Multimodal and cross-modal distributional semantics

Towards a common semantic space for words and things

Marco Baroni

Center for Mind/Brain Sciences
University of Trento
http://clic.cimec.unitn.it/marco

Dialogue 2014

In collaboration with

Elia Bruni

Angeliki Lazaridou

and Jasper Uijlings, Andrew Anderson, Georgiana Dinu, Gemma Boleda, Nam Khanh Tran, Giang Binh Tran, Adam Liska, Alessandro Lopopolo, . . .

Outline

Distributional semantics

Grounding with multimodal distributional semantics

Linking words and things by cross-modal mapping

The distributional hypothesis

Harris, Charles and Miller, Firth, Wittgenstein? ...

The meaning of a word is (can be approximated by, derived from) the set of contexts in which it occurs in texts

We found a little, hairy wampimuk sleeping behind the tree

See also MacDonald & Ramscar CogSci 2001

Distributional semantics

Landauer and Dumais PsychRev 1997, Turney and Pantel JAIR 2010, ...

he curtains open and the moon shining in on the barely
ars and the cold , close moon " . And neither of the w
rough the night with the moon shining so brightly , it
made in the light of the moon . It all boils down , wr
surely under a crescent moon , thrilled by ice-white
sun , the seasons of the moon ? Home , alone , Jay pla
m is dazzling snow , the moon has risen full and cold
un and the temple of the moon , driving out of the hug
in the dark and now the moon rises , full and amber a
bird on the shape of the moon over the trees in front
But I could n't see the moon or the stars , only the
rning , with a sliver of moon hanging among the stars
they love the sun , the moon and the stars . None of
the light of an enormous moon . The plash of flowing w
man 's first step on the moon ; various exhibits , aer
the inevitable piece of moon rock . Housing The Airsh
oud obscured part of the moon . The Allied guns behind

Distributional semantics

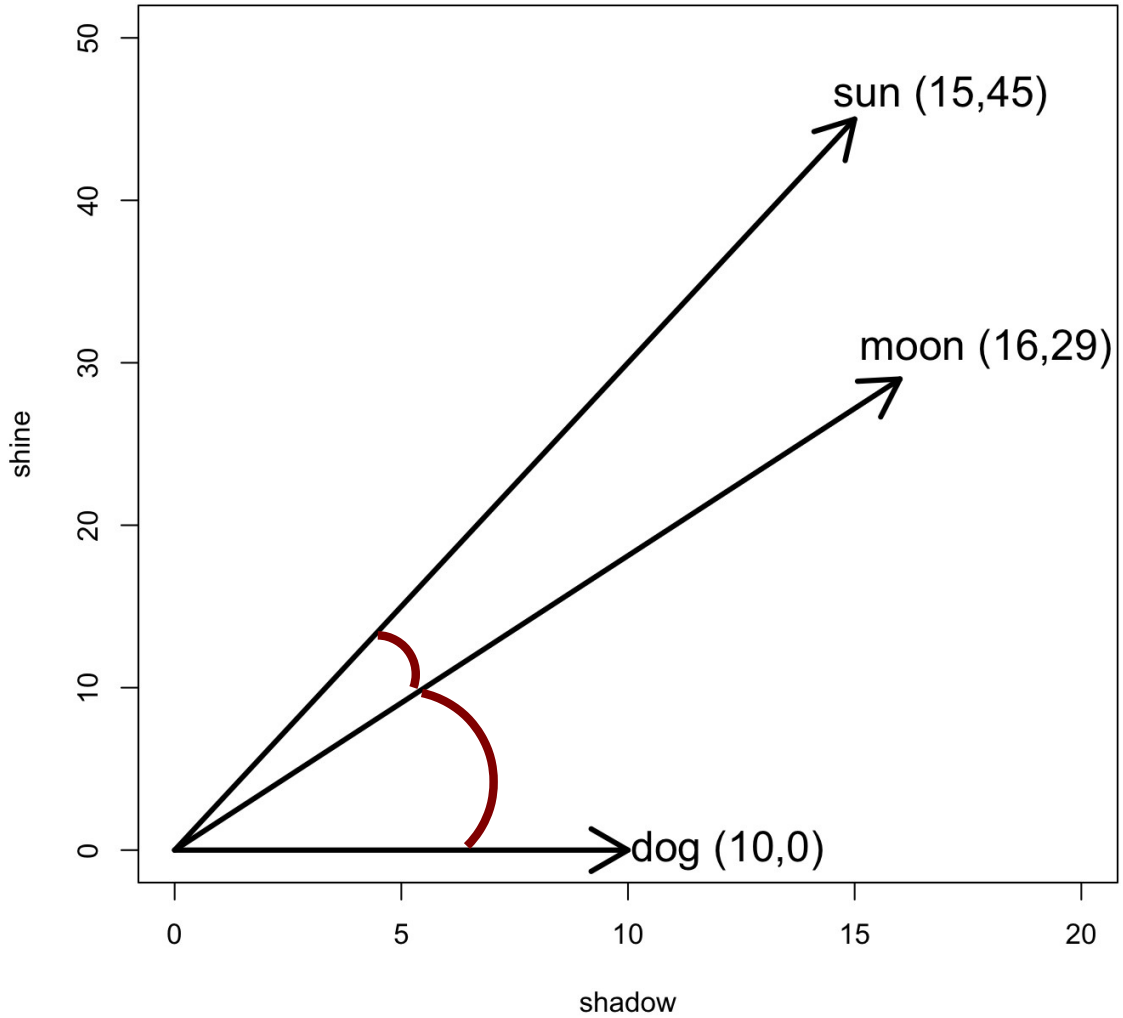
Distributional meaning encoded in co-occurrence vectors

	planet	night	full	shadow	shine	crescent
moon	10	22	43	16	29	12
sun	14	10	4	15	45	0
dog	0	4	2	10	0	0

Distributional semantics

The geometry of meaning

	shadow	shine
moon	16	29
sun	15	45
dog	10	0



Geometric neighbours $\approx$ semantic neighbours

rhino	fall	good	sing
woodpecker	rise	bad	dance
rhinoceros	increase	excellent	whistle
swan	fluctuation	superb	mime
whale	drop	poor	shout
ivory	decrease	improved	sound
plover	reduction	perfect	listen
elephant	logarithm	clever	recite
bear	decline	terrific	play
satin	cut	lucky	hear
sweatshirt	hike	smashing	hiss

Distributional semantics: A general-purpose representation of lexical meaning

Baroni and Lenci 2010

- Similarity (*cord-string* vs. *cord-smile*)
- Synonymy (*zenith-pinnacle*)
- Concept categorization (*car* ISA *vehicle*; *banana* ISA *fruit*)
- Selectional preferences (*eat topinambur* vs. **eat sympathy*)
- Analogy (*mason* is to *stone* like *carpenter* is to *wood*)
- Relation classification (*exam-anxiety* are in CAUSE-EFFECT relation)
- Qualia (TELIC ROLE of *novel* is *to entertain*)
- Salient properties (*car-wheels*, *dog-barking*)
- Argument alternations (*John broke the vase - the vase broke*, *John minces the meat - *the meat minced*)

Selectional preferences in semantic space

Padó et al. EMNLP 2007

To kill...

object	*cosine*	*with*	*cosine*
kangaroo	0.51	hammer	0.26
person	0.45	stone	0.25
robot	0.15	brick	0.18
hate	0.11	smile	0.15
flower	0.11	flower	0.12
stone	0.05	antibiotic	0.12
fun	0.05	person	0.12
book	0.04	heroin	0.12
conversation	0.03	kindness	0.07
sympathy	0.01	graduation	0.04

Outline

The grounding problem

Searle 1984, Harnad Physica 1990

The Chinese vector room

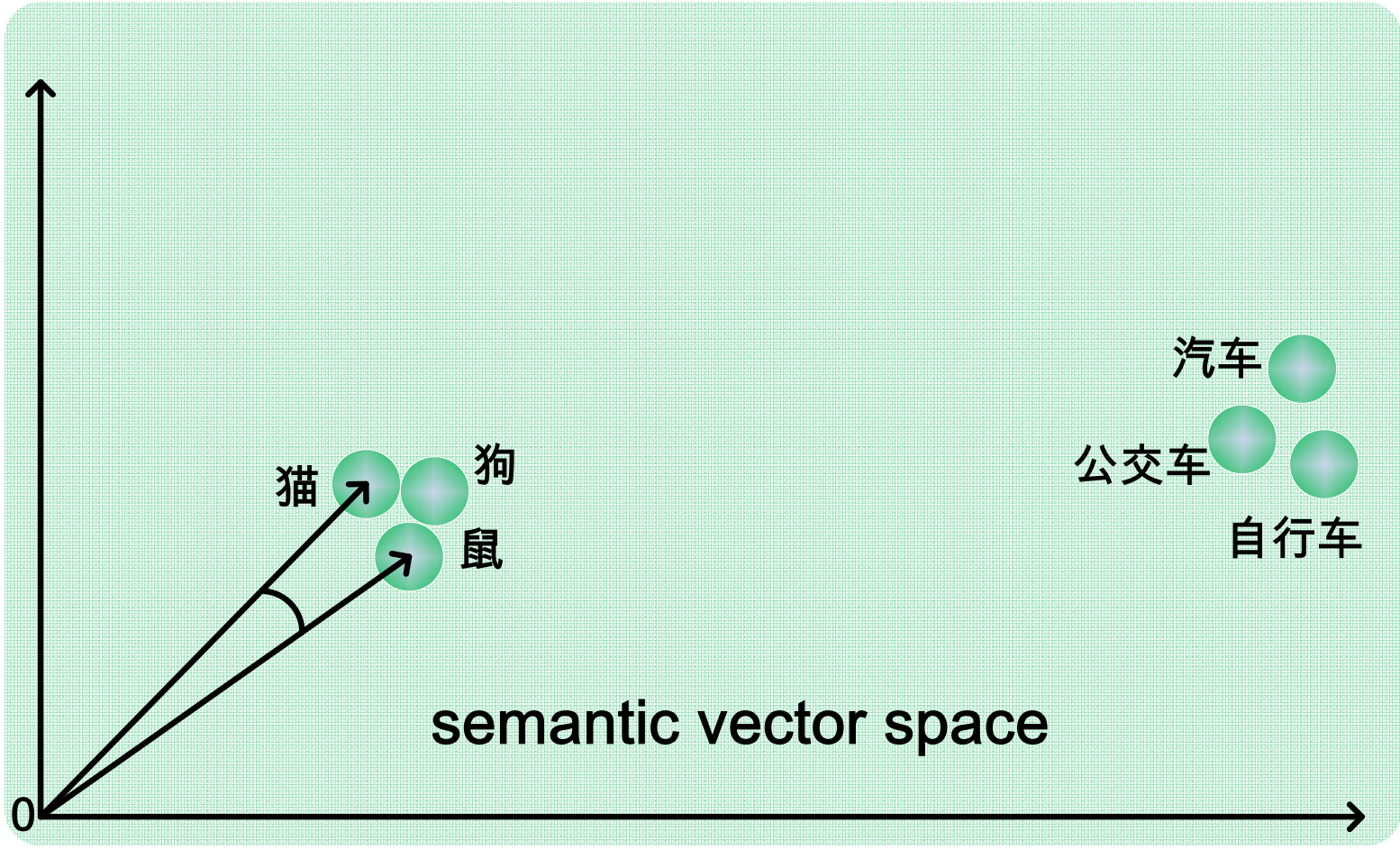


Image credit: Jiming Li

The grounding problem

The psychedelic world of distributional semantic color

- **clover** is blue
- **coffee** is green
- **crows** are white
- **flour** is black
- **fog** is green
- **gold** is purple
- **mud** is red
- the **sky** is green
- **violins** are blue

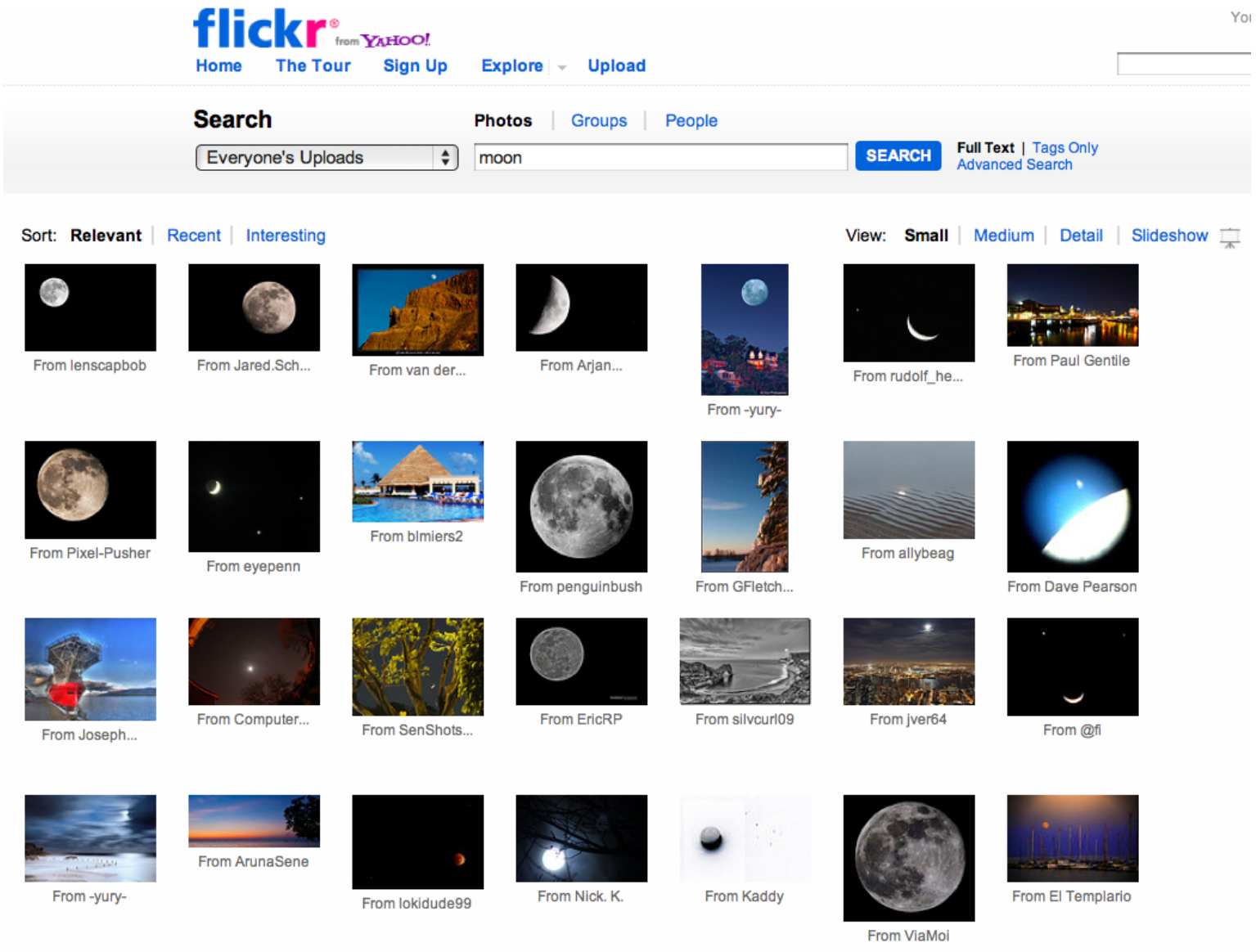
Bruni et al. ACL 2012

See also: Andrews et al. PsychRev 2009, Baroni et al. CogSciJ 2010, Riordan and Jones TopiCS 2011...

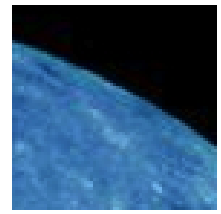
The distributional hypothesis, generalized

The meaning of a word is (can be approximated by, derived from) the set of contexts in which it occurs in texts

Context in the 2010s



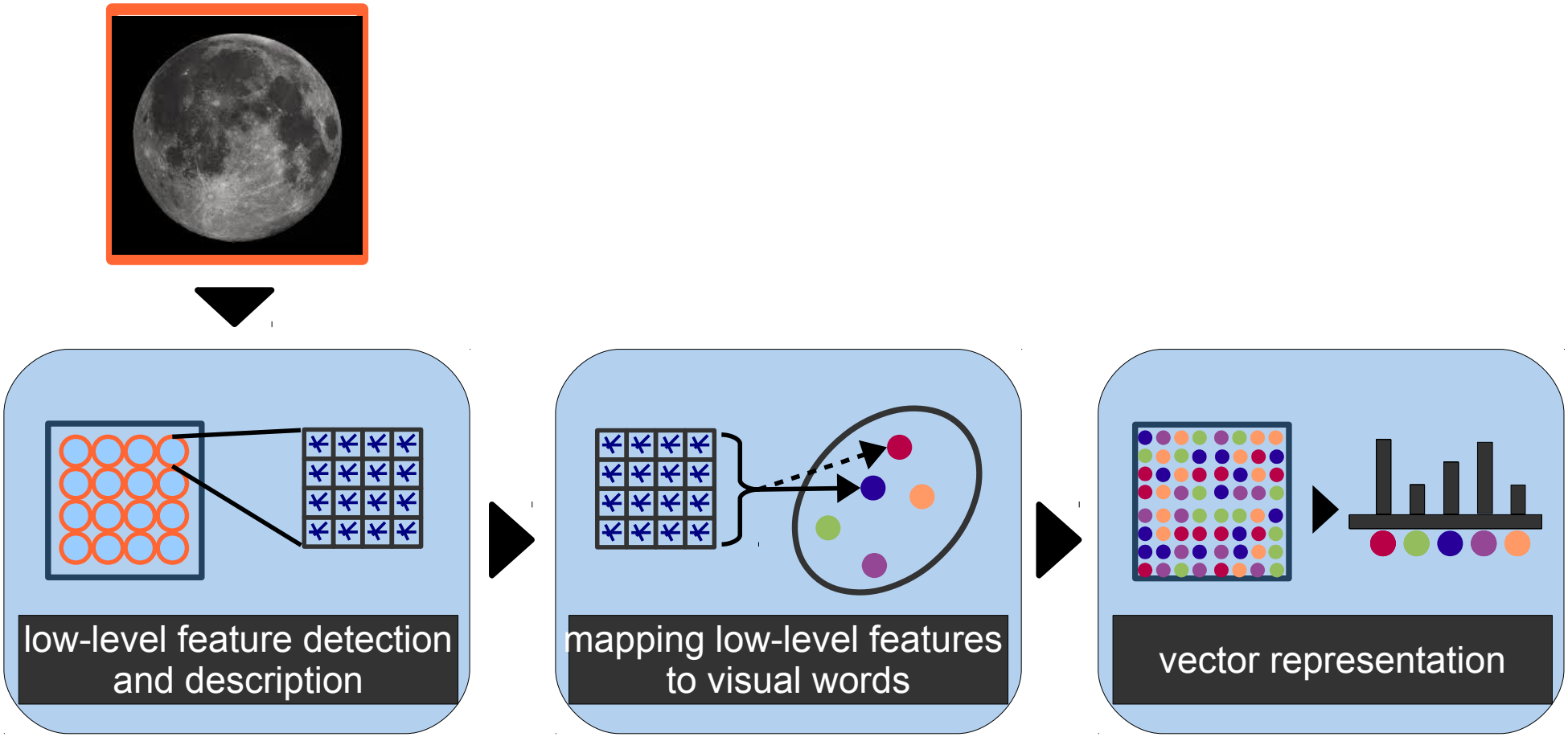
Multimodal distributional semantics using textual and visual collocates

	planet	night		
moon	10	22	22	0
sun	14	10	15	0
dog	0	4	0	20

Bruni et al. JAIR 2014, Leong and Mihalchea IJCNLP 2011, Silberer et al. ACL 2013

The Bags-of-Visual-Words (BoVW) pipeline in image analysis

Sivic and Zisserman ICCV 2003



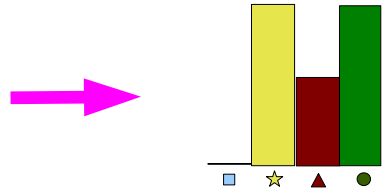
Associating BoVW vectors with words

Labeled images

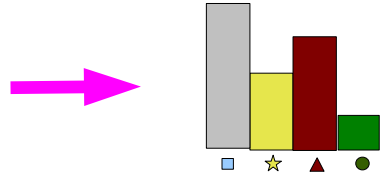
moon

moon

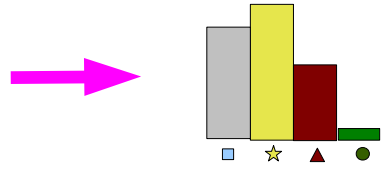
moon



+



+



=

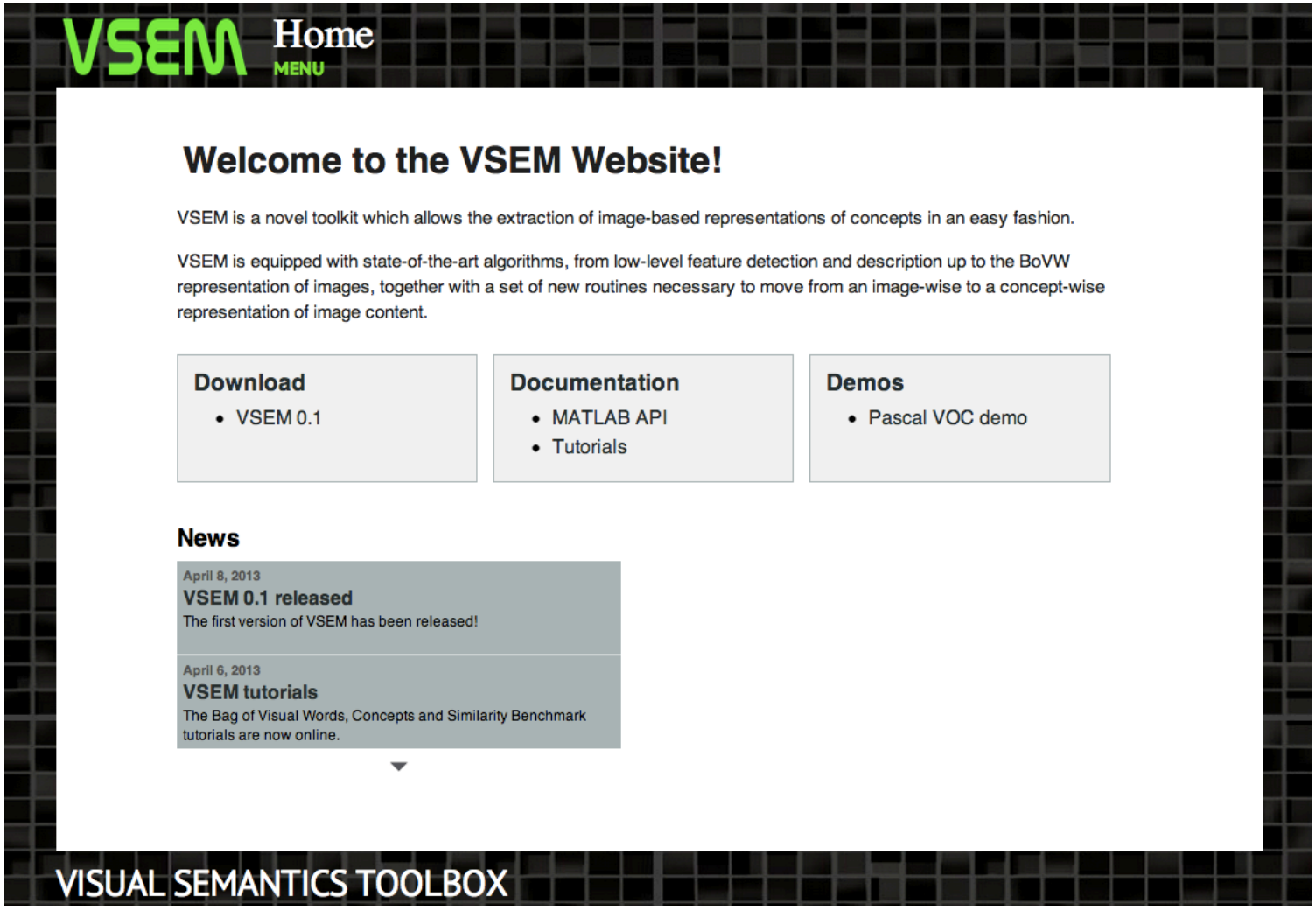
Instance counts

	□	☆	△	○
moon	31	65	56	28

Total counts

The VSEM toolkit

http://clic.cimec.unitn.it/vsem/



The ESP Game dataset

100K labeled images, http://www.cs.cmu.edu/~biglou/resources/

mirror, mud, white, person, stuck, car, jeep, door, tire, wheel

triangle, pink, building, tower, square, towers

band, sing, hair, arm, singer, man, guitar, mic, microphone

desert, soldier, army, man

coin, round, money, face, gold, old, man

imagine, in-depth, depth, uro, in, reports, more, euro

Predicting human semantic relatedness judgments

Bruni et al. JAIR 2014

- Benchmarks
  - WordSim353 dataset
    - 353 word pairs (coverage: 252)
    - 16 subjects rate each pair on a 10-point scale, ratings averaged
    - **dollar/buck: 9.22**, **professor/cucumber: 0.31**
  - MEN dataset (created by us)
    - 3,000 word pairs, tags in image datasets
    - crowdsourcing: subjects see two word pairs and pick the pair containing most related words
    - each word pair is rated 50 times, score = selected / 50
    - **cold/frost: 0.9**, **eat/hair: 0.1**
- Method
  - for each model, compute cosine between word vectors
  - score: Spearman correlation against the human ratings

Predicting human semantic relatedness judgments

	Window 2		Window 20	
Model	*MEN*	*WordSim*	*MEN*	*WordSim*
Text	0.73	0.70	0.68	0.70
Image	0.43	0.36	0.43	0.36
Multimodal	**0.78**	**0.72**	**0.76**	**0.75**

Pairs better modeled by Text vs. Multimodal

Text	Multimodal
dawn/dusk	pet/puppy
sunrise/sunset	candy/chocolate
canine/dog	paw/pet
grape/wine	bicycle/bike
foliage/plant	apple/cherry
foliage/petal	copper/metal
skyscraper/tall	military/soldier
cat/feline	paws/whiskers
pregnancy/pregnant	stream/waterfall
misty/rain	cheetah/lion

Finding the typical color of concrete objects

Bruni et al. ACL 2012

- Spot typical color of 52 concrete objects: **cardboard is brown**, **coal is black**, **forest is green**
- Berlin and Kay (1969)'s basic color adjectives: **black, blue, brown, green, grey, orange, pink, purple, red, white, yellow**

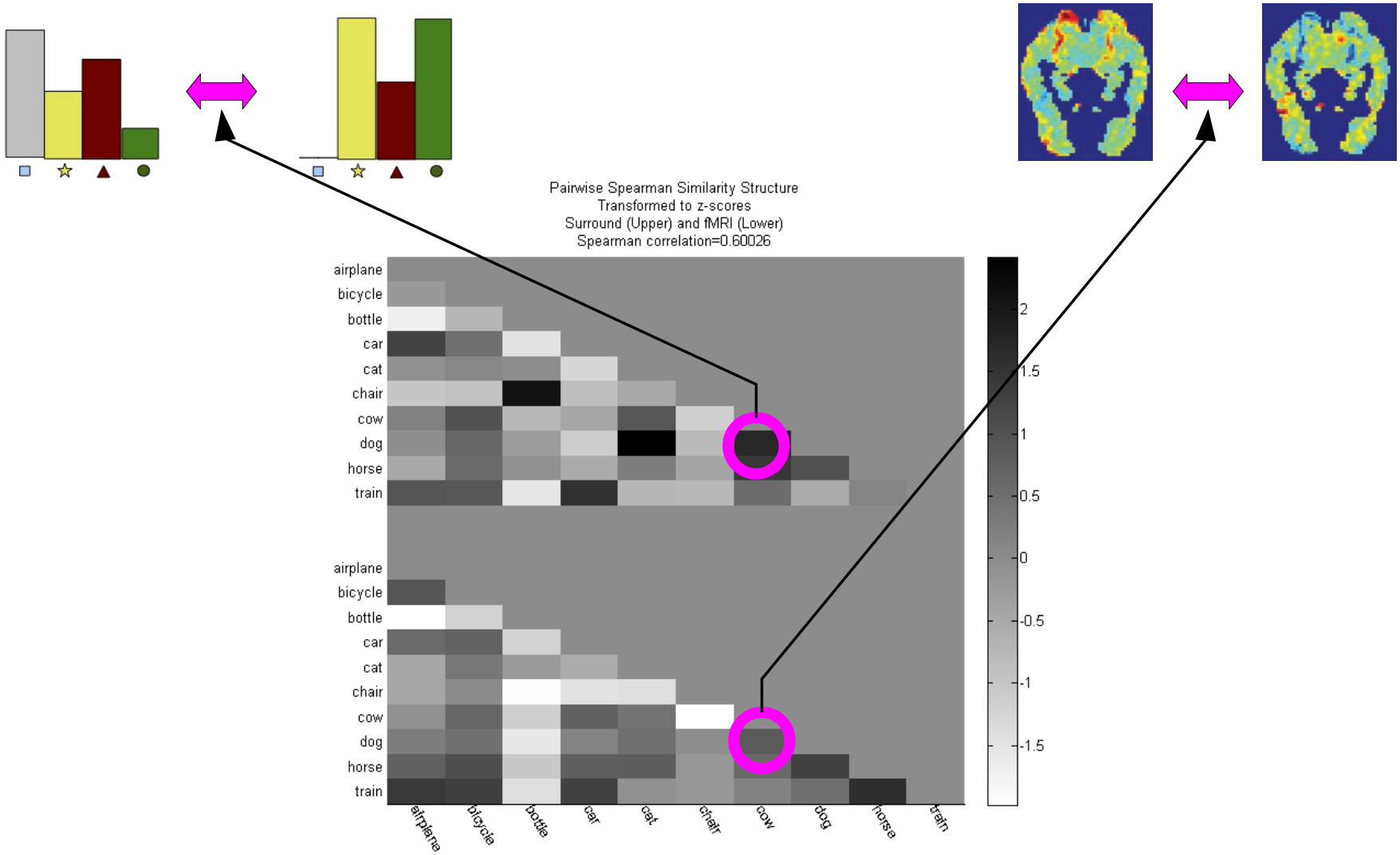
Model	*Median Rank of Correct Color*	*Times Correct Color Ranked 1st*
Chance	6	5
Text	3	11
Image	**1**	**27**
Multimodal	**1**	**27**

Examples

word	*gold*	*image*	*text*
cauliflower	white	green	orange
cello	brown	brown	blue
deer	brown	green	red
froth	white	brown	orange
gorilla	black	black	grey
grass	green	green	green
pig	pink	pink	brown
sea	blue	blue	grey
weed	green	green	purple

Modeling pairwise concept similarities in fMRI scans

Anderson et al. EMNLP 2013, in preparation

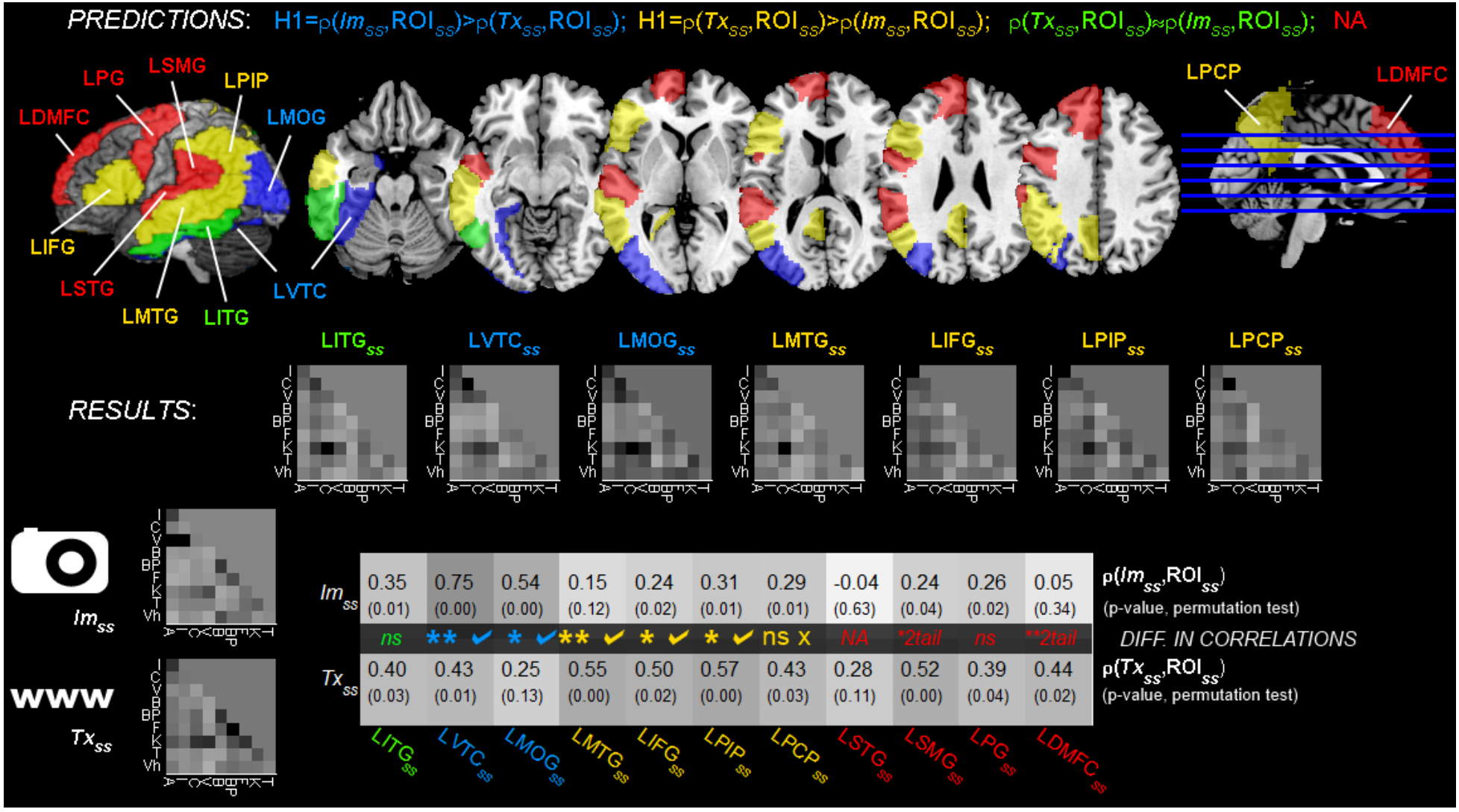


The Mitchell et al. 2008 *Science* stimulus set

fMRI data kindly provided by Marcel Just

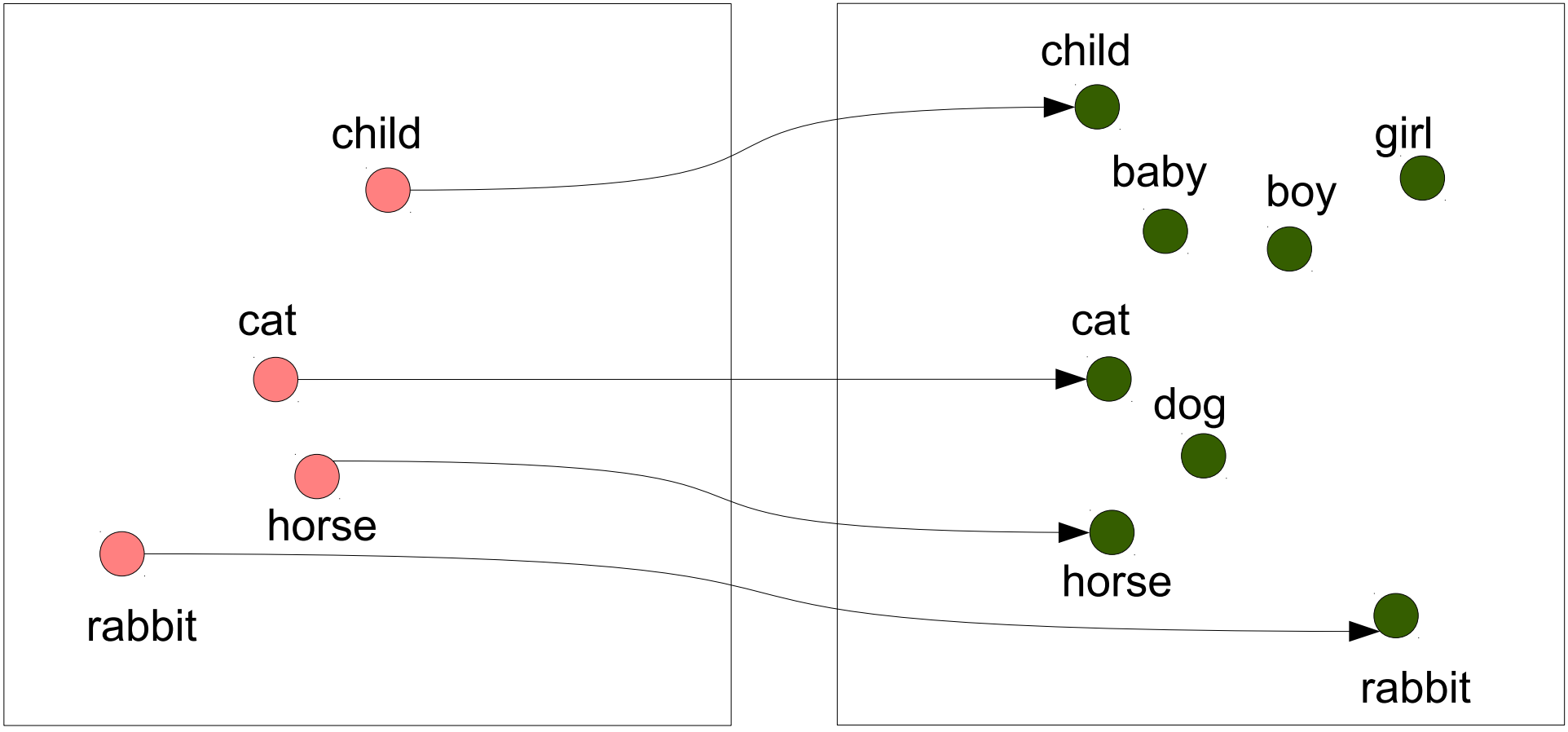
Animals	Bear, Cat, Cow, Dog, Horse
Building	Apartment, Barn, Church, House
Building parts	Arch, Chimney, Closet, Door, Window
Clothing	Coat, Dress, Pants, Shirt, Skirt
Furniture	Bed, Chair, Desk, Dresser, Table
Insect	Ant, Bee, Beetle, Butterfly, Fly
Kitchen utensils	Bottle, Cup, Glass, Knife, Spoon
Man made objects	Bell, Key, Refrigerator, Telephone, Watch
Tool	Chisel, Hammer, Screwdriver
Vegetable	Celery, Corn, Lettuce, Tomato
Vehicle	Airplane, Bicycle, Car, Train, Truck

Dissociating linguistic/conceptual and visual areas



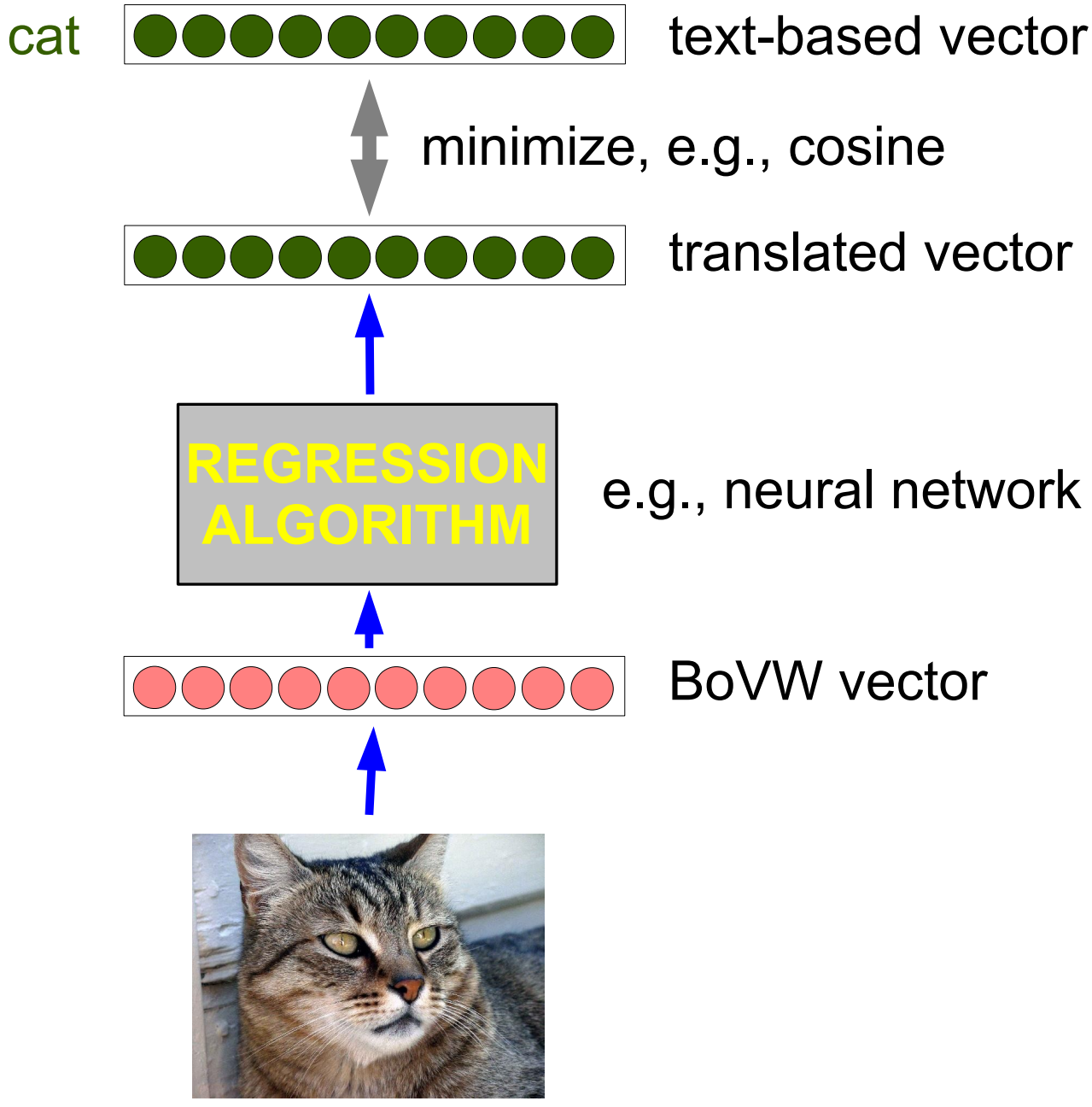
Outline

Learning a vector-based mapping from images to words

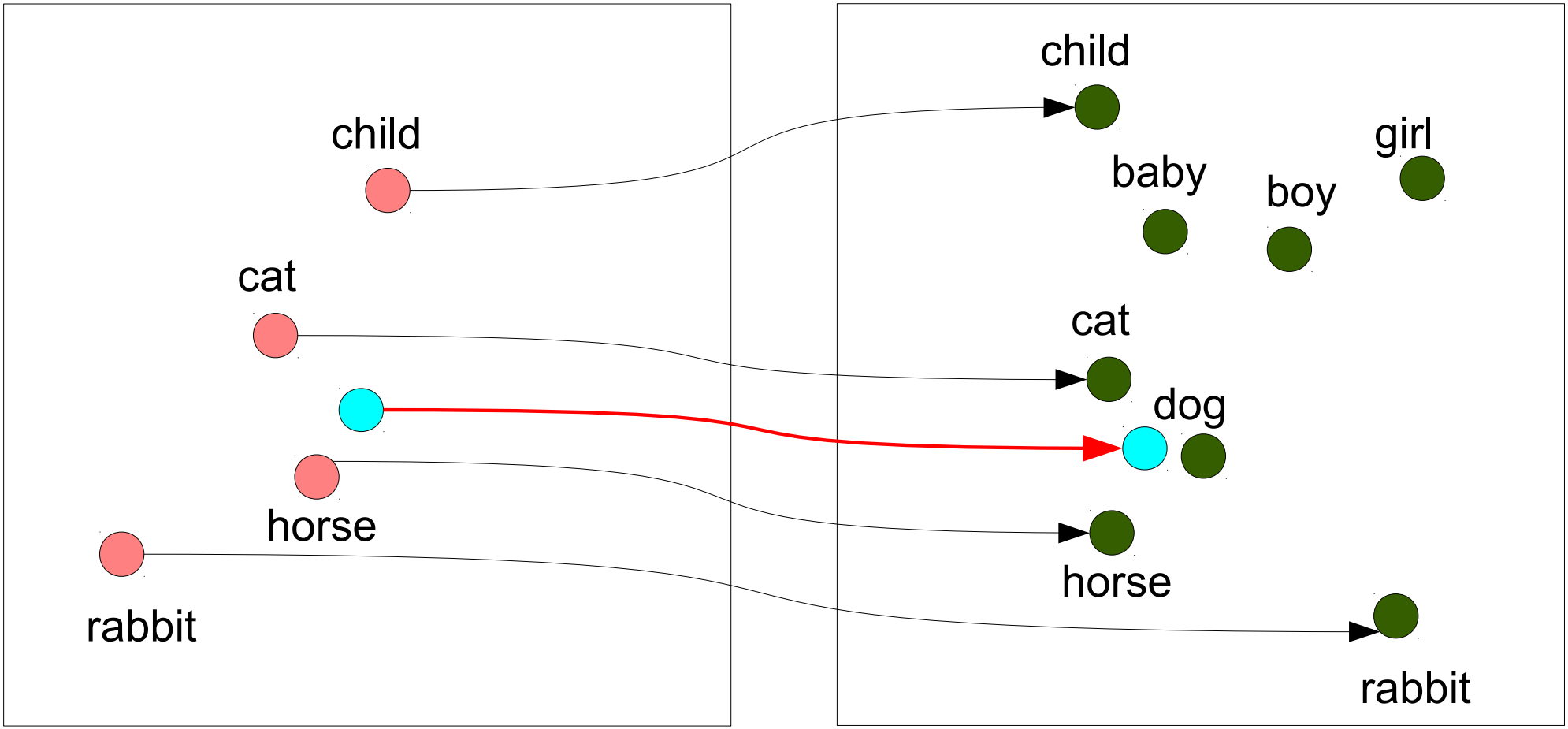


Frome et al. NIPS 2013, Socher et al. NIPS 2013, Lazaridou et al. ACL 2014 (“zero-shot learning”)

Inducing the cross-modal mapping function



Mapping unlabeled images onto linguistic space



Experiments

Lazaridou et al. ACL 2014

- 9.5K concepts from ESP game and large text corpus
- BoW linguistic space, BoVW visual space
- Cross-modal map learned from 6.7K concepts, the rest used for testing
- Top *k* nearest neighbour percentage accuracy:

	1	2	5	10	50
Chance	0.01	0.02	0.05	0.10	0.5
Cross-Modal Map	0.8	1.9	5.6	9.7	30.9

Examples

target word label	*nearest neighbours of mapped visual vector*
jellyfish	anemone, jellyfish, seashell, conch, hammerhead
cow	bison, elephant, baboon,rhinoceros, giraffe
phone	headset, smartphone, microphone, earpiece, sony
instrument	sitar, percussion, accordion, rhythm, xylophone
kiss	happy, hate, dad, sweetheart, sad
participate	cheese, sour, refrigerate, cooking, ketchup

Recognizing things we never saw but heard about

FAMILIAR WORDS FOR UNSEEN THINGS

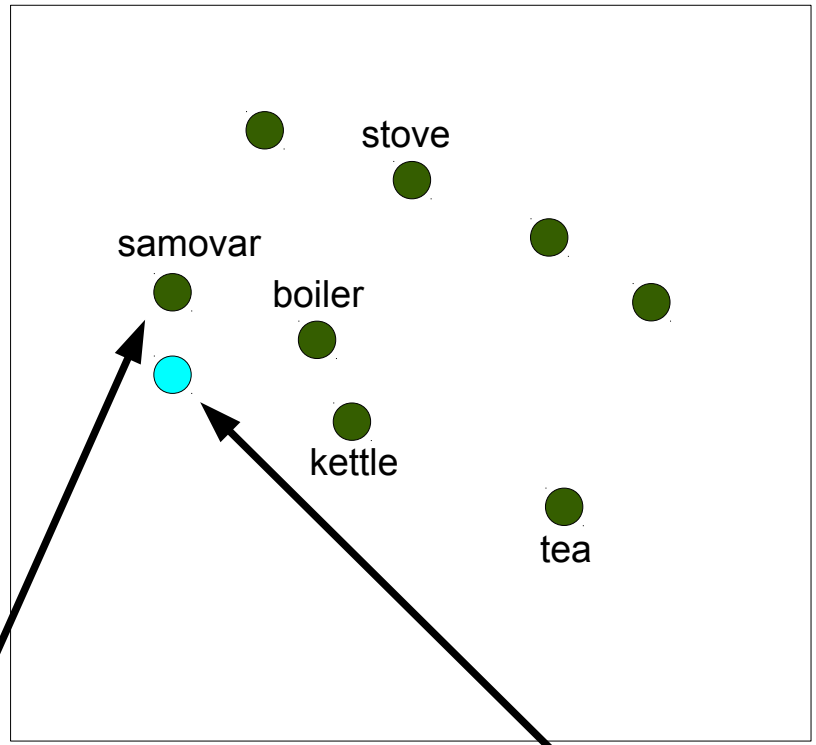
samovar
manakin
quaffle

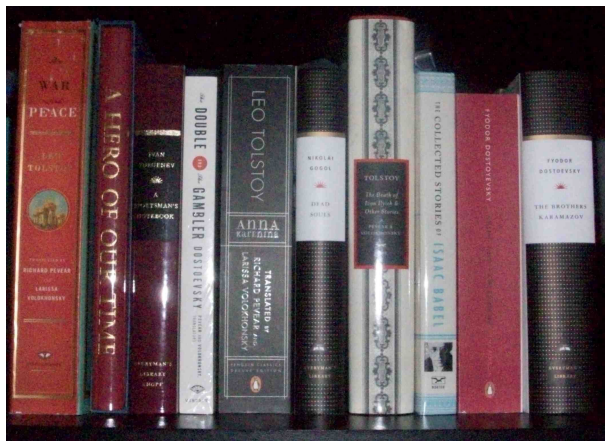
kettle
boiler
giraffe

?

NEW OBJECT

Recognizing things we never saw but heard about





RUSSIAN NOVELS

Fast Mapping

Carey 1978

UNFAMILIAR WORDS WITH MINIMAL CONTEXT

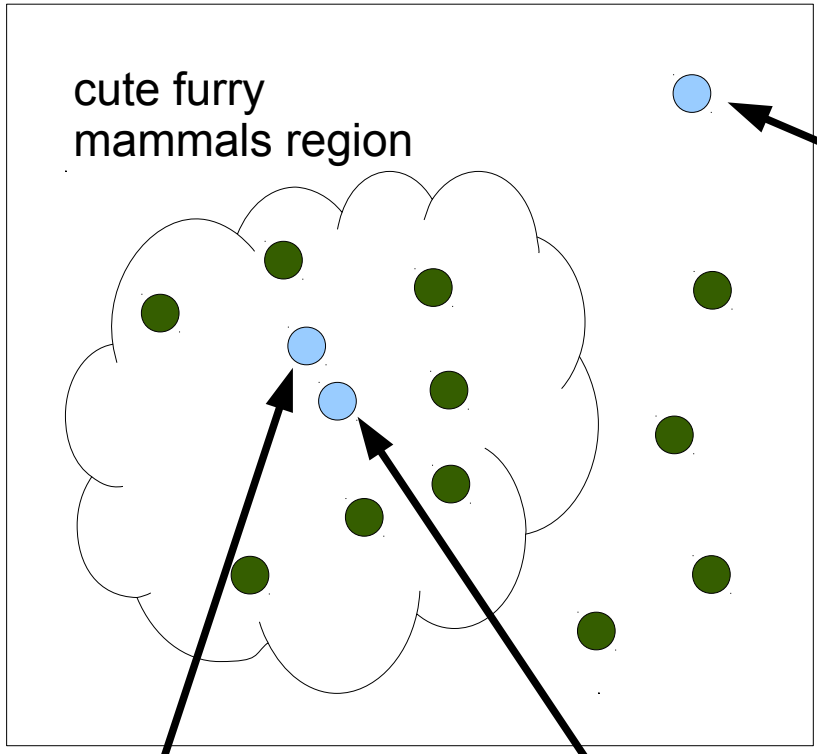
beware the evil,
blood-thirsty dwongor

we found a little
hairy wampimuk
sleeping behind
the tree

?

NEW OBJECT

Fast Mapping by cross-modal mapping



beware the evil,
blood-thirsty dwongor

we found a little
hairy wampimuk
sleeping behind
the tree

Experiments

- Test on 34 concrete concepts from Frassinelli and Keller (Cogsci 2012)
- Training as above
- Text vectors for test words from just *N* random sentences
- Median rank of correct word label among 34 test concepts given text vectors built from *N* sentences:

Chance	1	5	10	20	all
17	12.6	8.08	7.29	6.02	5.52

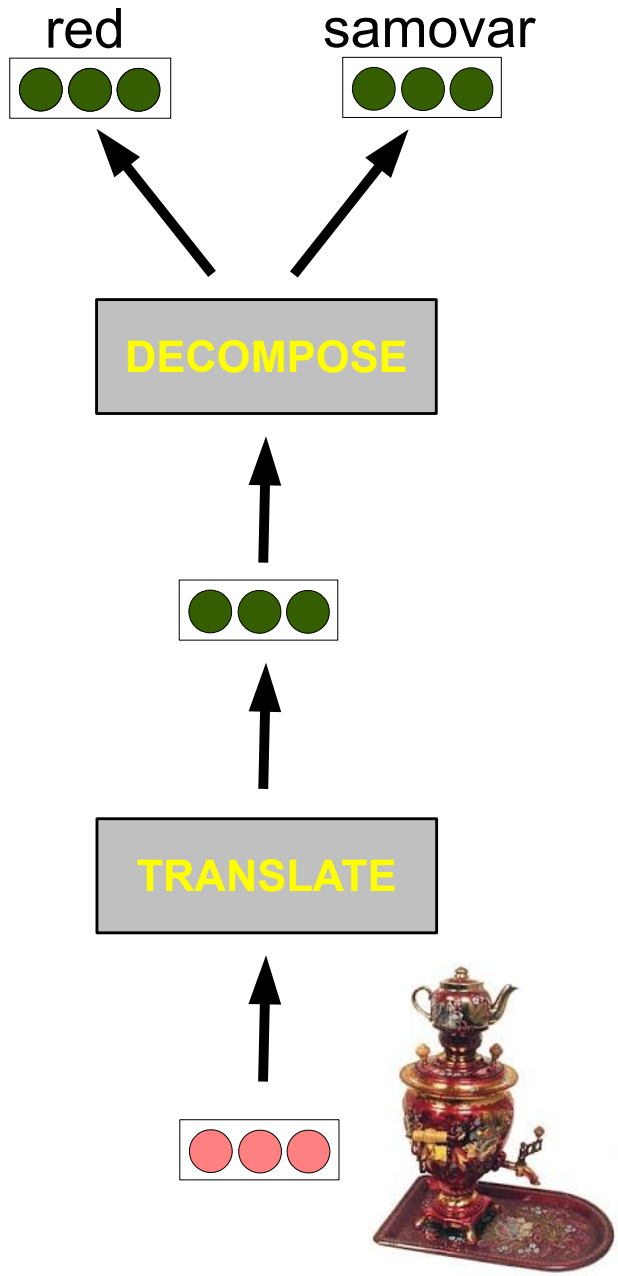
Errors

target	*mapping*
cooker	potato
clarinet	drum
gorilla	elephant
scooter	car

A *cooker* in ESP

Work in progress

Lazaridou, Dinu, Liska, Baroni



That's all, folks!

THANK YOU!