CLAMP-Cancer – an NLP tool to facilitate cancer research using EHRs

Hua Xu, PhD

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

Advancing Cancer Pharmacoepidemiology Research Through EHRs and Informatics (1 U24 CA194215-01A1)

- Team
  - UTHealth (Hua Xu)
  - Vanderbilt (Josh Denny)
  - Mayo Clinic (Ping Yang)
- Specific Aims of the study

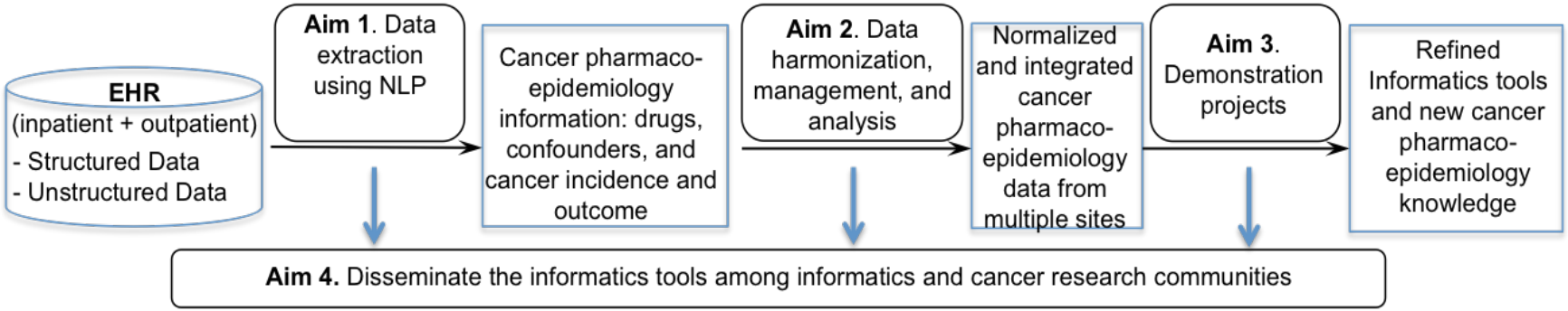


Figure 1. An overview of the proposed specific aims.

EHR - an enabling resource for clinical research

- **CTSA** – Clinical and translational science
- **eMERGE** – EMRs linked with genomic data
- **“All of US” -** Research Program of Precision Medicine Initiative (PMI)
- **PCORnet** - National Patient-Centered Clinical Research Network
- **OHDSI** – Observational health data sciences and informatics

Narratives in EHRs

Admit 10/23
71 yo woman h/o DM, HTN, Dilated CM/CHF, Afib s/p embolic event, chronic diarrhea, admitted with SOB. CXR pulm edema. Rx'd Lasix.
All: none
Meds Lasix 40mg IVP bid, ASA, Coumadin 5, Prinivil 10, glucophage 850 bid, glipizide 10 bid, immodium prn

- Clinical documents contain rich patient information needed for clinical research
- It is costly and time-consuming to extract such information manually

Clinical natural language processing (NLP)

- Methods
  - Named entity recognition (NER), concept encoding, relation extraction (e.g., temporal), co-reference resolution…
- Tools
  - General purpose : MedLEE, MetaMap, cTAKES….
  - Specific purpose : medication, diseases, labs ….
- Applications
  - Clinical research, clinical decision support, surveillance…

A challenge of clinical NLP - Portability

Need a solution for users to efficiently build high-performance NLP pipelines for individual applications!

CLAMP - Clinical Language Annotation, Modeling, and Processing

- A general purpose, high-performance clinical NLP system built on proven methods
- An Integrated development environment (IDE) for building customized clinical NLP pipelines
- A scalable enterprise solution for NLP needs in healthcare organizations
- Available at http://clamp.uth.edu

CLAMP – 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 default pipeline performance

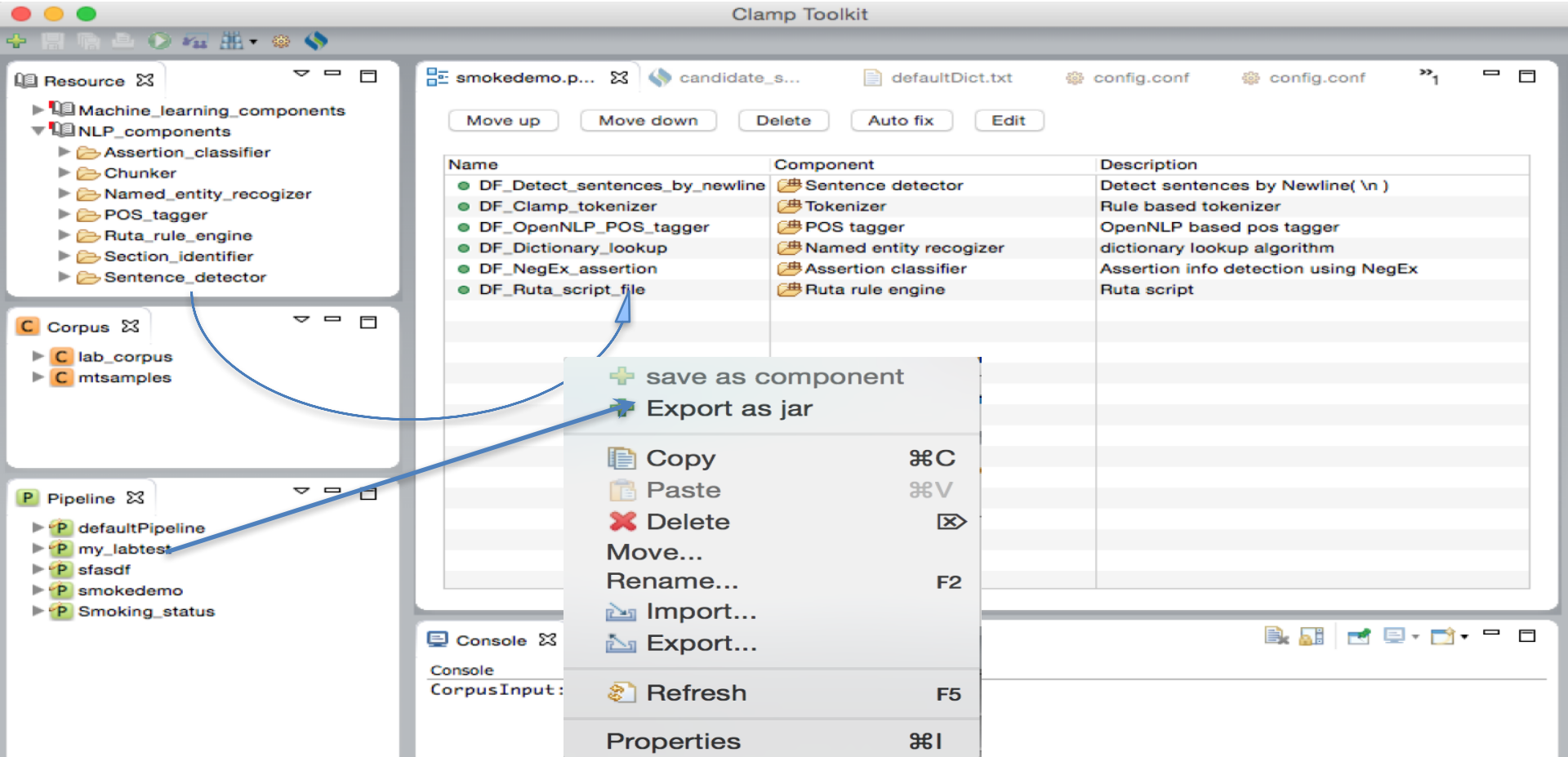
- Extract problems, treatments, and tests

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

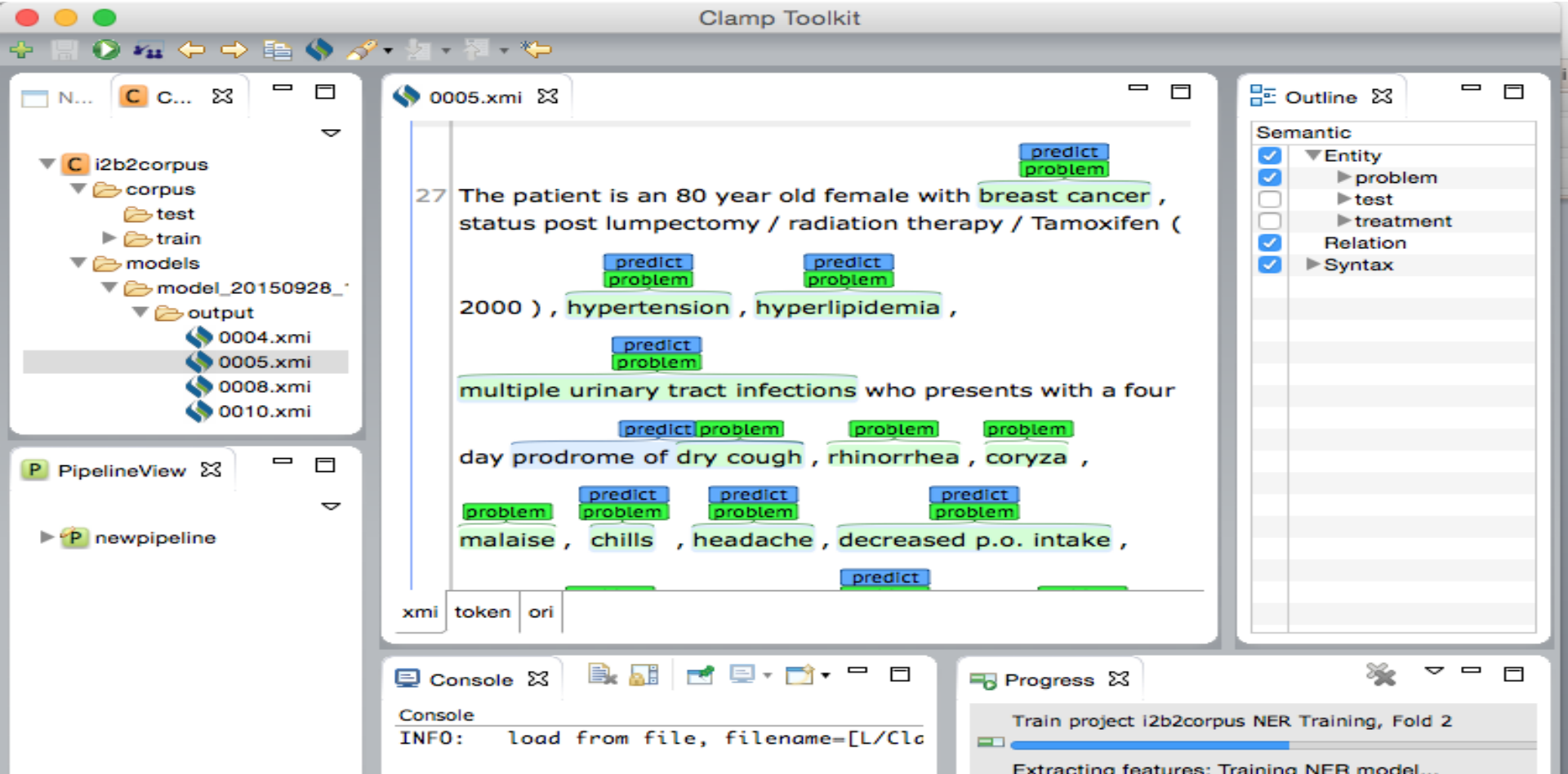
CLAMP GUI

- An IDE (integrated development environment) for building customized clinical NLP pipelines via GUIs
  - Annotating/analyzing clinical text
  - Training of ML-based modules
  - Specifying rules

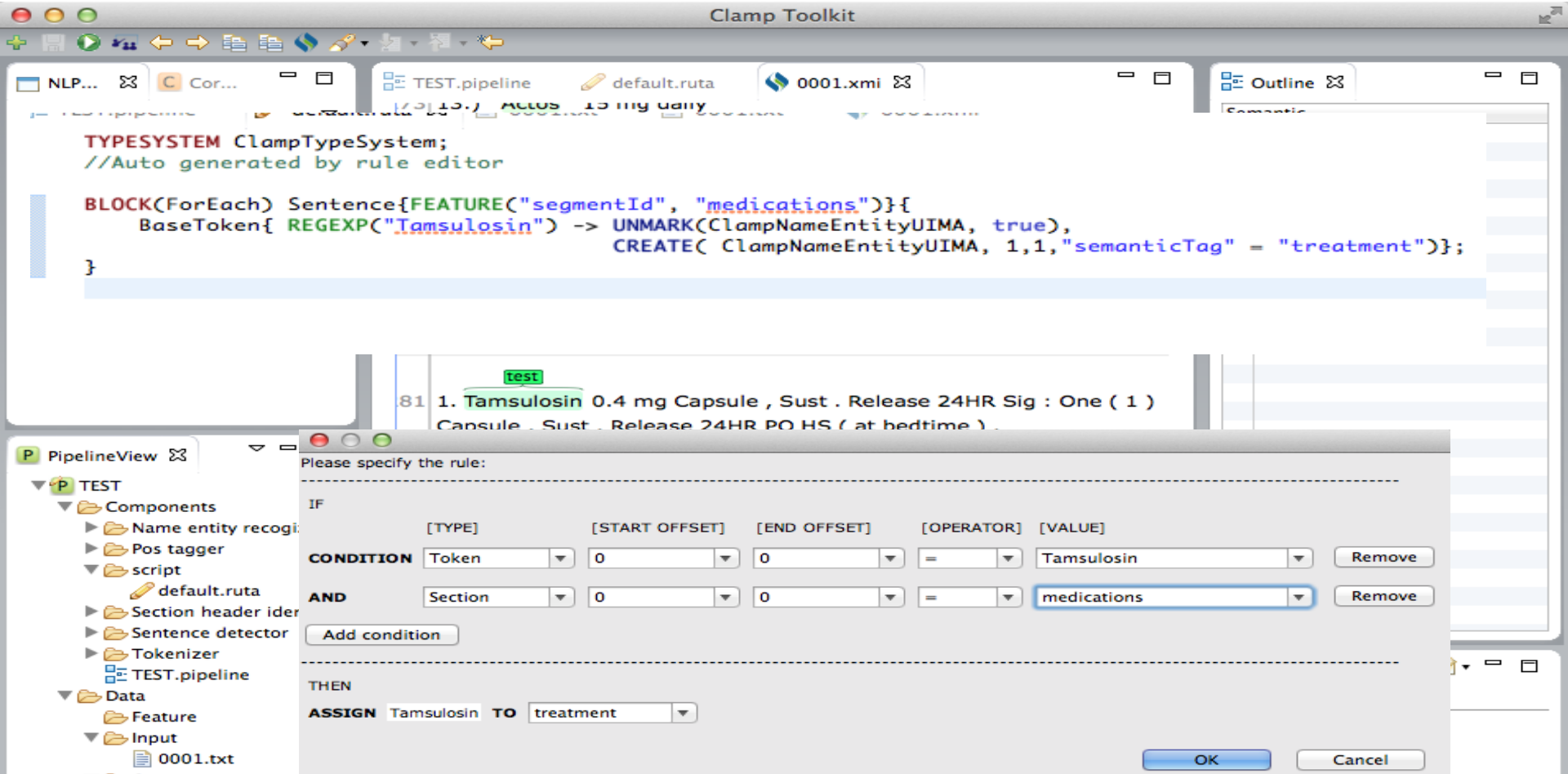
Building your own pipeline



Annotating/Re-training



Specifying rules

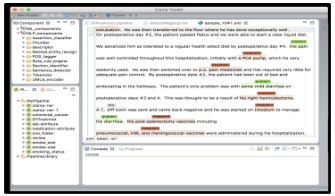


Two use cases

- 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

CLAMP Enterprise

CLAMP-GUI



NLP algorithm development

CLAMP-EE

- Deploy at different settings
  - Database
  - Websvc
  - FHIR
- Manage tasks
  - Coordinate execution
  - Monitor system health
- Visualize results
  - Search
  - Validate

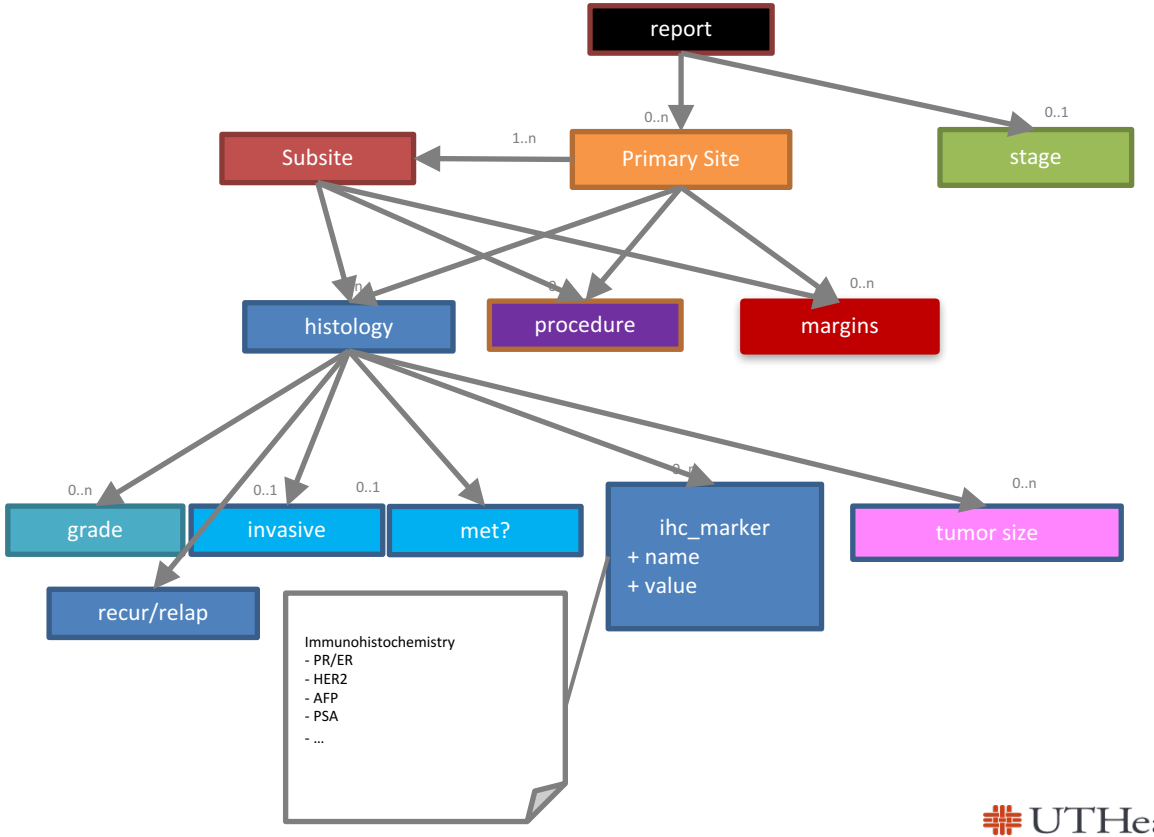
CLAMP-Cancer

- Focus on extracting cancer related information from EHRs
- High-performance default components
- Customizable by end-users
- User-friendly interface

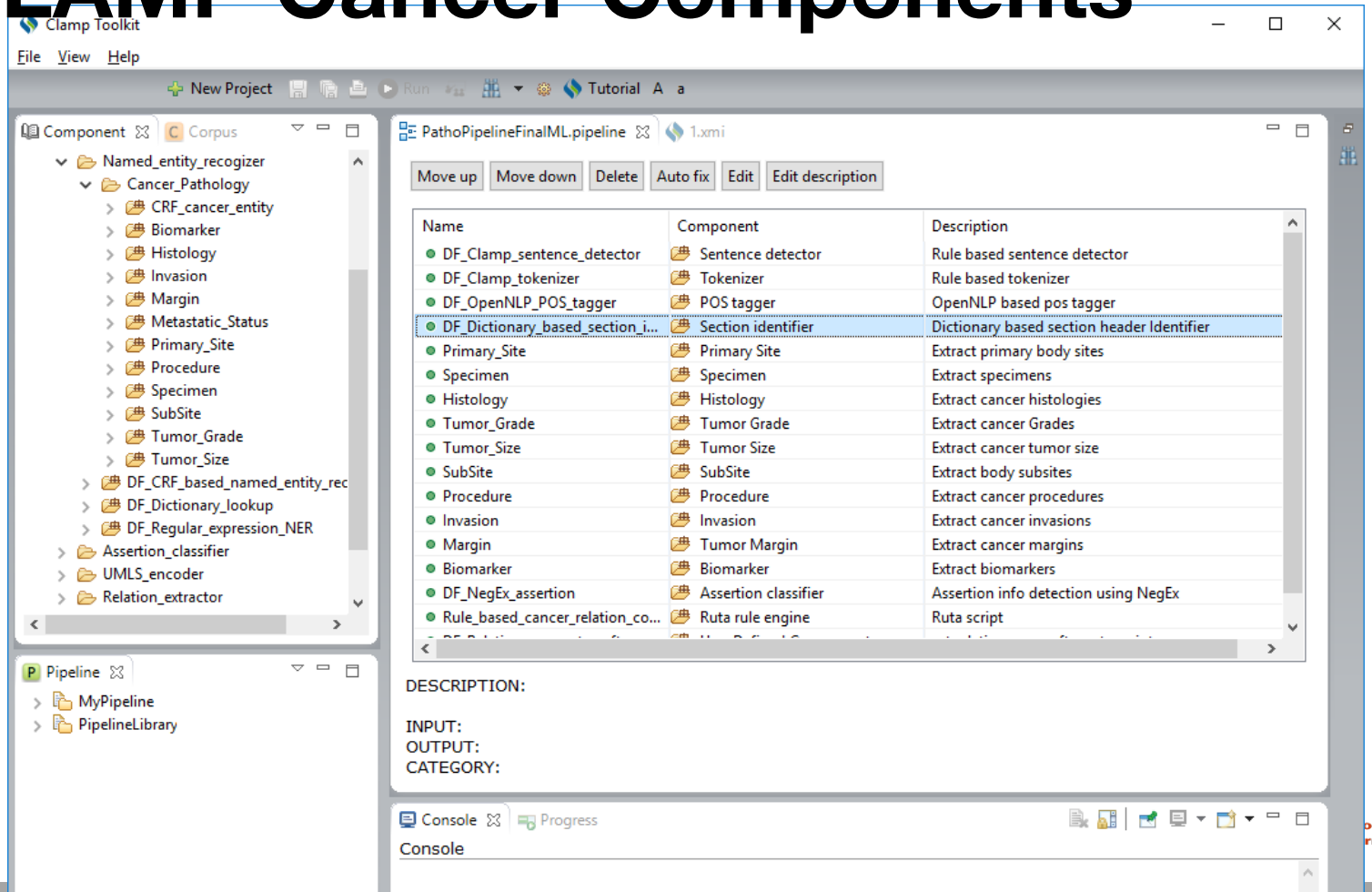
A first attempt – Diagnoses in pathology reports

- Developed an Information Model based on College of American Pathologists (CAP) Cancer Protocols
  - Primary sites, Histology/Grade, Procedure, Margin, Invasion, Metastatic Status, Tumor Size, Biomarkers
- Develop a set of NLP components combining machine learning and rule based approaches to extract entities and relations

CLAMP Cancer Information Model



CLAMP Cancer Components



ool of Biomedical
rmatics

CLAMP-Cancer (Vanderbilt)

Type of information	Exact Matching			Relaxed Matching		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Specimen	0.96	0.96	0.96	0.96	0.96	0.96
Primary Site	0.96	0.95	0.95	0.98	0.97	0.98
Sub-Site	0.91	0.82	0.86	0.94	0.85	0.90
Procedure	0.96	0.97	0.97	0.97	0.98	0.98
Histology	0.91	0.85	0.88	1.00	0.93	0.97
Tumor Grade	0.92	0.94	0.93	0.97	1.00	0.99
Tumor Size	1.00	0.83	0.91	1.00	0.83	0.91
Tumor Margin	0.93	0.93	0.93	1.00	1.00	1.00
Invasion	0.96	0.93	0.95	0.96	0.93	0.95
Biomarker	0.92	0.88	0.90	0.98	0.94	0.96

Soysal et al. *under review*

CLAMP-Cancer and MedKAT (Mayo)

	CLAMP-Cancer			MedKAT		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Tumor size	1.00	0.99	0.99	1.00	1.00	1.00
Dimension Extent	1.00	0.99	0.99	0.99	1.00	0.99
Dimension Unit	1.00	1.00	1.00	1.00	1.00	1.00
Tumor Site	0.94	0.89	0.92	0.96	0.95	0.96
Histology	0.91	0.92	0.92	0.96	0.98	0.97
Grade	1.00	0.88	0.94	0.93	0.97	0.99
Date	1.00	1.00	1.00	1.00	1.00	1.00

Soysal et al. *under review*

Future work

- Advanced methods to facilitate the portable NLP solutions
- New interfaces for cancer researchers
- More cancer specific NLP pipelines for diverse types of clinical notes
- Interoperability with other NLP systems: cTAKES and MetaMap

Acknowledgement

- Collaborators
  - Josh Denny, MD
  - Jeremy Warner, MD
  - Cindy Chen, PhD
  - Ping Yang, MD
  - Hongfang Liu, PhD
  - Serguei Pakhomov, PhD
  - Xianling Du, PhD

- Grants
  - NCI U24 CA194215
  - CPRIT R1307
  - NIGMS R01 GM102282
  - NLM R01 LM010681

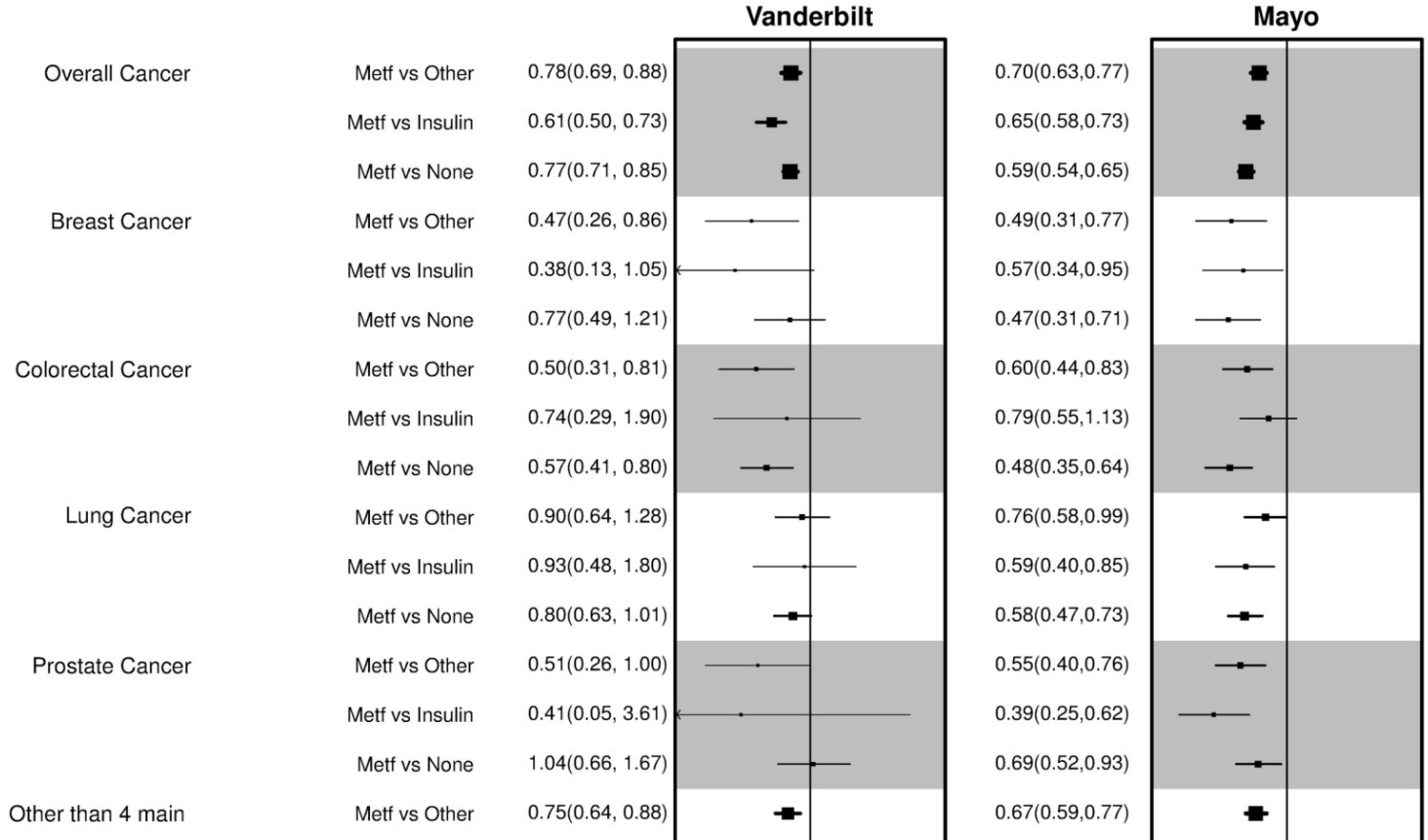
Team members:
- Jingqi Wang
- Min Jiang
- Ergin Soysal
- Sungrim Moon
- Jun Xu
- Yaoyun Zhang
- Anupama Gururaj
- Yonghui Wu
- Nina Slimi
- Kyle Nguyen
- Tolulola Dawodu
- Yukun Chen
- Qiang Wei
- Saied Pournejati
- Rui Li

Thank you!

Questions?

hua.xu@uth.tmc.edu

NLP to support clinical research - Metformin and Cancer Survival

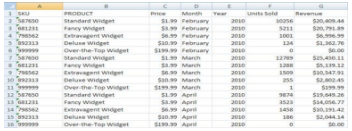


Screen 153 drugs for potential cancer therapeutic signals

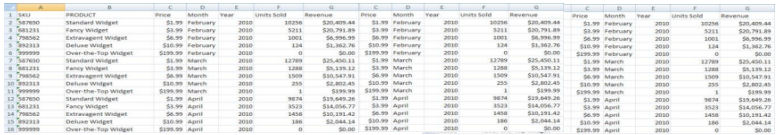
- Single Drug

153 drugs for chronic diseases

- Carefully selected covariates (30+)



all-inclusive covariates (3000+)



- Hypothesis-driven → Data-driven

Results

Rank	Drug	HR	P-value	Lower	Upper	Adj-p-value
1	sildenafil	0.704	0.000	0.621	0.797	0.000
2	olmesartan	0.714	0.001	0.580	0.879	0.017
3	thyroxine	0.791	0.000	0.734	0.852	0.000
4	carvedilol	0.803	0.003	0.693	0.930	0.033
5	alendronic.acid	0.812	0.000	0.725	0.908	0.004
6	amlodipine	0.839	0.000	0.782	0.901	0.000
7	epoetin.alfa.recom	0.840	0.000	0.772	0.914	0.001
8	ramipril	0.843	0.004	0.750	0.947	0.036
9	simvastatin	0.845	0.000	0.788	0.906	0.000
10	atorvastatin	0.861	0.000	0.797	0.930	0.003
11	metformin	0.862	0.003	0.781	0.952	0.033
12	esomeprazole	0.878	0.000	0.819	0.942	0.004
13	omeprazole	0.879	0.000	0.830	0.931	0.000
14	lisinopril	0.902	0.002	0.846	0.962	0.018
15	lansoprazole	0.908	0.005	0.849	0.971	0.037

4 drugs with strong evidence – ongoing clinical trials

7 drugs with weak evidence from literature