

# Improving Transportability of NLP Solutions using CLAMP

Hua Xu, PhD

School of Biomedical Informatics,  
University of Texas Health Science Center at Houston



UTHealth™  
The University of Texas  
Health Science Center at Houston

School of Biomedical  
Informatics

# The Transportability Problem of NLP

- Many NLP systems have been developed and shown good performance in different tasks
- However, performance drop when transporting these NLP tools
  - From one type of clinical notes to another
  - From one institute to another
  - From one application to another



UTHealth™  
The University of Texas  
Health Science Center at Houston

School of Biomedical  
Informatics

# An example of smoking status detection

- Mayo Clinical NLP System for smoking detection
  - I2b2 dataset, F-measure 85.5% (*Savova et al. JAMIA 2008*)
  - Vanderbilt dataset, F-measure 75% (*Liu et al. AMIA 2012*)
- Customize it to achieve an F-measure of 89%
  - Sentence boundaries
  - New lexicons (e.g., “Tob neg”)
  - Re-training of SVM classifiers by annotating local data
  - Add new rules





# What is CLAMP - Clinical Language Annotation, Modeling, and Processing?

- A general purpose clinical NLP system – “**CLAMP CMD**”
  - Built on proven methods
  - Good performance, high speed
- An IDE (integrated development environment) for building customized clinical NLP pipelines via GUIs – “**CLAMP GUI**”
  - Annotating/analyzing clinical text
  - Training of ML-based modules
  - Specifying rules
- An enterprise solution for NLP needs in healthcare organizations
  - “**CLAMP Enterprise**”
    - Task management
    - Visual analytics

# CLAMP CMD – built on proven methods

NLP Tasks		Ranking
Named entity recognition	2009 i2b2, medication	#2
	2010 i2b2 problem, treatment, test	#2
	2013 SHARe/CLEF abbreviation	#1
UMLS encoding	2014 SemEval, disorder	#1
Relation extraction	2012 i2b2 Temporal	#1
	2015 SemEval Disease-modifier	#1
	2015 BioCREATIVE Chemical-induced disease	#1



# CLAMP CMD – performance

- Extract problems, treatments, and tests

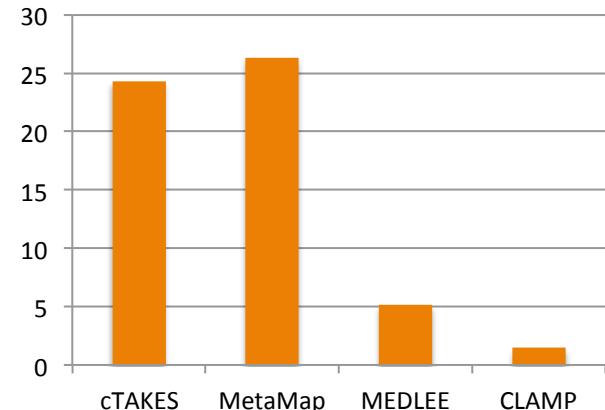
Corpus	Entity types	# entity	Exact match			Relaxed match		
			P	R	F1	P	R	F1
MTsamples	treatment, problem, test	25,531	0.841	0.811	0.826	0.921	0.890	<b>0.905</b>
i2b2	treatment, problem, test	72,846	0.891	0.861	0.876	0.958	0.925	<b>0.941</b>
UTNotes	treatment, problem, test	124,869	0.921	0.900	0.910	0.963	0.940	<b>0.951</b>
SemEval 2014	Disease_Disorder	10,077	0.861	0.791	0.824	0.870	0.799	<b>0.833</b>
SemEval 2015	Disease_Disorder	17,333	0.867	0.816	0.840	0.886	0.834	<b>0.859</b>

# CLAMP CMD - speed

- Thread-safe

Pipeline	MAC		Linux		Windows	
	Single thread	Multi threads	Single thread	Multi threads	Single thread	Multi threads
clamp-ner	0.72	0.28	1.25	0.17	0.69	0.302
clamp-ner-attribute	0.93	0.38	1.59	0.24	0.90	0.422
disease-attribute	0.62	0.26	1.06	0.17	0.58	0.296
lab-attribute	0.62	0.26	0.99	0.16	0.56	0.286
medication-attribute	0.67	0.29	1.15	0.18	0.61	0.3

Test Data Mimic2 Data set (500 documents) Number of multi-threads 10



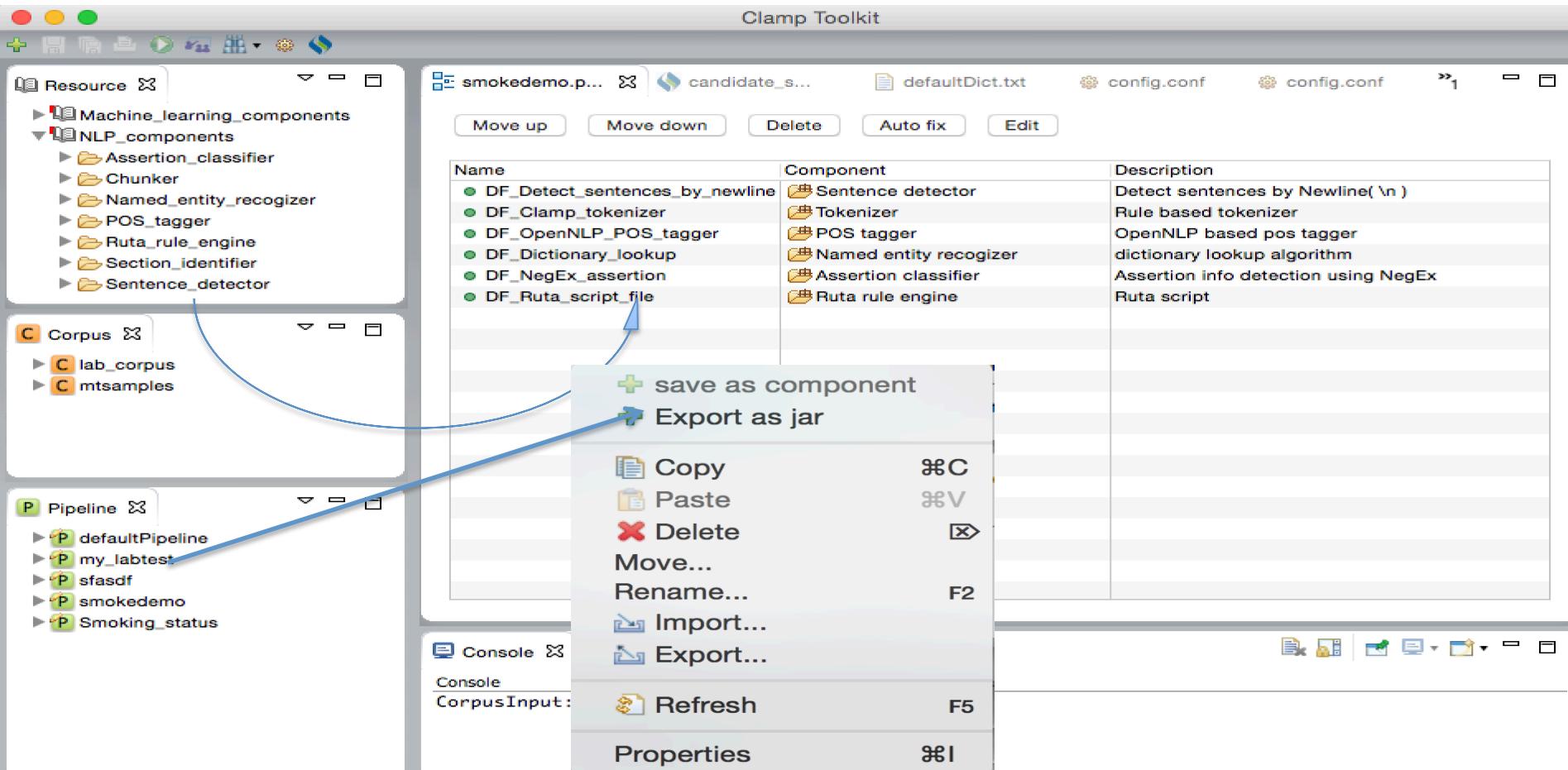
SemEval 2015 Corpus: 431  
documents, avg doc size: 9.38k  
SINGLE THREAD

# CLAMP GUI – two use cases

- Build a rule-based system to extract smoking status in clinical text and classify them into three categories: current smoker, past smoker, and non-smoker
- Build a hybrid (machine learning + rules) system for extracting lab tests and associated values from clinical text
- Videos: <http://clamp.uth.edu/tutorial.php>



# Building your own pipeline



# Annotating/Re-training

Clamp Toolkit

0005.xmi

27 The patient is an 80 year old female with **breast cancer**,  
status post lumpectomy / radiation therapy / Tamoxifen (2000 ), hypertension, hyperlipidemia,  
multiple urinary tract infections who presents with a four day prodrome of dry cough, rhinorrhea, coryza, malaise, chills, headache, decreased p.o. intake,

xmi token ori

Outline

Semantic

- Entity
  - problem
  - test
  - treatment
- Relation
- Syntax

P PipelineView

P newpipeline

Console

INFO: load from file, filename=[L/Clc]

Progress

Train project i2b2corpus NER Training, Fold 2

Extracting features: Training NER model...

The image shows the Clamp Toolkit software interface. The main window displays a clinical text document (0005.xmi) with various entities highlighted in green boxes and labeled with blue 'predict problem' tags. The left sidebar shows a file tree with 'i2b2corpus' and 'models' sections, and a 'PipelineView' tab. The bottom tabs show 'xmi', 'token', and 'ori'. The right sidebar contains an 'Outline' panel with a 'Semantic' section and a checklist for Entity, Relation, and Syntax categories. The 'Entity' category is expanded, showing 'problem', 'test', and 'treatment' sub-options, all of which are checked. The 'Relation' and 'Syntax' categories are also checked. The 'Console' and 'Progress' panels at the bottom provide system logs and training status information.

# Specifying rules

Clamp Toolkit

NLP... TEST.pipeline default.ruta 0001.xmi Outline

```
TYPESYSTEM ClampTypeSystem;
//Auto generated by rule editor

BLOCK(ForEach) Sentence{FEATURE("segmentId", "medications")}{
    BaseToken{ REGEXP("Tamsulosin") -> UNMARK(ClampNameEntityUIMA, true),
               CREATE( ClampNameEntityUIMA, 1,1,"semanticTag" = "treatment")};
}
```

test  
81 1. Tamsulosin 0.4 mg Capsule , Sust . Release 24HR Sig : One ( 1 )  
Capsule . Sust . Release 24HR PO HS ( at bedtime ).

P PipelineView TEST Components  
Name entity recogni...  
Pos tagger  
script  
default.ruta  
Section header iden...  
Sentence detector  
Tokenizer  
TEST.pipeline  
Data Feature Input 0001.txt

Please specify the rule:

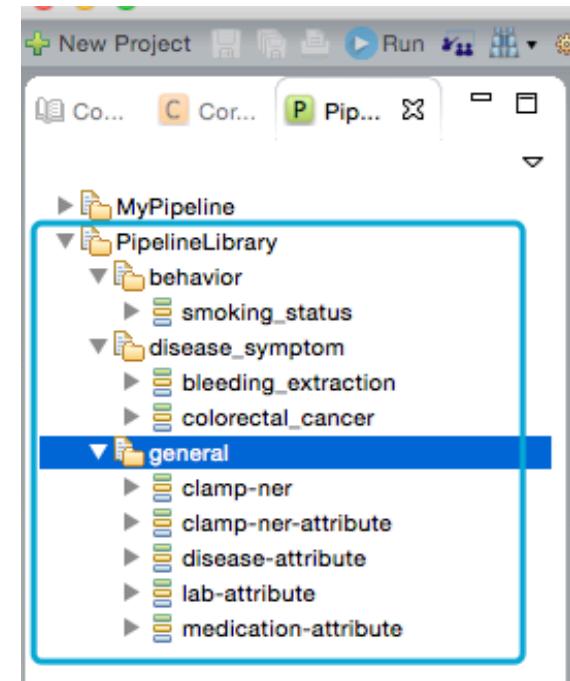
IF  
CONDITION Token [TYPE] [START OFFSET] [END OFFSET] [OPERATOR] [VALUE]  
AND Section 0 0 = medications

THEN  
ASSIGN Tamsulosin TO treatment

OK Cancel

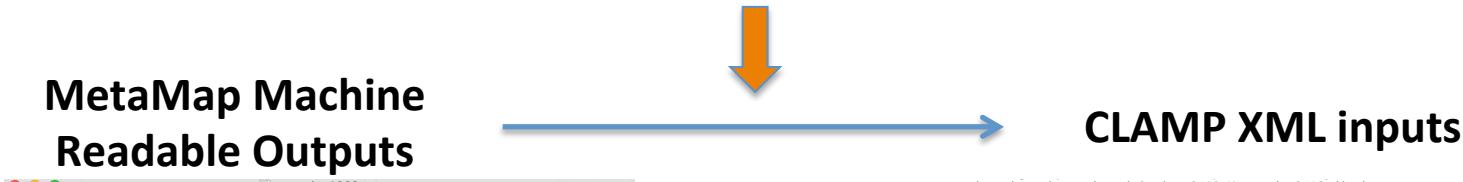
# A Library of NLP Pipelines

- General
  - Problem/treatment/test
  - Diseases with modifiers
  - Medications with signature
  - Lab tests and values
- Disease-specific
  - Colorectal cancer
  - Bleedings
  - ...
- Behavior
  - Smoking status
  - ...



# Interoperate with existing systems

- Compatible with cTAKES – UIMA
  - MetaMap2CLAMP – a command line program



urine throughout the admission.<sup>6,10</sup> Keywords: discharge summary, iron deficiency anemia, hypertension and obesity, iron deficiency, urine, anemia, hypertension, chemotherapy, discharge, ewing, sarcoma,<sup>6,10</sup> "/> **<tcas:DocumentAnnotation** xmi:id="8"

**sofa="1"** begin="0" end="195"**language="x-unspecified">** <textspan:Sentence xmi:id="13"

**sofa="1"** begin="0" end="51" sentenceNumber="0"/> **<textspan:Sentence** xmi:id="19" sofa="1"

**begin="52"** end="86" sentenceNumber="1"/> **<textspan:Sentence** xmi:id="25" sofa="1" begin="87"

end="208" sentenceNumber="2"/> **<textspan:Sentence** xmi:id="31" sofa="1" begin="209"

end="246" sentenceNumber="3"/> **<textspan:Sentence** xmi:id="37" sofa="1" begin="247"

end="278" sentenceNumber="4"/> **<textspan:Sentence** xmi:id="43" sofa="1" begin="279"

end="310" sentenceNumber="5"/> **<textspan:Sentence** xmi:id="49" sofa="1" begin="311"

end="344" sentenceNumber="6"/> **<textspan:Sentence** xmi:id="55" sofa="1" begin="345"

end="389" sentenceNumber="7"/> **<textspan:Sentence** xmi:id="61" sofa="1" begin="390"

end="410" sentenceNumber="8"/> **<textspan:Sentence** xmi:id="67" sofa="1" begin="411"

end="413" sentenceNumber="9"/> **<textspan:Sentence** xmi:id="73" sofa="1" begin="415"

end="429" sentenceNumber="10"/> **<textspan:Sentence** xmi:id="79" sofa="1" begin="430"

end="432" sentenceNumber="11"/> **<textspan:Sentence** xmi:id="85" sofa="1" begin="434"

end="441" sentenceNumber="12"/> **<textspan:Sentence** xmi:id="91" sofa="1" begin="442"

end="444" sentenceNumber="13"/> **<textspan:Sentence** xmi:id="97" sofa="1" begin="446"

end="459" sentenceNumber="14"/> **<textspan:Sentence** xmi:id="103" sofa="1" begin="460"

end="462" sentenceNumber="15"/> **<textspan:Sentence** xmi:id="109" sofa="1" begin="464"

end="477" sentenceNumber="16"/> **<textspan:Sentence** xmi:id="115" sofa="1" begin="478"

end="568" sentenceNumber="17"/> **<textspan:Sentence** xmi:id="121" sofa="1" begin="569"

end="601" sentenceNumber="18"/> **<textspan:Sentence** xmi:id="127" sofa="1" begin="602"

end="754" sentenceNumber="19"/> **<textspan:Sentence** xmi:id="133" sofa="1" begin="756"

end="842" sentenceNumber="20"/> **<textspan:Sentence** xmi:id="139" sofa="1" begin="844"

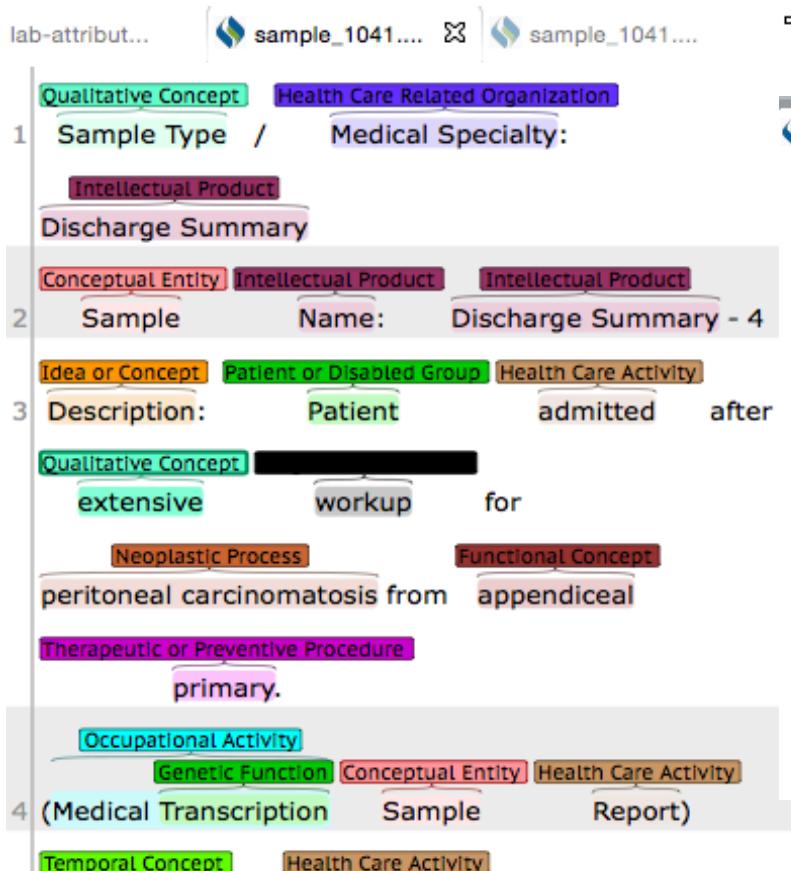
end="1016" sentenceNumber="21"/> **<textspan:Sentence** xmi:id="145" sofa="1" begin="1018"

end="1092" sentenceNumber="22"/> **<textspan:Sentence** xmi:id="151" sofa="1" begin="1094"

end="1119" sentenceNumber="23"/> **<textspan:Sentence** xmi:id="157" sofa="1" begin="1121"

end="1188" sentenceNumber="24"/> **<textspan:Sentence** xmi:id="163" sofa="1" begin="1190"

# MetaMap concept filtering



sample\_1041... sample\_1041... sample\_1054... sample\_1042... "1

13 4. Hyperkalemia.

14 PROCEDURES DURING HOSPITALIZATION: Cycle seven Ifosfamide, mesna, and VP-16 chemotherapy.

15 HISTORY OF PRESENT ILLNESS: Ms. XXX is a pleasant 37-year-old African-American female with the past medical history of Ewing sarcoma, iron deficiency anemia, hypertension, and obesity. She presented initially with a left frontal orbital swelling to Dr. XYZ on MM/DD/YYYY. A biopsy revealed small round cells and repeat biopsy on MM/DD/YYYY also showed round cells consistent with Ewing sarcoma, genetic analysis indicated a T1122 translocation. MRI on MM/DD/YYYY showed a 4 cm soft tissue mass without bony destruction. CT showed similar result. The patient received her first cycle of chemotherapy on MM/DD/YYYY. On MM/DD/YYYY, she was admitted to the ED with nausea

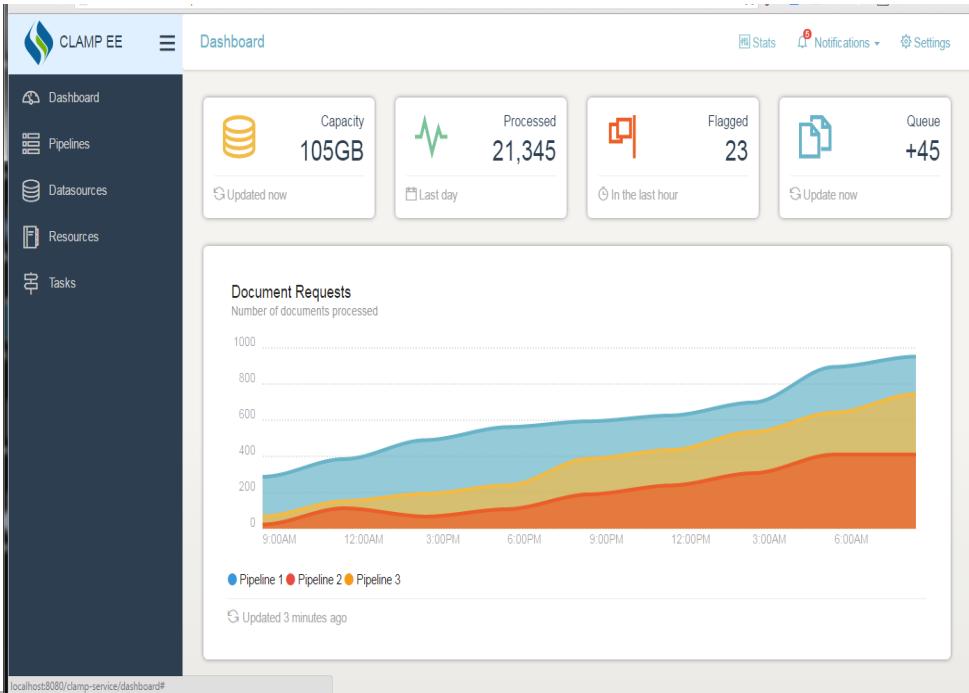
# Additional Features for Developers

- CLAMP APIs for different components
  - Document object encapsulates the UIMA JCAs object to overcome its complexity.
- CLAMP Extensions
  - Create custom components by implementing DocProcessor interface (UIMA compatible).



# CLAMP Enterprise

- Still under development



### Clinical Text Search Engine

hematuria

Select synonyms you want add to search

1-7 of 7 Patients for "hematuria" in Clinical Notes (found 78 notes)

Patient ID	Doc Num	Details
Patient_7	31	2010-05-09 Reason for Consultation: gross hematuria, pt with Foley for urinary retention 2010-05-09 [2010] DISCHARGED: "DATE[May 09 2010] PRINCIPAL DIAGNOSIS: 1. Hematuria Brief admission 2012-03-04 Called by ED re hematuria s/p SPT change. Mr. "NAME[AAA] is a pt. of Dr. "NAME[ZZT] who 2010-09-13 [2010] DISCHARGED: "DATE[Sep 13 2010] PRINCIPAL DIAGNOSIS: 1. Gross hematuria SECONDARY 2010-09-10 >Consult Request DetailsReason for Consultation: suprapubic catheter. Hematuria and penile More...
Patient_8	25	2006-01-20 good, without clear explanation for hematuria. It does show a cyst in the kidney, which is likely 2006-02-07 "PHONE FAX (615) "PHONE Primary Care Physician. "NAME[XXX WWW], M.D. Chief Complaint: Hematuria 2006-01-21 is worked in today after phoning this morning with gross hematuria. She had previously seen Dr 2011-06-04 3 0x3mm Vision bare metal stents. Bare metal stents were used due to a past history of hematuria and 2006-02-06 hematuria (note not yet back). Sounds like she is going to have cystoscopy. She has had no further hematuria More...
Patient_2	11	2008-04-22 hematuria evaluation--negative. PROBLEM LIST : Reviewed and updated. PE: on office chart. IMP: prostate 2008-07-01 visit and today 18 ALU back up to 10 Has had previous hematuria evaluation--negative. PROBLEM LIST 2011-10-21 medications. AUA symptom score 3. Post void residual 202 cc. He denies any gross hematuria or other 2010-08-24 symptom score 3. Post void residual 169 cc. He denies any gross hematuria or other complaints. He 2012-04-17. He denies any gross hematuria or other complaints. He reports good stream, feels empty after void More...
Patient_6	4	2003-01-11 hematuria, or joint pain. She has had "a little diarrhea" She came to the "NAME[XXX] ER and had 2002-02-20 patient denies dysuria. hematuria NEUROLOGIC: The patient denies dizziness, syncope, seizures 1997-12-14 urinate. She denies hesitancy, dysuria, hematuria, or abdominal discharge. MUSCULOSKELETAL 2009-10-14 hematemesis, no diarrhea, no melena, no hematemesis, no abdominal pain GU: no dysuria, no hematuria More...
Patient_10	3	2012-01-08 hematuria,no urgency) Musculoskeletal: no joint pains Neurologic: no dizziness, no seizures, no 2007-06-24 " no dysuria, hematuria or urinary incontinence. 2006-06-23 margin-top: 0pt; class="negative"> No hematuria or urinary More...
Patient_4	3	2005-07-09 the hematuria or no frequency. Nocturia x 1-2. Musculoskeletal: no joint pain no swelling, no 2005-12-22 " class="negative"> No dysuria or hematuria 2005-01-05, no constipation, no melena, no RBRPR GU endorses no dysuria, no hematuria. Musculoskeletal More...
Patient_3	1	2010-01-10 , no constipation, no melena, no abdominal pain GU: no dysuria, no hematuria. Neurologic: no More...

Note Type

- hp (66)
- ds (5)
- rh (3)
- rad (2)
- cc (1)
- fh (1)

# CLAMP users and use cases

- 20+ different organizations: academic + industry
- Many different use cases



# Use case # 1 – computer assisted coding of ICD 10 at MD Anderson

Transcribed Documents:

Date	Document ...	Status	Service	Responsible Clinician	Dictator
[Redacted]					

findings have ICD10 codes. Please read individual document for all NLP findings. Color legend: Problem, Treatment, Procedure, Test, Laterality.

CUI	Procedure	Doc Location	ICD10-CM	Diagnosis	Doc Location
C0398534	a matched unrelated donor transplant	[Redacted]	D64.9	anemia	[Redacted]

No Comorbidity found by NLP for this patient 7 days before Admission date upto Admission date.

Standard Annotated

BRIEF HISTORY:  
Mr. [Redacted] with history of B-ALL, who is status post a matched unrelated donor stem cell transplant, utilizing standard of care, busulfan, clofarabine, and ATG as conditioning regimen, and received his stem cell transplant on [Redacted] for patient's transplant history. Briefly, the patient's transplant course was complicated by acute skin GVHD; however, [Redacted] last month. He remained on tacrolimus for GVHD prophylaxis. There was concern for acute liver graft-versus-host disease in which patient's steroids were initiated at [Redacted] mg/kg for concern of graft-versus-host disease. Otherwise, the patient was admitted for further workup, as well as concerning circulating peripheral blasts.

HOSPITAL COURSE:  
1. Elevated LFTs. The patient did have persistent progressive hyperbilirubinemia with question of graft-versus-host disease versus leukemic infiltration of the liver. The patient did undergo a transjugular liver biopsy on [Redacted] and that pathology is pending. With question of graft-versus-host disease versus leukemic infiltration of the liver, the patient's Solu-Medrol was decreased to [Redacted] mg/kg and remains on tacrolimus. We will follow up with the pathology results on an outpatient basis. The patient's liver enzymes did trend upward in which patient's total bilirubin was [Redacted], indirect bilirubin was [Redacted] upon discharge, with alk phos of [Redacted], LDH of [Redacted], ALT of [Redacted].  
2. Patient did have questionable relapsed disease with circulating peripheral blasts. The patient did obtain a bone marrow biopsy and aspiration on [Redacted]. Patient did have [Redacted] blasts at that time, and the patient's differential from bone marrow differential did have [Redacted] blasts in his bone marrow. Leukemia was consulted for further recommendations. The patient will be seen by [Redacted].  
3. Counts. The patient's counts are stable. The patient does have some leukocytosis with white blood count [Redacted] peripheral blasts, platelets of [Redacted] hemoglobin [Redacted] on day of discharge. The patient did receive platelets for transjugular liver biopsy.  
4. Graft-versus-host disease. Currently, all other GVHD is quiescent; however, with questionable liver GVHD. The patient's voriconazole was placed on hold, and patient was initiated on caspofungin for antifungal coverage. Again, patient's total bilirubin upon discharge was [Redacted] direct bilirubin [Redacted]. Liver ultrasound was obtained on [Redacted] that just indicated hepatomegaly; however, without evidence of focal liver lesions or intra or extrahepatic biliary ductal dilatation.  
5. Infection. The patient will continue on valacyclovir, caspofungin, atovaquone, and Vantin as antimicrobial prophylaxis. The patient was asymptomatic for any signs or symptoms of infection during this hospital course.

# Use case # 2 – Meaningful use quality measurement – VTE detection at Memorial Hermann Hospital

https://localhost:8443/#/notes

VTE Web

Predicted 350502011 Acute massive pulmonary embolism

Annotation 350502011 Acute massive pulmonary embolism user ✓

Review Reviewer Co Reviewer Concept reviewer ?

**pulmonary angiography [pulmonary embolism protocol]. Subsequently, helical CT images were obtained from the thoracic inlet to the upper abdomen.**

**FINDINGS:** Extensive pulmonary embolism is present bilaterally. No there are some small lymph nodes in the AP window and paratracheal is present in the peripher distended. Continuation ex visualized portion of the lateral spleen that measure irregular lesion in the ri IMPRESSION: 1. Bilateral p 2. Irregular shaped infiltrate. 3. Mildly distended appearance of the right heart. 4. Mild mediastinal adenopathy. 5. Splenic hemangioma suspected with peripheral enhancement.

**Sensitivity (Recall): 0.98**  
**Specificity : 0.94**  
**PPV (Precision) : 0.89**  
**Accuracy : 0.95**

Review	Reviewer
None []	reviewer
None []	reviewer
Pulmonary embolism [98484016]	reviewer
Acute massive pulmonary embolism [350502011]	reviewer
None []	reviewer
None []	reviewer

Close Save

# Use case #3 – support clinical research and informatics education

- EHR-based drug studies
  - Pharmacovigilance
  - Pharmacogenomics
  - Drug repurposing
- Domain-specific studies
  - Cancer epidemiology
  - Psychiatric studies
  - .....
- Education for biomedical NLP



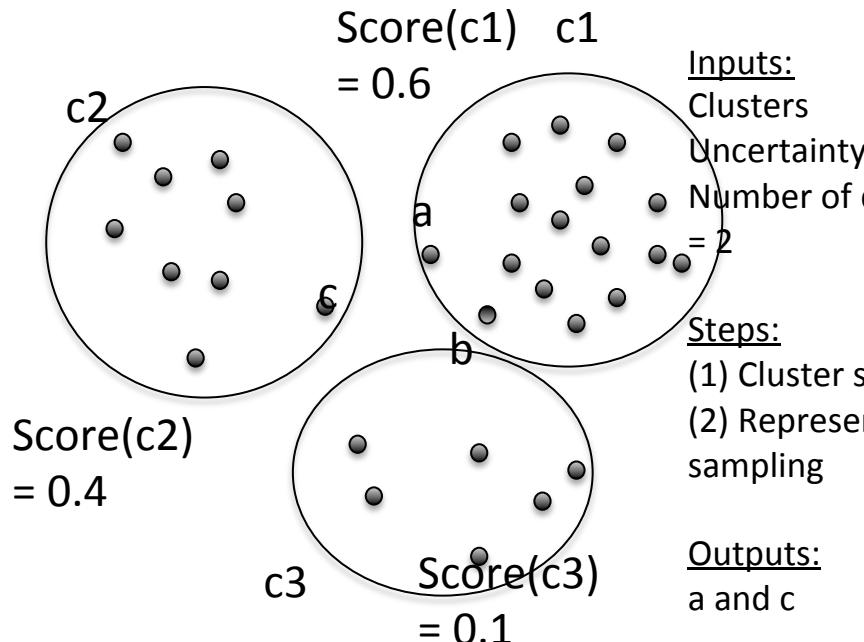
# More advanced algorithms to facilitate the transportability

- Active learning-based pre-annotation
- Deep learning-based NER
- Domain adaption
- Abbreviation handling

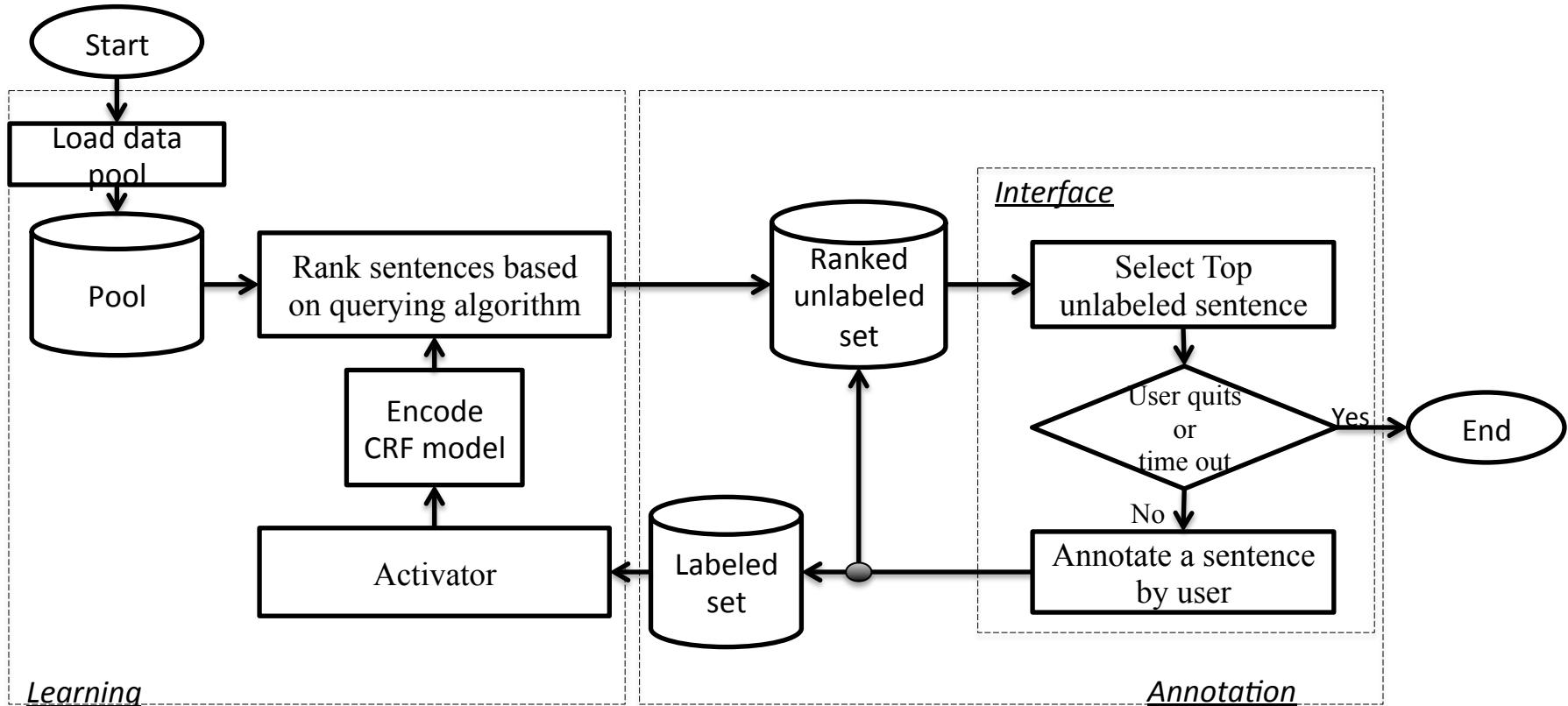


# Active learning based pre-annotation

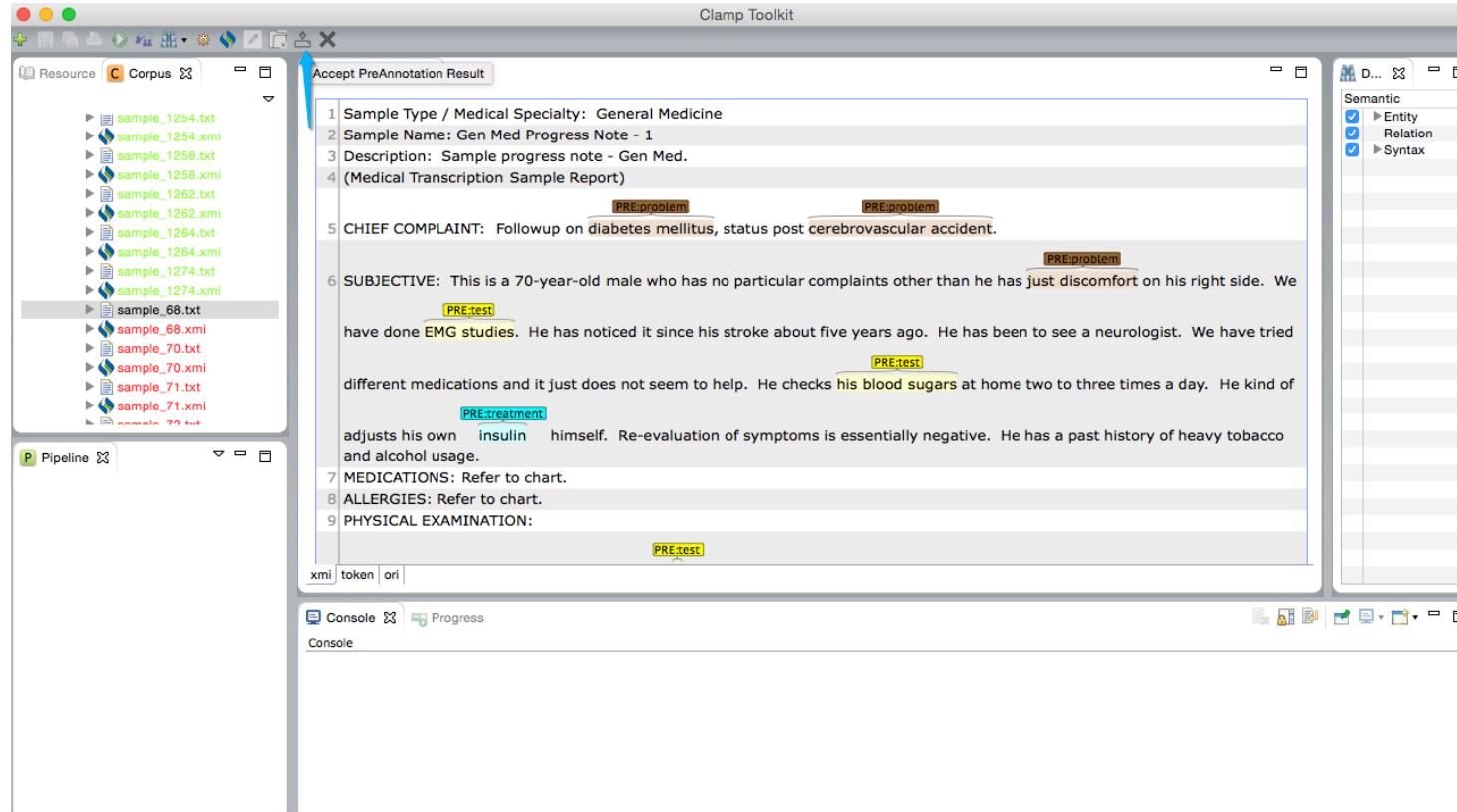
- Motivation
  - Annotation is costly
- Algorithm development
  - Clustering and uncertainty sampling engine (CAUSE)
  - Query the most uncertain and representative sentences



# Active learning workflow

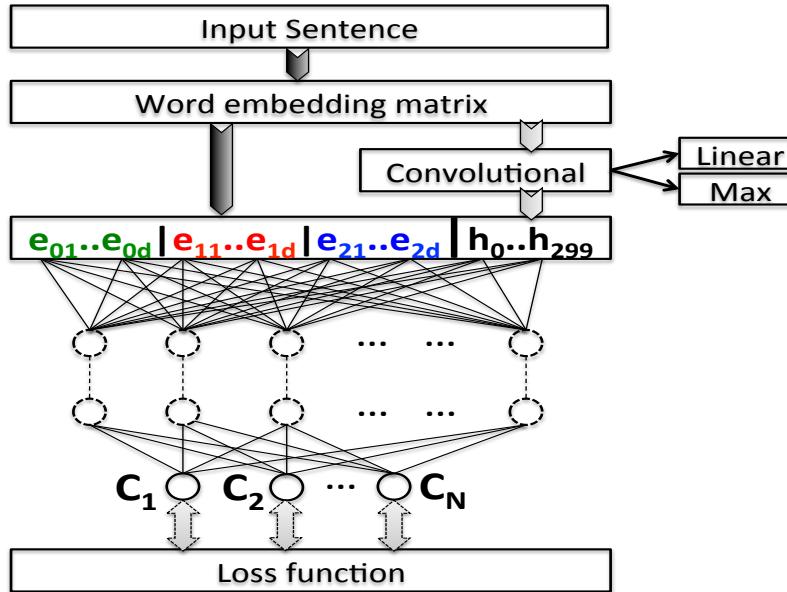


# Pre-annotation interface



# Deep neural networks for NER

The i2b2 corpus



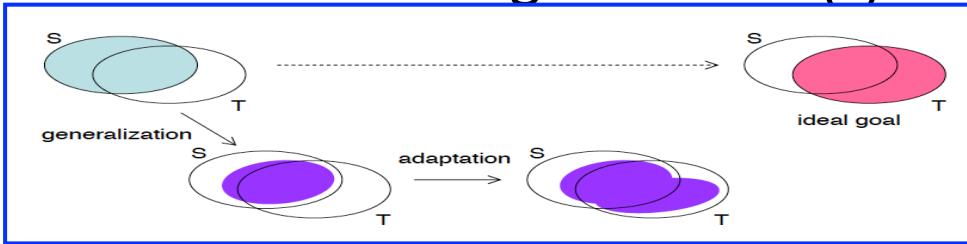
	Precision	Recall	F1-score
CRF-baseline	0.848	0.755	0.799
CRF-optimized	0.872	0.825	0.848
DNN	0.849	0.807	0.828

DNN Advantages:

- No human feature engineering
- No domain knowledge
- Simplified application system architecture
- Good baseline performance

# Domain Adaptation

- **Motivation:** deploy models from fixed source domain across different target domain(s)



- Different algorithms
  - **Instance pruning:** remove most different instances from source
  - **Transfer self-training:** add most similar source instances to the target training set iteratively for retrain
  - **Feature augment:** amplifying target training features with weight-adapted source domain features.



UTHealth™  
The University of Texas  
Health Science Center at Houston

School of Biomedical  
Informatics

# Domain Adaptation for semantic role labeling (SRL)

- significantly improved SRL performance
- reduced ~40% annotation cost
- directly using source domain dataset without domain adaptation may decrease the SRL performance
- different domain adaptation algorithms may be effective for different source domain datasets

Zhang Y et al, JAMIA 2015.

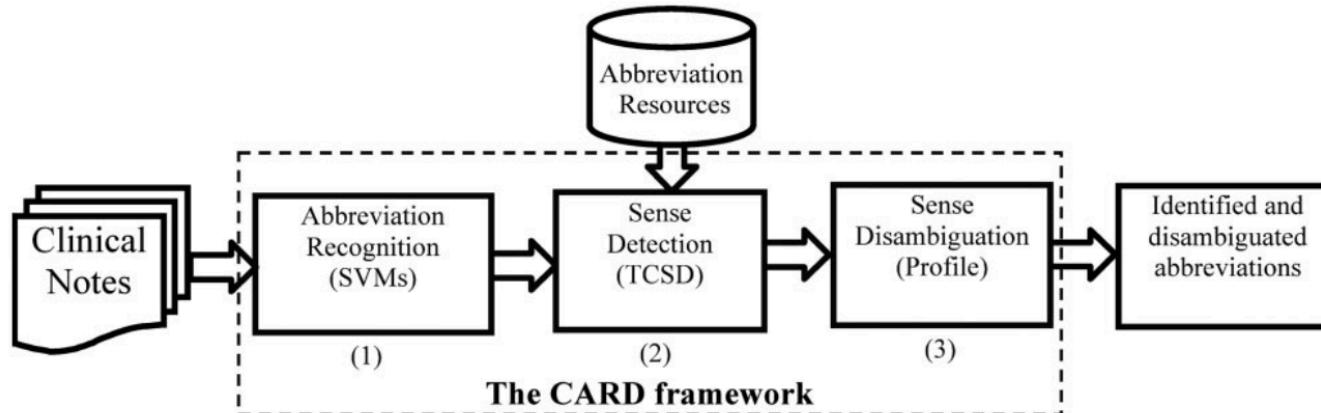


UTHealth™  
The University of Texas  
Health Science Center at Houston

School of Biomedical  
Informatics

# Handling abbreviations

- Motivation
  - Abbreviations are pervasive, important, ambiguous, dynamic
- CARD – Clinical Abbreviation Recognition and Disambiguation framework



# Work with CLAMP

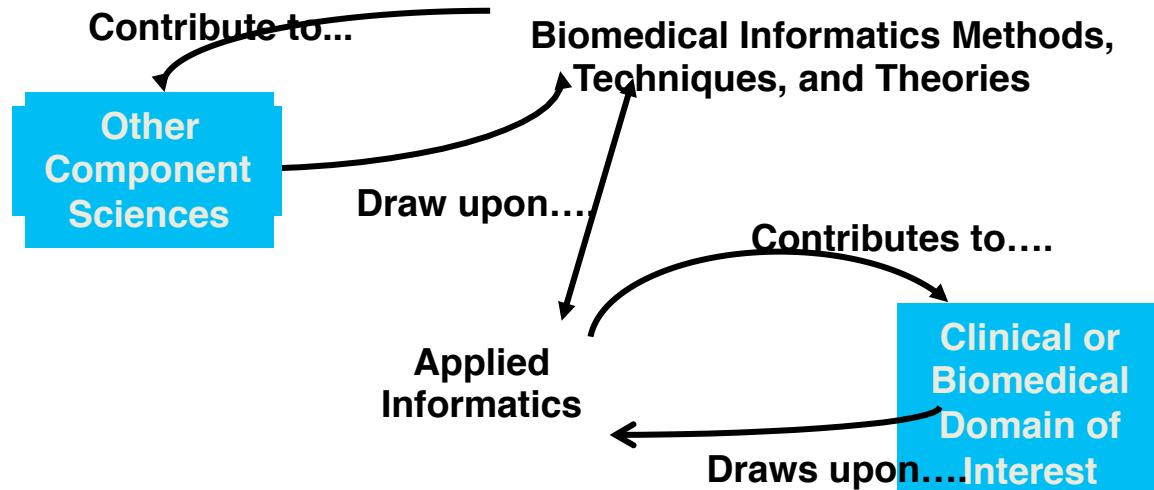
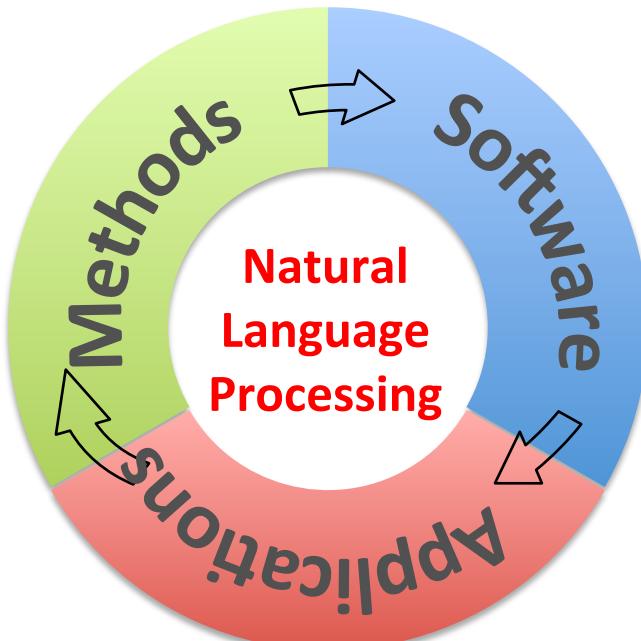
- License model
  - Research – free
    - User (without source code)
    - Contributor (with source code)
  - Operation/Commercial – paid
    - User (without source code)
    - Contributor (with source code)
- Support service
  - GitHub issue reporting
  - Supporting team: documentation, tutorial, QA...
- Available at <http://clamp.uth.edu/>



UTHealth™  
The University of Texas  
Health Science Center at Houston

School of Biomedical  
Informatics

# NLP and Biomedical Informatics



# Acknowledgement

- Collaborators
  - Robert Murphy, MD
  - Josh Denny, MD
  - Trent Rosenbloom, MD
  - Randy Miller, MD
  - Hongfang Liu, PhD
  - Qiaozhu Mei, PhD
  - Serguei Pakhomov, PhD
  - Jason Hou, MD
  - Tom Lasko, MD, PhD
- Grants
  - CPRIT R1307
  - NIGMS R01 GM102282
  - NLM R01 LM010681

## Team members:

- Jingqi Wang
- Min Jiang
- Ergin Soysal
- Sungrim Moon
- Jun Xu
- Yaoyun Zhang
- Anupama Gururaj
- Yonghui Wu
- Nina Slimi
- Kyle Nguyen
- Tolulola Dawodu
- Yukun Chen
- Qiang Wei
- Saied Pournejati
- Rui Li

# Thank you!

# Questions?

[hua.xu@uth.tmc.edu](mailto:hua.xu@uth.tmc.edu)