Abstractions, Concepts, and Machine Learning

Haixun Wang

Orders of magnitude (data)

MB = 10^6 bytes

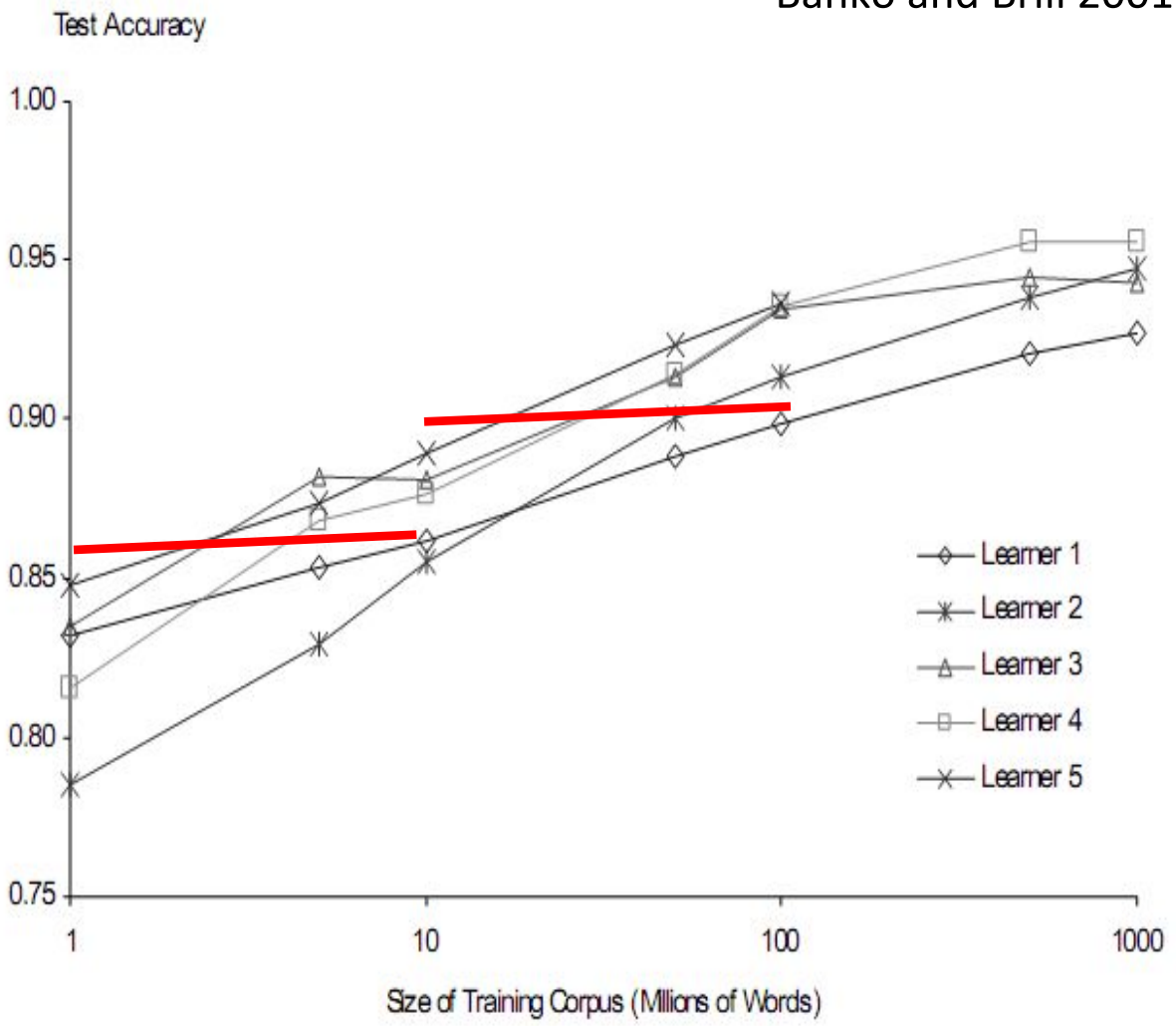
a typical book in text format

TB = 10^{12} bytes

astronomy data in one night;
US Library of Congress has 1000 TB data;
search log of Bing is 20 TB per day (2009)

More data beats better algorithms

Banko and Brill 2001

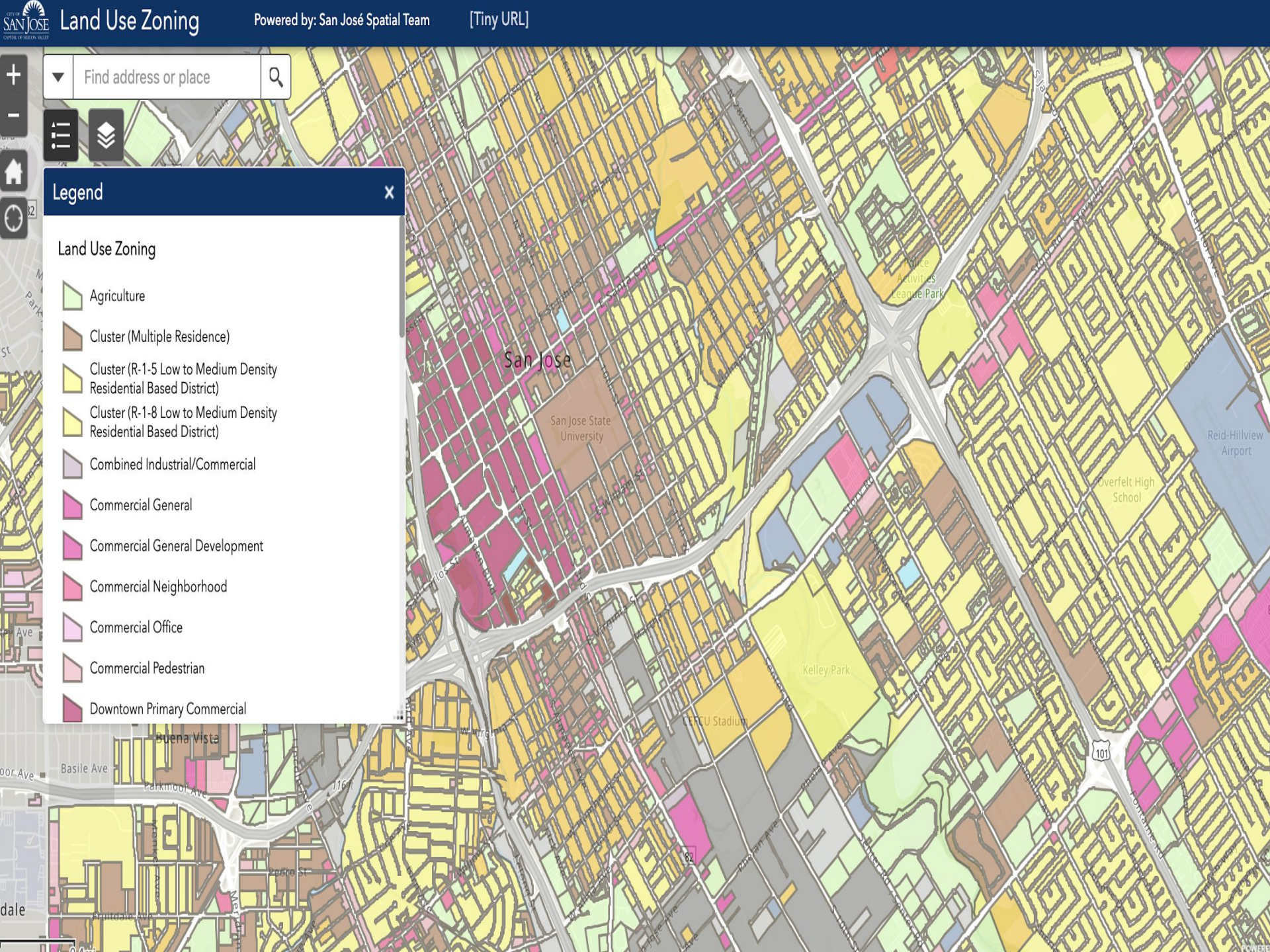


It’s a data driven world

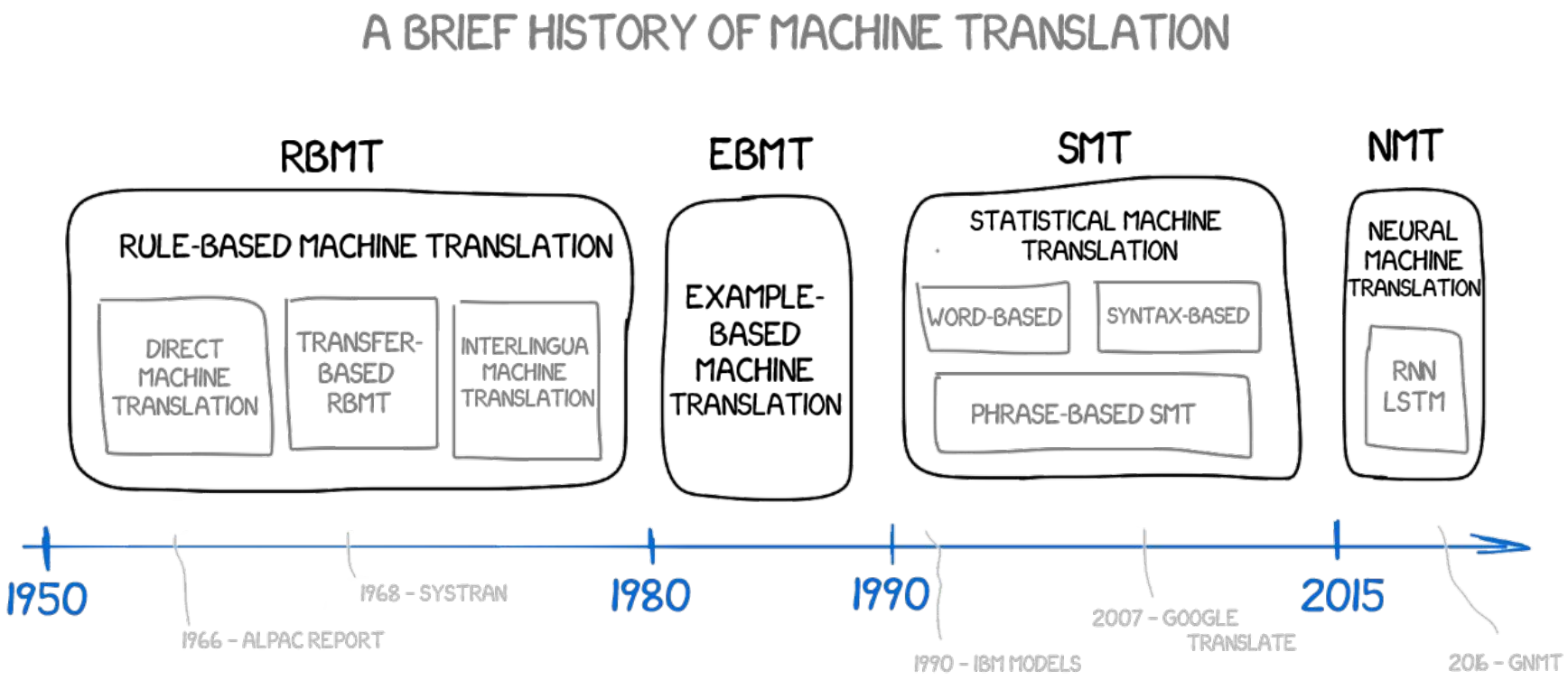
- Better telescopes, genome sequencers, fast computers, the Petabyte age, …
- Science today has turned into a data management problem

Satellite Imagery

Time lapse showing the construction of the Apple Headquarters in Cupertino, California. (Bird.I)

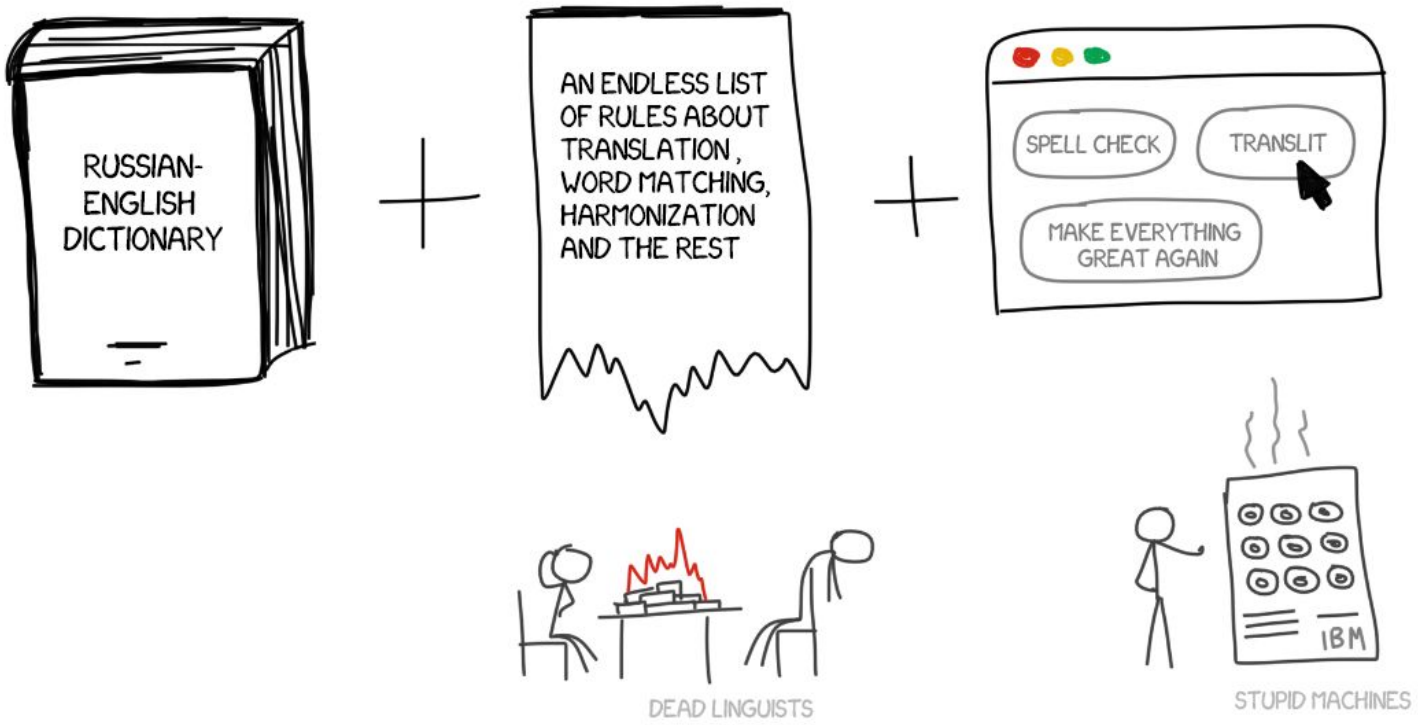

Land Use Zoning
Powered by: San José Spatial Team
[Tiny URL]
Find address or place
Legend
Land Use Zoning
Agriculture
Cluster (Multiple Residence)
Cluster (R-1-5 Low to Medium Density Residential Based District)
Cluster (R-1-8 Low to Medium Density Residential Based District)
Combined Industrial/Commercial
Commercial General
Commercial General Development
Commercial Neighborhood
Commercial Office
Commercial Pedestrian
Downtown Primary Commercial
San Jose
San Jose State University
Reid-Hillview Airport
Overfelt High School
Kelley Park
Buena Vista
Basile Ave
Parkmoor Ave

Machine Translation



from https://translartisan.wordpress.com/tag/example-based-machine-translation/

Rule based machine translation



from https://translartisan.wordpress.com/tag/example-based-machine-translation/

Parallel Texts

The Power of Data-Driven MT

- Systems improve by learning from human-produced translations
  - Adding more parallel data yields a better system
  - As the web grows, translation quality improves
  - Quality already exceeds best rule-based systems
- Given data, new language pairs can be launched very quickly
  - Haitian Creole <-> English: deployed in 4 days and 17 hours
  - A rule-based system would have taken months to build

The Shallowness of Google Translate

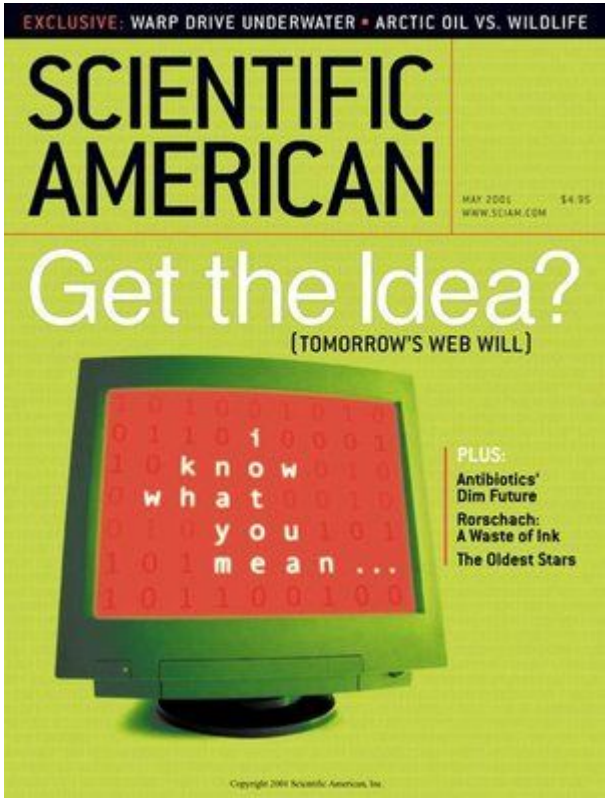
> In their house, everything comes in pairs. There's his car and her car, his towels and her towels, and his library and hers.

> Dans leur maison, tout vient en paires. Il y a sa voiture et sa voiture, ses serviettes et ses serviettes, sa bibliothèque et les siennes.

from https://www.theatlantic.com/technology/archive/2018/01/the-shallowness-of-google-translate/551570/

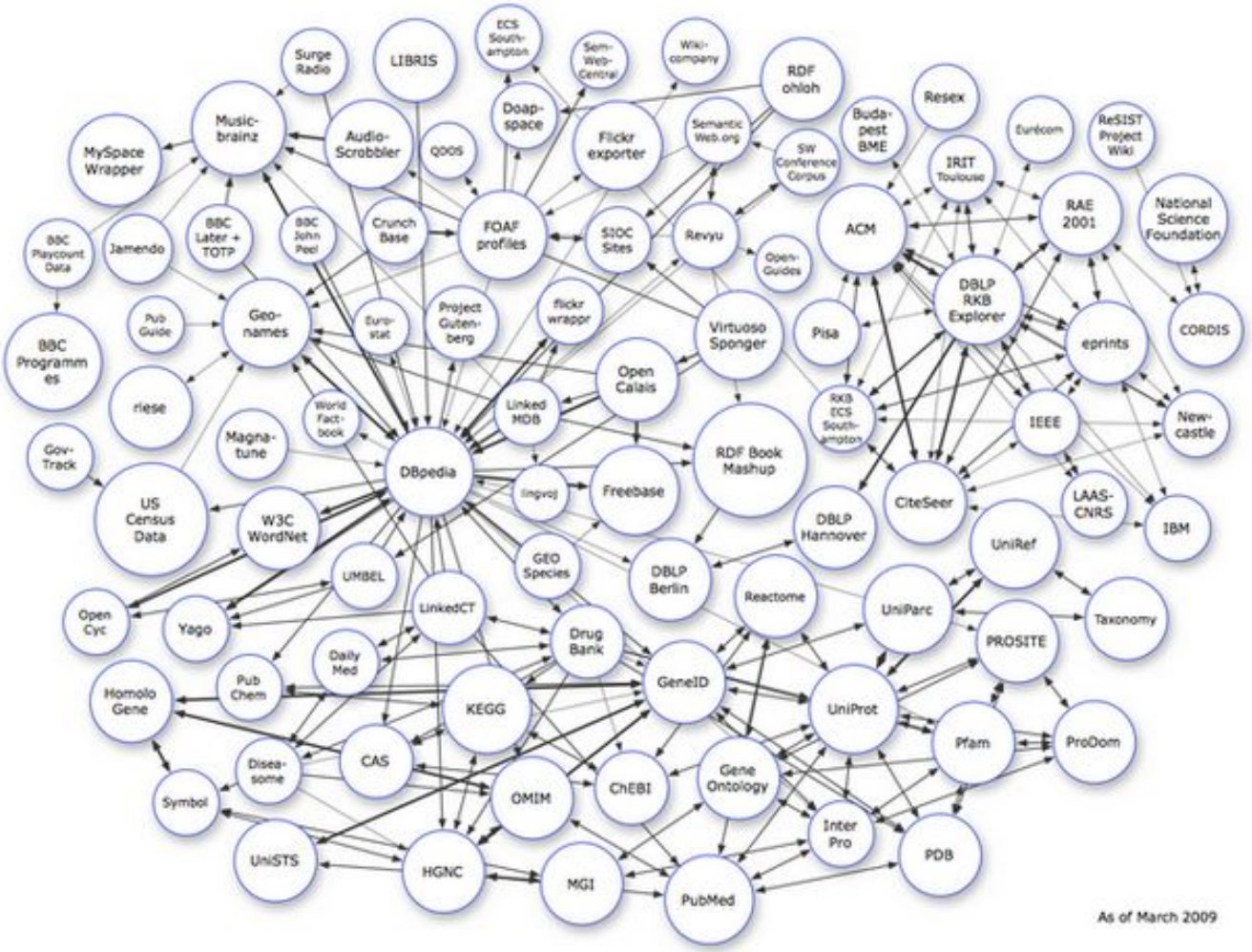
What Happened to the Semantic Web?

- The great promise of the Semantic Web was that it would be readable not just by humans but also by machines. -- 2001
- Unstructured information will give way to structured information – paving the road to more intelligent computing. -- Alex Iskold
- The infrastructure to power the Semantic Web is already here. -- Tim Berners-Lee



Turning the Web into a Database

We are now in a Web of Data



Linking Open Data community project are populating the Web with vast amounts of distributed yet interlinked RDF data.

According to Berners-Lee, Lassila, and Hendler ...

The entertainment system was belting out the Beatles' "We Can Work It Out" when the phone rang. When Pete answered, his phone turned the sound down by sending a message to all the other local devices that had a volume control. His sister, Lucy, was on the line from the doctor's office: "Mom needs to see a specialist and then has to have a series of physical therapy sessions. Biweekly or something. I'm going to have my agent set up the appointments." Pete immediately agreed to share the chauffeuring. At the doctor's office, Lucy instructed her Semantic Web agent through her handheld Web browser. The agent promptly retrieved the information about Mom's prescribed treatment within a 20-mile radius of her home and with a rating of excellent or very good on trusted rating services. It then began trying to find a match between available appointment times (supplied by the agents of individual providers through their Web sites) and Pete's and Lucy's busy schedules.[1]

GOOGLE ASSISTANT MAKING APPOINTMENTS THROUGH PHONECALL

"Do you have anything between 10 am and 12 pm?"

"Depending on what service she would like. What service is she looking for?"

Now... *that* should clear up a few things around here

It does not work

- The dream of the Semantic Web is to develop one ontology expressed in one language potentially covering everything that exists on the web
- Reality: there are several ontologies in several languages covering partly overlapping subdomains of the web
- integration problems, including structural heterogeneity, semantic heterogeneity, inconsistency and redundancy problems

Criticism

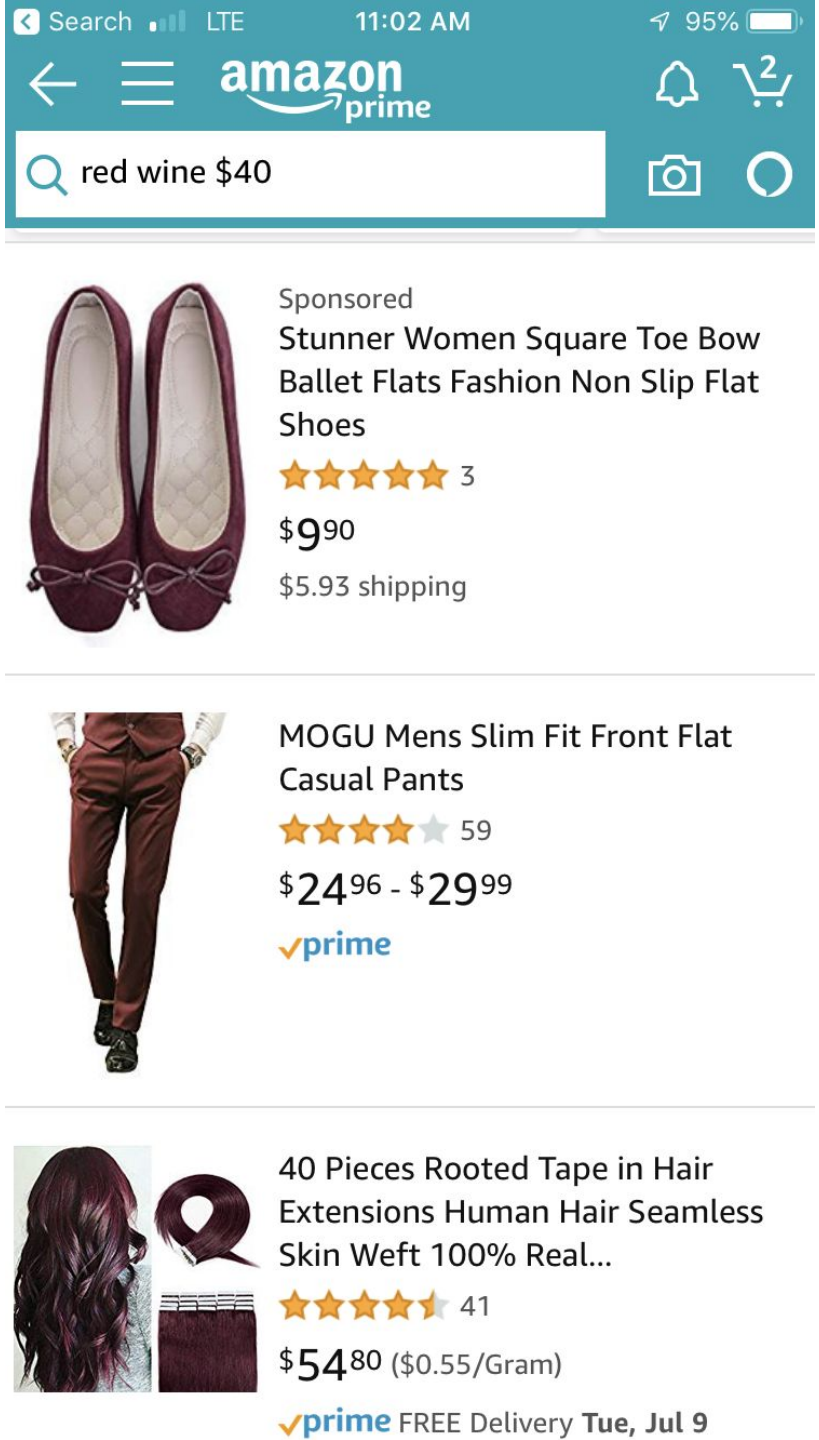
Like many visions that project future benefits but ignore present costs, it requires too much coordination and too much energy to effect in the real world, where deductive logic is less effective and shared world view is harder to create than we often want to admit.

- C. Shirky. The Semantic Web, Syllogism, and Worldview. http://www.shirky.com/writings/semantic\-syllogism.html, 2003

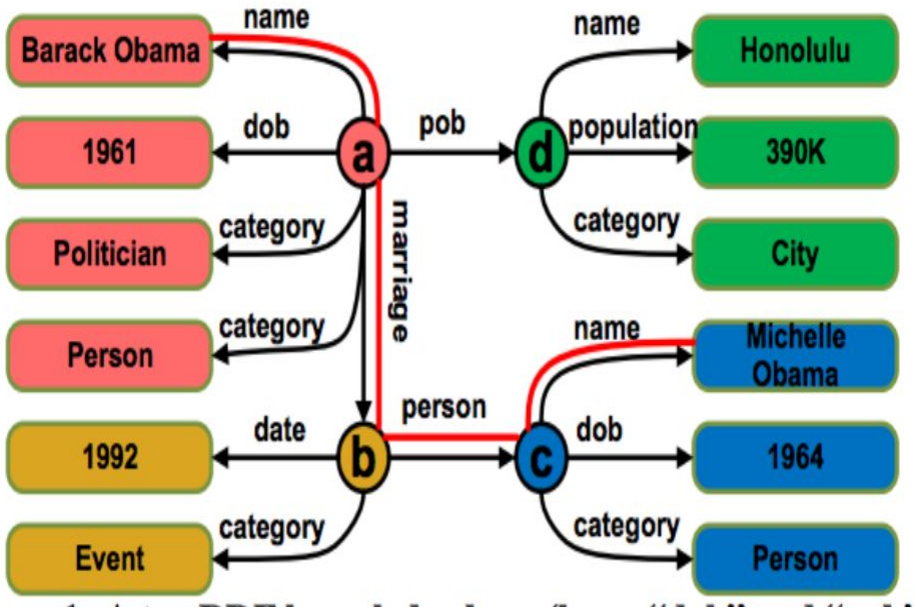
Text Understanding

Amazon Search “red wine $40”

(screenshot taken July 7, 2019)

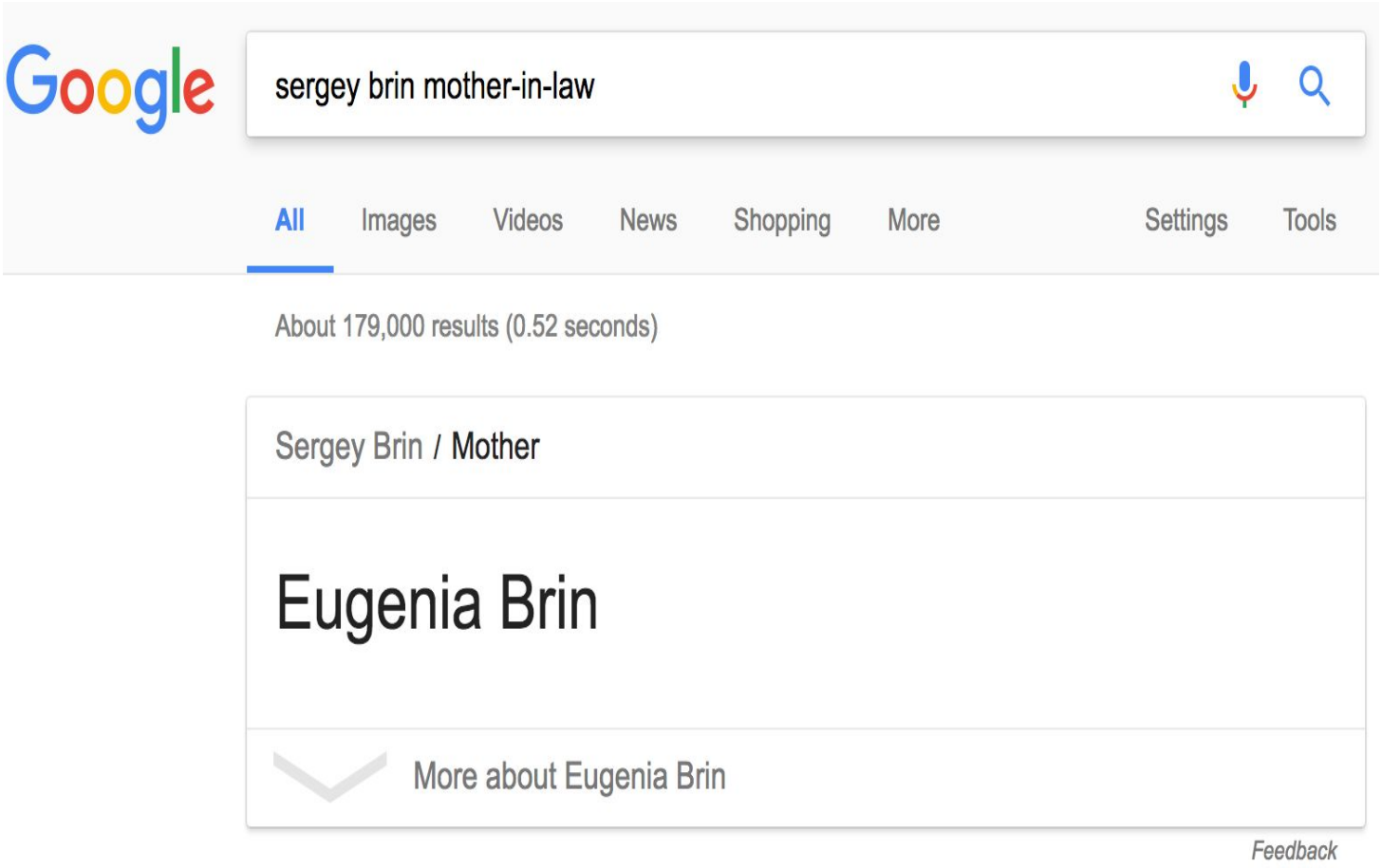


Knowledge-base Question Answering

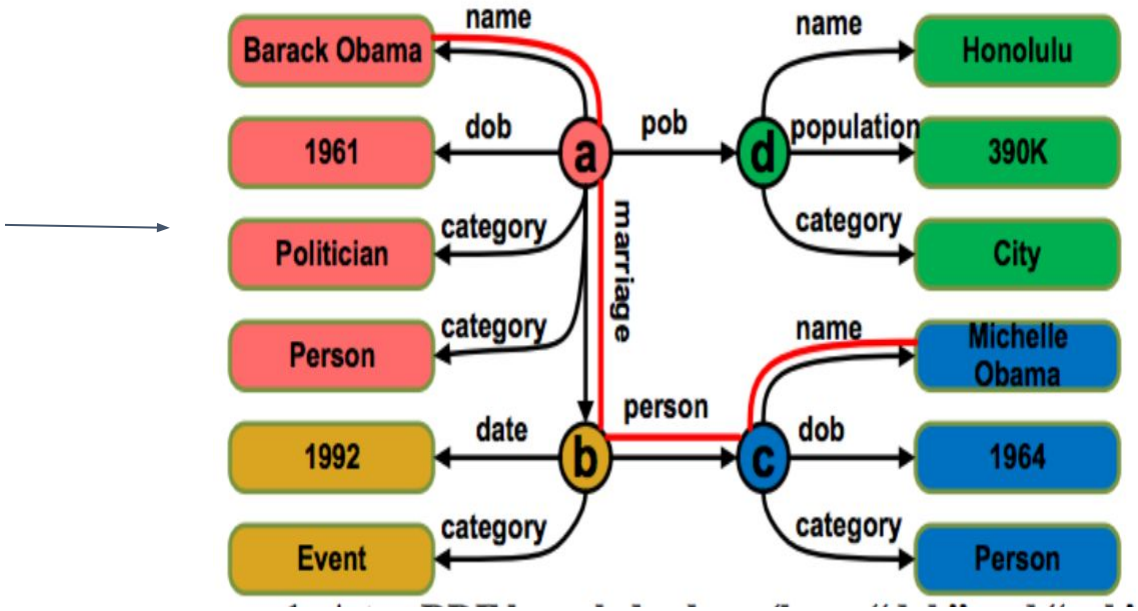


KBQA: Learning Question Answering over QA Corpora and Knowledge Bases, VLDB 2017

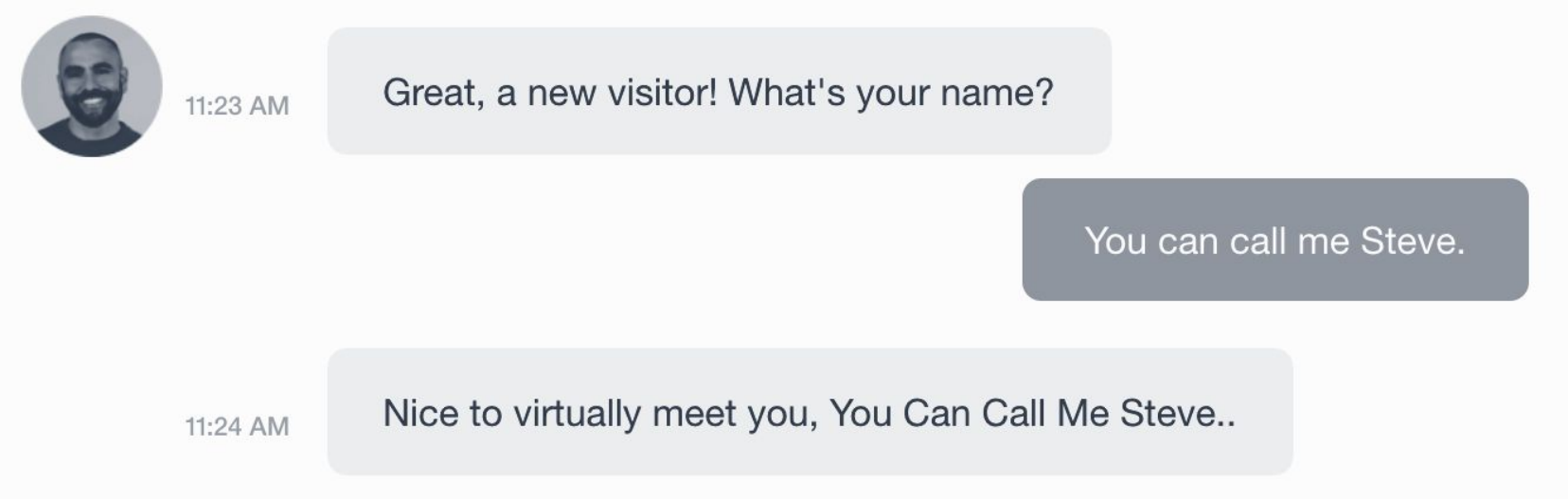
It’s hard …


Google
sergey brin mother-in-law
All
Images
Videos
News
Shopping
More
Settings
Tools
About 179,000 results (0.52 seconds)
Sergey Brin / Mother
Eugenia Brin
More about Eugenia Brin
Feedback


SALUT!
GRÜSS DICH!
HELLO!
CIAO!
HEJ!
SALUTON!
SELAM!
¡HOLA!
AMOJ!

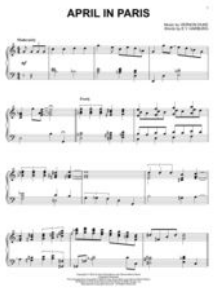

Barack Obama
name
1961
dob
Politician
category
Person
category
1992
date
Event
category
a
b
marriage
pob
person
d
c
Honolulu
name
population
390K
City
category
Michelle Obama
name
1964
dob
Person
category

Conversation, Chatbots



Text Understanding: Easy Tasks

- watch harry potter
- harry potter watch
- april in paris lyrics
- april in paris temperature



Text Understanding: Difficult Tasks

- eye drops off shelf
- time flies an arrow; fruit flies like a banana
- jordan 7 day weather forecast

The New York Times

Monday, March 24, 2014 | Today's Paper | Personalize Your Weather

Malaysia Says Jet Went Down in Ocean

Families Notified as New Analysis Shows Southern Path

By THOMAS FULLER and CHRIS BUCKLEY 10:48 AM ET

Based on satellite data, Prime Minister Najib Razak said Monday that there was no doubt that Flight 370 flew south into the Indian Ocean and could not have landed safely.

296 Comments

· **Video: Prime Minister's News Conference**

Rolex Dela Pena/European Pressphoto Agency

Relatives of passengers from the missing flight in Beijing.

The Opinion Pages

Vaccination and the Law

Do outbreaks of measles show why exemptions on immunization should end?

ROOM for DEBATE

- **Editorial: Willfully Endangering Drivers**
- **Krugman: Wealth Over Work**
- **Baird: Queen Victoria, Another Maligned Mother**

MARKETS » At 2:31 PM ET

S.&P. 500	Dow	Nasdaq
1,858.75	16,286.22	4,230.19
–7.77	–16.55	–46.60
–0.42%	–0.10%	–1.09%

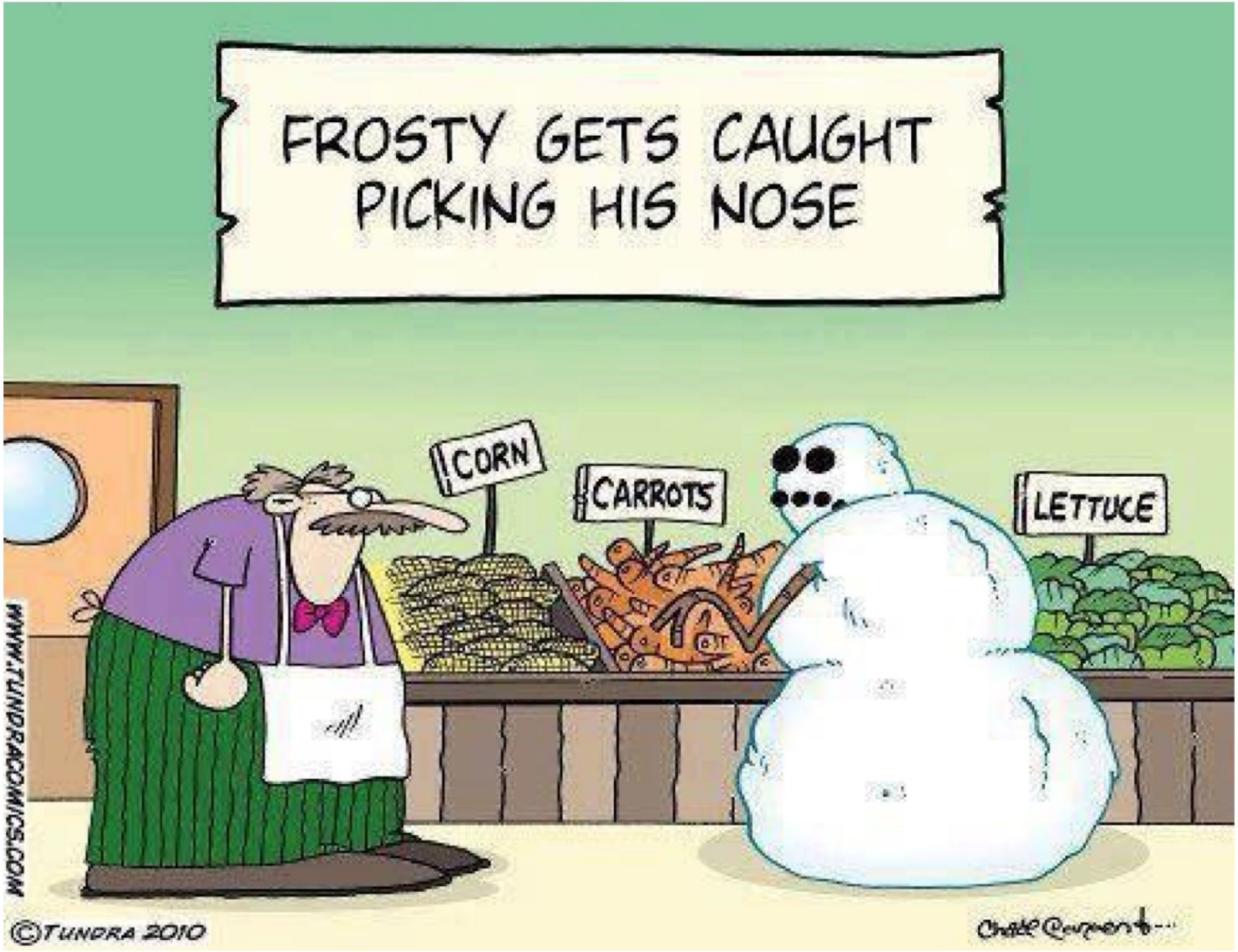
Get Quotes | My Portfolios »

TECHNOLOGY »

Web Fiction, Serialized and Social

With Wattpad, the once-solitary writing process has become informal, intimate and highly interactive.

TRAVEL »


FROSTY GETS CAUGHT PICKING HIS NOSE
CORN
CARROTS
LETTUCE
WWW.TUNDRACOMICS.COM
©TUNDRA 2010

Sometimes humans are clueless too

ASUS

ASUS Intel Core i5 8GB DDR3 1TB HDD Capacity Desktop PC Windows 7 Professional P8H61E (BP6320-I53470163B)

3 years next business day onsite service

(2) | Write a Review

In stock. Limit 5 per customer.

- Intel Core i5 3470(3.20GHz)
- 8GB DDR3 1TB HDD Capacity
- Windows 7 Professional
- Energy star certification

Knowledge of Language

Understanding
(internal representation)

Knowledge

Knowledge

Question

Understanding
(internal representation)

Answer

Knowledge of Language

- Mary puts on a coat every time she leaves the house.
  she = Mary (possible)
- She puts on a coat every time Mary leaves the house.
  she = Mary (impossible)
- Every time Mary leaves the house she puts on a coat.
  she = Mary (possible)
- Every time she leaves the house Mary puts on a coat.
  she = Mary (possible)

From Omer Preminger's 2012 MIT Lectures

The big question

- How does the mind get so much out of so little?
- Our minds build rich models of the world and make strong generalizations from input data that is *sparse, noisy, and ambiguous* – in many ways far too limited to support the inferences we make.
- How do we do it?

Science
AAAS
Science **331**, 1279 (2011);

How to Grow a Mind: Statistics, Structure, and Abstraction

Joshua B. Tenenbaum,[1*] Charles Kemp,[2] Thomas L. Griffiths,[3] Noah D. Goodman[4]

MIT CMU Berkeley Stanford

If the mind goes beyond the data given, *another source of information* must make up the difference.

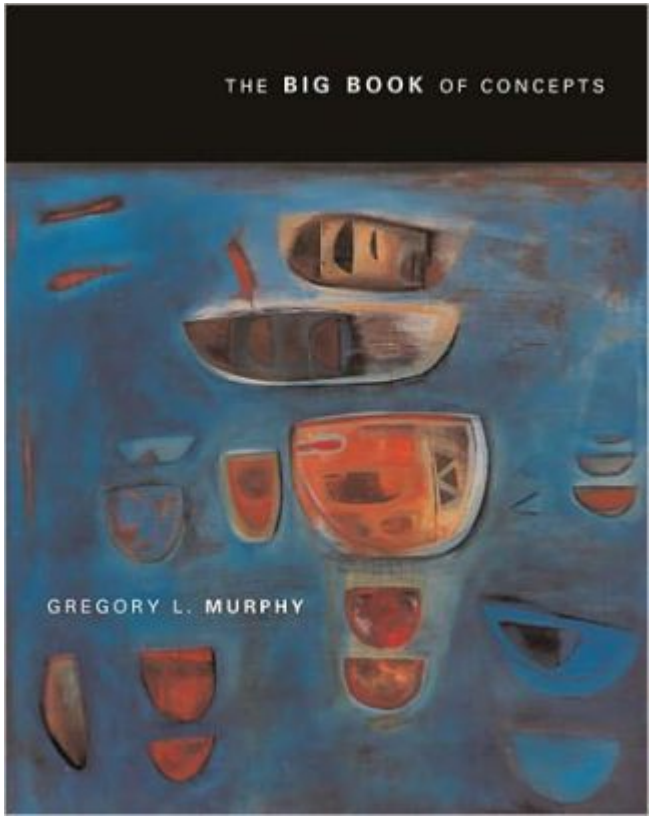
3 important topics

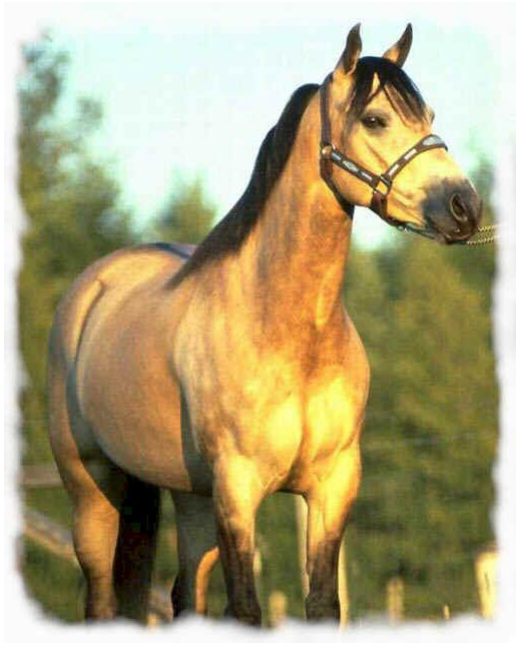
- Conceptualization
- Priors
- Structures

Why use the concept space?

"Concept are the glue that holds our mental world together."

Think of data sparseness.

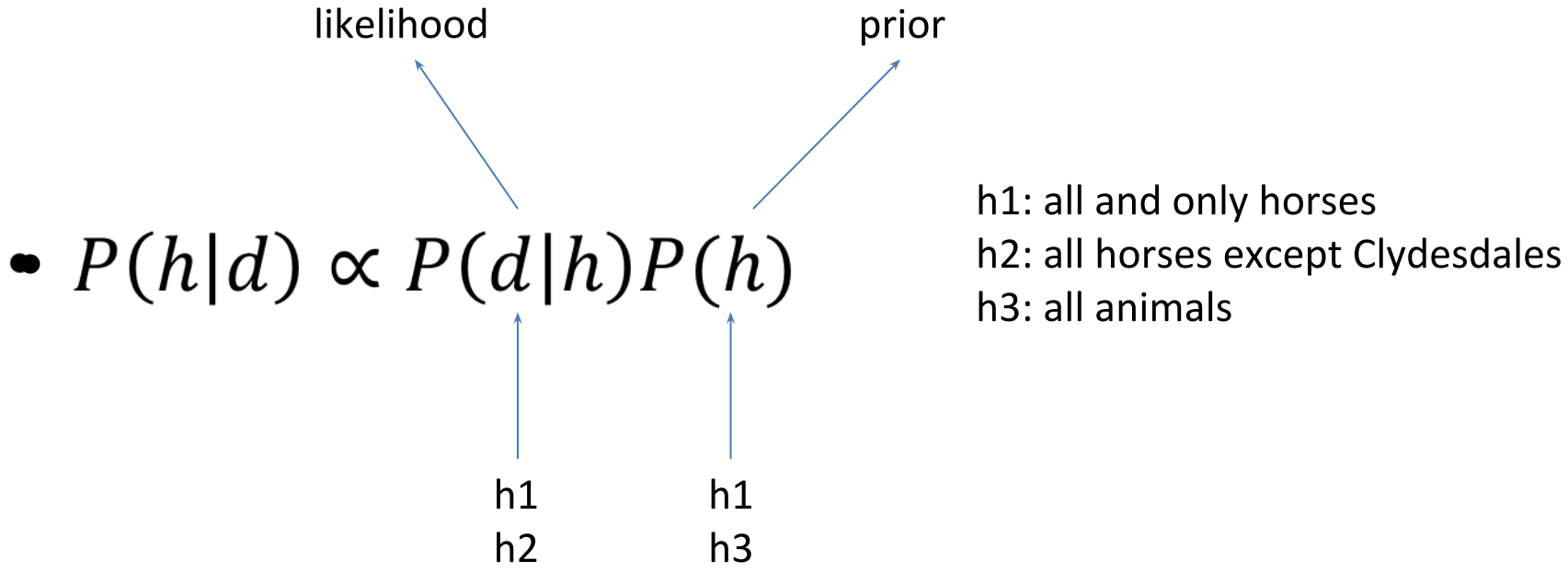




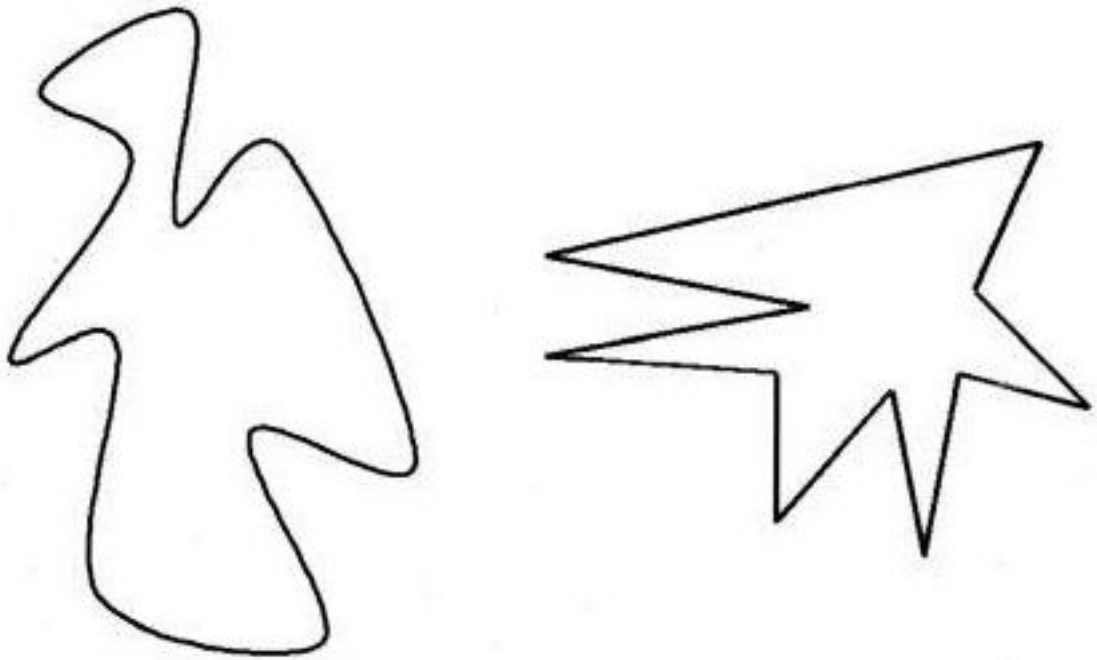
h1: all and only horses

h2: all horses except Clydesdales

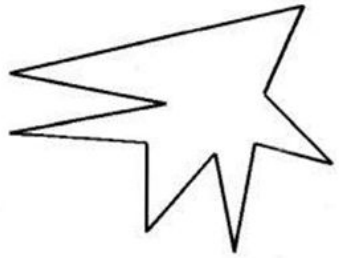
h3: all animals


likelihood
prior
$P(h|d) \propto P(d|h)P(h)$
h1
h2
h1
h3
h1: all and only horses
h2: all horses except Clydesdales
h3: all animals

Which is “kiki” and which is “bouba”?



\ˈkēkē



sound shape

zigzaggedness

What does this date signify?

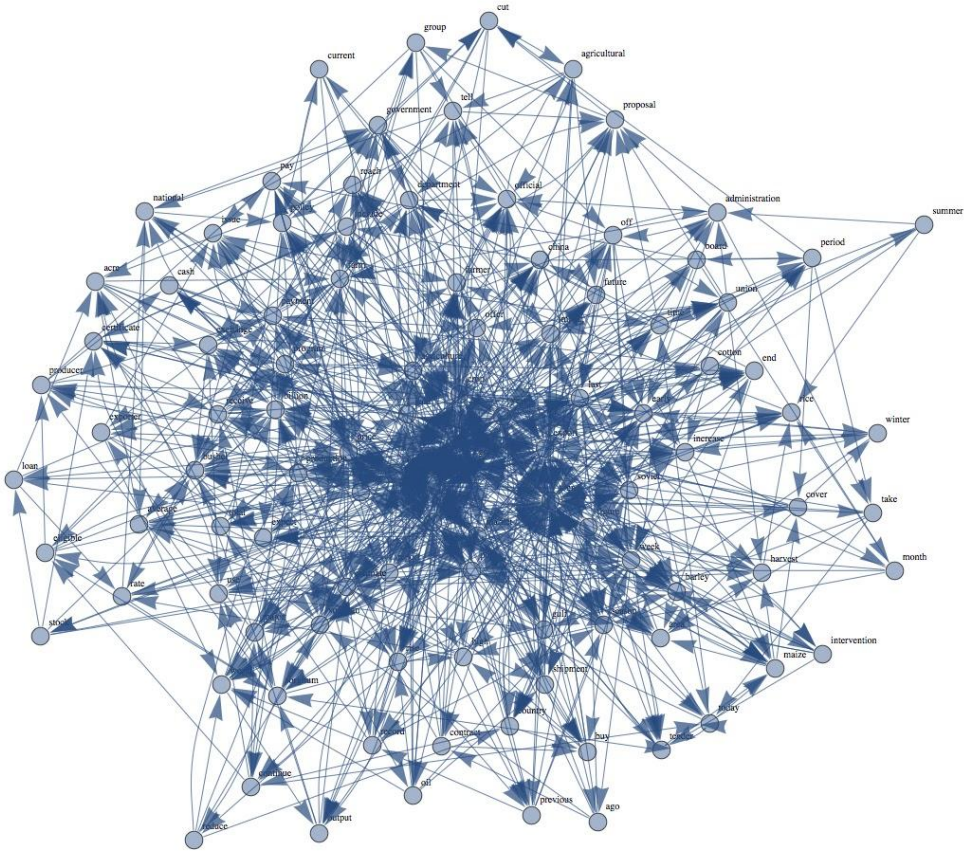
25 Oct 1881

What does the date signify?

Pablo Picasso	25 Oct 1881	Spanish

What triggers the “birthday” concept in our mind?

Probase: a semantic network for text understanding



Concepts Entities isA isPropertyOf Co-occurrence Verb/Adj

Knowledge in Probase

- What people know when they know a language?
- It is not about right or wrong. It's just usage.
- Words/phrases/constructs evoke a network of concepts which lead to understanding.

Probase vs. Knowledgebase

Probase	Knowledgebases
common/linguistic knowledge	facts
isA isPropertyOf co-occurrence …	DayOfBirth LocatedIn SpouseOf …
typicality, basic level of categorization	precision

Probase vs. Knowledge Bases

	WordNet	Wikipedia	Freebase	Probase
Cat	Feline; Felid; Adult male; Man; Gossip; Gossiper; Gossipmonger; Rumormonger; Rumourmonger; Newsmonger; Woman; Adult female; Stimulant; Stimulant drug; Excitant; Tracked vehicle; ...	Domesticated animals; Cats; Felines; Invasive animal species; Cosmopolitan species; Sequenced genomes; Animals described in 1758;	TV episode; Creative work; Musical recording; Organism classification; Dated location; Musical release; Book; Musical album; Film character; Publication; Character species; Top level domain; Animal; Domesticated animal; ...	Animal; Pet; Species; Mammal; Small animal; Thing; Mammalian species; Small pet; Animal species; Carnivore; Domesticated animal; Companion animal; Exotic pet; Vertebrate; ...
IBM	N/A	Companies listed on the New York Stock Exchange; IBM; Cloud computing providers; Companies based in Westchester County, New York; Multinational companies; Software companies of the United States; Top 100 US Federal Contractors; ...	Business operation; Issuer; Literature subject; Venture investor; Competitor; Software developer; Architectural structure owner; Website owner; Programming language designer; Computer manufacturer/brand; Customer; Operating system developer; Processor manufacturer; ...	Company; Vendor; Client; Corporation; Organization; Manufacturer; Industry leader; Firm; Brand; Partner; Large company; Fortune 500 company; Technology company; Supplier; Software vendor; Global company; Technology company; ...
Language	Communication; Auditory communication; Word; Higher cognitive process; Faculty; Mental faculty; Module; Text; Textual matter;	Languages; Linguistics; Human communication; Human skills; Wikipedia articles with ASCII art	Employer; Written work; Musical recording; Musical artist; Musical album; Literature subject; Query; Periodical; Type profile; Journal; Quotation subject; Type/domain equivalent topic; Broadcast genre; Periodical subject; Video game content descriptor; ...	**Instance of**: Cognitive function; Knowledge; Cultural factor; Cultural barrier; Cognitive process; Cognitive ability; Cultural difference; Ability; Characteristic; **Attribute of:** Film; Area; Book; Publication; Magazine; Country; Work; Program; Media; City; ...

isA Extraction

- Hearst pattern

  NP such as NP, NP, …, and|or NP

  such NP as NP,* or|and NP

  NP, NP*, or other NP

  NP, NP*, and other NP

  NP, including NP,* or | and NP

  NP, especially NP,* or|and NP

- … is a … pattern

  NP is a/an/the NP

- *domestic animals* such as *cats* and *dogs* …
- animals other than *cats* such as *dogs* …
- *China* is a *developing country.*
- *Life* is a box of *chocolate*.

How hard is it?

It has every problem NLP has.

For example, coref:

- Most frequently used metals in the catalysts are critical ones such as cobalt, nickel, …
- The breweries represented included some of Britain's finest such as Magic Rock, …
- Dig through the buckets of flowers and find the ones that look prettiest such as carnations, mums and daisies, …

How hard is it?

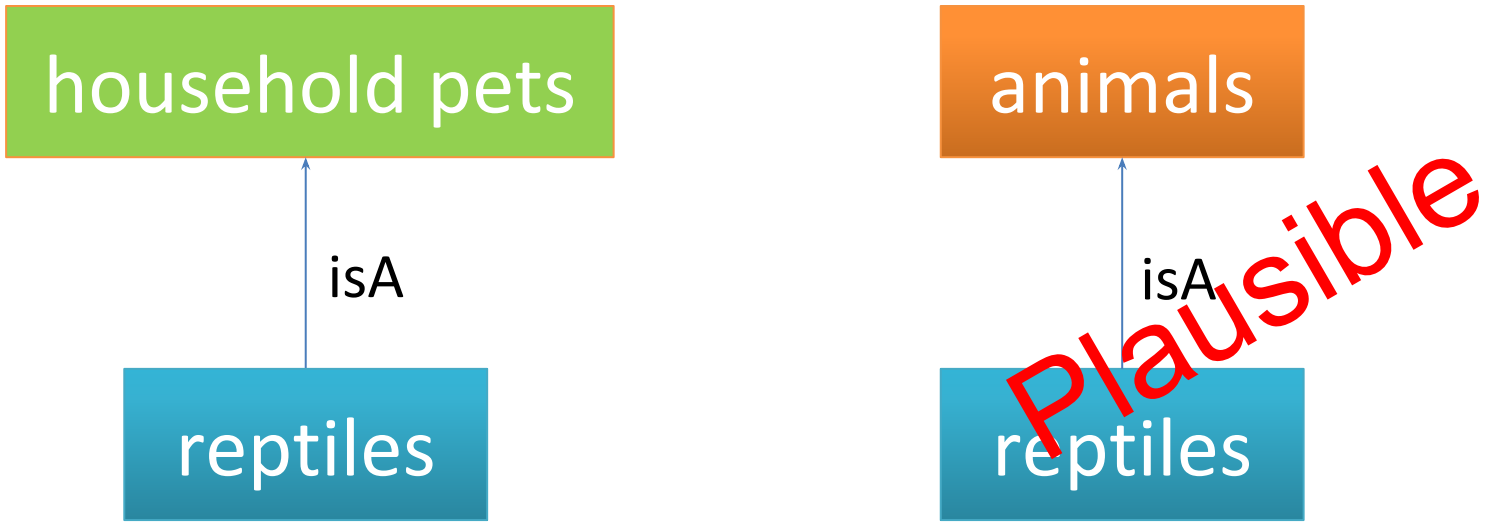
Example: entity and concepts

- sports such as track and field
- tv shows such as sex and city

… **animals** other than **cats** such as **dogs**

…

… **household pets** other than **animals** such as **reptiles**, aquarium fish …



Quiz: What X breaks the extraction?

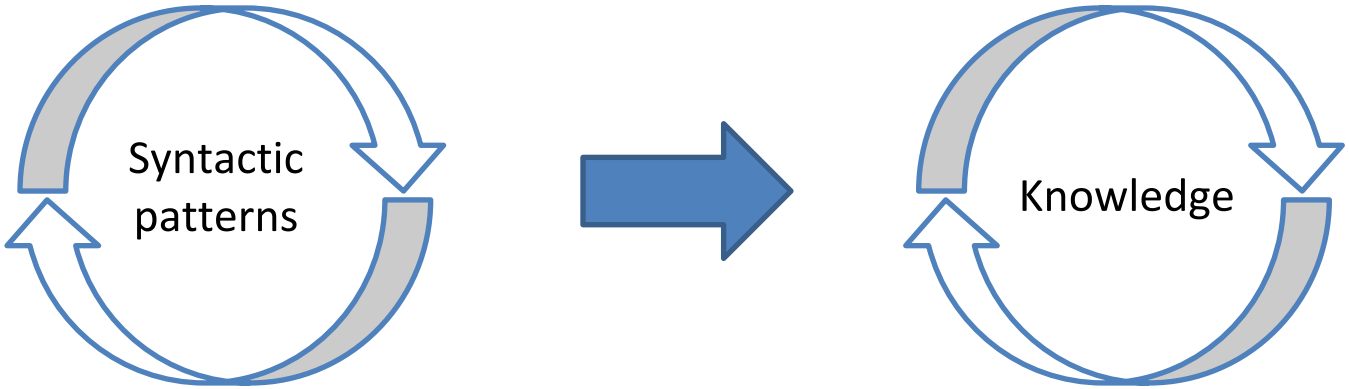
X such as **democracy** and **authoritarianism.**

Can we extract:

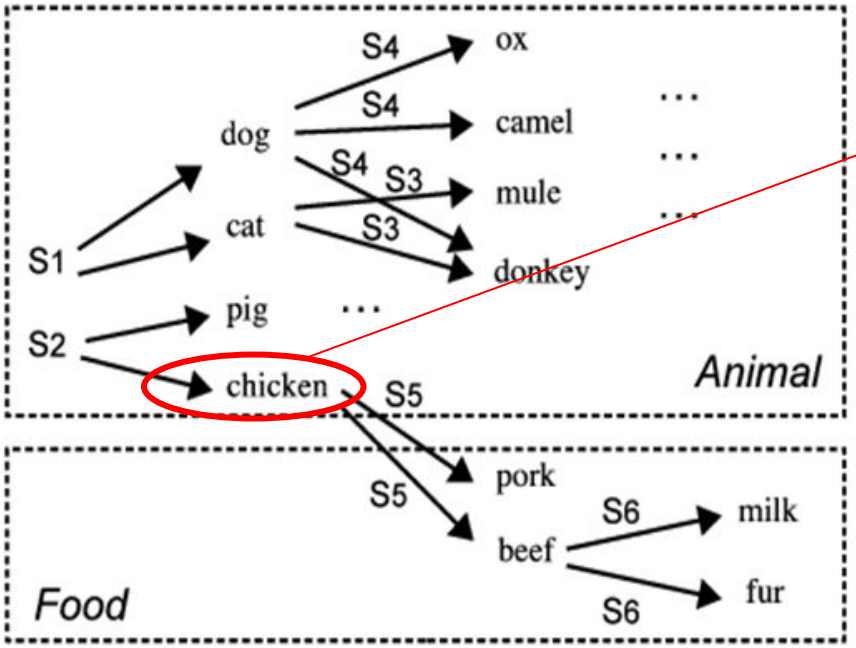
democracy isA **X**

authoritarianism isA **X**

Iterative Information Extraction



Semantic Drifts



A drifting point

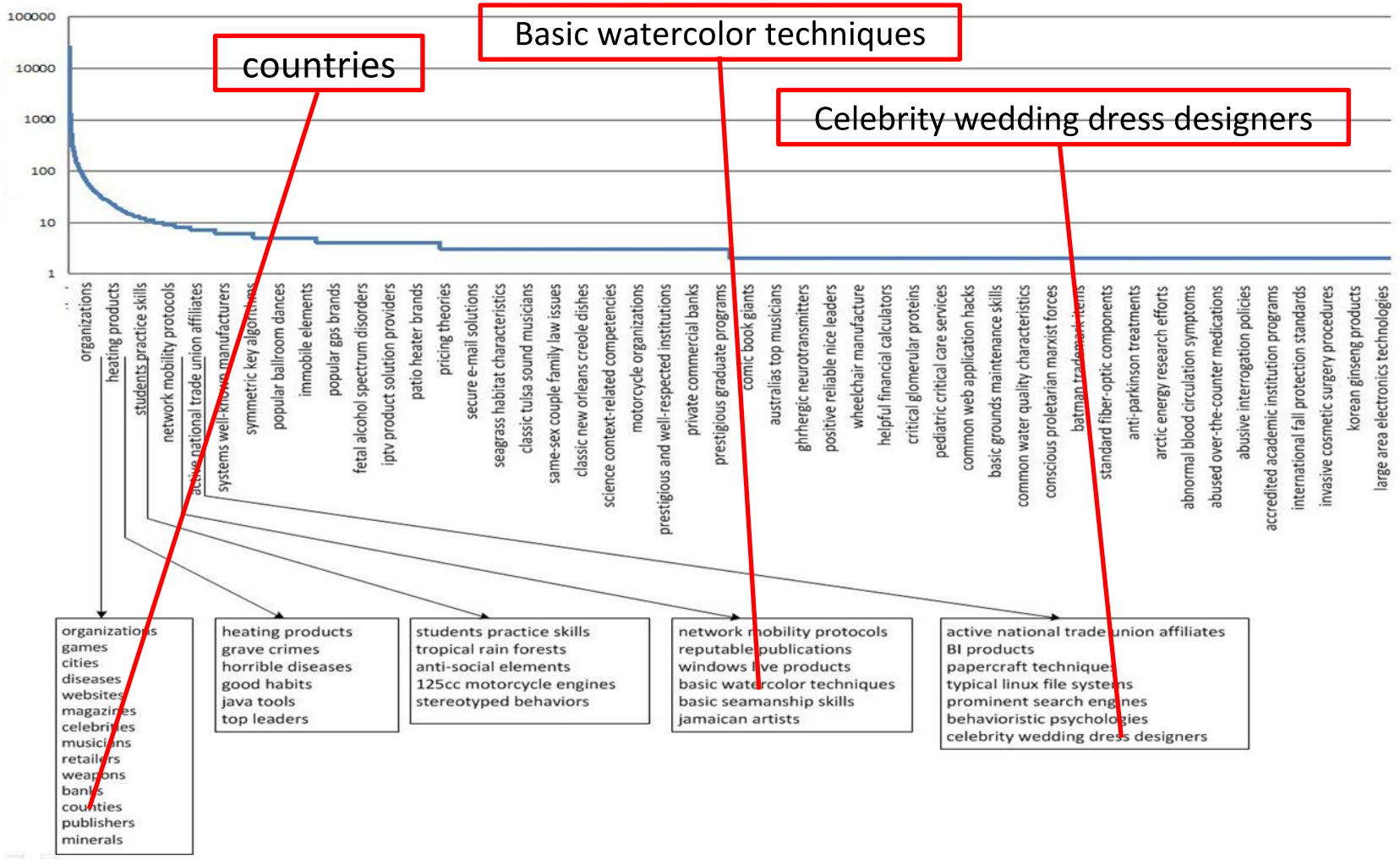
S1="Animals **such as** dogs and cats, grow fast."
S2="Land animals **such as** chicken and pigs – all of which live on land"

S3="Postures are often named after animals, **such as** mule, donkey and cat."
S4="... innkeeper, angels, and animals **such as** ox, camels, donkeys and dog"
S5="Common food from animals such as pork, beef and chicken"

S6="Products from animals **such as** fur, milk and beef are given to families..."

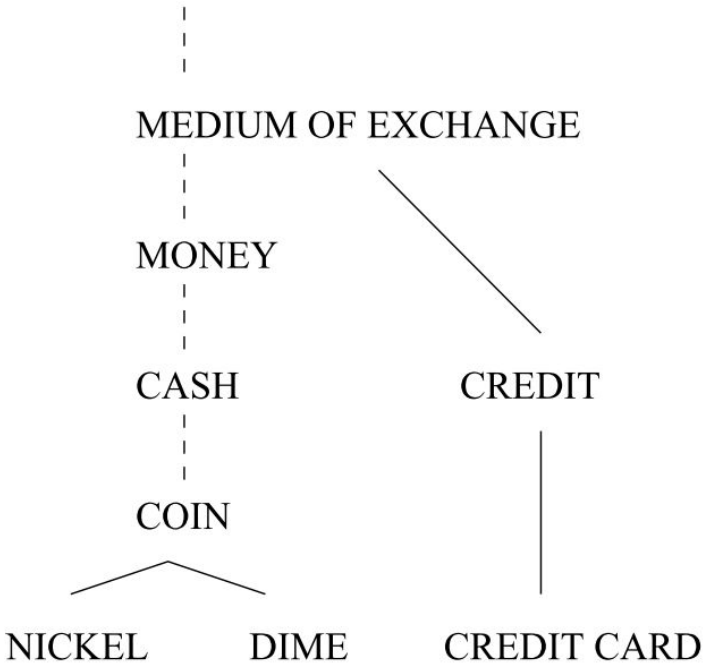
10%-20% Precision Improvement

Probase Concepts (2.7 million+)

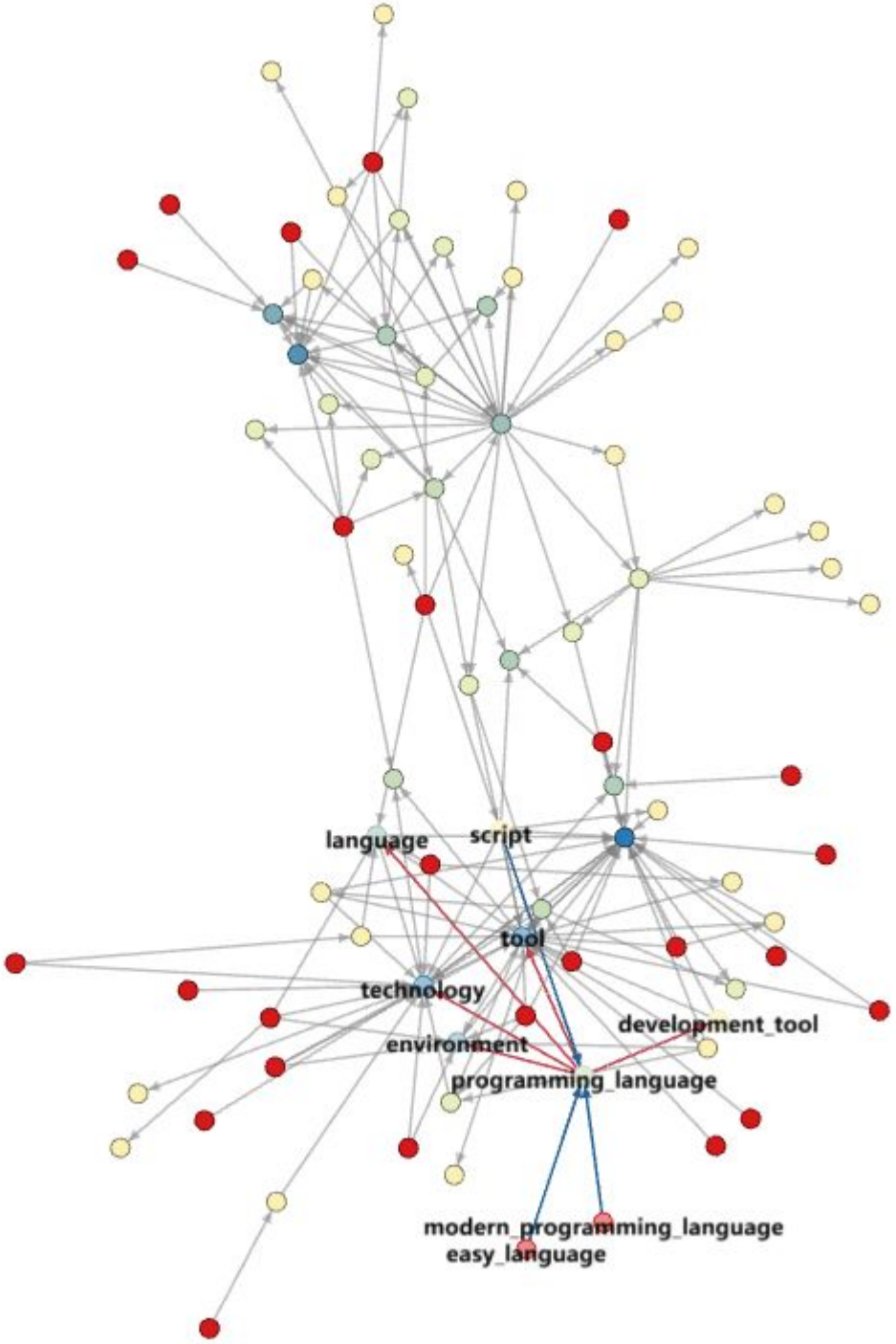


Probase isA error rate: <%1 @1 and <10% for random pair

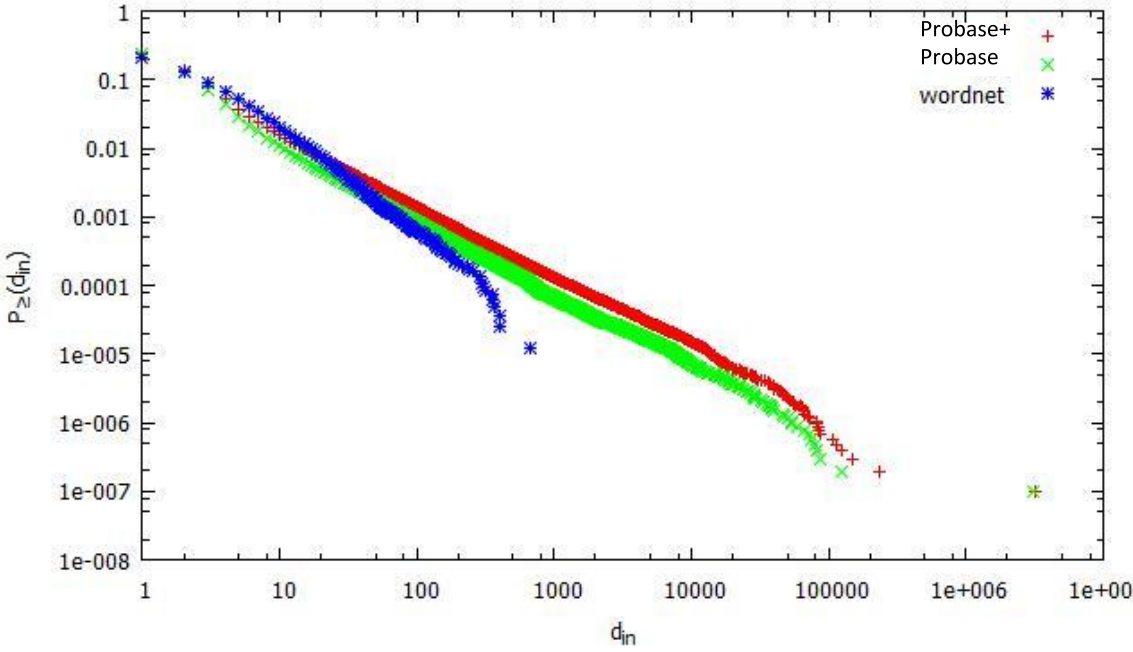
A traditional taxonomy



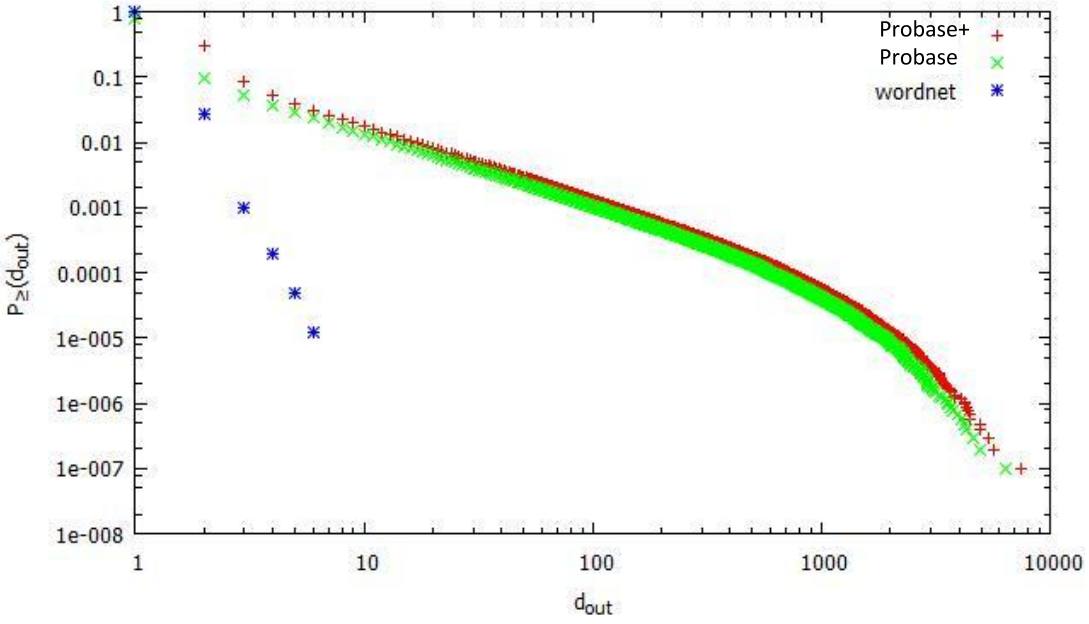
“python”


language
script
tool
technology
development_tool
environment
programming_language
modern_programming_language
easy_language

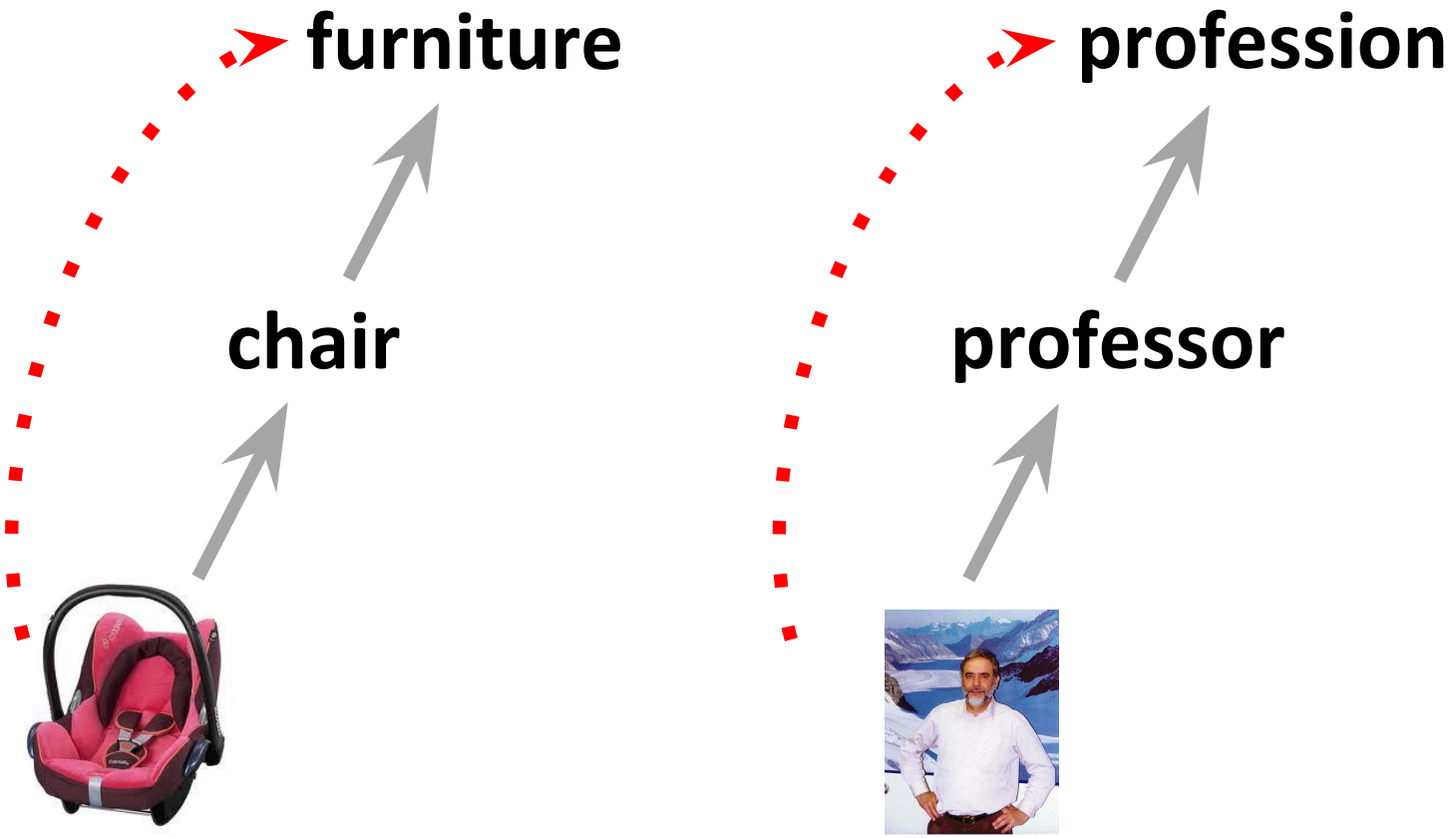
# of child concepts



# of parent concepts

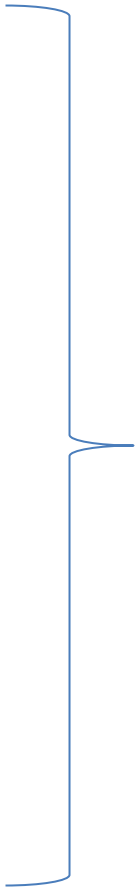


Transitivity does not always hold



Probase Scores

- Typicality
- Vagueness
- BLC (basic level of categorization)
- Ambiguity
- Similarity



foundation for inferencing

Typicality

bird

$$P(e|c) = \frac{n(c,e) + \alpha}{\sum_{e_i \in c} n(c,e_i) + \alpha N}$$

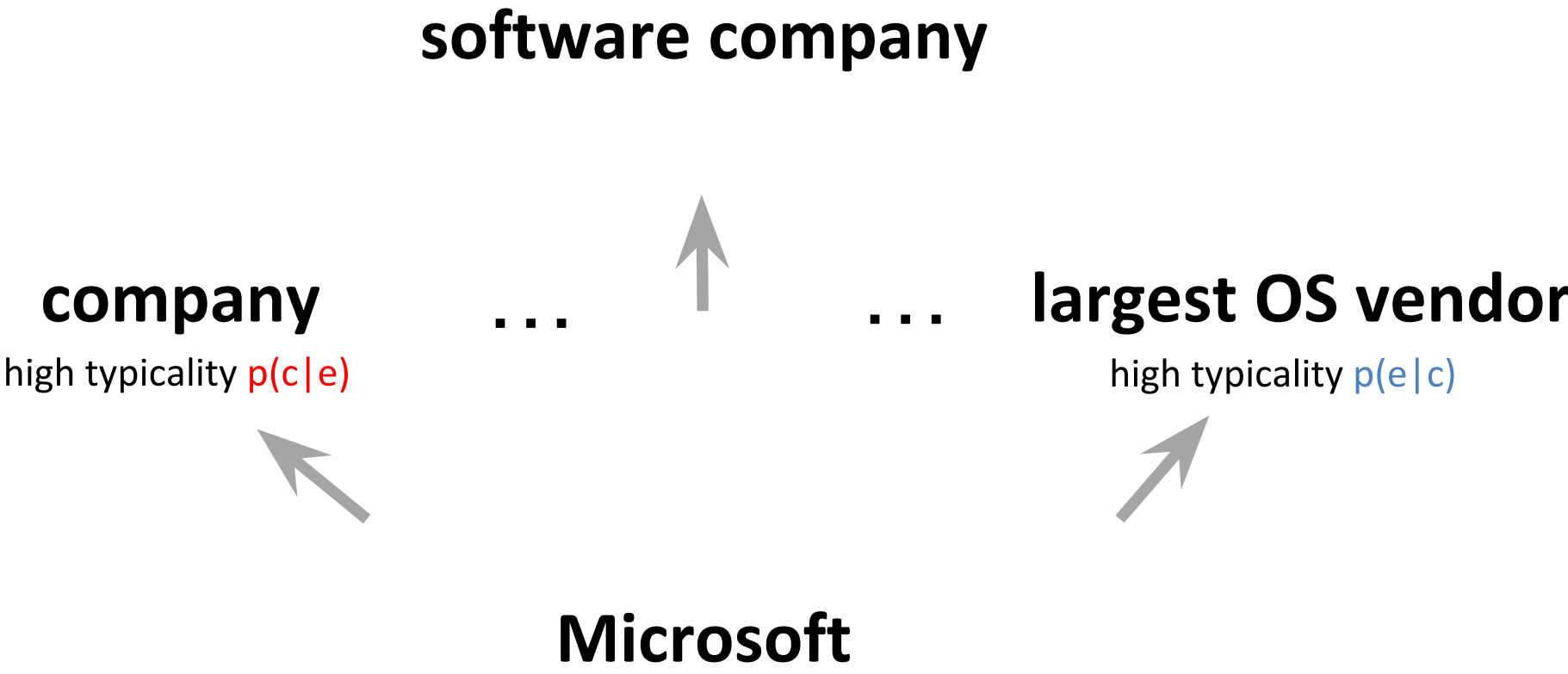
$$P(c|e) = \frac{n(c,e) + \alpha}{\sum_{e \in c_i} n(c_i,e) + \alpha N}$$

"robin" is a more *typical* bird than a "penguin" ➡ $p(robin|bird) > p(penguin|bird)$

BLC (basic level of categorization)

(In a vertical dimension) "… the most basic level of categorization will be the most inclusive (abstract) level at which the categories can mirror the structure of attributes perceived in the world."

BLC (basic level of categorization)



PMI

- $PMI(e, c) = \log \frac{P(e, c)}{P(e)P(c)} = \log P(e|c) - \log P(e)$

For given e, $\log P(e)$ is a constant.

PMI degenerates into log typicality.

NPMI (normalized PMI)

- $$NPMI(e,c) = \frac{PMI(e,c)}{-\log P(e,c)} = \frac{\log P(e|c) - \log P(e)}{-\log P(e,c)}$$

Still dominated by PMI, or typicality.

Our measure

$$R(e,c) = p(e|c) \cdot p(c|e)$$

$$\log R(e,c) = \log \frac{P(e,c)^2}{P(e)P(c)} = PMI(e,c) + \log P(e,c)$$

Another Explanation: Random Walk

- Compute *commute time* between *e* and *c*.
- It is the expected number of steps that a random walk starting at node *e*, going through *c* once, and returning to *e*.
- This converges to p(e|c)p(c|e)

Vagueness

key players
factors
items
things
reasons
…

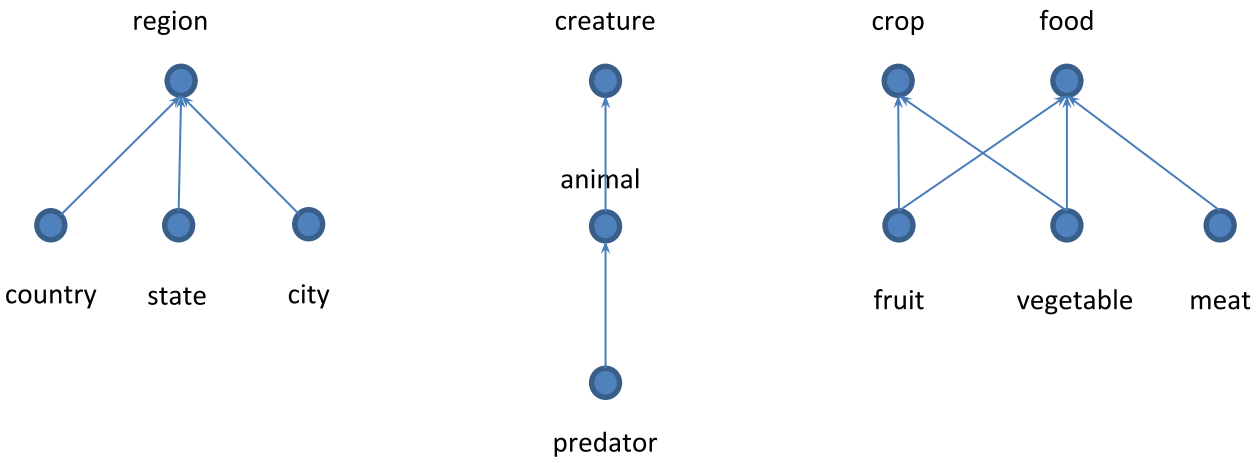
$$V(C) = \frac{|\{e_i | P(C|e_i) \geq c, \forall e_i \in C\}|}{N(C)}$$

(Do people whom you regard highly regard you highly?)

Ambiguity

- Probase defines 3 levels of ambiguity
  - Level 0 (1 sense): apple juice
  - Level 1 (2 or more related senses): Google
  - Level 2 (2 or more senses): python

- Concepts form clusters, clusters form senses (through isa relation)



Similarity

- microsoft, ibm → 0.933
- google, apple → 0.378 ??

$$sim(t_1, t_2) = \max_{x,y} cosine\,(c_x(t_1), c_y(t_2))$$

Many, many, applications

- Query Understanding
  - Head/Modifier/Constraint detection
- …
- SRL (semantic role labeling) with FrameNet
  - e.g. Tom broke the window.

Example: FrameNet

Frame: Apply_heat

FE1 **FE2** **FE3** **FE4**

She was FRYING eggs and bacon and mushrooms on a camp stove in Woolley 's billet .

Concept	P(c|FE)
heat source	0.19
Large metal	0.04
Kitchen appliance	0.02

Instance	P(w|FE)
Stove	0.00019
Radiator*	0.00015
Oven	0.00015
Grill*	0.00014
Heater*	0.00013
Fireplace*	0.00013
Lamp*	0.00013
Hair dryer*	0.00012
Candle*	0.00012

Vagueness

Representativeness indicates appropriate level of generality.

- a key player → High vagueness

$$V(C) = \frac{|\{e_i | P(C|e_i) \geq c, \forall e_i \in C\}|}{N(C)}$$

- a company → High typicality P(c|e)

$$P(c|e) = \frac{n(c,e) + \alpha}{\sum_{e \in c_i} n(c_i, e) + \alpha N}$$

Microsoft is …

- a software company → High representativeness

$$R_e(c) = P(c|e) \cdot P(e|c)$$

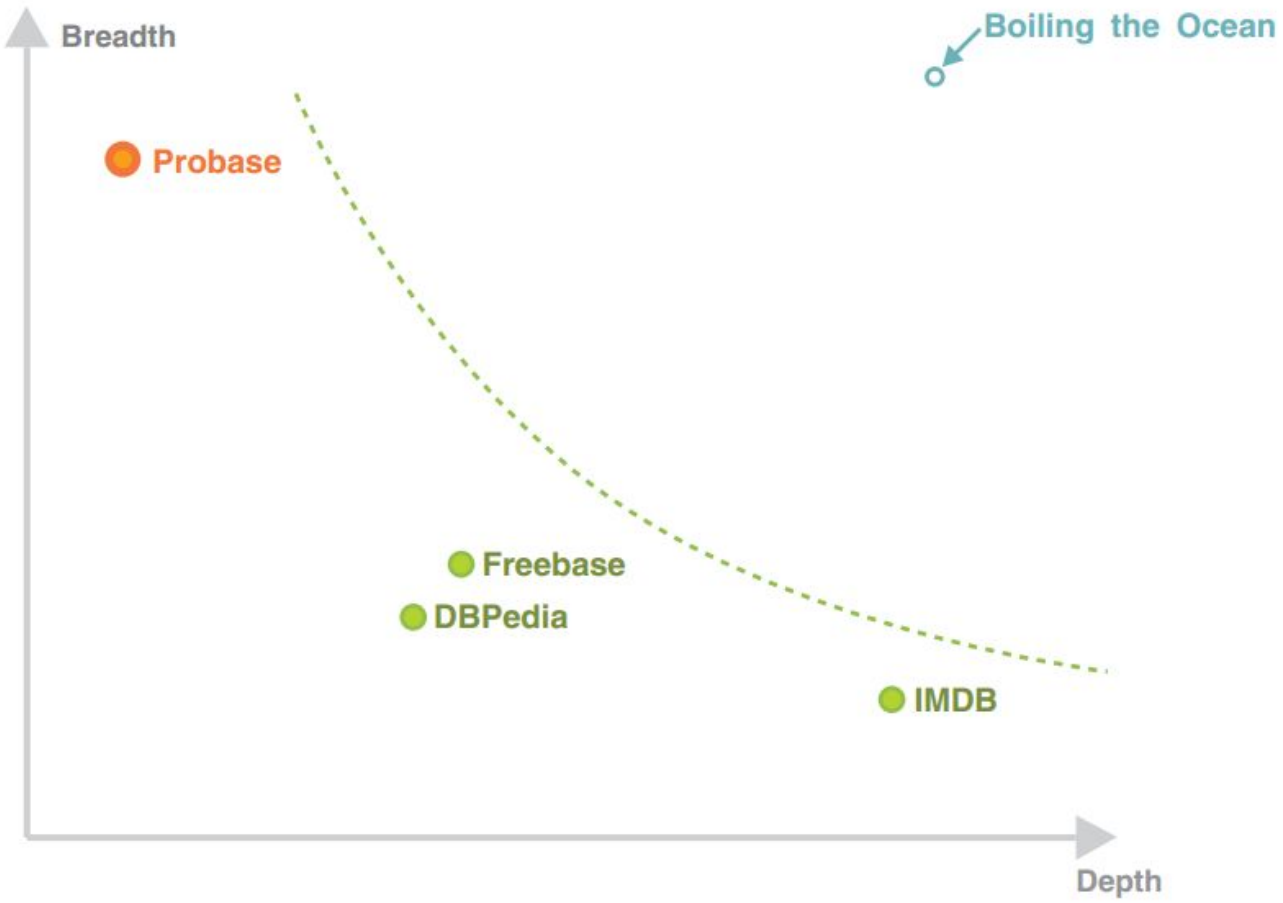
- a large seattle software firm → High typicality P(e|c)

$$P(e|c) = \frac{n(c,e) + \alpha}{\sum_{e_i \in c} n(c, e_i) + \alpha N}$$

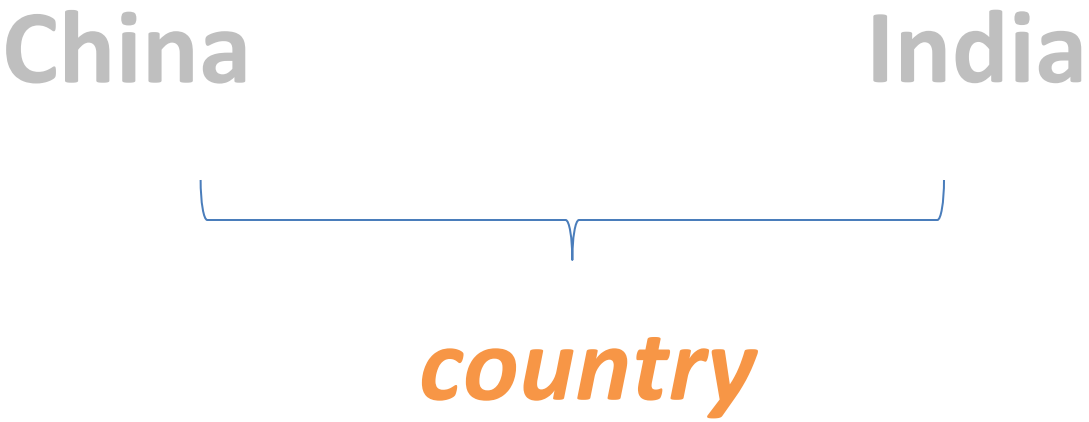
A concept is vague if it's not typical to its most typical instances

A robin is a more typical bird than a penguin.

What can Probase do?



Concept Learning



Concept Learning

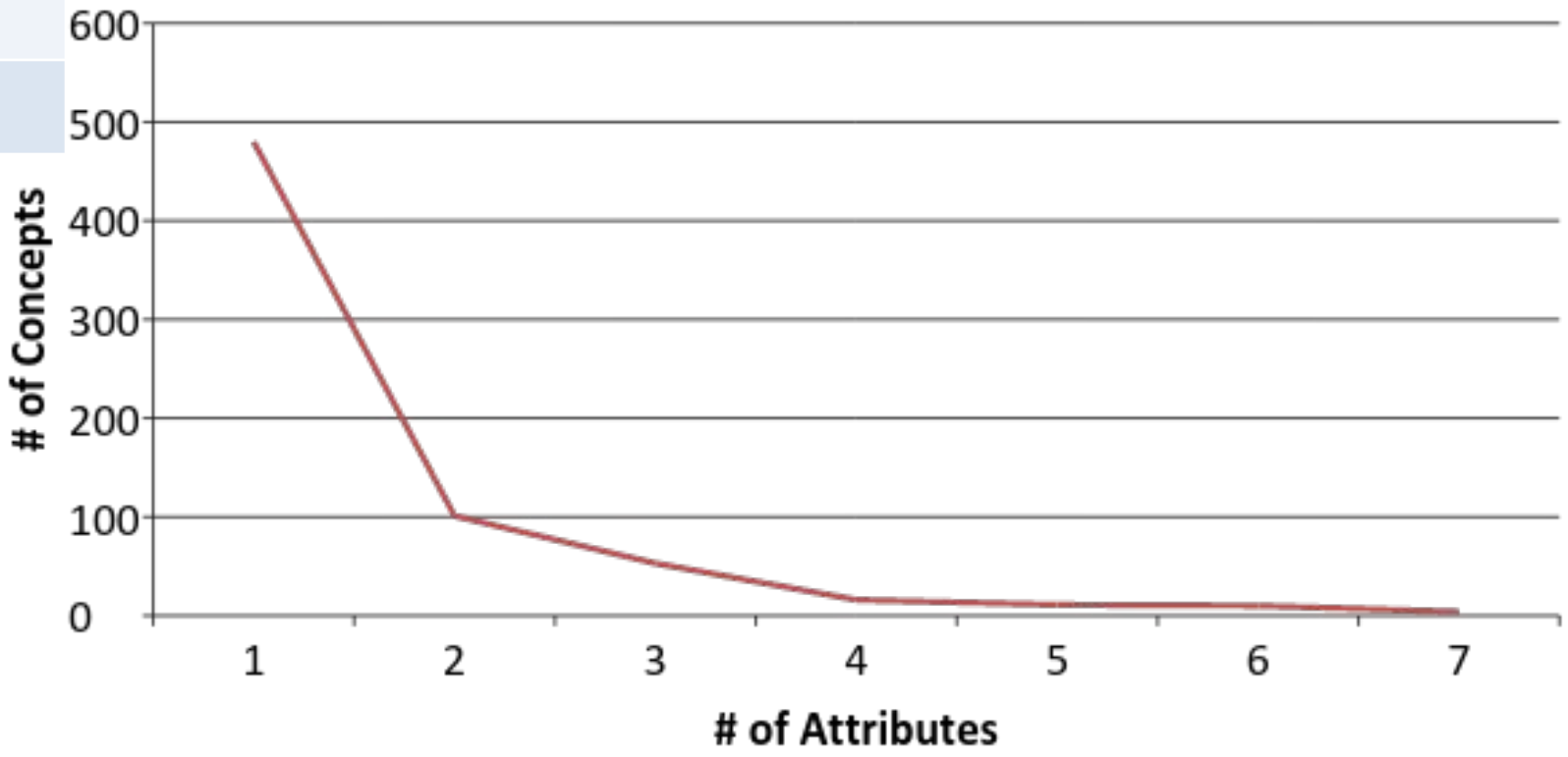
China Brazil India

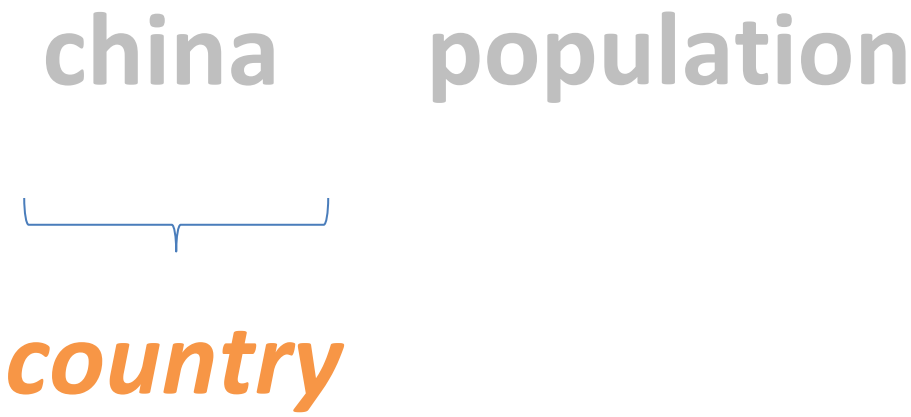
emerging market


body
smell
taste
wine

Understanding Web Tables

	website	president	city	motto	state	type	director




china
population
country

collector of fine china

earthenware

Bayesian

- $$P(c_k|E) = \frac{P(E|c_k)P(c_k)}{P(E)} \propto P(c_k)\prod_{i=1}^{M} P(e_i|c_k).$$

- For a mixture of instances and properties: Noisy-Or model

$$P(c|t_l) = 1 - \big(1 - P(c|t_l, z_l = 1)\big)\big(1 - P(c|t_l, z_l = 0)\big)$$

where $z_l = 1$ indicates t_l is an entity, $z_l = 0$ indicates t_l is a property

- Bayesian rule gives:

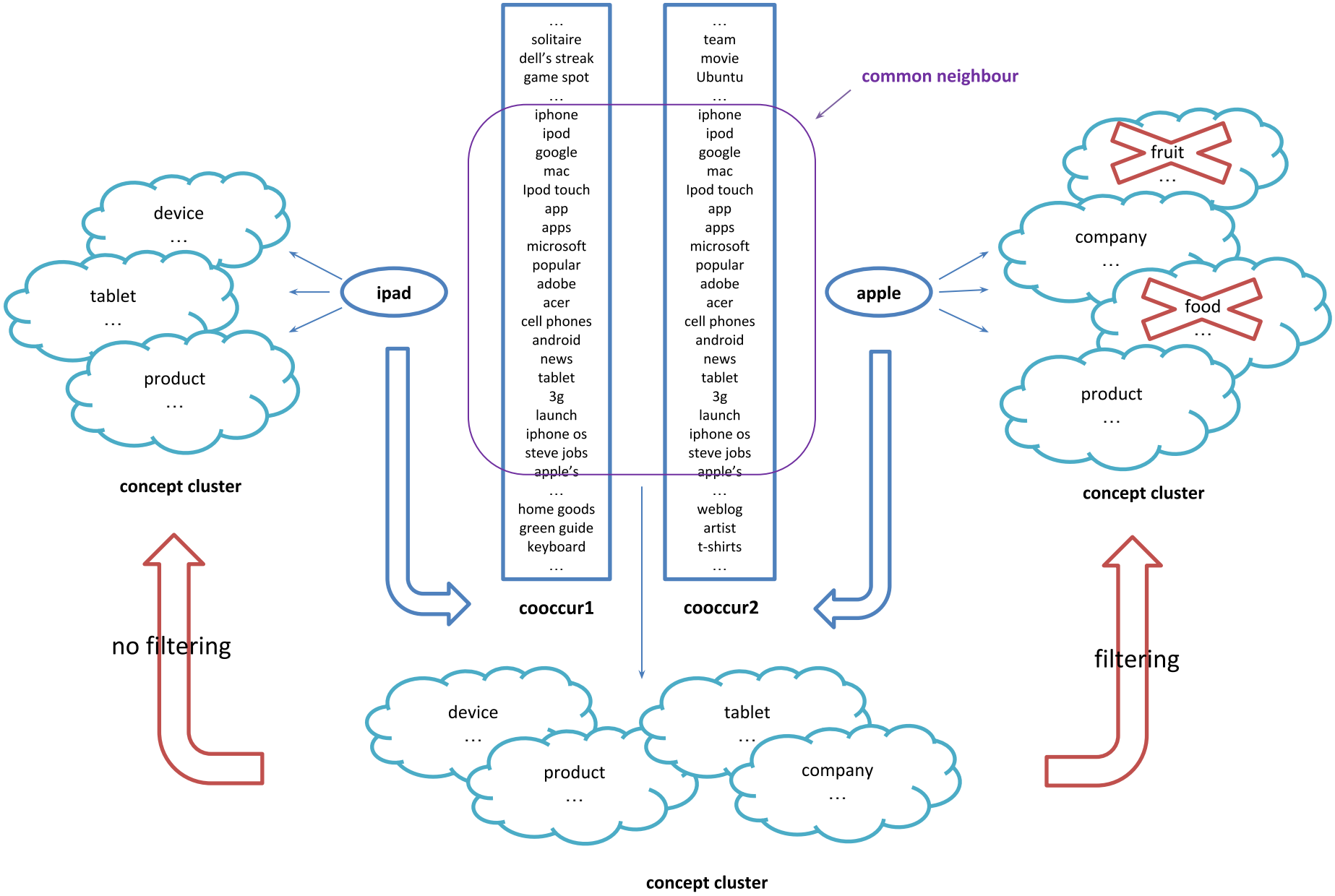
$$P(c|T) \propto P(c)\prod_l^L P(t_l|c) \propto \frac{\prod_l P(c|t_l)}{P(c)^{L-1}}$$

apple

company

iPad

device

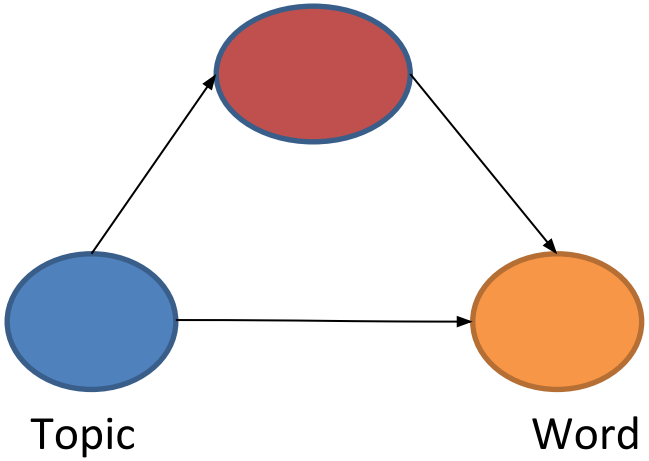

…
solitaire
dell's streak
game spot
…
iphone
ipod
google
mac
Ipod touch
app
apps
microsoft
popular
adobe
acer
cell phones
android
news
tablet
3g
launch
iphone os
steve jobs
apple's
…
home goods
green guide
keyboard
…
cooccur1
…
team
movie
Ubuntu
…
iphone
ipod
google
mac
Ipod touch
app
apps
microsoft
popular
adobe
acer
cell phones
android
news
tablet
3g
launch
iphone os
steve jobs
apple's
…
weblog
artist
t-shirts
…
cooccur2
common neighbour
ipad
device
…
tablet
…
product
…
concept cluster
apple
fruit
…
company
…
food
…
product
…
concept cluster
no filtering
filtering
device
…
product
…
tablet
…
company
…
concept cluster

Modeling Co-occurrence

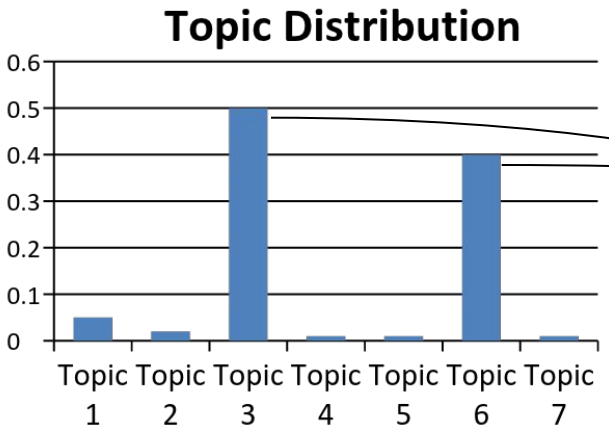
Probase

\+

Wikipedia



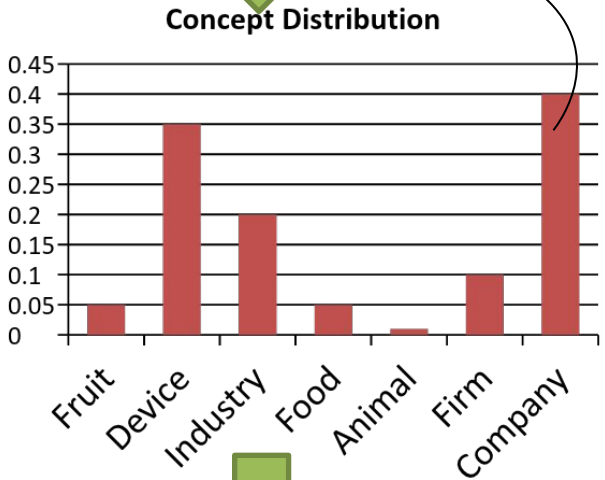
LDA model



Topic 3	Prob
software	0.0260
windows	0.0224
system	0.0184
version	0.0175
file	0.0172
user	0.0141
support	0.0115
microsoft	0.0114
os	0.0098
computer	0.0097
based	0.0089
available	0.0088
mac	0.0085
source	0.0081
linux	0.0079
operating	0.0077
open	0.0073
released	0.0072
server	0.0069
release	0.0066

Topic 6	Prob
company	0.1068
business	0.0454
companies	0.0186
inc	0.0167
corporation	0.0139
market	0.0138
founded	0.0136
based	0.0136
sold	0.0132
industry	0.0127
products	0.0126
firm	0.0125
group	0.0124
owned	0.0112
first	0.0111
largest	0.0101
management	0.0091
new	0.009
million	0.009
acquired	0.0085

$p(c|t)$



$p(w|topic)$

Company	Prob
Apple	0.214123
Google	0.122754
Microsoft	0.089717
affiliate company	0.073715
Sony	0.048214
ISP	0.038612
Internet Service Provider	0.036651
web host	0.036341
Nintendo	0.034999
HP	0.033760
Blizzard	0.031798
Toyota	0.028598

$p(w|c)$

$p(w|c)$

Apple, iPhone

$p(w|topic)$

- Infer topics z from text s using collapsed Gibbs sampling:

$$p(z_i = k|\vec{s}, z_{-i}, C) \propto (n_{.k} + \alpha) \times \frac{C_{s_i k} + n_{s_i k} + \beta}{\sum_w C_{wk} + n_{wk} + |W|\beta}$$

- Estimate the concept distribution for each term w in s:

$$p(c|w, z) \propto p(c|w) \sum_k \pi_{wk} \phi_{ck},$$

$$\phi_{ck} = \frac{C_{ck} + \beta}{\sum_w C_{wk} + |W|\beta},$$

Examples

ShortText: fox fur | Conceptualize

Show Parameters

Elapsed Time = 00:00:00.2360236

fox

[159/wild animal/pet/animal][v]

159/wild animal/pet/animal	0.5956765
wild animal	0.0169223
feral animal	0.01490341
introduced animal	0.01263432
pest animal	0.01216037
small animal	0.01138677
nocturnal animal	0.01060585
native animal	0.01022427
predatory animal	0.009197926
animal	0.008580011
large animal	0.007967799
wild animal	0.004003763
feral animal	0.003526101
introduced animal	0.00298924
pest animal	0.002877105
small animal	0.002694075
nocturnal animal	0.002509312
native animal	0.002419031
predatory animal	0.002176201

fur

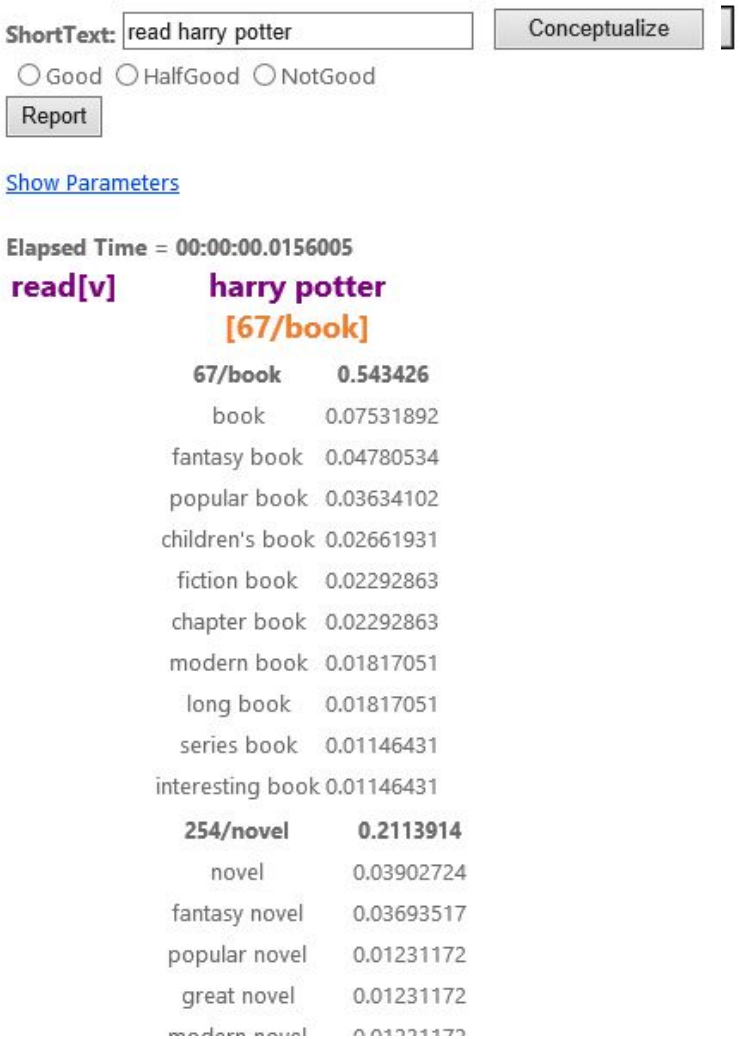
[4/texture/material][v]

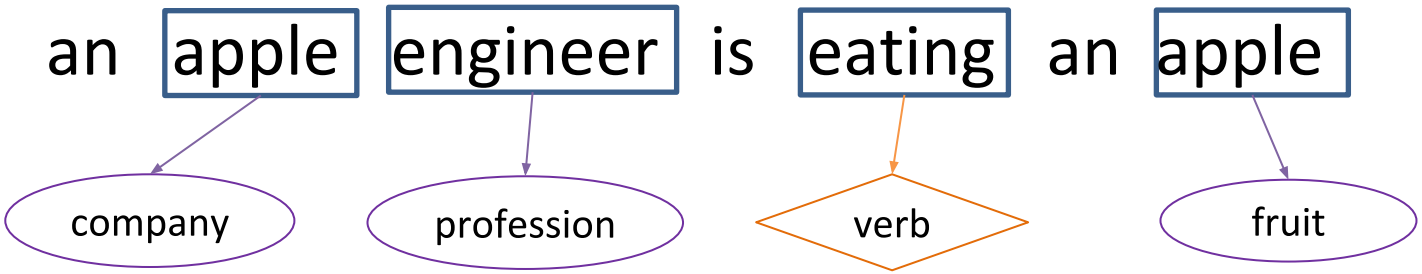
4/texture/material	0.2107609
texture	0.01112421
organic material	0.007871442
soft material	0.007446955
luxury material	0.007329956
luxurious material	0.007232076
raw material	0.006870993
natural material	0.006293016
real world surface	0.00589916
locally available raw material	0.005892543
dead material	0.005889004
electronic product	0.01436949
electronic good	0.01051342
high-tech product	0.009497663
electrical good	0.006462679
consumer electronic product	0.006424694
electrical product	0.006299805
consumer product	0.005079063
range electrical product	0.004229661

c/channel/network]

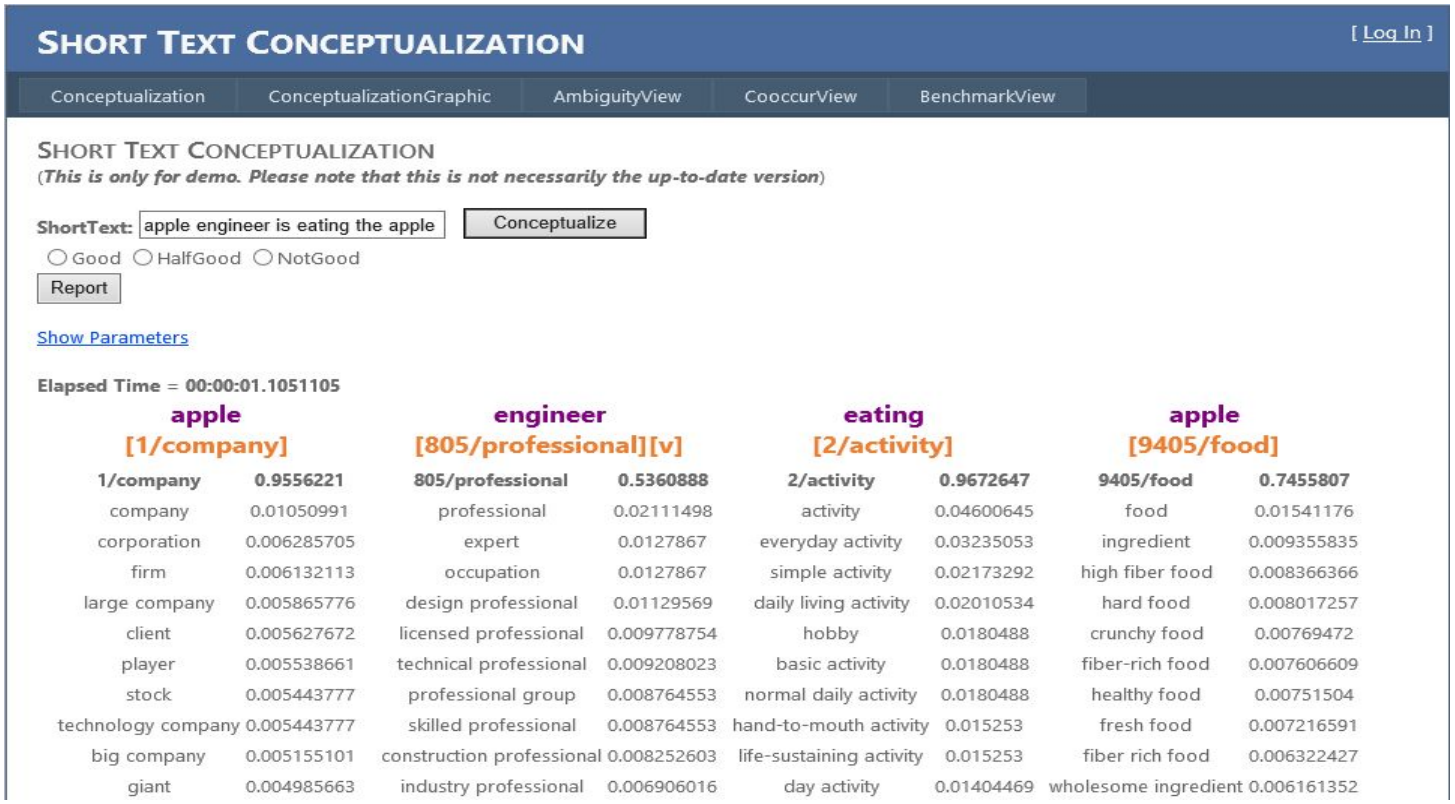
nel/network	0.6562241
:	0.1072035
	0.0970483
	0.06378444
	0.05830856
c	0.0403064
	0.0391444
	0.03133982
	0.0295761
	0.02876115
	0.02717765

Examples


ShortText: read harry potter
Conceptualize
Good HalfGood NotGood
Report
Show Parameters
Elapsed Time = 00:00:00.0156005
read[v]
harry potter
[67/book]
67/book 0.543426
book 0.07531892
fantasy book 0.04780534
popular book 0.03634102
children's book 0.02661931
fiction book 0.02292863
chapter book 0.02292863
modern book 0.01817051
long book 0.01817051
series book 0.01146431
interesting book 0.01146431
254/novel 0.2113914
novel 0.03902724
fantasy novel 0.03693517
popular novel 0.01231172
great novel 0.01231172


an
apple
engineer
is
eating
an
apple
company
profession
verb
fruit

Examples



SHORT TEXT CONCEPTUALIZATION [Log In]

Conceptualization | ConceptualizationGraphic | AmbiguityView | CooccurView | BenchmarkView

SHORT TEXT CONCEPTUALIZATION

(*This is only for demo. Please note that this is not necessarily the up-to-date version*)

ShortText: apple engineer is eating the apple [Conceptualize]

○ Good ○ HalfGood ○ NotGood

[Report]

Show Parameters

Elapsed Time = 00:00:01.1051105

apple [1/company]		engineer [805/professional][v]		eating [2/activity]		apple [9405/food]	
1/company	**0.9556221**	**805/professional**	**0.5360888**	**2/activity**	**0.9672647**	**9405/food**	**0.7455807**
company	0.01050991	professional	0.02111498	activity	0.04600645	food	0.01541176
corporation	0.006285705	expert	0.0127867	everyday activity	0.03235053	ingredient	0.009355835
firm	0.006132113	occupation	0.0127867	simple activity	0.02173292	high fiber food	0.008366366
large company	0.005865776	design professional	0.01129569	daily living activity	0.02010534	hard food	0.008017257
client	0.005627672	licensed professional	0.009778754	hobby	0.0180488	crunchy food	0.00769472
player	0.005538661	technical professional	0.009208023	basic activity	0.0180488	fiber-rich food	0.007606609
stock	0.005443777	professional group	0.008764553	normal daily activity	0.0180488	healthy food	0.00751504
technology company	0.005443777	skilled professional	0.008764553	hand-to-mouth activity	0.015253	fresh food	0.007216591
big company	0.005155101	construction professional	0.008252603	life-sustaining activity	0.015253	fiber rich food	0.006322427
giant	0.004985663	industry professional	0.006906016	day activity	0.01404469	wholesome ingredient	0.006161352

Bayesian inference allows a three-year-old to learn the concept of horse after seeing merely three pictures of horses.

But such inference may rely on *innate priors* that are hardwired into our brain through eons of evolution.

Example : Entity Linking

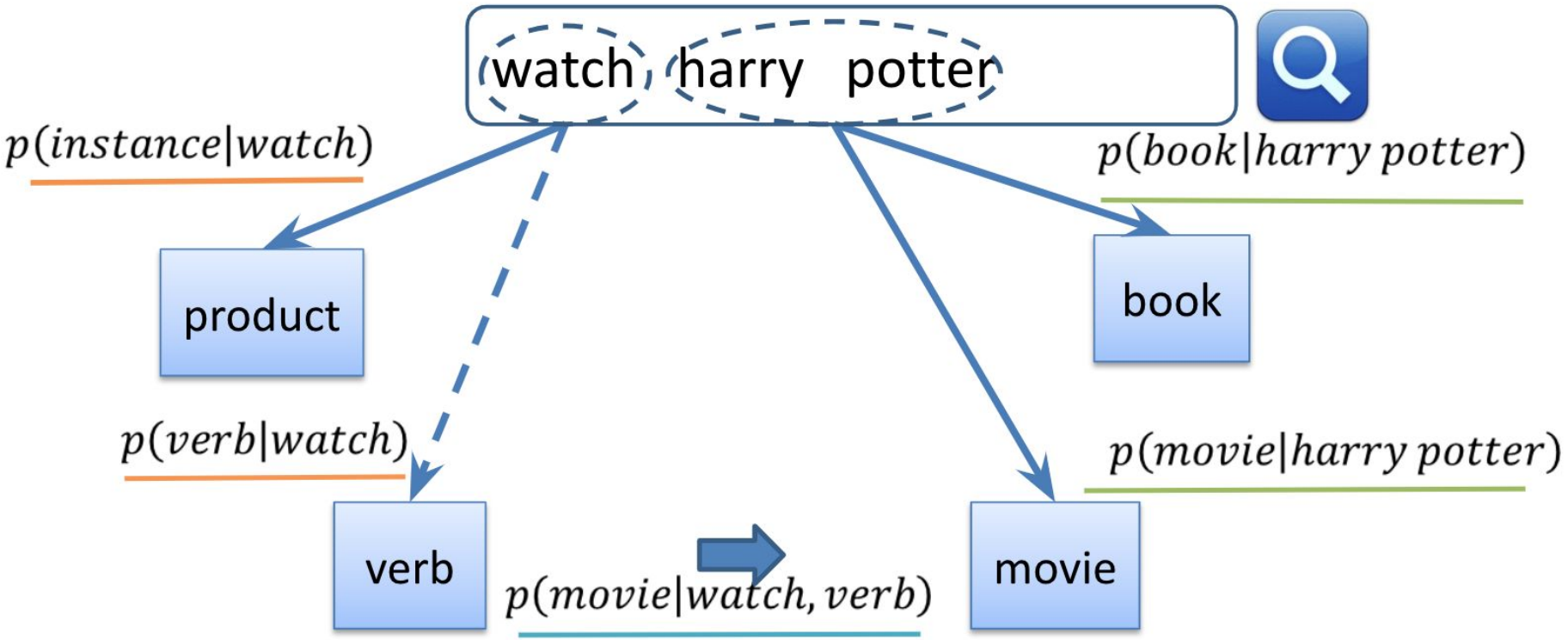
jordan 7 day weather forecast

P( |7 day weather forecast) $\propto$ P(7 day weather forecast | ) P()

P( |7 day weather forecast) $\propto$ P(7 day weather forecast | ) P()

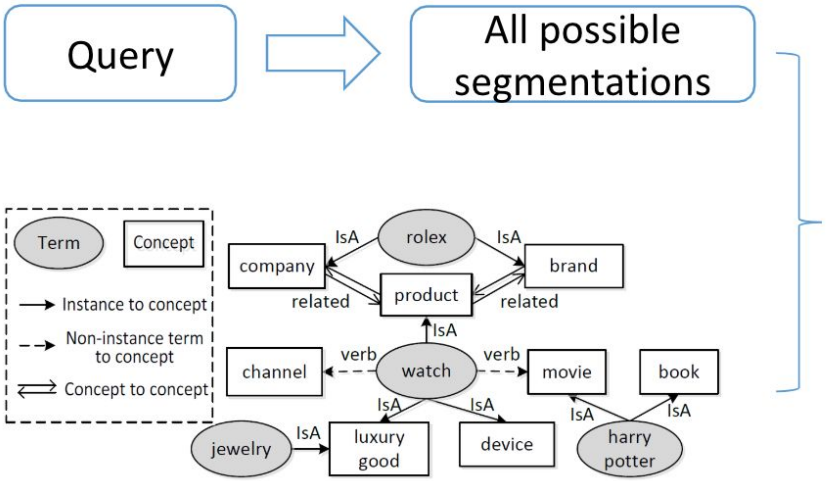
Understanding Short Text

[Wang et al. 2015b]

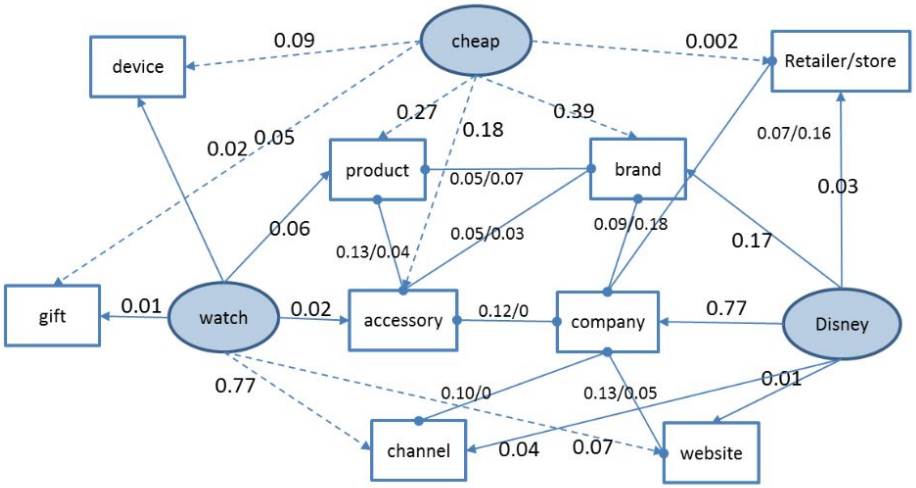


Understanding Short Text

$$\arg\max_{c} p(c|t,q)$$

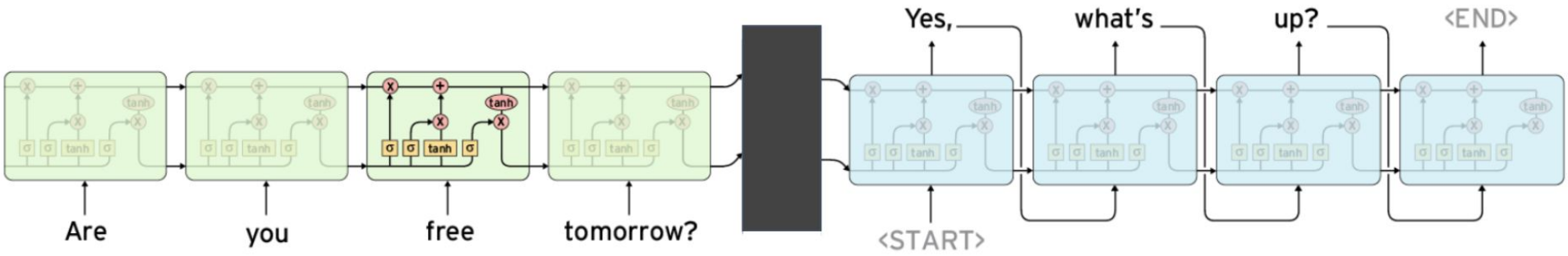


The offline semantic network



Random walk with restart [Sun et al., 2005] on the online subgraph

Implicit Approaches

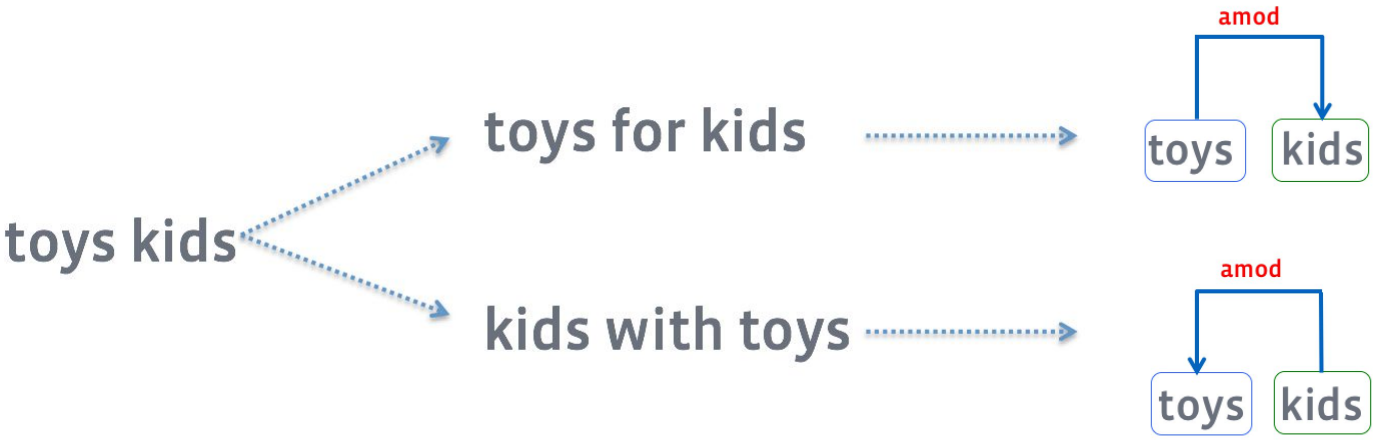


Life is like a **black box** of chocolates.

Structure

- red wine $40
- flight from JFK to SFO
- travel in arizona
- jordan 7 day weather forecast

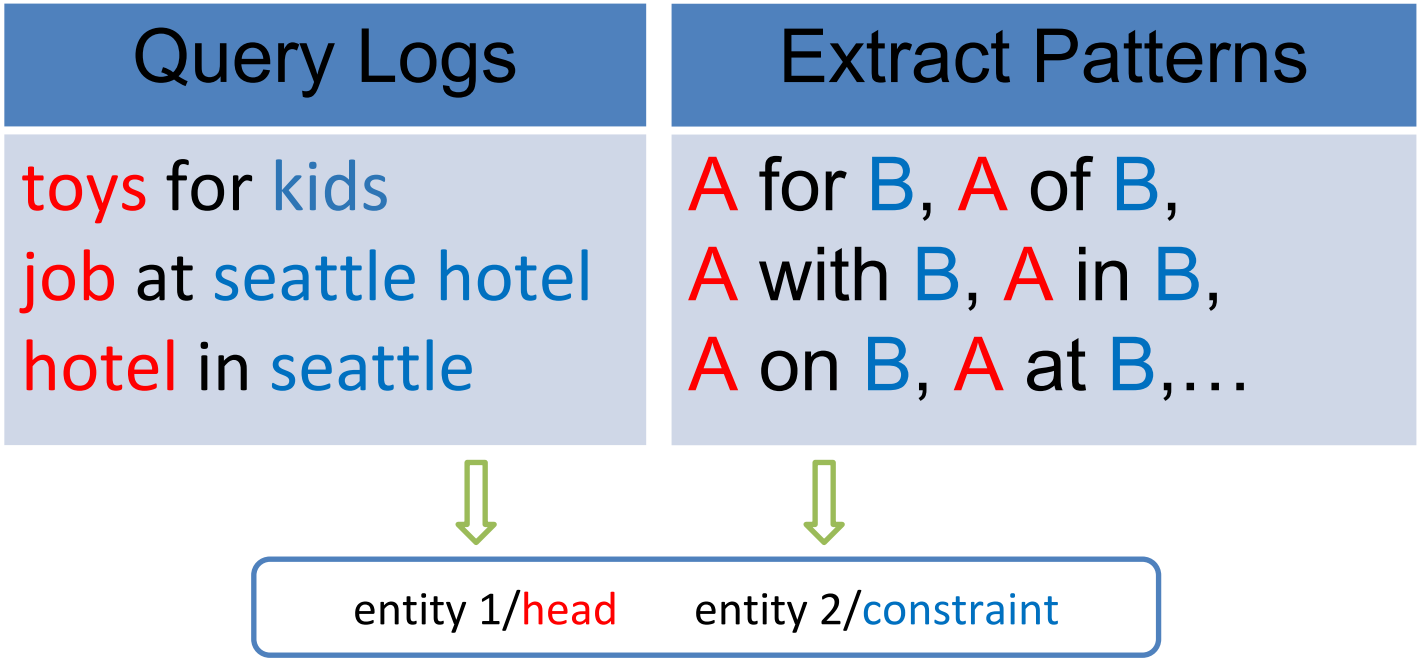
Structure in NL



Head, Modifier, and Constraint

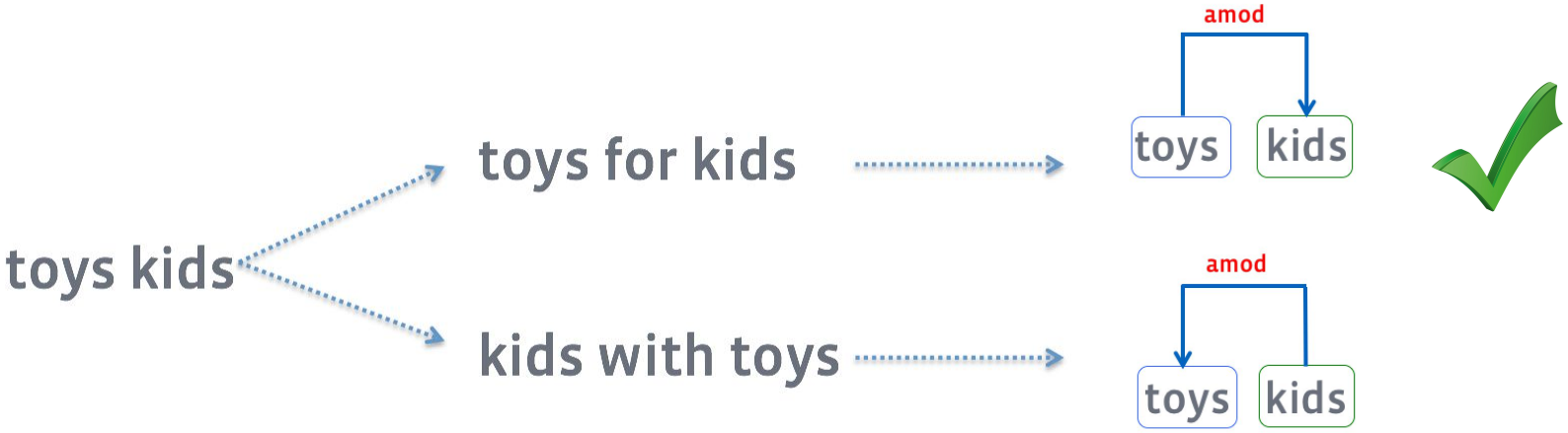
- Example: “luxury seattle hotel”, “seattle hotel job”
- ***Head:*** **intent/category** of the short text: “hotel”, “job”
- ***Constraints:*** **distinguish** this member from other members of the same category: “seattle” in “seattle hotel”
- ***Modifiers:*** **subjective:** “luxury”

Acquiring Patterns from Search Log

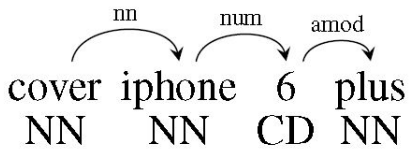


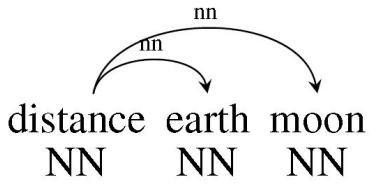
Performing conceptual analysis to “store” patterns at conceptual level

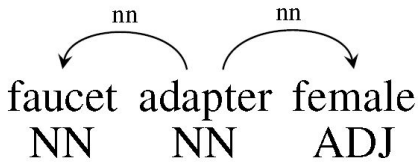
Structure


toys kids
toys for kids
kids with toys
amod
toys
kids
amod
toys
kids

Syntactic Parsing of Short Texts (EMNLP 2017)

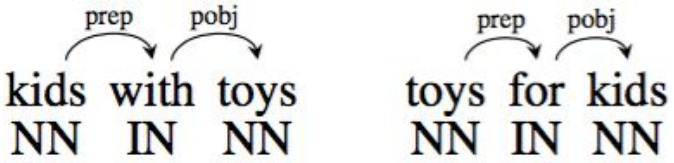






Challenges: Insufficient Grammatical Signals

- "toys kids" has ambiguous intent



- "distance earth moon" has clear intent
  - many equivalent forms: "earth moon distance", "earth distance moon", …

Challenges: Syntactic Parsing of Queries

- No standard
- No ground-truth

Why is syntactic parsing of queries even a legitimate problem?

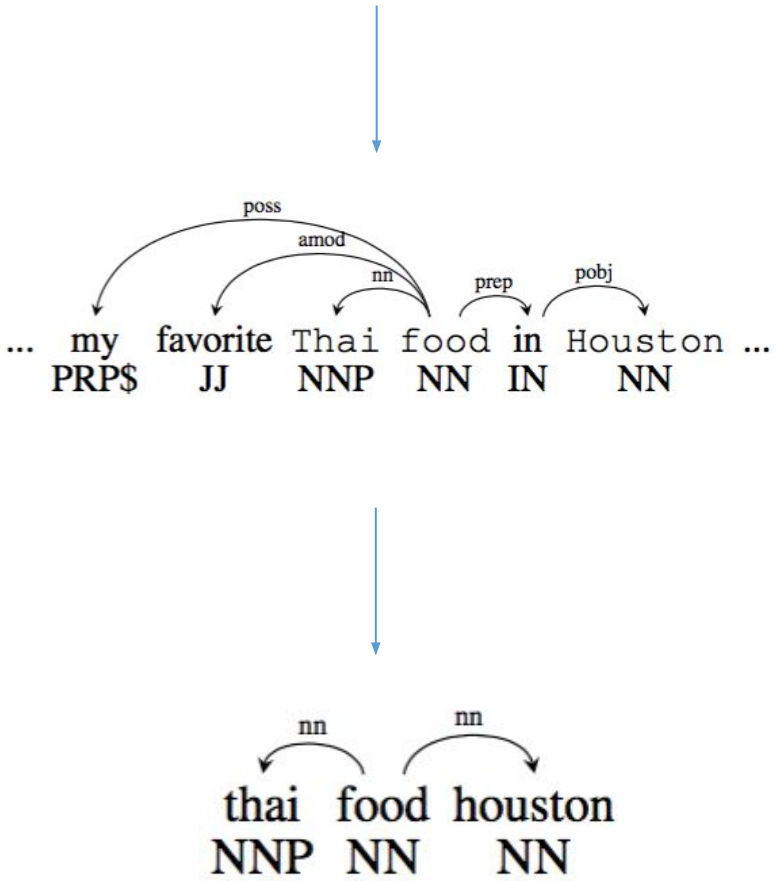
Our mind converts bag of words to meaning

“thai food houston”

… my favorite Thai food in Houston is ...

Your imagination is good but too costly

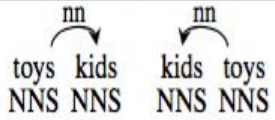
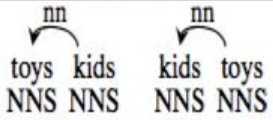
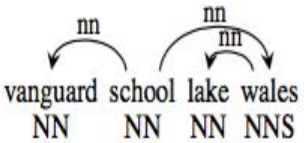
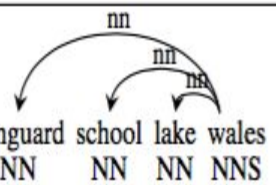
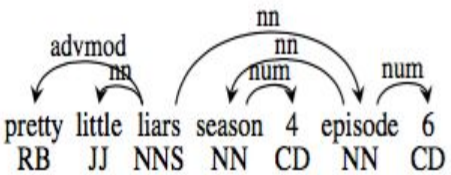
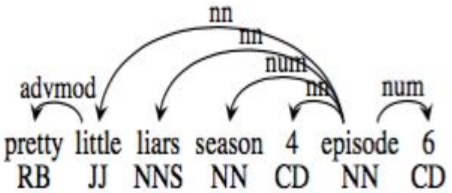
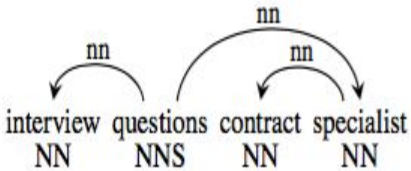
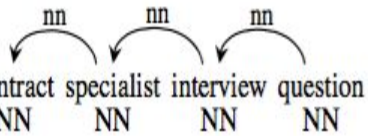
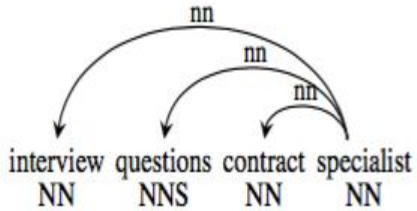
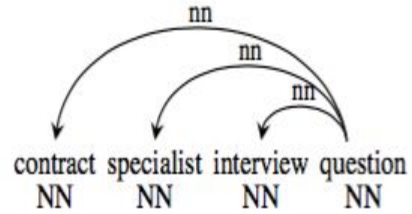
"thai food houston"



A Treebank for Short Texts

- Given query q
- Given q's clicked sentence $\{s\}$
- Parse each s
- **Project dependency from s to q**
- Aggregate dependencies

Result Examples

QueryParser	Stanford parser
	
	
	
 	 

Results

- Random queries:

  QueryParser UAS: 0.83, LAS: 0.75
  Stanford UAS: 0.72, LAS: 0.64

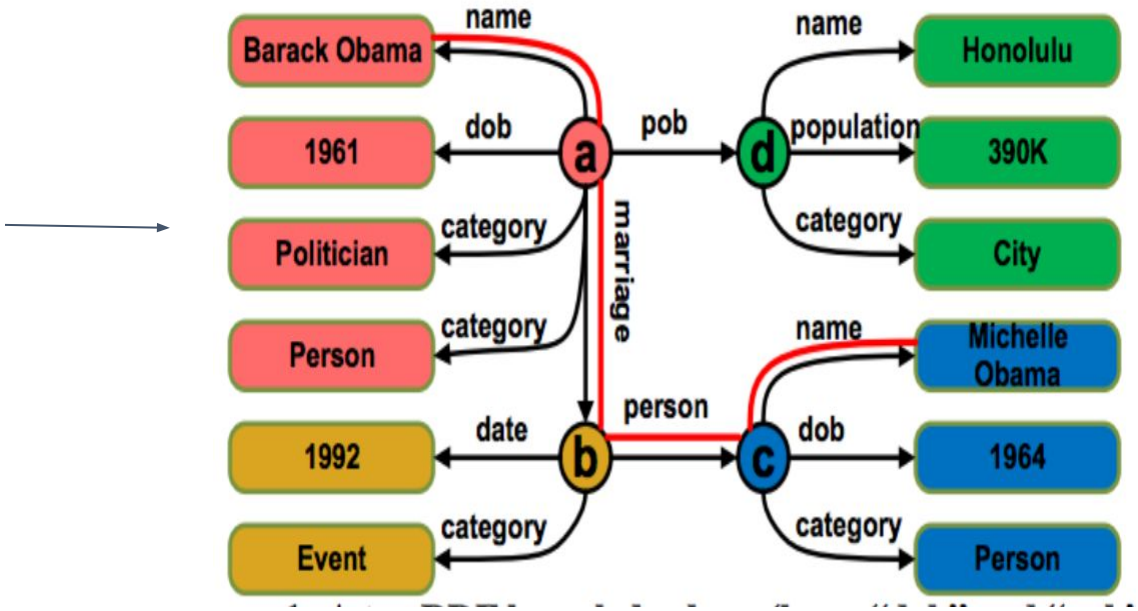
- Queries with no function words:

  QueryParser UAS: 0.82, LAS: 0.73
  Stanford UAS: 0.70, LAS: 0.61

- Queries with function words:

  QueryParser UAS: 0.90, LAS: 0.85
  Stanford UAS: 0.86, LAS: 0.80


SALUT!
GRÜSS DICH!
HELLO!
CIAO!
HEJ!
SALUTON!
SELAM!
¡HOLA!
AMOJ!


Barack Obama
name
1961
dob
Politician
category
Person
category
1992
date
Event
category
a
b
marriage
pob
person
d
c
Honolulu
name
population
390K
City
category
Michelle Obama
name
1964
dob
Person
category

Semantic Parsing

- What is the longest river in the smallest state?
- What is the average salary of the smallest department?
- How many people live in Menlo Park?

Operation (SELECT)	**Filtering (WHERE)**
SELECT column_name(s)	WHERE column_name <ops> value
SELECT COUNT(column_name)	WHERE COUNT(column_name) <ops> value
SELECT AVG(column_name)	WHERE column_name1 <ops1> value1 AND/OR column_name2 <ops2> value2
SELECT MAX(column_name)	WHERE column_name <ops> (SELECT ...)
...	...

Step 1: Separating Logic and Language

Q_1 : what's the **age** of *john smith*?
Q_2 : what's **employee**'s average **age**?

T_1 : what's the **age**:`c0` of *john smith*:`v1`?
T_2 : what's **employee**:`c0`'s average **age**:`c1` ?

T'_1 : what's the **area**:`c0` of *south america*:`v1`?
T'_2 : what's average **size**:`c0` of a **country**:`c1` ?

Q'_1 : what's the **area** of *south america*?
Q'_2 : what's average **size** of a **country**?

L_1 : select `c0` where `c1` = `v1`
L_2 : select avg(`c1`)

L'_1 : select `c0` where `c1` = `v1`
L'_2 : select avg(`c1`)

Step 2: Extend Schema to Include NL Signals

We do not have an all-inclusive Ontology that teaches us to how to talk about everything in the world.

If it's the first time to see "population of <city>", NL signals inform us it is equivalent to "number of people living in <city>"

If it's the first time to see stock price, NL signals may inform us people may say the "price soars above $$$"

Data Augmentation

- Bootstrap by creating a small set of examples that include all the NL nuances of talking about the new things
- Automatically generating a set of queries that including such NL nuances
- Using generated queries as new training dataset to update the model

Data Augmentation

- Reverse seq2seq training: from SQL to NL
- Manipulate SQL to generate more NL
- Use generated NL for forward seq2seq training

Result

	BASKETBALL	RESTAURANTS	CALENDAR	HOUSING	RECIPES
In-domain Performance					
(Su and Yan, 2017) (ES+I)	.780	.676	.556	.592	.857
Ours (In-domain)	**.816**	**.872**	**.738**	**.642**	**.878**
Cross-domain Performance					
(Su and Yan, 2017) (ES+X+I)	.837	.773	.788	.746	.877
Ours (In-domain + Out-domain)	**.848**	**.890**	**.825**	**.793**	**.887**

Thanks

https://medium.com/@haixun

- An Annotated Reading List of Conversational AI
- Getting NLP Ready for Business