Building a Knowledge Graph by Reading the Web

Estevam R. Hruschka Jr.

Federal University of Sao Carlos

Never-Ending Language Learner

Joint work with Carnegie Mellon Read The Web Project Group
(http://rtw.ml.cmu.edu/rtw/)
and MaLL (Machine Learning Lab) from Federal University of São Carlos
(http://www.dc.ufscar.br/MaLL/MaLL.html)

Humans learn many things, for years, and become better learners over time

Why not machines?

Never-Ending Learning

We'll never really understand learning until we build machines that

- learn many different things,
- over years,
- and become better learners over time.

Never-Ending Learning

We'll never produce natural language understanding systems until we have systems that react to arbitrary sentences by saying one of:

- I understand, and already knew that
- I understand, and didn't know, but accept it
- I understand, and disagree because …

Picture taken from [DARPA, 2012]

Machine Reading

Auto-Text to Knowledge

Picture taken from [DARPA, 2012]

NELL: Never-Ending Language Learner

Inputs:

- initial ontology
- handful of examples of each predicate in ontology
- the web
- occasional interaction with human trainers

The task:

- run 24x7, forever
- each day:
  1. extract more facts from the web to populate the initial ontology
  2. learn to read (perform #1) better than yesterday

NELL: Never-Ending Language Learner

Goal:

- run 24x7, forever
- each day:
  1. extract more facts from the web to populate given ontology
  2. learn to read better than yesterday

Today...

Running 24 x 7, since January, 2010

Input:

- ontology defining ~800 categories and relations
- 10-20 seed examples of each
- 1 billion web pages (ClueWeb – Jamie Callan)

Result:

- continuously growing KB with +90.000,000 extracted beliefs (different levels of confidence)

http://rtw.ml.cmu.edu

Read the Web

Research Project at Carnegie Mellon University

Home | Project Overview | Resources & Data | Publications | People

NELL: Never-Ending Language Learning

Can computers learn to read? We think so. "Read the Web" is a research project that attempts to create a computer system that learns over time to read the web. Since January 2010, our computer system called NELL (Never-Ending Language Learner) has been running continuously, attempting to perform two tasks each day:

Browse the Knowledge Base!

- First, it attempts to "read," or extract facts from text found in hundreds of millions of web pages (e.g., `playsInstrument(George_Harrison, guitar)`).
- Second, it attempts to improve its reading competence, so that tomorrow it can extract more facts from the web, more accurately.

So far, NELL has accumulated over 15 million candidate beliefs by reading the web, and it is considering these at different levels of confidence. NELL has high confidence in 1,471,011 of these beliefs — these are displayed on this website. It is not perfect, but NELL is learning. You can track NELL's progress below or @cmunell on Twitter, browse and download its knowledge base, read more about our technical approach, or join the discussion group.

NELL: Never-Ending Language Learner

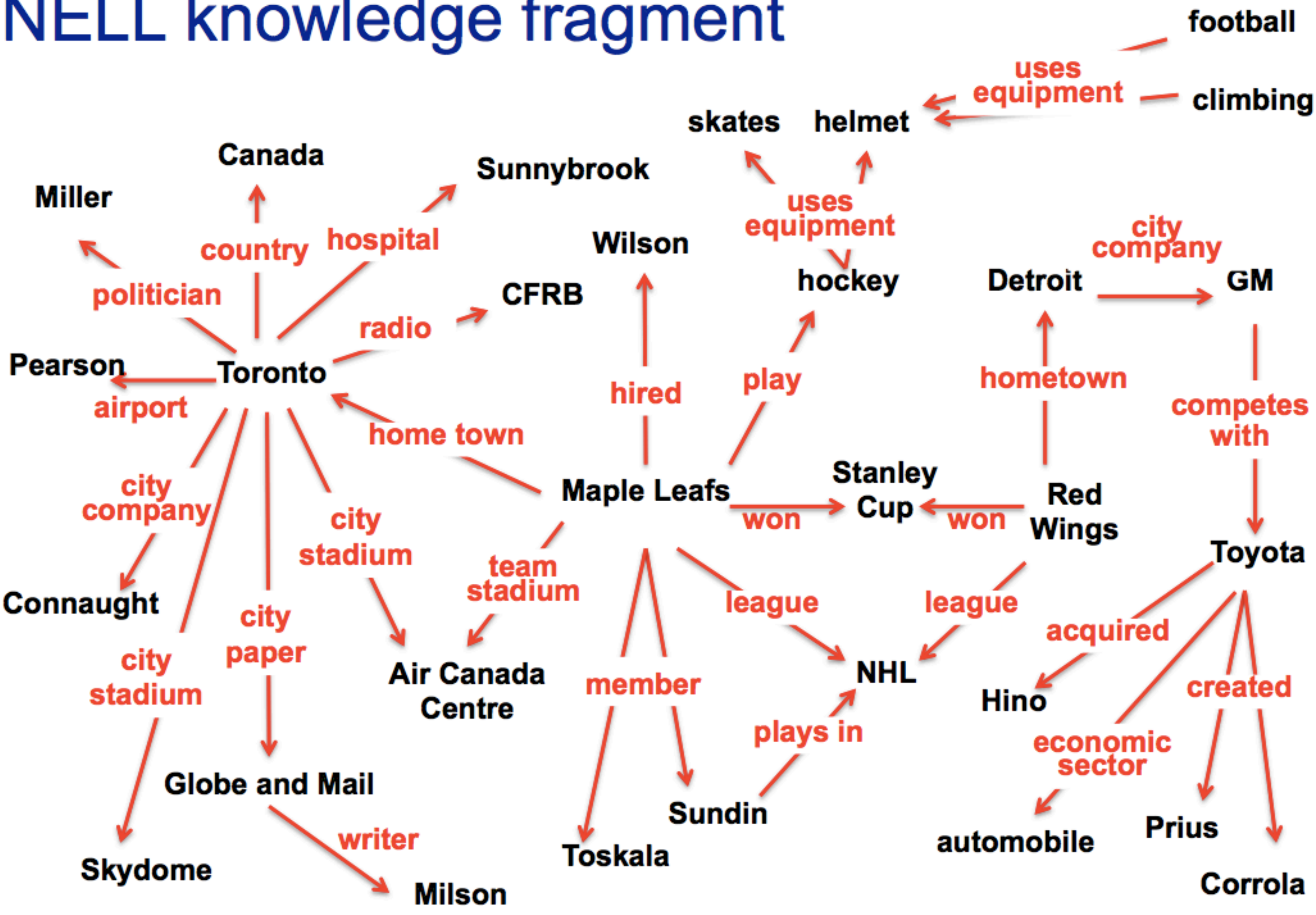
http://rtw.ml.cmu.edu

Recently-Learned Facts

Refresh

instance	iteration	date learned	confidence
thailand_philharmonic_orchestra is a musician	808	31-jan-2014	93.9
islamic_azad_university is a university	808	31-jan-2014	90.9
jesse_green is a chef	808	31-jan-2014	95.8
stinkpot_turtle is an amphibian	812	15-feb-2014	91.0
iaff is a trade union	809	03-feb-2014	92.1
mississippi empties into river st__croix_river	808	31-jan-2014	99.2
jim_plunkett plays in the league nfl	813	16-feb-2014	95.0
david_lean directed the movie doctor_zhivago	808	31-jan-2014	100.0
line is a role for players of ncaa_basketball	811	10-feb-2014	93.8
marc is the leader of the city neworleans	813	16-feb-2014	100.0

NELL knowledge fragment



Building the Knowledge Graph by Reading

1. Classify noun phrases (NP's) by category

The Problem with Semi-Supervised Bootstrap Learning

Paris
Pittsburgh
Seattle
Cupertino

The Problem with Semi-Supervised Bootstrap Learning

Paris
Pittsburgh
Seattle
Cupertino

mayor of arg1
live in arg1

The Problem with Semi-Supervised Bootstrap Learning

Paris
Pittsburgh
Seattle
Cupertino

San Francisco
Austin
denial

mayor of arg1
live in arg1

The Problem with Semi-Supervised Bootstrap Learning

Paris
Pittsburgh
Seattle
Cupertino

San Francisco
Austin
denial

mayor of arg1
live in arg1

The Problem with Semi-Supervised Bootstrap Learning

Paris
Pittsburgh
Seattle
Cupertino

San Francisco
Austin
denial

mayor of arg1
live in arg1

arg1 is home of
traits such as arg1

The Problem with Semi-Supervised Bootstrap Learning

Paris
Pittsburgh
Seattle
Cupertino

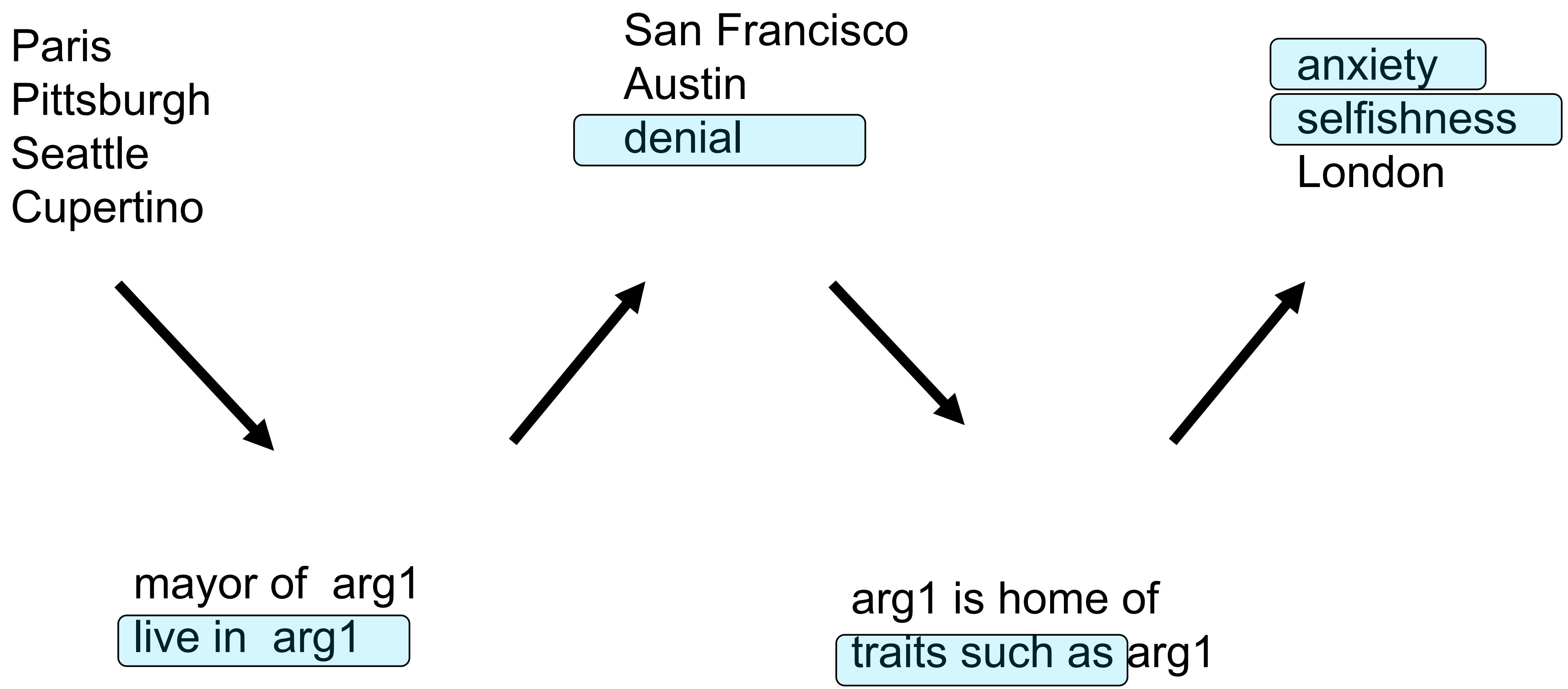
mayor of arg1
live in arg1

San Francisco
Austin
denial

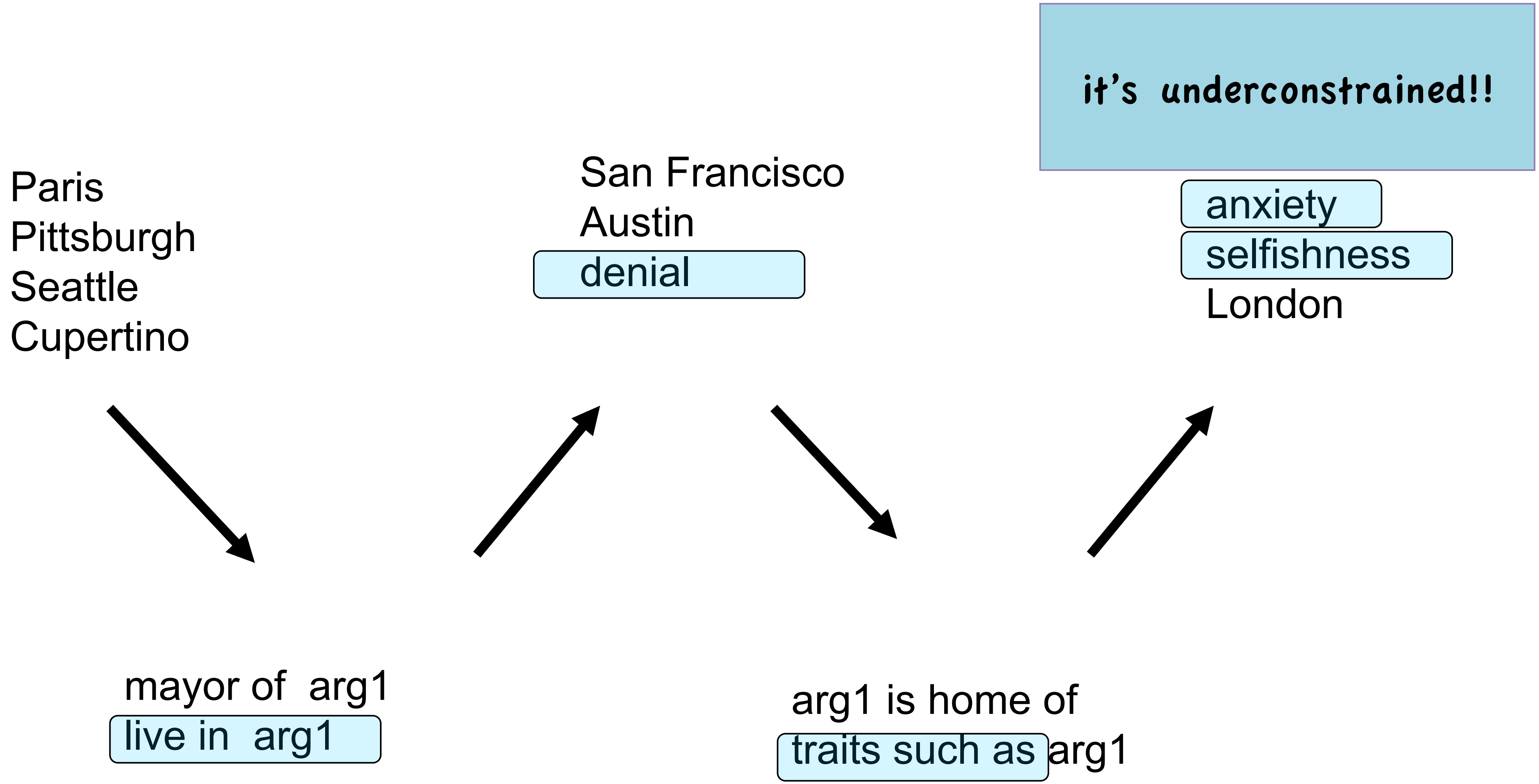
arg1 is home of
traits such as arg1

anxiety
selfishness
London

The Problem with Semi-Supervised Bootstrap Learning

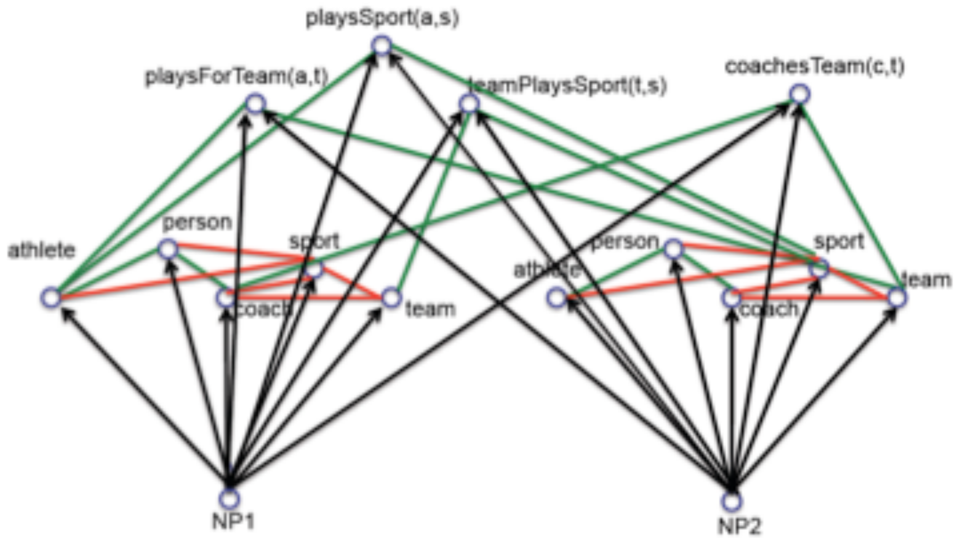


The Problem with Semi-Supervised Bootstrap Learning



Key Idea 1: Coupled semi-supervised training of many functions

hard
(underconstrained)
semi-supervised
learning problem



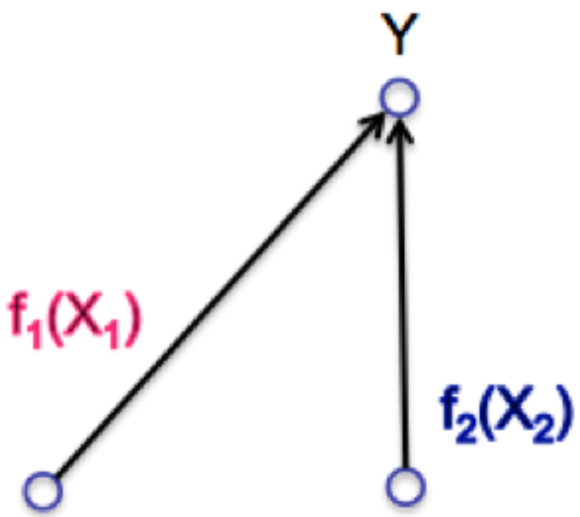
much easier (more constrained)
semi-supervised learning problem

Coupled Training Type 1: Co-training, Multiview, Co-regularization

[Blum & Mitchell; 98]
[Dasgupta et al; 01]
[Ganchev et al., 08]
[Sridharan & Kakade, 08]
[Wang & Zhou, ICML10]

Coupled Training Type 1: Co-training, Multiview, Co-regularization

[Blum & Mitchell; 98]
[Dasgupta et al; 01]
[Ganchev et al., 08]
[Sridharan & Kakade, 08]
[Wang & Zhou, ICML10]

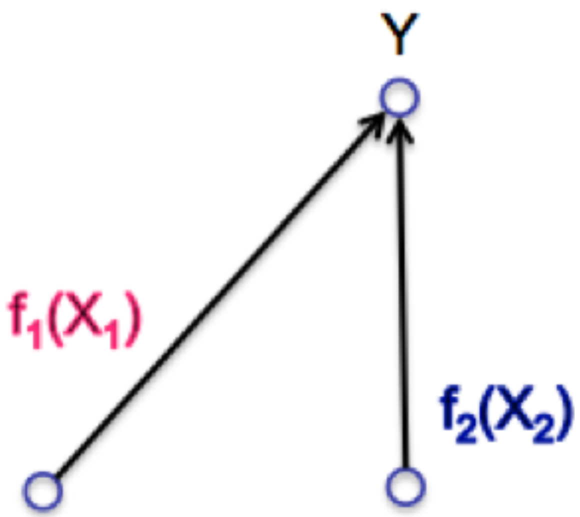


$\mathbf{X} = < X_1 , X_2 >$

Constraint: $f_1(x_1) = f_2(x_2)$

Coupled Training Type 1: Co-training, Multiview, Co-regularization

[Blum & Mitchell; 98]
[Dasgupta et al; 01]
[Ganchev et al., 08]
[Sridharan & Kakade, 08]
[Wang & Zhou, ICML10]



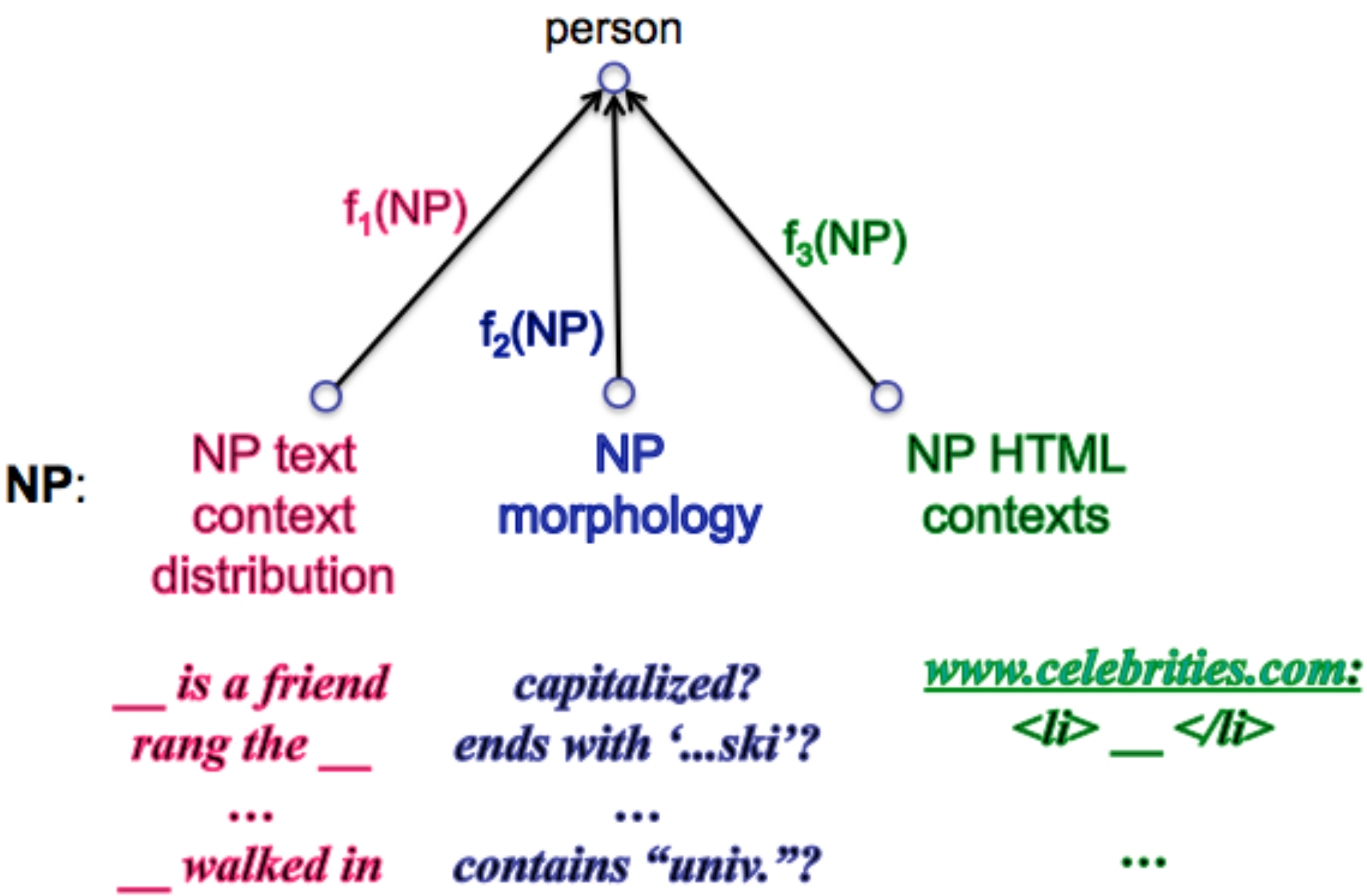
$\mathbf{X} = < X_1 \quad , \quad X_2 >$

Constraint: $f_1(x_1) = f_2(x_2)$

If f_1, f_2 PAC learnable,
X_1, X_2 conditionally indep
Then PAC learnable from *unlabeled* data and weak initial learner

and disagreement between f_1, f_2 bounds error of each

Type 1 Coupling Constraints in NELL



Coupled Training Type 2:
Structured Outputs, Multitask, Posterior Regularization, Multilabel

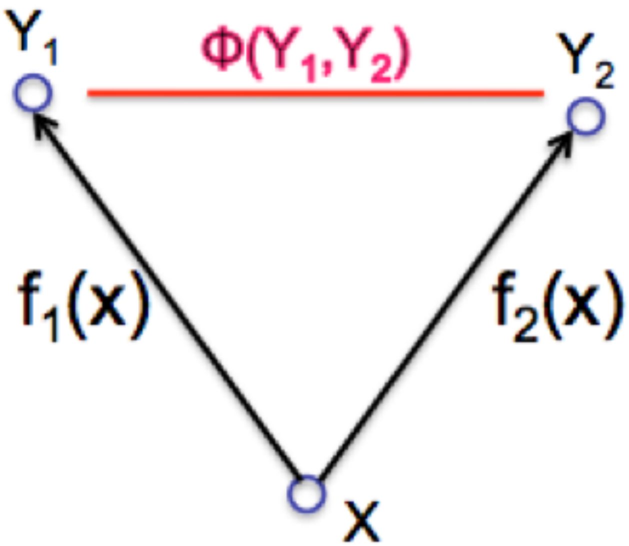
Learn functions with the same input, different outputs, where we know some constraint

[Daume, 2008]
[Bakhir et al., eds. 2007]
[Roth et al., 2008]
[Taskar et al., 2009]
[Carlson et al., 2009]

Coupled Training Type 2:
Structured Outputs, Multitask, Posterior Regularization, Multilabel

[Daume, 2008]
[Bakhir et al., eds. 2007]
[Roth et al., 2008]
[Taskar et al., 2009]
[Carlson et al., 2009]

Learn functions with the same input, different outputs, where we know some constraint

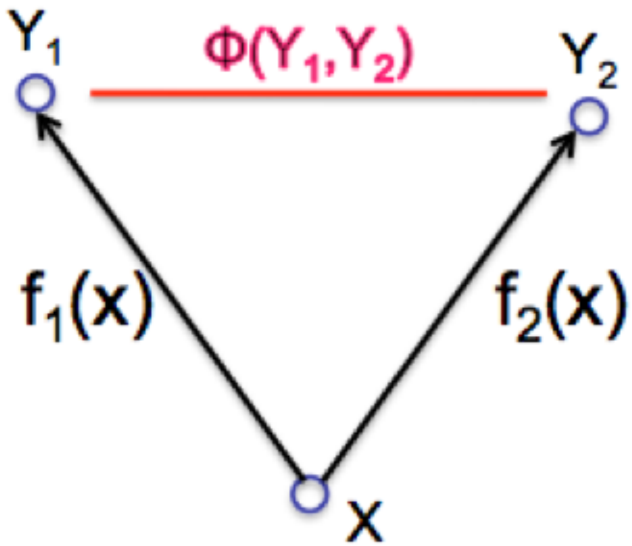


Constraint: $\Phi(f_1(x), f_2(x))$

Coupled Training Type 2:
Structured Outputs, Multitask, Posterior Regularization, Multilabel

[Daume, 2008]
[Bakhir et al., eds. 2007]
[Roth et al., 2008]
[Taskar et al., 2009]
[Carlson et al., 2009]

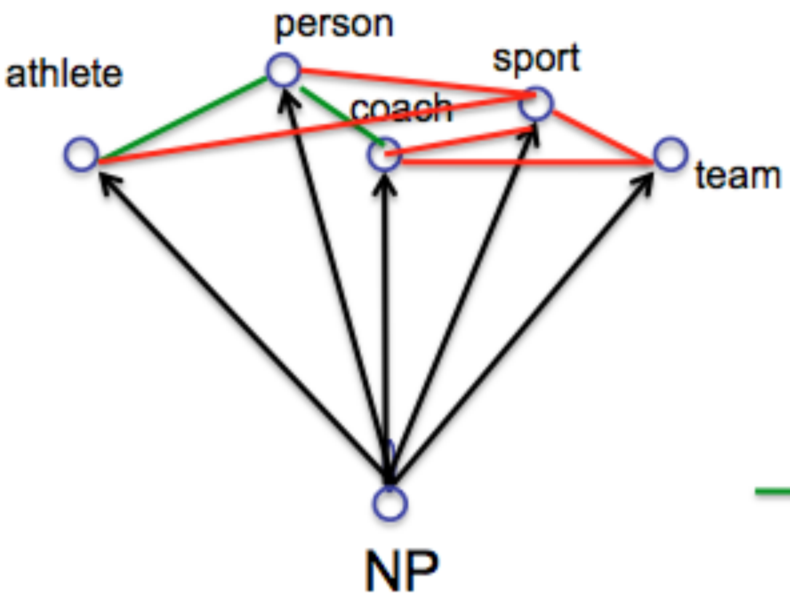
Learn functions with the same input, different outputs, where we know some constraint



Constraint: $\Phi(f_1(x), f_2(x))$

Effectiveness ~ probability that $\Phi(Y_1,Y_2)$ will be violated by incorrect f_j and f_k

Type 2 Coupling Constraints in NELL

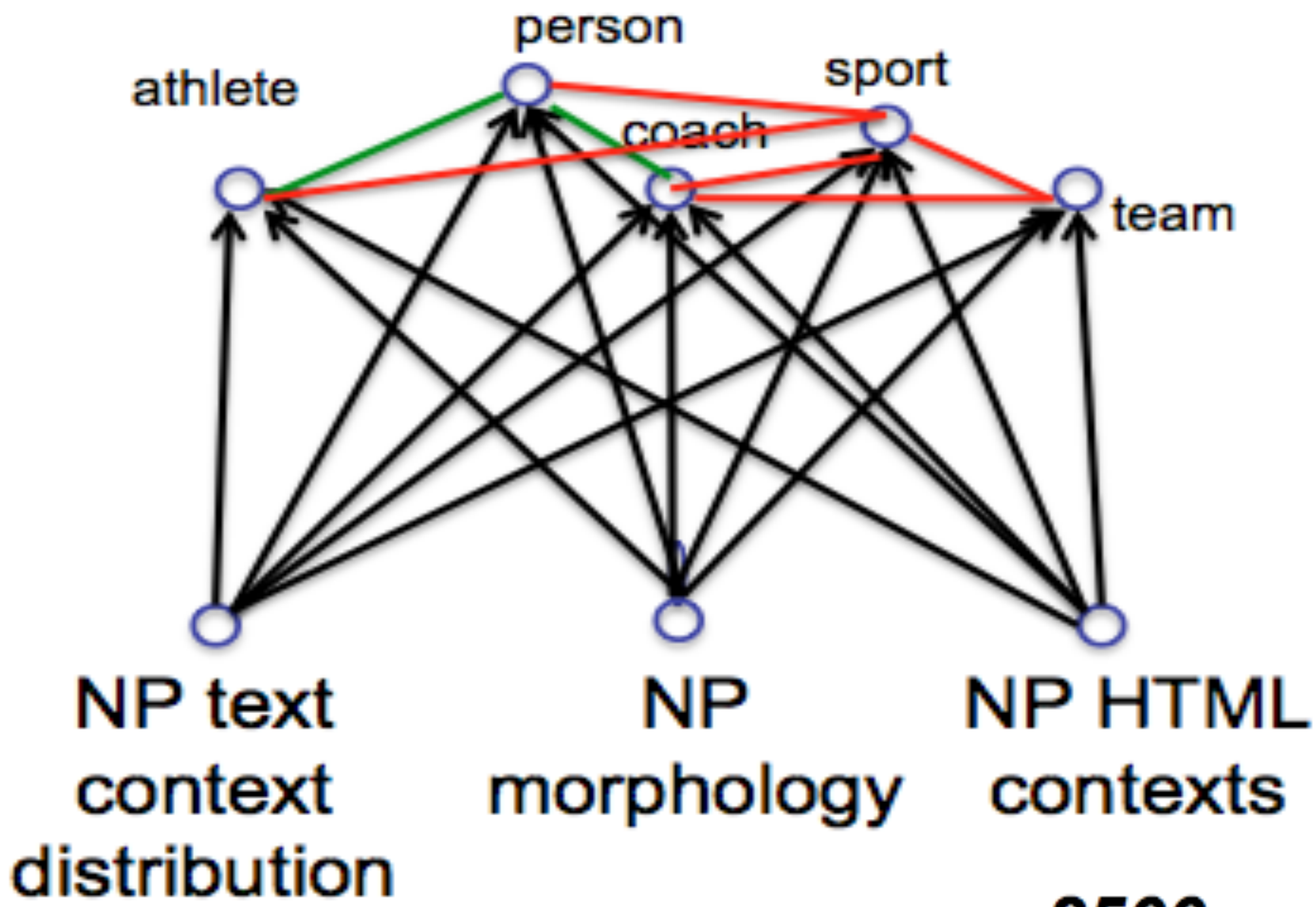


athlete(NP) → person(NP)

athlete(NP) → NOT sport(NP)

NOT athlete(NP) ← sport(NP)

Multi-view, Multi-Task Coupling



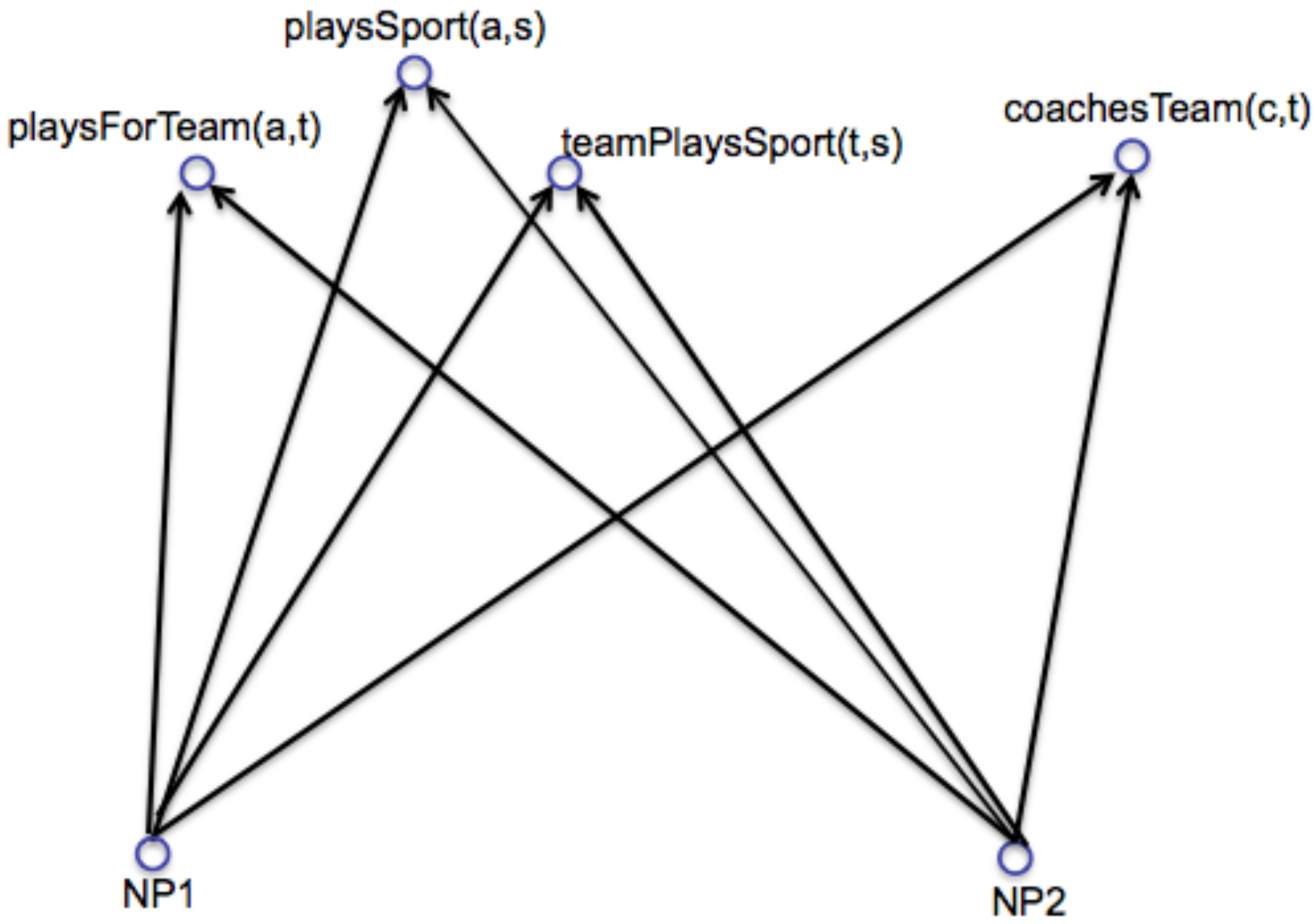
NP:

over 2500 coupled functions in NELL

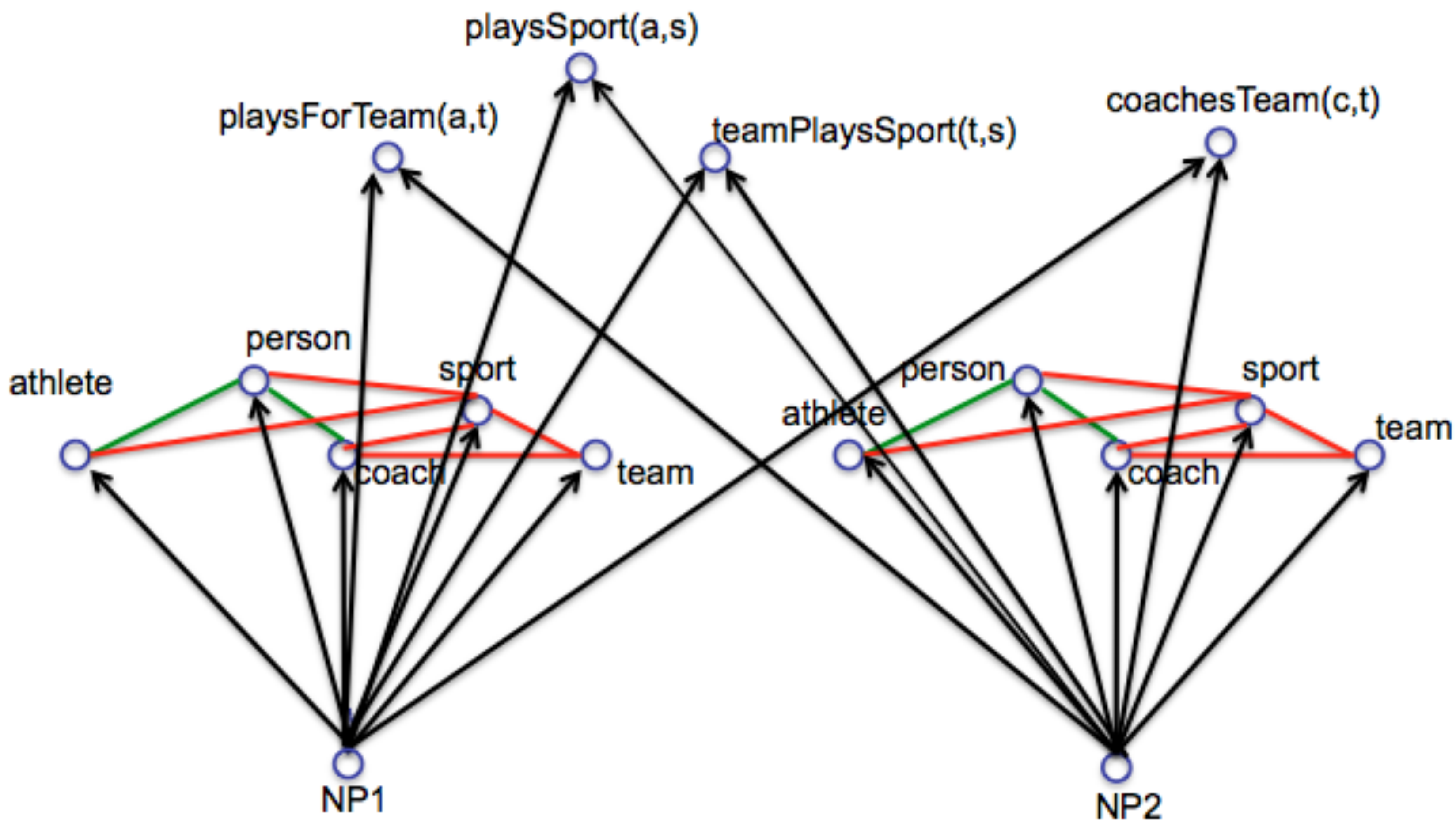
Building the Knowledge Graph by Reading

1. Classify noun phrases (NP’s) by category
2. Classify NP pairs by relation

Learning Relations between NP's

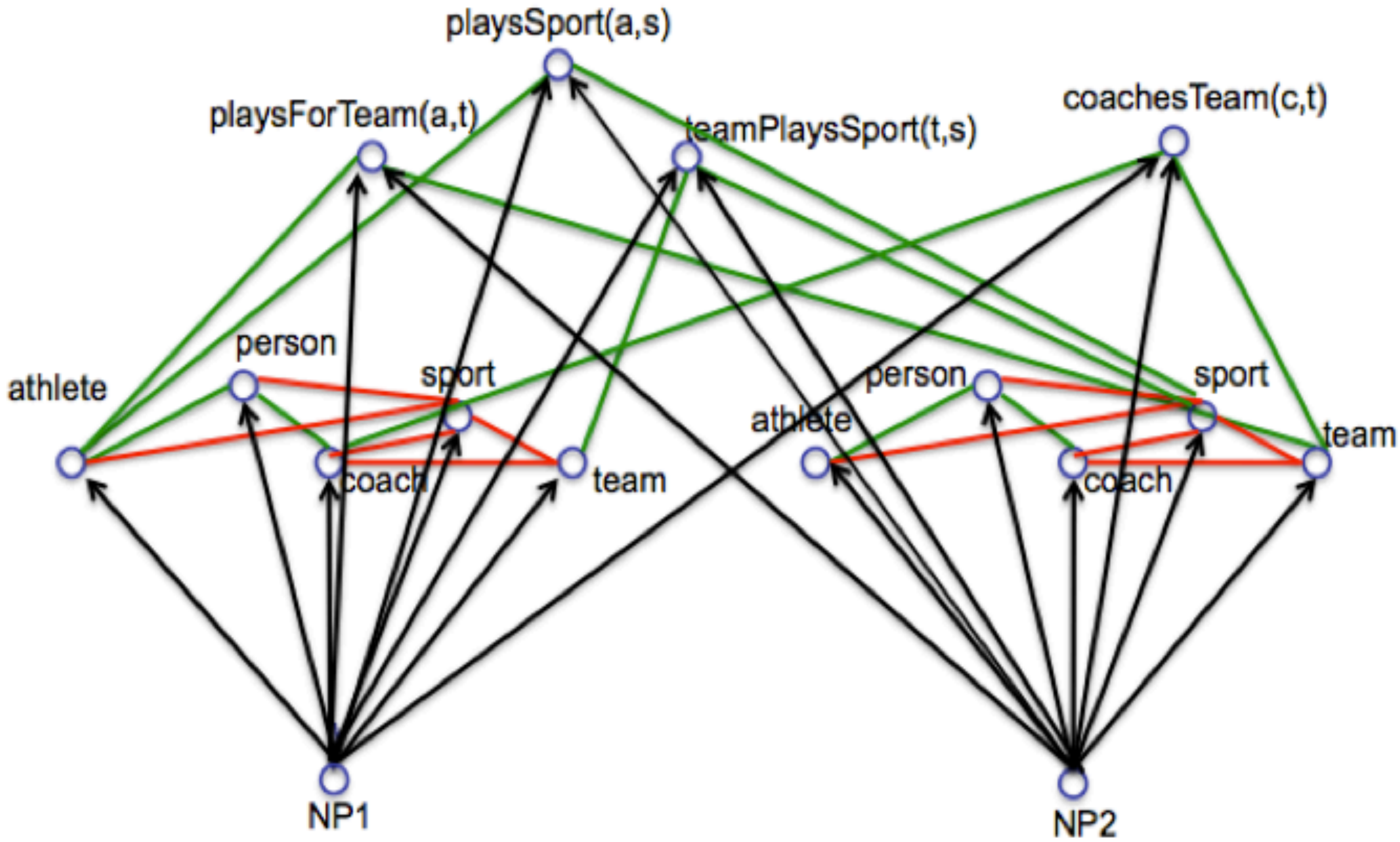


Learning Relations between NP's



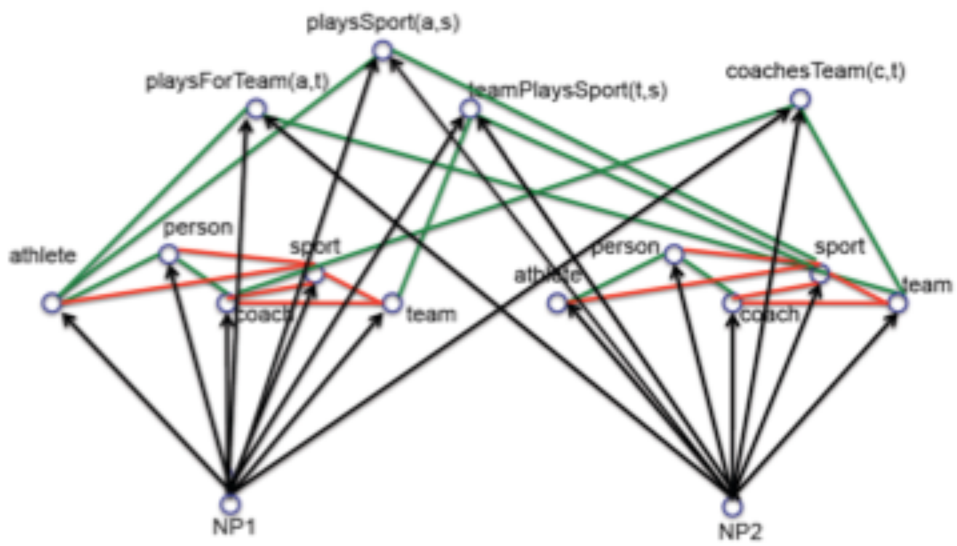
Type 3 Coupling: Argument Types

Constraint: f3(x1,x2) → (f1(x1) AND f2(x2))



playsSport(NP1,NP2) → athlete(NP1), sport(NP2)

Pure EM Approach to Coupled Training



E: jointly estimate latent labels for each function of each unlabeled example

M: retrain all functions, based on these probabilistic labels

Scaling problem:

- **E** step: 20M NP's, 10^{14} NP pairs to label
- **M** step: 50M text contexts to consider for each function → 10^{10} parameters to retrain
- even more URL-HTML contexts..

NELL's Approximation to EM

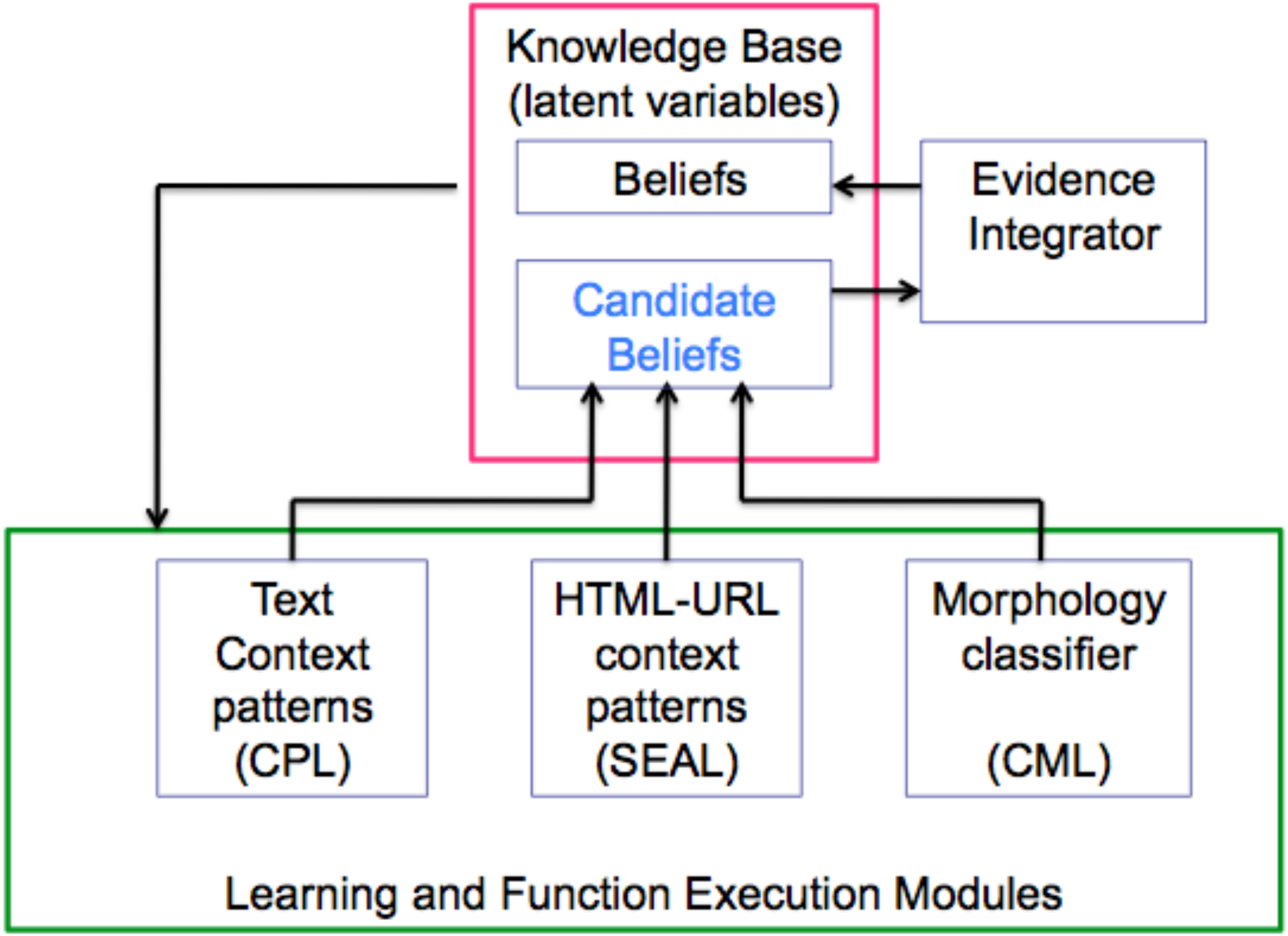
E' step:

- Consider only a growing subset of the latent variable assignments
  - category variables: up to 250 NP's per category per iteration
  - relation variables: add only if confident and args of correct type
  - this set of explicit latent assignments ***IS*** the knowledge base

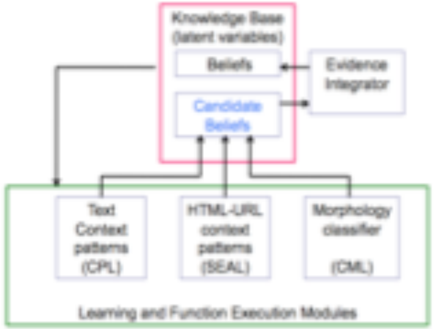
M' step:

- Each view-based learner retrains itself from the updated KB
- "context" methods create growing subsets of contexts

NELL Architecture



Never-Ending Language Learning



arg1_was_playing_arg2 arg2_megastar_arg1 arg2_icons_arg1
arg2_player_named_arg1 arg2_prodigy_arg1
arg1_is_the_tiger_woods_of_arg2 arg2_career_of_arg1
arg2_greats_as_arg1 arg1_plays_arg2 arg2_player_is_arg1
arg2_legends_arg1 arg1_announced_his_retirement_from_arg2
arg2_operations_chief_arg1 arg2_player_like_arg1
arg2_and_golfing_personalities_including_arg1 arg2_players_like_arg1
arg2_greats_like_arg1 arg2_players_are_steffi_graf_and_arg1
arg2_great_arg1 arg2_champ_arg1 arg2_greats_such_as_arg1
arg2_professionals_such_as_arg1 arg2_hit_by_arg1 arg2_greats_arg1
arg2_icon_arg1 arg2_stars_like_arg1 arg2_pros_like_arg1
arg1_retires_from_arg2 arg2_phenom_arg1 arg2_lesson_from_arg1
arg2_architects_robert_trent_jones_and_arg1 arg2_sensation_arg1
arg2_pros_arg1 arg2_stars_venus_and_arg1 arg2_hall_of_famer_arg1
arg2_superstar_arg1 arg2_legend_arg1 arg2_legends_such_as_arg1
arg2_players_is_arg1 arg2_pro_arg1 arg2_player_was_arg1
arg2_god_arg1 arg2_idol_arg1 arg1_was_born_to_play_arg2
arg2_star_arg1 arg2_hero_arg1 arg2_players_are_arg1
arg1_retired_from_professional_arg2 arg2_legends_as_arg1
arg2_autographed_by_arg1 arg2_champion_arg1

Predicate	Feature	Weight
mountain	LAST=peak	1.791
mountain	LAST=mountain	1.093
mountain	FIRST=mountain	-0.875
musicArtist	LAST=band	1.853
musicArtist	POS=DT_NNS	1.412
musicArtist	POS=DT_JJ_NN	-0.807
newspaper	LAST=sun	1.330
newspaper	LAST=university	-0.318
newspaper	POS=NN_NNS	-0.798
university	LAST=college	2.076
university	PREFIX=uc	1.999
university	LAST=state	1.992
university	LAST=university	1.745
university	FIRST=college	-1.381
visualArtMovement	SUFFIX=ism	1.282
visualArtMovement	PREFIX=journ	-0.234
visualArtMovement	PREFIX=budd	-0.253

Predicate	Web URL	Extraction Template
academicField	http://scholendow.ais.msu.edu/student/ScholSearch.Asp	[X] -
athlete	http://www.quotes-search.com/d_occupation.aspx?o=+athlete	<a href='d_author.aspx?a=[X]'>-
bird	http://www.michaelforsberg.com/stock.html	<option>[X]</option>
bookAuthor	http://lifebehindthecurve.com/	[X] by [Y] –

If coupled learning is the key idea, how can we get new coupling constraints?

Building the Knowledge Graph by Reading

1. Classify noun phrases (NP's) by category
2. Classify NP pairs by relation
3. Discover rules to predict new relation instances

Key Idea 2: Discover New Coupling Constraints

- first order, probabilistic horn clause constraints

0.93 athletePlaysSport(?x,?y) :- athletePlaysForTeam(?x,?z),
teamPlaysSport(?z,?y)

– connects previously uncoupled relation predicates
– infers new beliefs for KB

Example Learned Horn Clauses

0.95 athletePlaysSport(?x,basketball) :- athleteInLeague(?x,NBA)

0.93 athletePlaysSport(?x,?y) :- athletePlaysForTeam(?x,?z)
teamPlaysSport(?z,?y)

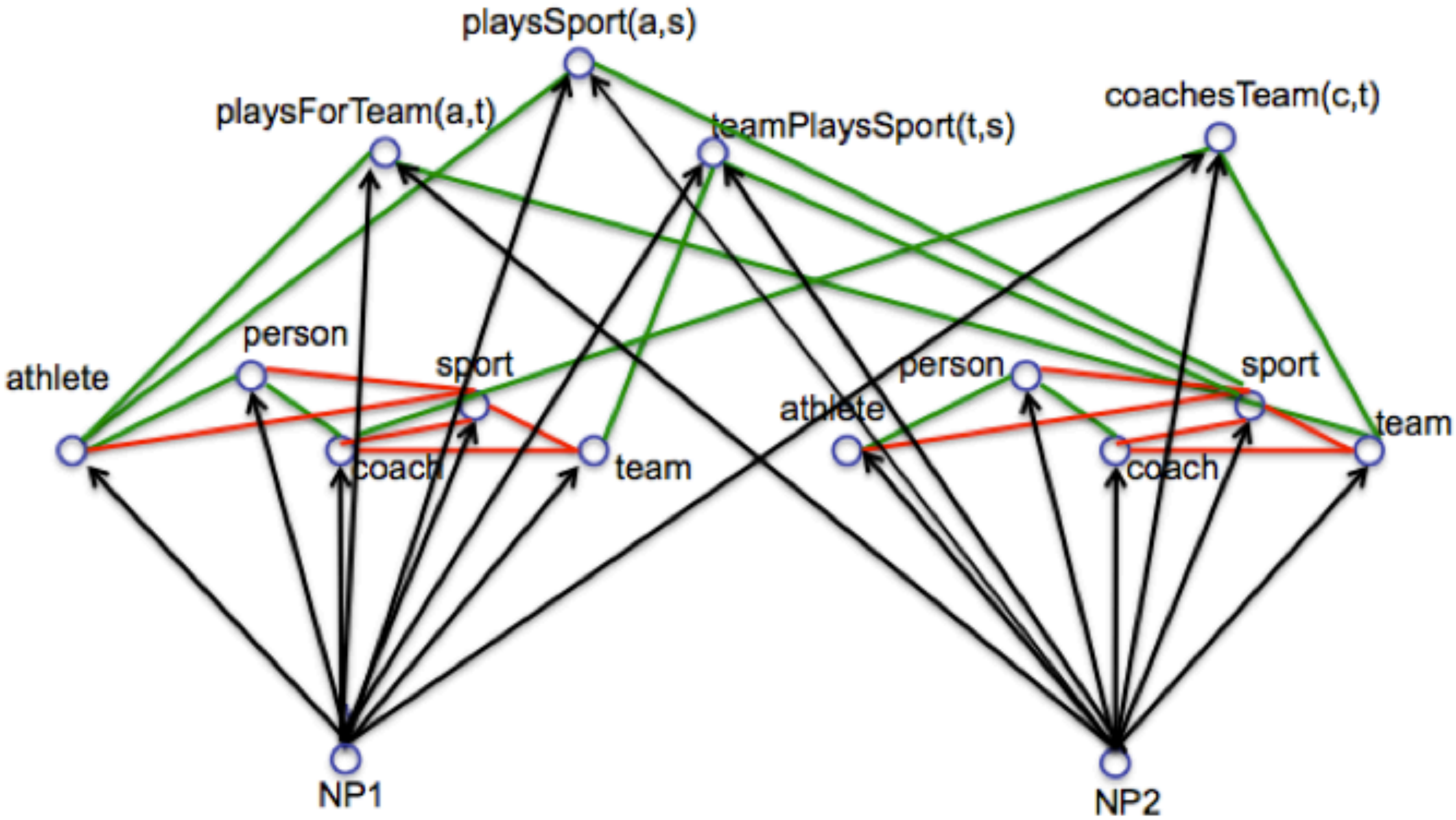
0.91 teamPlaysInLeague(?x,NHL) :- teamWonTrophy(?x,Stanley_Cup)

0.90 athleteInLeague(?x,?y):- athletePlaysForTeam(?x,?z),
teamPlaysInLeague(?z,?y)

0.88 cityInState(?x,?y) :- cityCapitalOfState(?x,?y),
cityInCountry(?y,USA)

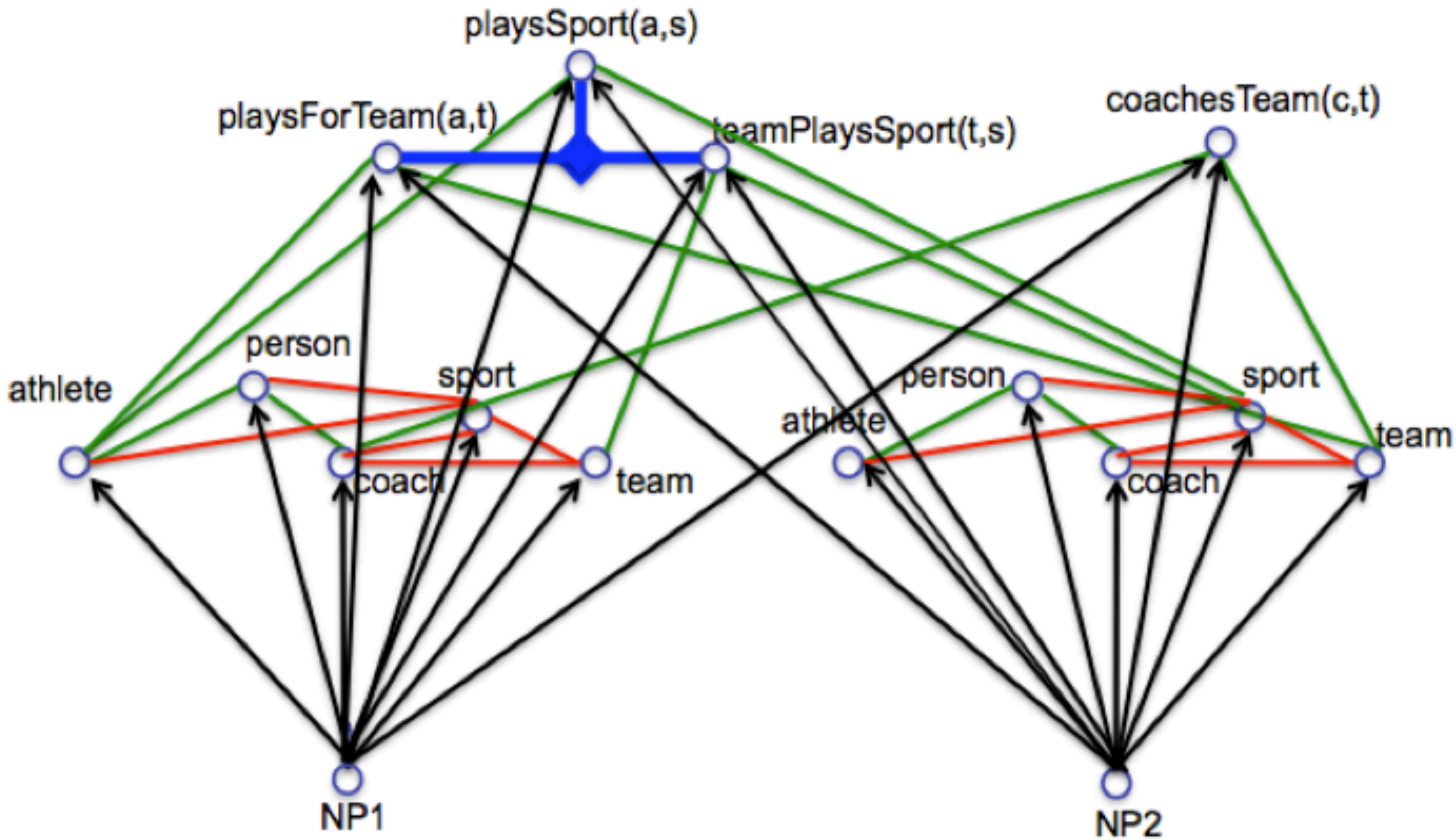
0.62* newspaperInCity(?x,New_York) :- companyEconomicSector(?x,media),
generalizations(?x,blog)

Learned Probabilistic Horn Clause Rules

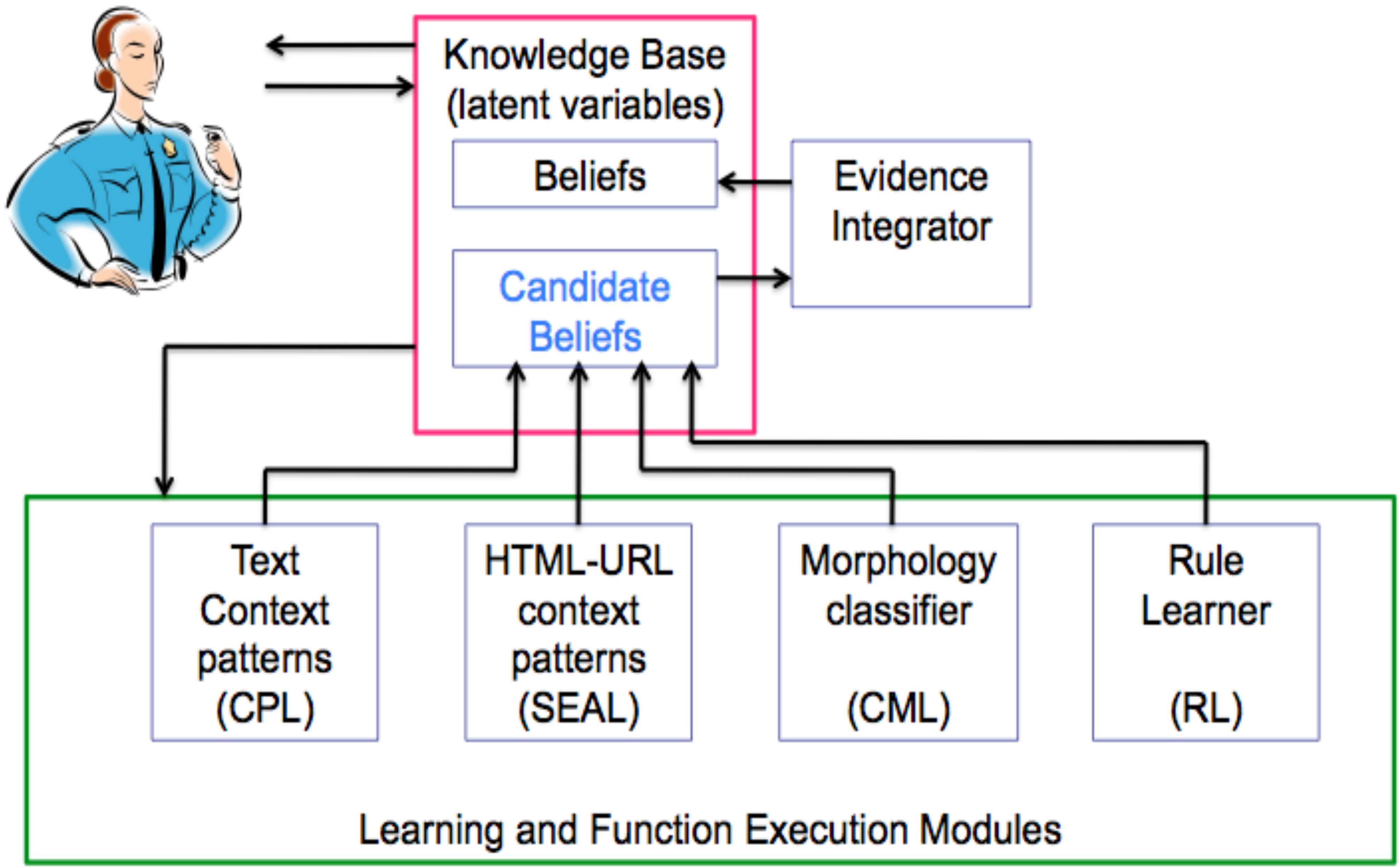


Learned Probabilistic Horn Clause Rules

0.93 playsSport(?x,?y) ← playsForTeam(?x,?z), teamPlaysSport(?z,?y)



NELL Architecture

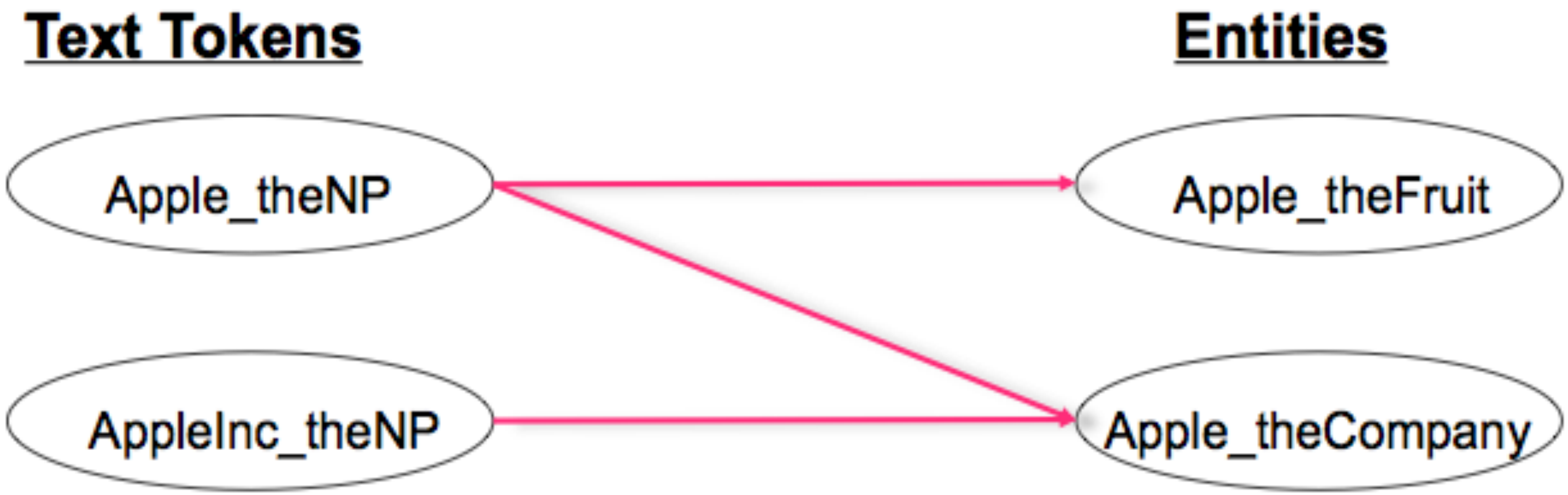


Building the Knowledge Graph by Reading

1. Classify noun phrases (NP’s) by category
2. Classify NP pairs by relation
3. Discover rules to predict new relation instances
4. Learn which NP’s (co)refer to which latent concepts

Distinguish Text Tokens from Entities

[Jayant Krishnamurthy]



Coreference Resolution:

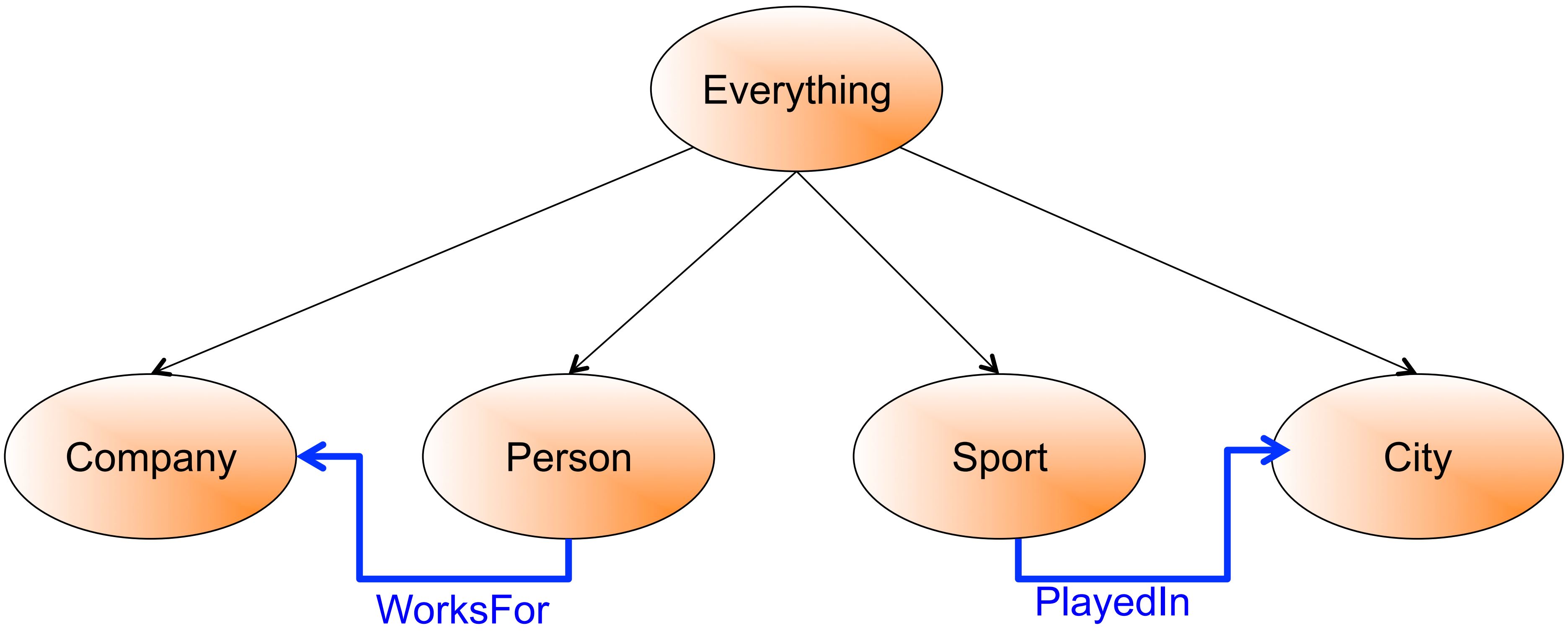
- Co-train classifier to predict coreference as f(string similarity, extracted beliefs)
- Small amount of supervision: ~10 labeled coreference decisions
- Cluster tokens using f as similarity measure

Building the Knowledge Graph by Reading

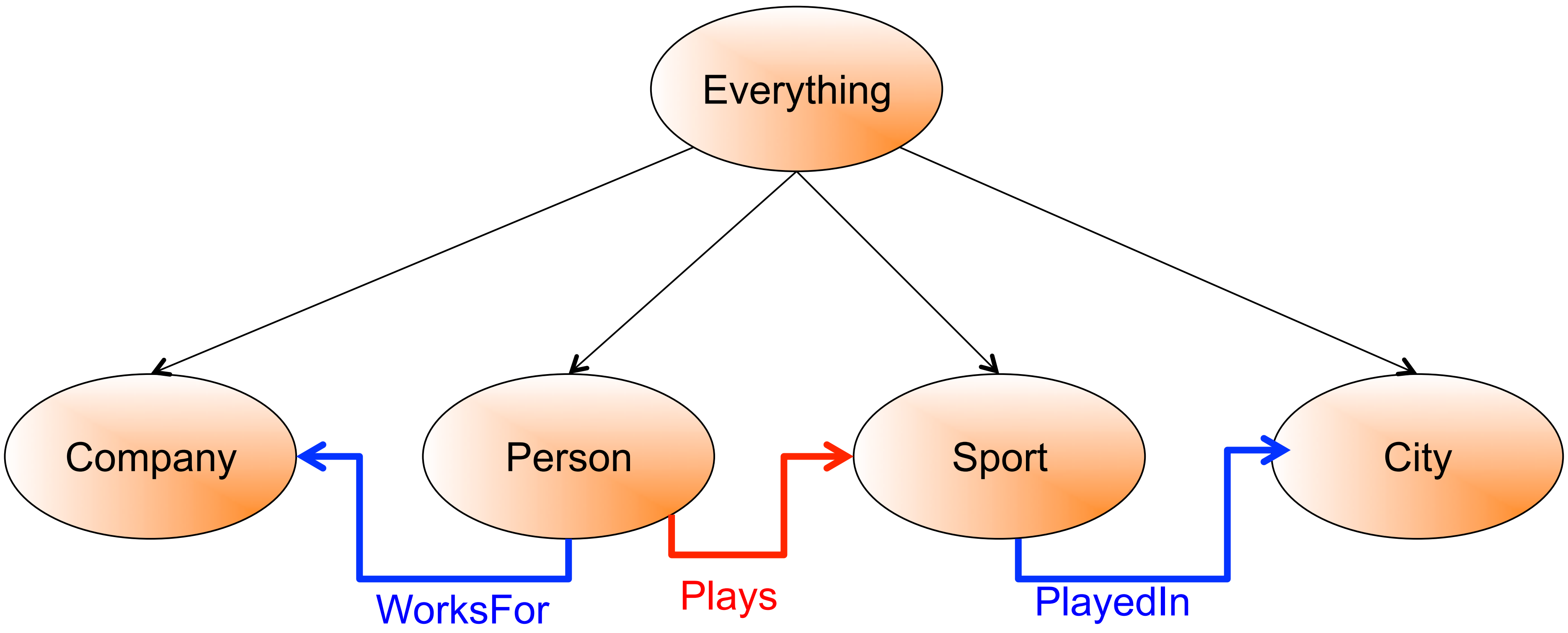
1. Classify noun phrases (NP’s) by category
2. Classify NP pairs by relation
3. Discover rules to predict new relation instances
4. Learn which NP’s (co)refer to which latent concepts
5. Discover new relations to extend ontology

Key Idea 3: Automatically Extending the Ontology

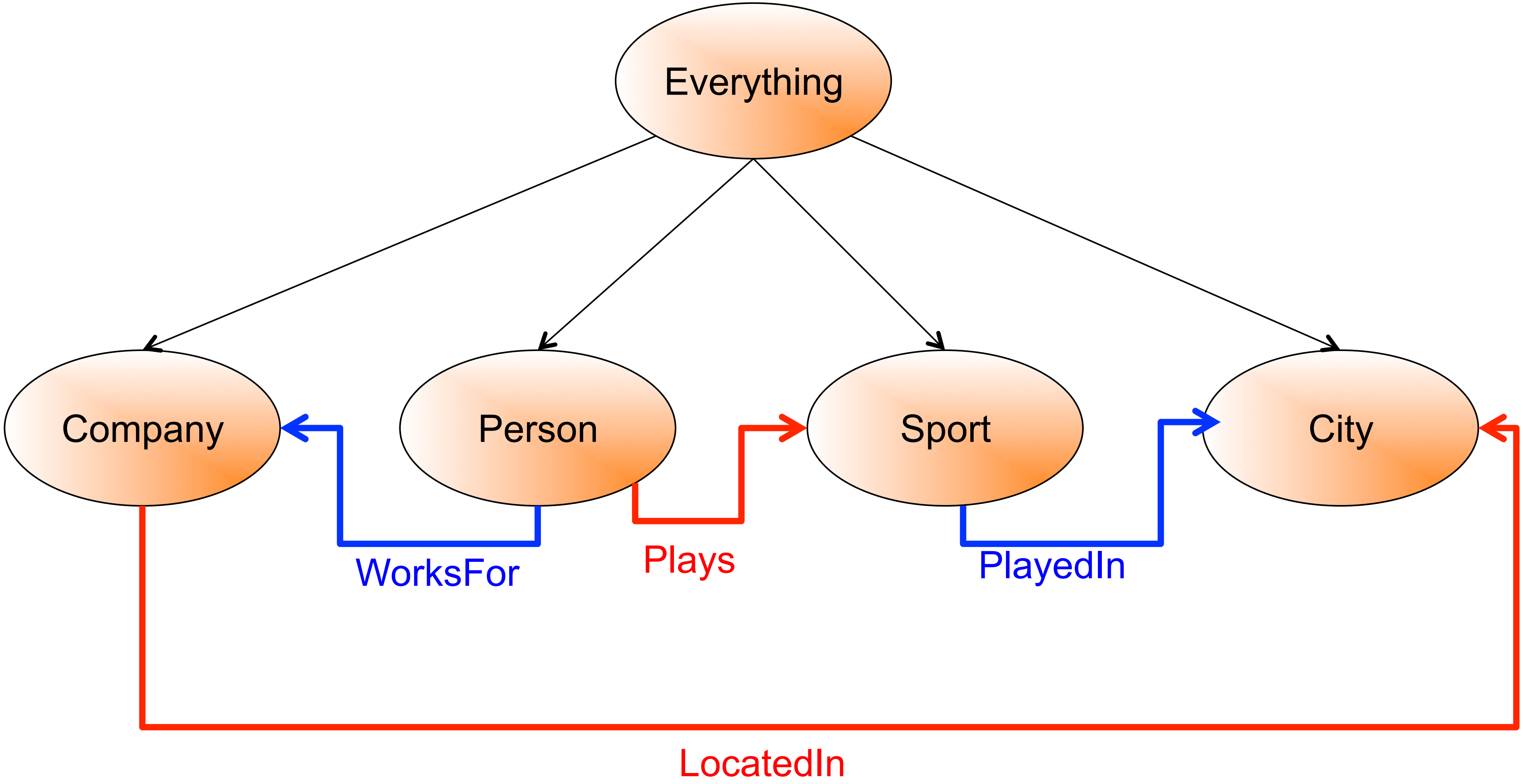
OntExt (Ontology Extension)



OntExt (Ontology Extension)



OntExt (Ontology Extension)



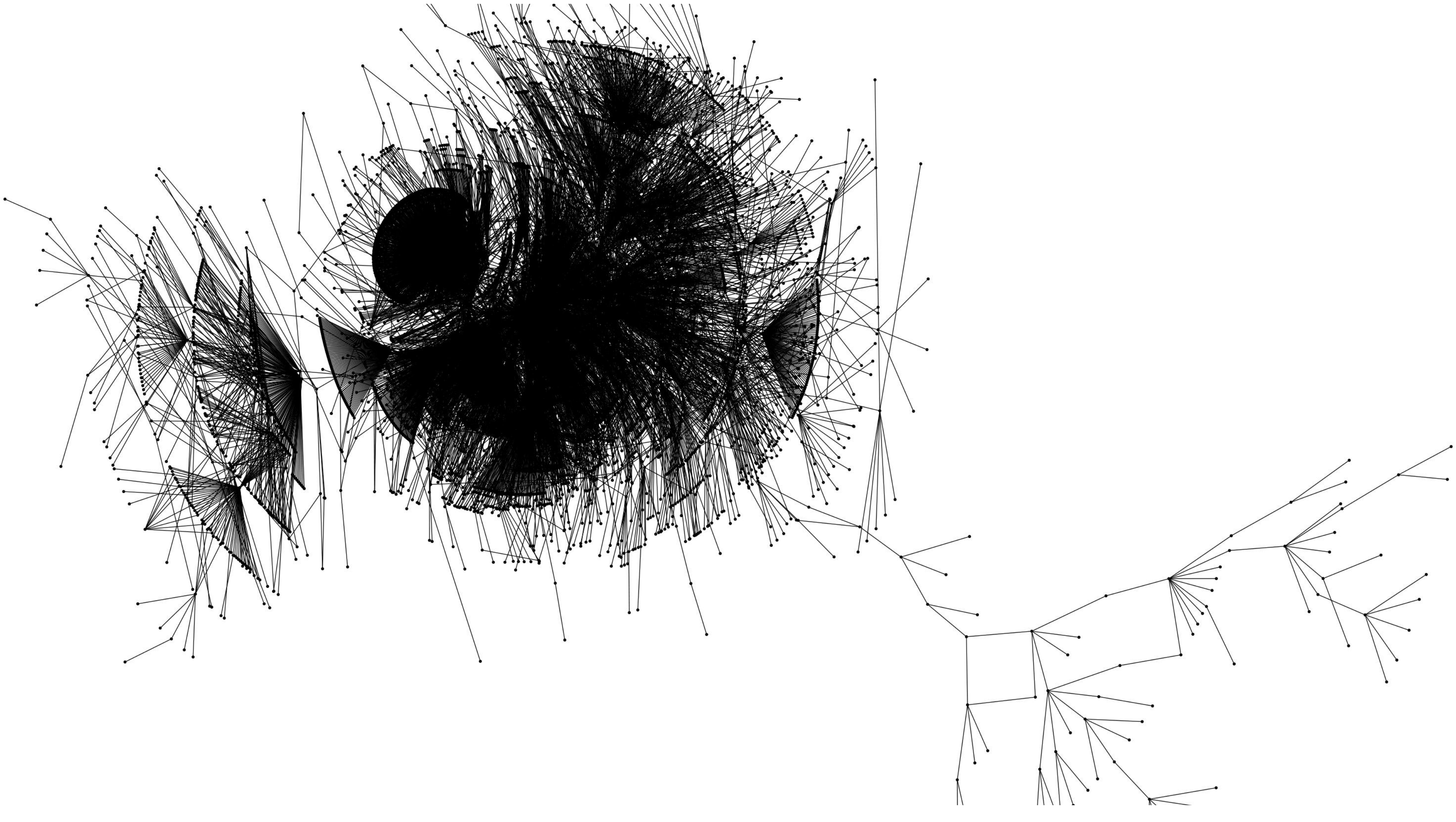
Prophet

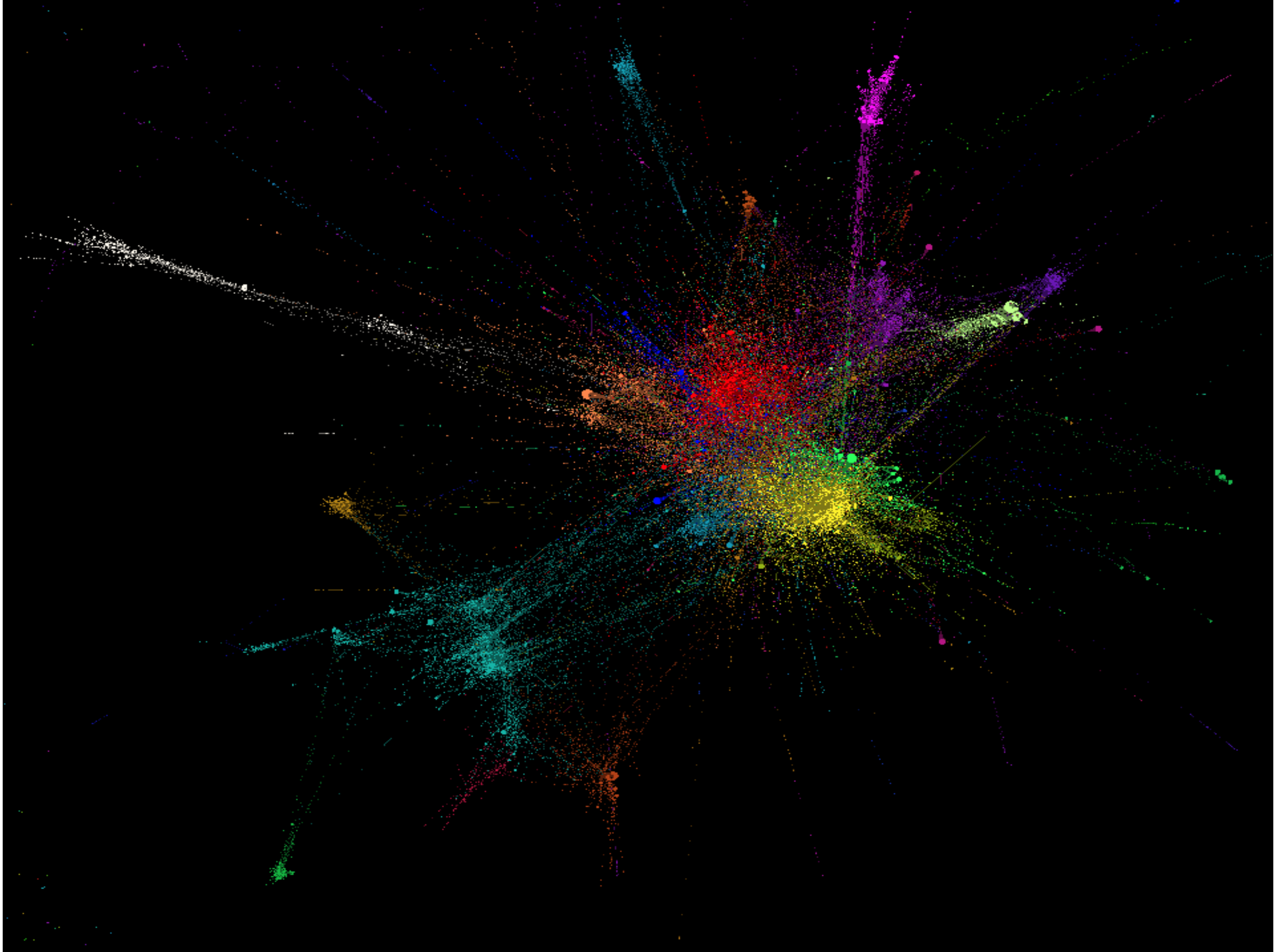
Mining the Graph representing NELL's KB to:

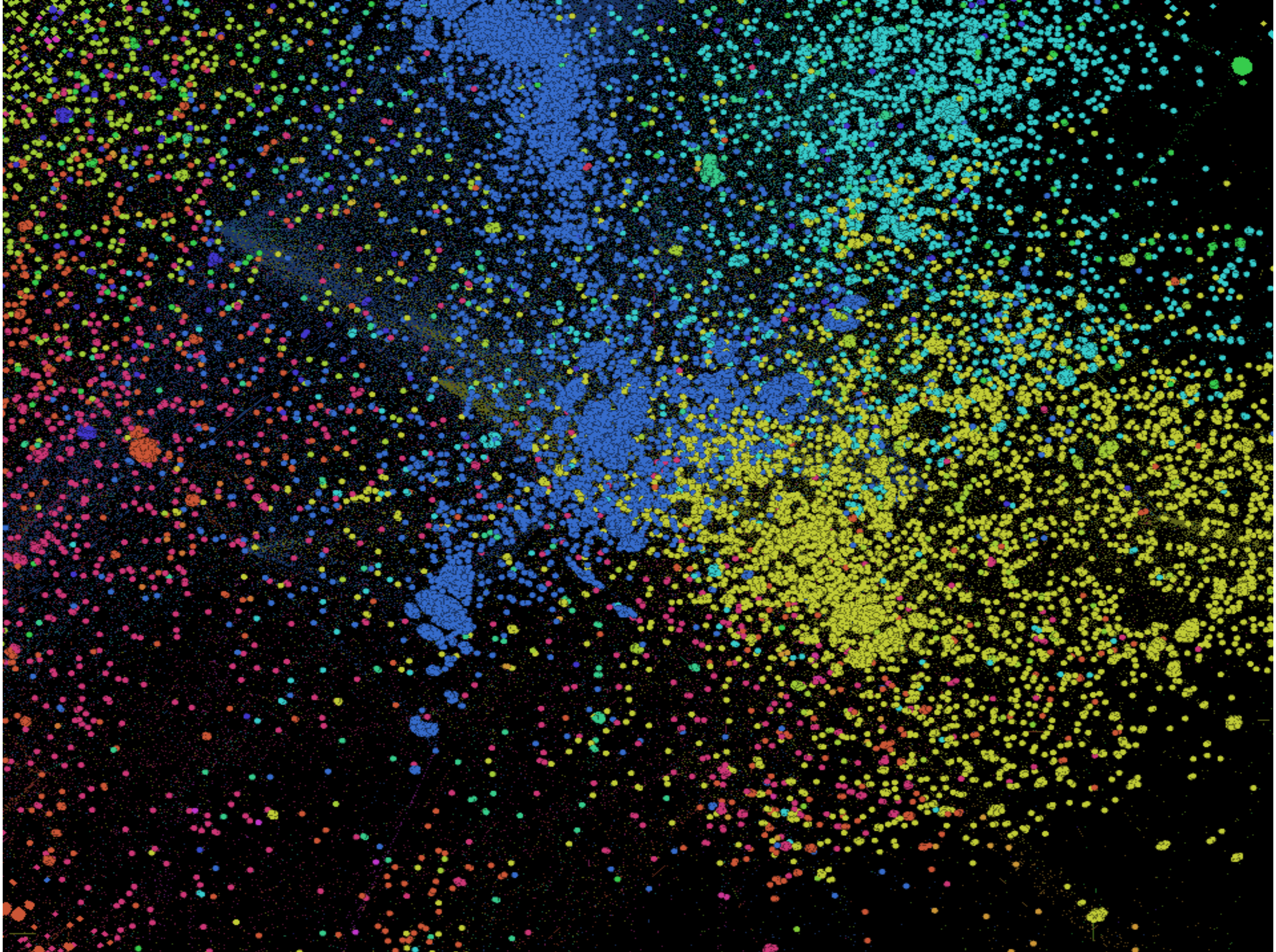
1. Extend the KB by predicting new relations (edges)that might exist between pairs of nodes;
2. Induce inference rules;
3. Identify misplaced edges which can be used by NELL as hints to identify wrong connections between nodes (wrong fats);

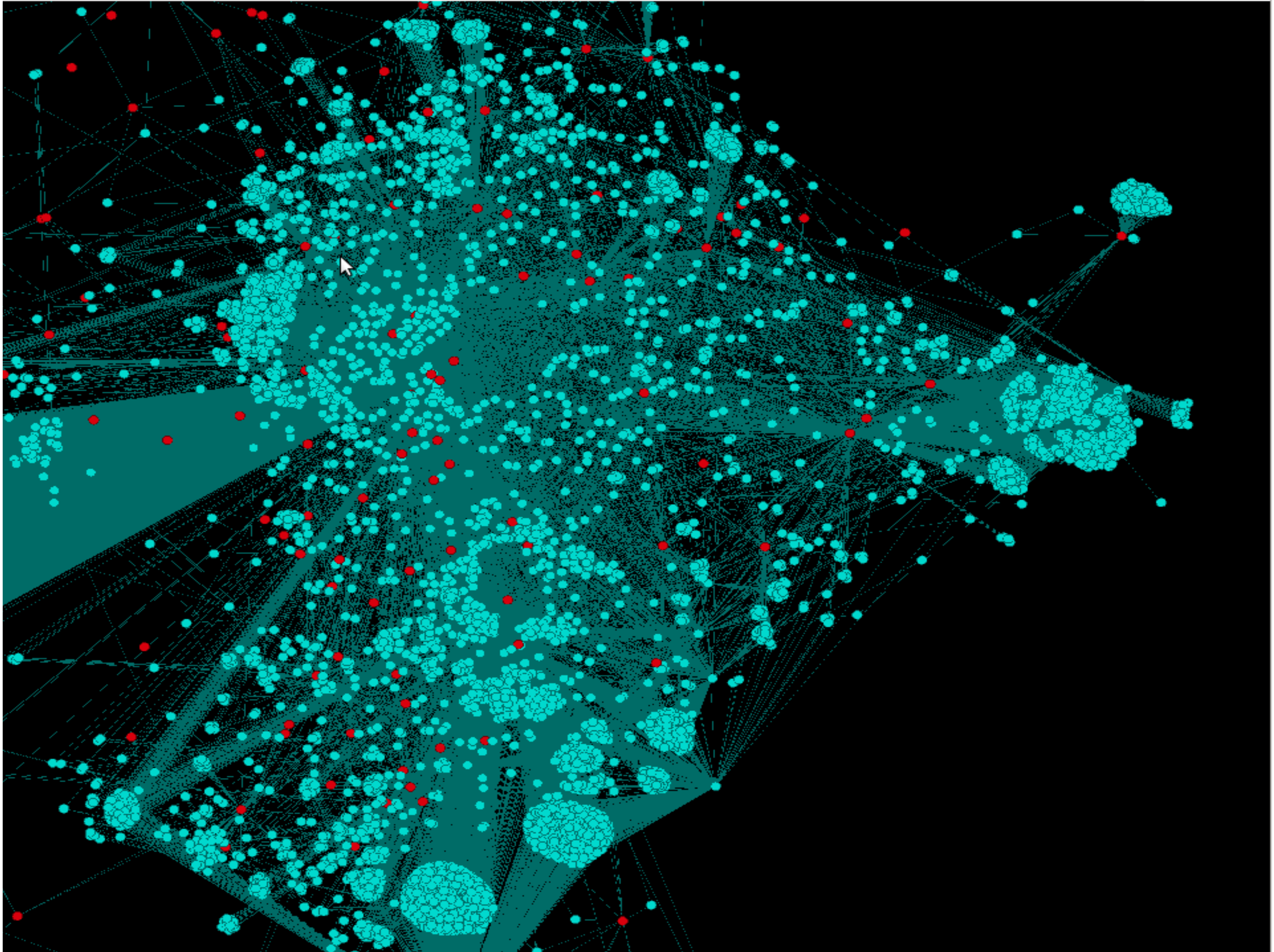
Prophet

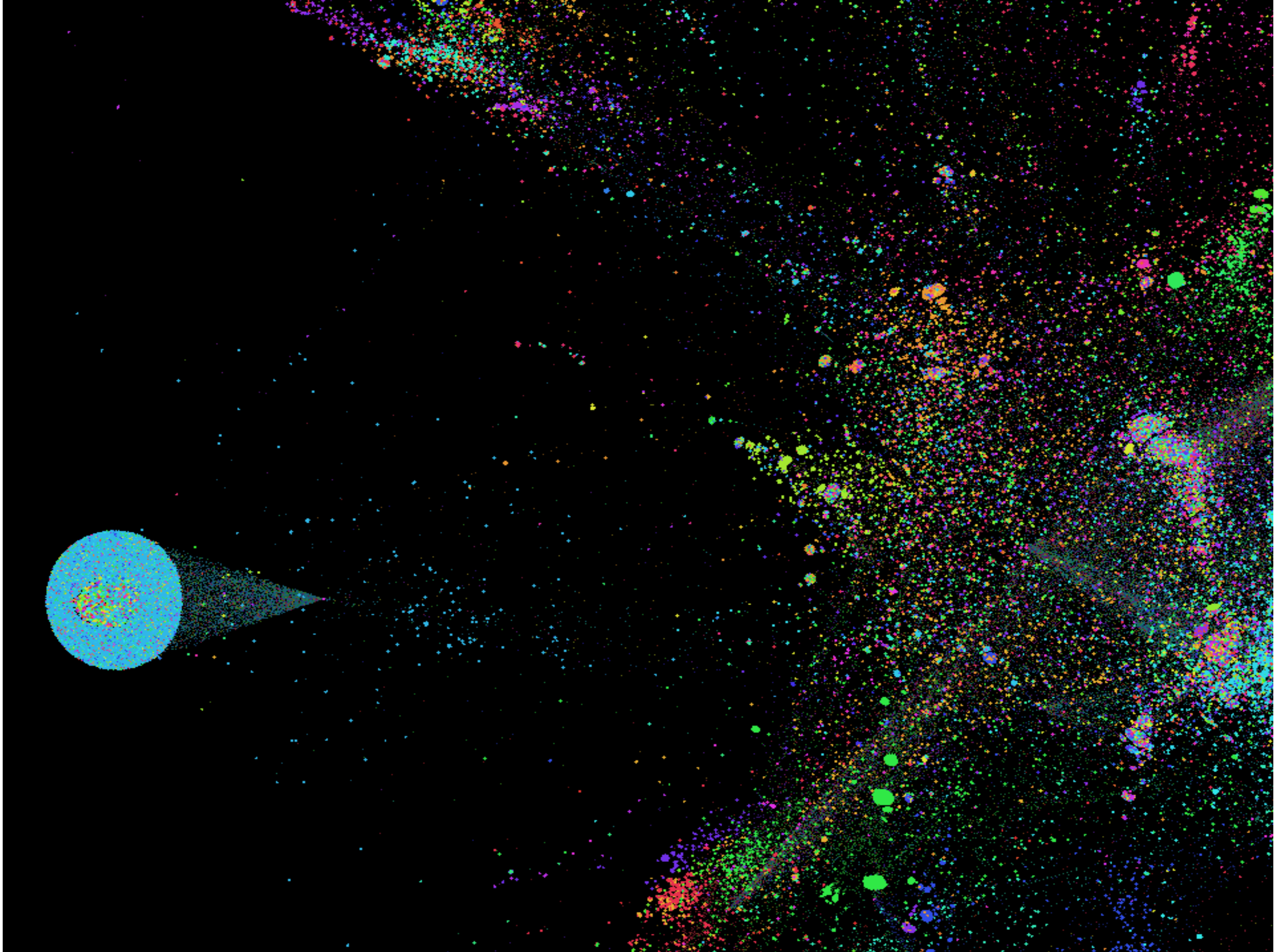
Find open triangles in the Graph





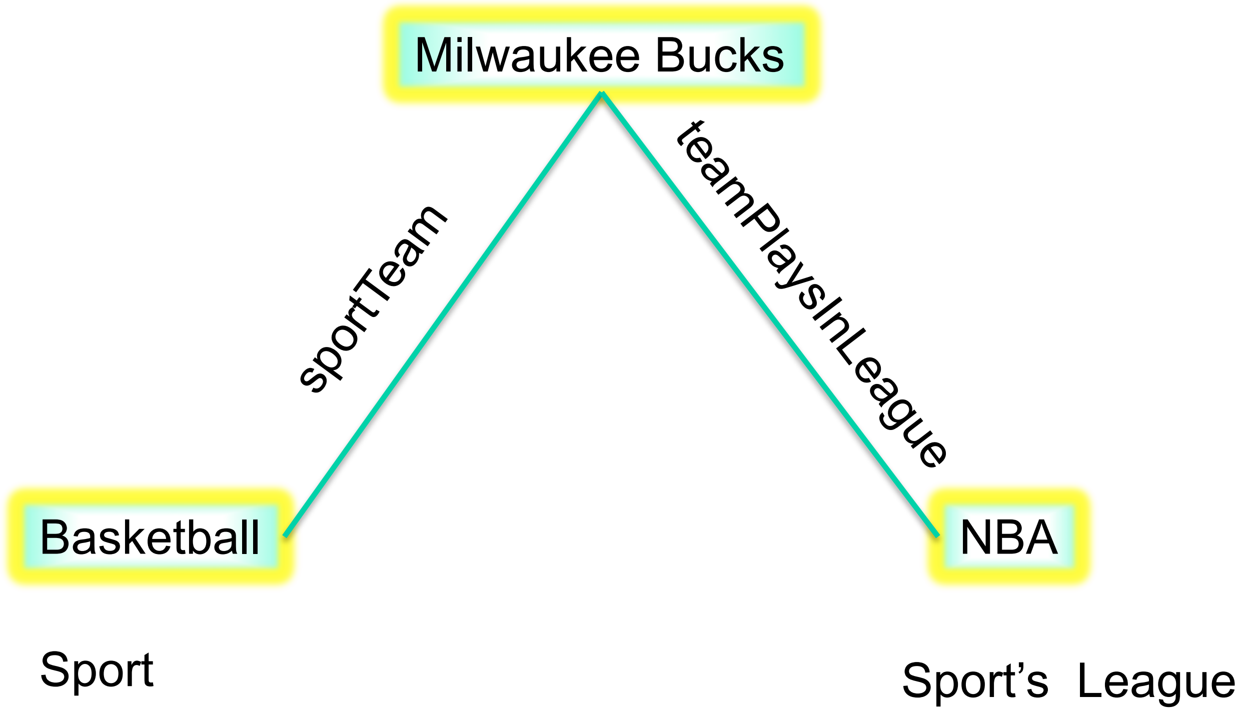






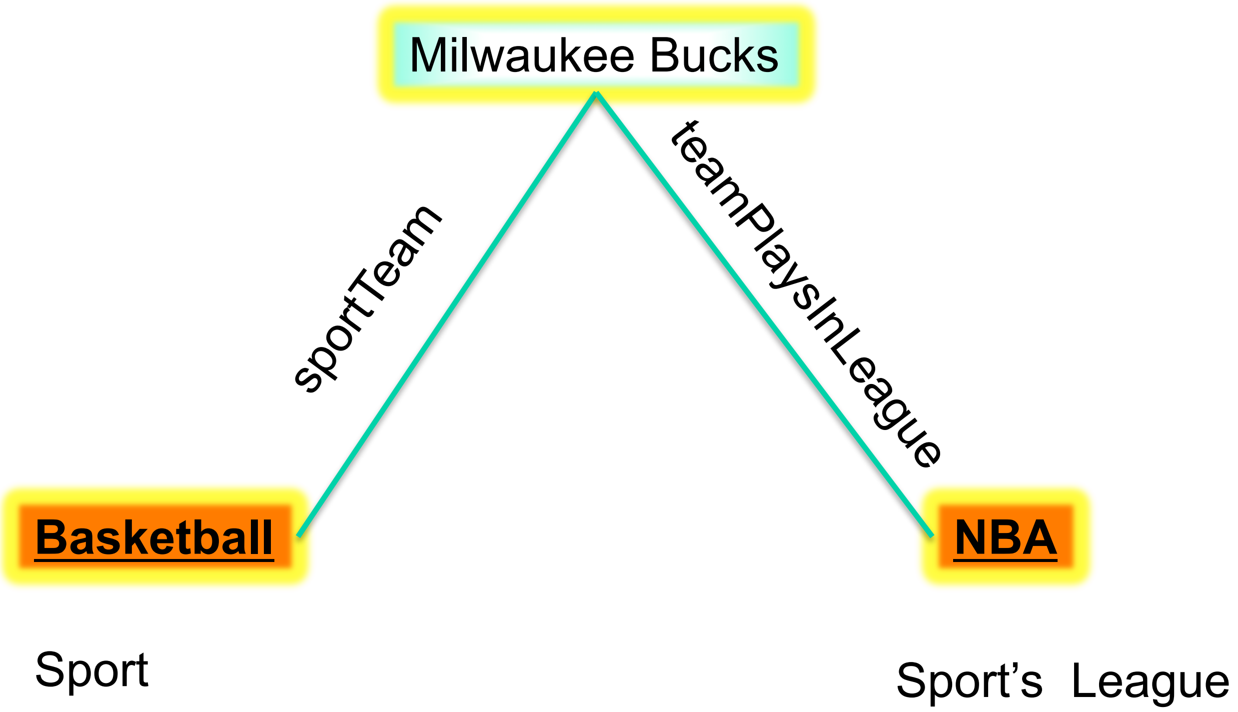
Prophet

open triangles



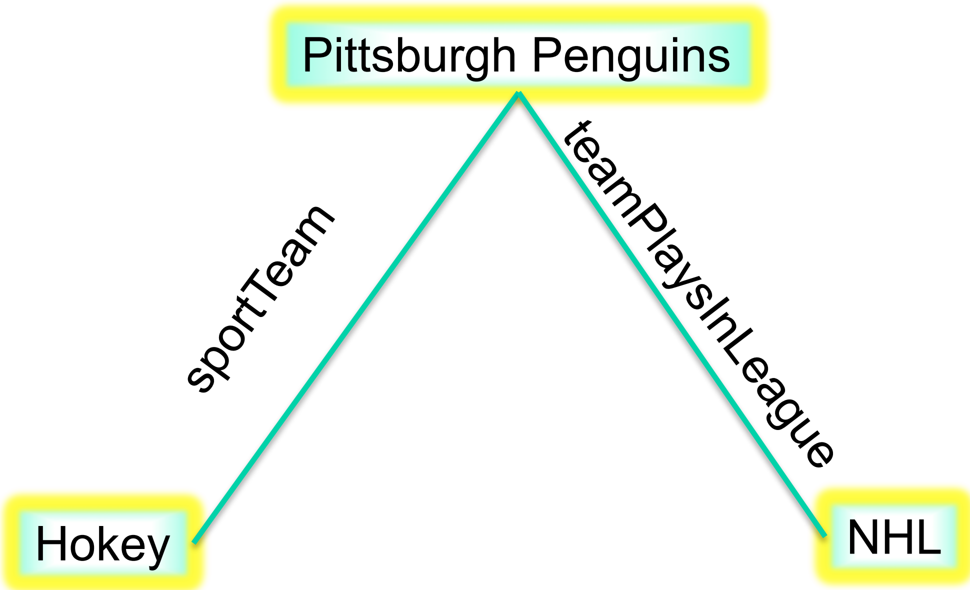
Prophet

open triangles



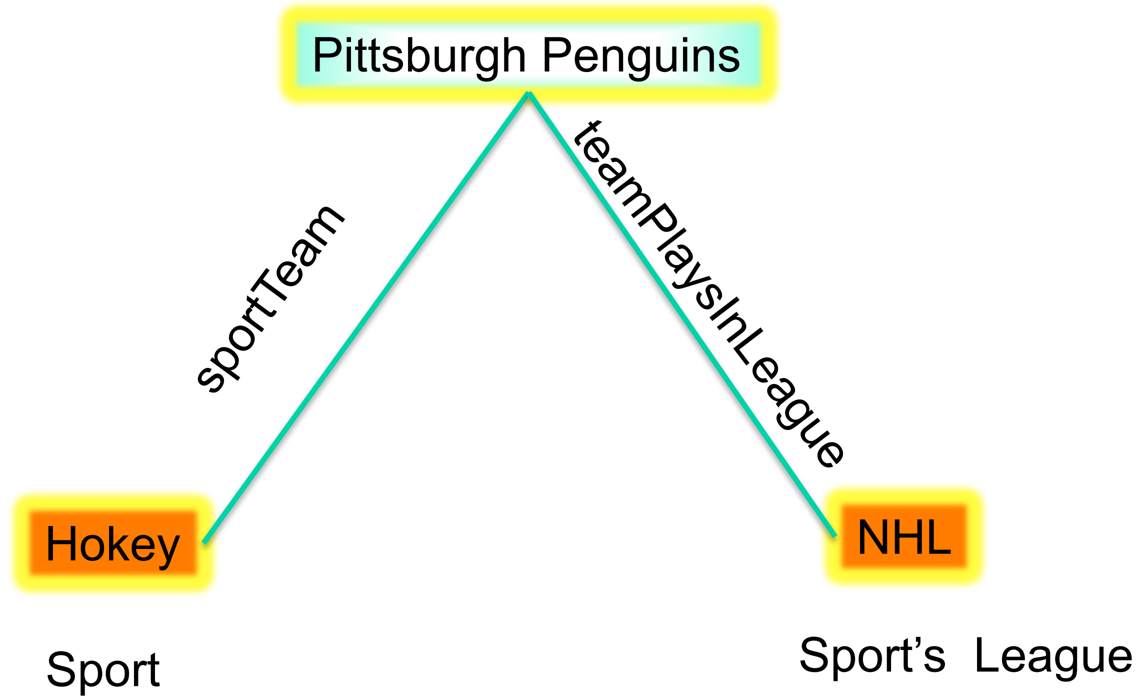
Prophet

open triangles



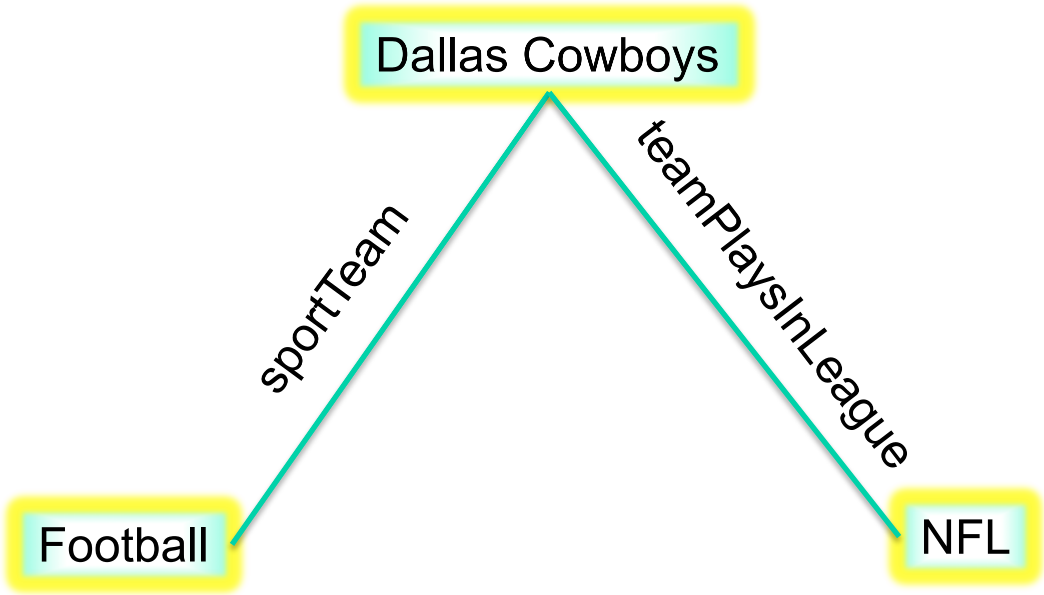
Prophet

open triangles



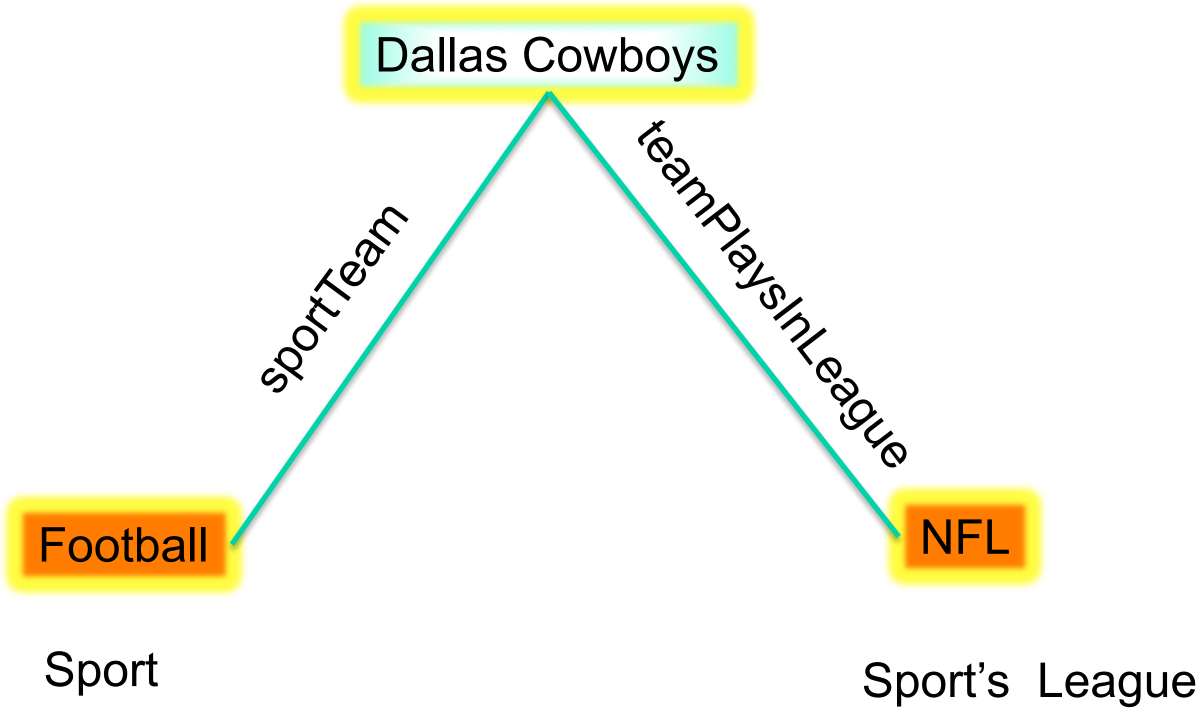
Prophet

open triangles



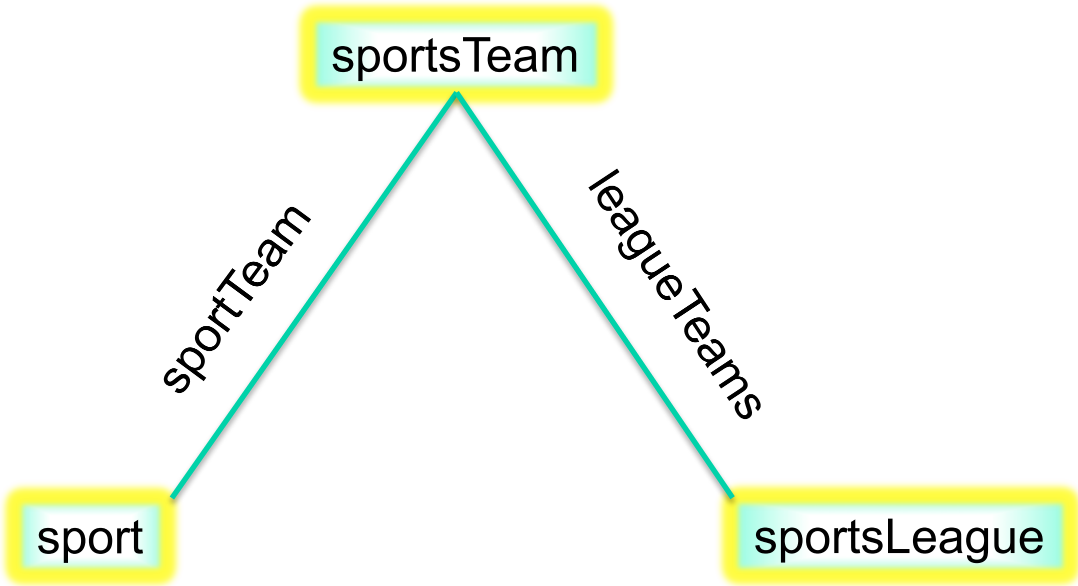
Prophet

open triangles



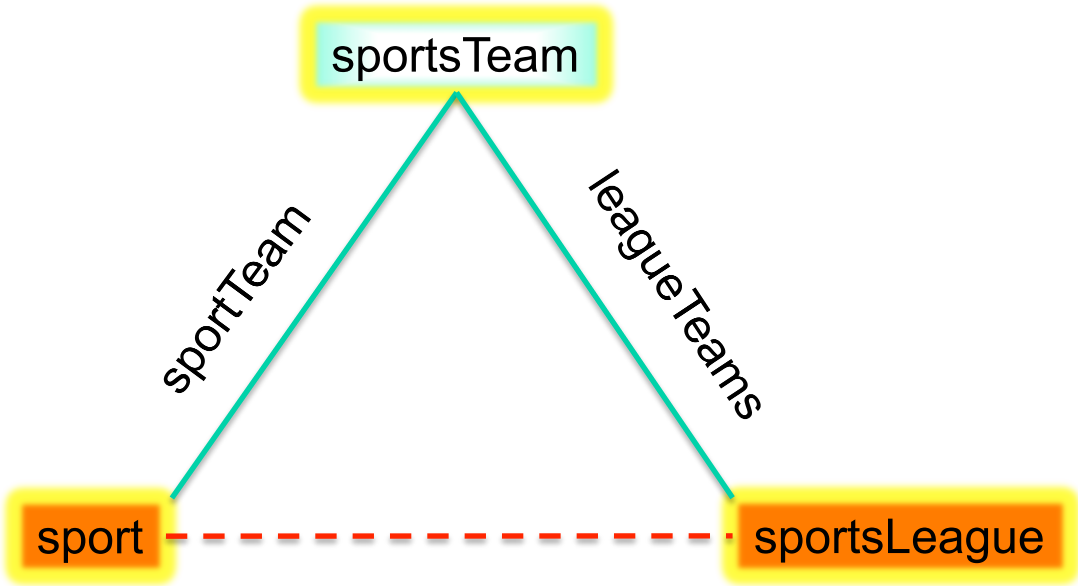
Prophet

open triangles



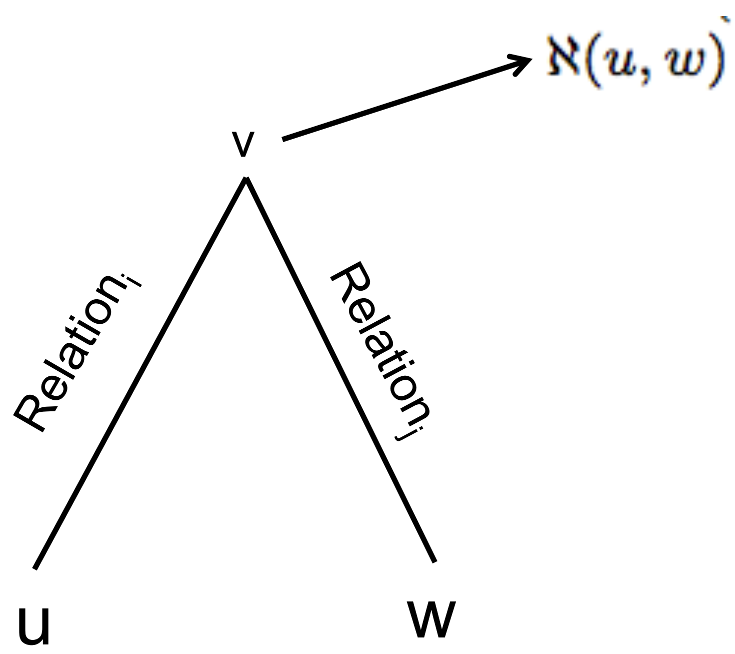
Prophet

open triangles



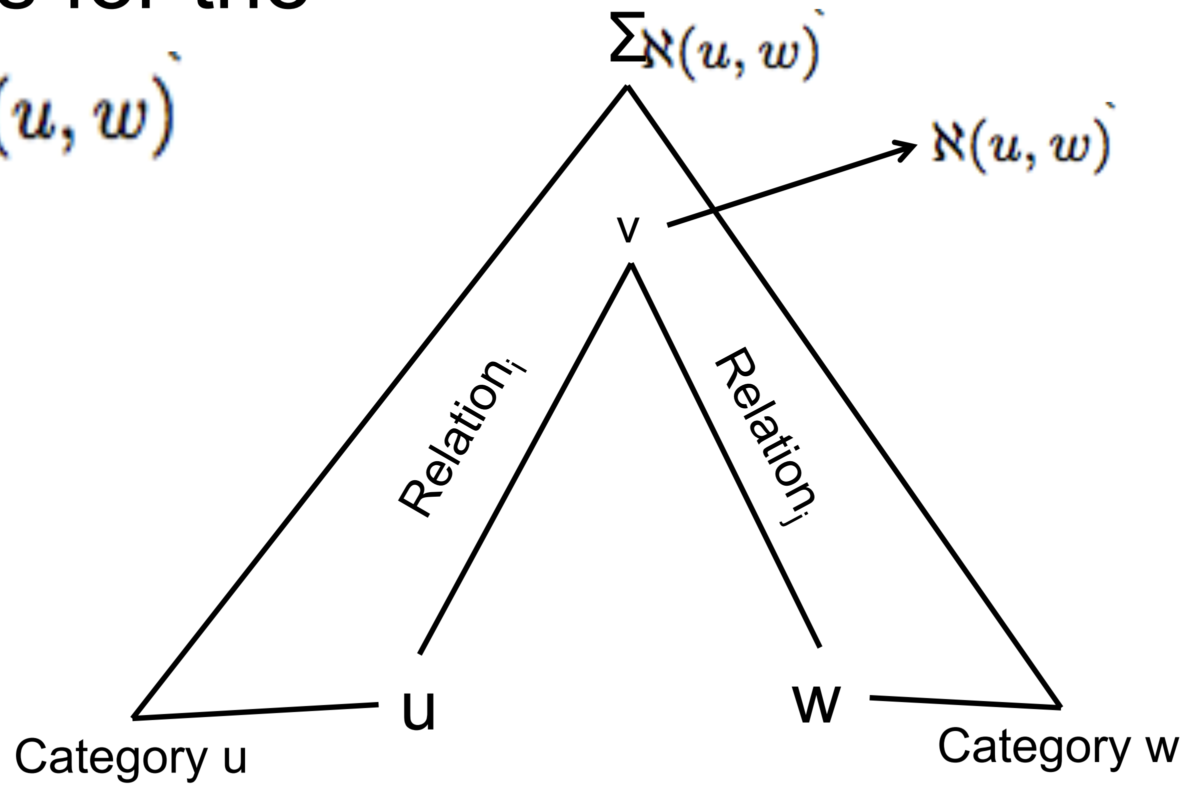
Prophet

- Compute the number of common neighbors $\aleph(u, w)$



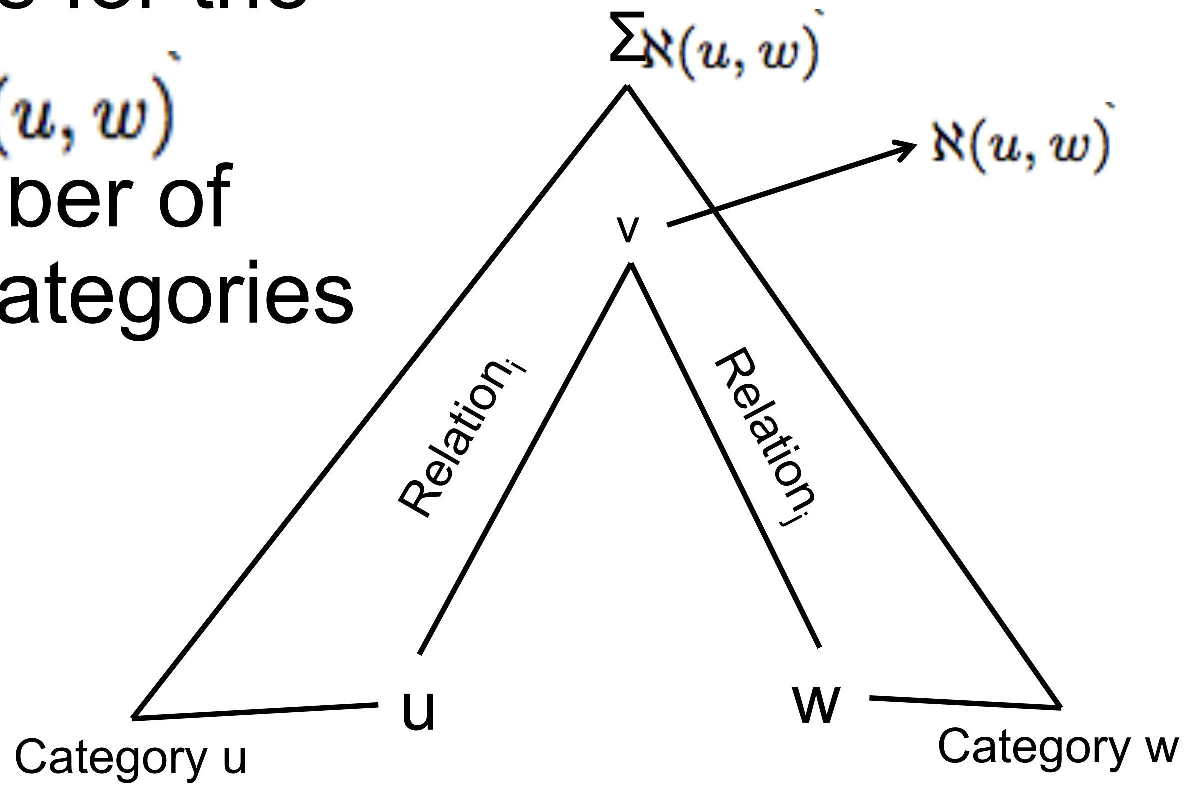
Prophet

- Compute the number of common neighbors $\aleph(u, w)$
- Compute the cumulative number of instances for the categories nodes $\bar{\aleph}(u, w)$



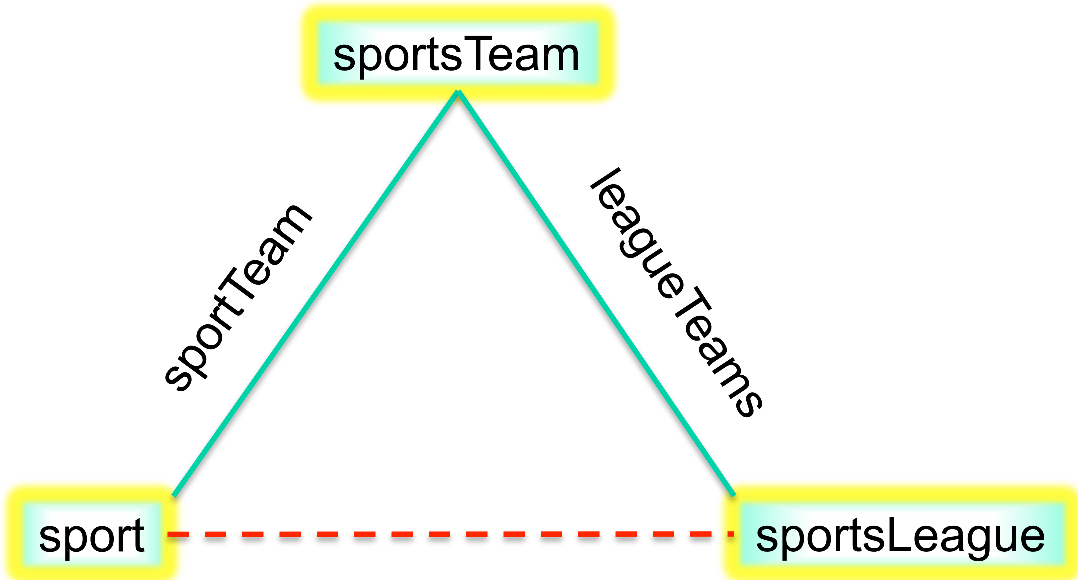
Prophet

- Compute the number of common neighbors $\aleph(u, w)$
- Compute the cumulative number of instances for the categories nodes $\bar{\aleph}(u, w)$
- $N_{\Lambda_C(u_C, w_C)}$ is the number of open triangles for categories u and w.



Prophet

$$\aleph_c(u_c, w_c) = \sum \aleph(u, w) - N_{\Lambda_c(u_c, w_c)}$$

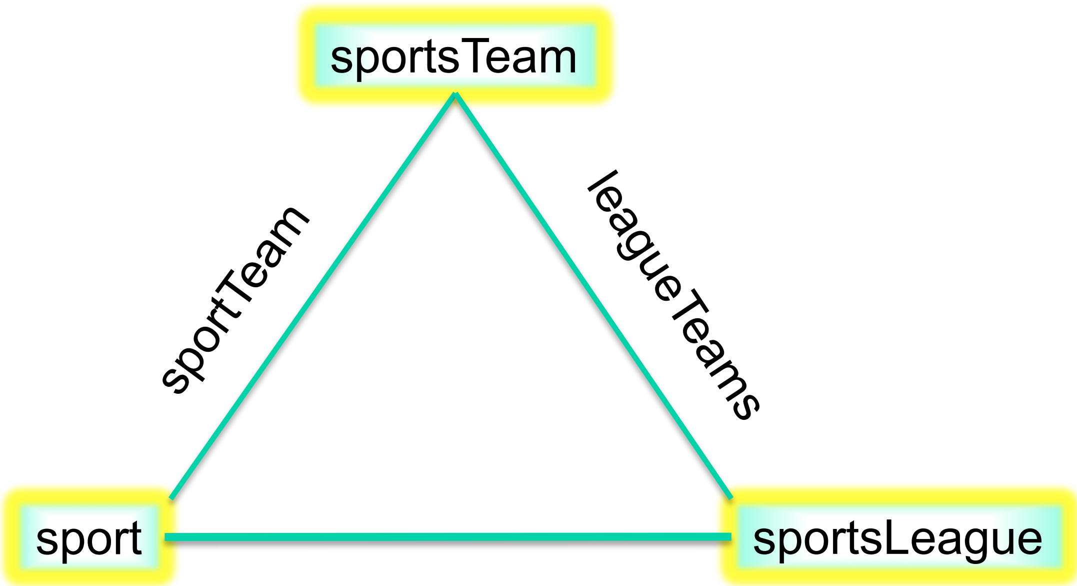


Prophet

$$\aleph_c(u_c, w_c) = \sum \aleph(u, w) - N_{\Lambda_c(u_c, w_c)}$$

If $\aleph_c(u_c, w_c)$ ξ then create the new relation

ξ = 10 (empirically)



Prophet

$$\aleph_c(u_c, w_c) = \sum \aleph(u, w) - N_{\Lambda_c(u_c, w_c)}$$

If $\aleph_c(u_c, w_c)$ ξ then create the new relation

ξ = 10 (empirically)

Name the new relation based on ReVerb

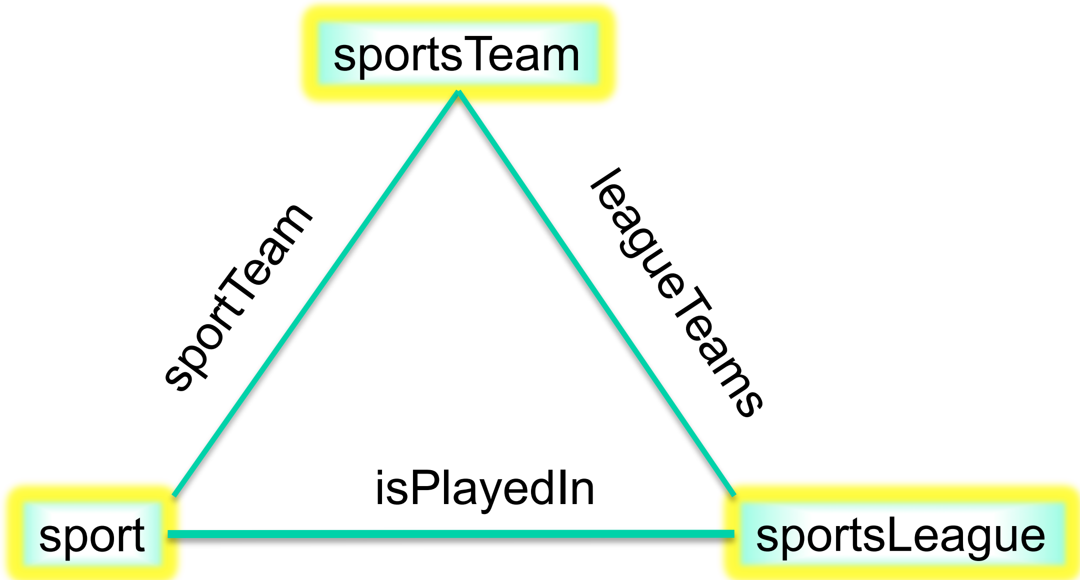


Table 1. Real datasets description and query time. Respectively the number of nodes ($|V|$), edges ($|E|$), triangles ($|\Delta|$), insertion time in *GraphDB-Tree* (I), time to query all triangles (Δ) , the transitivity ratio in each network ($T(G)$), Size in MB to store networks using *GraphDB-Tree* and time to query all triangles in R

name	$\|V\|$	$\|E\|$	$\|\Delta\|$	I	Δ	$T(G)$	size	R
ca-GrQc	5,242	28,980	48,260	1	1	0.6298	0.47	1
wiki-Vote	7,115	201,525	608,389	1	8	0.1255	1.78	22
Ca-HepPh	12,007	237,001	3358499	1	7	0.1457	4.21	18
Cit-HepTh	27,770	704,610	1,478,735	1	14	0.1196	6.38	60
Email-EuAll	265,214	730,052	267,313	1	36	0.0041	12.3	925
RoadNet-ca	1,965,206	5,533,214	120,676	3	9	0.0604	92.1	12
Web-google	875,713	8,643,937	13,391,655	3	83	0.0552	90.7	7021
WikiTalk	2,394,385	9,319,131	9,203,519	5	7,523	0.0011	132	43200
As-skitter	1,696,415	22,190,495	28,769,868	9	7,523	0.0054	219.5	+21600
Cit-Patents	3,774,768	33,037,896	7,515,023	15	121	0.0671	357	308
soc-Pokec	1,632,803	44,603,930	32,557,458	28	68,411	0.0161	398.2	9271
Com-LiveJournal	3,997,962	69,362,379	177.820.130	39	3,410	0.1154	654.8	19740
Soc-LiveJournal	4,847,570	86,054,328	285,030,584	42	13,382	0.2882	809.1	overflow
Com-Orkut	3,072,441	234,370,167	633,319,568	112	80,492	0.2303	1974.4	overflow

Navarro et al., 2013

How to Extract New Relations?

Proposed Approach - OntExt

- Traditional IE + Open IE
- Cluster context patterns which are semantically similar although they may be lexically dissimilar
- Scalability: Context-pattern X Context-pattern matrix
- Classifier learns to distinguish valid relations from semantically invalid relations

OntExt

Input:

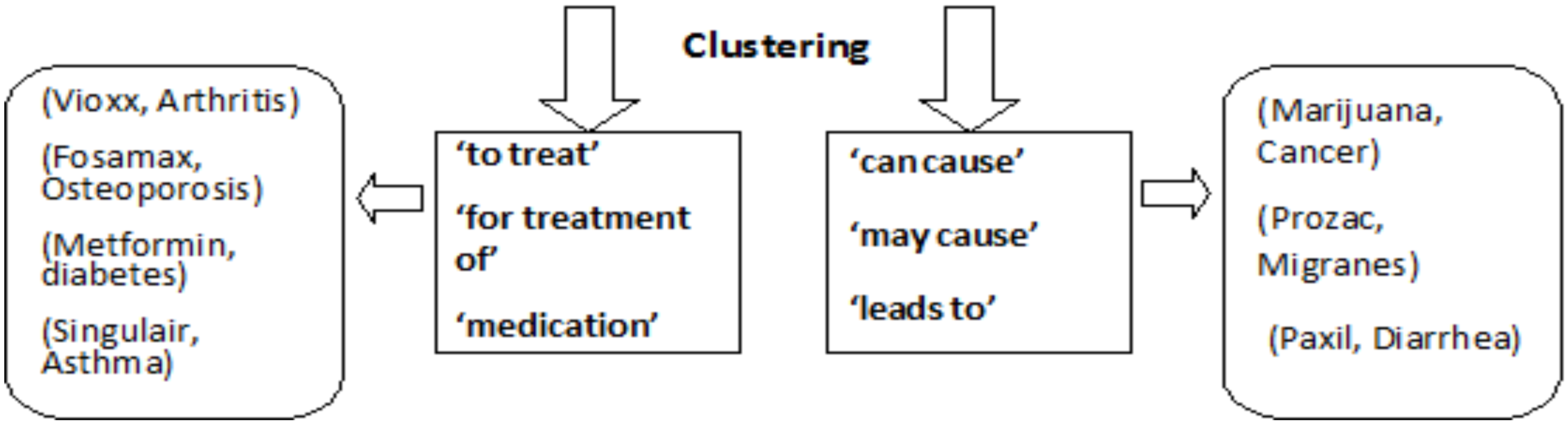
Preprocessed 2 billion sentences from ClueWeb09 data [Callan and Hoy, 2009].

Category instances (e.g. city(Ottawa), city(Berlin), country(Canada), etc.) are used to find context patterns

Context x Context Matrix

OntExt

Contexts/ Contexts	may cause	can cause	can lead to	to treat	for treatment of	medication
may cause	0.176	0.074	0.030	0.015	0.011	0.000
can cause	0.051	0.150	0.039	0.018	0.013	0.010
can lead to	0.034	0.064	0.189	0.019	0.021	0.018
to treat	0.006	0.011	0.007	0.109	0.043	0.015
for treatment of	0.005	0.008	0.008	0.045	0.086	0.023
medication	0.000	0.011	0.009	0.030	0.036	0.111



NELL: sample of self-added relations

- athleteWonAward
- animalEatsFood
- languageTaughtInCity
- clothingMadeFromPlant
- beverageServedWithFood
- fishServedWithFood
- athleteBeatAthlete
- athleteInjuredBodyPart
- arthropodFeedsOnInsect
- animalEatsVegetable
- plantRepresentsEmotion
- foodDecreasesRiskOfDisease
- clothingGoesWithClothing
- bacteriaCausesPhysCondition
- buildingMadeOfMaterial
- emotionAssociatedWithDisease
- foodCanCauseDisease
- agriculturalProductAttractsInsect
- arteryArisesFromArtery
- countryHasSportsFans
- bakedGoodServedWithBeverage
- beverageContainsProtein
- animalCanDevelopDisease
- beverageMadeFromBeverage

Building the Knowledge Graph by Reading

1. Classify noun phrases (NP's) by category
2. Classify NP pairs by relation
3. Discover rules to predict new relation instances
4. Learn which NP's (co)refer to which latent concepts
5. Discover new relations to extend ontology
6. Learn to infer relation instances via targeted random walks

CityLocatedInCountry(Pittsburgh) = ?

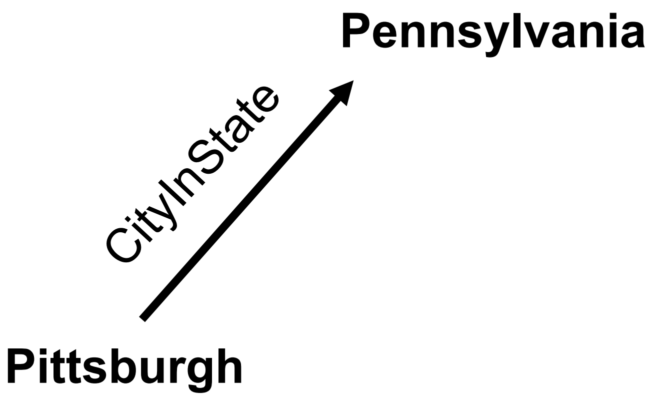
[Lao, Mitchell, Cohen, *EMNLP* 2011]

Pittsburgh

Feature = Typed Path	**Feature Value**	**Logistic Regresssion Weight**
CityInState, CityInstate^{-1}, CityLocatedInCountry	0.32	

CityLocatedInCountry(Pittsburgh) = ?

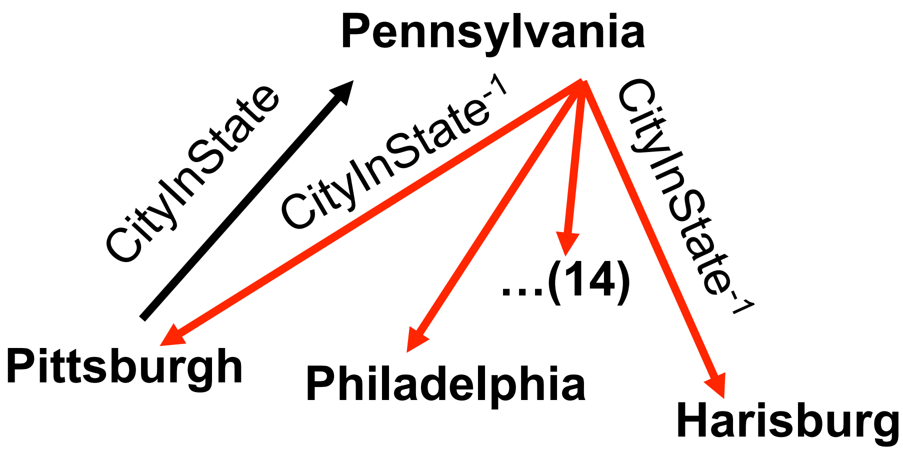
[Lao, Mitchell, Cohen, *EMNLP* 2011]



Feature = Typed Path	**Feature Value**	**Logistic Regresssion Weight**
CityInState, CityInstate^{-1}, CityLocatedInCountry		0.32

CityLocatedInCountry(Pittsburgh) = ?

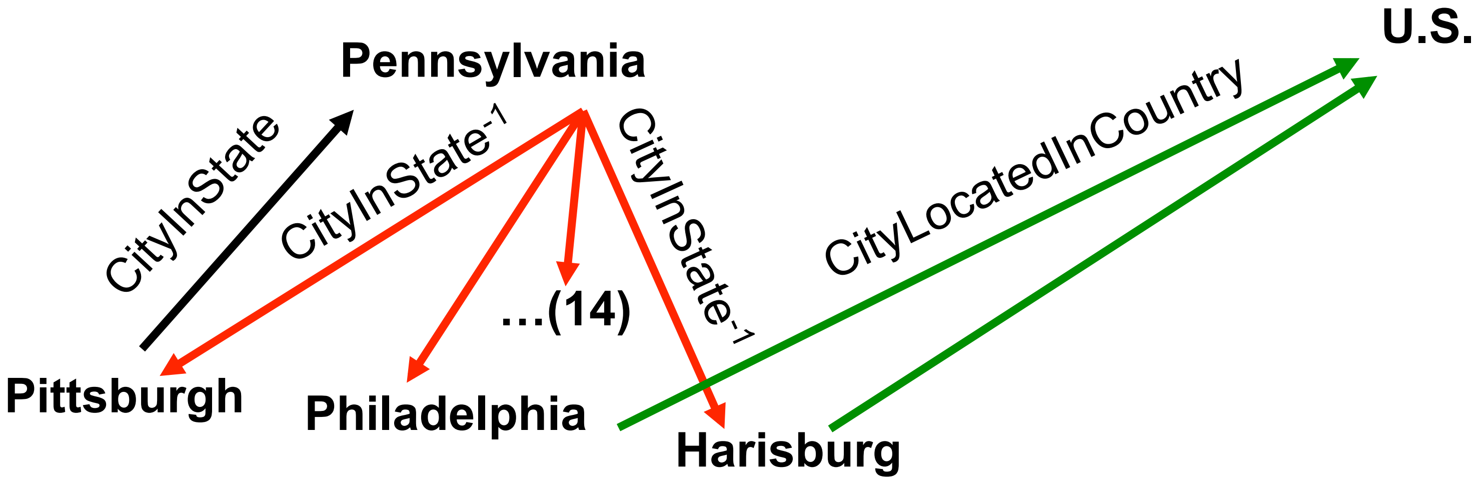
[Lao, Mitchell, Cohen, *EMNLP* 2011]



Feature = Typed Path	**Feature Value**	**Logistic Regresssion Weight**
CityInState, CityInstate^{-1}, CityLocatedInCountry		0.32

CityLocatedInCountry(Pittsburgh) = ?

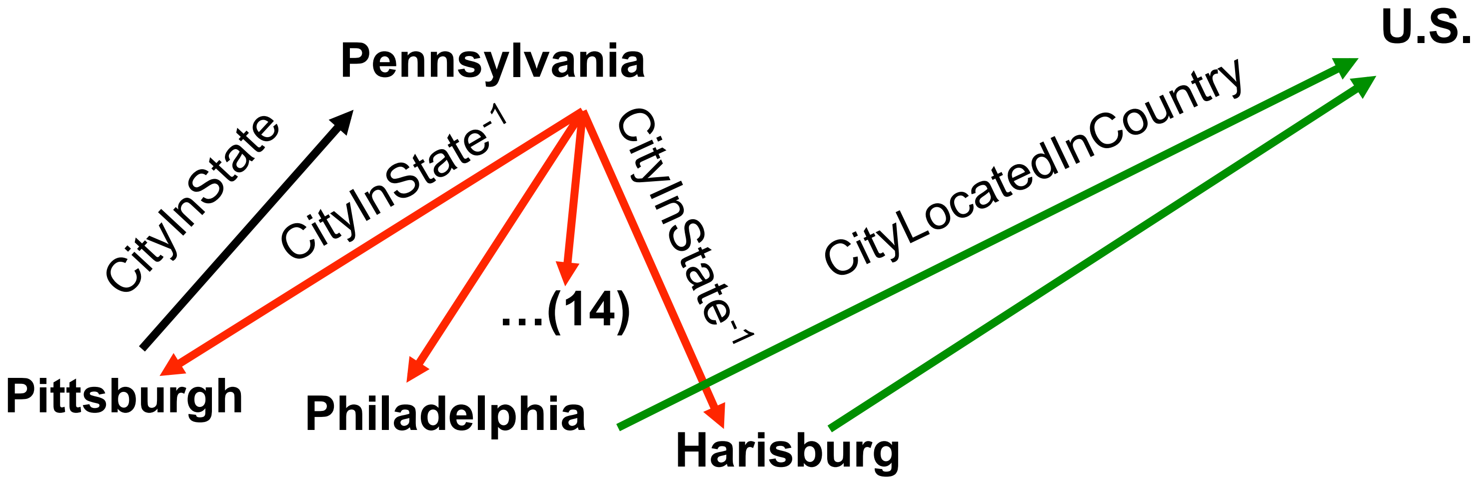
[Lao, Mitchell, Cohen, *EMNLP* 2011]



Feature = Typed Path	**Feature Value**	**Logistic Regresssion Weight**
CityInState, CityInstate^{-1}, CityLocatedInCountry		0.32

CityLocatedInCountry(Pittsburgh) = ?

[Lao, Mitchell, Cohen, *EMNLP* 2011]

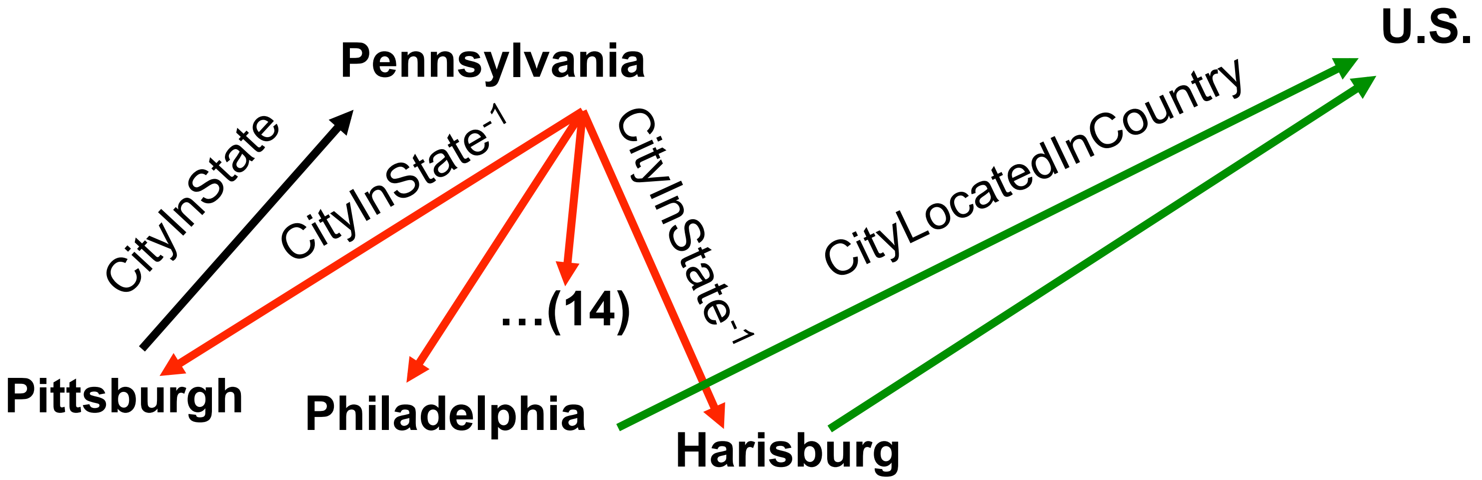


Pr(U.S. | Pittsburgh, TypedPath)

Feature = Typed Path	**Feature Value**	**Logistic Regresssion Weight**
CityInState, CityInstate^{-1}, CityLocatedInCountry	0.8	0.32

CityLocatedInCountry(Pittsburgh) = ?

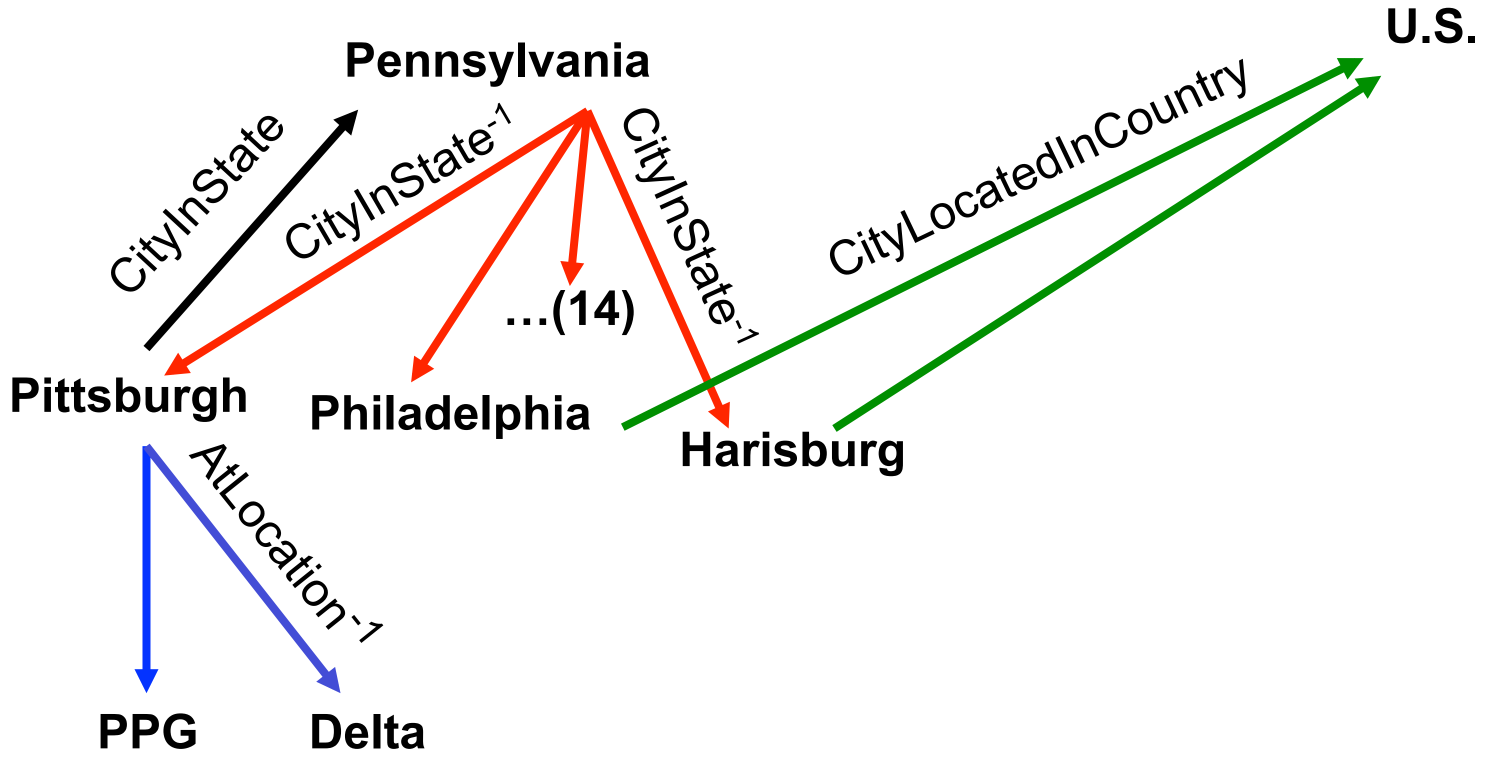
[Lao, Mitchell, Cohen, *EMNLP* 2011]



Feature = Typed Path	Feature Value	Logistic Regresssion Weight
CityInState, CityInstate^{-1}, CityLocatedInCountry	0.8	0.32
AtLocation^{-1}, AtLocation, CityLocatedInCountry		0.20

CityLocatedInCountry(Pittsburgh) = ?

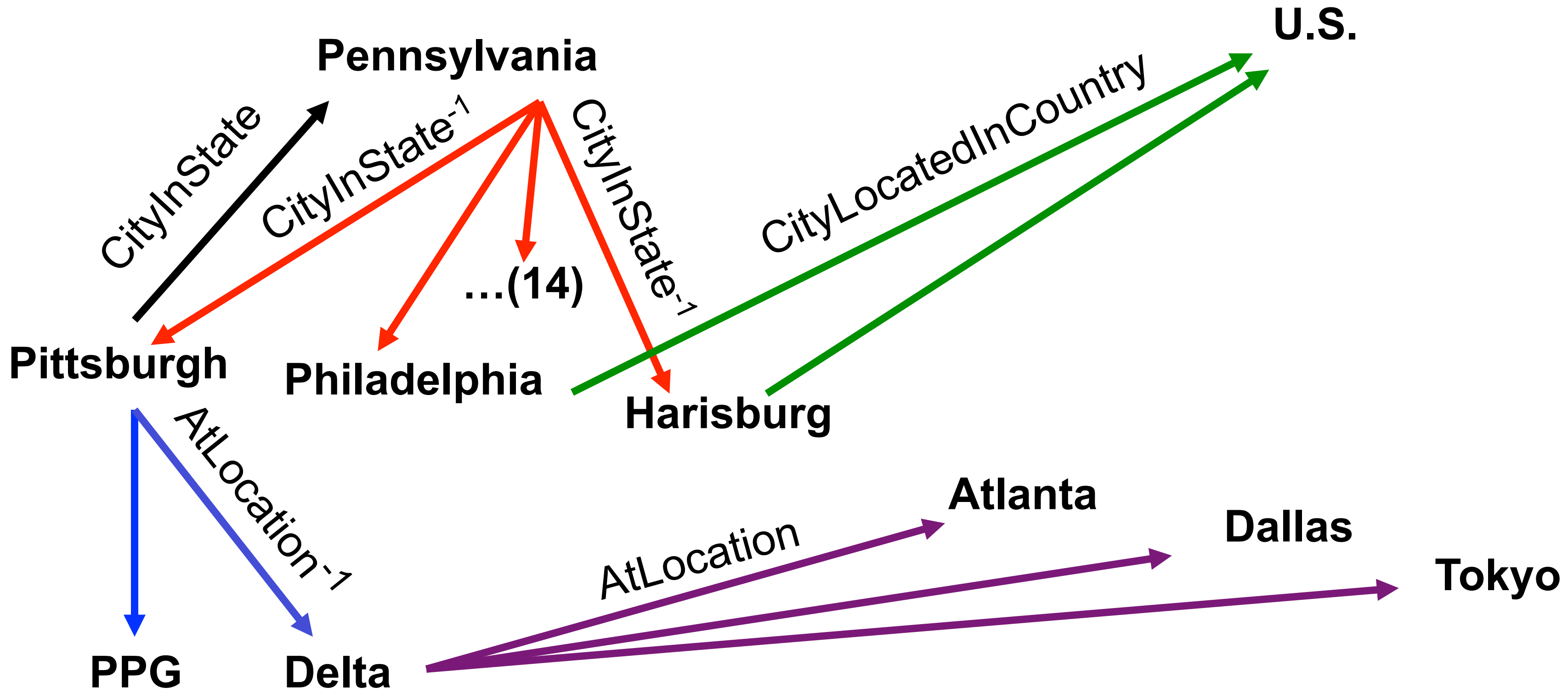
[Lao, Mitchell, Cohen, *EMNLP* 2011]



Feature = Typed Path	**Feature Value**	**Logistic Regresssion Weight**
CityInState, CityInstate^{-1}, CityLocatedInCountry	0.8	0.32
AtLocation^{-1}, AtLocation, CityLocatedInCountry		0.20

CityLocatedInCountry(Pittsburgh) = ?

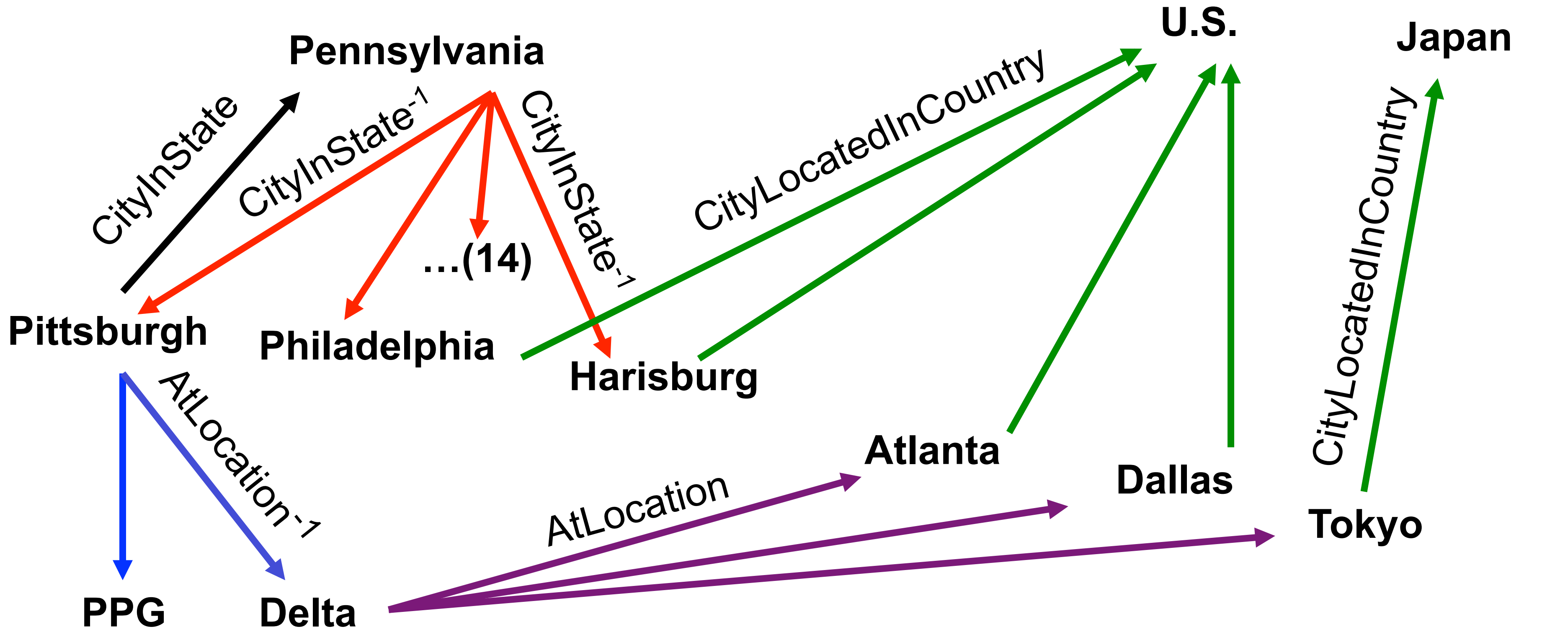
[Lao, Mitchell, Cohen, *EMNLP* 2011]



Feature = Typed Path	**Feature Value**	**Logistic Regresssion Weight**
CityInState, CityInstate^{-1}, CityLocatedInCountry	0.8	0.32
AtLocation^{-1}, AtLocation, CityLocatedInCountry		0.20

CityLocatedInCountry(Pittsburgh) = ?

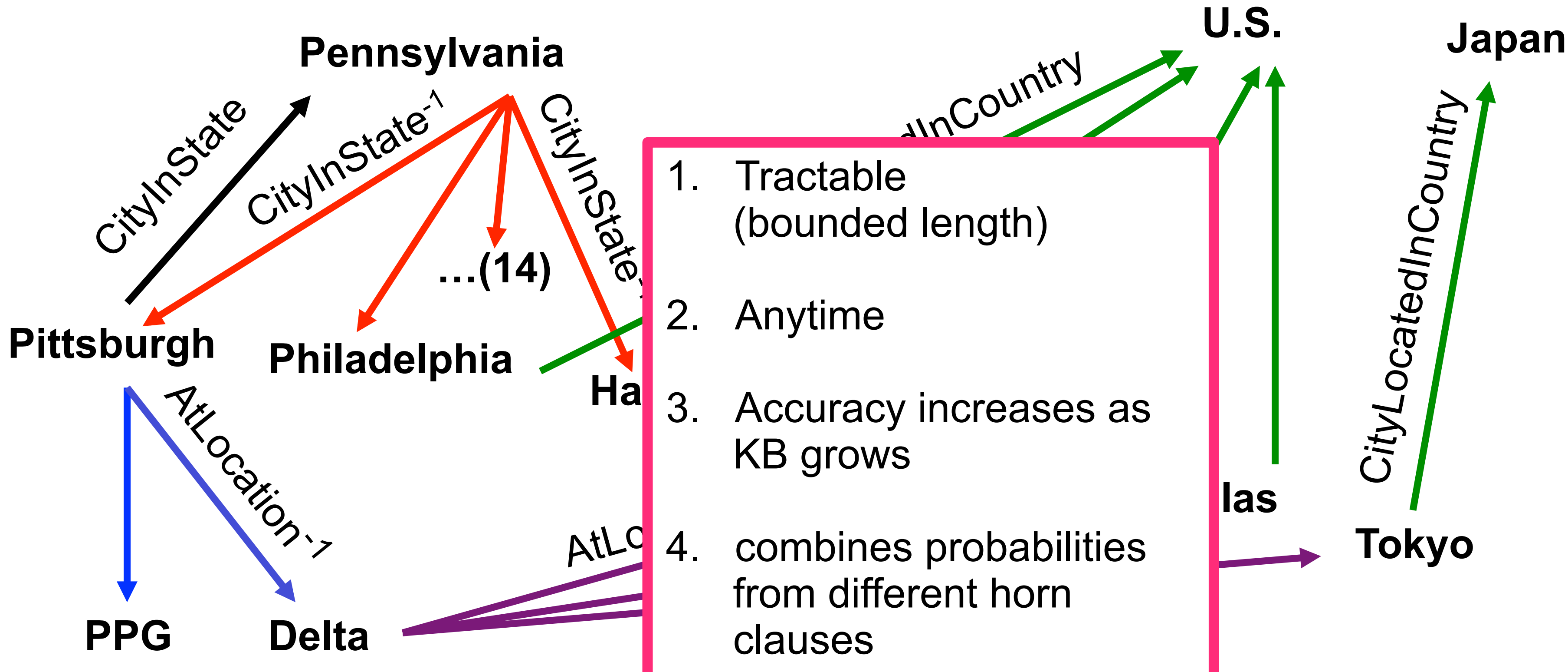
[Lao, Mitchell, Cohen, *EMNLP* 2011]



Feature = Typed Path	**Feature Value**	**Logistic Regresssion Weight**
CityInState, CityInstate^{-1}, CityLocatedInCountry	0.8	0.32
AtLocation^{-1}, AtLocation, CityLocatedInCountry	0.6	0.20

CityLocatedInCountry(Pittsburgh) = ?

[Lao, Mitchell, Cohen, *EMNLP* 2011]



1. Tractable (bounded length)
2. Anytime
3. Accuracy increases as KB grows
4. combines probabilities from different horn clauses

Feature = Typed Path	**Feature Value**	**Logistic Regresssion Weight**
CityInState, CityInstate^{-1}, CityLocatedInCountry	0.8	0.32
AtLocation^{-1}, AtLocation, CityLocatedInCountry	0.6	0.20
…	…	…

CityLocatedInCountry(Pittsburgh) = U.S. p=0.58

Random walk inference: learned rules

CityLocatedInCountry(*city, country*):

8.04 cityliesonriver, cityliesonriver^{-1}, citylocatedincountry

5.42 hasofficeincity^{-1}, hasofficeincity, citylocatedincountry

4.98 cityalsoknownas, cityalsoknownas, citylocatedincountry

2.85 citycapitalofcountry,citylocatedincountry^{-1},citylocatedincountry

2.29 agentactsinlocation^{-1}, agentactsinlocation, citylocatedincountry

1.22 statehascapital^{-1}, statelocatedincountry

0.66 citycapitalofcountry

.

.

.

[Gardner et al, 2014]

Opportunity:

Can infer more if we start with more densely connected knowledge graph

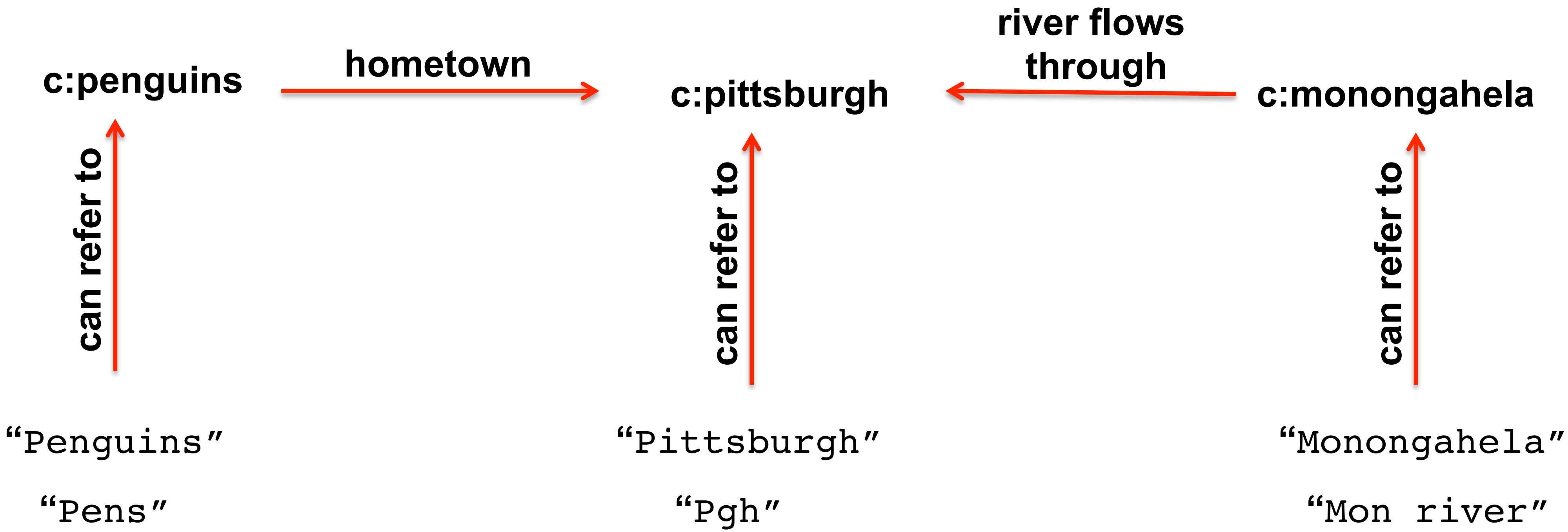
→ as NELL learns, it will become more dense

→ augment knowledge graph with a second graph of corpus statistics:

<subject, verb, object> triples

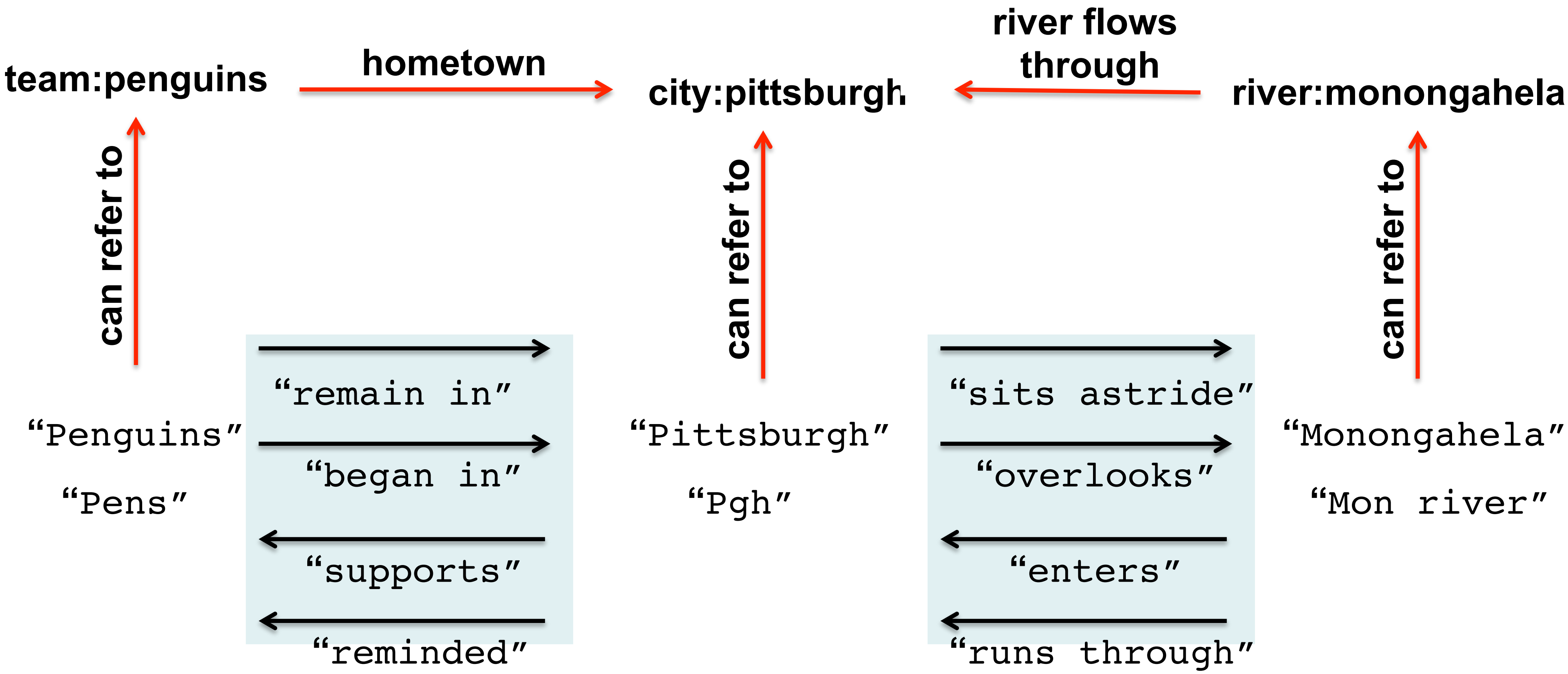
[Gardner et al, 2014]

NELL: concepts and "noun phrases"



[Gardner et al, 2014]

NELL: concepts and "noun phrases"



SVO triples from 500 M dependency parsed web pages (thank you Chris Re!)

NELL: concepts and "noun phrases"

- Circumvents NELL's fixed vocabulary of relations!
- Sadly, adding these does not help: too sparse
- But clustering verb phrases based on latent embedding (NNMF), produces significant improvement
  - {"lies on", "runs through", "flows through", …}
- Precision/recall over 15 NELL relations:
  KB only: 0.80 / 0.33
  KB + SVO_{latent}: 0.87 / 0.42

c:penguins **hometown**

can refer to

"Penguins"

"Pens"

"remain in"

"began in"

"supports"

"reminded"

"enters"

"runs through"

[Gardner et al., 2014]

SVO triples from 500 M dependency parsed web pages (thank you Chris Re!)

Building the Knowledge Graph by Reading

1. Classify noun phrases (NP’s) by category
2. Classify NP pairs by relation
3. Discover rules to predict new relation instances
4. Learn which NP’s (co)refer to which latent concepts
5. Discover new relations to extend ontology
6. Learn to infer relation instances via targeted random walks
7. Vision: connect NELL and NEIL

New Direction: Integrate Vision with Text

The problem:

Many things not learnable from text

New direction:

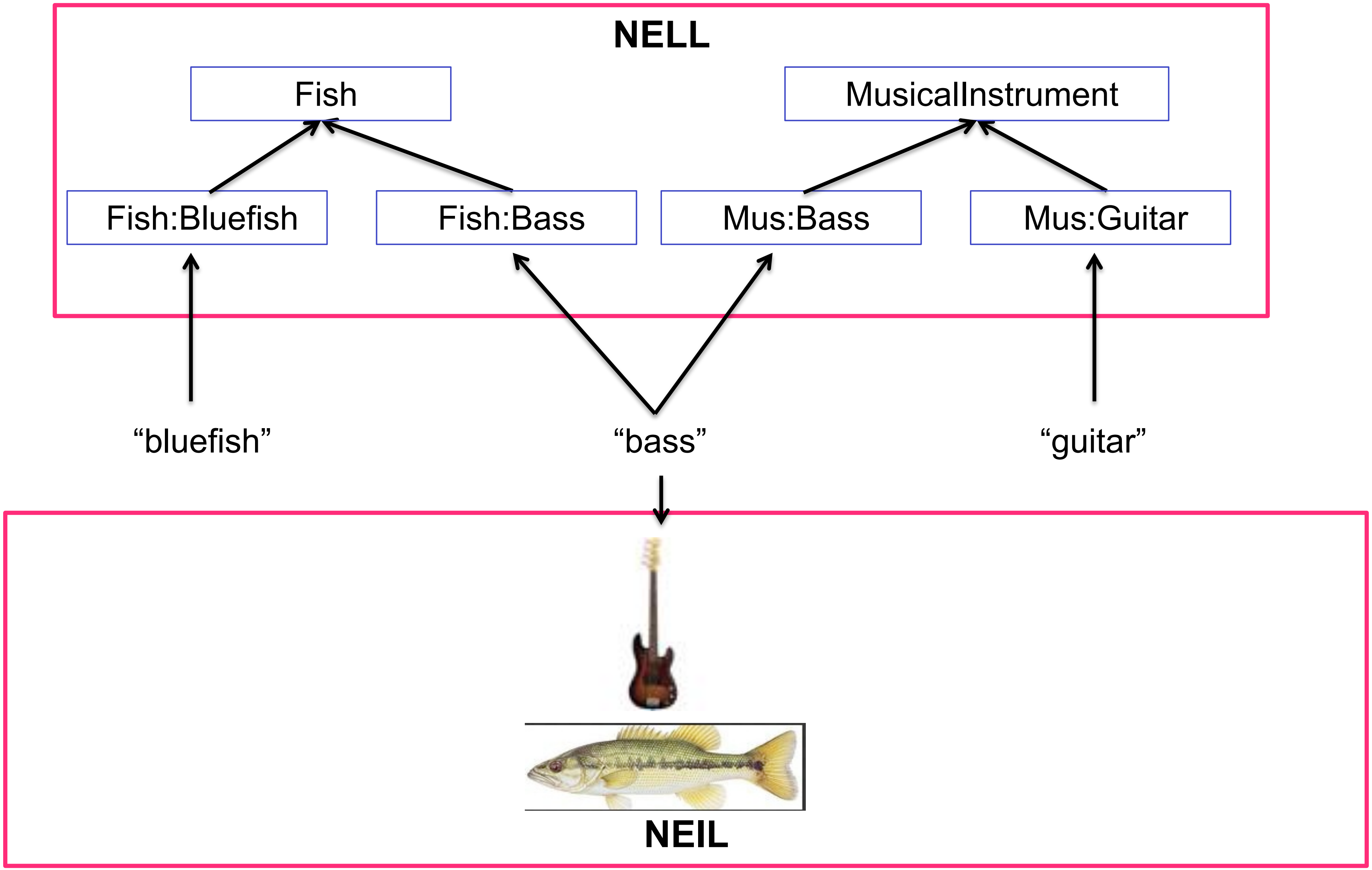
integrate NELL with NEIL (Never Ending Image Learner) [Gupta, Chen, 2013]

NELL gives noun phrases it understands to NEIL

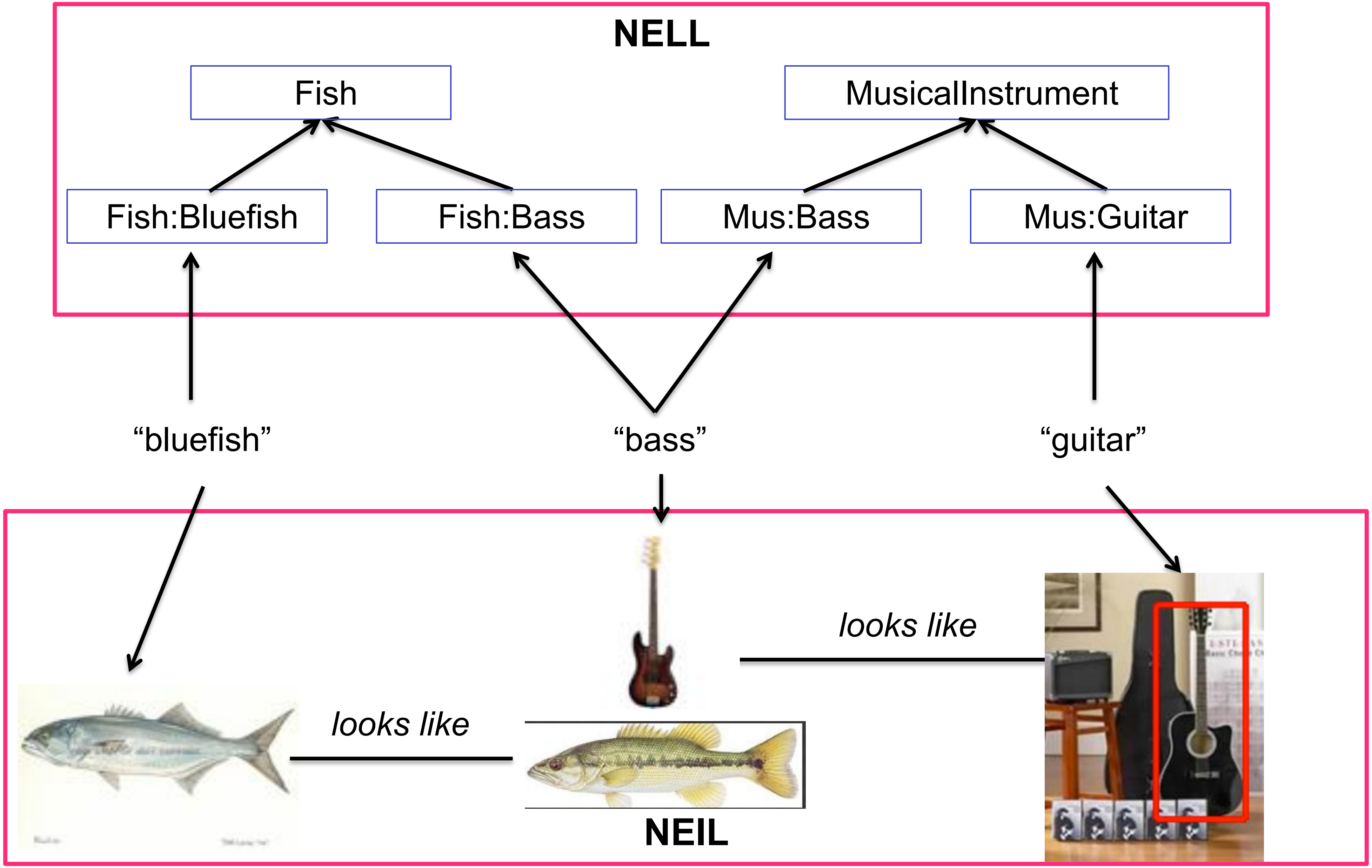
NEIL collects images associated with these, and analyzes

NELL, NEIL cotraining

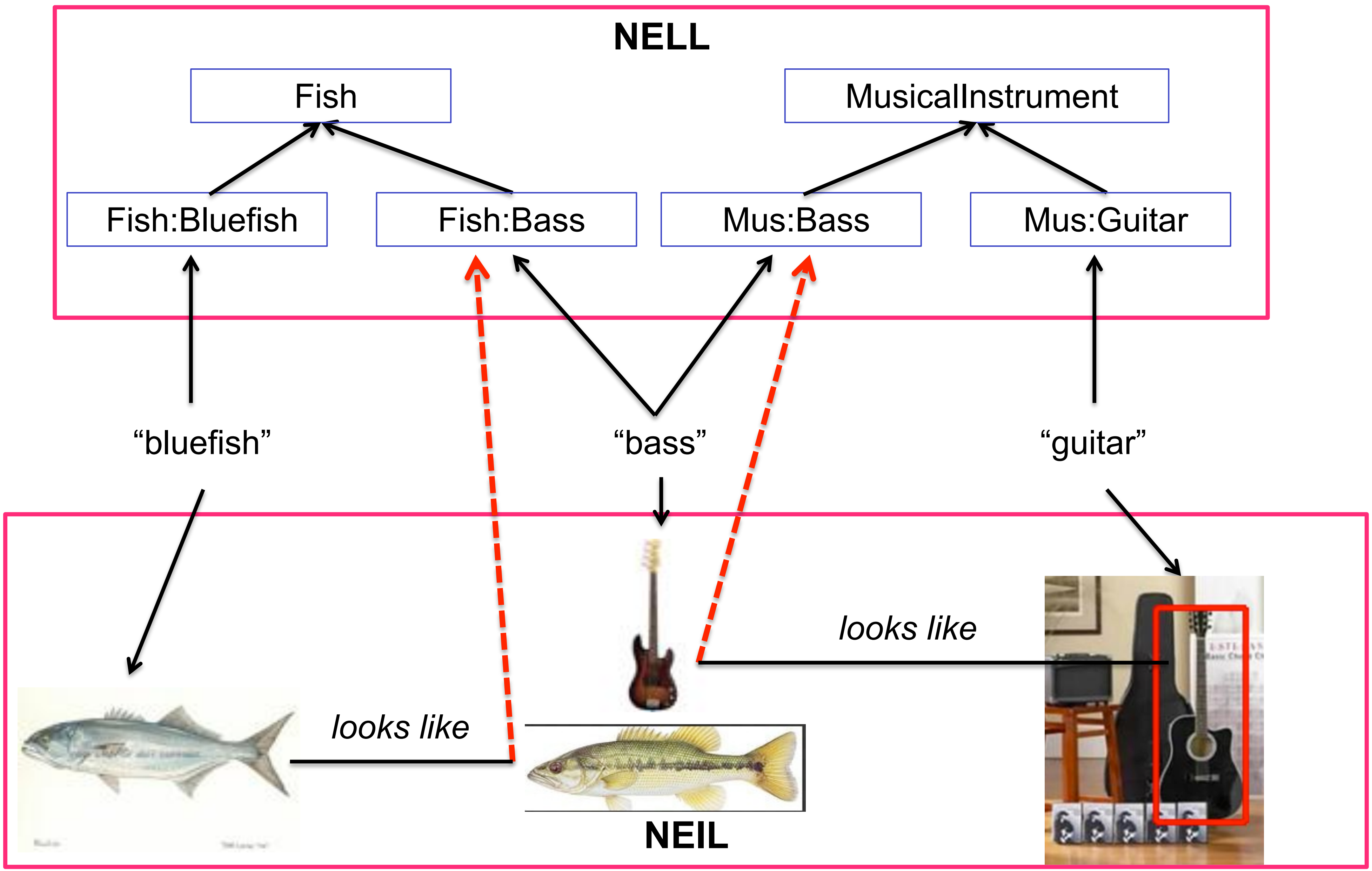
NEIL / NELL Polysemy: Bass



NEIL / NELL Polysemy: Bass



NEIL / NELL Polysemy: Bass



Building the Knowledge Graph by Reading

1. Classify noun phrases (NP’s) by category
2. Classify NP pairs by relation
3. Discover rules to predict new relation instances
4. Learn which NP’s (co)refer to which latent concepts
5. Discover new relations to extend ontology
6. Learn to infer relation instances via targeted random walks
7. Vision: connect NELL and NEIL
8. Mutilingual NELL (Portuguese)

Recently learned beliefs (from English text)

Refresh

instance	iteration	date learned	confidence
actimmune is a product	890	11-dec-2014	100.0
dogwood_drive is a street	892	30-dec-2014	100.0
the_news_progress is a newspaper	892	30-dec-2014	100.0
university_of_washington is a train station	892	30-dec-2014	100.0
iranian_real is a currency	892	30-dec-2014	91.5
lotronex is a drug worked on by glaxosmithkline	892	30-dec-2014	93.8
peter_finch starred in the movie network	892	30-dec-2014	100.0
bmw is a specific automobile maker dealer in tampa_bay	893	02-jan-2015	100.0
jeremy is a person who died at the age of 5	895	22-jan-2015	98.4
johannes_brahms is a person born on the date n1833	895	22-jan-2015	100.0

Recently learned beliefs (from Portuguese text)

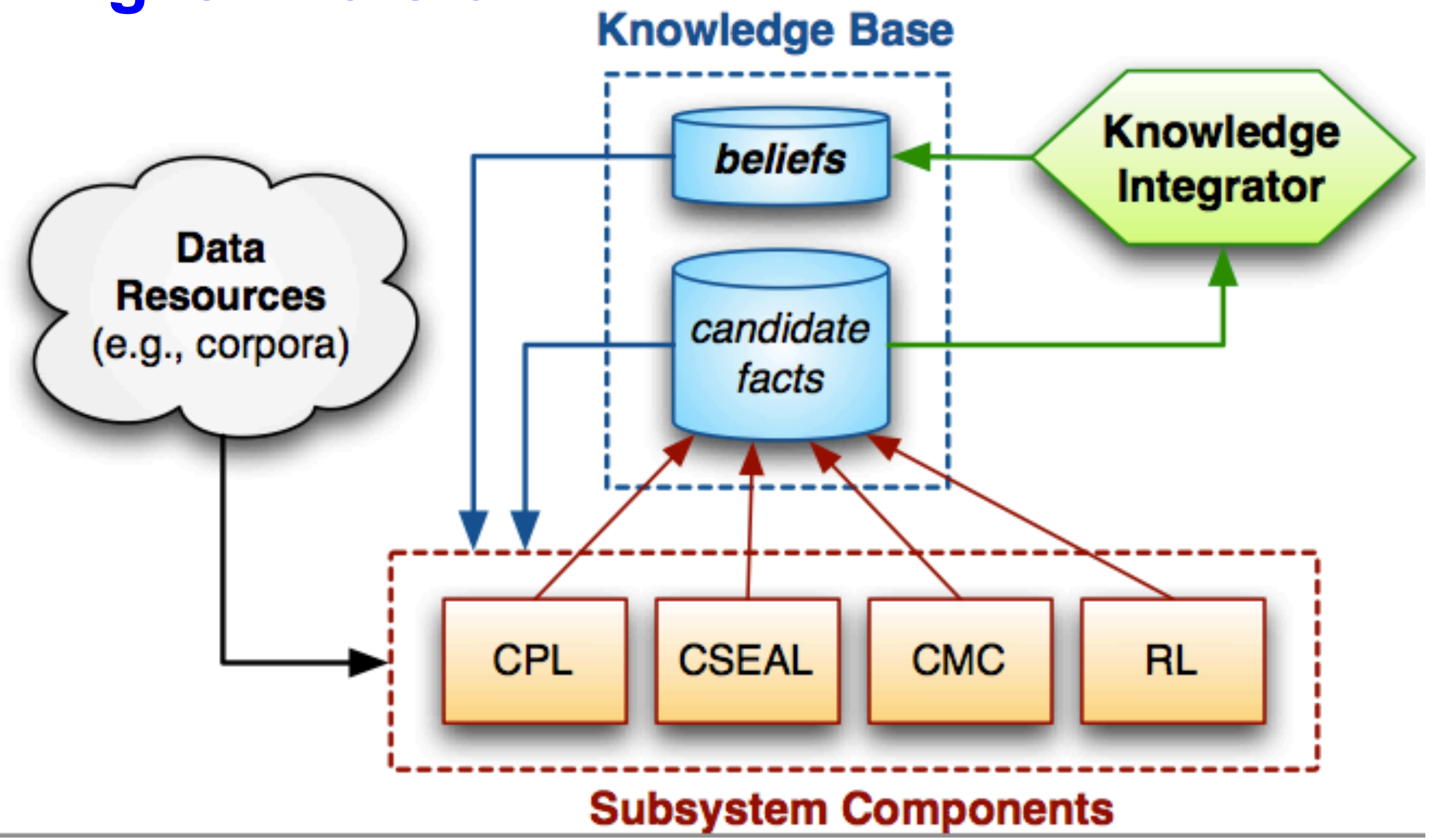
Refresh

instance	iteration	date learned	confidence
friboi is an organization	53	15-nov-2014	100.0
porto_alegre_ouro_preto_recife is a city	53	15-nov-2014	100.0
leis_do_poder is a book	54	13-dec-2014	100.0
primavera is a visualizable object	52	14-nov-2014	97.4
pirelli_general_motors_e_souza is a company	52	14-nov-2014	99.0
u_s__bancorp is a bank that has richard_k__davis as its CEO	57	12-jan-2015	100.0
curling is a sport with fans in the country canada	55	21-dec-2014	100.0

How to Read the Web in Many Languages?

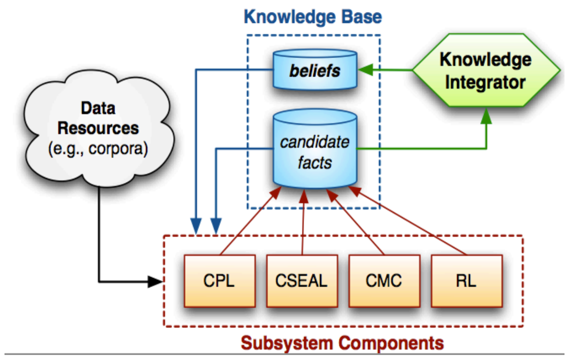
NELL: Never-Ending Language Learner

English Version



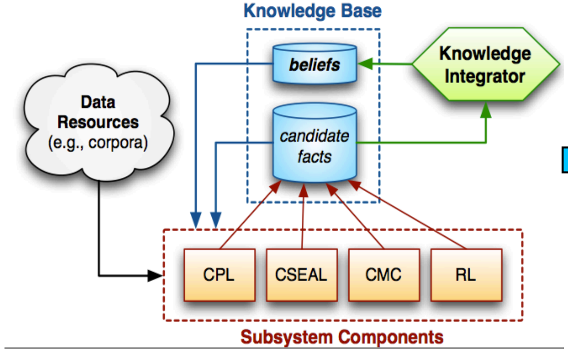
NELL: Never-Ending Language Learner

English NELL

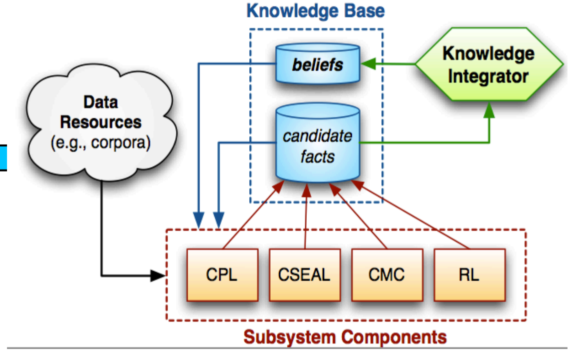


NELL: Never-Ending Language Learner

English NELL

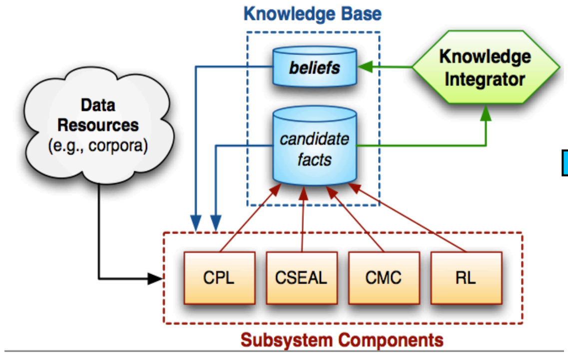


Portuguese NELL



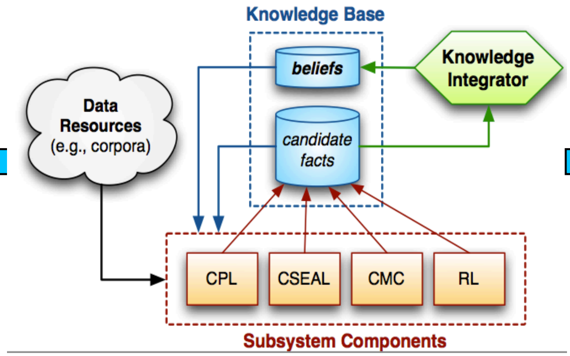
NELL: Never-Ending Language Learner

English NELL



+

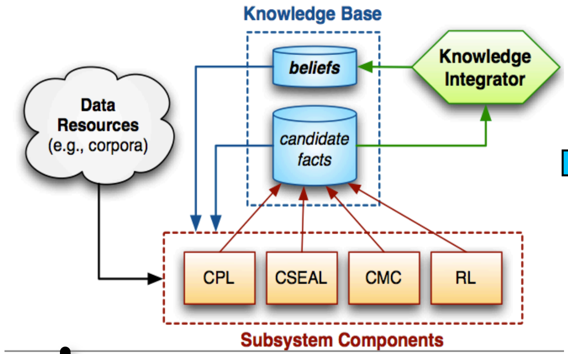
Portuguese NELL



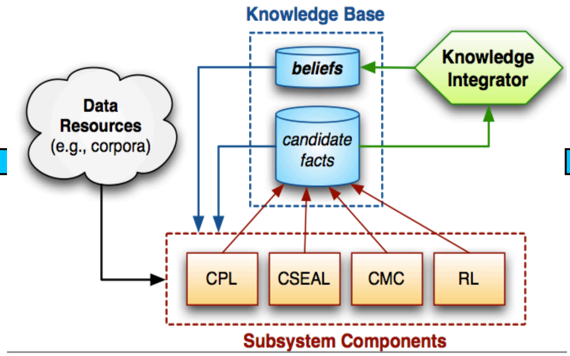
+ . . .

NELL: Never-Ending Language Learner

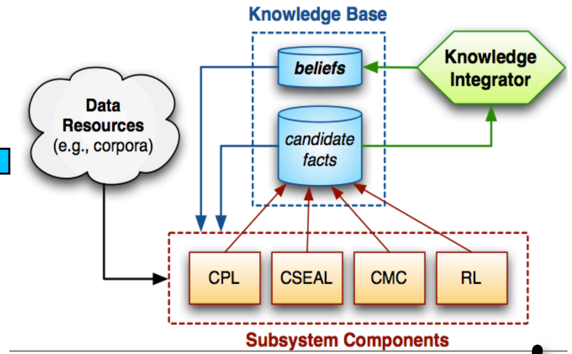
English NELL



Portuguese NELL

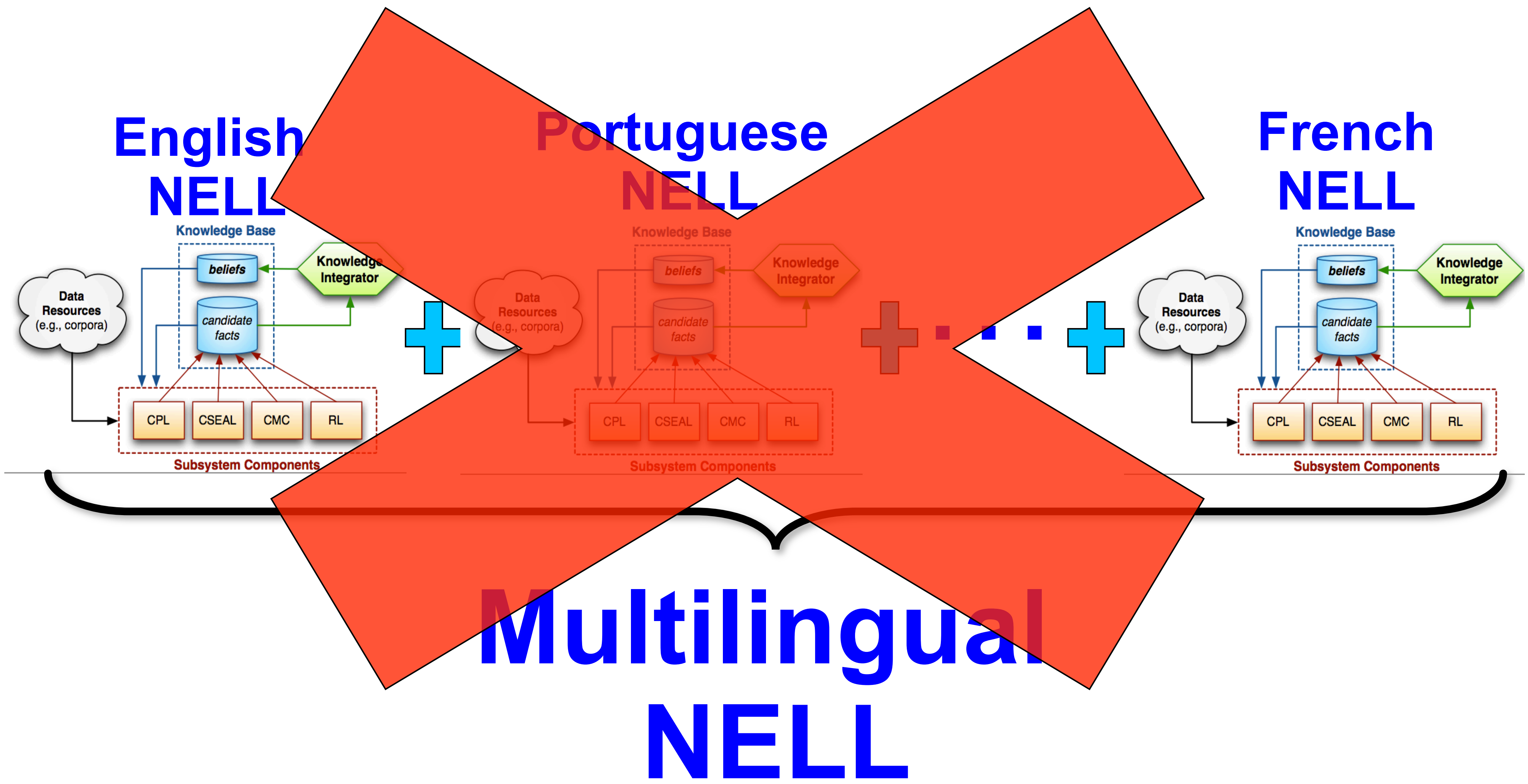


French NELL

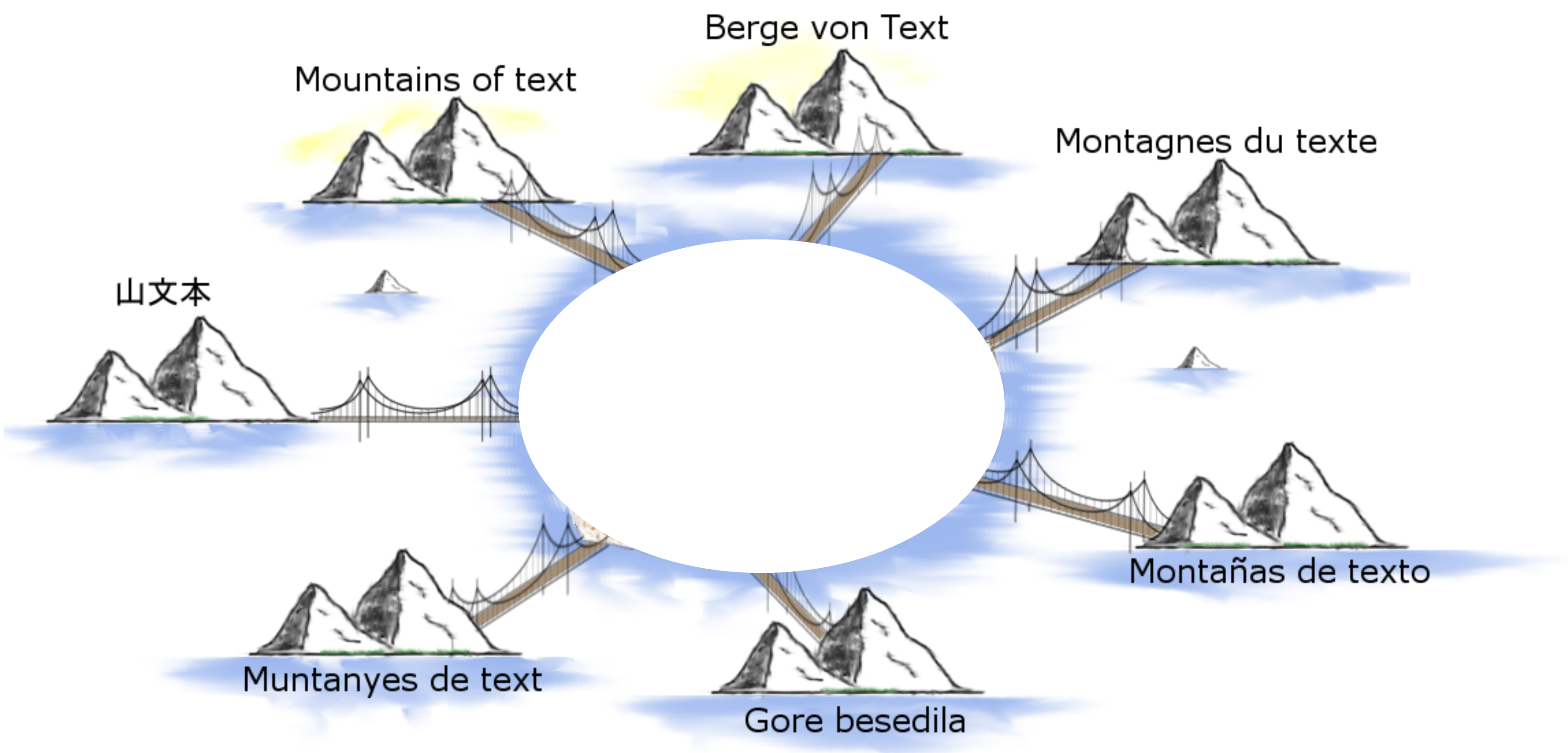


Multilingual NELL

NELL: Never-Ending Language Learner

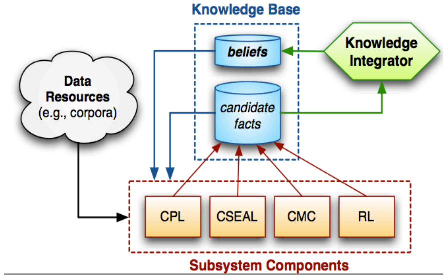


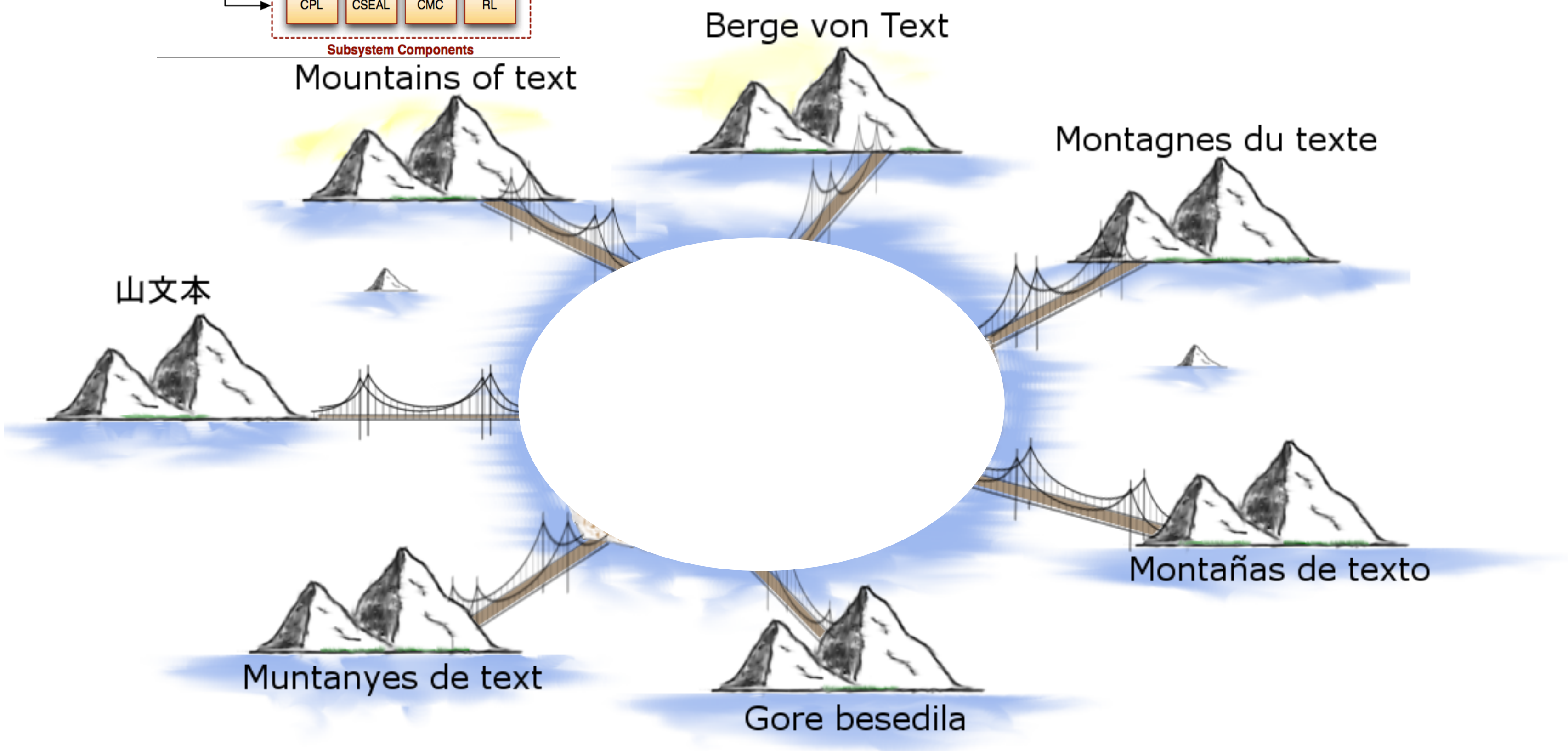
Multilingual Reading The Web



Picture Adapted from http://www.xlike.org

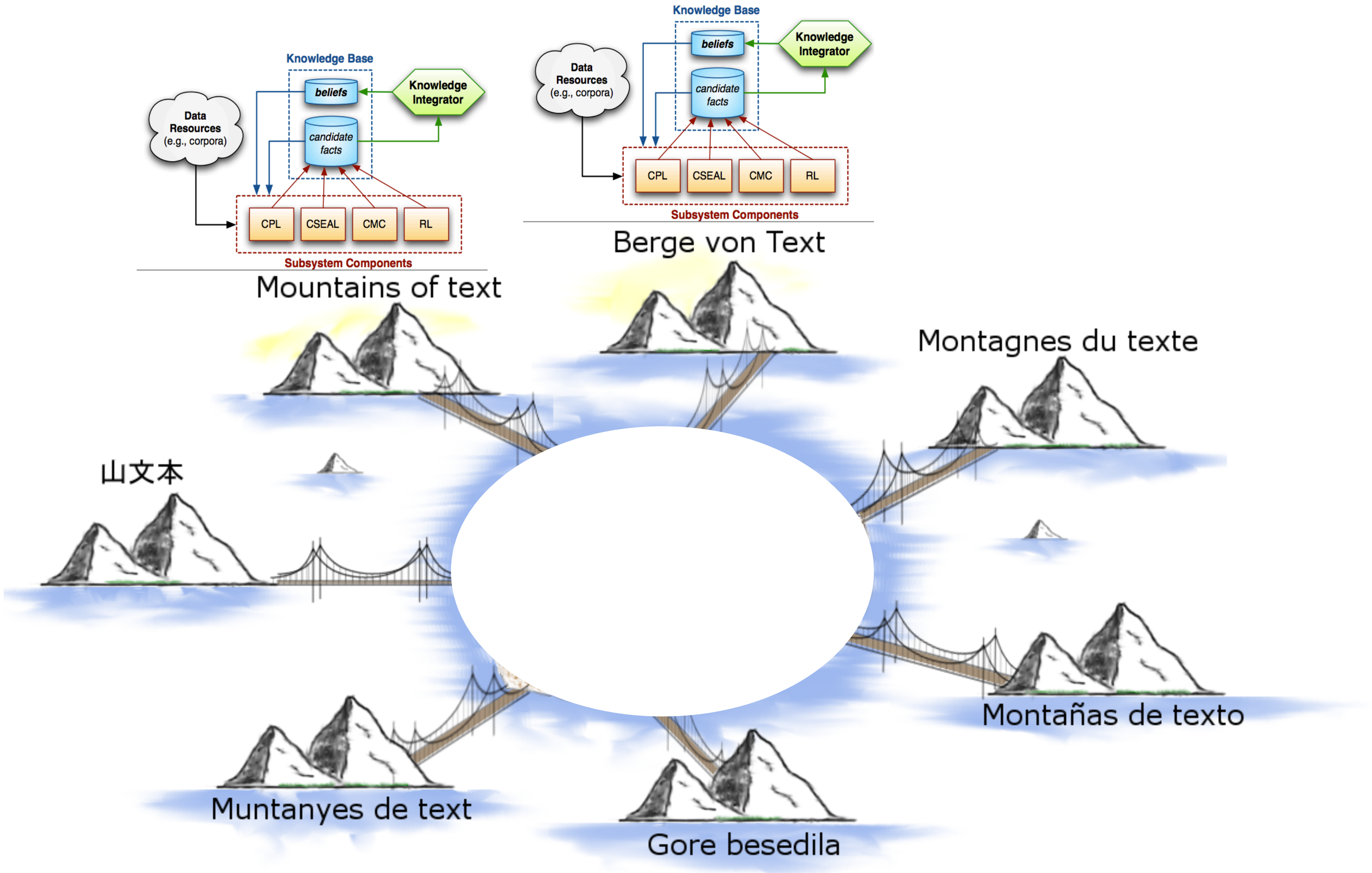
Multilingual Reading The Web





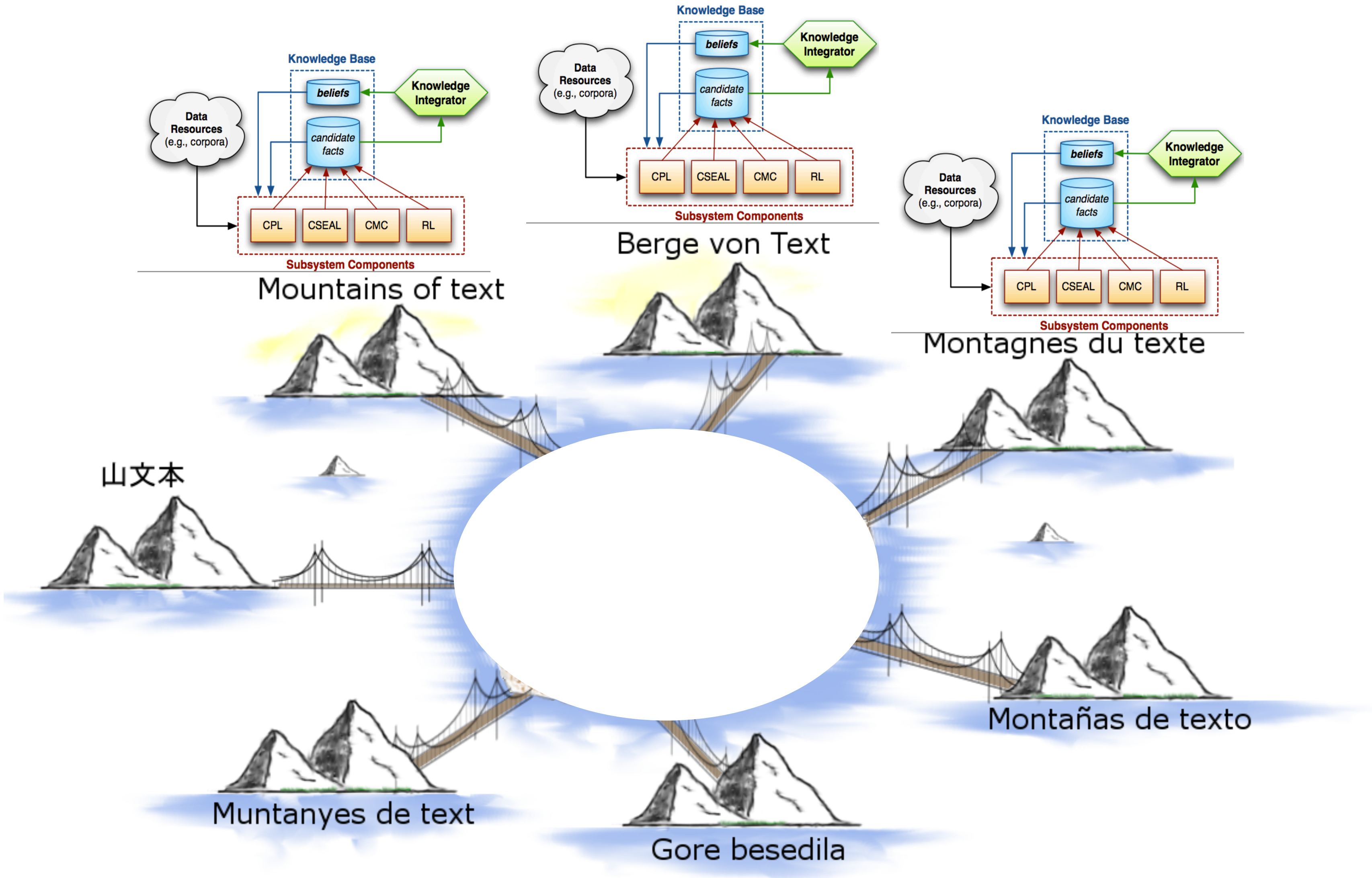
Picture Adapted from http://www.xlike.org

Multilingual Reading The Web



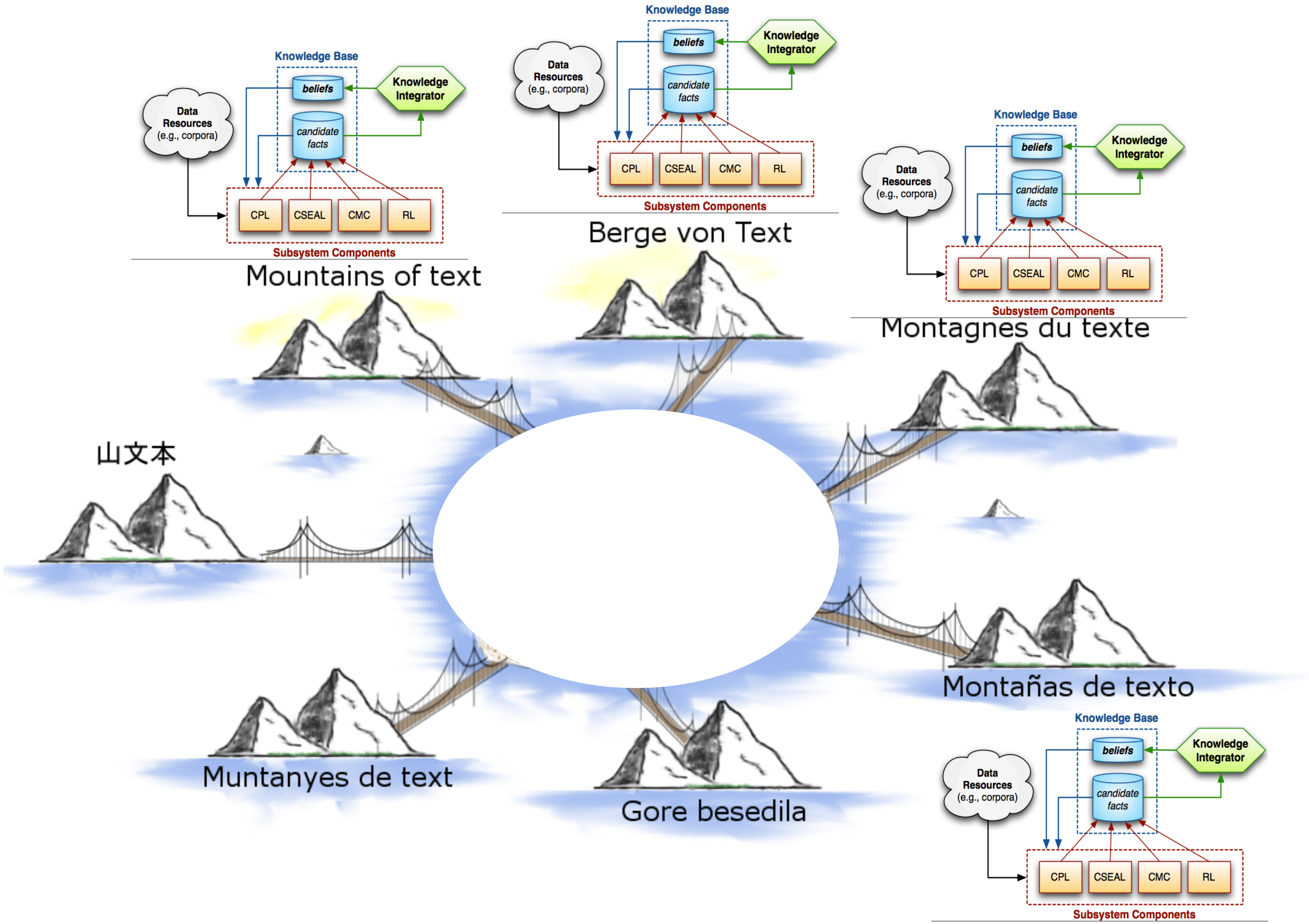
Picture Adapted from http://www.xlike.org

Multilingual Reading The Web



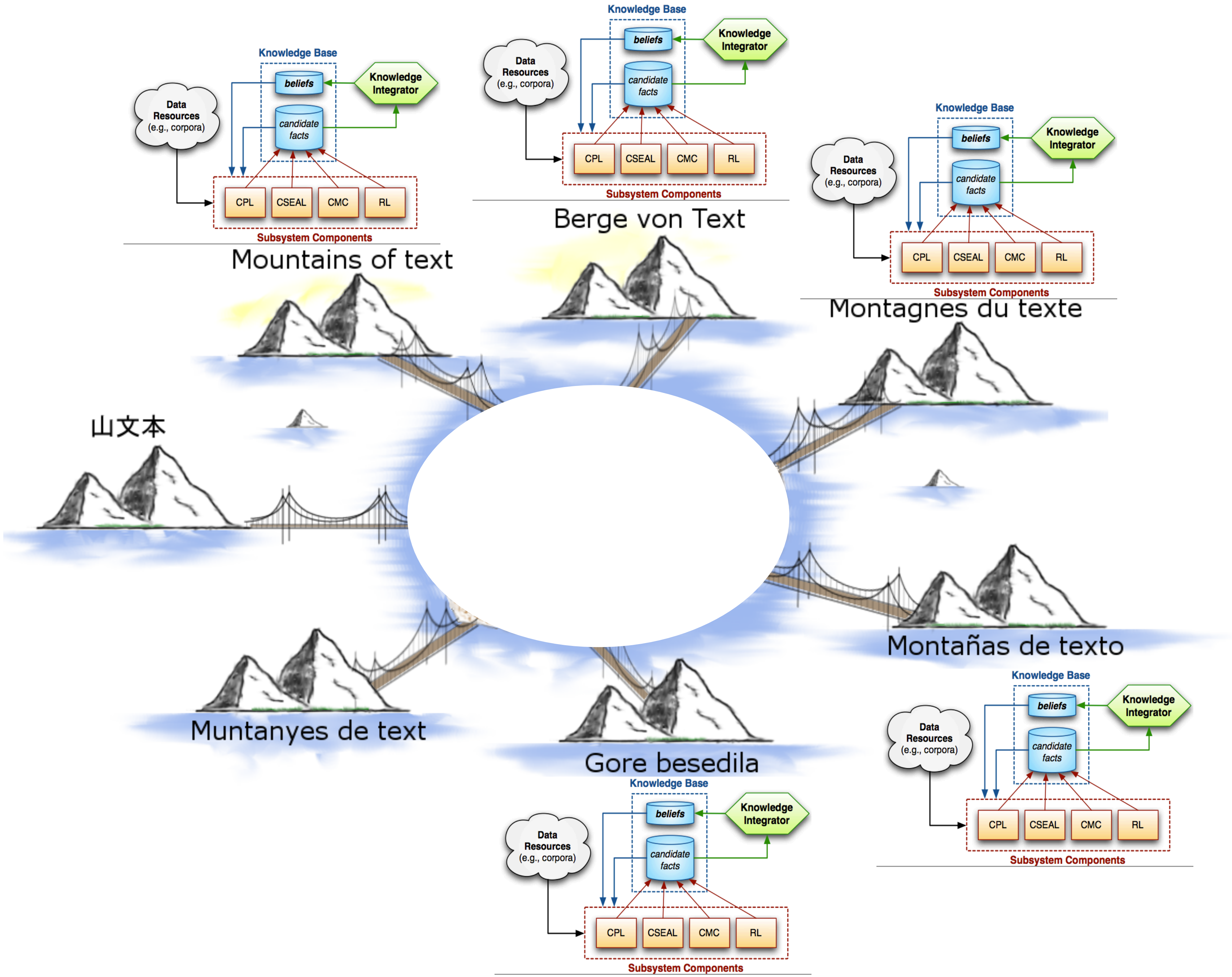
Picture Adapted from http://www.xlike.org

Multilingual Reading The Web



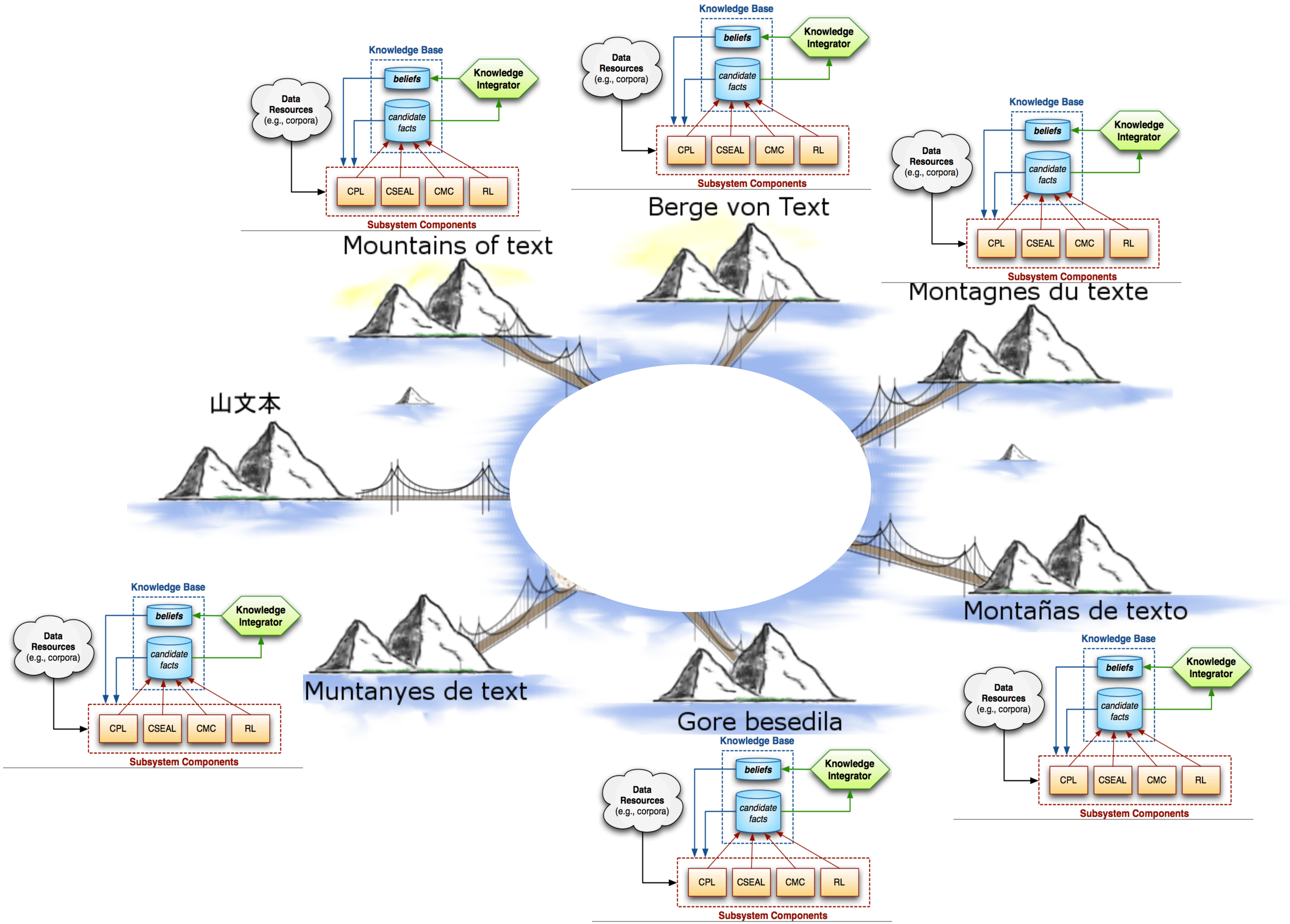
Picture Adapted from http://www.xlike.org

Multilingual Reading The Web



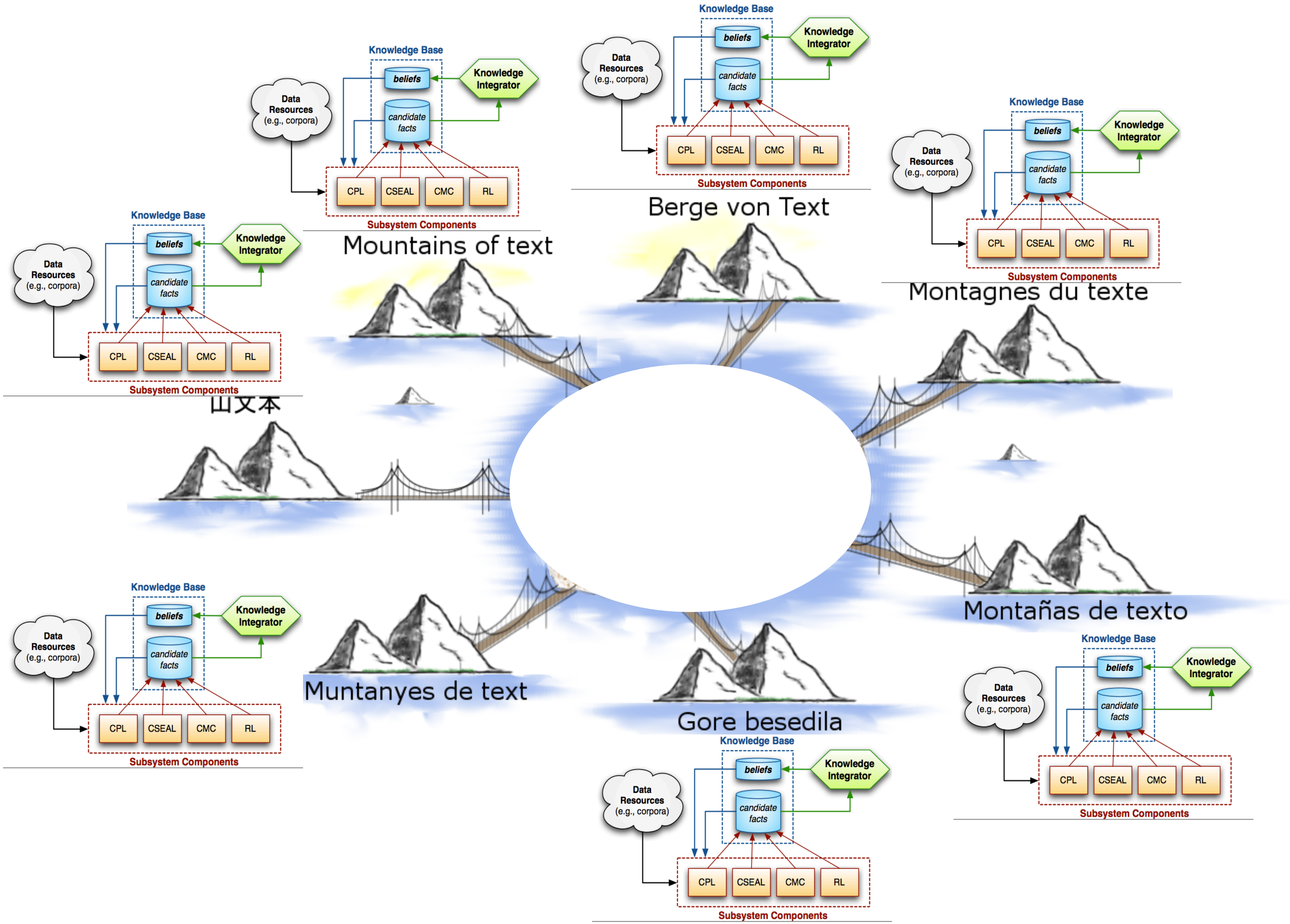
Picture Adapted from http://www.xlike.org

Multilingual Reading The Web



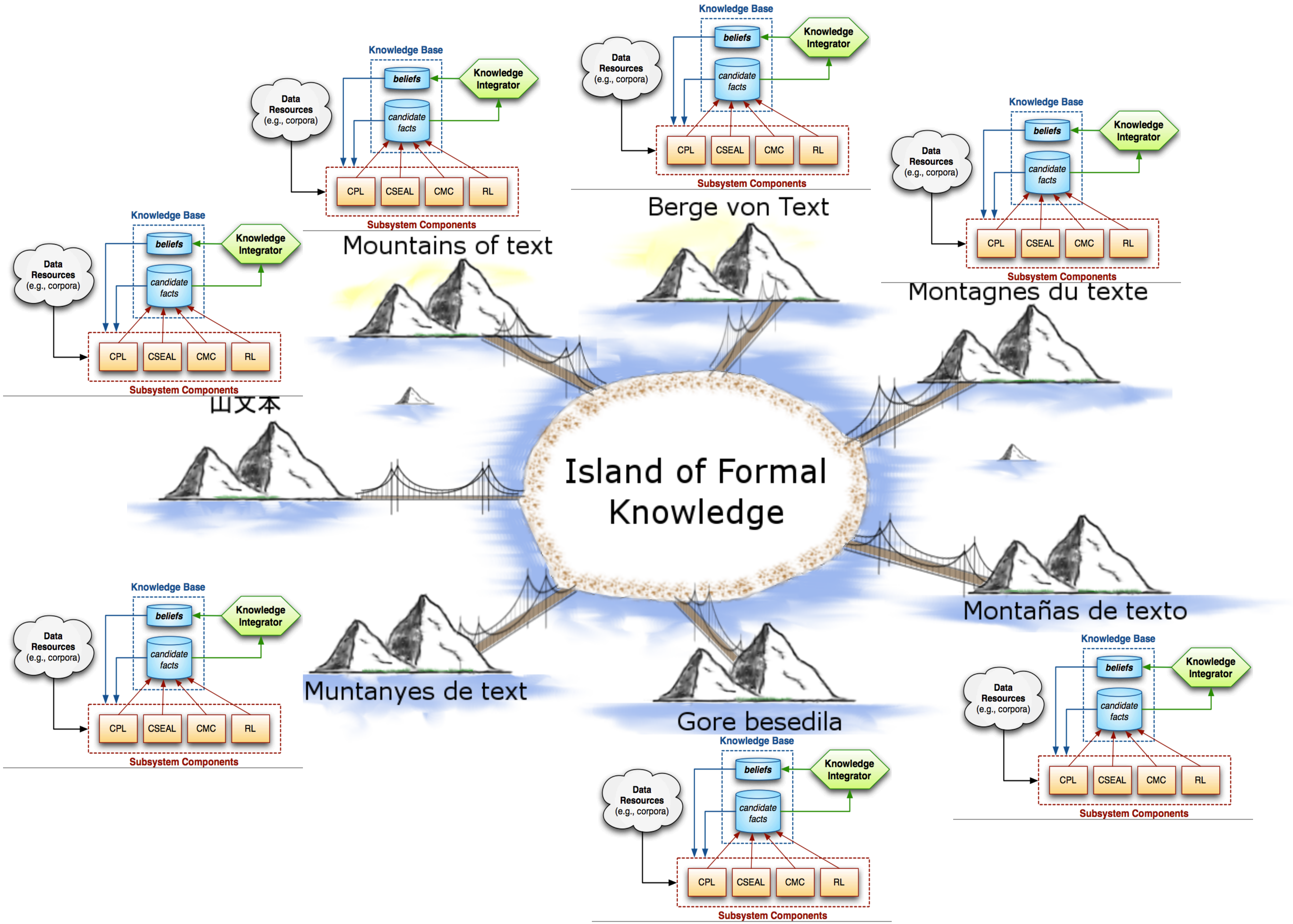
Picture Adapted from http://www.xlike.org

Multilingual Reading The Web



Picture Adapted from http://www.xlike.org

Multilingual Reading The Web



Picture Adapted from http://www.xlike.org

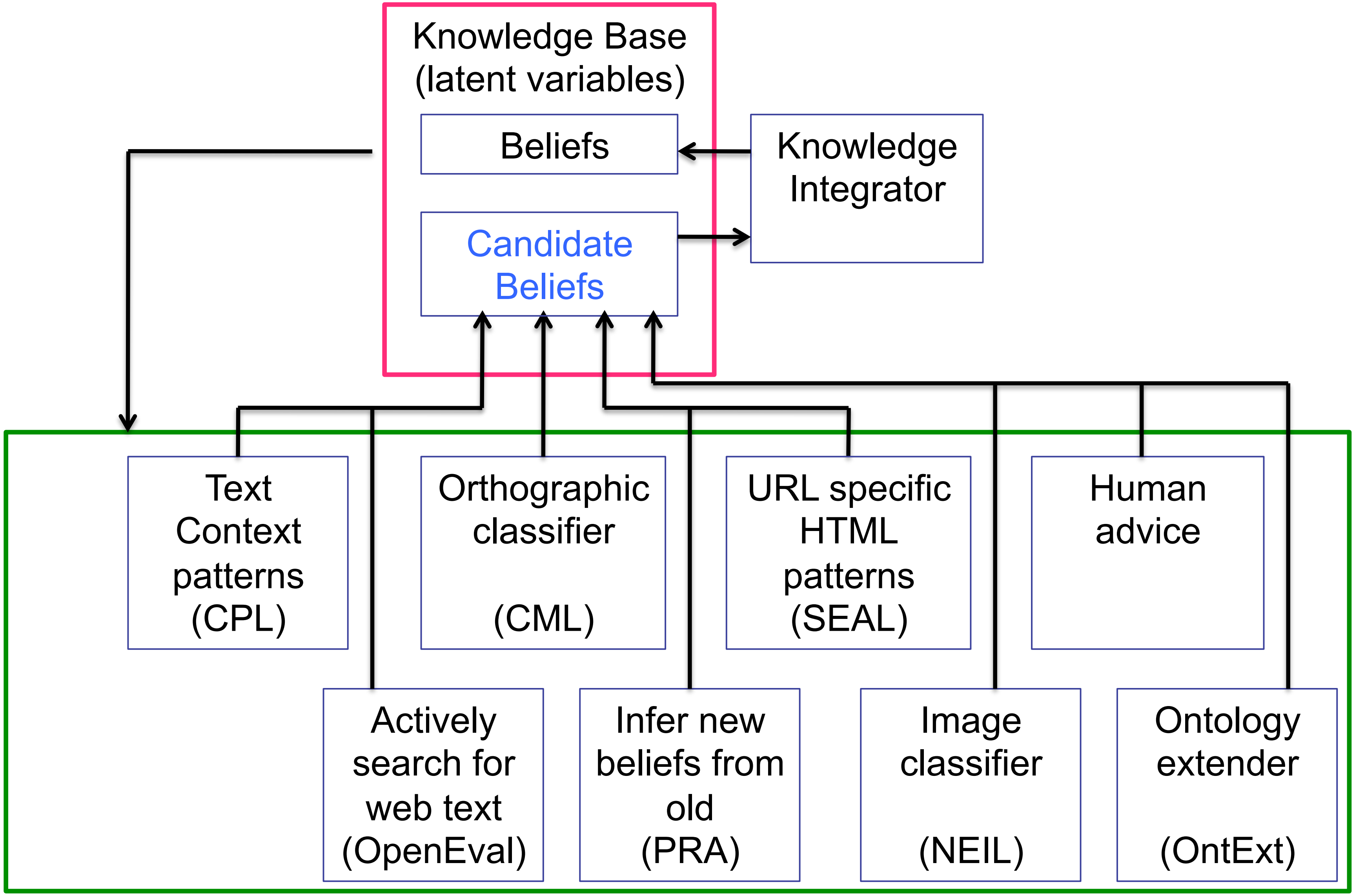
Key Idea 4: Cumulative, Staged Learning
Learning X improves ability to learn Y

1. Classify noun phrases (NP's) by category
2. Classify NP pairs by relation
3. Discover rules to predict new relation instances
4. Learn which NP's (co)refer to which latent concepts
5. Discover new relations to extend ontology
6. Learn to infer relation instances via targeted random walks
7. Vision: connect NELL and NEIL
8. Mutilingual NELL (Portuguese)

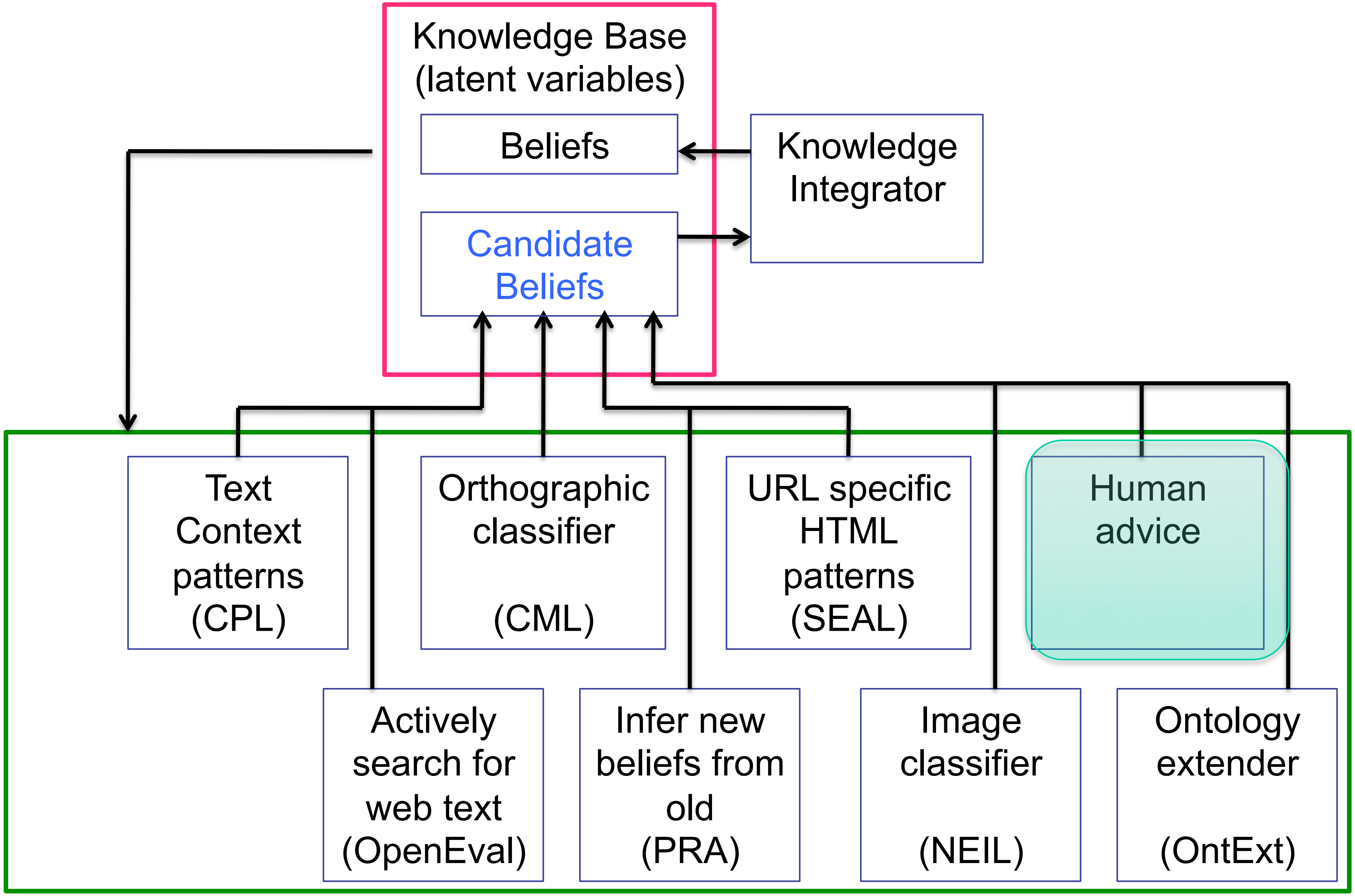
NELL is here

9. Learn to microread single sentences
10. Self reflection, self-directed learning
11. Goal-driven reading: predict, then read to corroborate/correct
12. Make NELL learn by conversation (e.g, Twitter)
13. Add a robot body, or mobile phone body, to NELL

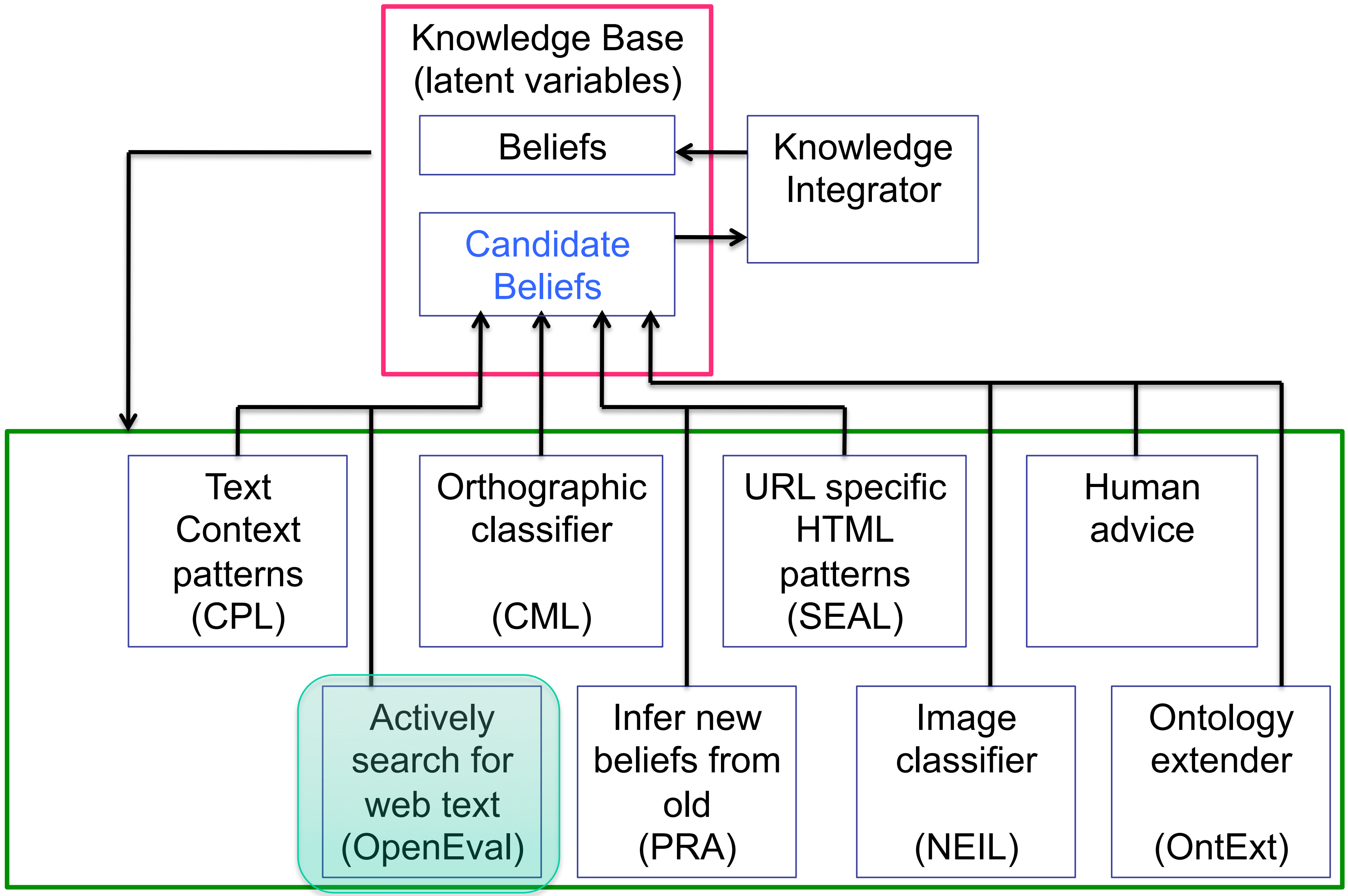
NELL Architecture



NELL Architecture



NELL Architecture



NELL: Never-Ending Language Learner

NELL is grown enough for new steps

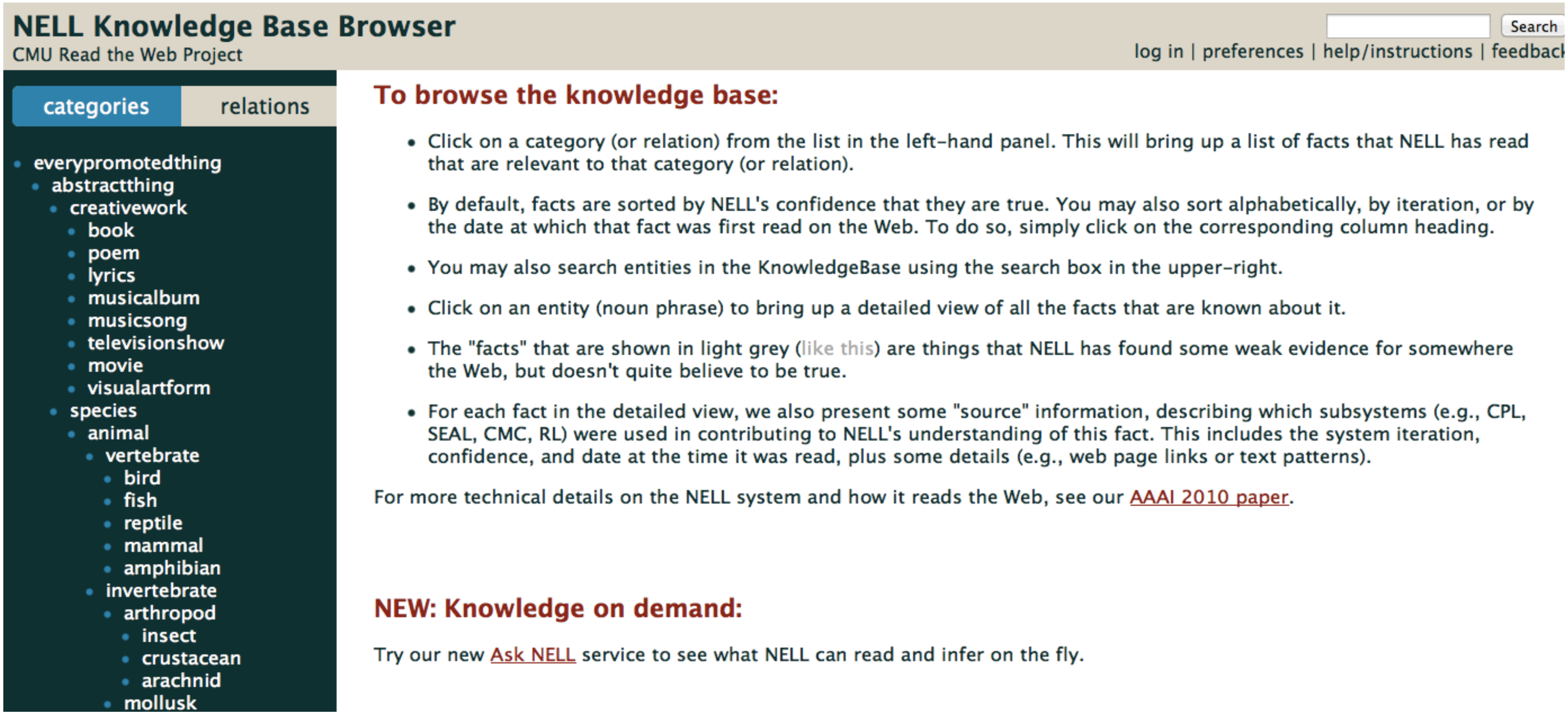
NELL turned 5 on Jan 12!
Congratulations NELL!!

NELL: Never-Ending Language Learner

NELL is grown enough for new st

NELL: Never-Ending Language Learner

NELL is grown enough for new steps



NELL: Never-Ending Language Learner

NELL is grown enough for new steps

NELL Knowledge Base Browser

CMU Read the Web Project

Search

log in | preferences | help/instructions | feedback

categories | relations

- everypromotedthing
  - abstractthing
    - creativework
      - book
      - poem
      - lyrics
      - musicalbum
      - musicsong
      - televisionshow
      - movie
      - visualartform
    - species
      - animal
        - vertebrate
          - bird
          - fish
          - reptile
          - mammal
          - amphibian
        - invertebrate
          - arthropod
            - insect
            - crustacean
            - arachnid
          - mollusk

To browse the knowledge base:

- Click on a category (or relation) from the list in the left-hand panel. This will bring up a list of facts that NELL has read that are relevant to that category (or relation).
- By default, facts are sorted by NELL's confidence that they are true. You may also sort alphabetically, by iteration, or by the date at which that fact was first read on the Web. To do so, simply click on the corresponding column heading.
- You may also search entities in the KnowledgeBase using the search box in the upper-right.
- Click on an entity (noun phrase) to bring up a detailed view of all the facts that are known about it.
- The "facts" that are shown in light grey (like this) are things that NELL has found some weak evidence for somewhere the Web, but doesn't quite believe to be true.
- For each fact in the detailed view, we also present some "source" information, describing which subsystems (e.g., CPL, SEAL, CMC, RL) were used in contributing to NELL's understanding of this fact. This includes the system iteration, confidence, and date at the time it was read, plus some details (e.g., web page links or text patterns).

For more technical details on the NELL system and how it reads the Web, see our AAAI 2010 paper.

NEW: Knowledge on demand:

Try our new Ask NELL service to see what NELL can read and infer on the fly.

NELL: Never-Ending Language Learner

NELL is grown enough for new steps

http://rtw.ml.cmu.edu

estevam.hruschka@gmail.com

Thank you very much!

and thanks to **Tsinghua University**, Google, Yahoo!, NSF, DARPA, Intel, Microsoft, Fulbright, Bloomberg, CNPq and FAPESP for partial funding and thanks to Carnegie Mellon University and thanks to Federal University of São Carlos

References

[Fern, 2008] Xiaoli Z. Fern, CS 434: Machine Learning and Data Mining, School of Electrical Engineering and Computer Science, Oregon State University, Fall 2008.

[DARPA, 2012] DARPA Machine Reading Program, http://www.darpa.mil/Our_Work/I2O/Programs/Machine_Reading.aspx.

[Mitchell, 2006] Tom M. Mitchell, The Discipline of Machine Learning, my perspective on this research field, July 2006 (http://www.cs.cmu.edu/~tom/pubs/MachineLearning.pdf).

[Mitchell, 1997] Tom M. Mitchell, Machine Learning. McGraw-Hill, 1997.

[Etzioni et al., 2007] Oren Etzioni, Michele Banko, and Michael J. Cafarella, Machine Reading.The 2007 AAAI Spring Symposium. Published by The AAAI Press, Menlo Park, California, 2007.

[Clark et al., 2007] Peter Clark, Phil Harrison, John Thompson, Rick Wojcik, Tom Jenkins, David Israel, Reading to Learn: An Investigation into Language Understanding. The 2007 AAAI Spring Symposium. Published by The AAAI Press, Menlo Park, California, 2007.

[Norvig, 2007] Peter Norvig, Inference in Text Understanding. The 2007 AAAI Spring Symposium. Published by The AAAI Press, Menlo Park, California, 2007.

[Wang & Cohen, 2007] Richard C. Wang and William W. Cohen: Language-Independent Set Expansion of Named Entities using the Web. In *Proceedings of IEEE International Conference on Data Mining* (ICDM 2007), Omaha, NE, USA. 2007.

[Etzioni, 2008] Oren Etzioni. 2008. Machine reading at web scale. In *Proceedings of the international conference on Web search and web data mining* (WSDM '08). ACM, New York, NY, USA, 2-2.

[Banko, et al., 2007] Michele Banko, Michael J. Cafarella, Stephen Soderland, Matthew Broadhead, Oren Etzioni: Open Information Extraction from the Web. IJCAI 2007: 2670-2676

References

[Weikum et al., 2009] G. Weikum, G., Kasneci, M. Ramanath, F. Suchanek. DB & IR methods for

knowledge discovery. Communications of the ACM 52(4), 2009.

[Theobald & Weikum, 2012] Martin Theobald and Gerhard Weikum. From Information to Knowledge: Harvesting Entities and Relationships from Web Sources. Tutorial at PODS 2012

[Hoffart et al., 2012] Johannes Hoffart, Fabian Suchanek, Klaus Berberich, Gerhard Weikum. YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia. Special issue of the Artificial Intelligence Journal, 2012

[Etzioni et al., 2011] Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, and Mausam "Open Information Extraction: the Second Generation“. *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011).*

[Hady et al., 2011] Hady W. Lauw, Ralf Schenkel, Fabian Suchanek, Martin Theobald, and Gerhard Weikum, "Semantic Knowledge Bases from Web Sources" at IJCAI 2011, Barcelona, July 2011

[Fader et al., 2011] Anthony Fader, Stephen Soderland, and Oren Etzioni. "Identifying Relations for Open Information Extraction”. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*

Settles, B.: Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In: Proc. of the EMNLP’11, Edinburgh, ACL (2011) 1467–1478 5.

Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Jr., E.R.H., Mitchell, T.M.: Toward an architecture for never-ending language learning. In: Proceedings of the Twenty-Fourth Conference on Artificial Intelligence (AAAI 2010).

Pedro, S.D.S., Hruschka Jr., E.R.: Collective intelligence as a source for machine learning self-supervision. In: Proc. of the 4th International Workshop on Web Intelligence and Communities. WIC12, NY, USA, ACM (2012) 5:1–5:9

References

[Appel & Hruschka Jr., 2011] Appel, A.P., Hruschka Jr., E.R.: Prophet – a link-predictor to learn new rules on Nell. In: Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops. pp. 917–924. ICDMW '11, IEEE Computer Society, Washington, DC, USA (2011)

[Mohamed et al., 2011] Mohamed, T.P., Hruschka, Jr., E.R., Mitchell, T.M.: Discovering relations between noun categories. In: Proceedings of the Conference on Empirical Methods in Nat- ural Language Processing. pp. 1447–1455. EMNLP '11, Association for Computa- tional Linguistics, Stroudsburg, PA, USA (2011)

[Pedro & Hruschka Jr., 2012] Saulo D.S. Pedro and Estevam R. Hruschka Jr., Conversing Learning: active learning and active social interaction for human supervision in never-ending learning systems. Xiii Ibero-american Conference On Artificial Intelligence, IBERAMIA 2012, 2012.

Krishnamurthy, J., Mitchell, T.M.: Which noun phrases denote which concepts. In: Proceedings of the Forty Ninth Annual Meeting of the Association for Compu- tational Linguistics (2011)

Lao, N., Mitchell, T., Cohen, W.W.: Random walk inference and learning in a large scale knowledge base. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. pp. 529–539. Associa- tion for Computational Linguistics, Edinburgh, Scotland, UK. (July 2011), http://www.aclweb.org/anthology/D11-1049

E. R. Hruschka Jr. and M. C. Duarte and M. C. Nicoletti. Coupling as Strategy for Reducing Concept-Drift in Never-ending Learning Environments. Fundamenta Informaticae, IOS Press, 2012.

Saulo D.S. Pedro, Ana Paula Appel, and Estevam R. Hruschka, Jr. Autonomously reviewing and validating the knowledge base of a never-ending learning system. In *Proceedings of the 22nd international conference on World Wide Web companion* (WWW '13 Companion), 1195-120, 2013.

S. Verma and E. R. Hruschka Jr. Coupled Bayesian Sets Algorithm for Semi-supervised Learning and Information Extraction. In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), 2012.

Navarro, L. F. and Appel, A. P. and Hruschka Jr., E. R., GraphDB – Storing Large Graphs on Secondary Memory. In New Trends in Databases and Information. Advances in Intelligent Systems and Computing, Springer, 177-186, 2013.

References

Assuming Facts Are Expressed More Than Once.
J. Betteridge, A. Ritter and T. Mitchell In Proceedings of the 27th International Florida Artificial Intelligence Research Society Conference (FLAIRS-27), 2014.

Estimating Accuracy from Unlabeled Data.
E. A. Platanios, A. Blum, T. Mitchell. In Uncertainty in Artificial Intelligence (UAI), 2014.

CTPs: Contextual Temporal Profiles for Time Scoping Facts via Entity State Change Detection.
D.T. Wijaya, N. Nakashole and T.M. Mitchell. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014.

Incorporating Vector Space Similarity in Random Walk Inference over Knowledge Bases.
M. Gardner, P. Talukdar, J. Krishnamurthy and T.M. Mitchell. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014.

Scaling Graph-based Semi Supervised Learning to Large Number of Labels Using Count-Min Sketch
P. P. Talukdar, and W. Cohen In 17th International Conference on Artificial Intelligence and Statistics (AISTATS, 2014.

Programming with Personalized PageRank: A Locally Groundable First-Order Probabilistic Logic.
W.Y. Wang, K. Mazaitis and W.W. Cohen. In Proceedings of the Conference on Information and Knowledge Management (CIKM), 2013.

Improving Learning and Inference in a Large Knowledge-base using Latent Syntactic Cues.
Matt Gardner, Partha Pratim Talukdar, Bryan Kisiel, and Tom Mitchell. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013), 2013.