from neural networks to the structure of language: a physicist's perspective

Eric DeGiuli

Institut de Physique Théorique Philippe Meyer
École Normale Supérieure, Paris

successes of machine learning

speech recognition

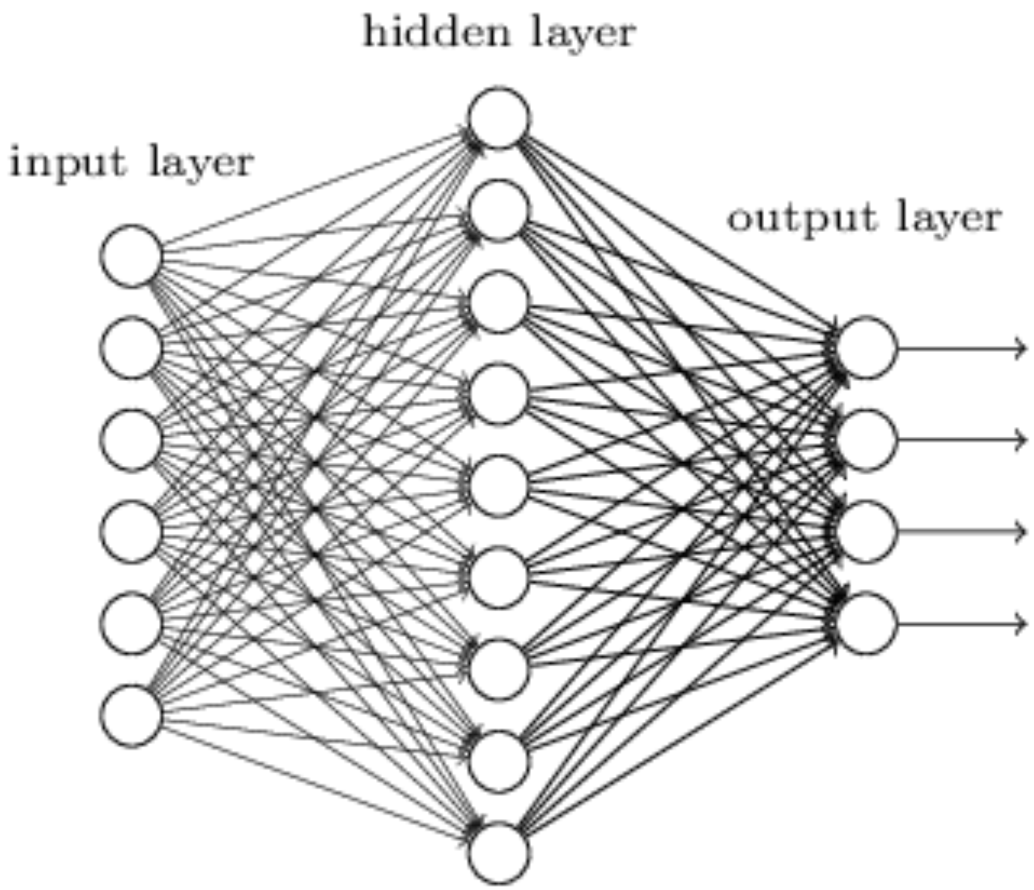
image recognition

Karpathy & Li. *Proc. IEEE CVPR* 2015

all using deep neural networks

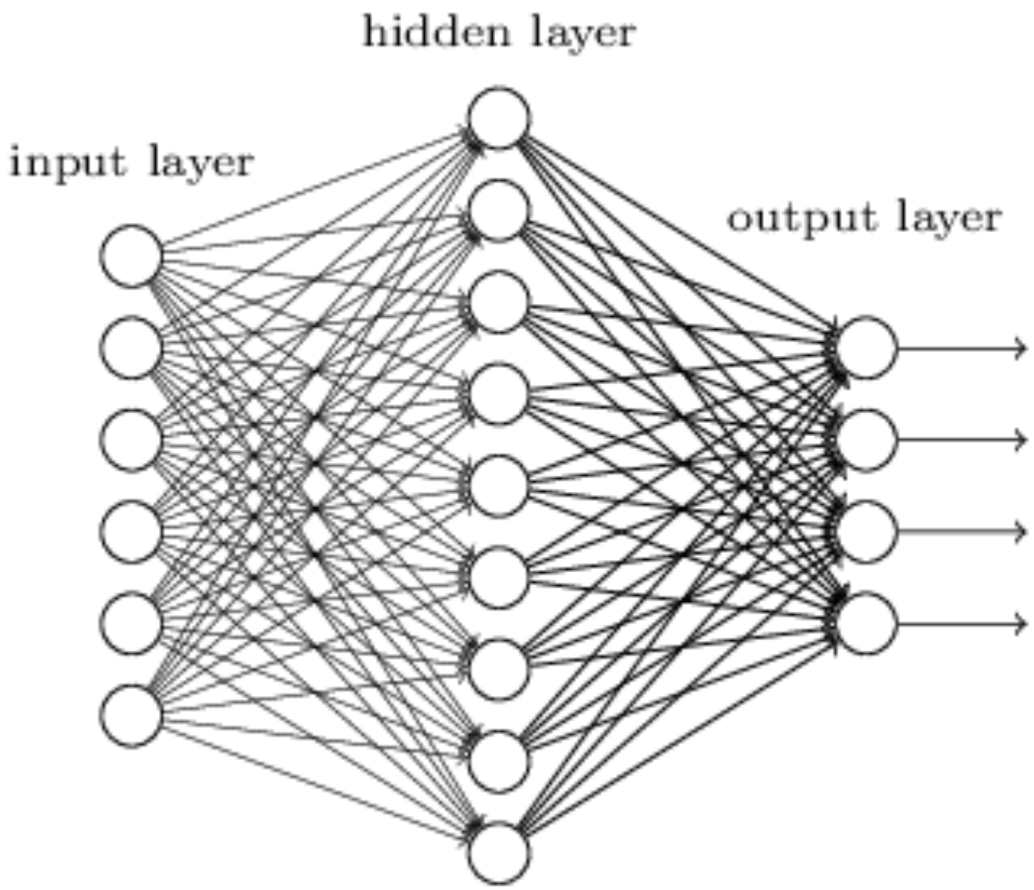
neural network architecture

x y p



neural network architecture

x y p



x = input

y = hidden variables
= f(Ax + b)

A = parameter matrix

b = offset parameter vector

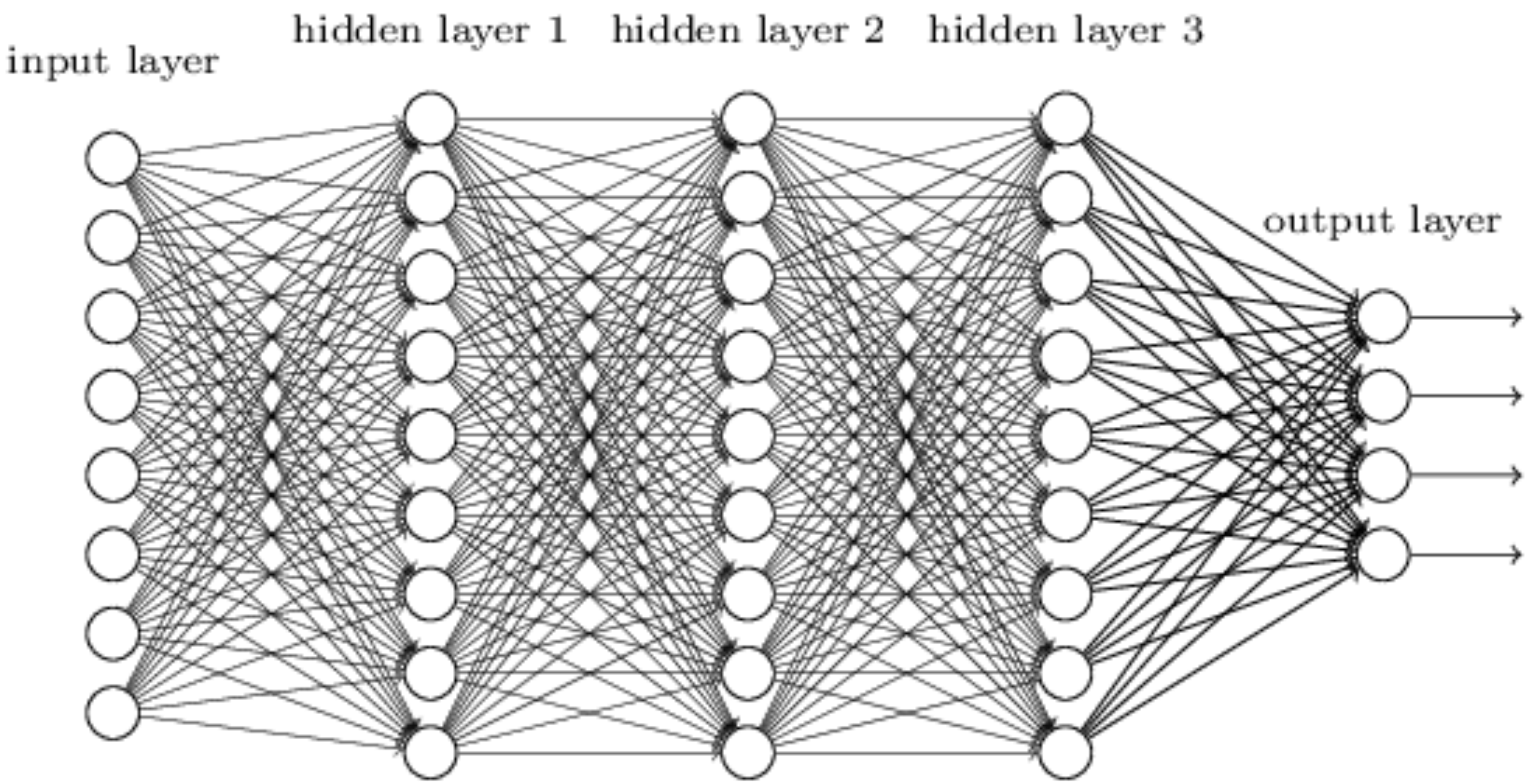
f = component-wise activation function

p = probabilities
= g(y)

goal of learning: determine optimal A, b

neural network architecture

x $y^{(1)}$ $y^{(2)}$ $y^{(3)}$ p



deep neural network = network with several hidden layers

activation

how to choose f ?
consider dp vs dx

x

y = f(Ax + b)

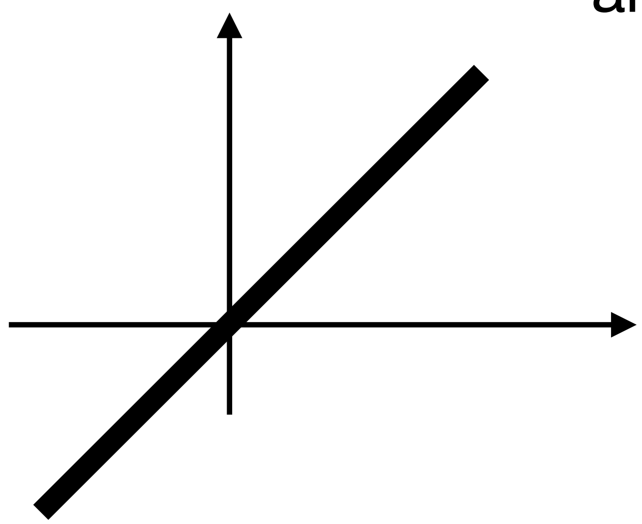
p=g(y)

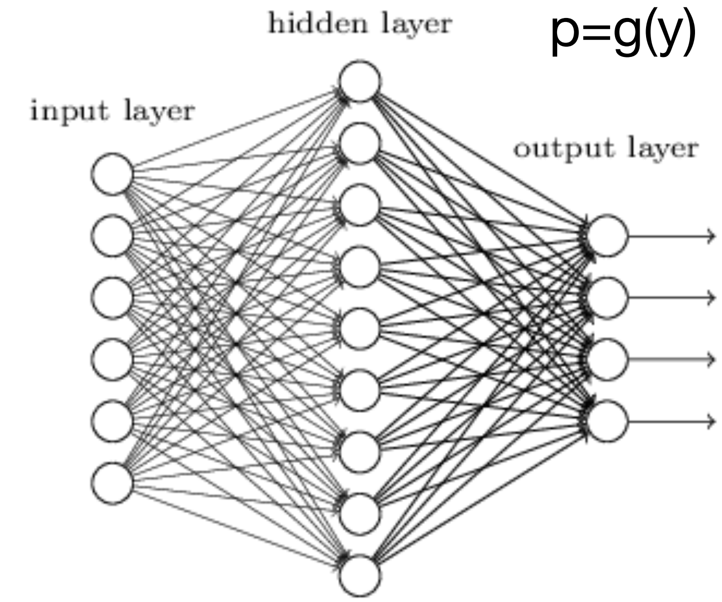
any change in input leads to change in y

all elements contribute to dp

$\boxed{\text{referendum machine}}$

linear





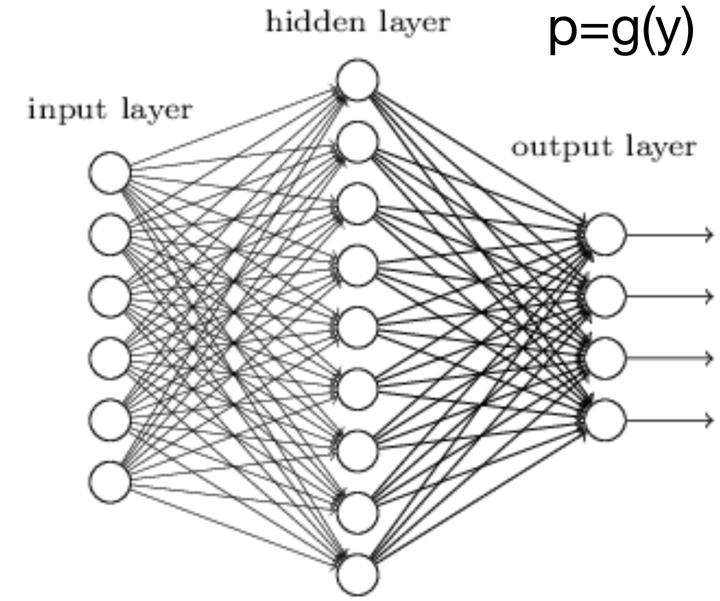
activation

how to choose f ?
consider dp vs dx

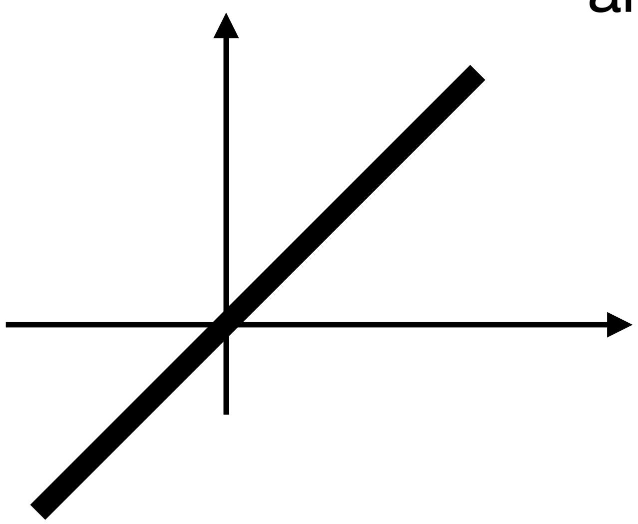
x

y = f(Ax + b)

p=g(y)



linear

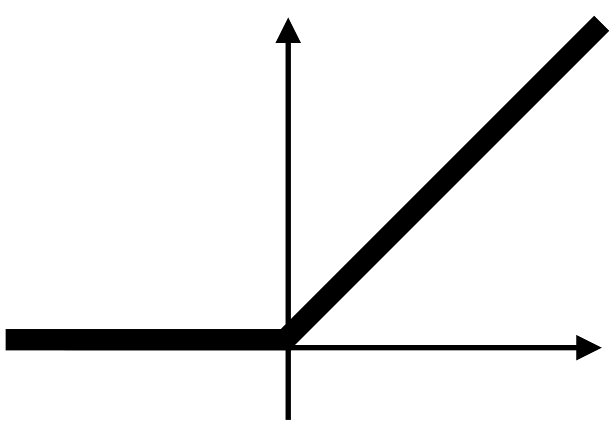


any change in input leads to change in y

all elements contribute to dp

referendum machine

nonlinear



common choice: clamp[1]

$$f(x) = \begin{cases} x & x \geq 0 \\ 0 & x < 0 \end{cases}$$

no change in output for x<0

elements can be indifferent

expert machine

[1] 'ReLU' = rectified linear

composition

- essential to have nonlinear activation
- saturation in activation ⇒ elements can act compositionally

  (expert elements rather than jack-of-all-trades)
- allows approximate factorization of data space ⇒ fewer parameter DOF
- theory? (see Tubiana Monasson PRL 2016)

1st principle of successful machine learning: composition

composition

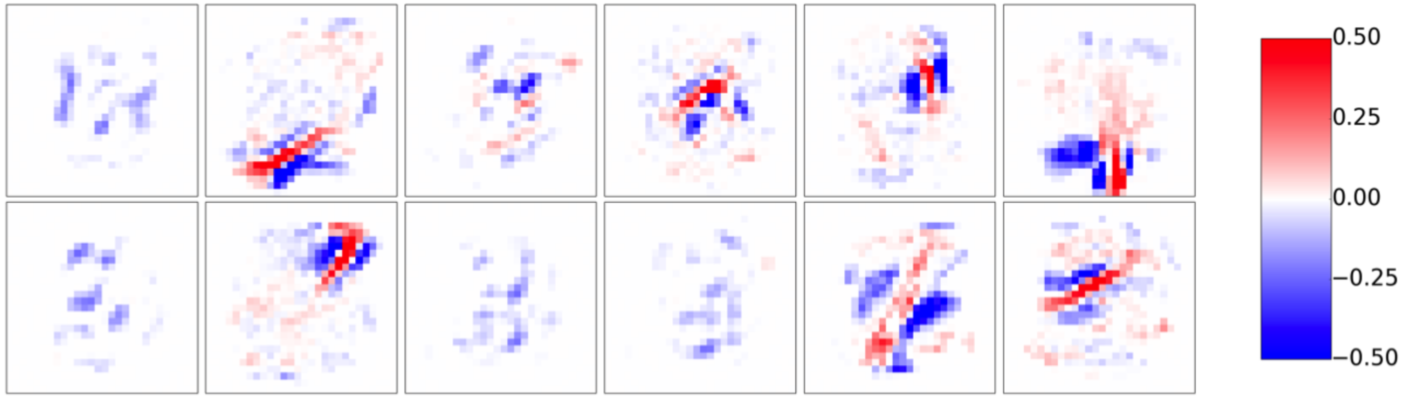
- essential to have nonlinear activation
- saturation in activation ⇒ elements can act compositionally

  (expert elements rather than jack-of-all-trades)
- allows approximate factorization of data space ⇒ fewer parameter DOF
- theory? (see Tubiana Monasson PRL 2016)

1[st] principle of successful machine learning: composition

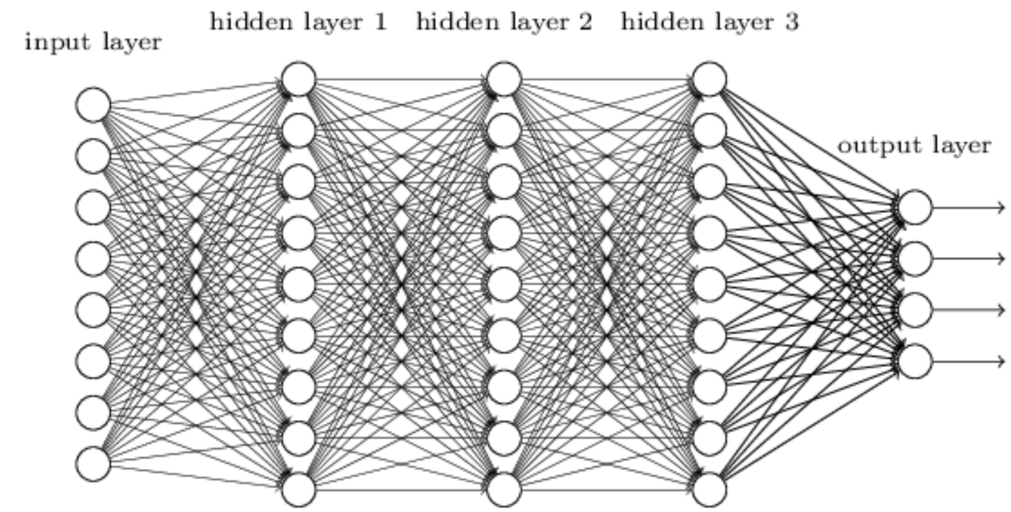
e.g. handwritten digits

hidden units ~ elementary strokes

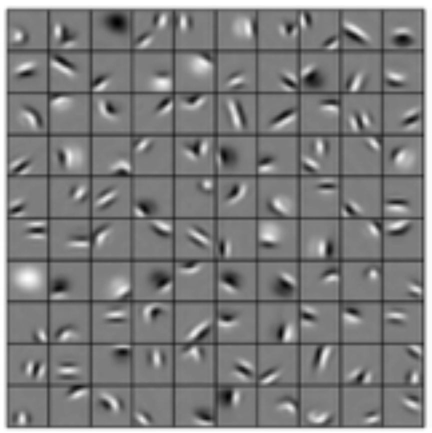


Tubiana Monasson 2016

hierarchy



what is role of deep structure?



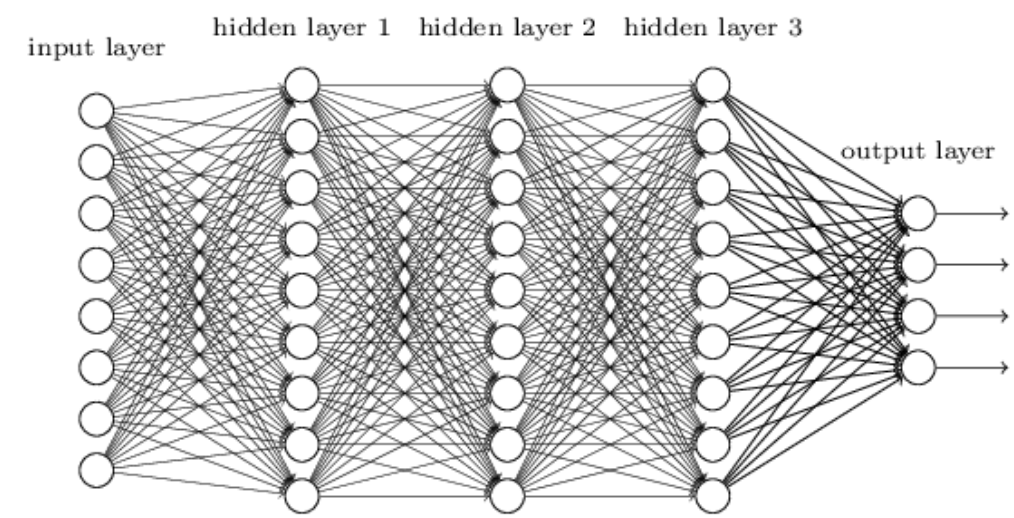
$y^{(1)}$

$y^{(2)}$

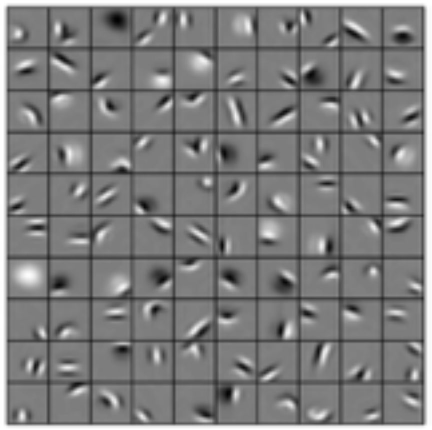
$y^{(3)}$

Lee, H, et al. *Comm. ACM* 54.10 (2011): 95-103.

hierarchy



what is role of deep structure?



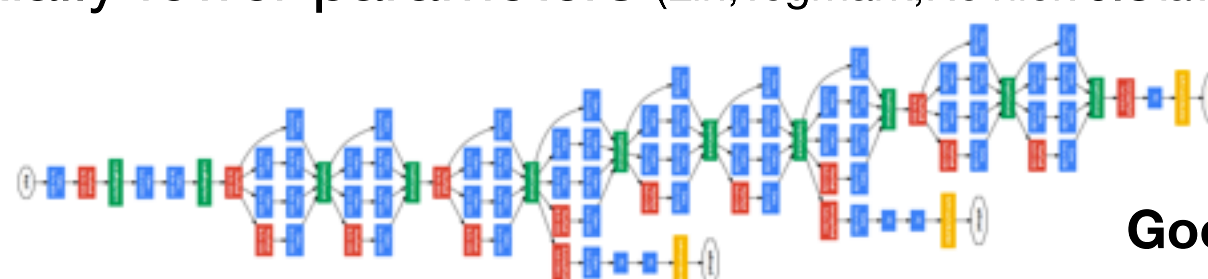
$y^{(1)}$

$y^{(2)}$

$y^{(3)}$

Lee, H, et al. *Comm. ACM* 54.10 (2011): 95-103.

- deep structure ⇒ hierarchical features
- can represent functions with exponentially fewer parameters (Lin,Tegmark,Rolnick J.Stat.Phys 2017)
- empirically, deeper = better

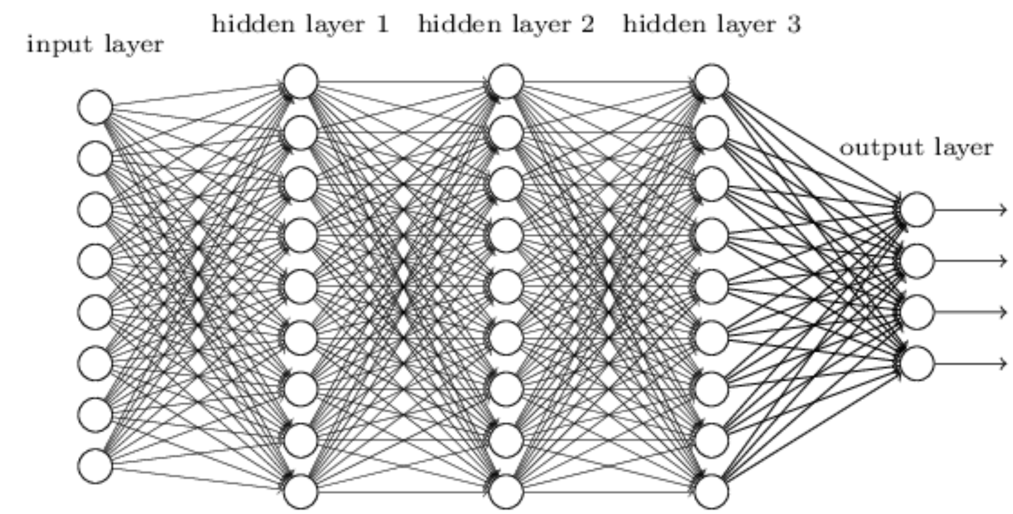


GoogLeNet

Szegedy et al. Cvpr, 2015.

2nd principle of successful machine learning: hierarchy

neural network paradigm



architecture: composition & hierarchy

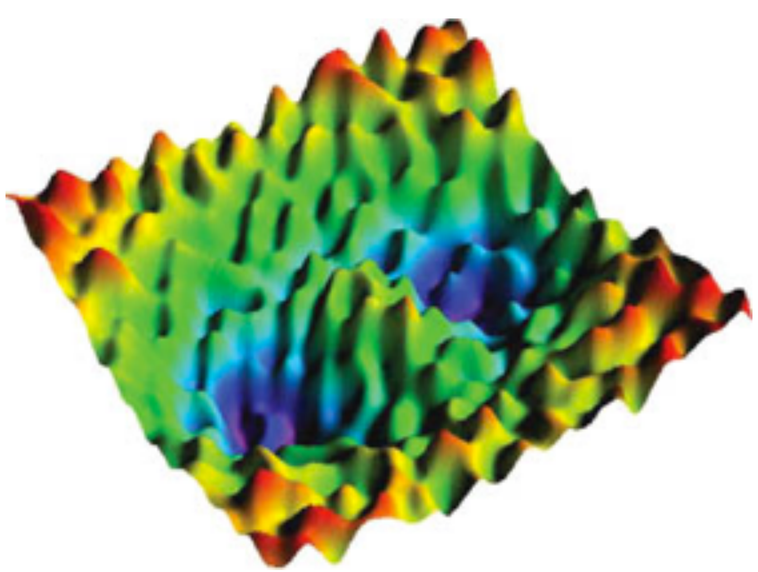
training?

consider `supervised' learning

have training data pairs (x_{data}, p_{data}) known exactly

define `energy' $E(A,b) = \Sigma\ (p(x_{data})-p_{data})^2$

- minimizing E is a disordered physics problem (disorder = fixed training data)
- do gradient descent
- phenomenology ~ classic glassy systems



what are the principles for learning?

limitations

neural networks now (2018) are still very far from human intelligence!

e.g. Winograd challenge

```
1. The city councilmen refused the demonstrators a permit because they feared violence.
2. The city councilmen refused the demonstrators a permit because they advocated violence.
```

Give 1. Ask `who feared violence?’

Give 2. Ask `who advocated violence?’

humans: > 90%

state-of-the-art (2016): 58%

what is the structure that makes these questions easy for us?

Quan Liu et al, of the University of Science and Technology, China

https://cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WS.html

a personal goal:

teach a machine to read & understand a book

why?

	2013	2014	2015	2016	
…	720,968	887,502	809,128	1,140,078	# articles added to PubMed each year

rigidity of language

1. Is John the man who is tall?

2. *Is John is the man who tall?

rigidity of language

1. Is John the man who is tall?
2. *Is John is the man who tall?
3. Colorless green ideas sleep furiously.
4. *Furiously sleep ideas green colorless.

Chomsky 1950s

rigidity of language

1. Is John the man who is tall?

2. *Is John is the man who tall?

3. Colorless green ideas sleep furiously.

4. *Furiously sleep ideas green colorless.

syntax = logical structure
semantics = `meaning' = connection to `truth'

Chomsky 1950s

formal grammars
(Pāṇini 400BC, Chomsky, Backus 1950s)

grammar[1] = set of string rewriting rules

A,B,C,…. hidden[2] symbols

a,b,c,…. observable[3] symbols

begin with start symbol, S

repeatedly apply rules until string of observables

[1] grammar = ‘generative grammar’ [2] ‘nonterminal’ [3] ‘terminal’

formal grammars

(Pāṇini 400BC, Chomsky, Backus 1950s)

grammar[1] = set of string rewriting rules

A,B,C,…. hidden[2] symbols

a,b,c,…. observable[3] symbols

e.g.
$S \rightarrow SS$
$S \rightarrow aSb$
$S \rightarrow ab$

begin with start symbol, S

repeatedly apply rules until string of observables

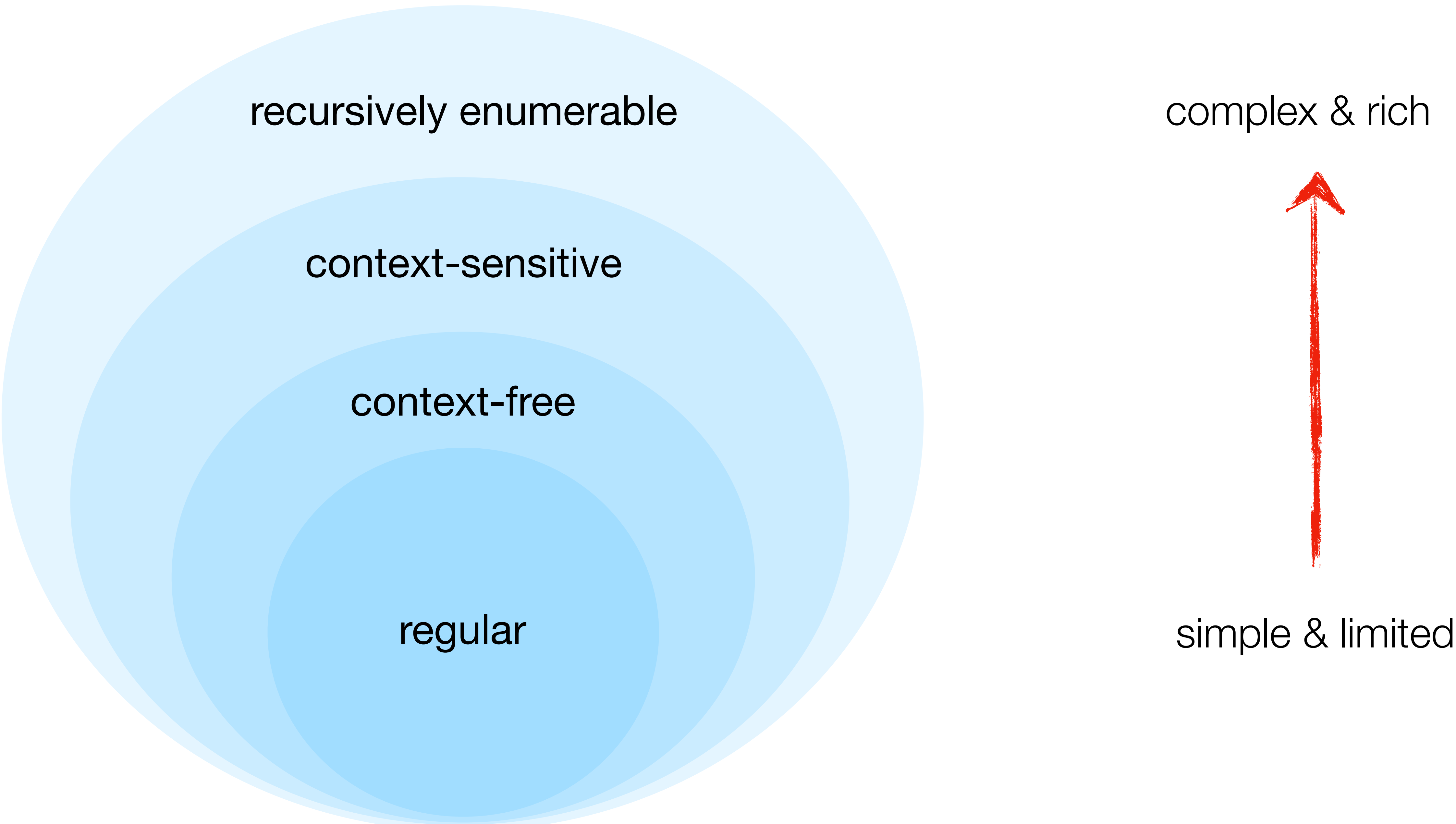
$S \rightarrow SS \rightarrow aSbS \rightarrow aabbS \rightarrow aabbab$

equivalent to (( )) ( )

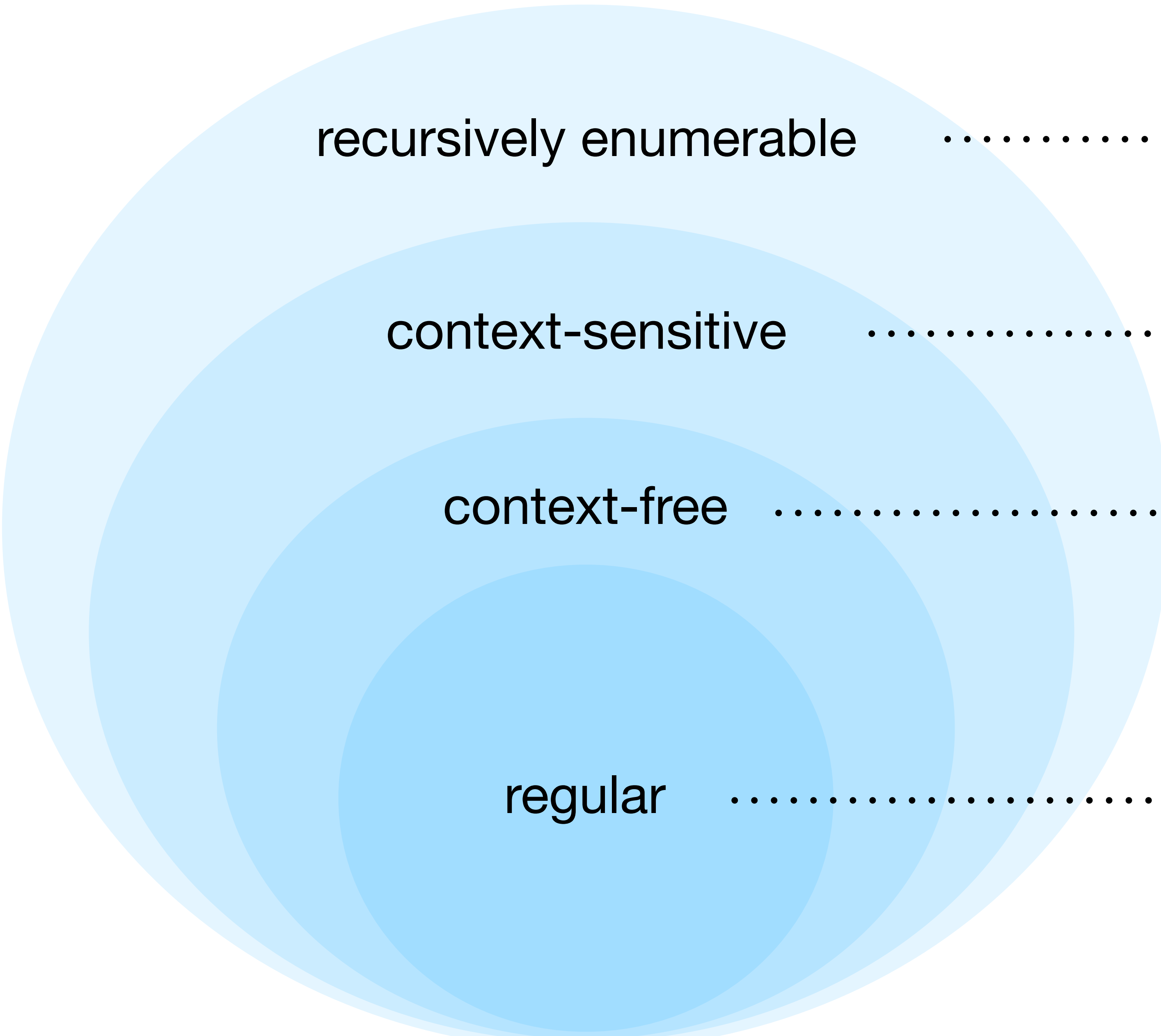
language = set of observable strings

[1] grammar = ‘generative grammar’ [2] ‘nonterminal’ [3] ‘terminal’

Chomsky hierarchy (1950’s)



Chomsky hierarchy (1950's)



automaton with infinite memory [1]

automaton with finite memory [2]

automaton with stack memory [3]

finite-state automaton

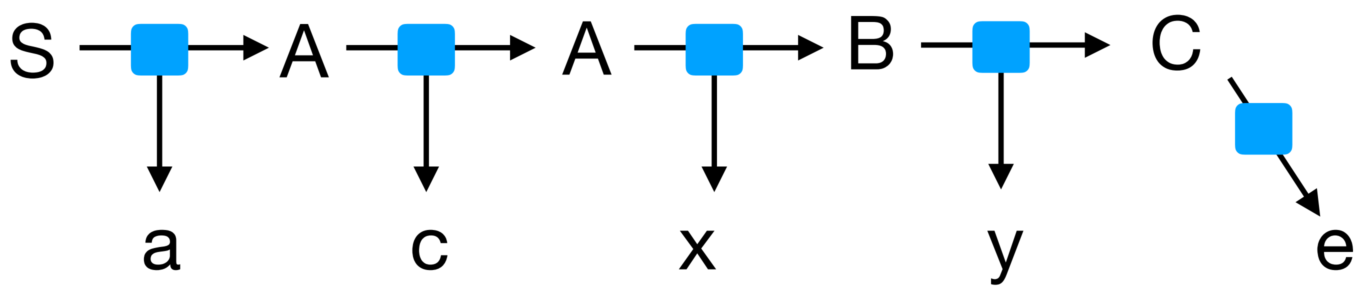
[1] Turing machine

[2] linear-bounded non-deterministic Turing machine

[3] non-deterministic pushdown automaton

structure of derivations

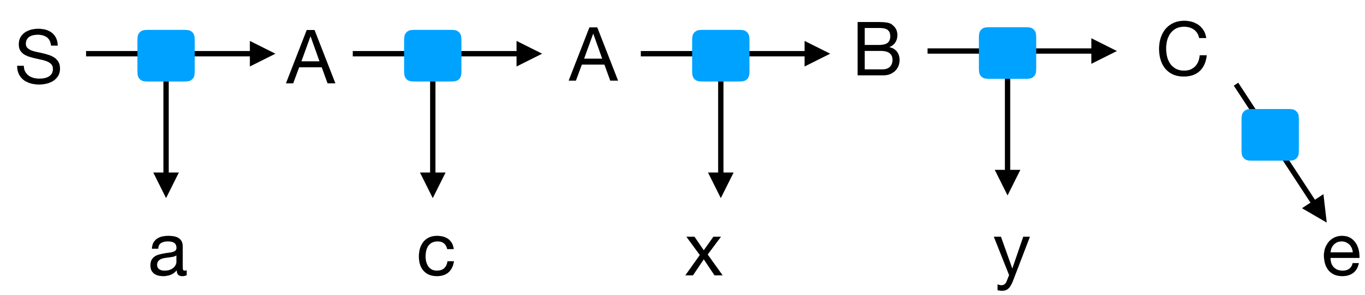
regular grammar:



- always linear
- used in computer science (e.g. search patterns)

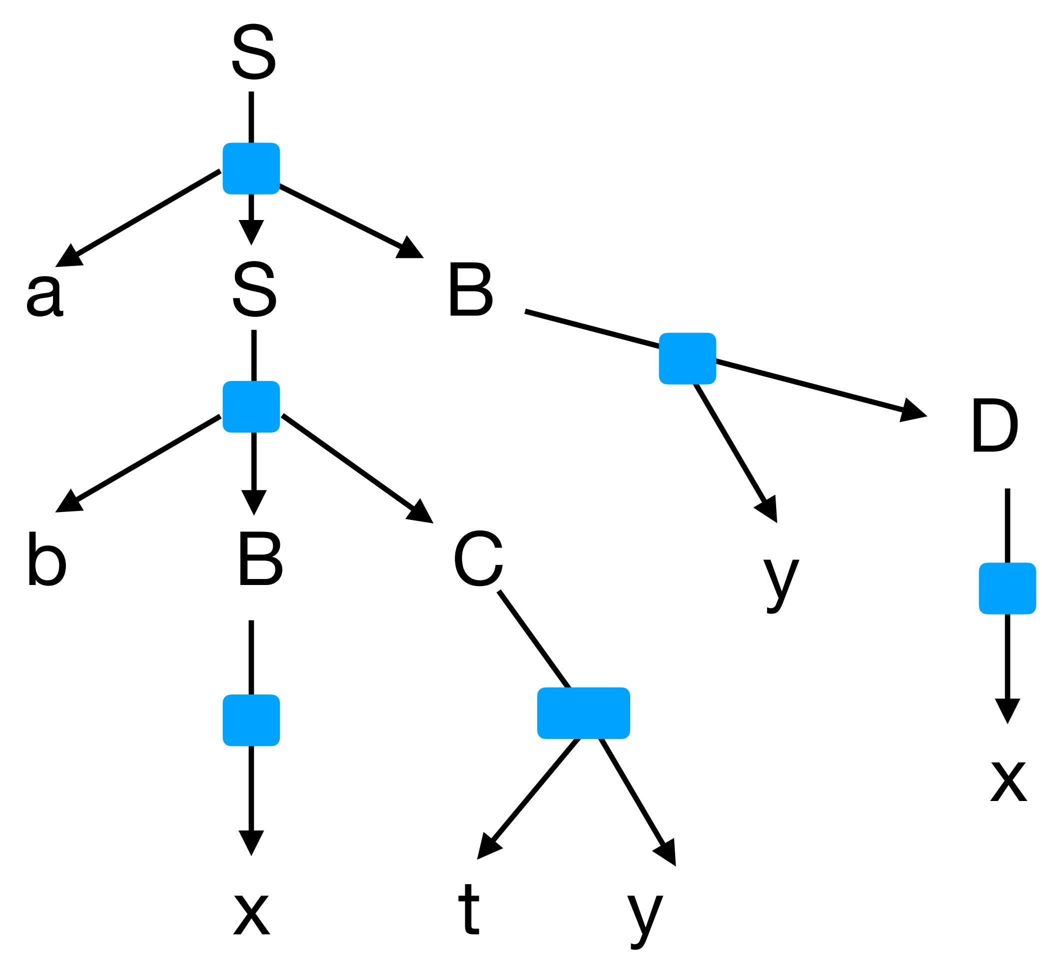
structure of derivations

regular grammar:



- always linear
- used in computer science (e.g. search patterns)

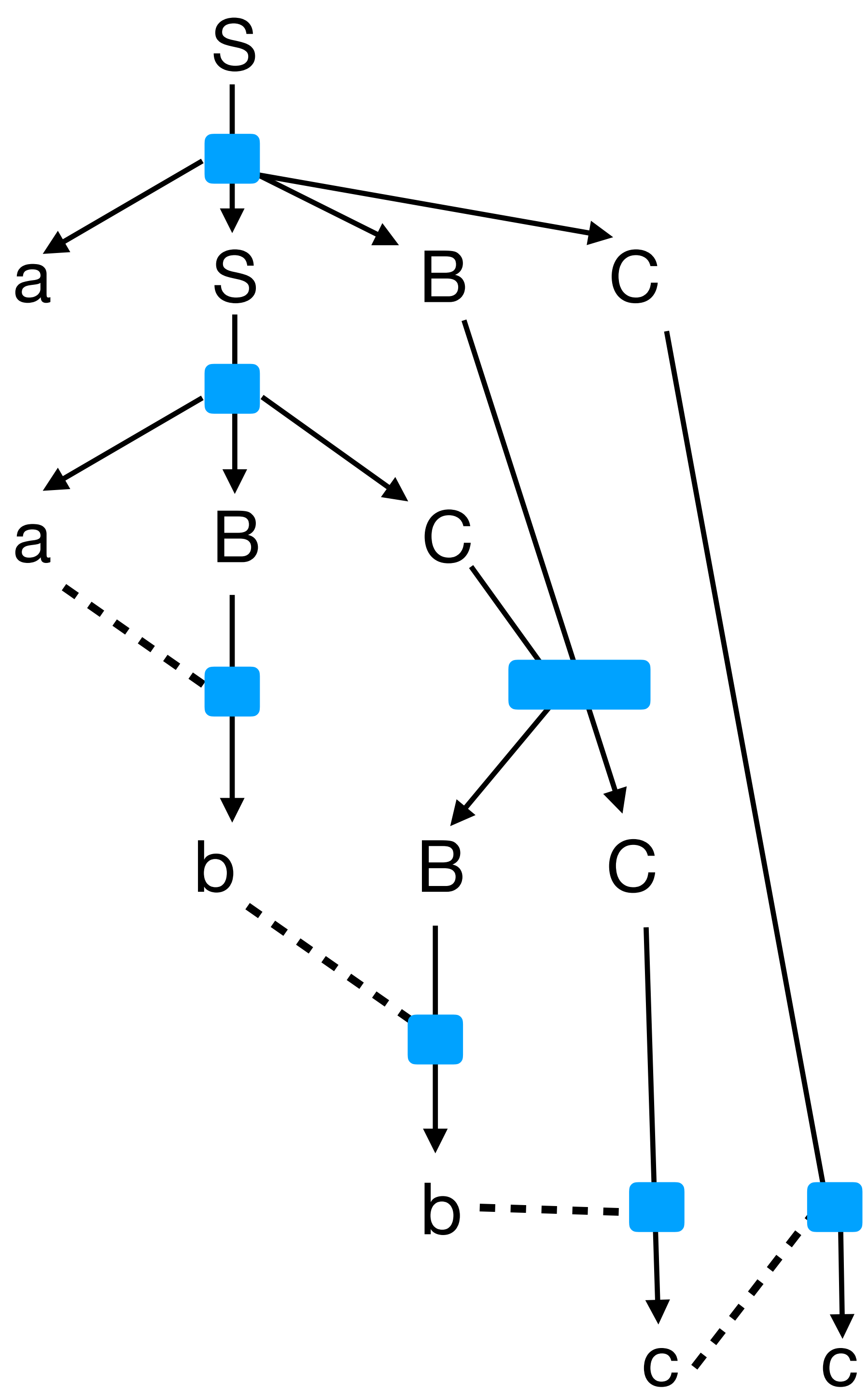
context-free grammar:



- always a tree
- used in linguistics for phrase structure (Chomsky 1956)
- central to computer science since Backus-Naur works ~1960

structure of derivations

context-sensitive grammar:



S ⇒ aSBC ⇒ aaBCBC ⇒ aabCBC
⇒ aabBCC ⇒ aabbCC ⇒ aabbcC
⇒ aabbcc

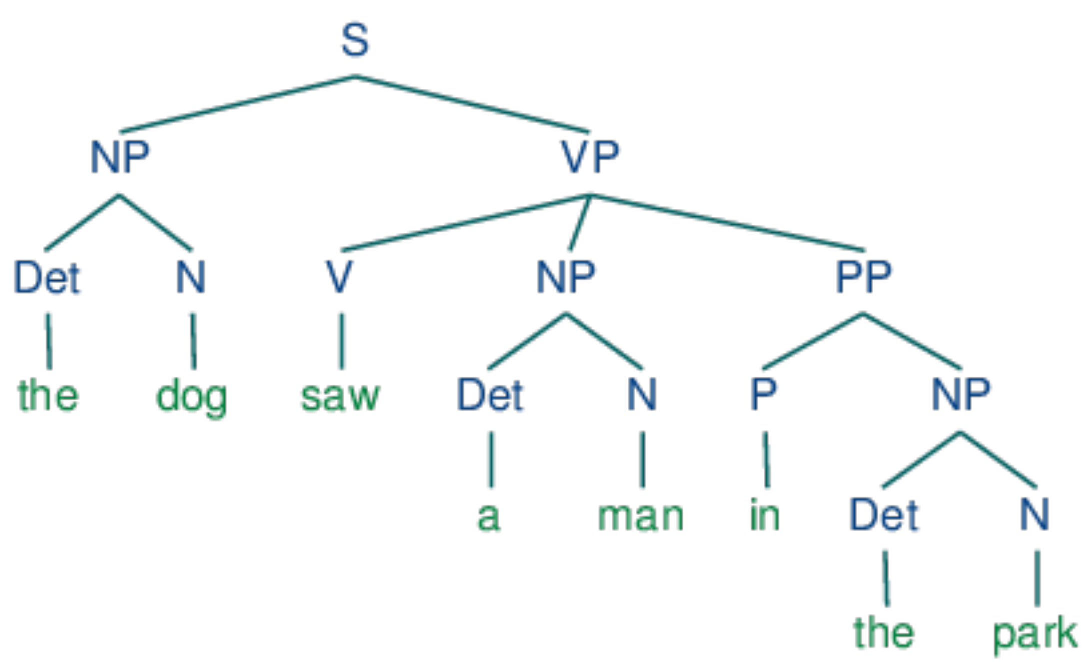
grammar:

S → aSBC
S → aBC
CB → BC
aB → ab
bB → bb
bC → bc
cC → cc

what about natural languages?

- ~7000 existing languages
- only 2 have confirmed non-context-free features (Swiss-German, Bambara)

i.e. context-free languages define an
ensemble for natural language syntax

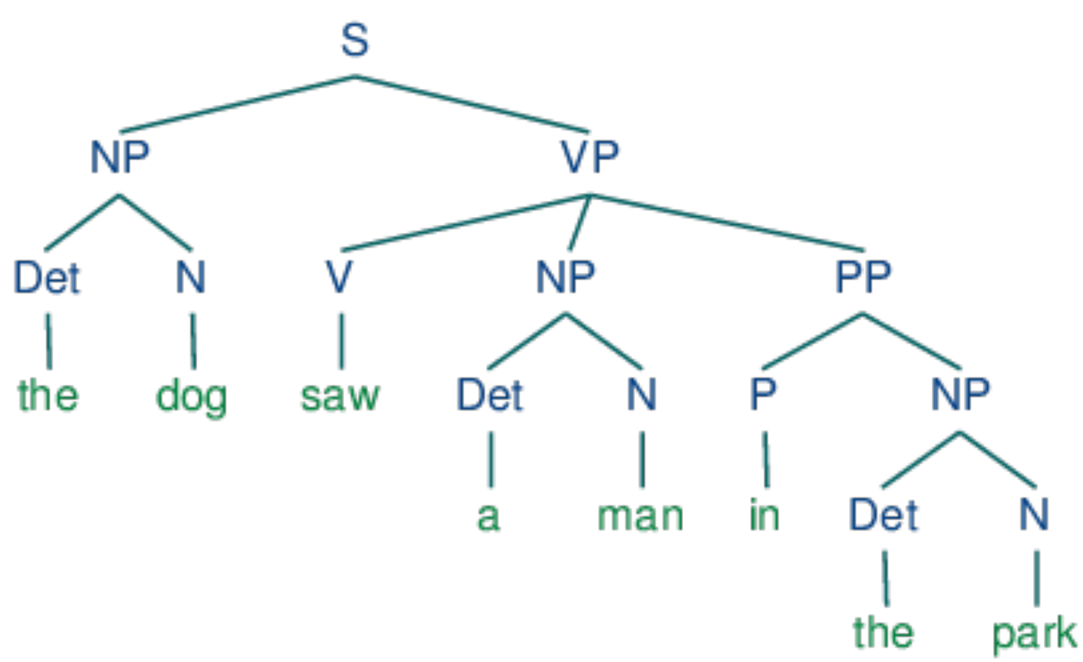


Pullum & Gazdar 1982, Shieber 1985, Culy 1985

what about natural languages?

- ~7000 existing languages
- only 2 have confirmed non-context-free features (Swiss-German, Bambara)

i.e. context-free languages define an *ensemble* for natural language syntax

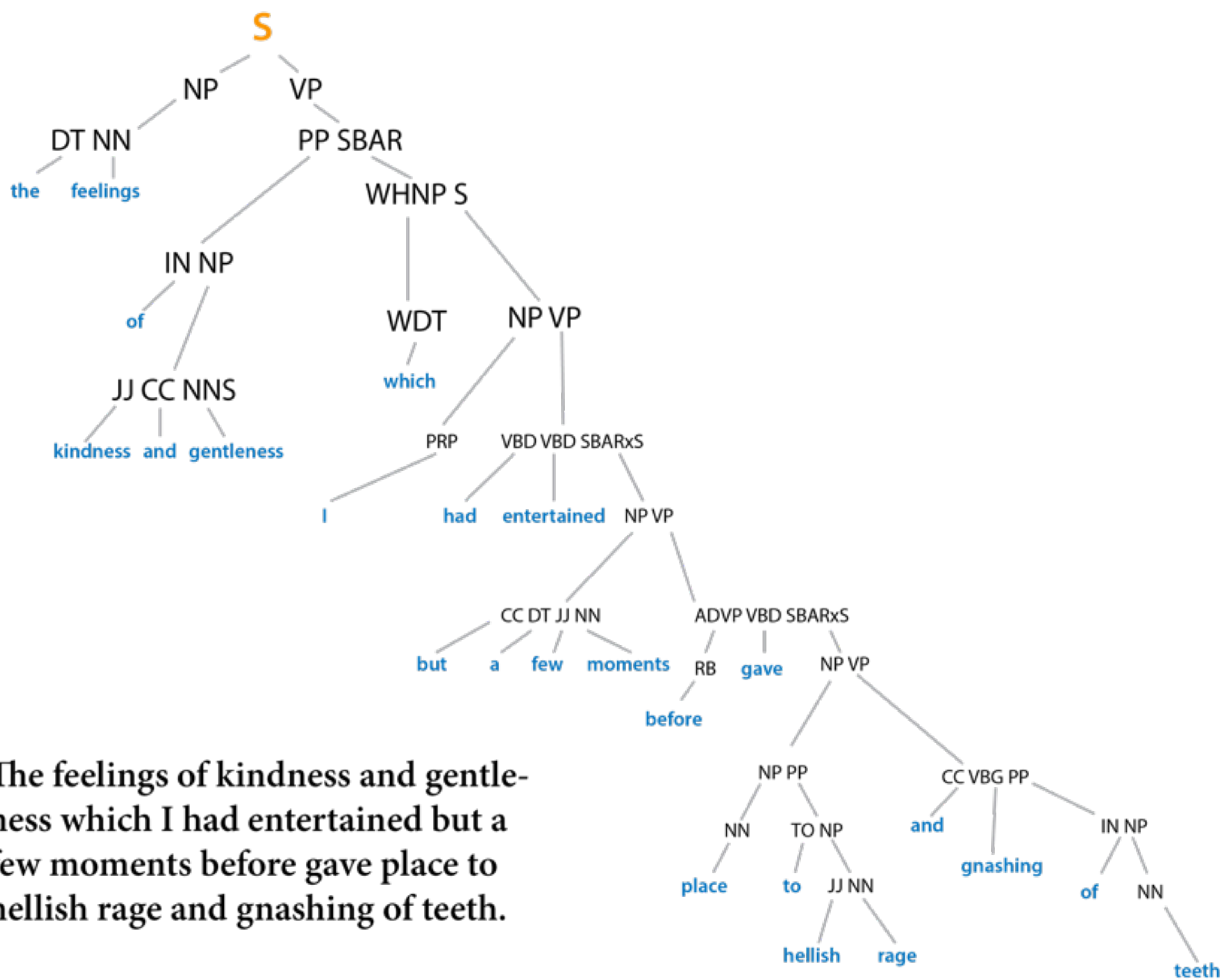


meaning of the tree?

‘the park’ behaves like ‘park’

‘in the park’ behaves like ‘in—noun’

Pullum & Gazdar 1982, Shieber 1985, Culy 1985



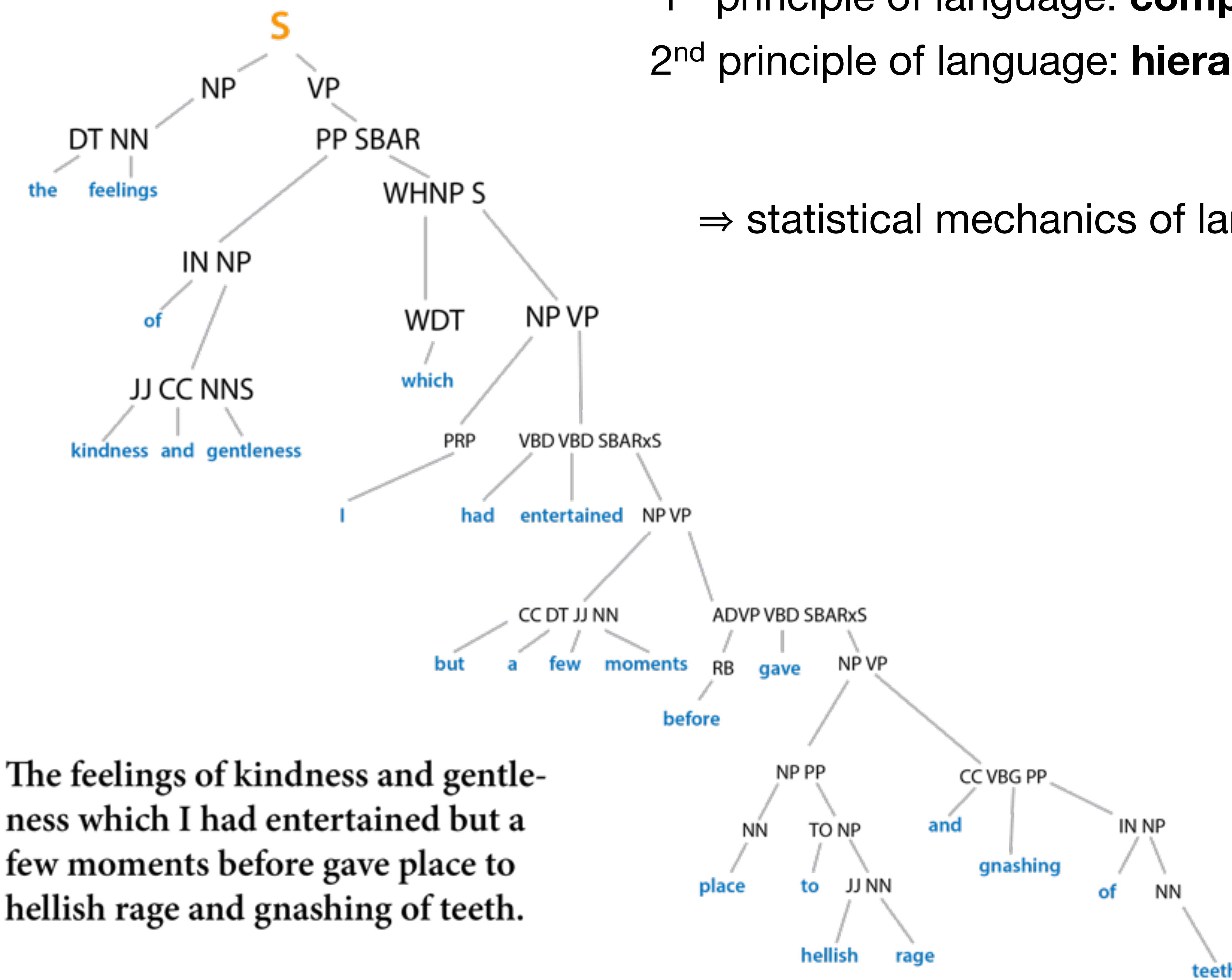
The feelings of kindness and gentleness which I had entertained but a few moments before gave place to hellish rage and gnashing of teeth.

W. Gilpin, online 2017

1st principle of language: **composition**

2nd principle of language: **hierarchy**

⇒ statistical mechanics of language !

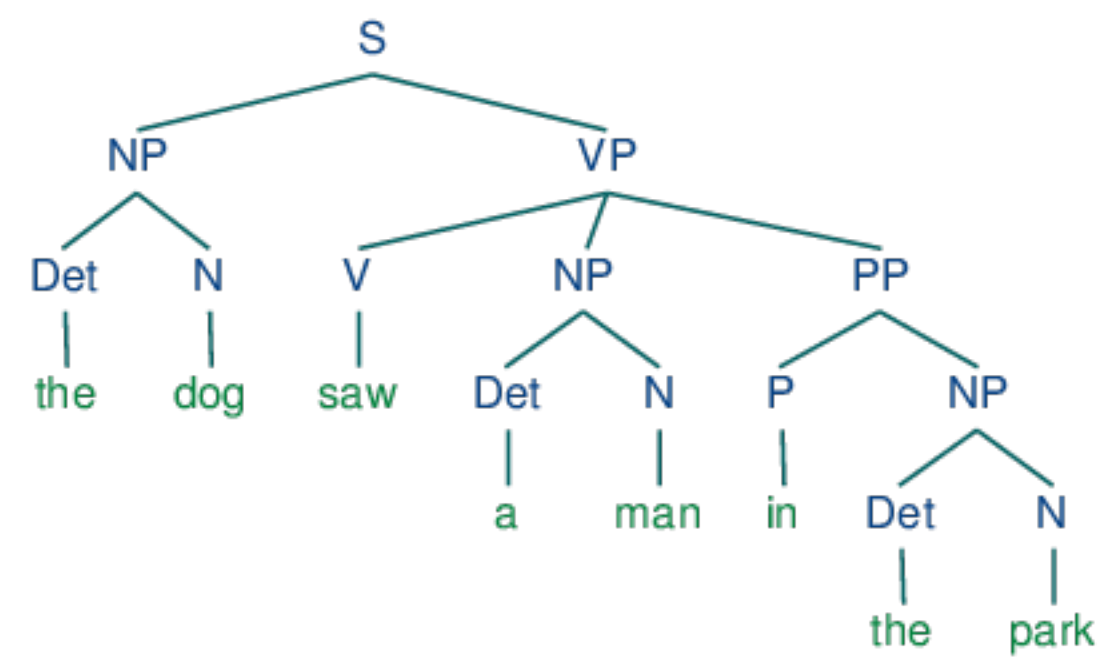


W. Gilpin, online 2017

random language model

can we understand something about *typical* context-free grammars?

1. can assume binary tree[1]



all rules either $A \rightarrow BC$ or $A \rightarrow b$

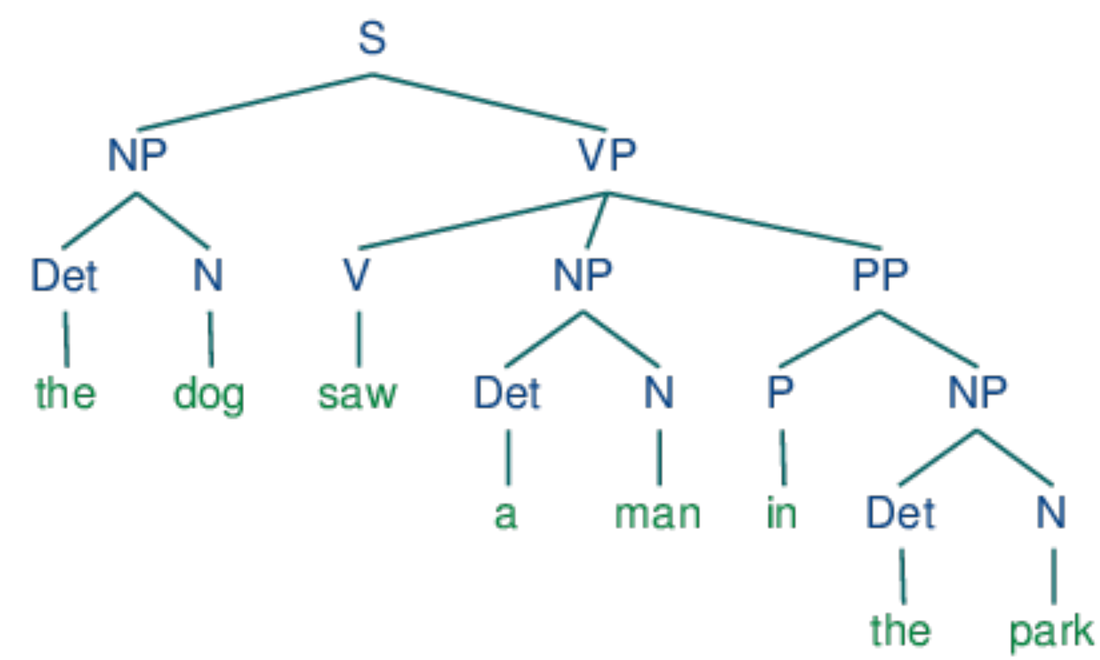
[1] binary tree = 'Chomsky normal form'

random language model

can we understand something about *typical* context-free grammars?

1. can assume binary tree[1]

all rules either A → BC or A → b



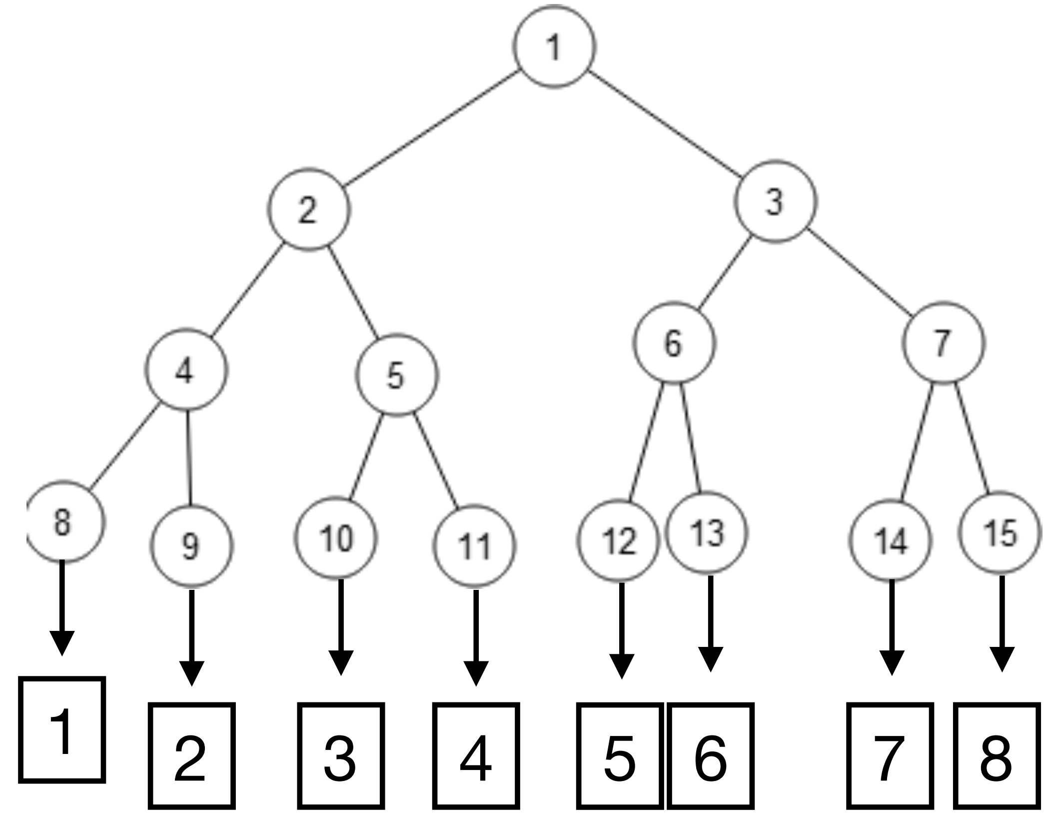
2. so far, rules have been yes/no. let rules → conditional probabilities

then a grammar is defined by

$$M_{ABC} = \mathbb{P}(A \to BC \mid A \to \text{hidden}),$$
$$O_{Ab} = \mathbb{P}(A \to b \mid A \to \text{observable}),$$

[1] binary tree = ‘Chomsky normal form’

random language model



for simplicity, fix tree topology

$$M_{ABC} = \mathbb{P}(A \to BC \mid A \to \text{hidden}),$$
$$O_{Ab} = \mathbb{P}(A \to b \mid A \to \text{observable}),$$

$$M_{\sigma_i\sigma_j\sigma_k} = \mathbb{P}(\sigma_i \to \sigma_j\sigma_k | \sigma_i, \sigma_j, \sigma_k \in \chi_N),$$
$$O_{\sigma_i o_j} = \mathbb{P}(\sigma_i \to o_j | \sigma_i \in \chi_N, o_j \in \chi_T),$$

$$\mathbb{P}(\{\sigma_i, o_t\}|M, O, \mathcal{T}) = \quad P_{\sigma_0} \prod_{\alpha\in\Omega} M_{\sigma_{\alpha_1}\sigma_{\alpha_2}\sigma_{\alpha_3}} \prod_{\alpha\in\partial\Omega} O_{\sigma_{\alpha_1}o_{\alpha_2}},$$

note: M,O are probabilities for a fixed grammar, then we have an ensemble of grammars

random language model

what is the physics?

$$M_{\sigma_i\sigma_j\sigma_k} = \mathbb{P}(\sigma_i \to \sigma_j\sigma_k | \sigma_i, \sigma_j, \sigma_k \in \chi_N),$$
$$O_{\sigma_i o_j} = \mathbb{P}(\sigma_i \to o_j | \sigma_i \in \chi_N, o_j \in \chi_T),$$

$$Z = \int DM \int DO \sum_{\mathcal{T}} \sum_{\{\sigma\}} \sum_{\{o\}} e^{\log \mathbb{P}}$$

Z contains all the context-free grammars & all grammatical sentences in the universe!

random language model

what is the physics?

$$M_{\sigma_i\sigma_j\sigma_k} = \mathbb{P}(\sigma_i \to \sigma_j\sigma_k | \sigma_i, \sigma_j, \sigma_k \in \chi_N),$$
$$O_{\sigma_i o_j} = \mathbb{P}(\sigma_i \to o_j | \sigma_i \in \chi_N, o_j \in \chi_T),$$

$$Z = \int DM \int DO \sum_{\mathcal{T}} \sum_{\{\sigma\}} \sum_{\{o\}} e^{\log \mathbb{P}}$$

impose normalization of probabilities
& # of nonzero rules

Z contains all the context-free grammars & all grammatical sentences in the universe!

random language model

what is the physics?

$$M_{\sigma_i\sigma_j\sigma_k} = \mathbb{P}(\sigma_i \to \sigma_j\sigma_k | \sigma_i, \sigma_j, \sigma_k \in \chi_N),$$
$$O_{\sigma_i o_j} = \mathbb{P}(\sigma_i \to o_j | \sigma_i \in \chi_N, o_j \in \chi_T),$$

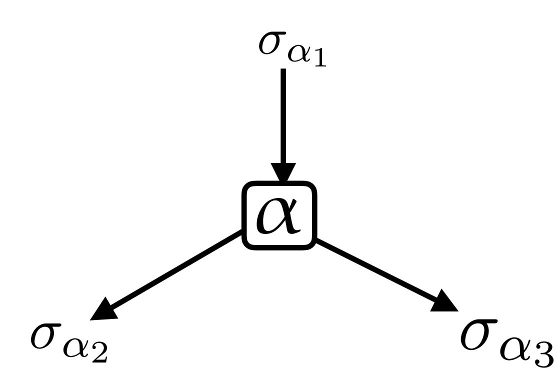
$$Z = \int DM \int DO \sum_{\mathcal{T}} \sum_{\{\sigma\}} \sum_{\{o\}} e^{\log \mathbb{P}}$$

Z contains all the context-free grammars & all grammatical sentences in the universe!

impose normalization of probabilities
& # of nonzero rules

$$\log \mathbb{P}(\{\sigma_i, o_t\} | M, O, \mathcal{T}) = \log P_{\sigma_0} + \sum_{\alpha \in \Omega} \log M_{\sigma_{\alpha_1}\sigma_{\alpha_2}\sigma_{\alpha_3}} + \sum_{\alpha \in \partial\Omega} \log O_{\sigma_{\alpha_1} o_{\alpha_2}}$$

looks a bit like a spin model … except



random language model

remarkably we can count all the context-free grammars in the universe

$$Z = \int DM \int DO \sum_{\mathcal{T}} \sum_{\{\sigma\}} \sum_{\{o\}} e^{\log \mathbb{P}}$$

what is the miracle? discrete Fourier transform

$$Z = Z_0 \sum_{\mathcal{T}} \sum_{\{\sigma\}} \sum_{\{o\}} e^{-H}$$

random language model

remarkably we can count all the context-free grammars in the universe

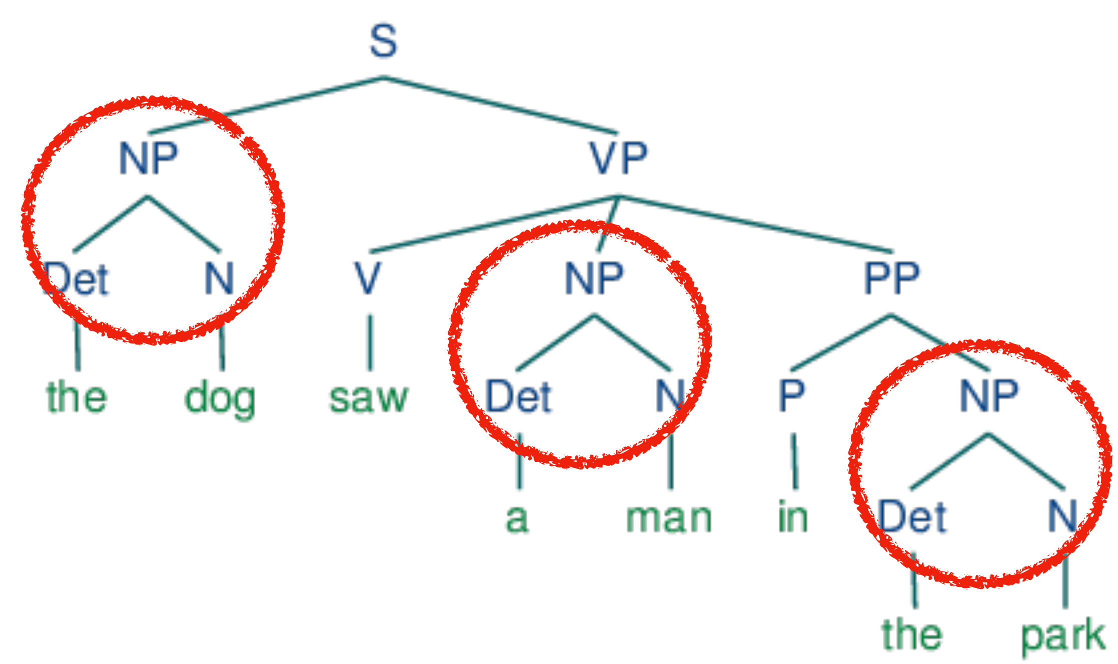
$$Z = \int DM \int DO \sum_{\mathcal{T}} \sum_{\{\sigma\}} \sum_{\{o\}} e^{\log \mathbb{P}}$$

what is the miracle? discrete Fourier transform

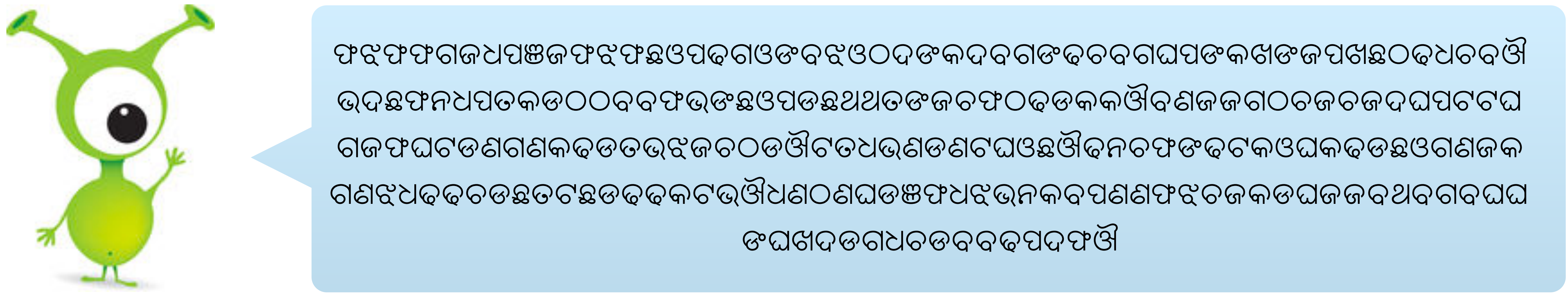
$$Z = Z_0 \sum_{\mathcal{T}} \sum_{\{\sigma\}} \sum_{\{o\}} e^{-H}$$

$$H = -k \sum_{\alpha,\beta \in \Omega} (N^2 \delta_{\sigma|\alpha,\sigma|\beta} - \delta_{\sigma_{\alpha_1},\sigma_{\beta_1}}) + \cdots$$

~ Potts model

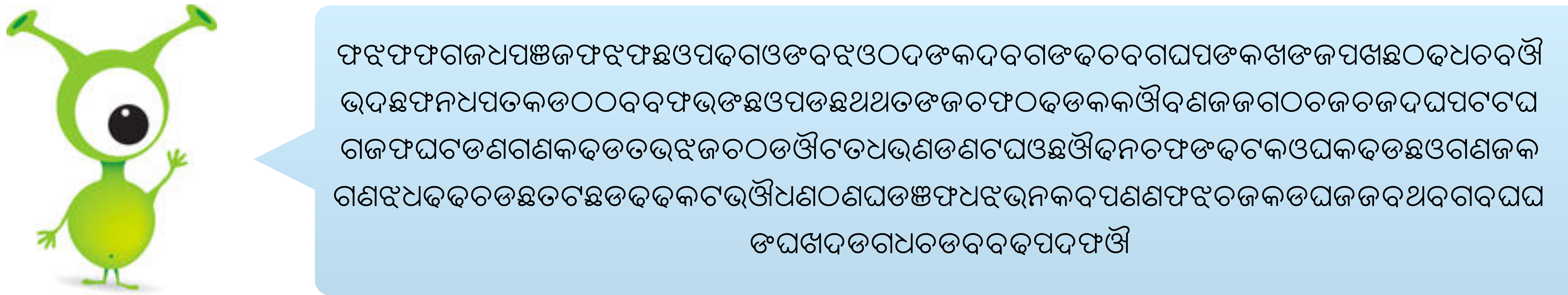


the SETI problem



can we learn its language?

the SETI problem

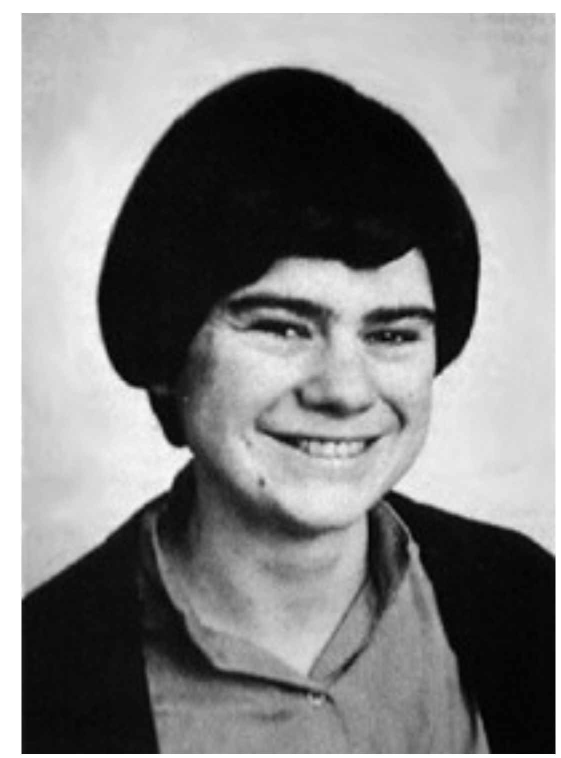


can we learn its language?

assume it is generated by a CFG

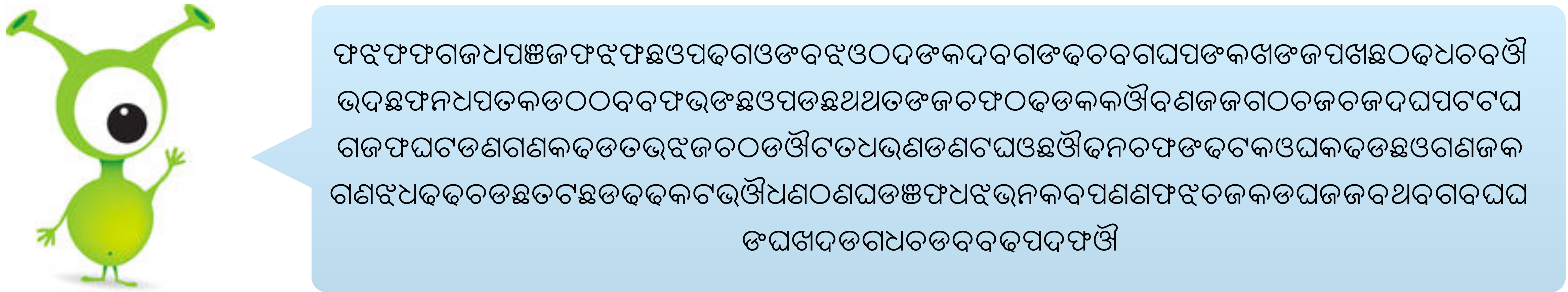
count number of grammars for which text is grammatical [1]

‘Gardner’ volume

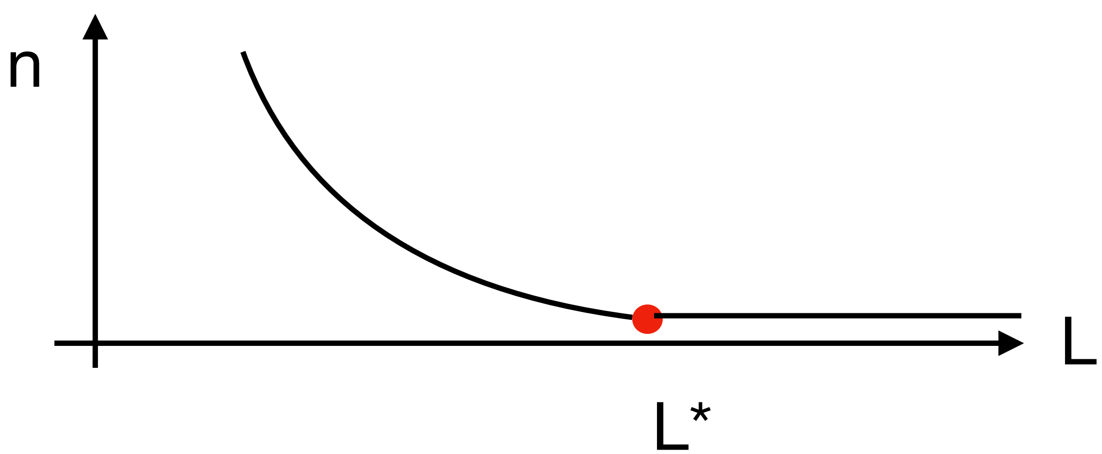


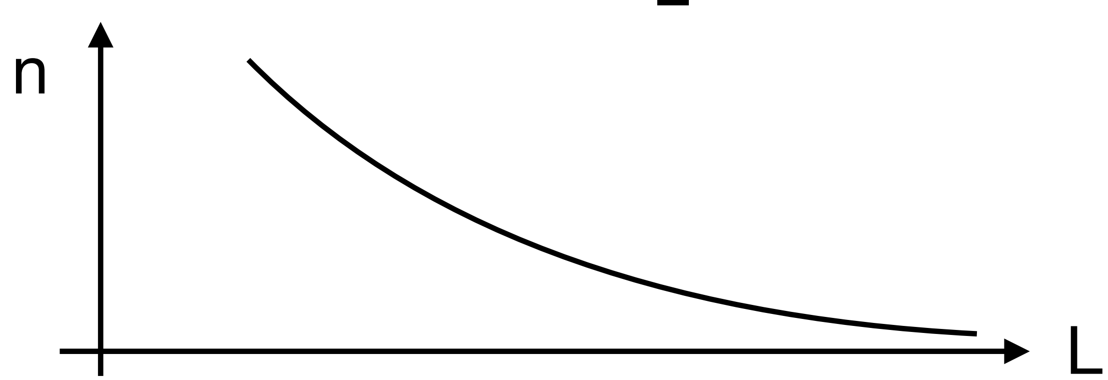
[1] more precisely, below some threshold K in probability

the SETI problem

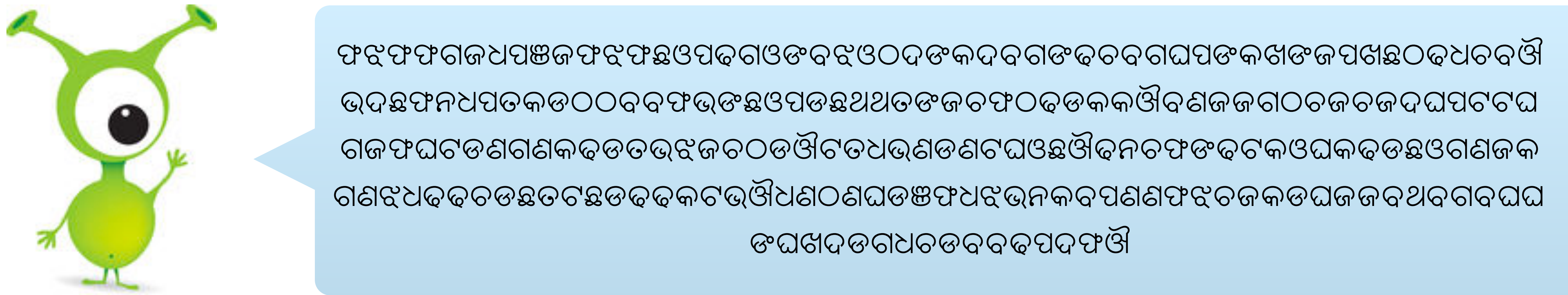


what do we expect?

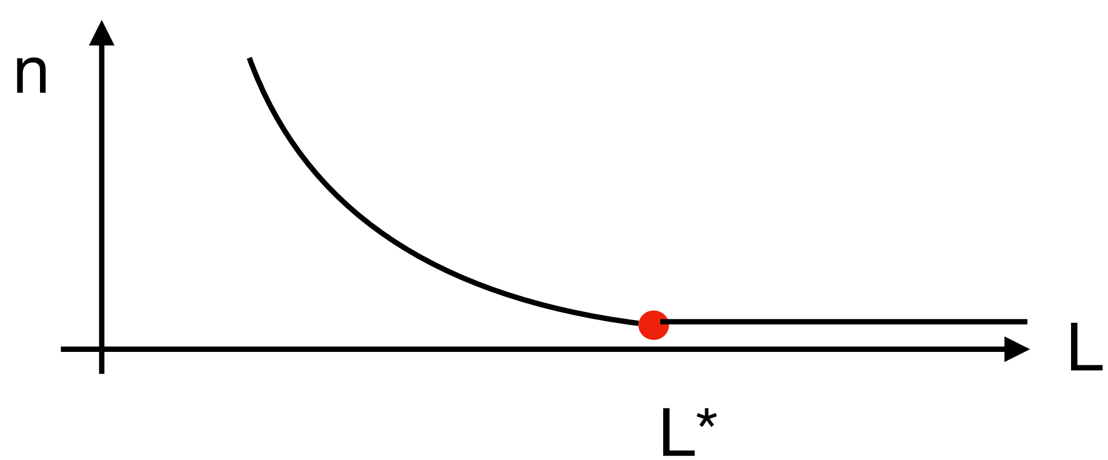




the SETI problem



what do we expect?



$$L = \frac{\ell}{N_{DOF}}$$

I am working on the full solution..

in a simple (wrong) approximation it is equivalent to Gardner's result for the perceptron

perspectives

ambiguity:

For a typical sentence, how many grammatical parses are there?

If n = 1, sentence is unambiguous
If n > 1, sentence is ambiguous
If n = 0, sentence is ungrammatical

Natural languages are typically ambiguous, e.g.
"Two cars were reported stolen by the Groverton police yesterday" [1]

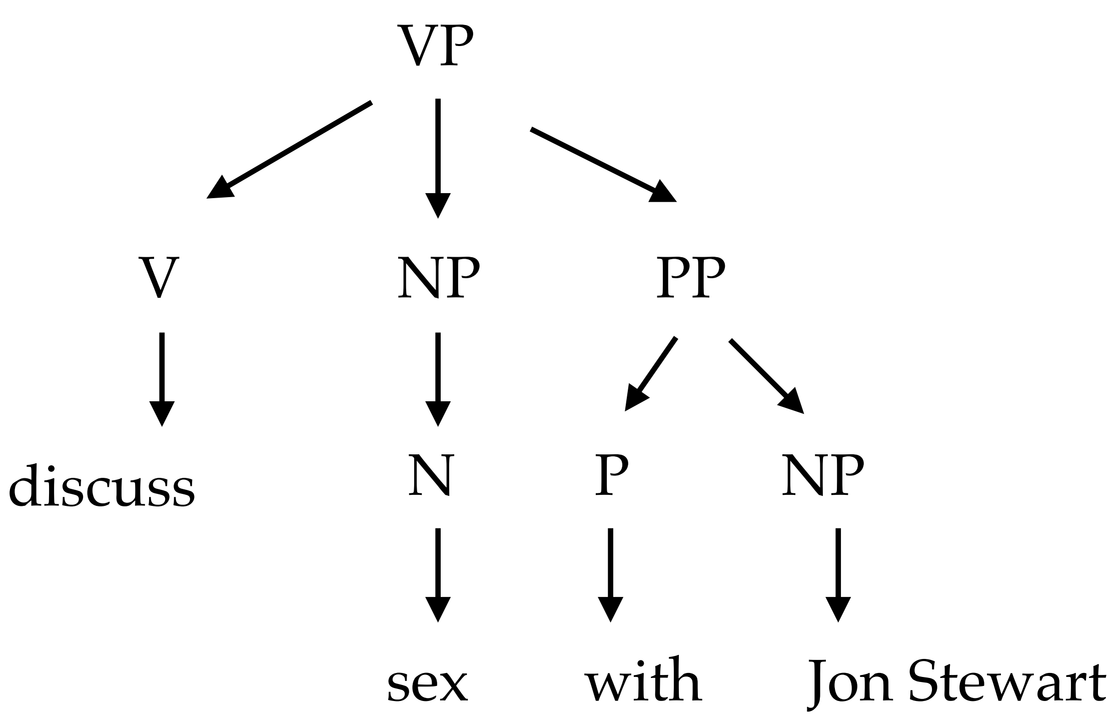
[1] from S Pinker, The Language Instinct

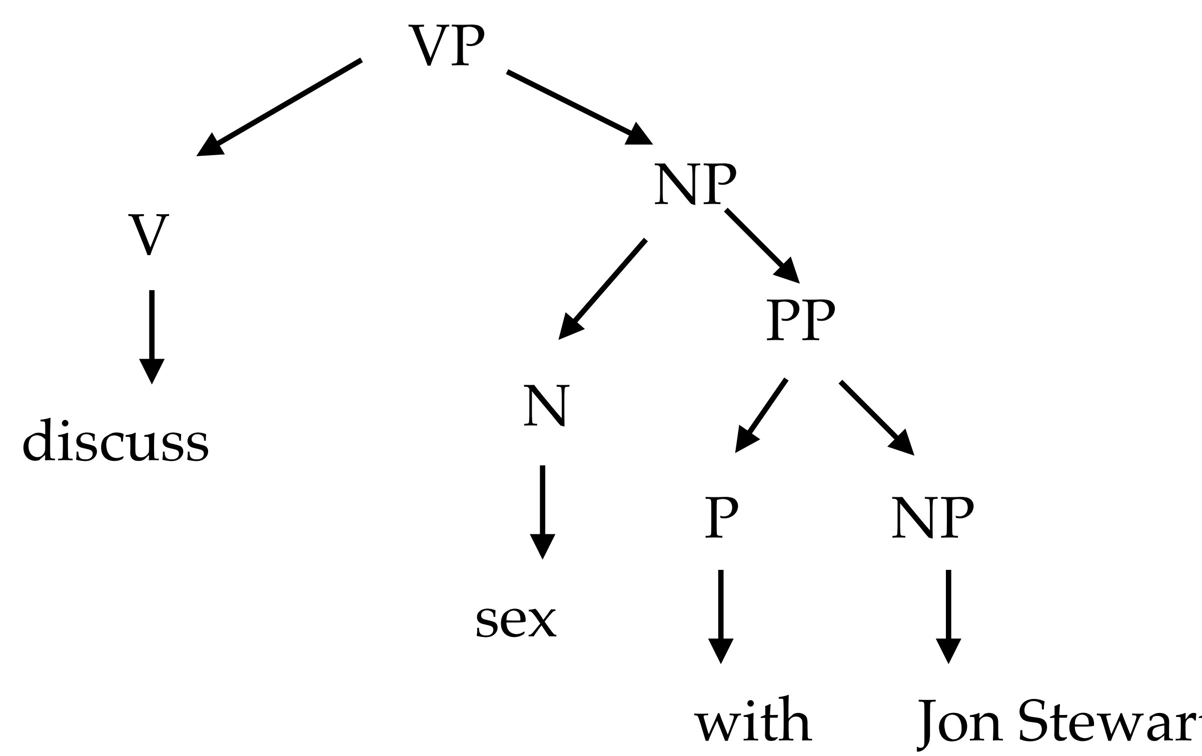
perspectives

ambiguity: For a typical sentence, how many grammatical parses are there?

If n = 1, sentence is unambiguous
If n > 1, sentence is ambiguous
If n = 0, sentence is ungrammatical

Natural languages are typically ambiguous, e.g.
“Two cars were reported stolen by the Groverton police yesterday” [1]





[1] from S Pinker, The Language Instinct

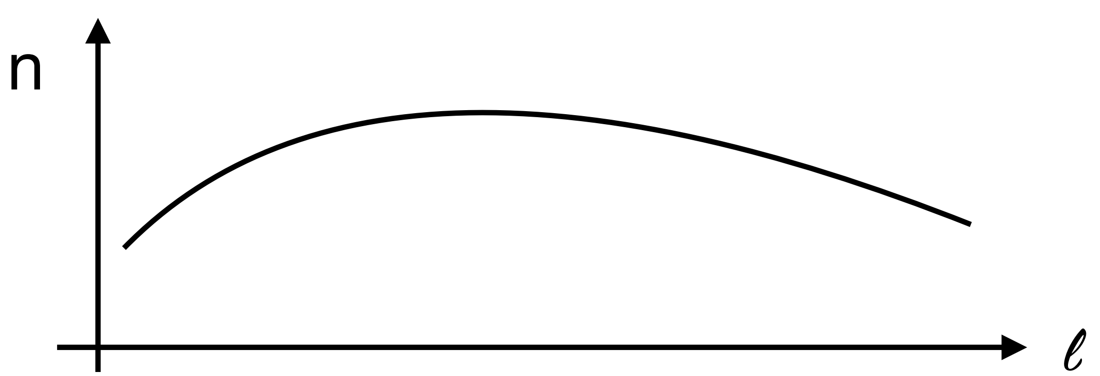
perspectives

ambiguity:

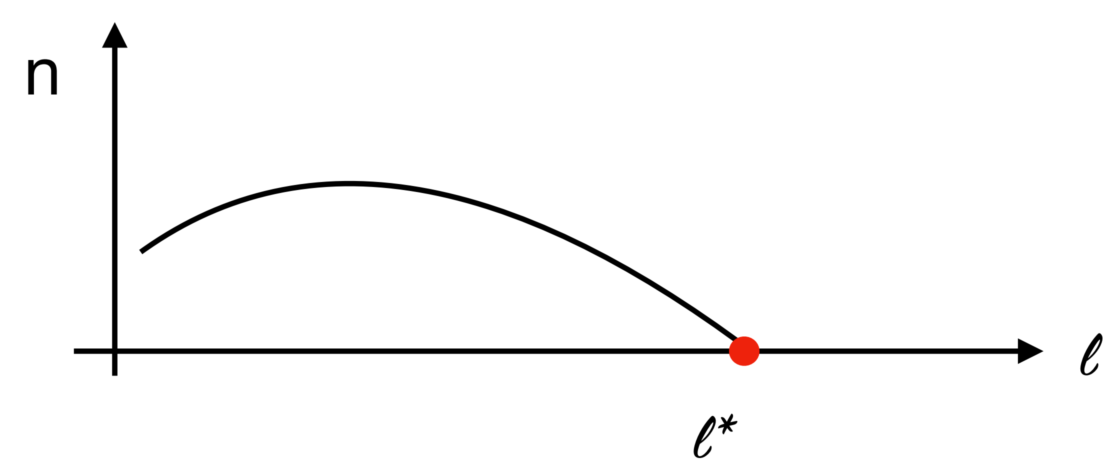
For a typical sentence, how many grammatical parses are there?

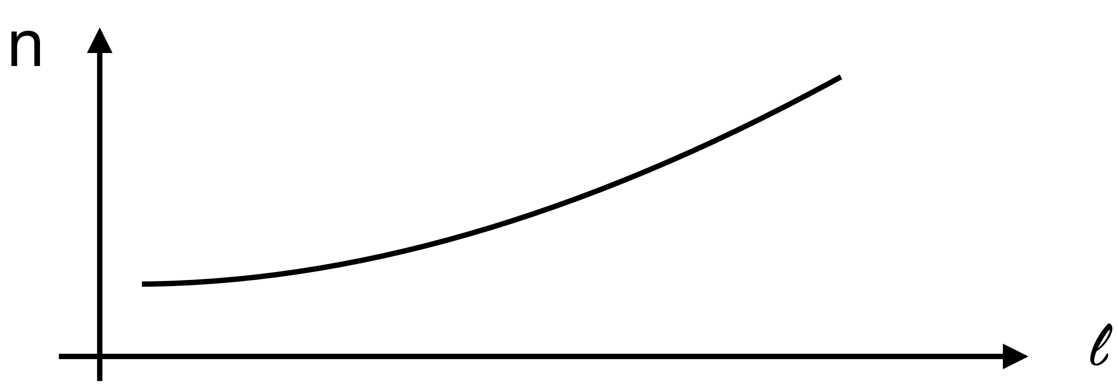
If n = 1, sentence is unambiguous
If n > 1, sentence is ambiguous
If n = 0, sentence is ungrammatical

Natural languages are typically ambiguous, e.g.
“Two cars were reported stolen by the Groverton police yesterday” [1]



which situation?





[1] from S Pinker, The Language Instinct

perspectives

phase diagram: What is the phase diagram of languages?

Are human languages atypical?

neural networks and learning:

What is the optimal architecture to learn highly compositional functions?

For natural language processing, how best to incorporate syntax into neural network approaches?

Can tools of disordered physics (e.g Thouless-Anderson-Palmer equations) help to learn languages?

perspectives

semantics: syntax isn't everything..

e.g. who is 'he' in this dialogue: [1]

Alice: I'm leaving you.
Bob: Who is he?!

Is there a physical approach to semantics?
c.f. dependency grammars, Montague grammars, …

[1] from S Pinker, The Language Instinct

conclusions

- successful machine learning architectures are *compositional* and *hierarchical*
- natural languages are also compositional and hierarchical
- context-free grammars define a simple model for these properties
- ensemble of grammars = random language model
- the statistical mechanical problem is not trivial, but not intractable

Mathematical linguistics has been around for 60 years.

It’s time for physical linguistics!

Thanks to my colleagues at ENS and elsewhere in Paris:

Remi Monasson, Jorge Kurchan, Francesco Zamponi,
Guilhem Semerjian, Pierfrancesco Urbani