Overview

- The NISS DSTK is a set of software programs and tools for performing and analyzing data swapping on categorical data within a risk-utility framework
- The DSTK was produced by the Digital Government Research Program at NISS, with support from the National Science Foundation and the National Center for Education Statistics.
- Written by Ashish Sanil, Jimmy Fulp, Shanti Gomatam, Charlie Liu and Alan Karr

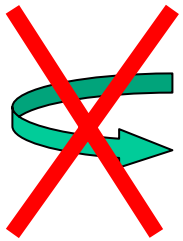
Outline

- Data Swapping overview
- DSTK description
- DSTK demo
- Concluding comments and future work

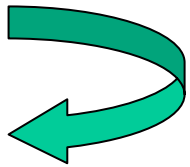
Data Swapping

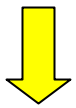
- Technique for statistical disclosure limitation (SDL), applied at microdata level
- Basic idea: switch subset of attributes between randomly selected pairs of records
- Rationale: intruder cannot be certain that any record is real
- Side effect: distorts data, reducing utility

Swapping Specifications: Example 1



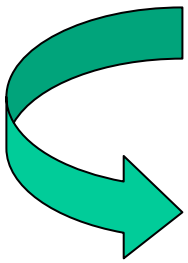
S	O	O	O	O	O
K-12	MA	M	0	L	53
K-12	IL	M		D	37
P	IL	F	1	D	42
HS	MA	M	1	K	48
SP	NC	F	1		45



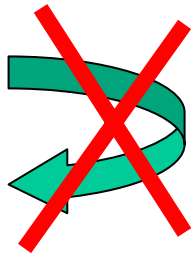


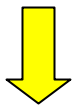
P	MA	M	0	L	53
K-12	IL	M		D	37
K-12	IL	F	1	D	42
HS	MA	M	1	K	48
SP	NC	F	1		45

Swapping Specifications: Example 2



S	F	O	O	O	O
K-12	MA	M	0	L	53
K-12	IL	M		D	37
P	IL	F	1	D	42
HS	MA	M	1	K	48
SP	NC	F	1		45





HS	MA	M	0	L	53
K-12	IL	M		D	37
P	IL	F	1	D	42
K-12	MA	M	1	K	48
SP	NC	F	1		45

Swapping Specifications

- S : variable will be swapped
- D : variable constrained to not be equal in the pair of records being swapped
- F : variable constrained to be equal in the pair of records being swapped
- O : unconstrained variable

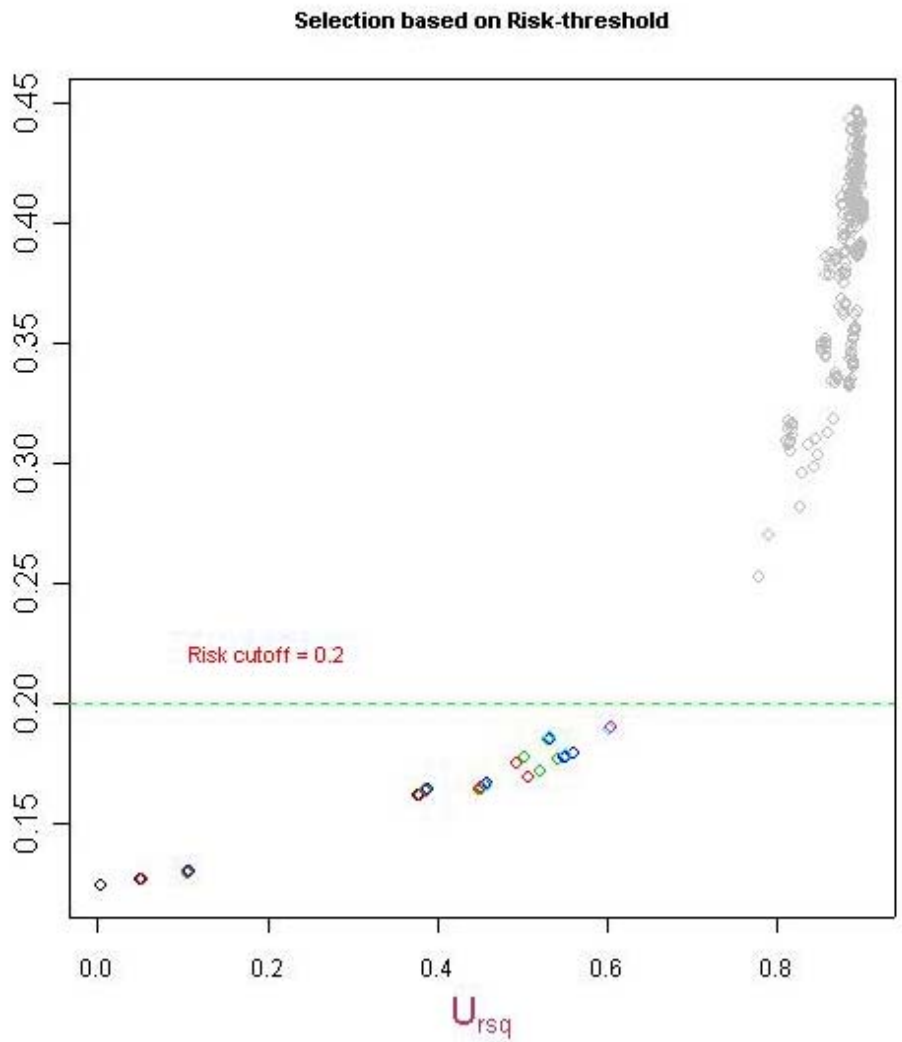
Data Swapping: Technical Aspects

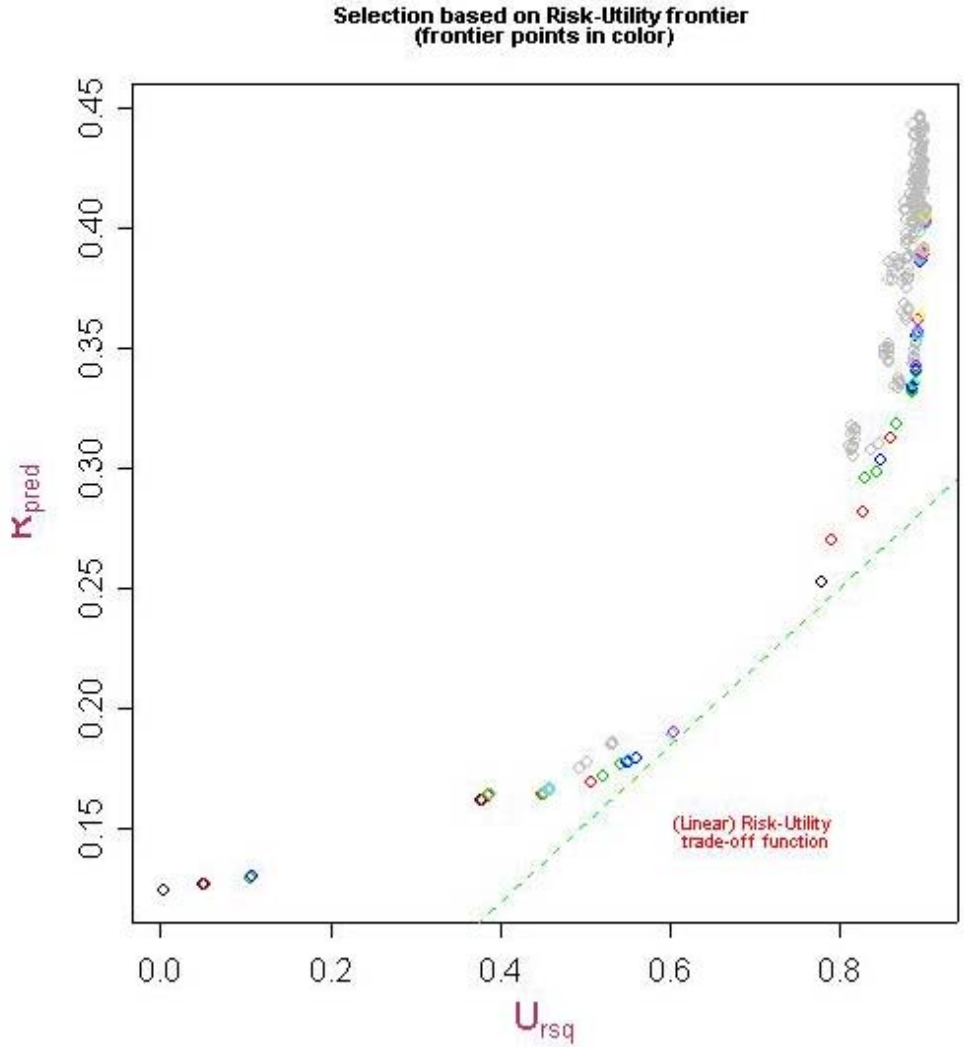
- Parameters
  - Swap rate (E.g, swap 2% of the records)
  - Swap attributes
  - Optionally, constraints on “non-swap” attributes
- Distortion effects
  - No change to joint distribution of swap attributes
  - No change to joint distribution of non-swap attributes
  - Change to joint distributions that involve both swap and non-swap attributes

Risk-Utility Framework

- Characterize each candidate release by
  - Disclosure **r**isk
  - Data **u**tility
- Agency could either
  - Maximize Utility subject to a Risk threshold
  - Jointly optimize over (Risk, Utility)
- We restrict attention to the frontier of un-dominated releases

Selection of Optimal Release based on Risk-Utility measures

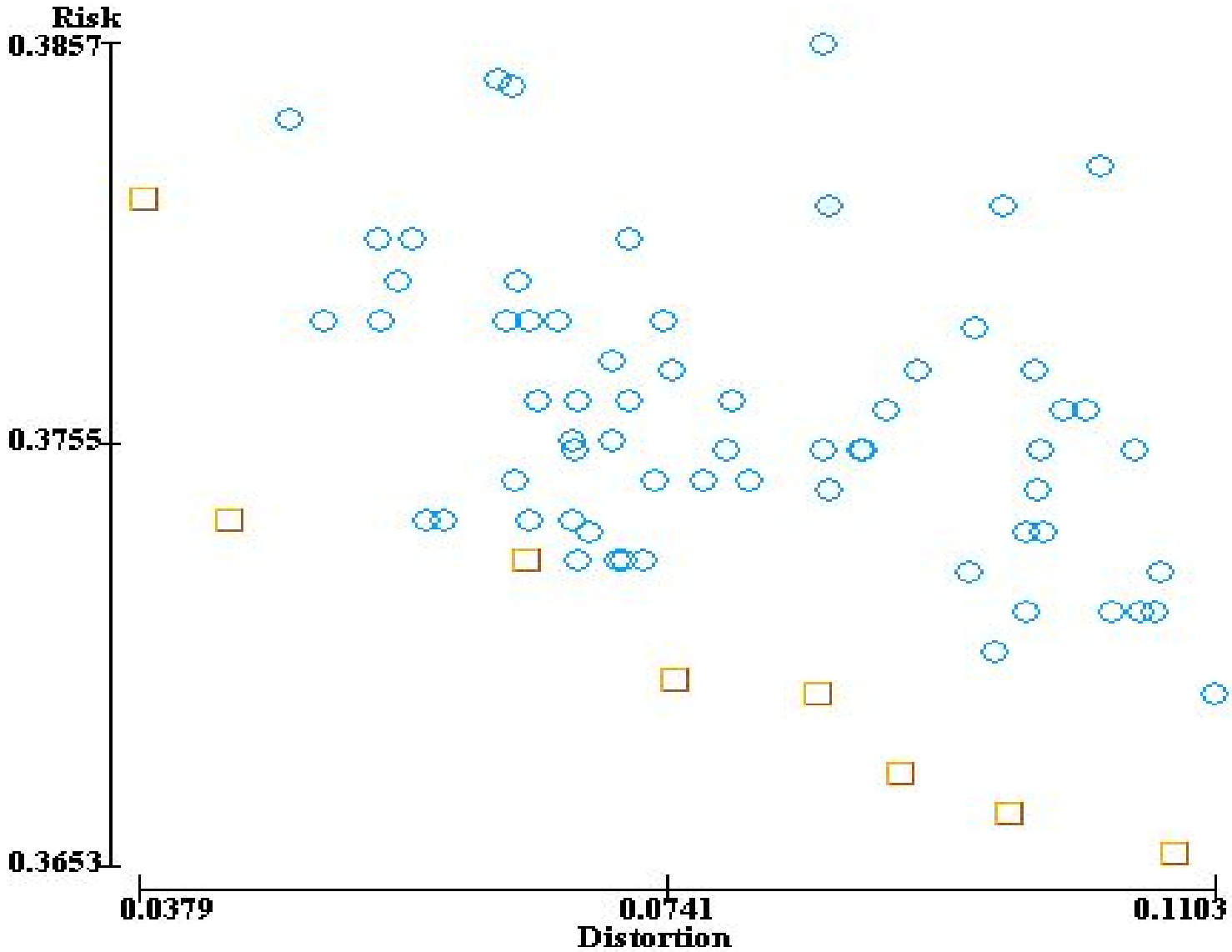




Distortion

- Operationally, we will minimize a distortion or dis-utility measure

Example Risk-Distortion Frontier



Risk and Distortion measures

- Risk: Fraction of unswapped records in low-count cells in post-swap data

$$\frac{\sum_{C1,C2} \text{Number of unswapped records}}{\text{Total number of unswapped records}}$$

- Distortion: Hellinger distance between pre- and post-swap data tables
- Weighted versions also supported

NISS DSTK Overview

- Functionality provided:
  - Single swaps, using a graphical user interface
  - Batch swaps
  - Risk-utility calculations
  - Visualization of (distortion, risk) frontiers
  - Java class library for performing customized data swapping tasks

Data File Format

- CSV text file (both ISO and Microsoft standards supported)
- First line is a header line of attribute names
- First three columns should be (ID, Weight, Categorized_Weight) -- all optional, but order should be maintained if any are present
- Attribute values are treated as categorical data (except for Weight, when present)

Data File Examples

RecordID	Weight	WtClass	Age	Work	Education	Status	Race	Sex	WrkHrs	Salary
1	4.34	a	25_55	Gov	Bach	UM	W	M	40	<50
2	96.12	d	25_55	SE	Bach	M	W	M	<40	<50
3	6.57	a	25_55	Pvt	HS	UM	W	M	40	<50
4	48.07	b	25_55	Pvt	<HS	M	NW	M	40	<50

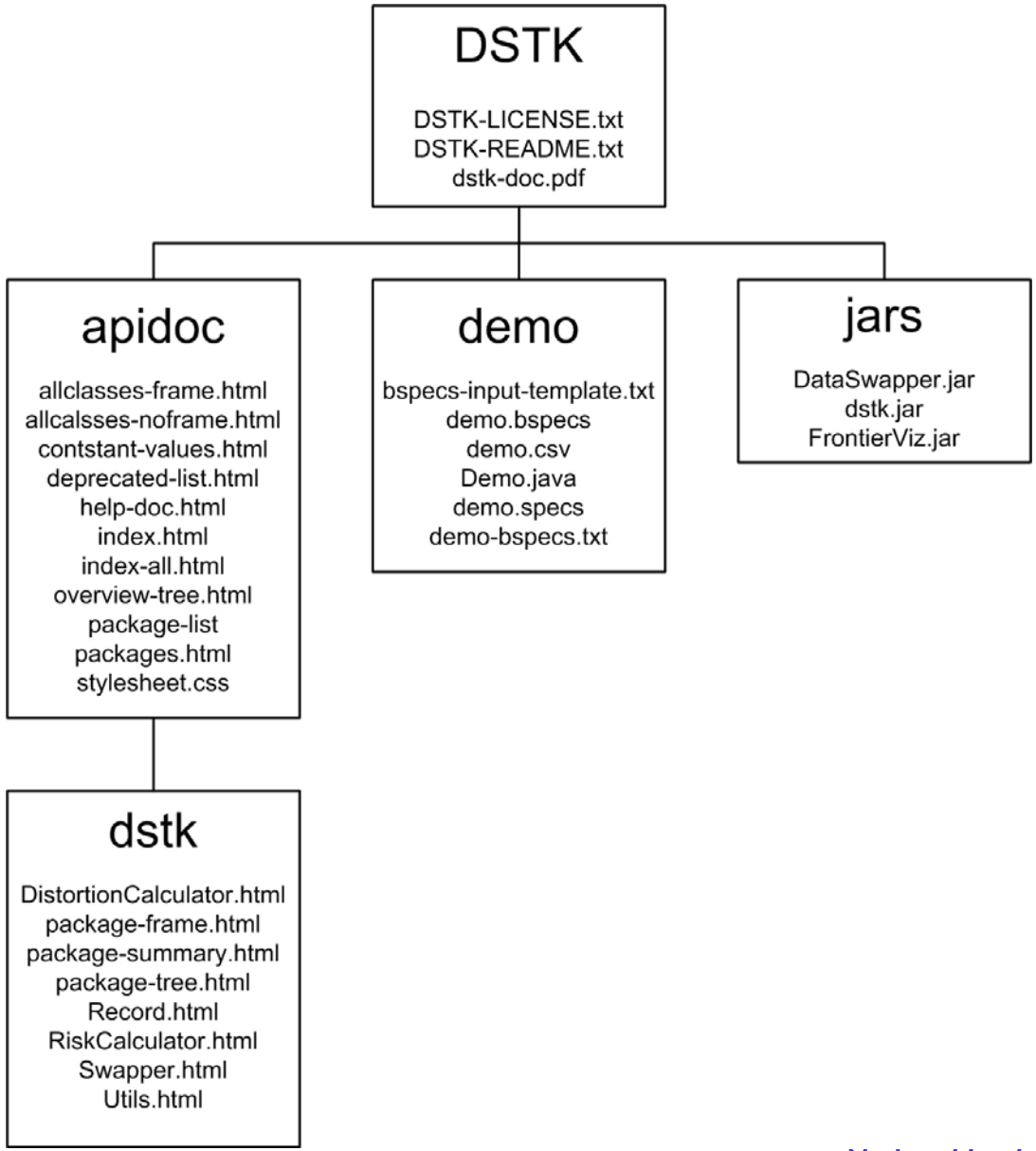
RecordID	WtClass	Age	Work	Education	Status	Race	Sex	WrkHrs	Salary
1	a	25_55	Gov	Bach	UM	W	M	40	<50
2	d	25_55	SE	Bach	M	W	M	<40	<50
3	a	25_55	Pvt	HS	UM	W	M	40	<50
4	b	25_55	Pvt	<HS	M	NW	M	40	<50

NISS

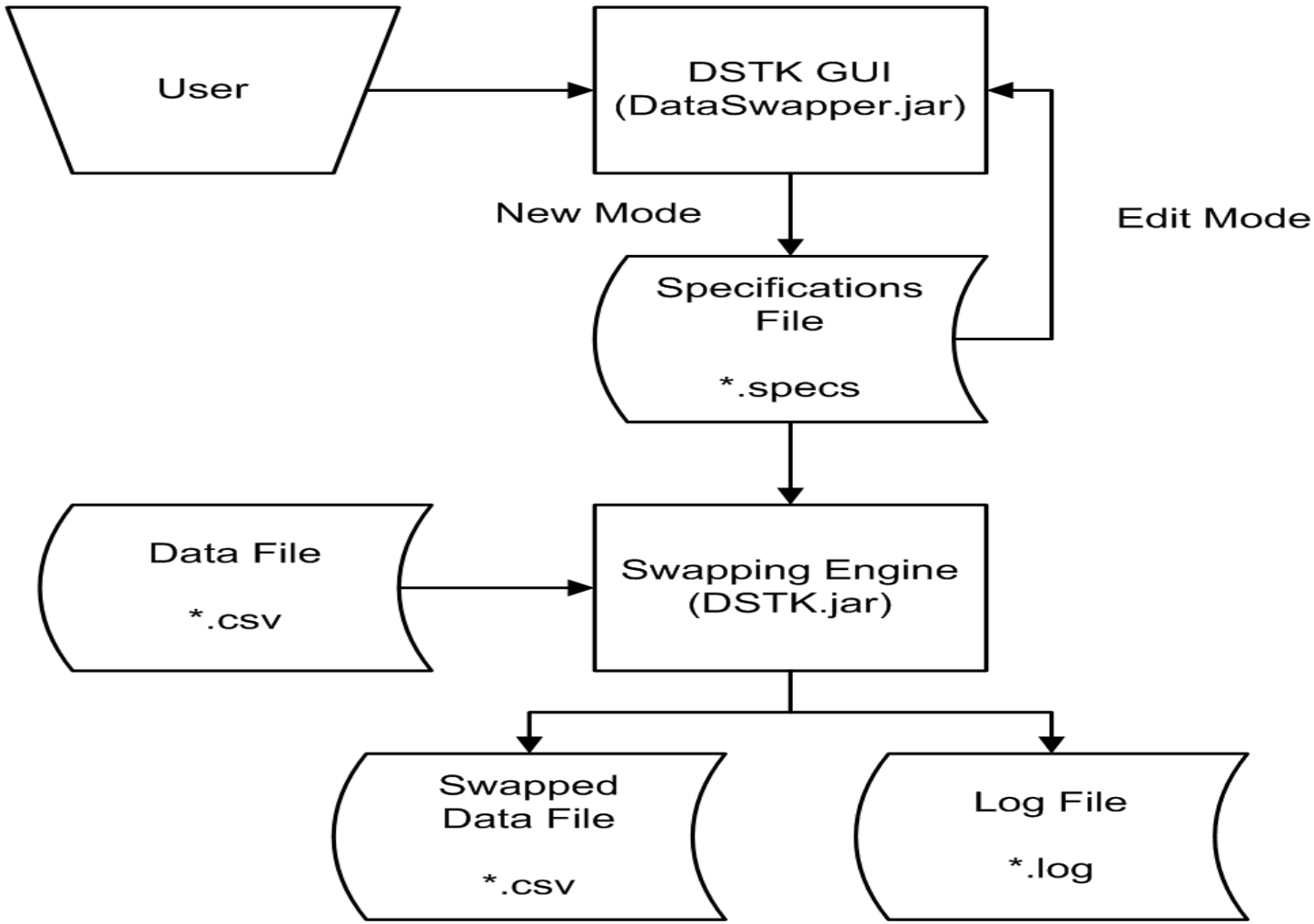
Data Swapping Toolkit

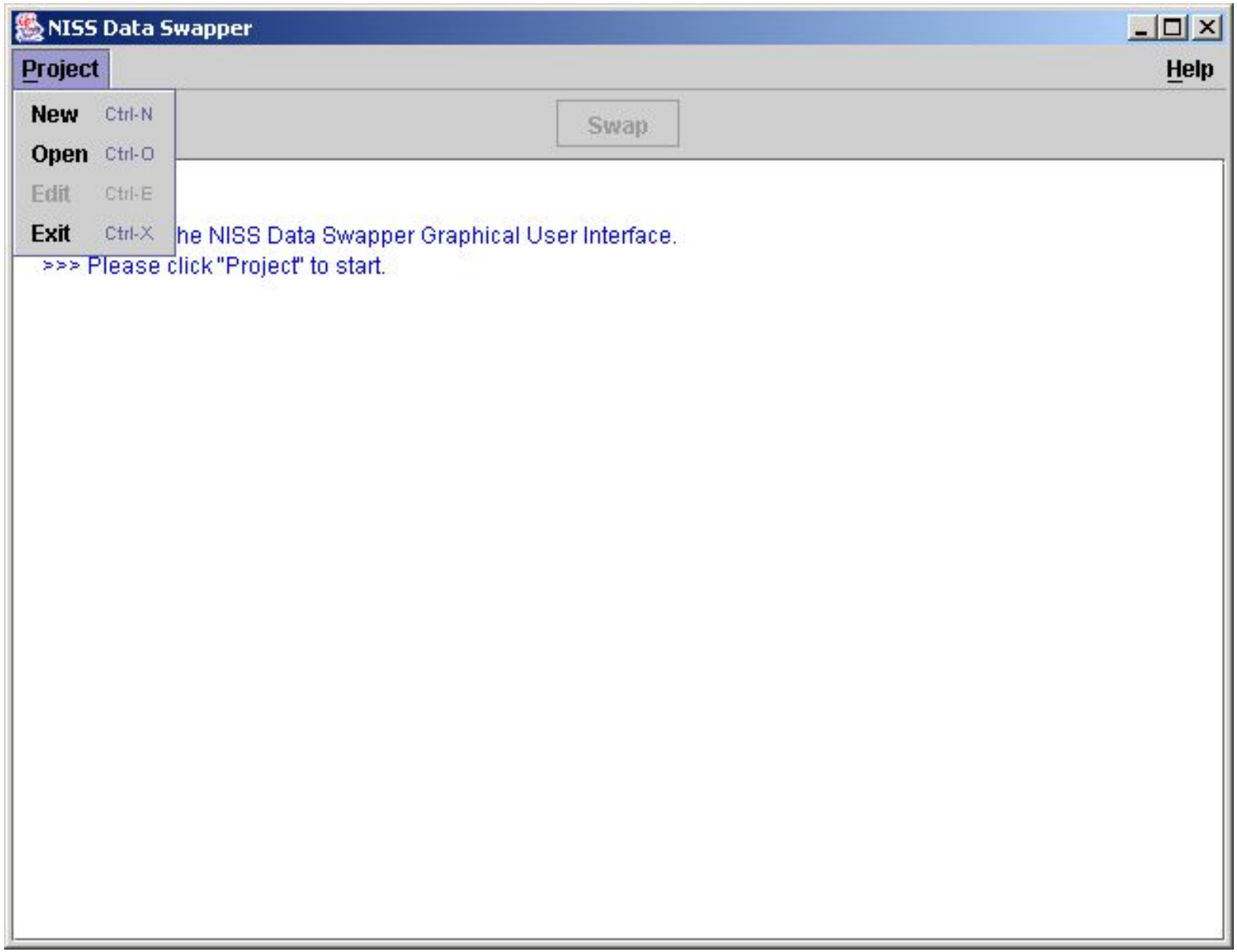
National Institute of Statistical Sciences
PO Box 14006, Research Triangle Park, NC 27709
www.niss.org

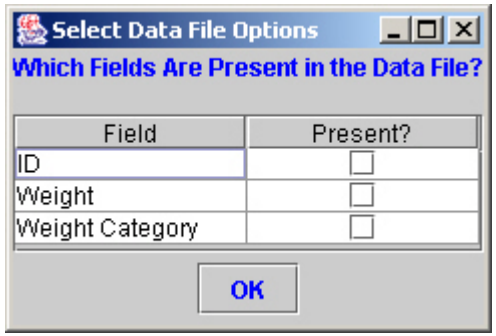
DSTK Package

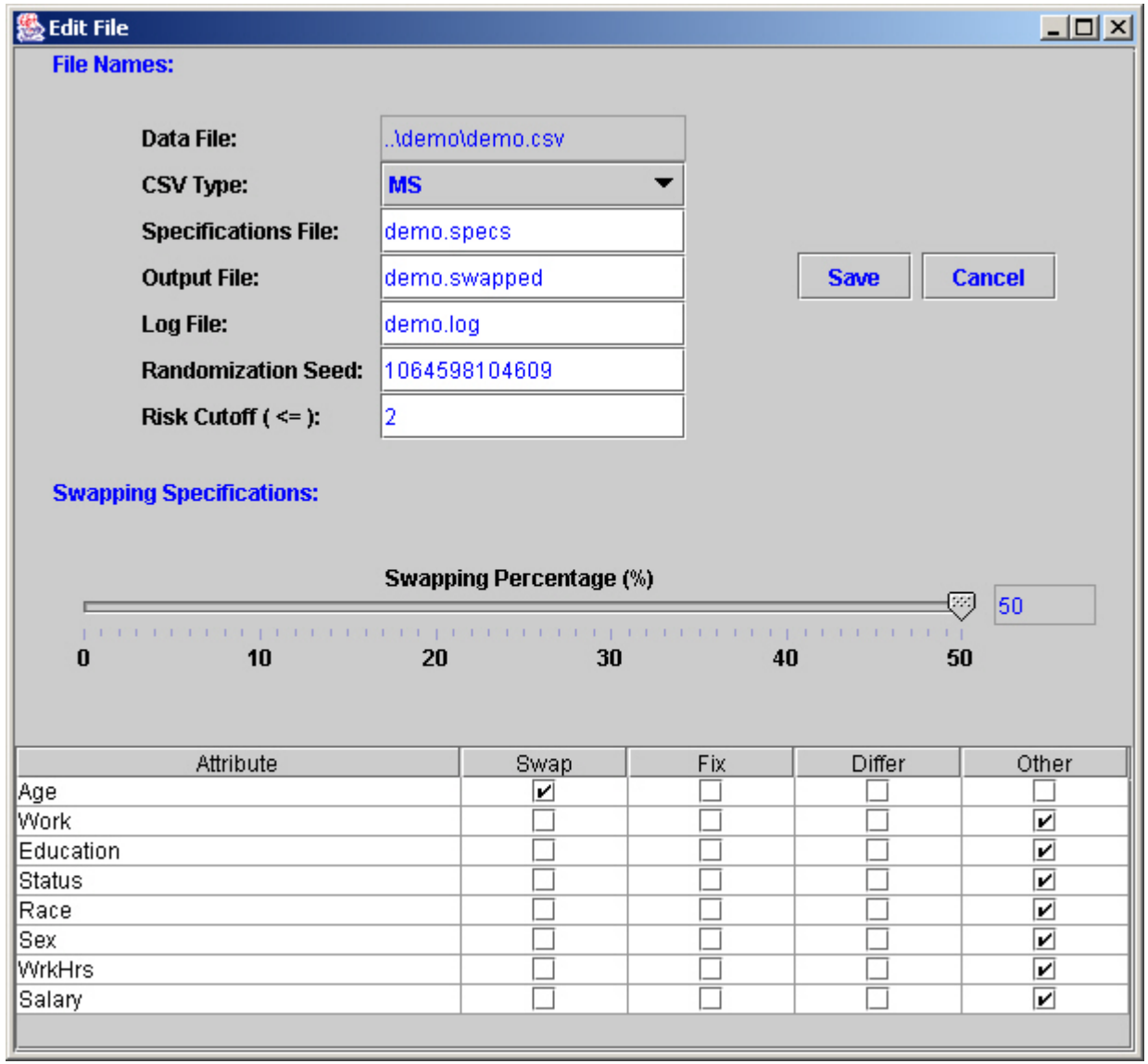


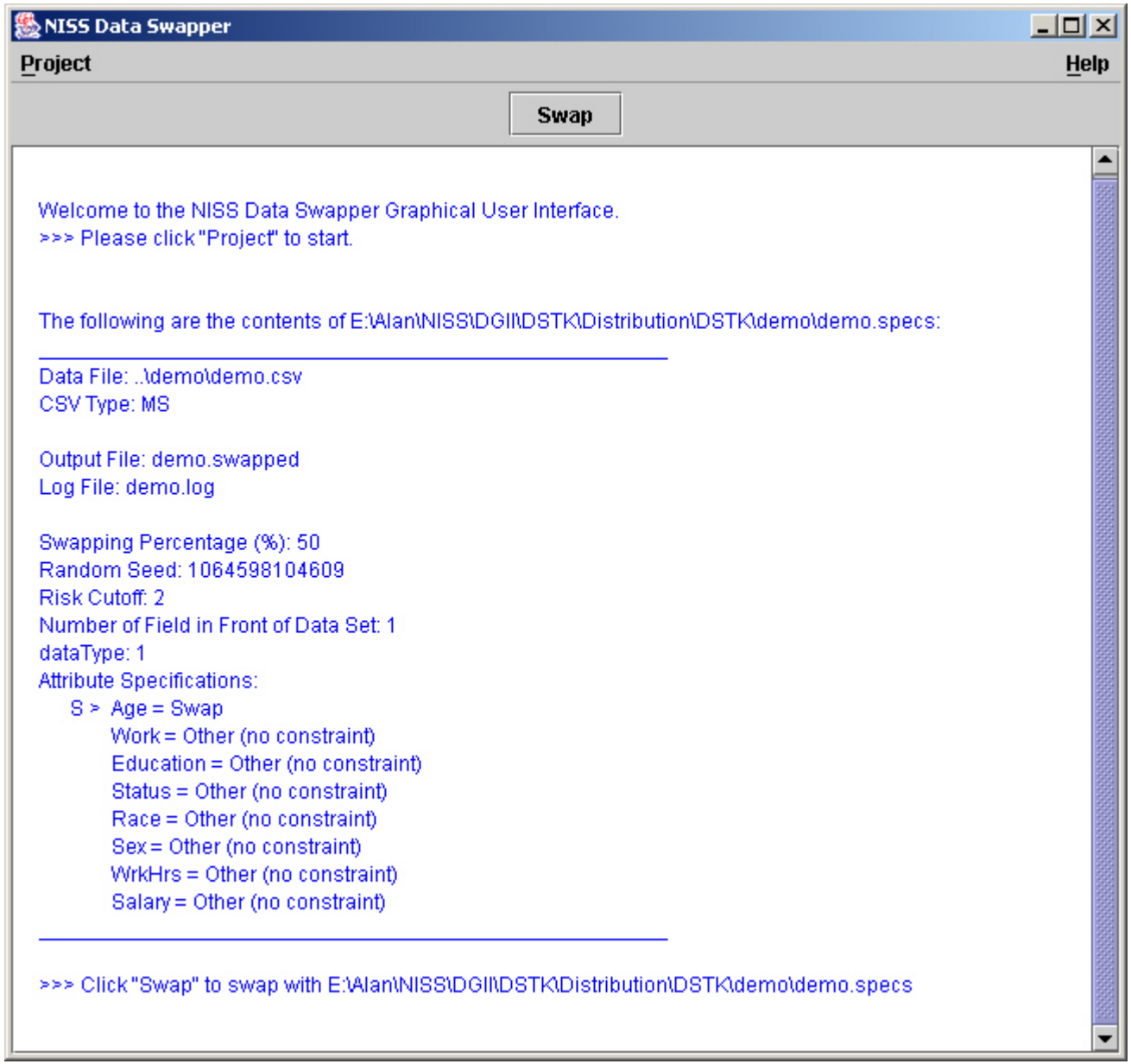
GUI-based Swapper

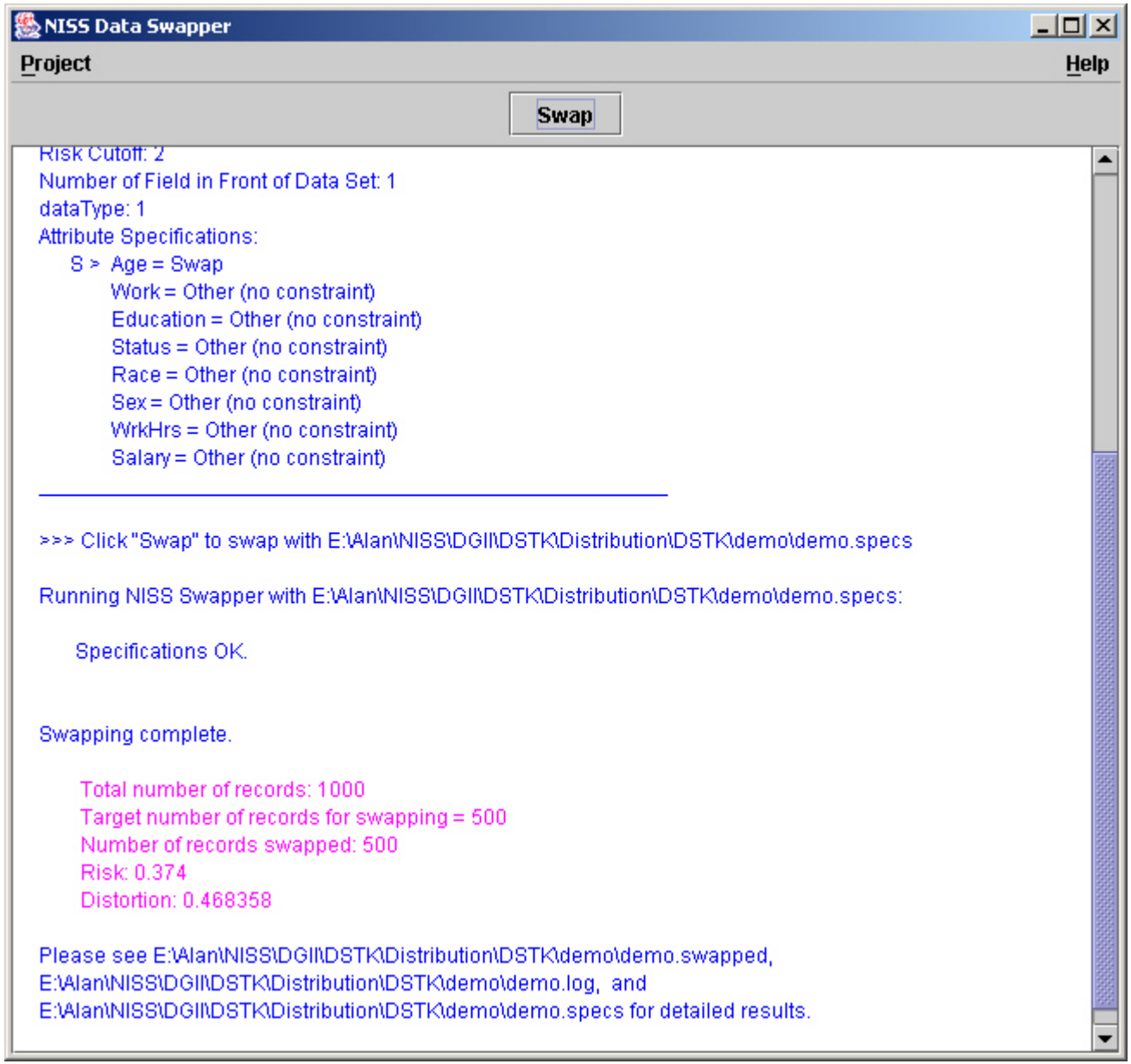



NISS Data Swapper
Project
Help
New Ctrl-N
Open Ctrl-O
Edit Ctrl-E
Exit Ctrl-X
Swap
he NISS Data Swapper Graphical User Interface.
>>> Please click "Project" to start.

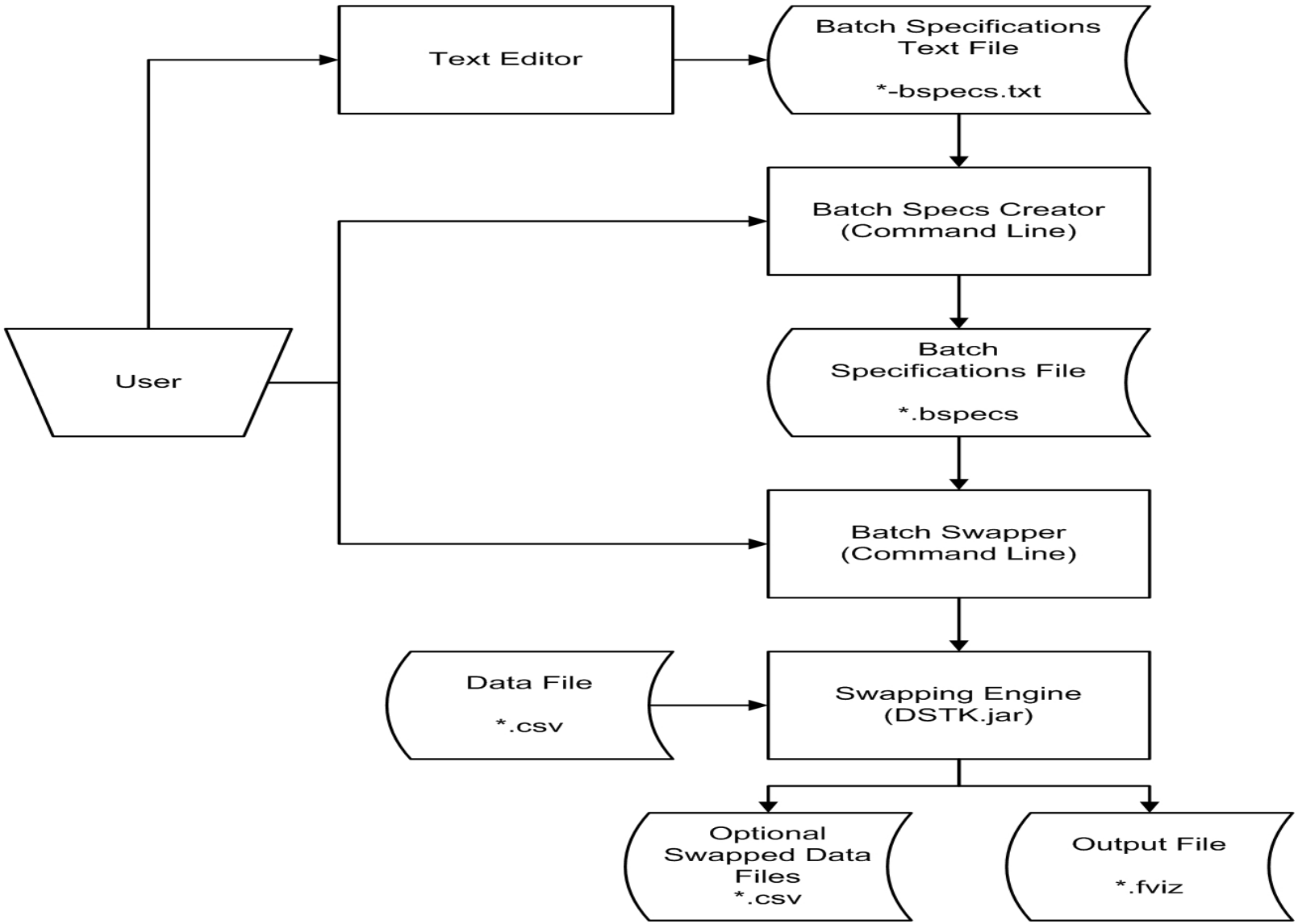

Select Data File Options
Which Fields Are Present in the Data File?
Field
Present?
ID
Weight
Weight Category
OK


Edit File
File Names:
Data File:
..\demo\demo.csv
CSV Type:
MS
Specifications File:
demo.specs
Output File:
demo.swapped
Log File:
demo.log
Randomization Seed:
1064598104609
Risk Cutoff (<=):
2
Save
Cancel
Swapping Specifications:
Swapping Percentage (%)
50
0
10
20
30
40
50
Attribute
Swap
Fix
Differ
Other
Age
Work
Education
Status
Race
Sex
WrkHrs
Salary


NISS Data Swapper
Project
Help
Swap
Welcome to the NISS Data Swapper Graphical User Interface.
>>> Please click "Project" to start.
The following are the contents of E:\Alan\NISS\DGII\DSTK\Distribution\DSTK\demo\demo.specs:
Data File: ..\demo\demo.csv
CSV Type: MS
Output File: demo.swapped
Log File: demo.log
Swapping Percentage (%): 50
Random Seed: 1064598104609
Risk Cutoff: 2
Number of Field in Front of Data Set: 1
dataType: 1
Attribute Specifications:
S > Age = Swap
Work = Other (no constraint)
Education = Other (no constraint)
Status = Other (no constraint)
Race = Other (no constraint)
Sex = Other (no constraint)
WrkHrs = Other (no constraint)
Salary = Other (no constraint)
>>> Click "Swap" to swap with E:\Alan\NISS\DGII\DSTK\Distribution\DSTK\demo\demo.specs


NISS Data Swapper
Project
Help
Swap
Risk Cutoff: 2
Number of Field in Front of Data Set: 1
dataType: 1
Attribute Specifications:
S > Age = Swap
Work = Other (no constraint)
Education = Other (no constraint)
Status = Other (no constraint)
Race = Other (no constraint)
Sex = Other (no constraint)
WrkHrs = Other (no constraint)
Salary = Other (no constraint)
>>> Click "Swap" to swap with E:\Alan\NISS\DGII\DSTK\Distribution\DSTK\demo\demo.specs
Running NISS Swapper with E:\Alan\NISS\DGII\DSTK\Distribution\DSTK\demo\demo.specs:
Specifications OK.
Swapping complete.
Total number of records: 1000
Target number of records for swapping = 500
Number of records swapped: 500
Risk: 0.374
Distortion: 0.468358
Please see E:\Alan\NISS\DGII\DSTK\Distribution\DSTK\demo\demo.swapped,
E:\Alan\NISS\DGII\DSTK\Distribution\DSTK\demo\demo.log, and
E:\Alan\NISS\DGII\DSTK\Distribution\DSTK\demo\demo.specs for detailed results.

Batch Swapper



Batch Specs Creator: Example Input File

```
# Required fields
data.file = demo.csv
swap.rates = 0.01,0.02
swap.options = oneway,twoway

# Optional fields (default value)
specs.file = (demo.bspecs)
output.file = (demo.fviz)
save.dir = ([none])
csv.type = (MS)
risk.cutoff = (2)
record.id = (false)          # else true
weight = (false)             # else true
weight.category = (false)    # else true
```

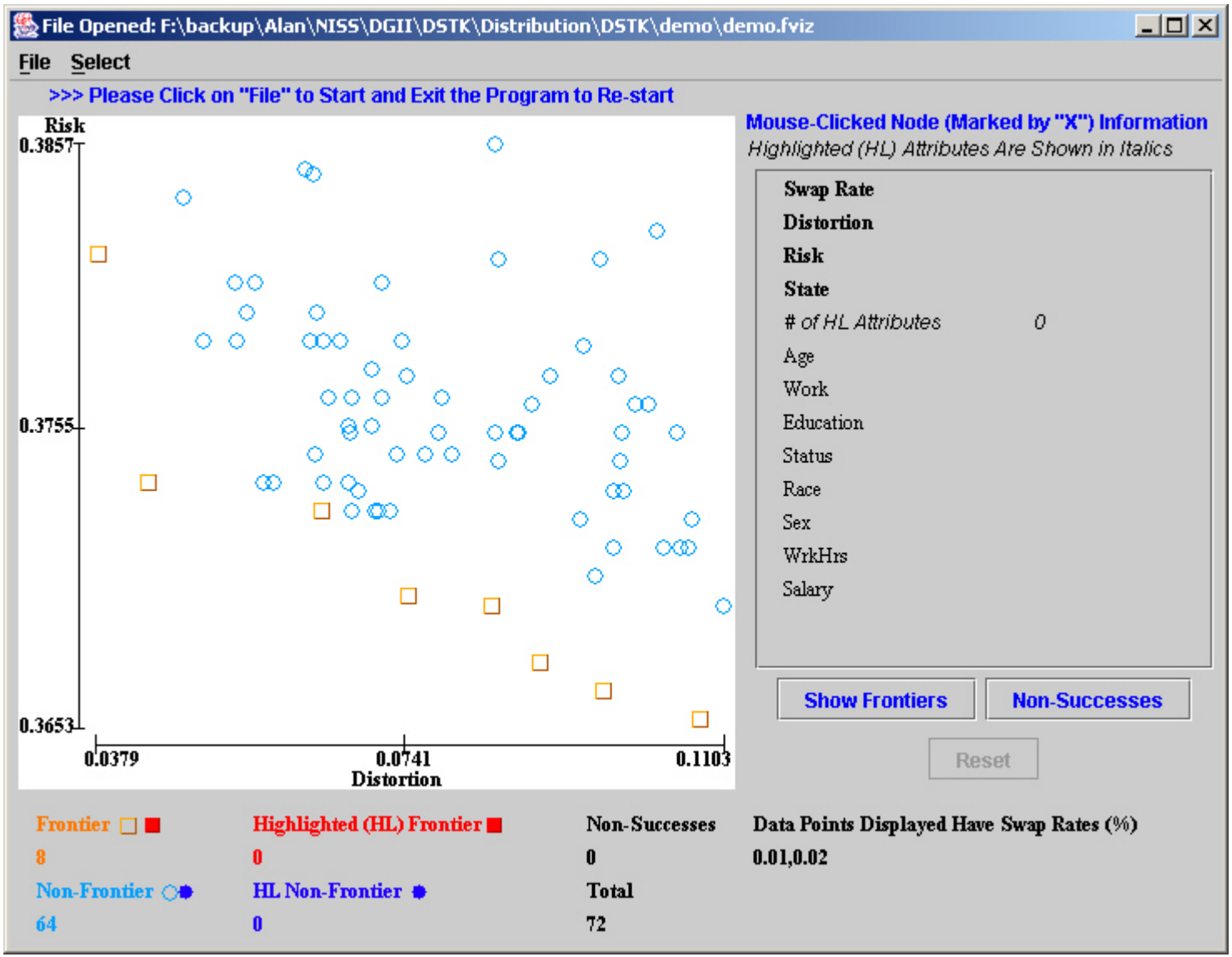
Batch Specs Creator: Example Output File

```
#demo.bspecs was created from demo-bspecs.txt
#Fri Nov 10 13:58:48 EDT 2003
record.id=true
output.file=demo.fviz
weight.category=false
weight=false
csv.type=MS
swap.options=oneway,twoway
data.file=demo.csv
specs.file=demo.bspecs
swap.rates=0.01,0.02
risk.cutoff=2
!0.01,S,O,O,O,O,O,O,O
!0.01,O,S,O,O,O,O,O,O
!0.01,O,O,S,O,O,O,O,O
……
……
```

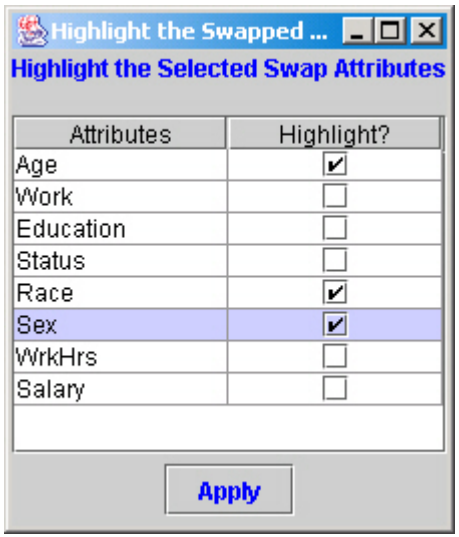
Batch Swapper: Example Output File

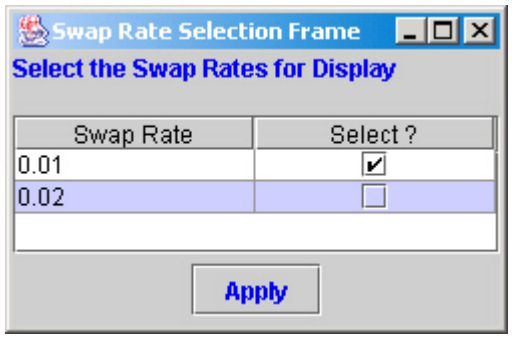
```
Age,Work,Education,Status,Race,Sex,WrkHrs,Salary,Rate,Dist,Risk,Flag,Seed
S,O,O,O,O,O,O,O,0.01,0.06649346464750473,0.3717171717171717,1,1065808728734
O,S,O,O,O,O,O,O,0.01,0.06699583262583089,0.3707070707070707,1,1065808728906
O,O,S,O,O,O,O,O,0.01,0.05693716357184327,0.3757575757575573,1,1065808729000
O,O,O,S,O,O,O,O,0.01,0.06341229767862364,0.3717171717171717,1,1065808729078
O,O,O,O,S,O,O,O,0.01,0.06581421012587736,0.3747474747474748,1,1065808729125
O,O,O,O,O,S,O,O,0.01,0.05828494798750034,0.3797979797979798,1,1065808729187
O,O,O,O,O,O,S,O,0.01,0.06713290866586832,0.3737373737373737,1,1065808729250
O,O,O,O,O,O,O,S,0.01,0.05881228421247015,0.3808080808080808,1,1065808729296
S,O,O,O,O,O,O,O,0.02,0.09044463780402506,0.3744897959183673,1,1065808729359
O,S,O,O,O,O,O,O,0.02,0.08664443689306965,0.3724489795918367,1,1065808729406
.....
.....
```

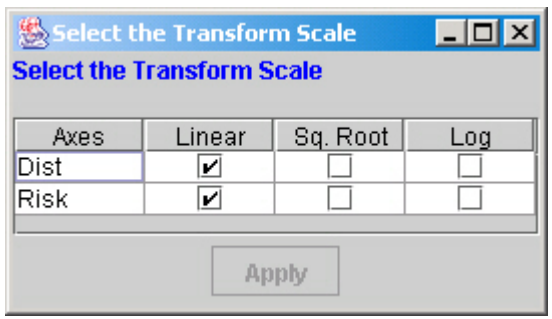
Frontier Visualizer

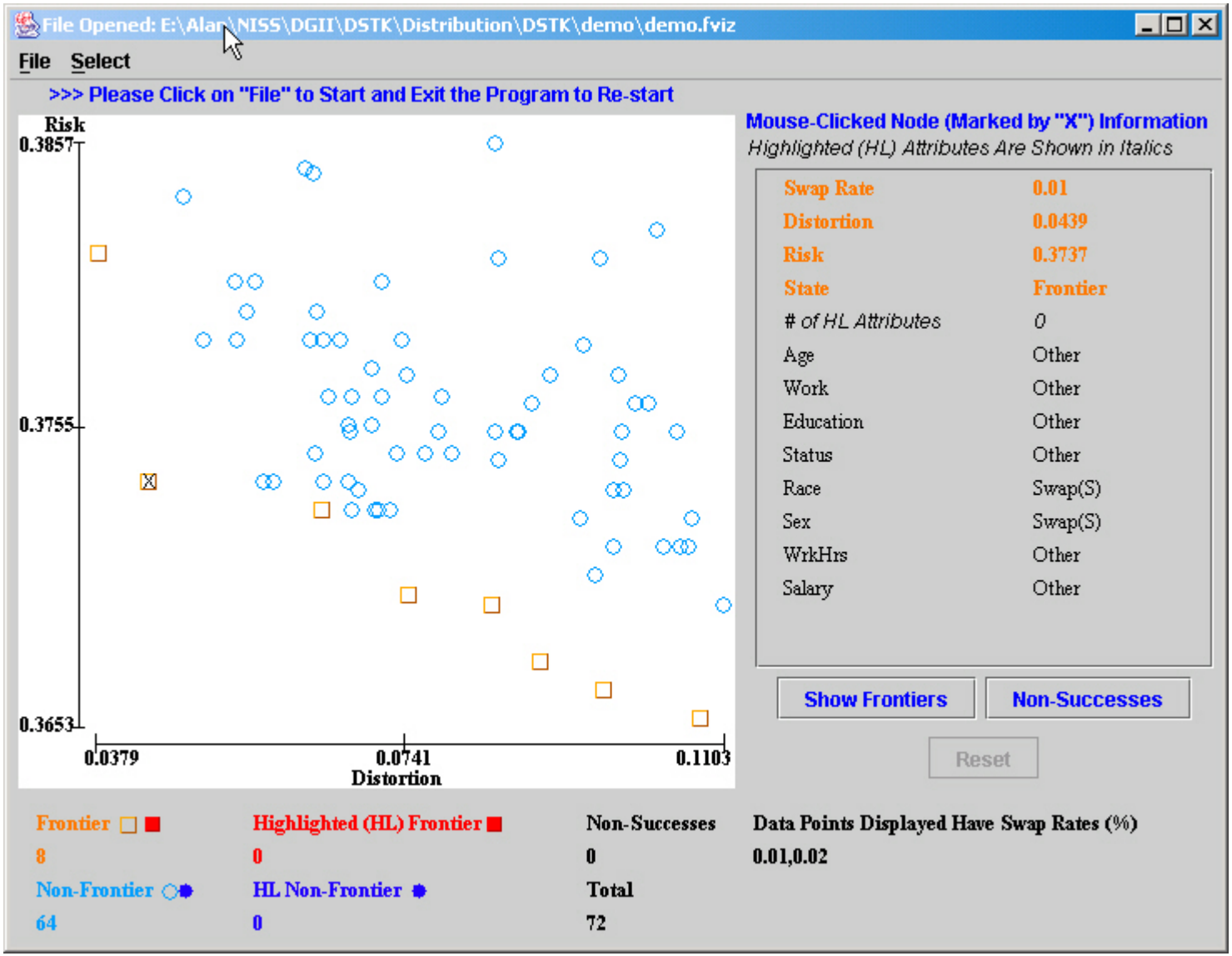


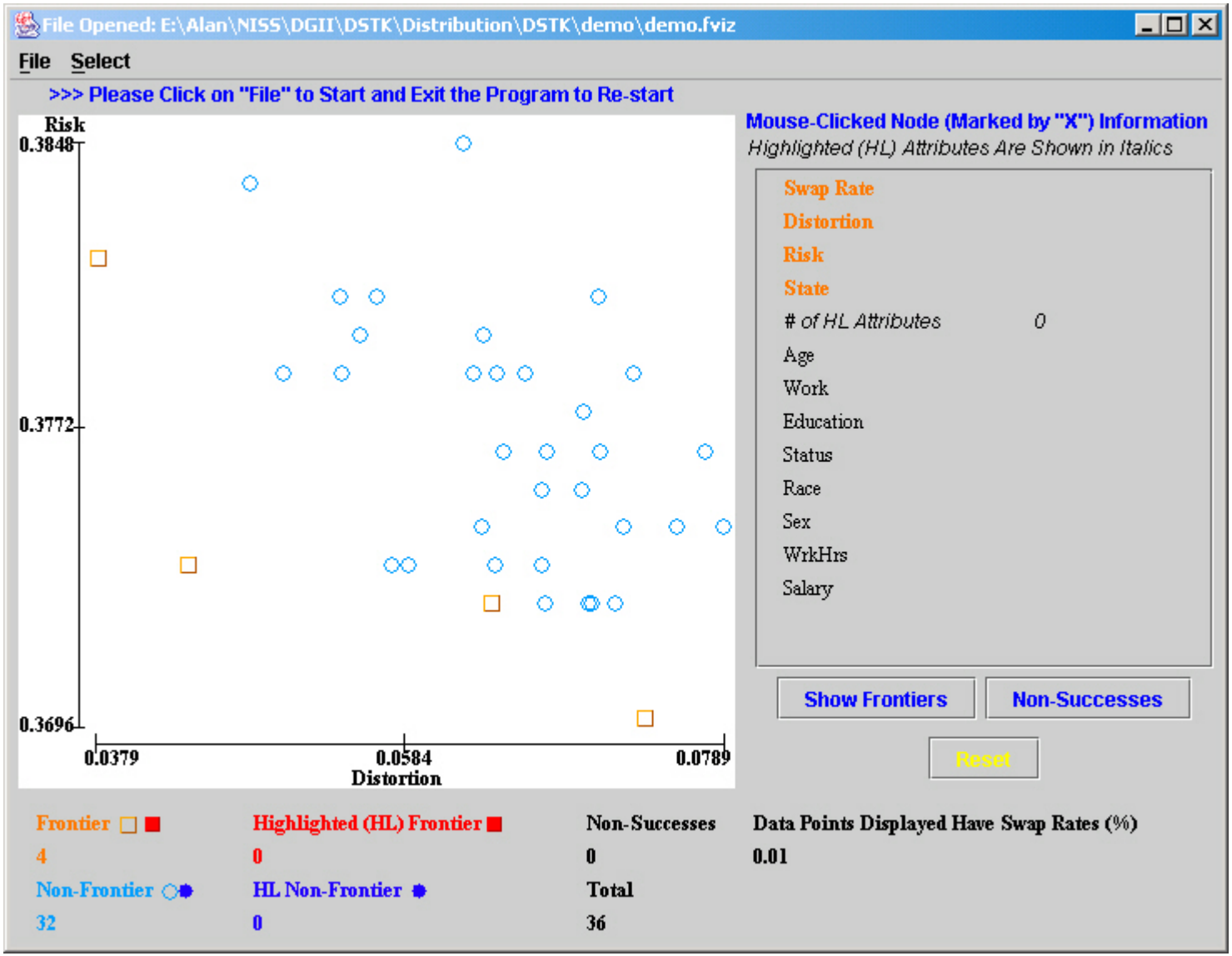
Selections

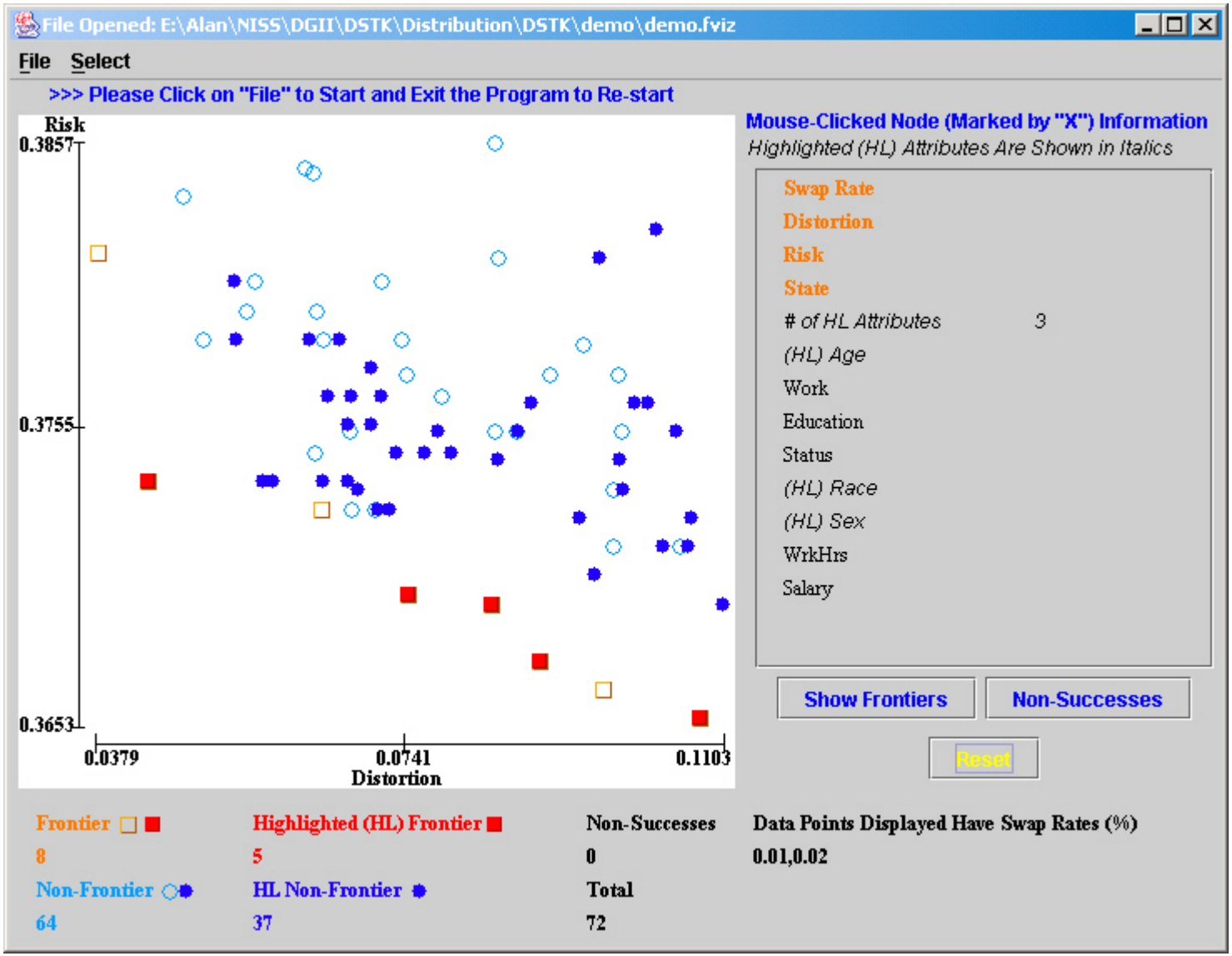

Highlight the Swapped ...
Highlight the Selected Swap Attributes
Attributes
Highlight?
Age
Work
Education
Status
Race
Sex
WrkHrs
Salary
Apply


Swap Rate Selection Frame
Select the Swap Rates for Display
Swap Rate
Select ?
0.01
0.02
Apply


Select the Transform Scale
Select the Transform Scale
Axes
Linear
Sq. Root
Log
Dist
Risk
Apply


File Opened: E:\Alan\NISS\DGII\DSTK\Distribution\DSTK\demo\demo.fviz
File Select
>>> Please Click on "File" to Start and Exit the Program to Re-start
Risk
0.3857
0.3755
0.3653
0.0379
0.0741
0.1103
Distortion
Mouse-Clicked Node (Marked by "X") Information
Highlighted (HL) Attributes Are Shown in Italics
Swap Rate 0.01
Distortion 0.0439
Risk 0.3737
State Frontier
# of HL Attributes 0
Age Other
Work Other
Education Other
Status Other
Race Swap(S)
Sex Swap(S)
WrkHrs Other
Salary Other
Show Frontiers
Non-Successes
Reset
Frontier
8
Highlighted (HL) Frontier
0
Non-Successes
0
Data Points Displayed Have Swap Rates (%)
0.01,0.02
Non-Frontier
64
HL Non-Frontier
0
Total
72


File Opened: E:\Alan\NISS\DGII\DSTK\Distribution\DSTK\demo\demo.fviz
File Select
>>> Please Click on "File" to Start and Exit the Program to Re-start
Risk
0.3848
0.3772
0.3696
0.0379
0.0584
0.0789
Distortion
Mouse-Clicked Node (Marked by "X") Information
Highlighted (HL) Attributes Are Shown in Italics
Swap Rate
Distortion
Risk
State
# of HL Attributes 0
Age
Work
Education
Status
Race
Sex
WrkHrs
Salary
Show Frontiers
Non-Successes
Reset
Frontier
4
Highlighted (HL) Frontier
0
Non-Successes
0
Data Points Displayed Have Swap Rates (%)
0.01
Non-Frontier
32
HL Non-Frontier
0
Total
36


File Opened: E:\Alan\NISS\DGII\DSTK\Distribution\DSTK\demo\demo.fviz
File Select
>>> Please Click on "File" to Start and Exit the Program to Re-start
Risk
0.3857
0.3755
0.3653
0.0379
0.0741
0.1103
Distortion
Mouse-Clicked Node (Marked by "X") Information
Highlighted (HL) Attributes Are Shown in Italics
Swap Rate
Distortion
Risk
State
# of HL Attributes 3
(HL) Age
Work
Education
Status
(HL) Race
(HL) Sex
WrkHrs
Salary
Show Frontiers
Non-Successes
Reset
Frontier
8
Highlighted (HL) Frontier
5
Non-Successes
0
Data Points Displayed Have Swap Rates (%)
0.01,0.02
Non-Frontier
64
HL Non-Frontier
37
Total
72

Frontier Visualizer Output

- Tables of frontier points
- JPEG images of the plots

DSTK Java Class library

- Provides set of classes
- Example code
- HTML docs

```
.....

  public static void main(String[] args) {
        String origFile = "demo.csv";
        String swappedFile = "demo-swapped.csv";
        String csvType = "MS";
        long seed = 1058215043231L;
        double rate = 0.01;
        int riskCutoff = 2;
        char[] constraints = {'S','F','O','O','O','O','O','O'};
        int dataType = Swapper.HAS_ID;
        Swapper swapper = new Swapper();
        try {
            swapper.setDataType(dataType);
            swapper.readOrigData(origFile,csvType,true);
            swapper.setRate(rate);
            swapper.setSeed(seed);
            swapper.setConstraints(constraints);
            swapper.doSwap();
            swapper.writeData(swapper.SWAPPED,swappedFile,true);
            String[] log = swapper.getLog();
            for(int i=0; i < log.length; i++) {
                System.out.println(log[i]);
            }
            RiskCalculator rc = new RiskCalculator(swapper,riskCutoff);
            DistortionCalculator dc = new DistortionCalculator(swapper);
            System.out.println("Risk: " + rc.risk());
            System.out.println("Distortion: " + dc.distortion());
        }
        catch (IOException ioe) {

.....
```

Future Work

- Provide other Risk measures and other Utility measures (specially inference-based ones)
- Support RDBMS as data sources
- Incorporate automatic aggregation functionality

NISS Data Swapping Toolkit

Available at:

http://www.niss.org/software/dstk.html