Raising a Computer

Tom M. Mitchell

Machine Learning Department
Carnegie Mellon University

Where I'm Placing My Bets

- 50% brain imaging studies of human language processing
  - http://cs.cmu.edu/~fmri
- 50% build and raise a never-ending learning agent to read the web
  - http://rtw.ml.cmu.edu
  - http://rtw.ml.cmu.edu/rtw/kbbrowser/city:san_juan

Points of this talk

1. A good path toward strong AI is to raise computers that learn many things over years, with some human guidance

2. Raising a computer to read the web is a path toward disruptive AI progress

Thesis: We will never really understand learning or intelligence until we raise machines that

- learn many different things,
- from years of diverse experience,
- in a staged, curricular fashion,
- and become better learners over time.

How many examples can we point to today?

NELL: Never-Ending Language Learner

Task:

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

Inputs:

- initial ontology (categories and relations)
- dozen examples of each ontology predicate
- the web
- occasional interaction with human trainers

NELL today

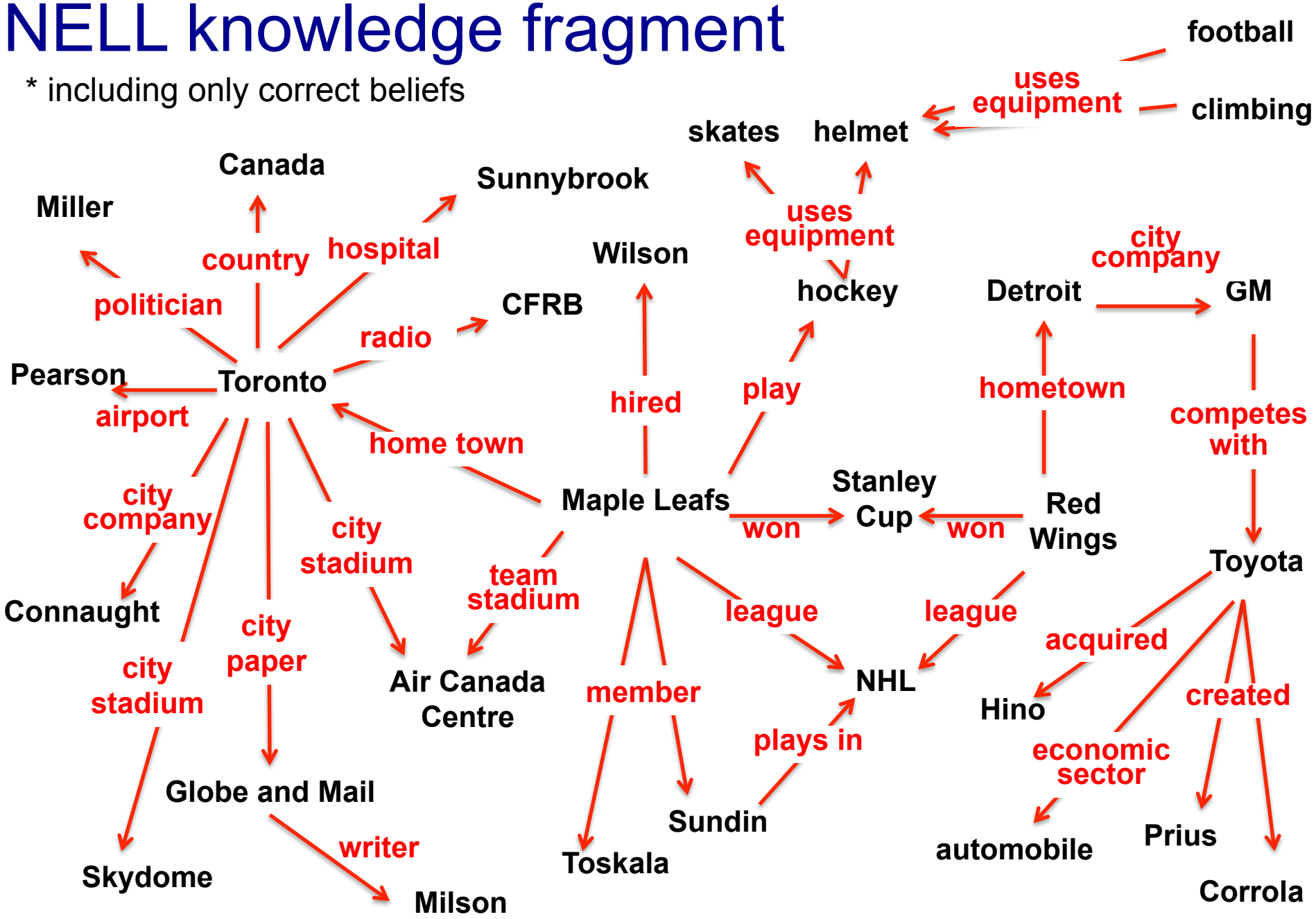
Running 24x7, since January, 12, 2010

Result:

- knowledge base with 90 million candidate beliefs
- learning to read
- learning to reason
- extending its ontology

NELL knowledge fragment

* including only correct beliefs



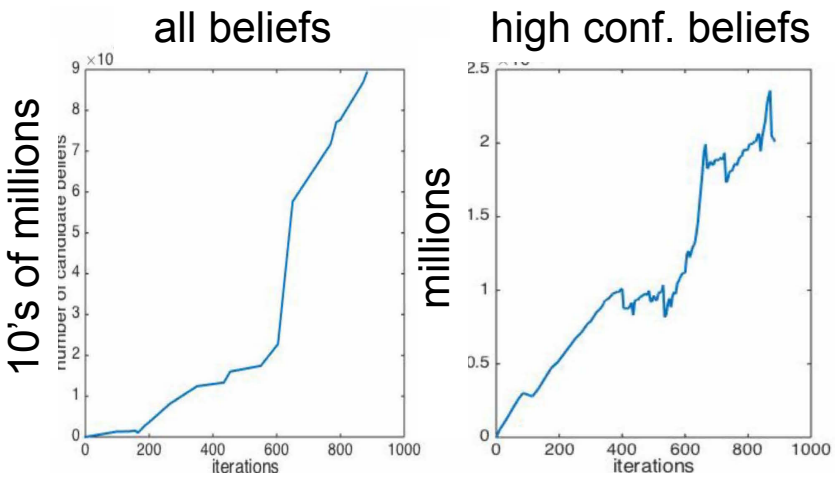
NELL Today

- eg. "diabetes", "Avandia", "tea", "IBM", "love" "baseball" "San Juan" "jeans" "BacteriaCausesCondition" "kitchenItem" "ClothingGoesWithClothing" …

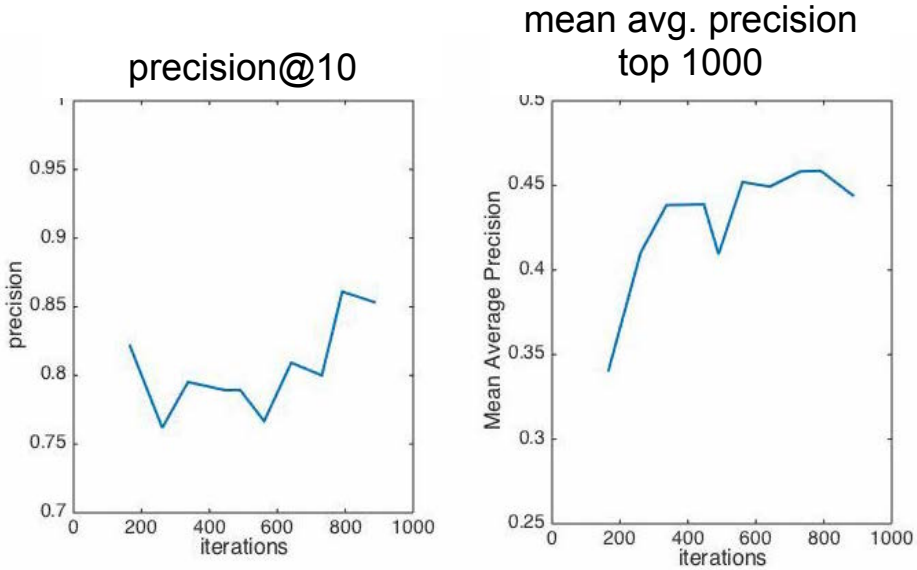
Recently-Learned Facts twitter

instance	iteration	date learned
zillion_stars is a geometric shape	893	02-jan-2015
many_other_books is a kind of media	892	30-dec-2014
street_fighter_2__champion_edition is software	889	07-dec-2014
spicy_coconut_yogurt_chicken_breasts is a type of meat	889	07-dec-2014
infill_walls is something found in or on buildings	889	07-dec-2014
state_university is a sports team also known as notre_dame	892	30-dec-2014
harrods is a tourist attraction in the city london	893	02-jan-2015
weiskopf plays the sport golf	893	02-jan-2015
hat is a clothing item to go with coveralls	889	07-dec-2014
james_cameron directed the movie titanic	892	30-dec-2014

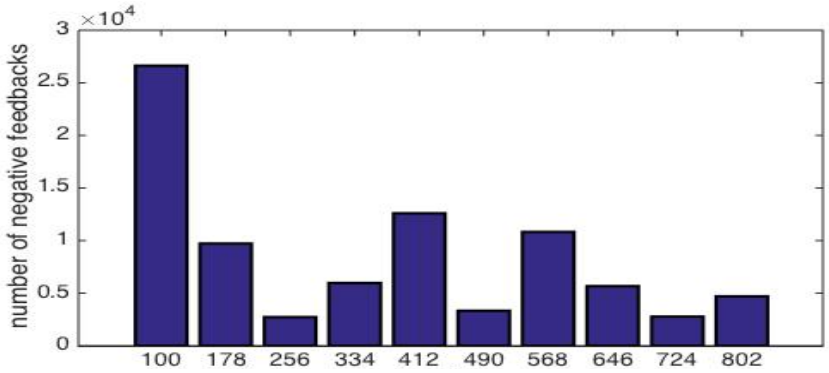
NELL Is Improving Over Time (Jan 2010 to Nov 2014)



number of NELL beliefs vs. time

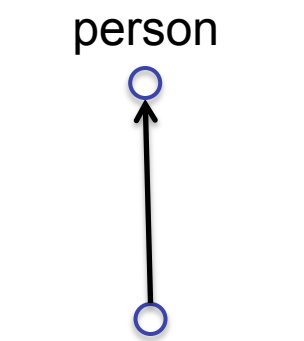


reading accuracy vs. time
(average over 31 predicates)

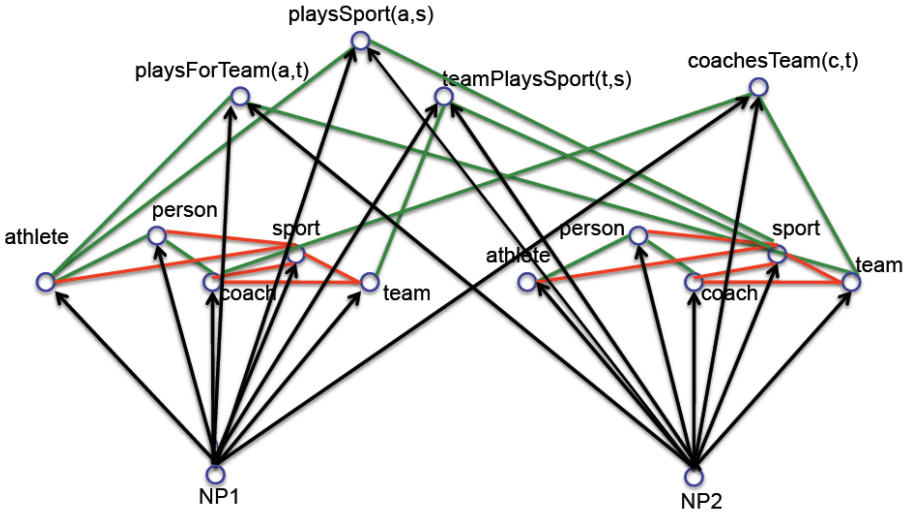


human feedback vs. time
(average 2.4 feedbacks per predicate per month)

Key Idea 1: Coupled semi-supervised training of many functions, from 99.9% unlabeled data



hard (underconstrained) semi-supervised learning problem



much easier (more constrained) semi-supervised learning problem

NELL: Learned reading strategies

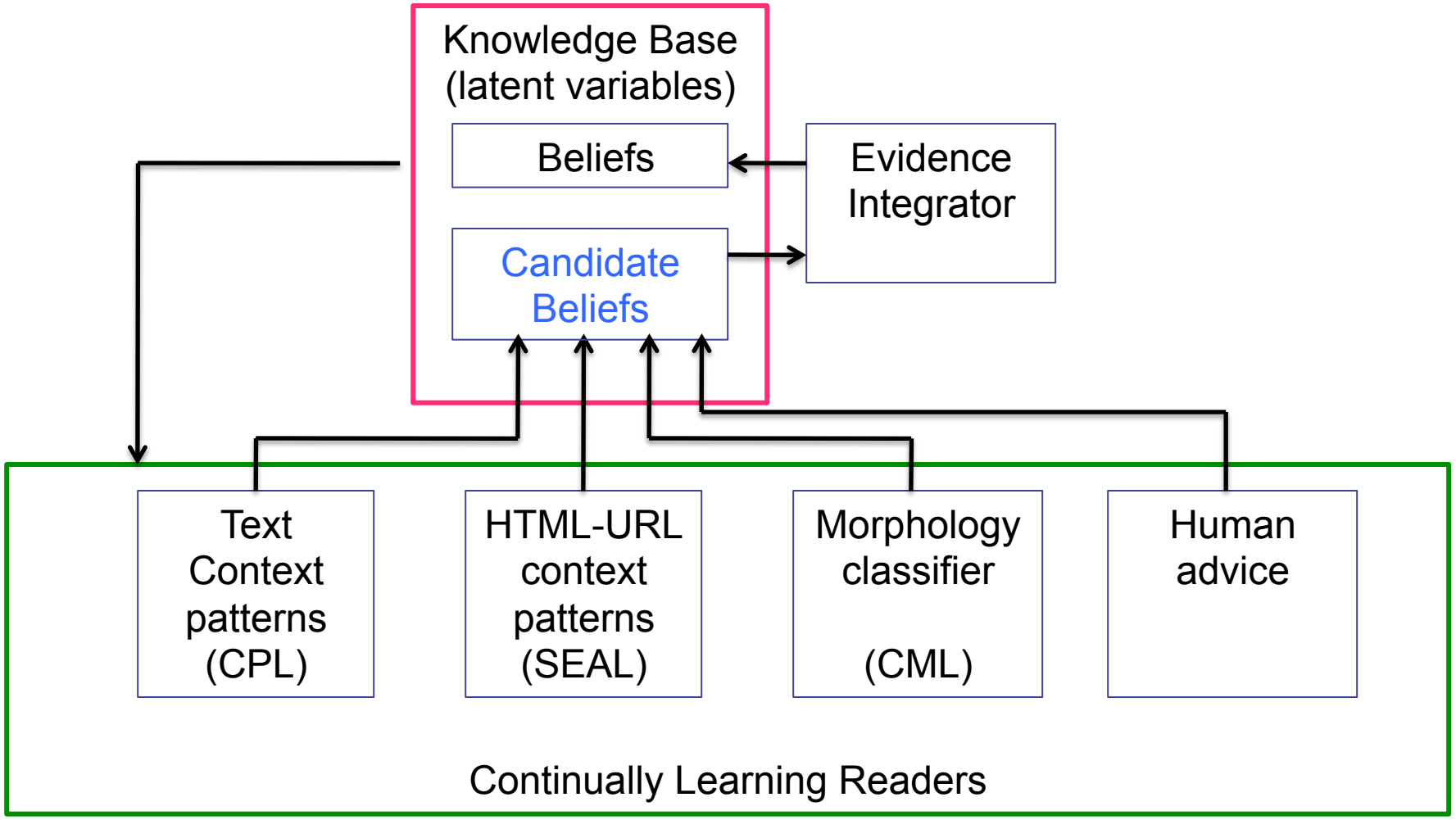
Mountain:

"volcanic crater of _" "volcanic erupt
region of _" "volcano , called _" "vo
"volcano known as _" "volcano Mt _
including _" "volcanoes , like _" "vo
_" "volcanoes including _" "volcano
"weather atop _" "weather station at
through _" "West face of _" "West r
ledge in _" "white summit of _" "wh
surrounding _" "wilderness areas ar
"winter ascents in _" "winter ascents
foothills of _" "world famous view of
popping by _" "you 've just climbed
"_ ' crater" "_ ' eruption" "_ ' foothills
Camp" "_ 's drug guide" "_ 's east r
Face" "_ 's North Peak" "_ 's North
southeast ridge" "_ 's summit calder
's west ridge" "_ (D,DDD ft" " "_ clim
consult el diablo" "_ cooking planks'

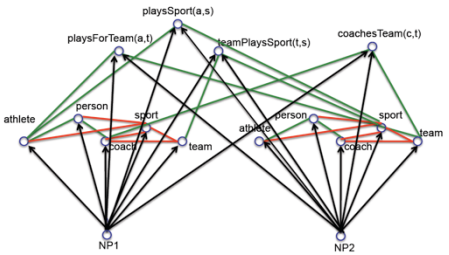
Predicate	Feature	Weight
mountain	LAST=peak	1.791
mountain	LAST=mountain	1.093
mountain	FIRST=mountain	-0.875
musicArtist	LAST=band	1.853
musicArtist	POS=DT_NNS	1.412
musicArtist	POS=DT_JJ_NN	-0.807
newspaper	LAST=sun	1.330
newspaper	LAST=university	-0.318
newspaper	POS=NN_NNS	-0.798
university	LAST=college	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

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/	`</li> <li>[X] by [Y] &#8211;`

Initial NELL Architecture



Key Idea 2:



Learn to infer not-yet-read beliefs

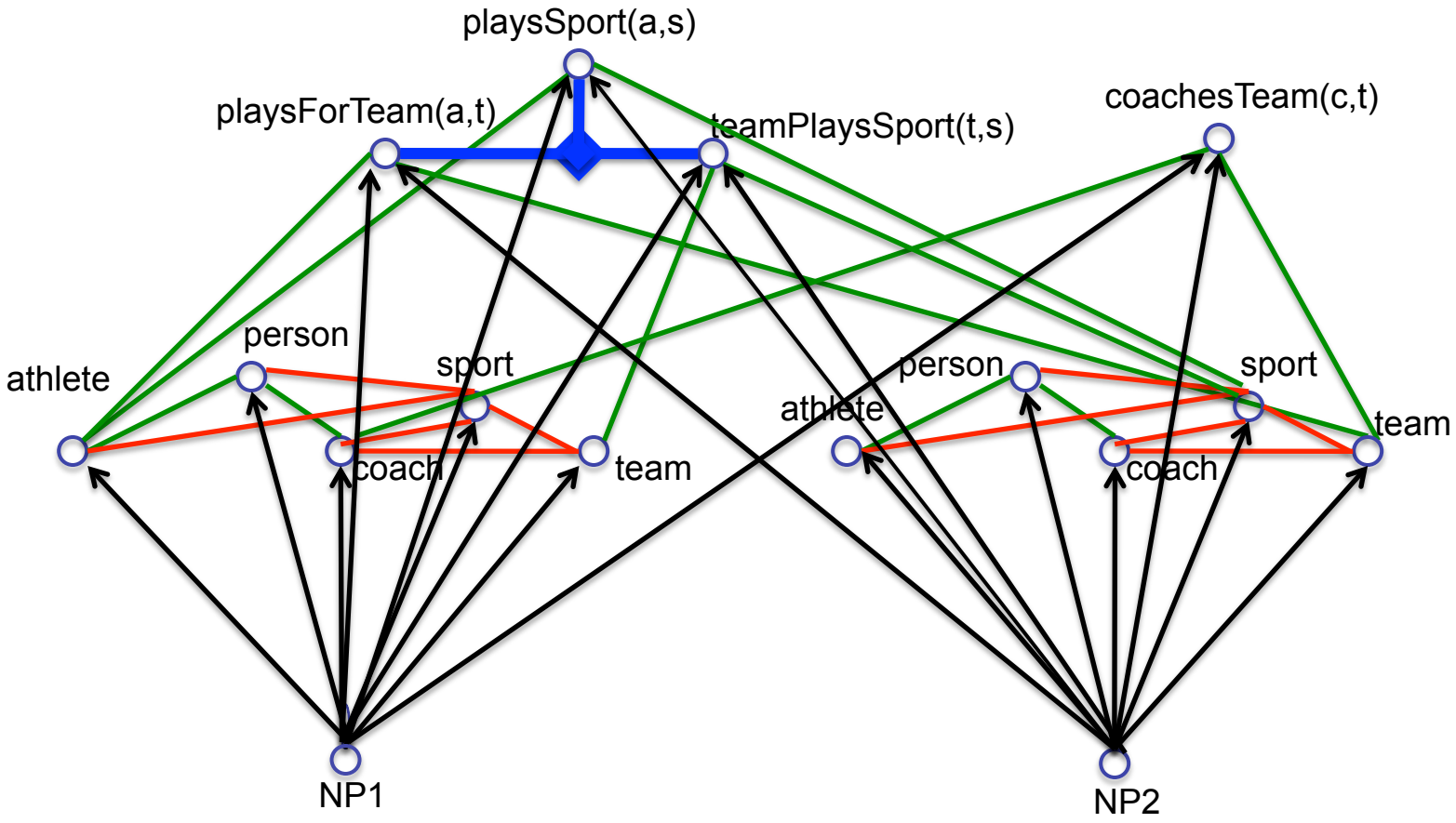
- first order, probabilistic horn clause constraints:

0.93 athletePlaysSport(?x,?y) ← athletePlaysForTeam(?x,?z)
teamPlaysSport(?z,?y)

  - learned by data mining the knowledge base
  - connect previously uncoupled relation predicates
  - infer new unread beliefs
  - NELL has 100,000s of learned rules
  - uses PRA random-walk inference [Lao, Cohen, Gardner]

Learned Probabilistic Horn Clause Rules

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



Key Idea 3:
Automatically extend ontology

Example Discovered Relations

[Mohamed et al. *EMNLP* 2011]

Category Pair	Frequent Instance Pairs	Text Contexts	Suggested Name
MusicInstrument Musician	sitar, George Harrison tenor sax, Stan Getz trombone, Tommy Dorsey vibes, Lionel Hampton	ARG1 master ARG2 ARG1 virtuoso ARG2 ARG1 legend ARG2 ARG2 plays ARG1	Master
Disease Disease	pinched nerve, herniated disk tennis elbow, tendonitis blepharospasm, dystonia	ARG1 is due to ARG2 ARG1 is caused by ARG2	IsDueTo
CellType Chemical	epithelial cells, surfactant neurons, serotonin mast cells, histomine	ARG1 that release ARG2 ARG2 releasing ARG1	ThatRelease
Mammals Plant	koala bears, eucalyptus sheep, grasses goats, saplings	ARG1 eat ARG2 ARG2 eating ARG1	Eat
River City	Seine, Paris Nile, Cairo Tiber river, Rome	ARG1 in heart of ARG2 ARG1 which flows through ARG2	InHeartOf

Sequence of Self-Reinforcing Competencies

1. Learning from 99.9% unlabeled data
   - learn thousands of different but coupled reading functions
   - redundancy on the web, seek internal consistency

2. Learn to infer/predict new unread beliefs
   - 100,000's of learned probabilistic inference rules
   - playsSport(a,s) ← playsForTeam(a,t), teamPlays(t,s)

3. Automatically extend representation
   - invent new relational predicates
   - riverFlowsThroughCity(x,y), bacteriaCausesCondition(x,y)

Key Idea: Curriculum for 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

NELL is here

8. Learn to microread single sentences
9. Self-reflection, goals, subgoals, self-directed activity
10. Goal-driven reading: predict, then read to corroborate/correct
11. Make NELL a conversational agent on Twitter
12. Add a robot body to NELL

Missing Competencies in NELL

1. Deep understanding of individual sentences
2. Self-reflection about current skill levels
3. Understanding of time
4. Goals, subgoals, focus of attention
5. Grounding in non-language data
6. No body
7. …, …, …, …!

Points of this talk

1. A good path toward strong AI is to raise computers that learn many things over years, with some human guidance

2. Raising a computer to read the web is a path toward disruptive AI progress

“The reason we don’t have computers that truly understand natural language is that it requires huge amounts of world knowledge.”

– thousands of NLU researchers & linguists

“Mary caught the butterfly with the net.”

“Mary caught the butterfly with the spots.”

thank you

and thanks to:
Darpa, Google, NSF, Yahoo!, Microsoft, Fulbright, Intel

follow NELL on Twitter: @CMUNELL
browse/download NELL's KB at http://rtw.ml.cmu.edu