

NISS

Data Confidentiality, Data Quality
and Data Integration
for Federal Databases

Alan F. Karr
karr@niss.org

Why Are We Doing This?



NORTH CAROLINA STATE BOARD OF ELECTIONS

SBOE Home :: Campaign Finance :: En Español :: Board Members :: SBOE Staff :: County Offices :: Search

[CHECK YOUR VOTER REGISTRATION HERE](#)

Voter Registration
Voting Information
Data and Statistics
Forms
Election Laws
SEIMS
Related Links

Voter Data Results From The NC Statewide Database	
Click Here to Search for Another Voter.	
Name:	KARR, ALAN FRANCIS
County Name:	ORANGE
Status:	ACTIVE
City:	CHAPEL HILL NC 27516
Race:	WHITE
Ethnicity:	NOT HISPANIC or NOT LATINO
Gender:	Male
Party:	



AnyBirthday.com

846 West St., New York, NY 10001 **Search using Age or birthday**

Born: Sep. 11, 1902



Locateme.com

Smith, John R.

[Click here for a Name and Age Search](#)

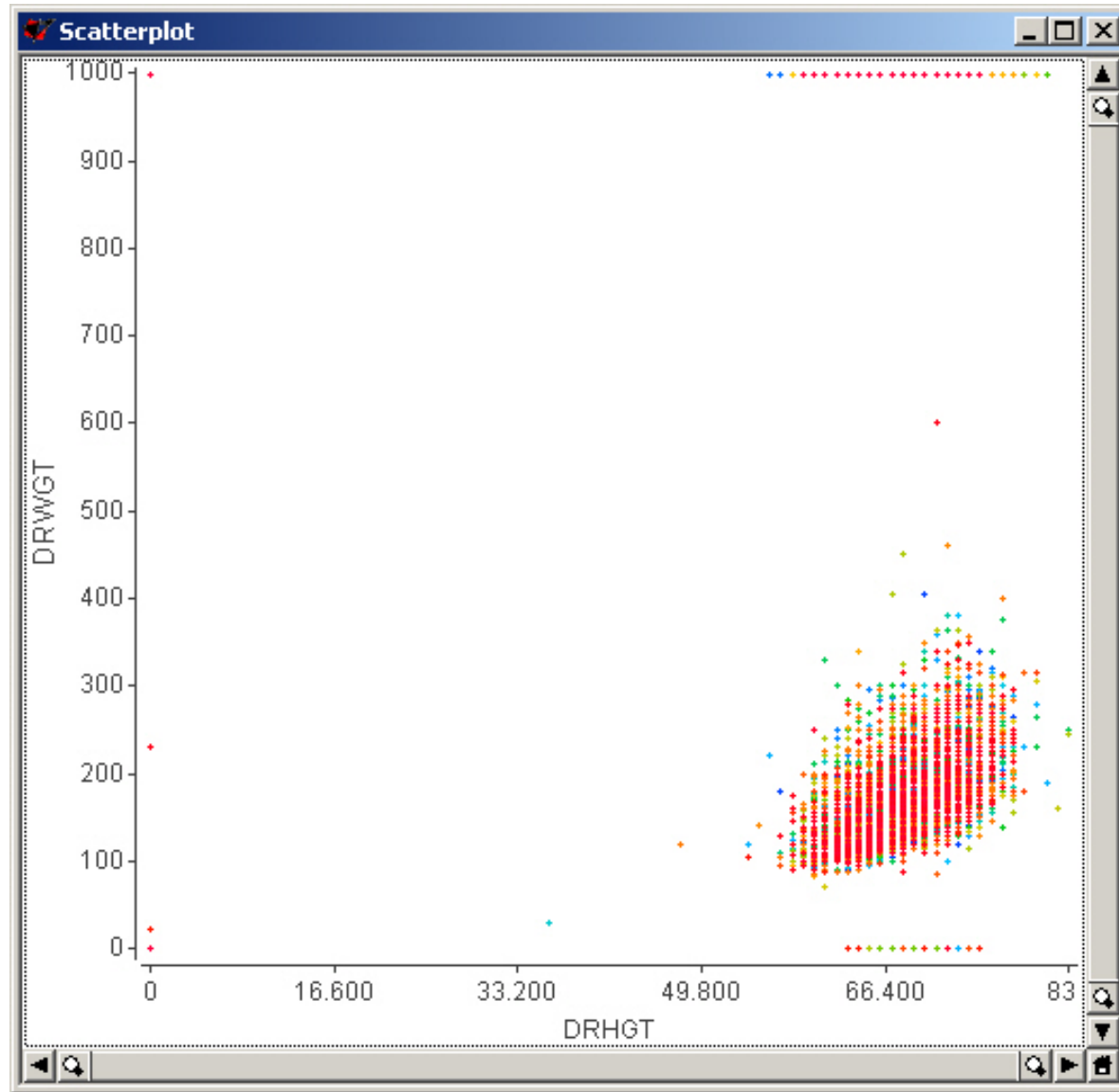
[Click here for Addresses and Phone Numbers of your search subject.](#)

NEW! Anybirthday.com PLUS lists Addresses!

Subject's Name	Birthday	Zip Code
ALAN F KARR		27516

ADDRESS: * Included for *Plus* Users Only [Click for Anybirthday PLUS](#)

Why Else?



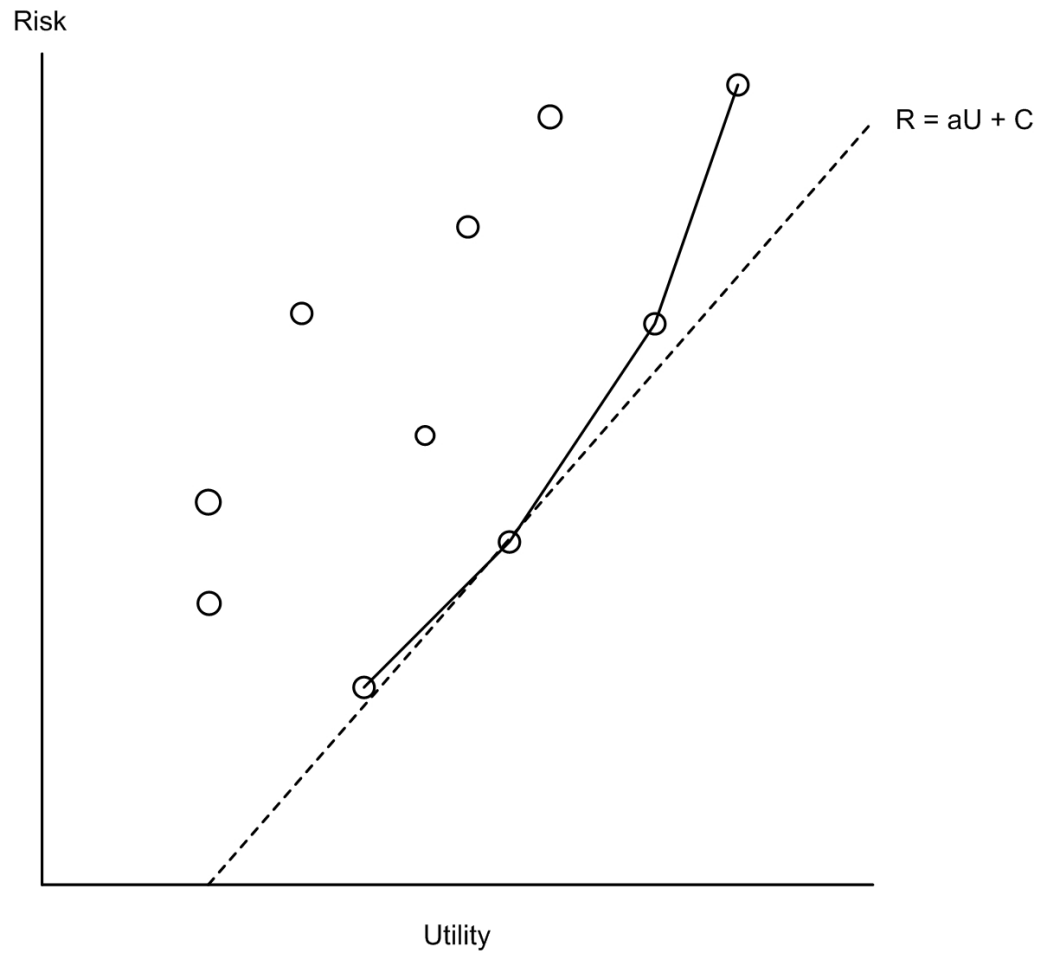
Why (3)?

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	CSTATE	CNUM	SEQNUM	VEHNUM	LNUM	PNUM	CITY	COUNTY	ACCDAT	ACCTIME	VEHFORM	PFORMS	NMOTFO	NHS
5369	53	183	0	0	1	0	690	61	6041999	327	1	1	0	1
5370	53	215	0	0	1	0	740	73	5261999	1205	1	2	0	1
5371	53	242	0	0	1	0	0	11	7101999	1125	2	5	0	1
5372	53	359	0	0	1	0	2230	53	9021999	2002	3	8	0	1
5373	53	383	0	0	1	0	1960	33	8261999	45	1	2	1	1
5374	53	412	0	0	1	0	0	61	8211999	30	1	3	0	1
5375	53	429	0	0	1	0	0	53	10171999	141	2	6	0	1
5376	53	431	0	0	1	0	0	67	10991999	9999	1	1	0	1
5377	53	446	0	0	1	0	2310	33	10231999	1815	4	6	0	1
5378	53	486	0	0	1	0	0	67	10111999	1910	3	5	0	1
5379	53	510	0	0	1	0	0	11	11261999	1359	2	6	0	1
5380	53	518	0	0	1	0	0	67	12101999	1347	1	1	0	1
5381	53	527	0	0	1	0	1000	15	12171999	1305	3	5	0	1
5382	6	1327	0	0	1	0	0	113	8011999	1	1	5	1	1
5383	32	34	0	0	1	0	0	3	1111999	945	1	1	0	1
5384	5	382	0	0	1	0	2320	119	9121999	1421	1	2	0	1
5385	17	4	0	0	1	0	0	63	1011999	2010	1	2	0	1
5386	17	16	0	0	1	0	3105	31	1051999	1415	1	2	1	1
5387	17	36	0	0	1	0	0	117	1131999	820	1	1	0	1
5388	17	45	0	0	1	0	0	167	1181999	355	1	2	0	1
5389	17	128	0	0	1	0	0	119	2211999	555	1	1	0	1
5390	17	238	0	0	1	0	0	43	4121999	2332	4	6	1	1
5391	17	380	0	0	1	0	0	135	5171999	45	1	1	0	1
5392	17	410	0	0	1	0	0	197	5291999	310	1	2	0	1
5393	17	459	0	0	1	0	0	197	6171999	102	4	4	0	1
5394	17	482	0	0	1	0	9340	167	6241999	1629	2	12	0	1
5395	17	659	0	0	1	0	8730	119	8071999	1352	1	2	0	1
5396	17	696	0	0	1	0	2610	163	8161999	5	2	7	0	1
5397	17	779	0	0	1	0	0	105	9011999	1435	1	1	0	1
5398	17	867	0	0	1	0	0	197	9301999	2332	2	2	0	1
5399	17	879	0	0	1	0	8410	31	10111999	2305	1	2	1	1
5400	17	1159	0	0	1	0	8410	31	11211999	544	2	3	0	1

Objectives

- High Level
 - Allow Federal statistical agencies to disseminate *useful* information derived from confidential data but *protect* the privacy of data subjects (individuals and establishments)
- Scientific
 - Problem formulations and scalable tools that accommodate both *disclosure risk* and *data/information utility*
 - Understanding *consequences of data integration* for data confidentiality, data quality and statistical inference
 - Creation of fundamental quantifications, usable models, scalable methods for *data quality*

Risk-Utility Formulations

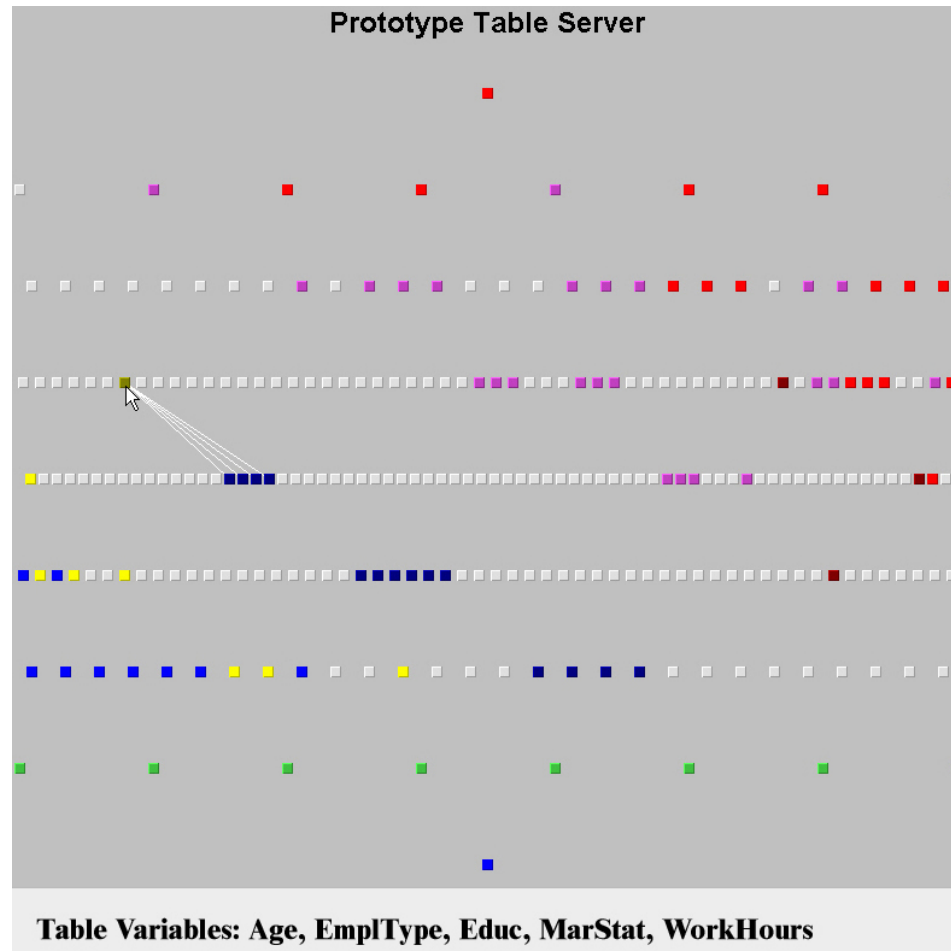


Accomplishments—1

Dissemination of Marginal Tables

- Scalable methods and software to compute bounds on cell entries from released marginals
- Initial methods to assess disclosure risk from released conditionals
- First successful technology that deals in a principled way with query interaction
- Scalable risk-utility formulations
 - Optimal tabular releases: maximize number of disclosed marginals subject to constraint on width of bounds for small count cells

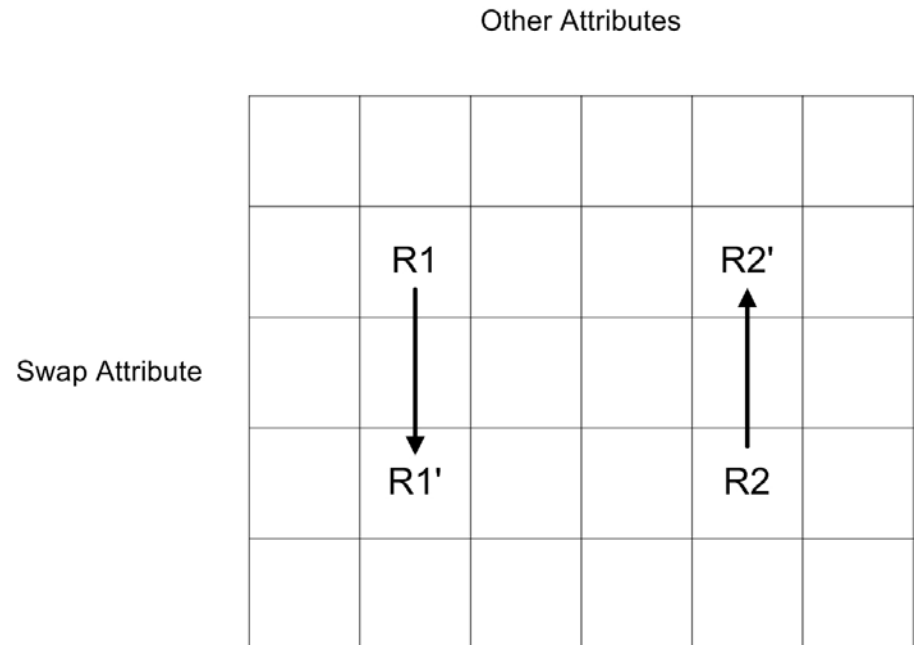
Software Product: Prototype Table Server



Accomplishments—2

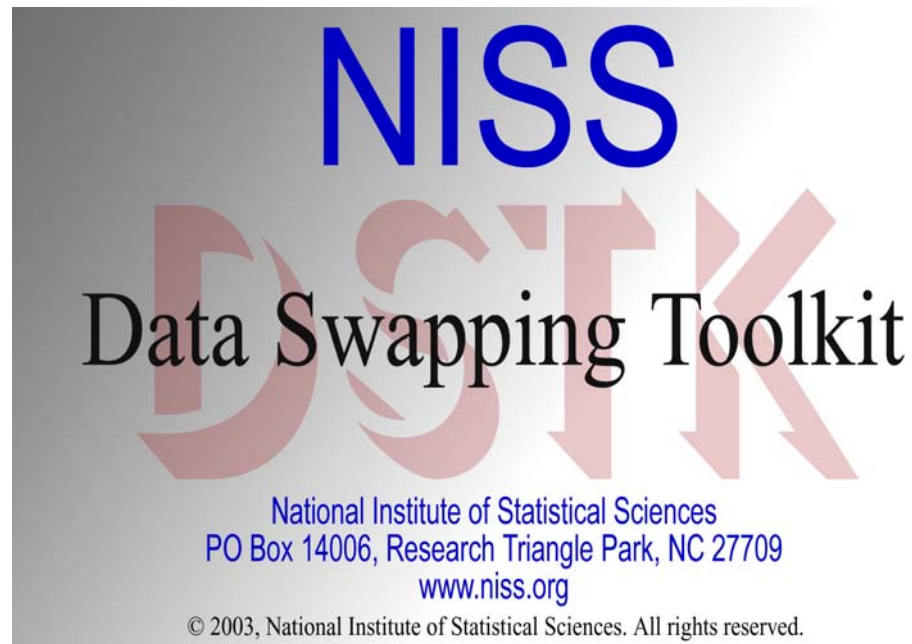
Data Swapping

- Complete risk-utility formulation for data swapping as a decision problem
 - Multiple measures of utility/distortion and disclosure risk
- Web service implementation of data swapping
 - Downloadable from www.niss.org/WebServices/dg/WebSwap.html

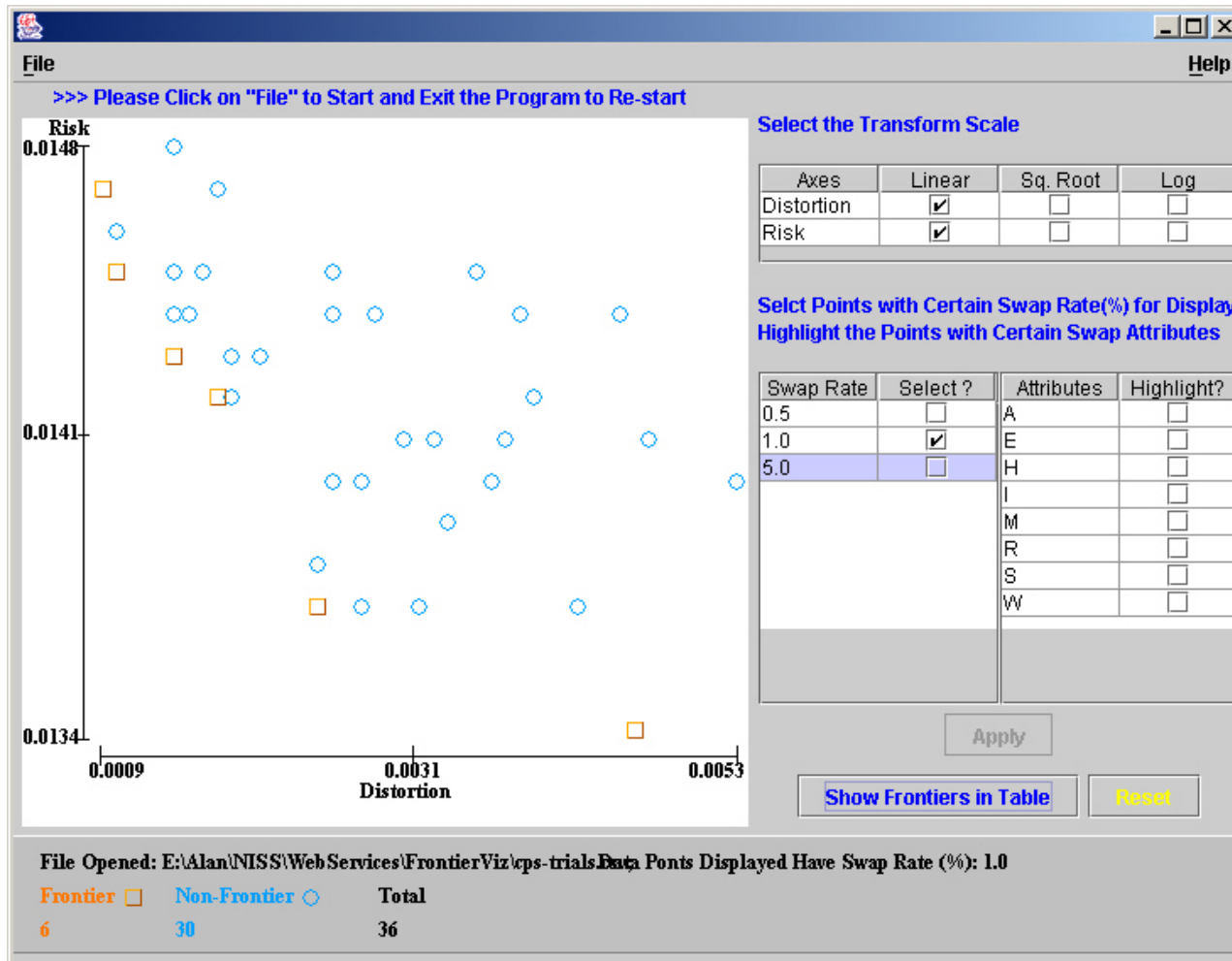


Software Product: NISS Data Swapping Toolkit

- Functionality
 - Perform swapping
 - Batch mode: large-scale studies to select swap attributes and rate
 - Visualization of results
- Downloadable from www.niss.org/software/dstk.html



Frontier Visualizer



Accomplishments—3

Remote Access Analysis Servers

- Fundamental abstractions
 - Query space
 - Answer space
 - Disclosure risk measure
 - Data utility measure
- Example: regression servers
 - Optimally protect sensitive variable
 - Risk: predictive, residual
 - Utility: unweighted, weighted
 - Software under development

$$\mathbf{S} = \left[\begin{array}{c|cc} s_{00} & \mathbf{s}_{\text{supp}}^t & \mathbf{s}_{\text{free}}^t \\ \hline \mathbf{s}_{\text{supp}} & & \\ \mathbf{s}_{\text{free}} & & \mathbf{S}_D \end{array} \right]$$

Accomplishments—4

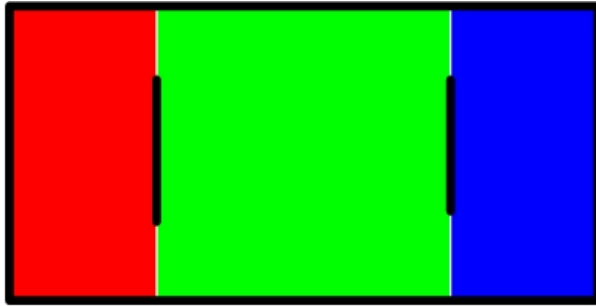
Secure Regression for Distributed Data

- Horizontally partitioned data
 - $K > 2$ agencies with the same attributes for different sets of data subjects
- Usual linear model $y = X\beta + \varepsilon$
- Use secure summation (a form of secure multi-party computation) to compute $X^T X$ and $X^T y$, from which agencies compute least squares estimators

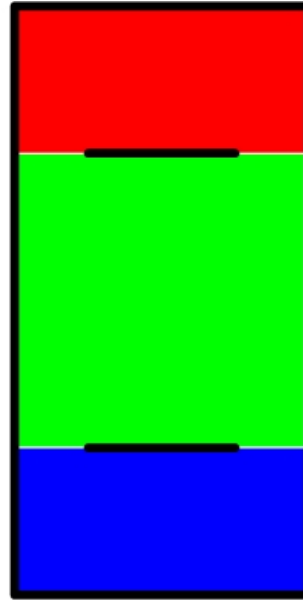
$$\hat{\beta} = (X^T X)^{-1} X^T y$$

- Can also use secure summation to calculate standard errors, R^2 , ...
- Diagnostics via securely integrated synthetic residuals

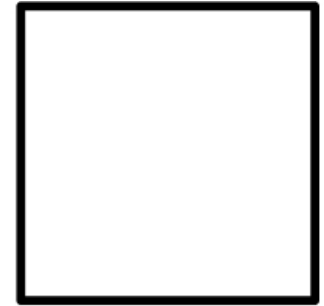
How Secure Regression is Done



$$X^T: p * (n_1 + n_2 + n_3)$$



$$X: (n_1 + n_2 + n_3) * p$$

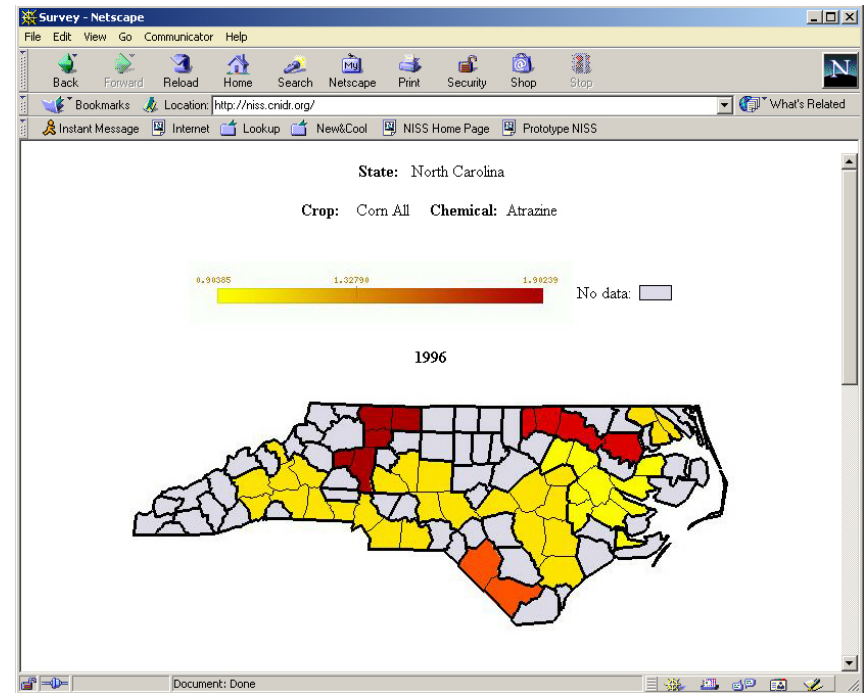
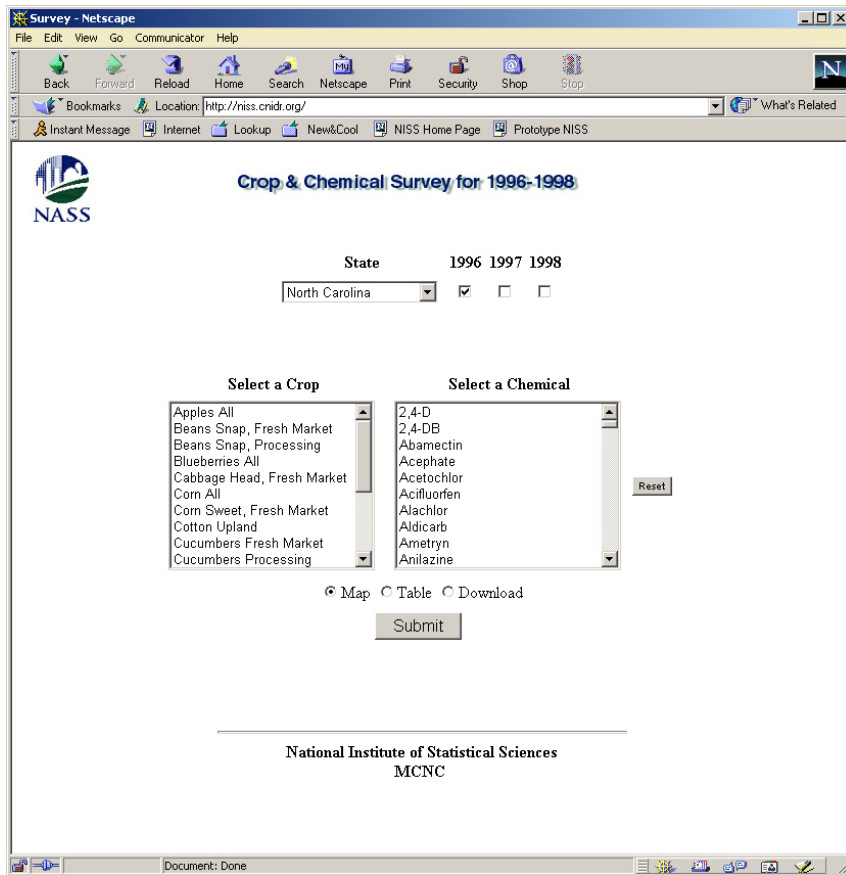


$$X^T X: p * p$$

Accomplishments—5

- Algorithms for geographical aggregation to achieve disclosability
 - Aggregate adjacent counties into supercounties that satisfy (n=3, p=60%)-rule
 - Developed on NASS chemical use survey data
 - Employed by NASS
- Bayesian analysis of consequences of aggregation for statistical inference
 - The “higher-order” the characteristic, the more aggregation hurts

Software Product: Aggregation and Visualization



Current Foci

- Additional secure analyses of distributed data
 - Regression for vertically partitioned data
 - More complex partitioning
 - Categorical data: secure contingency tables
 - Secure pooling to build list of non-zero cells
 - Secure summation to compute values of non-zero cells
 - Kernel methods
 - Principal components analysis
 - Support vector machines
 - Data mining
 - Classifiers

Current Foci—2

- Record linkage
 - Possible to do securely?
 - Consequences of incorrect linkage for inference
- Connections to computational algebra
 - See www.aimath.org/ARCC/workshops/compalgstat.html

Computational Algebraic Statistics

December 14 to 18, 2003

at the

[American Institute of Mathematics](http://www.aimath.org), Palo Alto, California

organized by

[Jesus A. De Loera](#), [Steven Fienberg](#), [Serkan Hosten](#), [Alan Karr](#), and [Bernd Sturmfels](#)

Challenges

- Scalability
 - Always an issue
 - Some good algorithms do not scale
- Complex remote servers
 - Query interaction
 - User equity
- Data integration
 - Distributed databases
 - Effects of incorrect integration
- Data quality
 - Problem completely overwhelms existing tools

Project Structure

- Lead institution: NISS
- University partners: Carnegie Mellon, Duke, Iowa State, Purdue, Southern Methodist
- Federal statistical agency partners providing testbed databases, collaborators and financial support: BLS, BTS, Census, NASS, NCES
- Shared problem definition
- Feedback
 - Ongoing conversations
 - Monthly E-mail updates
 - NISS Affiliates Data Confidentiality Technology Days

Broad Impact

- Protect government-collected data on individuals and establishments from increasingly severe threats to confidentiality
- Protect privacy of individuals and establishments
- Prevent Federal data warehouses from becoming data cemeteries
- Assist agencies in preparing for a “world without releasable microdata”