Speech, speech recognition, and dialog state tracking in spoken dialog systems

(and maybe situated language systems too)

Jason D. Williams
Microsoft Research

Quick intro

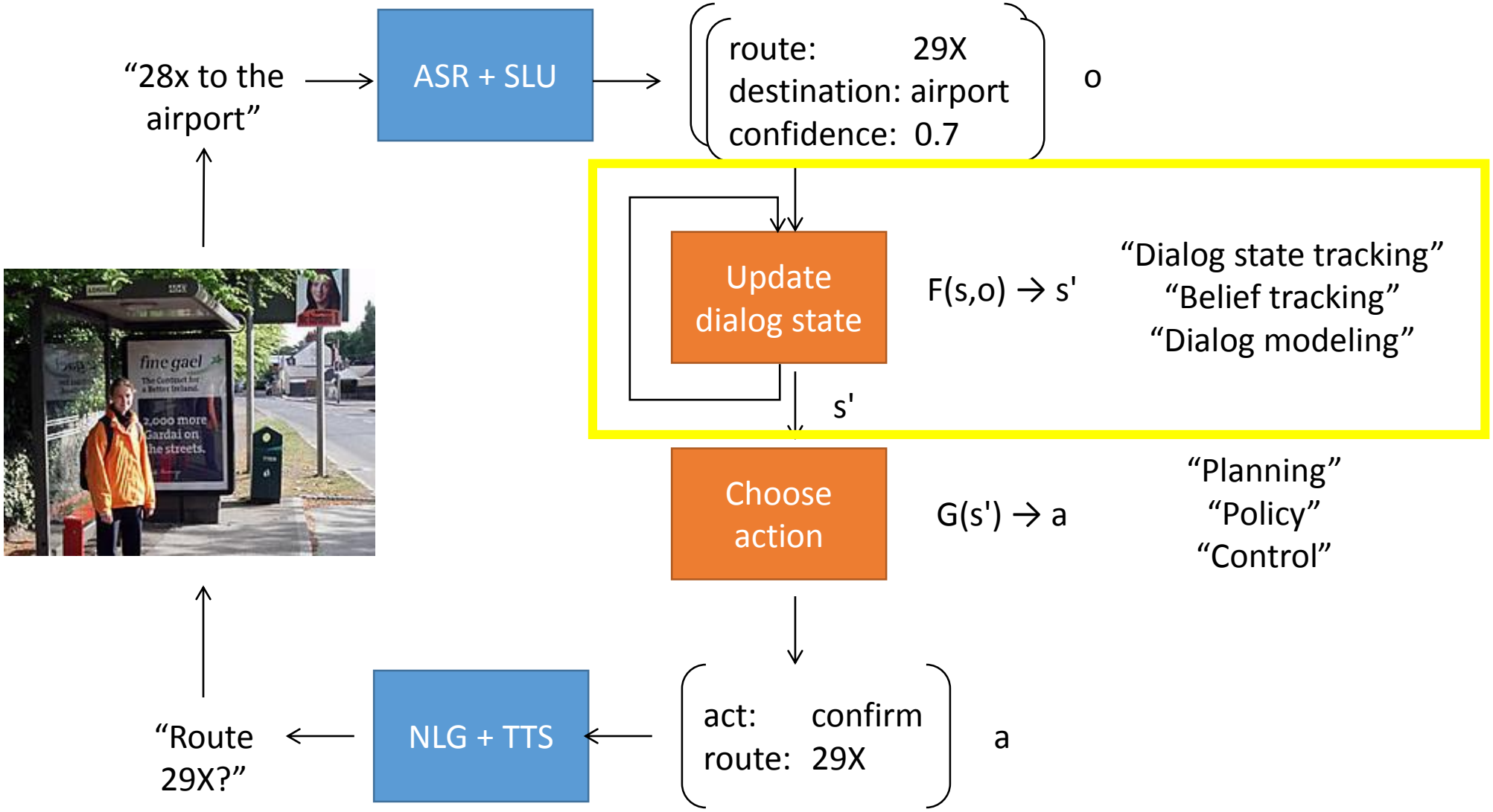
- After undergrad & masters…
- Worked for 3 years building spoken dialog systems in industry
  - Ran 100 usability tests
  - Listened to 1000s of calls
- PhD in statistical approaches to spoken dialog systems
- AT&T Research for 6 years
  - Also deployed 2 large-scale systems
- Now at MSR (1.5 years)

Aims for today

- Introduce spoken dialog systems
  - Emphasis on issues introduced by automatic speech recognition (ASR)
  - Directed at practitioners, not ASR researchers

- Present some techniques for state tracking in dialog systems
  - Improve robustness to ASR errors
  - And may also open doors to collaboration with the dialog community

Spoken dialog systems



Dialog state tracking

Dialog state tracking is difficult largely because errors in speech recognition and language understanding are error-prone, and render the true state of the dialog partially observable

Responses to "How may I help you?"

- Silences and hesitations while users think
  - Leads to end-pointing problems
  - Leads to users confusing themselves
- "Robot" language (hence examples, "speak naturally")
  - Example 1
  - Example 2
- Recognition errors confused with competences
  - > "i need to sign up for a get off benefit" *[no parse]*
  - > "i would like to enroll in a get one" *[no parse]*
  - > "i would like to get help with my dental insurance" <HELP>
  - > "dental insurance" <INSURANCE>

Source: Live calls, human resources dialog system

ASR/SLU errors are common

Grammar	Yes/no	City & state	How may I help you?

Source: Two different deployed commercial applications running two different speech recognizers

ASR/SLU errors are common

Grammar	Yes/no	City & state	How may I help you?
In-grammar/ in-domain accuracy	99.8%	85.1%	89.5%

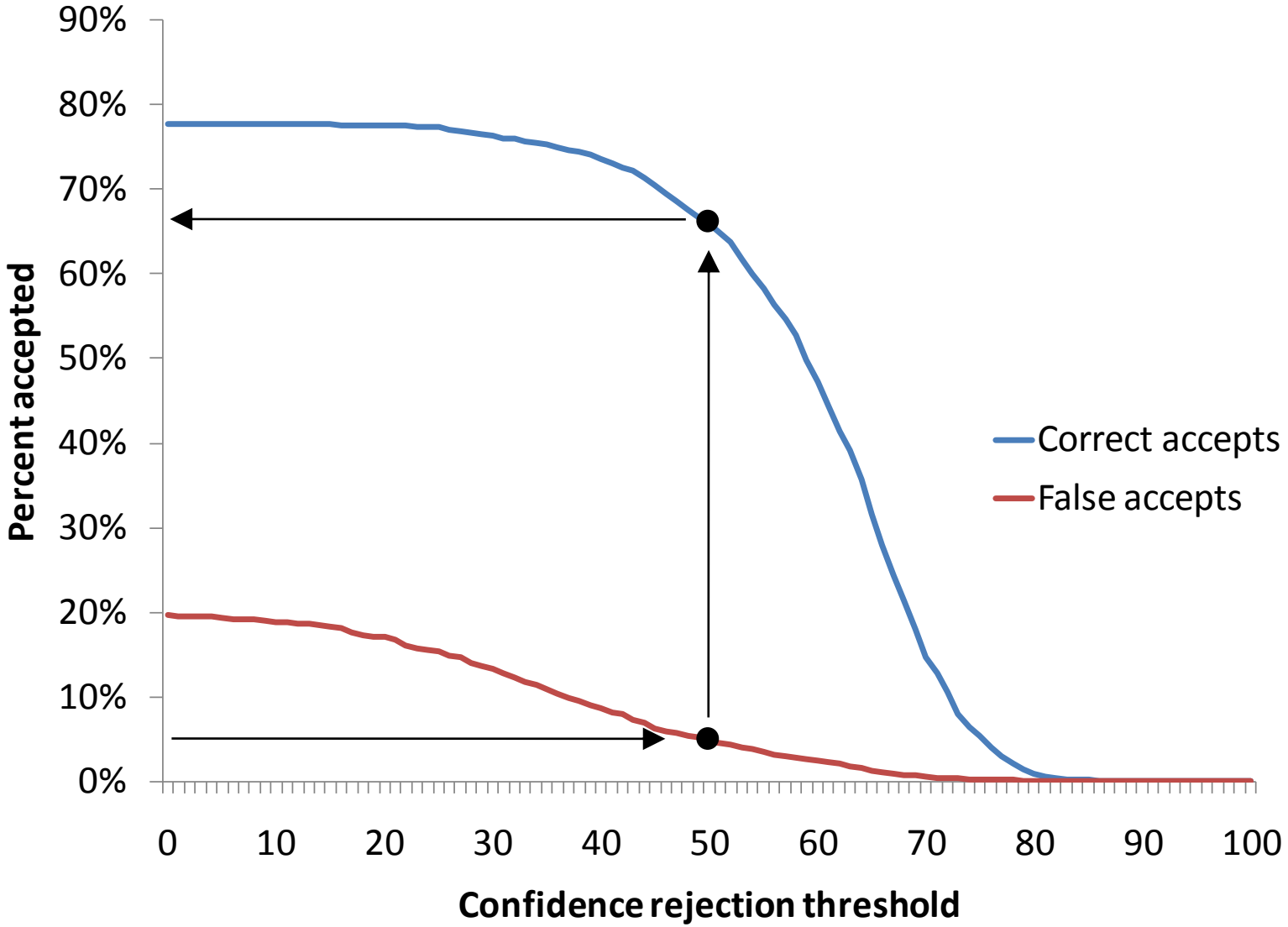
Source: Two different deployed commercial applications running two different speech recognizers

ASR/SLU errors are common

Grammar	Yes/no	City & state	How may I help you?
In-grammar/ in-domain accuracy	99.8%	85.1%	89.5%
% in-grammar/ in-domain	92.3%	91.0%	86.8%
Overall accuracy	92.1%	77.6%	77.7%

Source: Two different deployed commercial applications running two different speech recognizers

ASR errors are hard to detect



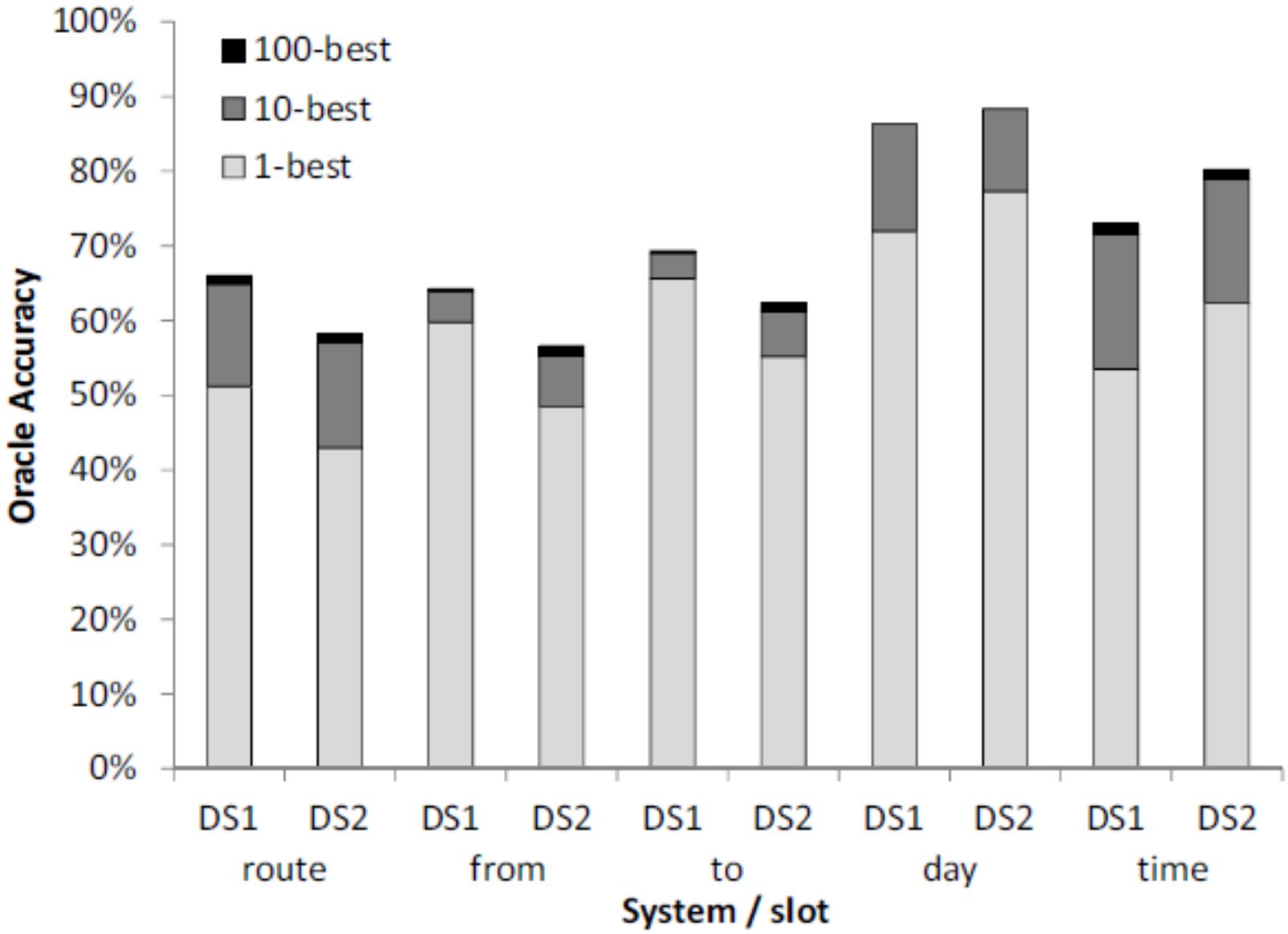
Source: Commercial deployed spoken dialog system

ASR/SLU errors are common

Grammar	Yes/no	City & state	How may I help you?
In-grammar/ in-domain accuracy	99.8%	85.1%	89.5%
% in-grammar/ in-domain	92.3%	91.0%	86.8%
Overall accuracy	92.1%	77.6%	77.7%
Accepted utts (False accepts)	89.6% (1.8%)	60.3% (4.9%)	73.3% (8.3%)

Source: Two different deployed commercial applications running two different speech recognizers

There is useful information on the N-Best list



Jason D. Williams, Challenges and Opportunities for State Tracking in Statistical Spoken Dialog Systems: Results From Two Public Deployments, in IEEE Journal of Selected Topics in Signal Processing, vol. 6, no. 8, pp. 959-970, IEEE SPS, December 2012

Users really care about recognition accuracy

- 3 spoken dialog systems: voice dialing and messaging, email, train schedules
- User satisfaction scores at end of each dialog
- Build regression models that predict user satisfaction as a function of dialog properties – main 3 are:
  - Speed, task completion, recognition accuracy

- Across all 3 systems, the results were the same:
  - **Most important: Reco accuracy**
  - Second more important: task completion
  - Distant third: speed

M Walker, C Kamm, D Litman. Towards developing general models of usability with PARADISE. 2000. Natural Language Engineering, vol 6, issue 3&4, pp 363-377.

Example: Rule-based state tracker

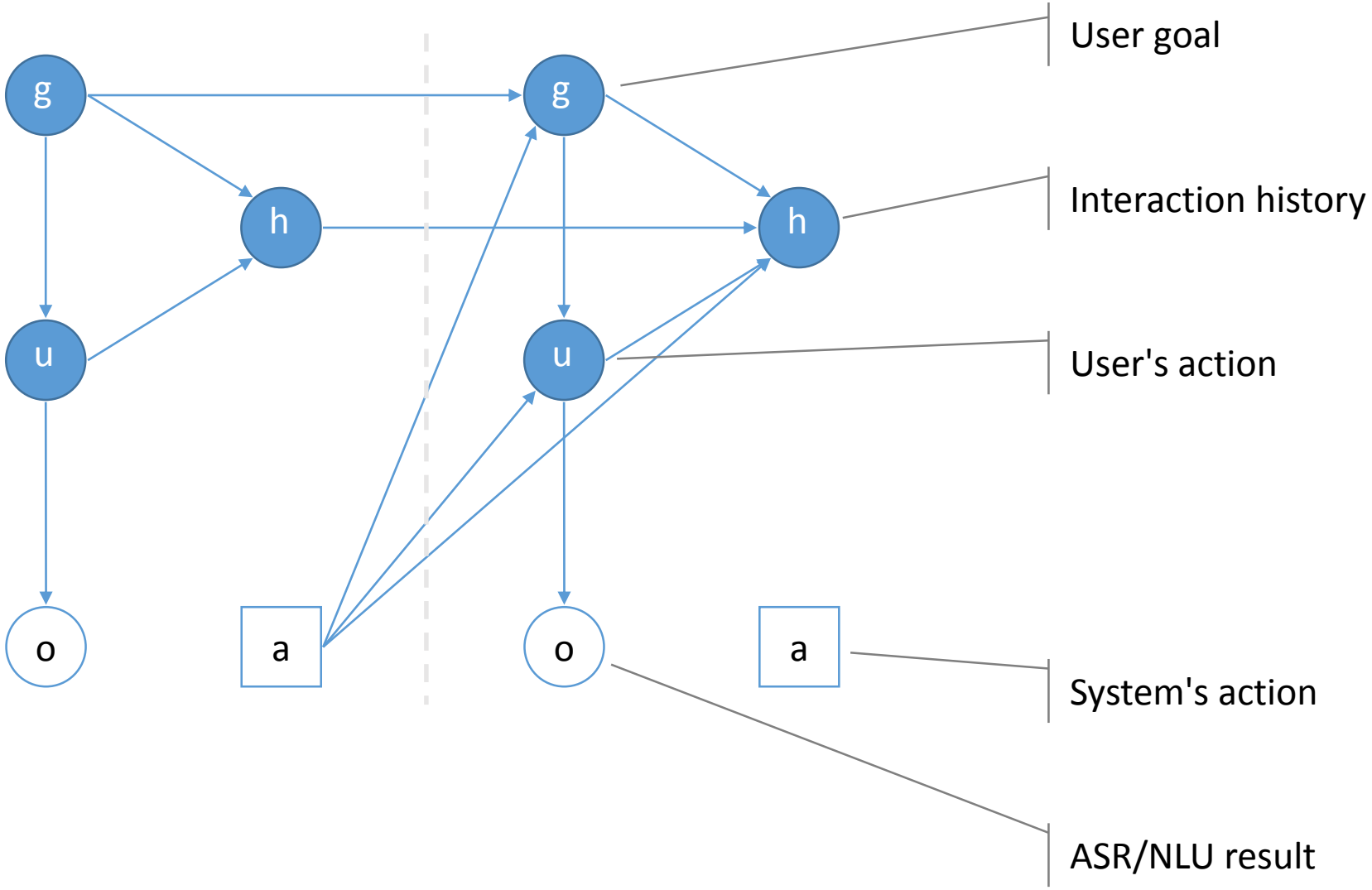
- In commercial systems, dialog state tracking is most commonly done with hand-coded rules
- Example:

```
if ($slu[0].score > 0.8):
  // keep SLU result
  $state.bus_route = $slu[0].bus_route
else:
  // discard SLU result
```

Can we do better?

- Toeholds:
  - Rules usually make decisions are made locally, but turns are correlated
  - N-Best lists contain information that most rules ignore
  - There are useful priors that rules usually don’t use
- Intuition
  - Cast state as a hidden variable and evidence as noisy emission from that state
  - View dialog state estimation as a tracking or filtering problem
  - Similar to particle filters or HMMs
- Methods
  - Generative models: partitions, bayes nets, more
  - Discriminative: maxent, SVMs, ANNs, more
  - More: heuristics

Tracking a distribution over interaction states



Generative approach

$$\underbrace{b'(g',h')}_{\text{new belief state}} = \underbrace{\eta}_{\text{normalizing constant}} \cdot \sum_{u'} \underbrace{P(o' \mid u',a)}_{\text{ASR model}} \sum_{h} \underbrace{P(u' \mid g',h,a)}_{\text{user action model}} \underbrace{P(h' \mid u',g',h,a)}_{\text{interaction history model}} \sum_{g} \underbrace{P(g' \mid g,a)}_{\text{user goal model}} \underbrace{b(g,h)}_{\text{old belief state}}$$

- Synthesizes:
  - Current belief over user's goal *b(g)*
  - Model of how user goal changes
  - Model of user producing utterances
  - ASR/SLU confidence

Jason D. Williams and Steve Young, Partially Observable Markov Decision Processes for Spoken Dialog Systems, in Computer Speech and Language, vol. 21, no. 2, pp. 393-422, 2007

AT&T Let's Go Call Visualizer

Call in /home/jdw/run/prod/logs/2010/07/23/default/0031/dt-jdw-201007230031-240B1

<< < 4 go > >>

to airport. Is that right?

ASR Result (play)	Origin	Destination
yes that's correct	carnegie mellon in oakland; carnegie mellon university...	to airport; airport; to the airport; the airport
-	-	-
-	-	-
-	-	-
-	-	-

Bus Route	Date	Time
twenty eight x; the twenty eight x	-	-
-	-	-
-	-	-
-	-	-
-	-	-

Do you want times for the next few buses? Say yes or no.

Jason Williams AT&T Labs - Research jdw@research.att.com www.research.att.com/people/Williams_Jason_D

Why aren't more people working on this?

- Building dialog systems is expensive -- difficult to obtain dialog data
- Can this be cast as a challenge task?

http://research.microsoft.com/en-us/events/dstc/

- 15K dialogs with real users
- Transcriptions, labels
- Common log format, with user and system semantics
- 3 totally different dialog systems
- Evaluation tools; baselines; template code

Jason D. Williams, Antoine Raux, Deepak Ramachandran, Alan Black. The Dialog State Tracking Challenge. SigDial 2013.

Dialog state tracking challenge data sets

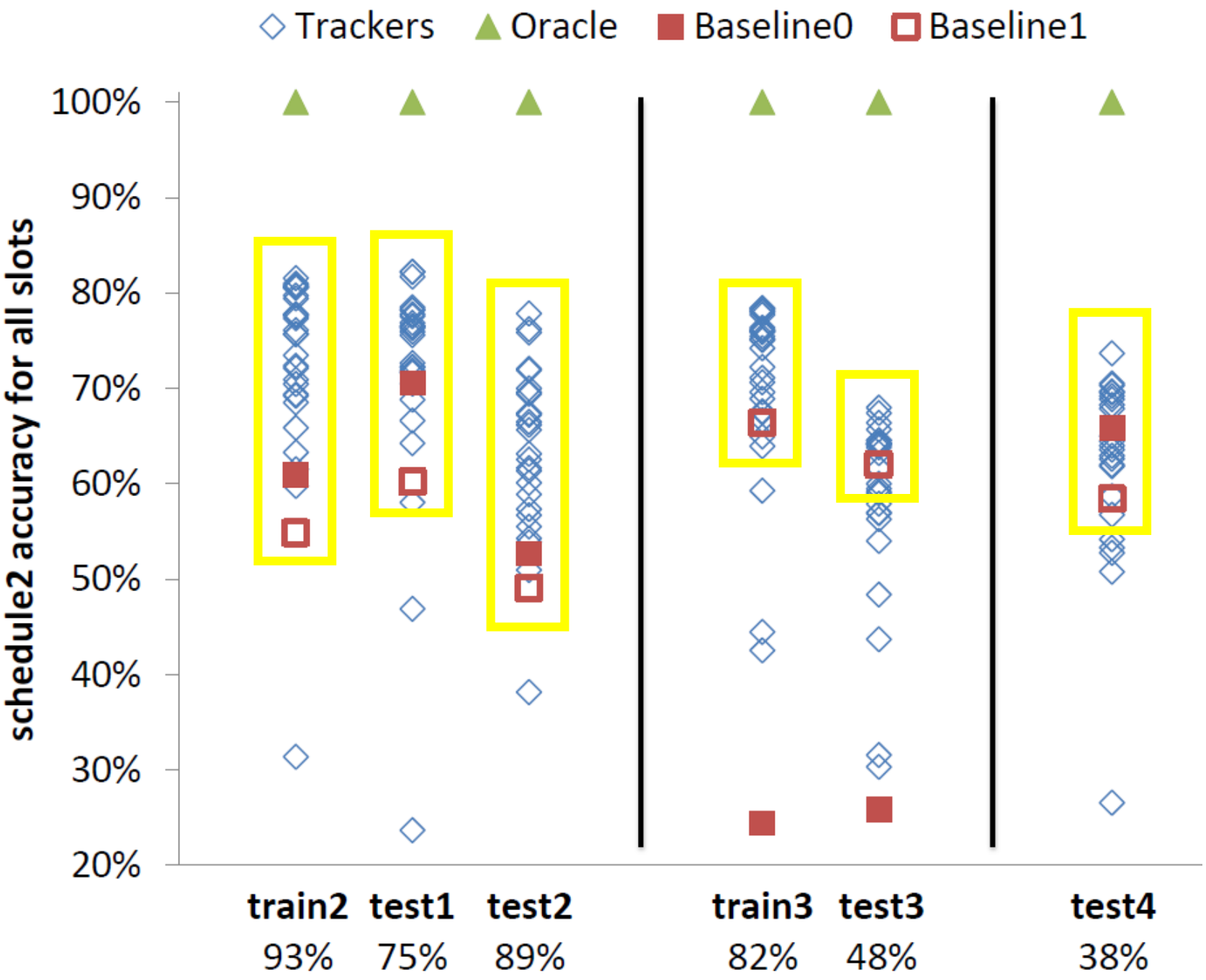
	Train					Test			
	1A	1B	1C	2	3	1	2	3	4
Group	A	A	A	A	B	A	A	B	C
Year(s)	2009	2009	2009	2010	2010	2011	2012	2011-2	2010
Dialogs	1013	1117	9502	643	688	715	750	1020	438
Turns/Dialog	14.7	13.3	14.5	14.5	12.6	14.1	14.5	13.0	10.9
Sys acts/turn	4.0	3.8	3.8	4.0	8.4	2.8	3.2	8.2	4.6
Av N-best len	21.7	22.3	21.9	22.4	2.9	21.2	20.5	5.0	3.2
Acts/N-best hyp	2.2	2.2	2.2	2.3	1.0	2.1	2.0	1.0	1.6
Slots/turn	44.0	46.5	45.6	49.0	2.1	41.4	36.9	4.3	3.5
Transcribed?	yes	yes	yes	yes	yes	yes	yes	yes	yes
Labelled?	yes	no	no	yes	yes	yes	yes	yes	yes
1-best WER	42.9%	41.1%	42.1%	58.2%	40.5%	57.9%	62.1%	48.1%	55.6%
1-best SLU Prec.	0.356	-	-	0.303	0.560	0.252	0.275	0.470	0.334
1-best SLU Recall	0.522	-	-	0.388	0.650	0.362	0.393	0.515	0.376
N-best SLU Recall	0.577	-	-	0.485	0.738	0.456	0.492	0.634	0.413

Results

9 teams participated (27 entries)

Graph shows accuracy of top-ranked dialog state

Results at SigDial (August, Metz, France – roughly co-located with Interspeech)

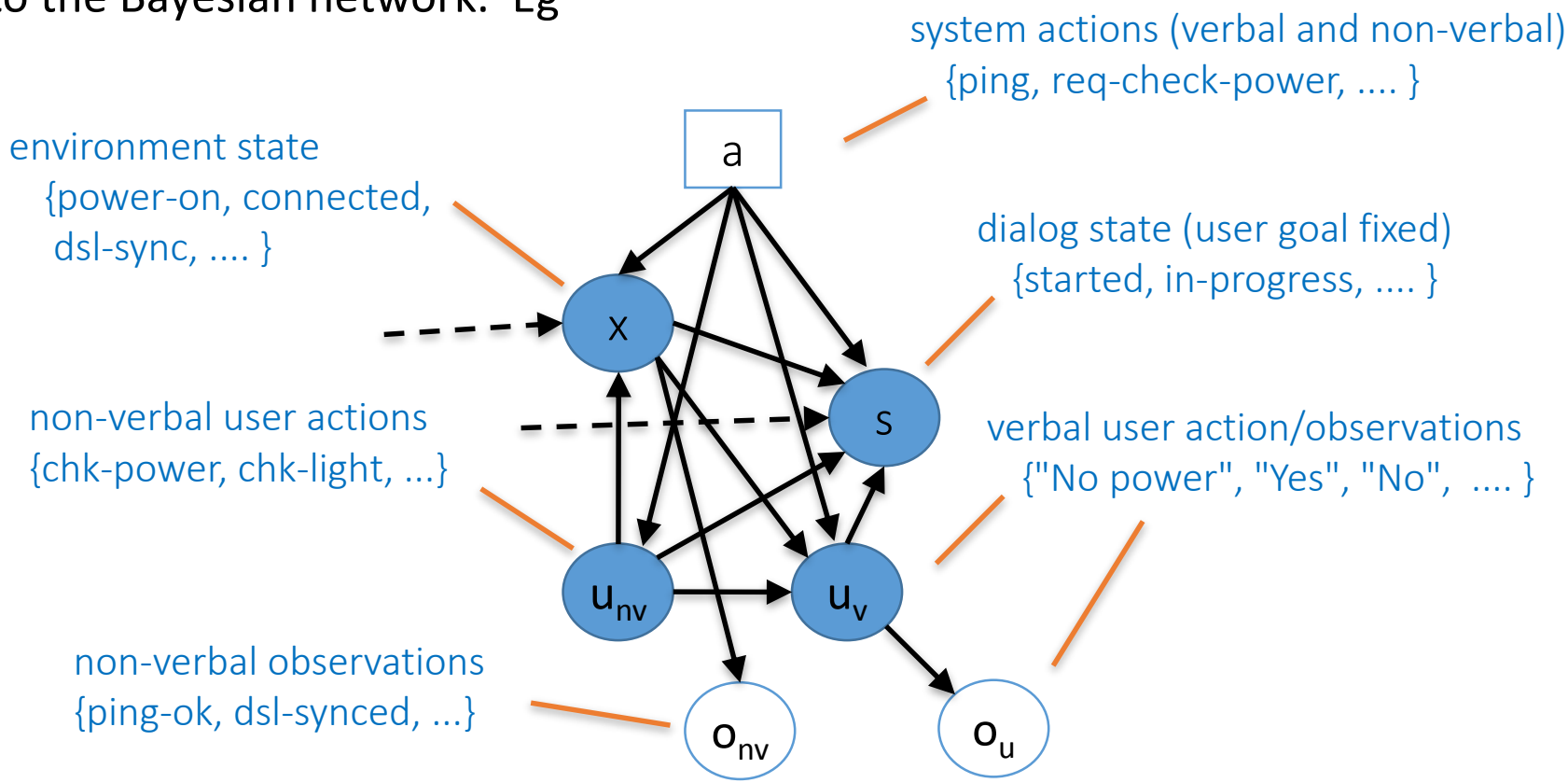


Integration point for situated language

- Observations can account for more than just speech sensor
  - Vision
  - Object attributes
  - Distance
- Hidden variable can account for more state than just user's goal
  - State of objects in a world
  - State of a robot in a world
  - State of currently inferred semantics about the world
- An instantiated example of this: troubleshooting dialog system

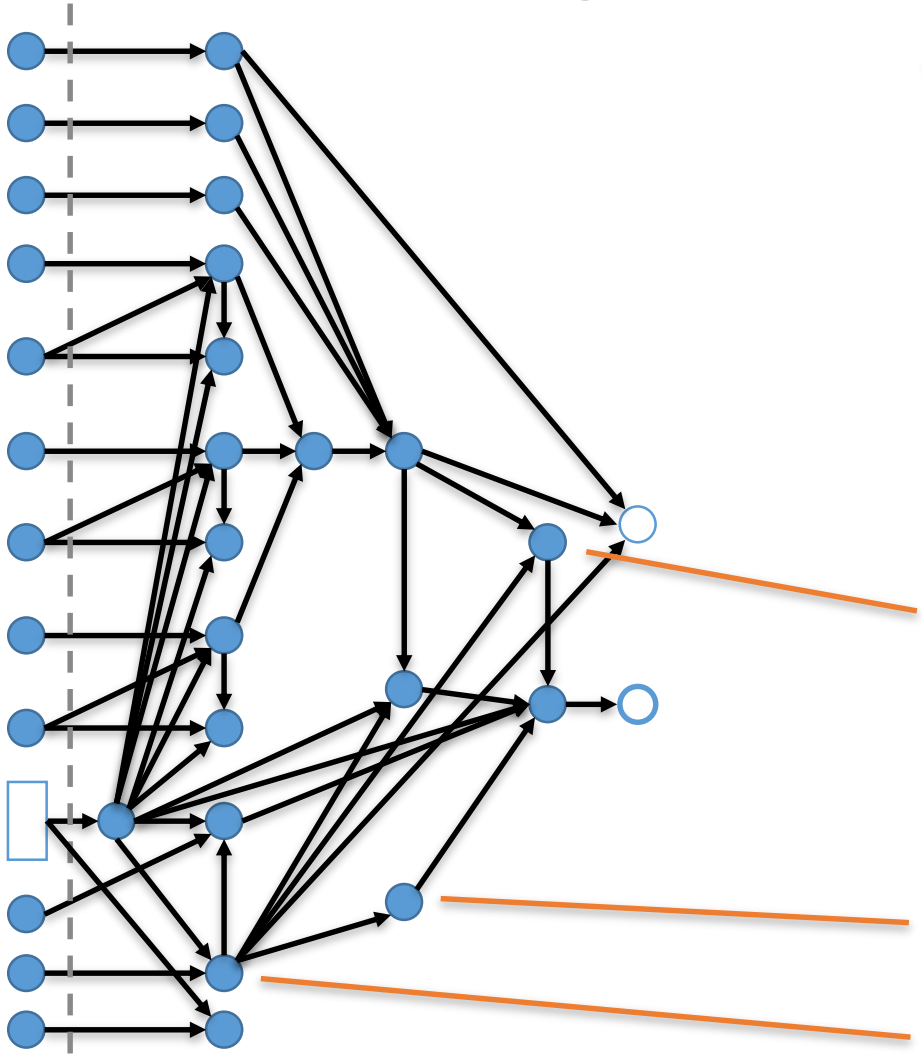
Troubleshooting spoken dialog system

In some applications such as help-lines for DSL modem faults, there are additional sources of uncertainty. These can be easily incorporated into the Bayesian network. Eg



Applying POMDPs to dialog systems in the troubleshooting domain, Williams (Proc W'Shop Bridging the Gap, ACL, 2007)

DSL troubleshooting SDS as a Bayesian network

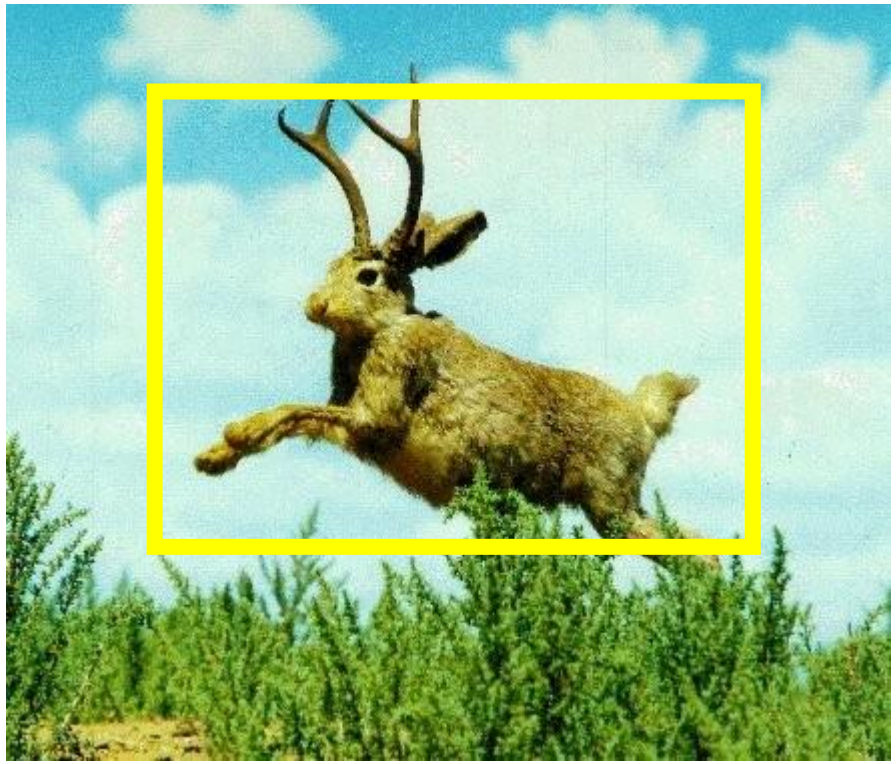


Hidden state components

- Service outage
- Upstream network failure
- Unknown, unfixable problem
- Correct username in browser
- Correct username on modem
- Correct password in browser
- Correct password on modem
- Correct service type in browser
- State of modem network light
- Correct service type on modem
- Config screen visible in browser
- Modem configuration is correct
- DSL connection is working
- State of modem power light
- User opened a webpage
- State of DSL modem

DSL troubleshooting
spoken dialog system
demonstration

Ideation: interactive image ID from attributes



Attribute detectors

has-antlers	0.9
has-paws	0.3
jumps	0.8
is-furry	0.7

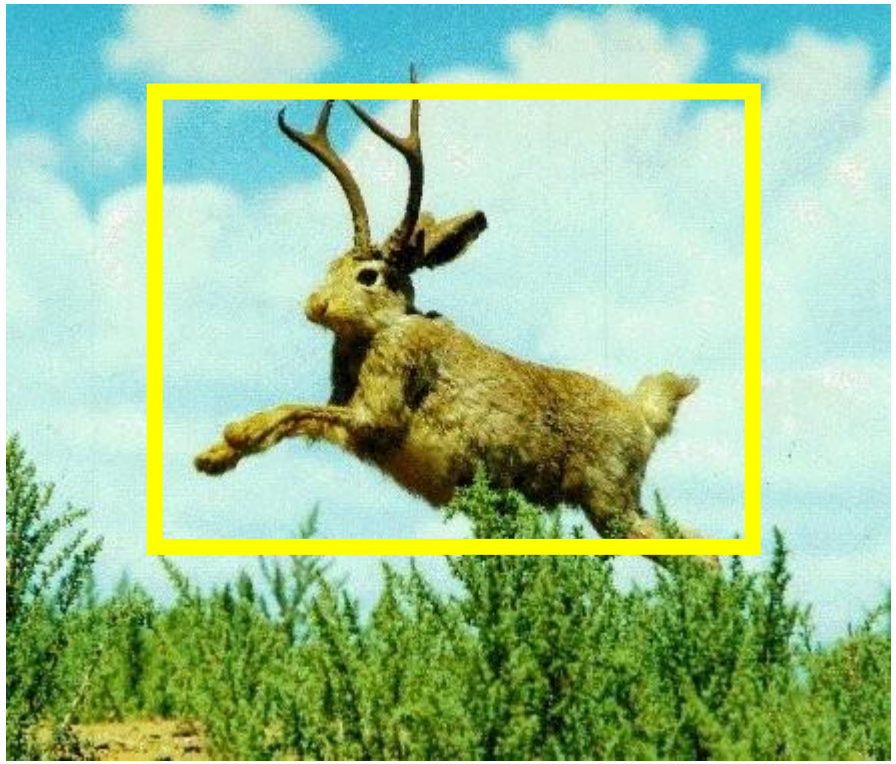
Inferred object class

is-deer	0.4
is-rabbit	0.3
is-moose	0.6

Dialog

S: Is this an animal?
U: Yes ~ 0.9

Ideation: interactive image ID from attributes



Attribute detectors

has-antlers	0.9
has-paws	0.3
jumps	0.8
is-furry	0.7

Inferred object class

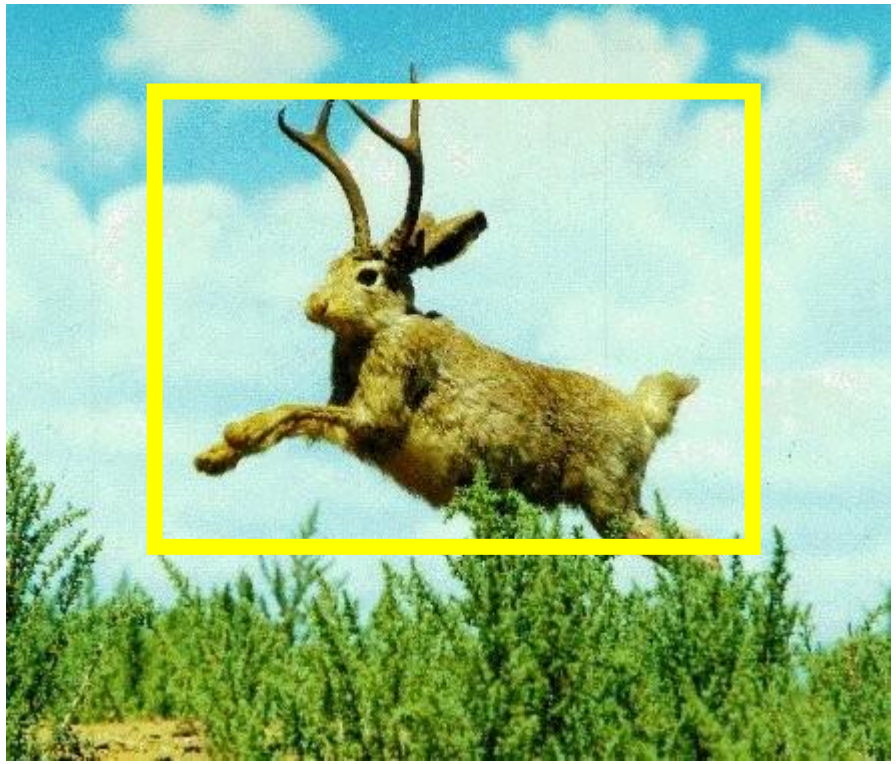
is-deer	**0.5**
is-rabbit	**0.4**
is-moose	**0.7**

Dialog

S: Is this an animal
U: Yes ~ 0.9

S: Is this a moose
U: ha! no way ~ 0.7

Ideation: interactive image ID from attributes



Attribute detectors

has-antlers	0.9
has-paws	0.3
jumps	0.8
is-furry	0.7

Inferred object class

is-deer	**0.6**
is-rabbit	**0.5**
is-moose	**0.1**

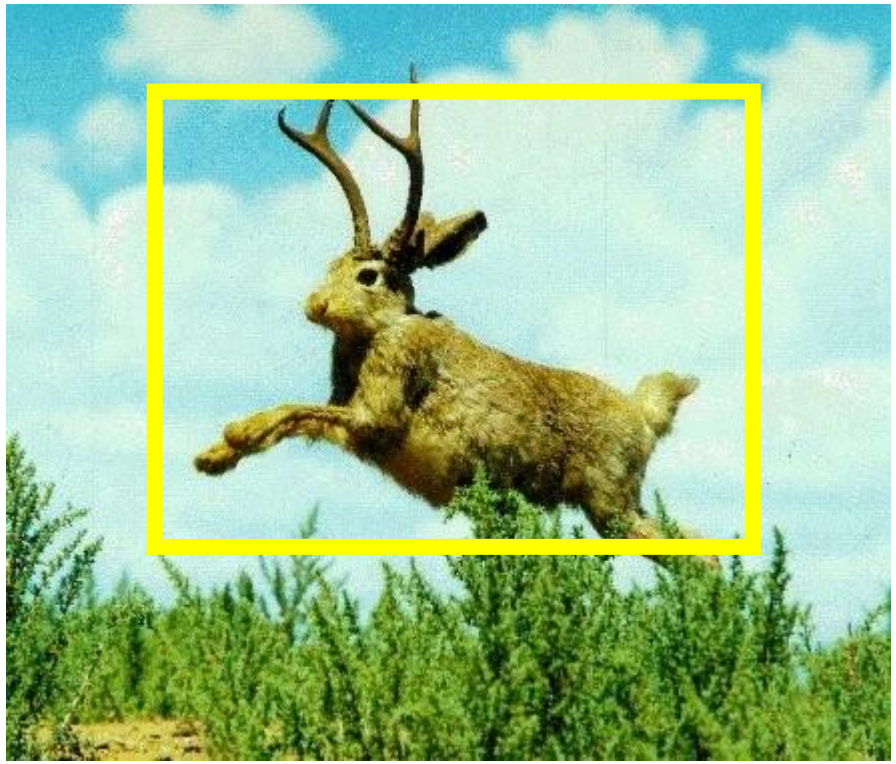
Dialog

S: Is this an animal
U: Yes ~ 0.9

S: Is this a moose
U: ha! no way ~ 0.7

S: Does it have hooves?
U: No ~ 0.9

Ideation: interactive image ID from attributes



Attribute detectors

has-antlers	0.9
has-paws	0.3
jumps	0.8
is-furry	0.7

Inferred object class

is-deer	**0.1**
is-rabbit	**0.7**
is-moose	0.1

Dialog

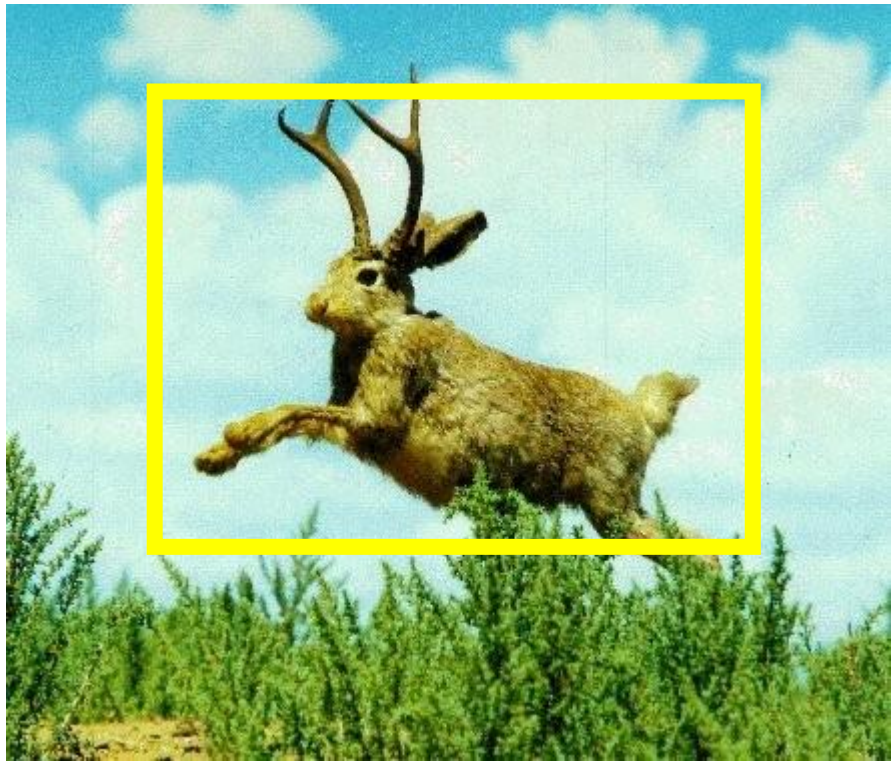
S: Is this an animal?
U: Yes ~ 0.9

S: Is this a moose?
U: ha! no way ~ 0.7

S: Does it have hooves?
U: No ~ 0.9

S: Is this a rabbit?
U: No, much more vicious ~0.6

Ideation: interactive image ID from attributes



Attribute detectors

has-antlers	0.9
has-paws	0.3
jumps	0.8
is-furry	0.7

Inferred object class

is-deer	0.1
is-rabbit	**0.2**
is-moose	0.1

Dialog

S: Is this an animal?
U: Yes ~ 0.9

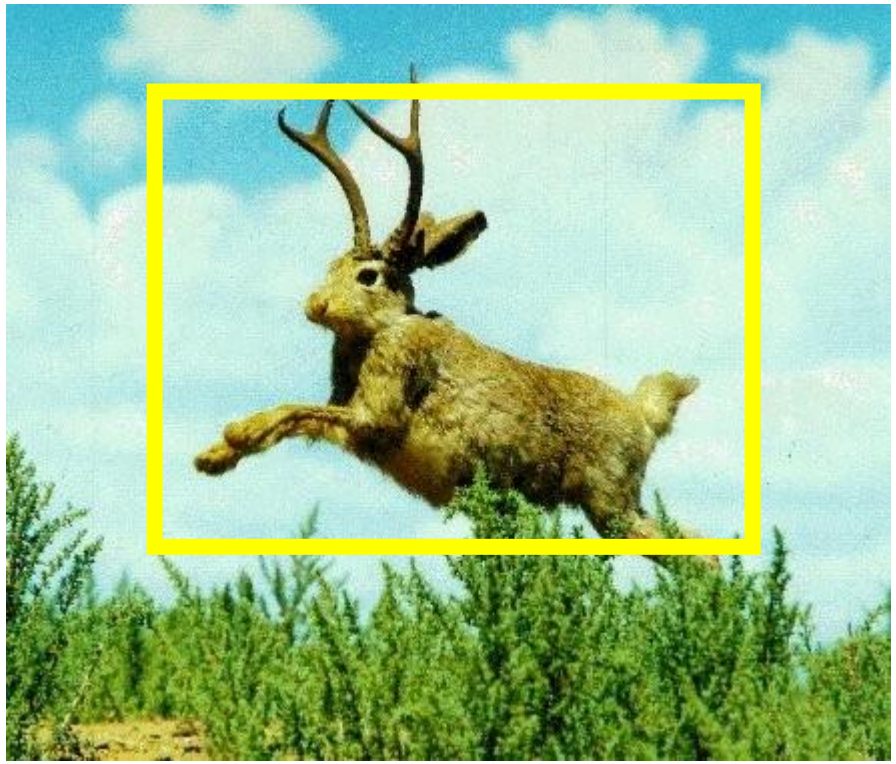
S: Is this a moose?
U: ha! no way ~ 0.7

S: Does it have hooves?
U: No ~ 0.9

S: Is this a rabbit?
U: No, much more vicious ~0.6

S: What kind of animal is it
U: Jackalope! ~ 0.7

Ideation: interactive image ID from attributes



Attribute detectors

has-antlers	0.9
has-paws	0.3
jumps	0.8
is-furry	0.7

Inferred object class

is-deer	0.1
is-rabbit	0.2
is-moose	0.1
is-jackalope	**0.8**

Dialog

S: Is this an animal?
U: Yes ~ 0.9

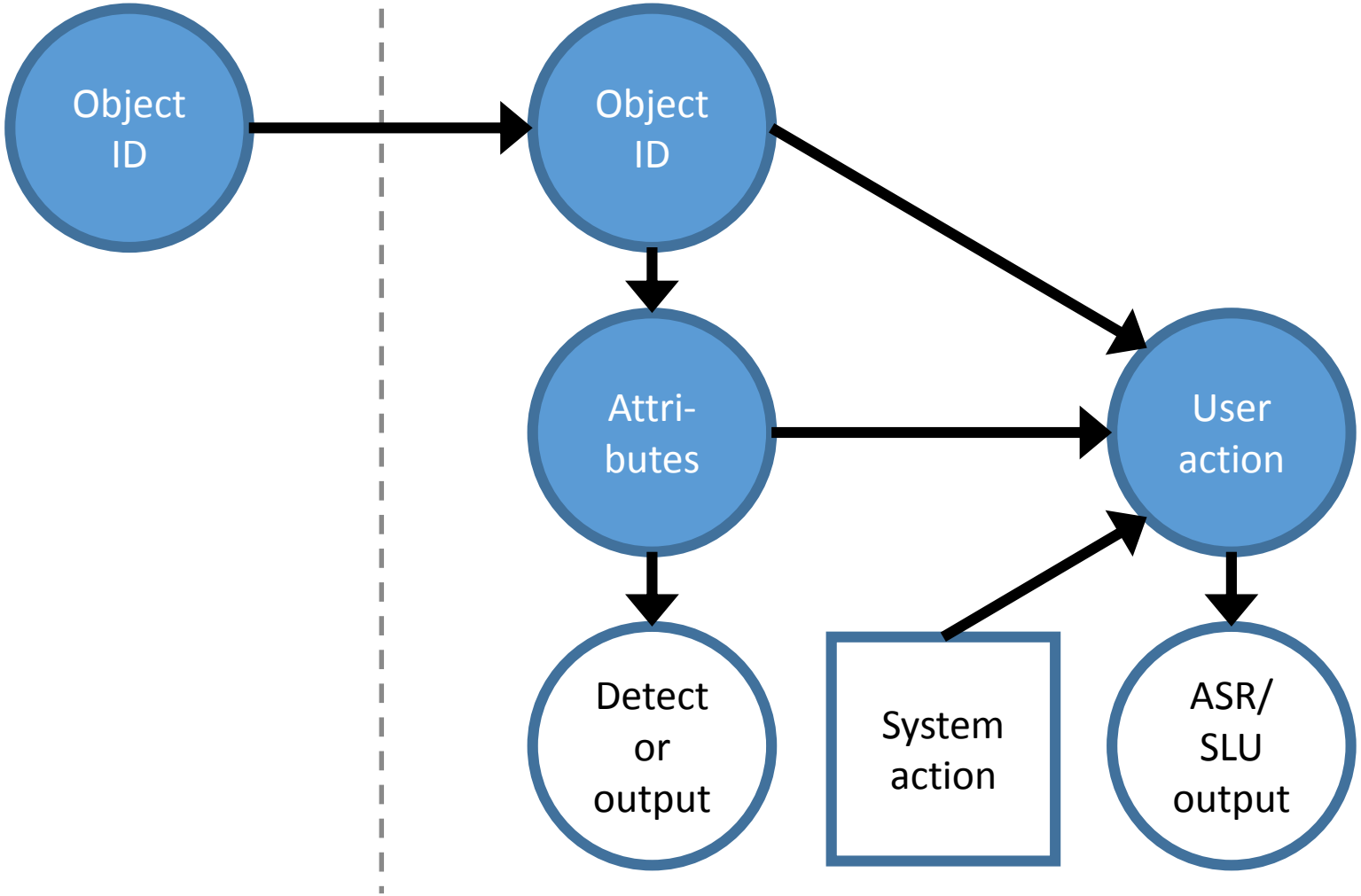
S: Is this a moose?
U: ha! no way ~ 0.7

S: Does it have hooves?
U: No ~ 0.9

S: Is this a rabbit?
U: No, much more vicious ~0.6

S: What kind of animal is it
U: Jackalope! ~ 0.7

Graphical model for conversations about objects and their attributes



Summary

- Speech recognition errors are **common and difficult to detect** reliably
- Dialog systems that do not handle errors well are **doomed to failure**
- Techniques exist to improve the accuracy of **dialog state tracking**
- These techniques suggest integration points for building **conversational systems in new and exciting domains**:
  - Machine vision conversational systems
  - Conversational robots
  - Learning within the dialog

Thanks!

Speech, speech recognition, and dialog state tracking
in spoken dialog systems

Jason D. Williams
Microsoft Research