Federal Register Notice: 89 FR 51554, Federal Register :: Networking and Information Technology Research and Development Request for Information on Digital Twins Research and Development, June 18, 2024.

Request for Information on the National Digital Twins R&D Strategic Plan

Dr. Ivo Dinov and Dr. Brian Athey

July 28, 2024

RFI Response: Digital Twins R&D Plan
Digital Twins R&D Strategic Plan Recommendations

by Ivo D. Dinov and Brian D. Athey, University of Michigan – Ann Arbor

Outline

This RFI Digital Twins R&D Strategic Plan response is focused on leveraging the advanced capabilities of modern AI services, tools for statistical obfuscation of sensitive data (e.g., DataSifter), and techniques for quantifying risks of deidentification (e.g., ϵ −differential privacy) to create, manage, and utilize synthetic digital-twin pairs for biomedical and health datasets. The primary goal is emphasizing the utility of desensitizing, sharing, aggregating, and performing AI-driven analysis on massive (human-identifiable) datasets without compromising participant identifiable human information. Integrating these technologies will facilitate powerful research innovation (preserving the information energy in the biomedical data) while ensuring data privacy and security. Generative Artificial Intelligence enabled Large Language Model platforms offer new opportunities to address these opportunities. This should reduce the barrier of utilizing personal health and other private information resources for research and for legitimate business uses currently hosted in overly protected IT environments built to be resistant to hackers and ransomware attacks.

Background

Digital twins, virtual representations of physical entities, have transformative potential across various domains, including biomedical research and healthcare. This RFI Digital Twins R&D Strategic Plan response outlines the development and implementation of a digital twin R&D framework that integrates foundational AI models and various techniques for data desensitization to generate realistic, representative, and safe (and synthetic) 'digital-twinned' versions of heterogeneous data, without compromising the fidelity or release of personal information. The focus is on creating synthetic digital-twin pairs to facilitate safe and effective research, data sharing, and AI-driven interrogation of biomedical datasets.

Goals & Objectives

1) *Enhance Data Privacy and Security*: Utilize epsilon-differential privacy to quantify risk of leaking personal information and ensure that synthetic digital twins do not contain identifiable human information.
2) *Enable Safe Data Sharing and Aggregation*: Develop protocols and frameworks that allow full-spectrum control of the balance between privacy-protection (security) and value (utility) of the desensitized digital twin computable objects. This would allow data governors (or Honest Brokers) to selectively dial up or down the level of privacy-protection and promote secure sharing and aggregation of biomedical datasets using DataSifting.
3) *Advance AI-Driven Biomedical and Healthcare Research*: Leverage contemporary model-based statistical techniques and model-free AI computational tools to perform advanced

AI-driven analytics on the synthetic digital twin objects and quantitatively compare the results to their counterparts computed on the native raw datasets.

4) *Provide a New Means to Study Bias in AI systems*: Systematic bias introduced by flawed training sets can be addressed by utilizing these Digital Twin methodologies to create synthetic populations of Digital Twins that can uncover bias while preserving privacy.
5) *Foster Interdisciplinary Collaboration*: Encourage collaboration between data scientists, healthcare professionals, and researchers to maximize the impact of digital twin technologies.

Key Elements

1) Existing and upcoming model-based statistical methods and model-free AI algorithms, accelerated by the proliferation of Generative AI platforms are the substrate of this opportunity. There is a wealth of offline tools, online services (including Cloud-services), and computational resources for statistical analysis, visualization, dynamic interrogation and computational modeling of both raw data and their corresponding derived digital twins. These resources form the backbone of the dynamic digital twin ecosphere where core data desensitization, synthetic data generation, and data analytics will take place.
2) DataSifting is a rigorous statistical obfuscation process designed to desensitize data elements, anonymize personal information, and synthetically generate realistic versions of digital twins that are complete simulated data archives matching the data type, characteristics, interdependencies, and features of the original (sensitive) data into the digital twins as computable simulated data objects. This sifting process controls the delicate balance between preserving the value (energy and utility) of the data while protecting the sensitive information (risk-reduction). DataSifting ensures that the synthetic digital twins generated from raw biomedical datasets are free from identifiable information, facilitating secure data sharing and aggregation. Use of Generative AI platform API functionality and deep prompt engineering can industrialize this process to create cohorts and populations of safe and secure synthetic digital twins.
3) Epsilon differential privacy (ϵ −DP) is a rigorous mathematical framework that ensures individual privacy in data analysis. By quantifying the risk of reidentification of sensitive information, the ϵ −DP framework provides strong privacy guarantees for the synthetic digital twins, enabling their use in sensitive biomedical research without compromising personal information.

Implementation Plan

1) Phase 1: Infrastructure Development
    i. Develop/deploy the computational infrastructure required to support the creation and analysis of synthetic digital twins.
    ii. Integrate AI services with DataSifter, and ϵ −DP to generate an end-to-end comprehensive (modular) platform.
    iii. Identify datasets and establish data pipelines for ingesting raw biomedical datasets and generating synthetic digital twins.
2) Phase 2: Digital Twin Pilot: Data Synthesis and Desensitization

i. Use DataSifter to desensitize raw biomedical datasets and compute a range of quantitative metrics quantizing the privacy-protection (risk) vs. data value (utility) retained in the digital twin objects.
ii. Generate synthetic digital twins using epsilon differential privacy and quantify the level of privacy guarantees.

3) Phase 3: AI-Driven Analytics
   i. Utilize existing computational tools to perform AI-driven analysis, powered by Generative AI platforms, on the derived synthetic digital twins and contrast these to their counterparts computed on the original raw data.
   ii. Develop and implement (novel) AI models for various biomedical research applications, such as disease progression modeling, treatment efficacy analysis, and patient outcome prediction.
4) Phase 4: Validation and Testing
   i. Conduct extensive validation and testing to ensure the accuracy and utility of synthetic digital twins.
   ii. Compare the results of AI-driven analysis on synthetic digital twins with those obtained from raw datasets to assess the fidelity of the synthetic data.
   iii. Systematically study bias in existing data sets leveraging social determinants of health (SDOH) and health outcomes knowledgebases.
5) Phase 5: Collaboration and Dissemination
   i. Foster interdisciplinary collaboration by organizing workshops, seminars, and collaborative research projects. Partner with NIH, NSF, DARPA, and industry participants, working with national consortia such as the Coalition for Health AI (CHAI) and other such organizations.
   ii. Disseminate findings through academic publications, conferences, and online platforms.

Expected Outcomes

1) *Enhanced Privacy and Security*: Robust privacy guarantees for biomedical datasets, enabling secure data sharing and aggregation.
2) *Increased Research Capabilities*: Advanced AI-driven analysis on synthetic digital twins, facilitating novel biomedical research and innovation.
3) *Broader Collaboration*: Increased collaboration between data scientists, healthcare professionals, researchers, and industry leading to more comprehensive and impactful research outcomes.

Challenges and Mitigation Strategies

1) *Data Privacy Concerns*
   Mitigation: Utilize epsilon differential privacy and DataSifter to ensure strong privacy guarantees for synthetic digital twins. Quantify the quality of the digital twin computational objects by computing a wealth of numerical metrics capturing the privacy protection (reidentification risk) and data value (digital twin utility).
2) *Computational Complexity*

Mitigation: Leverage existing AI computational resources and optimize data processing pipelines to handle large-scale data synthesis and analysis efficiently.

3) *Interdisciplinary Collaboration*
   Mitigation: Organize regular workshops, seminars, and collaborative projects to foster communication and collaboration between different stakeholders.

The integration of modern Generative AI platform services, DataSifting, and epsilon differential privacy algorithms into a comprehensive digital twin R&D framework holds immense potential for advancing future biomedical research and healthcare. This process offers enormous risk mitigation, cost-savings, increased ROI, and opportunities to expedite R&D and translation of basic data-STEM discoveries into biomedical and health applications. By creating synthetic digital-twin pairs that preserve privacy and enable advanced AI-driven analysis, we can unlock new opportunities for innovation while ensuring the security and confidentiality of sensitive data. This strategic plan outlines one clear and actionable roadmap to achieve these goals and drive transformative advancements in the field.