Streamlining rule-based NLP development and integration with production environment

HEALTH
UNIVERSITY OF UTAH

Jianlin Shi, MD, PhD

A background story

- 1st operational NLP project for billing assistant
- Collaborate with EDW, clinical analysts

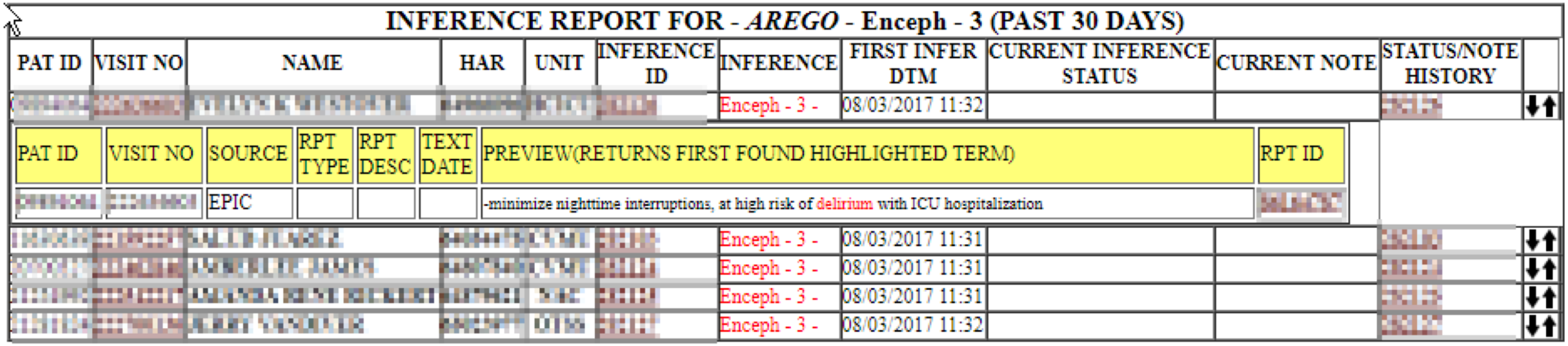
Use case

- **Encephalopathy**: A broad term for any brain disease that alters brain function or structure.
- **Missing code** means **revenue loss**
- **Old way:** database keywords search + manual review

Old Way

(encephalopathy, altered mental, delirium, confusion, AMS, confused, confussed, encephalopathic) ~near((no, encephalopathy), 5, TRUE) ~near(("not", encephalopathy), 8, TRUE) ~near((denies, encephalopathy), 5, TRUE) ~near((negative, encephalopathy), 5, TRUE) ~near((rule out, encephalopathy), 3, TRUE) ~near((mild, encephalopathy), 3, TRUE) ~near((borderline, encephalopathy), 5, TRUE)...

A web-based interface for code review (Warthog)



INFERENCE REPORT FOR - *AREGO* - Enceph - 3 (PAST 30 DAYS)

PAT ID	VISIT NO	NAME	HAR	UNIT	INFERENCE ID	INFERENCE	FIRST INFER DTM	CURRENT INFERENCE STATUS	CURRENT NOTE	STATUS/NOTE HISTORY	
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	Enceph - 3 -	08/03/2017 11:32			[redacted]	⬇⬆
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	Enceph - 3 -	08/03/2017 11:31			[redacted]	⬇⬆
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	Enceph - 3 -	08/03/2017 11:31			[redacted]	⬇⬆
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	Enceph - 3 -	08/03/2017 11:31			[redacted]	⬇⬆
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	Enceph - 3 -	08/03/2017 11:32			[redacted]	⬇⬆

PAT ID	VISIT NO	SOURCE	RPT TYPE	RPT DESC	TEXT DATE	PREVIEW(RETURNS FIRST FOUND HIGHLIGHTED TERM)	RPT ID
[redacted]	[redacted]	EPIC				-minimize nighttime interruptions, at high risk of delirium with ICU hospitalization	[redacted]

Questions were asked:

- Where to host the program ?
- How to add new inputs?
- How to trigger the execution?
- How to retrieve the NLP outputs?
- How much computing power is needed?
- Do we need redo everything for another project?

.

.

.

Questions were asked:

- Where to host the program ?
- How to add new inputs?
- How to trigger the execution?
- How to retrieve the NLP outputs?
- How much computing power is needed?
- Do we need redo everything for another project?

.

.

.

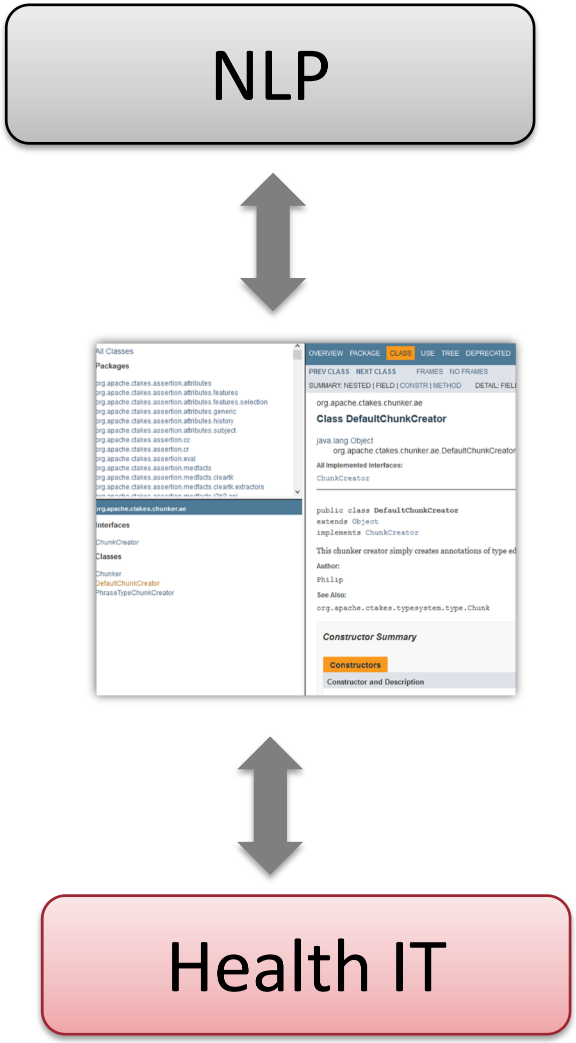
Research NLP VS Operational NLP

- Reusability — Development cost
- Computing cost
- Interoperability — with EHR and others
- Maintenance
- User support

One for “all” solution regarding

- Interaction with Health IT systems
- Supporting different NLP tasks—with extensibility

APIs



All Classes

Packages

org.apache.ctakes.assertion.attributes
org.apache.ctakes.assertion.attributes.features
org.apache.ctakes.assertion.attributes.features.selection
org.apache.ctakes.assertion.attributes.generic
org.apache.ctakes.assertion.attributes.history
org.apache.ctakes.assertion.attributes.subject
org.apache.ctakes.assertion.cc
org.apache.ctakes.assertion.cr
org.apache.ctakes.assertion.eval
org.apache.ctakes.assertion.medfacts
org.apache.ctakes.assertion.medfacts.cleartk
org.apache.ctakes.assertion.medfacts.cleartk.extractors

org.apache.ctakes.chunker.ae

Interfaces

ChunkCreator

Classes

Chunker
DefaultChunkCreator
PhraseTypeChunkCreator

OVERVIEW PACKAGE CLASS USE TREE DEPRECATED

PREV CLASS NEXT CLASS FRAMES NO FRAMES

SUMMARY: NESTED | FIELD | CONSTR | METHOD DETAIL: FIEL

org.apache.ctakes.chunker.ae

Class DefaultChunkCreator

java.lang.Object
    org.apache.ctakes.chunker.ae.DefaultChunkCreator

All Implemented Interfaces:

ChunkCreator

```
public class DefaultChunkCreator
extends Object
implements ChunkCreator
```

This chunker creator simply creates annotations of type ed

Author:

Philip

See Also:

org.apache.ctakes.typesystem.type.Chunk

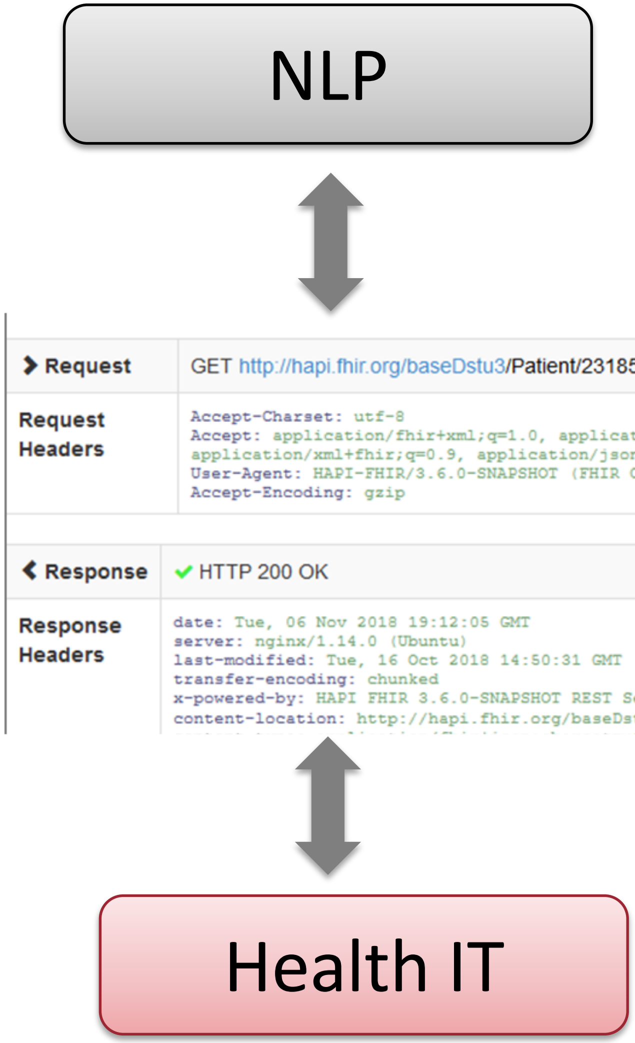
Constructor Summary

Constructors

Constructor and Description

- Powerful to access low level functionalities
- Language constrains
- Expert only
- Maintenance strait

Network Protocols



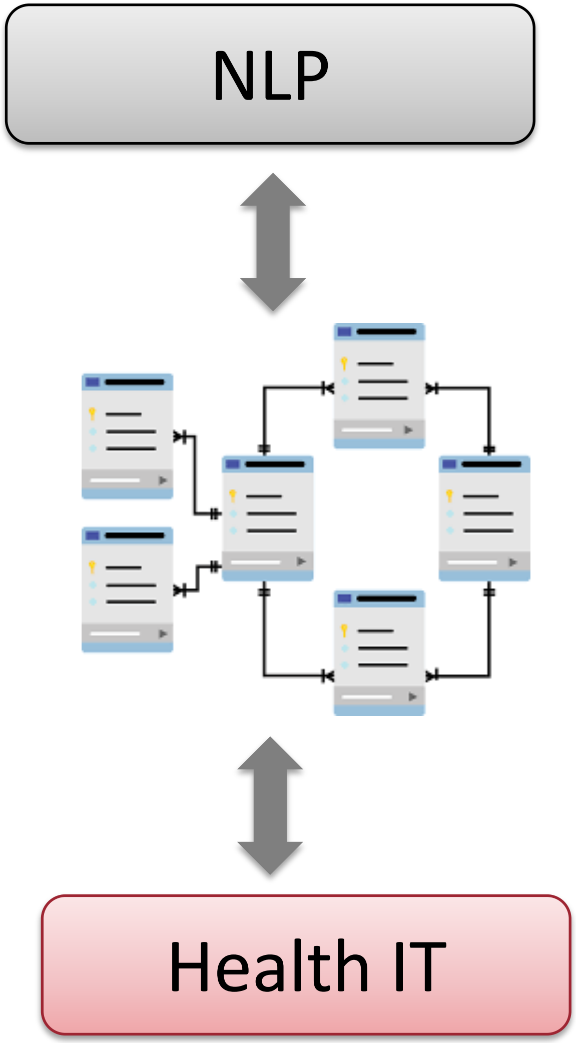
NLP

❯ Request	GET http://hapi.fhir.org/baseDstu3/Patient/231854/_history/
Request Headers	Accept-Charset: utf-8 Accept: application/fhir+xml;q=1.0, application/fhir+; application/xml+fhir;q=0.9, application/json+fhir;q=0. User-Agent: HAPI-FHIR/3.6.0-SNAPSHOT (FHIR Client; FHI Accept-Encoding: gzip
❮ Response	✓ HTTP 200 OK
Response Headers	date: Tue, 06 Nov 2018 19:12:05 GMT server: nginx/1.14.0 (Ubuntu) last-modified: Tue, 16 Oct 2018 14:50:31 GMT transfer-encoding: chunked x-powered-by: HAPI FHIR 3.6.0-SNAPSHOT REST Server (FHI content-location: http://hapi.fhir.org/baseDstu3/Patien

Health IT

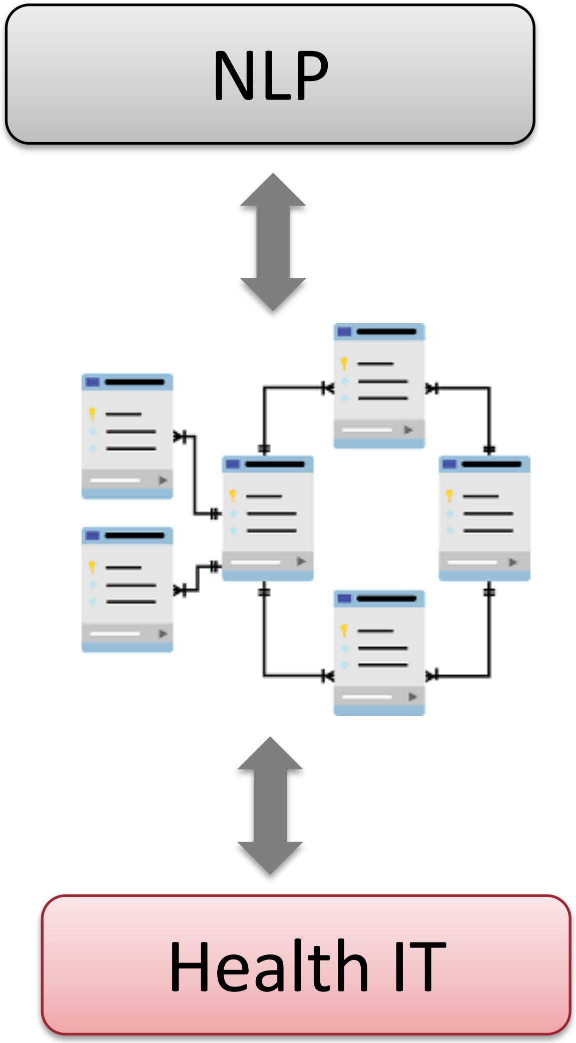
- Language independent
- Additional storage still needed
- Need client-side implementations
- Maintenance cost

Database



- Naturally integrated with EHR, EDW
- No technical barrier for EDW staff
- No programming workload for EDW staff
- No additional hurdle for current EDW users

Database

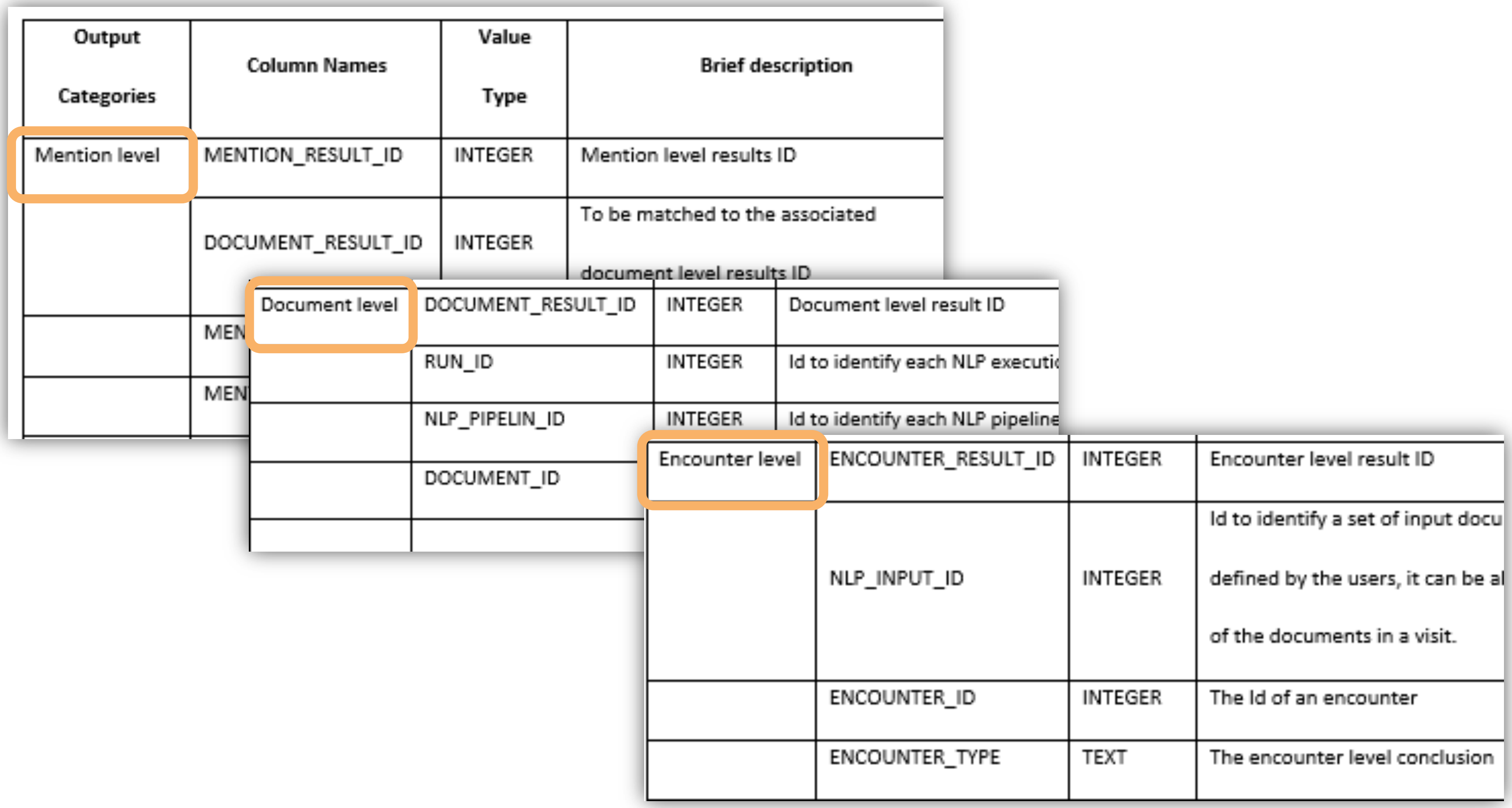


- Need a unified table structure / schema

Categories of Clinical NLP Tasks' Output

Output Categories	Description
Mention level	Identify the concepts or statements in a document
Document level	Classify a document based on the given conditions
Encounter level	Classify an encounter based on the given conditions
Patient level	Classify a patient based on the given conditions

Harmonized Database Schema for NLP output



Output Categories	Column Names	Value Type	Brief description
Mention level	MENTION_RESULT_ID	INTEGER	Mention level results ID
	DOCUMENT_RESULT_ID	INTEGER	To be matched to the associated document level results ID
	MEN		
	MEN		

Document level	DOCUMENT_RESULT_ID	INTEGER	Document level result ID
	RUN_ID	INTEGER	Id to identify each NLP executi
	NLP_PIPELIN_ID	INTEGER	Id to identify each NLP pipeline
	DOCUMENT_ID		

Encounter level	ENCOUNTER_RESULT_ID	INTEGER	Encounter level result ID
	NLP_INPUT_ID	INTEGER	Id to identify a set of input docu defined by the users, it can be al of the documents in a visit.
	ENCOUNTER_ID	INTEGER	The Id of an encounter
	ENCOUNTER_TYPE	TEXT	The encounter level conclusion

Harmonized Database Schema for NLP input

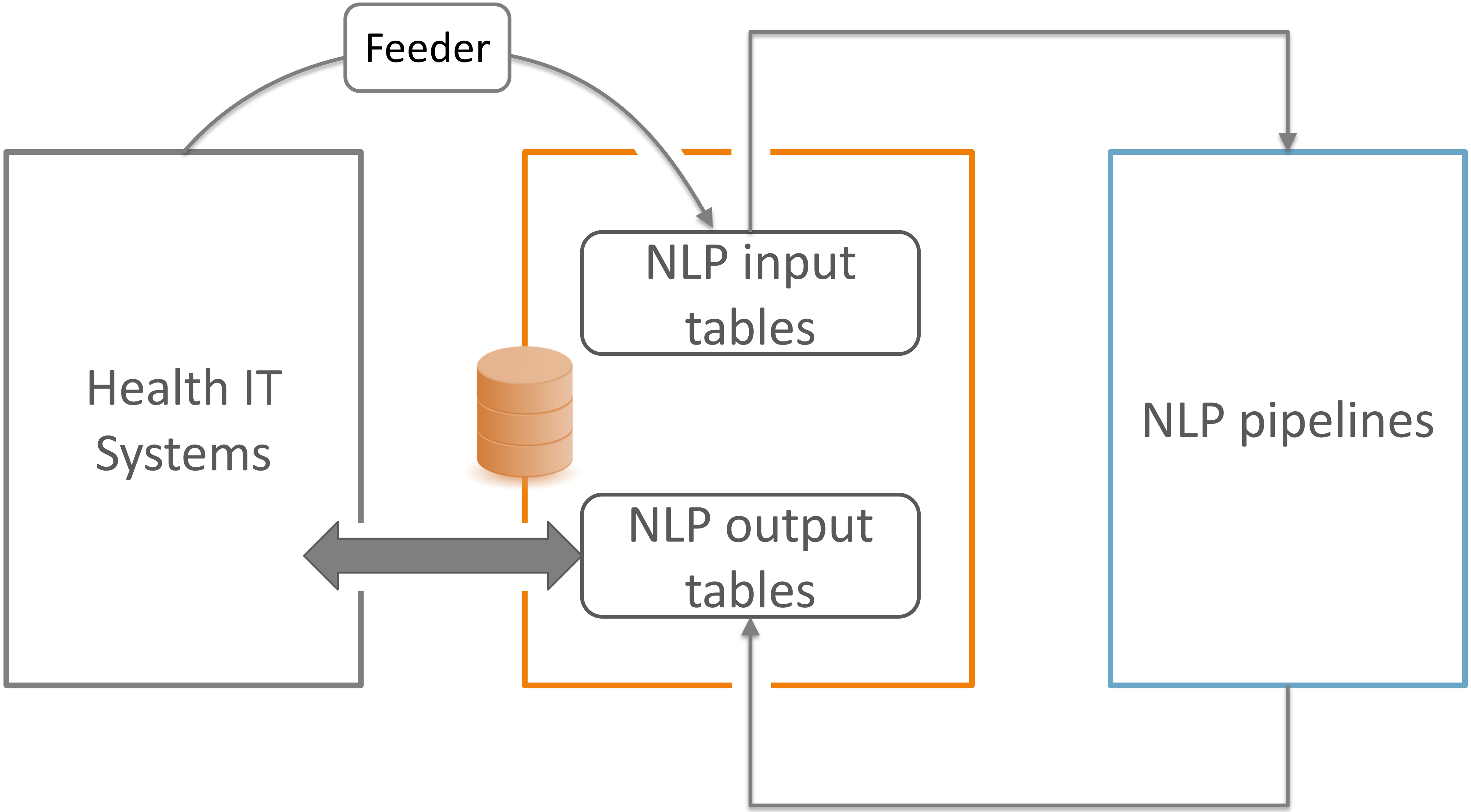
NLP input cohort table

Column Name	Value type	Brief Description
NLP_INPUT_ID	Integer	An ID for a group of documents that needs to be processed
PIPELINE_ID	Integer	NLP pipeline ID
…..		

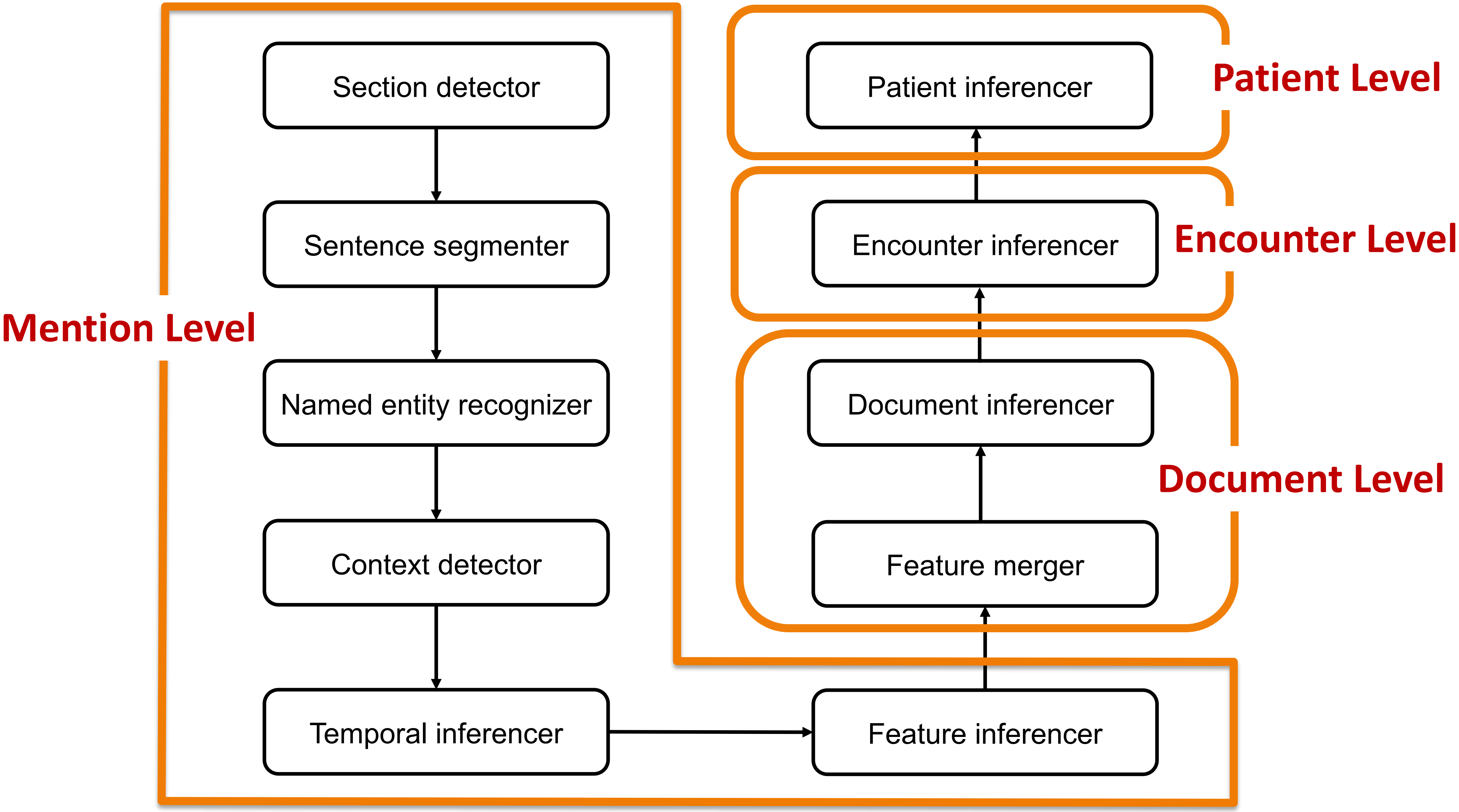
Column Name	Value type	Brief Description
NLP_INPUT_ID	Integer	An ID for a group of documents that needs to be processed
NOTE_ID	Integer	Note ID
…..		

NLP input document table

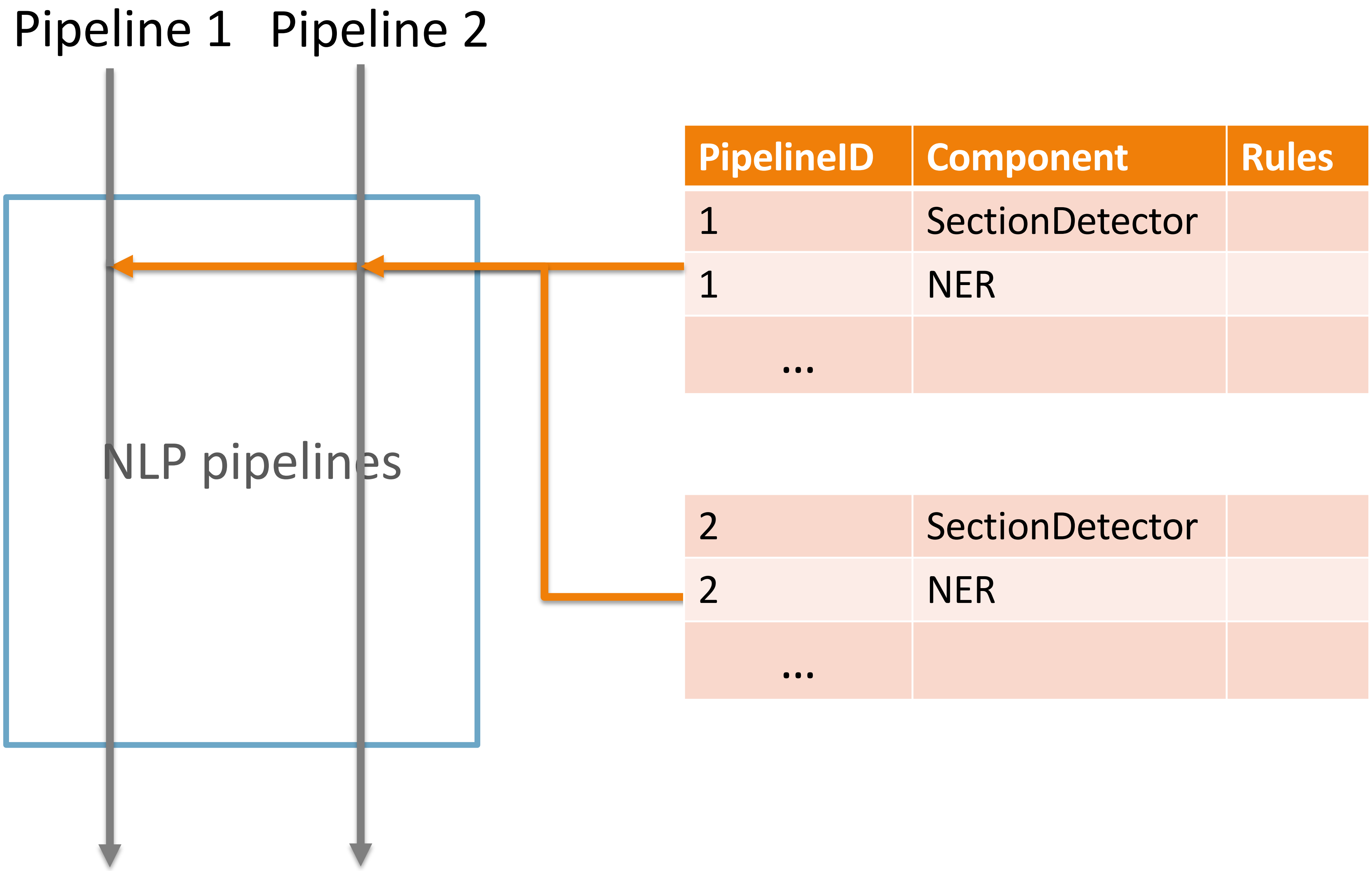
Overall setup



Workflow of a Generic Rule-based NLP Pipeline

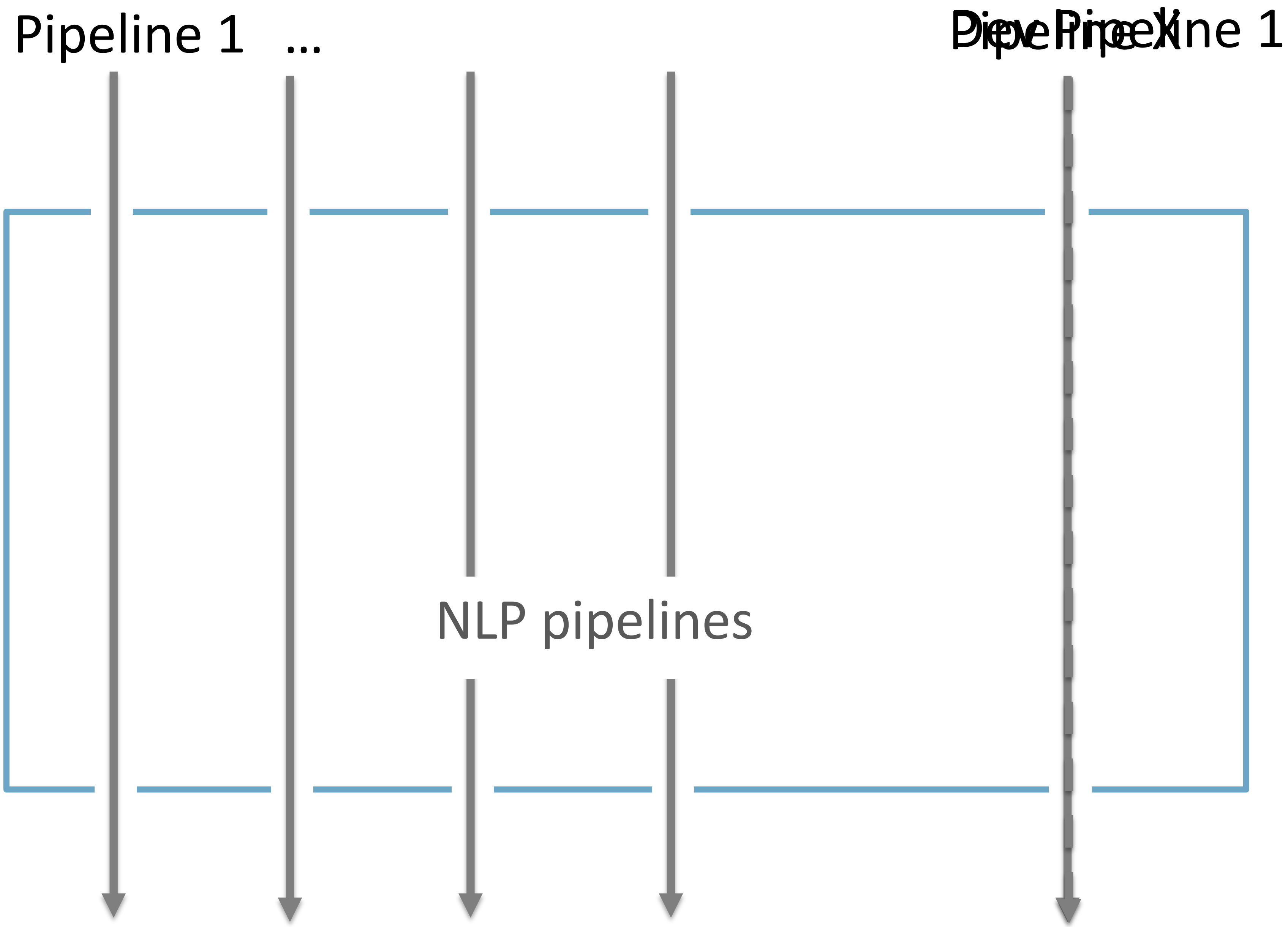


One for “all” rule-based NLP

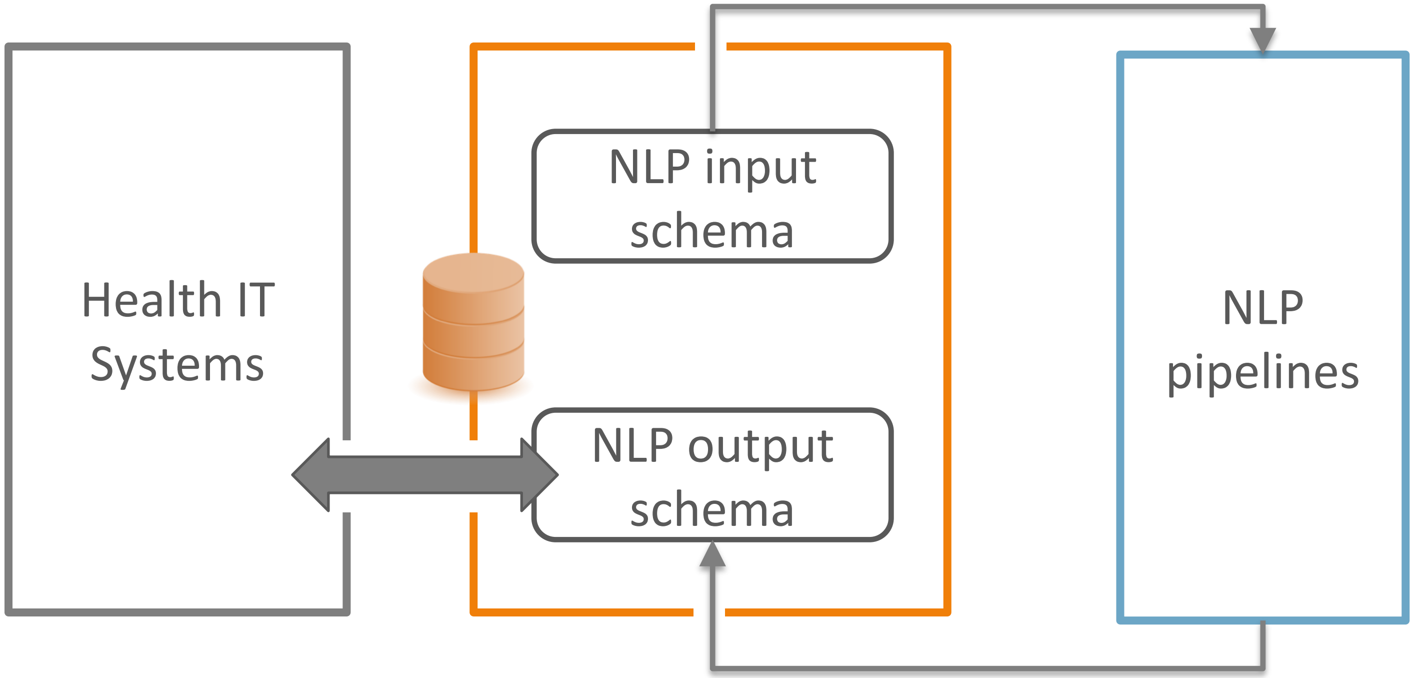


PipelineID	Component	Rules
1	SectionDetector	
1	NER	
…		
2	SectionDetector	
2	NER	
…		

From Dev To Production



Overall setup



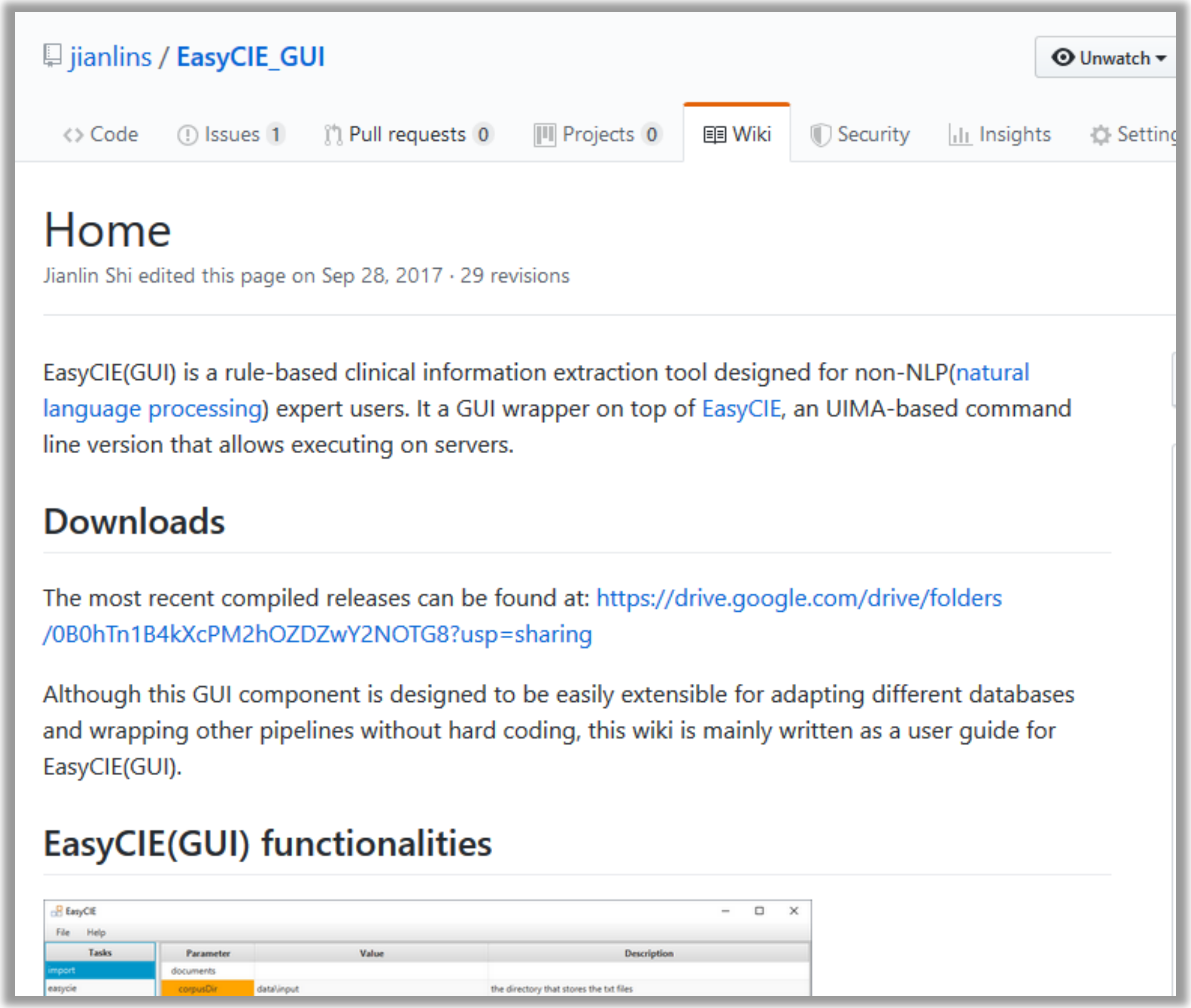
Rule Configuration

PipelineID	Component	Rules
1	SectionDetector	
1	NER	
…		

Rule development needs

- Focus on rule development rather than coding
- Operate with different databases
- Support debugging rules
- …

Tool for Rule Development



Home

Jianlin Shi edited this page on Sep 28, 2017 · 29 revisions

EasyCIE(GUI) is a rule-based clinical information extraction tool designed for non-NLP(natural language processing) expert users. It a GUI wrapper on top of EasyCIE, an UIMA-based command line version that allows executing on servers.

Downloads

The most recent compiled releases can be found at: https://drive.google.com/drive/folders/0B0hTn1B4kXcPM2hOZDZwY2NOTG8?usp=sharing

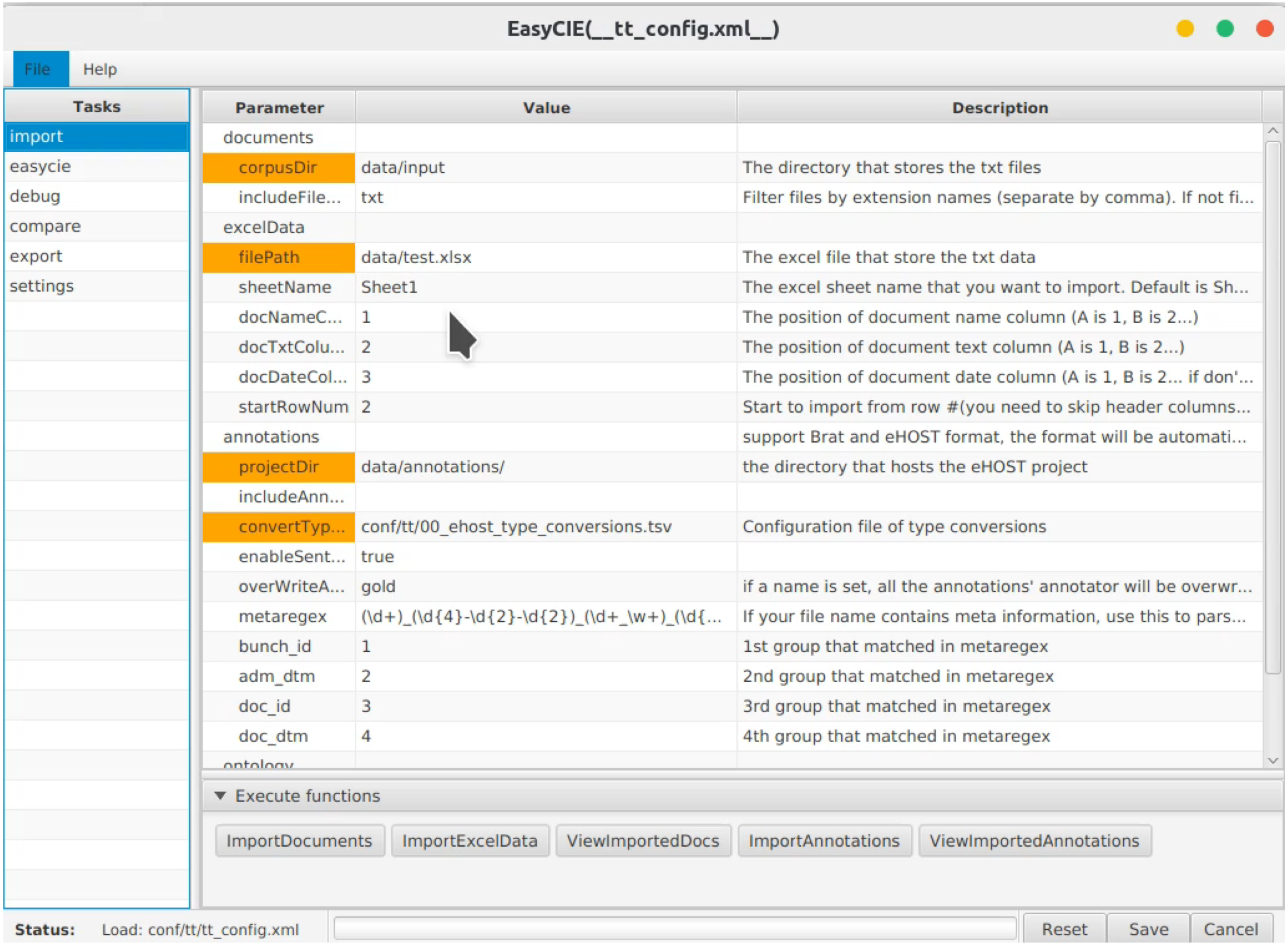
Although this GUI component is designed to be easily extensible for adapting different databases and wrapping other pipelines without hard coding, this wiki is mainly written as a user guide for EasyCIE(GUI).

EasyCIE(GUI) functionalities

Tasks
import
easycie

Parameter	Value	Description
documents		
corpusDir	data\input	the directory that stores the txt files

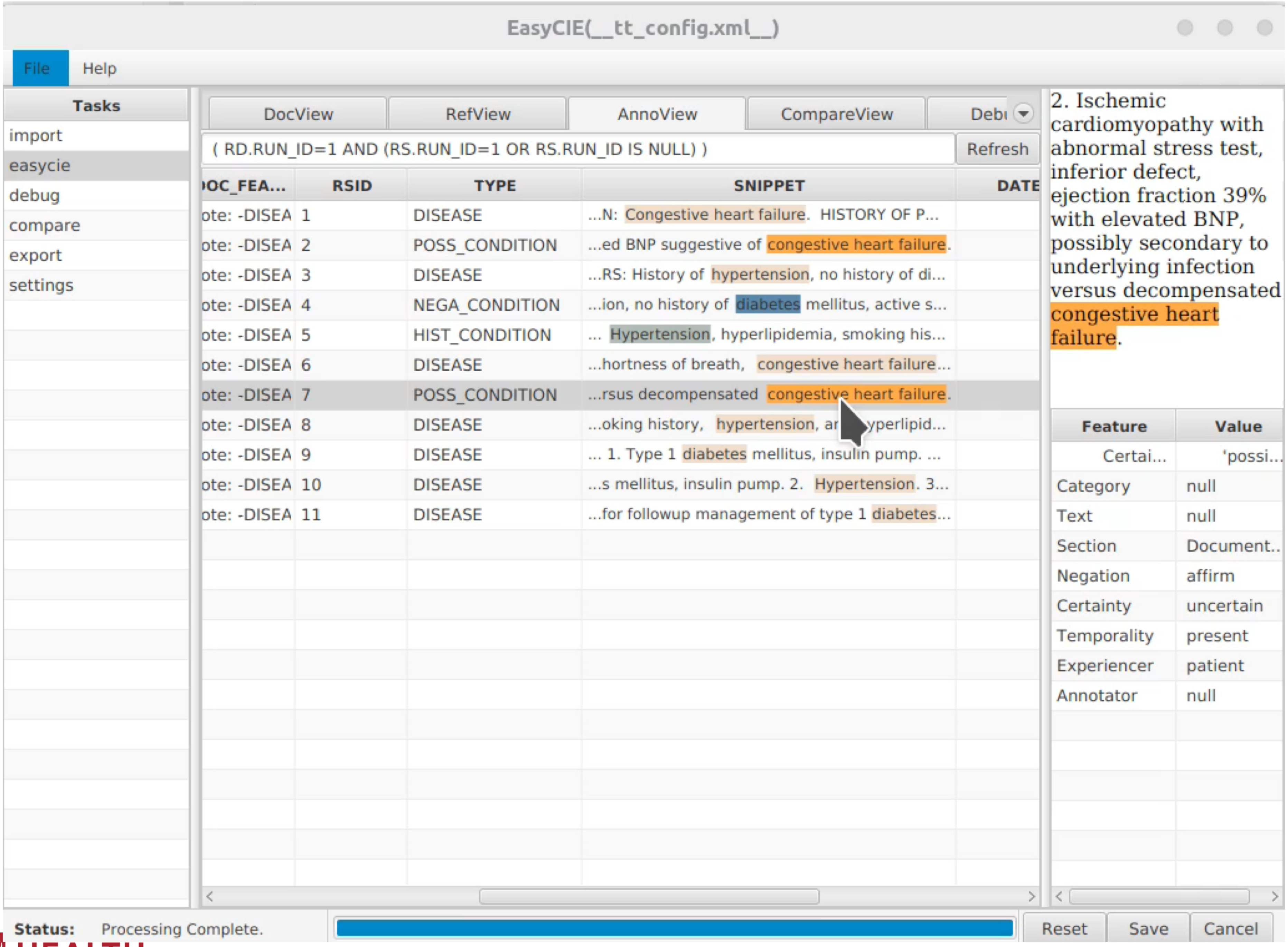
Import and Execute pipeline



Parameter	Value	Description
documents		
corpusDir	data/input	The directory that stores the txt files
includeFile...	txt	Filter files by extension names (separate by comma). If not fi...
excelData		
filePath	data/test.xlsx	The excel file that store the txt data
sheetName	Sheet1	The excel sheet name that you want to import. Default is Sh...
docNameC...	1	The position of document name column (A is 1, B is 2...)
docTxtColu...	2	The position of document text column (A is 1, B is 2...)
docDateCol...	3	The position of document date column (A is 1, B is 2... if don'...
startRowNum	2	Start to import from row #(you need to skip header columns...
annotations		support Brat and eHOST format, the format will be automati...
projectDir	data/annotations/	the directory that hosts the eHOST project
includeAnn...		
convertTyp...	conf/tt/00_ehost_type_conversions.tsv	Configuration file of type conversions
enableSent...	true	
overWriteA...	gold	if a name is set, all the annotations' annotator will be overwr...
metaregex	(\d+)_(\d{4}-\d{2}-\d{2})_(\d+_\w+)_(\d{...	If your file name contains meta information, use this to pars...
bunch_id	1	1st group that matched in metaregex
adm_dtm	2	2nd group that matched in metaregex
doc_id	3	3rd group that matched in metaregex
doc_dtm	4	4th group that matched in metaregex
ontology		

Status: Load: conf/tt/tt_config.xml

Debug



(RD.RUN_ID=1 AND (RS.RUN_ID=1 OR RS.RUN_ID IS NULL))

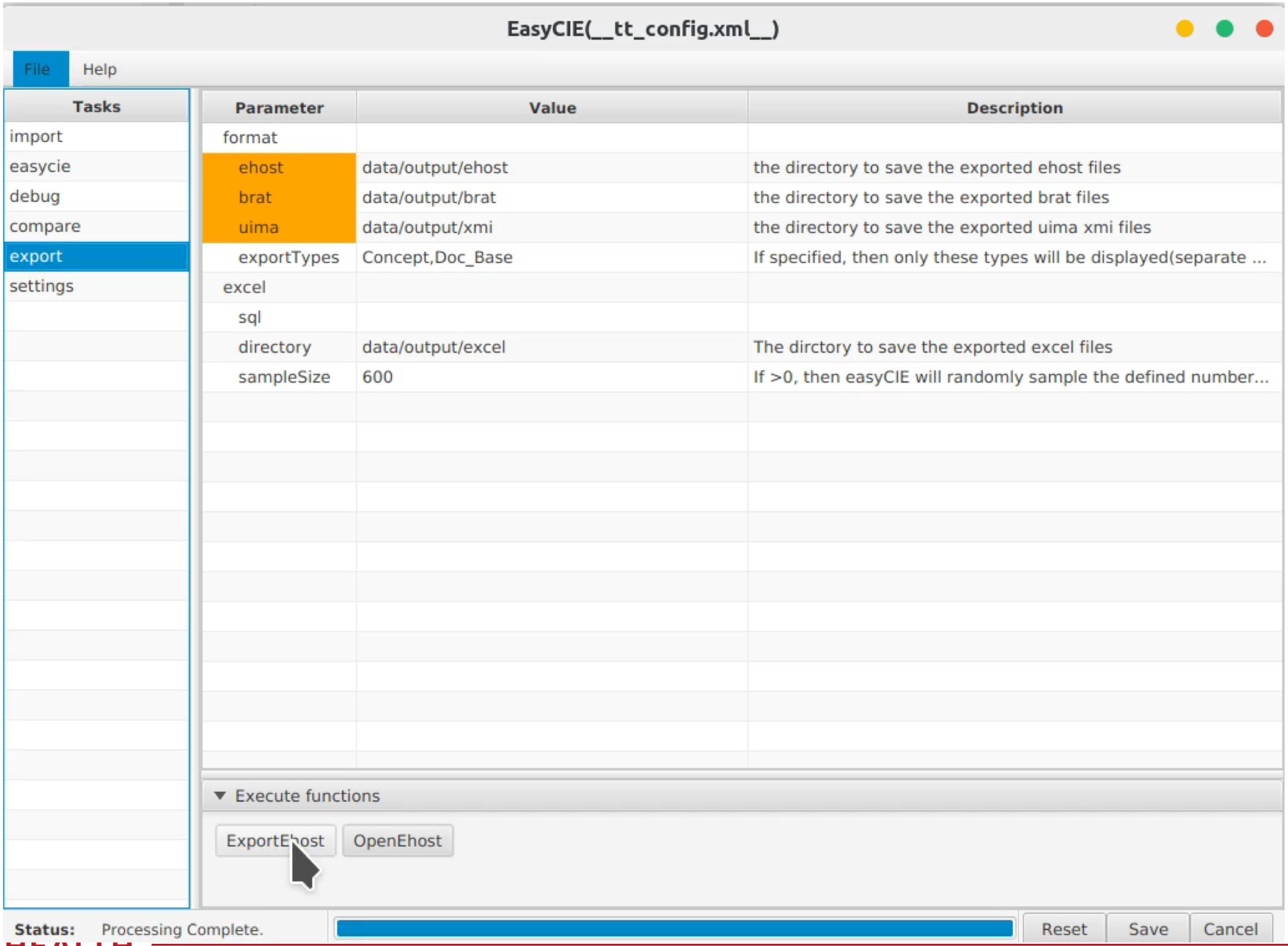
ɔOC_FEA...	RSID	TYPE	SNIPPET	DATE
ote: -DISEA	1	DISEASE	...N: Congestive heart failure. HISTORY OF P...	
ote: -DISEA	2	POSS_CONDITION	...ed BNP suggestive of congestive heart failure.	
ote: -DISEA	3	DISEASE	...RS: History of hypertension, no history of di...	
ote: -DISEA	4	NEGA_CONDITION	...ion, no history of diabetes mellitus, active s...	
ote: -DISEA	5	HIST_CONDITION	... Hypertension, hyperlipidemia, smoking his...	
ote: -DISEA	6	DISEASE	...hortness of breath, congestive heart failure...	
ote: -DISEA	7	POSS_CONDITION	...rsus decompensated congestive heart failure.	
ote: -DISEA	8	DISEASE	...oking history, hypertension, ar yperlipid...	
ote: -DISEA	9	DISEASE	... 1. Type 1 diabetes mellitus, insulin pump. ...	
ote: -DISEA	10	DISEASE	...s mellitus, insulin pump. 2. Hypertension. 3...	
ote: -DISEA	11	DISEASE	...for followup management of type 1 diabetes...	

2. Ischemic cardiomyopathy with abnormal stress test, inferior defect, ejection fraction 39% with elevated BNP, possibly secondary to underlying infection versus decompensated congestive heart failure.

Feature	Value
Certai...	'possi...
Category	null
Text	null
Section	Document..
Negation	affirm
Certainty	uncertain
Temporality	present
Experiencer	patient
Annotator	null

Status: Processing Complete.

Export

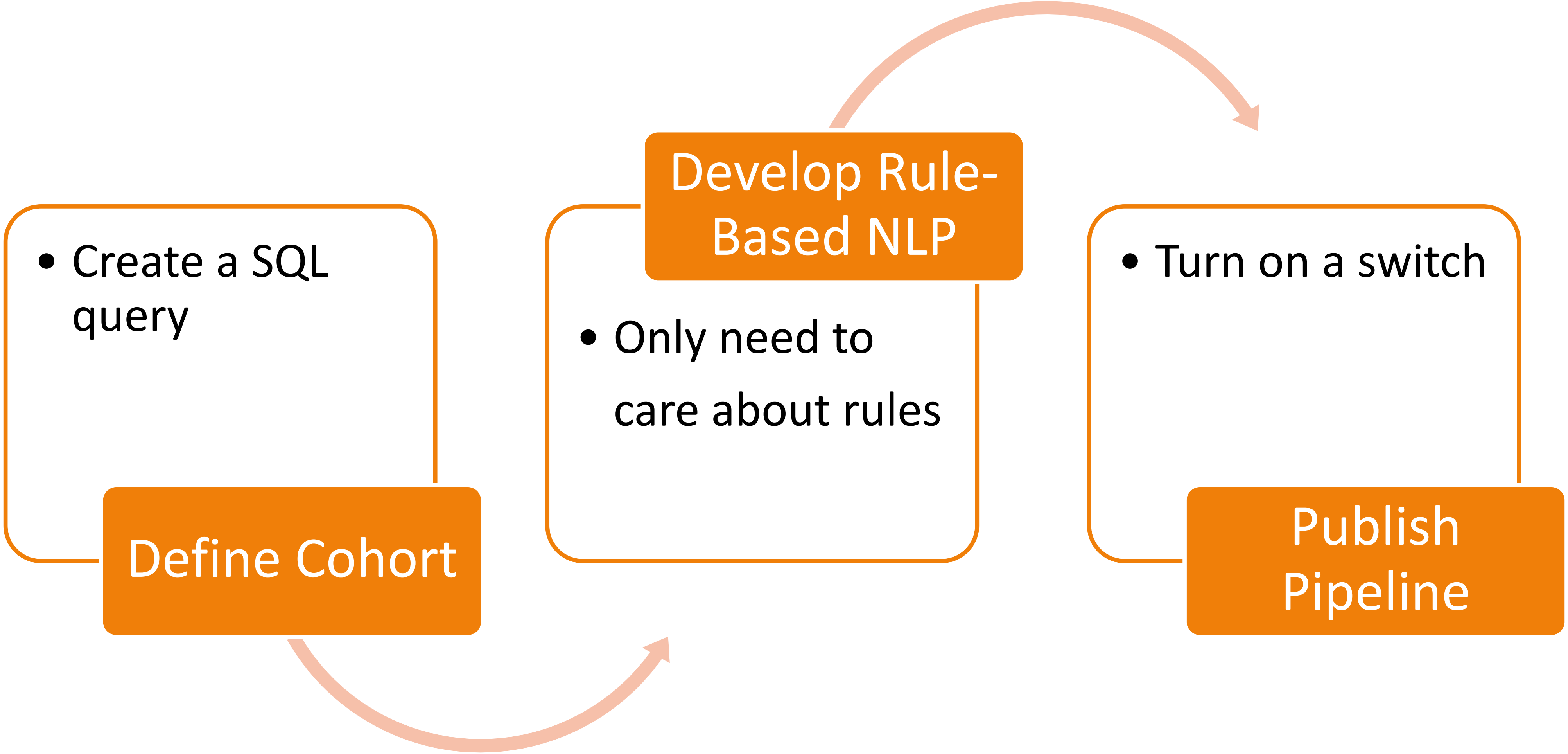


EasyCIE(__tt_config.xml__)

Parameter	Value	Description
format		
ehost	data/output/ehost	the directory to save the exported ehost files
brat	data/output/brat	the directory to save the exported brat files
uima	data/output/xmi	the directory to save the exported uima xmi files
exportTypes	Concept,Doc_Base	If specified, then only these types will be displayed(separate ...
excel		
sql		
directory	data/output/excel	The dirctory to save the exported excel files
sampleSize	600	If >0, then easyCIE will randomly sample the defined number...

Status: Processing Complete.

Streamlining rule-based NLP



Define Cohort

- Create a SQL query

Develop Rule-Based NLP

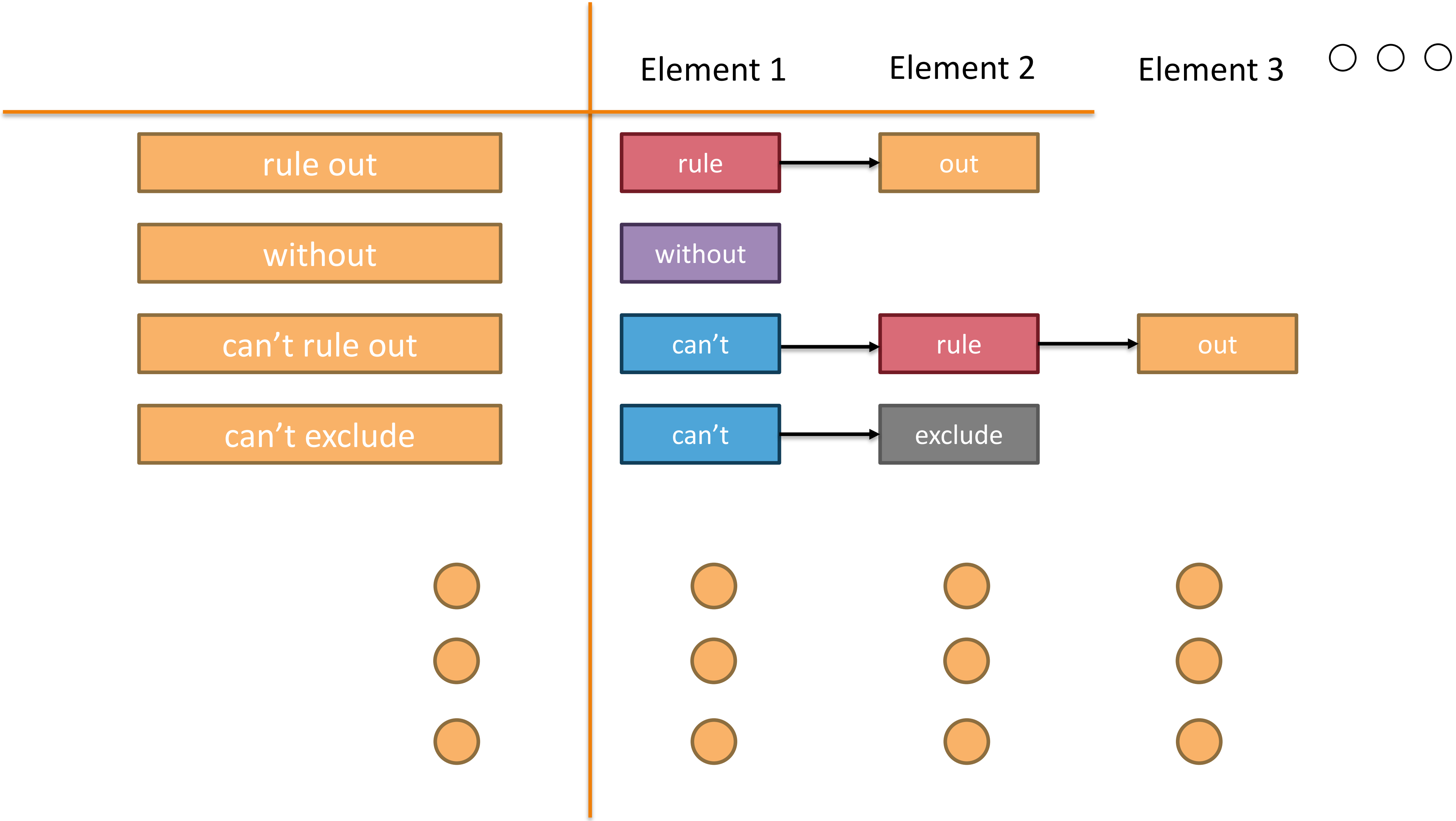
- Only need to care about rules

Publish Pipeline

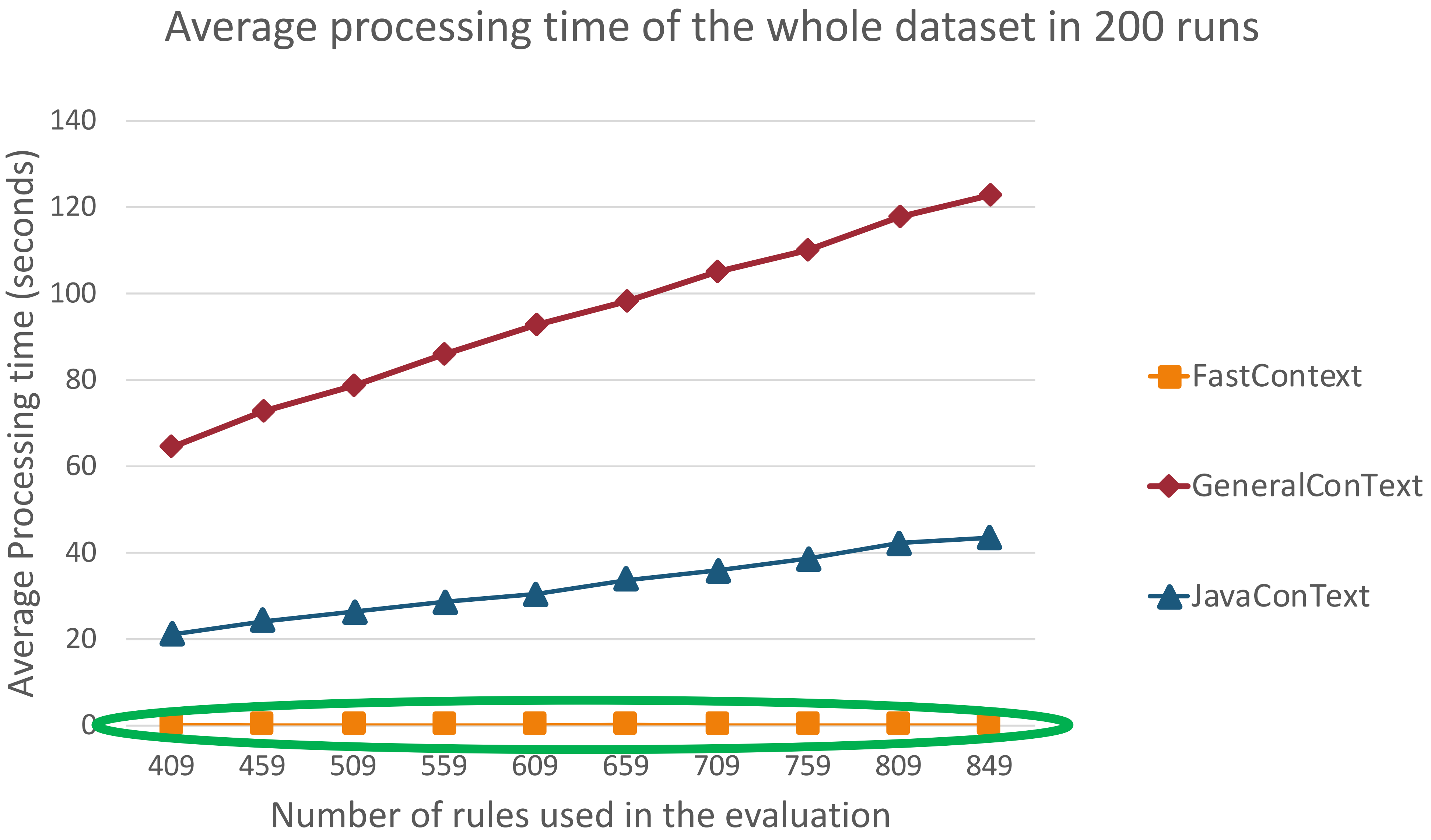
- Turn on a switch

Thank you!

n-Trie demo


Element 1
Element 2
Element 3
rule out
without
can’t rule out
can’t exclude
rule
out
without
can’t
rule
out
can’t
exclude

FastContext — Speed Evaluation



Encephalopathy pipeline evaluation

Totally sampled 665 visits,
including   8068 documents

	Found encephalopathy mentions (# of visit)	Verified by review (# of visits)
Old Warthog	50	13
NLP powered Warthog	208	178

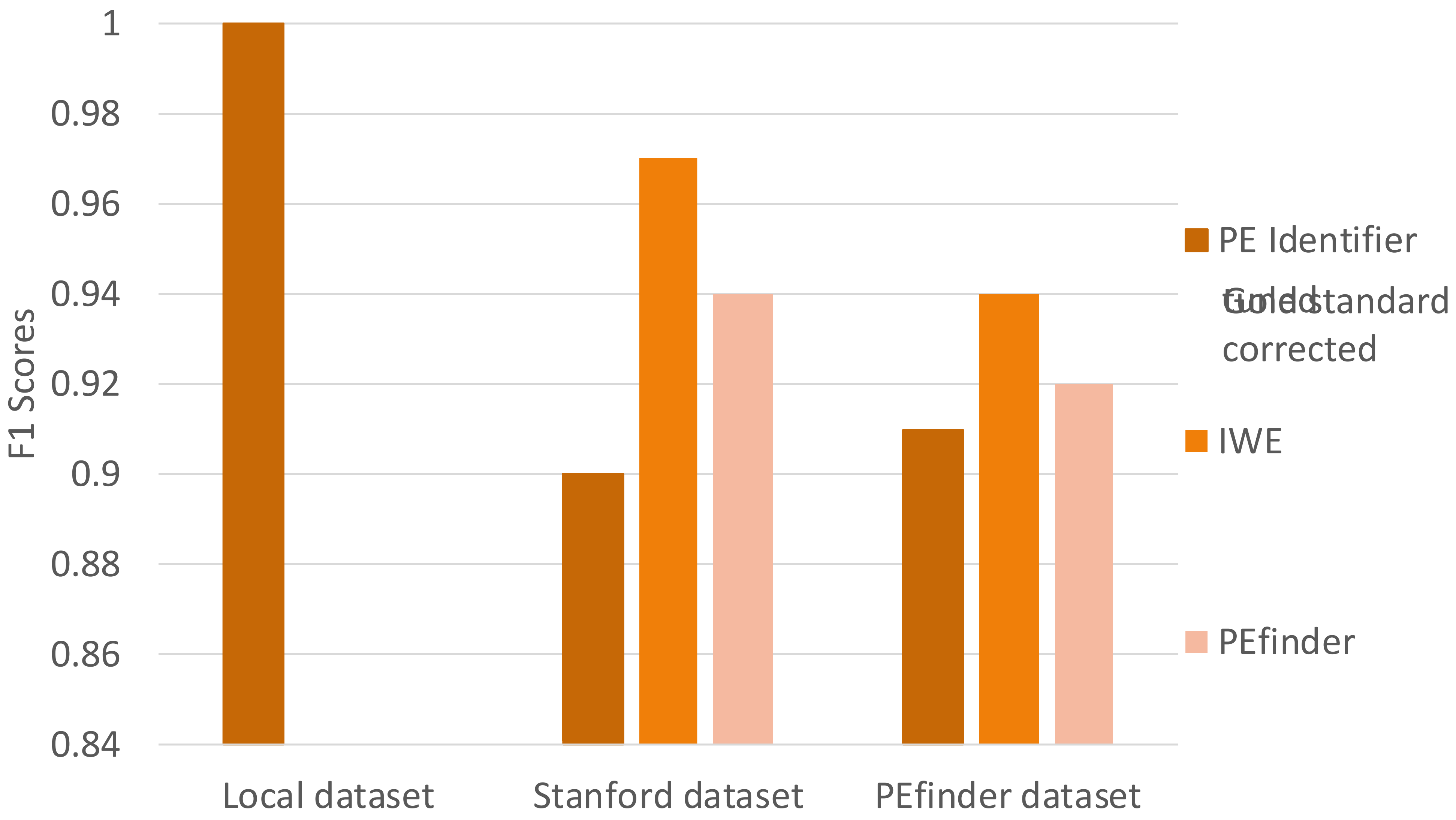
Precision improves 231%,
Recall improves 1269%

Methods---PE Identifier Evaluation

- A **local** dataset (400 annotated CT pulmonary angiography reports)
- **Stanford** dataset (944 reports)
- **PEfinder** dataset (859 reports)
- Compared with
  - Intelligent Word Embedding (IWE)[1]
  - PEfinder[2]

1. Banerjee I, Chen MC, Lungren MP, et al. Radiology report annotation using intelligent word embeddings: Applied to multi-institutional chest CT cohort. J Biomed Inform 2018;77:11–20. doi:10.1016/j.jbi.2017.11.012
2. Chapman BE, Lee S, Kang HP, et al. Document-level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm. J Biomed Inform 2011;44:728–37. doi:10.1016/j.jbi.2011.03.011

PE Identifier Evaluation


F1 Scores
Local dataset
Stanford dataset
PEfinder dataset
PE Identifier
tuned
Gold standard
corrected
IWE
PEfinder