Accelerating Research through Large Scale, Interpretable Machine Reading

By Enrique Noriega-Atala

Computational
Language
Understanding

- Postdoctoral Researcher @ Dept of Computer Science
- PhD from School of Information ’20
- Specialize in Natural Language Processing

Presentation Outline

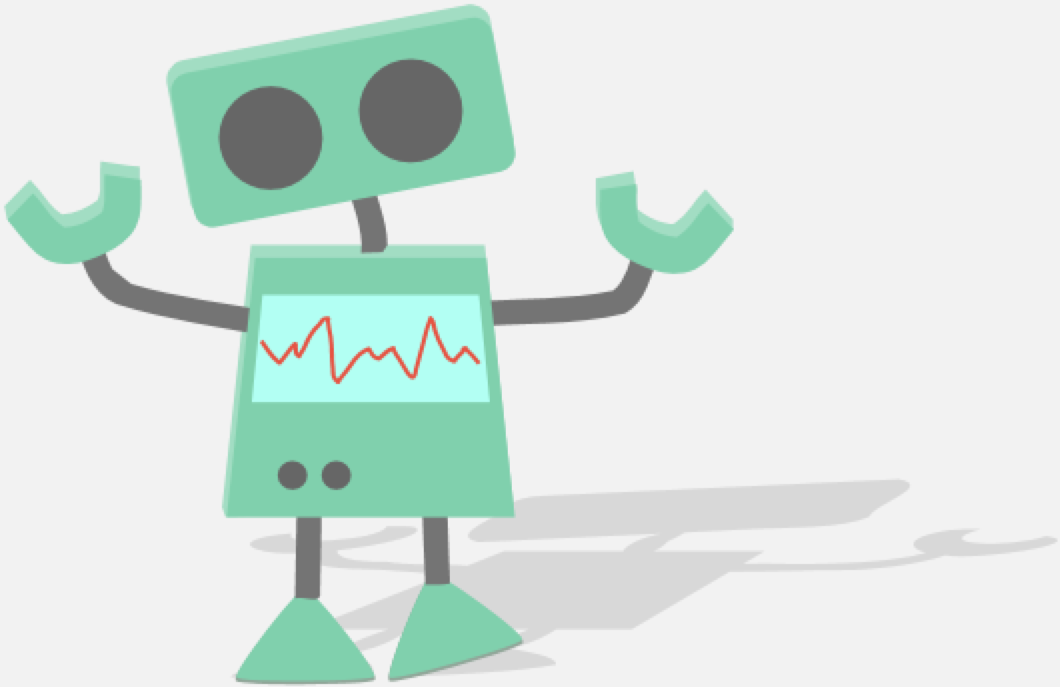
- Quick Overview of NLP
- Large Scale Information Extraction
- Learning to Generate Rules
- Visualization of Information Extraction

What is Natural Language Processing?

NLP, in part, is the application of computational techniques to understand texts written by people

Examples

- Sentiment classification
- Spam detection
- Machine translation
- Question answering



Enrique is presenting his research to the

Data Science community at the Kiva Room

Enrique is presenting his research to the Data Science community at the Kiva Room

Information Extraction

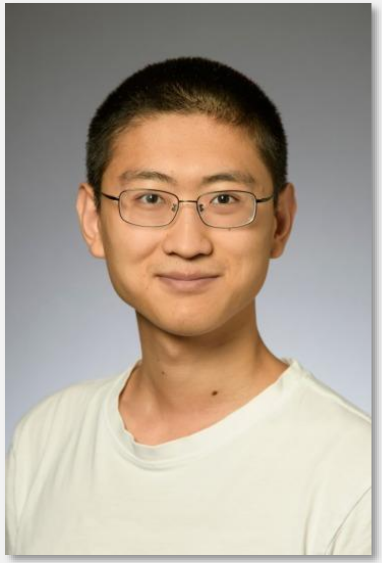
- Commonly trained using supervised machine learning

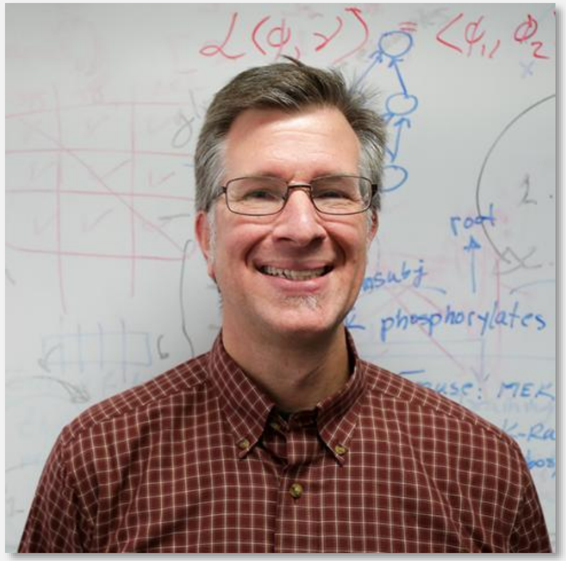
- Multi-step process:
  - Concept recognition
  - Normalization to database/ontology
  - Relation detection

Information Extraction

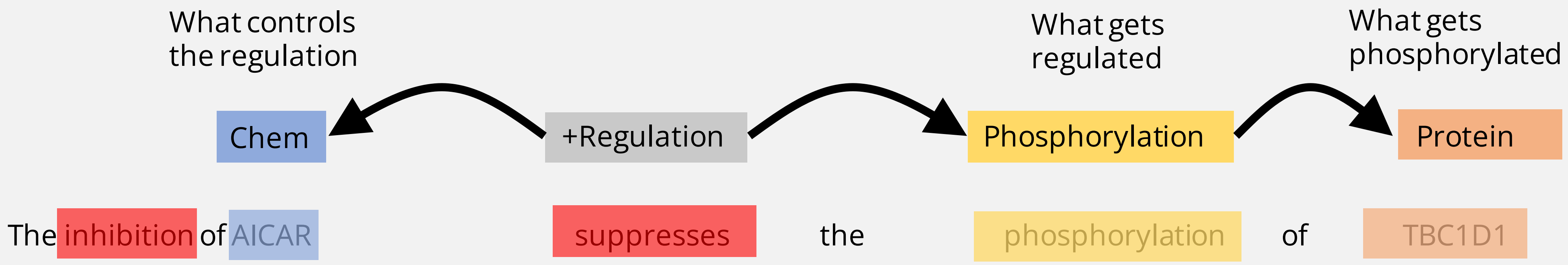
Large Scale Information Extraction




root
phosphorylates


IE in the Biomedical Domain



REACH: Biomedical Information Extraction System

Maintainability

Scalability

Maintainability

- Different requirements:
  - Cancer Biology
  - Hematology
  - Frailty Syndrome

- Similar, but not the same …

- Training data for each use case?

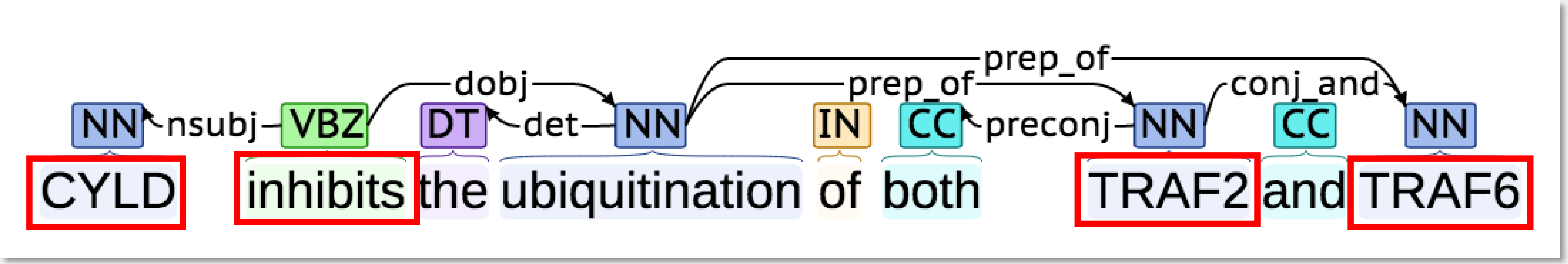
Statistical vs Rule-based

**Graphical representation of a domain expert*

Rule-based Information Extraction

```
- name: ${ eventName }_syntax_3_noun
  priority: ${ priority }
  example: "ERK- mediated serine ${ nominalTriggerLemma } of the GAB1 adaptor has been shown to ..."
  label: ${ label }
  action: ${ actionFlow }
  pattern: |
    trigger = [lemma=/${ nominalTriggerLemma }/ & ${triggerPrefix} & !outgoing=/${passive_agents}/] # nominal predicate
    cause:BioChemicalEntity = /${conjunctions}|${noun_modifiers}/{1,2} #or /${genitive_case_marker}/ /${passive_agents}/
    theme:BioChemicalEntity = /${genitive_case_marker}/ /${conjunctions}/? /${conjunctions}|${noun_modifiers}/{1,2} #{,2}
    site:Site? = /${any_preposition}|${conjunctions}|${noun_modifiers}/{1,2}
```

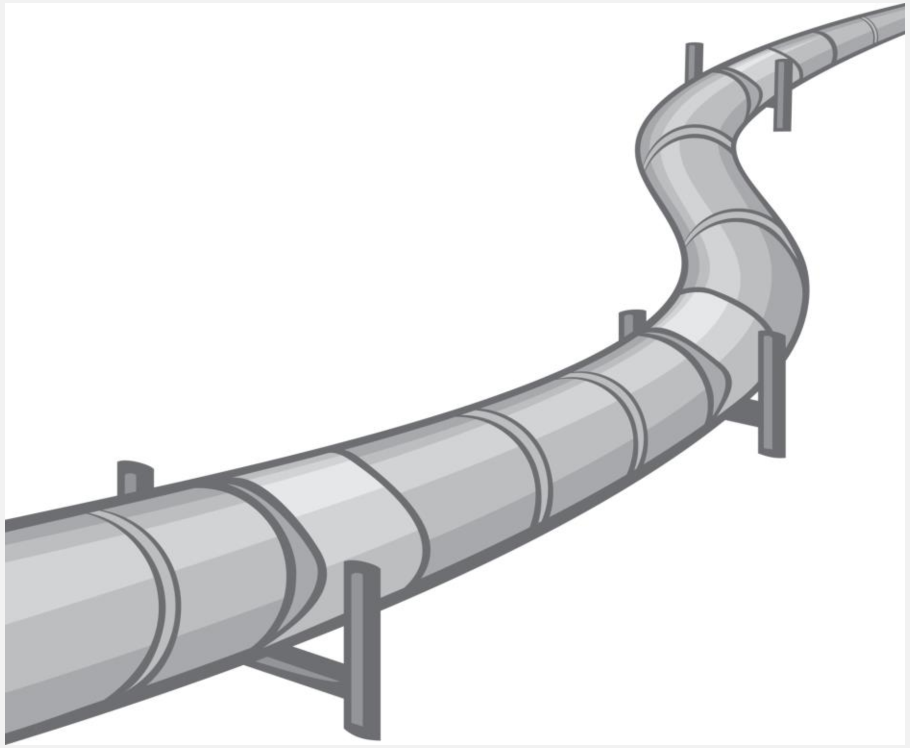
Rule-based Information Extraction



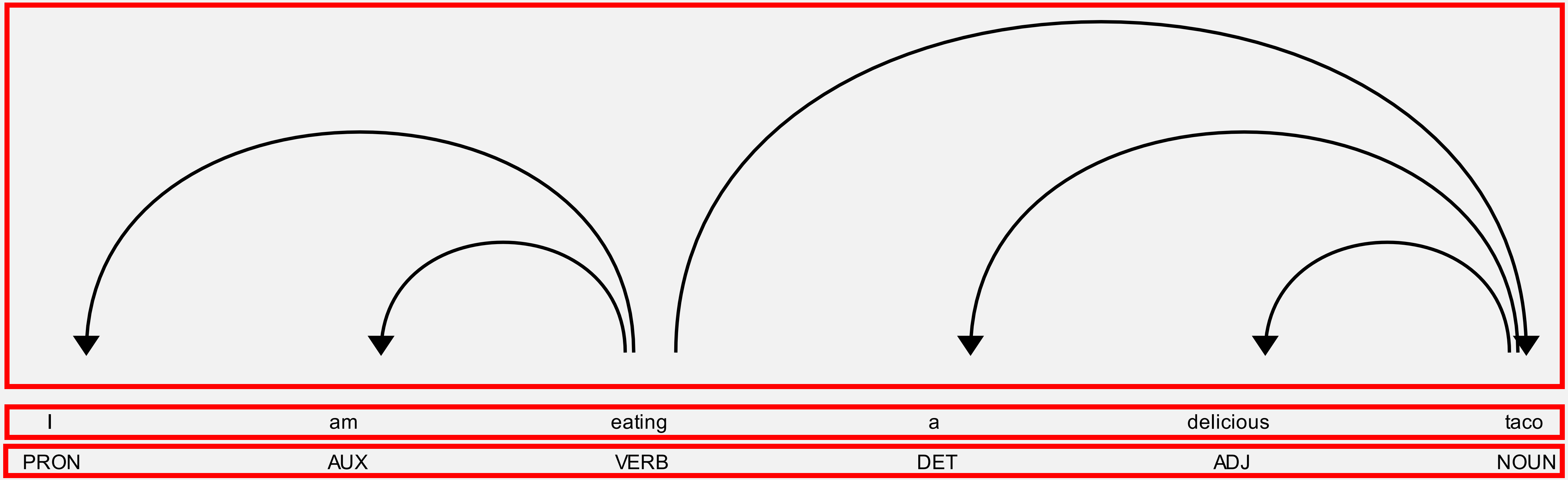
1. Match a Trigger: ***inhibits***
2. Find biochemical entity: *CYLD*
3. Find a second biochemical entity: *TRAF2* or *TRAF6*
4. Constructing the biochemical event(s)

NLP pipeline

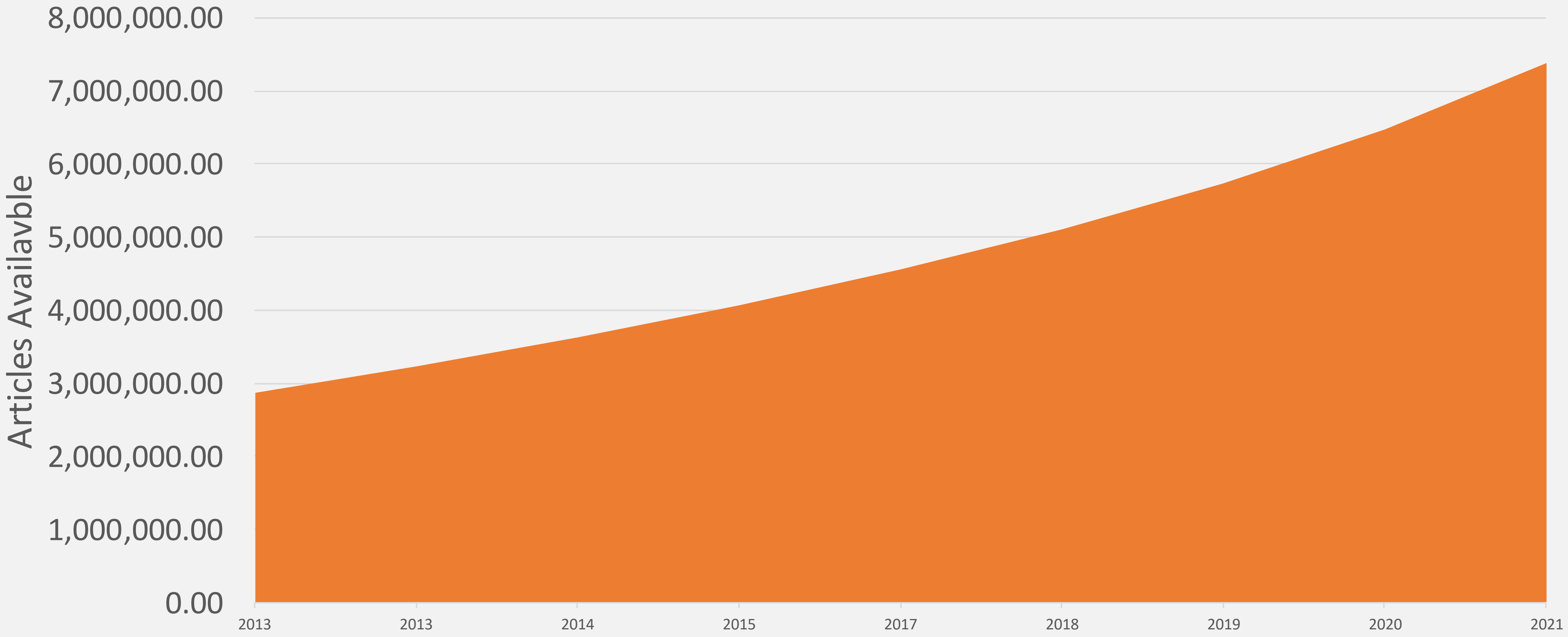
- Tokenization
- Part-of-speech tagging
- Syntactic parsing
- Named entity recognition



NLP pipeline

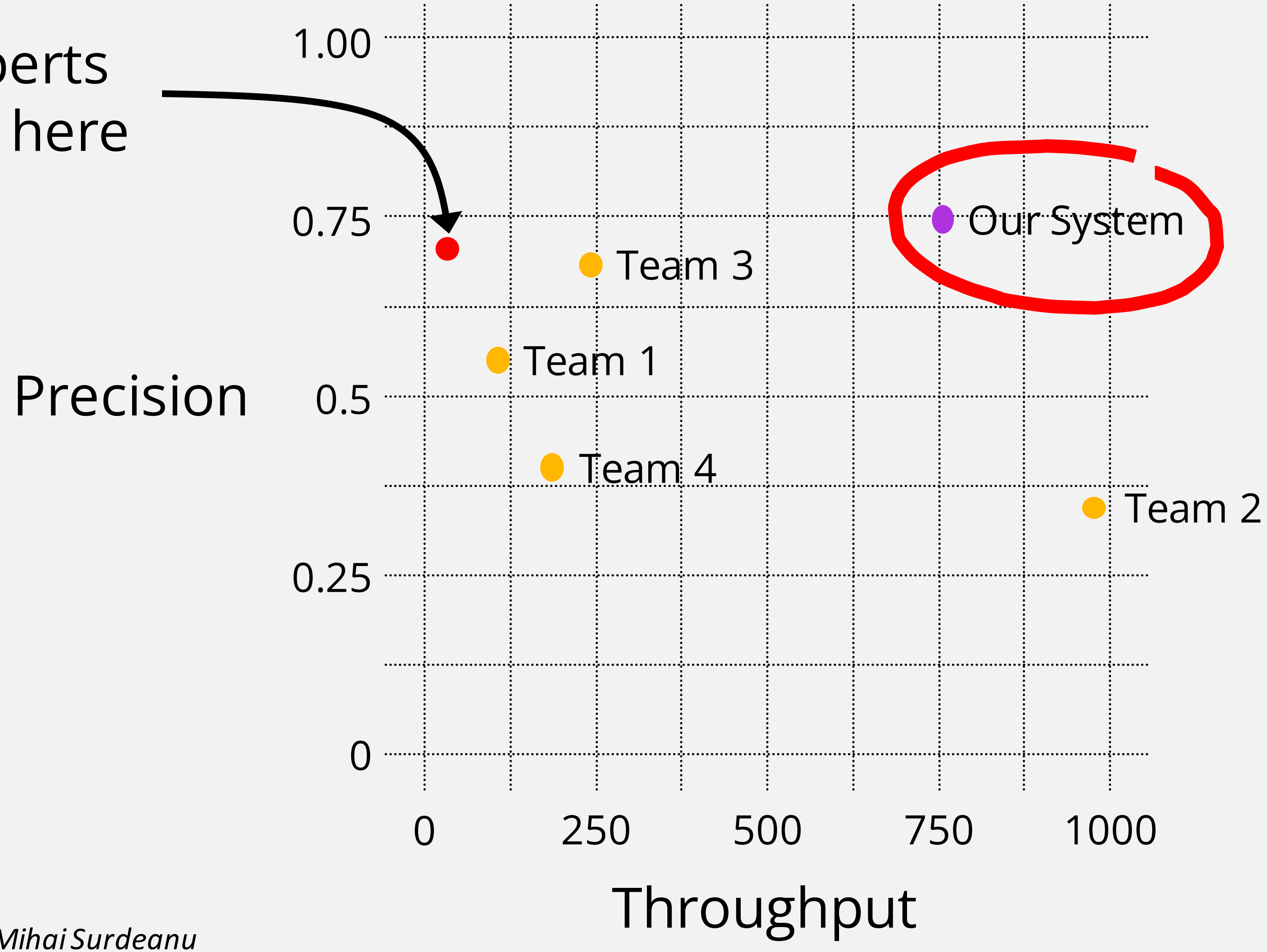


PubMed Size per Year



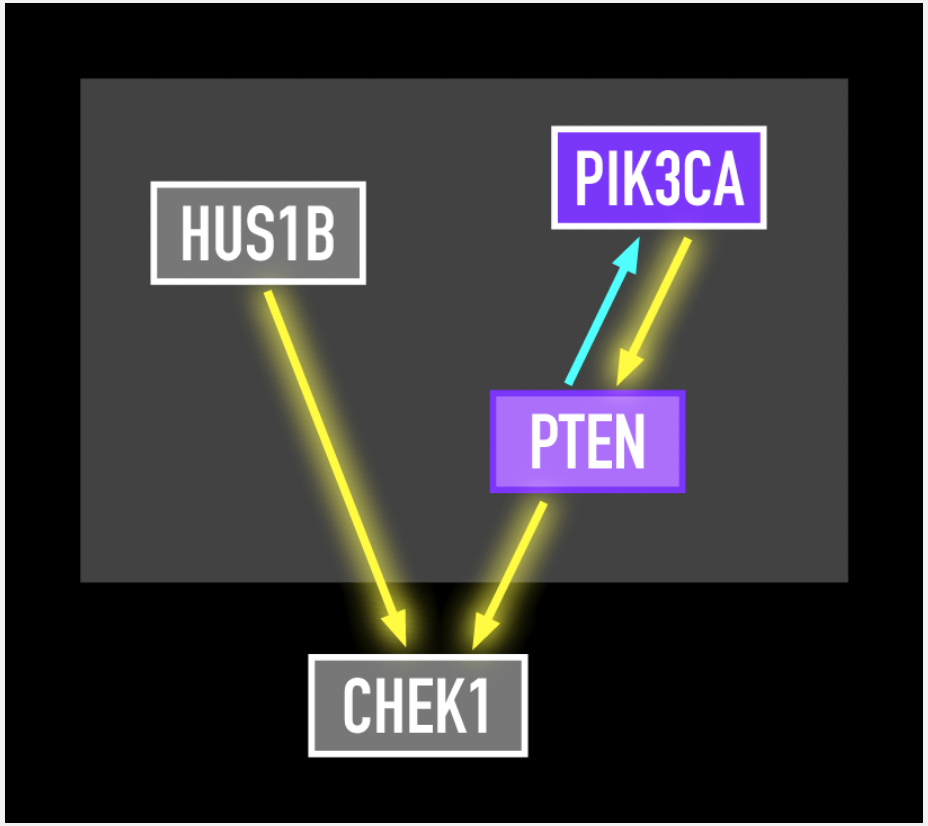
Machine reading performance

Human experts are around here



Slide courtesy of Dr. Mihai Surdeanu

Potential cancer-driving mechanisms discovered



Legend for gene alteration frequency:

0% 35%

Post-translational control (protein level)

New discovery

HUS1B → CHEK1: "Hus1 loss results in abnormal gammaH2AX localization and increased CHK1 phosphorylation."

PTEN → CHEK1: "PTEN also induces phosphorylation and monoubiquitination of DNA damage checkpoint kinase, Chk1"

PIK3CA → PTEN: "p110alpha inactivation can inhibit the impact of PTEN loss"

Slide courtesy of Dr. Mihai Surdeanu

WHAT RESEARCH IS LEFT?

- Contextualize extractions
- Automatically generating rules
- Visualizing and analyzing extractions

Contextualizing Information Extraction

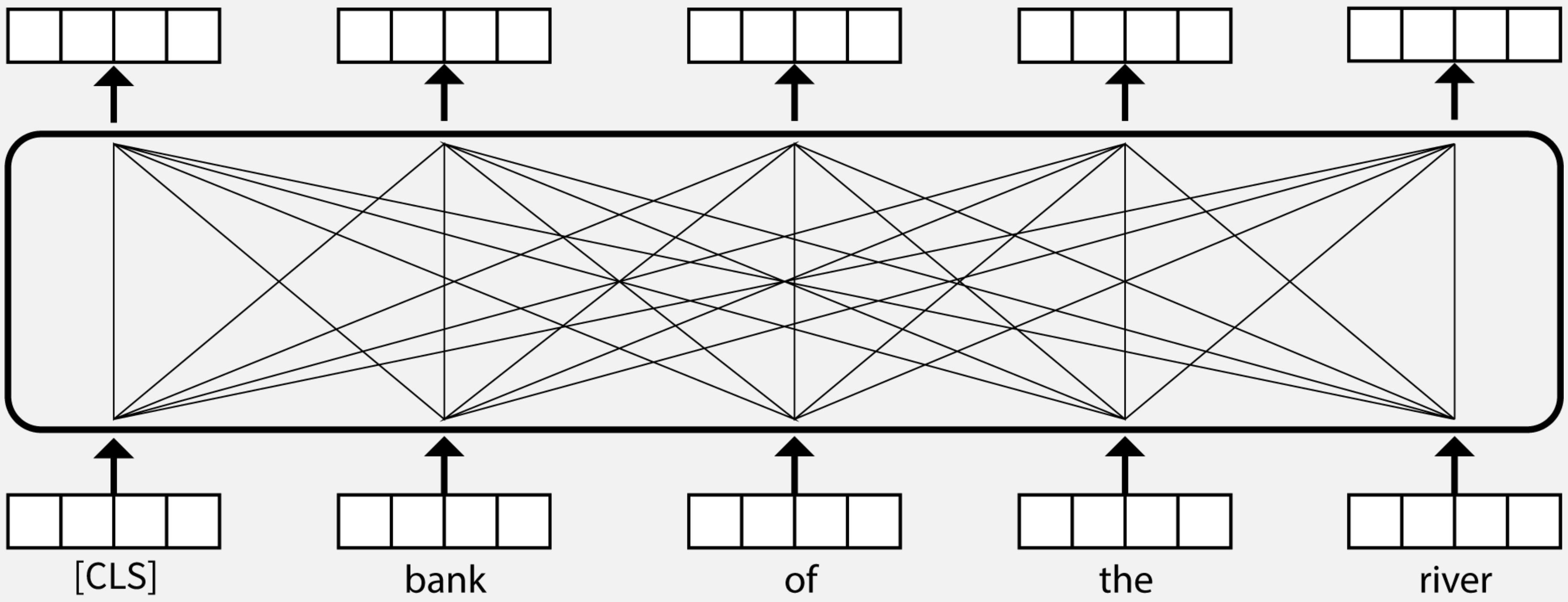
To date, the vast majority of experimental models are animal models, almost exclusively consisting of transgenic mice that express human genes that result in the formation of amyloid plaques (by expression of human *APP* alone or in combination with human *PSEN1*) and neurofibrillary tangles

Biological Context

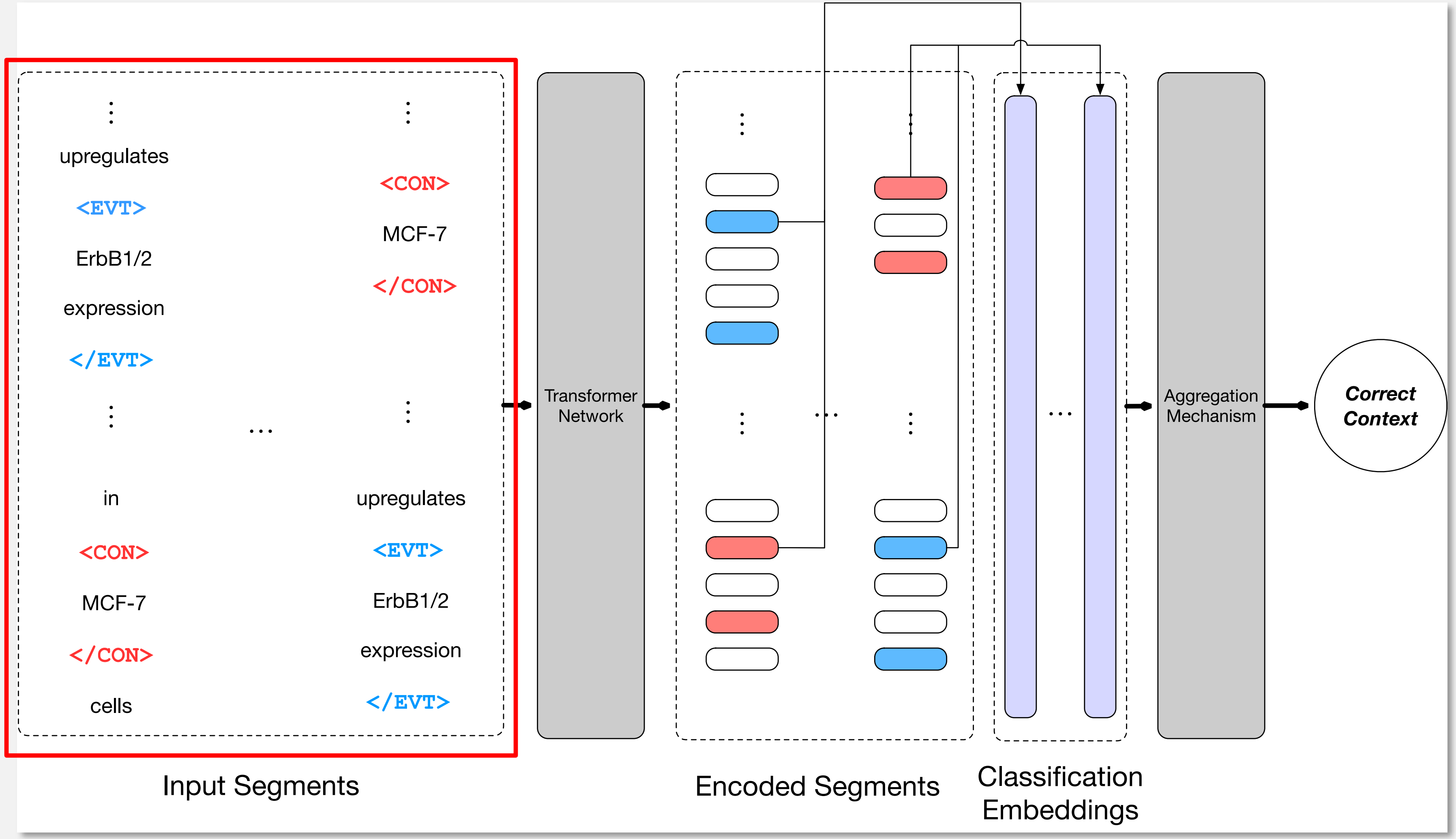
Chemical Event

Drummond E, Wisniewski T. Alzheimer's disease: experimental models and reality. Acta Neuropathol. 2017 Feb;133(2):155-175. doi: 10.1007/s00401-016-1662-x. Epub 2016 Dec 26. PMID: 28025715; PMCID: PMC5253109.

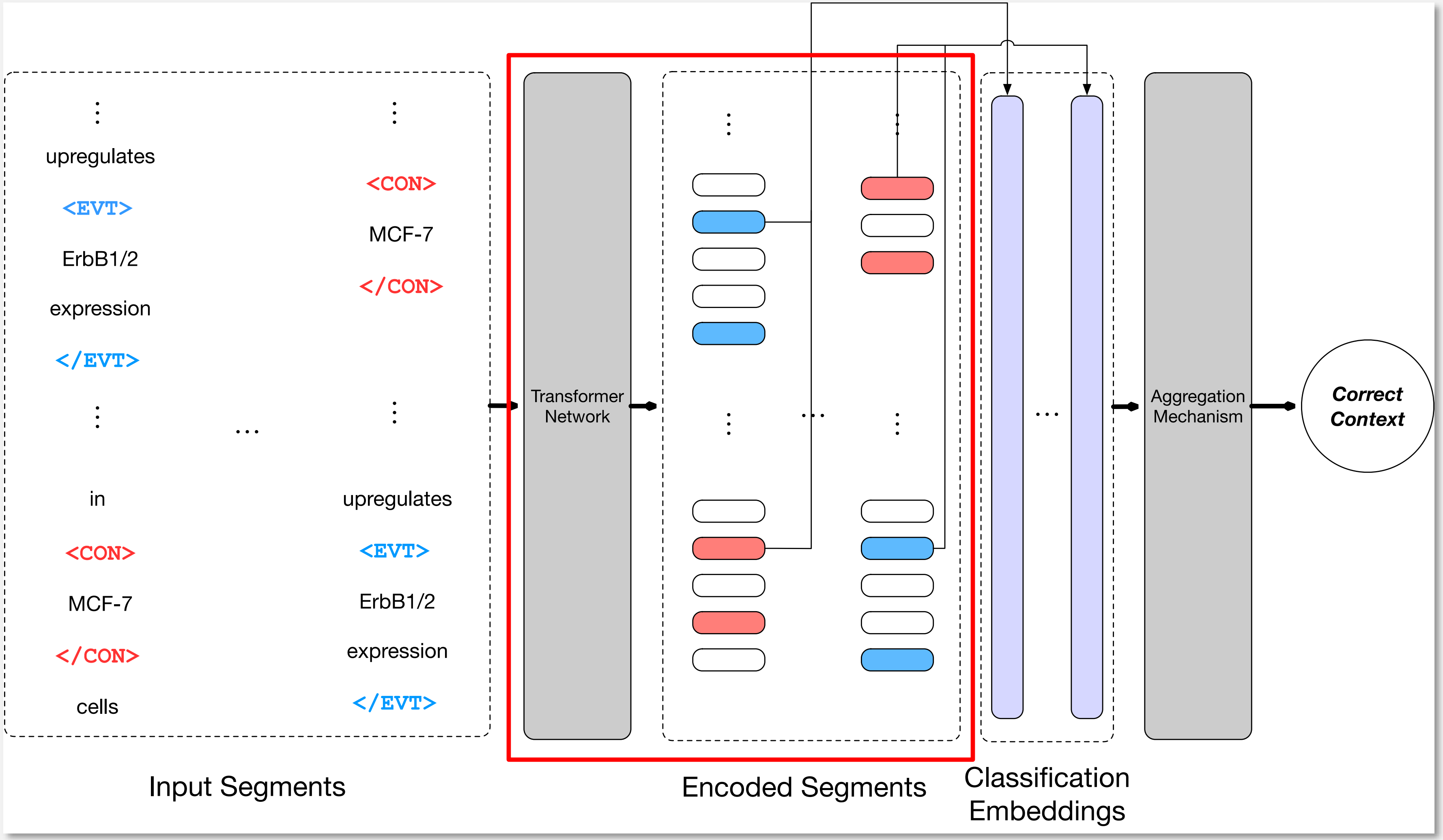
Transformer Neural Networks



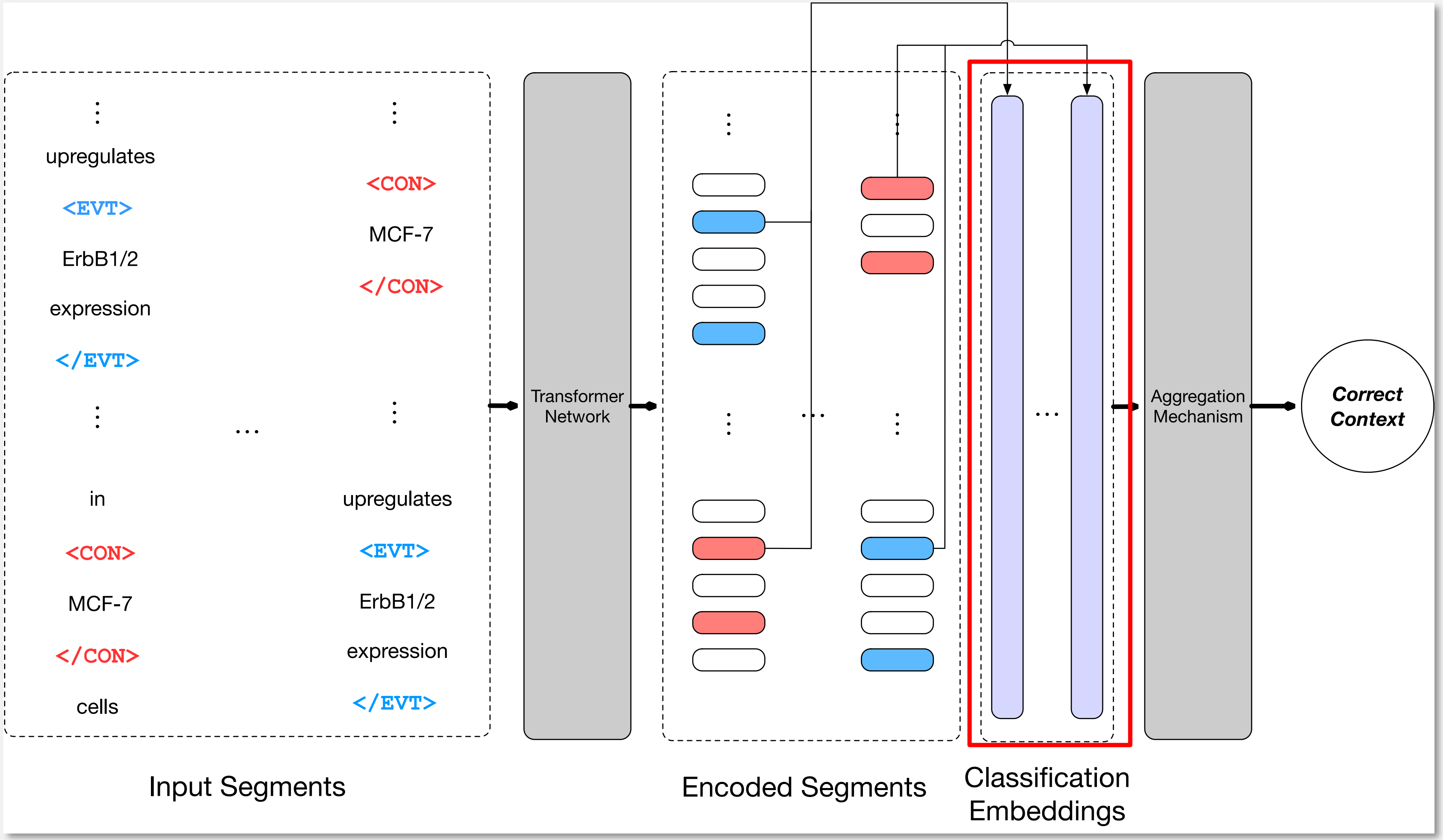
Context Detection Architecture



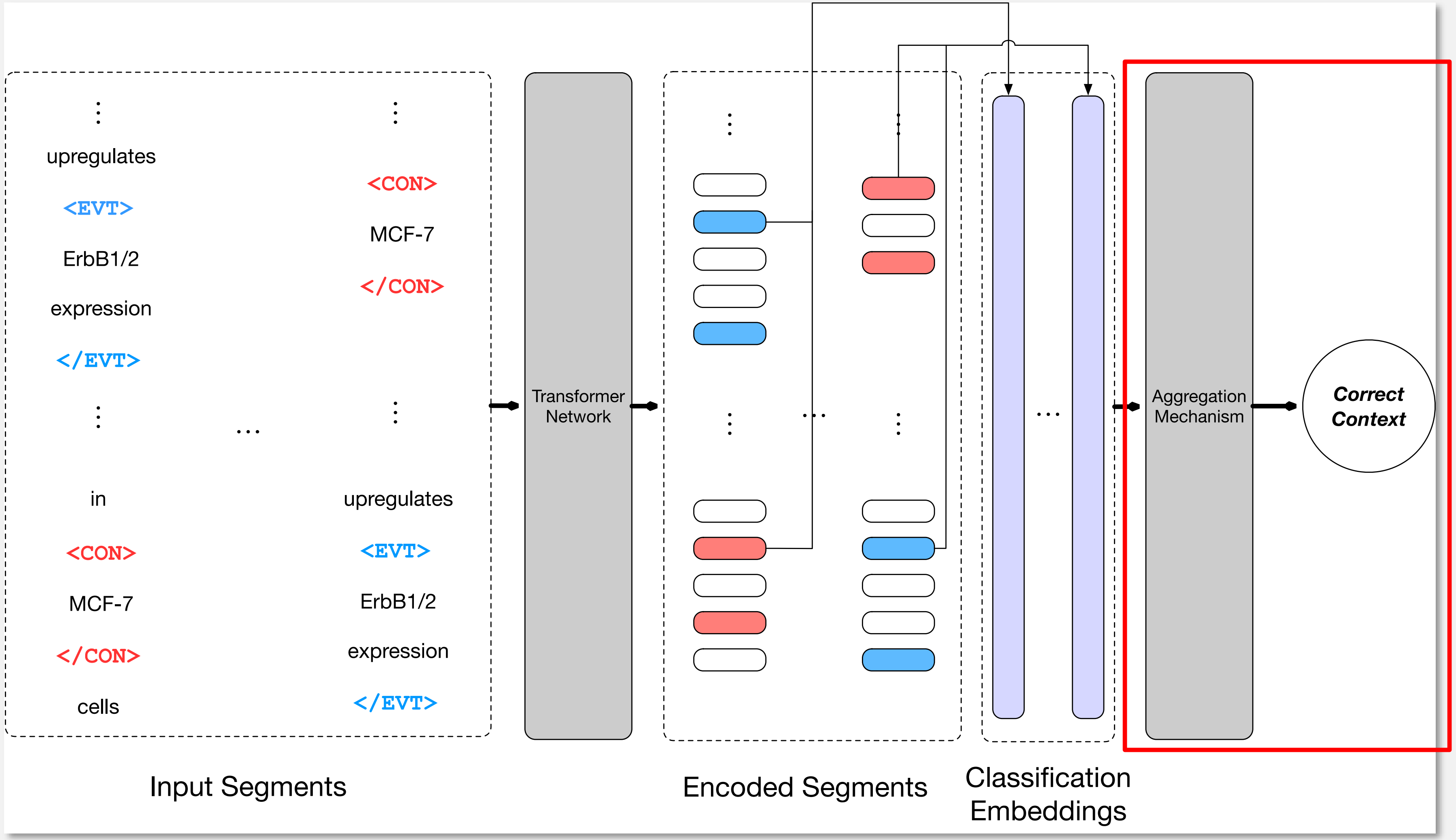
Context Detection Architecture



Context Detection Architecture



Context Detection Architecture



Context Classification Results

ENSEMBLE	PRECISION	RECALL	F1
Majority (3 votes)	**.58**	.50	**.54**
Parameterized aggregation	.54	.49	.51
One-hit	.41	.67	.50
Post inverse distance	.57	.45	.50
Nearest Mention	.54	.46	.49
BASELINES			
Random forest	.44	.54	.48
Logistic regression	.36	**.69**	.47
Heuristic	.42	.55	.47

Context Classification Results

ENSEMBLE	PRECISION	RECALL	F1
Majority (3 votes)	**.58**	.50	**.54**
Parameterized aggregation	.54	.49	.51
One-hit	.41	.67	.50
Post inverse distance	.57	.45	.50
Nearest Mention	.54	.46	.49
BASELINES			
Random forest	.44	.54	.48
Logistic regression	.36	**.69**	.47
Heuristic	.42	.55	.47

Context Classification Results

ENSEMBLE	PRECISION	RECALL	F1
Majority (3 votes)	**.58**	.50	**.54**
Parameterized aggregation	.54	.49	.51
One-hit	.41	.67	.50
Post inverse distance	.57	.45	.50
Nearest Mention	.54	.46	.49
BASELINES			
Random forest	.44	.54	.48
Logistic regression	.36	**.69**	.47
Heuristic	.42	.55	.47

Context Classification Results

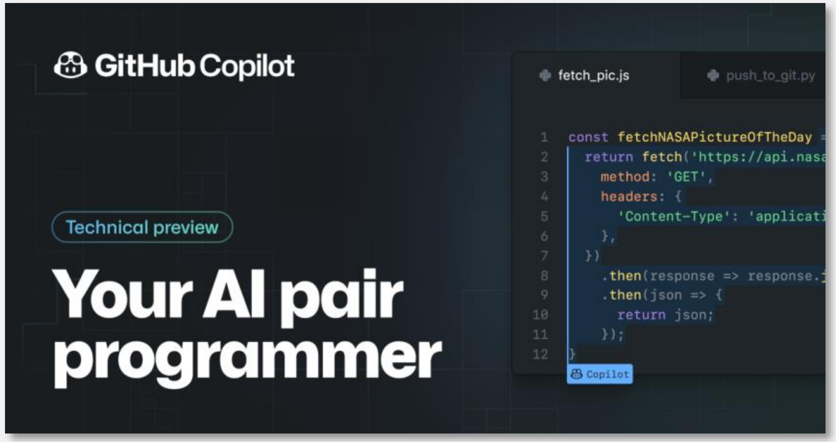
ENSEMBLE	PRECISION	RECALL	F1
Majority (3 votes)	**.58**	.50	**.54**
Parameterized aggregation	.54	.49	.51
One-hit	.41	.67	.50
Post inverse distance	.57	.45	.50
Nearest Mention	.54	.46	.49
BASELINES			
Random forest	.44	.54	.48
Logistic regression	.36	**.69**	.47
Heuristic	.42	.55	.47

Learning to Generate Rules

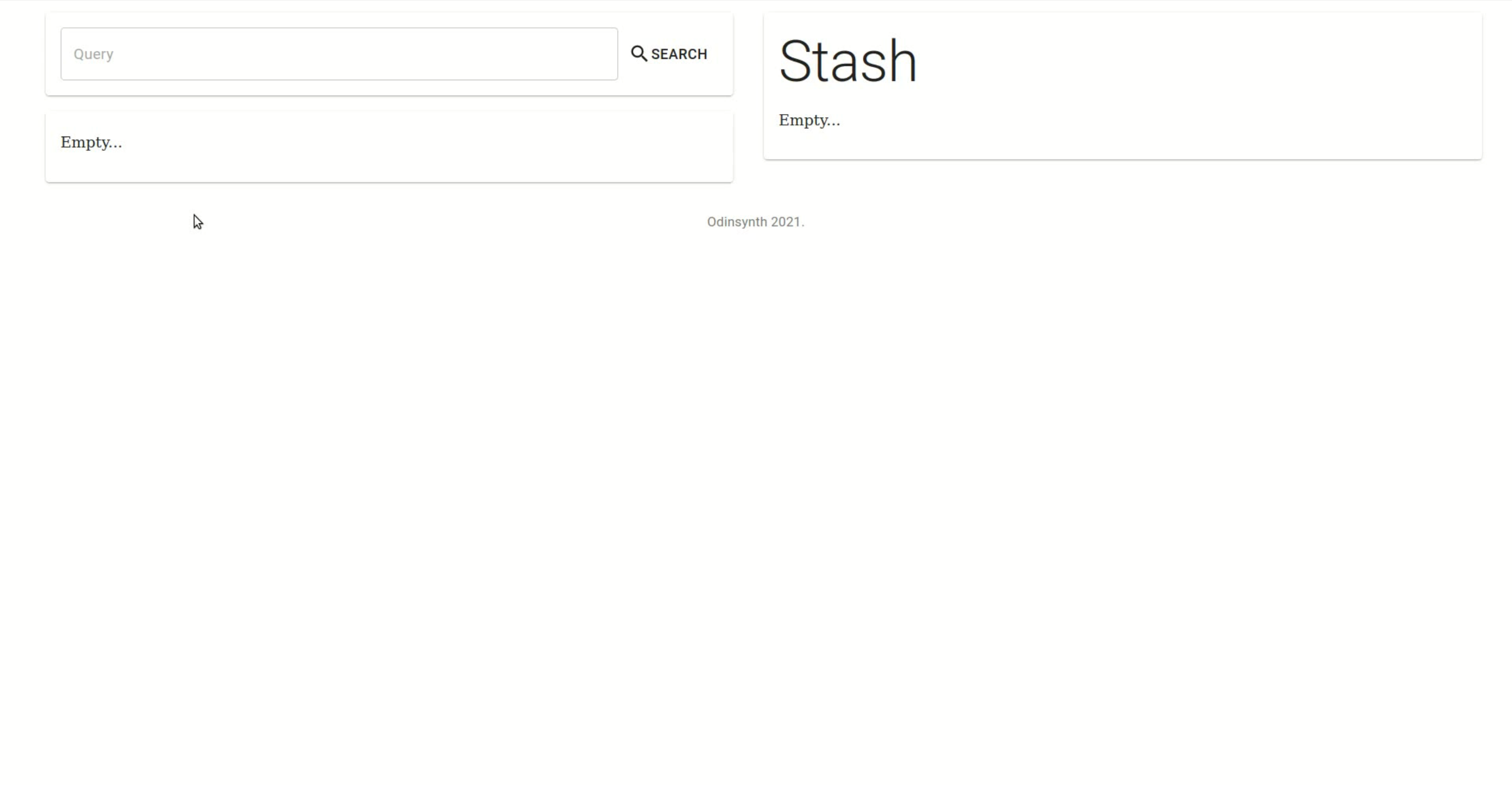



Rule Generation

Given a set of textual phrases with span annotations, generate a rule or set of rules that match the input



**but for information extraction rules*


Query
SEARCH
Empty...
Stash
Empty...
Odinsynth 2021.

Rule Generation

- Input:

She is already an honorary doctor of the <MATCH> Technical University of CLUJ-NAPOCA </MATCH>

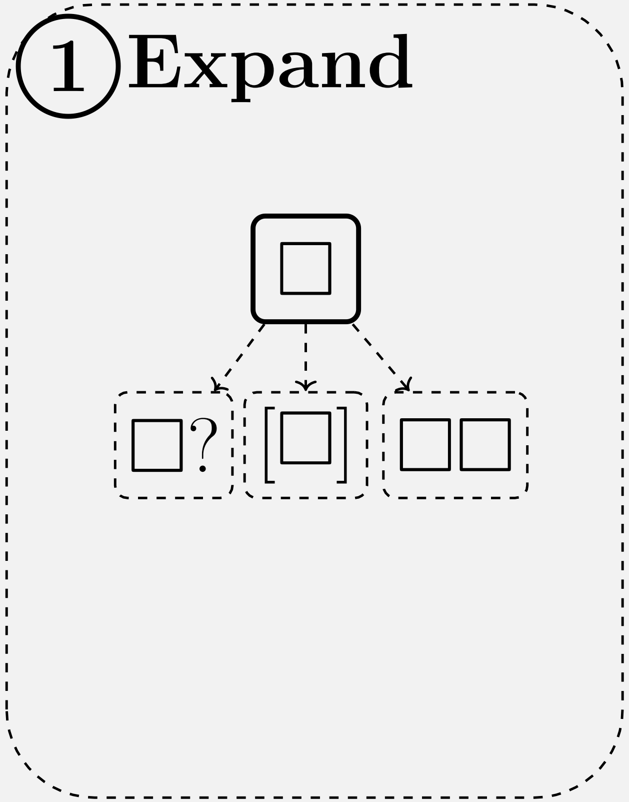
… using a modified version of the TUD (<MATCH> Technical University of Denmark </MATCH>) radar.

**User Specification*

- Output:

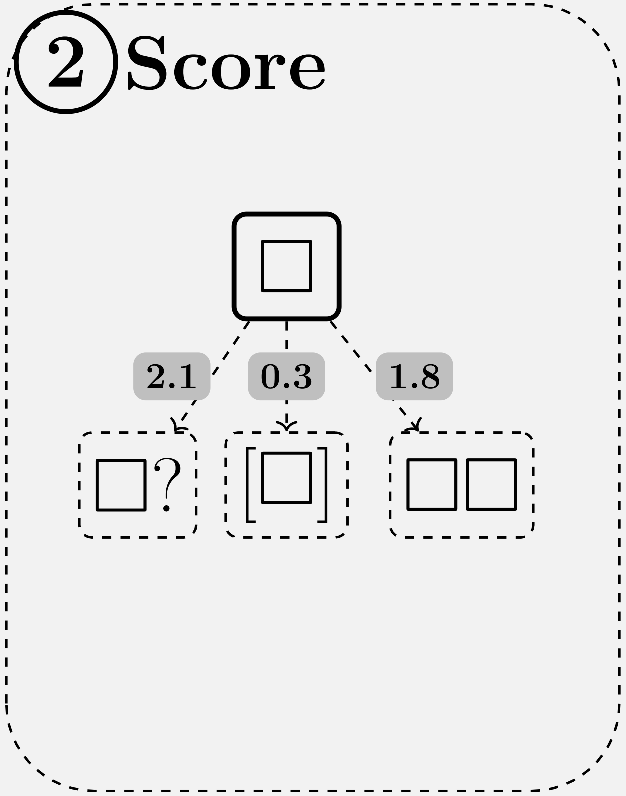
```
[word = Technical][tag = Noun] [word = of] [tag = Noun]
```

Algorithm: Expand, Score, Select



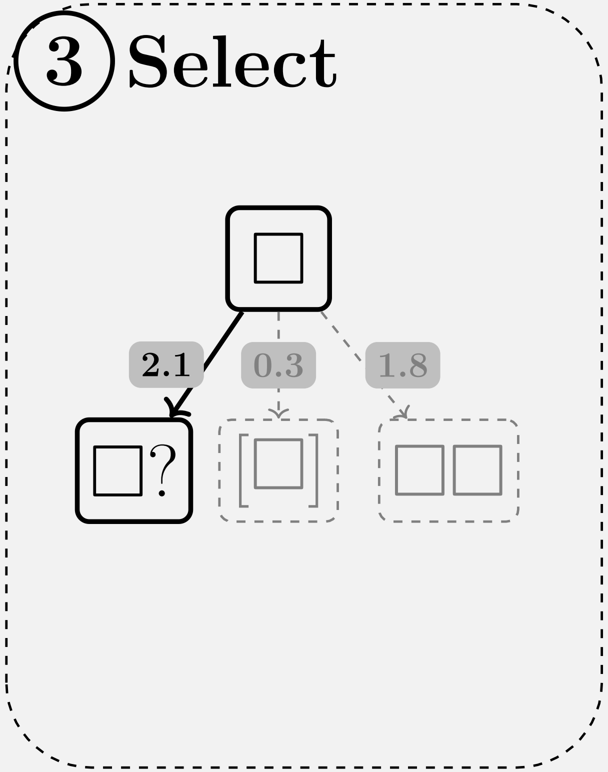
1

Algorithm: Expand, Score, Select



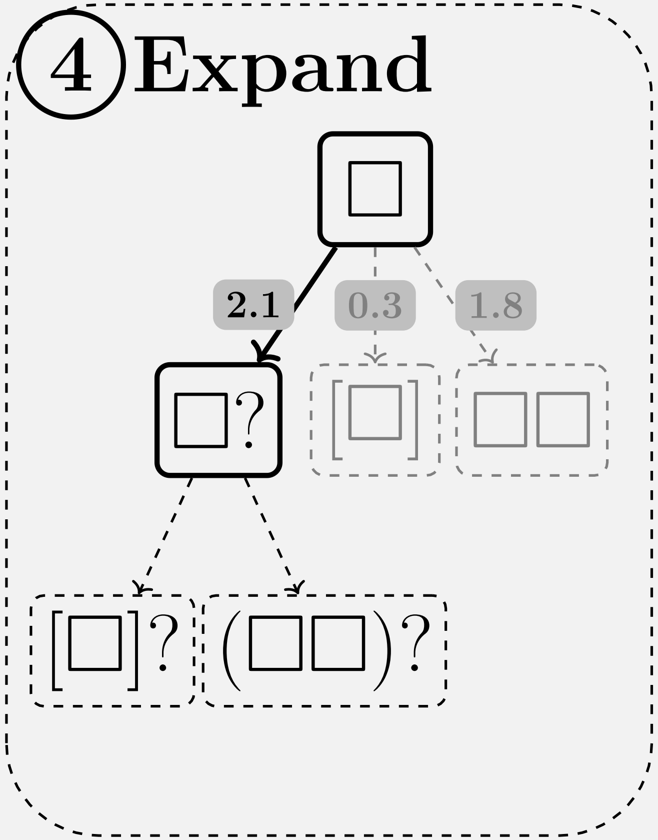
1

Algorithm: Expand, Score, Select



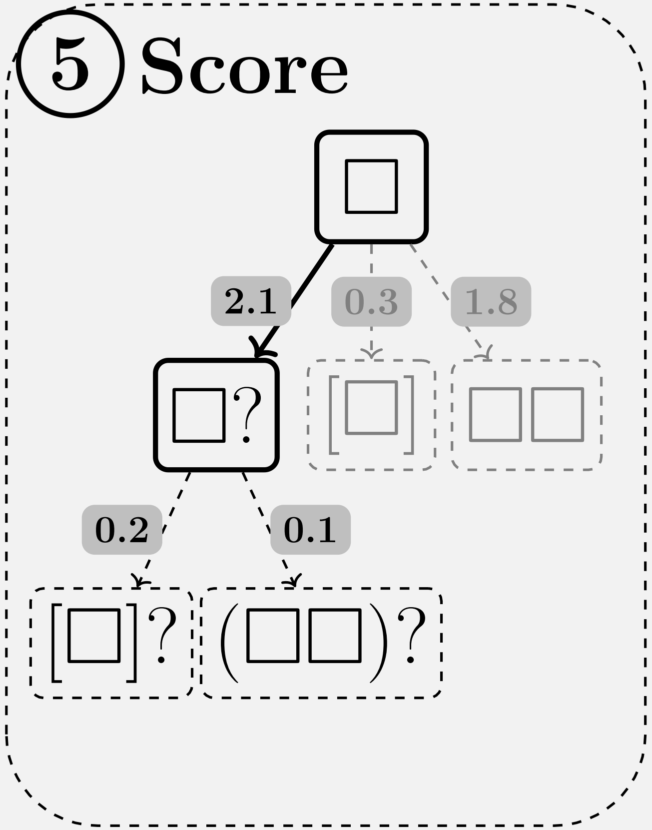
1

Algorithm: Expand, Score, Select



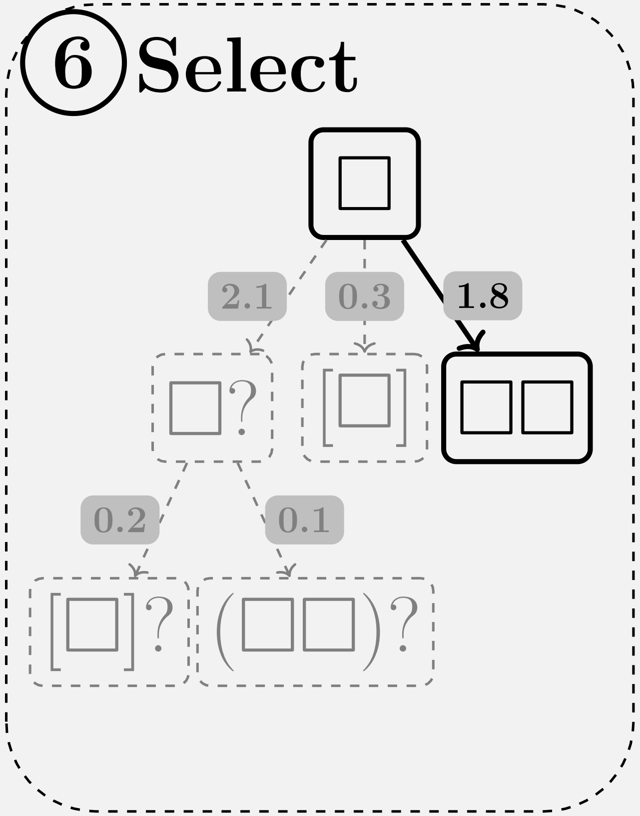
1

Algorithm: Expand, Score, Select



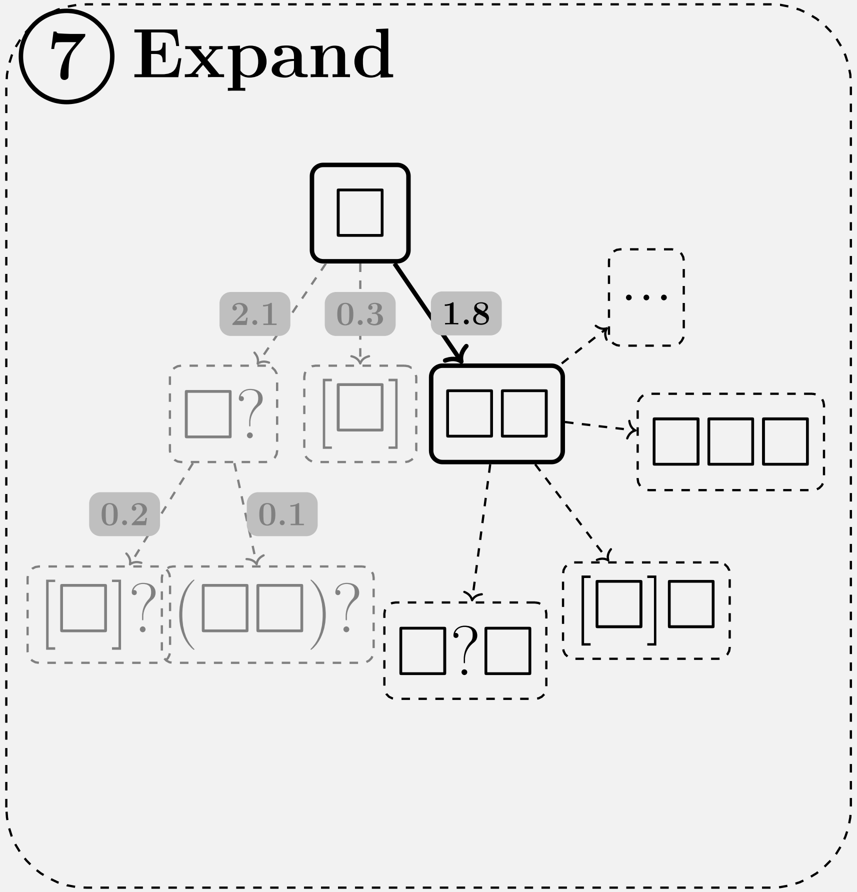
1

Algorithm: Expand, Score, Select

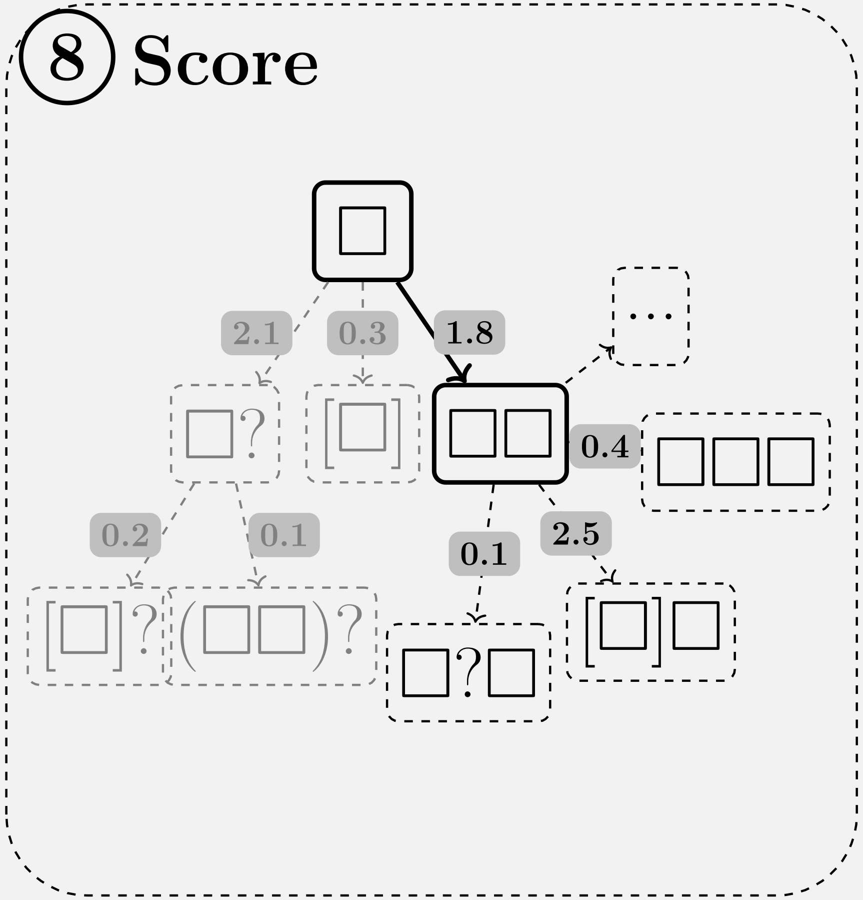


1

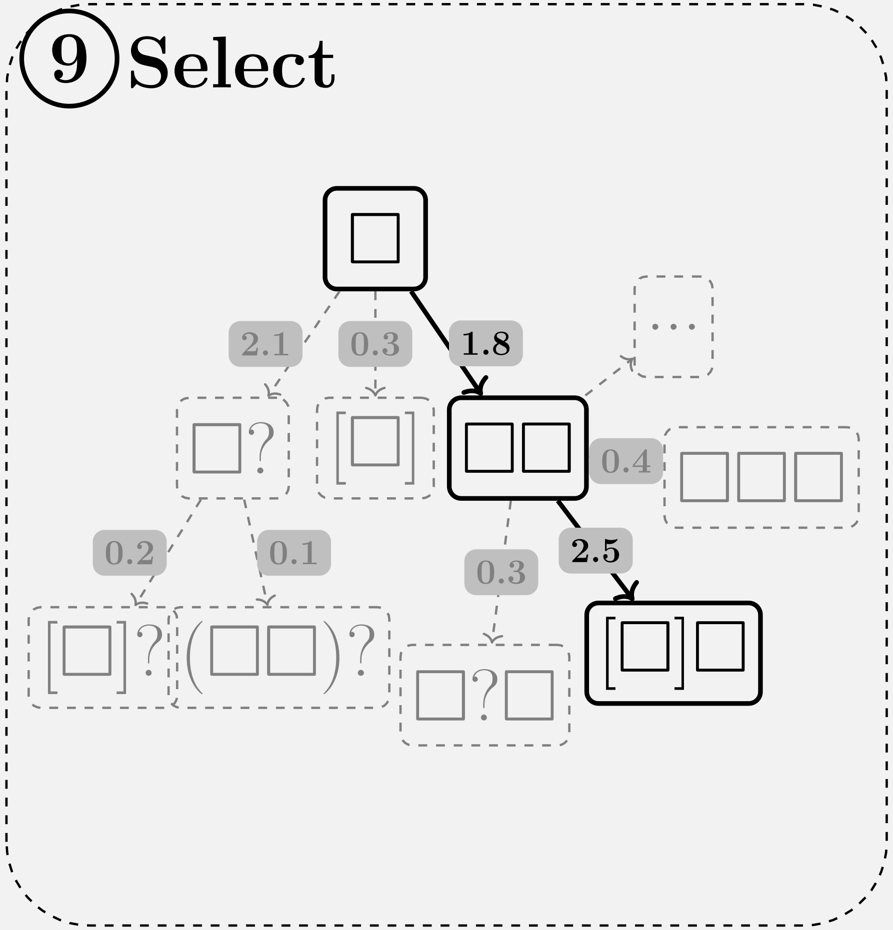
Algorithm: Expand, Score, Select



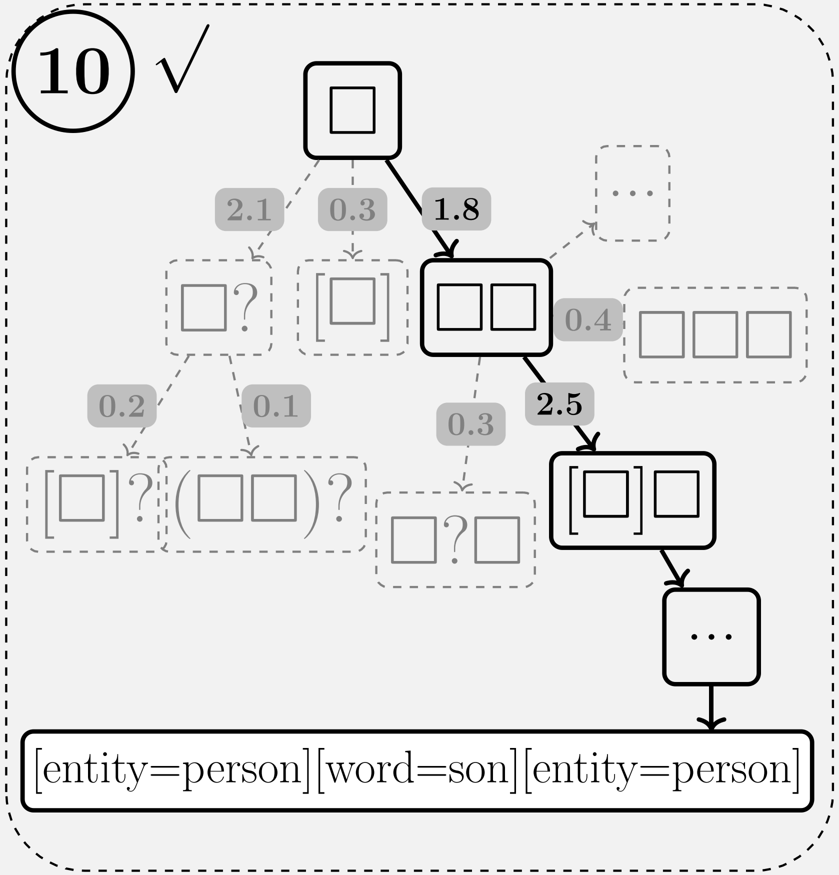
Algorithm: Expand, Score, Select



Algorithm: Expand, Score, Select



Algorithm: Expand, Score, Select

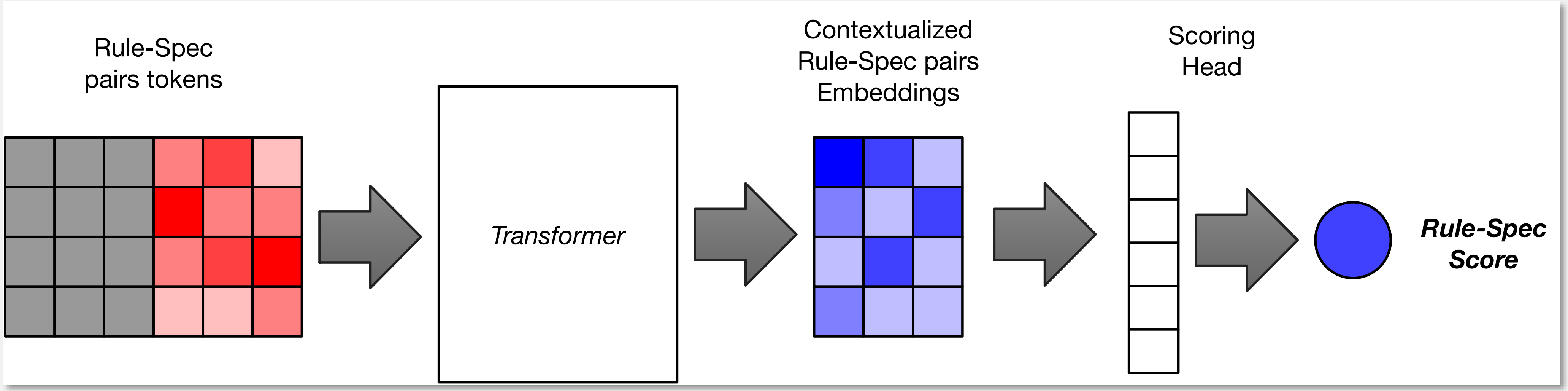


Rule Scoring

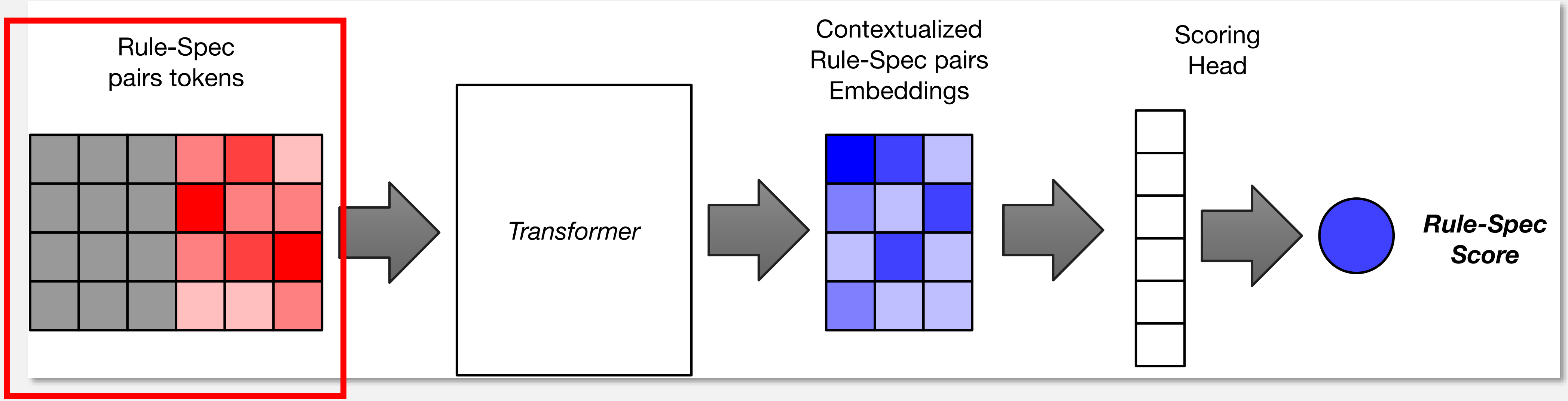
[CLS] [word = Technical][tag = Noun] [word = □] □[SEP] She is already an honorary doctor of the <MATCH> Technical University of CLUJ-NAPOCA </MATCH> [SEP]

[CLS] [word = Technical][tag = Noun] [word = □] □[SEP] … using a modified version of the TUD (<MATCH> Technical University of Denmark </MATCH>) radar. [SEP]

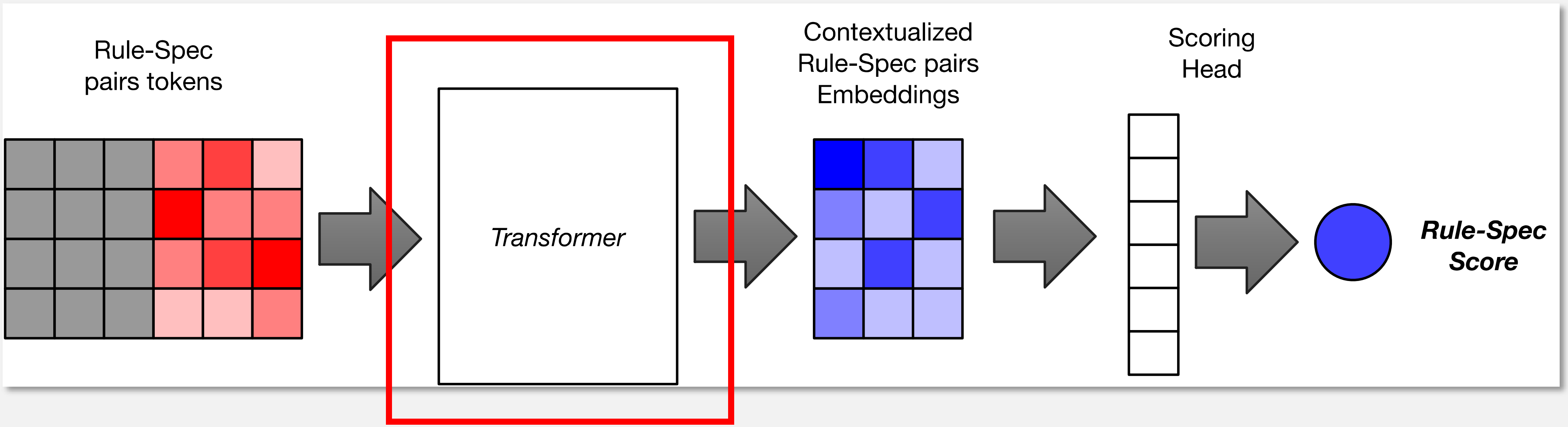
Rule Scoring Architecture



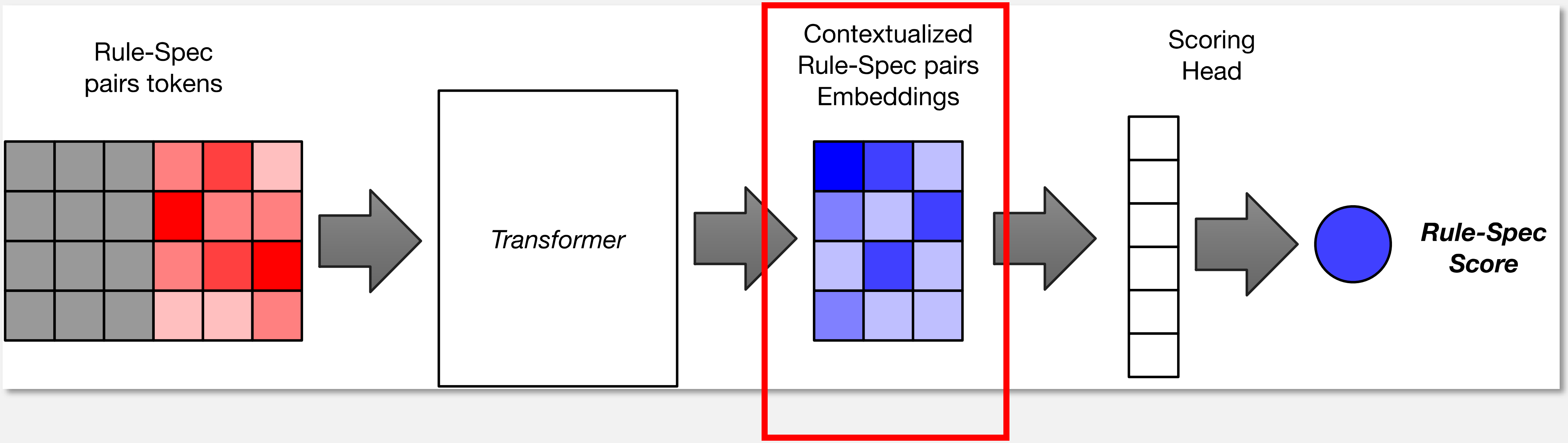
Rule Scoring Architecture



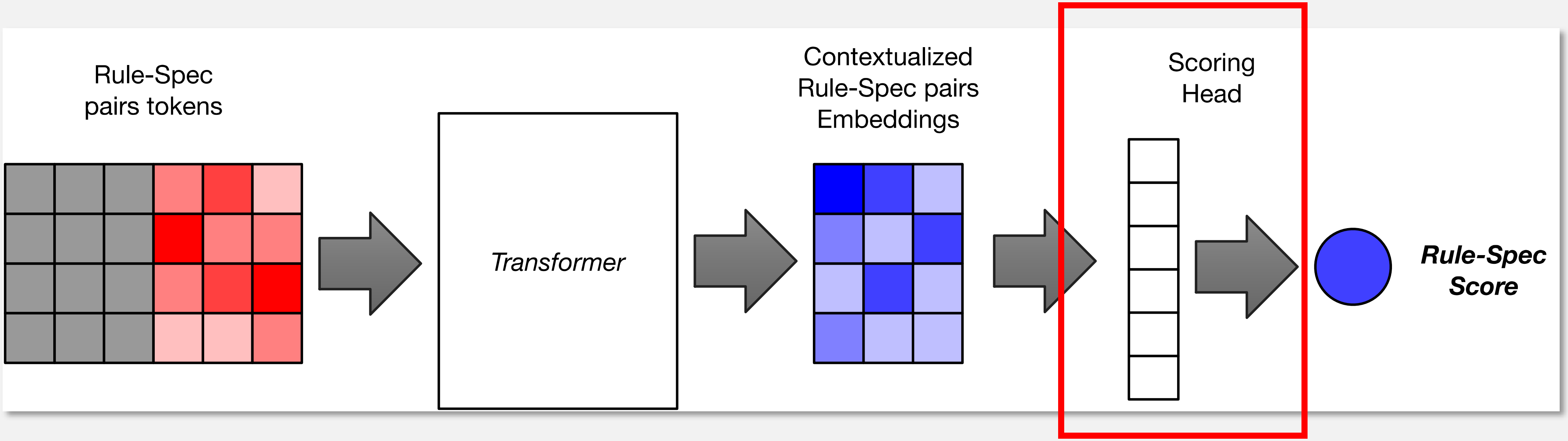
Rule Scoring Architecture



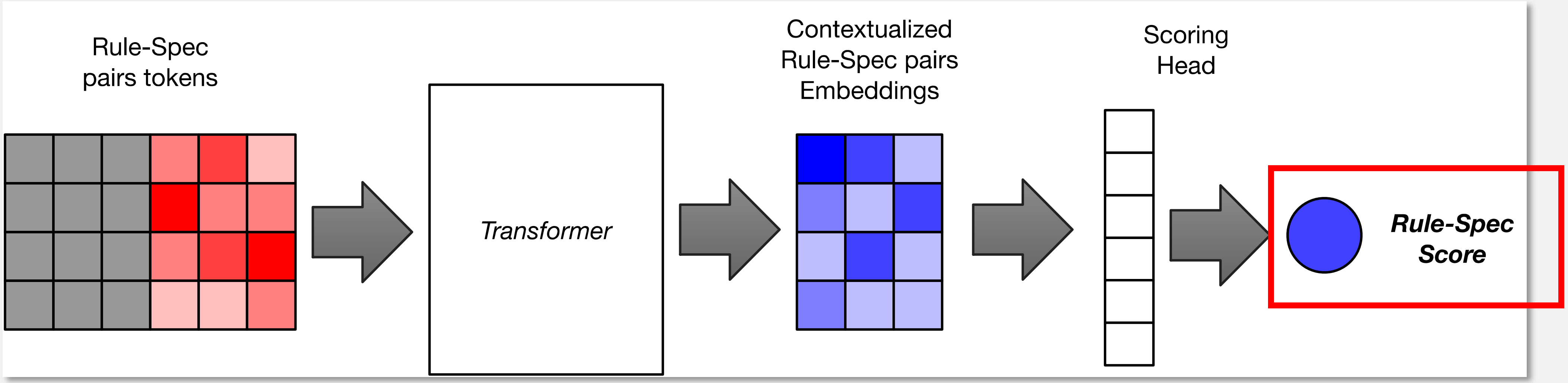
Rule Scoring Architecture



Rule Scoring Architecture



Rule Scoring Architecture



Rule Synthesis Performance

Loss Function	Spec Aggregation	Exact Matches	Partial Matches	Total Matches
Margin	Attention	13%	**54%**	67%
Margin	Average	11%	25%	36%
MSE	Attention	**19%**	51%	**70%**
MSE	Average	12%	49%	60%
Margin	No Specification	0%	6%	6%

Number of matches in the specifications of the testing set. Exact matches are cases in which all the spans in the specification are matched exactly. Partial matches are cases where a) a span is missing or b) an incorrect span is matched. The last column is the sum of both.

Rule Synthesis Performance

Loss Function	Spec Aggregation	Exact Matches	Partial Matches	Total Matches
Margin	Attention	13%	**54%**	67%
Margin	Average	11%	25%	36%
MSE	Attention	**19%**	51%	**70%**
MSE	Average	12%	49%	60%
Margin	No Specification	0%	6%	6%

Number of matches in the specifications of the testing set. Exact matches are cases in which all the spans in the specification are matched exactly. Partial matches are cases where a) a span is missing or b) an incorrect span is matched. The last column is the sum of both.

Rule Synthesis Performance

Loss Function	Spec Aggregation	Exact Matches	Partial Matches	Total Matches
Margin	Attention	13%	**54%**	67%
Margin	Average	11%	25%	36%
MSE	Attention	**19%**	51%	**70%**
MSE	Average	12%	49%	60%
Margin	No Specification	0%	6%	6%

Number of matches in the specifications of the testing set. Exact matches are cases in which all the spans in the specification are matched exactly. Partial matches are cases where a) a span is missing or b) an incorrect span is matched. The last column is the sum of both.

Automatically Generated Rules

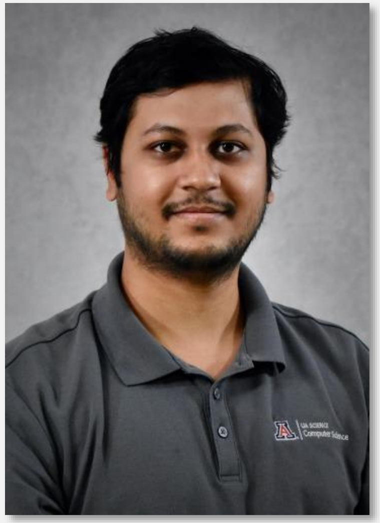
Target Rule	Generated Rule	User Specification Match Rate
[tag=Noun] [lemma=in] [lemma=many] [lemma=of] [tag=Prep]	[tag=Noun] [lemma=in] **[raw=many]** [word=of] [tag=Prep]	100%
[tag=Adj] [lemma=policy] [tag=EX] [tag=Verb]	**[lemma=same] [raw=policy]** [tag=EX] **[word=has]**	100%
[tag=Verb] [lemma=how] [tag=Deet]	**[lemma=describe] [raw=how] [word=the]**	58.8%
[lemma=under | lemma=water]	[lemma=latter | tag=Noun]	4.7%

Discrepancies with the target rule are highlighted

Future Directions

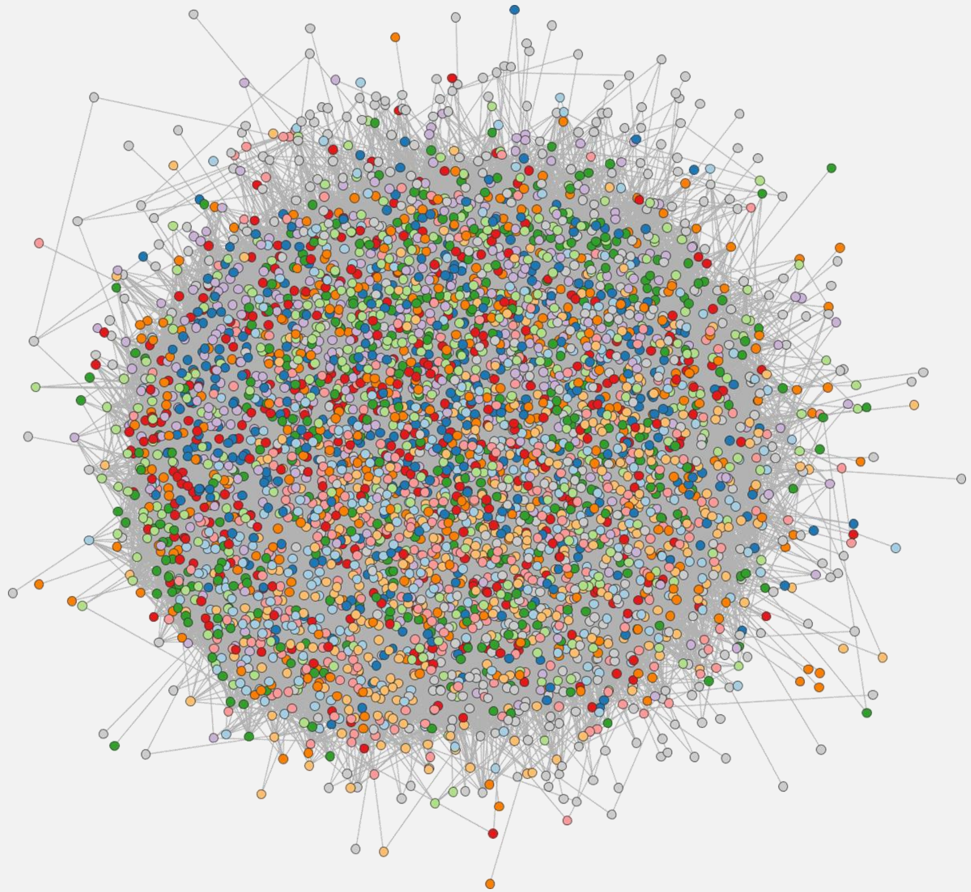
- Introduce syntactic constraints
- Improve run time and achieve interactivity
- Generate more than one rule to match a user specification

Visualization of Information Extraction



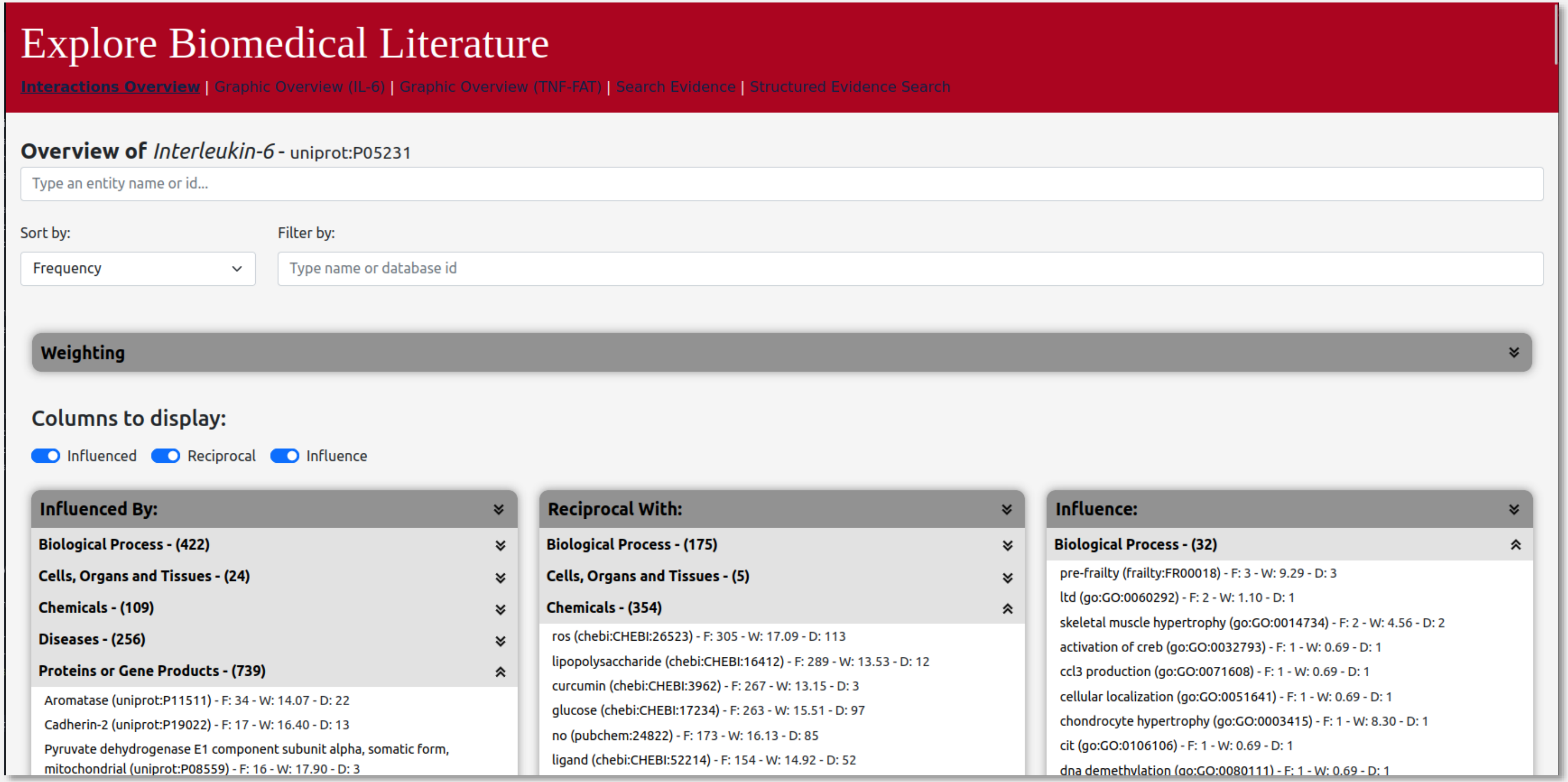
Visualization Design Goals

- Efficiently search and locate:
  - Mechanistic interactions
  - Underlying textual evidence
  - The pointer to the source of the information
- Reduce the *hairball* effect
- Search paradigm:
  - Narrow down search space
  - Iteratively search

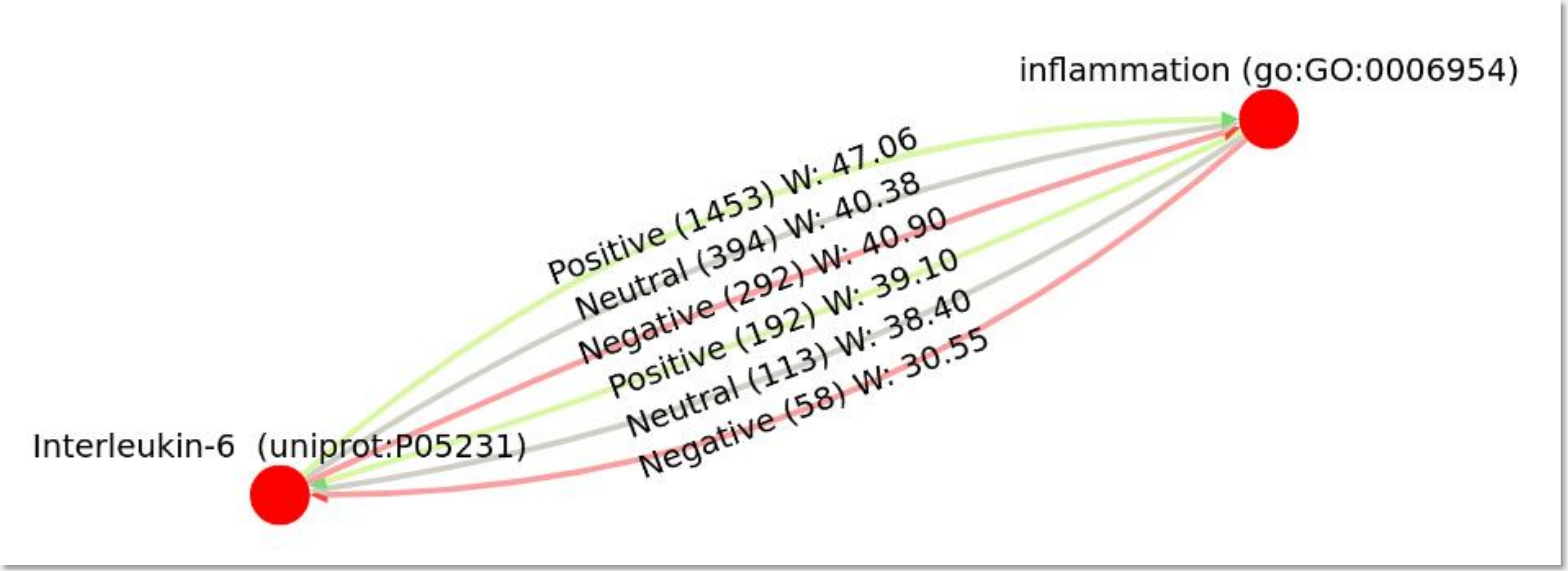


Structural Search

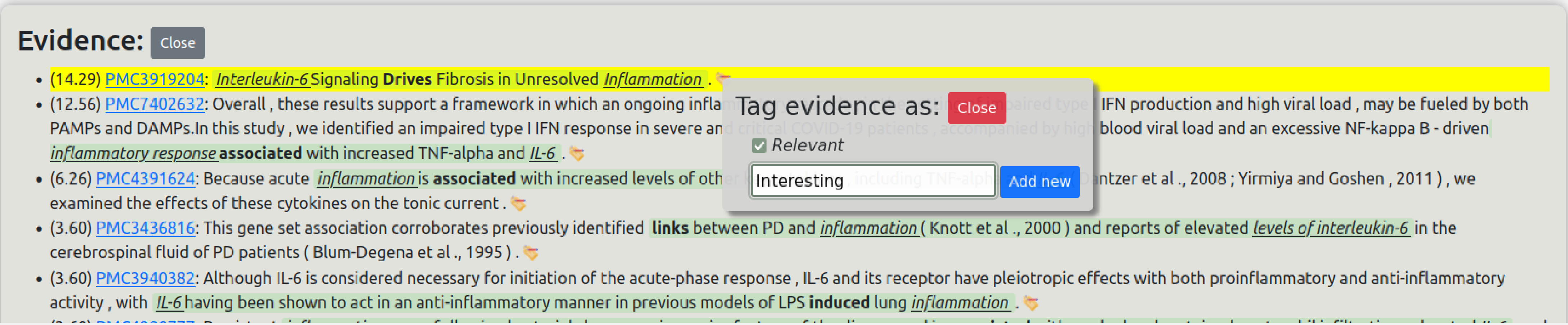
Search, navigate, and visualize exploiting underlying network structure



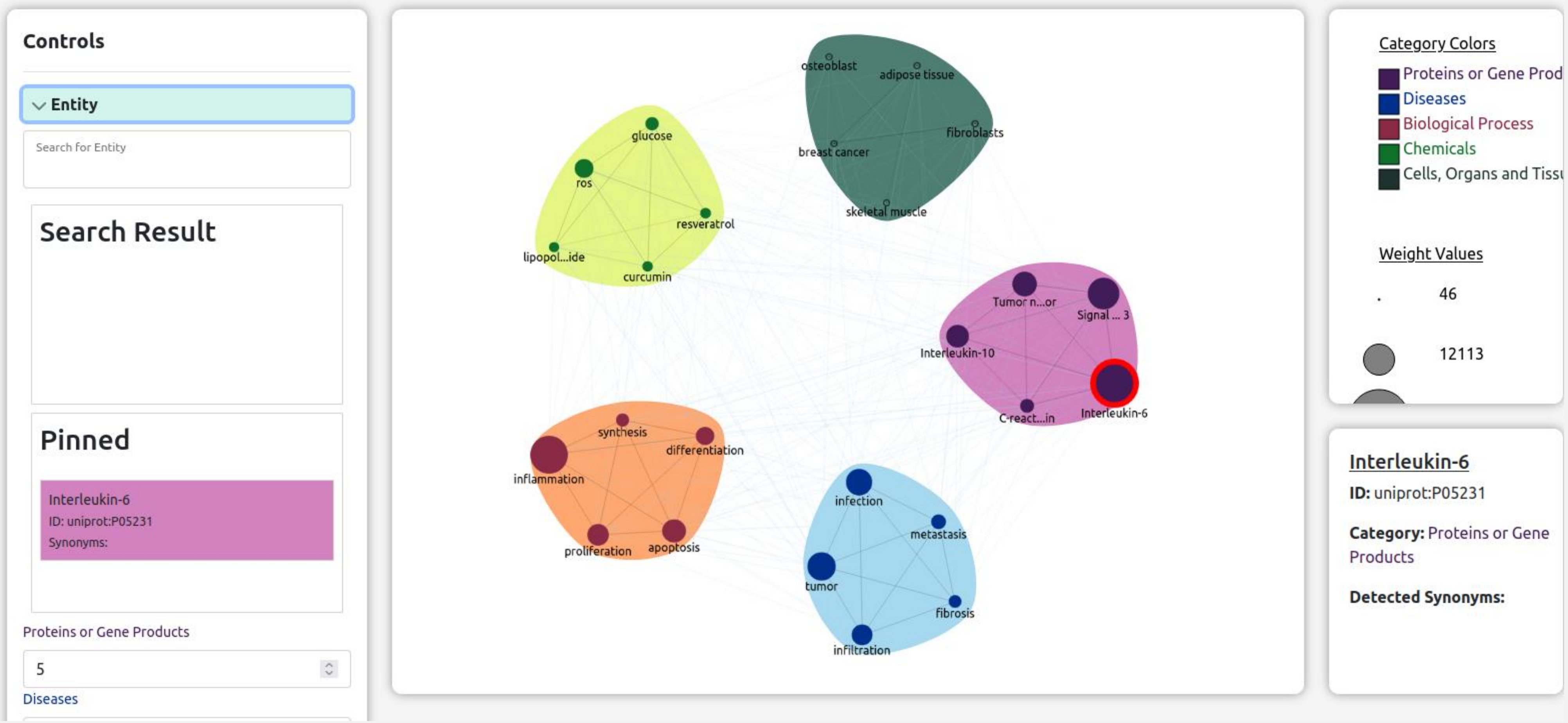
Node-Link Visualization



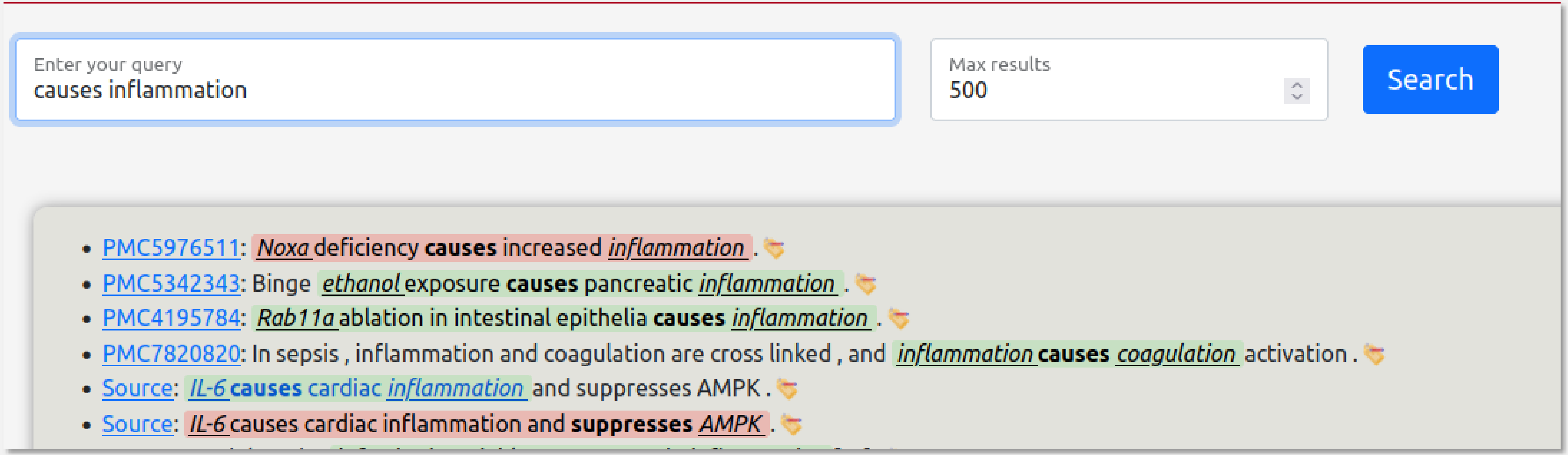
Evidence Panel



Graphical Overview



Textual Search



Conclusions

Conclusions

- NLP can *enhance* your research capabilities
- Don’t have to be an ML expert to leverage NLP
- Combining Viz + NLP helps drive new scientific discoveries

Thank You!