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

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

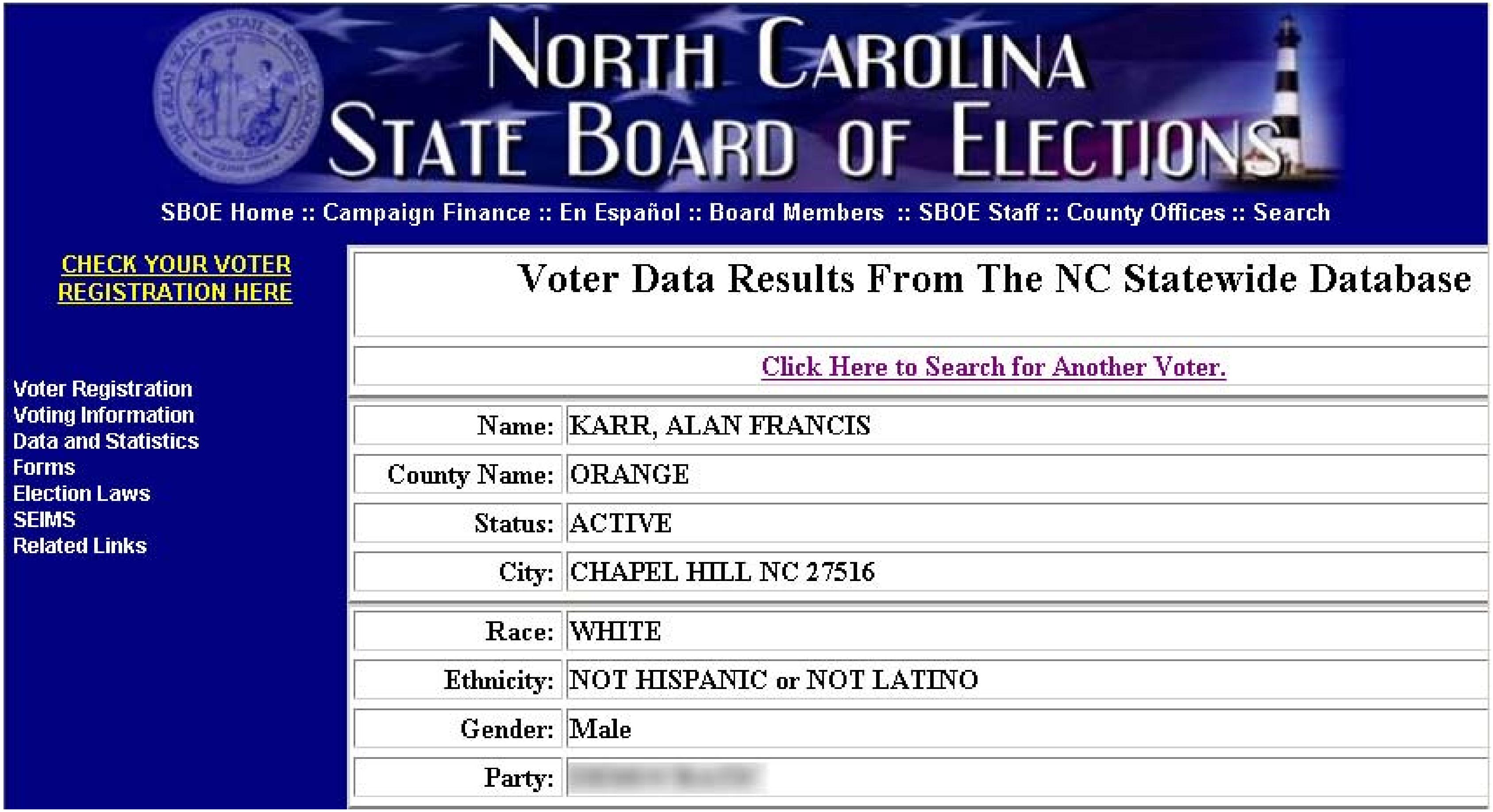
Forms of Disclosure

- 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)

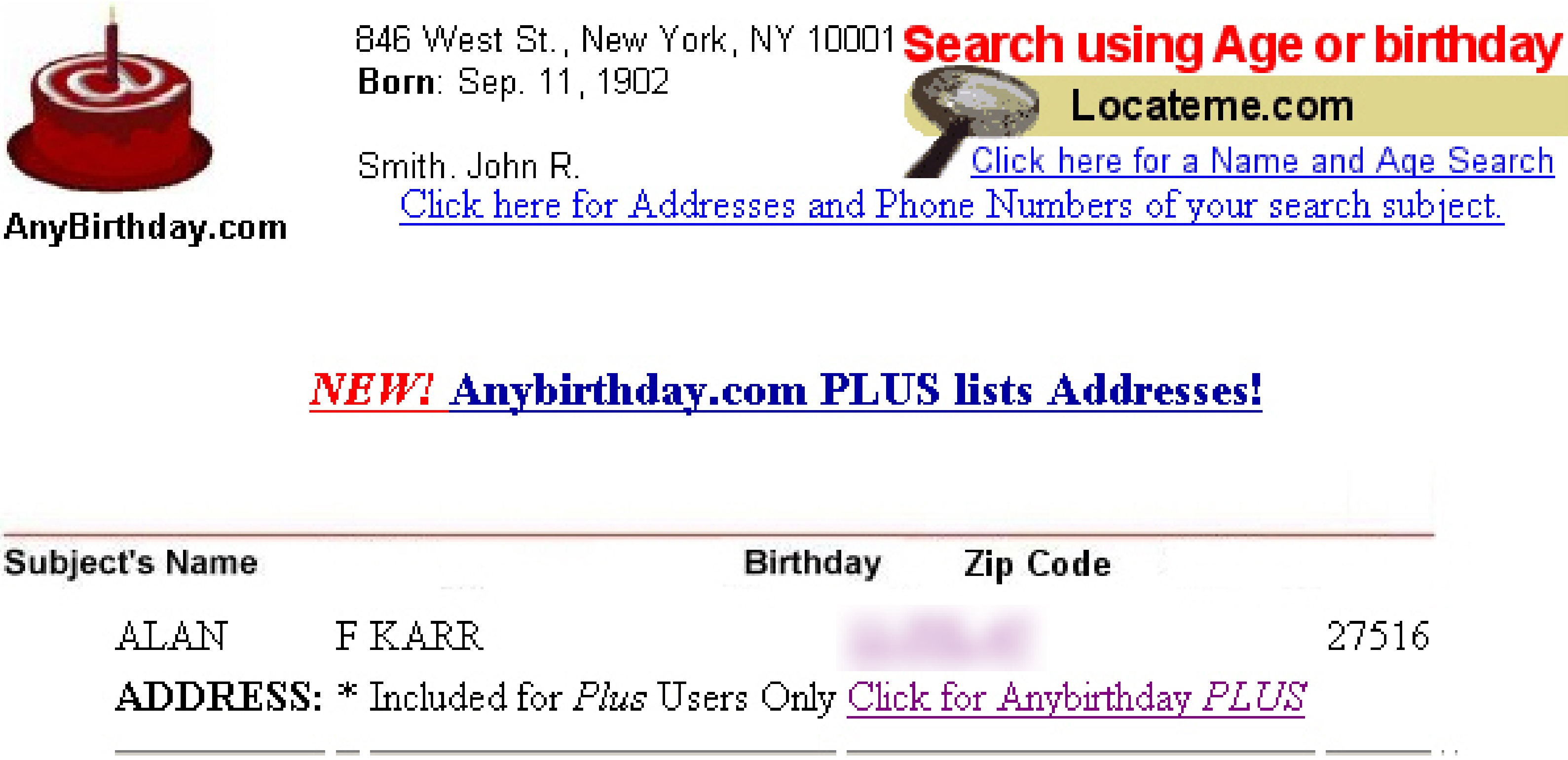
An Experiment


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.

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:	[redacted]



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.

Search using Age or birthday

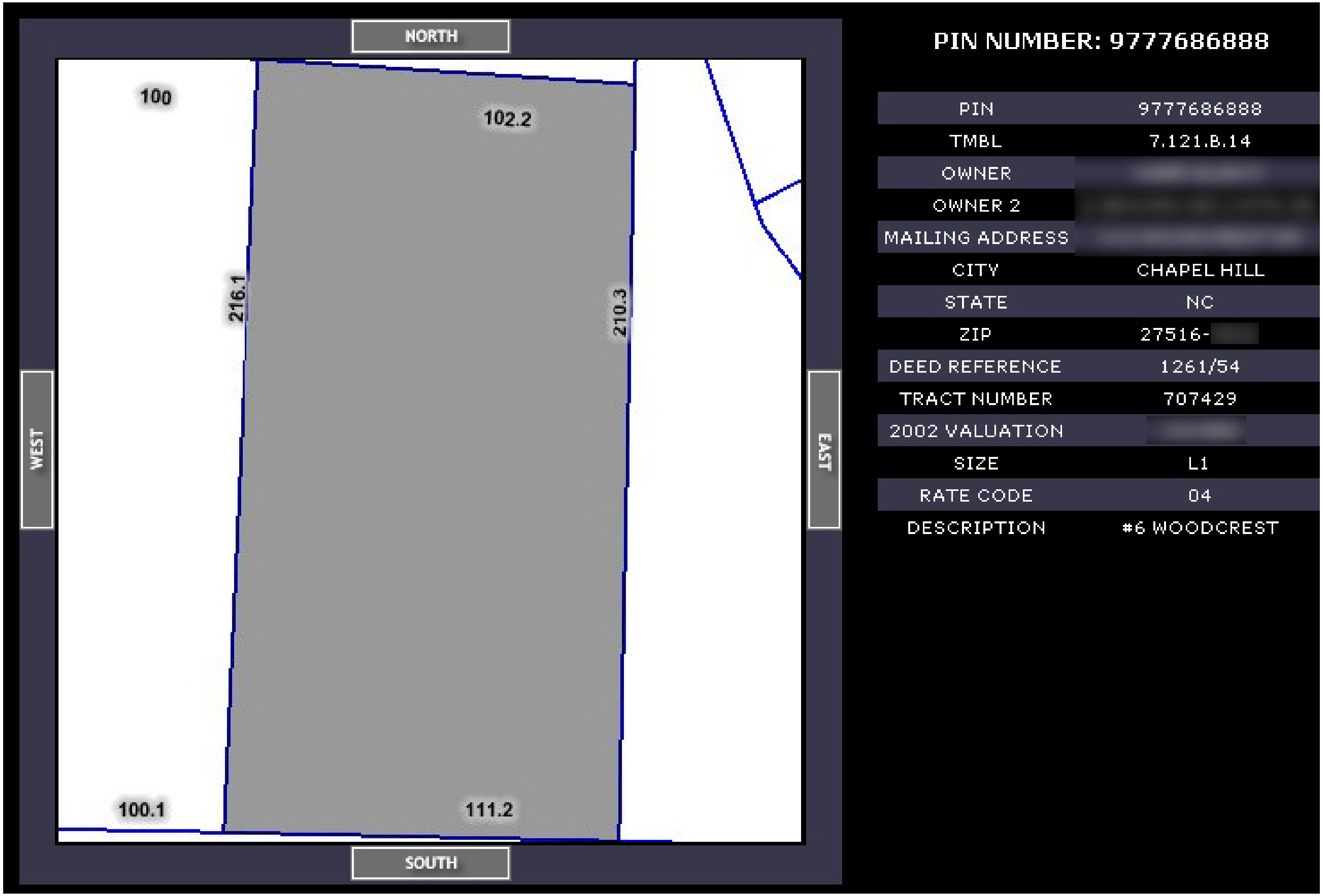
Locateme.com

Click here for a Name and Age Search

NEW! **Anybirthday.com PLUS lists Addresses!**

Subject's Name		Birthday	Zip Code
ALAN	F KARR	[redacted]	27516

ADDRESS: * Included for *Plus* Users Only Click for Anybirthday *PLUS*



PIN NUMBER: 9777686888

PIN	9777686888
TMBL	7.121.B.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* $R = f(\mathcal{D})$
  - Disclosure risk function **DR**(R)
  - Data utility function **DU**(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

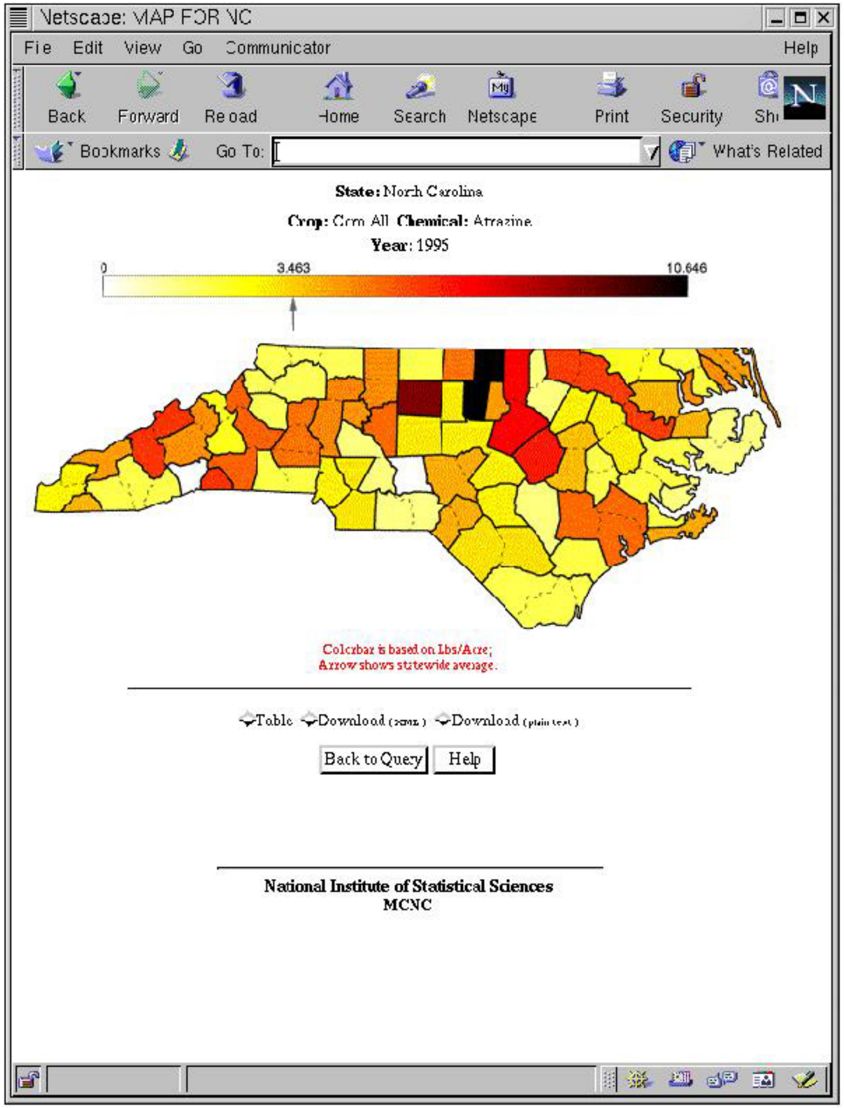
High-Level View of SDL—2

- 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


Netscape: MAP FOR NC
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop
Bookmarks Go To: What's Related
0
3.463
10.646

State: North Carolina

Crop: Corn All **Chemical:** Atrazine

Year: 1995

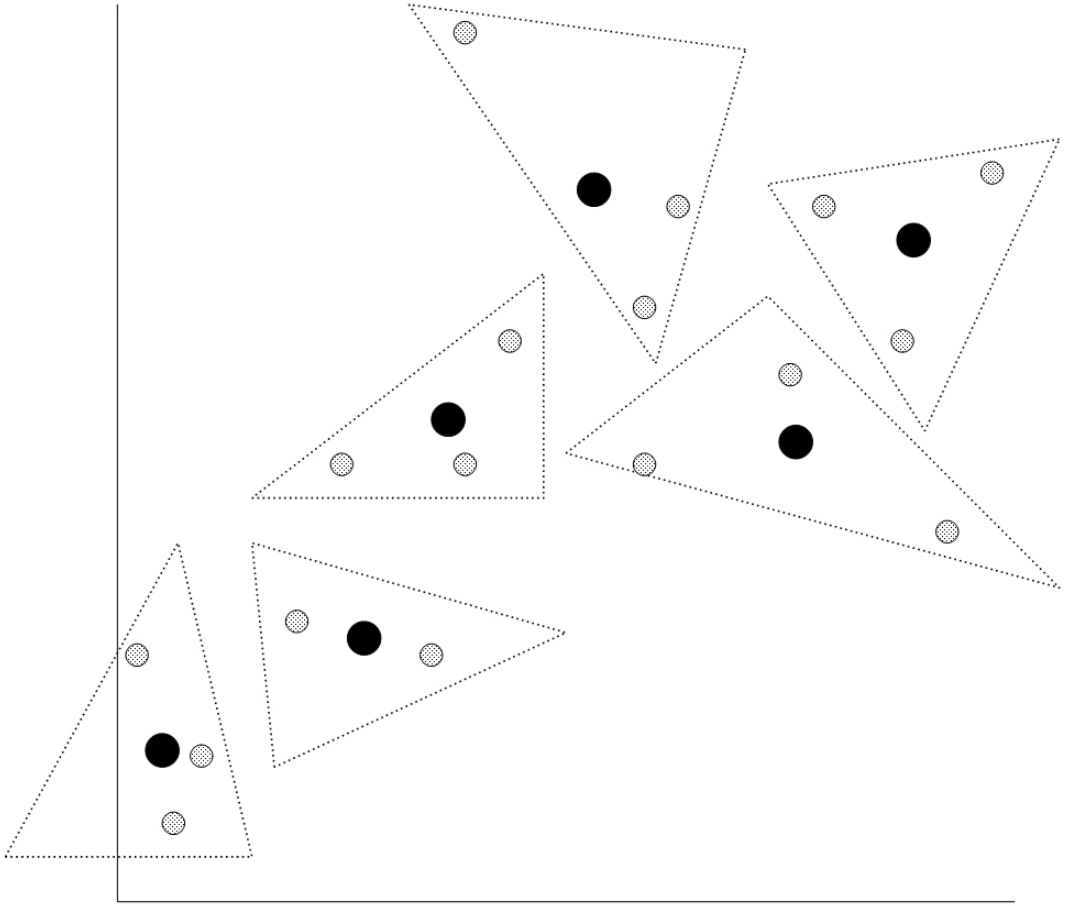
Colorbar is based on Lbs/Acre;
Arrow shows statewide average.

◇Table ◇Download (xgmx) ◇Download (plain text)

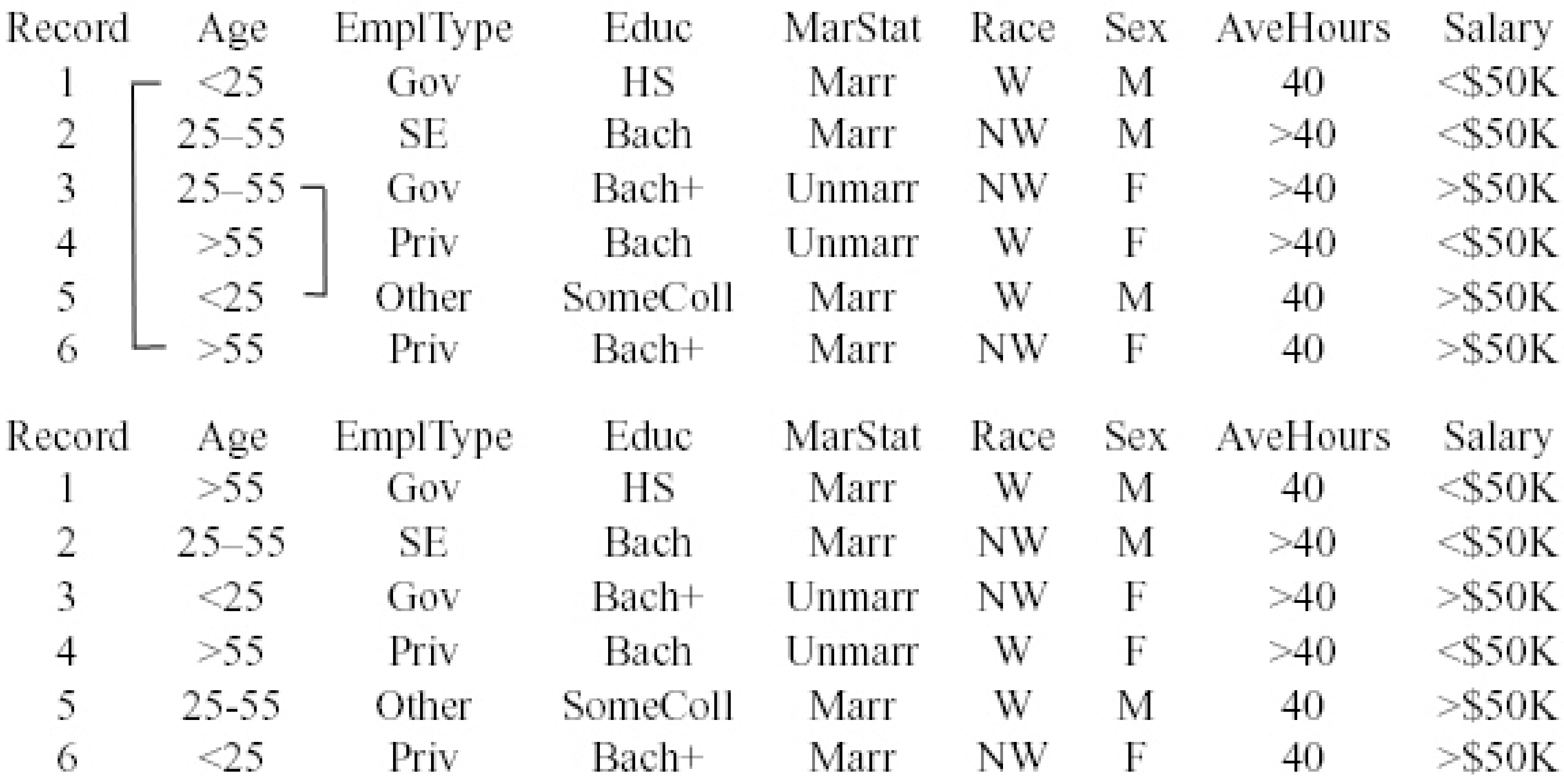
Back to Query Help

National Institute of Statistical Sciences
MCNC

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: does 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

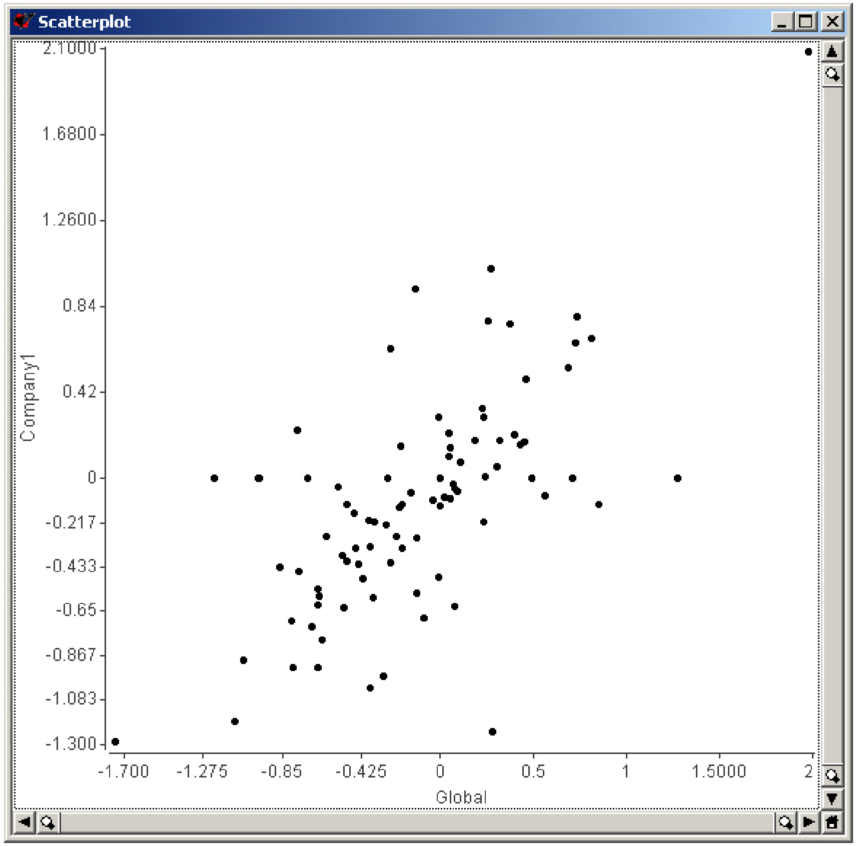
Emerging Idea 2:
Secure Analysis of Distributed Data

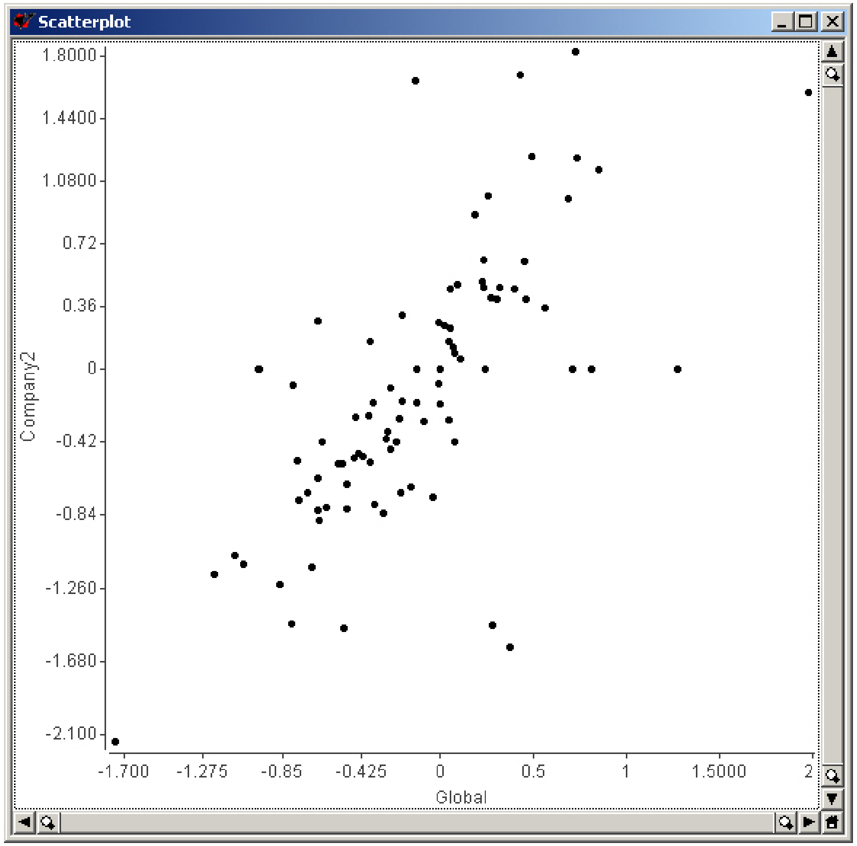
- Related databases held by multiple “agencies”
  - – Example: local employment data
- Agencies wish to perform sound statistical analyses on integrated data, but
  - – Actual data integration impossible
  - – Other constraints are present (no trusted third party)
- Approach: use secure multi-party computation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

Example: Chemical Data from Multiple Pharmaceutical Manufacturers

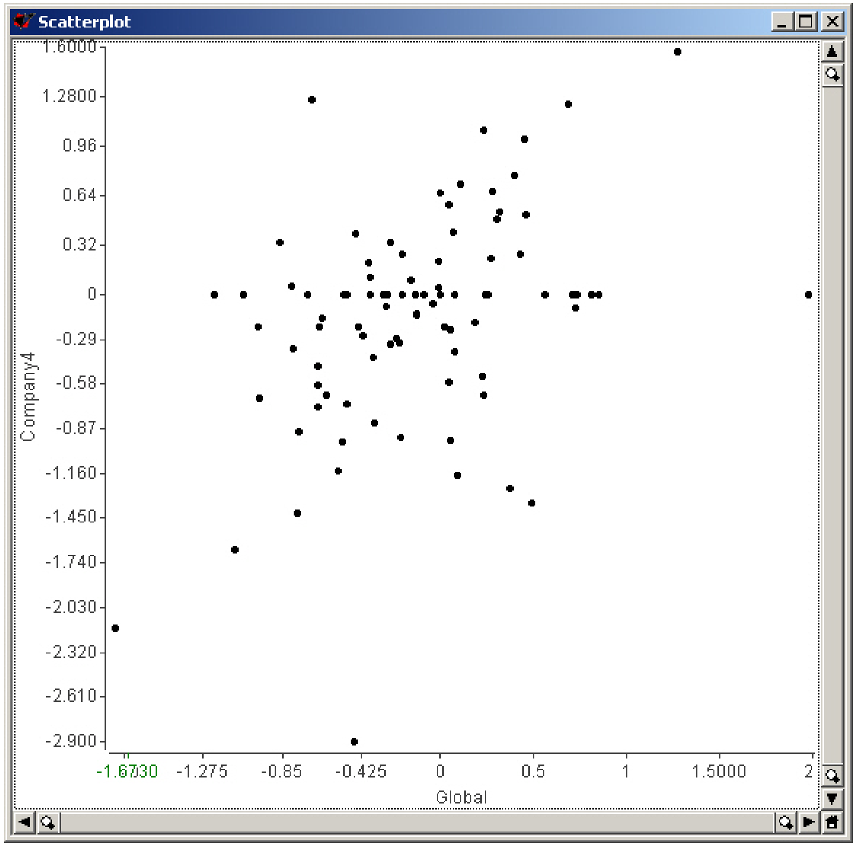
- Data
  - 1318 molecules
  - Response: water solubility
  - Predictors: 90 molecular descriptors + constant
- 4 companies
  - Each company’s data are relatively homogeneous, but with gaps!
  - Numbers of molecules = 499, 572, 16 (!), 231
- Analysis: linear regression

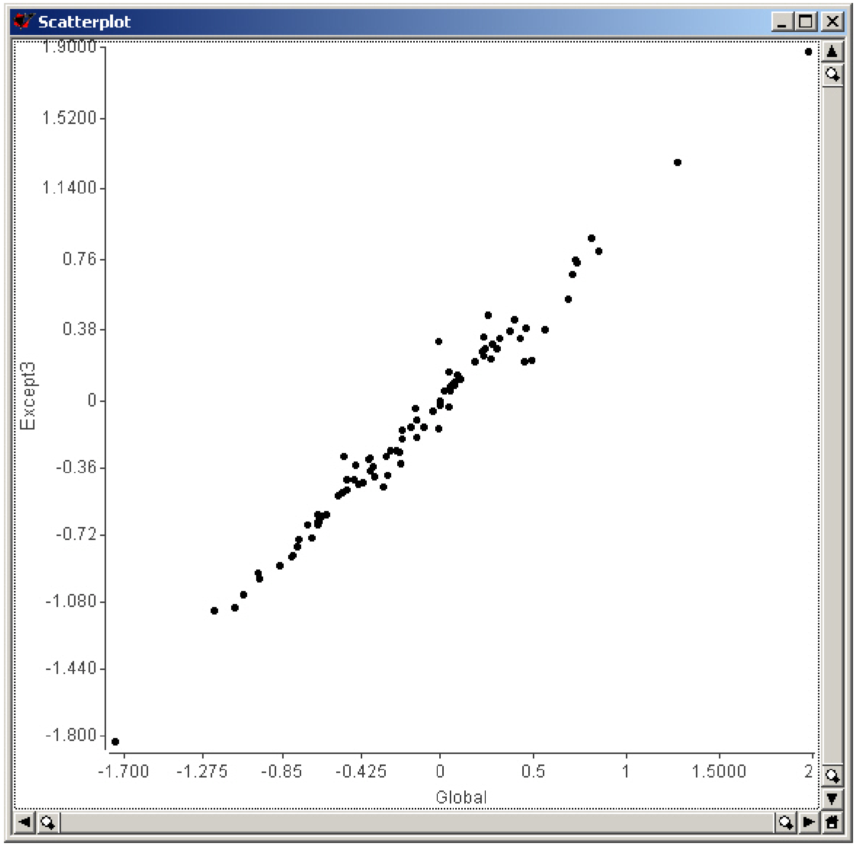
Results


Scatterplot
2.1000
1.6800
1.2600
0.84
0.42
0
-0.217
-0.433
-0.65
-0.867
-1.083
-1.300
Company1
-1.700
-1.275
-0.85
-0.425
0
0.5
1
1.5000
2
Global


Scatterplot
1.8000
1.4400
1.0800
0.72
0.36
0
-0.42
-0.84
-1.260
-1.680
-2.100
Company2
-1.700
-1.275
-0.85
-0.425
0
0.5
1
1.5000
2
Global

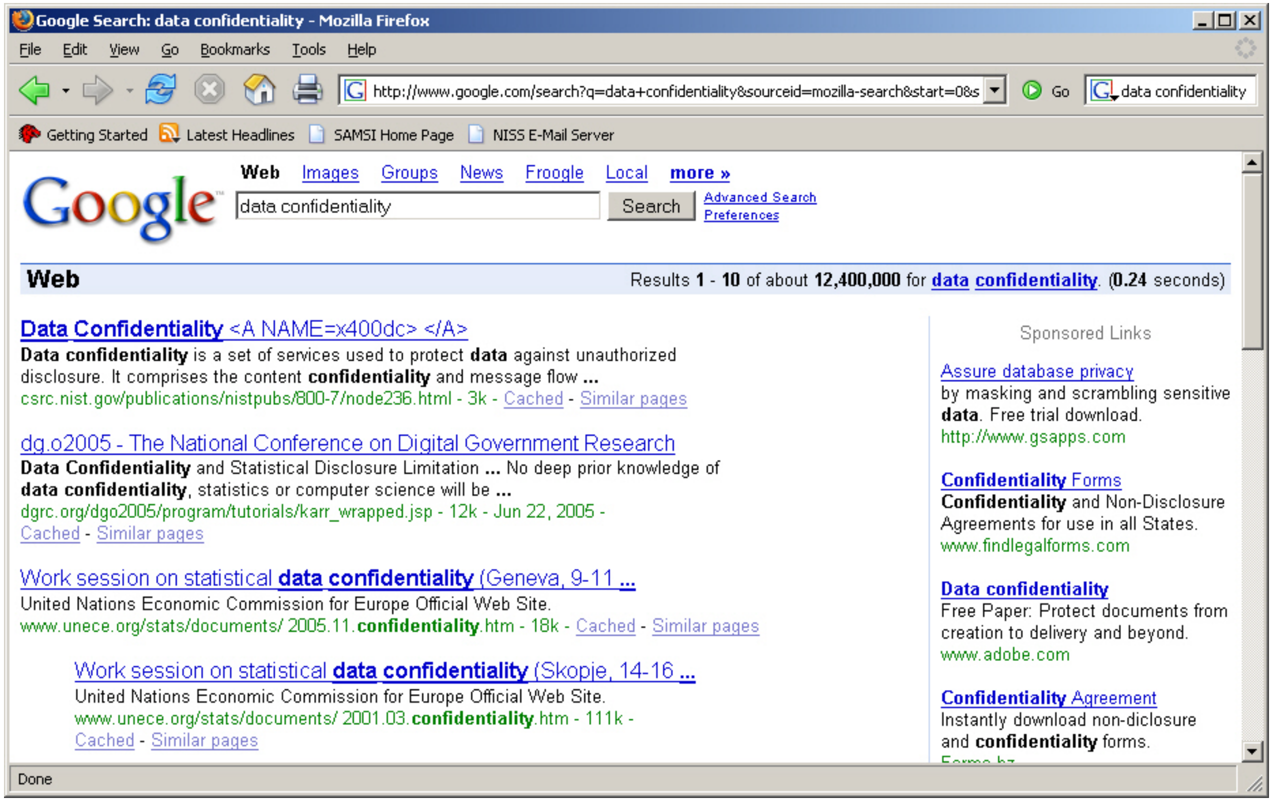
Results—2


Scatterplot
1.6000
1.2800
0.96
0.64
0.32
0
-0.29
-0.58
-0.87
-1.160
-1.450
-1.740
-2.030
-2.320
-2.610
-2.900
Company4
-1.6730
-1.275
-0.85
-0.425
0
0.5
1
1.5000
2
Global


Scatterplot
1.9000
1.5200
1.1400
0.76
0.38
0
-0.36
-0.72
-1.080
-1.440
-1.800
Except3
-1.700
-1.275
-0.85
-0.425
0
0.5
1
1.5000
2
Global

More Information

- NISS DG project web site: www.niss.org/dgii
- Google



Results **1** - **10** of about **12,400,000** for **data confidentiality**. (**0.24** seconds)

Data Confidentiality <A NAME=x400dc> </A>
Data confidentiality is a set of services used to protect **data** against unauthorized disclosure. It comprises the content **confidentiality** and message flow **...**
csrc.nist.gov/publications/nistpubs/800-7/node236.html - 3k - Cached - Similar pages

dg.o2005 - The National Conference on Digital Government Research
Data Confidentiality and Statistical Disclosure Limitation **...** No deep prior knowledge of **data confidentiality**, statistics or computer science will be **...**
dgrc.org/dgo2005/program/tutorials/karr_wrapped.jsp - 12k - Jun 22, 2005 - Cached - Similar pages

Work session on statistical **data confidentiality** (Geneva, 9-11 **...**
United Nations Economic Commission for Europe Official Web Site.
www.unece.org/stats/documents/ 2005.11.**confidentiality**.htm - 18k - Cached - Similar pages

Work session on statistical **data confidentiality** (Skopje, 14-16 **...**
United Nations Economic Commission for Europe Official Web Site.
www.unece.org/stats/documents/ 2001.03.**confidentiality**.htm - 111k - Cached - Similar pages

Sponsored Links

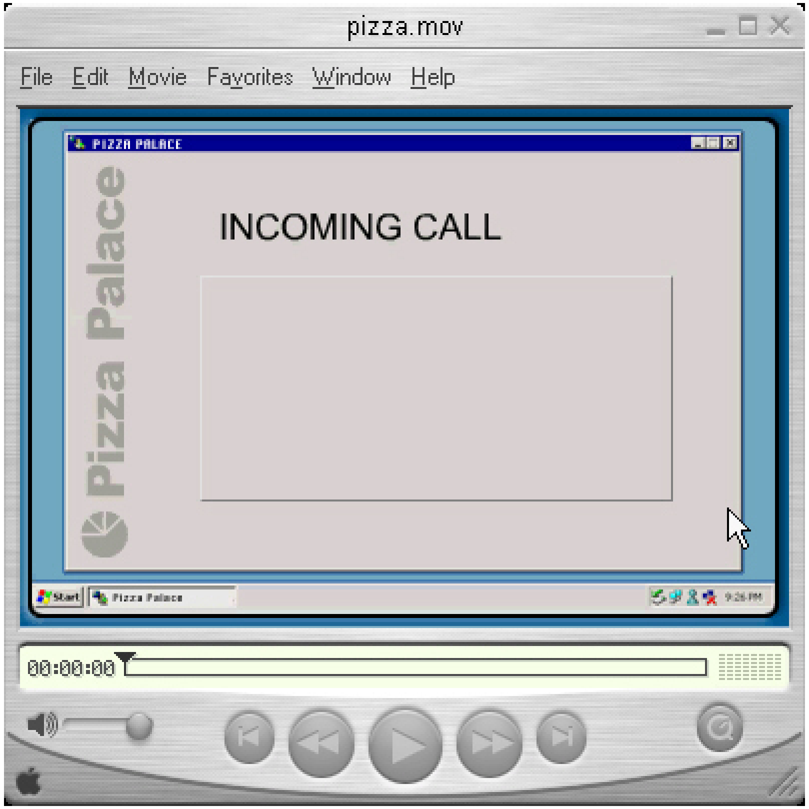
Assure database privacy
by masking and scrambling sensitive **data**. Free trial download.
http://www.gsapps.com

Confidentiality Forms
Confidentiality and Non-Disclosure Agreements for use in all States.
www.findlegalforms.com

Data confidentiality
Free Paper: Protect documents from creation to delivery and beyond.
www.adobe.com

Confidentiality Agreement
Instantly download non-diclosure and **confidentiality** forms.

What’s the Future?



INCOMING CALL

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