



## Session Report

*Please know you may design the structure of this report to better suit the session. It's important to capture the key outcomes and solutions proposed for the future.*

**Session Title:** From Detection to Deterrence: Unleashing the Potential of AI in Anticorruption

**Date & Time:** Thursday, 09.12.2022, 1:30 pm – 3:00 pm GMT -5

**Report prepared by:** Abid Hussain Shah, Operations Officer, Prevention, Risk & Knowledge, Integrity Vice Presidency, The World Bank

**Moderated by:** Alexandra Habershon, Manager, Prevention, Risk & Knowledge, Integrity Vice Presidency, The World Bank

### **Panellists:**

- David Szakonyi - Professor of Political Science, George Washington University, Washington DC
- Marcelo Donolo - Head of Data Lab, Integrity Vice Presidency, The World Bank
- Veronica Borysenko - Development Manager, Dozorro, Transparency International Ukraine
- Randy Repka - Director, Global TechSprint Strategy, Alliance for Innovative Regulation (AIR)

### **Share the thematic focus of the session, it's purpose and corruption risks?**

The **primary focus of the session** is gaining expert insight to distill *what it will take to unleash the potential of AI for anticorruption*.

Key questions covered in the session include: What conditions do governments need to put in place to enable the use of AI within public administration? What capacities do oversight bodies need to develop to undertake advanced data

analytics? What policy responses might help address political economy challenges to ensure these capacities can be used for enforcement and accountability? What are pitfalls to avoid?

The primary focus, and questions above, are complemented and illustrated by experiences shared by the panellists who are familiar with some of the most promising avenues for data sciences in anticorruption. The panel discusses experiences, and early impacts, in designing and using advanced analytics and AI tools for anticorruption. The examples shared of advanced analytics and AI used by governments, civil society, and international organizations cover public procurement, asset declarations, property registries, international development financing, and data sets relevant to detecting grand corruption risks.

The panel discussion and audience inputs aim to enable participants to learn how we can collectively unleash the potential of AI and imagine future uses and applications.

This session presents the academia, NGO, MDB, and activist perspectives in this field.

## Summary of panellists' contributions & discussion points (please be as detailed as possible)

### **PANELLIST CONTRIBUTIONS AND DISCUSSION**

*The session started with a question for three of the four panellists soliciting illustrative examples from their work related to Artificial Intelligence (AI). The fourth panellist was then asked about their experience in change management working with government agencies and regulators.*

**Question for Veronika Borysenko, Marcelo Donolo, and David Szakonyi. Please give examples of a promising use of advanced analytics from your own work. In this example:**

**What was the data source/type?**

**What problem were you were trying to solve?**

**Did your solution employ advanced analytics, Machine Learning (ML), or Artificial Intelligence (AI)?**

**Were you able to solve the problem?**

**What is the potential impact of this approach in the anticorruption sphere?**

### **Veronika Borysenko - Transparency International (TI) Ukraine**

*Note: Veronika discussed two methods, one developing risk indicators and the second a machine learning (ML) software, as part of an effort advancing public accountability.*

**Example: Developed a risk indicator method and Machine Learning (ML) software in Ukraine.**

In Ukraine, cooperation between the state, civil society, and the private sector led to major public procurement reform, including a creation of a transparent procurement portal named Dozorro. The portal offers a wide range of data on all stages of the public procurement cycle. Before the war started, the system had started 400

procurements packages, generating significant data which was available to all.

Transparency International (TI) Ukraine focused on providing an accountability function, essentially oversight and analysis of procurement. To effect oversight and analysis, TI focused on four items:

1. Developed a community of 25 Civil Society Organizations to monitor public procurement at the regional level
2. Developed an IT instrument to monitor procurement
3. Increase the efficiency of public procurement with the state
4. Work with different stakeholders to develop IT instruments.

***Developing Risk Indicators:***

The IT instrument is separate from the public procurement portal. It is a feedback mechanism on public procurement which is used by CSOs. CSOs use the portal, identify issues, and file appeals with public bodies for processing. To aid efficiency, TI developed and incorporated 35 risk indicators in the portal to highlight the riskiest procurement for CSOs. Once indicators were created many procurement processes were flagged.

***Developing a Machine Learning software:***

TI had the idea to develop a Machine Learning software. To train the software, the software was presented with two tenders and had to identify which tender was riskier - the process was called "Tinder for Tender". Civil society activists collaborated – essentially, they and their expertise were crowdsourced - in teaching the software to identify riskier procurements, and the software had a high accuracy rate. However, the software development stopped due to a lack of funding and human resources. The organization aims to return to this software and improve it further. It is important to invest money in developing a tool like this as it enables you to scrutinize billions of dollars - this is great Return on Investment. It is also an example of how to train a machine to look for corruption.

**Marcelo Donolo - Head of Data Lab, Integrity Vice Presidency, World Bank**

*Note: Marcelo gave two examples, one on text mining from documents and the second on developing pattern recognition systems.*

**Example 1: Text mining from documents.**

This is a 3-year project involving Machine Learning and natural language processing. The project identifies and extracts relevant content from documents (e.g., Project Audit Reports) given to the World Bank. The process summarizes the funding of audits without reading the reports.

This is a supervised learning approach. We built our system as an MOP network that uses Machine Learning. If you use supervised learning, you need a human, which means you do not have an efficiency gain over time. The accuracy of this algorithm is 86 to 92 percent. We must be able to scale this up. We are setting up a deep learning system to be able to move from supervised to unsupervised learning.

**Example 2. Development of pattern recognition systems.**

Pattern detection is critical for preventive risk management work. We examine procurement data recorded by WB through the procurement process in its projects and extract information via a pattern recognition system. We have created data models and put them in digital visualization engines. We are trying to get patterns of bid rigging and shell companies.

We are using off-the-shelf software, building algorithms that build data models. You can program patterns you are looking for, e.g., the price differences between bids and bid estimates; shell companies; and frequencies of bidding of suppliers of interest in sectors and locations.

To build this, we followed these steps:

- (i). Look at patterns of complaints.
- (ii). Examine actual, past investigations to mimic findings.
- (iii). For preventive purposes, we ask: what are systemic patterns we see?

Essentially, we are building a tool that allows an investigator to not spend time going after data but to receive a data set that comes with analyses and does it in a way that can be a self-serviced type of tool.

The team is training the model by replicating what investigators do in their daily work. It is a good way to augment human intelligence and to find things because of large volumes of data.

**David Szakonyi - Professor, George Washington University (GWU)**

*Note: David gave an example on cross referencing efforts to confirm public data.*

**Example: Validating Public Data Registries**

Background/Disclosure data: Two million Russian officials disclose their income each year. TI Russia put this information in a database. The information is in 1000000 documents in Excel, PDF, and other formats. etc. The team built an open-source software, analyzed the structure of data and tried to come up with centralized format to store it in. The software used Machine Learning techniques, including pattern learning. The task allows us to feed any new document in the system and put it online with checks.

Cross referencing the disclosure data/the car analysis:

People lie on their tax/income forms, so submitted data needs to be cross referenced. We have designed an algorithm that draws on the make and model of every car manufactured. The algorithm finds the market value of these cars, then it takes the data on luxury car ownership from the car insurance portals, and then compares this to the disclosure data. This way, rather than finding red flags in the disclosure data, the team validates the data against external registries; this can be scaled for millions of officials.

The method can assist policy making decisions, governance initiatives, etc. It can set the ground for Machine Learning inspired tools. It is also a training dataset.

There is a question of enforceability. Is this method translatable elsewhere? Can this achieve greater accountability? We can do it where we can find the data; and can we transmit this method to other entities?

**Question for Randy Repka - Alliance for Innovative Regulation (AIR)**

**You don't work on data science specifically, or on anticorruption, until recently. Your area of expertise could be described as change management. Your organization has started working with the US govt to support public oversight authorities/ regulators leverage data analytics and AI for anticorruption. Can you tell us in general terms what needs your organization helps to meet in this context?**

Alliance for Innovative Regulation (AIR) has started working with US govt. The organization niche is demonstrating the value of technology, and these types of solutions, to the government. Specifically, we help government agencies and regulators become familiar with and understand the power of these technologies.

The issue: The global financial system is modernizing. It is important for government agencies to keep pace. AIR has a global remit and has seen progress in the space in the past few years - e.g., governments have innovation offices. We have sandboxes to test new technologies in safe environments. COVID has been especially beneficial in this regard, shedding light on the need for technology advances. We have identified four needs:

1. Need to understand the basics: Government regulators and agencies need to **understand** new technology. e.g., crypto etc. are foreign terms to many regulators.
2. Understanding how to deploy technology: Rather than just provide oversight functions, government agencies need to understand it to advance their own operations. This includes using technology to aid human efforts, understanding how to perform oversight and be able to review all the data available to an agency, for example. Often, agencies do not have the technology to review available data, and ML and AI can be effective in this case.
3. Cultural shift: In the case of regulatory agencies, change management is necessary.
4. Need for action: AIR does a lot of educating - writing white papers, and getting policymakers involved in actionable Text Sprints (p11) to demonstrate the value of technology.

## **Main outcomes of session (include quotes/highlights and interesting questions from the floor)**

**Q&A**

**In your work on Russian asset declarations, were you able to expose corruption and was its impact? Did it help get officials listed on OFAC list?**

**David Szakonyi - Professor, GWU**

We are trying to get data out. In our experience, Russian corruption investigations started by CSO, do not lead anywhere. Prosecutions in country are usually targeted against certain persons. That is the politics of anticorruption.

**Is your system able to scale across different countries?****David Szakonyi - Professor, GWU**

All the code is on GitHub so it depends. You will have to write new code for different countries as you are solving other problems along the way. For example, you are deciding on formats and compatibility. I would put less emphasis on getting coders. First, you need a large workflow approach to anticipate problems as they occur.

**How do we take these tools to developing countries? We need to transition tools to Spanish and other languages as it will explode capacity.****Marcelo Donolo - Head of Data Lab, Integrity Vice Presidency, World Bank**

World Bank Integrity Vice Presidency works with data but it has limitations because of confidential nature of information (in audits, etc.) Consequently, it makes sense for to join forces and solve problems.

All of the Integrity Vice Presidency work doing so far has been in English. We will have to deal with the problem of multiple languages. We are trying to build up a community of practice; there is a lot happening right now and we are all trying to solve the same kinds of problems. Data science is new, so it takes to time to organize around issues and determine which issues to tackle first.

**We have been working with data to promote openness. We have been trying to introduce a multi-stakeholder approach emphasizing collaboration, not confrontation. We try to instill a philosophy of dialogue between actors across a broad range. What are you specific experiences if you are trying to promote a collaborative approach between government and civil society.**

**Randy Repka - Alliance for Innovative Regulation (AIR)**

A multistakeholder approach looks similar to the approach of Open Gov Partnership. It is the only way to get work done in some contexts. You have to figure out how to bring civil society in as monitors and ensure that public data is validated by experts. How do we know that the data is reliable? Open Government Partnership does this via co-creation and so avoids confrontation. It tries to close gaps with findings that researchers come in.

**Veronika Borysenko - TI Ukraine**

Cocreation and cooperation was essential to our work, as we got a lot of information from risk indicators. Our cooperation with civil society, government, and private sector gave us the space to develop risk indicators for state audit service.

**How do you shape strategic interventions at the early stages of an activity?****Marcelo Donolo - Head of Data Lab, Integrity Vice Presidency, World Bank**

We have been building **pattern detection systems**. In this exercise, we have to determine the type of cases we use to pursue to test hypothesis. We leverage the knowledge of teams within the World Bank integrity Vice Presidency to understand the risks in specific geographies. We work on developing a hypotheses and several types of use cases to build data models. You need a Subject Matter Expert (SME) to talk closely with data team to make the team understand how fraud and corruption issues work into pattern recognition processes.

We have 10 use cases, where we examine what is working, and determine the level of granularity we are looking at. The experience of SMEs helps as we are trying to look at systemic risks that have been a problem. The issue is how to detect this via the data. Is it enough to prove a risk exists in procurement space? One way to think about how to structure an approach to detecting integrity risks.

In the space of government ethics we can use advanced analytics to look at outliers to see where to invest in training.

**David Szakonyi - Professor, GWU**

You have to ask: do we have the right data? You have to go back and keep digging in - what were they doing. Having high quality data through these techniques is the first step; forensics is a good input in this regard.

**What should we keep in mind before algorithm goes to stage/launch? We are in the last mile of developing a procurement monitoring system.**

This is different from a data issue. What is the process for making decisions once high-risk tenders are identified when AI system identifies risks?

**When you perform a Text Sprint (p11), how do you structure the phase after ideation?**

**Randy Repka - Alliance for Innovative Regulation (AIR)**

Part of a sprint design is what happens after the sprint. There are a few options.

Option A. Use open-source software.

Option B. Identify incubator programs so winners are fed into cohorts. The programs can match the subject of sprint. This is useful for social enterprise solutions.

Option C. A partner can take the learning solution and can put it out to RFP. In this case, you must make the decisions upfront.

**When we talk about data, accessibility of data for end users is the challenge. What are they doing in their different spheres to make data more accessible and simpler?**

**David Szakonyi - Professor, GWU**

The TI Russia team commissioned a sociologist with a 30-question instrument. The interviewer talked to people across the spectrum to see what they were getting out of the TI Russia site - and consequently the user interface changed. It is a conversation with users.

**What are the mistakes you have had in the field?**

**Marcelo Donolo - Head of Data Lab, Integrity Vice Presidency, World Bank**

We use the hypothesis approach, and when using it you have to ensure have to address assumptions thoroughly. Address all the doubts and the law of unintended consequences.

**We are a social enterprise and have difficulties with funding and resources. How can we ensure long term and sustainable funding and partners, and how do we select the right partners?**

**Veronika Borysenko - TI Ukraine**

For CSOs it is difficult to get funds for ML software. The solution is to work closely with the government. We can get the funds but the government will have sustainability, especially if the government is committed to doing this (but it needs political will).

In Ukraine we have CSOs but we developed their capacity, conducting events, gathering them and sharing methodologies to build their capacities.

## Key recommendations for the future and concrete follow-up actions

**LESSONS LEARNED AND RECOMMENDATIONS**

*The moderator asked a question of each panellist to elicit lessons learned and recommendations.*

**Question for Veronika Borysenko - TI Ukraine**

**Procurement oversight in Ukraine has earned a well-deserved reputation for breaking new ground in the use of advanced analytics for monitoring corruption risks in public procurement. What did it take to develop this capacity? Looking at the results, is there anything you would consider doing differently? What lessons from this experience would be helpful for other public authorities?**

Our lessons are:

1. Digitalization of data: Once you perform this you need to understand the different risks at each stage of public procurement process. You need different IT instruments and analytical tools for each risk.
2. Constantly improve developed tools: We realized we need to constantly improve the ML instrument. This also meant that it was important to continuously use the ML software.
3. Engage communities: Our community comprised up to 25 CSOs which monitor public procurement in different regions of Ukraine. These CSOs helped us improve the quality of ML software.
4. Sustainability is everything: We develop this tool but we also need to understand that these kinds of tools require a lot of funds and human resources. You will need to invest resources and money so that system functions, and you will need to keep tinkering for improvements, etc. At the end you will get a lot of results and will be able to save public funds.

**Q. When a suspicious tender is identified. Have you been able to track impact of work?**

We do not have all the time to track all the appeals, but we have examples when our appeals ended starting criminal investigations. There are 70 criminal investigations. It is very difficult to track but there must be political will, which is still an issue, to start an investigation.

**Question for Marcelo Donolo - Head of Data Lab, Integrity Vice Presidency, World Bank**

**You lead the Data Lab in the Integrity Vice Presidency at the WB. This is still a fairly**



**new development in the WB - you are actively leading the creation of new tools and processes. What are some takeaways from your experience developing the unit and approaches or new tools that would be helpful for other multilateral institutions or public authorities thinking about doing the same?**

***Creation of a data team is critical.***

If you are going to start working on advanced type of analytics you need to centralize data function within the unit. You need a dedicated team. In the recent past at the WB, there were no data teams outside an IT unit. The WB's integrity vice presidency is one of the first teams to place a data team inside its unit with the need to manage data differently to accelerate the experimentation and deployment of new ways of leveraging technology. This team is called the Data Lab.

***For a data lab, the skill mix is important.***

We have a small data lab of 3 to 6 people, supporting all the integrity functions from prevention to compliance.

The skill mix needed is:

Understand technology, and be able to code in several languages. Someone should be able to do data science work as a data scientist is different. The team needs subject matter experts (SMEs) in all areas important to lab work. The team also needs a coordinator, with experience of the corporate environment. For example, since the WB has its own central IT department, the data lab needs to collaborate with the central IT team (the lab cannot function in a silo).

***Strategic role of data lab.***

The team needs to also have a specific role to meet the unit strategic objectives. There needs to be an understanding of what infrastructure is needed to be in place to make it work, and how to deploy solutions. There also needs to be a strategic understanding as to how to scale up from using a laptop and server to the institution. The team needs to also work with institution as the whole, and to make sure it can do these things securely.

***Space for creativity.***

A lab needs to have space to explore, test, and fail. The data science field is evolving so fast that by the time you figure out what to do something better and new comes along.

***Infrastructure.***

Data science works at scale. To scale you need robust infrastructure, and to leverage cloud computing. The data lab built a cloud infrastructure within our unit to be able to scale, and used technology to automate the process from collection to output. Hence, thinking about infrastructure needs to work at scale is important.

***Key point.***

Invest in creating teams with right profiles as well as invest in infrastructure.

**A two-part question for David Szakonyi - Professor, GWU**

**As an academic your research has relied on working with open data. Can you talk us through some of the key requirements to do this work (data quality being an**

**obvious one) and perhaps more crucially share your thoughts on what it will take for this work to have an impact in terms of enhanced accountability or anticorruption enforcement?**

***Requirements for open data work.***

We are dependent on data available to us, and our nonprofit struggles to find data sets we need to answer our questions.

We also need to evaluate policies and examine references to financial flows and corruption in these policies. For example, these include accessing basic data from the US - even income and asset disclosures in government. When we access data (when it is shared with us, for example) the government gets the utility and we surprise it with the results and this enables us to combat money laundering in public and private sectors.

***Requirements for Beneficial Ownership registries.***

Governments are realizing we are doing the work they should be doing and doing it more creatively. My plea is we get the government on board with the idea for further insights. Not just to uncover corruption, but to go beyond - we need the data.

Regarding Russia, a lot of our tools are mimicking what the Russian government is making. The tools are not that unique and government are not unaware of them. In the US we work with media - we have seen low adoption by government of techniques advocated. We are sharing methods and information but through journalists. The model of cooperation with government does not always exist.

**Question for Randy Repka - Alliance for Innovative Regulation (AIR)**

**Can you share with us some examples from the work your organization has been doing with national authorities, and in partnership with the US government, to support innovation within public authorities/ regulators? What are some of the key lessons for organizations in our audience today in creating space and capacity for innovation?**

***Organize a Text Sprint.***

It is similar to a hackathon but with unique differences. In a hackathon we just use tech people. In a Text Sprint we bring together disparate parts of ecosystem - policymakers, subject matter experts, academics (all unlikely collaborators) - to focus on a single product. In these sprints we work with a government partner. Last year's Text Sprint was with the law enforcement department of the US State Department. Demonstration that work is being done.

To do a sprint, you need to plan. First, ask for the (partner) state department objectives. In this instance, they were looking to foster multilateral cooperation, and wanted to use the sprint to demonstrate power of tech in the space.

We worked on asset anticorruption solutions for emerging technologies. The sprint aimed at the role of multilateral cooperation; 73 different countries were represented. We brought on board experts and judges, and had speakers to educate a broader population. The sprint had 5 teams, with each team bringing a different

solution.

As an example, one solution, called Chains of Corruption, focused on a coalition for post disaster recovery. The solution promotes transparency for dispersal of funds. It uses blockchains to make dissemination transparent, with funds dissemination upon meeting demonstrated milestones.

A lesson learned is: when you do idea development, you need to think of users of solutions and prepare, in the Text Sprint, to lay out what happens next to further develop solutions and scale them up.

## What can be done to create opportunities for scaling up the solutions discussed in the session? And by whom?

**The discussion touched on the following areas where initiatives could be scaled up:**

**From Supervised to Unsupervised Learning: Text mining from documents (p3).**

Currently, the World Bank Integrity Vice Presidency is text-mining a range of Bank project documents. The machine learning method uses a supervised learning approach which involves humans which restricts efficiency gains over time. The team is setting up a deep learning system to move from supervised learning to unsupervised learning.

**Validating public data registries by cross referencing with other databases (p4 and p6).**

David Szakonyi's team has designed an algorithm that cross references luxury car insurance data against tax disclosures of Russian officials. This effort and method can be scaled up.

To scale up to different countries, however, a different approach is needed. The code is on GitHub, but a new code needs to be written different countries as in each effort new problems need to be solved (e.g., formats and compatibility). Scaling up relies less on coders than on a methodical workflow approach that anticipates problems.

**Data labs or units within an MDB context have potential to scale (p9 and p10).**

To be able to scale the work of an institutional data lab, and understanding of technological infrastructure needs (e.g., cloud computing) is critical, as is an understanding of the institutional landscape and structure.

**When organizing a Text Sprint, plan for eventual scale up (p11).**

When you do idea development for a Text Sprint think of users of solutions and prepare to lay out what happens next to further develop solutions and scale them up.

**Is there a specific call to action to key stakeholders, such as governments, businesses, funders, civil society, young people,**

**journalists or any other stakeholder that should be noted?  
Please specify if relevant.**

There was no specific call to action. The session, however, explored exciting possibilities in technology and Artificial Intelligence that can be leveraged in the fight against corruption.

**Rapporteur's name and date submitted**

**Abid Hussain Shah, 12.12.2022**