Statistical Disclosure Limitation & the Challenge of Societal-Scale Data

Stephen E. Fienberg

Department of Statistics, Heinz College, Machine Learning Department, and Cylab

Carnegie Mellon University

Pittsburgh, PA 15213-3890 USA

fienberg@stat.cmu.edu

Census Confidentiality I

- **1790-1840 Census data publically posted.**
- **1850 Census includes new demographic data, no public posting but unrestricted access continued.**
- **1910 President Taft unequivocally promised confidentiality for all census information collected.**
- **1929 Census Act made individually identifiable information confidential.**
- **1940 Census long form done on a sample basis.**
- **1942 Under War Powers Act, Census Bureau released individually identifiable information to assist in the internment/relocation of Japanese Americans.**

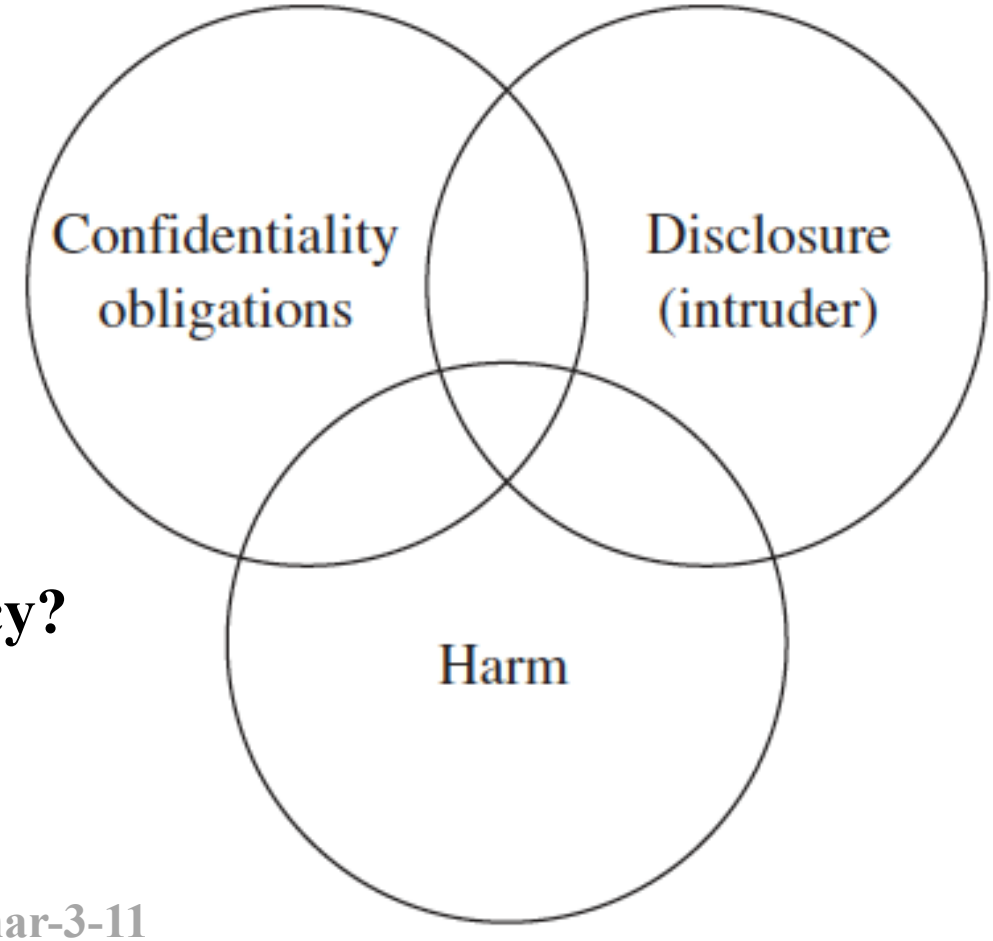
Census Confidentiality II

- **1952 Census data available from Archives after 72 years.**
- **1954 Title 13 of US Code describes confidentiality of Census data for 1960 census forward.**
- **1958 dispute over access to enterprise data from Economic Census.**
- **1960-present Census Bureau releases sample census microdata files, PUMS, without individually identifiable information (but with minimum population threshold for geography−from 1 million to 250,000 to 100,000).**

Census Confidentiality III

- **1994 *Report on Statistical* Disclosure Limitation Methodology, Statistical Policy Working Paper 22 (2nd version, 2005)**
- **2000 *AmericanFactfinder* for release of census tabulations and microdata files.**
- **2000-2010 *American Community Survey* replaces long form.**

Disclosure Limitation, Confidentiality & Harm



Where is privacy?

Outline

- **The Census Bureau snafu.**
  - **Principles of data sharing and statistical disclosure limitation.**
  - **Risk-Utility trade-off.**
- **Differential Privacy (DP) in a focused statistical problem:**
  - **Protecting contingency table data.**
- **Moving to societal-scale data.**

THE WALL STREET JOURNAL.
WSJ.com

THE NUMBERS GUY | FEBRUARY 6, 2010

Census Bureau Obscured Personal Data—Too Well, Some Say

By CARL BIALIK

Errors in some U.S. Census Bureau data are sending researchers inside and outside government scrambling to check whether some key findings need to be reassessed.

After the Census Bureau compiles overall counts in its decennial population surveys and other studies, it releases additional details about respondents to outside researchers. But in order to protect respondents' privacy, the bureau masks some of the personal information in these so-called microdata.

A study has found the agency went too far hiding individual identities, introducing errors that might lead economists and demographers astray. By relying on the microdata, researchers would have found, for example, evidence of a steep drop-off in marriage rates for women at age 65, or of a big rise in the proportion of women in their early 70s who are working—both false conclusions.

The anomalies highlight how vulnerable research is to potential problems with underlying numbers supplied by other sources, even when the source is the government. And they illustrate how tricky it can be to balance privacy with accuracy.

Census Costs & Products

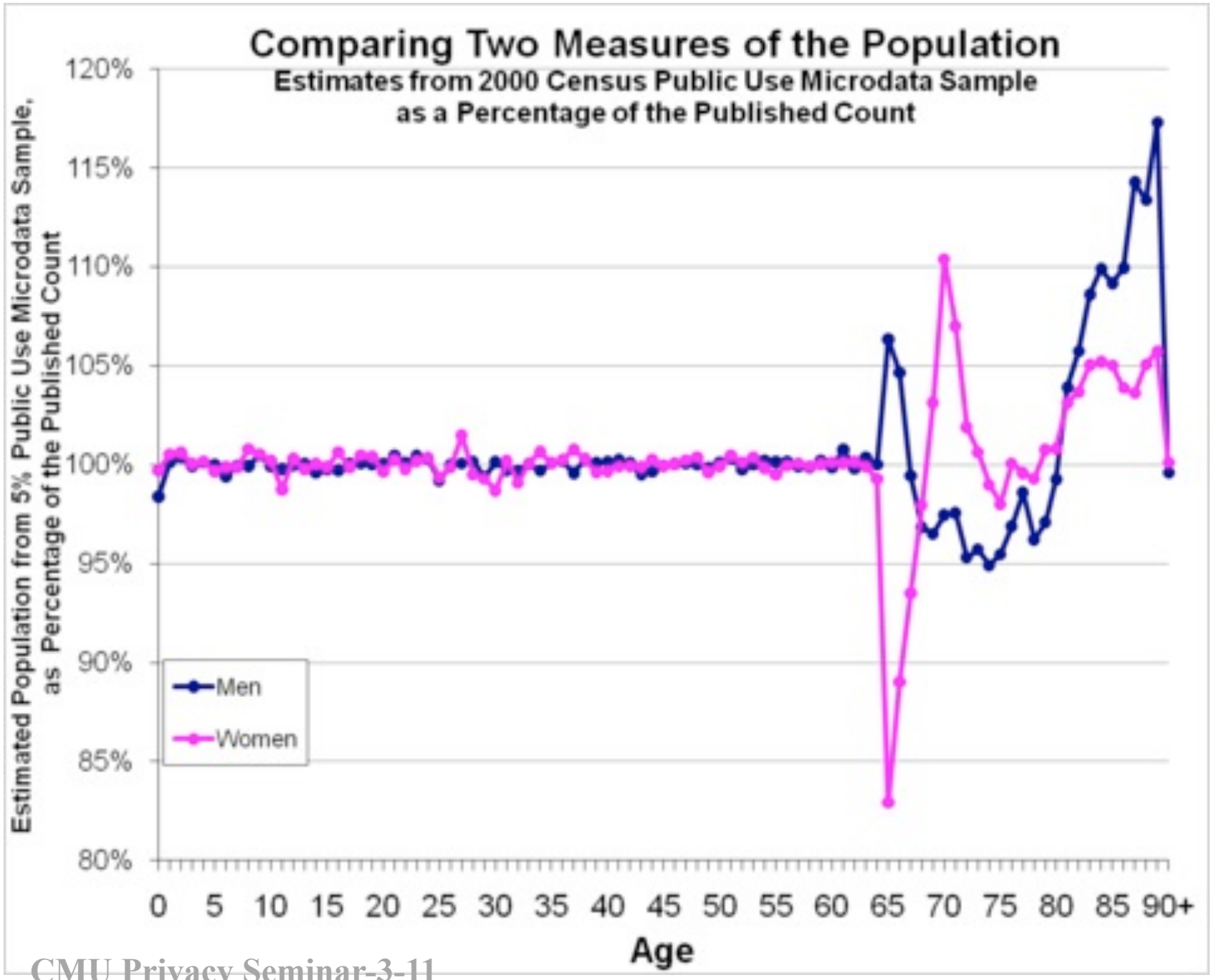
- **Costs: $6.5 billion in 2000; $14 billion+ in 2010**
- **Short form data (100%)**
  - **State totals by Dec. 31 for reapportionment**
  - **Age (<18, ≥18) × Gender × Race for each census block to states for redistricting**
- **Long form data (sample of 1 in 6) via American Factfinder (replaced by ACS data in 2010):**
  - **Allocation of funds: $400 billion in 2010**
  - **Tables and special packages (e.g., travel-to work info for urban planners, etc.)**
  - **1% and 5% PUMS**

Public Use Microdata Samples (PUMS)

- **5% PUMS Files**
  - **PUMS contain individual data for geographic units known as super-Public Use Microdata Areas (super-PUMAs) and Public Use Microdata Areas (PUMAs). Each PUMA must have a minimum of 100,000 population and each super-PUMA contains a minimum population of 400,000.**

Census Disclosure Protection Approach

- Data swapping & Sampling & Imputation/ Editing
- PUMS files
  - Top-coding for variables like income
  - Population controls for geography
  - Some outlier values are averaged together, and that average is assigned to every one of those outliers.
  - Addition of statistical noise to the subset of older respondents
- No details on properties of each of these components, e.g. % of swapped files


Comparing Two Measures of the Population
Estimates from 2000 Census Public Use Microdata Sample
as a Percentage of the Published Count
Estimated Population from 5% Public Use Microdata Sample,
as Percentage of the Published Count
120%
115%
110%
105%
100%
95%
90%
85%
80%
Men
Women
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90+
Age

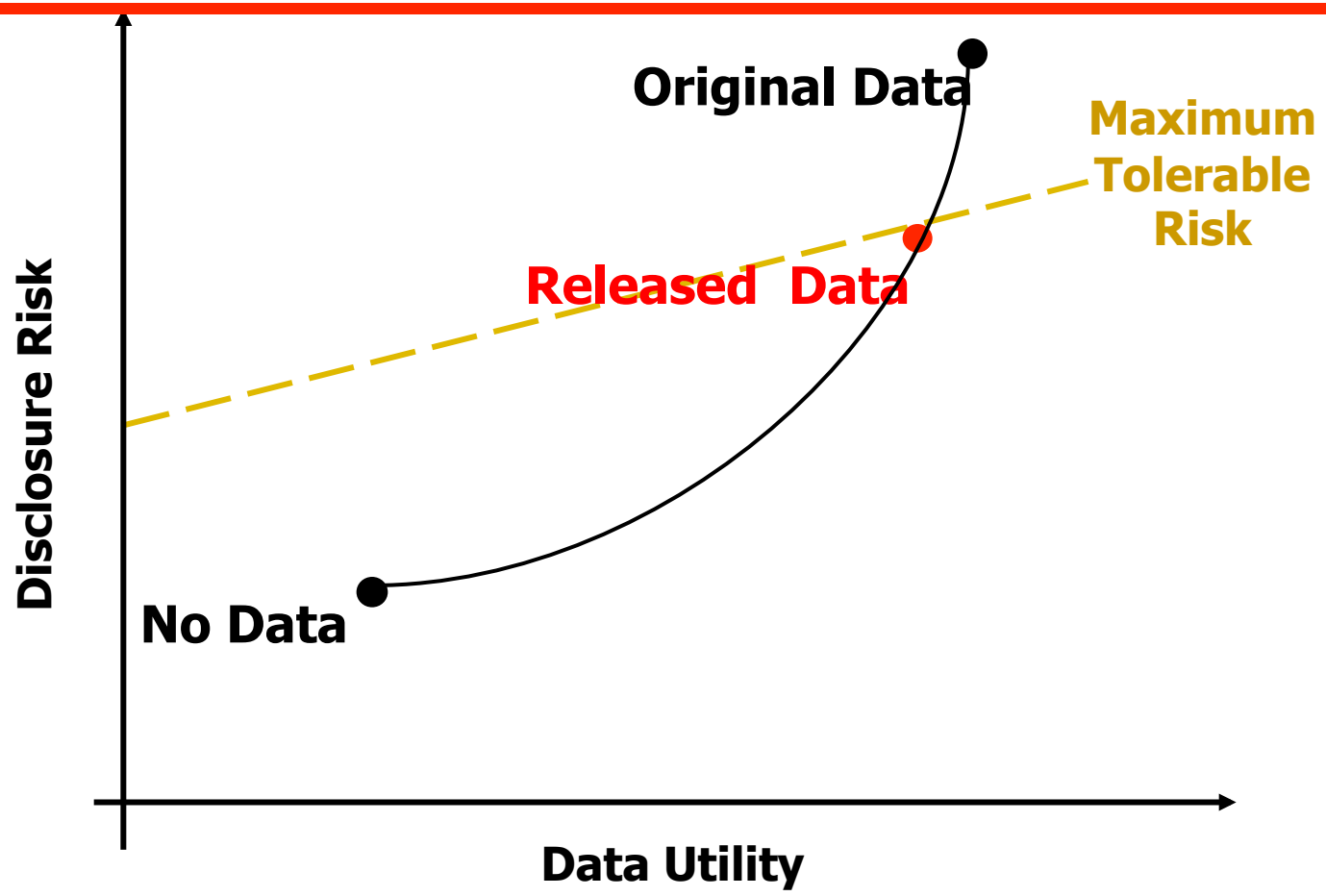
Census Bureau Response

- **“We want to preserve confidentiality, and we want to maximize utility of our data. This tension is inherent in everything we do,” says Robert M. Groves, director of the Census Bureau.**
- **“Flawed software code designed to add the statistical noise to the subset of older respondents should have offset those changes with opposite adjustments made elsewhere in the data sample. This didn't happen as it should have, so that ages and other attributes were skewed.”**
- **Before the data were released in 2003, the Census Bureau's diagnostic tools flagged the problem, but it “didn't seem large enough in the judgment of our analysts to stop the release,” says Dr. Groves.**

Morales of Story

- **For the Census Bureau:**
  - **Ad hockery in DL can lead you astray.**
  - **Not releasing the details of DL methodology will likely get you in trouble in the long run.**
- **For us:**
  - **"Accuracy" of "released" statistical data matters to both users and data owners.**
  - **Privacy protection is for the data at hand and not for possible replications that we will never see.**

R-U Confidentiality Map



(Duncan, et al. 2001, 2004,
Trottini, 2002; Ting et al. 2008)

Usability, Transparency, & Duality in Privacy Protection

- **Usability:** extent to which released data are free from systematic distortions that impair inference.
- **Transparency:** extent to which methodology provides direct or implicit information on bias and variability resulting from disclosure limitation mask.
- **Duality:** extent to which methods aim at both disclosure limitation and making the maximal amount of data available for analysis.

General Methods for Protection

- **Removing obvious identifiers/near-identifiers**
  - **Names, geography, birthdate, etc.**
- **Data transformations:**
  - **Matrix masking X → AXB + C**
    - **e.g., noise addition**
  - **Data suppression**
    - **Deleting cases / sampling**
    - **Cell suppression**
- **Synthetic data**
  - **Sampling from posterior distribution**

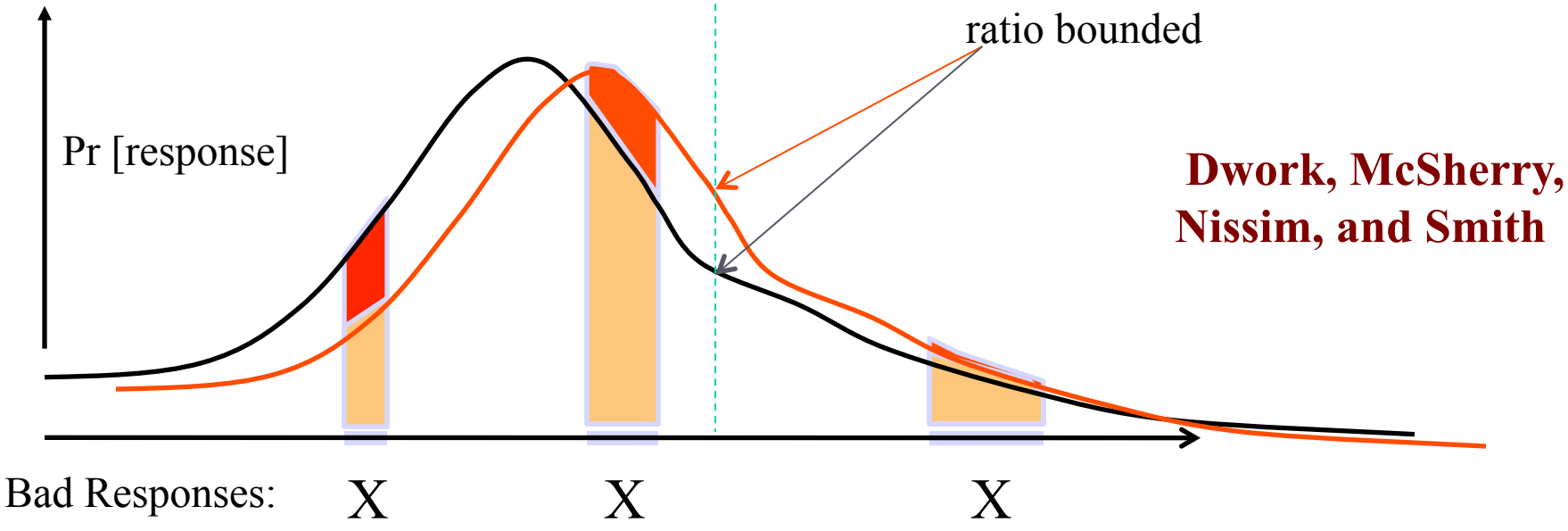
Inferential Utility

- **"Statistical reversibility" of data transformation:**
  - **Need (a) released data and (b) likelihood function including full information on transformation applied.**
  - **For noise addition this may involve using "measurement error model" since most (all) of variables are measured with error.**

ε-Differential Privacy

Randomized function $\mathcal{K}$ gives ε-differential privacy if for all neighboring D_1 and D_2, and all $C \in \text{range}(\mathcal{K})$:

$$\Pr[\mathcal{K}(D_1) \in C] \leq e^{\varepsilon} \Pr[\mathcal{K}(D_2) \in C]$$



Differential Privacy

- **The standard "DP mechanism" is the addition of doubly exponential noise, with a parameter ε.**
- **The more data or statistics you protect the larger the noise required.**

Differential Privacy

- **DP offers strong privacy "guarantees," through all possible violations, but…**
  - **Strong privacy "guarantees" may destroy utility of the data.**
  - **Does not recognize the iterative and possibly unstructured nature of statistical data analysis.**
- **Research users want data sets to analyze, not DP-protected coefficients.**

Differential Privacy

- **DP is fundamentally a *frequentist* notion:**
  - **Privacy resides in the method that generates the altered data, as well as extremal aspects of data themselves.**
  - **Has the flavor on minimax approaches.**
- **But for "my problems," data are in hand when we begin to consider data release and disclosure limitation (not privacy).**

Protecting Contingency Tables

Barak et al. (2007)

- **Want to release a set of altered MSS marginals.**
  - **Use Fourier coefficient basis for noise addition.**
  - **This produces non-integer and inconsistent margins.**
  - **Consistency of margins doesn't guarantee existence of a table satisfying released margins.**
  - **Barak et al. find "nearby" set of consistent integer margins which preserve DP property.**
- **What about**
  - **releasing *n*? Known in all of my applications!**
  - **utility?**

Yang, Fienberg, & Rinaldo: Examined DP Approach

Robustness of approach for RU tradeoff

- **Edwards 2^6 genetics table, with n=70.**
- **Czech auto workers 2^6 heart attack risk table, with n=1,841.**
- **Rochdale 28 survey data on women's work, with n=665;very sparse structure.**
- **American Community Survey 4 × 4 × 16 travel to work table.**

- **National Long Term Care Survey**
  - **2^{16} disability table with n=21,574.**
  - **2^{96+5} version based on 6 waves (plus mortality), n~45,000. Our models have no MSSs!**

Our Approach

- **We used an *ad hoc* approach to utility by looking at**
  - **For each noise level, we compute the deviance (KL-distance) between the MLE and 100 tables perturbed at this noise level.**

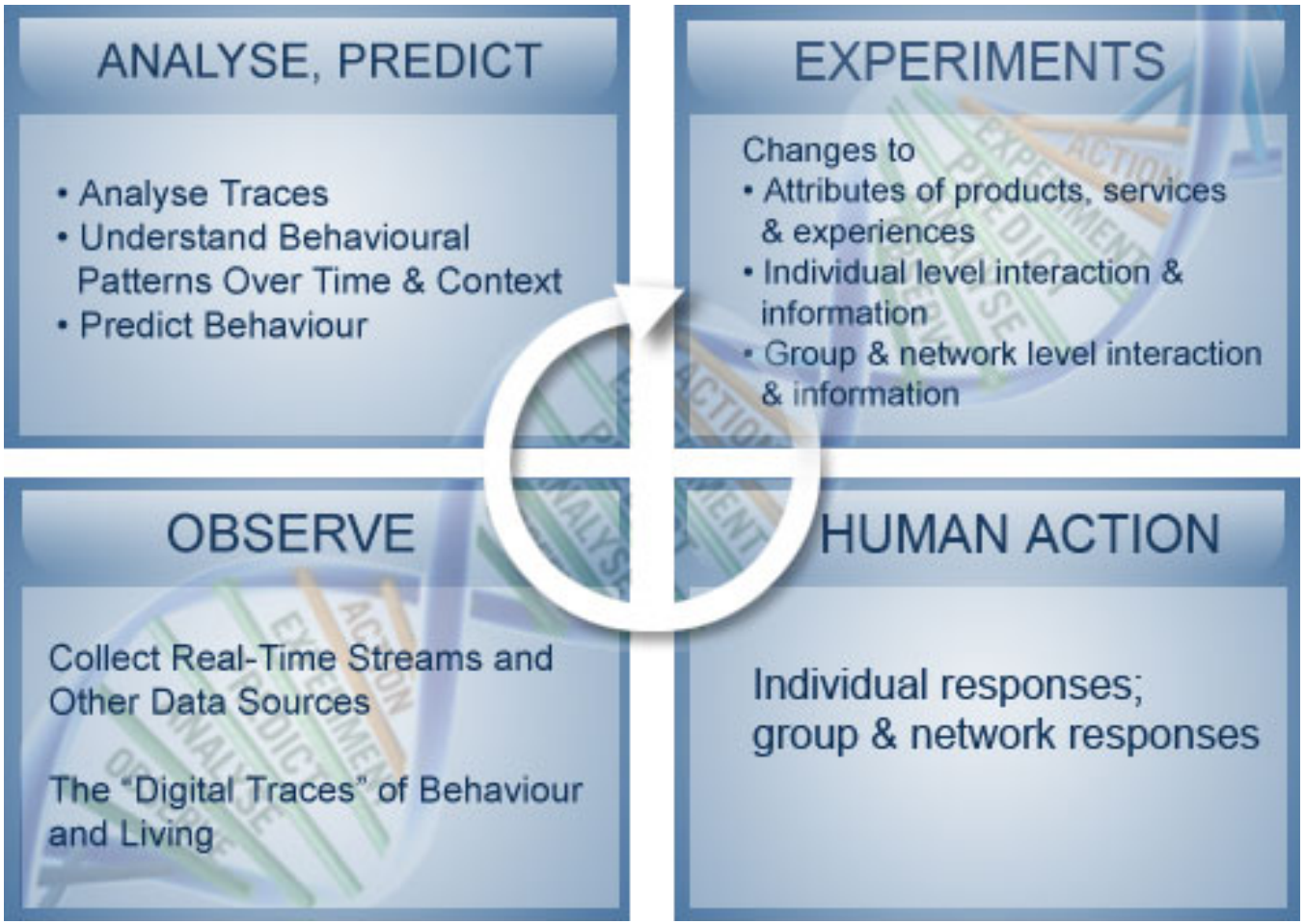
Lessons Learned

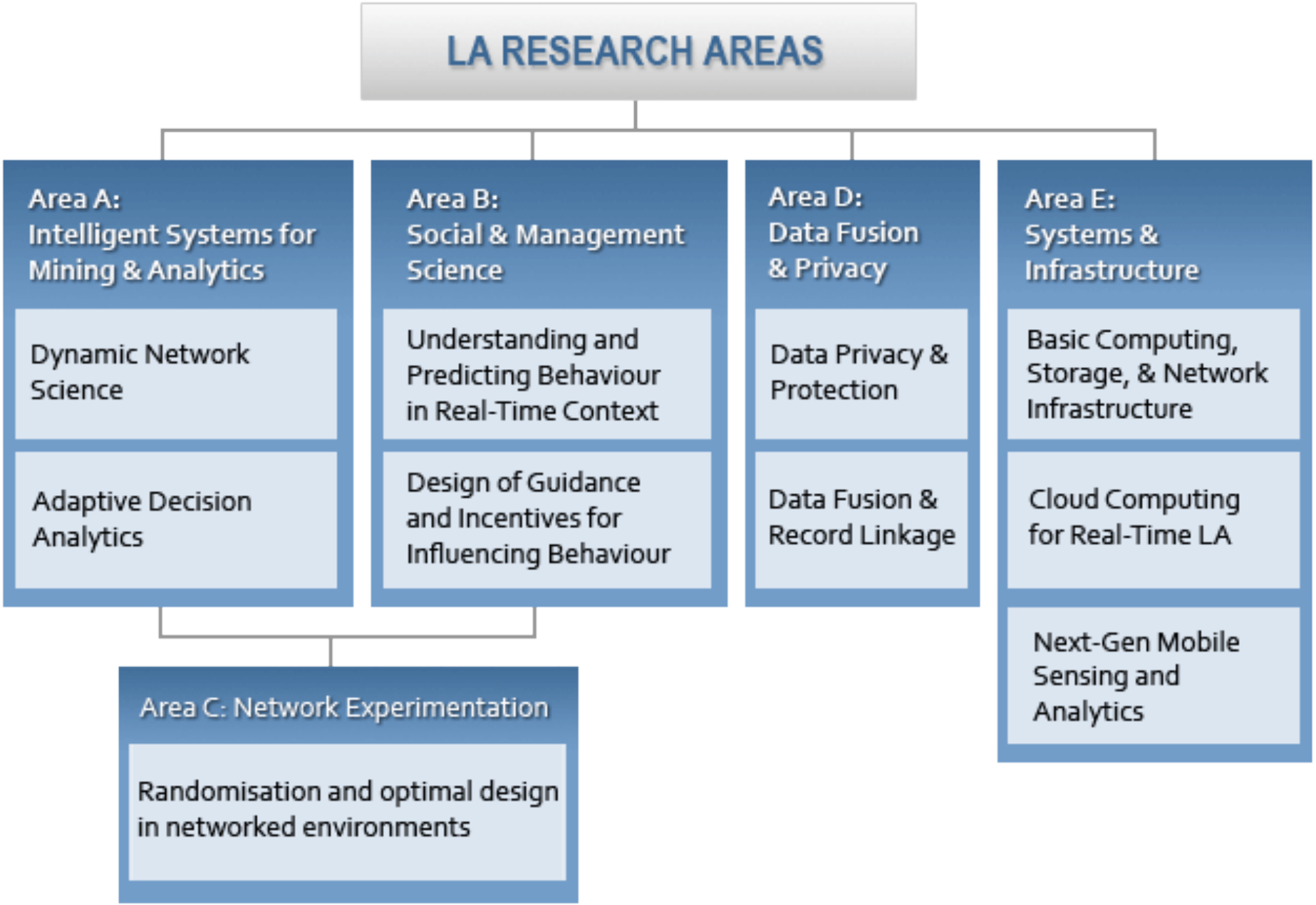
- As ε increases, amount of noise added decreases
  - Deviance between DP generated tables and real MLEs gets smaller.
  - If we add a lot of noise, it has strong privacy guarantees but the statistical inference becomes infeasible.
  - When we add little noise, the statistical inference is better but no privacy guarantees.
- DP struggles with releasing useful information associated with large sparse contingency tables.

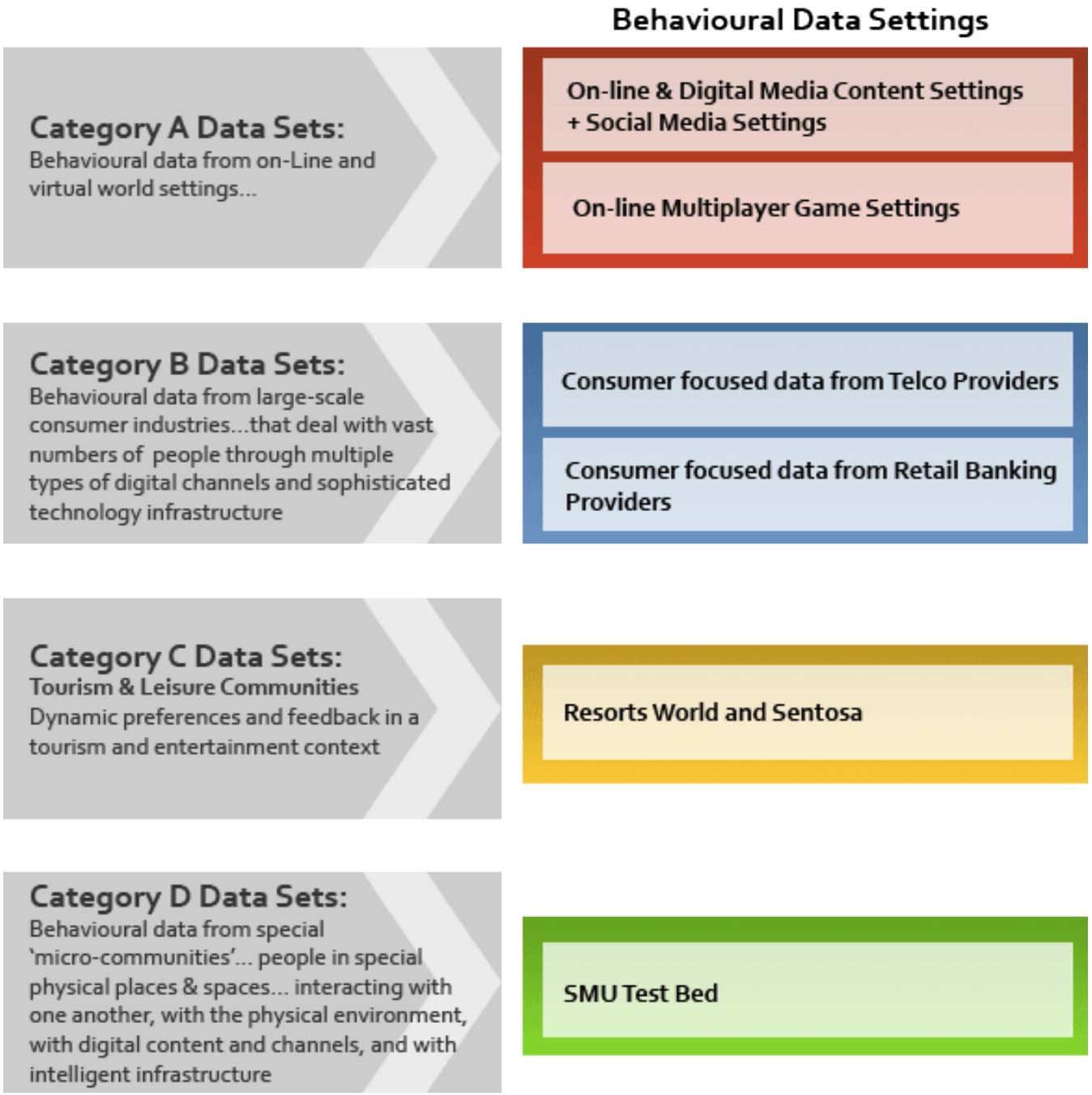
Possible Implications

- **We need to:**
  - **Incorporate RU ideas into DP formulation so that data releases have real utility:**
    - **Learn how to draw inferences from privacy-protected releases.**
    - **Focus on model search processes, not simply reporting one set of summary statistics.**
  - **Move from frequentist to Bayesian formulation:**
    - **Provide protection for actual data at hand.**
    - **Identify inferences from "record linkage"?**

CMU-SMU Living Analytics Research Centre




LA RESEARCH AREAS
Area A: Intelligent Systems for Mining & Analytics
Dynamic Network Science
Adaptive Decision Analytics
Area B: Social & Management Science
Understanding and Predicting Behaviour in Real-Time Context
Design of Guidance and Incentives for Influencing Behaviour
Area D: Data Fusion & Privacy
Data Privacy & Protection
Data Fusion & Record Linkage
Area E: Systems & Infrastructure
Basic Computing, Storage, & Network Infrastructure
Cloud Computing for Real-Time LA
Next-Gen Mobile Sensing and Analytics
Area C: Network Experimentation
Randomisation and optimal design in networked environments


Behavioural Data Settings
Category A Data Sets:
Behavioural data from on-Line and virtual world settings...
On-line & Digital Media Content Settings + Social Media Settings
On-line Multiplayer Game Settings
Category B Data Sets:
Behavioural data from large-scale consumer industries...that deal with vast numbers of people through multiple types of digital channels and sophisticated technology infrastructure
Consumer focused data from Telco Providers
Consumer focused data from Retail Banking Providers
Category C Data Sets:
Tourism & Leisure Communities
Dynamic preferences and feedback in a tourism and entertainment context
Resorts World and Sentosa
Category D Data Sets:
Behavioural data from special 'micro-communities'... people in special physical places & spaces... interacting with one another, with the physical environment, with digital content and channels, and with intelligent infrastructure
SMU Test Bed

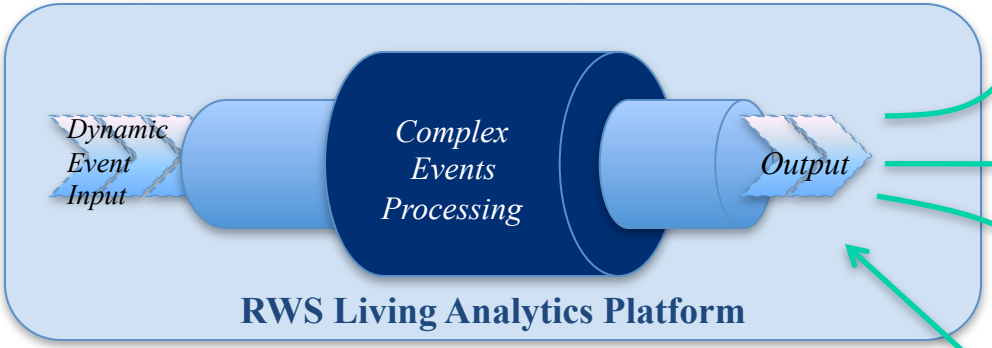
UNIVERSAL/RESORTS WORLD

Ride Status (Wait Time)

Show Schedules

Dynamic Day Itinerary Planner

Customer Profile, Preferences & Trip Info (Check in/out date, Hotel Location, Single/Two Day Pass)



Optimize Stay & Maximize Trip Experience

Recommendations & Incentives

Restaurant Wait Queue

Weather Condition

Traffics Condition

Management Benefits:
- Load balance traffics & manage congestion
- Gain customer behavior insights
- Improve customer services
- Improve operational efficiency & productivity

LiveLabs @ SMU

Insight at the Intersection of the Physical & Virtual Social World

Focus: Youth Activities & Lifestyles

- Socially-networked interactions & peer-driven behavior
- Rich digital media consumption
- Smartphone-centric worldview

5000 person Campus Testbed

Capture Daily Activities in Indoor Spaces

- Fine-grained indoor location & activity history
- Continuous mobile sensing
- Campus, shopping centers, hawker centers, leisure & sports venues

Real-time analytics on combined virtual & physical context

- Next-gen IDM & Telco architecture to enable data fusion & mining.
- Incorporation of Internet, cable, phone & physical activity

LARC Privacy & Security

Summary

- **The Census Bureau snafu.**
  - **Principles of data sharing and statistical disclosure limitation.**
  - **Risk-Utility trade-off.**
- **Differential Privacy (DP) in a focused statistical problem:**
  - **Protecting contingency table data.**
- **Moving to societal-scale data.**

End

- **My CMU privacy collaborators:**
  - **Rob Hall, Jiashin Jin, Alessandro Rinaldo, Xiaolin Yang, Larry Wasserman**
- **Joint CMU/PSU/Cornell collaboration**
- **LARC**
  - **http://www.larc.smu.edu.sg/**

References

- **Barak, B., Chaudhuri, K., Dwork, C., Kale, S., McSherry, F., and Talwar, K. (2007). Privacy, accuracy, and consistency too: a holistic solution to contingency table release. *PODS* 2007: 273–282.**
- **Dobra, A., Fienberg, S.E., Rinaldo, A., Slavkovic, A.B. and Zhou, Y. (2008). Algebraic statistics and contingency table problems: Log-linear models, likelihood estimation, and disclosure limitation. In *Emerging Applications of Algebraic Geometry* (M. Putinar and S. Sullivant, eds., Springer, New York, 63–88.**
- **Fienberg, S.E., Rinaldo, A., and Yang, X. (2010). Differential privacy and The risk-utility tradeoff for multi-dimensional contingency tables, In J. Domingo-Ferrer and E. Magkos, eds., *Privacy in Statistical Databases* 2010 (PSD 2010), LNCS Vol. 6344, Springer, pp. 187–199.**