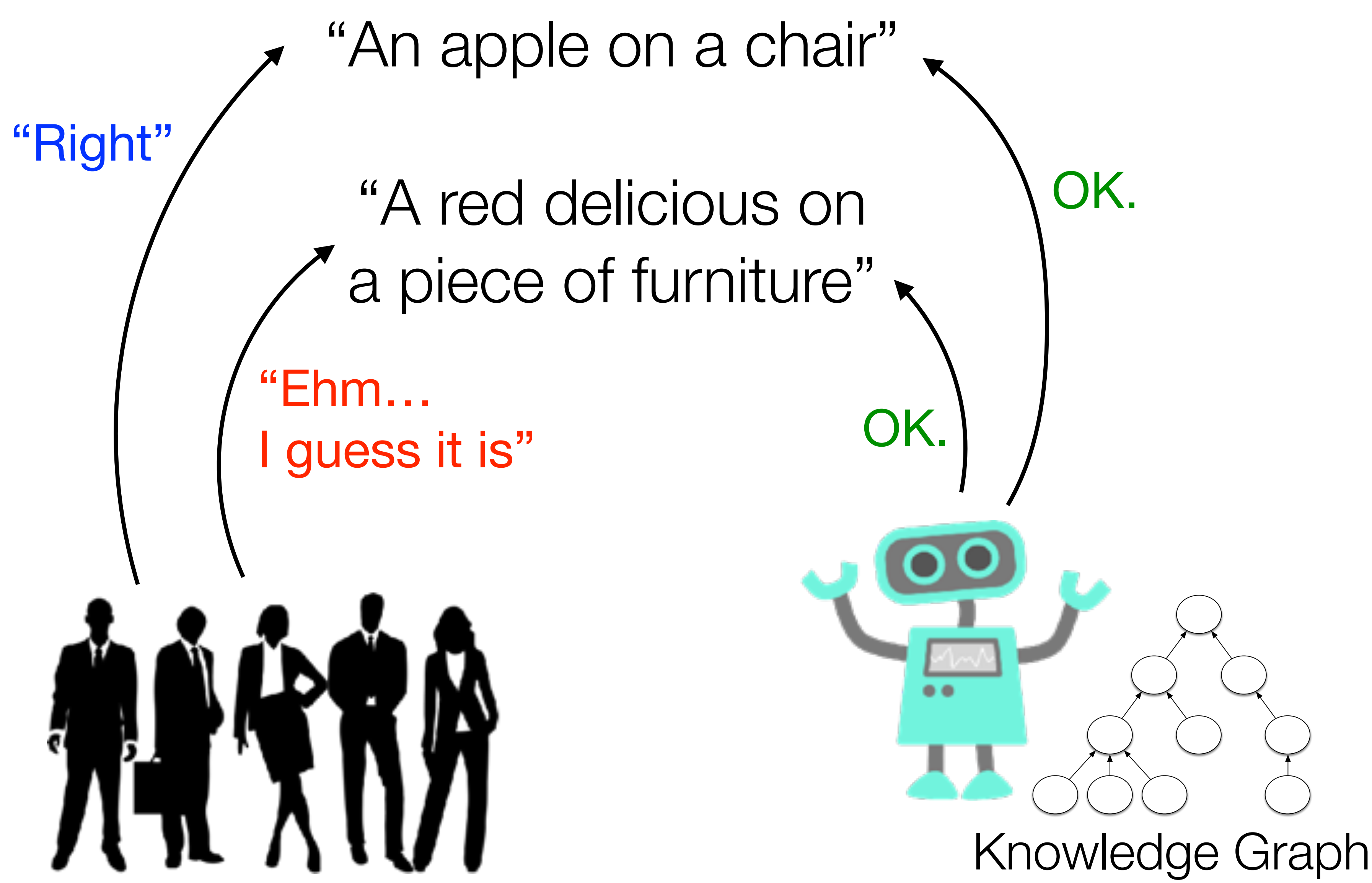
Predicting the basic level in a hierarchy of concepts

Laura Hollink, Aysenur Bilgin, Jacco van Ossenbruggen
Human Centered Data Analytics
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands

Motivation

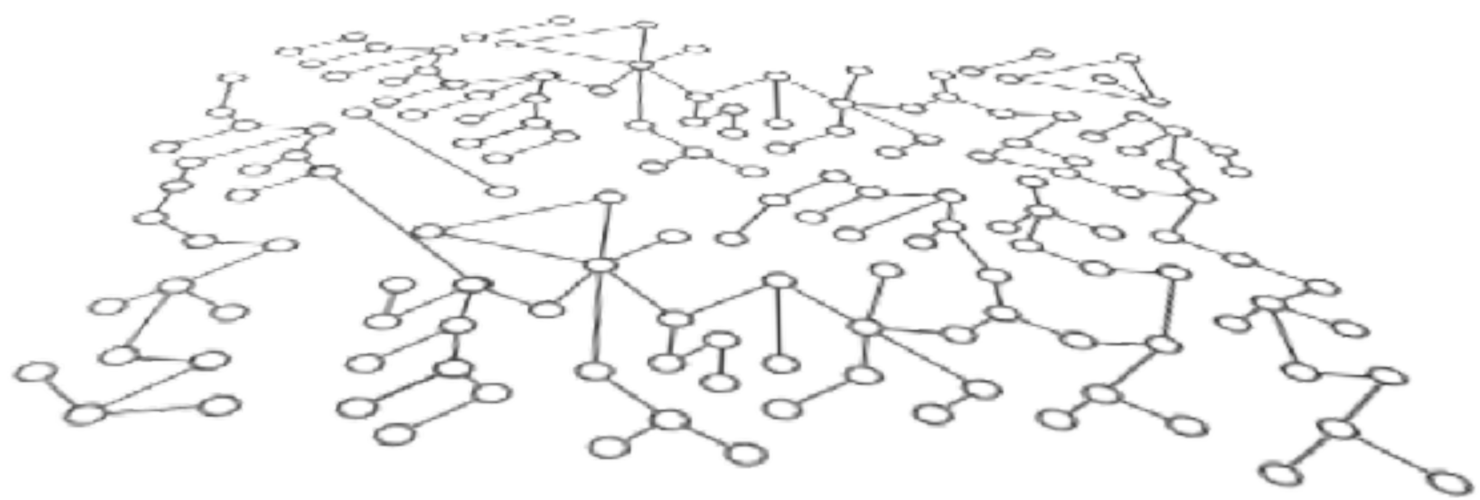
Basic level theory

What is this?



Source: Industry-scale Knowledge Graphs: Lessons and Challenges. Natasha Noy at al. 2019. Queue 17-2.

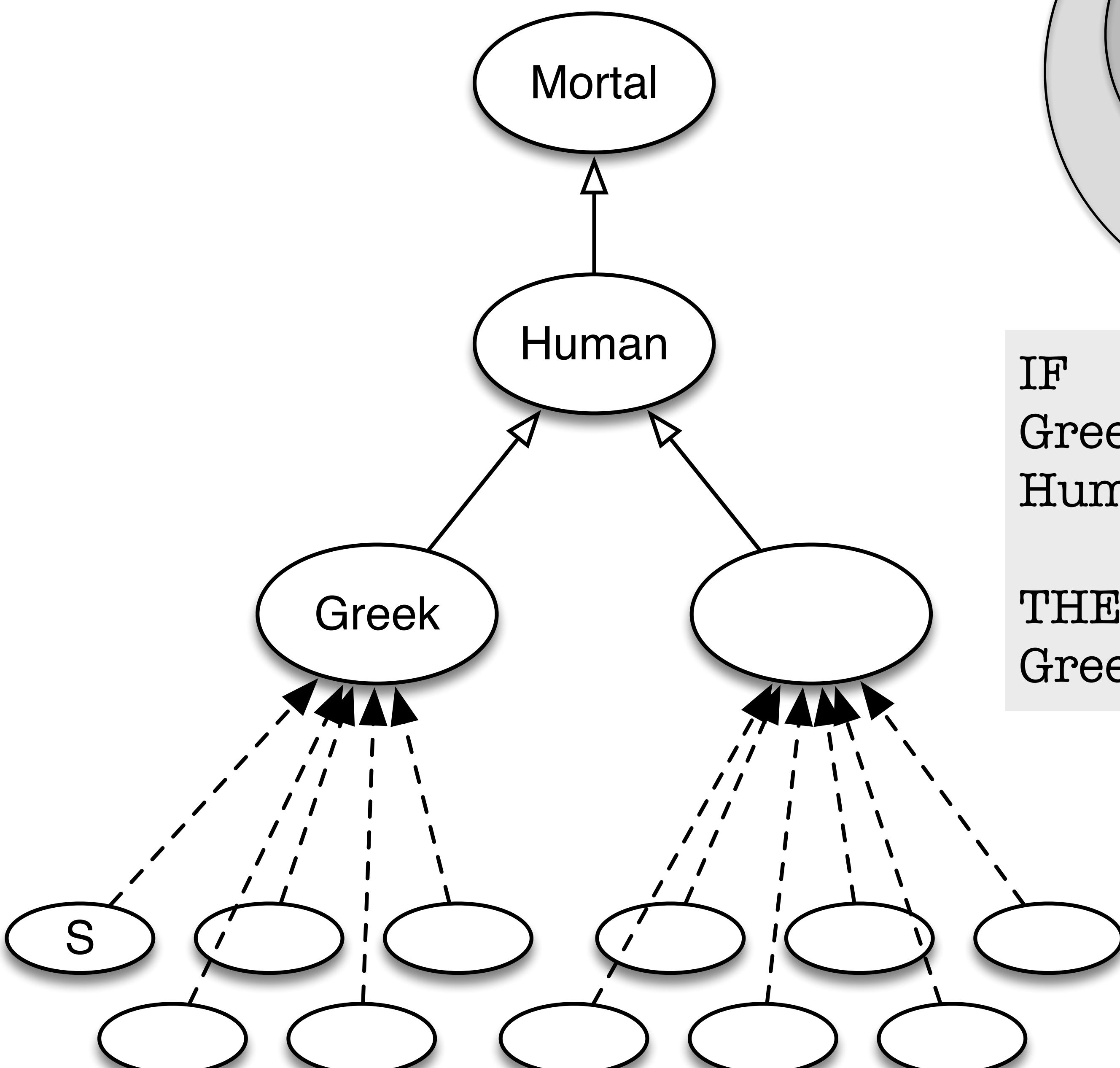
Knowledge graphs

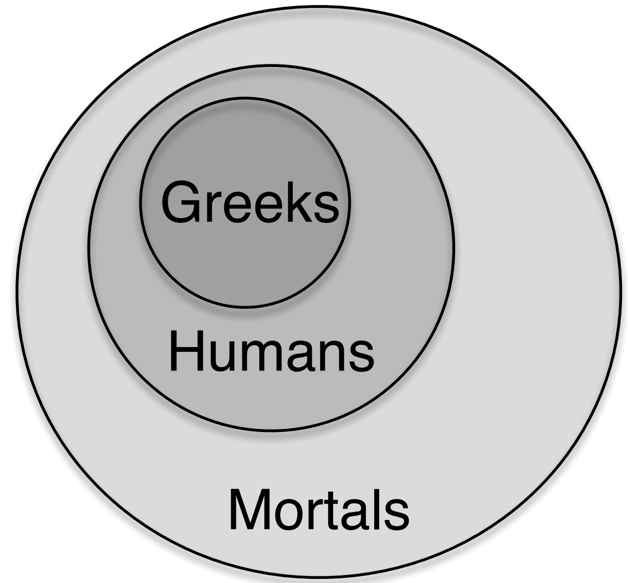


```
tenoroon is_a double-reed instrument.
double-reed instrument is_a woodwind-instrument.
```

```
dyshidrotic_eczema is_a skin_condition.
dyshidrotic_eczema occurs_on extremities.
```

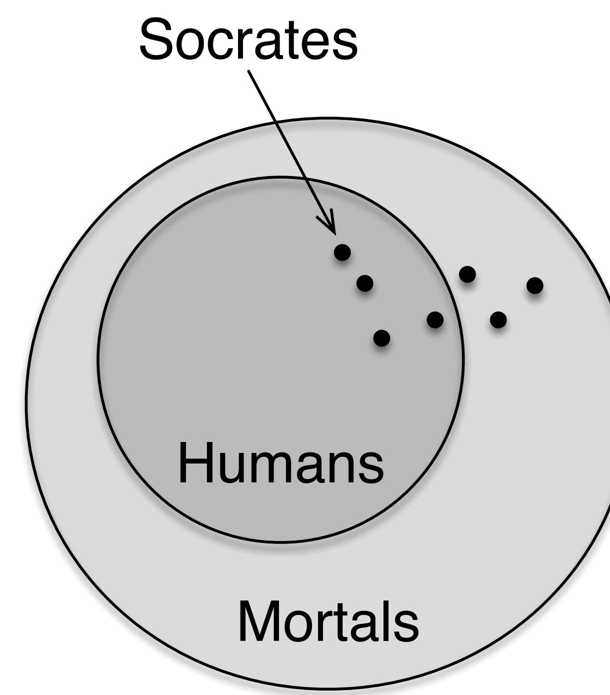
Data models in knowledge graphs





IF
Greeks are Humans
Humans are Mortals

THEN
Greeks are Mortals



IF
Socrates is-a Human
Humans are Mortals

THEN
Socrates is-a Mortal

‘Data models’ in the human mind

- No necessary and sufficient conditions, but something like “family resemblances”
- Members of a class may not share any characteristics

Wittgenstein

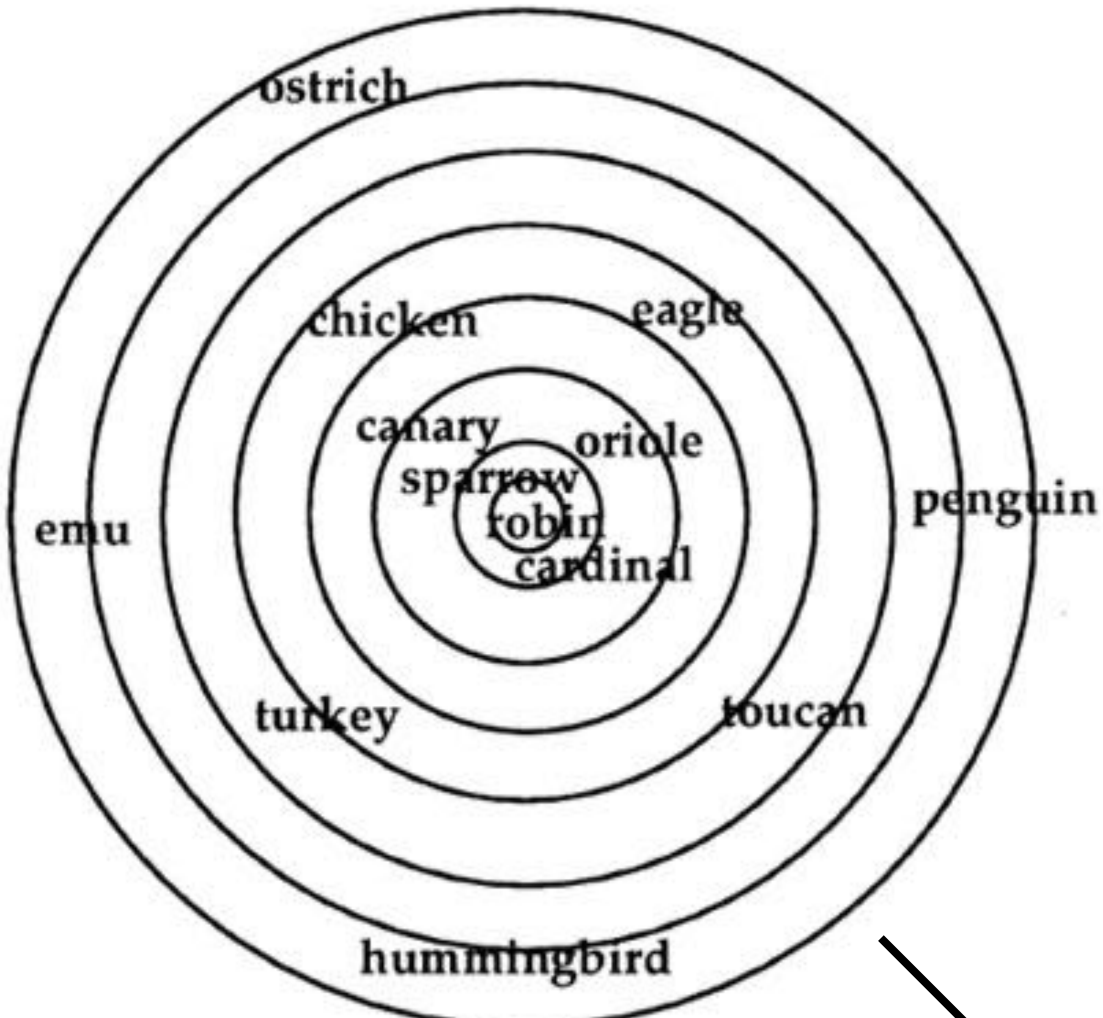
Rosch

Empirical evidence

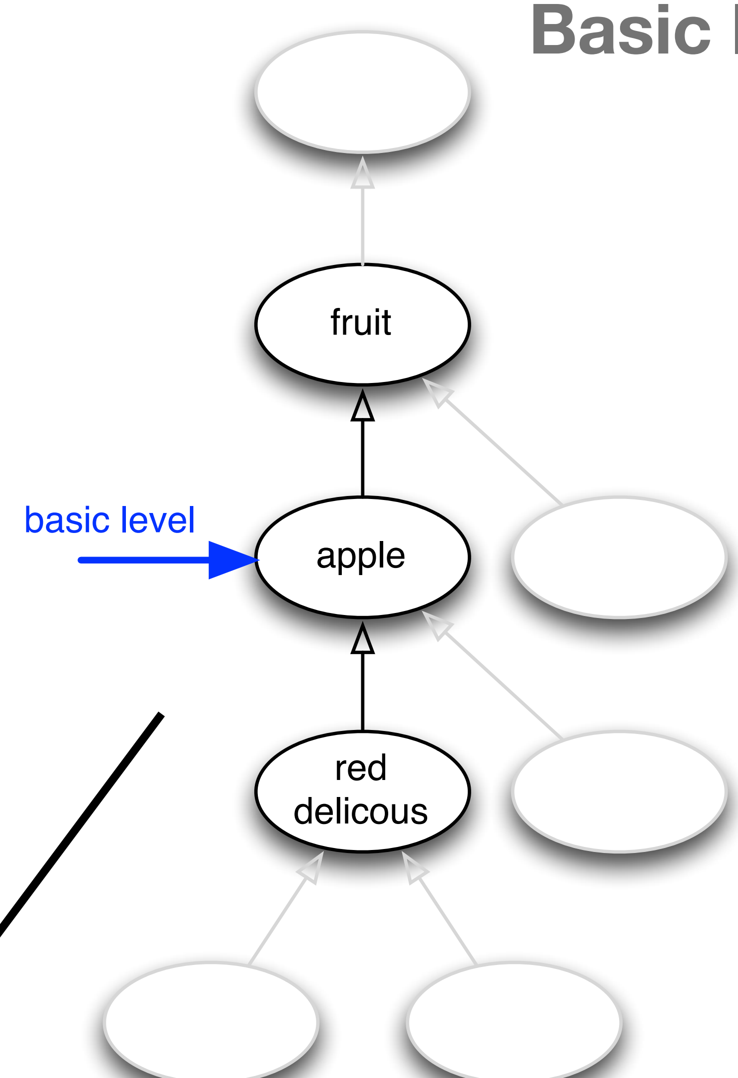
- Prototype theory
- Notion of the basic level

Empirical evidence of this human 'data model'

Prototype theory



Basic level



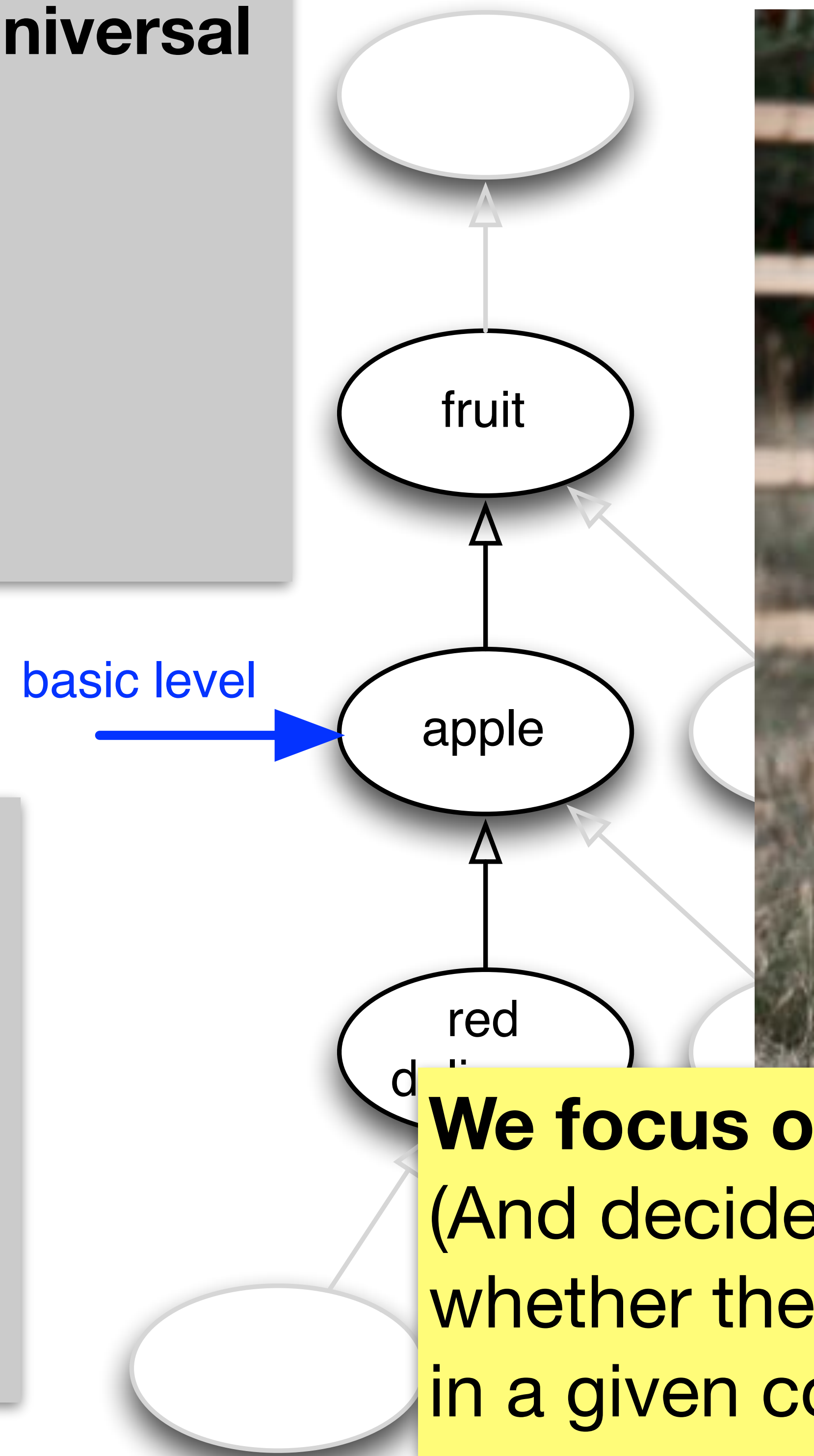
Lab experiments showed:

Effects: people react **quicker**, more **accurately**, more **consistently**

Large agreement between people

Basic level across people and cultures

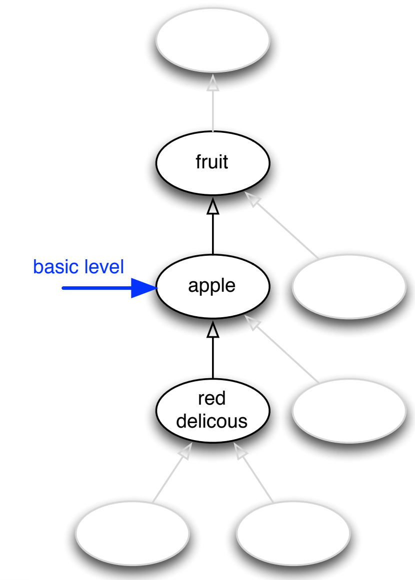
Basic level is almost universal
Because:
- ‘Gestalt’
- Muscle movement
- Most new information
- Most used words
- First learned words



A few things are known to alter basic level effects somewhat:
- Expertise
- Familiarity
- Prototypicality

We focus on universal effect.
(And decide on a case-by-case basis whether the basic level is the right level in a given context)

Can we predict
which concepts in a knowledge graph
are basic level concepts?

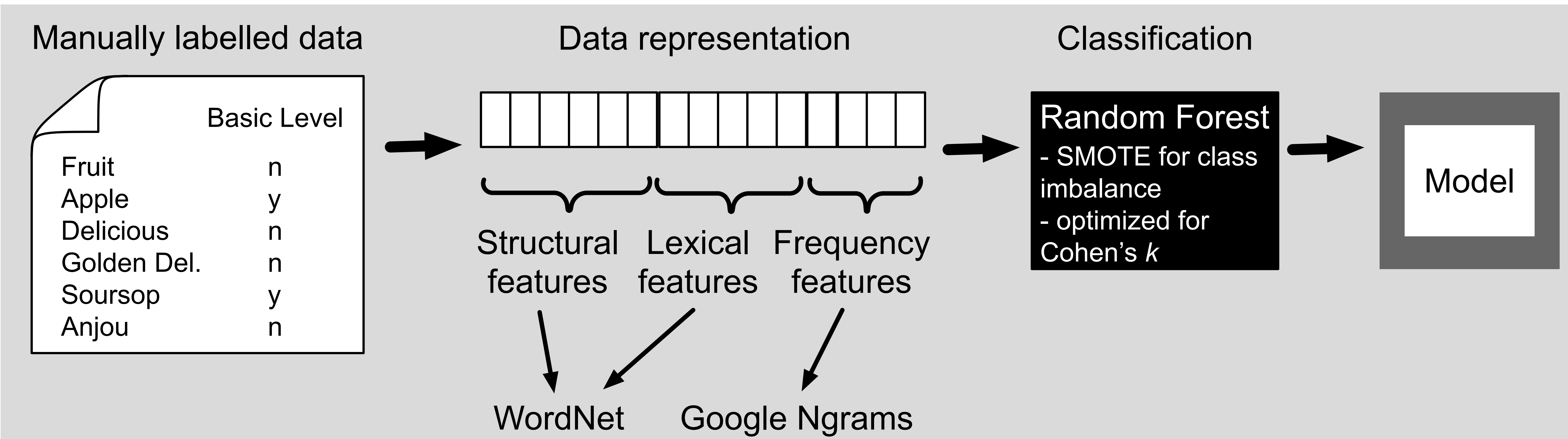


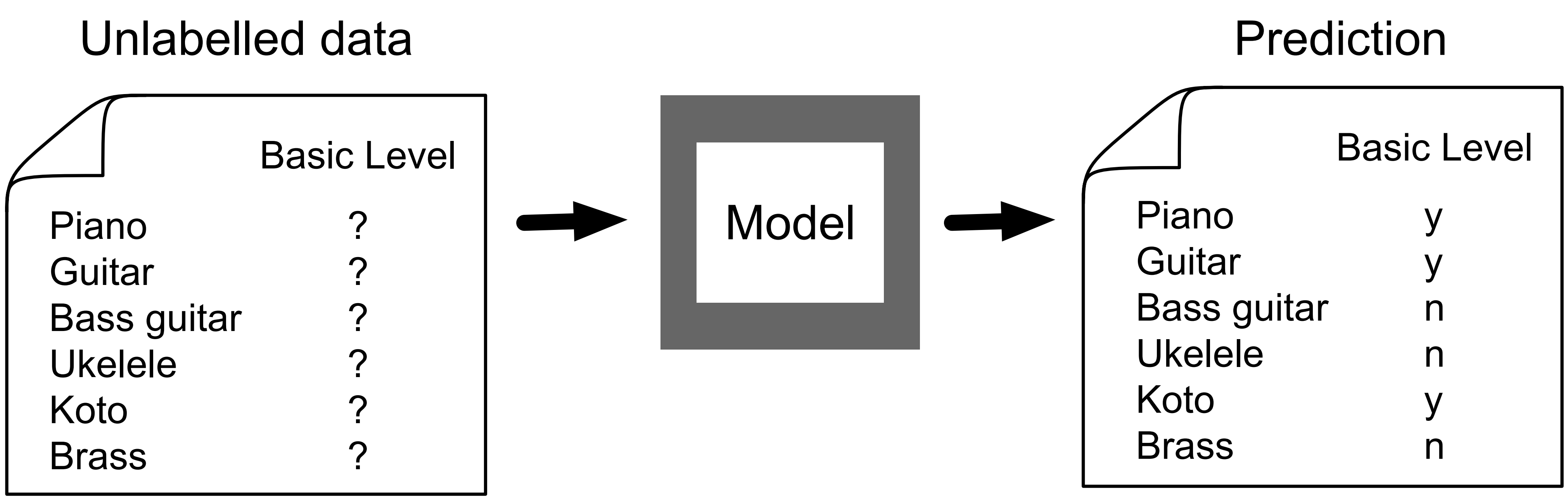
Or, rephrased:
Can we predict
for which concepts in a Knowledge graph
can users be expected to display **basic level effects**?

Hypothesis:
Instead of lab experiments with human subjects,
we can learn this from ‘**human-produced data’.**

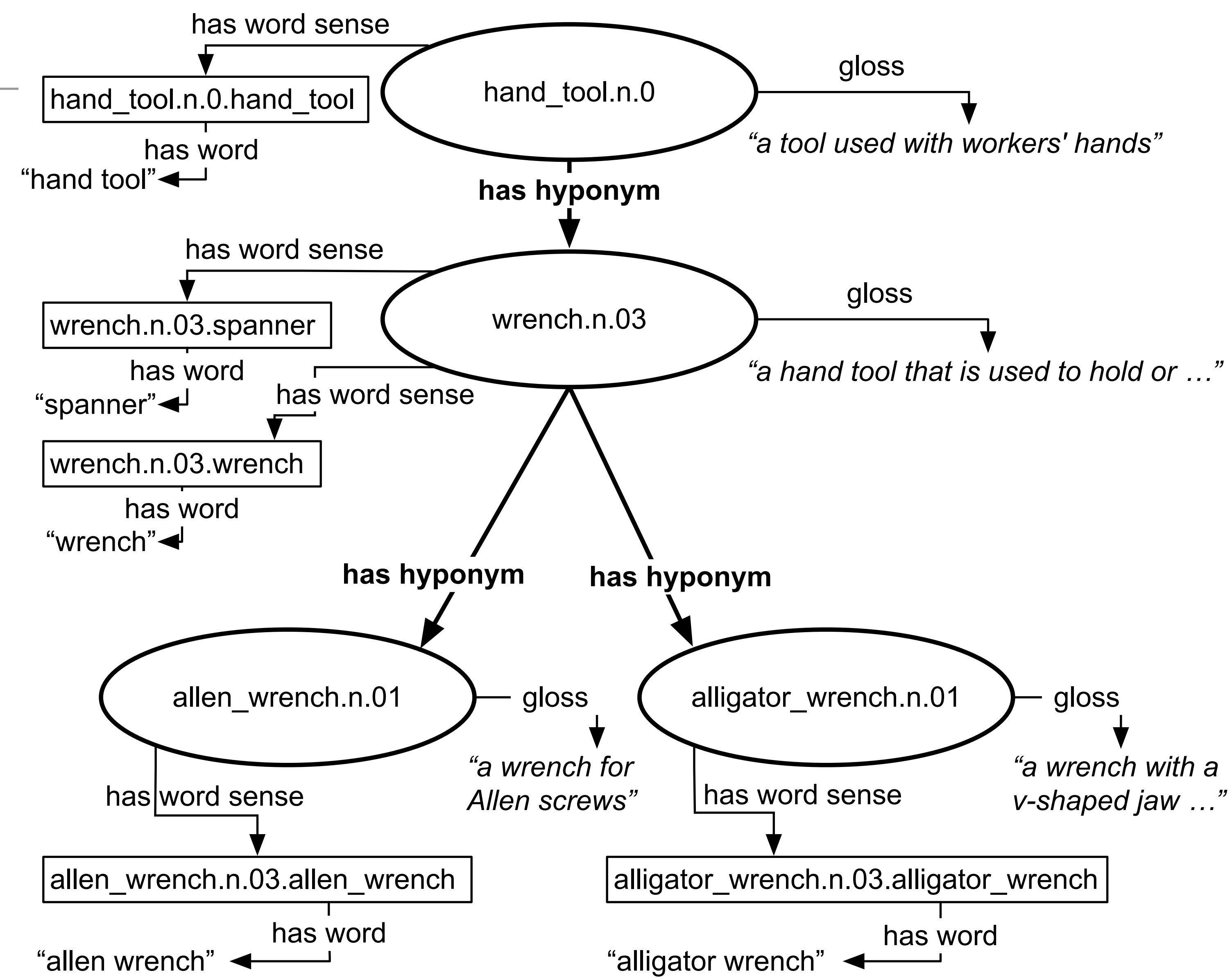
L Hollink, A Bilgin, J van Ossenbruggen. *Predicting the Basic Level in a Hierarchy of Concepts*. Metadata and Semantics Research Conference, Nov/Dec 2020.

Predicting the basic level based on three types of human-produced data





Features



Structural features: from WordNet

"the level at which people can name most properties"

- Nr. of subconcepts
- Nr. of direct superconcepts
- Nr. of part-of properties
- Depth in hierarchy
- Length of the description ("gloss")

Lexical features: from WordNet

"the level with shortest, most polysemous words"

- Word_length
- Nr. of meanings
- Nr. of synonyms

Frequency features: from Google Ngrams

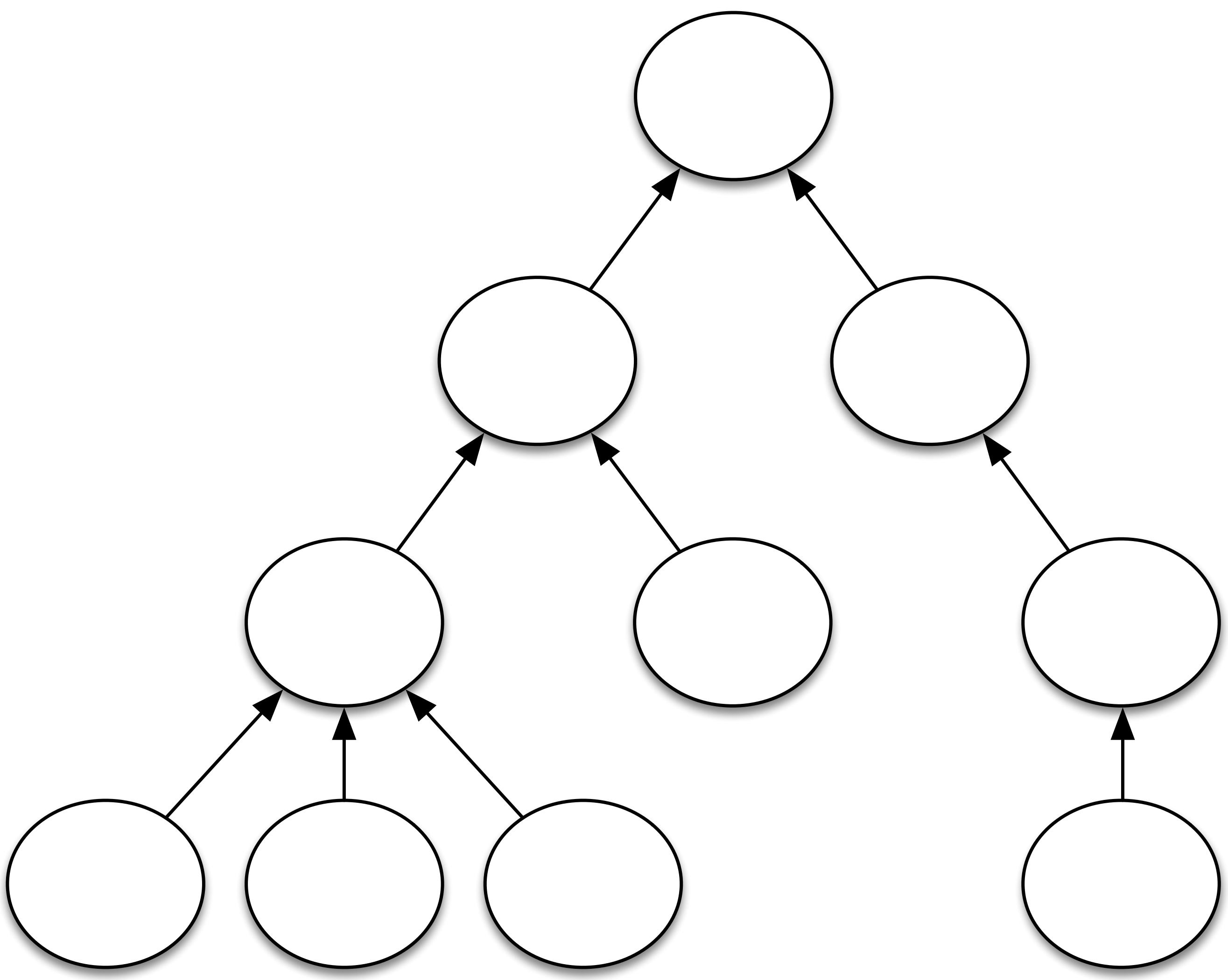
"the level which is named most often by people"

- Frequency of occurrence of the word in the Google Books corpus

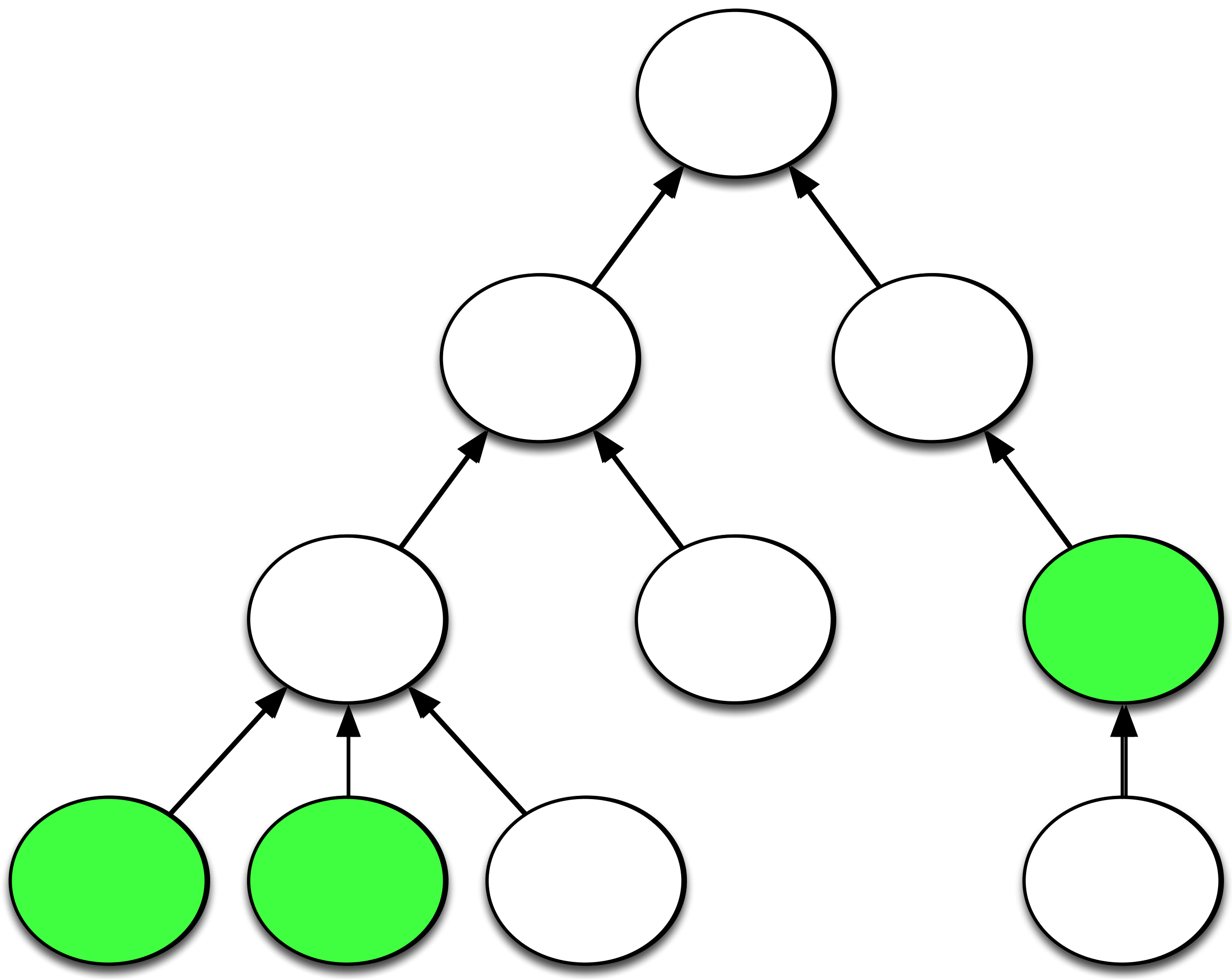
Google Books

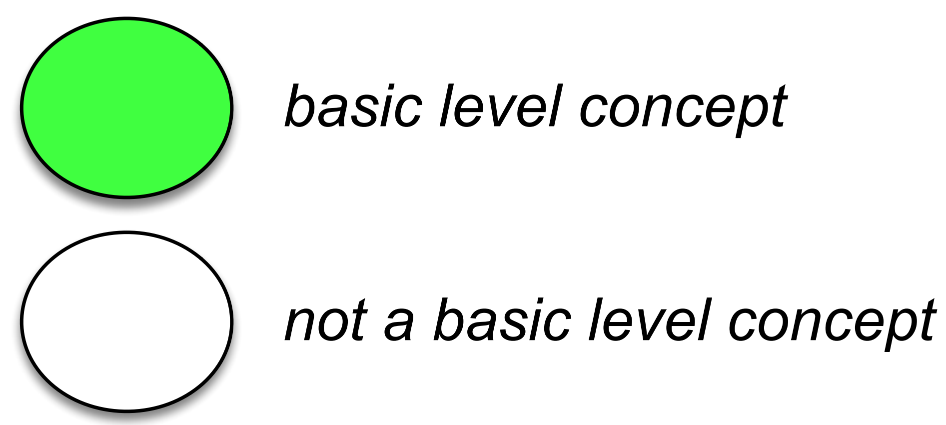
Corpus 'English 2012'
4.5M books, 1800-2008

A binary classification problem



A binary classification problem





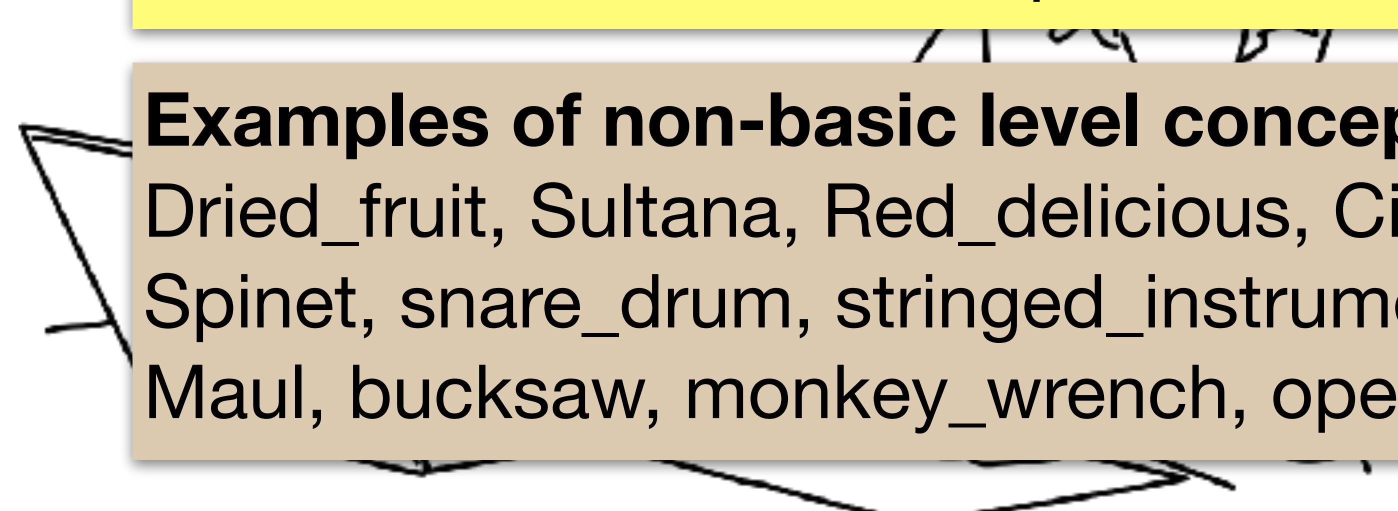
Experiment: basic level prediction in WordNet
Training and testing on manually labelled concepts

- 518 concepts in WordNet
- From 3 branches / “domains” that correspond to categories in Rosch’ experiments.
- all labelled by 3 raters
- Krippendorff α = 0.73
- 1/3 labelled as basic level

Measuring
basic level effects

≠

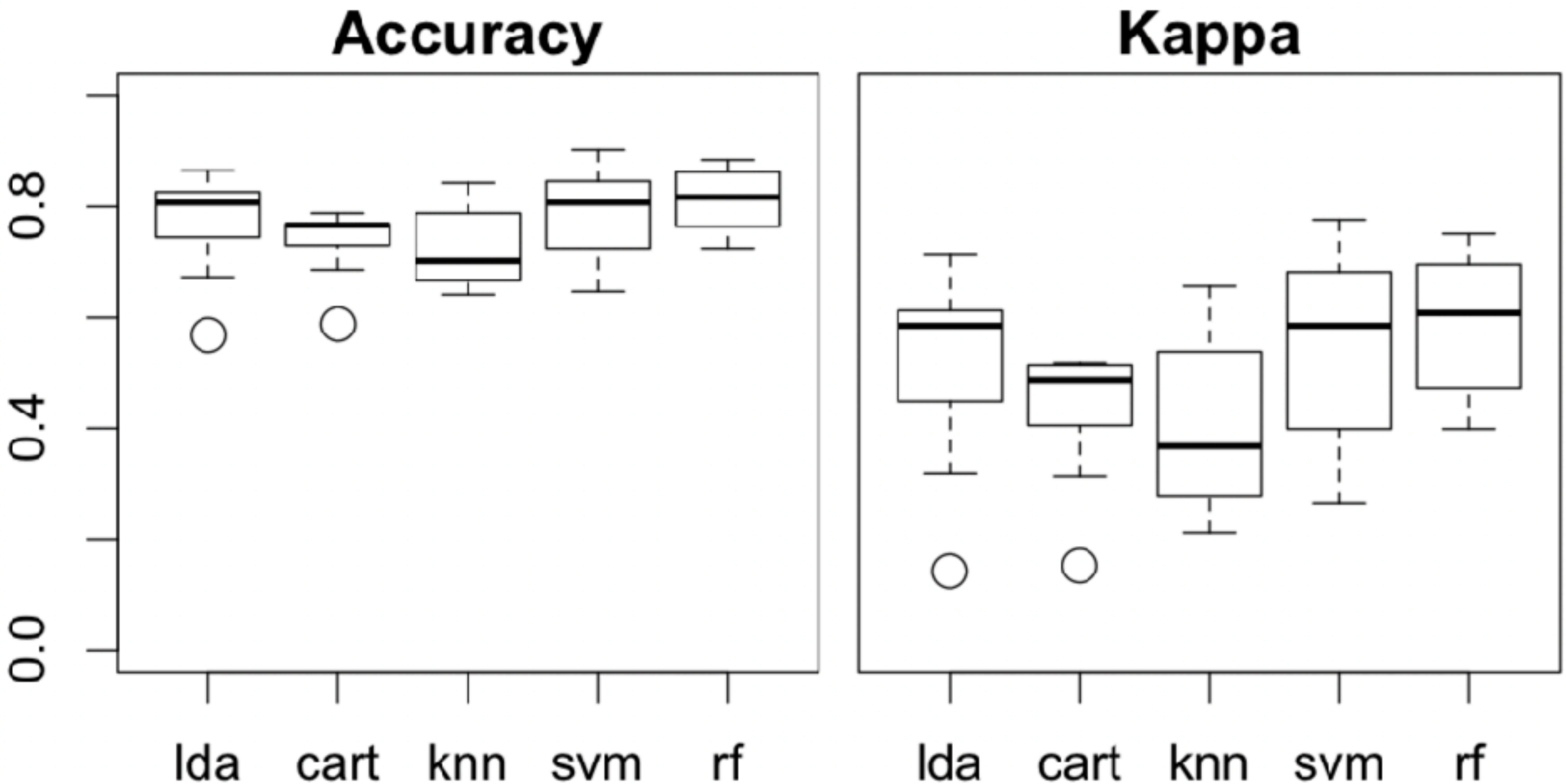
Asking
“do you



Examples of basic level concepts:
Apple, Apricot, Avocado, Carambola
Drum, Flute, Guitar, Harp,
Screwdriver, Shovel, Toothpick, Wrench

Examples of non-basic level concepts:
Dried_fruit, Sultana, Red_delicious, Citrus
Spinet, snare_drum, stringed_instrument
Maul, bucksaw, monkey_wrench, opener

Experimental results:
A comparison of classification algorithms



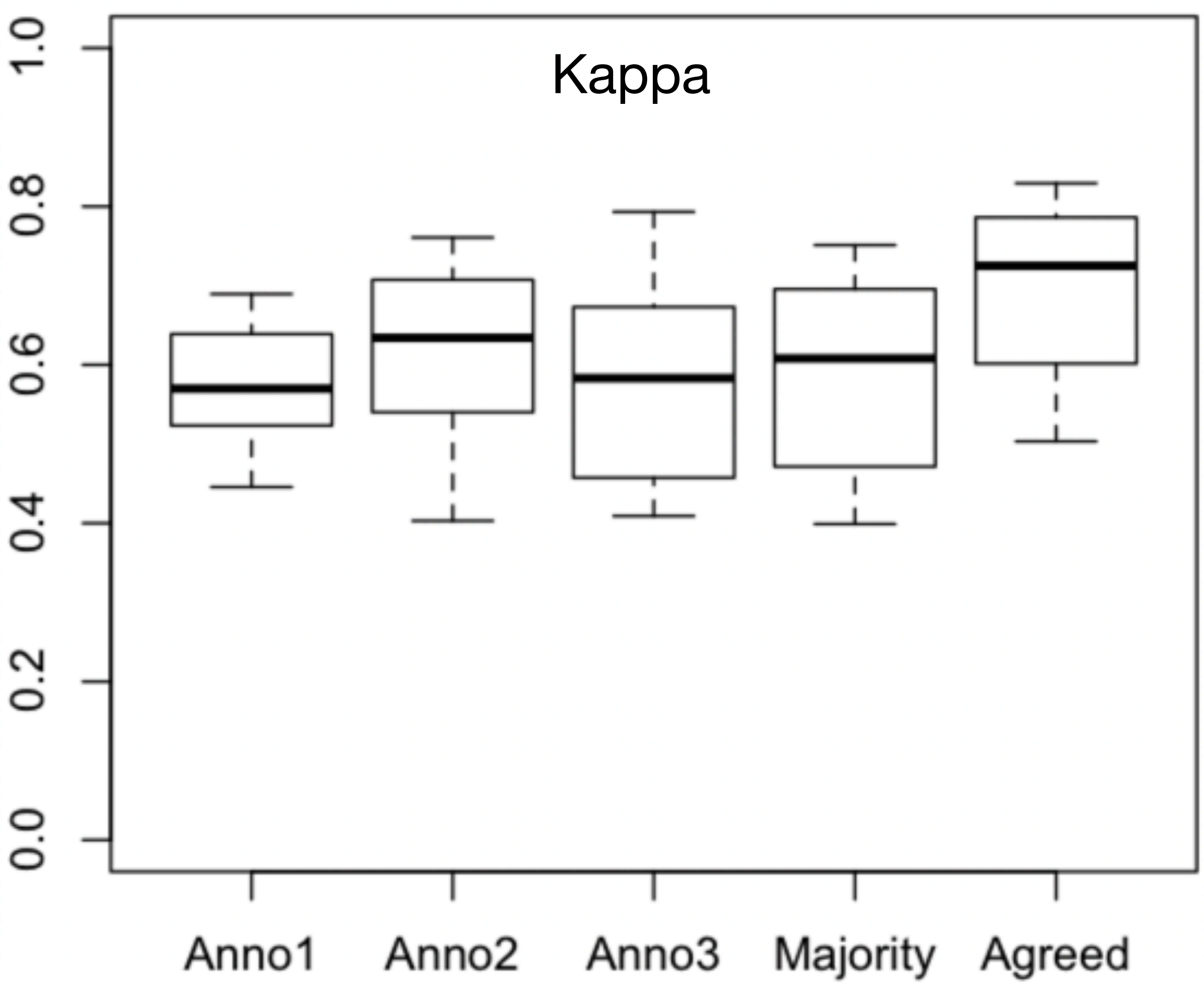
Further experiments will be run using a random forest (RF).

Experimental results:
A comparison to baselines

		Accuracy	Kappa
	Random Forest	0.82	0.61
Manual:	basic level at fixed depth	0.64	0.17
Randomly guessing:	all as basic level	0.36	0.00
	none as basic level	0.64	0.00
	50% as basic level	0.49	-0.02
	36% as basic level	0.54	0.01

Accuracy is not a helpful measure in this case

Experimental results:
A comparison of human annotators



- Raisin, Prune, Dried Apricot

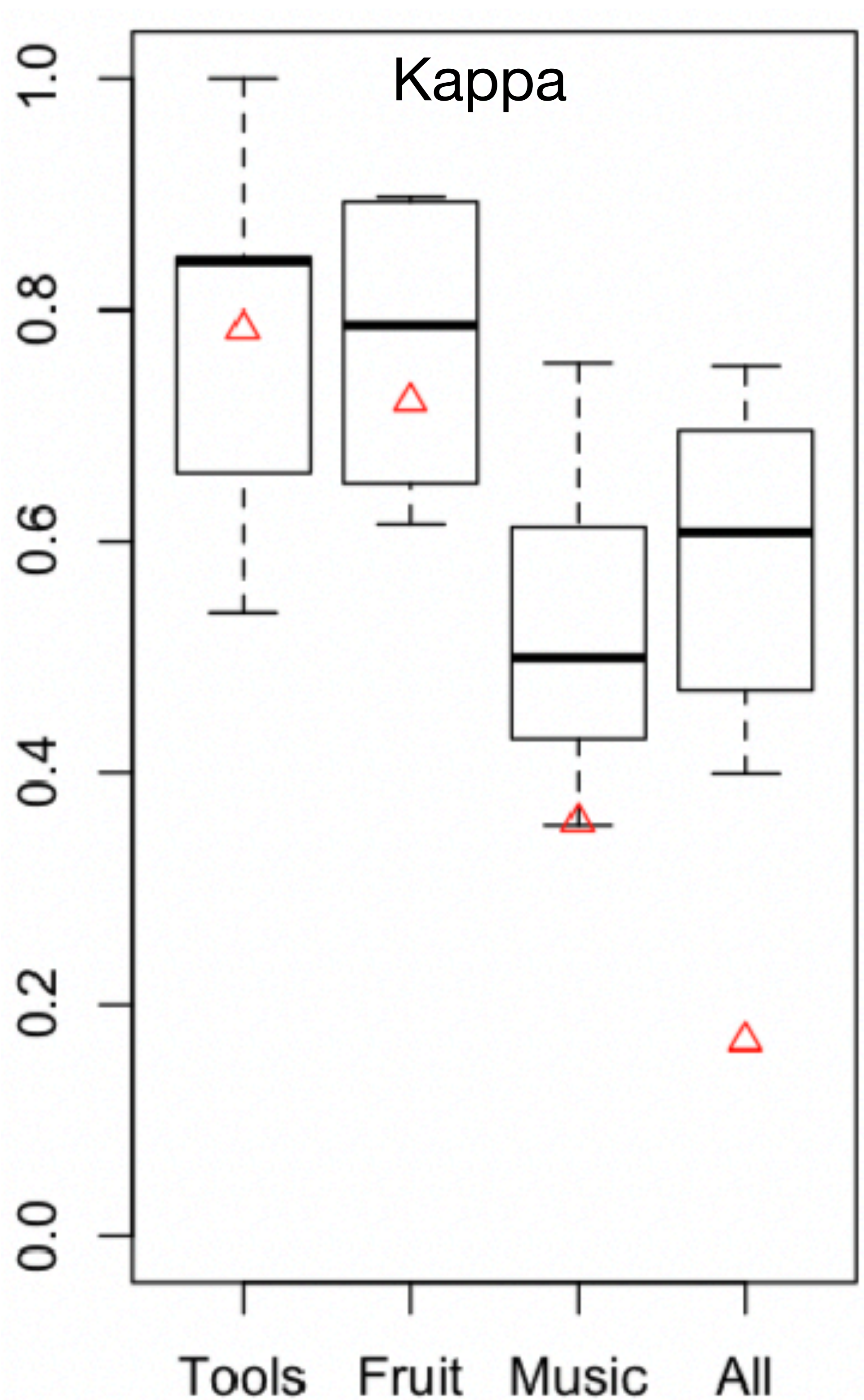
- No significant difference between raters.
- Better performance on concepts on which raters agreed.

Cases that are clear-cut for humans are also easier for machines?

Examples on which raters disagreed:

- Berry
  - Strawberry
  - Blackberry

Experimental results: A comparison of domains



- Training/test set includes concepts from 3 domains:
  - Hand tools
  - Edible fruit
  - Musical Instruments

- Large differences between domains
- Manual method scores reasonably well within one domain.

Cases that are clear-cut for humans are also easier for machines?

Experimental results: Feature importance in the three domains

Feature	All	Tool	Fruit	Music
depth_in_hierarchy	1	1	1	2
G.Ngram_score	2	2	3	3
gloss_length	3	4	4	1
word_length_min	4	5	6	4
polysemy_max	5	3	5	7
nr_of_partOfs	6	8	2	8
nr_of_hyponyms	7	6	8	5
nr_of_synonyms	8	7	7	6
nr_of_direct_hypern.	9	9	9	9

- There are some differences between domains
- All features types are needed

Experimental results: Prediction in a new domain

- What happens when we predict in a new domain, for which we don't have manually labelled examples in the training set?
    - Performance drops.
- Normalisation: divide each feature value by the average feature value ***within the domain***.
    - After per-domain normalisation, performance drop is much smaller.

New domain	Trained on	RF	Manual
Tools	Fruit+Music	0.37	0.02
Fruit	Tools+Music	-0.1	-0.42
Music	Tools+Fruit	0.35	-0.01

Experimental results: Prediction in a new domain

- What happens when we predict in a new domain, for which we don't have manually labelled examples in the training set?
  - Performance drops.
- Normalisation: divide each feature value by the average feature value ***within the domain***.
  - After per-domain normalisation, performance drop is much smaller.

				After normalisation of features:		
New domain	Trained on	RF	Manual	Structural	Lexical	Frequency
Tools	Fruit+Music	0.37	0.02	0.62	0.43	0.34
Fruit	Tools+Music	-0.1	-0.42	0.41	0.06	-0.13
Music	Tools+Fruit	0.35	-0.01	0.32	0.21	0.34

Applying the model knowledge-graph scale

- Applied to WordNet (74k noun synsets)
- With best preforming settings
- How to split a Knowledge Graph up into domains?

Result: 16k basic level concepts (21%)

Available in RDF from:

https://github.com/jrvosse/wordnet-3.0-rdf/tree/master/basiclevels

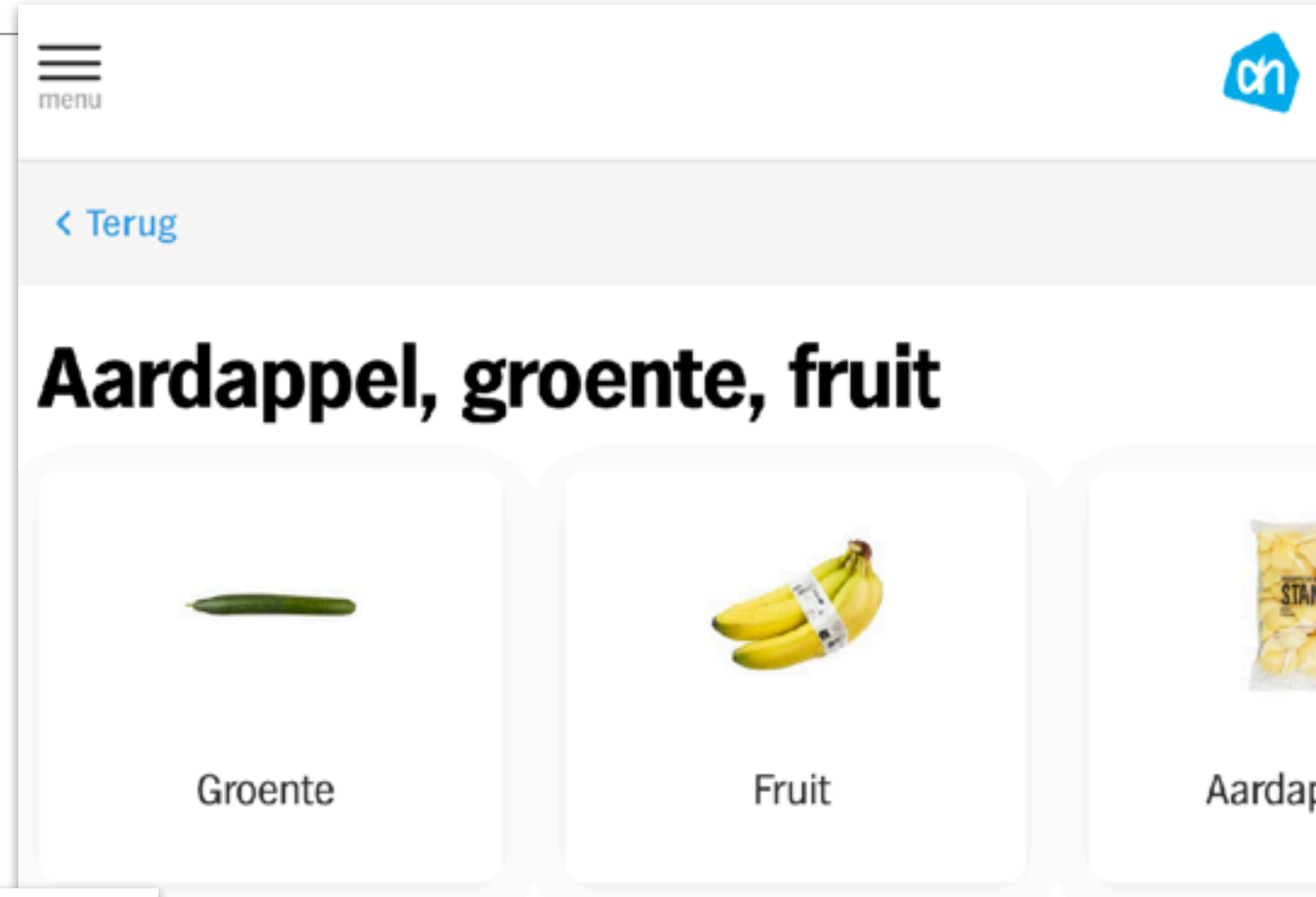
Purpose:

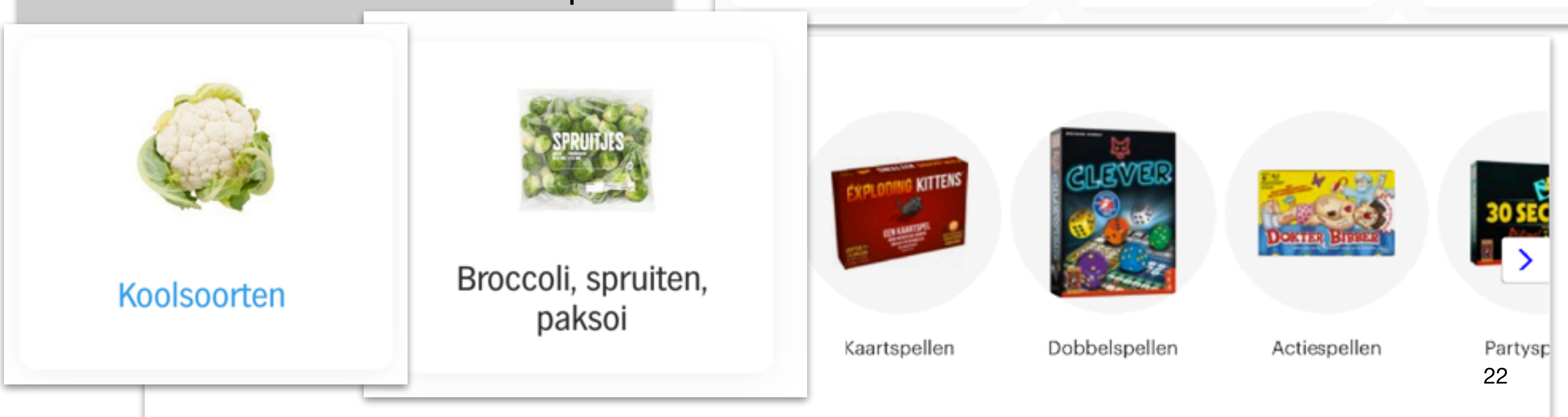
- to enable research into the use of basic level concepts in applications
- for us to further finetune the algorithm
    - e.g. to remove cases where two basic levels are in a hierarchical relation.

Possible use cases

- When displaying categories to customers, which product image should be used to represent a category?
  - Prototype-scores could help!

- Which names to choose for the categories?
  - basic level terms could help.

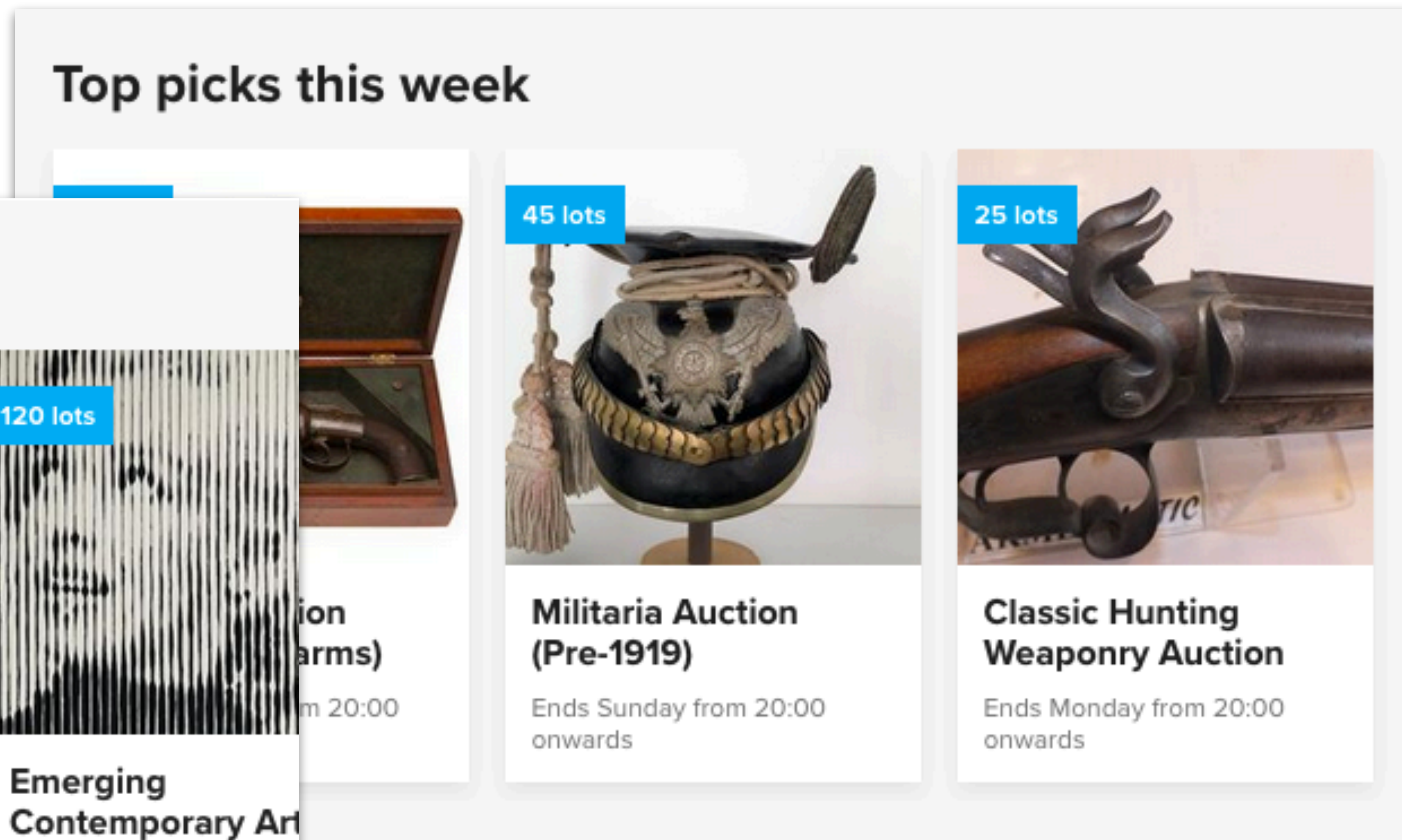


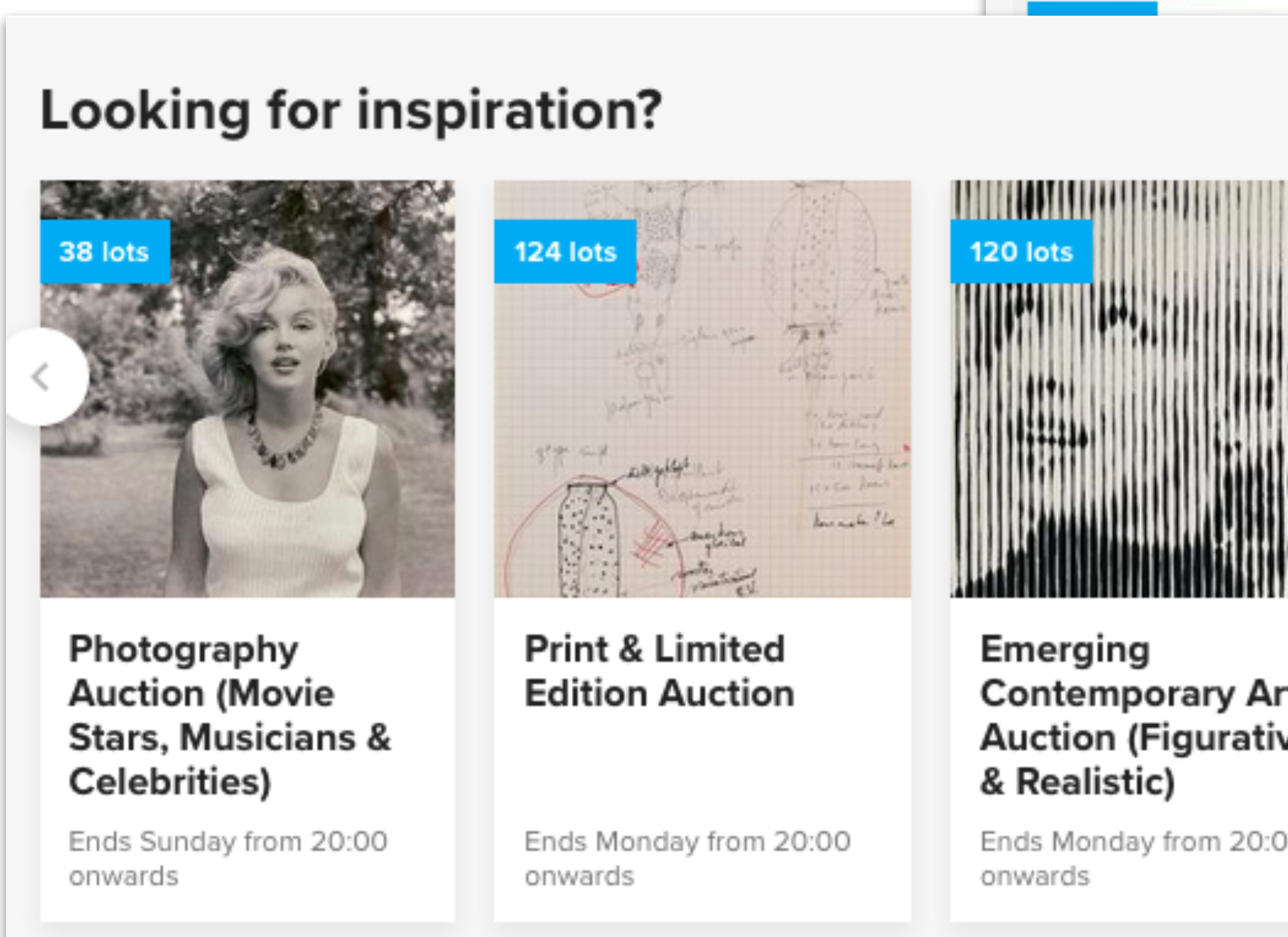


Possible use cases

- When displaying categories to customers, which product image should be used to represent a category?
    - Maybe prototype-scores could help!

- Which names to choose for the categories?
    - basic level terms could help.





Conclusions - where do we stand now?

- We can predict basic level concepts based on human produced data
  - if we have a representative training set.
  - If not, in a new domain, we need per-domain normalisation
    - Domain splitting / “ontology modularisation” is crucial.
  - Open question:
    - Some cases are easy for both humans and machines; some are hard.

    is it true that the easy ones are worth most for an application?

Conclusions - where do we want to go?

- More sources:
  - “basic” text corpora (children’s books, language learning resources)
  - distributional models
  - structure of wikipedia lemmas
  - image repositories (e.g. ImageNet)
- Better training sets
  - larger, e.g. with crowd-sourcing
  - measuring basic level effects instead of asking a rater.
- Wider applicability
  - Test in other knowledge graphs
- Also predict ***prototypes.***

…so that machines can better anticipate human behaviour.