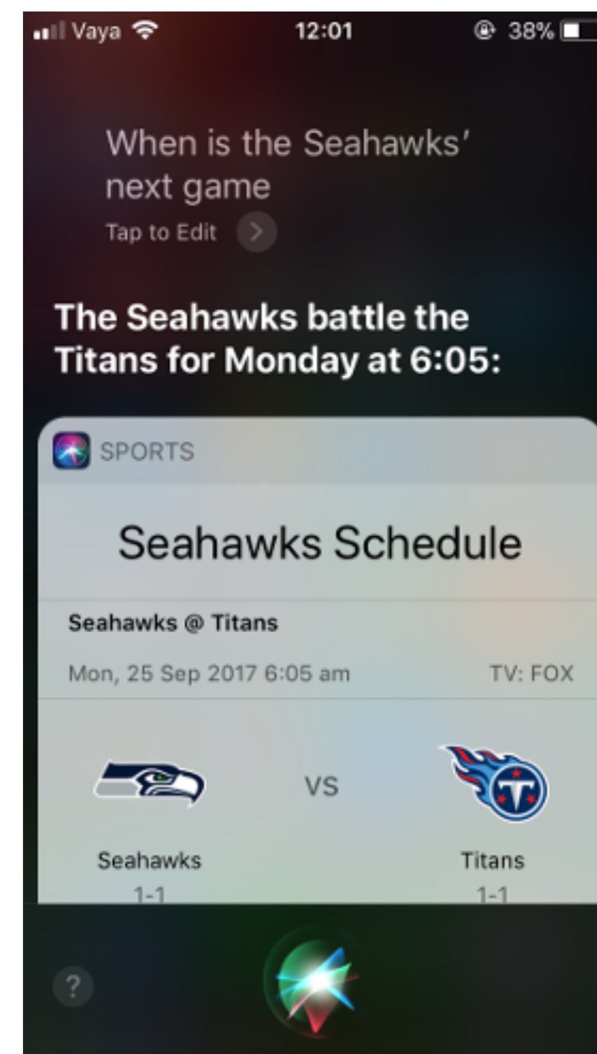
Natural Language Processing

George Konidaris
gdk@cs.brown.edu

Fall 2021


amazon


Vaya
12:01
38%
When is the Seahawks' next game
Tap to Edit
The Seahawks battle the Titans for Monday at 6:05:
SPORTS
Seahawks Schedule
Seahawks @ Titans
Mon, 25 Sep 2017 6:05 am
TV: FOX
VS
Seahawks
1-1
Titans
1-1

Natural Language Processing

Understanding spoken/written sentences in a natural language.

Major area of research in AI.

Why?

- Humans use language to communicate.
- Most natural interface.
- Huge amounts of NLP "knowledge" around.
  - E.g., books, the entire internet.
- Generative power

- Key to intelligence?
  - Hints as to underlying mechanism
  - Key indicator of intelligence

Natural Language Processing

It is also *incredibly hard*. ***Why?***

I saw a bat.

Lucy owns a parrot that is larger than a cat.

John kissed his wife, and so did Sam.

Mary invited Sue for a visit, but she told her she had to go to work.

I went to the hospital, and they told me to go home and rest.

The price of tomatoes in Des Moines has gone through the roof.

Mozart was born in Salzburg and Beethoven, in Bonn.

(examples via Ernest Davis, NYU)

Natural Language Processing

"If you are a fan of the justices who fought throughout the Rehnquist years to pull the Supreme Court to the right, Alito is a home run - a strong and consistent conservative with the skill to craft opinions that make radical results appear inevitable and the ability to build trusting professional relationships across ideological lines." (TNR, Nov. 2005)

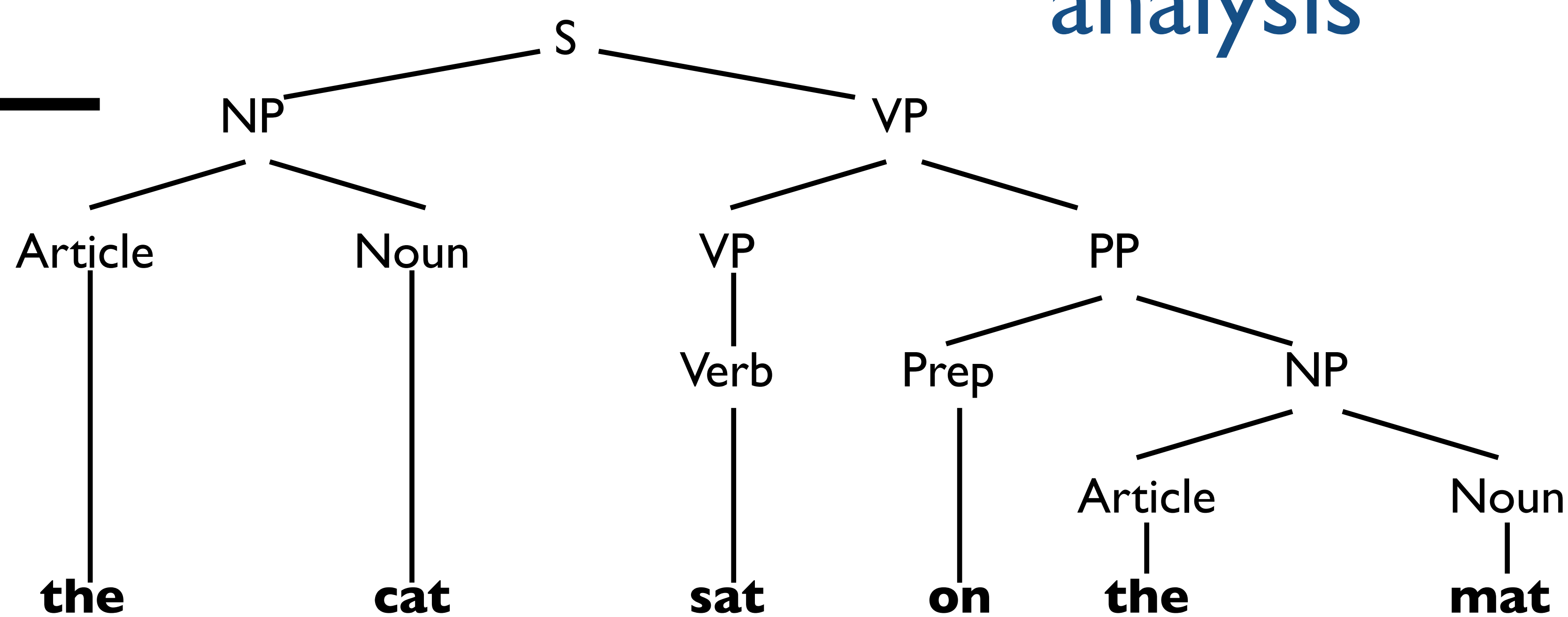
(examples via Ernest Davis, NYU)

Component Problems

perception

"the cat sat on the mat"

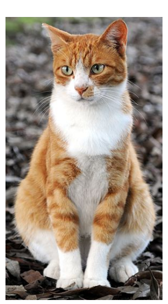
syntactic analysis



semantic analysis

SatOn(x = Cat, y = Mat)

disambiguation

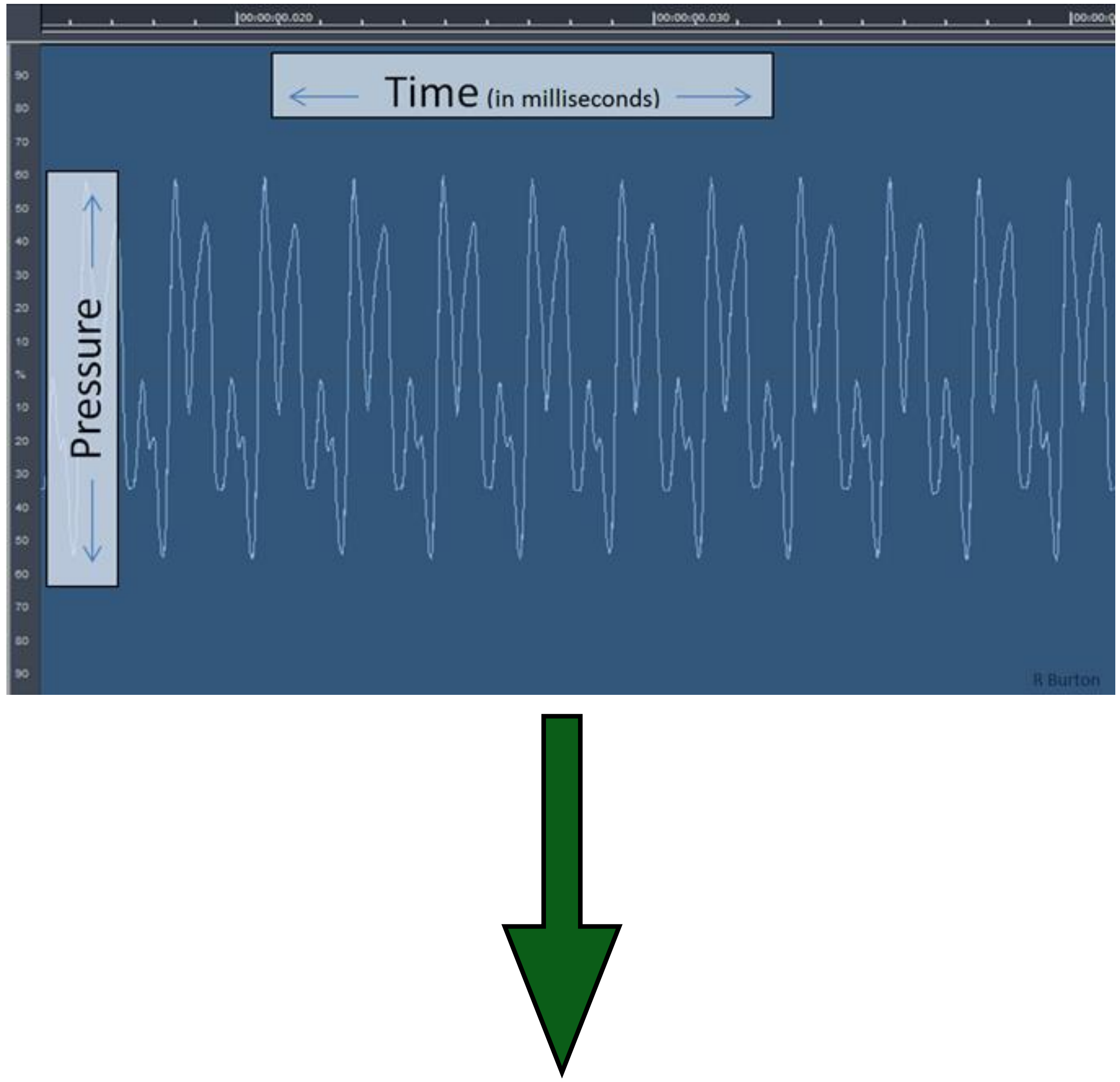
Cat?  

Mat?  

incorporation

SatOn(cat3, mat16)

Perception



"The cat sat on the mat."

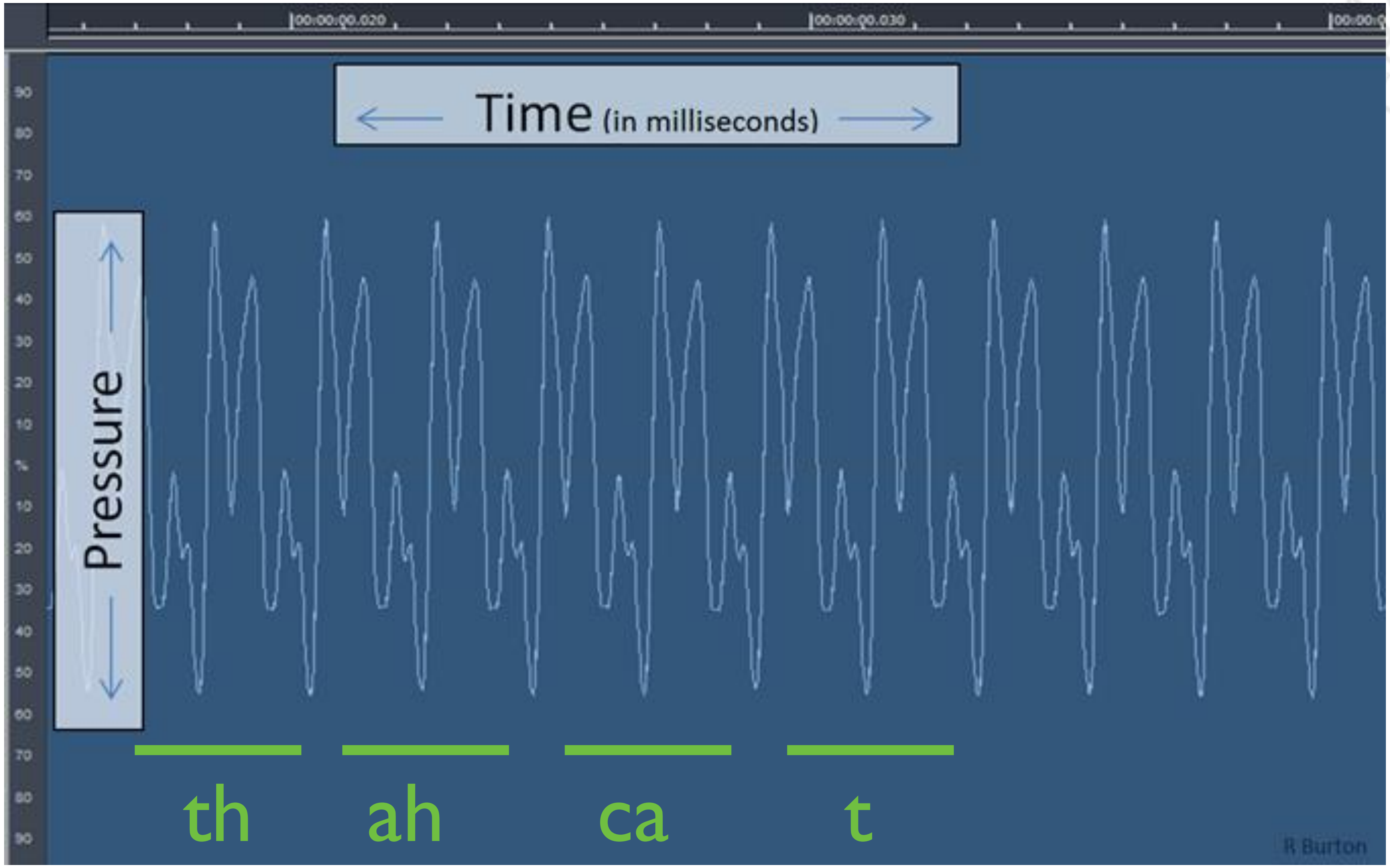
Major Challenges

Speaker accent, volume, tone.

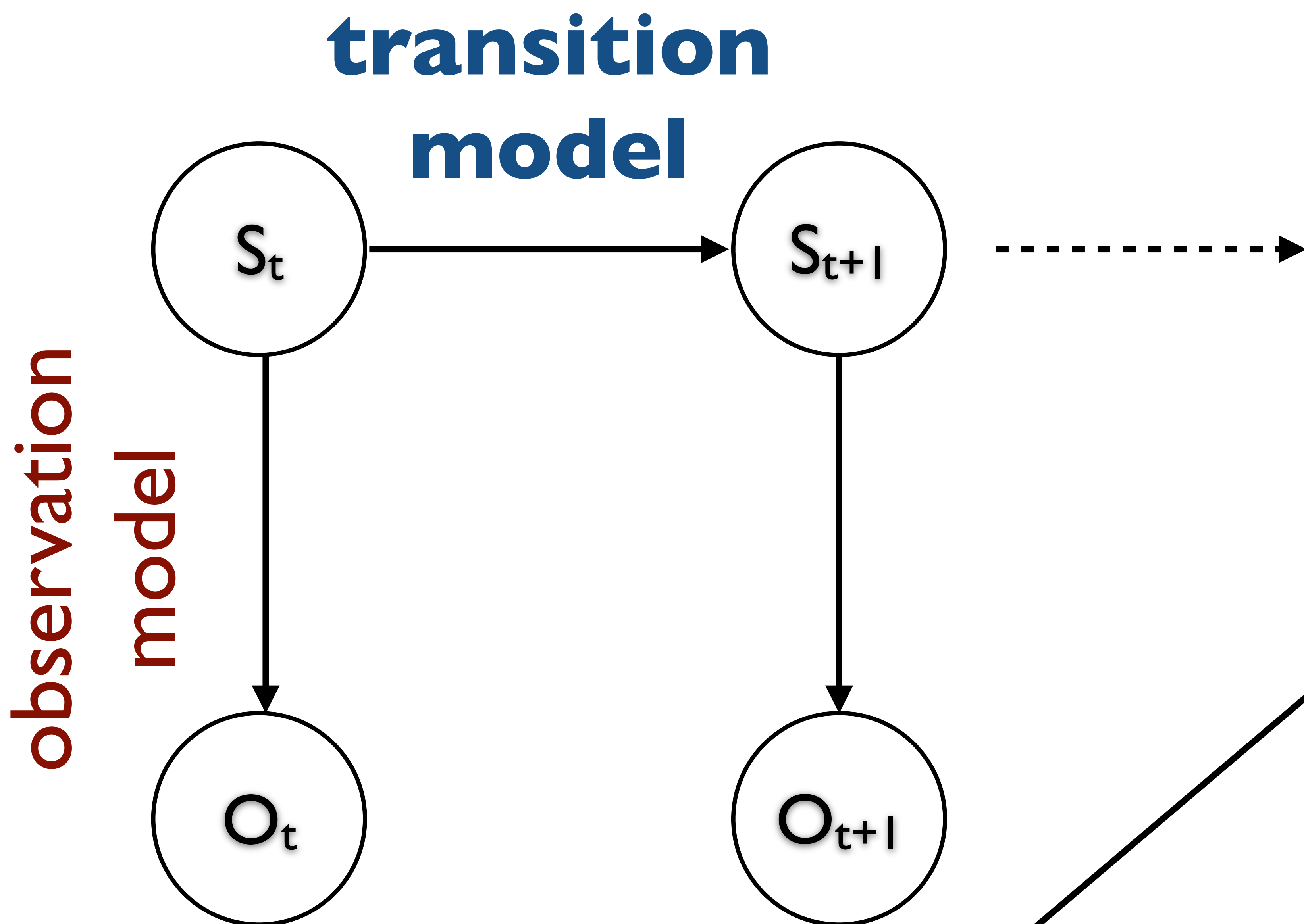


No pauses - word boundaries?
Noise.
Variation.

Speech Recognition



Speech Recognition Using HMMs



Must store:

- $P(O \mid S)$
- $P(S_{t+1} \mid S_t)$

prob. of observed audio given phoneme

prob. of one phoneme following another

Issues

Phoneme sequence not Markov

- Must introduce memory for context
- k-Markov Models

People speak faster or slower

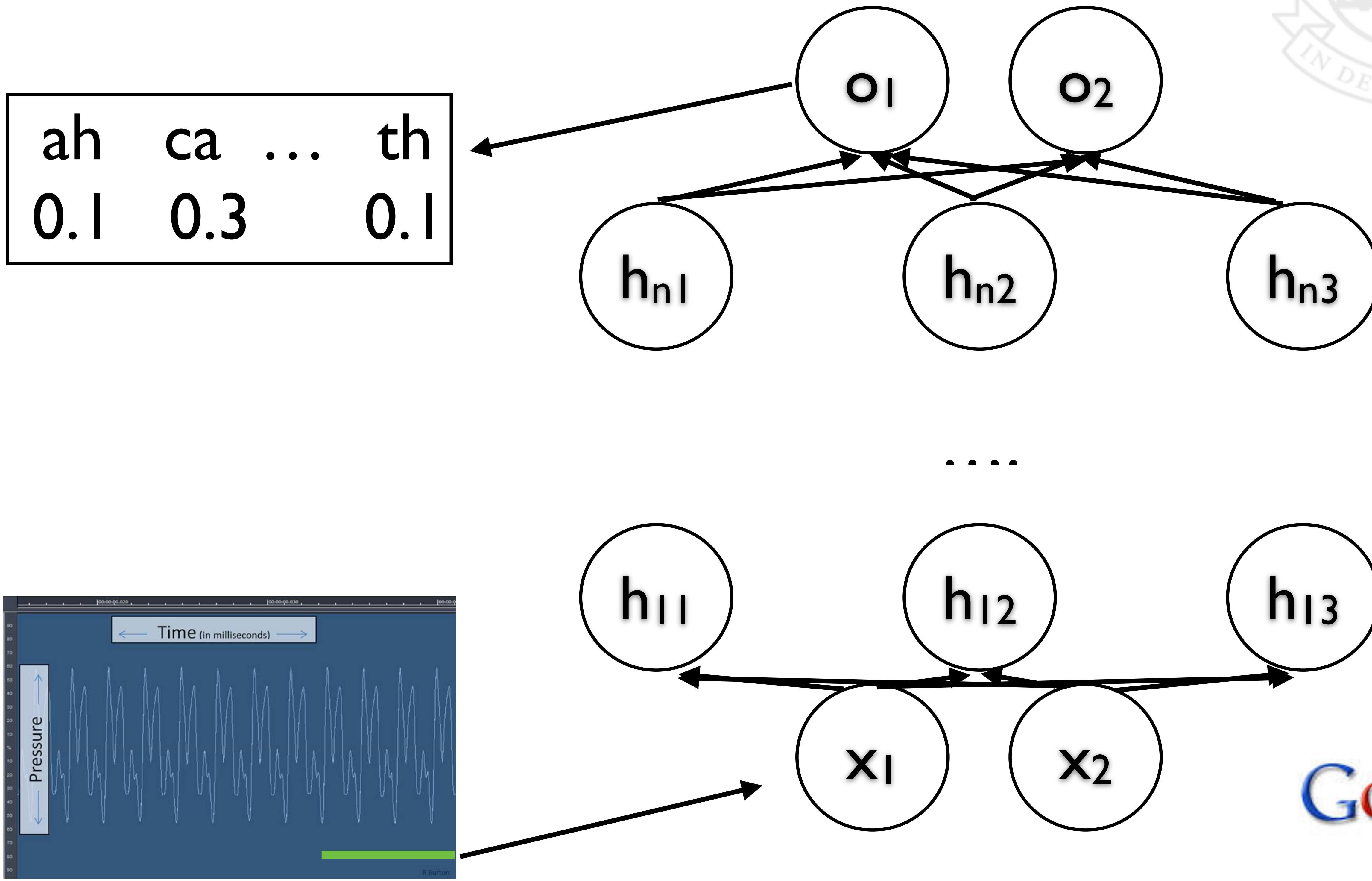
- “Window” does not have fixed length
- Dynamic Time Warping

Quite a simplistic model for a complex phenomenon.

Nevertheless, speech recognition tech based on HMMs commercially-viable mid-1990s.

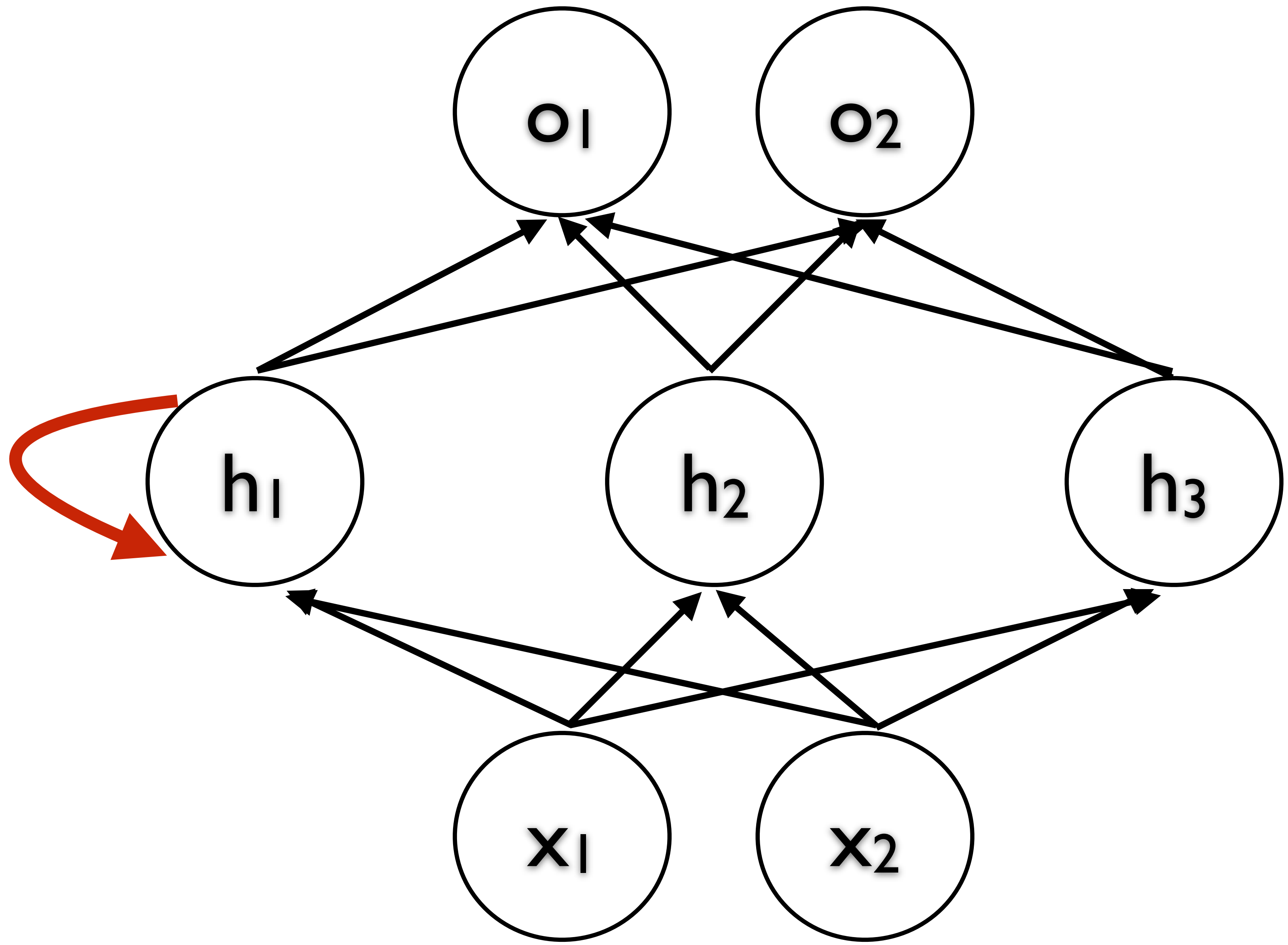
Speech Recognition with Deep Nets

Mid-to-late 2000s: replace HMM with Deep Net.



Speech Recognition with Deep Nets

How to deal with dependency on prior states and observations?



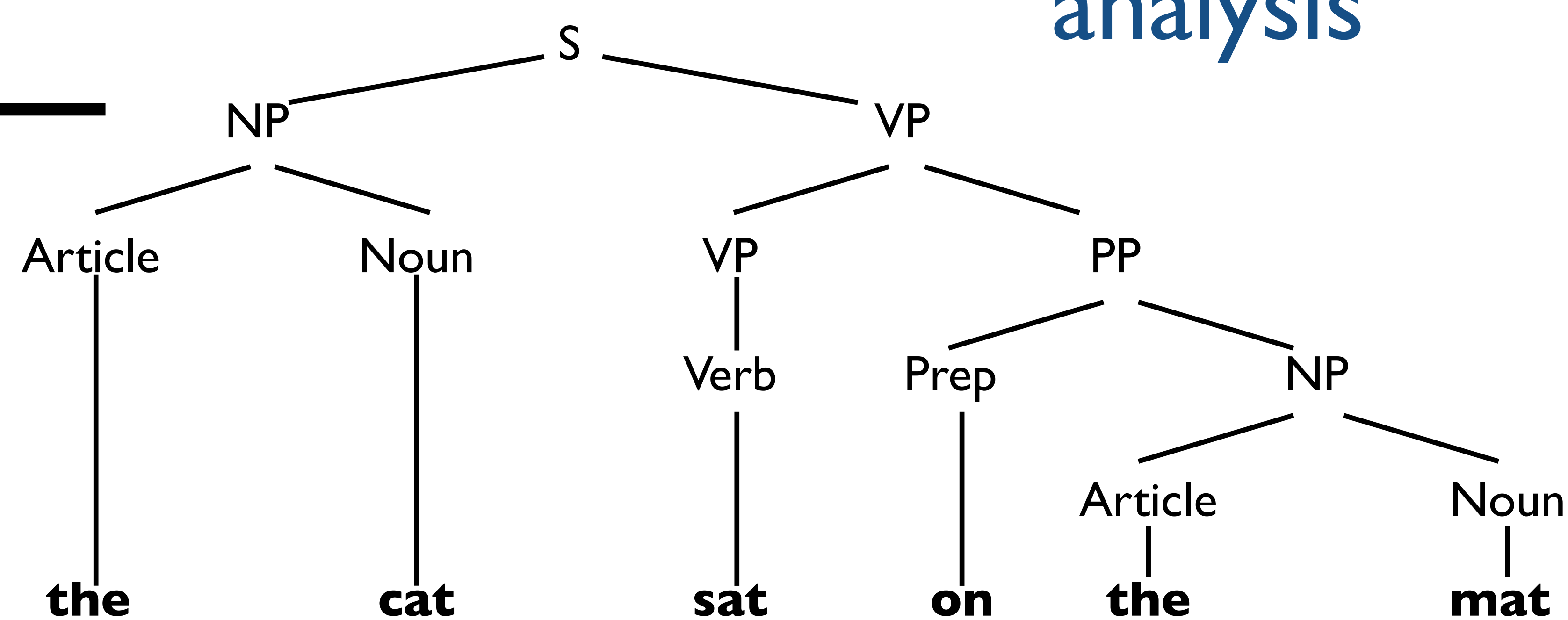
Recurrent nets: form of memory.

Component Problems

perception

"the cat sat on the mat"

syntactic analysis



semantic analysis

SatOn(x = Cat, y = Mat)

disambiguation

Cat?

Mat?

incorporation

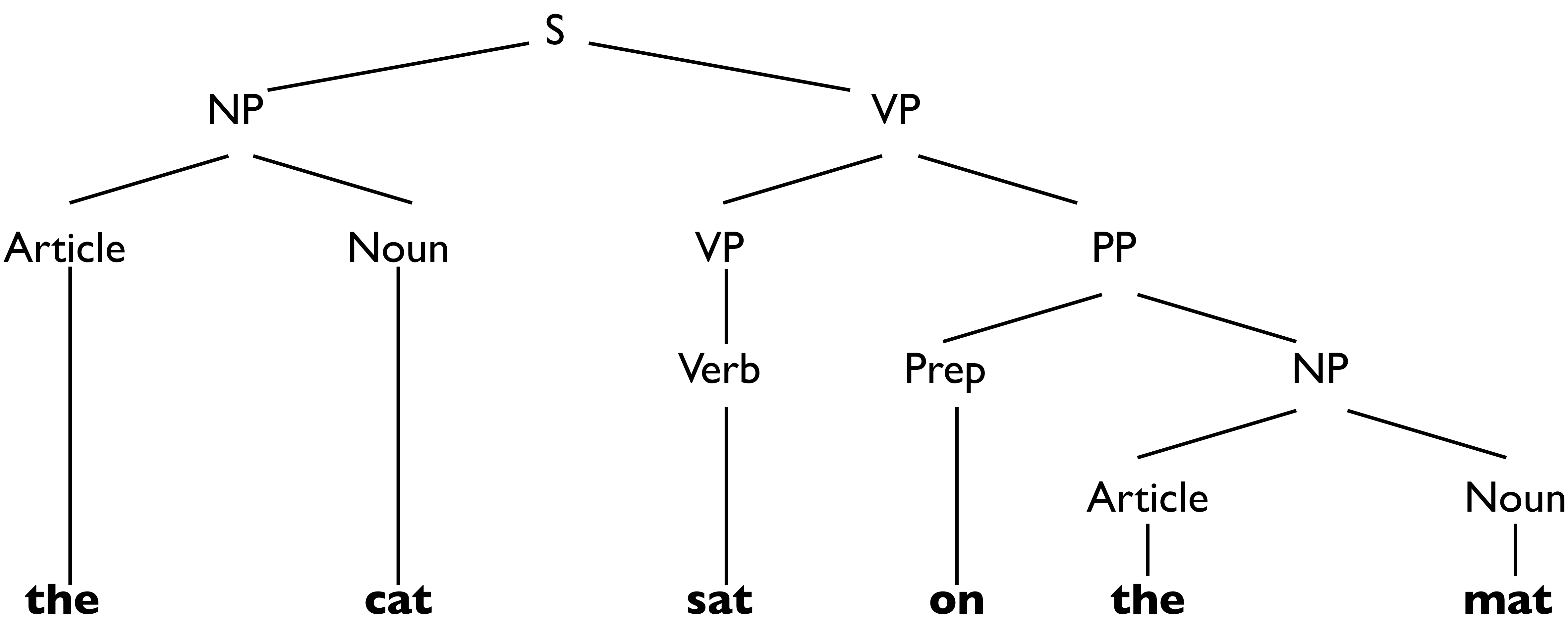
SatOn(cat3, mat16)

Syntactic Analysis

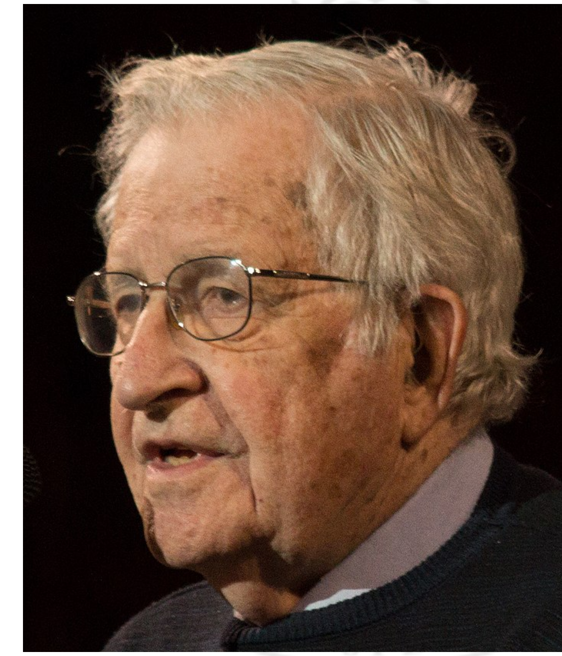
Syntax: characteristic of language.

- Structure.
- Composition.

But observed in linear sequence.



Syntactic Analysis



How to describe this structure?

Formal grammar.

- Set of rules for generating sentences.
- Varying power:
  - Recursively enumerable (equiv. Turing Machines)
  - Context-Sensitive
  - Context-Free
  - Regular

Each uses a set of rewrite rules to generate *syntactically correct* sentences.

Colorless green ideas sleep furiously.

Syntax

$\mathcal{E}_0$:	S	$\rightarrow$	$NP\ VP$	[0.90]	I + feel a breeze
		$\vert$	$S\ Conj\ S$	[0.10]	I feel a breeze + and + It stinks
	NP	$\rightarrow$	$Pronoun$	[0.30]	I
		$\vert$	$Name$	[0.10]	John
		$\vert$	$Noun$	[0.10]	pits
		$\vert$	$Article\ Noun$	[0.25]	the + wumpus
		$\vert$	$Article\ Adjs\ Noun$	[0.05]	the + smelly dead + wumpus
		$\vert$	$Digit\ Digit$	[0.05]	3 4
		$\vert$	$NP\ PP$	[0.10]	the wumpus + in 1 3
		$\vert$	$NP\ RelClause$	[0.05]	the wumpus + that is smelly
	VP	$\rightarrow$	$Verb$	[0.40]	stinks
		$\vert$	$VP\ NP$	[0.35]	feel + a breeze
		$\vert$	$VP\ Adjective$	[0.05]	smells + dead
		$\vert$	$VP\ PP$	[0.10]	is + in 1 3
		$\vert$	$VP\ Adverb$	[0.10]	go + ahead
	$Adjs$	$\rightarrow$	$Adjective$	[0.80]	smelly
		$\vert$	$Adjective\ Adjs$	[0.20]	smelly + dead
	PP	$\rightarrow$	$Prep\ NP$	[1.00]	to + the east
	$RelClause$	$\rightarrow$	$RelPro\ VP$	[1.00]	that + is smelly

Formal Grammars

Two types of symbols:

- Terminals (stop and output this)
- Non-terminals (one is a *start symbol*)

Production (*rewrite*) rules that modify a string of symbols by matching expression on left, and replacing it with one on right.

$S \rightarrow AB$	ab
$A \rightarrow AA$	$aaaaaab$
$A \rightarrow a$	$abbb$
$B \rightarrow BBB$	$aabbbbb$
$B \rightarrow b$	

Context-Free Grammars

Rules must be of the form:

$$A \rightarrow B$$

where *A* is a ***single*** non-terminal and *B* is any sequence of terminals and non-terminal.

Why is this called *context-free*?

Probabilistic CFGs

Attach a probability to each rewrite rule:

$$A \to B[0.3]$$
$$A \to AA[0.6]$$
$$A \to a[0.1]$$

Probabilities for the same left symbol sum to 1.

Why do this?

More vs. less likely sentences.

Probability distribution over valid sentences.

E_0

Noun	→	**stench** [0.05] | **breeze** [0.10] | **wumpus** [0.15] | **pits** [0.05] | . . .
Verb	→	**is** [0.10] | **feel** [0.10] | **smells** [0.10] | **stinks** [0.05] | . . .
Adjective	→	**right** [0.10] | **dead** [0.05] | **smelly** [0.02] | **breezy** [0.02] . . .
Adverb	→	**here** [0.05] | **ahead** [0.05] | **nearby** [0.02] | . . .
Pronoun	→	**me** [0.10] | **you** [0.03] | **I** [0.10] | **it** [0.10] | . . .
RelPro	→	**that** [0.40] | **which** [0.15] | **who** [0.20] | **whom** [0.02] ∨ . . .
Name	→	**John** [0.01] | **Mary** [0.01] | **Boston** [0.01] | . . .
Article	→	**the** [0.40] | **a** [0.30] | **an** [0.10] | **every** [0.05] | . . .
Prep	→	**to** [0.20] | **in** [0.10] | **on** [0.05] | **near** [0.10] | . . .
Conj	→	**and** [0.50] | **or** [0.10] | **but** [0.20] | **yet** [0.02] ∨ . . .
Digit	→	**0** [0.20] | **1** [0.20] | **2** [0.20] | **3** [0.20] | **4** [0.20] | . . .

Lexicon

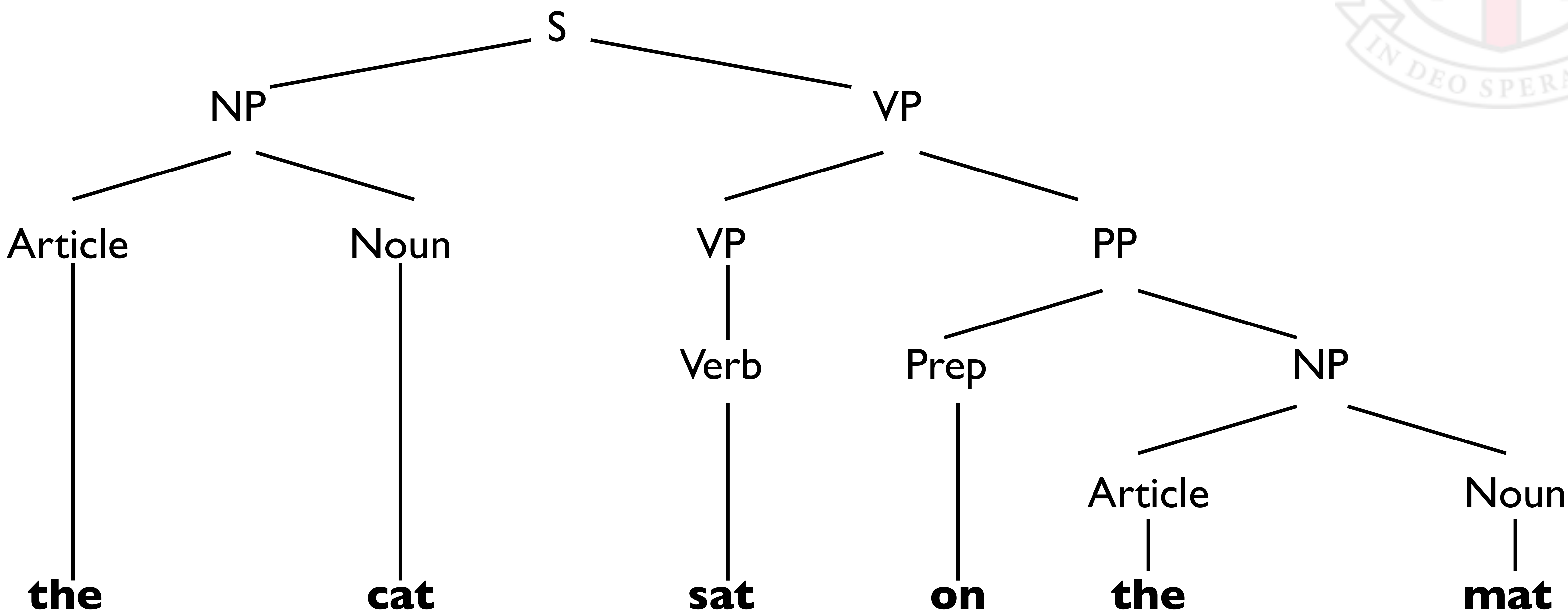
(R&N)

E_0

$\mathcal{E}_0$:	*S*	→	*NP VP*	[0.90]	I + feel a breeze
		|	*S Conj S*	[0.10]	I feel a breeze + and + It stinks
	NP	→	*Pronoun*	[0.30]	I
		|	*Name*	[0.10]	John
		|	*Noun*	[0.10]	pits
		|	*Article Noun*	[0.25]	the + wumpus
		|	*Article Adjs Noun*	[0.05]	the + smelly dead + wumpus
		|	*Digit Digit*	[0.05]	3 4
		|	*NP PP*	[0.10]	the wumpus + in 1 3
		|	*NP RelClause*	[0.05]	the wumpus + that is smelly
	VP	→	*Verb*	[0.40]	stinks
		|	*VP NP*	[0.35]	feel + a breeze
		|	*VP Adjective*	[0.05]	smells + dead
		|	*VP PP*	[0.10]	is + in 1 3
		|	*VP Adverb*	[0.10]	go + ahead
	Adjs	→	*Adjective*	[0.80]	smelly
		|	*Adjective Adjs*	[0.20]	smelly + dead
	PP	→	*Prep NP*	[1.00]	to + the east
	RelClause	→	*RelPro VP*	[1.00]	that + is smelly

Grammar

(R&N)

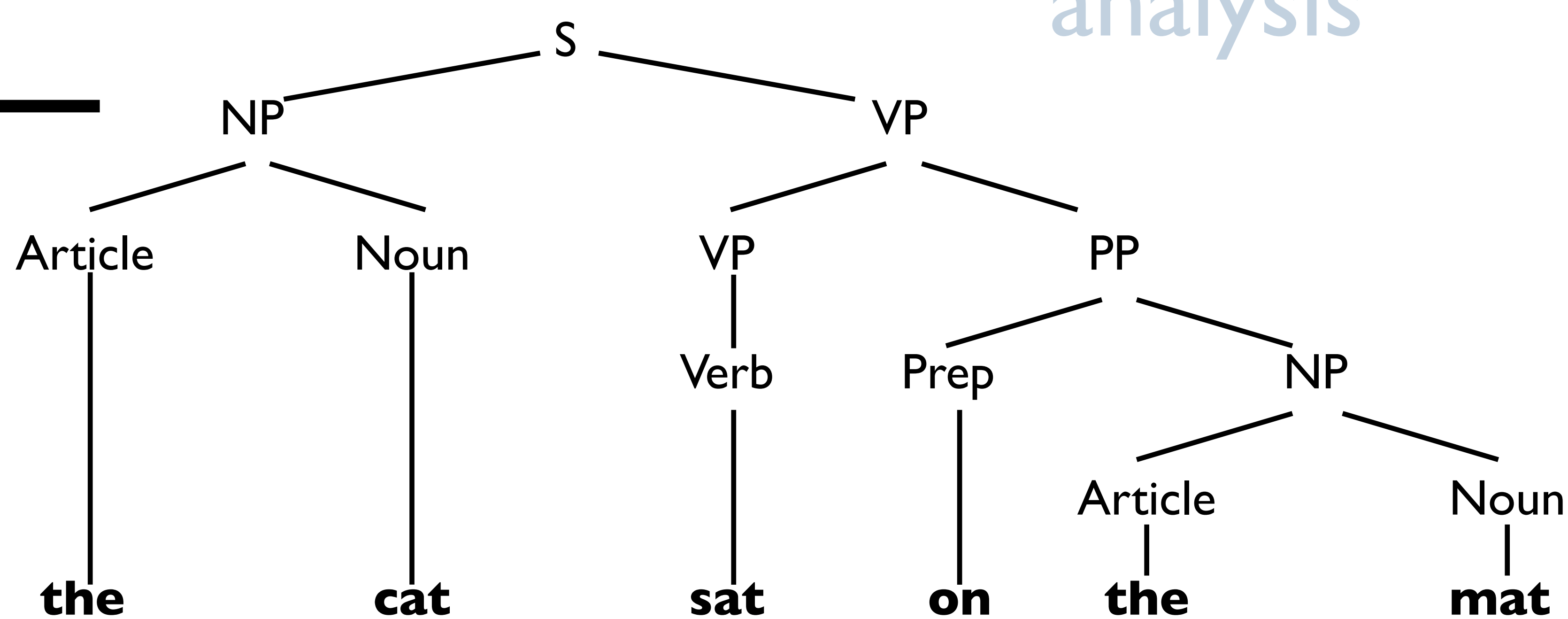

S
NP
VP
Article
Noun
VP
PP
Verb
Prep
NP
Article
Noun
the
cat
sat
on
the
mat

Component Problems

perception

“the cat sat on the mat”

syntactic analysis



semantic analysis

SatOn(x = Cat, y = Mat)

disambiguation

Cat?

Mat?

incorporation

SatOn(cat3, mat16)

Semantic Analysis

Semantics: what the sentence actually means, eventually in terms of symbols available to the agent (e.g., a KB).

"the cat sat on the mat"

↓

SatOn(x = Cat, y = Mat)

SatOn(cat3, mat16)

Semantic Analysis

Key idea: ***compositional semantics.***

The semantics of sentences are built out of the semantics of their constituent parts.

"The cat sat on the mat."

Therefore there is a clear relationship between syntactic analysis and semantic analysis.

Semantic Analysis

Useful step:

- Probability of parse depends on words
- Lexicalized PCFGs

$$VP(v) \rightarrow Verb(v) NP(n) [P_1(v, n)]$$

variables

probability depends on variable bindings

ate bandanna

vs.

ate banana

Semantic Analysis

"John loves Mary"
Desired output: *Loves(John, Mary)*

Semantic parsing:

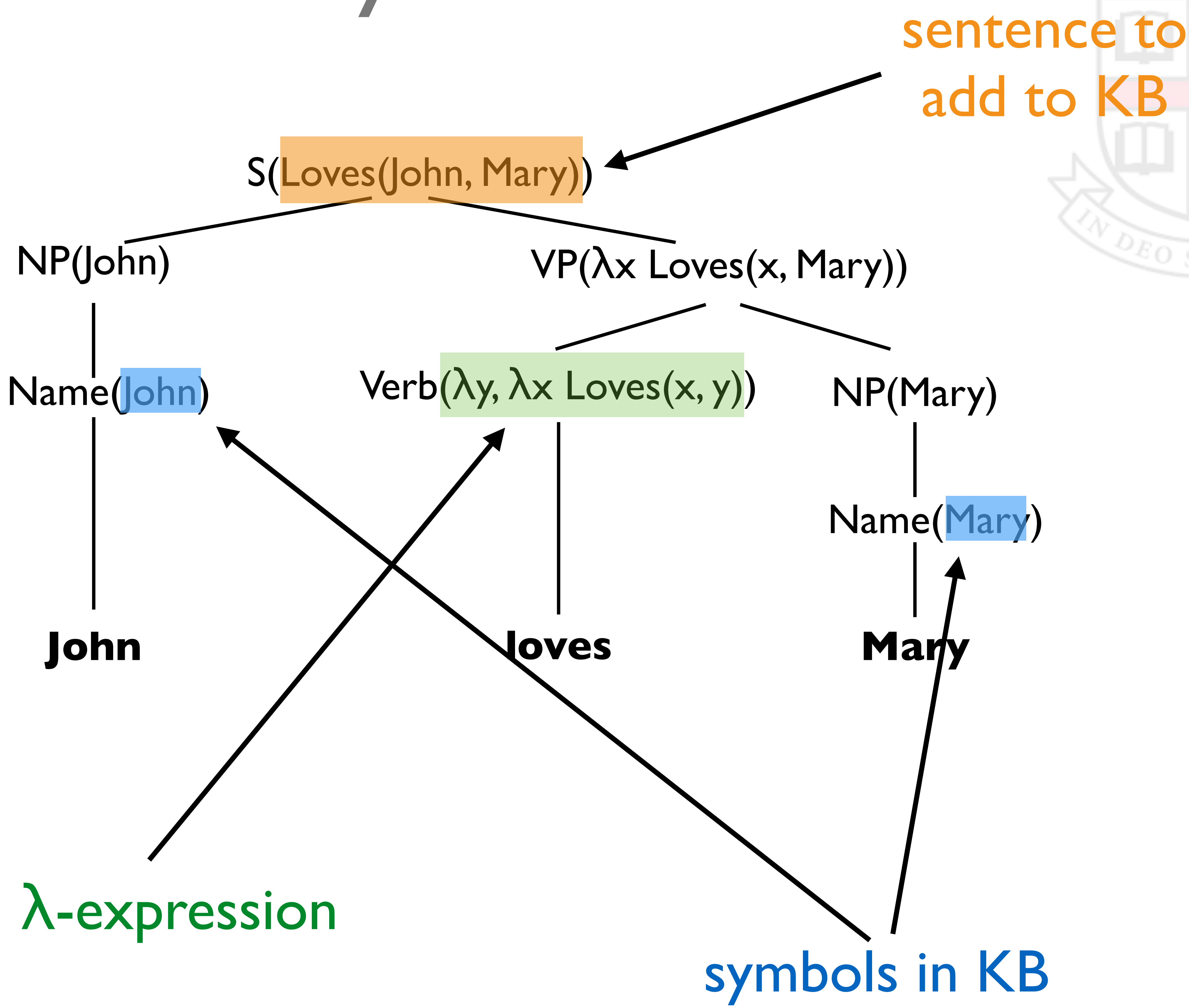
- Exploit compositionality of parsing to build semantics.

$$S(pred(obj)) \rightarrow NP(obj)\ VP(pred)$$
$$VP(pred(obj)) \rightarrow Verb(pred)\ NP(obj)$$
$$NP(obj) \rightarrow Name(obj)$$

$$Name(John) \rightarrow \textbf{John}$$
$$Name(Mary) \rightarrow \textbf{Mary}$$
$$Verb(\lambda y\ \lambda x\ Loves(x, y)) \rightarrow \textbf{loves}$$

(R&N)

Semantic Analysis



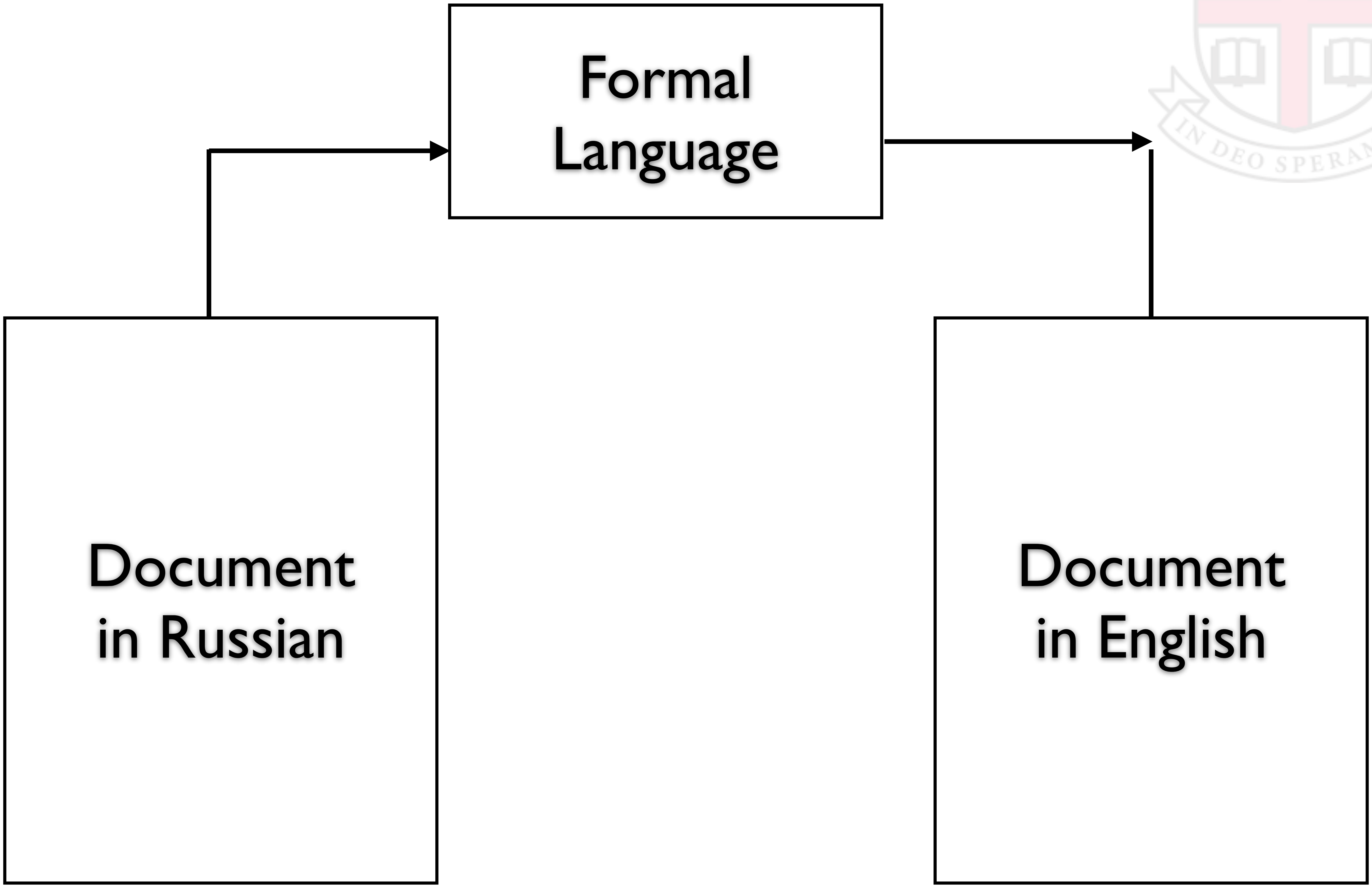
Machine Translation

Major goal of NLP research for decades.

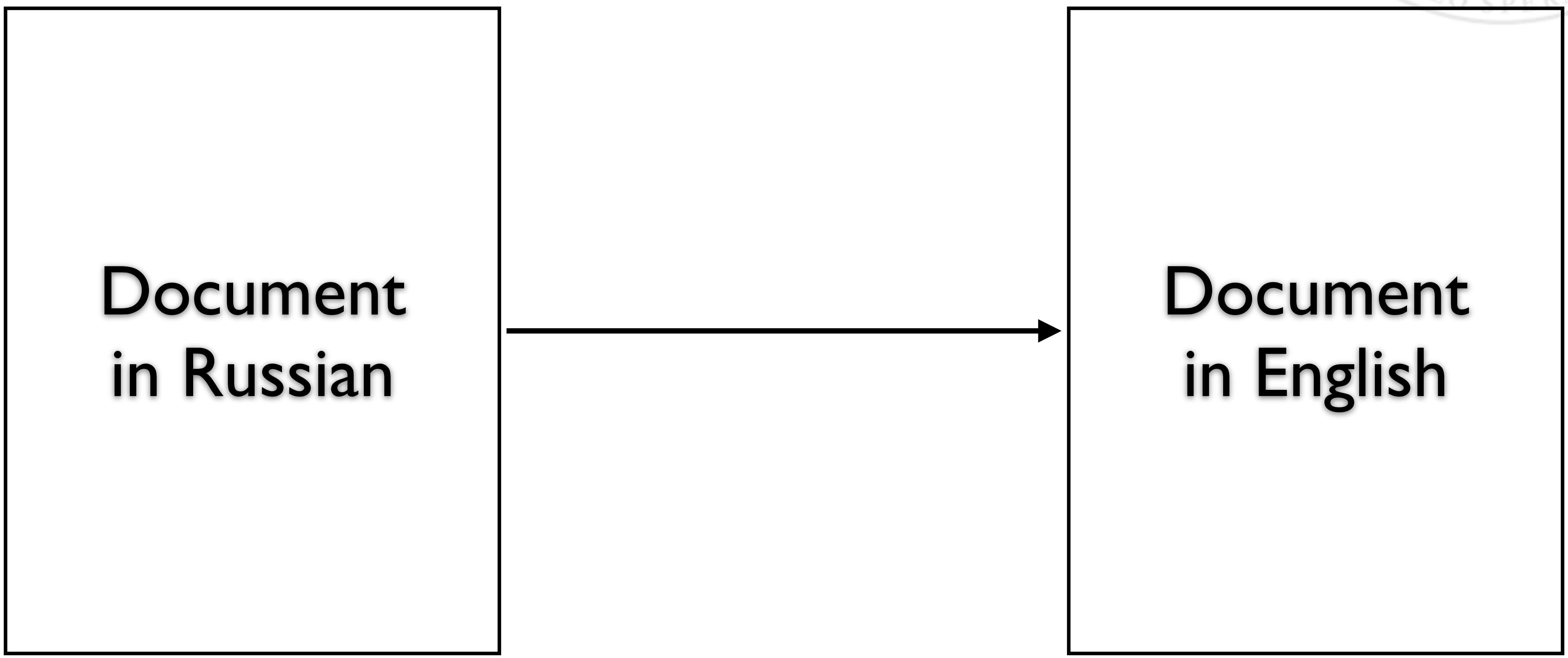
Document in Russian

Document in English

Competing Approaches



Competing Approaches



Google Translate

100 languages, 200 million people, 100 billion words daily