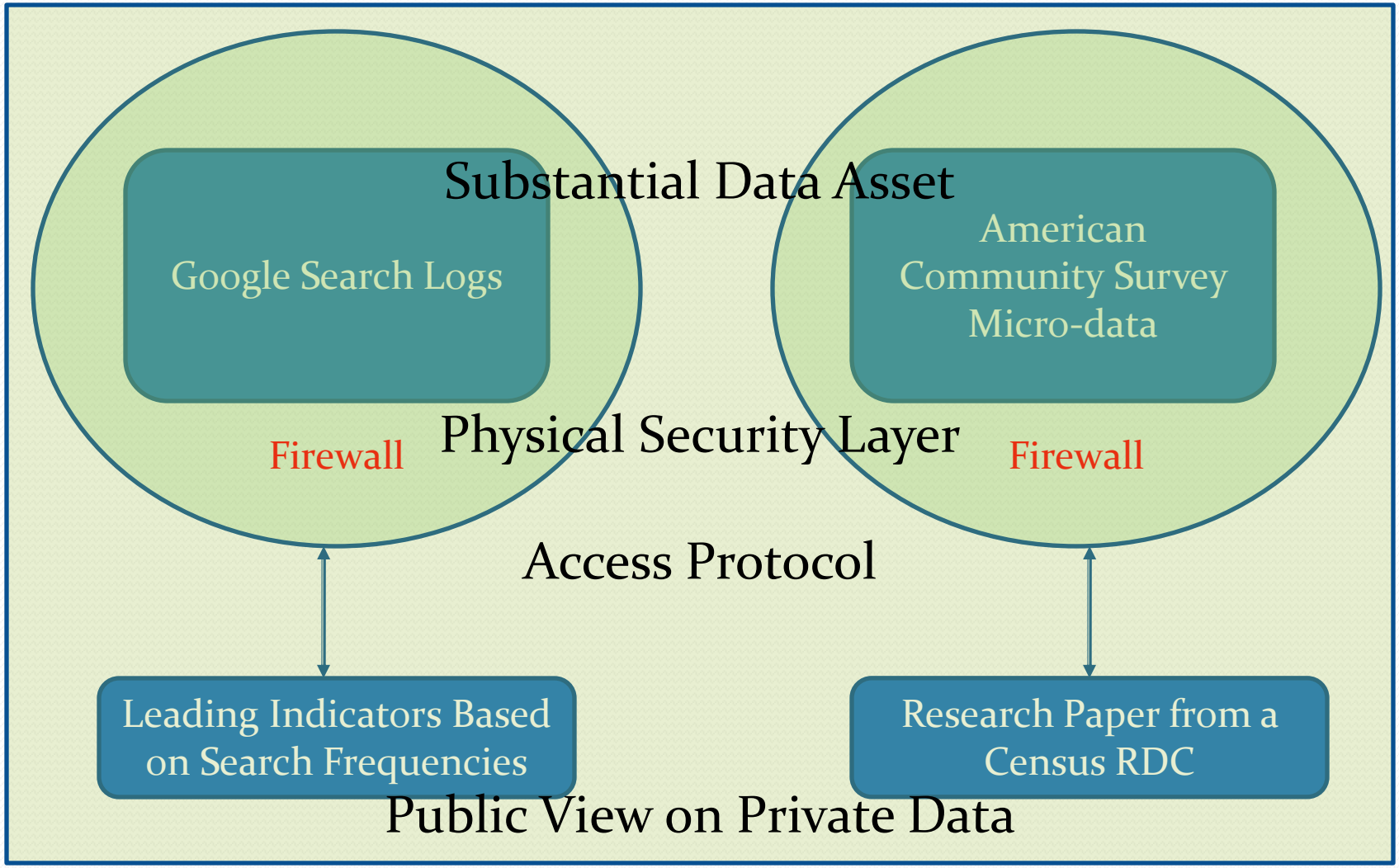
Providing Secure Access to Sensitive Data

John M. Abowd
Cornell University
IASSIST 2010
June 4, 2010

Cornell University

Everybody's Got the Same Problem



But They Don't Use the Same Strategies

- Microsoft, Google, Yahoo!, Amazon tend to rely on formal privacy methods
- Scientific validity and future data integrity are corporate assets
- Most strategies explicitly integrate content and privacy functions
- Distinguish between types of data releases:
  - Official statistics
  - Special tabulations
  - Public-use files
  - Protocol data access (RDC, AFF Advanced Query)
- Most strategies isolate the content and disclosure limitation functions

And They Don't Make the Same Mistakes

- AOL and Netflix search log releases were fully preventable
- Information loss from randomized sanitizers is fully quantifiable but scientifically difficult to digest
- Not as concerned about the public good aspect of data release
- Research papers viewed as usually safe when based on statistical modeling
- Concerned about the public good from statistical publications but limited by ad hoc disclosure avoidance procedures that provide security through obscurity

What To Do Next

- Recognize that every public view from official statistics to model output in scientific papers can be assessed for both statistical validity and privacy/confidentiality leakage
  - There is a continuum from official statistics through "need-to-know" privileged access that all research paradigms touch—think more about the substance and less about the form
- Integrate research teams primarily interested in content with teams primarily interested in privacy protection
  - Be certain that both groups know what the other is doing