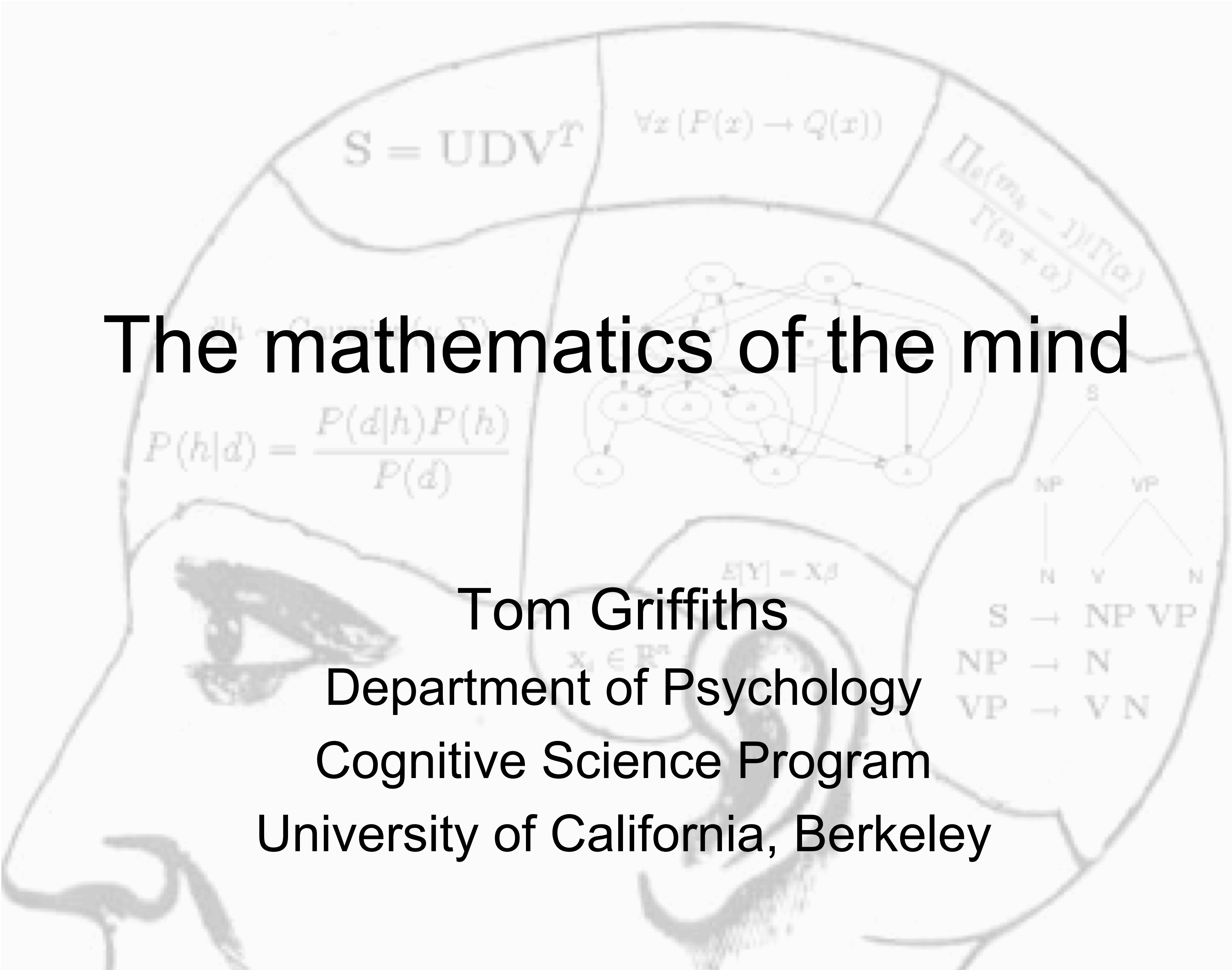
The mathematics of the mind

Tom Griffiths

Department of Psychology

Cognitive Science Program

University of California, Berkeley



Why apply math to the mind?

$$F = ma \qquad\qquad \frac{dx_i}{dt} = \sum_j q_{ij} f_j x_j - \phi x_i$$

Prediction and explanation

Mysteries of the mind

Artificial intelligence

Computational problems

- Easy:
  - arithmetic, algebra, chess
- Difficult:
  - learning and using language
  - sophisticated senses: vision, hearing
  - similarity and categorization
  - representing the structure of the world
  - scientific investigation

human cognition sets the standard

Three approaches

Rules and symbols

Networks, features, and spaces

Probability and statistics

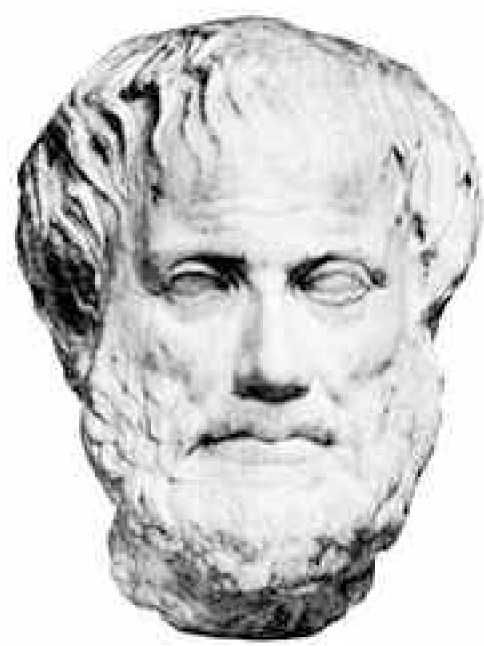
Three approaches

Rules and symbols

Networks, features, and spaces

Probability and statistics

Logic

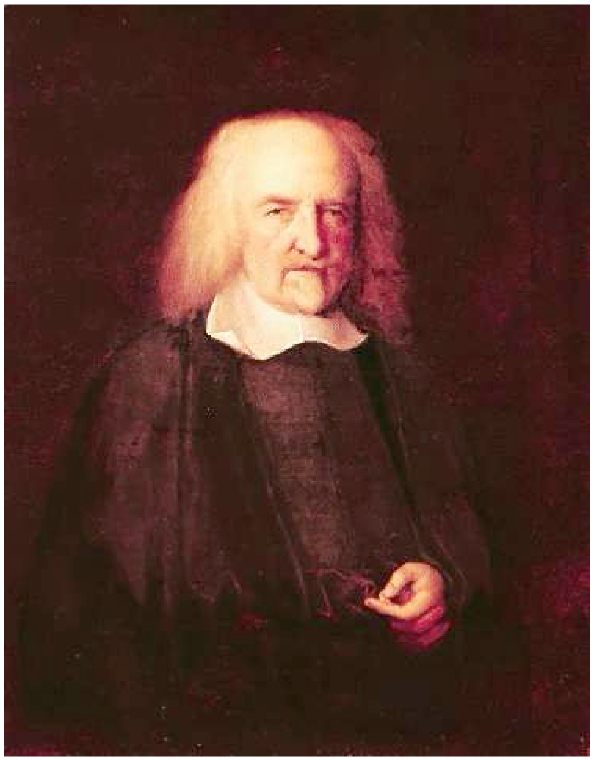


Aristotle
(384-322 BC)

All As are Bs
All Bs are Cs

All As are Cs

The mathematics of reason



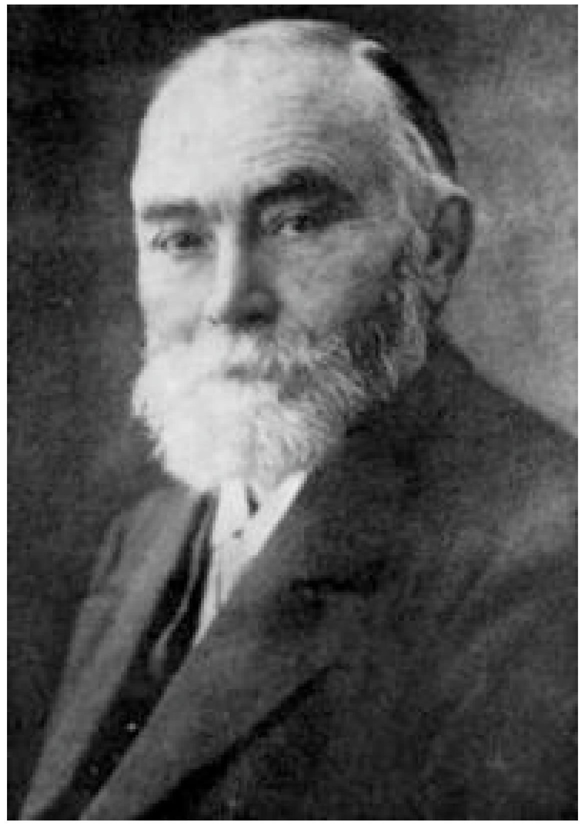
Thomas Hobbes
(1588-1679)

Rene Descartes
(1596-1650)

Gottfried Leibniz
(1646-1716)

Modern logic

George Boole
(1816-1854)



Gottlob Frege
(1848-1925)

$$\frac{\begin{array}{l} P \rightarrow Q \\ P \end{array}}{Q}$$

Syntax and semantics

Syntax

$$\frac{\begin{array}{l} P \rightarrow Q \\ P \end{array}}{Q}$$

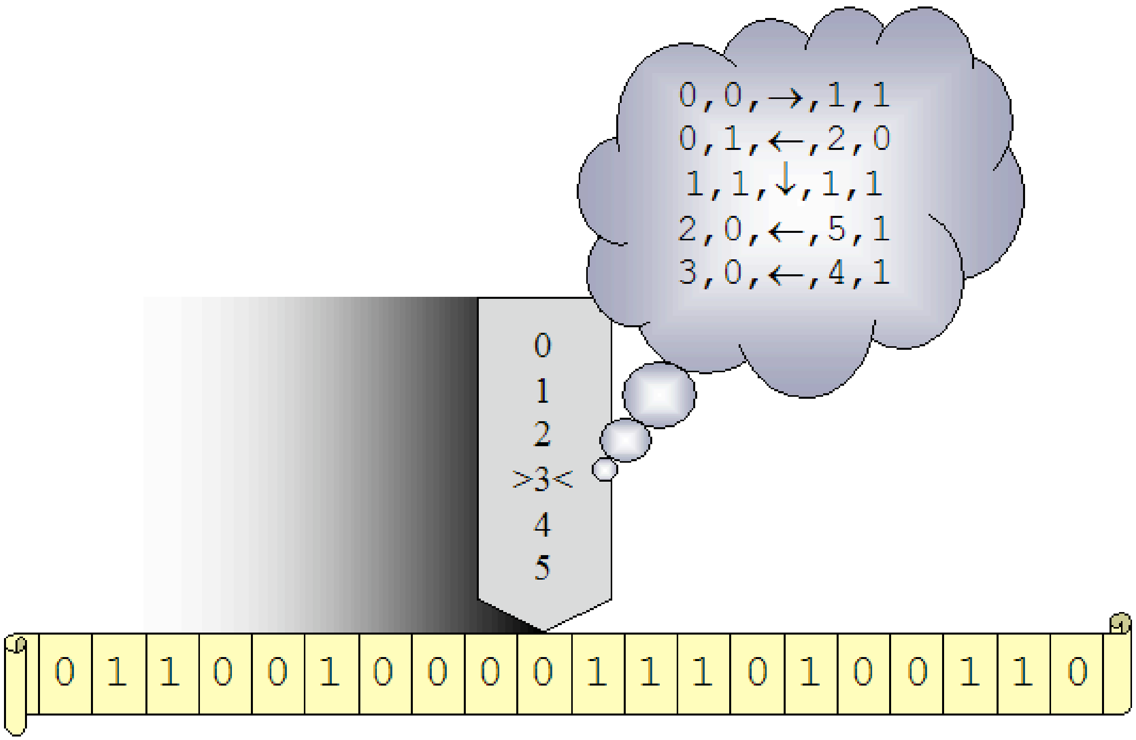
Semantics

P	Q	P→Q
T	T	T
T	F	F
F	T	T
F	F	T

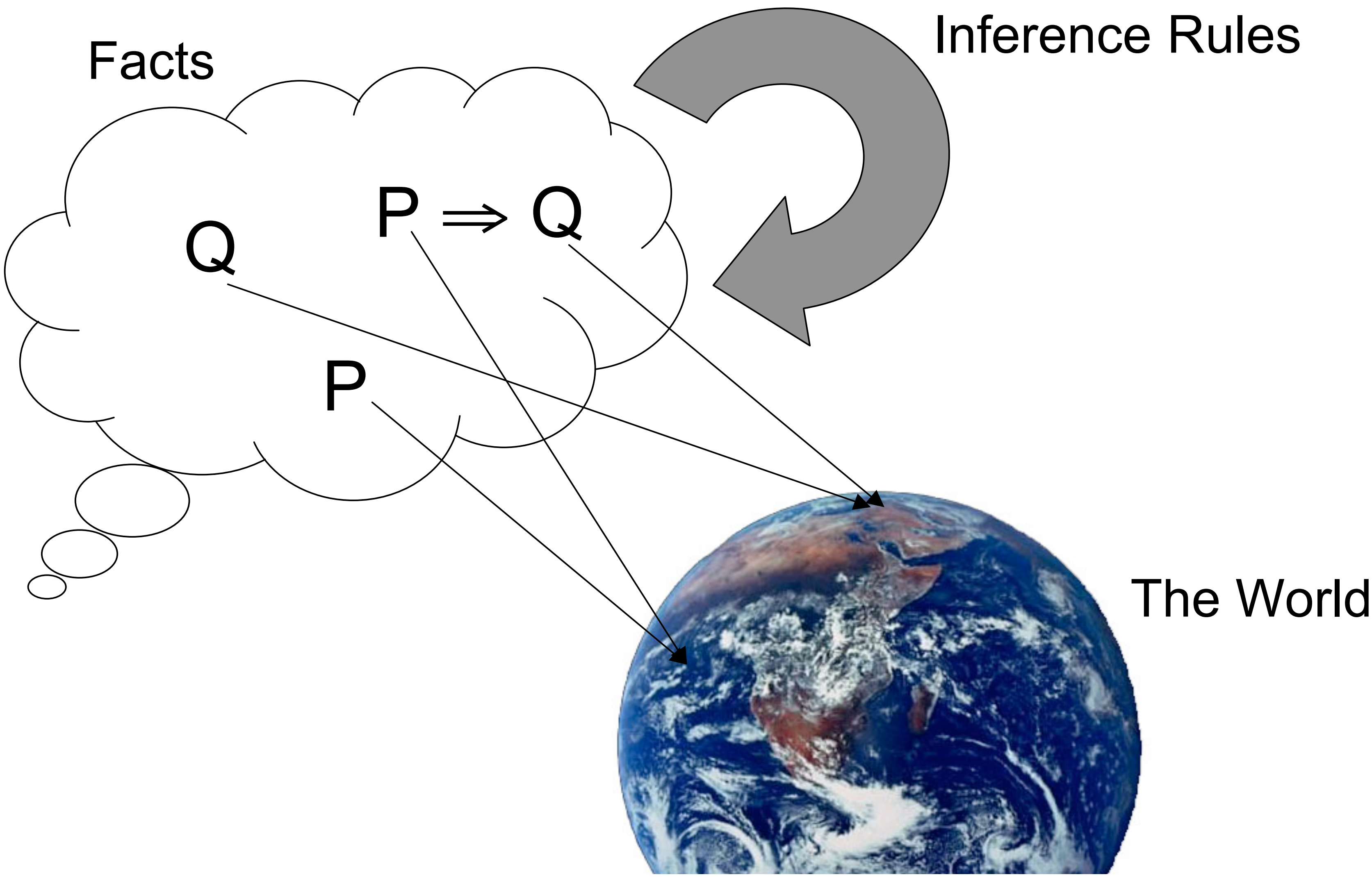
Can discover new truths through syntactic operations

Computation

Alan Turing
(1912-1954)



A logical view of the mind



Categorization











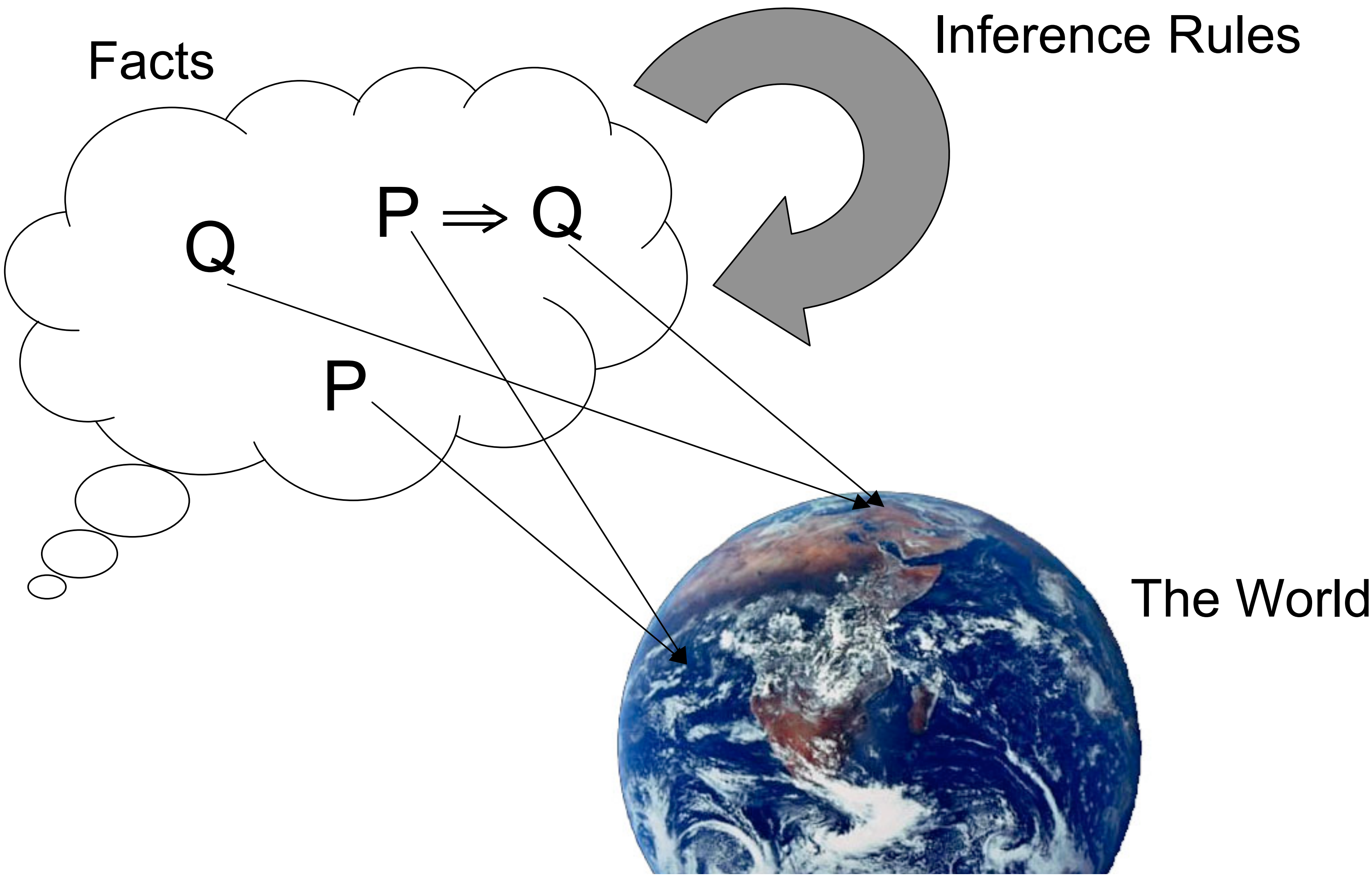



Categorization

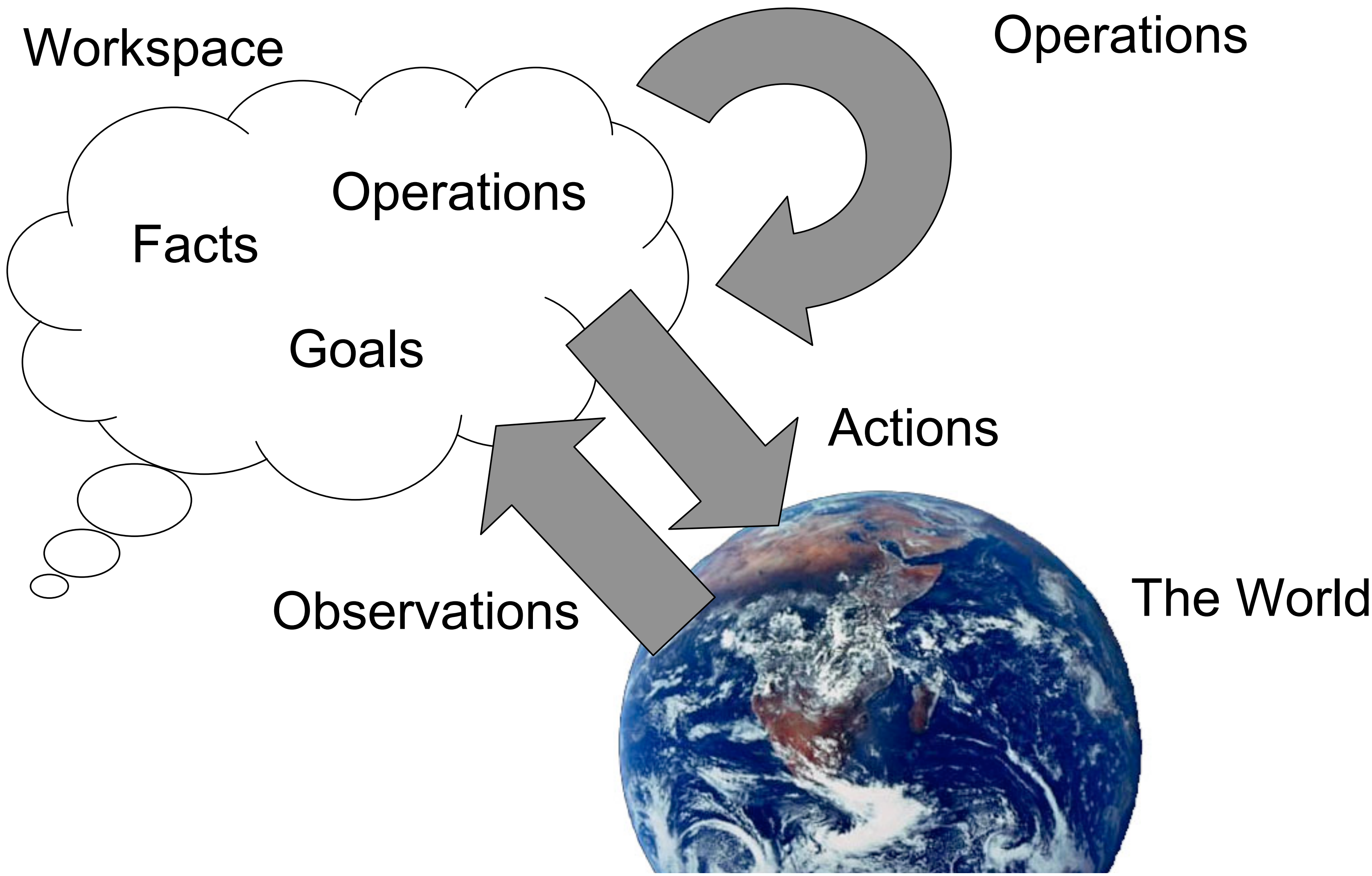
cat $\Leftrightarrow$ small $\wedge$ furry $\wedge$ domestic $\wedge$ carnivore



A logical view of the mind



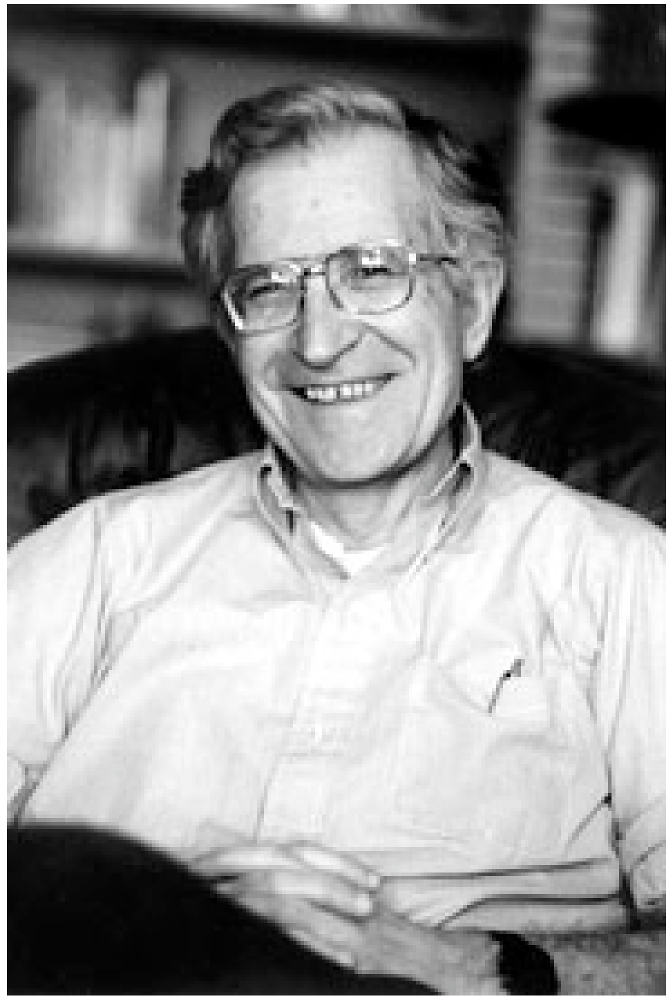
Early AI systems…



Rules and symbols

- Perhaps we can consider thought a set of rules, applied to symbols…
  - generating infinite possibilities with finite means
- This idea was applied to:
  - deductive reasoning (logic)
  - language (generative grammar)
  - problem solving and action (production systems)

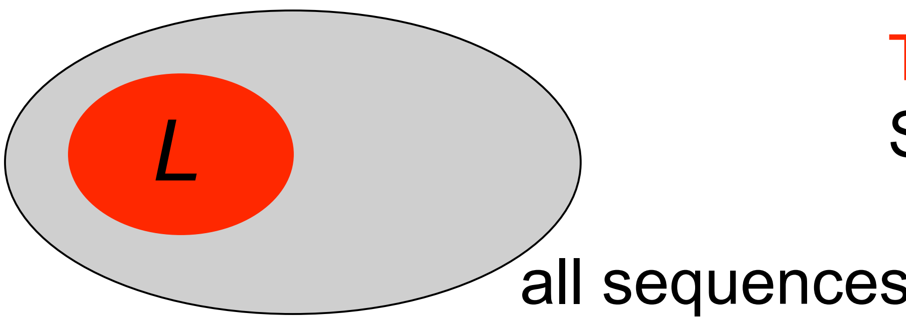
The rules of language



Noam Chomsky

Language

"a set (finite or infinite) of sentences, each finite in length and constructed out of a finite set of elements"

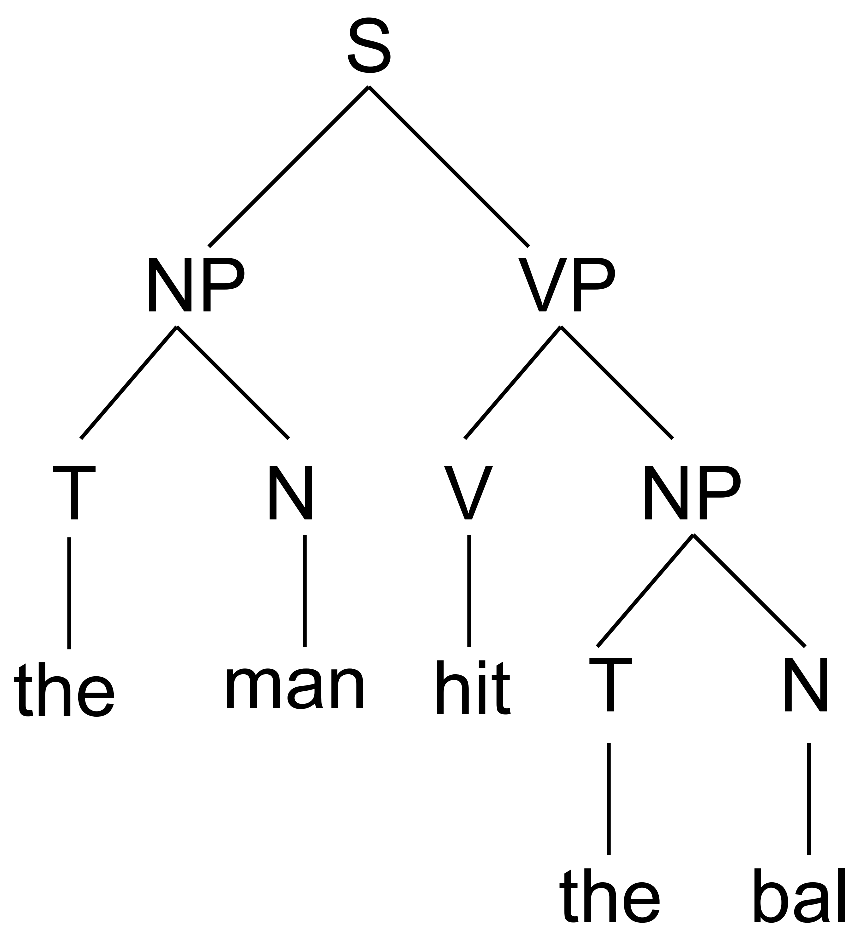


This is a good sentence	1
Sentence bad this is	0

linguistic analysis aims to separate the *grammatical* sequences which are sentences of *L* from the *ungrammatical* sequences which are not

A context free grammar

S	→ NP VP
NP	→ T N
VP	→ V NP
T	→ the
N	→ man, ball, …
V	→ hit, took, …



Rules and symbols

- Perhaps we can consider thought a set of rules, applied to symbols…
  – generating infinite possibilities with finite means
- This idea was applied to:
  – deductive reasoning (logic)
  – language (generative grammar)
  – problem solving and action (production systems)
- *Big question:* what are the rules of cognition?

Computational problems

- Easy:
  - arithmetic, algebra, chess
- Difficult:
  - learning and using language
  - sophisticated senses: vision, hearing
  - similarity and categorization
  - representing the structure of the world
  - scientific investigation

human cognition sets the standard

Inductive problems

- Drawing conclusions that are not fully justified by the available data
  - e.g. detective work

"In solving a problem of this sort, the grand thing is to be able to reason backward. That is a very useful accomplishment, and a very easy one, but people do not practice it much."

- Much more challenging than deduction!

Challenges for symbolic approaches

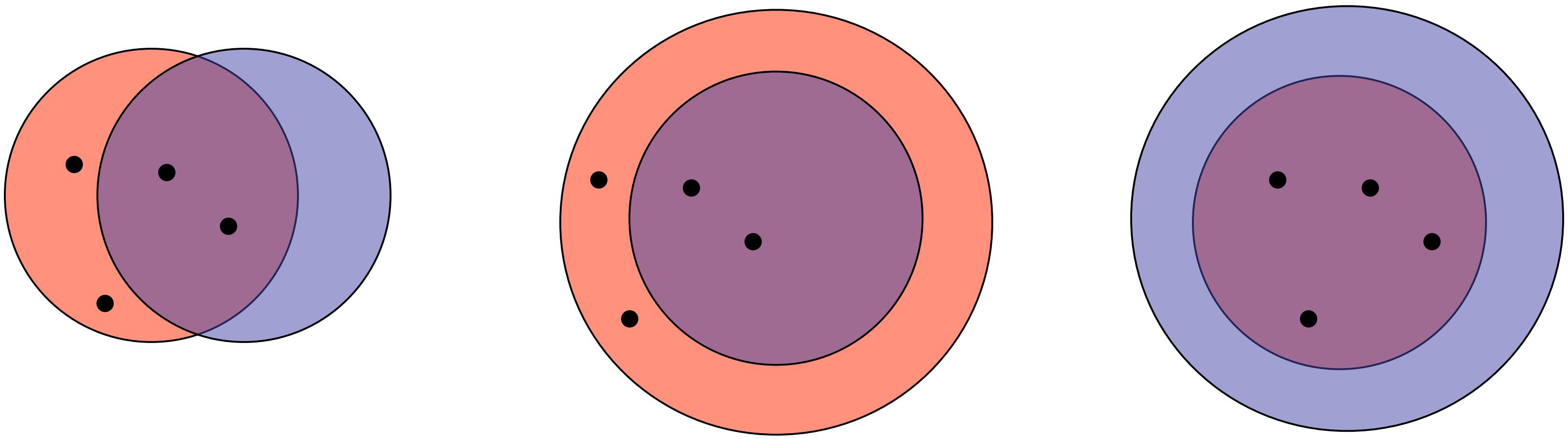
- Learning systems of rules and symbols is hard!
  - some people who think of human cognition in these terms end up arguing against learning…

The poverty of the stimulus

S → NP VP
NP → T N
VP → V NP
T → the
N → man, ball, …
V → hit, took, …

The logical problem

Red: Target language   Blue: Current hypothesis



If target language is a subset of the current hypothesis, no positive evidence can definitely rule it out

Challenges for symbolic approaches

- Learning systems of rules and symbols is hard!
  - some people who think of human cognition in these terms end up arguing against learning…
- Many human concepts have fuzzy boundaries
  - notions of similarity and typicality are hard to reconcile with binary rules


HUMANDESCENT.COM


HUMANDESCENT.COM

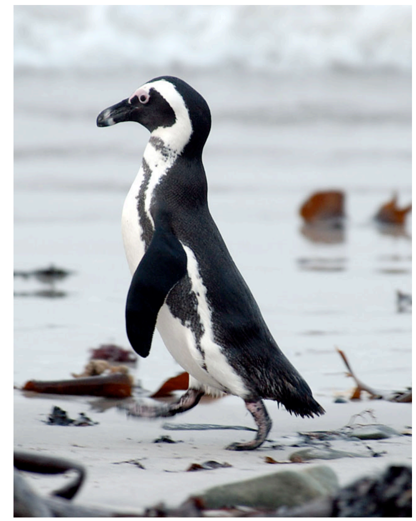
HUMANDESCENT.COM

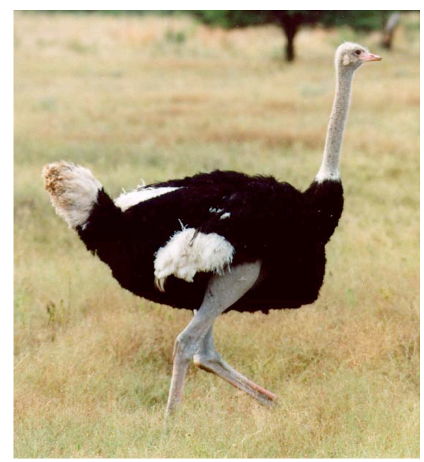
HUMANDESCENT.COM

HUMANDESCENT.COM

Typical

Atypical





Challenges for symbolic approaches

- Learning systems of rules and symbols is hard!
  - some people who think of human cognition in these terms end up arguing against learning…
- Many human concepts have fuzzy boundaries
  - notions of similarity and typicality are hard to reconcile with binary rules
- Solving inductive problems requires dealing with uncertainty and partial knowledge

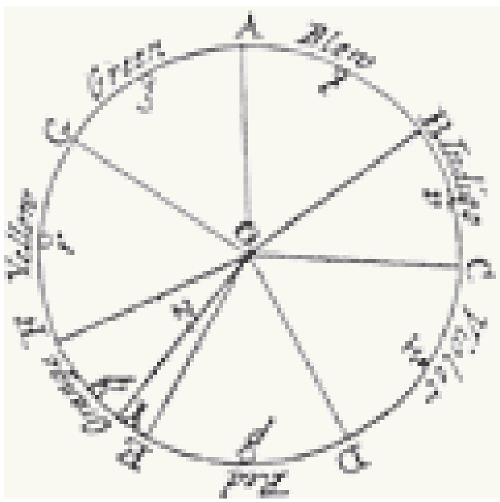
Three approaches

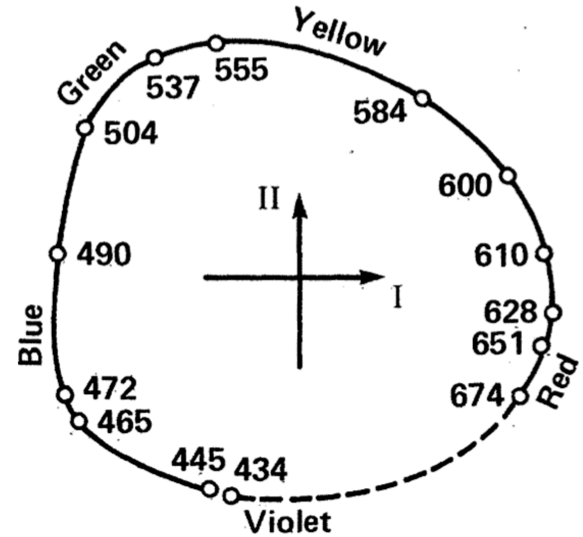
Rules and symbols

Networks, features, and spaces

Probability and statistics

Spatial representations





Categorization







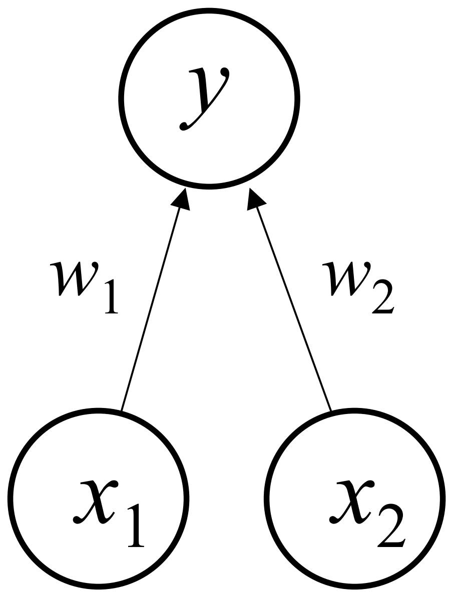







Perceptrons

+1 = cat, -1 = dog

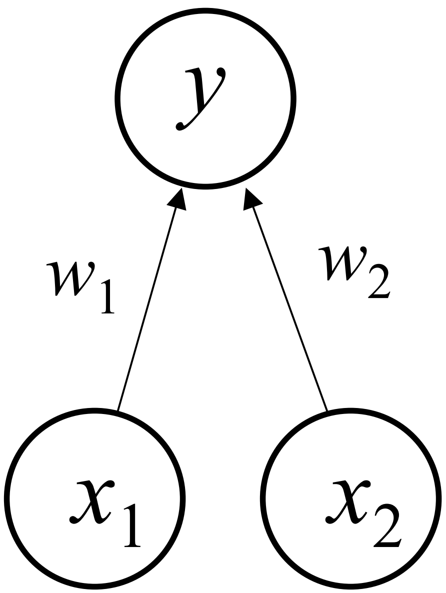


perceptual features

Frank Rosenblatt

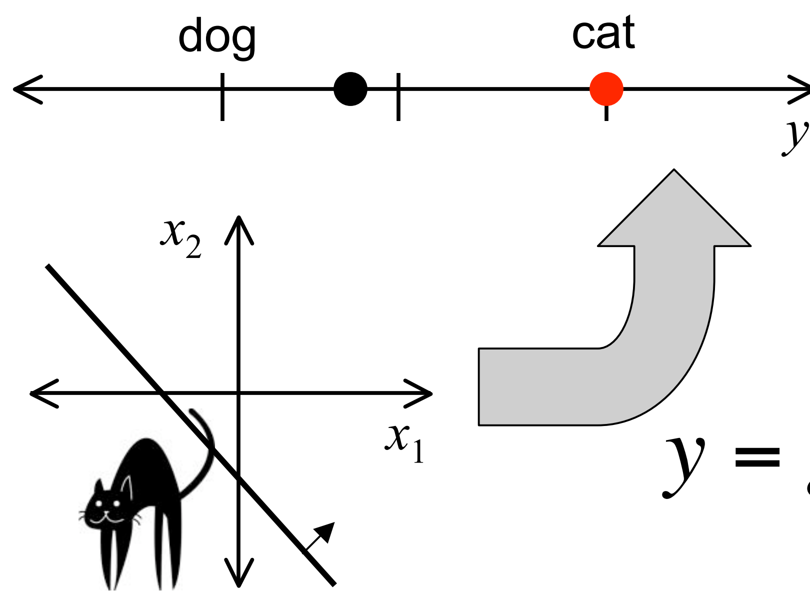
Computing with spaces

+1 = cat, -1 = dog



perceptual features

error: $E = \left(y - g(\mathbf{Wx})\right)^2$

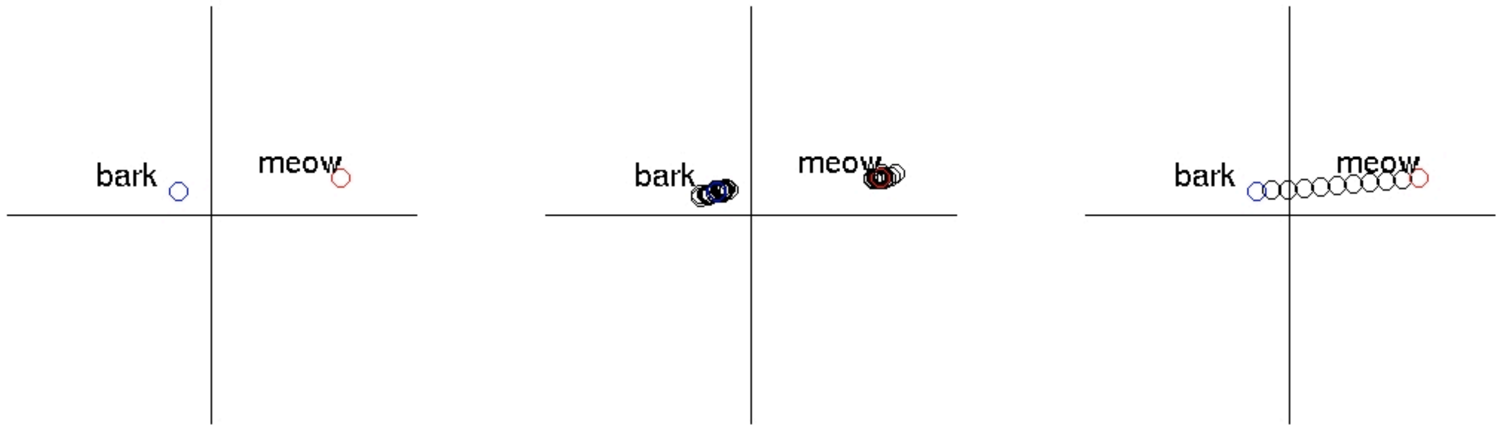


$y = g(\mathbf{Wx})$

Networks, features, and spaces

- Can capture the effects of typicality, similarity, uncertainty, and prior knowledge

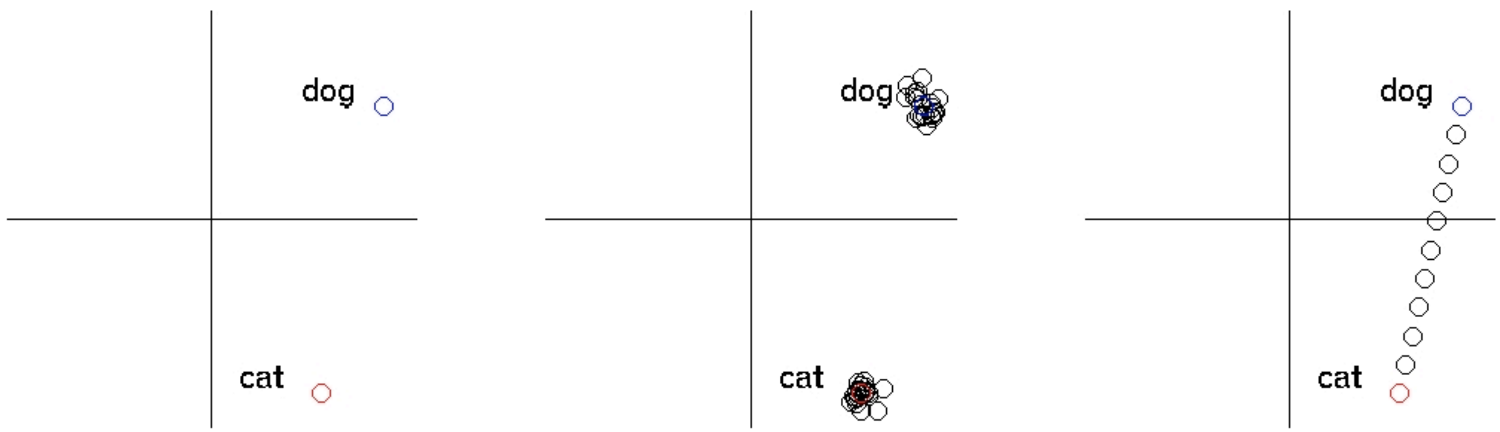
Computing with spaces



representation

noise
tolerance

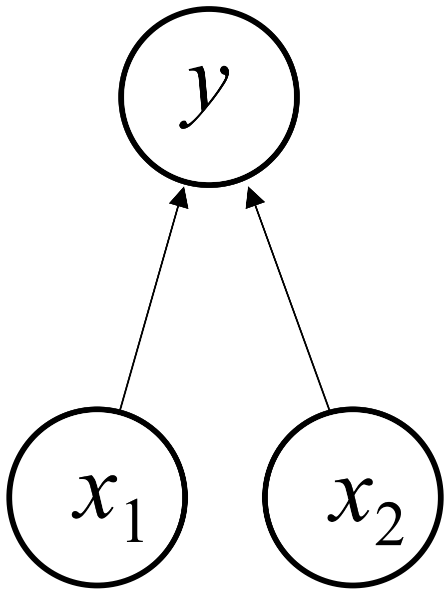
interpolation

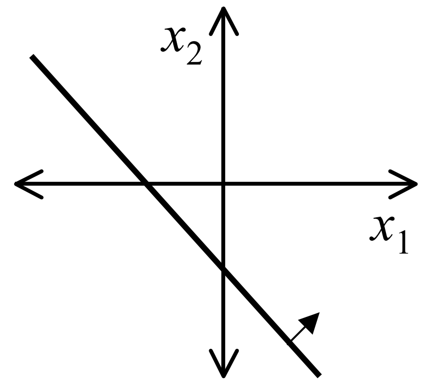


Networks, features, and spaces

- Can capture the effects of typicality, similarity, uncertainty, and prior knowledge
- Can represent any continuous function

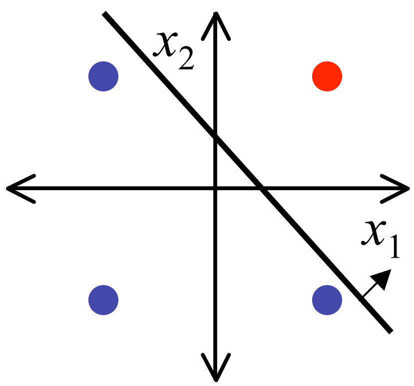
Problems with simple networks



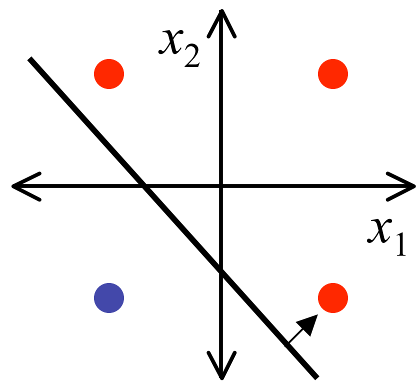


Some kinds of data are not linearly separable

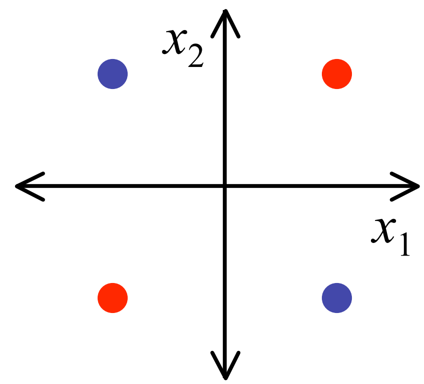
AND



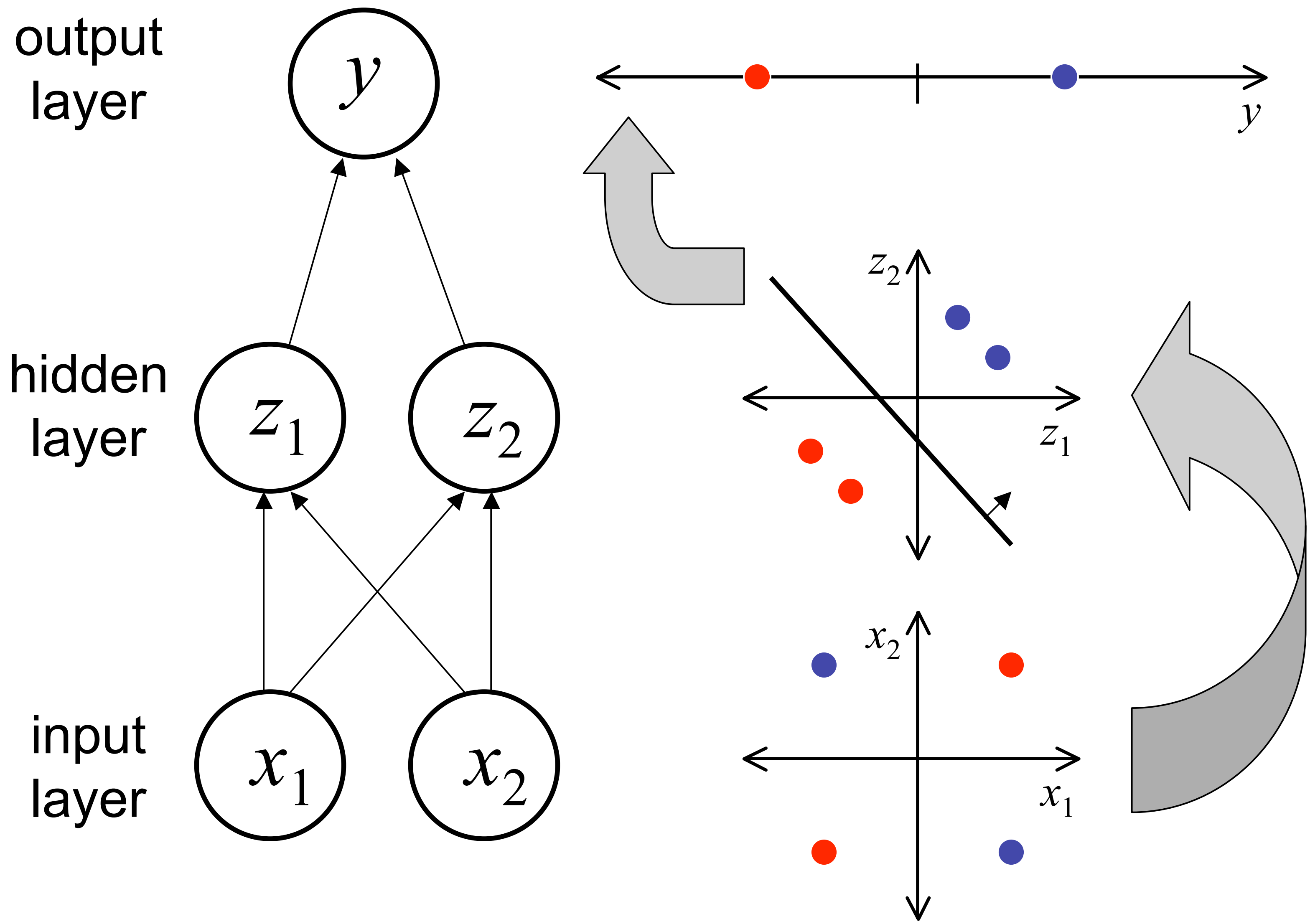
OR



XOR



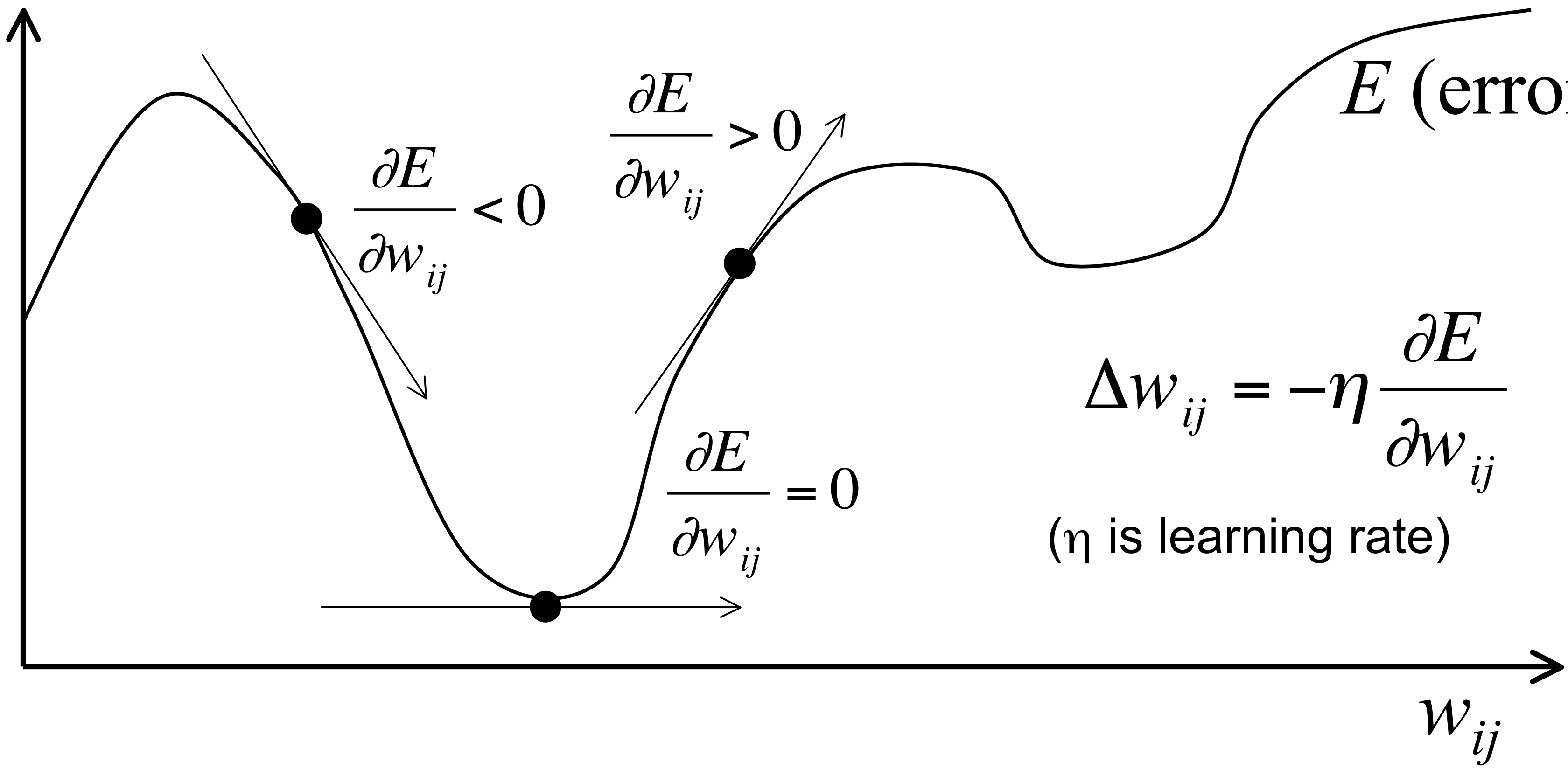
A solution: multiple layers



Networks, features, and spaces

- Can capture the effects of typicality, similarity, uncertainty, and prior knowledge
- Can represent any continuous function
- Simple algorithms for learning from data

General-purpose learning mechanisms



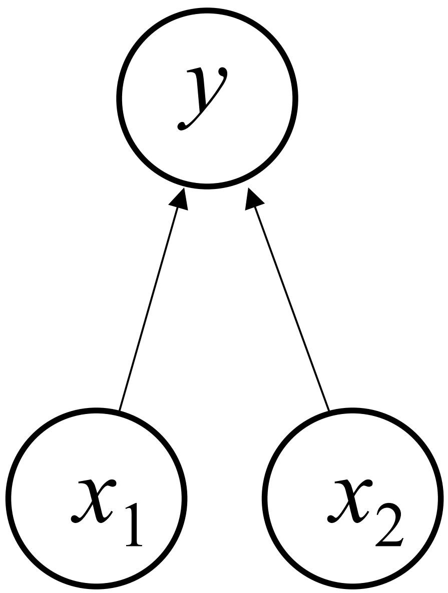
$$\Delta w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}}$$

(η is learning rate)

The Delta Rule

$$\Delta w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}}$$

+1 = cat, -1 = dog



perceptual features

$$E = \left(y - g(\mathbf{Wx})\right)^2$$

for any function g with derivative g'

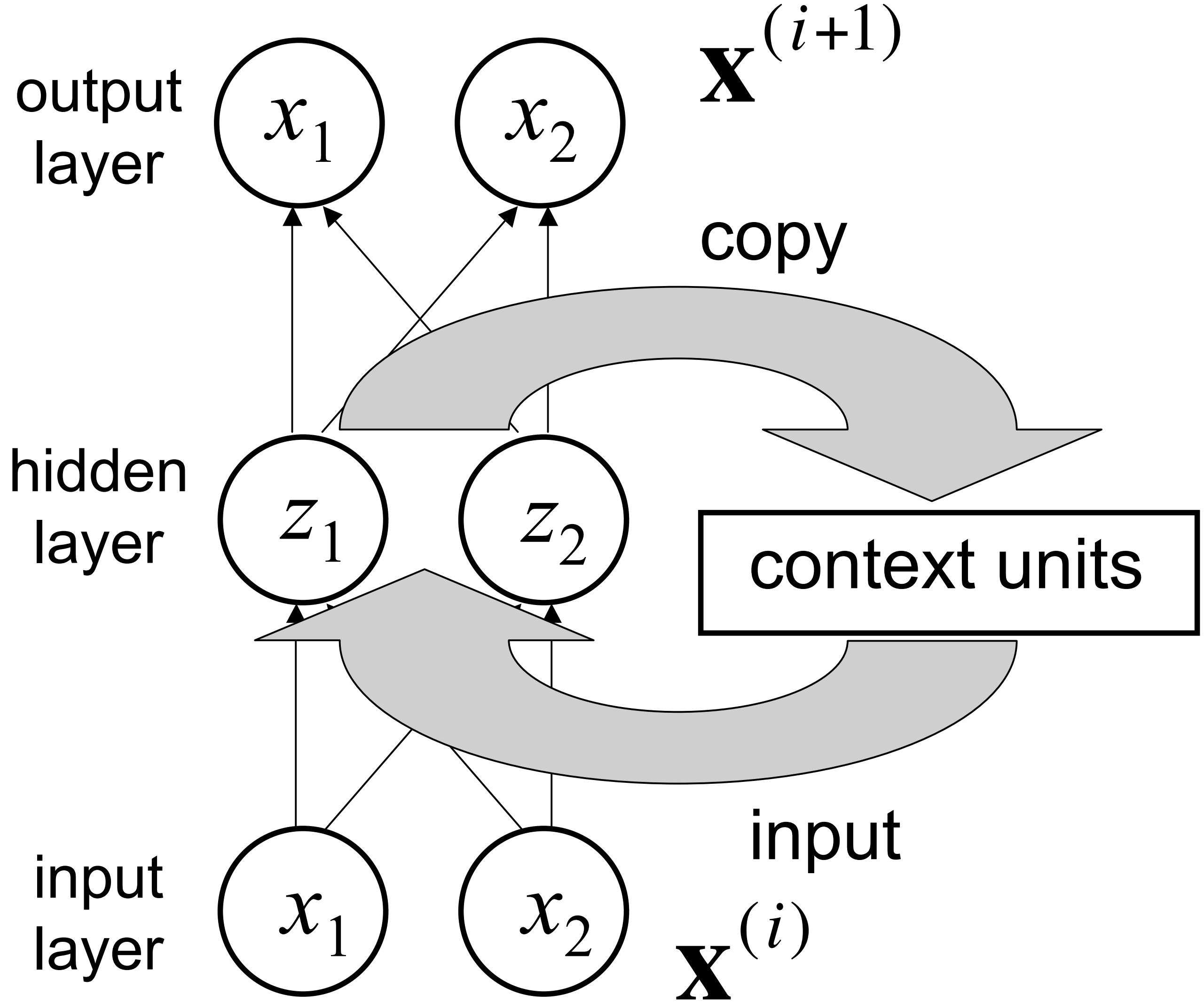
$$\frac{\partial E}{\partial w_{ij}} = -2\left(y - g(\mathbf{Wx})\right) g'(\mathbf{Wx})\, x_j$$

$$\Delta w_{ij} = \eta \underbrace{\boxed{\left(y - g(\mathbf{Wx})\right)}}_{\text{output error}} \underbrace{\boxed{g'(\mathbf{Wx})\, x_j}}_{\text{influence of input}}$$

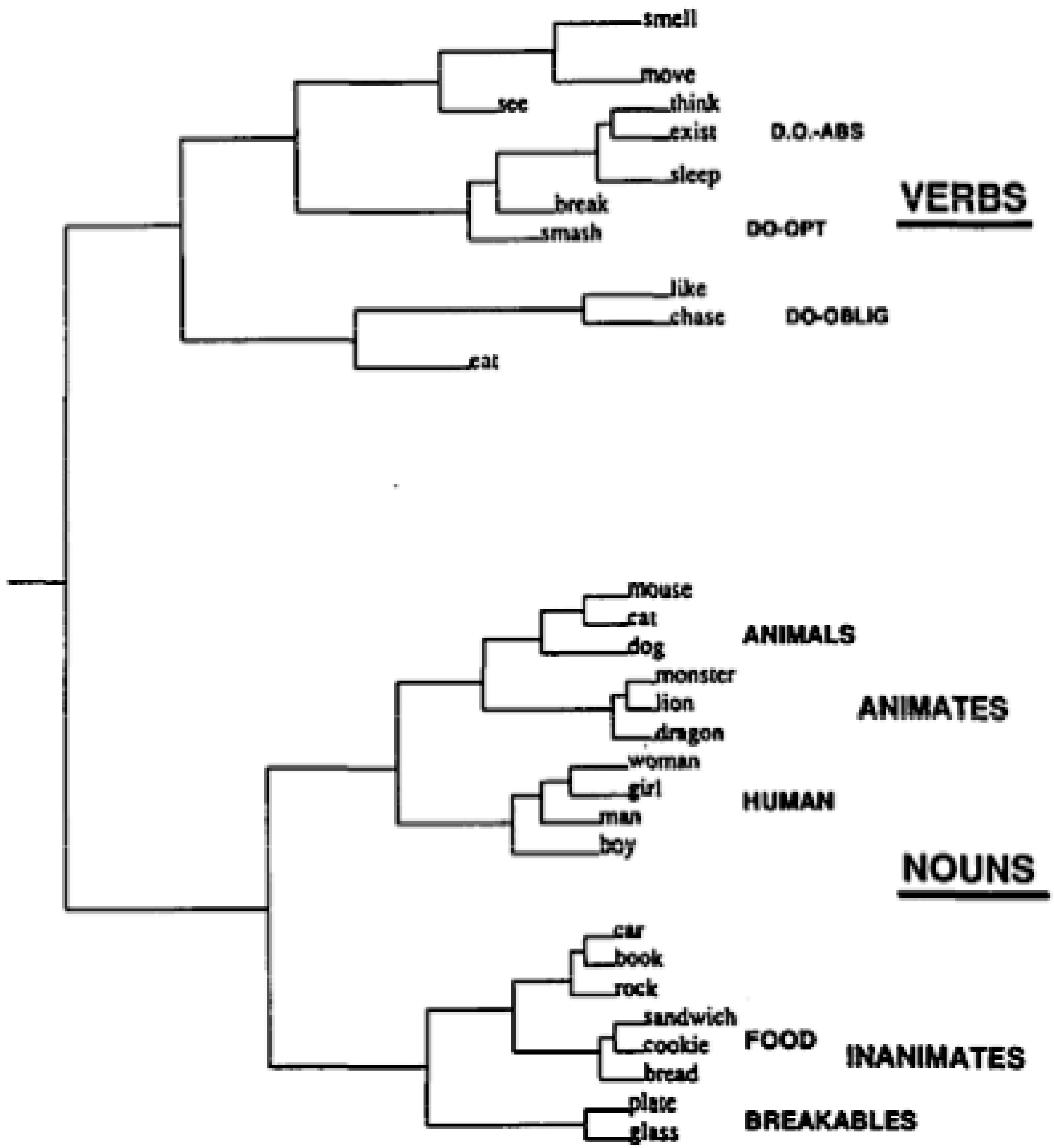
Networks, features, and spaces

- Can capture the effects of typicality, similarity, uncertainty, and prior knowledge
- Can represent any continuous function
- Simple algorithms for learning from data
- A way to explain how people could learn things that look like rules and symbols…

Simple recurrent networks



(Elman, 1990)



Hidden unit activations after 6 iterations of 27,500 words

(Elman, 1990)

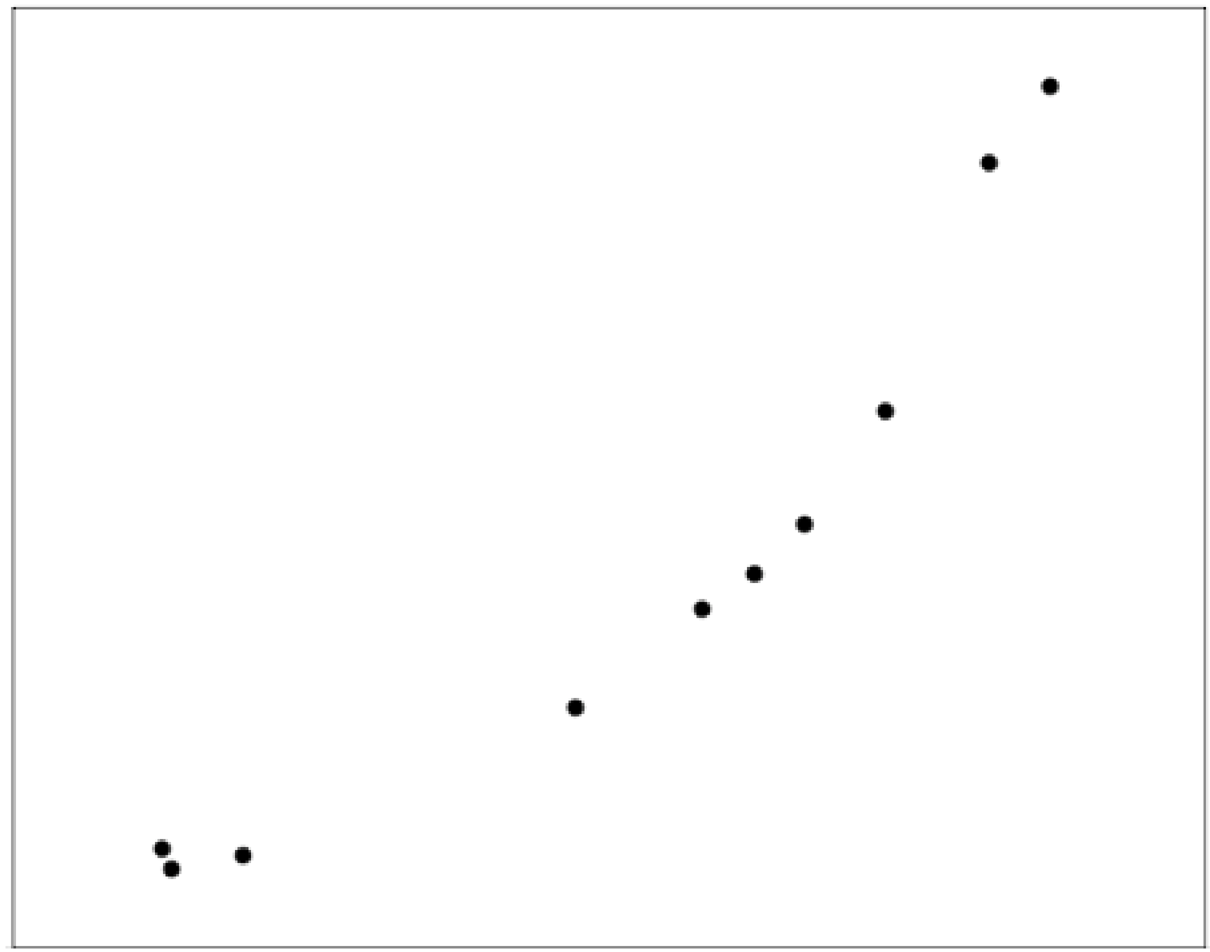
Networks, features, and spaces

- Can capture the effects of typicality, similarity, uncertainty, and prior knowledge
- Can represent any continuous function
- Simple algorithms for learning from data
- A way to explain how people could learn things that look like rules and symbols…
- *Big question:* how much of cognition can be explained by the input data?

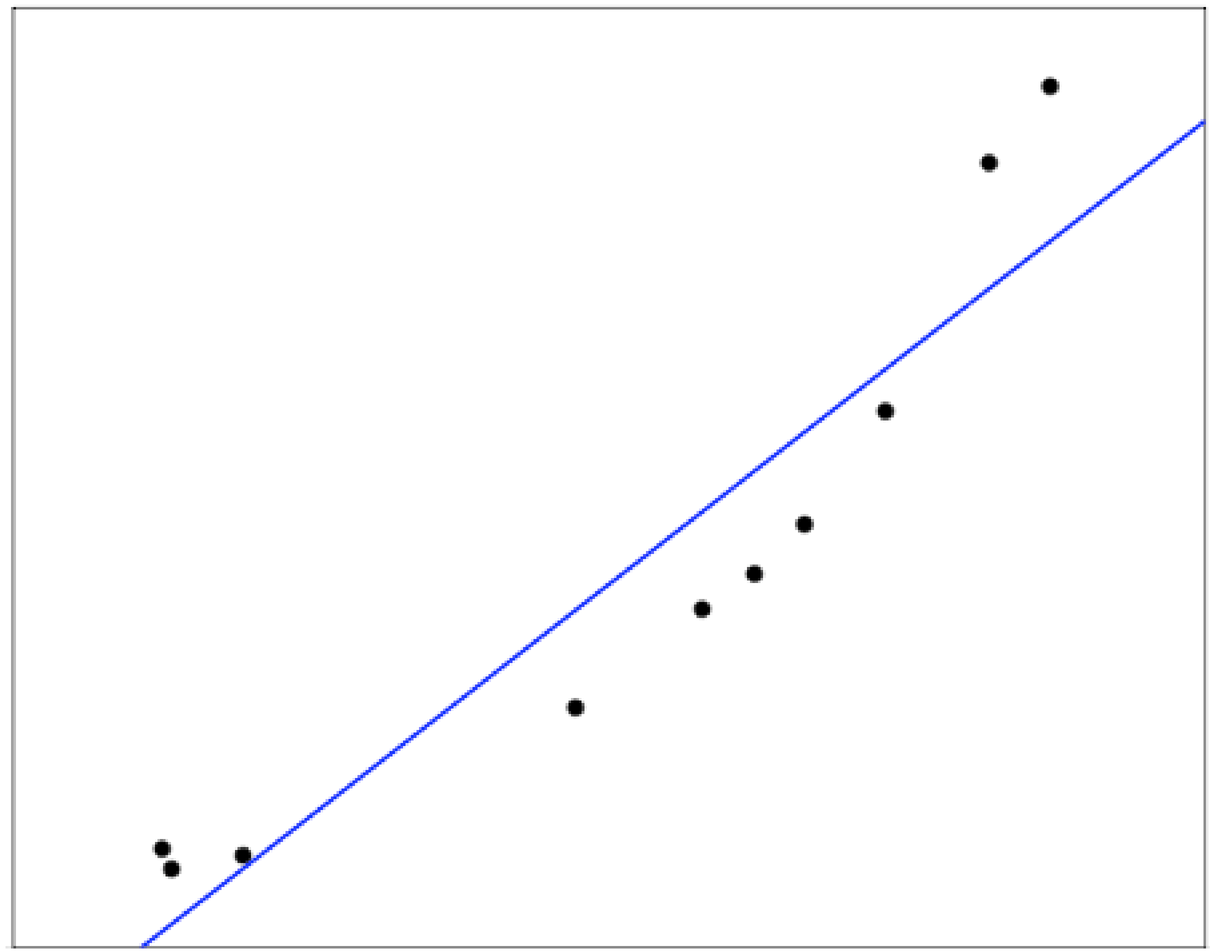
Challenges for neural networks

- Being able to learn anything can make it harder to learn specific things
  - this is the “bias-variance tradeoff”

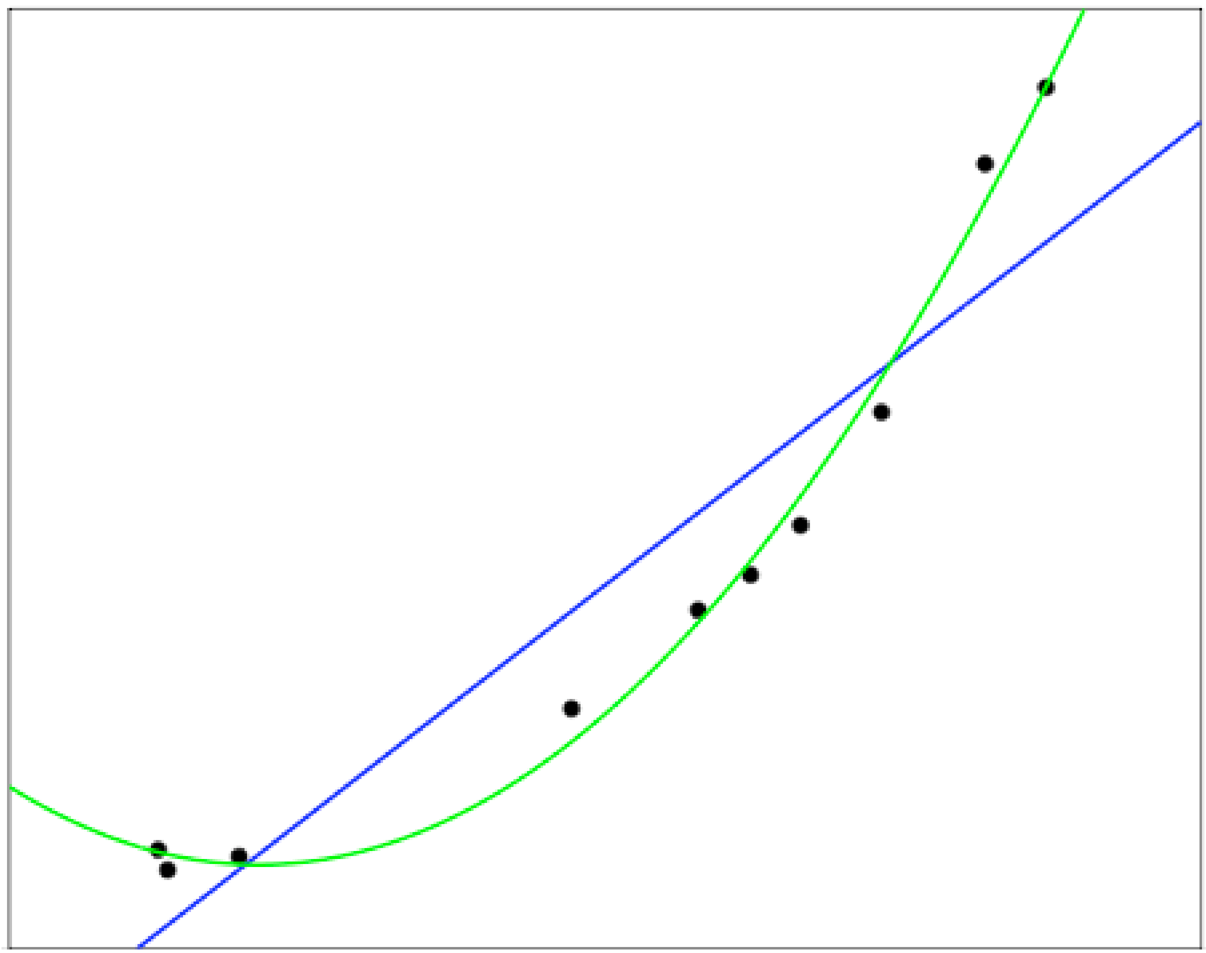
Bias-variance tradeoff



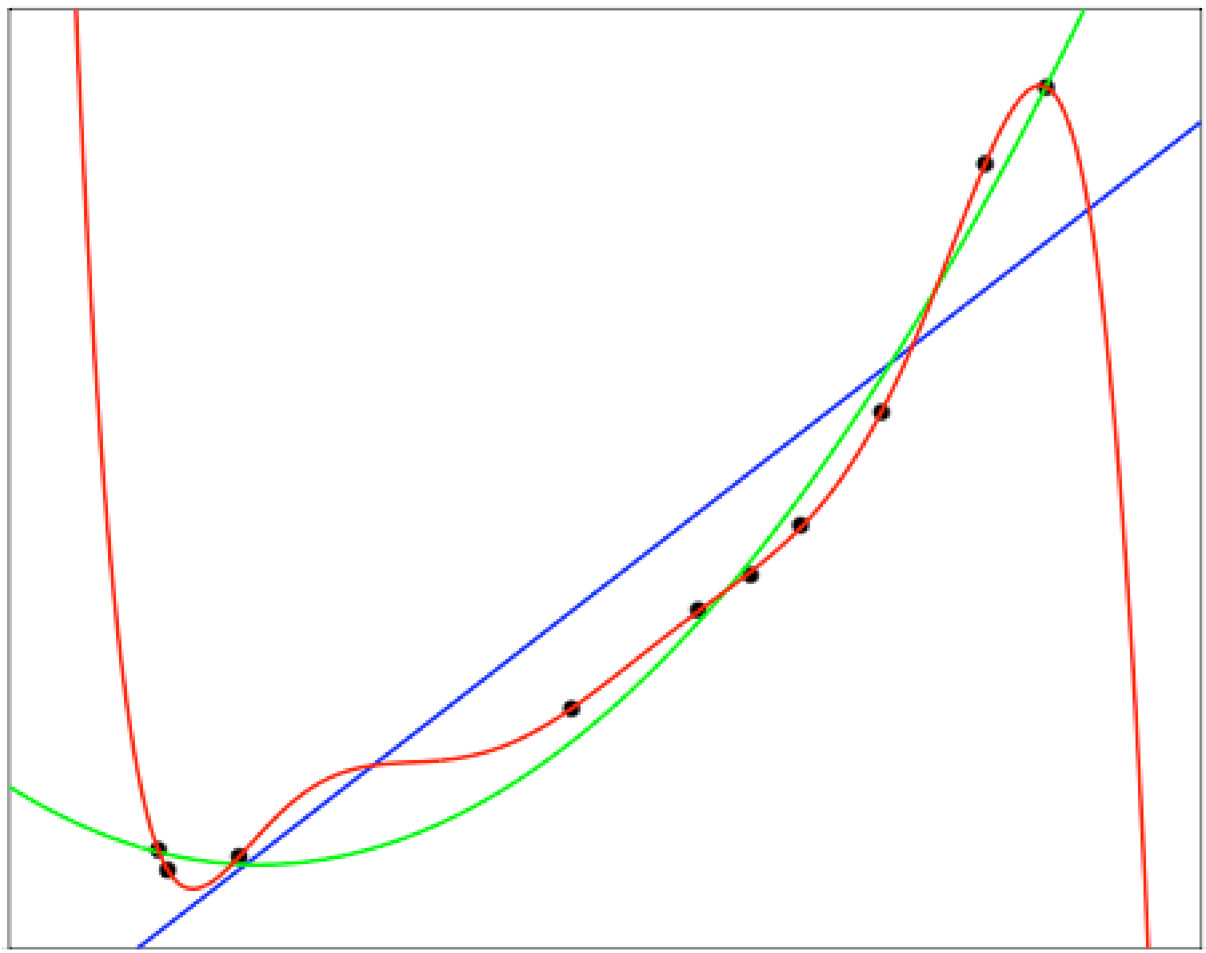
Bias-variance tradeoff



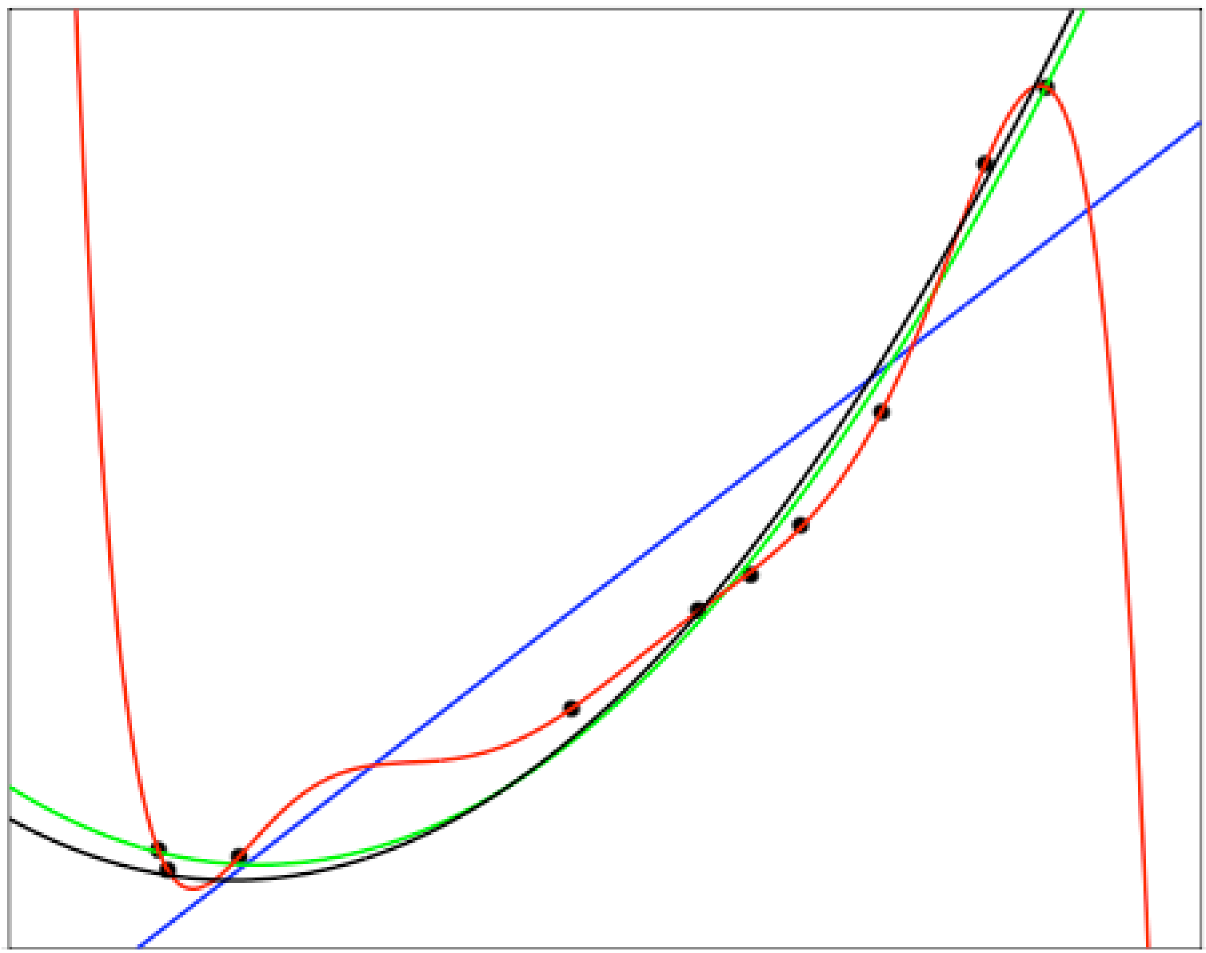
Bias-variance tradeoff



Bias-variance tradeoff



What about generalization?



What happened?

- The set of 8th degree polynomials contains almost all functions through 10 points
- Our data are some true function, plus noise
- Fitting the noise gives us the wrong function
- This is called *overfitting*
  - while it has low bias, this class of functions results in an algorithm that has high variance (i.e. is strongly affected by the observed data)

The moral

- General purpose learning mechanisms do not work well with small amounts of data
  (the most flexible algorithm isn't always the best)

- To make good predictions from small amounts of data, you need algorithms with bias that matches the problem being solved

Challenges for neural networks

- Being able to learn anything can make it harder to learn specific things
  - this is the "bias-variance tradeoff"
- Neural networks allow us to encode constraints on learning in terms of neurons, weights, and architecture, but is this always the right language?

Three approaches

Rules and symbols

Networks, features, and spaces

Probability and statistics

Probability

Gerolamo Cardano
(1501-1576)

Probability

Thomas Bayes
(1701-1763)

Pierre-Simon Laplace
(1749-1827)

Bayes' rule

How rational agents should update their beliefs in the light of data

Posterior probability

Likelihood

Prior probability

$$P(h \mid d) = \frac{P(d \mid h)P(h)}{\sum_{h' \in H} P(d \mid h')P(h')}$$

Sum over space of hypotheses

h: hypothesis
d: data

Bayes makes sense

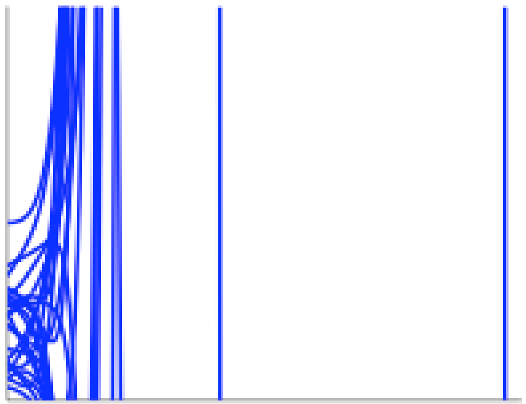
- Your friend coughs (the data d)
- Which of three hypotheses h is best?
  - a cold — medium prior, medium likelihood
  - lung cancer — low prior, high likelihood
  - a headache — high prior, low likelihood

Cognition as statistical inference

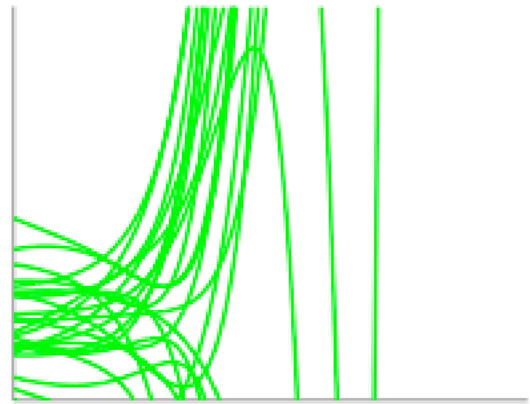
- Bayes' theorem tells us how to combine prior knowledge with data
  - a different language for describing the constraints on human inductive inference

Prior over functions

$k = 8,\ \alpha = 5,\ \beta = 1$

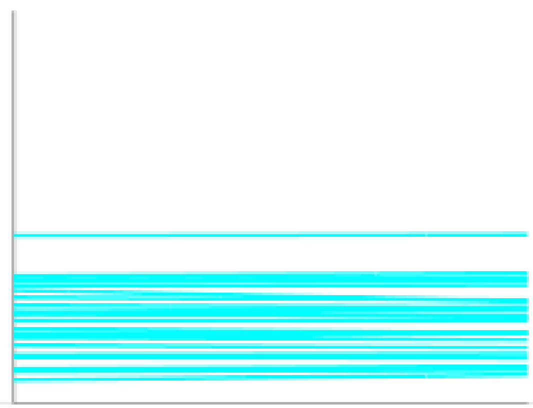


$k = 8,\ \alpha = 5,\ \beta = 0.3$

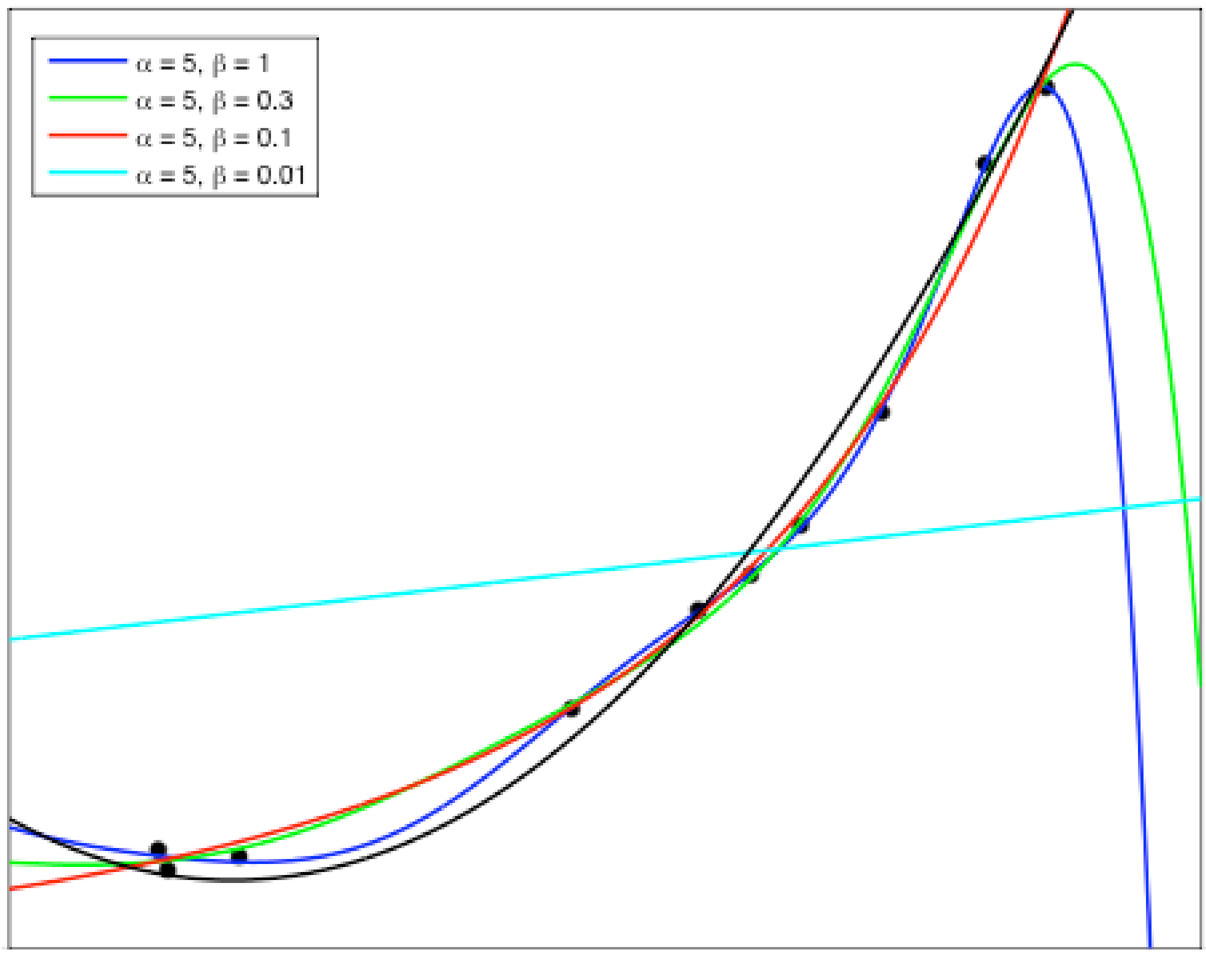


$k = 8, \alpha = 5,\ \beta = 0.1$

$k = 8,\ \alpha = 5,\ \beta = 0.01$



Maximum *a posteriori* (MAP) estimation

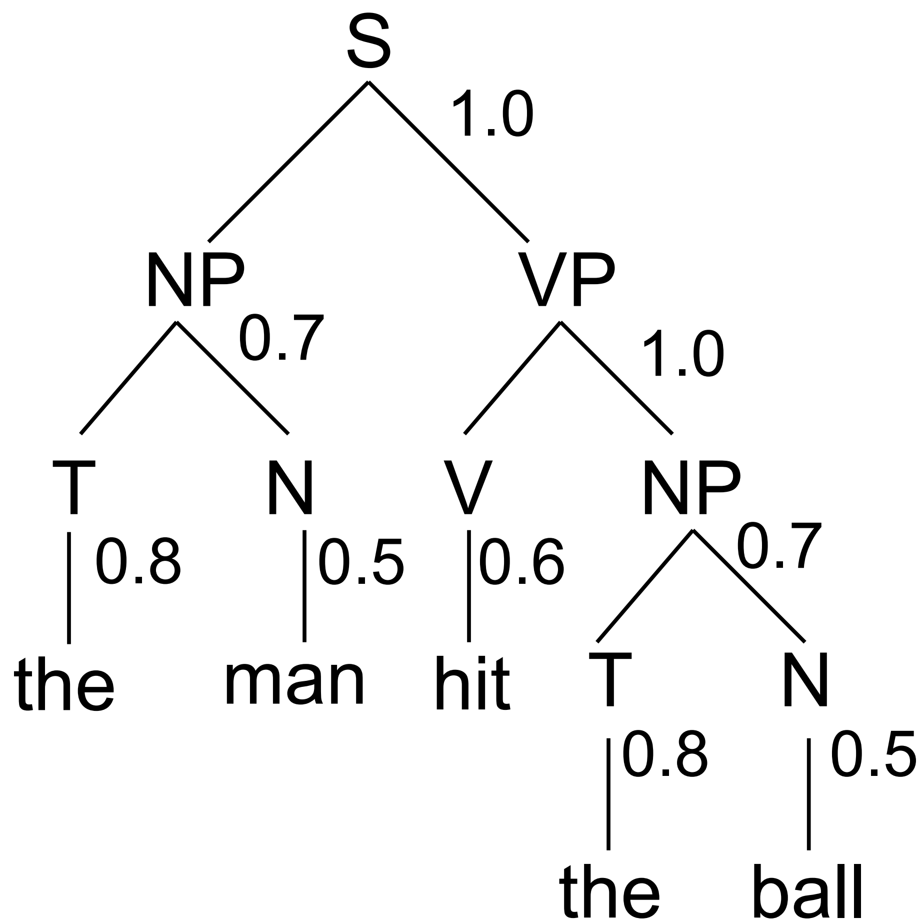


Cognition as statistical inference

- Bayes' theorem tells us how to combine prior knowledge with data
  - a different language for describing the constraints on human inductive inference
- Probabilistic approaches also help to describe learning

Probabilistic context free grammars

S	$\rightarrow$	NP VP	1.0
NP	$\rightarrow$	T N	0.7
NP	$\rightarrow$	N	0.3
VP	$\rightarrow$	V NP	1.0
T	$\rightarrow$	the	0.8
T	$\rightarrow$	a	0.2
N	$\rightarrow$	man	0.5
N	$\rightarrow$	ball	0.5
V	$\rightarrow$	hit	0.6
V	$\rightarrow$	took	0.4



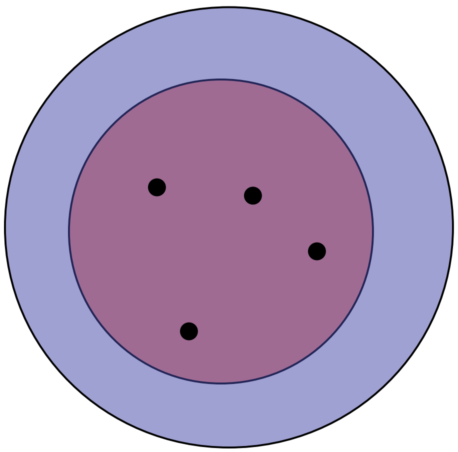
$$P(\text{tree}) = 1.0\times0.7\times1.0\times0.8\times0.5\times0.6\times0.7\times0.8\times0.5$$

Probability and learnability

- Any probabilistic context free grammar can be learned from a sample from that grammar as the sample size becomes infinite

Bayesian inference

Red: h_1 Blue: h_2



Assume sentences are sampled uniformly from each set

$$P(d \mid h) = \begin{cases} 1/|h| & d \in h \\ 0 & \text{otherwise} \end{cases}$$

$|h_2| > |h_1|$, so $P(d|h_1) > P(d|h_2)$ for d from h_1

So… the posterior probability of h_1 increases with each sentence consistent with h_1 (even though these sentences are consistent with h_2 as well)

Probability and learnability

- Any probabilistic context free grammar can be learned from a sample from that grammar as the sample size becomes infinite
- Prior probability trades off with how much data needs to be seen to believe a hypothesis

Cognition as statistical inference

- Bayes’ theorem tells us how to combine prior knowledge with data
  - a language for describing the constraints on human inductive inference
- Probabilistic approaches also help to describe learning
- *Big question:* what do the constraints on human inductive inference look like?

Challenges for probabilistic approaches

- Computing probabilities is hard… how could brains possibly do that?

- How well do the “rational” solutions from probability theory describe how people think in everyday life?

Three approaches

Rules and symbols

Networks, features, and spaces

Probability and statistics