



Anonymized Target List Expansion for Name Vetting

Christopher Thorpe

703-983-1090

cthorge@mitre.org

MSR

Problem

- **Name vetting is the process of comparing two lists of names (e.g., passenger and terrorists)**
- **Name-vetting systems vary widely in throughput and accuracy**
- **Agencies use various commercial off-the-shelf (COTS) systems, with inconsistent results**
- **Watchlists are distributed across agencies and companies, risking unauthorized disclosure**
- **Improved process required to increase accuracy, improve consistency, and reduce risk of exposure**

Background

Osama bin Laden

Usama bin Ladin

Ossama Bin-Laden

Osamah Bin Ladenas

Oussama Ben Laden

Ousama Binlادن

Usamah Binlادن

Ussama Benlادن

Oossama Benlادن

Osaama ibn Ladin

Usame ibn Laden

Osaamah ibn Laden

Osamaa ibn Ladin

Ossamah Benlادن

Ousaama Benlادن

Ousamah Binlادن

Oussamah Binlادن

Ussamah Ben Laden

Oszama Bin Ladenas

Uszama Bin-Laden

Usamat bin Ladin

Usoma bin Laden



U. Bin Laden

O. Bin Laden

Sheikh Usamah Bin-Muhammad Bin-'Awad Bin-Ladin

Usoma Muhammad 'Avad bin Ladin

Bin Ladens Osama Muhameds Avads

Objective

- Apply MITRE's unique joint expertise in name-matching technologies (CAASD Transportation Security + C2C Human Language Technology)
- Methodically and scientifically explore the challenges of anonymized name vetting
- Build a limited prototype that incorporates name variant generation and exact match
- Evaluate performance against existing name-matchers

Activities



- **Researched and wrote “Name Variant Generation and Normalization for Anonymized Vetting”**
- **Designed prototype name-vetting architecture**
- **Designed evaluation methodology**
- **Leveraged Federal Identity Matching Working Group evaluation tools and large dataset of deceased persons**
- **Surveyed variant generation tools and dictionaries**
- **Implemented exact-match prototype with primitive normalization, variant generation, and scoring**
- **Evaluated multiple runs against ground truth data sets**
- **Submitted MITRE Intellectual Property Disclosure for “multilateral variant generation” process**

Highlight

Prototype v6, Test Set #3

Run date: 7 MAR 2008

Query list: 200 names (deceased; profile = domestic aviation)

Watchlist: 5,000 names (deceased; profile = terrorist database)

Expanded Watchlist: 2.2 million names (5,000 + variants)

Results: Precision = .71, Recall = .83

Sample matches (expanded query was exact match to EWL)

JAMES LEWI COOK

JAMES COOK

J F RODRIGUEZ

JOSE RODRIGUEZ

ABRAHAM MOHAMED

AMIN A MOHAMMED

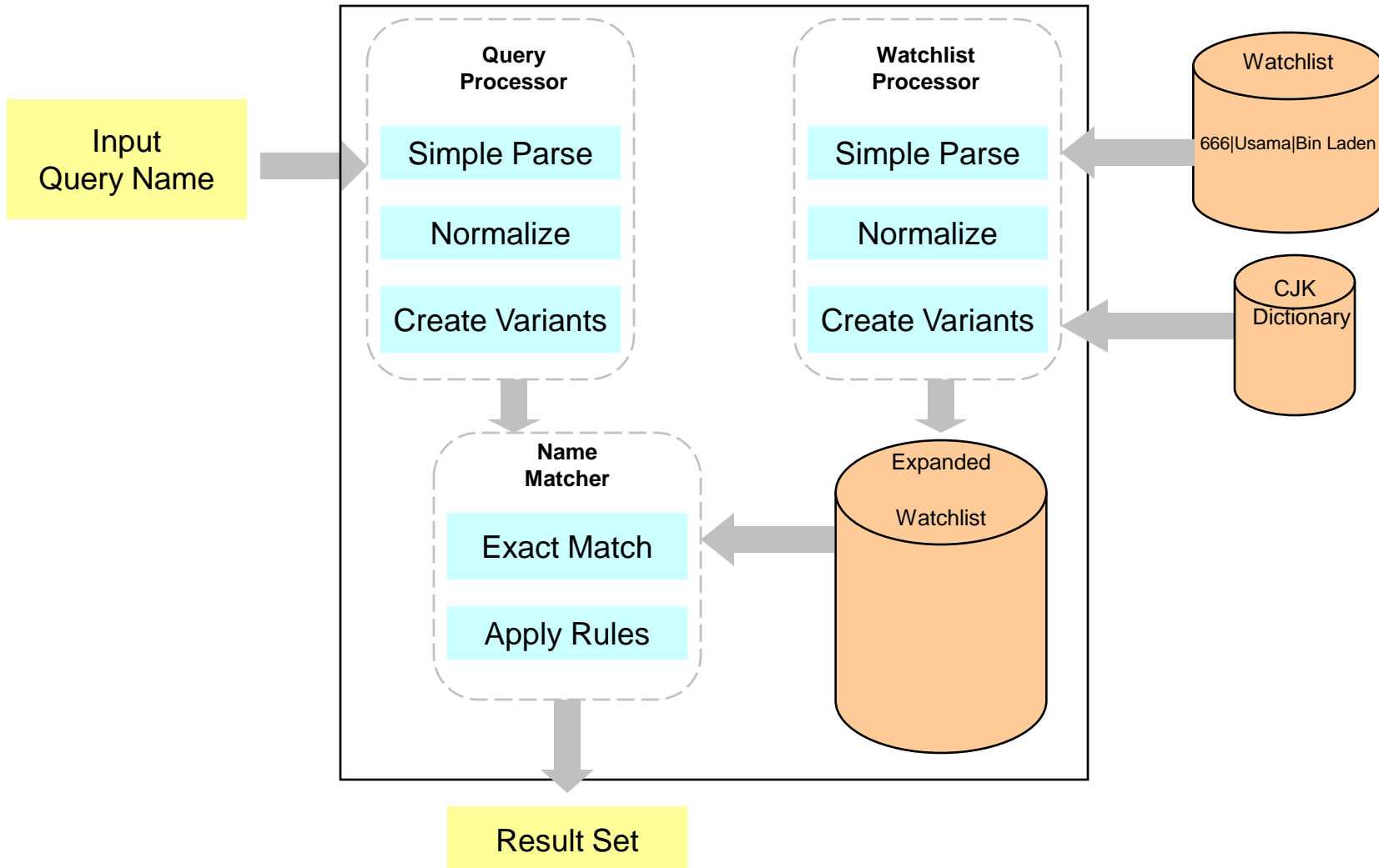
ABRAHAM MOHAMED

SHARIFA A MUHAMMAD [gender mismatch]

A HAKIM

LOREEN A HAKEEM

Demonstration



Impacts



- **Research community**
 - **Contributes to an understanding of the theoretical issues in anonymized matching of highly variable data**
- **Customers**
 - **Highly focused matching reduces misidentifications**
 - **Intelligent variant generation improves detection**
 - **Anonymization reduces risk of exposure**
- **Vendors**
 - **Potential technology transfer**

Future Plans

