The Language Application Grid and Galaxy

Nancy Ide, Keith Suderman
Vassar College

James Pustejovsky, Marc Verhagen
Brandeis University

Christopher Cieri
Linguistic Data Consortium

Eric Nyberg
Carnegie-Mellon University

The LAPPS Grid

- Framework for development of Natural Language Processing (NLP) applications
  - Information extraction/Question answering
    - Google search, mining information in unstructured textual data
  - Machine translation
  - Speech recognition
  - Sentiment analysis
    - “What is the current attitude toward Trump on Twitter?”
  - … and many more

LAPPS Galaxy Interface

http://galaxyproject.org

- The LAPPS Grid recently adopted the GALAXY workflow engine as a front end for construction of pipelines etc.

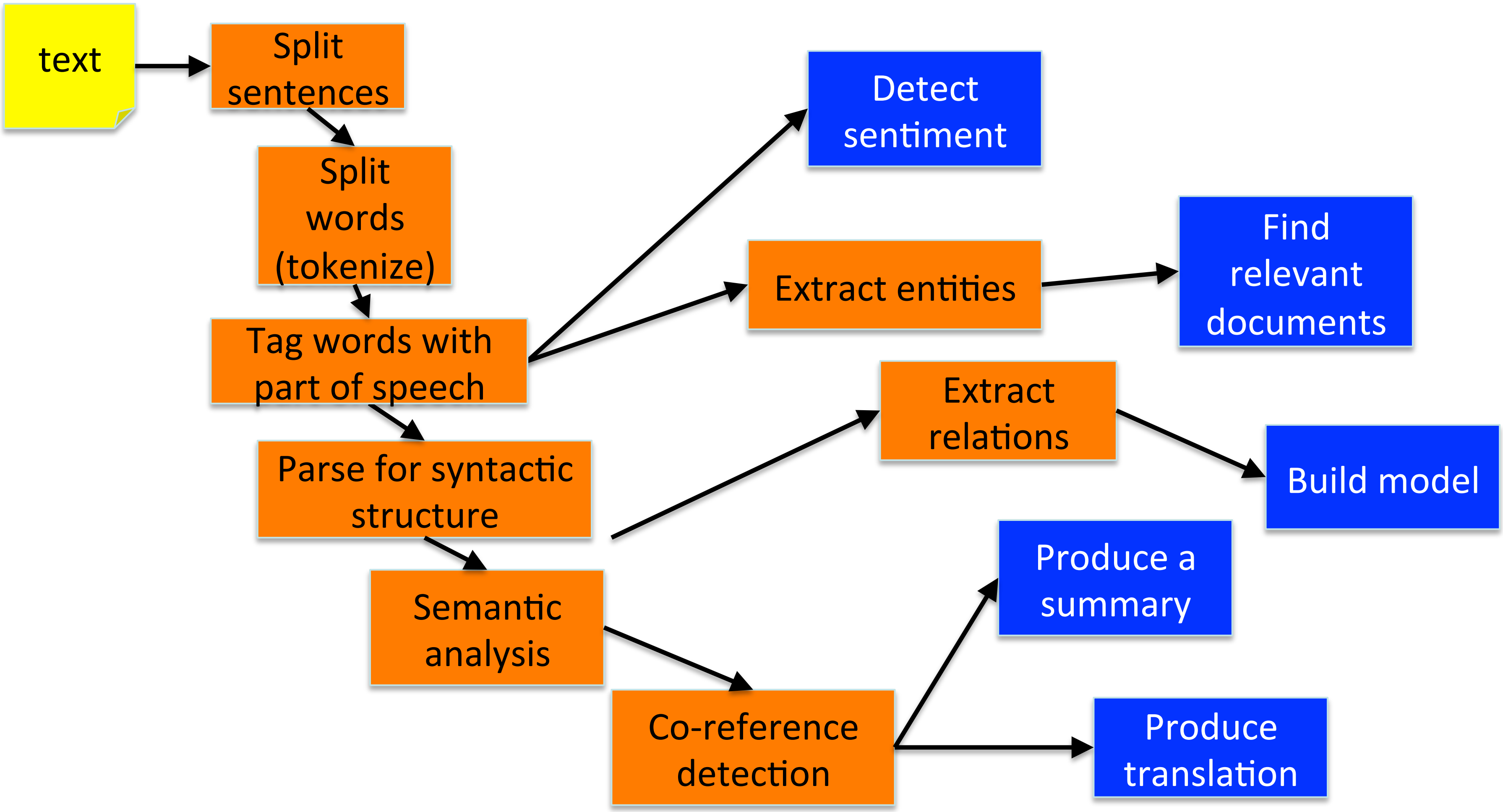
What we have in common

- Processing large amounts of (mostly text) data
- Data needs to be run through a pipeline of processors
  - pipelines are application dependent
  - Searching for patterns
- Processors in the pipeline may not have been designed to work together.

What is different

- Data sets
  - Genomes vs millions of small texts
- Data
  - You have genes, proteins, etc. (ACGT)
  - We have Twitter and Emoji
    - The OED added 2,000 new entries in the June 2016 update
  - Most of our processors are web services
    - The network is the bottleneck

Pipeline architecture for loosely-coupled linguistic analyses



The LAPPS Grid

- Provides interoperable access to
  - Wide array of NLP processing tools and components
  - language resources such as mono- and multi-lingual corpora and lexicons
- Enables pipelining tools to create custom NLP applications and “black box” composite services
- Provides an open advancement (OA) framework for component- and application-based evaluation
- Actively pursuing creation of an interoperable global network of grids and frameworks

Interoperability

- LAPPS Interchange Format (LIF)
  - allows services to exchange information
  - Syntactic interoperability
    - handled by JSON-LD
    - enforced by the LIF JSON schema
  - Semantic interoperability
    - enhanced by using the Linked Data aspect of JSON-LD to link to the LAPPS Web Services Exchange Vocabulary

Not a problem for genomic research!

Current collaborations/projects

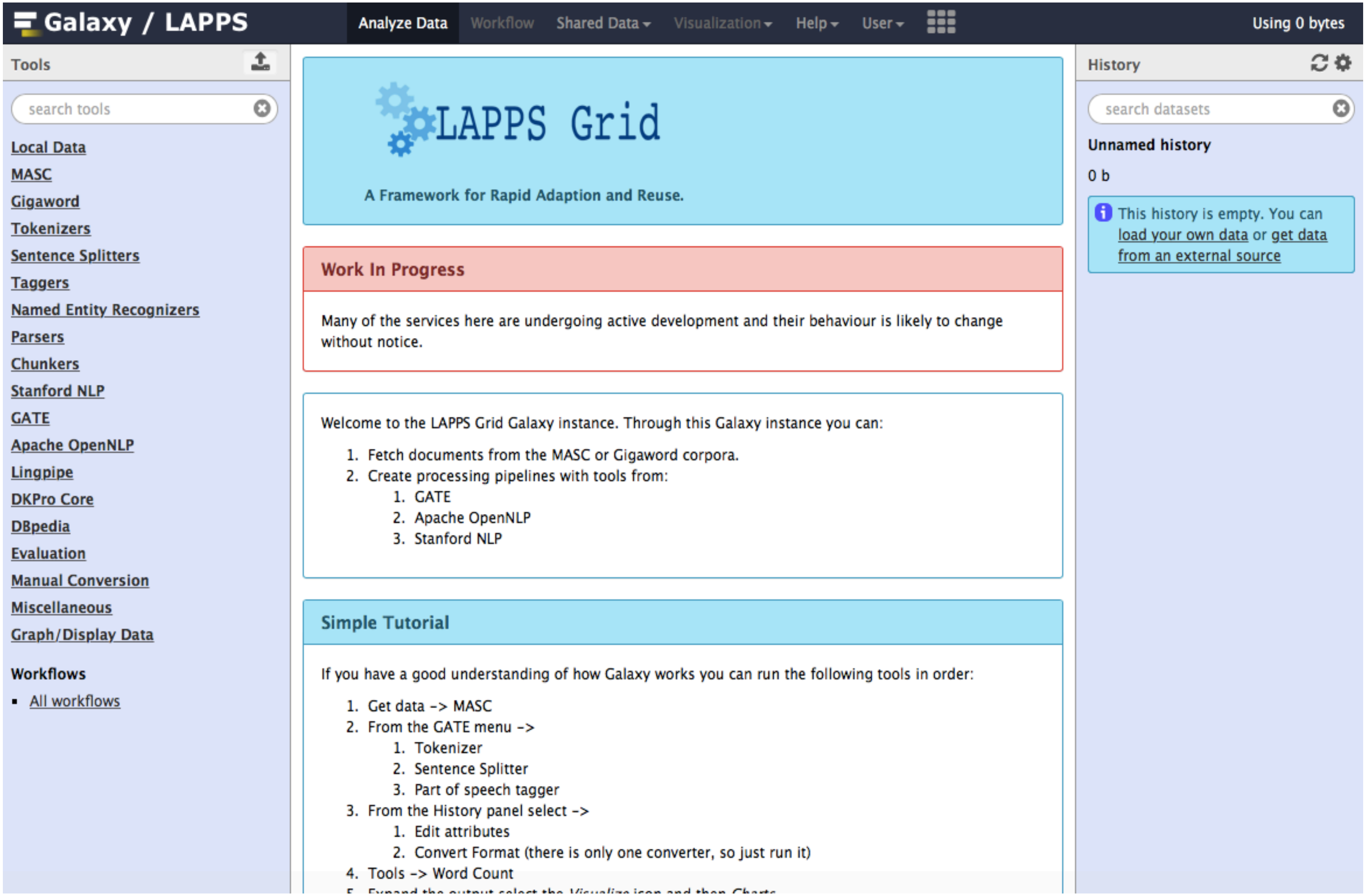
- Federation of Service Grids
  - LAPPS Grid, Language Grid (Kyoto University, Japan), NECTEC (Thailand), University of Indonesia, Xinjiang University (China), ELRA Grid
  - Access to all tools, applications, and resources on any grid through any portal
- LAPPS/CLARIN
  - CLARIN/WebLicht (Tubingen) and LINDAT/CLARIN (Prague)
  - Mellon Foundation proposal to create a trust network between LAPPS and CLARIN
- OpenMinted
  - Advisory board—work together on harmonization
- LAPPS Grid used in
  - DARPA LORELEI project for under-resourced languages
  - HathiTrust Research Center (HTRC) text mining project
  - Multi-day course for government analysts
  - Undergraduate and graduate Computational Linguistics courses at CMU, Brandeis, Vassar
  - ? IBM Watson

Replicability and Sharing

- The field of NLP research and development has been plagued by a chronic lack of replicability of results
  - A great deal of re-inventing of the wheel and wasted effort
  - Evaluation of results hampered when details of a study (including versions and parameters for data, software) are not included in papers
- The field of NLP is still hampered by a lack of widespread sharing of resources that are the basis of research results

LAPPS/Galaxy

- LAPPS/Galaxy components are LAPPS web services
- Access to 100+ LAPPS services plus those of federated partners
- Interoperability
  - LAPPS services among each other
  - LAPPS services and external grid components
    - handled by LAPPS converters



This history is empty. You can load your own data or get data from an external source

A Framework for Rapid Adaption and Reuse.

Work In Progress

Many of the services here are undergoing active development and their behaviour is likely to change without notice.

Welcome to the LAPPS Grid Galaxy instance. Through this Galaxy instance you can:

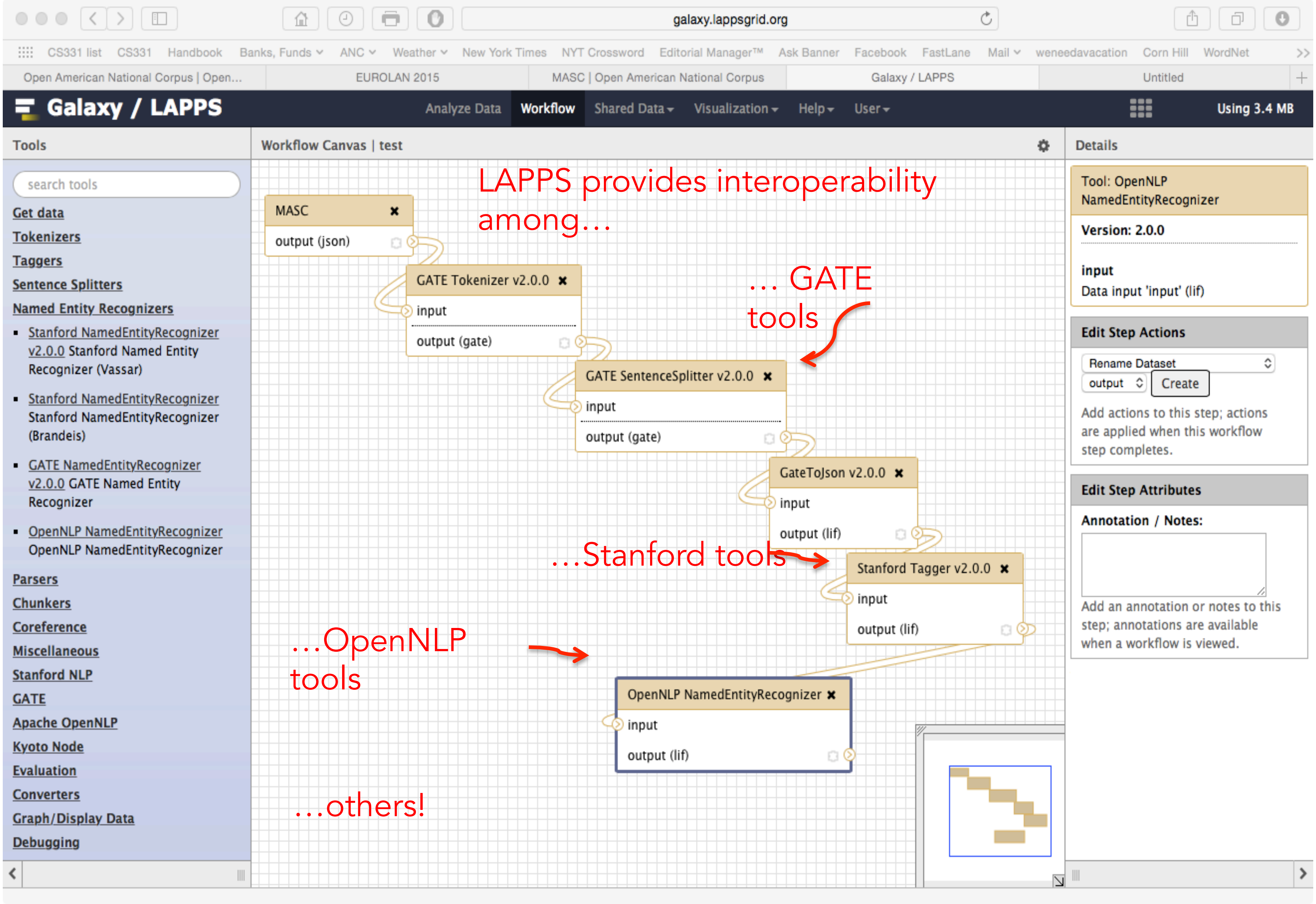
1. Fetch documents from the MASC or Gigaword corpora.
2. Create processing pipelines with tools from:
    1. GATE
    2. Apache OpenNLP
    3. Stanford NLP

Simple Tutorial

If you have a good understanding of how Galaxy works you can run the following tools in order:

1. Get data -> MASC
2. From the GATE menu ->
    1. Tokenizer
    2. Sentence Splitter
    3. Part of speech tagger
3. From the History panel select ->
    1. Edit attributes
    2. Convert Format (there is only one converter, so just run it)
4. Tools -> Word Count
5. Expand the output select the *Visualize* icon and then *Charts*

Workflow construction



Tool: OpenNLP NamedEntityRecognizer

Version: 2.0.0

input
Data input 'input' (lif)

Add actions to this step; actions are applied when this workflow step completes.

Annotation / Notes:

Add an annotation or notes to this step; annotations are available when a workflow is viewed.

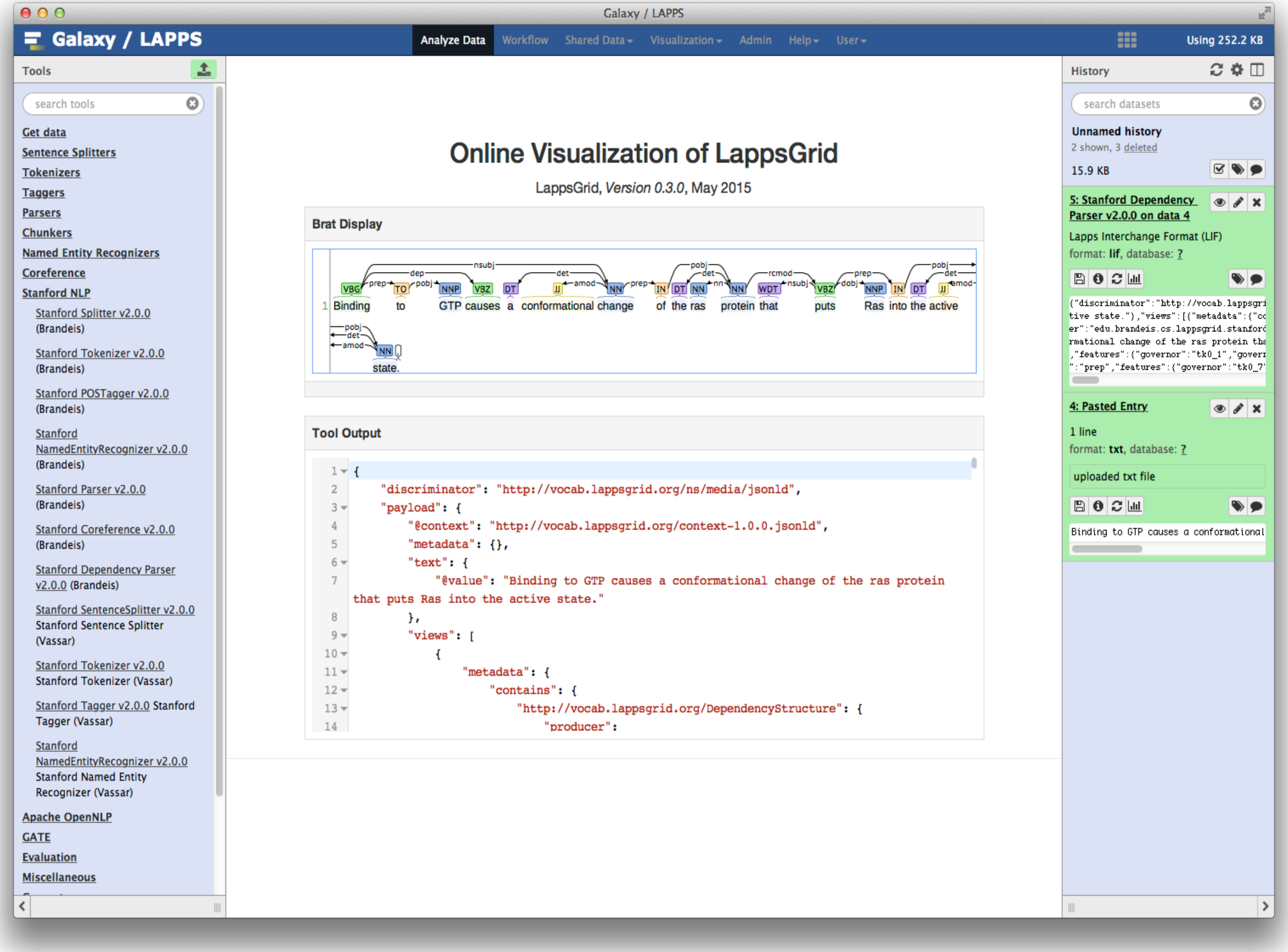
LAPPS provides interoperability among…

… GATE tools

…Stanford tools

…OpenNLP tools

…others!


Galaxy / LAPPS
Galaxy / LAPPS
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 252.2 KB
Tools
search tools
Get data
Sentence Splitters
Tokenizers
Taggers
Parsers
Chunkers
Named Entity Recognizers
Coreference
Stanford NLP
Stanford Splitter v2.0.0 (Brandeis)
Stanford Tokenizer v2.0.0 (Brandeis)
Stanford POSTagger v2.0.0 (Brandeis)
Stanford NamedEntityRecognizer v2.0.0 (Brandeis)
Stanford Parser v2.0.0 (Brandeis)
Stanford Coreference v2.0.0 (Brandeis)
Stanford Dependency Parser v2.0.0 (Brandeis)
Stanford SentenceSplitter v2.0.0 Stanford Sentence Splitter (Vassar)
Stanford Tokenizer v2.0.0 Stanford Tokenizer (Vassar)
Stanford Tagger v2.0.0 Stanford Tagger (Vassar)
Stanford NamedEntityRecognizer v2.0.0 Stanford Named Entity Recognizer (Vassar)
Apache OpenNLP
GATE
Evaluation
Miscellaneous

Online Visualization of LappsGrid

LappsGrid, *Version 0.3.0*, May 2015

Brat Display

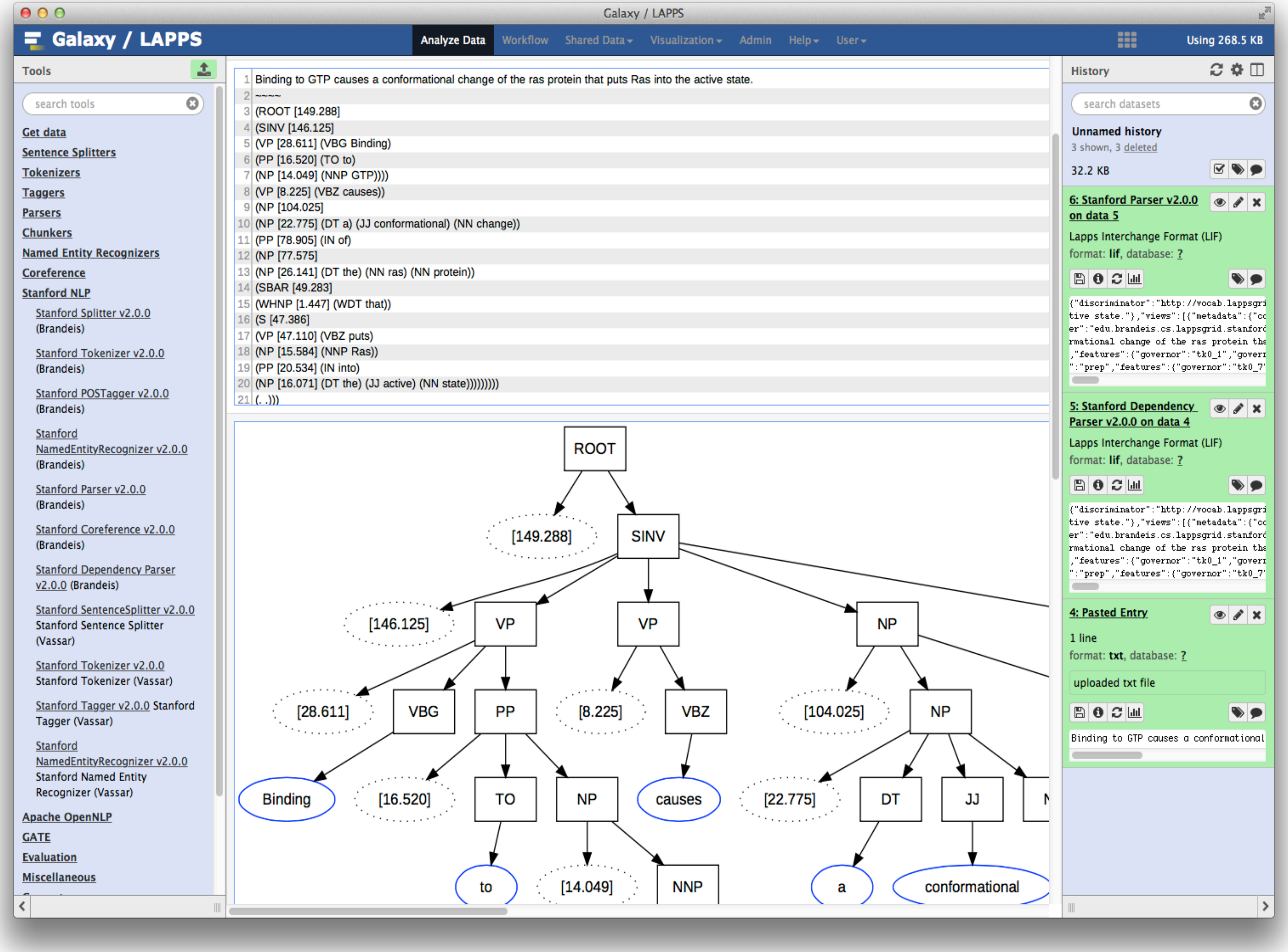
nsubj
dep
det
pobj
det
rcmod
prep
pobj
det
prep
pobj
amod
prep
nn
nsubj
dobj
amod
VBG
TO
NNP
VBZ
DT
JJ
NN
IN
DT
NN
NN
WDT
VBZ
NNP
IN
DT
JJ
1
Binding to GTP causes a conformational change of the ras protein that puts Ras into the active
pobj
det
amod
NN
state.

Tool Output

```
{
    "discriminator": "http://vocab.lappsgrid.org/ns/media/jsonld",
    "payload": {
        "@context": "http://vocab.lappsgrid.org/context-1.0.0.jsonld",
        "metadata": {},
        "text": {
            "@value": "Binding to GTP causes a conformational change of the ras protein that puts Ras into the active state."
        },
        "views": [
            {
                "metadata": {
                    "contains": {
                        "http://vocab.lappsgrid.org/DependencyStructure": {
                            "producer":
```

```
{"discriminator":"http://vocab.lappsgri
tive state."},"views":[{"metadata":{"co
er":"edu.brandeis.cs.lappsgrid.stanford
rmational change of the ras protein tha
,"features":{"governor":"tk0_1","govern
":"prep","features":{"governor":"tk0_7"
```

```
Binding to GTP causes a conformational
```



Get data
Sentence Splitters
Tokenizers
Taggers
Parsers
Chunkers
Named Entity Recognizers
Coreference
Stanford NLP

- Stanford Splitter v2.0.0 (Brandeis)
- Stanford Tokenizer v2.0.0 (Brandeis)
- Stanford POSTagger v2.0.0 (Brandeis)
- Stanford NamedEntityRecognizer v2.0.0 (Brandeis)
- Stanford Parser v2.0.0 (Brandeis)
- Stanford Coreference v2.0.0 (Brandeis)
- Stanford Dependency Parser v2.0.0 (Brandeis)
- Stanford SentenceSplitter v2.0.0 Stanford Sentence Splitter (Vassar)
- Stanford Tokenizer v2.0.0 Stanford Tokenizer (Vassar)
- Stanford Tagger v2.0.0 Stanford Tagger (Vassar)
- Stanford NamedEntityRecognizer v2.0.0 Stanford Named Entity Recognizer (Vassar)

Apache OpenNLP
GATE
Evaluation
Miscellaneous

```
Binding to GTP causes a conformational change of the ras protein that puts Ras into the active state.
~~~~
(ROOT [149.288]
(SINV [146.125]
(VP [28.611] (VBG Binding)
(PP [16.520] (TO to)
(NP [14.049] (NNP GTP))))
(VP [8.225] (VBZ causes))
(NP [104.025]
(NP [22.775] (DT a) (JJ conformational) (NN change))
(PP [78.905] (IN of)
(NP [77.575]
(NP [26.141] (DT the) (NN ras) (NN protein))
(SBAR [49.283]
(WHNP [1.447] (WDT that))
(S [47.386]
(VP [47.110] (VBZ puts)
(NP [15.584] (NNP Ras))
(PP [20.534] (IN into)
(NP [16.071] (DT the) (JJ active) (NN state))))))))
(. .)))
```

Unnamed history
3 shown, 3 deleted
32.2 KB

6: Stanford Parser v2.0.0 on data 5

Lapps Interchange Format (LIF)
format: lif, database: ?

{"discriminator":"http://vocab.lappsgri
tive state."},"views":[{"metadata":{"cc
er":"edu.brandeis.cs.lappsgrid.stanforc
rmational change of the ras protein tha
,"features":{"governor":"tk0_1","goverr
":"prep","features":{"governor":"tk0_7'

5: Stanford Dependency Parser v2.0.0 on data 4

Lapps Interchange Format (LIF)
format: lif, database: ?

{"discriminator":"http://vocab.lappsgri
tive state."},"views":[{"metadata":{"cc
er":"edu.brandeis.cs.lappsgrid.stanforc
rmational change of the ras protein tha
,"features":{"governor":"tk0_1","goverr
":"prep","features":{"governor":"tk0_7'

4: Pasted Entry

1 line
format: txt, database: ?

uploaded txt file

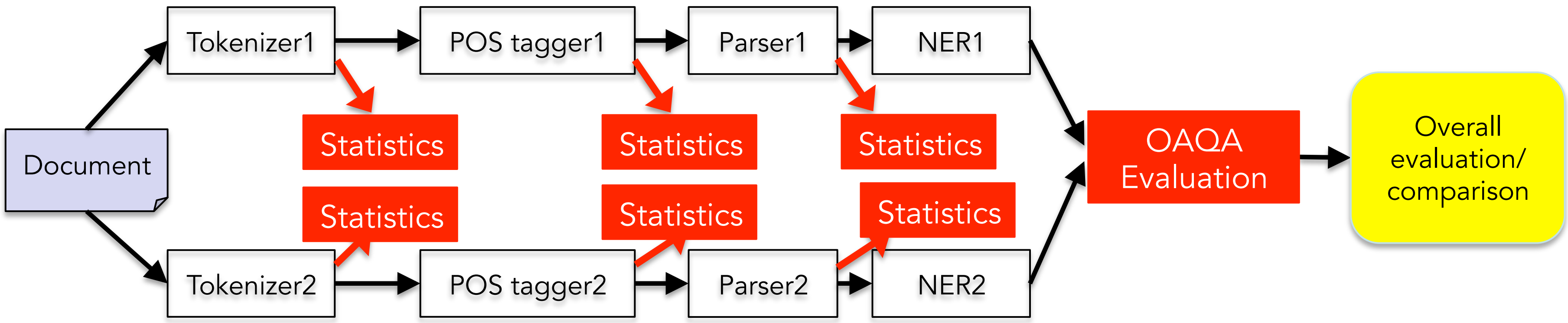
Binding to GTP causes a conformational

Evaluation in LAPPS/Galaxy

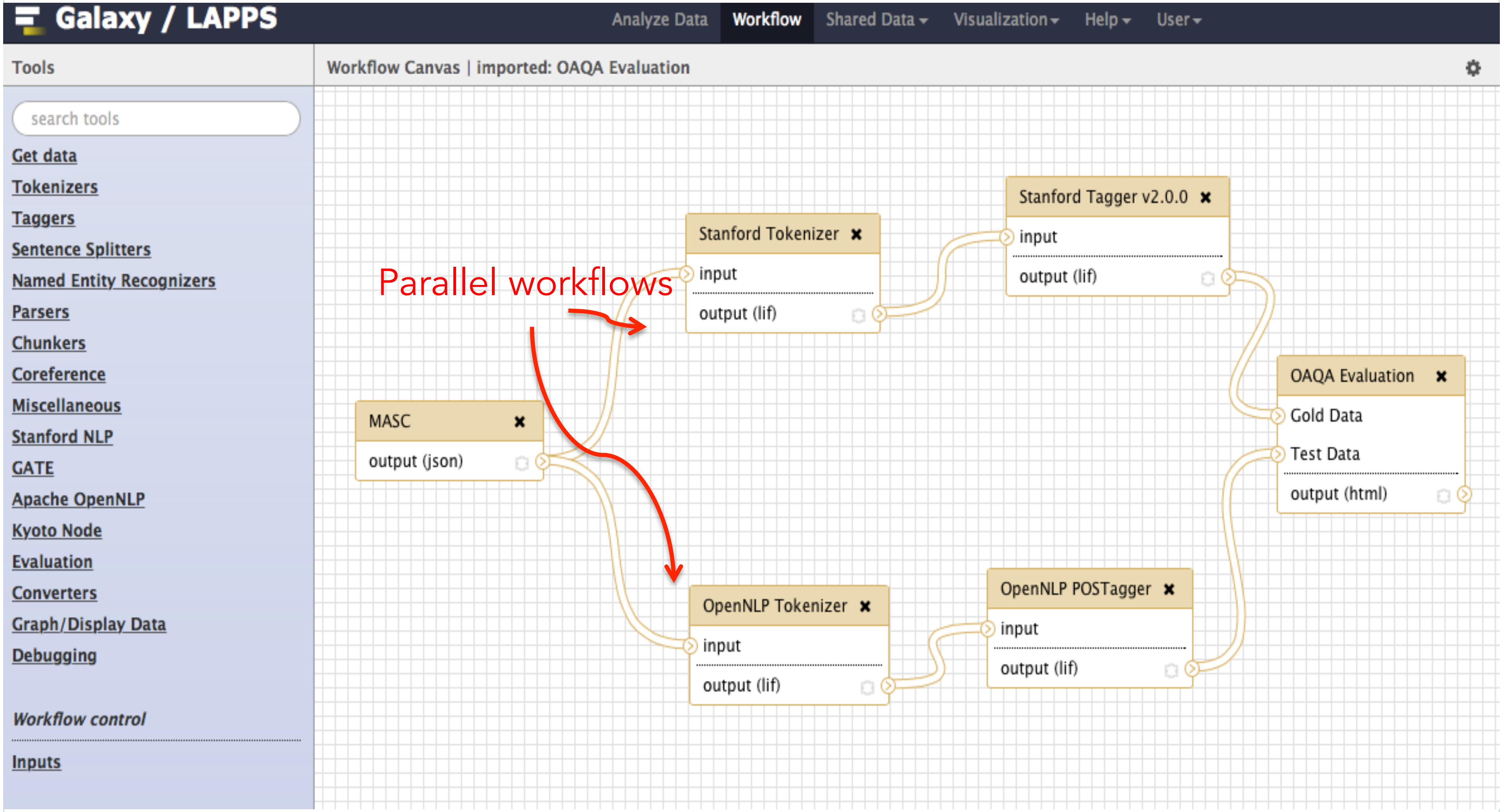
- CMU has implemented services for state-of-the-art Open Advancement techniques
- Enables rapid identification of
  - frequent error categories within modules and documents
  - which module(s) and error type(s) have the greatest impact on overall performance
- Used in the development of IBM's Watson to achieve steady performance gains over the four years of its development

Open Advancement in a Nutshell

- Analyzes results in/from alternative pipelines



- Can be comparison to gold standard, or comparison to another pipeline or pipelines
- Potentially any number of pipelines can be compared
  - CMU working on methods for finding an optimal solution among all multiple possible paths



Parallel workflows

Potential benefits of LAPPS/Galaxy collaboration

- Galaxy contains a huge number of tools for analyzing genomic and other biomedical data
- LAPPS includes tools to perform NLP analyses on unstructured textual data
- Combining LAPPS services with Galaxy tools can allow for analysis of data mined from the vast stores of biomedical literature (Biomed, PubMed, PLOS, etc.)

- BIONLP meets bio-analysis!

Machine Reasoning 101

- Literally dumber than a bag of light switches
  - Actually literally is a bag of light-switches
- Symbolic reasoning using *triples*
  - OWL/RDF expressed as triples
- *subject predicate object*
- The system doesn't actually understand anything
  - follows simple rules to manipulate symbols

Machine Reasoning 101

- Given a simple Knowledge Base
  - A P B
  - A Q C
  - D P B
  - D Q E
  - F P G
  - H P G
  - F Q A
  - H Q D

Machine Reasoning 101

- Can construct a simple query
  - ?x P G   (1)
  - ?x Q ?y   (2)
  - ?y Q C   (3)
- Solve for ?x and ?y
  - Two triples match (1)
  - F P G (x=F)
  - H P G (x=H)

Machine Reasoning 101

- Given x=F or x=H solve ?x Q ?y
  - F Q A (x=F, y=A)
  - H Q D (x=H, y=D)
- Given y=A or y=D solve ?y Q C
- One triple matches : A Q C
  - x=F and y=A
- If unable to match any triple the answer is False (no)

Machine Reasoning 101

- Human readable knowledge base
  - Toronto is-a City
  - Toronto is-in Canada
  - New-York is-a City
  - New-York is-in USA
  - Pearson is-a Airport
  - LaGuardia is-a Airport
  - Pearson is-in Toronto
  - LaGuardia is-in New-York

Machine Reasoning 101

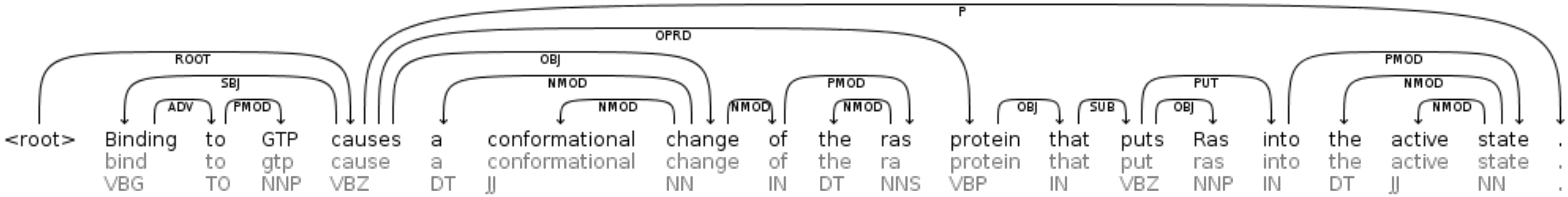
- Find all the airports in Canada
  - ?x is-a Airport
  - ?x is-in ?city
  - ?city is-in Canada
- Supporting facts (triples)
  - Pearson is-a Airport
  - Pearson is-in Toronto
  - Toronto in-in Canada

Example

Binding to GTP causes a conformational change of the ras protein that puts Ras into the active state. GTP-bound ras binds to the raf protein kinase. This binding of raf to ras has the effect of activating the raf kinase and localizing the raf kinase to the cell membrane. Activated raf now phosphorylates and activates the Mek1 kinase. The Mek1 kinase then phosphorylates the ERK kinase on both threonine and tyrosine residues which activate ERK kinase activity. The phosphorylated ERK protein then translocates to the nucleus where it regulates gene expression in part by phosphorylating the Elk1 transcription factor. Phospho-Elk then upregulates the gene expression of target genes such as the proto- oncogene c-fos. The entire signaling cascade is terminated by the intrinsic GTPase activity of ras which hydrolyzes the bound GTP into GTP, thus returning ras to the GDP bound state where it releases bound raf. The GTPase activity of ras is accelerated by interaction with another protein called GAP. The oncogenic rasv12 mutant has diminished GTPase activity and therefore stays in the active GTP bound state constitutively. Deletion of GAP or the related NF1 genes will also enhance ras activity by slowing the rate of ras-GTP hydrolysis.

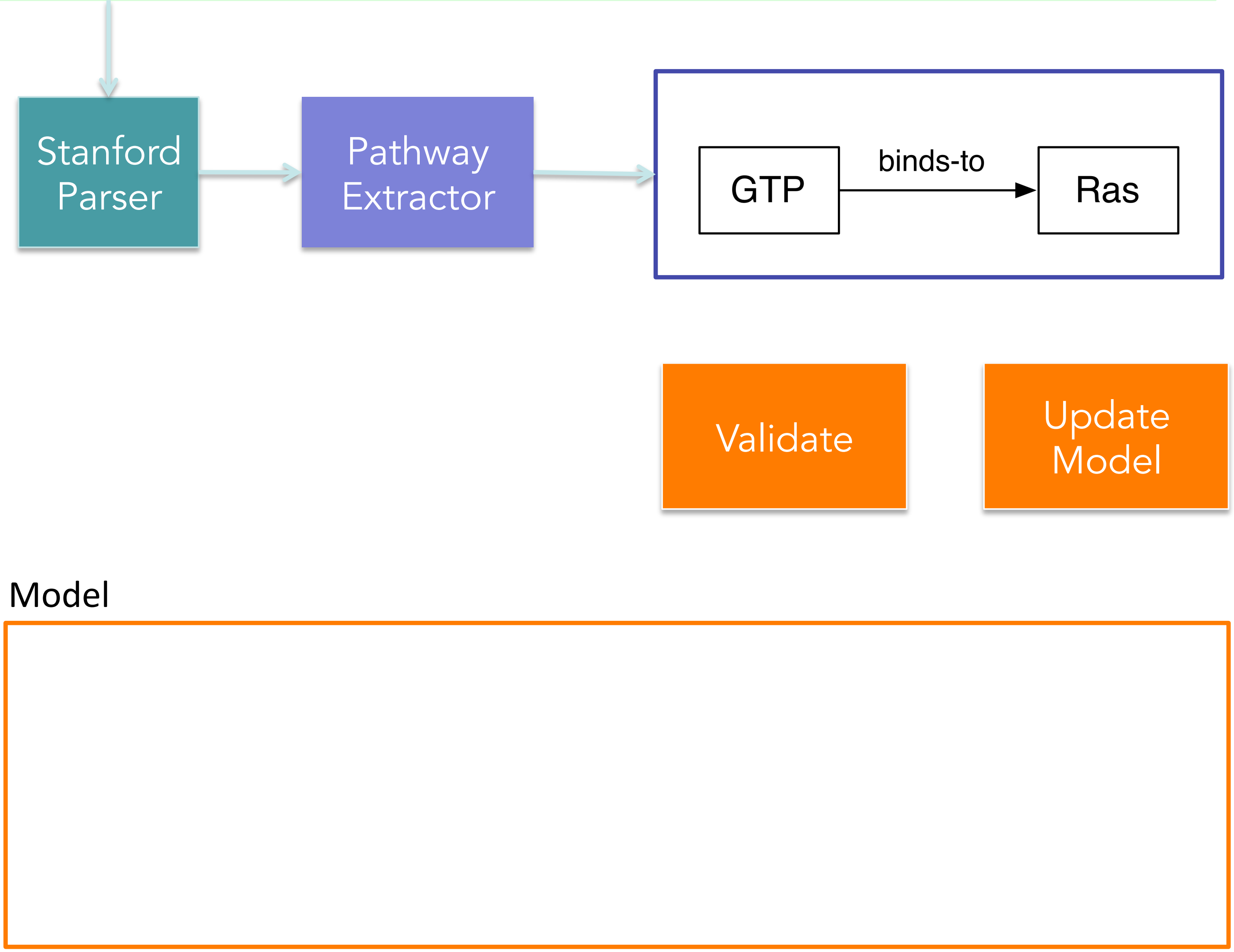
Information Extraction

- Binding to GTP causes a conformational change of the ras protein that puts Ras into the active state.



<root>	Binding	to	GTP	causes	a	conformational	change	of	the	ras	protein	that	puts	Ras	into	the	active	state	.
	bind	to	gtp	cause	a	conformational	change	of	the	ra	protein	that	put	ras	into	the	active	state	.
	VBG	TO	NNP	VBZ	DT	JJ	NN	IN	DT	NNS	VBP	IN	VBZ	NNP	IN	DT	JJ	NN	.

Binding to **GTP** **causes** a **conformational change** of the **ras protein** that puts **Ras** into the active state.



Stanford Parser

Pathway Extractor

GTP

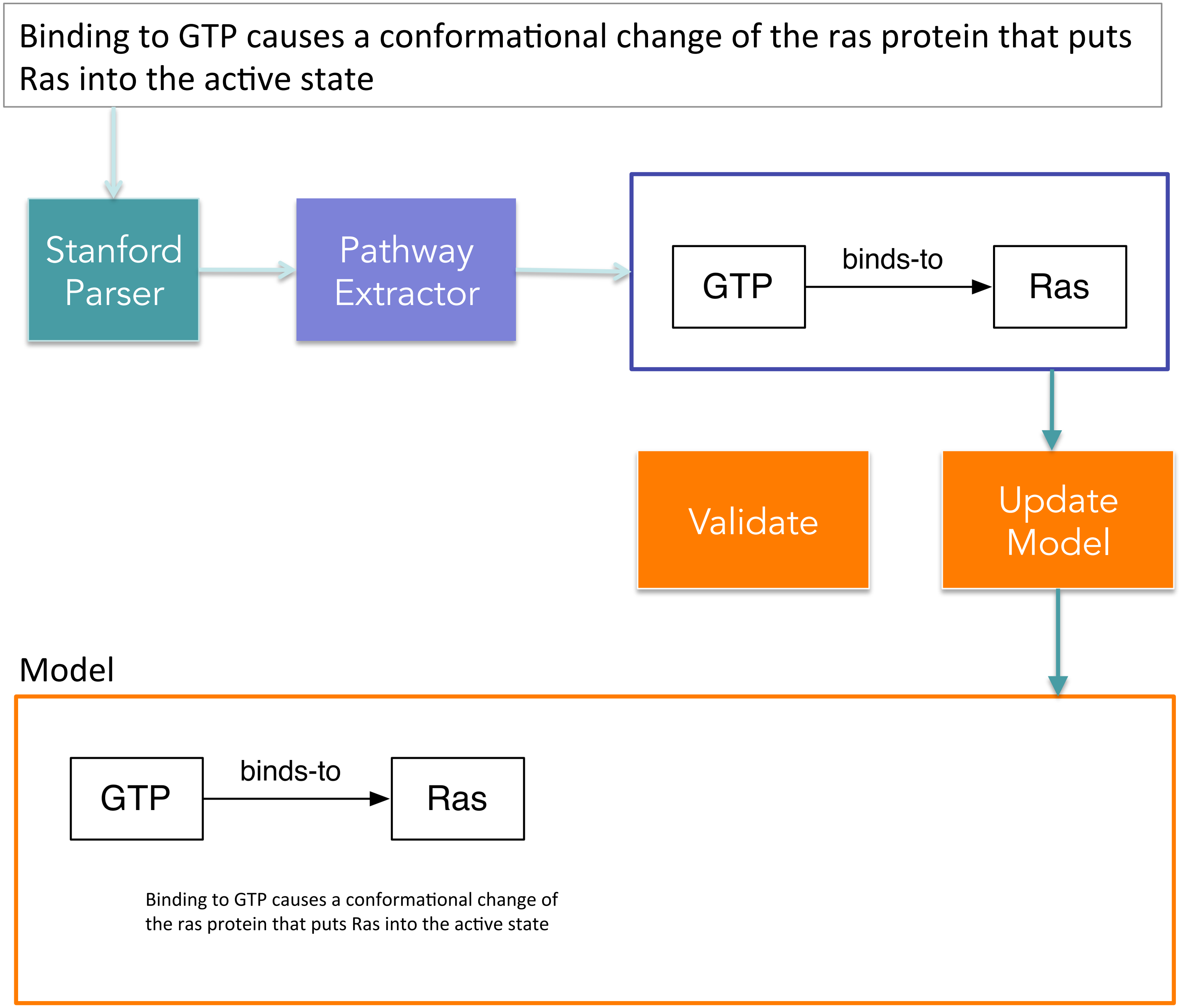
binds-to

Ras

Validate

Update Model

Model



Binding to GTP causes a conformational change of the ras protein that puts Ras into the active state

Stanford Parser

Pathway Extractor

GTP binds-to Ras

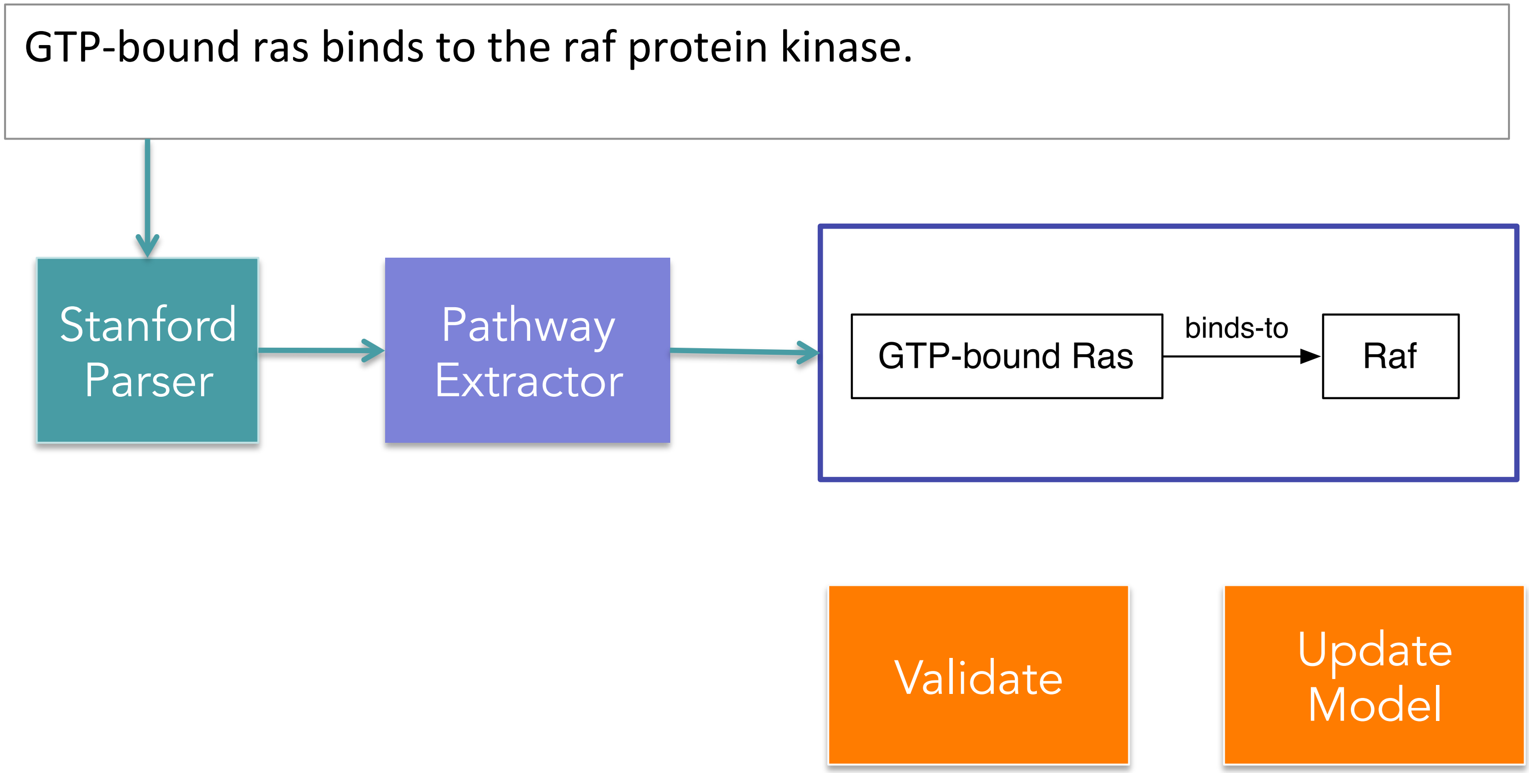
Validate

Update Model

Model

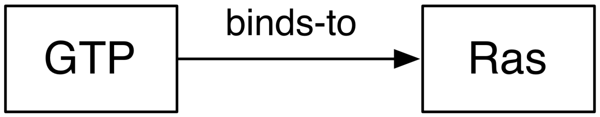
GTP binds-to Ras

Binding to GTP causes a conformational change of the ras protein that puts Ras into the active state

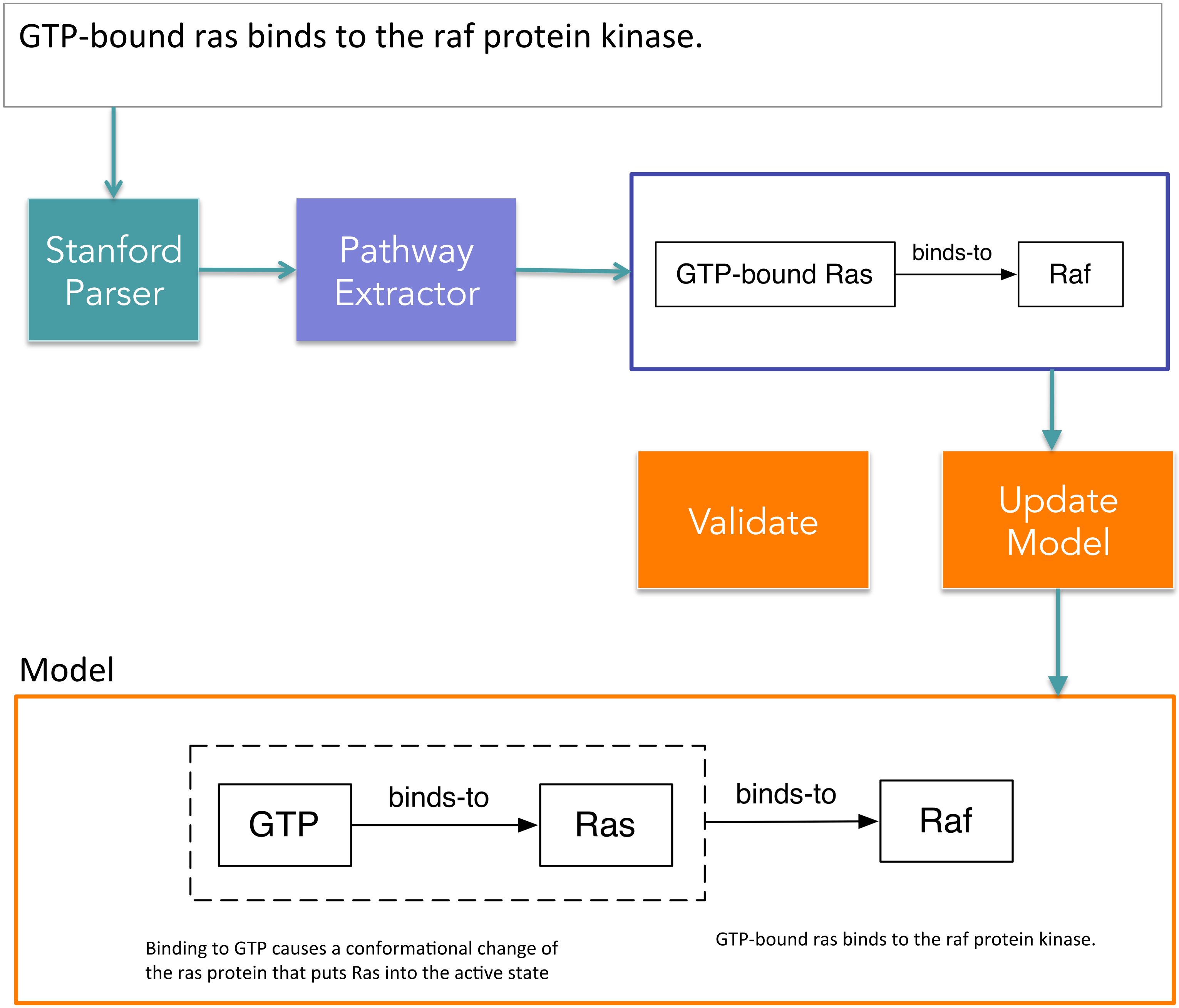


GTP-bound ras binds to the raf protein kinase.

Model



Binding to GTP causes a conformational change of the ras protein that puts Ras into the active state



GTP-bound ras binds to the raf protein kinase.

Stanford Parser

Pathway Extractor

GTP-bound Ras binds-to Raf

Validate

Update Model

Model

GTP binds-to Ras

binds-to Raf

Binding to GTP causes a conformational change of the ras protein that puts Ras into the active state

GTP-bound ras binds to the raf protein kinase.

Potential

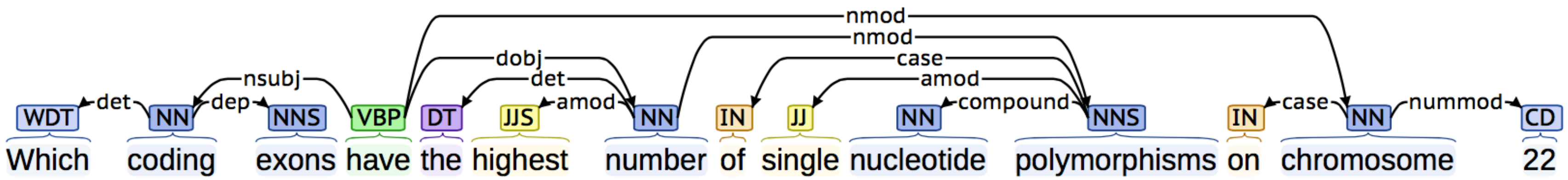
- Feed model /extracted information to tools in Galaxy for analysis/visualization

- The LAPPS Grid team is very interested in collaborations to explore and develop such capabilities!

Potential

- Siri, which coding exon has the highest number of single nucleotide polymorphisms on chromosome 22?
- If you ask Siri or Google this question you get the Galaxy 101 Tutorial page.
- Wouldn't it be nice to get the Galaxy pipeline that answered the question?
  - Or better, just the output from the pipeline

Parse the Question

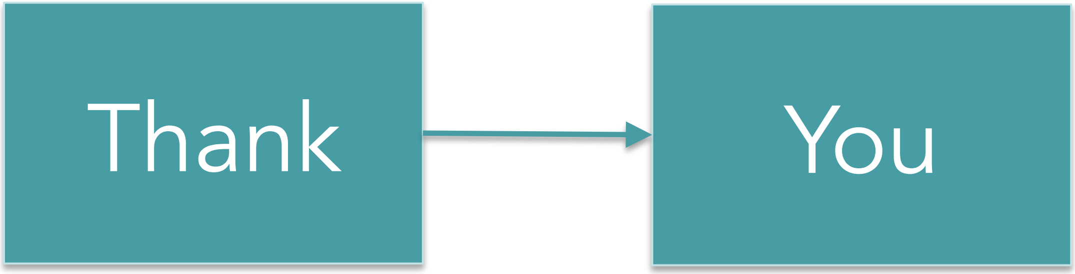


Parse the Question

- Predicate: *have*
- Subject: *coding exon*
- Object: *number of*
  - SNP on *chromosome 22*
- Plan steps to answer question
  - data required
  - processing steps required

Answer the Question

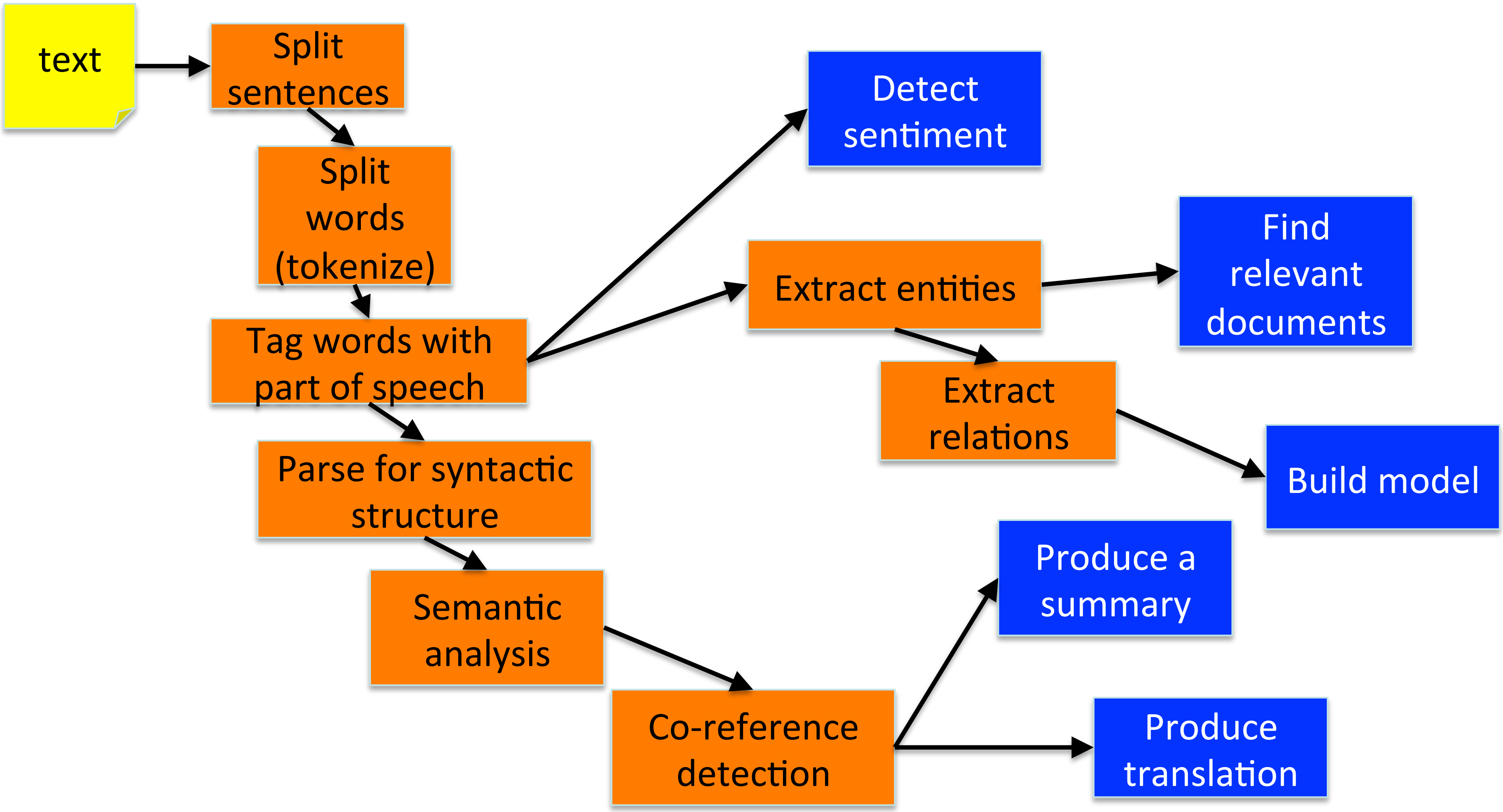
- Data required
  - Coding Exons
  - Single Nucleotide Polymorphisms
  - How do we know where to find this data?
- Steps required
  - Join exons with SNP
  - Group by column 4 (how do we know this?)
  - Count SNP per exon
  - Sort and return top five.



Thank

You

Pipeline architecture for loosely-coupled linguistic analyses



Natural Language Processing

- LAPPS provides a component framework for developing NLP applications
- What is an NLP application
  - Question Answering
  - Sentiment Analysis
  - Machine Translation
  - Message Understanding

Question Answering

- Siri/Google do a reasonable job.
- Typically need to be trained for a specific domain.
- When they are wrong they are spectacularly wrong
  - IBM Watson got the Final Jeopardy! question hilariously wrong because it did not *know* Toronto Canada was not a US city.

Sentiment Analysis

- Does the message express a positive or negative viewpoint?
- Challenges
  - Sarcasm
  - People make things up
    - I don't want to harsh on your mellow but words have no meaning.

Machine Translation

- Getting better
  - Mostly "not-gibberish" now.
- Same problems with sentiment analysis
  - English, "I have a bone to pick with you."
  - German, "I have a chicken to pluck with you."
- Starts to perform very badly in multi-step translations
  - Japanese-> English -> Arabic

Message Understanding

- Topic Analysis
- Summarization
  - What are the main points of the article
  - Typically represented as a "bag of words"

NLP Pipeline

- Segmentation
  - Split the text into sentences
  - Split the sentences into words
- You would think this would be easy
  - You would be wrong
  - gonna, wanna, ur, even things like don't
    - do not
    - do n't
    - don 't

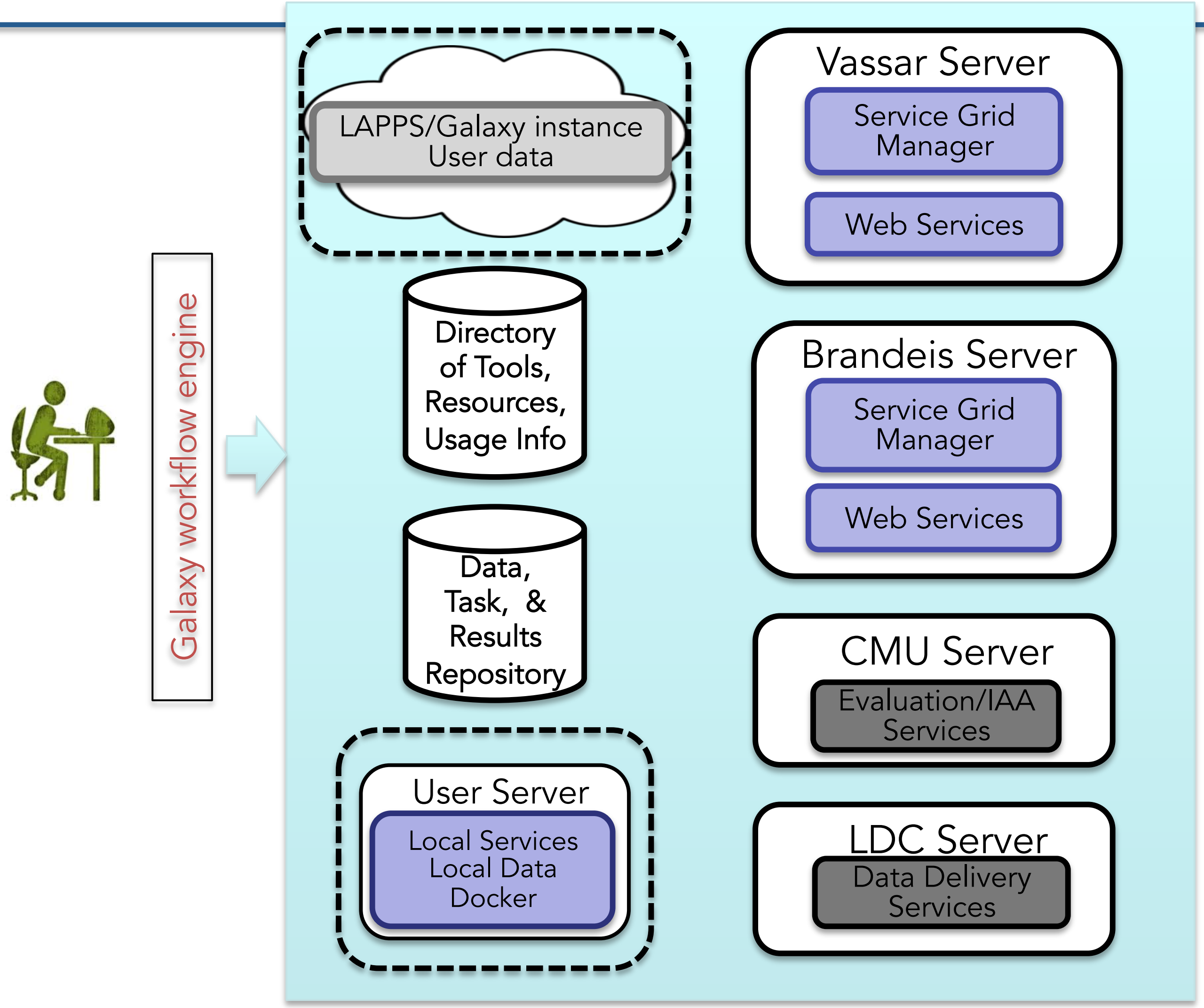
NLP Pipeline

- Tagging
  - Part-of-speech
  - Named entities
  - Semantic Role Labeling
    - John sold the book to Mary.
    - Verb "to sell"
    - "John" is the seller
    - "Mary" is the recipient
    - "the book" was the thing that was sold

Knowledge Bases

- Where do they come from?
  - Hand crafted
    - very time consuming (expensive)
  - Machine generated
    - *Easy* to generate large KB
      - full of noise
    - More data == Better data
      - use statistics to reduce noise

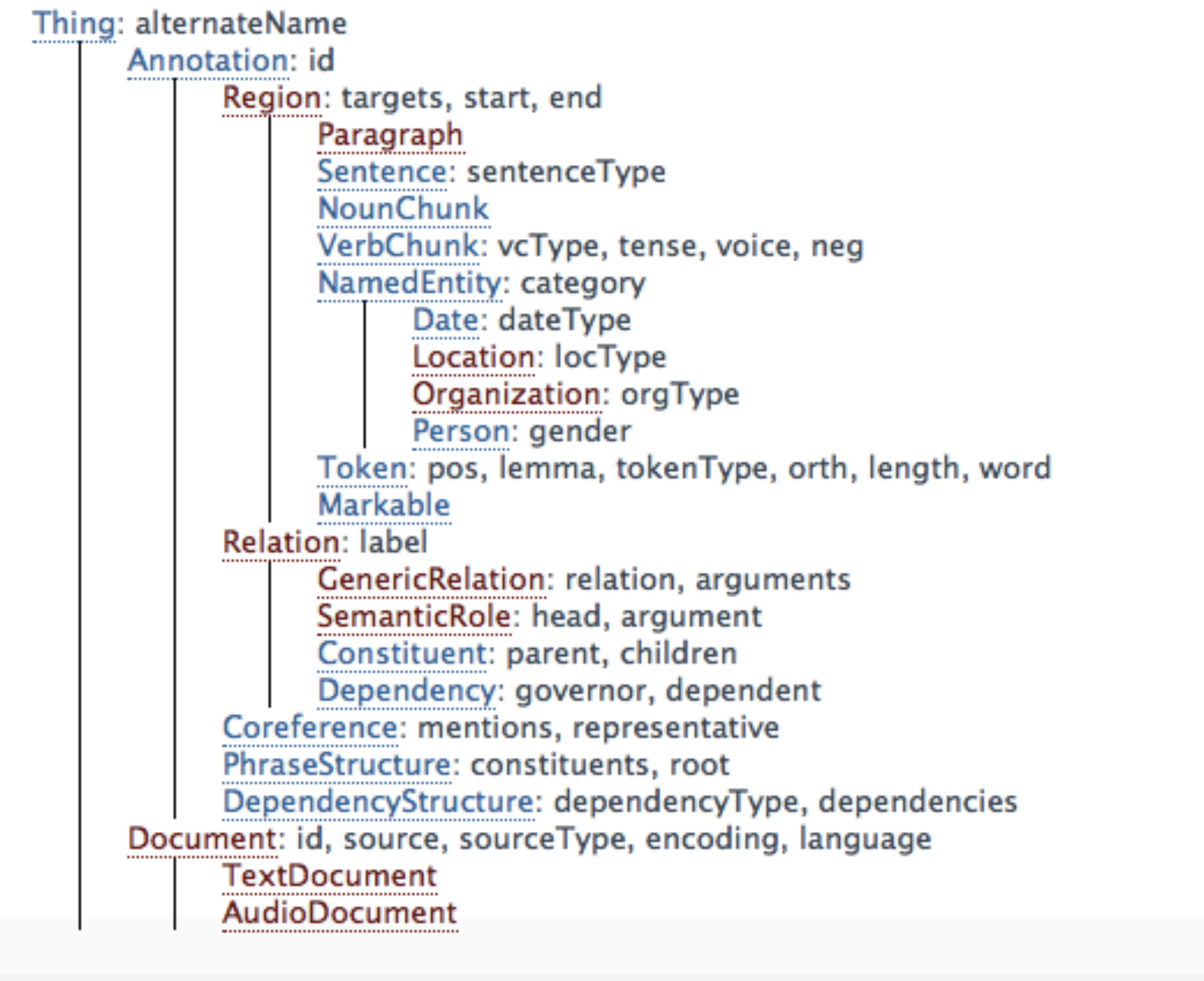
LAPPS Grid Overview



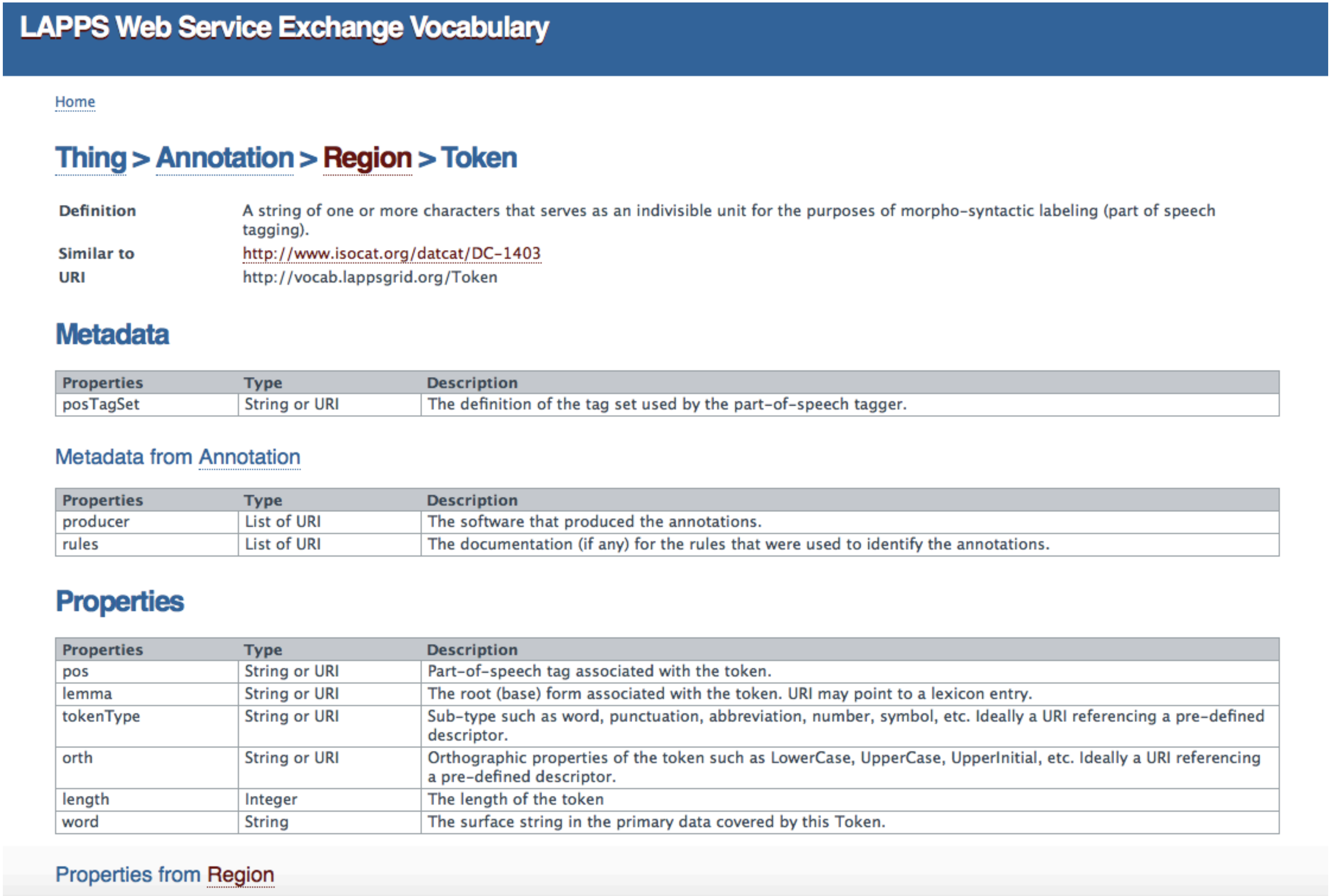
Why Galaxy?

- Accessible: Accommodates users with a broad range of expertise (non-computational to expert programmer)
- Reproducible: Galaxy captures information so that any user can understand and repeat a complete computational analysis
- Transparent: Users share and publish analyses via the web and create interactive, web-based documents that describe a complete analysis
- Well-developed, supported, open !

LAPPS Exchange Vocabulary Type Hierarchy



- Thing: alternateName
  - Annotation: id
    - Region: targets, start, end
      - Paragraph
      - Sentence: sentenceType
      - NounChunk
      - VerbChunk: vcType, tense, voice, neg
      - NamedEntity: category
        - Date: dateType
        - Location: locType
        - Organization: orgType
        - Person: gender
      - Token: pos, lemma, tokenType, orth, length, word
      - Markable
    - Relation: label
      - GenericRelation: relation, arguments
      - SemanticRole: head, argument
      - Constituent: parent, children
      - Dependency: governor, dependent
    - Coreference: mentions, representative
    - PhraseStructure: constituents, root
    - DependencyStructure: dependencyType, dependencies
  - Document: id, source, sourceType, encoding, language
    - TextDocument
    - AudioDocument



LAPPS Web Service Exchange Vocabulary

Home

Thing > Annotation > Region > Token

Definition	A string of one or more characters that serves as an indivisible unit for the purposes of morpho-syntactic labeling (part of speech tagging).
Similar to	http://www.isocat.org/datcat/DC-1403
URI	http://vocab.lappsgrid.org/Token

Metadata

Properties	Type	Description
posTagSet	String or URI	The definition of the tag set used by the part-of-speech tagger.

Metadata from Annotation

Properties	Type	Description
producer	List of URI	The software that produced the annotations.
rules	List of URI	The documentation (if any) for the rules that were used to identify the annotations.

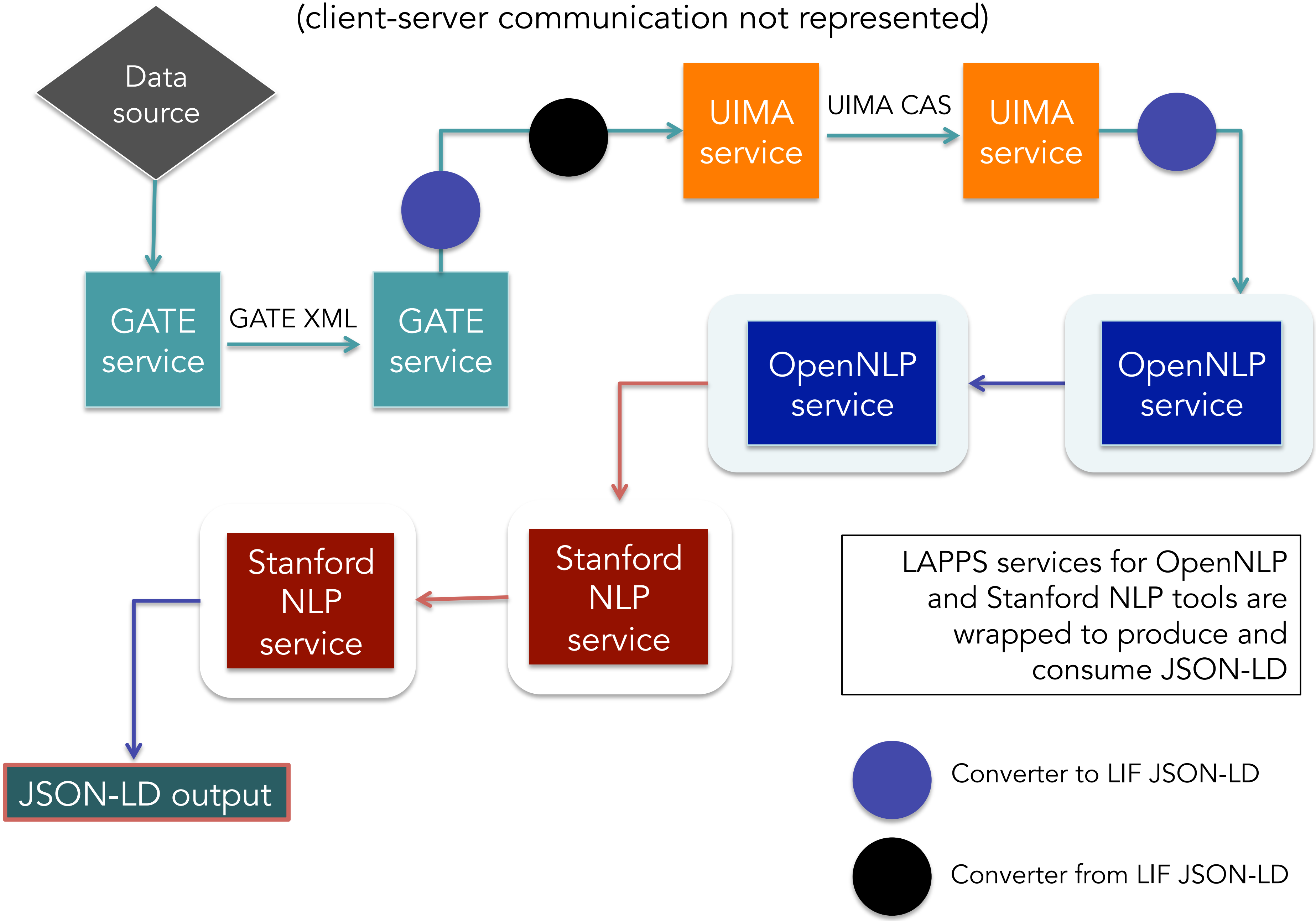
Properties

Properties	Type	Description
pos	String or URI	Part-of-speech tag associated with the token.
lemma	String or URI	The root (base) form associated with the token. URI may point to a lexicon entry.
tokenType	String or URI	Sub-type such as word, punctuation, abbreviation, number, symbol, etc. Ideally a URI referencing a pre-defined descriptor.
orth	String or URI	Orthographic properties of the token such as LowerCase, UpperCase, UpperInitial, etc. Ideally a URI referencing a pre-defined descriptor.
length	Integer	The length of the token
word	String	The surface string in the primary data covered by this Token.

Properties from Region

Logical flow

(client-server communication not represented)



Data source

GATE service

GATE XML

GATE service

UIMA service

UIMA CAS

UIMA service

OpenNLP service

OpenNLP service

Stanford NLP service

Stanford NLP service

JSON-LD output

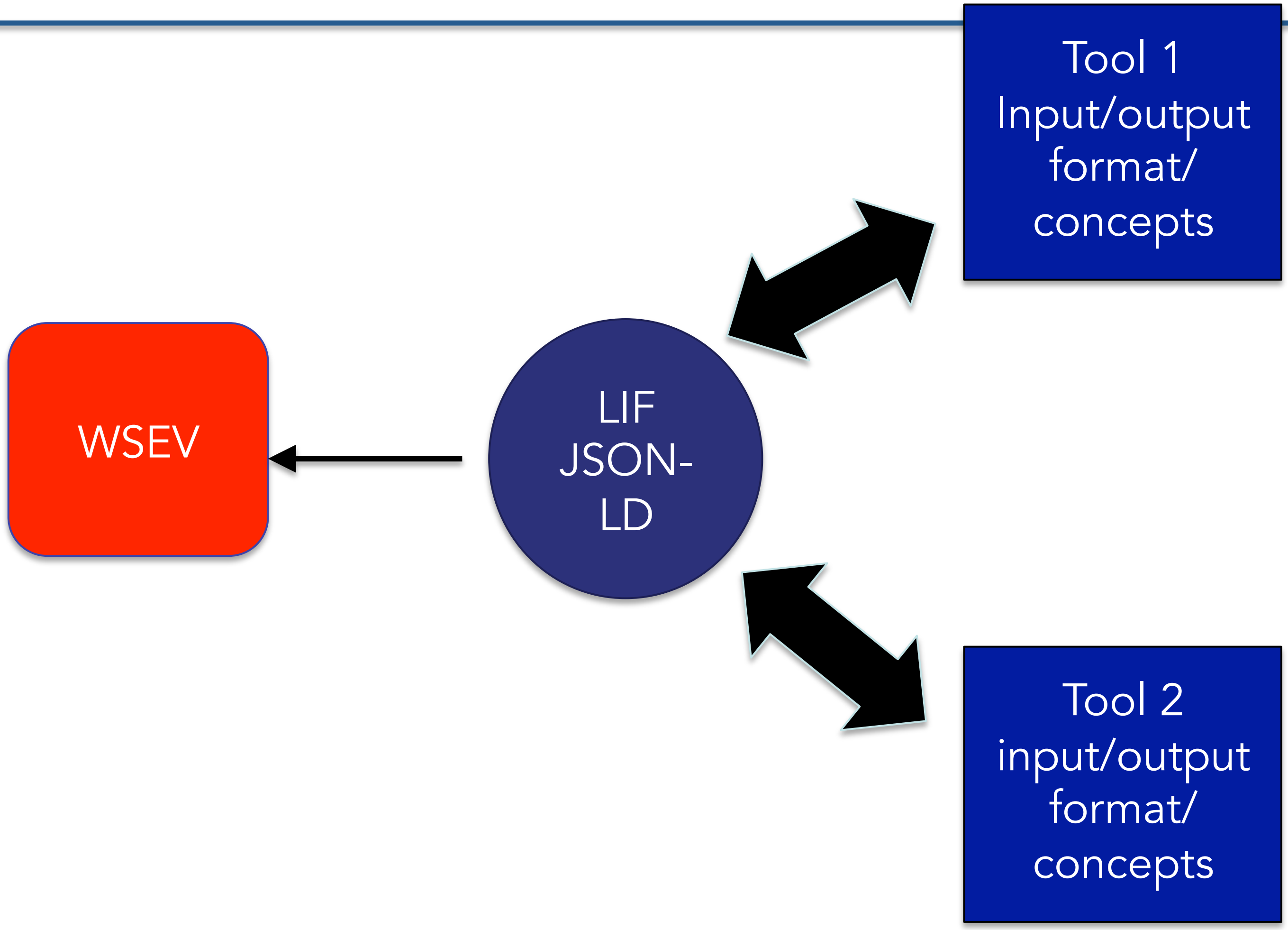
LAPPS services for OpenNLP and Stanford NLP tools are wrapped to produce and consume JSON-LD

● Converter to LIF JSON-LD

● Converter from LIF JSON-LD

All tools' input/output formats mapped into and out of LIF

Linguistic categories etc. mapped to WSEV



LAPPS Web Service Exchange Vocabulary

- Specifies a terminology for a core of linguistic objects and features exchanged among NLP tools that consume and produce linguistically annotated data
- Linked wherever possible to existing repositories such as ISOCat (CLARIN Concept Repository), schema.org, FoLiA categories, etc.
- References in JSON-LD representation point to URIs providing **definitions** for specific linguistic categories in the WS-EV

Interoperability

- LAPPS Interchange Format (LIF)
  - allows services to exchange information
  - Syntactic interoperability
    - handled by JSON-LD
    - enforced by the LIF JSON schema
  - Semantic interoperability
    - enhanced by using the Linked Data aspect of JSON-LD to link to the LAPPS Web Services Exchange Vocabulary
      - ➢ Not yet-another-repository! Linked to others where possible