Integrating **Public** and **Private** Medical Texts for Patient De-Identification with Apache cTAKES

Presented By: Andrew McMurry & Britt Fitch
(Apache cTAKES committers)

Co-authors:
Guergana Savova, Ben Reis, Zak Kohane

*Clinical variables in **patient notes** often NOT available in coded EHR*

- ✓ BMI
- ✓ Smoking status
- ✓ Family history of disease
- ✓ Pathology lab results
- ✓ Medication adherence
- ✓ Lifestyle factors
- ✓ …

High Quality Phenotypes

*Clinical variables in **patient notes** often NOT available in coded EHR*

- ✓ BMI
- ✓ Smoking status
- ✓ Family history of disease
- ✓ Pathology lab results
- ✓ Medication adherence
- ✓ Lifestyle factors
- ✓ …

X

Confidential Patient Identifiers

De-identifying Private Medical Text

Human experts

- Laborious $$
- Fatigue errors

Automation (machine learning)

- Training set from human annotators = small
- Training local features limits general utility

Reversing the De-ID task

"What are the chances that a word or phrase would occur in a medical journal or medical dictionary?"

P (~phi | PublicText)

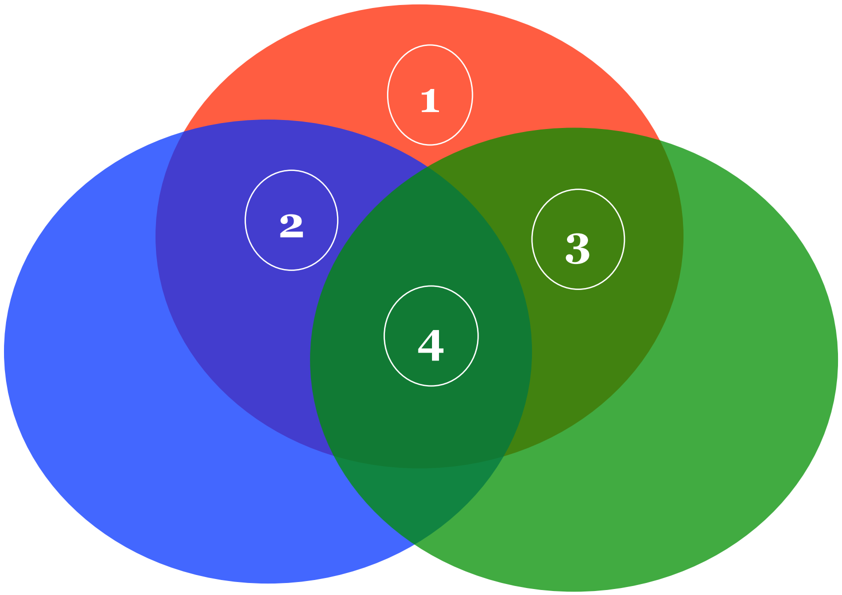
vs

P (phi | PrivateText)

Public Medical Text

- Open Access on the rise !
- Publications: natural language examples of medical topics
- Ontologies: concept codes for medical topics
- Automation: millions of training examples available

Physician Notes



Medical Publications

Medical Ontologies

1. Nouns and Numbers that only occur in Physician Notes are probably PHI.
2. Words that occur frequently in medical publications are probably NOT PHI.
3. Words and phrases in in medical ontologies are probably not PHI.
4. Words shared in all three medical text sources are very unlikely to contain PHI.

Data

Public Medical Text

- 10,000 Journal Publications
- 10 UMLS Ontologies

Private Medical Text

- I2b2 De-Identification Challenge

APPROACH

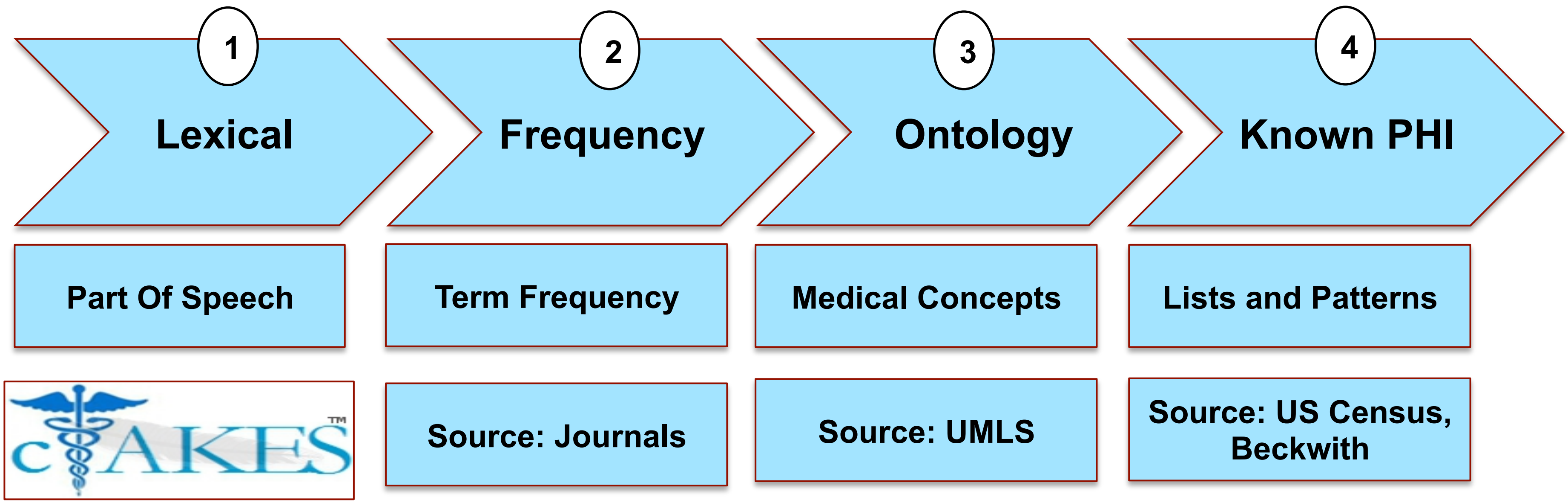
- **Annotate**
  - Sentence boundaries
  - Tokenization
  - Part of Speech tagging
  - Named Entity Recognition

- Learn ***background*** distribution from PUBLIC text

- Learn ***properties*** of PHI from fewer human annotations

- Classify new data: more like public text or private text ???

Annotation Pipeline



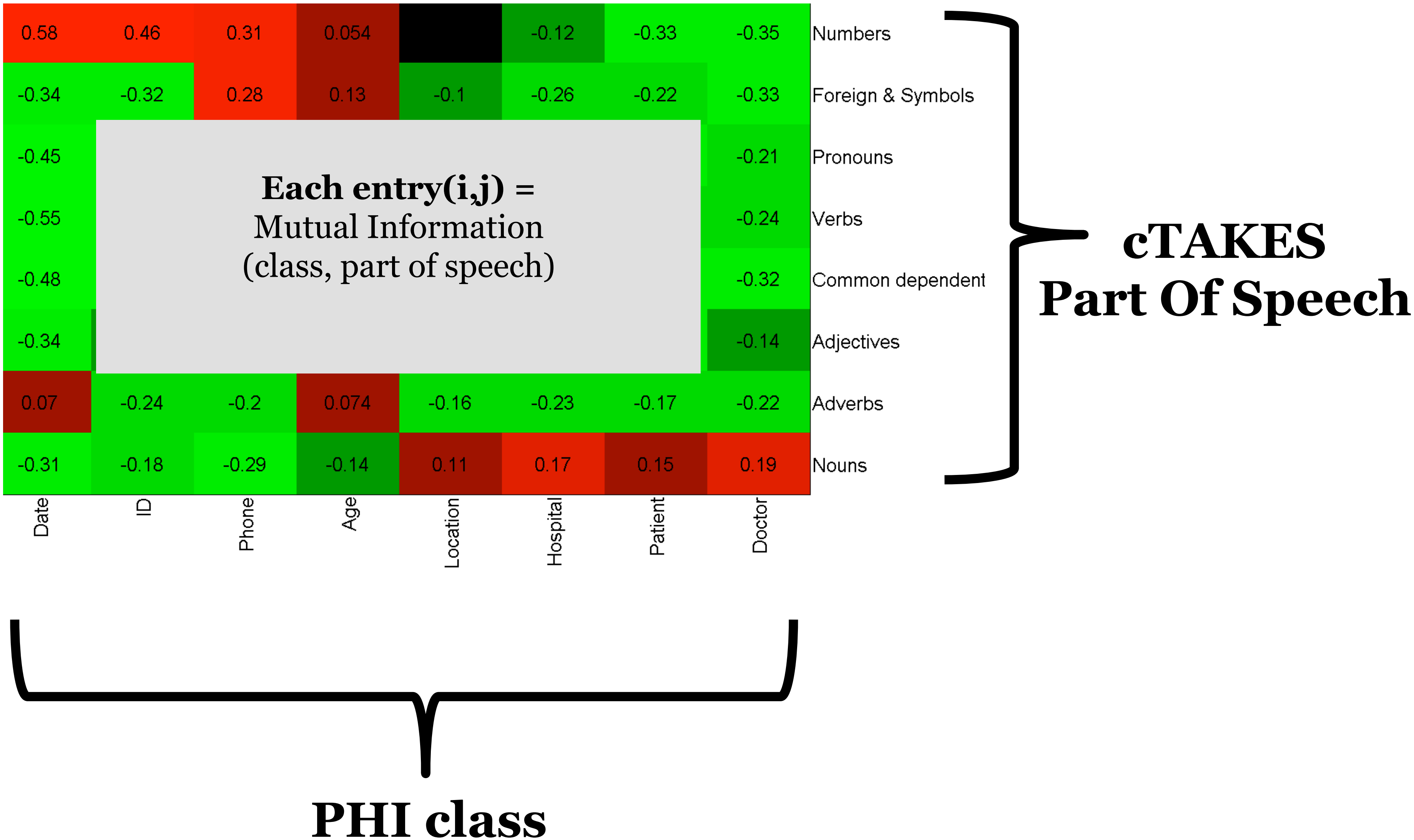
Annotation Pipeline

① Lexical Phase: split document into sentences, tag part of speech for each token.
② Frequency Phase: calculate term frequency with and without part of speech tag.
③ Ontology Phase: search for each word/phrase in ten UMLS ontologies
④ Known PHI Phase: match US census names and textual patterns for each PHI type.

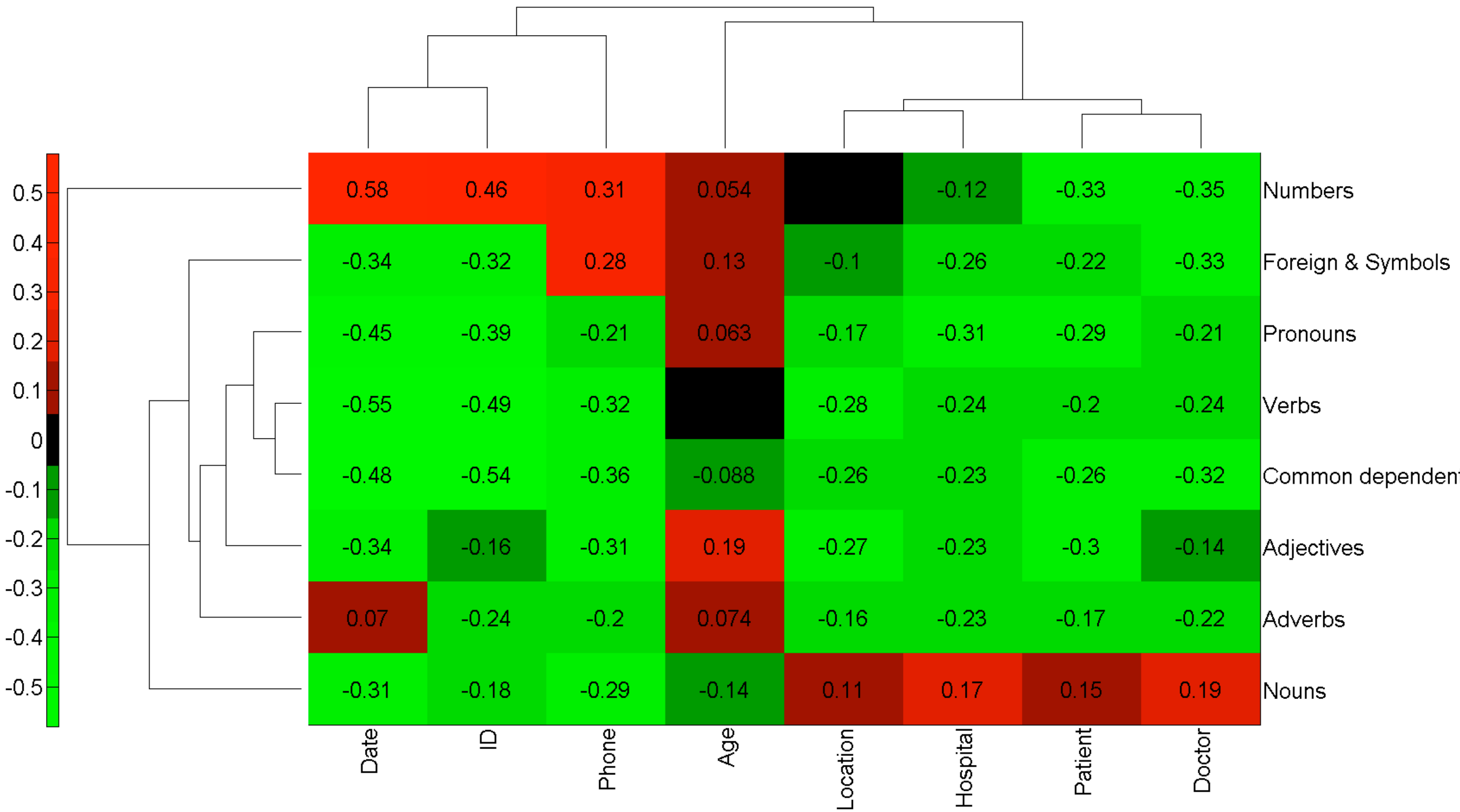
Results

1 Lexical	2 Frequency	3 Ontology	4 Known PHI
Part Of Speech	Term Frequency	Medical Concepts	Lists and Patterns
cTAKES™	Source: Journals	Source: UMLS	Source: US Census, Beckwith

Part Of Speech highly informative for PHI classification



PHI classes cluster by Part Of Speech



(Biclustering normalized MI distance)

	Date	ID	Phone	Age	Location	Hospital	Patient	Doctor
Numbers	0.58	0.46	0.31	0.054		-0.12	-0.33	-0.35
Foreign & Symbols	-0.34	-0.32	0.28	0.13	-0.1	-0.26	-0.22	-0.33
Pronouns	-0.45	-0.39	-0.21	0.063	-0.17	-0.31	-0.29	-0.21
Verbs	-0.55	-0.49	-0.32		-0.28	-0.24	-0.2	-0.24
Common dependent	-0.48	-0.54	-0.36	-0.088	-0.26	-0.23	-0.26	-0.32
Adjectives	-0.34	-0.16	-0.31	0.19	-0.27	-0.23	-0.3	-0.14
Adverbs	0.07	-0.24	-0.2	0.074	-0.16	-0.23	-0.17	-0.22
NOUNS	-0.31	-0.18	-0.29	-0.14	0.11	0.17	0.15	0.19

People and Places

	Date	ID	Phone	Age	Location	Hospital	Patient	Doctor
NUMBERS	0.58	0.46	0.31	0.054		-0.12	-0.33	-0.35
Foreign & Symbols	-0.34	-0.32	0.28	0.13	-0.1	-0.26	-0.22	-0.33
Pronouns	-0.45	-0.39	-0.21	0.063	-0.17	-0.31	-0.29	-0.21
Verbs	-0.55	-0.49	-0.32		-0.28	-0.24	-0.2	-0.24
Common dependent	-0.48	-0.54	-0.36	-0.088	-0.26	-0.23	-0.26	-0.32
Adjectives	-0.34	-0.16	-0.31	0.19	-0.27	-0.23	-0.3	-0.14
Adverbs	0.07	-0.24	-0.2	0.074	-0.16	-0.23	-0.17	-0.22
Nouns	-0.31	-0.18	-0.29	-0.14	0.11	0.17	0.15	0.19

Patient Numbers

★ **NOUNS**: people and places

★ **NUMBERS**: patient IDs and dates

✓ **ADJECTIVES**: clinical description

"*severe* chronic obstructive pulmonary disease"

✓ **VERBS**: clinical action

"*decreased* white blood cell count"

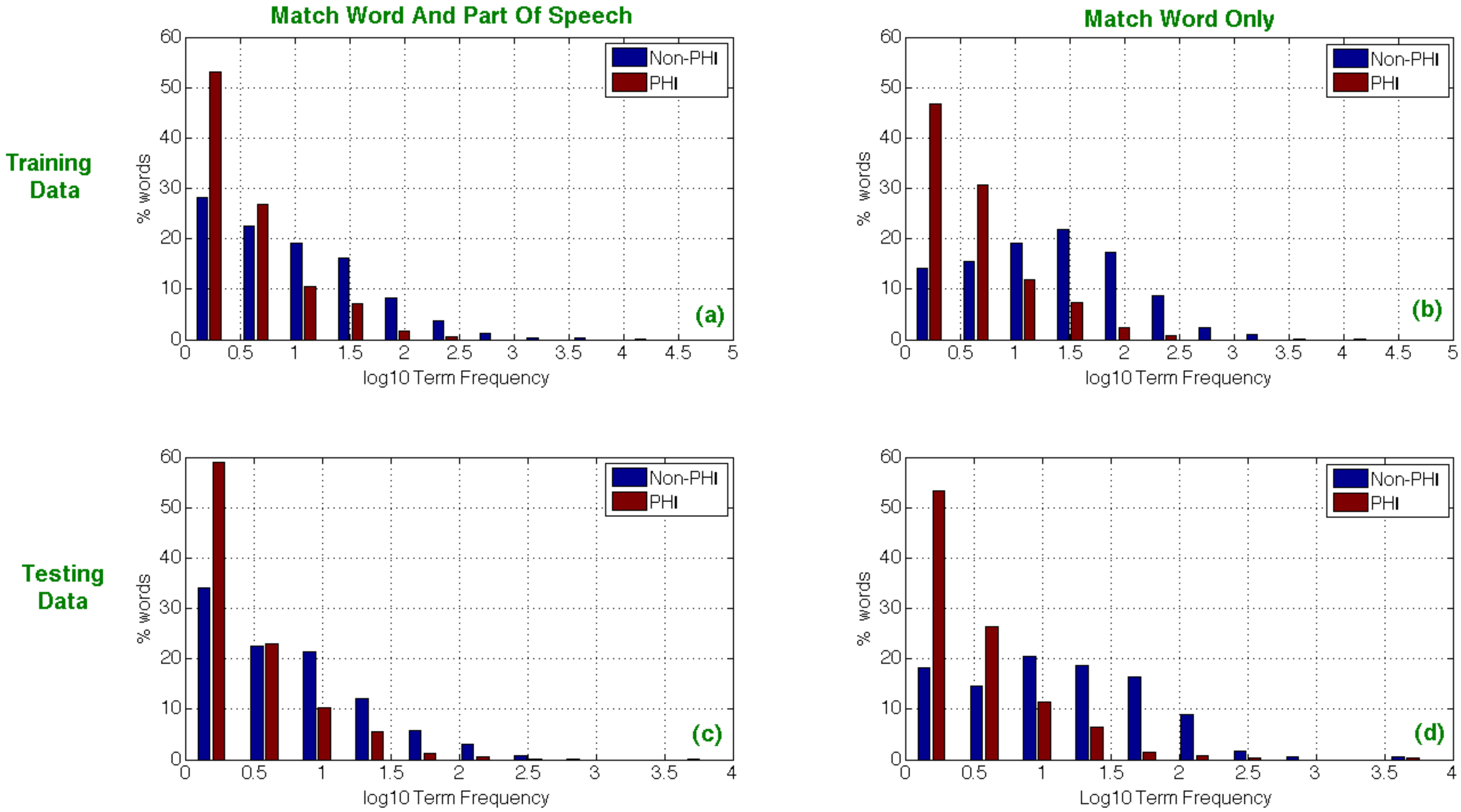
Other part of speech

Common words and dependencies, etc.

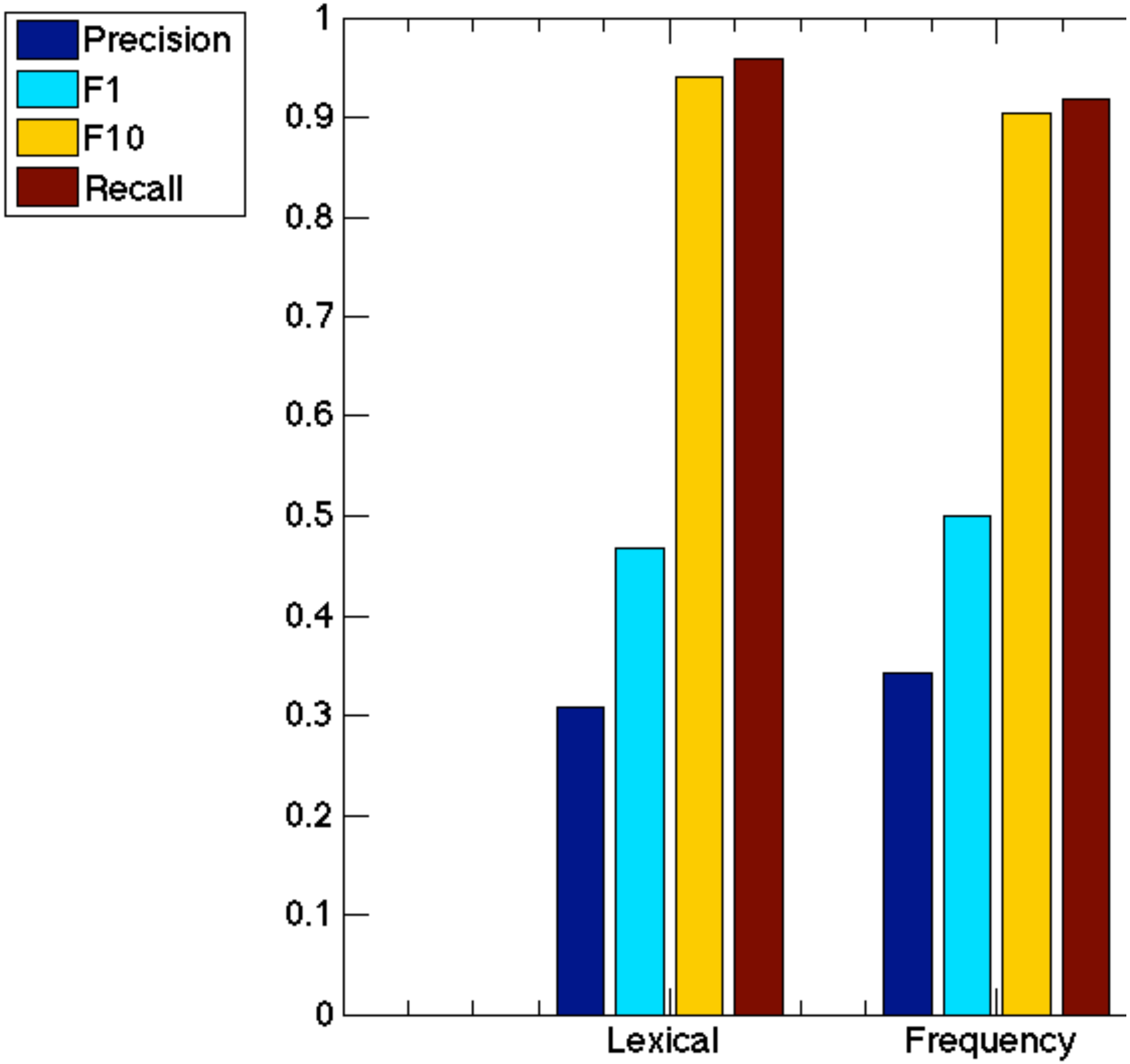
Results

1. **Lexical**
   - Part Of Speech
   - cTAKES™
2. **Frequency**
   - Term Frequency
   - Source: Journals
3. **Ontology**
   - Medical Concepts
   - Source: UMLS
4. **Known PHI**
   - Lists and Patterns
   - Source: US Census, Beckwith

Term Frequency in 10,000 Journal Articles



Classifier Performance



Results

1 Lexical	2 Frequency	3 Ontology	4 Known PHI
Part Of Speech	Term Frequency	Medical Concepts	Lists and Patterns
cTAKES™	Source: Journals	Source: UMLS	Source: US Census, Beckwith

Part of Speech + TF not enough

"Mr. **Huntington** suffers from **Huntington's** disease. He was admitted to a hospital on **Huntington** street"

--Professor Szolovits Example*

Part of Speech + TF not enough

"Mr. **Huntington** suffers from **Huntington's** disease. He was admitted to a hospital on **Huntington** street"

--Professor Szolovits Example*

①Noun Person
②Noun Disease

③Noun Location

same Term Frequency

Match Longest UMLS Phrase

# Concepts	Dictionary
3,461	COSTAR
5,020	HL7V2.5
8,062	HL7V3.0
102,048	ICD10CM
253,708	ICD10PCS
40,491	ICD9CM
327,181	LOINC
739,161	MESH
437,307	RXNORM
1,170,855	SNOMEDCT

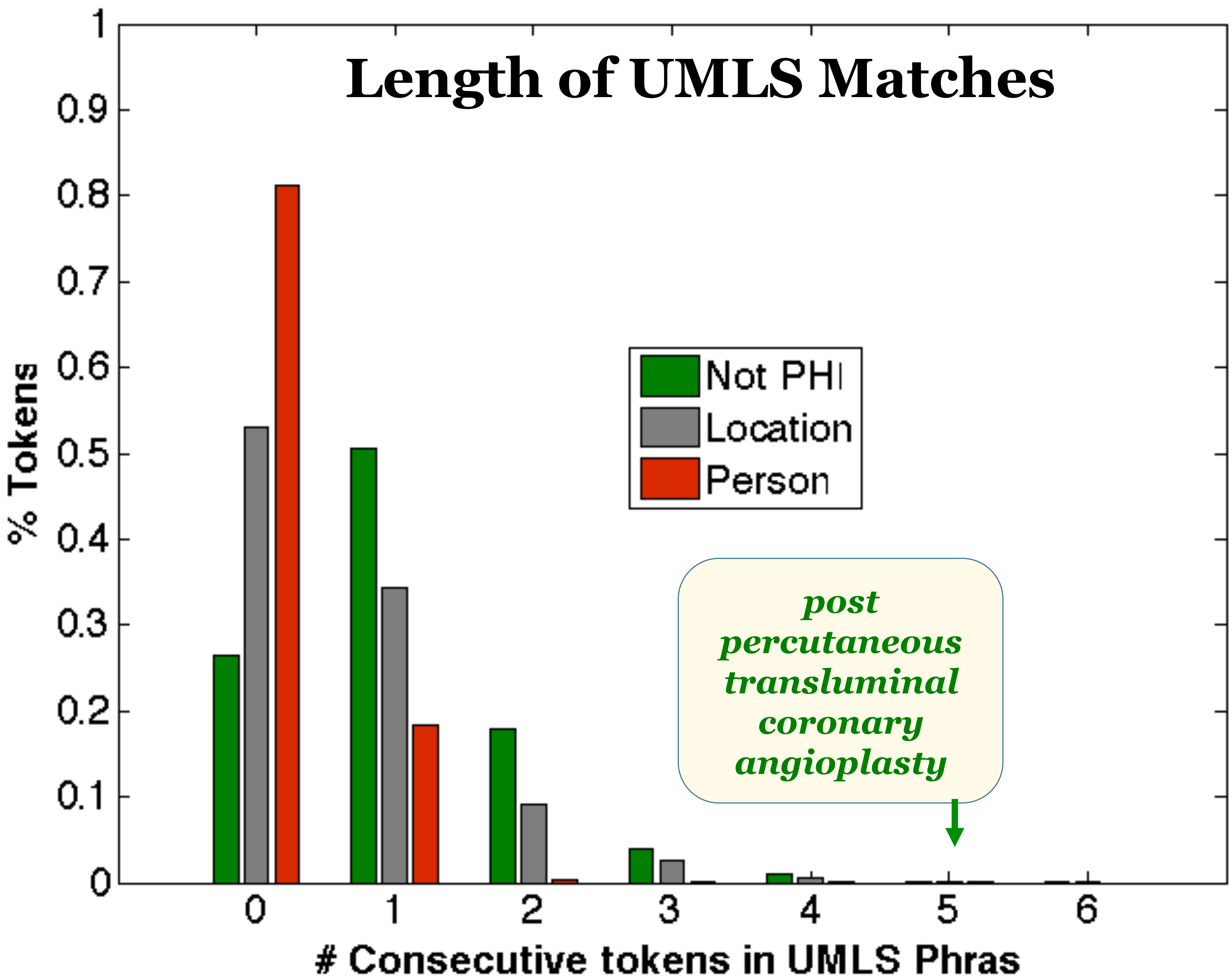
HUNTINGTON DIS
HUNTINGTONS DIS

Huntington Chorea
Huntington's Chorea

Huntington's dementia

Huntington's disease*

Huntington's disease in last 7D
Huntington's disease Multiple sclerosis in last 7 days


Length of UMLS Matches
% Tokens
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
1
2
3
4
5
6
Consecutive tokens in UMLS Phras
Not PHI
Location
Person
post
percutaneous
transluminal
coronary
angioplasty

Results

1 Lexical	2 Frequency	3 Ontology	4 **Known PHI**
Part Of Speech	Term Frequency	Medical Concepts	**Lists and Patterns**
cTAKES™	Source: Journals	Source: UMLS	**Source: US Census, Beckwith**

Known PHI lists

Software

Highly accessed

Open Access

Development and evaluation of an open source software tool for deidentification of pathology reports

Bruce A Beckwith[1,2]*, **Rajeshwarri Mahaadevan**[2], **Ulysses J Balis**[2,3] and **Frank Kuo**[2,4]

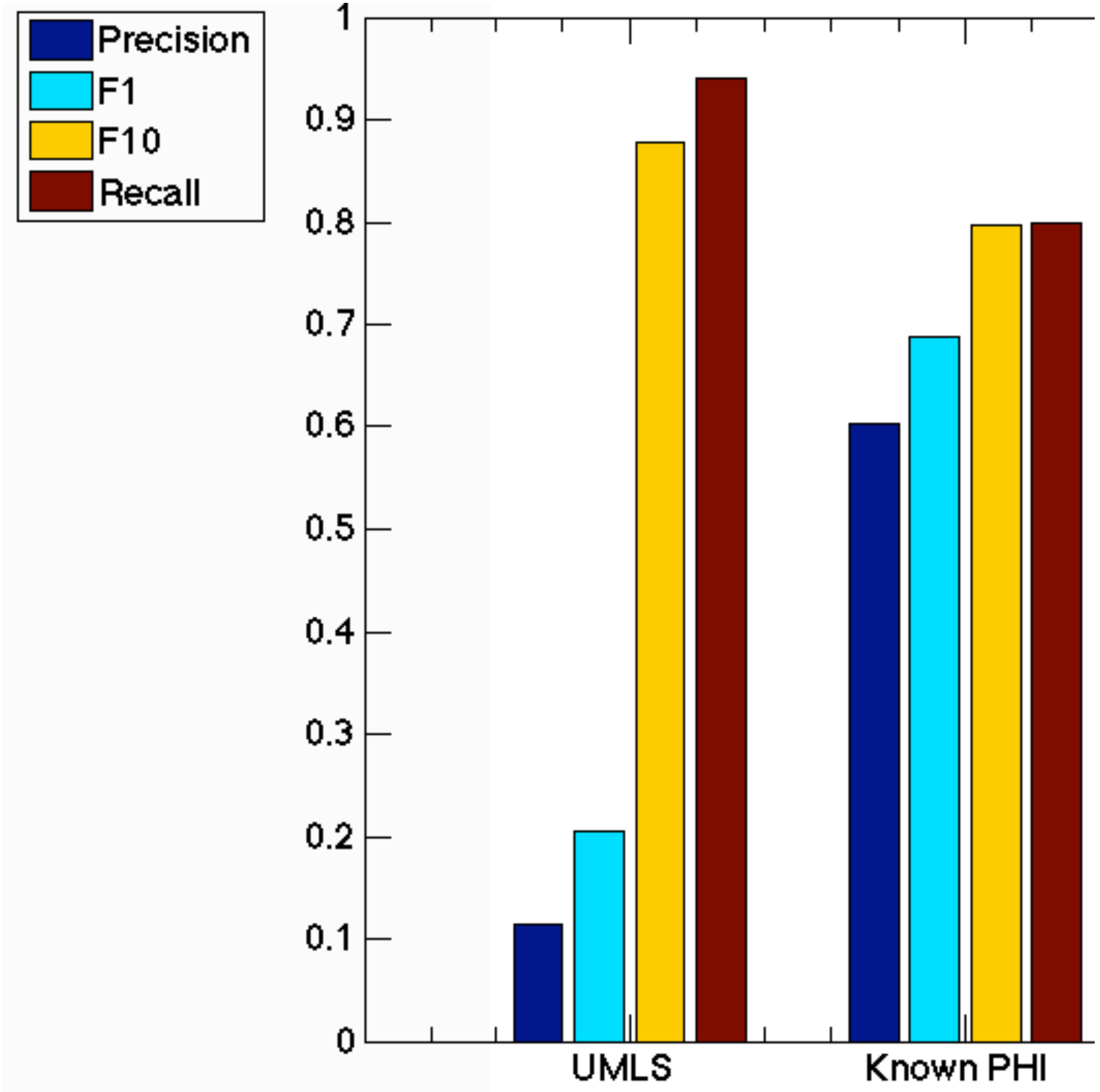
* Corresponding author: Bruce A Beckwith bruce_beckwith@bidmc.harvard.edu

Email: Bruce A Beckwith bruce_beckwith@bidmc.harvard.edu - Rajeshwarri Mahaadevan rajeshwarri@yahoo.com - Ulysses J Balis balis@helix.mgh.harvard.edu - Frank Kuo fkuo@partners.org

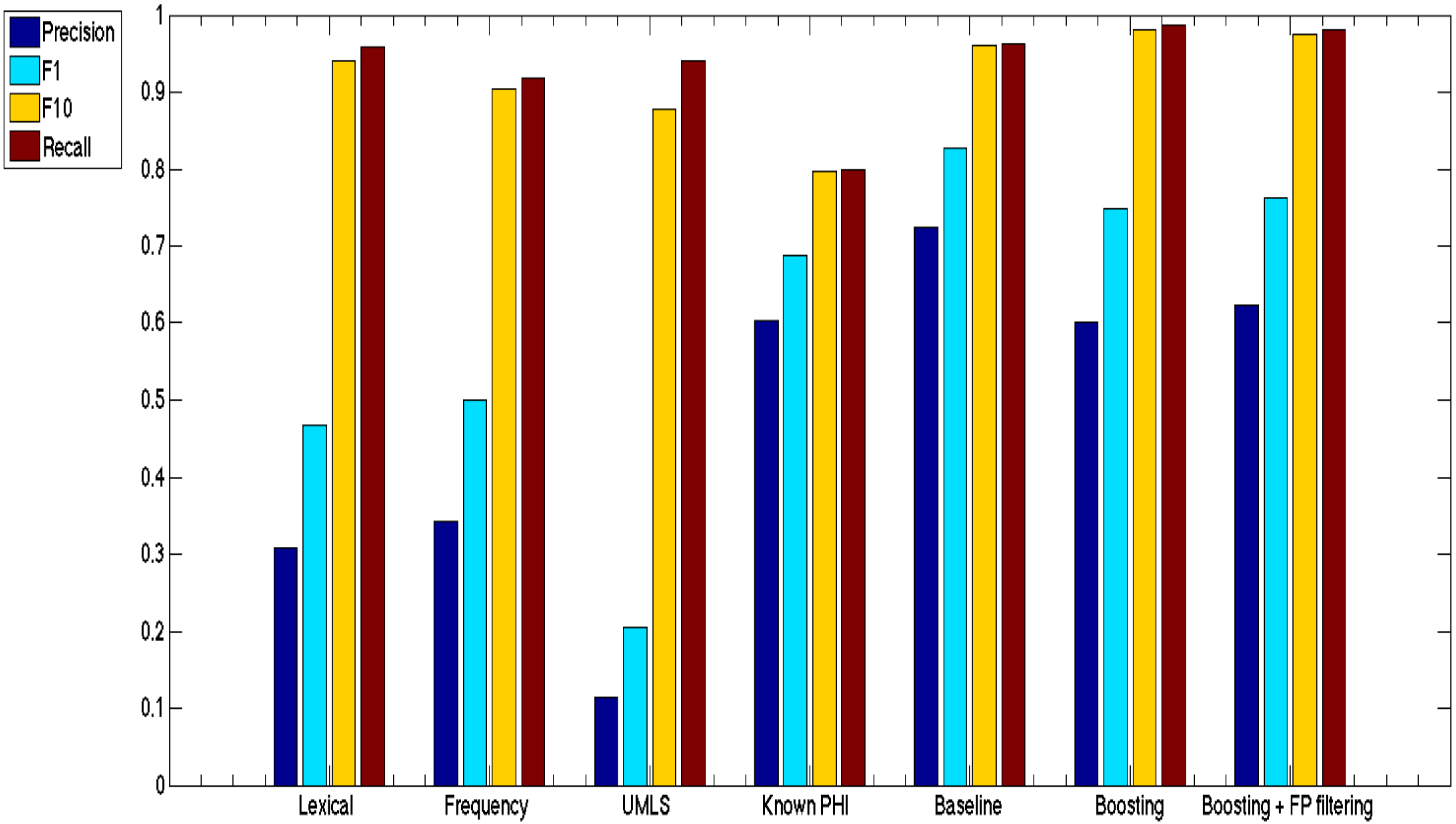
BMC Medical Informatics and Decision Making 2006, **6**:12 doi:10.1186/1472-6947-6-12

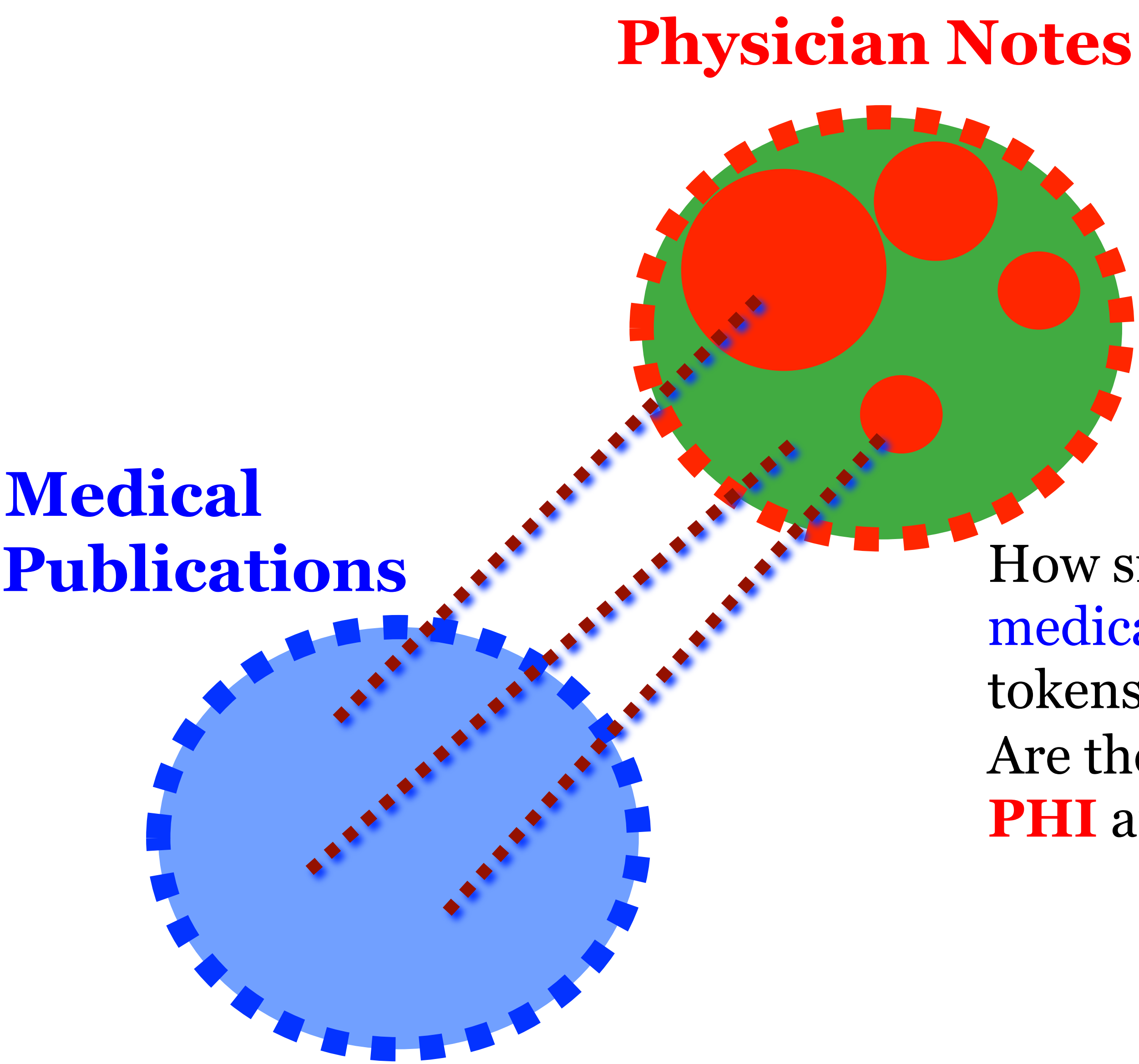
Published: 6 March 2006

Classifier Performance



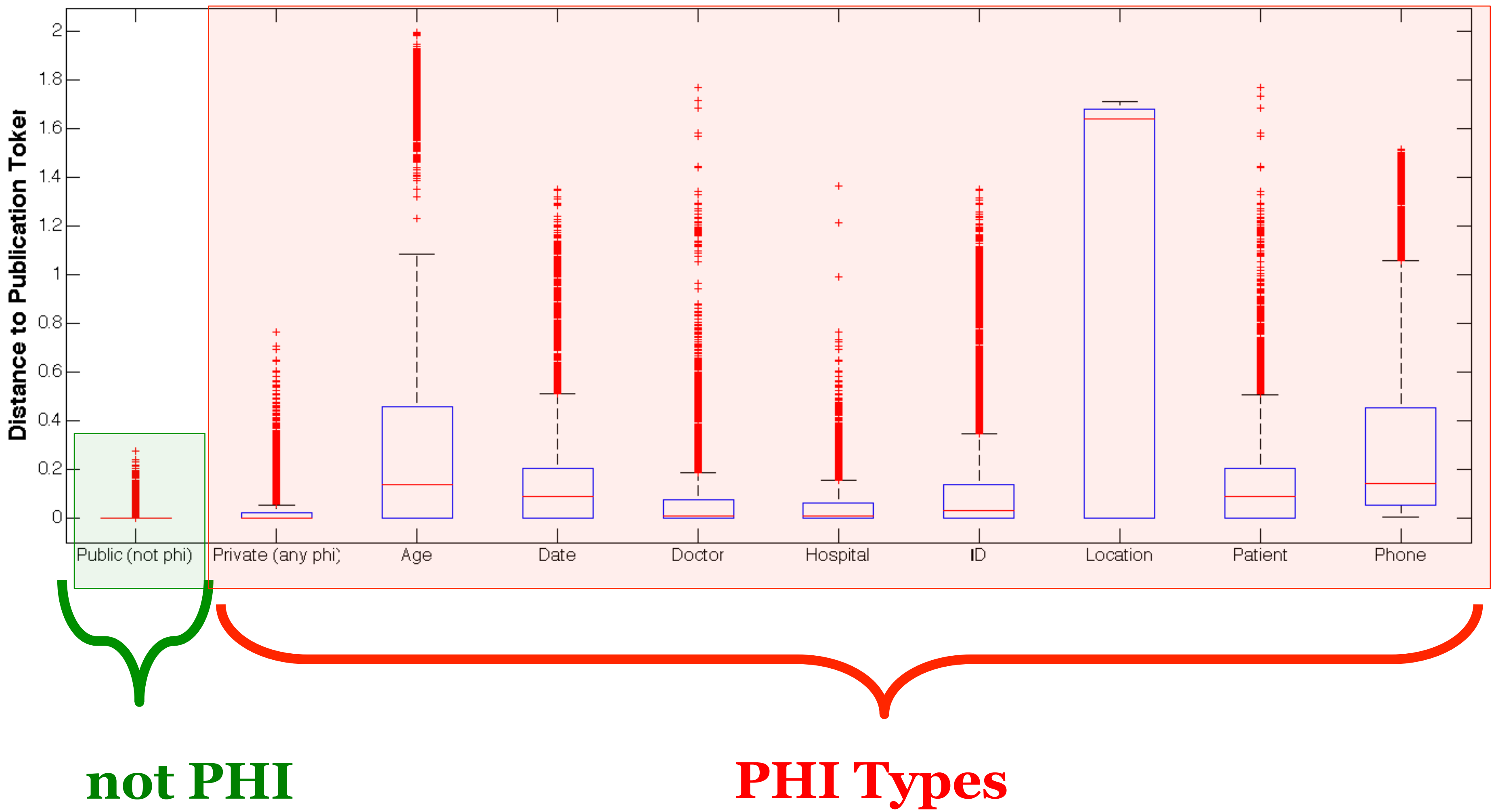
Classifier Performance



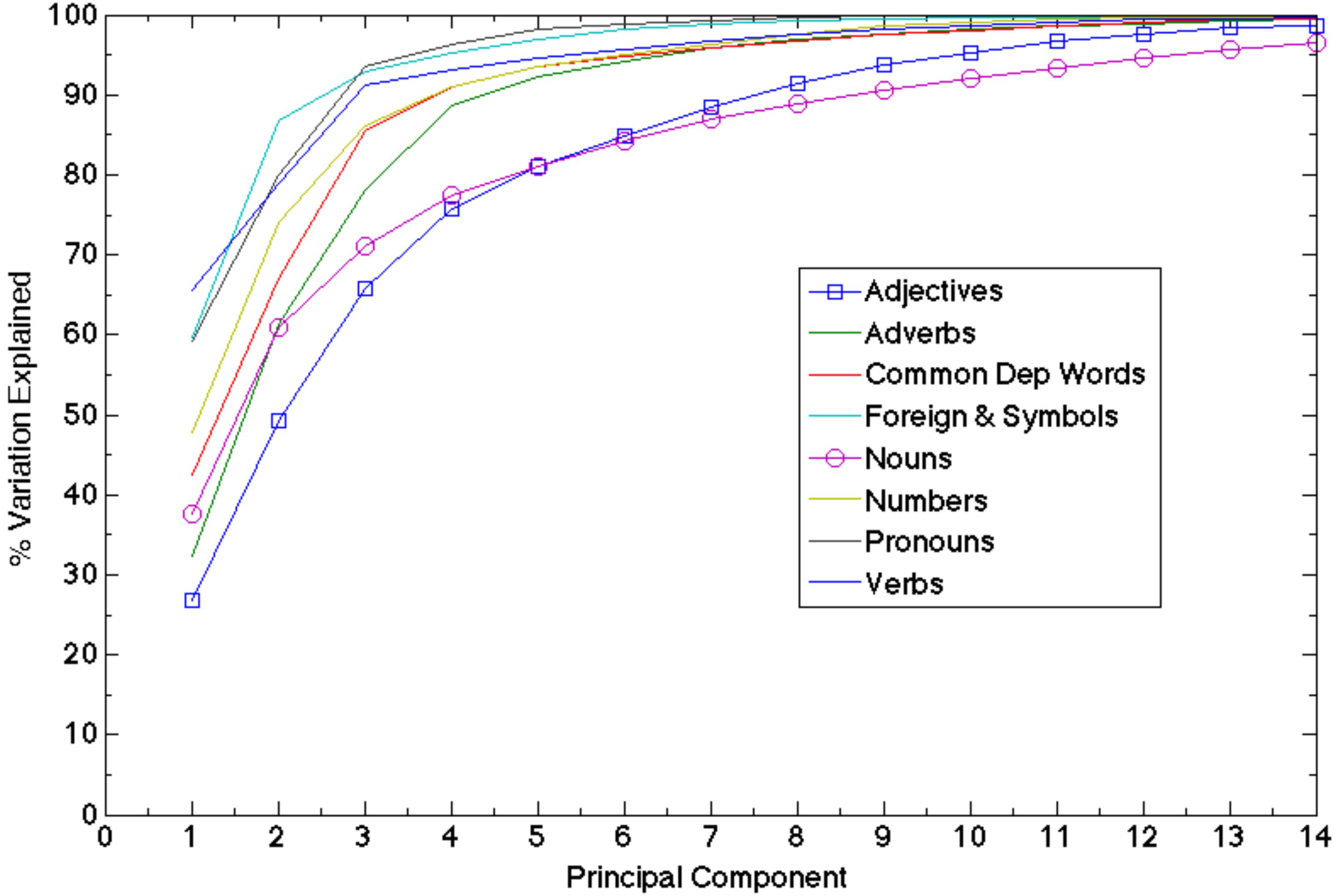


How similar are tokens in medical publications to tokens in physician notes?
Are the distances the same to **PHI** and **non-PHI** tokens?

Distance Measure for public and private text tokens



Variation for each Part Of Speech



False Positive Filtering with kNN

```
CM =

      119896           0           5          74          25           2           0           0           0
           0           0           0           0           0           0           0           0           0
           2           0        1570           0           0           0           0           0           0
          15           0           0         430           0           3           0           0           0
          31           0           0           1         212           0           0           0           0
           1           0           0           0           0         583           0           0           0
           9           0           0           0           0           0           0           0           0
          10           0           0           0           0           0           0           0           0
           3           0           0           0           0           1           0           0          72

% Vector Space Model (100NN)
=============================================

TP=2872 , TN=119896, FP=106, FN=71,
Sens    0.976 Specificity 0.999  Accuracy 0.999
Recall  0.976 Precision   0.964  F-Score  0.970
F1      0.970 F10         0.976  F100     0.976
```

83% of test set was mapped to 100NN
All sharing the same class label

Favor Recall (Boosted Decision Tree)

```
CM =

      131327           58         2847         1100          872          621          211          561           20
           0            3            0            0            0            0            0            0            0
          16            0         3654            0            0            1            0            0            0
          17            0            1         1846           44           77           26           86            0
          58            0           23           95         1381            2           13           61            0
           7            0           43            1            0         1401            1            2            0
          19            0            6           43           49            2           93            3            2
          14            0            0          119           11            0            7          251            0
           3            0            0            0            7            2            2            0          151

% Boosted Decision Tree (high recall)
=============================================

TP=9509 , TN=131327, FP=6290, FN=134,
Sens    0.986 Specificity 0.954  Accuracy 0.956
Recall 0.986 Precision    0.602  F-Score  0.748
F1      0.748 F10          0.980  F100     0.986

NA=0.046
Age=1.000
Date=0.996
DR=0.992
Hosp=0.964
ID=0.995
Loc=0.912
Pat=0.965
Phone=0.982
```

Boosting + False Positive Filtering

```
ConfusionMatrix order
        NA      Age     Date    DR      Hosp    ID      Loc     Pat     Phone

CM =

      131897          53        2775         950         688         575         170         497          12
           0           3           0           0           0           0           0           0           0
          22           0        3648           0           0           1           0           0           0
          35           0           1        1834          43          77          25          82           0
          81           0          23          93        1367           2          12          55           0
           7           0          43           1           0        1401           1           2           0
          21           0           8          43          48           2          90           3           2
          19           0           0         118           8           0           7         250           0
           3           0           0           0           7           2           2           0         151

% Classifier Summary
=============================================

TP=9455 , TN=131897, FP=5720, FN=188,
Sens    0.981 Specificity 0.958  Accuracy 0.960
Recall  0.981 Precision   0.623  F-Score  0.762
F1      0.762 F10         0.975  F100     0.980

NA=0.042
Age=1.000
Date=0.994
DR=0.983
Hosp=0.950
ID=0.995
Loc=0.903
Pat=0.953
Phone=0.982
```

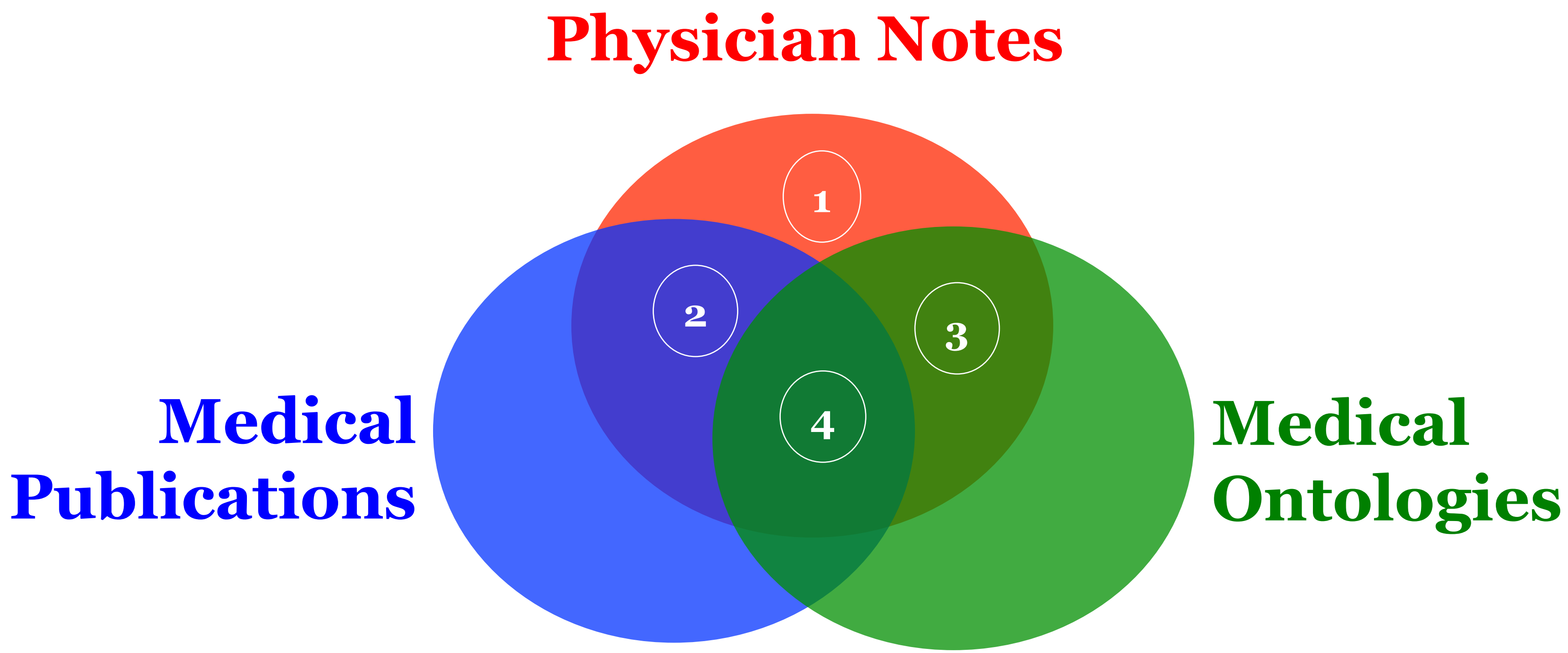
Summary

- Human annotations = $$, *rare*, hard to get
- Automated De-ID tends to overfit local training instances

- Learn ***background*** distribution from PUBLIC text
- Learn ***properties*** of PHI from fewer human annotations

- Apache cTAKES ***lexical*** annotations very informative for DeID
- Scrubber pipeline open source with example training set

- Classify new data: more like public text or private text ???



1. Nouns and Numbers that only occur in Physician Notes are probably PHI.
2. Words that occur frequently in medical publications are probably NOT PHI.
3. Words and phrases in in medical ontologies are probably not PHI.
4. Words shared in all three medical text sources are very unlikely to contain PHI.