Statistical Dialogue Systems

Alexandros Papangelis

2021

About me

NCSR Demokritos & U.T. Arlington

MultiModal Adaptive Dialogue Systems


ArticuLab Carnegie Mellon
Socially-Skilled Virtual Agents


Toshiba Research Europe
Goal-Oriented Statistical Dialogue

Uber AI

Conversational AI


Amazon Alexa AI
Conversational AI

Today's Schedule

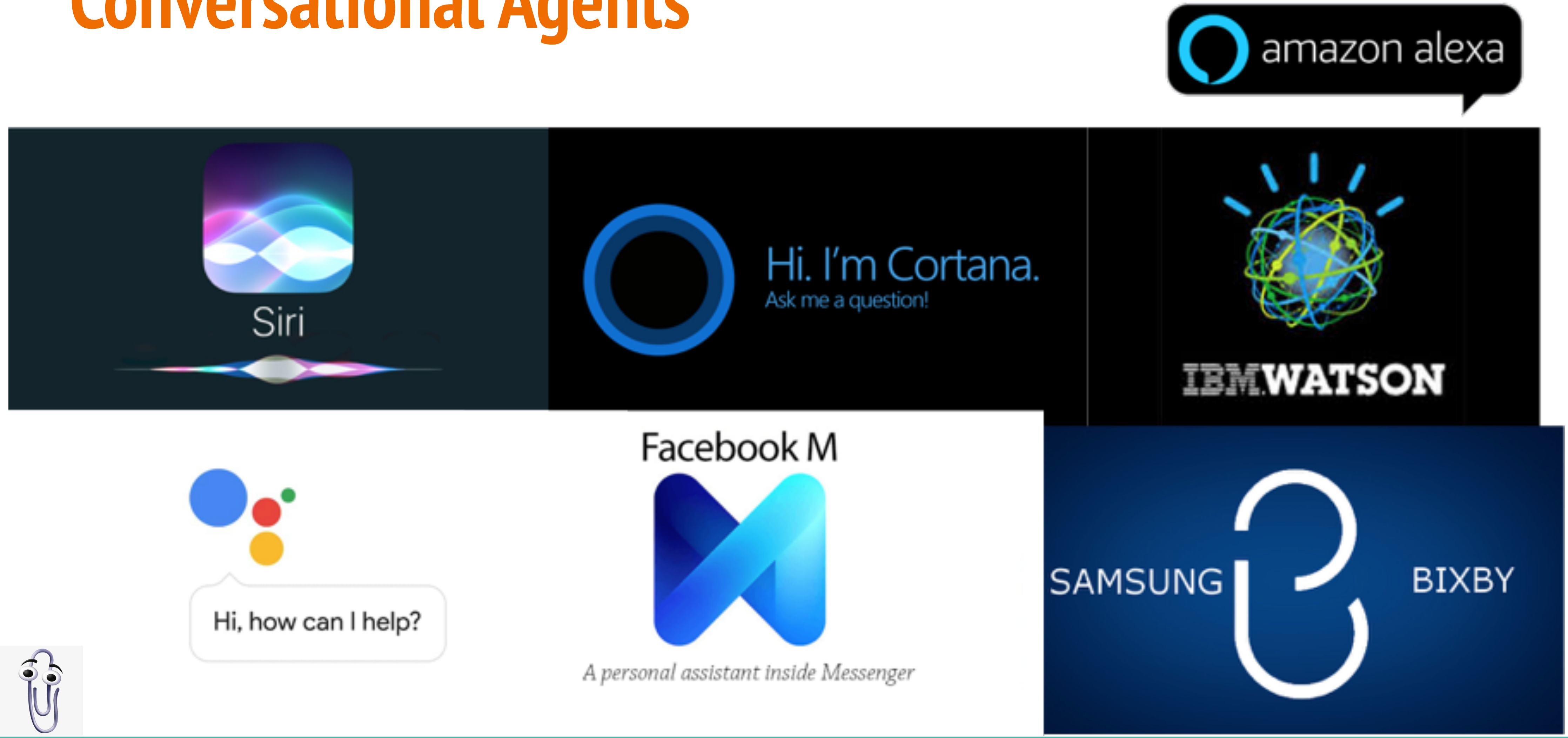
16:15 - **Intro to Conversational Agents**

17:00 - Break

17:15 - **Other aspects and challenges**

Friday: Deep Learning in dialogue

Conversational Agents



Conversational Agents

Different challenges for each application

THOMSON REUTERS | OPEN CALAIS

AI at Bloomberg

intel

NUANCE

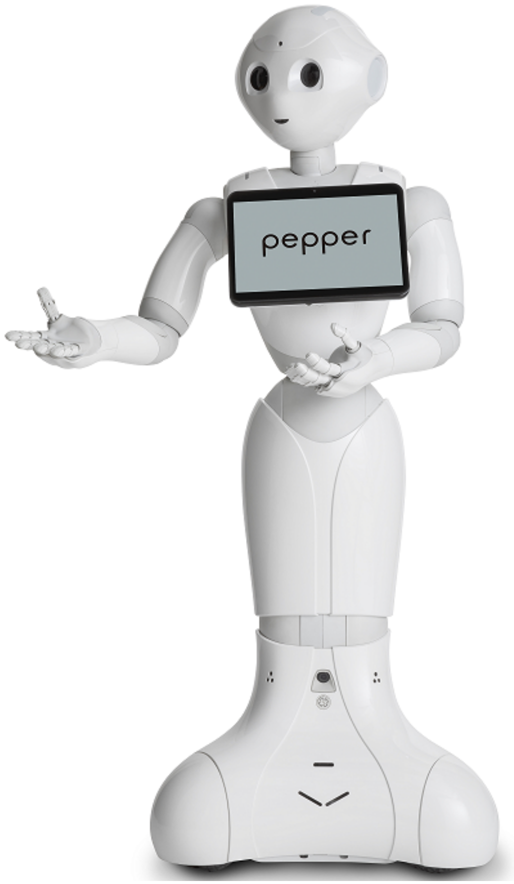
Uber

trio.ai

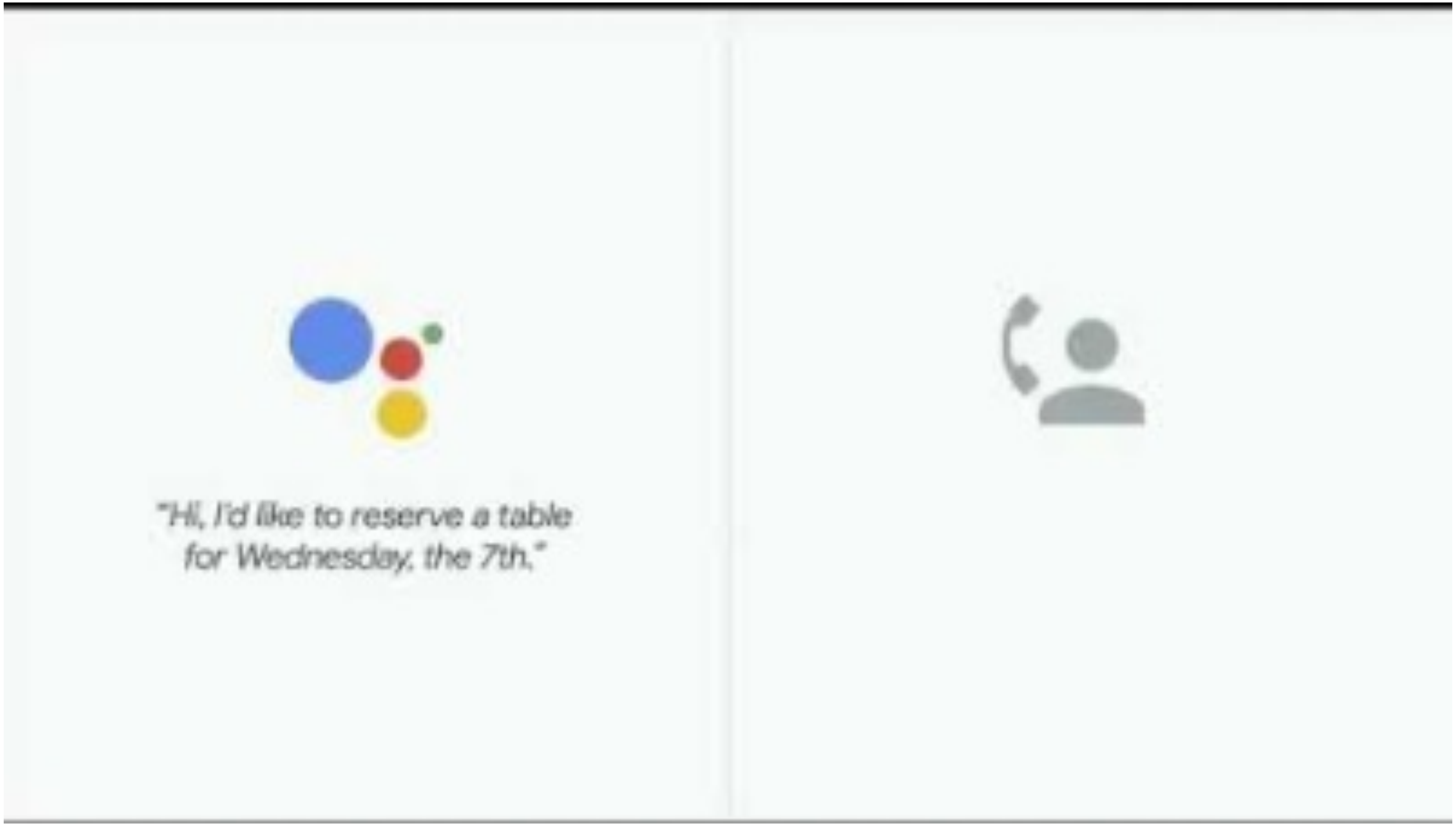
Adobe

Plus countless speech & language startups

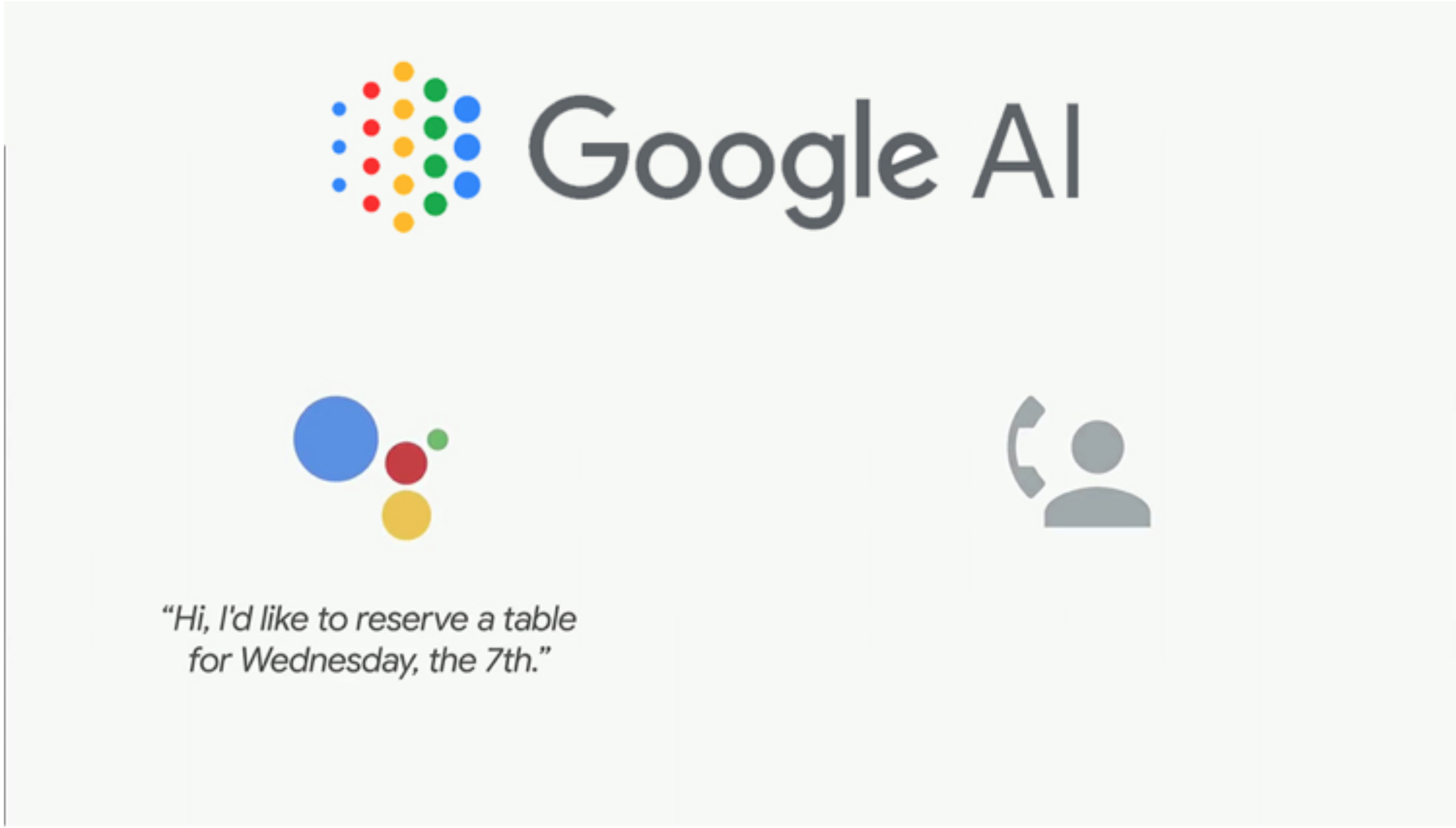
Embodied Conversational Agents



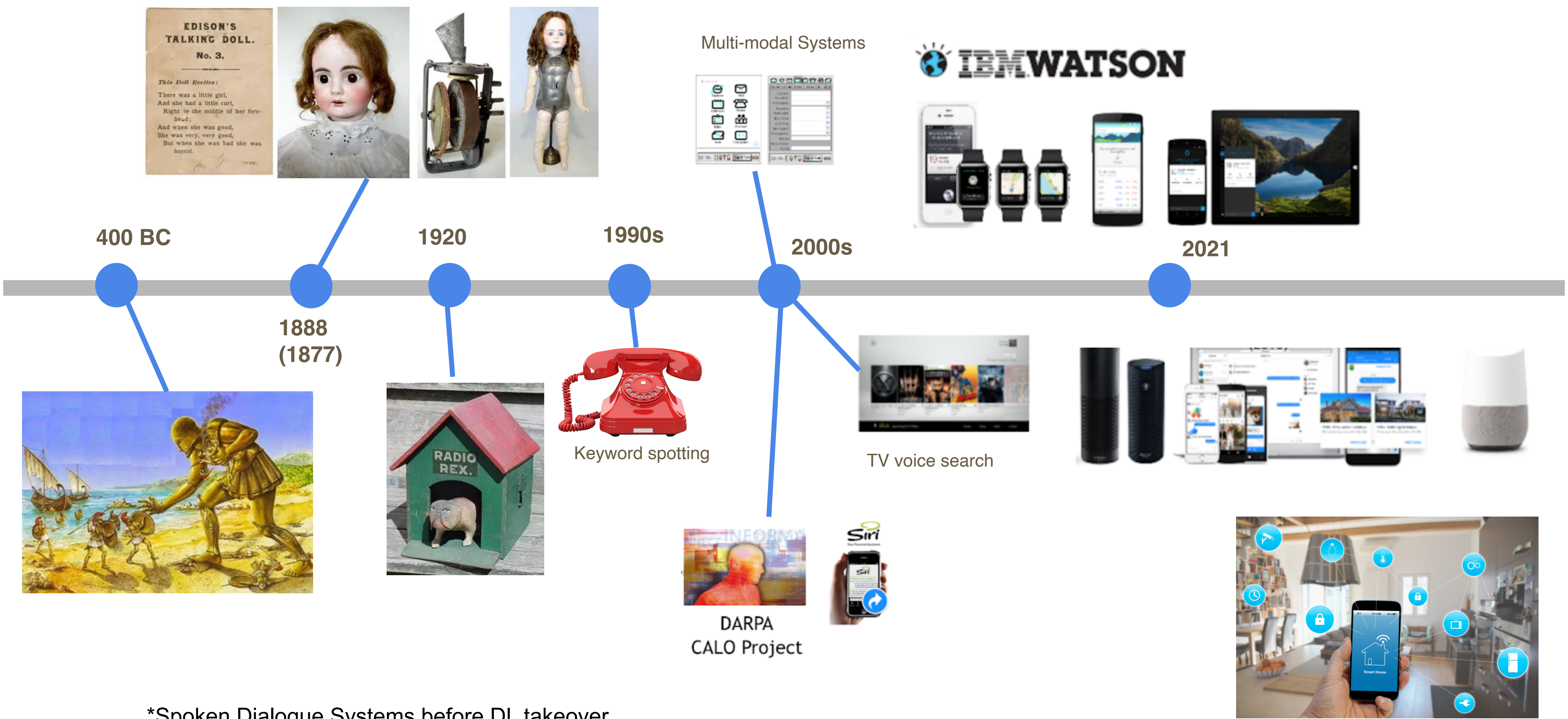
Google Duplex



Google Duplex Gone Wrong (fake)



Commercial Conversational Agents* Timeline



*Spoken Dialogue Systems before DL takeover

How is this different from other forms of communication?

- Input
    - Many **disfluencies** (uh, um, mh-m, hmm, …)
    - Word or phrase **repetitions** (false starts, re-starts, …)
    - Ill-defined (**omissions**, common knowledge, …)
    - **Noise** in channel (others talking, bad quality phone line, music, …)
    - One partial solution: **grounding**
        - Confirm user's intentions
- Turn-taking
    - Barge-ins, interruptions
    - Know when to take turn, system / user / mixed initiative
- Paralinguistic information *(Communication is not only text & speech)*
    - What we say, how we say it, why we say it, what effect does it have, where we look at

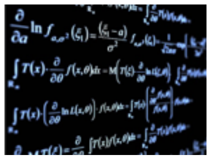
Small changes in language → big effect

More intelligent → higher expectations!

What does it take?

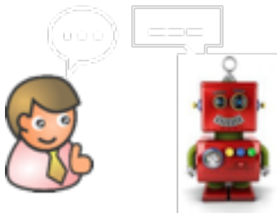

Dialogue Theory

Language Technology


Mathematics


Human Interaction

Social Science


Human-Machine Dialogue

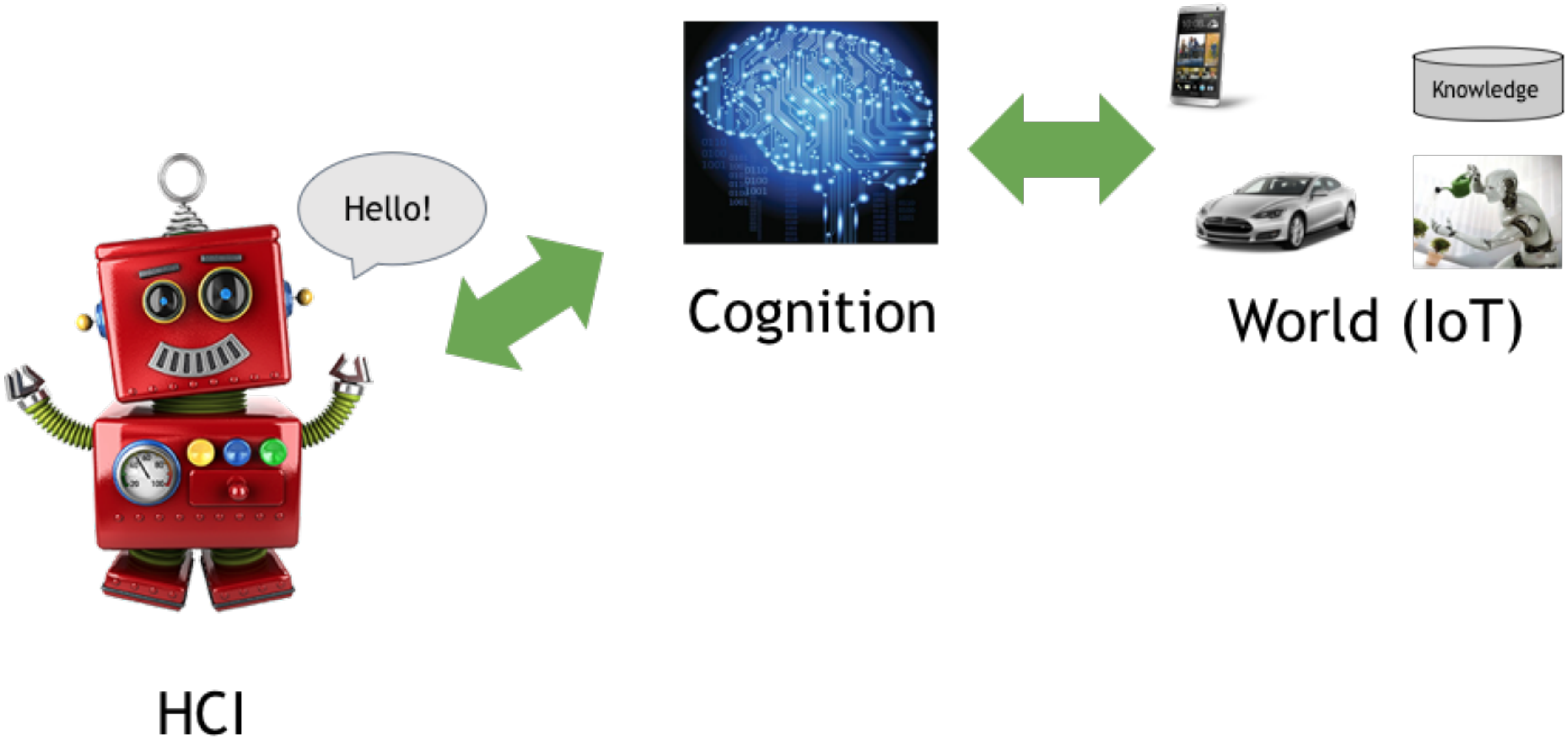

Assistive Technology


Education / Training

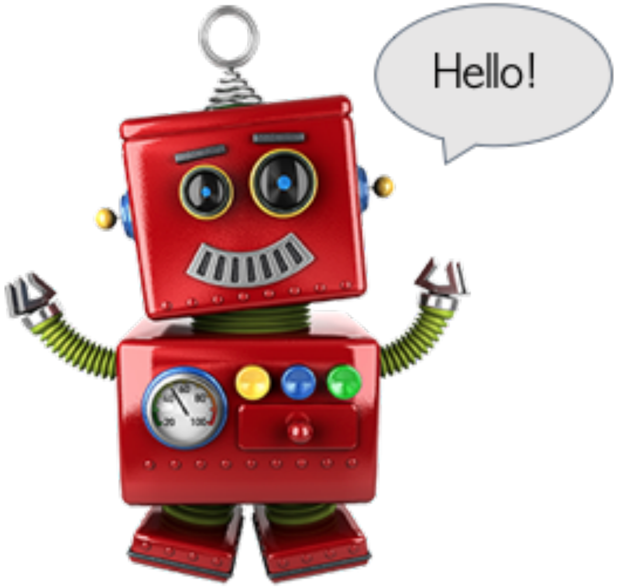

Robotics

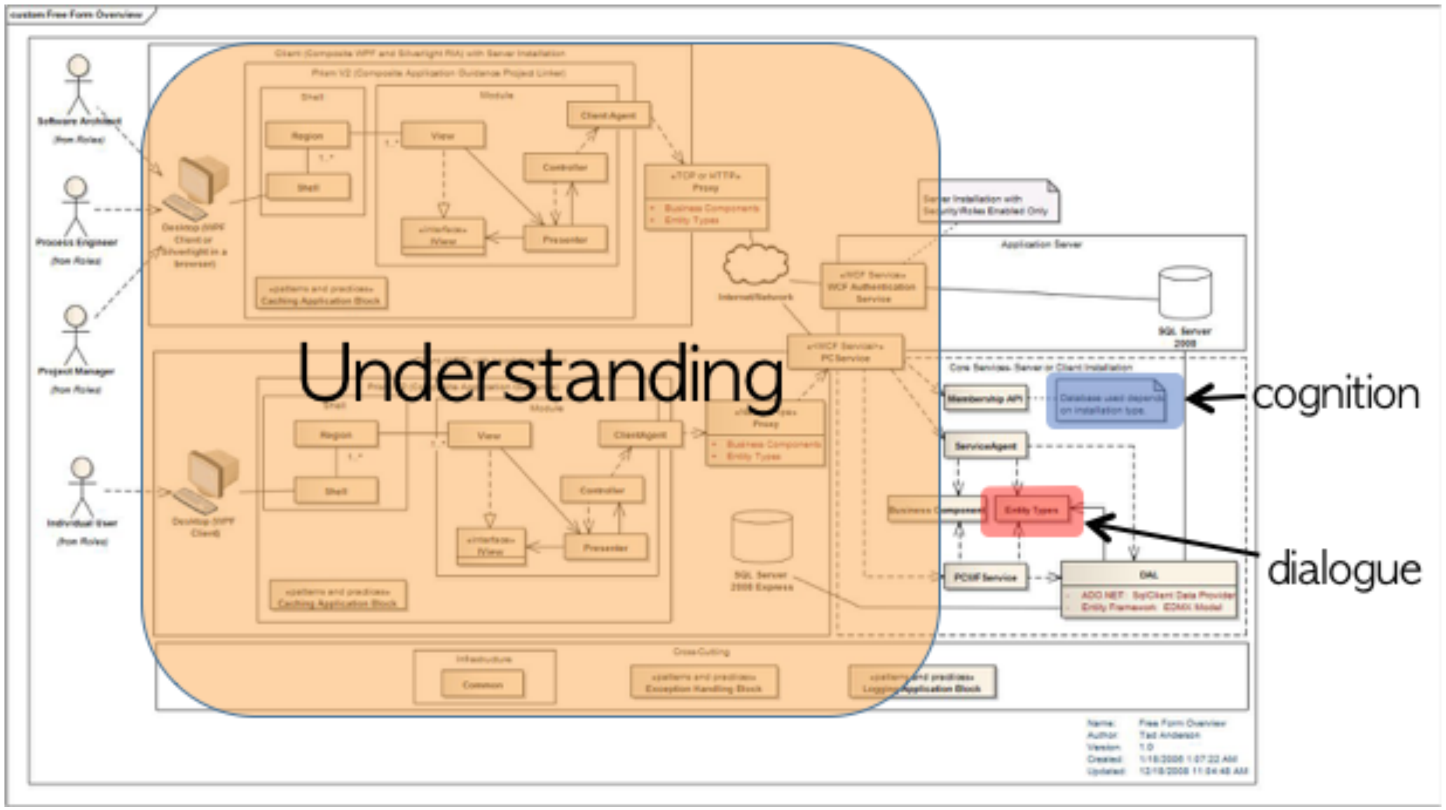

(Social) Intelligent Agent

Conversational Agents

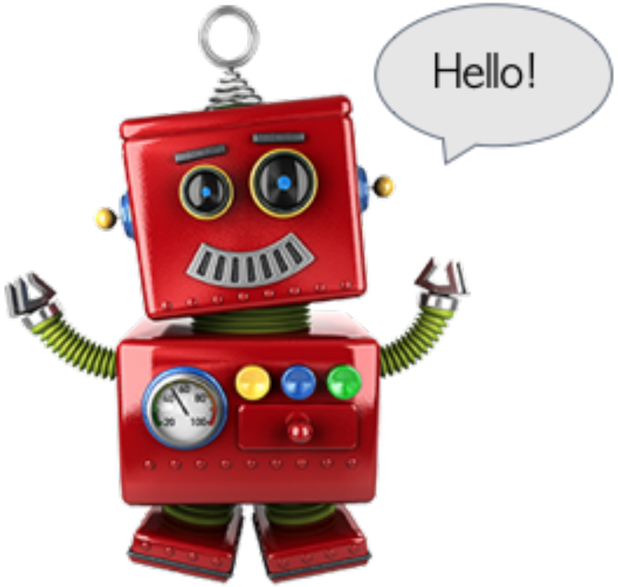


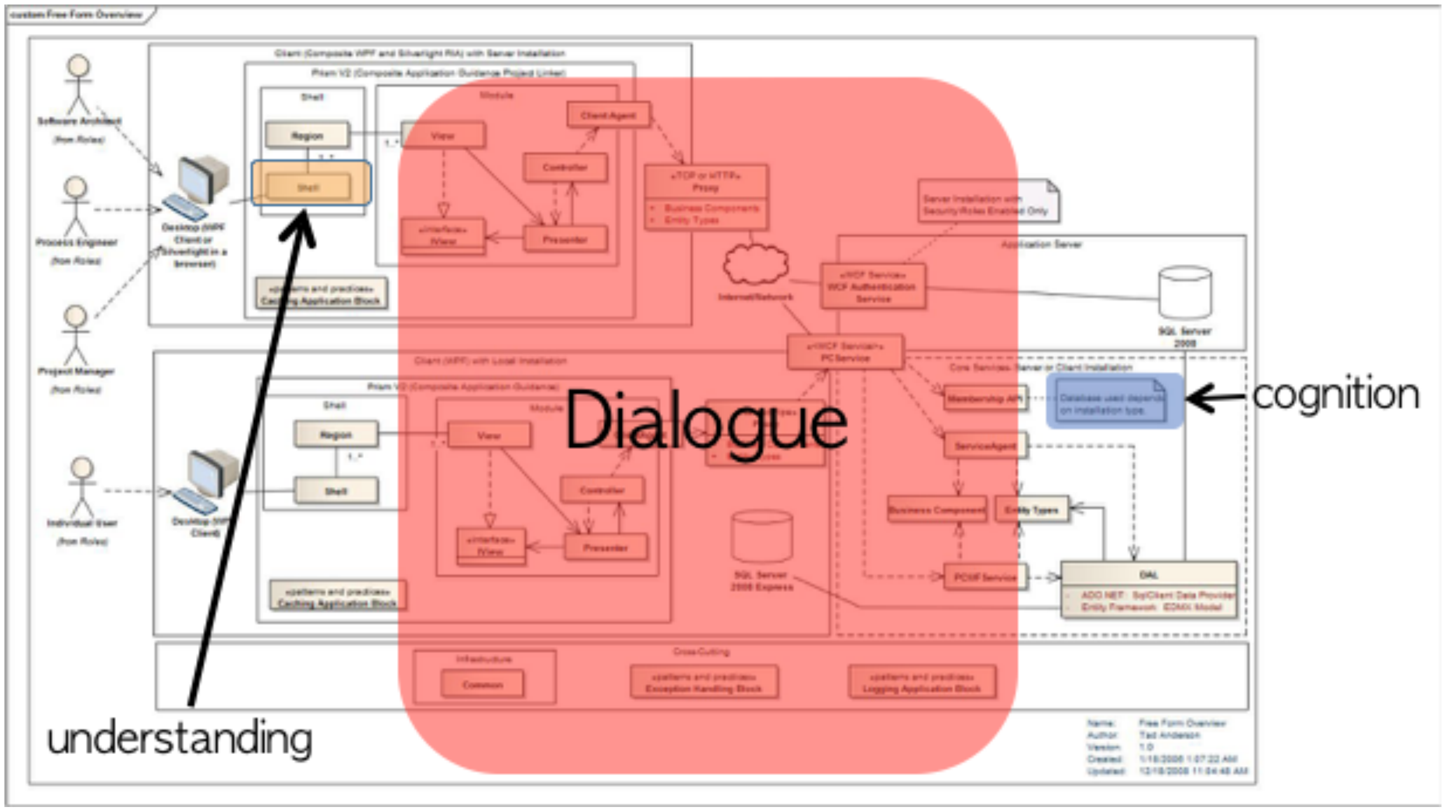
Conversational Agents



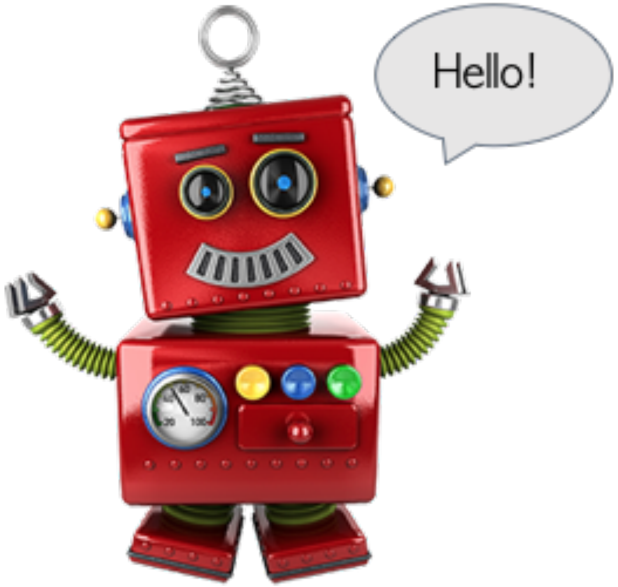


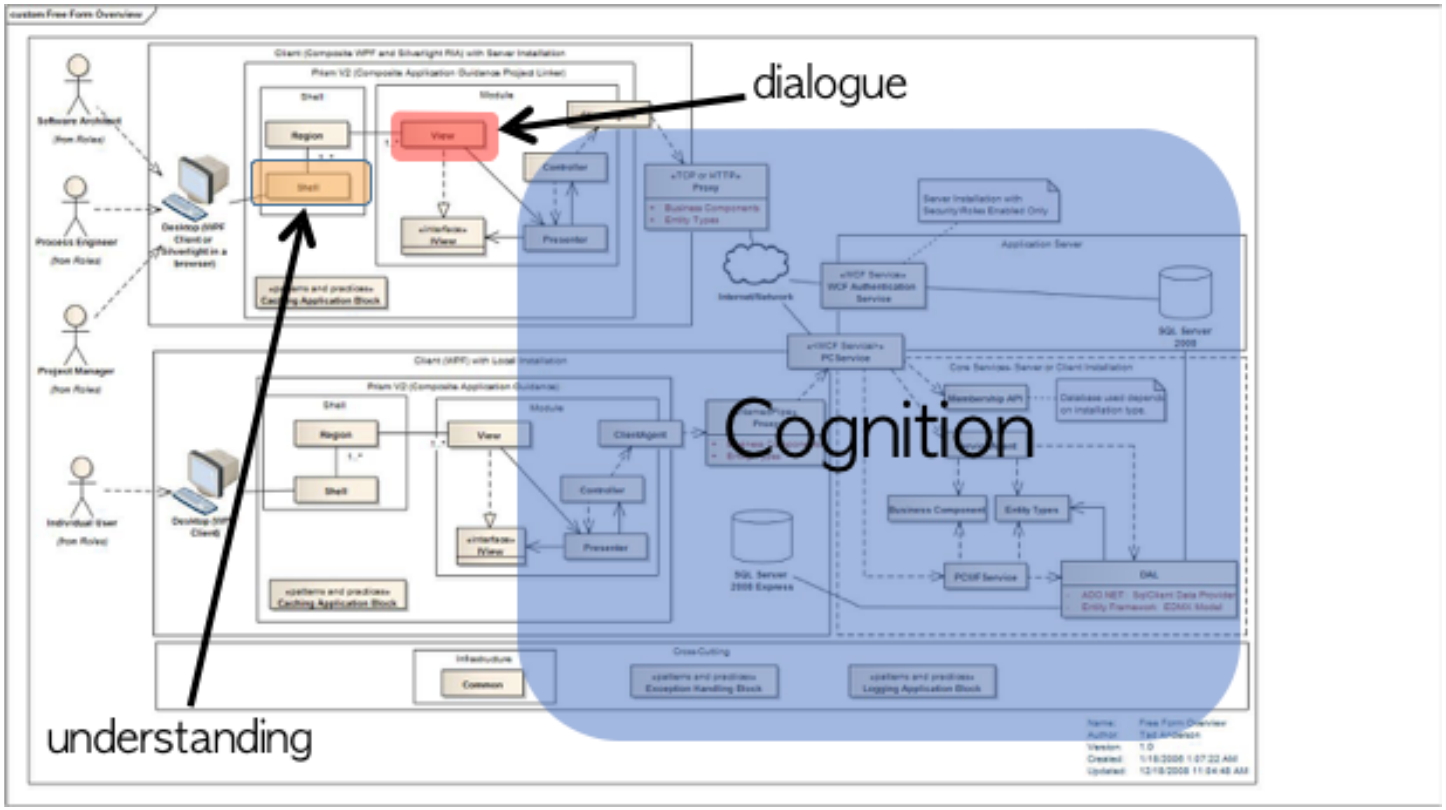
Conversational Agents





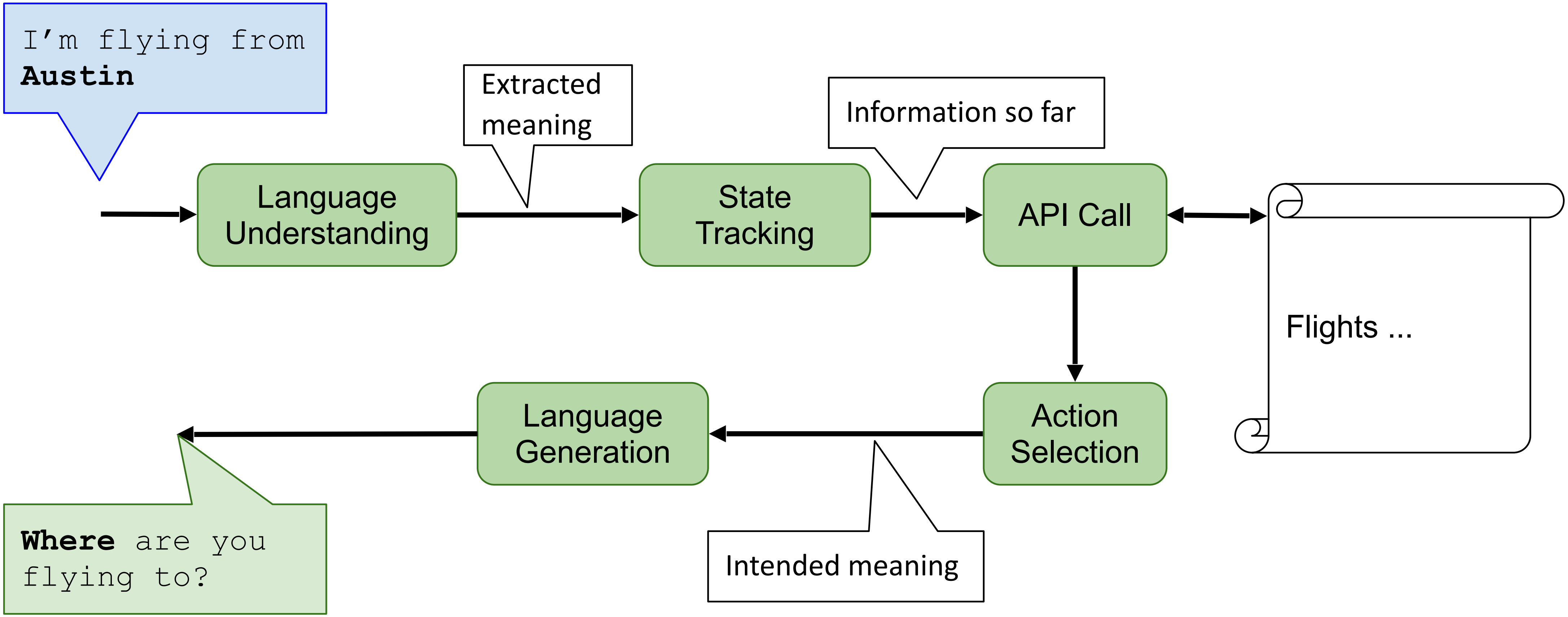
Conversational Agents





Guess which group coined the term “Conversational AI”

Conversational Agent Architecture



Some basic concepts

- Dialogue Turn
  - Turn-taking, overlaps, …
  - Utterance (what was said)
- Dialogue Act
  - *Communication Act (attr op val)*
  - Inform (price = cheap, time = early)
  - Confirm (price < 150)
  - Request (location)
  - Hello ()
  - Book ()
  - ...
- Dialogue History
  - What has been said and done so far
  - Past interactions
- Dialogue State
  - Encoding of relevant parts of history
  - Used by the system to determine most appropriate response
- Domain
  - What the conversation is about
  - Multi-domain
  - “Open” domain

Slot Filling / Information Seeking Dialogues

User: *“I’m looking for a hotel in Heraklion with free parking and at least 3 stars.”*

Slots:[type=hotel, location=Heraklion, services=free-parking, stars >= 3]

SELECT * FROM DataBase WHERE type=hotel …;

Agent: *“Deluxe is a nice 4-star hotel in Heraklion with free parking.”*

In reality:

User: *“Ummm.. so like what about the one with the pool? How much was breakfast there?”*

Slot Filling / Information Seeking Dialogues

The domain: tuple {S, V, A, D}

- S: a set of **slots**
- V: a set of **values for each slot**
  - (a set of sets)
- A: a set of **actions**
  - Dialogue Actions
- D: a **database** of items / records

Example: **Cambridge Restaurants**

System-requestable

User-requestable

Slots: [Food: African, Italian, …
Price: Cheap, Expensive, $15, …
Area: Centre, North, …
Name: Café Uno, Cote, …
Address: Regent Street, …
Phone: 01223 …
Postcode: CB2 1AB, …]

Actions: request(food), inform(name),
…

Dialogue Act Examples

Table 2
An example dialogue and its representation at the dialogue act level.

	Utterance	Dialogue act
U:	Hi, I am looking for somewhere to eat.	hello(task = find,type=restaurant)
S:	You are looking for a restaurant. What type of food do you like?	confreq(type = restaurant,food)
U:	I'd like an Italian somewhere near the museum.	inform(food = Italian,near=museum)
S:	Roma is a nice Italian restaurant near the museum.	inform(name = "Roma", type = restaurant, food = Italian, near = museum)
U:	Is it reasonably priced?	confirm(pricerange = moderate)
S:	Yes, Roma is in the moderate price range.	affirm(name = "Roma", pricerange = moderate)
U:	What is the phone number?	request(phone)
S:	The number of Roma is 385456.	inform(name = "Roma", phone = "385456")
U:	Ok, thank you goodbye.	bye( )

Many datasets to explore nowadays

- Frames
  - Human to human travel booking
- MultiWOZ 2.0
  - Human to human info seeking
- MetalWOZ
  - 47 transational tasks
- Ubuntu
  - Chat about support
- Negotiation
- PolyAI Datasets
  - Reddit
  - Amazon
  - OpenSubtitles
- Twitter
- Multi-party interactions
- Situated understanding
- ...

Spoken Language Understanding

Automatic Speech Recognition

ASR Hypotheses:
["I'm flying from Austin", 0.87]
["I'm flying from Boston", 0.35]
[…]

Natural Language Understanding

NLU Hypotheses:
[inform(origin=Austin), 0.79]
[inform(origin=Boston), 0.13]
[request(destination), 0.05]
[…]

...

Natural Language Understanding

Part of speech tags (e.g. n-grams)
Custom parsing
…

...

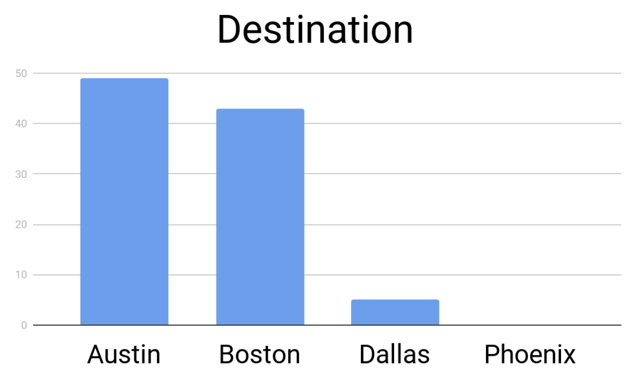
Shallow Understanding

Deep Understanding

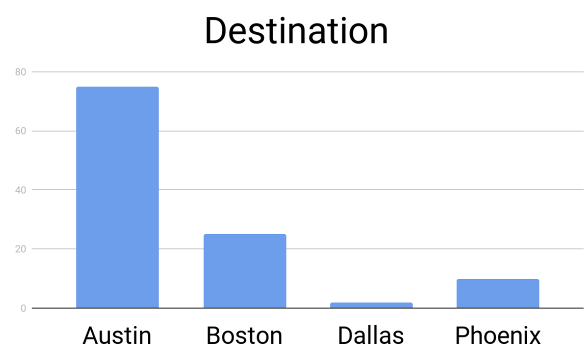
Dialogue State Tracking

[inform(task=flight), 0.95]
[hello(), 0.03]

[inform(origin=Austin), 0.49]
[inform(origin=Boston), 0.43]
[request(destination), 0.05]

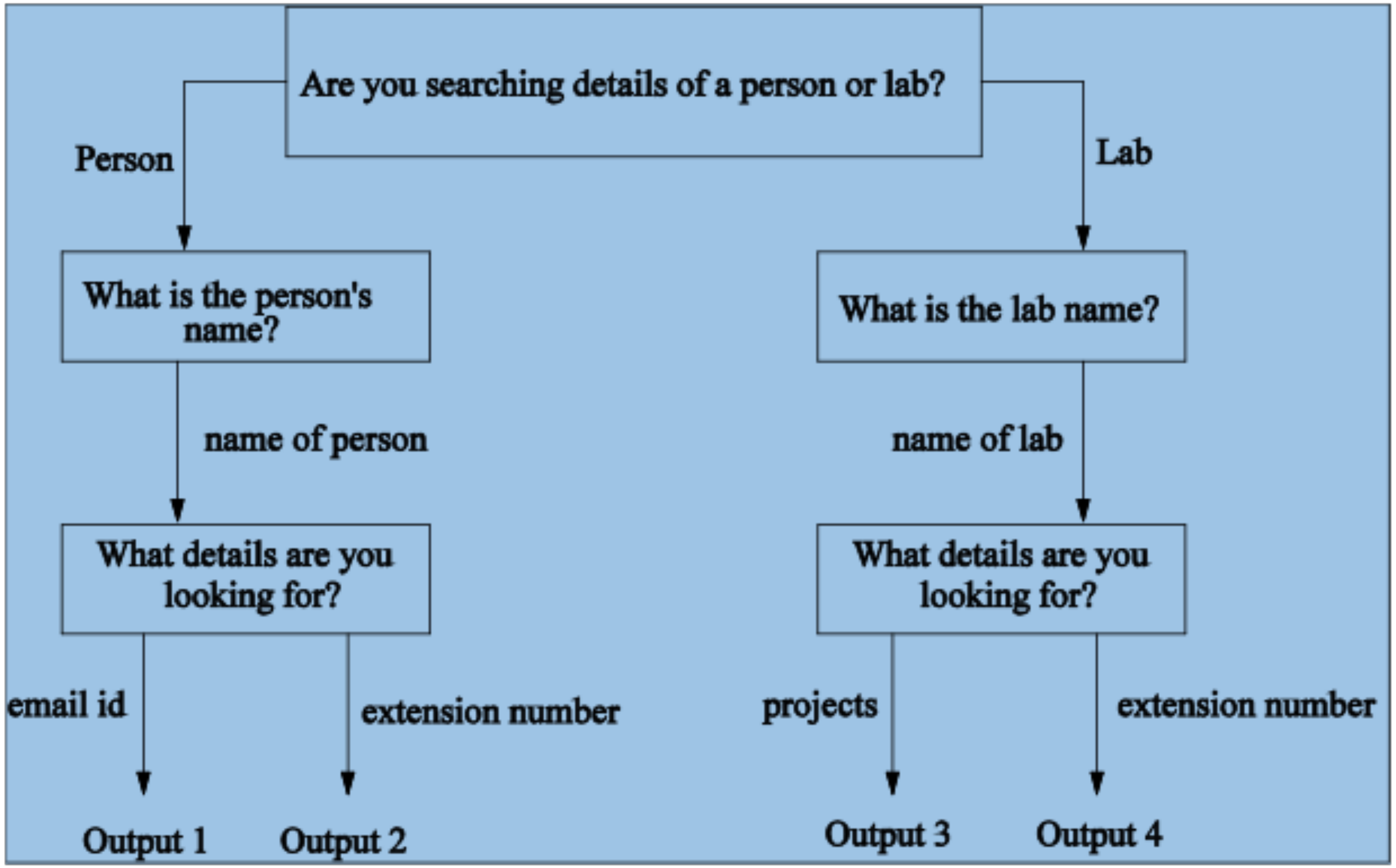


[inform(origin=Austin), 0.81]
[inform(origin=Phoenix), 0.13]



The same principle applies to intent (inform, request …)

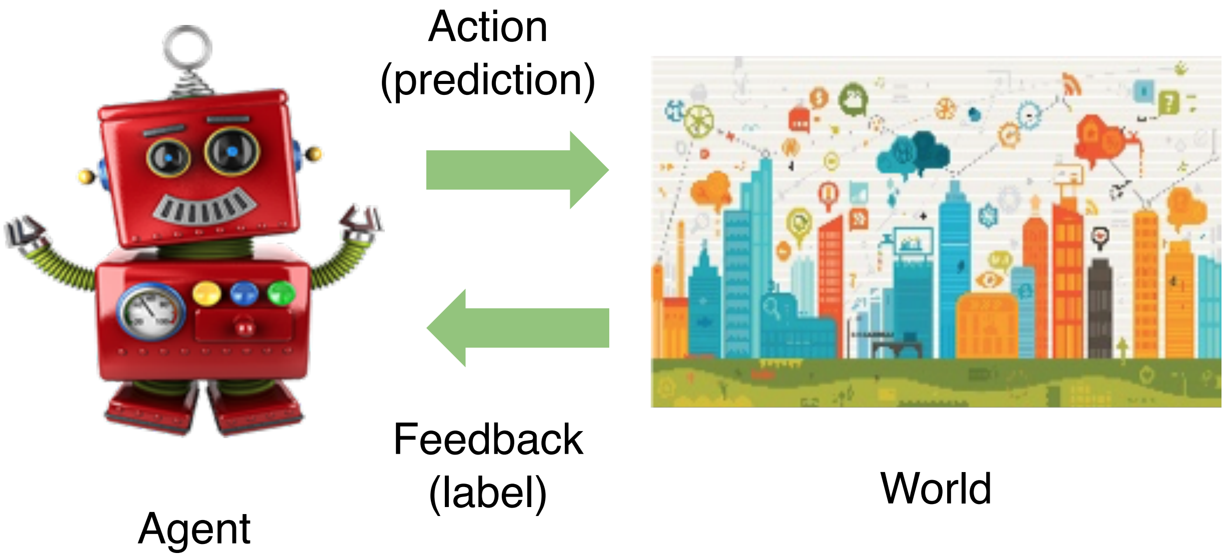
Dialogue Management - Finite State Machine

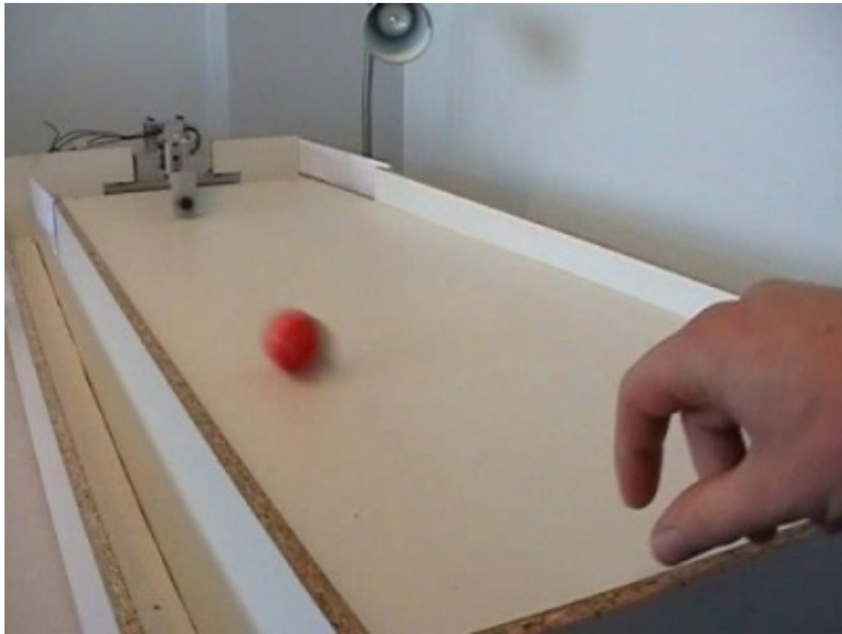


[+] Easy to interpret
[+] Predictable behaviour

[-] Does not scale
[-] Cannot adapt easily
[-] Hard to recover from errors

Reinforcement Learning





Behaviour (policy): Way of choosing an action from a given state
Objective: Learn a good policy!
Method: Explore!

Markov Decision Process

- Lecturer is paid $20,000 per year
- How much will they make in their lifetime?

$$\sum_{years} 20,000 \to \infty$$

- What's wrong here?

- (taxes and)
- Rewards in the future are not worth the same as today!
    - Inflation, things go wrong, ...

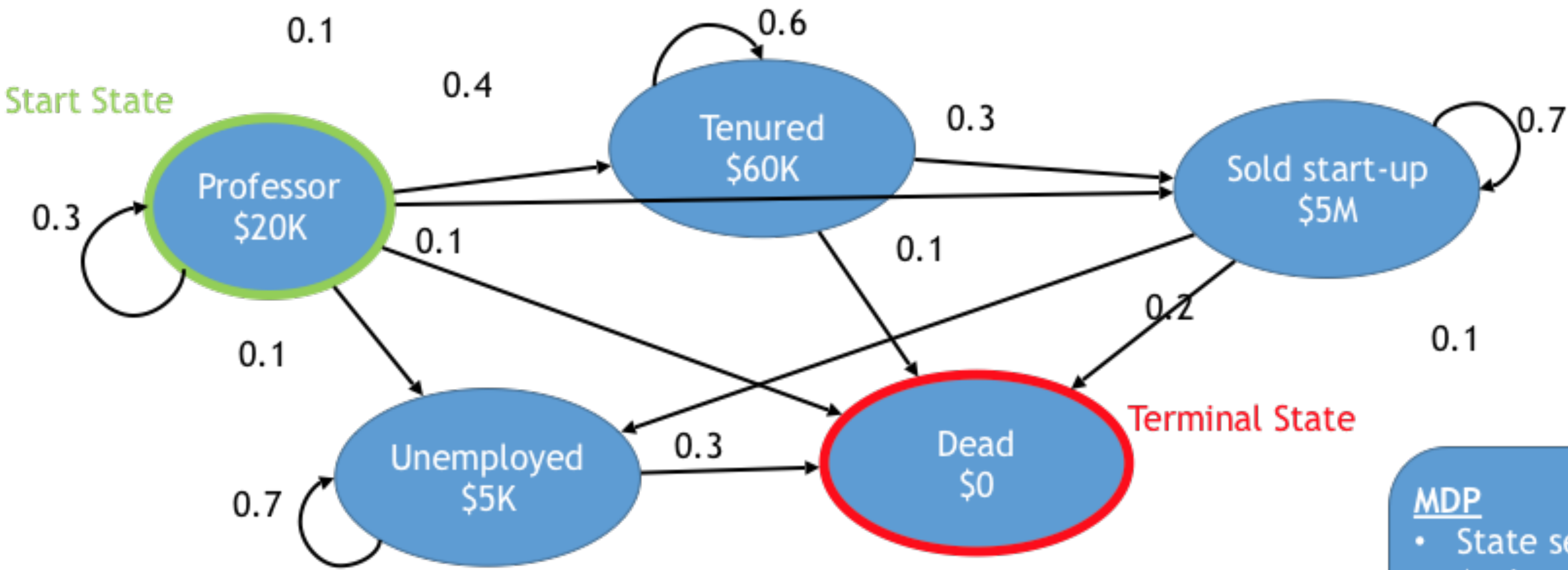
Discounted Rewards

- Money one year from now is worth 90% of today's value
- Money n years from now is worth 0.9^n 0.9 is the discount factor γ
- What is the lecturer's **Future Discounted Sum of Rewards**?
  - J = Reward now + 0.9 · (Reward 1 year from now) + 0.9^2 · (Reward 2 years from now) + …

- Things change in the future!

- Academic life changes
  - On the street - $5,000 per year
  - Lecturer - $20,000 per year
  - Tenured Professor - $60,000 per year
- Other life changes
  - Successful startup - $5,000,000 per year
  - Dead - $0 per year

States and Transitions



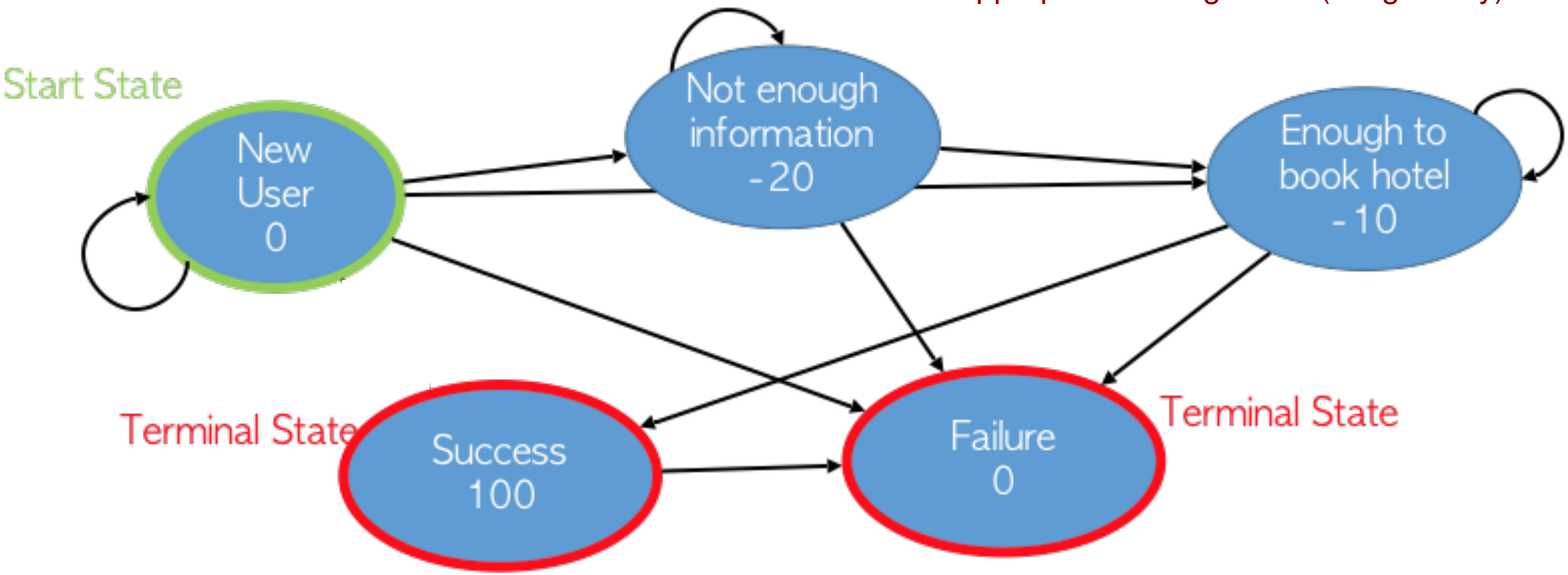
J: Expected discounted future rewards.
Policy: For any state, compute an action to take.
Optimal Policy: For any starting state, the best we can do is follow this policy.
Reinf. Learning: Find policy that maximises J from all starting states.

MDP
- State set
- Action set
- Reward function
- Transition function
- Discount factor

Dialogue as an Optimisation Problem [Levin, 1998]

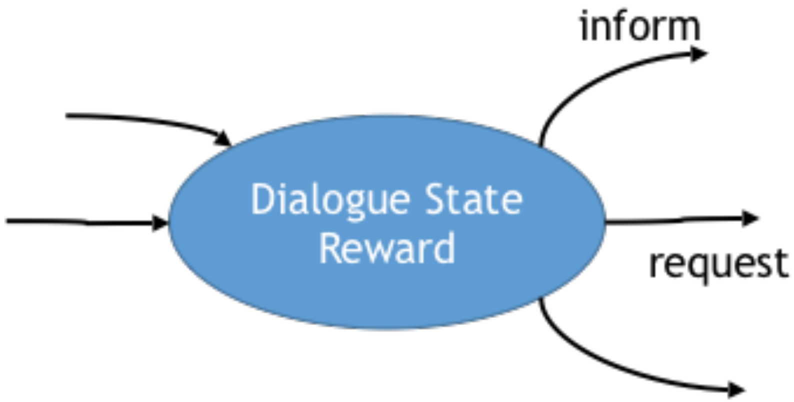
Hotel booking assistant

From each Dialogue State, select the most appropriate Dialogue Act (thing to say)




Questions?

Dialogue as an Optimisation Problem [Levin, 1998]



Agent: Dialogue Manager

State: Dialogue State

Action: Dialogue Act (inform, request, …)

Feedback: Reward Function (dialogue success, user satisfaction, …)

Training: Reinforcement Learning

Data: Human – Human, Human – Machine, Simulation

Reinforcement Learning

- Optimal Control: well-understood dynamics
- Cannot solve MDP for large problems
- Explore state-action space and exploit

- Value function

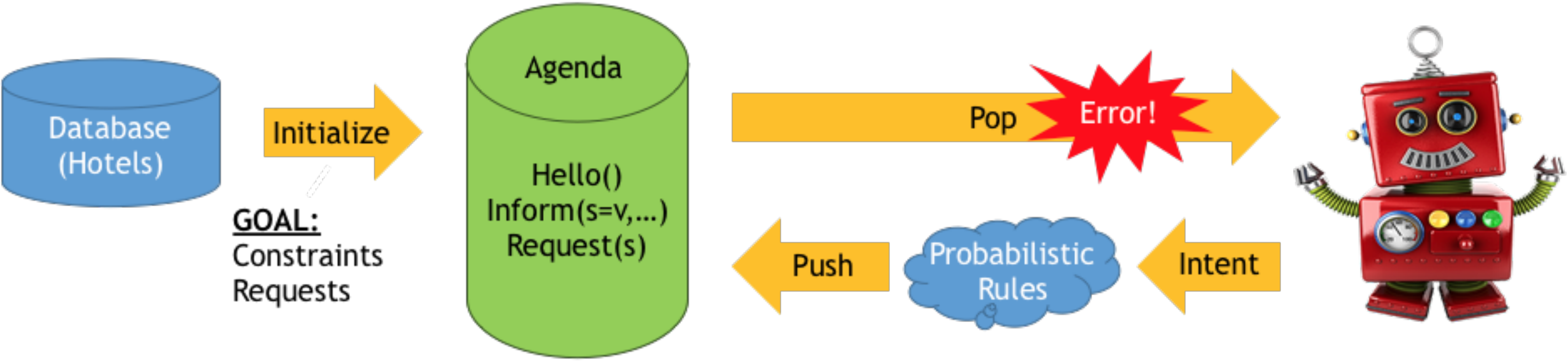
$$V^{\pi}(s_t) = E[\sum_{t=0}^{\infty} \gamma^t R_{t+1} | s_0 = s_t] \quad Q^{\pi}(s_t, a_t) = E[\sum_{t=0}^{\infty} \gamma^t R_{t+1} | s_0 = s_t, a_0 = a_t]$$

- Q-Learning update

$$Q(s_t, a_t) \leftarrow (1 - \alpha) Q(s_{t-1}, a_{t-1}) + \alpha[r_t + \gamma max_a \{Q(s_t, a)\}]$$

Simulated Users

[Schatzmann, 2007]



[+] Allows fast and cheap training
[+] Rules can be learned from data

[-] Usually not realistic
[-] Does not capture linguistic variability

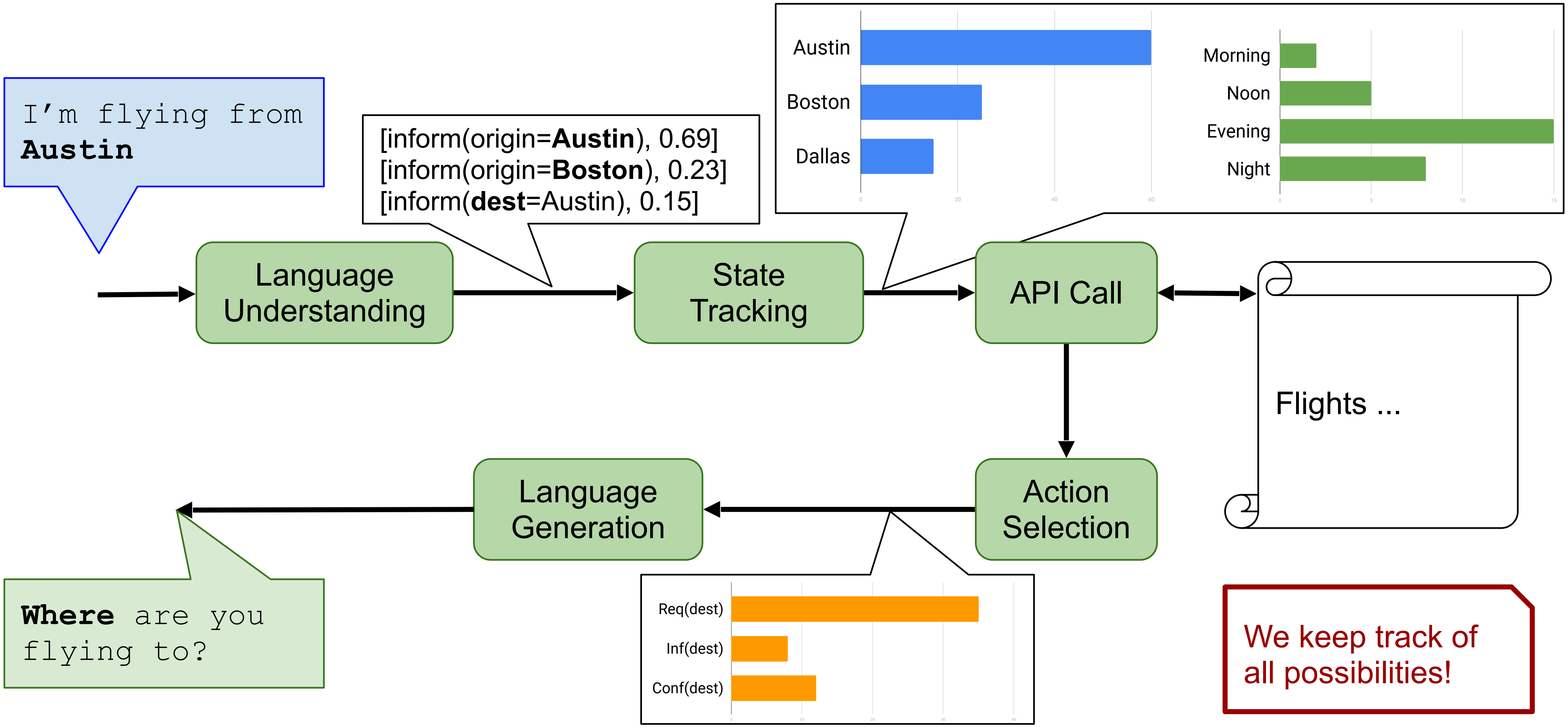
Use to learn a good policy,
then train with humans (crowd-sourcing) to learn optimal policy

Partially-Observable MDP




- Model uncertainty in the environment (e.g. observations, actions, …)
- Belief State: Probability distribution over states
- Policy: Select action from any given belief state

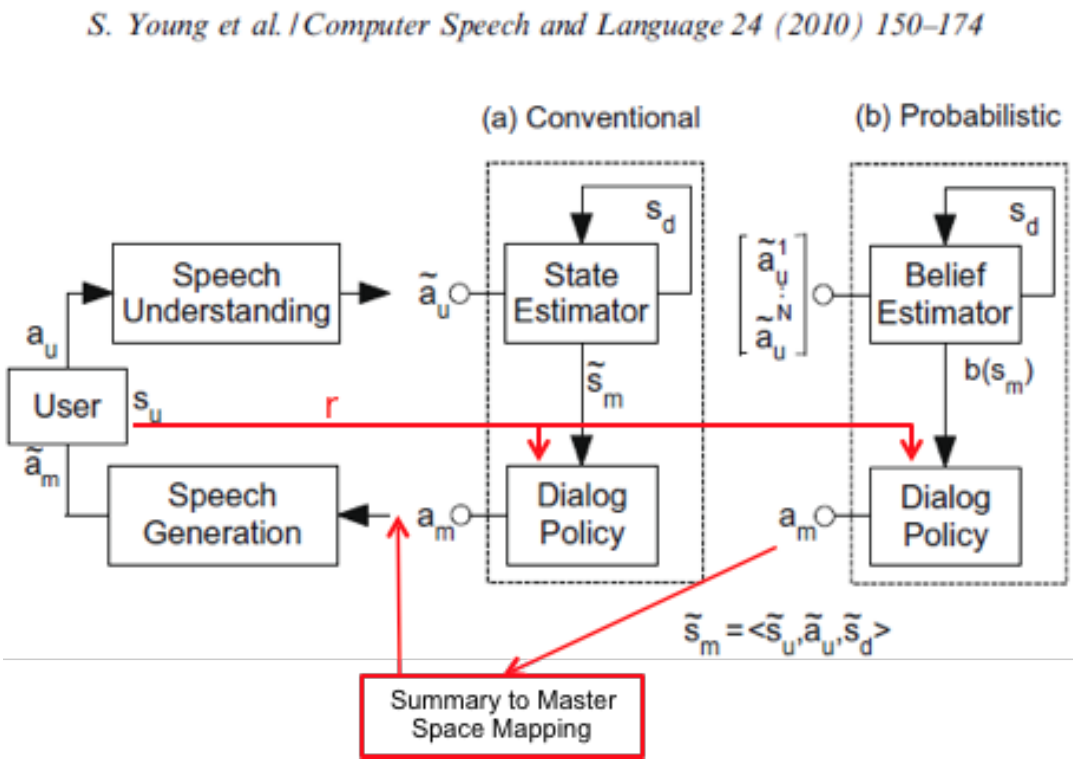
Conversational Agent Architecture With Noise



The Hidden Information State Model [Young et al, 2010]

Slots: pieces of information, e.g. location of hotel
Slot Values: e.g. centre, north, east, …
Actions: dialogue acts e.g. greet, inform, request, …
Dialogue State: information about the dialog so far
History: summarized information
User Goals: estimate of user's intentions
…

Summary State-Action space:
E.g. grid-based approximation
Belief space: over the Summary dialogue state
Action space: Summary actions
Observations: SLU hypotheses
Reward: Arbitrary
Learning: GP-SARSA



The Hidden Information State Model [Young et al, 2010]

POMDP-based dialogue management
Group states with same (similar) belief
Split partition upon new information

Map states and actions to summary space, eg:
use top-2 beliefs
inform(si = v1), …, inform(si = vN) → inform(si = top belief)

Learn in summary space
NAC, GP-SARSA, etc.

Map back to master space
Using history, ontologies, rules, etc.

S. Young et al. / Computer Speech and Language 24 (2010) 150–174

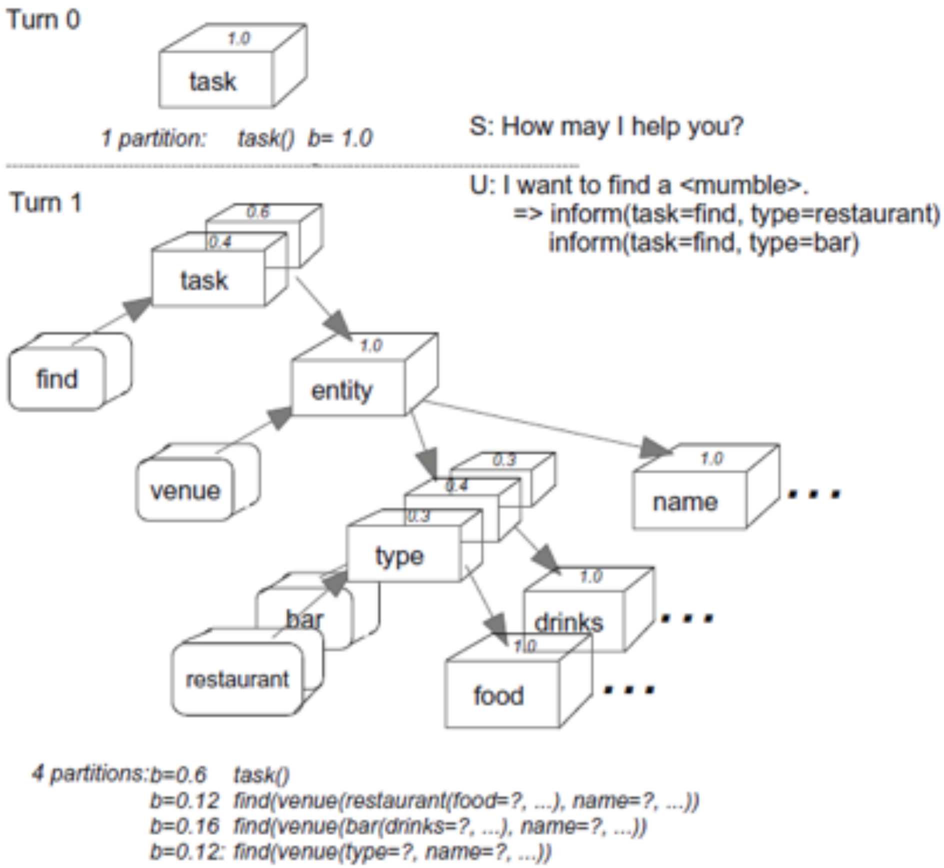
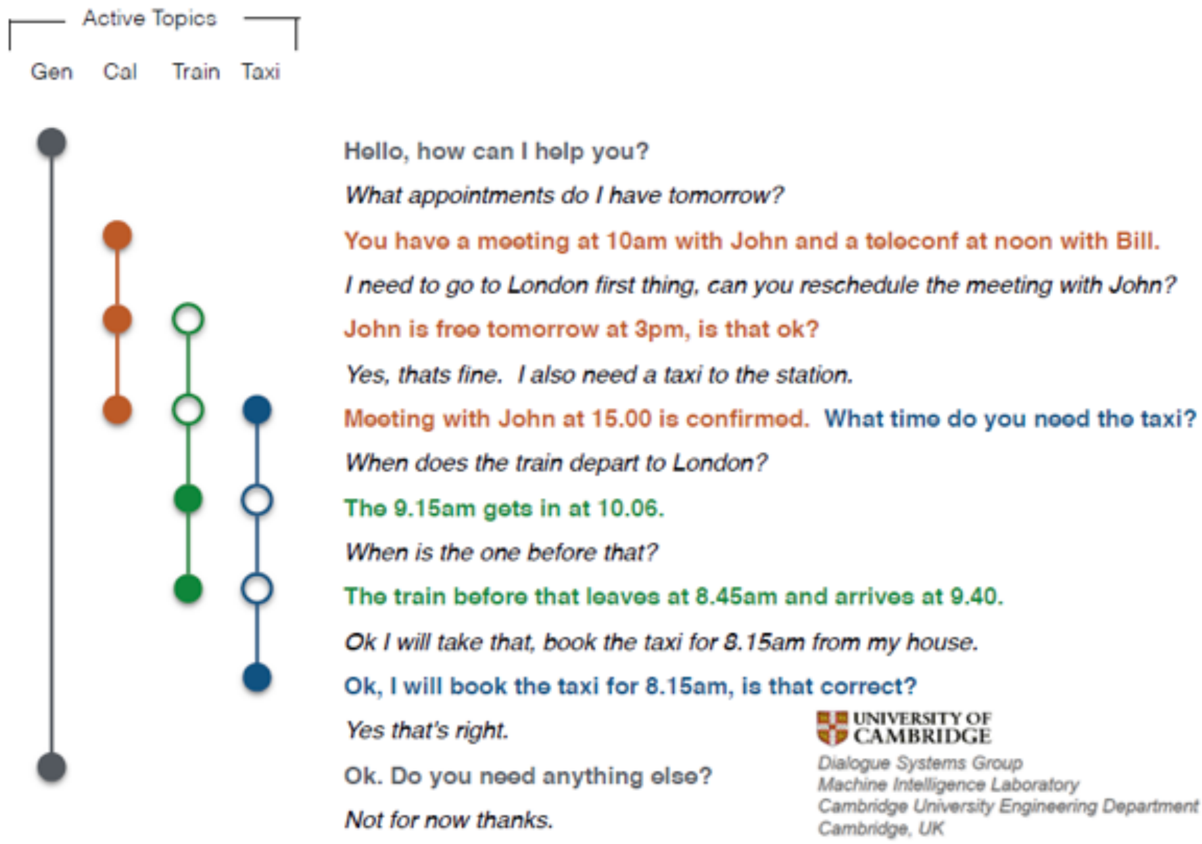


Fig. 5. Illustration of partition splitting.

Multi-Domain Dialogues

- Conversation as multiple (overlapping) topics
- Each topic is a set of nodes in a knowledge graph / ontology
- On topic invocation, only the relevant partition of the belief state is considered



Domain Independent Features

[Wang, 2015]

Traditional speech acts:

$$a(s_1 = v_1, \dots, s_n = v_n)$$

Summary acts:

$$a(s = b_s^{top}, s = b_s^{sec})$$

Descriptive

- #values, value distribution, ..

Impact on results

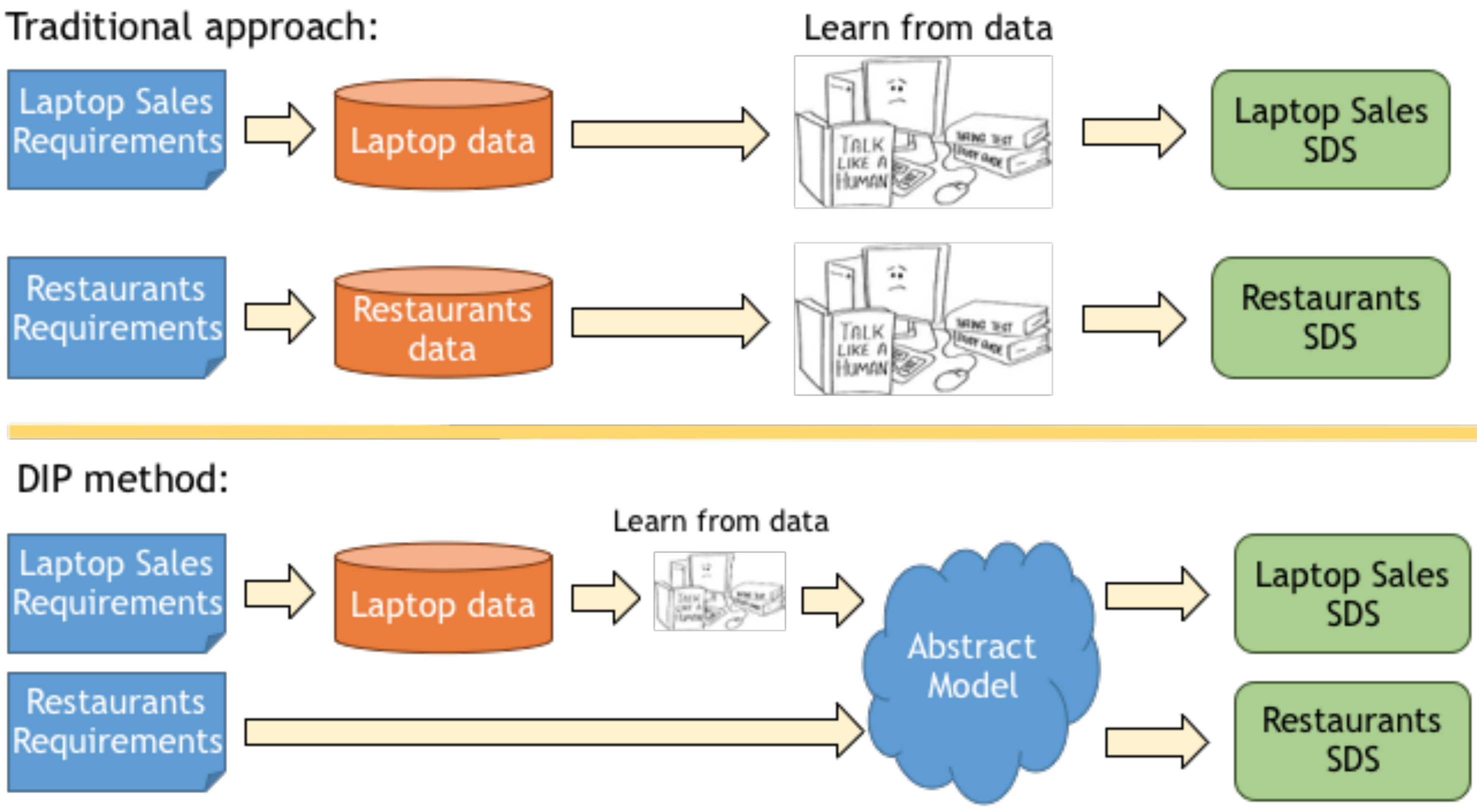
- DB Entropy, if-filled, …

Dialogue

- last user act, top belief, …

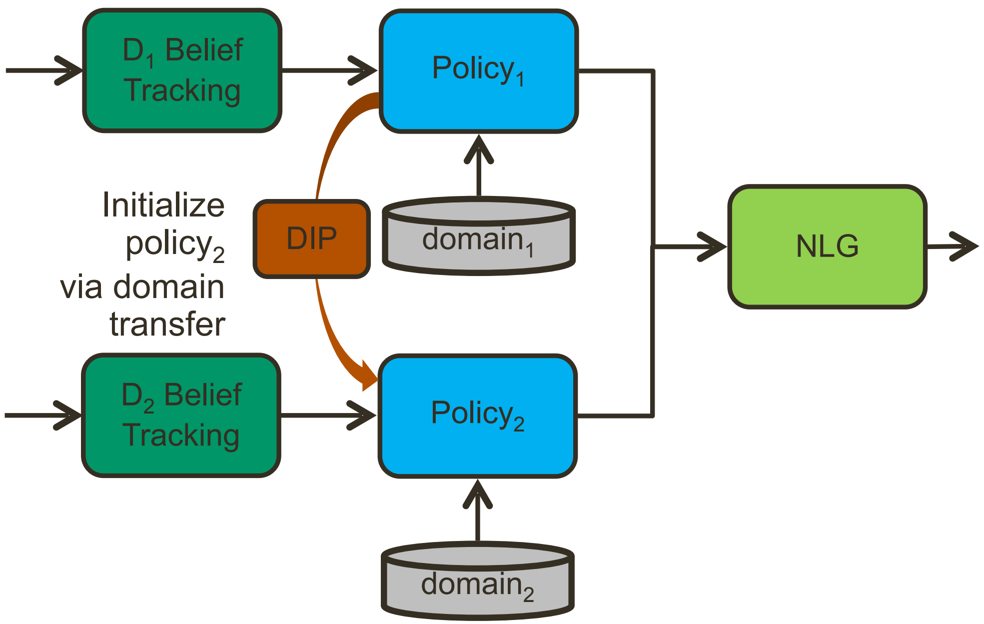
Metrics

- Importance
- Priority

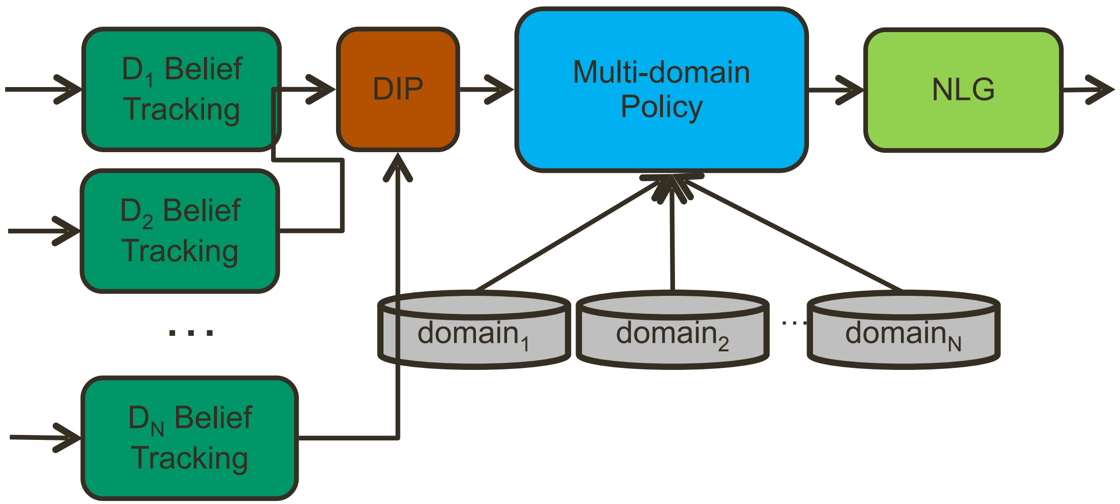


[Papangelis, 2017]

DIP for multiple domains



Domain transfer methods



DIP method

Language Generation

- Template-based

inform(name=Pho, location=centre, price=expensive) → Pho is an expensive restaurant located at the centre.

<NAME> is <PRICE> restaurant located <LOCATION>.
<NAME> is located at <LOCATION> and is <PRICE>.
…
What kind of <SLOT-NAME> are you looking for?
…

- “Traditional”
  - Content planning (what to say)
  - Referring expression generation (pronouns, anaphora, …)
  - Surface realisation (orthography, syntax, …)
- Statistical
  - Select best template
  - Generate words

Kim et al., SIGDIAL 2020

Knowledge-grounded conversations

- (Kim et al, SigDial, 2020)

- DSTC-9 Track on “Beyond Domain APIs: Task-oriented Conversational Modeling with Unstructured Knowledge Access”

- Topical Chat Dataset
  - Conversation data linked to relevant knowledge
  - Entities:
    - 300 common entities
    - 8 topics
  - Facts (for each entity):
    - Crowdsourced fun facts
    - Articles

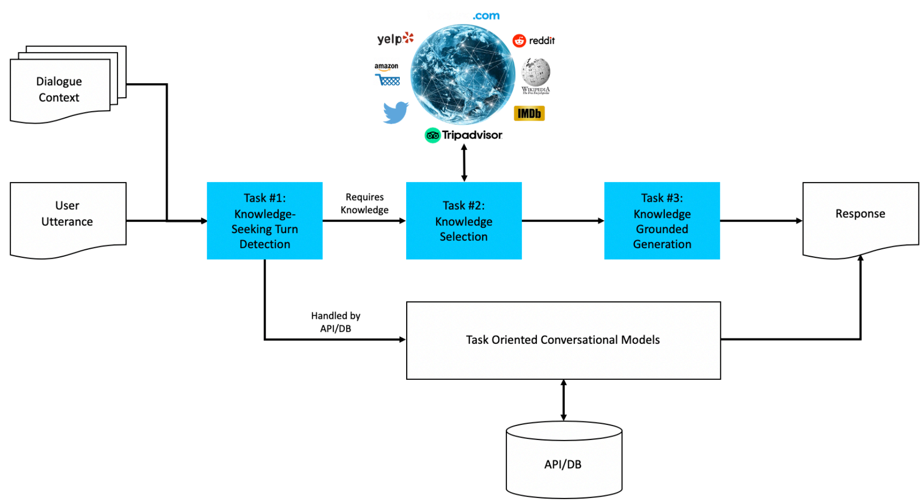
Looking for an upscale restaurant for our anniversary tomorrow.

Do you have an area of interest?

Check near Los Altos

Pompei is a highly-rated Italian restaurant in downtown area, would you like me book there?

Do they have outdoor dining options?



Commonsense reasoning

Images from tutorial: https://homes.cs.washington.edu/~msap/acl2020-commonsense/

2018 human-level performance on reading comprehension on SQuAD (Stanford QA dataset)

2017 super-human performance on speech recognition

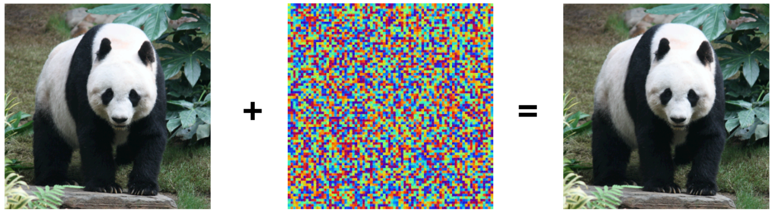
2016 Google neural machine translation

2015 super-human performance on image captioning

super-human performance on object recognition

Commonsense reasoning ≠ Solving a dataset!

Images from tutorial: https://homes.cs.washington.edu/~msap/acl2020-commonsense/



Giant panda

Gibbon

Object Recognition

Szegedy et al, 2014….

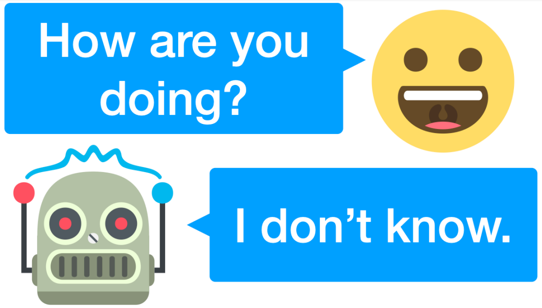
VQA

Jabri et al, 2017

A horse standing in the grass.

Captioning

MacLeod et al, 2017



Dialogue

Li et al, 2016

Open-ended Generation

Holtzman et al, 2018

…. Nikola Tesla moved to Prague in 1880. … Tadakatsu moved to Chicago in 1881.

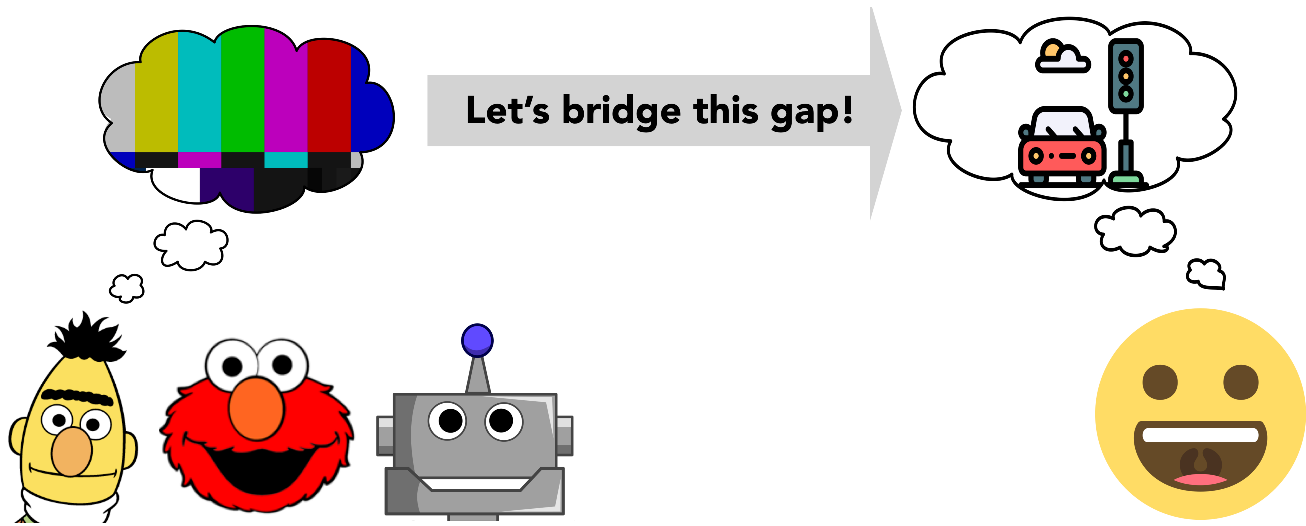
Where did Tesla move in 1880? Chicago

QA

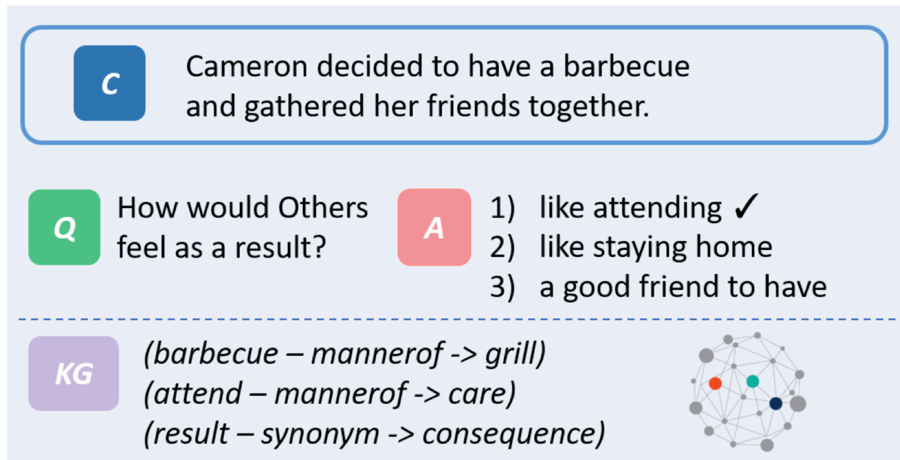
Jia et al, 2017

Commonsense reasoning ≠ Solving a dataset!

Images from tutorial: https://homes.cs.washington.edu/~msap/acl2020-commonsense/



Peters et al., 2018;
Devlin et al., 2018



Social Interaction QA Benchmark

Speech problems

- More uncertainty!
  - One extra layer of processing
  - Very easy to pick up other signals (other people talking, radio, …)

- More phenomena
  - Turn-taking
  - Multi-party conversations
  - Multi-modal conversations

- We’re only scratching the surface here, many of these are fields of their own

Turn Taking

[Examples from: Khouzaimi PhD thesis]

Dialogue Turn: One person speaking

Not as clear cut in spoken language!

Cultures perceive overlap / barge-in differently

HECTOR: I would like to try some exotic destination this summer where I can ...

TANIA: ... Have you ever been to India?

H: We have apple juice...tomato juice...

T: Oh Yeah! That is my favorite, plus, my doctor advised me to have some every day.

Turn Taking

[Examples from: Khouzaimi PhD thesis]

HECTOR: First you put the meat in the oven ...

TANIA: ...aha...

HECTOR: ...then you start preparing the salad...

	T_REF_IMPL	T_REF_RAW	T_REF_INTERP	T_MOVE
H_NONE	FLOOR_TAKING_IMPL	FLOOR_TAKING_RAW	FLOOR_TAKING_INTERP	INIT_DIALOGUE
H_FAIL	FAIL_IMPL	FAIL_RAW	FAIL_INTERP	FAIL_MOVE
H_INCOHERENCE	INCOHERENCE_IMPL	INCOHERENCE_RAW	INCOHERENCE_INTERP	INCOHERENCE_MOVE
H_INCOMPLETE	BACKCHANNEL	FEEDBACK_RAW	FEEDBACK_INTERP	BARGE_IN_CHANGE
H_SUFFICIENT	REF_IMPL	REF_RAW	REF_INTERP	BARGE_IN_RESP
H_COMPLETE	REKINDLE_IMPL	REKINDLE_RAW	REKINDLE_INTERP	END_POINT

Table 3.2: *Turn-taking phenomena taxonomy. The rows/columns correspond to the levels of information added by the floor holder/taker.*

Turn Taking

[Examples from: Khouzaimi PhD thesis]

SYSTEM: Welcome to the Dictanum service. Please dictate your number.

USER: 01 45 *(short silence)*

SYSTEM: 01 45

USER: 65 79 *(short silence)*

SYSTEM: 67 79

Pause length matters!

USER: No, 65 79 *(long silence)*

SYSTEM: Sorry, 65 79

USER: 98 *(short silence)*

SYSTEM: 98

USER: ... *(long silence)*

SYSTEM: The dictated number is 01 45 65 79 98. Is that correct?

USER: Yes.

SYSTEM: Thank you for using Dictanum. Bye.

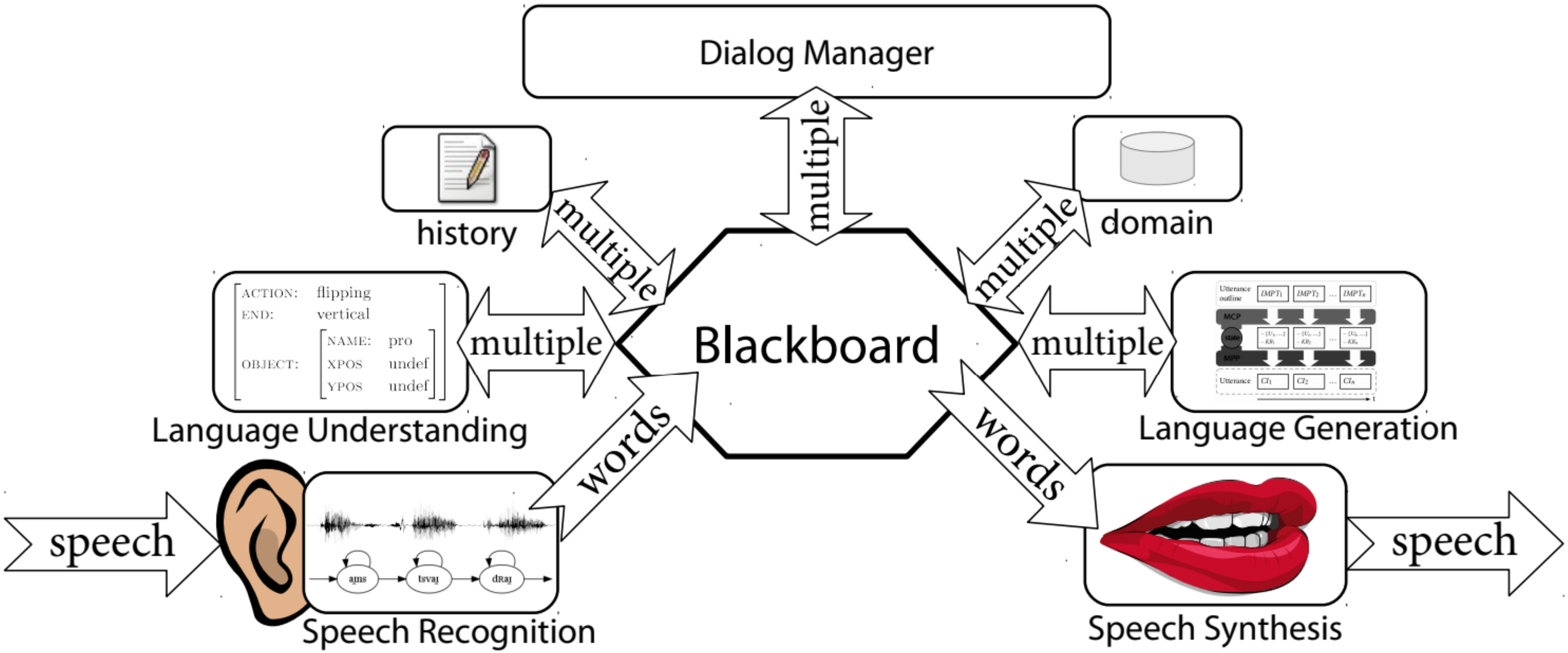
Turn Taking

[Examples from: Khouzaimi PhD thesis]

- Scheduler
  - Dialogue turn → micro turn
  - Actions: Take Floor, Release Floor, Barge In, Backchannel, …
  - Input: speech features, text, …

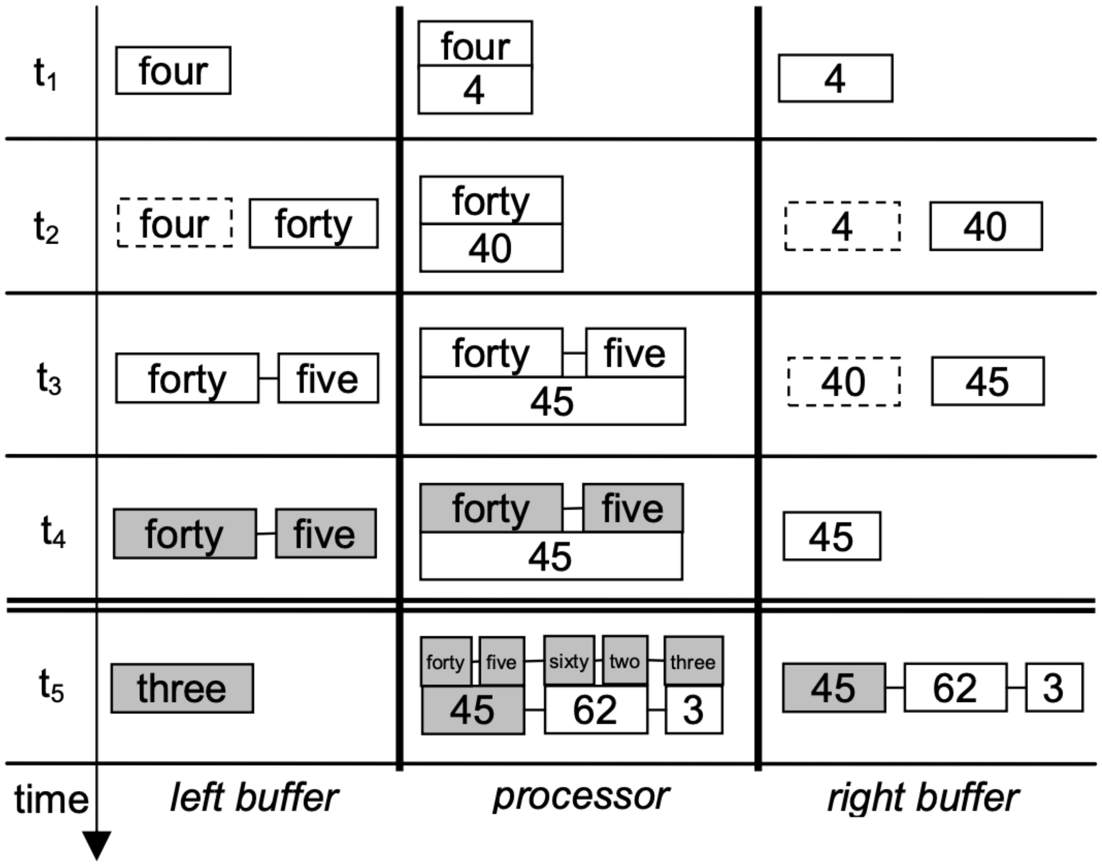
- The Dialogue System must support incremental processing!
  - Must very quickly produce backchannels
  - Must be able to stop LG and TTS almost immediately
  - LU, DST, Policy must support incremental input

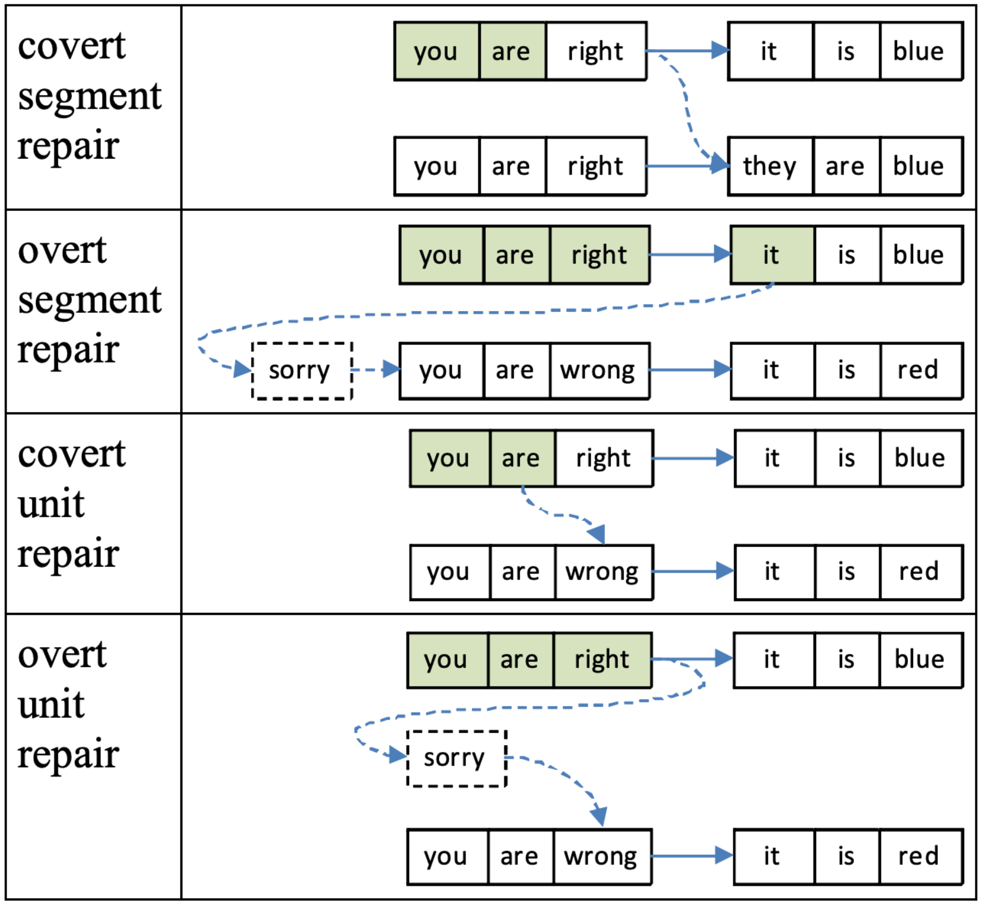
Incremental Dialogue Systems [Baumann, PhD Thesis, 2013]



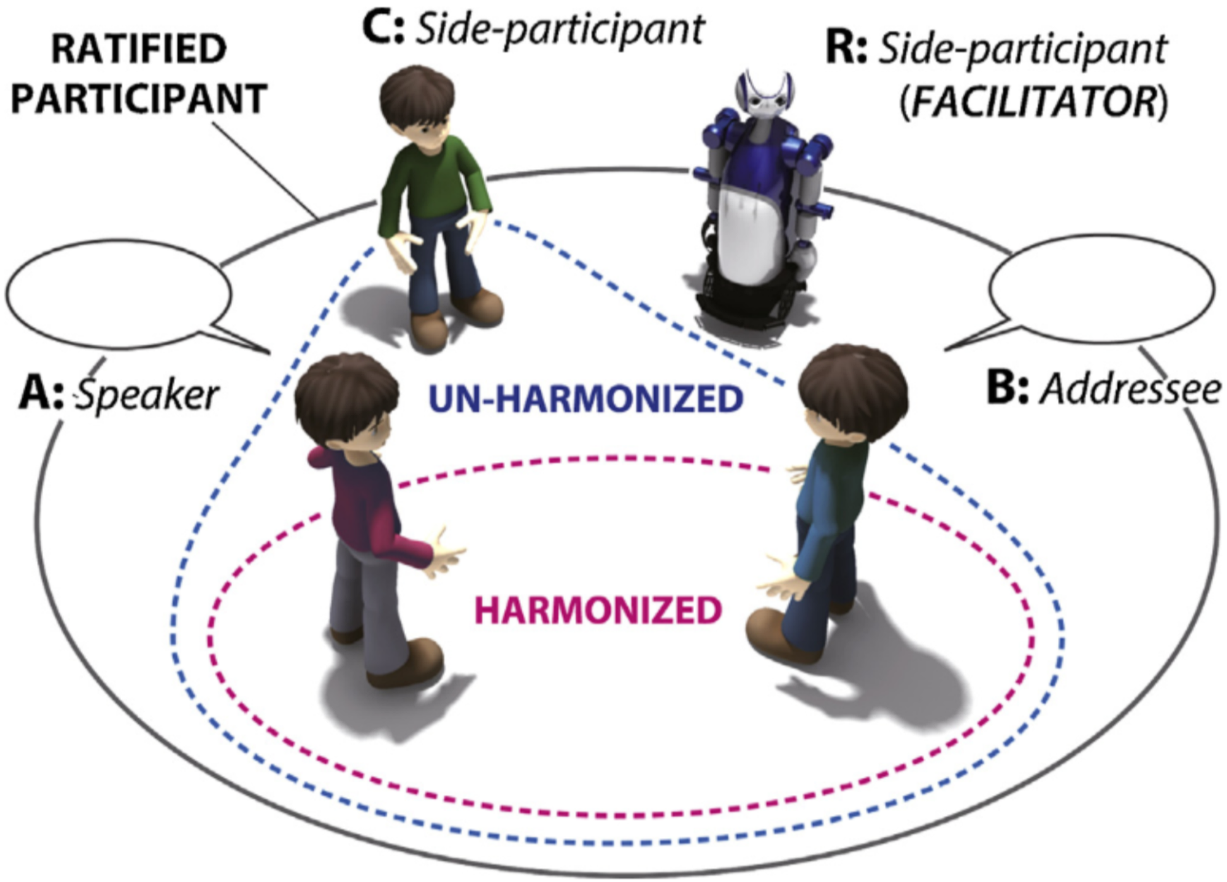
Incremental Dialogue Systems

[Schlangen, 2009][Skantze, 2010]





Multi-Party Dialogues



Multi-Modal Dialogues

Eye contact for turn taking! (e.g. avert eyes = keep turn)

Embodied conversational agents

Virtual or physical

User expectations

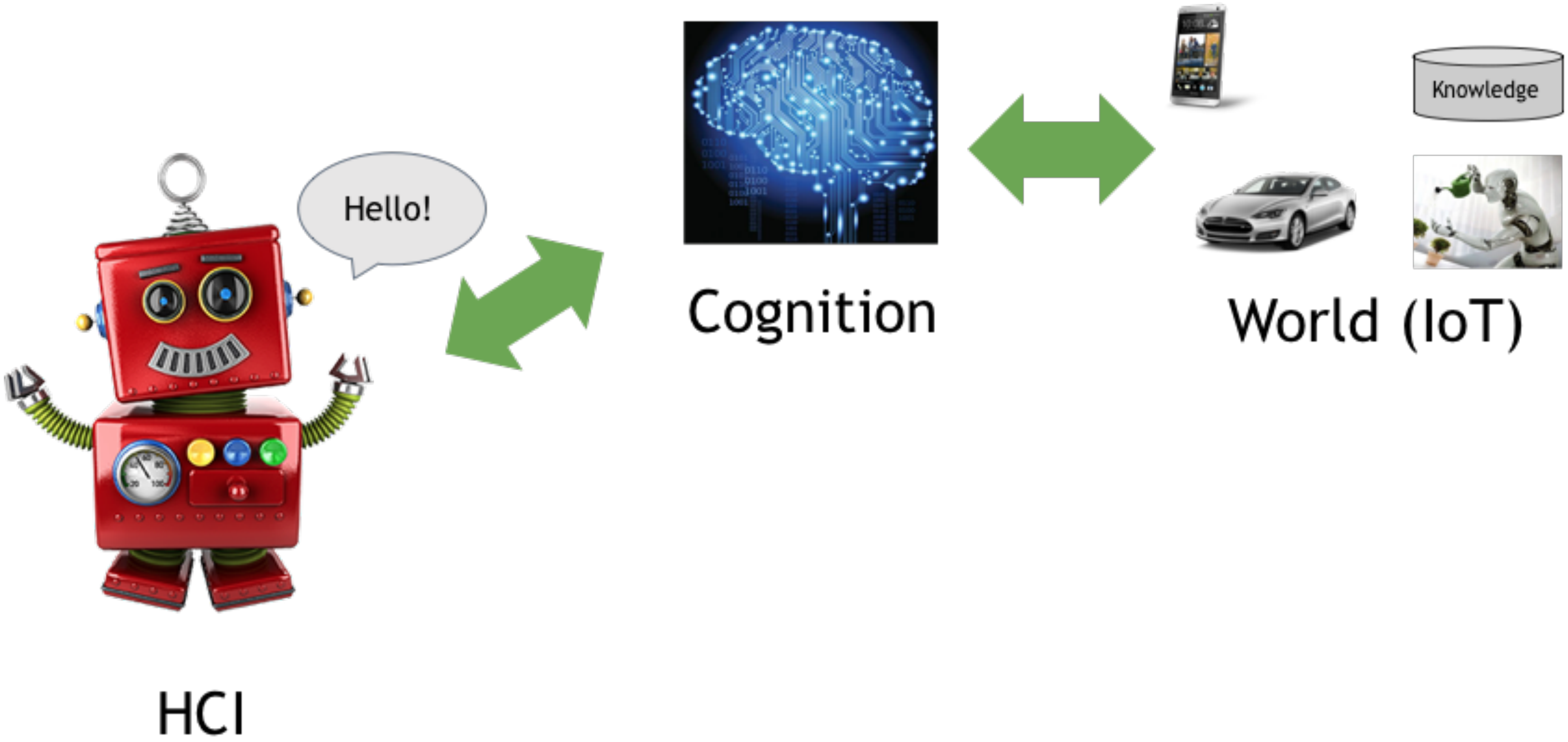
Appearance and personality matter!

Other devices / sensors (fuse information, show or tell, etc)

Subtle cues from multiple modalities


So far...
How to model simple transactional dialogues
...mostly from a DM perspective
But what about other kinds of dialogues?
Long-term? Non-goal oriented?

Socially-skilled Conversational Agents

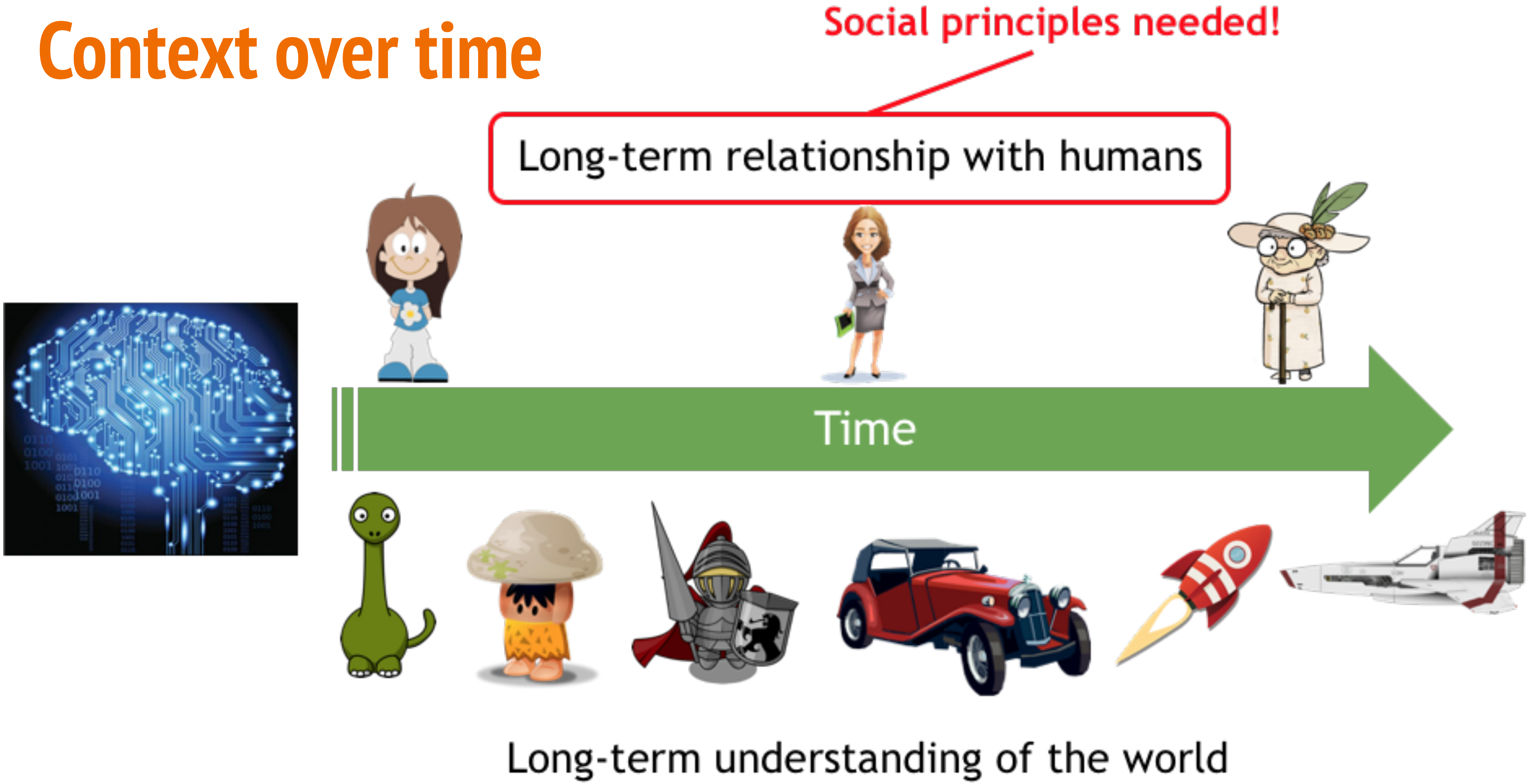


Example: How to build long-term relationships

- Agents that are:
  - Personal assistants
  - Companions
  - Tutors
  - ...

Image from here

Context over time



How has this understanding been used?

- Rapport often (mis)interpreted as helpful or polite, and not viewed within a cultural and situational context.

- Expectation of a one-size-fits-all model of relationships

- However, rapport:
  - is not always polite
  - is not always explicitly helpful
  - fosters unique conversational patterns
  - can be built and demonstrated with technology

More challenges

Uncertainty over effects of actions

Planning ahead over multiple conversations (not just multiple turns)

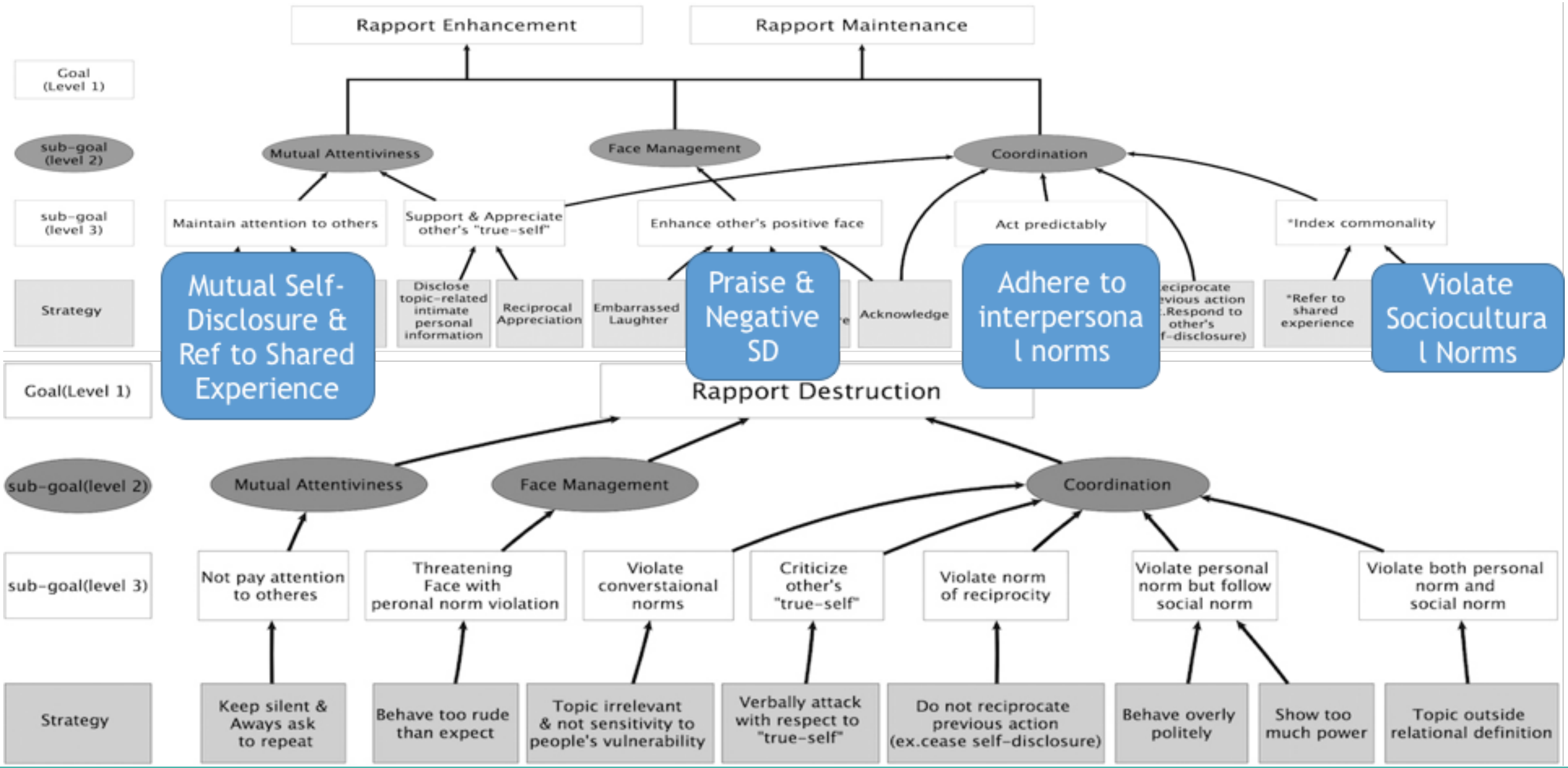
Locutionary, Illocutionary, perlocutionary acts

Non-verbal behaviour (backchannels, head / eye / hands movement, pose, …)

Collecting and annotating the data!

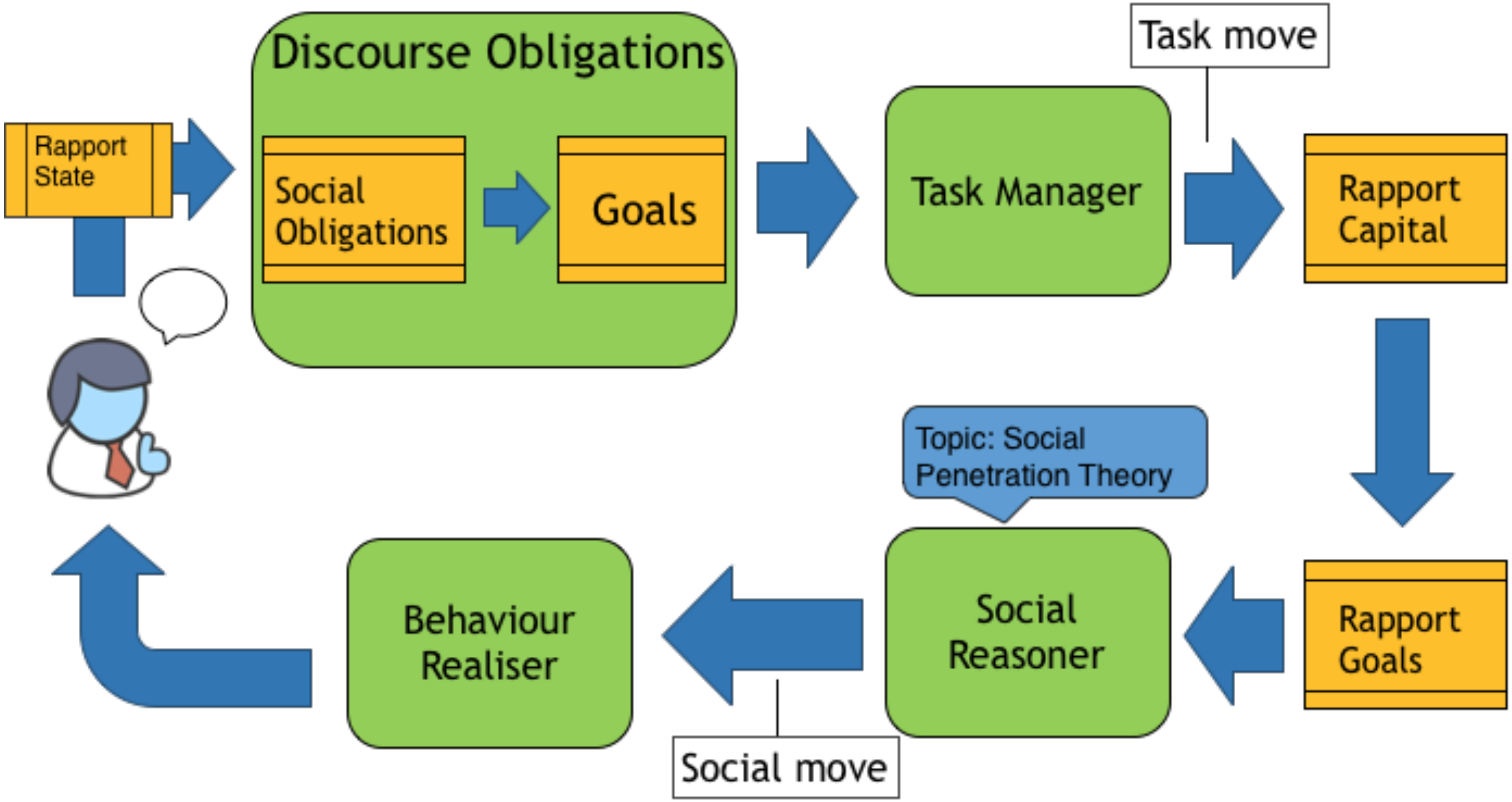
Theoretical Model of Rapport

[Zhao, 2014]



Conversation Flow

[Papangelis, 2014]



Example systems that implement that architecture

Reciprocal Peer Tutoring

CMU InMind

Long-term relationships
With a personal assistant

SARA (Socially-Aware Robot Assistant)
Demo at WEF 2016, China

Later Demo @ World Economic Forum (2017)

Alexa prize

A global university competition committed to advancing the field of conversational AI.

Hold a coherent and engaging conversation for 20 minutes on popular topics

Questions?