TOWARDS A MINIMAL MODEL FOR NLP OUTPUTS

ALEXANDRE YAHI PHD STUDENT IN BIOMEDICAL INFORMATICS

OHDSI NLP WORKGROUP - FEBRUARY 3, 2016

COLUMBIA UNIVERSITY MEDICAL CENTER

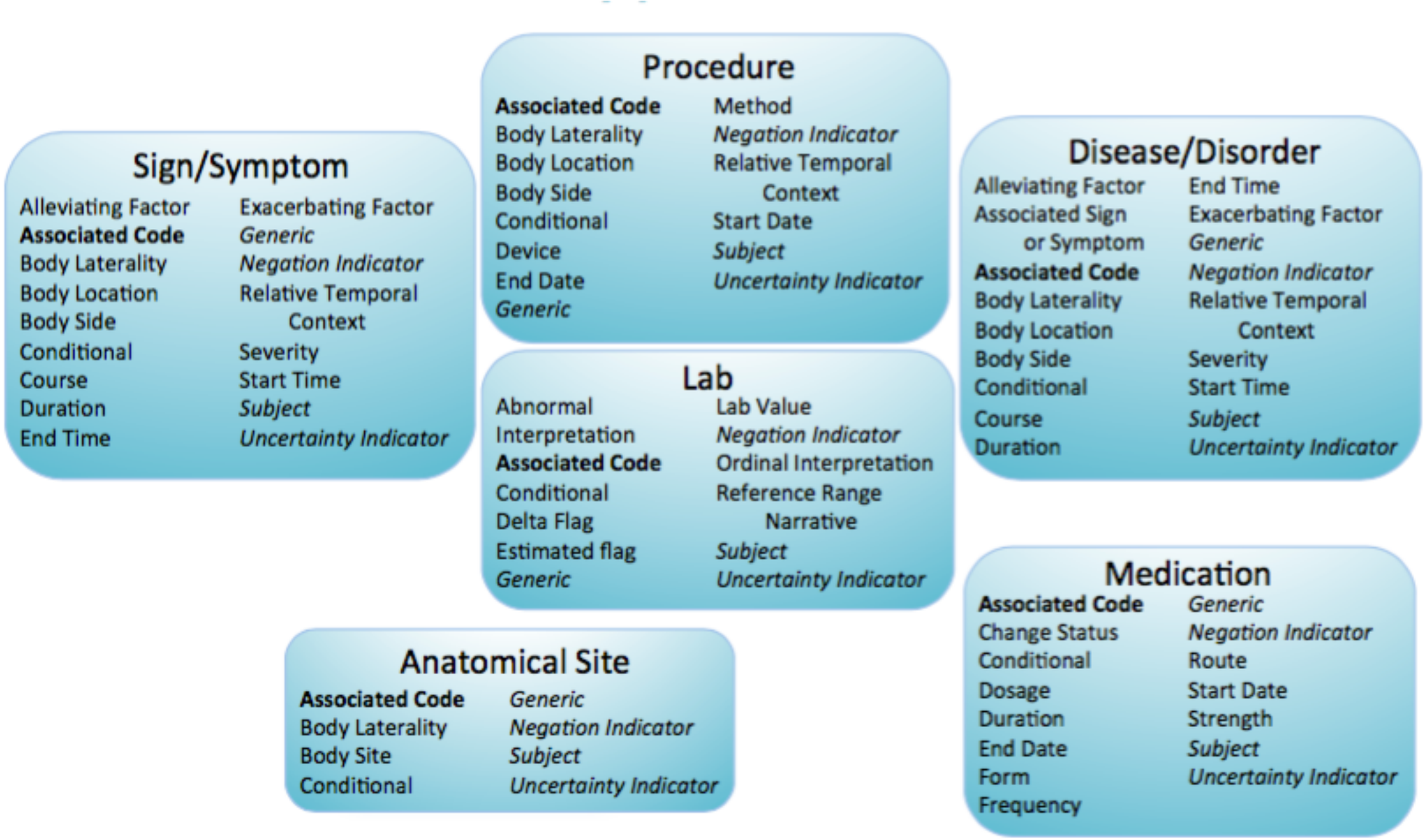
OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

CURRENT DATA STRUCTURE (SHARE): DISORDER ANNOTATIONS

- CUI (normalization)
  "presented with facial rash"
  Facial rash (CUI C0239521)
- Negation
  "patient denies numbness"
- Subject
  "son has schizophrenia"
- Uncertainty
  "evaluation of MI"
- Course
  "The cough got worse over the next two weeks."
- Severity
  "slight bleeding"
- Conditional
  "Pt should come back if any rash occurs"
- Generic
  "she went to the HIV clinic"
- Body Location
  "patient presented with facial rash"
  Face (CUI: C0015450)

Slide courtesy of Noémie Elhadad

OTHER SEMANTIC TYPES



Slide courtesy of G. Savova

CTAKES: MAYO CLINICAL TEXT ANALYSIS AND KNOWLEDGE EXTRACTION SYSTEM

- cTakes is an open source NLP tool relying on existing open-source technologies such as the Unstructured Information Management Architecture (UIMA) framework and OpenNLP toolkit
- Used the command line Collection Processing Engine (CPE), wrapped it in a simple Python script that creates .xml parameter file and .sh launch file.

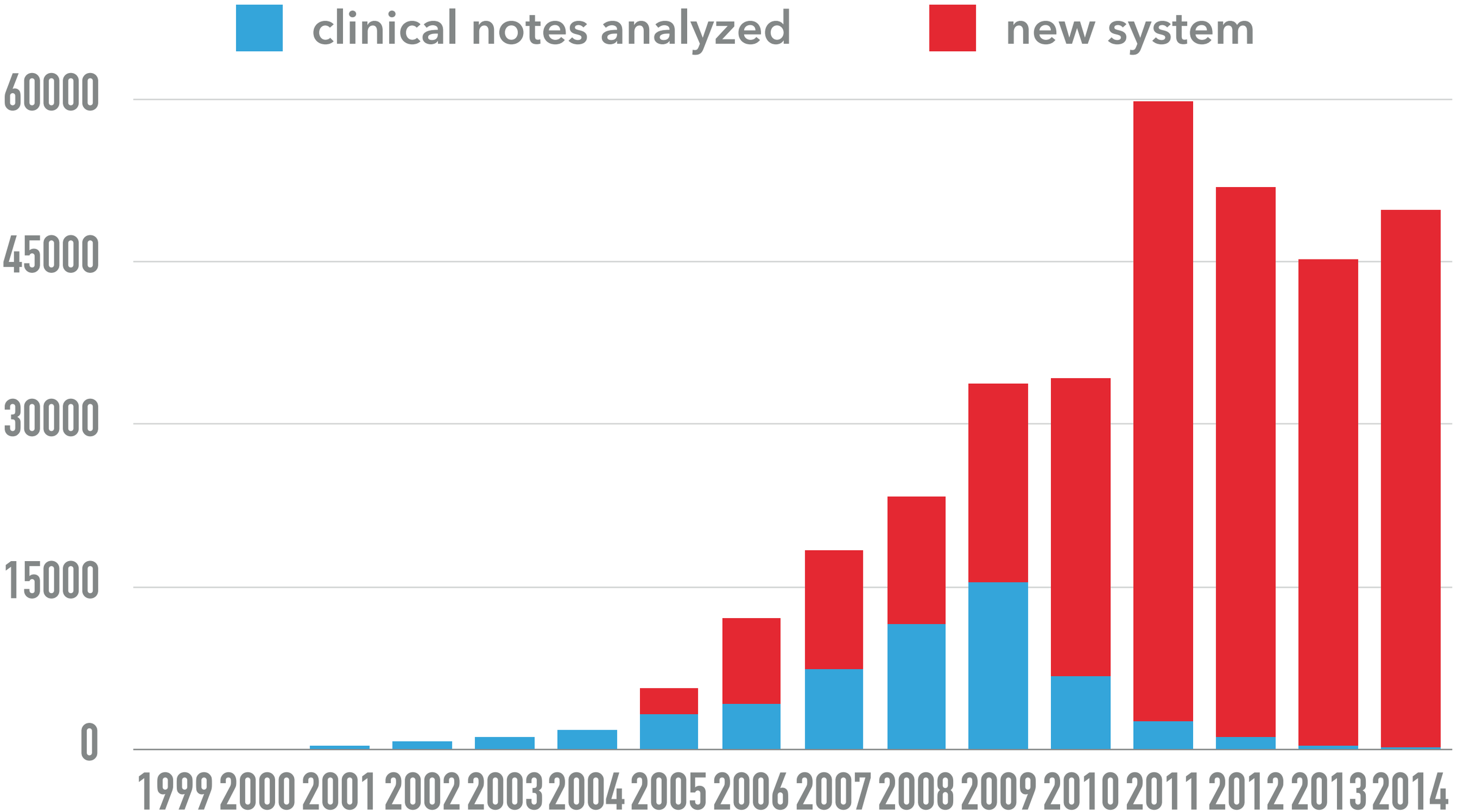
PROCESSED NOTES – FOCUS ON THE EMERGE COHORT

- Total number of processed notes: 58,785
- Total number of patients: 1,757
- Note types (top 10):

  101818 - Nephrology : 9660
  67759 - Clinical Note : 6883
  67761 - Progress Note : 5868
  67760 - Admission Note : 5529
  81271 - Telephone Call Documentation : 3506
  67769 - Follow up : 2520
  67765 - Consult Note : 2084
  75854 - Hospitalist Attending Initial Visit : 1781
  67764 - Discharge Note : 1495
  75852 - Resident Initial Visit : 1384

- About 280k notes not analysed yet for this cohort in the new system (Eclipsys): it's a combination of semi-structured data and free-text that needs pre-processing before using any NLP tool

NOTE COUNTS BY YEAR



SEMANTIC TYPES AND SEMANTIC GROUPS

- Only parsed 2 semantic types: Diseases and Symptoms
- Top-5 Semantic groups: Disorders

T033 – Finding : 990286

T047 - Disease or Syndrome : 758480

T184 - Sign or Symptom : 419907

T046 - Pathologic Function : 259461

T191 - Neoplastic Process : 97139

MODIFIERS TUPLES

- Most common is what we want:
- 'confidence','polarity','uncertainty','conditional','generic','subject','historyOf'

```
0.0|1|0|false|false|patient|0,2393283
0.0|-1|0|false|false|patient|0,390837
```

DISCUSSIONS

- cTakes is not user friendly - many issues due to a lack of legible documentation
- Some doubts about the negation engine
- Need to identify notes' sections - not sure cTakes uses the right modifier for term in the history section for example
- New notes structure (Eclipsys) might need a combination of regular expression parsing and NLP for the free-text chunks