

# NISS

## Connections of Data Mining to Other Problems

Alan F. Karr

National Institute of Statistical Sciences

[karr@niss.org](mailto:karr@niss.org)

# Outline

- Commercials
  - NISS
  - SAMSI
- The other problems: DC, DQ, DI
- Illustrative examples of what we can and CANNOT do

# NISS



19 T. W. Alexander Drive, Research Triangle Park, NC 27709

# What NISS Does

- Mission: identify, catalyze and foster high-impact, cross-disciplinary research involving the statistical sciences
- Modus Operandi: build new bridges between statistical and disciplinary sciences, using multiple mechanisms
  - Government- or industry-funded projects driven by disciplinary issues, usually involving multiple institutions
  - Affiliates Program: corporations, government agencies, university departments

# Who NISS Is

- Director (Karr), three Assistant Directors (Gerig, McDonald, Young)
- Research staff (Sanil)
- Support staff
- Senior project participants
  - Duke, UNC, NCSU, CMU, Purdue, SMU, UMD, GA Tech, Vanderbilt, UWA
  - GM, SecondSight, BLS, BTS, Census, NASS, NCES
- Postdoctoral Fellows (9 next year)
  - 2-3 year project-tied appointments
- Graduate students
  - With advisors
  - As interns
- Undergraduate interns

# Current Activities -1

- Computer model evaluation
  - Characterize uncertainty in model predictions where field data are absent
- Software engineering
  - Lightweight instrumentation of fielded software for performance evaluation, testing, user profiling
- Web data
  - Identify inconsistencies between user behavior and Web site structure

# Current Activities - 2

- Data quality
  - Design of data quality evaluation process and instruments for BTS
  - Embedded in DG
- Digital Government (DG): T&M to software
  - Data confidentiality
    - Tabular data, data swapping, aggregation, regression servers
  - Data quality
  - Data integration

samsi

NSF ● Duke ● NCSU ● UNC ● NISS

# SAMSI Basics

- **Vision:** forge a new synthesis of the statistical sciences and the applied mathematical sciences with disciplinary science to confront data- and model-driven scientific challenges
- **Structure**
  - Partnership of Duke, NCSU, UNC, and NISS
  - In collaboration with Mathematical Sciences Research Institutes program of the NSF
  - In cooperation with the William R. Kenan, Jr. Institute for Engineering, Technology and Science (Raleigh, NC)
- **Management:** James Berger (Duke), H. T. Banks (NCSU), J. S. Marron (UNC), and Alan F. Karr (NISS)
- **Location:** NISS building

# SAMSI Modus Operandi

- 2-3 research programs per year
  - At interfaces: Involve both statistics and applied math (also probability, computational sciences, operations research)
  - Framed and guided by disciplinary needs
  - “Catalytic rather than conclusional”
  - Multiple types: Focused Study Programs, Synthesis Programs, Pilot Programs

# SAMSI Programs for 2003-04

- **Internet Traffic** (Fall, 2003)
  - Measurement, modeling, heavy traffic workshop: 9/17-20/03
  - Internet tomography workshop: 10/12-14/03
- **Data Mining and Machine Learning** (Entire year)
  - Tutorials and opening workshop: 9/6-10/03
  - Closing workshop: 5/20-22/04
  - *Possible* foci include:
    - Sampling, model selection and search, robustness and data quality
    - Inference for high-dimensional, small sample ("large  $p$ , small  $n$ ") data
    - Scalability and other computational issues
    - Text mining
    - Novel, complex forms of data, such as images and space-time data
- **Multiscale Modeling and Control Design** (Spring, 2004)
  - Opening workshop: 1/17-21/04
  - Soft-matter and nano-materials workshop: 2/15-17/04
  - Granular flow workshop: 4/15-17/04

# Opportunities for Participation

- Program proposals and leadership
- Research visits (short- and long-term)
- Workshops
- Postdoctoral fellowships (2+ years, in collaboration with NISS, universities, ...)
- SAMSI-University Fellowships
- Research programs and summer interdisciplinary workshops for graduate students

# More Information

NISS

[www.niss.org](http://www.niss.org)

SAMSI

[www.samsi.info](http://www.samsi.info)

SAMSI DM&ML

[www.samsi.info/200304/dmml/dmml-home.html](http://www.samsi.info/200304/dmml/dmml-home.html)

DC, DI and DQ

# DC: Data Confidentiality

- Fundamental Issue: Tradeoffs between
  - Disclosure risk: data subject IDs, attribute values
  - Data utility: to researchers, public, ...
- Principal approaches to statistical disclosure limitation (SDL)
  - Restricted access: special sites, licensing, ...
  - Restricted data
    - Tables: release only selected marginals
    - PRAM: data swapping, synthetic data, ...
    - Aggregation: top-coding, geographical, ...
    - Disseminate “safe analyses” rather than data

# DI: Data Integration

- Fundamental Issue: Management of and inference from data created by combining multiple, “related” databases, often assembled by different organizations for different purposes

# DQ: Data Quality

- Fundamental Issue: Characterize and improve capability of data to be used effectively, economically and rapidly to inform and evaluate decisions
  - Multi-dimensional: accuracy, accessibility, relevance, timeliness, metadata, documentation, user capabilities, user expectations, cost, domain knowledge
  - Multi-disciplinary: statistics, computer science, total quality management (TQM)

# The Thesis of This Talk

- DC, DI, DQ are known to interact
  - Some aspects identified but none well understood
  - Examples
    - Poor DQ protects DC
    - Tools for DI also can be used to break DC
- All of {DC, DI, DQ} interact with DM
  - Not only understanding but also even identification of central issues lacking

# Testbed Databases

- **CPS** (Current Population Survey): 48,842-element, 8-attribute excerpt from 1993 CPS
- **FARS** (Fatality Analysis and Reporting System): 1999 Data
  - FARS-A: Accident Table
  - FARS-D: Driver Table
  - FARS-P: Person Table
  - FARS-V: Vehicle Table

# DM Purposes

- Pattern discovery
- Identification of anomalous data (“looking for needles in haystacks”)
- Scalability

# Association Rules

- $D$  = database
- $A, B$  subsets of  $D$
- Association rule:  $A \Rightarrow B$ 
  - Confidence:  $C = |A \cap B| / |A| = P(B | A)$
  - Support:  $S = |A \cap B| / |D| = P(A \cap B)$

# Example: CPS

Age

Salary

	<25	25-55	>55
<\$50K	8339	23912	4904
>=\$50K	93	9629	1965

Support

	<25	25-55	>55
<\$50K	.170	.490	.100
>=\$50K	.002	.197	.040

Confidence given Age

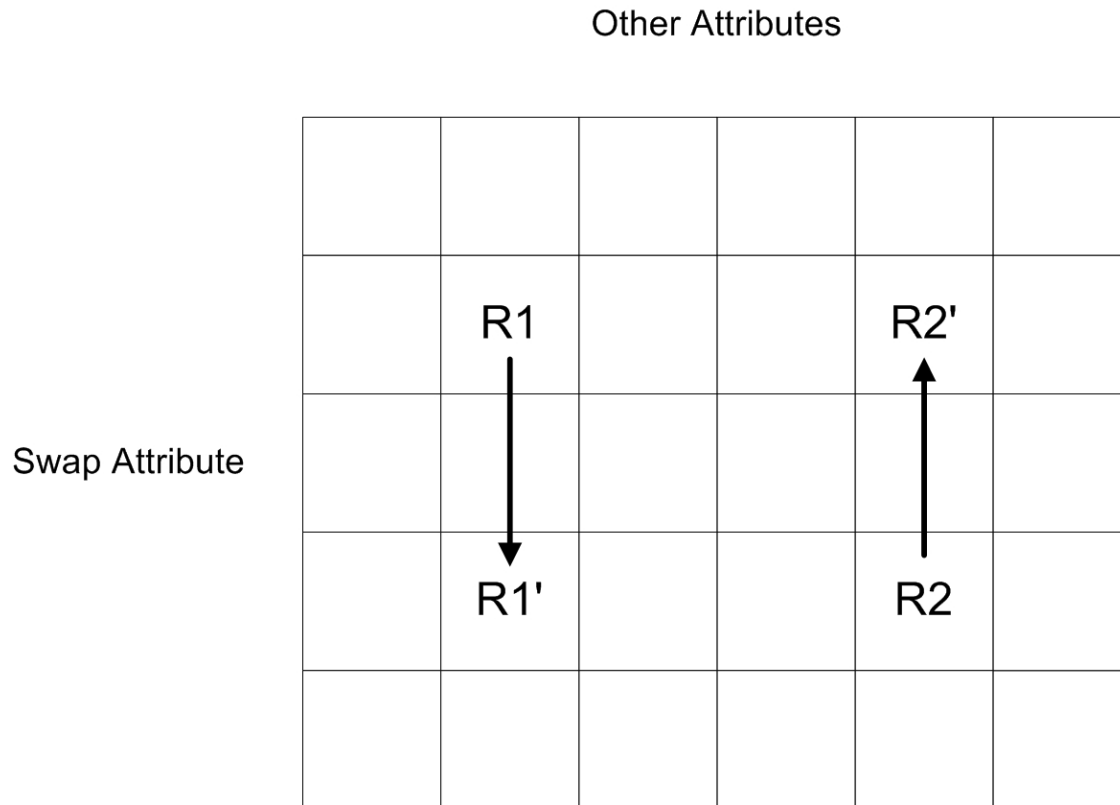
	<25	25-55	>55
<\$50K	0.99	.71	.71
>=\$50K	.01	.29	.29

DC: Data Confidentiality

# DM $\cap$ DC: What We Do Know

- Very complex problem
  - Highly charged politically
  - Multiple stakeholders: individuals, data holders, government agencies, ...
- SDL creates anomalies
- Initial Step: Privacy-preserving data mining
  - Local computation (Clifton, et al.)
  - “PRAM”-based approaches (Agrawal & Srikant, Privacy Preserving Data Mining, Proc. ACM SIGMOD Conference on Management of Data, 2000)

# SDL-Created Anomalies



# Privacy-Preserving Association Rules

- Problem
  - Multiple but “identical” databases
  - Find item pairs  $(A_i, A_j)$  with *global* support  $\geq s\%$
  - Only local computations, performed by database owners
  - Protect
    - Data items
    - Support at each site
    - Database sizes

# Local Computation: The Algorithm

- Compute:  $1(\sum_k C_k(i, j) \geq s \sum_k N_k)$
- Procedure
  - Site 1
    - Generate (large) random number  $R$
    - Calculate and transmit to site 2:  $X_1 = R + C_1(i, j) - sN_1$
  - Site  $m$ 
    - Calculate and transmit to site  $m+1$ :  $X_m = X_{m-1} + C_m(i, j) - sN_m$
  - Site 1
    - Check whether  $X_k \geq R$
    - Distribute result

# DM $\cap$ DC: What We Don't Know

- Abstractions for disclosure risk associated with DM
- Whether DM is inherently more threatening to confidentiality than other statistical analyses
- Effect of most SDL strategies on DM
- Whether DM can be used to defeat SDL
  - Example: Can DM detect swapped records?

# Example: CPS After Swapping

Age

	<25	25-55	>55
Salary <\$50K	7848	24292	5015
Salary >=\$50K	584	9249	1854

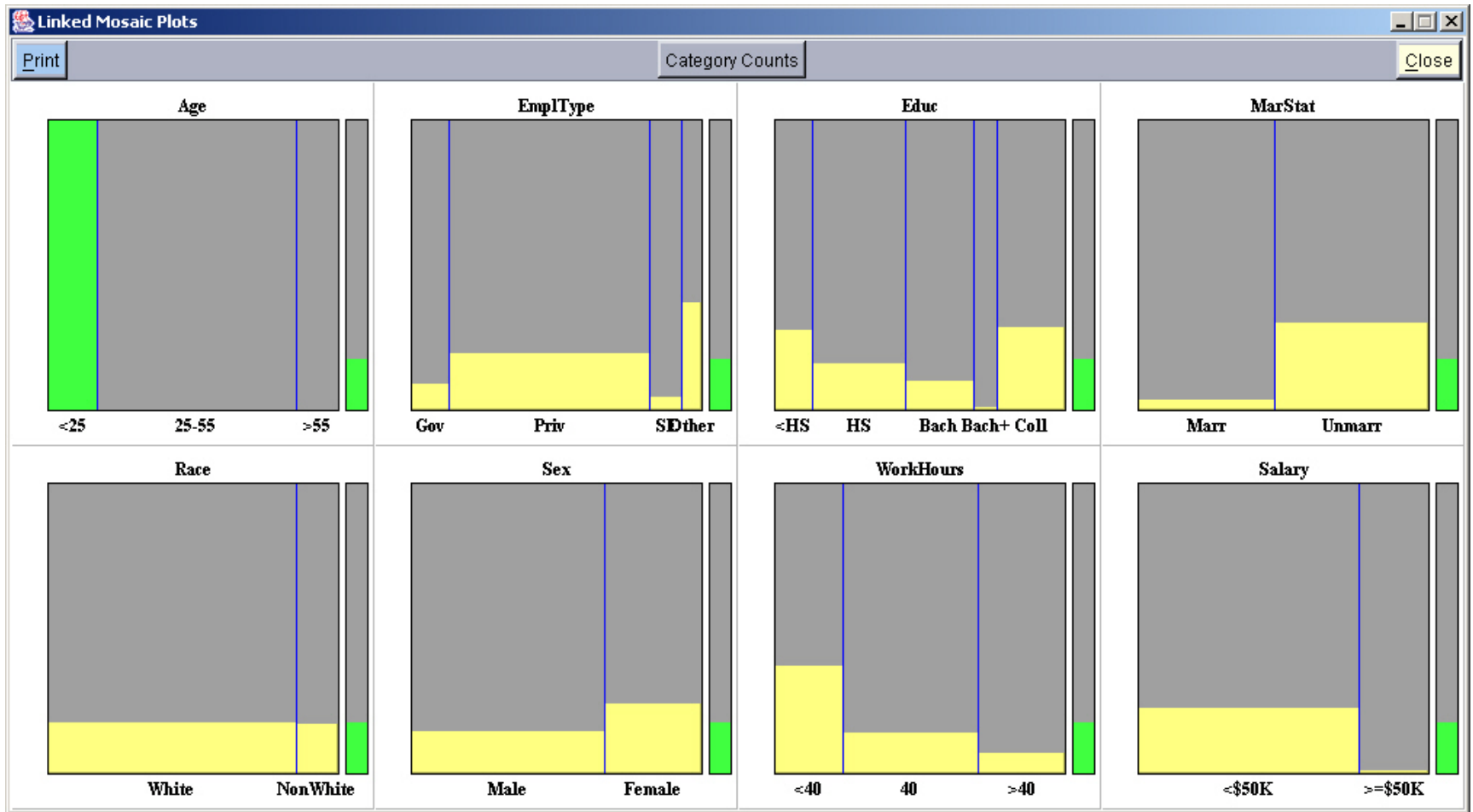
Support

	<25	25-55	>55
<\$50K	0.161	.497	.103
>=\$50K	.012	.189	.038

Confidence given Age

	<25	25-55	>55
<\$50K	.93	.72	.73
>=\$50K	.07	.28	.27

# Alternative View of CPS Data



# DI: Data Integration

# DM $\cap$ DI: What We Do Know

- Methods for DI are limited
  - Database joins
  - Record linkage
- Difficulties with DI inhibit DM
  - Example: Ford Explorer/Firestone tire problem
    - FARS
    - Warranty/service data
    - Manufacturing data
    - Common thread = VIN (Example: 1FTDF15Y0KNB)

# DM $\cap$ DI: What We Don't Know

- How to track the effects of particular DI methods
- How to do DM when DI is impossible
  - Example: patterns in fatal accidents involving transit vehicles
    - FARS-A: CITY, COUNTY, TRID
    - FARS-V: MAKE, MODEL, BDTYP, MODYR
    - NTD: TRS\_ID, cPER\_VEH, dPER\_VEH (patron and employee in-vehicle fatalities, per year)
    - Neither joins nor record linkage is possible!
    - What to do???
  - Example: what if the relevant join is “too big?”

DQ: Data Quality

# DM $\cap$ DQ: What We Do Know

- DQ problems are massive
- Many data sets are not (completely) real
- In real data sets, many points are anomalous
- Anomalous data are more likely to be wrong
- How to use EDA to understand DQ problems

# The Scale of DQ Problems

- FARS-A, 1999
  - 18,433 records, 49 attributes
- Intactness: Ignoring lat/long, which are present in only 71 records, only 30% of records
  - Have no missing values
  - Pass 3 simple consistency checks
- 4 attributes have the same value for all records
- Variety of other problems

# FARS-A 1999 Excerpt

	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

# “Unreal” Data

- SDL applied
  - Data swapping
  - Synthetic data
- Statistical adjustments
  - Imputation of missing responses
- Data clean-up
  - DQ edits. Examples:
    - Address corrections
    - (male,hysterectomy) → (female,hysterectomy)

# Prevalence of Anomalous Points

- CPS (not realistically sparse)
  - 361 data elements in cells with count 1
  - 374 data elements in cells with count 2
- In higher dimensions, nearly all cell counts are 0 or 1, and most are 0
  - Example: 14-dimensional CPS data
    - 435,000,000 cells
    - 299,285 data elements
    - 72,739 data elements in cells with count 1 or 2

# Anomalous Data Are Often Wrong

- Example
  - FARS LTIME = [FARS-A:(ACCDATE,ACCTIME)  
– FARS-P:(DEATHDATE,DEATHTIME)]
  - (CSTATE=54, CNUM=321)
    - FARS-A: ACCDATE = *11*/7/99, ACCTIME = 19:27, NFAT = 2
    - FARS-P: For both fatalities, DEATHDATE = *12*/7/99,  
DEATHTIME = 19:27, LTIME = 72000 (largest value in file)
- Implication: Large numbers of false positives in low quality data: lots of things look like needles, but aren't needles!

# DQ $\cap$ DM: What We Don't Know

- How to quantify DQ
- How to model the causes or effects of poor DQ
- How to build robust DM tools
- How to deal with costs

# The Grand Challenge

$DM \cap DC \cap DI \cap DQ$