Project Name: Saskatchewan Police Predictive Analytics Lab Missing Persons Project

[CSSP-TI-2017-2245]

Project Progress Report (PR1)

Reporting Period: July 25, 2017 to October 27, 2017

Date of Report: November 3, 2017

Project Status:

Since the date of MOA award (July 25, 2017), the Saskatchewan Police Predictive Analytics Lab (SPPAL) Missing Persons Project Team has made significant progress towards deliverables outlined in the Defence Research and Development Canada (DRDC) Project Charter.

First, a project team complete with Project Leads from the Saskatchewan Ministry of Justice (MOJ) and Saskatoon Police Service (SPS), as well as three Principal Investigators (PIs) from the University of Saskatchewan (U of S), and project supports (finance, legal, and project management) has been created.

A project governance model for this project/team has also been finalized by the joint Project Champions, Dr. Rector, MOJ and Deputy Chief Bent, SPS. Project agreements (e.g., data sharing) are currently being finalized and all identified project personnel have successfully completed SPPAL security clearance procedures.

A schedule of regular project meetings has been established and work has commenced in earnest on project deliverables 1 through 6 inclusive, and 8 (as identified in the Work Breakdown Structure on pages 12-14 of the MOA with Public Works and Government Services Canada or PWGSC).

Given the delayed project start date and interconnected nature of project deliverables, the Missing Persons Project Team has commenced work on both immediate (1-4) and intermediate (5-6 and 8) project deliverables simultaneously.

The first four project deliverables were to have been completed by October 31, 2017. We are currently finalizing deliverables 1 through 3 inclusive, for submission to DRDC/PWGSC in November 2017. Progress updates on deliverables 1 through 5 inclusive, including projected time to completion, are reviewed below.

Project Deliverables:

Deliverables #1 and #5 - Data Integration Workshop; Expansion of SPPAL Development Environment

Progress on these related deliverables build on discussions had with private industry 17(1)(d) and technology partners at the U of S and include the completion of a needs assessment for SPPAL - for both the lab in general and the Missing Persons Project.

Specifically, Dr. Osgood and his team of highly qualified persons (HQPs) from the U of S have engaged with the Technology Services Division at SPS, where SPPAL is housed, to explore technologies that will permit the lab to perform interactive and exploratory analyses of large volume data in a flexible, efficient, and rigorous manner within a secure data environment that protects the confidentiality and privacy of underlying data.

Technologies currently being explored for further SPPAL development include: metadata curation and updates (e.g., maintenance of up-to-date sustainable documentation to facilitate analysis), SPPAL infrastructure configuration (i.e., computational service providers with network and human protocols for conducting work), server configuration (e.g., internal data and analysis servers, 14(1)(m) [redacted]), programming technologies and supporting libraries (e.g., Java, Python), data analysis, simulation, and visualization software (e.g., AnyLogic, Tableau), and advanced security mechanisms (e.g., technical "DMZ" with firewalls, "containerized software" such as Docker). Plans for experimentation and configuration tuning for lab optimization and expansion are also being explored.

The constituent elements of Deliverable #1 including: i) phased roadmap; ii) architecture design document; iii) security summary document; and iv) vision scope and requirements document, are currently being finalized. Documentation will include a proposed "resources-service map" as well as "data flow and protection" diagrams which include planned security and quality control mechanisms (e.g., plans for user access control, account management, audit trails).

Estimated date of completion: November 2017

Deliverable #2 - User Interface Needs Assessment

Dr. Spiteri and his team of HQPs at the U of S have begun the user interface assessment process beginning with the SPS Missing Persons Unit. Frontline officers have provided input regarding their day-to-day work, potentially useful variables and risk factors as related to missing persons, and possible elements to be incorporated in applied tools.

Preliminary interfaces have now been developed and specialized code is also being written so as to incorporate future statistical models. Potential plans to demonstrate these computer interfaces to additional community safety partners (possibly as part of Deliverable #3 - Missing Persons Workshop) are currently being considered.

Estimated date of completion: November 2017

Deliverable #3 - Missing Persons Workshop

A draft agenda and participant list for this workshop is currently being finalized with plans to hold this workshop at SPS in November 2017. The workshop is an opportunity to consider relevant risk factors from a multidisciplinary perspective to inform work related to development of risk models (e.g.,

Deliverable #6), learn about current processes used by community safety partners to prioritize and respond to missing persons cases and share preliminary work from the Missing Persons Project (Deliverable #2) for feedback. Representatives from police, social services, non-government community-based organizations, and academia will be invited presenters and participants.

Estimated date of completion: November 2017

Deliverable #4 - Data Coding (Police Data)

Work on the creation of a living data dictionary for police missing persons data is ongoing.

Estimated date of completion: December 2017

Additional Updates:

The SPPAL Missing Persons Project has generated considerable interest and excitement among police and community safety partners in Saskatchewan (SK). Principal Investigators from the U of S and HQPs, including leads from SPS Technology Services, are exploring innovative and cutting edge security technology for the Lab development environment. Members of the Missing Persons Project Team have also provided several invited talks in this reporting period (e.g., Prevention Matters Conference, Provincial Child Death Review Research Sub-committee).

The SK MOJ and SPPAL, including members of the Missing Persons Project Team, have also partnered with Mr. N. Taylor and the Canadian Association of Chiefs of Police - Information and Communications Technology Committee (CACP-ICT) to host an Open Analytics for Community Safety and Wellbeing Conference. Current discussions with CACP are looking to finalize dates for this conference in April 2018 in Toronto, Ontario. Conference goals and objectives include establishing a network for community safety analytics research and practice in Canada (Deliverable #8). Researchers embarking on similar analytic work in Ontario in the area of criminal intelligence have also recently requested a site visit to SPPAL in late November - early December 2017.

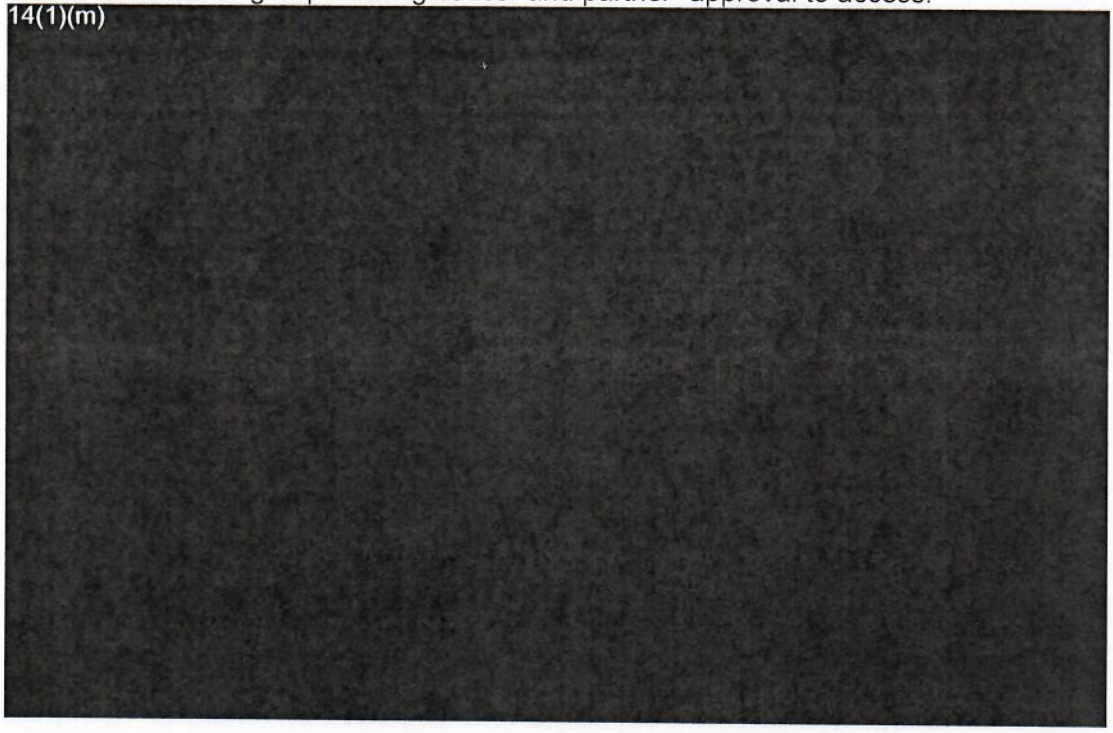
Data Integration Workshop and Development

(Deliverable 1; Part iii)

Privacy/security summary

Many of the essential laboratory privacy and security mechanisms originate in a combination of rules extending from sources: 1) In its capacity as the host organization, Saskatoon Police Service 2) Other partner organizations, and by the SPS in its partner capacity 3) From involved institutional ethics boards (on a project-by-project basis). Many elements of such privacy characterized elsewhere in this document. Grouped, they are as follows, with many accompanied by additional commentary.

- The laboratory technology shall prevent removal of any data in electronic form from its security premises without express approval of an appointed officer approved by the SPPAL oversight committee. The security checks will confirm compliance of the data proposed for removal with stipulations extending from any of the groups above, including those associated with institutional ethics board approvals for the particular project. These may include rules related to presence of purely data aggregated over sufficiently large periods of time, geographic regions, age ranges or other aspects of heterogeneity. It also involves handling or elimination of small cell sizes. Such approved processes will, in a growing number of occasions as time passes, support export of data for operational use by partners such as the Saskatoon Police Service, Ministry of Justice and Ministry of Social Services.
- Data provided by partners shall be accessible in decrypted (plain-text) form only to validated members of groups having ethics- and partner- approval to access.



14(1)(m) [redacted]

- Data shall remain in encrypted format until time of use, so as to allow system administrators to manage the data files without compromising the security of the data.
- 14(1)(m) [redacted]
- The individual members of the group shall be physically verified to be within the appropriate groups prior to getting access to the data in the bed.
- As a corollary to the above, all sensitive data provided by partners shall be unable to have decrypted or otherwise accessed in plain-text even by SPPAL system administrators, staff and officers, unless such officers are official members of the appropriate groups.
- There are to be no printers or other devices to create removal physical reproductions of the electronic data present in the SPPAL.
- Any backups made by SPPAL data shall be made on encrypted data stores and will be transmitted for archiving in an encrypted fashion. Such backups will then not be accessible to system administrators during any restoration processes.
- Physical security for all individuals accessing SPPAL shall first be enforced by standard SPS mechanisms. This includes location-specific access cards issued following background checks, fingerprint recording, oath taking and other processes.
- All individuals making use of SPPAL facilities within the lab shall be required to sign in and out of the SPPAL.
- The SPPAL shall record detailed auditing information on laboratory use, including -- but not limited to -- logs of log-ins and log-outs, access to particular computational packages (e.g., containers), queries executed on databases, command histories of analysis, all submissions and pipeline results for continuous integration pipelines.
- Individuals operating within the SPPAL shall have write access only to project-specific electronic resources within the laboratories.
- Sensitive data within the laboratory shall be stored in encrypted volumes, with the encryption key only available to those operating within the project.
- While a given user may be a member of multiple project groups, a given log-in to the lab shall take place within the context of a given project, and only provide access to data for that project. This restriction will further prevent such individuals from cross-linking different datasets.
- 14(1)(m) [redacted]
- 14(1)(m) [redacted]
- 14(1)(m) [redacted]

- 14(1)(m) [redacted]
- 14(1)(m) [redacted]
- As required by Research Ethics Board and partners requirements, a variety of automatic de-identification mechanisms shall be applied on the unencrypted data, ranging from the most simple (separation of a master list from the de-identified data, with de-identified data using scrambled unique identifiers that can be cross-linked to the master list) to more refined (such as use of cryptographic salts).
- 14(1)(m) [redacted]

DRDC MP/SPPAL PROPOSAL - FINANCIAL INFORMATION AT A GLANCE FOR 2 YEAR FUNDING CYCLE

Specific funding earmarked in proposal for Financial, Legal, and IT Support:

$10,000	Finance
$20,000	Privacy/Legal
$20,000	IT/Installation/Support
$50,000	

Funding for development of Lab (architectural design, software, and data integration platform):

$322,500	External Consultation (Horton/Yoppworks)
$140,400	Software
$462,900	

Other IT related funding:

~$195,000	HQPs in Computer Science
$15,000	SPS Data Coding
$10,000	Internal Review (of developed Missing Persons user interface prototype)
~$220,000	

TOTAL INVESTMENT INTO SPS: ~$512,000 to $732,000 over 2 years

-**$50,000** to approximately **$220,000** available for direct project support and technological services

-**$462,000** committed for Lab development including cutting edge technological services and software

-Funding provided over a two year funding cycle, with funding being heavily weighted in the first year

-Networking and professional development opportunities are also available (e.g., working with specialized external technological consultants, workshop/conference attendance, funds for travel)

TOTAL DRDC FUNDING PROVIDED OVER 2 YEARS: $937,900

Project Name: Saskatchewan Police Predictive Analytics Lab Missing Persons Project

[CSSP-TI-2017-2245]

Project Progress Report (PR2) – Year 1 Reporting

Reporting Period: July 25, 2017 to March 31, 2018

Date of Report: March 29, 2018

Project Status:

Since the date of MOA award (July 25, 2017), the Saskatchewan Police Predictive Analytics Lab (SPPAL) Missing Persons Project Team has made significant progress towards deliverables outlined in the Defence Research and Development Canada (DRDC) Project Charter.

First, a project team complete with Project Leads from the Saskatchewan Ministry of Justice (MOJ) and Saskatoon Police Service (SPS), as well as three Principal Investigators (PIs) from the University of Saskatchewan (U of S), and project supports (finance, legal, and project management) was created.

A project governance model for this project/team was also finalized by the joint Project Champions, Dr. Rector, MOJ and Deputy Chief Bent, SPS. All identified project personnel have successfully completed SPPAL security clearance procedures and supporting project agreements (e.g., data sharing) are currently being finalized.

Given the delayed project start date and interconnected nature of project deliverables, the Missing Persons (MP) Project Team commenced work on both immediate and intermediate project deliverables simultaneously (e.g., expansion of the SPPAL development environment and MP user interface development). 17(1)(d) [redacted] A Contactor Report highlighting work completed to date was also submitted to DRDC on February 28, 2018. This report is included in Appendix A.

The majority of tasks comprising Milestones 1 through 5 inclusive as identified on page 4 of the MOA with Public Works and Government Services Canada (PWGSC) have now been completed. A Missing Persons Workshop (Task 3) has been planned and will take place on April 6, 2018. Otherwise, tasks comprising the fifth and final Milestone for Year 1 (Task 6 and 11) have been completed and are currently under review by the joint Project Champions. There are plans to submit these by the end of the day today (March 28, 2018) or immediately following the holiday weekend (April 3, 2018). The CSSP Progress Report (PR3) will be submitted by April 6, 2018.

In conclusion, the SPPAL Missing Persons Project has generated considerable interest and excitement among police and community safety partners in Saskatchewan. Principal Investigators from the U of S and HQPs, including leads from SPS Technology Services, are exploring innovative and cutting edge security technology for the Lab development environment. Members of the Missing Persons Project

Team have also provided several invited talks in this reporting period (e.g., Prevention Matters Conference, Provincial Child Death Review Research Sub-committee).

Finally, the MOJ and SPPAL, including members of the Missing Persons Project Team, have also partnered with Mr. N. Taylor and the Canadian Association of Chiefs of Police - Information and Communications Technology Committee (CACP-ICT) to host an Open Analytics for Community Safety and Wellbeing Conference. This conference will take place on April 23-25, 2018 in Toronto, Ontario. Conference goals and objectives include establishing a network for community safety analytics research and practice in Canada consistent with Year 2 project goals.

Appendix A: Contactor Report

2018-02-28

CSSP-TI-2017-2245

Produced for: DRDC CSS Executive

Contractor Report

Saskatchewan Police Predictive Analytics Lab Missing Persons Project

Background

In November 2015, the Saskatchewan Police Predictive Analytics Lab (SPPAL) was established through a partnership between the Saskatchewan Ministry of Justice; Corrections and Policing (MOJ), the University of Saskatchewan (U of S), and the Saskatoon Police Service (SPS). Municipal police agencies in Saskatchewan, RCMP "F" Division, and the Saskatchewan Ministry of Social Services, Child and Family Programs have all agreed in principle to share information with SPPAL. The Canadian Safety Knowledge Alliance (CSKA) is also a SPPAL collaborator. Potential collaborations with other jurisdictions are also being explored.

A primary objective of SPPAL is to increase capacity to integrate and examine large amounts of police and community safety data in order to develop predictive models and applied tools that can assist police and partner agencies to intervene effectively to reduce risk and promote community safety. In other words, SPPAL is intended to be an operational lab working with identified data from multiple data sources which is required for police and partner agencies to intervene to reduce risk in high risk cases and situations.

These developments provided the foundation to explore SPS missing persons data with special emphasis on vulnerable and/or marginalized groups who may be at increased risk for harm including, children in care, "habitual runaways," and Indigenous persons reported missing in the province of Saskatchewan (SK). Additional SK municipal police data and SK child welfare data will be integrated with existing data as key technologies representing project deliverables are further advanced (e.g., methodologies to support secure data transfer and linkage). Ultimately, advanced predictive analyses and development of computer interfaces will assist SPS and other public safety and social service agencies in responding to a high volume of calls for service and developing coordinated interventions for persons reported missing.

Approach and Methodologies

Applied community safety analytics requires ongoing engagement with police agencies, subject matter experts and highly qualified persons (HQPs) in the areas of policing, mathematics, statistics, computer science, and the human services (e.g., psychology, social work). A SPPAL Missing Persons (MP) Project Team complete with Project Leads from the MOJ and SPS, as well as three Principal Investigators (PIs) from the University of Saskatchewan (U of S), and project supports (e.g., financial and legal services) was created.

Major tasks to be completed by the MP Project Team over the 2-year funding cycle for this project are depicted in *Figure 1* on the following page.

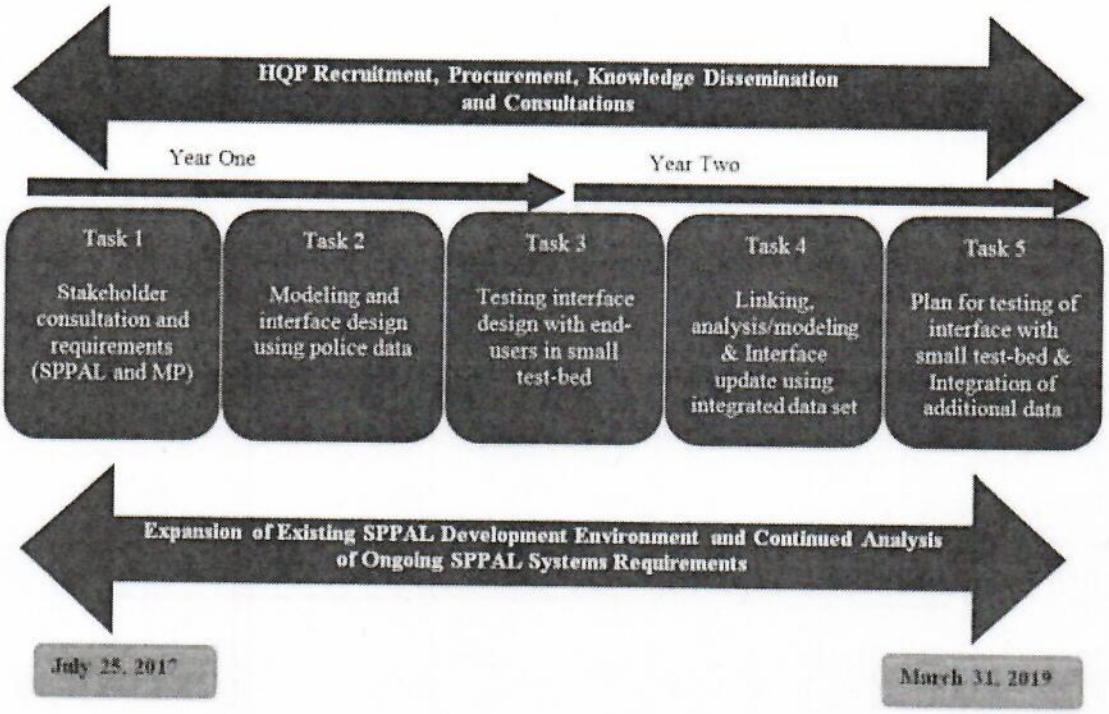


Figure 1: Major tasks to be completed within the scope of the SPPAL MP Project

An important first task was to complete an updated assessment of systems requirements in the areas of privacy, security, data sharing, hardware, software, data integration, data streaming, human resources/HQPs, and potential future needs for SPPAL in order to support the MP Project Plan and achieve project deliverables. These results are currently being used to create a data development environment for the MP project that can support analyses with municipal police data in order to develop and test advanced predictive models for MP cases and also build an integrated database containing information from collaborating agencies complete with requisite support tools and structures.

The creation of applied tools for end users (e.g., police and community safety partners) has been simultaneously occurring alongside the development of the data environment described above. This has included an initial assessment of user needs (e.g., municipal police) to inform the development of computer interfaces incorporating analytical models for purposes of communicating results to frontline community safety partners. A follow-up workshop has also been planned with representatives from police, social services, non-government community-based organizations, and academia. This will further assist the MP Project Team in understanding the processes and practices currently used by community safety partners to prioritize and respond to missing persons cases and considering relevant risk factors to be considered for use in advanced modeling from a multidisciplinary perspective. Members of the Project Team will also be invited to share work completed to date (e.g., preliminary computer interfaces) and obtain feedback.

Testing, evaluation, and updating the user interface using models developed on police and enriched data sources (e.g., social services data) will occur in Year 2. A proposed framework for evaluating applied tools/prototypes and updated requirements analyses will also be completed.

Preliminary Results

Work completed thus far on the SPPAL MP Project provides a solid foundation for developing predictive models and applied tools for police and community safety partners to inform decision-making, guide preventative interventions, and assist in reducing time missing and associated negative outcomes for persons reported missing.

Initial projects tasks have been completed or are nearing completion (see Tasks 1, 2, and 3 as depicted in *Figure 1*) and include the following key deliverables: development of a phased roadmap for further development of the SPPAL data environment, architecture design documents, proposed privacy and security mechanisms, preliminary user interface needs assessment and development, planning for multidisciplinary stakeholder workshop, development of living data dictionary, consideration of relevant data coding standards, exploratory data analysis, and expansion of the SPPAL environment. Key findings are presented below.

Phased Roadmap for SPPAL Development

The proposed roadmap for laboratory infrastructure and evolution is as follows:

1. Initial laboratory establishment with no external access or access to external data stores. Offers encrypted project-specific data access, verification of ethics and authentication processes, Docker container-based dependency and security management.
2. Expansion of SPPAL to include secure DMZ, with security ensured via multiple levels of firewall, including software and hardware (router) based mechanisms. The resulting system will support external support of Docker-based analysis containers via cryptographically secure lab-specific gitlab instance, and access to external social media feeds and selected government-based servers.
3. Creation synthetic datasets adhering to schemas of underlying data sources but which avoid privacy and confidentiality concerns by use of artificial data. Larger investments in this area will support capturing key statistical and structural patterns within the data.
4. Creation of mechanisms for automatically validating data structures proposed for export, thus easing manual verification.
5. With an eye towards further security and support cross-linking databases, further investigation of the following:
    a. Programmatic separation of (identifiable) master-list from rigorous de-identified data for analysis.
    b. Exploration of homomorphic encryption for the laboratory, which will allow an important subset of analyses to be performed on the encrypted data. We will explore the use of subsets of such encrypted data outside of the laboratory.
    c. In context of non-encrypted data, implementation of privacy-preserving data mining infrastructure.

Proposed Architecture/Design

Consistent with the roadmap described above, the design and development of the SPPAL data environment has involved investments in technologies focusing on secure efficiency, scalability, performance, and robustness. These investments will permit the lab to perform interactive and exploratory analyses of large volume, flexible, independent, robust and systematic analyses in a more timely fashion. These expansions also support a more secure and autonomous environment by lessening vulnerability to outside attack. Critically given the nature of the data being analyzed, such advantages are being secured in an environment which provides strict and rigorous guarantees concerning the confidentiality and privacy of the underlying data.

Metadata curation and updates

Maintenance of accessible, informative and up-to-date sustainable documentation will serve as an important facilitator for analysis within the SPPAL. Because of the secure and isolated nature of the SPPAL intranet, it is important that such documentation be available offline while within the laboratory itself. Given that the SPPAL infrastructure -- with the sole and specialized exception of the external work package submission system -- is not accessible from outside the SPPAL, and that public hosting of SPPAL metadata is inadvisable, it has been proposed to use the popular offline documentation system Zeal to view lab metadata and documentation, with appropriate docsets available only to SPPAL associates. For any subsets of such documentation with particularly restrictive guidelines, it has been proposed that the documentation be available via a lab-internal document-based HTTP server.

SPPAL infrastructure configuration

The SPPAL Infrastructure plan underway makes central use of computational service providers (virtualized servers), and a set of network and human protocols for conducting work. This plan is discussed and depicted below (see *Figures 2 and 3*).

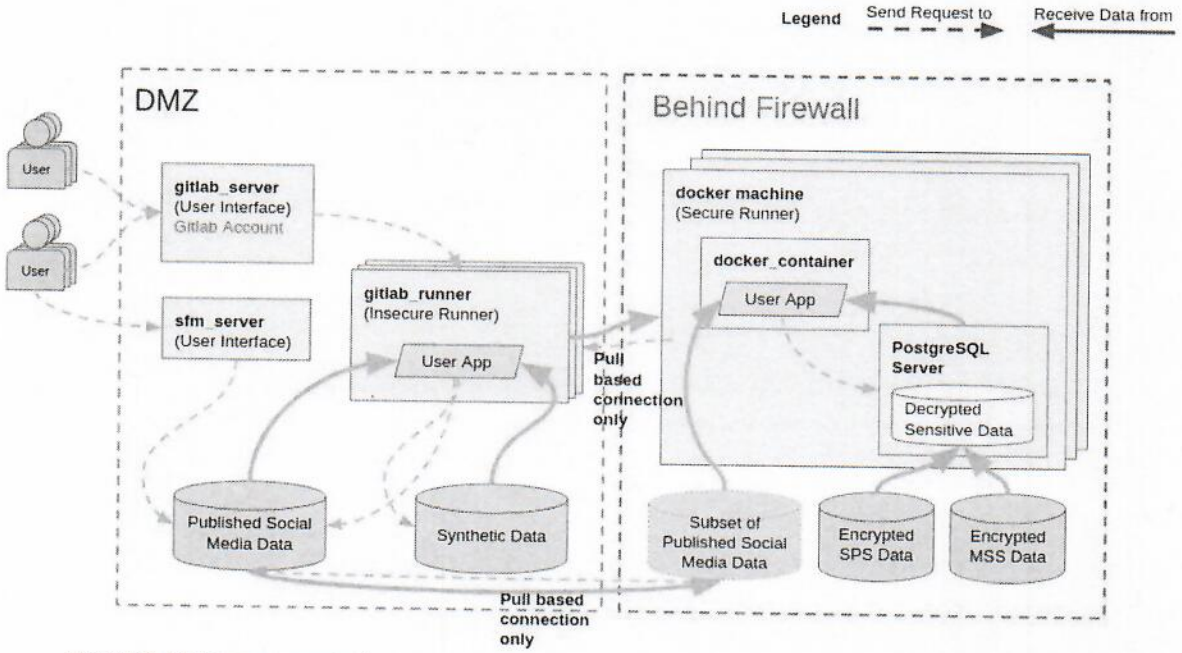


Figure 2: User access control

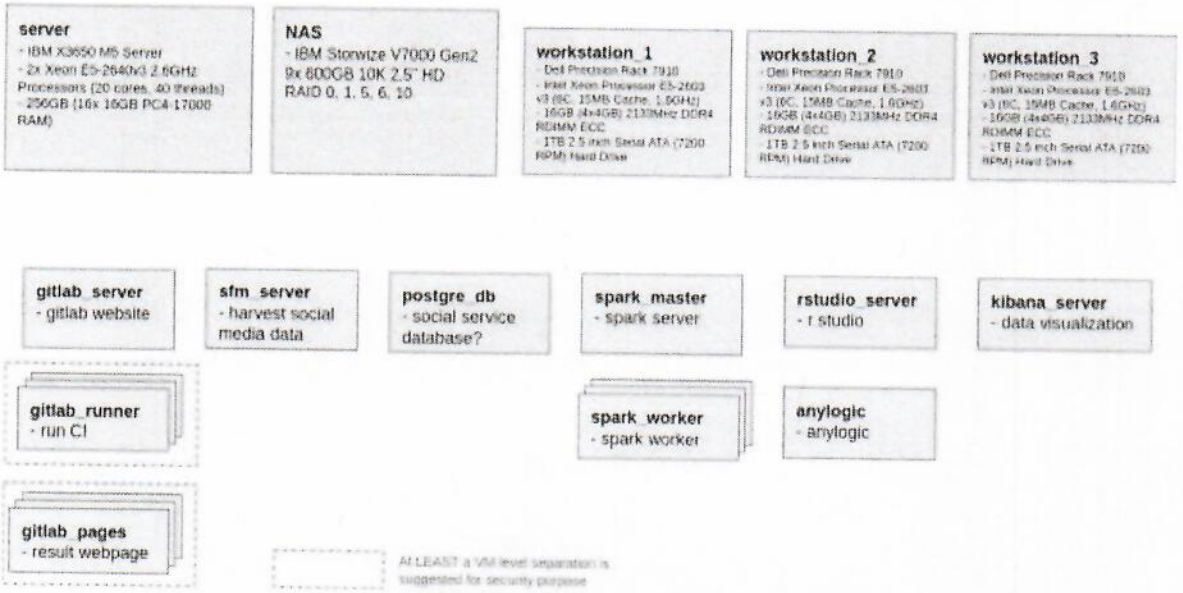


Figure 3: Resources-service map: Hardware and logical machines

Server configuration

The laboratory architecture will make use of a set of virtualized servers (depicted within the figure above), recognizing that two elements will likely be provided by a cluster of machines for performance reasons. Specifically, the laboratory architecture establishes a division into internal data servers hosting encrypted data, analysis servers serving as computational engines but holding no persistent data, and (in the DMZ) a bulkheaded and process-isolated external work container gitlab-based submission server. All such servers will be set up to maintain an extensive audit travel concerning work being conducted. This is supported by configuration to support detailed multi-level logging, both through traditional Linux system logs as well as container-based logging, database logging of queries, and code uses of the programming libraries on that machine via the application programming interface (API), such as through *ltrace* and *strace*. Particular classes of servers will be equipped with added protection. Each such type of servers is described briefly below.

Data server

For security reasons, the architecture plan is to have SQL Postgres databases running inside encrypted Docker images, rather than connecting to external database machines. Analysis will pull data from such machines and cannot update it (via a "read only mount"). We anticipate initially limiting the database availability on such machines with the traditional. Following experimentation and evaluation, the popular No SQL database Cassandra may also be used. Beyond data, the servers will further host restricted-access documentation and metadata and an associated HTTP-based browsing system whose circulation is limited to the laboratory (see below). A key part of the audit trail for the databases will be provision of system log entries that record all queries, the time at which such requests take place, the user involved and identity of the verified analysis application making the request.

Analysis infrastructure

The analysis infrastructure for SPPAL will consist of a Spark cluster of one or more machines. The purpose of such machines is to serve as scalable computational engines for (potentially large-scale) analyses and for exploratory data analysis, all while providing strict guarantees ensuring privacy and confidentiality. The storage of data on such analysis machines is not allowed, and the system will be configured to prevent storage of data and results on the machine resources; such storage can only take place within the containers themselves, with the result only being exported for analyst use if SPS approval is granted. There are several mechanisms anticipated to further enhance the security of such analysis servers. In addition to the audit trail supported through system logging, such servers will further support detailed audit trails via jobs submitted to Spark via the Spark history server mechanisms. In addition, while the analysis server will act as clients of the external data server (specifically for automatically importing security-checked packages) and of the data analysis containers, by controlling send requests (easily confirmed as secure), risks of leaking any internal data server to external data server will be essentially non-existent due to the one-way connection provided. Further, as noted below, this server will blacklist all connections from the external-facing server, thus preventing connections from coming in that direction. Most storage directories are stored purely within the Docker container context and will thus automatically be cleared at the time that the container completes execution.

External submission and secure harvesting server

The external submission and secure harvesting server is a small virtual machine that is recycled frequently. While very narrow and highly restricted, the exposure of this machine to the external world requires extraordinary security measures. These are reflected in two ways. The first is its isolation within a "DMZ" by 2 firewalls (one separating from other SPPAL machines, and one separating it from the external world). The second is the fact that all other SPPAL and SPS machines will be set to ignore all connections from the IP of this server. This allows the connections to be unidirectional, in that the other servers can send requests to this server (e.g., to obtain the validated packages), while this server is unable to contact any of the other machines.

This machine serves two functions. Firstly, the server will run a web (HTTPS) server accessible only via virtual private network (VPN) from the external world. (IP filtering options further limiting connections to certain selected partner sites are also being considered). The sole web content served by this web server consists of a gitlab instance "work container submission" system, which allows external analysts to submit a secure Docker "work container" specification that packages up an analysis script with the resources that it requires (where the required dependencies are only downloaded from available sources). The system will undertake its normal continuous integration pipeline on submitted work packages submission server, validating the integrity of the submitted work packages, sequestering failed packages for system administrator study, and sending mail to the claimed submitter that indicates the outcomes of the validation. For packages that fail validation, that information will contain indication as to some causes for the validation failure. Successful work packages passing authenticity and validation tests will be turned into Docker containers for operation (in a protected space) within the SPPAL analysis area. Because of the one-way connection from the SPPAL analysis area, such compiled Docker containers will only be retrieved (pulled) by the analysis components of the laboratory, rather than pushed to the laboratory.

The second major function served by this server is hosting of a virtualized machine providing selective access to external data specifically from validated websites (e.g., Facebook, Twitter, publically available Government of Saskatchewan data, etc.). An important need for such access comes in the form of "streaming" services that provide just-breaking data for identification of important trends (e.g., copycat suicide phenomena, discussion of

restricted substances). For the near future, we anticipate this machine specifically hosting Social Feed Manager for the purposes of harvesting such data as well as custom harvesting scripts emerging from other laboratories.

To provide the necessary level of security that is advised, this system will further be located in a technical "DMZ" that features two firewalls: One that separates it from the external world (and only supports connections via virtual private network [VPN, and possibly from only certain partner IP addresses]) and another that separates the work package submission server from the internal network of SPS. The machine will further be configured to only make use of a secure domain name system (DNS) server, to avoid possible DNS redirection attacks.

Lab-Wise Security Mechanisms

It is essential that the SPPAL support analyst engagement that is rigorously and robustly secure from both technical and human vulnerabilities, while at not so burdensome that it prevents sustainable engagement with the lab. To achieve this, in addition to physical security mechanisms, we propose the use of fine-grained virtualization in the form of containers, as supported through widely-used Docker technology. Such "containerized software" provides a means of creating user-and-project-specific "sandboxes" in which analyst-supplied analysis scripts can safely run, but also a means of packaging up such scripts with the requisite resources, such as software libraries (obviating the need to download such mechanisms to the SPPAL, with resulting loss to security). Such "sandboxes" are by default not able access system resources, and cannot do so unless explicitly and transparently configured so as to do so. The integrity of the containers can also be validated explicitly prior to running using encryption mechanisms. Even more importantly, such containers can be executed in a controlled fashion that prevents the container execution from binding to any inappropriate resources (e.g., network, folders outside of the user's home folder), and to govern the output of any data (in this case, limiting it to output to particular directories to which only SPPAL staff have read-access, for the purpose of checking).

Researchers be provided with a means of uploading (only) via a cryptographically signed connections specifications for Docker containers created with the Docker container management system to a secure server located behind the SPPAL firewall, and accessible only via password-protected HTTP on a non-standard port. The server will validate the Docker scripts and confirm that the resources proposed to be accessible via these containers fall within the acceptable guidelines, with the results of that validation being reported to the users. Only specifications for containers whose resource requests are acceptable for a given project will then be available for in-lab execution in a secure user-and-project specific sandbox by the user. Rejected resources will be deleted from the server. Only approved SPS employees will have access to the upload server.

Users themselves will not be able to issue commands (e.g., "dock") that lead to execution of the software containers; such commands will not be available for their access on the SPPAL analysis servers. Instead, the sole route to users running an approved container on the SPPAL analysis server will be for that user to run that container via a simple web-based specifically for use in the SPPAL. This execution will execute the container in a "sandboxed" fashion that denies it access to all but encrypted (using a project-specific encryption key) read-access directories within or beneath the user's project-specific folder. Moreover, output can only occur to directories on a filesystem to which the user lacks read access, and only approved SPPAL personnel enjoy such access. Such output files must then be inspected by approved SPPAL personnel for security, privacy and confidentiality reasons before their final placement in a time-limited user folder for export from the lab. All intermediate files output by the container will be cleared immediately following execution of the container.

Once transmitted to the lab, the user is not permitted to transport containers to outside of the SPPAL. Just as restricting the inputs from the container provides an extra degree of security within the lab environment, effective regulation of outputs from the container is important to ensuring that output data matches ethical guidelines for use. All data written by the container (e.g., output files giving summary statistics and other aggregated information in accordance with ethics guidelines) must reside within the container itself until the point where contents proposed for export are validated as being in compliance with Ethical Approvals and approved by SPS staff.

SPPAL Technologies

The laboratory expansion involves installation, configuration, and optimization of a variety of groups of technologies as outlined below. The container management system will be highlighted here as it is critical to proposed security mechanisms and other technologies will be briefly reviewed.

Container management system and infrastructure

The container management system being used for the SPPAL expansion is based upon Docker (version 17.04.0-ce). This technology will aid enhanced security by greatly lessening the need to connect to external resources for dependency (e.g., library, package, and module) management, allow different analyses to run in isolation, etc. In addition to basic configuration, a Docker library is being be set up to allow lab users to employ a standardized templates for forming their Docker images for particular types of analysis. A critical additional element of the work involves web server to support the submission system. The following diagram demonstrates several stages associated execution of Docker containers. The Docker container is involved in a set of test, validation and verification steps prior to execution. Once inside the firewall, a researcher physically within the analysis area of the laboratory or through external cryptographically signed connections can provide a private key that supports decrypting the project-specific data, populating the databases and running the analysis containers

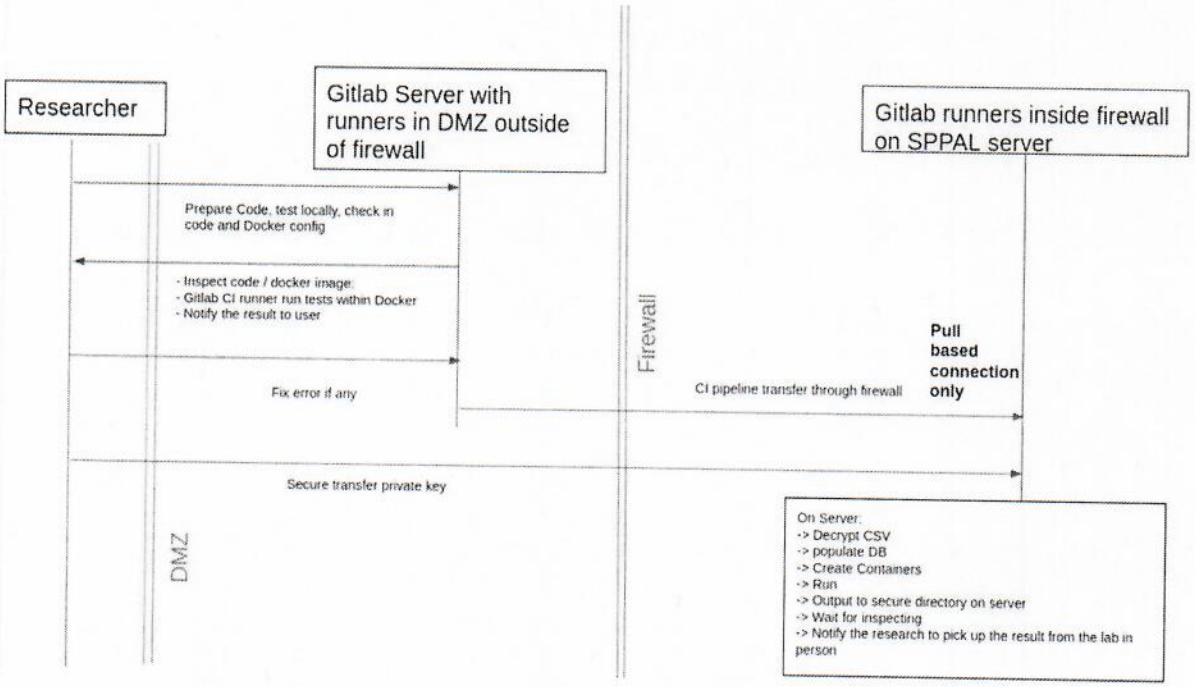


Figure 4: Day-to-day workflow using Docker

Additional technologies include:

Java based technologies: In the SPPAL expansion, Java serves both as a supported programming language and (via the Java Virtual Machine) as a platform for many additional technologies.

Python based technologies: The programming language Python is highly popular for large-scale data analysis. As such, there are a large set of Python technologies that we are installing for the SPPAL.

R based technologies: The software package R is a highly popular tool for data analysis, and offers some support for large-scale data analysis. The laboratory expansion involves installing R version 3.4.0, together with other supporting packages.

Supporting libraries and tools: To support the above, the lapack.o/blas.o libraries will be installed and as well as various supporting tools (e.g., documentation tools, tools to take visual differences between text files, software development tools).

Simulation software: In order to create synthetic ground-truth data for machine learning models, for testing such models, and for understanding the patterns seen in some of the data, AnyLogic will be installed.

Exploratory data analysis tools: The software Tableau is widely used as for exploratory data analysis with large volumes of data. This software is particularly valuable for rapid exploration of data and showing data to non-technical stakeholders. We also plan to install and configure the increasingly popular and powerful Apache Kibana software.

In addition to setting up the technologies above, it is important for SPPAL to have in place two mechanisms ("servers") for providing common services and tasks. The database server selected for this work is to be used for an important subset of data, such as publically available datasets, including those brought in from the DMZ via Social Feed Manager and from other non-sensitive data sources. We have selected Ubuntu LTS ("Long Term Service") 16.04 as the base operating system.

Finally, iterative, incremental software development is a widely used approach for creation of reliable software. While use of external services (such as GitHub's build system) is widespread among software developers, the lack of access to external resources rules out such use for SPPAL. We are using Gitlab as the build server to manage continuous integration pipelines, together with a Git repository and Wiki.

Privacy/Security

Many elements relating to privacy and security are discussed elsewhere in this summary document however, a primary focus of initial deliverables involved working to "making SPPAL safe." These processes have been well articulated by the Project Team and approved by Project Champions and several key mechanisms will be highlighted here.

- The laboratory technology shall prevent removal of any data in electronic form from its security premises without expressed approval of an appointed officer approved by the SPPAL oversight/steering committee.

- Data provided by partners shall be accessible in decrypted (plain-text) form only to validated members of groups having ethics- and partner- approval to access. Physical security checks shall take place and confirmation will be provided from SPPAL Partners and Research Ethics Board Approvals as to what data (at a field- and table- level) can be accessed. A private key/public key pair shall be generated at this point, with the

private key being then being stored in a Docker image within the continuous integration mechanisms for a particular project.

- A passphrase for a project shall then be shared between the researchers associated with the project. When a project team member leaves the project, the passphrase shall be changed to a new passphrase. When accessing the SPPAL remotely or in person, beyond the standard researcher-specific login, the project-specific passphrase shall be provided by researcher, in order to further prove their affiliation with the project. Such provision of the passphrase together with their login credentials shall support the capacity to specify analysis specification to run in a container for that particular project. The analysis script shall undergo multiple levels of verification prior to execution, and then will be run in fashion that provides only access to the associated resources approved for that project. The individual who runs this script and others in the same project shall only be able to inspect the output in person in the laboratory.

- Data shall remain in encrypted format until time of use, so as to allow system administrators to manage the data files without compromising the security of the data. Any backups made by SPPAL data shall be made on encrypted data stores and will be transmitted for archiving in an encrypted fashion. Such backups will then not be accessible to system administrators during any restoration processes.

- The SPPAL shall record detailed auditing information on laboratory use, including -- but not limited to -- logs of log-ins and log-outs, access to particular computational packages (e.g., containers), queries executed on databases, command histories of analysis, all submissions and pipeline results for continuous integration pipelines.

- Individuals operating within the SPPAL shall have write access only to project-specific electronic resources within the laboratories. While a given user may be a member of multiple project groups, a given log-in to the lab shall take place within the context of a given project, and only provide access to data for that project. This restriction will further prevent such individuals from cross-linking different datasets.

- The laboratory shall remain inaccessible at a hardware level from any access from outside of the physical premises of the laboratory. All access to external data sources (such as those listed above) shall be made from a dedicated DMZ region of SPS information technology infrastructure. The DMZ shall be separated by from the main analysis area of the laboratory via multiple levels of firewall. These must include both software and hardware (router) based mechanisms.

- Researchers shall be capable of submitting batch packages of work from outside of the laboratory, using cryptographically protected methods. All submission of batch work packages (both from outside and from within the analysis components of the laboratory) must take place via the DMZ. Submission of mechanisms to the DMZ itself will ensure that malicious code cannot be injected within the laboratory itself. All packages of work (whether submitted within or external to the laboratory) shall be verified to only access acceptable data sources both statically (from Docker container specification) and dynamically (when run). These multiple lines of checking will help further elevate confidence in the security of the mechanisms involved.

- As required by Research Ethics Board and SPPAL partner requirements, a variety of automatic de-identification mechanisms shall be applied on the unencrypted data, ranging from the most simple (separation of a master list from the de-identified data, with de-identified data using scrambled unique identifiers that can be cross-linked to the master list) to more refined (such as use of cryptographic salts).

User Interface Needs Assessment and Development

A preliminary needs assessment was informed through consultation with end users (e.g., police officers in the Missing Persons Unit, Targeted Enforcement Section, SPS). This document concludes with a list of User Interface needs which will be further refined as additional consultation occurs, including information collected from additional stakeholders (e.g., Social Services) at the upcoming Missing Persons Workshop.

In short, it has been proposed that the User Interface be a lightweight web application, where all functionalities are grouped visually and logically into thematic units, according to the types of data, analyses, or reporting with which they are associated. The main design principles of the User Interface are ease of use: to provide visual grouping of data, analyses, and reporting with minimal effort required regarding user actions; and flexibility: to permit inter-operation with other databases as required.

The following is a list of User Interface functionalities deemed basic or essential:

- Common login screen; user is able to enter username and password to access MPP data, analysis tools, and report creation.
- Homepage; user is able to navigate from here to access the main features of the software, i.e., loading / viewing / analyzing data, creating reports.
- Administrative structure; administrator is able to assign the level of data accessibility for users; user is only able to access data for which permission has been granted.
- Data access; user is able to access appropriate data bases or parts thereof according to permissions granted.
- Data analysis; user is able to perform analysis on data ranging is scope from time slicing or other data masking to machine learning or other advanced statistical analysis.
- Reporting structure; user is able to create reports based on data accessed through the User Interface, including convenient customization of common or frequent reports.

The following is a list of User Interface functionalities deemed advanced or potentially desirable:

- In conjunction with identified security requirements and processes allow login through social media such as Facebook, Twitter, etc.
- Convenient and dynamic visualization of data such as through bar charts, pie charts, and heat maps.
- Internal communication with team members/community safety partners such as via secure local e-mail or messaging.

Preliminary interfaces have now been developed and specialized code is also being written so as to incorporate future statistical models.

Data Coding - Police Data

A living data dictionary for SPS missing person data has now been created so as to define and structure police data for use in predictive analyses and interface development. Potential applicability of relevant standards and guidelines for the sharing of information (e.g., the National Information Exchange Model or NIEM) are currently being considered.

Conclusions and Future Directions

The SPPAL Missing Persons Project has generated considerable interest and excitement among police and community safety partners in SK. In addition to the work described above, members of the Missing Persons Project Team have also provided several invited talks in this reporting period (e.g., Prevention Matters Conference, Provincial Child Death Review Research Sub-committee). Researchers embarking on similar analytic work in Ontario in the area of criminal intelligence requested a site visit to SPPAL which took place in November 2017 and there are plans for continued collaboration.

The SK MOJ and SPS have also partnered with Mr. N. Taylor and the Canadian Association of Chiefs of Police - Information and Communications Technology Committee (CACP-ICT) to host an Open Analytics for Community Safety and Wellbeing Conference. The conference will take place in April 2018 in Toronto, Ontario. MP Project Team Members will be presenting at this conference. In addition to knowledge dissemination, conference objectives include working to establish a network for community safety analytics research and practice in Canada in keeping with project deliverables for in Year 2.

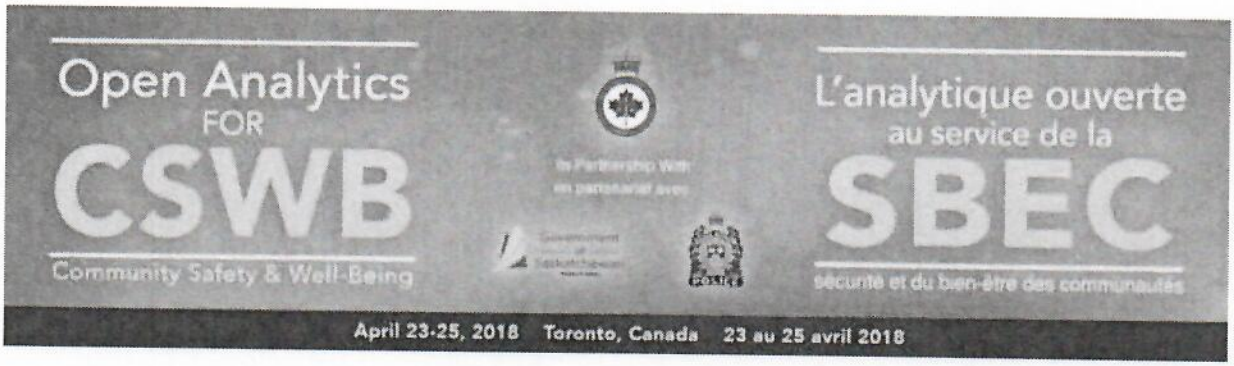


Figure 5: Open Analytics Conference