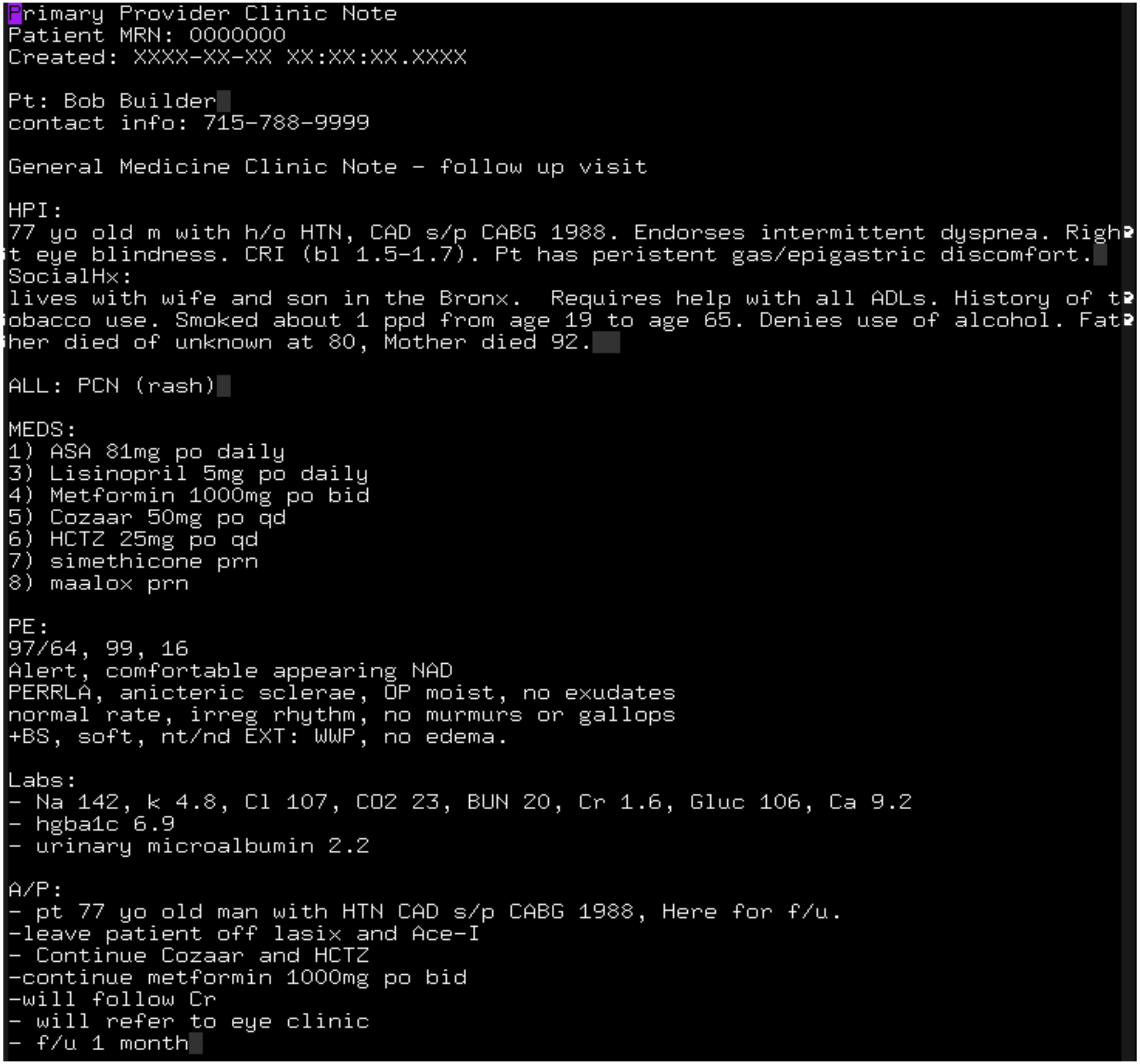
Clinical NLP schemas

Noémie Elhadad
noemie@gmail.com
Columbia University

Outline

- Types of NLP outputs
  - Unstructured text → structured output
  - Unstructured text → bag of words
  - Unstructured text → word embeddings
- The ShARe schema for structured output
- Some low level details

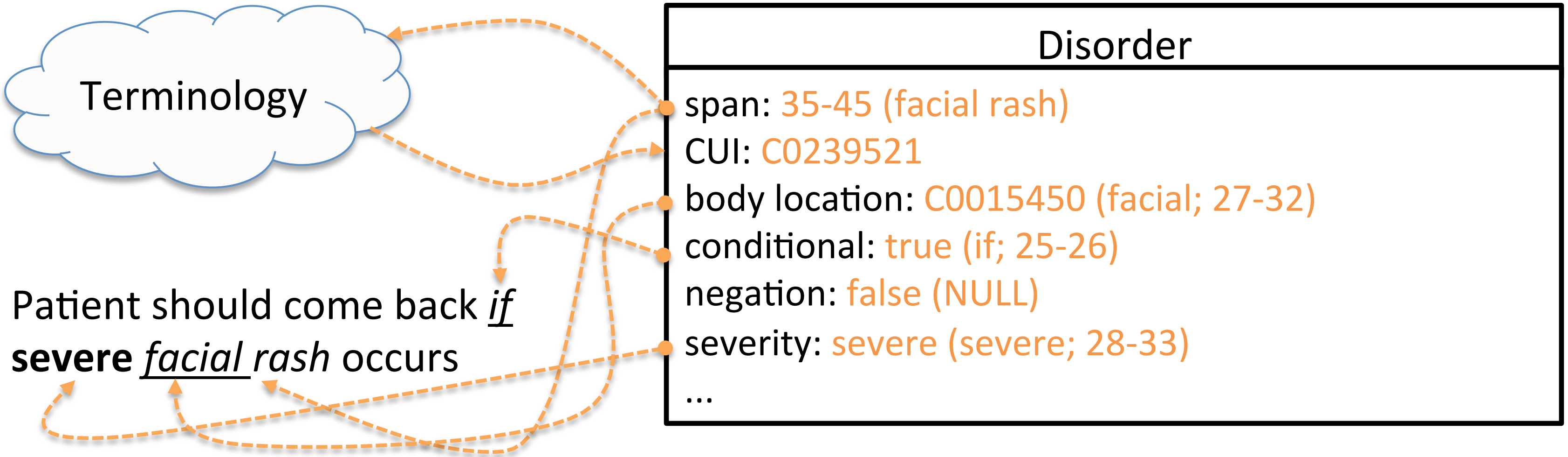
Unstructured text



```
Primary Provider Clinic Note
Patient MRN: 0000000
Created: XXXX-XX-XX XX:XX:XX.XXXX

Pt: Bob Builder
contact info: 715-788-9999

General Medicine Clinic Note - follow up visit

HPI:
77 yo old m with h/o HTN, CAD s/p CABG 1988. Endorses intermittent dyspnea. Right
 eye blindness. CRI (bl 1.5-1.7). Pt has peristent gas/epigastric discomfort.
SocialHx:
lives with wife and son in the Bronx.  Requires help with all ADLs. History of tobacco
 use. Smoked about 1 ppd from age 19 to age 65. Denies use of alcohol. Father
 died of unknown at 80, Mother died 92.

ALL: PCN (rash)

MEDS:
1) ASA 81mg po daily
3) Lisinopril 5mg po daily
4) Metformin 1000mg po bid
5) Cozaar 50mg po qd
6) HCTZ 25mg po qd
7) simethicone prn
8) maalox prn

PE:
97/64, 99, 16
Alert, comfortable appearing NAD
PERRLA, anicteric sclerae, OP moist, no exudates
normal rate, irreg rhythm, no murmurs or gallops
+BS, soft, nt/nd EXT: WWP, no edema.

Labs:
- Na 142, k 4.8, Cl 107, CO2 23, BUN 20, Cr 1.6, Gluc 106, Ca 9.2
- hgba1c 6.9
- urinary microalbumin 2.2

A/P:
- pt 77 yo old man with HTN CAD s/p CABG 1988, Here for f/u.
-leave patient off lasix and Ace-I
- Continue Cozaar and HCTZ
-continue metformin 1000mg po bid
-will follow Cr
- will refer to eye clinic
- f/u 1 month
```

Structured output

- Clinical NLP pipeline output



Terminology

Patient should come back *if*
severe *facial rash* occurs

Disorder

span: 35-45 (facial rash)
CUI: C0239521
body location: C0015450 (facial; 27-32)
conditional: true (if; 25-26)
negation: false (NULL)
severity: severe (severe; 28-33)
...

Structured output

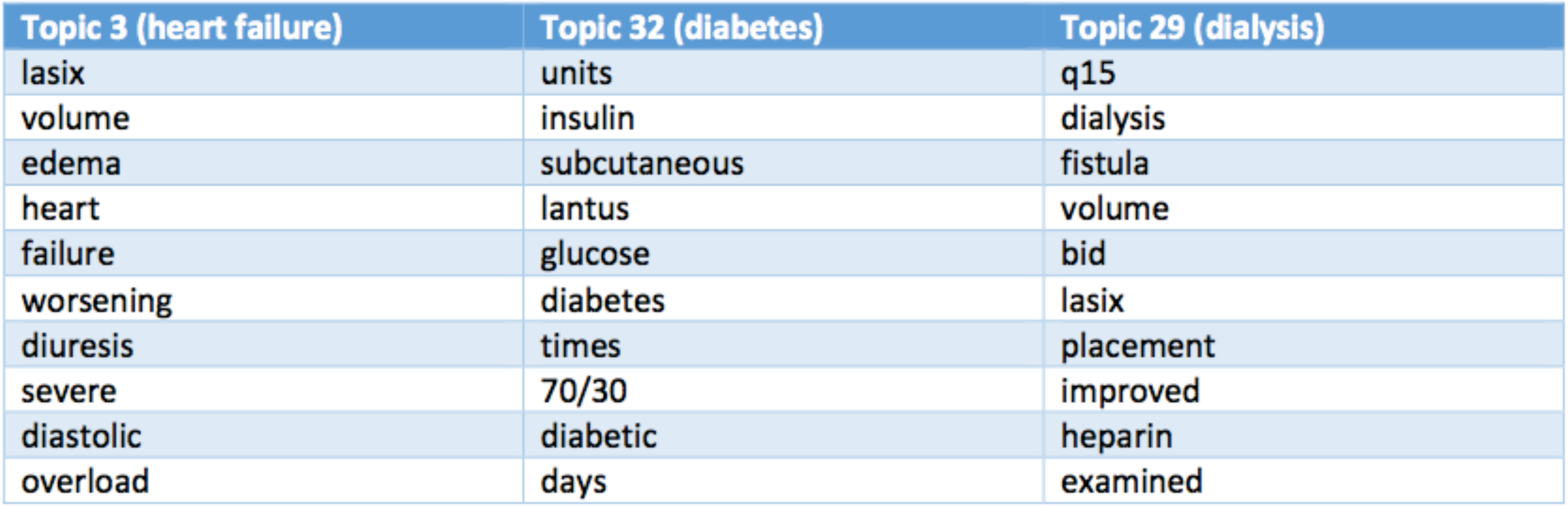
- Useful for
  - phenotyping
  - cohort identification
  - information extraction
  - …

Bag of words (observations)

- Vocabulary of all the words in the given notes of an institution
  - Filter out infrequent words, stop words, identifiers
  - Words are not filtered according to a terminology
- Each note is represented as a bag of words
  - Note n = w_{43}:3, w_{118}:9, w_{210}:2, w_{534}:10, …
  - Lose the sequence of the words
  - Less semantic interpretation, more raw individual observations

Bag of words (observations)

- Let the burden of identifying features to further processes

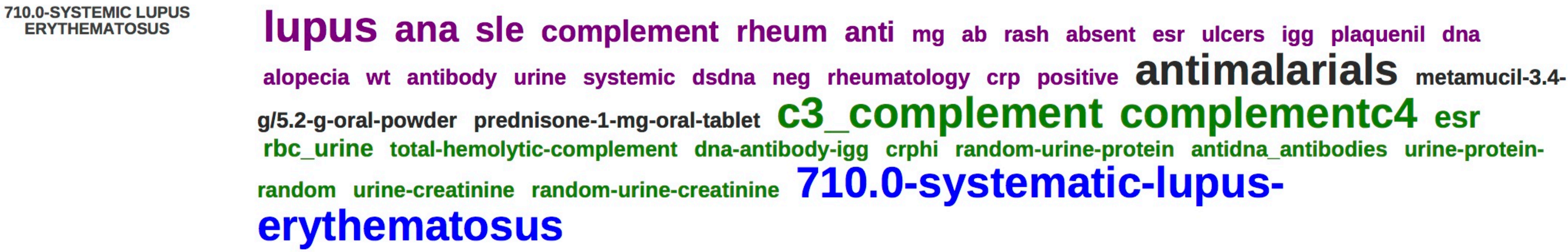


Topic 3 (heart failure)	Topic 32 (diabetes)	Topic 29 (dialysis)
lasix	units	q15
volume	insulin	dialysis
edema	subcutaneous	fistula
heart	lantus	volume
failure	glucose	bid
worsening	diabetes	lasix
diuresis	times	placement
severe	70/30	improved
diastolic	diabetic	heparin
overload	days	examined

- Perotte et al (2015) Risk Prediction for Chronic Kidney Disease Progression Using Heterogeneous Electronic Health Record Data and Time Series Analysis. J Am Med Inform Assoc.
- Pivovarov et al (2015) Learning Probabilistic Phenotypes from Heterogeneous EHR Data. J Biomed Inform. In Press.

Bag of observations (words)

- Let the burden of identifying features to further processes


710.0-SYSTEMIC LUPUS ERYTHEMATOSUS
lupus ana sle complement rheum anti mg ab rash absent esr ulcers igg plaquenil dna
alopecia wt antibody urine systemic dsdna neg rheumatology crp positive antimalarials metamucil-3.4-g/5.2-g-oral-powder prednisone-1-mg-oral-tablet
c3_complement complementc4 esr rbc_urine total-hemolytic-complement dna-antibody-igg crphi random-urine-protein antidna_antibodies urine-protein-random urine-creatinine random-urine-creatinine
710.0-systematic-lupus-erythematosus

- Perotte et al (2015) Risk Prediction for Chronic Kidney Disease Progression Using Heterogeneous Electronic Health Record Data and Time Series Analysis. J Am Med Inform Assoc.
- Pivovarov et al (2015) Learning Probabilistic Phenotypes from Heterogeneous EHR Data. J Biomed Inform. In Press.

Word embeddings

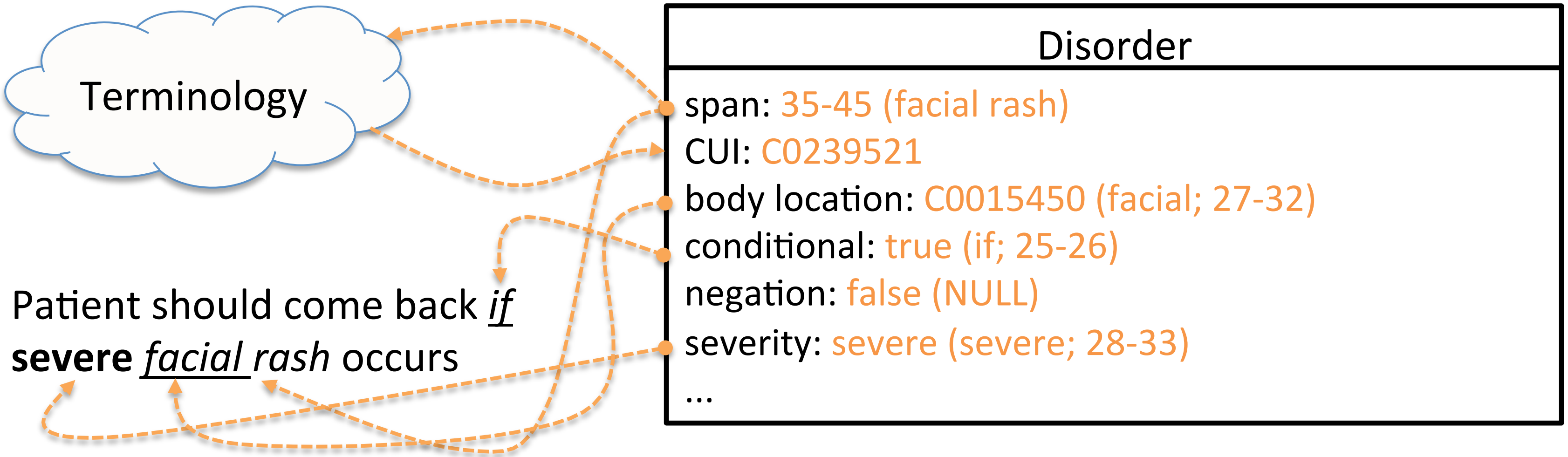
- Observations are still words, but now a word is represented as a vector
  - Dimension reduction
  - Distributional semantics, word embeddings
- For each word, the representation is learned optimized for a particular task
  - Optimize for language model
  - Optimize for phenotype recognition
  - …
- Often need to keep some sequential information

Outline

- Types of NLP outputs
  - Unstructured text → structured output
  - Unstructured text → bag of words
  - Unstructured text → word embeddings
- The ShARe schema for structured output
- Some low level details

Structured output schema

- Clinical Element Model templates
- Data modeling across several initiatives and institutions (ShARe, SHARP, THYME)



Terminology

Patient should come back *if* **severe** *facial rash* occurs

Disorder

span: 35-45 (facial rash)
CUI: C0239521
body location: C0015450 (facial; 27-32)
conditional: true (if; 25-26)
negation: false (NULL)
severity: severe (severe; 28-33)
...

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)

Other semantic types

Sign/Symptom

- Alleviating Factor
- **Associated Code**
- Body Laterality
- Body Location
- Body Side
- Conditional
- Course
- Duration
- End Time
- Exacerbating Factor
- *Generic*
- *Negation Indicator*
- Relative Temporal Context
- Severity
- Start Time
- *Subject*
- *Uncertainty Indicator*

Procedure

- **Associated Code**
- Body Laterality
- Body Location
- Body Side
- Conditional
- Device
- End Date
- *Generic*
- Method
- *Negation Indicator*
- Relative Temporal Context
- Start Date
- *Subject*
- *Uncertainty Indicator*

Lab

- Abnormal Interpretation
- **Associated Code**
- Conditional
- Delta Flag
- Estimated flag
- *Generic*
- Lab Value
- *Negation Indicator*
- Ordinal Interpretation
- Reference Range Narrative
- *Subject*
- *Uncertainty Indicator*

Anatomical Site

- **Associated Code**
- Body Laterality
- Body Site
- Conditional
- *Generic*
- *Negation Indicator*
- *Subject*
- *Uncertainty Indicator*

Disease/Disorder

- Alleviating Factor
- Associated Sign or Symptom
- **Associated Code**
- Body Laterality
- Body Location
- Body Side
- Conditional
- Course
- Duration
- End Time
- Exacerbating Factor
- *Generic*
- *Negation Indicator*
- Relative Temporal Context
- Severity
- Start Time
- *Subject*
- *Uncertainty Indicator*

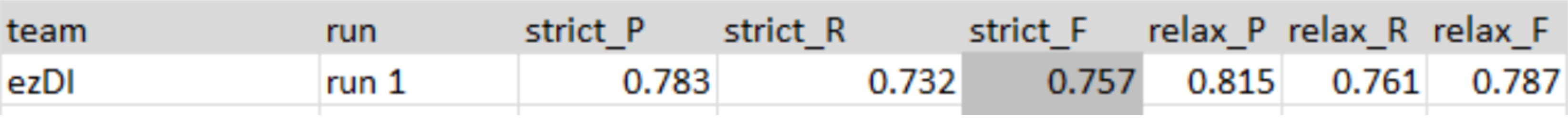
Medication

- **Associated Code**
- Change Status
- Conditional
- Dosage
- Duration
- End Date
- Form
- Frequency
- *Generic*
- *Negation Indicator*
- Route
- Start Date
- Strength
- *Subject*
- *Uncertainty Indicator*

Slide courtesy of G. Savova

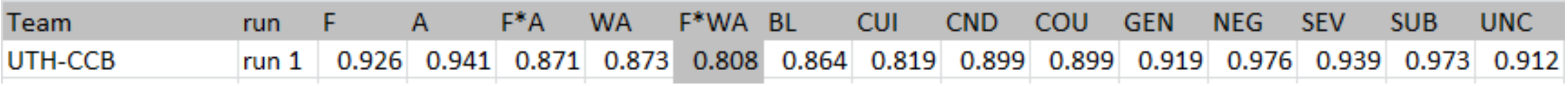
ShARe community task (disorders only)

- Task 1 – 16 teams (concept recognition and normalization)



team	run	strict_P	strict_R	strict_F	relax_P	relax_R	relax_F
ezDI	run 1	0.783	0.732	0.757	0.815	0.761	0.787

- Task 2b – 9 teams (concept + attributes normalization)



Team	run	F	A	F*A	WA	F*WA	BL	CUI	CND	COU	GEN	NEG	SEV	SUB	UNC
UTH-CCB	run 1	0.926	0.941	0.871	0.873	0.808	0.864	0.819	0.899	0.899	0.919	0.976	0.939	0.973	0.912

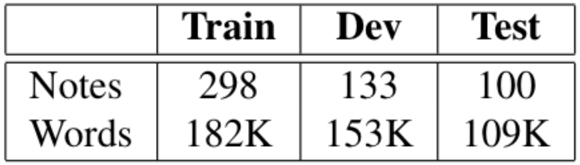
- Elhadad et al (2015) SemEval-2015 Task 14: Analysis of Clinical Text. Proc. SemEval'15.
- Pradhan et al (2015) Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. J Am Med Inform Assoc.

Outline

- Types of NLP outputs
  - Unstructured text → structured output
  - Unstructured text → bag of words
  - Unstructured text → word embeddings
- The ShARe schema for structured output
- Some low level details

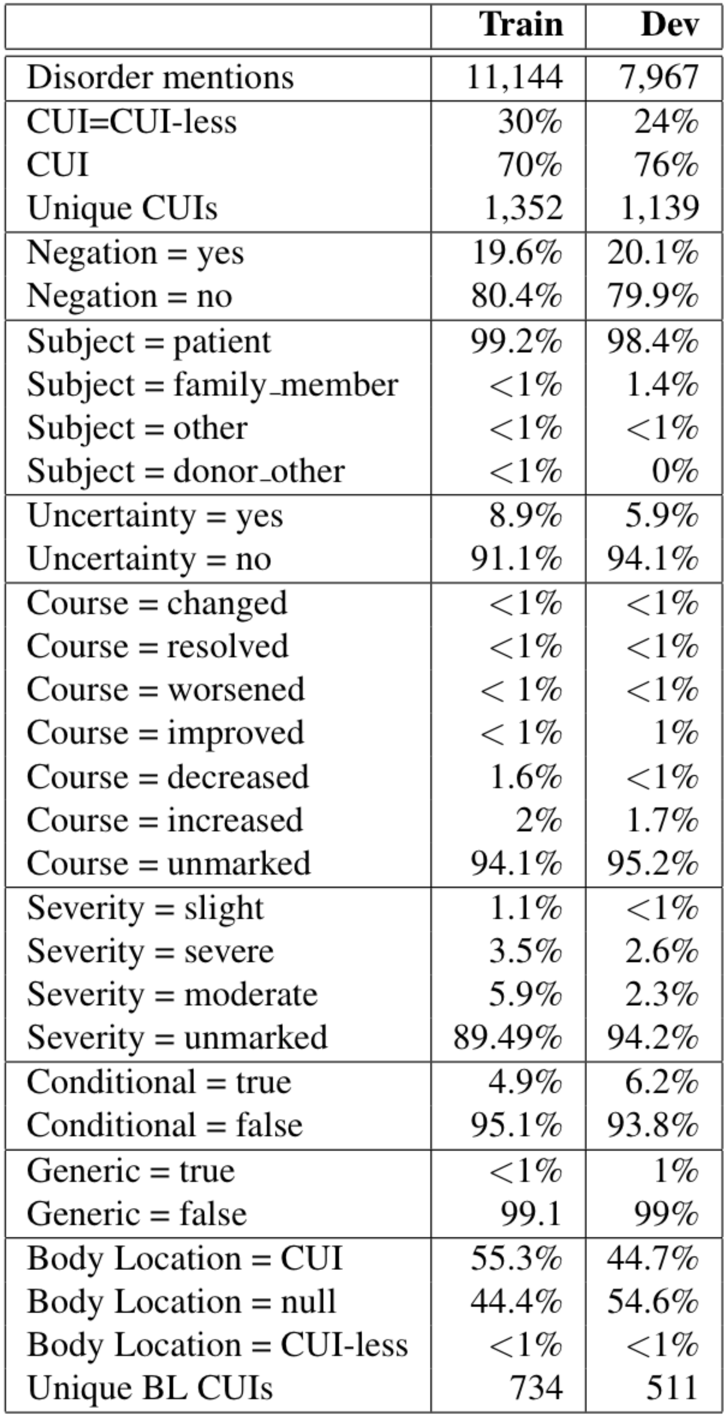
ShARe dataset

- Annotated clinical notes



	Train	Dev	Test
Notes	298	133	100
Words	182K	153K	109K

- Un-annotated clinical notes
  - 400,000+ notes
  - 122 M words



	Train	Dev
Disorder mentions	11,144	7,967
CUI=CUI-less	30%	24%
CUI	70%	76%
Unique CUIs	1,352	1,139
Negation = yes	19.6%	20.1%
Negation = no	80.4%	79.9%
Subject = patient	99.2%	98.4%
Subject = family_member	<1%	1.4%
Subject = other	<1%	<1%
Subject = donor_other	<1%	0%
Uncertainty = yes	8.9%	5.9%
Uncertainty = no	91.1%	94.1%
Course = changed	<1%	<1%
Course = resolved	<1%	<1%
Course = worsened	< 1%	<1%
Course = improved	< 1%	1%
Course = decreased	1.6%	<1%
Course = increased	2%	1.7%
Course = unmarked	94.1%	95.2%
Severity = slight	1.1%	<1%
Severity = severe	3.5%	2.6%
Severity = moderate	5.9%	2.3%
Severity = unmarked	89.49%	94.2%
Conditional = true	4.9%	6.2%
Conditional = false	95.1%	93.8%
Generic = true	<1%	1%
Generic = false	99.1	99%
Body Location = CUI	55.3%	44.7%
Body Location = null	44.4%	54.6%
Body Location = CUI-less	<1%	<1%
Unique BL CUIs	734	511

Many concepts and attributes can be…

- Pipe delimited

```
report name|disorder-span|cui|Norm_NI|Cue_NI|Norm_SC|Cue_SC|Norm_UI|Cue_UI|
Norm_CC|Cue_CC|Norm_SV|Cue_SV|Norm_CO|Cue_CO|Norm_GC|Cue_GC|Norm_BL|Cue_BL|
Norm_DT|Norm_TE|Cue_TE
```

```
09388-093839-DISCHARGE_SUMMARY.txt|30-36|C0040128|*no|*NULL|*patient|*NULL|*no| *NULL|*false|
*NULL| *unmarked|*NULL|severe|*NULL|*false|*NULL|C0040132|*NULL| Before|*None|*NULL
```

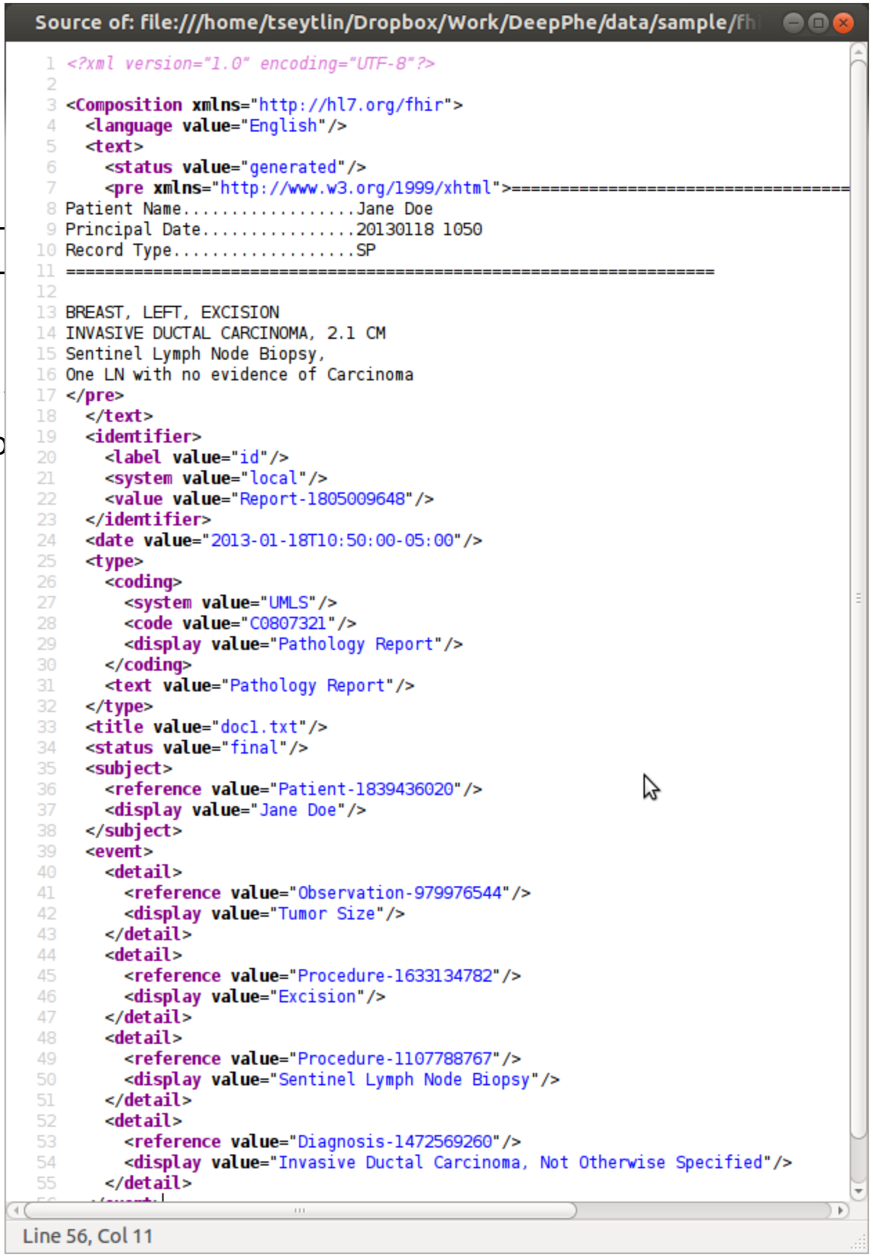
Many concepts and attributes can be…

- Pipe delimited

```
report name|disorder-span|cui|Norm_NI|Cue_
Norm_CC|Cue_CC|Norm_SV|Cue_SV|Norm_CO|Cue_
Norm_DT|Norm_TE|Cue_TE
```

09388-093839-DISCHARGE_SUMMARY.txt|30-36|C0040128|
*NULL| *unmarked|*NULL|severe|*NULL|*false|*NULL|Coo

- Composed in FHIR®©



Source of: file:///home/tseytlin/Dropbox/Work/DeepPhe/data/sample/fh

```
<?xml version="1.0" encoding="UTF-8"?>

<Composition xmlns="http://hl7.org/fhir">
  <language value="English"/>
  <text>
    <status value="generated"/>
    <pre xmlns="http://www.w3.org/1999/xhtml">=====================================
Patient Name..................Jane Doe
Principal Date................20130118 1050
Record Type...................SP
=================================================================

BREAST, LEFT, EXCISION
INVASIVE DUCTAL CARCINOMA, 2.1 CM
Sentinel Lymph Node Biopsy,
One LN with no evidence of Carcinoma
</pre>
  </text>
  <identifier>
    <label value="id"/>
    <system value="local"/>
    <value value="Report-1805009648"/>
  </identifier>
  <date value="2013-01-18T10:50:00-05:00"/>
  <type>
    <coding>
      <system value="UMLS"/>
      <code value="C0807321"/>
      <display value="Pathology Report"/>
    </coding>
    <text value="Pathology Report"/>
  </type>
  <title value="doc1.txt"/>
  <status value="final"/>
  <subject>
    <reference value="Patient-1839436020"/>
    <display value="Jane Doe"/>
  </subject>
  <event>
    <detail>
      <reference value="Observation-979976544"/>
      <display value="Tumor Size"/>
    </detail>
    <detail>
      <reference value="Procedure-1633134782"/>
      <display value="Excision"/>
    </detail>
    <detail>
      <reference value="Procedure-1107788767"/>
      <display value="Sentinel Lymph Node Biopsy"/>
    </detail>
    <detail>
      <reference value="Diagnosis-1472569260"/>
      <display value="Invasive Ductal Carcinoma, Not Otherwise Specified"/>
    </detail>
```

Line 56, Col 11

Screenshot courtesy of G. Savova

Many concepts and attributes can be…

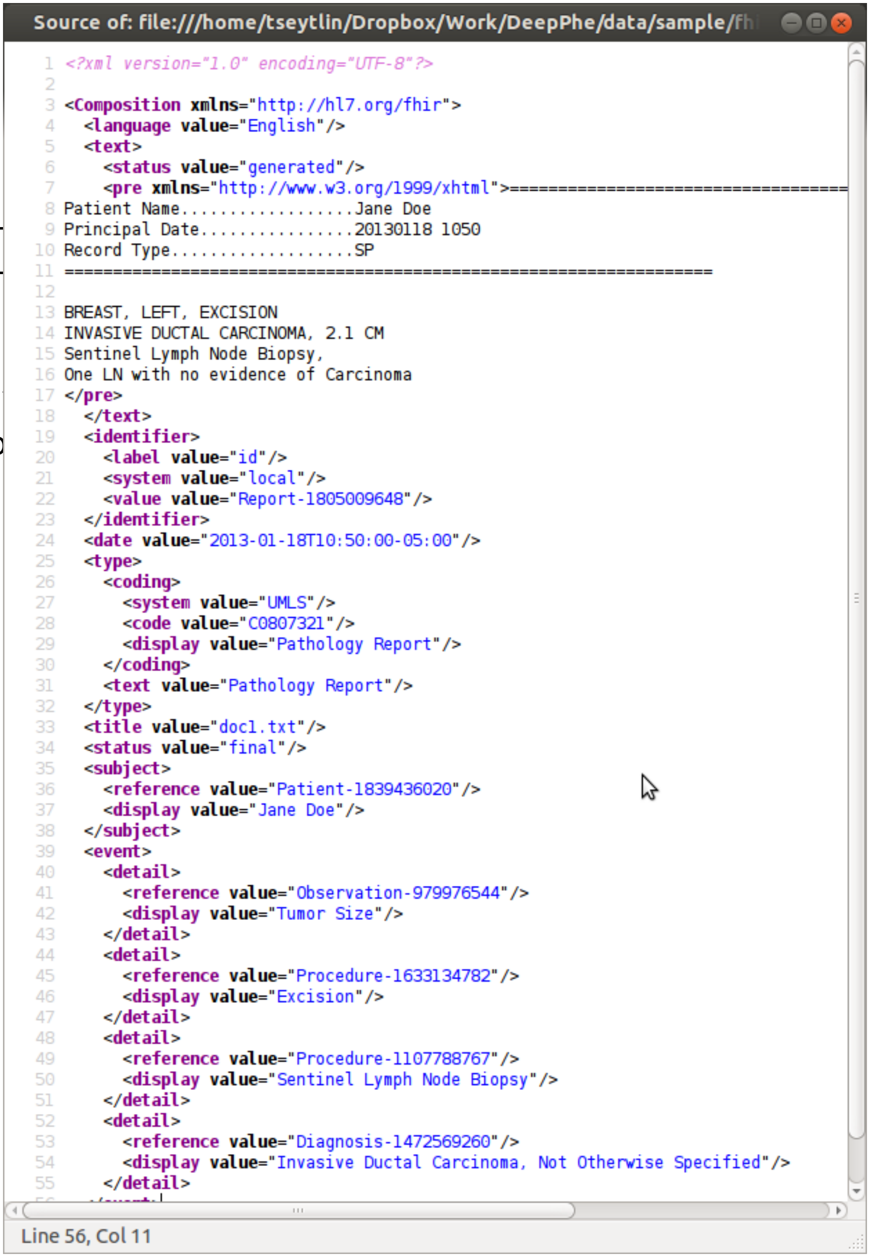
- Pipe delimited

```
report name|disorder-span|cui|Norm_NI|Cue_
Norm_CC|Cue_CC|Norm_SV|Cue_SV|Norm_CO|Cue_
Norm_DT|Norm_TE|Cue_TE
```

09388-093839-DISCHARGE_SUMMARY.txt|30-36|C0040128|
*NULL| *unmarked|*NULL|severe|*NULL|*false|*NULL|Coo

- Composed in FHIR®©

- Lucene indexes


Source of: file:///home/tseytlin/Dropbox/Work/DeepPhe/data/sample/fh
Line 56, Col 11

```xml
<?xml version="1.0" encoding="UTF-8"?>

<Composition xmlns="http://hl7.org/fhir">
  <language value="English"/>
  <text>
    <status value="generated"/>
    <pre xmlns="http://www.w3.org/1999/xhtml">=====================================
Patient Name..................Jane Doe
Principal Date................20130118 1050
Record Type...................SP
===================================================================

BREAST, LEFT, EXCISION
INVASIVE DUCTAL CARCINOMA, 2.1 CM
Sentinel Lymph Node Biopsy,
One LN with no evidence of Carcinoma
</pre>
  </text>
  <identifier>
    <label value="id"/>
    <system value="local"/>
    <value value="Report-1805009648"/>
  </identifier>
  <date value="2013-01-18T10:50:00-05:00"/>
  <type>
    <coding>
      <system value="UMLS"/>
      <code value="C0807321"/>
      <display value="Pathology Report"/>
    </coding>
    <text value="Pathology Report"/>
  </type>
  <title value="doc1.txt"/>
  <status value="final"/>
  <subject>
    <reference value="Patient-1839436020"/>
    <display value="Jane Doe"/>
  </subject>
  <event>
    <detail>
      <reference value="Observation-979976544"/>
      <display value="Tumor Size"/>
    </detail>
    <detail>
      <reference value="Procedure-1633134782"/>
      <display value="Excision"/>
    </detail>
    <detail>
      <reference value="Procedure-1107788767"/>
      <display value="Sentinel Lymph Node Biopsy"/>
    </detail>
    <detail>
      <reference value="Diagnosis-1472569260"/>
      <display value="Invasive Ductal Carcinoma, Not Otherwise Specified"/>
    </detail>
  </event>
```

Screenshot courtesy of G. Savova

Points for discussion

- NLP for what tasks and requirements on NLP output
- Tables and schema as minimum viable products given NLP technology
  - Note table vs/and NLP output table
- How to store many observations and their attributes