ACDEB
Advisory Committee on Data for Evidence Building

Meeting 6 Notes and Actions
March 19, 2021

Next Meeting: April 23, 2021 (9 AM - Noon (EDT) -- Scientific Information, Public Comments Initial Review, Data for Evidence and Infrastructure Framework, Requirements Definition Next Steps

Meeting Agenda:

1. Privacy & Confidentiality Concepts – Charles Cutshall, Shawn Davis
2. Privacy & Confidentiality Technologies – Michael Hawes, Mayank Varia, Len Burman
3. Data Ethics – Ted Kwartler, Haniyeh Mahmoudian
4. Executive Order Discussion

I. Committee Updates

The Chair provided key staffing updates including announcing Alyssa Holdren as the new Designated Federal Officer, Avi Alpert as the new rapporteur, and Emilda Rivers as the new Chair of the Committee.

II. Privacy & Confidentiality Concepts -- Charles Cutshall (Commodity Futures Trading Commission), Shawn Davis (Edelson PC)

The first presenter, Charles Cutshall, provided a history of data privacy concepts in government dating back to the HEW report in 1973 and the subsequent Privacy Act of 1974 which codified the report's principles into law. He noted that ACDEB can build on the previous work and set a foundation for future thinking on how to expand access while respecting privacy and maintaining trust in government.

The second presenter, Shawn Davis, provided an overview of how individuals can be identified through data and described how maintaining data security can be seen as a balancing act between confidentiality, data integrity, and data availability. He described the different types of direct and indirect identifiers and gave examples of the disparate agency interpretations of deidentification. Finally, he provided examples of how linkages across data, particularly when geolocation data is included, increases reidentification risk.

(See presentation at bea.gov/evidence under Meetings tab.)

Committee Feedback/Discussions:

→ **Privacy is Contextual Not Absolute:**

- Privacy is not binary but exists on a continuum; it has multiple tiers for which different groups can be given different levels of privileges.
- The use of tiered access for different purposes, with different value and different security risk, can help in balancing the trade-offs.

- These tiers should look beyond just government and research access and consider industry access as well.
- A data inventory is critical to understanding all the metadata out there, and particular focus should be on reducing or eliminating public release of most risky metadata such as geolocation.

→ **Bridging Different Standards Domains:**

- The Federated nature of U.S. governance means the Committee needs to consider how the impact across various standard setters (such as states).
- For OMB, the challenge is drafting guidance that can be applied to all agencies without being too broad thus mitigating some of the disparate definitions used today.
- OMB should leverage its budgetary function to impact data privacy technology development and investment across agencies.

III. Privacy & Confidentiality Technologies -- Michael Hawes (Census Bureau), Mayank Varia (Boston University), Len Burman (Urban Institute's Tax Policy Center / Syracuse University)

The first presenter, Michael Hawes, noted that every data release contains some degree of private information leakage and introduced the 'triple constraint of data': Confidentiality, Access, and Accuracy. He described how emerging technologies, such as synthetic data, query systems, secure multiparty computation, and differential privacy, particularly when used in combination, can help mitigate the risks. He noted that none of these technologies provide off the shelf solutions today and require proper implementation and continued research and development. Finally, he provided a case example from the Census Bureau and the use of differential privacy to protect 2020 Census data.

The second presenter, Mayank Varia, introduced Secure Multi-Party Computation (SMC), which allows parties to aggregate data sets without transferring or revealing private data to each other or a third party. He provided a case example from the Boston Women's Workforce Council. Finally, he raised the law and policy questions that need to be answered to use this solution more broadly with government data.

The third presenter, Len Burman, provided more detail on the emerging technologies that enable increased data release while maintaining privacy including: Fully Synthetic Data, Validation Servers, and SMC. He noted how Fully Synthetic Data can be safe-to-release with no restrictions since it is impervious to linkage attacks and could be used as a training dataset for the Validation Server which allows researchers to run statistical programs on synthetic data. He described SMC as a possible 'game changer' because it could provide data matching across datasets without ever combining the datasets. Finally, while recognizing that the application of these technologies still faces real challenges, he noted that NSDS can play a role in spurring

innovation that serves both agencies and data users in producing timely, manageable datasets that can be used to produce high-quality evidence.

(See presentation at bea.gov/evidence under Meetings tab.)

Committee Feedback/Discussions:

→ **Coordinating Data Releases:**
- Today different agencies have different rules for releasing data, and even within agencies data releases may not be coordinated.
- This can increase re-identification risk due to data linkages.
- NSDS can play a role both in developing standard disclosure risk assessments and acceptable risk levels and serving as release coordinator and central repository for linked releases.

→ **Technology's Double-Edged Sword:**
- The technologies such as AI and machine computing that are making data privacy more challenging can also serve to make it more secure.
- There is an opportunity to transform how agencies think about privacy in line with implementing new technologies.
- Further, there is an opportunity to help researchers think about data in new ways such as when it may be appropriate to substitute microdata with synthetic or aggregated data.

IV. Data Ethics – Ted Kwartler (DataRobot, Inc.), Haniyeh Mahmoudian, (DataRobot, Inc.)

The presenters, Ted Kwartler and Haniyeh Mahmoudian, provided a definition of data ethics as responsible and sustainable data collection, processes, and storage practices as well as the ethical use of data. They expanded on this definition for the use of AI and Machine Learning in data analysis. They recognized that AI has great power to help the world, but it must be done in an ethical, explainable, and equitable manner to realize its potential. They used examples to illustrate how it is challenging to determine fairness in the application of AI for data-based decision-making, raising considerations such as whether it is fair to target sub-group outcomes or error rates. Finally, they laid out the implications of this for the Committee as it considers recommendations regarding a Data Ethics Framework with an emphasis on any outcome being explainable, equitable, and accessible.

(See presentation at bea.gov/evidence under Meetings tab.)

Committee Feedback/Discussions:

- **NSDS' Role in Data Ethics, If Any:**
  - Is NSDS downstream of the data ethics decisions?
  - Should NSDS be tasked with developing guardrails for ethical use of data and how would this work in coordination with Agencies?

V. Executive Order Discussion

The Chair facilitated a discussion on the impact of various Executive Orders (EO) on the Committee's work. It was noted that several EOs are quite relevant such as those on Scientific Integrity, Equity, and Covid-19 Data. The NSDS can serve an important role in accomplishing the goals of these EOs as they require linking and sharing administrative data across agencies in a secure platform. It is important that this Committee supports the working groups implementing the EOs and highlights considerations that were discussed today such as balancing privacy and specificity in linked data.

V. Conclusion

The Committee closed by highlighting the planned topics to cover in the next information gathering meeting.

April
Data Challenges in the Physical Sciences (Otis Brown)
Using the Hart/Potok Report Framework as an Organizing Principle (Nick Hart)
Public Comment Discussion