

Apache clinical Text

Analysis cTAKES Extraction System



ctakes.apache.org

Natural Language Processing

The patient underwent a CT scan in April which did not reveal lesions in his liver.

Entity Recognition	CT scan Procedure UMLS ID: C0040405	Lesion Disease / Disorder UMLS ID: C0022198	Liver Anatomy UMLS ID: C0023884
Entity Properties	Negated: no Subject: patient	Negated: yes Subject: patient	Negated: no --
UMLS Relation	Lesions <i>LOCATED AT</i> liver		
Temporal Relations	CT scan <i>WITHIN</i> April		Lesions <i>WITHIN</i> CT scan
Coreferences	patient <i>SAME AS</i> his		

Natural Language Processing

Service Date/Time: 02-May-2014 11:33
 Provider: Johnny Cash, MD. Pager: 4-5555
 Section: GI_Type/Desc: MIS Status: Trx Revision #: 2
 [end section id="20114"]

[start section id="20112"]
 CHIEF COMPLAINT/PURPOSE OF VISIT
 #1 Colorectal Cancer
 [end section id="20112"]

[start section id="20103"]
 HISTORY OF PRESENT ILLNESS
 The patient is a 55-year-old man referred by Dr. Good for recently diagnosed colorectal cancer. The patient was well till 6 months ago, when he started having a little blood with stool. He initially thought it was hemorrhoids related, and saw his primary physician after a few weeks of symptoms. As physical examination by the primary care provided did show small hemorrhoids but, as stool was strongly positive for blood, he was referred for a repeat colonoscopy. Due to family circumstances this was postponed for a few months. He had a prior colonoscopy at age 50, which revealed 5 or 6 polyps, all adenomatous, and the maximum size of the largest polyp was about 1 cm. This polyp was located in the cecum.
 The repeat colonoscopy on March 16, 2014, showed a 4 cm likely colorectal cancer in the cecum, just behind the ileocecal valve. Biopsies were positive for grade 3 out of 4 adenocarcinoma. Stains for HNPCC genes were negative. Patient did bring the slides of the colonoscopy biopsies but left these in his hotel room.
 The patient underwent a staging CT scan early April, which did not reveal obvious intra-abdominal spread and large lymph nodes, or metastatic lesions in his liver. We do not have the images available today.
 The patient is now here at our institution for evaluation of possible surgery.
 [end section id="20103"]

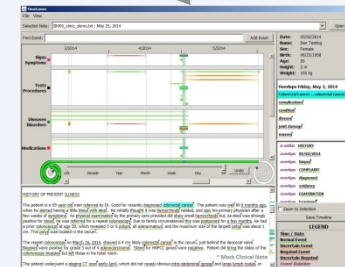
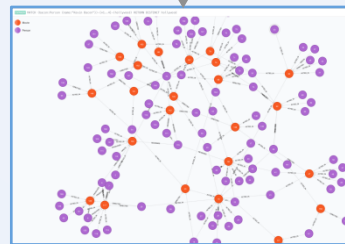
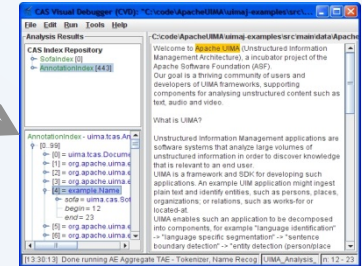
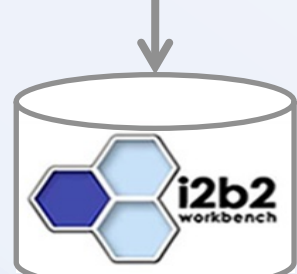


Storage

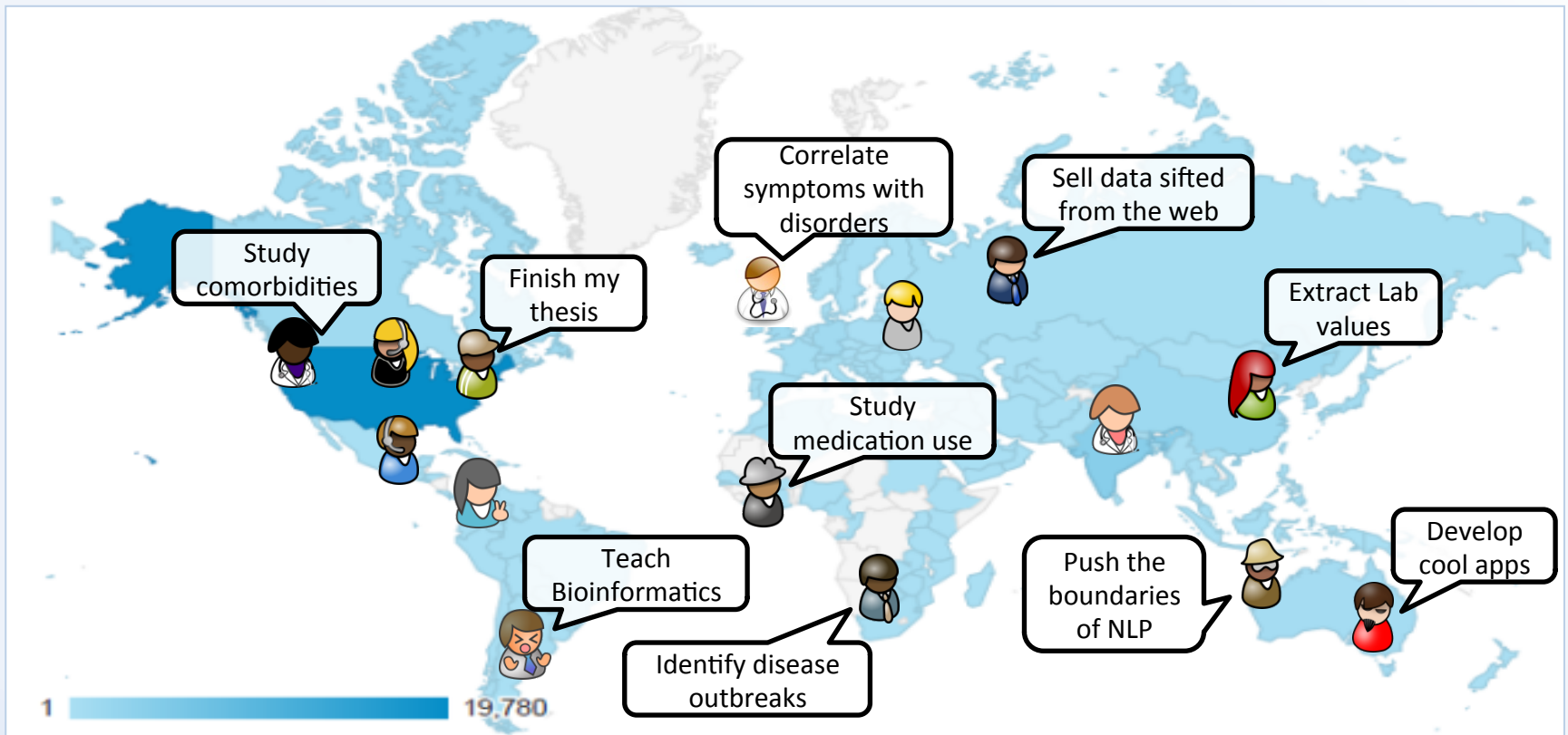


Application

```
<?xml version="1.0" encoding="UTF-8" standalone="yes" ?>
<us-bibliographic-data-grant>
  <publication-reference>
    <document-id>
      <country>US</country>
      <doc-number>08309744</doc-number>
      <kind>B2</kind>
      <date>20121113</date>
    </document-id>
  </publication-reference>
  <application-reference appl-type="utility">
    <document-id>
      <country>US</country>
      <doc-number>13081794</doc-number>
      <date>20110407</date>
    </document-id>
  </application-reference>
  <us-application-series-code>13</us-application-series-code>
  <priority-claims>
    <priority-claim kind="national" sequence="01">
      <country>CA</country>
      <doc-number>2609240</doc-number>
      <date>20080229</date>
    </priority-claim>
  </priority-claims>
</us-bibliographic-data-grant>
```



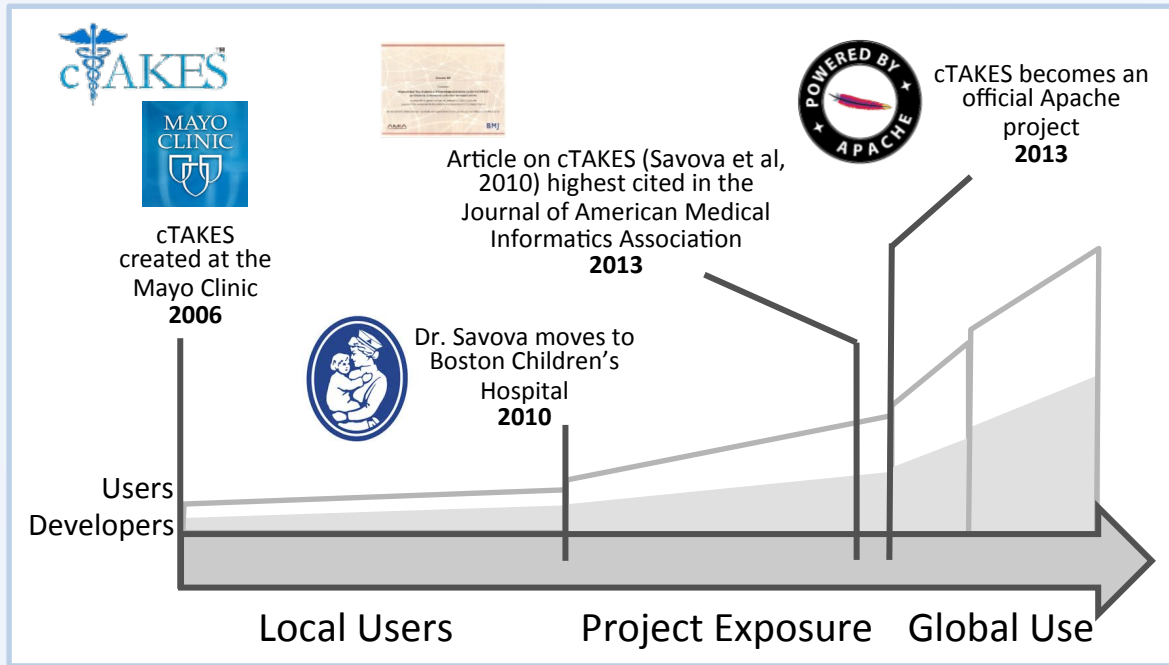
cTAKES Global Users





Users have many use cases ...

Point of Care is not (yet) one of them

cTAKES Growth



Participating in Core Projects   

Participating in NLP challenges  

Why Apache



Global Presence
Name Recognition

Free Resources and Support

- Version Control
- Web Server
- Wiki / Documentation
- Email List Servers
- Access to Software
- Legal Support

So we are golden, right?

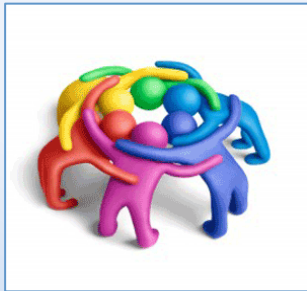
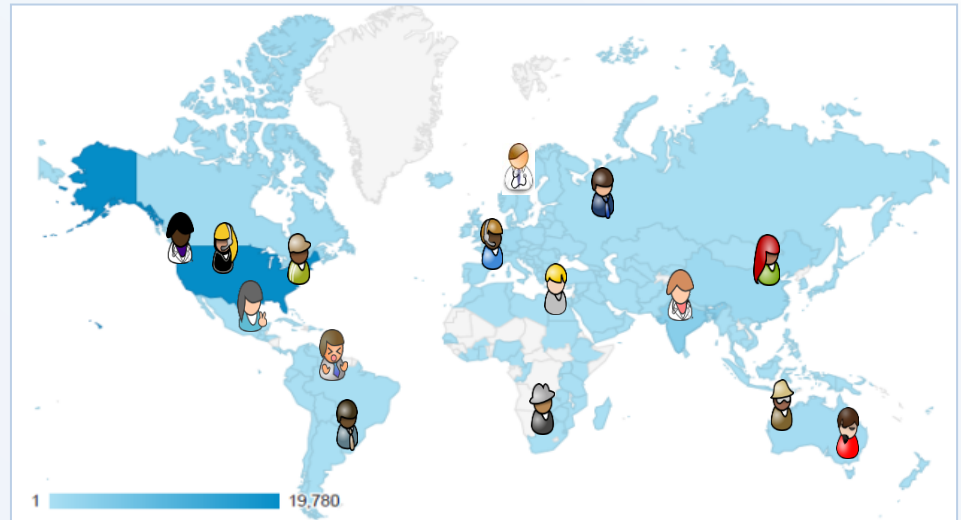


Core, Dedicated Group
of
NLP Researchers and
Software Developers

Maintaining Quality

Open Source Community

- Test cTAKES
- Report and Fix Bugs
- Add Enhancements
- Provide email Help
- Write Documentation
- Contribute to Official Site
- Contribute Entire Modules



Core Group handles most of the Effort

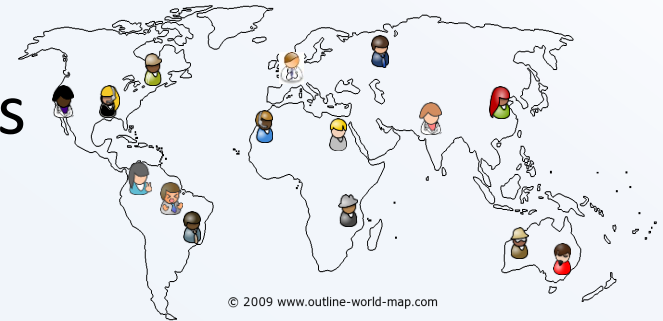
Funding



Apache Provides Support

Open Source Community Contributes

The Core Group is not free ...

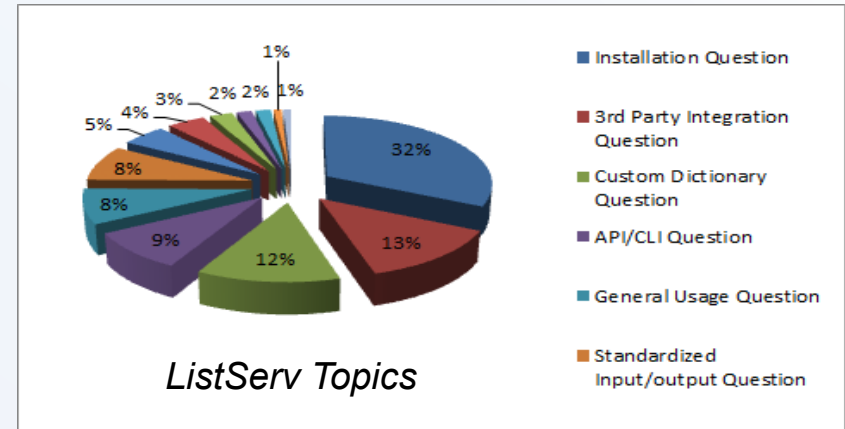


Funding through

- Grants, with Specific Aims on improving Clinical NLP
- Running cTAKES for large and small Funded Projects
- No Funding from our Institutions
 - But we are working on it ...
- Creating a Data Distribution Center
- Creating a Processing Center

Challenges

- Answering Questions
 - Wait for Community
- Fixing
 - Bait the Community
- Enhancing
 - Ask Community for similarities
 - Create seed, ask Community for assistance
 - Offer to absorb 3rd party tools as parts of cTAKES
 - yTEX NLP Tools
 - Scrubber De-Identification
- Funding
 - See previous slide

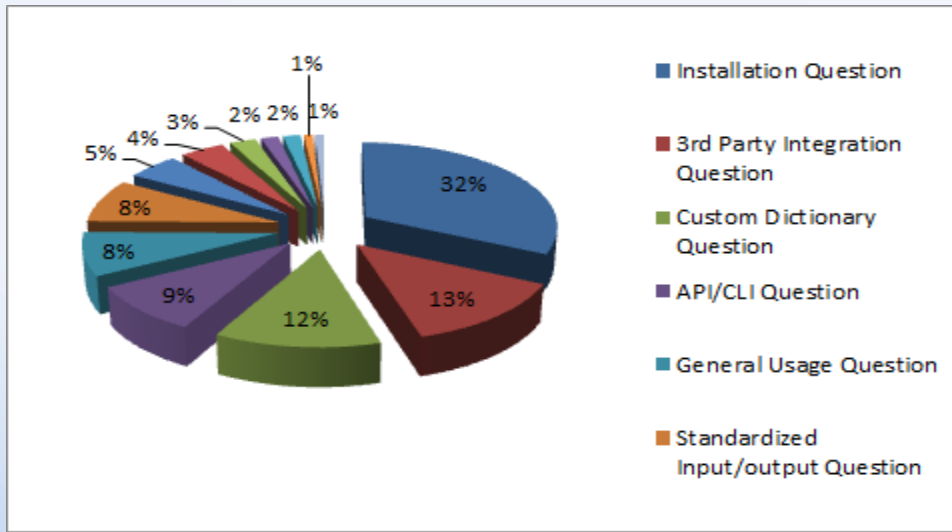


Questions ?



User Community

- Types of users
 - Developers
 - End users -- biomedical investigators, data warehouse managers, point-of-care clinicians
 - Active user and developer mailing lists



Applications

- Patient cohort identification from the EMR – eMERGE, PGRN, i2b2
- Analysis of rare diseases – Phelan McDermid Syndrome
- Pharmacovigilance – adverse events from twitter, MTX liver toxicity
- Patient-facing applications – patient-interpretable clinical notes
- Point-of-care – summarization
- Question-answering
- Quality metrics

Apache cTAKES

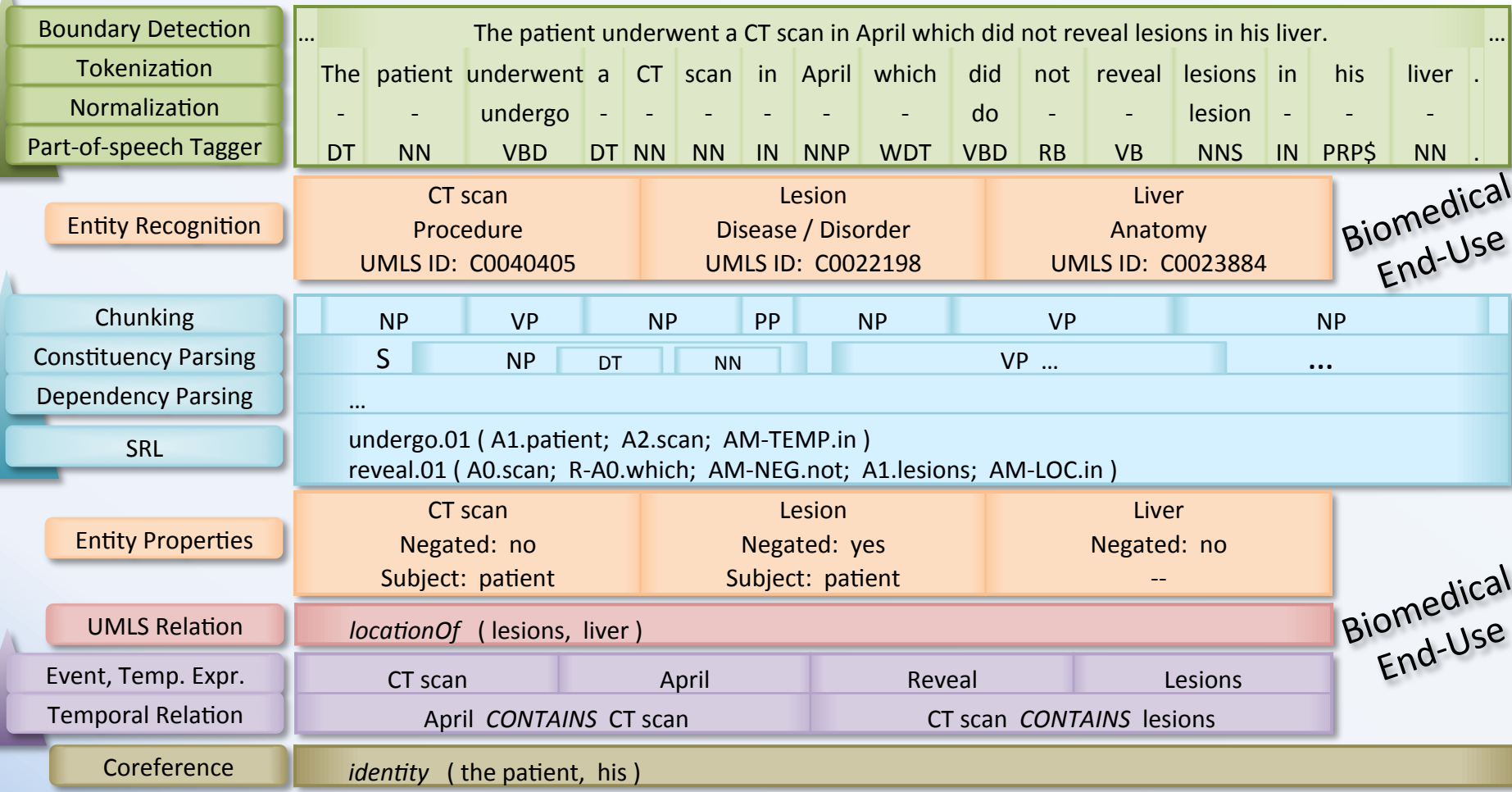
- Natural language processing system (NLP)
- Extracts information from electronic medical record clinical text
- Top level open source project in the Apache Software Foundation
- Diverse and Global User Base

Key Features

- **Powerful:** discover codable entities, events, attributes and relations
- **Fast:** process batches at 50,000 clinical notes per hour
- **Scalable:** run on clusters, queue systems and cloud computing services
- **Modular:** use only the components needed
- **Portable:** run on any major computer platform
- All machine learning models are trained on expert-annotated gold standard data

Sample Pipeline

The patient underwent a CT scan in April which did not reveal lesions in his liver.



Sample Pipeline part 1

The patient underwent a CT scan in April which did not reveal lesions in his liver.

Boundary
Detection

... The patient underwent a CT scan in April which did not reveal lesions in his liver. ...

Tokenization

The	patient	underwent	a	CT	scan	in	April	which	did	not	reveal	lesions	in	his	liver	.
-----	---------	-----------	---	----	------	----	-------	-------	-----	-----	--------	---------	----	-----	-------	---

Normalization

-	-	undergo	-	-	-	-	-	-	do	-	-	lesion	-	-	-
---	---	---------	---	---	---	---	---	---	----	---	---	--------	---	---	---

Part-of-speech
Tagging

DT	NN	VBD	DT	NN	NN	IN	NNP	WDT	VBD	RB	VB	NNS	IN	PRP\$	NN	.
----	----	-----	----	----	----	----	-----	-----	-----	----	----	-----	----	-------	----	---

Sample Pipeline part 1

The patient underwent a CT scan in April which did not reveal lesions in his liver.

Boundary Detection

Tokenization

Normalization

Part-of-speech Tagger

...	The patient underwent a CT scan in April which did not reveal lesions in his liver.															...
The	patient	underwent	a	CT	scan	in	April	which	did	not	reveal	lesions	in	his	liver	.
-	-	undergo	-	-	-	-	-	-	do	-	-	lesion	-	-	-	.
DT	NN	VBD	DT	NN	NN	IN	NNP	WDT	VBD	RB	VB	NNS	IN	PRP\$	NN	.

Entity
Recognition

CT scan

Procedure

UMLS ID: C0040405

Lesion

Disease / Disorder

UMLS ID: C0022198

Liver

Anatomy

UMLS ID: C0023884

cTAKES can normalize to domain ontologies such as SNOMED-CT and RxNORM

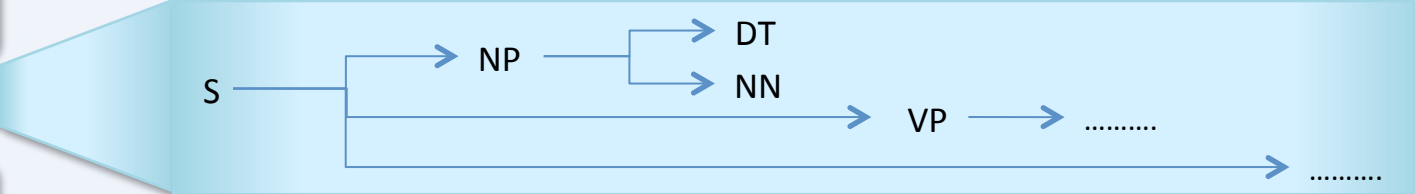
Sample Pipeline part 2

The patient underwent a CT scan in April which did not reveal lesions in his liver .
 DT NN VBD DT NN NN IN NNP WDT VBD RB VB NNS IN PRP\$ NN .

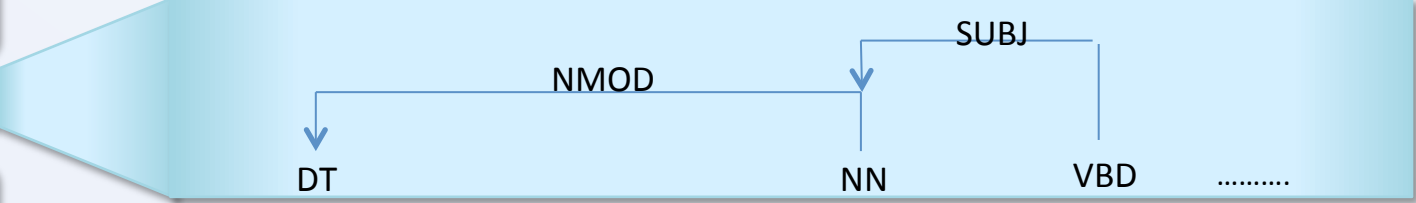
Chunking

NP VP NP PP NP VP NP

Constituency Parsing



Dependency Parsing



Semantic Role Labeling

undergo.01 (A1.patient; A2.scan; AM-TEMP.in)
 reveal.01 (A0.scan; R-A0.which; AM-NEG.not; A1.lesions; AM-LOC.in)

Sample Pipeline part 2

The patient underwent a CT scan in April which did not reveal lesions in his liver.

Chunking		NP	VP	NP	PP	NP	VP	NP
Constituency Parsing	S	NP	DT	NN		VP
Dependency Parsing	...							
SRL	undergo.01 (A1.patient; A2.scan; AM-TEMP.in) reveal.01 (A0.scan; R-A0.which; AM-NEG.not; A1.lesions; AM-LOC.in)							

Entity Properties	CT scan	Lesion	Liver
	Negated: no	Negated: yes	Negated: no
	Subject: patient	Subject: patient	--

UMLS Relation	<i>locationOf</i> (lesions, liver)
---------------	--------------------------------------

Sample Pipeline part 3

The patient underwent a CT scan in April which did not reveal lesions in his liver.

Events, Temporal Expressions

CT scan

April

Reveal

Lesions

Temporal Relations

April *CONTAINS* CT scan

CT scan *CONTAINS* lesions

Coreferences

identity (the patient, his)

Clinical Element Model Template



Sign/Symptom

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

Procedure

Associated Code	Method
Body Laterality	<i>Negation Indicator</i>
Body Location	Relative Temporal
Body Side	Context
Conditional	Start Date
Device	<i>Subject</i>
End Date	<i>Uncertainty Indicator</i>
<i>Generic</i>	

Disease/Disorder

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

Lab

Abnormal	Lab Value
Interpretation	<i>Negation Indicator</i>
Associated Code	Ordinal Interpretation
Conditional	Reference Range
Delta Flag	Narrative
Estimated flag	<i>Subject</i>
<i>Generic</i>	<i>Uncertainty Indicator</i>

Anatomical Site

Associated Code	<i>Generic</i>
Body Laterality	<i>Negation Indicator</i>
Body Site	<i>Subject</i>
Conditional	<i>Uncertainty Indicator</i>

Medication

Associated Code	<i>Generic</i>
Change Status	<i>Negation Indicator</i>
Conditional	Route
Dosage	Start Date
Duration	Strength
End Date	<i>Subject</i>
Form	<i>Uncertainty Indicator</i>
Frequency	





FHIR Composition representing the document

```
Source of: file:///home/tseytin/Dropbox/Work/DeepPhe/data/sample/fh...
1 <?xml version="1.0" encoding="UTF-8"?>
2
3 <Composition xmlns="http://hl7.org/fhir">
4   <language value="English"/>
5   <text>
6     <status value="generated"/>
7     <pre xmlns="http://www.w3.org/1999/xhtml">-----
8 Patient Name.....Jane Doe
9 Principal Date.....20130118 1050
10 Record Type.....SP
11 -----
12 BREAST, LEFT, EXCISION
13 INVASIVE DUCTAL CARCINOMA, 2.1 CM
14 Sentinel Lymph Node Biopsy,
15 One LN with no evidence of Carcinoma
16 </pre>
17 </text>
18 <identifier>
19   <label value="id"/>
20   <system value="local"/>
21   <value value="Report-1805009648"/>
22 </identifier>
23 <date value="2013-01-18T10:50:00-05:00"/>
24 <type>
25   <coding>
26     <system value="UMLS"/>
27     <code value="C0807321"/>
28     <display value="Pathology Report"/>
29   </coding>
30   <text value="Pathology Report"/>
31 </type>
32 <title value="doc1.txt"/>
33 <status value="final"/>
34 <subject>
35   <reference value="Patient-1839436020"/>
36   <display value="Jane Doe"/>
37 </subject>
38 <event>
39   <detail>
40     <reference value="Observation-979976544"/>
41     <display value="Tumor Size"/>
42   </detail>
43   <detail>
44     <reference value="Procedure-1633134782"/>
45     <display value="Excision"/>
46   </detail>
47   <detail>
48     <reference value="Procedure-1107788767"/>
49     <display value="Sentinel Lymph Node Biopsy"/>
50   </detail>
51   <detail>
52     <reference value="Diagnosis-1472569260"/>
53     <display value="Invasive Ductal Carcinoma, Not Otherwise Specified"/>
54   </detail>
55 </event>
56 </Composition>
```

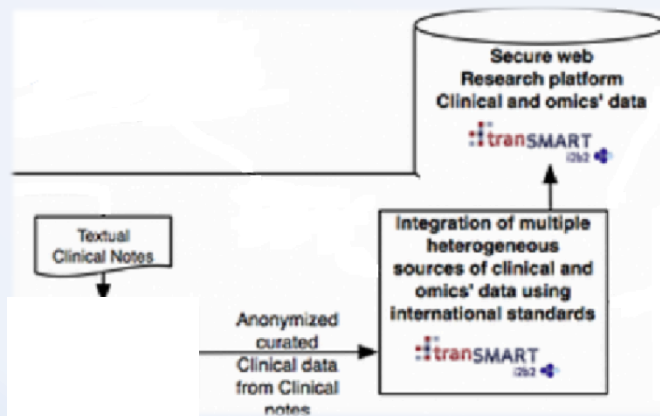
Line 56, Col 11



System Integration

cTAKES has been integrated with

- i2b2 platform
- TranSMART



Partial TranSMART environment



i2b2 platform Cells

END

Measures of Success

- Goal – enable groundbreaking medical research
- Publications using the tool
- Mailing list activity around the tool

Sustainability Plans

- The Apache Community
- In an active state of fund-raising

Lower Entry Barrier

- Based on community feedback
 - Improve ease of use
 - API
 - GUI
 - Documentation and how to videos

Community Contributions

- Apache Software Foundation principles
 - Do-ocracy
 - Community testing