DANIEL SONNTAG

RESEARCH FELLOW TALK

MULTIMODAL–MULTISENSOR INTERFACES IN INDUSTRIAL AND MEDICAL APPLICATION DOMAINS

FIRST PART OF ABSTRACT

- Multimodal-multisensor interfaces combine one or more user input modalities with sensor information (e.g., location, proximity, tilt). Sensor-based cues may be used to interpret a user's physical state, health status, mental status, current context, engagement in activities, and many other types of information.
- Sensor input aims to transparently facilitate user-system interaction, and adaptation to users' needs. The type and number of sensors incorporated into multimodal interfaces has been expanding rapidly, resulting in explosive growth of multimodal-multisensor interfaces.
- **We believe that multimodal-multisensor interfaces can be designed to most effectively advance human performance during the next decade.**
- We demonstrate this by recent DFKI projects in industrial and medical application domains.

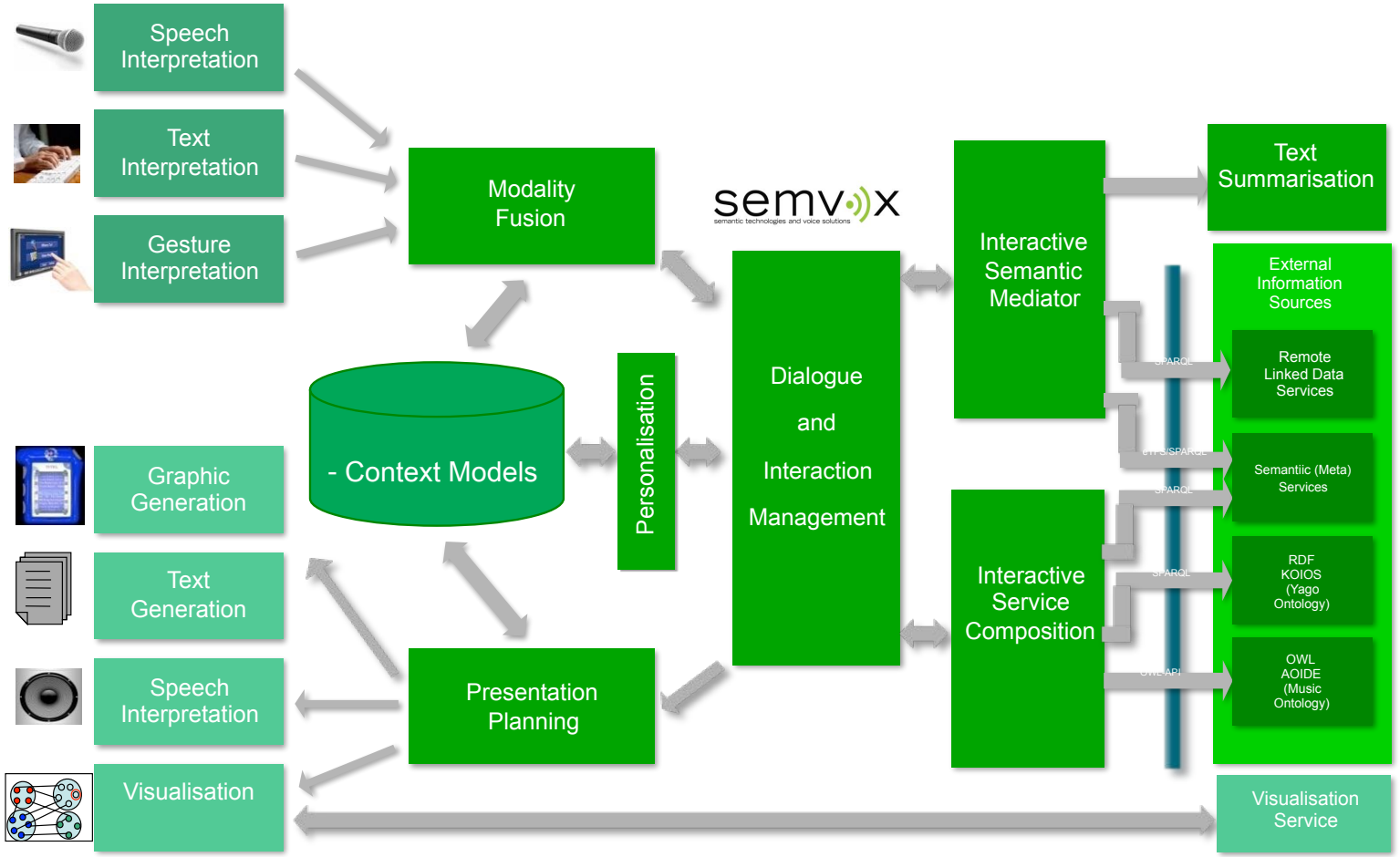
IUI AND ADVANCES IN INFORMATION EXTRACTION: FROM TEXT AND IMAGE TO KNOWLEDGE

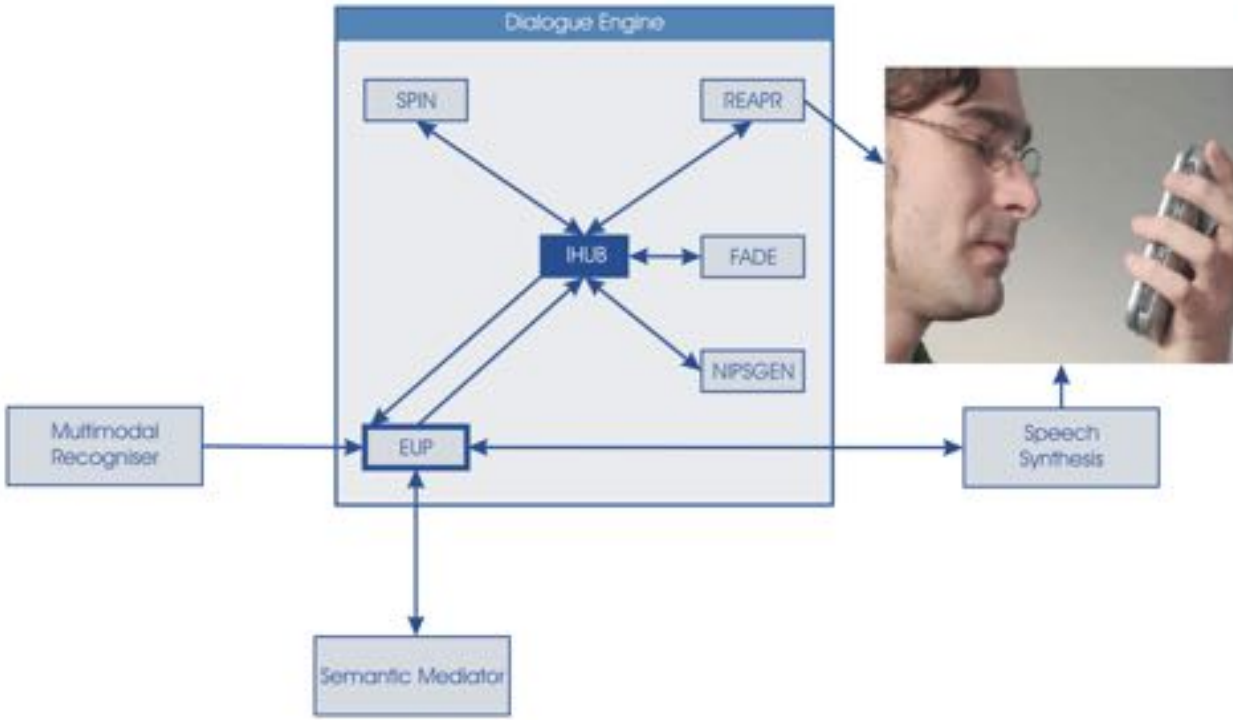
- http://www.dfki.de/%7Esonntag/courses/WS15/IUI.html
- http://www.dfki.de/~sonntag/courses/SS14/IE.html

Who was world champion in 1990 ?

Question Answering Functionality

MULTIMODAL DIALOGUE SHELL WORKFLOW





Definition 5 Knowledge Base

A Knowledge Base (KB) is a structure

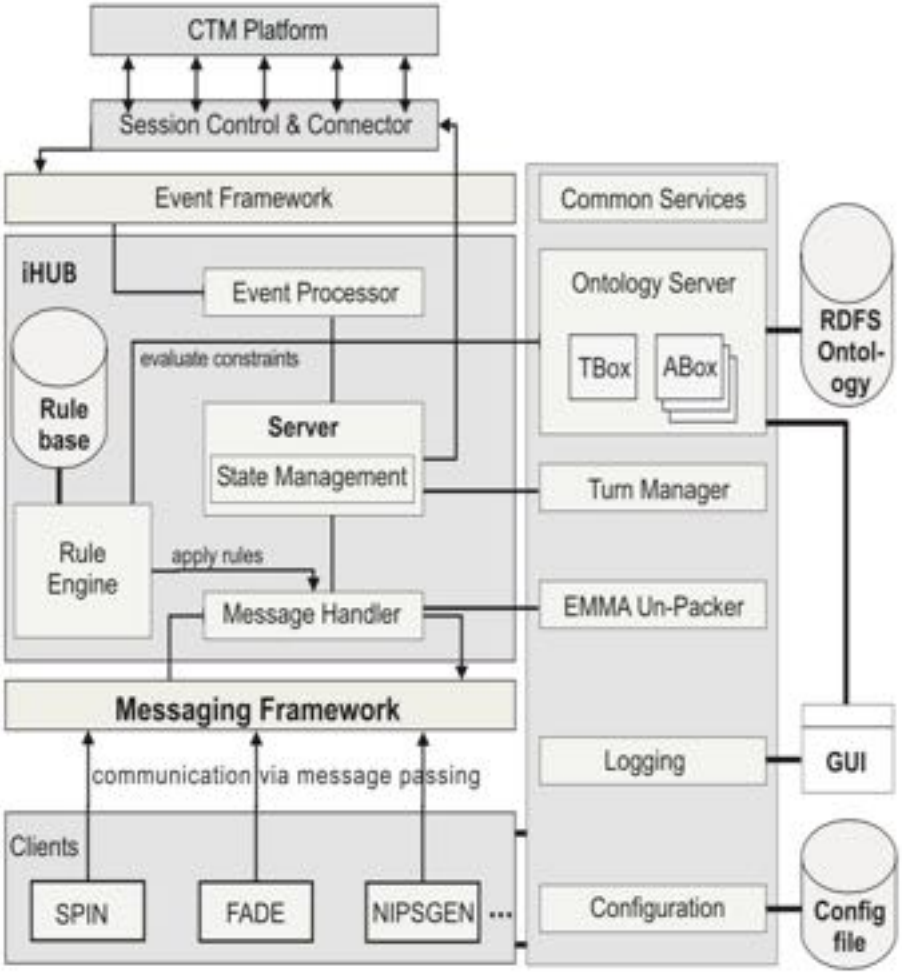
$$KB := (C_{KB}, R_{KB}, I, \imath_C, \imath_R) \tag{2.12}$$

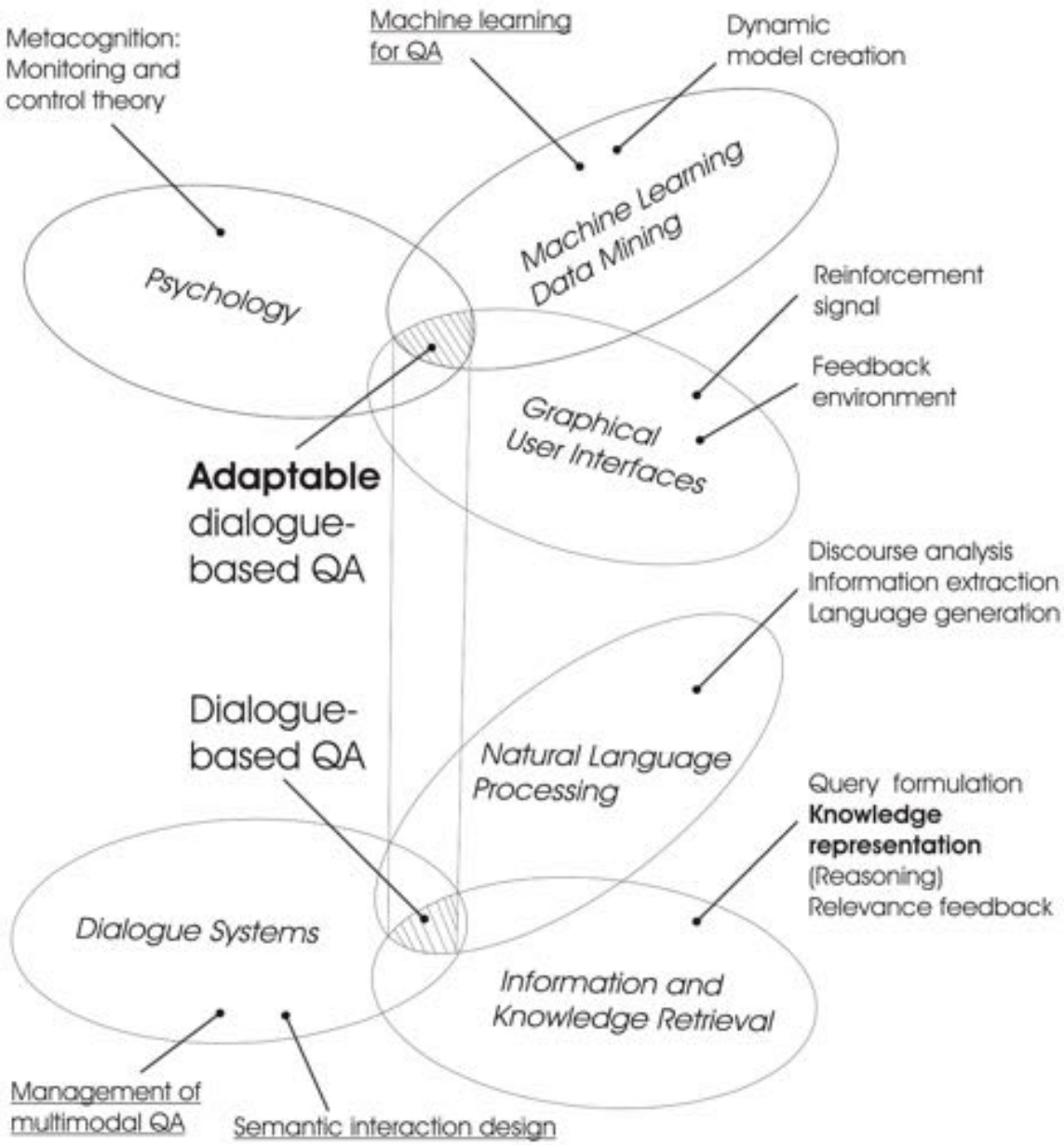
consisting of

- the sets C_{KB} and R_{KB} of supported concepts and relations within the knowledge base;
- the set I of instance identifiers (also called instances or objects);
- a function $\imath_C : C_{KB} \rightarrow 2^I$ called *concept instantiations*;
- a function $\imath_R : R_{KB} \rightarrow 2^{I^+}$ called *relation instantiations*; $\forall r \in R : \imath_R(r) \subseteq \prod_{1 \leq i \leq |\sigma(r)|} \imath_C(\pi_i(\sigma(r)))$
- a function $\imath_A : R_{KB} \rightarrow I \times v_p(\pi_2(\sigma(a)))$ called *attribute instantiation*, with $\imath_A(r) \subseteq \imath_C(\pi_1(\sigma(a))) \times v_p(\pi_2(\sigma(a)))$; $v_p()$ being the values of the primitive datatypes $p \in P$.

MULTIMODAL DIALOGUE SYSTEMS

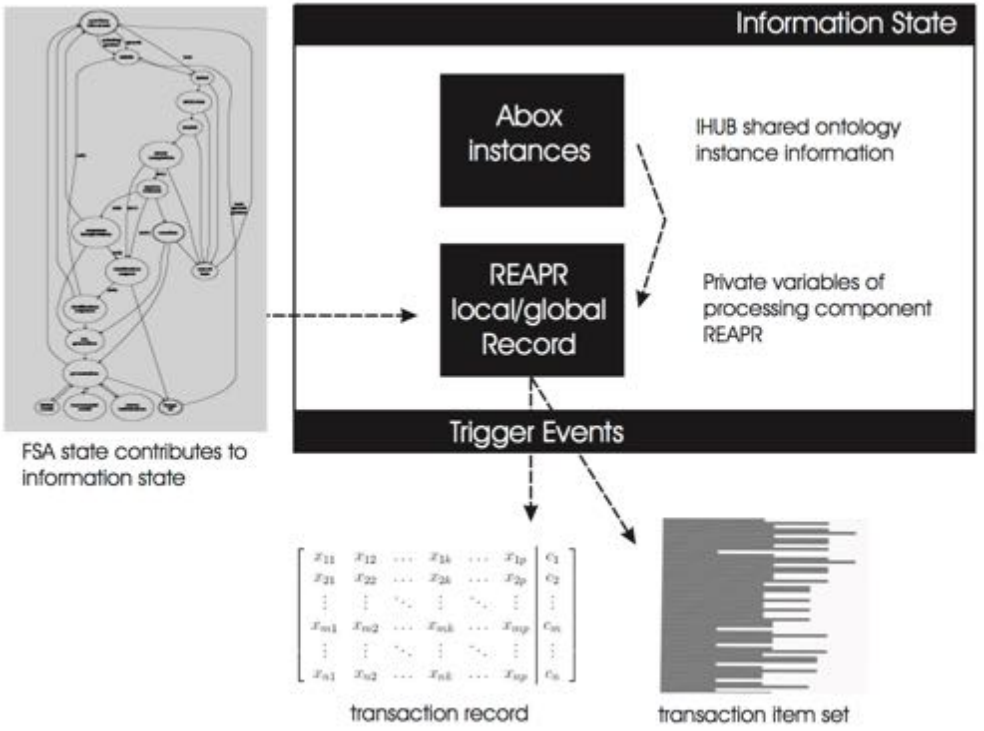
ONTOLOGIES AND ADAPTIVITY IN DIALOGUE FOR QUESTION ANSWERING



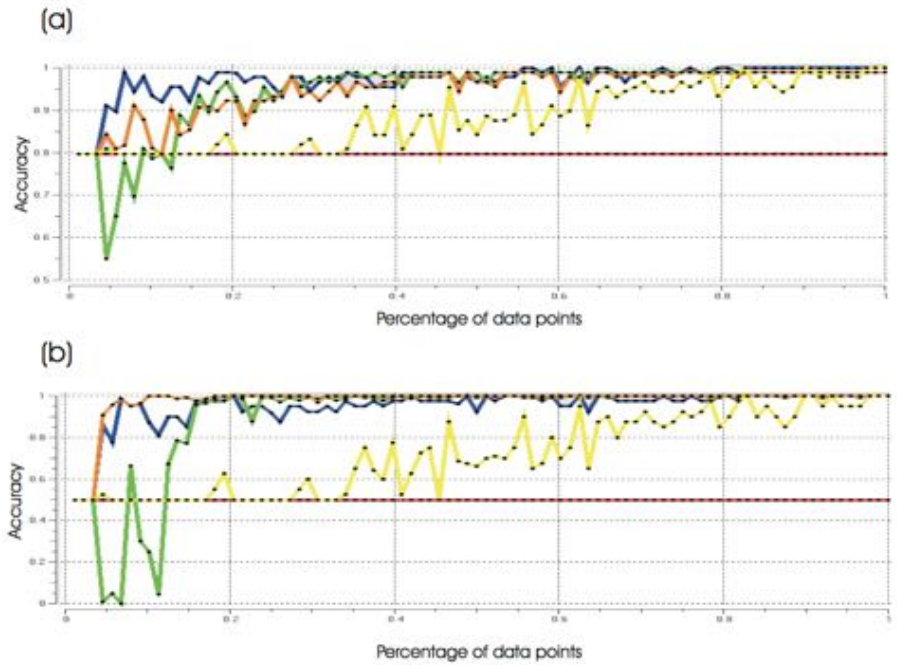


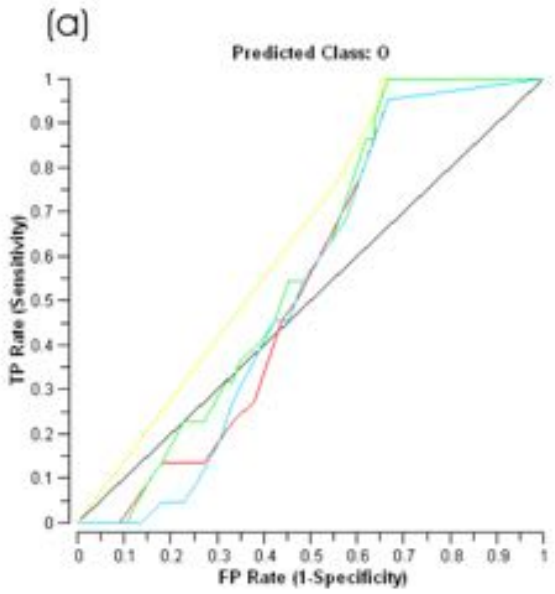
INFORMATION STATE

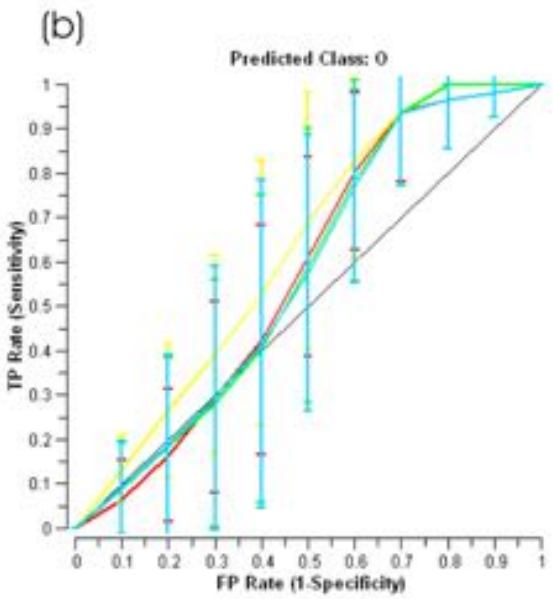
Feature Class	IS State Features
MMR	*Listening, Recording, Barge-in, Last-ok, Input dominance (text or voice)*
NLU	*Confidence, Domain relevance*
Query	*Dialogue act, Question Foci, Complexity, Context object, Query text*
Fusion	*Fusion act, Co-reference resolution*
Answer	*Success, Speed, Answer streams, Status, Answer type, Content, Answer text*
Manager	*Turn/Task numbers, Idle states, Waiting for Results, User/system turn, Elapsed times: input/output, Dialogue act history (system and user) e.g. reject, accept, clarify*

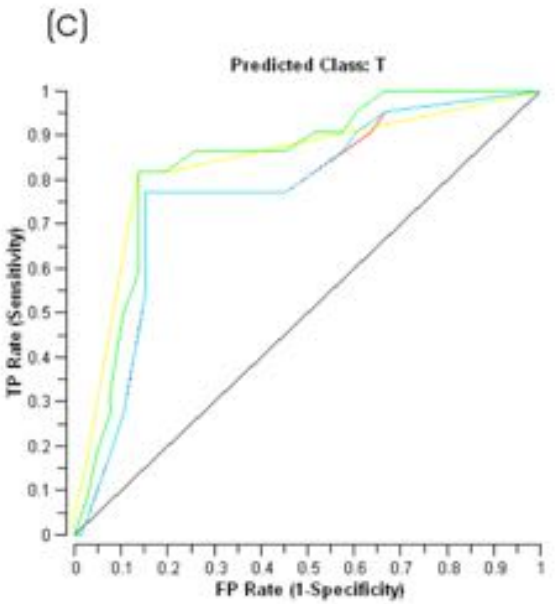


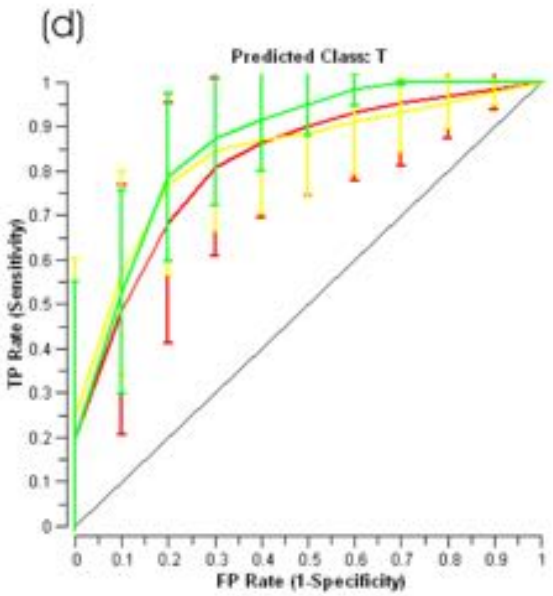
ANSWER STREAM PREDICTION





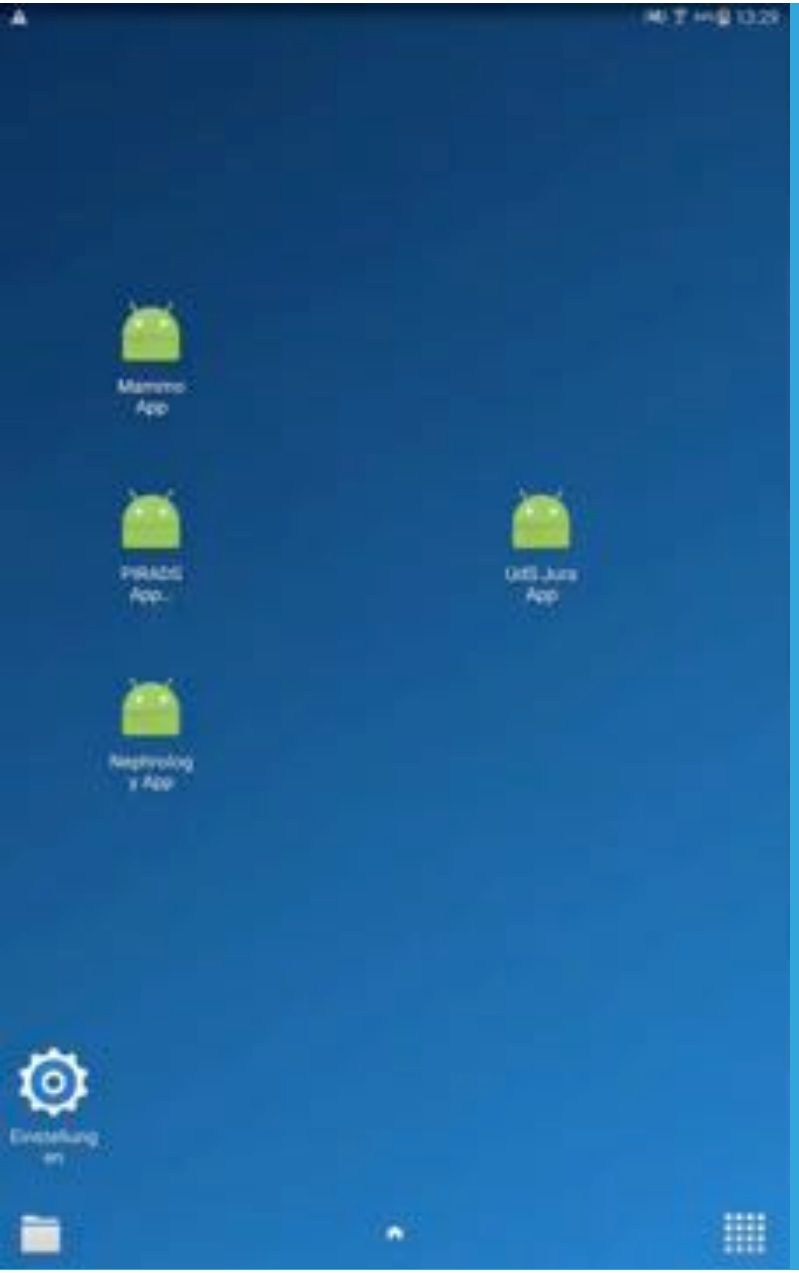






- Make dialogue-based radiology image reporting possible
- Reduce turn-over times and annotation errors
- Facilitate structured reporting

With the iPad's FDA approval, a breakthrough for mobile medical imaging, especially in the U.S., can be expected. With RadSpeech, we aim to build the next generation of intelligent, scalable, and user-friendly mobile semantic search and image annotation interfaces for the medical imaging domain.



DIGITISATION AND DIGITALISATION

Knowledge Acquisition
Pi-Rads

ONLINE FEATURES

$$\vec{\mu} = \frac{1}{n}\sum_{i=0}^{n-1}\vec{s}_i$$

where n is the number of samples used for the classification, the mean radius μ_r (standard deviation) as

$$\mu_r = \frac{1}{n}\sum_{i=0}^{n-1}\| \vec{s}_i - \vec{\mu} \|,$$

and the angle φ_{si} as

$$\varphi_{s_i} = \cos^{-1}\left(\frac{(s_i - s_{i-1})\cdot(s_{i+1} - s_i)}{\| s_i - s_{i-1} \| \| s_{i+1} - s_i \|}\right).$$

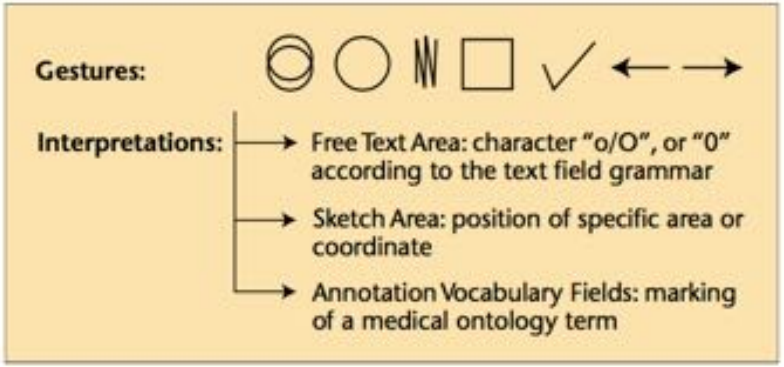
stamp. A stroke is a sequence S of samples,

$$S = \{\vec{s}_i \mid i \in [0, n-1], t_i < t_{i+1}\}$$

where n is the number of recorded samples. A sequence of strokes is indicated by

$$D = \{S_i \mid i \in [0, m-1]\},$$

where m is the number of strokes. The area A covered by the sequence of strokes D is defined as the area of the bounding box that results from a sequence of strokes.



ID	Feature	Description	Note
0	Number of Strokes	N	
1	Length	$\lambda = \sum_{i=0}^{n-1} \| vecs_i - vecs_{i+1} \|$	s_i denotes a sample.
2	Area	A	
3	Perimeter Length	λ_c	Length of the path around the convex hull.
4	Compactness	$c = \frac{\lambda_c^2}{A}$	
5	Eccentricity	$e = \sqrt{1 - \frac{b^2}{a^2}}$	a and b denote the length of the major or minor axis of the convex hull, respectively.
6	Principal Axes	$e_r = \frac{b}{a}$	
7	Circular Variance	$v_r = \frac{1}{n\mu_r^2}\sum_{i=0}^{n}(\| \vec{s}_i - \vec{\mu} \| - \mu_r)^2$	μ_r denotes the mean distance of the samples to the centroid μ_c.
8	Rectangularity	$r = \frac{A}{ab}$	
9	Closure	$c_l = \frac{\| \vec{s}_0 - \vec{s}_n \|}{\lambda}$	
10	Curvature	$\kappa_s = \sum_{i=1}^{n-1}\varphi_{s_i}$	φ_{s_i} denotes the angle between the $\overline{s_{i-1}s_i}$ segments and $\overline{s_i s_{i+1}}$ at s_i.
11	Perpendicularity	$p_r = \sum_{i=1}^{n-1}\sin(\varphi_{s_i})^2$	
12	Signed Perpendicularity	$p_{sr} = \sum_{i=1}^{n-1}\sin(\varphi_{s_i})^3$	
13	Angles after Equidistant Resampling (6 line segments)	$\sin(\alpha), \cos(\alpha)$	The five angles between succeeding lines are considered to make the features scale and rotation invariant (normalization of writing speed).

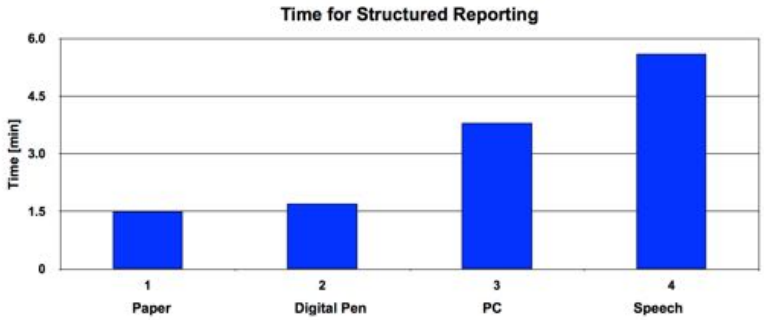
EXAMPLES OF BLENDED INTERACTION RANGE FROM DIGITAL PEN & PAPER AND PEN & MULTI-TOUCH INTERACTION TO TANGIBLE DISPLAYS IN ROOMS OF MIXED OR AUGMENTED REALITY.

Combines the virtues of physical and digital artifacts.

Cognitive theories and post-WIMP designs and technologies.

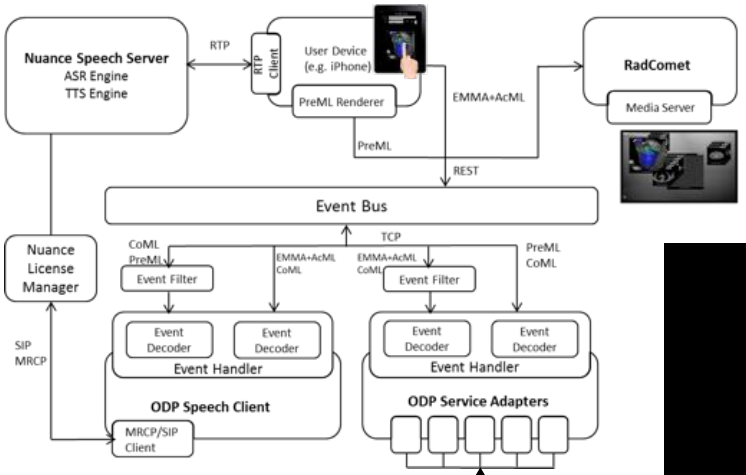
The quality of *Blended Interaction* is judged by its compatibility with our natural cognitive processes when we interact and collaborate in the real non-digital world.

System Features	Paper	Mammo Digital Paper	PC (iSoft)	ASR (Nuance)
Pen-on-paper interface	x	x		
Immediate validations		x	x	x
Offline validation (of digital content)		x	x	x
Realtime recognition (text)		x	x	x
Realtime recognition (gestures)		x		
Online correction of recognition errors		x	x	
Real-time capture to structured database		x	x	
Forward capture to database		x	x	x
Source Document (Certificate)	x	x	(x)	
Digital Source Document (Certificate)		x		
Training hours before effective usage	10	10	30	35
No user distraction from primary task	x	x		
No distraction from patient	x	x		(x)
Average time to complete one predefined Radlex entry	3 sec	3 sec	5 sec	2 sec

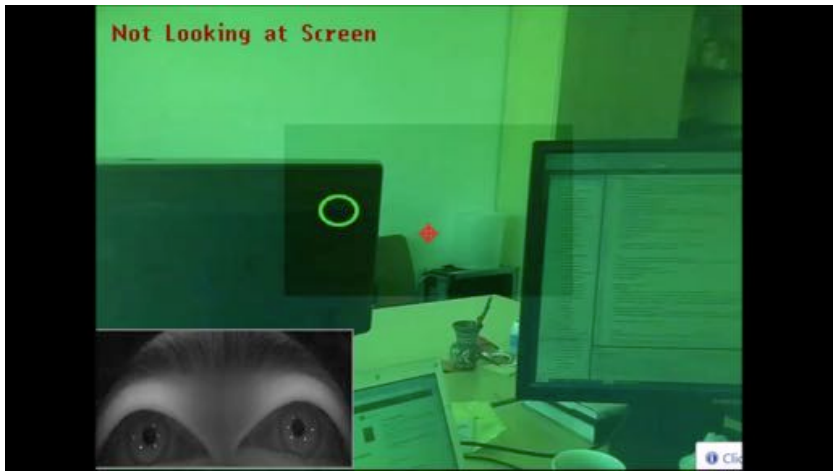


Digital manufacturing

COMPLEX MULTIMODAL MULTISENSOR SYSTEMS



ERmed
Erweiterte Realität in der Medizin



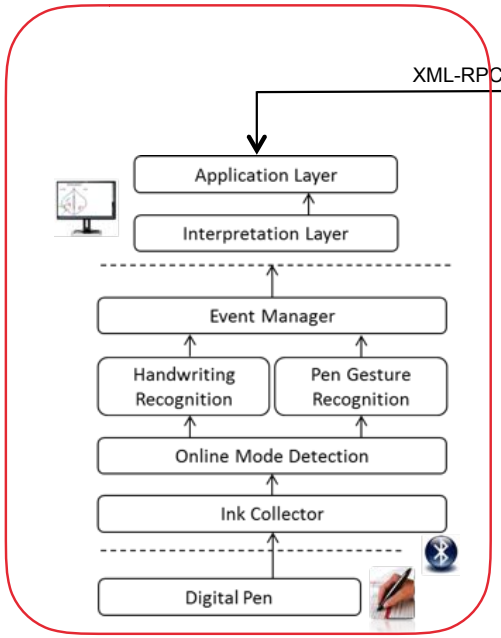
XML-RPC

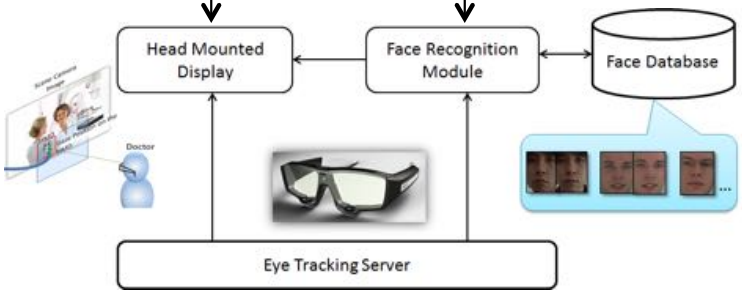
XML-RPC

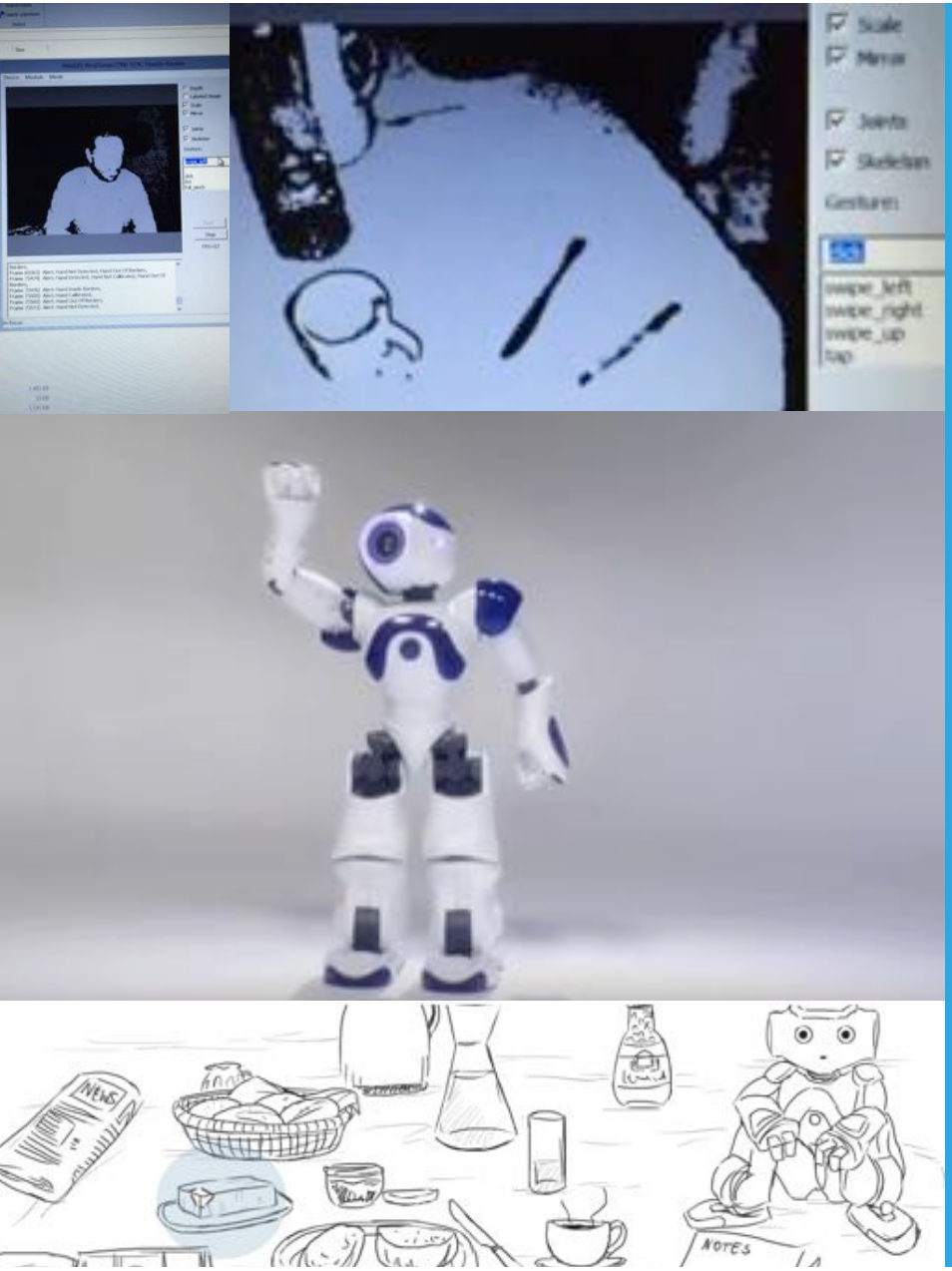
ERmed Proxy

Ermed Bridge

TCP

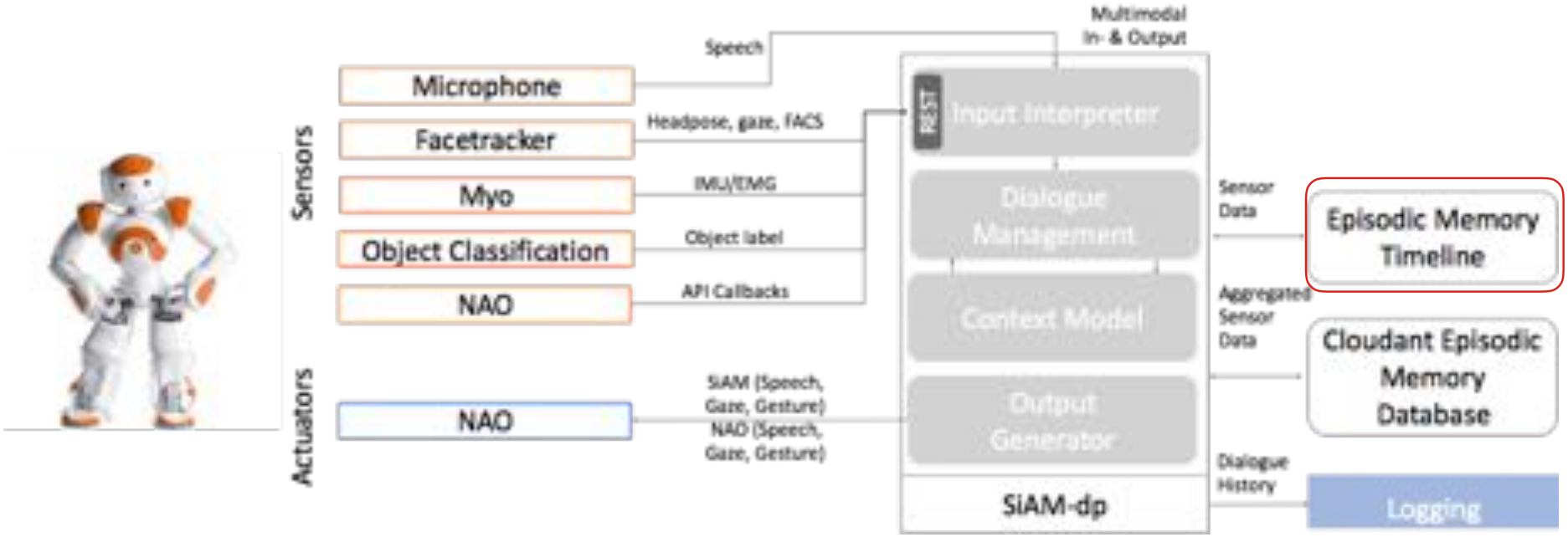




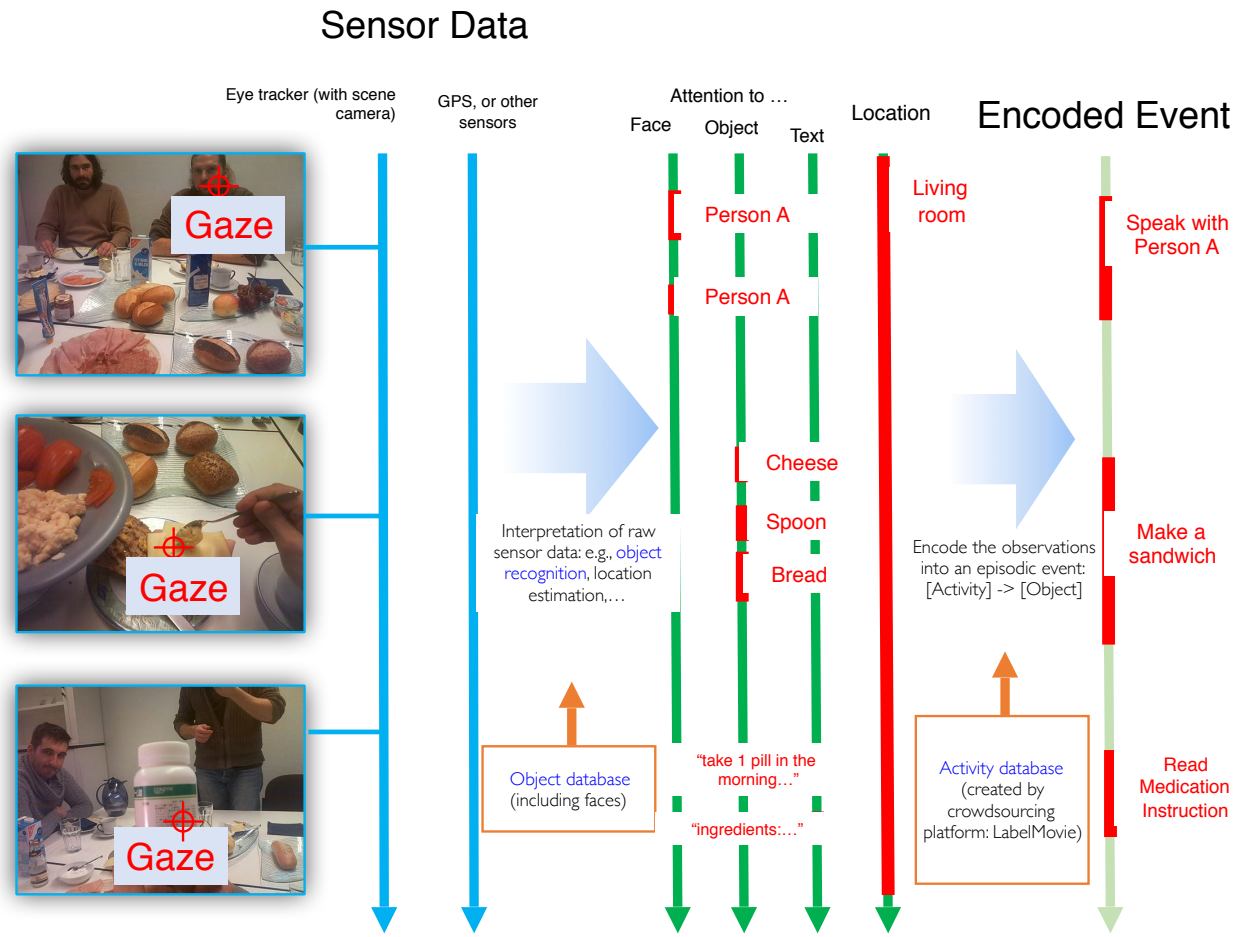


ACTIVITY RECOGNITION

Companion technology and social interaction

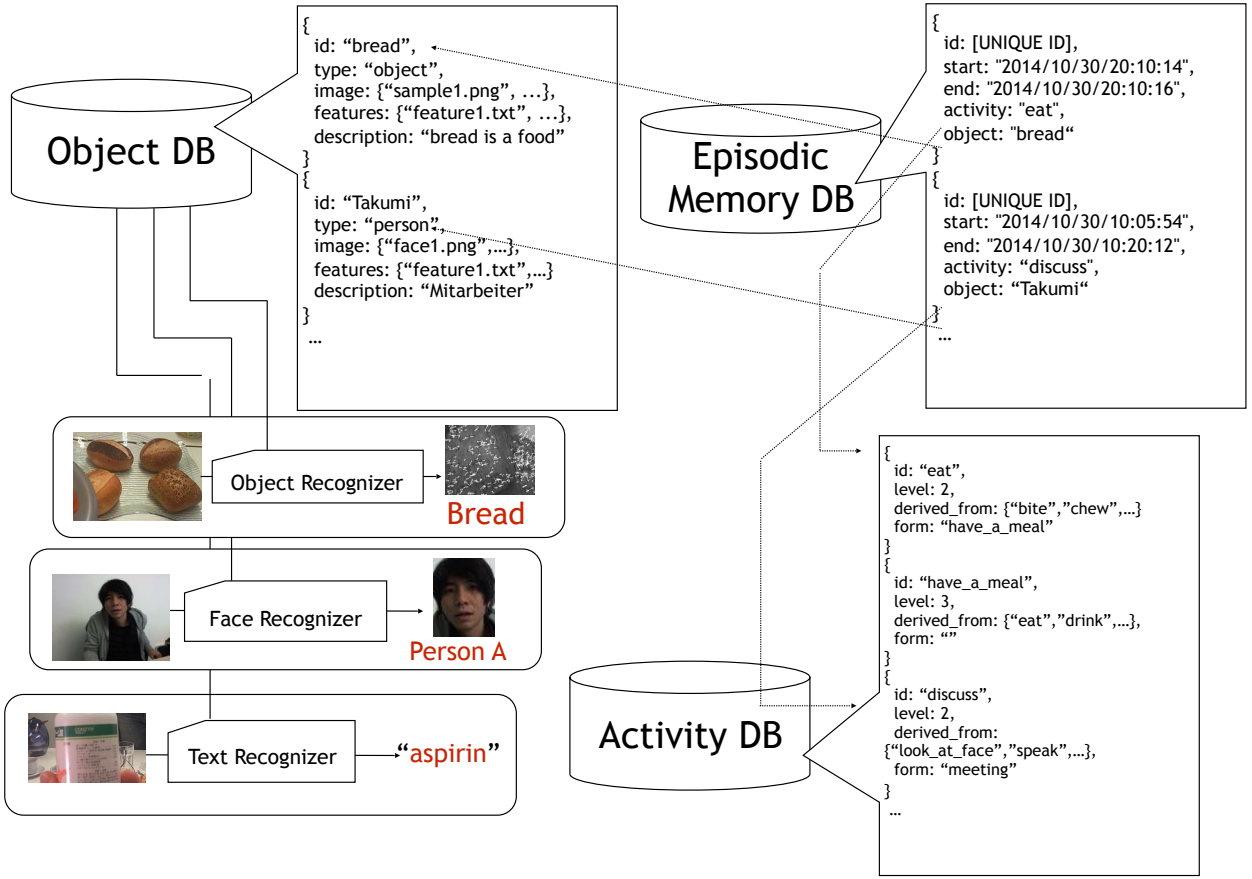

Sensors
Actuators
Microphone
Facetracker
Myo
Object Classification
NAO
NAO
Speech
Headpose, gaze, FACS
IMU/EMG
Object label
API Callbacks
SiAM (Speech, Gaze, Gesture)
NAO (Speech, Gaze, Gesture)
Multimodal In- & Output
REST
Input Interpreter
Dialogue Management
Context Model
Output Generator
SiAM-dp
Sensor Data
Aggregated Sensor Data
Dialogue History
Episodic Memory Timeline
Cloudant Episodic Memory Database
Logging

Episodic memory event encoding model (Breakfast Scenario)



Databases and recognition modules

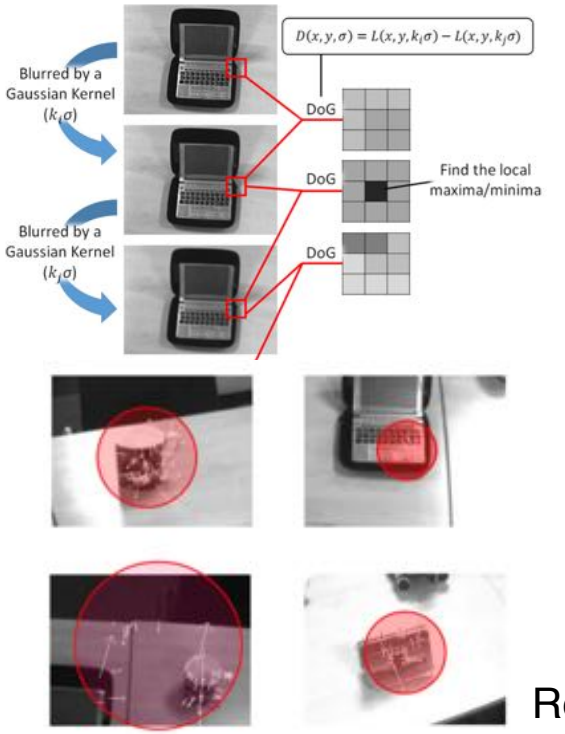
Kognit Cloudant Database: https://kognit-tt.cloudant.com/



- By combining image analysis technologies with eye gaze analysis technologies, a computer can recognise the visual content the user is attending to.

- Analysis of user eye gaze is useful to present information of attended content in an adequate way.

- Gaze Videos (x2)



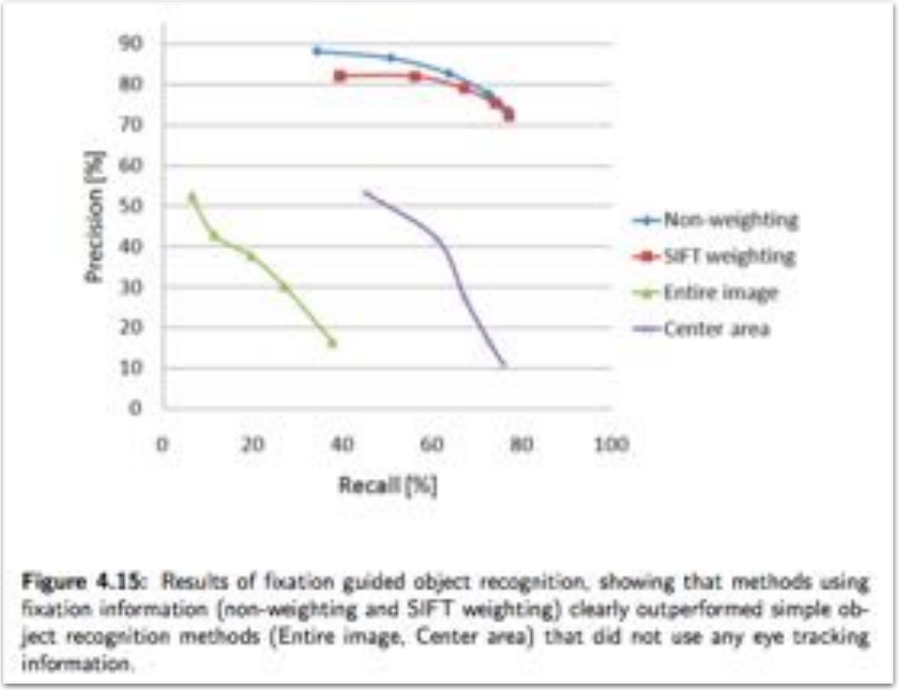
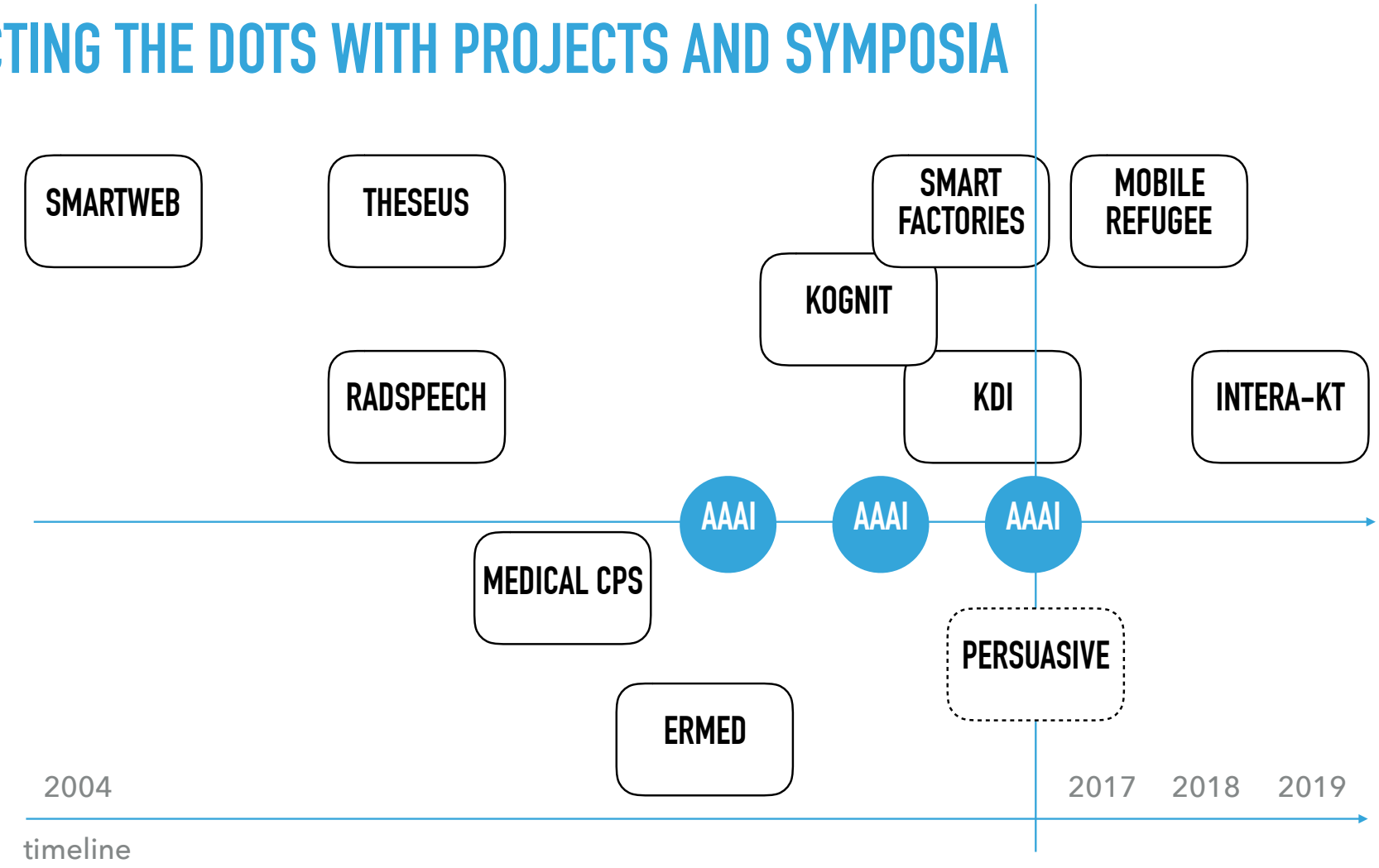
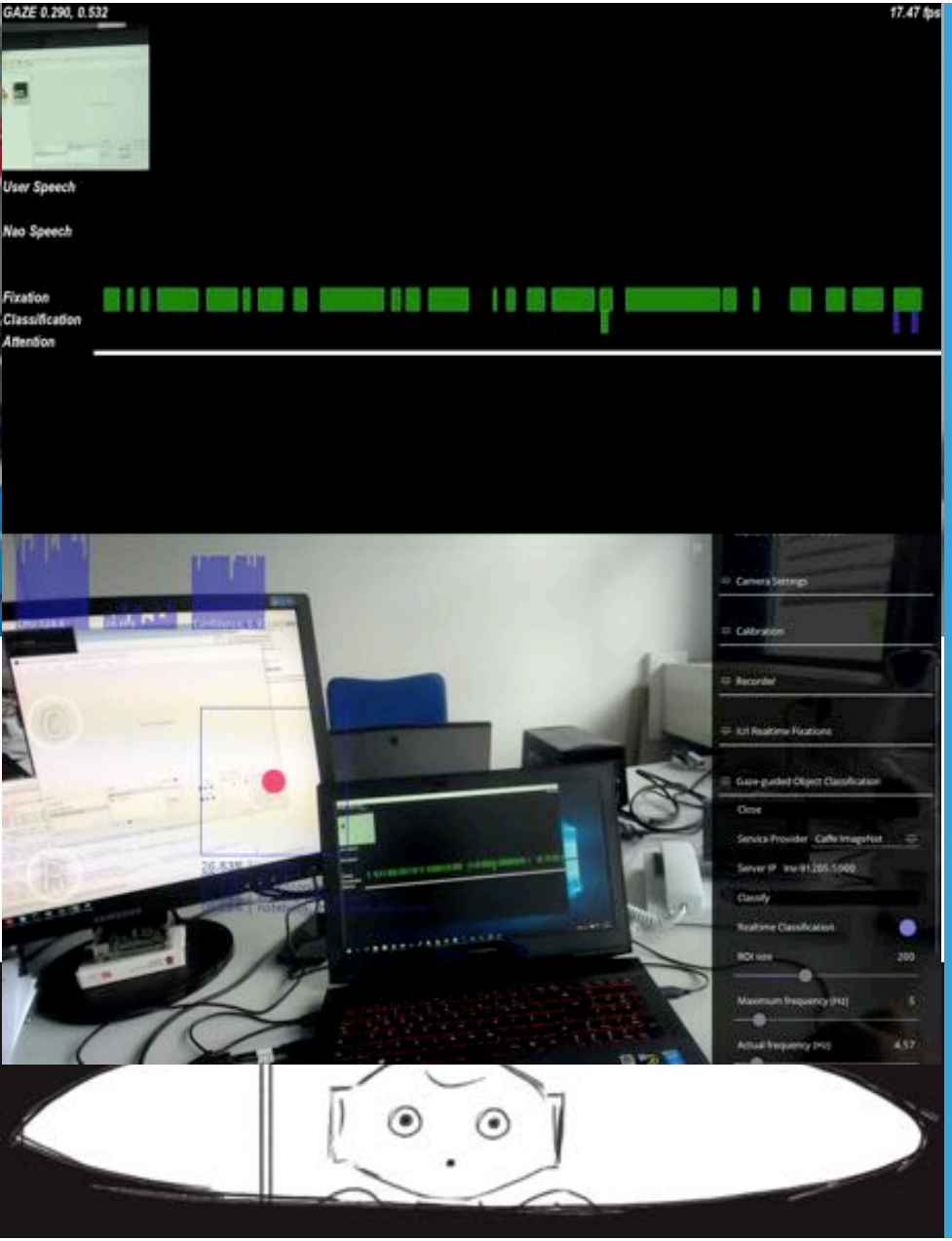


Figure 4.15: Results of fixation guided object recognition, showing that methods using fixation information (non-weighting and SIFT weighting) clearly outperformed simple object recognition methods (Entire image, Center area) that did not use any eye tracking information.

Real-Time Gaze-based Object Recognition

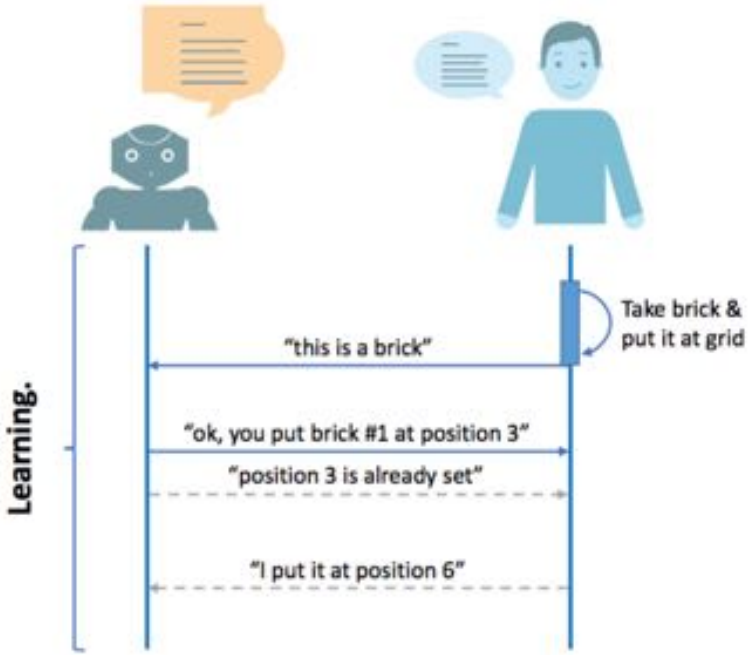
CONNECTING THE DOTS WITH PROJECTS AND SYMPOSIA

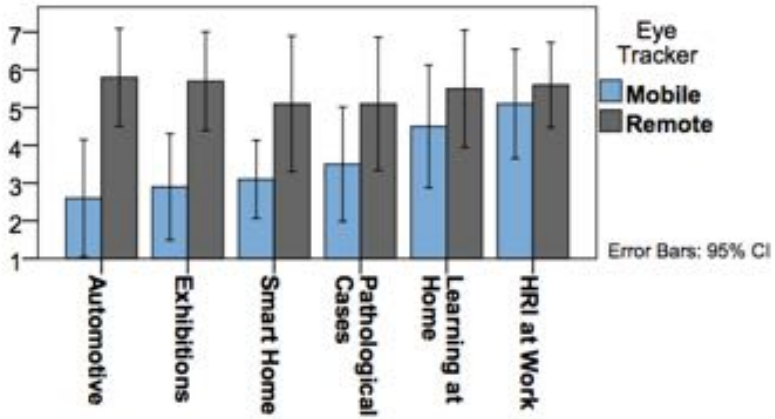


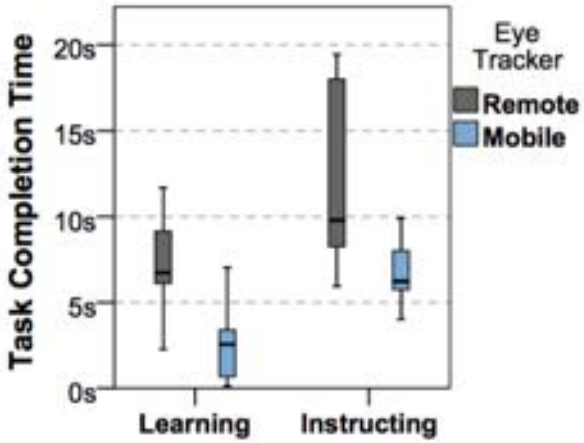


DEEP LEARNING BASED EPISODIC MEMORY

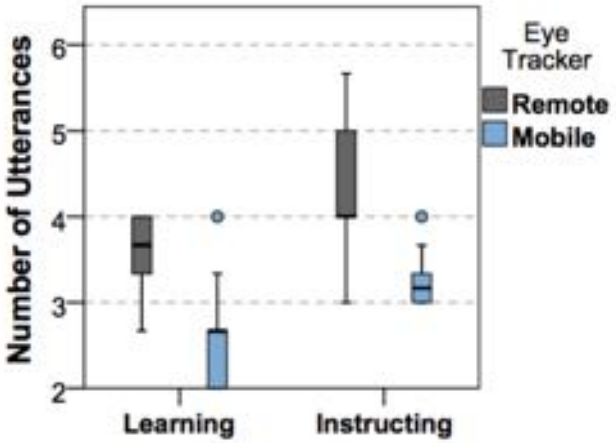
Object Recognition
20 -> 2000 objects





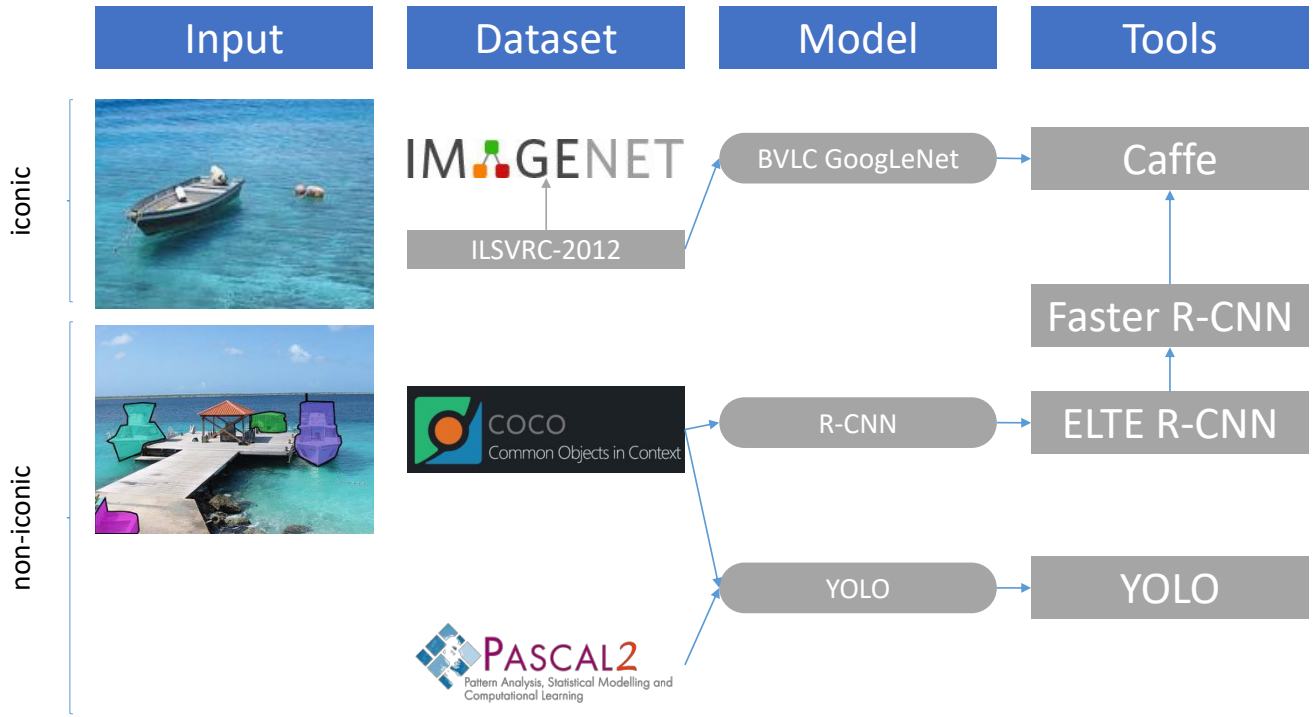


(a) Task completion time

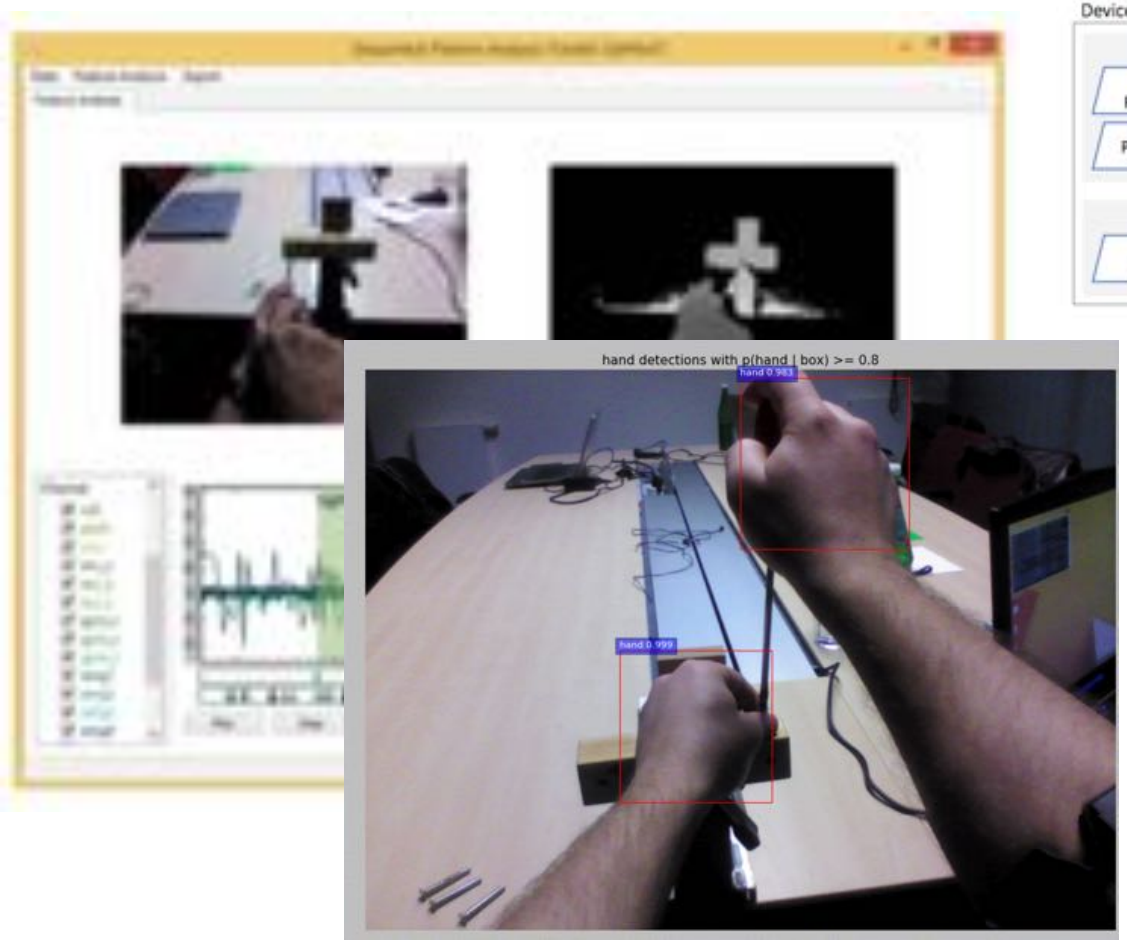


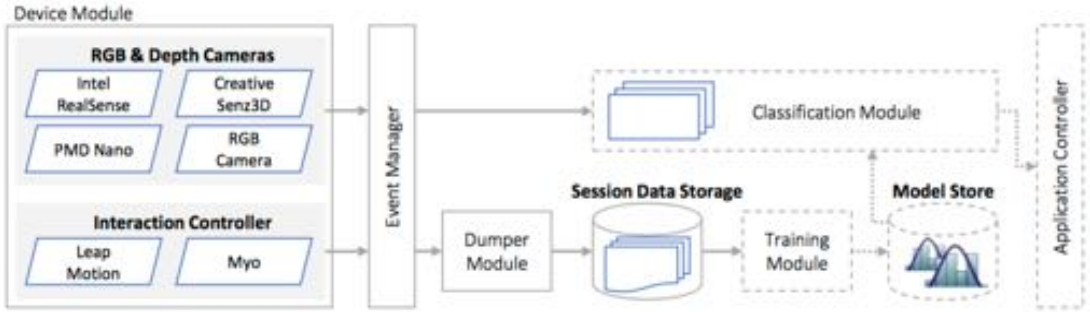
(b) Number of utterances

DEEP LEARNING FOR OBJECT DETECTION



MULTIMODAL MULTISENSOR ACTIVITY ANNOTATION TOOL





Cameras

Device	*Output*	*Platform*
Webcam (UVC)	RGB	Windows, Linux
Intel RealSense F200	RGB, depth, point cloud	Windows
Creative Senz3D	RGB, depth	Windows, Linux
PMD Nano	RGB, depth, amplitude	Windows, Linux

Interaction Devices

Device	*Output*	*Platform*
Leap Motion	hand tracking data	Windows, Linux
Myo	IMU, EMG data	Windows

WHY MULTIMODAL-MULTISENSOR INTERFACES HAVE BECOME DOMINANT

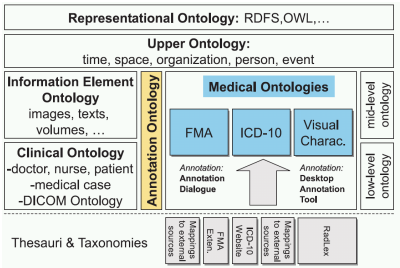
- Flexibility
- They support users' ability to select a suitable input mode, or to shift among modalities as needed, during the changing physical contexts and demands of continuous mobile use.
- Likewise there are ideal for supporting individual differences among users.
- This has stimulated the paradigm shift toward multimodal-multisensor interfaces on computers today, which often is further enhanced by either multimodal output or multimedia output.

2 PART OF ABSTRACT

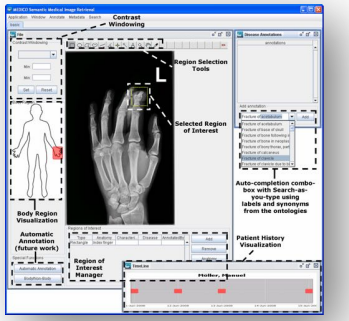
- Do not think about the best method of machine learning—the central issue is the application domain and domain problems such as integrated decision support for doctors.
- More of IUI's effort should go into real-life problems, approaches, and architectures.
- Our cognitive computing approach is an incremental knowledge acquisition process rather than almost exclusively using a data-driven engineering model of research.

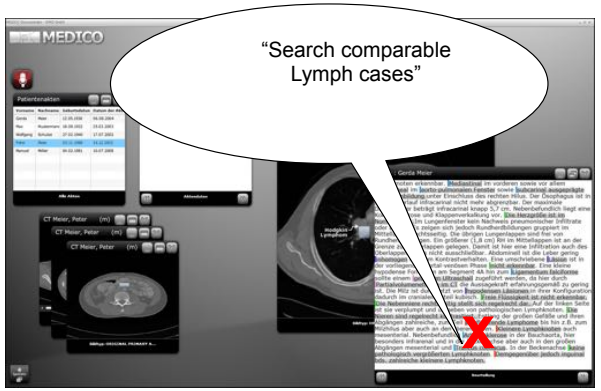
Incremental Knowledge Acquisition

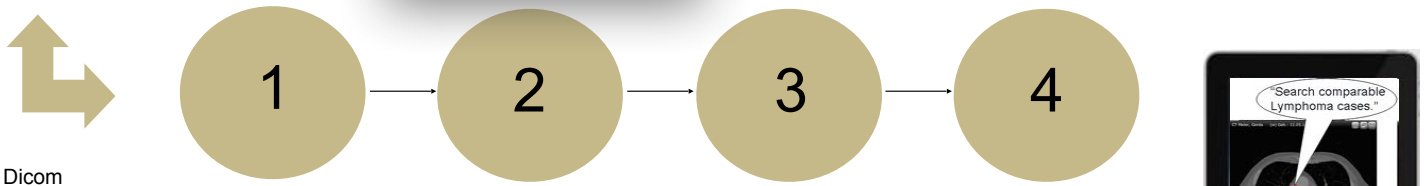
Structured/Structural Knowledge



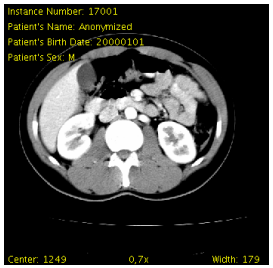
Desktop Annotation



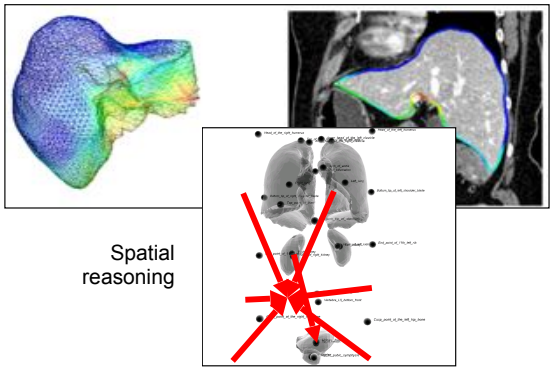




Dicom

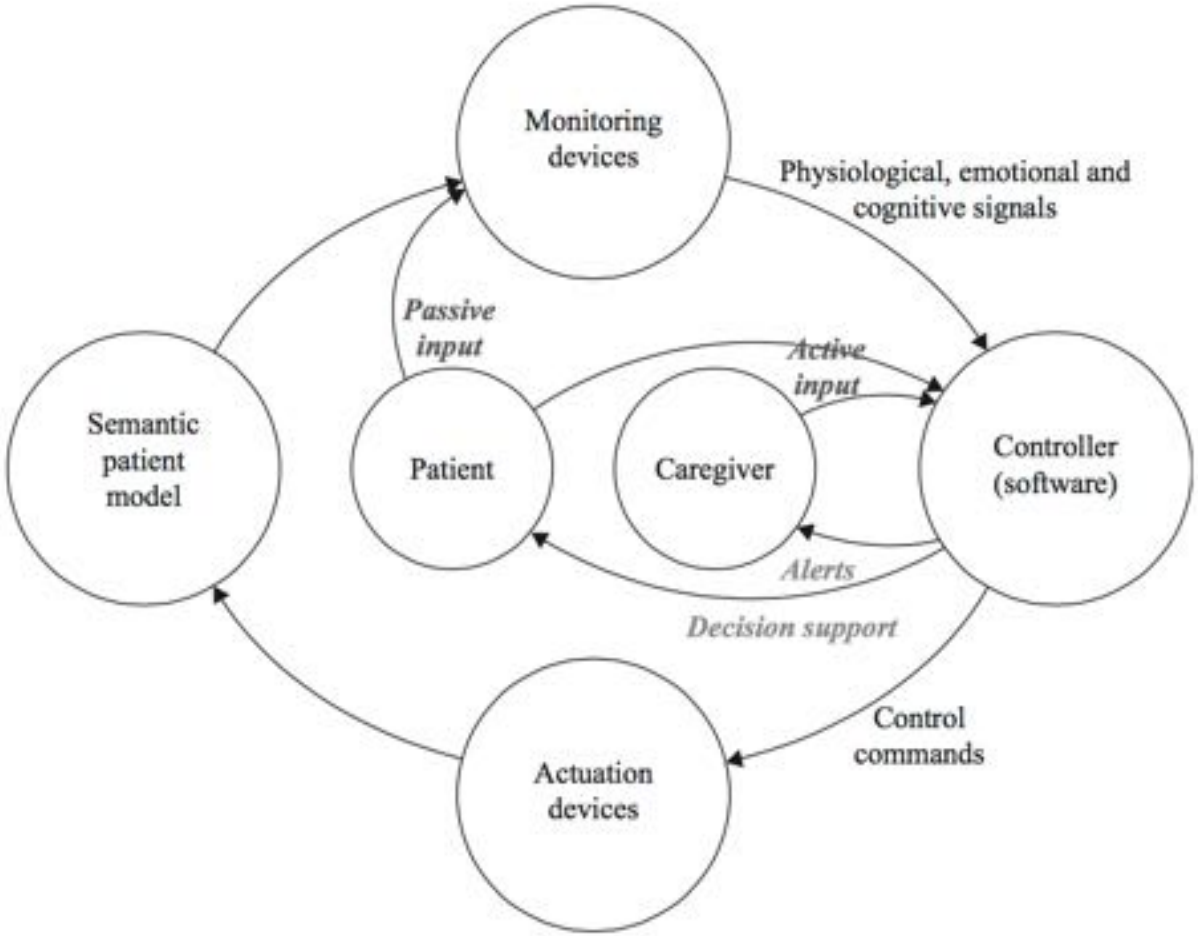


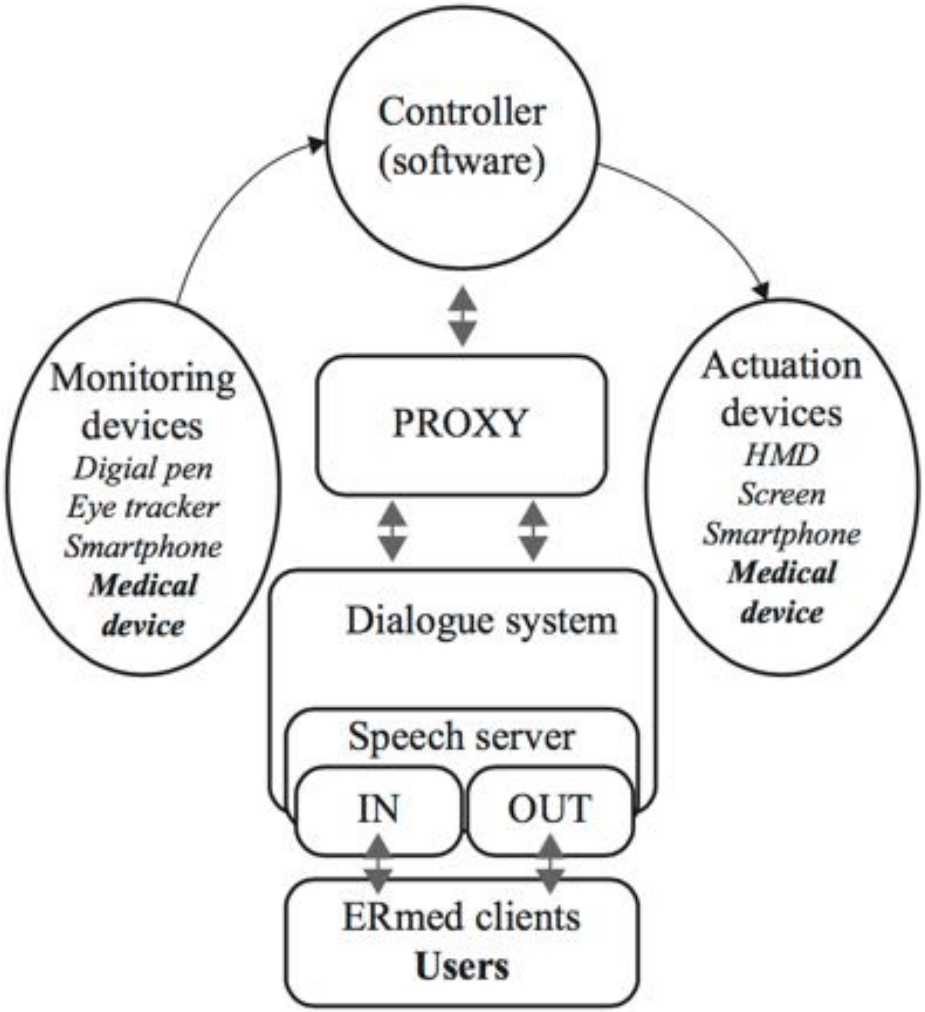
Automatic Image recognition



Spatial reasoning

SENSOR-ACTION-LOOP IN CPS (IN KOGNIT)

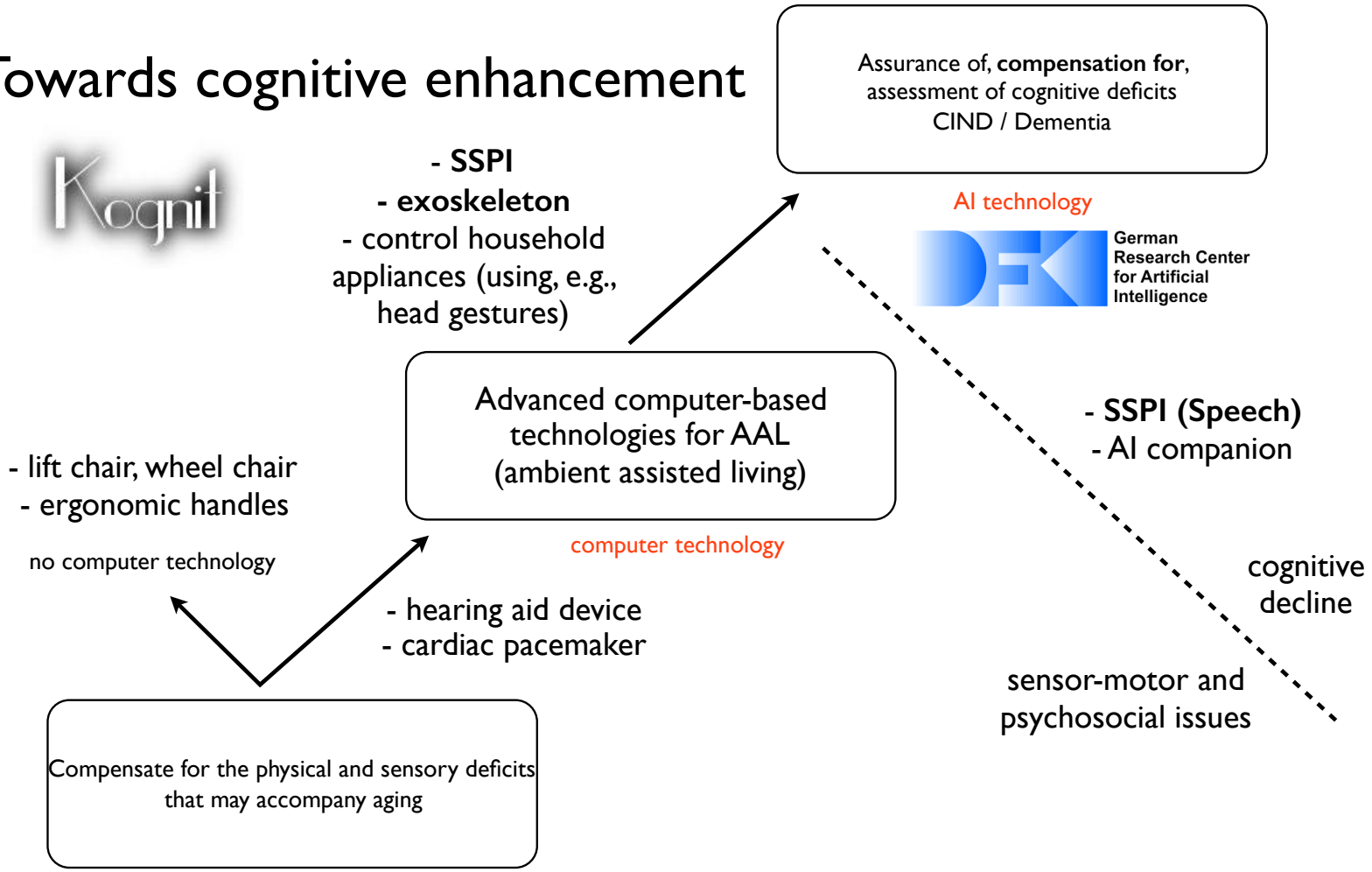




COGNITIVE ENHANCEMENT

improving cognitive abilities

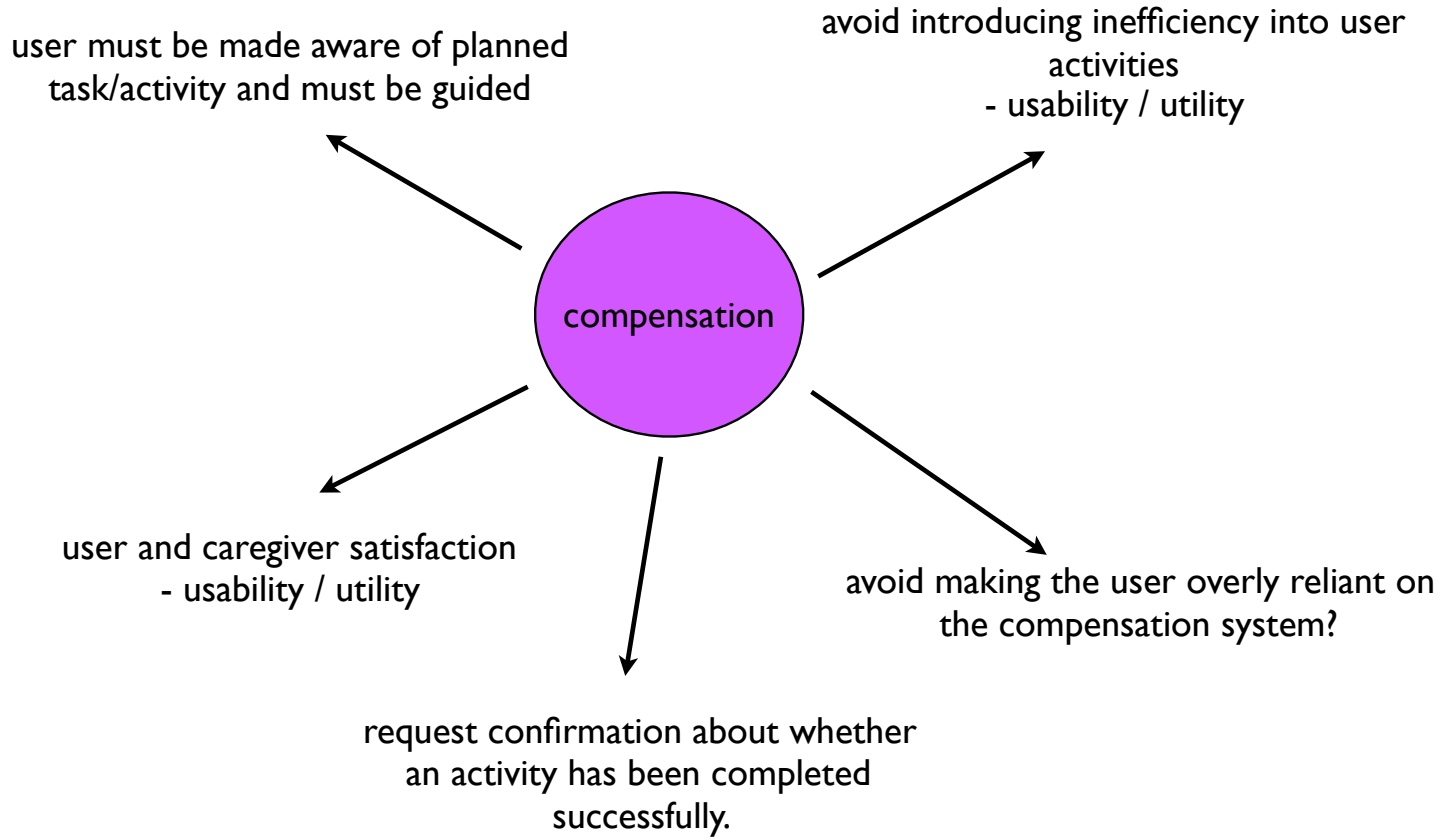
Towards cognitive enhancement



https://dl.dropboxusercontent.com/u/48051165/ISMAR-2015-IUI-TUTORIAL.pdf

https://dl.dropboxusercontent.com/u/48051165/kognit.pdf

Compensation Paradox

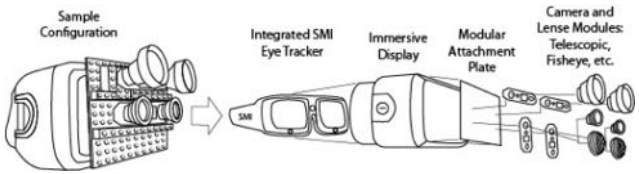


Kognit Hardware Overview

Narrative Clip

Pupil Labs
Eye-tracker

SMI Eye-tracker



Samsung Galaxy Note
4

Space Glasses Meta Pro
(3D cam)

Cybershot scene cam

Anoto

Oculus DK 2

WheelPhone

NAO Humanoid

Leap Motion

Accu LED projector

Epson Moverio BT-200

Low range 3D cam

Tobii EyeX

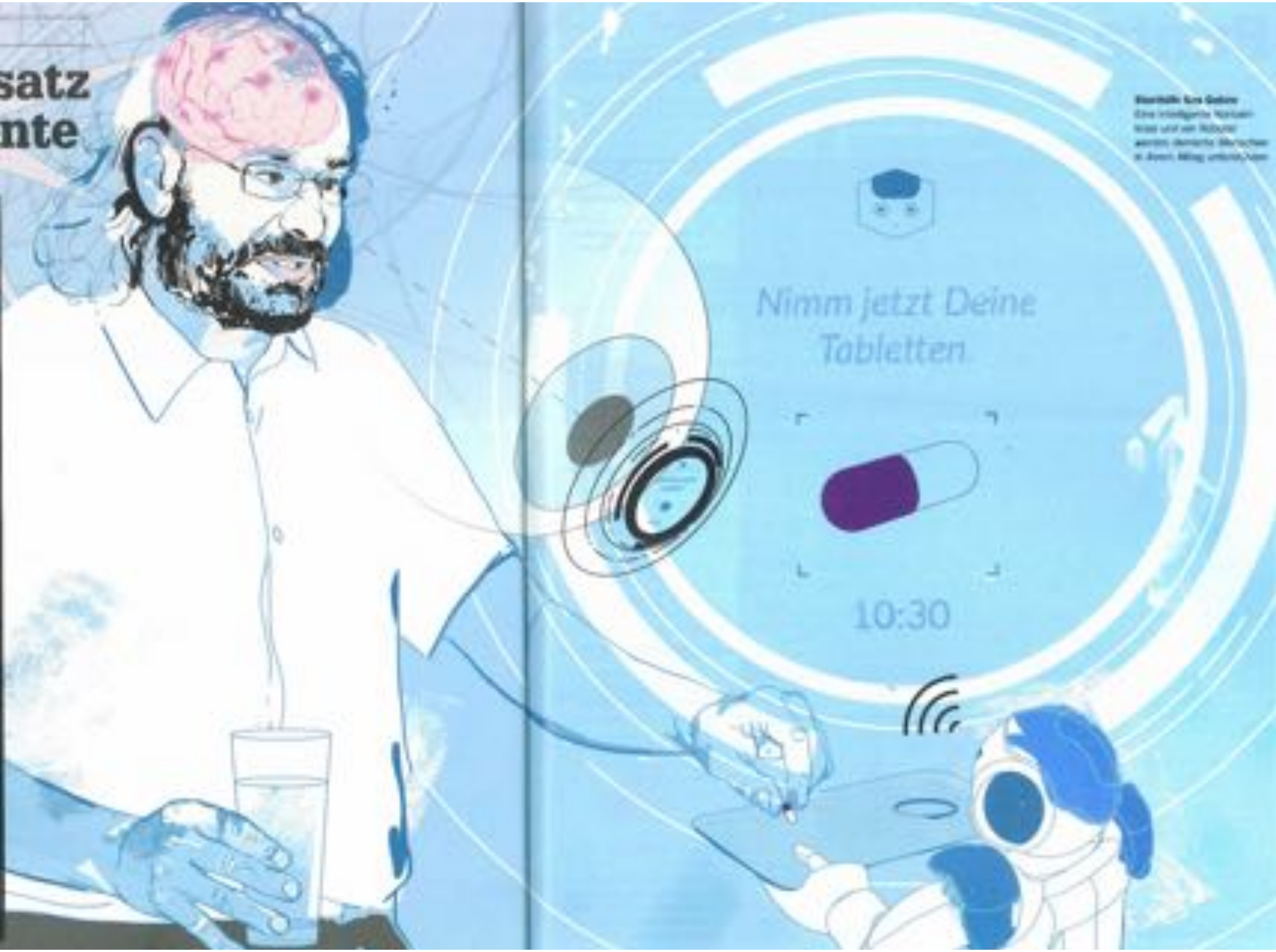
Structure Sensor

Brother Airscouter

Medizin der Zukunft III

Gehirnersatz für Demente

Daniel Sonntag Experte für künstliche Intelligenz



CONCLUSION, LESSONS LEARNED AND FUTURE INVESTIGATIONS

ML IS NOT A SILVER BULLET

- The interest has also being fueled by the recent research breakthroughs brought about by deep learning.
- Currently ML has several limitations in complex real-life situations. Some of these limitations include:
  - many ML algorithms require large number of training data that are often too expensive to obtain in real-life,
  - significant effort is often required to do feature engineering to achieve high performance,
  - many ML methods are limited in their ability to exploit background knowledge,
  - lack of a seamless way to integrate and use heterogeneous data.
  - do not provide end-to-end intelligent user interface systems.

DEEP LEARNING

- Success stories: supervised learning, backpropagation, stochastic gradient decent, massive amounts of data. Convolutional NN, RNN.
- Deep Learning in practice can be a game changer (95%->99%).
  - Parallel is good, layering is good.
  - Important Tool box
- **But: technology can be replicated, data cannot.**
  - **data acquisition (digitisation and digitalisation) is a strategic goal**
  - **what's your multimodal-multisensor data strategy over the long run?**

our data strategy

annotate images from, e.g., Google Images

annotate images from own application domain

Collect annotated images (segmentation / bounding boxes) from relevant application domains.

increased dataset

rotate

set background

set objects

occlude with additional objects

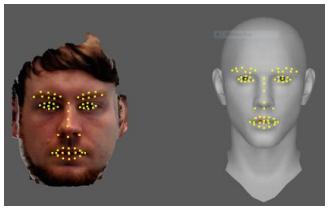
(if possible from application domain)

add surrogate semantics to Faster R-CNN and R-FCN

add models and semantics

head / hand model

conceptNet



interactive machine learning,

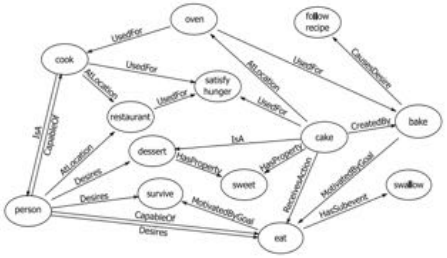
iteratively add training examples.

Tight interaction loop.

Computer is part of a human design process.

Interactively edit data.

How does a human-centred approach change the way machine learning is done?



FUTURE INVESTIGATIONS

- rule-based systems, ontologies and deep learning
- self-training in narrower contexts
- anomaly detection and single-shot-learning
- more sensors
- goal-orientation
- episodic memory in VR
- providing examples of cognitive computing