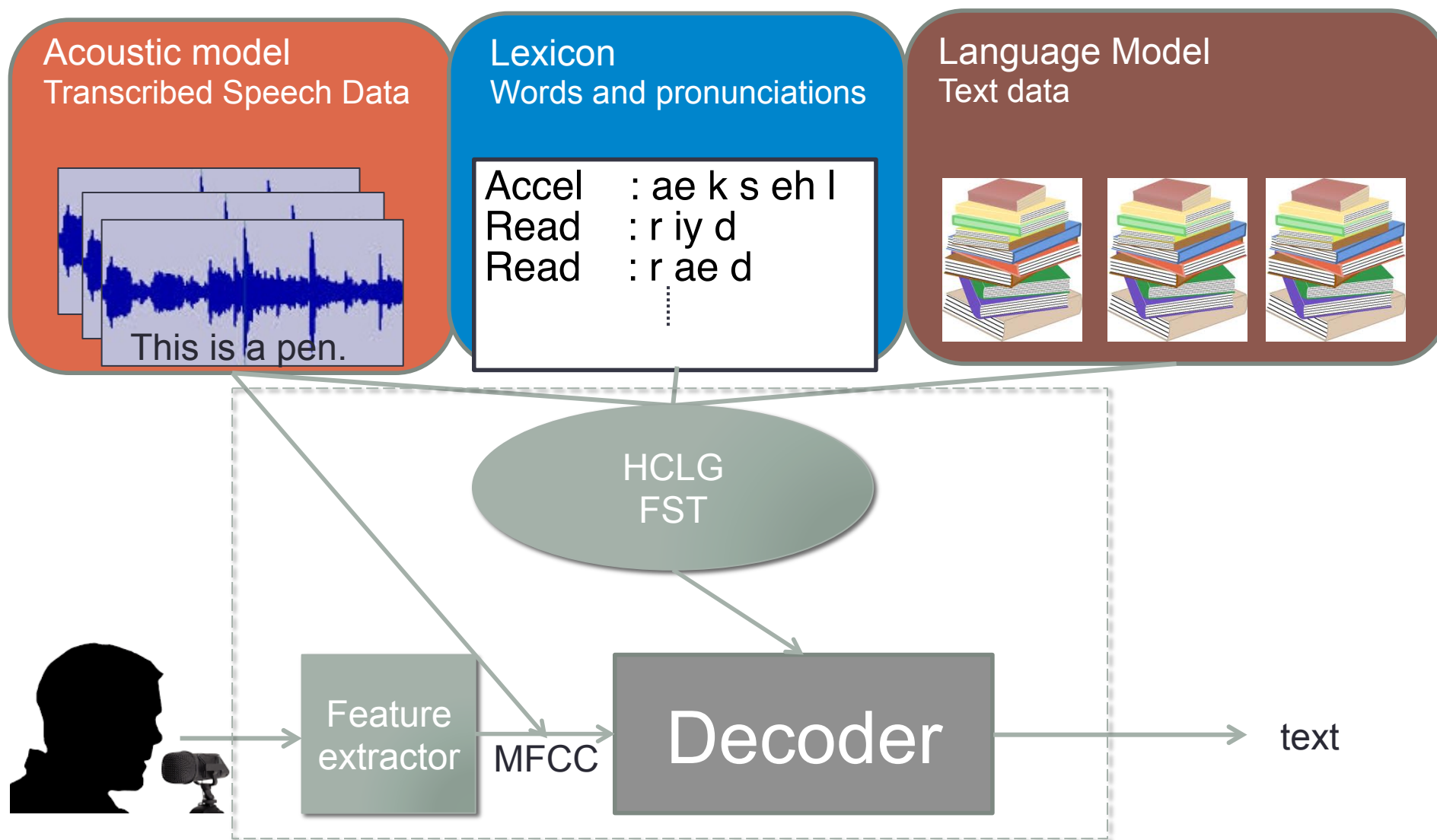




DECODING

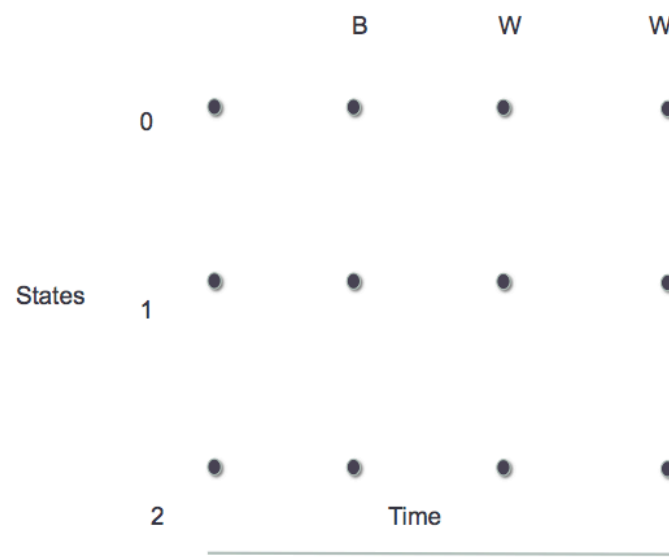
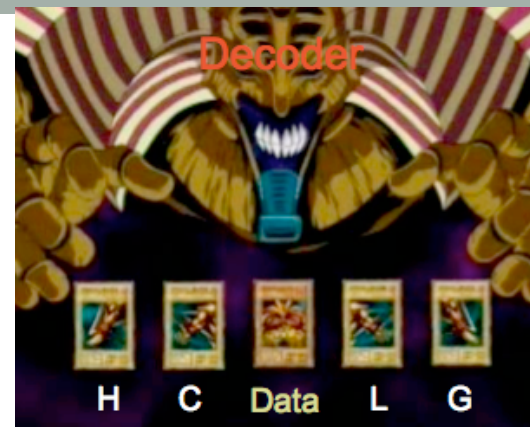
Many illustrations provided by courtesy of James Glass

Inside the recognizer



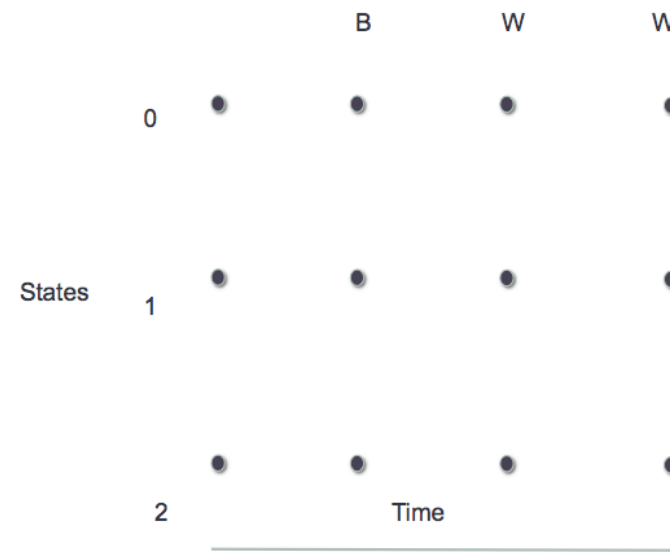
Decoder

- Goes along the FST arcs adding weights (negative log prob)
- Additional weights for emission probabilities
 - $P(\text{MFCC} \mid \text{GMM id})$
 - The GMM id is the input of the path to take in the FST
- ASR output is the path with minimum distance
- Done by Viterbi with pruning



Pruning

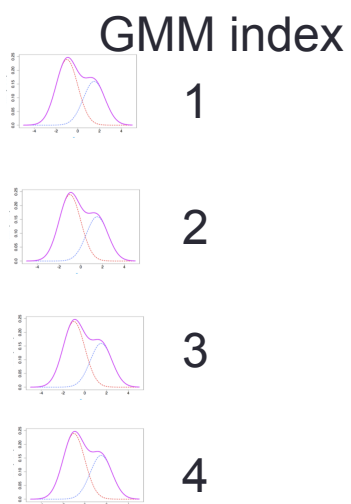
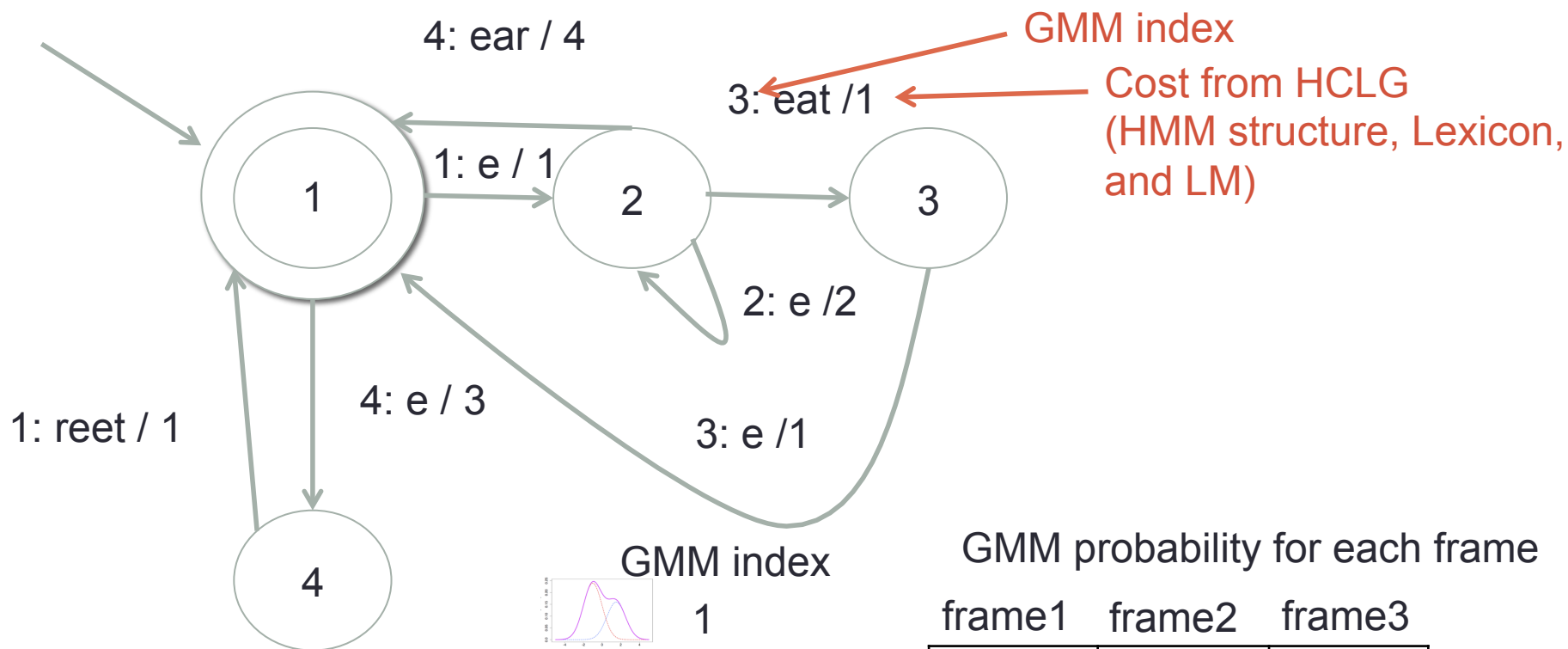
- Search space is too large!
- Usually has 1000-10000 GMM ids. So each step needs to compute 1000×1000 possibilities
- Needs pruning
- Key idea: remove paths unlikely to be the solution from consideration
- Downside : can prune the correct answer



Time synchronous beam search

- Time synchronous – consider one frame at a time
- Only consider search from **active nodes**
- Any nodes with distance more than **(best score + beam)** are pruned -- relative pruning
- Other criteria
 - Absolute pruning
 - Maximum number of active nodes
 - Minimum number of active nodes

Time synchronous beam search example



GMM probability for each frame

frame1	frame2	frame3
1	2	1
2	4	3
3	5	7
4	6	2

Other output alternatives

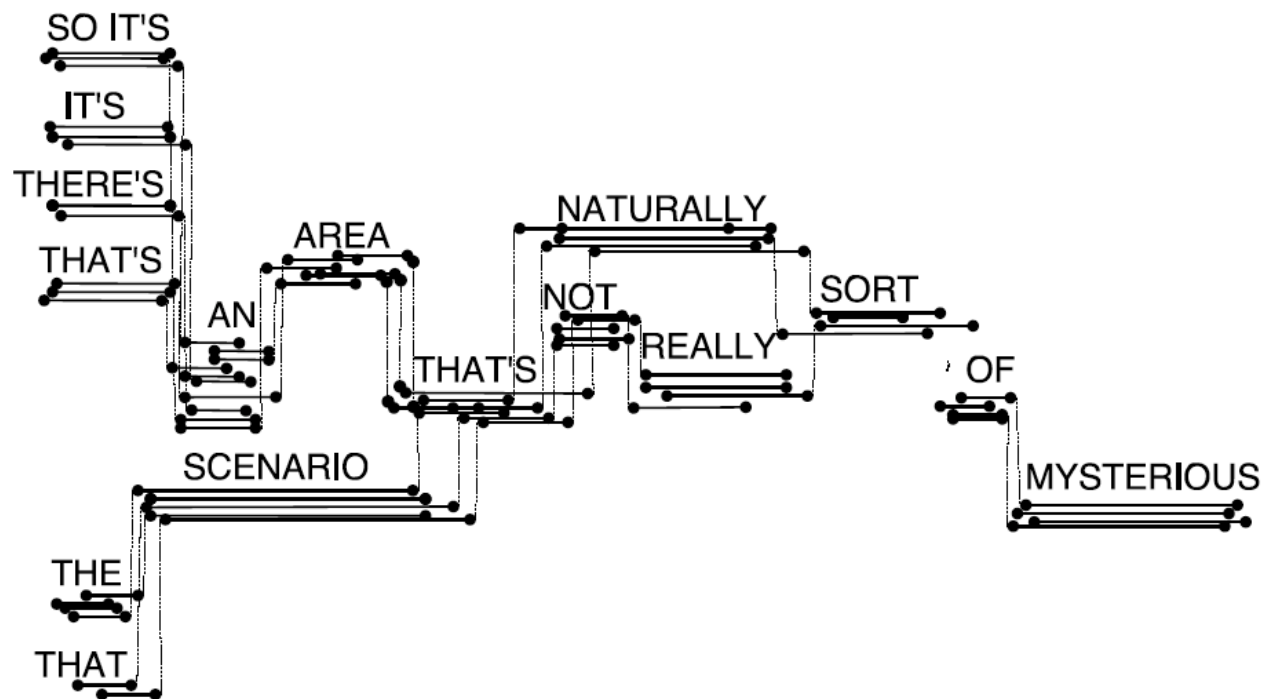
- n-best

Rank	Path	AM logprob	LM logprob
1.	it's an area that's naturally sort of mysterious	-7193.53	-20.25
2.	that's an area that's naturally sort of mysterious	-7192.28	-21.11
3.	it's an area that's not really sort of mysterious	-7221.68	-18.91
4.	that scenario that's naturally sort of mysterious	-7189.19	-22.08
5.	there's an area that's naturally sort of mysterious	-7198.35	-21.34
6.	that's an area that's not really sort of mysterious	-7220.44	-19.77
7.	the scenario that's naturally sort of mysterious	-7205.42	-21.50
8.	so it's an area that's naturally sort of mysterious	-7195.92	-21.71
9.	that scenario that's not really sort of mysterious	-7217.34	-20.70
10.	there's an area that's not really sort of mysterious	-7226.51	-20.01

- ASR has errors, n-best can recover the correct answer for downstream applications.

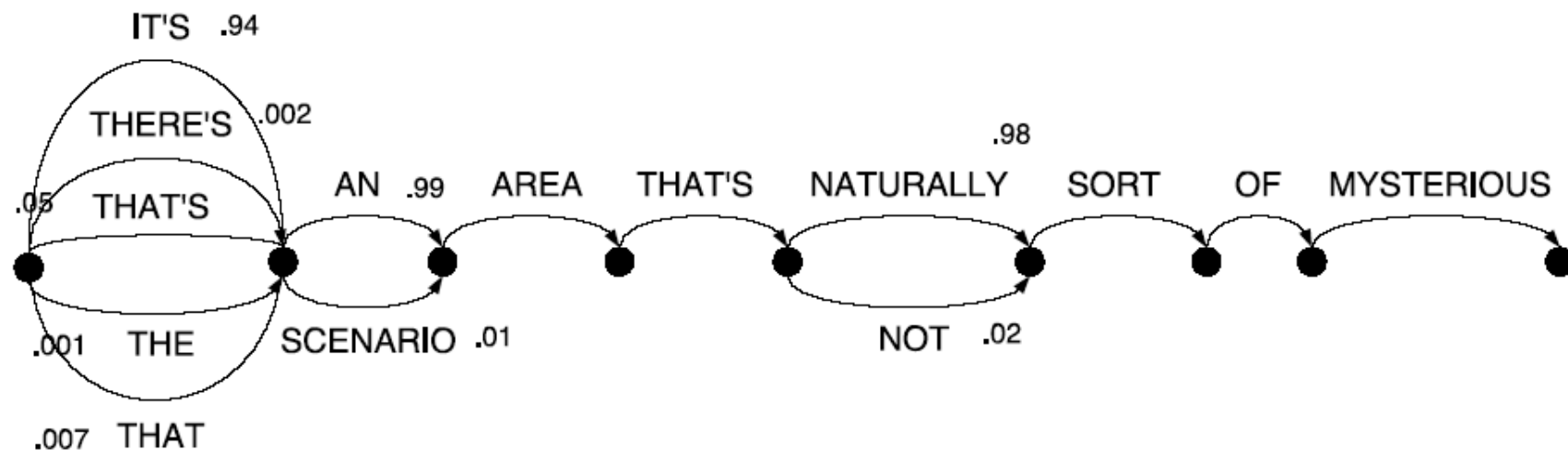
Other output alternatives

- Word lattice
 - A structure of the likely paths from decoding
 - Still have path scores
 - Each arc can have different timings for the same word.



Other output alternatives

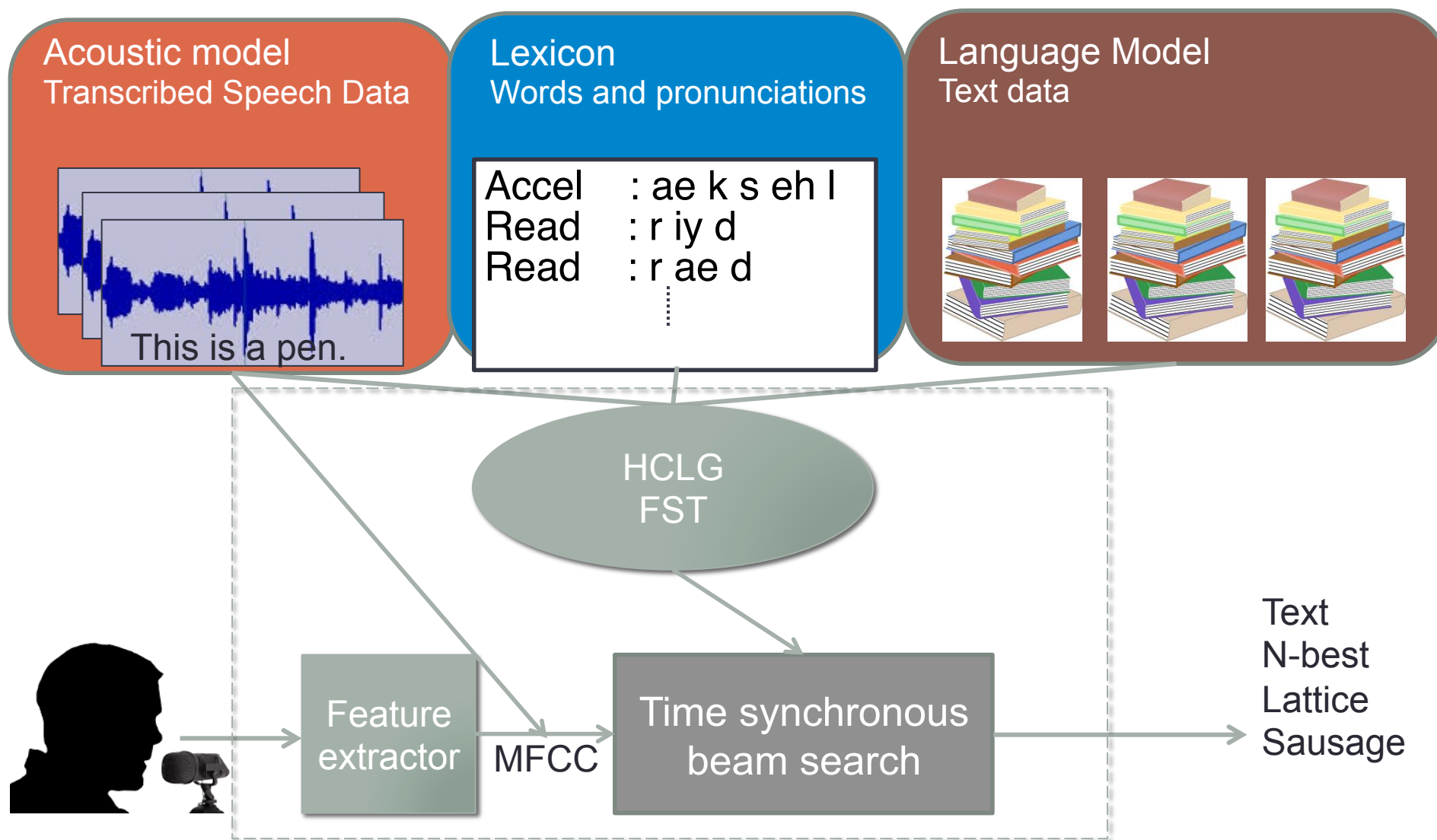
- Confusion networks (sausages)
 - A compact representation of word lattices
 - Many timing approximations
 - Some path is gone (not-really)
 - Some new path (It's scenario)
 - Easier for downstream application to process than lattices.



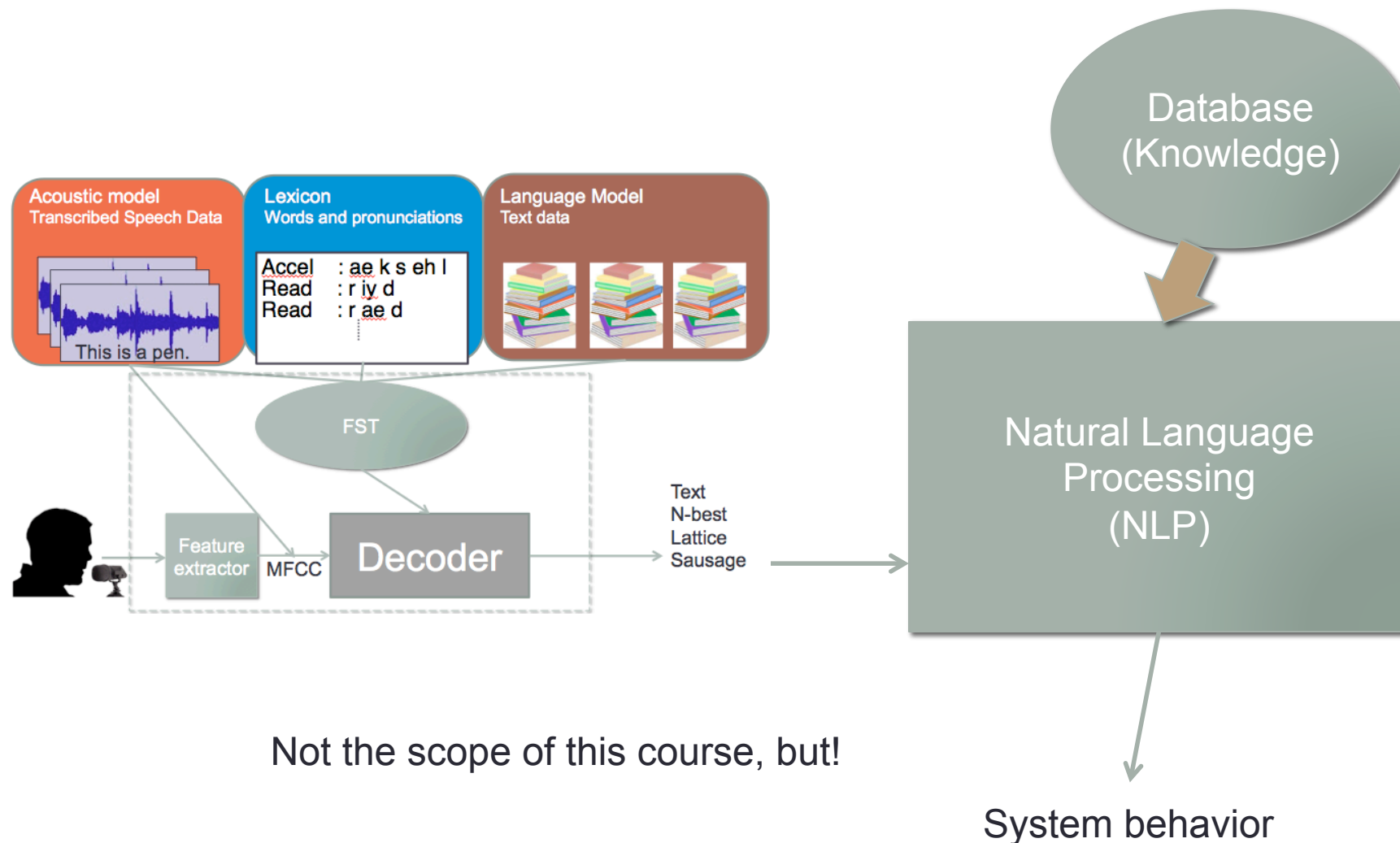
Search consideration

- Weighting between AM and LM
 - $P^{AMW}(X | L) P(L | W) P^{LMW}(W)$
- $P(X|L)$ and $P(W)$ very has different dynamic range.
 - $P(X|L)$ computed every frame. $P(W)$ computed every word.
 - $P(X|L)$ prefers more words. $P(W)$ prefer less words.
 - Increasing AMW gives more insertion error but less deletion error and vice versa.
- Can also add **word insertion penalty** (P_{in})
 - A constant is multiplied for every word
 - $P^{AMW}(X | L) P(L | W) P^{LMW}(W) P_{in}^{\#words}$
- These parameters are tuned using the dev set

Inside the recognizer



Understanding the output



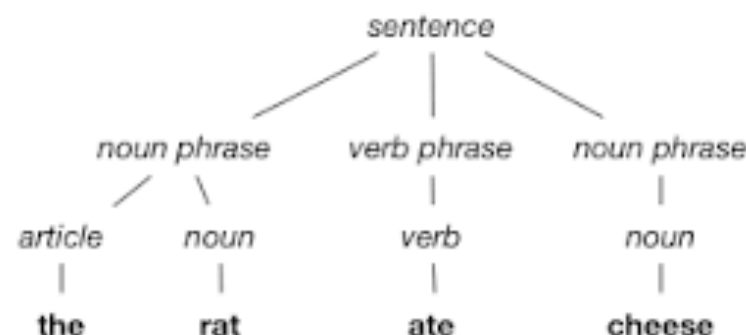


NLP

In 10 minutes

NLP tasks

- Syntactic Understanding
 - Parse Tree
 - Identify sentence structure
- Semantic Understanding
 - Map words to concepts



“Find me a highly rated robot sci-fi movie”

↓ ↓ ↓ ↓

<other> <rating> <plot> <genre>

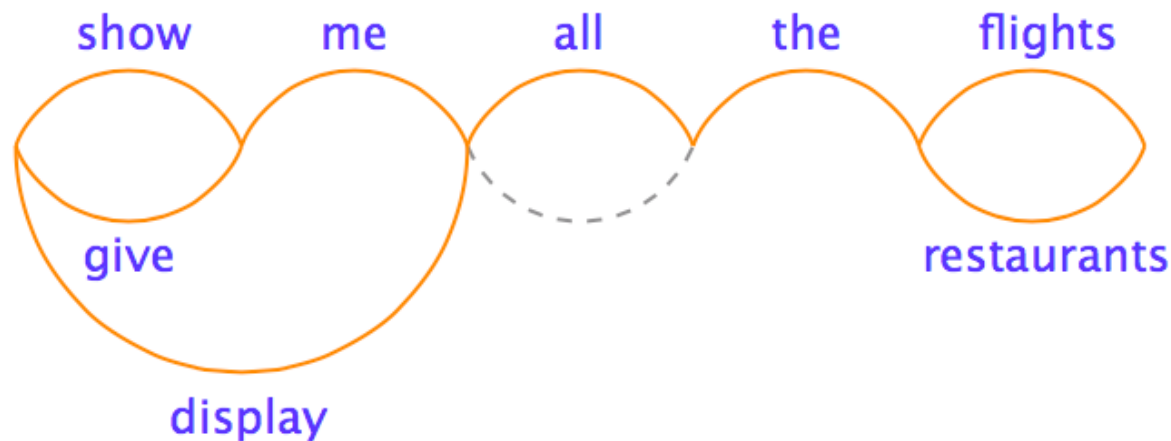
- And more... word segmentation, morphology, translation, summarization, sentiment analysis,...

NLP approaches (for ASR)

- Rule-based NLP
 - Human designs what are possible inputs and what's the meaning
 - Finite state networks and Context-free grammars (CFGs) for parsing and rules/templates for semantic interpretation
- Statistical NLP
 - Trained a classifier based on label data
 - Handles free form queries

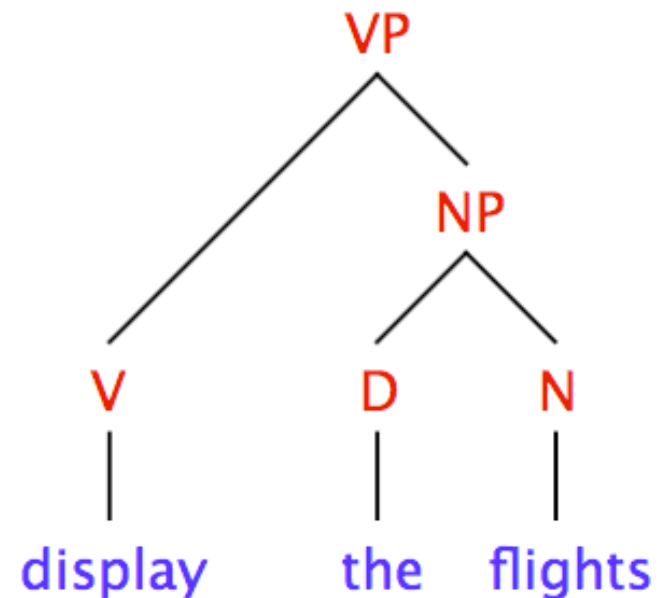
Finite State Networks

- Human expert defines all the possible input as a graph/ FST
- Replace the G FST
- Cannot handle unexpected grammars, and speaking errors.



Context Free Grammars (CFG)

- Describe by rewrite rules
 - $VP \rightarrow V NP$
 - $NP \rightarrow D N$
 - $V = \{\text{display, show}\}$
 - $D = \{\text{the, a}\}$
 - $N = \{\text{flights, flight, airport, plane}\}$
- Recursive structure makes it bad for FST format.
- Finite coverage. Does not handle speaking errors well.



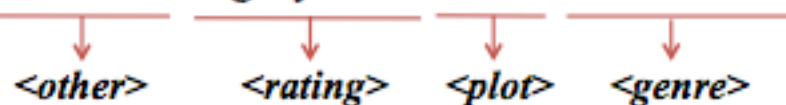
Comparison

	Rule-based	Statistical
Rules or templates	✓	
Data dependency		✓
Expert knowledge	✓	
Human efforts	✓	
Time consuming	✓	
In-domain accuracy	✓	
Domain-independent		✓
Scalability		✓
Out-of-vocabulary words		✓
Recognition errors		✓
Incremental improvement	✓	
Expert control	✓	

Semantic Tagging

- Segmentation and classification problems
 - Segment the sentence into phrases and classify into classes

“Find me a highly rated robot sci-fi movie”

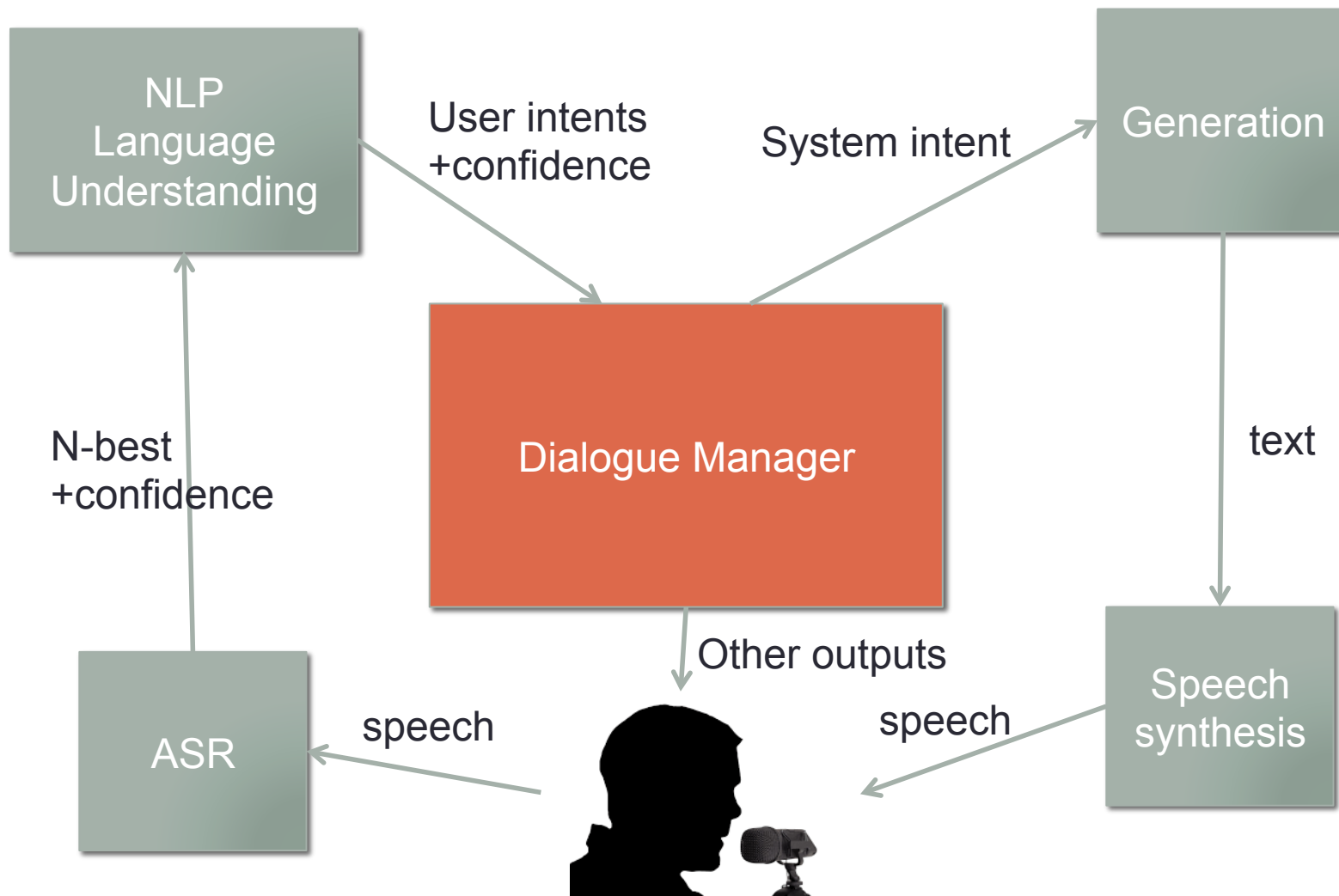

<other> <rating> <plot> <genre>

“Book me a double room for 2 at Marriott Bellevue on Friday”


<other> <room_type> <#people> <hotel_name> <location> <reservation_date>

- Need to collect labeled data
- Popular method: Conditional Random Fields (CRF) and its variants
 - Being replaced by Deep Learning
- Beyond the scope of this course

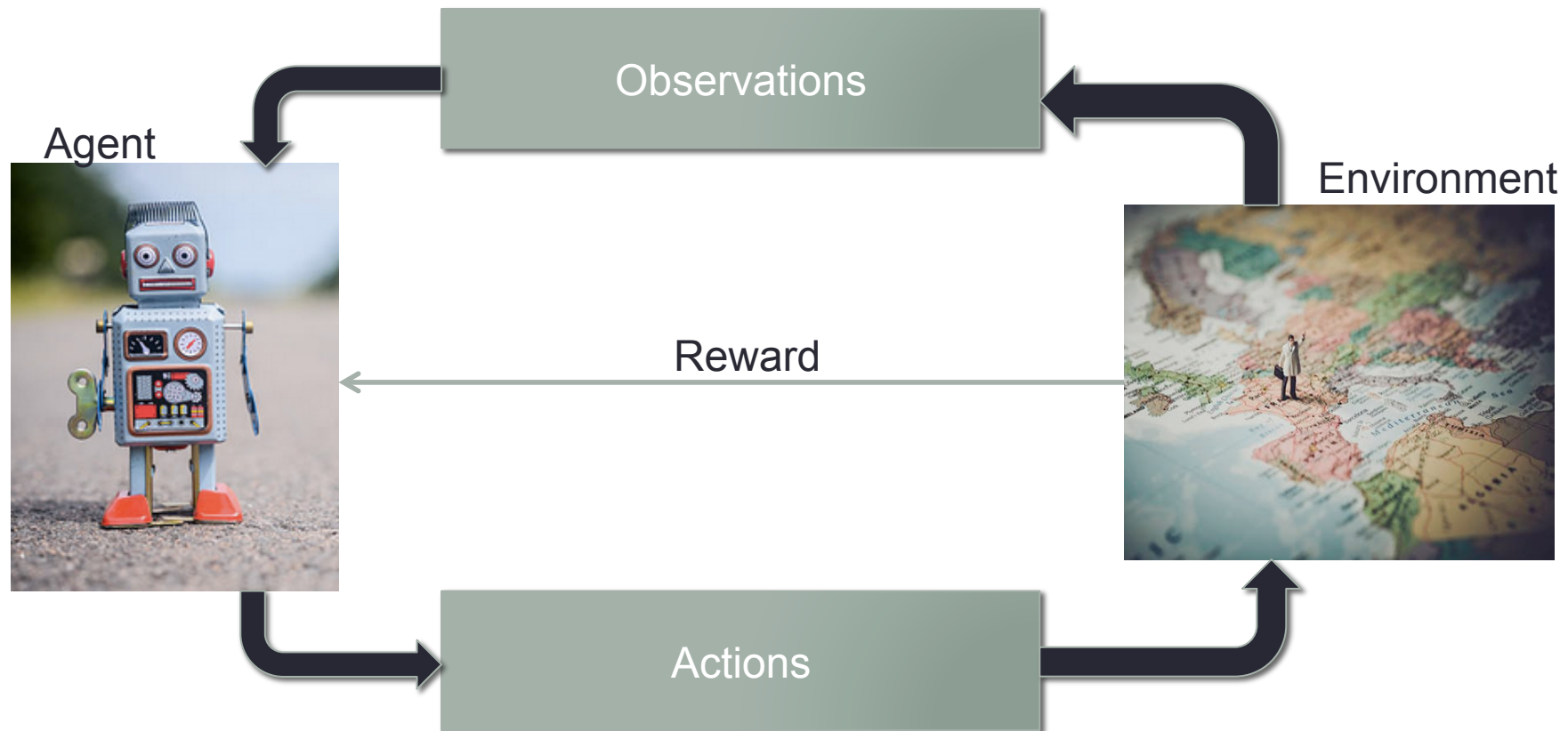
Dialogue systems



Dialogue management

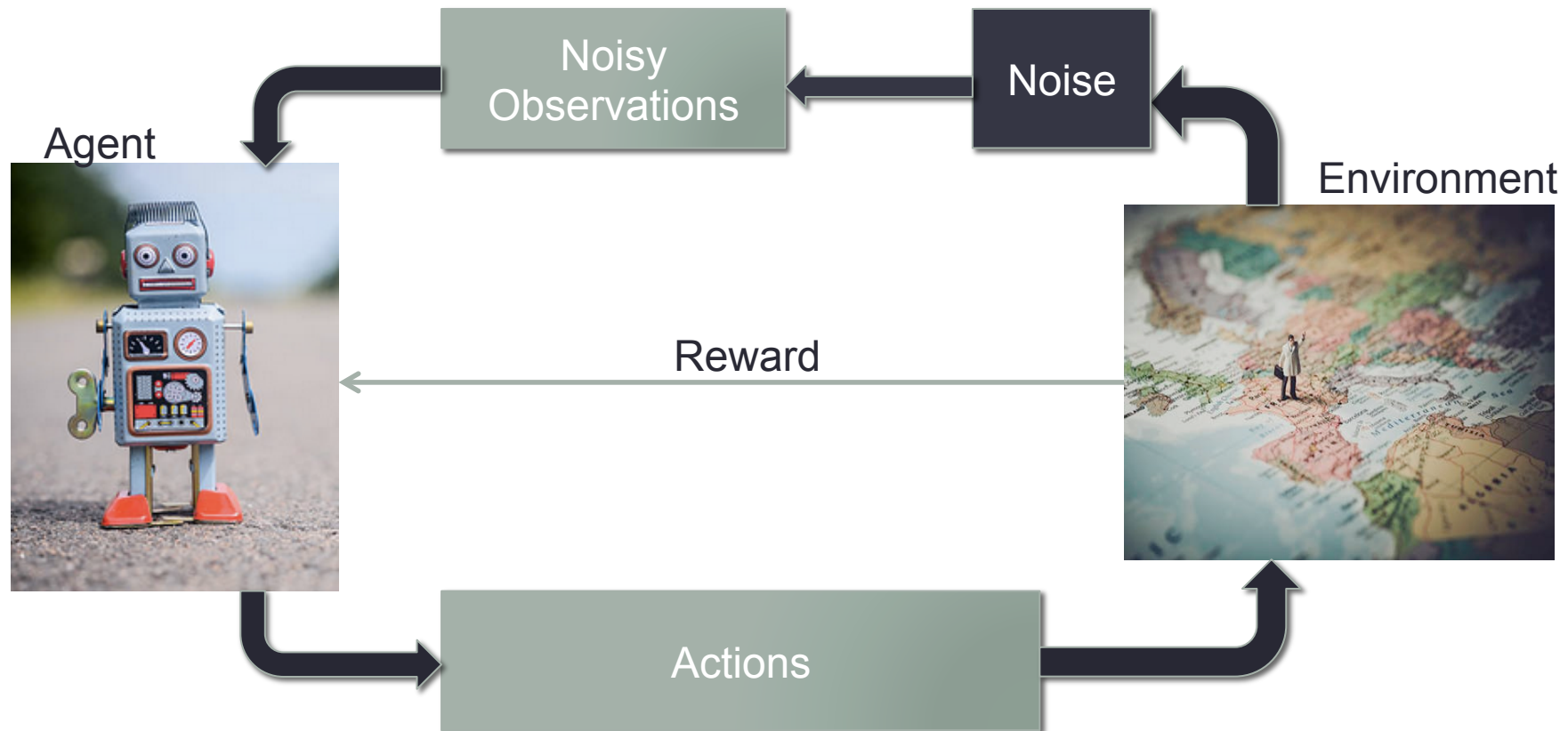
- Dialogue management is in charge of controlling the conversation, and making decisions to say/do things
- Action based on some representation of the current dialogue state
- State changes according to the language understanding output
- Needs to cope with errors from ASR and NLP
 - Compare cost of
 - Doing something wrong
 - Asking the user to confirm
 - Asking the user to repeat the sentence

Reinforcement learning framework - Markov Decision Process (MDP)



Learning through trial and error

Partially Observable Markov Decision Process (POMDP)



Agent does not directly observe the environment
Agent does not know the intent of the user, but sees his input

Dialogue management methods

- Partially Observable Markov Decision Process (POMDP)
 - The user state (intent) is unobserved, but the system observed a noisy user input (speech).
 - The model tradeoffs between the cost of different actions depending on the estimated state
 - Used for other applications such as human robot interaction
 - Very similar to HMM: key difference
 - Your action influences the distribution of the next state
- Deep learning on the rise
- Unsolved problem – how to learn the appropriate cost from data

POMDP application in decision making

- Saving endanger species
- Problem : We want to save an endanger species. We do not know if it still exists.
- Actions
 - A) continue surveying till you find species
 - B) Protect the species (laws, make the forest national park)
 - C) Do nothing
- Hidden state : Species exists or not
- Cost function : Heuristics about the value of the species and how threatened it is



Chadès, Iadine, et al. "When to stop managing or surveying cryptic threatened species." *Proceedings of the National Academy of Sciences* 105.37 (2008): 13936-13940.

Kaldi demo

- https://github.com/ekapolc/ASR_classproject

Project

- What we give
 - The AM (based on the class recordings)
 - How to hook Kaldi to get ASR results
 - How to use the G2P
- What we expect
 - Refine your scope
 - Create a set of dev and test sentences (10 dev 10 test per person)
 - Generate LM training text
 - Limit your vocabulary – generate lexicons based on G2P + human correction
 - Measure performance in perplexity (and WER in the future)
 - Aim small. AM is bad. Compensate with low perplexity and low vocabulary size
 - Can you parse the ASR output using simple rules for your task
- What we don't expect
 - Statistical NLP
 - Dialogue systems
 - Gowajee wake word

Next time

- Deep learning
 - Why deep learning is on the rise
 - Deep learning variants
 - CNN, RNN, GRN, LSTM, attention model
 - Applications in ASR
 - AM
 - LM
 - End-to-end
 - Other related applications
 - Word2vec
 - Summarization\Translation
 - Deep learning tricks
 - What's a good candidate for deep learning applications
 - Transfer learning
 - Multi task learning
 - Unsupervised training

