HCI in an IOT World

Jim Glass

Senior Research Scientist

Computer Science and Artificial Intelligence Laboratory (CSAIL)
Massachusetts Institute of Technology

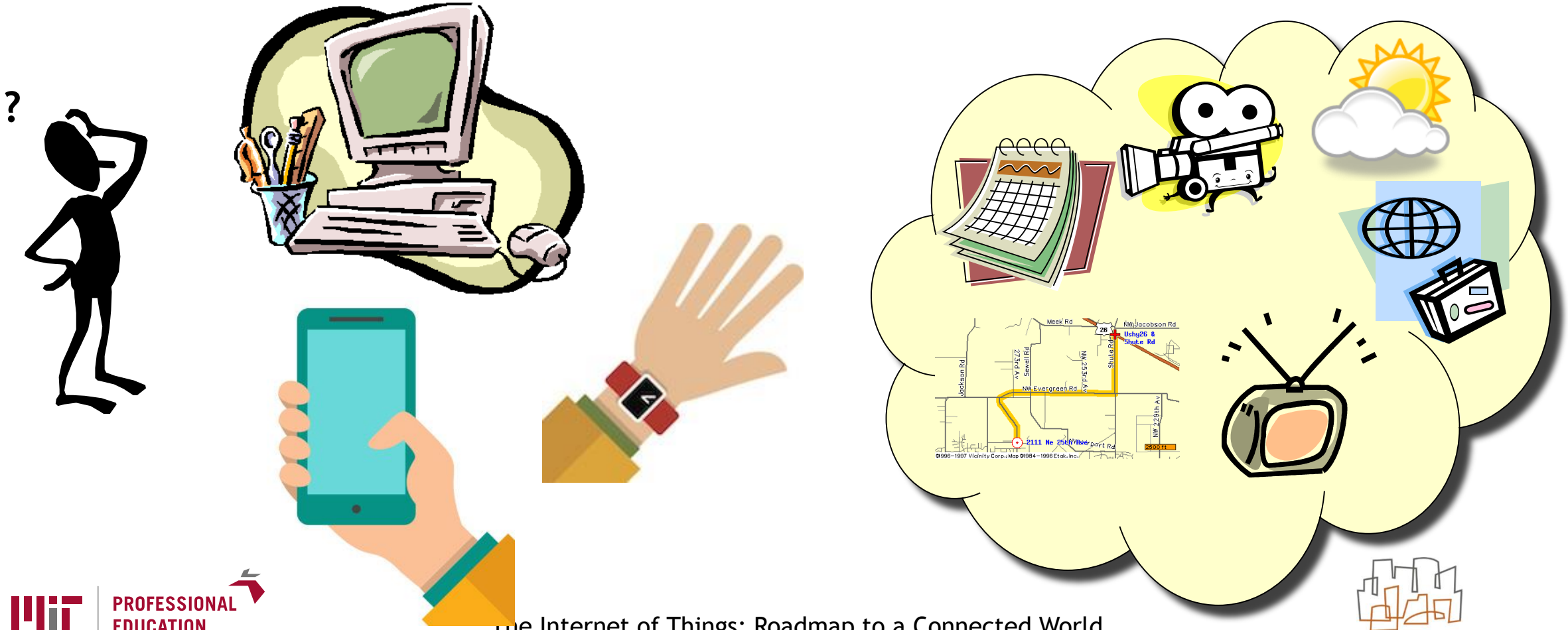
CSAIL
MIT COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE LABORATORY

OUTLINE

- **HCI for IOT**
- **Prototypes**
- **Technologies**
- **Multi-modality**
- **Computation**
- **Challenges**

HUMAN-COMPUTER INTERACTION (HCI)

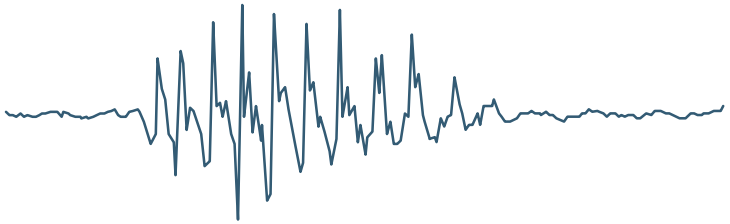
- **HCI focuses on the interface between people and computers**



HUMAN-COMPUTER INTERACTION FOR IOT

- **Q: How are we going to communicate with all our devices?**

- **A: Speak with them!**

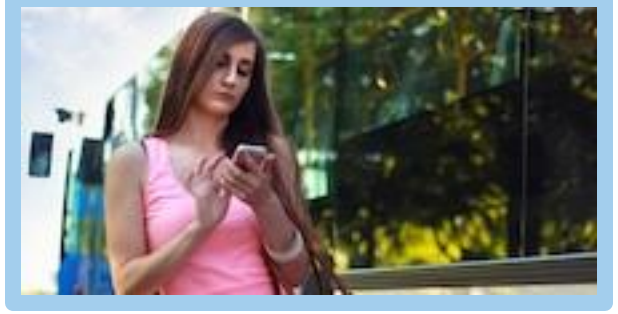


THE NEED FOR SPEECH

- **Speech has reached a tipping point**
- **Users want to speak to their devices**

Are there any good Thai restaurants near here?

I need a flight to Chicago this afternoon around 4.



Any recommendations for good action movies with car chases?

- **"Speech" means more than speech**

Understanding

Dialogue

Generation

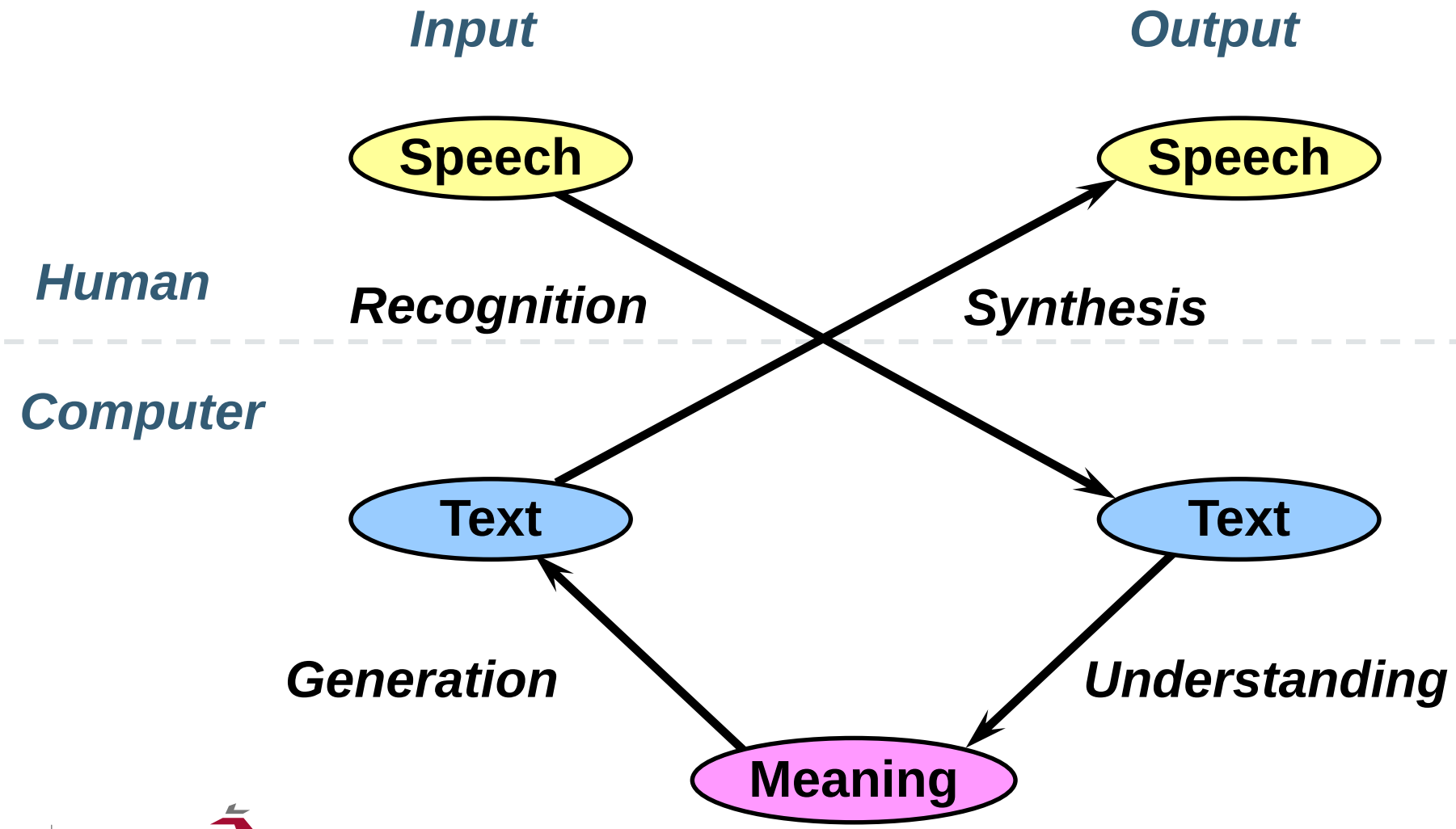
VIRTUES OF SPOKEN LANGUAGE

Natural	**Requires no special training**
Flexible	**Leaves hands and eyes free**
Efficient	**Has high data rate**
Economical	**Can be transmitted and received inexpensively**

Speech is ideal for information access & management when:

- The information space is broad and complex
- The users are technically naïve, or
- The device is small

COMMUNICATION VIA SPOKEN LANGUAGE



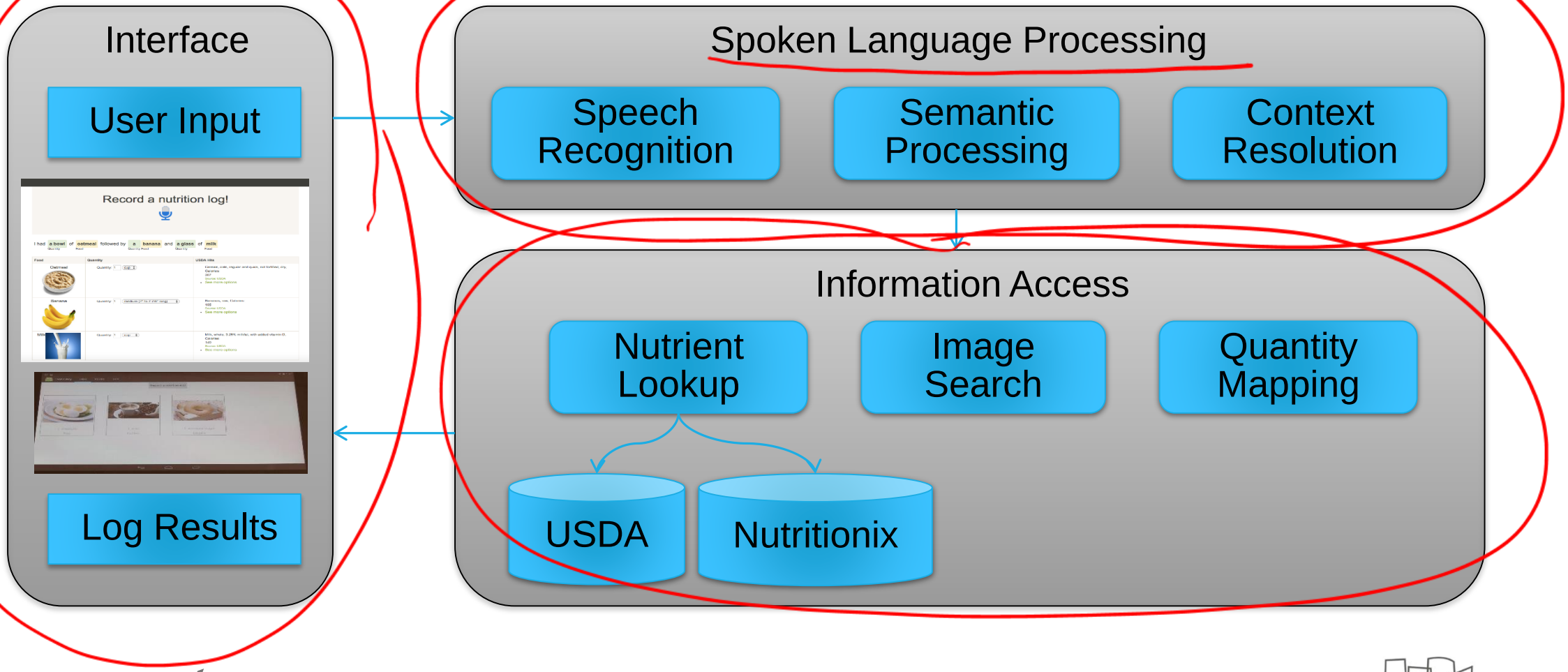
ADVANCED SPEECH INTERFACES

- **Can communicate with users through *conversation***
- **Can *understand* verbal input**
  - Speech recognition
  - Language understanding (in context)
- **Can *verbalize* response**
  - Language generation
  - Speech synthesis
- **Can engage in *dialogue* with a user during the interaction**

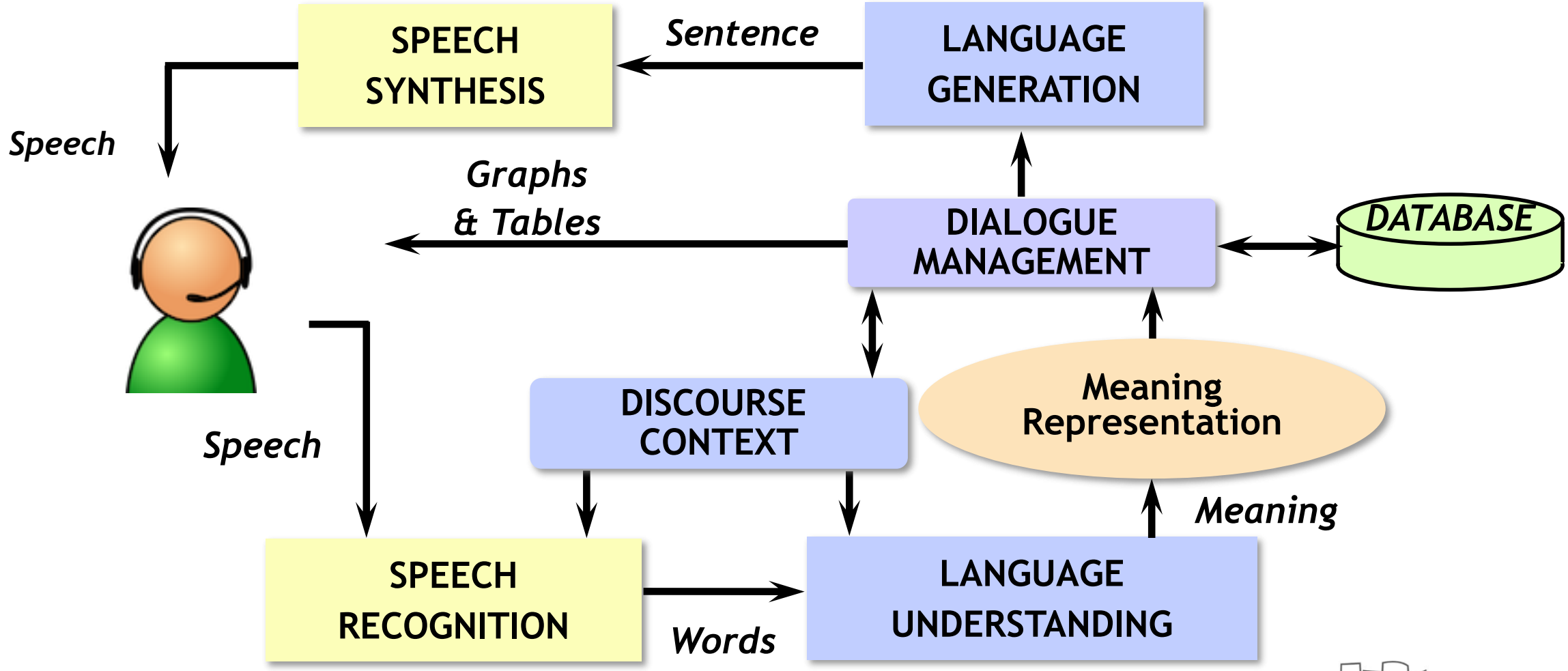
SPOKEN LANGUAGE PROTOTYPES

SPOKEN LANGUAGE TECHNOLOGIES

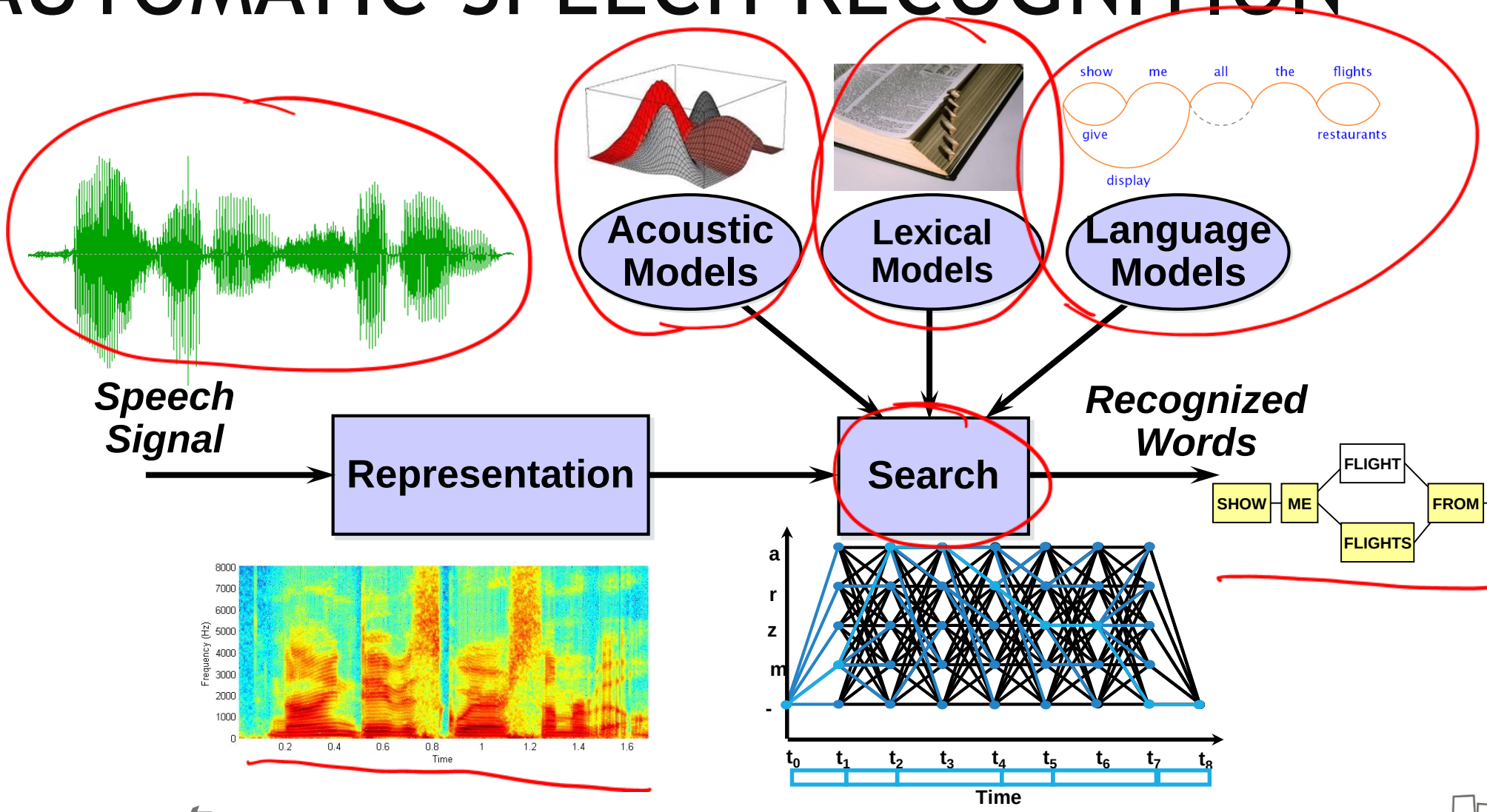
EXAMPLE SYSTEM ARCHITECTURE



SPOKEN LANGUAGE TECHNOLOGIES



AUTOMATIC SPEECH RECOGNITION

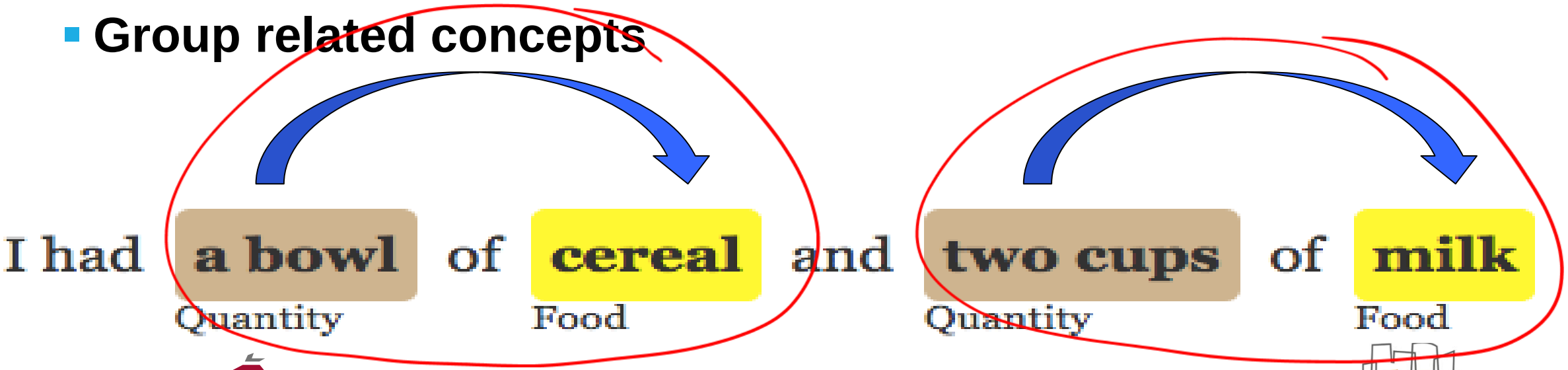


SEMANTIC PROCESSING

- **Identify and label concepts**

I had **a bowl** of **Kellogg's** **frosted** **flakes**
Quantity Brand Description Food

- **Group related concepts**



CONTEXT RESOLUTION

I had **a cup** (Quantity) of **yogurt** (Food) and **a strip** (Quantity) of **bacon** (Food) for breakfast

It was **greek** (Description) **yogurt** (Food)

Food	Quantity	USDA Hits
Yogurt 	Quantity: 1 container	Yogurt, Greek, plain, nonfat, Calories: 59 Source: USDA • See more options • Back
Bacon 	Quantity: 1 slice raw	Pork, cured, bacon, unprepared, Calories: 116.76 Source: USDA • See more options

I had **some** (Quantity) **milk** (Food)

It was actually **3 cups** (Quantity) of **milk** (Food)

Food	Quantity	USDA Hits
Milk 	Quantity: 3 cup	Milk, whole, 3.25% milkfat, with added vitamin D, Calories: 61 Source: USDA • See more options • Back

DIALOGUE MANAGEMENT

- **Pre-Retrieval: Ambiguous Input => Unique Query to DB**

U: I need a flight from Boston to San Francisco
C: Did you say Boston or Austin?
U: Boston, Massachusetts
C: I need a date before I can access the database
U: Tomorrow
C: Hold on while I retrieve the flights for you

Clarification (recognition errors)

Clarification (insufficient info)

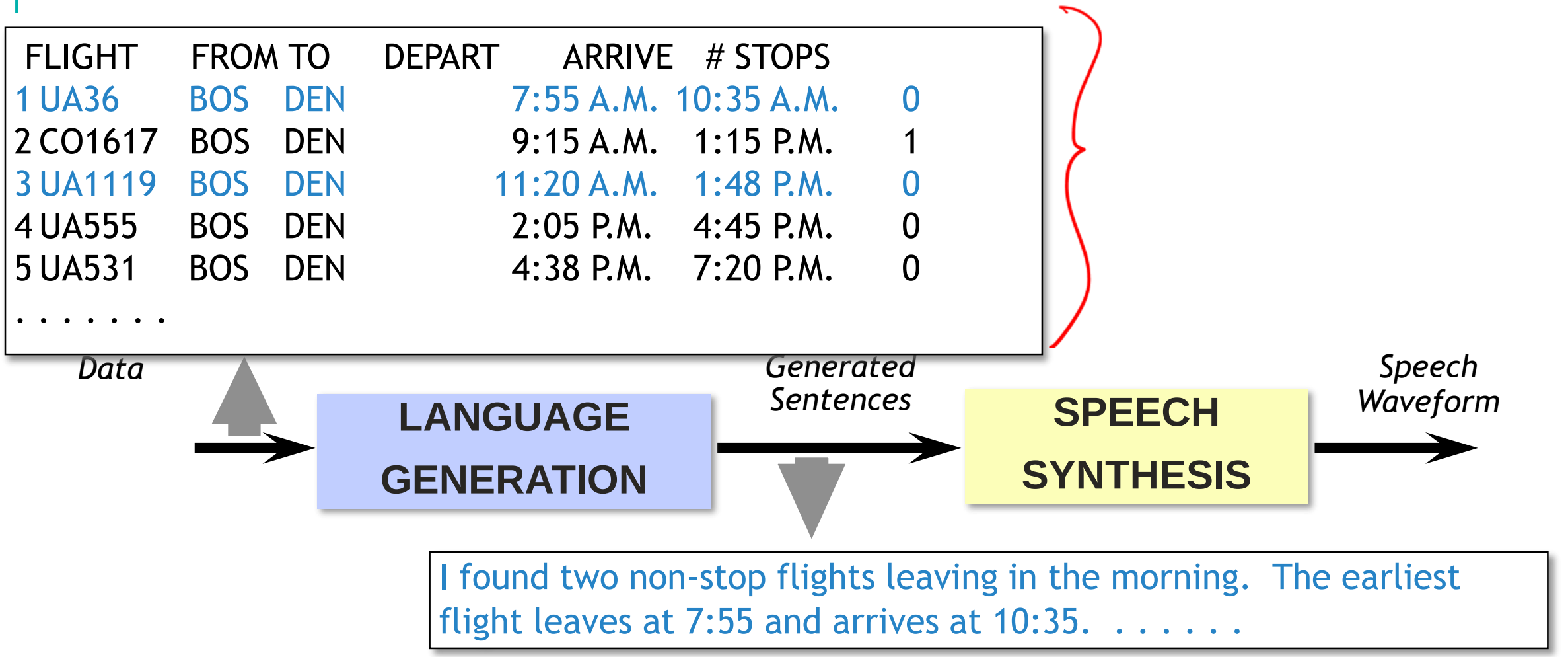
- **Post-Retrieval: Multiple DB Retrievals => Unique Response**

C: I have found 10 flights meeting your specification. When would you like to leave?
U: In the morning.
C: Do you have a preferred airline?
U: United
C: I found two non-stop United flights leaving in the morning …

Help the user narrow down the choices

SPEECH GENERATION

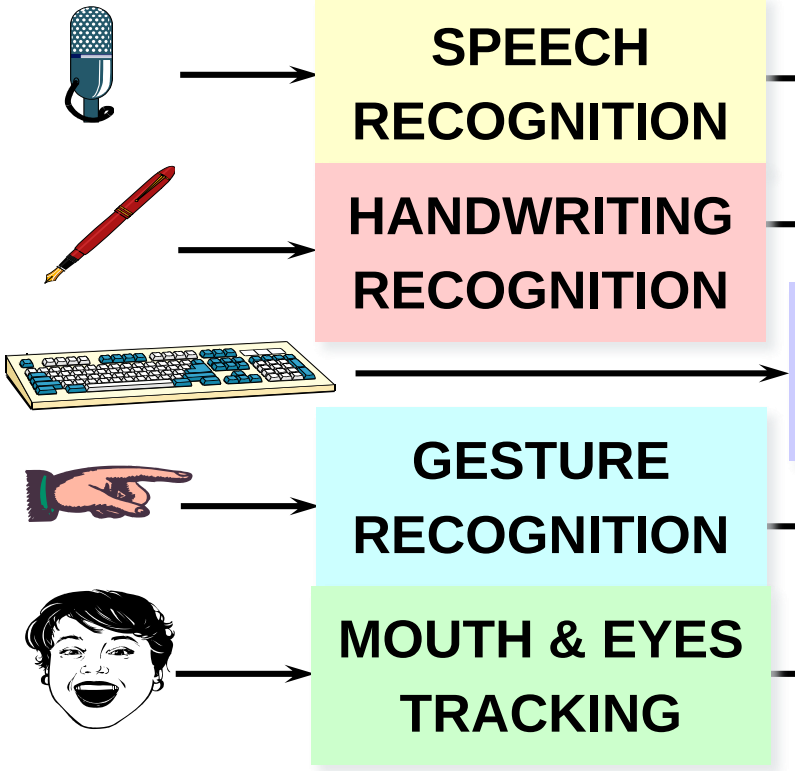
FLIGHT	FROM	TO	DEPART	ARRIVE	# STOPS
1 UA36	BOS	DEN	7:55 A.M.	10:35 A.M.	0
2 CO1617	BOS	DEN	9:15 A.M.	1:15 P.M.	1
3 UA1119	BOS	DEN	11:20 A.M.	1:48 P.M.	0
4 UA555	BOS	DEN	2:05 P.M.	4:45 P.M.	0
5 UA531	BOS	DEN	4:38 P.M.	7:20 P.M.	0
. . . . .					

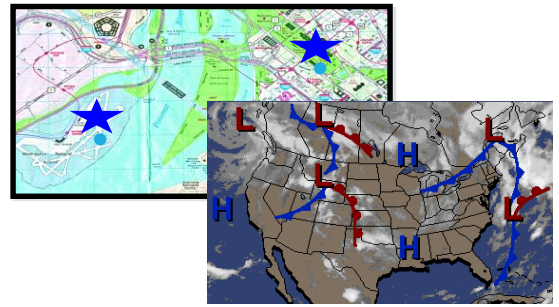


MULTI-MODAL INTERACTION

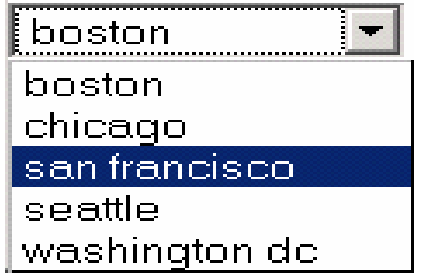
MULTI-MODAL INTERACTION

- Flexible: Users select preferred modalities
- Efficient: Language + gestures can be simpler than uni-modal interfaces
- Robust: Inputs are complementary and consistent





SPEECH AND GESTURE INTEGRATION

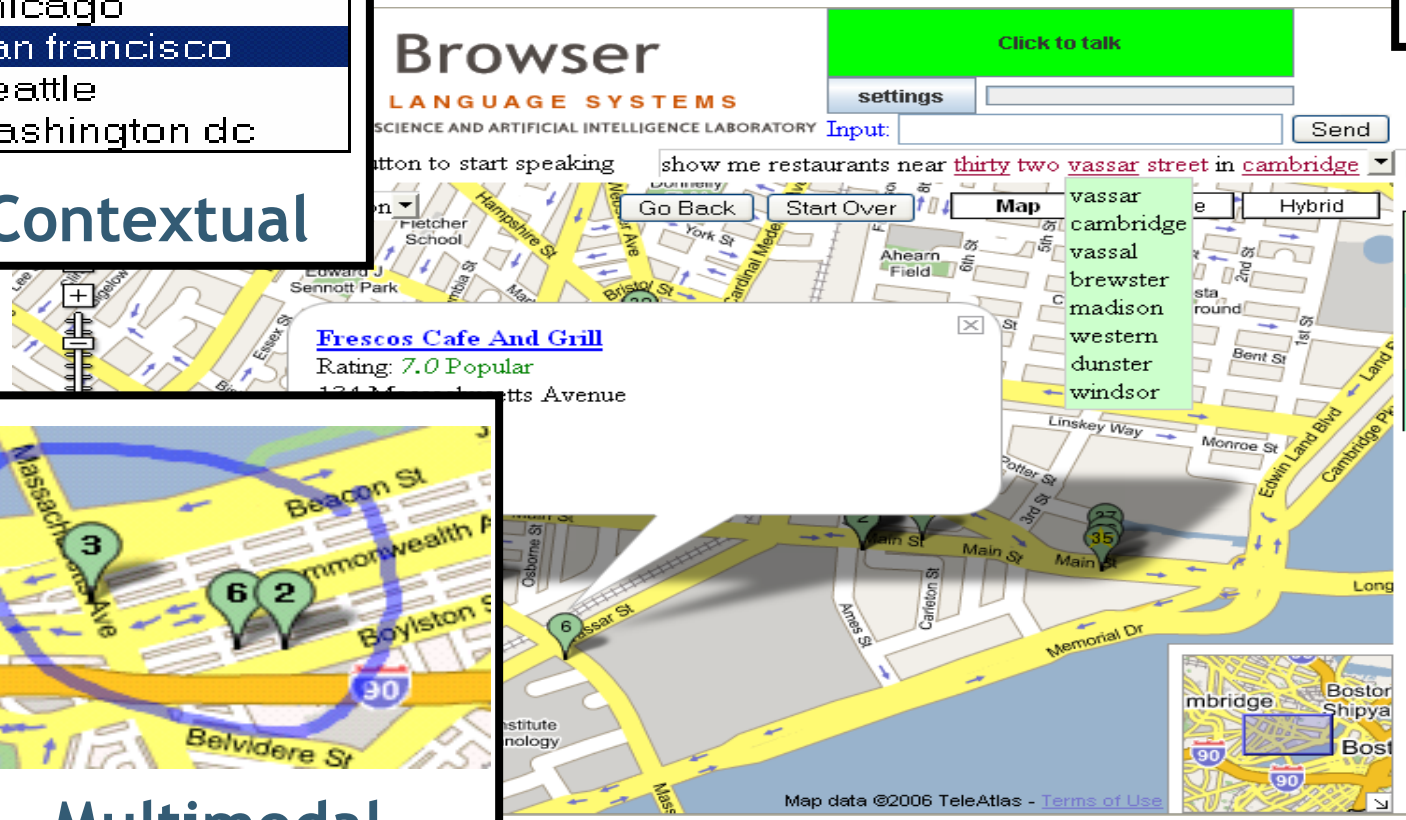


Contextual

Accessible

Top Speaker: Jim_Glass (99%)

Personalizable

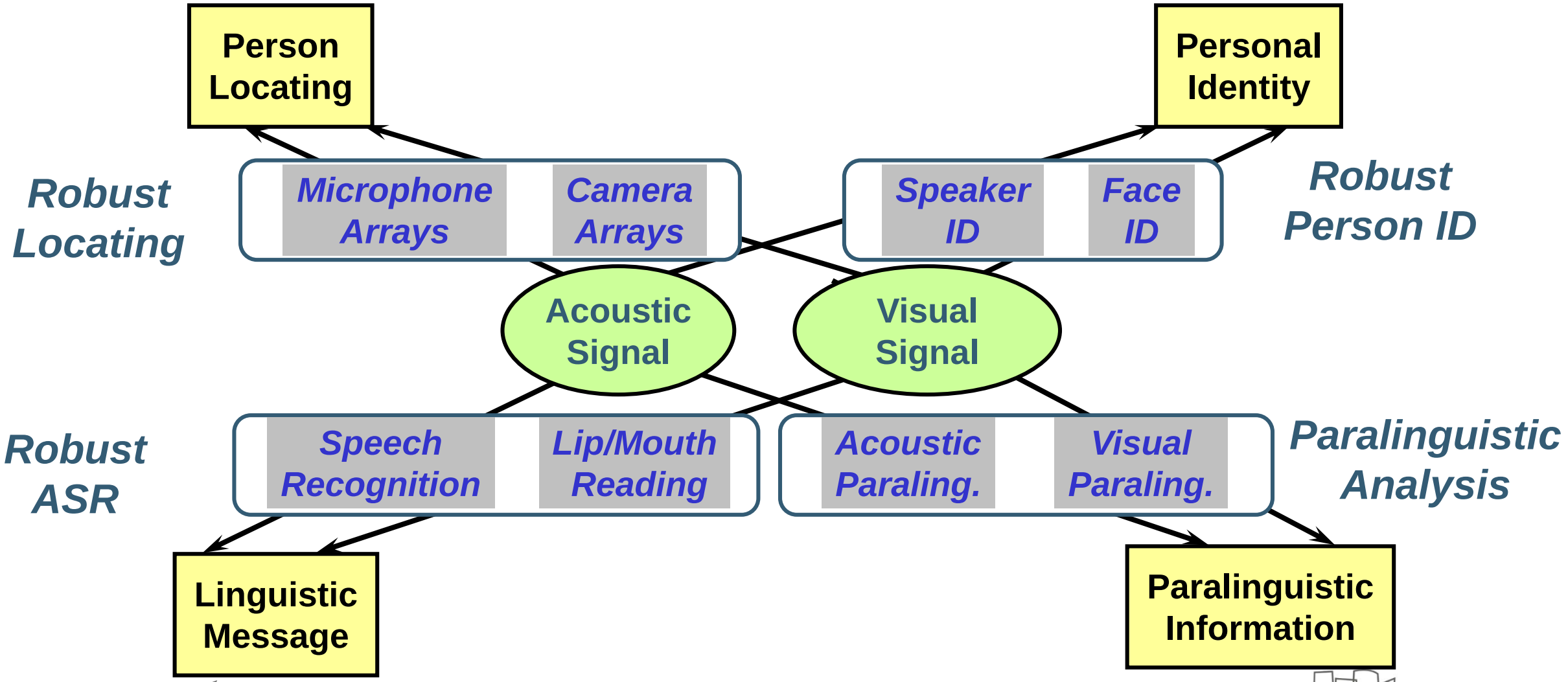


Multimodal

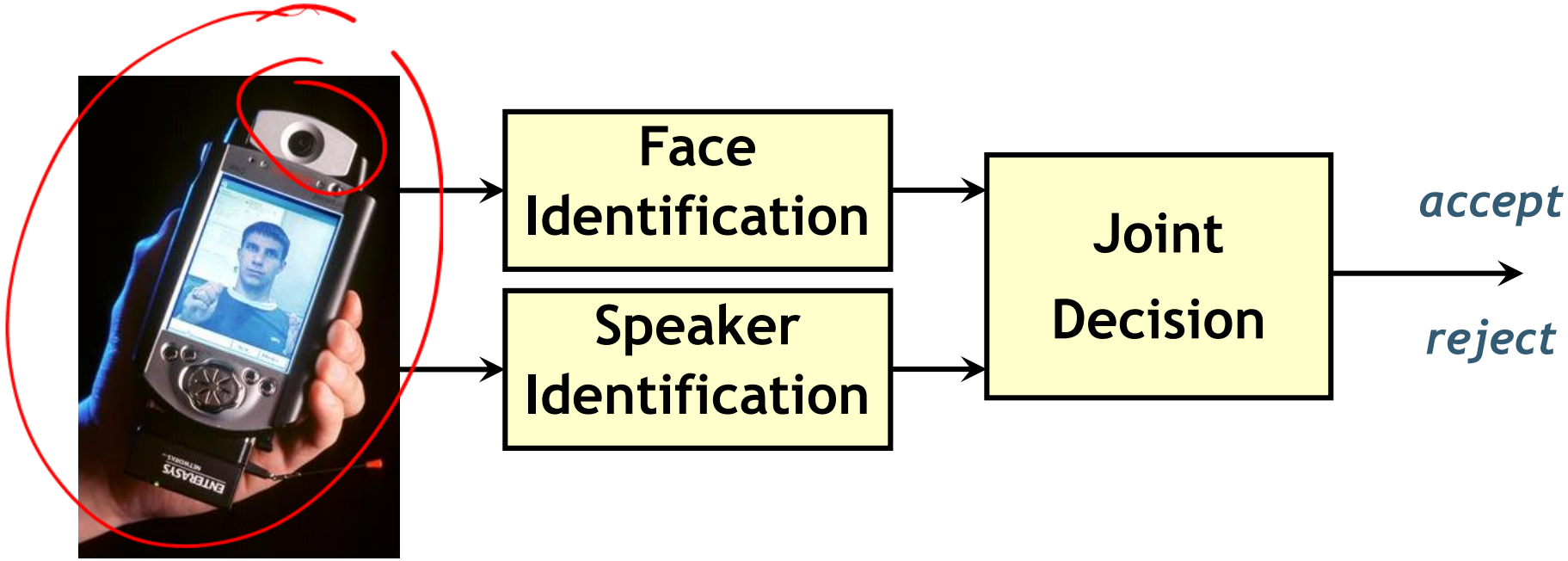
Customizable

MULTI-MODAL SYMBIOSIS

MULTI-MODAL SYMBIOSIS



MULTI-MODAL PERSON VERIFICATION

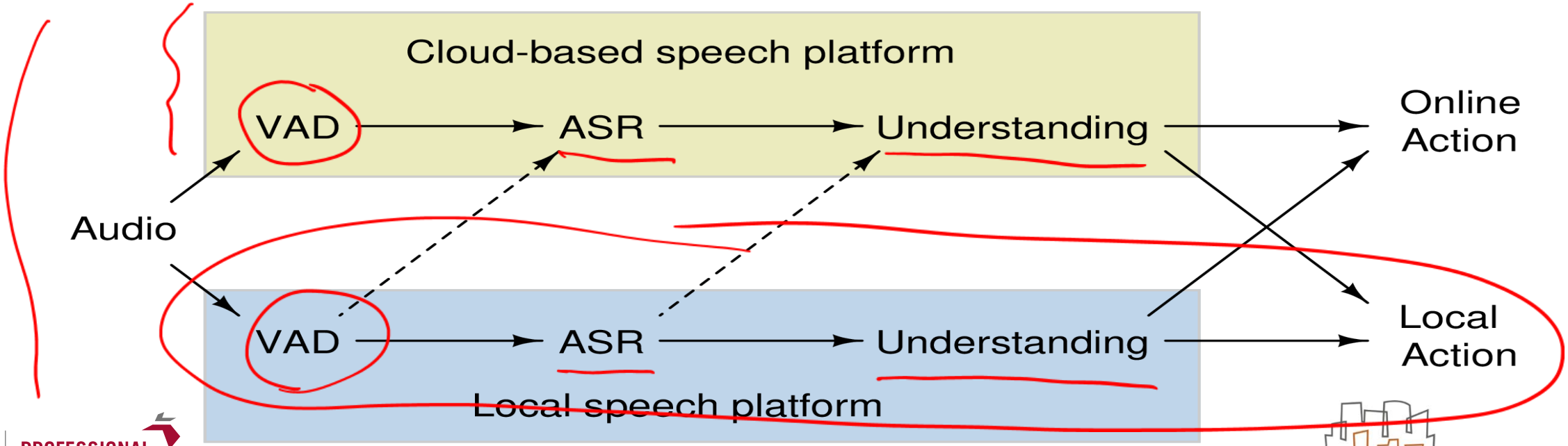


- **Information from different modalities is complementary**
  - State-of-the-art face recognition and speaker verification can jointly reduce task error rates by an order of magnitude

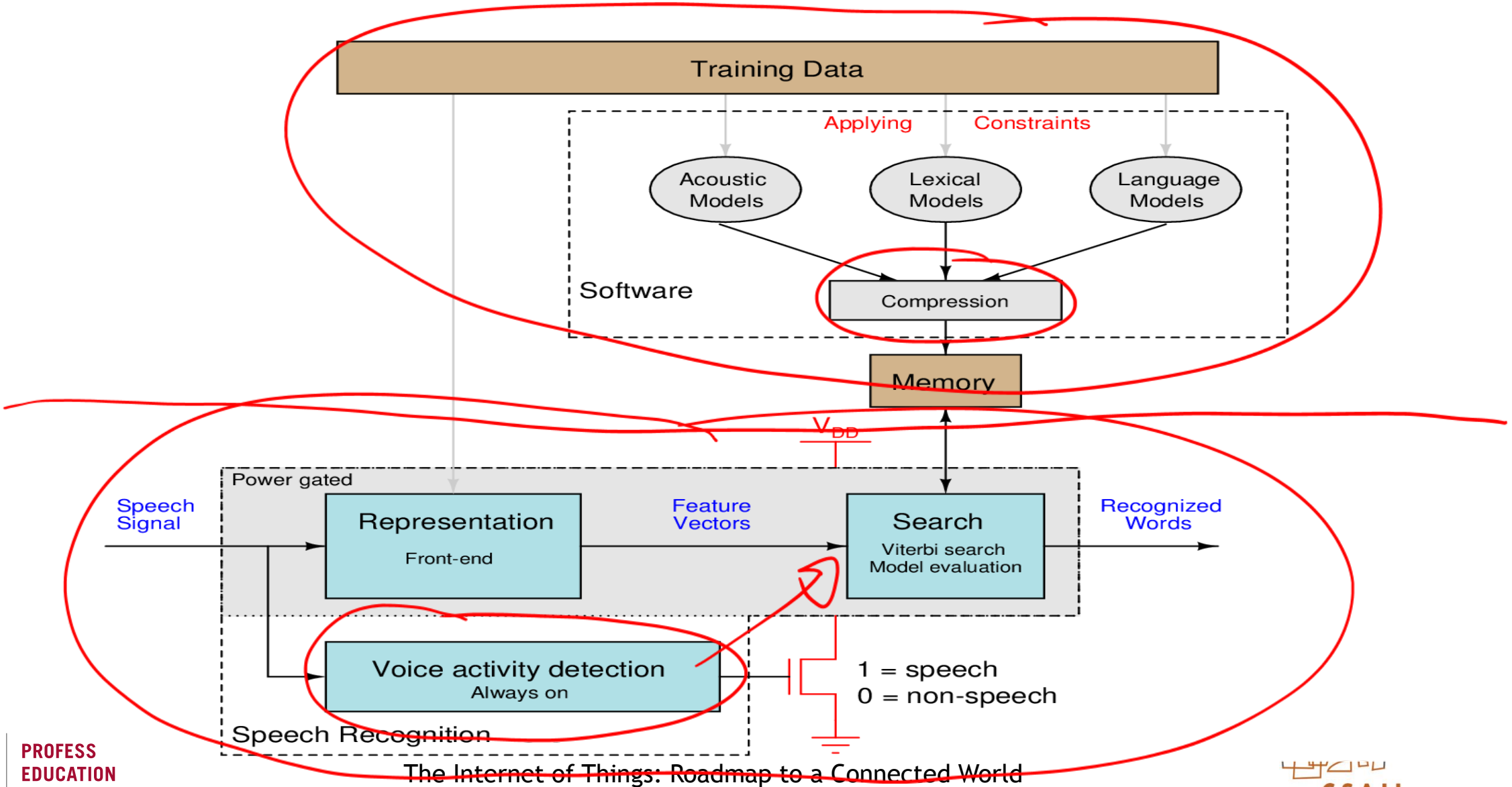
COMPUTATION

LOW-POWER LANGUAGE PROCESSING

- **Spoken interaction with low-power devices**
  - e.g., self-powered, wearable watch, bracelet, etc
- **Some operations can be performed locally**
  - e.g., Voice activity detection (VAD), automatic speech recognition (ASR)

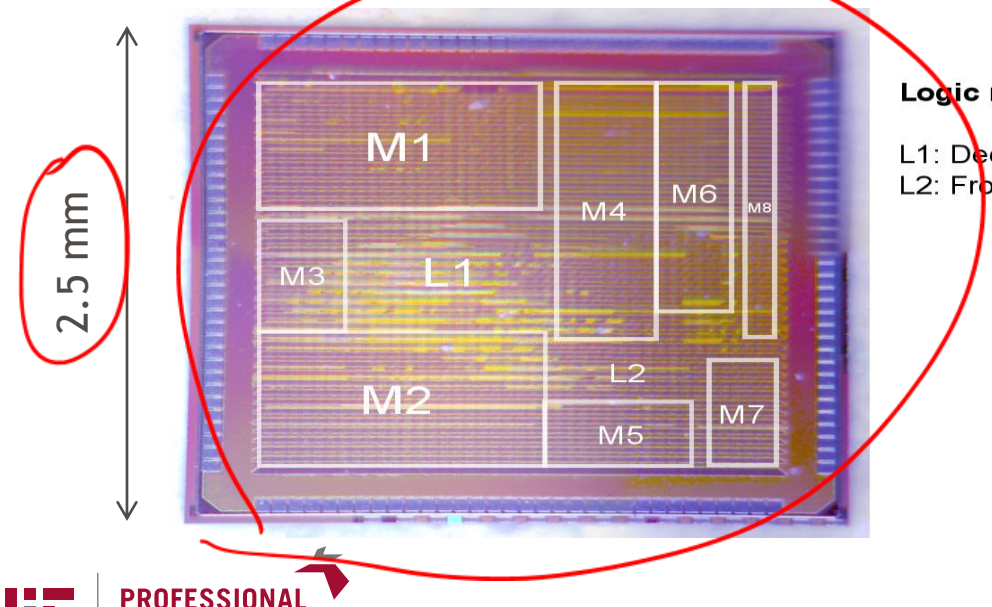


SPEECH RECOGNITION ON A CHIP



ULTRA LOW-POWER SPEECH PROCESSING

- **Goal: Enable spoken interaction with low-power devices**
  - Leverage recent advances in ultra low-power circuit design
- **Prototype examples:**
  - Chip implementations of speech recognizer & voice activity detection
  - Ongoing research into chip implementation of speaker verification

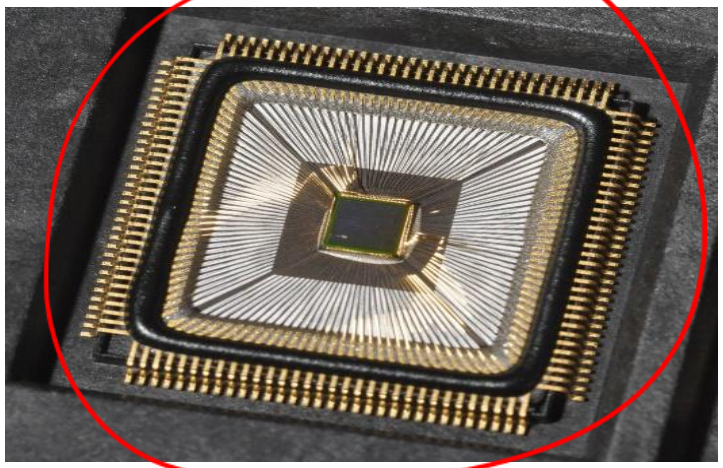


Logic regions

L1: Decoder
L2: Frontend

Memories

M1: Active state list 1
M2: Active state list 2
M3: GMM quantization tables and cache
M4: WFST cache data table
M5: Feature vector buffer
M6: WFST cache hash table
M7: Feature and audio log
M8: Frontend and FFT scratch memories

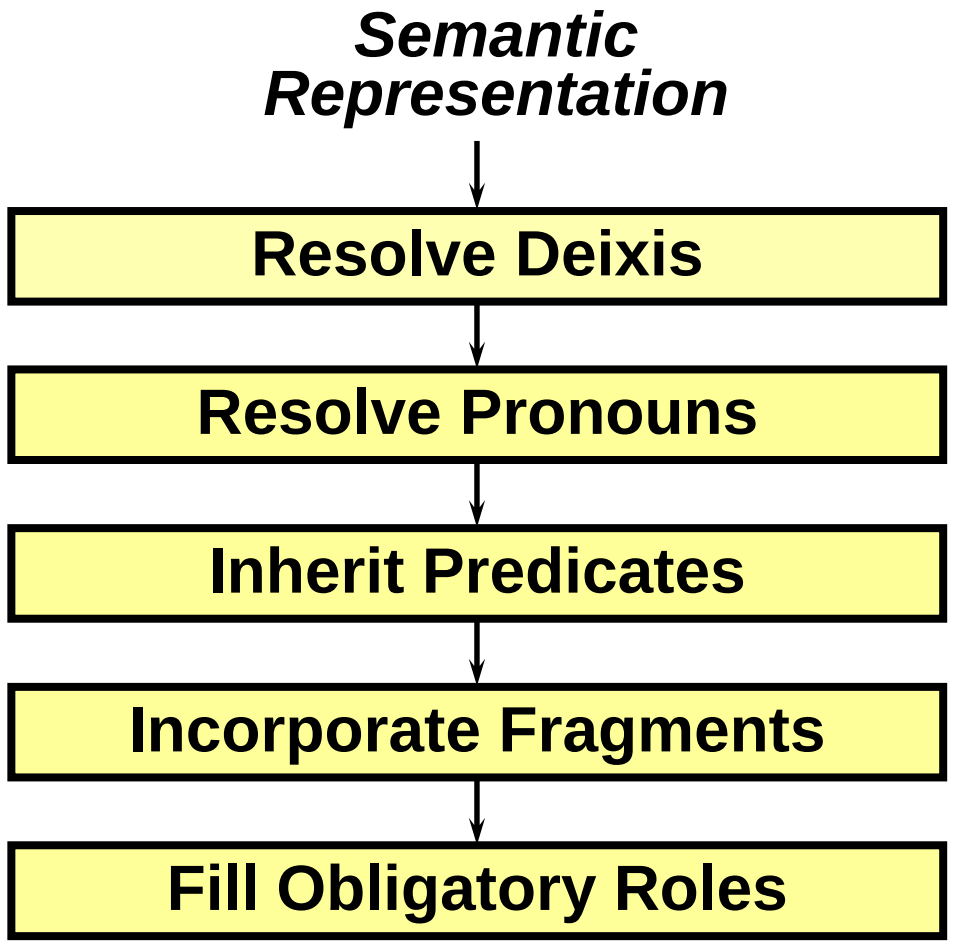


CHALLENGES

ADVANCED MULTI-MODAL INTERACTION

- **Enable natural, flexible, robust human-computer interaction**
- **Interact with people much like they interact with each other**
- **Integrate many technologies to understand context:**
  - what's going on (e.g., activity detection)
  - who is there (e.g., person identification/verification)
  - who is talking (e.g., audio/visual sound localization)
  - what they are saying (e.g., audio/visual speech recognition)
  - what they are doing (e.g., writing, gaze, pose, gesture)
  - what they are intending (e.g., multi-modal understanding)
  - how they are feeling (e.g., emotion, cognitive state)

CONTEXTUAL UNDERSTANDING



"Flights from Frankfurt to Boston"

"When does *this one* leave?"

"What meal does *it* serve?"

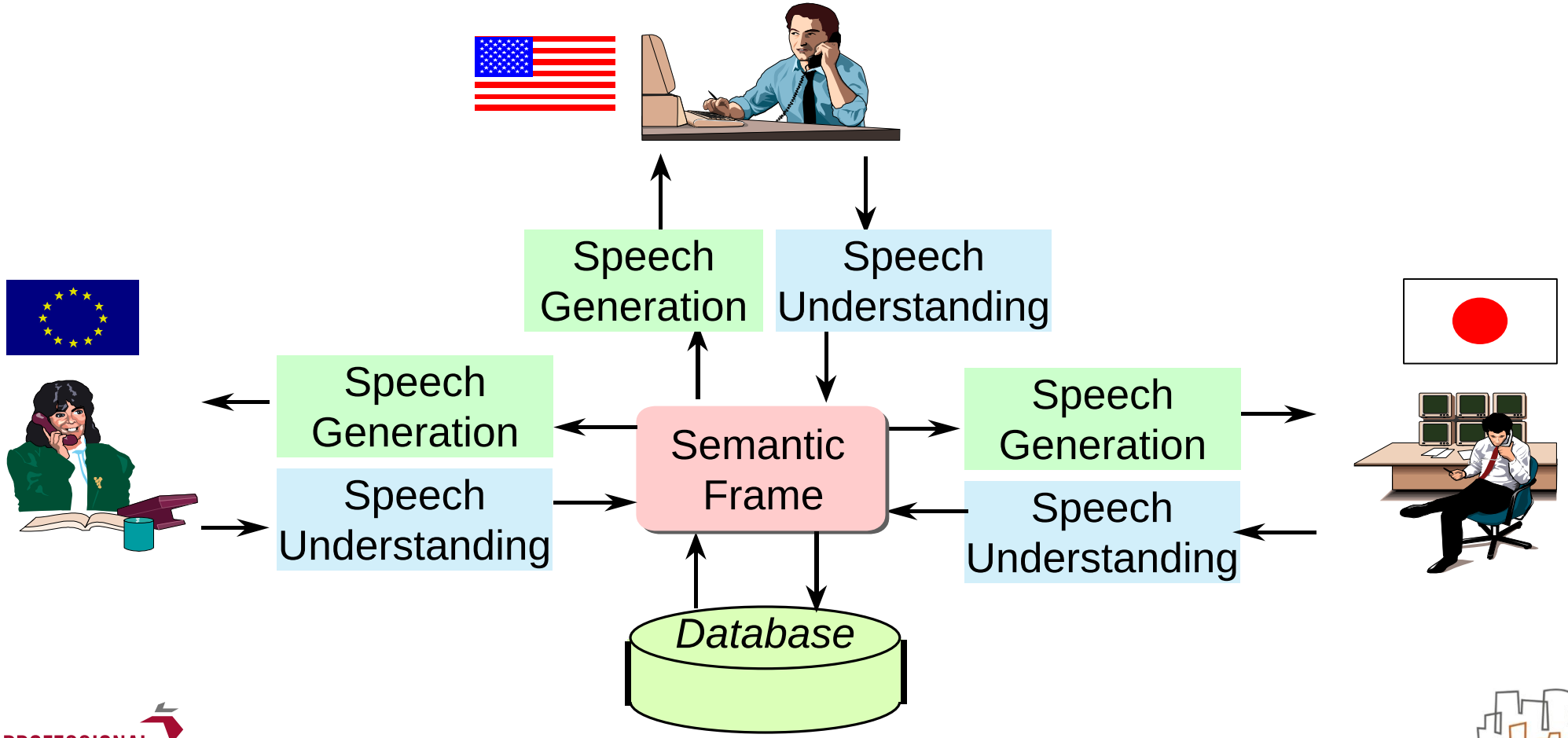
"Show me the ones *on United*"

"What about *Lufthansa*?"

"Give me flights to New York"

MULTI-LINGUAL INTERACTION

- **An *interlingua* approach is useful for multilingual HCI**



FINAL MESSAGE

- **Speech-based interfaces for IOT are inevitable, driven by**
  - The need for mobility and connectivity
  - The miniaturization of devices
  - Humans' innate ability and desire to speak
- **Tomorrow's interfaces must**
  - Be increasingly untethered and robust to different environments
  - Understand and respond in context
  - Incorporate multiple modalities and languages
- **Recent speech interfaces are just the beginning**
- **Many challenges remain!**

THANK YOU!

Jim Glass

Senior Research Scientist

Computer Science and Artificial Intelligence Laboratory (CSAIL)
Massachusetts Institute of Technology