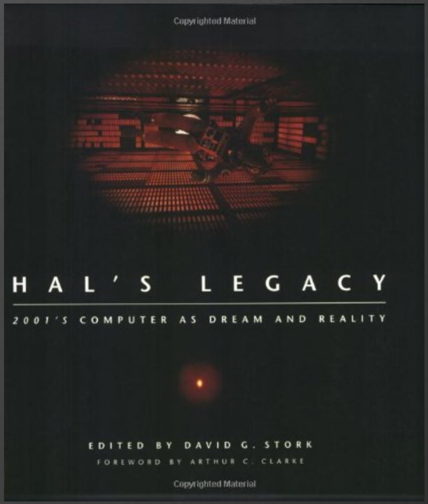
Open the pod bay doors…Siri!

Roberto Pieraccini
CEO
ICSI, The International Computer Science Institute
Berkeley

2001 a space Odyssey
The visionary talent of Clarke and Kubrick

A 1960s prediction of the technology of 40 years later



2001 a space Odyssey
The visionary talent of Clarke and Kubrick

A 1960s prediction of the technology of 40 years later

Computer chess

Audio-video communication

On board entertainment

Computer graphics

Tablet devices

…and bad airline food

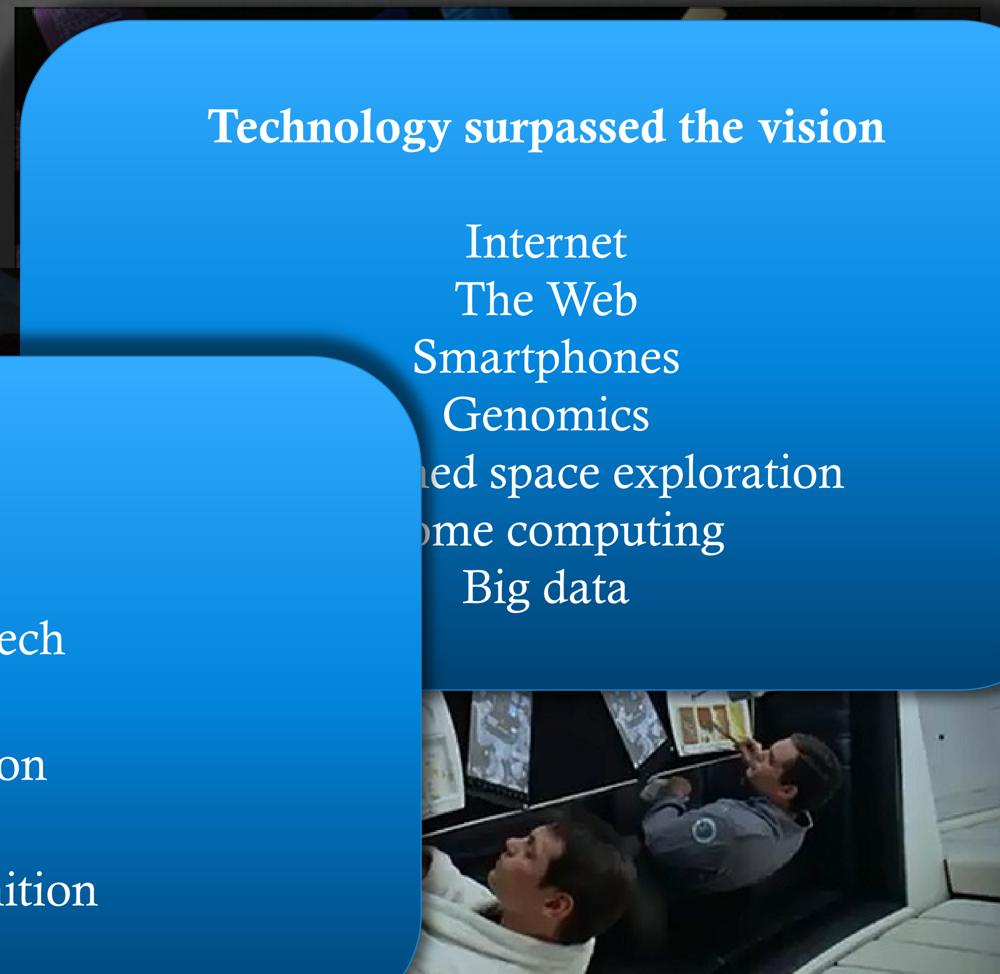
2001 a space Odyssey
The visionary talent of Clarke and Kubrick

A 1960s prediction of the technology of 40 years later

Computer chess

Audio-video communication

On board entertainment



Technology surpassed the vision

Internet
The Web
Smartphones
Genomics
ed space exploration
ome computing
Big data

Except for

Computer Speech

Computer Vision

Computer cognition

Speech technology in 2001: the vision

Speech technology in 2001: the reality

1997, 2003, 2009 Speech Recognition Prediction Surveys

Roger K. Moore, Interspeech 2011

In which year each one of these statements will be true

1. More than 50% of new PCs have dictation on them, either at purchase or shortly after.
2. Most telephone Interactive Voice Response systems accept speech input (and more than just digits).
3. TV closed captioning is automatic and pervasive.
4. Voice recognition is commonly available at home (e.g. interactive TV, control of home appliances and home management systems).
5. Automatic airline reservation by voice over the telephone is the norm.
6. It is possible to hold a telephone conversation with an automatic chat-line system for more than 10 minutes without realizing it isn't human.
7. Voice-enabled command, control and communication in cars becomes as common as intermittent wiper, power window or power door lock.
8. No more need for speech research.
9. A leading cause of time away from work is being hoarse from talking all the time, and people buy keyboards as an alternative to speaking.
10. Public proceedings (e.g. courts, public inquiries, parliament etc.) are transcribed automatically.
11. First legal case in which a recording of a person's voice is thrown out because it cannot be proved whether a computer or a person said it.
12. Speech recognition accuracy equals that of the average (individual) human transcriber.

NEVER!

20+ years from now
….always

1997, 2003, 2009 Speech Recognition Prediction Surveys

Roger K. Moore, Interspeech 2011

Bill Gates, 1 October 1997: "**In this 10-year time** fram key tim spe sta

Bill Gates, 28 July 2003: "It's all the dreams of software, of vision and speech recognition and busi **Som** **migh** and l machine lea things, I think make some in

Bill Gates, 25 February 2004: "Now, with speech it's not as easy. Speech is another one that will be solved, and w **withi** peop **fram** And s handy even ha it can se those things **undoubtedly will be solved in the next decade.**"

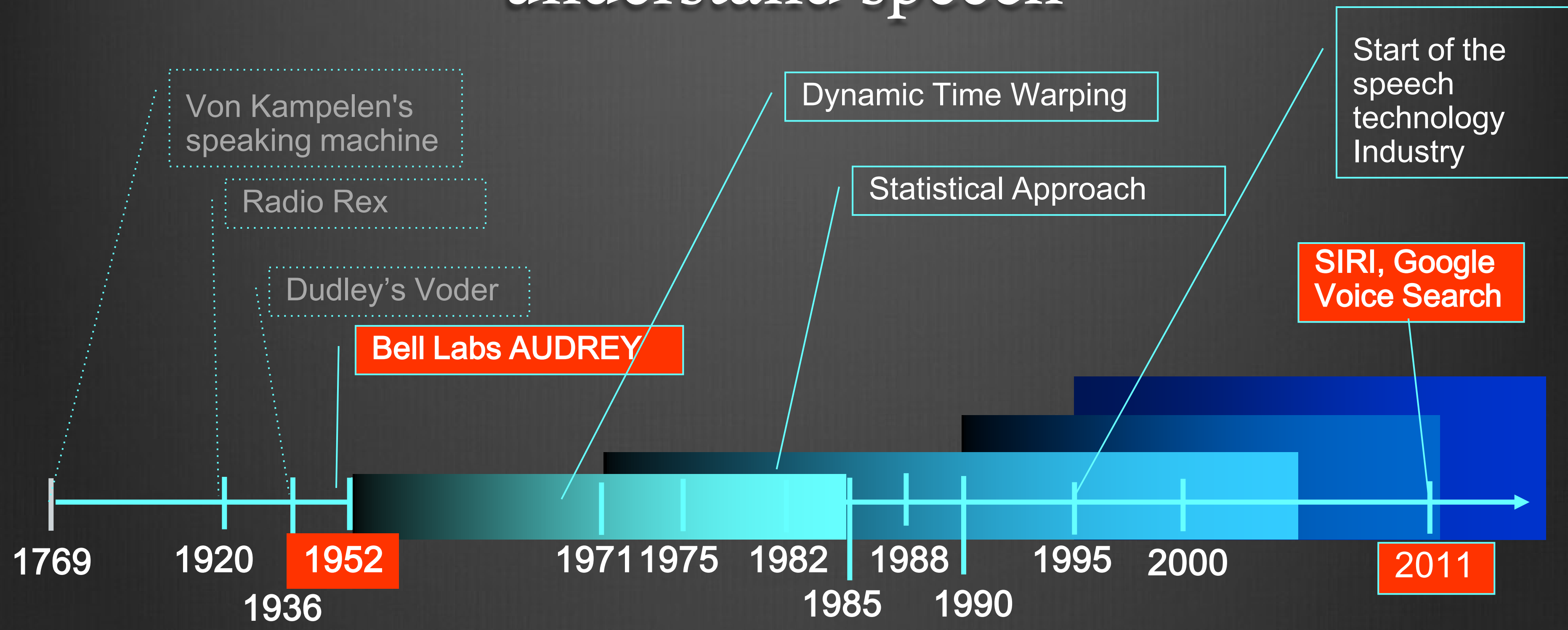
Bill Gates, 14 September 2005: "We totally believe speech recognition will go mainstream somewhere **ov**

Bill Gates, 14 October 2005: "Another big change you'll see is that we'll have microphones on PCs and th fe **n**

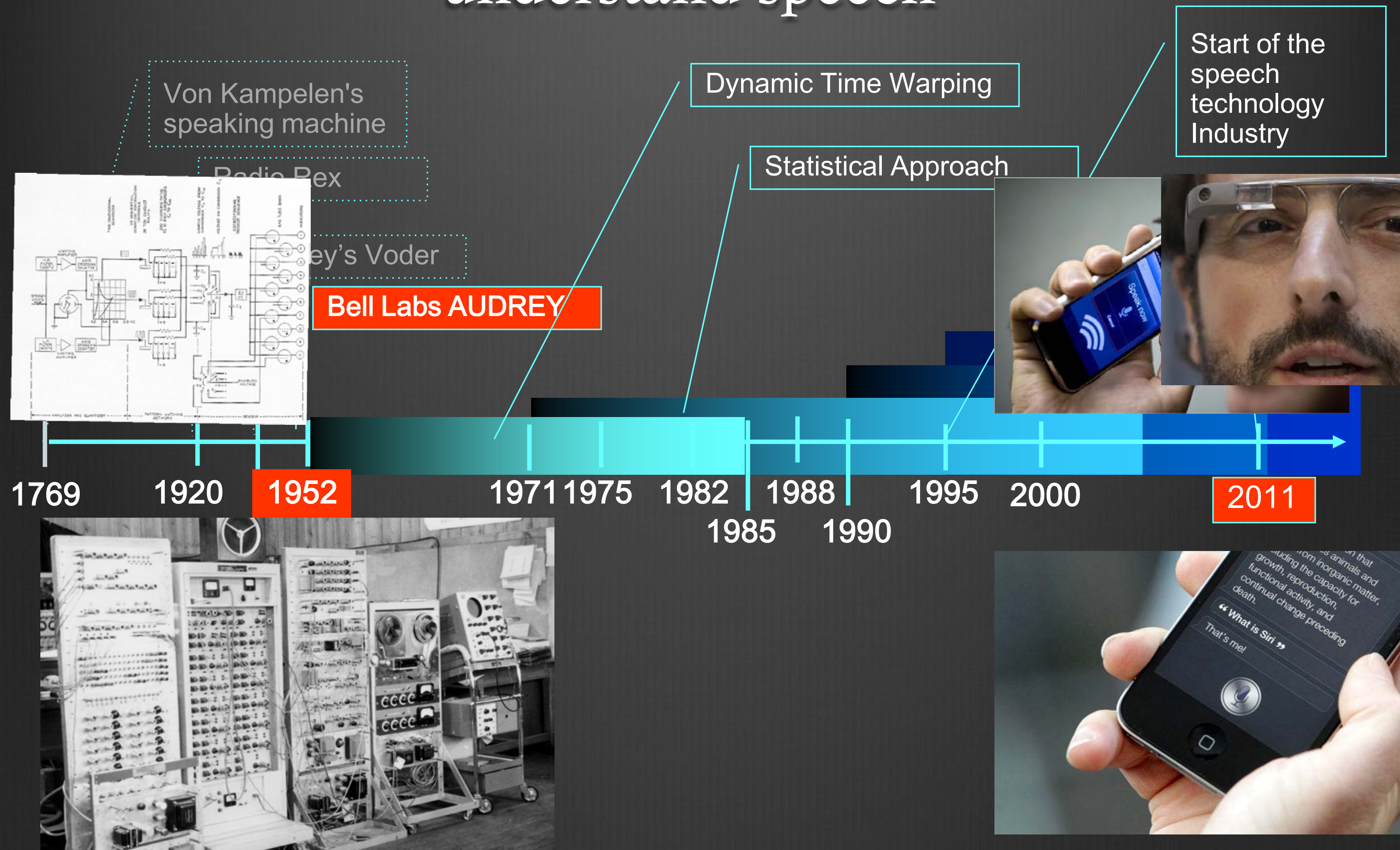
Bill Gates, 9 June 2011: "**The next big thing** is definitely speech and voice recognition. You'll be able to touch that board or speak to it and get your message to colleagues around the world. Screens are cheap."

8. No
9. A leading cau talking all the speaking.
10. Public procee are transcribe
11. First legal case out because it said it.
12. Speech recognition accuracy equals that of the average (individual) human transcriber.

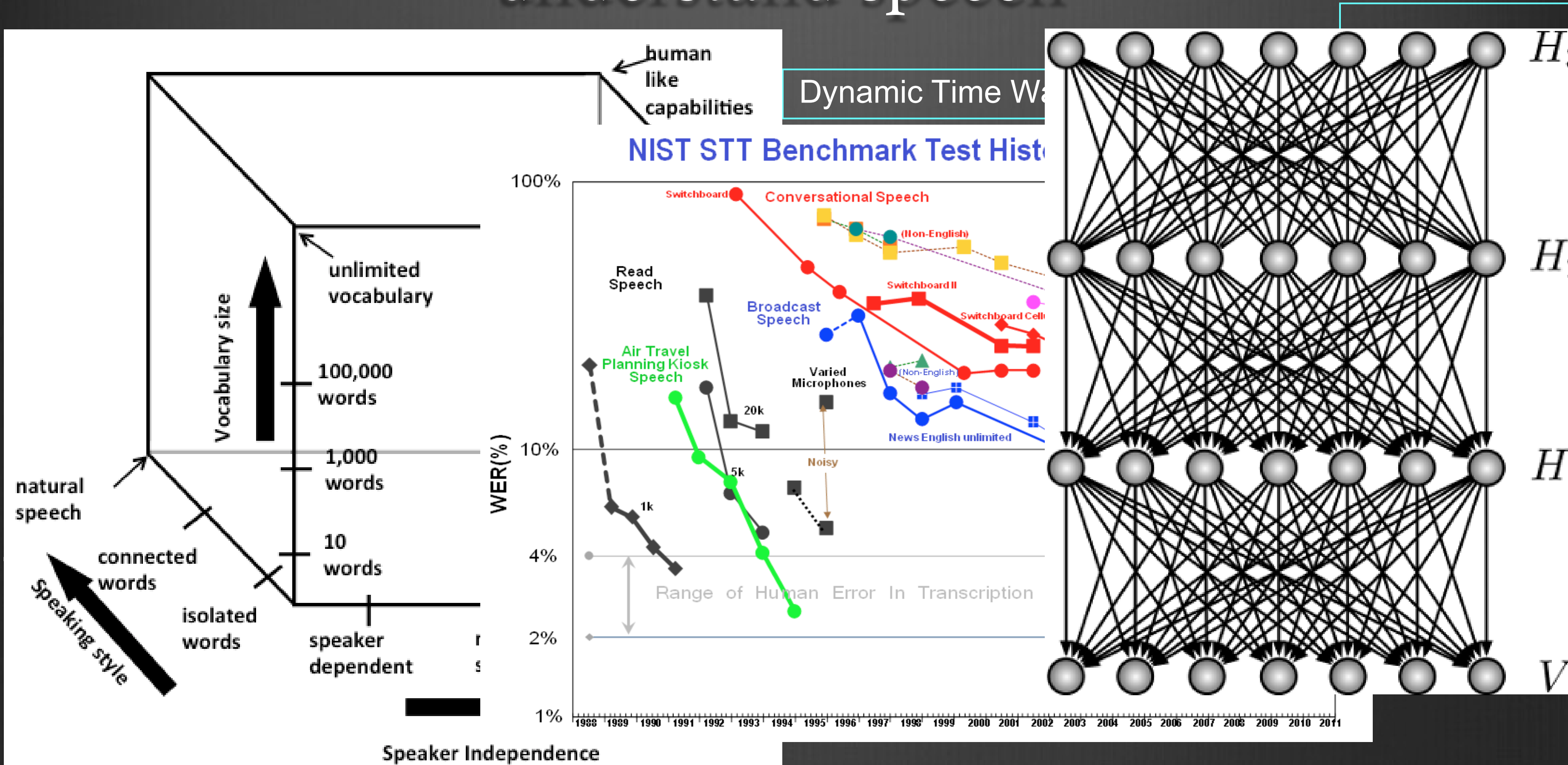
The 60 years journey of computers that understand speech



The 60 years journey of computers that understand speech



The 60 years journey of computers that understand speech

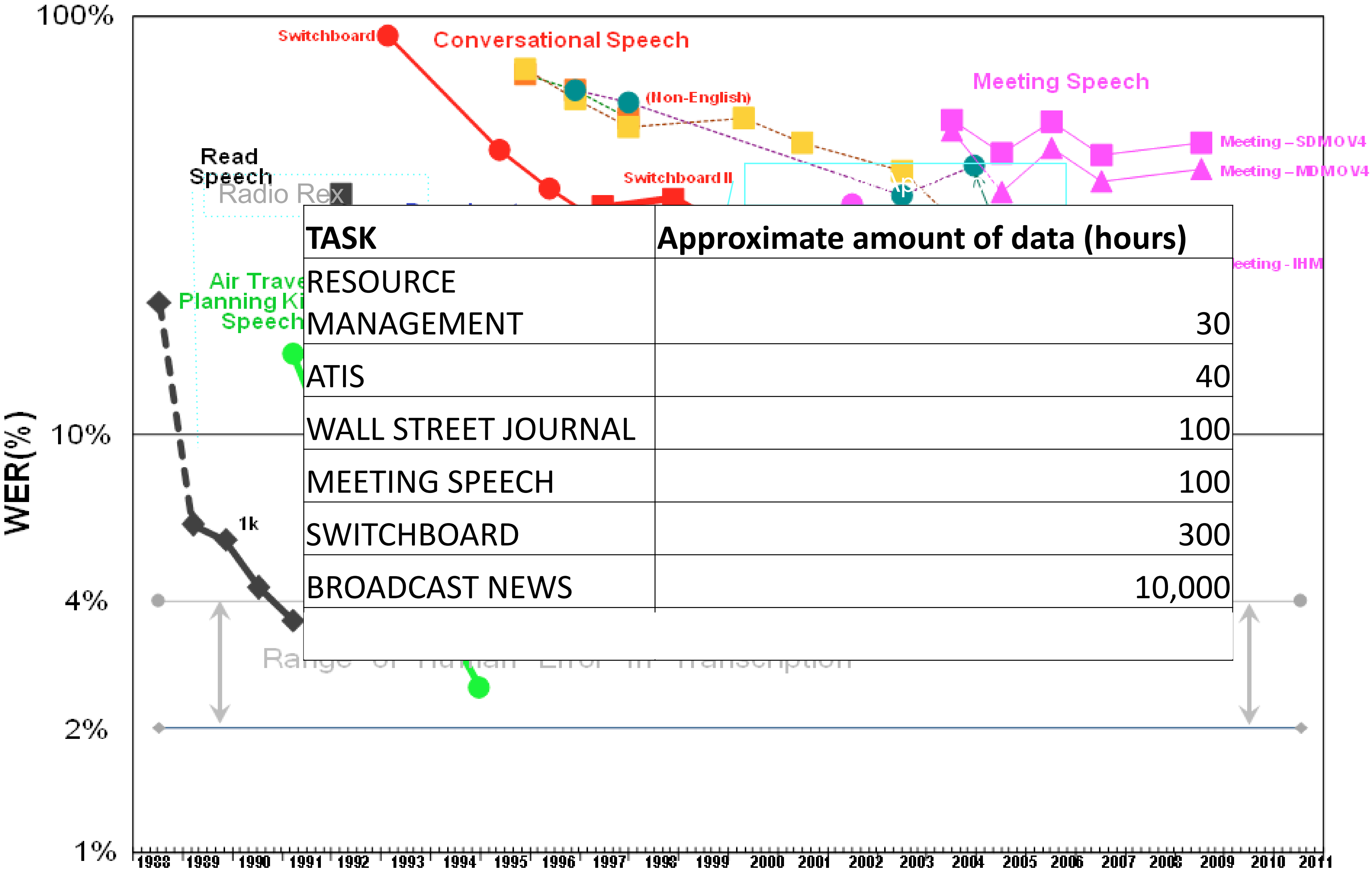


1980s: STRUGGLE FOR CAPABILITIES

1990s: STRUGGLE FOR ACCURACY

2000s: STRUGGLE FOR ROBUSTNESS

NIST STT Benchmark Test History – May. '09



TASK	Approximate amount of data (hours)
RESOURCE MANAGEMENT	30
ATIS	40
WALL STREET JOURNAL	100
MEETING SPEECH	100
SWITCHBOARD	300
BROADCAST NEWS	10,000

So … why is ASR so difficult?

- Humans use speech in the most adverse conditions, and yet they can communicate
- They expect machines to be able to do the same
- ASR technology is very good today, but still very brittle when
  - Noise is high
  - Accent is strong
  - Microphone is far
  - Words are unknown
  - Voice is mixed with sounds, music, or other voices
  - Speaker is a goat

Speech recognition in noise

Digit recognition accuracy (AURORA-2)

Noise level

SNR/dB	Restaurant	Street	Airport	Train-station	Average
clean	98.68	98.52	98.39	98.49	98.52
20	96.87	97.58	97.44	97.01	97.22
15	95.30	96.31	96.12	95.53	95.81
10	91.96	94.35	93.29	92.87	93.11
5	83.54	85.61	86.25	83.52	84.73
0	59.29	61.34	65.11	56.12	60.46
-5	25.51	27.60	29.41	21.07	25.89
Average between 0 and 20dB	85.39	87.03	87.64	85.01	**86.27**

Hirsch, Pearce, ISCA ITRW ASR2000

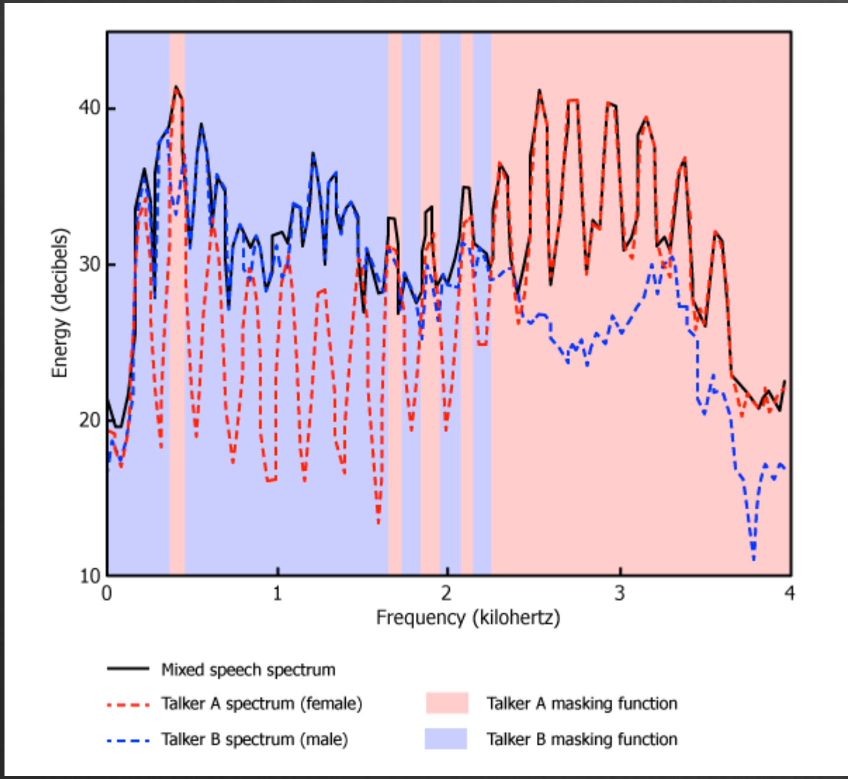
Dealing with unknown words

The cocktail party effect

Source separation today

From: Audio Alchemy: Getting Computers to Understand Overlapping Speech
J. R. Hershey, P. A. Olsen, S. J. Rennie, A. Aaron, Scientific American, April 2011

SPEAKER MASKING ALGORITHM



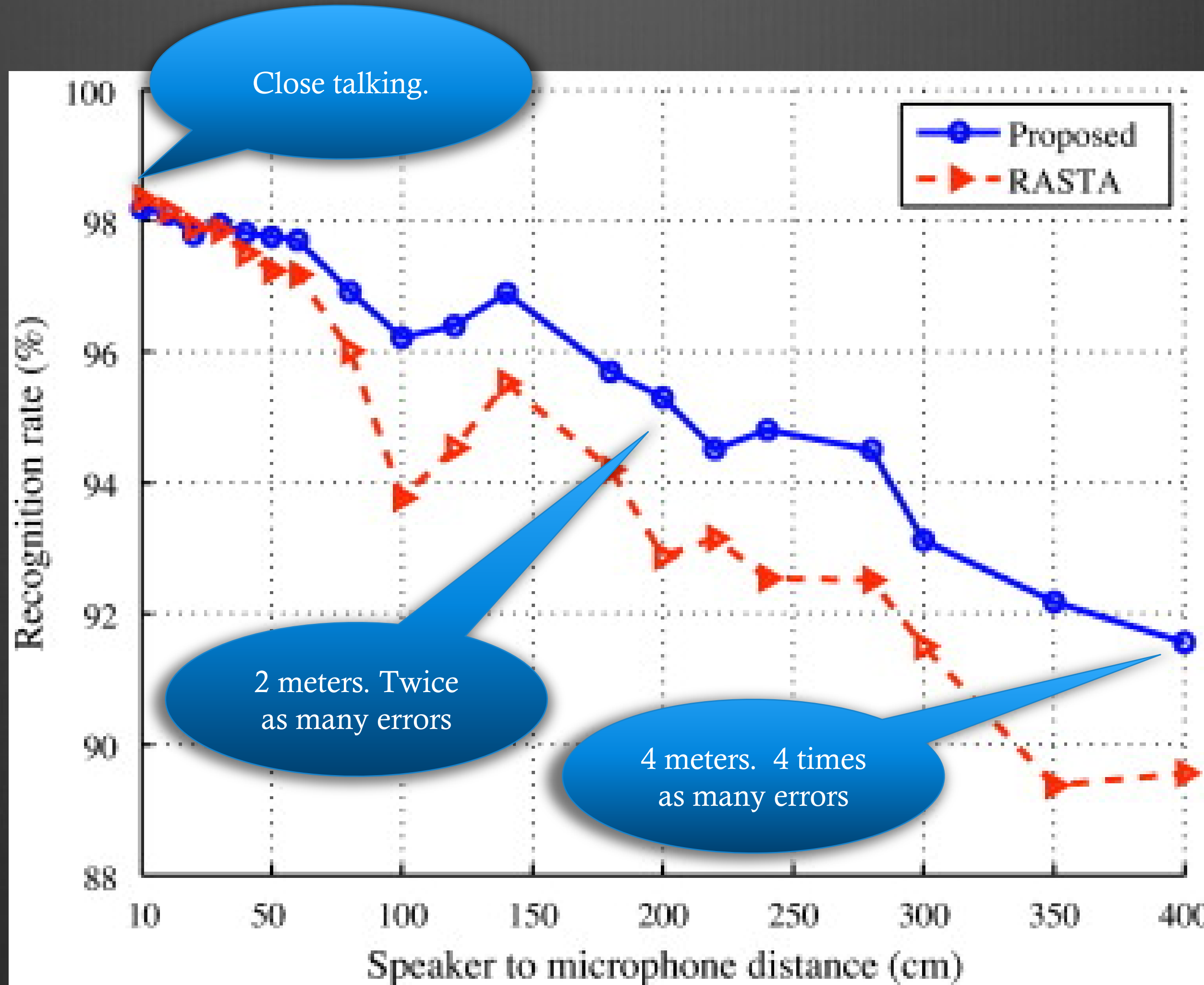
MIXED SPEECH

Speaker 1: Lay white at K 5 again.
Speaker 2: Bin blue by M zero now.
Speaker 3: Set green in M 7 please.
Speaker 4: Lay green with S 7 please

SEPARATION BY SPEAKER MASKING

Speaker 1: Lay white at K 5 again.
Speaker 2: Bin blue by M zero now.
Speaker 3: Set green in M 7 please.
Speaker 4: Lay green with S 7 please

Reverberation



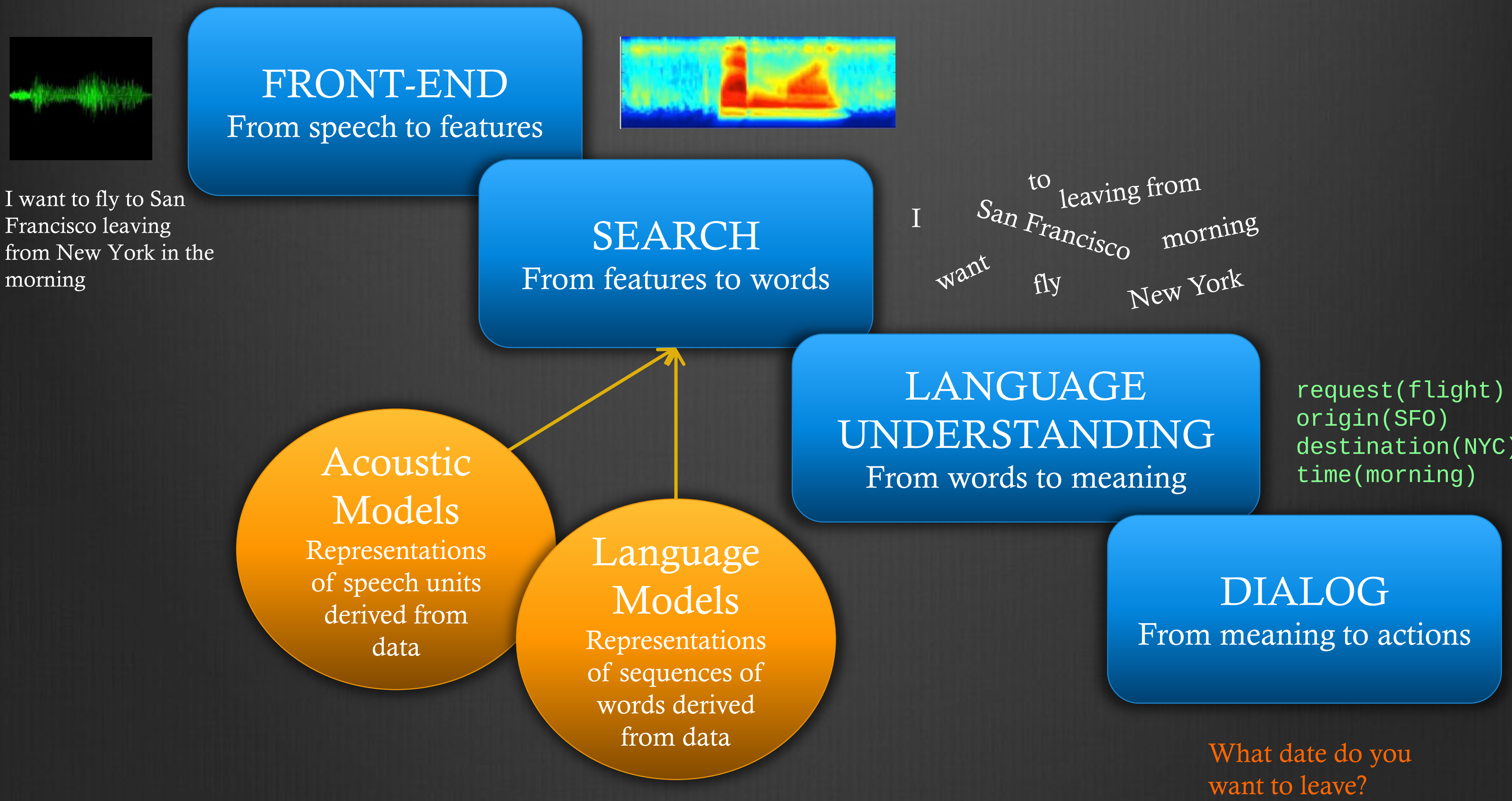
From: *Sub-band temporal modulation envelopes and their normalization for automatic speech recognition in reverberant environments*, X. Lu, M. Unoki, S. Nakamura, Computer Speech and Language, July 2011

Sheep and goats

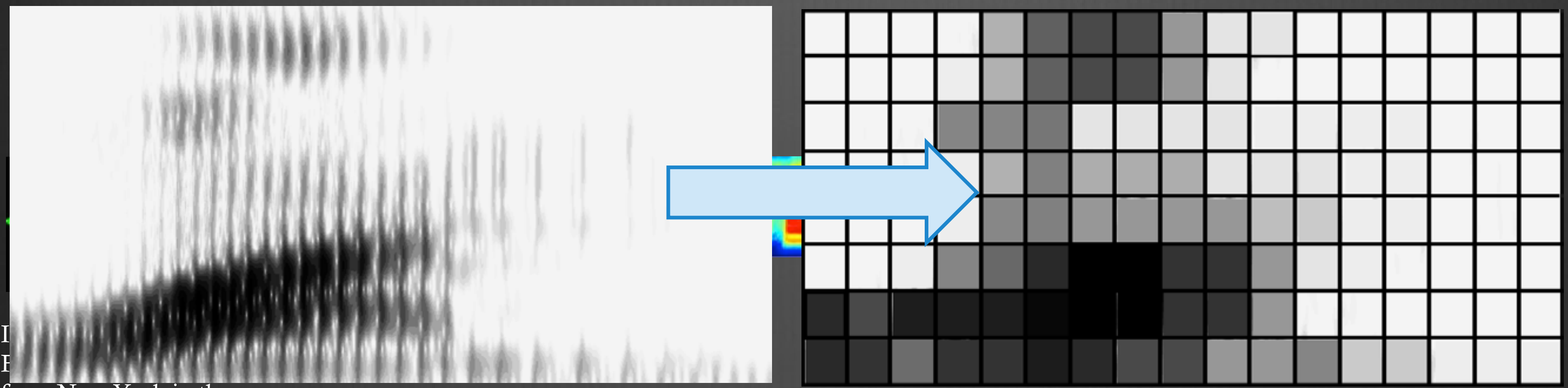
We at Texas Instruments symbolize the skewed distribution [of performance among speakers] by categorizing speakers as either "sheep" or "goats". The sheep, for whom the system works well, comprise the bulk of the population, say 80-90%. But the goats, despite their minority, have the greatest influence on the performance of the system, because most of the recognition errors are attributed to them.

G. R. Doddington, "Whiter Speech Recognition?" in *Trends in Speech Recognition,* Wayne A Lea Editor, Prentice Hall, 1980.

The parts of a speech understanding system



The parts of a speech understanding system



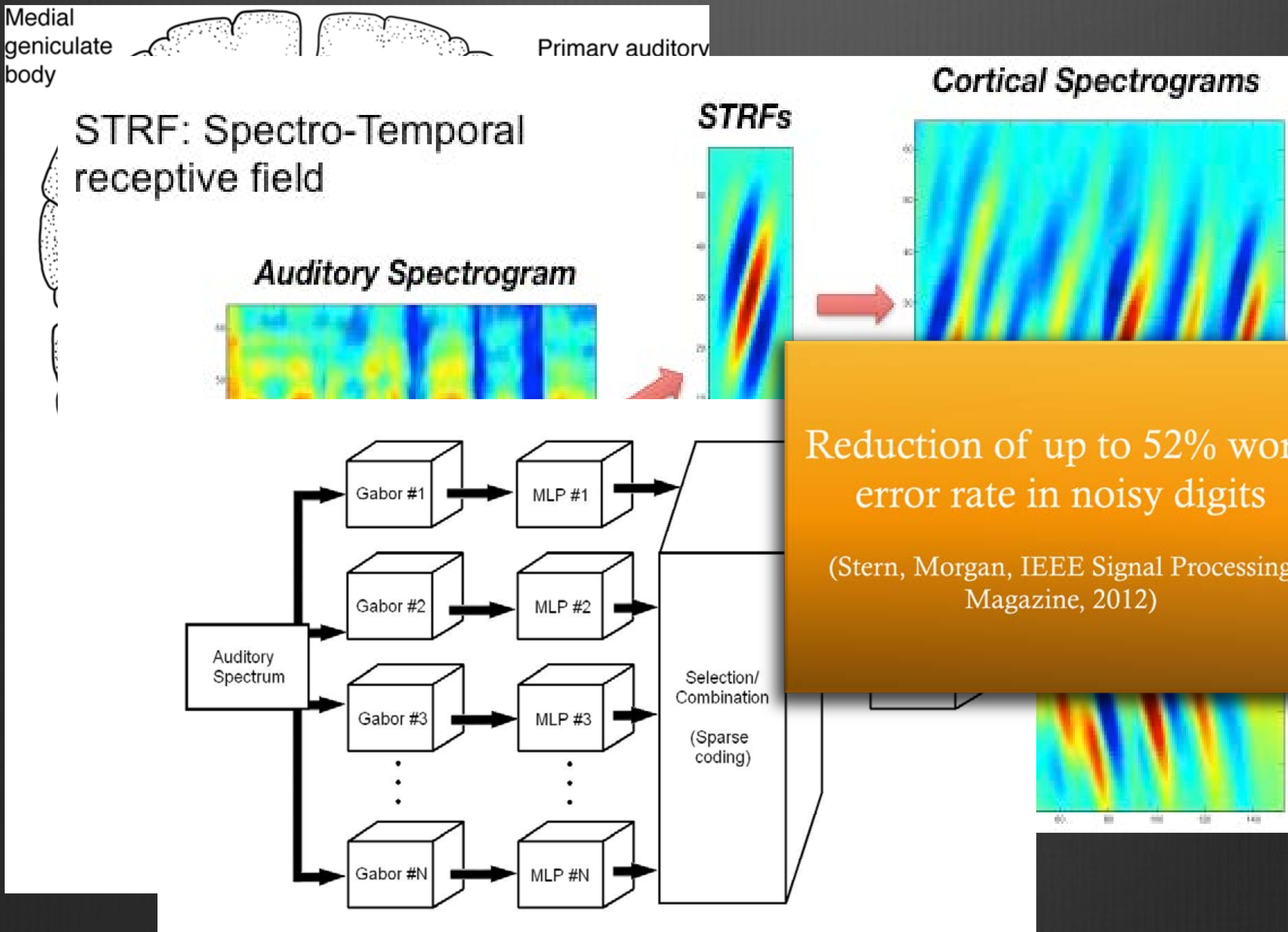
I
F
from New York in the
morning

For decades we have been using variations and transformations of a coarse spectral representation of each 10 msec. segment of speech (frames)

That is called MFCC, or Mel-frequency Cepstral Coefficients

But … what is the human brain using as its front-end? After all the human brain is the best "speech recognizer" we know

Emulating the human brain cortex



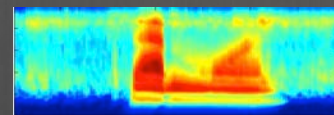
Reduction of up to 52% word error rate in noisy digits

(Stern, Morgan, IEEE Signal Processing Magazine, 2012)

The parts of a speech understanding system

I want to fly to San Francisco leaving from New York in the morning

FRONT-END
From speech to features



SEARCH
From features to words

I want to fly San Francisco leaving from New York morning

LANGUAGE UNDERSTANDING
From words to meaning

```
request(flight)
origin(SFO)
destination(NYC)
time(morning)
```

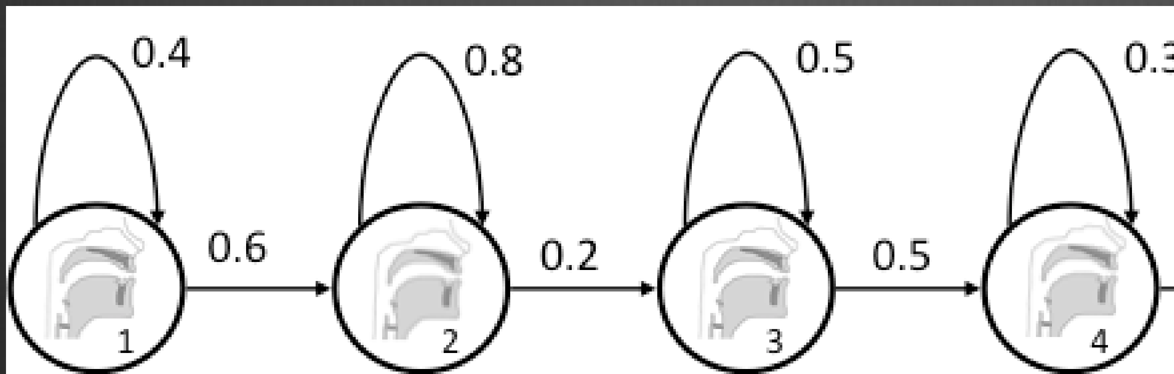
DIALOG
From meaning to actions

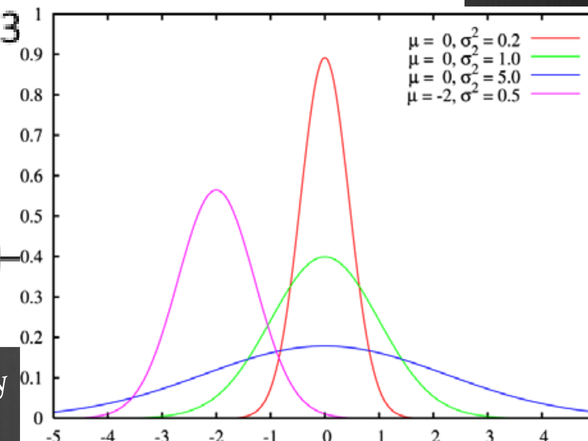
What date do you want to leave?

Acoustic Models
Representations of speech units derived from data

Language Models
Representations of sequences of words derived from data

The parts of a speech understanding system





From features to words

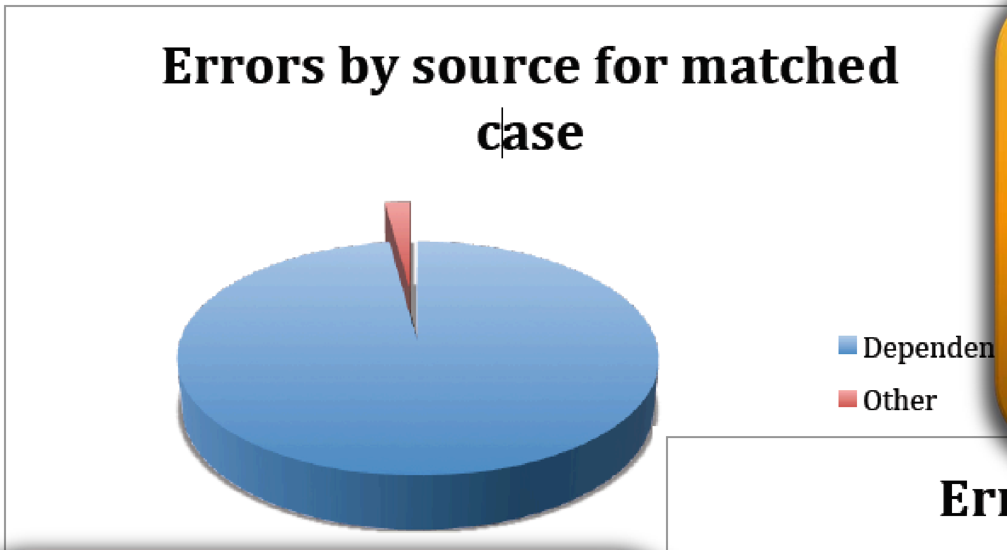
want fly

Since the 1970s, the leading approach to acoustic modeling has been that of Hidden Markov Models (HMM) based on parametric statistical distributions (Gaussian Mixture Models or GMMs)

Both assumptions are know to be wrong with respect to the properties of human speech, but useful to simplify the models so as we could use them.

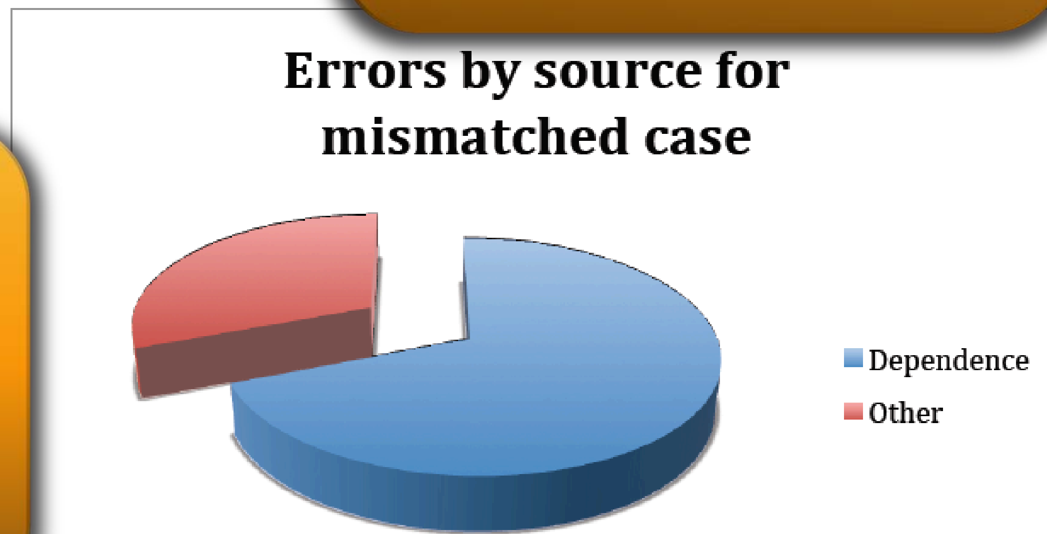
But now with have so much more data and so much computer power that we could try to find better and more fit models

Hidden Markov Models?
The effect of wrong model assumptions



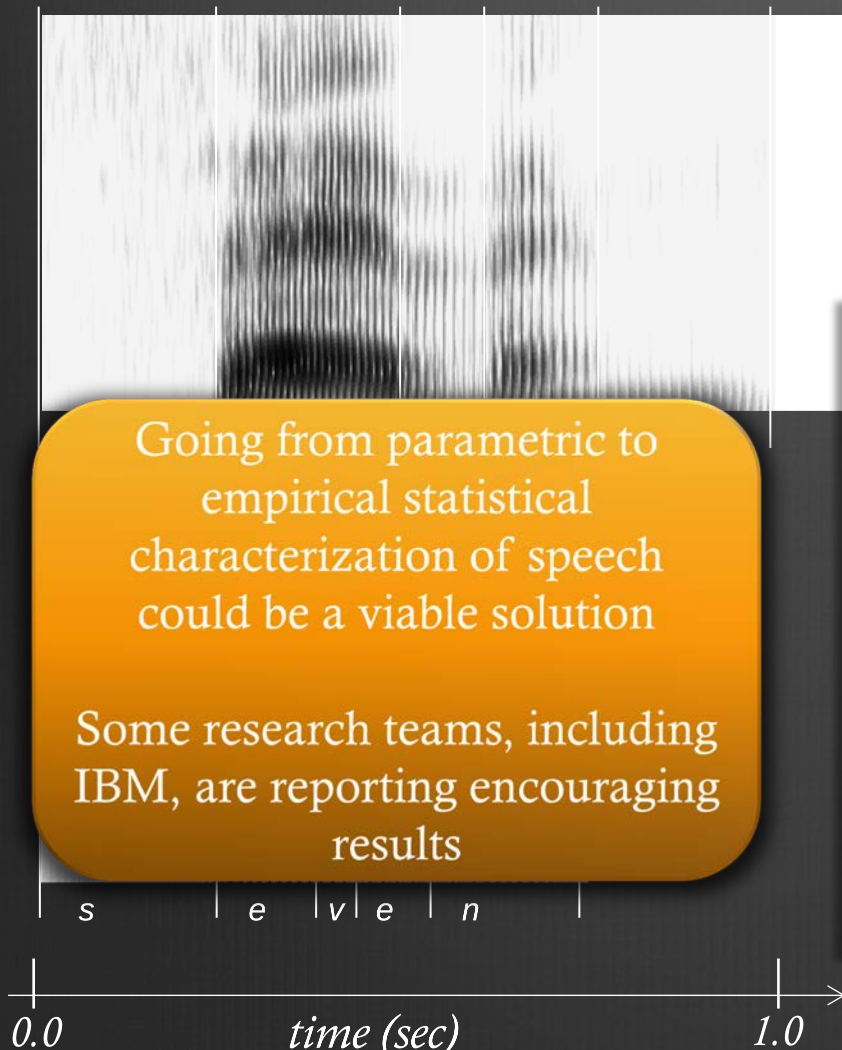
The wrong HMM assumptions account for most of the errors when models are trained in similar (matched) conditions as the test utterances

…and for more than 50% of the errors when the conditions are different (mismatched)



Wegmann, Morgan, Cohen, 2013

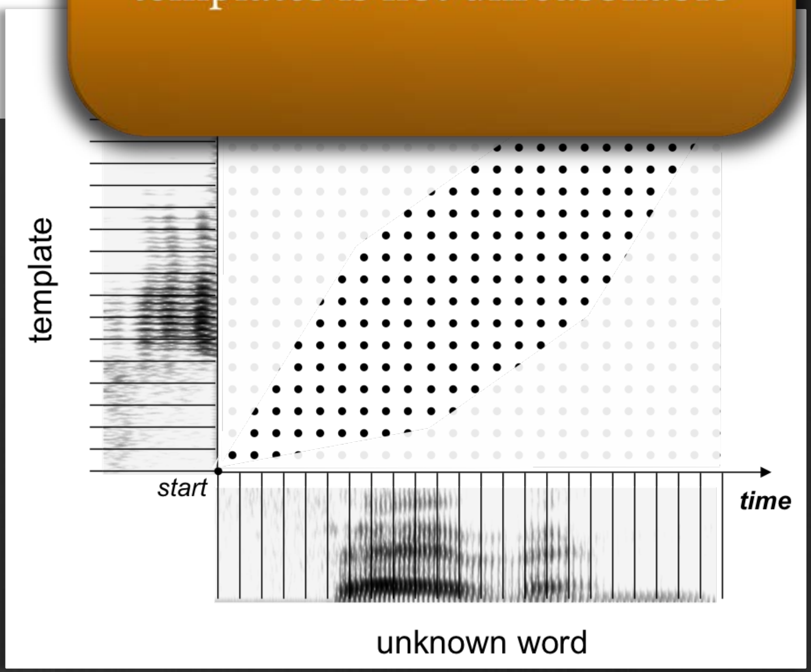
Going back to templates?



Going from parametric to empirical statistical characterization of speech could be a viable solution

Some research teams, including IBM, are reporting encouraging results

Storage is so cheap and computers are so fast that thinking of using millions of templates is not unreasonable

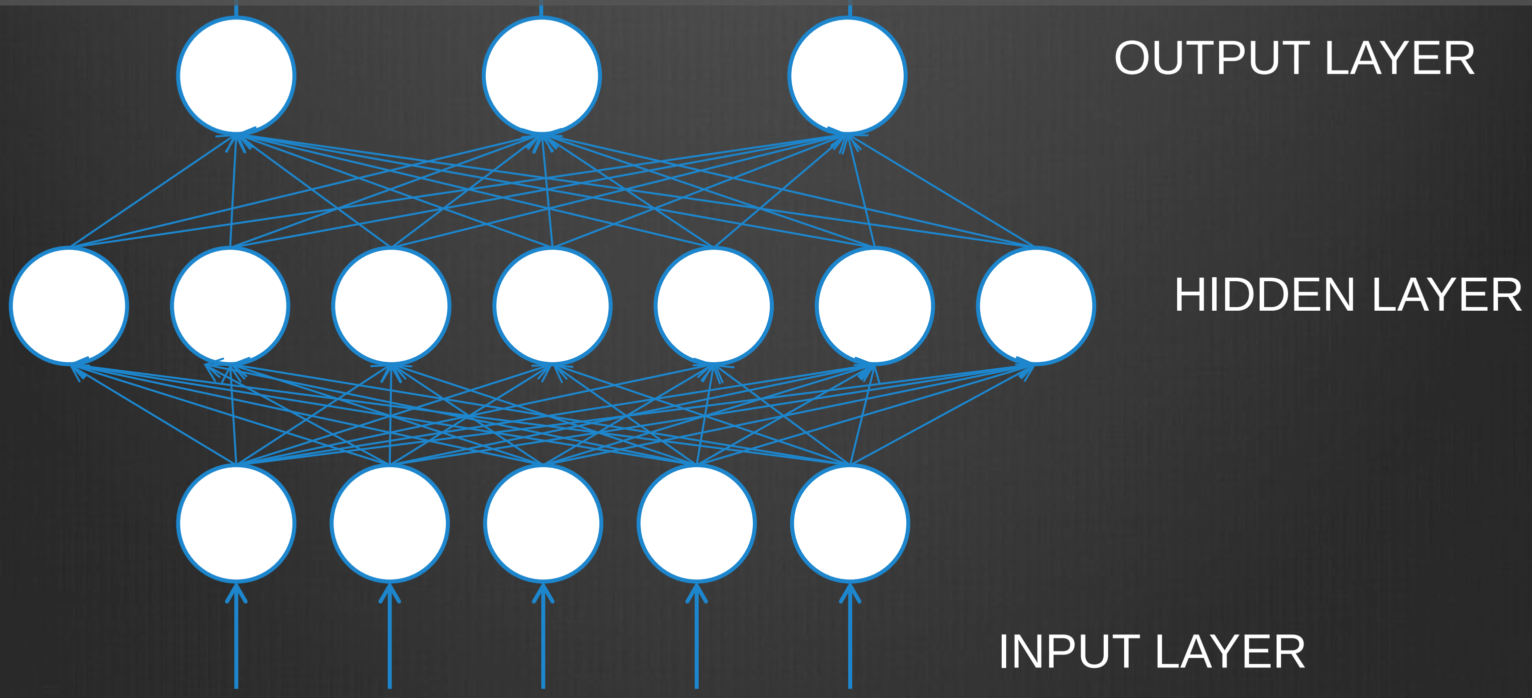


The return of Artificial Neural Networks

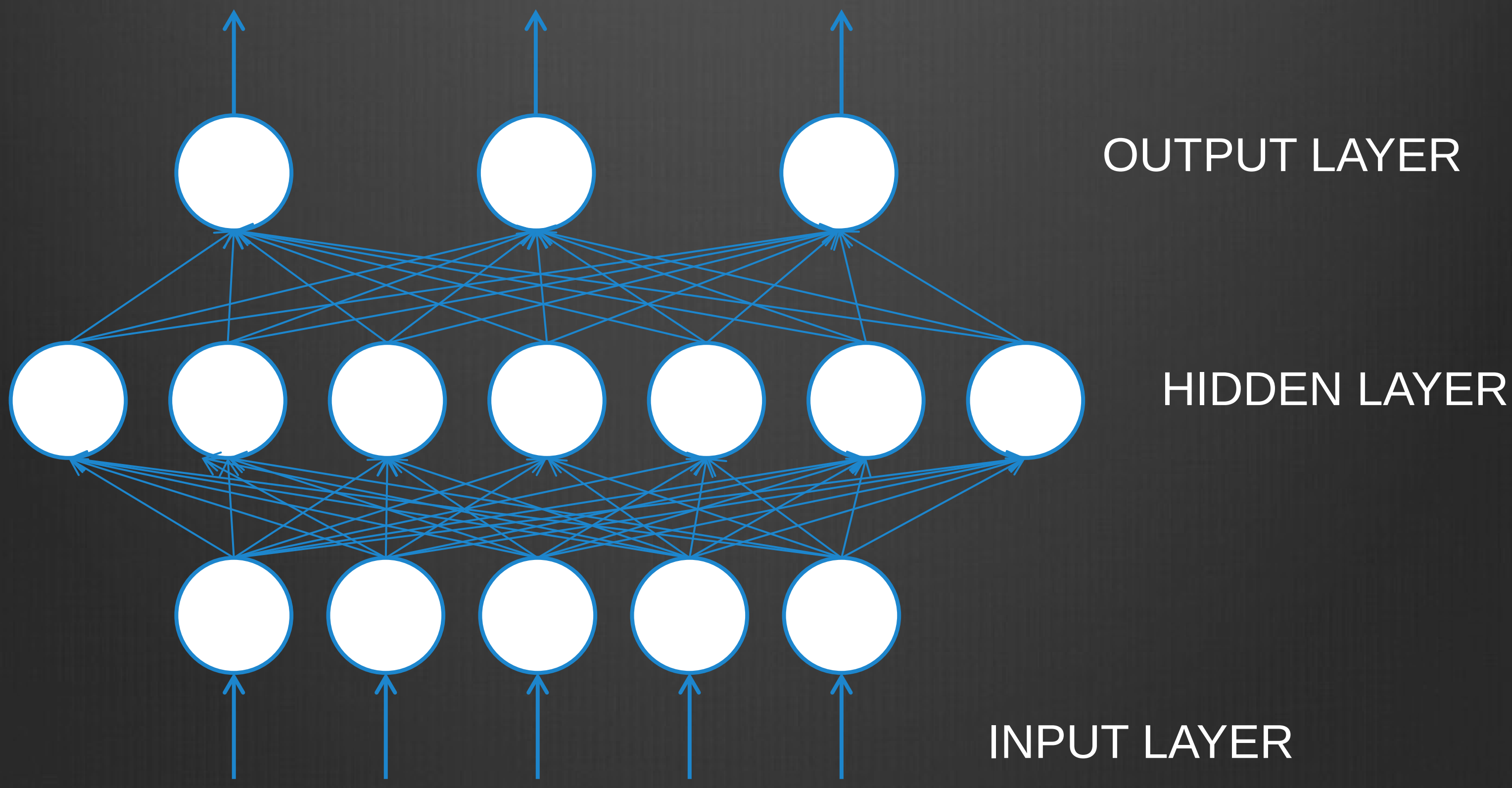
Although many tried to use Artificial Neural Networks as an alternative to Hidden Markov Models, no one could really outperform the mighty HMMs

Through the years, the only successful use of neural network was as probability density estimators in hybrid HMMs

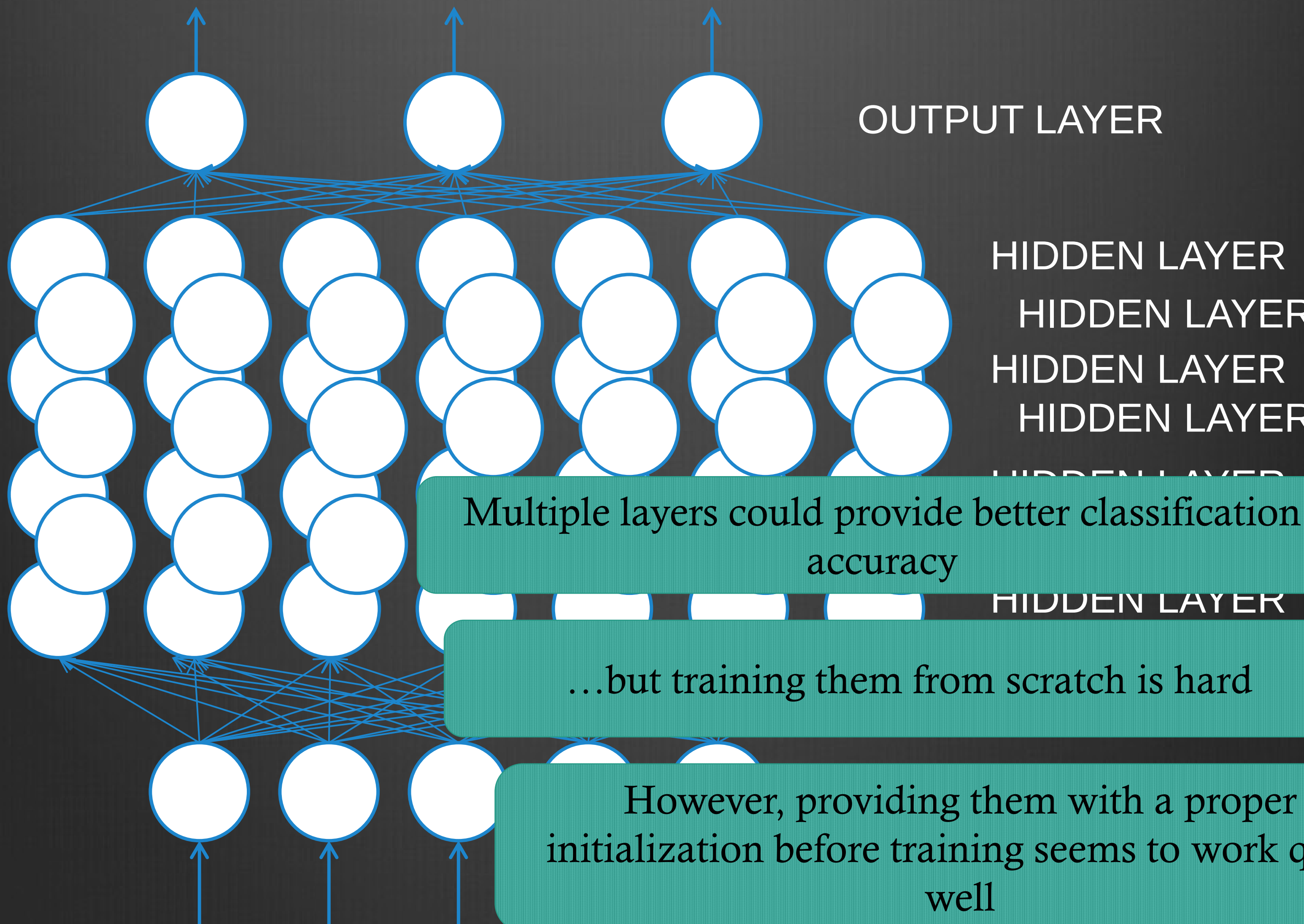
… speech research forgot about them … until recently, when some tried to go deeper … as in DEEP NEURAL NETWORKS



Deep Neural Networks



Deep Neural Networks



Multiple layers could provide better classification accuracy

…but training them from scratch is hard

However, providing them with a proper initialization before training seems to work quite well

Do deep neural networks help speech recognition?

Configuration	Test WER
CD-GMM-HMM (BMMI)	34.8%
2kx5	27.4%
2kx2-(64:64)x1-2kx2	26.8%
2kx4-(64:64)x1	26.4%
2kx4-(96:96)x1	26.2%

Microsoft Research, Switchboard (Dong, Deng, Seide, Interspeech 2012)

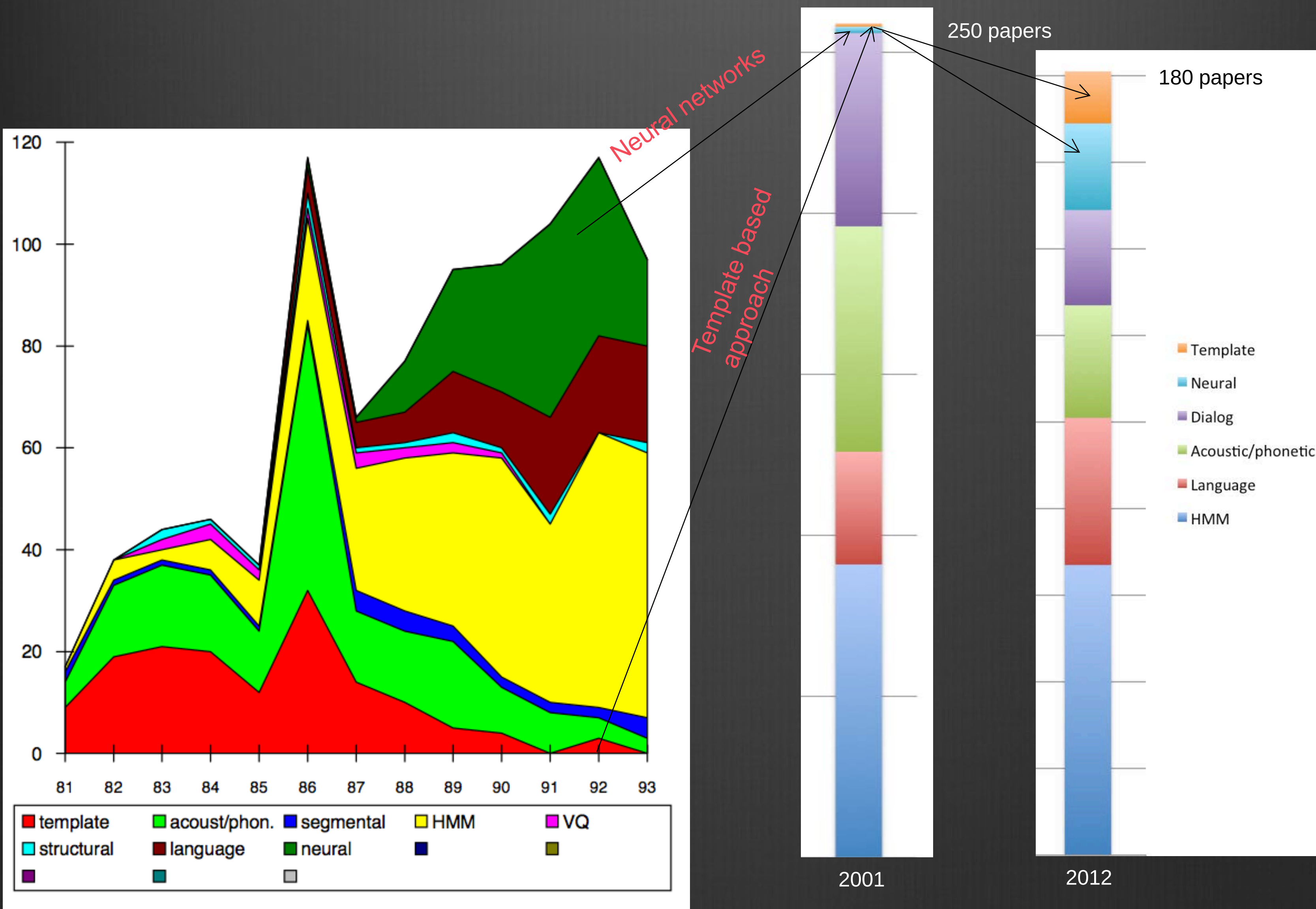
Training	WER	BMMI objfun.
ML	23.6%	0.16
FMMI	20.3%	0.18
FMMI-BMMI	18.7%	0.20

IBM research, broadcast news (Saon, Kingsbury, Interspeech 2012)

Name	Model	WER(%)
Voice Search	GMM-HMM baseline	16.0
	DBN pretrained ANN/HMM with sparsity	12.3
	+ *MMI*	12.2
	+ *system combination with SCARF*	**11.8**
YouTube	GMM-HMM baseline	52.3
	DBN pretrained ANN/HMM with sparsity	47.6
	+ *MMI*	47.1
	+ *system combination with SCARF*	**46.2**

Google (Jaitly et als, Interspeech 2012)

Trends in speech recognition research according to topics in major speech technology conferences



what about language models?

THE HEAD AND IN FRONTAL
AN ENGLISH WRITER THA
CHARACTER OF THIS POI
ANOTHER METHOD FOR
THAT THE TIME OF WHO
PROBLEM FOR AN UNEXP

Claude Shannon, 1947

SEARCH
From features to

Google report that using **19B 4-grams** reduces the error rate of a modest amount as compared with using "only" **14M 3-grams** (from 34.6 to 31.8)

Many have tried to create hybrid models with N-grams and linguistic rules.

But no-one outperformed N-grams so far

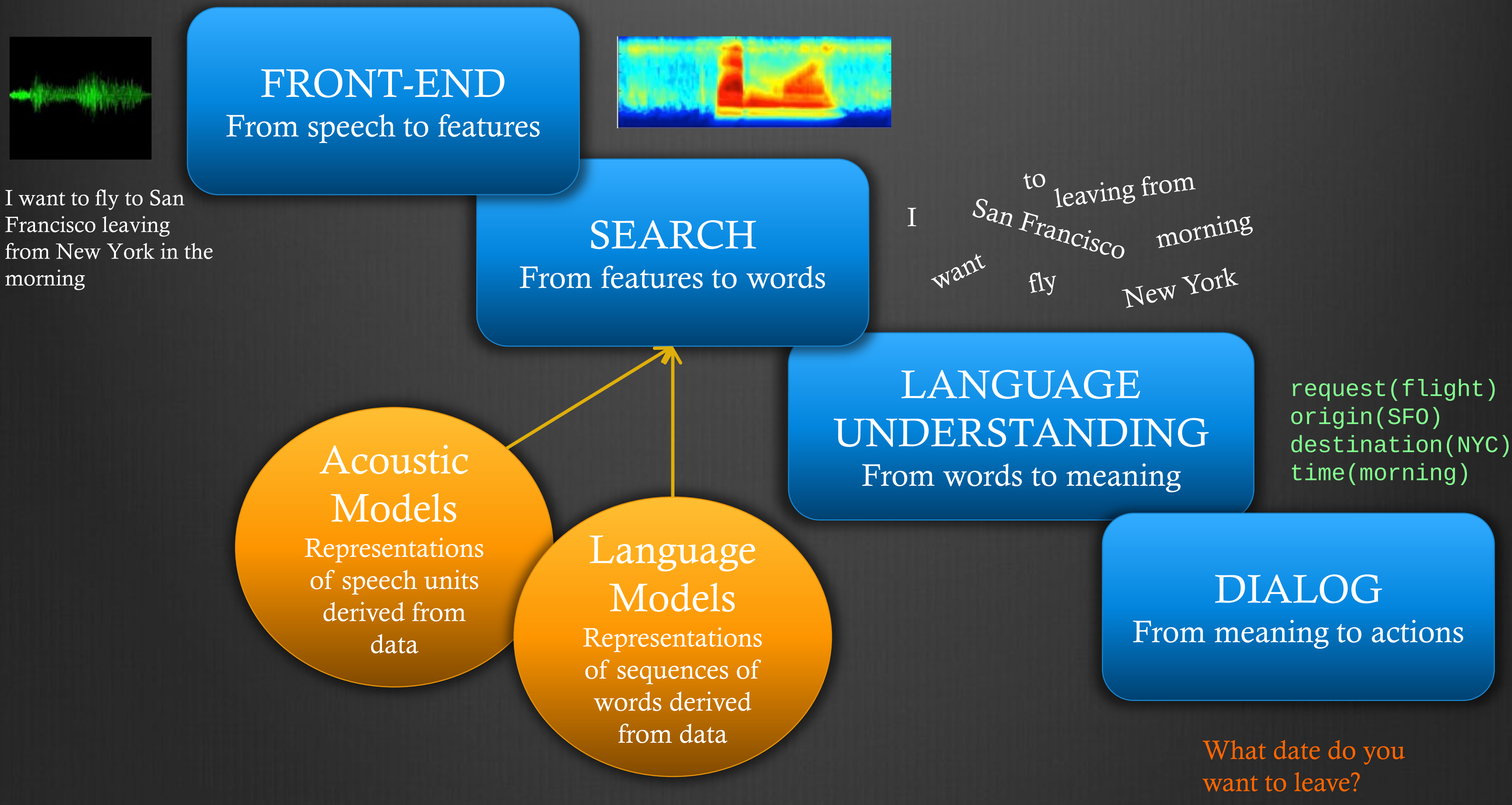
… the leading approach to language modeling.

… ord given the 2 preceding words

… if you have a 10,000 word vocabulary, you … -grams. Most of the work is around how to …

Today's storage and CPU allow the use of massive tri-grams, four-grams, five-grams, and above

The parts of a speech understanding system



The parts of a speech understanding system

Language understanding

It is highly domain depe d
from scratch.

Commercial systems ext
augmented with code to

```
$ITINERARY = $FROM $TO;
<script>
        origin = $FROM.VALUE;
        destination = $FROM.VALUE;
        INVALIDATE = $origin == $destination;
</script>

$FROM = from $AIRPORT;
<script>
        VALUE = $AIRPORT.VALUE;
</script>

$TO = to $AIRPORT;
<script>
        VALUE = $AIRPORT.VALUE;
</script>

$AIRPORT = [new york] (J F K )|(kennedy) [airport];
<script>
        VALUE = JFK;
</script>

$AIRPORT = (boston | logan) [airport];
<script>
        VALUE = BOS;
</script>
```

ght)

(NYC)
g)

The parts of a speech understanding system

Language understanding is generally one of the weakest links of the chain

It is highly domain dependent. Every different domain generally has to be hacked from scratch.

Commercial systems use handcrafted grammars augmented with code to represent meaning

… or statistical classifiers expensively built by manually annotating hundreds of thousands of in-domain sentences

Statistical Semantic Classification

TRANSCRIPTIONS	ANNOTATIONS
want to cancel the account	CANCEL_ACCOUNT
cancel service	CANCEL_ACCOUNT
I cant send a particular message to a certain group of people	CANNOT_SEND_RECEIVE_EMAIL
cancellation of the service	CANCEL_ACCOUNT
I need to setup my email	EMAIL_SETUP
they registered my modem in from my internet and I need to get my email address	EMAIL_SETUP
my emails are not been received at the address I sent it to	CANNOT_SEND_RECEIVE_EMAIL
…	

Language Model for Speech Recognition

Statistical Semantic Classifier

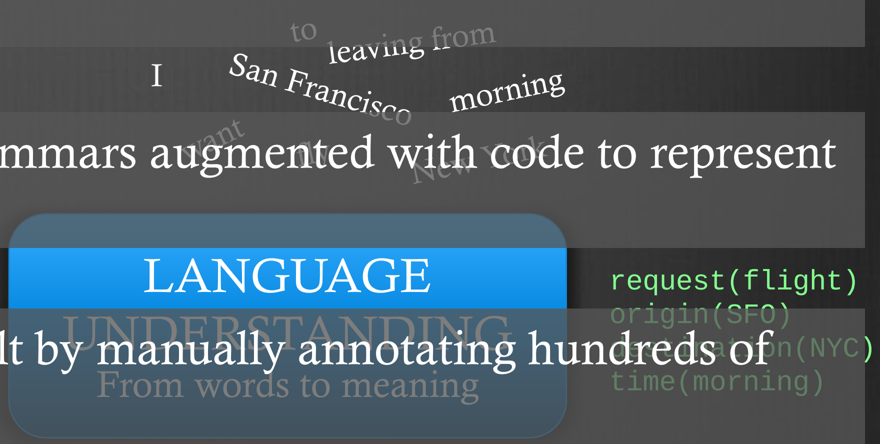
The parts of a speech understanding system

Language understanding is generally one of the weakest links of the chain

It is highly domain dependent. Every different domain generally has to be hacked from scratch.

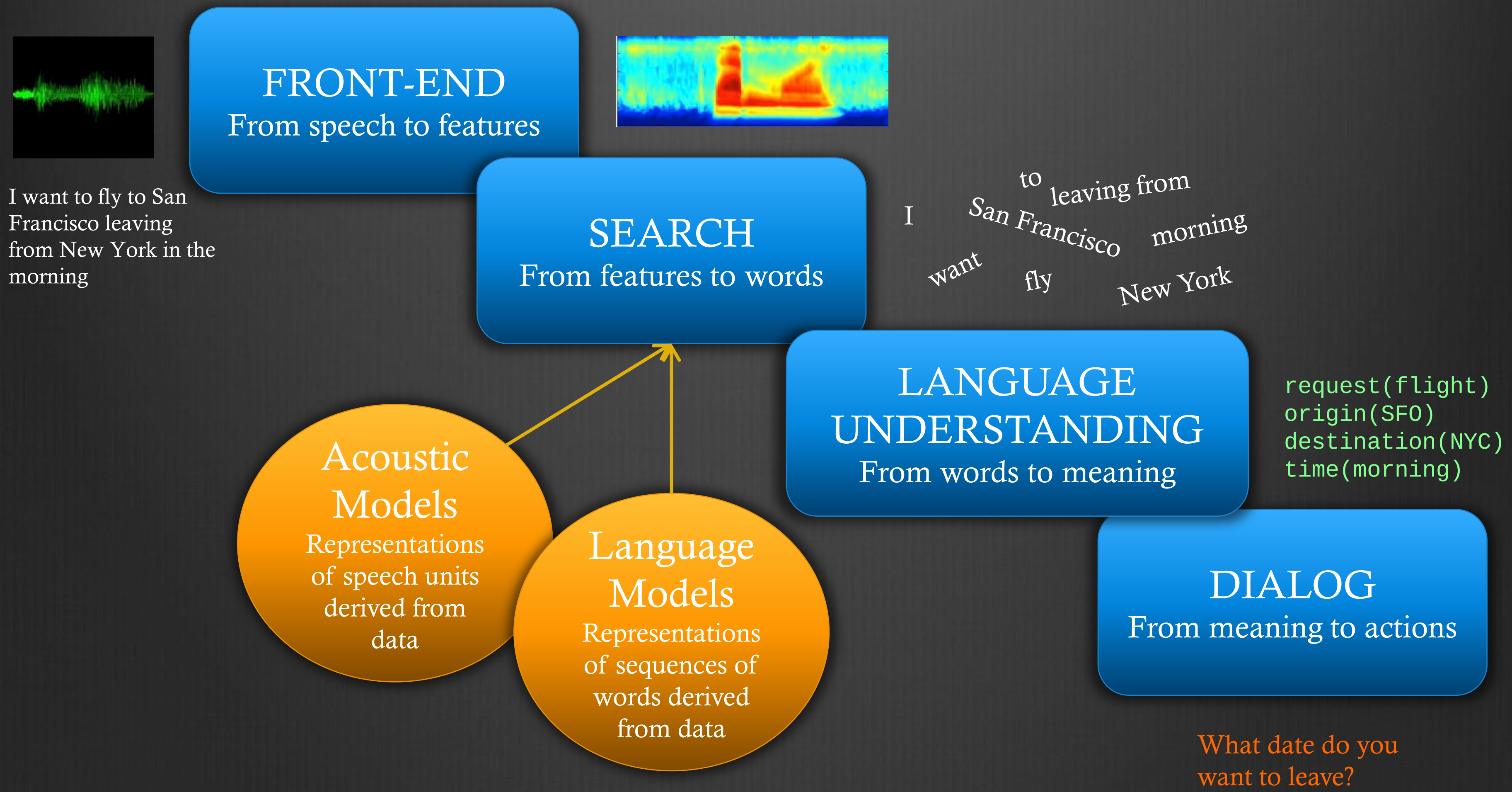
Commercial systems use handcrafted grammars augmented with code to represent meaning

… or statistical classifiers expensively built by manually annotating hundreds of thousands of in-domain sentences



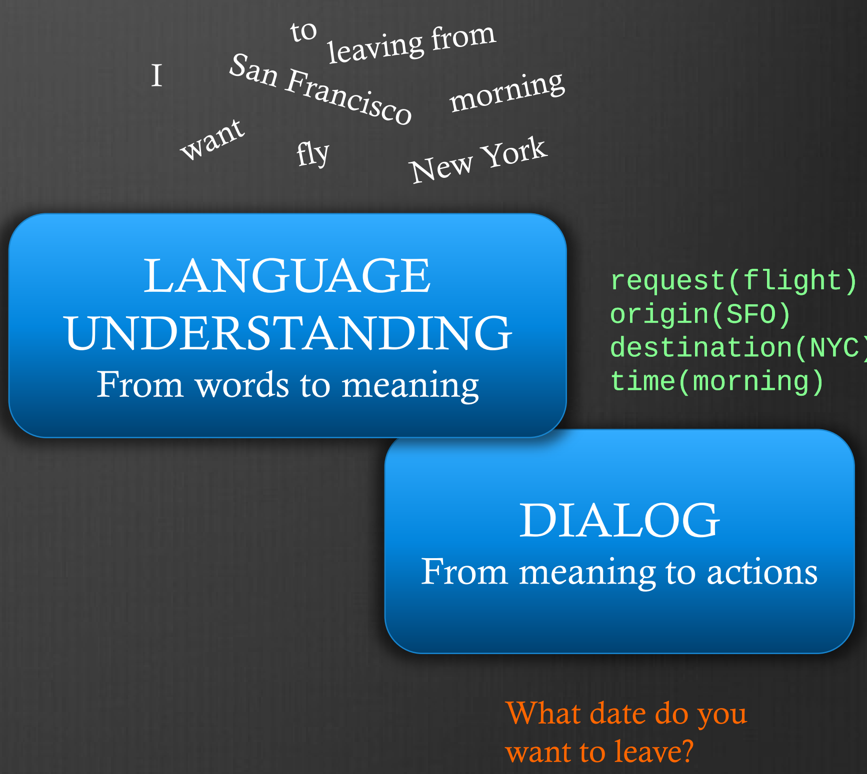
As of today we do not have domain independent language understanding systems

The parts of a speech understanding system

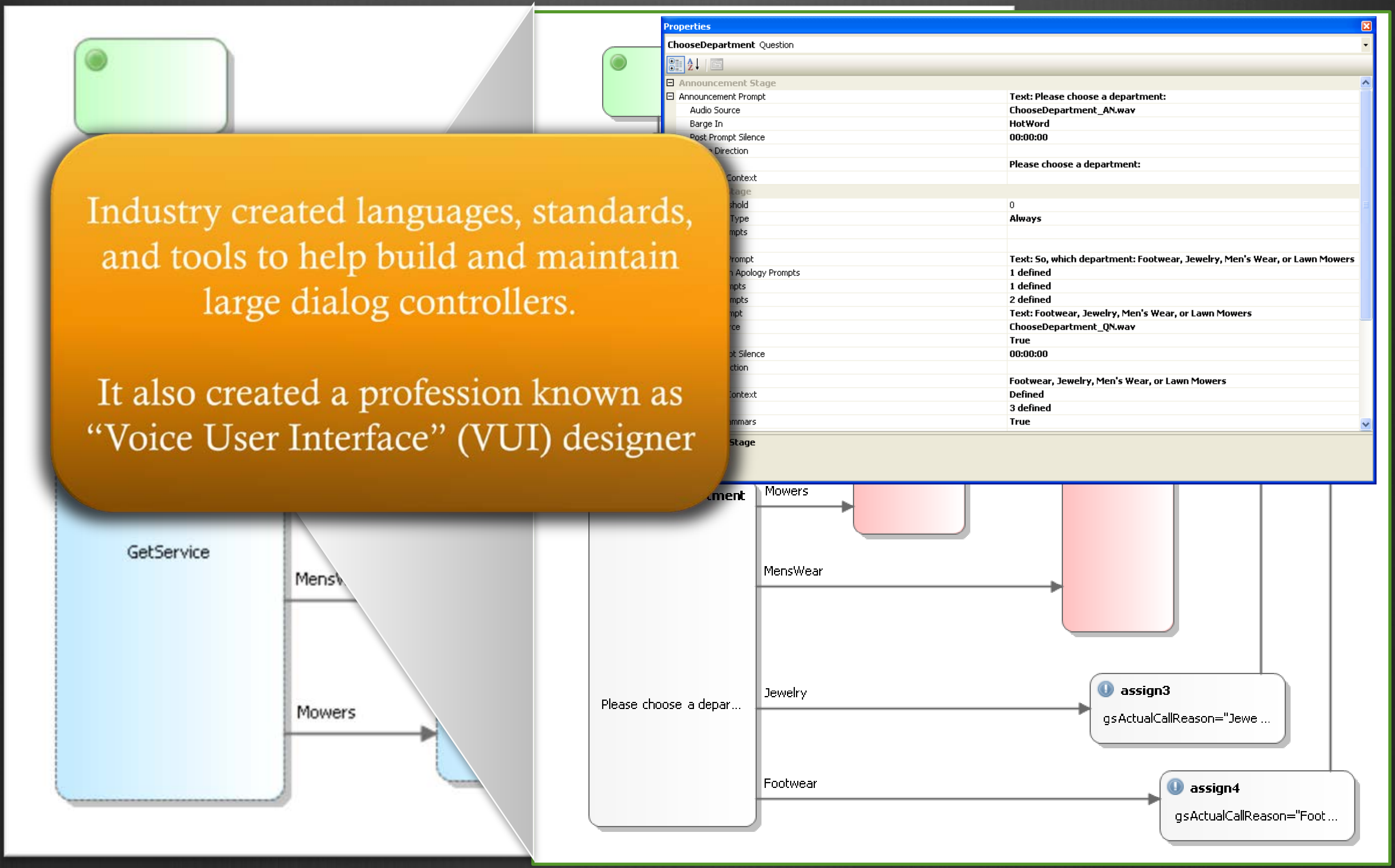


The parts of a speech understanding system

The dialog manager is often a giant finite state machine controller built by hand (called *call flow)*



Call-flow



Industry created languages, standards, and tools to help build and maintain large dialog controllers.

It also created a profession known as "Voice User Interface" (VUI) designer

The parts of a speech understanding system

The dialog manager is often a giant finite state machine controller built by hand (called *call flow)*

Learning dialog by reinforcement learning is the new holy grail

I want to fly from San Francisco leave morning New York

LANGUAGE UNDERSTANDING
From words to meaning

```
request(flight)
origin(SFO)
destination(NYC)
time(morning)
```

DIALOG
From meaning to actions

What date do you want to leave?

Reinforcement learning for dialog

Reinforcement learning for dialog

UNTRAINED STRATEGY

S: RELEASE AIRLINE	*Do you want to choose another airline?*
U:	What?
S: RETRIEVAL	
S: CONSTRAIN DEPART. TIME	*When do you want to leave?*
U:	UH? I want to go to Boston
S: OUTPUT DATA	*There are 23,457 flights...*
U:	I don't understand, can you please show me the flights from San Francisco to Boston?
S: CLOSE DIALOGUE	*Thank you for using AT&T.*

TRAINED STRATEGY

S: GREETING	*This is the AT&T flight information. How may I help you?*
U:	I want to go to Boston.
S: CONSTRAIN ORIGIN	*Where do you want to leave from?*
U:	San Francisco
S: CONSTRAIN AIRLINE	*Which airline?*
U:	Delta
S: RETRIEVAL (30 flights)	
S: CONSTRAIN DEPART. TIME	*When do you want to leave?*
U:	In the late afternoon.
S: RETRIEVAL (3 flights)	
S: OUTPUT DATA	*Flight ... leaves at ..., flight ... leaves at ...*
U:	Thanks.
S: CLOSE DIALOGUE	*Thank you for using AT&T.*

CURRENT DIALOG STATE

Machine learning and dialog

UNTRAINED STRATEGY

S: RELEASE AIRLINE	*Do you want to choose another airline?*
U:	What?
S: RETRIEVAL	
S: CONSTRA… DEPART… TIME	*When do you want to leave?*
U:	
S: OUT… DATA	
U:	
S: CLO… DIALOG…	

Dealing with Uncertainties

POMDPs (Partially Observable Markov decision Processes)

Problems:

- Design the proper reward schema
- Create artificial user for training

DE… TIME	
U:	In the late afternoon.
S: RETRIEVAL (3 flights)	
S: OUTPUT DATA	*Flight … leaves at …, flight … leaves at …*
U:	Thanks.
S: CLOSE DIALOGUE	*Thank you for using AT&T.*

CURRENT DIALOG STATE

ELIZA meets PARRY

ELIZA (1964) simulates a Rogerian psychotherapist

PARRY (1972) simulates a paranoid schizophrenic

This is what happens when the two programs talk to each other

. . .

PARRY: I went to the races.
ELIZA: Please go on.
PARRY: Do you know anything about bookies?
ELIZA: What makes you think I know anything about bookies?
PARRY: A bookie didn't pay me off once.
ELIZA: What does that suggest to you?
PARRY: I beat up the bookie who did not pay off, and I thought he might use his friends in the underworld to get even with me

. . .

So .. Where are we now?

- 60 year of research brought great progress in computer speech understanding
- However, still far from human ability, and from the vision of a truly conversational computer
- No major breakthroughs since 30 years go
- Just more data, and more computer power
- Lots of hard hand-crafting for the higher levels of knowledge (language understanding and dialog management)

Where are we going?

- No data like more data is still a valid paradigm for incremental improvement
- Bu we start experiencing the asymptotic limitations of the strong model assumptions we made 30 or more years ago
- Trying to understand how the brain works and applying that to machine can hopefully bring new results
- Trying to move move away from highly handcrafted systems.