

# NISS

## Data Confidentiality and Statistical Disclosure Limitation: A Quick Overview

Alan F. Karr  
National Institute of Statistical Sciences  
Research Triangle Park, NC 27709  
[karr@niss.org](mailto:karr@niss.org)

# DC from Multiple Perspectives

- Official statistics agencies must
  - Preserve confidentiality of data
  - Preserve privacy of data subjects
  - Maintain quality of data
  - Disseminate useful information
- Holders of proprietary data want to
  - Safeguard IP
  - Advance research to create new products
- Data subjects want protection from threats to
  - Privacy
  - Economic interests

# Forms of Disclosure

- Identity disclosure
  - Record is associated with a particular subject, typically by *record linkage* to another database containing an ID
- Attribute disclosure
  - Value of sensitive attribute is disclosed
- Inferential disclosure
  - Identity or attribute disclosure on a statistical basis
- False positive
  - Intruder acts on basis of incorrect disclosure

# How Easy is It?

- Most people can be identified by
  - Date of birth (MM/DD/YYYY)
  - Gender
  - 5-digit ZIP code
- Finding these items on the web is
  - Easy
  - Generally free (ChoicePoint, crooks and others charge)

# An Experiment



## 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.](#)

<b>Name:</b>	KARR, ALAN FRANCIS
<b>County Name:</b>	ORANGE
<b>Status:</b>	ACTIVE
<b>City:</b>	CHAPEL HILL NC 27516
<b>Race:</b>	WHITE
<b>Ethnicity:</b>	NOT HISPANIC or NOT LATINO
<b>Gender:</b>	Male
<b>Party:</b>	



AnyBirthday.com

846 West St., New York, NY 10001

Born: Sep. 11, 1902

Smith, John R.

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



Locateme.com

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

**[NEW! Anybirthday.com PLUS lists Addresses!](#)**

---

Subject's Name

Birthday

Zip Code

ALAN

F KARR

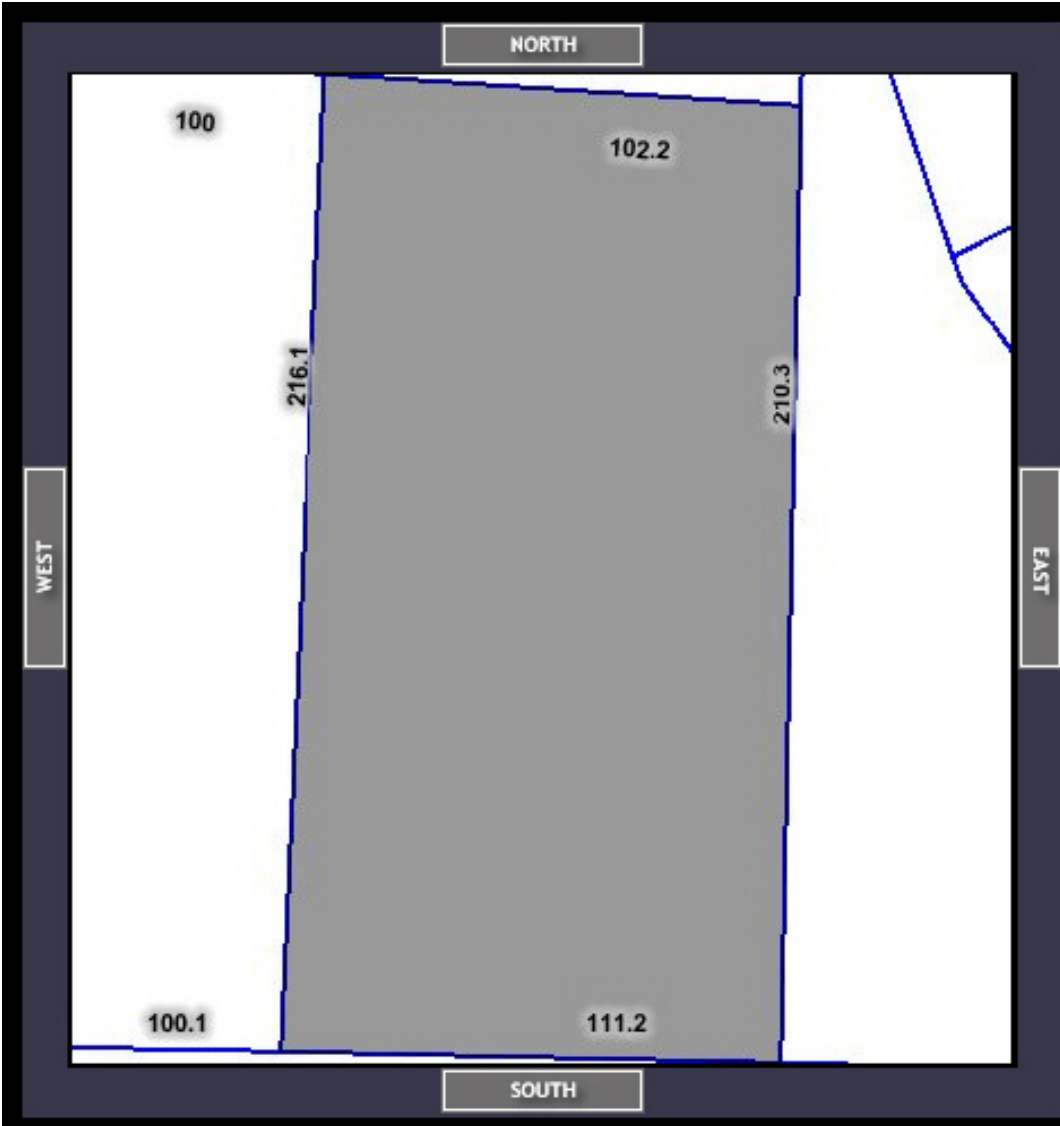
27516

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

---

**PIN NUMBER: 9777686888**

PIN	9777686888
TMBL	7.121.8.14
OWNER	[REDACTED]
OWNER 2	[REDACTED]
MAILING ADDRESS	[REDACTED]
CITY	CHAPEL HILL
STATE	NC
ZIP	27516-[REDACTED]
DEED REFERENCE	1261/54
TRACT NUMBER	707429
2002 VALUATION	[REDACTED]
SIZE	L1
RATE CODE	04
DESCRIPTION	#6 WOODCREST



# The Fundamental Issue: Tradeoffs Between

- Confidentiality protection
  - Mandated by law
  - Imposed by regulation
  - To maintain quality
- Data utility, to support
  - Policy formulation and evaluation
  - Research, especially statistical inference

# Risk-Utility Formulations

- Components
  - Database  $\mathcal{D}$
  - Set  $\mathcal{R}$  of *candidate releases*  $\mathbf{R} = f(\mathcal{D})$
  - Disclosure risk function  $\mathbf{DR}(\mathbf{R})$
  - Data utility function  $\mathbf{DU}(\mathbf{R})$
- Goal: Select the “best release”
  - Maximize utility subject to constraint on risk
  - Select from risk-utility frontier

# High-Level View of SDL

- Restricted access
  - To approved individuals, for approved analyses, at a restricted data center, at a cost, under additional restrictions
- Restricted data: “the truth but not the whole truth”
  - Drop attribute
  - Coarsen categories: Geographical aggregation, top-coding
- Altered data: not the truth
  - Microaggregation
  - Data swapping
  - Perturbation
  - Synthetic data

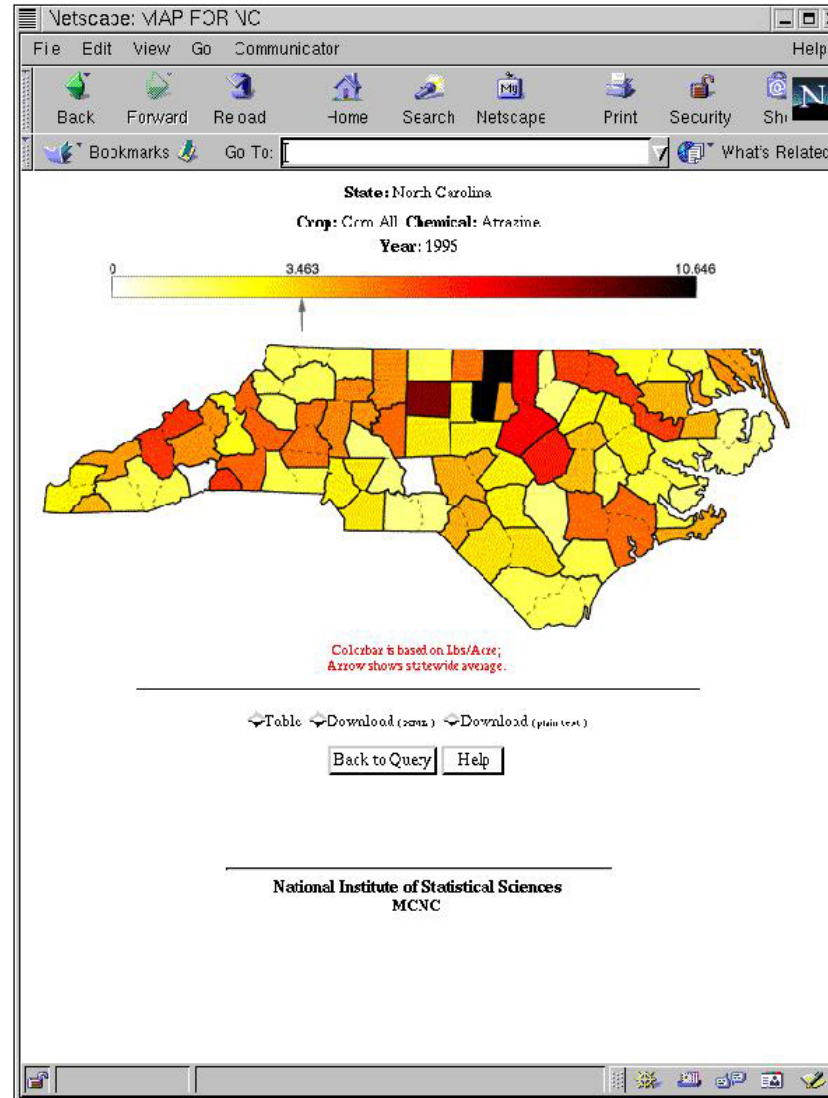
# High-Level View of SDL—2

- Servers
  - Disseminate analyses rather than data
- Poor quality data = “the best defense” ?
- Hope to err on the side of confidentiality

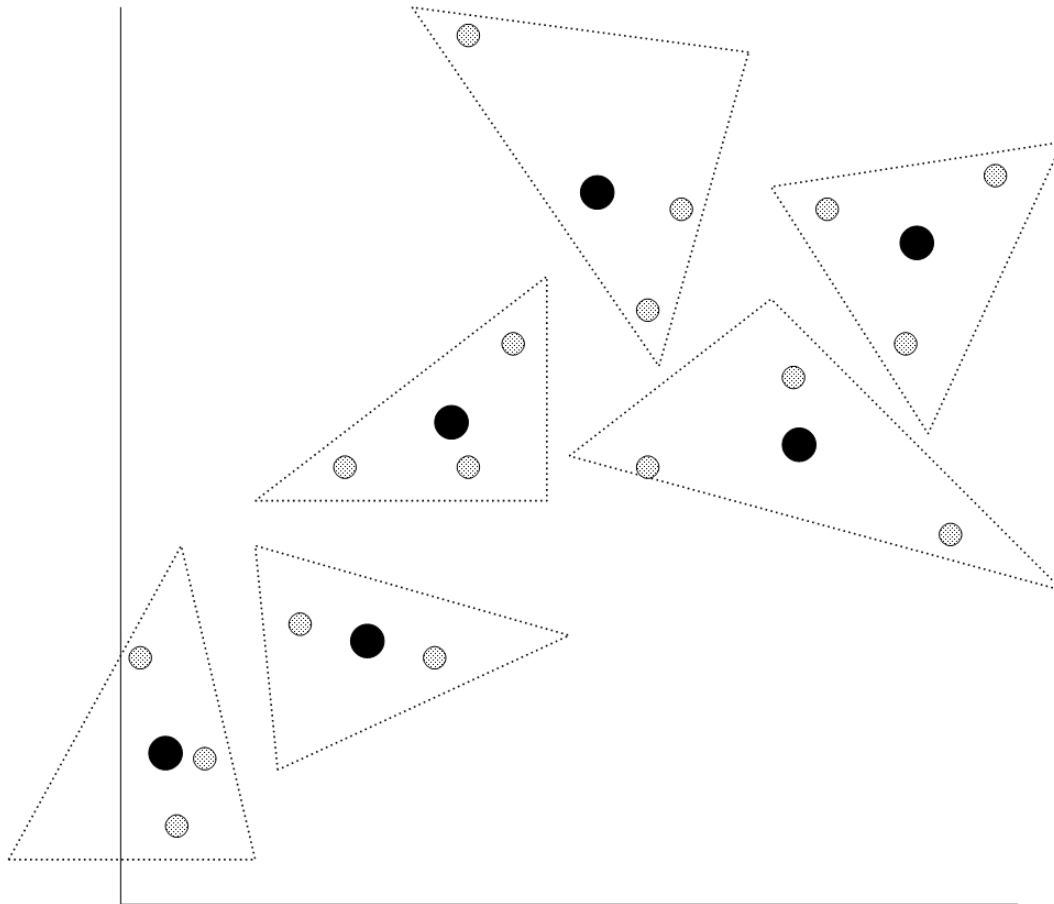
# Sampler of SDL Techniques

- To be illustrated
  - Geographical aggregation
  - Microaggregation
  - Data swapping
  - Servers
- Others include
  - Sample from the data
  - Cell suppression for tabular data
  - Jittering

# Geographical Aggregation



# Microaggregation



# Data Swapping (CPS data)

Record	Age	EmplType	Educ	MarStat	Race	Sex	AveHours	Salary
1	<25	Gov	HS	Marr	W	M	40	<\$50K
2	25-55	SE	Bach	Marr	NW	M	>40	<\$50K
3	25-55	Gov	Bach+	Unmarr	NW	F	>40	>\$50K
4	>55	Priv	Bach	Unmarr	W	F	>40	<\$50K
5	<25	Other	SomeColl	Marr	W	M	40	>\$50K
6	>55	Priv	Bach+	Marr	NW	F	40	>\$50K

Record	Age	EmplType	Educ	MarStat	Race	Sex	AveHours	Salary
1	>55	Gov	HS	Marr	W	M	40	<\$50K
2	25-55	SE	Bach	Marr	NW	M	>40	<\$50K
3	<25	Gov	Bach+	Unmarr	NW	F	>40	>\$50K
4	>55	Priv	Bach	Unmarr	W	F	>40	<\$50K
5	25-55	Other	SomeColl	Marr	W	M	40	>\$50K
6	<25	Priv	Bach+	Marr	NW	F	40	>\$50K

# Synthetic Data

- Basic paradigm
  - Fit a statistical model to the confidential data
  - Use the model in Monte Carlo mode to synthesize a database of the same size as the original one
  - Disseminate the synthetic data
- Advantages
  - Risk low: records aren't real
- Disadvantages
  - Utility may be low: analysis on synthetic data may not yield same result as on original data

# Emerging Idea 1: Servers

- Web-based systems to which users submit queries for analyses of a confidential database
- Servers must
  - Assess risk, taking into account interactions with previously answered queries
  - Assess utility, accounting for queries that become unanswerable
  - Decide whether and how to respond, keeping in mind that a denial may be informative

# Emerging Idea 2:

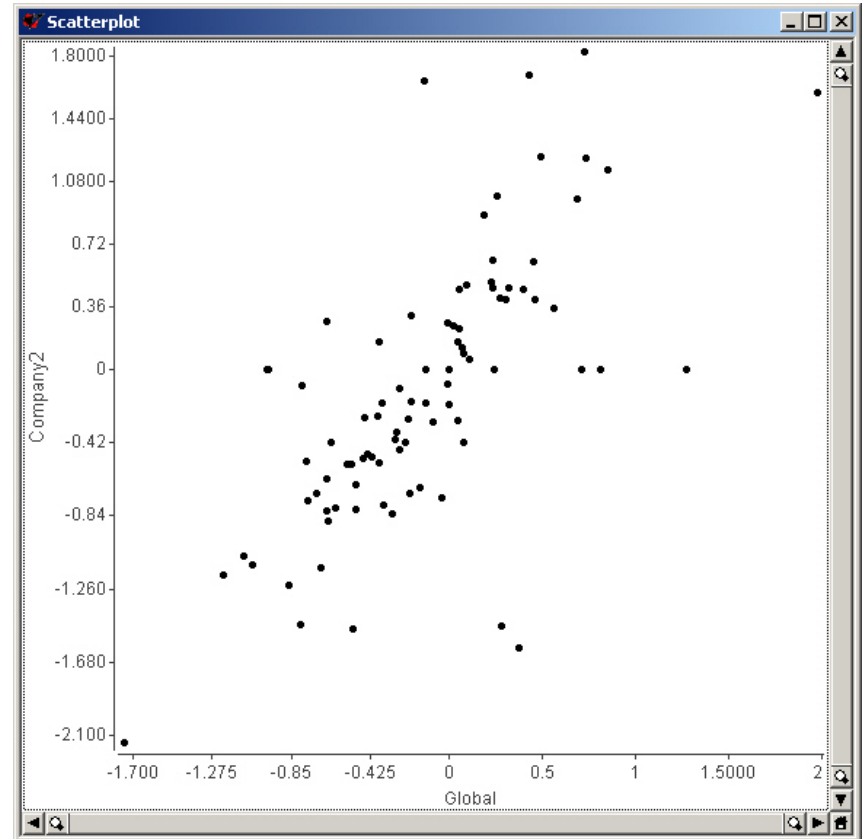
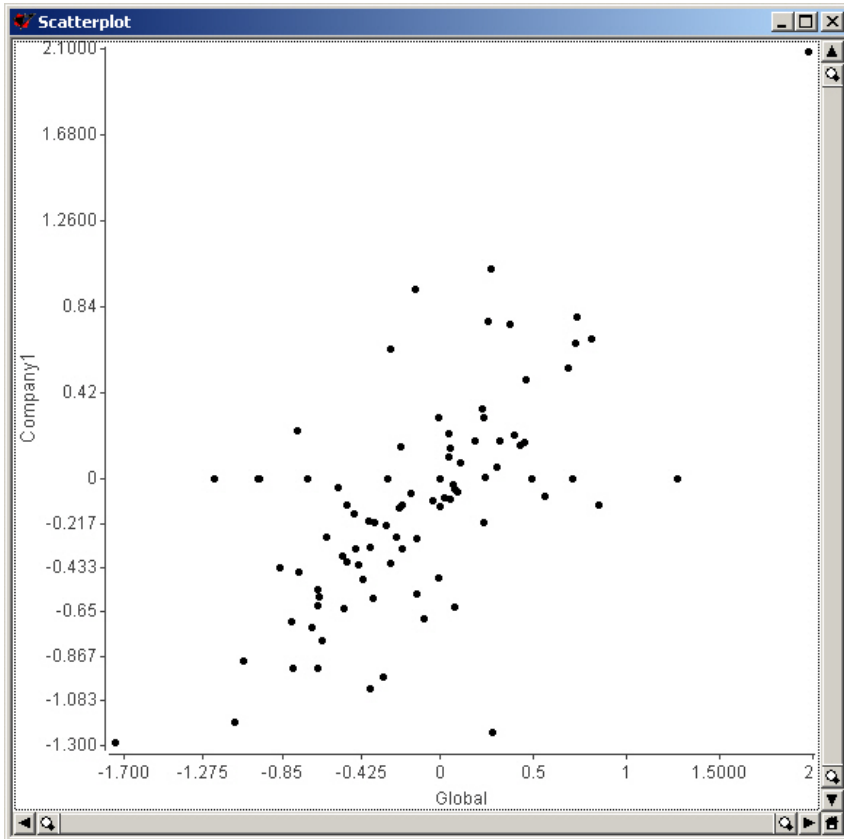
## Secure Analysis of Distributed Data

- Related databases held by multiple “agencies”
  - Example: local employment data
- Agencies wish to perform sound statistical analyses on integrated data, but
  - Actual data integration impossible
  - Other constraints are present (no trusted third party)
- Approach: use secure multi-party computation to share data summaries that are sufficient to perform the analysis

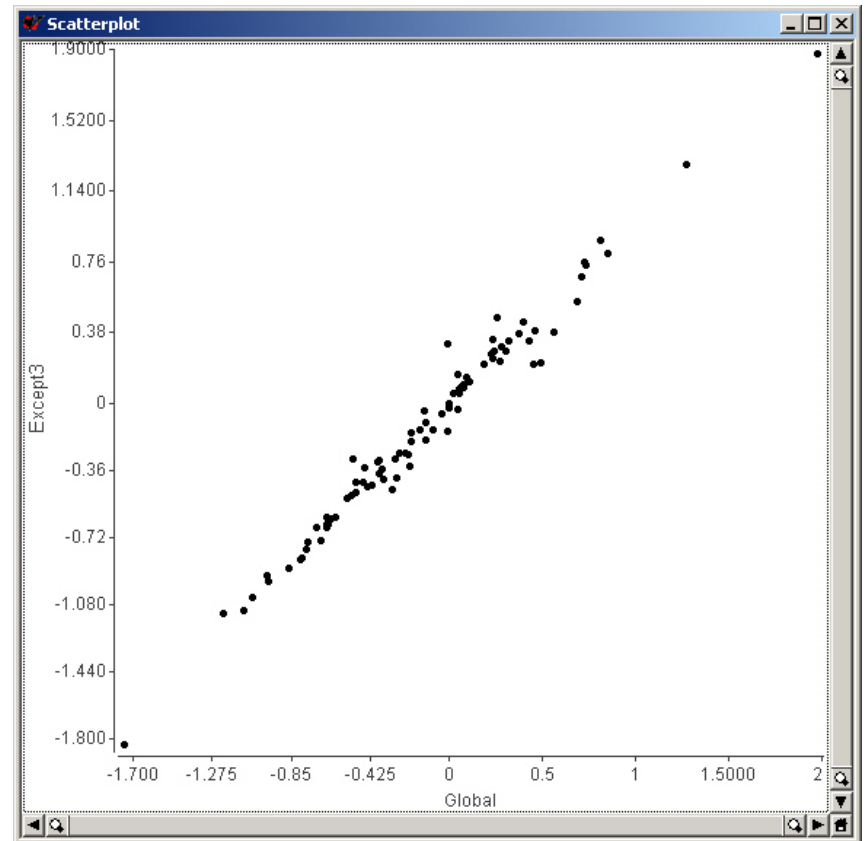
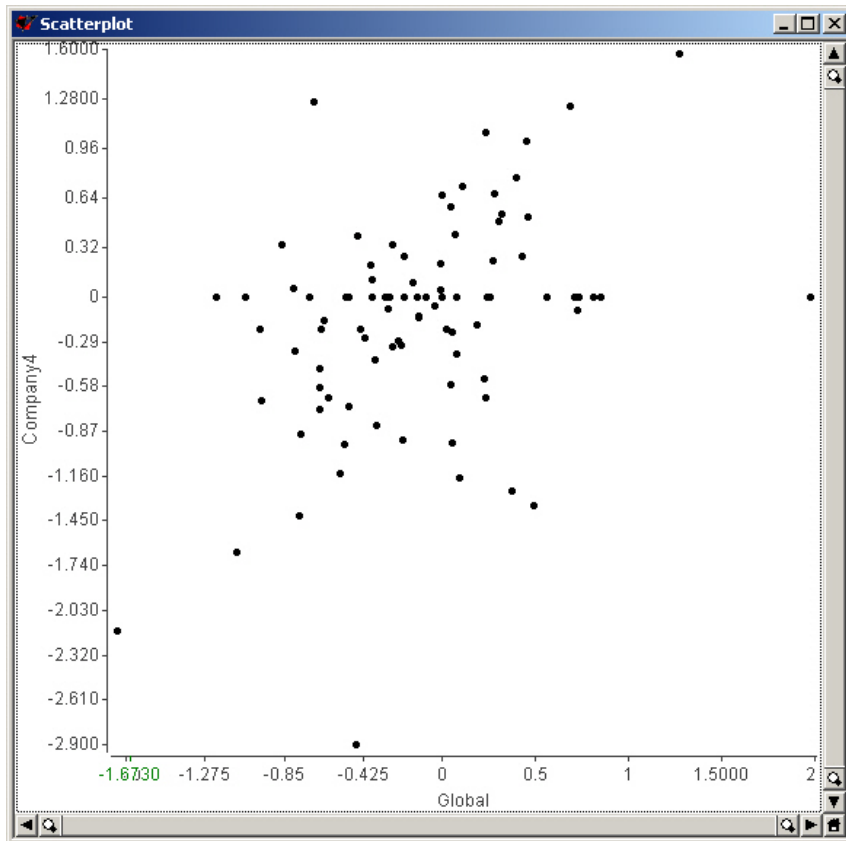
# Example: Chemical Data from Multiple Pharmaceutical Manufacturers

- Data
  - 1318 molecules
  - Response: water solubility
  - Predictors: 90 molecular descriptors + constant
- 4 companies
  - Each company's data are relatively homogeneous, but with gaps!
  - Numbers of molecules = 499, 572, 16 (!), 231
- Analysis: linear regression

# Results

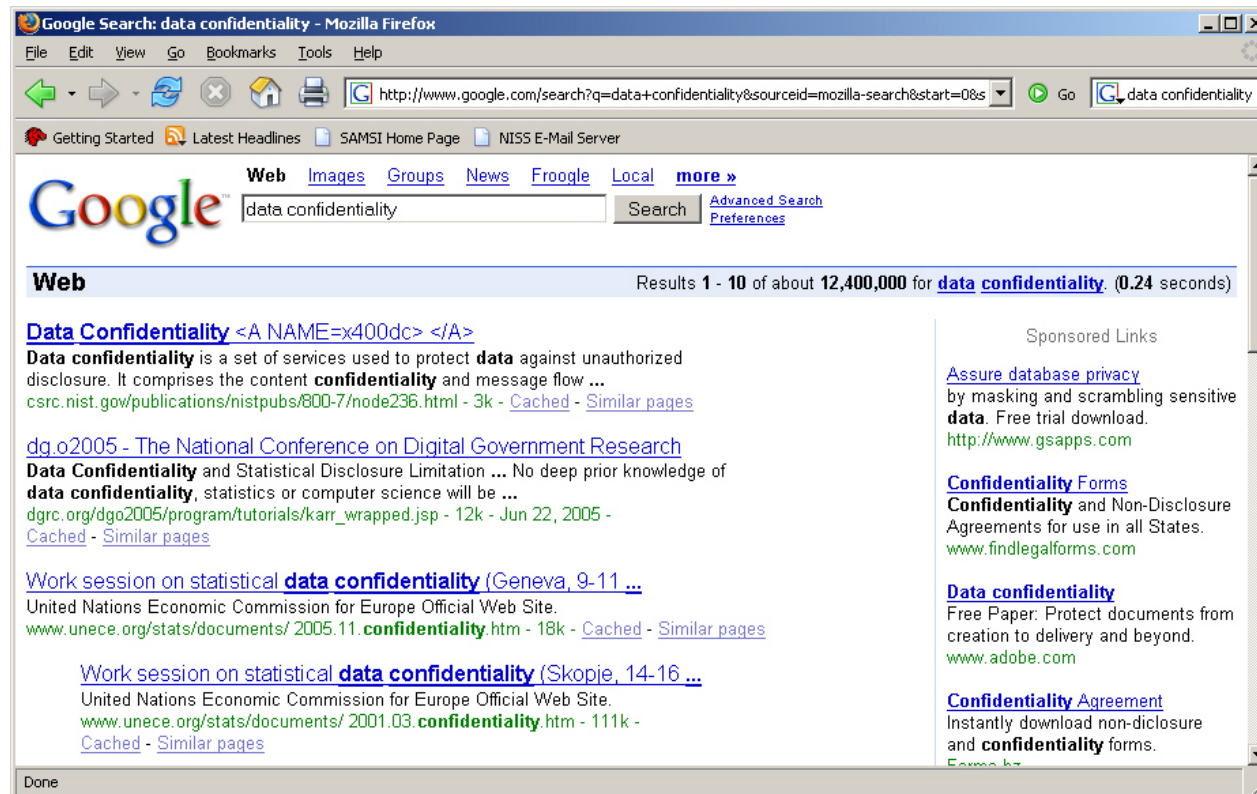


# Results—2

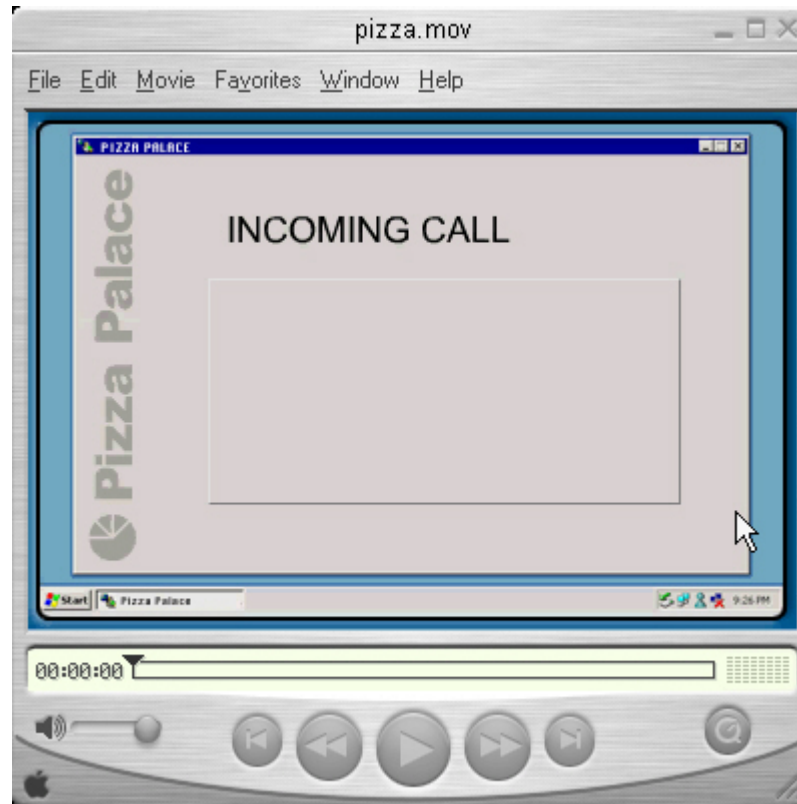


# More Information

- NISS DG project web site: [www.niss.org/dgii](http://www.niss.org/dgii)
- Google



# What's the Future?



<http://www.aclu.org/pizza/index.html?orgid=EA071904&MX=1414&H=1>