# Natural Language Processing for Anonymization

John Aberdeen & Lynette Hirschman

781 271 2840 • aberdeen@mitre.org

MSR

MITRE Innovation Program

# Problems

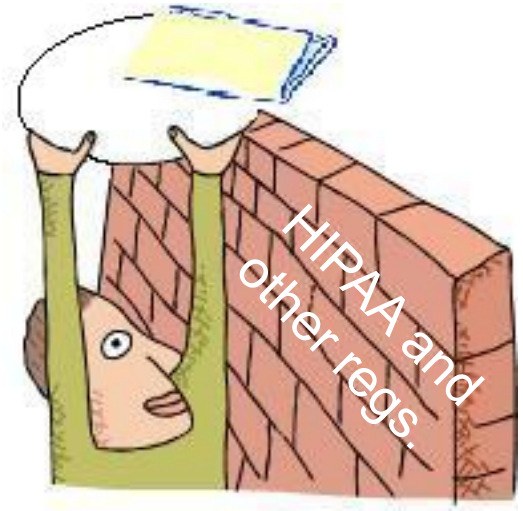
- Privacy constraints are a bottleneck to sharing data containing personally identifiable information (PII), e.g., medical or surveillance records
  - Data holders need tools to remove sensitive information - but researchers/tool developers cannot see the real data.
- Machine-learning based Natural Language Processing (NLP) techniques work well for de-identification of named entities
  - But they require ground truth data to train them.
- End users need a robust, adaptable turn key system that they can train and adapt to their data and their requirements.

© 2009 The MITRE Corporation. All rights reserved

# Background

**Records with personally identifiable information (PII) cannot be shared due to privacy constraints.**

Medical or Surveillance Records containing PII



- **Medical studies**
- **Analyst reports**

- Medical record processing is labor-intensive—automated techniques have been slow to develop.
- Analysts must present reports with identities masked.



© 2009 The MITRE Corporation. All rights reserved

# Objectives

- Create a de-identification system that can be rapidly tailored to multiple domains and record types.
- Partner with real-world users of de-identification to tailor and evaluate systems.
- Package the system to enable data holders and end users to train the system on their data.
- Achieve performance levels that are useful for real applications.

© 2009 The MITRE Corporation. All rights reserved

# Activities

- We have created a domain-portable de-identification system MAT: MITRE Anonymization Tool.
- We have partnered with two medical providers who are real-world users of de-identification technology.
- We have developed resynthesis techniques to convert anonymized data from research partners into realistic, usable training data
  - Because of privacy considerations, we only have access to scrubbed data in most instances.
- We have evaluated 2 applications and are publishing results
  - Accuracy on 1st partner's real hospital records: 0.99
  - 2nd partner has used our tools to build a new corpus and de-identification system with ~12 person-hours of annotation (accuracy 0.99).

© 2009 The MITRE Corporation. All rights reserved

# Highlight

- **Before:**

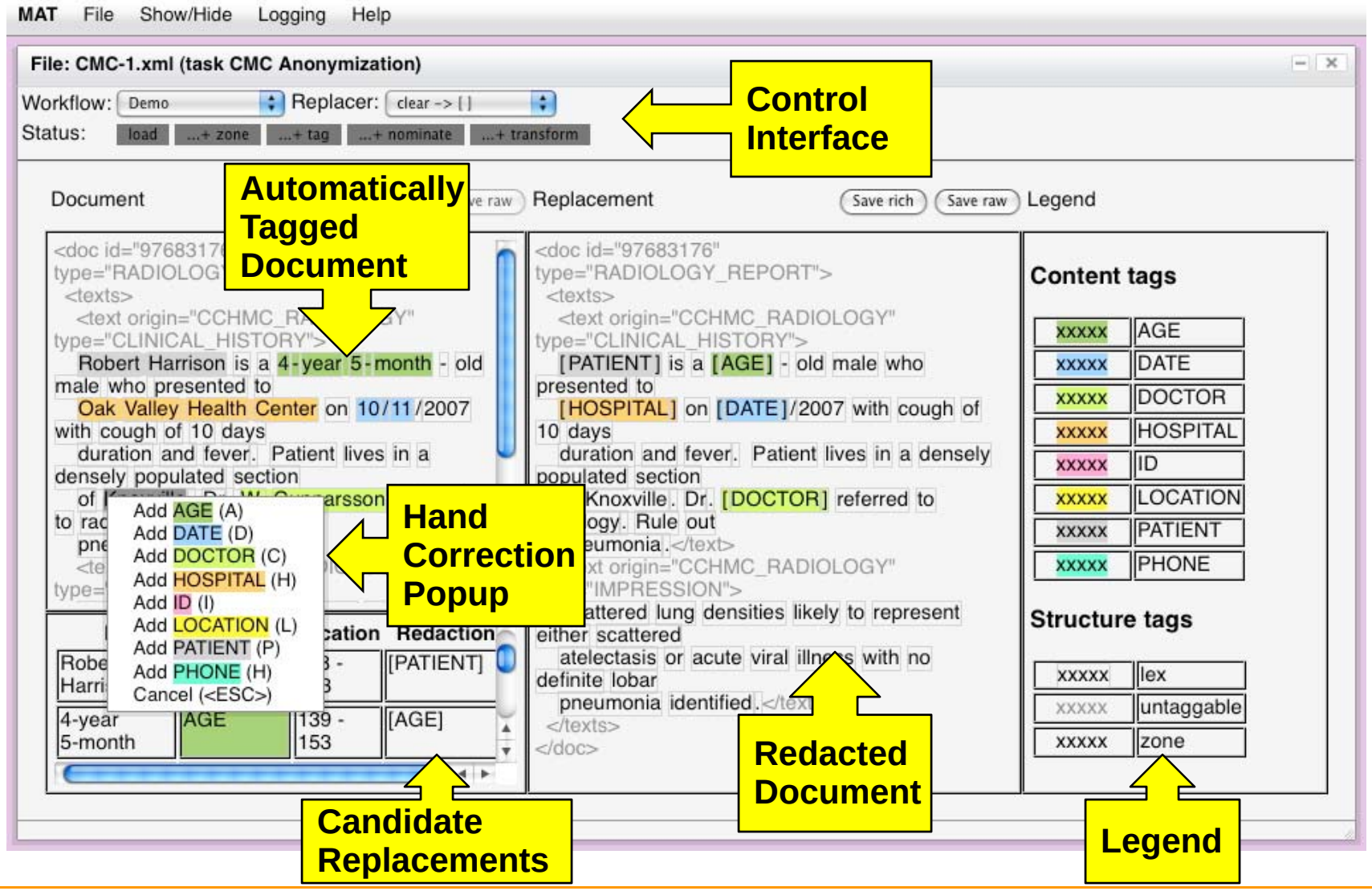
**HISTORY OF PRESENT ILLNESS**: The patient is a 77-year-old-woman with long-standing hypertension who presented as a walk-in to me at the **Oak Valley Health Center** on **July 9th** . Recently had been started q.o.d. on Clonidine since **May 5th** to taper off of the drug. Was told to start Zestril 20 mg. q.d. again. The patient was sent to the **Smith Cardiac Unit** for direct admission for cardioversion and anticoagulation, with the Cardiologist, Dr. **Pearson** to follow.

- **After:**

**HISTORY OF PRESENT ILLNESS**: The patient is a 77-year-old-woman with long-standing hypertension who presented as a walk-in to me at the **[HOSPITAL XX XX XX XX]** on **[DATE - YY]** . Recently had been started q.o.d. on Clonidine since **[DATE - ZZ]** to taper off of the drug. Was told to start Zestril 20 mg. q.d. again. The patient was sent to the **[HOSPITAL VV VV VV]** for direct admission for cardioversion and anticoagulation, with the Cardiologist, Dr. **[DOCTOR QQ]** to follow.

© 2009 The MITRE Corporation. All rights reserved

# Demonstration




© 2009 The MITRE Corporation. All rights reserved

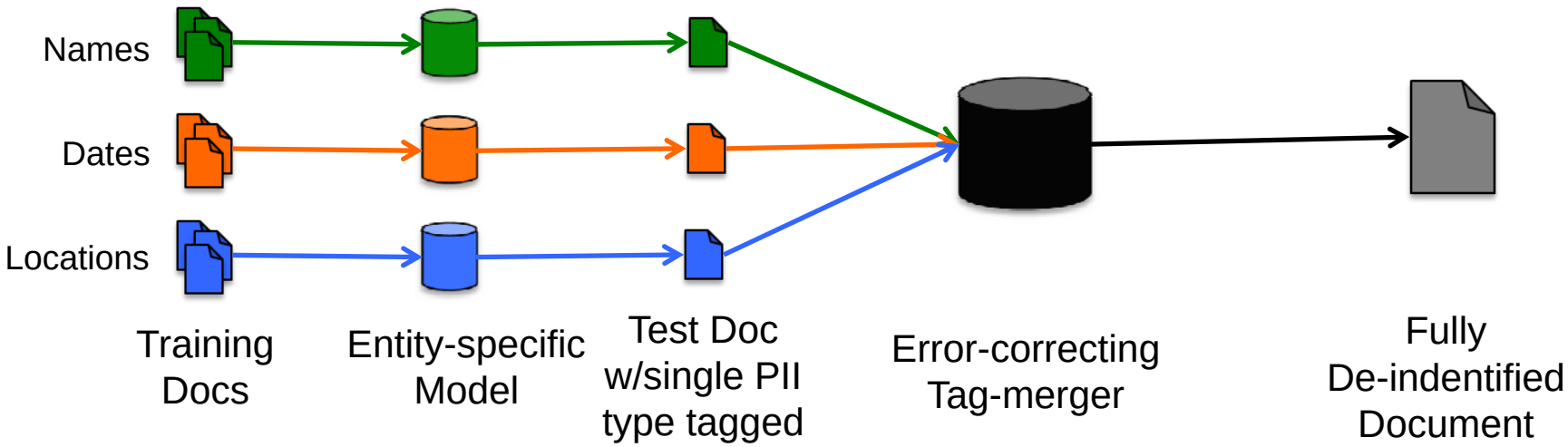
# Impacts

- Breaking through a critical bottleneck in handling records with privacy constraints, such as medical records and surveillance reports.
- Building partnerships with key stakeholders in the medical community and in government agencies that use data containing PII.
- Positioning MITRE as a leader in developing, evaluating, integrating, and disseminating NLP technology for de-identification of free text records.

© 2009 The MITRE Corporation. All rights reserved

# Future Plans

- **Combine entity-specific taggers to increase performance, training efficiency, and rapid adaptation to new document types.**




- **Enables reuse of taggers across applications (e.g., for date or phone number - or even person name identification).**
- **Enables rapid reconfiguration to a new tag set.**
- **Enables precise fine-tuning of taggers for specific types of PII**
  - **Human annotation can be focused on PII types that are particularly important or need improvement.**

© 2009 The MITRE Corporation. All rights reserved