Opportunity Title: Information Security Classification of Disparate Data: Artificial Intelligence/Machine Learning
Opportunity Reference Code: ICPD-2020-30

Organization Office of the Director of National Intelligence (ODNI)

Reference Code ICPD-2020-30

How to Apply

Create and release your Profile on Zintellect – Postdoctoral applicants must create an account and complete a profile in the on-line application system. **Please note: your resume/CV may not exceed 2 pages.**

Complete your application – Enter the rest of the information required for the IC Postdoc Program Research Opportunity. The application itself contains detailed instructions for each one of these components: availability, citizenship, transcripts, dissertation abstract, publication and presentation plan, and information about your Research Advisor co-applicant.

Additional information about the IC Postdoctoral Research Fellowship Program is available on the program website located at: https://orise.orau.gov/icpostdoc/index.html.

If you have questions, send an email to ICPostdoc@orau.org. Please include the reference code for this opportunity in your email.

Application Deadline 2/28/2020 6:00:00 PM Eastern Time Zone

Description

Research Topic Description, including Problem Statement:

Data, and the insights analysts obtain from it, are crucial for Intelligence Community (IC) agencies to perform their mission. The volume and variety of data is increasing, and they are interconnected so that insights is obtained from the combination of data from many sources. Data classification is traditionally based on the content of the data, although context and metadata may also have an impact on its sensitivity. Typically, classification of the data is based on the potential impact on the national interest, organizations or individuals if the data is compromised. Classifications range from no business impact for unclassified data to catastrophic impact for top secret data. In some cases appropriate classification of data is straightforward since either the nature of the data or the way in which it was collected clearly indicate its level of sensitivity. Increasingly, organizations in the IC are drawing on a variety of data derived from unclassified or low classification sources. In this case, the level of sensitivity of the derived data is not clear, particularly when it is comprised of a range of data-types including structured, unstructured and multimedia data.

The classification level of data has substantial implications on its ability to be shared and analyzed or combined with data from other sources, which can limit its usefulness and the ability of IC agencies to collaborate with other agencies, industry or academia. At present, the risk-based guidelines do not provide clear guidelines as to the sensitivity of derived collections of disparate data. Hence the goal of this project is to use mathematical and statistical principles to establish a framework for classifying disparate collections of security relevant data based on its importance, value or sensitivity, taking into consideration the need to maximize the availability and hence usefulness of the data.

In addition, any artificial intelligence (AI) or machine learning (ML) model developed needs to be trained against a representative data set. Often, combining two disparate data sets to train a model will greatly affect its sensitivity and therefore its classification. And in some cases data sets cannot be combined between agencies due to legislative or policy restrictions. A related area of research for this proposal is the ability to train a model based on siloed or disparate data sets in a manner that preserves privacy or other security restrictions but still enables the model developed to be trained against an appropriate representative and complete data set.

Example Approaches:

- Graph networks are widely used for social network analysis. When applied to entities extracted from text-based data (for example) they can help to quantify the amount of information within a given dataset, providing guidelines for the scope of potential damage if different types or quantities of data are compromised. There are already many publicly available datasets than can be used to test these methods and develop principles for the potential impact of a data breach. An important aspect of this work will be to identify the type and extent of damage and to relate that back to statistical properties and characteristics of the data.
- ML and AI can be used to classify the content of data collection into relevant groupings and to identify outliers and anomalies. These methods show substantial promise for analyzing aggregated datasets of disparate data to find sensitive information they may contain. Applying these methods to disparate collections of data will help quantify the level of risk associated with these collections and hence inform the appropriate classification of the data.

Opportunity Title: Information Security Classification of Disparate Data: Artificial Intelligence/Machine Learning
Opportunity Reference Code: ICPD-2020-30

- Historic data breaches and unauthorized disclosures provide an opportunity to evaluate the amount of damage that can be attributed to a given volume and type of data. Methods for evaluating identification of sensitive information stemming for privacy research, as well as methods outlined above, can be allied to these datasets to quantify the probability and extent of compromise for a given dataset (which may depend on the type, volume and nature of the data), providing empirical indicators of damage.
- An alternative approach could be to consider the potential level of compromise if sensitive attributes were made available at a low classification (for example if shared between agencies or made available to industry partners) with or without context and in either an open or encrypted form as a reference for AI or ML analysis, or for context based searching.

Relevance to the Intelligence Community:

This is an escalating problem for IC agencies as there is an increasing need to collaborate across agencies, and with industry and academia. Higher classification of data restricts availability, usefulness and value. Moreover, the classification of data has an impact on its use for internet of things applications, edge technologies and AI. Having a well-defined set of objective principles for classifying disparate collections of security-relevant data would assist in balancing the risks associated with sharing data against the benefits of sharing the data.

Key Words: Information Classification, Information Security, Data Classification Standards, Machine Learning, Artificial Intelligence, Prediction, Analytics, Graph Networks

Qualifications

Postdoc Eligibility

- U.S. citizens only
- Ph.D. in a relevant field must be completed before beginning the appointment and within five years of the application deadline
- Proposal must be associated with an accredited U.S. university, college, or U.S. government laboratory
- Eligible candidates may only receive one award from the IC Postdoctoral Research Fellowship Program

Research Advisor Eligibility

- Must be an employee of an accredited U.S. university, college or U.S. government laboratory
- Are not required to be U.S. citizens

Eligibility Requirements

- **Citizenship:** U.S. Citizen Only
- **Degree:** Doctoral Degree.
- **Discipline(s):**
  - **Communications and Graphics Design** (2 👁)
  - **Computer, Information, and Data Sciences** (16 👁)
  - **Earth and Geosciences** (21 👁)
  - **Engineering** (27 👁)
  - **Environmental and Marine Sciences** (14 👁)
  - **Life Health and Medical Sciences** (45 👁)
  - **Mathematics and Statistics** (10 👁)
  - **Other Non-S&E** (2 👁)
  - **Other Physical Sciences** (12 👁)
  - **Other S&E-Related** (1 👁)
  - **Physics** (16 👁)
  - **Social and Behavioral Sciences** (27 👁)