

Ritesh Soni

Masters of Information and Data Science (candidate)
School of Information
University of California, Berkeley

October 4, 2014

Federal Trade Commission
600 Pennsylvania Ave, NW
Washington, DC 20001

Re : *Big Data: A Tool for Inclusion or Exclusion?* (Project No P145406) Workshop Comment

Dear Commissioners,

Thank you for organizing this workshop and raising awareness of an important issue of our times as big data and related policy frameworks. In the context of FTC's workshop and its focus on examining the potentially positive or negative effects of big data on low income and underserved populations, I submit these comments for the record.

These comments are focussed on two of principles of Fair Information Practice Principles¹ in the age of Big Data :

- **There must be a way for the individual to find out what information about him is in a record and how it is used.**
- **There must be a way for an individual to correct or amend a record of identifiable information about him**

As noted in the FTC's report on the data broker industry², companies both direct to consumer companies in various sectors together with the data broker industry collect vast amount of data on individuals with multiple layers of data brokers sharing data. There is little transparency if any about both the source of data and models created to infer additional attributes. Given lack of standards around enabling the FIPPs principles above the entire industry operates far below the consciousness of the human subjects it affects.

Professor Alessandro Acquisti (Panel 2) suggested in his closing remarks the crucial role Data Provenance can play enable transparency and accountability in this space. I whole heartedly agree with him that data provenance will be a significant step in the right direction, although it faces limitations some of which I highlight here. The primary

¹Records, Computers and the Rights of Citizens <http://epic.org/privacy/hew1973report/>

² Data Brokers : A call for Transparency and Accountability <http://www.ftc.gov/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014>

limitation is due to the inverse relationship between model flexibility and interpretability for inferred data. If data was simply sourced and traded the application of a data provenance framework would enable clear traceability and transparency. It is due to the introduction of inferred data, and the complex models used to infer such data, that breaks the chain of information and severely restricts our ability to trace back. The following section helps both understand these limitations, and proposes an approach forward

Data provenance and big data

Provenance information explains the creation process and origin of data by recording which data items a given data item is derived and which transformations were responsible in creating a certain piece of data (a so-called data item)³. Both data provenance and transformation provenance enable an individual to find out what information about him is in a record and how it is used. However the transformations can range from basic operations to fairly advanced predictive analytics that can have direct impact on the *granularity* of such provenance information.

Prediction accuracy and model interpretability

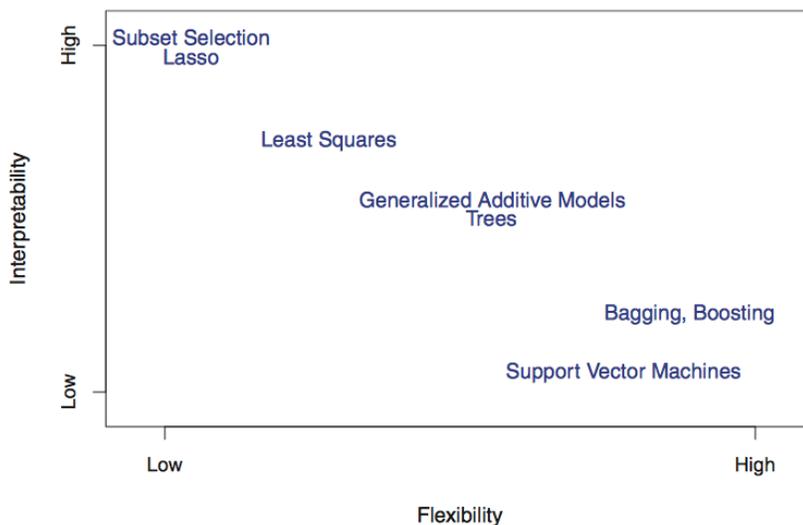


Figure 1 : Tradeoff between Flexibility an Interpretability

The race to find the next best prediction models is often challenged by the inverse relationship between how flexible and powerful a model is vs the ability to interpret the output generated by the model (See Figure 1⁴). Very often interpretability is sacrificed for higher accuracy leading to models that the creators themselves don't fully understand. Netflix's VP of product innovation, Todd Yellin who conceived the *alt genre* to predict newer genres that people would like, isn't himself quite sure why there are so many

³ Big Data Provenance <http://cs.iit.edu/~dbgroup/pdfpubls/G13.pdf>

⁴ An Introduction to Statistical Learning <http://www-bcf.usc.edu/~gareth/ISL/>

altgenres that feature Raymond Burr and Barbara Hale⁵. It's inexplicable with human logic. Areas where such interpretability is being sacrificed is of concern.

This is of concern as simpler parametric models (e.g. Lasso, Least squares) are able to provide more clear insight into the relationship between an inferred data-item (predicted value) and the data-items used for the analysis (predictor variables) vs non parametric models. Not only is our ability to interpret the ability stronger but our ability to quantify the effect of each predictor variable much more easier.

Extending transformation provenance

As Solon in his opening remarks called out that unintentional discrimination is likely to be far more common than the kinds of discrimination that could be pursued intentionally. Fine grained data provenance combined with a common taxonomy to standardize the industry wide data elements and segments would be a start. It follows from the second principle that in order to allow for an individual to correct a piece of information about him, the relationship between the final predicted category or score and the various data elements used for this purpose must additionally be clear.

Companies engaged in creating complex models for high accuracy should make efforts to supplement these models with one's that are perhaps lower in accuracy but provide high degrees of interpretability for provenance purposes. In the end creating two models, the first model (Model 1) to enable clear auditing and traceability and the second (Model 2) for high accuracy for core market competitiveness. Transformation provenance should be extended with information about Model 1 and relative predictor weights in the model. This enables a low fidelity traceability of interpreting a final inferred data item.

Data usage provenance

In many modern data collection mechanisms, users are notified of the data being collected and the use for such data collection. This usage needs to be formally defined and data provenance extended with *usage provenance*⁶. This extension would serve the purpose of

- enabling data consumers such as agencies, marketers, retailers to ensure that their use of data through the entire data chain (primary, traded and inferred) is consistent with the consumers usage agreement and
- enabling consumers to understand the reasons behind a certain data attribute, the related usage rights and the ability to act on such usage rights contextually
- Establishes an even more vibrant data trading market and a confident use of consumer data by the industry that is unambiguously tied to usage rights

⁵ How Netflix reverse engineered Hollywood http://www.theatlantic.com/technology/archive/2014/01/how-netflix-reverse-engineered-hollywood/282679/?single_page=true

⁶ Fine-Grained User Privacy from Avenance Tags <https://www.cs.cornell.edu/fbs/publications/avenanceHotPET.pdf>

Consumer Record

Primarily Sourced Data

Provide the ability to validate, modify or opt out of sharing this data directly with the entity

Traded Data

Provide the ability to traverse this hierarchy to get to primary source data. The ability to easily act on the

Inferred Data

Understand the variables contributing to this inferred data item

Model r : The actual model being used described in a standard format (PMML)

Predictors : Predictors with weighted contribution based on the model

Usage Rights

Usage model attached to every data element

Inferred data should inherit the intersection of usage items associated with each predictor

Additionally, the usage attributes for the inferred data should be the intersection of the usage attributes attached with each of its predictor variables. This would not only ensure that the usage for the inferred data item is in fact what was intended, but also allow for usage information to flow through inference steps in the same manner as interpretability. This approach would also serve as a market force to encourage data brokers to use the minimum viable set of predictors in creating inferred values in order to retain the maximum usage rights of the inferred data-item. The lesser sources of data they use, the more rights they'll retain post analysis and the more the value of the resulting dataset.

Concluding remarks

The proliferation of data, advancements in computational and statistical techniques and the application of advanced analytics is here to stay as a technological force defining societal evolution, economic progress and new research frontiers. This can either act as a democratic equalizer in identifying and offering greater protection against discrimination of underserved and low income populations, or can act as a catalyst to further digital polarization. Enabling a foundation for transparency based on a data provenance standard, that retains interpretability through the entire data chain, is a proposed mechanism to direct these advances to enabling greater human potential.

Sincerely Yours

Ritesh Soni