Automated anonymization of protected personal data in clinical reports

AZAD DEHGHAN, PHD

CATHAL GALLAGHER

deepCognito

d-wise

Artificial intelligence

Natural Language Processing (NLP) / Information Extraction (IE)

- A brief History
  - The Turing Test: “Computing Machinery and Intelligence”, 1950
  - Message Understanding Conferences (end of 80’s/90’s)

Methods

- Data-driven: statistical learning also known as “machine learning”
- Knowledge-driven: information extraction rules, lexical resources, and ontologies
- Hybrid (combination of above)

Background

Secondary Use

- Public Domain Sharing
- Internal/External Controlled Re-use
- Strategic Sharing
- Academic Sharing

EMA Policy 0070

- EMA will publish anonymized CSRs in the public domain
- Phase 1: publication of clinical reports
- Phase 2: publication of individual patient data
- Requirement to measure patient re-identification risk

GDPR & Industry Dynamics

- Transparency has ascended the agenda
- Regulatory landscape continues to evolve
- Safe sharing requires anonymization & risk management

Blur

Data Anonymization

Risk Management

CSR Anonymization

What are protected personal data?

PPD IDENTIFIERS
ANTHROPOMETRIC DATA
DATE
FREE TEXT
CONTACT INFORMATION
IDENTIFICATION NUMBERS
LOCATION DATA
NAME
SENSITIVE DATA
SOCIAL IDENTIFY
SOCIOECONOMIC DATA
OTHER PERSONAL DATA

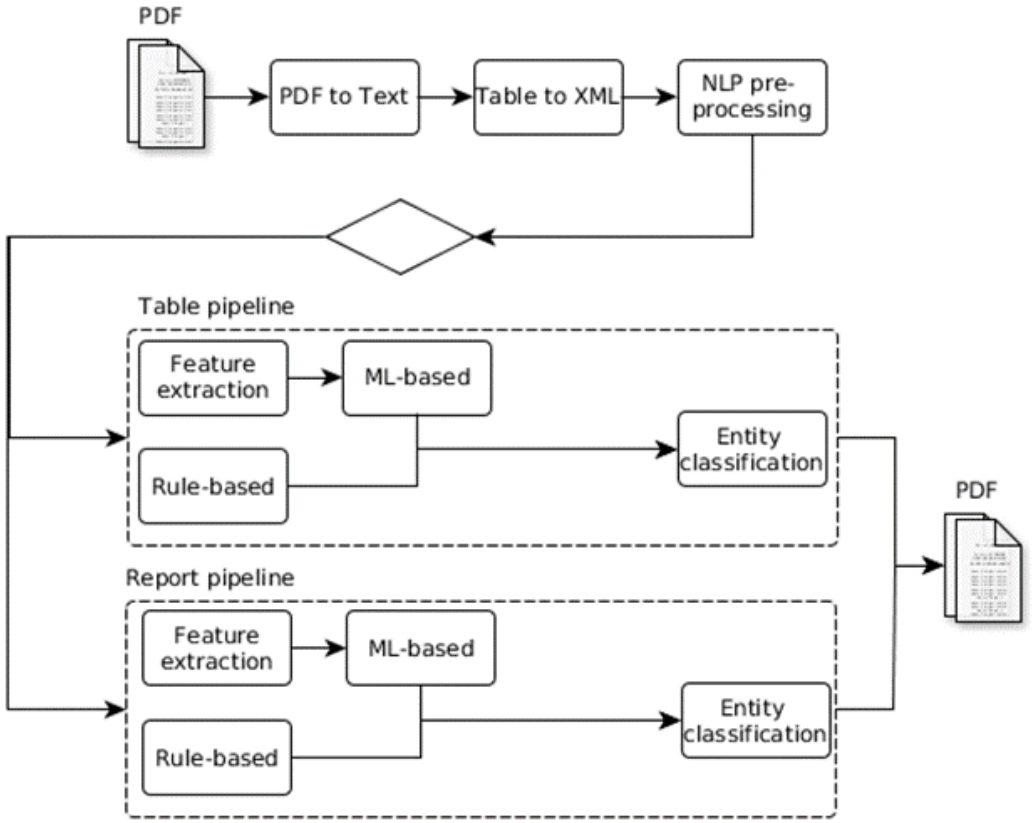
What are protected personal data?

Enabling propagation Blur Risk

Entities

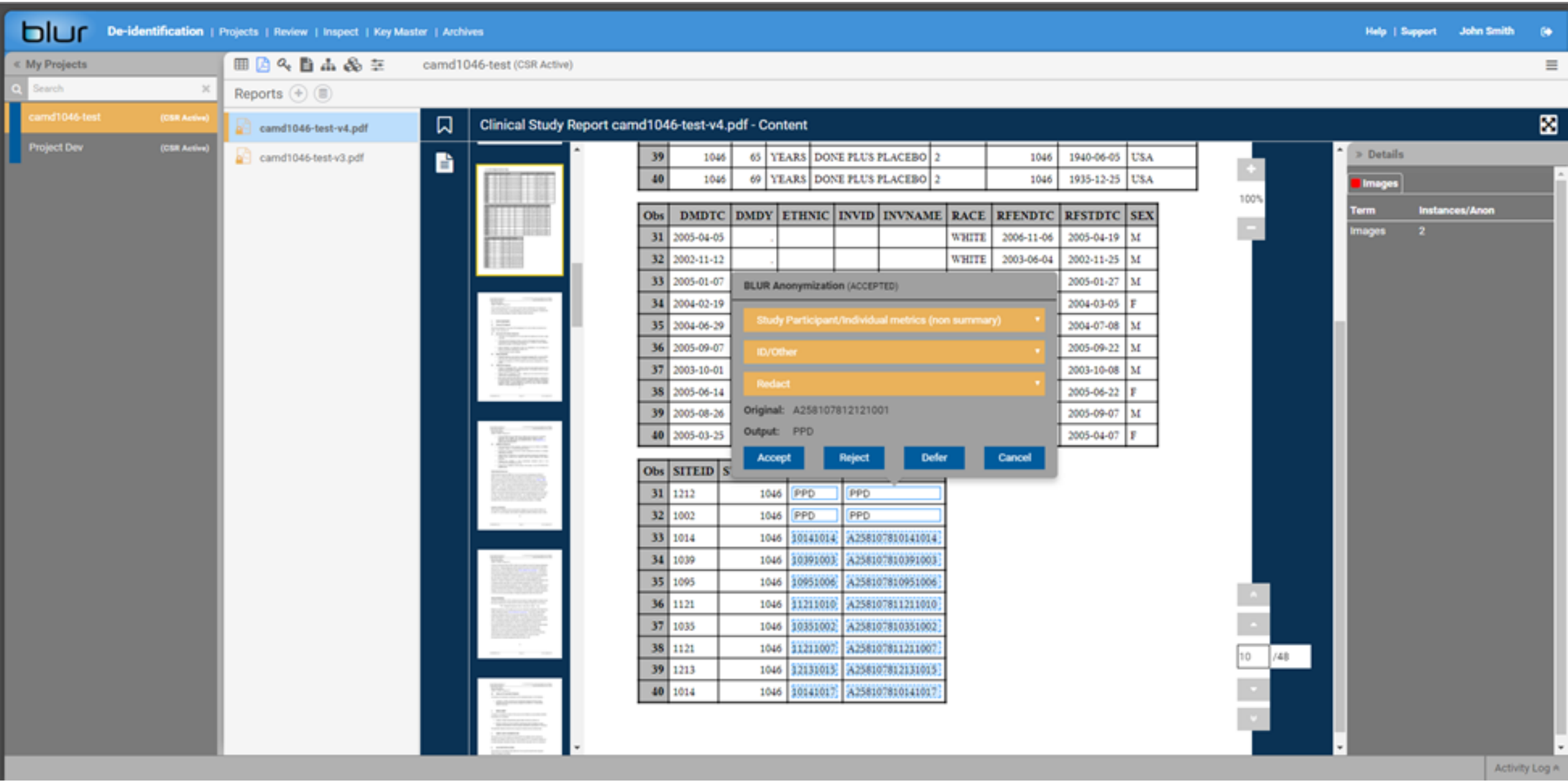
- Aggregate or summary data
- Contract research organization or vendor
- Other persons (e.g., medical monitors, committee members, etc.)
- Other pharmaceutical staff members
- Principal or coordinating Investigator
- Sponsor non-signatory
- Sponsor signatory
- Study participant
- Study site staff
- Vendor (as principal investigator)

NLP Engine Architecture



Blur workflow

1. User uploads study data and documents
2. User performs anonymization work on study data
3. User transition to CSR anonymization
4. User executes “find PPD” operation
5. System submits anonymized study data and CSR content to Blur NLP
6. User accepts/rejects annotations using GUI (see Figure 2)
7. User downloads proposed document for EMA approval
8. User makes modifications based on EMA feedback.
9. User downloads final redacted CSR


blur
De-identification | Projects | Review | Inspect | Key Master | Archives
Help | Support
John Smith
My Projects
Search
camd1046-test (CSR Active)
Project Dev (CSR Active)
camd1046-test (CSR Active)
Reports
camd1046-test-v4.pdf
camd1046-test-v3.pdf
Clinical Study Report camd1046-test-v4.pdf - Content
39 1046 65 YEARS DONE PLUS PLACEBO 2 1046 1940-06-05 USA
40 1046 69 YEARS DONE PLUS PLACEBO 2 1046 1935-12-25 USA
Obs DMDTC DMDY ETHNIC INVID INVNAME RACE RFENDTC RFSTDTC SEX
31 2005-04-05 WHITE 2006-11-06 2005-04-19 M
32 2002-11-12 WHITE 2003-06-04 2002-11-25 M
33 2005-01-07 2005-01-27 M
34 2004-02-19 2004-03-05 F
35 2004-06-29 2004-07-08 M
36 2005-09-07 2005-09-22 M
37 2003-10-01 2003-10-08 M
38 2005-06-14 2005-06-22 F
39 2005-08-26 2005-09-07 M
40 2005-03-25 2005-04-07 F
BLUR Anonymization (ACCEPTED)
Study Participant/Individual metrics (non summary)
ID/Other
Redact
Original: A258107812121001
Output: PPD
Accept
Reject
Defer
Cancel
Obs SITEID
31 1212 1046 PPD PPD
32 1002 1046 PPD PPD
33 1014 1046 10141014 A258107810141014
34 1039 1046 10391003 A258107810391003
35 1095 1046 10951006 A258107810951006
36 1121 1046 11211010 A258107811211010
37 1035 1046 10351002 A258107810351002
38 1121 1046 11211007 A258107811211007
39 1213 1046 12131015 A258107812131015
40 1014 1046 10141017 A258107810141017
100%
10 /48
Details
Images
Term Instances/Anon
Images 2
Activity Log

NLP Performance

Using customary Information extraction metrics:

$$\textbf{Precision} = True\ positives / True\ positives + False\ positives$$

$$\textbf{Recall} = True\ positives / True\ positives + False\ negatives$$

$$F_\beta\text{-score} = (1+\beta\uparrow 2)\, Precision \times Recall / \beta\uparrow 2 \times Precision + Recall;$$

where β=1, as we accept equivalent importance between Precision and Recall.

Human benchmark

Human annotators

- Inter annotator agreement (IAA)
  - Human IAA: 99% micro F_1-score

Blur NLP engine

- 99% micro F_1-score

Discussion

Why not **accuracy**?

- Technically flawed and inaccurate
  - Textual data are skewed toward true-negatives
- Difficult to not achieved 99%
  - 500 subject identifiers distributed evenly
  - 100 pages with an average of 500 words per page

Precision, Recall, F-score

- Do not consider true-negatives
- Not sensitive to skewed distributions

Summary

- **Artificial Intelligence** / Natural Language Processing
- Policy 0070
- **Blur de-identification** and **anonymization**
- **Blur Risk**
- Protect Personal Data
- **Blur** workflow
- **Blur** NLP engine
- **Blur** NLP performance