Multimodal Language Learning

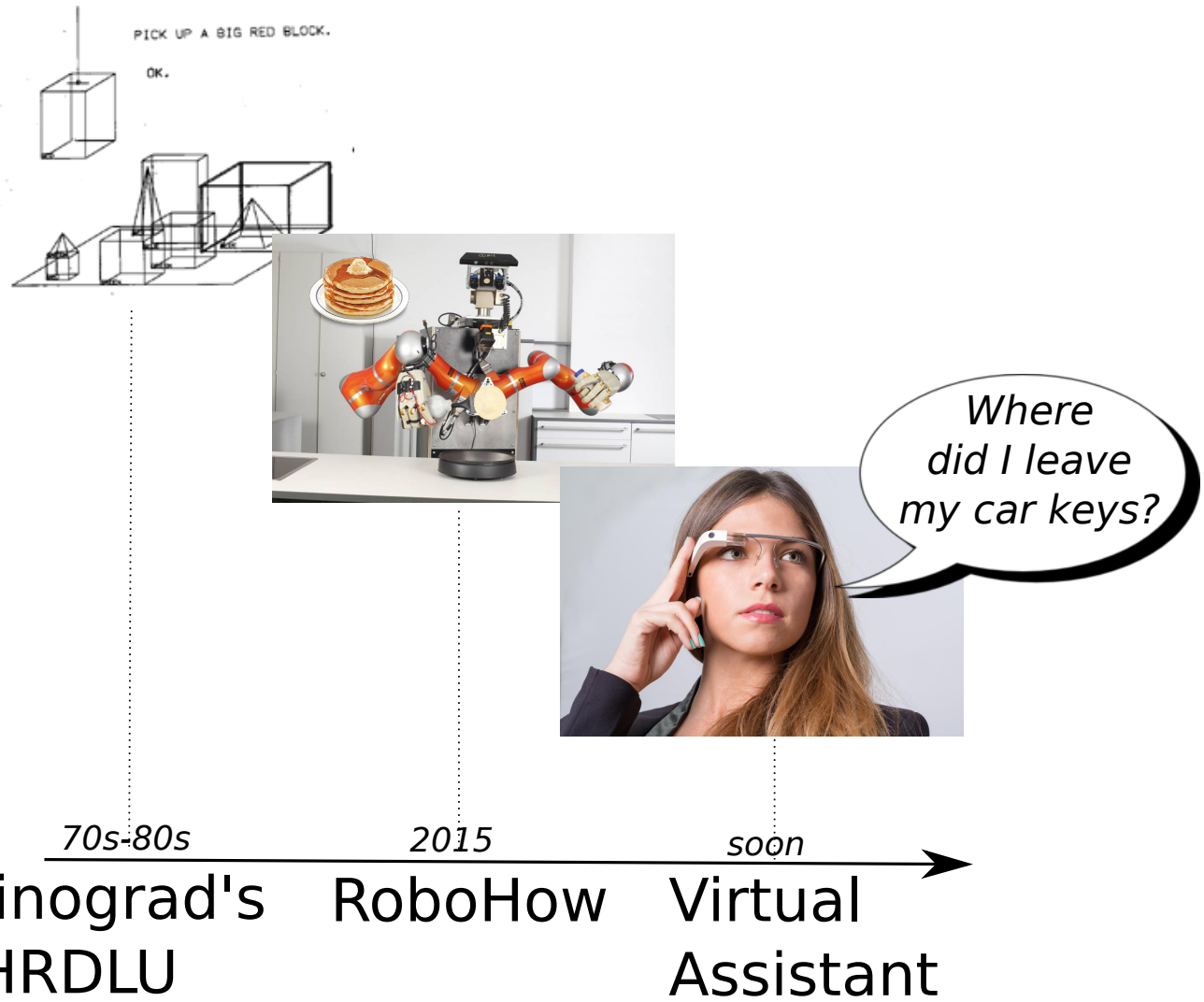
Angeliki Lazaridou

Center for Brain/Mind Sciences
University of Trento

September 14, 2015

The AI Dream

Clever agents interacting with humans



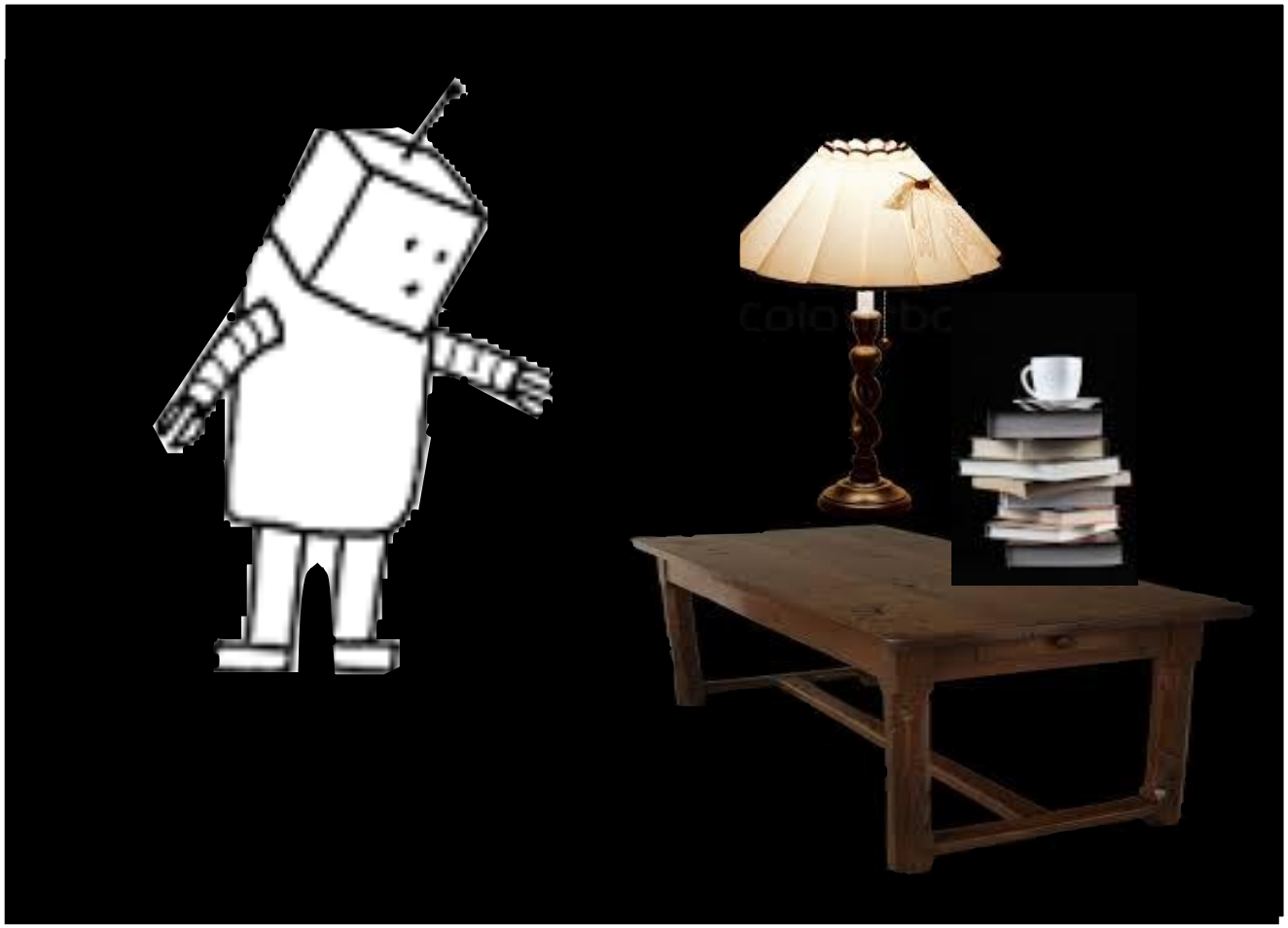
How should these systems learn?

- Interactive Learning
  - Involve human in the learning loop
  - *Pre-school for agents* paradigm where the user explicitly teaches the agents
  - Less detailed supervision via evaluations, feedback ...
- Non-interactive Learning
  - Equip agents with a good starting point

What is a good starting point?

- interacting with humans requires strong NLU
- need of a way to represent meaning
  - Logic-based representations
  - WordNet, FrameNet
  - distributional models

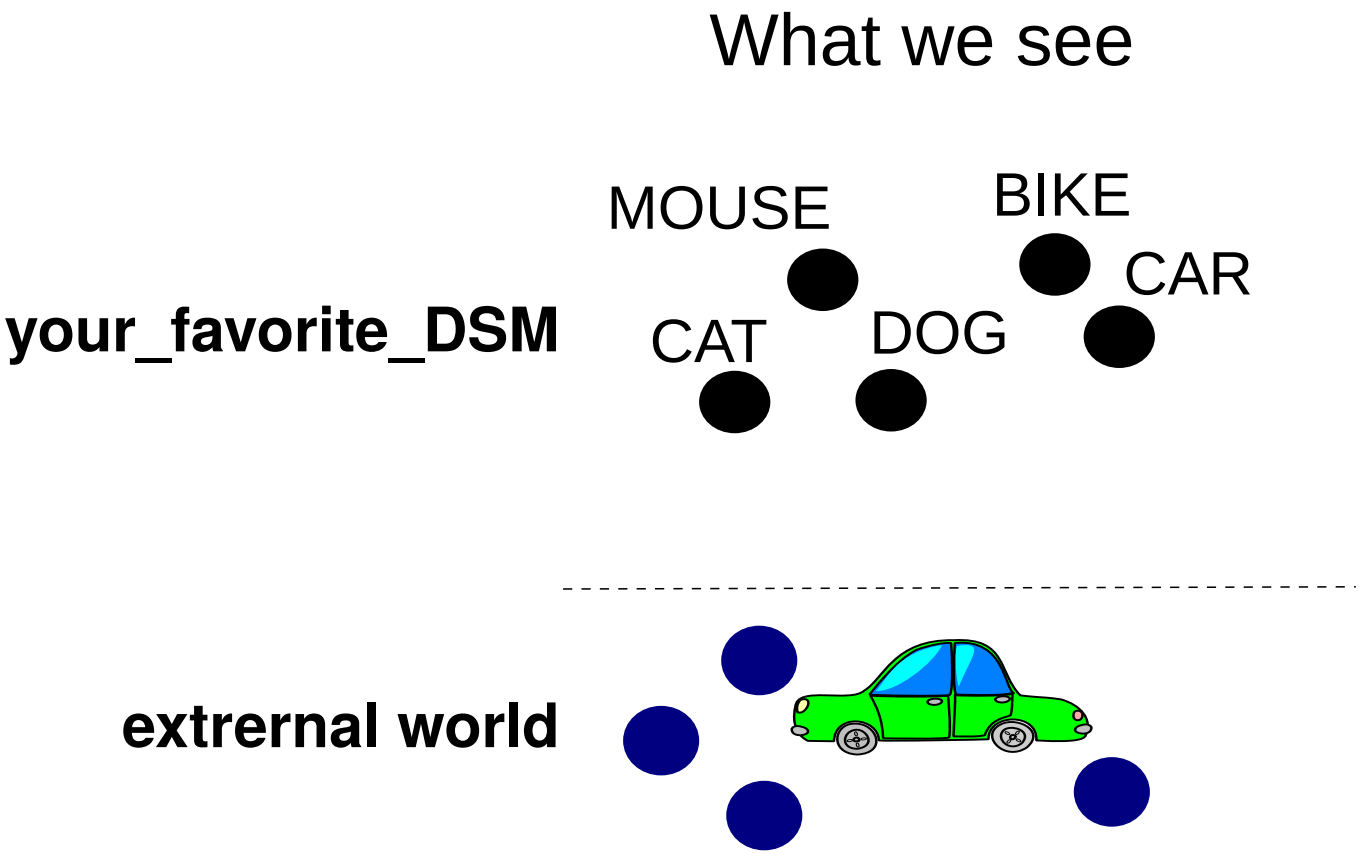
Distributional Semantic Models



Landauer and Dumais (1997), Count-based Models, Bengio et al. (2003), Collobert and Weston (2008), Mikolov et al. (2013), Pennington et al. (2014), Levy and Goldberg (2014), ...

Distributional Semantic Models

All about semantic relations

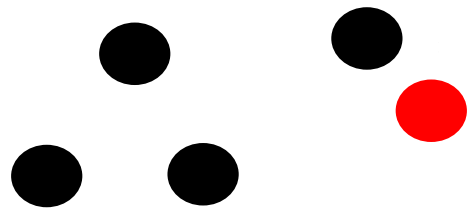


Distributional Semantic Models

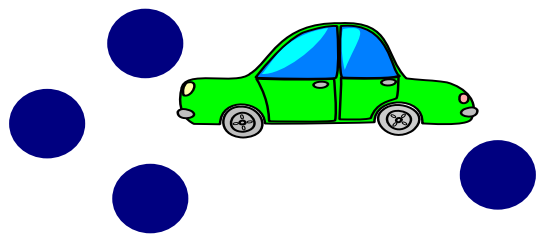
Searle's Chinese Room Metaphor

What DSM see

your_favorite_DSM



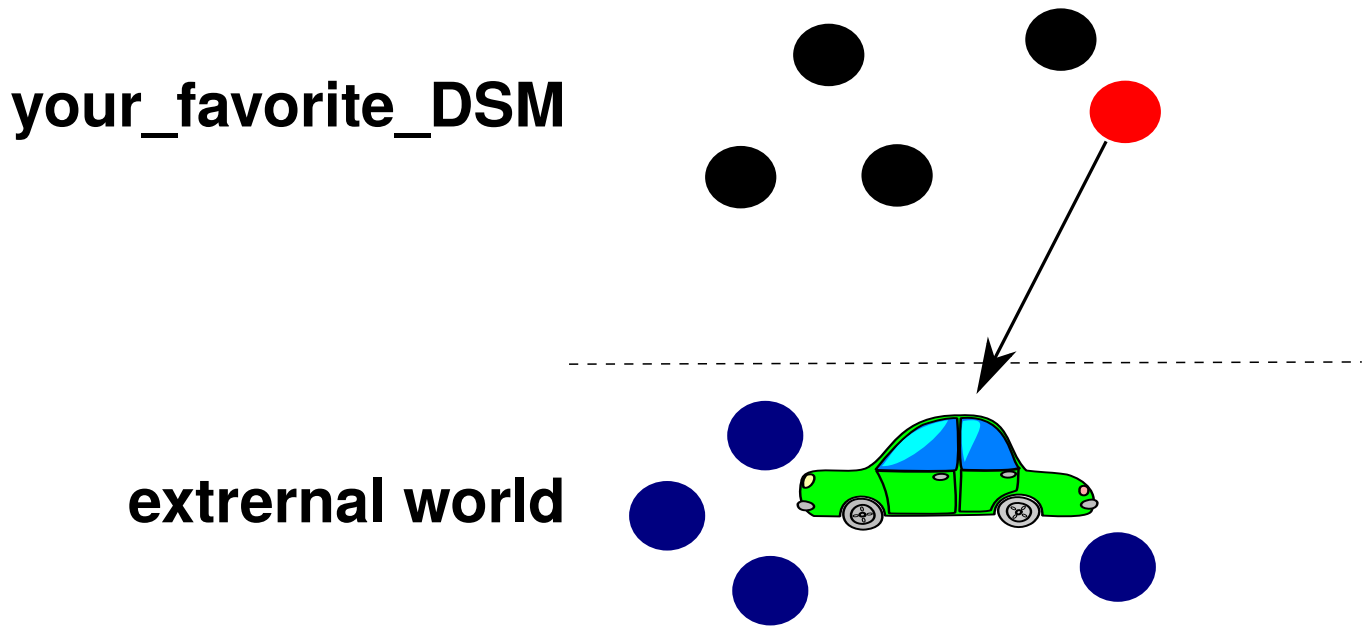
extrernal world



Distributional Semantic Models

All about semantic relations...

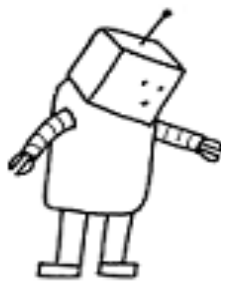
DSM reaching out to the external world?



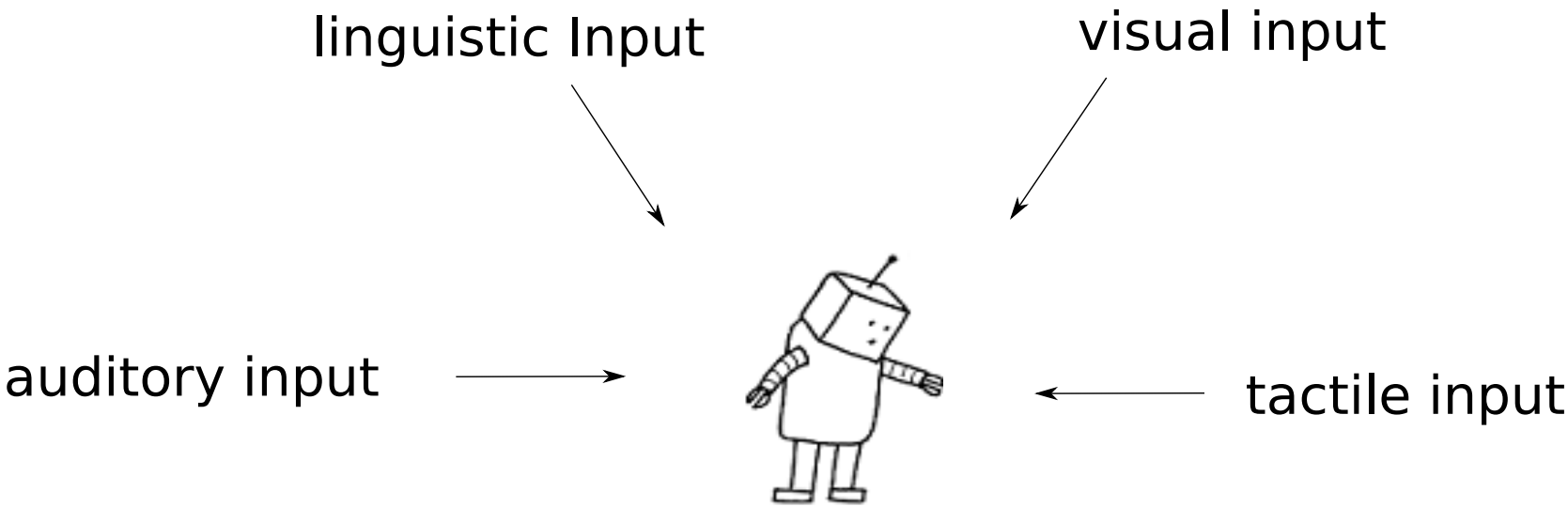
that must be fleshed out in terms of non-linguistic structures!
(Roy, 2008)

How DSM learn the meaning of words

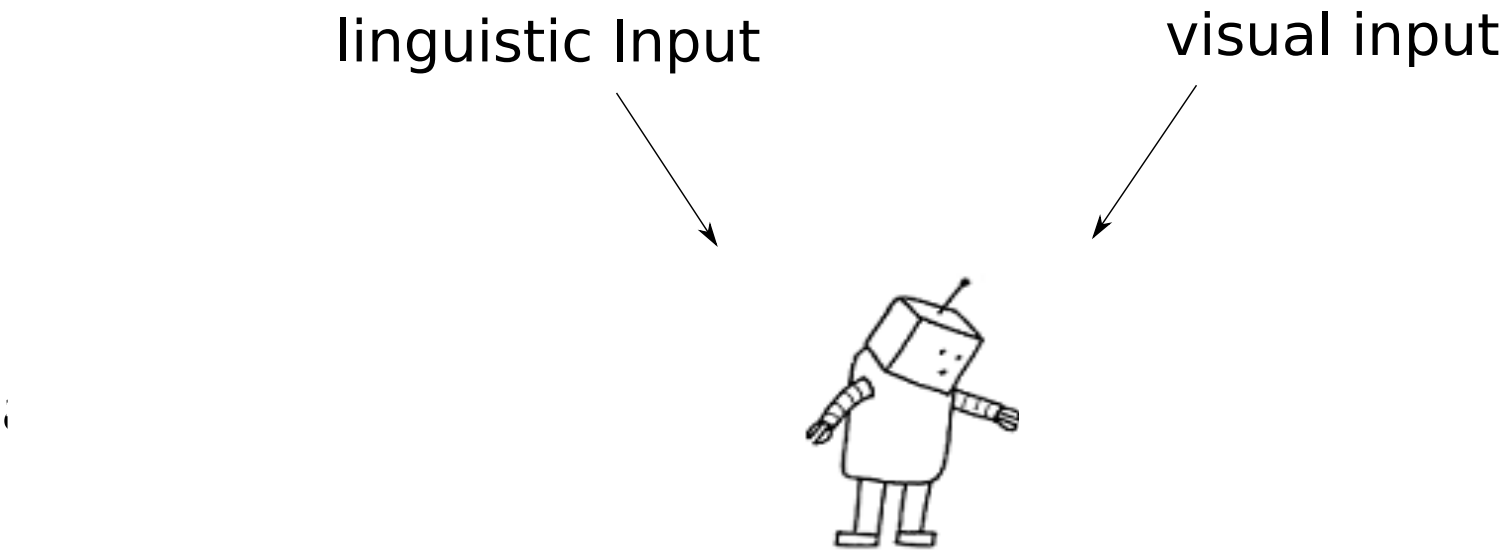
linguistic Input



How DSM **should** learn the meaning of words



This work: Exposing models to multi-modal data



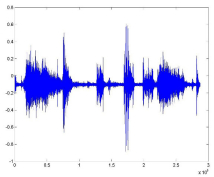
Text+Images: Feng and Lapata (2010), Leong and Mihalcea (2011), Weston et al. (2011), Bruni et al. (2014), Kiela and Bottou (2014)
Text+Feature norms: Howell et al. (2005), Andrews et al. (2009), Johns and Jones (2012), Steyvers et al. (2010), Hill and Korhonen (2014), Silberer and Lapata (2014)
Text+Image tags: Baroni and Lenci (2008), Hill and Korhonen (2014), Young et al. (2014)

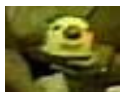
How are we going to implement this?

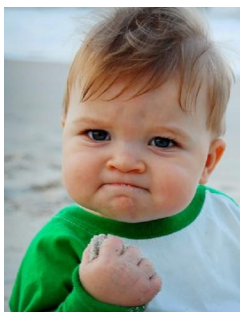
Children learning the meaning of words

Children learn in a cross-situational environment

What the child hears **What the child sees**

audio signal   sequence of scenes

word segmemtation *"Look at the yellow duck!"*  object segmentation



Simulate cross-situational environment

What thel model "sees" **What thel model "hears"**

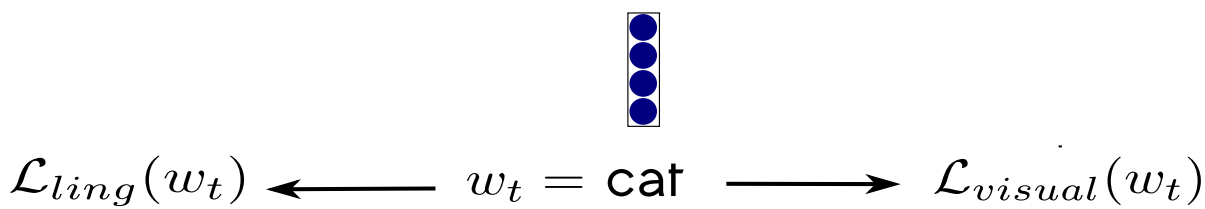
Wikipedia

Cats *are often valued by humans for companionship and their ability to hunt*

ImageNet

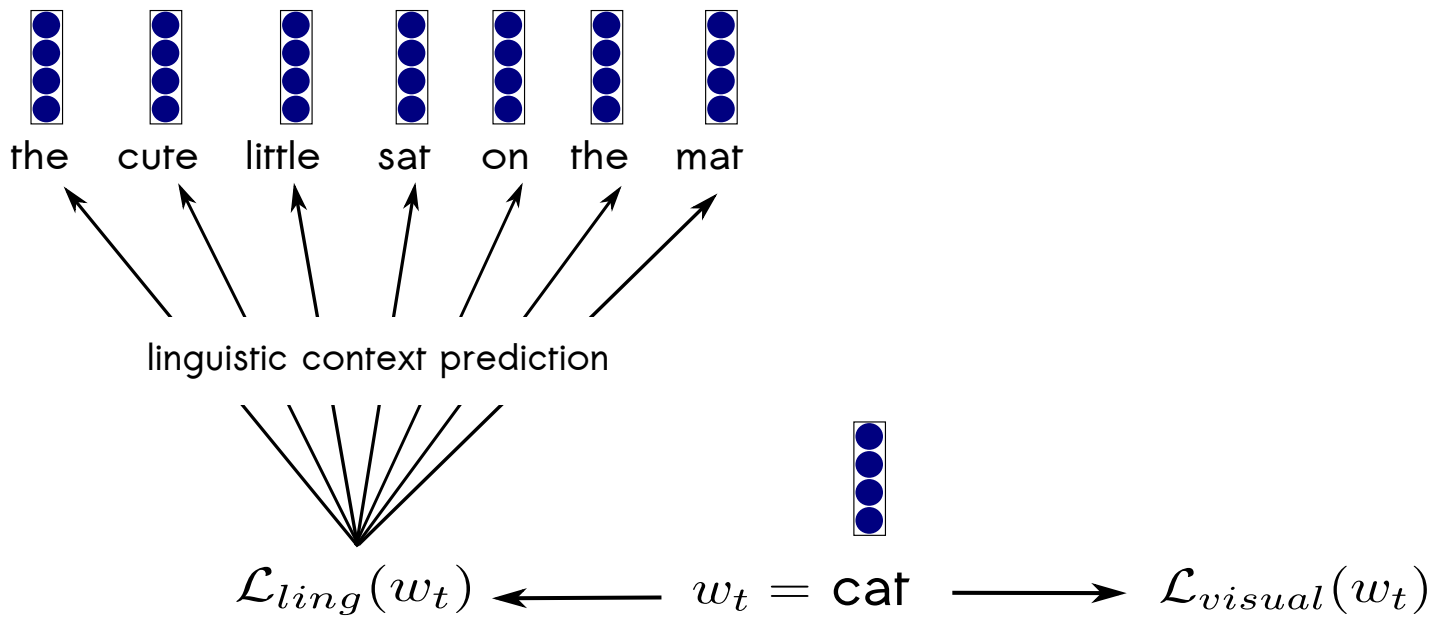
Multimodal Skip-gram Model (aka MSG)

Simple multi-task framework



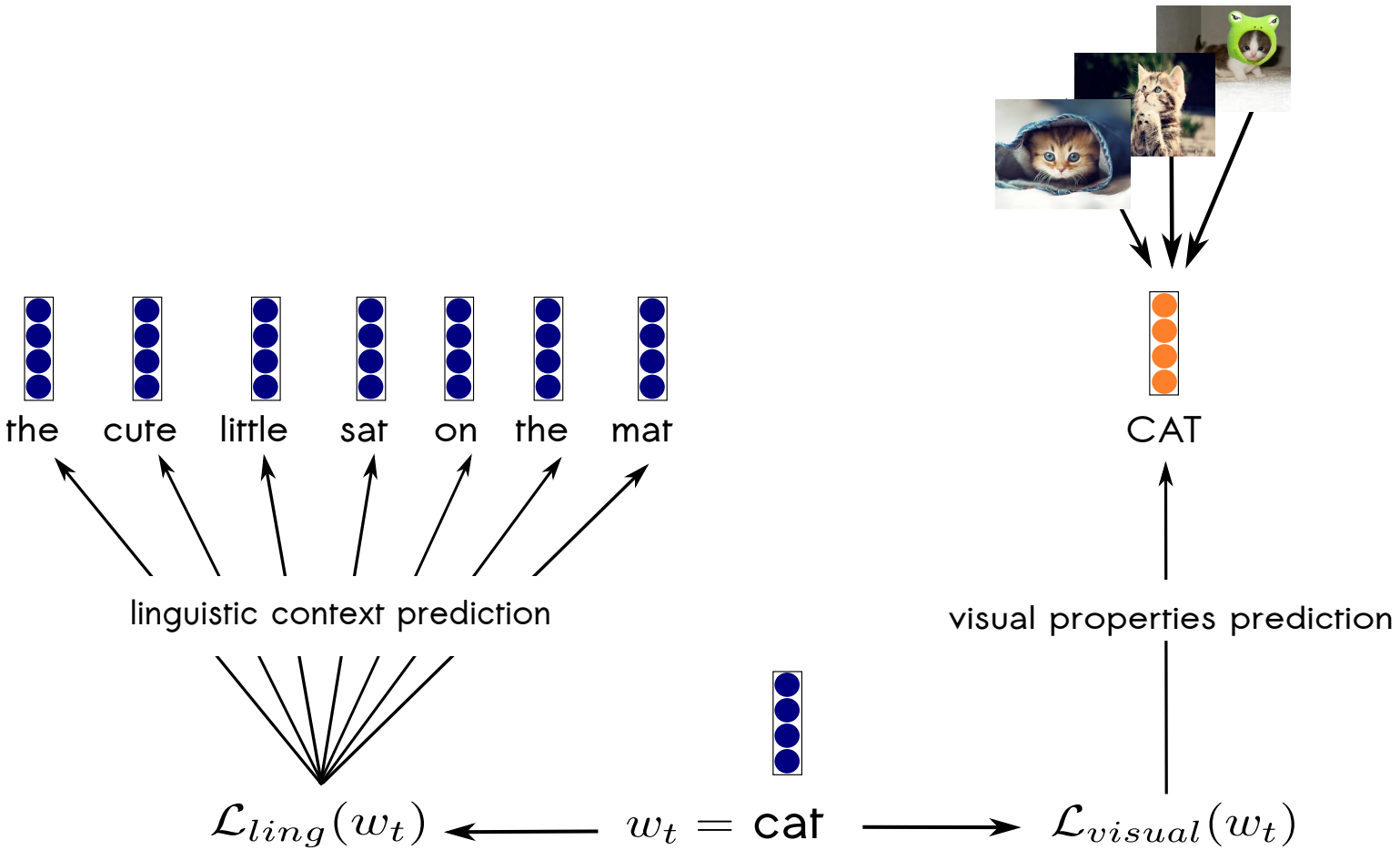
Multimodal Skip-gram Model (aka MSG)

Linguistic objective similar to Skipgram (Mikolov et al., 2013)



Multimodal Skip-gram Model (aka MSG)

For **some** concrete words, also a visual objective that predicts visual features



Properties of MSG

- Information Propagation
  - Grounding of all words (e.g., verbs, *abstract* words)
  - Truly multimodal space
- Incremental learning
  - Dynamic updates – how does each context shift meaning?
  - Support for more principled learning – Curriculum Learning (Bengio et al., 2009)

Approximating human similarity judgments

Figure of merit: Spearman's ρ

	MEN	**Simlex-999**	**SemSim**	**VisSim**
examples	bakery bread	happy cheerful	jeans sweater	donkey horse
Bruni et al.	**0.78**			
Hill et al.		**0.41**		
Silberer and Lapata			0.70	**0.64**
visual vectors	0.62*	0.54*	0.55*	0.56*
linguistic vectors	0.70	0.33	0.62	0.48
multimodal SVD	0.61	0.28	0.65	0.58
multimodal skip-gram	0.75	0.37	**0.72**	0.63

Reaching out to the external world: 0-shot **image retrieval**

DOG= 

CAT = 

HYRAX = ?

Reaching out to the external world: 0-shot **image retrieval**

Search space: 5.1K images with unique labels; percentage precision

	P@1	**P@10**	**P@20**	**P@50**
chance	<0.1	0.2	0.4	1.0
skip-gram/supervised cross-modal mapping	**2.3**	11.9	17.9	30.9
multimodal skip-gram/direct retrieval	2.0	**14.1**	**20.1**	**33.0**

Nearest visual neighbours of abstract words

freedom

theory

wrong

god

together

place

Subjects' significant preference for true neighbour over confounder:

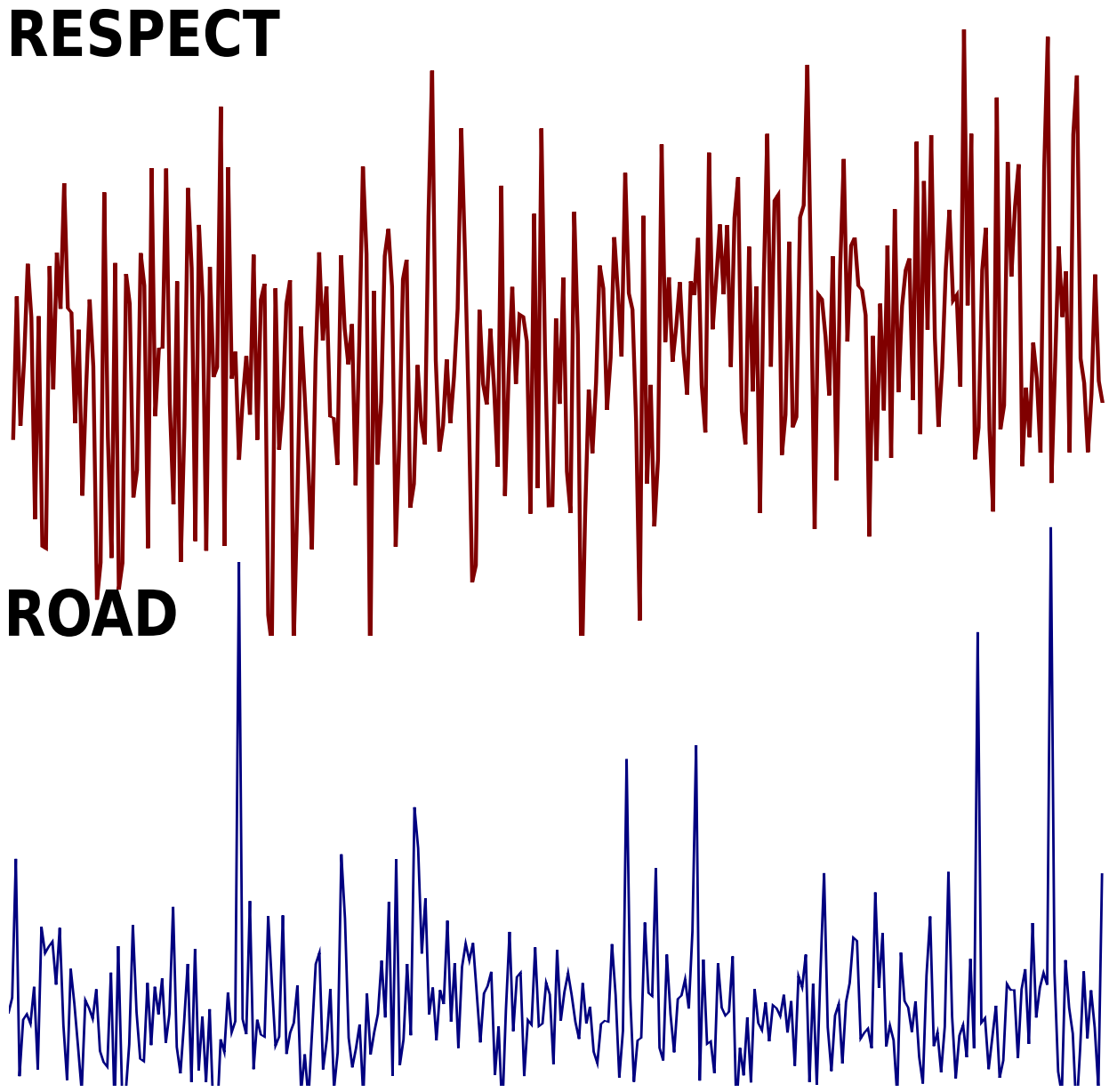
random level: 0%

unseen abstract: 23%

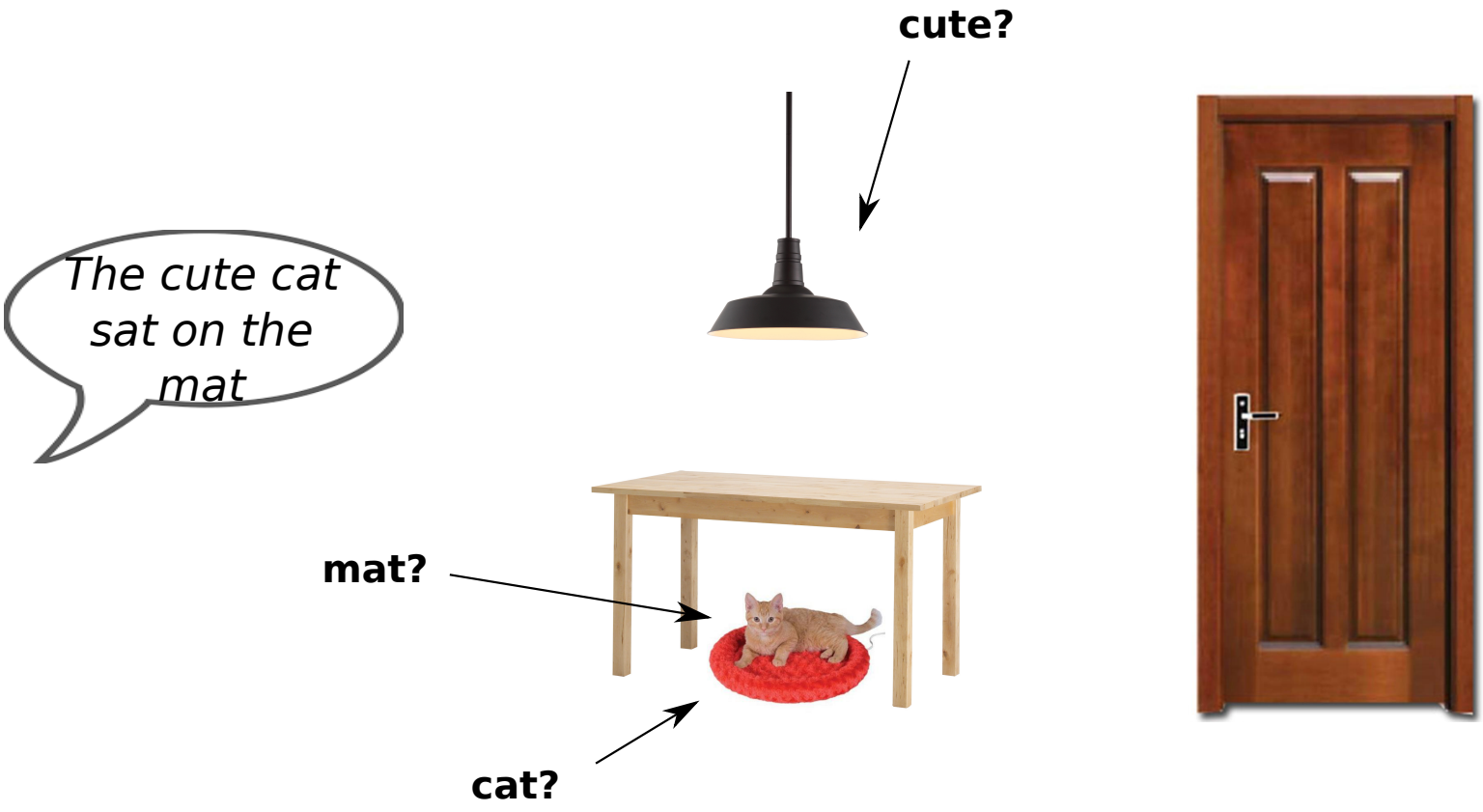
unseen concrete: 53%

Abstractness correlates with MSG entropy

$\rho > 0.7$ on Kiela et al. ACL 2014 data set, no correlation for skip-gram vectors!



Referential Uncertainty



The dream that never came true

CHILDES Video

The Frank corpus

http://langcog.stanford.edu/materials/nipsmaterials.html

```
*mot    let me have that
%ref:   RING
*mot    ahhah whats this
%ref:   RING HAT
*mot    what does mom look like with the hat on
%ref:   RING HAT
*mot    do i look pretty good with the hat on
%ref:   RING HAT
*mot    hmm
%ref:   RING HAT
*mot    hmm
%ref:   RING HAT
*mot    do i look pretty good
%ref:   RING HAT
*mot    peekaboo
%ref:   RING HAT
```

The Frank corpus

Our version

let me have that

ahhah whats this


what does mom look like with the hat on


do i look pretty good with the hat on


hmm


The model

cat mat MAT CAT

$$\frac{1}{T}\sum_{t}\left(\mathcal{L}_{ling}(w_t)+\mathcal{L}_{vision}(w_t)\right) \quad (1)$$

$$\mathcal{L}_{vision}(w_t)=\sum_{i} score(w_t, I_i) \quad (2)$$

score: some loss function, here max-margin+cosine

Simple attention

- force the model learn meaningful word-object associations
- force the model to avoid preferring uniform word-object matchings
  - cat mat MAT CAT

	Model A		Model B	
	CAT	MAT	CAT	MAT
cat	1.0	1.0	1.5	0.5
mat	1.0	1.0	0.5	1.5

$$\mathcal{L}_{vision}(w_t) = \sum_i \alpha(w_t, l_i) * score(w_t, l_i)$$

$$\alpha(w_t, l_i) = \frac{exp(score(w_t, l_i)}{\sum_{t'} exp(score(w_{t'}, l_i))} \quad (3)$$

Matching words with objects

36 test words, 17 test objects

Model	*Best F*
MSG+shuffled visual vectors	.53
MSG+shuffled order	.55
MSG	.75
MSG+attention	.78
BEAGLE	.55
PMI	.53
Bayesian CSL	.54
(BEAGLE+PMI	.83)

BEAGLE, PMI: Kievit-Kylar et al. CogSci 2013
Bayesian CSL: Frank et al. NIPS 2007

Generalization to new instances of learned objects

look at the kitty!

look at the oink!

MSG object identification after a single exposure

The case of fast-mapping

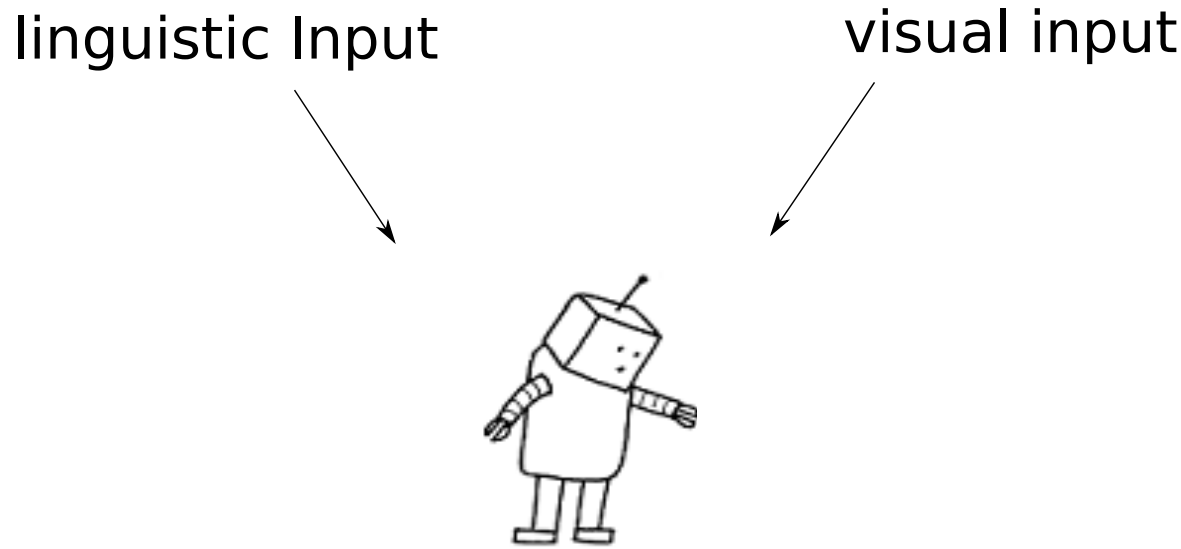
word	*gold object*	*17 objects*	*5K objects*
bunny	bunny	bunny	hare
cows	cow	cow	heifer
duck	duck	hand	chronograph
duckie	duck	hand	chronograph
kitty	kitty	kitty	kitten
lambie	lamb	lamb	lamb
moocows	cow	pig	bison
rattle	rattle	hand	invader

Reasoning from minimal exposure

*The **wampimuk** stared at me with its big scared eyes.*
With my sudden move, it started climbing on the tree.
It had now completely hid in the trees...



Non-interactive Multimodal Language Learning



- better datasets of language+vision, e.g., movies
  - "natural" spoken data → akin to situated dialogue in real life
  - attention mechanism → here-and-now reference (*Oh, look at this book*)
  - memory component → absent reference (*She gave me a book yesterday*)
  - temporal dimensions of videos → grounding of actions
- one step closer to realistic simulations, still further away from real learning

Interactive Language Learning

Learning by interacting with environment or users

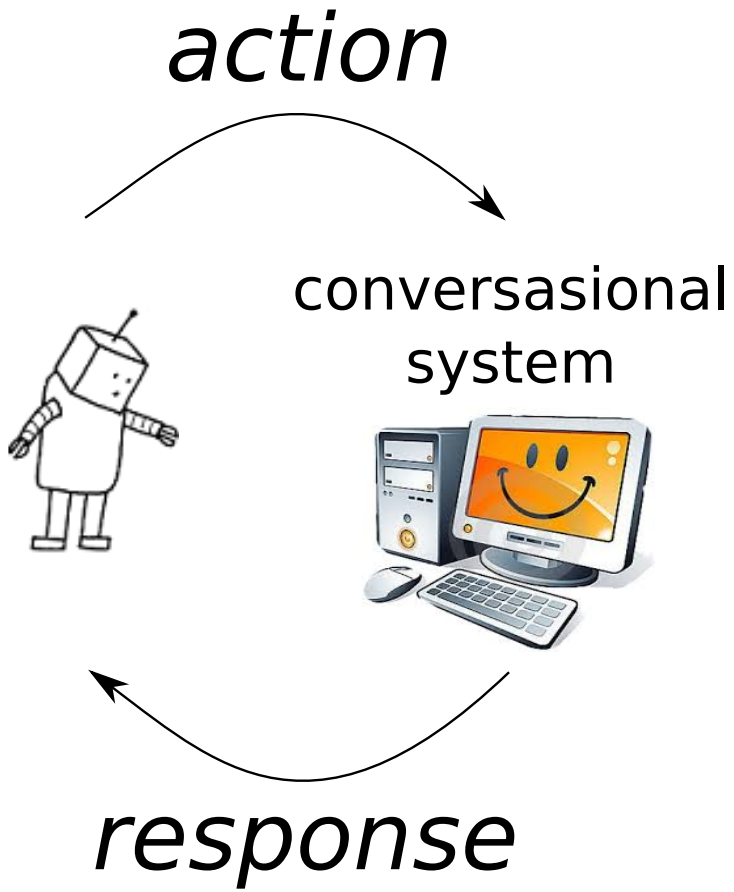
- Artificial intelligence: Learning to see and **act**[1]
- Learning to play games: Deep Q-network (Mnih et al., 2015)
- Example: Goal-drive learning of COMPOSITION FUNCTIONS
  - Given some instructions in NL, learn how to understand AND COMPOSE the instructions, by executing them
  - If successful, positive reward.

[1] Schölkopf, Nature 2015

Interactive Language Learning

Obtaining reward signal

- Problem Potentially very expensive using humans interacting with agents
- Solution Simulate humans



- don't adequately keep track of state of the world

+ can pick up the statistical regularities of the trained corpora

Interactive Multimodal Language Learning

Reward by synthesizing a scene (in collaboration with the user) given a short story

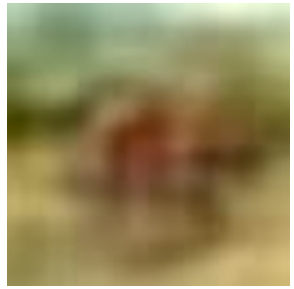
Alice brought all her toys to play with
Bob. After a while, Bob took some of her
toys and started playing alone.
Alice was really upset!

A modern block world: Abstract Scenes (Zitnick and Parikh, 2013)

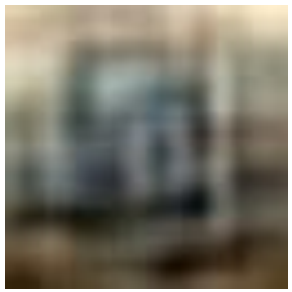
Multimodal Learning is fun: Computational Imagery

(Cherry-picked :-)) Examples

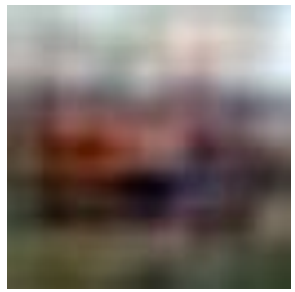

flamingo




camel



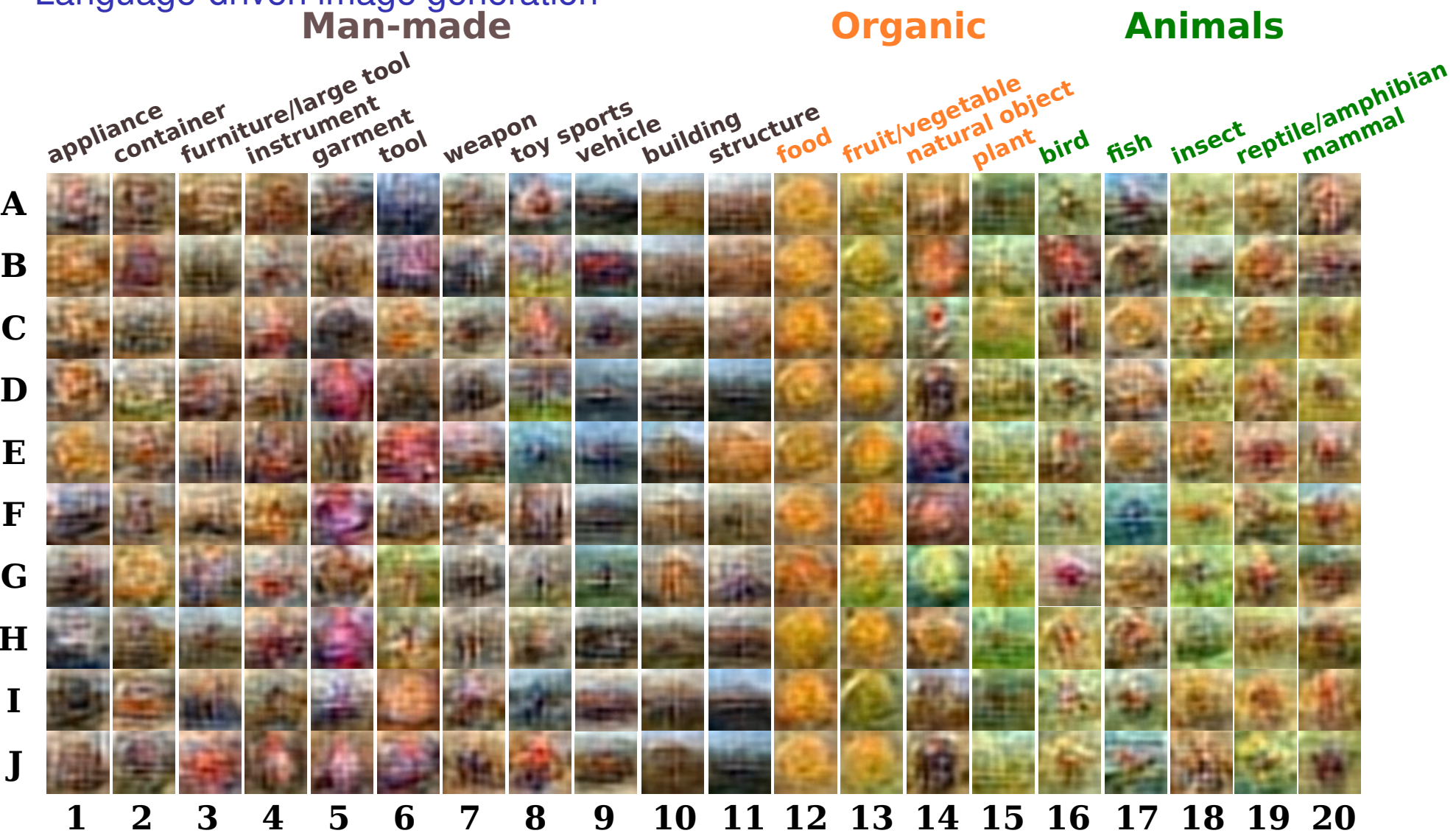

telephone




ambulance

Multimodal Learning is fun: Computational Imagery

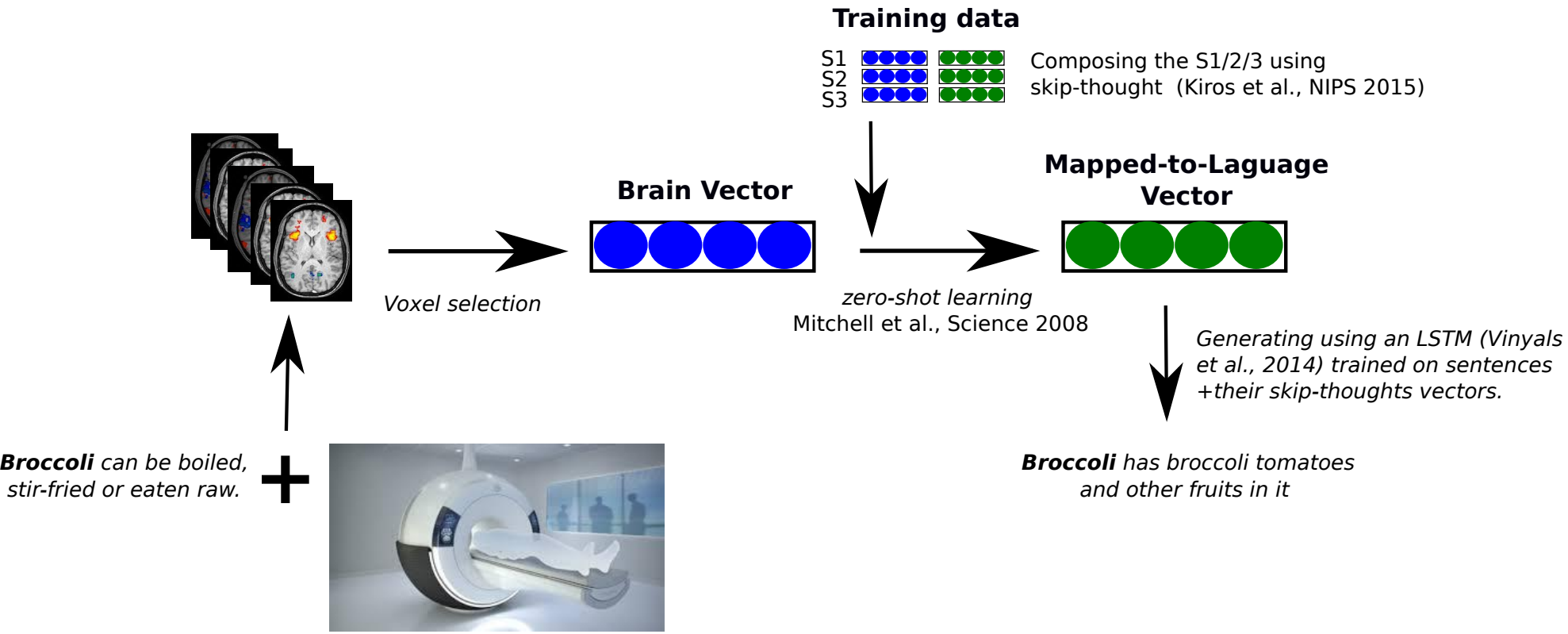
Language-driven image generation



Applications: Automatic book illustration, computational art (images from poems)

Multimodal Learning is fun: From *thoughts* to *words*

Reconstructing text from brain activation



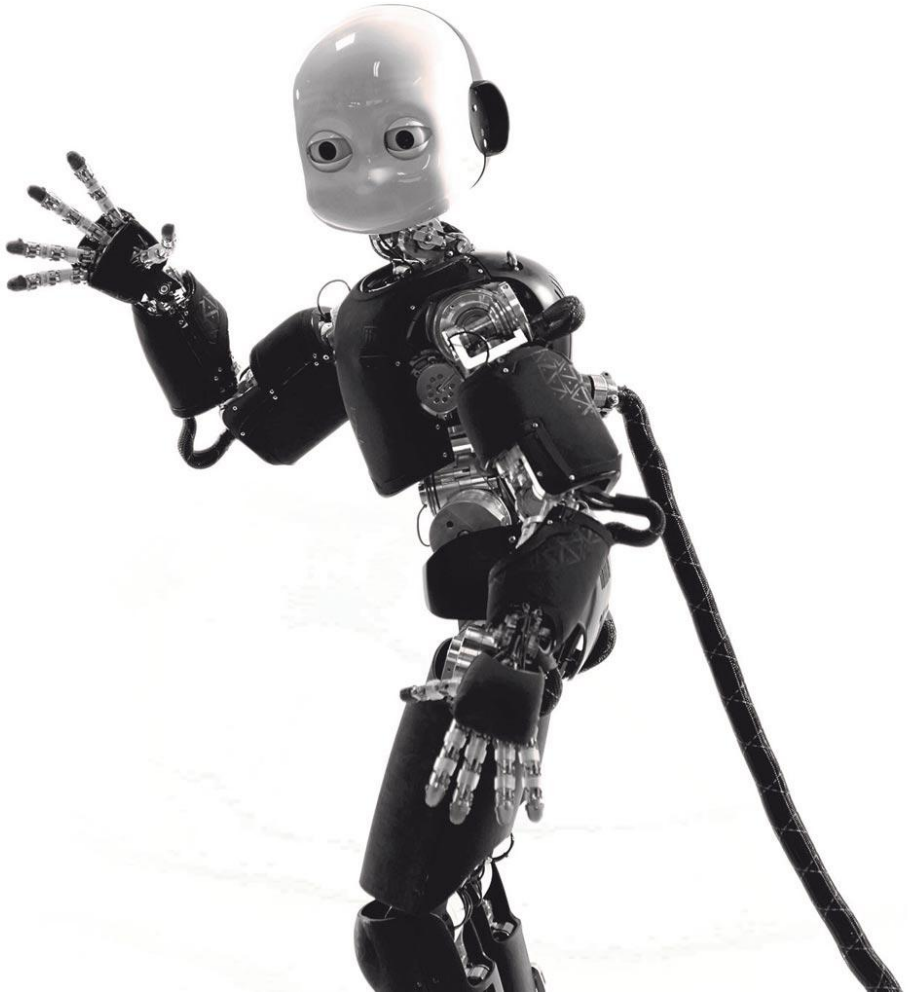
Multimodal Learning is fun: From *thoughts* to *words*

Examples

stimulus: Horses have been used for draft work, travel and entertainment
generated from:
– language: Horses have been used for entertainment center and other food
– brain: It may be used to be a horse or museum

stimulus: Broccoli can be boiled, stir-fried or eaten raw
generated from:
– language: Broccoli can be eaten broccoli and nuts
– brain: Broccoli has broccoli tomatoes and other fruits in it

stimulus: Arson can be done to cause damage to others or collect insurance
generated from:
– language: Baseball uniform can be used to be some baseball uniform and fire hydrant
– brain: A fire hydrant is a fire hydrant is being used as the safety



Thanks ...

...Nghia The Pham

...Marco Marelli

...Dat Tien Nguyen

...Raffaela Bernardi

...Raquel Fernadez

...Grzegorz Chrupala

...Francisco Pereira

...Marco Baroni