# Web-Based Systems that Disseminate Information but Protect Confidential Data

Alan F. Karr
Ashish P. Sanil
National Institute of Statistical Sciences
{karr, ashish}@niss.org

# Fundamental Tension

- Confidentiality of Federal data (e.g., Census)
- Privacy of data subjects (individuals, establishments)
- Agencies' mandates to disseminate information
- Citizens' right to know

# High-Level Goals

Merge statistics and IT to

- Disseminate information derived from data
- Preserve confidentiality
- Serve customer needs in new ways
- Build systems that work at realistic scales and serve as testbeds for development of new disclosure limitation methodology

# Examples

- NASS system
- Table servers

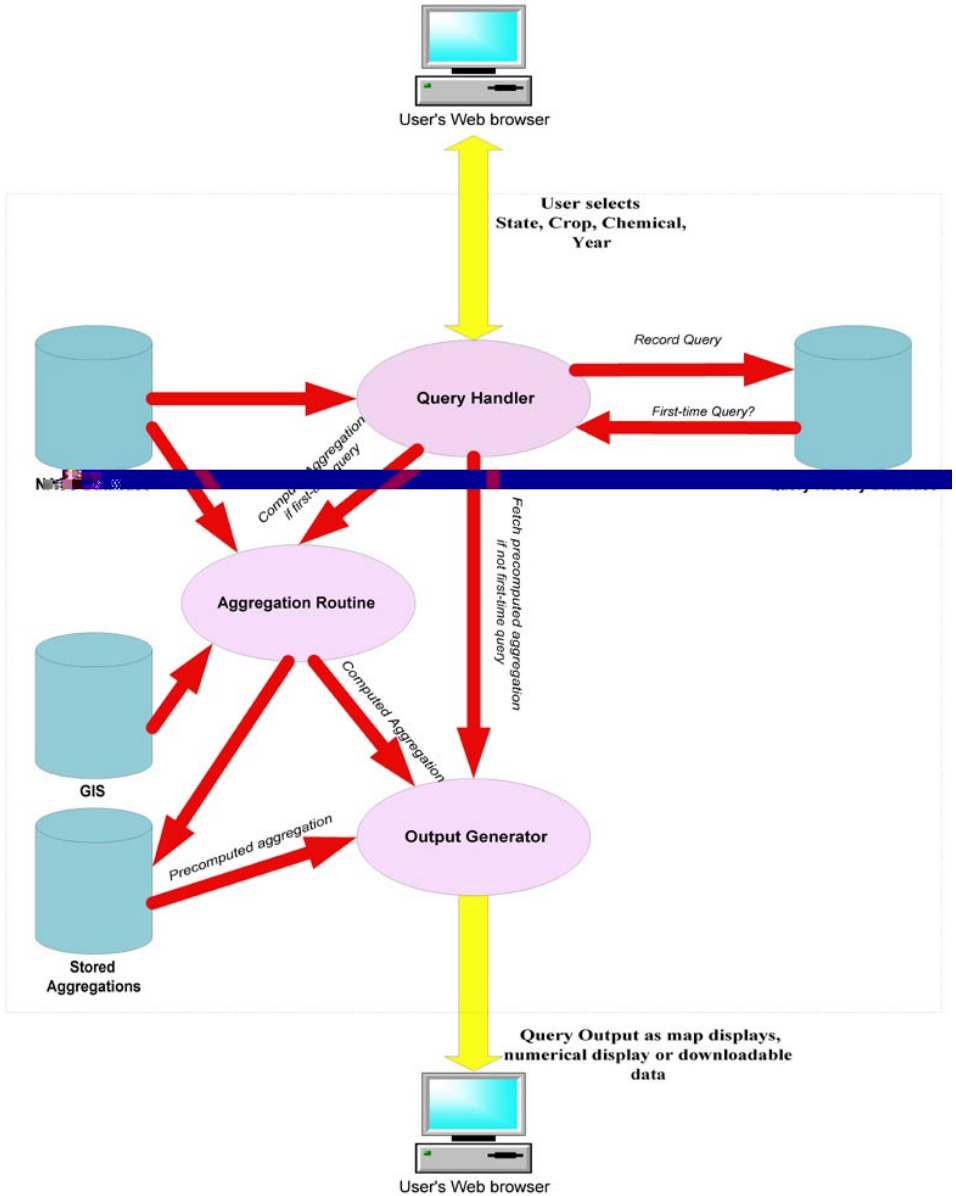
# NASS Data

- Survey of on-farm use of chemicals in 1996-98:
  - Farm ID
  - State and County
  - Year
  - Farm size in acres
  - Crop
  - Chemical
  - Pounds of chemical applied
- 194,410 records: 30,500 farms, 322 chemicals, 67 crops
- Currently, application rates released only at state level. Ideal is to release at county level.

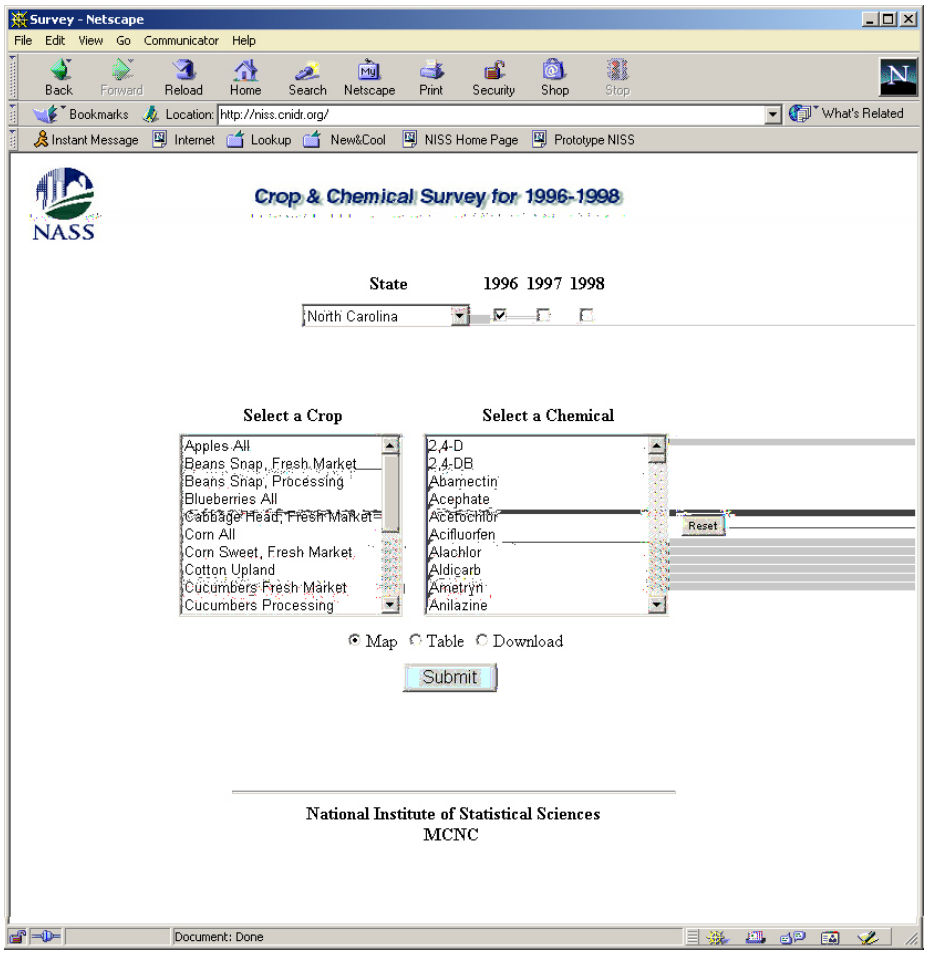
# Confidentiality

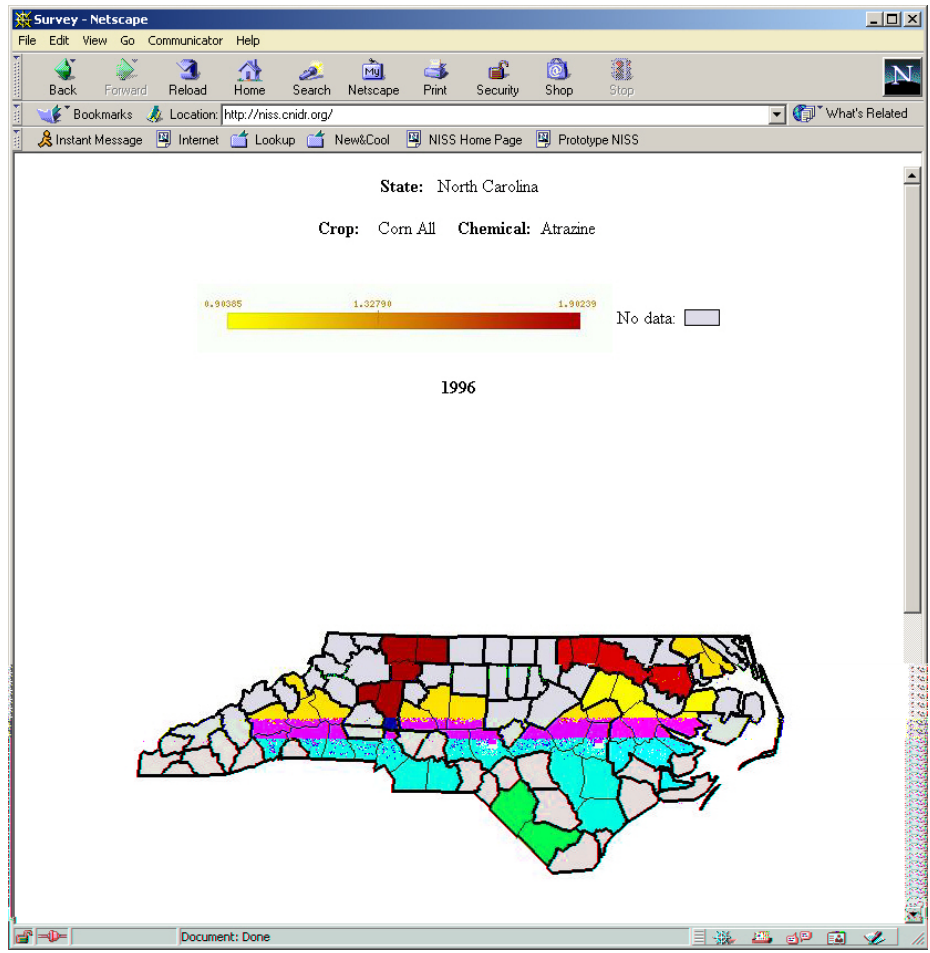
- **Criterion:** (N,p)-rule: Released unit must contain at least N = 3 farms, no one of which comprises more than p = 60% of the total acreage
- **Strategy:** Aggregate adjacent counties into disclosable supercounties, using fast, automatic, effective heuristic methods.

# NASS System Architecture



# System Operation

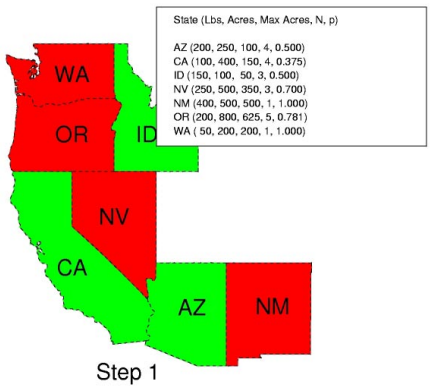


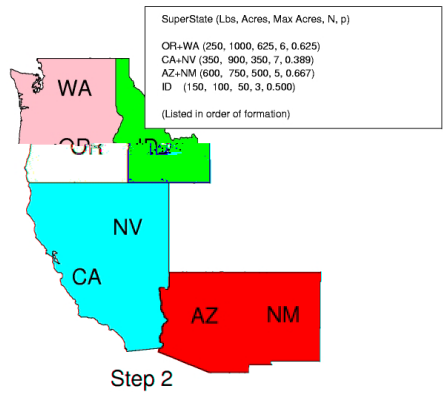
# Algorithm Details

- **Small** version: Create small supercounties
- **Pure** version: Preserve purity of disclosable counties
- **Hybrid**: Small, then pure
- Randomization of order of proposed merges
- Select among candidates randomly or based on similarity of application rates

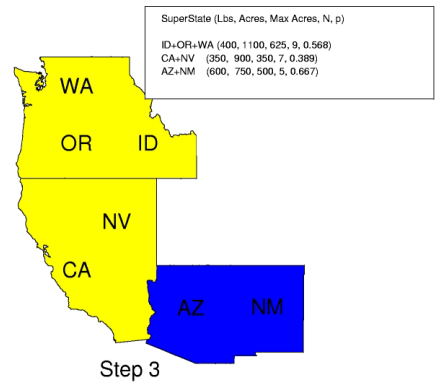
# Example



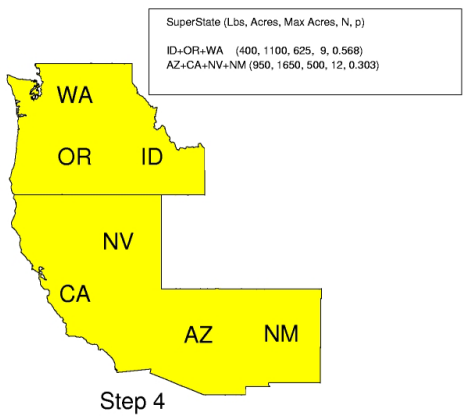
Step 1



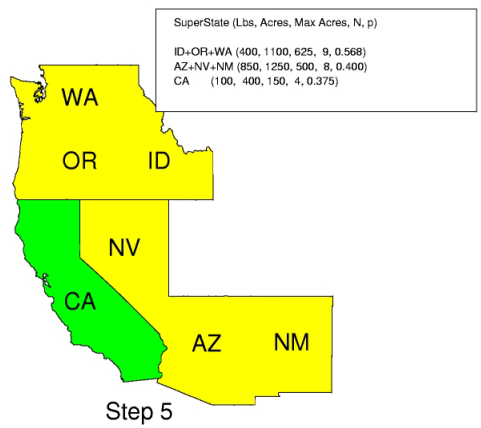
Step 2



Step 3



Step 4



Step 5

# Table Servers

- **Data:** Large (40 variables x 4 categories each) contingency table of counts or sums
- **Queries:** Marginal sub-tables (cross-tabs)
- **Risk:** Identity or attribute disclosure
- Risk is *dynamic*: risk of responding to one query depends on (all) queries that were answered previously

# Key Abstractions

- **Query space Q:** Massive ($2^{40}$), but partially ordered by inclusion
- **Core Releases:** at the start of system operation
- **Released Frontier:** Maximal elements of releases to date
- **Unreleasable Frontier:** Minimal elements of set of unreleasable tables

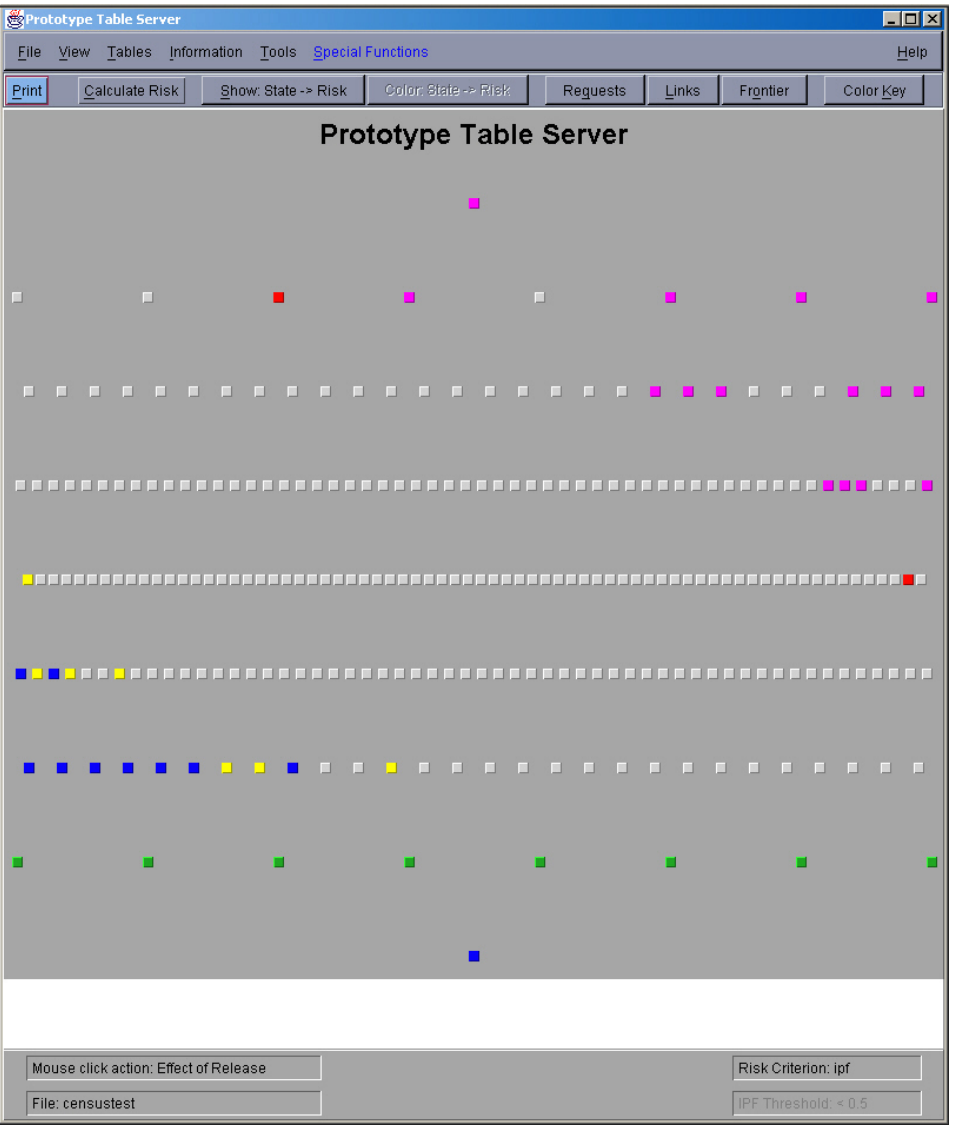
# Key Abstractions - 2

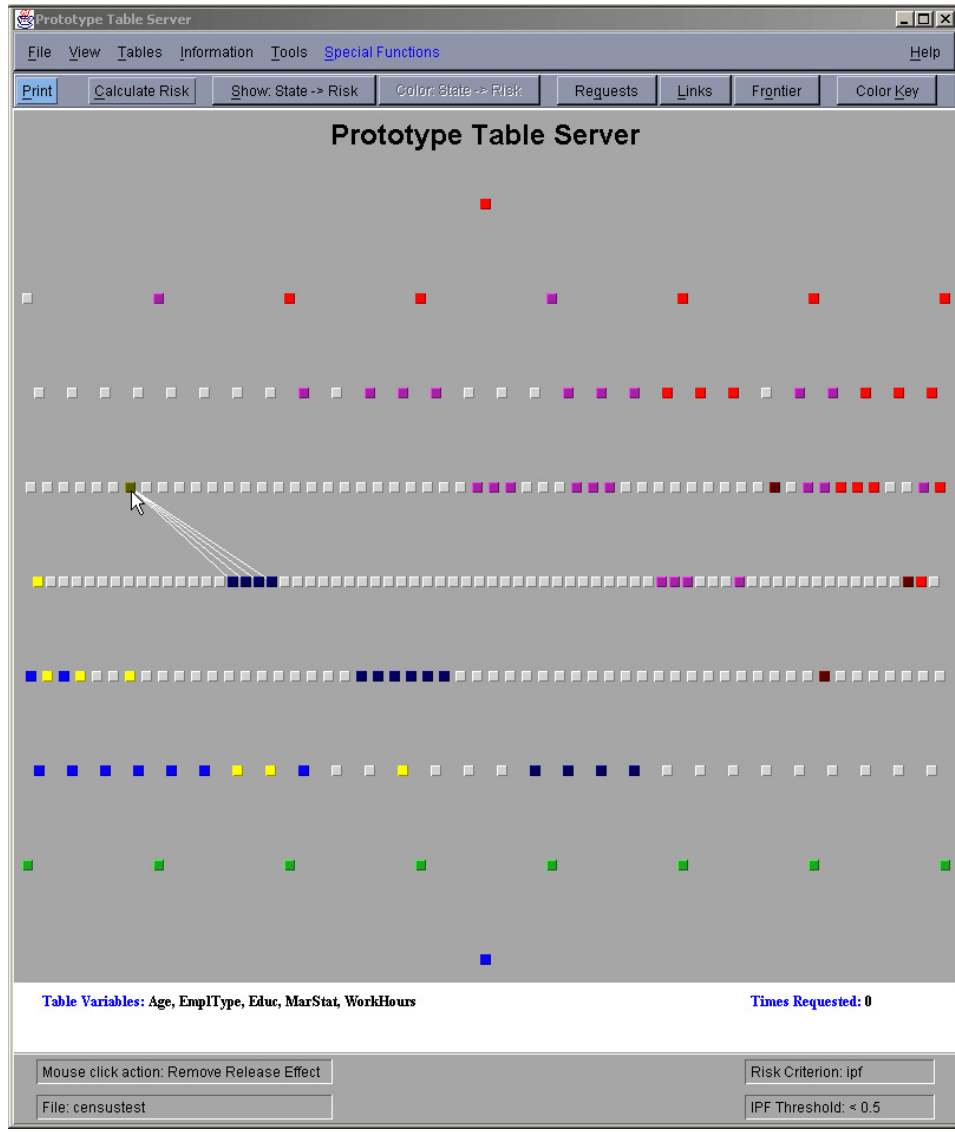
- **Risk Criteria:** Apply to subsets of Q.
  - Bounds for sensitive (e.g., small) entries in full table
  - IPF reconstruction of full table
  - # of tables whose marginals agree with releases
- **Release Rule:** To respond to queries, accommodating
  - Risk threshold if query were answered
  - Increase in unreleasable set
  - *Value* (to whom?) of releasing information

# Other Issues

- **Equity:** Who can move the frontier, and under what circumstances?
- **Scalability:** Most techniques don't scale
- **Risk Reduction:** Responses other than refusing queries.

# Java Prototype



# Additional Information

- Project Web site: www.niss.org/dg
- NASS system: niss.cnidr.org
- Karr, *et al.*, *IEEE Computer* **34 (2)** (2001) 36-37.
- Lee, *et al.*, *Review of Official Statistics* (to appear)