Microsoft Research FacultySummit

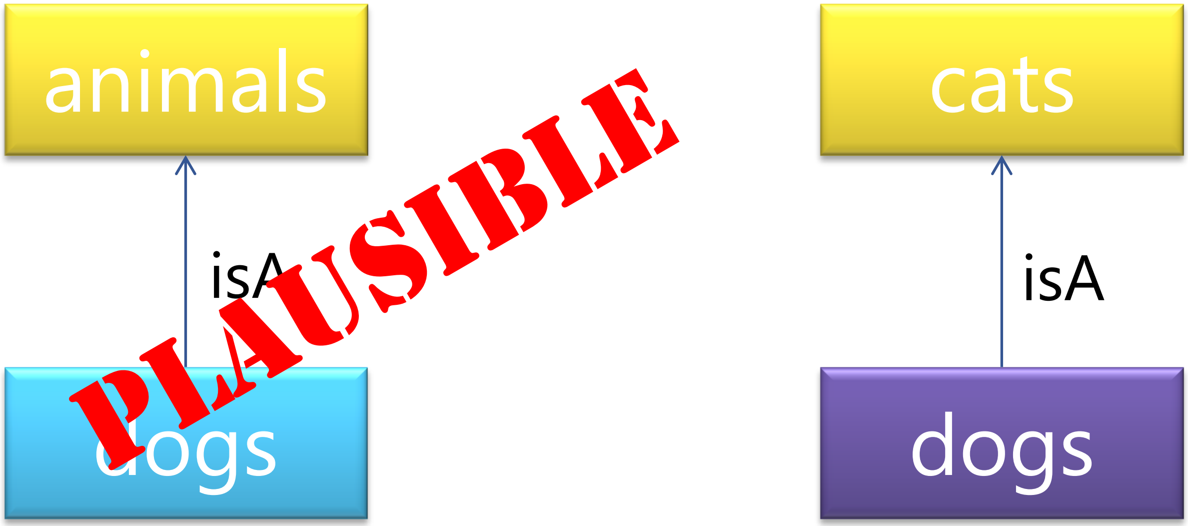
Probase : A Knowledge Base for Text Understanding

Haixun Wang
Microsoft Corporation

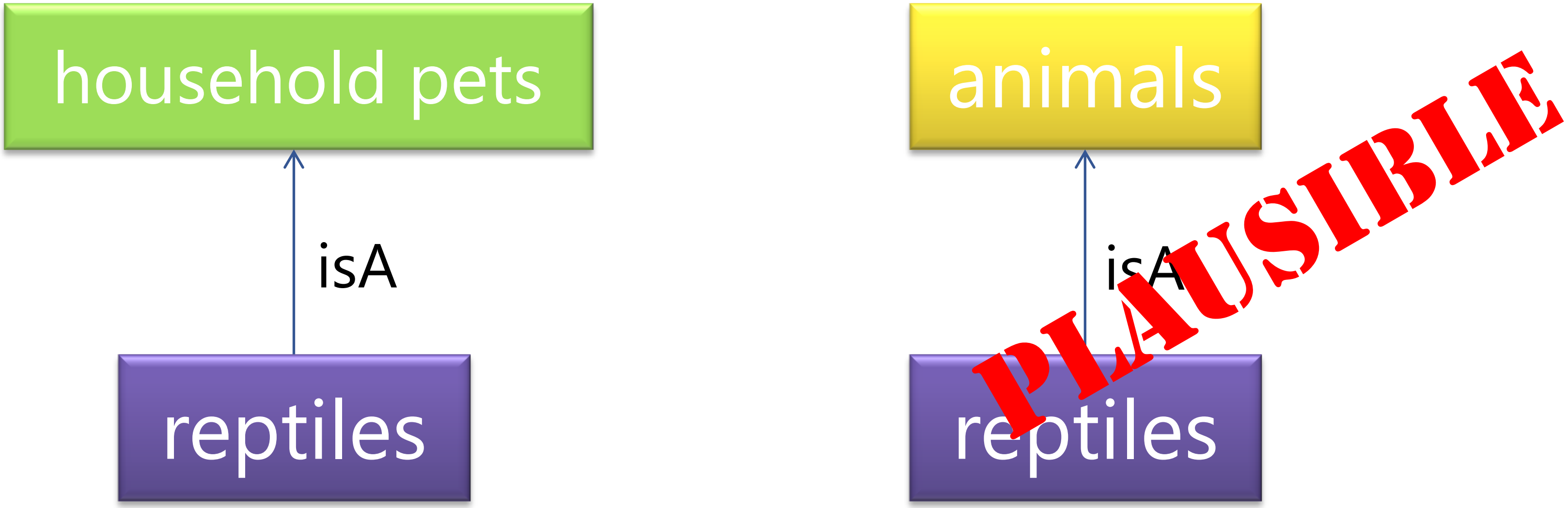
FUTURE WORLD 2011 2031

Pablo Picasso	25 Oct 1881	Spanish

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



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



ProBase

1 More than 2.7 million concepts automatically harnessed from 1.68 billion documents

2 Computation/Reasoning enabled by scoring:

Consensus:
e.g., is there a company called Apple?

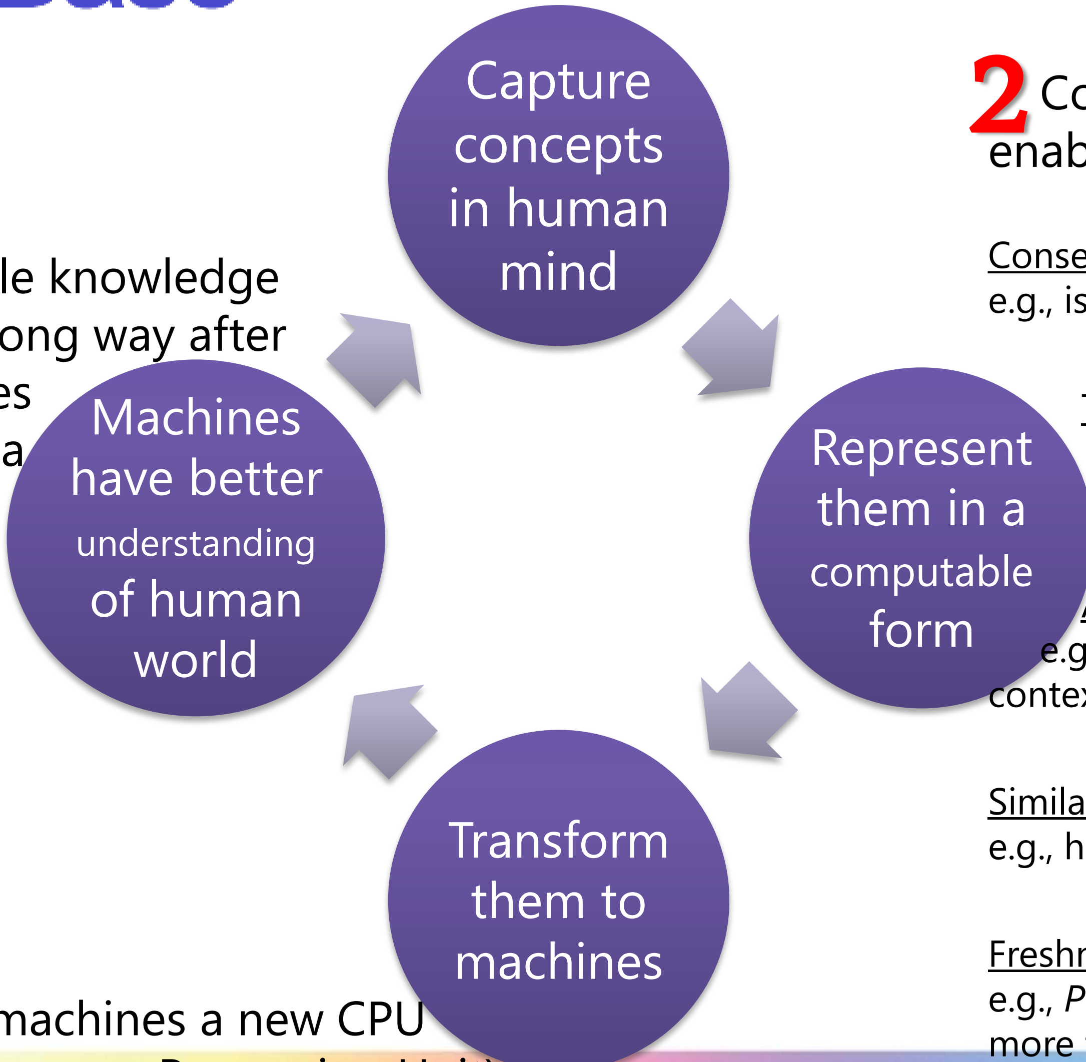
Typicality:
e.g. how likely you think of Apple when you think about companies?

Ambiguity:
e.g., does the word *Apple*, sans any context, represent *Apple the company*?

Similarity:
e.g., how likely is an actor also a celebrity?

Freshness:
e.g., *Pluto as a dwarf planet* is a claim more fresh than *Pluto as a planet*.

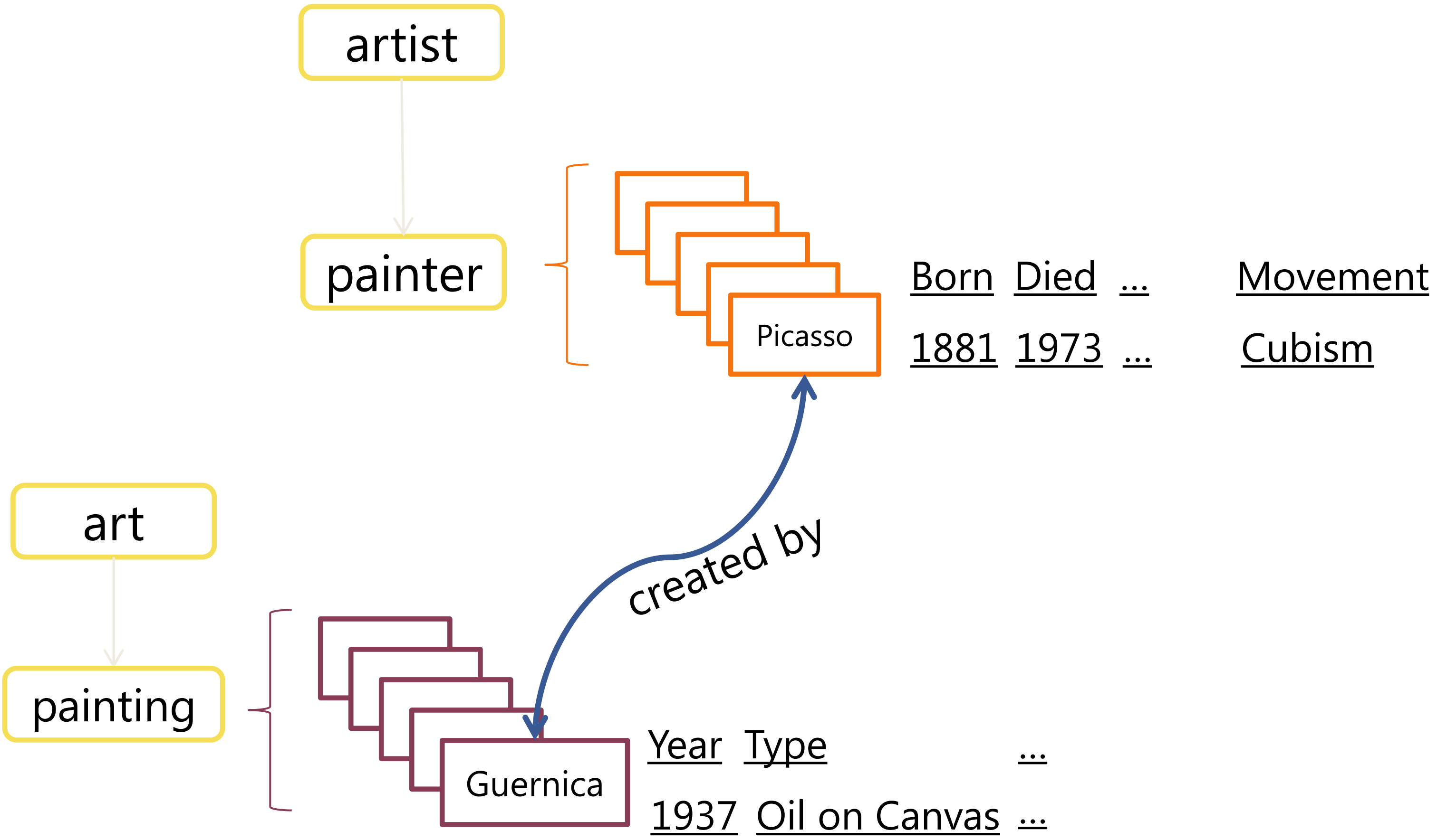
…



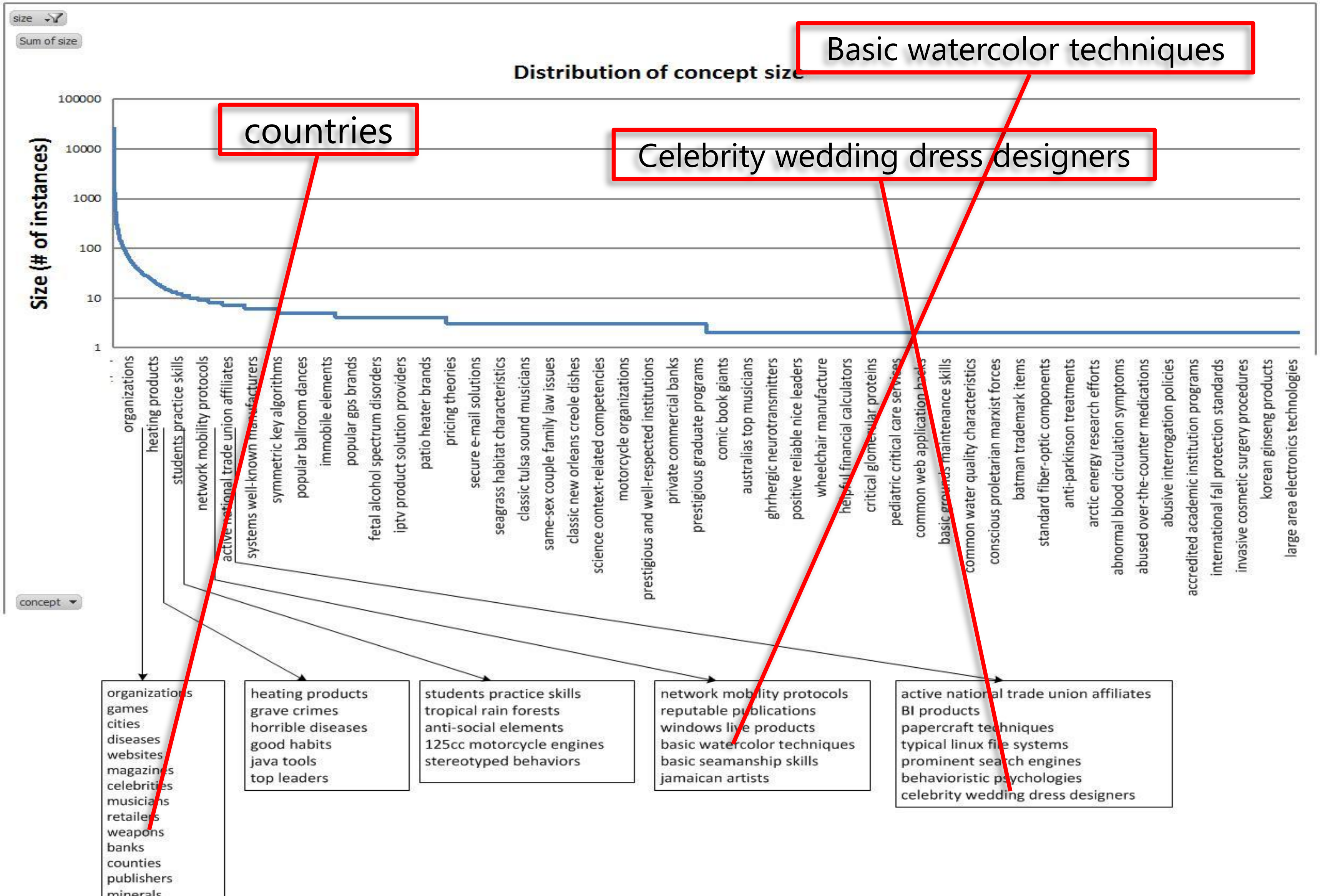
3 Give machines a new CPU (Commonsense Processing Unit) powered by a distributed graph engine called Trinity.

4 A little knowledge goes a long way after machines acquire a human touch

Probase Internals



2.7 million concepts



Data Sources

- Patterns for single statements

  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

- Examples:
  - Good: “rich countries such as USA and Japan …”
  - Tough: “animals other than *cats* such as *dogs* …”
  - Hopeless: “At Berklee, I was playing with *cats such as Jeff Berlin*, *Mike Stern*, *Bill Frisell*, and *Neil Stubenhaus.”*

Properties

- Given a class, find its properties
- Candidate seed properties:
  - "What is the [property] of [instance]?"
  - "Where", "When", "Who" are also considered

Similarity between two concepts

- Weighted linear combinations of
  - Similarity between the set of instances
  - Similarity between the set of attributes
- (nation, country)
- (celebrity, well-known politicians)

Beyond noun phrases

- Example: the verb “hit”
  - Small object, Hard surface
    - (bullet, concrete), (ball, wall)
  - Natural disaster, Area
    - (earthquake, Seattle), (Hurricane Floyd, Florida)
  - Emergency, Country
    - (economic crisis, Mexico), (flood, Britain)

Quantify Uncertainty

- Typicality

  P(concept | instance)

  P(instance | concept)

  P(concept | property)

  P(property | concept)

- Similarity

  sim(concept_1, concept_2)

the foundation of text understanding and reasoning

Text Mining / IE: State of the Art

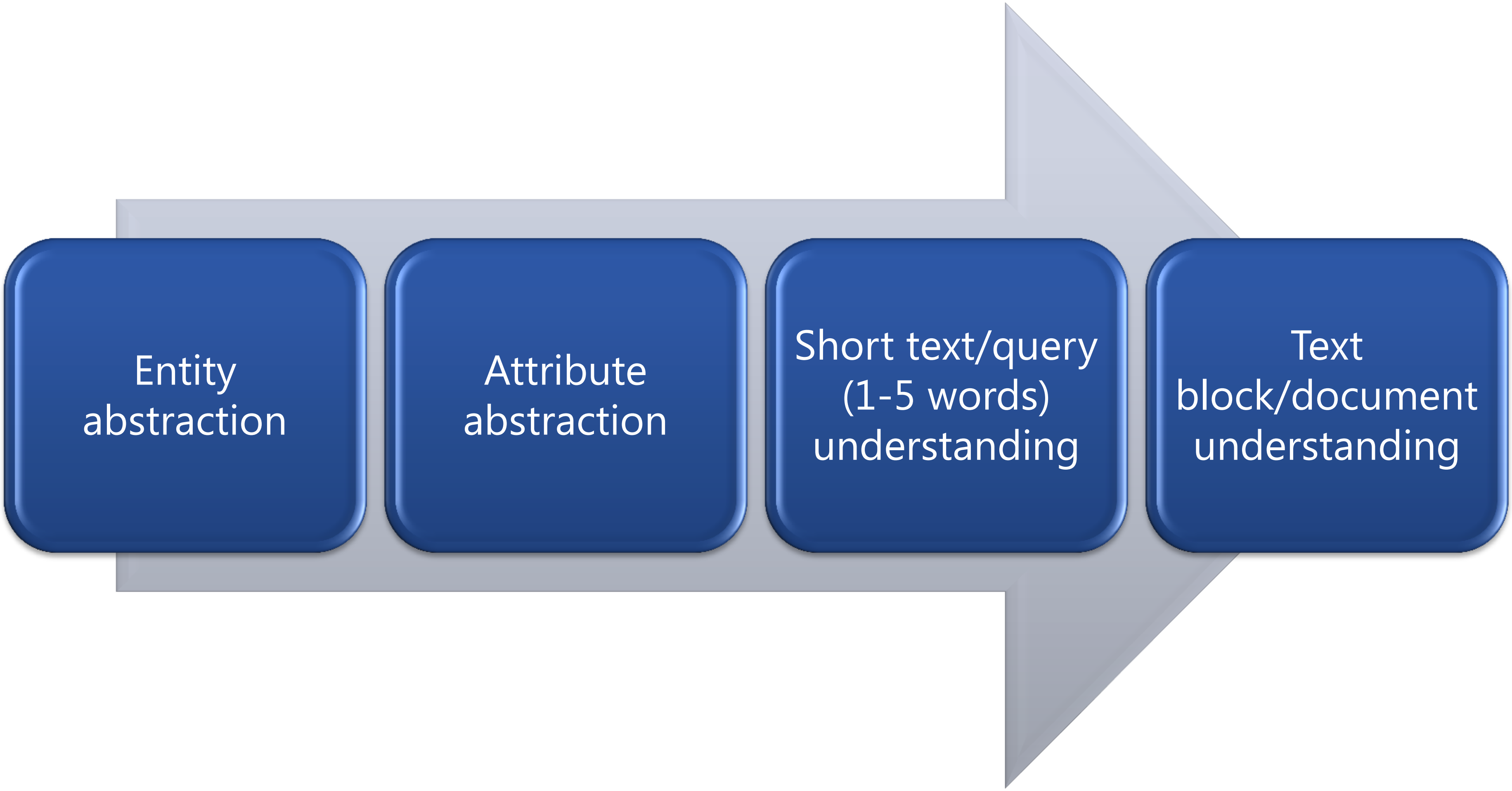
- Bag of words based approach: e.g., LDA
  - Based on multiple document statistics
  - Simple bag-of-words, no semantics
- Supervised learning: e.g., CRF
  - Labeled training data required
  - Difficulty for out-of-sample features
- Lack of semantics
- What role can a knowledgebase play?

Shopping

Five of us bought 5 Kinects and posed in front of an Apple store.

Apple store that sells fruits or apple store that sells iPads?

Step by Step Understanding



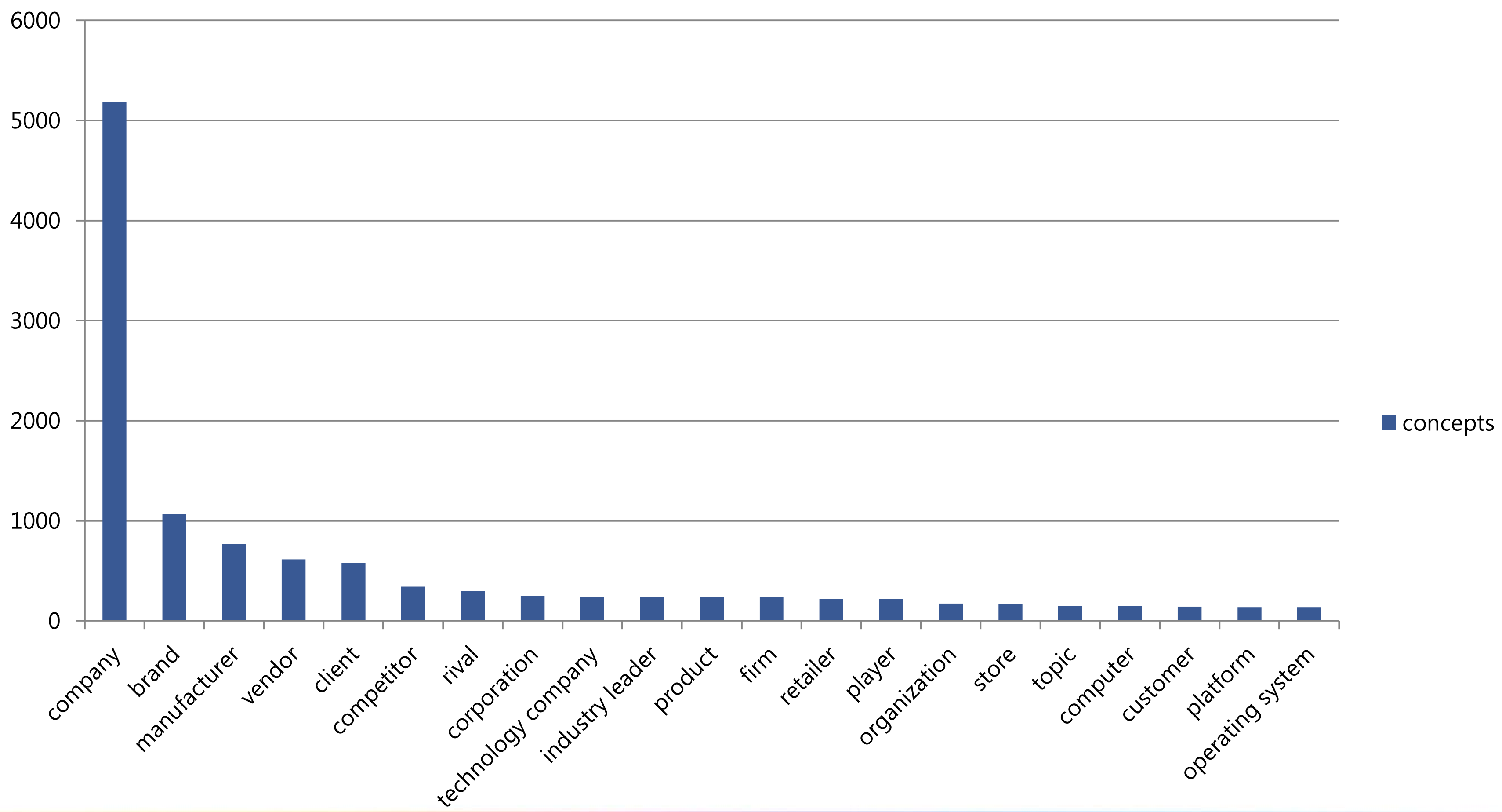
Short Text

- Challenge:
  - Not enough statistics
- Applications
  - Twitter
  - Query/Search Log
  - Anchor Text
  - Image/video tag
  - Document paraphrasing and annotation

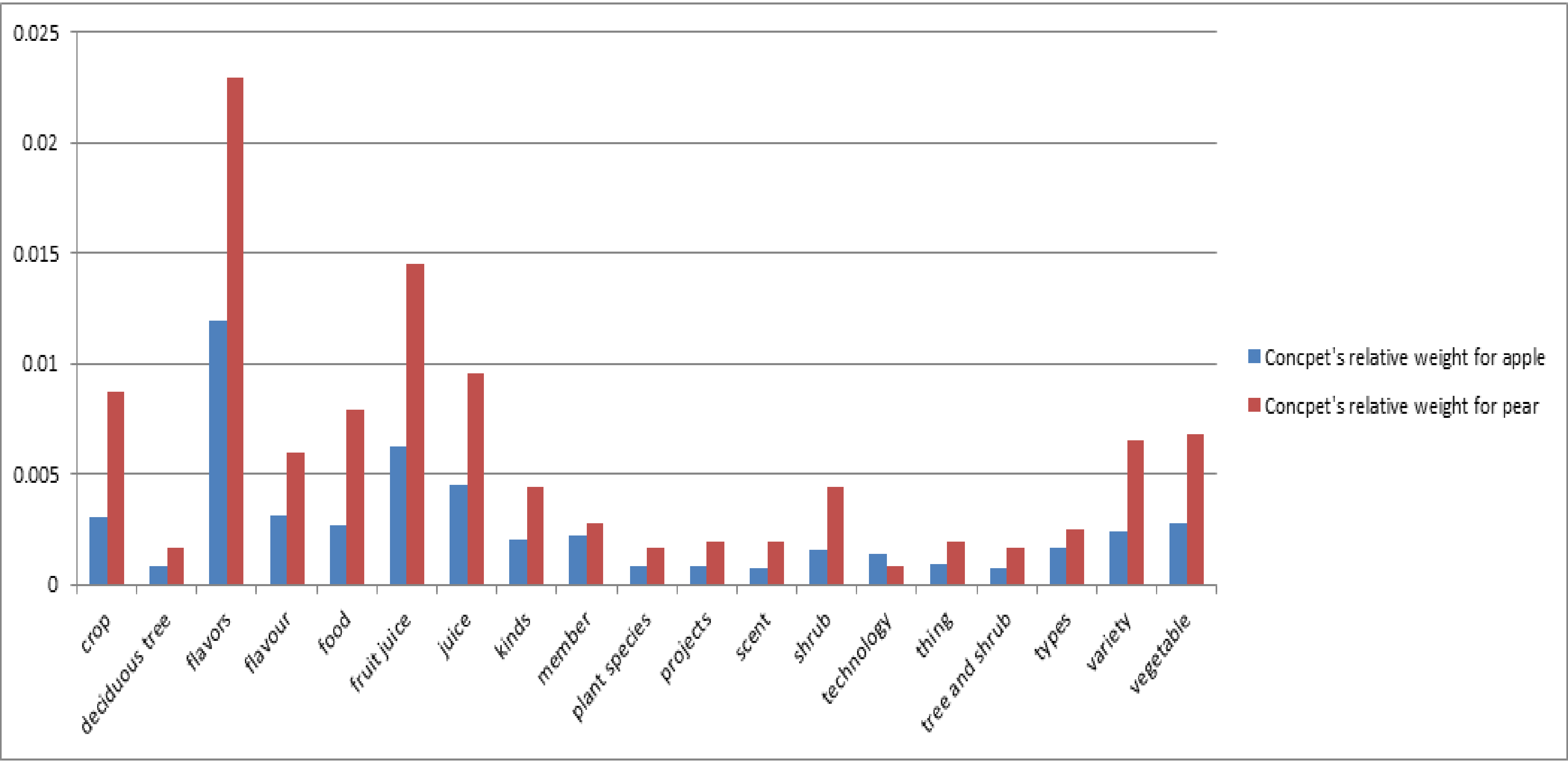
Comparison of Knowledge Bases

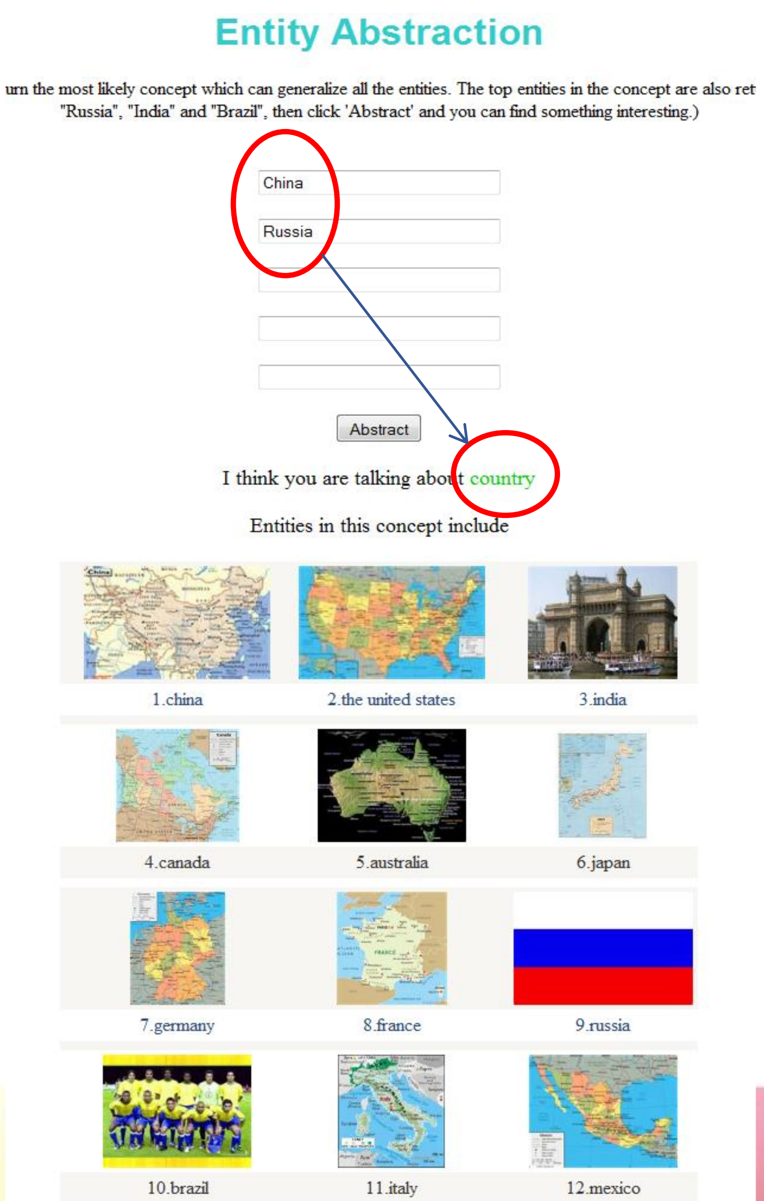
	WordNet	**Wikipedia**	**Freebase**	**Probase**
Cat	Feline; Felid; Adult male; Man; Gossip; Gossiper; Gossipmonger; Rumormonger; Rumourmonger; Newsmonger; Woman; Adult female; Stimulant; Stimulant drug; Excitant; Tracked vehicle; …	Domesticated animals; Cats; Felines; Invasive animal species; Cosmopolitan 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; …

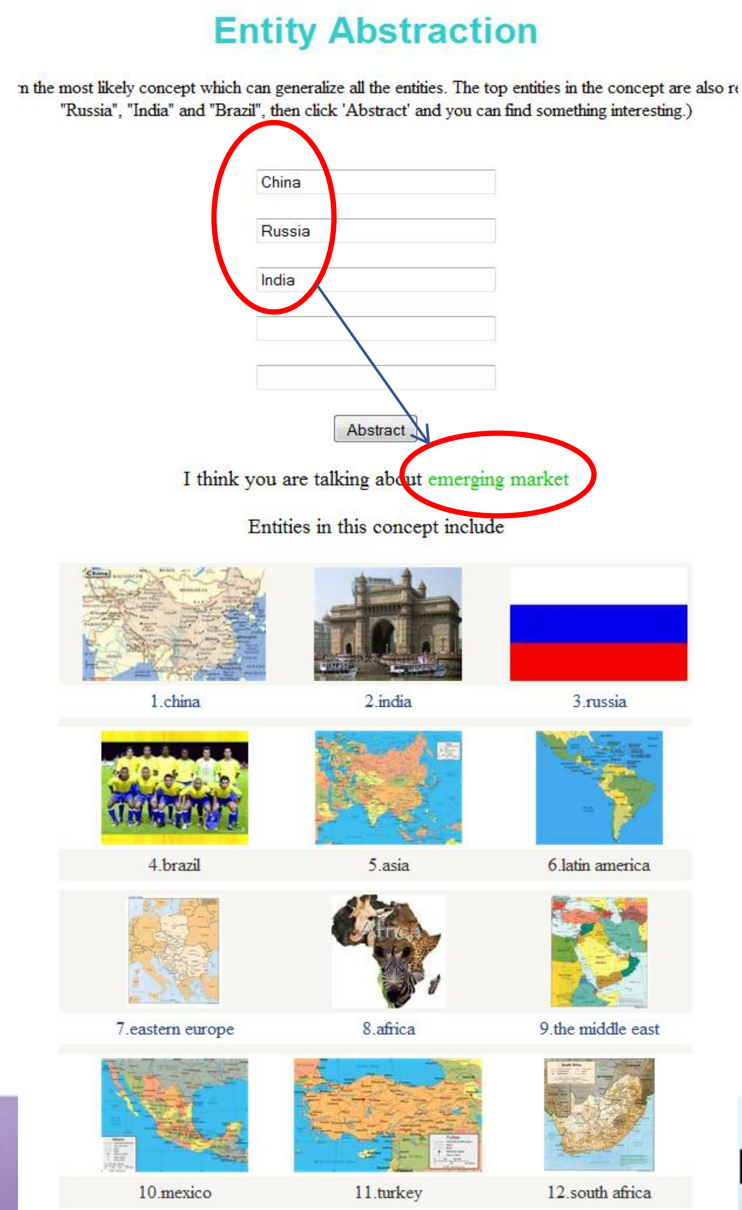
In the mind of the machine: when it sees the word ‘apple’



… when it sees the words ‘apple’ and ‘pear’ together




Entity Abstraction
urn the most likely concept which can generalize all the entities. The top entities in the concept are also ret
"Russia", "India" and "Brazil", then click 'Abstract' and you can find something interesting.)
China
Russia
Abstract
I think you are talking about country
Entities in this concept include
1.china
2.the united states
3.india
4.canada
5.australia
6.japan
7.germany
8.france
9.russia
10.brazil
11.italy
12.mexico


Entity Abstraction
n the most likely concept which can generalize all the entities. The top entities in the concept are also r
"Russia", "India" and "Brazil", then click 'Abstract' and you can find something interesting.)
China
Russia
India
Abstract
I think you are talking about emerging market
Entities in this concept include
1.china
2.india
3.russia
4.brazil
5.asia
6.latin america
7.eastern europe
8.africa
9.the middle east
10.mexico
11.turkey
12.south africa

mmit

Entity Abstraction

- Given a set of entities

$$E = \{e_i, i \in 1, ..., M\}$$

- Target Concept (Naïve Bayes Rule)

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

- Where c_k a concept, and

$$P(e_i|c_k) = \frac{P(e_i, c_k)}{P(c_k)}$$

- is computed ba ' co-occurrence

How to Infer Concept from Attribute?

- Given a set of attributes
- The Naïve Bayes Rule gives

$$A = \{a_j, j \in 1, ..., N\}.$$

- where

$$P(c_k|A) = \frac{P(A|c_k)P(c_k)}{P(A)} \propto P(c_k) \prod_{j=1}^{N} P(a_j|c_k),$$

$$P(a_j|c_k) = \sum_{i:e_i \in E} P(a_j|e_i)P(e_i|c_k),$$

(university, florida state university, 75)
(university, harvard university, 388)
(university, university of california, 142)
(country, china, 97346)
(country, the united states , 91083)
(country, india , 80351)
(country, canada , 74481)

\+

(florida state university, website, 34)
(harvard university, website, 38)
(university of california, city, 12)
(china, capital, 43)
(the united states , capital, 32)
(india , population, 35)
(canada , population, 21)

⇨

(university, website, 4568)
(university, city, 2343)
(country, capital, 4345)
(country, population, 3234)
……

When Type of Term is Unknown:

- Given a set of terms with unknown types $T = \{t_l\}, l = 1, \ldots, L$
- Generative model

$$P(t_l|c_k) = P(t_l|z_l = 1, c_k)P(z_l = 1|c_k) + P(t_l|z_l = 0, c_k)P(z_l = 0|c_k)$$

Using Naive Baye

$$P(c_k|T) = \frac{P(T|c_k)P(c_k)}{P(T)} \propto P(c_k)\prod_{l}^{L} P(t_l|c_k)$$

- Discriminative model (Noisy-OR)

$$P(c_k|t_l) = 1 - (1 - P(c_k|t_l, z_l = 1))(1 - P(c_k|t_l, z_l = 0))$$

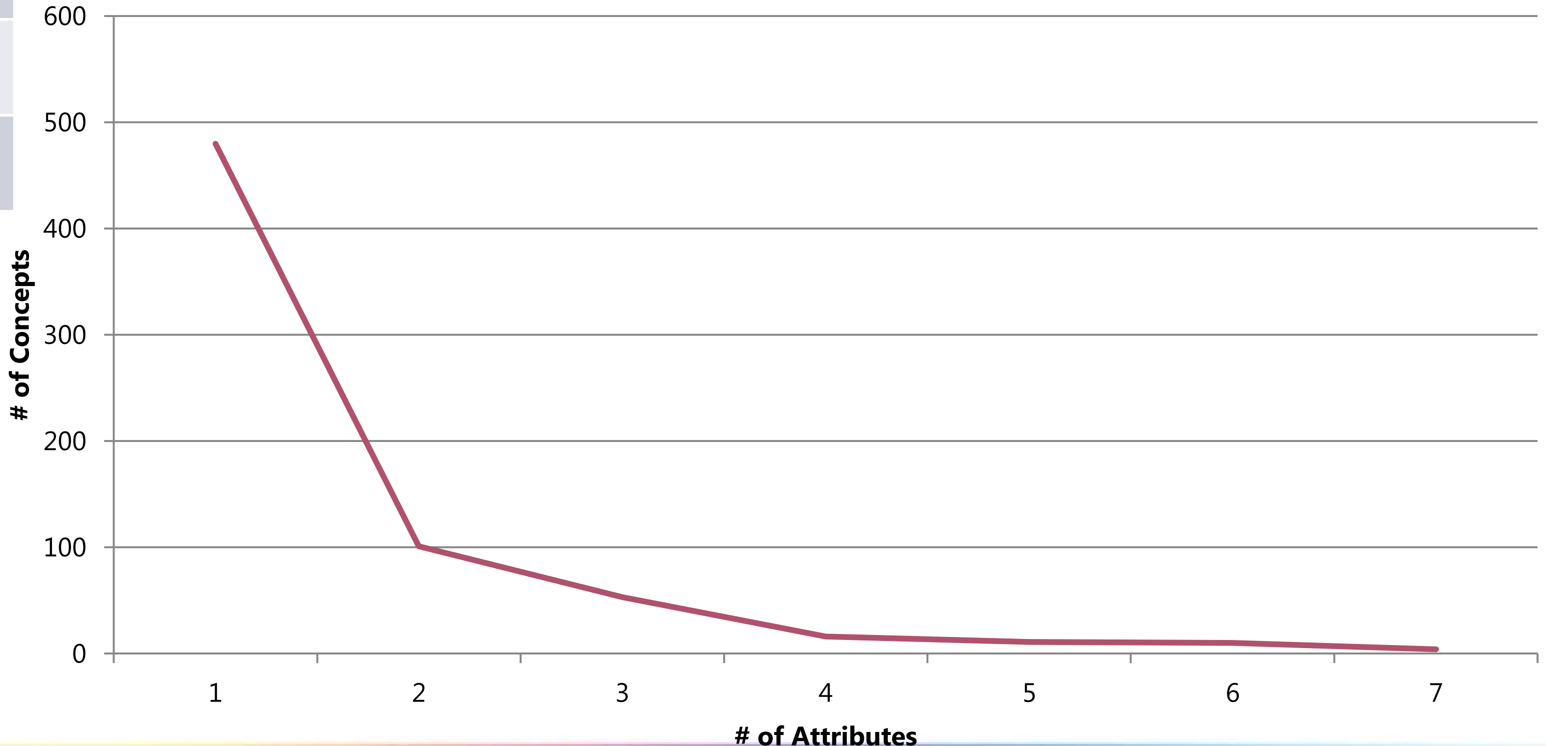
And using twice

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

- where $z_l = 1$ indicate “entity” and $z_l = 0$ indicate “attribute”

When you see attributes …

	website	president	city	motto	state	type	director



Understanding = Concept Forming

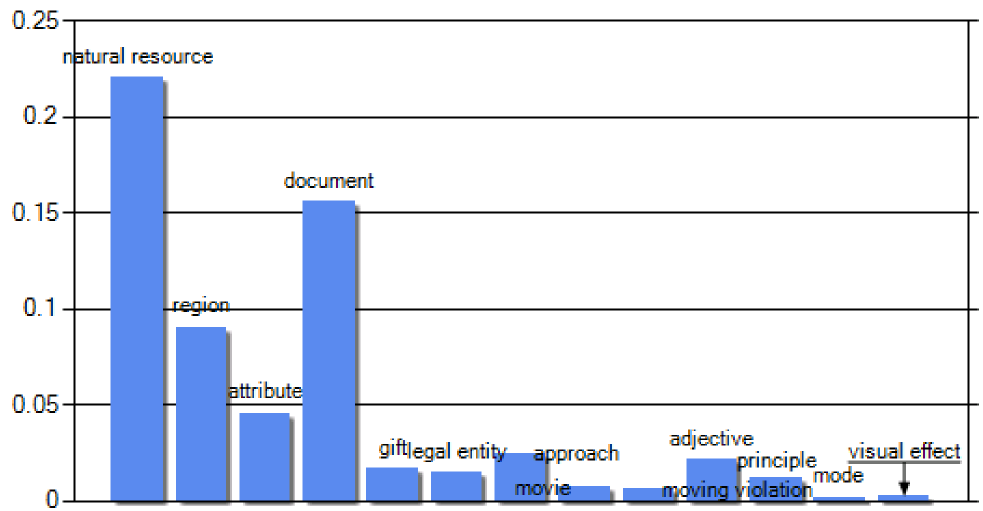
apple			
pear			

Short Text Conceptualization

Please input your query/text

```
What happens to lakes in an area hit by forest fires and floods?
Some will glow in the dark.
```

Parse



Parsed text: **area; hit; forest; will; glow; dark;**

Recommend Attribute:

Recommend Concept:

Recommend Entity:

Clustering Twitter Messages

- Problem 1 (unique concepts): use keywords to retrieve tweets in 3 categories:
  - 1. Microsoft, Yahoo, Google, IBM, Facebook
  - 2. cat, dog, fish, pet, bird
  - 3. Brazil, China, Russia, India
- Problem 2 (concepts with subtle differences): use keywords to retrieve tweets in 4 categories:
  - 1. United states, American, Canada
  - 2. Malaysia, China, Singapore, India, Thailand, Korea
  - 3. Angola, Egypt, Sudan, Zambia, Chad, Gambia, Congo
  - 4. Belgium, Finland, France, Germany, Greece, Spain, Switzerland

Comparison Results

Clustering NMI scores on Twitter data.

Method	@Problem1	@Problem2
Original Data	0.215±0.010	0.452±0.076
LDA (1×Cluster Num)	0.161±0.065	0.114±0.037
LDA (2×Cluster Num)	0.067±0.022	0.069±0.024
WordNet	0.195±0.070	0.074±0.074
Freebase	0.531±0.164	0.204±0.037
Wikipedia (Category-Link)	0.540±0.077	0.336±0.089
Wikipedia (ESA)	0.351±0.132	0.340±0.800
Probase (Top 10)	0.318±0.110	0.490±0.029
Probase (Top 20)	0.479±0.111	0.555±0.019
Probase (Top 50)	0.559±0.123	**0.632±0.066**
Probase (Top 500)	**0.826±0.062**	0.301±0.189
Probase (Top 5000)	0.690±0.176	0.095±0.084

Many Applications …

- Mapping questions to knowledge
  - How many people are in China? → entity: China, Attribute: population
  - Where is MSR? → entity: MSR, Attribute: location
  - How long does it take for Asclepius to take effect? → entity: Asclepius, Attribute: pharmaceutical effect
- Synonym
  - China national song → entity: China, Attribute: national anthem
  - USA headline → entity: USA, Attribute: news
  - India demographic → entity: India, Attribute: population
- Misspelling
  - Japan poulation → entity: Japan, Attribute: population
- Correlated indirectly
  - google earth China → entity: China, Attribute: map
  - China dishes → entity: China, Attribute: food
  - what is the exchange rate for UK → entity: UK, Attribute: currency

Summary

- A little knowledge goes a long way
- A concept space large enough to model the concepts in a human mind
- Scores and weights that enable Bayesian reasoning.
- Many applications

Thanks!

Examples

Concept	Entity	Co-occurrence	Concept Number	Entity Number	P(e|c)	P(c|e)
country	india	80905	2262485	197915	0.03576	0.40879
country	china	98517	2262485	269127	0.04354	0.36606
emerging market	china	6556	29298	269127	0.22377	0.02436
emerging market	india	5702	29298	197915	0.19462	0.02881
area	china	2231	2525020	269127	0.00088	0.00829
area	india	1797	2525020	197915	0.00071	0.00908

Concept	Attribute	P(c, a)	P(c)	P(a)	P(a|c)	P(c|a)
country	population	4.08183	173.44931	41736.78060	0.02353	0.00010
country	language	1.48795	173.44931	58584.50905	0.00858	0.00003
emerging market	language	4.52949	402.13772	58584.50905	0.01126	0.00008
emerging market	population	16.54701	402.13772	41736.78060	0.04115	0.00040

Examples

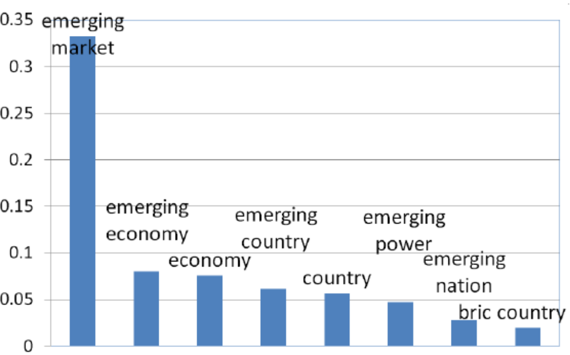
- Given "china", "india", "language" and "population", "emerging market" will be ranked as 1st

Concept	Entity	Co-occurrence	Concept Number	Entity Number	P(e|c)	P(c|e)
country	india	80905	2262485	197915	0.03576	0.40879
country	china	98517	2262485	269127	0.04354	0.36606
emerging market	china	6556	29298	269127	0.22377	0.02436
emerging market	india	5702	29298	197915	0.19462	0.02881
area	china	2231	2525020	269127	0.00088	0.00829
area	india	1797	2525020	197915	0.00071	0.00908

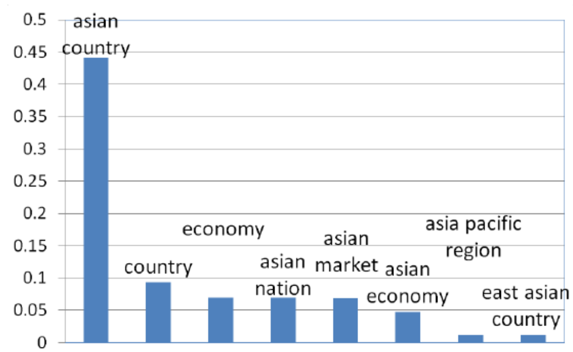
Concept	Attribute	P(c, a)	P(c)	P(a)	P(a|c)	P(c|a)
factor	population	75.74704	71073.46656	41736.78060	0.00107	0.00181
factor	language	113.32628	71073.46656	58584.50905	0.00159	0.00193
countries	population	4.08183	173.44931	41736.78060	0.02353	0.00010
countries	language	1.48795	173.44931	58584.50905	0.00858	0.00003
emerging market	language	4.52949	402.13772	58584.50905	0.01126	0.00008
emerging market	population	16.54701	402.13772	41736.78060	0.04115	0.00040

mit

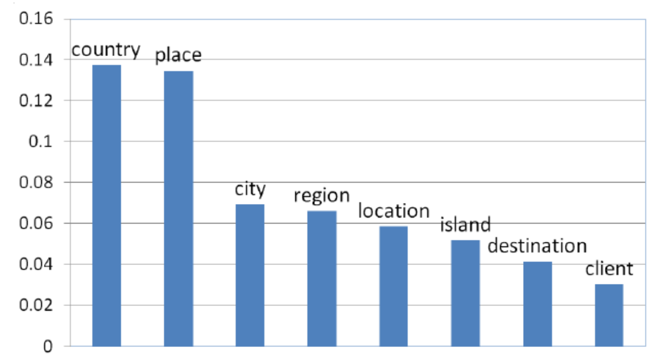
Example (Cont'd)



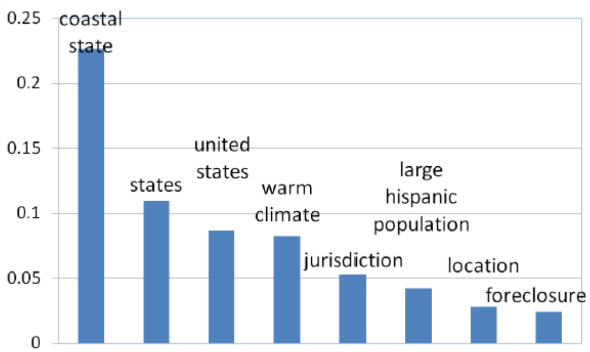
(a) China (I), Russia (I), India (I), Brazil (I)



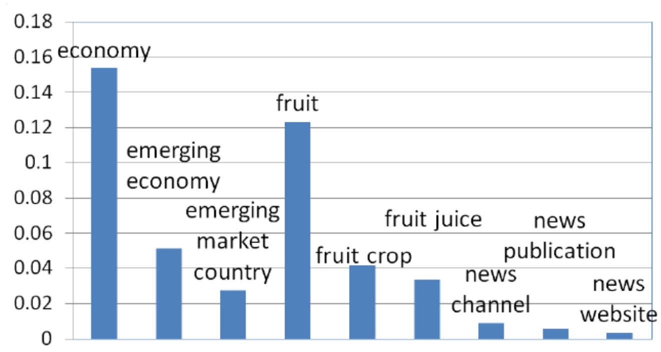
(b) China (I), India (I), Japan (I), Singapore (I)



(c) population (A), location (A), president (A)



(d) California (U), Florida (U), population (U)



(e) China (U), Brazil (U), Russia (U), apple (U), banana (U), BBC (U), New York Time (U)

Microsoft® Research
FacultySummit

2031
FUTURE WORLD
2011