Assessing Data Quality

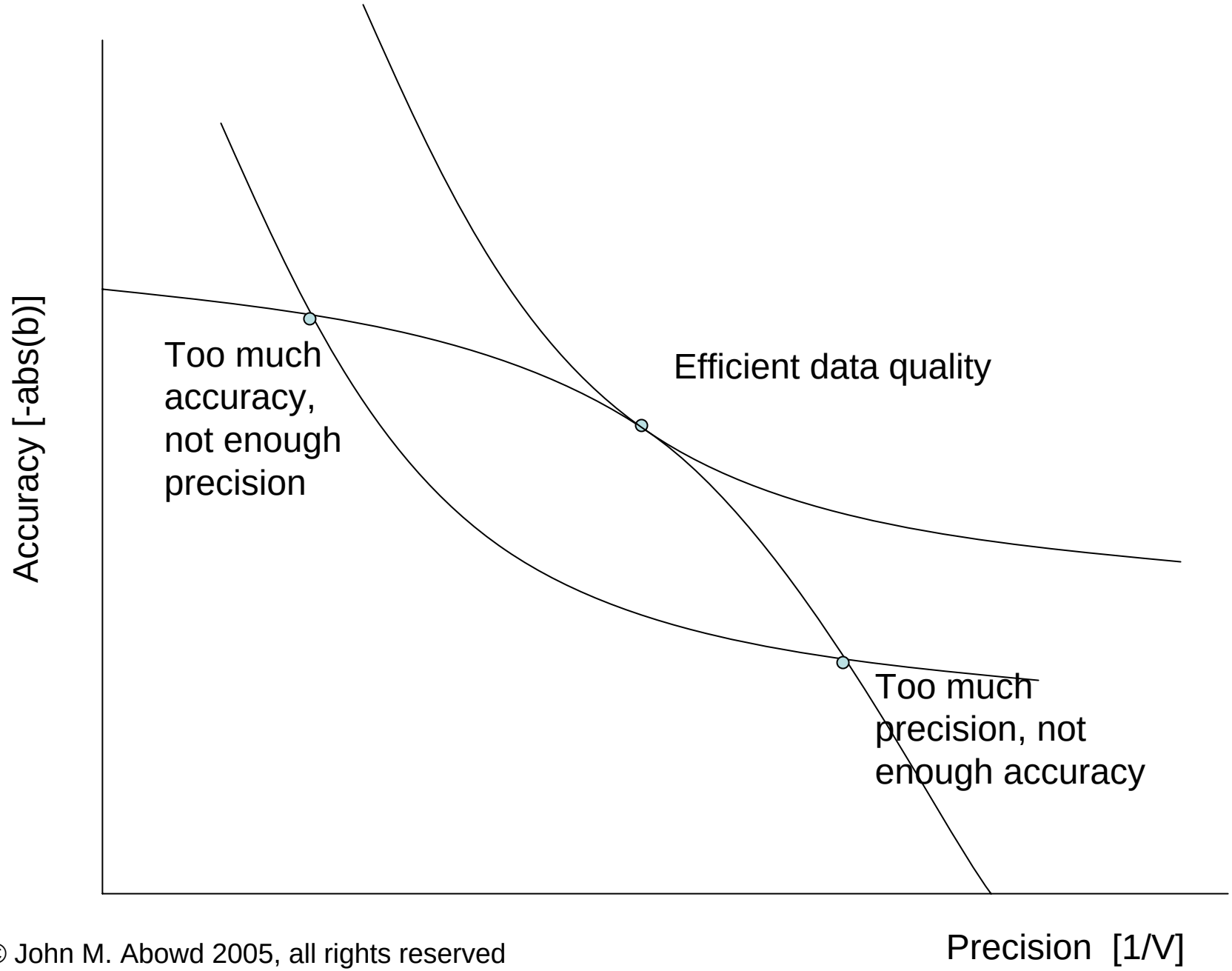
John M. Abowd
April 2005

Outline

- Summary
- Primary collection objective
- Secondary data analysis
- In-scope population(s)
- Sampling frame(s)
- Completed data
- Benchmarking
- Publications

Summary

- Data quality can be given a precise definition (simplified here)
- For a given hypothesis, the utility associated with a given data collection is a function of the accuracy (-abs(bias)) and precision (inv(var)) for a given estimand Q
- The resources associated with generating the estimand Q can be used to reduce bias or increase precision



Primary Data Collection

- Bias is controlled by
  - Instrument design
  - Sampling frame maintenance
  - Editing
  - Confidentiality protections
- Precision is controlled by
  - Sampling scheme
  - Complexity of the concept under measure

Secondary Data Analysis

- Most bias considerations are hidden in the preparation of the analysis file
  - Non-response
  - Item missing and editing
  - Accuracy of the instrument
- Most precision gains are illusory
  - Sampling design is given
  - Assessments of precision often ignore design

QA of In-scope Population

- Recall our discussion of the Decennial Census
  - Original frame
  - Initial Census results
  - Census 2000 Supplemental Survey
  - Independent demographic analysis

QA of Frame Maintenance

- Diverging trends
- Household measures of employment v. establishment measures of employment
- How do you determine which is correct?

QA of Data Completion

- Basic non-response (missing from frame or sample)
- Item non-response
- Incomplete response
- Confidentiality protections

Benchmarking

- For household surveys: uses independent population estimates
- For establishment surveys: uses independent business size estimates

Publications

- How do the factors we have identified here affect official publications?
- Often unknown but very important