Lecture 16:
Introduction to Natural Language Processing (NLP)

Agenda

- Introduction
- NLP problems
  - A lot of material covered at a high level
- NLP in biomedical and clinical settings

Credit

- The material in these slides are partially based on:
  1. “An Introduction to Clinical Natural Language Processing”, Leonard D’Avolio Dina Demner-Fushman Wendy W. Chapman
  2. “Machine Learning Methods in Natural Language Processing”, Michael Collins
  3. “Biomedical and Clinical Natural Language Processing”, Ozlem Uzuner Meliha Yetsgen Amber Stubbs
  4. “Speech and Language Processing” - Jurafsky and Martin
  5. Natural Language Processing for Precision Medicine, Hoifung Poon, Chris Quirk, Kristina Toutanova, Scott Wen-tau Yih

Applications

Why NLP?

- Increasing amounts of biomedical literature
  - Extracting facts, relations, events into knowledge repositories (text mining)
  - Model organism database curation
  - Question answering (TREC Genomics track)
  - Literature based discovery

Biomedical NLP

- Increasing demands for use of EHR data
  - Phenotyping for genomic-related analysis
  - Linking evidence for Evidence-based medicine
  - Biosurveillance
  - Quality measures
- Majority of EHR data is free text!

Clinical NLP

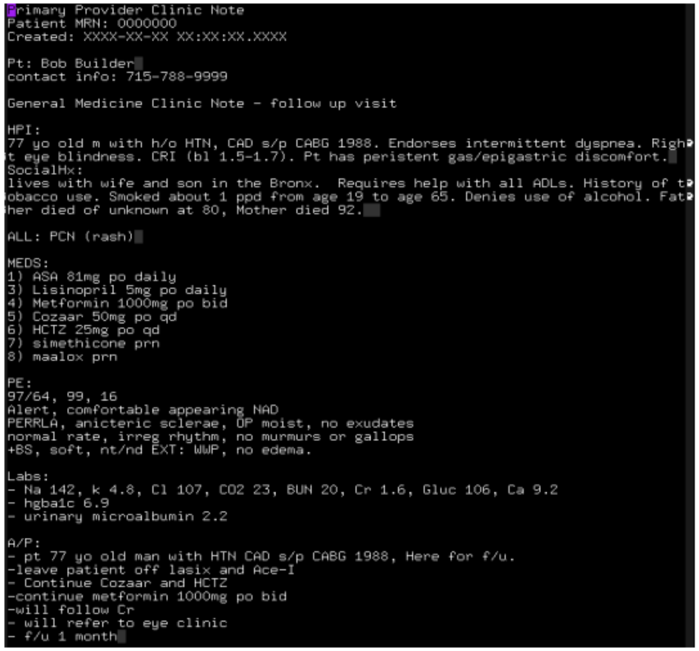
Sample ER Reports

1. **HISTORY OF PRESENT ILLNESS:** The patient is a (XX)-month-old child. Family notes that since last night, the patient was fussy and irritable, possibly pulling at her ears. No obvious pain or discomfort with urination. No chest pain, cough. No abdominal pain. No nausea, vomiting or diarrhea and is having usual stool and wet diapers. Child, however, has been much more fussy at times and now comes to the ER for further evaluation. This is the mother's first child. The child is consolable and not constantly crying in the ER and is able to be addressed and approached by me without being upset. Family notes the child has been fully immunized and is compliant with scheduled appointments.

2. **HISTORY OF PRESENT ILLNESS:** This (XX)-year-old very pleasant gentleman presents to the emergency room with a one day complaint of pain in his right ear. The patient states last evening he thought he had wax. He used a wax softener. He has had lots of drainage from his ear today. He is still having pain. He attributes it to his new hearing aid. He also is complaining of pain now in the right side of his head. There is no nausea, no vomiting, no tinnitus, visual, olfactory or auditory changes. He states the headache is behind the ear, and it is related directly to the pain in his ear. There is no chest pain, no shortness of breath, nausea, vomiting. No other complaints. He is very affable and in no apparent distress.

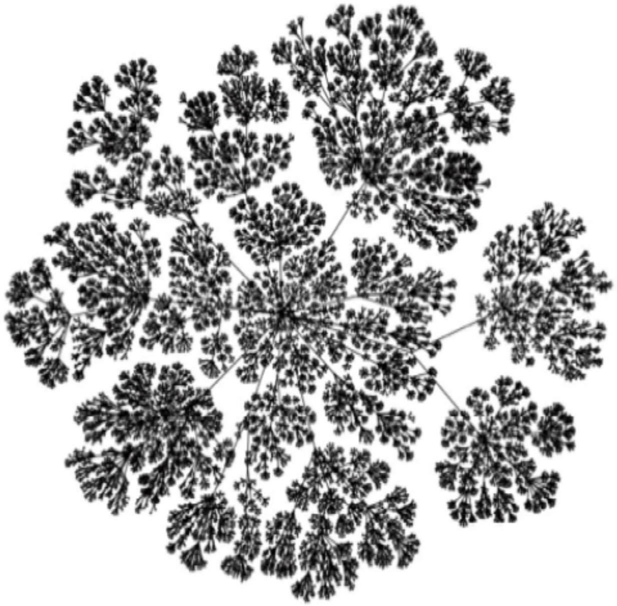
Example Applications

Application: automated coding



```
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
```

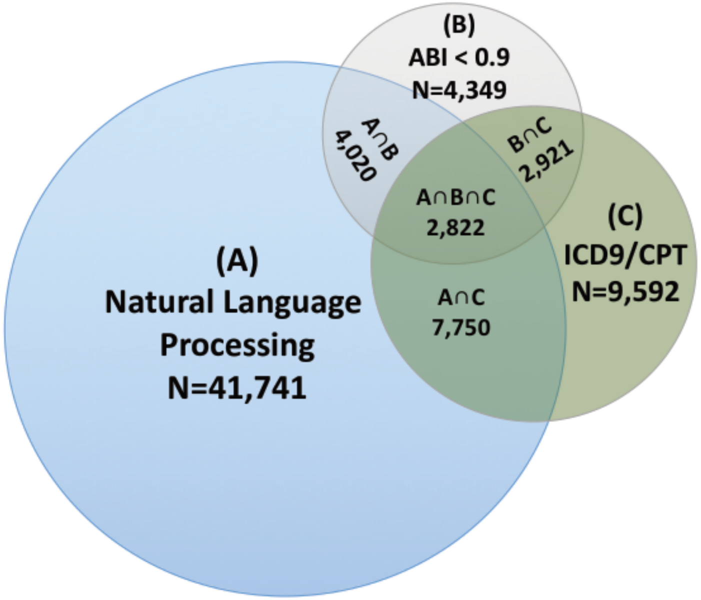
+



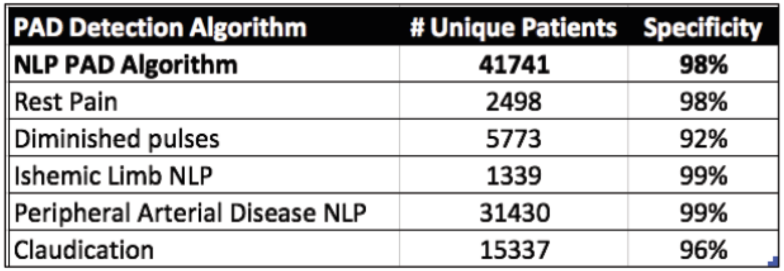
→ 401.1 (Hypertension);
428.0 (Congestive heart failure);
369.6 (One eye blindness)

Application: cohort detection

Patient cohort identification - define a set of inclusion/exclusion criteria and enable investigators to query both structured and unstructured clinical information simultaneously



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



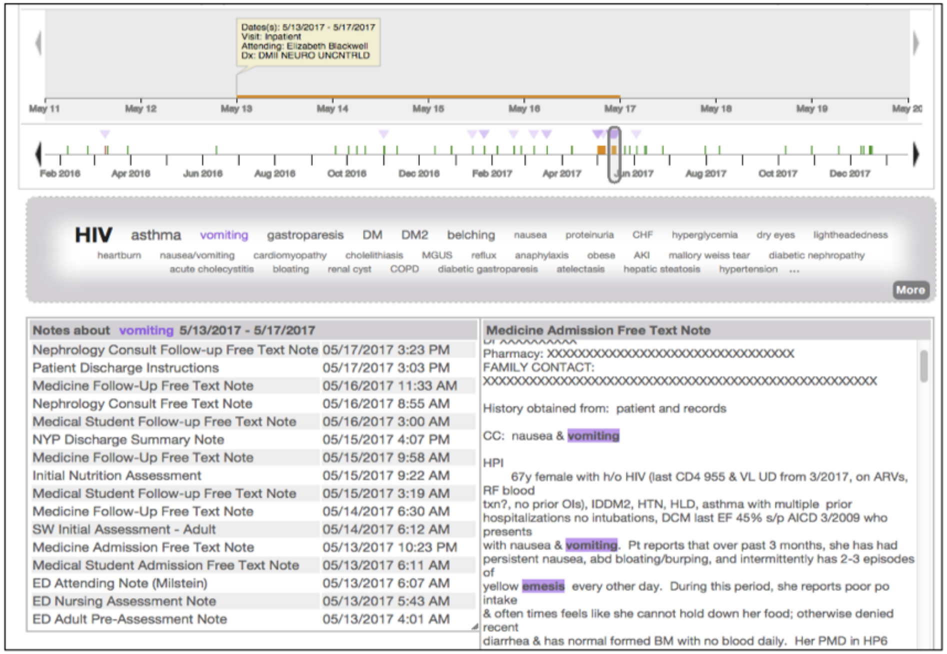
PAD Detection Algorithm	# Unique Patients	Specificity
NLP PAD Algorithm	**41741**	**98%**
Rest Pain	2498	98%
Diminished pulses	5773	92%
Ishemic Limb NLP	1339	99%
Peripheral Arterial Disease NLP	31430	99%
Claudication	15337	96%

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*.

Application: clinical decision support

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

Clinical decision support - Developing automated systems to assist decision making in clinical settings by utilizing clinical narratives, e.g. automating colonoscopy follow-up where NLP can be applied to extract necessary information for decision support from pathology and colonoscopy reports.



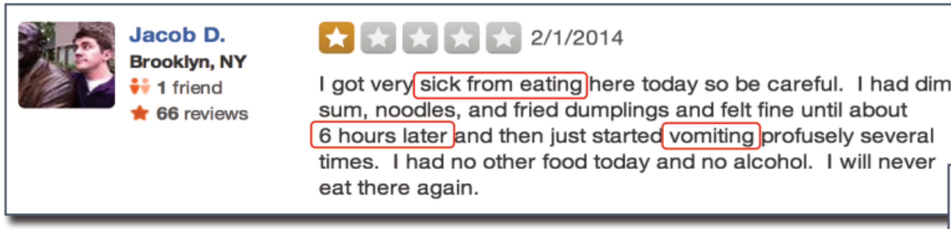
Notes about vomiting 5/13/2017 - 5/17/2017	
Nephrology Consult Follow-up Free Text Note	05/17/2017 3:23 PM
Patient Discharge Instructions	05/17/2017 3:03 PM
Medicine Follow-Up Free Text Note	05/16/2017 11:33 AM
Nephrology Consult Free Text Note	05/16/2017 8:55 AM
Medical Student Follow-up Free Text Note	05/16/2017 3:00 AM
NYP Discharge Summary Note	05/15/2017 4:07 PM
Medicine Follow-Up Free Text Note	05/15/2017 9:58 AM
Initial Nutrition Assessment	05/15/2017 9:22 AM
Medical Student Follow-up Free Text Note	05/15/2017 3:19 AM
Medicine Follow-Up Free Text Note	05/14/2017 6:30 AM
SW Initial Assessment - Adult	05/14/2017 6:12 AM
Medicine Admission Free Text Note	05/13/2017 10:23 PM
Medical Student Admission Free Text Note	05/13/2017 6:11 AM
ED Attending Note (Milstein)	05/13/2017 6:07 AM
ED Nursing Assessment Note	05/13/2017 5:43 AM
ED Adult Pre-Assessment Note	05/13/2017 4:01 AM

Medicine Admission Free Text Note

```
Pharmacy: XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
FAMILY CONTACT:
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

History obtained from:  patient and records

CC:  nausea & vomiting

HPI
     67y female with h/o HIV (last CD4 955 & VL UD from 3/2017, on ARVs,
RF blood
txn?, no prior OIs), IDDM2, HTN, HLD, asthma with multiple  prior
hospitalizations no intubations, DCM last EF 45% s/p AICD 3/2009 who
presents
with nausea & vomiting.  Pt reports that over past 3 months, she has had
persistent nausea, abd bloating/burping, and intermittently has 2-3 episodes
of
yellow emesis  every other day.  During this period, she reports poor po
intake
& often times feels like she cannot hold down her food; otherwise denied
recent
diarrhea & has normal formed BM with no blood daily.  Her PMD in HP6
```

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.

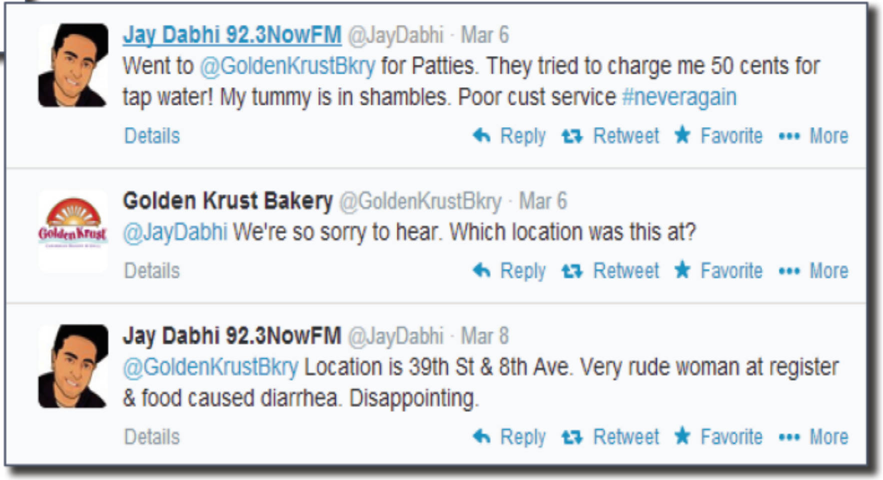
Application: info-surveillance from public social media



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

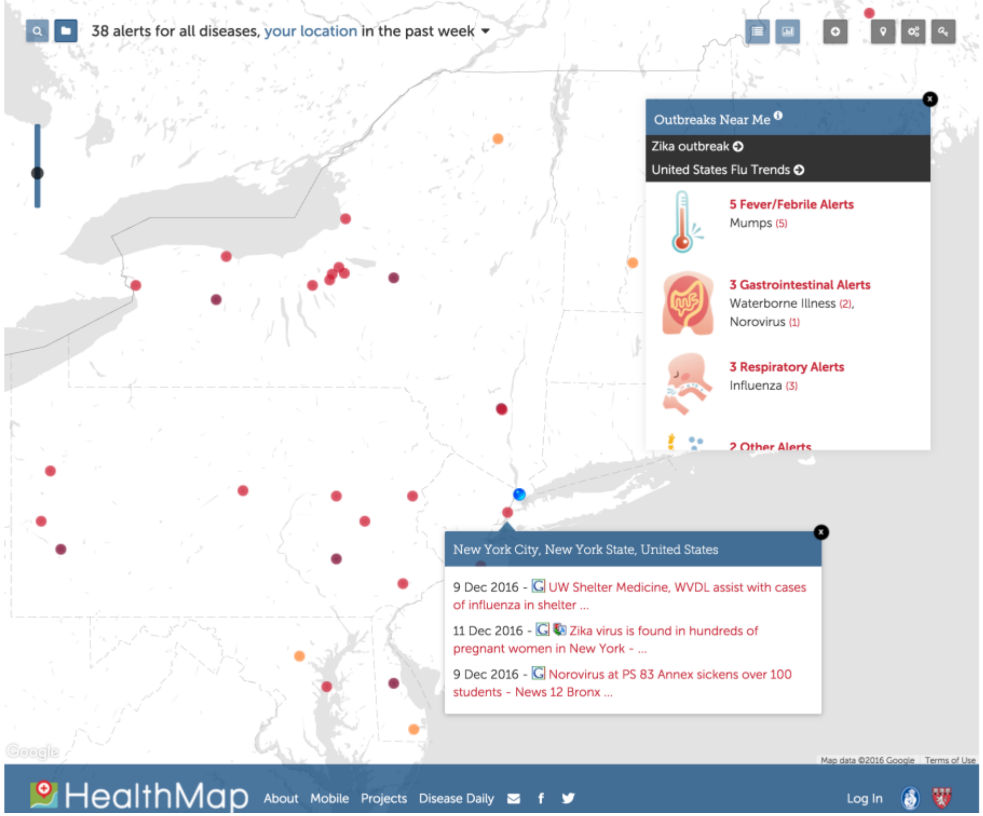
Golden Krust Bakery @GoldenKrustBkry · Mar 6
@JayDabhi We're so sorry to hear. Which location was this at?

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

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



38 alerts for all diseases, your location in the past week

Outbreaks Near Me

Zika outbreak

United States Flu Trends

5 Fever/Febrile Alerts
Mumps (5)

3 Gastrointestinal Alerts
Waterborne Illness (2), Norovirus (1)

3 Respiratory Alerts
Influenza (3)

2 Other Alerts

New York City, New York State, United States

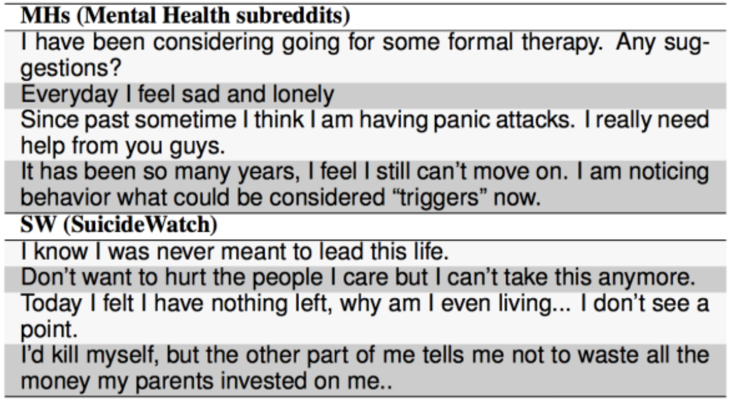
9 Dec 2016 - UW Shelter Medicine, WVDL assist with cases of influenza in shelter ...

11 Dec 2016 - Zika virus is found in hundreds of pregnant women in New York - ...

9 Dec 2016 - Norovirus at PS 83 Annex sickens over 100 students - News 12 Bronx ...

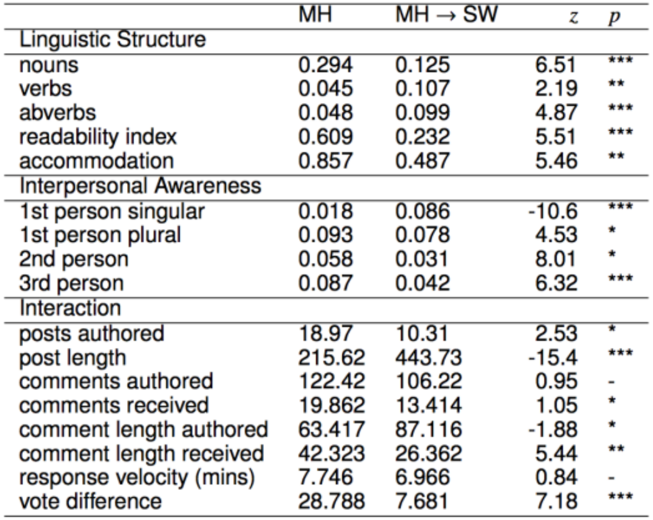
BioSurveillance - for detecting emerging infectious diseases and acts of bioterrorism. These data include the chief complaint fields from outpatient encounters and emergency department visits. Currently, biosurveillance systems retrieve chief complaint section of the encounter note and administrative codes such as ICD codes. Keyword searches look for the occurrence of such word forms as “sore throat” but may miss such related notions as “pain upon swallowing” or “throat feels raw”.

Application: predictive analytics



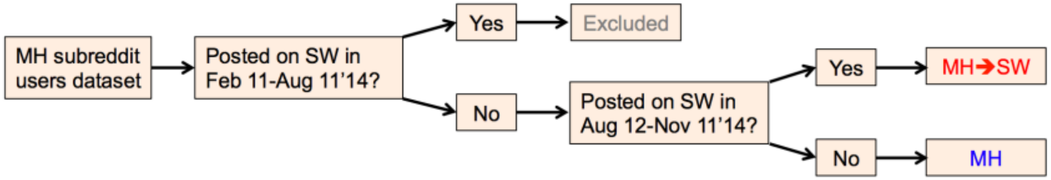
MHs (Mental Health subreddits)
I have been considering going for some formal therapy. Any suggestions?
Everyday I feel sad and lonely
Since past sometime I think I am having panic attacks. I really need help from you guys.
It has been so many years, I feel I still can't move on. I am noticing behavior 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.



	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.

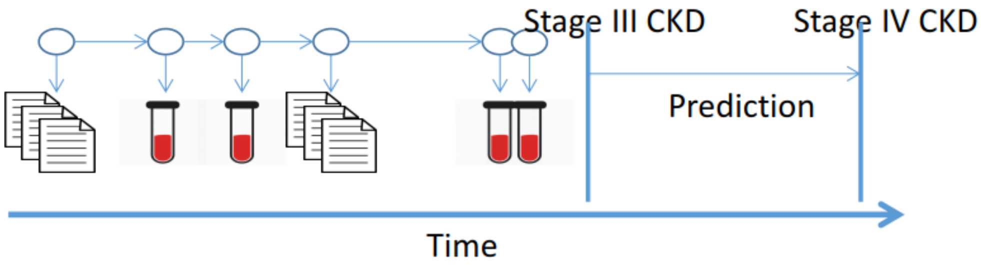


MH subreddit users dataset
Posted on SW in Feb 11-Aug 11'14?
Yes
Excluded
No
Posted on SW in Aug 12-Nov 11'14?
Yes
MH→SW
No
MH

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

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



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

Biomedical Documents

MEDLINE

PubMed

NATIONAL LIBRARY OF MEDICINE

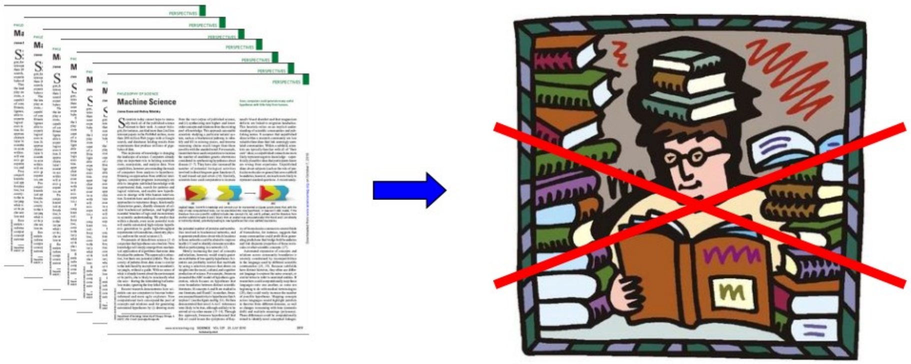
Grows in size at a dramatic pace

PubMed

27 million abstracts

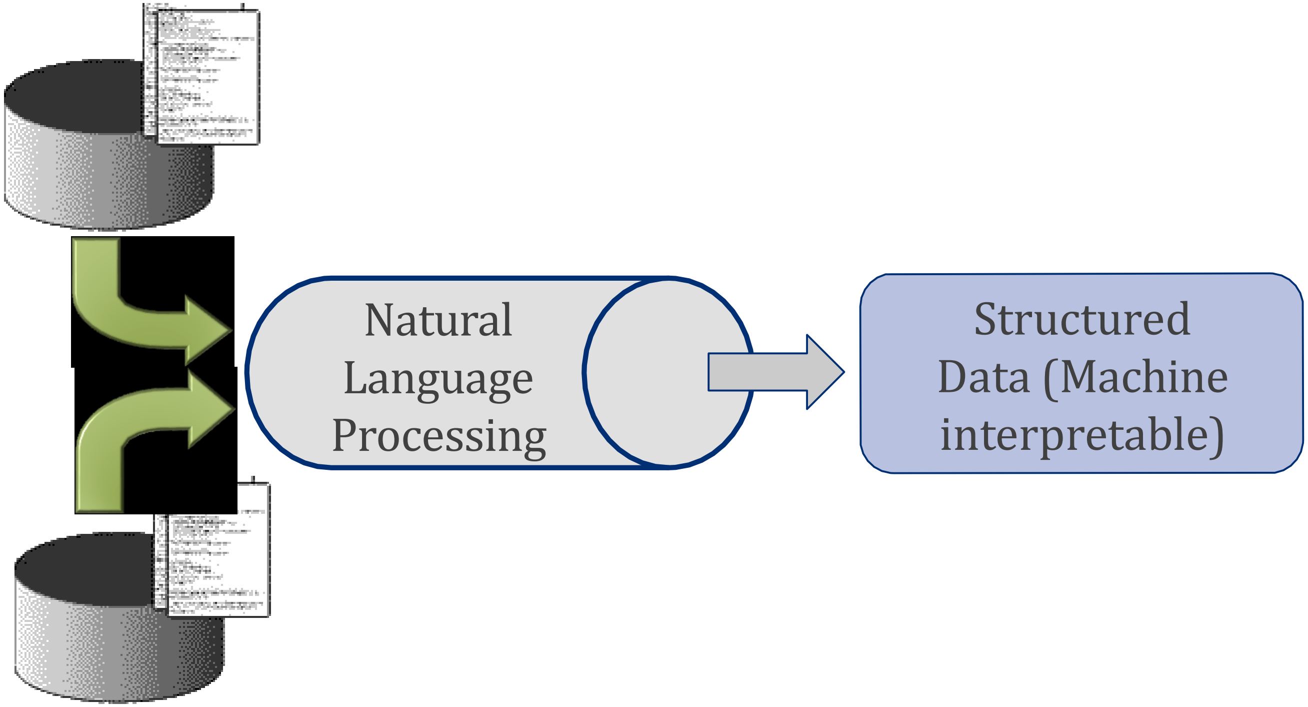
Two new abstracts every minute

Adds over one million every year



Clinical & Biomedical NLP

Clinical NLP: Electronic Medical Records



Natural Language Processing

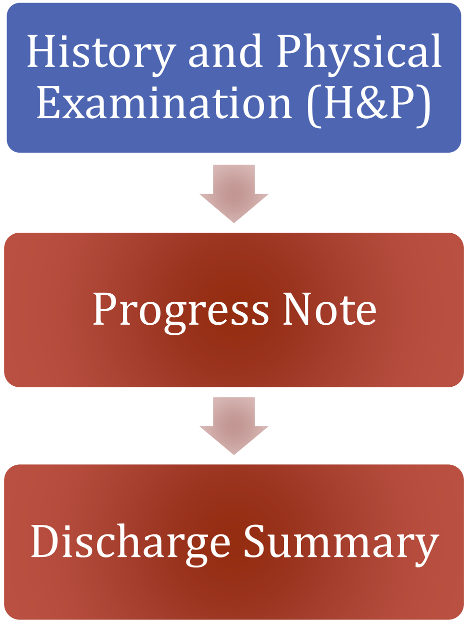
Structured Data (Machine interpretable)

- Classify
- Extract
- Summarize

BioNLP: PubMed Articles

Closer Look at Clinical Notes

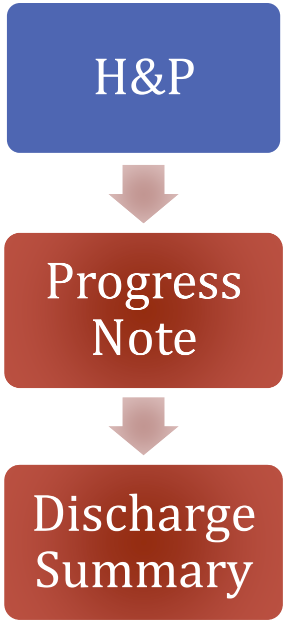
Clinical Documentation



http://clerkship.medicine.ufl.edu/files/2012/01/Complicated_admission.pdf

Clinical Documentation: H&P

- History & Physical:
  - Chief Complaint
  - History of Present Illness
  - Past Medical, Surgical, Psychiatric Histories
  - Family & Social History
  - Medications and Allergies
  - Vitals & Labs & Imaging
  - Review of Systems
  - Problem List
  - Assessment & Plan

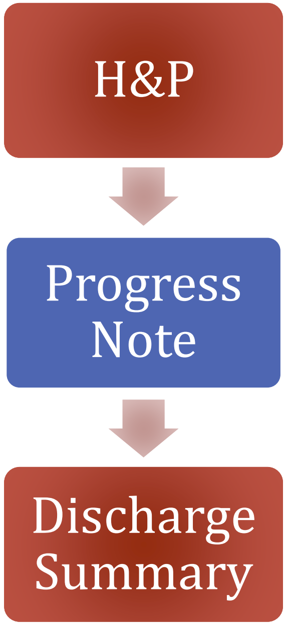


H&P

Progress Note

Discharge Summary

Clinical Documentation: Progress Note

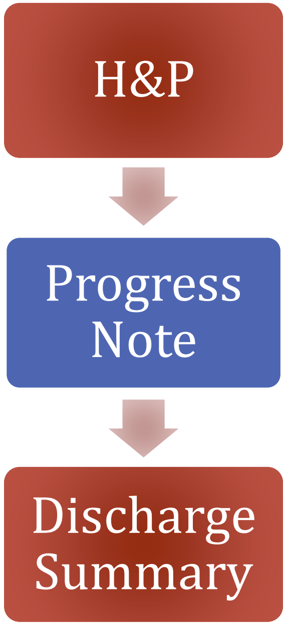

H&P

Progress Note

Discharge Summary

- SOAP
  - Subjective: Qualitative narrative
  - Objective: Vitals, Exam, Labs, Imaging, etc.
  - Assessment: What does it all suggest?
  - Plan: Next steps.

Clinical Documentation: Discharge Summary



H&P

Progress Note

Discharge Summary

Final Diagnosis:
Procedures:
History of Present Illness
Laboratory/Data Hospital Course (by PROBLEM LIST.... NOT BY DATE ---)

Discharge Medications
Discharge Instructions (diet, activity, discharged to home/nursing facility, etc)

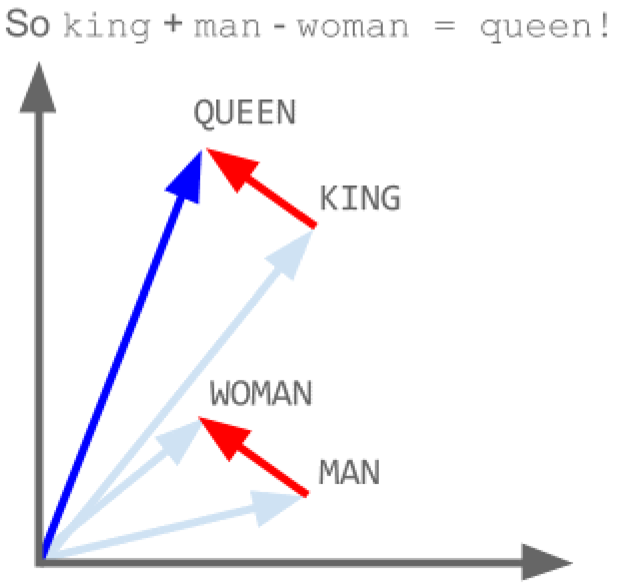
NLP Challenges

Major Elements

1. Words
2. Syntax
3. Meaning
4. Discourse

From ML to NLP

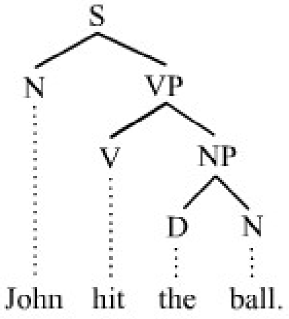
- Features from words and sentences?
  - Bag of words
  - TF-IDF
  - N-grams
  - Word vectors: word2vec, doc2vec
- Models?
  - Linear?
  - Deep learning?



http://multithreaded.stitchfix.com/blog/2015/03/11/word-is-worth-a-thousand-vectors

Common NLP Tasks

- Co-reference resolution
  - Given a sentence, determine which words refer to the same objects: He entered John's house through the front door
- Named entity recognition (NER)
  - Given a sentence, determine which items map to proper names, such as people or places. Mr. Jones, Epilepsy
- Part-of-Speech tagging (POS)
  - Given a sentence, determine the part of speech for each word. Discharge (verb/Noun)
- Parsing
  - Determine the parse tree (grammatical analysis) of a given sentence.



Common NLP Tasks

- Relationship extraction
  - Given a sentence, identify the relationships among named entities (e.g. what is inhibiting expression of what).
- Word sense disambiguation
  - Many words have more than one meaning (e.g. Discharge)
- Speech processing
  - This covers speech recognition, text-to-speech and related tasks.
- Discourse Parsing
  - Identifying the discourse structure of connected text (e.g. elaboration, contrast), or classifying the speech acts (e.g. assertion, question).

Clinical vs. Biomedical NLP

- Clinical NLP
  - Task: Focus on documentation related to patient care
  - Data source: Electronic Health Records (EHR)
    - Both inpatient and outpatient documentation
- Biomedical NLP
  - Task: Focus on scientific discoveries about biology, physiology, and medicine
  - Data source: Journal articles, clinical trials, webpages, …

Example: Classification

Clinical Examples

- Classify
- Extract
- Summarize

Classify a chief complaint into a syndrome category

“coughing” to Respiratory

Biomedical Examples

- Classify
- Extract
- Summarize

Triage of articles likely to have experimental evidence

Find evidence to assign top-level GO (Gene Ontology) terms

Example: Extraction

Clinical Examples

- Classify
- Extract
- Summarize

# of lymph nodes removed during colorectal cancer surgery

Biomedical Examples

- Classify
- Extract
- Summarize

#Extract bio-molecular events
What [ANTIBODIES] have been used to detect protein TLR4?

Example: Summarization

Clinical Examples

- Classify
- Extract
- Summarize

From a H&P note, list chronic condition

Summarize family history of prostate cancer

Biomedical Examples

- Classify
- Extract
- Summarize

Summarize full text documents

Gene Reference into function (GeneRef)

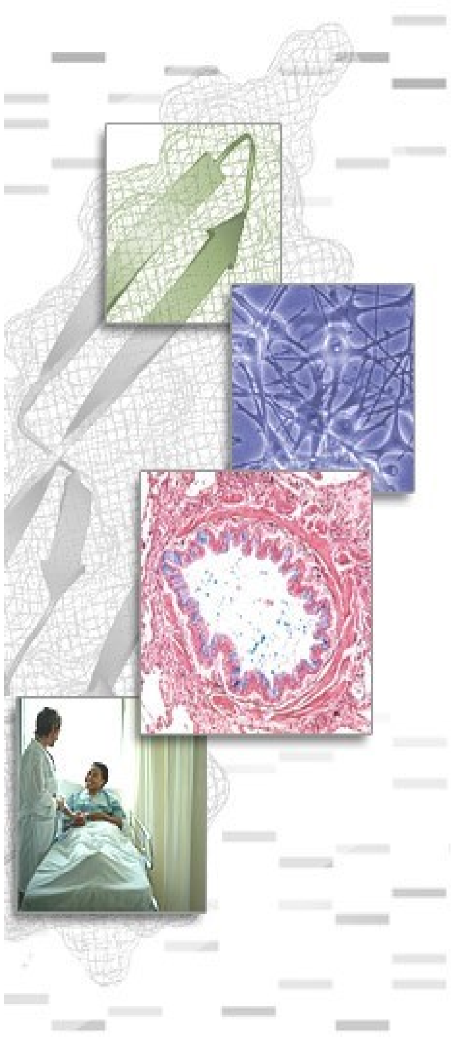
Biomedical Language

- Contains domain specific rich and evolving vocabulary
- Journal abstracts and articles usually follow similar section structure
- Sentences are very grammatical but include highly ambiguous terms
  - Neurofibromatosis 2 [disease]
  - Neurofibromin 2 [protein]
  - Neurofibromatosis 2 gene [gene]

Clinical Language

- Domain-specific, jargon, idioms
- Telegraphic, with misspellings, incomplete sentences
- Speculations, hypotheses, and negations
- Some structure

Sample Clinical Documents

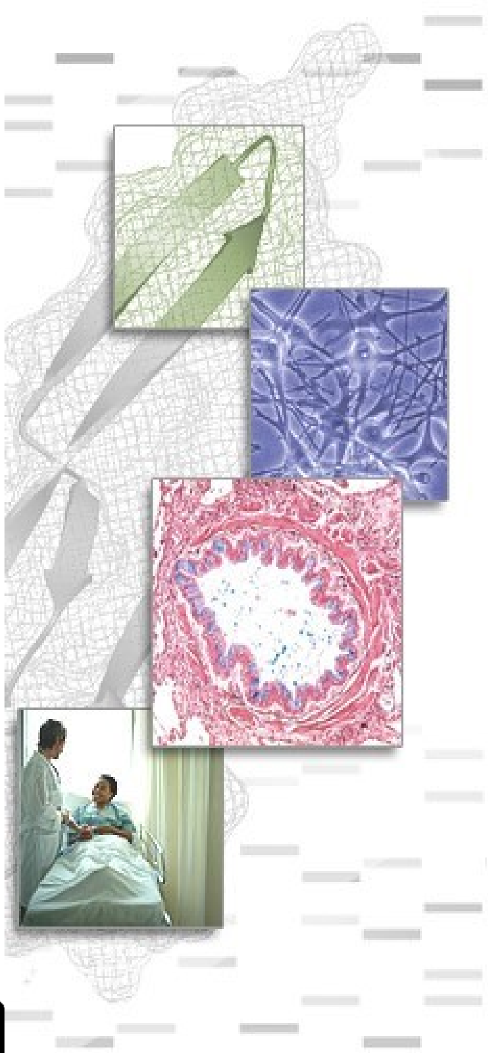


HISTORY OF PRESENT ILLNESS: Mrs. [**Hun]** is a 77 year-woman with long standing **hypertension** who presented as a Walk-in to me at the [Bronx] Health Center on September 15. **Recently** had been started **q.o.d**. on Clonidine since August 2 to taper off of the drug. **Was** told to start Zestril 20 **mg. q.d**. again. The patient was sent to the Emergency Unit for direct admission for cardioversion and anticoagulation, with the Cardiologist, Dr. [Swasissz] to follow.

SOCIAL HISTORY: **Lives** alone, **has** one daughter living in [Spring]. **Is** a non-smoker, and **does** not drink alcohol.

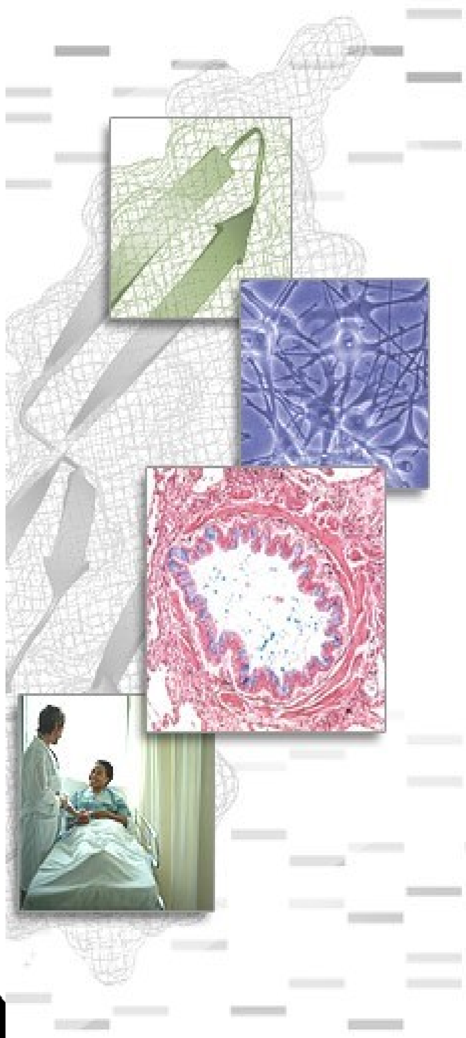
HOSPITAL COURSE AND TREATMENT: During admission, the patient was seen by Cardiology, Dr. [Tylenol], was started on **IV** Heparin, Sotalol 40 **mg PO b.i.d**. increased to 80 **mg b.i.d**., and had an echocardiogram. By the next day the patient had better rate control and blood pressure control but remained in atrial fibrillation. On Saturday the patient was felt to be medically stable…

Sample Clinical Documents



The patient is a 46 year old woman with a history **of Q wave myocardial infarc6on** with right ventricular **infarct** in October 1992. Peak **CK's** were 2300. Catheterization showed 100% **RCA** lesion which was treated with angioplasty **reduced** to 20-30% stenosis. Subsequent catheterization October 92 , July 92 and September 92 for atypical chest pain , showed clean coronaries. **Exercise tread mill test in September 92 ,** the patient went three minutes and 31 seconds with standard Bruce protocol and stopped secondary to atypical chest pain**. Maximum heart rate 162 , blood pressure 176/90 , no ST or T wave changes**. In April 92 she ruled out for myocardial infarction **by enzymes** and EKG , after presenting with prolonged chest pain. **VQ scan was low probability**. Chest CT ruled out aortic dissection. The patient now presents to the hospital with 24 hours of right sided chest pain , stating that it was squeezing in her right breast , **felt to be between the shoulder blades**. She complained of shortness of breath , dizziness , weakness and nausea , **no palpitations were noted**…

Sample Clinical Documents



Pt recently hospitalized 7/19/06 for **chf** exacerbation (diastolic dysfunction) **2nd** to dietary and medicine noncompliance (salty foods , stopped her **HCTZ**) and continued to smoke. **Pt diuresed** and sent home on new **lasix 60qam 40qpm regimen**. **Pt** noticed steady decline in functional status during the last 3 weeks because of **SOB. at** baseline should **sat** 85% on **ra** , 95% on **6L02NC** at rest and ambulation. (on home **o2**) but now , can't ambulate , **sating** 83-89% on 6l at rest. **also** notes **pnd** , orthopnea. **Pt** notes intermittent chest pain on and off lasting 5 minutes not associated with exertion or any other cardiac **sx**. 8/15 dobuta mibi-> ischemia in d1 territory. **11/19 :echo-->ef 60% , Pa pressure 48 + RA. no valve dz. rv enlarged and hypokinetic. A/P: pump: decompesated CHF (diastolic dysfxn , ? cor pulmonale component) 2nd to diet/med non-compliance. uptitrate captopril , continue iv lasix 60 qd with goal net neg 2 liters , daily weights , strict Iand O. check cxray. Switched to po lasix 10/06 , back to lisinopril for d/c Fri. ischemia: has + mibi in past , but no further workup to d1 lesion. can't get ecasa 2nd to vWD. continue BB , will hold off on statin since not hyperlipidemic. rate:tele.** …

Why is NLP Difficult?

Named entity recognition

Linguistic variation
Polysemy Finding
validation
Implication

Contextual attribute assignment

Negation
Uncertainty
Temporality

Discourse processing

Report structure
Coreference

Linguistic Variation

Different Words with the Same Meaning

Derivation

mediastinal = mediastinum

Inflection

opacity = opacities; cough = coughed

Synonymy

Addison's Disease: Addison melanoderma, adrenal insufficiency, adrenocortical insufficiency, asthenia pigemntosa, bronzed disease, melasma addisonii, …

Chest wall tenderness: chest wall did demonstrate some slight tenderness when the patient had pressure applied to the right side of the thoracic cage

Polysemy

One Word With Multiple Meanings

General polysemy

Patient was prescribed codeine upon discharge

The discharge was yellow and purulent

Acronyms and Abbreviations

APC: activated protein c, adenomatosis polyposis coli, adenomatous polyposis coli, antigen presenting cell, aerobic plate count, advanced pancreatic cancer, age period cohort, alfalfa protein concentrated, allophycocyanin, anaphase promoting complex, anoxic preconditioning, anterior piriform cortex, antibody producing cells, atrial premature complex, …

Negation

Approximately half of all clinical concepts in dictated reports are negated*

Explicit negation

“The mediastinum is not widened”

Mediastinal widening: absent

Implied absence without negation

“Lungs are clear upon auscultation”

Rales/crackles:

absent

Rhonchi: absent

Wheezing: absent

*Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. Evaluation of negation phrases in narrative clinical reports. Proc AMIA Sym. 2001:1059.

Uncertainty

Unsure

treated for a presumptive sinusitis

Reasoning

It was felt that the patient probably had a cerebrovascular accident involving the left side of the brain. Other differentials entertained were perhaps seizure and the patient being postictal when he was found, although this consideration is less likely.

Reason for exam

R/O out pneumonia.

Temporality

Clinical reports tell a story

Past medical history

History of CHF presenting with shortness of left-sided chest pain.

Hypothetical or non-specific mentions

He should return for fever or increased shortness of breath.

Temporal course of disease

Patient presents with chest pain … After administration of nitroglycerin, the chest pain resolved.

Finding Validation

Mention of a finding in the text does not guarantee the patient has the finding

She received her influenza vaccine

His temperature was taken in the ED

Some findings require values

Fever

Temperature 38.5C

Oxygen desaturation

Oxygen saturation low

Oxygen saturation 85% on room air

Implication

Audience for patient reports is physicians

- Pneumonia not mentioned in many positive reports

- Sentence level inference
  - “There were hazy opacities in the lower lobes” -> Localized infiltrate
- Report level inference
  - Localized infiltrates -> Probable pneumonia

Report Structure

- Anatomic Location sometimes in section header

  NECK: no adenopathy.

- Some sections carry more weight

  IMPRESSION: atelectasis

- Some reports contain pasted text difficult to process

  Cardiovascular: [ ] Angina [ ] MI [x] HTN [ ] CHF [ ] PVD [ ] DVT [ ] Arrhythmias [ ] Previous PTCA [ ] Previous Cardiac Surgery [ ] Negative -Denies CV problems

Coreference

Chest x-ray again shows a well-circumscribed nodule located in the left upper lobe. The tumor has increased in size since the last exam with a diameter of approximately 2 cm.

- How big is the nodule?
- Has the nodule increased in size?
- Where is the tumor?

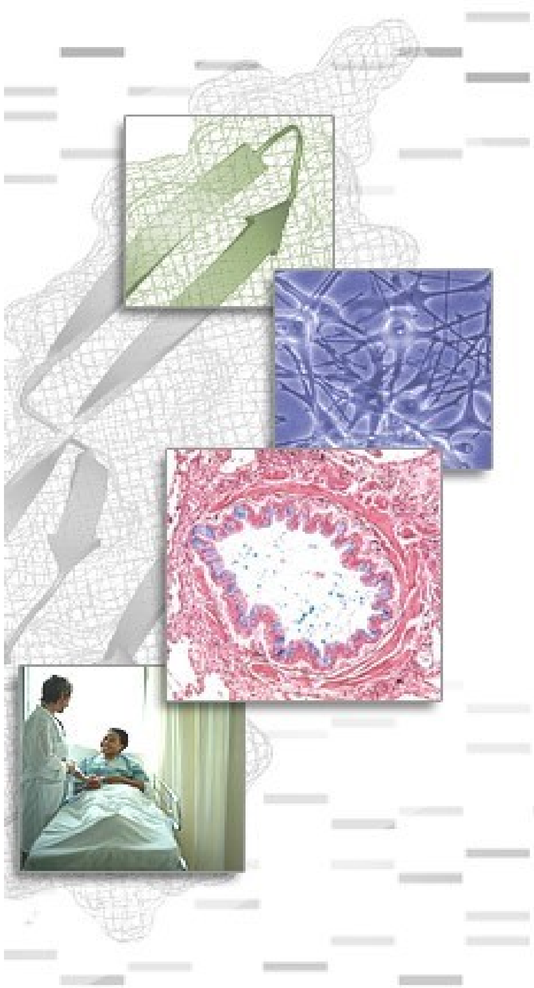
Sample Clinical NLP Application

- Does the patient drink?
  - Classify patient into 3 classes: Heavy consumption, Moderate consumption, None
- Application:
  - Retrospective cohort study as ICU mortality predictor
- Structured data:
  - ICD9 codes of alcohol-based illness, e.g., delirium tremens (291*)

Sample Clinical Documents

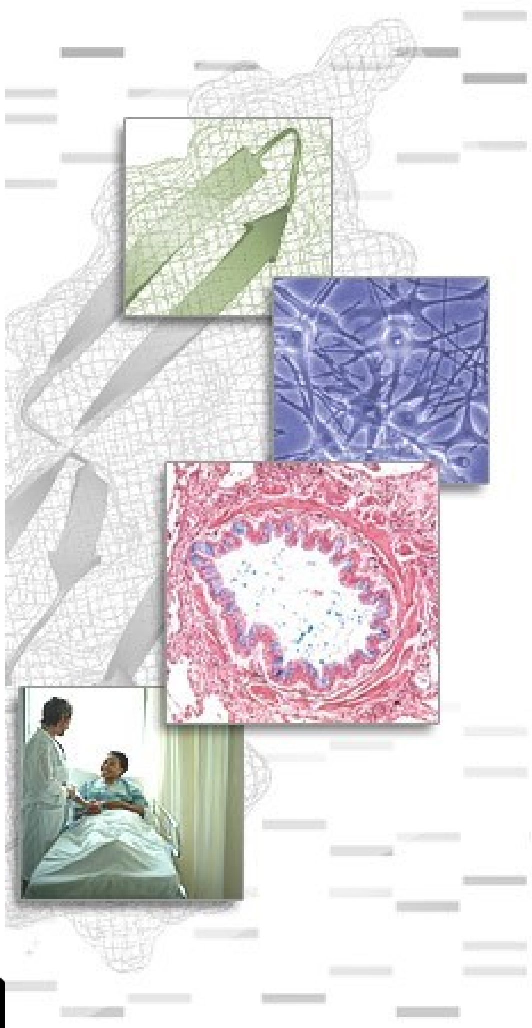
- Search for alcohol OR drink OR…
- Retrieval results:
  - Yes: The patient is known to have a history of alcohol abuse
  - Maybe: tox screen only significant for alcohol level of 273
  - No: She denied any intravenous drug use, tobacco, or alcohol
  - Maybe: He was counseled against the use of alcohol
  - Maybe: He stated that he would not drink alcohol in the future
  - No: Major Surgical or Invasive Procedure: Alcohol septal ablation.
  - Yes: Quit smoking 'many' yrs ago. smoked 1 pp week. used to drink 10--12 hard drinks every other day. last drink before goig to OSH. resolved to quit now. no IVDU.
  - No: her husband is a heavy drinker and often isolates himself to drink alone
  - No: the patient was seen to drink a tremendous amount of water

Clinical NLP Tasks



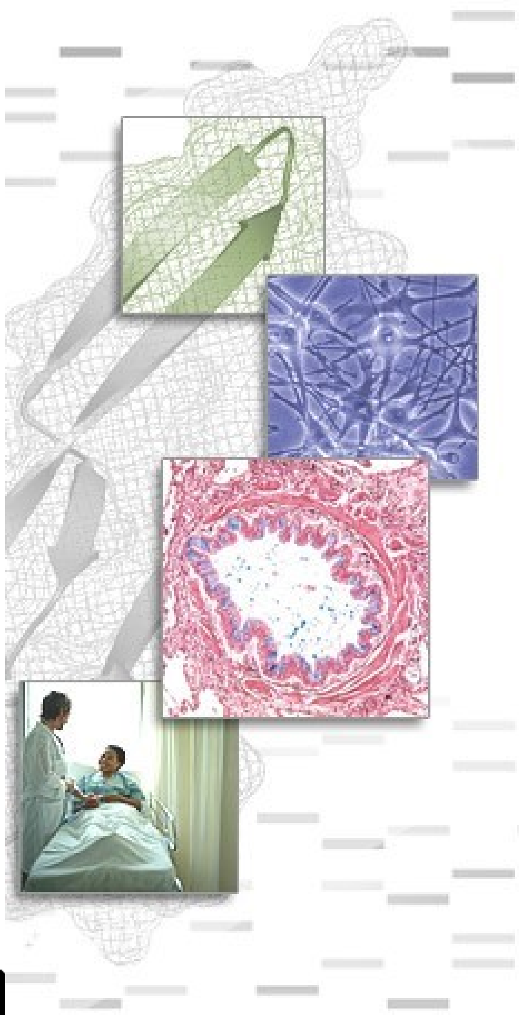
- Does the patient drink alcohol or not?
  - Negation detection
- Is alcohol use asserted, equivocal, modal or hypothetical?
  - Uncertainty
- Who is drinking?
  - Experiencer, relation extraction
- What is the patient drinking?
  - Relation, event extraction
- Is the patient drinking currently and regularly?
  - Temporal relations, timeline generation

More Clinical NLP Tasks



- Low-level tasks
  - Sentence boundary detection
  - Part of speech tagging
  - Shallow parsing (chunking)
- High-level Tasks
  - Spelling/grammatical error identification and recovery
  - Named entity recognition and information extraction
  - Word sense disambiguation

Challenges in Clinical NLP Tasks



- Sentence boundary detection: complicated by
  - abbreviations and titles, e.g., m.g., Dr.
  - lists and templates, e.g., MI [x], SOB[]
- Tokenization: complicated by
  - characters typically used as token boundaries, e.g.,
  - 10 mg/day, N--acetylcysteine.
- Morphological decomposition: complicated by
  - compound words, e.g., nasogastric
- Text segmentation: complicated by
  - problem–specific needs, e.g., sections, including Chief Complaint, Past Medical History, HEENT, etc.

- In general, systems developed for non-clinical text often work less well on clinical narratives.

Biomedical and Clinical Corpora

- There are various biomedical corpora annotated for syntax and semantics
  - MedTag: A collection of biomedical annotations (MEDLINE abstracts): the AbGene corpus of annotated sentences of genes and protein named entities, the MedPost corpus of part of speech tagged sentences and the GENETAG corpus for named entity identification used for BioCreAtIvE I.
  - TREC Genomics Track: A set of data collecions provided by TREC Genomics Track useful for development and evaluation of retrieval and text categorization strategies in the biomedical domain.
  - BioCreative corpus: Dataset produced by the BioCreative assessment, text passages relevant for GO annotations of human proteins.
  - GENIA corpus: Annotated corpus of literature related to the MeSH terms: Human, Blood Cells, and Transcription Factors.
  - Yapex corpus: Training and test data for the protein tagger (NER) YAPEX.
  - PASBio: Predicate-argument structures of biomedical literature.
  - LLLti5 dataset: Genic Interaction Extraction Challenge: protein/gene interactions IE data set
  - IEPA corpus: The Interaction Extraction Performance Assessment corpus
  - BioText Data: Dataset for extraction of disease/treatment entities relations
  - BioText NC Semantics Dataset: Dataset of Noun Compound Semantics used in experiments described in articles
  - PennBioIE: UPenn Biomedical Information Extraction datasets of annotated PubMed abstracts: CYP45ti domain and oncology domain
  - Medstract corpus: Biomedical annotation corpus useful for acronym definition and coreference resolution
  - Medstract corpus: Biomedical annotation corpus useful for acronym definition and coreference resolution
  - OHSUMED text collection: Document collection used for the TREC-9 contest.
  - BMC corpus: Open access corpus of full text articles provided by BioMed Central.
  - FetchProt corpus: Full text journal articles from the biological domain analyzed for experiments on proteins.
  - PDG Bio-sentence splif er corpus: Small collection of text data sets derived from PubMed abstracts to develop and assess sentence spliV n g tools.
  - Bio1 corpus: annotated corpus, same field as GENIA, but annotated to small top-level ontology.
  - …