Data Sharing & Differential Privacy PSCR Data Analytics Portfolio

July 10th, 2019 1-1:45pm

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**

PSCR Public Safety Data Analytics

Data Sharing & Differential Privacy:
Imagine a world where data is shared safely

Terese Manley, Prize Manager
Christine Task, Project Lead

Team DPSyn, Challenge Winner

John Garofolo
Data Analytics Portfolio Lead
NIST/PSCR

Mary Theofanos
Principal Investigator
NIST/ITL

Data Sharing Differential Privacy

Public Safety has an immediate need for data analysis

- Agencies are using advanced communications technology
- Inform decision-making and increase safety
- Share data freely for better predictions of incidents
- Real-time analytics
- Ensure data privacy

Researchers have an immediate need for utility

- Datasets should not be shared without privacy protection
- Differential Privacy is a growing standard for de-identification
- Trade-offs between data privacy and utility
- Benchmarking needed to take theory to practical application

Public Safety Data News Headlines

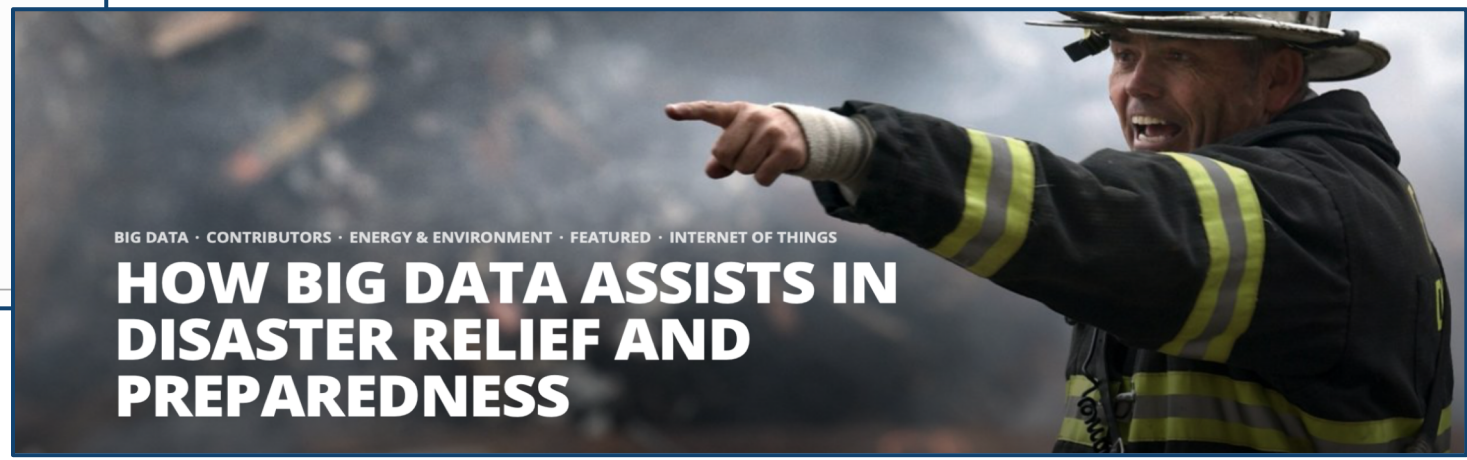
Is Big Data Analytics The Secret To Successful Fire Fighting?

Bernard Marr Contributor ⓘ

Public Safety Data News Headlines

Is Big Data Analytics The Secret To Successful Fire Fighting?

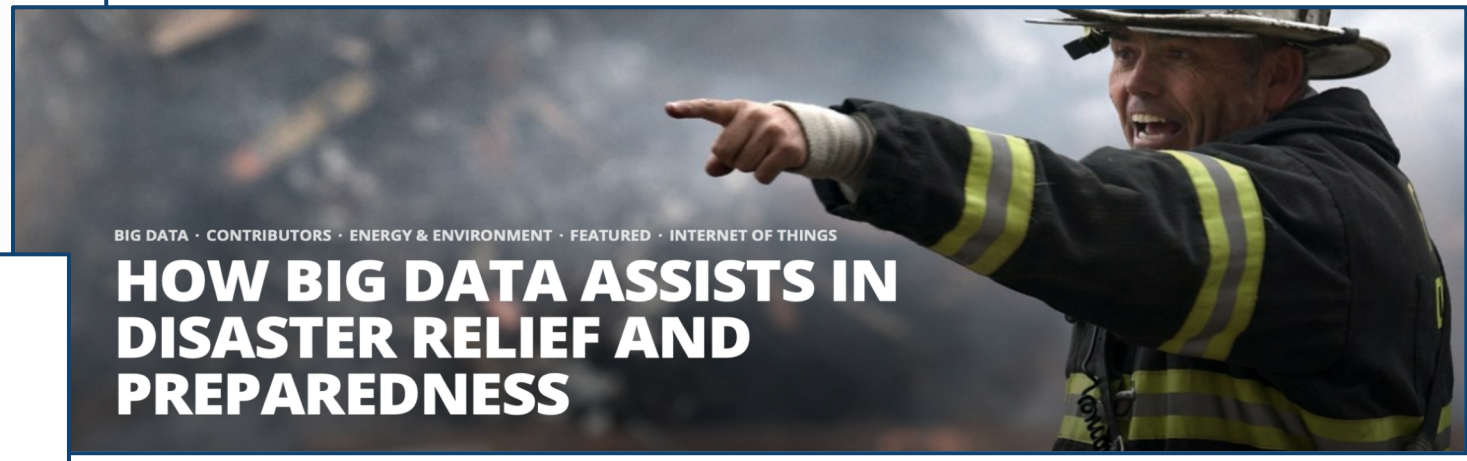
Bernard Marr Contributor ⓘ



Public Safety Data News Headlines

Is Big Data Analytics The Secret To Successful Fire Fighting?

Bernard Marr Contributor



Observations

How Big Data Can Help in Disaster Response

Technology is enabling better management of risks and crises

By Amir Elichai on December 13, 2018

Public Safety Data News Headlines

Is Big Data Analytics The Secret To Successful Fire Fighting?

Bernard Marr Contributor

Observations

How Big Data Can Help in Disaster Response

Technology is enabling better management of risks and crises

By Amir Elichai on December 13, 2018

The benefits of leveraging data and analytics in EMS

Improving patient outcomes, hospital relations and the community narrative with data analysis

Public Safety Data News Headlines

Is Big Data Analytics The Secret To Successful Fire Fighting?

Bernard Marr Contributor

Observations

How Big Data Can Help in Disaster Response

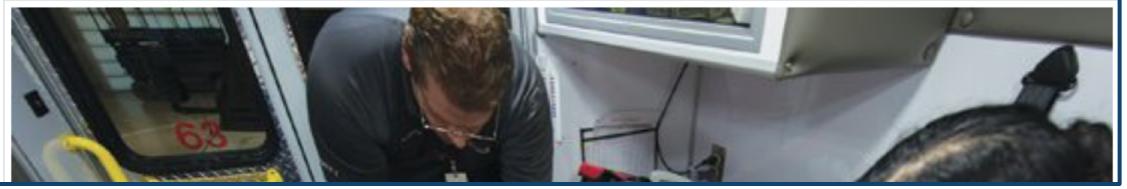
Technology is enabling better management of risks and crises

The benefits of leveraging data and analytics in EMS

oving patient outcomes, hospital relations and the community
ative with data analysis

How Everyday Data Improves EMS and Patient Care

Sun, May 1, 2016 | By Alisa Habeeb Williams, BS, NRP



Public Safety Data News Headlines

Is Big Data Analytics The Secret To Successful Fire Fighting?

Bernard Marr Contributor

Observations

How Big Data Can Help in Disaster Response

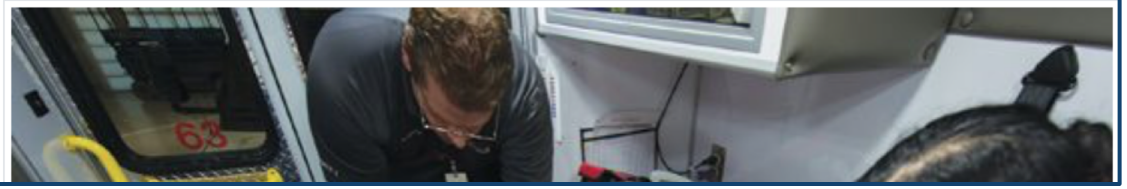
Technology is enabling better management of risks and crises

The benefits of leveraging data and analytics in EMS

oving patient outcomes, hospital relations and the community
ative with data analysis

How Everyday Data Improves EMS and Patient Care

Sun, May 1, 2016 | By Alisa Habeeb Williams, BS, NRP



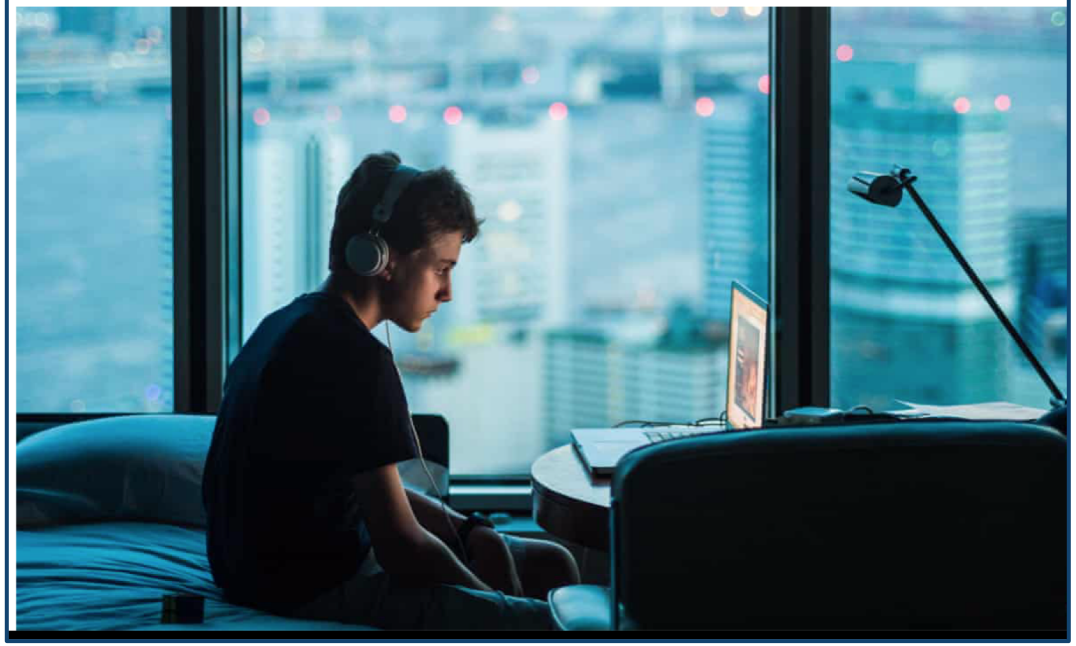
How big data analytics can be the difference for law enforcement

Finding a way to help law enforcement agencies do more with less

Data Privacy Threats News Headlines

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

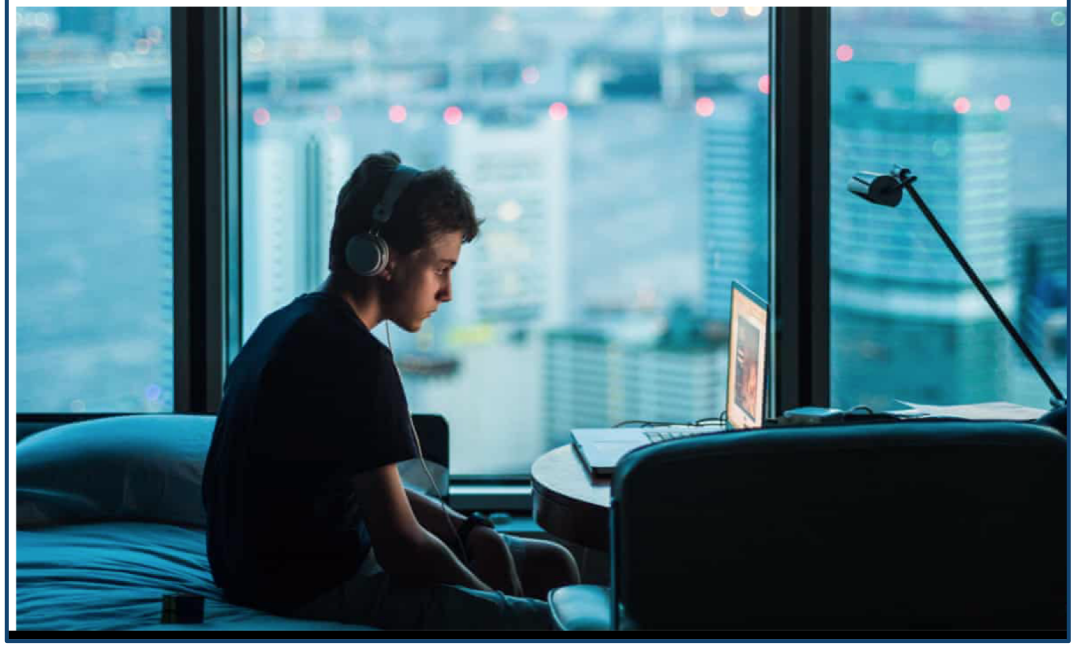
So-called 'anonymous' data can be easily used to identify everything from our medical records to purchase histories



Data Privacy Threats News Headlines

'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

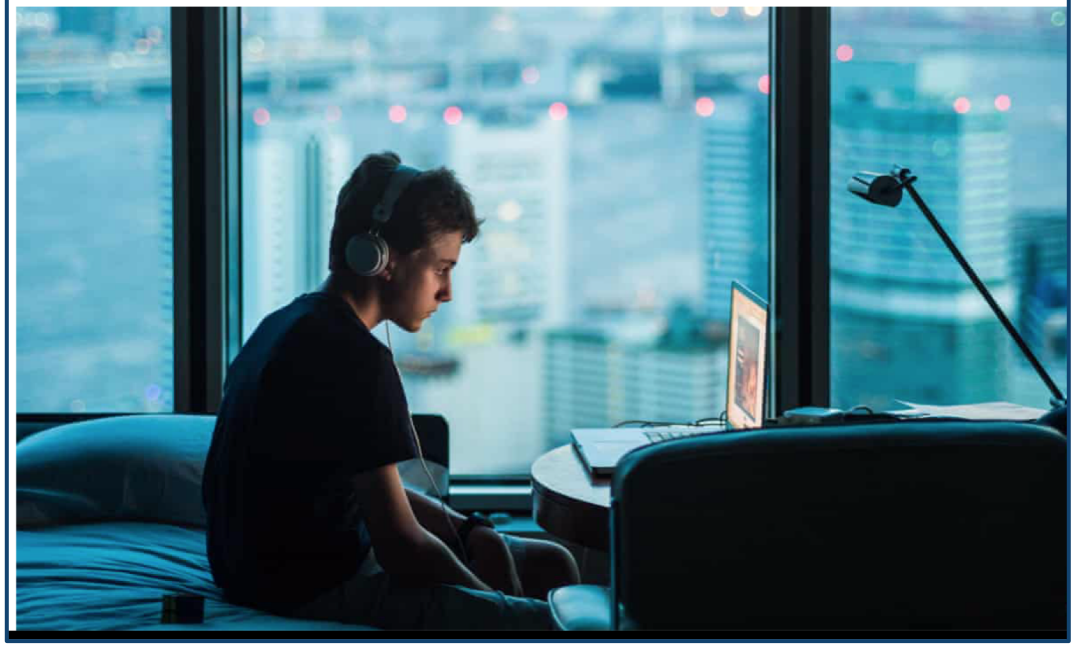


Keeping Secrets: Anonymous Data Isn't Always Anonymous

Data Privacy Threats News Headlines

'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



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 " ...

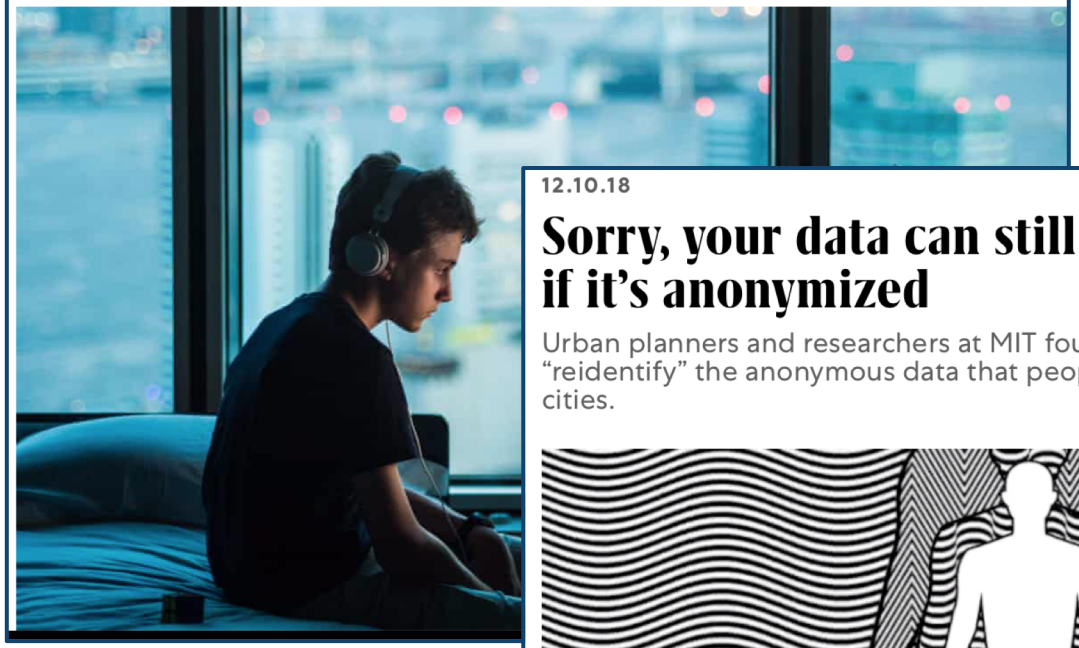
NATE ANDERSON - 9/8/2009, 7:25 AM

Keeping Secrets: Anonymous Data Isn't Always Anonymous

Data Privacy Threats News Headlines

'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



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

POLICY —

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

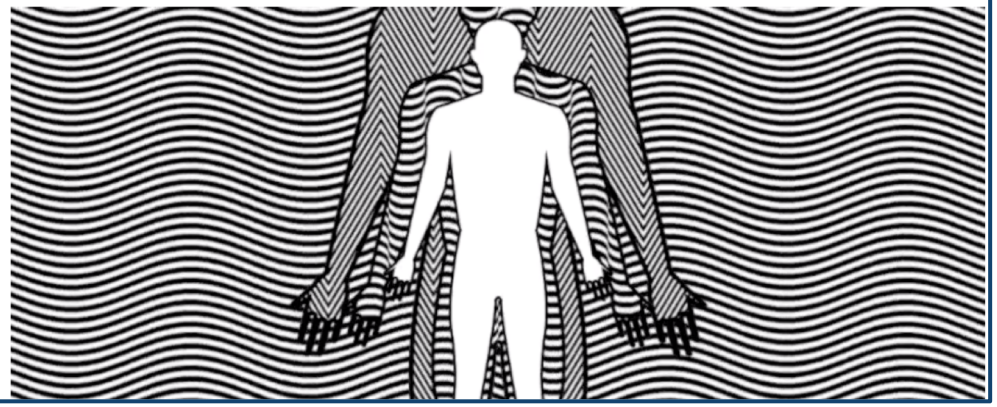
Companies continue to store and sometimes release vast databases of " ...

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

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.

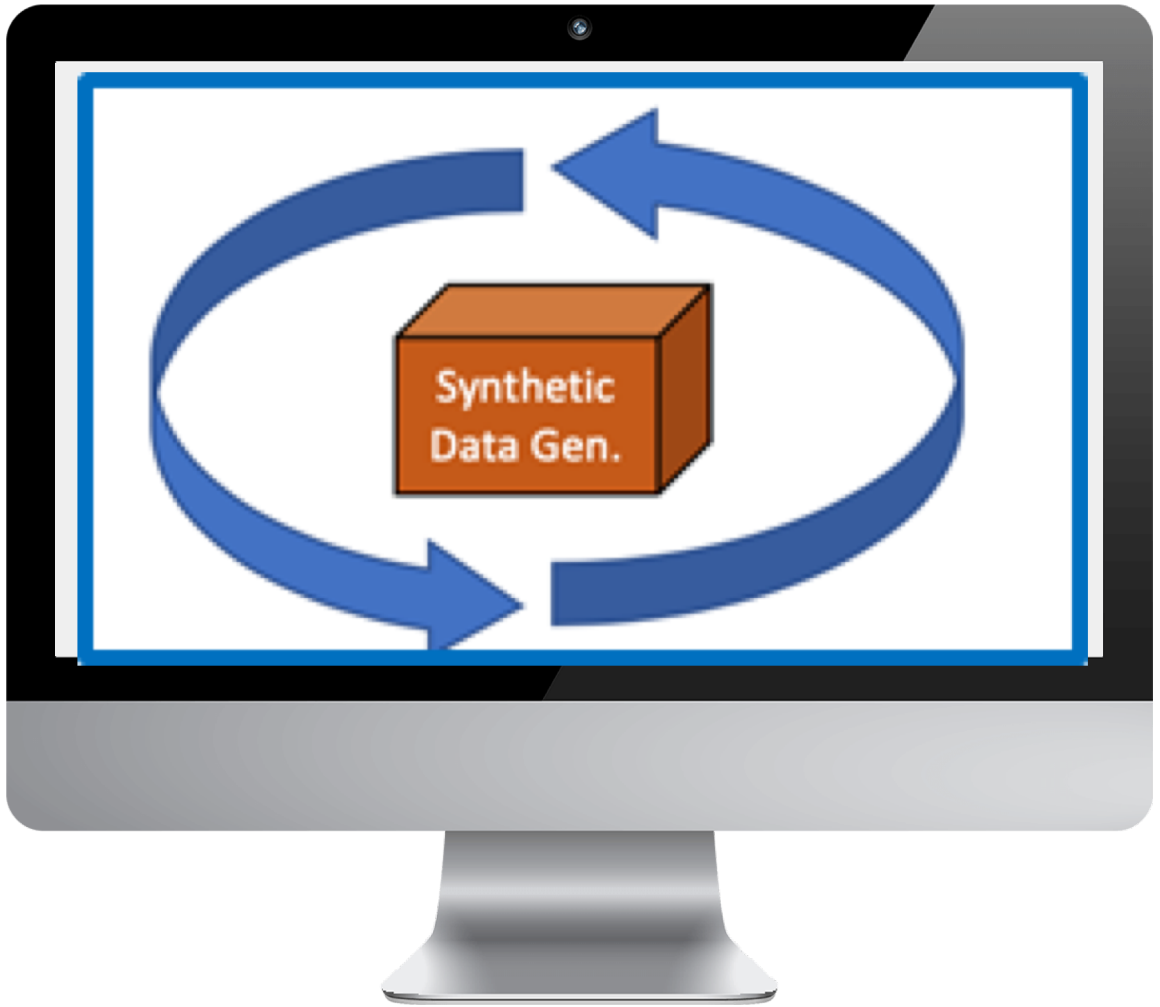


Keeping Secrets: Anonymous Data Isn't Always Anonymous

Differential Privacy The Tutorial

To Corner Alliance:
This page will have an embedded
Video but still waiting on the video.
It will arrive by 7/5/19 and I
Will put it in this same huddle
folder once it's received and
Notify Brianna. Once the video
Is embedded, slides 16 and 17
Can be removed (they were
Added for backup purposes).

Terese Manley



Differential Privacy What do we mean by Privacy?

Privacy-preserving data-mining algorithms allow trusted data-owners to release useful, aggregate information about their data-sets (such as common user behavior patterns) while at the same time protecting individual-level information.

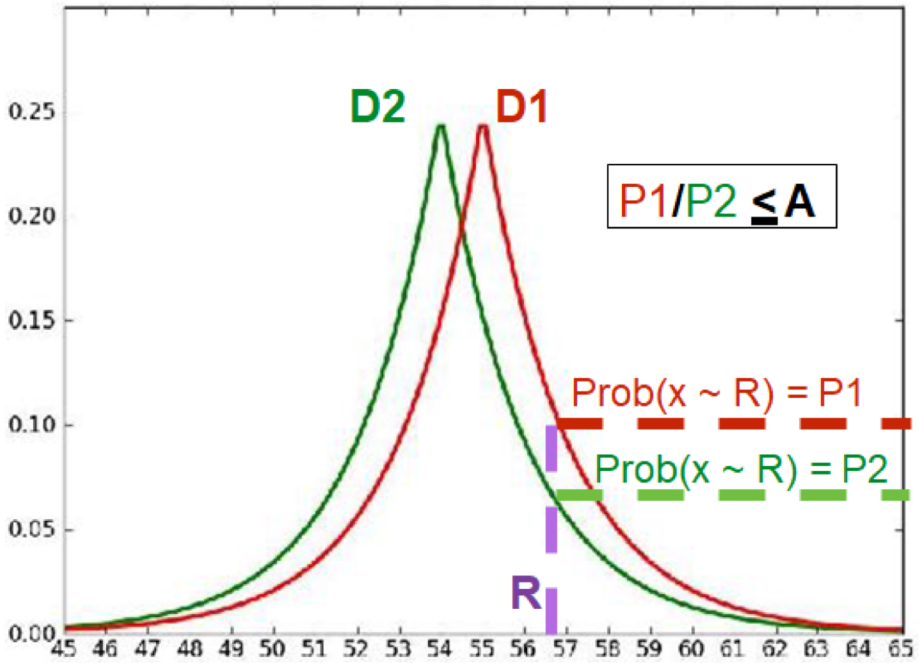
Intuitively, the concept of making large patterns visible while protecting small details makes sense. You just 'blur' things a bit:

http://fryeart1.weebly.com/journals.html

If we refine this idea into a mathematically formal definition, we can create a standard for individual privacy.

Differential Privacy This is the Laplace Mechanism

Adding **Laplacian noise** to the true answer means that the distribution of possible results from any data set overlaps heavily with the distribution of results from its neighbors.

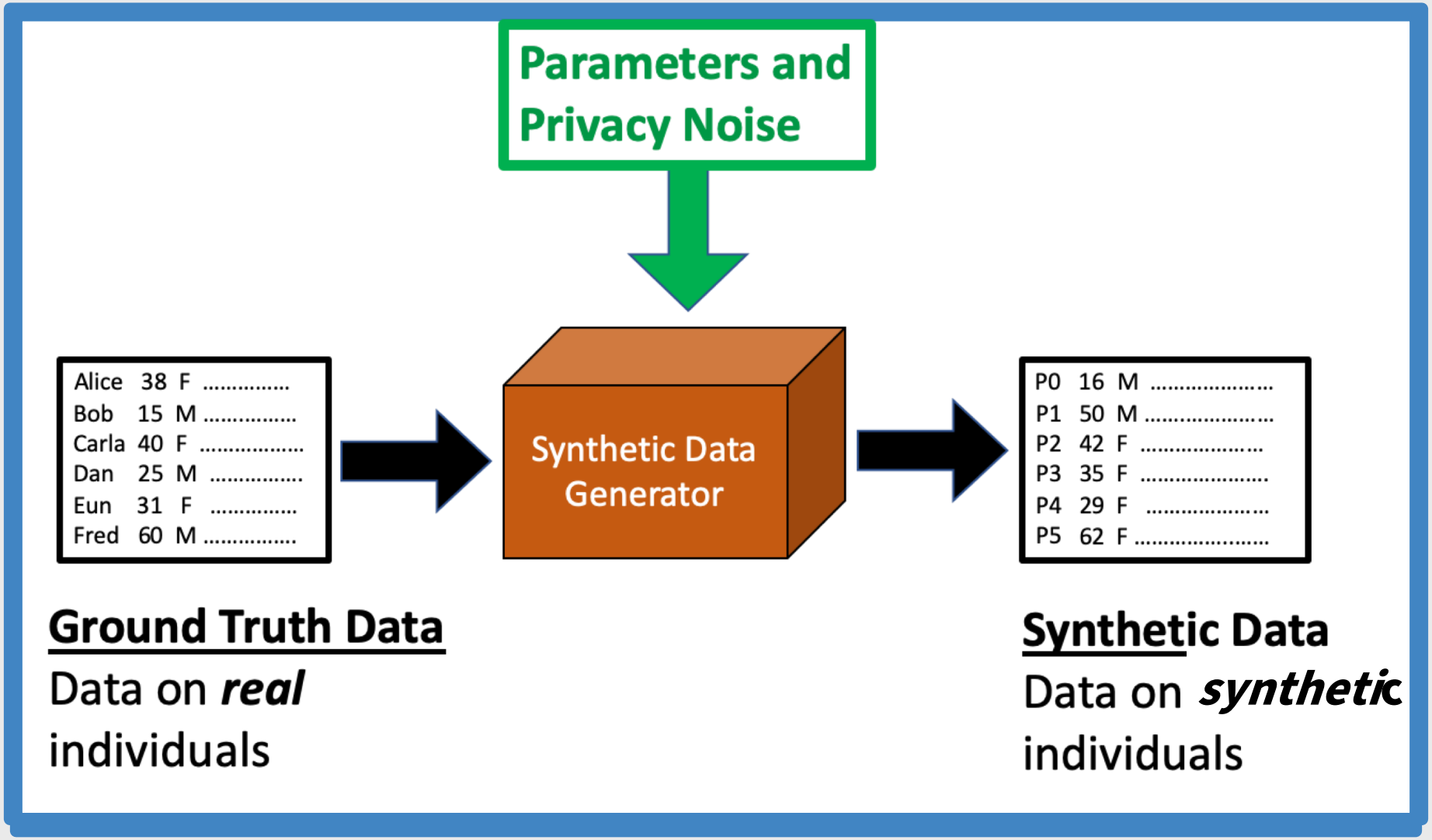


$$Prob(R = x \mid D \text{ is the true world}) = \frac{\varepsilon}{2\Delta F} e^{\frac{|x - F(D)|\varepsilon}{\Delta F}}$$

R can be any publishable data structure (not just a simple count), usually formally represented as a k-dimensional vector. It encapsulates all privatized published data.

A provides a formal measure of individual privacy. The larger A is, the farther apart P1 and P2 can be, resulting in less overlap between possible realities and a weaker privacy guarantee.

Differential Privacy Synthetic Data Generation



Challenge Need Advancing Differential Privacy

Quickly worsening privacy risks have brought Differential Privacy into a period of rapid advancement and adoption

- **Planned use in the 2020 Census**
- **Currently used in tools by Google and Apple**
- **Increase in state and local data sharing initiatives, requiring reliable privacy tools.**

Challenge Need Advancing Differential Privacy

Quickly worsening privacy risks have brought Differential Privacy into a period of rapid advancement and adoption

- **Planned use in the 2020 Census**
- **Currently used in tools by Google and Apple**
- **Increase in state and local data sharing initiatives, requiring reliable privacy tools.**

When new tech moves from theory to practice, benchmarking and competitive algorithm development are crucial. The community needed NIST's metrology expertise.

Challenge Need Advancing Differential Privacy

Quickly worsening privacy risks have brought Differential Privacy into a period of rapid advancement and adoption

- **Planned use in the 2020 Census**
- **Currently used in tools by Google and Apple**
- **Increase in state and local data sharing initiatives, requiring reliable privacy tools**

When new tech moves from theory to practice, benchmarking and competitive algorithm development are crucial. The community needed NIST's metrology expertise.

The Objective of the Differential Privacy Synthetic Data Challenge is to support rapid advancement in the development of high quality, practically usable differentially private data release tools over a year long challenge

The Challenge Bringing in the Experts

Subject Matter Experts

Christine Task, PhD

- Senior Computer Scientist
- NIST Contractor – Project Lead
- Knexus Research Corporation

Joe Near, PhD

- Assistant Professor of Computer Science
- University of Vermont

Claire McKay Bowen, PhD

- Postdoctoral Research Associate
- Statistical Sciences Group (CCS-6)
  Los Alamos National Laboratory

The Challenge Bringing in the Experts

Subject Matter Experts

Joshua Snoke, PhD

- Associate Statistician
- RAND Corporation

Changchang Liu, PhD

- Research Staff Member
- IBM Thomas J. Watson Research Center

Om Thakkar

- Graduate Student, Data Science
- Boston University

The Challenge Developing a Design

Not all algorithms are created equal: Sanitizing the data using algorithms that satisfy Differential Privacy will prevent re-identification attacks but…

The Challenge Developing a Design

Not all algorithms are created equal: Sanitizing the data using algorithms that satisfy Differential Privacy will prevent re-identification attacks but…

- **A poorly designed algorithm can require adding so much randomized "noise" to protect the data that it becomes useless for analysis**

The Challenge Developing a Design

Not all algorithms are created equal: Sanitizing the data using algorithms that satisfy Differential Privacy will prevent re-identification attacks but…

- **A poorly designed algorithm can require adding so much randomized "noise" to protect the data that it becomes useless for analysis**
- **Which algorithms are well designed?**

The Challenge Developing a Design

Not all algorithms are created equal: Sanitizing the data using algorithms that satisfy Differential Privacy will prevent re-identification attacks but…

- **A poorly designed algorithm can require adding so much randomized "noise" to protect the data that it becomes useless for analysis**
- **Which algorithms are well designed?**
- **What tricks work well to preserve utility while protecting privacy?**

The Challenge Developing a Design

Not all algorithms are created equal: Sanitizing the data using algorithms that satisfy Differential Privacy will prevent re-identification attacks but…

- **A poorly designed algorithm can require adding so much randomized "noise" to protect the data that it becomes useless for analysis**
- **Which algorithms are well designed?**
- **What tricks work well to preserve utility while protecting privacy?**

To find out, the challenge needed to be designed so that teams could repeatedly evaluate their algorithms on real world problems, use results to improve their algorithms, and then evaluate and improve again. We wanted to provide motivation and tools for them to steadily refine their solutions.

The Challenge Developing a Design

Not all algorithms are created equal: Sanitizing the data using algorithms that satisfy Differential Privacy will prevent re-identification attacks but…

- **A poorly designed algorithm can require adding so much randomized "noise" to protect the data that it becomes useless for analysis**
- **Which algorithms are well designed?**
- **What tricks work well to preserve utility while protecting privacy?**

To find out, the challenge needed to be designed so that teams could repeatedly evaluate their algorithms on real world problems, use results to improve their algorithms, and then evaluate and improve again. We wanted to provide motivation and tools for them to steadily refine their solutions.

We needed to create a boot camp

The Challenge Developing a Design

Challenge Objective: **Support rapid advancement in the development of high quality, practically usable differentially private data release tools**

Phase 1

hero^x

- Summer 2018
- Conceptual phase
- Teams proposed DP algorithms as white papers, explaining solution features
- Winners chosen by judges panel and people's choice vote

Phase 2

topcoder™

- Oct 2018 - May 2019
- Empirical phase
- Teams developed software solutions
- Sequence of three 2-month Matches
- Leaderboard shows synthetic data quality scores

The Challenge Phase 2 Design

What's a Marathon Match?

- Head-to-head algorithm competition
- Provisional Part: First 5 weeks
  - Pushes teams to improve solutions
  - Teams repeatedly submit synthetic data sets to be scored
  - Online leaderboard shows current score and team rankings
- Sequestered Part: Last 3 weeks
  - Final scoring and validation
  - Teams submit software, source code and extensive documentation.
  - Solutions are tested on secret data-sets.
  - Final scores determine prizes
- Solutions SME certified for Differential Privacy, received a 1000x point boost

Standings

Handle	Score	Rank
rmckenna	934788.86	1
ninghui	930228.00	2
privbayes	839445.10	3
gardn999	805170.88	4
manisrivastava	652890.07	5
rachelcummings	407420.08	6

The Challenge Phase 2 Design

Match Design: **Each Match introduced a new scoring metric (on top of previous metrics) to increase difficulty. Data included both emergency incident data and population data.**

Match 1:

Data: San Francisco Fire Data
Scoring: 3-Marginals

Match 2:

Data: San Francisco Fire Data
Scoring: 3-Marginals *and* Row Pool

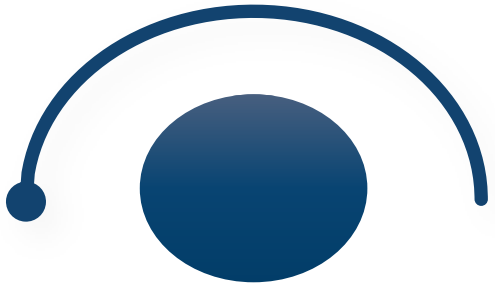
Match 3:

Data: 1940's Census Data
Scoring: 3-Marginals *and* Row Pool *and* Analytics Check (Gini Index, Pay Gap)

The Challenge Prize Design

Total Prize Purse $185,000

Phase 1



Concept Paper

1st place $15k

2nd place $10k

3rd place $5k

People's Choice $5k

$35k

The Challenge Prize Design

Total Prize Purse $185,000

Phase 1 → **Phase 2**

Concept Paper

1st place $15k

2nd place $10k

3rd place $5k

People's Choice $5k

$35k

Match #1

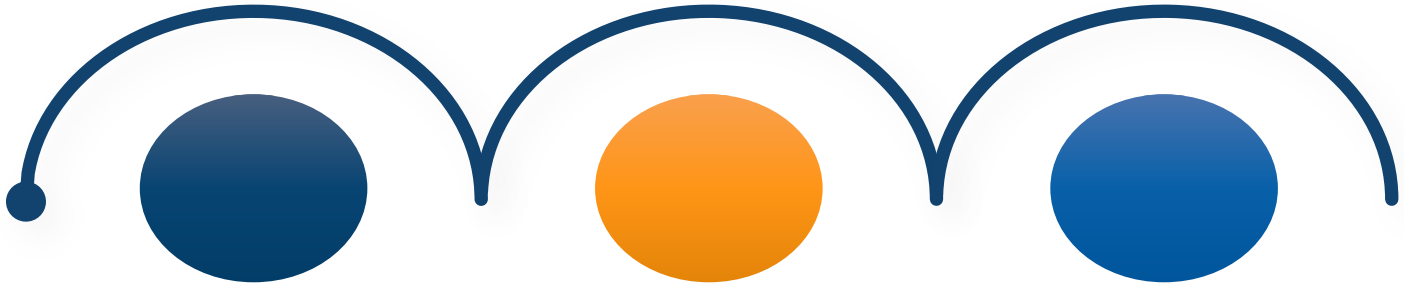
1st place $10k
2nd place $7k
3rd place $5k
4th place $2k
5th place $1k
Progressive 4 x $1k

$29k

The Challenge Prize Design

Total Prize Purse $185,000

Phase 1 → **Phase 2**



Concept Paper	Match #1	Match #2
1st place $15k	1st place $10k	1st place $15k
2nd place $10k	2nd place $7k	2nd place $10k
3rd place $5k	3rd place $5k	3rd place $5k
People's Choice $5k	4th place $2k	4th place $3k
	5th place $1k	5th place $2k
	Progressive 4 x $1k	Progressive 4 x $1k
$35k	**$29k**	**$39k**

The Challenge Prize Design

Total Prize Purse $185,000

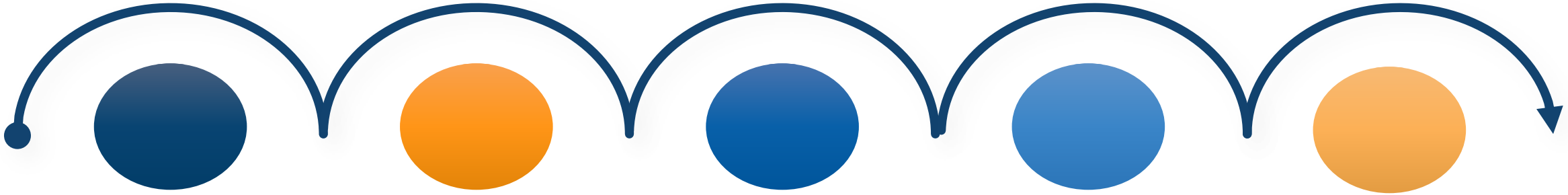
Phase 1 → **Phase 2**

Concept Paper	Match #1	Match #2	Match #3
1st place $15k	1st place $10k	1st place $15k	1st place $25k
2nd place $10k	2nd place $7k	2nd place $10k	2nd place $15k
3rd place $5k	3rd place $5k	3rd place $5k	3rd place $10k
People's Choice $5k	4th place $2k	4th place $3k	4th place $5k
	5th place $1k	5th place $2k	5th place $3k
	Progressive 4 x $1k	Progressive 4 x $1k	Progressive 4 x $1k
$35k	**$29k**	**$39k**	**$62k**

The Challenge Prize Design

Total Prize Purse $185,000

Phase 1 → **Phase 2**



Concept Paper	Match #1	Match #2	Match #3	Open Source
1st place $15k	1st place $10k	1st place $15k	1st place $25k	Additional $4k/team
2nd place $10k	2nd place $7k	2nd place $10k	2nd place $15k	
3rd place $5k	3rd place $5k	3rd place $5k	3rd place $10k	
People's Choice $5k	4th place $2k	4th place $3k	4th place $5k	
	5th place $1k	5th place $2k	5th place $3k	
	Progressive 4 x $1k	Progressive 4 x $1k	Progressive 4 x $1k	
$35k	**$29k**	**$39k**	**$62k**	**$20k**

The Challenge Top 5 Winners

Final winners of Match 3

1ST $25K

TEAM Ryan McKenna

UMass Amherst

Marginals

2ND $15K

TEAM DPSyn

PURDUE UNIVERSITY

Marginals

3RD $10K

TEAM PrivBayes

Graphical Model

4TH $5K

TEAM John Gardner

Histogram

5TH $3K

TEAM UCLANESL

Ucla IBM Research

W-GAN

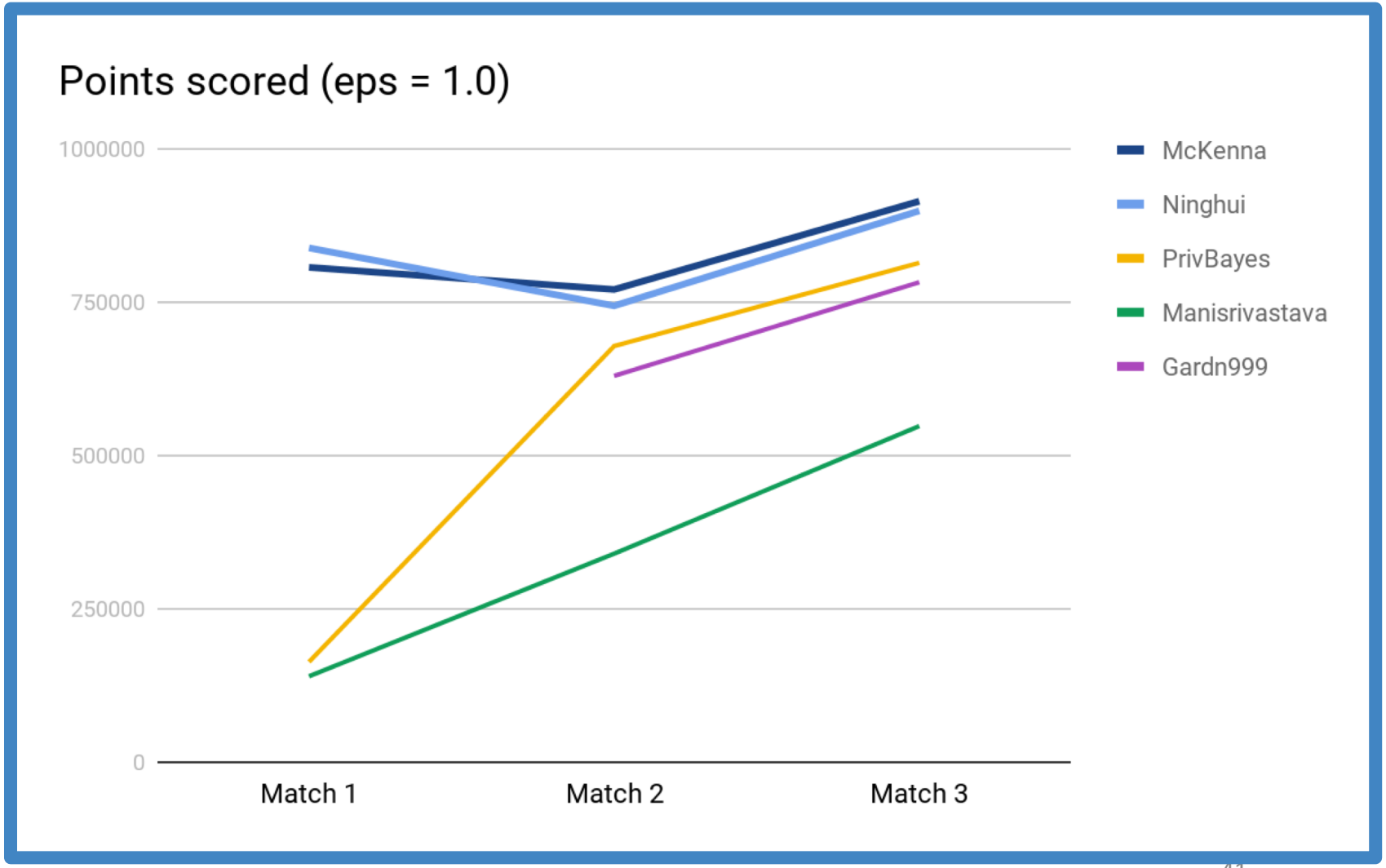
Contestants included an international innovator community from academia, research institutes and industry

Challenge The Results

- **Even though the difficulty increased each match, the teams continued to maintain and *improve* their performance**

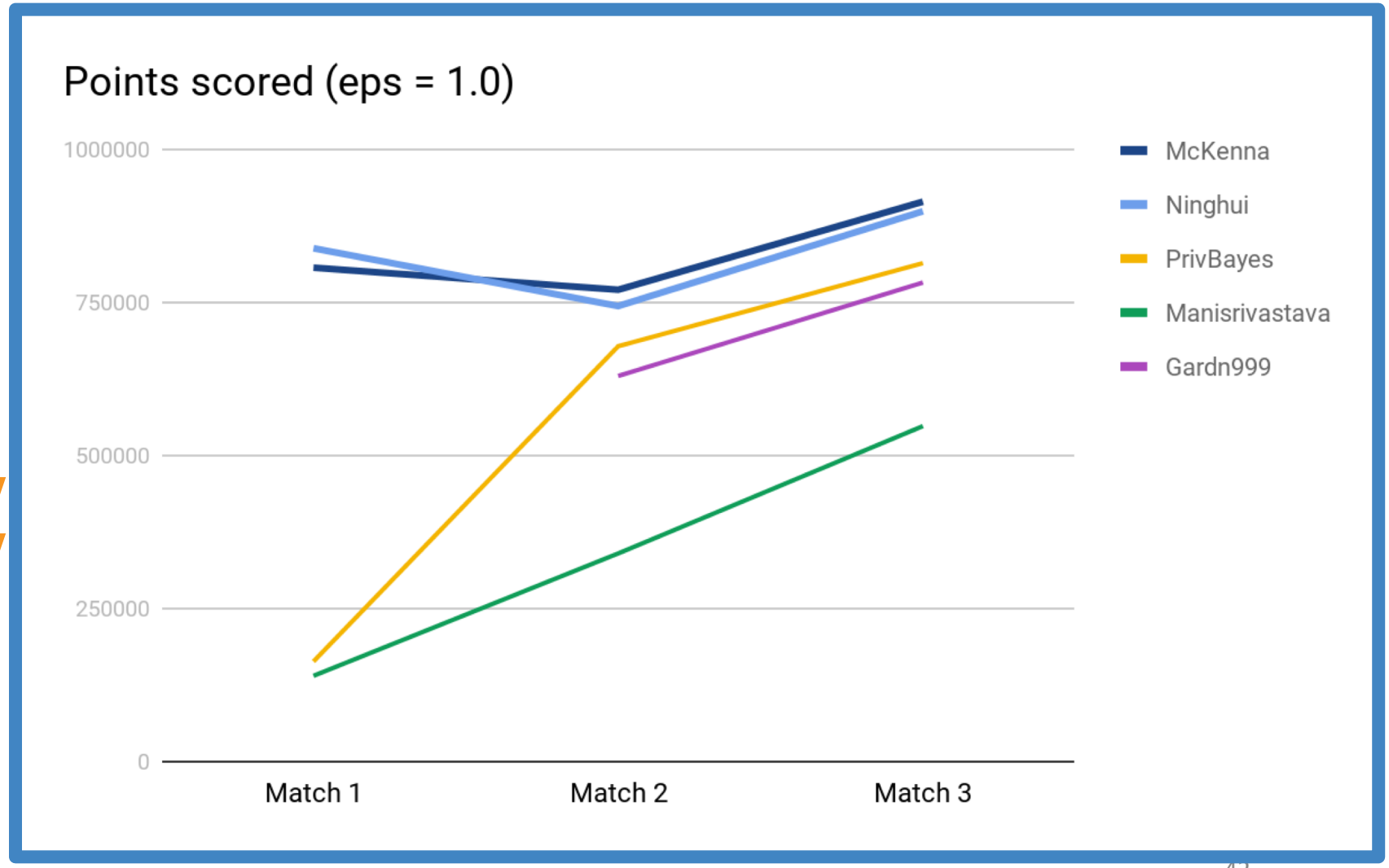
Challenge The Results

- **Even though the difficulty increased each match, the teams continued to maintain and *improve* their performance**



Challenge The Results

- **Even though the difficulty increased each match, the teams continued to maintain and *improve* their performance**
- **Each team made many new discoveries about what tricks improved their chosen approaches**
- **These discoveries are *vital to the progress of the field***



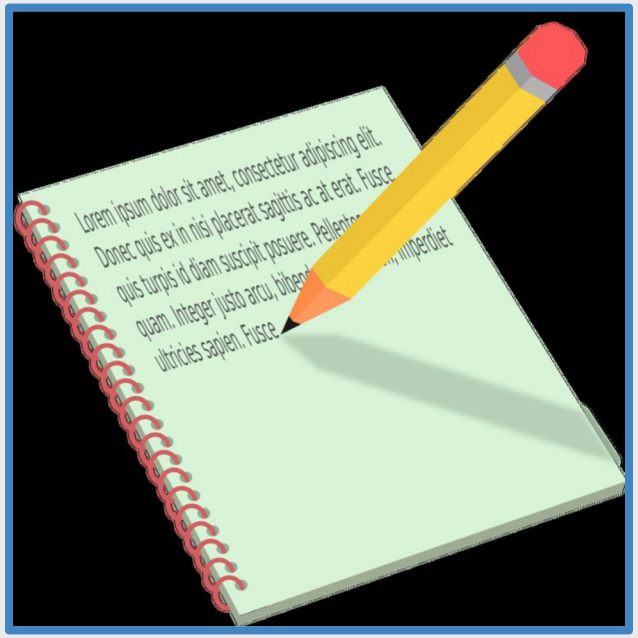
Challenge The Results

- **Even though the difficulty increased each match, the teams continued to maintain and *improve* their performance**
- **Each team made many new discoveries about what tricks improved their chosen approaches**
- **These discoveries are *vital to the progress of the field***
- **Prototyped solutions, open sourced and well documented, will continue to be improved and feed into downstream research**

The Challenge Significant Results

Lessons Learned in this Challenge

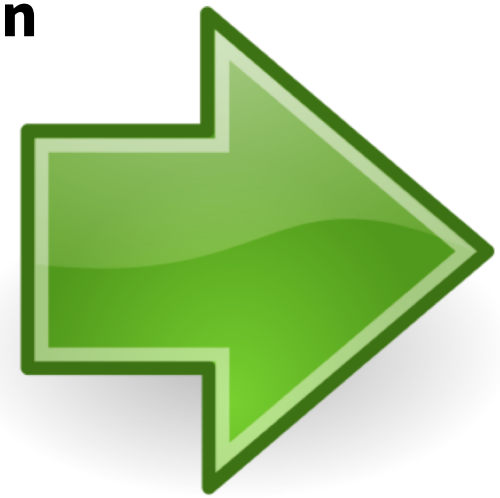
- ➢ **The significance of taking theory and moving it to practical, applied algorithms**
- ➢ **New Benchmarking techniques were developed, and competitive benchmarking had a significant effect on the research community**
- ➢ **Expansion of the DP community and recruitment of new data scientists**
- ➢ **Future challenges might use automated DP validation to assist the manual process of DP certification**



The Challenge Significant Results

What's next?

- Developing two NIST-IR research publications
  - Summary Report: Challenge design, lessons, solutions & results
  - Metrology Study: Further work on synthetic data evaluation
- Exploring possibility of an academic journal special issue
  - "Journal of Privacy and Confidentiality "
- Contestant source code, voluntarily open-sourced for research
  - NIST Privacy Collaboration Space repository
  - Academics are testing automated DP validation systems
- Evaluating future NIST Differential Privacy workshop and Prize Challenge



2nd Place Winner Team DPSyn

DPSyn is an algorithm for synthesizing microdata while satisfying differential privacy

Rapid Progress on their algorithm due to the prize challenge

Source Code is posted publicly for use by the research community

46

Call To Action

Do you have public safety datasets available for use in research?
(example datasets: 911 call data, EMS call data)

Do you want to be involved with PSCR Data Analytics as a technical or public safety expert?

Stop by our demo table today or contact us!

Mary Theofanos
Principal Investigator
mary.theofanos@nist.gov

Terese Manley
Prize Manager
terese.manley@nist.gov

Christine Task
Project Lead
christine.task@knexusresearch.com

THANK YOU