Thesis Oral

Grounded Knowledge Bases for Scientific Domains

Dana Movshovitz-Attias

August 26, 2015

Thesis Committee:
William Cohen, Chair
Tom Mitchell
Roni Rosenfeld
Alon Halevy, Google Research

Carnegie Mellon University Computer Science Department

Who is Barack Obama's wife?

Who is Barack Obama's wife?

Michelle Obama (m. 1992)

Barack Obama, Spouse

Michelle LaVaughn Robinson Obama is an American lawyer and writer. She is married to the 44th and current President of the United States, Barack Obama, and is the first African-American First Lady of the United States. Wikipedia

More about Michelle Obama

Family of Barack Obama - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Family_of_**Barack**_**Obama** Wikipedia
Michelle Obama, née Robinson, the **wife** of **Barack Obama**, was born on January 17, 1964, in Chicago, Illinois. She is a lawyer and was a University of Chicago ...
Sidwell Friends School - Marian Shields Robinson - Bo - Charles T. Payne

Michelle Obama - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Michelle_**Obama** Wikipedia
Michelle LaVaughn Robinson **Obama** (born January 17, 1964) is an American lawyer and writer. She is the **wife** of the 44th and current **President** of the United ...
Craig Robinson (basketball) - Hyde Park, Chicago - Sidley Austin - Valerie Jarrett

Michelle Obama - Biography - U.S. First Lady, Lawyer ...
www.biography.com/people/michelle-**obama**-307592 FYI
Explore the life of **Michelle Obama**, the 44th first lady and **wife** of President **Barack Obama**. Learn more at Biography.com.

First Lady Michelle Obama | whitehouse.gov
https://www.whitehouse.gov/.../first-lady-michelle-**obama** White House
First Lady Michelle LaVaughn Robinson Obama is a lawyer, writer, and the **wife** of the 44th and current President, **Barack Obama**. She is the first ...

Barack Obama was asked about 'his first wife' by a woman ...
www.telegraph.co.uk › ... › Barack Obama The Daily Telegraph
Jan 22, 2015 - **Barack Obama** was asked about his 'first **wife**' by a woman who took a ... when she gave Obama a gift of green lipstick for **Michelle Obama**, who ...

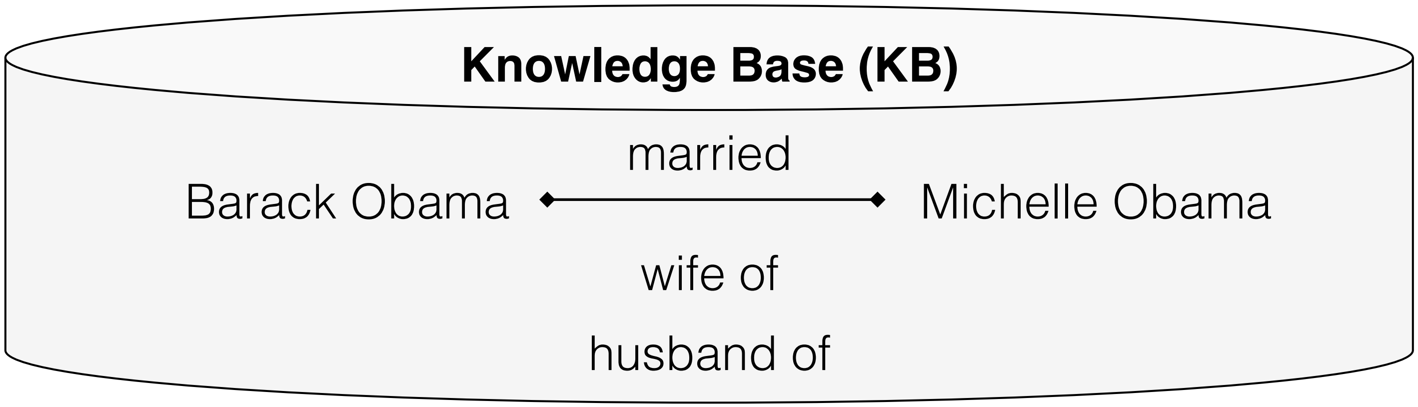
Michelle Obama - First Lady and wife of President Barack ...
www.telegraph.co.uk › News › World News The Daily Telegraph

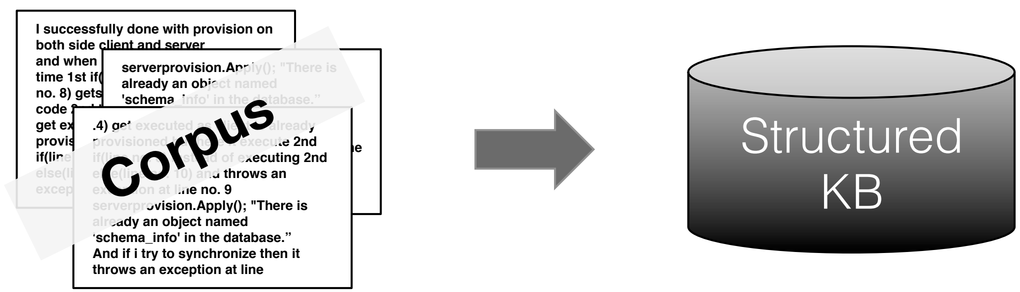
Michelle Obama - Biography - U.S. First Lady, Lawyer ...

www.biography.com/people/michelle-**obama**-307592
Explore the life of Michelle Obama, the 44th first lady and **wife** of President **Barack Obama**. Learn more at ...

Michelle Obama - First Lady and wife of President Barack ...
www.telegraph.co.uk › News › World News
Jul 1, 2015
Michelle LaVaughn Obama, First Lady and **wife** of US President **Barack Obama**: All the latest news and ...

Who is **Barack Obama**'s **wife**?





DBpedia

NELL

Knowledge Vault

TextRunner

YAGO

Freebase

Software

“What is the **run time** of **quick sort**?”

Quicksort - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/**Quicksort** ▾ Wikipedia ▾
Animated visualization of the **quicksort** algorithm. The horizontal lines are pivot values. **Quicksort** (sometimes called partition-exchange **sort**) is an efficient **sorting** algorithm, serving as a systematic method for placing the elements of an array in order.
Tony Hoare - Dutch national flag problem - Robert Sedgewick - Flashsort

Running Time of Quicksort - HackerRank
https://www.hackerrank.com/challenges/**quicksort**4 ▾
Mar 4, 2013 - The **running time of Quicksort** will depend on how balanced the partitions are. If you are unlucky and select the greatest or the smallest element ...

Quicksort Running Time - Math StackExchange
math.stackexchange.com/.../**quicksort-running-time** ▾ Stack Exchange ▾
Feb 18, 2011 - ... the recurrence will be: The above recurrence has the solution (I will prove this later): Hence the **running time of QUICKSORT** in this case is ...

Quick Sort - Personal.kent.edu
www.personal.kent.edu/~rmuhamma/Algorithms/.../**Sorting/quickSort**.ht... ▾
The **running time of quick sort** depends on whether partition is balanced or unbalanced, which in turn depends on which elements of an array to be **sorted** are used for partitioning. A very good partition splits an array up into two equal sized arrays.

[PDF] Quick Sort
https://www.cse.ust.hk... ▾ Hong Kong University of Science and Technology ▾
P, followed by the results of **quicksort**(S. 2.)) ... void **quicksort**(int A[ ], int left, int right).

…the average case running time is θ (n logn)…

[PDF] Quicksort
www.bowdoin.edu/~ltoma/...**quicksort/quicksort**.pdf ▾ Bowdoin College ▾
Quicksort(A, q + 1,r). Fl. Sort using **Quicksort**(A,1,n). If q = n/2 and we divide in Θ(n) time, we again get the recurrence T(n)=2T(n/2) + Θ(n) for the **running time** ...

Biomedical

“What is the **molecular mass** of **BamA**?”

[PDF] Quiz 2 answers - with worked solutions - Bama.ua.edu
www.**bama**.ua.edu/.../quiz2_key_with_solutions.p... ▾ University of Alabama ▾
Quiz 2 answers - with worked solutions. 1. What is average mass, in grams, of one atoms of iron? use mol wt of iron (= **molar mass** = mass for one mol of Fe ...

[PPT] Polymers: Introduction - Bama.ua.edu
bama.ua.edu/~kshaughn/ch338/.../poly-lecture.ppt ▾ University of Alabama ▾
Monomer: Low **molecular weight** compound that can be connected together to give a poymer; Oligomer: Short polymer chain; Copolymer: polymer made up of 2 ...

[PDF] Hydrogen/Deuterium exchange mass spectrometry - Bam...
bama.ua.edu/.../Busenlehner_ABBreview_2005.p... ▾ University of Alabama ▾
by LS Busenlehnera - 2005 - Cited by 143 - Related articles
Oct 5, 2004 - exchange mass spectrometry (H/D exchange MS) is emerging as an efficient ... proteins having **molecular masses** in excess of 50 kDa.

TP0326, a Treponema pallidum β-Barrel Assembly ...
www.ncbi.nlm.nih.gov/... ▾ National Center for Biotechnology Information ▾
by DC Desrosiers - 2011 - Cited by 32 - Related articles
Apr 27, 2011 - In E. coli, **BamA** is the central component of a multi-protein complex consisting ..... Native TP0326 forms part of a high **molecular mass** complex.

The crystal structure of BamB suggests interactions with ...
www.ncbi.nlm.nih.gov/... ▾ National Center for Biotechnology Information ▾
by N Noinaj - 2011 - Cited by 47 - Related articles
Jan 26, 2011 - It interacts with the periplasmic domain of **BamA**, an integral outer .... We determined the **molecular mass** of BamB in solution using size ...

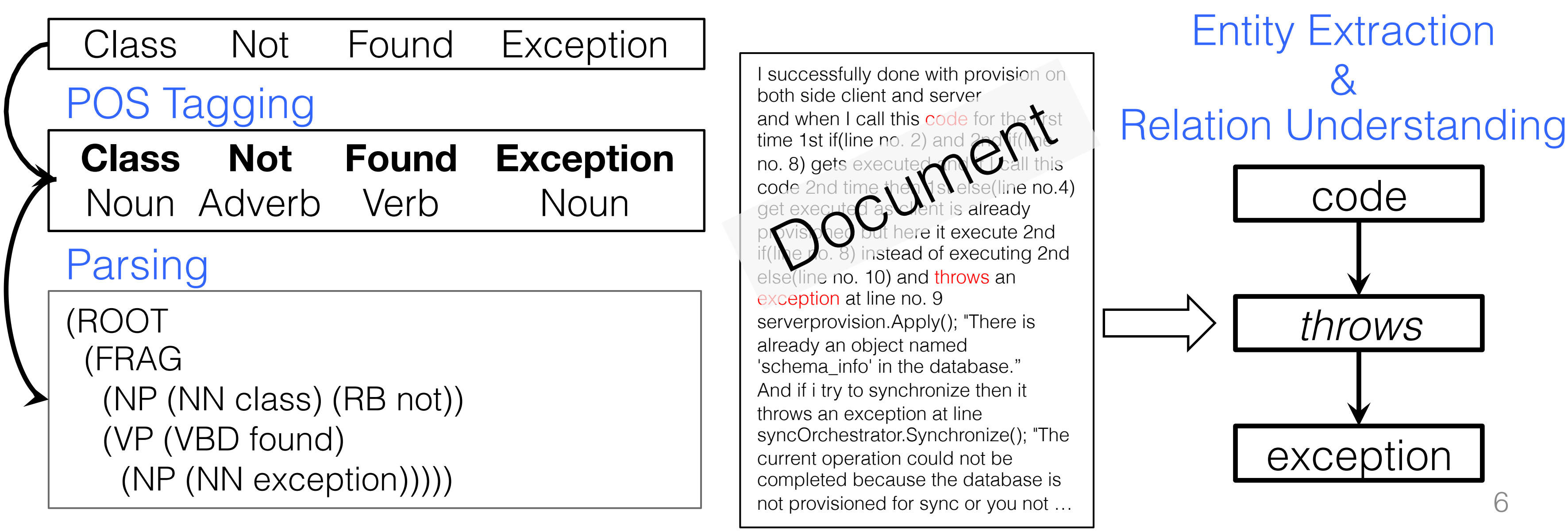
…BamA at its expected molecular weight (~90 kDa)…

BAMA Course Survey: ... Use masses liste above to calculate the **molecular mass**. ... This keeps the shape of the molecule "stra " or "linear" allowing these ...

Challenges in Mining Scientific Text

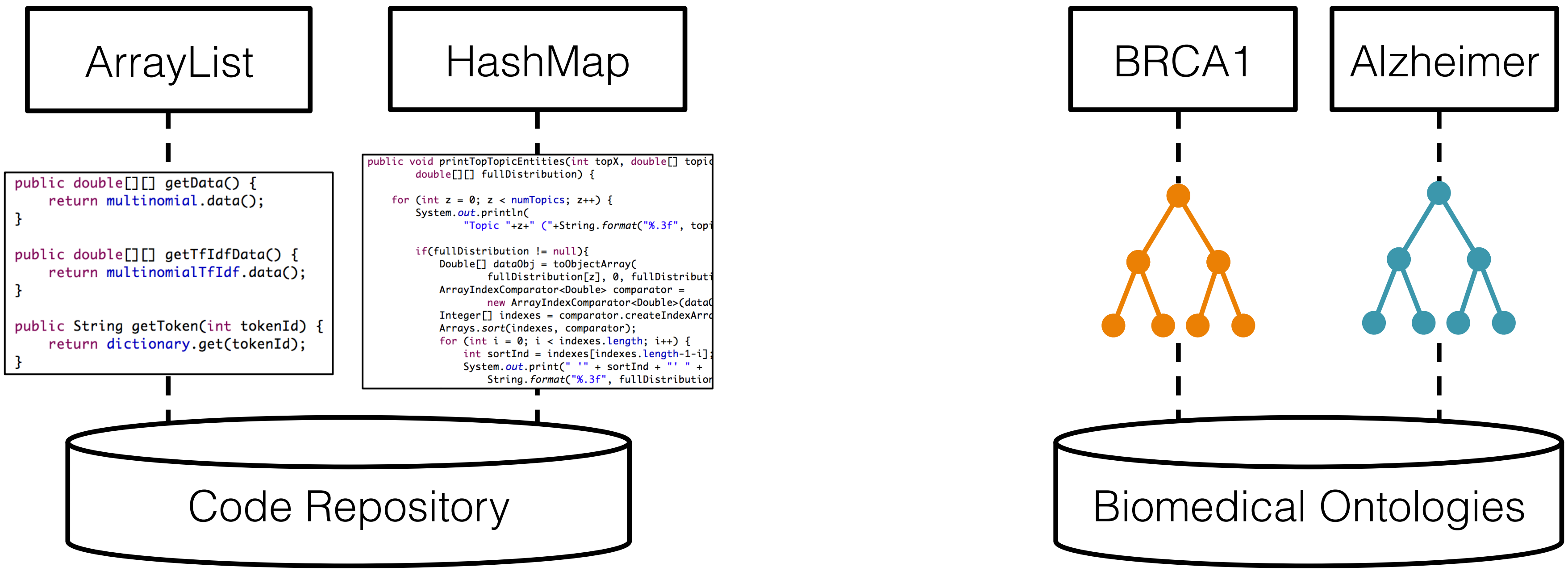
- Specialized terminology
- Domain-specific language constructs

Affected NLP Techniques



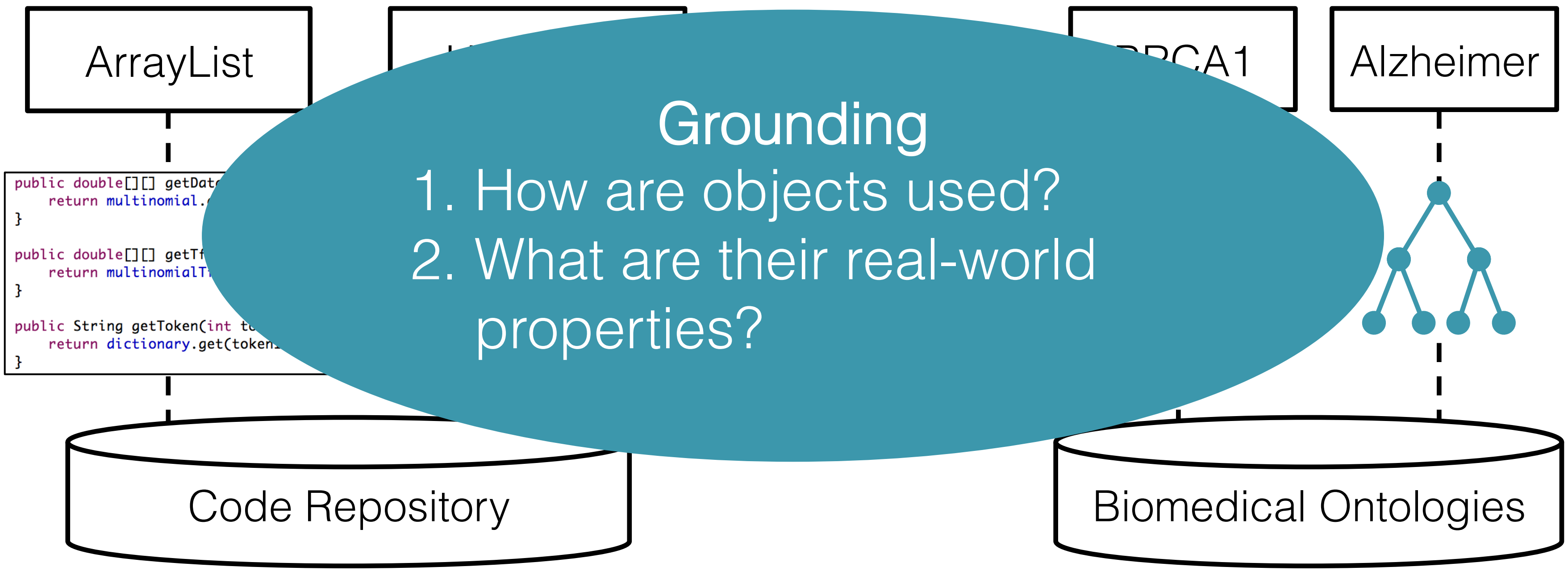
Grounding Scientific Entities

Scientific data is not found only in text → Opportunity: Domain-specific resources



Grounding Scientific Entities

Scientific data is not found only in text → Opportunity: Domain-specific resources

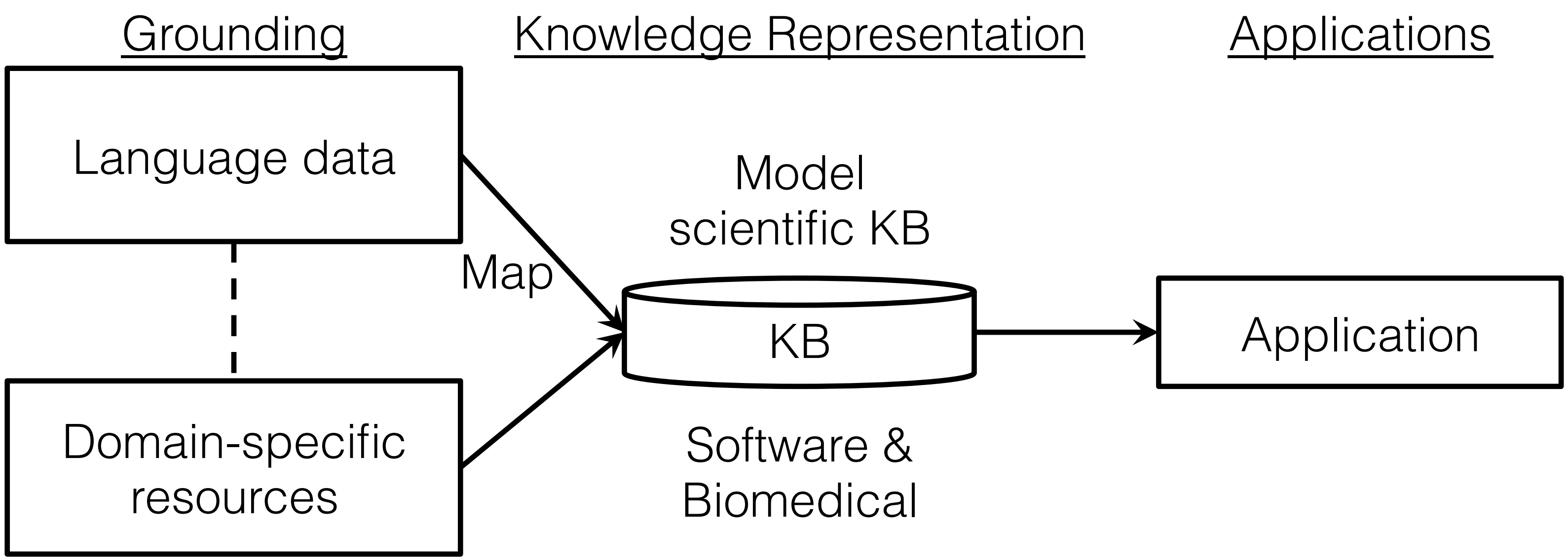


Grounding

1. How are objects used?
2. What are their real-world properties?

Thesis Statement

"**Grounding entities** to specialized data from a scientific domain facilitates improved unsupervised and semi-supervised algorithms for **Knowledge Base construction** for that domain"



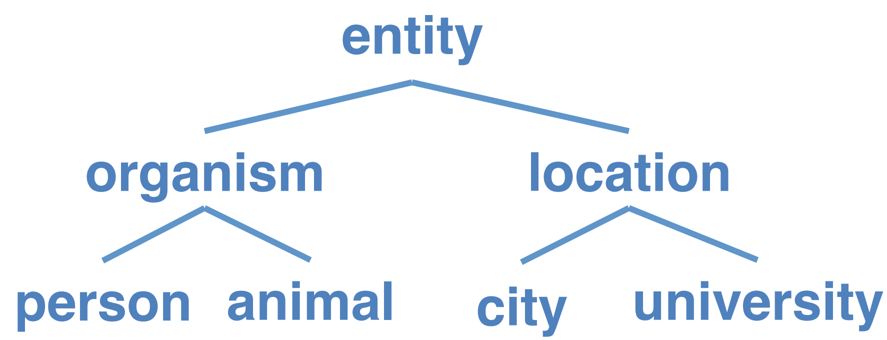
Knowledge Base Construction

Open IE

(Beijing, *is the capital of***, China)**
(Penticton, *has very***, warm summers)**
(Goods, *can be defined in***, a variety of ways)**

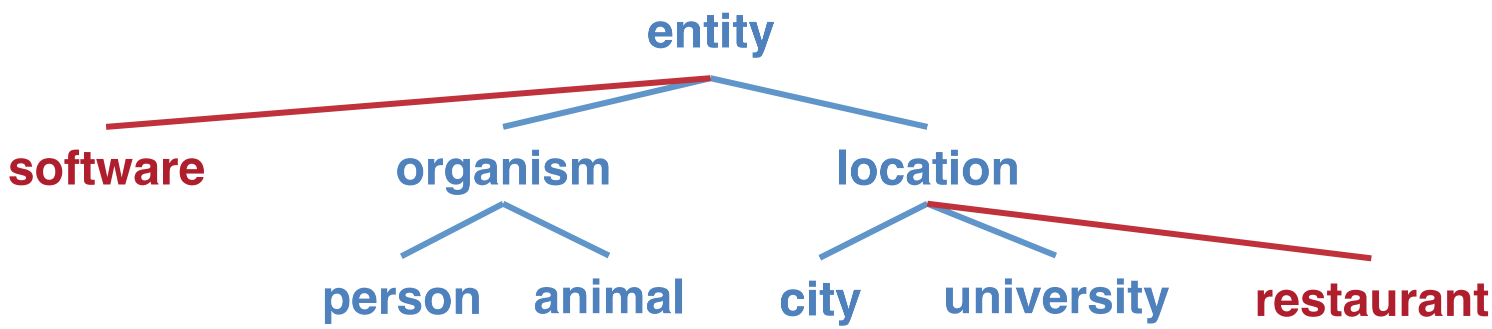
- ReVerb [Fader et al., 2011]
- TextRunner [Yates et al., 2007]

Ontology-Guided Construction



- NELL [Carlson et al., 2010]
- FreeBase [Google, 2011]
- Yago [Suchanek et al., 2007, 2008]
- Knowledge Vault [Dong et al., 2014]

Reasoning with Ontologies



✔ Ontologies give information context

✔ Easy to extract domain-specific information

✘ Ontologies are
- expensive
- require prior knowledge

✘ Manual ontology does not reflect language statistics

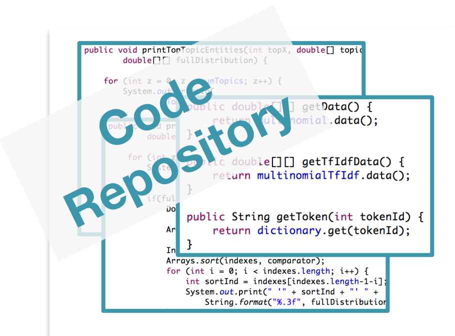
Roadmap

Statistical Language Model for
Software Domain Application

```
/* comment prediction */
```

Predicting Code Comments

Model code with statistical language models



Predict class comment

Code

```
package org.apache.lucene.index;
import java.io.IOException;
import java.io.Closeable;

public abstract class TermEnum
implements Closeable {
public abstract boolean next() throws
IOException; }
```

⇩

Comment

Abstract class for enumerating terms.
<p>Term enumerations are always ordered by Term.compareTo(). Each term in
the enumeration is greater than all that precede it.

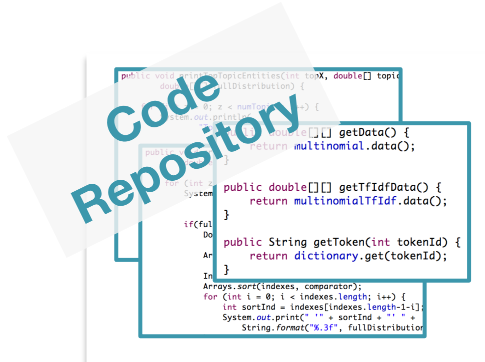
Evaluate how much typing can we save?

Up to 47% !

Train a named-entity extractor

[Movshovitz-Attias & Cohen, ACL13]

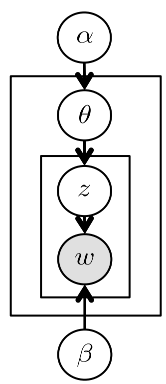
Code Modeling



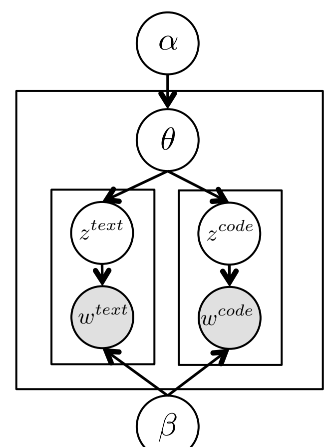
Language and domain understanding

1. Shallow statistical model
N-gram

2. BOW topic model
LDA



3. Basic domain encoding (code vs. text entities)
Link-LDA



Lessons

Predicting Code Comments

- Comments are highly predictable
  - 47% of characters predicted
- Un-intuitively: Shallow statistical method performed best

3-gram: Train a named-entity extractor

link-LDA: Train a named-entity extractor

- Intuitively: Domain encoding improves performance

Link-LDA > LDA

Hypothesis

We Need Deeper Understanding of Domain Entities

```
/* This method reads the next double from stdin */
                              float
                              long
                              int
```

Data Type

Numerical Data Type

Signed Floating-Point Type

`double` `float`

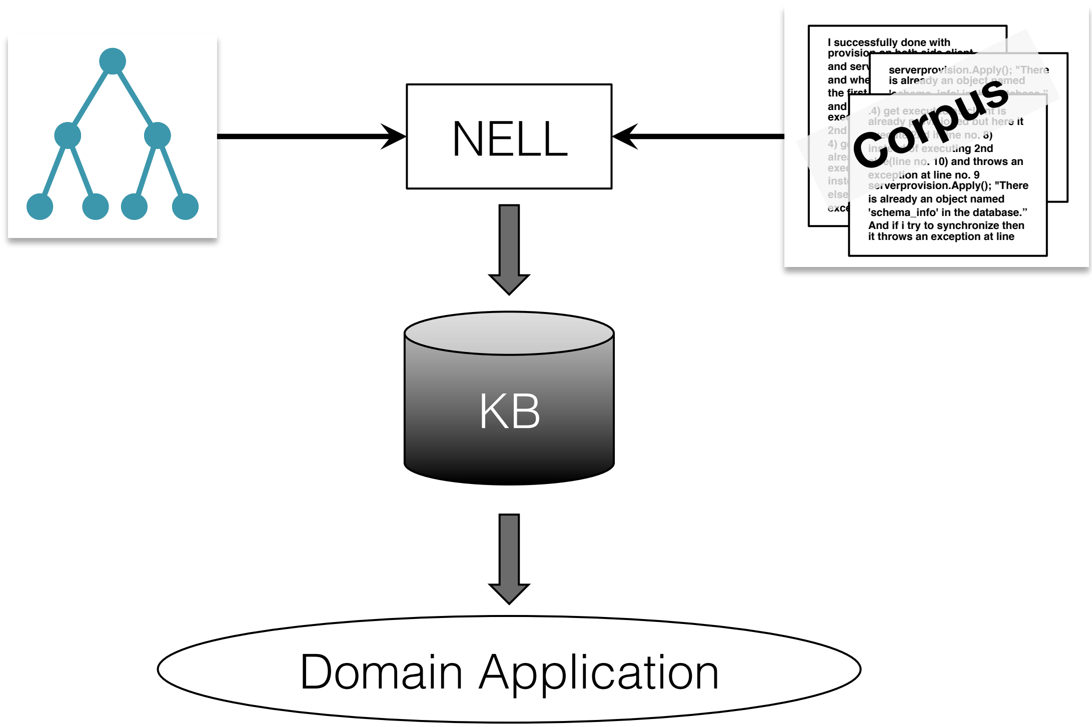
Signed Integer Type

`long` `int`

Semantics and ***categorical understanding*** contribute to language modeling task

Hypothesis

We Need Deeper Understanding of Domain Entities

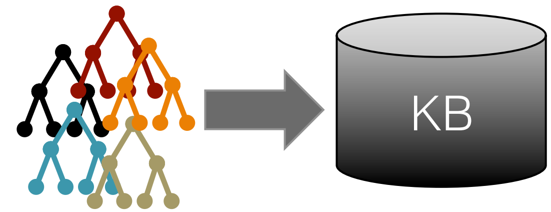


Semantics and ***categorical understanding*** contribute to language modeling task

Roadmap

Statistical Language Model for
Software Domain Application

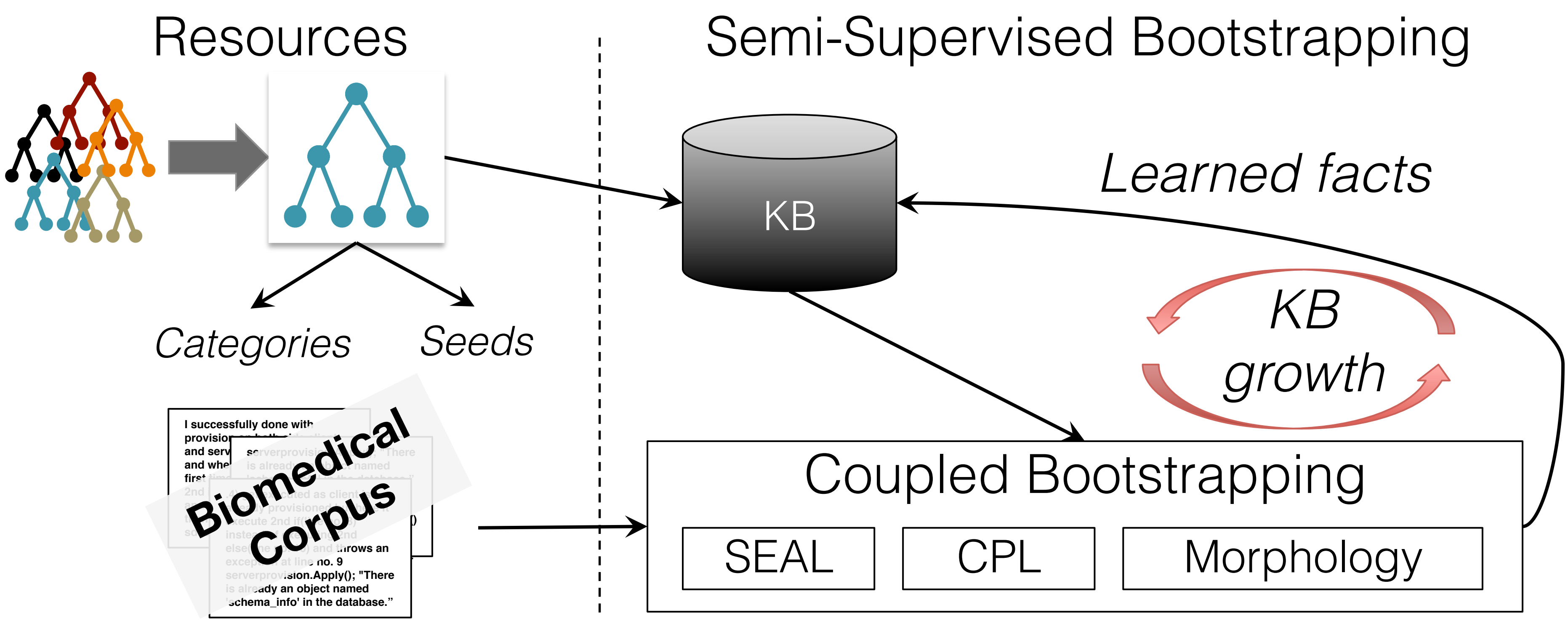
```
/* comment prediction */
```



Bootstrap KB Learning for
the Biomedical Domain

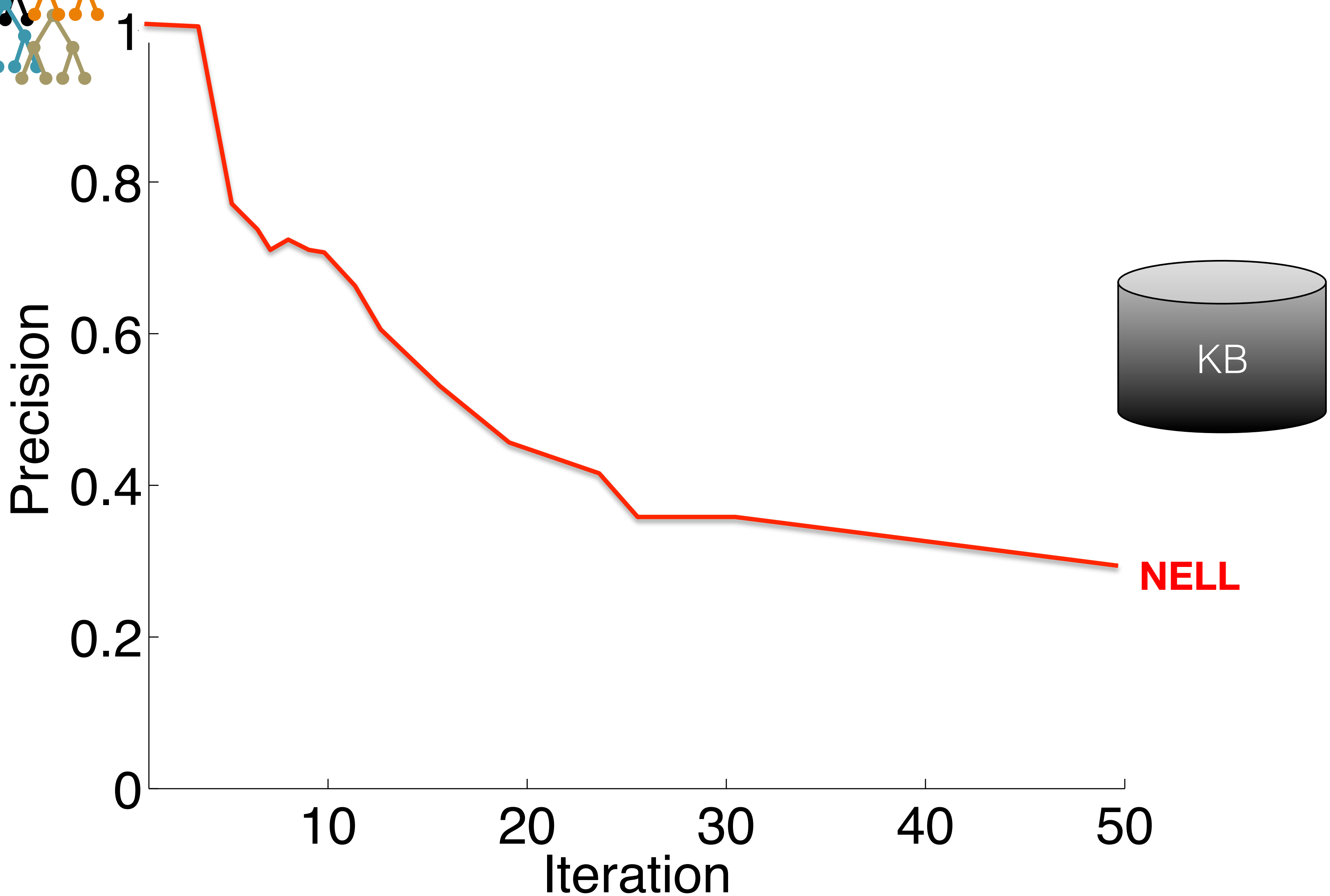
Bootstrap KB Learning for the Biomedical Domain

- Modify KB learning system (NELL) for biomedical domain



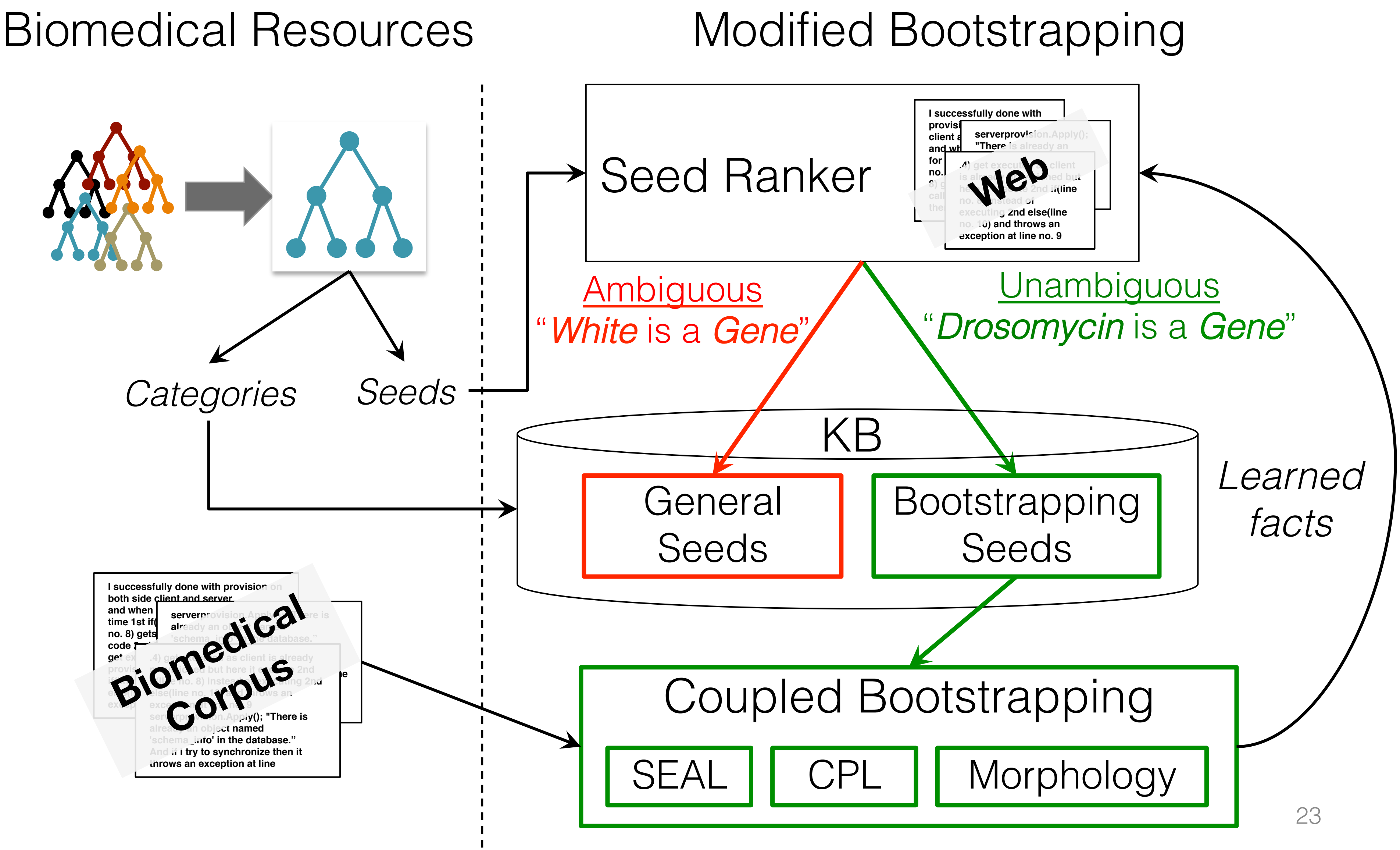
[Movshovitz-Attias & Cohen, BioNLP @ NAACL12]

Semantic Drift in NELL

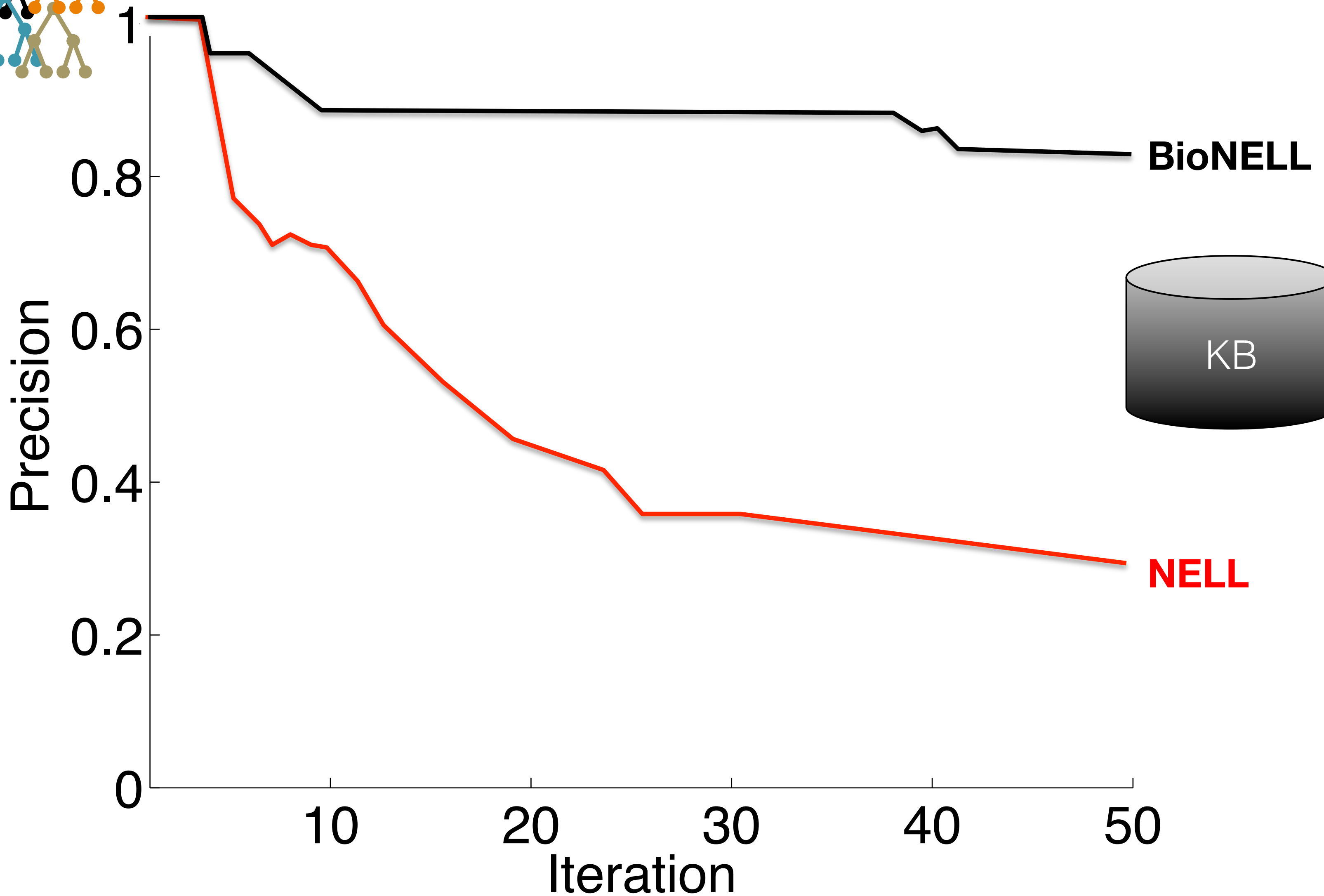


BioNELL

Constrained KB growth



Biomedical Resources
Modified Bootstrapping
Categories
Seeds
Seed Ranker
Web
Ambiguous
"White is a Gene"
Unambiguous
"Drosomycin is a Gene"
KB
General Seeds
Bootstrapping Seeds
Learned facts
Biomedical Corpus
Coupled Bootstrapping
SEAL
CPL
Morphology



1
0.8
0.6
0.4
0.2
0
Precision
10
20
30
40
50
Iteration
BioNELL
NELL
KB

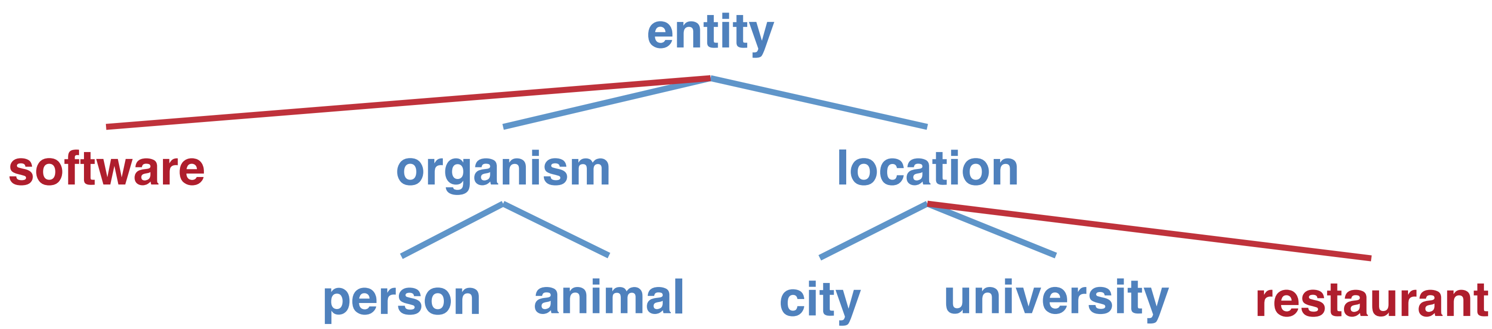
Lessons

BioNELL: Lessons

- Biomedical ontologies + filtering bootstrapping seeds lead to:
  - High-precision biomedical KB
  - Improves domain applications (NER)

- Disadvantage: Relies on input ontologies
  - No existing Software ontologies

Reasoning with Ontologies



✔ Ontologies give information context

✔ Easy to extract domain-specific information

✘ Ontologies are
- expensive
- require prior knowledge

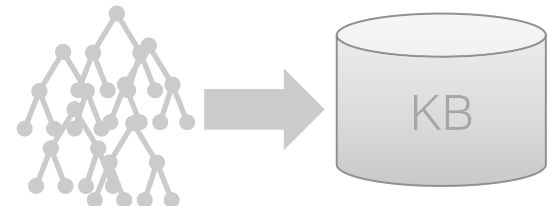
✘ Manual ontology does not reflect language statistics

✘ **Ontology and facts are often drawn from different sources**

Roadmap

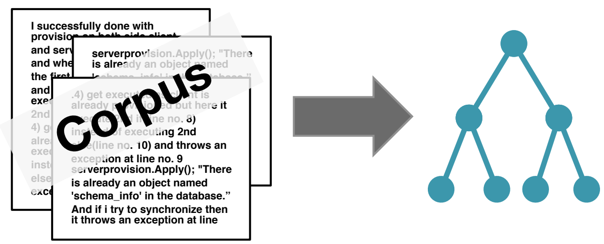
Statistical Language Model for
Software Domain Application

```
/* comment prediction */
```



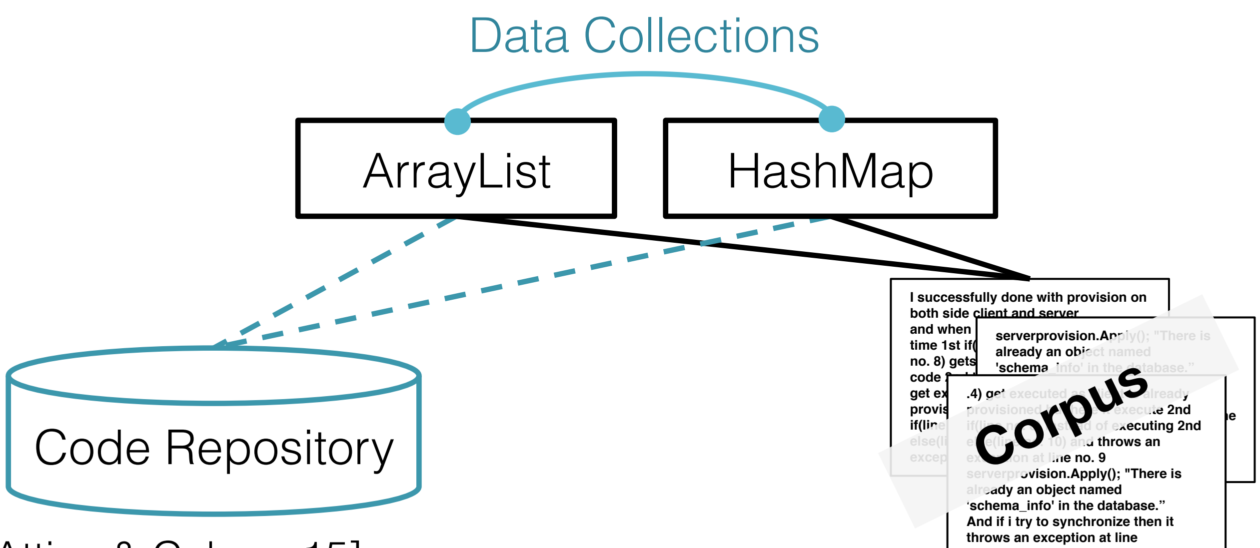
Bootstrap KB Learning for
the Biomedical Domain

**Grounded Software
Ontology Construction**



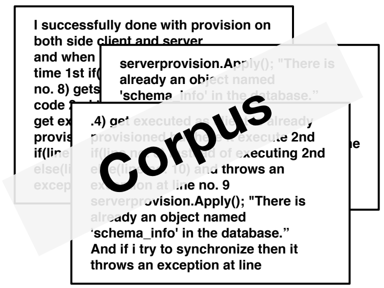
Grounded Software Ontology Construction

- We detect coordinate relations (similarity) between Java classes
  - "This method iterates over ArrayLists **and** HashMaps"



[Movshovitz-Attias & Cohen, 15]

Grounding



“…Is a root node an internal node?…”

```
public interface TreeNode
{
    /**
     * Returns the child <code>TreeNode</code> at index
     * <code>childIndex</code>.
     */
    TreeNode getChildAt(int childIndex);

    /**
     * Returns the number of children <code>TreeNode</code>s the receiver
     * contains.
     */
    int getChildCount();

    /**
     * Returns the parent <code>TreeNode</code> of the receiver.
     */
    TreeNode getParent();

    /**
     * Returns the index of <code>node</code> in the receivers children.
     * If the receiver does not contain <code>node</code>, -1 will be
     * returned.
     */
    int getIndex(TreeNode node);

    /**
     * Returns true if the receiver allows children.
     */
    boolean getAllowsChildren();

    /**
     * Returns true if the receiver is a leaf.
     */
```

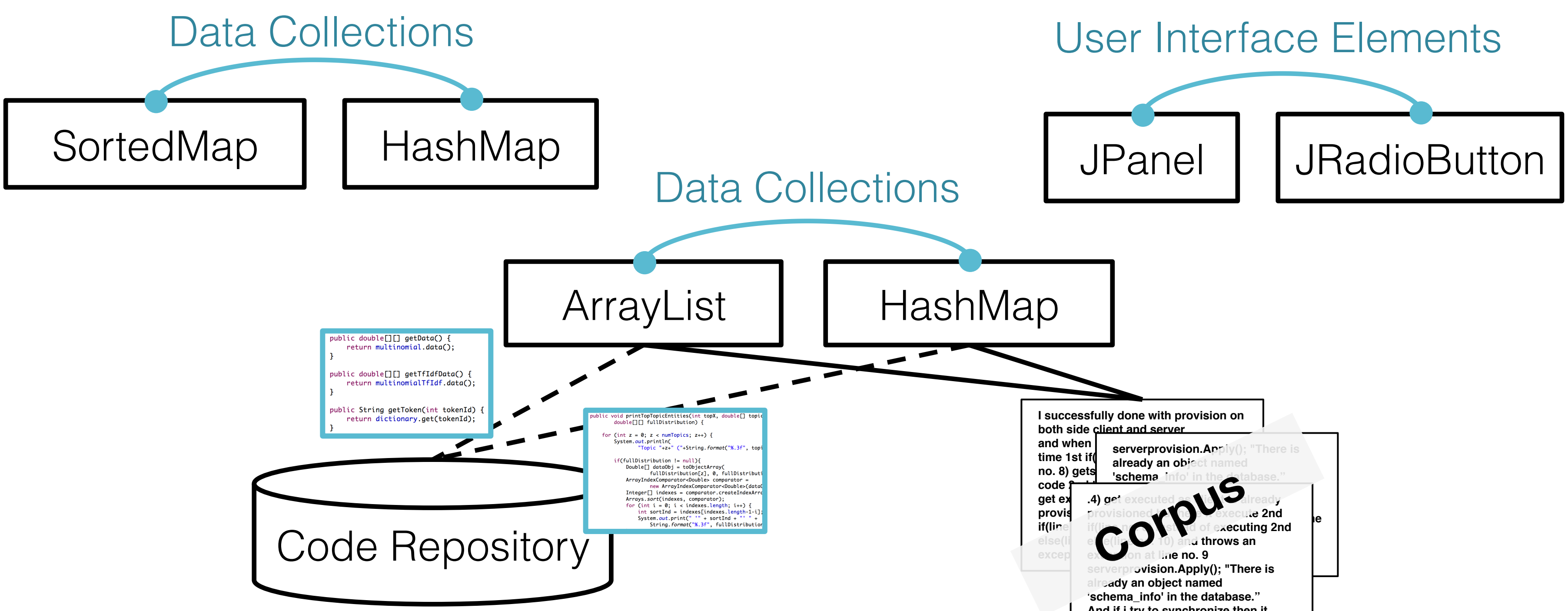
Code Repository

81 implementations of Node

```
javax.xml.soap.Node
javax.imageio.spi.DigraphNode
javax.swing.tree.TreeNode
javax.swing.tree.MutableTreeNode
com.sun.org.apache.xerces.internal.dom.NodeImpl
```

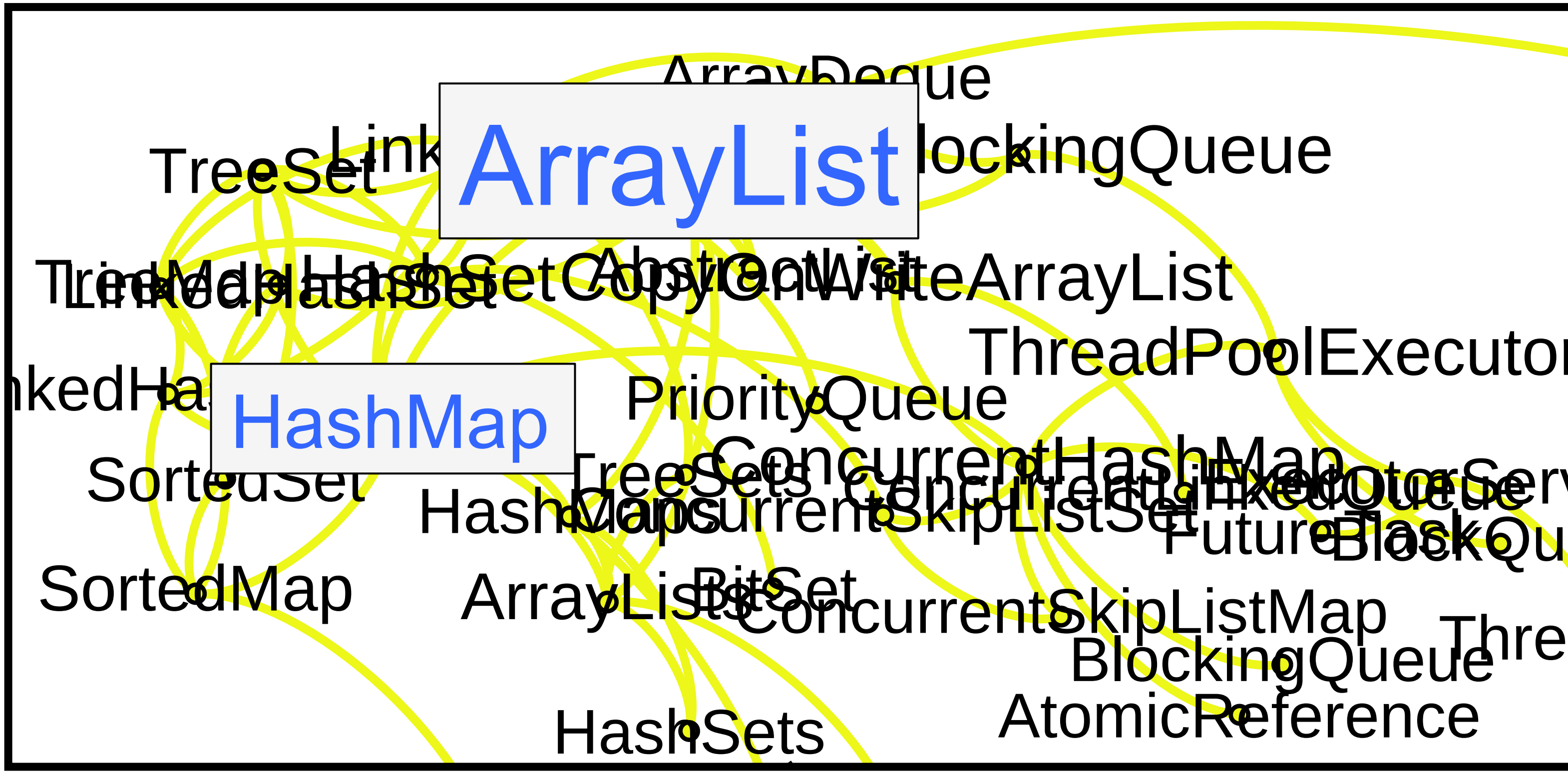
Grounded Software Ontology Construction

- We detect coordinate relations (similarity) between Java classes
  - “This method iterates over ArrayLists **and** HashMaps”



[Movshovitz-Attias & Cohen, 15]

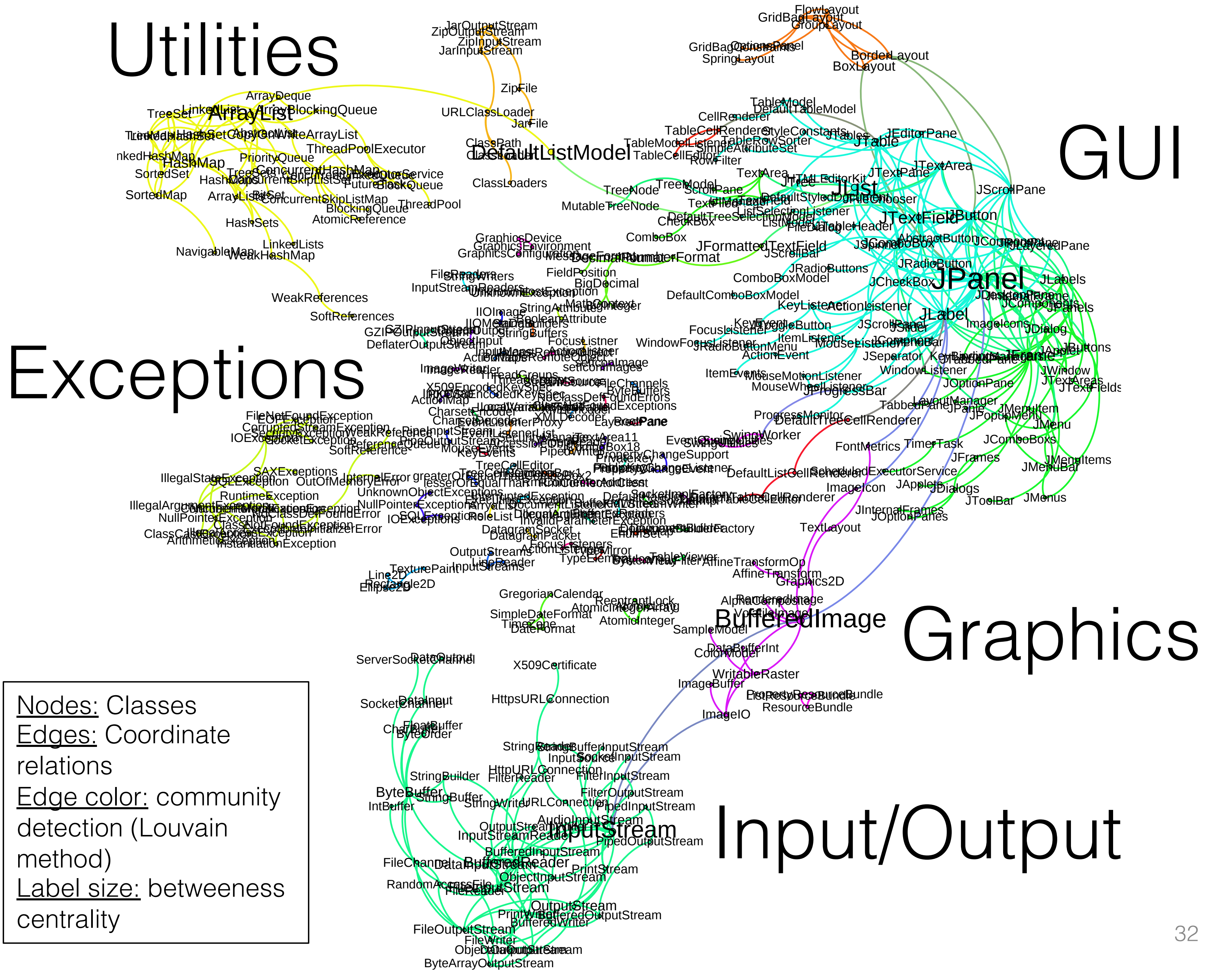
Utility Classes



- TreeSet
- SortedMap
- SortedSet
- HashSet
- BitSet

- ArrayDeque
- PriorityQueue
- BlockingQueue
- ArrayBlockingQueue
- LinkedHashMap

<u>Nodes:</u> Classes
<u>Edges:</u> Coordinate relations
<u>Edge color:</u> community detection (Louvain method)
<u>Label size:</u> betweeness centrality



Nodes: Classes
Edges: Coordinate relations
Edge color: community detection (Louvain method)
Label size: betweeness centrality

Lessons

Contributions & Lessons

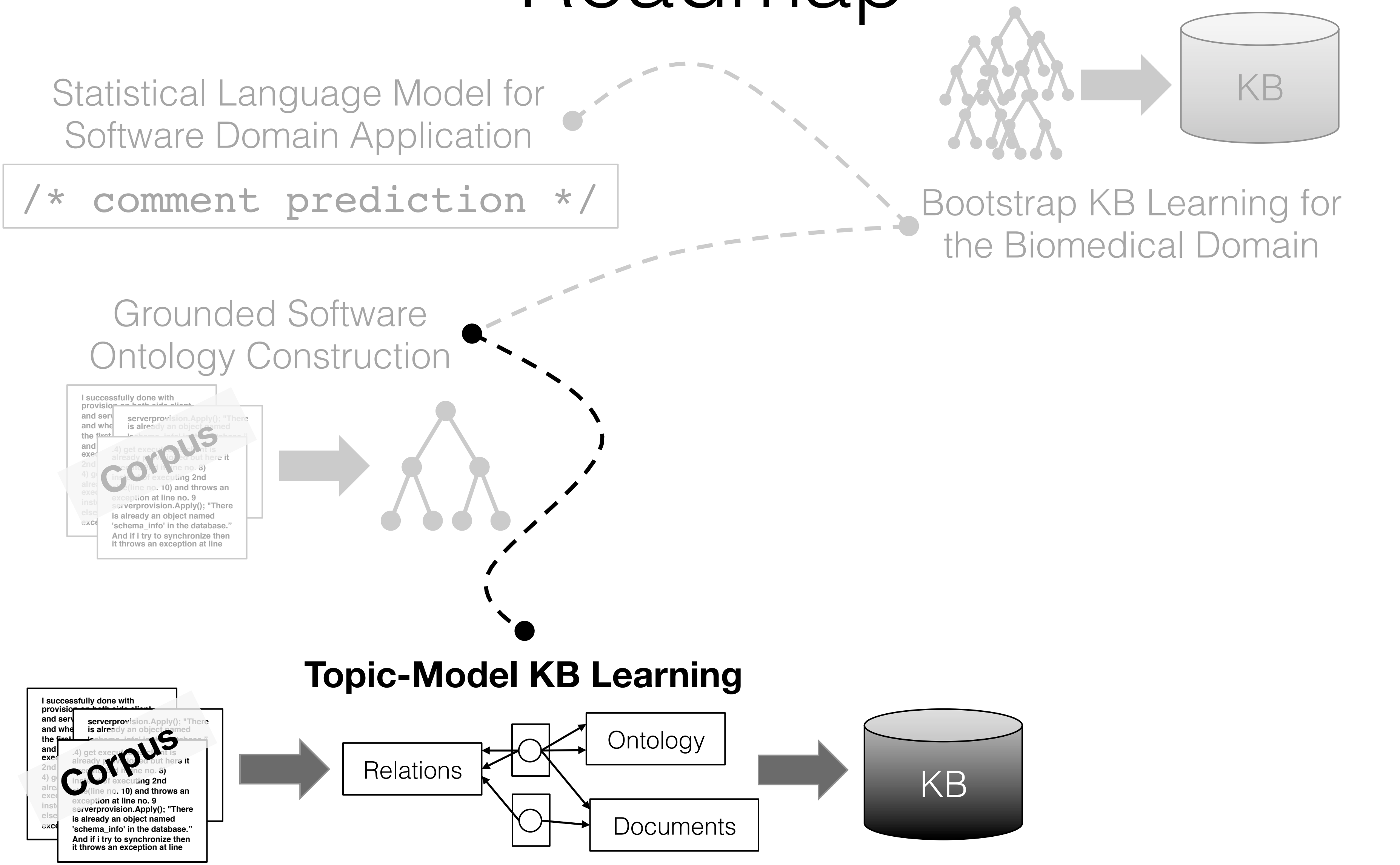
- Linked class entities to code implementation
- Defined distributional similarity for code
- By combining code and text similarities we learned relations (and ontology)

- Advantage: This ontology reflects statistics in language and code
  - In contrast to manually built ontologies

- Grounding to code limits scope of learned ontology to code entities
  - What's missing?

Users
Computer resources
Design patterns

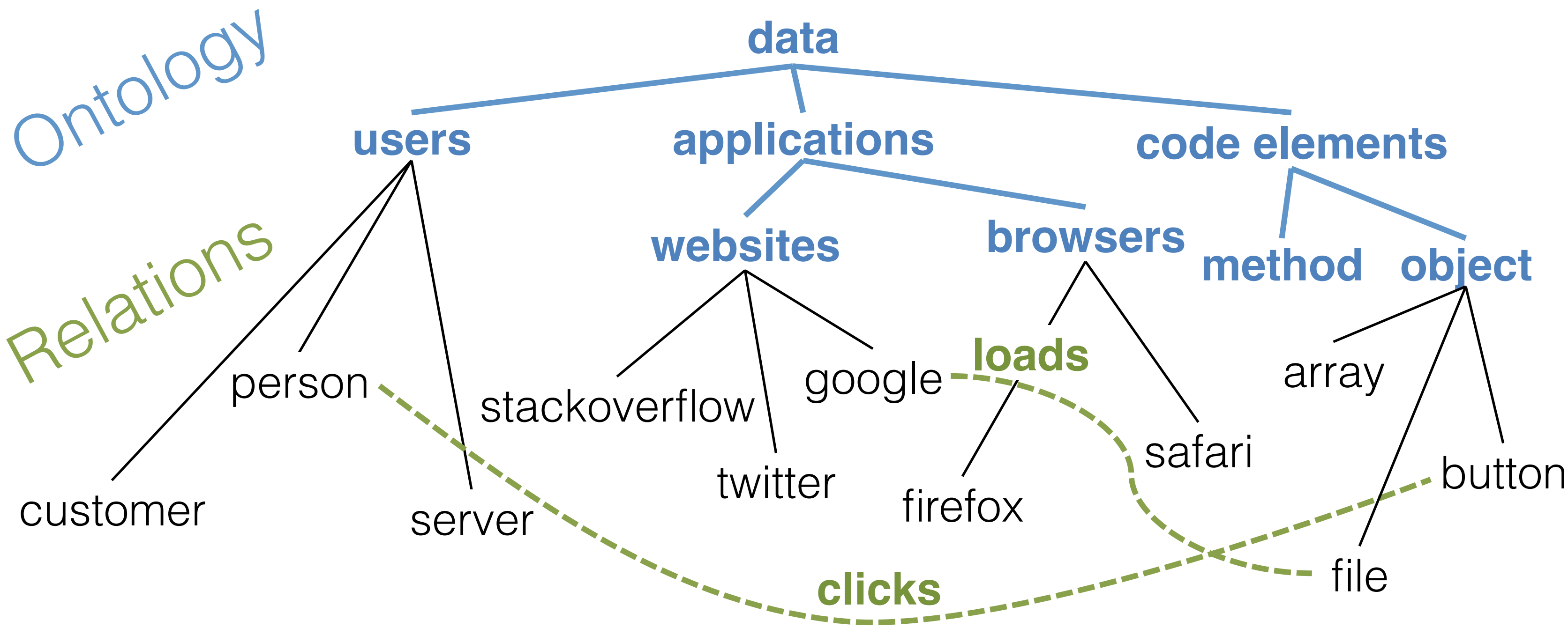
Roadmap



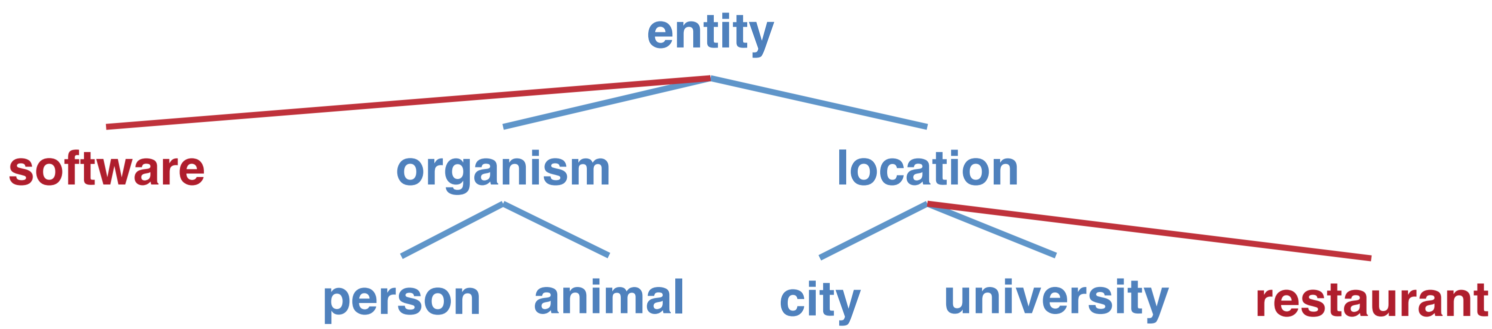
Statistical Language Model for
Software Domain Application
/* comment prediction */
KB
Bootstrap KB Learning for
the Biomedical Domain
Grounded Software
Ontology Construction
Corpus
Topic-Model KB Learning
Corpus
Relations
Ontology
Documents
KB

Goal: Corpus-Driven Knowledge Base

- Schema and facts are drawn from corpus
- Unsupervised: learn optimal latent corpus structure together with best-matching facts



Reasoning with Ontologies



- ✔ Ontologies give information context
- ✔ Easy to extract domain-specific information

- Ontologies are
  - expensive
  - require prior knowledge
- Manual ontology does not reflect language statistics
- Ontology and facts are often drawn from different sources

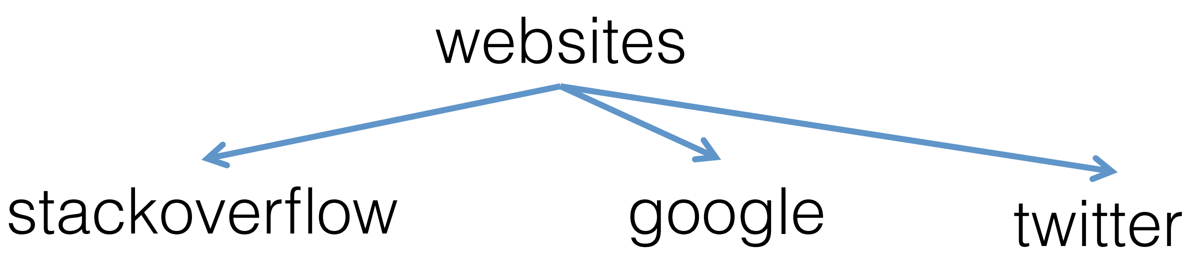
Pattern-based Relation Extraction

1. Hypernym-hyponym

"**websites** such as **stackoverflow**"

"**websites** including **google** and **twitter**"

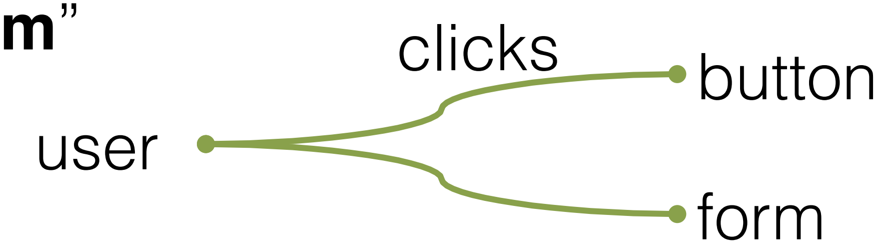
Y such as **X**
X is a **Y**
Y including **X**



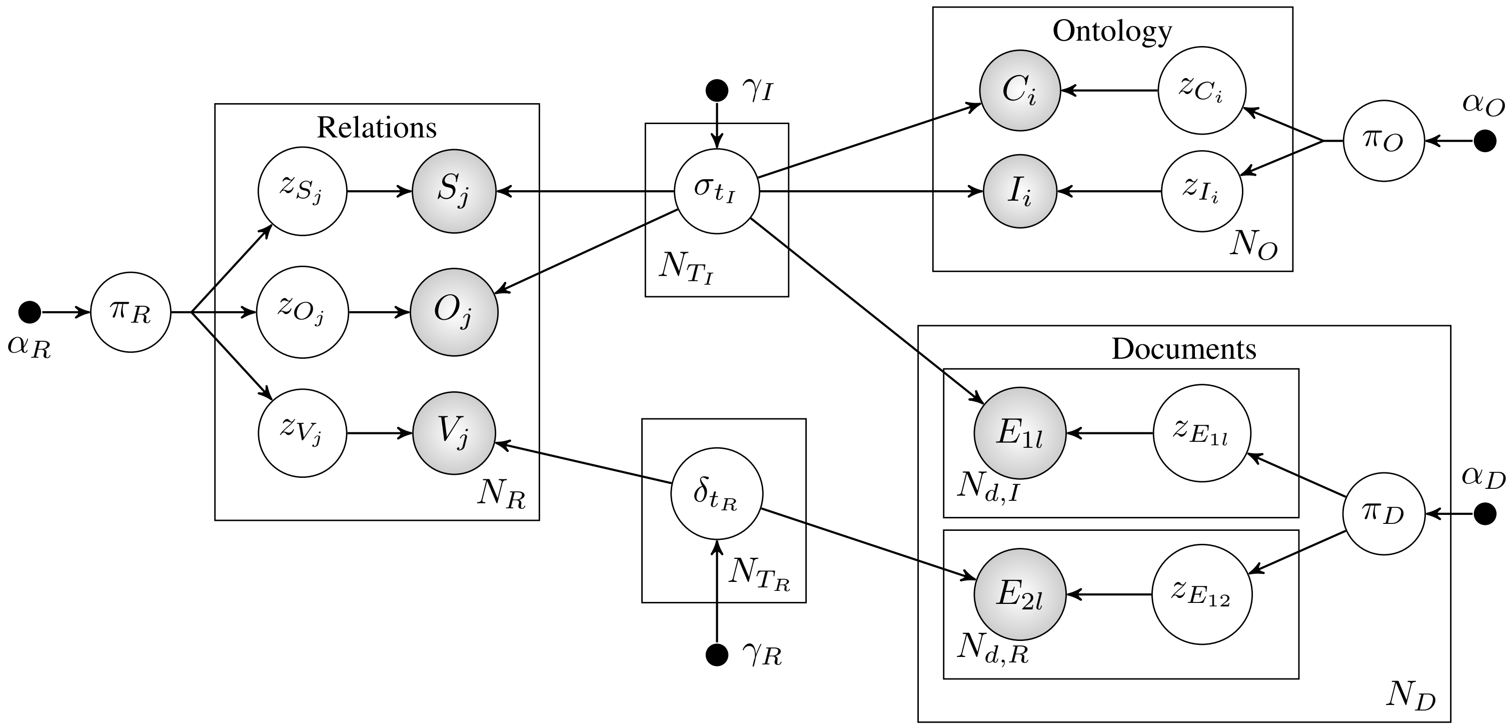
2. Subject-Verb-Object

"**user** clicks **button**"

"**user** clicks **form**"

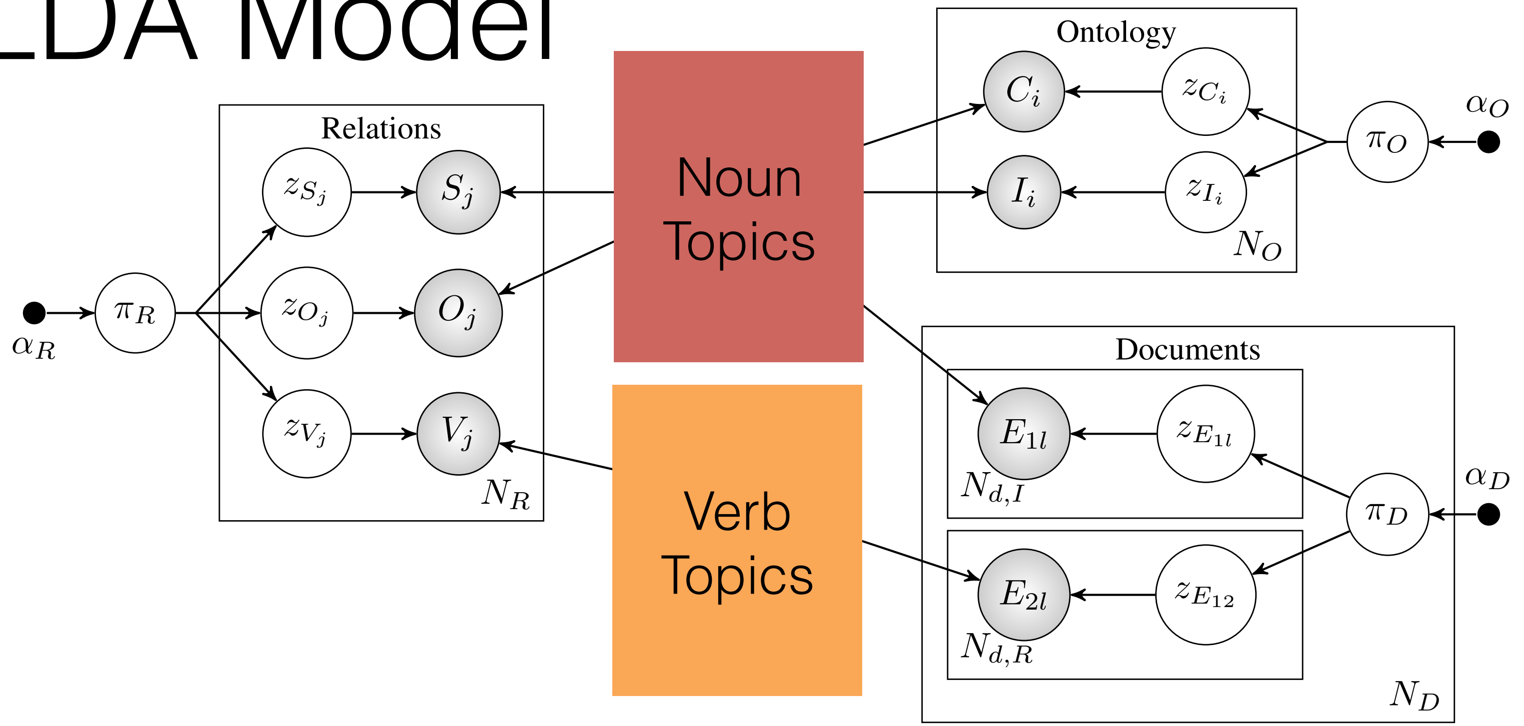


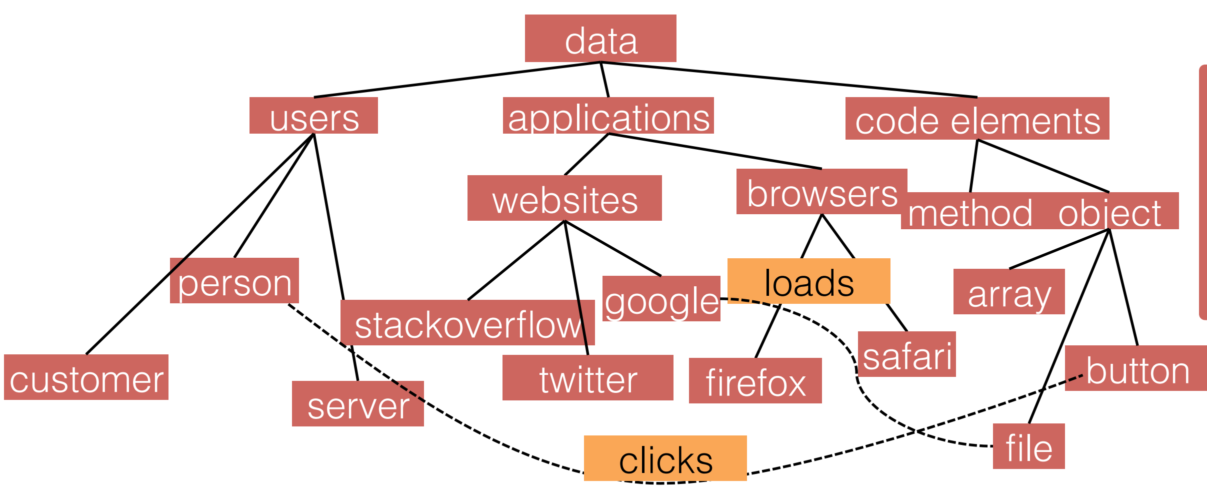
KB-LDA Model



Relations
Ontology
Documents
z_{S_j}
S_j
z_{O_j}
O_j
z_{V_j}
V_j
N_R
π_R
α_R
γ_I
σ_{t_I}
N_{T_I}
δ_{t_R}
N_{T_R}
γ_R
C_i
z_{C_i}
I_i
z_{I_i}
N_O
π_O
α_O
E_{1l}
$z_{E_{1l}}$
$N_{d,I}$
E_{2l}
$z_{E_{12}}$
$N_{d,R}$
N_D
π_D
α_D

KB-LDA Model

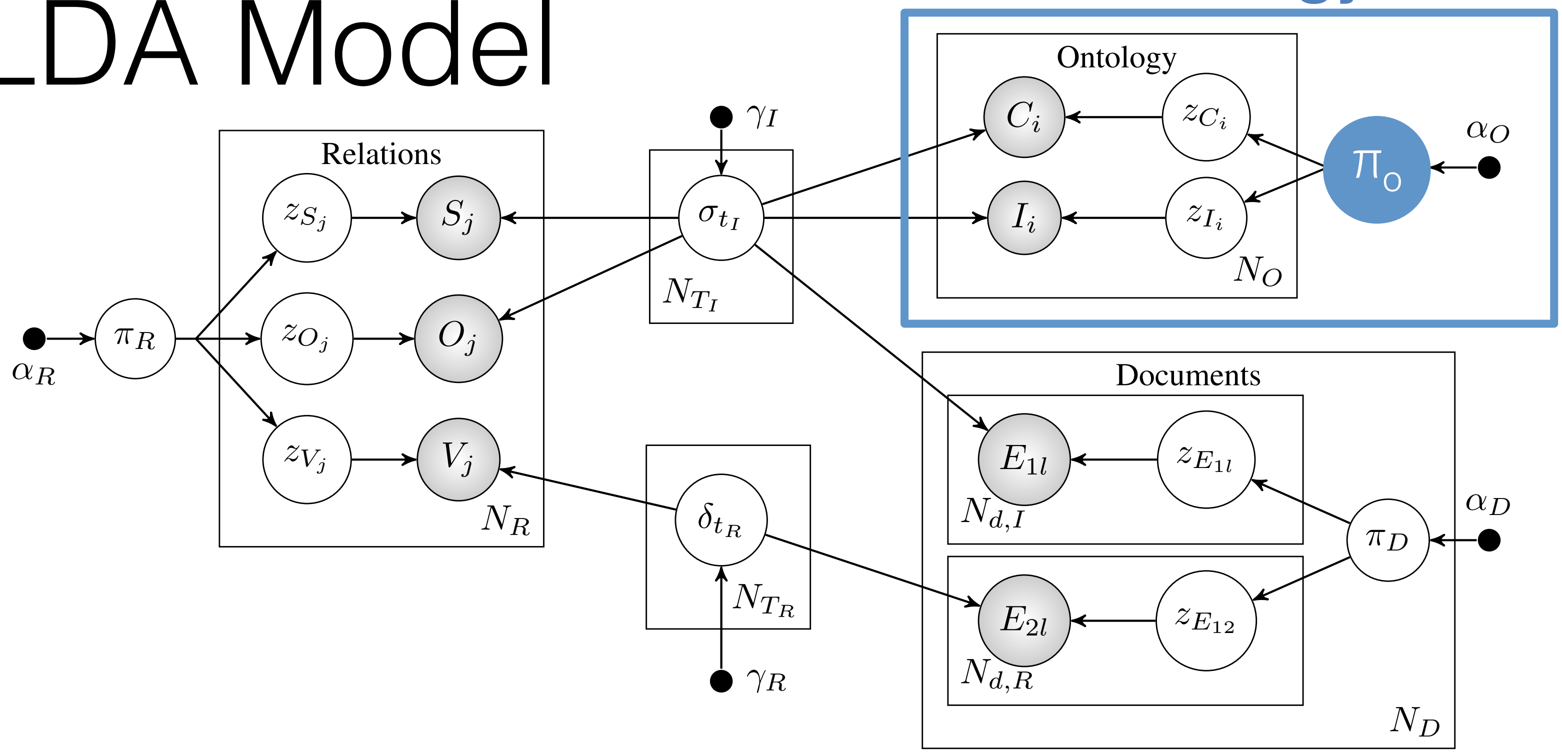


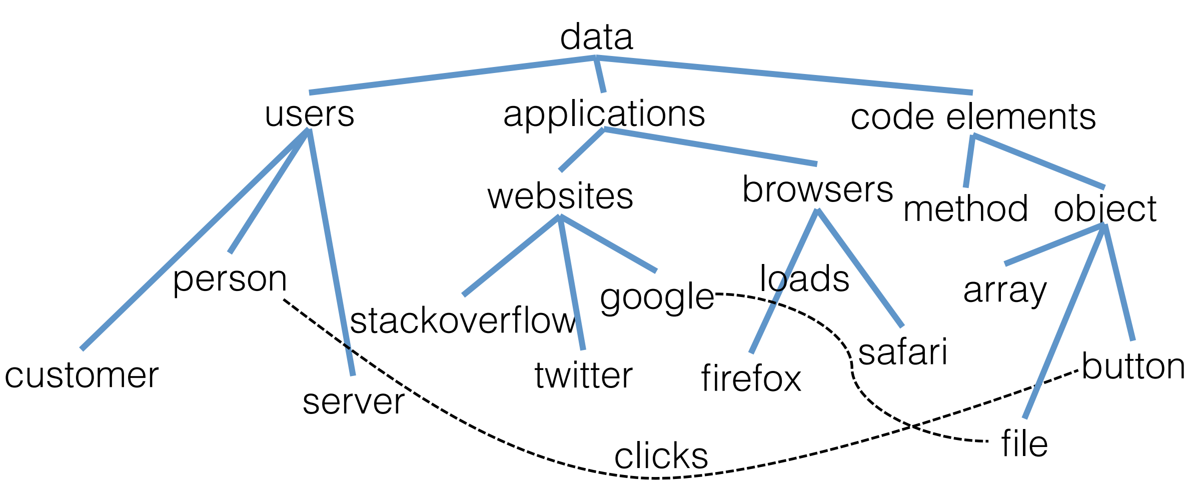


Categories	Typed Relations
stackoverflow	clicks
google	selects
twitter	hits

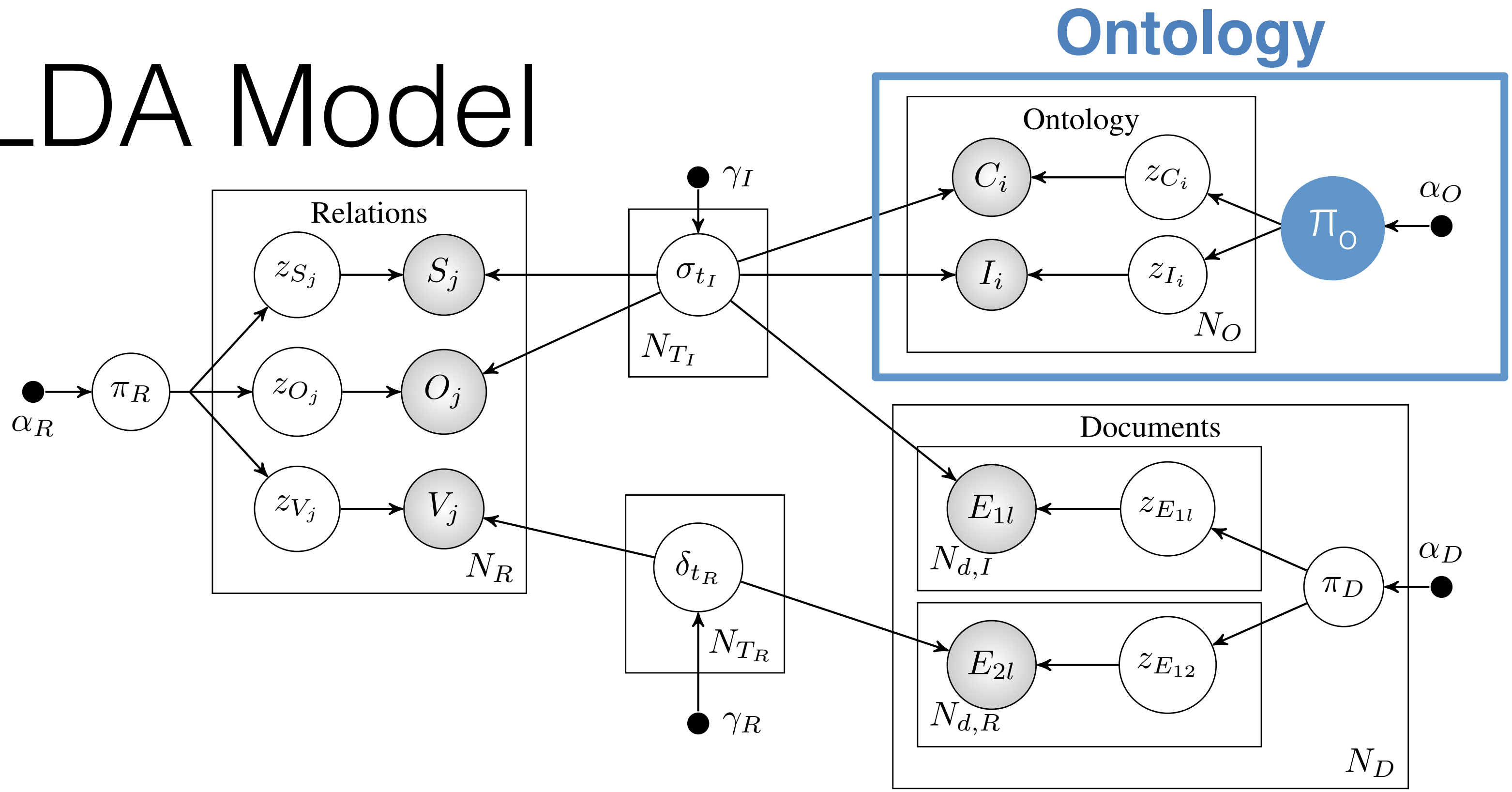
KB-LDA Model

Ontology





KB-LDA Model

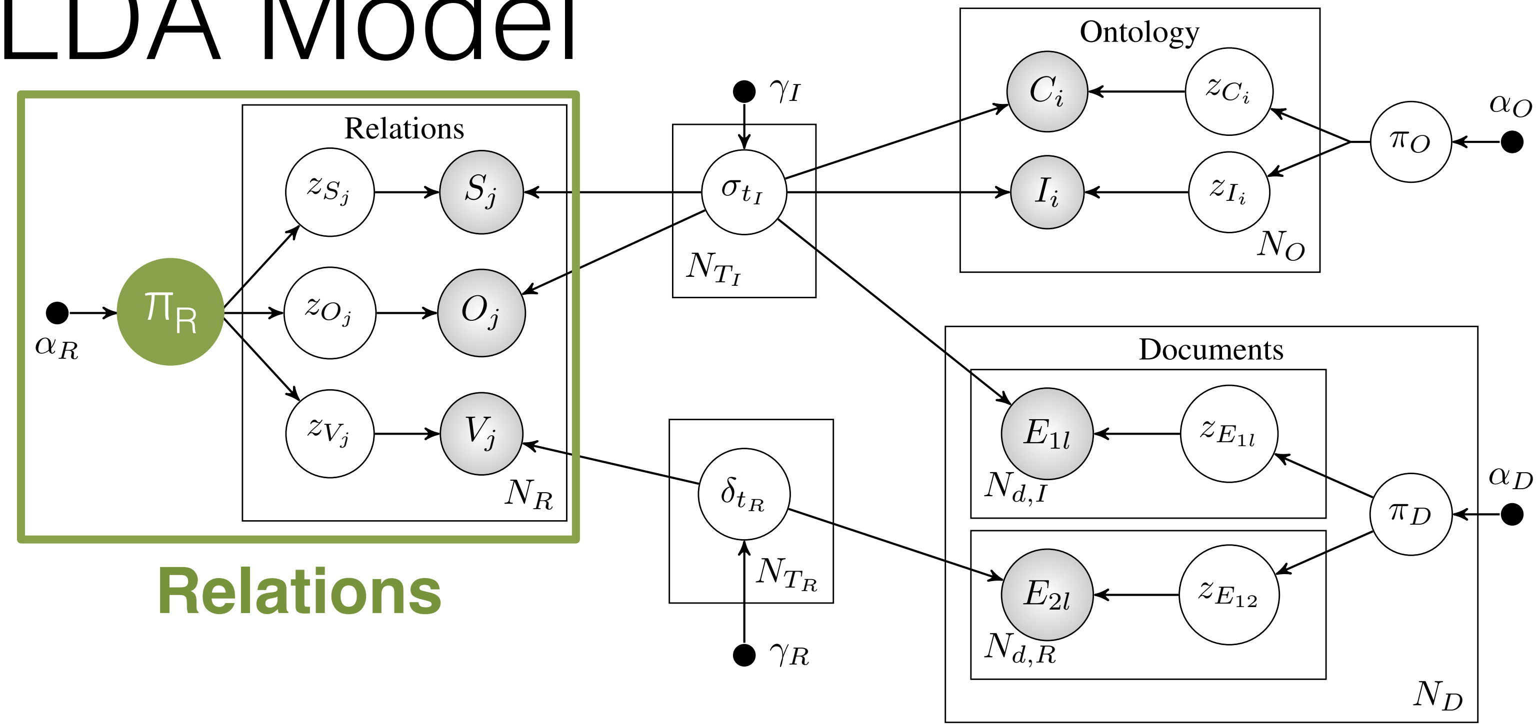


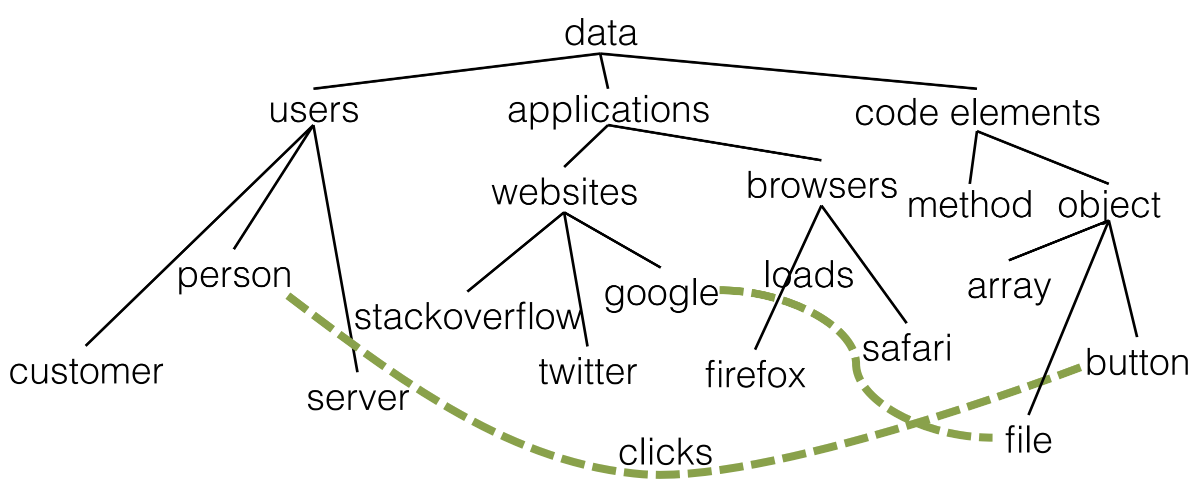
Hypernym-hyponym relations:

websites → google
platforms → stackoverflow

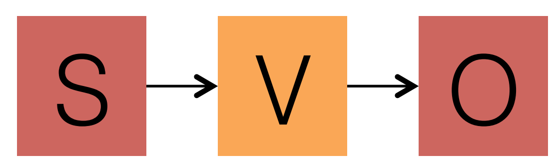
websites platforms applications	Noun Topic 1
↓	
stackoverflow google twitter	Noun Topic 2

KB-LDA Model

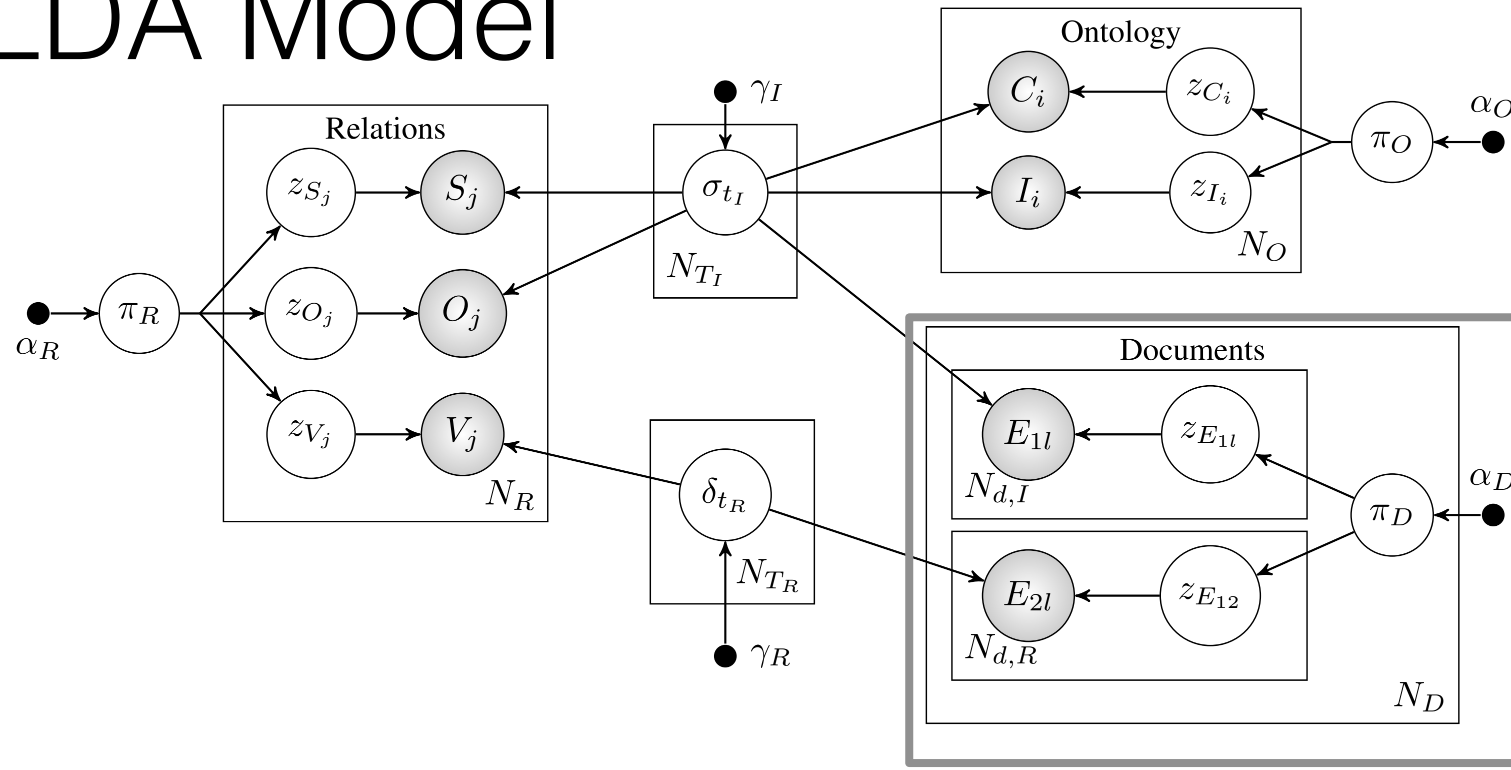


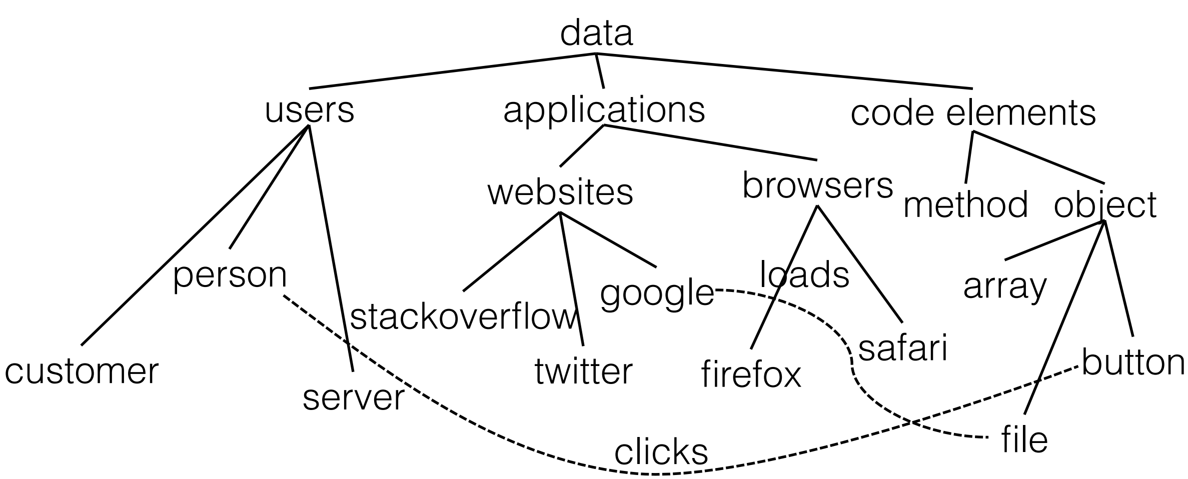


Extracted SVO relation:
person, *clicks*, button



KB-LDA Model

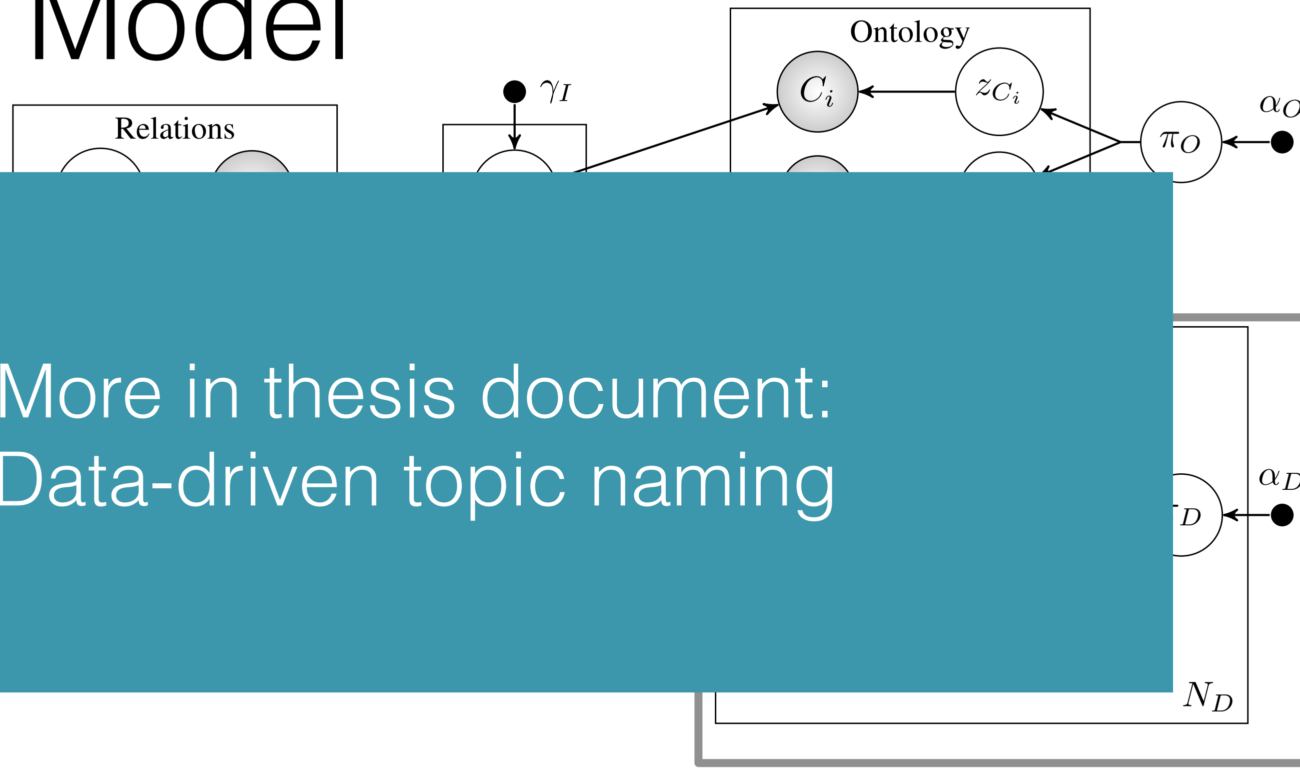




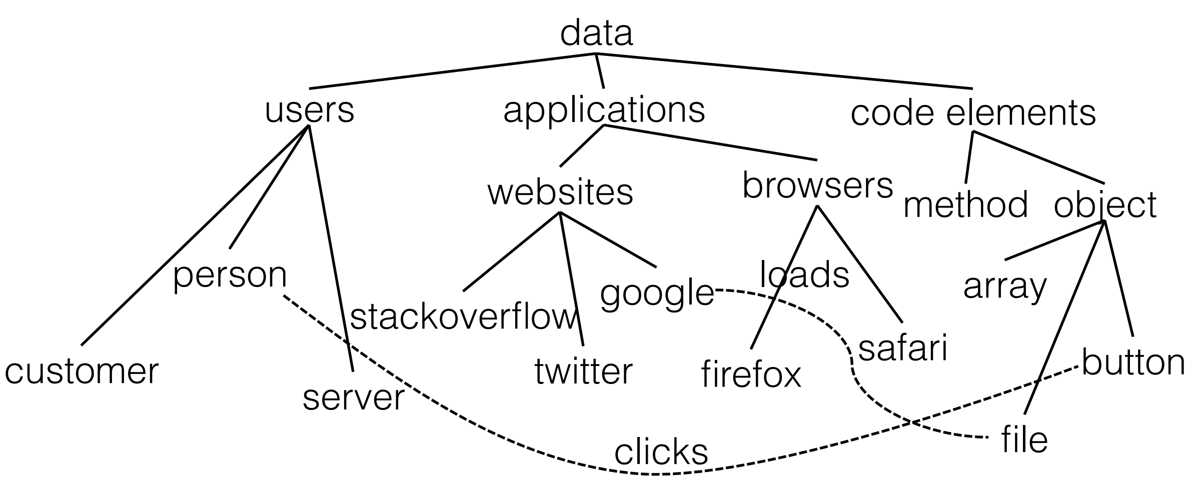
Documents

Downloads on websites sometimes have an **MD5** checksum, **allowing people** to confirm the integrity of the file. I have heard this is to allow not only corrupted files to be instantly identified before they cause a problem but also for for any **malicious changes** to be easily **detected**.

KB-LDA Model

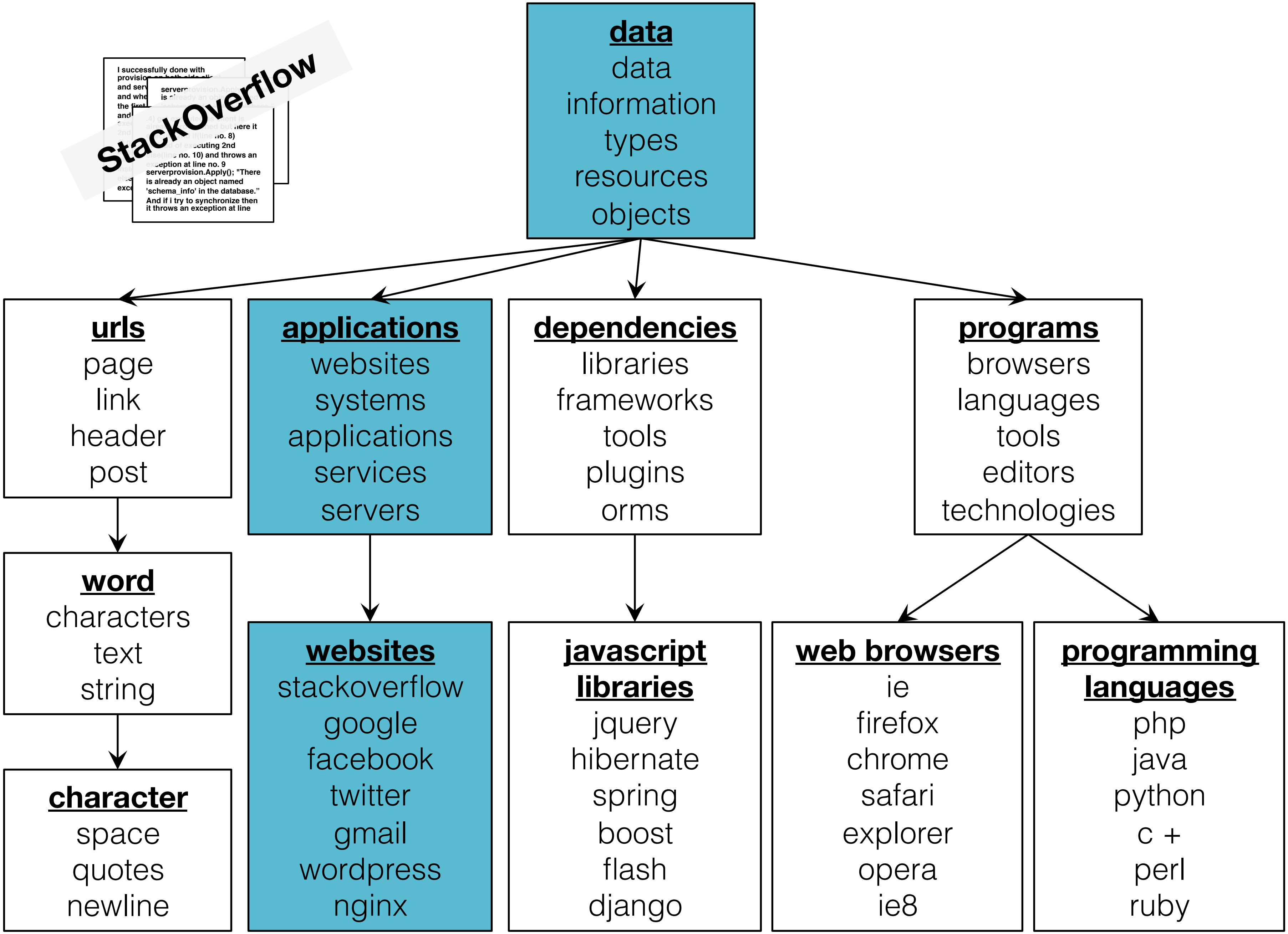


More in thesis document:
Data-driven topic naming



Documents

Downloads on websites sometimes have an **MD5** checksum, **allowing people** to confirm the integrity of the file. I have heard this is to allow not only corrupted files to be instantly identified before they cause a problem but also for for any **malicious changes** to be easily **detected**.


StackOverflow
data
data
information
types
resources
objects
urls
page
link
header
post
applications
websites
systems
applications
services
servers
dependencies
libraries
frameworks
tools
plugins
orms
programs
browsers
languages
tools
editors
technologies
word
characters
text
string
character
space
quotes
newline
websites
stackoverflow
google
facebook
twitter
gmail
wordpress
nginx
javascript libraries
jquery
hibernate
spring
boost
flash
django
web browsers
ie
firefox
chrome
safari
explorer
opera
ie8
programming languages
php
java
python
c +
perl
ruby

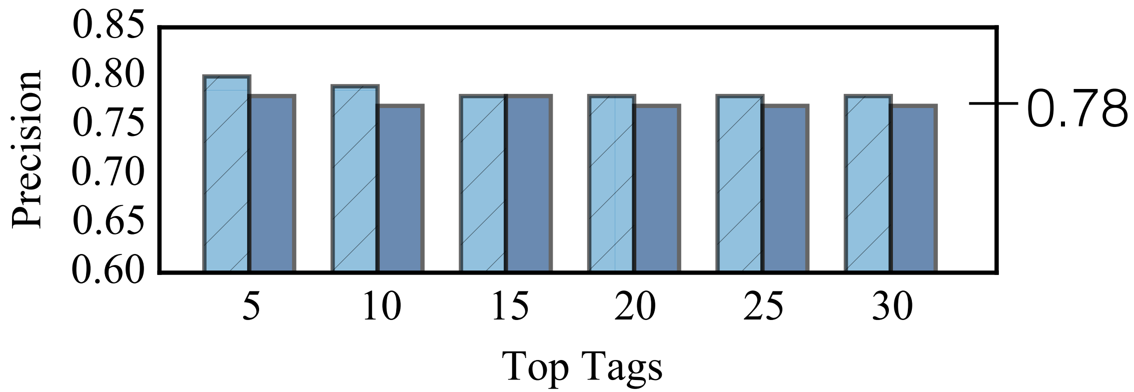
Subject Topic	user people customer client person	function method class object query	code server compiler client program	user people customer client person	user people customer client person
Verb Topic	clicks selects submits hits changes	has returns takes implements supports	makes throws sends gives receives	uploads opens closes visits leaves	enters defined changes add edit
Object Topic	button form link item file	value place property interface pointer	error exception request message warning	app page application file site	name images id password address

M-Turk Evaluation of Noun Topics

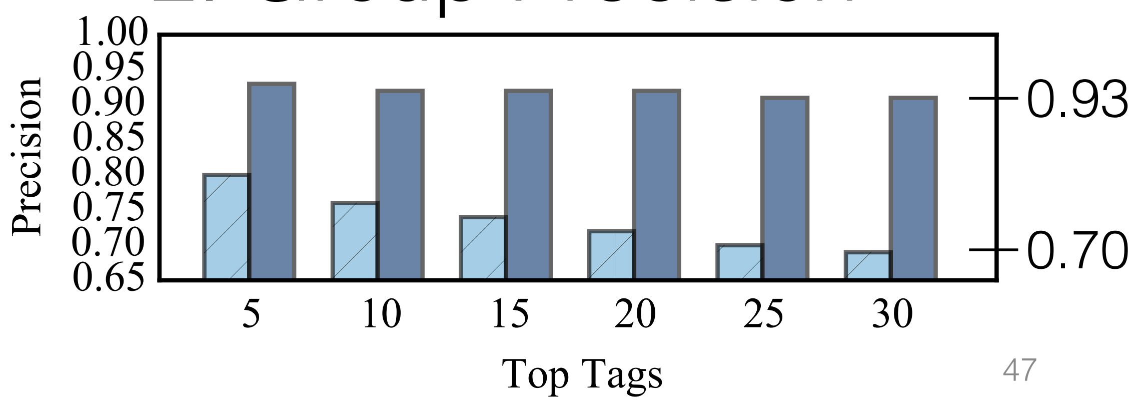
"Which words are ***not*** related to **programming languages**?"

java
python
javascript
firefox
ruby
perl

1. Word Intrusion



2. Group Precision



Domain-Specific Extraction from Open IE

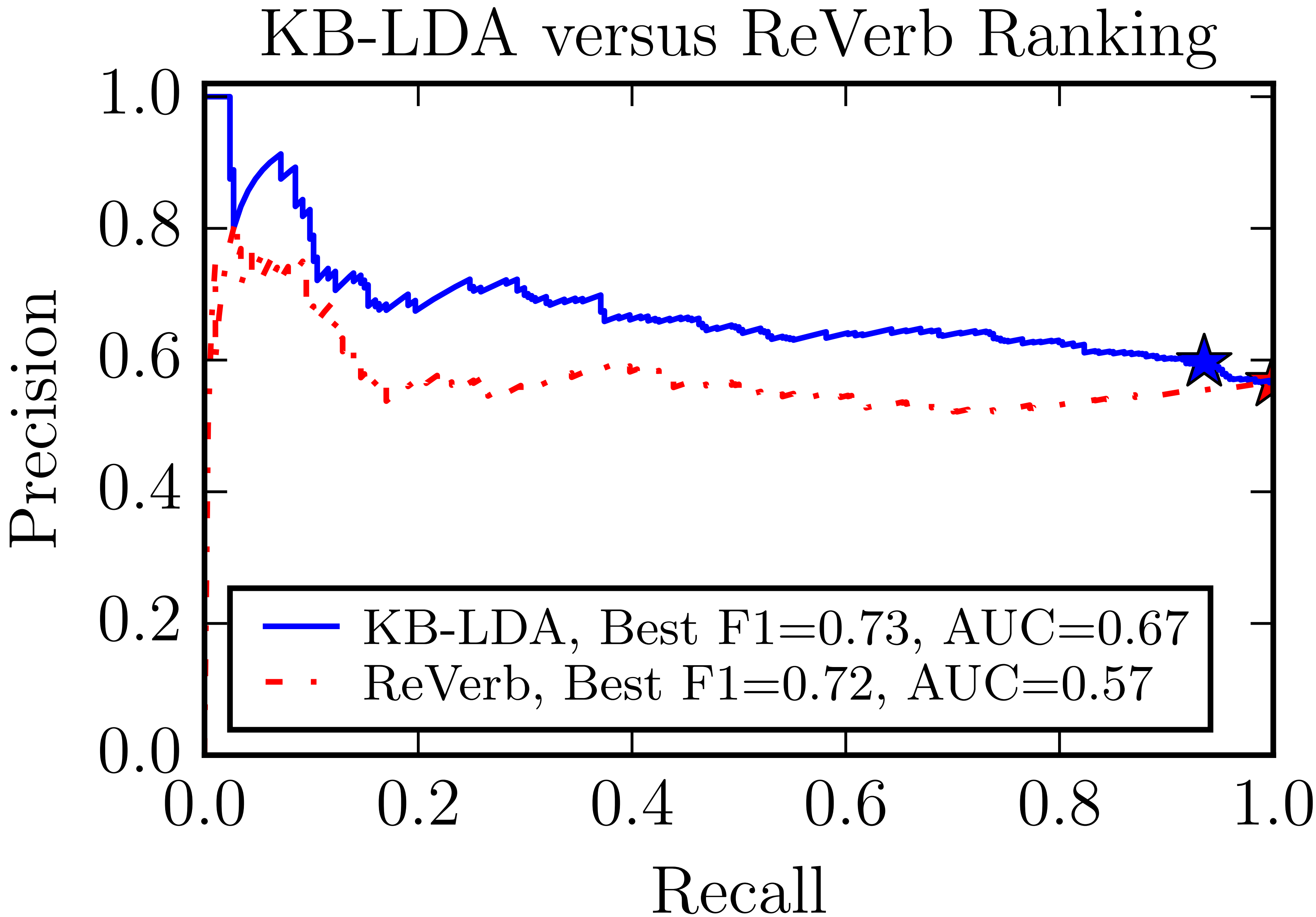
ReVerb
15m SVO triples

Triples with software entities
5k triples

SVOs extracted directly from StackOverflow
37k triples

✔ (safari, *supports*, svg)
✔ (computer, *is running*, xp)

✘ (people, *can read*, italian)
✘ (view, *looks*, south)

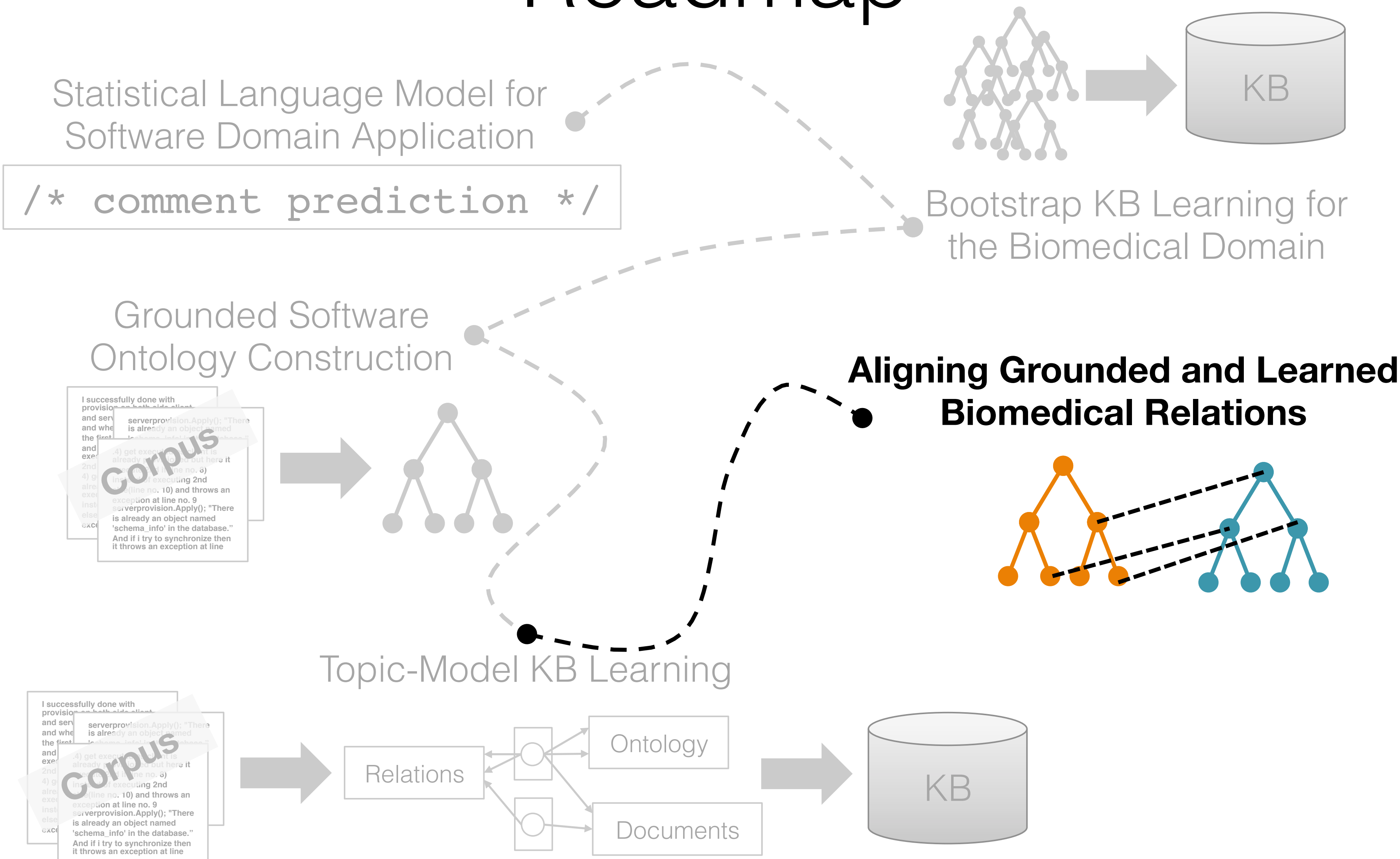

KB-LDA versus ReVerb Ranking
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
KB-LDA, Best F1=0.73, AUC=0.67
ReVerb, Best F1=0.72, AUC=0.57

Lessons

KB-LDA Lessons

- Corpus-driven KB construction: Jointly optimizes schema and facts
- Unsupervised: Useful for exploration of new domains (Software)
- Can pre-existing domain knowledge improve KB learning? How much? (Biomedical ontologies)

Roadmap



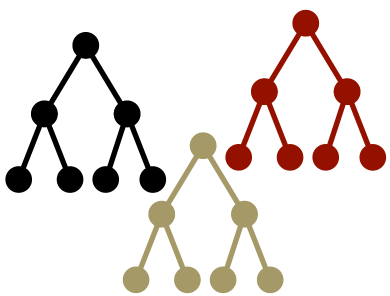
Statistical Language Model for
Software Domain Application
/* comment prediction */
KB
Bootstrap KB Learning for
the Biomedical Domain
Grounded Software
Ontology Construction
Corpus
Aligning Grounded and Learned
Biomedical Relations
Topic-Model KB Learning
Corpus
Relations
Ontology
Documents
KB

Aligning Grounded and Learned Relations

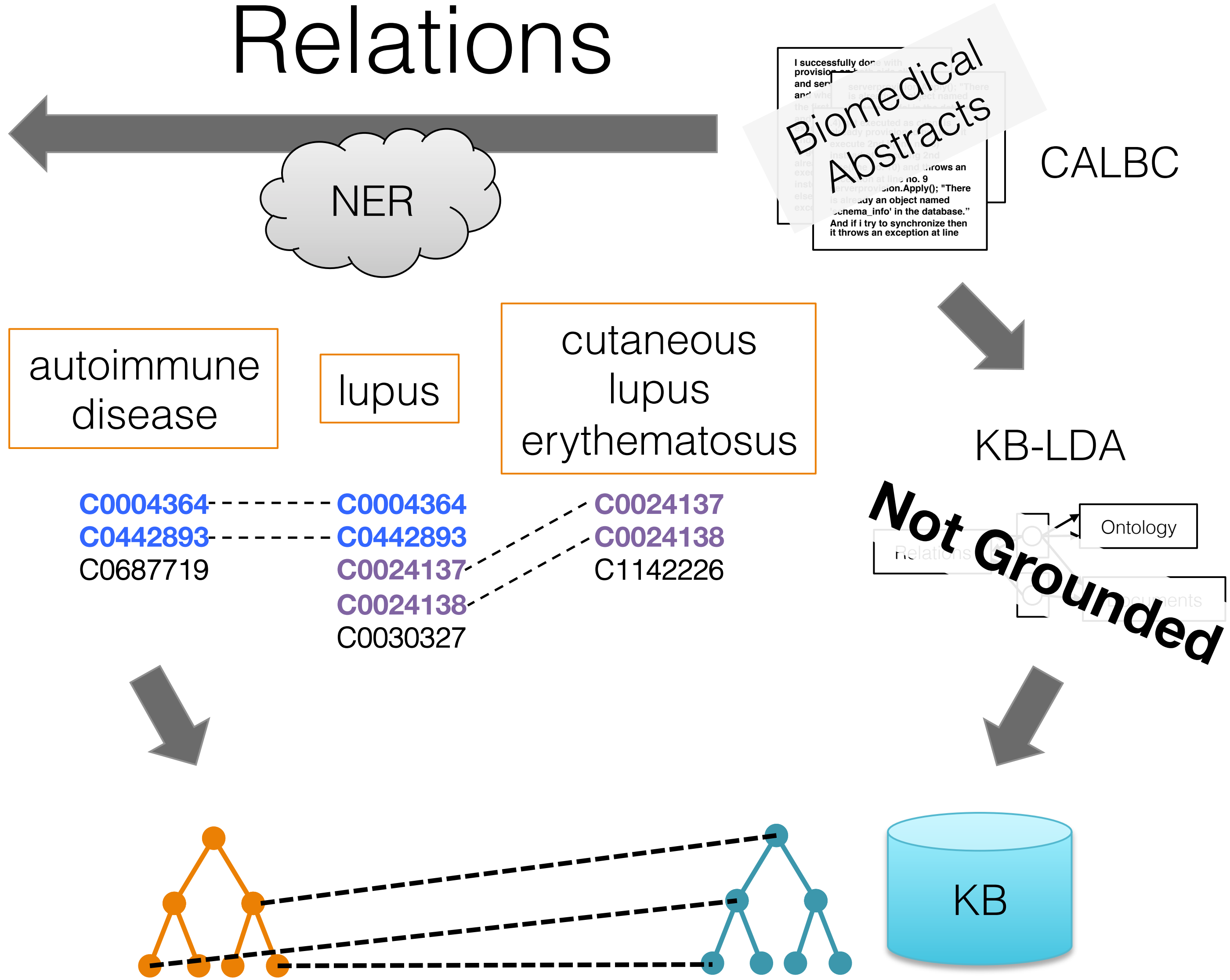
- Evaluation of KB-LDA relations compared to known relations
- Investigation of the potential of grounding

Aligning Grounded and Learned Relations

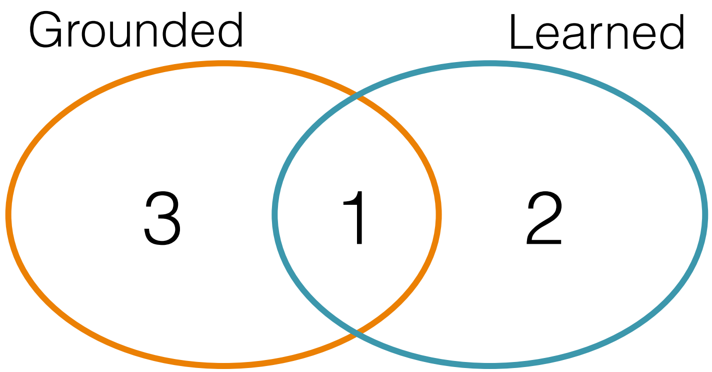
Grounded



- Proteins and Genes (PRGE)
- Chemicals (CHED)
- Diseases and disorders (DISO)
- Living Beings (LIVB)
- Unified Medical Language System (UMLS)
- …



Possible Alignment Outcomes



1. Does KB-LDA learn existing relations and concepts?
   – Model Validation

2. Does KB-LDA discover “new” relations and concepts?
   – Added value of language statistics

3. What is missing?
   – Can be added through grounding

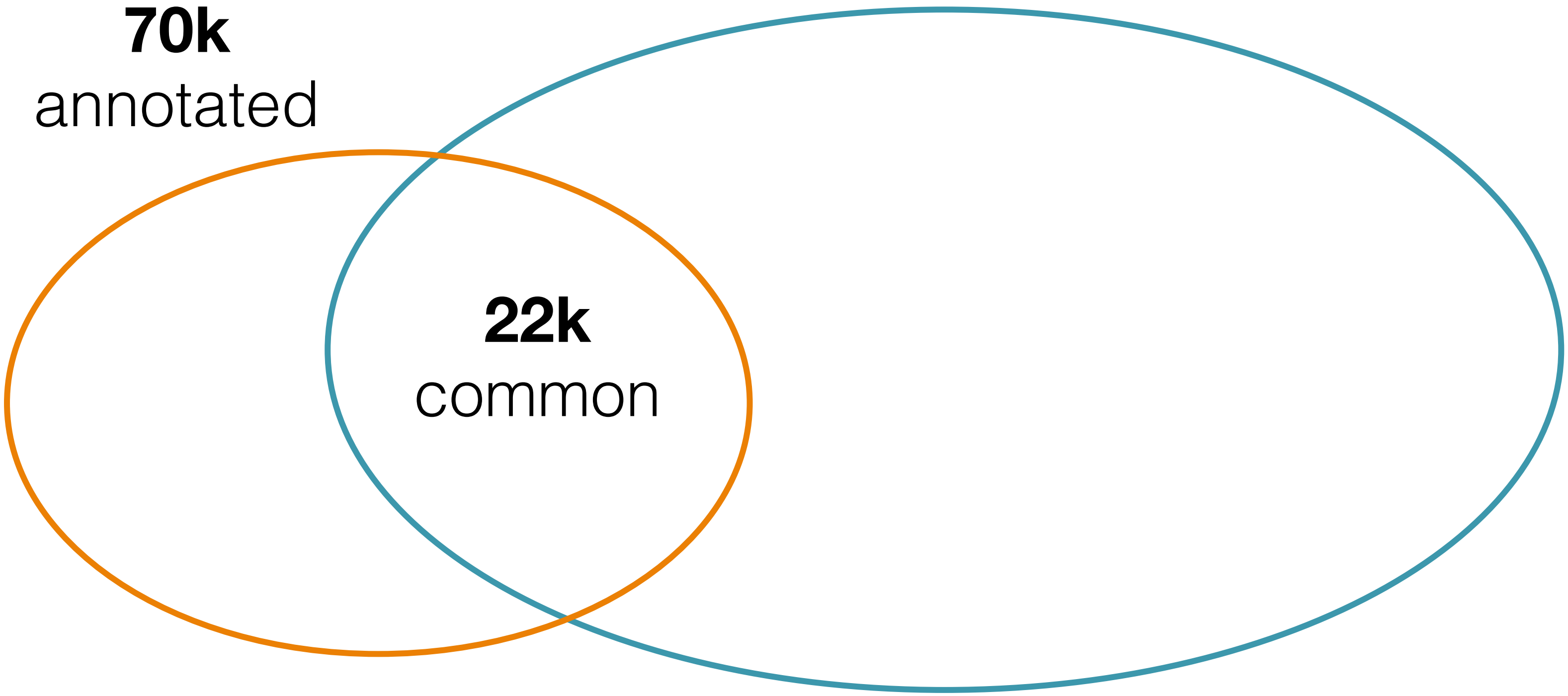
Grounded and Learned Entities

Grounded

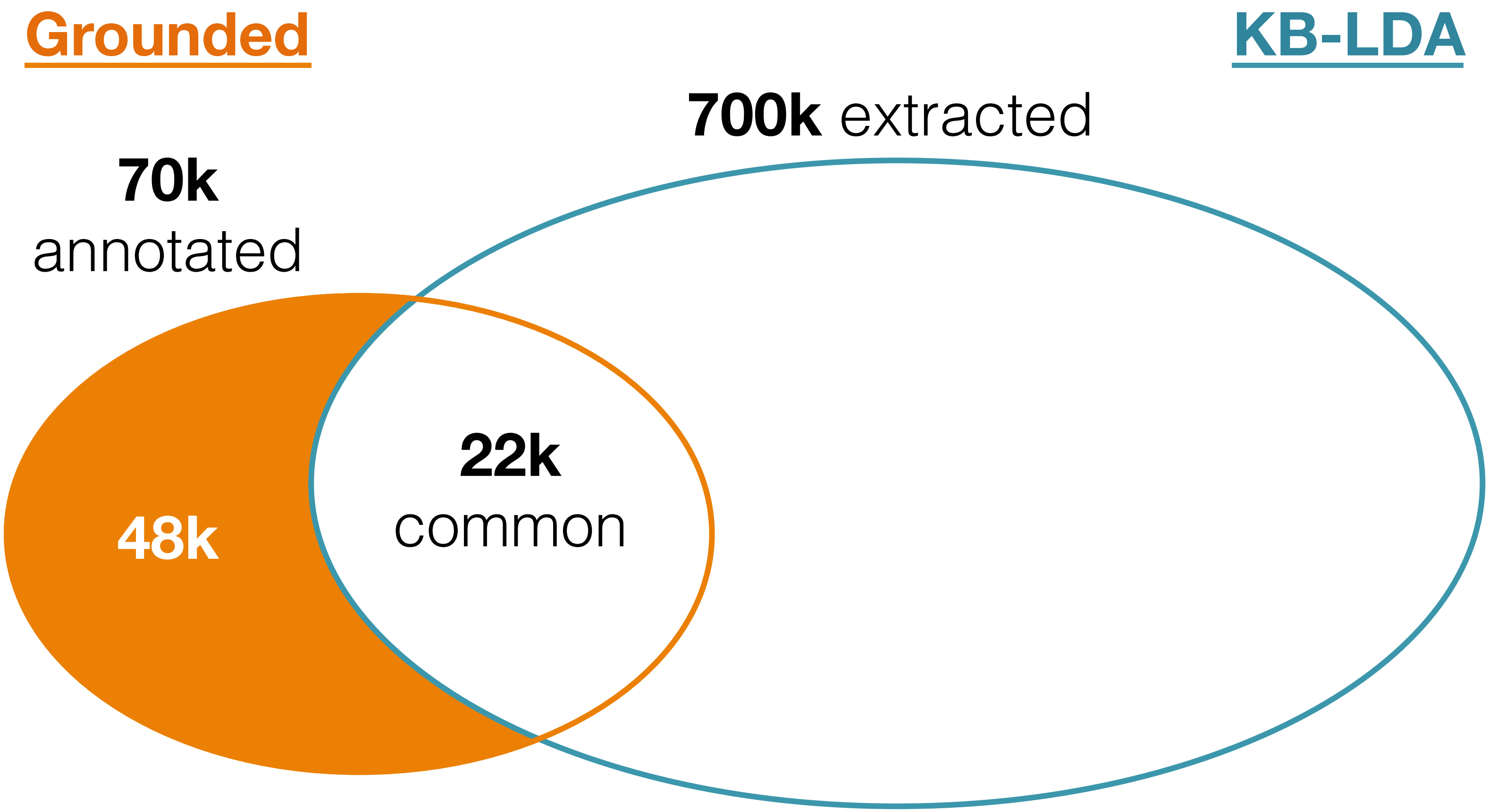
KB-LDA

700k extracted

70k
annotated



Grounded and Learned Entities



Grounded Entities: Manual evaluation of sample

Partial or incorrect parse
35/100

- ✘ bronchial asthma
- ✘ pneumoniae
- ✘ beta-1,2-mannotriose

Average corpus frequency
4.87

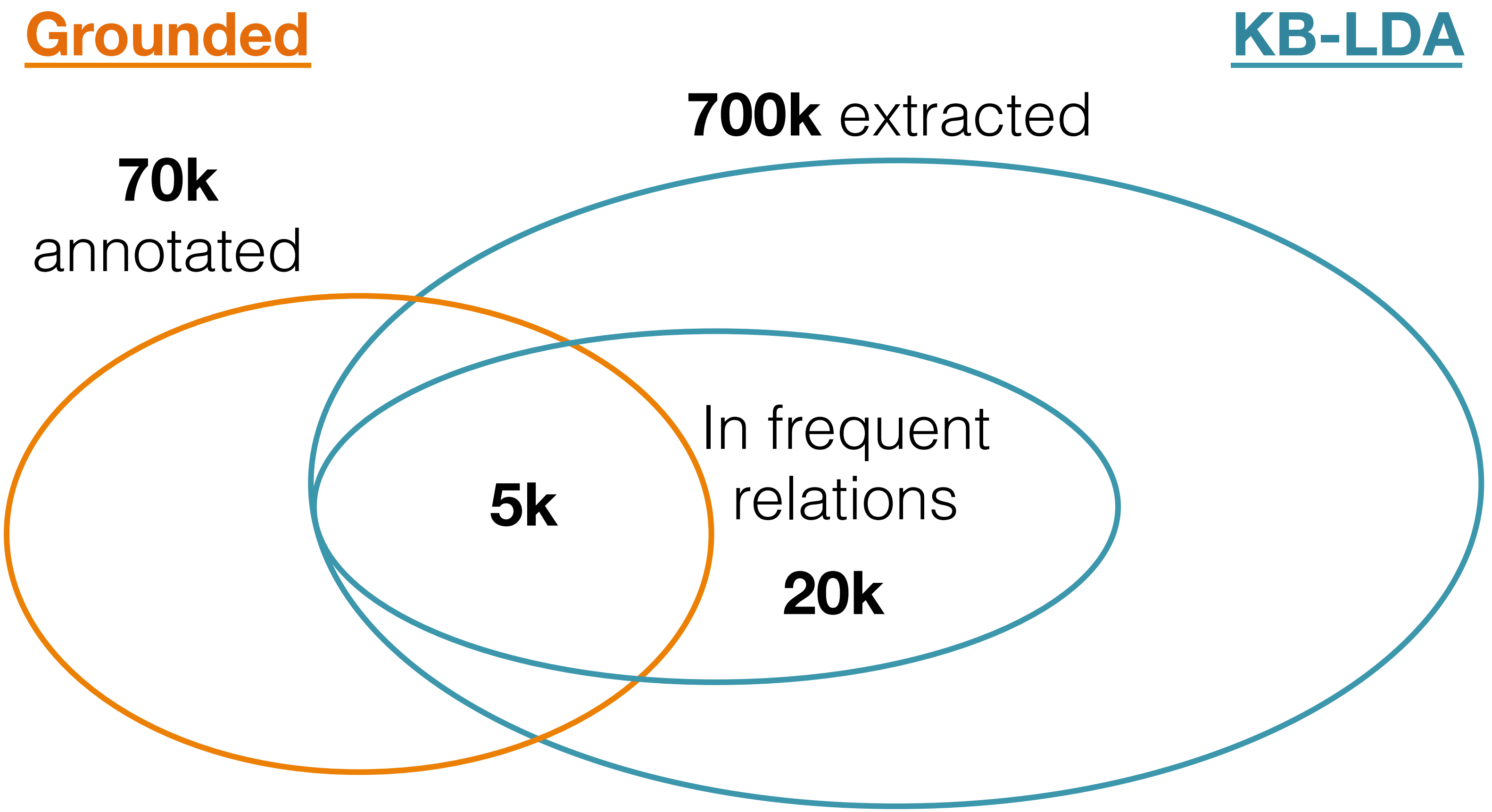
Only in incorrect form
20/35

Frequent entities (f>10)
12

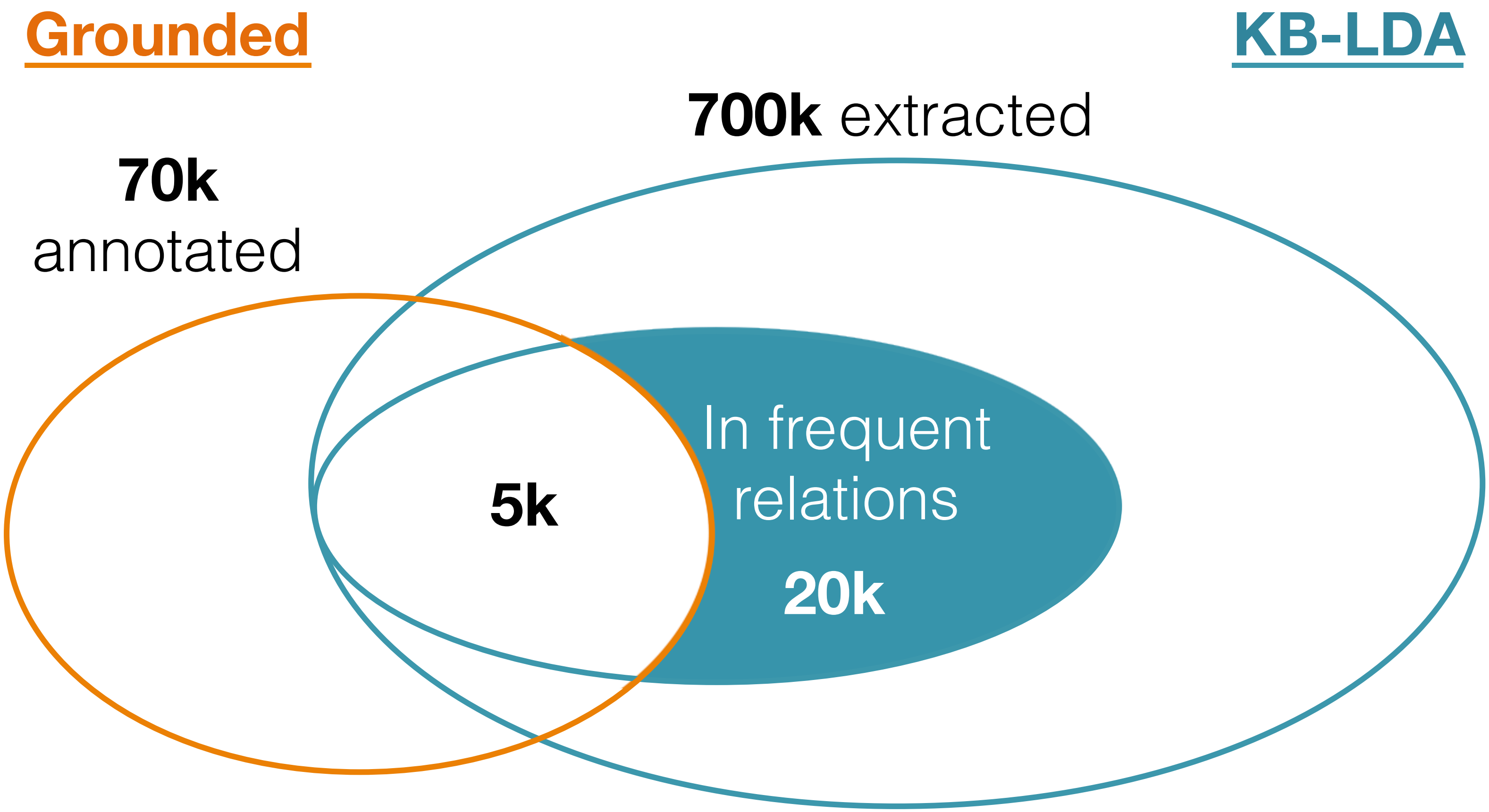
Limitation of grounding

Potential advantage of grounding

Grounded and Learned Entities



Grounded and Learned Entities



KB-LDA Entities: Manual evaluation of sample

Correct
97/100

Added value of language

Parse errors
3/100

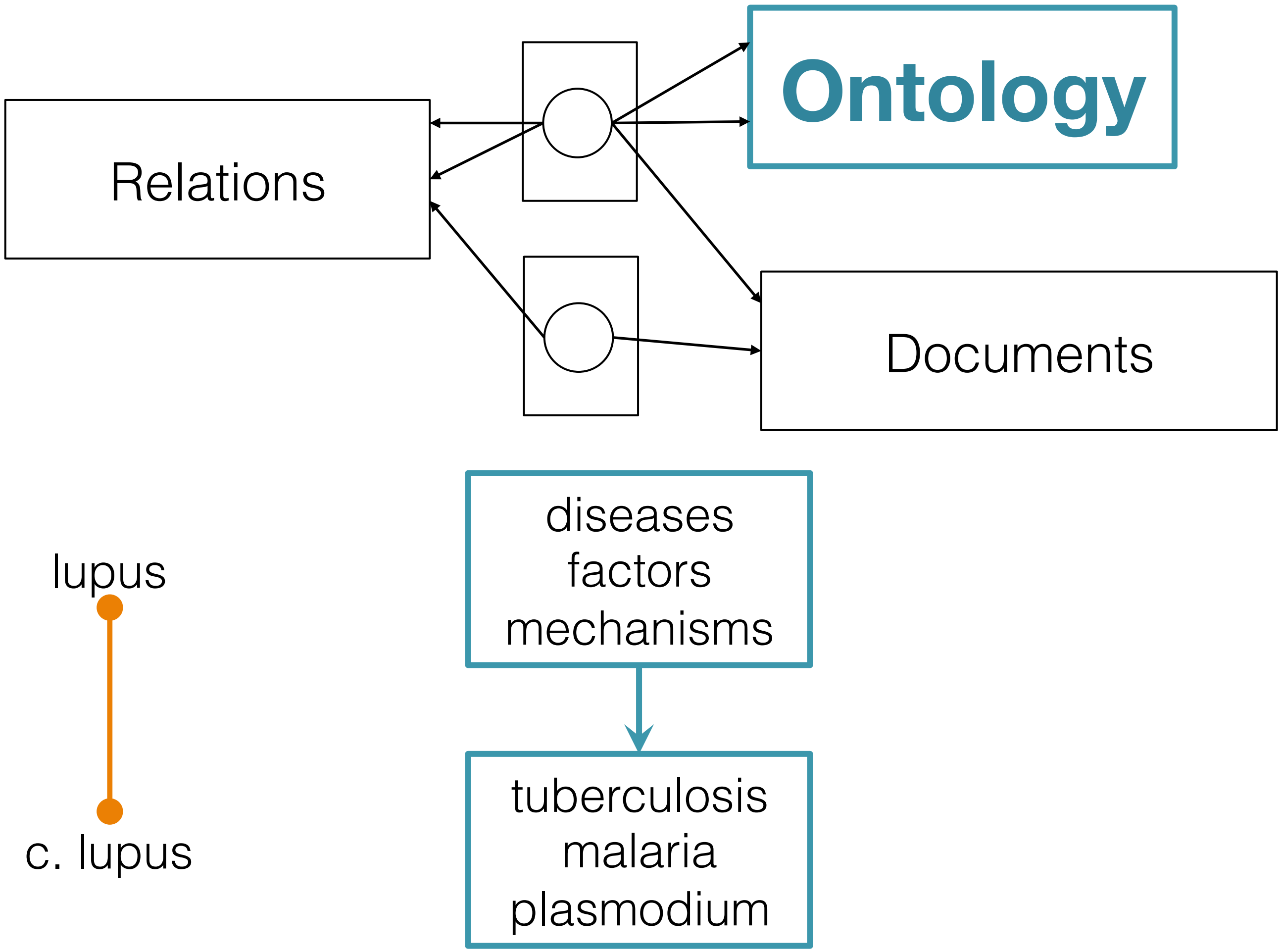
Experimental Terminology
16/100

- ✓ techniques
- ✓ samples
- ✓ hapten inhibition experiments
- ✓ sodium dodecyl sulfate-polyacrylamide gel

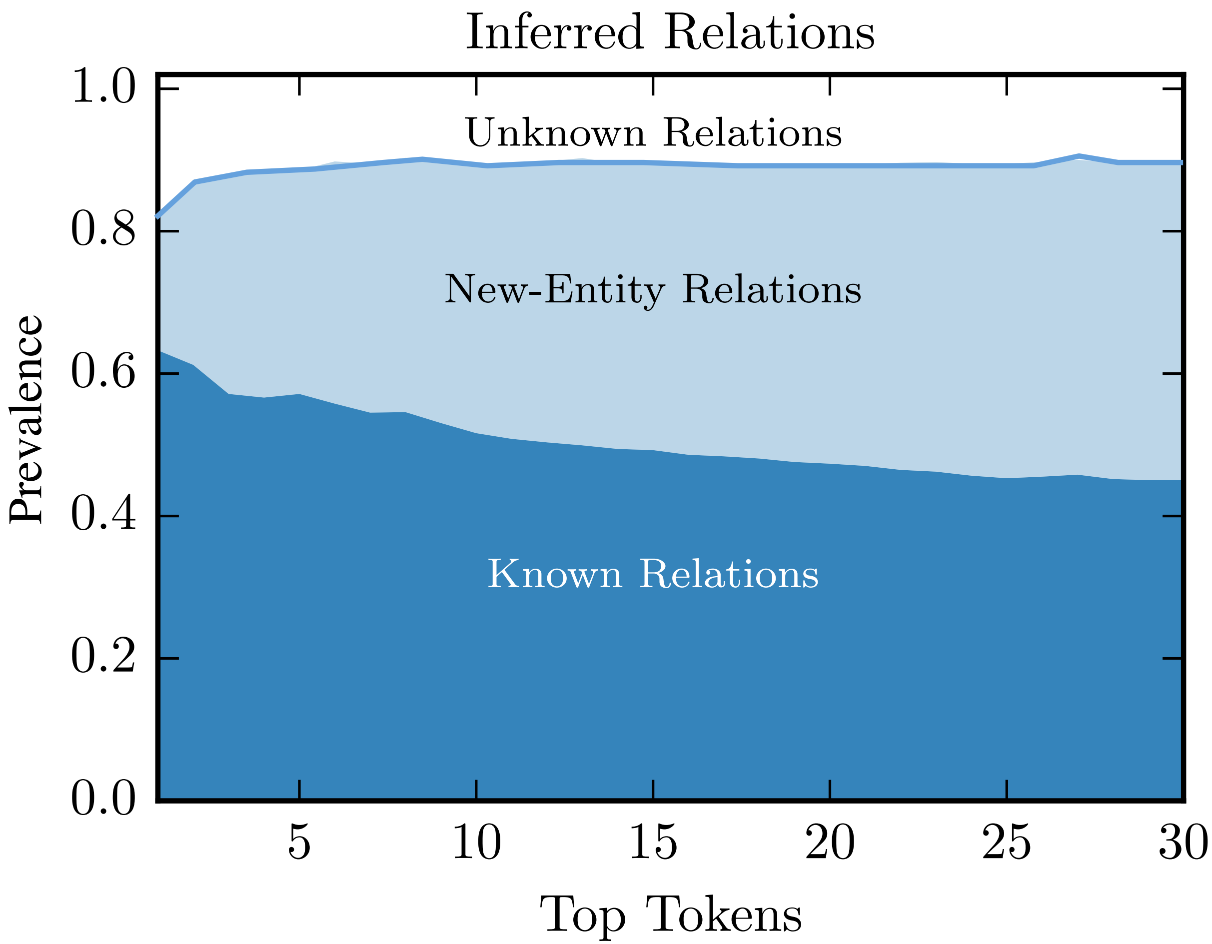
Biological Entities and Processes
81/100

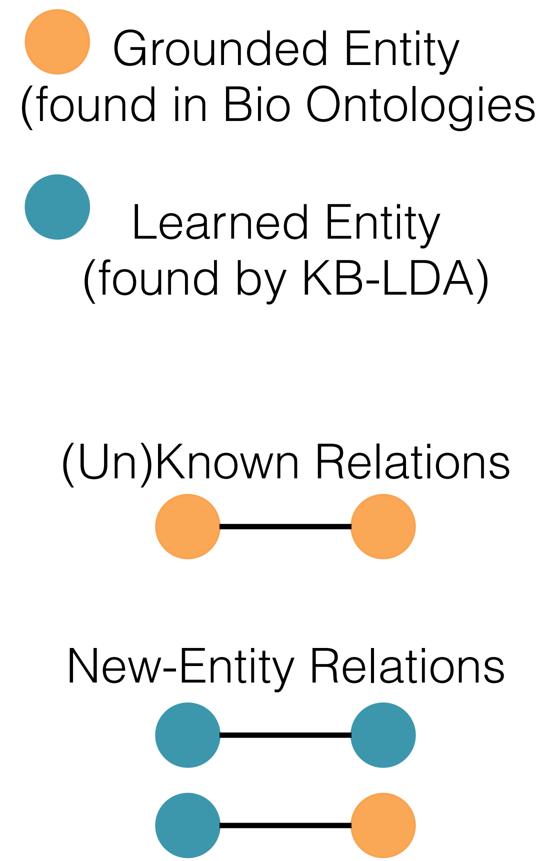
- ✓ linkage
- ✓ leukotoxin
- ✓ chemotactic response
- ✓ plasma cell-associated markers

Ontology

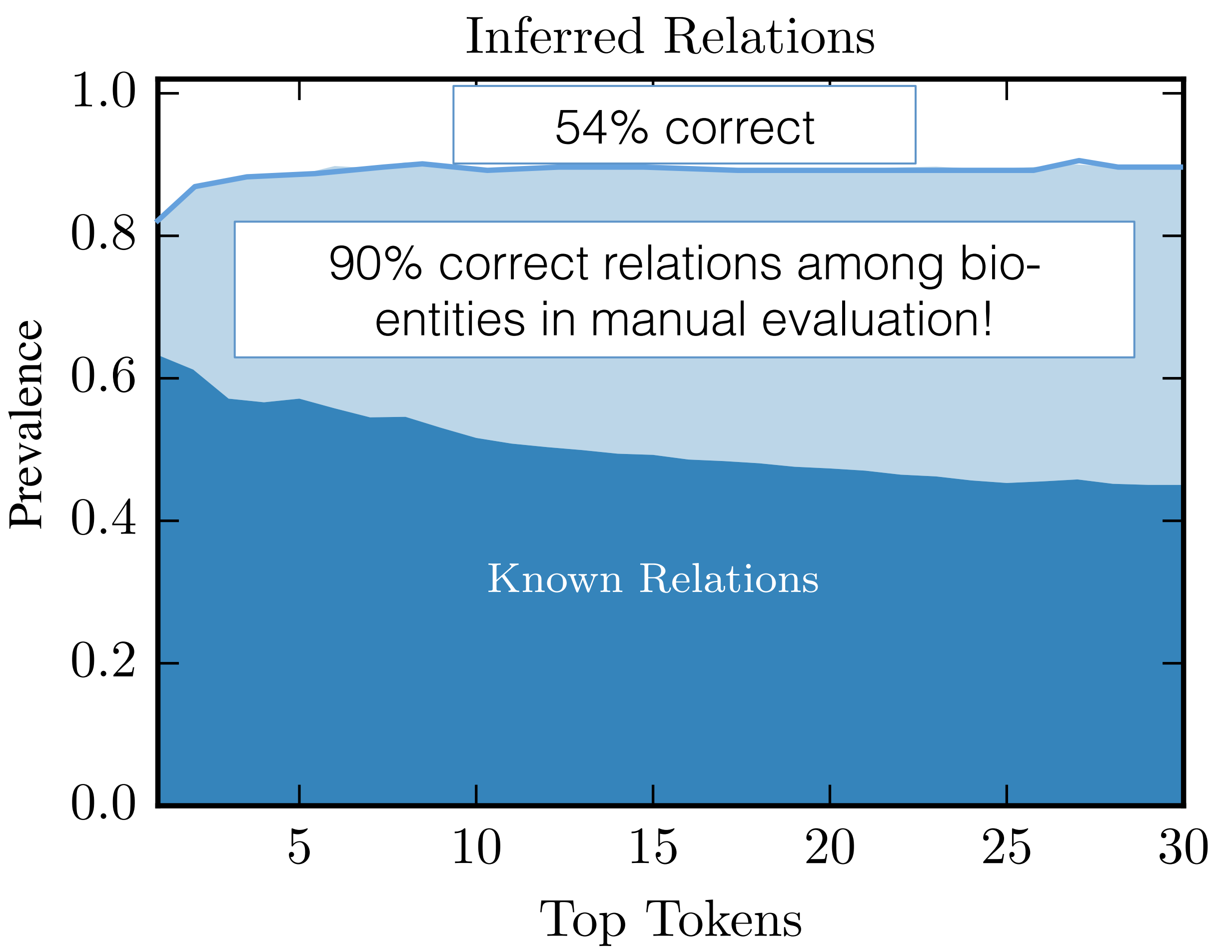

Relations
Ontology
Documents
lupus
c. lupus
diseases
factors
mechanisms
tuberculosis
malaria
plasmodium

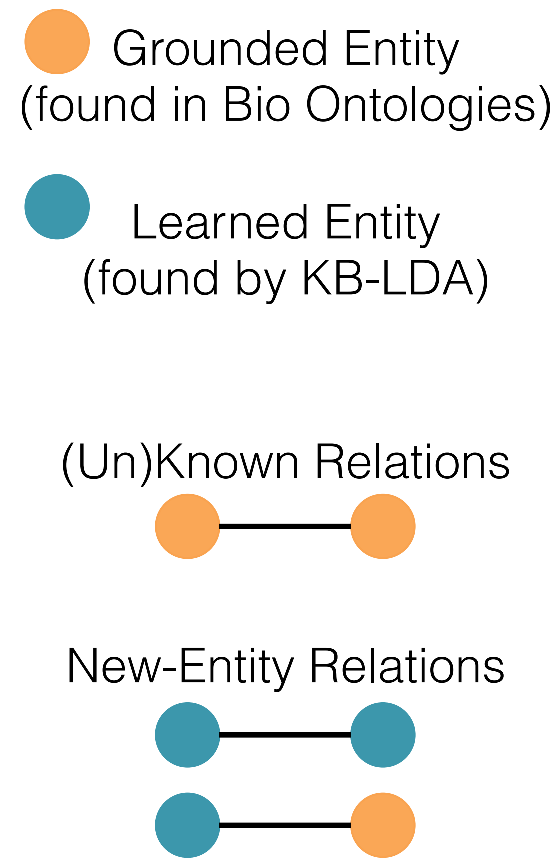
Discovered Ontology Relations



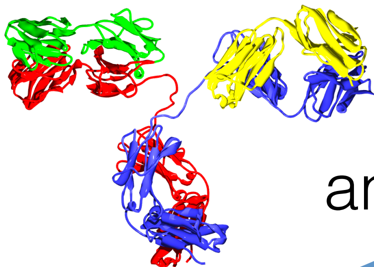


Discovered Ontology Relations





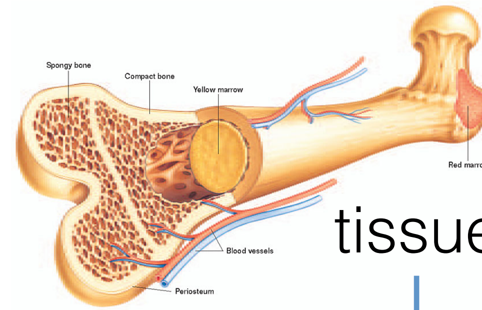
Discovered Ontology Relations



antibody

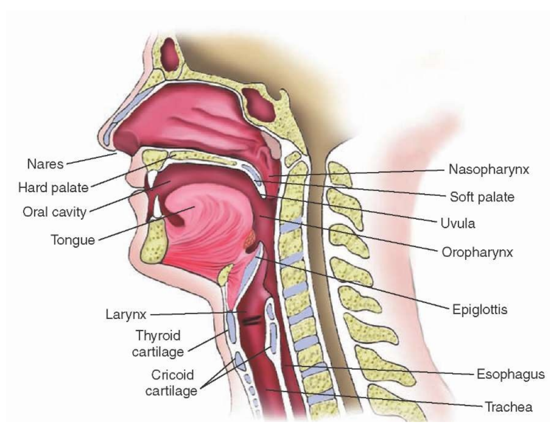
EMA (Endomysial autoantibodies)

DCS-6 (Anti-Cyclin D1 antibody)



tissue

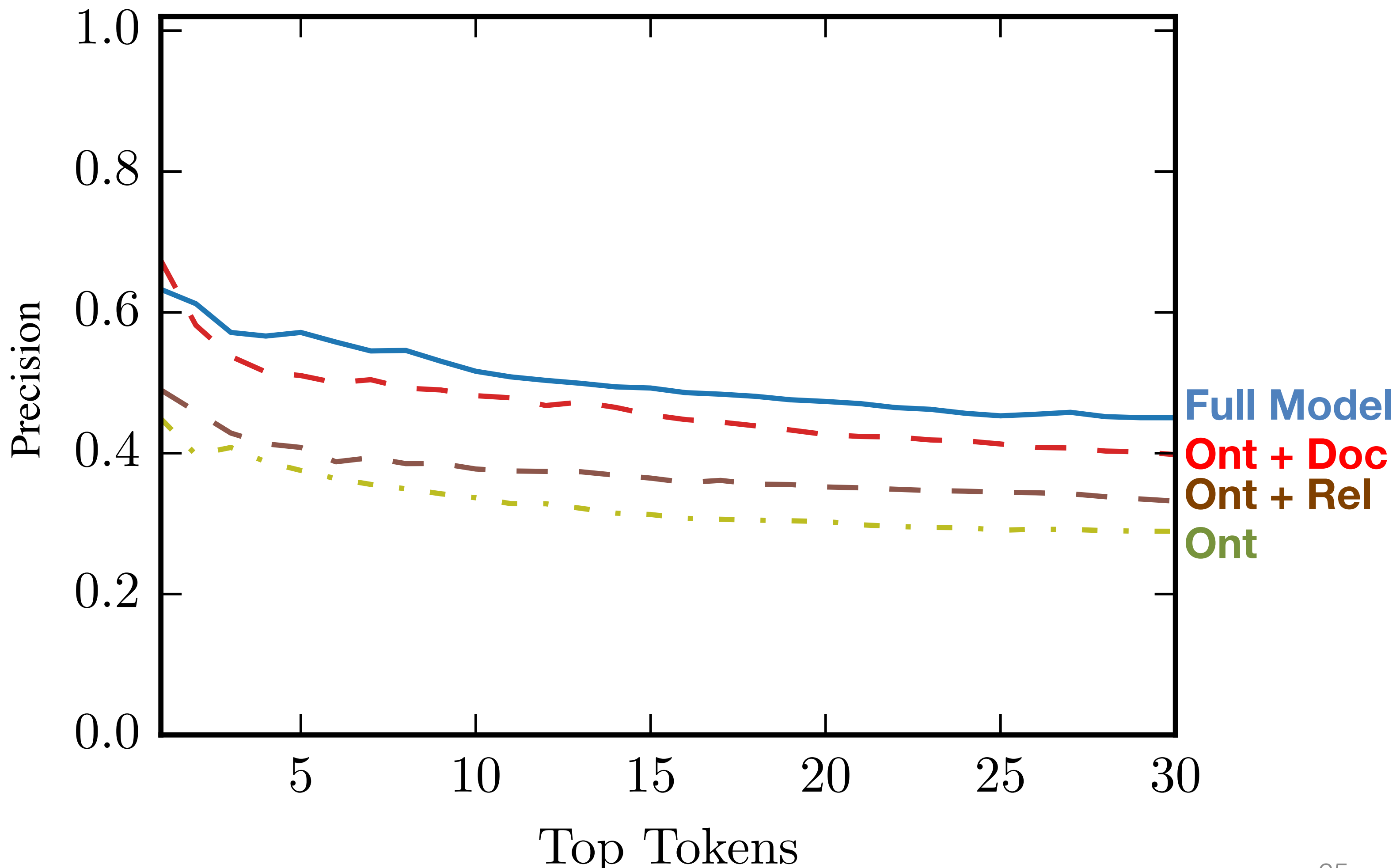
bone marrow



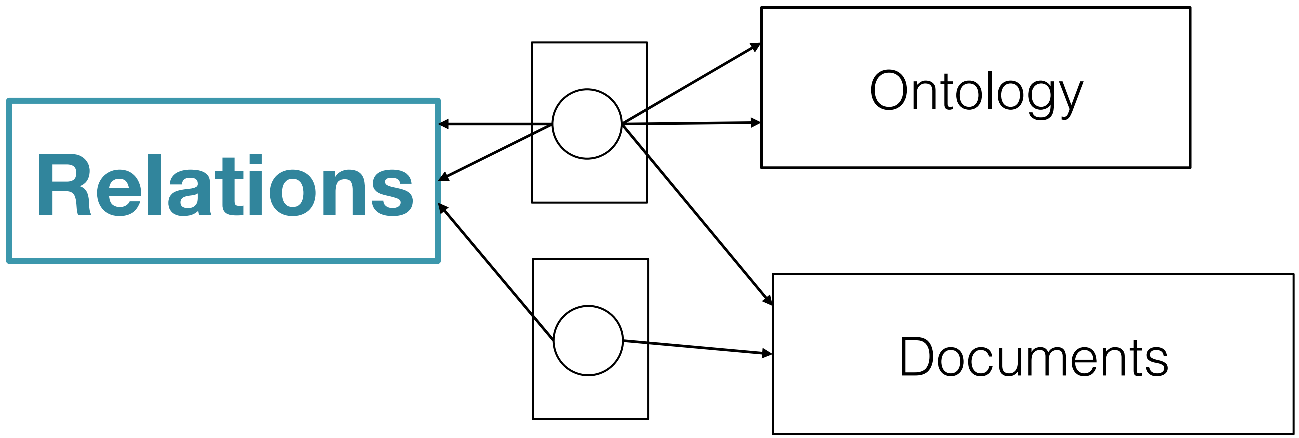
organ

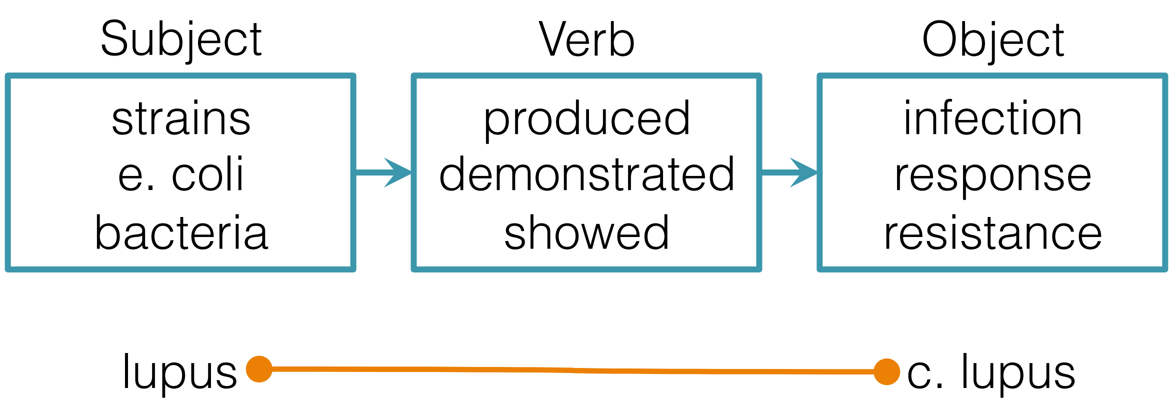
airway

Contribution of Model Components to Learning Ontology

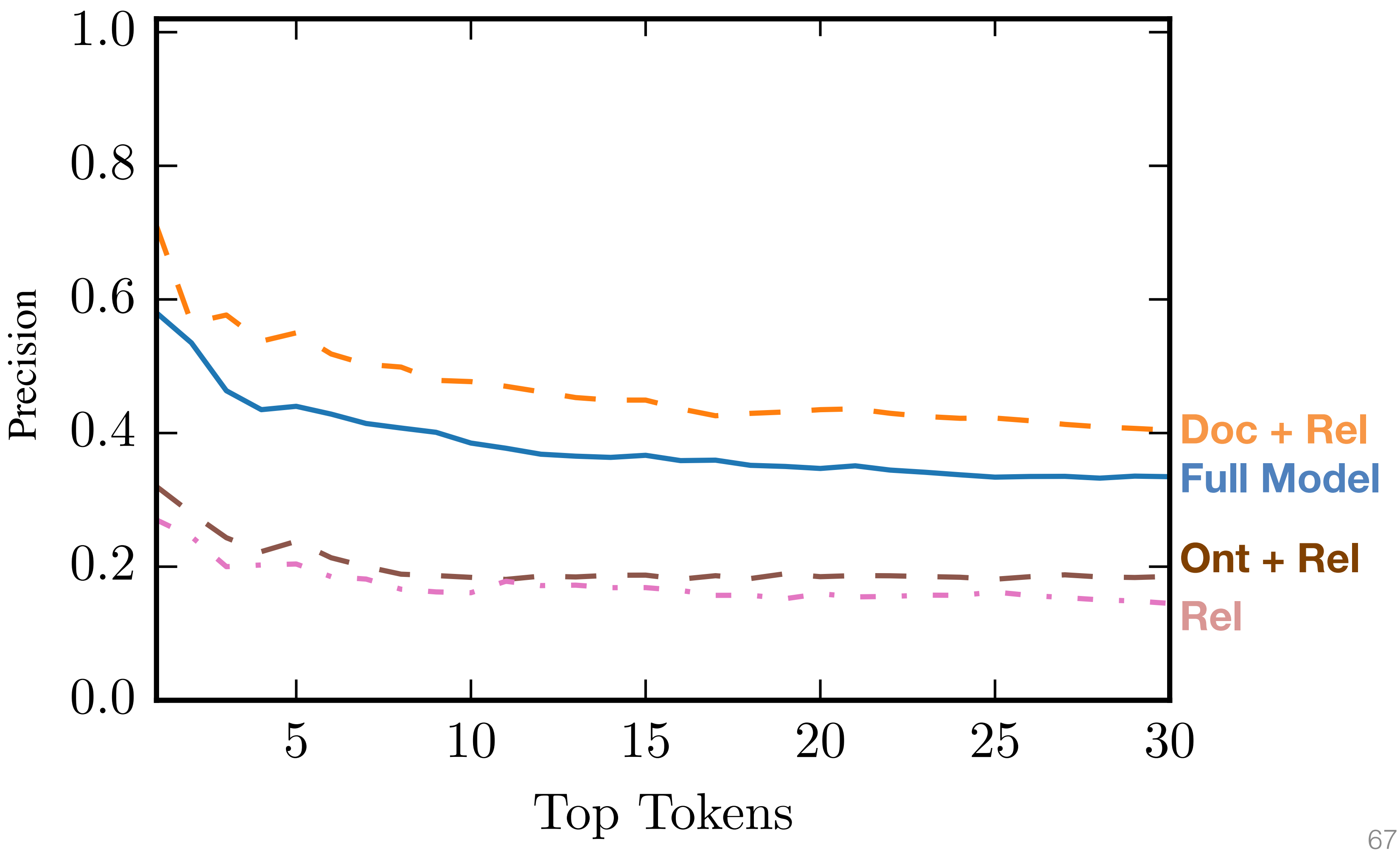


Contribution of Model Components to Learning General Relations

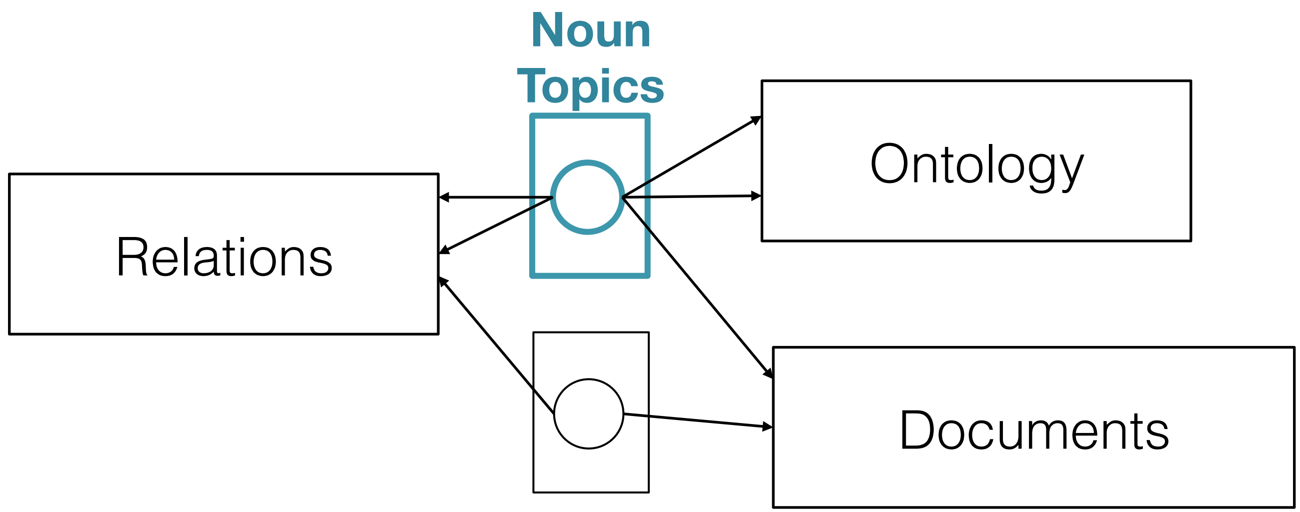


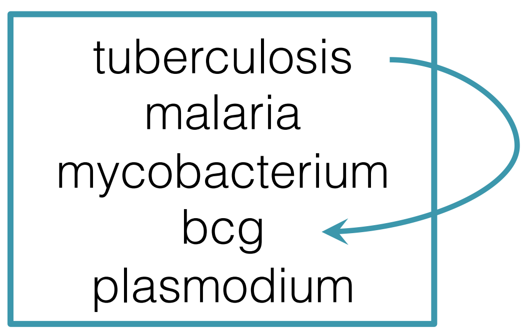


Contribution of Model Components to Learning General Relations

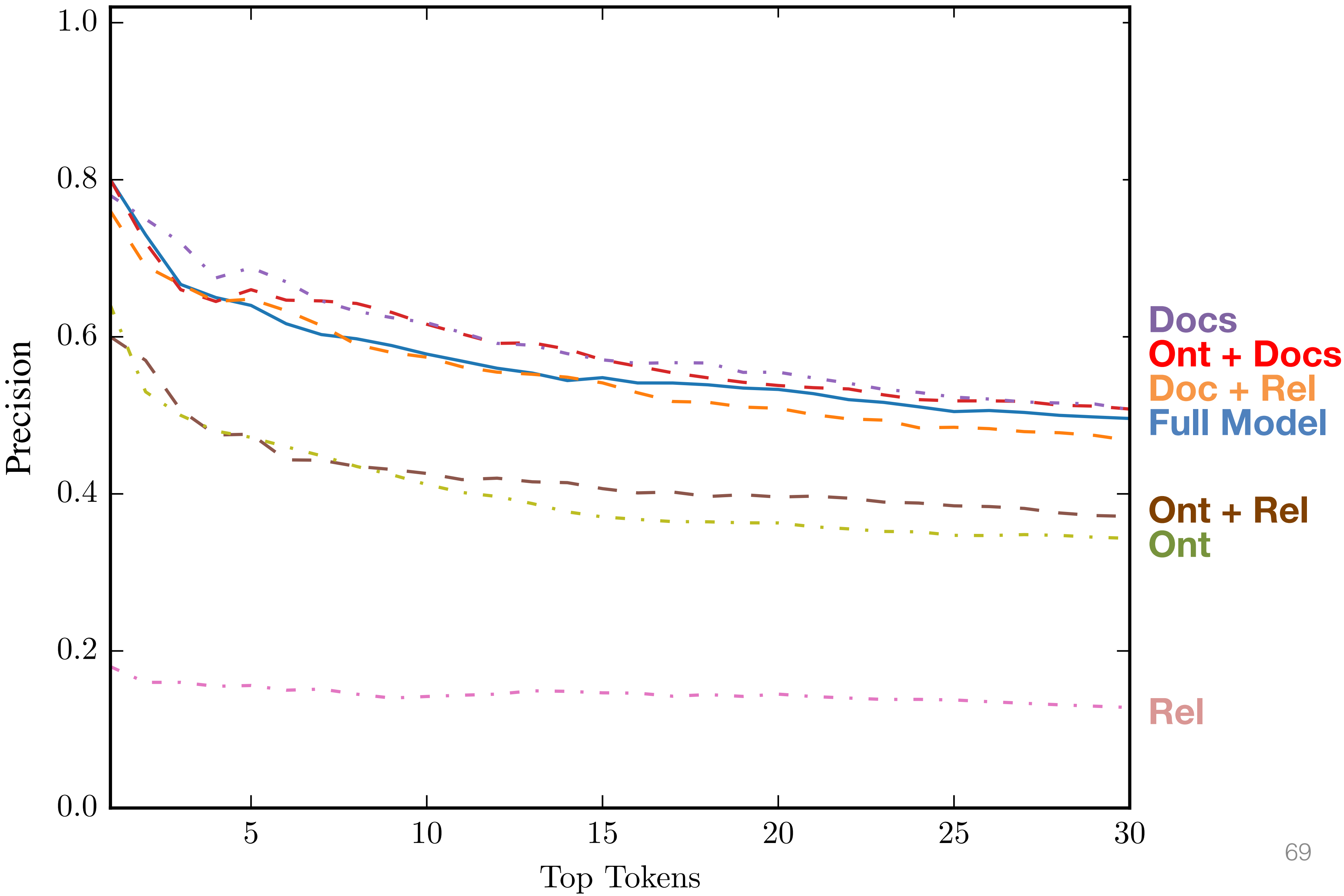


Contribution of Model Components to Learning Intra-Topic Relations

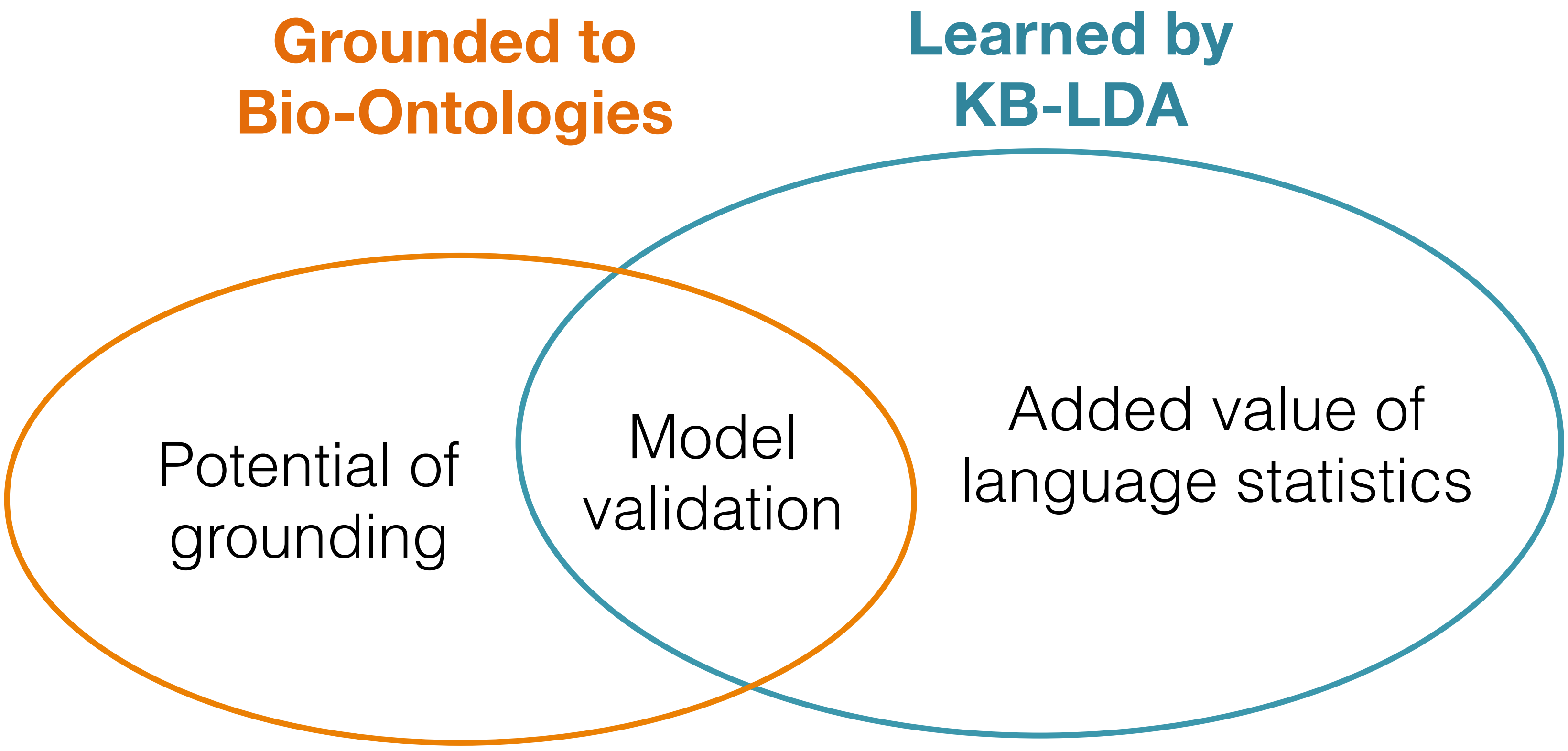




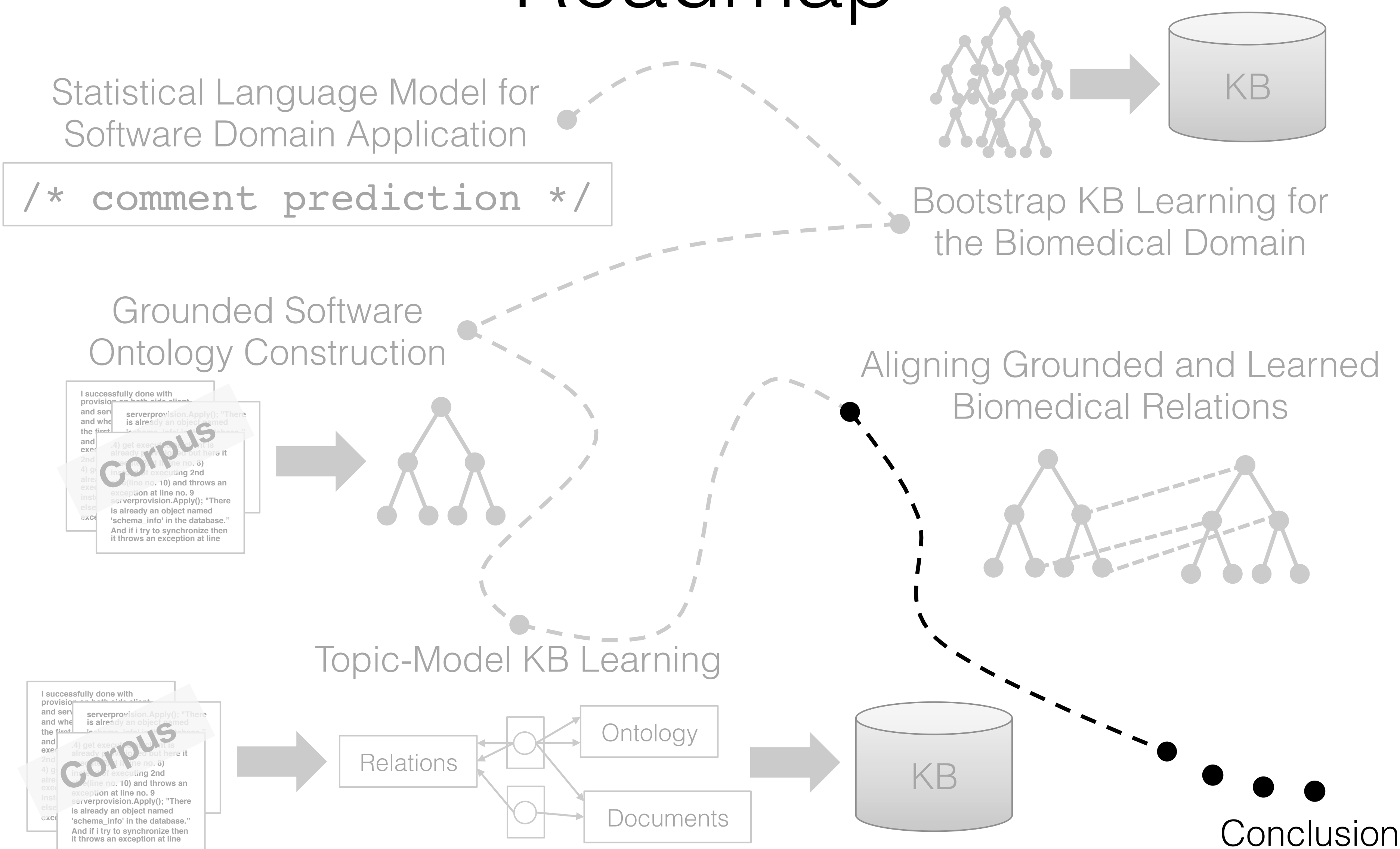
Intra-Topic Relations



Lessons


Grounded to
Bio-Ontologies
Learned by
KB-LDA
Potential of
grounding
Model
validation
Added value of
language statistics

Roadmap

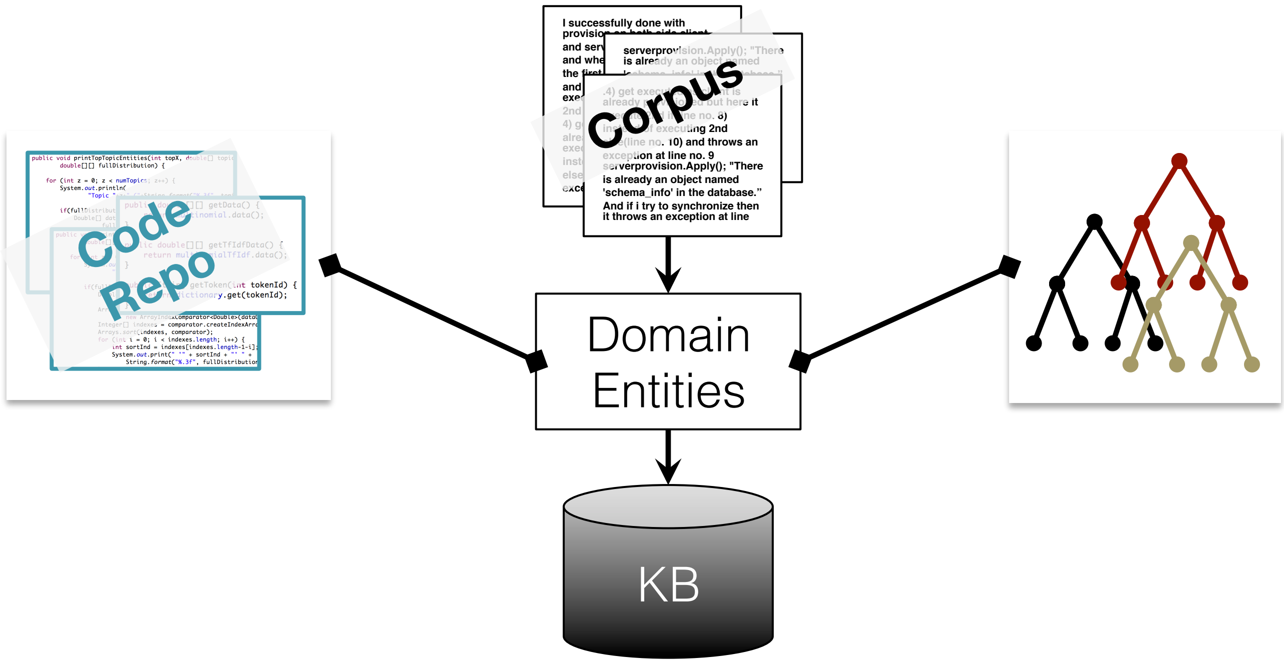


Statistical Language Model for Software Domain Application
/* comment prediction */
KB
Bootstrap KB Learning for the Biomedical Domain
Grounded Software Ontology Construction
Corpus
Aligning Grounded and Learned Biomedical Relations
Topic-Model KB Learning
Corpus
Relations
Ontology
Documents
KB
Conclusion

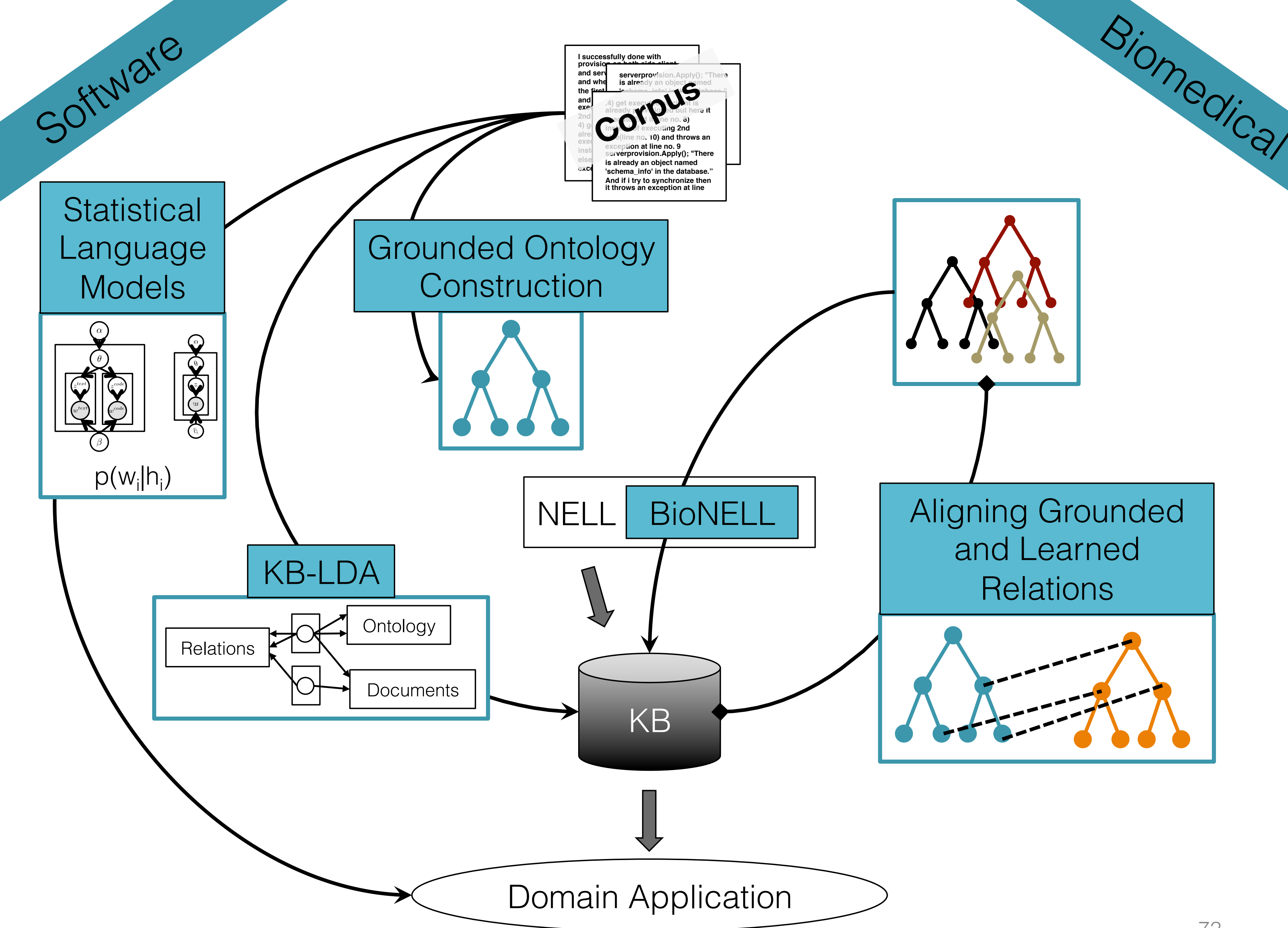
Software

Biomedical

Key Idea



"**Grounding entities** to specialized data from a scientific domain facilitates improved unsupervised and semi-supervised algorithms for **Knowledge Base construction** for that domain"

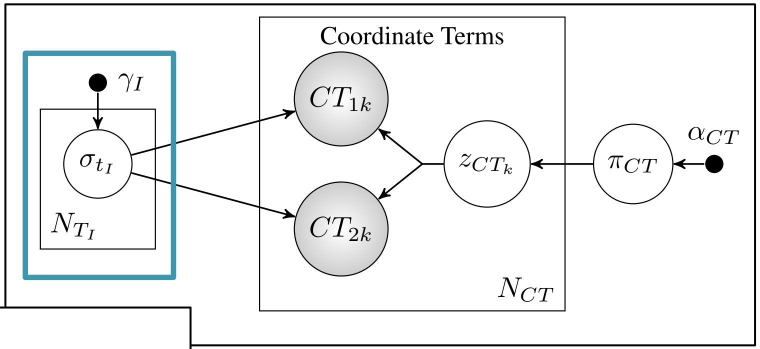


Software
Biomedical
I successfully done with
serverprovision.Apply(); "There
is already an object named
Corpus
exception at line no. 9
serverprovision.Apply(); "There
is already an object named
'schema_info' in the database."
And if i try to synchronize then
it throws an exception at line
Statistical Language Models
p(w_i|h_i)
Grounded Ontology Construction
NELL
BioNELL
KB-LDA
Relations
Ontology
Documents
KB
Aligning Grounded and Learned Relations
Domain Application

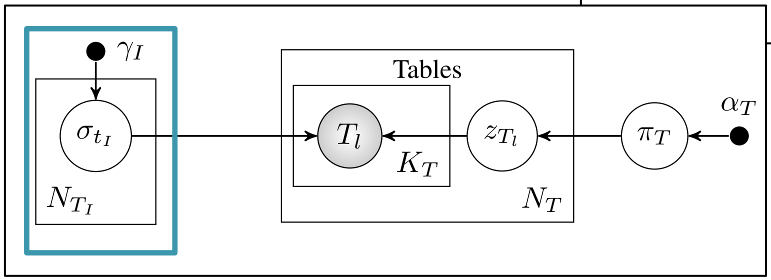
What's Next?

Grounded Knowledge Base Learning

Learned Coordinate Terms



Tables



Revisit application improvement using learned KBs