

Center for Data Science at RTI International

**UNC-CH Intro to Data Science Class
November 25, 2019**



***Jason
Nance***



**Data
Scientist**

***Gayle
Bieler***



**Founding
Director**



delivering **the promise of science**
for global good



RTI International is an independent, nonprofit research institute dedicated to improving the human condition.

We combine scientific rigor and technical expertise in social and laboratory sciences, engineering, and international development to deliver solutions to the critical needs of clients worldwide.

Worldwide Presence and Financial Strength

\$972 M 
FY2017 Revenue

3,852 
Projects
(fiscal year 2017)

1,198 
Clients
(fiscal year 2017)

12 
U.S. Offices

Research Triangle Park, NC

Ann Arbor, MI
Atlanta, GA
Berkeley, CA
Chicago, IL
Fort Collins, CO
Portland, OR
Rockville, MD
San Francisco, CA
Seattle, WA
Waltham, MA
Washington, DC

12 
International
Offices

Abu Dhabi, United Arab Emirates
Barcelona, Spain
Beijing, China
Belfast, Northern Ireland
Jakarta, Indonesia
Kuala Lumpur, Malaysia
Ljungskile, Sweden
Manchester, United Kingdom
Nairobi, Kenya
New Delhi, India
San Salvador, El Salvador
Toronto, Canada

Practice Areas

Multidisciplinary expertise and research that informs policy, practice, and decision-making

Public Health

Social and Justice policy

Education and workforce
development

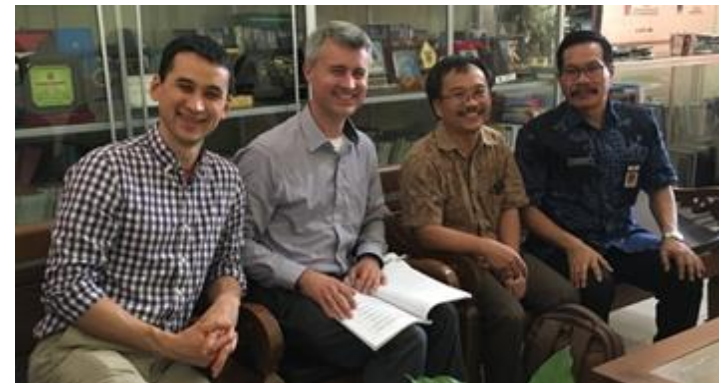
International development

Energy research

Environmental sciences

Food security and agriculture

Innovation ecosystems



Statistics and Data Science

Ensuring the quality, validity, and reliability of data from research studies of all types.

Turning data into actionable insights.



Gertrude Cox, 1960s



Center for Research in Statistics, 1980s

Data Science
Survey statistics
Biostatistics
Environmental statistics

A Transformation

Persistent feeling that we could use data not exclusively for acquiring knowledge but for informing ***action*** and ***decision-making***

Solve thorny problems

Work directly with everyday people

Evolve quickly through iteration

Understand and respond to human need

Make government work better for all of us

We can do this!



*DJ Patil, former US
Chief Data Scientist*

CDS@5: Data Science is a Team Sport



We are a vibrant team with a compelling social mission...

Data for Good



Continuous learning and innovation

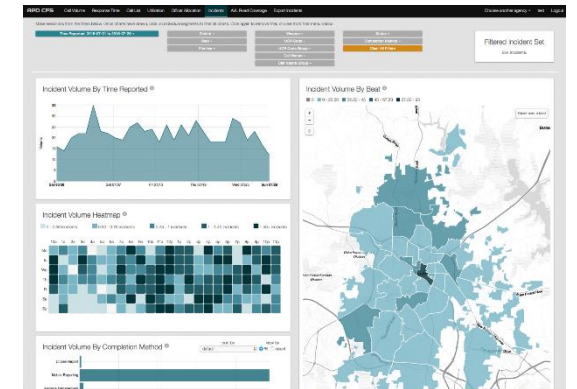
*Data science needs to learn
from itself*

Creating products that people use

*Human-centered design,
Agile software development*

Practicing data science as a team sport

*Team culture,
Cross-disciplinary
collaboration*



Integration is Key

Data scientists...

Analytical work



Software developers

Data engineers...

*Form the backbone of
a data science team*

Visual designers...

*Visual communication
of complex concepts*

Backgrounds in other fields a huge plus

We Solve Problems Across RTI: From the Social Sciences to the Lab Sciences

Research Domains

- Justice
- Clinical Studies
- Public Health
- Environment
- Education
- Surveys
- Lab sciences (sensors)
- **Data Science Tools**
 - Open source projects
 - Publicly available data products

For Whom?

The logo for the National Institute of Justice (NIJ), featuring the letters "NIJ" in a bold, black, sans-serif font.The logo for the Bureau of the Census (BJS), featuring the letters "BJS" in blue, with a yellow swoosh above them.The logo for the Food and Drug Administration (FDA), featuring the letters "FDA" in a stylized, blue, sans-serif font.The logo for the Substance Abuse and Mental Health Services Administration (SAMHSA), featuring a stylized figure and the text "SAMHSA" in a bold, black, sans-serif font.The logo for the National Institutes of Health (NIH), featuring the letters "NIH" in white, with a blue arrow pointing to the right.The logo for the Centers for Disease Control and Prevention (CDC), featuring the letters "CDC" in white, with a blue background.The logo for the Defense Threat Reduction Agency (DTRA), featuring a globe and the text "DTRA" in a bold, black, sans-serif font.The logo for the North Carolina Department of Health and Human Services (NC), featuring the letters "NC" in a bold, black, sans-serif font.The logo for the Environmental Protection Agency (EPA), featuring a stylized green plant and the letters "epa" in a bold, black, sans-serif font.The logo for the City of Medicine in Durham, featuring a blue square with white stars and the text "DURHAM" and "1869 CITY OF MEDICINE" in a bold, black, sans-serif font.The logo for the Greensboro Police, featuring the text "Greensboro Police" in a bold, black, sans-serif font, with a police badge icon.The logo for the Laura and John Arnold Foundation (LJAF), featuring the letters "ljaf" in a stylized, blue, sans-serif font.The logo for the Bill & Melinda Gates Foundation, featuring the text "BILL & MELINDA GATES foundation" in a bold, black, sans-serif font.

Who Do We Look For?

People who are:

- Analytically minded
- Curious
- Continuously learning
- Team-oriented
- Communicative
- Entrepreneurial
- Humble / good humor



And share our interest...

Building a vibrant data science team within an organization with a compelling social mission...*Data for Good*



Identifying Arrest-Related Deaths: The Problem



President Obama, in response to his *Task Force on 21st Century Policing*, March 2, 2015:

“Right now, we do not have a good sense, and local communities do not have a good sense, of how frequently there may be interactions with police and community members that result in a death”.

Crowd-Sourced Lists

At least 304 black people were killed by police in the United States in 2014.

Get the facts about police violence in your community to make the case for change.

On August 9, 2014, Michael Brown Jr. was murdered by Officer Darren Wilson, sparking nationwide protests against police killings of black people. This map, a project of [WeTheProtesters.org](#), bears witness to the black men and women who have been killed at the hands of law enforcement in 2014. And while a focus on police killings cannot capture the full scale of the violence our communities face at the hands of police, we hope this data helps communities better understand the problem and begin to make progress towards addressing it.



- At least 1149 people were killed by police in 2014. 304 (26%) were black.
- Black people were nearly 3x more likely than whites to be killed by police in 2014.
- At least 101 unarmed black people were killed by police in 2014. 40% of all unarmed victims were black.
- Police killed at least 16 more black people in 2014 than in 2012.
- Where you live matters. A black person is 5x more likely to be killed by police in St. Louis than in New York City. A black person is 10x more



M

Search Medium

Sign in / Sign up



Guardian US on Jun 1 · 7 min

Nitish Pahwa and 207 others recommended



NEXT STORY

Killed By Police

Stephen Meyer
July 13 at 6:21pm

Perhaps the US police should change their motto from "To serve and protect" to "kill or incarcerate"?

Like · Comment 2

Charlie Guy
July 9 at 2:19pm

Justice for all justice for Tiphne walk downtown Jacksonville if you have lost a loved one please come join us or a friend due to an act of violence or senseless violence please come join us July 20th 5:15 p.m. Behind the police station at 501 East Bay Street please this is justice for all walk I need everyone who have lost their loved ones or There friends to a senseless act of violence to come out and be a part of this walk on July 20th supporters are welcome to come out and walk with us as well please share. Representative of the the TDH Foundation and hosted by the TDH foundation

Killed By Police
July 15 at 4:19pm · Edited

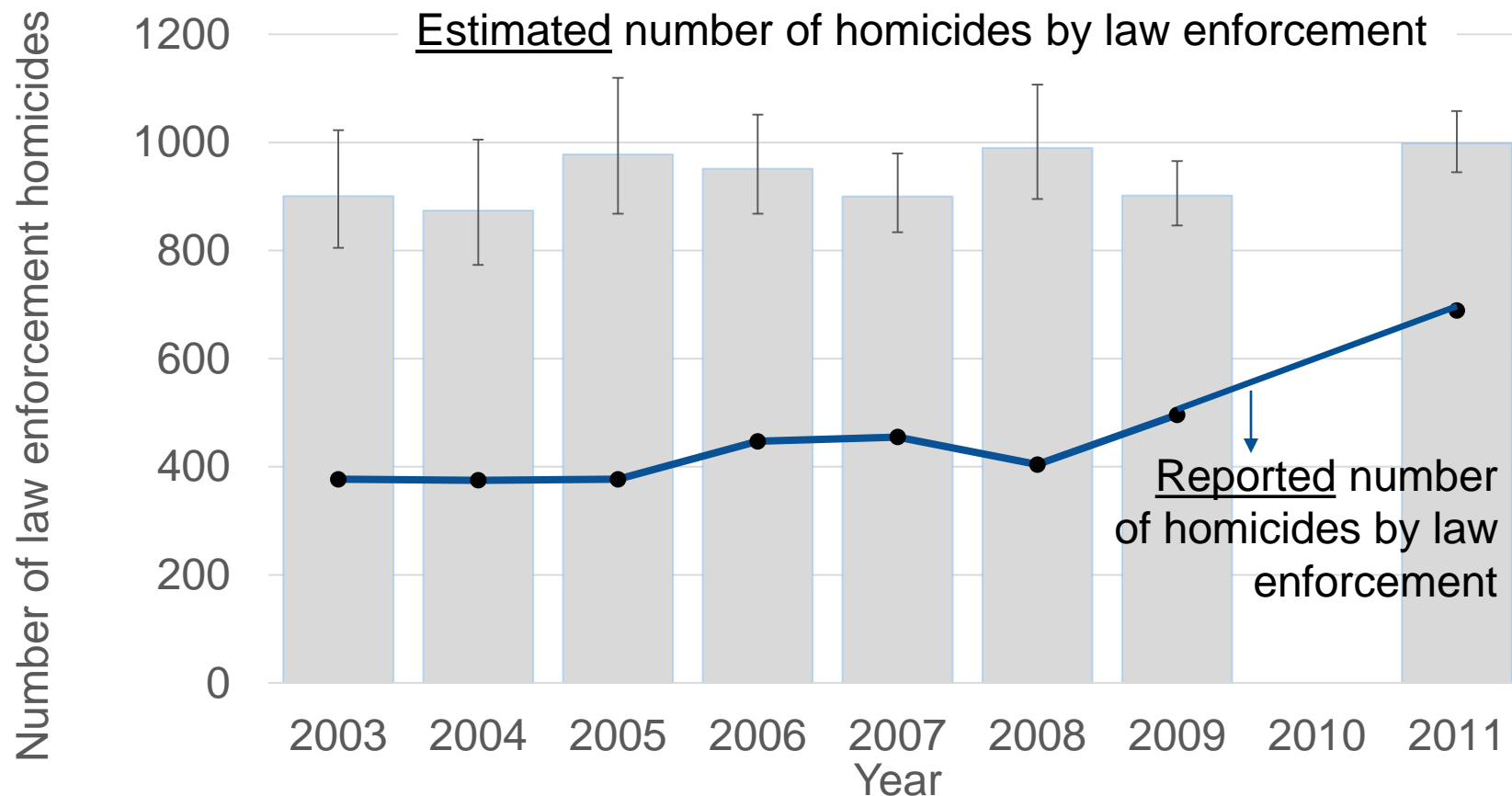
(2494) July 15, 2015
Eugene Kailing, 43



MSP trooper fatally shoots 'erratic' person in Osceola Co. | Michigan

BJS coverage assessment findings

Number of homicides by law enforcement reported to the ARD program and estimated coverage, 2003–2009 and 2011



Source: Banks, D., Couzens, G.L. & Planty, M.G. (2015, October). *Assessment of Coverage in the Arrest-Related Deaths Program* (NCJ 249099). Bureau of Justice Statistics, US Department of Justice.

Hybrid Approach to Data Collection

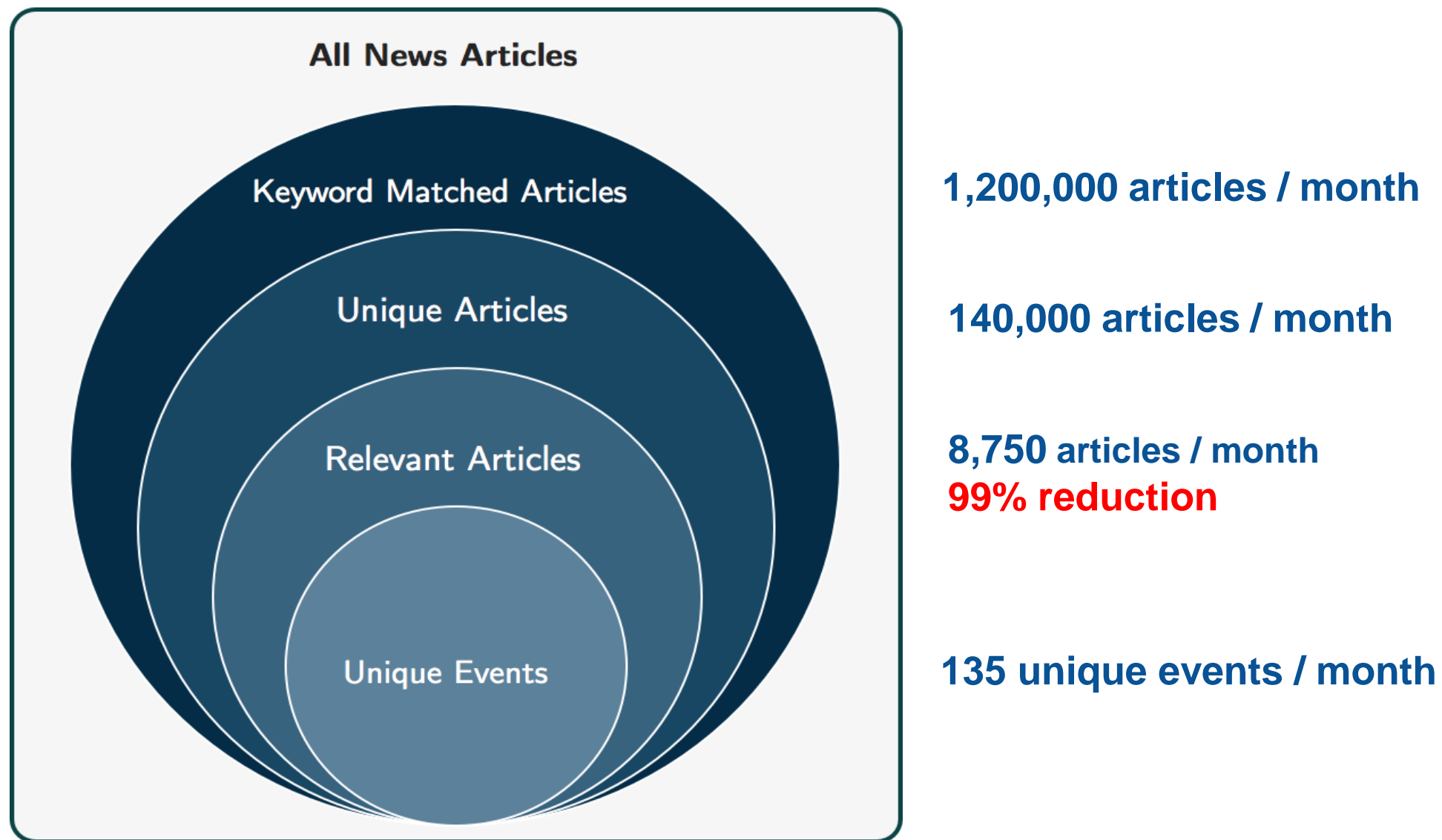
Phase 1: Identification

- Review news alerts and other **public sources** for potential arrest-related deaths.
- Track efficiency and coverage of various approaches for identifying arrest-related deaths.
- **Compile quarterly lists** of potential arrest-related deaths.

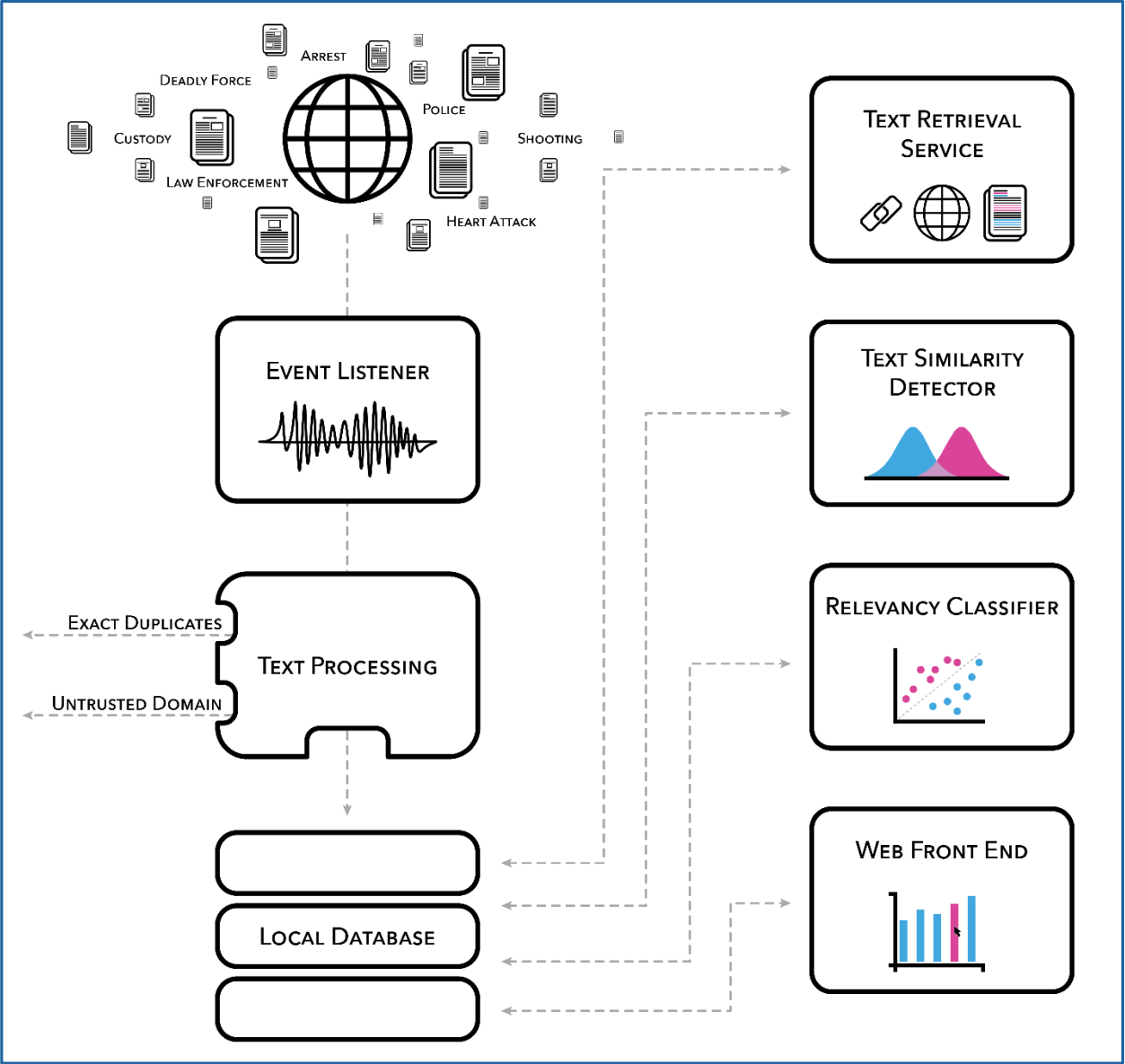
Phase 2: Agency survey

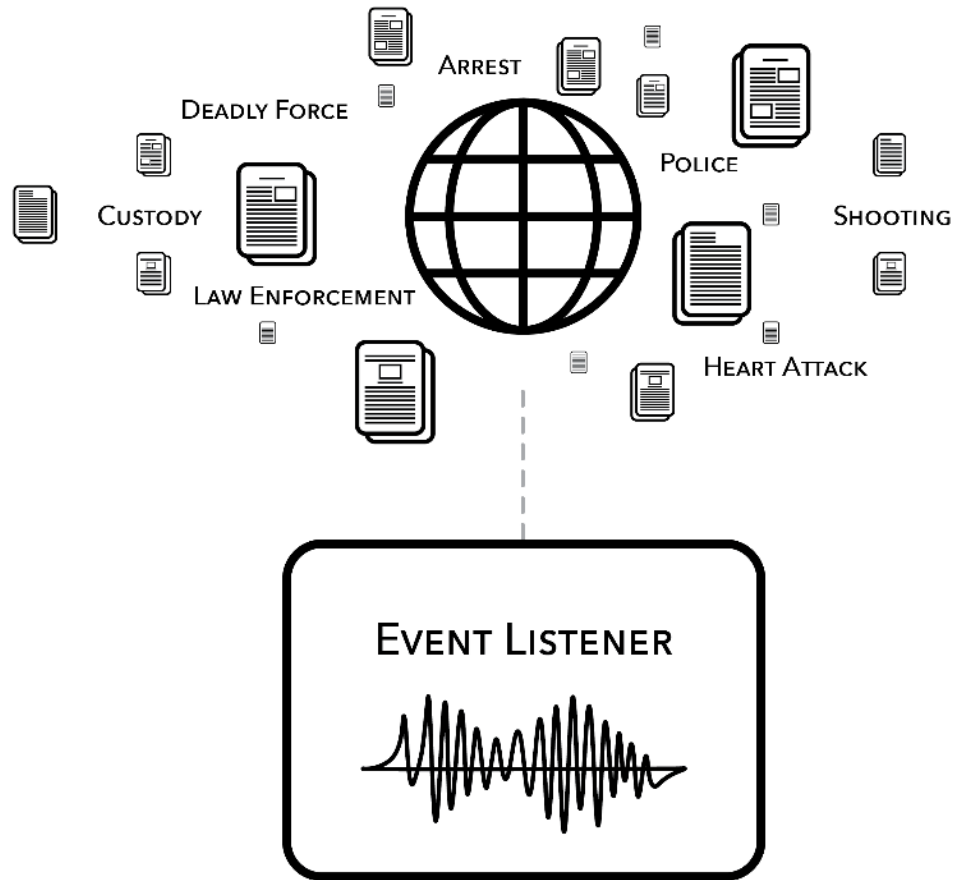
- Conduct follow-up with LE agencies and medical examiner offices to confirm identified deaths and collect more info about circumstances surrounding deaths.
- Survey agencies with identified deaths and others to assess and adjust Phase 1 identification methods.

Using Open Information Sources to Estimate Arrest-Related Deaths



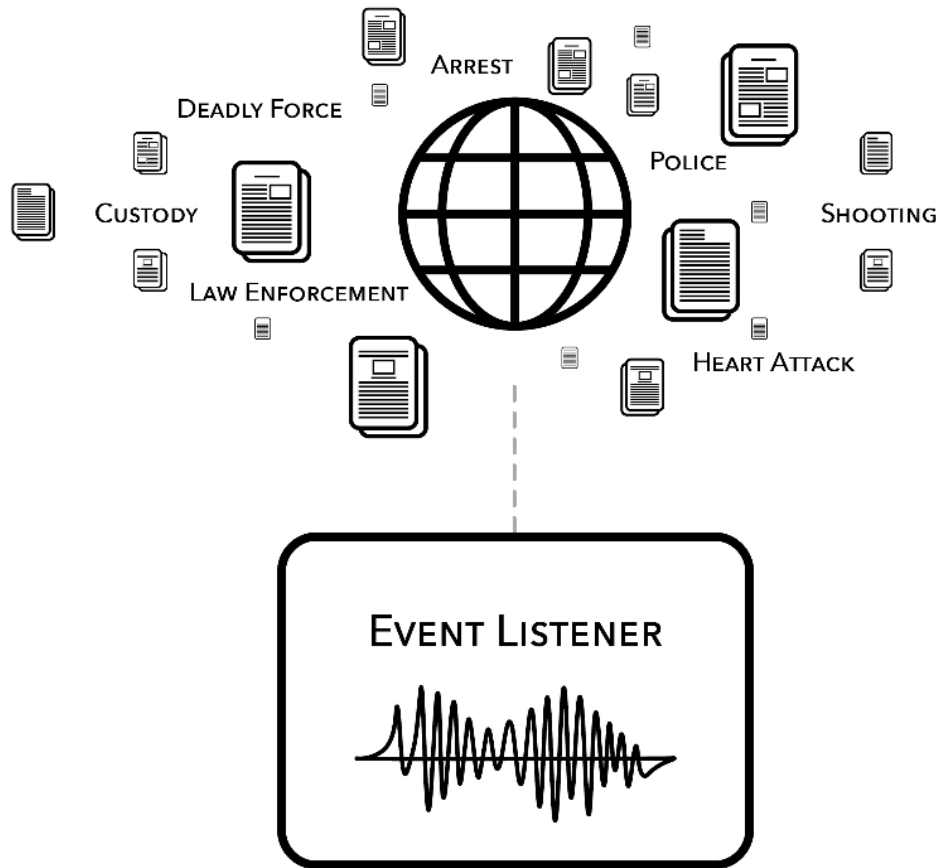
Media Alert Coding and Classification Pipeline





Returns articles with a combination of the primary search term and another term indicating law enforcement involvement such as "police", "officer", "arrested"

- Shoot(ing,er)
- Kill(ed,ing)
- Death, dead, died
- Deadly/lethal force
- Use of force
- Heart attack
- Accidental
- Overdose
- Taser, stun gun
- Standoff

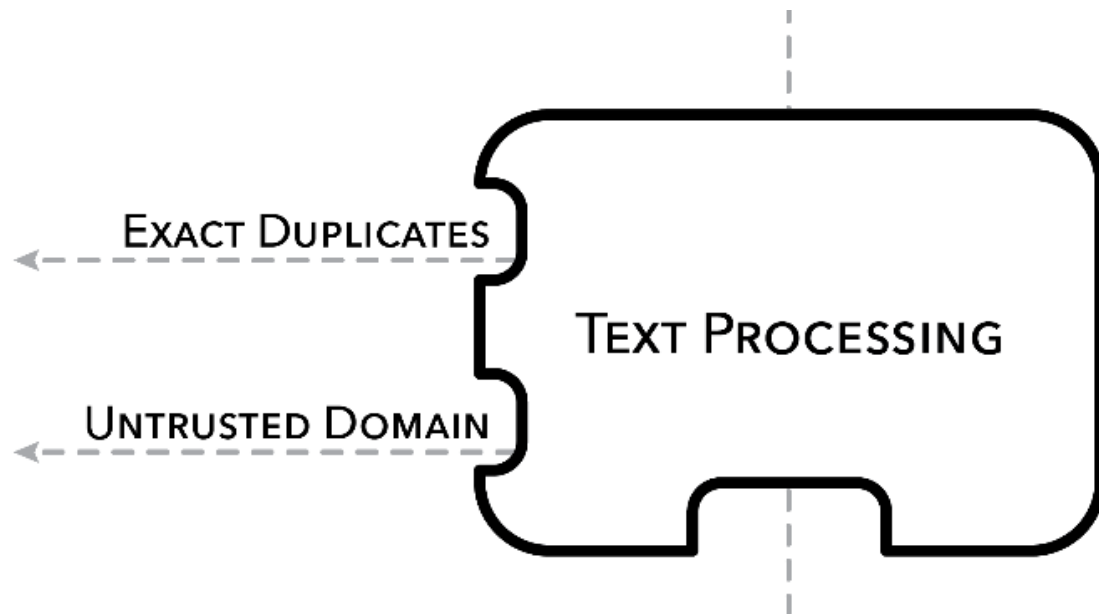


Data Science Problem

- Many potential news stories to review
- Many coders, many spreadsheets
- **Goal:** Reduce human labor without reducing number of individuals identified

1,200,000 articles / month

Text Processing: Exact Duplicates / Untrusted Domains



Remove articles from certain URLs:
530k / mo

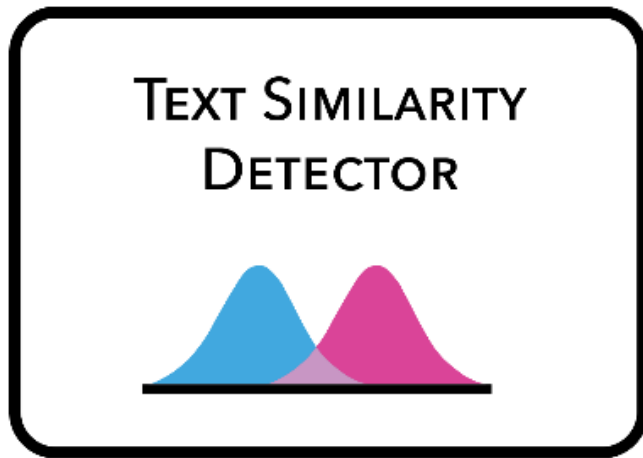
- National news outlets
- Non-news websites

Eliminate exact duplicates:

- URLs: **80k / mo**
- Article texts >40 char long: **77k / mo**

1,200,000 → 515,000 articles / mo

Text Similarity Detector: TF-IDF Similarity



- Compute TF-IDF* similarity for all pairs of articles in 10-day sliding windows
- If a pair has high similarity, drop one
- Compare both texts and titles (above length limit)
- **270k / month reduction**

*Term Frequency-Inverse Document Frequency

A former **South Carolina** police chief has dodged prison time in the 2011 shooting death of an unarmed Black man, receiving just one year of home detention, according to **NPR**.

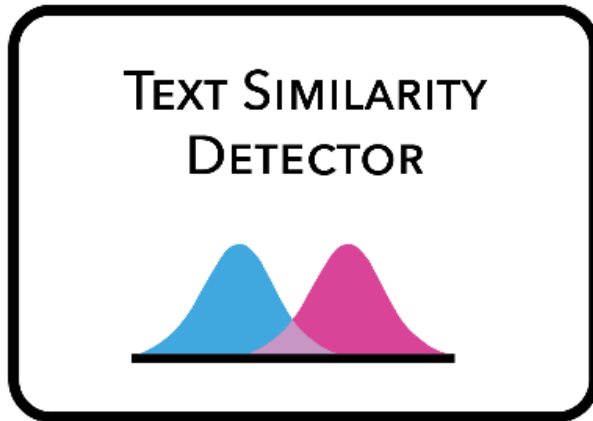
Former Eutawville, S.C. police chief **Richard Combs**, 38, shot and killed 54-year-old **Bernard Bailey** as he left the Eutawville Police Department in May 2011.

A former South Carolina police chief who was charged with murder in the shooting death of a black man in 2011 pleaded guilty to misconduct charges after two separate trials ended in hung juries.

The former Eutawville chief, Richard Combs, had been indicted more than three years after he shot and killed Bernard Bailey, 53, outside the police department. Combs argued that he feared Bailey would use a vehicle as a weapon against him.

1,200,000 → 515,000 → 245,000 articles / month

Text Similarity Detector: Entity Similarity



San Bernardino, California (CNN)—With the investigation still unfolding, much is unclear about Wednesday's deadly San Bernardino shooting at a center for people with developmental disabilities.

[See latest developments](#)

[Update 12:22 a.m ET Thursday]

The family has not been able to track down recently named suspect [Syed Farook](#) or his wife since Wednesday morning, said Hussam Ayloush, executive director of CAIR (the Council on American Islamic Relations).

- Generate set of named entities mentioned in each article
- If sets are similar enough, drop one article
- Compare texts only
- **105k / month reduction**

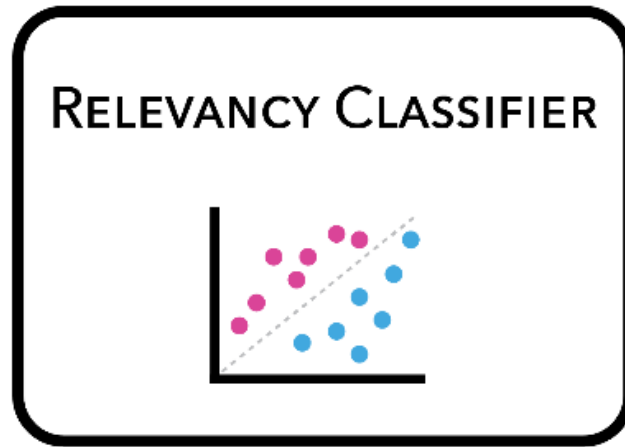
Assailants Syed Farook and Tashfeen Malik killed in shootout with police after deadly attack at holiday party that killed 14 and injured at least 17 more

In the latest burst of deadly gun violence in the U.S., two heavily armed assailants entered a social-services agency in San Bernardino, California, on Wednesday and opened fire during a holiday party.

At least 14 people were killed in the initial attack and another 17 were injured, according to San Bernardino police officials.

About four hours later, local police located a dark-colored SUV and engaged in a shootout that killed two suspects, one male and one female. They were later identified as Syed Farook, 28, and Tashfeen Malik, 27, the latter believed to be Farook's girlfriend or wife.

1,200,000 → 515,000 → 245,000 → 140,000 articles / month




- **Ensemble model** (logistic regression, decision tree classifiers, and a neural network)
- Separate models applied to text and title
- Words positively associated with arrest-related death
 - “shot and killed”, “armed”, “died”, “suspect”
- Words negatively associated with arrest-related death
 - “victim”, “anyone with”, “police investigate”

Remaining volume after all steps: 8,750 articles / mo (99.2% reduction)

1,200,000 → 515,000 → 245,000 → 140,000 → 8,750 articles / mo


ARD

Last Decedent Reviewed:183565 | Articles Reviewed Today: 3

Duren  Log Out

Agent: Officer had conflicting accounts of black man's death

MONTGOMERY, Ala. — A white Alabama police officer charged with killing a black man gave two conflicting accounts of what happened to a state investigator, the agent testified at a hearing Thursday. Officer Aaron Smith is charged with murder in the ...

 [Open source article in a new window](#)

☒ Does Not Meet ARD Definition

☐ Unknown Whether Meets ARD Definition (Needs Follow-Up)

☐ Meets ARD Definition

☐ Meets ARD Definition But Does Not Fall Within Our Date Scope

Save and close

Follow-Up Notes

Notes

First Name

First name

Middle Name


Middle name

Last Name

Last name

Agency Name

Agency name

 News Local and State Education Business Arts Health Entertainment Lifestyle Leisure

24 March 2016 Thursday 17:49

This news clip was read 29 times.


Agent: Officer had conflicting accounts of black man's death

State Bureau of Investigation Agent Jason DiNunzio testified Smith didn't suspect Gunn of a crime when he initially stopped him, but Smith immediately confronted Gunn and told him to put his hands on the hood of the patrol car. DiNunzio said he doesn't...

Share on Facebook

Tweet

Share on Google+


T- T+ A 

Agent: Officer had conflicting accounts of black man's death

MONTGOMERY, Ala. — A white Alabama police officer charged with killing a black man gave two conflicting accounts of what happened to a state investigator, the agent testified at a hearing Thursday.

Officer Aaron Smith is charged with murder in the Feb. 25 death of 58-year-old Greg Gunn. After Thursday's hearing, Judge Jimmv Pool ruled there was probable cause

★ Feature



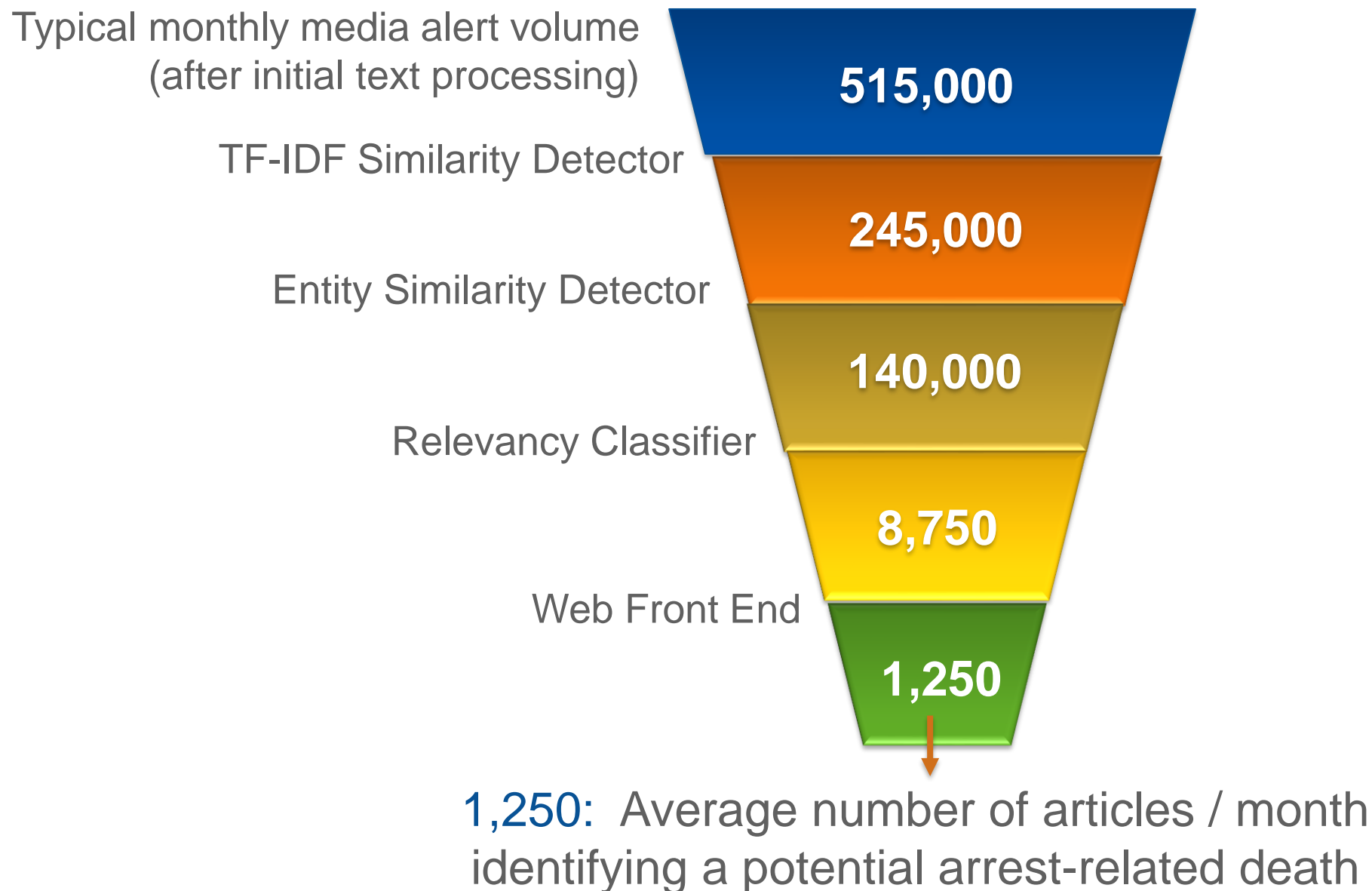
Matt Had hustles t Funny C

Gregory Gunn | Montgomery, AL | Montgomery PD | 2016-02-25 | (96%)

Select Known Decedent

24

Overall reduction in article volume



- What to build vs. what to buy?
- Acceptable amount of loss
- Monitoring of manual coding (quality control)
 - "Unknown" queue
 - "Known decedent" identification and linking
 - Resolving disagreements between coders and the Relevancy Classifier
 - Comparing with external sources

Identifying Arrest-Related Deaths: Results

U.S. Department of Justice
Office of Justice Programs
Bureau of Justice Statistics

Revised December 22, 2016



TECHNICAL REPORT

December 2016, NCJ 250112

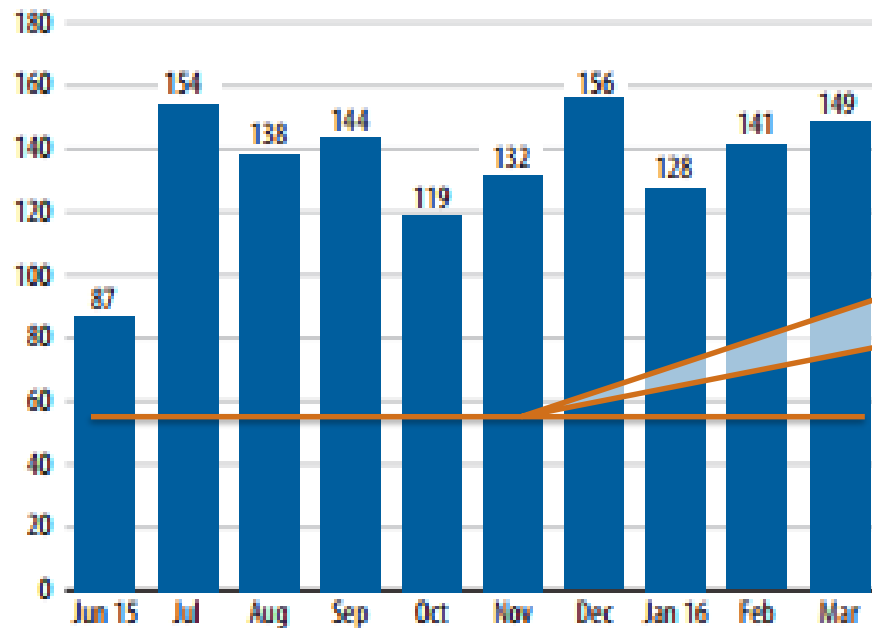
Arrest-Related Deaths Program Redesign Study, 2015–16: Preliminary Findings

Duren Banks, Ph.D., Paul Ruddle, and Erin Kennedy, *RTI International*
Michael G. Planty, Ph.D., *BJS Statistician*

FIGURE 1

Potential arrest-related deaths identified through open source review, June 2015–March 2016

Number of deaths



Average number of law enforcement homicides identified per month in 2011 through the previous ARD program methodology

Source: Bureau of Justice Statistics, Arrest-Related Deaths Program Redesign Study, 2015–16.

Program Findings

Related News



Study estimates 1,900 arrest-related deaths occurred in US between June 2015-May 2016

December 15, 2016

The BJS Arrest-Related Deaths program includes all persons who died during the process of arrest or while in police custody including deaths due to homicide, including justifiable homicide by a law enforcement officer, suicide, accidental injury and natural causes

[View all news for this expert](#)

Technical Report:

Banks, Ruddle, Kennedy, & Planty (2016). *Arrest-related Deaths Program Redesign Study, 2015-16: Preliminary Findings*. Dept of Justice, Office of Justice Programs, Bureau of Justice Statistics.

In the Media

NBC NEWS

New Government Report Records Twice as Many Police Related Deaths as FBI Stats

December 18, 2016

THE GUARDIAN

Killings by US police logged at twice the previous rate under new federal program

December 15, 2016

THE GUARDIAN

US government database hopes to tell 'whole story' of police killings after year of Guardian count

December 12, 2015

Identifying Arrest-Related Deaths: Impact



From *The Guardian*, Dec 13, 2015:

“We should be able to know how many people die in law enforcement custody”

“We couldn’t have done this 30 years ago. Now, because we have all this information available online in a machine-readable context, we are able to narrow down the universe of articles or posts initially without having a person read the universe of articles.”



FiveThirtyEight



Thanks to the open source community for:

- Python
- pandas
- spaCy
- scikit-learn
- Luigi
- Keras
- nltk
- Jupyter
- numpy
- scipy
- TensorFlow
- PostgreSQL

What else could we track from news articles?

Police use of force

- Monthly volume (after initial text processing): 375,000
- Deep learning models based on Facebook's FastText embedding classification system
 - Title: character-level up to 4-grams
 - Text: word level up to 2-grams (bigrams)
- Monthly volume (after applying classifier): 22,000

Hate / bias crimes

- Monthly volume (after initial text processing): 560,000
- Also based on Facebook's FastText embedding classification system
- Monthly volume (after applying classifier): 2,000

In both cases, difficult to verify results from external sources

How Would We Update Our Approach?



<https://github.com/RTIInternational/gobbli/>

Thank You!

**Gayle Bieler
Jason Nance**



Questions?