Approved by the Data subcommittee on January 24, 2020

**Office of the Child Advocate**

**Juvenile Justice Policy and Data Board- Subcommittee Meeting**

**October 10th, 2019**

**Members and Designees in Attendance:**

- David Chandler (DYS)
- Lydia Todd (CLM)
- Abigail Averbach (DPH)
- Kristi Polizzano (Probation)
- Naoka Carey (CfJJ)
- Patricia Bergin (EOPPS)
- Eneida Anjos (DCF)
- Joshua Dohan (CPCS)
- Mike Glennon (MDAA)
- Leon Smith (CfJJ)

**Other Attendees:**

- Azer Bestavros (Boston University)
- Mayank Varia (Boston University
- Melissa Threadgill (OCA)
- Lindsay Morgia (OCA)
- Members of the public

**Meeting Commenced:** 2:07pm

**Welcome and Introductions**

Melissa Threadgill welcomed the members of the board and the members of the public who attended the meeting. Everyone introduced themselves and which department/organization they are representing. She said that the first two-thirds of the meeting would be for the presentation from Boston University's Azer Bestavros. The last half hour will be to walk through some data.

**Approval of Minutes from September Meeting**

---Ms. Threadgill asked for a motion to approve the minutes. The group approved the minutes without objection.

Approved by the Data subcommittee on January 24, 2020

**Guest Speakers from Boston University**

*Sharing Knowledge without Sharing Data: Platforms for resolving the false dichotomy between privacy and utility of information*

Ms. Threadgill welcomed Dr. Bestavros, the Director of the Hariri Institute for Computing at Boston University, and his colleague, Mayank Varia. The Office of the Child Advocate has been fortunate to meet Dr. Bestavros and learn about his technology that enables data sharing amongst agencies/organizations while honoring client confidentiality.

Dr. Bestavros thanked the group for attending and Ms. Threadgill for believing in his technology. Dr. Bestavros and his colleague are both in the computer science department at Boston University. Their work is less about producing papers for academic journals and more about applying technology to real-world problems.

Dr. Varia introduced himself. He works in cybersecurity and is interested in how to take research and technology and apply it to problems.

Dr. Bestavros began by describing data as a function of knowledge. The key question is, can we get the knowledge without exposing the data? He provided an example of four gentlemen. Each knows their net worth, but they do not want to share that information. The knowledge we want is, who is the wealthiest? He provided additional examples from a Department of Labor lawsuit against Google regarding equal pay. The Department wants Google to hand over HR data. What is the DOL attorney trying to do? Determine if there is a pay different with statistical significance.

Another example is a bill in the Senate that asks if it is worth it to attend college. The IRS and the Department of Education have data that can help answer this question, but federal regulation prohibits sharing. Can we share the knowledge without exposing the private data?

Finally, Dr. Bestavros shared examples from the Juvenile Justice Leadership Forum. The Department of Children and Families has a database, as does the Department of Elementary and Secondary Education. If we want to determine if certain programs are helping to reduce absenteeism, for example, we would need to join the databases, and mine the data that is currently separate.

Dr. Bestavros said that knowledge can be derived from data "without requiring owners of the data to share it or to trust anything other than the mathematics". He uses a real-life example of his experience with Mayor Menino helping to address the gender wage gap in Boston. The Boston Women's Workforce Council wanted to learn more about pay disparities. They were able to get companies to agree to a compact to answer questions, but the project did not get off the ground, as companies were hesitant to give their data to a third party. In addition, lawyers

from a local educational institution, which was ready to do the analysis, said the project could not move forward because of the dangers of losing the data. One year later, under Mayor Walsh, Dr. Bestavros demonstrated his technology to show how they could do the analysis without any of the parties involved knowing the outcomes for any individual company. The first time they did they analysis was only to show how it worked; the results were not published. Now, the group is on its fourth year. They can analyze the data by race, gender, and type of job. People who did not know about the technology before now advocate for it.

Dr. Bestavros said that his programming, known as multi-party computation (MPC), computes a function based on the data from multiple parties each with private data. Nothing is revealed about the inputted data beyond what the output of *f*, an orthogonal question, reveals.

Why use MPC?

- Multiple parties have data that they do not want to share in fear of exposing the identities or sensitive information about their clients
- As a collaborative group, involved parties agree to compute something to discover aggregate data
- MPC is "the collaborative analysis of multiple siloed data sets that are never communicated nor trusted to any central authority or database"

Hoes MPC work?

Dr. Bestavros walked the attendees through an example of how the city of Boston utilized MPC to study the gender wage gap. The analyst has a question - what is the difference between salaries for men and women? The analysts gets two computers from two different organizations, BU and MIT, which are operated independently. Neither BU nor MIT have information about the actual salaries of employees in Boston; they receive random data.

- For example, if the salaries of males were $9, Boston gives BU the number 10 and MIT the number 11. 10 +11= 21, and if 21 was translated onto a clock, it would be 9.
- If the salaries of females was $7, BU gets number 4 and MIT gets number 3. 4+3= 7.
- To find the difference in wages, subtract the difference so 9-7=2. MPC would report that the wage difference in Boston is $2.

The above example is simple, but it can be more complex. Essentially, data is split into shares, distributed to computers, and combined for an answer. Other companies, such as Partisia [rate credit of farmers], Unbound [protect cryptographic keys], Boston University [equity in pay], Cybernetica [VAT audio], and Google [federated machine learning], have successfully used MPC. All that gets published is the answer to the question.

The technology itself is not new. The theory was developed in 1979, but no one cared to make it practical. With developments in technology and increased processing speeds, BU believes that it can be transformational. Dr. Bestavros invited the group to look up the video about MPC for more information.

Attendees were invited to ask questions. Attendees, including members of the public, asked if government agencies could use MPC to answer specific questions, such as DYS finding if a correlation exists between the number of nights youth spend in DYS facilities and outcomes down the road.

A member of the public asked if DCF and DESE could be providers, similar to BU and MIT. Dr. Bestravos said yes. It is also possible to use 10 providers instead of two; the more providers there are, the harder it is for anyone to find identifying information in the data.

Another person asked if MPC was similar to block chain. Dr. Bestavros said no; the purpose is different. Block chain is not about confidentiality.

Dr. Bestavros reiterated that MPC used to be a very slow process, but the gap is shrinking with advancement in technology. Now it has been commoditized, and they are trying to make it a service. They have open-source platforms available to use. MPC does not change anything about the agencies' operations. They will give you the plug-ins that will allow the participants to do the function. MPC has been referenced as a best practice in the context of the college spending bill at the federal level. Because MPC removes access to data, it provides a workaround for disclosure regulations. However, people do need to agree to the use of their data.

Someone asked about potential risks - what can go wrong? Dr. Bestavros said BU and MIT could collude. However, this can be avoided by having at least three parties and making yourself one of the parties. Dr. Bestavros also gave this example - say there are 100 banks and a hacker wants to steal information from all of them. Do they try to hack each individual bank, or hack a third party with all of the information? The third party becomes a single point of failure.

Dr. Bestavros explained that we need both transparency and exploration. Historically, this was a binary choice, but with MPC, there is a third way. Trust in the system is built through use. The BU center can do a proof of concept using synthetic data, which can be adapted to agency needs.

Ms. Carey asked a question about how much work needed to be done to set up questions and then conduct follow-up. Dr. Bestavros said that he can work with the group to figure out the first question, and then if the group agrees on a second question, they can answer it immediately. It is up to the group how far they wish to do.

Ms. Todd said that she was still not completely clear on the matching question. She asked how detailed of an analysis we could get using MPC - could it include cross-tabs and regressions?

Dr. Bestavros said yes. Dr. Varia said that if you take any complicated computer function, it's really just a fast calculator. We already know how to break it down into parts and transform into secure pieces. The only "exposed" variable is the answer. The individual parts become meaningless.

A member of the public asked how we determine that we are asking the right question or interpreting the data correctly. Dr. Bestavros asked how one would solve that problem without the technology. Ms. Todd said that the question is more about research design, validity, and reliability, which is beyond the scope of the technology. Ms. Threadgill asked the group to consider how much more powerful we could be if we spent more time on developing questions rather than trying to gain access to data. Often, researchers come with a filter, and ask for what they think they can get rather than what is needed to truly answer the questions. This technology could remove that filter. We can add layers of data to understand a range of outcomes.

Mr. Chandler said that DYS is very interested in trying this technology. They are interested to know how many of their youth are still enrolled in school and have debt. They need buy-in with another partner to do a proof of concept.

Ms. Averbach mentioned that two concerns are confidentiality and exposure of "real information". Mr. Chandler suggested starting small. Ms. Averbach said that we could completely mask the data. Dr. Bestravos said the proof of concept is about building confidence and getting comfortable with the technology.

A member of the public asked about computation time and data size. Dr. Varia said the speed depends data size and the complexity of the calculation. They have tested it up to 150 gigabytes. Dr. Bestavros said the first time they tried it, it took one day. It was difficult to coordinate schedules. It took a month to get the data, but about 10 minutes to do the computation.

A member of the public expressed concern about bias in non-matched names. They ran into an issue with hyphenated names, and it was a person who found the issue. What happens when you don't have eyes on the data? Dr. Bestavros said the use of synthetic data can help with this. Ms. Carey asked if there were any basic checks. Dr. Bestavros said that tsometimes, people type the wrong thing. In the first version, they assumed correct entry, and in the 2nd version, there were common sense checks. Dr. Varia also said that there have colleagues at BU that can help with issues of fairness.

A member of the public asked if they had used MPC with PHI data. Dr. Varia said that they have a grant to work with the BU School of Law to address legal issues. They have not had to consider HIPAA yet, as their healthcare work has been international.

Another member of the public asked about the cost of the implementation process. Dr. Bestavros said that agencies can produce a report in a spreadsheet, and the software will start

from there. If you want it on your computers, you would bring in people in terms of staff time. Mr. Azer states how BU has funding for MPC. In particular, DARPA is interested in making this usable and want to test it on nonpartisan issues. Ms. Threadgill also says that the OCA would like to help support a pilot project. The only current cost to participate for an agency would be staff time. CPCS expressed that they would be very interested in this programming if it will lead to better outcomes for clients. Dr. Bestavros offered to do multiple proofs of concept. Ms. Threadgill thanked the presenters.

**Fall Data Report**

Ms. Threadgill reminded the group that we have a report deadline coming up, and that the purpose of this report is to show the potential impacts of the criminal justice reform bill. The report includes system metrics, such as race and ethnicity data, in addition to aggregate data. Data that we do not have includes arrest data, arraignments, and adjudications.

Ms. Threadgill quickly review the overnight arrest admissions (ONA) data, as we discussed it at our last meeting. Applications for complaint have been decreasing over time, and this is also echoed in delinquency filings. Juvenile detention decreased by 28%. Based on data from one month snapshots, probation caseloads, which are related to delinquency, slightly increased pre-trial conditions.

Ms. Threadgill said that there has been a smaller decrease in DYS first-time commitments over the past two fiscal years. Ms. Todd clarified that all other slides were based on three-year data, which they are. Mr. Chandler said that the decrease is not a sudden drop.

Ms. Threadgill said that big picture, the numbers are all down. She then presented data related to specific statutory changes. The number of total arrests of youth under 12 decreased from 53 to 32. Ms. Threadgill noted that the 2018 data includes the months prior to the passage of the bill. Ms. Threadgill moved on to first offense low-level misdemeanors. While we cannot isolate the specific charges, the data shows that arrests, applications, filings, and commitments have all decreased.

Ms. Threadgill then shared the significant drops in school-based offenses from FY18 to FY19.

Ms. Threadgill reviewed the ONA data from last time and showed the large drops in school disturbances, alcohol, and property offenses. Ms. Carey wondered, regarding property crimes, if police officers may believe it is not worth filing if it is no longer a felony. Ms. Todd expressed interest in hearing from police on the matter; anecdotally, some police state that they no longer do anything with kids. Mr. Glennon says that he thinks numbers may rise again due to confusion, but Mr. Dohan said the numbers are consistent with a 10 year drop. Clarifying a question about the grid system, Mr. Chandler said that it is based on the adult system. In Grid 2, 47% of the population is in for assault and battery.

Regarding use of other community-based services, CRA filings have been stable, though truancy has increased slightly. There was some concern that if kids are arrested less, then there would be more CRAs, but this does not appear to be the case. BSAS has closed programs, and we do not have recovery high school data. Ms. Anjos says that DCF refers to BSAS, but there are waitlists. Ms. Threadgill said that she did not want to speak for DPH, but her understanding from Brian Jenney at BSAS is that demand is down and programs were operating significantly under capacity. Ms. Anjos said DCF sees a different picture that includes waitlists and staffing issues. Ms. Todd noted the staffing crisis; it is challenging to do this work for $14/hr. Ms. Carey suggested that this information might be good to put into the report. Ms. Threadgill said that most of the BSAS referrals come from non-court sources.

The last remaining slides are regarding race/ethnicity data. ONA arrests are down for white youth. Ms. Todd asked about the use of the term "non-white." Ms. Threadgill said that this is how the data was categorized when OCA received it from the courts. Ms. Todd noted that it was very outdated. Ms. Threadgill said that our best guess is that the courts collapsed everything into one category. Mr. Glennon noted that rates of non-reporting were very high, and Mr. Smith noted reliability issues with identifying race on police reports. Mr. Dohan asked if we could add a column that shows the percent of cases where race is not reported.

Regarding drug offenses, Ms. Threadgill noted that while not every child belongs in an inpatient program, there does appear to be two different systems for kids with substance abuse issues based on race. Mr. Dohan suggested including data from other sources, such as the Youth Risk Behavior Survey, to help counter stereotypes of who uses and sells drugs.

Finally, Ms. Threadgill noted that for first-time commitments, there was a 5% increase for Hispanic youth. Ms. Todd said that we have seen this every time since the start of JDAI. Ms. Carey asked if anyone has spoken with judges, which they have not. Ms. Threadgill said once she receives the outstanding data, she will turn the slides into narratives, and welcomes any feedback.

**Adjournment: 4:05PM**