Optimizing the Use of Microdata:

Julia Lane

Adapted from ASA presentation in honor of Pat Doyle

Overview

- Benefits and Costs of Microdata Access
- Example of Consequences of Current Practice
- Current and Future Challenges
- Developing an Economic Framework
- Using the Framework to Shape a Research Agenda
- Next Steps

Benefits Of Microdata Access

- Permits Analysis of Complex Questions
  - Tabular data answers predefined questions
  - Micro data "drills down" to basic decision-making unit
  - Heterogeneous behavior of economic agents
- Ability to Estimate Marginal Effects
- Scientific Safeguard
- Data Quality
- Development of Core Constituency for Statistical Agencies

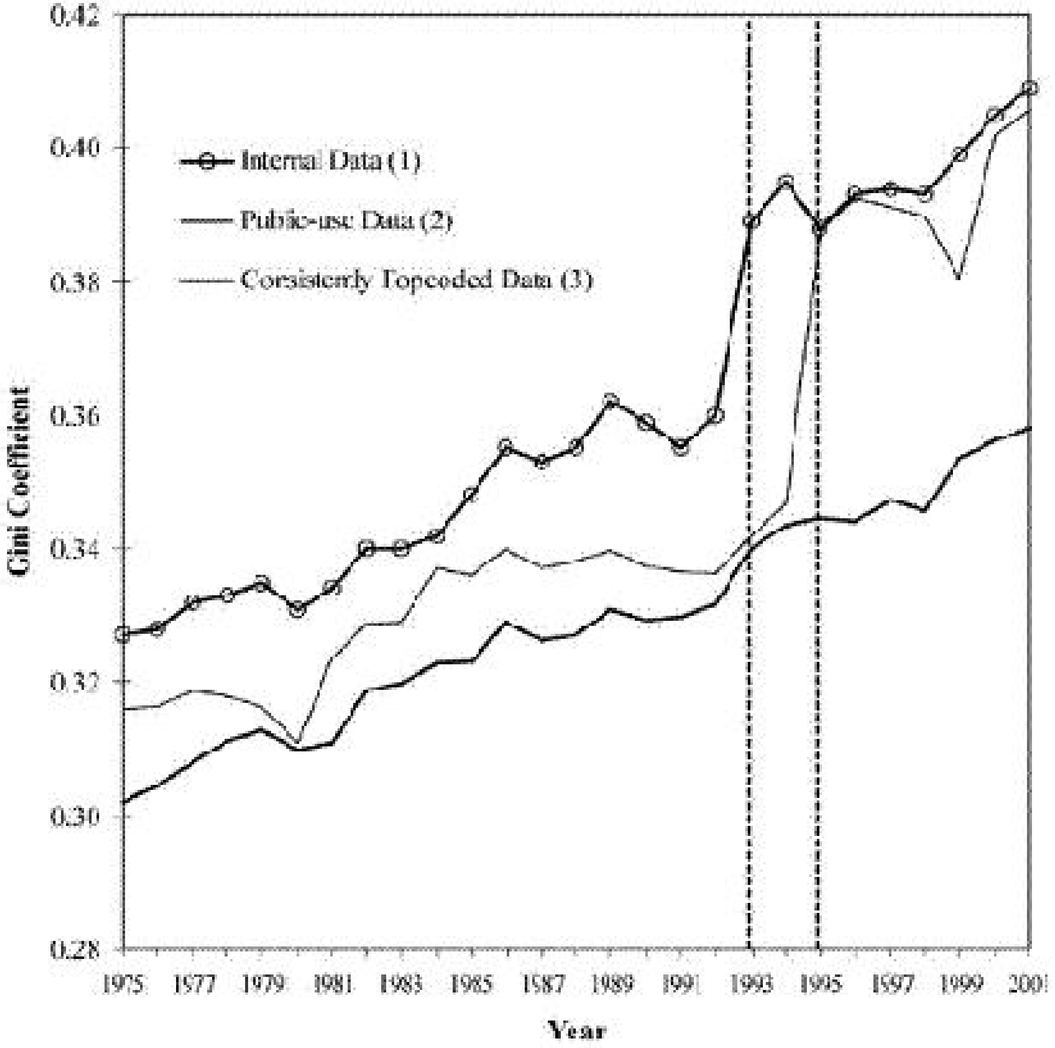
Costs Of Microdata Access

- Different modalities
  - Research Data Centers
    - cost of safeguards
  - Licensing
    - cost of monitoring
  - Remote Access
    - cost of developing and updating
  - Public Use Files
    - cost of developing and updating
- Reputation Costs
  - “Official” statistics?
  - Role of work in progress
  - Authorized purpose?
- Disclosure
  - Legal liability
  - Ethical
  - Response rates

Example of Impact of One Approach: Public Use Files

- Reduce Information
  - variable deletion
  - recoding categorical variables into larger categories
  - recoding continuous variables into categories
  - rounding continuous variables
  - using top and bottom code
  - using local suppression and enlarging geographic areas
- Perturb Data
  - noise addition
  - record swapping
  - rank swapping
  - blanking and imputation
  - micro-aggregation
  - multiple imputation/modeling to generate synthetic data

Consequences of Topcoding for Data Quality



g. 1. Trends in Gini coefficients of earnings for full-time year-round workers (1975–2001). (1) Jones and Weinberg (2000), ble 1, page 3 and US Census Bureau (2002). (2 and 3) Authors' calculations using CPS public-use data, 1976–2000.

Consequences of Topcoding for Decisionmaking

- Earnings inequality increasing
  - Steadily?
  - Sharply?
  - When?
- Inference for policy makers?

Consequences of Topcoding for Data Quality

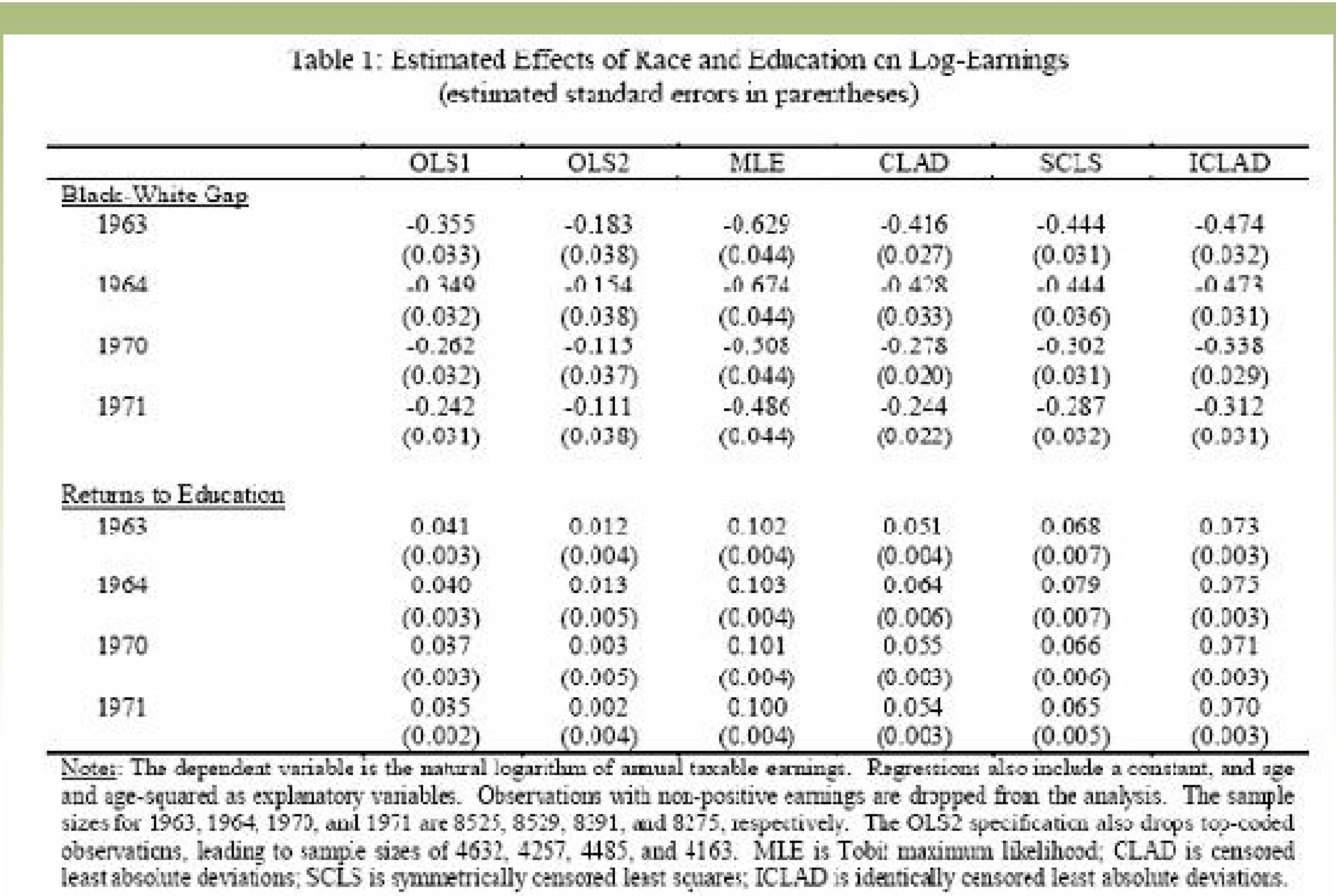


Table 1: Estimated Effects of Race and Education on Log-Earnings
(estimated standard errors in parentheses)

	OLS1	OLS2	MLE	CLAD	SCLS	ICLAD
Black-White Gap						
1963	-0.355 (0.033)	-0.183 (0.038)	-0.629 (0.044)	-0.416 (0.027)	-0.444 (0.031)	-0.474 (0.032)
1964	-0.349 (0.032)	-0.154 (0.038)	-0.674 (0.044)	-0.428 (0.033)	-0.444 (0.036)	-0.473 (0.031)
1970	-0.262 (0.032)	-0.115 (0.037)	-0.508 (0.044)	-0.278 (0.020)	-0.302 (0.031)	-0.338 (0.029)
1971	-0.242 (0.031)	-0.111 (0.038)	-0.486 (0.044)	-0.244 (0.022)	-0.287 (0.032)	-0.312 (0.031)
Returns to Education						
1963	0.041 (0.003)	0.012 (0.004)	0.102 (0.004)	0.051 (0.004)	0.068 (0.007)	0.073 (0.003)
1964	0.040 (0.003)	0.013 (0.005)	0.103 (0.004)	0.064 (0.006)	0.079 (0.007)	0.075 (0.003)
1970	0.037 (0.003)	0.003 (0.005)	0.101 (0.004)	0.055 (0.003)	0.066 (0.006)	0.071 (0.003)
1971	0.035 (0.002)	0.002 (0.004)	0.100 (0.004)	0.054 (0.003)	0.065 (0.005)	0.070 (0.003)

Notes: The dependent variable is the natural logarithm of annual taxable earnings. Regressions also include a constant, and age and age-squared as explanatory variables. Observations with non-positive earnings are dropped from the analysis. The sample sizes for 1963, 1964, 1970, and 1971 are 8525, 8529, 8391, and 8275, respectively. The OLS2 specification also drops top-coded observations, leading to sample sizes of 4632, 4257, 4485, and 4163. MLE is Tobit maximum likelihood; CLAD is censored least absolute deviations; SCLS is symmetrically censored least squares; ICLAD is identically censored least absolute deviations.

Consequences of Topcoding for Decisionmaking

- Standard Censored Regression Problem
- Black/white earnings
  - Gap of .35 or .63 log points in 1963?
  - Change in gap between 1963 and 1971 .06 log points or .15 log points?

  ⇒ Policy maker?

    ⇒Racial earnings gap closing rapidly

    ⇒Racial earnings gap closing slowly?
- Return to Education
  - First column: Dropped from 1% in 1963 to approximately zero in 1973?
  - Final column Consistent at 7%.

  ⇒ Policy maker?

    ⇒Stop investing in education?

    ⇒Investment in education should increase?

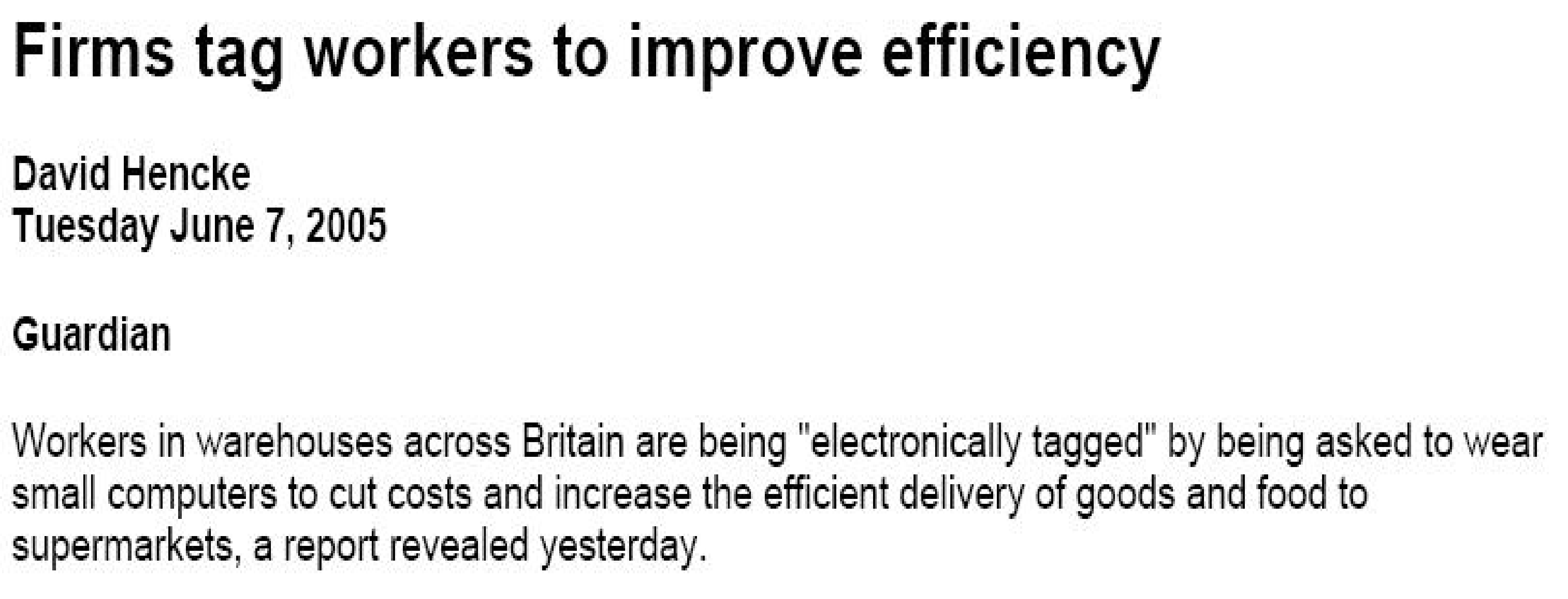
New Challenges: The Basic Issue

"A recent book and conference on confidentiality and data access brought home the growing challenge facing the Census Bureau …. It is becoming clear that advances in technology and increased use of administrative records may, at some point in the future, render our current disclosure avoidance procedures inadequate. At the same time … the larger federal statistical system face increasing demands for more, better and more recent data to meet critically important public policy and research needs."

Pat Doyle, 2001

New Challenges:
New Data Collection Modalities

- Surveys/censuses/admin data and..
- Textual corpora
- Videotapes
- wireless network embedded devices
    - increasingly sophisticated phones
    - RFIDs
    - sensor webs
    - smart dust
- Cognitive neuroimaging records

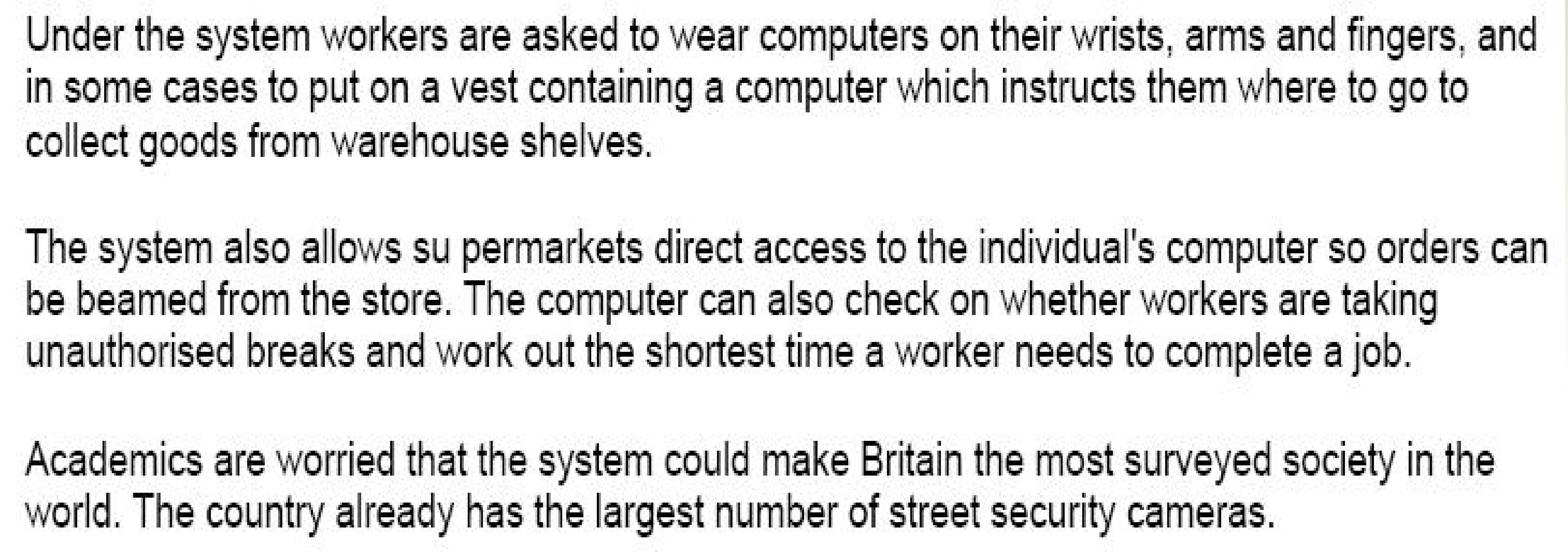


Firms tag workers to improve efficiency

David Hencke
Tuesday June 7, 2005

Guardian

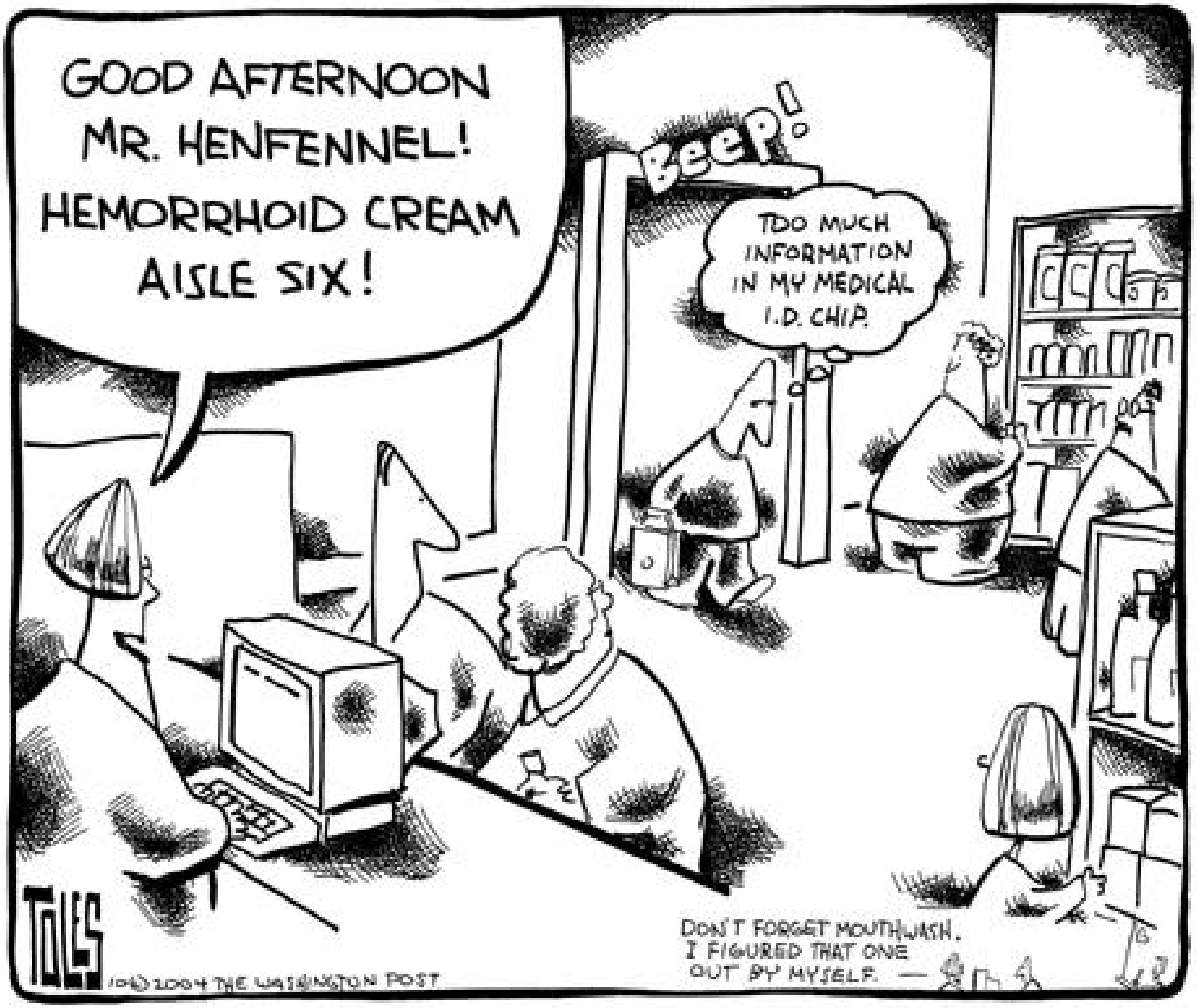
Workers in warehouses across Britain are being "electronically tagged" by being asked to wear small computers to cut costs and increase the efficient delivery of goods and food to supermarkets, a report revealed yesterday.



Under the system workers are asked to wear computers on their wrists, arms and fingers, and in some cases to put on a vest containing a computer which instructs them where to go to collect goods from warehouse shelves.

The system also allows su permarkets direct access to the individual's computer so orders can be beamed from the store. The computer can also check on whether workers are taking unauthorised breaks and work out the shortest time a worker needs to complete a job.

Academics are worried that the system could make Britain the most surveyed society in the world. The country already has the largest number of street security cameras.


GOOD AFTERNOON MR. HENFENNEL! HEMORRHOID CREAM AISLE SIX!
BEEP!
TOO MUCH INFORMATION IN MY MEDICAL I.D. CHIP.
DON'T FORGET MOUTHWASH. I FIGURED THAT ONE OUT BY MYSELF.
TOLES
10/6 © 2004 THE WASHINGTON POST

Proposed Approach

- Formalize currently piecemeal approach to core problem:
  - Optimize data quality
  - Protect Confidentiality
- Respond to Changing World
- Exploit existing knowledge in other areas
- Develop approach that is responsive to overwhelming demand for information but recognizes constraints

Economic Framework

Maximize $U= u(Q, R, N)$,

U is Data Utility

$Q=$ Data quality,

$R=$Researcher quality, and

$N=$number of times the data are accessed

If $Mi =$ modality i, then we can write $Q(Mi)$.

R and N are both determined by the access costs, A, imposed by the access modality, so $R(Ai)$ and $N(Ai)$.

Economic Framework

Subject to

$S = H. D + C$

S = social cost

H is harm

D is disclosure risk

C is cost to government

Economic Framework

$D^* = z(E, I, Z, Mi)$

- E is the existence and accessibility of other data sources that can be used for reidentification. The relationship between this and re-identification is affected by technology, T, and can be written $E(T)$
- I is the existence of malevolent interlopers. This relationship is affected by technology, legal penalties, L, and the characteristics of the population, X and can be written $I(T, L, X)$
- Z is researcher error. This is affected by technology, legal penalties, training and adoptable protocols, P and can be written $Z(T,L, P)$
- M, as before, is the set of access modalities

Constrained Optimization

$L = U – \lambda (H\ z(E,I,Z, Mi) + pt\ T + \Sigma Mi\ pAiMi\ – S\ )$

Using Framework to Shape a Research Agenda

1. *Developing metrics of data quality* **Q**
   - Domingo-Ferrer/Torra/Winkler/Shlomo/Haworth
2. *Quantifying the effect of the cost of access* ***A*** *on usage* ***N*** *and researcher quality* ***R***
   - Dunne/Seastrom
3. *Measuring harm* **H**
   - Madsen/Singer/Greenia (CDAC, 2005)
4. *Quantifying the relationship between other data sources* ***E*** *and disclosure* ***D***
   - Winkler/Domingo-Ferrer/Torra
5. *Modelling malevolent behavior I and researcher error* ***Z***
   - Feigenbaum/Agarawal/PORTIA project
6. *Investigating alternative technological approaches* ***T*** *to providing new access modalities* ***M***
   - *Cybertrust/Defense Department/RDC's/NSF funded researchers*

Next Steps

- Need active funding within statistical community
  - Consider portfolio approach – multiple modalities, human AND physical infrastructure (Portia Project)
  - Consortium of agencies (Census, BLS, BEA etc) to fund research agenda
- Leverage research outside statistical community
  - Conference of European Statisticians Statistical Confidentiality And Microdata Access – Principles And Guidelines Of Good Practice
  - Engagement with other academic communities (e.g. cybertrust/IIS (Information, Privacy and Security) initiatives at NSF; DARPA); IASSIST
  - Role of supercomputer centers