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

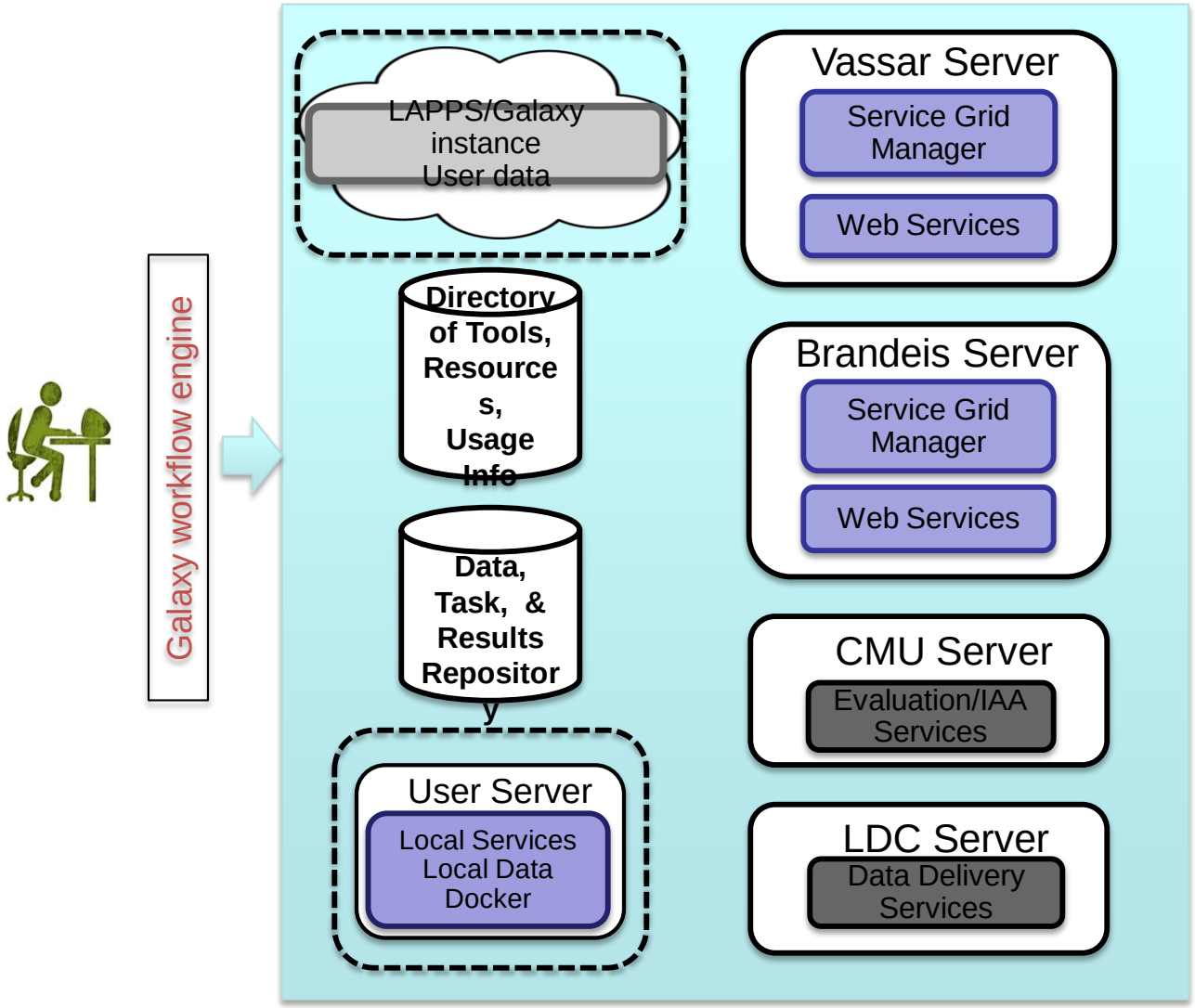
The LAPPS Grid

- A framework to
  - enable language service discovery, composition, and reuse
    - For both NLP researchers and others (who may use pre-developed composite services)
  - promote sustainability, manageability, usability, and **interoperability** of NLP components
- Based on the **service-oriented architecture** (SOA)
  - Web-oriented version of the “pipeline” architecture for sequencing 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
- Will enable **easy navigation through licensing issues**
- Actively pursuing creation of an **interoperable global network of grids and frameworks**

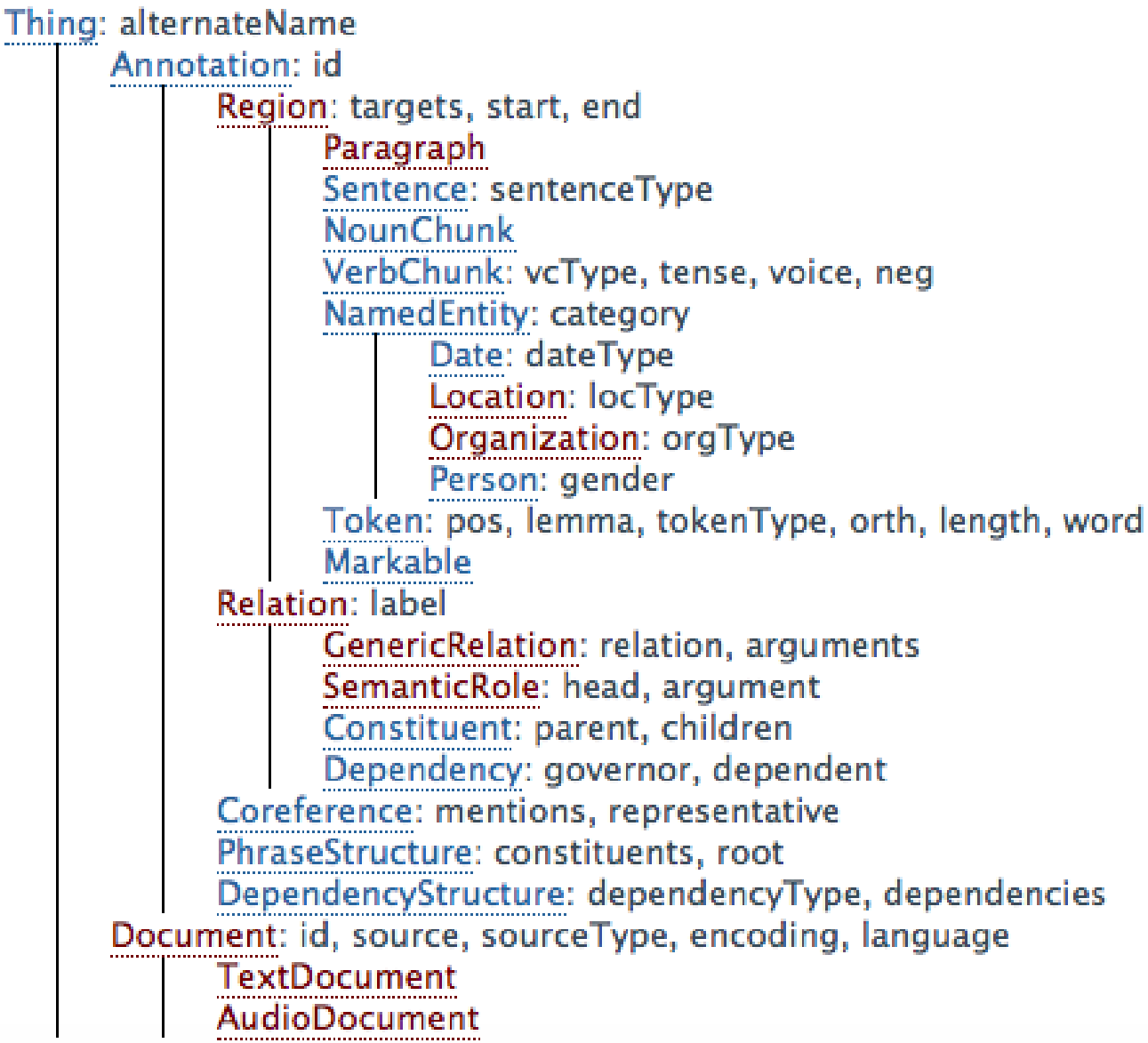
LAPPS Grid Overview



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**

LAPPS Exchange Vocabulary Type Hierarchy



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.

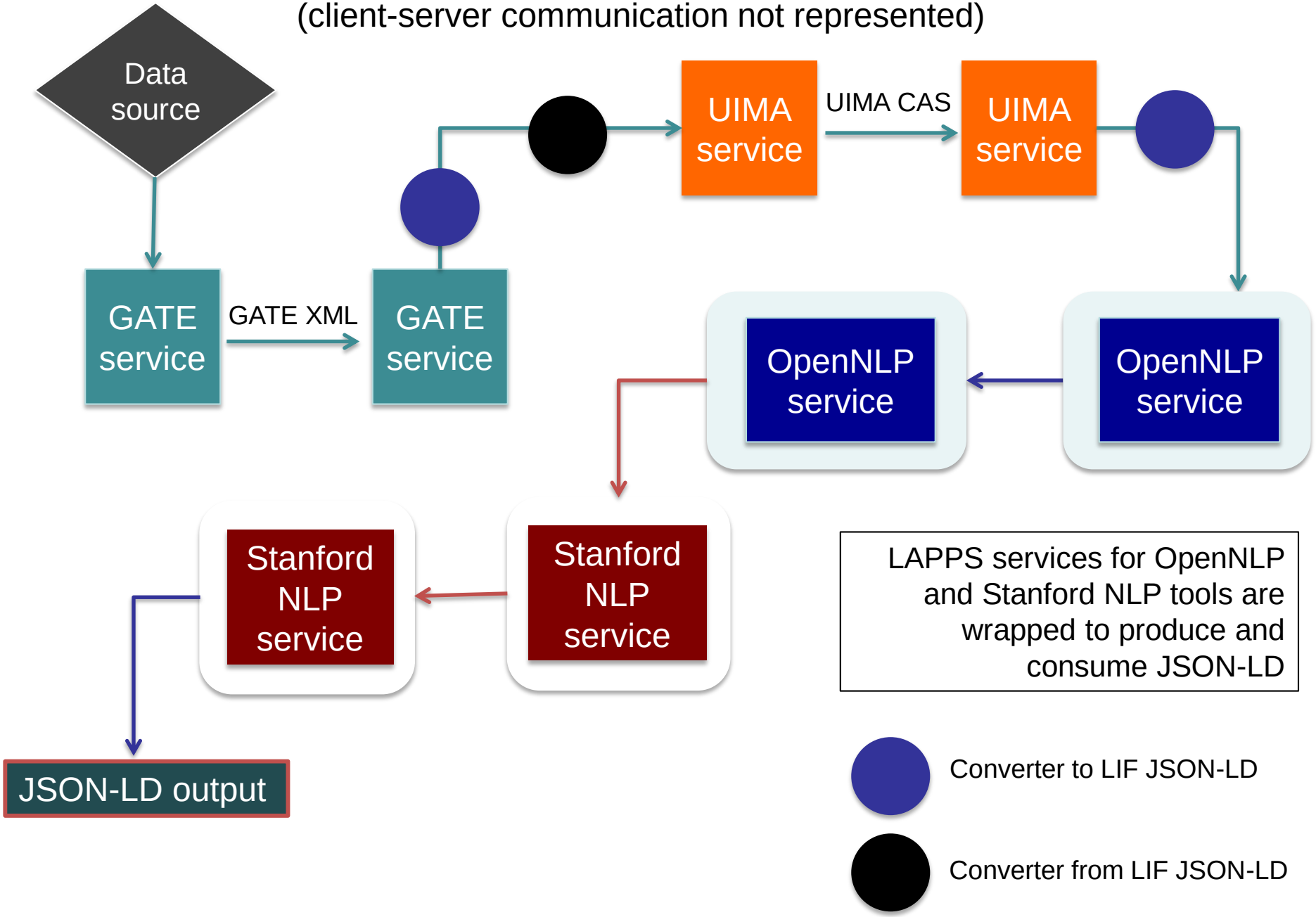
Properties

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

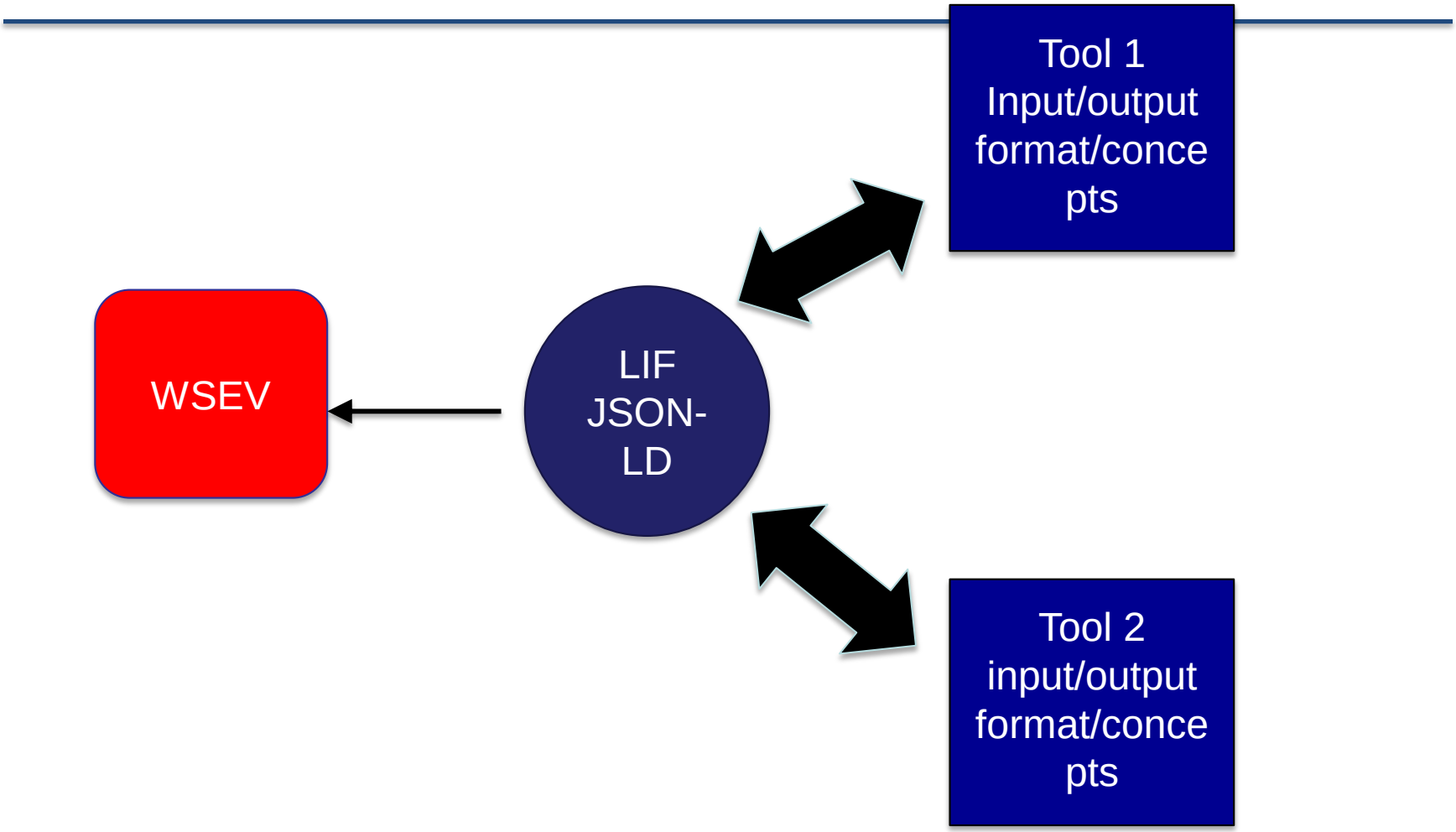
Logical flow

(client-server communication not represented)



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

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 CL courses at CMU, Brandeis, Vassar
  - ? IBM Watson

LAPPS Galaxy Interface

http://galaxyproject.org

- LAPPS recently adopted the **GALAXY workflow engine** as a front end for construction of pipelines etc.
  - open, web-based platform developed for computational genomics/biomedical research

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 !**

Galaxy

- Available as
  - A free public web server supported by the Galaxy Project (Johns Hopkins, Penn State)
    - Includes widely used bioinformatics tools
    - Users save histories, workflows, and datasets on the server, all can be shared with others
  - Open source software that can be downloaded and installed locally or in a cloud, and customized to address specific usages
  - Public web servers hosted by other organizations

LAPPS/Galaxy

- Multiple options for running a LAPPS/Galaxy instance
  - Use the **LAPPS/Galaxy web interface**
    - http://galaxy.lappsgrid.org
  - Create a **local Galaxy instance**:
    - Clone our fork of the Galaxy project, or
    - Run the LAPPS Grid appliance
      - a series of **docker images** that is a self-contained vm running Galaxy and all LAPPS services
      - useful when privacy required, no network connection available, etc.
  - Create an **instance in the cloud**
    - Useful for large datasets, computationally intense applications

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

Galaxy as Promoter of Open, Replicable Research

- The field of NLP research and development has been plagued by a chronic lack of **replicability** of results and information about provenance
- Galaxy provides for:
  - automatic recording of inputs, tools, parameters and settings used for each step in an analysis in a publicly viewable history
  - sharing datasets, histories, and workflows via web links
  - creation of custom web-based documents to communicate about an experiment or result
- **Encourages open publication, sharing of data and results**
- **Fosters replicability and reuse by providing a physical infrastructure to enable it**

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 Galaxy components
    - handled by LAPPS converters

Tools

search tools

Local Data
MASC
Gigaword
Tokenizers
Sentence Splitters
Taggers
Named Entity Recognizers
Parsers
Chunkers
Stanford NLP
GATE
Apache OpenNLP
Lingpipe
DKPro Core
DBpedia
Evaluation
Manual Conversion
Miscellaneous
Graph/Display Data

Workflows

- All workflows

LAPPS Grid

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

History

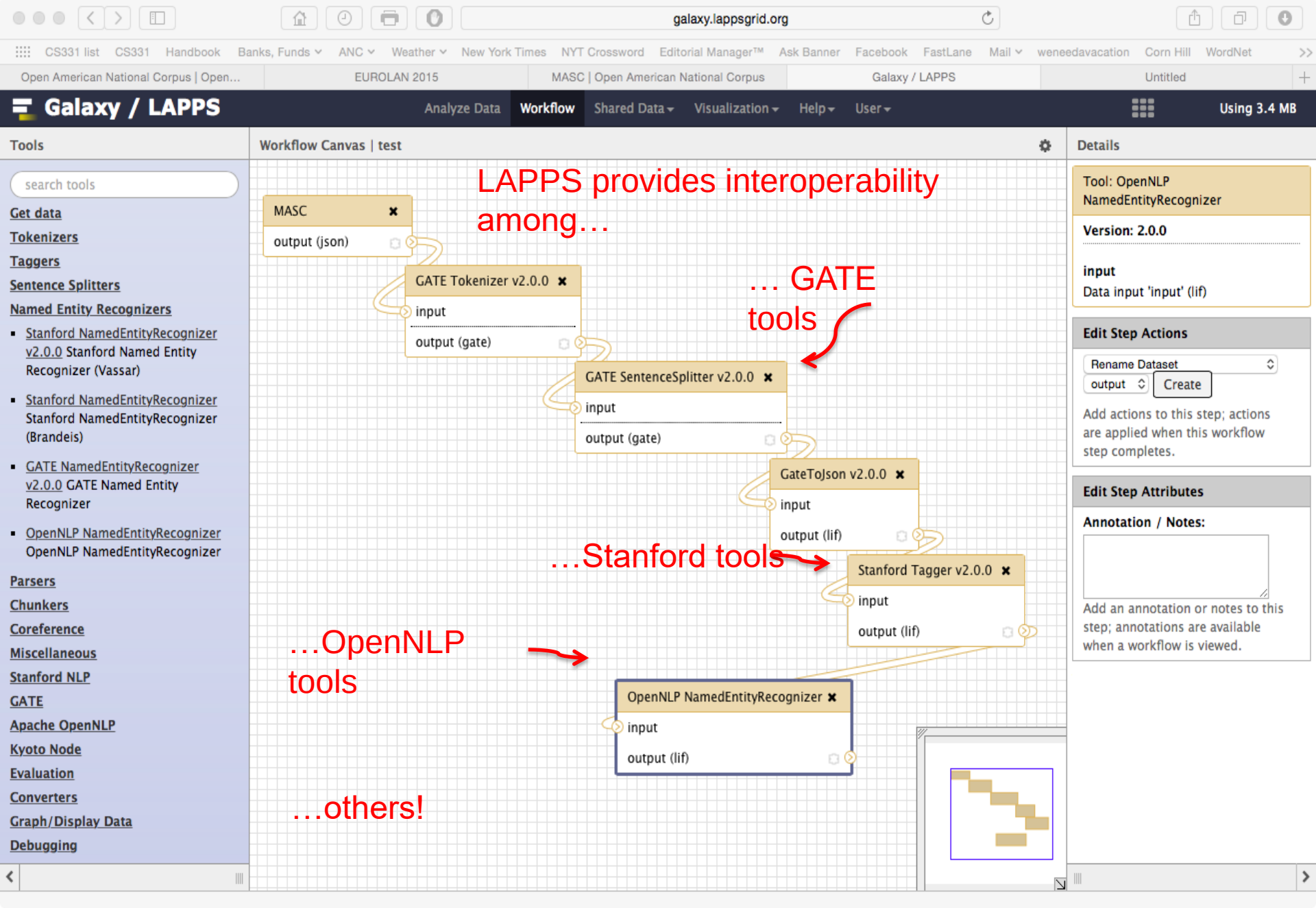
search datasets

Unnamed history

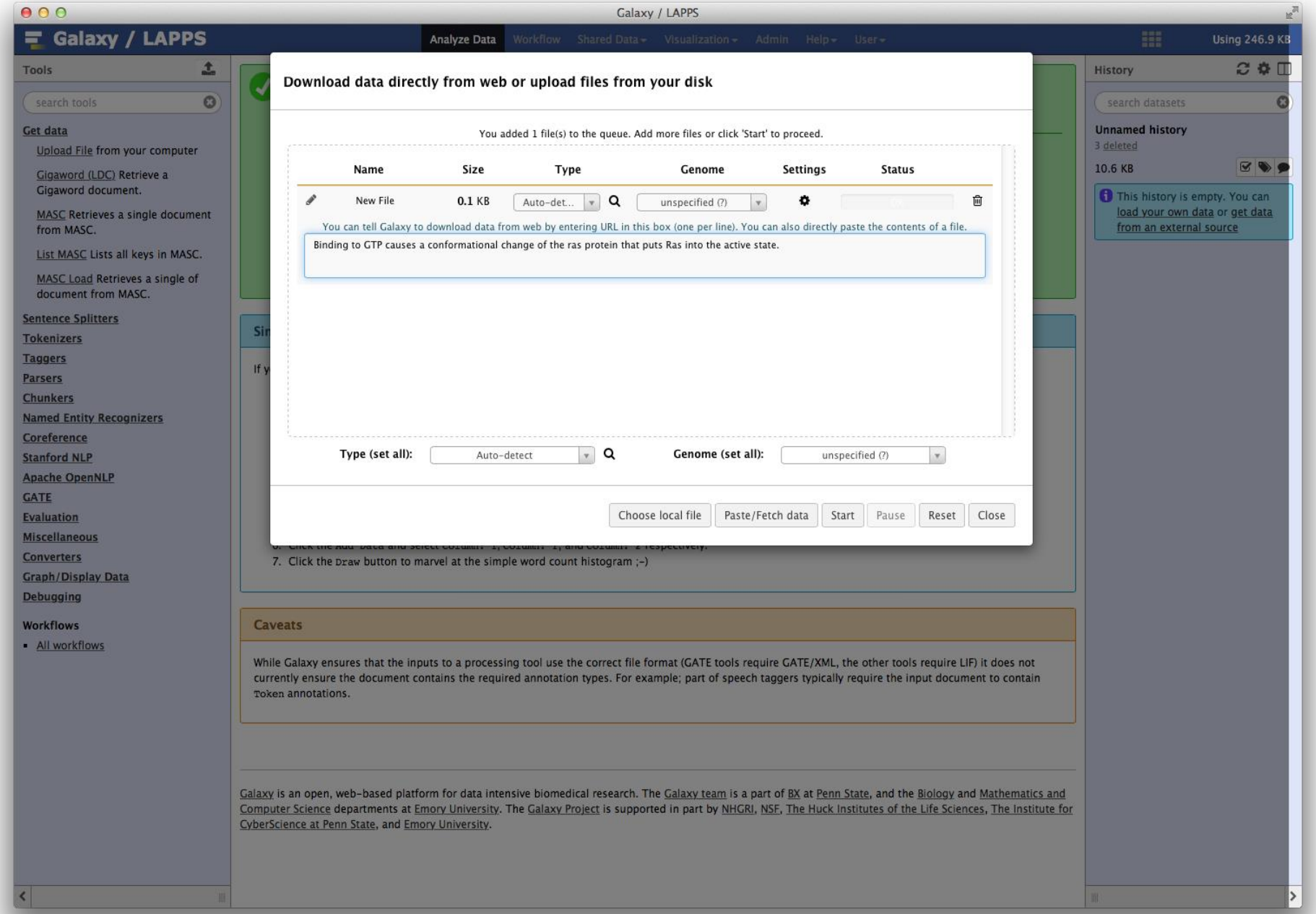
0 b

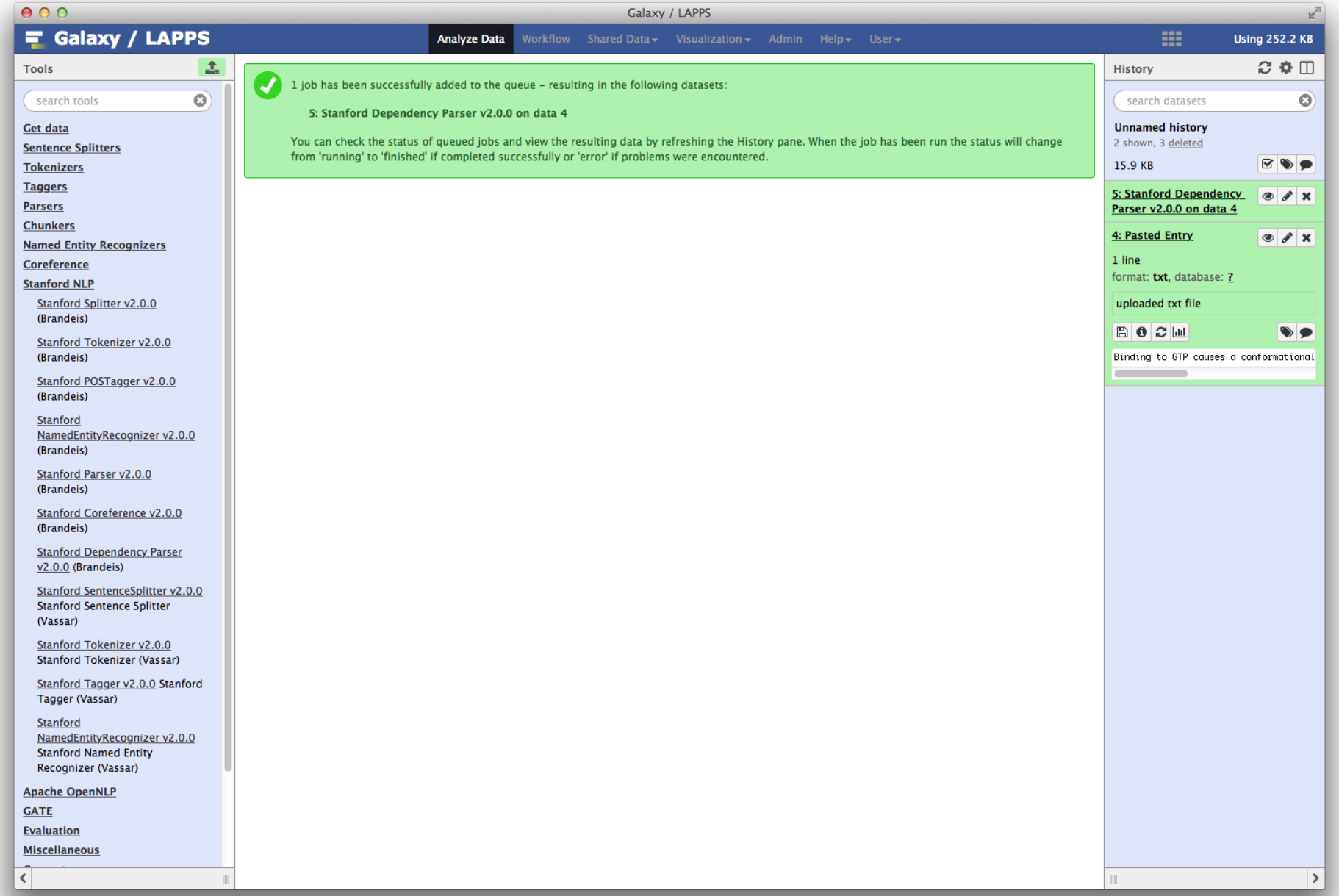
This history is empty. You can load your own data or get data from an external source

Workflow construction



Step-by-step analysis




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
1 job has been successfully added to the queue – resulting in the following datasets:
5: Stanford Dependency Parser v2.0.0 on data 4
You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.
History
search datasets
Unnamed history
2 shown, 3 deleted
15.9 KB
5: Stanford Dependency Parser v2.0.0 on data 4
4: Pasted Entry
1 line
format: txt, database: ?
uploaded txt file
Binding to GTP causes a conformational

Tools

search tools

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

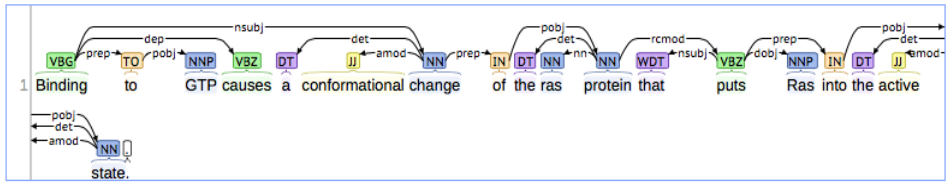
- Stanford Splitter v2.0.0 (Brandeis)
- Stanford Tokenizer v2.0.0 (Brandeis)
- Stanford POSTagger v2.0.0 (Brandeis)
- Stanford NamedEntityRecognizer v2.0.0 (Brandeis)
- Stanford Parser v2.0.0 (Brandeis)
- Stanford Coreference v2.0.0 (Brandeis)
- Stanford Dependency Parser v2.0.0 (Brandeis)
- Stanford 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



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":
```


History

search datasets

Unnamed history
2 shown, 3 deleted

15.9 KB

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":{"co
er":"edu.brandeis.cs.lappsgrid.stanford
rmational change of the ras protein tha
,"features":{"governor":"tk0_1","governo
":"prep","features":{"governor":"tk0_7"

4: Pasted Entry

1 line
format: **txt**, database: ?

uploaded txt file

Binding to GTP causes a conformational

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

1 job has been successfully added to the queue – resulting in the following datasets:

6: Stanford Parser v2.0.0 on data 5

You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

History

search datasets

Unnamed history

3 shown, 3 deleted

32.2 KB

6: Stanford Parser v2.0.0 on data 5

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":{"co
er":"edu.brandeis.cs.lappsgrid.stanford
rmational change of the ras protein tha
,"features":{"governor":"tk0_1","governo
":"prep","features":{"governor":"tk0_7"

4: Pasted Entry

1 line

format: **txt**, database: ?

uploaded txt file

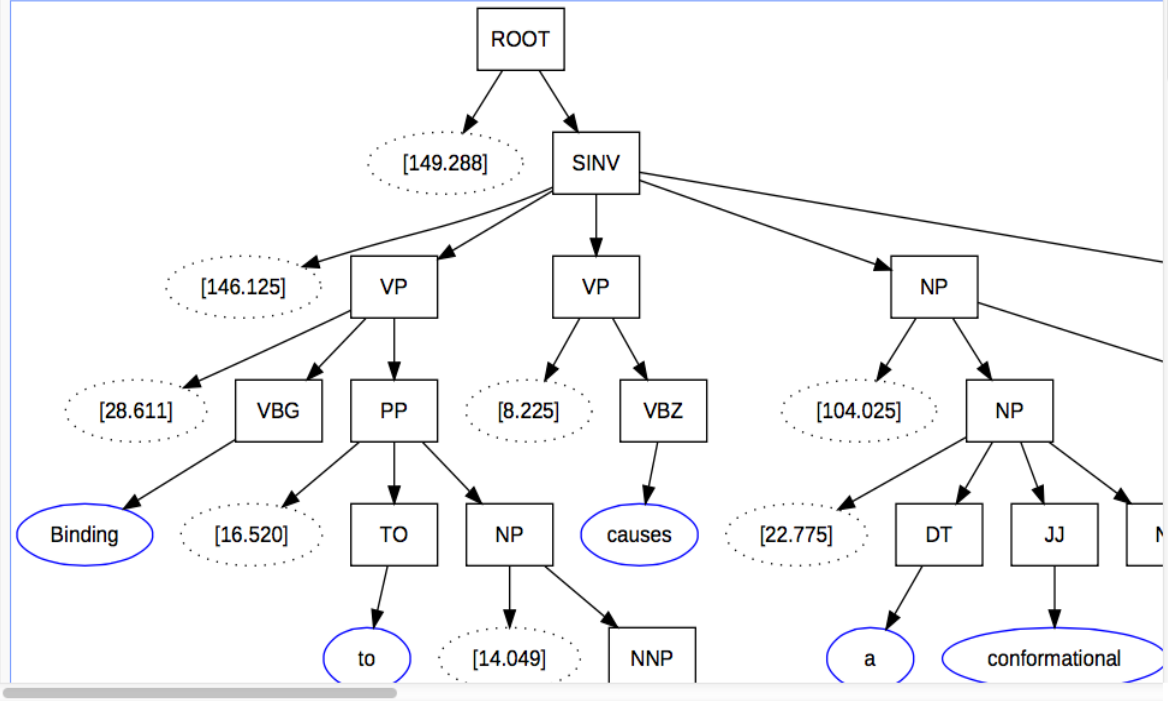
Binding to GTP causes a conformational

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

```
1  Binding to GTP causes a conformational change of the ras protein that puts Ras into the active state.
2  ~~~~
3  (ROOT [149.288]
4  (SINV [146.125]
5  (VP [28.611] (VBG Binding)
6  (PP [16.520] (TO to)
7  (NP [14.049] (NNP GTP))))
8  (VP [8.225] (VBZ causes))
9  (NP [104.025]
10 (NP [22.775] (DT a) (JJ conformational) (NN change))
11 (PP [78.905] (IN of)
12 (NP [77.575]
13 (NP [26.141] (DT the) (NN ras) (NN protein))
14 (SBAR [49.283]
15 (WHNP [1.447] (WDT that))
16 (S [47.386]
17 (VP [47.110] (VBZ puts)
18 (NP [15.584] (NNP Ras))
19 (PP [20.534] (IN into)
20 (NP [16.071] (DT the) (JJ active) (NN state)))))))))
21 (. .)))
```



History

search datasets

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":{"co
er":"edu.brandeis.cs.lappsgrid.stanford
rmational change of the ras protein ths
,"features":{"governor":"tk0_1","governor
":"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":{"co
er":"edu.brandeis.cs.lappsgrid.stanford
rmational change of the ras protein ths
,"features":{"governor":"tk0_1","governor
":"prep","features":{"governor":"tk0_7"

4: Pasted Entry
1 line
format: txt, database: ?

uploaded txt file

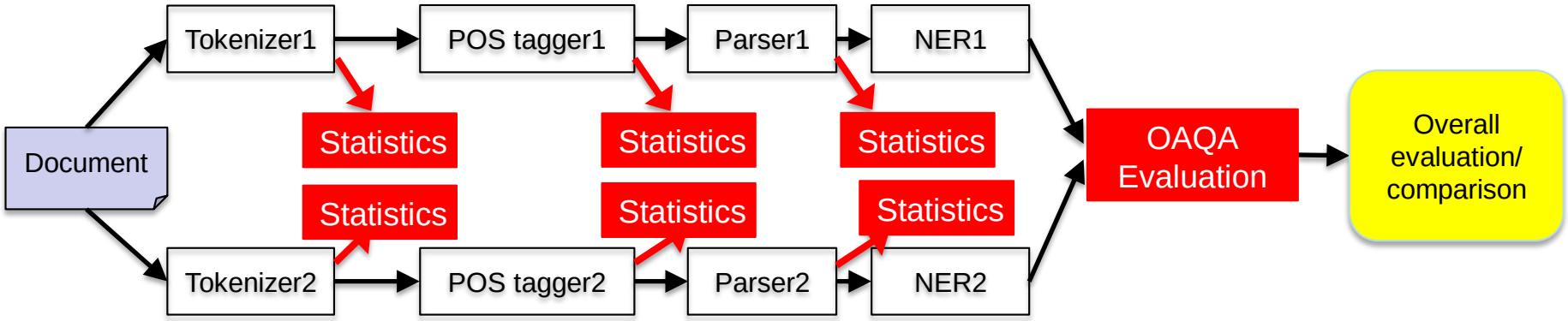
Binding to GTP causes a conformational

Evaluation in LAPPS/Galaxy

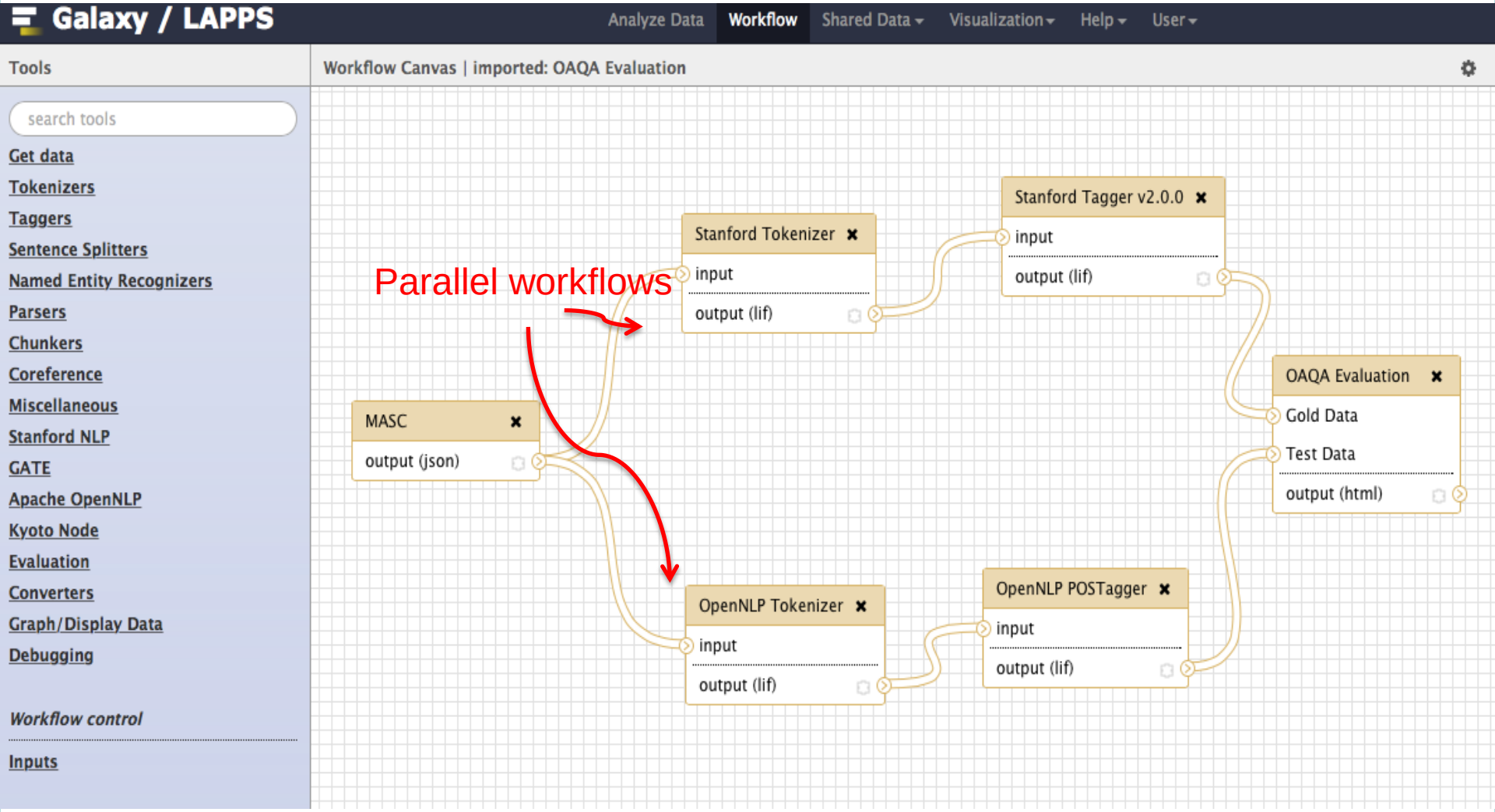
- CMU has implemented services for state-of-the-art **Open Advancement** techniques
- Enables rapid identification of
  - **frequent error categories** within modules and documents
  - which module(s) and error type(s) have the greatest **impact on overall performance**
- Used in the development of IBM's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


Galaxy / LAPPS
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Tools
Workflow Canvas | imported: OAQA Evaluation
search tools
Get data
Tokenizers
Taggers
Sentence Splitters
Named Entity Recognizers
Parsers
Chunkers
Coreference
Miscellaneous
Stanford NLP
GATE
Apache OpenNLP
Kyoto Node
Evaluation
Converters
Graph/Display Data
Debugging
Workflow control
Inputs
Parallel workflows
MASC
output (json)
Stanford Tokenizer
input
output (lif)
Stanford Tagger v2.0.0
input
output (lif)
OpenNLP Tokenizer
input
output (lif)
OpenNLP POSTagger
input
output (lif)
OAQA Evaluation
Gold Data
Test Data
output (html)

MASC gold standard vs. Stanford NEs

Precision: 0.391
Recall: 0.473
F1: 0.428

Reference Outputs			Predicted Outputs			
Start	**End**	**Features**	**Start**	**End**	**Features**	**Text**
70	80	LOCATION				Asia Minor
			70	74	LOCATION	Asia
			82	85	DATE	now
93	99	LOCATION	93	99	LOCATION	Turkey
104	110	LOCATION	104	110	LOCATION	Greece
145	151	LOCATION	145	151	LOCATION	Aegean
389	402	LOCATION				Mediterranean
429	438	DATE				7000 b.c.
			429	433	DATE	7000
			434	438	DATE	b.c.
			478	481	DATE	now
490	494	LOCATION	490	494	LOCATION	Iran
647	657	LOCATION				Aegean Sea
			647	653	LOCATION	Aegean
			654	657	LOCATION	Sea
738	743	LOCATION	738	743	PERSON	Milos
899	904	LOCATION	899	904	PERSON	Milos

Potential benefits of LAPPS/Galaxy collaboration

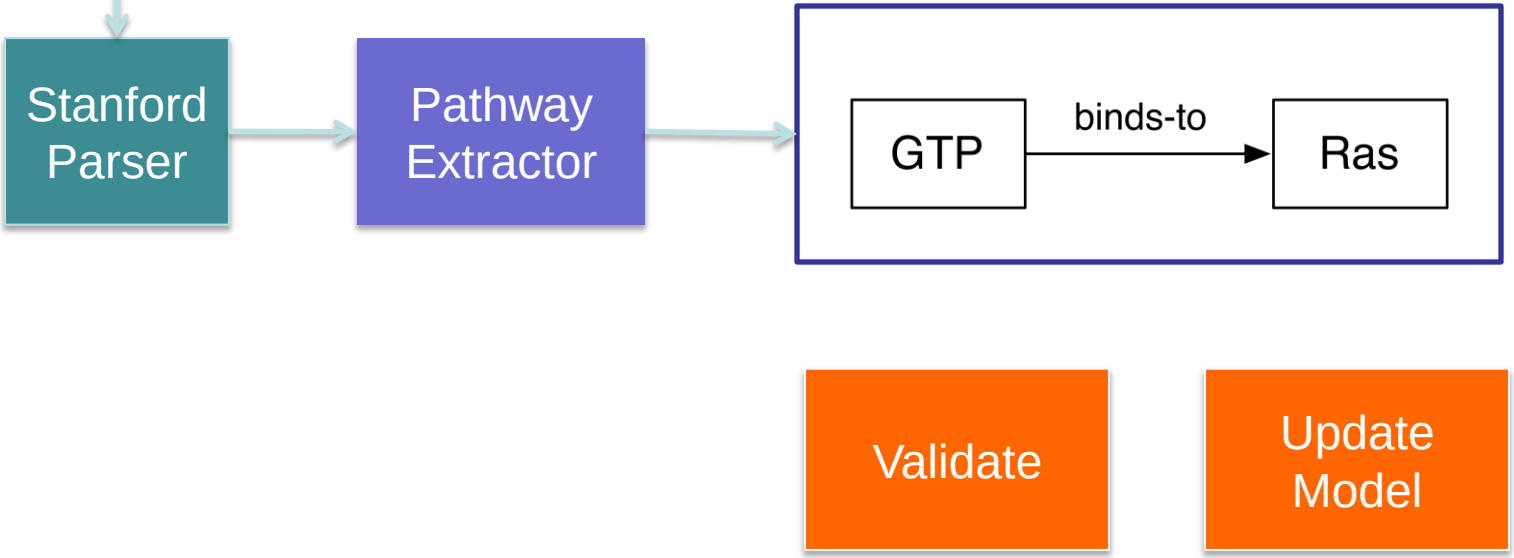
- Galaxy contains a huge number of tools for analyzing genomic and other biomedical data
- LAPPS includes tools to perform NLP analyses
- Combining LAPPS services with Galaxy tools can allow for analysis of data mined from the vast stores of biomedical literature (Biomed, PubMed, PLOS, etc.)

- **BIONLP meets bio-analysis!**

Example

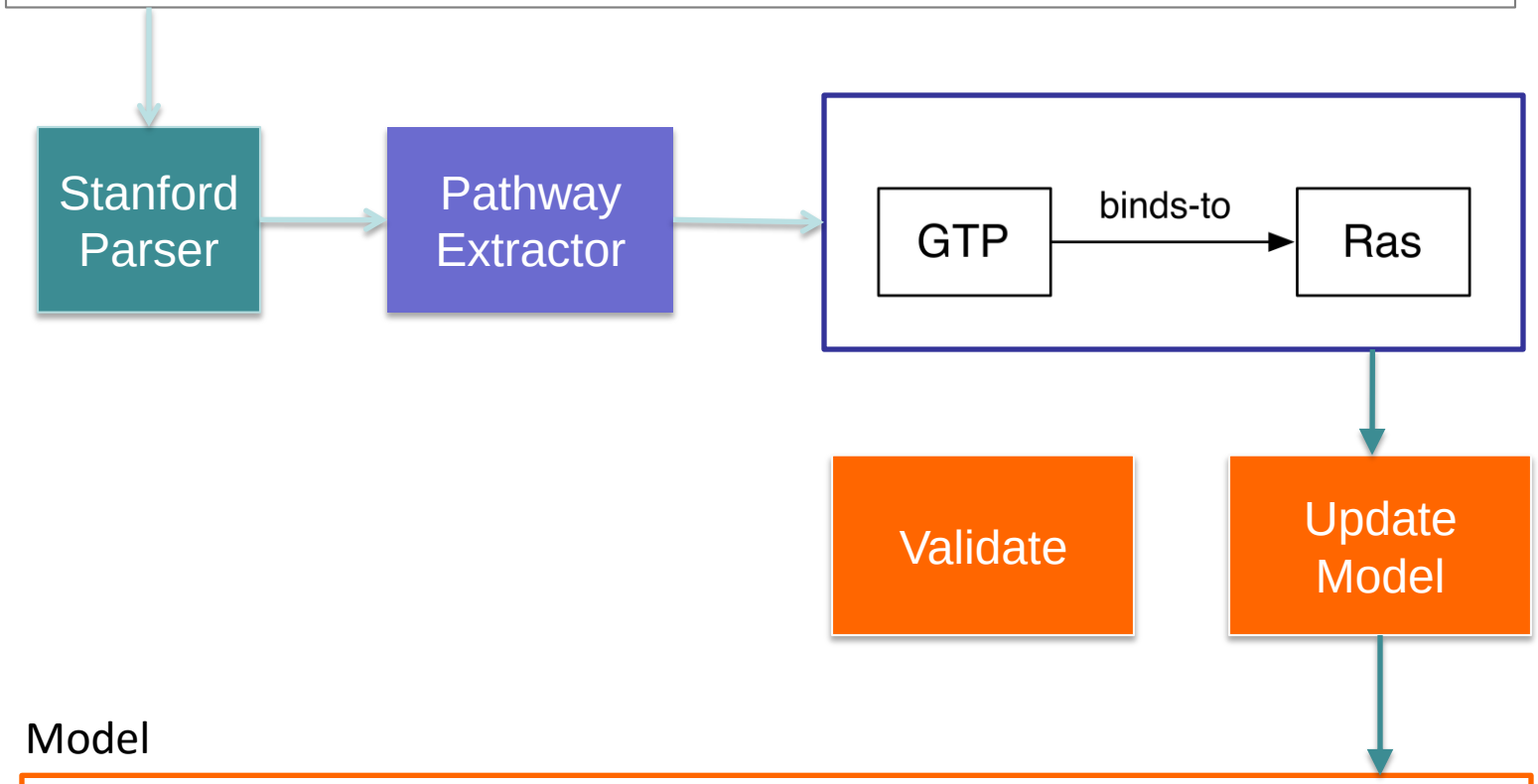
Binding to GTP causes a conformational change of the ras protein that puts Ras into the active state. GTP-bound ras binds to the raf protein kinase. This binding of raf to ras has the effect of activating the raf kinase and localizing the raf kinase to the cell membrane. Activated raf now phosphorylates 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.

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

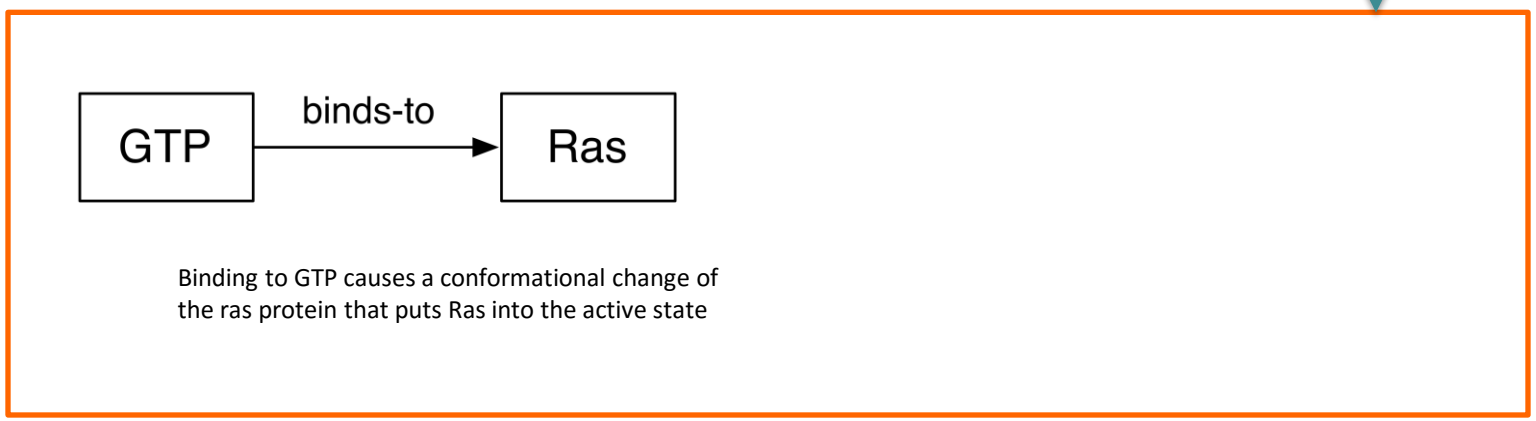


Model

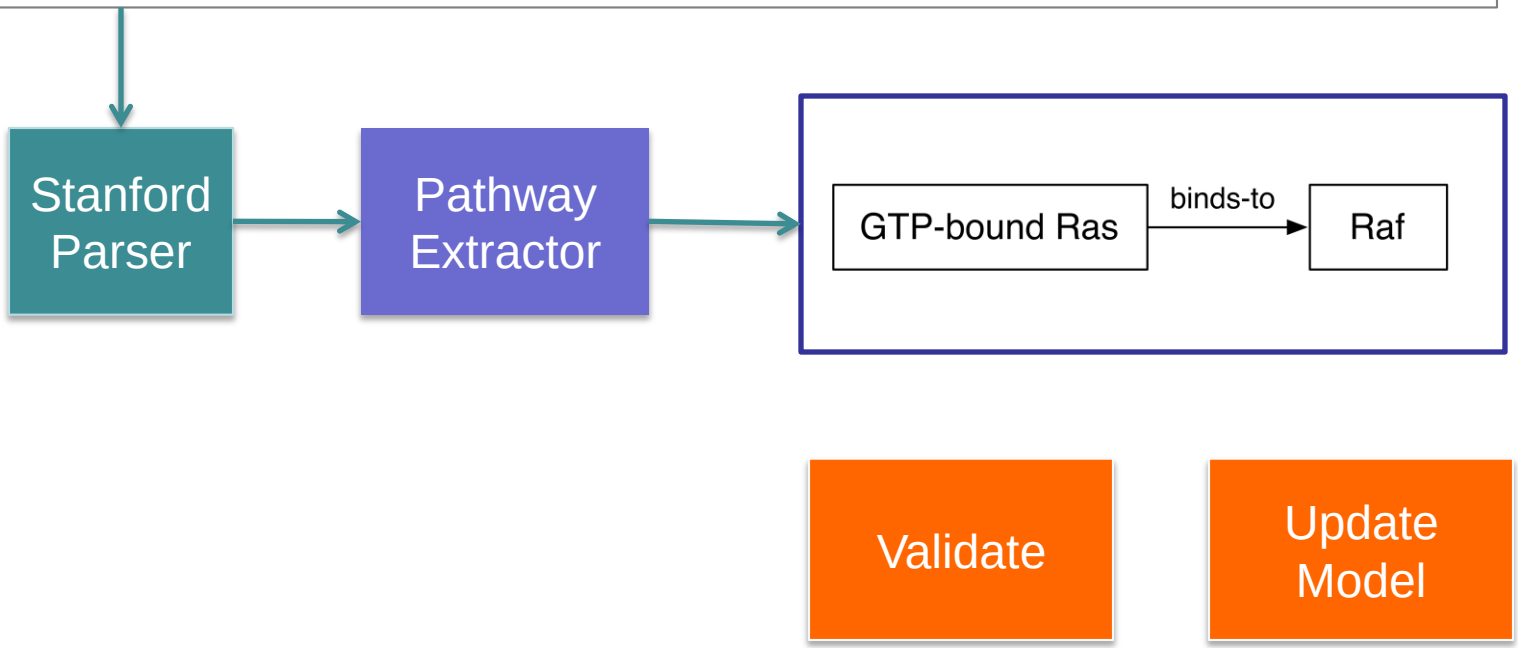
Binding to GTP causes a conformational change of the ras protein that puts Ras into the active state



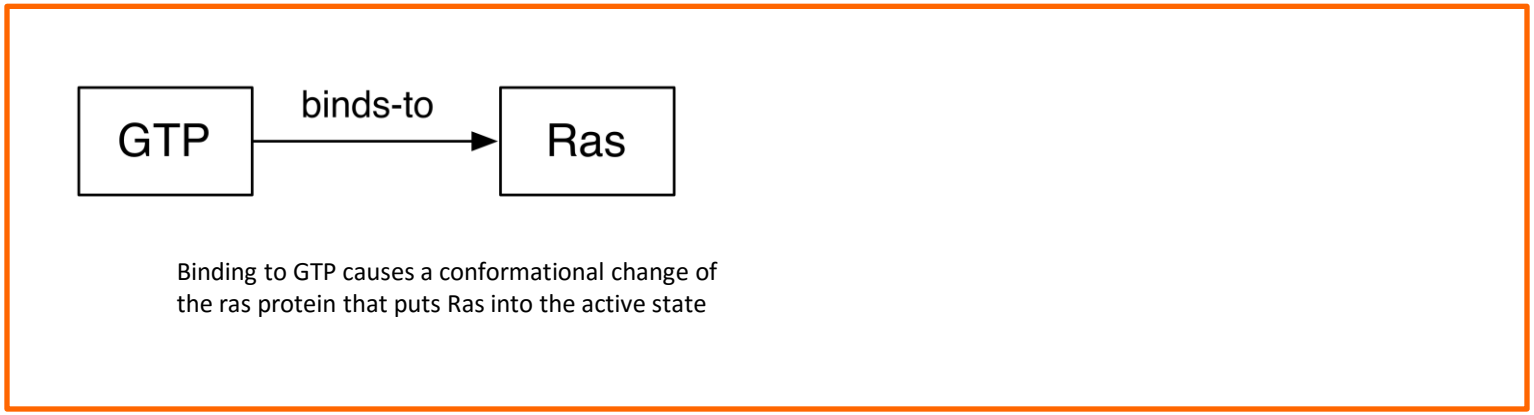
Model

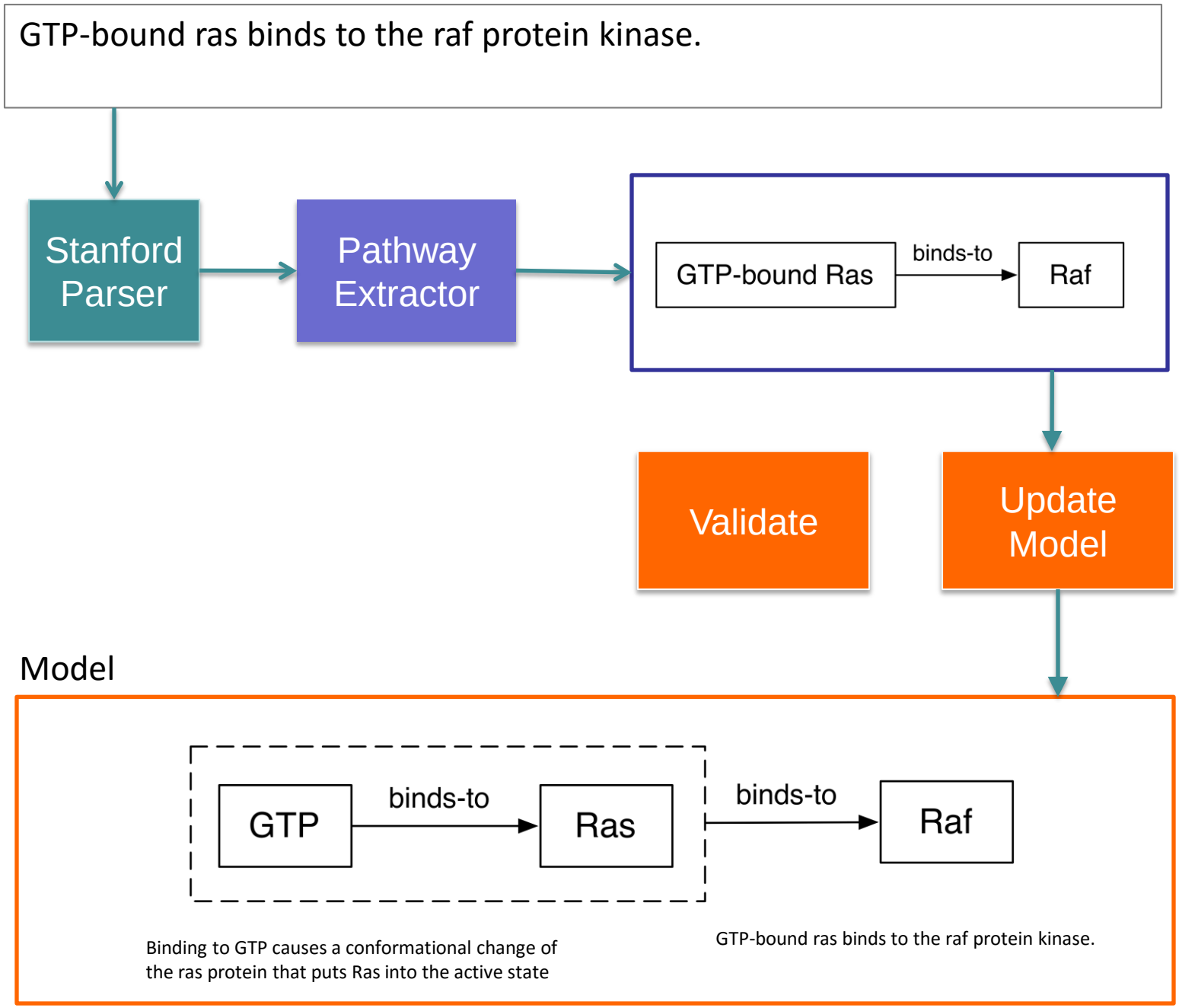


GTP-bound ras binds to the raf protein kinase.



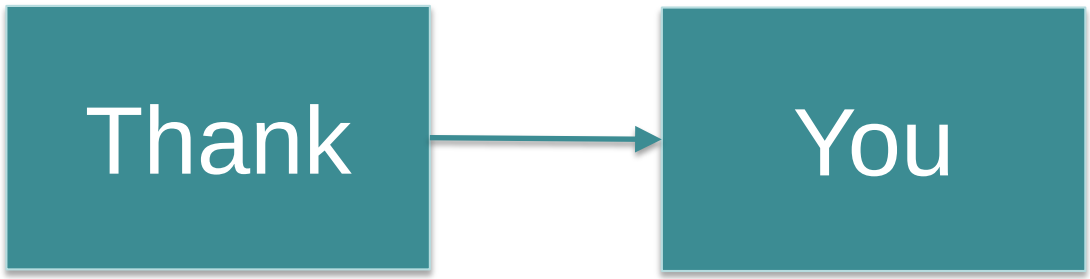
Model




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.

LAPPS/GALAXY Demo

P31
LR Infrastructures and Architectures
Thursday May 26
14:55-16:35
Poster Area 1


Thank
You