MACHINE LEARNING FOR HEALTHCARE
6.S897, HST.S53

Massachusetts Institute of Technology

Lecture 9: Clinical text & natural language processing

Prof. David Sontag

MIT EECS, CSAIL, IMES

Acknowledgements: Slides adapted from Nigam Shah (Stanford School of Medicine) and Noemie Elhadad (Columbia University)

Anatomy of a clinical note

(Slide credit: Nigam Shah)

Free-text and semi-structured texts

(Slide credit: Noemie Elhadad)

```
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. Righ
t 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 t
obacco use. Smoked about 1 ppd from age 19 to age 65. Denies use of alcohol. Fat
her 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
```

CT scan of liver

single hypervascular liver lesion within the medial segment ofthe left lobe of the liver abutting and compressing the middlehepatic vein as described above, concerning for hepatocellularcarcinoma. two nonspecific pulmonary nodules as described above.

(Slide credit: Nigam Shah)

MRI of wrist

subtle non-displaced transverse fracture through the proximalpole of the scaphoid with surrounding edema. edema of the dorsal capsule of the wrist consistent with asprain. small ganglion cyst of the palmar aspect of the wrist. preliminary report faxed to the referring physician in sportsmedicine clinic at this time.

(Slide credit: Nigam Shah)

X-ray of foot

three views of the left foot without comparison show normal mineralization and alignment. projecting along the lateral aspect of the calcaneus are two ovoid-appearing ossific structures which are partially corticated. this is more posterior than the typical location of an os peroneum, and could represent injury to the peroneus longus tendon or possibly fracture of an os peroneum though the acuity is uncertain. please correlate with specific site of patient's reported foot pain. incidental note of small ossicle in the first-second interface likely due to an os intermetatarsarium there are no other acute findings. recommendation: if the lateral hind foot corresponds with pain or there is tendon dysfunction, consider magnetic resonance imaging for further evaluation.

(Slide credit: Nigam Shah)

Progress note: SOAP format

Diagnosis	Procedures
839 OT DISLOCATION	12345 Test for notes
839.21 DISLOCAT THORACIC VERT CLOSE	
847 SPRAINS OT/UNSPEC BACK	
723.1 CERVICALGIA	

Subjective:

The patient indicated on her visit today that she has been feeling a slight bit better in the right cervical area. Also, the pain on the right in the upper back area has been feeling slightly better. She also states that the head pain hasn't been nearly as bad lately. Mrs. Patient reported her neck pain at 6, upper back pain at 4. and headache at 1, based on a 1 to 10 pain scale and a percentage of improvement of her headache at 90%.

Objective:

The 1st cervical vertebra is found to be in a right posteriorly rotated subluxation with a moderate amount of spinal joint fixation. Cervical segment C5 is found to be in a left posterior malaligned position with a moderate degree of fixation. The T1 segment was noted to be subluxated posteriorward on the right with a moderate fixation of the spinal joints. On examination of the spinal joints, a fixation of a moderate degree at T8 was detected. A very intense level of pain and discomfort at C1 to T1 on the right was found on palpation of the spine.

Assessment:

The patient is approaching MMI.

Plan:

The appointment schedule is for Monday, Wednesday, and Friday treatment. Adjustment was recommended to correct misalignment and relieve joint fixation in the full spine region. In order to promote healing and reduce inflammation, electro stimulation of the muscles was administered to the neck area, and region of the thoracic spine.

(Slide credit: Nigam Shah)

Progress note: example

Patient 1
8/15/2003
Denies heroin or other illicit drug use. Drinks occasional beer. Last two urine tests have been negative for drugs.

No new psychosocial difficulties. Seemingly spending more time at home and no reports no difficulties at work.

Brief physical shows that his BP continues elevated 152/92.

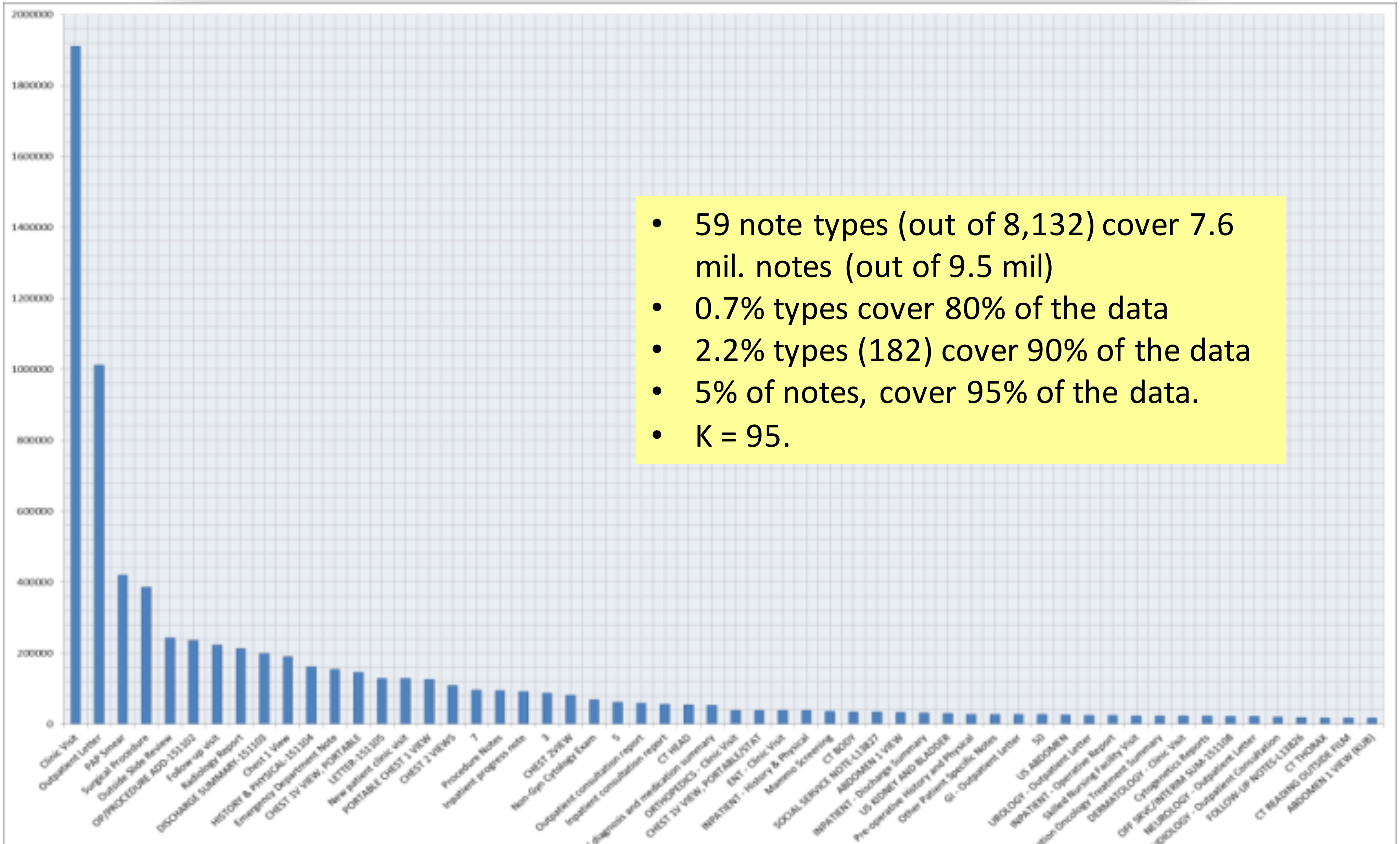
Reports that he attends NA weekly and continues in the weekly support group (which is confirmed by the group leader.)

Impression
Heroin use currently in remission.
Participating in program of recovery and by self-report is using Subutex as directed.
Mild blood pressure elevation

Rx Plan
Continue Subutex 16 mg daily
Discussed BP elevation and the importance of developing an exercise program and low salt diet.
Return visit 3 weeks

(Slide credit: Nigam Shah)

STRIDE Note Types



- 59 note types (out of 8,132) cover 7.6 mil. notes (out of 9.5 mil)
- 0.7% types cover 80% of the data
- 2.2% types (182) cover 90% of the data
- 5% of notes, cover 95% of the data.
- K = 95.

(Slide credit: Nigam Shah)

Clinical vs. Biomedical Text

- **Biomedical text** appears in books, articles, literature abstracts, posters.
- **Clinical text** is written by clinicians or healthcare providers. Describes patients, their pathologies, their personal, social and medical histories, findings made during interviews or procedures, etc.

(Slide credit: Nigam Shah)

What could we do this this clinical text? What are examples where it provides complementary or distinct information from other structured data that might be available?

Take two minutes, and brainstorm with a partner ideas for how to use the clinical text

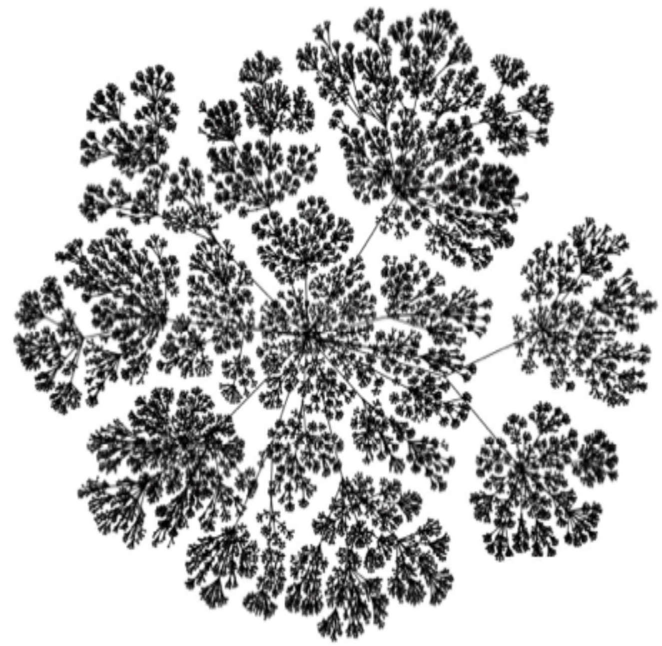
Applications of clinical NLP

Application: automated coding

(Slide credit: Noemie Elhadad)

```
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. Righ
t 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 t
obacco use. Smoked about 1 ppd from age 19 to age 65. Denies use of alcohol. Fat
her 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
```

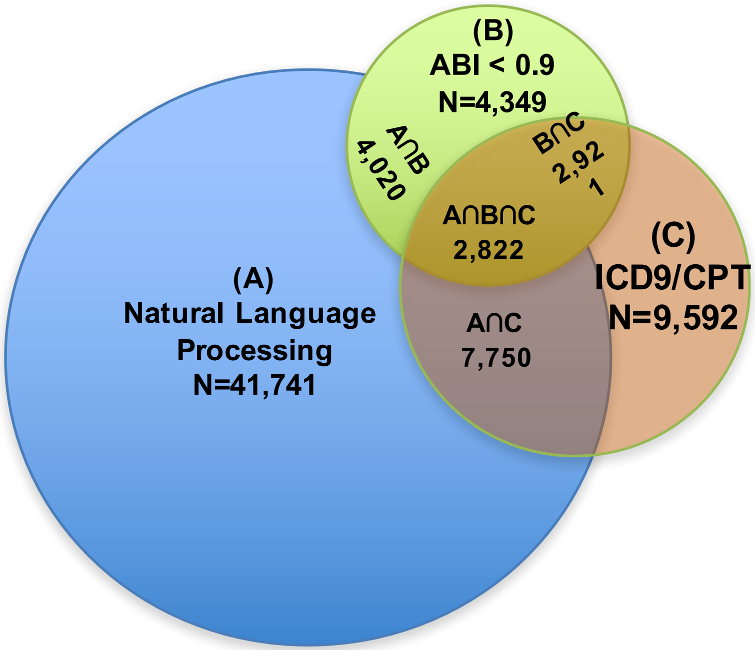
+



→ 401.1 (Hypertension);
428.0 (Congestive heart failure);
369.6 (One eye blindness)

Application: cohort detection

(Slide credit: Noemie Elhadad)

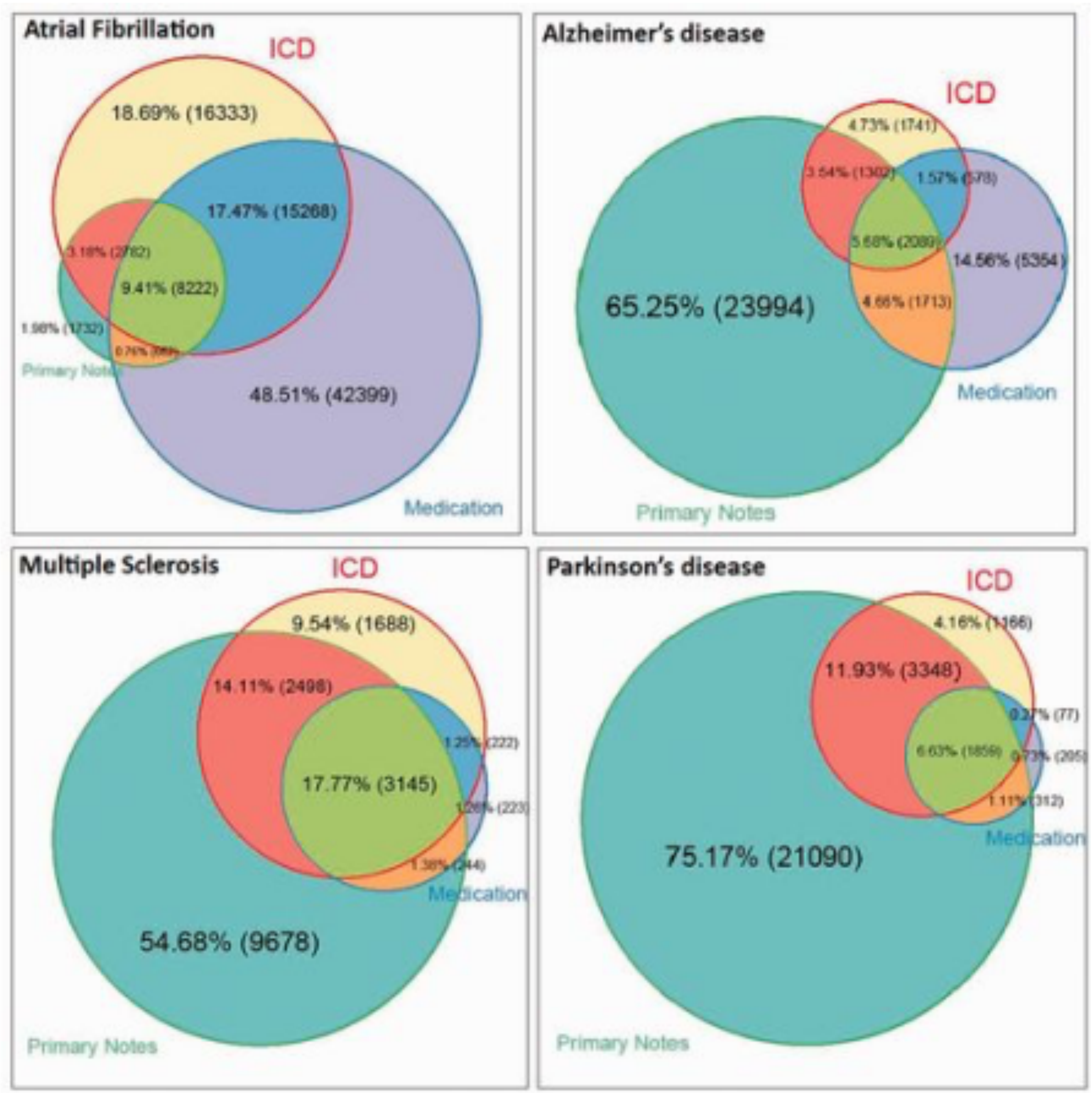


NLP detected 4x more patients than traditional algorithms. More importantly, many patients with Peripheral Arterial Disease (PAD) are missed using standard approaches.

PAD Detection Algorithm	# Unique Patients	Specificity
NLP PAD Algorithm	**41741**	**98%**
Rest Pain	2498	98%
Diminished pulses	5773	92%
Ischemic Limb NLP	1339	99%
Peripheral Arterial Disease NLP	31430	99%
Claudication	15337	96%

Duke JD, Chase M, Ring N, Martin J, Fuhr R, Hirch A. (2016) Natural Language Processing to Augment Identification of Peripheral Arterial Disease Patients in Observational Research. *American College of Cardiology Annual Symposium*.

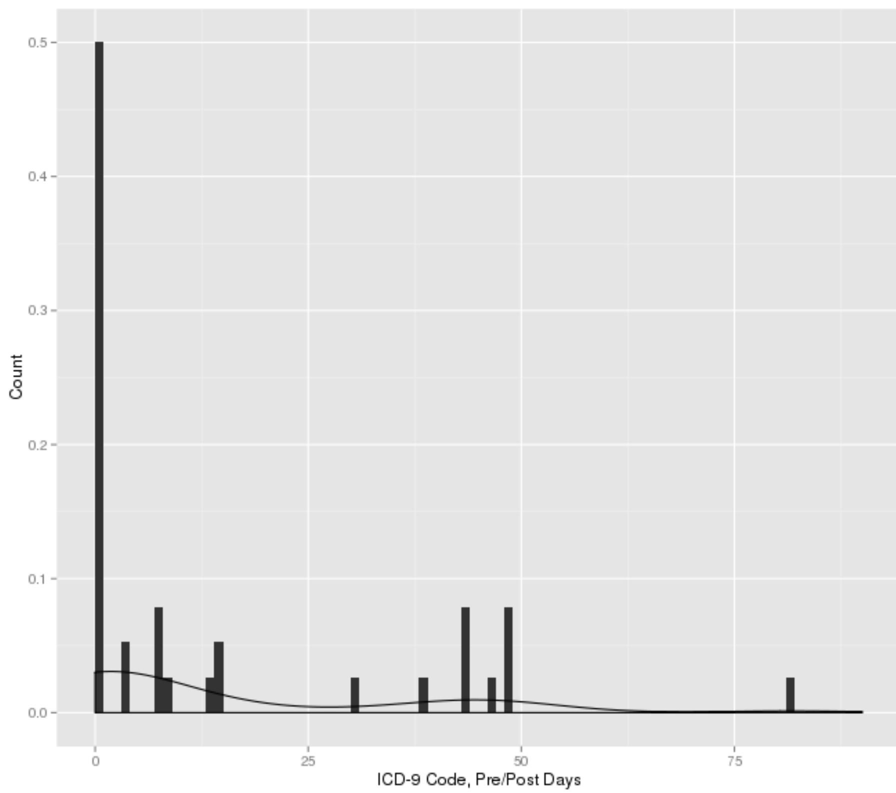
The utility of looking into notes

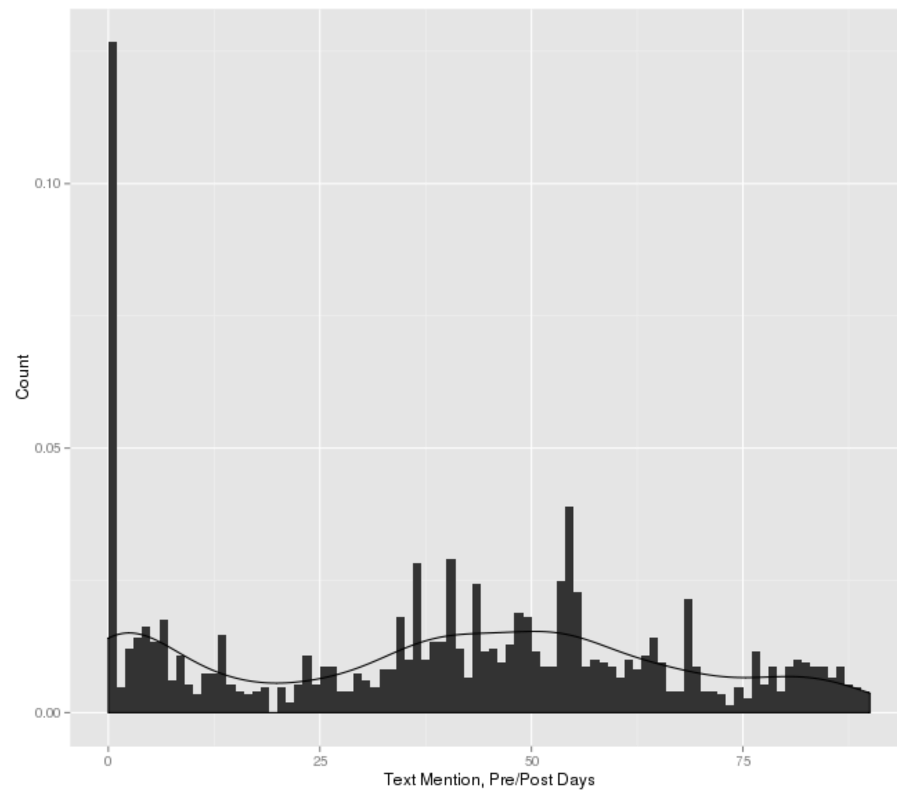


Wei WQ, et al 2015 JAMIA

(Slide credit: Nigam Shah)

The utility of looking into notes





(Slide credit: Nigam Shah)

Application: clinical decision support

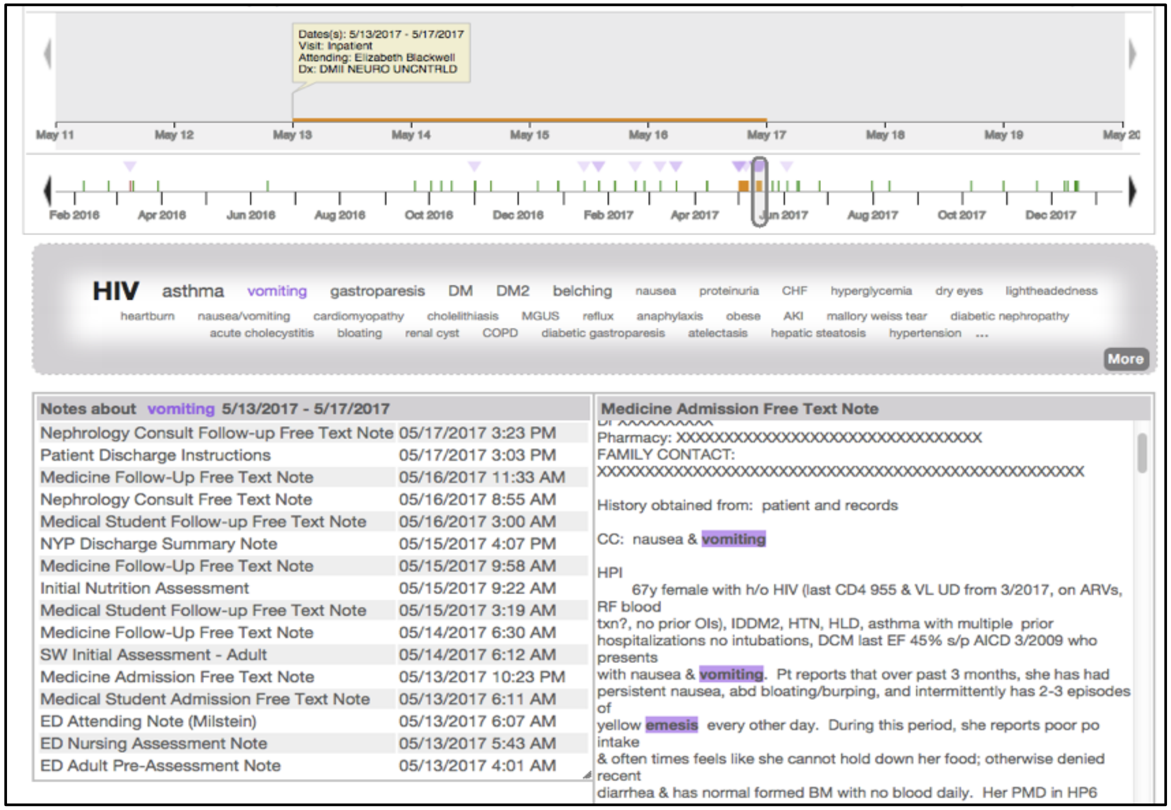
(Slide credit: Noemie Elhadad)

- Leverage information from the clinical notes within logic of clinical decision support
  - –Drug –drug interactions
  - –Allergies
  - –…

Demner-Fushman D, Chapman W, McDonald C. (2009) What can natural language processing do for clinical decision support? J Biomed Inform. 42(5):760-762
Demner-Fushman D, Elhadad N. (2016) Aspiring to unintended consequences of natural language processing: a review of recent developments in clinical and consumer-generated text processing. IMIA Yearbook of Medical Informatics.

Application: data exploration

(Slide credit: Noemie Elhadad)

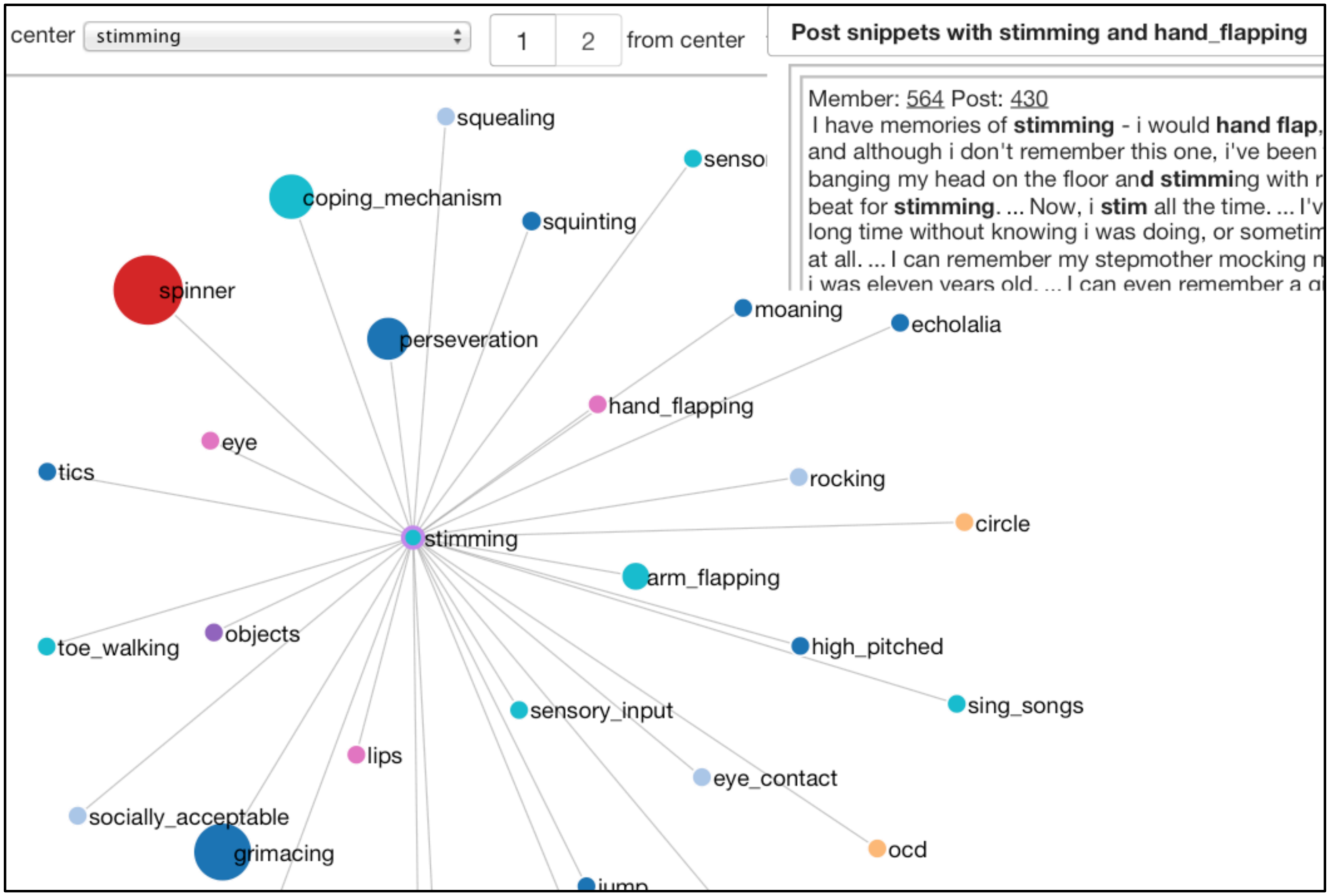


Hirsch J, Tanenbaum J, Lipsky Gorman S, Liu C, Schmitz E, Hashorva D, Ervits A, Vawdrey D, Sturm M, Elhadad N. (2015) HARVEST, a longitudinal patient record summarizer. *J Am Med Inform Assoc*. 22(2):263-274.

Pivovarov R, Coppleson Y, Lipsky Gorman S, Vawdrey D, Elhadad N. (2016) Can patient record summarization support quality metric abstraction? Am Med Inform Assoc Symp.

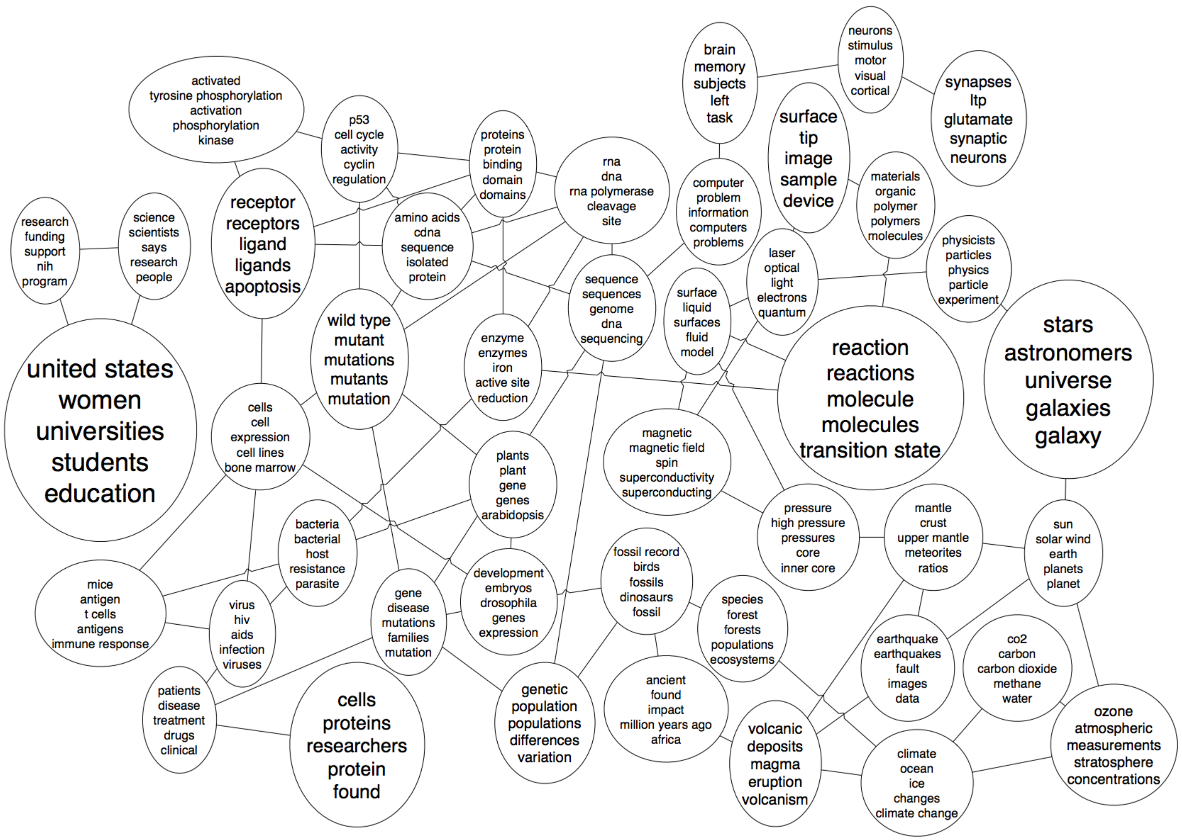
Application: data exploration

(Slide credit: Noemie Elhadad)



Application: data exploration

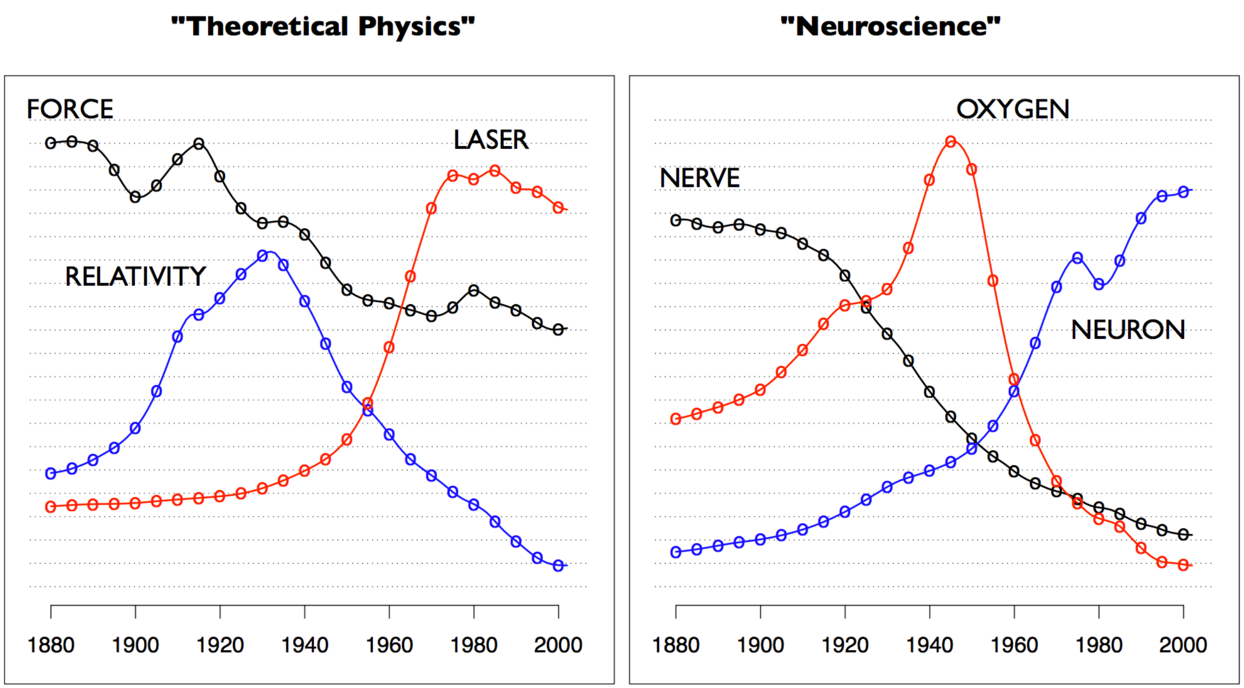
(Slide credit: Noemie Elhadad)



Blei D, Lafferty J. (2007) A correlated topic model of Science. Annals of Applied Statistics. 1(1):17-35.

Application: data exploration

(Slide credit: Noemie Elhadad)



Blei D, Lafferty J. (2007) A correlated topic model of Science. Annals of Applied Statistics. 1(1):17-35.

Application: info-surveillance from public social media

(Slide credit: Noemie Elhadad)


Jacob D.
Brooklyn, NY
1 friend
66 reviews

2/1/2014

I got very sick from eating here today so be careful. I had dim sum, noodles, and fried dumplings and felt fine until about 6 hours later and then just started vomiting profusely several times. I had no other food today and no alcohol. I will never eat there again.


Jay Dabhi 92.3NowFM @JayDabhi · Mar 6
Went to @GoldenKrustBkry for Patties. They tried to charge me 50 cents for tap water! My tummy is in shambles. Poor cust service #neveragain
Details Reply Retweet Favorite More

Golden Krust Bakery @GoldenKrustBkry · Mar 6
@JayDabhi We're so sorry to hear. Which location was this at?
Details Reply Retweet Favorite More

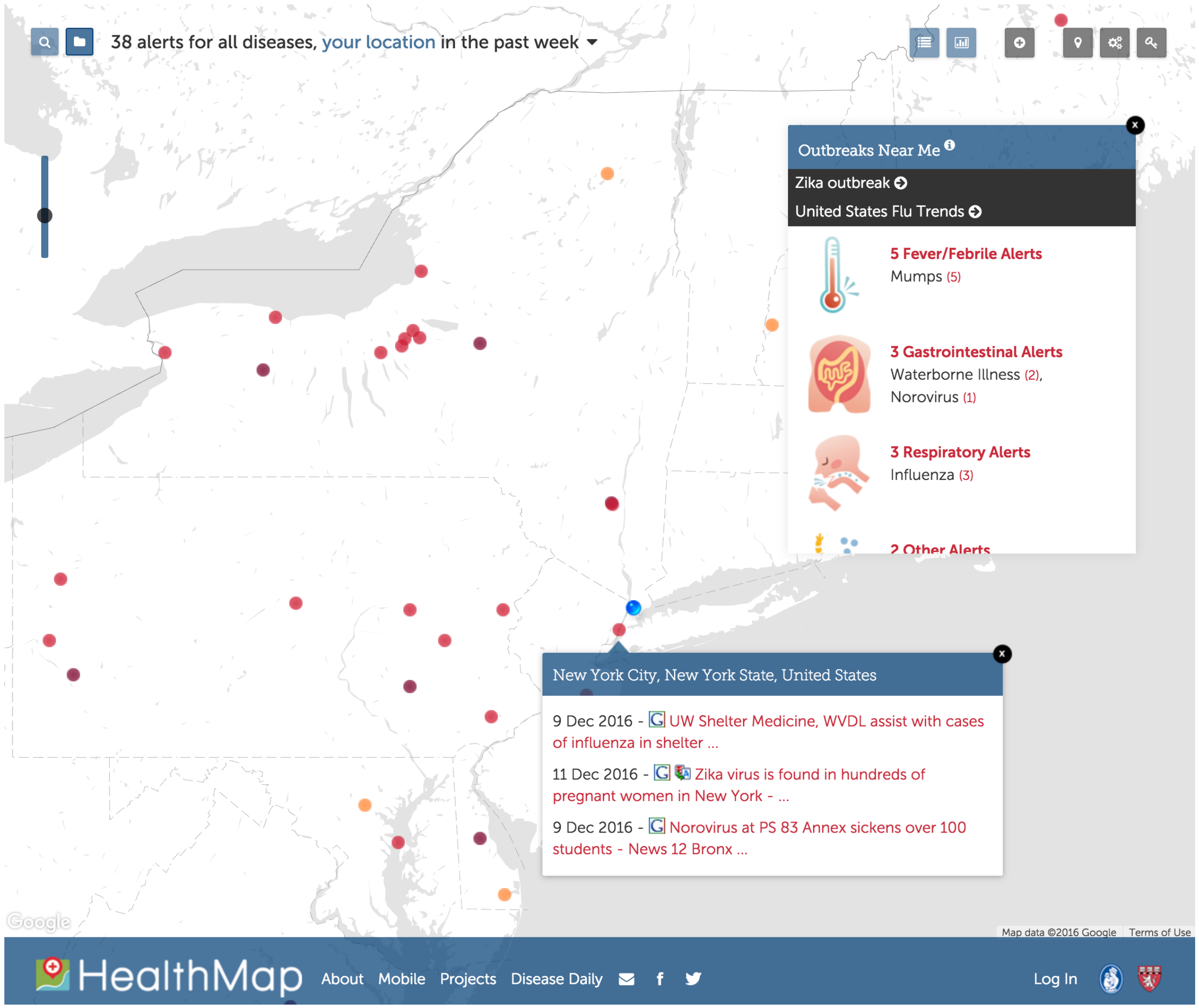
Jay Dabhi 92.3NowFM @JayDabhi · Mar 8
@GoldenKrustBkry Location is 39th St & 8th Ave. Very rude woman at register & food caused diarrhea. Disappointing.
Details Reply Retweet Favorite More

Harrison C, Jorder M, Stern H, Stavinsky F, Reddy V, Hanson H, Waechter H, Lowe L, Gravano L, Balter S. (2014) Using online reviews by restaurant patrons to identify unreported cases of foodborne illness– New York City, 2012-2013. Centers for Disease Control and Prevention's Morbidity and Mortality Weekly Report (MMWR), 63(20):441–445.

Paul M, Dredze M. (2011) You are what you tweet: Analyzing Twitter for public health. In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media.

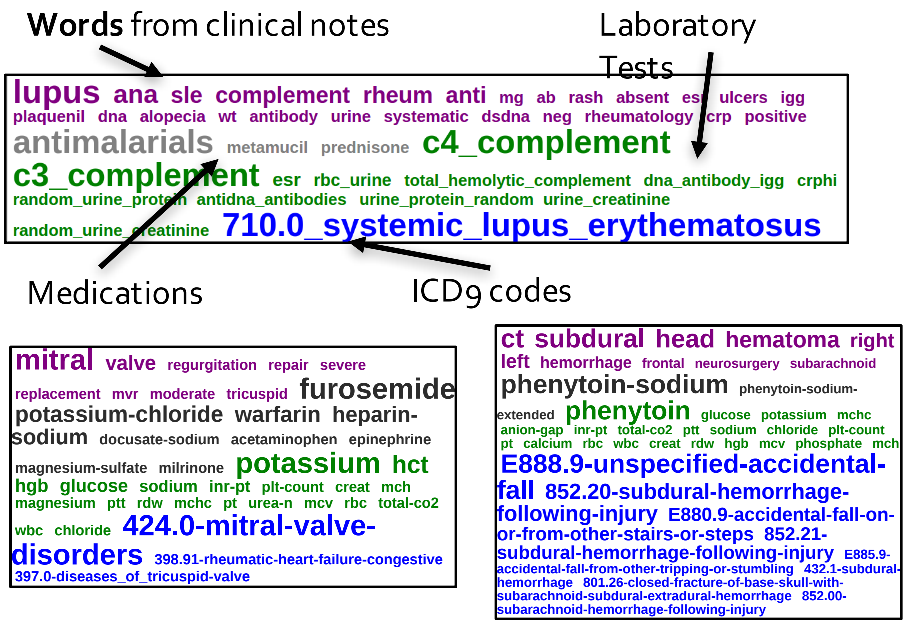
Application: info-surveillance

(Slide credit: Noemie Elhadad)

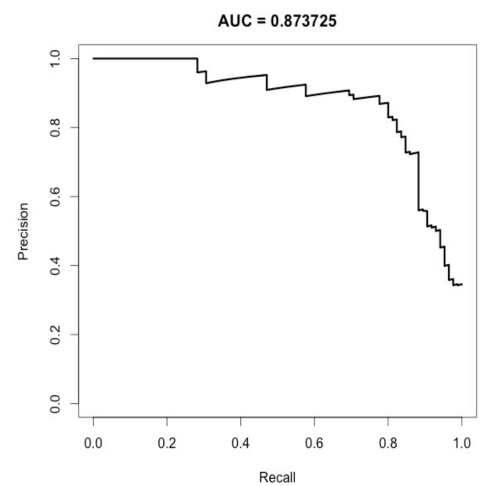


Application: high throughput phenotyping

(Slide credit: Noemie Elhadad)



DM2 cohort identification (n=2,500)



Pivovarov R, Perotte A, Grave E, Angiolillo J, Wiggins C, Elhadad N. (2015) Learning probabilistic phenotypes from heterogeneous EHR data. *J Biomed Inform.* 58:156-165.

Application: predictive analytics

(Slide credit: Noemie Elhadad)

MHs (Mental Health subreddits)
I have been considering going for some formal therapy. Any suggestions?
Everyday I feel sad and lonely
Since past sometime I think I am having panic attacks. I really need help from you guys.
It has been so many years, I feel I still can't move on. I am noticing behavior what could be considered "triggers" now.
SW (SuicideWatch)
I know I was never meant to lead this life.
Don't want to hurt the people I care but I can't take this anymore.
Today I felt I have nothing left, why am I even living... I don't see a point.
I'd kill myself, but the other part of me tells me not to waste all the money my parents invested on me..

Table 1: Example titles of posts in the MHs and SW datasets; content has been carefully paraphrased to protect the privacy of the individuals.

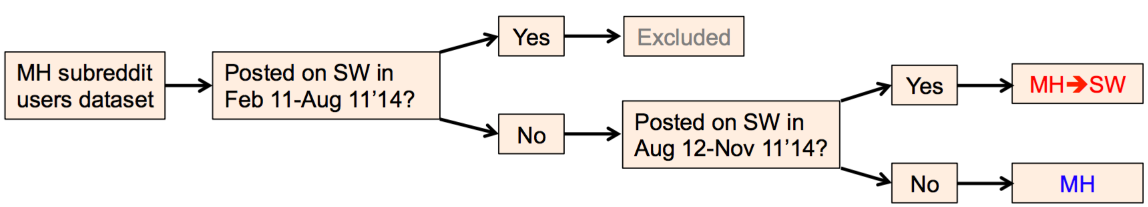


Figure 1: Schematic diagram of obtaining MH → SW and MH classes of users.

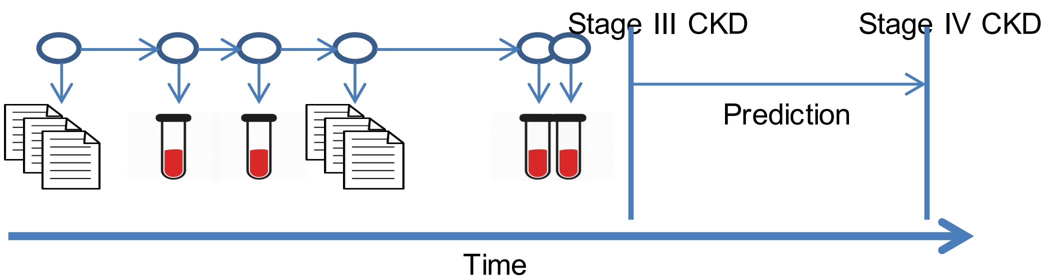
	MH	MH → SW	z	p
Linguistic Structure				
nouns	0.294	0.125	6.51	***
verbs	0.045	0.107	2.19	**
abverbs	0.048	0.099	4.87	***
readability index	0.609	0.232	5.51	***
accommodation	0.857	0.487	5.46	**
Interpersonal Awareness				
1st person singular	0.018	0.086	-10.6	***
1st person plural	0.093	0.078	4.53	*
2nd person	0.058	0.031	8.01	*
3rd person	0.087	0.042	6.32	***
Interaction				
posts authored	18.97	10.31	2.53	*
post length	215.62	443.73	-15.4	***
comments authored	122.42	106.22	0.95	-
comments received	19.862	13.414	1.05	*
comment length authored	63.417	87.116	-1.88	*
comment length received	42.323	26.362	5.44	**
response velocity (mins)	7.746	6.966	0.84	-
vote difference	28.788	7.681	7.18	***

Table 2: Differences between MH → SW and MH user classes based on linguistic structural, interpersonal awareness and interaction measures. Statistical significance is reported based on Wilcoxon signed rank tests at levels $p = .05/N; .01/N; .001/N$, $(N = 17)$, following Bonferroni correction.

De Choudhury M, Kiciman E, Dredze M, Coppersmith G, Kumar M. (2016) *Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media*. CHI'16.

Application: predictive analytics

(Slide credit: Noemie Elhadad)



Survival Model (n=2,617)	Concordance (n=291)
(Text + Lab) Kalman Filter	0.849
Lab Kalman Filter	0.836
Recent Labs	0.819
Text Kalman Filter	0.733
eGFR risk score	0.779

(Heart Failure)
Lasix
Volume
Edema
Heart
Failure
Worsening
Diuresis
Severe
Diastolic
Overload

(Diabetes)
Units
Insulin
Subcutaneous
Lantus
Glucose
Diabetes
Times
70/30
Diabetic
Days

(Dialysis)
q15
Dialysis
Fistula
Volume
Bid
Lasix
Placement
Improved
Heparin
Examined

(Health Maintenance)
Flu
Visit
Fasting
Colonoscopy
Year
Shot
Vaccine
wnl
Check
Primary

(Gynecological)
Breast
Vaginal
Mammo
Cancer
Hx
Pap
nl
Age
will
Endometrial

(Asthma)
Albuterol
Asthma
Inhaled
Lung
Obstructive
Wheezing
Advair
Pulm
Restrictive
Puffs

Perotte A, Ranganath R, Hirsch J, Blei D, Elhadad N (2015). Risk Prediction for Chronic Kidney Disease Progression Using Heterogeneous Electronic Health Record Data and Time Series Analysis. *J Am Med Inform Assoc*. 22(4):8720

Now that we have seen a few example documents and applications, what do you think are the main problems we might face in trying to do "natural language processing" on such content.

Take two minutes, and discuss with a partner the issues we might face.

(Slide credit: Nigam Shah)

My dream: text *understanding*

My dream: text *understanding*

“pt with fever, chills, N/V since friday after eating what hethought was undercooked meat. Unable to hold po's down. Fevers to 103”

↓

Medical history / context
- Possible food poisoning

Symptoms:
- fever and chills (now)
- nausea and vomiting (for previous X days)
- unable to keep any foods/liquids down (recent past)
- fever as high as 103 (recent past)

My dream: text *understanding*

“89 yo f s/p esophageal hernia repair 3/09 w/ ?g-tube placement now w/ c/o's n&v. family reports pt's appetite is decreased, no BM x3d. generally not feeling well, had a bad day”

↓

Medical history / context:
- 89 years old
- female
- recent hernia repair; has feeding tube

Symptoms:
- nausea and vomiting
- decreased appetite
- no bowel movements (for 3 days)
- malaise (today)

My dream: text *understanding*

“from the scene fall of horse landed on r thigh deformity iv fluid 100 fentanyl/ morhpine 4. no head or neck pain/”

↓

Medical history / context:
- very recent trauma injury
- currently on pain killers

Symptoms:
- thigh deformity (since accident)
- no head pain (since accident)
- no neck pain (since accident)

Problems unique to clinical text

- Ungrammatical, has misspellings and concatenations. Contains short telegraphic phrases, acronyms, abbreviations, which are often overloaded **= haiku of acronyms**
- Some sources are dictated and composed deliberately for clear communication (radiology reports) while others are written for documentation (progress notes) **= high variance in quality**
- Can contain many things that can be typed or pasted, such as long sets of lab values or vital signs **= pasted in junk**
- Idiosyncratic and institution-specific template-use is common **= lot of copy-pasting**
- Pervasive fear, misunderstanding, and confusion around security, privacy, de-identification, and anonymization **= ridiculous amount of agony in getting access**

(Slide credit: Nigam Shah)

How do we get there?

First question: how do we **represent** the structured data?

The UMLS consists of

Metathesaurus

1 million+ biomedical concepts from over 100 sources

Semantic Network

135 broad categories and 54 relationships between categories

SPECIALIST Lexicon & Tools

lexical information and programs for language processing

3 Knowledge Sources

used separately or together

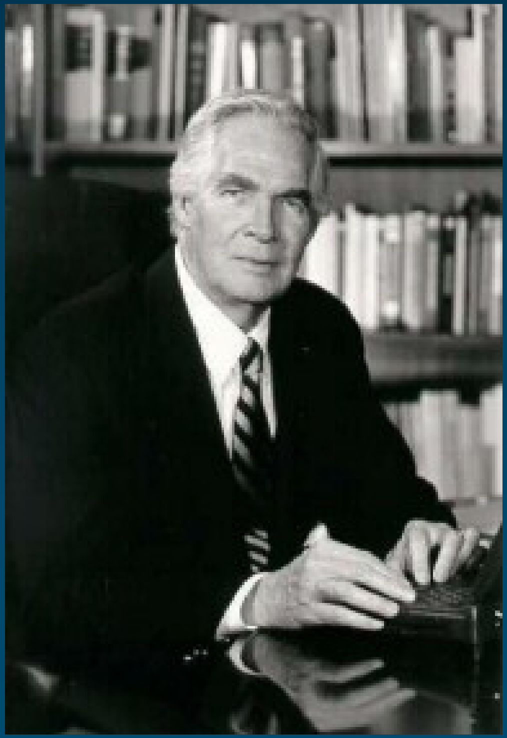
(Slide credit: Rachel Kleinsorge and Jan Willis, “UMLS Basics class”)

History of the UMLS

[Lindberg & al., *Methods*, 1993]
[Humphreys & al., *JAMIA*, 1998]

- Started at National Library of Medicine, 1986
- "Long-term R&D project"
- Complementary to IAIMS

(Integrated Academic Information Management Systems)



«[...] the UMLS project is an effort to overcome two significant barriers to effective retrieval of machine-readable information.

- The first is the variety of ways the same concepts are expressed in different machine-readable sources and by different people.
- The second is the distribution of useful information among many disparate databases and systems.»

(Slide credit: Rachel Kleinsorge and Jan Willis, "UMLS Basics class")

Metathesaurus: clusters terms by meaning

- Synonymous terms clustered into a concept
- Preferred term is chosen
- Unique identifier (CUI) is assigned

Addison's disease	Metathesaurus	PN	
Addison's disease	SNOMED CT	PT	363732003
Addison's Disease	MedlinePlus	PT	T1233
Addison Disease	MeSH	PT	D000224
Bronzed disease	SNOMED Intl 1998	SY	DB-70620
Deficiency; corticorenal, primary	ICPC2-ICD10 Thesaurus	PT	MTHU021575
Primary Adrenal Insufficiency	MeSH	EN	D000224
Primary hypoadreanlism syndrome, Addison	MedDRA	LT	10036696

C0001403 | Addison's disease

(Slide credit: Rachel Kleinsorge and Jan Willis, "UMLS Basics class")

Semantic Network

- 135 Semantic Types
  - Broad subject categories (Clinical Drug, Virus)
  - Ex:
    - **Addison's Disease**
    - **Semantic Type: Disease or Syndrome**
- 54 Semantic Relationships
  - Links between categories (isa, causes, treats)
  - Ex:
    - Virus **causes** Disease or Syndrome
- Types + Relationships
  - Form the structure of the semantic network
  - Broadly categorize the biomedical domain

(Slide credit: Rachel Kleinsorge and Jan Willis, "UMLS Basics class")

Concept cluster of synonymous terms

Concept C0001621

Term adrenal disease gland L0001621
- S0011232 *Adrenal Gland Diseases*
- S0011231 Adrenal Gland Disease
- S0000441 Disease of adrenal gland
- S0481705 Disease of adrenal gland, NOS
- S0220090 Disease, adrenal gland
- S0044801 Gland Disease, Adrenal

Term adrenal disorder gland unspecified L0041793
- S0860744 *Disorder of adrenal gland, unspecified*
- S0217833 Unspecified disorder of adrenal glands

Term adrenal disorder L0161347
- S0225481 *ADRENAL DISORDER*
- S0627685 DISORDER ADRENAL (NOS)

Term adrenal disorder gland L0181041
- S0632950 *Disorder of adrenal gland*
- S0354509 Adrenal Gland Disorders

Term L0162317
- S0226798 *SURRENALE, MALADIES* FRE

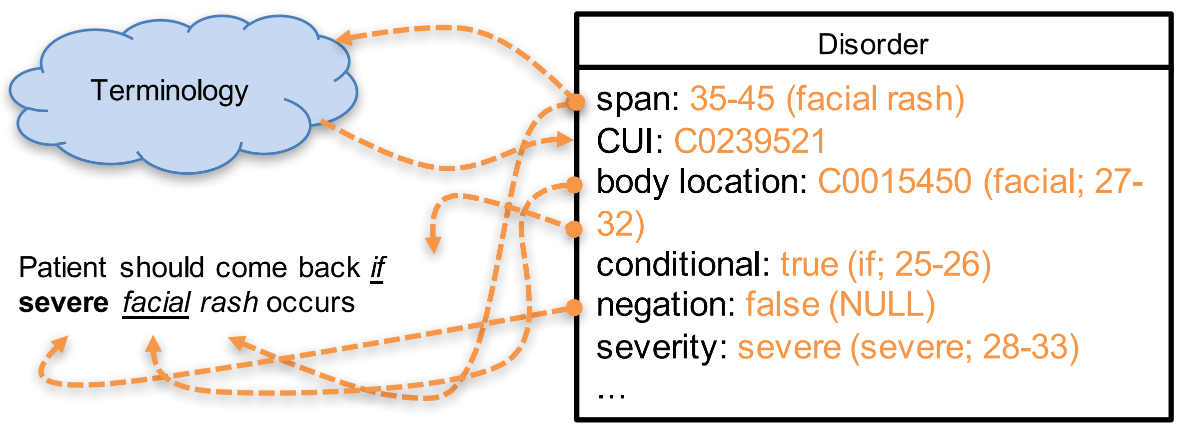
(Slide credit: Rachel Kleinsorge and Jan Willis, "UMLS Basics class")

How do we get there?

How do we extract the relevant concepts and understand their broader context?

Information extraction

(Slide credit: Noemie Elhadad)



Example of mapping to UMLS (work in my lab)

- Goal: identify mentions of medical problems in text and map them to UMLS concepts

- First step: identify the mentions in text
  - Tag tokens in the input text to indicate their involvement in a mention

(Slide credit: Ankit Vani and Yacine Jernite, NYU)

Tagging scheme

B	First token of the mention
I	Other tokens of the mention
O	Everything else
OD	Within scope of a mention but not part of the mention itself
ID	Tokens which are part of a discontinuous mention
In	Identifying token in overlapping mentions
Bn	Identifying token in overlapping mentions, first word of the mention
Ip	Part of only one of two overlapping mentions, but not the identifying token

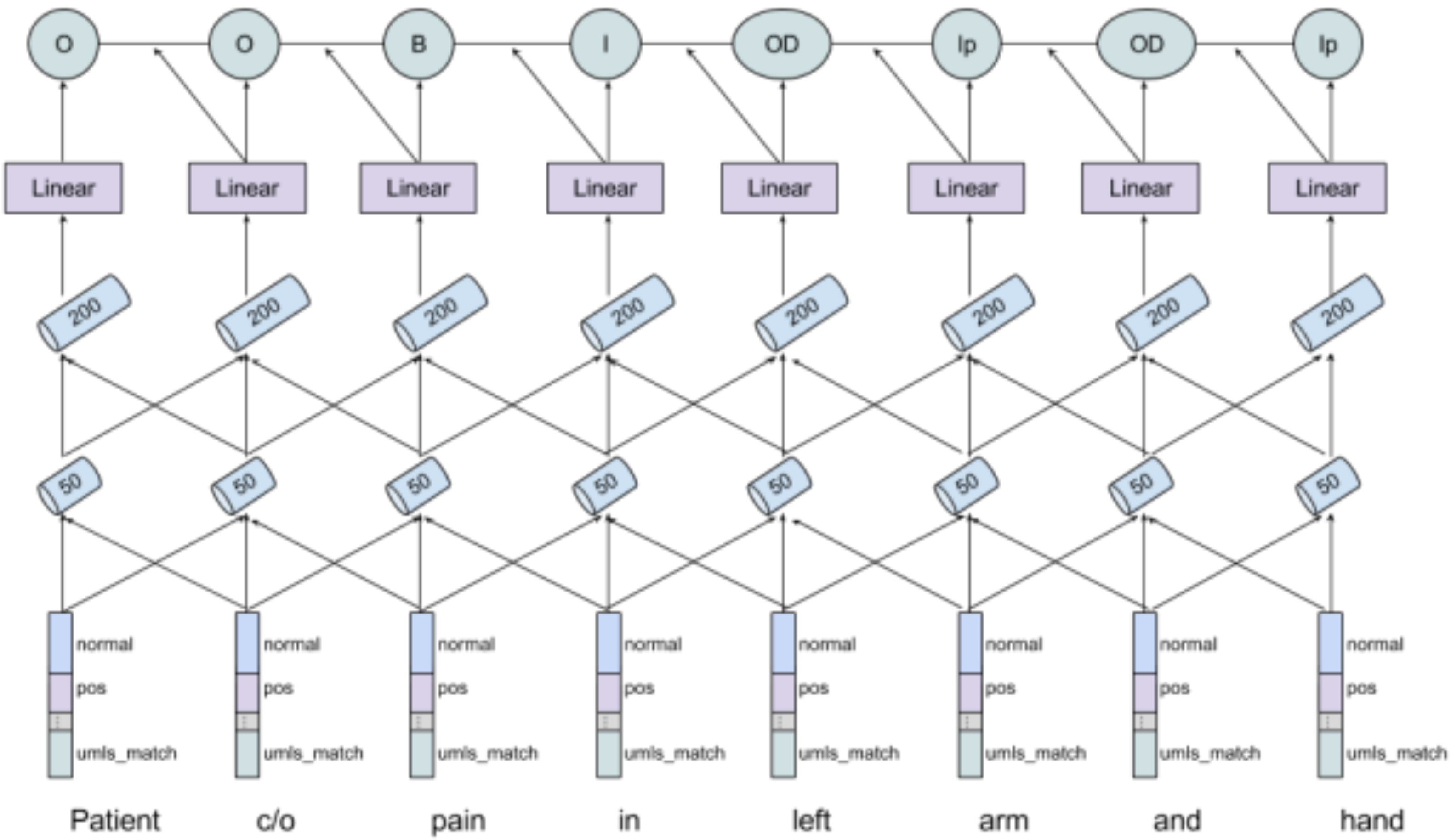
(Slide credit: Ankit Vani and Yacine Jernite, NYU)

Tagging examples

- `the patient suffers from a broken jaw .`
  `O   O   O   O O B   I O`
- `the pain is strongest in the arm .`
  `O   B  OD   OD   OD  OD  ID O`
- `left arm and shoulder are swollen`
  `B   In  OD  In   OD  ID`
- `elbow and wrist broken`
  `Bn  OD  Bn   ID`
- `inflammation of left` *`kidney`* `and spleen`
  `B   OD  In  Ip  OD  In`

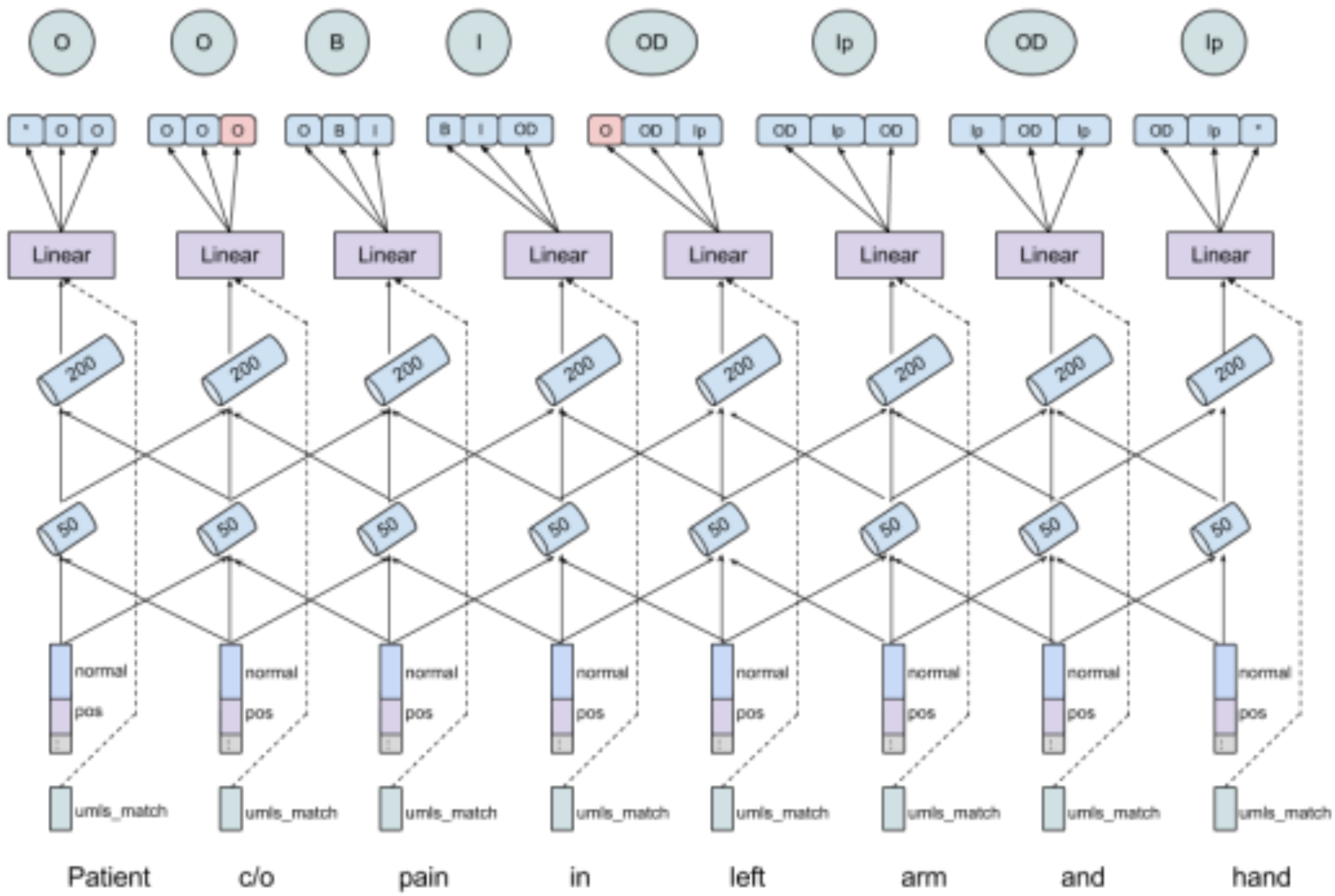
(Slide credit: Ankit Vani and Yacine Jernite, NYU)

Deep conditional random field



(Slide credit: Ankit Vani and Yacine Jernite, NYU)

Window prediction without CRF



(Slide credit: Ankit Vani and Yacine Jernite, NYU)

Information extraction (Modifiers)

(Slide credit: Noemie Elhadad)

- CUI (normalization)

  "presented with facial rash"

  Facial rash (CUI C0239521)

- Negation

  "patient denies numbness"

- Subject

  "son has schizophrenia"

- Uncertainty

  "evaluation of MI"

- Course

- 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)

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.

Negation and Context detection

no abnormal [PREN]
no cause of [PREN]
no complaints of [PREN]
no evidence [PREN]
no new evidence [PREN]
no other evidence [PREN]
no evidence to suggest [PREN]
no findings of [PREN]
no findings to indicate [PREN]
no mammographic evidence of
no new [PREN]
no radiographic evidence of
no sign of [PREN]
no significant [PREN]
no signs of [PREN]
no suggestion of [PREN]
no suspicious [PREN]
not [PREN]
not appear [PREN]
not appreciate [PREN]
not associated with [PREN]
not complain of [PREN]
not demonstrate [PREN]
not exhibit [PREN]
not feel [PREN]
not had [PREN]

```
T/SICU Nursing Admission Note:

This is a 31 year old male s/p seizure on ladder with resulting fall
15-20 feet on [**09-17**] now presenting to the T/SICU post surgical
repair of multiple facial fractures, right mandibular fracture, and
left distal radius fracture. He needs to remain intubated for 48
hours post-op. His past medical history is significant only for
seizure disorder, and his only medication is depakote. He has no
known allergies.
```

```
     Nursing Admission Note:

  is a    year old male     seizure    ladder                    fall
 20 feet                  presenting                     post surgical
repair   multiple facial fractures, right mandibular fracture,
left distal radius fracture.    needs                            48
hours post-op.     past medical history    significant
seizure disorder,             medication    depakote.            no
known allergies.
```

Negex and Contex projects
https://github.com/chapmanbe/pyConTextNLP

(Slide credit: Nigam Shah)

Example NLP pipeline (cTAKEs)

An example of a sentence discovered by the sentence boundary detector:

Fx of obesity but no fx of coronary artery diseases.

Tokenizer output – 11 tokens found:

```
Fx  of  obesity  but  no  fx  of  coronary  artery  diseases  .
```

Normalizer output:

```
Fx  of  obesity  but  no  fx  of  coronary  artery  disease  .
```

Part-of-speech tagger output:

```
Fx  of  obesity  but  no  fx  of  coronary  artery  diseases  .
NN  IN  NN       CC   DT  NN  IN  JJ        NN      NNS       .
```

Shallow parser output:

```
Fx  of  obesity  but  no  fx  of  coronary  artery  diseases  .
NP  PP  NP            NP      PP            NP
```

Named Entity Recognition – 5 Named Entities found:

```
Fx of obesity but no fx of coronary artery diseases .
      obesity  (type=diseases/disorders, UMLS CUI=C0028754, SNOMED-CT codes=308124008 and 5476005)
                           coronary artery diseases (type=diseases/disorders, CUI=C0010054, SNOMED-CT=8957000)
                           coronary artery (type=anatomy, CUI(s) and SNOMED-CT codes assigned)
                                    artery (type=anatomy, CUI(s) and SNOMED-CT codes assigned)
                                    diseases (type=diseases/disorders, CUI = C0010054)
```

Status and Negation attributes assigned to Named Entities:

```
Fx of obesity but no fx of coronary artery diseases .
      obesity (status = family_history_of; negation = not_negated)
                           coronary artery diseases (status = family_history_of, negation = is_negated)
```

Figure 1 Example sentence processed through cTAKES components 'family history of obesity but no family history of coronary artery diseases.' Fx, family history.

(Slide credit: Nigam Shah)

How do we get there?

What data is available for training and evaluating clinical NLP algorithms?

https://www.i2b2.org/about/index.html

i2b2

A National Center for Biomedical Computing

Informatics for Integrating Biology & the Bedside

About Us | Driving Biology Projects | Software | Resources | Events | Training | News | Collaborations | Publications

About Us

- Overview
- Introduction
- Mission Statement
- Structure
- Contact Info
- Sponsors
- Links

Overview

Informatics for Integrating Biology and the Bedside (i2b2) is an NIH-funded National Center for Biomedical Computing (NCBC) based at Partners HealthCare System in Boston, Mass. Established in 2004 in response to an NIH Roadmap Initiative RFA, this NCBC is one of four national centers awarded in this first competition (http://www.bisti.nih.gov/ncbc/); currently there are seven NCBCs. One of 12 specific initiatives in the New Pathways to Discovery Cluster, the NCBCs will initiate the development of a national computational infrastructure for biomedical computing. The NCBCs and related R01s constitute the National Program of Excellence in Biomedical Computing.

The i2b2 Center, led by Director Isaac Kohane, M.D., Ph.D., Professor of Pediatrics at Harvard Medical School at Children's Hospital Boston, is comprised of six cores involving investigators from the Harvard-affiliated hospitals, MIT, Harvard School of Public Health, Harvard Medical School and the Harvard/MIT Division of Health Sciences and Technology. This Center is funded under a Cooperative agreement with the National Institutes of Health.

The i2b2 Center is developing a scalable computational framework to address the bottleneck limiting the translation of genomic findings and hypotheses in model systems relevant to human health. New computational paradigms (Core 1) and methodologies (Cores 3) are being developed and tested in several diseases (airways disease, hypertension, type 2 diabetes mellitus, Huntington's Disease, rheumatoid arthritis, major depressive disorder, inflammatory bowel disease, multiple sclerosis) (Core 2 Driving Biological Projects).

https://www.i2b2.org/NLP/DataSets/Main.php

2008 Obesity Challenge

[Obesity Challenge Participants]

NLP Data Set #2:

Please cite as:

- Uzuner Ö. (2009). "Recognizing Obesity and Co-morbidities in Sparse Data". *Journal of the American Medical Informatics Association*. July 2009; 16(4): 561-570. http://jamia.bmj.com/content/16/4/561.full.pdf.

2009 Medication Challenge

[Medication Challenge Participants]

NLP Data Set #3:

Please cite as:

- Uzuner Ö, Solti I, Xia F, Cadag E. (2010). "Community Annotation Experiment for Ground Truth Generation for the i2b2 Medication Challenge". *Journal of the American Medical Informatics Association*. 2010;17:519-523 doi:10.1136/jamia.2010.004200. http://jamia.bmj.com/content/17/5/519.full.pdf.
- Uzuner Ö, Solti I, Cadag E. (2010). "Extracting Medication Information from Clinical Text". *Journal of the American Medical Informatics Association*. 2010;17:514-518 doi:10.1136/jamia.2010.003947. http://jamia.bmj.com/content/17/5/514.full.pdf.

2010 Relations Challenge

[Relations Challenge Participants]

NLP Data Set #4:

alt.qcri.org/semeval2015/task14/

Home | Task Description | Data and Tools | Results | Relevant Publications | Important Dates | FAQs

SemEval-2015 Task 14

SemEval-2015 Task 14: Analysis of Clinical Text

The purpose of this task is to enhance current research in natural language processing methods used in the clinical domain. The second aim of the task is to introduce clinical text processing to the broader NLP community. The task aims to combine supervised methods for text analysis with unsupervised approaches. More specifically, the task aims to combine supervised methods for entity/acronym/abbreviation recognition and mapping to UMLS CUIs (Concept Unique Identifiers) with access to larger clinical corpus for utilizing unsupervised techniques. It also comprises the task of identifying various attributes of the disorders and normalizing their values. We refer to this as the template filling task.

Contact Info

Organizers (in alphabetical order)

- Wendy W. Chapman, University of Utah
- Noemie Elhadad, Columbia University
- Suresh Manandhar, University of York, UK
- Sameer S. Pradhan, Harvard University
- Guergana K. Savova, Harvard University

Contact:

- Guergana.Savova@childrens.harvard.edu
- Noemie.Elhadad@columbia.edu

Task 1: Disorder Identification

In the disorder identification task, the goal is to recognize the span of a disorder mention and its normalization to a unique CUI in the UMLS/SNOMED-CT terminology in a set of clinical notes. (UMLS/SNOMED-CT is the set of CUIs in UMLS restricted to concepts that are part of the SNOMED-CT vocabulary).

Here are a few examples—more are provided in the annotation guidelines and in the page on Datasets. Given the following three sentences:

1. The rhythm appears to be *atrial fibrillation*.
2. The *left atrium* is moderately *dilated*.
3. 53 year old man s/p *fall from ladder*.

The spans of the disorder mentions are identified as follows: In examples 1. and 3., the phrases *atrial fibrillation* and *fall from ladder* fall in the disorder semantic group in the UMLS. Example 2. is a case of discontigous mentions represented by *left atrium...dialated*. This phenomena where a discontiguous phrase is the best representative of the disorder occurs more commonly in the clinical domain than in the general domain, and therefore is annotated as such.

The disorder entities identified in the examples above map to the following CUIs:

1. *atrial fibrillation* - C0004238; UMLS preferred term atrial fibrillation
2. *left atrium...dilated* - C0344720; UMLS preferred term left atrial dilatation
3. *fall from ladder* - C0337212; UMLS preferred term is accidental fall from ladder

Task 2: Disorder Slot Filling

For a given disorder mention, there are several attributes one can identify. This task focuses on identifying the normalized value for nine modifiers in a disorder mentioned in a clinical note: the CUI of the disorder (very much like in Task 1), as well as the potential attributes (negation indicator, subject, uncertainty indicator, course, severity, conditional, generic indicator, and body location) as described in Table 1. The Clinical Element Models are the original source of all of the attributes.

Attributes Types	**Example Sentence**	**Normalized Values**	**Cue word**
Disorder CUI	The left atrium is moderately dilated	***C0344720*** (UMLS CUI)	**4-15, 31-37** *(left atrium…dilated)*
Negation Indicator (NI)	*Denies* numbness	*no, **yes**	***0-5*** *(Denies)*
Subject Class (SC)	*Son* has schizophrenia.	*patient, family_member, donor_family_member, donor_other, null, and other	***0-2*** *(Son)*
Uncertainty Indicator (UI)	*Evaluation of* MI.	*no, **yes**	***0-9*** *(Evaluation)*
Course Class (CC)	The cough *worsened* over the next two weeks.	*unmarked, changed, increased, decreased, improved, **worsened**, and resolved	***11-18***
Severity Class (SC)	He noted a *slight* bleeding.	*unmarked, **slight**, moderate, and severe	***12-17*** *(slight)*
Conditional Class (CO)	The patient should come back if any rash occurs.	**true**, *false	***30-31*** *(if)*
Generic Class (GC)	The patient was referred to the Lupus *Clinic*.	**true**, *false	***38-43*** *(Clinic)*
Body Location (BL)	Patient has *facial* rash.	***C0015450*** (UMLS CUI)	***12-17*** *(facial)*

Bold indicates the values for the example

Default values indicated with *

Health Natural Language Processing (hNLP) Center

The Health Natural Language Processing (hNLP) Center targets a key challenge to current hNLP research and health-related human language technology development: the lack of health-related language data.

The Center's primary activities are to:

1. Provide a repository and a data curation, distribution and management point for health-related language resources
2. Support sponsored research programs and health-related language-based technology evaluations
3. Engage in collaborations with US and foreign researchers, institutions and data centers
4. Host and participate in various workshops

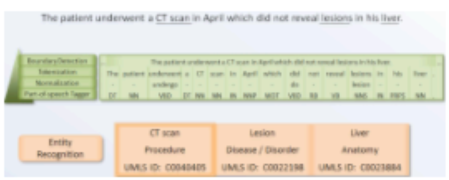
The data consists of de-identified clinical notes from several institutions. We have paid special attention to the de-identification process which included a combination of automatic and manual redacting of information.

To obtain a data set, you must be a member.

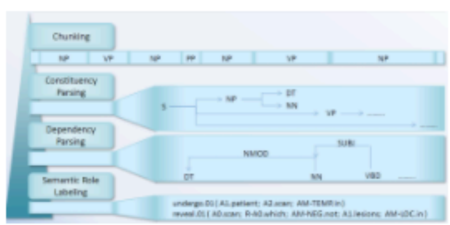
Layered Annotations

Some data sets contain layers of annotations. Click an image below to expand it.

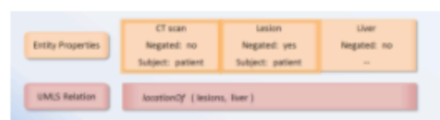
Entity Recognition



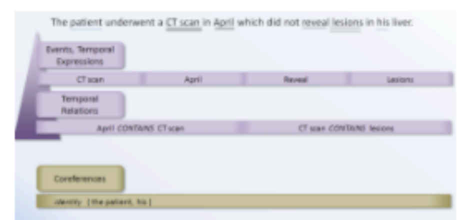
Semantic Role Labelling



Properties and Relations



Temporal



Boston Children's Hospital Until every child is well · chip · Columbia University in the City of New York · University of Colorado Boulder

For many predictive tasks, simple bag-of-words models work well

Example: Triage nurse notes

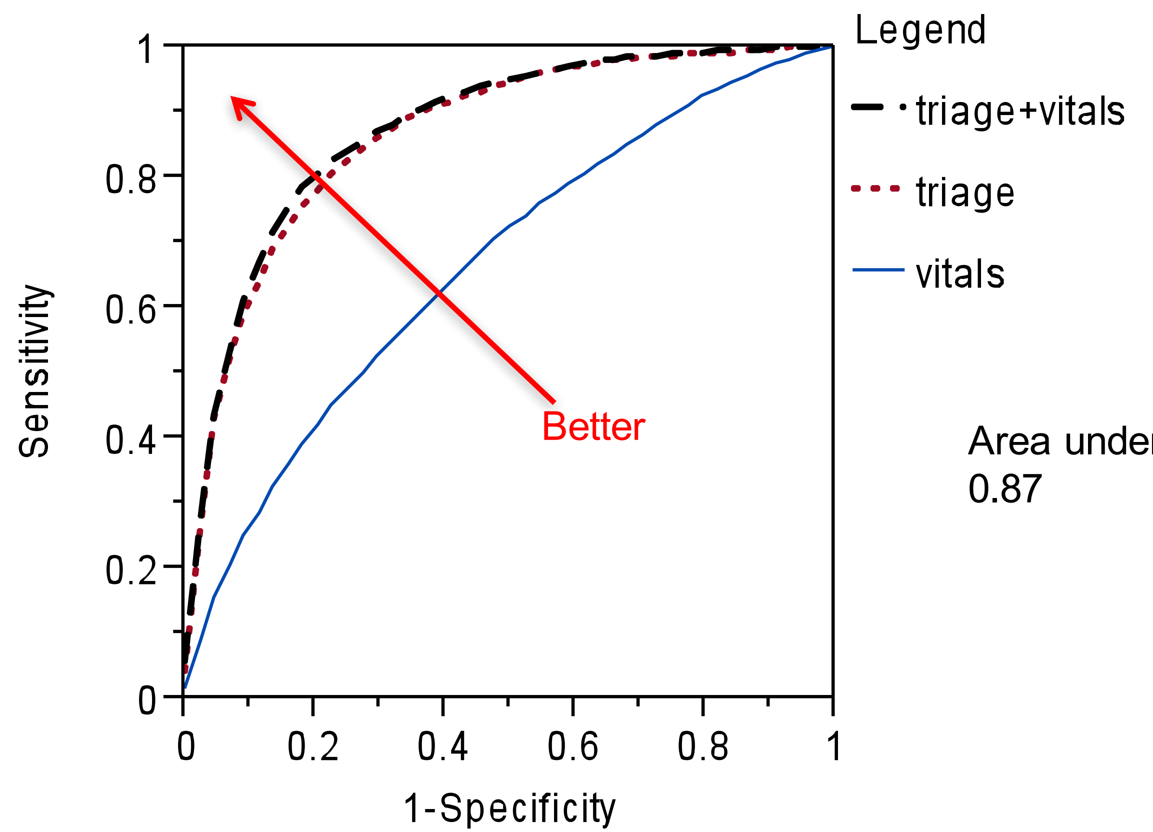
Which of these is likely to develop sepsis?

- pt with fever, chills, N/V since friday after eating what hethought was undercooked meat. Unable to hold po's down. Fevers to 103
- 89 yo f s/p esophageal hernia repair 3/09 w/ ?g-tube placement now w/ c/o's n&v. family reports pt's appetite is decreased, no BM x3d. generally not feeling well, had a bad day.
- from the scene fall of horse landed on r thigh deformity iv fluid 100 fentanyl/ morhpine 4. no head or neck pain/
- cantonese speaking with numness right arm blurred vision dizziness lack of focus SOB since8 am. tongue midline. no facial droop. same sxs as strok in 08.

Text is much more valuable than structured data

ROC curve on test data (~19,000 patients)

(Proportion of actual infection cases that are detected)



(Proportion of falsely predicted infections)

Area under the curve: 0.87

[Horng, Sontag, et al. "Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning". PLOS ONE, 2017]

Learning a representation for reasoning about a patient

- The goal of the triage note is to summarize a patient’s state to provide maximal *context* in which to understand future data
- Can we learn the latent space directly from the triage text?
- Our approach is to try to tease out this latent space using a type of dimensionality reduction
- We use a “topic” model called latent Dirichlet allocation

Latent Dirichlet allocation

- Generative model for documents (patient's triage text)
- Assume there are T topics (for us, T=500), and the variable z_i denotes the assignment of a topic to the i'th word
- Generative model for single patient's triage text:
  - $\theta \sim \mathrm{Dir}(\alpha)$ (θ is a distribution over the T topics)
  - For each word i,

$$z_i \sim \mathrm{Multinomial}(\theta) \quad \text{(choose a topic for i'th word)}$$
$$w_i \sim \Pr(w \mid z = z_i) \quad \text{(sample a word)}$$

- We learn the distributions Pr(w | z = t) and the "priors" α_t

What do we learn?

α_t	**Topic distributions**
.0013	facial numbness droop weakness sided speech slurred face…
.0004	rabies bat vaccine exposure shot here for in room prophylaxis…
.0023	shoulder pain rom arm decreased limited pulse injury …
.0237	etoh found admits unable ambulate trauma fs no on drinking…
.0068	gait unsteady steady dizziness feet ha stable alert well oriented…
.0041	vaginal discharge bleeding vag d/c gyn itching pelvic foul…
.0032	throat sore swallowing voice fevers ear difficulty st swallow…
.0027	cellulitis swelling redness with lle rle leg and fevers l lower…
.0009	pna cough on pneumonia with cxr dx recent levaquin r/o…

We discover synonyms

[Horng, Sontag, et al. "Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning". PLOS ONE, 2017]

Results make sense

Topic distributions	Less likely
facial numbness droop weakness sided speech slurred face…	
rabies bat vaccine exposure shot here for in room prophylaxis…	
shoulder pain rom arm decreased limited pulse injury …	
etoh found admits unable ambulate trauma fs no on drinking…	
gait unsteady steady dizziness feet ha stable alert well oriented…	Infection
vaginal discharge bleeding vag d/c gyn itching pelvic foul…	
throat sore swallowing voice fevers ear difficulty st swallow…	
cellulitis swelling redness with lle rle leg and fevers l lower…	
pna cough on pneumonia with cxr dx recent levaquin r/o…	More likely

[Horng, Sontag, et al. “Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning”. PLOS ONE, 2017]

Current developments in clinical NLP research

(Slide credit: Noemie Elhadad)

- Improved language models
  - Better contextual representations of what sequences of words (or characters) represent
- Improved sequence models
  - RNN (on words and characters) can capture rich, long-distance dependencies in text
- Models for mixed modalities
  - Text + images, text + laboratory tests, text +…