PSCR 2020: THE DIGITAL EXPERIENCE

August 2020

NIST

#PSCR2020

PSCR

Crisis Collaborations: Challenges for Safe Data Sharing with Differential Privacy

Diane Ridgeway
Christine Task
Gary Howarth
David Van Ballegooijen

NIST

#PSCR2020

PSCR

DISCLAIMER

Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately.

Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

*** Please note, unless mentioned in reference to a NIST Publication, all information and data presented is preliminary/in-progress and subject to change**

Overview

- **Public Safety Data and Open Data Initiatives**
- **Privacy Risks and Approaches**
- **Differential Privacy and PSCR Challenge**
- **Potential Impacts of Differential Privacy Research**

Data Collected by Public Safety

Calls for Service

- Calls to “911” for emergency assistance
- May include calls non-emergency calls
- Typically maintained in law enforcement computer-aided dispatch systems

Incidents

- Collected by an agency for management
- Stored in Records Management Systems (RMS)
- Officer reports on crimes, situations, concerns, suspects, citizen public safety issues, etc.

Stops, Citations, Arrests

- Proactive and reactive stop of pedestrians or motor vehicles
- May be resolved through warnings, citations, summons, or physical arrests
- Data may be overlapping such as a stop followed by a citation or arrest

Complaints

- Potential mistreatment by authorities
- Policy, procedure, and legal violations
- May include internal affairs investigations
- Collection process required by national law and accreditation standards

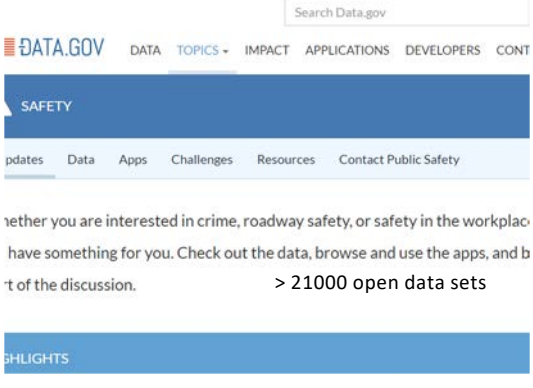
Public Safety Data De-Identification Use Cases

Analytics

Many cities are developing algorithms to analyze crime, fire, and health data. Developers would like to access other localities' data for training, analysis, and validation.

Open Access to Data

Many public safety agencies are required to report certain data. Others wish to share data with the public and researchers.



> 21000 open data sets

POLICE DATA INITIATIVE

> 150 Agencies
> 200 open datasets

FIRE DATA LAB

~ 3M incident reports

Open Data Initiatives

Transparency vs. Privacy

Risks

Personally Identifiable Information Protection

"PII is any information about an individual maintained by an agency, including (1) any information that can be used to distinguish or trace an individual's identity, such as name, social security number, date and place of birth, mother's maiden name, or biometric records; and (2) any other information that is linked or linkable to an individual, such as medical, educational, financial, and employment information."

- NIST Special Publication 800-122 Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)

Personally Identifiable Information Risk

Harm to organizations:

Low	Mod	High
LIMITED	**SERIOUS**	**SEVERE or CATASTROPHIC**
Mission Impact: Reduced effectiveness **Asset Damage:** Minor **Financial Loss:** Minor **Personnel:** Minor harm	**Mission Impact:** Reduced functionality **Asset Damage:** Significant **Financial Loss:** Significant **Personnel:** Significant non-life threatening	**Mission Impact:** Loss of a primary function **Asset Damage:** Major **Financial Loss:** Major **Personnel:** Catastrophic life threatening injury or death

PSCR ANALYTICS

Approaches to Maintaining Privacy

Redact
Eliminate sensitive data

Anonymize
Mask sensitive data

K-Anonymization
Change the individual data, but maintain statistical relevancy of the overall data set

Differential Privacy
Modify data sets so they no longer link to individual responses

Attacks on Privacy: De-anonymization

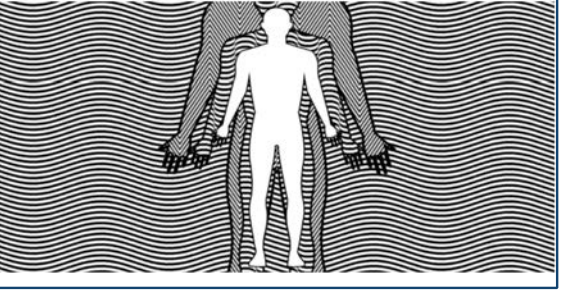
'Data is a fingerprint': why you aren't as anonymous as you think online

So-called 'anonymous' data can be easily used to identify everything from our medical records to purchase histories

12.10.18

Sorry, your data can still be identified even if it's anonymized

Urban planners and researchers at MIT found that it's shockingly easy to "reidentify" the anonymous data that people generate all day, every day in cities.



Keeping Secrets: Anonymous Data Isn't Always Anonymous

March 12, 2014 by datascience@berkeley Staff

ars TECHNICA BIZ & IT TECH SCIENCE POLICY CARS GAMING & CULTURE STO

POLICY —

"Anonymized" data really isn't—and here's why not

Companies continue to store and sometimes release vast databases of " ...

NATE ANDERSON - 9/8/2009, 7:25 AM

De-anonymization New York Taxi Data

•“Using a simulation of the medallion data, we show that our attack can re-identify over 91% of the taxis that ply in NYC even when using a perfect pseudonymization of medallion numbers.”

•Douriez, Marie, et al. "Anonymizing nyc taxi data: Does it matter?." *2016 IEEE international conference on data science and advanced analytics (DSAA)*. IEEE, 2016.

New York taxi details can be extracted from anonymised data, researchers say

FoI request reveals data on 173m individual trips in US city - but could yield more details, such as drivers' addresses and income

▲ Data about New York city taxi drivers and rides could be de-anonymised, researchers warn. Photograph: Jan Johannessen/Getty Images Photograph: Jan Johannessen/Getty Images

Alex Hern

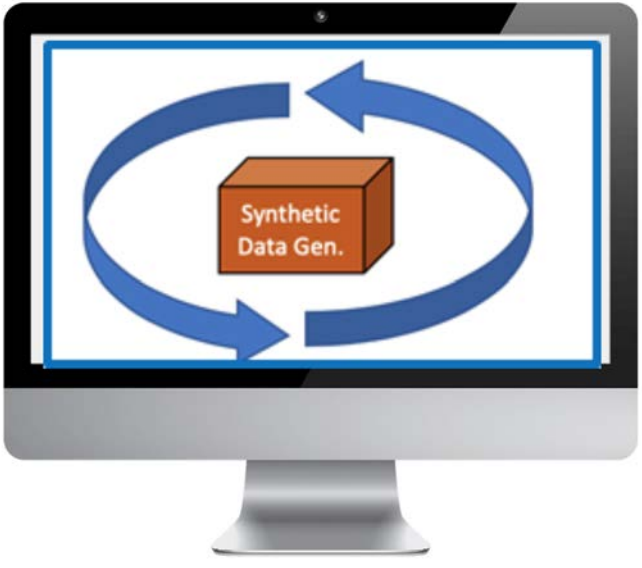
@alexhern

Fri 27 Jun 2014 10.57 EDT

Formal Privacy Differential Privacy Guarantee

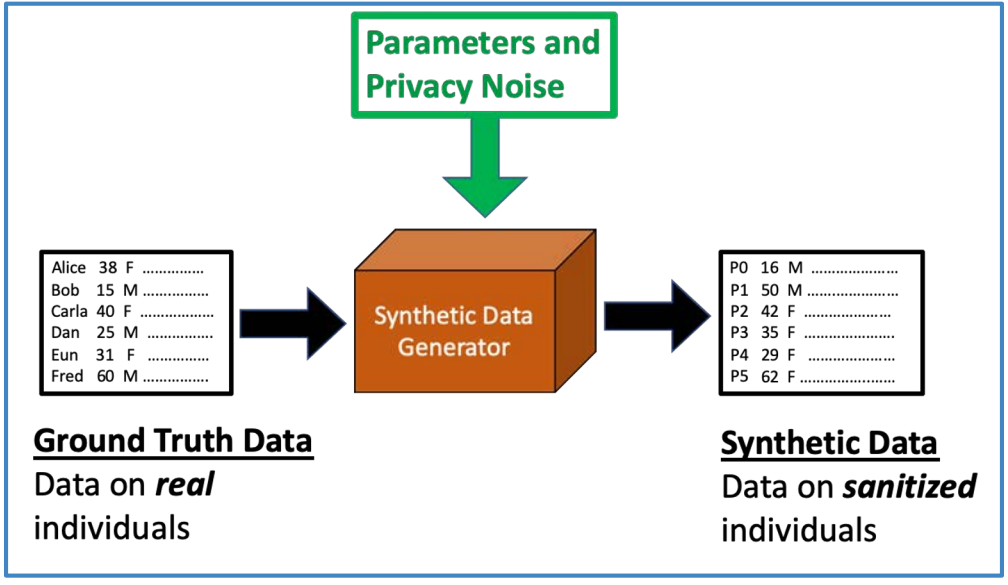
"Differential Privacy is a standard that protects privacy no matter what third-party data is available. It does so by strictly limiting what it is possible to learn about any individual in the data set."

Formal Privacy Differential Privacy Tutorial



PSCR ANALYTICS

PSCR Differential Privacy Challenges: 2018 Differential Privacy Synthetic Data Challenge



PSCR ANALYTICS

PSCR Differential Privacy Challenges

2018: Synthetic Data

- Generated synthetic Personally Identifiable Information (PII) data
- Tabular data

2020: Map Data

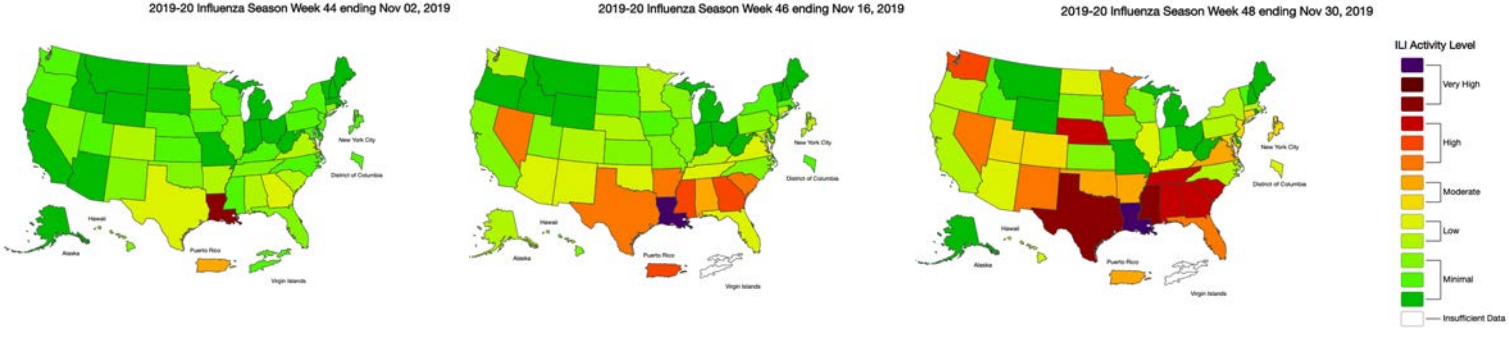
- Generated synthetic Analytics and PII data
- Map/Geographical data

2020: Temporal Data

- Time Series Synthetic Maps
- Applications: Situational awareness, planning, model training and prediction

Privatizing Temporal Map Data

This challenge will follow the success of DEID1 by analyzing differential privacy models tailored to share spatial (e.g. map) and temporal data (changes over time).



FluView, CDC

PSCR OPEN INNOVATION
Solving the Public Safety needs of tomorrow...today

Privatizing Temporal Map Data

Synthetic map data requires quality results across the ***entire map***.

02

Time sequences increase the data space, and the difficulty ***exponentially.***

Judging Maps | **Map Diversity** | **Adding Sequences**

01

Dense urban, sparse rural, and other variations require ***flexible algorithms***.

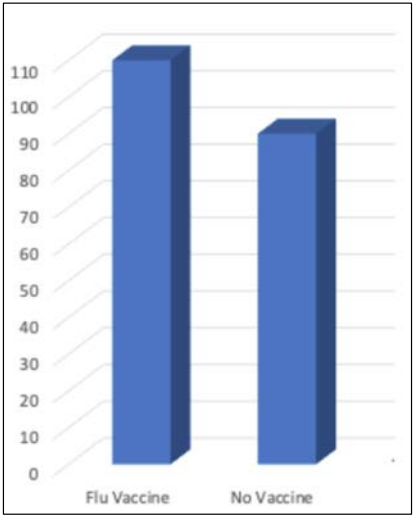
03

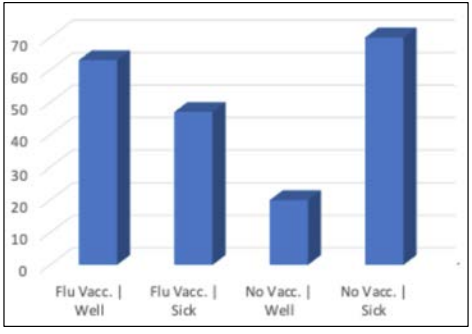
Technical Challenges

PSCR ANALYTICS

Privatizing Temporal Map Data

Problem size and complexity increase with amount of information shared and number of map locations





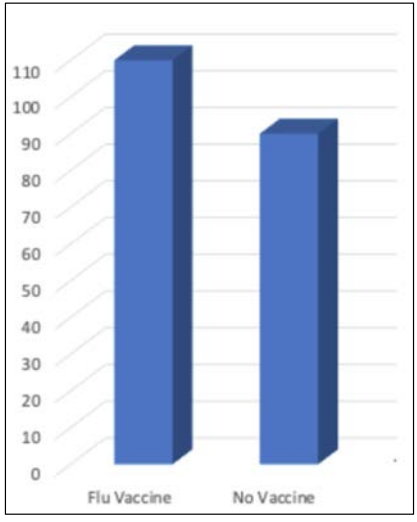
Technical Challenges

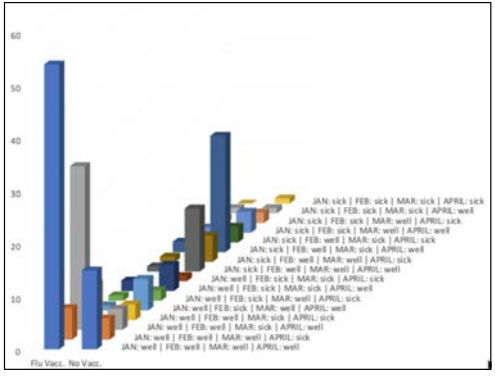
PSCR ANALYTICS

Privatizing Temporal Map Data

Problem size and complexity increase with amount of information shared and number of map locations

Problem size and complexity increase <u>exponentially</u> with number of time steps (per individual).





Technical Challenges

PSCR ANALYTICS

2020 Differential Privacy Temporal Map Challenge

- 1 October 2020 launch date
- Three contests:
  - Data de-identification algorithm challenge
  - Metric challenge for scoring algorithm accuracy
  - Open Source and Development Contest
- Multi-phase challenge
- Up to $300,000 in prize money

Visit: www.nist.gov/ctl/pscr/open-innovation-prize-challenges

DISCLAIMER

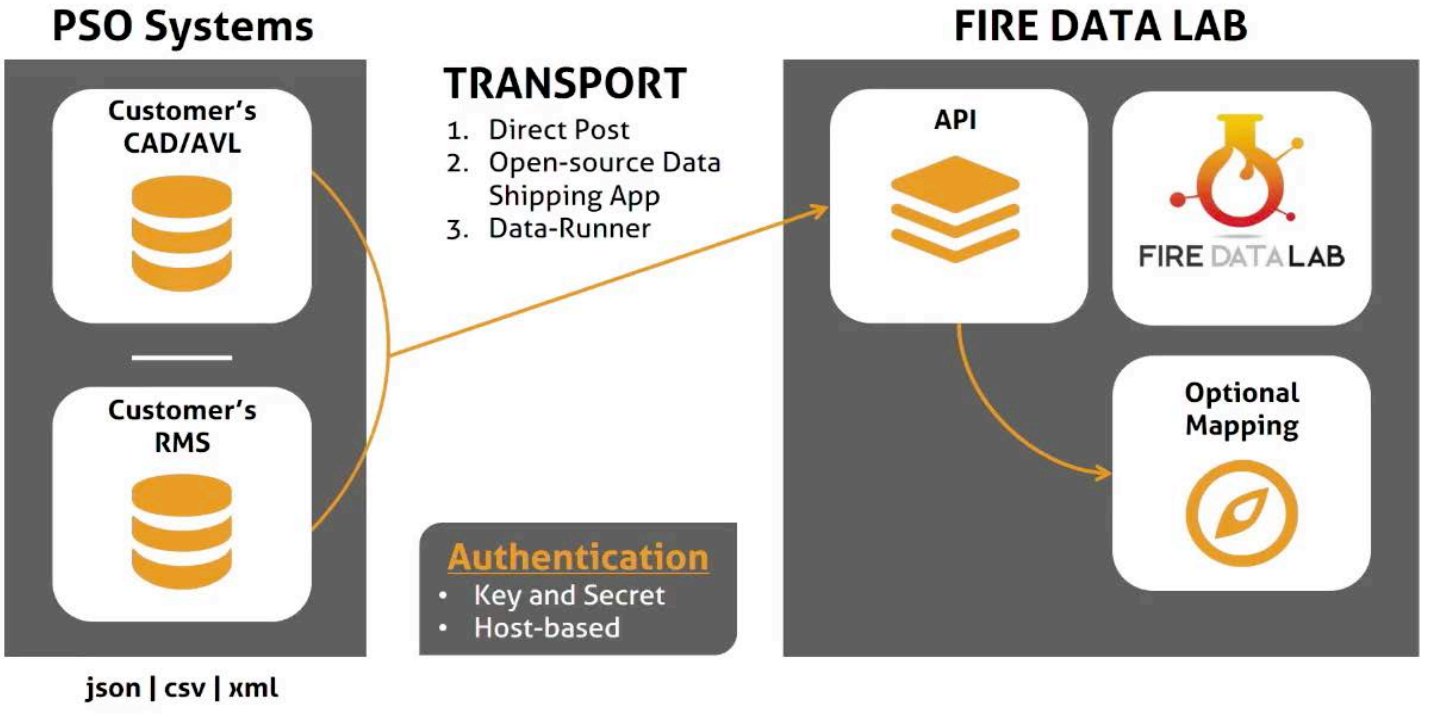
The following slides, 24 to 30 are presented by a guest speaker and presented for publication in the National Institute of Standards and Technology's PSCR 2020: The Digital Experience. The contents of this presentation do not necessarily reflect the views or policies of the National Institute of Standards and Technology or the U.S. Government

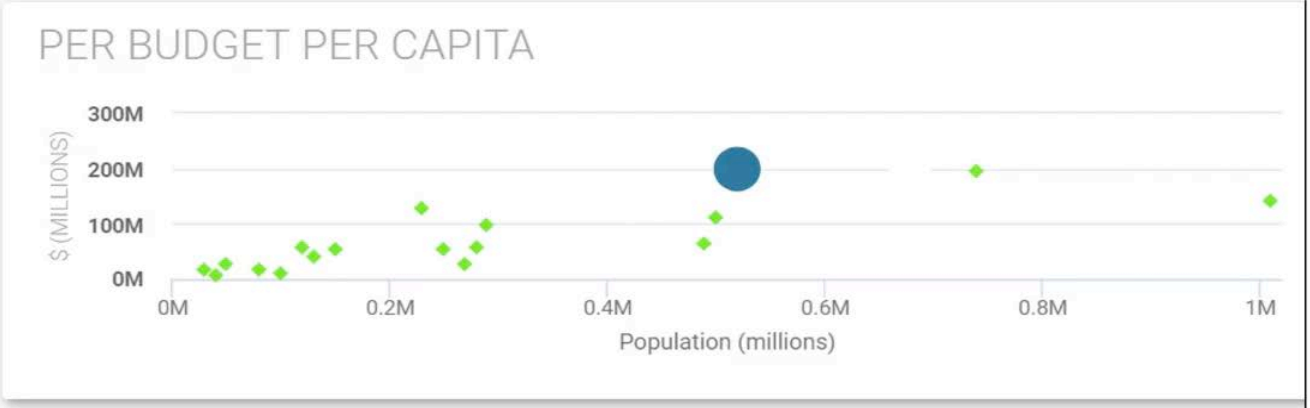
Posted with Permission.

2020 PSCR

FIRE DATA LAB

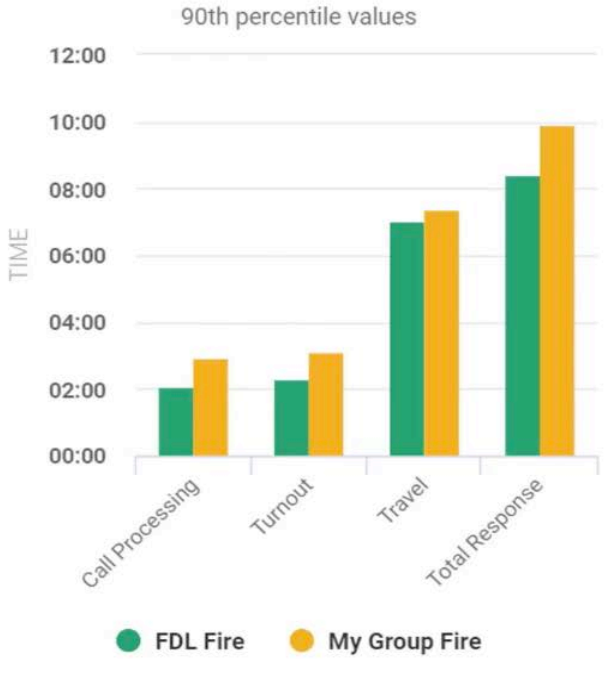
David Van Ballegooijen
Western Fire Chiefs Association


PSO Systems
Customer's CAD/AVL
Customer's RMS
TRANSPORT
1. Direct Post
2. Open-source Data Shipping App
3. Data-Runner
Authentication
• Key and Secret
• Host-based
FIRE DATA LAB
API
FIRE DATA LAB
Optional Mapping
json | csv | xml

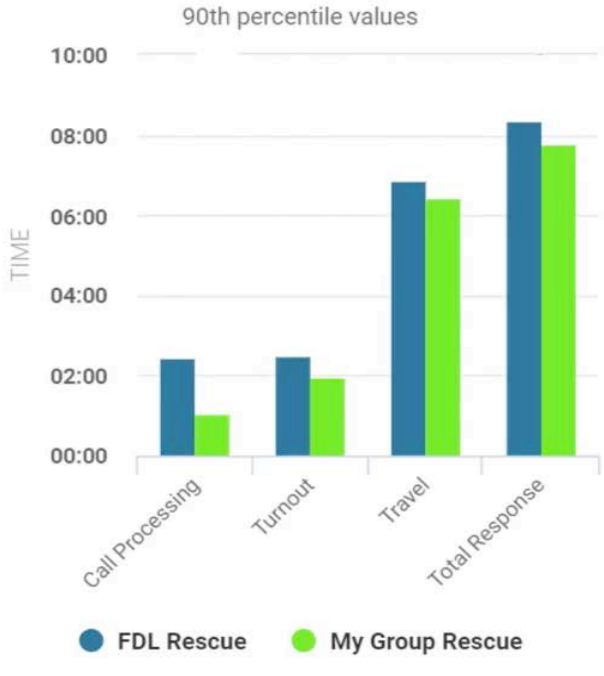

PER BUDGET PER CAPITA
$ (MILLIONS)
300M
200M
100M
0M
0M
0.2M
0.4M
0.6M
0.8M
1M
Population (millions)


EMERGENCY RESPONDERS PER CAPITA
0.5/1k
0.75/1k
1/1k
1.25/1k
1.5/1k
1.75/1k
FTE's per thousand

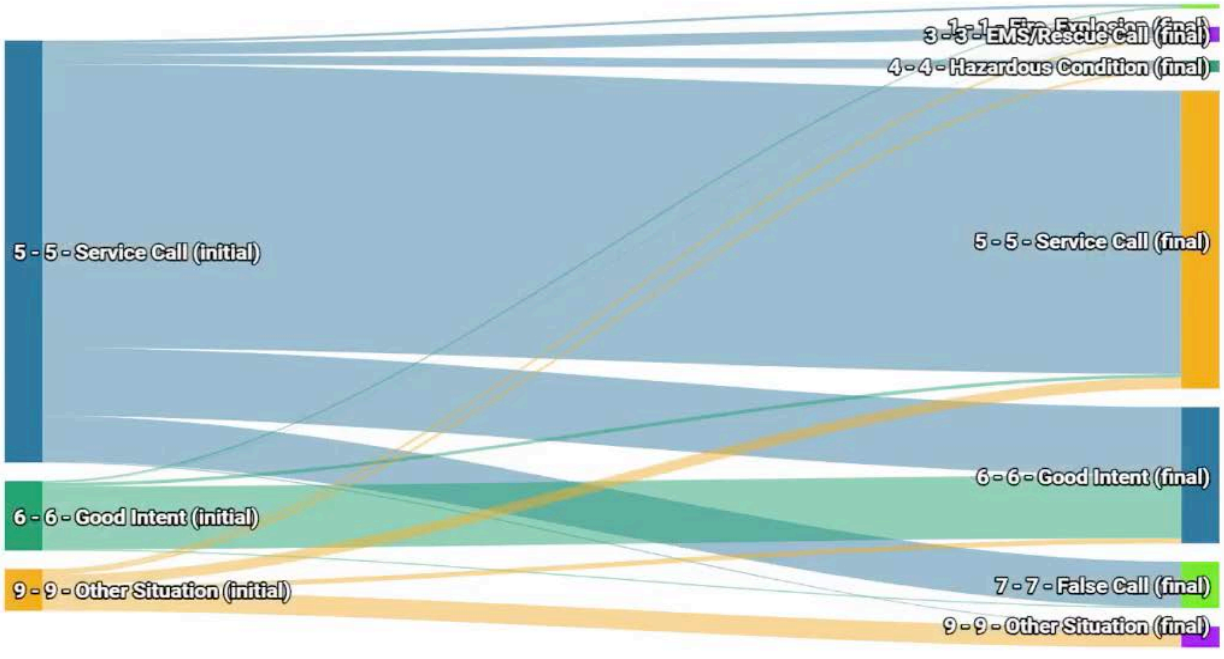
EMERGENT FIRE CALL RESPONSE



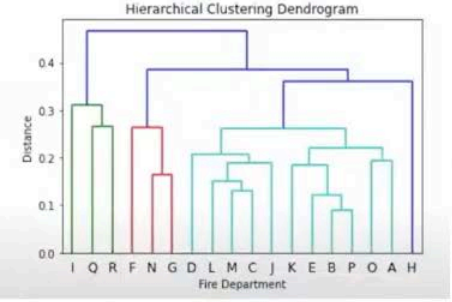
EMERGENT RESCUE CALL RESPONSE



AMBIGUOUS NFIRS CLASSIFICATIONS

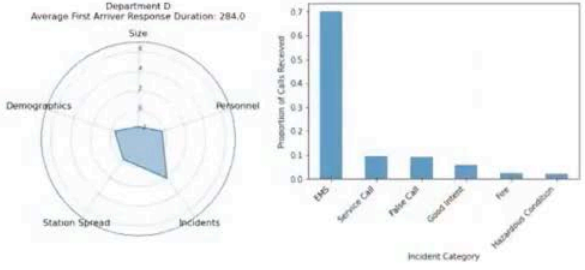


Visualization



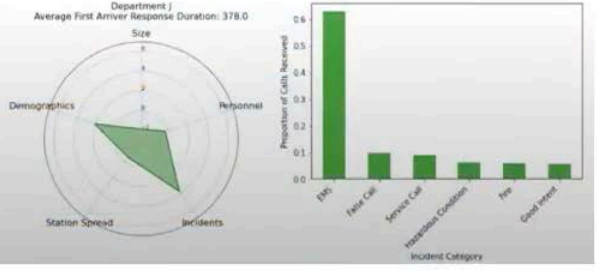
Chosen Department:

Department D



Most similar department by clustering:

Department J



Potential Impacts: Crisis

Granular Cohorts	Intervention Analysis
Defined value of service	Research Access

FIRE DATA LAB

Diane Ridgeway,
Project Manager, NIST ITL
- diane.ridgeway@nist.gov

Gary Howarth, PhD
Prize Challenge Manager, NIST PSCR
- gary.howarth@nist.gov

Christine Task, PhD
Differential Privacy Challenges Technical Lead, Knexus Research, Inc.
- christine.task@knexusresearch.com

David Van Ballegooijen
General Manager, Western Fire Chiefs Association
dave@wfca.com

John Garofolo
Analytics Portfolio Lead, NIST ITL
john.garofolo@nist.gov

THANK YOU