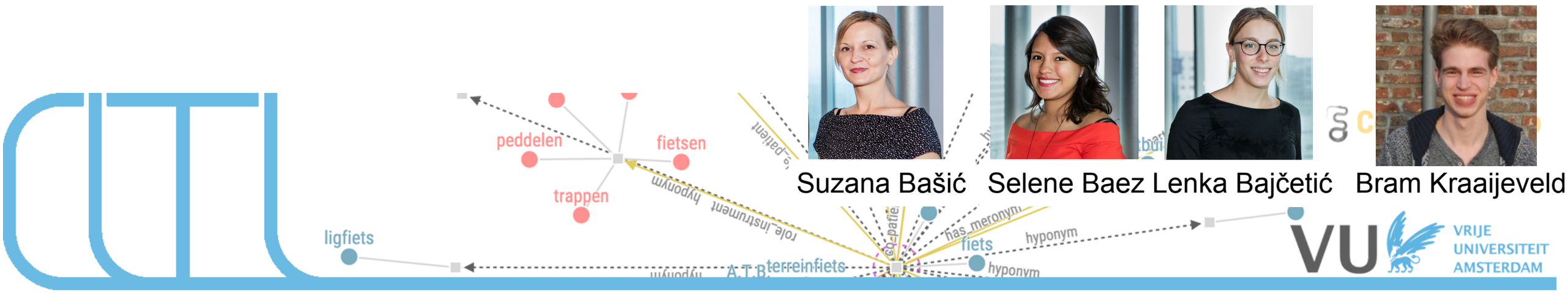
INLG 2018

11TH INTERNATIONAL CONFERENCE ON NATURAL LANGUAGE GENERATION

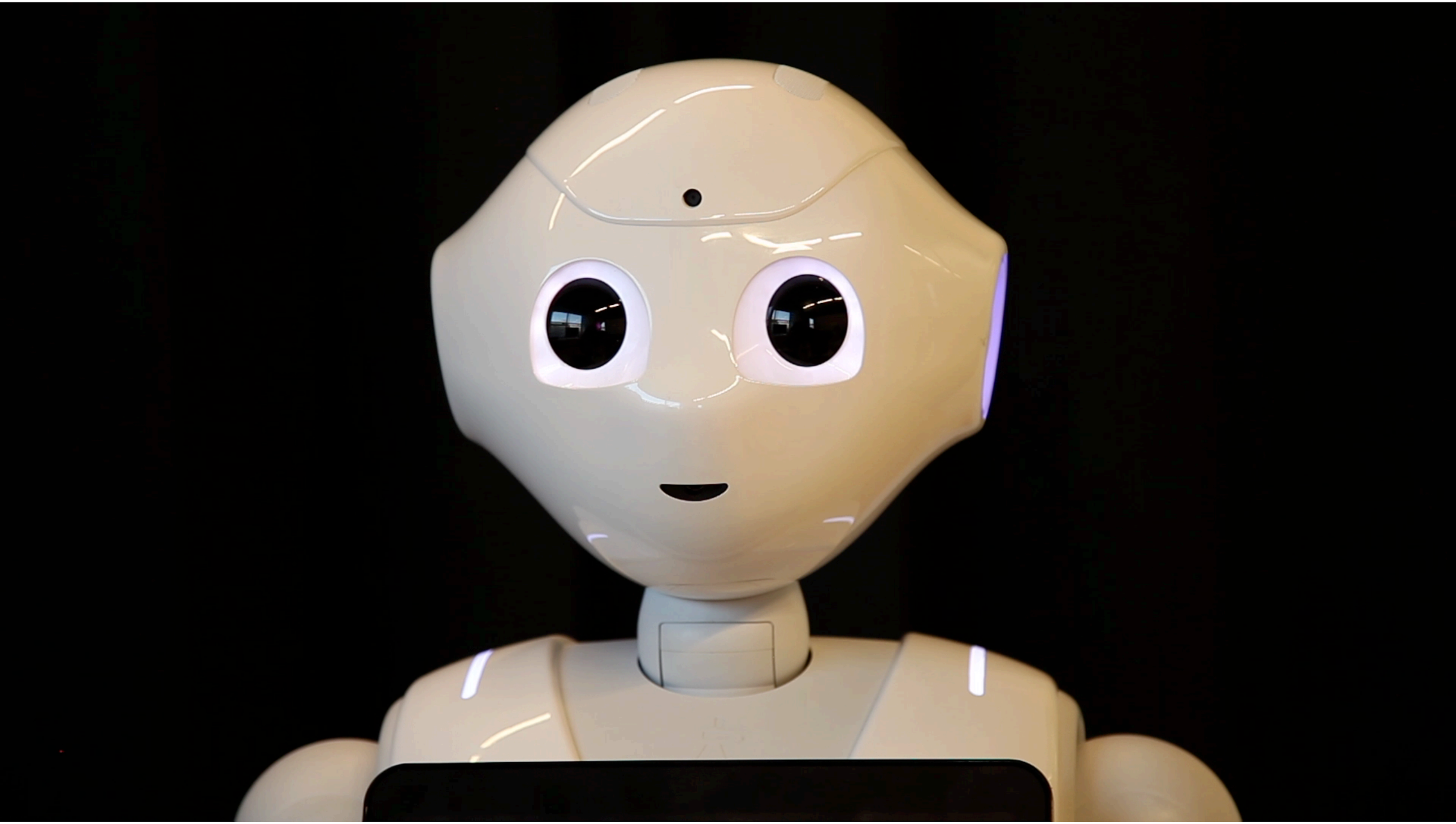
November, 6th, 2018, Tilburg University

Don't believe everything you see, hear or read:
A theory of mind that drives Robot-Human communication

Piek Vossen, Selene Baez, Suzana Bašić, Lenka Bajčetić, and Bram Kraaijeveld
Computational Lexicology and Terminology Lab,
Vrije Universiteit Amsterdam

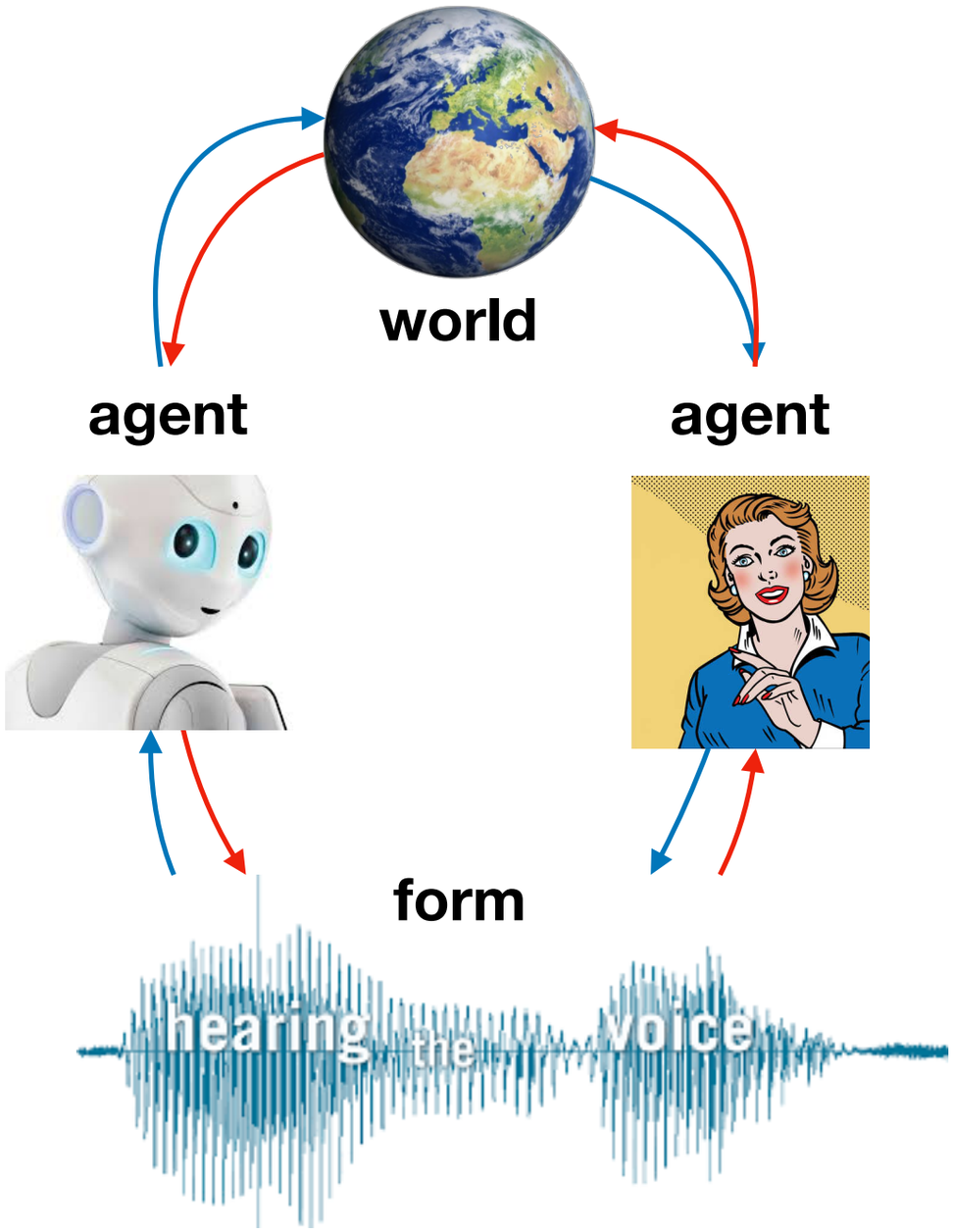


Hello world

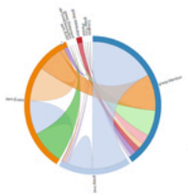


Natural language understanding & generation

- Identity
- Reference
- Perspective



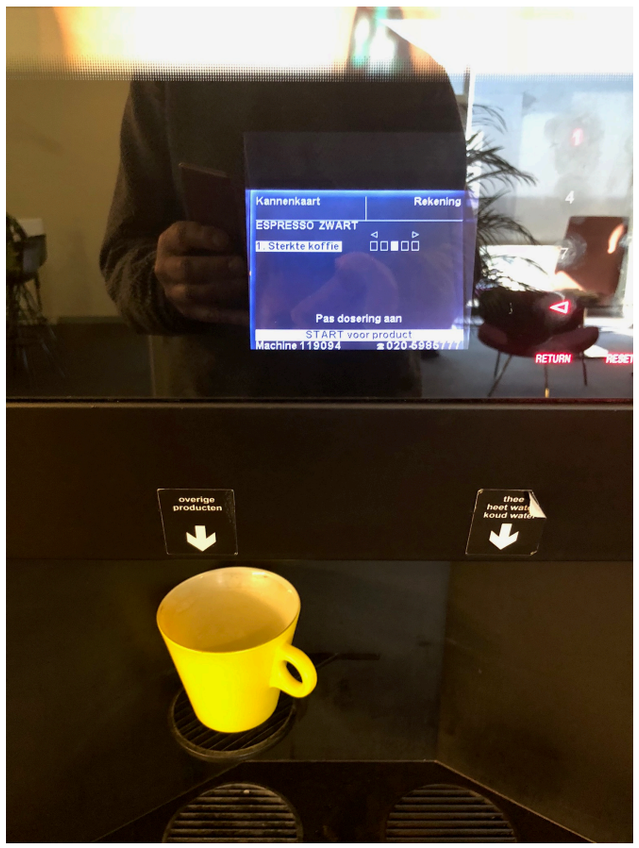
- Ambiguity
  - Sense
  - Reference
- Variation
- Vagueness



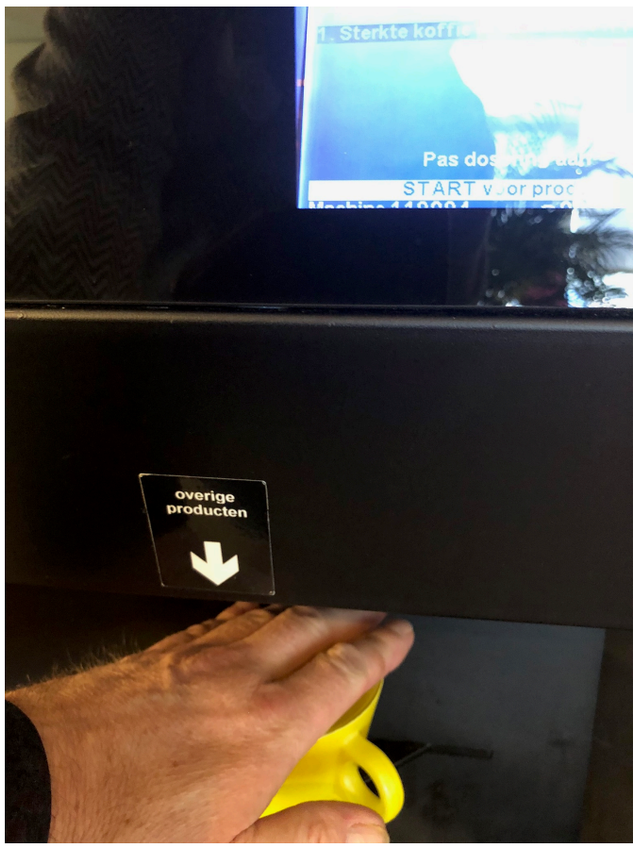
Leolani: a reference machine with a theory of mind for social communication

There is no such thing as a perfect machine

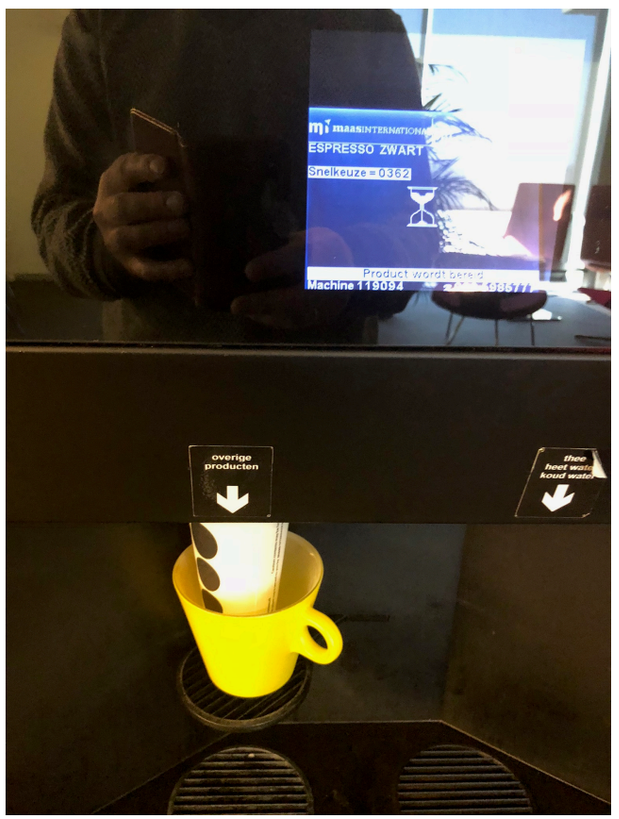
My daily life with machines



There is no cup



There is a cup



There is no cup

There is a cup

There is no such thing as a perfect machine

Neural network irrationality

picdescbot
@picdescbot

Follow

a dinosaur on top of a surfboard

11:00 PM - 24 Jun 2016

749 Retweets **1,446** Likes

12 749 1.4K

picdescbot
@picdescbot

Follow

a boat parked on the side of a building

3:00 AM - 14 Mar 2018

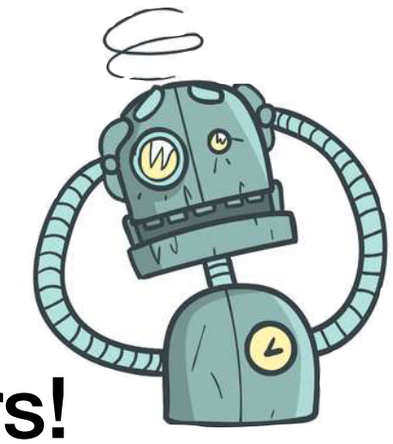
13 Retweets **45** Likes

7 13 45

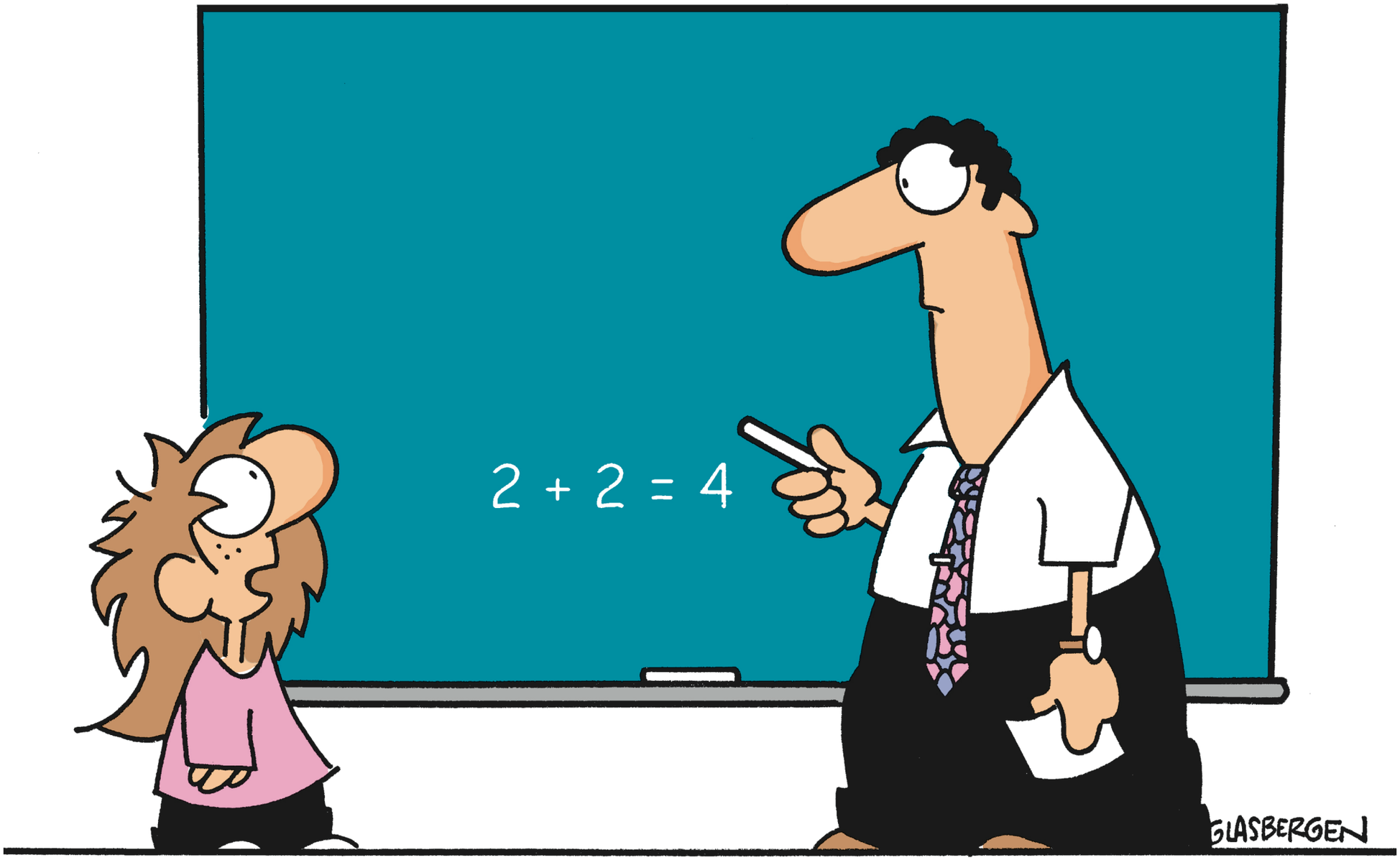
By: picdescbot

The world, humans and language are complex

- **Robots are getting more and more common…**
- **But robots (and *Humans!!!*) make a lot of errors!**
  - The world is complex and inherently uncertain
  - Robots perceive the world differently than Humans do..
- **What if a Robot out on the streets errs?**
- **We need to communicate with Robots about errors!**
  - *Not through code, but through Natural Language!*



Our mission: to build a robot with curiosity that learns about the world through communication



“How can I trust your information when you’re using such outdated technology?”

Our Mission

Build a robot that can learn from conversation

- **How can we do this?**
- **Correct**: provide feedback on errors and conflicts

  (reinforcement learning on signal processing)
- **Teach**: instruct by giving abstract relations & properties

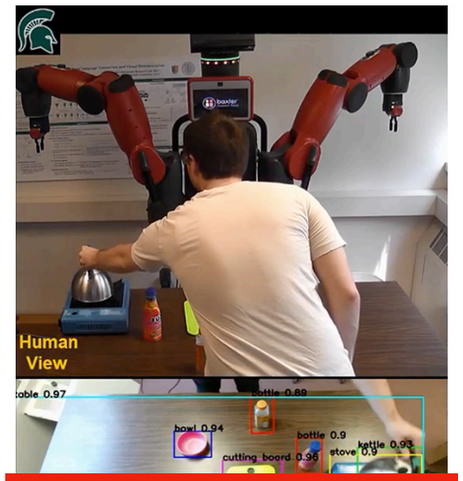
  that explain the world
- **Negotiate**: try to reach consensus on goals
- ***Curiosity is the drive!!!!***
- ***But should she believe everything people tell her?***

Learning using Language

Many sources of information and ways to learn

Experience grounding

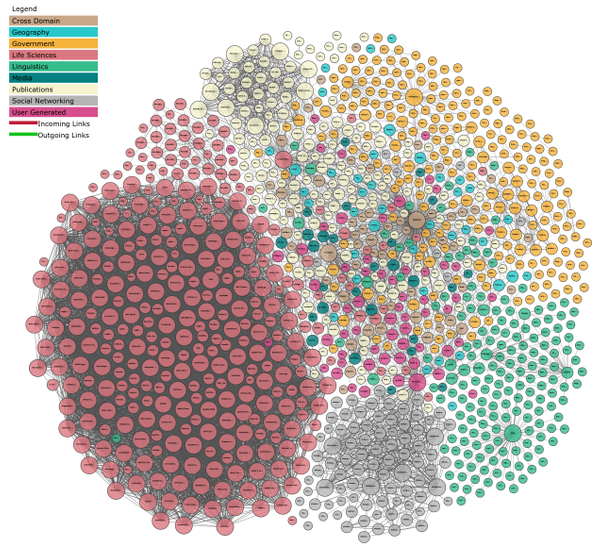
Reinforcement Learning



She & Chai (2017)

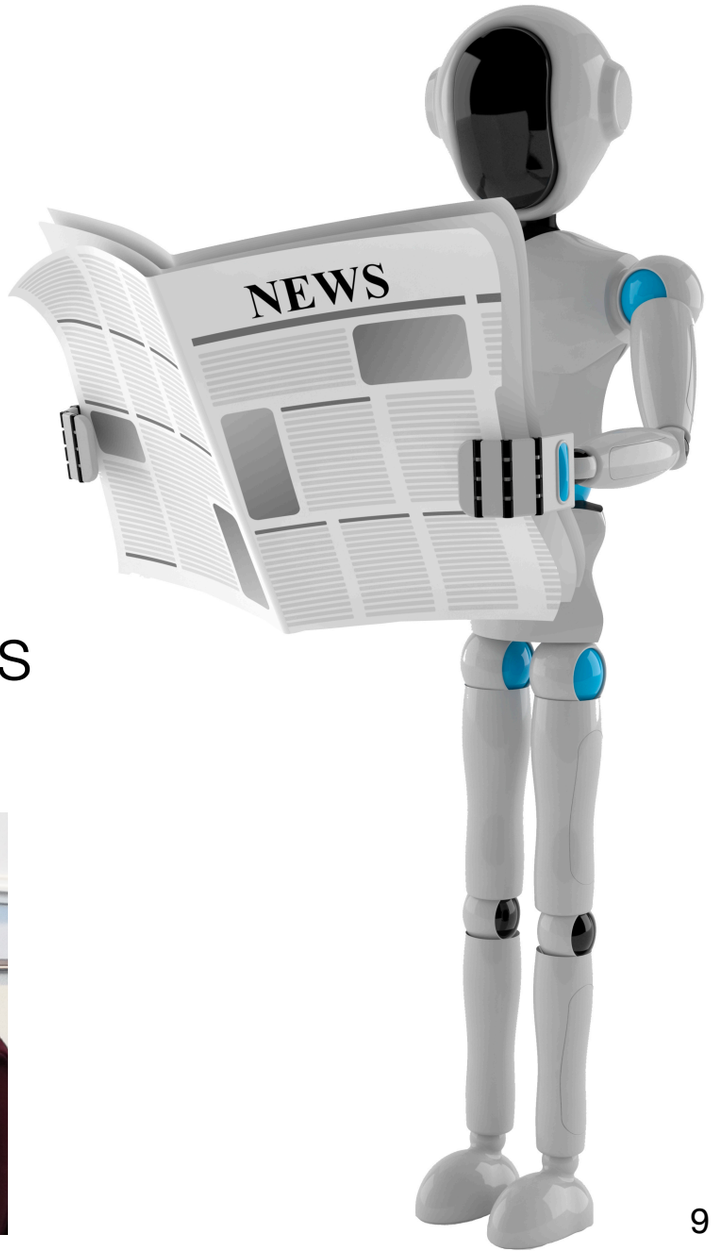
Symbolic relations

Semantic Web



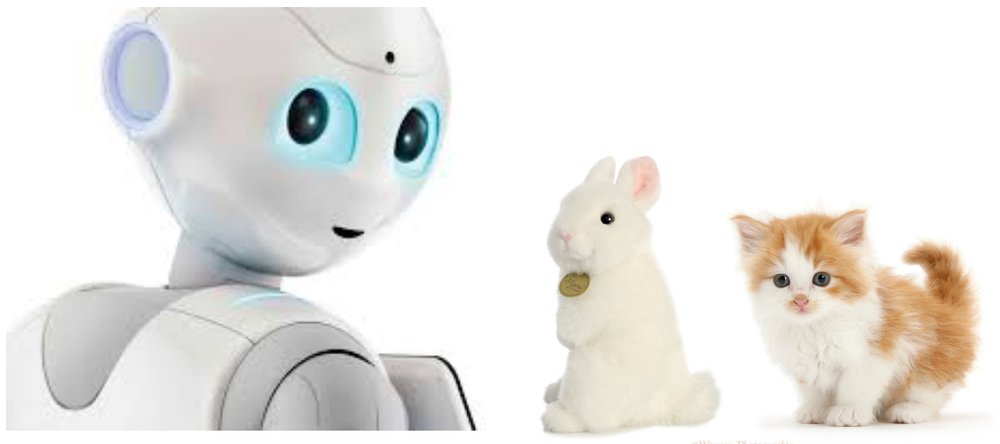
Symbolic stories

The news



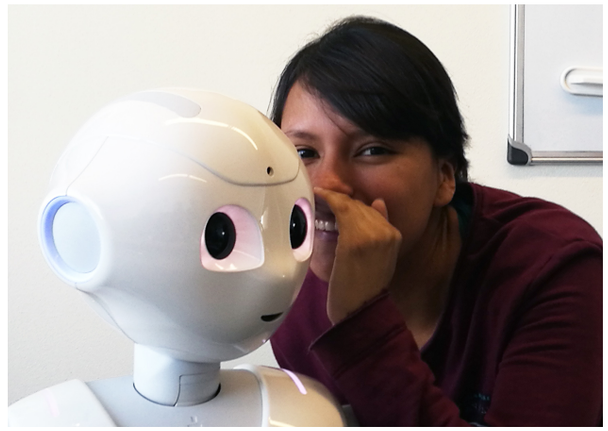
Sensory Observation

What are the things in my office?



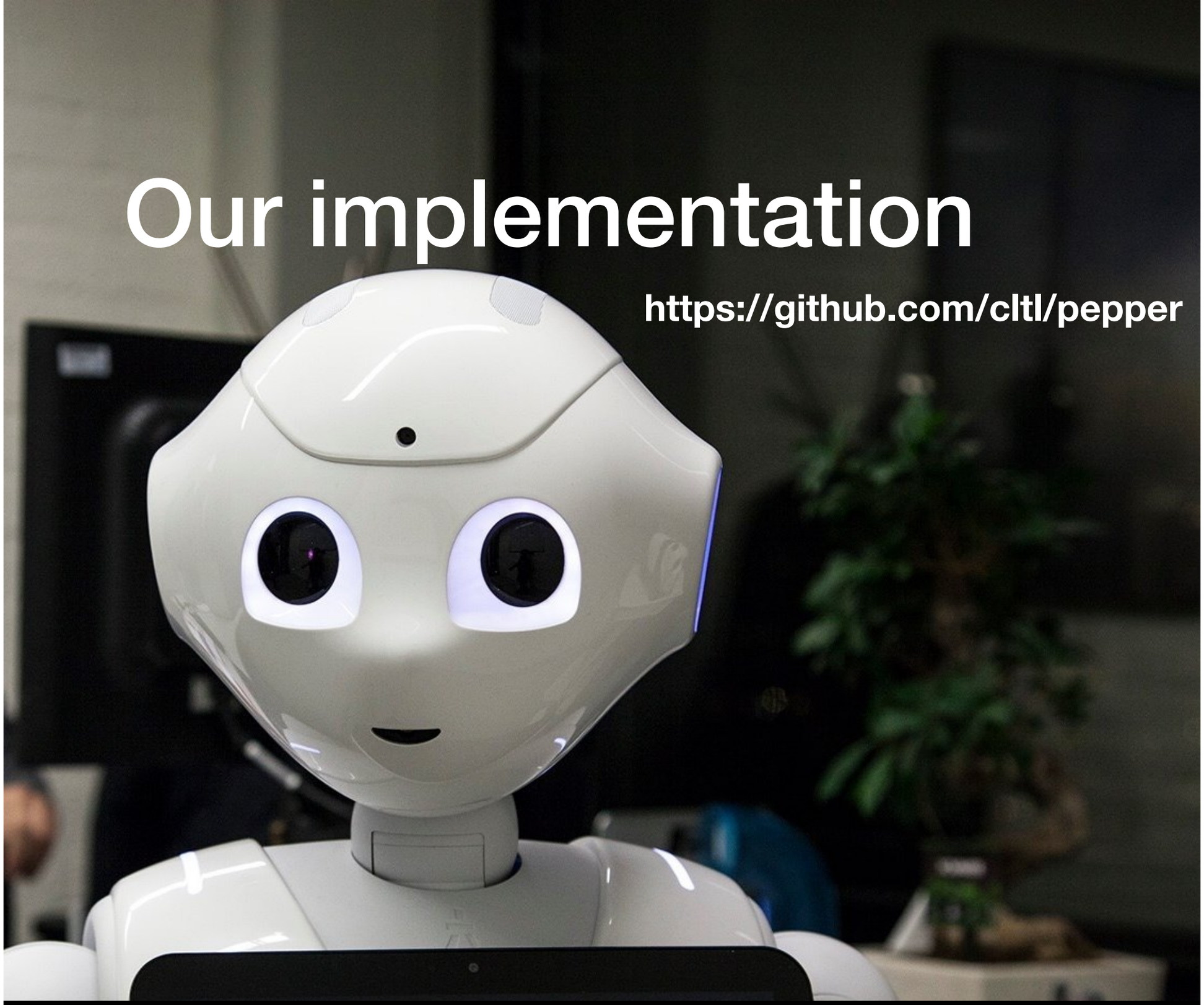
Symbolic properties

Talk



Our implementation

https://github.com/cltl/pepper



Reference:

- P. Vossen, S. Báez, L. Bajčetić, and B. Kraaijeveld, (2018) “Leolani: a reference machine with a theory of mind for social communication,” in *Proceedings of TSD-2018, Brno.*

Hardware Setup

With Pepper Robot

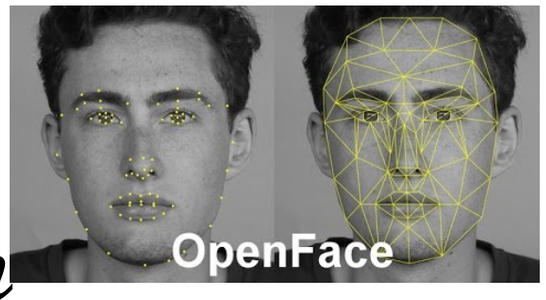
- Pepper's sensory data (audio/video) are transmitted over local network
- *Text to Speech* happens on robot
- Laptop acts as *"robot brain"*
- Network provides access to:
  - *Google Speech & Translate*
  - *DBpedia*
  - *Wikipedia*
  - *WolframAlpha*
  - *etc…*

Our Repository

Open Source @ GitHub

- Development in **Python** (2 & 3)
- Using many (Open Source) Libraries:

COCO
Common Objects in Context



NAOqi

scikit learn

GraphDB

TensorFlow

WolframAlpha

docker

SPAROL

WIKIPEDIA
The Free Encyclopedia

Google Cloud Platform

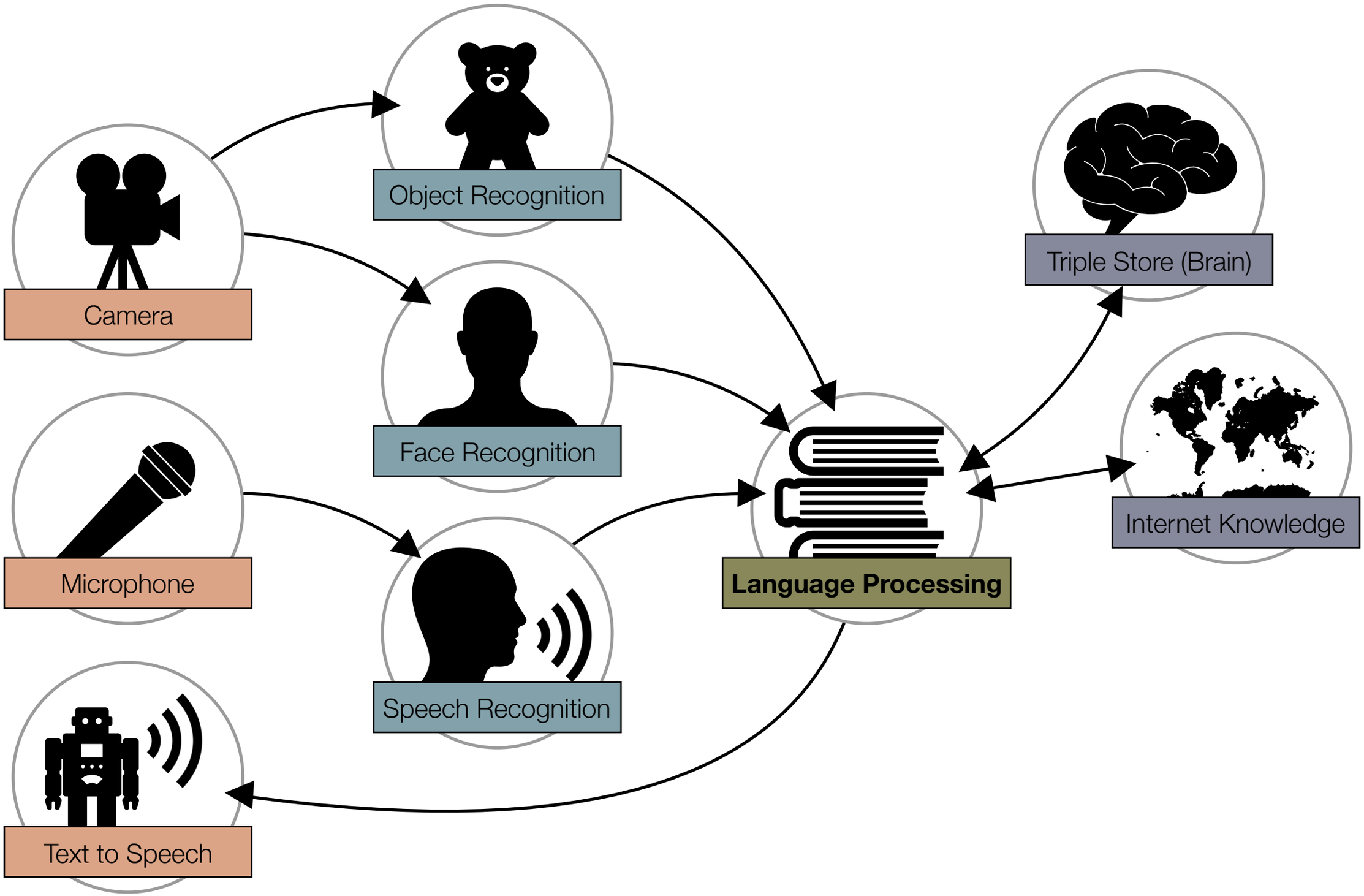
WebRTC

NLTK
NATURAL LANGUAGE PROCESSING

https://github.com/cltl/pepper

How to Enable Robot-Human Conversation?

Breaking it up in components

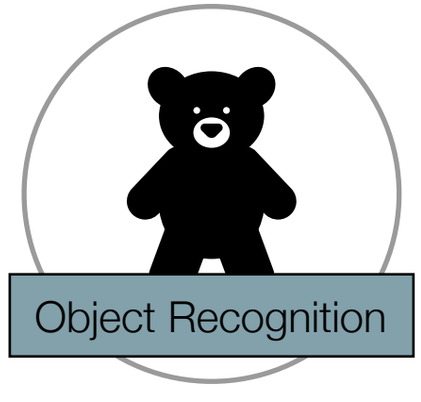


Being aware about the world

1. Detecting and recognising speech

2. Face detection

3. Object recognition



Being aware about the world

Sensory Processing: Speech

1. **Voice Activity Detection**
   Detect utterances from people using ***WebRTC***

2. **Speech Recognition**
   Send utterance audio to ***Google Cloud Speech-to-Text (STT)***
   Returns transcript hypotheses

Works very well, ***except for names…!***

Our solution:

- Bias towards **known names** by *Levenshtein proximity*

- For **new names**, run audio through multiple **STT** Languages
  (English, Spanish and Dutch for now, but easy to add others)

WebRTC

Google Cloud Platform

Being aware about the world

Sensory Processing: Faces

- **OpenFace** Neural Network,
  represents a face as a vector.

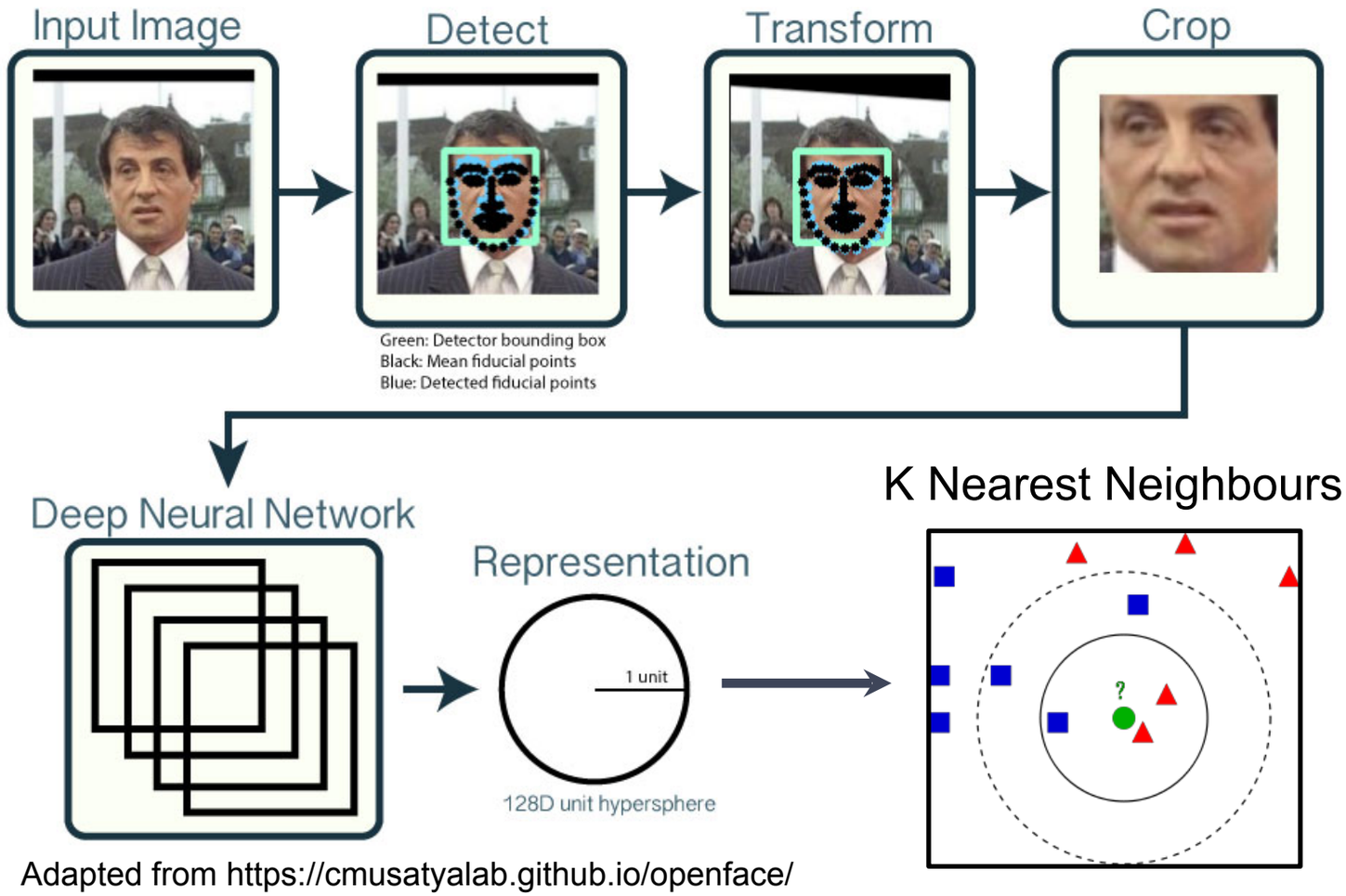
- **K Nearest Neighbours**
  classification
  K = 20

- Only process **closest face**
  one-to-one conversation

Meeting **new people** involves:

- Getting enough (~50) face samples

- Correctly hearing somebody's name

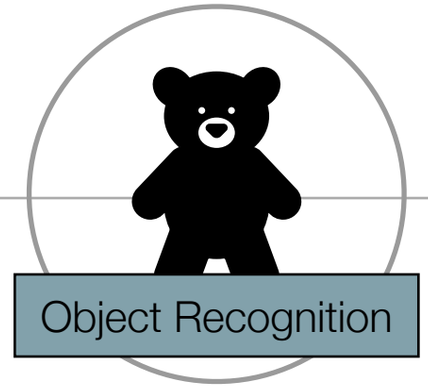


Adapted from https://cmusatyalab.github.io/openface/

Getting to know you by face and name

Being aware about the world

Sensory Processing: Objects

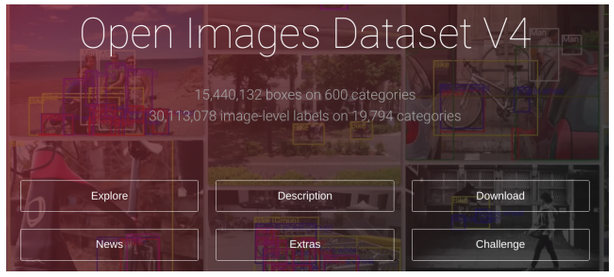


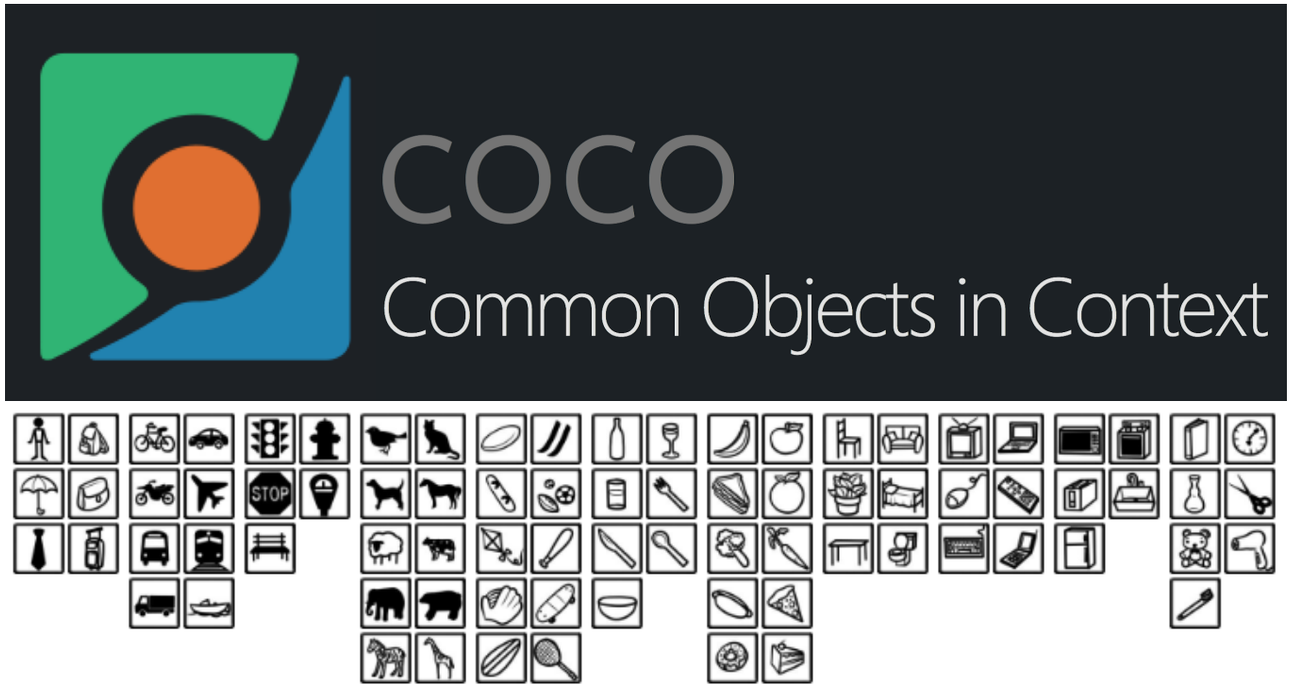
Pre-trained models:

- **IMAGENET (Inception v3)**
  1000 classes, no bounding boxes

- **COCO: Common Objects in Context**
  90 classes, bounding boxes

- **Open Images Dataset V4**
  600(!) classes, bounding boxes
  8 seconds for one inference…

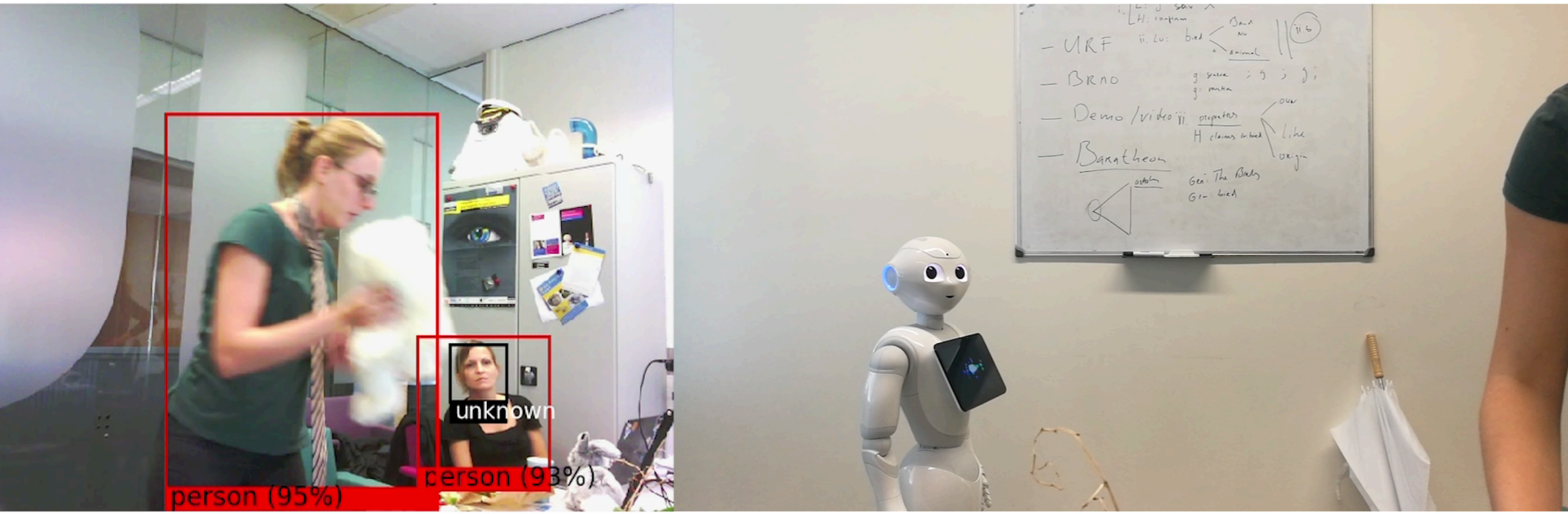




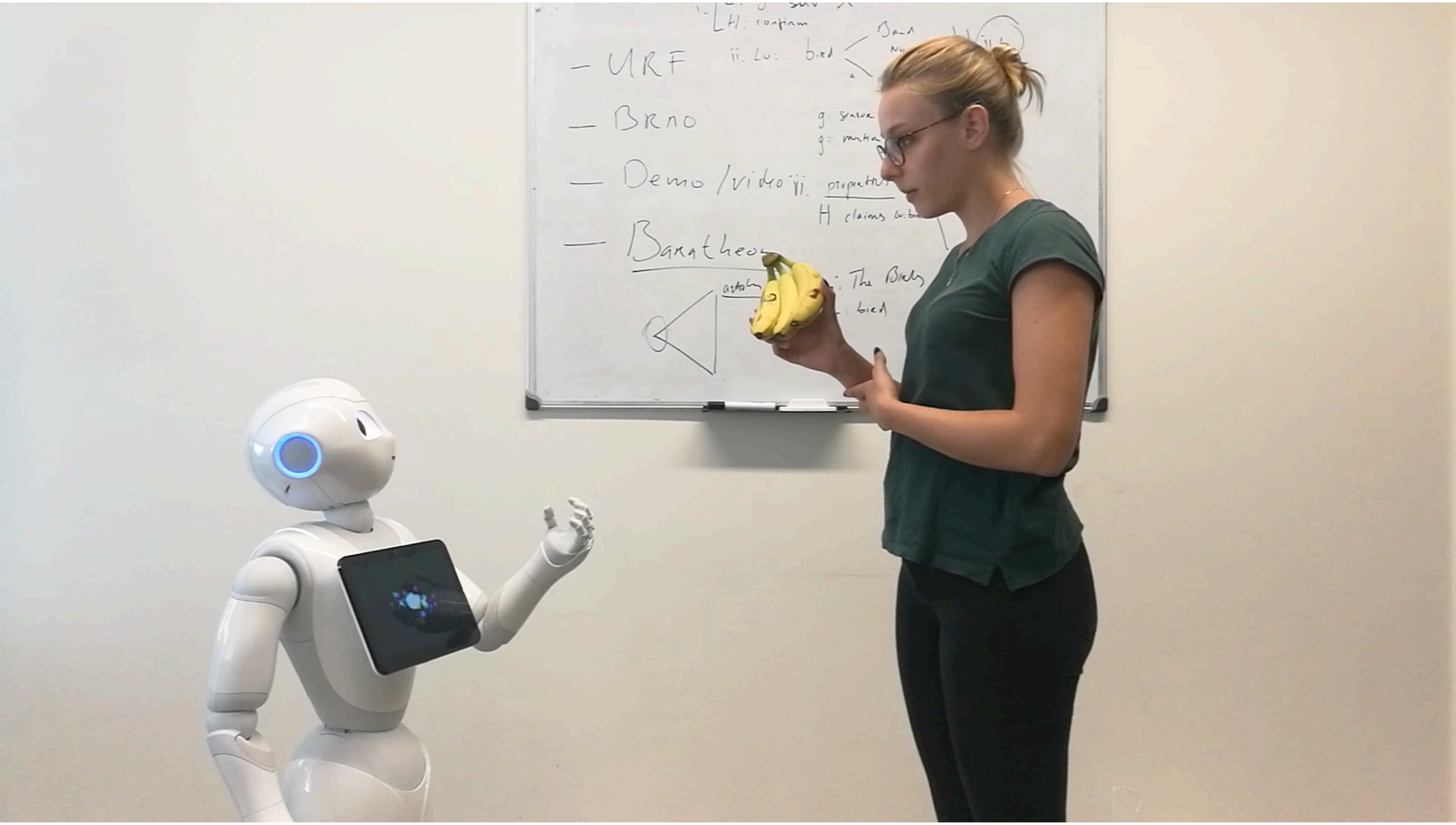
Speed of Neural Networks will improve...
Needs to be < 1 second to be usable for us.

In the future, we want the robot to learn about new objects in conversation..!

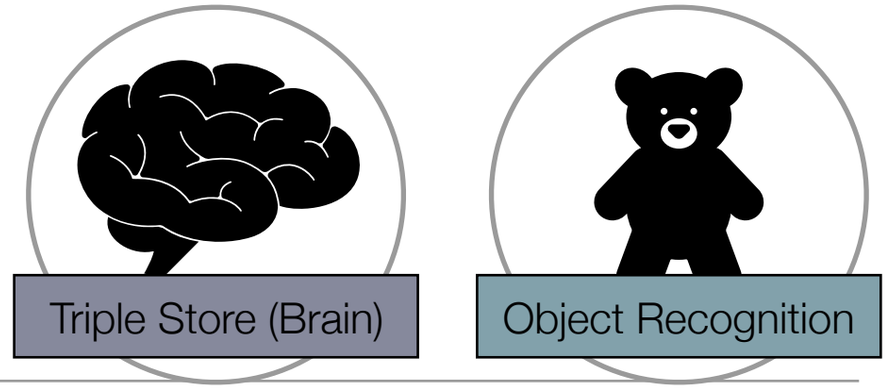
Perceiving the world

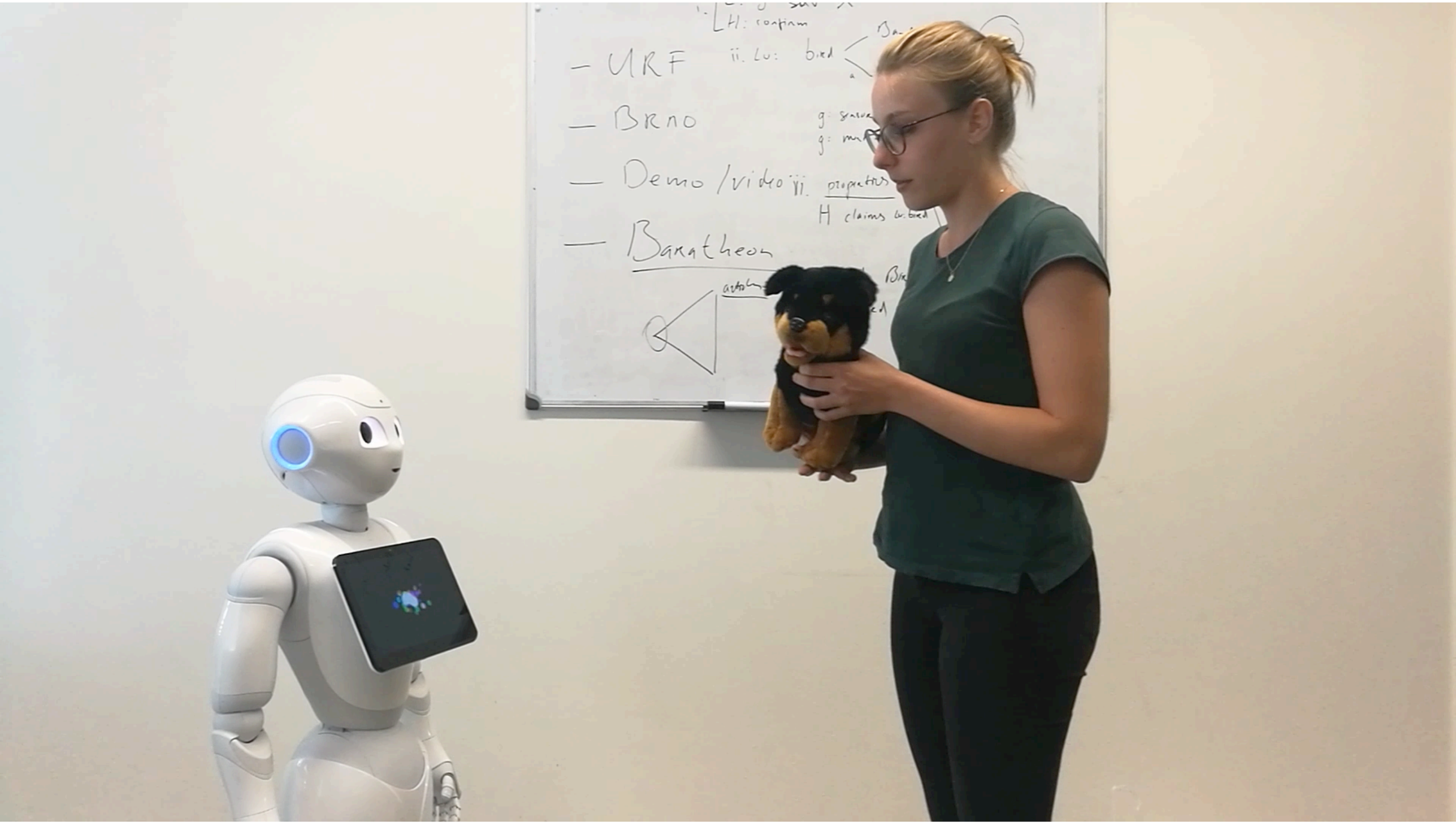


Perceiving the world

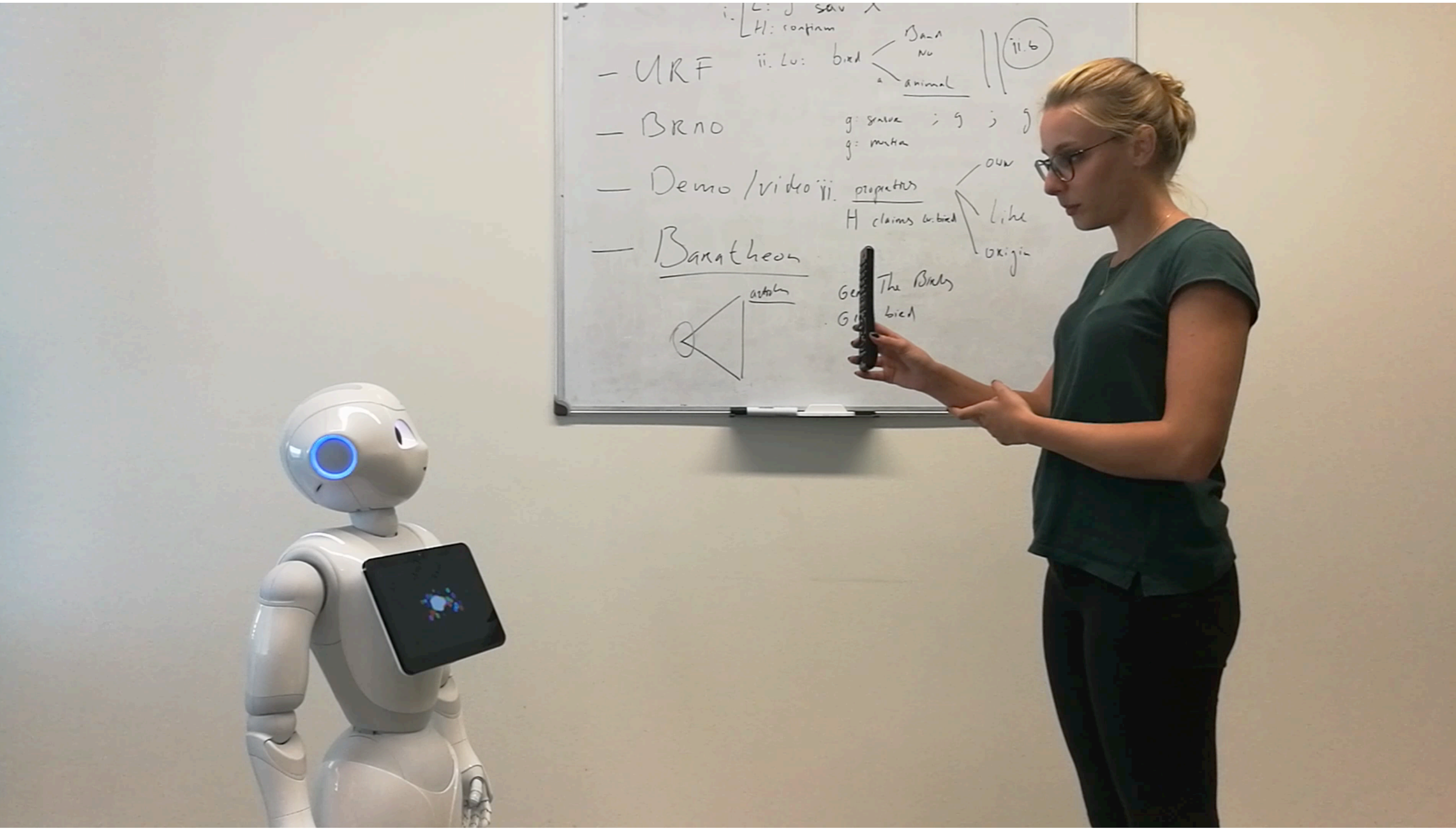


Learning about the world

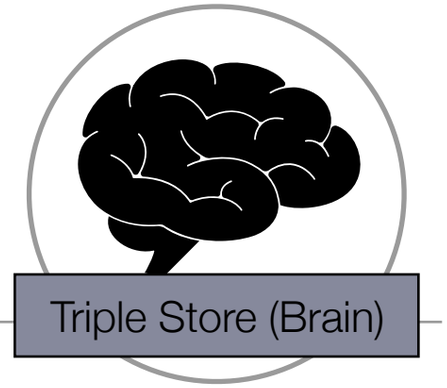


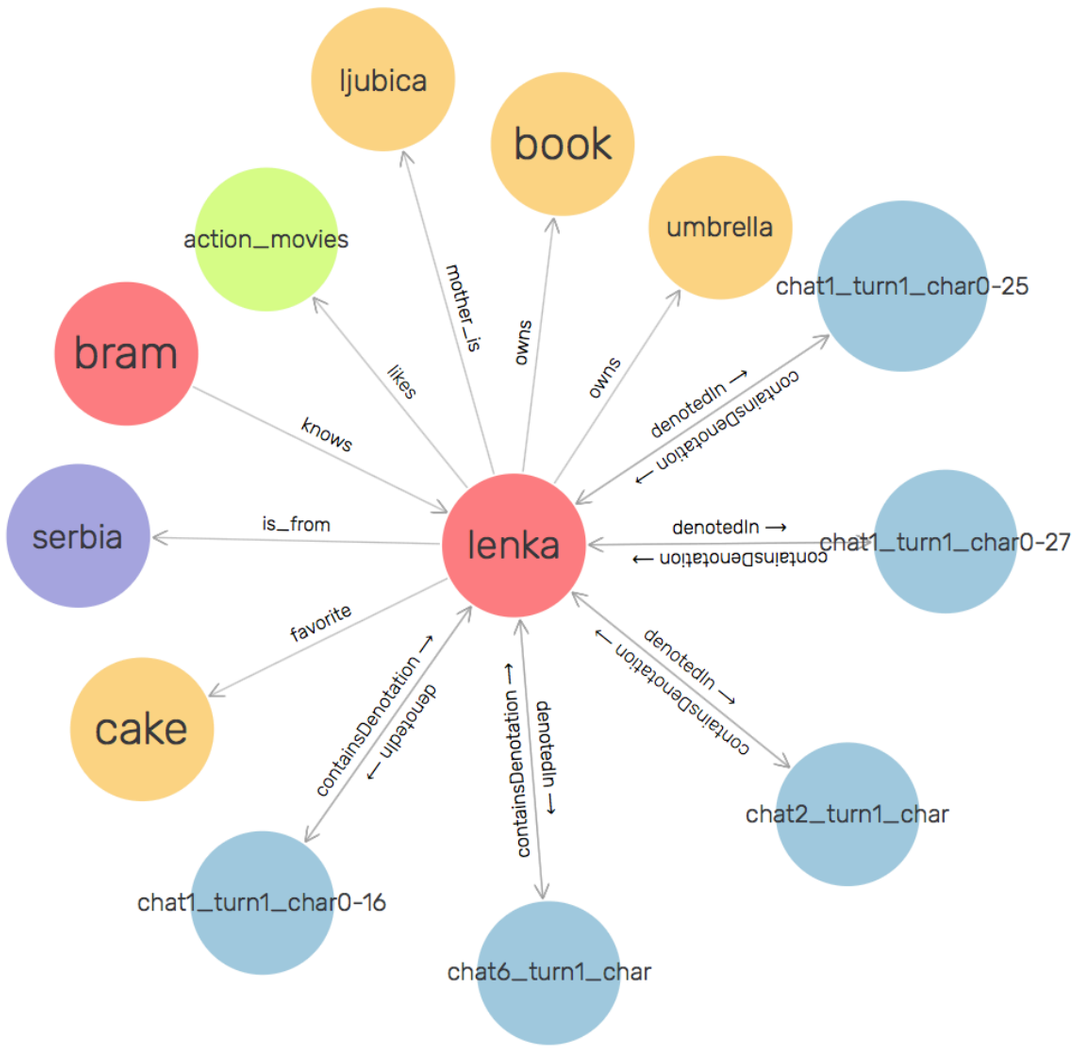


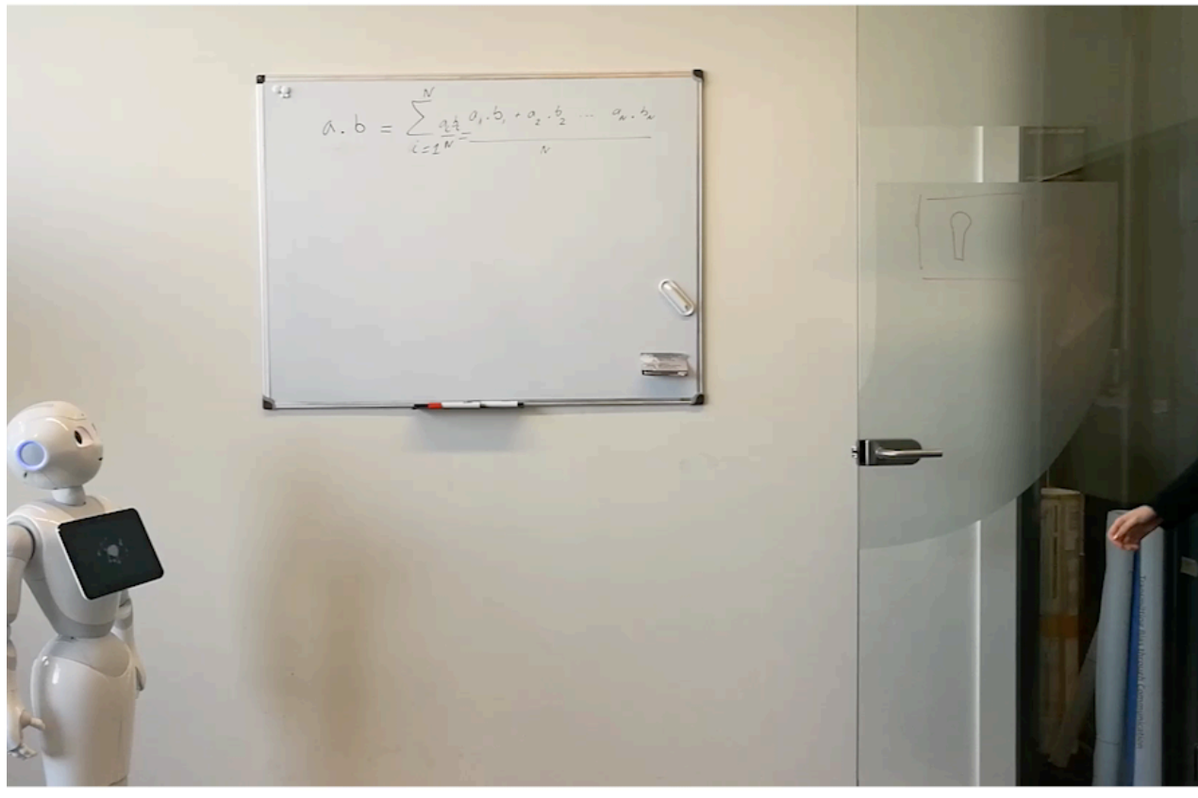
Learning about the world



Learning about people







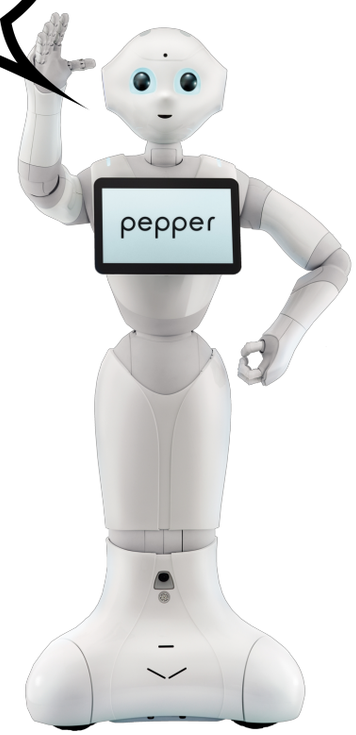
Communicating about the world

Natural Language Processing

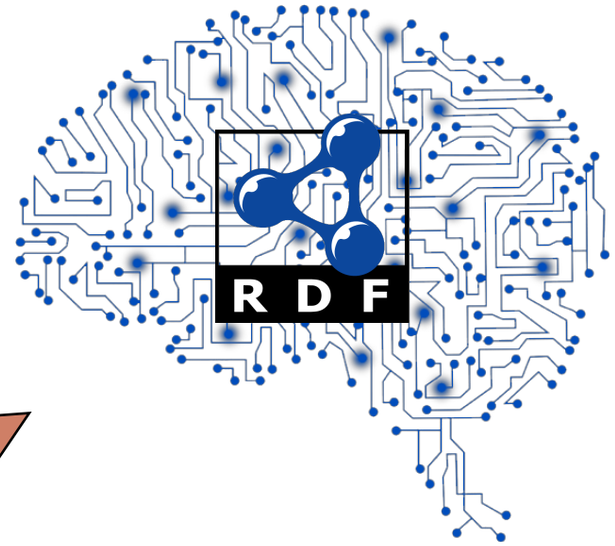
Language Processing

I like pizza too, Bram!

I like pizza!



RDF triple(s)



subject: “Bram”
predicate: “likes”
object: “pizza”

RDF triple

Utterance
Chat ID
Speaker info

Statement

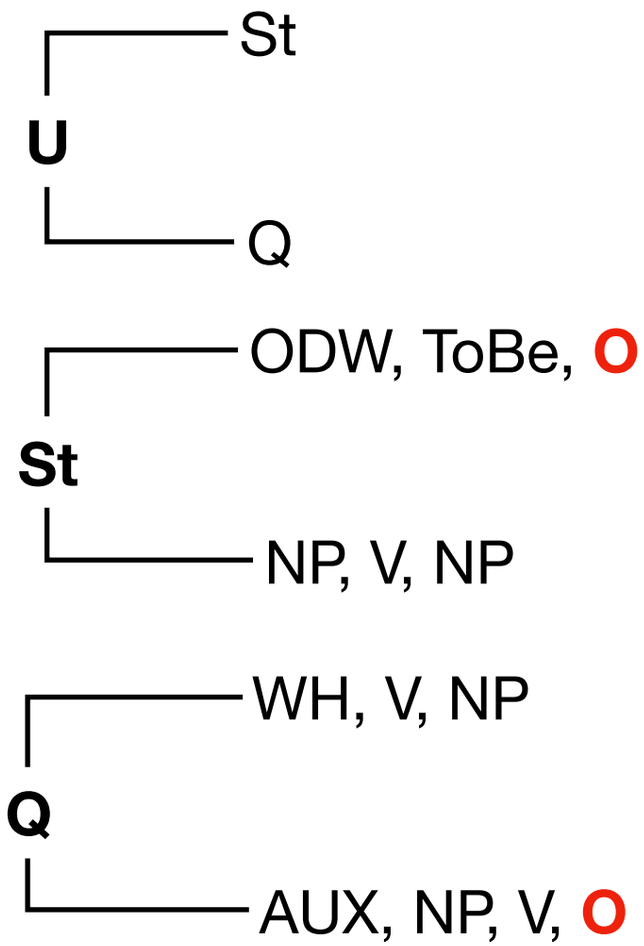
Question

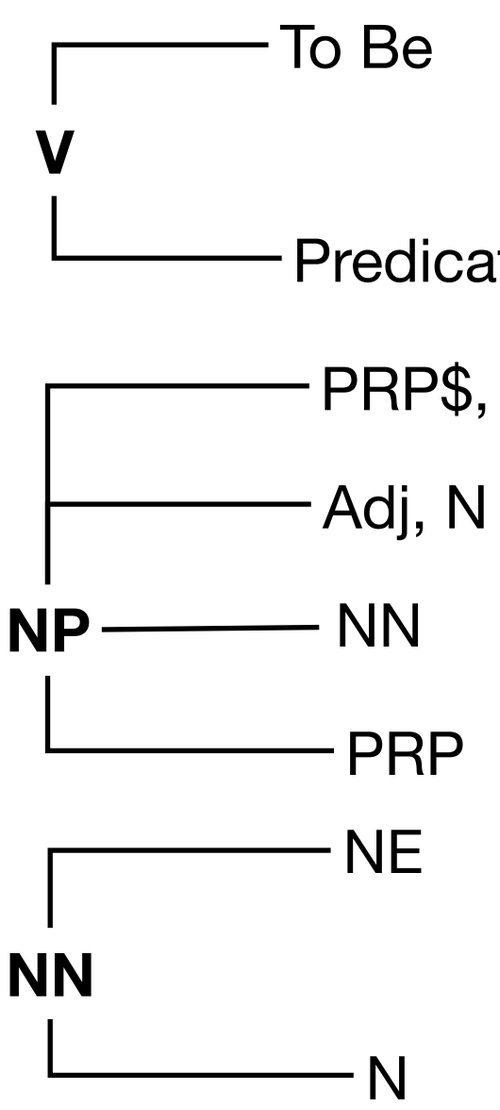
WH-question

Verb question

SPARQL query

Perceptual phrase grammar





ODW → there, this, it

O → objects from COCO

ToBe → present tense

P → likes, sees, knows, …

PRP$ → my, your, his, hers, …

NE → Stanford NE, names of friends

N, Adj → can be anything

- Grammar (SVO order, deictic relations) to detect references (speaker & hearer identification, entities, **objects detected!**) and subject-property-object relations
- Resolving self-reference (Leolani = I/me), speaker reference (you) in language generation

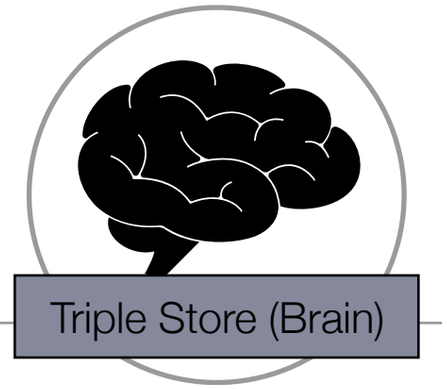
Learning from conversation

Pepper's Brain

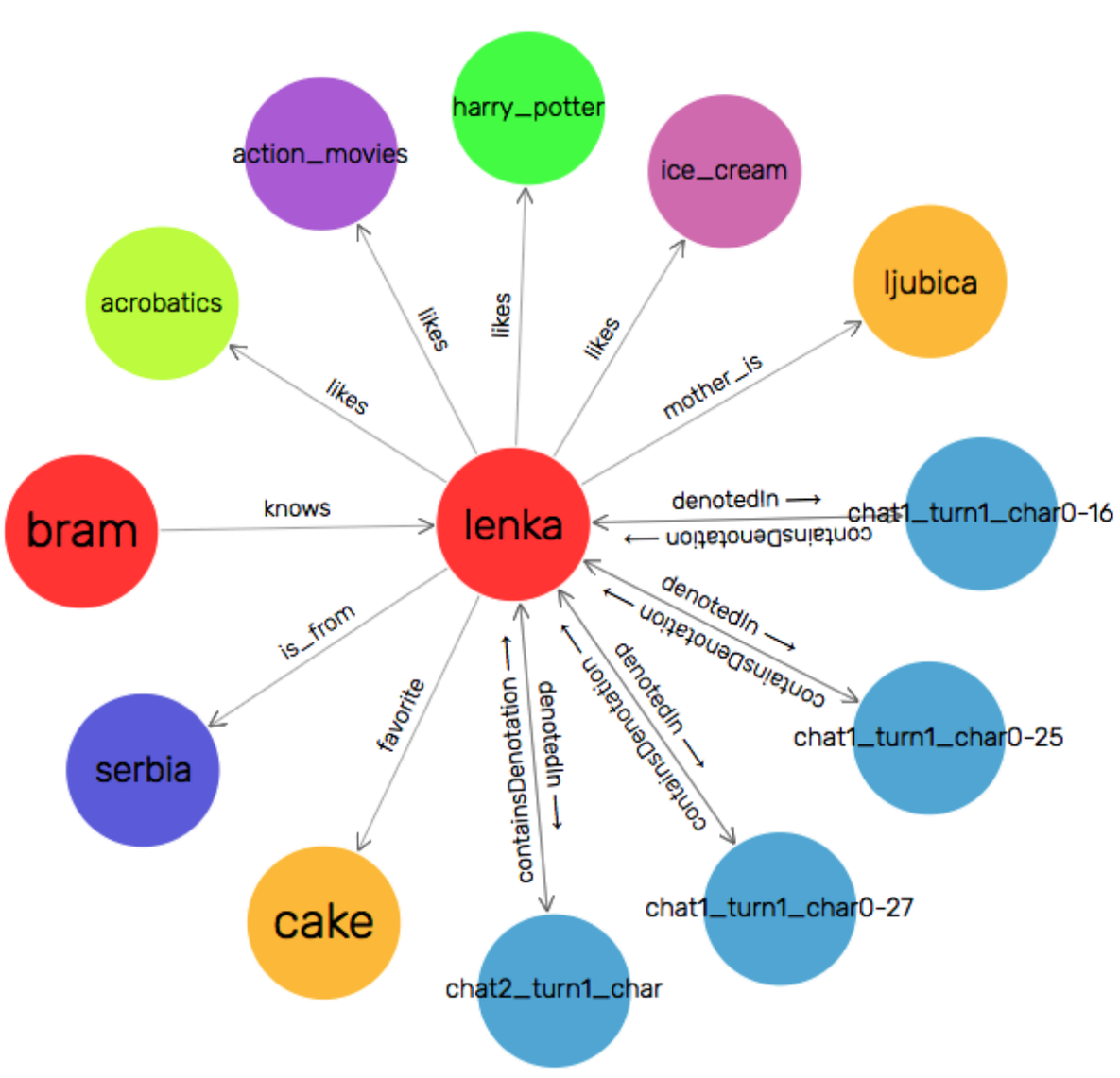
- **Perceived objects:**
  - Learning about perceived objects from the web
  - Telling her what an observed object is: *a dog is also a pet*
- **Any string can be the value as an object of a predicate** *(no slot types)***:**
  - Bram likes romantic movies, science fiction movies, chocolate, pizza, ….
  - Next any object can become subject of a typing triple: *tweet, mail, post* subClassOf *message*
- ***"X can do-action Y/X can have-property Y"* used to learn new properties:**
  - You can ***send/email/receive/read***; book can have ***colour***
- ***"X is same as Y"* used to learn synonymous expressions:**
  - *mail* is same as *send*

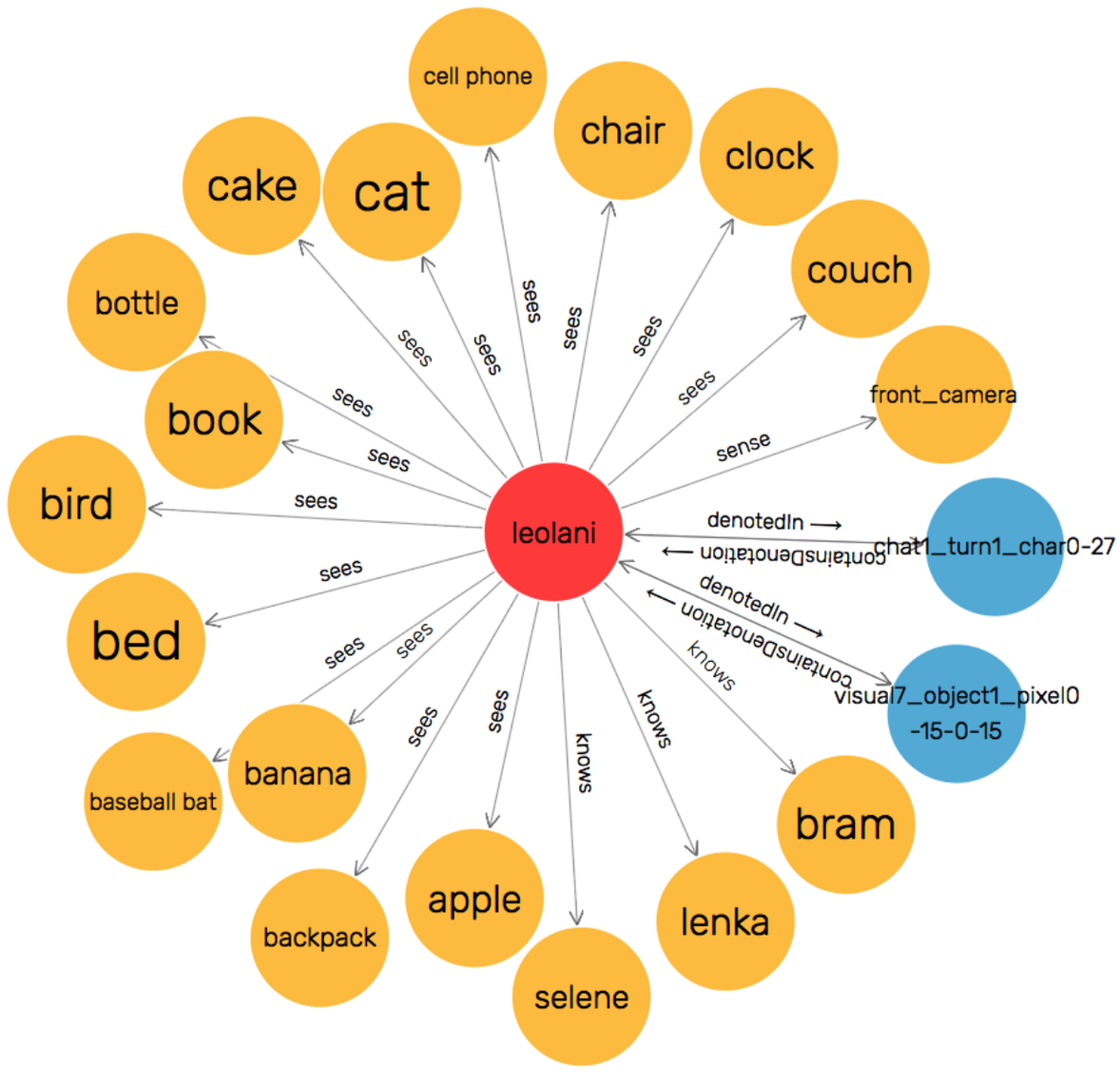
Storing Knowledge

Pepper's Brain



- Circles: ***Things***
- Arrows: ***Relations*** between ***Things***





What was seen when?

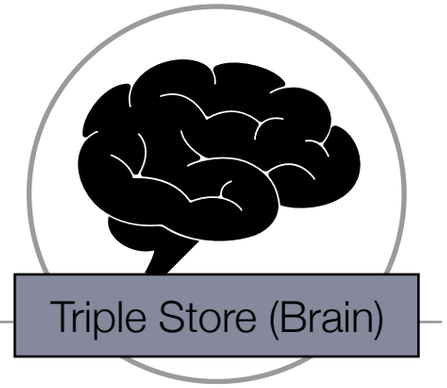
```
PREFIX sesame: <http://www.openrdf.org/schema/sesame#>
PREFIX grasp: <http://groundedannotationframework.org/grasp#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX sem: <http://semanticweb.cs.vu.nl/2009/11/sem/>

SELECT ?node ?a ?t ?s {
    ?node rdf:type grasp:Experience .
    ?node sem:hasTime ?t .
    ?node sem:hasActor ?a .
    ?node grasp:denotedBy ?s
} ORDER BY ?node
```

What was seen when?

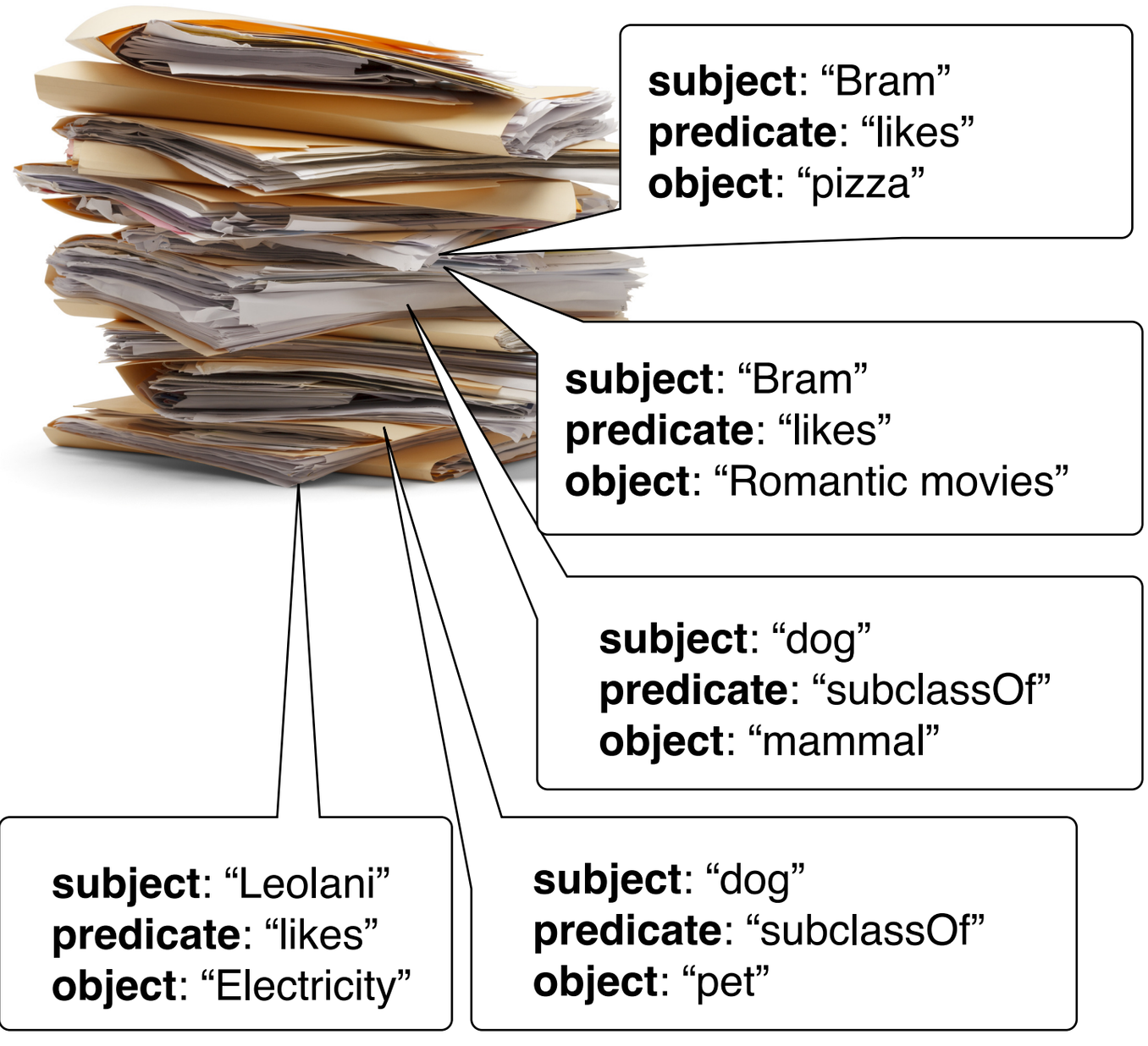
	node	a	t	s
1	leolaniWorld:leolani-sees-apple	leolaniInputs:front_camera	leolaniTime:2018-08-31	leolaniTalk:visual7_object1_pixel0-15-0-15
2	leolaniWorld:leolani-sees-backpack	leolaniInputs:front_camera	leolaniTime:2018-08-31	leolaniTalk:visual7_object1_pixel0-15-0-15
3	leolaniWorld:leolani-sees-banana	leolaniInputs:front_camera	leolaniTime:2018-08-31	leolaniTalk:visual7_object1_pixel0-15-0-15
4	leolaniWorld:leolani-sees-baseball_bat	leolaniInputs:front_camera	leolaniTime:2018-08-31	leolaniTalk:visual7_object1_pixel0-15-0-15
5	leolaniWorld:leolani-sees-bed	leolaniInputs:front_camera	leolaniTime:2018-08-31	leolaniTalk:visual7_object1_pixel0-15-0-15
6	leolaniWorld:leolani-sees-bird	leolaniInputs:front_camera	leolaniTime:2018-08-31	leolaniTalk:visual7_object1_pixel0-15-0-15
7	leolaniWorld:leolani-sees-book	leolaniInputs:front_camera	leolaniTime:2018-08-31	leolaniTalk:visual7_object1_pixel0-15-0-15
8	leolaniWorld:leolani-sees-bottle	leolaniInputs:front_camera	leolaniTime:2018-08-31	leolaniTalk:visual7_object1_pixel0-15-0-15
9	leolaniWorld:leolani-sees-cake	leolaniInputs:front_camera	leolaniTime:2018-08-31	leolaniTalk:visual7_object1_pixel0-15-0-15
10	leolaniWorld:leolani-sees-cat	leolaniInputs:front_camera	leolaniTime:2018-08-31	leolaniTalk:visual7_object1_pixel0-15-0-15
11	leolaniWorld:leolani-sees-cell_phone	leolaniInputs:front_camera	leolaniTime:2018-08-31	leolaniTalk:visual7_object1_pixel0-15-0-15
12	leolaniWorld:leolani-sees-chair	leolaniInputs:front_camera	leolaniTime:2018-08-31	leolaniTalk:visual7_object1_pixel0-15-0-15
13	leolaniWorld:leolani-sees-clock	leolaniInputs:front_camera	leolaniTime:2018-08-31	leolaniTalk:visual7_object1_pixel0-15-0-15
14	leolaniWorld:leolani-sees-couch	leolaniInputs:front_camera	leolaniTime:2018-08-31	leolaniTalk:visual7_object1_pixel0-15-0-15

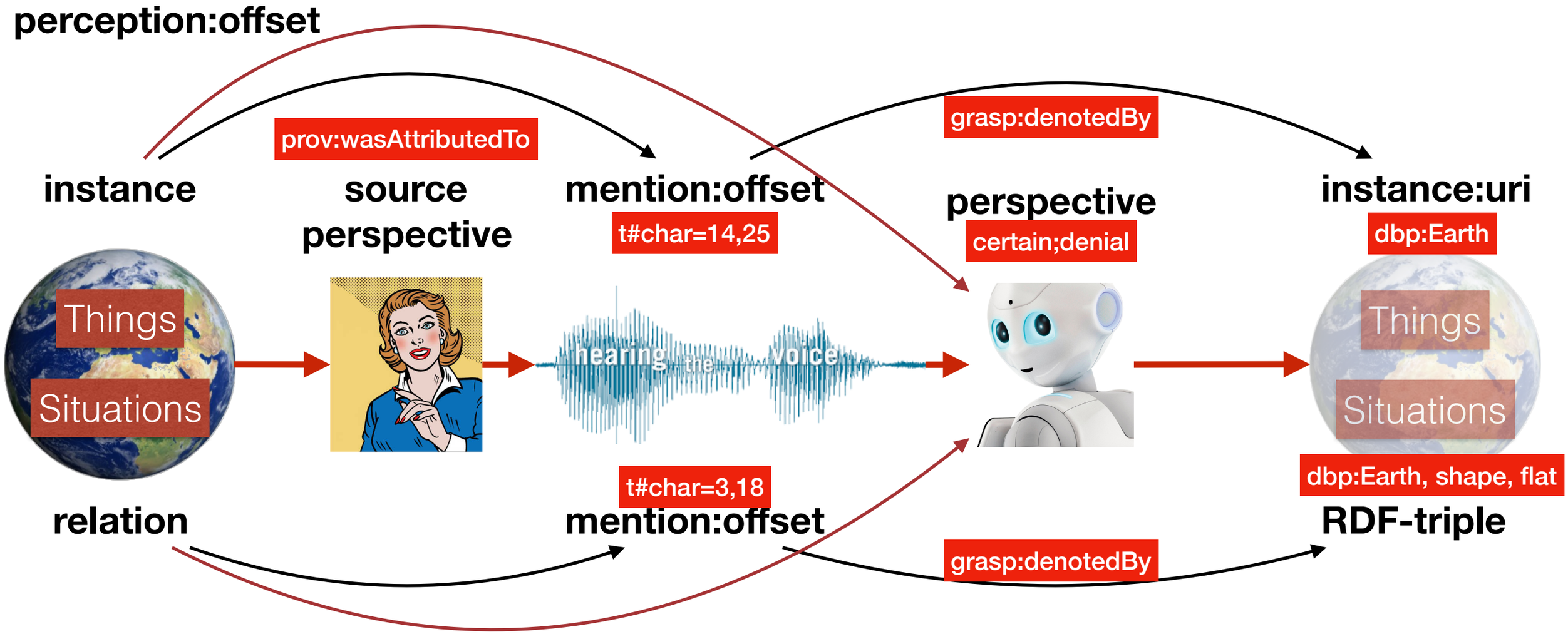
The brain is just a huge collection of RDF triples

But should she believe everything said and seen ….



- We do not just store triples as facts…
- but also who said it, when and in which context
- ***Leolani needs a so-called theory of mind***





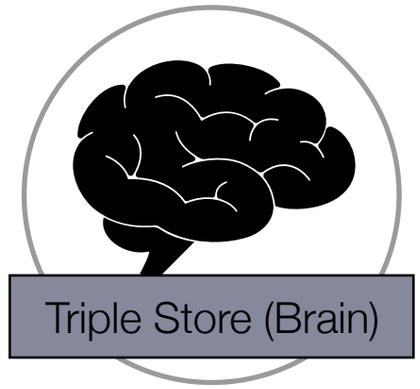
GRaSP

Grounded **R**epresentation **a**nd **S**ource **P**erspective

https://github.com/cltl/GRaSP

A. Fokkens, P. Vossen, M. Rospocher, R. Hoekstra, and W. van Hage, “Grasp: grounded representation and source perspective,” in *Proceedings of knowrsh*, Varna, Bulgaria, 2017.

Statement



Claim

Instance	Property	Value
Bram	actor	laugh

Instance: Lenka

Source

Perspective: confirm, uncertain

face-detection

mood-detection

Natural Language Processing

“Bram is laughing”

Utterance

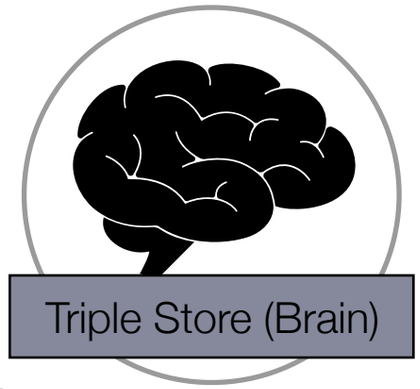
20180512T10:05am.

leolaniWorld:instances		
leolaniWorld:Lenka	rdfs:label	“Lenka”;
leolaniWorld:Bram	rdfs:label	“Bram”;
	grasp:denotedIn	leolaniTalk:chat1_turn1_char0-16.
leolaniWorld:laugh	a	sem:Event;
	rdfs:label	“laugh”;
	grasp:denotedIn	leolaniTalk:chat1_turn1_char0-16.

leolaniWorld:claims		
leolaniWorld:claim1	a	grasp:Statement;
	grasp:subject	leolaniWorld:laugh;
	grasp:predicate	sem:hasActor;
	grasp:object	leolaniFriends:Bram.

leolaniTalk:perspectives		
leolaniTalk:chat1_turn1	a	grasp:Turn;
	sem:hasActor	leolaniFriends:Lenka;
	sem:hasTime	leolaniTime:20180512T10:05am.
leolaniTalk:chat1_turn1_char0-16	a	grasp:Mention;
	grasp:denotes	leolaniWorld:claim1 ;
	prov:wasDerivedFrom	leolaniTalk:chat1_turn1 ;
	prov:wasAttributedTo	leolaniFriends:Lenka .
leolaniTalk:chat1_turn1_char0-16_ATTR1	a	grasp:Attribution;
		grasp:CONFIRM,
	rdf:value	grasp:UNCERTAIN,
		grasp:SURPRISE;
	grasp:isAttributionFor	leolaniTalk:chat1_turn1_char0-16.

Statement



Claim

Instance	Property	Value
Bram	actor	laugh

Instance: Selene

Source

Perspective: denial, certain

face-detection

mood-detection

Natural Language Processing

“No, Bram is not laughing”

Utterance

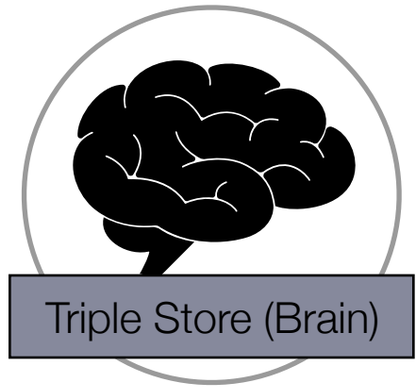
20180512T10:06am.

leolaniWorld:claims		
leolaniWorld:claim1	a	grasp:Statement;
	grasp:subject	leolaniWorld:laugh;
	grasp:predicate	sem:hasActor;
	grasp:object	leolaniFriends:Bram.

leolaniTalk:perspectives		
leolaniTalk:chat1_turn1	a	grasp:Turn;
	sem:hasActor	leolaniFriends:Lenka;
	sem:hasTime	leolaniTime:20180512T10:05am.
leolaniTalk:chat1_turn1_char0-16	a	grasp:Mention;
	grasp:denotes	leolaniWorld:claim1 ;
	prov:wasDerivedFrom	leolaniTalk:chat1_turn1 ;
	prov:wasAttributedTo	leolaniFriends:Lenka .
leolaniTalk:chat1_turn1_char0-16_ATTR1	a	grasp:Attribution;
		grasp:CONFIRM,
	rdf:value	grasp:UNCERTAIN,
		grasp:SURPRISE;
	grasp:isAttributionFor	leolaniTalk:chat1_turn1_char0-16.

leolaniTalk:perspectives		
leolaniTalk:chat2_turn1	a	grasp:Turn;
	sem:hasActor	leolaniFriends:Selene;
	sem:hasTime	leolaniTime:20180512T10:06am.
leolaniTalk:chat2_turn1_char0-24	a	grasp:Mention;
	grasp:denotes	leolaniWorld:claim1 .
	prov:wasDerivedFrom	leolaniTalk:chat2_turn1 .
	prov:wasAttributedTo	leolaniFriends:Selene .
leolaniTalk:chat2_turn1_char0-24_ATTR1	a	grasp:Attribution;
	rdf:value	grasp:DENY, grasp:CERTAIN;
	grasp:isAttributionFor	leolaniTalk:chat2_turn1_char0-24.

Statement



Claim

Instance	Property	Value
Bram	actor	laugh

Instance

Lenka

Source

Perspective

denial, certain

face-detection

mood-detection

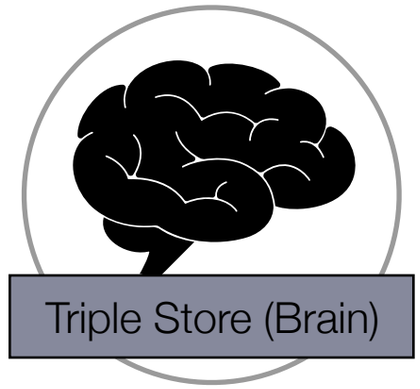
Natural Language Processing

“You are right, he is not”

Utterance

20180512T10:07am.

Perception

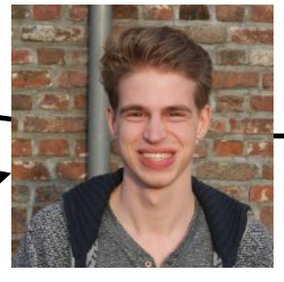


Perception

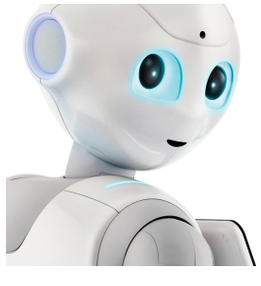
Instance	Property	Value
Bram	mood	joy

Perspective
confirm, uncertain

face-detection



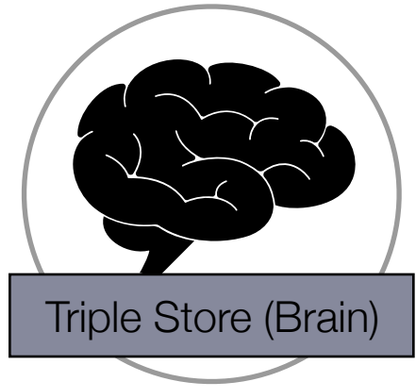
mood-detection



Perception

20180512T10:05am.

Brain



Utterance

20180512T**10:05**am.

Perspective

confirm, uncertain

Claim

Instance	Property	Value
Bram	actor	laugh

Utterance

20180512T**10:07**am.

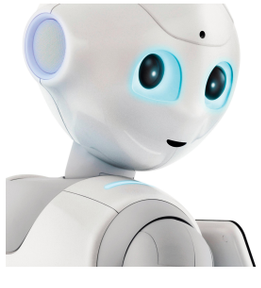
Utterance

20180512T**10:06**am.

Perspective

denial, certain

Belief



Perception

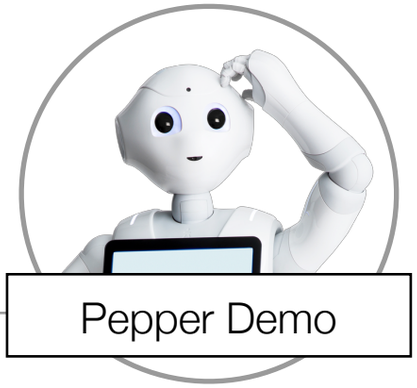
20180512T10:05am.

Perspective

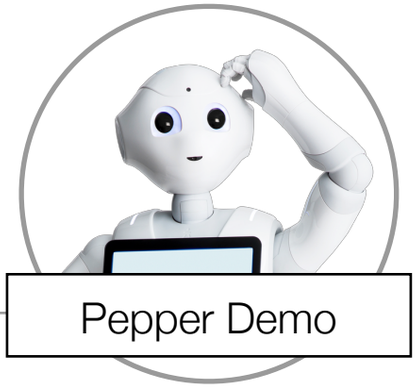
confirm, uncertain

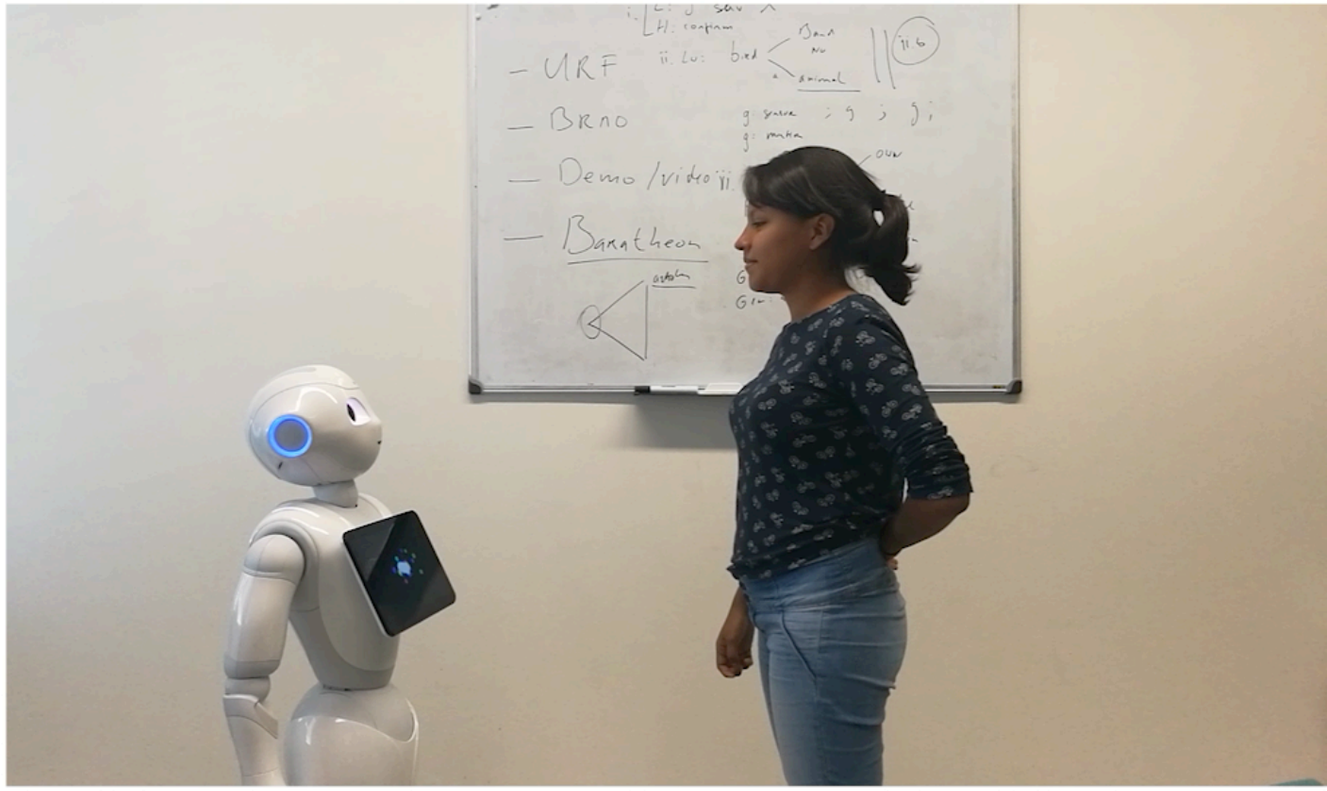
Instance	Property	Value
Bram	actor	joy

Disagreeing about properties



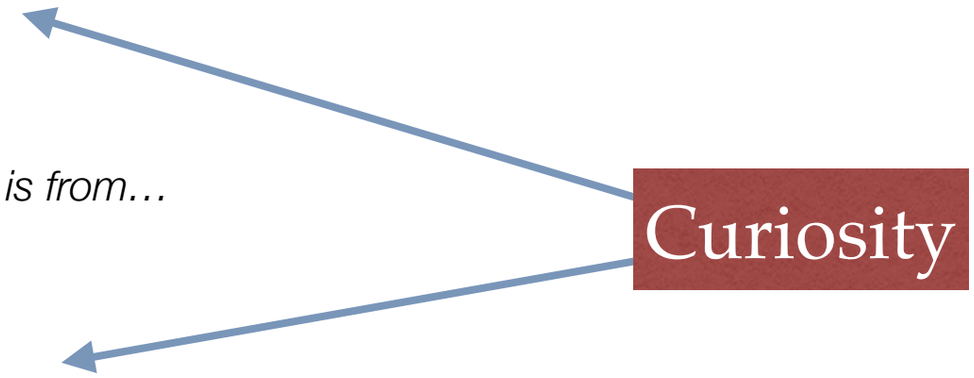
Telling different things about objects





Generation perspective: drive to communicate

- **Knowing people** (through face recognition)
  - Refer to latest conversation or perception:
    - *seeing Bram while Lenka just told her where Bram is from…*
  - Get to know new people
- **Fill gaps in knowledge:**
  - *Where are you from?*
  - *What is pizza? Where is a pizza from?*
  - Typical things people do/like/have/own (profiling), ask for the obvious
    - *Do you like romantic movies?*
- **Resolve uncertainties and conflicts** *(people are confusing)***:**
  - *Romantic movies or science fiction movies*
  - *Lenka told me, she owns a book, you told me you own a book*



What to communicate: properties of answers

- Coming from the brain or from the web:
  - *I saw a dog, people told me, I looked on the web….*
- Multiple results, one result, no result:
  - *You are the first person I met from Serbia*
- Yes/no questions:
  - *I did not know that/ yes I do/ I do not know but I can look for it on the web*
- Provenance and source:
  - *who said when what, Yesterday you told me that Bram….*
  - *which perception, Look I saw a dog yesterday and learned from the web it is a mammal*

But there is more:

from detection to a world model….

- Relevance
- Permanence
- Disambiguating reference and quantification
- Situational and personal contexts and knowledge

Relevance

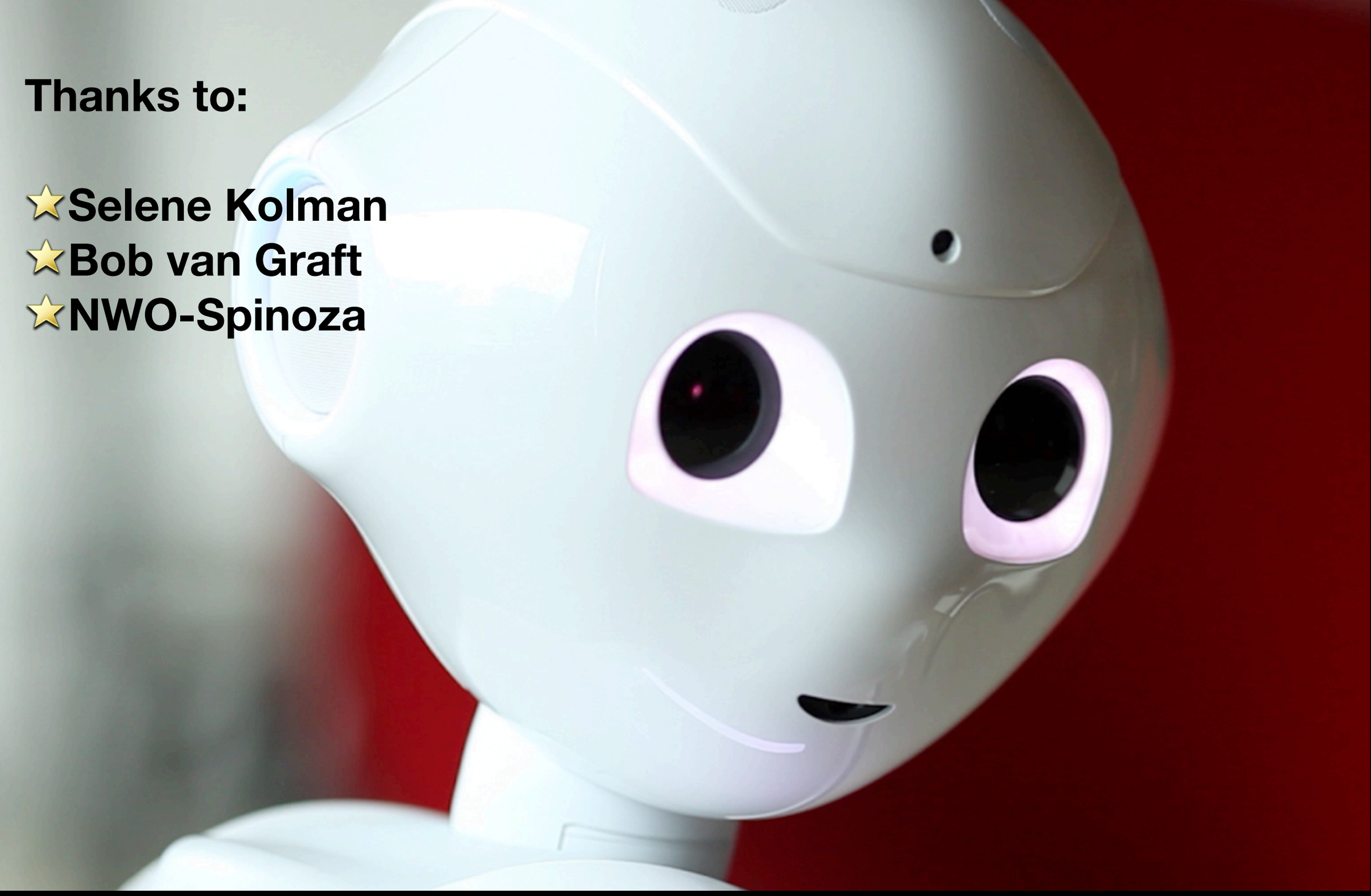
- **3 images per second (adjustable),**
  - *but not everything she sees is relevant*
- Reasons to **remember people**: *saying hello, giving consent, your name*
- Reasons to **remember objects** (create a URI, label and image vectors):
  - Somebody recently attributed properties to an object: e.g. *ownership*, *type*
  - You start talking about an object in the current conversation: *do you see a clock?*
  - Corrected image recognition: *no, this is an apple*
  - Learning properties of a new object from the web: *a dog is a mammal*
  - Conflict or uncertainty between robot and human or across humans
- ***Otherwise: ignore the images and the people ….***

Permanence

- Keep multiple objects of the same type apart?
  - *Lenka's book, Selene's book*
- How to recognise objects from previous interaction?
  - ***Count*** *the books, the perceptions of a book, copies of a book*
- How to know that the same object changed location?
- How to detect properties of objects?
  - *Conflicting (where are you from? ownership)*
  - *Discriminating (age, gender, colour, size) versus variable (ownership, location, condition)*

Permanence

- How to connect object recognition to personal relationships and previous experiences?
  - Mixing small data with big data:
    - *My office should be known after a few sessions*
    - *Learning personal vocabulary and ways of reference*
- Personal & situational disambiguation:
  - *in my office it is likely my coffee mug*



Thanks to:

- ★Selene Kolman
- ★Bob van Graft
- ★NWO-Spinoza

Reference:

- P. Vossen, S. Báez, L. Bajčetić, and B. Kraaijeveld, (2018) “Leolani: a reference machine with a theory of mind for social communication, invited keynote speech,” in *Proceedings of TSD-2018, Brno.*
- http://www.understandinglanguagebymachines.org/tsd-2018-brno/

VU VRIJE UNIVERSITEIT AMSTERDAM

Understanding Language by Machines