Institute for Employment Research

The Research Institute of the Federal Employment Agency

IAB

# My understanding of the differences between the CS and the statistical approach to data confidentiality

**The 4th IAB workshop on confidentiality and disclosure - bridging approaches from statistics and computer science**

Jörg Drechsler

30. June 2011, Nürnberg

# Outline

- History of Statistical Disclosure Control
- Some Aspects of Data Privacy in CS (to the best of my knowledge)
- Differences Between the Two Approaches
- What (I Think) We Can Learn from Each Other
- Further Thoughts

# History of Statistical Disclosure Control

- Most data collected by statistical agencies in the early days
- All information concerning the data was published in tables
- Access to the microdata for external researchers was unthinkable
- Research on data confidentiality mainly focused on tabular data
- Confidentiality for tabular data still a very important topic for statistical agencies
- First papers on microdata confidentiality in the early eighties (Data swapping, Dalenius and Reiss (1982))
- Data confidentiality for microdata can be achieved in two ways
  - Information reduction
  - Data perturbation

# Information Reduction

- Information that poses a possible risk of re-identification is suppressed
- Possible methods:
  - top coding
  - global recoding
  - local suppression
  - droping variables
- Advantage
  - all released information is unaltered
- Disadvantage
  - important information is lost
  - information reduction might be so severe for sensitive data that the dataset will become useless

# Data perturbation

- all variables remain in the dataset but individual records are altered to guarantee data confidentiality
- Possible Methods:

  - swapping
  - noise addition
  - micro aggregation
  - data shuffling
  - synthetic data

- Advantage
  - all information is still available in the released data
- Disadvantage
  - data have been altered
  - important relationships found in the original data might be distorted

# Data Dissemination in Practice

- Most users are sceptical about perturbation approaches
- Situation in Germany
  - Almost all released datasets are only protected by information reduction
  - Almost no business data available
- Situation in the US and other countries
  - Perturbation methods widely used
  - Methods are better accepted by the users
  - Statistical agencies still mostly use traditional techniques
  - negative consequences of these techniques have been repeatedly shown (Winkler, 2007)

# Latest developments

- Third alternative to research data centers and data dissemination
- Much research on remote analysis servers and remote data access
- Remote analysis servers
  - User doesn't have access to the microdata
  - Can select his analysis from a drop-down menu
- Full Remote Access
  - User connects to a server from his desktop computer
  - User has full access to the microdata
  - Server automatically suppresses output that might violate confidentiality requirements
  - Data never leave the secure environment

# Some aspects of data privacy in CS

- Not limited to data from statistical agencies
- Broader perspective
- Encryption
- Privacy-preserving data mining
  - External user submits queries to a system
  - Queries are only answered if privacy is protected
- Secure multiparty computation
  - Two or more parties would like to analyze their datasets jointly
  - Datasets can’t be shared
  - Aim is to perform the analysis without the need to share the data

# Privacy-preserving data publishing (PPDP)

- Goal is to publish microdata
- Strong focus on algorithms (sanitization mechanisms)
- General aim is to offer some formal privacy guarantee
- Examples of privacy definitions
  - *k*-anonymity
  - *l*-Diversity
  - ($\alpha$,$\beta$)-privacy
  - $\varepsilon$-differential-privacy

# Differences between the two approaches

## Statistical Inference

- SDC deals with data privacy mostly from the survey statistics perspective
- PPDP deals with any kind of data
- PPDP not interested in making statistical inference

## Privacy guarantees

- very vague in SDC
- no *ex ante* guarantees
- level of protection has to be evaluated for each dataset
- PPDP offers formal privacy guarantees

# Differences between the two approaches

## Analytical validity

- SDC more concerned about analytical validity
- Valid inferences should be obtainable
- PPDP focuses mainly on formal privacy guarantees

# What can we learn from each other?

## SDC

- formal privacy definitions are very helpful
- most disclosure risk evaluations in SDC are subjective
- results from PPDM might be useful in the remote access context

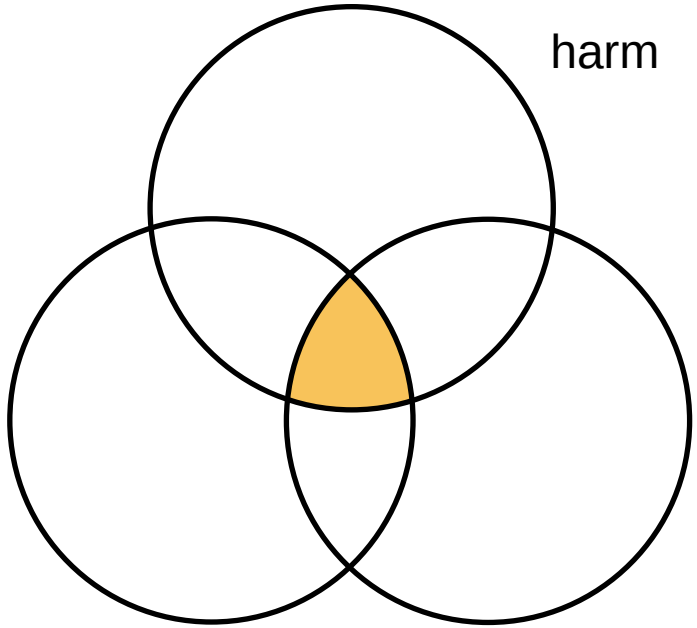
## PPDP

- formal privacy definitions alone are not sufficient
- preserving analytical validity equally important
- current assumptions about intruder knowledge are too restrictive
- include statistical inference in data utility evaluations

# Further thoughts

- Current goal: No risk of disclosure if the data are released
- Goal might be too restrictive

Steve Fienberg:

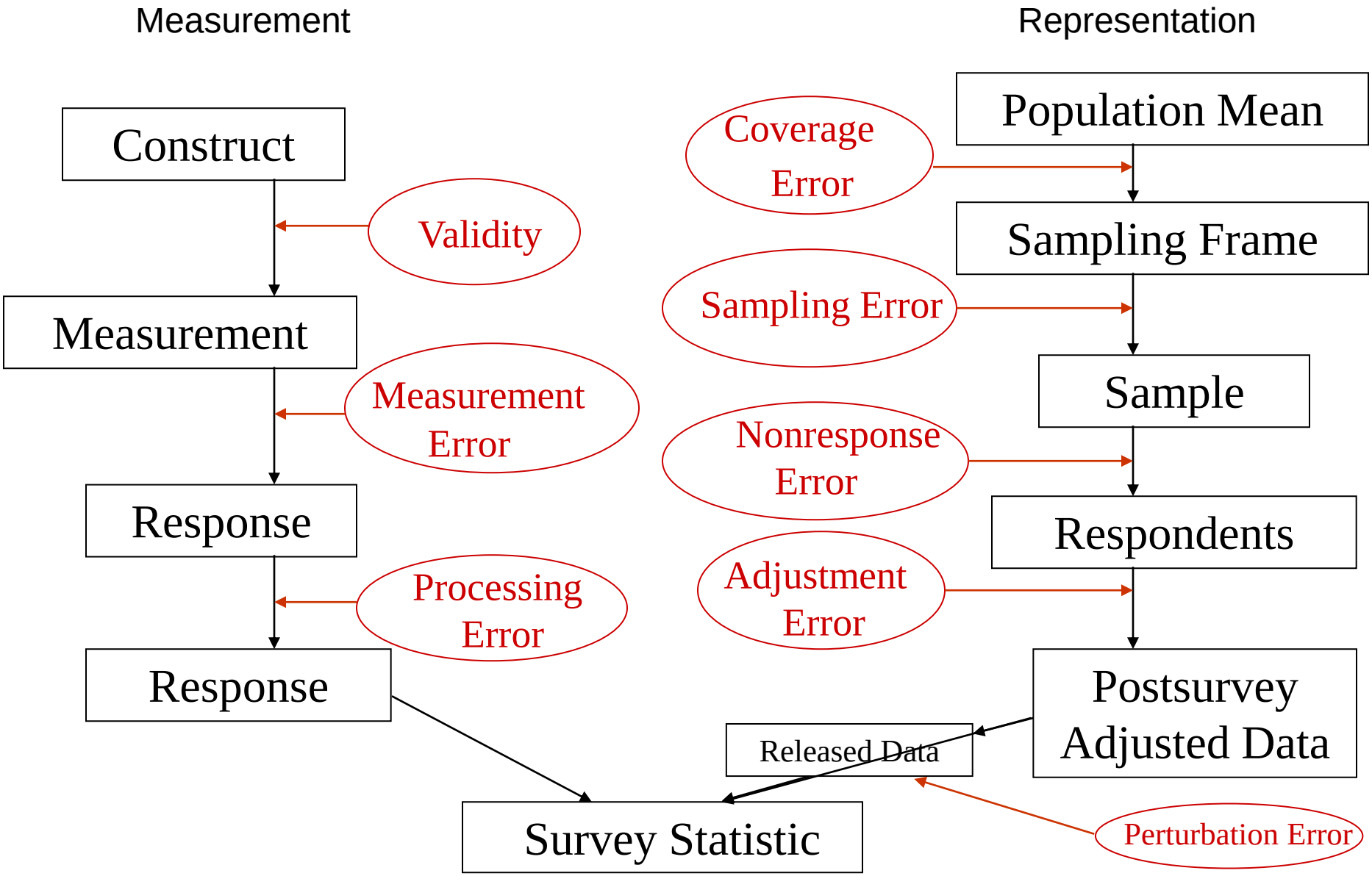



- But how do we measure harm?

# Valid Statistical Inference

- Users are reluctant to use disclosure protected data, especially if data have been protected by perturbation methods.
- Generating disclosure protected data that will provide valid results for any possible query is impossible.
- Useful SDC method should provide information for the user, which analyses might provide valid results.
- Bias from protection methods are by no means the only source for potential bias in the results.
- Other sources for bias might easily dwarf the bias from protecting the data.

# Total Survey Error



Institute for Employment Research

The Research Institute of the Federal Employment Agency

IAB

# Thank you for your attention