

NISS

Interactions Among Data Confidentiality, Data Integration, Data Mining, Data Quality: Challenges for Statisticians

Alan F. Karr

National Institute of Statistical Sciences

karr@niss.org

The Take Away Message

- Four important problems
- Each understood to some degree, but by no means completely
- Interactions critical in multiple settings, but hardly understood at all
- Therefore, great set of challenges, with major policy implications, for statistical scientists!

The Problem Areas

- DC = data confidentiality
 - Protect data subjects and attribute values, yet disseminate useful information
- DI = data integration
 - Combine data across multiple databases not designed with DI in mind
- DM = data mining
 - Discover patterns, information and knowledge in large, complex, unstructured databases
- DQ = data quality
 - Cope with errors and anomalies in real databases

Testbed Databases

- **CPS-8**: excerpt from 1993 CPS
 - 48,842 data records (not realistic!)
 - 8 categorical attributes (not realistic!)
 - 2880 cells in full table (not realistic!)
 - 1695 cells with non-zero counts (not realistic!)
- **FARS** (Fatality Analysis and Reporting System): 1999
 - FARS-A: Accident Table
 - FARS-D: Driver Table
 - FARS-P: Person Table
 - FARS-V: Vehicle Table

DC: Overview

- Fundamental issue: Tradeoffs between
 - Disclosure risk: data subject IDs, attribute values
 - Data utility: to researchers, public, other agencies, ...
- Approaches to statistical disclosure limitation (SDL)
 - Restricted access: special sites, licensing, ...
 - Restricted data
 - Tables: release only selected marginals, cell suppression, ...
 - Aggregation: top-coding, geographical, ...
 - Altered data, typically by introduction of randomness
 - PRAM: data swapping, synthetic data, jittering, ...
 - Disseminate “safe analyses” rather than data
 - Regression servers

DC: The Threat



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
Born: Sep. 11, 1902



Search using Age or birthday
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

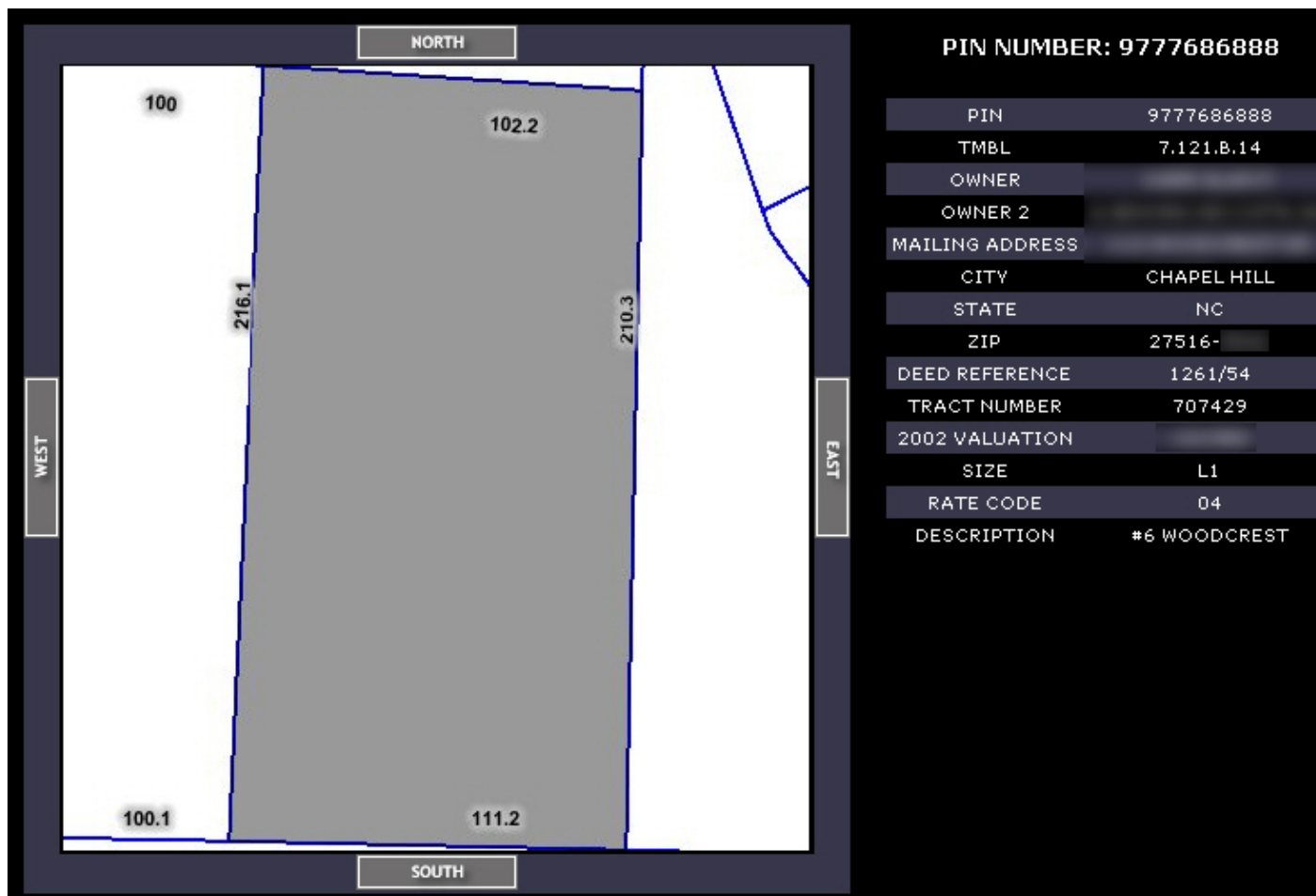
Zip Code

ALAN F KARR

27516

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

And More ...

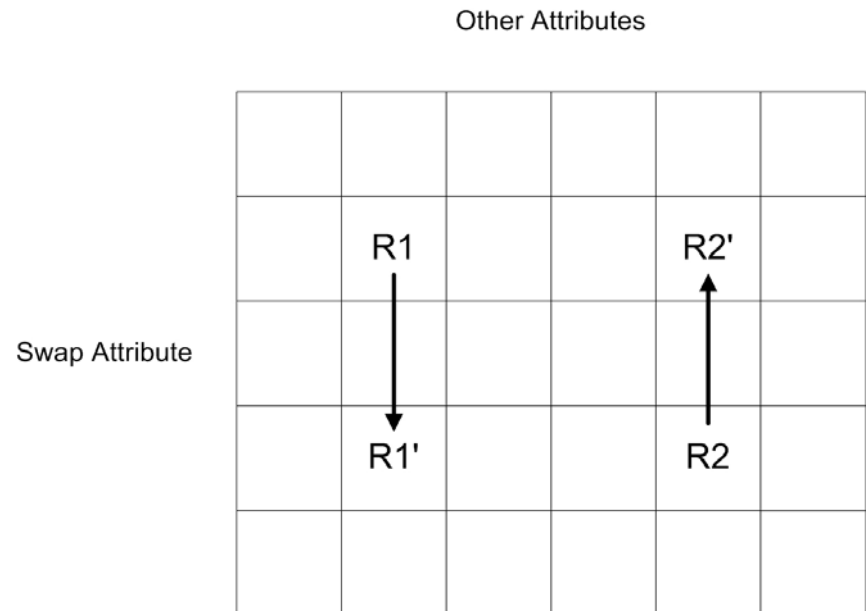


Risk-Utility Formulations

- Define release space \mathcal{R} (may be partially ordered)
- Quantify, with monotone functions on \mathcal{R} ,
 - Disclosure risk
 - [Transparency risk: Are statements of the form “Data swapping was employed for SDL” risky?]
 - Data utility
- Two basic paradigms
 - Maximize utility subject to risk constraint(s)
 - Identify risk-utility frontier

Example: Data Swapping

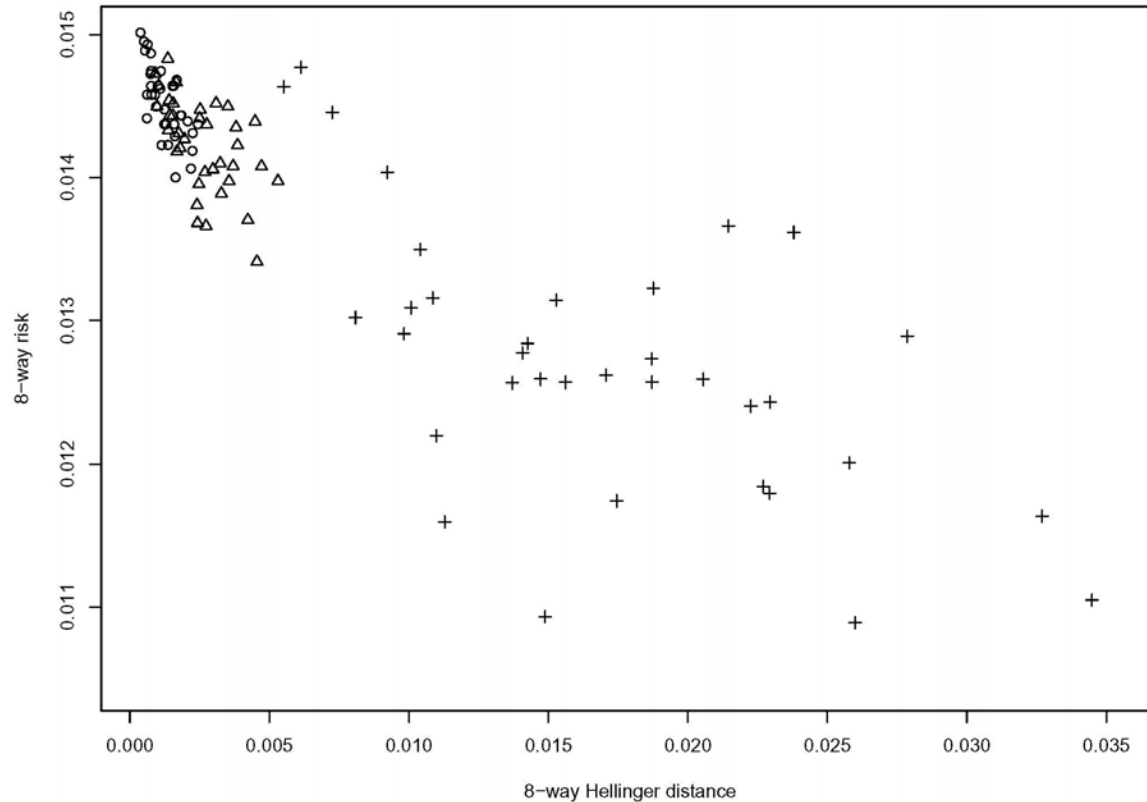
- Basic idea: switch subset of attributes between randomly selected pairs of records at microdata level
- Rationale: reduces disclosure risk
 - Intruder cannot be certain that any record is real
- Side effect: distorts data, reducing utility
 - Changes (only) joint distributions that involve both swapped and unswapped attributes



Risk-Utility Formulation

- Decision problem: select
 - Swap rate
 - Swapped attributes
 - Optionally, constraints on unswapped attributes
- Characterize each candidate release by
 - Disclosure risk. Example: number of unswapped records in small count cells in post-swap table
 - Data utility. Example: dis-utility = data distortion, as measured by Hellinger distance, or ...
- No unique solution, but restrict attention to frontier of undominated releases

Example Frontier: CPS-8



- 1% frontier = {AveHours, Educ, AveHours+Educ, AveHours+Sex, Sex, AveHours+MarStat, EmplType+MarStat}
- 2% frontier = {Educ, AveHours, Race, AveHours+Race, EmplType+Sex, EmplType+Race, AveHours+MarStat, Age+Income}
- 10% frontier = {Educ, Race, Educ+Race, AveHours+Race, EmplType+Race, Age+Race}

DC: Challenges

- Inference-based measures of data utility
 - Example: do released marginals lead to a good log-linear model of the full table?
- Transparency risk
 - How dangerous is it to make statements of the form “The released marginals lead to a good log-linear model of the full table”?
- Analysis servers
 - A “world without microdata” is possible

DI: Overview

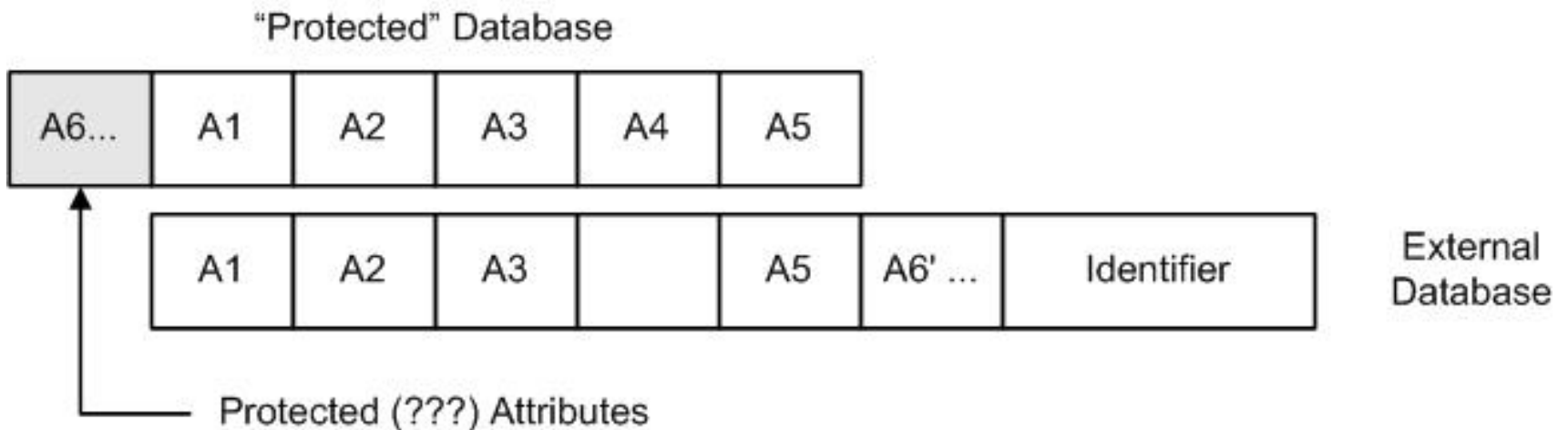
- Fundamental issue: Management of and inference from databases created by combining multiple, databases assembled by different organizations for different purposes
- Approaches from multiple disciplines
 - Database: joins
 - Data warehousing: parsing and standardization
 - Statistics: probabilistic record linkage
 - AI/NL: machine translation

DI: Challenges

- Paucity of methods
 - Automatic methods
 - Scale is a significant issue
 - Metadata quality is a major impediment
 - *Any* methods
 - Kean report: "Although relevant information . . . regarding the attacks was available to the Intelligence Community prior to September 11, 2001, the Community too often failed to [...] *consider and appreciate its collective significance.*"
- Effects of DI on inference
 - Example: quantification of uncertainty introduced by incorrect record linkage

DC \cap DI: We Know That

- Tools for DI also can be used to break DC
 - Example: Record linkage
 - Medical database from Cambridge, MA contained gender, date of birth, 5-digit zip code
 - Linkage (by hand) to public voter list identified 90+% of records
- Databases to integrate with abound



DC \cap DI: Challenges

- Right abstractions for the problem
 - “Correct” universe of DI candidates to break DC
 - Risk: (Probability of) re-identification too simplistic
 - Alternatives: Cost, effort, ???
 - Must deal with incorrect linkages
- How to differentiate among methods for DI
- Privacy preserving alternatives to DI
 - Algorithmic issues
 - Conceptual issues: PPAs require advance agreement that result is not too risky

Privacy Preserving Integration

- Concept: pooled analyses without
 - Shared data
 - Trusted third parties
- Example: local computation
 - Analysis possible from $(\sum p_{i1}, \dots, \sum p_{in})$, where (p_{i1}, \dots, p_{in}) are characteristics of DB_i
 - DB_1 : compute $(p_{11} + R_1, \dots, p_{1n} + R_n)$, send to DB_2
 - DB_2 : add (p_{21}, \dots, p_{2n}) , producing $(p_{11} + p_{21} + R_1, \dots, p_{1n} + p_{2n} + R_n)$
 - DB_1 : receive $(\sum p_{i1} + R_1, \dots, \sum p_{in} + R_n)$, remove R_1, \dots, R_n , circulate $(\sum p_{i1}, \dots, \sum p_{in})$

DM: Overview

- Fundamental issue: discovery of information and knowledge in large, complex, unstructured databases
- Purposes
 - Pattern discovery
 - Example: transactions
 - Identification of anomalous data
 - Example: “looking for needles in haystacks”
- Often, use simple analyses, in order to overcome scalability problems with complex statistical procedures

Example: Association Rules

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

Example: CPS-8

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

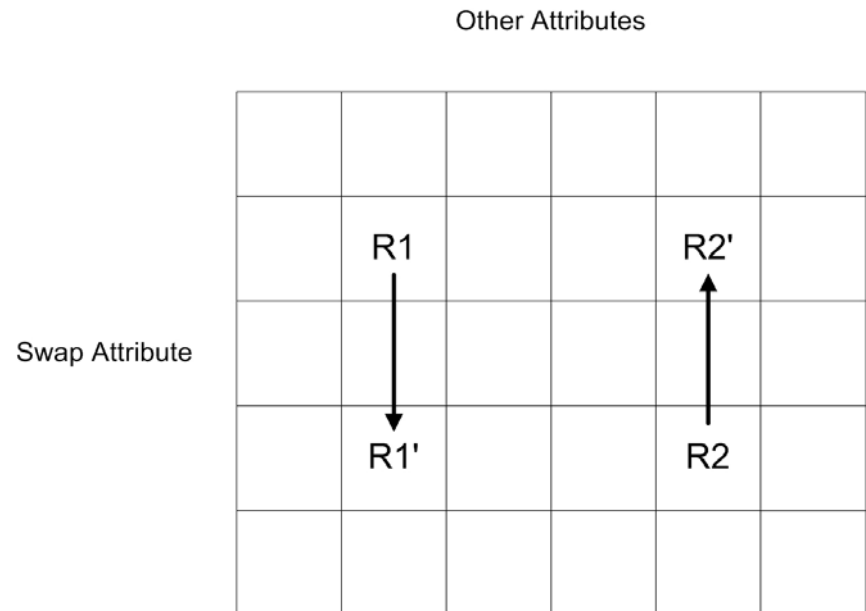
	<25	25-55	>55
<\$50K	0.99	.71	.71
>=\$50K	.01	.29	.29

DM: Challenges

- Statistical properties of DM tools are not understood
- DM tools are
 - Too blunt for *very rare* patterns (anomalies)
 - Susceptible to false positives: lots of things look like needles, but aren't needles
- In high dimensions, anomalies are inherent
 - CPS-8 (2880 cells, 48,842 records): 735 (1.5%) in cells with count 1 or 2
 - 14-dimensional CPS data (435,000,000 cells; 299,285 records): 72,739 (24.3%) in cells with count 1 or 2
- Concern about DC and poor DQ both create additional anomalies

DC \cap DM: We Know That

- SDL creates anomalies
- Privacy-preserving data mining is claimed to be possible to some extent
 - Additive noise
 - Swapping
 - Local computation



Privacy Preserving Association Rules

- Problem
 - Multiple but identical (!?) databases
 - Find item pairs (A_i, A_j) with *global* (across all the databases) support $\geq s\%$
 - Protect: Data items, value of support at each site, database sizes
- Solution: privacy preserving local computation of

$$1(\sum_k C_k(i, j) \geq s \sum_k N_k)$$

DC \cap DM: Challenges

- Abstractions for disclosure risk from DM
 - Inferential disclosure +: relationships in the data
- Whether, compared to other analyses, DM is
 - Inherently more threatening to DC
 - A different kind of threat
- Effect of most SDL strategies on DM
- Whether DM can be used to defeat SDL
 - Example: Can DM detect swapped records?

Example: CPS-8 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

DI \cap DM: What Know That

- Many disconnects exist between DI and DM
 - Database joins: scalability problem
 - Record linkage: presumes clean correspondence between attributes [, good model for $P\{\text{Match}|S_1, S_2\}$]
- Difficulties with DI inhibit DM
 - Example: Ford Explorer/Firestone tire problem
 - FARS-V (NHTSA)
 - Warranty/service data (Manufacturers, dealers, JD Power,...)
 - Manufacturing data (Manufacturers)
 - Common thread = VIN (Example: 1FTDF15Y0KNB)

DI \cap DM: Challenges

- How to track the effects of particular DI methods
- How to do DM (except possibly by hand) when DI is “impossible”
 - Example: 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-employee in-vehicle fatalities, per year)
 - Neither joins nor record linkage is possible!

DQ: Overview

- 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)

Scale and Ubiquity of DQ Problems

- FARS-A, 1999 (one version)
 - 18,433 records, 49 attributes
- Intactness: Ignoring (lat,long), which is present in only 71 records, only 30% of records
 - Have no missing values
 - Pass 3 simple consistency checks (Example: temporal precedence)
- 4 attributes have the same value for all records
 - Produces high confidence, high support, but meaningless association rules
 - True dimension of the data is at most 45 (and is actually less)

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

Other FARS DQ Problems

- Numerical codes for categorical variables
 - Number of lanes: 7 means “7 or more”
- Multiple representations for missing attributes, some of which are valid data values
 - Age = 99
 - Also, [blank], [space], *, 0
- Partially missing attributes
 - ACCDATE = 10991999
- “Unjoinable” tables that should be in 1-1 correspondence
 - FARS-D and FARS-V: in some versions, join is empty

DQ: Challenges

- Quantification: few meaningful DQ metrics
- Models of the causes or effects of poor DQ
- Costs
 - How much does good DQ cost?
 - How much does bad DQ cost?

DC \cap DQ: We May Know That

- Poor DQ protects DC
 - “Poor data quality is the best form of SDL”
- Perceived lack of DC decreases DQ
 - “Respondents lie when they don’t think their information will be protected”
 - Example: SAMHSA data on teen drug use

Household Information, recognizable (only) by parents	Teenage respondent’s age, use of drugs
--	---

DC \cap DQ: Challenges

- Whether poor DQ really does protect DC
 - Can “false” records be identified (e.g., by DM)?
- Whether or how perceived protection of confidentiality affects DQ
 - Evidence is largely but not entirely anecdotal
- Right level to think about DC \cap DQ
 - Individual?
 - Population?

DI \cap DQ: We Know That

- Inability to do DI is a DQ problem
- Metadata quality may be the real issue
 - Absence of foreign keys
 - Inability to match “the same attributes”
 - Example: gasoline and petrol
 - Example: different units
 - Different definitions of “the same” attribute
 - Different attribute categories
- Changes over time compound the problem

DI \cap DQ: Challenges

- Whether DI improves DQ
 - More precisely, what effects do different methods for DI have DQ?
- How to do DI in ways that don't impair DQ
- MDQ

DM \cap DQ: We Know That

- DQ problems create anomalies
 - Example: FARS-P LTIME (between accident and death)
 - Sort FARS-P by LTIME: Largest value = 72000 (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

DM \cap DQ: Challenges

- Wrong data are often anomalous, but are anomalous data more likely to be wrong?
- Is DM an effective approach to detecting or ameliorating DQ problems?

Overarching Challenges

- $D\{C, I, M, Q\}$ for
 - Data that are not categorical or numerical
 - Examples: Geospatial data, images, video, audio
 - Data streams: large quantities of ephemeral, machine-generated data
- How to deal with costs
- Benchmarks to evaluate different strategies
- Effective, automatic use of domain knowledge

New Data



How Do These Rate on D{C,I,M,Q}?

