

machine learning conference



Learning Robust Dialog Policies for Conversational Error Recovery

Maryam Fazel-Zarandi @fazelzar
Machine Learning Scientist | Alexa



Data Driven Dialog Management (D3M)



Maryam Fazel



Daniel Li



Jin Cao



Peter Henderson



Jared Casale

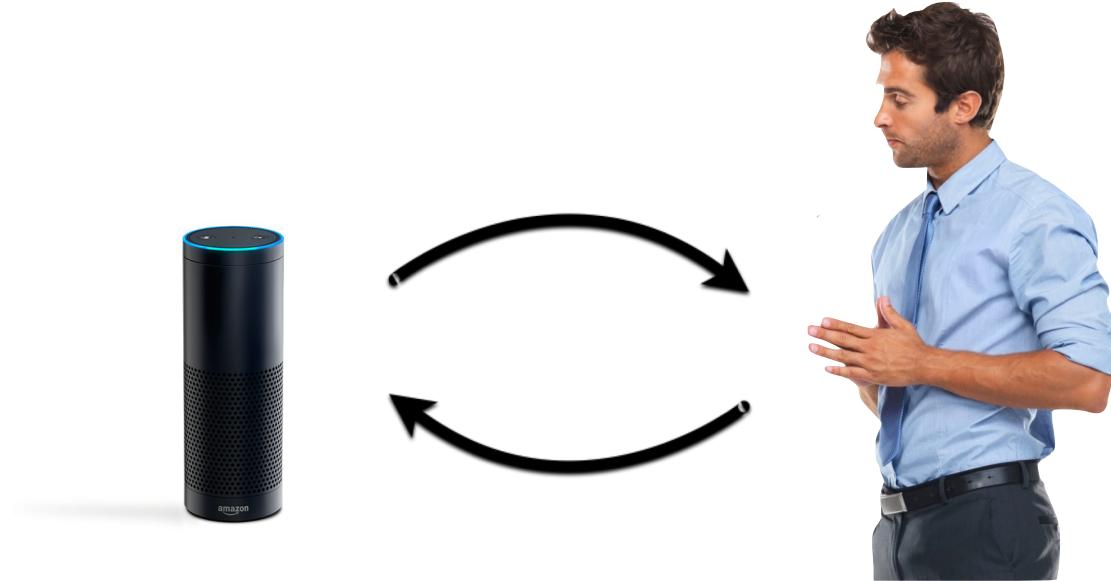


David Whitney



Alborz Geramifard

Motivation



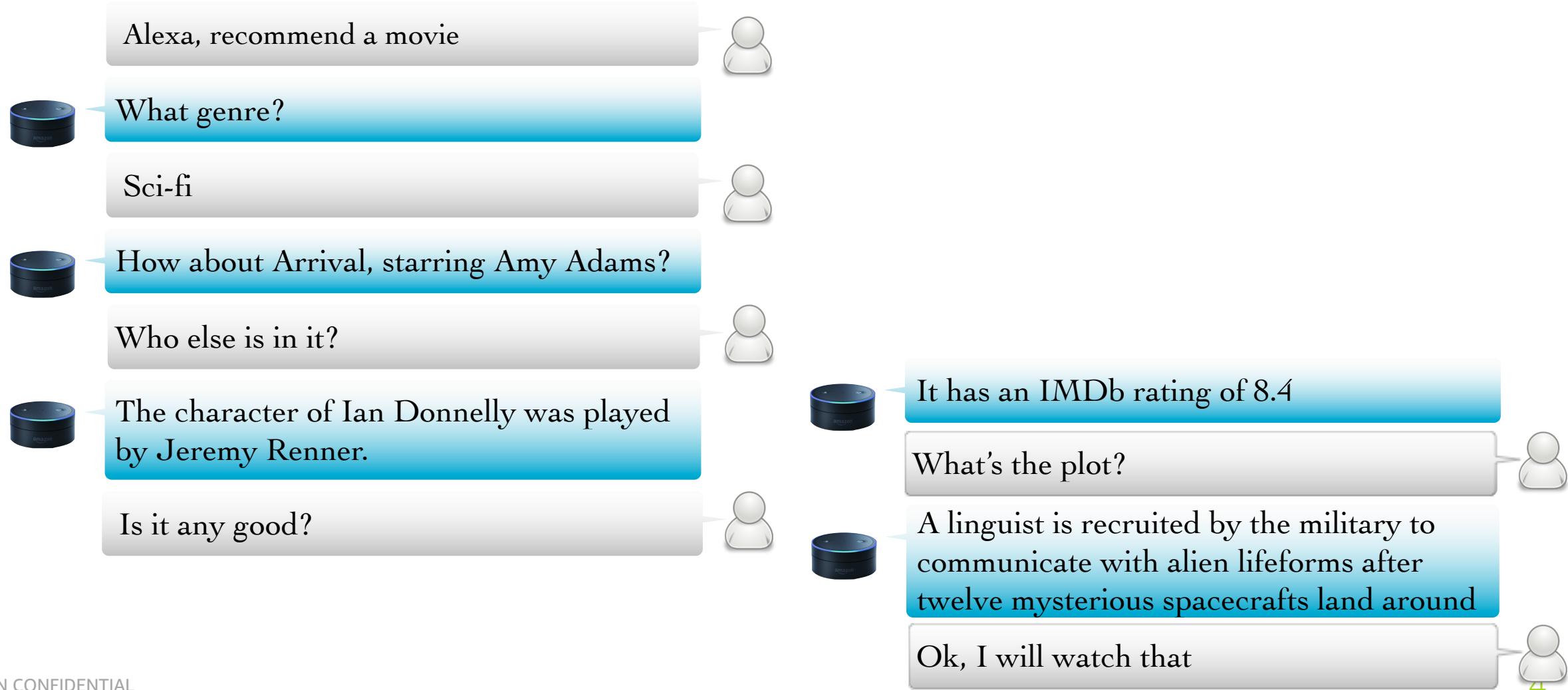
Understand

Understand the customer's need
in a sequence of interactions

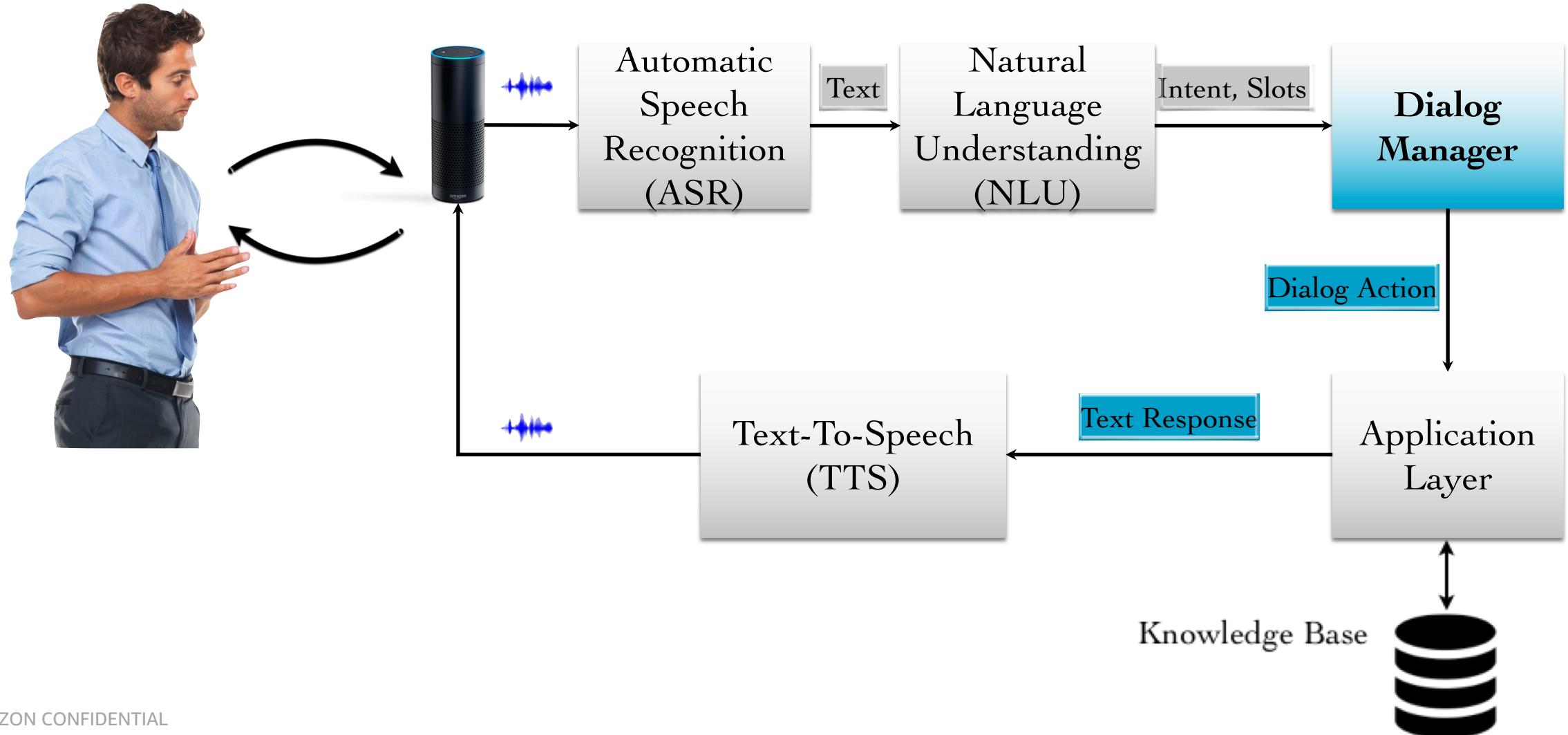
Minimize

Minimize a notion of
accumulated frustration

Example (MovieBot)



Big Picture





Existing Gaps

Manual: Hand written rules → Time consuming

Limited: Only few domains with multi-turn interactions

Non-Adaptive: Can not learn from interactions

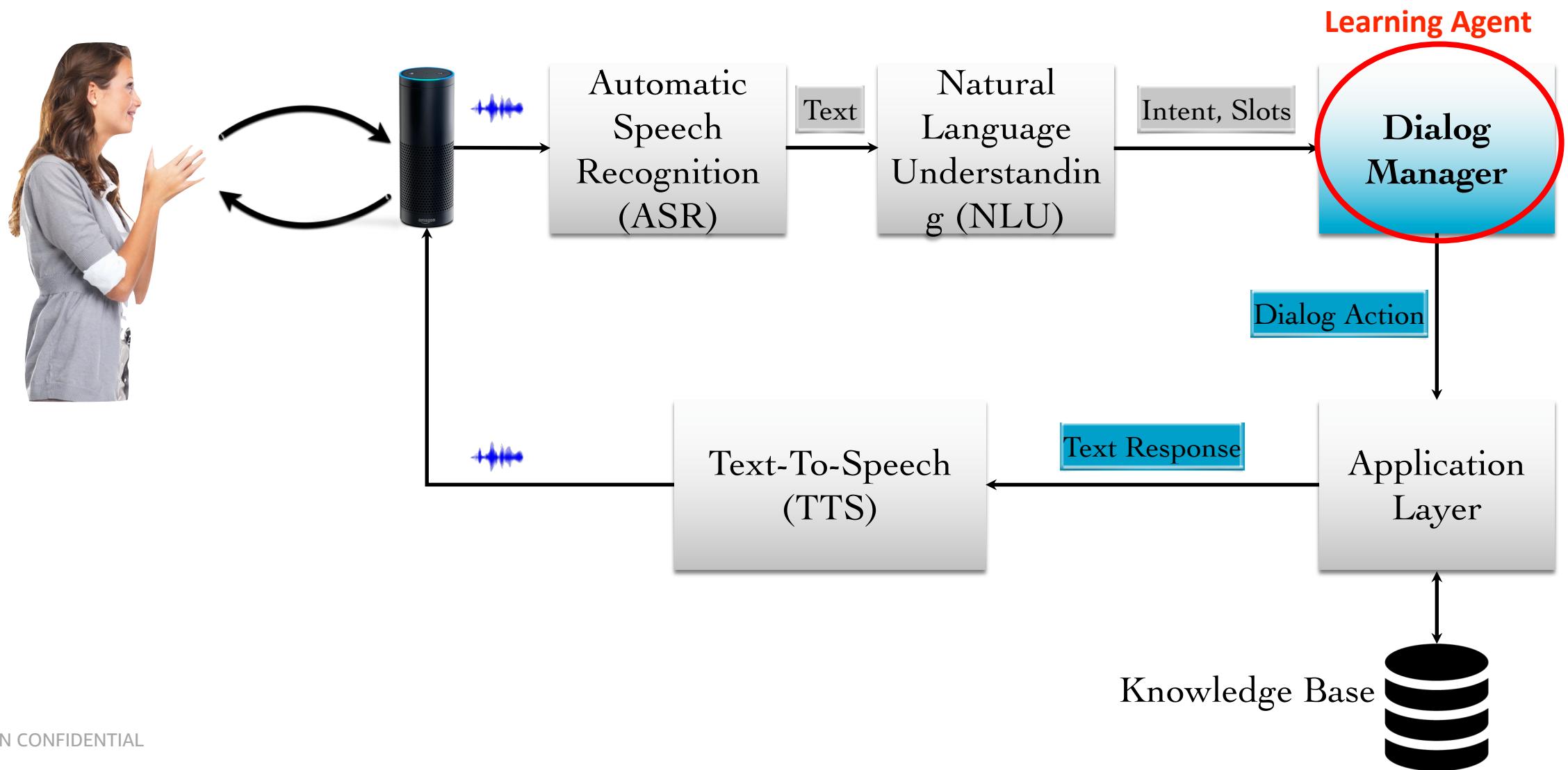
Goal

Automated: Learn the policy using the data

Scalable: Small cost for expanding

Adaptive: Gets smarter with more interactions

Big Picture



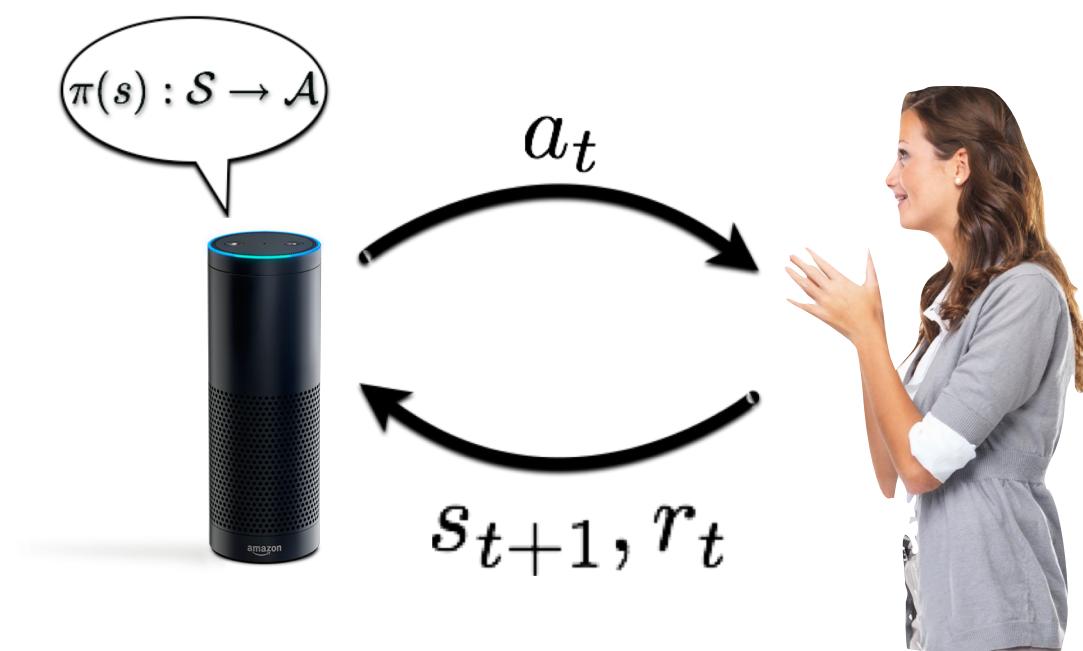


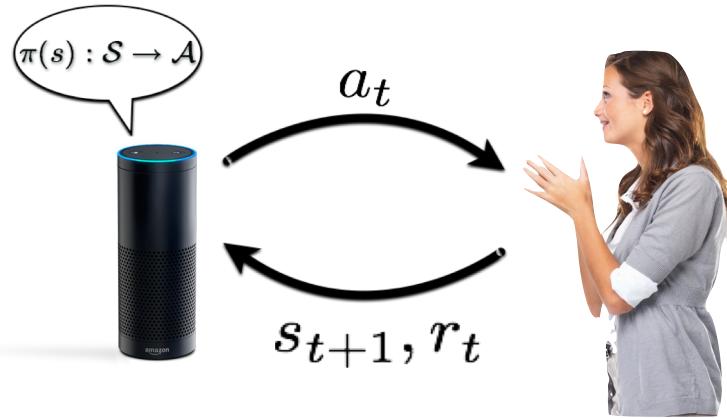
Outline

- Motivation
- Problem Formulation
- Conversational Error Recovery
- Summary and Conclusion

Markov Decision Process (MDP)

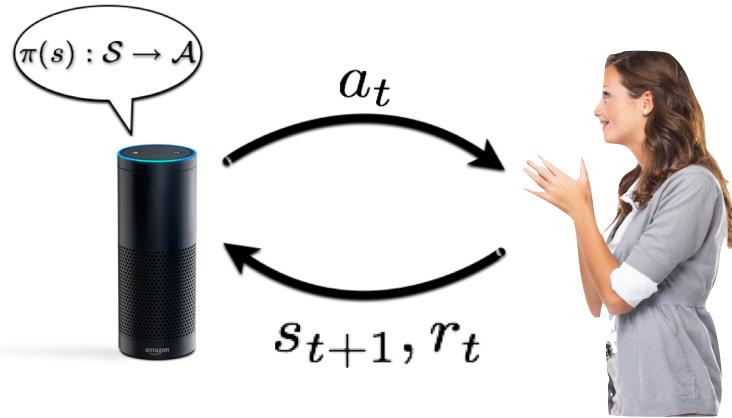
$$(S, A, P_{ss'}^a, R_{ss'}^a, \gamma)$$





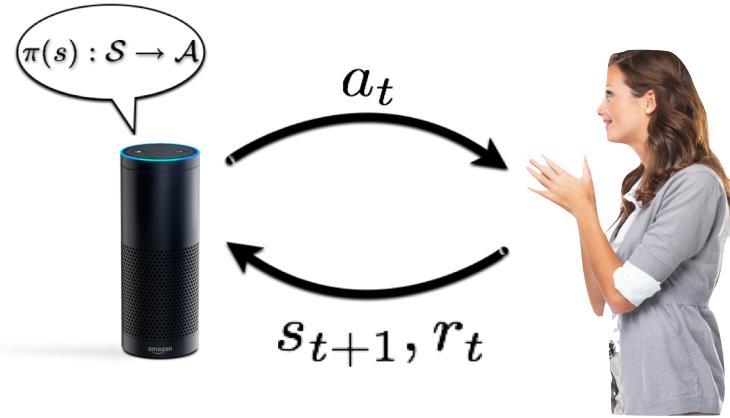
$$(\mathcal{S}, \mathcal{A}, P_{ss'}^a, R_{ss'}^a, \gamma)$$

- **Sufficient statistics** for the past history
- Example dimensions:
 - NLU Output: Intent, Slot, Confidence, etc.
 - ASR Output: Transcription, Lattice, etc.
 - Confirmation History
 - User Preferences



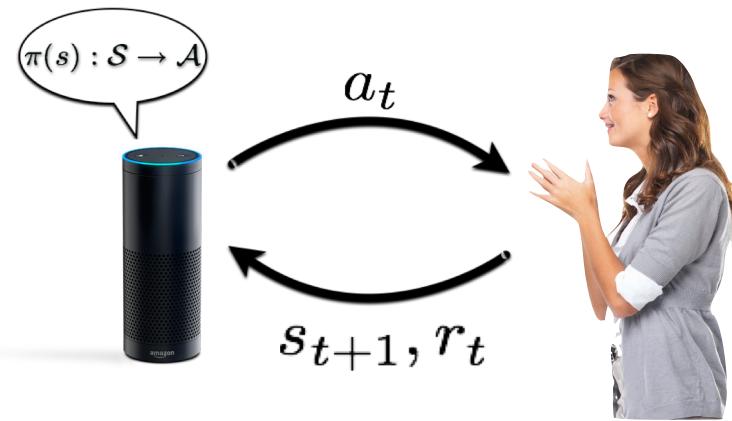
$$(S, \mathbf{A}, P_{ss'}^a, R_{ss'}^a, \gamma)$$

- **Possible Actions** from the current state
- Examples:
 - Confirm the movie title
 - Ask for a genre
 - Recommend a popular movie



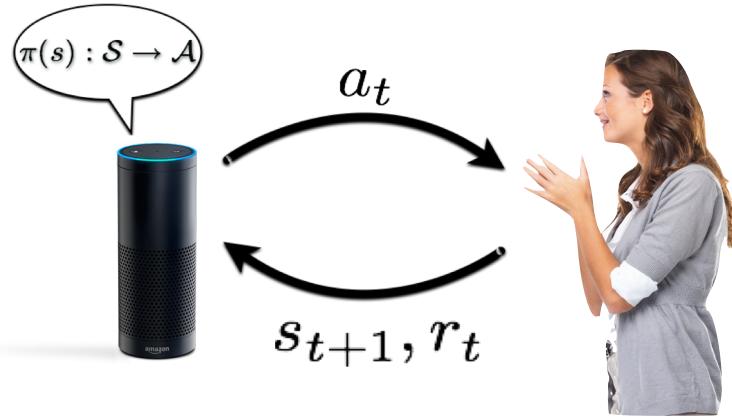
$$(S, A, P_{ss'}^a, R_{ss'}^a, \gamma)$$

- **Transition Probability:** Not Known!
- Distribution over the next state (s') given the current state and action (s, a)
- Dependent on the **user**, ASR, NLU



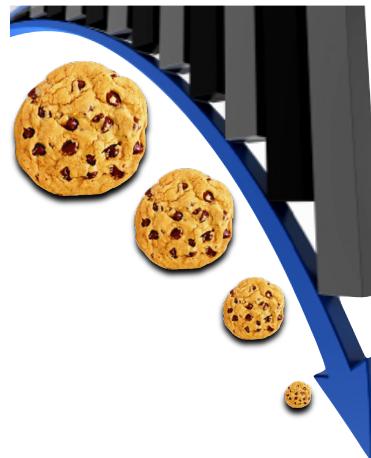
$$(S, A, P_{ss'}^a, R_{ss'}^a, \gamma)$$

- **Reward Function:** Rule-based or User Satisfaction
- Evaluated **manually** or **automatically** (ideal)
- Examples:
 - Successful execution
 - Sentiment Analysis: "Good job!"
 - Voice Cues: "Tuurn oof thee laamp"

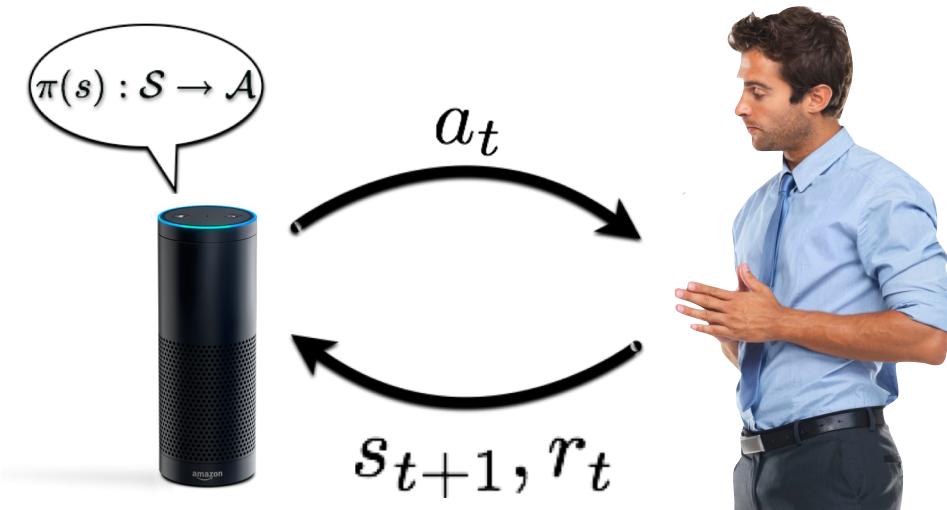


$$(S, A, P_{ss'}^a, R_{ss'}^a, \gamma)$$

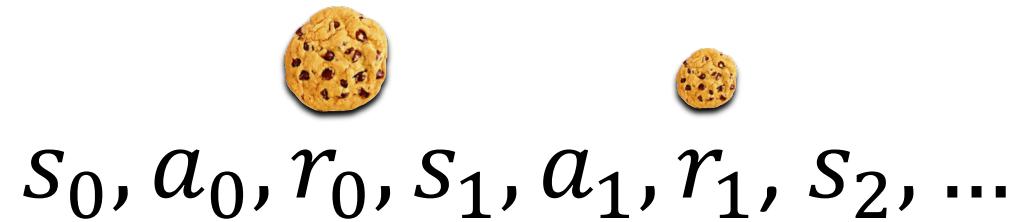
- **Discount Factor:** Future reward vs. immediate reward



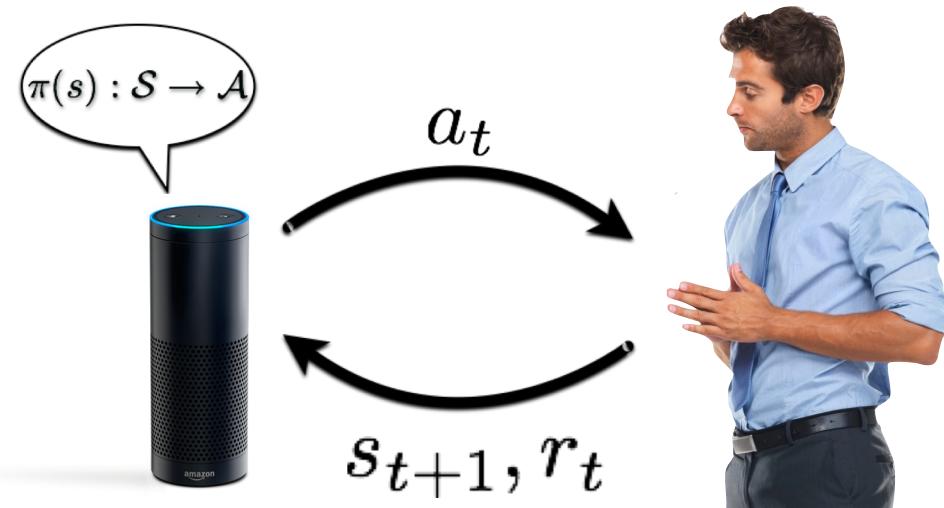
Reinforcement Learning (RL)



- We only see this:



Reinforcement Learning (RL)



- Optimization:

The diagram shows a series of four chocolate chip cookies decreasing in size, followed by a small circle representing a continuation of the sequence. Below this, the Q-learning formula is displayed:
$$Q^\pi(s, a) = E \left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \middle| s_0 = s, a_0 = a, \pi \right]$$

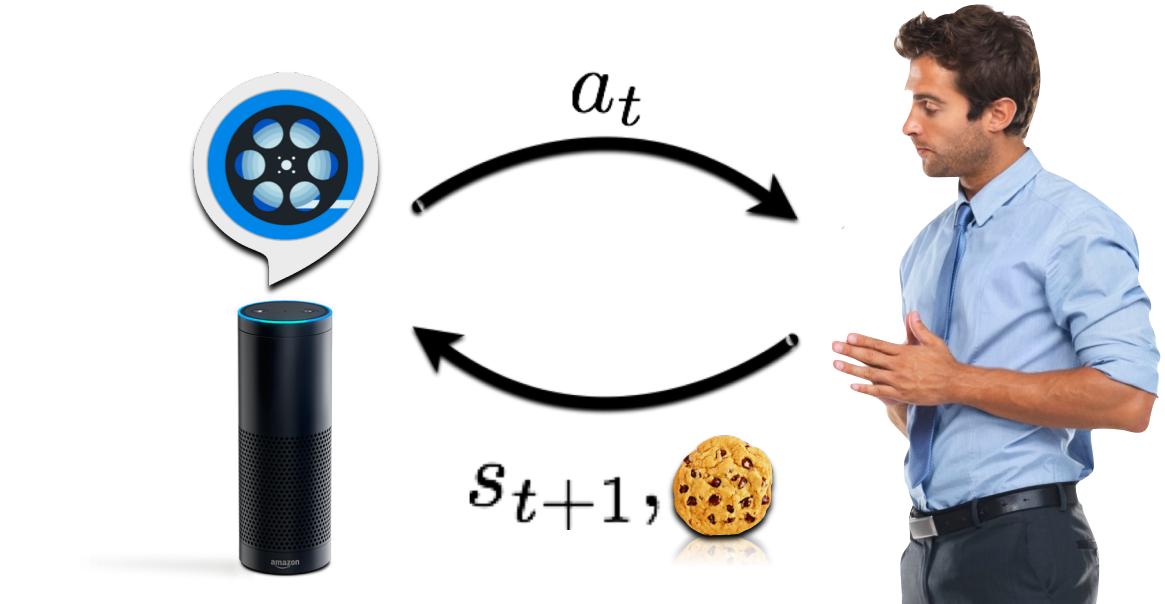


Outline

- Motivation
- Problem Formulation
- Conversational Error Recovery
- Summary and Conclusion

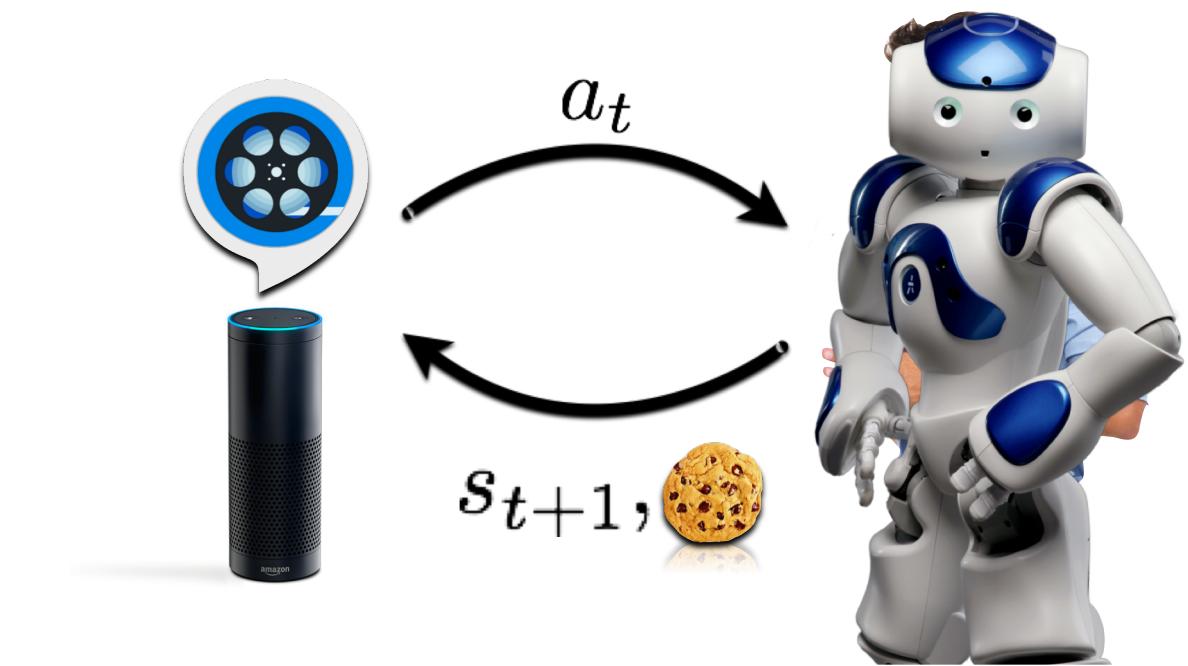
Chicken and Egg Problem

- Need a **good policy** to appeal to the user to collect data
- Need **data** to train a good policy

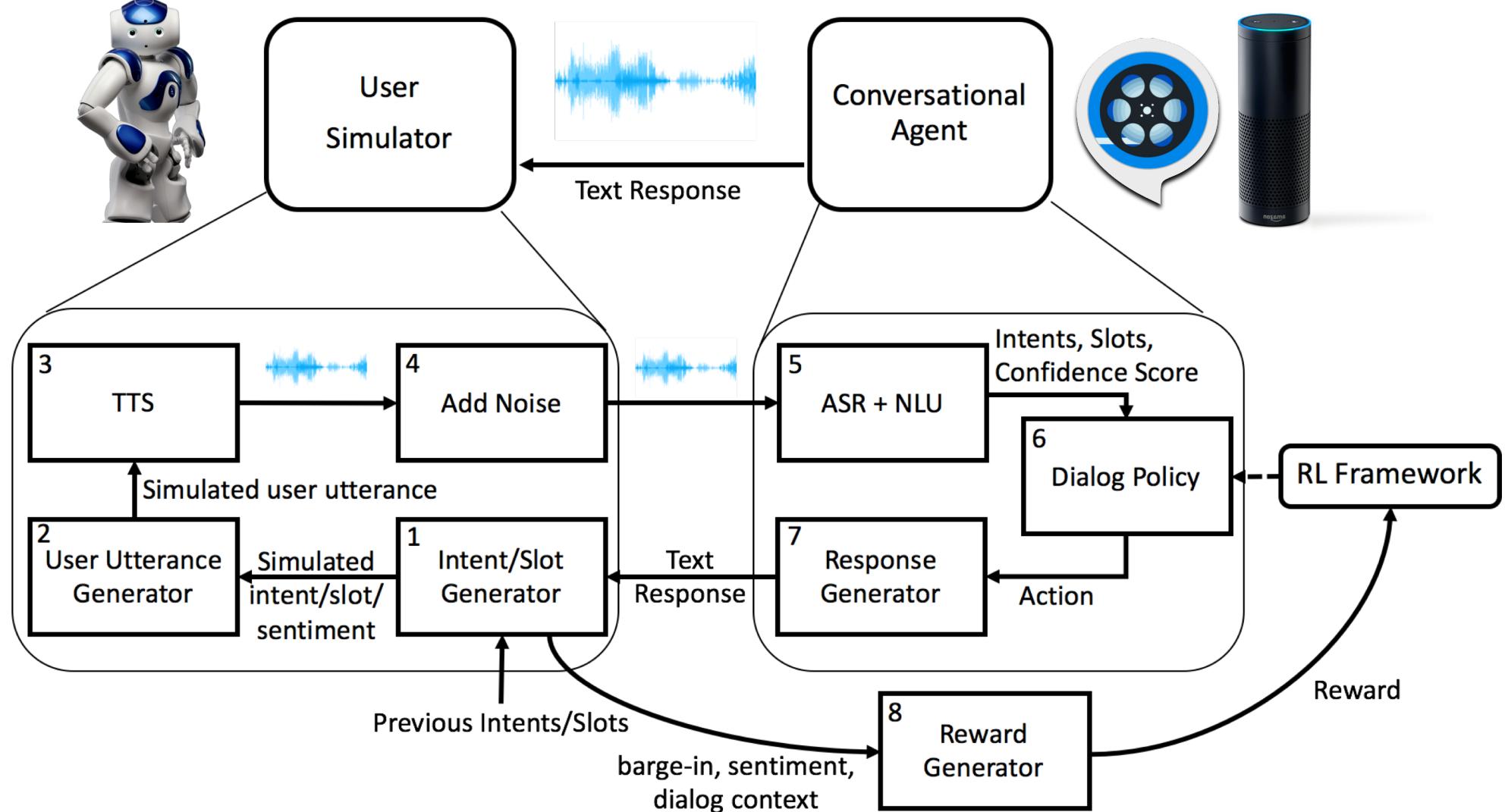


Chicken and Egg Problem

- Need a **good policy** to appeal to the user to collect data
- Need **data** to train a good policy



Simulated User – Agent Interaction





Is it a good user simulator?

- Confusability with real subjects

Number of participants	78
Number of dialogs	780
Human identified as human	58.97%
Simulator identified as human	67.18%



Is it a good user simulator?

- Confusability with real subjects

Criteria for Selecting Human	Criteria for Selecting Simulator
Expression of emotions	Simple or formal conversation
Logical progression of the dialog	Asking a question without context
Goals that humans have regardless of agent's response	Disconnected topics
Naturalness of the dialog	Random responses by the user

Now to the RL Part

- **State:** {Hypothesis intent and slot, ASR and NLU intent/slot confidence scores, the previous agent action}
- **Action:** {execute, confirm, elicit, elicit-intent, elicit-slot}

Do you want movies directed by Christopher Nolan?

Can you repeat?

What do you want to know about Tom Hanks?

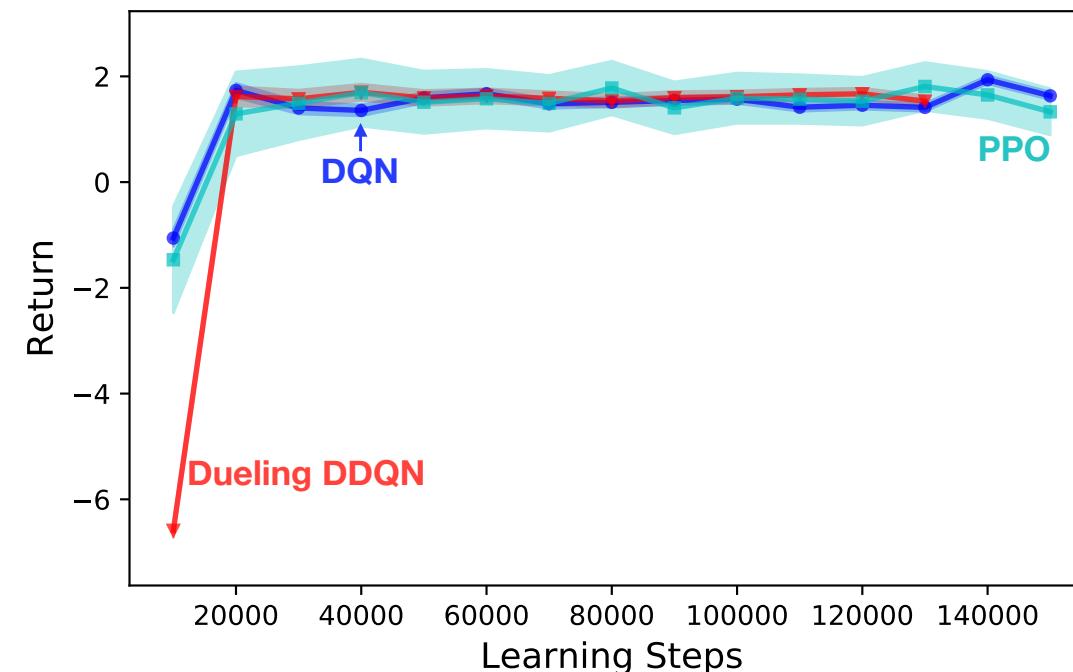
Which movie's plot are you asking for?

Reward Function

action	state	reward
execute(<i>hypothesis</i>)	<i>hypothesis</i> = <i>reference</i>	+1
execute(<i>hypothesis</i>)	<i>hypothesis</i> ≠ <i>reference</i>	-1
confirm	*	-0.33
elicit-intent	*	-0.33
elicit-slot	*	-0.33
elicit	*	-0.50
*	positive sentiment	+0.17
*	negative sentiment	-0.17
*	barge-in	-0.17

RL Agents

- Deep Q-Network (DQN) [Mnih et al. 2013]
- Dueling Double DQN (Dueling DDQN) [Wang et al. 2016]
- Proximal Policy Optimization (PPO) [Schulman et al. 2017]

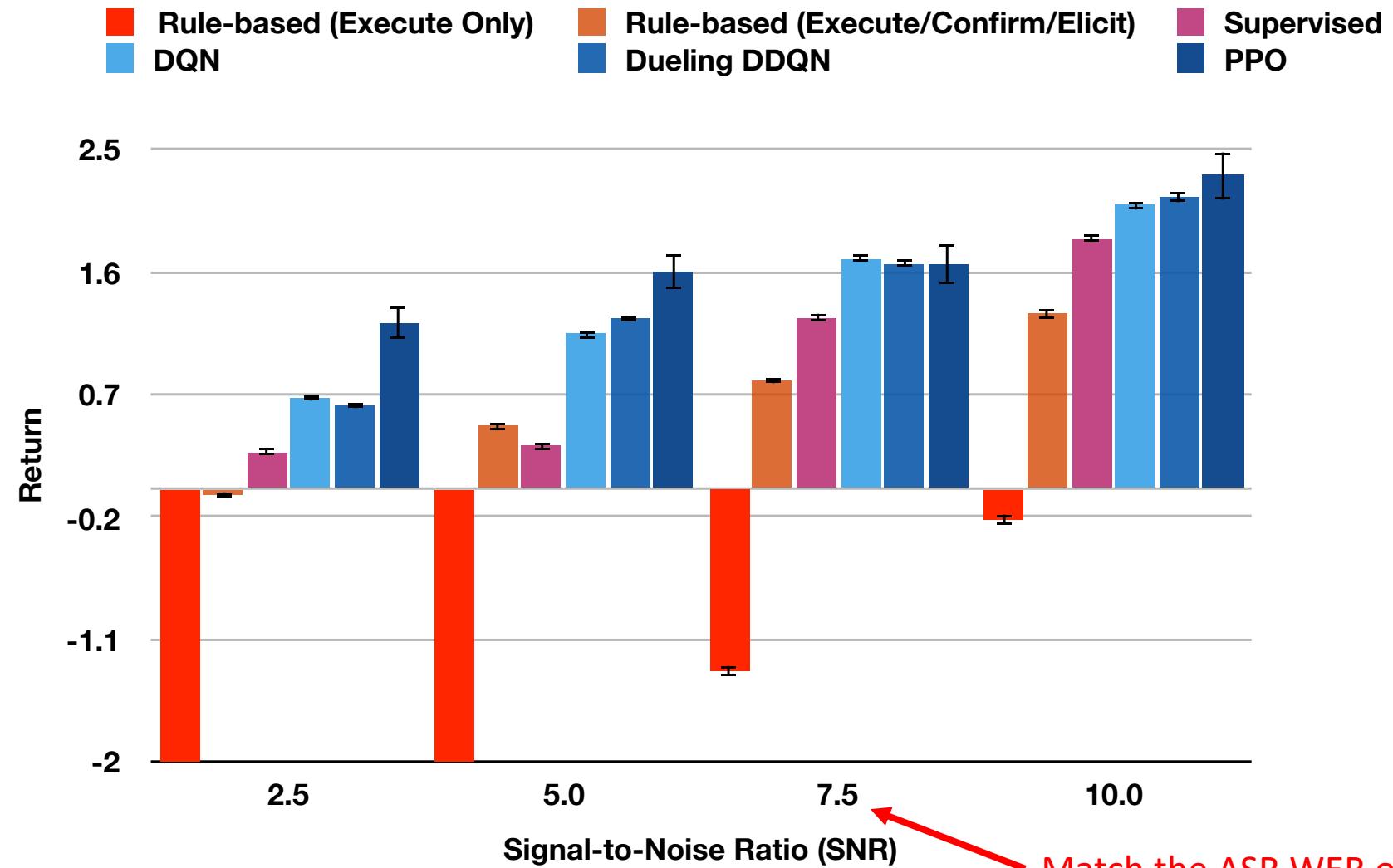




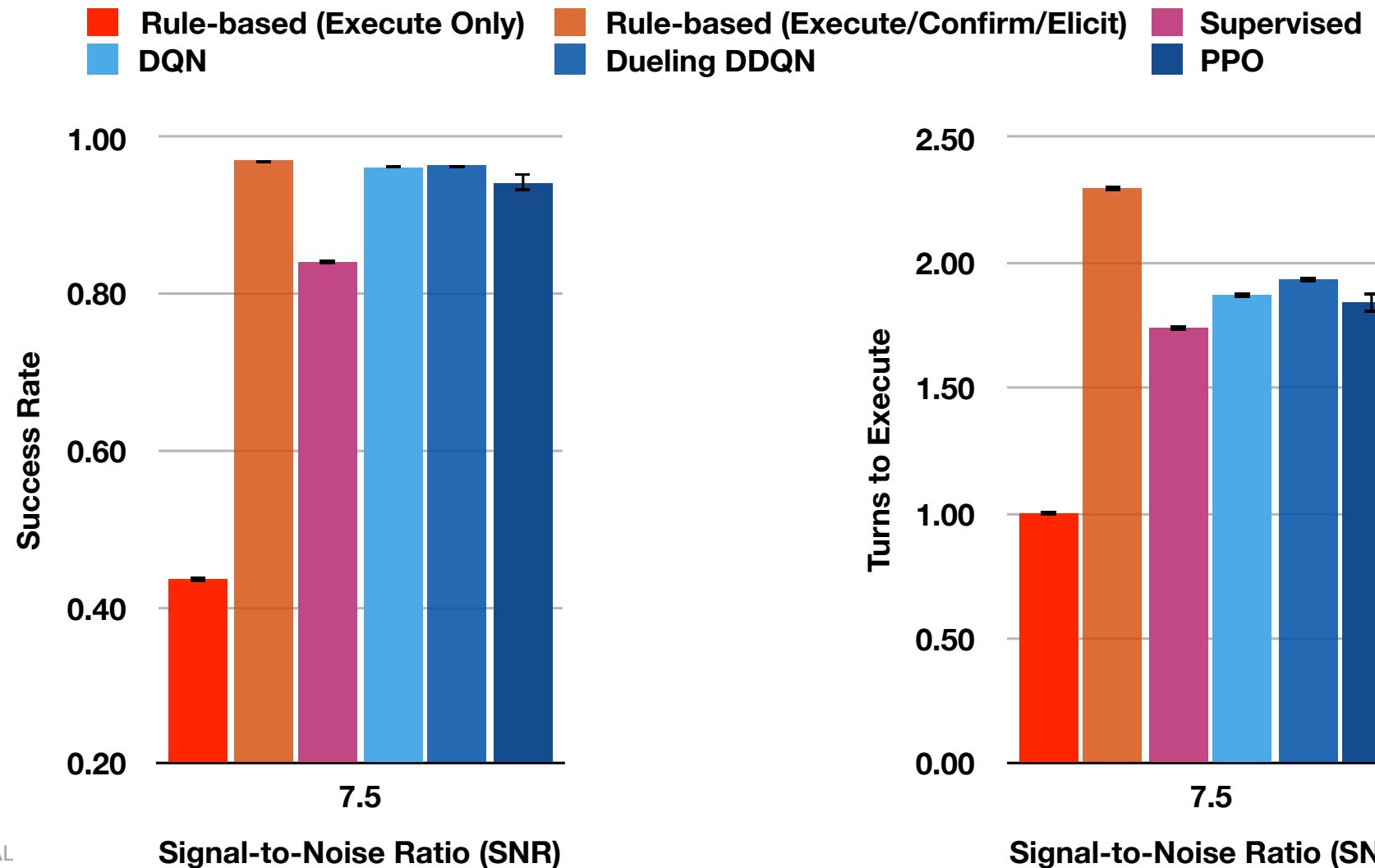
Baseline Agents

- Rule-based Agent 1: Always Execute
- Rule-based Agent 2: Learn optimal ASR/NLU thresholds
- Supervised Agents:
 - Random Forest, Gradient Boosting, Support Vector Machine (SVM), Multi-layer Perceptron (MLP)

RL Methods Outperform Rule-Based and Supervised Policies



Higher/Same Success Rate, Fewer Turns per Execution





Outline

- Motivation
- Problem Formulation
- Conversational Error Recovery
- Summary and Conclusion



Summary and Conclusion

- Developed a user simulator which mimics realistic dialogs with noisy audio
- RL techniques achieved the same success rate with 35% fewer clarification questions relatively compared to the best rule based technique

Summary and Conclusion

- Ongoing and future work:
 - Other clarification policies: disambiguation, search-based elicitation
 - Working with domain teams (MovieBot, Communications, Conversational Discovery) to deliver the learned policies to production
 - Evaluating the reward automatically by estimating user satisfaction
- How can we help you?
 - User simulator is easily adaptable to new domains
 - RL framework





**Tutorial: Introduction to Markov Decision Processes
and Reinforcement Learning (Friday 1:15-3:00PM)**

Thank You!

alborzg-dialog@amazon.com

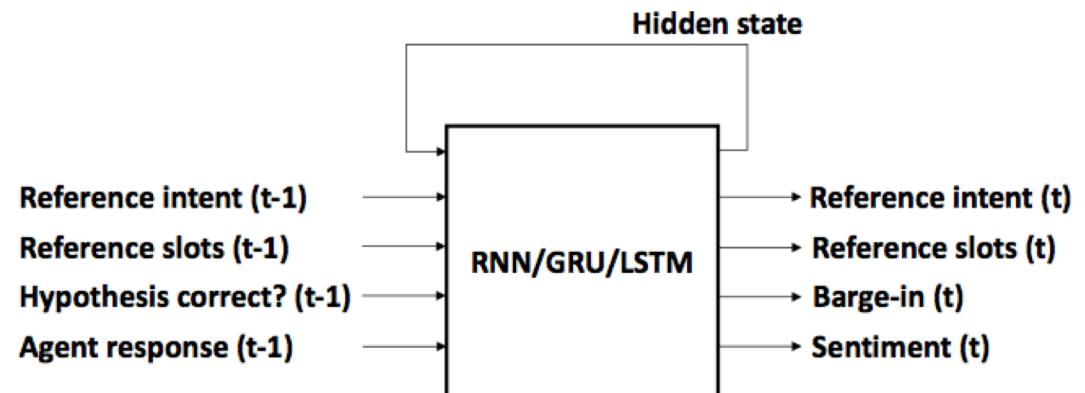




Backup Slides

Intent/Slot Generator

- ⌚ Data:
 - ⌚ ~1.4K dialogs, > 10K user turns
 - ⌚ 56 intents, 13 slot types
- ⌚ Methods:
 - ⌚ RNN: Vanilla, Gated Recurrent Unit (GRU), Long Short Term Memory (LSTM)
 - ⌚ Bigram: Previous Intent and Slot





Utterance Generator

- Uniformly sample one utterance from existing templates using the generated intent/slot/sentiment

Noise Addition

Contaminate the synthesized output with white noise by setting the desired signal-to-noise ratio (SNR)

Example of User Simulated Dialogs

User ummm comedy
MovieBot I think you might like Cars 3, a comedy adventure film, starring Owen Wilson.

User How long is the movie?

MovieBot The movie duration is 1 hour and 42 minutes.

User Who wrote the soundtrack?
MovieBot I don't know how to handle that yet. You can also say, can children watch this movie.

User What is a popular romantic comedy movie?
MovieBot I'd recommend Blind, a romance drama film starring Demi Moore.

User Tell me about it.
MovieBot Here is a summary of blind from IMDb.
gives plot summary

User That is enough.

MovieBot Thank you for using MovieBot.

User Science fiction
MovieBot I think you might like Geostorm, a sci-fi action film., starring Gerard Butler.

User Repeat that.
MovieBot I've sent a card to your Alexa App to help you remember the title.

User Are you listening to me?
MovieBot I'm sorry, there was an error, please try again or you can ask me something else.

User Thank you, MovieBot stop.
MovieBot Goodbye.

RL Agents

- Deep Q-Network (DQN) [Mnih et al. 2013]
- Dueling Double DQN (DDQN) [Wang et al. 2016]
- Proximal Policy Optimization (PPO) [Schulman et al. 2017]

Setup

- ϵ -greedy ($1 \mapsto 0.1$)
- 150K interactions
- $\gamma = 0.97$
- Shared random seeds across methods
- Optimized parameters using HyperOpt
- Evaluated each method after 10K with $\epsilon=0$

Rule-based Agents

- Rule-based Policy 1: Always Execute
- Rule-based Policy 2: Learn optimal ASR/NLU thresholds

a	s
execute(hyp)	$ASR_score \geq 0.76 \text{ AND } NLU_intent_score \geq 0.91 \text{ AND } NLU_slot_score \geq 0.77$
elicit	$ASR_score < 0.75 \text{ OR } (NLU_intent_score < 0.05 \text{ AND } NLU_slot_score < 0.71)$ OR user responded no to <i>confirm</i>
elicit-intent	$NLU_intent_score < 0.05 \text{ AND } NLU_slot_score \geq 0.77$
elicit-slot	$NLU_intent_score \geq 0.91 \text{ AND } NLU_slot_score < 0.71$
confirm	Otherwise

Supervised Agents

- Random Forest, Gradient Boosting, Support Vector Machine (SVM), Multi-layer Perceptron (MLP)
- Setup
 - 10K dialog episodes with simulator
 - Manually labeled action for each turn in each dialog episode
 - 5-fold cross validation

	Weighted Precision	Weighted Recall	Weighted F1-Score
Random Forest	0.798	0.819	0.804
Gradient Boosting	0.805	0.825	0.809
SVM	0.773	0.801	0.776
MLP	0.789	0.810	0.784