Data Protection and Dissemination: Discussant Remarks

Matt Williams

09/26/24

Future Directions for Social and Behavioral Science Methodologies in the Next Decade: A Workshop

RTI INTERNATIONAL

Remarks

- Highlight some themes across the presentations
- Pose some questions for presenters and participants
- Yield remaining time to the discussion

Themes

- High Dimensionality and Complexity (data and products)
- Complex Pipelines (data collection, integration, analysis)
- Proliferation of variations of (formal) privacy
- Critical Role of Purposeful Scientific Data Collections (e.g. surveys)

High Dimensionality and Complexity of Data

- Proliferation of complex measurements
  - Structured data (paper forms)
  - Unstructured (open text, sensors, scraped data, audio, video, GPS, etc)
- Complex structures and collections of measurements
  - (e.g. electronic health records)
  - Relational databases instead of flat/rectangular files
- Curse of Dimensionality
  - High Dimensional Space is Sparse – even when we have LOTS of data.
  - Every individual is unique. High dimensional summary is impossible.
  - Consider lower dimensional projections/views of the data (many, many of them)
  - Prof. Chen example - propensity as a one-dimensional summary from linked admin data.

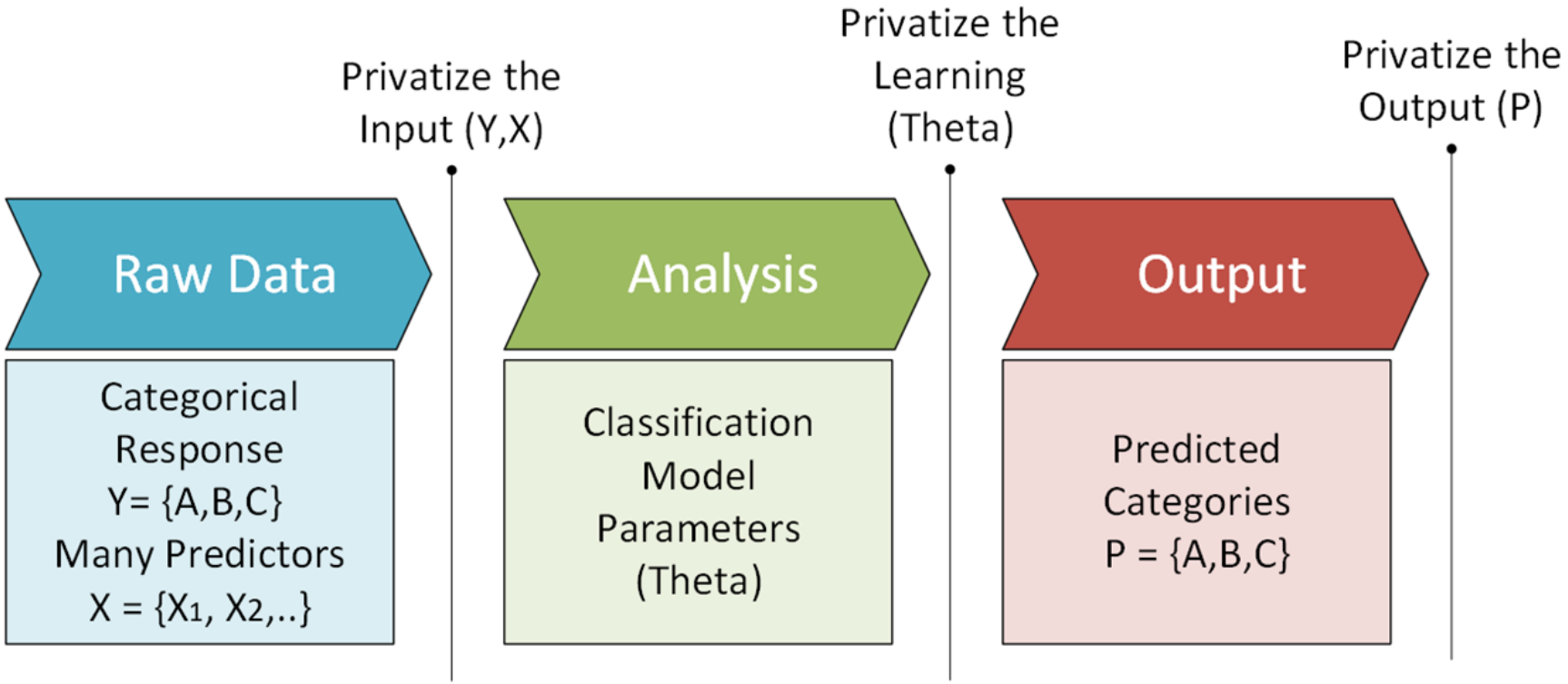
High Dimensionality and Complexity of Products

- Statistical agencies: Mandated to maximize use of current data.
  - Encouraged to link and integrate multiple data sources (increasing complexity).
- Tracking data usage once it is released is its own industry
  - Search engine /page view analytics
  - Bibliometrics – traditionally – scholarly/peer reviewed
  - Policy/Legal uses (federal, state, local)
- Statistical Agencies – diversifying dissemination methods
  - Tabular data and table generators (APIs)
  - Interactive Visualizations (e.g. maps)
  - Charts and Dashboards
  - Research Data Centers – bibliographies of uses

High Dimensionality and Complexity of Data (Products)

- The legal mandates for statistical agencies to collect data don't enumerate all data they should collect and the products they should release.
  - **Which** frameworks and structures can help prioritize all these (competing) uses?
- We know that collapsing across dimensions leads to confounding – for example Simpson's paradox or the modifiable areal unit problem.
  - **What** frameworks exist for identifying the appropriateness/risk of using low dimensional/marginal distributions? (e.g. age age-adjusted mortality rates instead of single year of age rates)
  - There is a trade-off between intersectionality research and personalized interventions with privacy concerns due to unique/rare values.

Complex Pipelines (Data Analysis)



Complex Pipelines (Upstream Processes “Raw”)

- We can randomize/protect at the point of data collection (randomized response), but it is less efficient that protecting learning or output
- The supply chain of collection and processing to get “raw” data is complicated and uses individual information multiple times.
  - Drechsler, J., & Bailie, J. (2024). The Complexities of Differential Privacy for Survey Data (No. w32905). National Bureau of Economic Research.
- **How** can we track the “entire” influence an any given individual on the final data set (sample design, non-response, edits, imputation, weighting adjustments)?

Complex Pipelines (Downstream Processes “Output”)

- Proliferation of Uses
  - Integrated Data Products (e.g. probability and non-probability data)
  - Benchmarking other studies (e.g. denominator for mortality rates)
- Is it really possible to “future proof”?
  - **How** can we anticipate all uses?
    - DP proponents – privacy protections can be future proof
    - What about utility?
  - **Can** we consider reducing privacy over time – (1950 census vs. 2030 census) Does data need the same privacy protection forever?
  - **Are there** future disruptive technologies (e.g. quantum computing) that would accelerate learning/inference attacks?

Proliferation of formulations of (formal) privacy

- Traditional Statistical Disclosure Control (SDC)
  - Diverse set of approaches – tend to focus on the realized product.
- Differential Privacy
  - Emphasis on the process not the product. Complementary to SDC.
  - Can it unify/replace a diverse set of privacy risk approaches and measures?
  - Prof. Gong: variations on different “design choices” leading to inclusion and exclusion of what is protected and at what level of `guarantee’
  - **How** can we expand tools to convert from one DP currency to another? (Renyi DP to Epsilon, Delta) or are these variations fundamentally incompatible?

Proliferation of formulations of (formal) privacy

- Personal (Entity) view of privacy
  - Context and population specific
  - Notions of harm, legal rights, and societal imperatives
  - **How** best to algin SDC and DP with these societal definitions?
  - Do people (entities) actually make rational choices based balancing risk?
- Equity
  - Is it equitable for everyone (entity) to get the same level of protection?
  - **How** can we capture and integrate a plurality of privacy-utility trade-offs?

Critical Role of Purposeful Scientific Data Collections (e.g. surveys)

- In spite of declining response rates, probability-based sample surveys are critical (Profs. Shlomo and Chen)
  - They allow us to mitigate representation/selection bias for non-probability data
  - **How** can we increase our investment in probability surveys?
    - The value of probability surveys is amplified, not diminished by integrating with non-probability collections
    - Is it better to reduce our probability sample sizes and `double-down' on higher response rates, knowing we'll integrate with non-probability data?
  - **Are** (new) mandatory data collections possible (Prof. Shlomo)?
  - **What** is the right incentive for participation (cash, tax credit/fines, community grants\benefits)?

Summary - Speculation

- Are we too reliant on statistical tools to rescue us?
  - (Prof Gong: SIPP and ACS – technology has not advanced enough)
  - **What** policy and societal levers are available?
- We focus on preventing all possible privacy exposures.
  - Low-trust frameworks are inefficient. (Loss of effective sample size)
  - Is this an impossible burden on a statistical agency?
- **How** can we advance shared accountability frameworks for privacy/appropriate use of data?
  - **How** can we incentivize reporting of inappropriate use (whistle-blowing)?
  - **How** can we evaluate policies requiring algorithmic transparency/appropriate use of data – hiring, insurance claims, loan applications, etc?

Thank you

Contact: Matt Williams| email: mrwilliams@rti.org