Commonsense for Machine Intelligence: Text to Knowledge and Knowledge to Text

Gerard de Melo, Niket Tandon and Aparna S. Varde

RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

AI2 ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

MONTCLAIR STATE UNIVERSITY

acm Association for Computing Machinery

Research Tutorial at ACM CIKM 2017

Singapore, Nov 6 to 10, 2017

CIKM 2017 SINGAPORE

Tutorial Presenters

Gerard de Melo, Assistant Professor, Rutgers University, New Brunswick, NJ, USA

RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

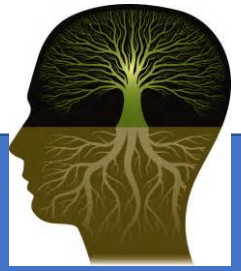
Niket Tandon, Research Scientist, Allen Institute for Artificial Intelligence, Seattle, WA, USA

AI2 ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

Aparna S. Varde, Associate Professor, Montclair State University, Montclair, NJ, USA

MONTCLAIR STATE UNIVERSITY

Tutorial Agenda



Part 1: Acquiring Commonsense Knowledge

Part 2: Detecting and Correcting Odd Collocations in Text

Part 3: Applications and Open Issues

Part 1: Acquiring Commonsense Knowledge

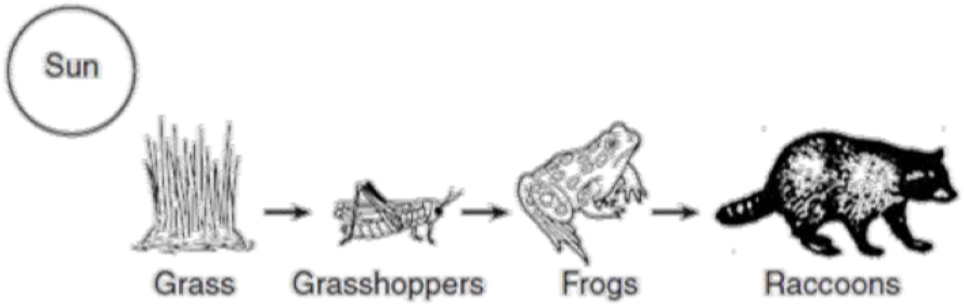
Commonsense for Machine Intelligence: Text to Knowledge and Knowledge to Text

Part 1: Acquiring Commonsense Knowledge

- **introduction**
  - introduction to csk
  - csk unimodal and multimodal kbs
- csk representation
  - discrete and continuous representations
  - multimodal continuous representations
- acquisition methods
  - different levels of supervision and modalities
  - from facts to rules
- csk evaluation
  - explicit evaluation techniques: sampling, turked
  - challenge sets and problems in text and vision

What is commonsense?

Questions from Aristo challenge (allenai.org/data)

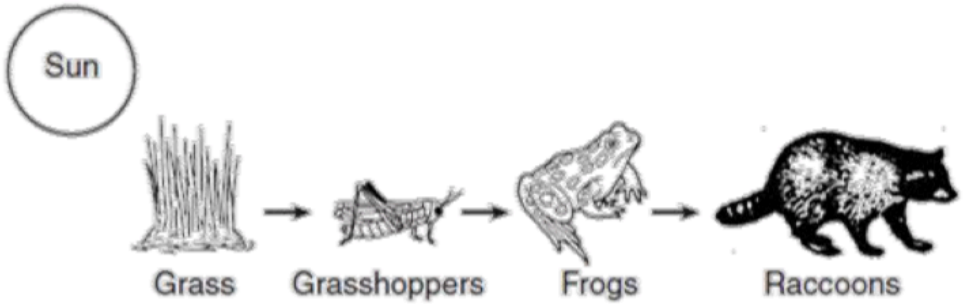


If all the frogs died, the raccoon population would most likely
(A) **decrease** (B) increase (C) remain the same

For roller-skate race, what is the best surface?
(A) sand (B) grass (C) **blacktop**

Common knowledge about things in the world, their associations, and interactions.
Commonsense knowledge is mostly location and culture independent.

Machines cannot reason like humans, because they lack commonsense.



If all the frogs died, the raccoon population would most likely
(A) **decrease** (B) increase (C) remain the same

For roller-skate race, what is the best surface?
(A) sand (B) grass (C) **blacktop**

Picture depicts food web. Arrow indicates consumes. If frogs die, raccoons won't get food and die-- so their population will decrease.

Humans

Human- Machine Knowledge Gap

Machines

Roller skate is best on a smooth surface. Blacktop surfaces are shiny and shiny surfaces are smooth.

Not seen those arrows enough so cannot generalize. Haven't seen raccoons, frogs in a sentence frequently

grass – related to – field – race , so "grass"
blacktop is not related to race.

What about the Knowledge graphs?

Machines

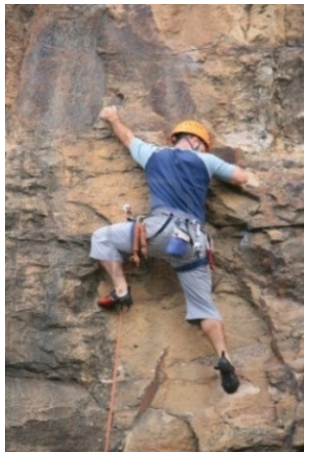
Alexander J. Honnold

Personal information	
Born	August 17, 1985 (age 29)
Education	UC Berkeley (dropped out)
Occupation	Professional rock climber
Climbing career	
Type of climber	• Free solo • Big wall
Highest grade	Redpoint: 5.14c (8c+) Bouldering: V12 (8A+)
Known for	Big Wall Free Soloing Speed record on *The Nose* of El Capitan

Machines can surpass most humans on Encyclopedic knowledge about popular "named entities"

Humans

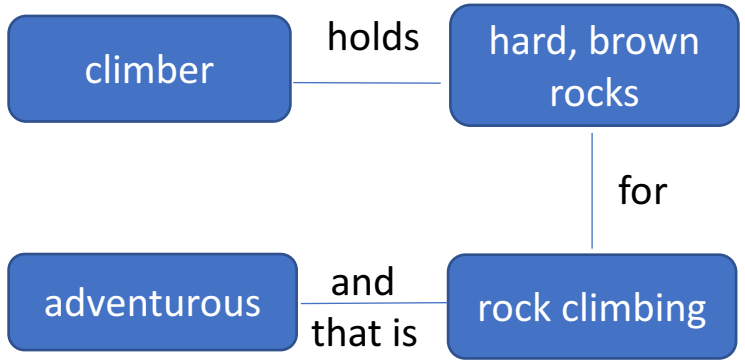
Personal information	
Born	?
Education	
Occupation	
Climbing career	
Type of climber	?
Highest grade	
Known for	



1 Rock | 2 Hands

1 Person | 2 Legs

Machines cannot surpass any human on commonsense knowledge about "common nouns"



Commonsense knowledge acquisition is hard

Elusive

Less, implicitly expressed

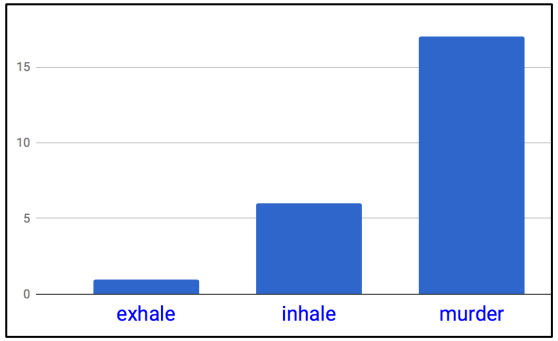
[Tandon et. al Ngram workshop 2010]

I touched a table that was hard.

roots absorb water → water is at the roots

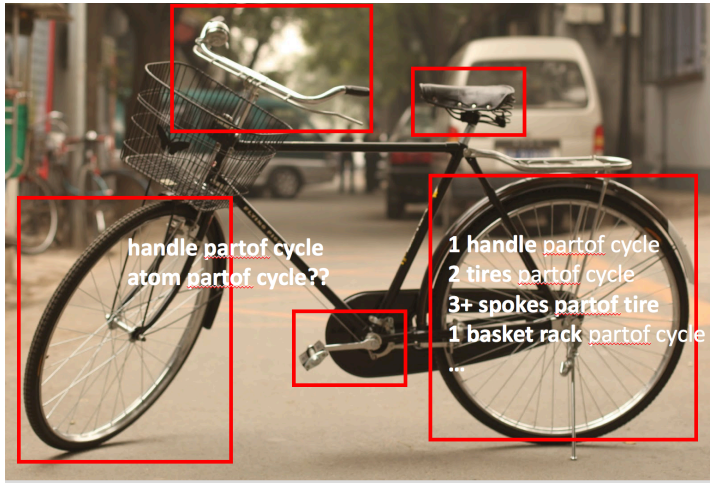
Reporting Bias

[Parikh EACL 2017, Gordon et. al AKBC 2013]



Multimodal

[Tandon et. al AAAI 2016]



Commonsense knowledge acquisition is hard

Elusive

Less, implicitly expressed

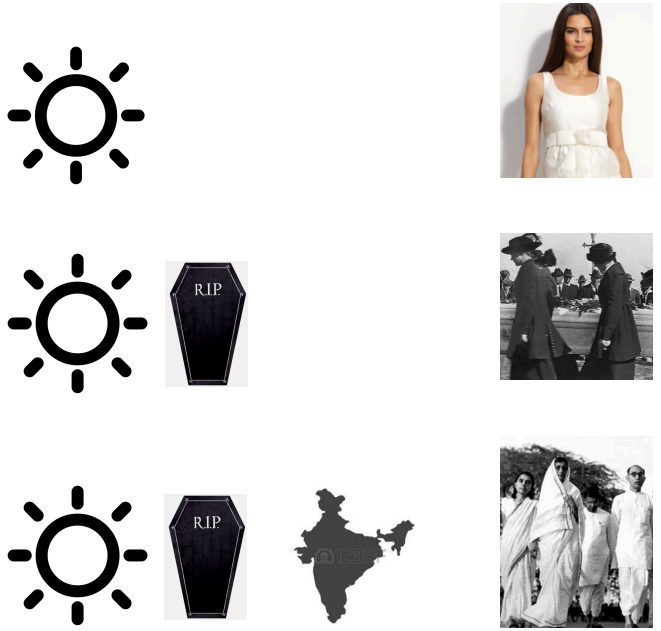
Reporting Bias

Multimodal

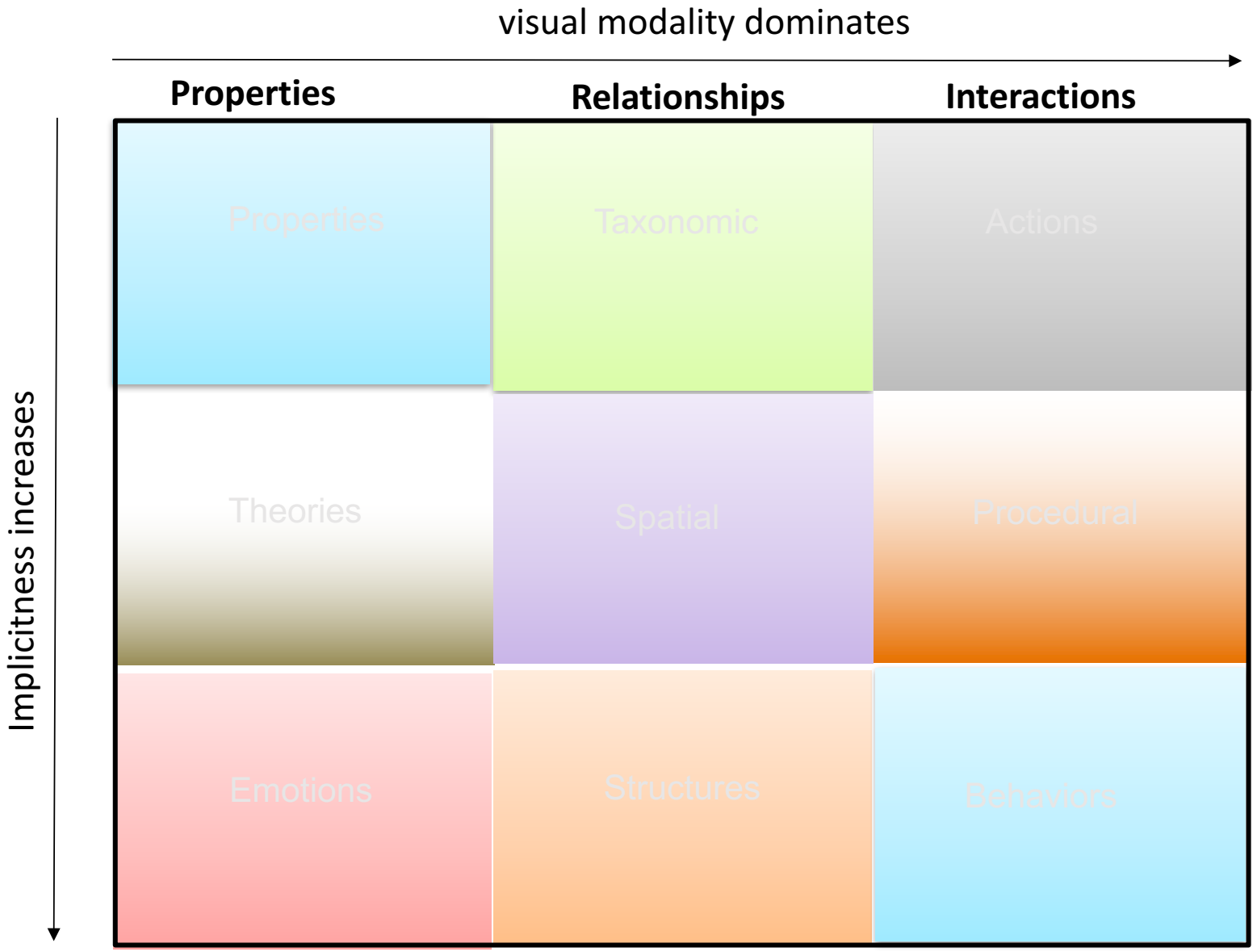
Contextual

Depends on the context

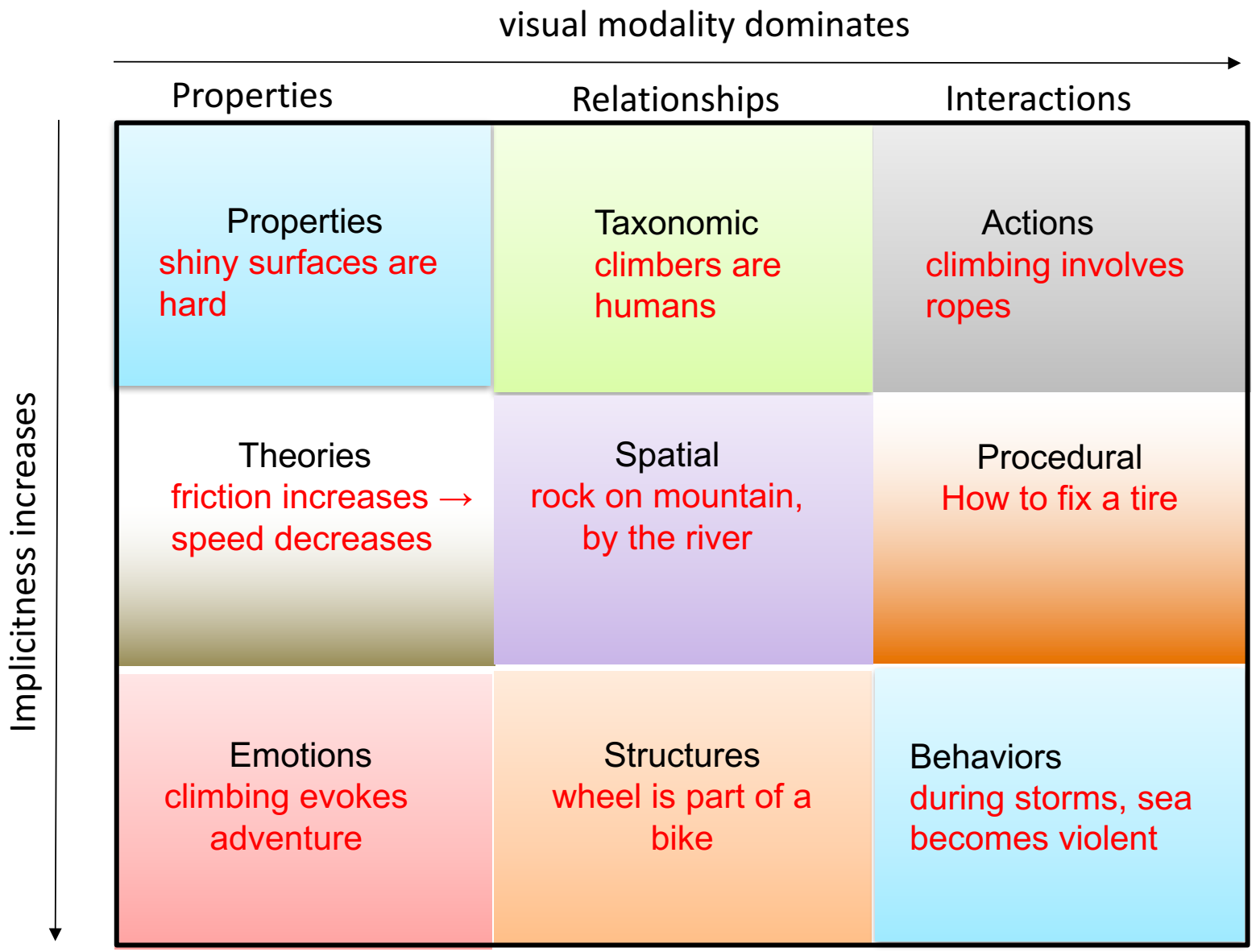
[Liu et. al 2004]



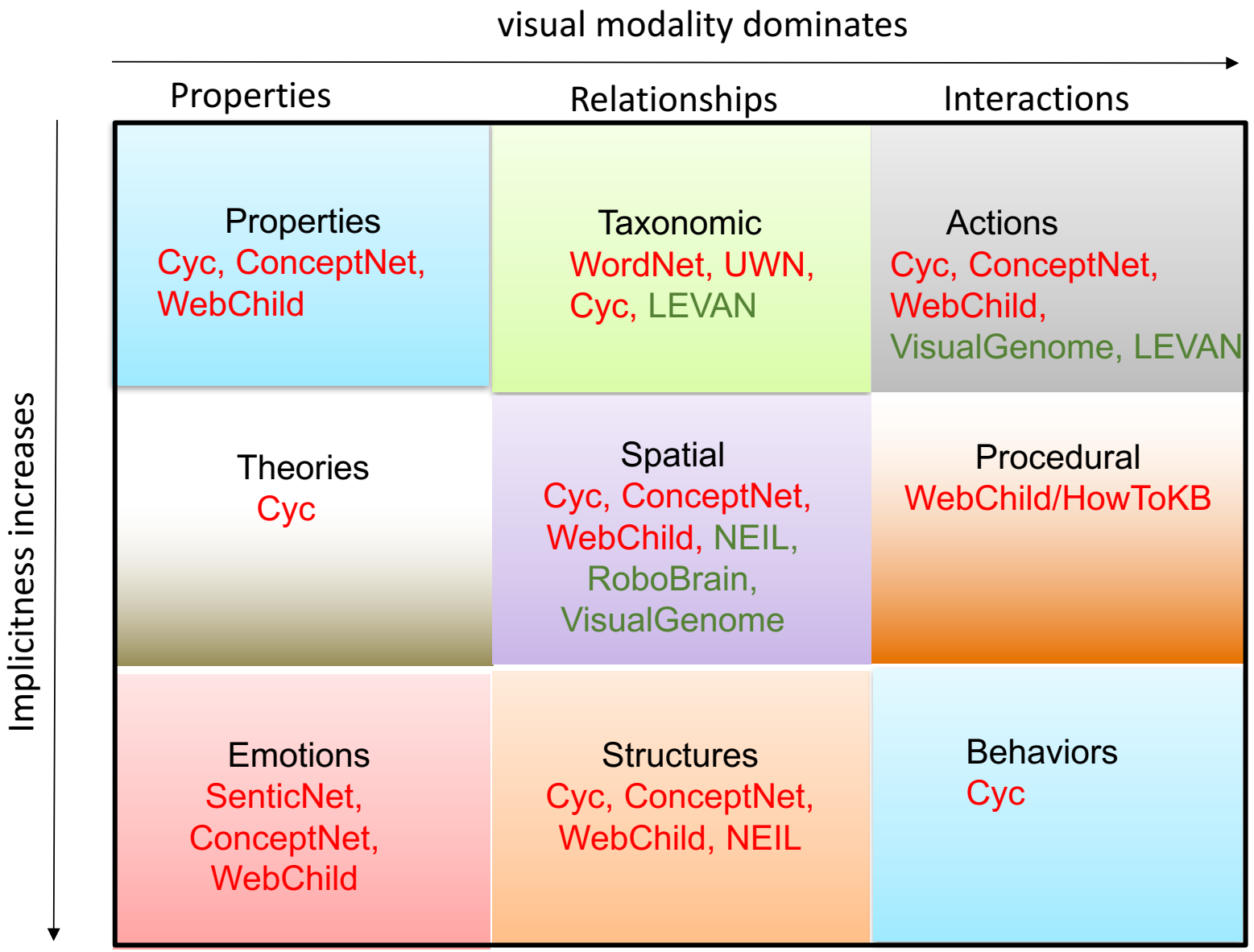
Commonsense knowledge types in multiple modalities



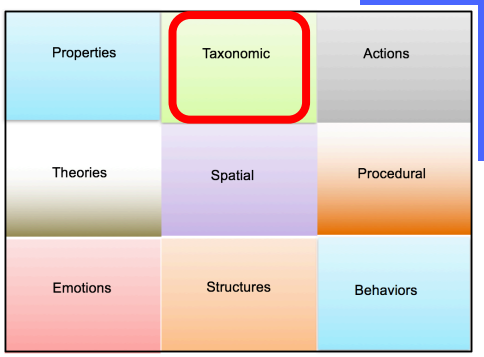
Commonsense Knowledge types

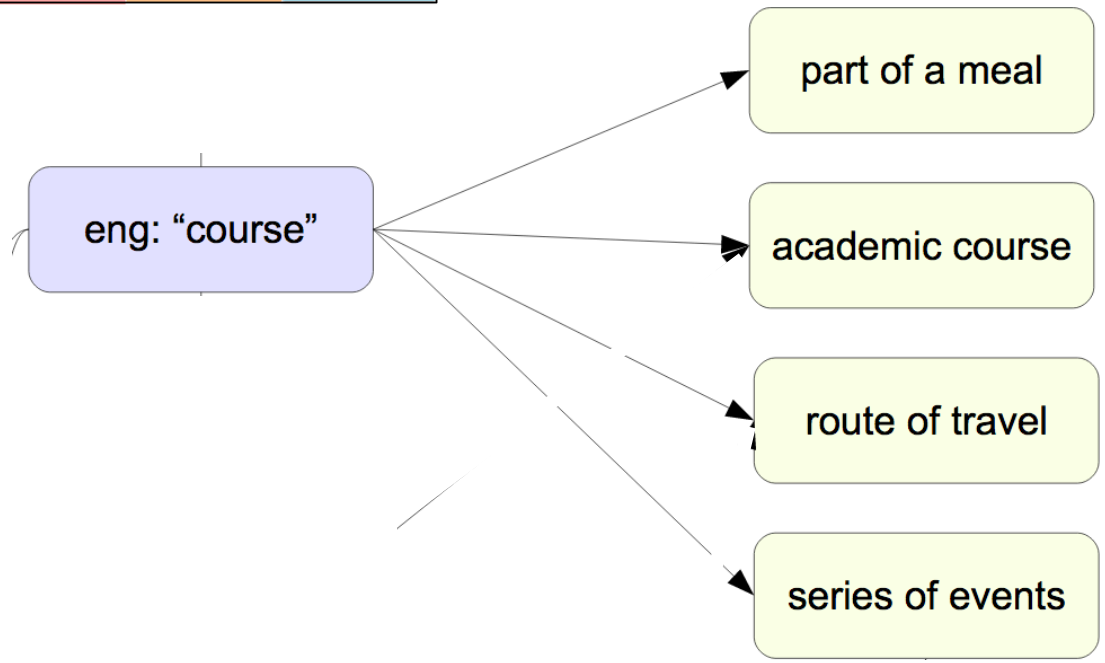


Commonsense Knowledge KBs

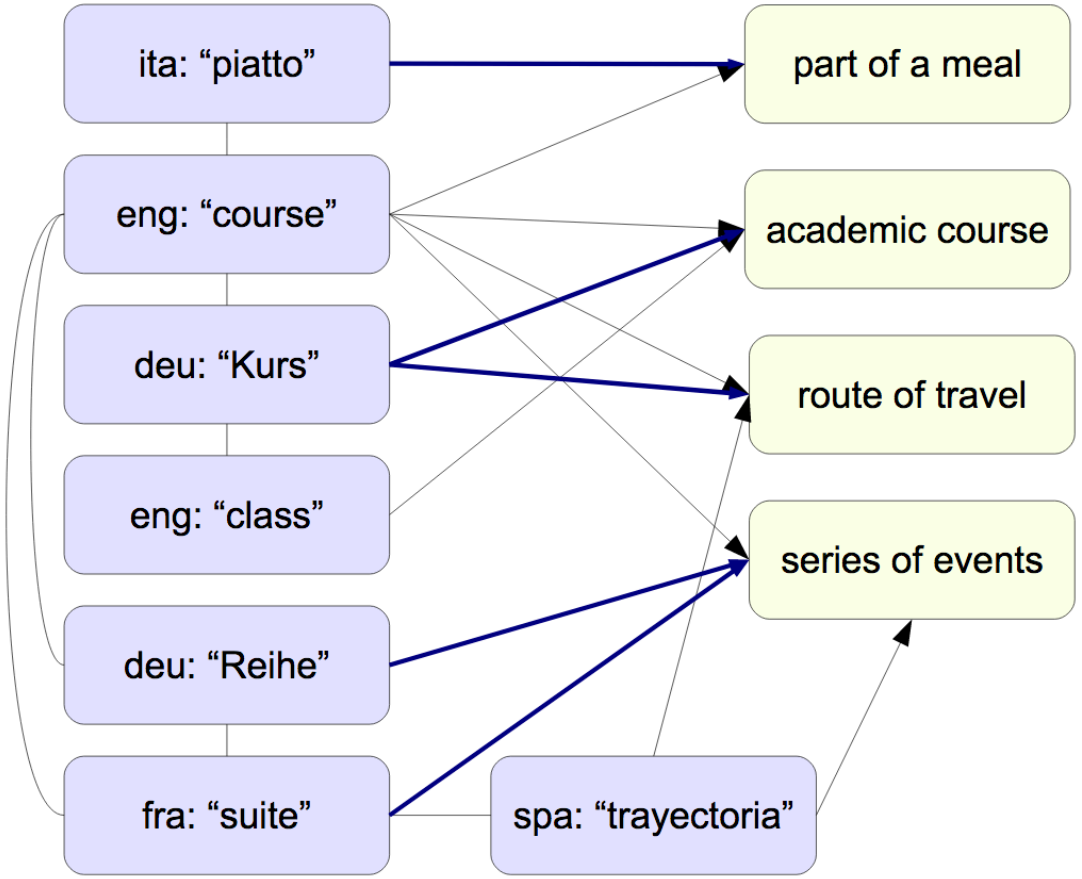


WordNet
[Miller 1998]

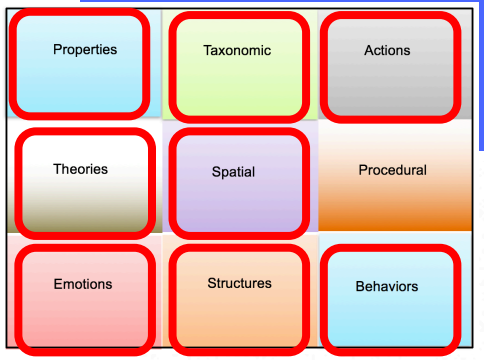




Universal WordNet
[de Melo et. al 2009]



Note: There are many more taxonomies such as Microsoft ProBase



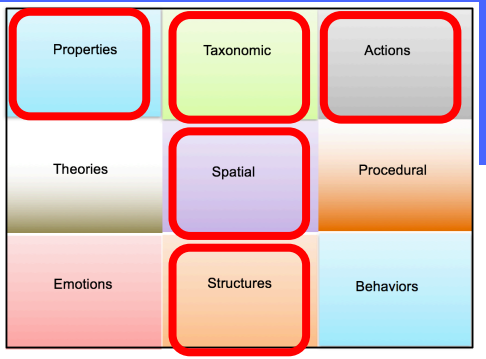
Cyc

[Lenat 1995]

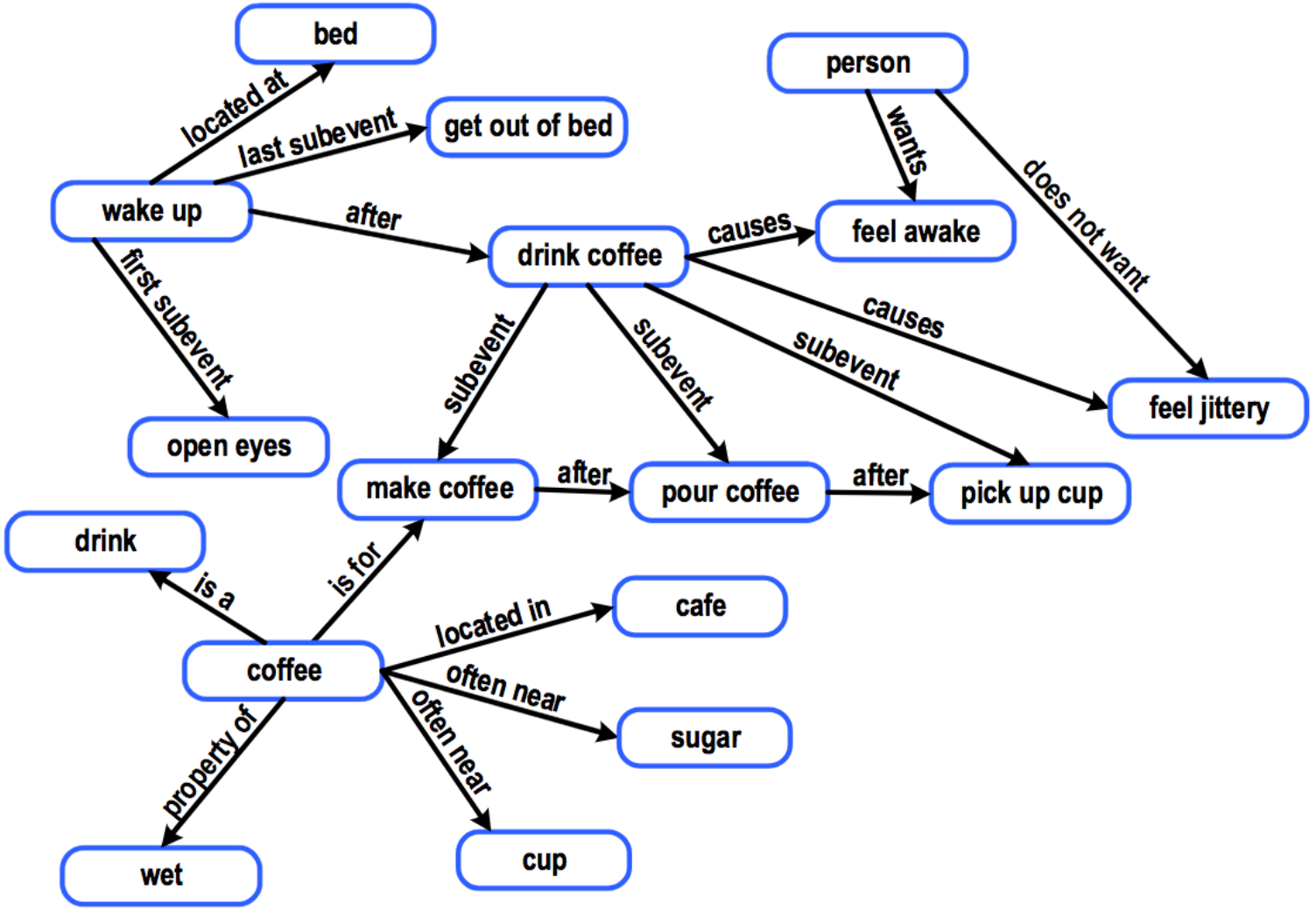
Thing
Intangible Thing
Individual
Relations
Sets
Collections
Space
Paths
Logic
Time
Spatial Parts
Math
Events & Scripts
Physcal Objects
Borders Geometry
Agents
Materials, Parts & Statics
Artifacts
Weather
Movement
Actors & Actions
Organization
State Change Dynamics
Natural Geography
Living Things
Plans & Goals
Organizational Actions
Types of Organization
Chemistry
Sentient Beings
Ecology
Physical Agents
Organizational Plan
Human Activitis
Political Geography
Physical Artifacts
Human Anatomy & Physiology
Politics & Warefare
Nations Governments and Geopolitics
Emotion, Perception & Belief
Plants Animals
Law
Business & Commerce
Purchasing Shopping
Astronomy
Conceptual Works
Products & Devices
Literature & Works of Art
Human Behavior & Actions
Social Behaviour
Sports, Recreation & Entertainment
Professions Occupations
Business
Earth & Solar System
Software
Vehicles, Buildings & Weapons
Mechanica & Electrical Devices
Language
Social Relations & Culture
Social Activities
Transportaiton & Logistics
Travel & Communication
Everyday Living
Military Organizations

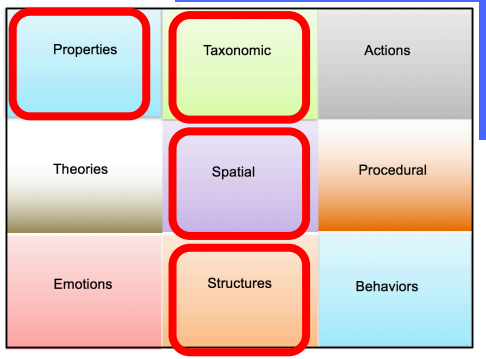
Domain-Specific Knowledge

(e.g., Healthcare, Computer Security, Command and Control, Mortgage Banking, ...)



ConceptNet [Liu et. al 2004]

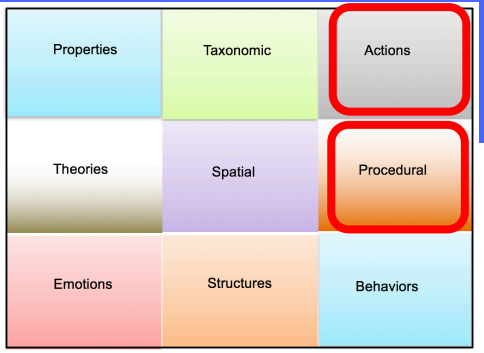




Aristo tuple KB [Dalvi et. al 2017]

Tree

has-part	trunk, bark, vacuole, cell membrane, cell, nucleus, treetop, xylem, stump, cytoplasm, tree branch, corpus, plasma membrane, section, leaf node
feature	trunk, massive trunk
carry	leaf
occupy	habitat, rocky habitat
use	photosynthesis



Action and procedural knowledge KBs

Generic

HowToKB

Syntactic structures

OpenIE

ConceptNet

Fillmore'76

Schank'75

Minsky'74

Reduced expressivity

Semantic expressivity

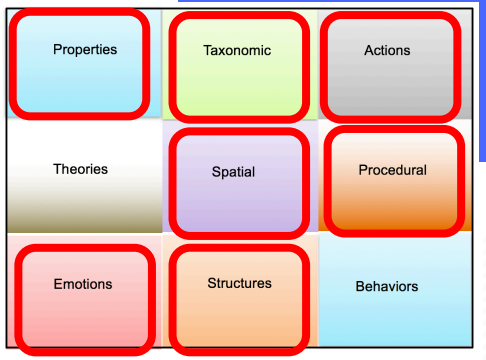
PropBank

VerbNet

FrameNet

Knowlywood

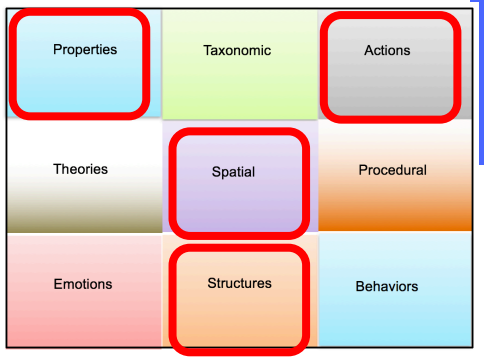
Domain specific



WebChild [Tandon et. al 2014]

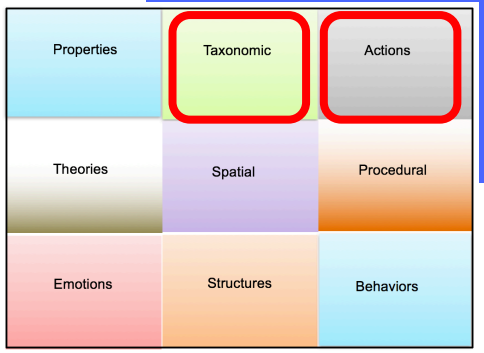
mountain: *a land mass that projects well above its surroundings; higher than a hill*

TYPE OF	natural elevation
	size to **object**, under the category of **mountaineering**
PHYSICAL PROPERTIES	large high heavy cold hard More
ABSTRACT PROPERTIES	elegant old safe holy risky More
COMPARABLES	mountain,hill mountain,mount mountain, high hill valley,mountain More
HAS PHYSICAL PARTS	mountain peak mountainside slope tableland hill More
HAS SUBSTANCE	mixture metallic element material page wood More
IN SPATIAL PROXIMITY	coast tunnel lake sea river More
ACTIVITIES	climb mountain cross mountain move mountain see mountain ascend mountain



Visual Genome [Krishna et. al 2016]

Regions	Attributes	Relationships
A wall on the side of a building	leaf is green and yellow	light reflecting on leaves
green leaf with yellow spots	stems is green	
yellowish green stems	light is bright	
bright light reflecting on leaves	leaf is dark	
a dark green leaf with brown edges	leaf is green	
a large green leaf	leaf is brown edge	
a large dark green leaf	leaf is large	
	leaf is dark green	
	leaf is yellow, brown, and green	
	pale flower is pale	



LEVAN

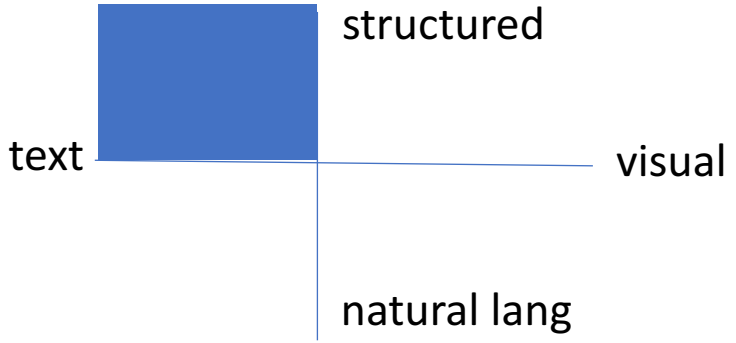
[Divvala et. al 2014]

object (horse), scene (kitchen), event (xmas), action (walking)

Part 1: Acquiring Commonsense Knowledge

- introduction
  - introduction to csk
  - csk unimodal and multimodal kbs
- **csk representation**
  - discrete and continuous representations
  - multimodal continuous representations
- acquisition methods
  - different levels of supervision and modalities
  - from facts to rules
- csk evaluation
  - explicit evaluation techniques: sampling, turked
  - challenge sets and problems in text and vision

Reasoning inspired formal representations



Formal representations

Classical deductive reasoning requires formal representations for syllologisms
Cyc's microtheories is a classic example of formal representations.

All trees are plants
(#$genls #$Tree-ThePlant #$Plant) \;

For roller-skate race, what is the best surface?
(A) sand (B) grass (C) **blacktop**

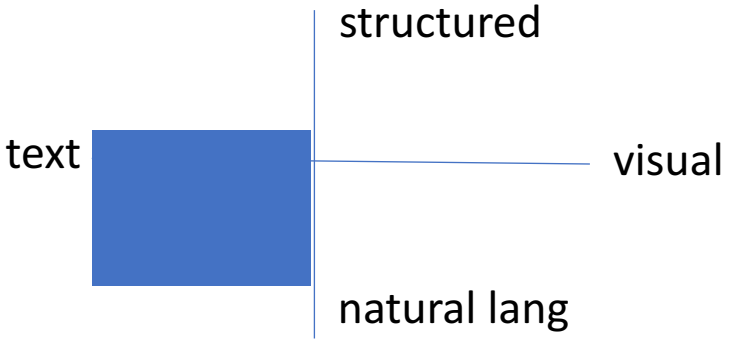
blacktop surfaces are shiny
shiny surfaces are smooth
smooth surfaces have less friction
less friction speeds up roller-skates
you win race when fastest

Problems

- Require perfect knowledge for heavy duty deductive inference, and,
- Far from natural language, query must be translated to this representation.

Representation in Natural Language

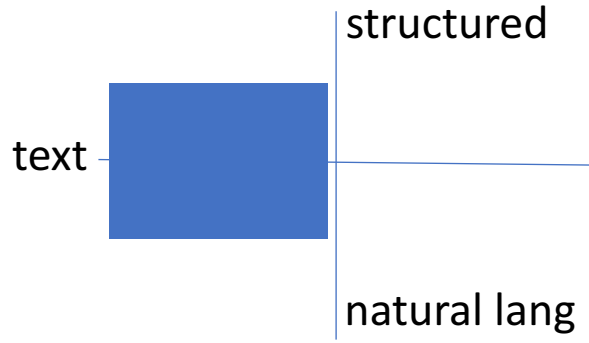
ConceptNet has a string representation,
over a fixed set of relations.



+ Admits an open vocabulary

- **difficult knowledge retrieval and generalization,** especially under reporting bias.
  - plants, absorb, solar energy
  - trees, take in, sunlight

Representation in Natural Language



Aristo TupleKB and WebChild use ILPs
to cluster for structure by maximizing for coherence,
overcoming reporting bias by aggregating frequencies.

<wet wood, softer than, dry wood> ≡ <dry wood, harder than, wet wood>

max (triple internal coherence + sense frequency)

max (triple internal coherence + sense frequency)

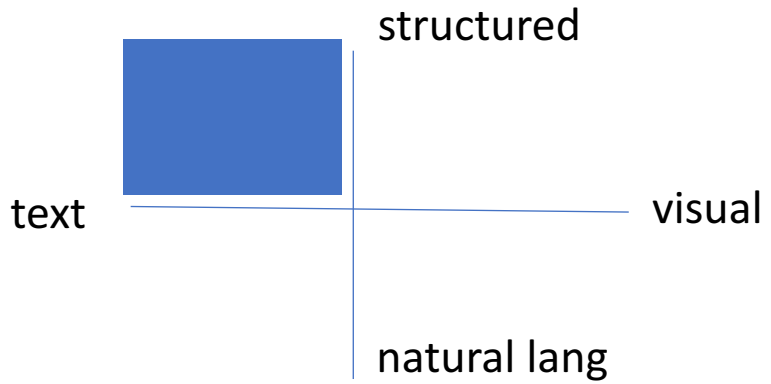
<wet wood-n-1,
softer-a-1 than,
dry wood-n-1>

≡ in WN space

<dry wood-n-1,
harder-a-1 than,
wet wood-n-1>

Frame based representation

Higher arity tuples,
such as OpenIE [Etzioni et. al 2011] :
<s, p ,o, location ,time>



further structured into frame based representations
to retain more context.

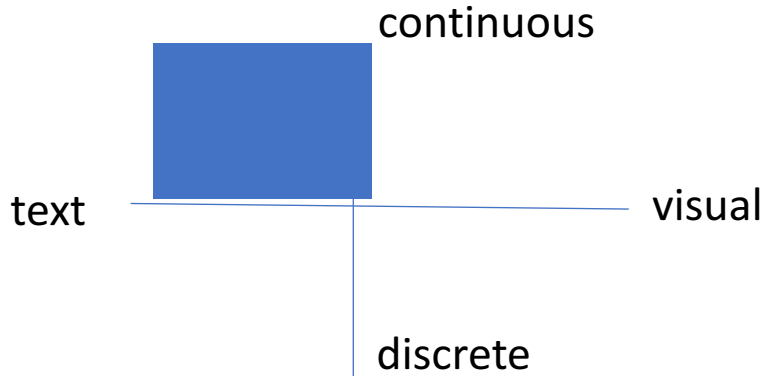
+ useful when a relation admits multiple values
+ allows for maintaining top-k values by salience

{Climb up a mountain	, Hike up a hill}
Participants	climber, boy, rope
Location	camp, forest, sea shore
Time	daylight, holiday

Continuous representations.

Why continuous representation?

Knowledge retrieval over discrete representations suffers from linguistic variations



Using matrix and tensor factorizations [Bordes et. al 2014] , continuous representations in the embedding space has made progress for Encylopedic KB completion. [Jain et. al 2017] found that progress across popular data sets does not generalize.

A recent approach injects knowledge in a memory cell, KB-LSTMs [Yang et. al 2017] such that the hidden vector learned per word is biased with KB context. Gains were limited, but this is a promising direction.

Negative data assumption in training these models

Encyclopedic Knowledge

If (e_i, r_k, e_j) holds, then

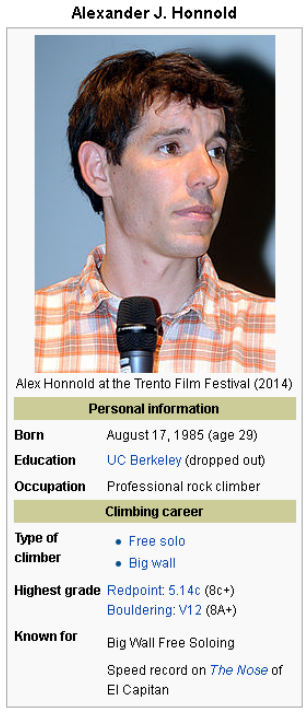
(e_i, r_k, $e_{j'}$!= e_j) is -ve

A. Honnold, bornIn, US

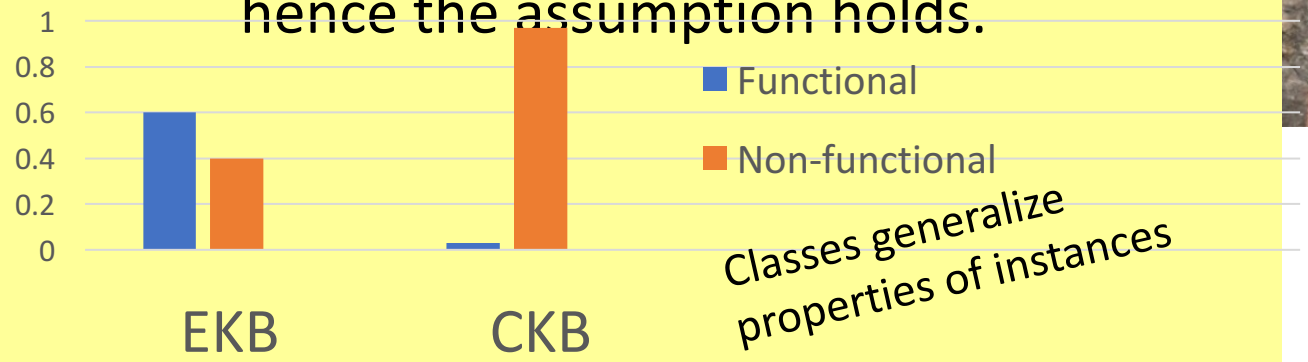
A. Honnold, bornIn, UK

CKB: ***assumption fails***

Commonsense Knowledge



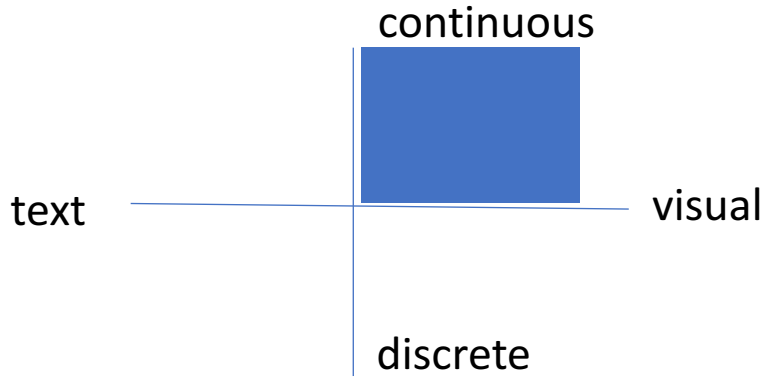
EKBs have several functional relations hence the assumption holds.



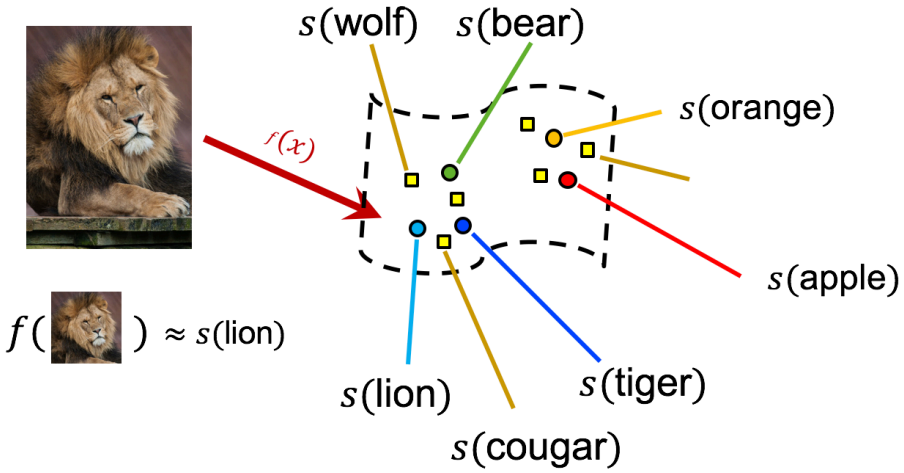
Classes generalize properties of instances

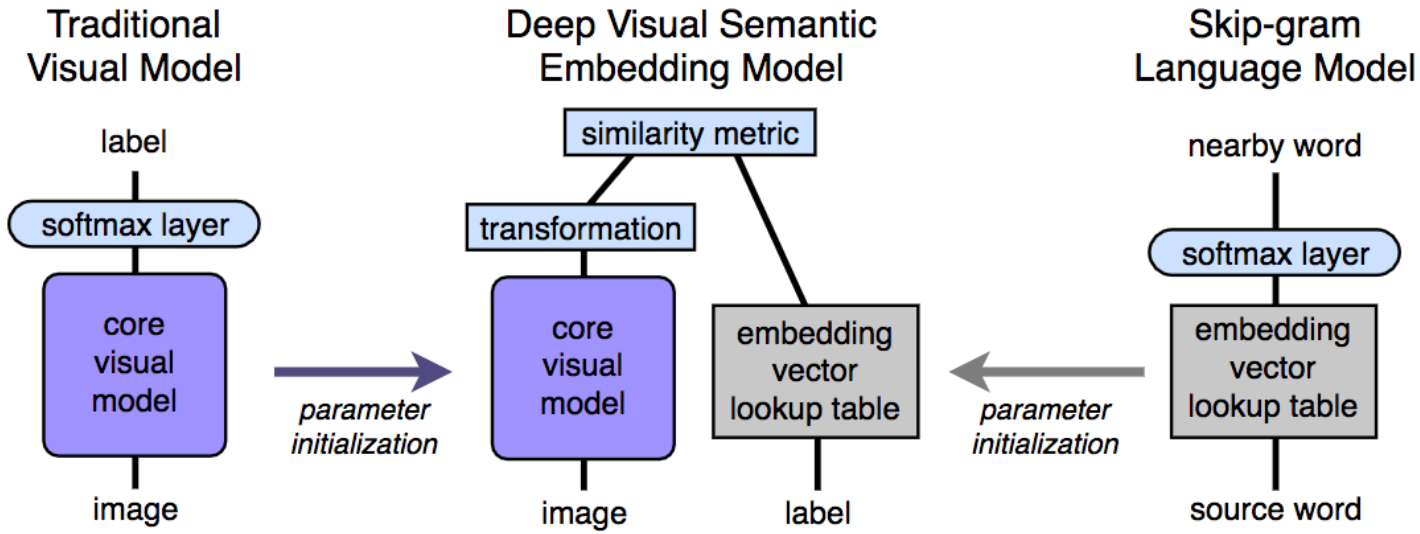
DeViSE [Frome et. al 2013] - continuous multimodal representation

Trains a joint embedding model of both images and labels (text), with non-linear mappings from image features to the embedding space.



The joint model is initialized with parameters pre-trained at the lower layers of traditional image, text models





Part 1: Acquiring Commonsense Knowledge

- introduction
  - introduction to csk
  - csk unimodal and multimodal kbs
- csk representation
  - discrete and continuous representations
  - multimodal continuous representations
- **acquisition methods**
  - different levels of supervision and modalities
  - from facts to rules
- csk evaluation
  - explicit evaluation techniques: sampling, turked
  - challenge sets and problems in text and vision

Sources of acquisition

- Authored
  - experts, crowd, games
- Text
  - Web, ngrams, books, tables, moviescripts, Wikipedias
- Images:
  - Flickr annotations, clip art, real images, videos

Elusive

remember the CSK challenge?

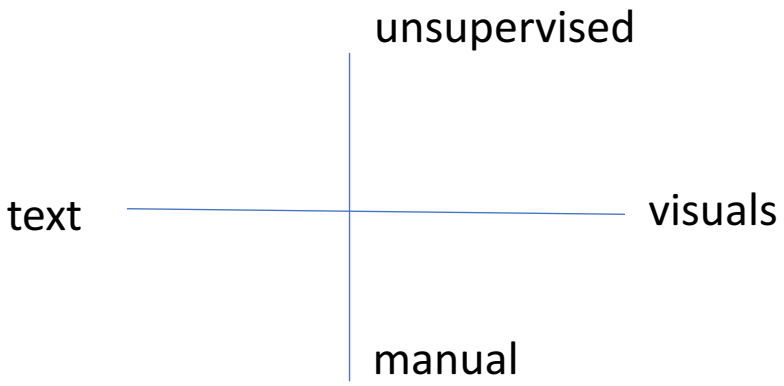
Less, implicitly expressed

Reporting Bias

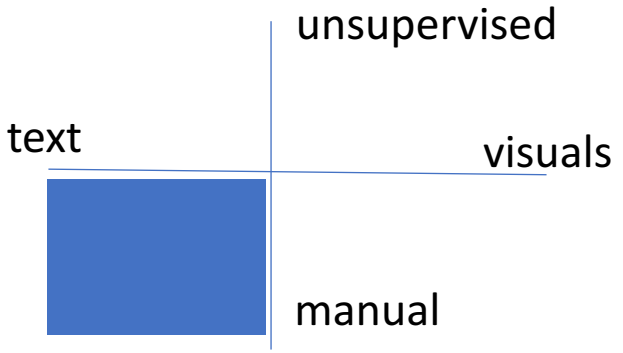
Multimodal

Acquisition methods

- different levels of supervision and modalities
- from facts to rules



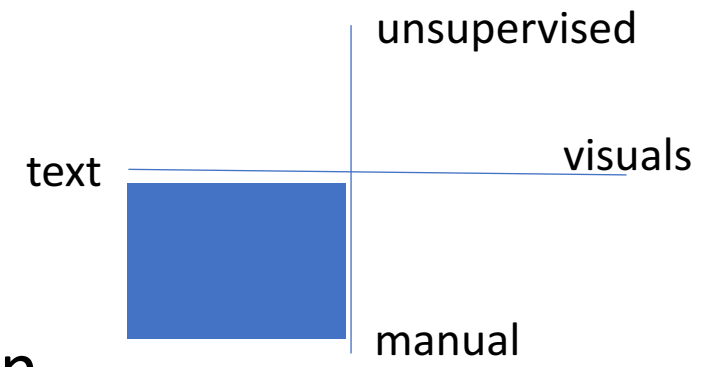
WordNet– hand authored



- WordNet is a lexical DB with taxonomies of nouns, verbs.
- Linguistics authored WN DAG, to a very fine-grained level.
- Sense orderings in WN can be uncommon with freq. statistics from an annotated corpora.
  - the first sense of “tiger” is … a fierce man.

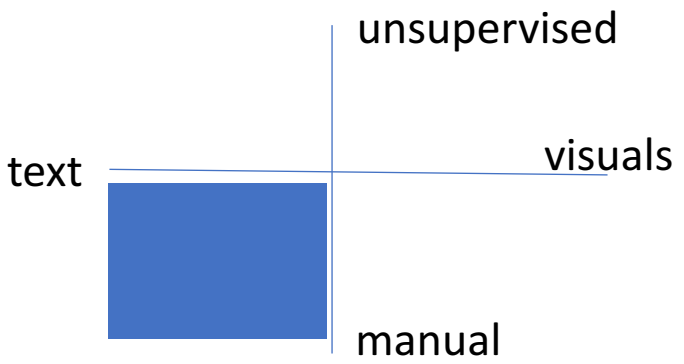
S: (n) **roller coaster**, big dipper, chute-the-chute
- *direct hyponym* / *full hyponym*
- ***direct hypernym*** / *inherited hypernym* / *sister term*
  - S: (n) elevated railway, elevated railroad, elevated, el, overhead railway
    - S: (n) railway, railroad, railroad line, railway line, railway system
      - S: (n) line
        - S: (n) carrier, common carrier
          - S: (n) business, concern, business concern, business organization, business organisation
            - S: (n) enterprise
              - S: (n) organization, organisation
                - S: (n) social group
                  - S: (n) group, grouping
                    - S: (n) abstraction, abstract entity
                      - S: (n) entity
  - S: (n) ride
    - S: (n) mechanical device
      - S: (n) mechanism
        - S: (n) device
          - S: (n) instrumentality, instrumentation
            - S: (n) artifact, artefact
              - S: (n) whole, unit
                - S: (n) object, physical object
                  - S: (n) physical entity
                    - S: (n) entity

Cyc – hand authored



- Cyc is being built for 30 years by several experts, still about a milllion facts
- Cyc is a collection of microtheories partitioned by domain
- CycL is a language in which these microtheories can be operated, to perform very powerful deductive reasoning.
- The reasons for the failure of Cyc has been that:
  - it is far from language: Input query must to be translated into CycL first.
  - it is very difficult to know which microtheory is applicable for a scenario.
  - deductive reasoning is powerful, but brittle: commonsense tends to be contextual

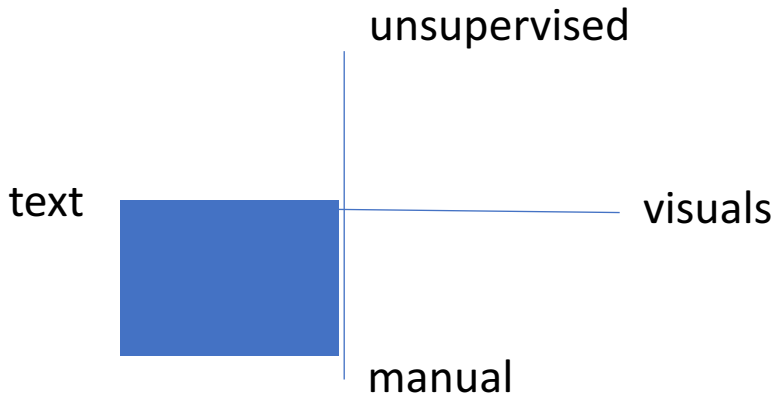
ConceptNet – crowdsourced



- Step 1: Collect OMCS corpus of CSK sentences
  - Templatized– closed set of relations.
  - ___ can be used to ___ expected to filled as e.g., pen, write.
- Step 2: Relation extraction becomes a regex match
  - Post-processed strings to minimize noise, scores = agreements.
  - Negations are a highlight of ConceptNet.
- Step 3: Repeated Step 1, 2 for multiple languages
  - These are independent efforts.
- Step 4: Vectorized the factorized concepts in concept x attributes matrix
  - Recently, ConceptNet is retrofitted to Glove embeddings.
- Note: ConceptNet5 onwards is a mix of non-commonsense and commonsense knowledge. E.g., >99% of the ~5 million part-whole relations are geo-locations.
  - Marina Bay is part of Asia – whereas, one would expect leaf is part of a plant.

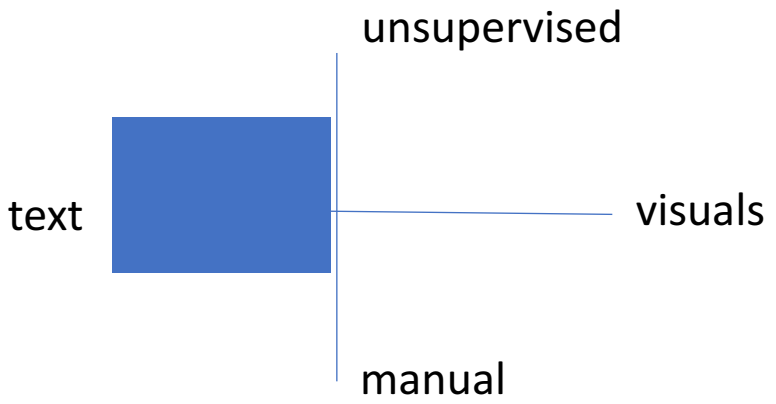
Verbosity – game [Ahn et. al 2006]

- Player 1 (describer) tells a secret word to player 2 (guesser) by filling in provided templates.
  - “It is a type of ___”,
  - “About the same size as ___”.
- Correct guesses lead to a fact.
- Note: Engagement in commonsense games is difficult.



Pattern based methods – distant supervision.

- For a fixed set of relations, works in two steps:
  - Step 1: Find patterns that cover seeds facts.
  - Step 2: Apply patterns to find facts.
- Two considerations:
  - Typed (e.g., POS tagged) patterns for better accuracy
  - Pattern scoring is very important when relations are related (as in CSK)
- Usually:
  - Semantic drift after initial round(s).
  - Curating after each round leads to higher accuracy.



Pattern based methods – distant supervision [Tandon et. al 2016].

Noisy patterns must be pruned

* ***are essential parts of****
* ***on sale along with*** *

PE are essential parts of PE
seat, cycle
dollars, ~~**life**~~
snake, **machine**

$$\sigma(a_k) = \frac{e^{supp(a_k)}}{1 + e^{supp(a_k)}} \frac{e^{str(a_k)}}{1 + e^{str(a_k)}}$$

Score of a candidate a_k (either pattern or assertion)

Seed Support: many distinct seed matches implies high score

Strength: Penalize semantic drift of candidates matching seeds of diff. relations

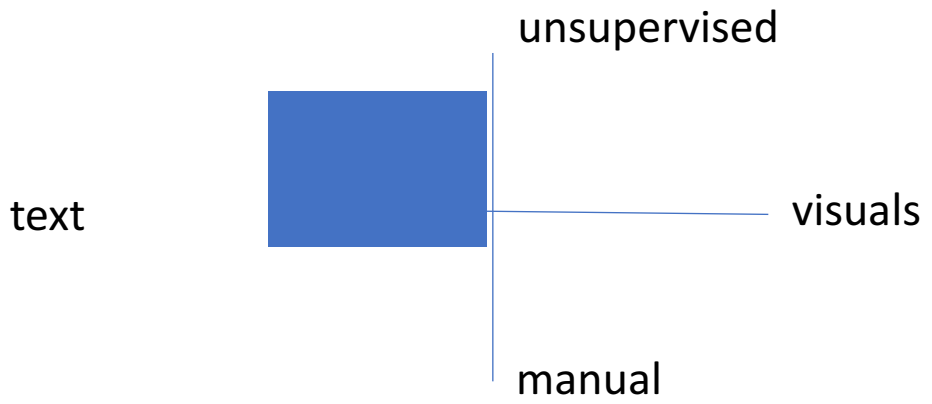
Type checking always helps

Physical entity,
abstract entity help tell apart easily confused relations

Domain (r)	Relation: r	Range (r)
Physical Entity	**Physical part of (P)** wheel, cycle	Physical Entity
Physical + Abstract Entity	**Member of (M)** cyclist, team	Abstract Entity
Substance Entity	**Substance of (S)** rubber, wheel	Physical Entity

Semi-supervised learning for knowledge triples

- When the amount of training data is limited/ expensive
- The space is naturally clusterable
- Well studied machine learning methods to leverage.



Semi-supervised learning in WebChild [Tandon et. al 2014]

Task: Input = web corpus, bootstrapping patterns to extract adjective noun pairs

e.g., <summit, crisp>

Output: *triples* $< w1_n^s , \quad r, \quad w2_a^s >$

$summit_n^2$ $hasTemperature$ $crisp_a^3$

disambiguated n

1.) 

2.) 

3.) 

...

fine-grained relations:
r∈R

hasAppearance
hasSound
hasTaste
hasTemperature
hasSound
evokesEmotion

...

disambiguated a

1.) 

2.) 

3.) 

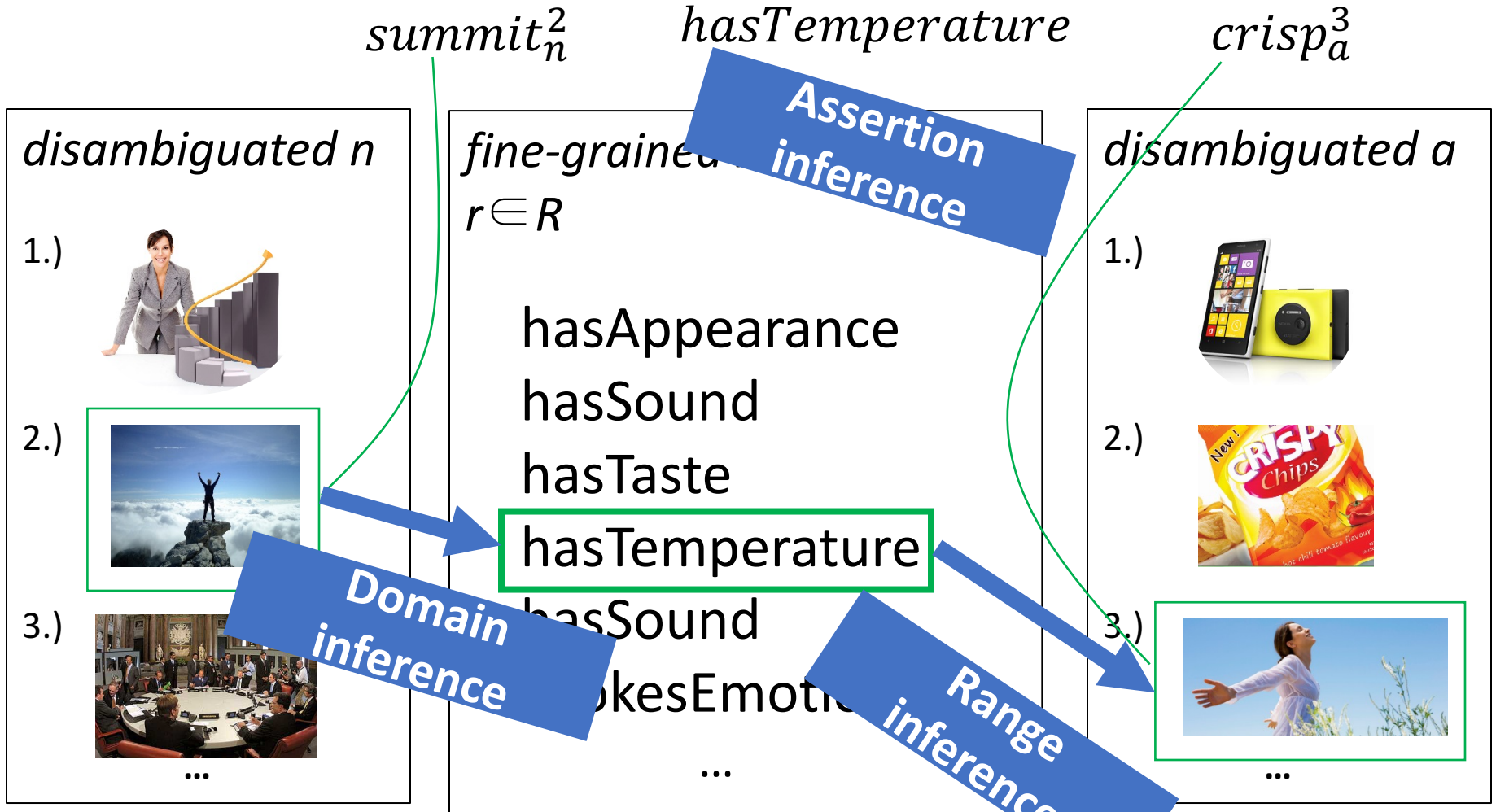
...

Semi-supervised learning in WebChild [Tandon et. al 2014]

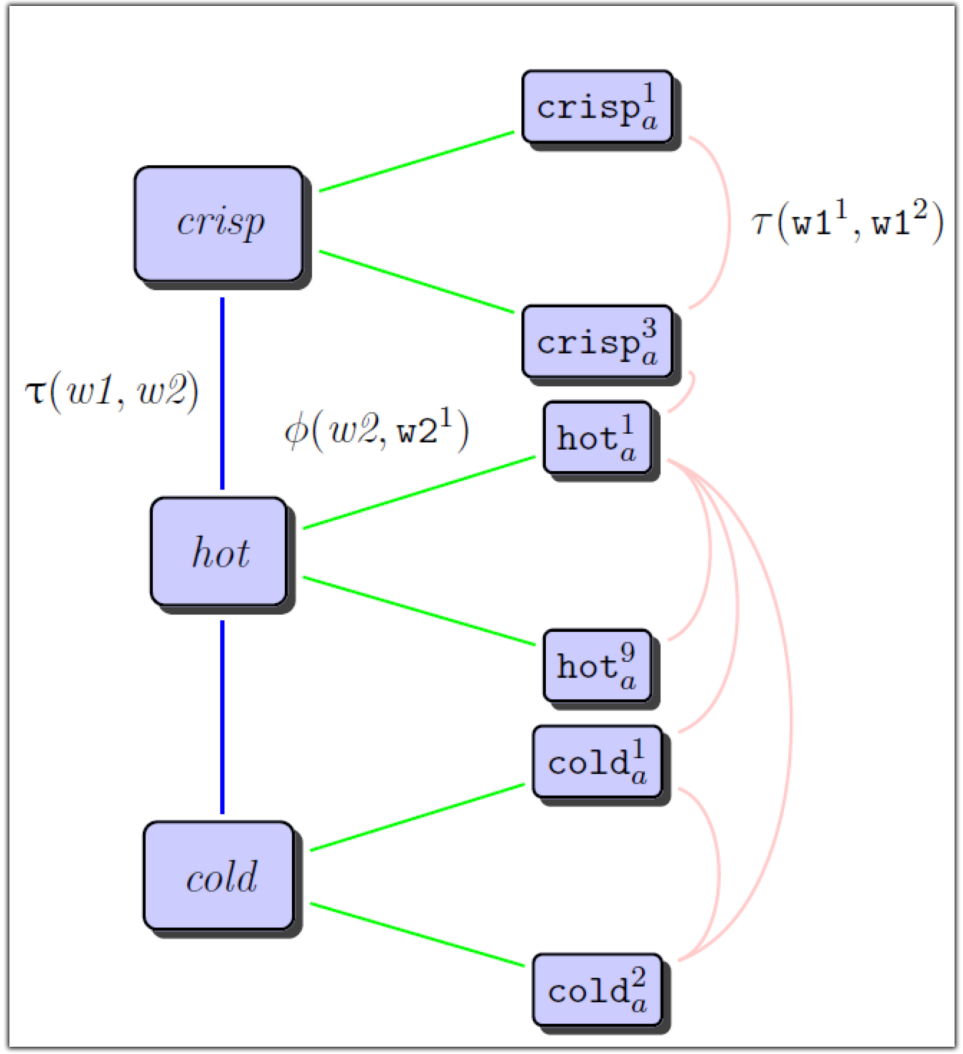
Task: Input = web corpus, bootstrapping patterns to extract adjective noun pairs

e.g., <summit, crisp>

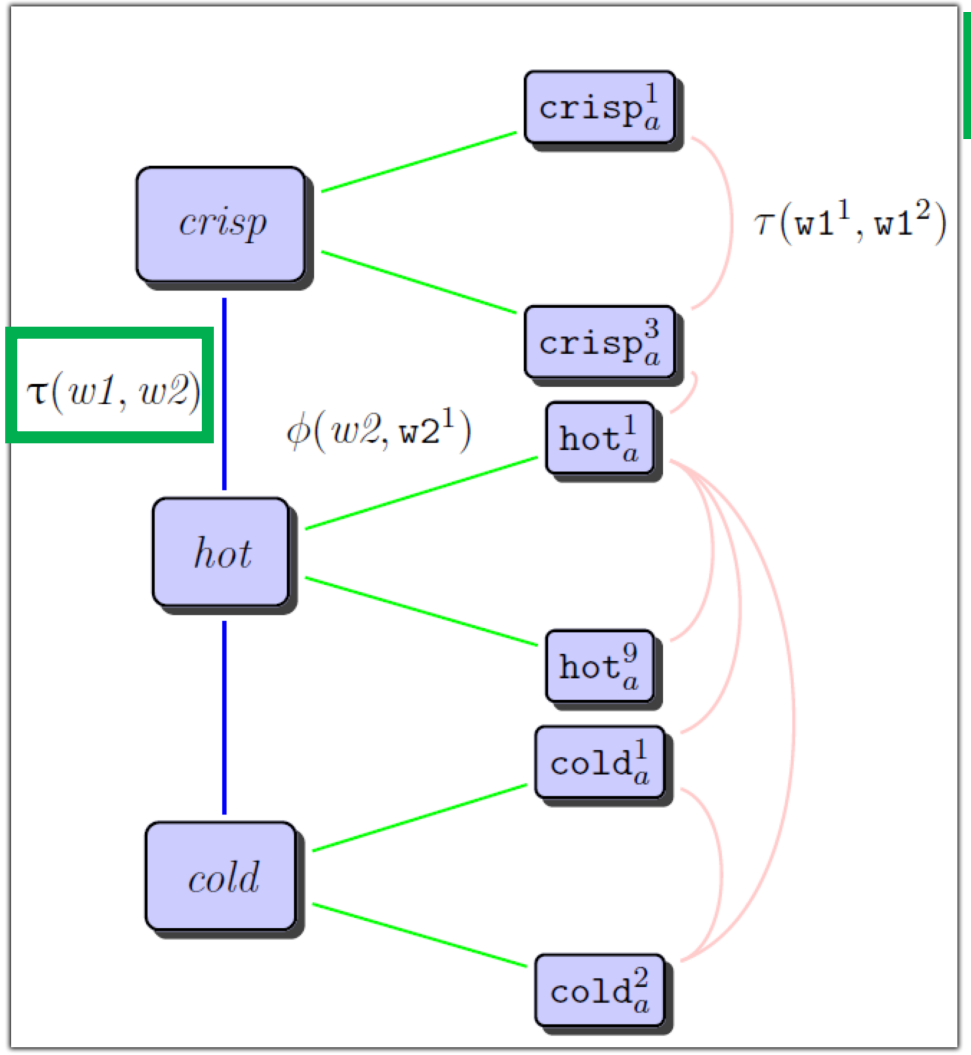
Output: *triples* $< w1_n^s , \quad r, \quad w2_a^s >$



An instance of the problem: $range(r = hasTemperature)$



An instance of the problem: $range(r)$



$$\tau_{\mathrm{AA}}[a1\ ,\ a2] = \alpha\ O^T O + (1-\alpha) P^T P$$

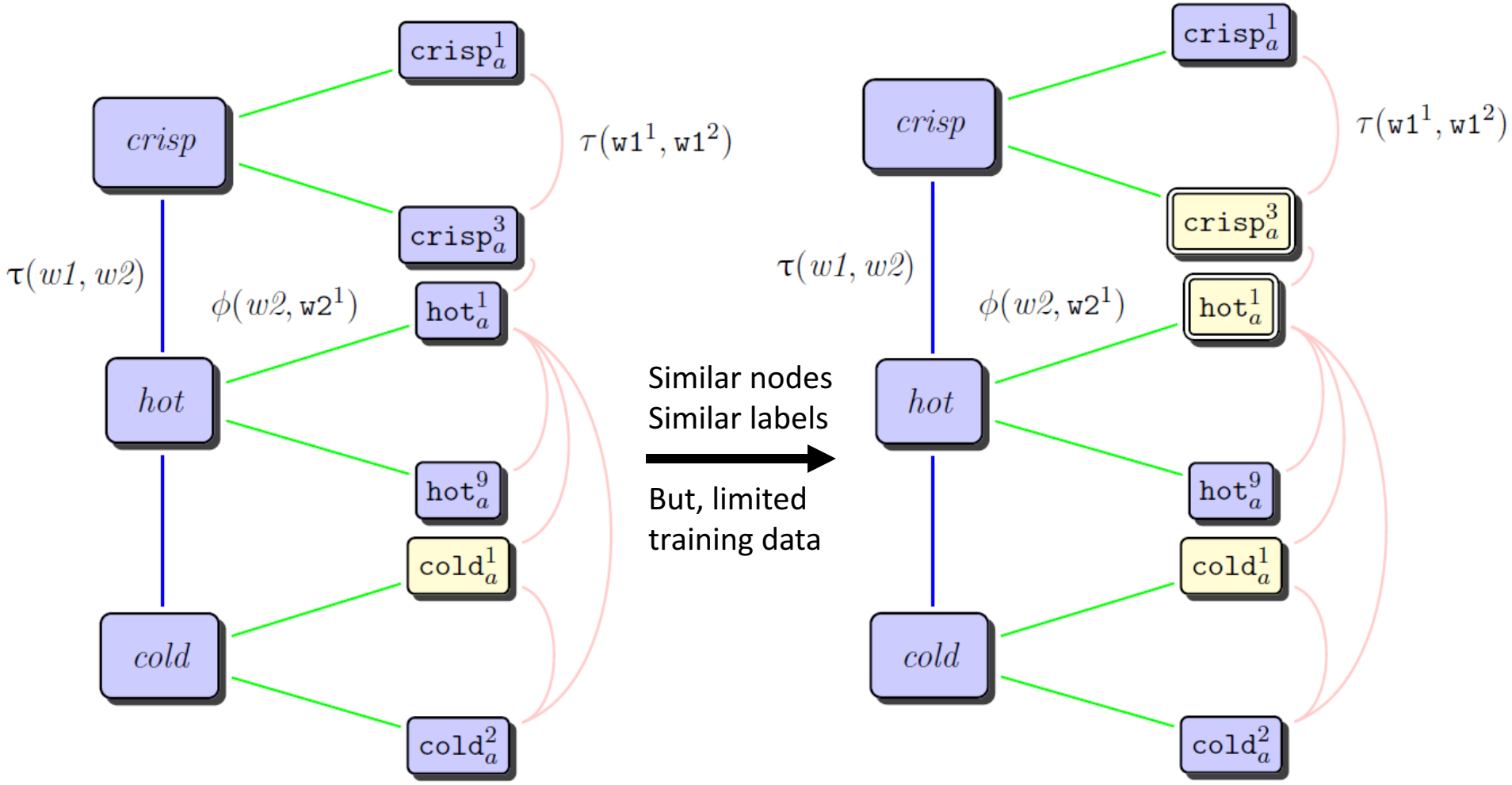
$$\tau_{\mathrm{AA}}[\mathtt{a}^i\ ,\ \mathtt{a}^j] = \beta\ \mathrm{hirst}[\mathtt{a}^i, \mathtt{a}^j] + (1-\beta) G^T G$$

$$\phi[a\ ,\ \mathtt{a}^i\ ] = \gamma \frac{1}{1+\mathrm{i}} + (1-\gamma) O^T G$$

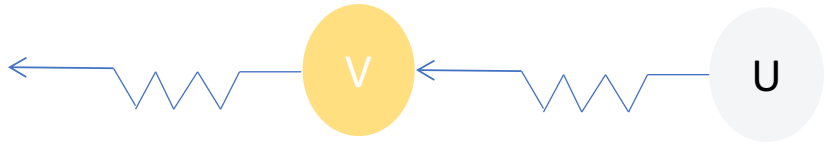
	summit	mountain	dancer
cold	20	50	3
hot	30	40	10
crisp	15	15	1

$$O_{i,j}: \quad freq(noun_i, adj_j)$$
$$P_{i,j}: \quad \#\mathrm{patt}(noun_i, adj_j)$$
$$G_{i,j}: \quad freq(adj_i, glossword_j)$$

Label propagation for graph inference, given few seeds.
- Label per node = in/not in range of hasTemperature



Label Propagation: Loss function [Talukdar et. al 2009]



Seed label loss

+

Similar node diff label loss

+

Label prior loss (high degree nodes are noise)

$$\left[\mu_1 \sum_v p_v^{inj} \sum_l (Y_{vl} - \hat{Y}_{vl})^2 + \mu_2 \sum_{v,u} p_v^{cont} W_{vu} \sum_l (Y_{vl} - \hat{Y}_{ul})^2 + \mu_3 \sum_{vl} (\hat{Y}_{vl} - R_{vl})^2\right]$$

WebChild: resulting KB

Relation	Range	Domain	Assertions
hasTaste	$\texttt{sweet}^1_a$	$\texttt{strawberry}^1_n$	$(\texttt{chocolate}^1_n, \texttt{creamy}^2_a)$
	$\texttt{hot}^9_a$	$\texttt{chili}^1_n$	$(\texttt{pizza}^1_n, \texttt{delectable}^1_a)$
	$\texttt{sour}^2_a$	$\texttt{salsa}^1_n$	$(\texttt{salsa}^1_n, \texttt{spicy}^2_a)$
	$\texttt{salty}^3_a$	$\texttt{sushi}^1_n$	$(\texttt{burger}^1_n, \texttt{tasty}^1_a)$
	$\texttt{lemony}^1_a$	$\texttt{java}^2_n$	$(\texttt{biscuit}^2_n, \texttt{sweet}^1_a)$
hasShape	$\texttt{triangular}^1_a$	$\texttt{leaf}^1_n$	$(\texttt{palace}^1_n, \texttt{domed}^1_a)$
	$\texttt{meandering}^1_a$	$\texttt{circle}^1_n$	$(\texttt{table}^2_n, \texttt{flat}^1_a)$
	$\texttt{crescent}^1_a$	$\texttt{ring}^8_n$	$(\texttt{jeans}^2_n, \texttt{tapered}^1_a)$
	$\texttt{obtuse}^2_a$	$\texttt{egg}^1_n$	$(\texttt{tv}^2_n, \texttt{flat}^1_a)$
	$\texttt{tapered}^1_a$	$\texttt{face}^1_n$	$(\texttt{lens}^1_n, \texttt{spherical}^2_a)$

Semi-supervised learning for knowledge frame [Tandon et. al 2015]

Go up an elevation
.. ..

Parent activity

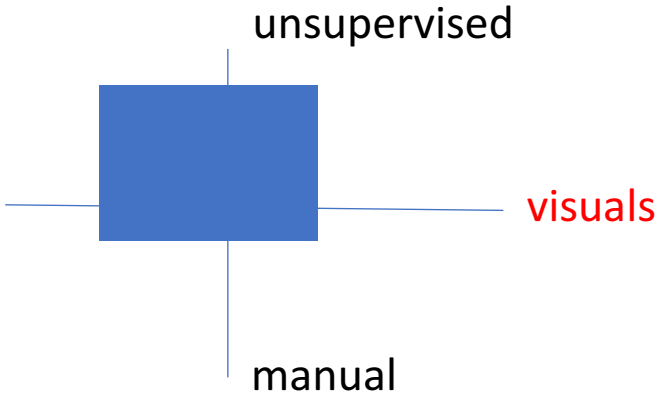
Previous activity

Next activity

{Climb up a mountain , Hike up a hill}	
Participants	climber, boy, rope
Location	camp, forest, sea shore
Time	daylight, holiday
Visuals	

Get to village
.. ..

Drink water
.. ..



Judicious choice of multimodal dataset

Blogger

May contain events or activities but varying granularity and no visuals. No clear scene boundaries.

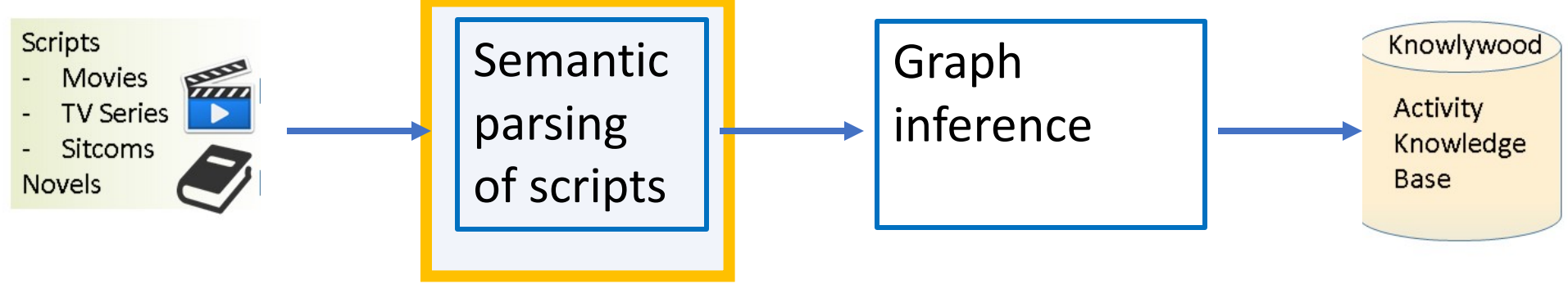
Hollywood narratives are easily available and meet the desiderata

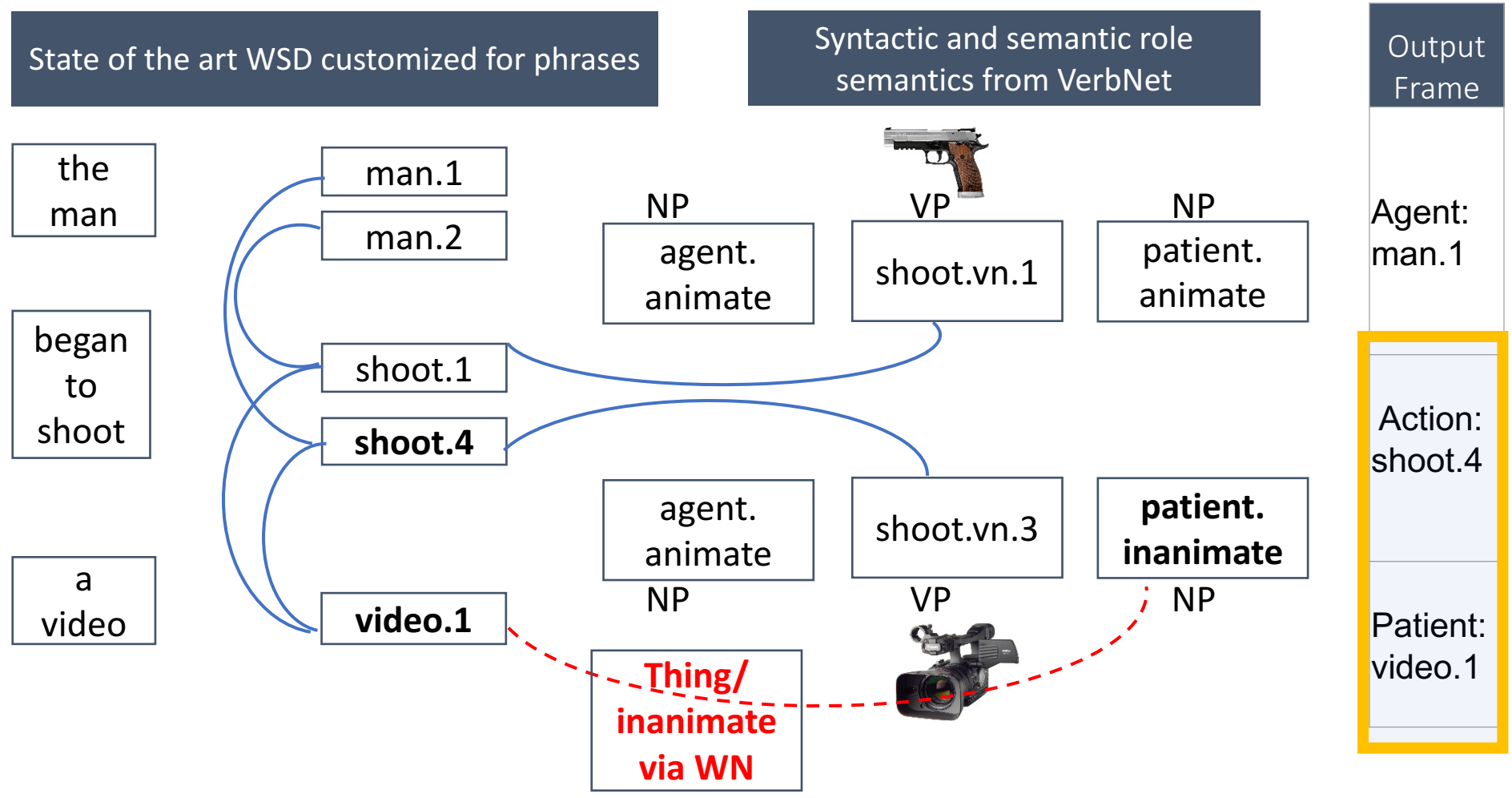
EXT. SMALL MOUNTAIN--DAY

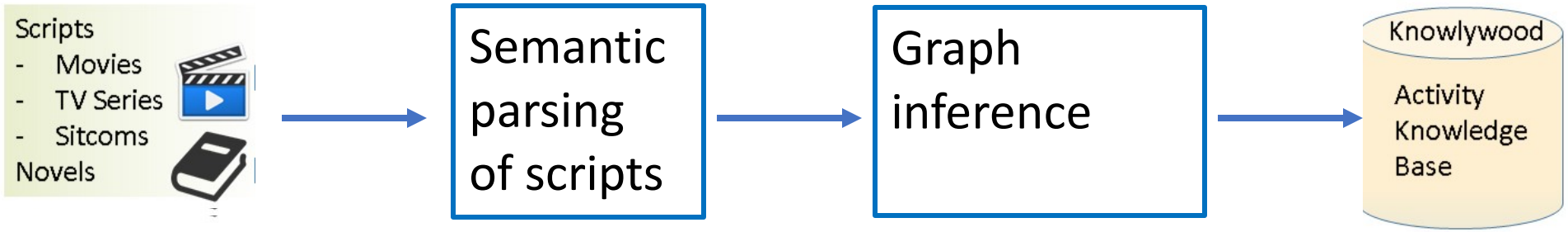
Wichita charges up the rockage of a small mountain-hill-type thing. The image repeats itself over and over--each time Wichita is more sweaty, gasping, sneering.

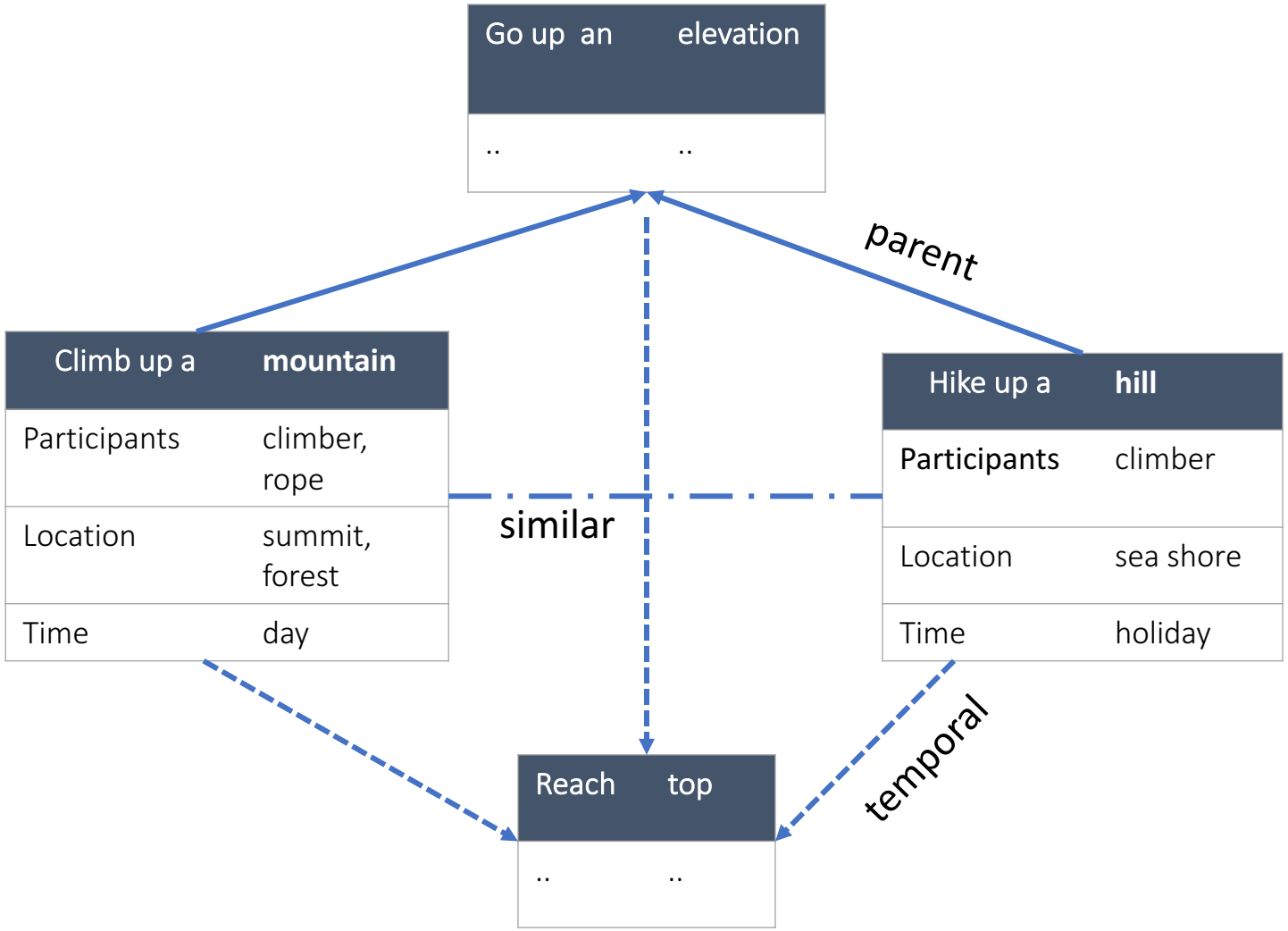
Wichita (V.O.)
The rules forbid anyone from the climbing the camp's mountain.

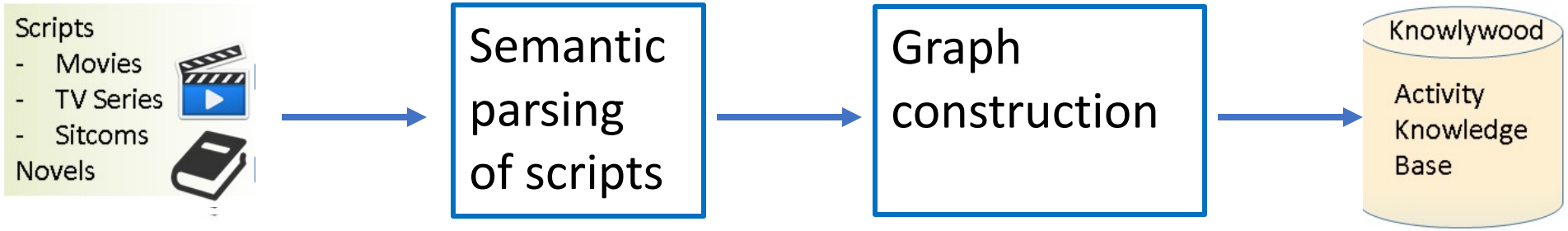
align via subtitles with approximate dialogue similarity

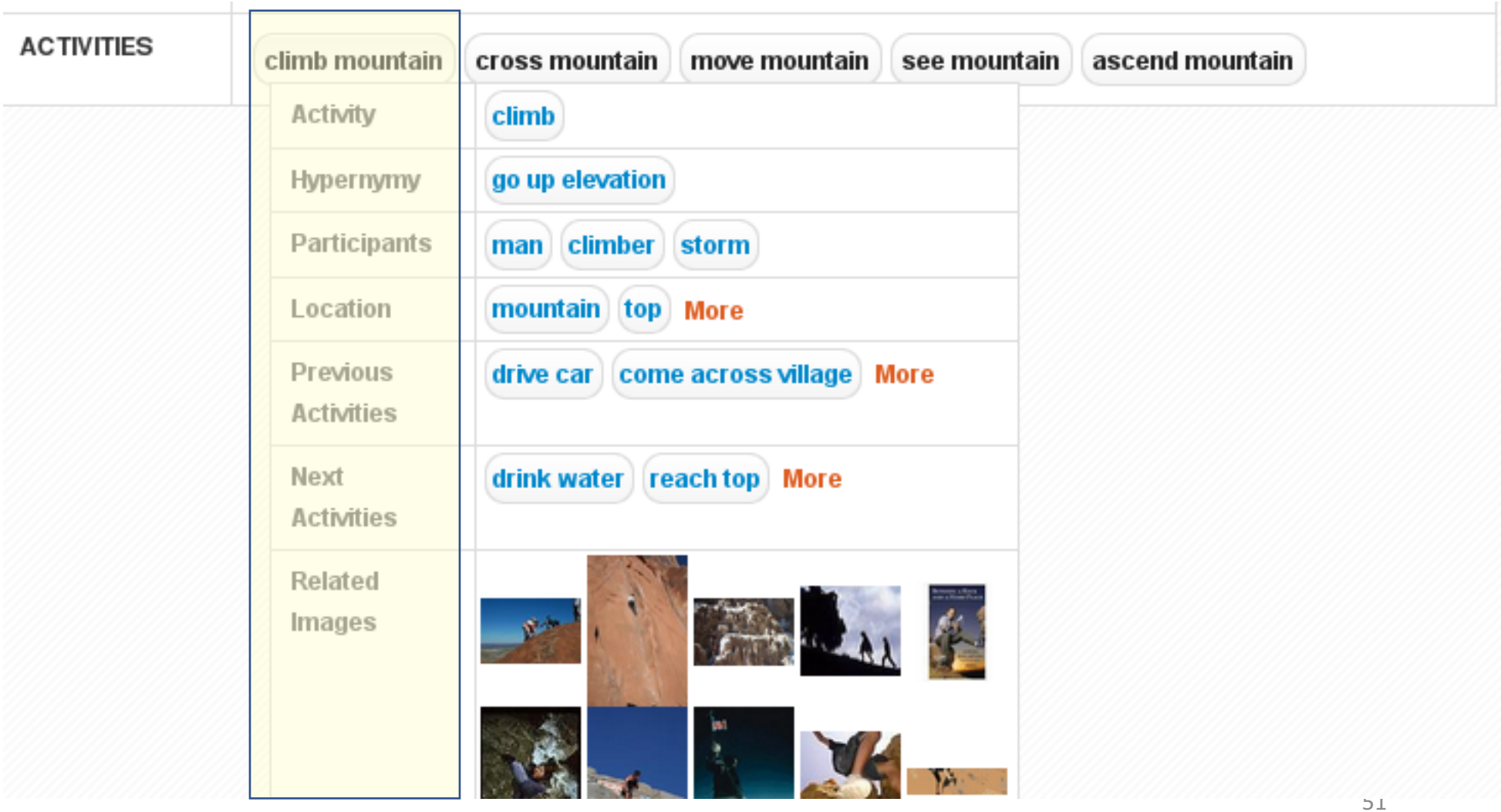

Scripts
- Movies
- TV Series
- Sitcoms
Novels
Semantic parsing of scripts
Graph inference
Knowlywood
Activity Knowledge Base


State of the art WSD customized for phrases
Syntactic and semantic role semantics from VerbNet
Output Frame
the man
began to shoot
a video
man.1
man.2
shoot.1
shoot.4
video.1
NP
agent. animate
VP
shoot.vn.1
NP
patient. animate
agent. animate
NP
shoot.vn.3
VP
patient. inanimate
NP
Thing/ inanimate via WN
Agent: man.1
Action: shoot.4
Patient: video.1


Scripts
- Movies
- TV Series
- Sitcoms
Novels
Semantic parsing of scripts
Graph inference
Knowlywood
Activity Knowledge Base

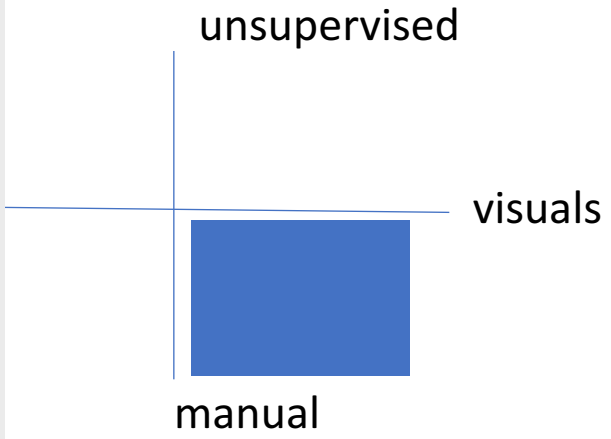

Go up an elevation
.. ..
parent
Climb up a mountain
Participants climber, rope
Location summit, forest
Time day
similar
Hike up a hill
Participants climber
Location sea shore
Time holiday
temporal
Reach top
.. ..


Scripts
- Movies
- TV Series
- Sitcoms
Novels
Semantic parsing of scripts
Graph construction
Knowlywood
Activity Knowledge Base


ACTIVITIES
climb mountain
cross mountain
move mountain
see mountain
ascend mountain
Activity
climb
Hypernymy
go up elevation
Participants
man
climber
storm
Location
mountain
top
More
Previous Activities
drive car
come across village
More
Next Activities
drink water
reach top
More
Related Images

Visual Genome – crowdsourced visual KB

Regions	Attributes	Relationships
A wall on the side of a building	leaf is green and yellow	light reflecting on leaves
green leaf with yellow spots	stems is green	
yellowish green stems	light is bright	
bright light reflecting on leaves	leaf is dark	
a dark green leaf with brown edges	leaf is green	
a large green leaf	leaf is brown edge	
a large dark green leaf	leaf is large	
	leaf is dark green	
	leaf is yellow, brown, and green	
	pale flower is pale	



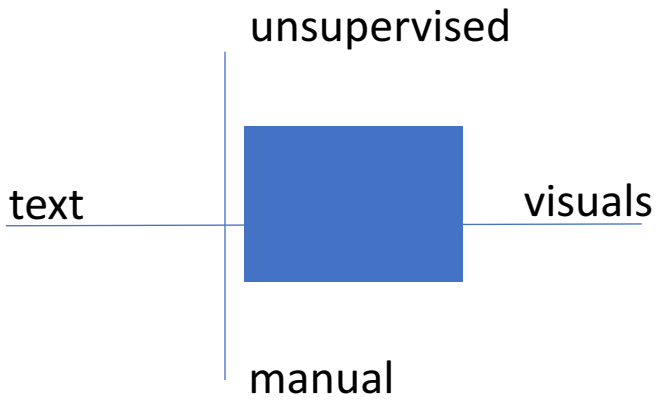
- Crowdsourced on mechanical turk.
- Contains bounding boxes annotated with relationships.

Learn commonsense through visual abstraction [Vedantam et. al 2015]

new assertion:
squirrels *look at* nuts

similar to?

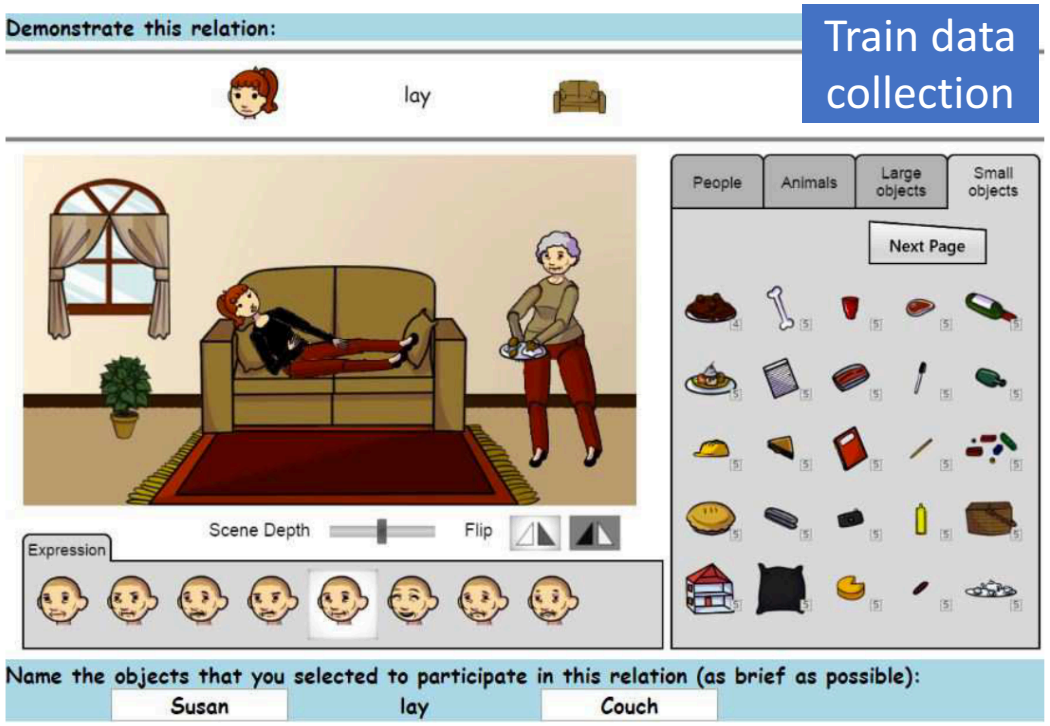
known plausible assertion:
squirrels *want* nuts



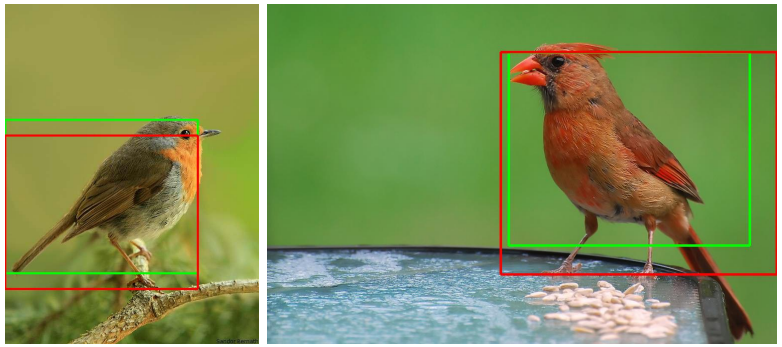
- "look at" and "want" not semantically sim.
- Input = {squirrel (s), looks at (p), nuts(o)}
- Output = plausible/ not

ß	text plausibility w.r.t. train	+	1-ß	visual plausibility w.r.t. train

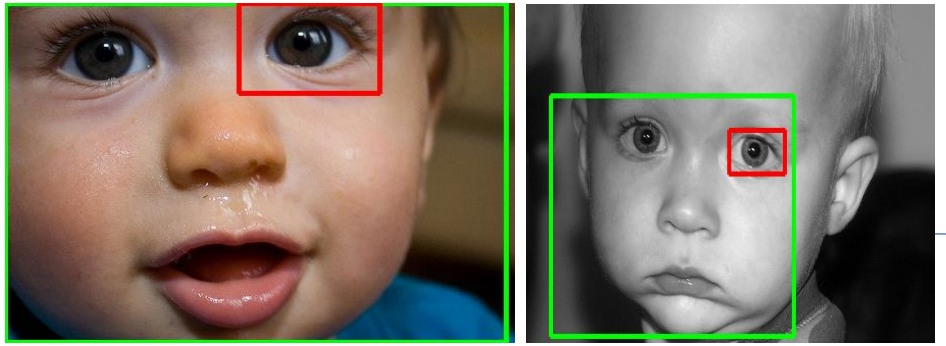
ß	vector space sim. $V(s_1, s_1') + V(p_1, p_1') + V(o_1, o_1')$	+	1-ß	Object features + Scene features + Interaction features



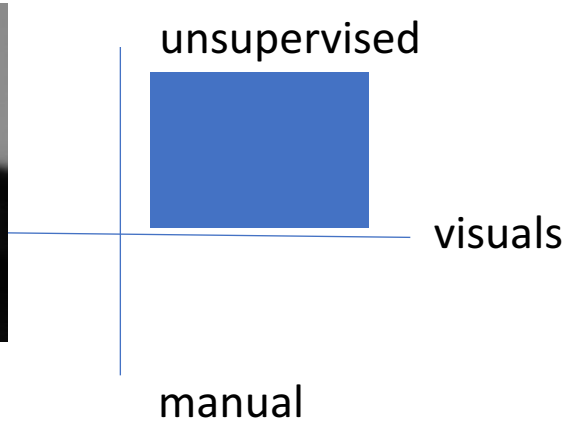
Learning commonsense directly through images [Shrivastava et. al 2014]



Sparrow is a kind of/looks similar to **bird**



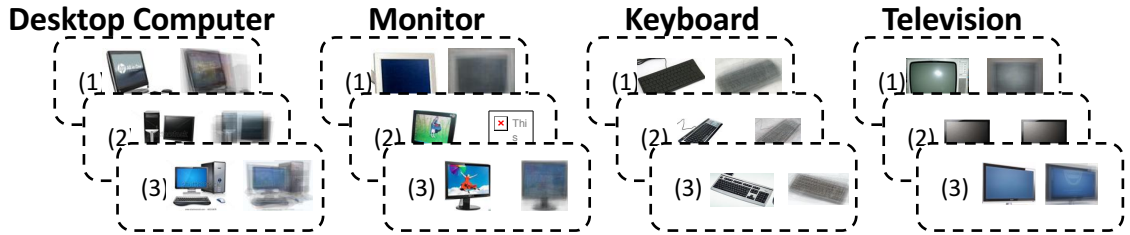
Eye is a part of **Baby**



Object - Scene

Helicopter is found in **Airfield**

Ferris wheel is found in **Amusement park**



(0) Seed Images

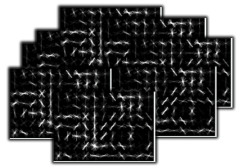
Desktop Computer

Monitor

Keyboard

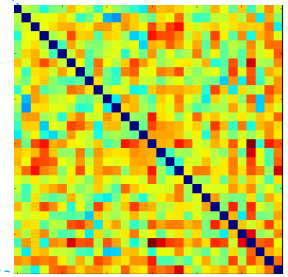
(1) Subcategory Discovery

(2) Train Models



Desktop Computer (1)
Desktop Computer (2)
Desktop Computer (3)
…
Monitor (1)
…

(3) Relationship Discovery



Learned relationships:

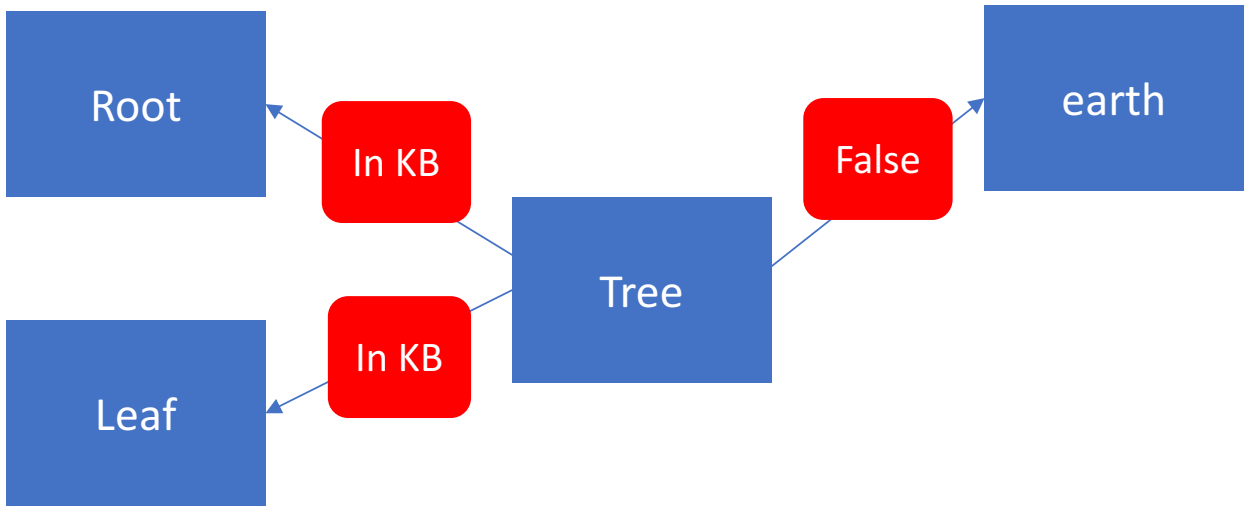
- **Keyboard** is a part of **Desktop Computer**
- **Monitor** is a part of **Desktop Computer**
- **Television** looks similar to **Monitor**

(4) Add New Instances

Desktop Computer Monitor Television

From facts to rules: AMIE

- AMIE is a system that mines Horn rules from a KB.
  - hasChild(z, y) $\wedge$ married(x, z) $\Rightarrow$ hasChild(x, y)
- Rules can be used to make predictions and in reasoning.
- Performs exhaustive top-down search based on partial closed world assumption, minimum support threshold.



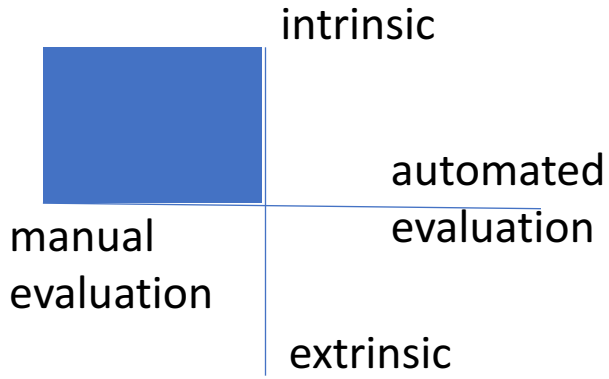
- Scoring of candidate rules based on PCA confidence: the ratio of predicted positive examples, out of all predicted examples.

Part 1: Acquiring Commonsense Knowledge

- introduction
  - introduction to csk
  - csk unimodal and multimodal kbs
- csk representation
  - discrete and continuous representations
  - multimodal continuous representations
- acquisition methods
  - different levels of supervision and modalities
  - from facts to rules
- **csk evaluation**
  - explicit evaluation techniques: sampling, turked
  - challenge sets and problems in text and vision

Evaluation of acquired knowledge

- Commonsense relations are typically non-functional (e.g., location of a car – roads, parking lot, highway)

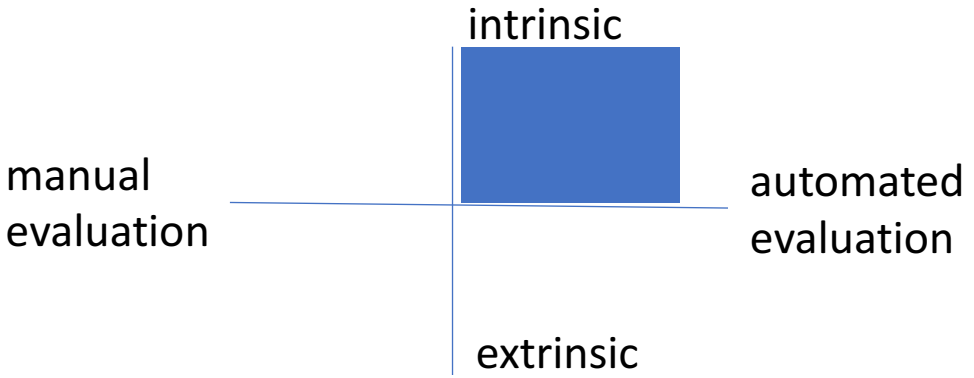


- Most approaches (e.g., Tandon 2014) are based on a manual assessment including from turkers, where the task is to identify if a relation “can hold”.

- The non-functional slice of encyclopedic knowledge (e.g., childrenOf) does not require “prominent” values. Commonsense acquisition must capture the most salient answers, however. Newer intrinsic evaluations have started asking “What is the 2-3 most prominent values for this concept, under a certain relation).

- Estimating recall remains an open problem because we do not know the possible commonsense knowledge. Dalvi et. al 2017 propose completeness w.r.t. a corpus, of prominent facts from a book.

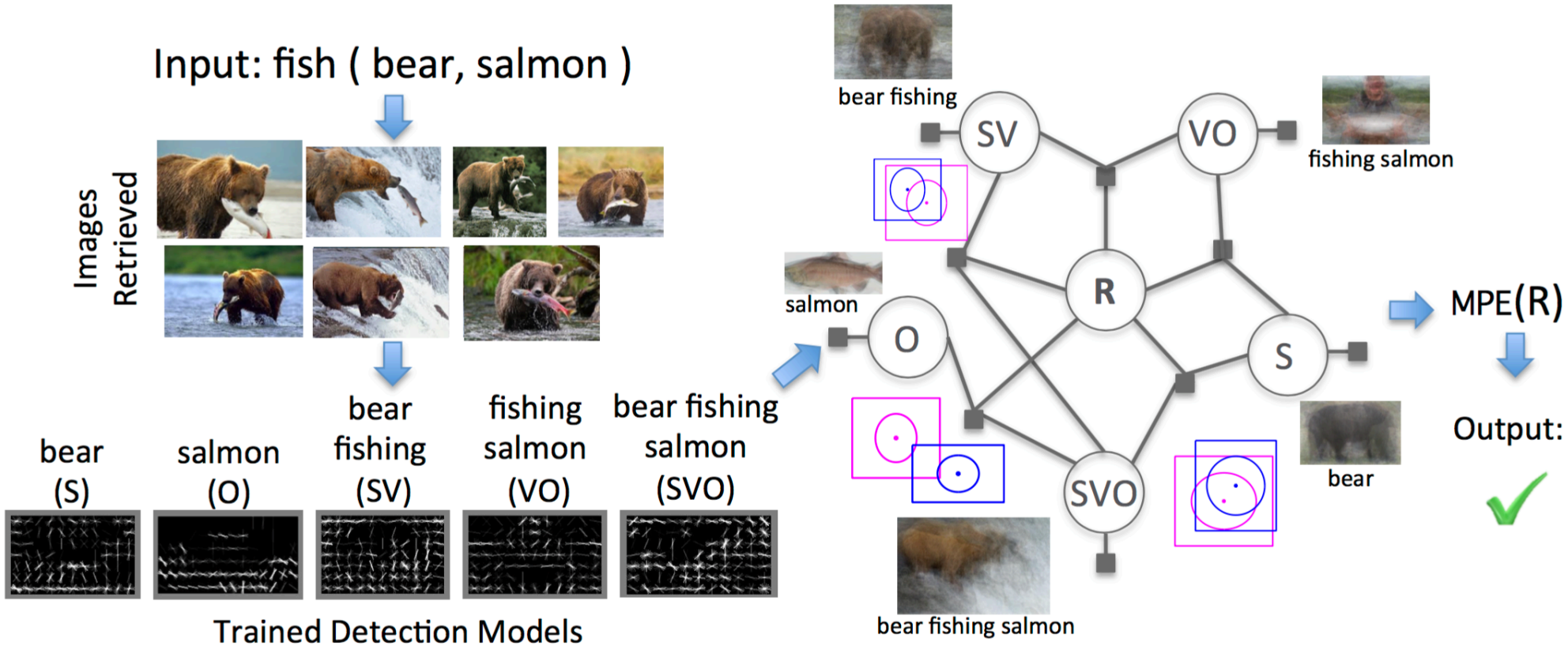
Evaluation of acquired knowledge



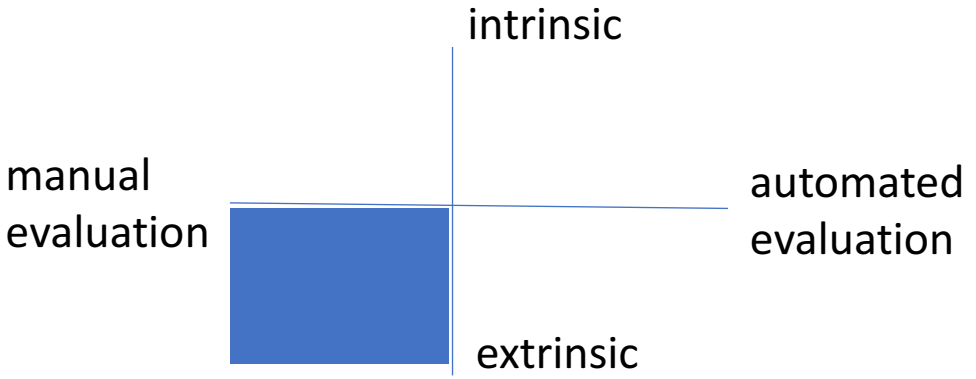
- Verification of commonsense knowledge from text, by detecting inconsistencies based on manually specified rules such as transitivity:
  - e.g., spoke part of wheel ^ wheel part of cycle → spoke part of cycle
- Due to reporting bias, visual verification is becoming popular. Visual verification is not an option for Encyclopedic knowledge.
- Tandon et. al 2016 described how part-whole relationships (such as, seat is part of cycle) can be verified using images or more scalably using image tags.
  - **cycle**, fun, trip, go, **seat**, niket
    seat is visible part of cycle

Visual verification using low level image features

[Sadeghi et. al 2015]



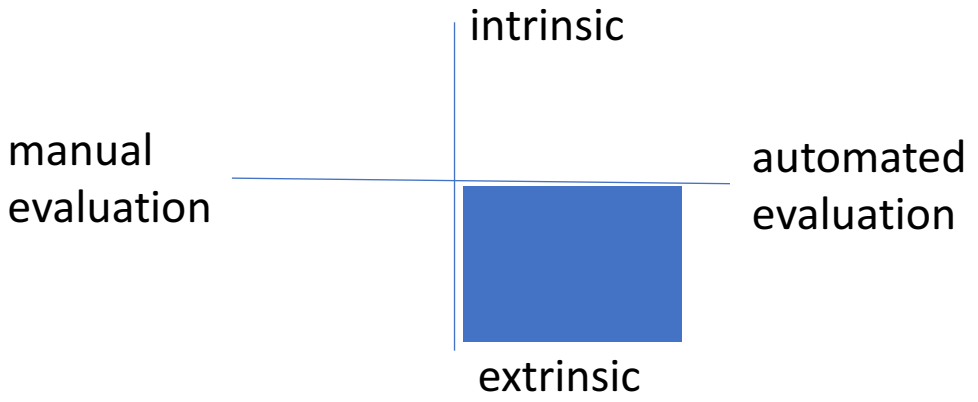
Evaluation of acquired knowledge



Commonsense for better concept-concept similarity, but this evaluation is limited by scale and subjectivity (Tandon et al. 2014)

Top 10 adjectives	universal, magnetic, small, ornamental, decorative, solid, heavy, white, light, cosmetic
Top 5 expansions	wall mount, mounting bracket, wooden frame, carry case, pouch

Evaluation of acquired knowledge



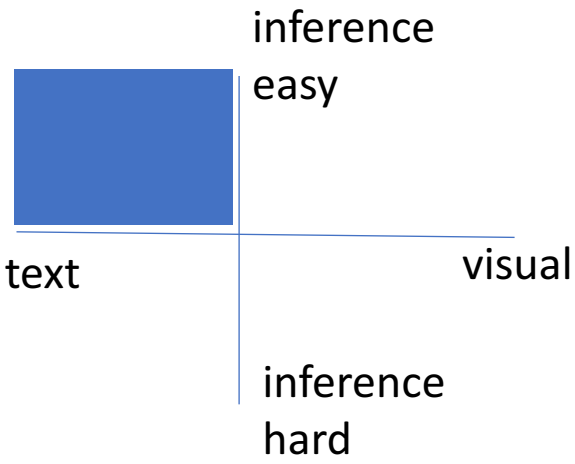
A number of disparate large-scale annotated challenge sets for commonsense exist.

However, unlike some standard tasks like reading comprehension or object detection, these have not been the heralds of commonsense knowledge as they require a reasoner which makes the evaluation subjective.

More end to end or task-oriented evaluation of commonsense knowledge is needed.

Challenge sets and problems for commonsense

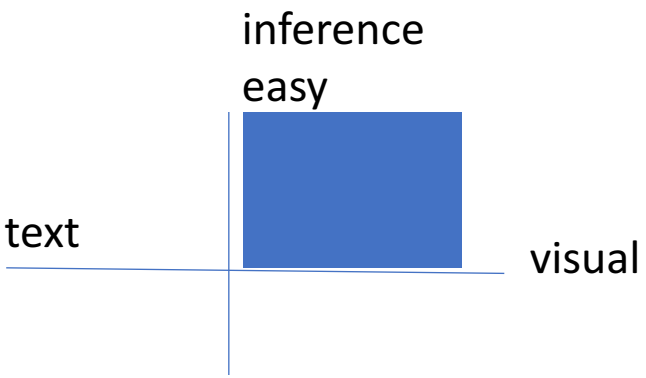
The StoryCloze dataset [Mostafazadeh et. al 2016] requires predicting the conclusion of a story. However, elimination of answer choices is an effective approach for this dataset, rendering this task inference easy.

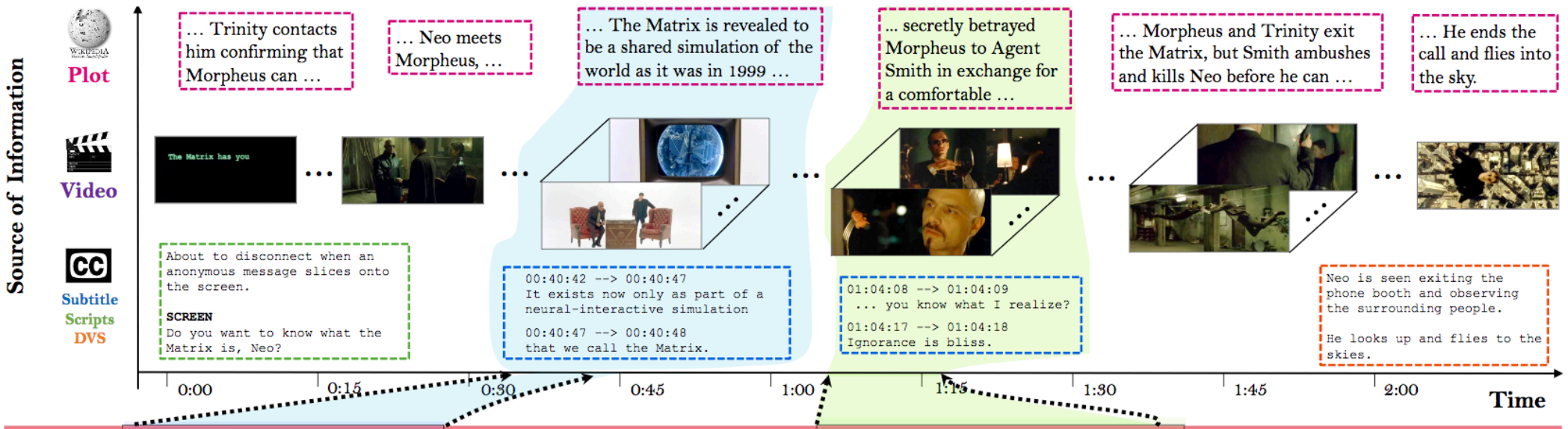


Context	Right Ending	Wrong Ending
Sammy's coffee grinder was broken. He needed something to crush up his coffee beans. He put his coffee beans in a plastic bag. He tried crushing them with a hammer.	It worked for Sammy.	Sammy was not that much into coffee.
Gina misplaced her phone at her grandparents. It wasnt anywhere in the living room. She realized she was in the car before. She grabbed her dads keys and ran outside.	She found her phone in the car.	She didnt want her phone anymore.

Challenge sets and problems for commonsense

MovieQA [Tapaswi et. al 2016] poses QA over movie video clips, plots, subtitles, scripts, and DVS [Rohrbach et. al 2015]





	What is the Matrix?	Who kills Neo in the Matrix?	Why does Cypher betray Morpheus?	How does the movie end?
	A: A shared simulation of the world	A: Smith kills Neo	A: In exchange for a comfortable life	A: With Neo flying into the sky
Quiz	A: A group of robots A: A human body A: A set of numbers stored as a table	A: Trinity kills Neo A: Morpheus kills Neo after he realizes that Neo is not the one	A: In exchange for money A: Because he is threatened by Agent Smith	A: With the Machines chasing after Neo A: We see Mr. Smith torture Morpheus

Challenge sets and problems for commonsense

Stories (Winograd [Liu et. al 2016], SemEval 2018 Task 11[Ostermann 2017])

Mariano fell with a crash and lay stunned on the ground. Castello instantly kneeled by his side and raised his head.

His head: Mariano/ Castello?

Consider the following reading text from the *planting a tree* scenario...

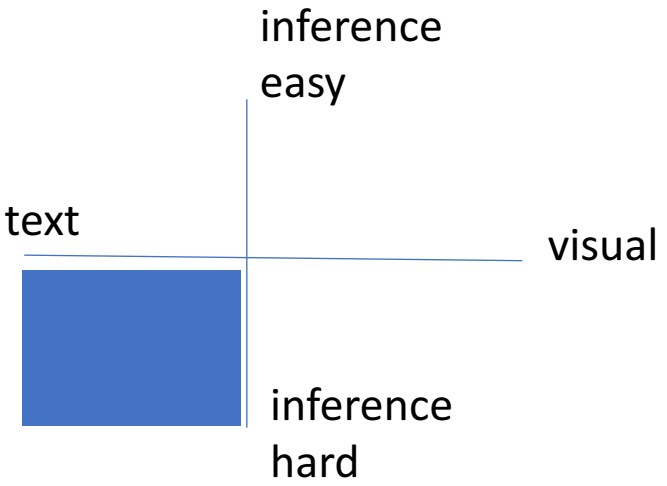
My backyard was looking a little empty, so I decided I would plant something. I went out and bought tree seeds. I found a spot in my yard that looked like it would get enough sunshine. There, I dug a hole for the seeds. Once that was done, I took my watering can and watered the seeds.

... and the following questions on the text.

A. Why was the tree planted in that spot?
   1. to get enough sunshine
   2. there was no other space

B. What was used to dig the hole?
   1. a shovel
   2. their bare hands

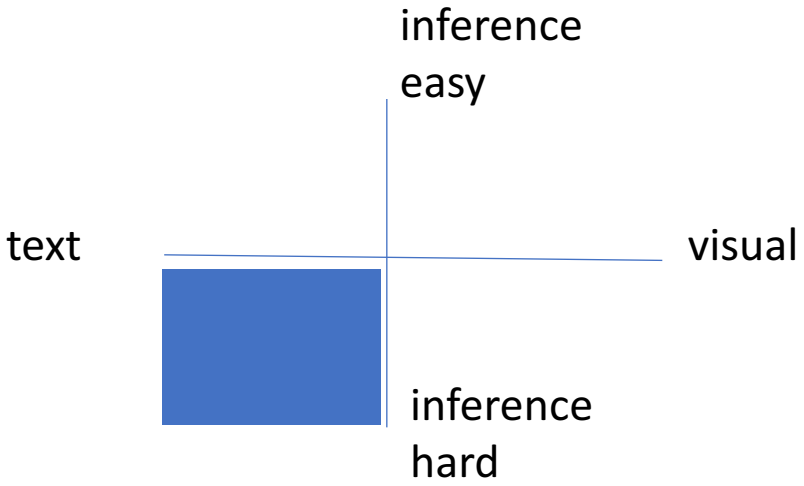
C. Who took the watering can?
   1. the grandmother
   2. the gardener



Challenge sets and problems for commonsense

QA/Comprehension (TQA, Aristo [Clark 2014])

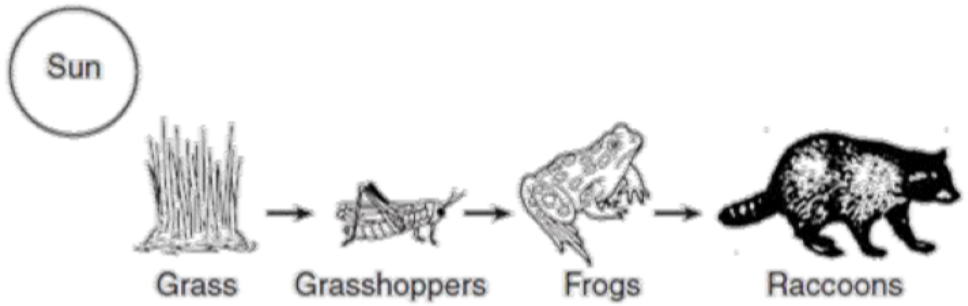
For roller-skate race, what is the best surface?
(A) sand (B) grass (C) **blacktop**



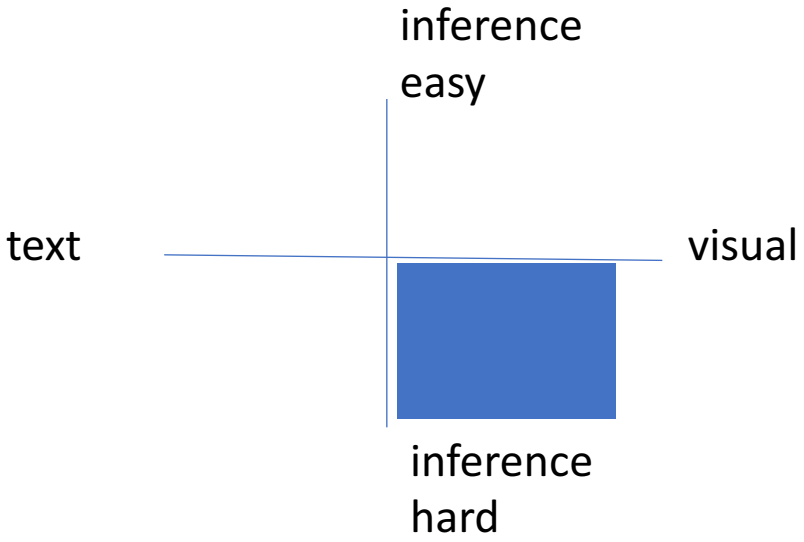
Challenge sets and problems for commonsense

QA (TQA [Kembhavi et. al 2016], Aristo [Clark 2014], FVQA)

Aristo challenge *allenai.org/data*

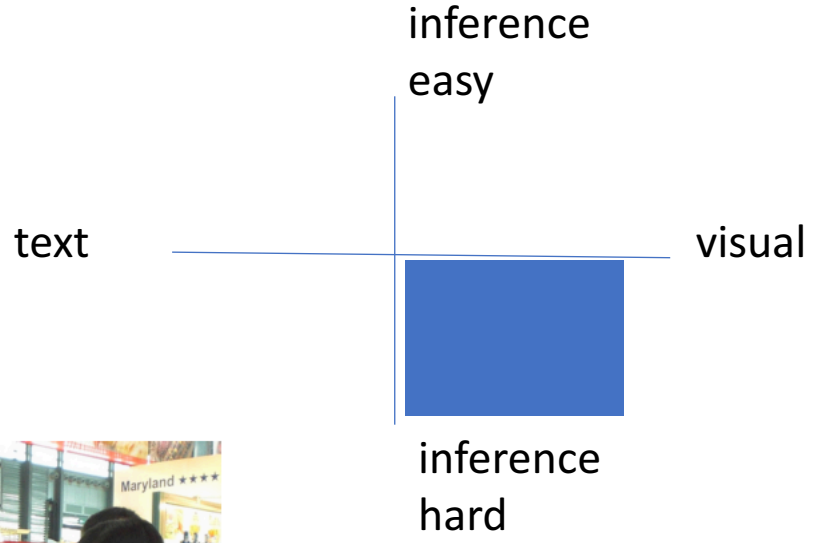


If all the frogs died, the raccoon population would most likely
(A) **decrease** (B) increase (C) remain the same



Challenge sets and problems for commonsense

TQA/ Aristo, FVQA [Wang et. al 2016]
includes supporting facts from KB



Q: What things in this image are eatable ?

A: Apples

Q: What is the order of the animal described in this image ?

A: Odd toed ungulate

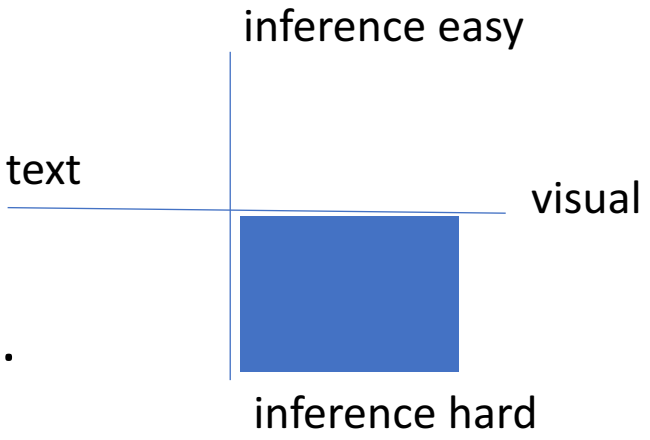
Q: What thing in this image is helpful for a romantic dinner ?

A: Wine

Challenge sets and problems for commonsense

TQA, Aristo, FVQA

Uses an LSTM and a data-driven approach to learn the mapping of images/questions to queries. Uses DBpedia, ConceptNet and WebChild.

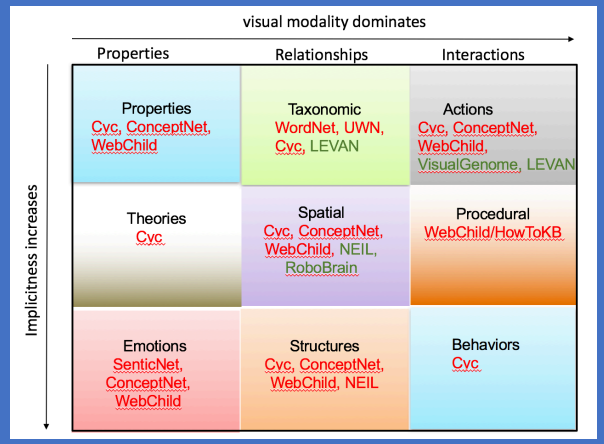


Question: What the red object on the ground can be used for?
Answer: Firefighting
Supporting Fact: Fire hydrant can be used for fighting fires.

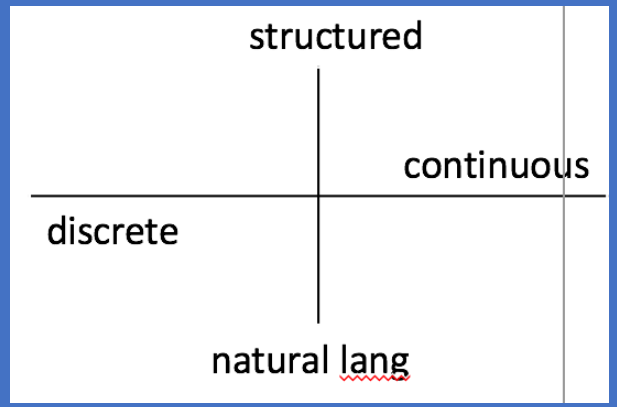
KB	Predicate	#Facts	Examples
DBpedia	Category	35152	(Wii,Category,VideoGameConsole)
ConceptNet	RelatedTo	79789	(Horse,RelatedTo,Zebra),(Wine,RelatedTo,Goblet)
	AtLocation	13683	(Bikini,AtLocation,Beach),(Tap,AtLocation,Bathroom)
	IsA	6011	(Broccoli,IsA,GreenVegetable)
	CapableOf	5837	(Monitor,CapableOf,DisplayImages)
	UsedFor	5363	(Lighthouse,UsedFor,SignalingDanger)
	Desires	3358	(Dog,Desires,PlayFrisbee),(Bee,Desires,Flower)
	HasProperty	2813	(Wedding,HasProperty,Romantic)
	HasA	1665	(Giraffe,HasA,LongTongue),(Cat,HasA,Claw)
	PartOf	762	(RAM,PartOf,Computer),(Tail,PartOf,Zebra)
	CreatedBy	96	(Bread,CreatedBy,Flour),(Cheese,CreatedBy,Milk)
WebChild	Smaller, Better, Slower, Bigger, Taller, ...	38576	(Motorcycle,Smaller,Car),(Apple,Better,VitaminPill), (Train,Slower,Plane),(Watermelon,Bigger,Orange), (Giraffe,Taller,Rhino)

Summary of Part 1

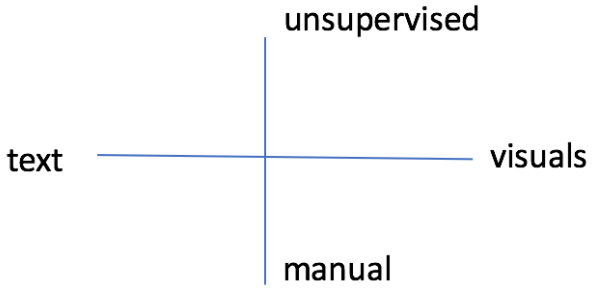
CSK types



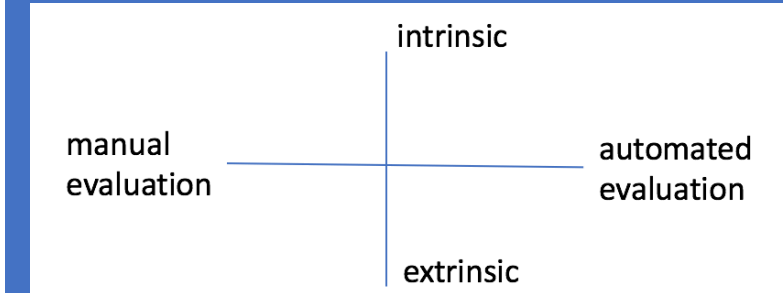
CSK representation



CSK acquisition



CSK evaluation



Future directions

1. Modeling implicit states through simulations:
   - Physical CSK – implicit states: roots absorb water → water is at roots
   - Social CSK– man: I broke up with Jenny, robot: do you want to go out with her?

2. Multimodal CSK (vision, text, and more e.g. audio for continuous learning)
   - Can commonsense be derived only from videos, how do we know what relations to focus?

3. Salient and concise KBs: efficient computations, quality control.
   - dogs have eyes, cats have eyes, squirrels have eyes.. vs. mammals have eyes.

4. Better evaluation metrics:
   - KB comprehensiveness as a metric– task independent
   - Extrinsic evaluations to continuously track progress

References

- Berant, J. (2012). Global Learning of Textual Entailment Graphs.
- Berant, J., & Clark, P. (2014). Modeling Biological Processes for Reading Comprehension (EMNLP)
- Bordes, A., & Gabrilovich, E. (2014). Constructing and Mining Web-scale Knowledge Graphs.
- Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference.
- Cambria, Eric (2012). SenticNet. http://sentic.net
- Y.-W. Chao, Z. Wang, R. Mihalcea, J. Deng. (2015) Mining semantic affordances of visual object categories. CVPR
- Chambers, N., & Jurafsky, D. (2010). A Database of Narrative Schemas. LREC
- Chang, A. X., Savva, M., Manning, C. (2014). Learning Spatial Knowledge for Text to 3D Scene Generation. EMNLP
- Chen, J., Tandon, N., & De Melo, G. (2016). Neural word representations from large-scale commonsense knowledge. WI
- Clark, P. (2014) Elementary School Science and Math Tests as a Driver for AI: Take the Aristo Challenge! IAAI
- Dalvi, B., Tandon, N., & Clark, P. (2017). Domain-Targeted, High Precision Knowledge Extraction. TACL
- de Melo, G., & Weikum, G. (2009). Towards a universal wordnet by learning from combined evidence. *CIKM*.

- Divvala, S. K., Farhadi, A., & Guestrin, C. (n.d.). Learning Everything about Anything: Webly-Supervised Visual Concept Learning.
- Galárraga, L., Teflioudi, C., Hose, K., & Suchanek, F. M. (2015). Fast rule mining in ontological knowledge bases with AMIE+. VLDB Journal
- Gordon, J., Durme, B. Van, & Schubert, L. K. (2010). Evaluation of Commonsense Knowledge with Mechanical Turk. Naacl-Hlt ’10
- Gordon, J., & Van Durme, B. (2013). Reporting bias and knowledge acquisition. In Proceedings of the 2013 workshop on Automated knowledge base construction - AKBC ’13
- Hajishirzi, A. K. and M. S. and D. S. and J. C. and A. F. and H. (2017). Are You Smarter Than A Sixth Grader ? Textbook Question Answering for Multimodal Machine Comprehension. CVPR
- Jain, P., Murty, S., Mausam, & Chakrabarti, S. (2017). Joint Matrix-Tensor Factorization for KB Inference.
- Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., et al. (2016). Visual genome: Connecting language and vision using crowdsourced dense image annotations.
- Lai, G., Xie, Q., Liu, H., Yang, Y., & Hovy, E. (2017). RACE: Large-scale ReAding Comprehension Dataset From Examinations.
- Li, X., Vilnis, L., & McCallum (2017). Improved Representation Learning for Predicting Commonsense Ontologies
- H. Liu and P. Singh (2004). ConceptNet: a practical commonsense reasoning tool- kit. BT Technology Journal.

- Liu, Q., Jiang, H., Ling, Z.-H., Zhu, X., Wei, S., & Hu, Y. (2016). Combing Context and Commonsense Knowledge Through Neural Networks for Solving Winograd Schema Problems
- Long, T., Bengio, E., Lowe, R., Chi, J., Cheung, K., & Precup, D. (2017). World Knowledge for Reading Comprehension: Rare Entity Prediction with Hierarchical LSTMs Using External Descriptions. In EMNLP
- Louvan, S., Naik, C., Lynn, V., Arun, A., Balasubramanian, N., & Clark, P. (2009). Semantic Role Labeling for Process Recognition Questions.
- Mesnil, G., Bordes, A., Weston, J., Chechik, G., & Bengio, Y. (2014). Learning semantic representations of objects and their parts. Machine Learning
- Modi, A., Titov, I., Demberg, V., Sayeed, A., & Pinkal, M. (2017). Modeling Semantic Expectation: Using Script Knowledge for Referent Prediction.
- Mostafazadeh, N., Chambers, N., He, X., Parikh, D., Batra, D., Vanderwende, et al. (2016). A Corpus and Evaluation Framework for Deeper Understanding of Commonsense Stories
- Palmer, M., Wu, S., & Titov, I. (2013). Semantic Role Labeling Tutorial. Naacl
- Pichotta, K., & Mooney, R. J. (2016). Learning Statistical Scripts With LSTM Recurrent Neural Networks (AAAI)
- Rohrbach, A., Rohrbach, M., Tandon, N., & Schiele, B. (2015). A dataset for Movie Description. CVPR
- Saxena, A., Jain, A., Sener, O., Jami, A., Misra, D. K., & Koppula, H. S. (2014). RoboBrain: Large-Scale Knowledge Engine for Robots.

- Tapaswi, M., Bäuml, M., & Stiefelhagen, R. (n.d.). Book2Movie: Aligning Video scenes with Book chapters.
- Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., & Fidler, S. (2015). MovieQA: Understanding Stories in Movies through Question-Answering.
- Trummer, I., Halevy, A., Lee, H., Sarawagi, S., & Gupta, R. (2015). Mining Subjective Properties on the Web. In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data - SIGMOD '15
- Vedantam, R., Lin, X., Batra, T., Zitnick, C. L., & Parikh, D. (2015). Learning common sense through visual abstraction. In Proceedings of the IEEE International Conference on Computer VisionWang, P., Wu, Q., Shen, C., Hengel, A. van den, & Dick, A. (2016). FVQA: Fact-based Visual Question Answering.
- Welbl, J., Stenetorp, P., & Riedel, S. (2017). Constructing Datasets for Multi-hop Reading Comprehension Across Documents.
- Wu, Q., Teney, D., Wang, P., Shen, C., Dick, A., & van den Hengel, A. (2016). Visual question answering: A survey of methods and datasets. Computer Vision and Image Understanding.
- Yang, B., & Mitchell, T. (2017). Leveraging Knowledge Bases in LSTMs for Improving Machine Reading. ACL
- Zhang, Y., Zhong, V., Chen, D., Angeli, G., & Manning, C. D. (n.d.). Position-aware Attention and Supervised Data Improve Slot Filling