ETM
ENGINEERING & TECHNOLOGY MANAGEMENT

“Big Data and Privacy”

Course: Science and Technology Policy ETM 575
Term: Winter 2016
Instructor: Prof. Elizabeth Gibson, PhD

Author: Team 1
Maoloud Dabab and Rebecca Craven

Table of Contents:

Abstract:

Big data poses serious challenges to the privacy and security of individuals and their data. This research considers how to best address the social problem that the pervasiveness of data collection, analysis, and storage creates with regard to individuals' ability to control their own data. Using Quality Function Deployment (QFD) and Technology Roadmapping analysis methods, we assess the social problems, technologies, resources, and industries that are most relevant to addressing data privacy. We find that the medical industry is the most important industry to consider because of the nature of the data generated through medical processes and technologies, and that enforcement mechanisms, specifically in the form of federal enforcement agencies, are the most effective way to ensure compliance by actors. However, there are extenuating political circumstances and increased costs that make the implementation of policies difficult in the United States that also need to be considered. Future research should further address some of these elements as big data technologies continue to be adopted across corporations and organizations.

Introduction:

Big data does not have a single agreed upon definition in the literature, but is most often characterized by its volume, velocity, variety, variability and complexity [10]. Unlike analog data, this digital native data allows faster access, processing and analysis in greater quantities previously possible. Big data depends on four main steps of data processing: collection, storage, analysis, and use of big data. The technologies used in these steps are no longer novel and emergent; Gartner has not included big data from its annual hype cycle report since 2015 because big data has "gone mainstream" [11]. Big data technology depends on continuous streams of data, generated by individuals that may or may not be aware of their data's place in datasets or analysts' hands. The concerns over the privacy and security of individuals' data echo those of previous eras of non-digital data, which are at this point well-legislated and litigated in the American system [1]. As such, the PCAST report on big data and privacy notes that "…it is the use of data (including born-digital or born-analog data and the products of data fusion and analysis) that is the locus where consequences are produced" (xii). This report considers solutions to social problems that arise from the increasing ubiquity of big data. We proceed by first determining the policy background and social problems related to privacy and big data, and then analyzing potential solutions to these problems through House of Quality and Technology Road Mapping analysis.

Social Problems:

As big data increasingly becomes a part of every industry and every sector, it makes new solutions to a plethora of social problems possible, from improving food security to preventing human trafficking [22]. However, it also brings new social problems with it. These social problems can be grouped into five general categories: privacy and security, data reuse, data accuracy, data access, and archiving and preservation [1]. Each of these social problem areas represent a point in the big data process at which social externalities can occur, and are discussed in both the PCAST report and in the broader big data literature. Privacy and security refers primarily to the initial generation of data by individuals, including secondary generation through association with other existing data sets. Data reuse, by contrast, is concerned with the repurposing of data from its intended recipients and processes for other uses. Data accuracy issues can arise when multiple data sources with differing controls and verification processes impact the overall quality of the data, and the degree to which the data is correct. Data access concerns the individuals and organizations that have access to any data that is part of the big data process, including the archiving and preservation of data, which refers to the historical cataloging of data once its initial use has passed. In all of these cases, individuals that generate data have little to no control over that data once it comes into digital existence unless technology processes are constrained by social and political processes.

In the literature, concerns about big data privacy has been specifically singled out for additional consideration by industry, government, and research consortia [2]. Many aspects of technology have been party to the challenges to the bounds of the American right to privacy [1]. However, others have noted that information technology poses privacy issues that are unique to digital-native data [9], and that big data is particularly pertinent to technology privacy discussions [8]. Because big data has implications for both public and private uses, there are especially large implications for government if policies are not sufficient to address privacy issues with big data [7], especially in light of the historical precedent favoring the individual's right to privacy in the United States. Thus, while many aspects of big data pose potential risks and social problems, the issues concerning data privacy and the protection of individual data generators are the social problems that are focused on for the remainder of this report.

Policy background:

The fully policy background on big data in the United States includes aspects of cybersecurity and privacy policies that are not necessarily specific to big data. Cybersecurity-specific policies are particularly well documented in the literature [5]. Additionally, much of the analysis of big data is from a legal, and not technical, perspective [3]. As such, policy infrastructure primarily refers to existing privacy protection policies that are in place and may not deal with digital privacy at all. In general, ingrained policies and protections are neither big data nor even technology specific. In the United States, policies regarding privacy and big data can be implemented at two levels: federal and state. Treatments of US big data policy generally show a lack of coherent federal policy that crosses sectors [6] [3], with most policies specifically focusing on particular sectors like healthcare, education, and financial institutions. Data policies are contained within broader privacy legislation like HIPAA, FERPA, and FCRA for those sectors respectively. There are some precedents for state-level policies regarding data protection and privacy issues in the United States. California is especially active in this regard, with legislation like the California Online Privacy Protection Act (CalOPPA) addressing digital data privacy [12]. This process of distinguishing between states and has the effect of further fragmenting data privacy policies and their reach. Just as there are no comprehensive big data regulations at the federal level, there are also no bodies tasked specifically with big data compliance, monitoring, or enforcement. Legislation concerning data privacy is full of recommendations and self-enforcement requirements, but little in the way of coercive inducements to follow the guidelines. There are federal agencies that are engaged in big data research in cooperation with private industry (OSTP and NITRD are particularly important in this respect), though the research is actually carried out by legislatively-created agencies that have much broader missions (NASA, EPA, NOAA, etc. all fit this characterization). Thus, there are a multitude of agencies and laws that impact data privacy, though the effects are inconsistent across jurisdictions and industries.

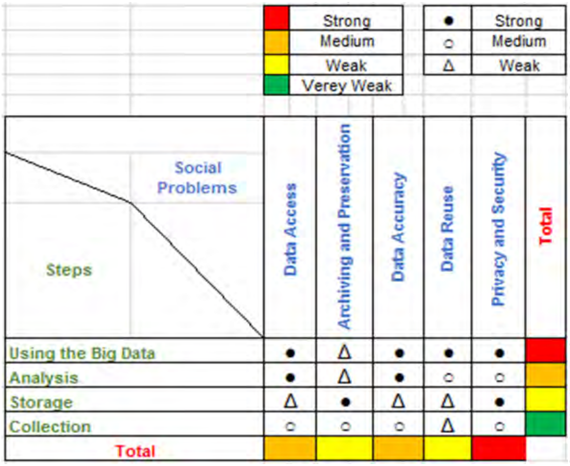
Some countries and regions are ahead of the United States in terms of the data privacy policies that they have either implemented or are in the process of implementing. Greenleaf [4] offers a comparative look at privacy laws, though most are not actually big data specific. Perhaps the best example of data privacy policies from which the US might learn is the European Union. Industry-spanning comprehensive big data policies that affect all data within a geographical territory are being considered in the EU these would affect anyone generating, transmitting, or storing data in the EU’s geographical territory. Other countries are also considering implementing similar laws, which indicates that these types of policies might be possible to consider in the US context.

The gaps in the current US policies and mechanisms regarding big data privacy are identified in the PCAST report and what is implemented elsewhere. The US has only focused on industry-specific privacy and data protection legislation and regulation (HIPAA, Ferpa, etc.) and not the comprehensive approaches as seen elsewhere. Given the recommendations of the PCAST report [1], especially recommendations 2 and 5, it is apparent that in order to meet increasing social needs regarding technology in the United States, a comprehensive policy is needed to address privacy issues that exist across jurisdictions and sectors.

Analysis

We utilized the technology roadmapping methodology to analyze the policy solutions that could be applied to the social problems created by big data, taking the multi-organizational approach as outlined in Phaal et al [14]. Technology roadmapping is a comprehensive approach for strategy planning to integrate science/technological considerations into business planning, and provides a way to identify new opportunities to achieve a desired objective from the development of new technologies. A technology roadmap of social problems and technologies can provide insight for policy makers, industry leaders, and private citizens regarding the expected developments in the field of big data. However, big data has vast effects on privacy across all sectors, and the literature varies considerably regarding projections and levels of certainty. Additionally, as outlined above, the groundwork is already laid for sector-specific approaches to data and privacy issues. "…Because no two businesses are the same, if privacy policies are the same or substantially similar, at least one of the privacy policies is not on point" [15]. As such, we first utilize Quality Function Deployment (QFD) analysis to identify an initial industry in which the need for policy change is most acute [16] [17]. We continue with this analysis method to narrow down the aspects of big data and the policy environment that are most pertinent to a roadmap.

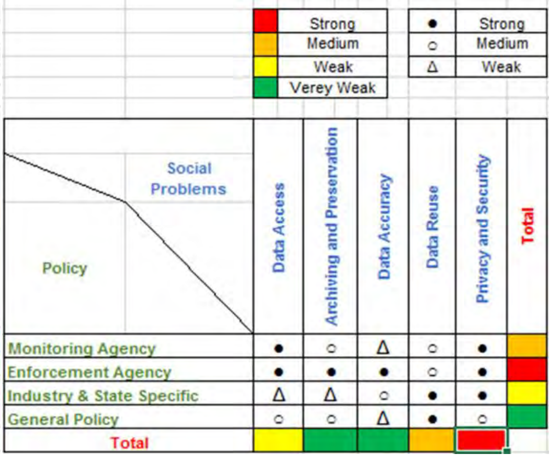
Five phases of QFD analysis were used to generate findings. The purposes of this phase of analysis were to prioritize social problems and policy needs, translate those needs into a specific industry, and identify resources necessary for policy change to occur. The categories included in each analysis are based on the literature consulted in the social problem definition and policy background stage of analysis (see above), the PCAST report [1], and additional literature on big data [20] [21] and in the healthcare industry particularly [18]. Each stage ranks between four and seven characteristics on a scale of strong, medium, and weak, with strong indicating the most pronounced effects based on each pairing. The purpose of this Prioritize the social problems and the policy needs.



🟥	Strong
🟧	Medium
🟨	Weak
🟩	Verey Weak

●	Strong
○	Medium
Δ	Weak

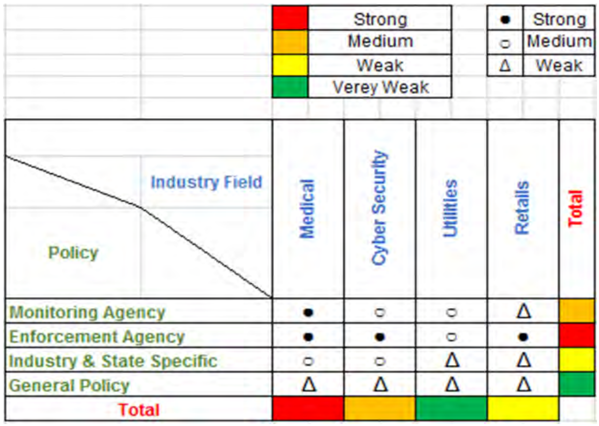
Social Problems Steps	Data Access	Archiving and Preservation	Data Accuracy	Data Reuse	Privacy and Security	Total
Using the Big Data	●	Δ	●	●	●	🟥
Analysis	●	Δ	●	○	○	🟧
Storage	Δ	●	Δ	Δ	●	🟨
Collection	○	○	○	Δ	○	🟩
Total	🟧	🟨	🟧	🟨	🟥	



🟥	Strong
🟧	Medium
🟨	Weak
🟩	Verey Weak

●	Strong
○	Medium
Δ	Weak

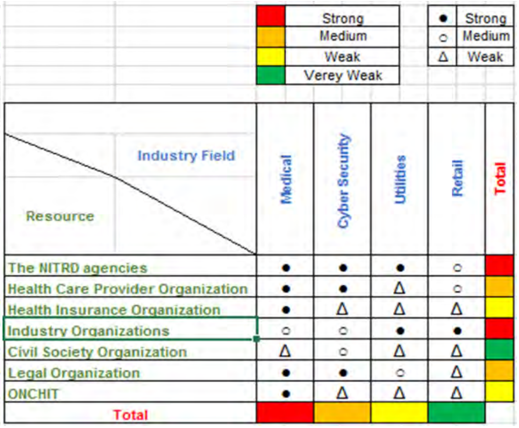
Social Problems Policy	Data Access	Archiving and Preservation	Data Accuracy	Data Reuse	Privacy and Security	Total
Monitoring Agency	●	○	Δ	○	●	🟧
Enforcement Agency	●	●	●	○	●	🟥
Industry & State Specific	Δ	Δ	○	●	●	🟨
General Policy	○	○	Δ	●	○	🟩
Total	🟨	🟩	🟩	🟧	🟥	



🟥	Strong
🟧	Medium
🟨	Weak
🟩	Verey Weak

●	Strong
○	Medium
Δ	Weak

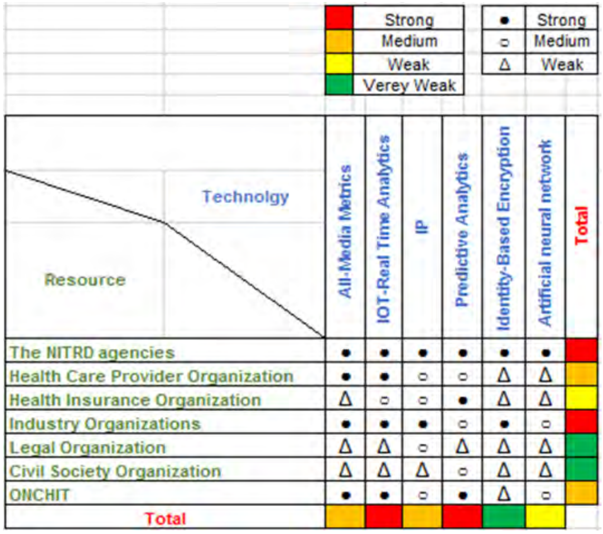
Industry Field Policy	Medical	Cyber Security	Utilities	Retails	Total
Monitoring Agency	●	○	○	Δ	🟧
Enforcement Agency	●	●	○	●	🟥
Industry & State Specific	○	○	Δ	Δ	🟨
General Policy	Δ	Δ	Δ	Δ	🟩
Total	🟥	🟧	🟩	🟨	



🟥	Strong
🟧	Medium
🟨	Weak
🟩	Verey Weak

●	Strong
○	Medium
Δ	Weak

Industry Field Resource	Medical	Cyber Security	Utilities	Retail	Total
The NITRD agencies	●	●	●	○	🟥
Health Care Provider Organization	●	●	Δ	○	🟧
Health Insurance Organization	●	Δ	Δ	Δ	🟨
Industry Organizations	○	○	●	●	🟥
Civil Society Organization	Δ	○	Δ	Δ	🟩
Legal Organization	●	●	○	Δ	🟧
ONCHIT	●	Δ	Δ	Δ	🟨
Total	🟥	🟧	🟨	🟩	



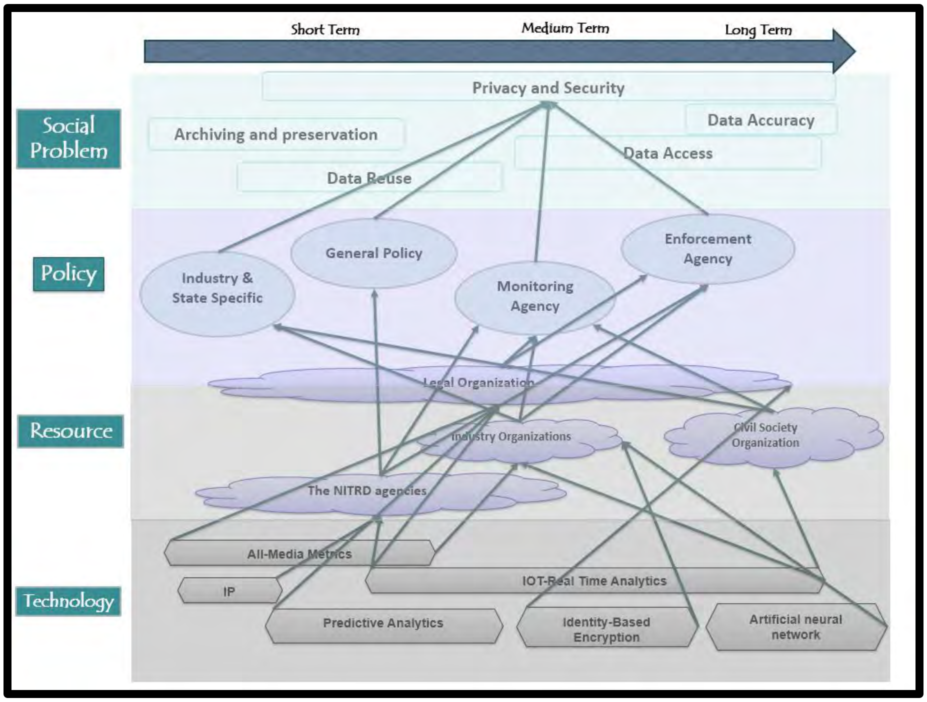
🟥	Strong
🟧	Medium
🟨	Weak
🟩	Verey Weak

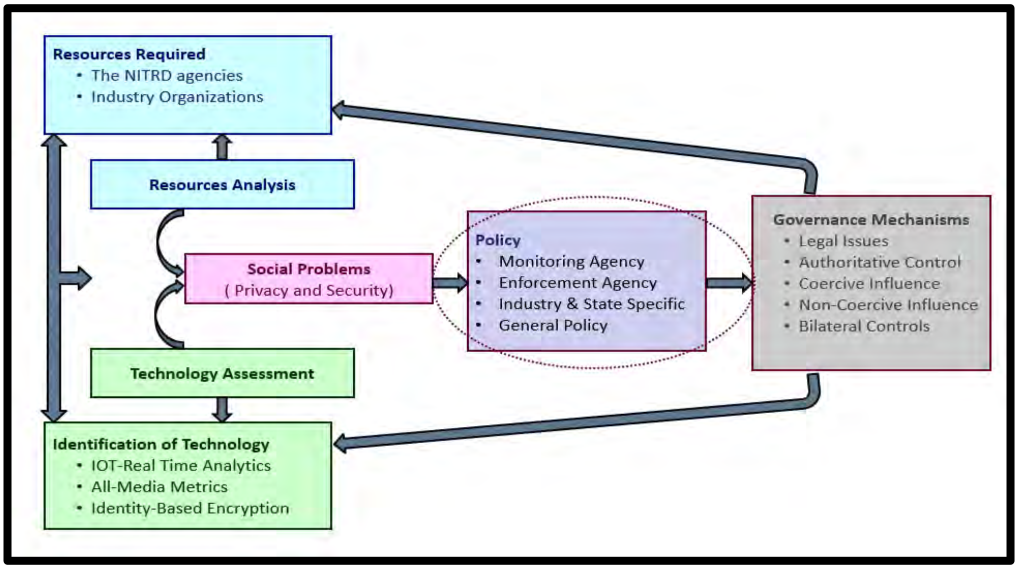
●	Strong
○	Medium
Δ	Weak

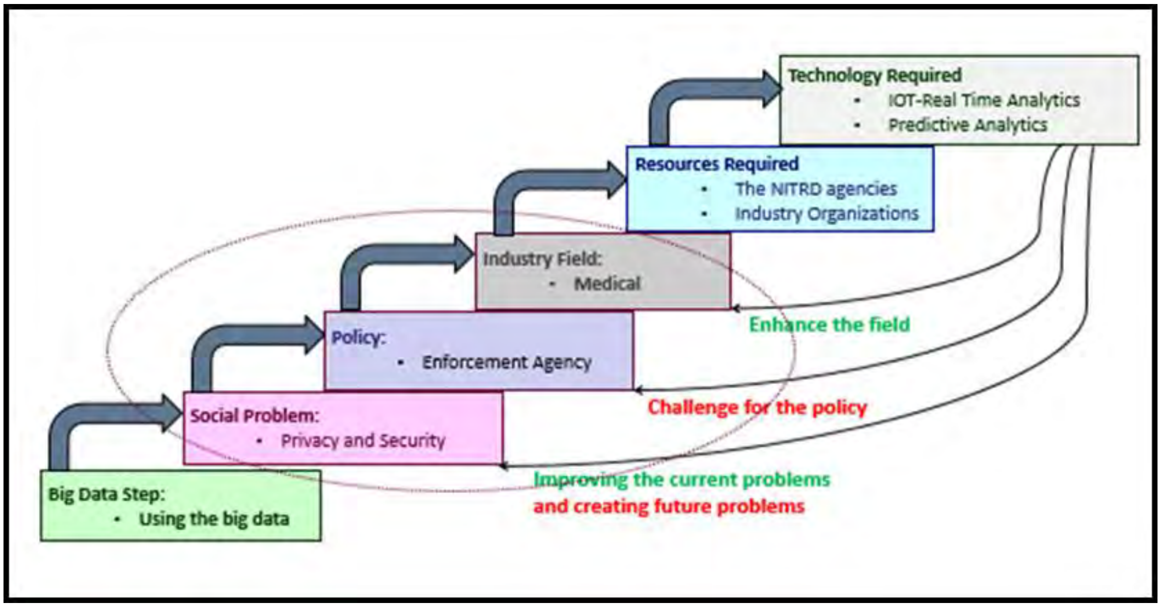
Technolgy Resource	All-Media Metrics	IOT-Real Time Analytics	IP	Predictive Analytics	Identity-Based Encryption	Artificial neural network	Total
The NITRD agencies	●	●	●	●	●	●	🟥
Health Care Provider Organization	●	●	○	○	Δ	Δ	🟧
Health Insurance Organization	Δ	○	○	●	Δ	Δ	🟨
Industry Organizations	●	●	●	○	●	○	🟥
Legal Organization	Δ	Δ	○	Δ	Δ	Δ	🟩
Civil Society Organization	Δ	Δ	Δ	○	Δ	Δ	🟩
ONCHIT	●	●	○	●	Δ	○	🟧
Total	🟧	🟥	🟧	🟥	🟩	🟨	

The analysis tables for all five phases are shown above. In the first phase, privacy and security in the context of the big data use phase receive the most "strong" rankings. Phase two shows that policies that include enforcement mechanisms or enforcement agencies are the strongest for addressing privacy and security issues. Phase three identifies the medical industry as the industry in which the effects of data policies are most felt. In terms of policy resources analyzed in the fourth phase, industry organizations and NITRD agencies have the strongest role to play in policy making that will impact data privacy and security in the medical field. And finally, the specific applications of big data technology that are strongest in this field, and therefore require the most consideration and social policy, are Internet of Things Real Time Analytics and Predictive Analytics. These technologies are specifically mentioned in the healthcare context by Roski et al [18] as providing the greatest opportunities for advancement but also the greatest threats to privacy and security for individuals.

The information from the QFD analysis was then compiled into the technology roadmap framework to provide a pictorial representation of the analysis. Using the typology of roadmaps articulated by Phaal et al [14], these are best described as strategic planning roadmaps that use four main elements (social problem, policy, technology, and resources) that are then extended with elements specific to this project: big data steps and industry. Because there are a variety of roadmap types, several depictions of technology roadmaps are included here. Each is informed by the QFD analysis and simply organizes the resulting information slightly differently. These roadmaps provide insight into the feedback loops that exist between technology, industry, and policy as big data becomes more ingrained in the processes of the healthcare industry. The steps in the development and adoption of any technology necessarily inform each other, and our analysis of relevant literature revealed the linkages shown in these roadmapping graphics. The biggest difference between these roadmaps


Short Term
Medium Term
Long Term
Social Problem
Policy
Resource
Technology
Privacy and Security
Data Accuracy
Archiving and preservation
Data Access
Data Reuse
General Policy
Enforcement Agency
Industry & State Specific
Monitoring Agency
Legal Organization
Civil Society Organization
Industry Organizations
The NITRD agencies
All-Media Metrics
IP
IOT-Real Time Analytics
Predictive Analytics
Identity-Based Encryption
Artificial neural network


Resources Required
• The NITRD agencies
• Industry Organizations
Resources Analysis
Social Problems
(Privacy and Security)
Technology Assessment
Identification of Technology
• IOT-Real Time Analytics
• All-Media Metrics
• Identity-Based Encryption
Policy
• Monitoring Agency
• Enforcement Agency
• Industry & State Specific
• General Policy
Governance Mechanisms
• Legal Issues
• Authoritative Control
• Coercive Influence
• Non-Coercive Influence
• Bilateral Controls



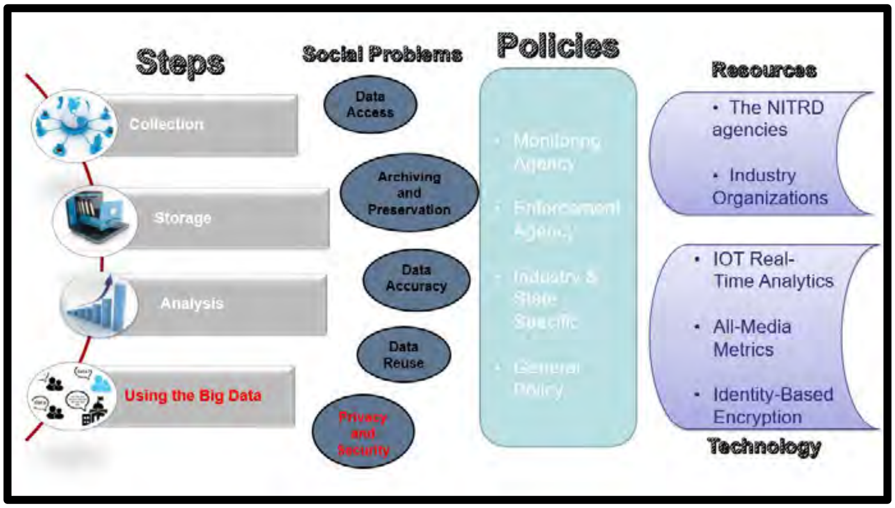
Discussion

There are several interesting points that arise from this analysis. First, while many fields are affected by big data, the healthcare industry poses both the biggest opportunities for big data to add value, and the biggest threats to individual privacy. Personal health data is considered among the most private and protected by strict laws. Big data, particularly the compilation of diverse data via typical processes and advancements like IoT-connected devices through processes that may or may not be evident to individuals generating data, are new challenges to data that do not have an analog counterpart. The use of these data streams for predictive analytics purposes also is a bigger concern given the sensitivity of this data. As the history of privacy in the United States shows, this is an evolving legal process that will continually be challenged as technologies develop and new uses for data are found. Data governance is a relatively new concept in the public sector in particular, and this will undoubtedly result in the emergence of additional social problems as values, risks, and costs are better assessed and big data governance policies implemented [19].

The policy tool with the most support for effectively protecting individual privacy and providing data security is an enforcement agency with the ability to impose penalties on corporations, organizations, and industries that fail to adequately protect information. This is similar to the enforcement mechanisms that are in the process of being implemented in the European Union [13]. This is a major departure from current US policies at all levels, which provide guidelines to industries but ultimately rely on self-reporting for enforcement and civil litigation. Criminalization of negligent data handling like that proposed in the European Union, while providing greater incentives to comply, also requires political capital to facilitate such a major policy shift. Thus, while this policy tool emerged from the QFD analysis as the one with the most potential benefits, it also may face political feasibility issues that could derail it in its entirety.

Conclusion and recommendations

Big data poses new security and privacy risks as greater quantities of data are generated, processed, and used, often without the knowledge of the individuals creating the data streams. The United States has a complex history with privacy that is only exasperated by the speed at which big data technology is adopted in industries that already collect and compile sensitive data. Big data policies that provide consistency across industries are desirable, as they create consistent expectations for individuals that are generating the data that is used in datasets and predictive analytics. For this reason, progress in the healthcare sector can be beneficial in other industries as well, as healthcare provides a blueprint for other sectors seeking to protect individuals' data. The use of big data is a primary concern for consumers and those managing data privacy in general, and modernization of privacy policies and data protection measures is needed to ensure individuals that their data is safe. The summary of the background information, QFD analysis and key aspects of the technology roadmapping are shown in the following figure.



There are three primary recommendations that arise from the analysis. First, the new presidential administration in the United States should renew the funding of those agencies that participate in big data research and policy formation, including the NITRD in particular. NITRD is an organization that is created through executive order, so it is subject to presidential renewal with each new administration. NITRD also created a 5-year research plan for big data, much of which focuses explicitly on the privacy and security of data. Continuing to research the impacts and possible policies associated with big data in a collaborative way is imperative to protecting current and future data streams, and ensuring the continued existence and funding of agencies that are committed to this mission seems the best way to do this.

Additional research on big data's relationship to data privacy and security should also continue. Future research should incorporate the costs of technologies and policies that can ensure data security, which are acknowledged but not a part of this analysis. The legal and political costs of policy change in this area also warrants further analysis, as even the PCAST report acknowledges that "privacy protection cannot be achieved by technical measures alone" (xii). The complexity of the policy environment surrounding big data governance is a factor that makes a single analysis method unlikely to capture the full scope of the opportunities and threats that are facing all actors. We therefore recommend that additional research continue combining methods of analysis to attempt to better understand the extent of the interacting aspects of the ever-changing technology environment.

While the opportunities that come from big data analytics are well documented, the particular technology and policy methods that can be used to best ensure data security and privacy remain unclear. This analysis shows that enforcement mechanisms in the form of agencies that are capable of leveraging civil and criminal penalties against those that fail to adequately guard the data generated by individuals provide the strongest protections. However, the social, political, and technical complexity of the policy environment continues to increase, and only time will tell if this analysis provides insight that is either possible or feasible to implement in the United States.

References:

[1]. PCAST. (2014). “Big Data and Privacy: A Technological Perspective.” Published May 2014. Web. https://bigdatawg.nist.gov/pdf/pcast_big_data_and_privacy_-_may_2014.pdf

[2]. NITRD. (2016). “The Federal Big Data Research and Development Strategic Plan.” Published May 2016. https://www.nitrd.gov/Publications/PublicationDetail.aspx?pubid=63 .

[3]. Jahanian, F. (2015). “The Policy Infrastructure for Big Data: From Data to Knowledge to Action.” *I/S: A Journal of Law and Policy for the Information Society* 10(3): 865-880.

[4]. Greenleaf, G. (2014). “Sheherezade and the 101 Data Privacy Laws: Origins, Significance and Global Trajectories.” *Journal of Law, Information & Science, Special Edition: Privacy in the Social Networking World* 23(1):

[5]. Hart, C., Jin, D.Y., & Feenberg, A. (2014). “The Insecurity of Innovation: A Critical Analysis of Cybersecurity in the United States.” *International Journal of Communication* 8(1): 2860-2878.

[6]. Stough, R. & McBride, D. (2014). “Big Data and U.S. Public Policy.” *Review of Policy Research* 31(4): 339-342.

[7]. Williamson, A. (2014). *Big Data and the Implications for Government.* Legal Information Management 14(): 253-257.

[8]. Van Dijck, F. (2014). “Datafication, dataism and dataveillance: Big data between scientific paradigm and ideology.” *Surveillance & Society: Newcastle upon Tyne.”* 12(2): 197-208.

[9]. Nissenbaum (1997). “Toward an Approach to Privacy in Public: Challenges of Information Technology.” *Ethics & Behavior* 7(3): 207-219.

[10]. Kshetri, N. (2014). “Big data’s impact on privacy, security and consumer welfare.” *Telecommunications Policy* 38(11): 1134-1145.

[11]. Kho, N. (2016). “The State of Big Data.” EContent, Jan/Feb 2016, 28-29.

[12]. Consumer Federation of California. (2017). “California Online Privacy Protection Act (CalOPPA).” Web. < https://consumercal.org/about-cfc/cfc-education-foundation/california-online-privacy-protection-act-caloppa-3/ >

[13]. Raymond, A. (2013). “Data management regulation: Your company needs an up-to-date/information management policy.” Business Horizons 56(4): 513-520.

[14]. Phaal, R., Farrukh, C., & Probert, D. (2004). “Technology roadmapping – A planning framework for evolution and revolution.” *Technological Forecasting & Social Change* 71: 5-26.

[15]. Pavolotsky, J. (2013). “Privacy in the Age of Big Data.” *The Business Lawyer* 69(1): 217-225.

[16]. Adiano, C. & Roth, A. (1994). “Beyond the House of Quality: Dynamic QFD.” *Benchmarking for Quality Management & Technology* 1(1): 25-37.

[17]. Chan, L.-K., & Wu, M.-L. (2002). “Quality function deployment: A literature review.” *European Journal of Operations Researce* 143(3): 463-497.

[18]. Roski, J., Bo-Linn, G., & Andrews, T. (2014). “Creating Value in Health Care Through Big Data: Opportunities and Policy Implications.” *Health Affairs* 33(7): 1115-1122.

[19]. Tallon, P. (2013). “Corporate Governance of Big Data: Perspectives on Value, Risk, and Cost.” *Computer* 46(6): 32-38.

[20]. Kaisler, M., Amour, F., Espinosa, J., & Money, W. (2013). “Big Data: Issues and Challenges Moving Forward.” *46th Hawaii International Conference on System Sciences* Conference Proceedings, 995-1004.

[21]. Assuncao, M., Calheiros, R., Bianchi, S., Netto, M., & Buyya, R. (2015). “Big Data computing and clouds: Trends and future directions.” *Journal of Parallel and Distributed Computing* 79: 3-15.

[22]. Desouza, K. & Smith, K. (2014). “Big Data for Social Innovation.” Stanford Social Innovation Review. Web. Published Summer 2014. < https://ssir.org/articles/entry/big_data_for_social_innovation >.