Software & Information
Industry Association
www.siia.net
SIIA

Comments of the
Software & Information Industry Association
on the
FTC Workshop on Big Data: A Tool for Inclusion or Exclusion?
September 15, 2014
(Comments filed October 31, 2014)

The Software and Information Industry Association (SIIA) appreciates the opportunity to comment on the Federal Trade Commission's (FTC) Workshop on "Big Data: A Tool for Inclusion or Exclusion?" The workshop usefully framed important developments in the use of data analytics for providing services to low income and underserved consumers and provided a forum for discussion of the possibility of unfair or discriminatory use of data analytics.

As the principal trade association of the software and digital information industry, the more than 500 members of SIIA develop and market software and electronic content for business, education, government, healthcare and consumers. SIIA's members are software companies, data analytics firms, compilers of data, information service companies, and digital publishers of business-to-business content. Software is a $425 billion industry that directly employs 2.5 million U.S. workers and supports millions of other jobs by driving American productivity.[1] As leaders in the global market for software and information products and services, our membership consists of some of the largest and oldest technology enterprises in the world, as well as many smaller and newer companies.

Mark MacCarthy, SIIA's Vice President of Public Policy, participated on the workshop's second panel relating to new developments in data analytics. SIIA's comments summarize his discussion in that panel and contain additional reflections on the issues raised by the workshop. The main points of these comments are:

- SIIA members are industry leaders using analytics to promote social and economic opportunity through, for example, alternative credit scoring models, cognitive computing in health care, predictive analytics in education, and detecting discrimination.

- Adequate and appropriate due diligence and transparency requirements already apply in regulated eligibility contexts such as lending, insurance, housing and employment.

- Due diligence and transparency mandates should not be applied to new contexts absent a showing that they are needed to prevent tangible harms.

[1] Software & Information Industry Association, The U.S. Software Industry: An Engine for Economic Growth and Employment, Prepared for SIIA by Robert J. Shapiro, 2014 at http://www.siia.net/index.php?option=com_docman&task=doc_download&gid=5446&Itemid=318

- Trade secrets and business confidentiality should continue to protect algorithms from disclosures. Such disclosures do not accomplish any meaningful public policy objective.
- The focus of policy concern should be on the uses of algorithms, not on the algorithms themselves.
- The FTC should continue the dialogue and discussion with other expert agencies and stakeholders to promote the positive use of data rather than seek additional laws or regulations.

General Remarks

Big data analytics is a natural evolution of older data analytics methodologies. It involves new processing techniques for analyzing data of increased variety, velocity and volume. Big data sets often consist of unstructured data such as text, images or video, or semi-structured data such as web logs. As a result, they require analytical techniques different from those typically used to analyze structured data bases.

Big data is a crucial development that allows analysts to detect patterns in data without first having to develop and test hypotheses. For instance, hospitals working with data analytics firms were able to pore through millions of data points from new born infants to discover the pattern of stable medical device readings up to 24 hours in advance of the onset of potentially deadly fevers. This pattern is based on real, but poorly understood causal mechanisms, and has enabled doctors and nurses to intervene before the crisis became apparent. The pattern could not have been detected by human observation alone, since it involved so many data points, and could not have been predicted in advance. Instead, it is a new insight that must be integrated into existing medical knowledge even as it is acted upon to save lives.[2]

The need for big data analytics will increase as the availability of data and the size of data sets increases. Some relevant facts:

- 90% of the world's data was created in the last two years
- 80% of the world's data today is unstructured
- 1 trillion connected devices generate 2.5 quintillion bytes data / day

Accompanying this increased availability of data will be an increasing demand for better data-driven decisions in healthcare, energy, education, public safety, transportation, marketing, and financial services. Indeed, the availability of data itself has never been sufficient. The key thing is the capacity to use analytical techniques to glean new insights relevant to organizational decisions.

[2] Victor Mayer-Schonberger and Kenneth Cukier, Big Data: A Revolution That Will Transform How We Live, Work, and Think, Houghton Mifflin Harcourt, 2013, pp. 60-61

It is important to distinguish different stages of the data analytics process. In 2013 the Center for Information Policy Leadership divided the process of robust analytics into two phases: discovery and implementation. The exploration or insight discovery phase is the initial step, where algorithms are developed, tested and refined; some analytics projects never move beyond this initial discovery phase. The application or use of algorithms and data in particular contexts is a separate and additional step. This second step is when laws, policy and consumer expectations are also applied to give guardrails and real-world meaning to the application of the algorithm.[3] This distinction between discovery and implementation mirrors the divisions made by leading privacy analysts such as Tal Zarsky (data gathering, data analysis and data use) [4] and Helen Nissenbaum (data monitoring, data analysis and data dissemination).[5]

Recently, the Information Accountability Foundation updated this distinction and applied it to the issue of ethics, fairness and value in data analytics. The Foundation's major conclusion is that ethics, fairness and value to the individual are criteria to be applied primarily at the stage of applying the algorithm and data use, not at the initial stage of exploration and discovery. Since discovery involves detecting patterns in data sets and testing the viability of algorithms, not in using the algorithm and applying data patterns for a particular purpose "the discovery phase is not usually personally impactful."[6] This is a key theme of these comments – policy issues such as fairness arise in the context of data use and not in the prior stage of the development of knowledge or insight.

It is also important to limit the FTC's policy concerns to the use of algorithms that involve personal information or that have a direct impact on individuals. Some of statistical models that are not consumer-facing and so fall outside this purview include:

> "…models used for financial analysis, financial management, or financial reporting; loss forecasting; capital adequacy; asset valuation; portfolio management or monitoring; market risk management; operational risk management (except perhaps customer-facing fraud models);

[3] Center for Information Policy Leadership, Big Data and Analytics: Seeking Foundations for Effective Privacy Guidance, February 2013 http://www.hunton.com/files/Uploads/Documents/News_files/Big_Data_and_Analytics_February_2013.pdf

[4] T.Z. Zarsky, "Desperately Seeking Solutions: Using Implementation-Based Solutions for the Troubles of Information Privacy in the Age of Data Mining and the Internet Society," 56 Me. L. Rev. 13, 2004, p. 30–32, available at http://www.mainelaw.maine.edu/academics/maine-law-review/pdf/vol56_1/vol56_me_l_rev_013.pdf

[5] Helen Nissenbaum, Privacy In Context: Technology, Policy, And The Integrity Of Social Life, Stanford University Press, 2010 p. 11

[6] Information Accountability Foundation, A Unified Ethical Frame for Big Data Analysis, October 8, 2014 at http://informationaccountability.org/wp-content/uploads/IAF-Unified-Ethical-Frame-v1-08-October-2014.pdf

> reporting; and other back-office analysis or risk management purposes that do not directly affect decisions regarding actual or potential customer accounts."[7]

SIIA members are industry leaders using analytics to promote social and economic opportunity through, for example, alternative credit scoring models, cognitive computing in health care, predictive analytics in education, and detecting discrimination

As detailed through the examples described below, data analytics are often used to promote inclusion. Indeed, an important function of public policy should be to endorse and incentivize the increased use of data analytics for the important public policy purpose of promoting social and economic opportunity.

Cognitive Computing in Health Care

New cognitive computing systems can understand natural language. They are big data systems that can process up to 60 million pages of text per second. They can answer questions, formulate hypothesis and test them, and they learn through interaction with users. They can be "trained" on large bodies of knowledge. IBM's Watson technology, the computer system that beat Jeopardy champions, is the best known example.

In healthcare, IBM has worked with hospitals to train its Watson technology to be an Oncology Diagnosis and Treatment Advisor. The system synthesizes vast amounts of data from textbooks, guidelines, journal articles, and clinical trials to help physicians make diagnoses and identify treatment options for cancer patients.

This system is a tool that can help health-care professionals, rather than replace them. It is a "clinical support" tool rather than a "decision making" tool. The doctor is always in charge.

The Watson system is in use today at Memorial Sloan Kettering, The University of Texas MD Anderson Cancer. At the Mayo clinic, it is helping to select patients for clinical trials.

How does Watson help the underserved?

There are severe shortages of some specialty providers in rural and low-income areas. Some 50 to 60 percent of community hospitals, for example, do not have an oncologist on staff. So people with complex conditions such as cancer often have to travel long distances or face long waits to receive treatment. Patients who face these barriers often forgo treatment or wait until they have severe complications before seeking medical intervention. The result is a higher cost of treatment and lower chance of survival.

[7] David Skanderson and Dubravka Ritter, Fair Lending Analysis of Credit Cards, Payment Card Center Federal Reserve Bank of Philadelphia, August 2014 at http://www.philadelphiafed.org/consumer-credit-and-payments/payment-cards-center/publications/discussion-papers/2014/D-2014-Fair-Lending.pdf

But now suppose that these new cognitive computing systems can be made available to community hospitals throughout the country? This provides the underserved with better access to the best cancer care available.

The promise of cognitive computing systems that can aid diagnosis and treatment of cancer is better quality care and lower cost for all patients, but especially for those who currently face these healthcare barriers.

Alternative Data Credit Scoring Models

Credit scoring models have been used for decades to increase the accuracy and efficiency of credit granting. They help as many people as possible to receive offers of credit on terms they can afford; and they allow lenders to efficiently manage credit risk. They improve upon the older judgmental systems that relied excessively on subjective assessments by loan officers.

The traditional credit scores are built from information in credit bureau reports and typically use variables relating to credit history. But these traditional credit scores have well-known limitations. They are not able to score approximately 70 million individual who lack credit reports or have "thin" credit reports without enough data to generate a credit score.

This inability to score no-file or thin-file individuals differentially affects historically disadvantaged minorities. A recent Lexis-Nexis study found that 41% of historically underserved minority populations of Hispanics and African-Americans could not be scored using traditional methods, while the unscorable rate for the general population was only 24%. Minorities face an unscorable rate that is 1.7 times the rate – almost twice – the rate for the general population.[8]

To remedy this limitation, companies are looking beyond the information contained in credit reports to alternative data sources and they are building credit scores based on this additional data. For instance, an alternative credit score, called RiskView, built by Lexis-Nexis relies on data such as public and institutional data such as educational history and professional licensing, property asset and ownership data such as home ownership, and court-sourced items such as foreclosures, evictions, bankruptcies, and tax liens.

The Lexis-Nexis report demonstrated the extent to which credit risk scores built from alternative data can help to extend credit to unscorable consumers, finding that fully 81% of the unscorable minorities received a RiskView score. A major benefit of alternative credit scores is the improvement in the availability of credit for historically underserved minority groups.

Continued data innovation for individual scoring products will continue and accelerate as competitive pressures drive other companies to develop and improve their own alternative data scores. In addition,

[8] Jeffrey Feinstein, Alternative Data and Fair Lending, Lexis-Nexis, August 2013 available at http://www.lexisnexis.com/risk/downloads/whitepaper/fair_lending.pdf

companies will be developing scoring techniques that improve the chances of small businesses to gain access to larger lines of credit to grow their businesses and compete directly with larger companies.

Predictive Analytics for Schools

A study by Johns Hopkins University research professor Robert Balfanz shows that most students who eventually drop out of high school can be identified as early as the sixth grade by their attendance, behavior and course performance. Even more can be identified by the middle of ninth grade.[9]

What should be done with this knowledge? Joel Allen, CEO of Amplify and former Chancellor of the New York City Department of Education, identified the path of inclusion and the use of data for promoting social and economic opportunity in education:

> "Using data to help identify these students and give them meaningful supports and interventions as early as possible would have a significant impact on the number of students that graduate ready for success in either college or career. This isn't the stuff of science fiction. These are actionable steps we can take right now, thanks to the power of technology."[10]

In fact, many schools throughout the country use data to identify students to improve their chances of graduating. Some use an early warning indicator and intervention system that uses attendance records, behavior problems and course performance to measure dropout risk. [11] In one school in 2013, one-third of students flagged for missing school got back on track to graduation. Two-thirds of the students who were having behavioral problems made a turnaround.[12]

IBM's Predictive Analytics Solution for Schools and Educational Systems (PASSES) uses a broader range of factors including demographic variables to identify at-risk students. Timely identification enables schools to intervene early to provide students with the right support and intervention. In Hamilton County Board of Education in Tennessee, for example, graduation rates increased by more than 8 percentage points and standardized test scores in math and reading increased by more than 10

[9] Robert Balfanz, Stop Holding Us Back, New York Times, June 4, 2014 at http://mobile.nytimes.com/blogs/opinionator/2014/06/07/stop-holding-us-back/?_php=true&_type=blogs&emc=edit_tnt_20140608&nlid=50637717&tntemail0=y&_r=0

[10] Joel Klein, Time to Disrupt Class, U.S. Chamber of Commerce Foundation, September 29, 2014 at http://www.uschamberfoundation.org/time-disrupt-class

[11] Mary Brucef and John M. Bridgeland, "The Use of Early Warning Indicator and Intervention Systems to Build a Grad Nation," Johns Hopkins University November 2011 at http://www.civicenterprises.net/MediaLibrary/Docs/on_track_for_success.pdf

[12] Sammy Mack, "Putting Student Data To The Test To Identify Struggling Kids," National Public Radio, April 08, 2014 at http://www.npr.org/blogs/ed/2014/04/08/300587823/putting-student-data-to-the-test-to-identify-struggling-kids

percent.[13] In Mobile County, Alabama the dropout rate has been nudged downward by three percent since the system's introduction.[14]

Detecting and Remedying Discrimination

As noted by several participants at the workshop, data analytics can be used to detect or remedy unfairness or discrimination. One example is a software recruiting tool that can help employers diversify their workforce:

> "…[R]ecruiting software company Entelo recently launched a "Diversity" product to help companies correct the underrepresentation of certain groups in their workforces. The tool allows employers to search for candidates with their desired professional qualifications and then filter by gender, race and military history, helping expand and diversify applicant pools. The software is designed to prevent recruiters from using it to discriminate against protected groups.[15]

Moreover, statistical techniques can be used to assess whether a statistical model has disproportionate adverse effects on protected classes. For instance, non-mortgage financial institutions do not have information about the race and ethnicity of their applicants and customers. To assess whether their statistical models comply with fair lending rules they need a way to infer race and ethnicity. Publicly available information from the U.S. Census Department on surnames and geo-location are reasonably reliable predictors of these characteristics, and advanced statistical techniques can improve the predictiveness of these factors.[16]

Using these proxies as factors in making lending decisions is exactly the kind of use of analytical techniques that the law is designed to prevent. But using the same correlations and statistical techniques to assess whether a statistical model has a disparate impact is a progressive use of data analytics that can detect and reduce potential discrimination. By publishing its compliance methodology, CFPB is encouraging companies to carefully assess their models for disparate impact.

[13] IBM, IBM Predictive Analytics Solution for Schools and Educational Systems, http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=YTS03068USEN&appname=wwwsearch

[14] IBM, Mobile County Public Schools: Analytical insights help keep students on track, IBM, 2011 p. 5 available at http://www.ibm.com/smarterplanet/us/en/leadership/mobilecounty/assets/pdf/IBM_MobileCounty.pdf

[15] Jules Polonetsky and Chris Wolf, Fighting Discrimination – With Big Data, The Hill, September 15, 2015 at http://thehill.com/blogs/pundits-blog/technology/217680-fighting-discrimination-with-big-data. See also Future of Privacy Forum, "Big Data: A Tool for Fighting Discrimination and Empowering Groups, September 2015.

[16] CFPB recently revealed the methodology it uses to assess disparate impact for fair lending compliance. Consumer Financial Protection Board, Using Publicly Available Information to Proxy for Unidentified Race and Ethnicity, Summer 2014 at http://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf . It does not mandate that anyone use this methodology but companies seeking to assess fair lending compliance risk are now in a position to more reliably make these assessments.

Adequate and appropriate transparency and due diligence requirements already apply in regulated eligibility contexts such as lending, insurance, housing and employment.

Much discussion at the workshop concerned the need for due diligence and transparency measures as a way to combat potentially discriminatory uses of data analytics. Adequate and appropriate measures already apply in regulated eligibility contexts such as lending, insurance, housing and employment. For instance, statutory constraints on discrimination in these areas include:

- Title VII of the Civil Rights Act of 1964 makes it unlawful for employers and employment agencies to discriminate against an applicant or employee because of such individual's "race, color, religion, sex, or national origin."[17] This is enforced by the Equal Employment Opportunity Commission and state fair employment practices agencies.

- The Equal Credit Opportunity Act makes it unlawful for any creditor to discriminate against any applicant for credit on the basis of "race, color, religion, national origin, sex or marital status, or age,"[18] which is enforced by the Consumer Financial Protection Bureau.[19]

- Title VIII of the Civil Rights Act of 1968, the Fair Housing Act**,** prohibits discrimination in the sale, rental or financing of housing "because of race, color, religion, sex, familial status, or national origin."[20] The act also protects people with disabilities and families with children. It is enforced by the Department of Housing and Urban Development.

- The Genetic Information Nondiscrimination Act of 2008 prohibits U.S. health insurance companies and employers from discriminating on the basis of information derived from genetic tests.[21] Enforcement is divided among a number of agencies including the Department of Health and Human Services (for health insurance) and the Equal Employment Opportunity Commission (for employment).

[17] 42 U.S.C. §2000e-2 available at http://www.law.cornell.edu/uscode/text/42/2000e-2

[18] 15 U.S.C. § 1691 available at http://www.law.cornell.edu/uscode/text/15/1691

[19] The Federal Reserve Board originally enforced the Equal Credit Opportunity Act, but the Dodd-Frank Act of 2011 transferred jurisdiction to CFPB. See Consumer Financial Protection Bureau, CFPB Consumer Protection Laws: ECOA, June 2013 p. 1 available at http://files.consumerfinance.gov/f/201306_cfpb_laws-and-regulations_ecoa-combined-june-2013.pdf

[20] 42 U.S.C. 3604 available at http://www.law.cornell.edu/uscode/text/42/3604

[21] Pub. L. No. 110-233, 122 Stat. 881 available at http://www.gpo.gov/fdsys/pkg/PLAW-110publ233/pdf/PLAW-110publ233.pdf

In addition, there are requirements under the Fair Credit Reporting Act when information is used for employment, insurance or credit granting.[22] These requirements include notice of adverse action, disclosures, access and correction, use limitations, consent, and redress.

These laws provide for reasonable, contextually appropriate amounts of due diligence and transparency. Importantly, they impose legal obligations on the providers of data, compilers of data, and users of data and data analytics, but not directly on the providers of analytics services. FCRA, for example, contains a regulatory structure applicable to credit reporting agencies, users of credit reports and furnishers of information to credit reporting agencies and provides different and discrete requirements for each group tailored to the relevant risk of harm. Providers of analytics services are not directly within the regulatory ambit of FCRA, the fair lending laws or other anti-discrimination, but are brought into the regulatory sphere through the need to provide products and services which their clients can use in a way that complies with their regulatory requirements.

Even in these eligibility contexts, the legal framework appropriately limits its due diligence and transparency requirements. For instance, there is no absolute requirement for users of data analytics to conduct disparate impact assessments. Instead, such procedures are encouraged by the need to demonstrate compliance to bank examiners and to qualify as empirically derived, demonstrably and statistically sound ("EDDSS") models. Also, in the case of lending decisions, companies routinely perform internal tests to assure themselves that the models they use are predictive of creditworthiness, including both traditional and alternative scoring models. Reviews of these studies have been made public.[23]

The transparency requirements are limited to obligations to disclose credit scores in certain cases such as where the scores were used for adverse decisions. But there is no legal obligation to disclose the statistical models or scoring algorithms used in particular cases. Trade secrets with respect to algorithms are permitted in these regulated contexts even when those algorithms are used for eligibility decisions.

These tailored and targeted limitations have allowed companies to employ consumer reports effectively while minimizing the risk of harmful uses of that data. Indeed, federal agencies have found that credit scores accurately predict creditworthiness and insurance risk. In 2007, the FTC evaluated credit insurance scores and found that

> Credit-based insurance scores are effective predictors of risk under automobile policies. They are predictive of the number of claims consumers file and the total cost of those claims. The use of scores is therefore likely to make the price of insurance better match the risk of loss posed by

[22] http://www.consumer.ftc.gov/sites/default/files/articles/pdf/pdf-0096-fair-creditreporting- act.pdf

[23] See, for instance, Center for Financial Services Innovation, The Predictive Value of Alternative Credit Scores, November 26, 2007 at http://www.cfsinnovation.com/node/330262?article_id=330262

> the consumer. Thus, on average, higher-risk consumers will pay higher premiums and lower-risk consumers will pay lower premiums.[24]

Also in 2007, the Federal Reserve Board of Governors assessed credit scores used for lending purposes and found that

> "[t]he credit history scores … are predictive of credit risk for the population as a whole and for all major demographic groups…[with] the higher (better) the credit score, the lower the observed incidence of default.[25]

As Commissioner Julie Brill noted in her speech at the FTC's workshop, the question of whether credit scores act as proxies for protected classes has also been addressed by "extensive and rigorous" FTC and Fed studies that "found that the scores they examined largely did *not* serve as proxies for race or ethnicity."[26]

Due diligence and transparency mandates should not be applied to new contexts absent a showing that they are needed to prevent real harms.

The legal requirements appropriate for preventing unfairness and discrimination were imposed to cover specific cases where it was thought the dangers of harm were the greatest and where there was evidence that these harms were in fact occurring. Housing, employment, insurance, and credit granting are so important to the life prospects of individuals that unfair treatment in these areas could substantially reduce their chances of economic and social success. To be subject to adverse action in these areas in an unfair manner or merely because of membership in a protected class was a substantial

[24] Federal Trade Commission, Credit-Based Insurance Scores: Impacts On Consumers Of Automobile Insurance (July 2007), p. 3, available at http://www.ftc.gov/sites/default/files/documents/reports/credit-based-insurance-scores-impacts-consumers-automobile-insurance-report-congress-federal-trade/p044804facta_report_credit-based_insurance_scores.pdf

[25] Board Of Governors Of The Federal Reserve System, Report To Congress On Credit Scoring And Its Effects On The Availability And Affordability Of Credit (Aug. 2007), p. S-1 at http://www.federalreserve.gov/boarddocs/rptcongress/creditscore/creditscore.pdf

[26] Commissioner Julie Brill, Remarks at the FTC Workshop on "Big Data: A Tool for Inclusion or Exclusion?" Federal Trade Commissioner Julie Brill September 15, 2014, available at http://www.ftc.gov/system/files/documents/public_statements/582331/140915bigdataworkshop1.pdf See also Board Of Governors Of The Federal Reserve System, Report To Congress On Credit Scoring And Its Effects On The Availability And Affordability Of Credit (Aug. 2007), *available at* http://www.federalreserve.gov/boarddocs/rptcongress/creditscore/creditscore.pdf and also Federal Trade Commission, Credit-Based Insurance Scores: Impacts On Consumers Of Automobile Insurance (July 2007), *available at* http://www.ftc.gov/sites/default/files/documents/reports/credit-based-insurance-scores-impacts-consumers-automobile-insurance-report-congress-federal-trade/p044804facta_report_credit-based_insurance_scores.pdf . These questions are also addressed in the Federal Reserve Board's 2010 staff report, Does Credit Scoring Produce a Disparate Impact? available at http://www.federalreserve.gov/pubs/feds/2010/201058/201058pap.pdf

harm that policy had to address. Moreover, there was ample evidence that without significant legal protection members of protected classes would face discriminatory barriers, unfair treatment and exclusionary practices that would in fact lead to the substantial diminishment of their chances for economic success. In short, there was evidence of actual harm that led to the imposition of these legal requirements.

This complex of legal requirements does not apply outside specific protected classes and specific contexts. And it should not be extended to new protected classes or to new contexts without a similar showing that the new requirements are need to prevent real harm.

Nothing at the FTC's workshop suggested the presence in today's marketplace of significant harmful discriminatory practices outside the traditional areas that need a remedy in the form of new law or regulation. The examples that were discussed were either already amply covered under existing anti-discrimination and unfairness law or were not evidence of substantial harm. In the first case, the remedy is to enforce existing law. In the second case, there is no problem for law or regulation to address.

Trade secrets and business confidentiality should continue to protect algorithms from disclosure. Such disclosure does not accomplish any meaningful public policy objective.

During the workshop, several participants called for requiring companies to disclose their algorithms. However, complete algorithmic transparency would prevent the use of trade secrets in the area of data analytics. Such a proposal is harmful and it does not meaningfully advance any public policy purpose. Even the industries regulated for anti-discrimination purposes and for fairness are permitted to maintain their algorithms as trade secrets.

The fundamental rationale for trade secrets is to provide those who invent something the ability to exploit it without others being able to take advantage of the fruits of their innovative efforts. Trade secrets promote innovation and competition.

Imagine how a proposal for algorithmic transparency would work in practice. Consider a company that invested billions of dollars developing software and analytical capabilities to provide new insights into healthcare. Under algorithm transparency, as soon these capacities are brought to market they must be revealed to the general public. This revelation would include the formulas used, the software code involved, and perhaps even the software engineering notes used to develop these analytical capabilities. The information would then be available to the company's competitors, who would immediately copy the successful techniques and rush to market with an alternative product that does much the same thing. But having spent no resources to develop the product, they can offer it at a fraction of the cost of the original developer. Why would analytics companies invest huge sums in developing, improving, updating their algorithms if they were required to disclose them immediately to competitors?

Such a proposal is affirmatively harmful to the industries involved and to the public interest in competition and innovation in the data analytics industry.

The rapid pace of change for today's algorithms provides another reason for rejecting calls for even one-time disclosure. Any such disclosure of formulas, source code and related features of the operation of an algorithm would merely provide a snap-shot representing what the analytical machine does at a particular point in time. But one of the features of today's algorithms is that they learn and update themselves in real time. They can give different weights to different data elements, discard some completely or add new ones all depending on continuous feedback from users. A snapshot disclosure would provide an incomplete picture, giving distorted importance to a moment in time that is often quickly superseded by the next improvement in the algorithm.

Data journalists interested in understanding the uses of algorithms for decision making recognize the difficulties of public transparency as a solution. One noted that companies are reluctant to make their statistical models public since

> "…exposing too many details of their proprietary systems (trade secrets) may undermine their competitive advantage, hurt their reputation and ability to do business, or leave the system open to gaming and manipulation."

Gaming and manipulation are "real issues" he says, quoting Goodhart's rule that when a measure becomes a target it ceases to be a good measure. For example, measures for fraud prevention and identity authentication would become difficult if not impossible if the fraudsters and identity thieves knew the details of the algorithms used to detect them. Spammers and criminals already spend billions trying to game search results, a task that would be much easier if they know more about the internal workings of search algorithms.

To protect trade secrets, U.S. law does not and should not provide a requirement to disclose the formulas that power data analytic techniques. Even in the cases where the context is protected by law such as employment, lending, and insurance disclosure of algorithms is not required. Moreover, a change in this legal regime to mandate the disclosure of this analytic information should be contemplated only when it is necessary to prevent substantial harm to consumers. No showing has been made that such a change is necessary.

The focus of policy concern should be on the uses of algorithms, not on the algorithms themselves

The idea that an "algorithm" makes a decision is a concept that needs to be rejected clearly and completely if there is to be any progress in understanding the issues raised by data analytics. In many cases, the notion of algorithmic decision-making is just shorthand for the idea that organizations use algorithms to aid them in making decisions. But like many such pieces of shorthand it can lead policymakers to focus in the wrong area. In particular, it seems to suggest that that the analytic tool is the issue rather than the use of the tool for particular purposes.

The analytical process can be divided in many ways, but a helpful one in the policy context is the distinction between the stage at which knowledge and insight is gathered and the stage at which such knowledge is put to use for a particular purpose. As mentioned earlier, the Center for Information

Policy Leadership, the Information Accountability Foundation and other privacy scholars have made this distinction a core part of the way in which they analyze the data analytics process for policy purposes. Several examples illustrate the difference.

An audio file recognition program is a system that matches a new audio file against a library of such files and estimates the probability that the new file is the same as one of the files in its library. An organization determines what level of probability it will accept as an indication of a match, and that depends on the purposes for which it is attempting to match the files. The purposes could include letting people know what music they happen to be listening to or detecting copyright infringement. The algorithm determines neither the acceptable probability level nor the purpose for which a match will be used. If there is a policy issue, it relates to the use of the algorithm, not to the algorithm itself.

As mentioned earlier, some characteristics reveal race and ethnicity so clearly that they effectively function as proxies for them. Two well-known examples are surname and census geography. There are advanced statistical techniques that can combine these two characteristics to generate a new variable that is an even better statistical proxy for race and ethnicity. The correlations and algorithms that create proxy variables do not determine the use of proxies. If there is a policy concern about proxies, it is not in the research and knowledge discovery process that leads data scientists to develop these proxies but in the particular use to which they are put.

The use of proxies for race and ethnicity to make eligibility decisions in the areas of housing, lending, insurance and employment would seem to violate antidiscrimination rules. If organizations cannot discriminate in these contexts on the basis of membership in a protected class, then it is highly likely that antidiscrimination enforcement agencies would raise significant questions if they based their decisions on use a proxy for membership in a protected class. An approach that says to focus on the algorithm might say that the construction and development of these proxies is a matter of concern and should be discouraged. Yet, these proxies are useful in assessing whether lending decisions have a disproportionate adverse impact on protected classes. As the CFPB recently explained:

> “Information on consumer race and ethnicity is generally not collected for non-mortgage credit products. However, information on consumer race and ethnicity is required to conduct fair lending analysis. Publicly available data characterizing the distribution of the population across race and ethnicity on the basis of geography and surname can be used to develop a proxy for race and ethnicity. Historically, practitioners have relied on proxies based on geography or surname only. A new approach proposed in the academic literature—the BISG (Bayesian Improved Surname Geocoding) method—combines geography- and surname-based information into a single proxy probability. In supervisory and enforcement contexts, Office of Research (OR) and Division of Supervision, Enforcement, and Fair Lending (SEFL) rely on a BISG proxy

> probability for race and ethnicity in fair lending analysis conducted for non-mortgage products."[27]

Models and statistical analysis involving proxies can be used to assess the presence of disproportionate adverse impacts on protected classes and can guide lenders who might want to avoid lending practices that have such impacts. The issue for policymakers is not the existence of proxies for race and ethnicity but how these proxies are used. Some uses promote exclusion; some are essential in the fight against it.

In general, policymakers should be concerned with how algorithms are used. The research and discovery process that leads to their construction should not be their focus.

FTC should continue the dialogue and discussion with other expert agencies and with stakeholders to promote the positive use of data rather than seek additional laws or regulations.

The FTC has done the public and the business community a great service by drawing attention to this important issue. Public discussion of the extent to which data analytics can be used to exclude people from important dimensions of our nation's social, cultural, economic and political life are absolutely essential to make sure that we get ahead of any potential problems associated with the use of these powerful tools. As we have mentioned throughout this submission, such a discussion has to be balanced with sufficient attention to the substantial benefits for social and economic progress made possible by data analytics. And it has to focus on significant risks of substantial harm rather than on speculative or hypothetical concerns that have no basis in actual or likely business practices. Still a focus on ways to avoid any exclusionary effects of the use data analytics is vitally important and welcome.

One possible way forward in this area is to continue conversations and discussions, perhaps in the form of additional workshops, to explore the boundaries of business responsibility in this area. Businesses have stepped up to the plate with voluntary efforts in the past when challenged by new technologies that can pose public risks. The work by the Direct Marketers Association to establish codes of conduct is important.[28] For instance, its principle in article 32 that "marketing data should be used only for marketing purposes" assures the public that the data bases compiled by their member companies will be used for marketing purposes only and will not be used against them in other contexts such as employment or insurance. In a similar way, the Digital Advertising Alliance established self-regulatory principles for online advertising, including the pledge that tracking information gathered by online advertising networks would not be used for eligibility decisions.[29] Most recently, the Council of Better

[27] Consumer Financial Protection Board, Using Publicly Available Information to Proxy for Unidentified Race and Ethnicity, Summer 2014 at http://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf. p. 23

[28] Direct Marketing Association, DMA Guidelines for Ethical Business Practice January 2014 at https://thedma.org/wp-content/uploads/DMA-Ethics-Guidelines.pdf

[29] Digital Advertising Alliance, DAA Self-Regulatory Principles at http://www.digitaladvertisingalliance.org/content.aspx?page=principle

Business Bureaus launched a program called Digital IQ to explore the ways the business community could empower people to make smart choice about their data in an age of ubiquitous data collection.[30]

Stakeholders could explore, encourage and highlight continued business efforts in this area to make sure that established social norms of fairness and non-discrimination are appropriately implemented in all aspects of the data analytics process. For instance, the development of models might need to be more closely associated with those who are using the models. This suggestion was helpfully made in a recent paper published by the Federal Reserve of Philadelphia in regard to the use of credit scores for assessing credit card risks.[31] The point was that the compliance, marketing and analytical functions within a financial service organization could usefully cooperate at an early stage of model construction so that fair lending compliance was accomplished more easily.

The suggestion is extendable to other contexts of data analysis. Privacy by design programs and privacy impact assessments have been features of privacy practice for many years now.[32] The key insight of these techniques is that it is easier and more efficient to design privacy protections into a product, process or service at the beginning, rather than try to graft it on to something that has been constructed without those considerations in mind. These same techniques can be used to assure that statistical models build in fairness and non-discriminatory elements at the beginning

As noted before data analytics software can be used to help employers diversify their workforce, to assess compliance with fair lending laws and in other ways help to detect and remedy discrimination. While disproportionate adverse impacts on protected classes are legally prohibited only in certain industries and contexts, all business have an interest in knowing whether their use of data and algorithms might have these effects. Without seeking new legal mandates and prohibitions, the FTC could provide encouragement and support to organizations who take the extra step of using data to assess the fairness of their own decisionmaking techniques.

The OECD’s recent Global Forum on the Knowledge Economy – Data – Driven Innovation for a Resilient Society reached a conclusion that SIIA can endorse:

> Governments and stakeholders need to develop a coherent policy approach to harness the economic benefits of data driven innovation. They need to assess the context for data

[30] Council of Better Business Bureaus, “Better Business Bureau to Launch “Digital IQ” Initiative with Axciom,” July 1, 2014 at http://www.bbb.org/council/news-events/news-releases/2014/07/better-business-bureau-to-launch-digital-iq-initiative-with-acxiom/

[31] David Skanderson and Dubravka Ritter, Fair Lending Analysis of Credit Cards, Payment Card Center Federal Reserve Bank of Philadelphia, August 2014 at http://www.philadelphiafed.org/consumer-credit-and-payments/payment-cards-center/publications/discussion-papers/2014/D-2014-Fair-Lending.pdf

[32] See, for instance, the seven foundational principles for privacy by design at http://www.privacybydesign.ca/index.php/about-pbd/7-foundational-principles/

> collection, analysis and use to ensure that data - driven innovation serves societal values in an ethical and equitable manner.[33]

Stakeholders should continue their discussions on these important issues. The objective should be for government, industry, academics, and advocates from all sides to gain a shared understanding of how business and other organizations should use the powerful new analytics tools at their disposal to promote inclusion and economic opportunity. Looking for new legal tools to punish businesses is not the way forward. Constructive and open dialogue would be.

[33] Organization for Economic Cooperation and Development, Global Forum on the Knowledge Economy – Data-Driven Innovation for a Resilient Society, October 2 – 3, 2014, Highlights at http://www.oecd.org/sti/GFKE2014_highlights.pdf