Anonymity and Risk

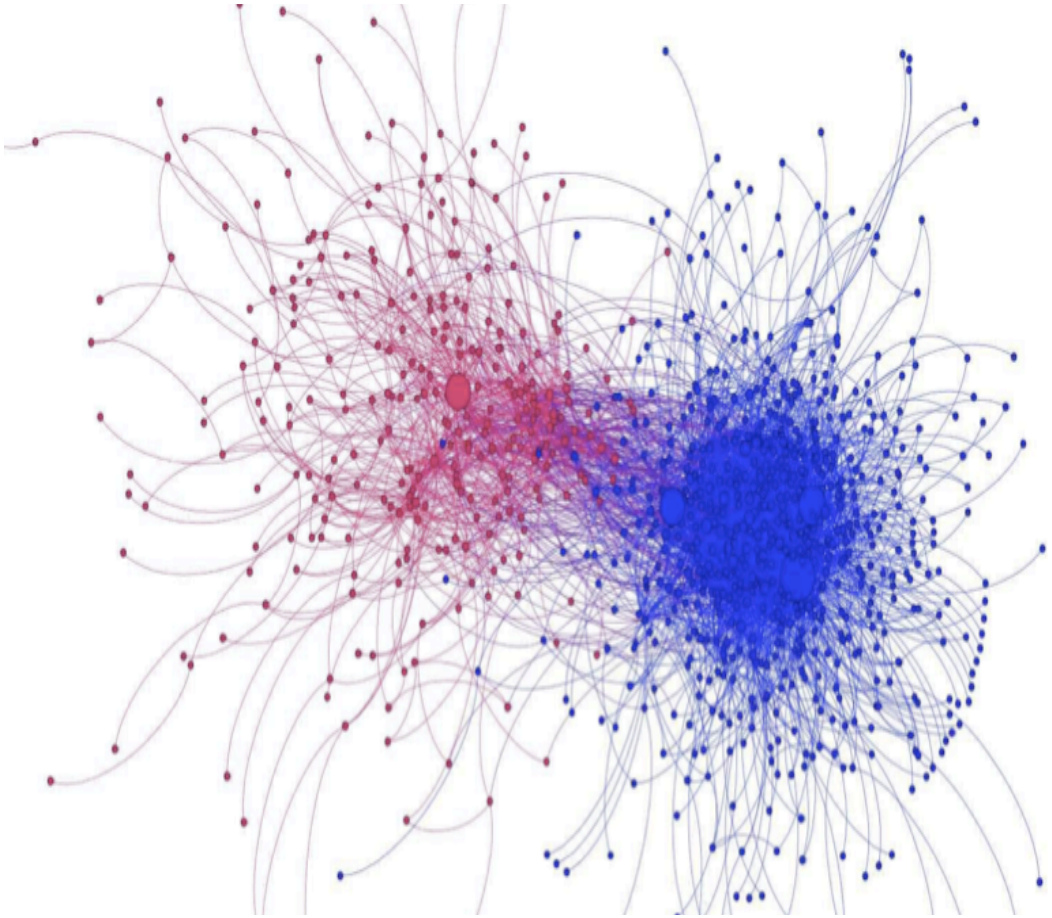
Ira S. Rubinstein

NYU School of Law
Information Law Institute
and

Woodrow Hartzog

Samford University's Cumberland School of Law

The Re-Identification Debate



1. Anecdotes Drive the Re-ID Narrative
2. The Scholarship is Divided
3. …and Highly Technical
4. The Deeper Debate is Between Pragmatists and Formalists

Argument by Anecdote: Two Not Very Troubling Anecdotes

- Exposing Governor Weld's Medical Records
  - Classic linkage attack
  - Barth-Jones: Unfair advantages?
  - Implications?
- Identifying AOL Searcher No. 4417749
  - AOL released search data for research purposes; replaced identifying info with unique ID
  - This proved to be weak protection (duh!)
  - Implications?

More Troubling Anecdotes

- Breaking the Anonymity of the Netflix Dataset
  - Public release of dataset for algorithm contest
  - N&S developed an algorithm for de-identifying Netflix data by comparing it with a publicly available dataset of movie ratings (IMDb)
  - Major technical breakthroughs:

    (1) Determined that less "background" info needed than previously thought and (2) developed a "robust" algorithm
  - Implications?
- Genetic Privacy Breaches

  [under development]

Divided Scholarship

Paul Ohm

Jane Bamberger

Felix Wu

Highly Technical Arguments

- The re-ID controversy sits on top of a more fundamental technical debate between statistics and computer science
- The two communities overlap but in the popular treatment that filters out to legal scholars, advocates of the two sides occupy black and white positions and remain very far apart. Why?
- (And is it turtles all the way down?)

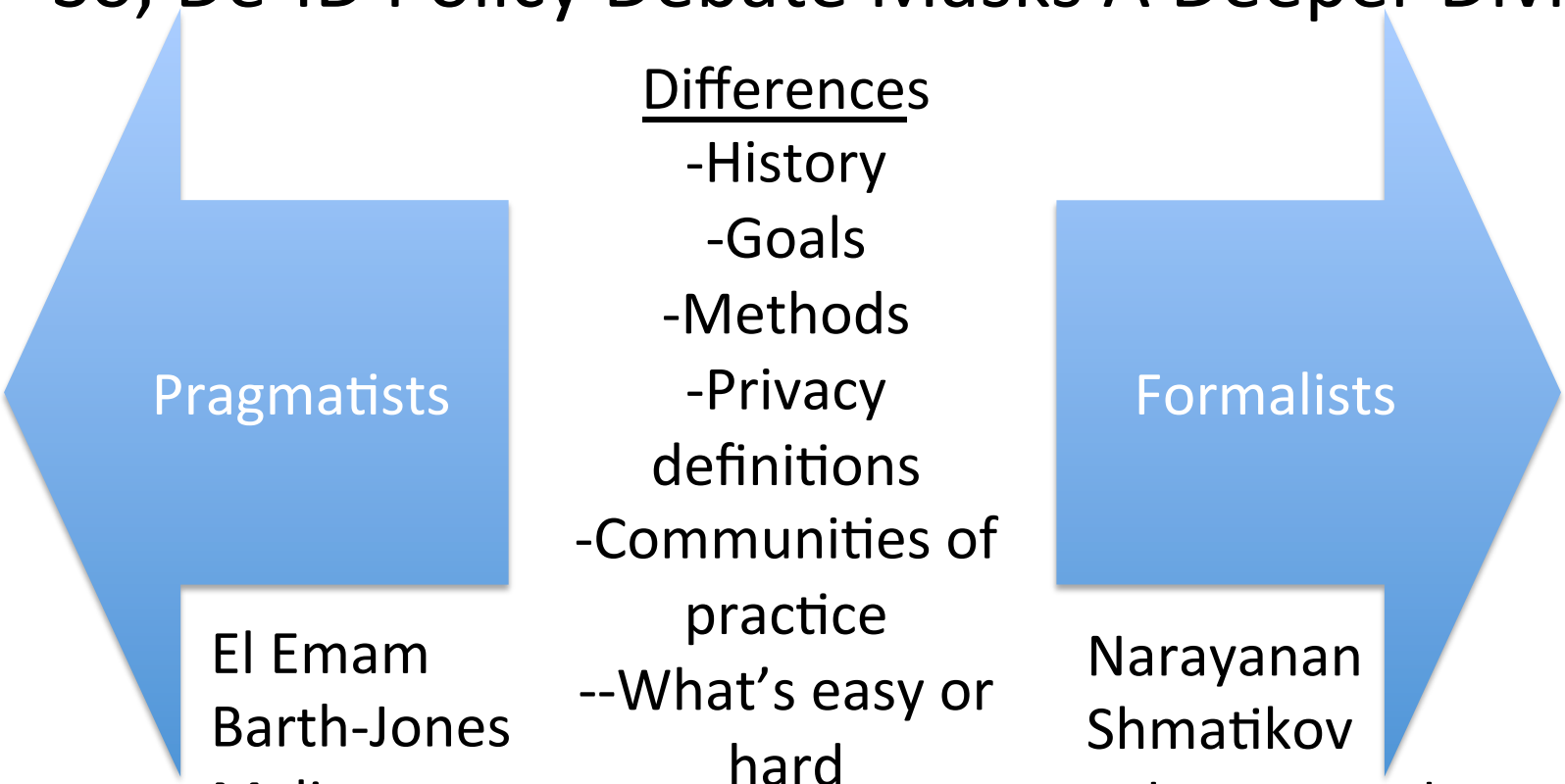
Statistical Disclosure Limitations

- Focuses on privacy & confidentiality of data collected into statistical databases for research purposes
- Goals:
  - Preserve confidentiality and provide access to useful statistical data
  - *Balance* data utility and disclosure risk
- Not a primary goal:
  - Mathematical rigor in defining privacy, modeling adversaries, and quantifying re-ID probabilities
  - Guarantees of confidentiality or privacy notwithstanding availability of background information

DP and Other Formalist Approaches

- Goal: Place privacy/confidentiality on a *mathematically rigorous* foundation
- Dwork:
  - Absolute privacy protection impossible due to background knowledge
  - But DP guarantees that "almost, and quantifiably, no risk [to an individual] is incurred by joining a statistical database"
  - This *provable* privacy guarantee is independent of the availability of background information
- Other approaches with a formalistic "mindset" include homomorphic encryption, privacy-preserving data publishing, and secure multi-party computation

So, De-ID Policy Debate Masks A Deeper Divide



Differences

-History

-Goals

-Methods

-Privacy definitions

-Communities of practice

--What's easy or hard

-Measures of "success"

-Access to data/ views on open data

El Emam
Barth-Jones
Malin

Question for formalists:
Is it practical?

Narayanan
Shmatikov
Felten, Dwork

Question for pragmatists?
Is it provable?

This leave us with a dilemma (because the background info. problem is not going away!)

- What To Do?
  - Ignore this problem? Restrict types of data? Suspend public releases of de-IDed data absent formal privacy guarantees?
- There is also a "Dworkian" dilemma
  - Is SDL "privacy-supportive" but lacking "rigorous definitions of privacy and modeling of the adversary" or an instance of "the sanitation pipedream"?
  - Granted, DP is "mathematically rigorous" but much useful research today relies on SDL.
  - So what to do? Suspend use of SDL and rely solely on DP and other formal methods?

A New Direction: Re-Frame the Policy Debate

- Build on the *three* major forms of interaction between researchers and data

  [see Model for User-Data Interaction or slides 19 and 20]

  - Mode 1: Direct access (walled garden)
  - Mode 2: Dissemination-based access (public release)
  - Mode 3: Query-based access (trusted curator)
- Develop a new, security-based approach that
  - Reflects strengths & weaknesses of all *three* modes
  - Takes advantage of all available technical AND legal tools (including statutory regulation, contract law and criminal law)

Towards a Security-Based Approach

- The formalists marginalize risk
- The pragmatists over-leverage risk
- Data security offers a better framework:
  - Tolerates a certain amount of risk but does not make risk the trigger for protection.
  - Any non-publicly disclosed database presents a classic *security* problem
  - Security allows an intellectual shift from magic bullets to security *processes.*

Applying Security Processes to Anonymization

Data security law focuses on 4 main processes:

1. Asset and risk identification
2. Data minimization
3. Technical, physical, and administrative/ procedural safeguards
4. Response plans

How This Changes the Debate

- Data security policy almost uniformly relies on a reasonableness standard
  - NB: Technical standards like NIST 800-53 are crucial
- This allows organizations to institute data security practices in their own setting, rather than rely on a "one size fits all" standard
- Our proposal emphasizes "reasonable" de-identification procedures, not only in practice but in re-framing how we **talk** about the de-identification which also matters:
  - Explains why we're stuck in the present debate
  - Allows us to shift the debate towards a better policy resolution
  - BTW: The FTC gets both points

Necessary Legal Reforms: Existing Laws

- Transition from PII to "PII 2.5"
  - How to demarcate 3 sub-categories?
  - Which FIPs apply to "identifiable data"?
- Regulate the process of anonymization as applied to *all* data by establishing a "reasonableness" standard
  - In theory, the FTC could encourage the development of such standards and enforce them
  - Will the FTC also extend its unfairness jurisdiction to "bad anonymization"?
- Modify HIPAA De-ID Rule
  - Replace safe harbor method with a "reasonable" security standard
  - Incentivize expert determinations

Necessary Legal Reforms: New Laws

- Enact a baseline privacy law
- Limit or bar the use of "release-and-forget" anonymization
- Incentivize Data Use Agreements
- Make disclosers and recipients of de-IDed data accountable for data processing
  - Regulate unauthorized re-identification of de-identified datasets
  - Enforce new rules via civil and criminal penalties and civil recovery in damages
  - see Robert Gellman, The Deidentification Dilemma: A Legislative and Contractual Proposal (2011)

What About Open Data?

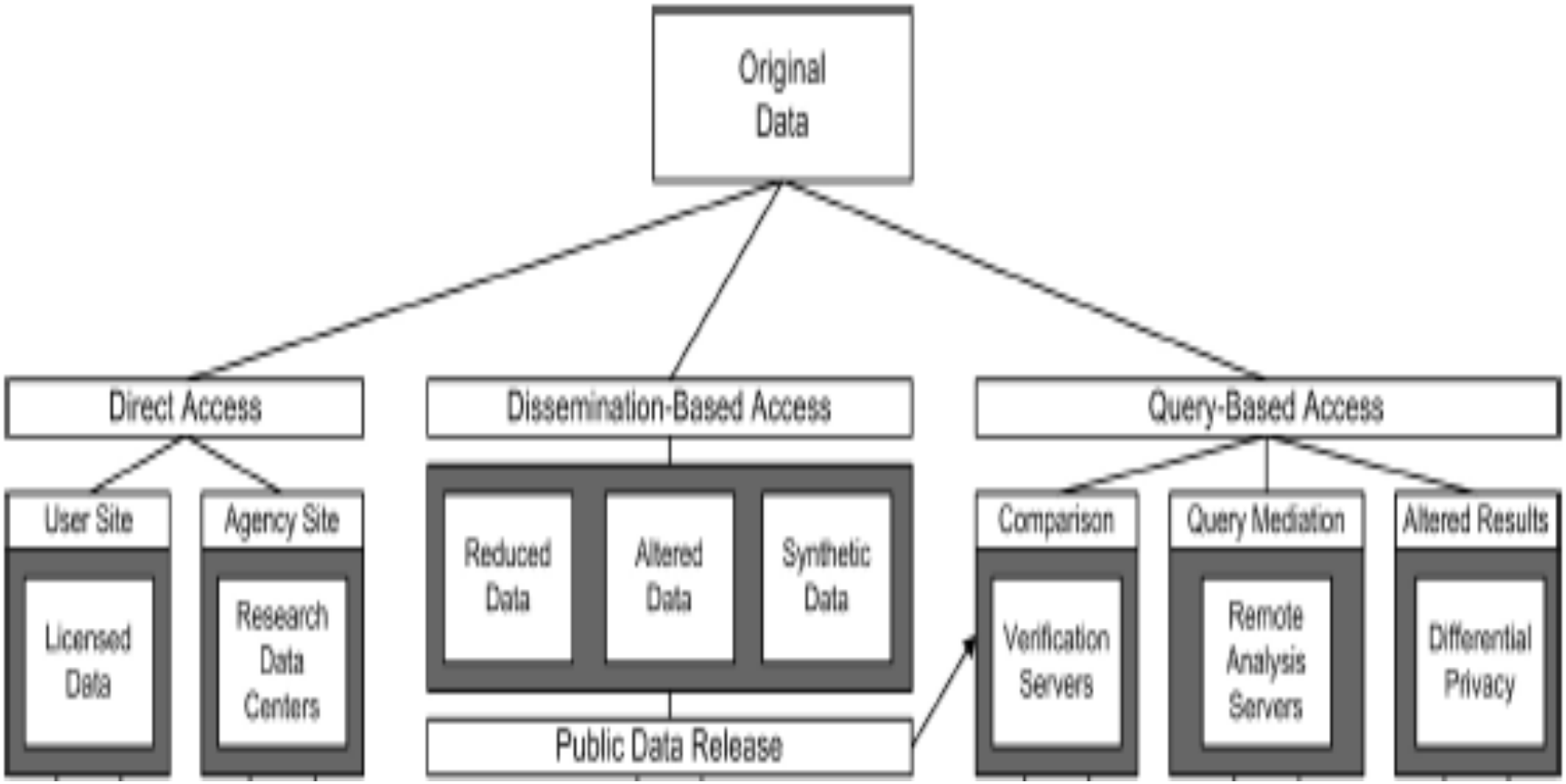
- Benefits of "open data"
  - Helps ensure accountability in research by allowing others access to researchers' data;
  - Allows researchers to build on the work of others more efficiently and helps to speed the progress of science;
  - Facilitates trust between researchers and with the public;
  - Allows for secondary analyses that expand the usefulness of datasets and the resulting knowledge gained;
  - Lowers burden on research participants through the reuse of existing research data and the decrease in the cost of data
- Is it possible to implement our proposal without undermining these benefits of open data, perhaps by finding functional equivalents to openness?

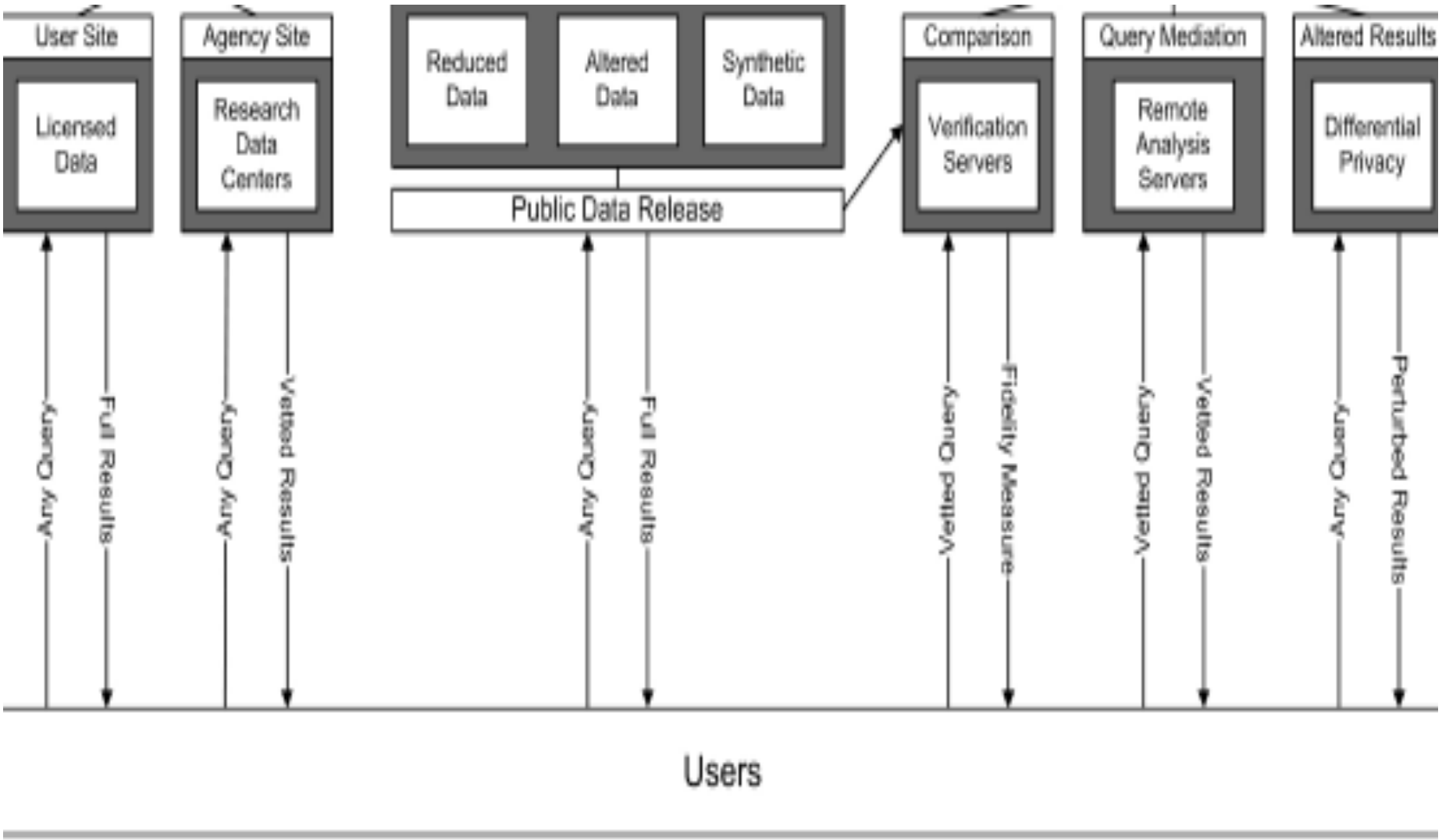
Risk-Based Anonymization in Practice: Revisiting the Anecdotes

[under development]

- Weld
- AOL
- Netflix
- Genetic Privacy Breaches
  - [But mention NIH policy of "tier-based access"]

Model for User-Data Interaction




Direct
Dissemination-Based
Query-Based
User Site
Licensed Data
Agency Site
Research Data Centers
Reduced Data
Altered Data
Synthetic Data
Public Data Release
Comparison
Verification Servers
Query Mediation
Remote Analysis Servers
Altered Results
Differential Privacy
Any Query
Full Results
Any Query
Vetted Results
Any Query
Full Results
Vetted Query
Fidelity Measure
Vetted Query
Vetted Results
Any Query
Perturbed Results
Users