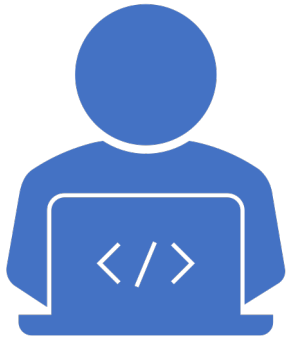
Challenges for Scientific Publication Mining

Nancy Ide

Department of Computer Science

Vassar College

Poughkeepsie, New York USA



The Problem

Scientific literature is growing at an exponential rate

- Too much information for anyone to read, much less understand
- Researchers become **increasingly specialized**
- Rise of specialized, non-interacting literatures
  - Create **islands of knowledge**, discoveries in one area not known outside of it
- Difficult for researchers to stay current in even their narrow discipline

Scientific Information Overload

The global research community generates ~2.5 million new scholarly papers per year (English only)

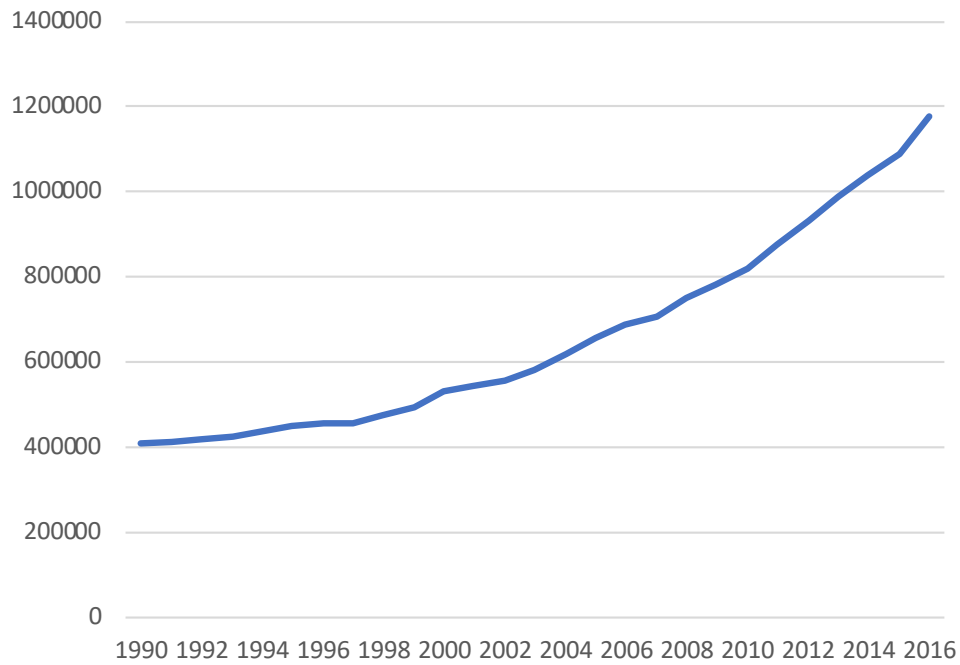
A new research paper is published every 12 seconds

70,000 papers published on a single protein

Challenge to scientists:

Keep updated on new developments, paper writing, project proposal preparation, paper reviewing, peer assessment, etc.

Publications Added to PubMed 1990-2016



Total for all years close to 30 million articles

PubMed is now accumulating over 1,000,000 new entries every year

Drowning in data, starving for knowledge

Herbert A. Simon (1916–2001) pointed out 40 years ago:

> "A wealth of information creates a poverty of attention and a need to allocate that attention efficiently among the overabundance of information sources that might consume it"

Simon, H.A.: Designing Organizations for an Information-Rich World. In: Greenberger, M. (ed.) Computers, Communication, and the Public Interest, pp. 37–72. The Johns Hopkins Press, Baltimore (1971)

Solution

Exploit techniques from the field of Natural Language Processing (NLP)

1

Information Retrieval yields all relevant texts

- Gathers, selects, filters documents that may prove useful
- Finds what is known

2

Information Extraction extracts facts and events of interest to user

- Finds relevant concepts, facts about concepts
- Finds only what we are looking for

3

Text (document) Mining discovers unsuspected associations

- Combines and links facts and events
- Discovers new knowledge, finds new associations

Processes

A large number of linguistic approaches to processing of scientific publications

- Extensive use of **linguistic information** such as grammatical relations and word order together with **semantic resources** such as ontologies and controlled vocabularies

Major technologies:

- **Named entity recognition**
- **Relation extraction**
- **Event extraction**

Supported by **statistical analysis** and **machine learning**

Focus : Biomedical Publications +BioNLP

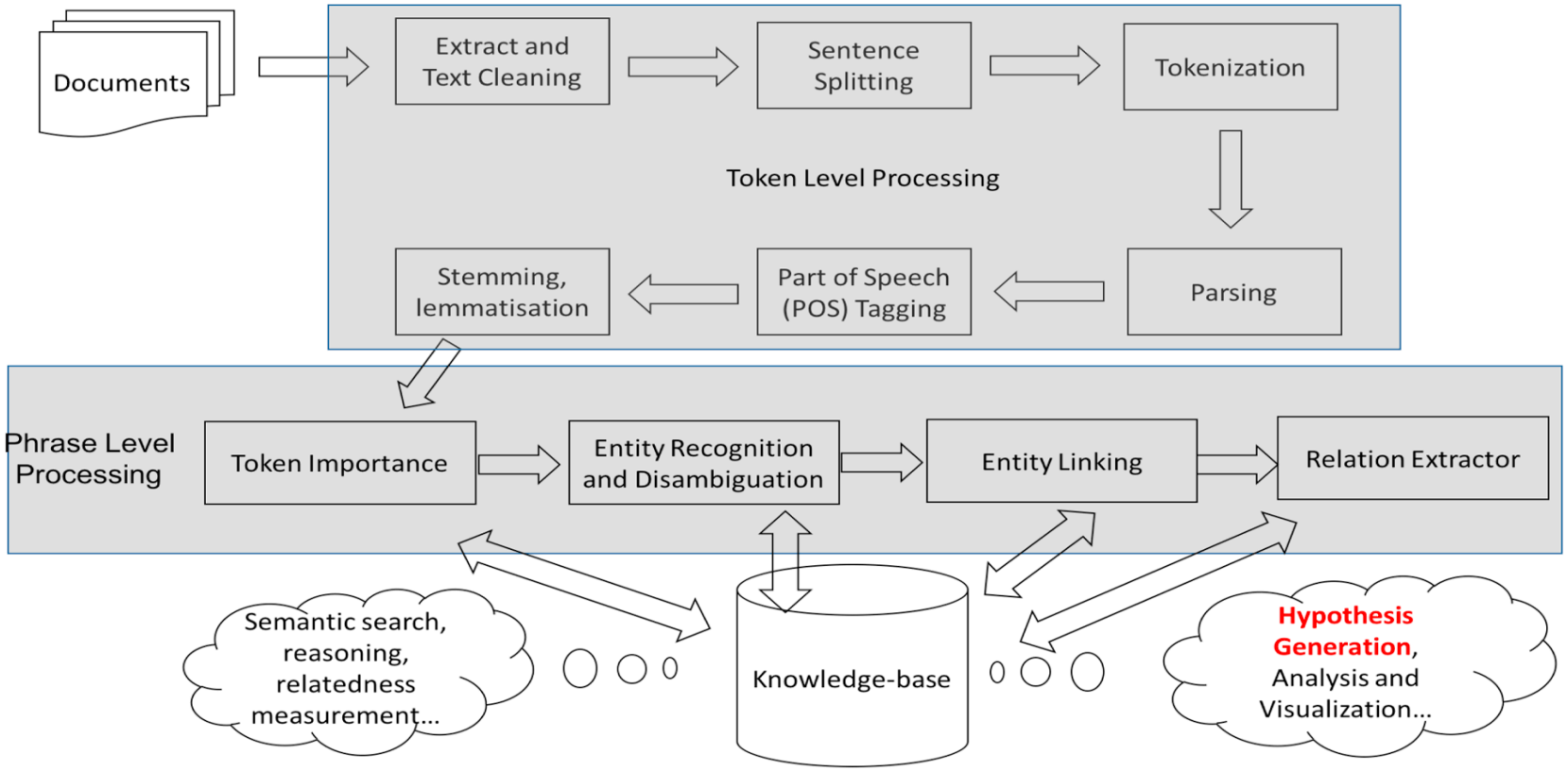
Biomedical literature offers a rich set of knowledge sources to discover important facts and find associations among them

Demonstrate the range of issues, obstacles to text mining

Major processing tasks performed on biomedical text

1. Identify and classify **biomedical entities** (NER) into predefined categories such as proteins, genes, or diseases
2. Infer **pair-wise relationships** among named entities e.g., protein-protein interaction gene-protein, and medical problem-treatment

Typical Framework



Major processing tasks performed on biomedical text

1. Identify and classify **biomedical entities** (NER) into predefined categories such as proteins, genes, or diseases
2. Infer **pair-wise relationships** among named entities e.g., protein-protein interaction gene-protein, and medical problem-treatment

Named Entity Recognition

- The most fundamental task in biomedical text mining is the recognition of named entities (called **Named Entity Recognition** or **NER**), such as proteins, species, diseases, chemicals or mutations
- Commonly approached as a **supervised learning problem**
- NER systems may require considerable **manual feature engineering** to learn robust models using **hand-labeled training data**

Challenges

NER is made challenging by the nature of biomedical texts, e.g.

- Heavy use of **domain specific terminology** (12% biochemistry-related technical terms)
- Constant introduction of **new terms** and **short forms or abbreviations**
- Most words have low frequency (**data sparseness**)
- **Complex co-referential links**
- **Complex mapping from syntax to semantics**

Traditional Biomedical NER Methods

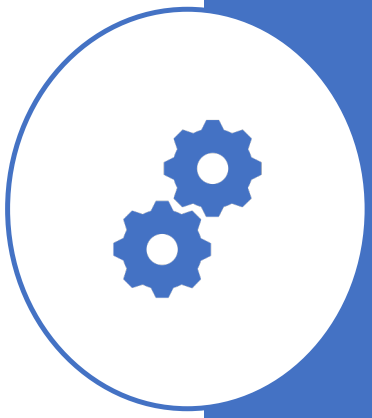
Rule-based techniques

- Recognize biomedical entities using **manually defined rules** based on textual patterns
  - E.g., the suffix '-ase' is more frequent in *protein names* than in *diseases*

Dictionary-based methods

- Extract named entities by **searching for them in dictionaries** constructed for each entity type

- Time consuming to create rules and dictionaries, requires domain-expert knowledge
- Recall obtained using these methods is generally low due to the inherent difficulty of capturing new entities



Machine Learning

Over the last years, pattern- and dictionary- based methods superseded by approaches relying on **sequential classification algorithms**

ML-based methods for BioNLP dominated by **feature-based** and **kernel-based** methods

- **Supervised Learning**
  - De facto standard model: Conditional Random Fields (CRFs)
- **Semi-supervised learning**
  - Use small amount of labeled data with a large amount of unlabeled data
  - Use assumptions about smoothness, low dimensional structure, or distance metrics to leverage unlabeled data

Feature-based Methods

- Deriving good features is difficult, time-consuming, and requires expert knowledge
  - Currently more of an art than a science
  - Incurs extensive trial-and-error experiments

Kernel-based Methods

- Attempt to solve this problem by implicitly calculating dot products for every pair of examples
  - Apply a similarity function between examples and use a discriminative method to label new examples
  - Requires manual effort to design an appropriate similarity function
  - High computational complexity

Feature Engineering in NER

State-of-the-art tools are entity-specific

- Empirically optimal feature sets differ between entity types
  - Costly to develop

Features are often optimized for a specific gold standard corpus

- Not reusable
- Extrapolation of quality measures difficult

Paradigm Shift

Important recent developments

1. Word embeddings

- Represent a single word by a low-dimensional vector capturing the frequencies of co-occurring adjacent words
  - Vs. bag-of-words approach underlying conventional methods
- Capture semantic similarities between words (as mathematical similarities between their vectors) not visible from surface
  - E.g., ‘enables’ and ‘allows’ are syntactically different, but meaning is related (thus similar sets of co- occurring words, vs. co-occurrences of the word ‘swim’)

The underlying idea of representing words ‘by the company they keep’ is an old concept in linguistics, usually called **distributional semantics**

Important recent developments

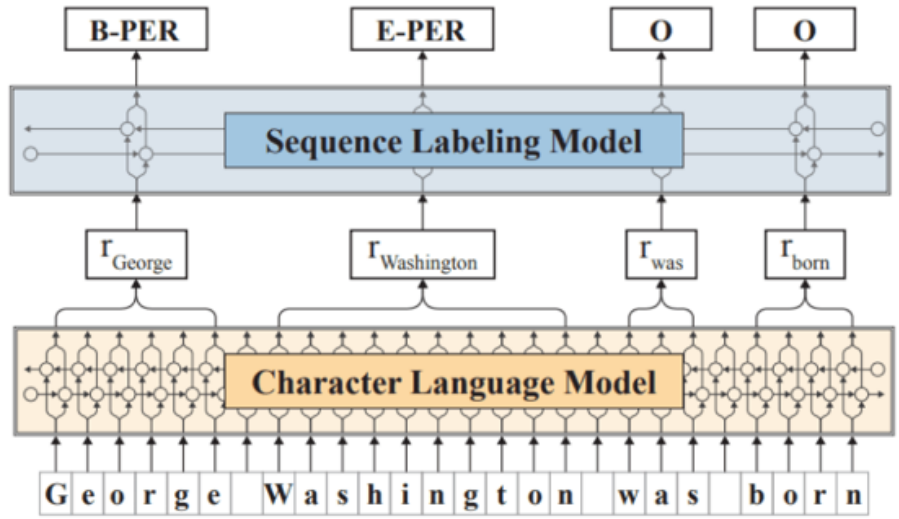
2. Artificial neural networks (ANNs)

- Automatically learn non-linear combinations of features
- Better recognition results than CRFs, which only learn (log-)linear combinations of features
- **Deep neural networks** -- especially bidirectional long short-term memory networks (BiLSTM) -- learn efficiently and effectively

Idea is not new, but recent progress in the size of available data and machine capabilities make it applicable to practically relevant problems

LSTM-CRF

- Most commonly used in recent work: bidirectional LSTMs with a sequential conditional random layer above
- Method:
  - Character language model *C* pre-trained on huge corpora
  - *C* is used to create contextual word embeddings *W*
  - *W* fed to a BiLSTM+CRF classifier that classifies the input tokens one by one



No More Feature Engineering?!

- Recent success in deep learning for NER (Lample et al., 2016) suggests that **automatic feature extraction will largely replace feature engineering**
- Semi-supervised methods that augment labeled datasets with **word embeddings** outperform supervised baselines in tasks like gene name recognition

However, this shifts the burden to constructing the massive hand-labeled training sets needed for robust deep models

The Annotation Bottleneck

High accuracy NER systems *still* require **manually annotated named entity datasets** for training and evaluation

- Deep learning models are massively more complex than traditional models
  - May have hundreds of millions of free parameters
  - Require commensurately more labeled training data
- Need is even more pronounced for biomedical language
  - General-purpose annotated corpora (e.g., product reviews, Wikipedia articles) are not specific for biomedical language
    - Rarely contain concepts of interest to biologists or clinicians
    - Must develop specialized corpora

Gold Standard Resources are expensive to create!

- Annotated corpora and knowledge sources such as lexicons, ontologies typically contain manual input by highly trained domain specialists
- Cost dictates that resources are
  - limited in size
  - not available for many sub-domains and specialized areas
- **Result: many NER systems suffer from poor performance**

?

So, how do we obtain enough training data to fit complex deep learning models?

Crowdsourcing

- One way of generating large-scale labeled data
- Can be expensive
  - Annotators may require specialized domain knowledge
- Even expert inter-annotator agreement rates can be low for certain tasks

Distant supervision

Leverage structured resources like ontologies and knowledge bases to label training data

- Noisy, but has shown empirical success

- Drawback for BioNLP: the **wide space of curated resources**
  - NCBO Bioportal (Whetzel et al., 2011) currently houses 541 distinct biomedical ontologies
    - Contain **different hierarchical structures, concept granularities**, and otherwise **overlap or conflict** in their definitions of 8 million entities
    - Any single ontology may have widely varying accuracy depending on the target task
  - **Difficult to combine** using simple methods like majority vote

Active Learning

- After a round of supervised learning, select additional data points for labeling that are estimated to be **most valuable for improving the model**

Transfer Learning

- "Pre-train" a model on one or more datasets, and "fine-tune" it on the task of interest on another dataset

Multi-task Learning

- Use multiple annotated datasets together to train a model for improved performance on a single dataset

Weak Supervision

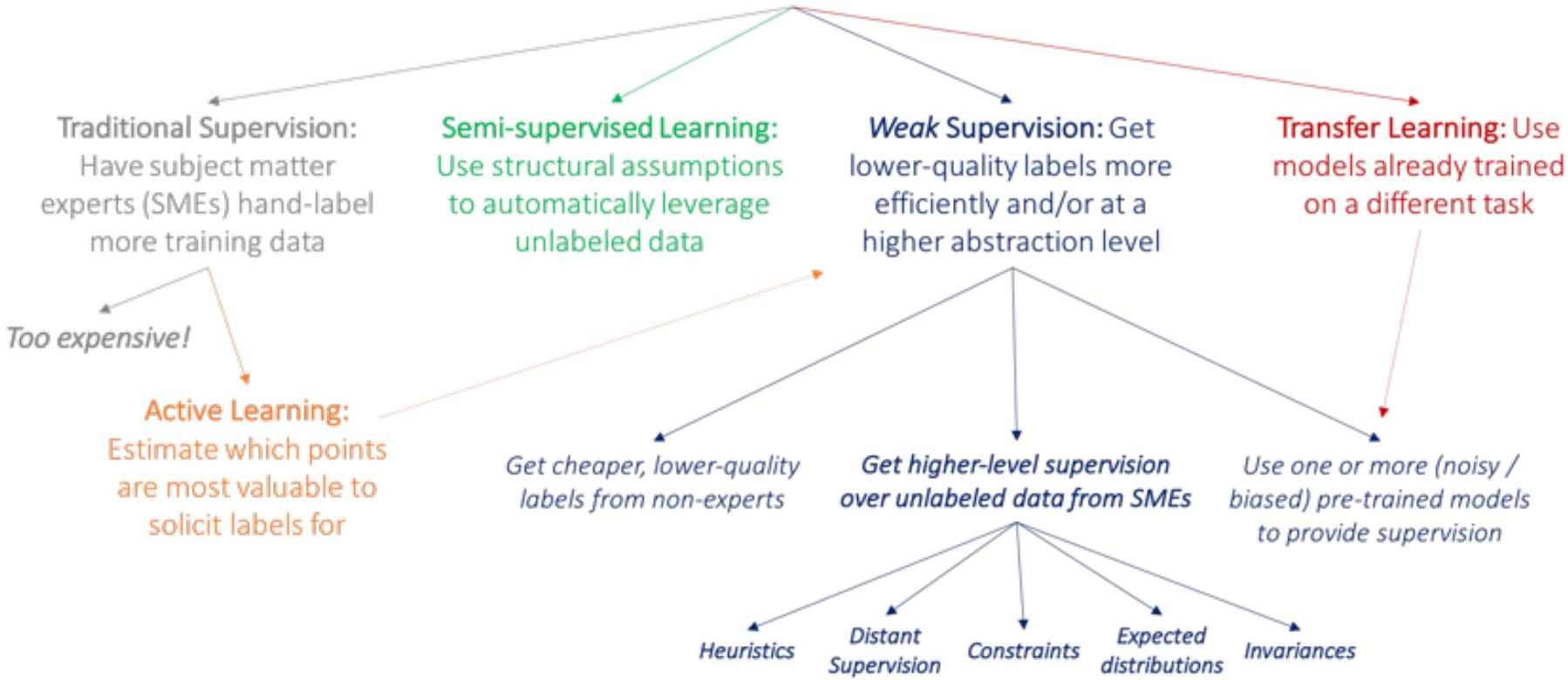
The current trend

- Create noisier, lower-quality, but larger-scale training sets
  - Constructed via strategies such as
    - using cheaper annotators
    - programmatic scripts
    - more creative and high-level input from domain experts
    - etc.

Advantages of Weak Supervision

- Annotators can provide higher-level, more expressive input
- **Can be robust to inevitable lack of precision, coverage, or conflict resolution in this input**
- Can define flexible and interpretable paradigms for how to interact with, supervise, and essentially "program" machine learning models
  - See e.g., Ratner et al. (2016), Data Programming: Creating Large Training Sets, Quickly. *Advances in neural information processing systems*. 29.

Overview of Methods



Major processing tasks performed on biomedical text

1. Identify and classify **biomedical entities** (NER) into predefined categories such as proteins, genes, or diseases
2. Infer **pair-wise relationships** among named entities e.g., protein-protein interaction gene-protein, and medical problem-treatment

Relation Extraction

The task of extracting semantic relationships from a text

- Usually occur between two or more entities of a certain type
- General RE
  - Entity types e.g. Person, Organization, Location
  - Semantic categories e.g., married to, employed by, lives in
- RE from biomedical texts
  - Interactions between biomolecules
  - Events occurring subsequently over time (temporal relationships)
  - Causal relationships

Relations in UMLS: Unified Medical Language System

Injury	disrupts	Physiological Function
Bodily Location	location-of	Biologic Function
Anatomical Structure	part-of	Organism
Pharmacologic Substance	causes	Pathological Function
Pharmacologic Substance	treats	Pathologic Function

Relation extraction from text

Doppler echocardiography can be used to **diagnose** left anterior descending artery **stenosis** in patients with type 2 diabetes

Echocardiography, Doppler DIAGNOSES Artery stenosis

Methods for Building Relation Extractors

- Hand-written patterns
- Supervised machine learning
- Semi-supervised and unsupervised
- Bootstrapping (using seeds)
- Distant supervision
- Unsupervised learning from the web

Hand-built patterns for relations

Plus:

- Human patterns tend to be high-precision
- Can be tailored to specific domains

Minus

- Human patterns are often low-recall
  - A lot of work to think of all possible patterns
- Don’t want to have to do this for every relation
- Need better accuracy

Supervised relation extraction

- Train classifier with gold standard data annotated for entities and their relations
- **Gazeteer** and **trigger word features** for relation extraction
  - Trigger list for *family*: kinship terms
    - parent, wife, husband, grandparent, etc. [from WordNet]
  - Gazetteer: Lists of useful geo or geopolitical words
    - Country name list

Semi-supervised

Seed-based or bootstrapping approaches to relation extraction

- Bootstrapping: use seeds to directly learn to populate a relation

Author	Book
Isaac Asimov	The Robots of Dawn
David Brin	Startide Rising
James Gleick	Chaos: Making a New Science
Charles Dickens	Great Expectations
William Shakespeare	The Comedy of Errors

- Extract patterns

  The Comedy of Errors, by **William Shakespeare**, was

  The Comedy of Errors, by **William Shakespeare**, is

  The Comedy of Errors, one of **William Shakespeare**'s earliest attempts

  The Comedy of Errors, one of **William Shakespeare**'s most

- Iterate, finding new seeds that match the pattern

Distant Supervision

Combine bootstrapping with supervised learning

- Instead of a few seeds, use a large database to get huge number of seed examples
- Create features from these examples
- Combine in a supervised classifier

Unsupervised relation extraction

- **Open Information Extraction**
  - Extract relations from the web with no training data, no list of relations
  - Use parsed data to train a “trustworthy tuple” classifier
  - Single-pass extract all relations between NPs, keep if trustworthy
  - Assessor ranks relations based on text redundancy
- Drawbacks
  - No gold set of correct instances of relations
  - Cannot compute precision and recall

Deep Learning for Relation Extraction

- Like NER, deep learning enables relation classification without handcrafted features
- Architectures include **RNN-based** (LSTM, bi-LSTM) and **CNN-based** (CNN,PCNN)
- Typically use **word embeddings**
- Also use **positional embeddings**: relative distance of each word from the entities in the sentence
  - Assumption: words closer to the target entities usually contain more useful information regarding the relation class

Multi-instance Learning

Exploit the large amount of training data created by distant supervision while being robust to the noise in the labels

- Method:
  - For every entity pair, defines a **bag consisting of all sentences that contain a mention of the entity pair**
  - Label is given to each bag of the relation entity rather than each sentence
  - **Assumption: at least one sentence that mentions two entities will express their relation**
  - Select the most likely sentence for each entity pair in training and prediction
- Drawback: the method loses a large amount of rich information contained in neglected sentences.

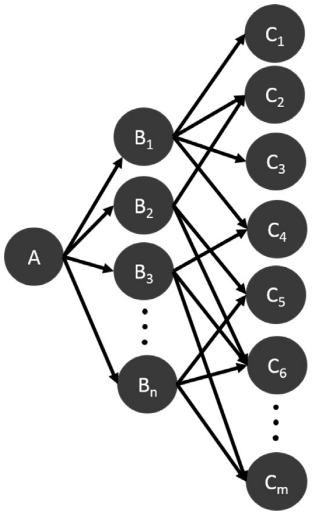
Tweaks for Improvement

Recent attempts to handle the noise from distant supervision use mechanisms like

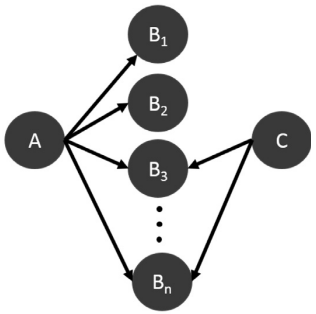
- **selective attention** over instances
- **max pooling**
- exploit **interaction between relations**
  - E.g., relations like *Father of* and *Mother of* can be exploited to extract instance for *Spouse of*

- These tweaks only work on the training and inference parts of the model
- ANN architecture used to encode the sentences remains the same

Literature Based Discovery (LBD)



Open Discovery



Closed Discovery

Explicit knowledge is found in text to generate "A implies B" and "B implies C" relationships

Two main ways to perform LBD

- **Open discovery**
  - user inputs a start term, system outputs a list of target terms
  - used to generate new discoveries
- **Closed discovery**
  - user inputs both a start term and a target term, system outputs a set of linking terms
  - used to explain correlations or observations

Swanson's Pioneering Work

Based on the literature published until 1985, Swanson postulated that there is a **connection between Fish Oils (FO) and Raynaud's Disease (RD)**

- Proposed **blood viscosity** as the concept that connects these two terms
  - Documents on FO consistently referred to its effect on blood viscosity
  - Documents related to RD noted a correlation between blood viscosity and RD
- Later clinically corroborated

- Known as **A-B-C model**
  - identifies plausible *B* terms that connect the *A* with the *C* term

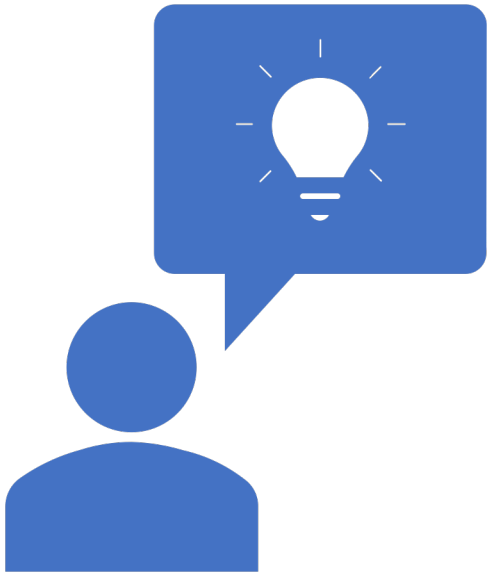
So, what do you need to perform text mining on biomedical documents?

Checklist

Access to

- Basic NLP software for linguistic processing
- Trainable NER and Relation Extraction Software
- Traditional and Deep learning frameworks
- Domain-appropriate lexicons, dictionaries, ontologies, etc.
- Large bodies of biomedical publications
- Sophisticated annotation editor

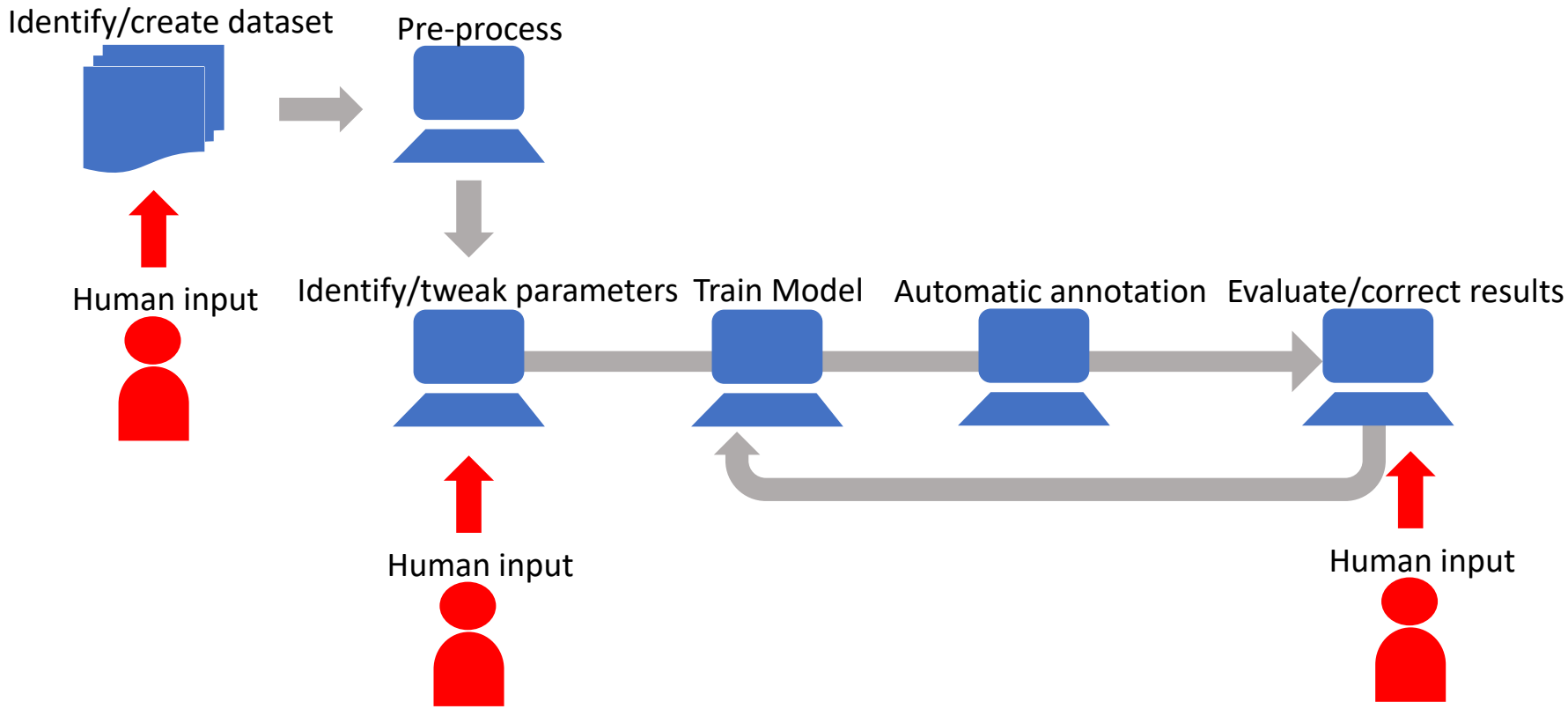
PLUS a good amount of knowledge about how to appropriately acquire, apply, evaluate, and improve these tools and resources!



And also…

The Human-in-the-Loop

Domain adaptation process



Domain Adaptation

- Critical capability for biomedical text mining
- Existing gold standard corpora and lexicons, ontologies cover entities like genes, proteins, diseases, etc.
- Researchers generally interested in entities relevant to specific subject areas
- Must augment existing resources or create new ones for text mining geared to specific domains

Existing Resources

- Gold standard corpora developed to support shared tasks/challenges
  - Informatics for Integrating Biology and the Bedside (i2b2)
  - BioNLP
  - BioCreative
  - CRAFT, GENIA, corpora developed by the community
- Frequently combine corpora with controlled vocabularies and ontologies
  - E.g., National Library of Medicine's Unified Medical Language System (UMLS) and Medical Subject Headings (MeSH)

Typical Scenario

- A scientist wants to apply text mining techniques to find articles including references to certain entities (e.g., proteins, genes) and their interactions
  - Knows nothing about NLP or Computer Science
  - Unfamiliar with NLP technologies
- Searches for NLP software that might help

Typical Scenario

- Finds existing tools and frameworks that are freely available

Not to mention several commercial (i.e., pricey) options

- Questions
  - Do these things all do the same thing, or do they differ in some way?
  - Do some work better than others?
  - Are some easier to use than others?
  - How does one choose?

Confused scientist

Problem

- Many existing tools, including some specifically created for BioNLP, are **difficult to install, configure, and use without some computational expertise**
- Even **more difficult to modify or adapt** without computational expertise and some knowledge of NLP
- Also: which tools performing the same task perform best and/or are best suited to a given task?

Another Sneaky Underlying Problem

- Input and output of tools from different sources differ dramatically!!!
- Often demands significant effort and expertise to adapt tools from different sources to work together

  ...if it is possible at all

- I.e., tools are not **interoperable**

Resource Interoperability

- The same interoperability problem exists for resources
  - Different physical formats
    - PDF, XML, plain text…
    - Extraction of text from PDF can be unreliable
  - Different representations for annotations
    - Different physical formats
      - XML, JSON, brackets, BIO
  - **Different semantic categories**

What is Needed?

A one-stop platform where scientists can readily access resources and tools and

- plug-and-play both tools and resources **interoperably**, i.e., without the need to convert formats etc.
- experiment with different tools, scenarios
- leverage support for **human-in-the-loop**

The Language Applications (LAPPS) Grid

Nancy Ide, Keith Suderman
Vassar College

James Pustejovsky, Marc Verhagen, Keigh Rim
Brandeis University

Christopher Cieri, Denise DiPersio, Jonathan Wright
Linguistic Data Consortium (Penn)

Eric Nyberg, Di Wang
Carnegie Mellon University

What is the LAPPS Grid?

Funded by US National Science Foundation and the Andrew K. Mellon Foundation

- Collaborative among Vassar College, Brandeis University, University of Pennsylvania, and Carnegie Mellon University
- Goal: Provide an infrastructure that facilitates
  - Retrieving large text collections from providers and repositories
  - Devising pipelines (workflows) of **interoperable** web services that automatically annotate data, provide evaluation metrics for the results, etc.
  - Saving, storing, and sharing pipelines and results for later use by yourself or others

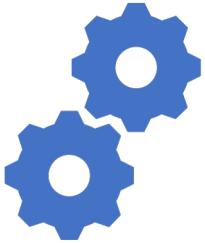
LAPPS/Galaxy Interface

Galaxy is an open, web-based platform designed primarily for computational genomics research

Accessible: Users without programming experience can easily specify parameters and run tools and workflows

Reproducible: Galaxy captures information so that any user can repeat and understand a complete computational analysis

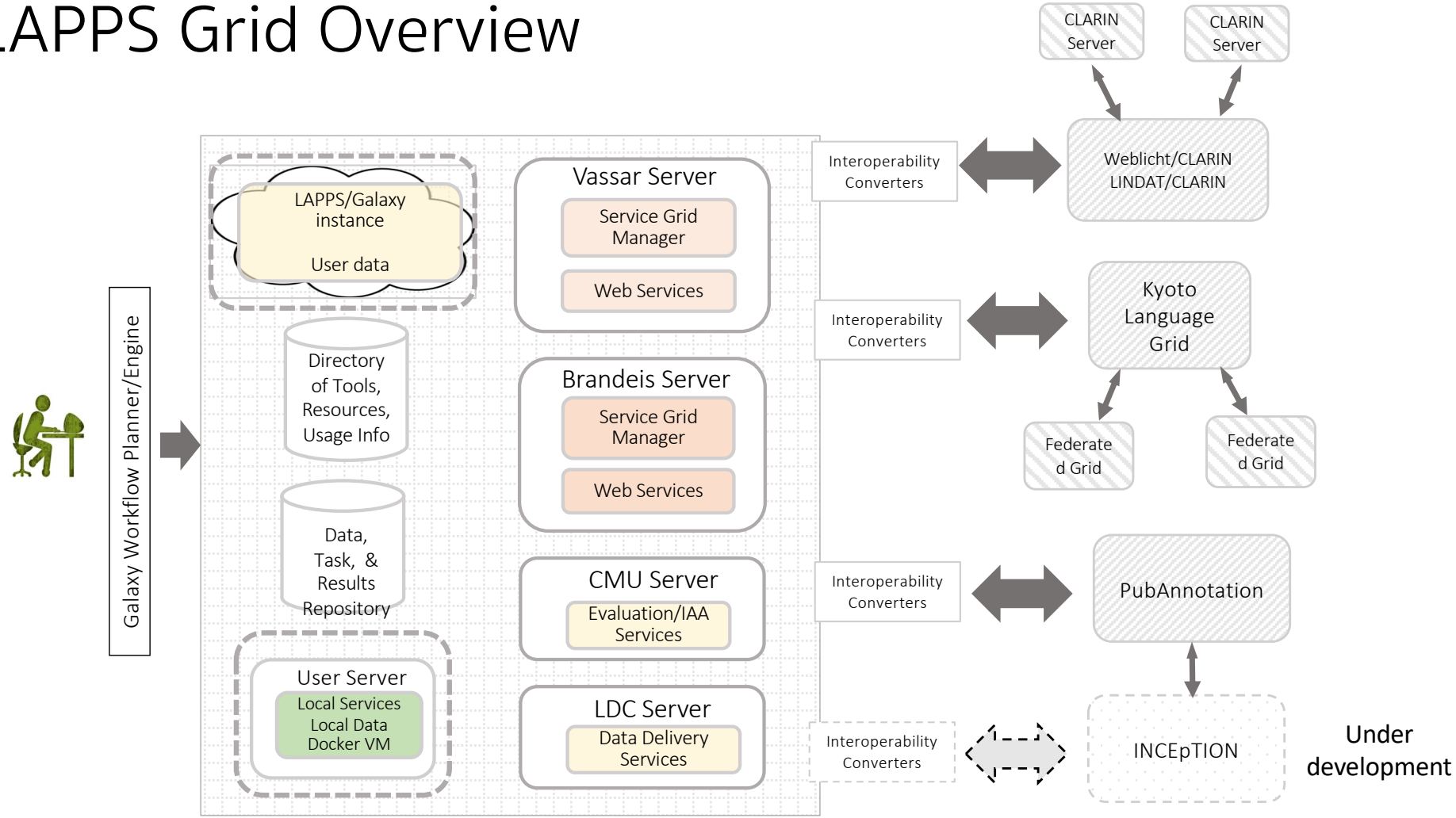
Transparent: Users share and publish analyses via the web and create interactive, web-based documents that describe a complete analysis



The LAPPS Grid uses the GALAXY framework as a vehicle to combine services of the Language Application Grid

Text processing pipelines, components wrapped as services, visualization of component output, evaluation of alternate pipelines, saving and sharing workflows, etc.

LAPPS Grid Overview



Galaxy

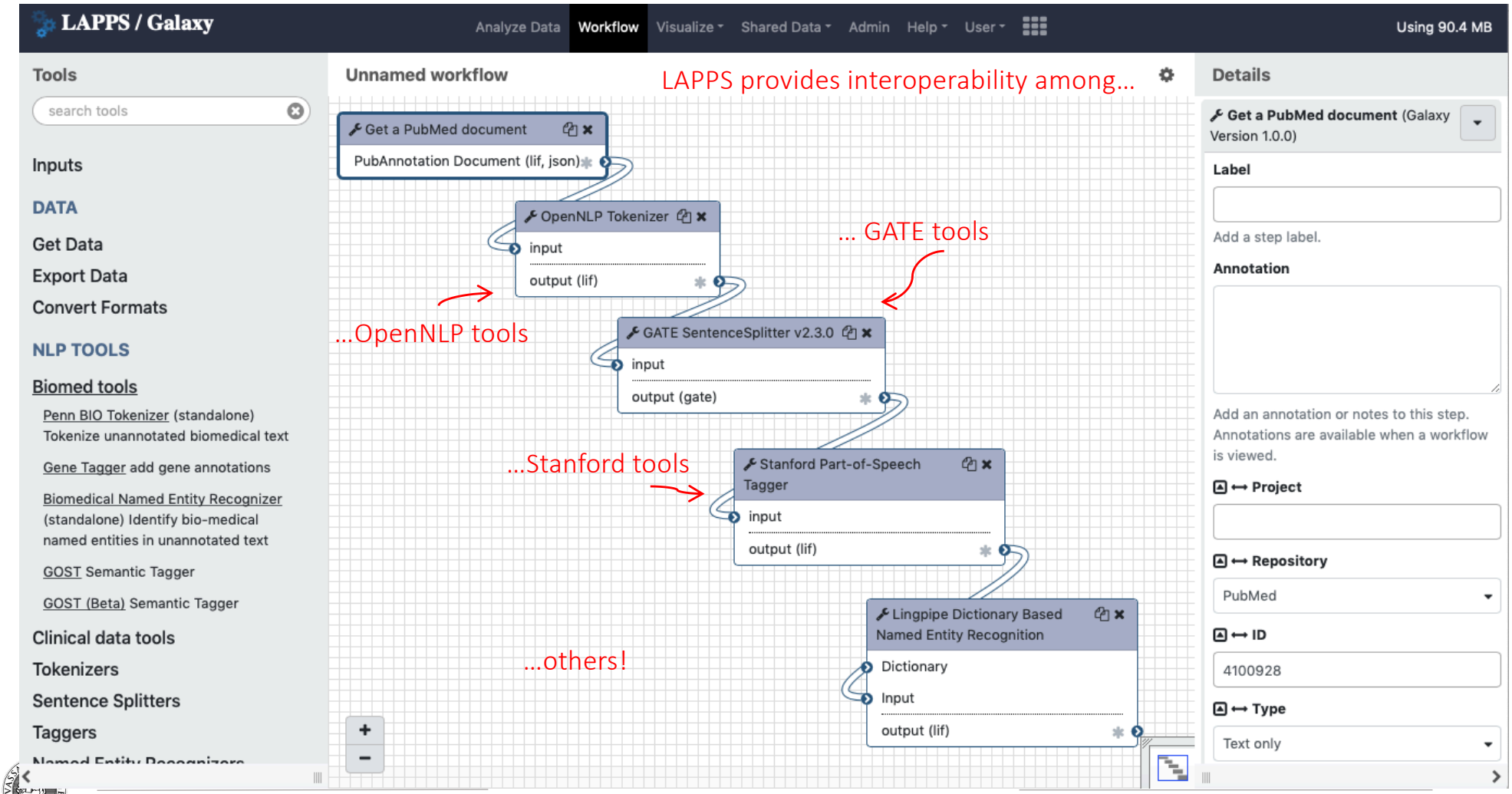
http://galaxyproject.org

LAPPS/GALAXY

Multiple options for running a LAPPS/Galaxy instance:

1. Use the **LAPPS/Galaxy web interface**
   - http://galaxy.lappsgrid.org
2. Create a **local Galaxy instance** including:
   - All of Galaxy, or
   - The Galaxy "NLP Flavor" with only LAPPS tools
3. Create a **docker image that is a self-contained vm** running LAPPS/Galaxy
   - Useful when privacy required, no network connection available, etc.
4. Create a **Galaxy instance in the cloud**
   - Useful for large datasets, computationally intense applications
   - https://jetstream.lappsgrid.org

Workflow construction



How Does the LAPPS Grid Enable Interoperability?

- **LAPPS Interchange Format (LIF)**
  - Format that allows web services to exchange detailed information about data and its annotations
  - "Pivot" into and out of which other formats are converted
  - **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**

BioNLP-oriented Tools in the LAPPS Grid

- Penn BioTokenizer
- Biomedical NER — Annotates proteins, DNA, RNA, cellLines, cellTYpes
- Gene annotator
- CDC/FDC CTakes
- GOST Semantic Tagger

Other LAPPS Grid Tools Useful for BioNLP

- TimeML Events
- LingPipe Dictionary-based NER
- Several different NER modules, tokenizers, parsers, chunkers, etc.
- HeidelTime
- Evaluation tools (Open Advancement)

Gold Standard Biomedical Data in the LAPPS Grid

BIONLP 2016 Reference Corpus

- 14 full paper PubMed articles about NFκB proteins

  Annotations for token+pos, dependency parse, event annotations, named entity annotations for proteins
- Annotates relations between events and proteins (themeOf, causeOf, locationOf, equivalentTo), and modification (negation, speculation),

BIONLP 2016 Protein Corpus

- Annotations for token+pos, dependency parse, proteins

BIONLP 2016 Coreference Corpus

- Annotations for anaphors bound by protein or event references, produced semi-automatically.
- Includes tokens+pos, dependency parses, coreference, relation (boundBy)

Access to Biomedical data from the LAPPS Grid

PubAnnotation

- Currently, all PubMed abstracts and PMC texts with annotations created and curated by users

PubMed

- All PubMed abstracts and PMC texts, solr indexed for search; automatically annotated versions (token, sentence, pos); word embeddings for all data

PubDictionaries

- Biomedical dictionaries etc. created and curated by users

The Language Applications Grid

Ask Me (almost) Anything

version

I am eager to help

Title

Enable	Algorithm	Weight
☑	consecutive terms	1.0
☑	total search terms	1.0
☑	position	1.0
☑	% search terms	1.0
☑	term order	1.0
☑	1st sentence	1.0
☑	sentence count	1.0

Select All | Clear All

Weight 0.9

Abstract

Enable	Algorithm	Weight
☑	consecutive terms	1.0
☑	total search terms	1.0
☑	position	1.0
☑	% search terms	1.0
☑	term order	1.0
☑	1st sentence	1.0
☑	sentence count	1.0

Select All | Clear All

Weight 1.1

LAPPS Grid Q/A to create a custom corpus

The Question

Question	What kinases phosphorylate AKT1 on threonine 308?
Query	body:kinases AND body:phosphorylate AND body:akt1 AND body:threonine AND body:308
Size	102
Time	0:00:05.542

Send Results To Galaxy

To send data to LAPPS/Galaxy you must be a registered user. Enter you Galaxy username (email address) below and the files will be available in th
(click the *Choose FTP files* button). If files with the same name already exists on the Galaxy server they will be overwritten.

Enter your Galaxy username

Send to Galaxy

Rate These Answers

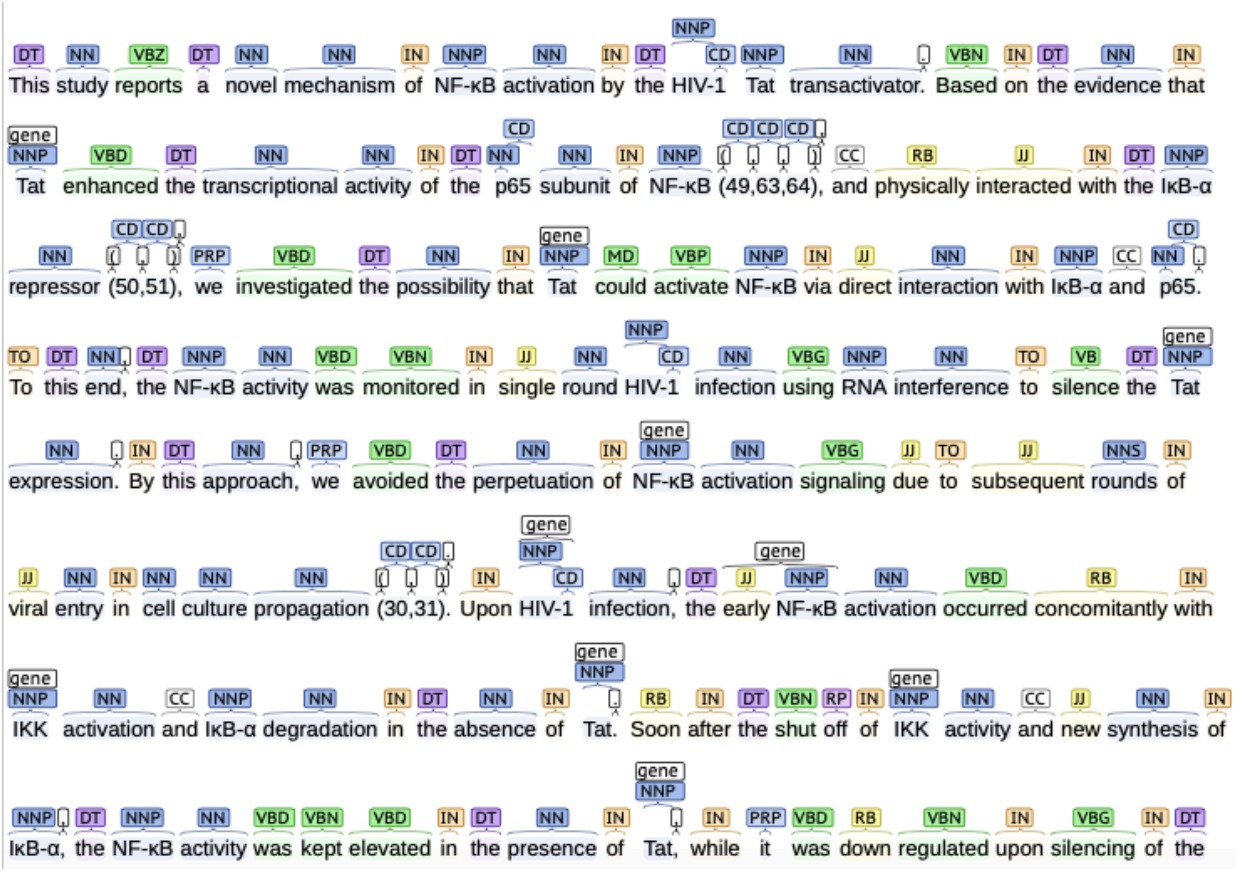
Good Meh Bad

Results

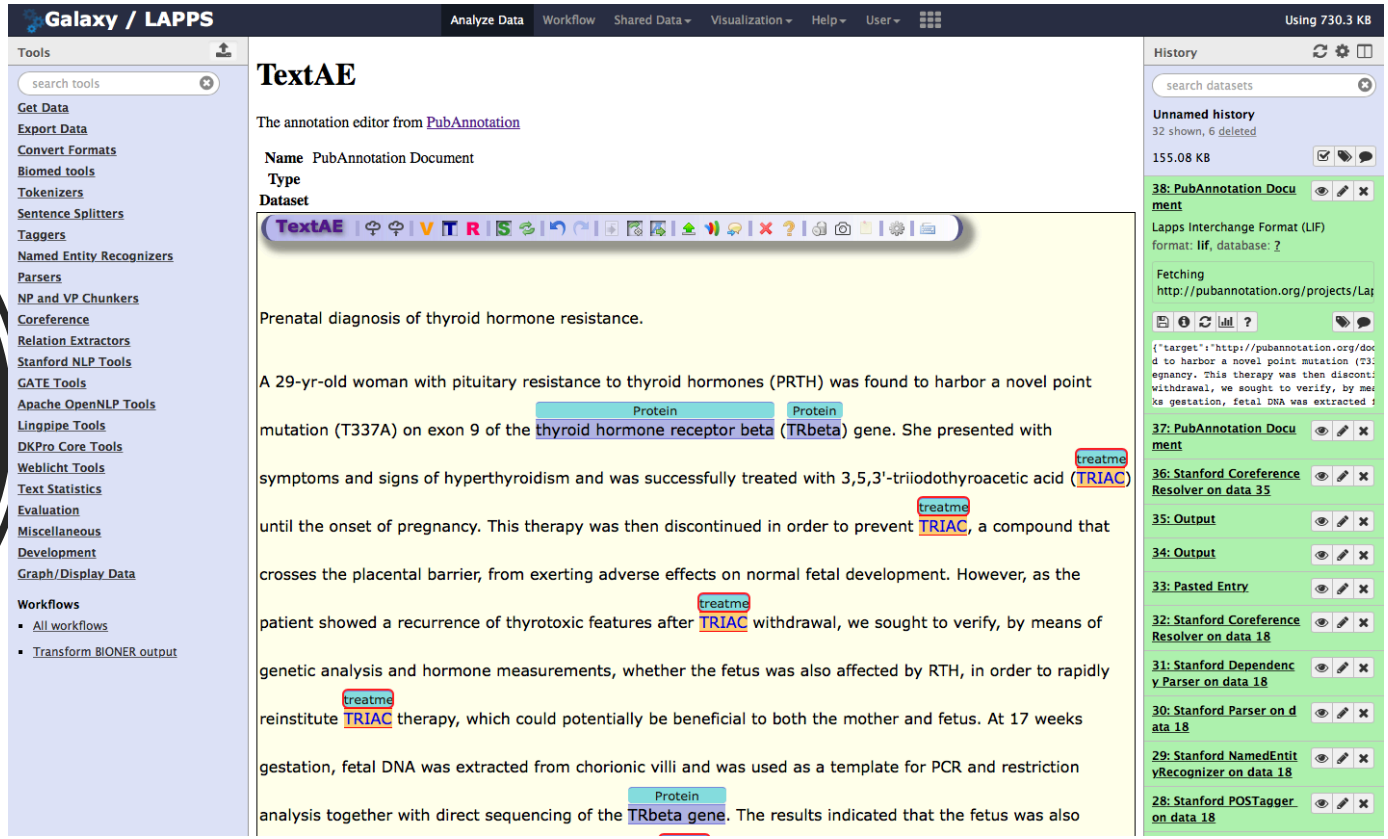
The Answers

Index	Score	PMID	Year	Title	ngrams	pterms	position	freq	order	1stSent	sents	title	ngrams
0	6.778	5342720	2016	SMYD3-mediated lysine methylation in the PH domain is critical for activation of	0.000	0.200	0.000	0.077	1.000	0.200	1.000	2.477	0.012

Visualization in the LAPPS Grid



TextAE Visualization and Editing in the LAPPS Grid



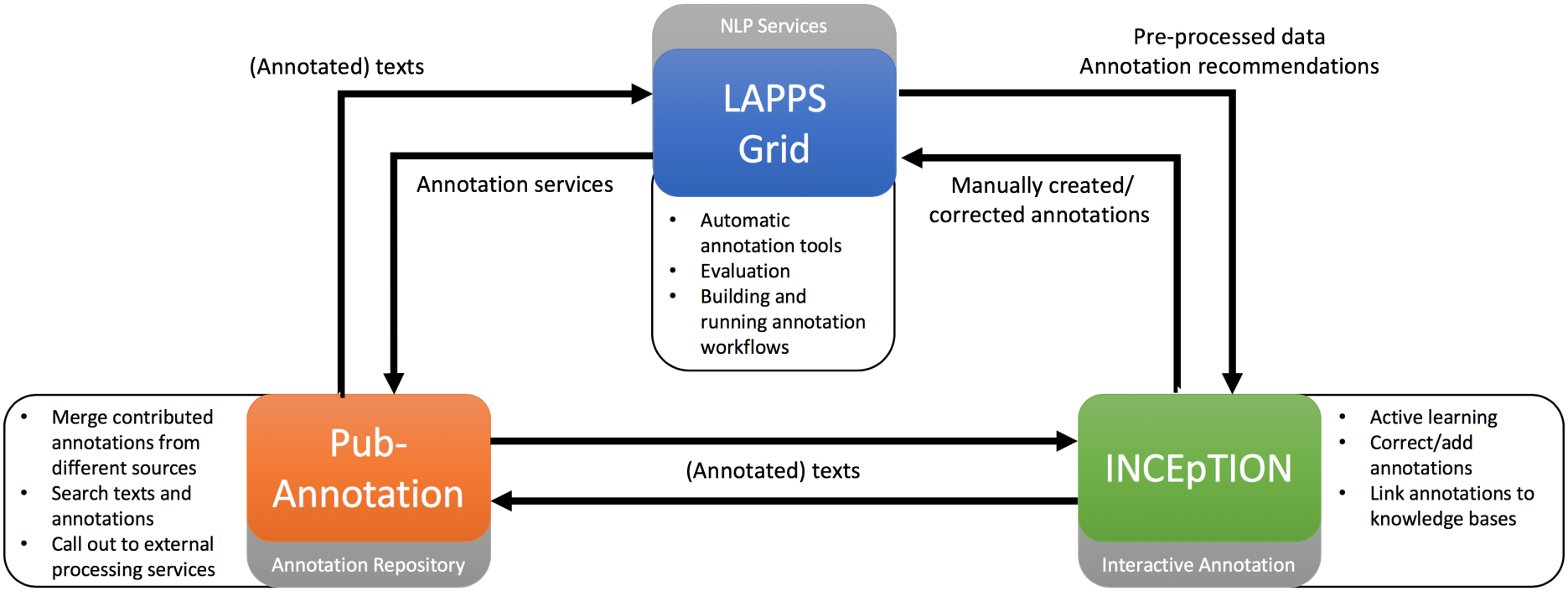
Current Activities

NSF EAGER grant (Vassar, Brandeis, Tufts, Penn State) to develop and implement methods for domain adaptation to accommodate specific areas of scientific text mining research

Collaboration with PubAnnotation and INCEpTION to fully integrate the three platforms to enable iterative development of language models via "on the fly" machine learning

Nascent collaboration with University of Wisconsin's "Geo Deep-Dive" project, access to millions of scientific publications (many copyrighted) using their extensive HPC facilities

Interaction among PubAnnotation, INCEpTION, and LAPPS Grid



Current Activities

NSF ABI grant

- Collaboration between Vassar College and Galaxy Principal Investigators to
  - Develop tools, ready-made workflows, etc. for mining biomedical publications
  - Provide seamless integration of text mining capabilities and the vast array of tools provided in Galaxy

Collaboration with the US government Centers for Disease Control and Food and Drug Administration to adapt the LAPPS Grid for summarization and mining of clinical reports

LAPPS Grid is a Work in Progress

- Recent shift to scientific text mining
- Establishing an increasing number of fruitful collaborations
- Seeking contributions of software, data, resources, ideas

Thank you!

Questions?