Autonomous and Adaptive Systems

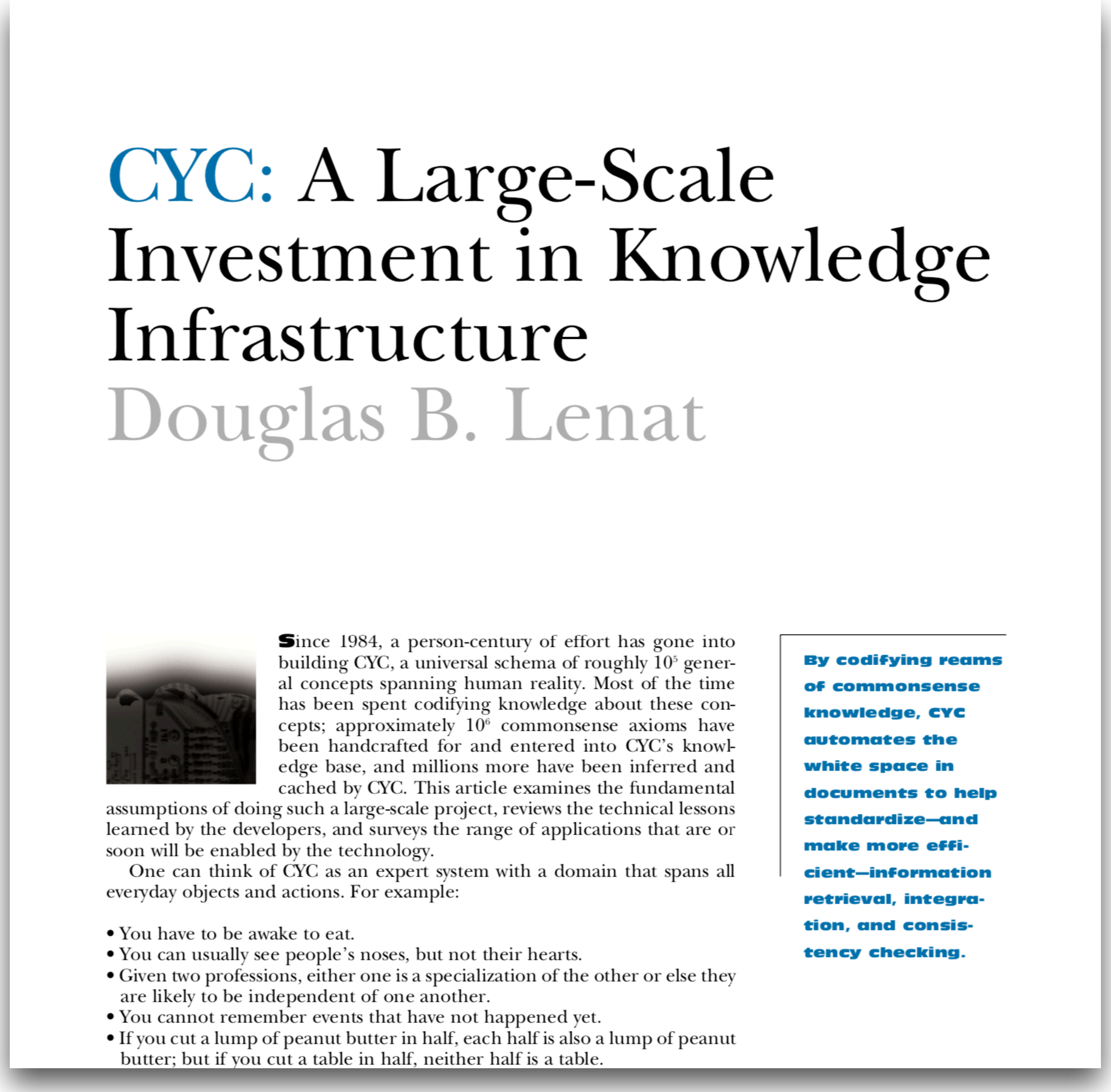
Introduction to Deep Learning I

Mirco Musolesi

mircomusolesi@acm.org

Overview and Historical Notes

- Initial AI systems are based on the definition of formal systems (logic, knowledge base, etc.).
  - Several artificial intelligence projects have sought to hard-code knowledge about the world in formal language.
  - Difficult to list rules for very large number of situations.
  - Some rules might also not be possible to be codified given the sheer complexity of the world.
- However note: recent developments in combining deep learning and symbolic AI.
  - Very open field of research at the moment.



CYC: A Large-Scale Investment in Knowledge Infrastructure

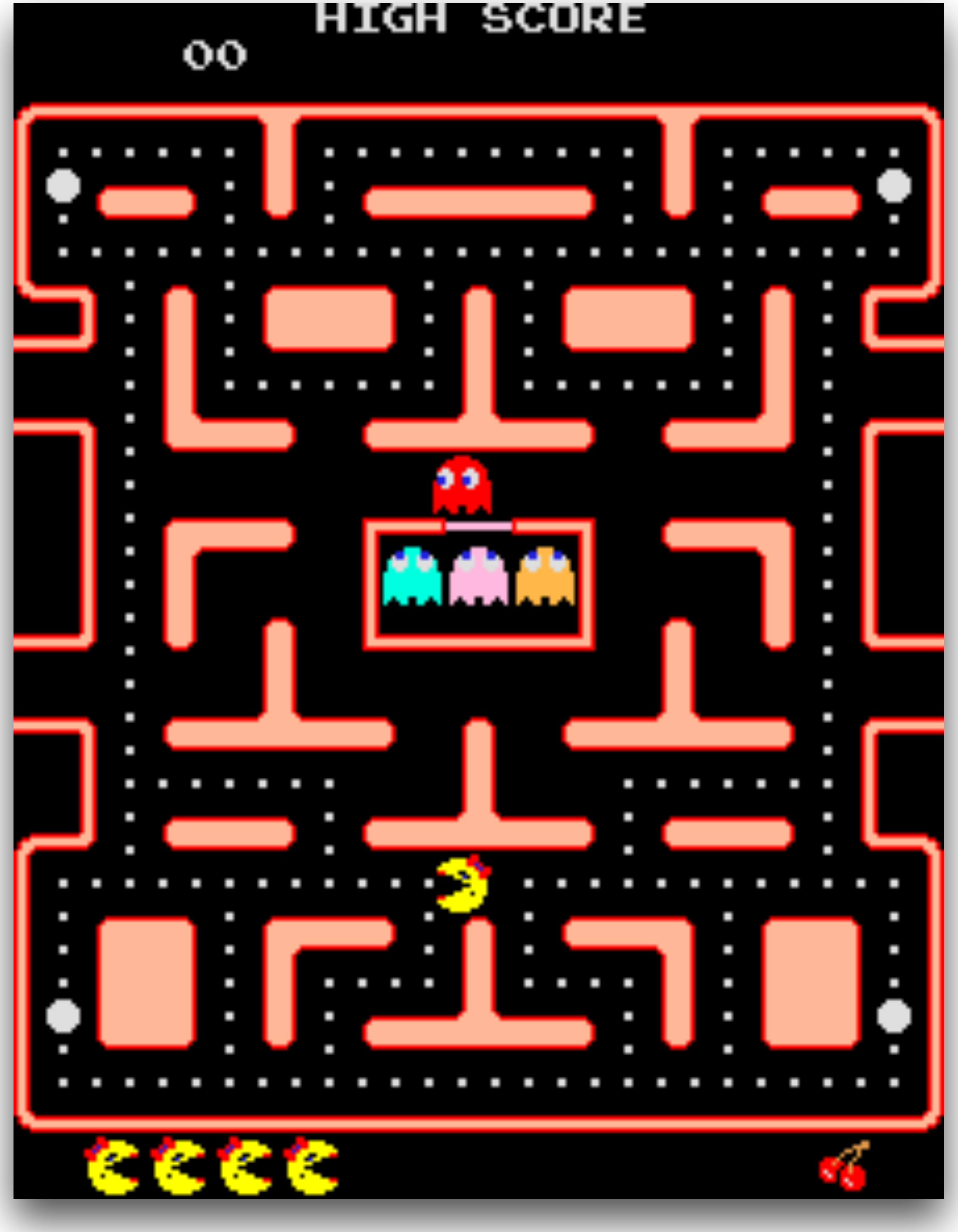
Douglas B. Lenat

Since 1984, a person-century of effort has gone into building CYC, a universal schema of roughly 10^5 general concepts spanning human reality. Most of the time has been spent codifying knowledge about these concepts; approximately 10^6 commonsense axioms have been handcrafted for and entered into CYC's knowledge base, and millions more have been inferred and cached by CYC. This article examines the fundamental assumptions of doing such a large-scale project, reviews the technical lessons learned by the developers, and surveys the range of applications that are or soon will be enabled by the technology.

One can think of CYC as an expert system with a domain that spans all everyday objects and actions. For example:

- You have to be awake to eat.
- You can usually see people's noses, but not their hearts.
- Given two professions, either one is a specialization of the other or else they are likely to be independent of one another.
- You cannot remember events that have not happened yet.
- If you cut a lump of peanut butter in half, each half is also a lump of peanut butter; but if you cut a table in half, neither half is a table.

By codifying reams of commonsense knowledge, CYC automates the white space in documents to help standardize—and make more efficient—information retrieval, integration, and consistency checking.



HIGH SCORE
00

Abstract/Formal Tasks vs Intuition

- Abstract and formal tasks that are among the most difficult undertakings for a human being are among the easiest for a computer.
- Computers have long been able to defeat even the best chess player but only recently have begun matching some of the abilities of average human beings to recognise objects or speech.
- Much of human knowledge is about unstructured “inputs” (e.g., sensory data).
- Computers need to capture this same knowledge in order to behave in an intelligent way.

Extracting Patterns from Raw Data

- AI systems need the ability of acquiring their own knowledge by extracting patterns from raw data.
  - Usually, this capability is referred to as *machine learning*.
  - Machine learning allows computer systems to learn through data and experience.
  - Examples: naive Bayes, logistic regression, decision tree, random forest, etc.

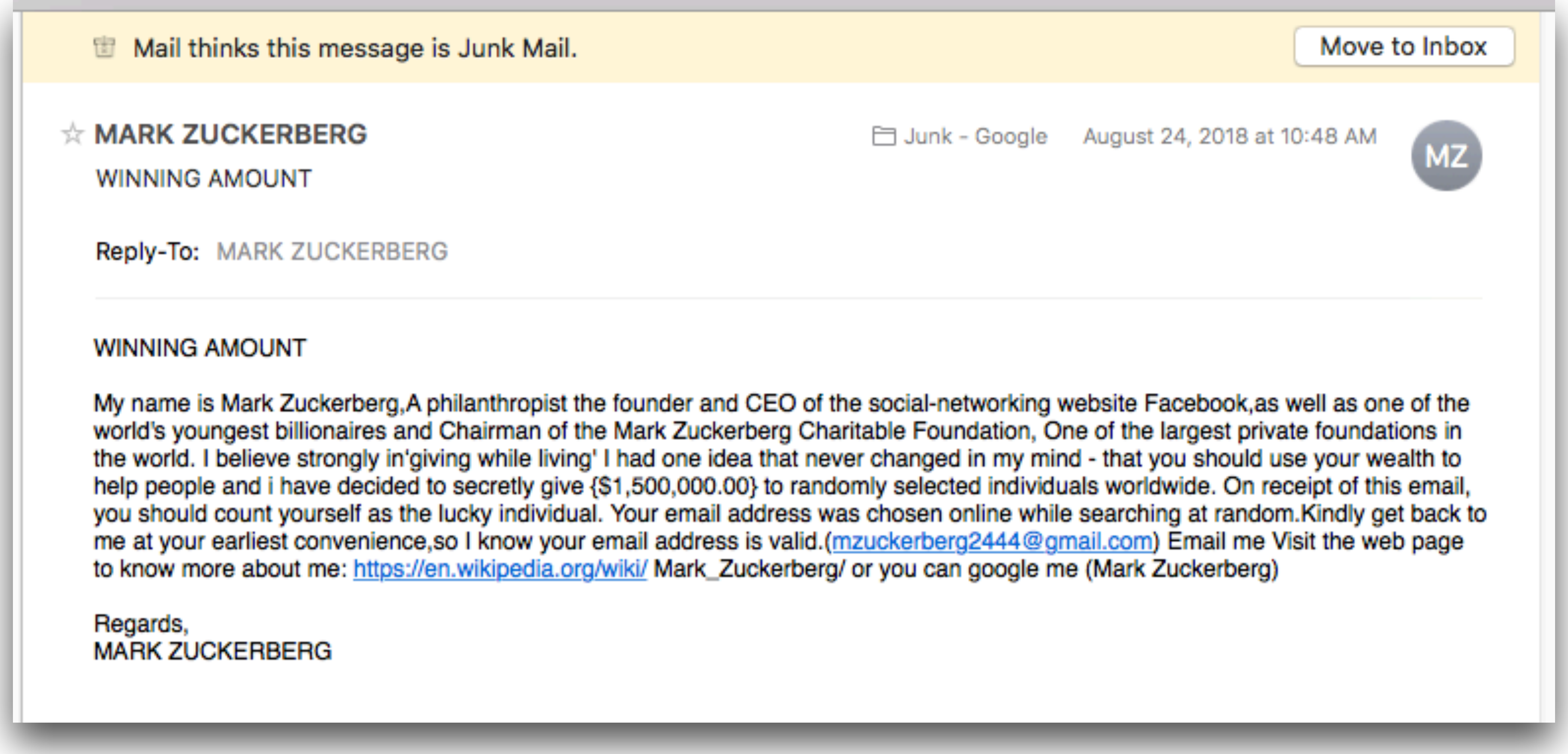
Representation and Features

- The performance of these algorithms depends heavily on the *representation* of the data they are given.

- For example, in order to determine if a patient has a certain disease or not we can input certain physiological indicators (with thresholds, for example, temperature higher than 37.5C) and/or the presence of absence of certain symptoms.

- Each piece of information included in the representation of the patient is called known as a *feature*.

- A machine learning algorithm (let's say logistic regression) learns how each of these features of the patient



MARK ZUCKERBERG

August 24, 2018 at 10:48 AM

WINNING AMOUNT

Reply-To: MARK ZUCKERBERG

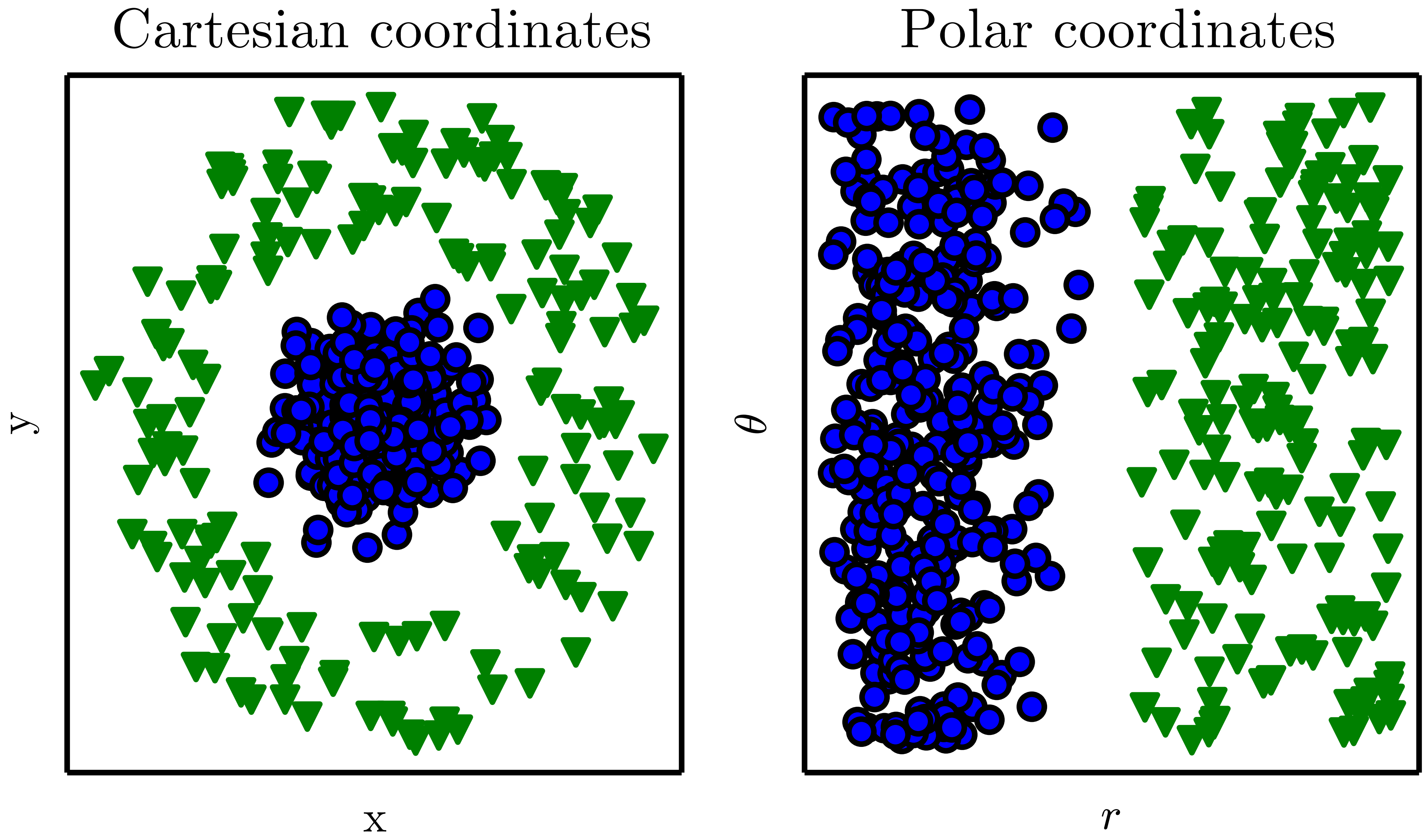
WINNING AMOUNT

My name is Mark Zuckerberg,A philanthropist the founder and CEO of the social-networking website Facebook,as well as one of the world's youngest billionaires and Chairman of the Mark Zuckerberg Charitable Foundation, One of the largest private foundations in the world. I believe strongly in'giving while living' I had one idea that never changed in my mind - that you should use your wealth to help people and i have decided to secretly give {$1,500,000.00} to randomly selected individuals worldwide. On receipt of this email, you should count yourself as the lucky individual. Your email address was chosen online while searching at random.Kindly get back to me at your earliest convenience,so I know your email address is valid.(mzuckerberg2444@gmail.com) Email me Visit the web page to know more about me: https://en.wikipedia.org/wiki/ Mark_Zuckerberg/ or you can google me (Mark Zuckerberg)

Regards,
MARK ZUCKERBERG

Representation and Features

- In this case features might be occurrences or not of certain words, formatting, length of the message or other information related to the email protocols.
- We can build a vector of values (continuous and discrete) representing each email. Each element of the vector will be associated to one feature.
- Many artificial intelligence tasks can be solved by designing the right set of features.
- However, for many tasks it is difficult to know what features should be extracted.
  - Thinks about identification of a photo, emotion in a voice of a speaker, understanding images of a road, playing a complex videogame (e.g., Starcraft).
- Note: it is only about the features themselves, but how the information is structured and "represented":
  - Machine learning on Roman numbers is probably not a good idea.



Credit: Goodfellow et al. 2016

Representation Learning

- One solution to this problem is to use machine learning *to discover not only the mapping from representation to output but also the representation itself*.

- This approach is usually known as *representation learning*.

- Learned representation often results in much better performance than can be obtained with hand-made representations.

Factors of Variation

- When designing features or algorithms for learning features, our goal is to separate the *factors of variation* that explain the observed data.
- In this context, we use word *factors* to refer to separate sources of information that are useful for the machine learning task at hand.
- Such factors are often quantities that are not directly observed.
  - They are often unobserved (or latent) and they affect the observable ones.
  - Some of them might be linked to human constructs (e.g., the colour of an object) and other might not. In the latter case the factors might not easily interpreted by a human (see also the problem of AI interpretability).

Factors of Variation

- Some factors of variations affect all the piece of information we have (for example angle of view of a car).
- We need to disentangle the factors that allow us to successfully perform the machine learning task and extract representations that are not affected by factors of variation that are not “useful” for the task.
- For example, if we need to classify a car vs truck, the angle itself is not fundamental for the classification.
- We need a tool that learns to “ignore” that factor of variation.

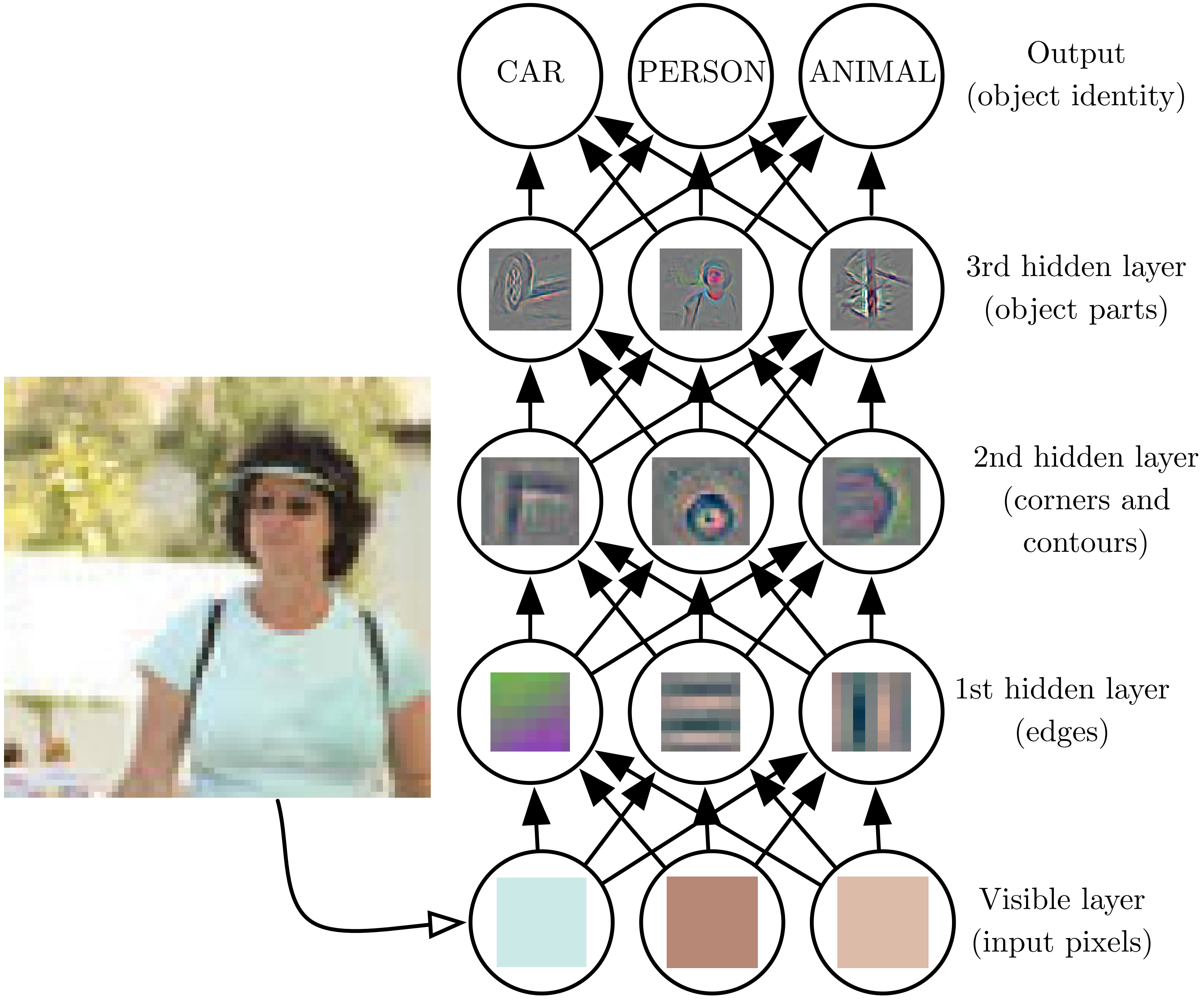
Deep Learning

- Deep learning address this problem of representation learning by introducing representations that are expressed in terms of other, simpler representations.

- Classic example of deep learning model is the feedforward deep network (or multi-layer perceptron).

- Please note: a multilayer perceptron at the end is a (complex) mathematical function mapping input values to output values.

- This function is the result of combining several simpler functions in the intermediate nodes.



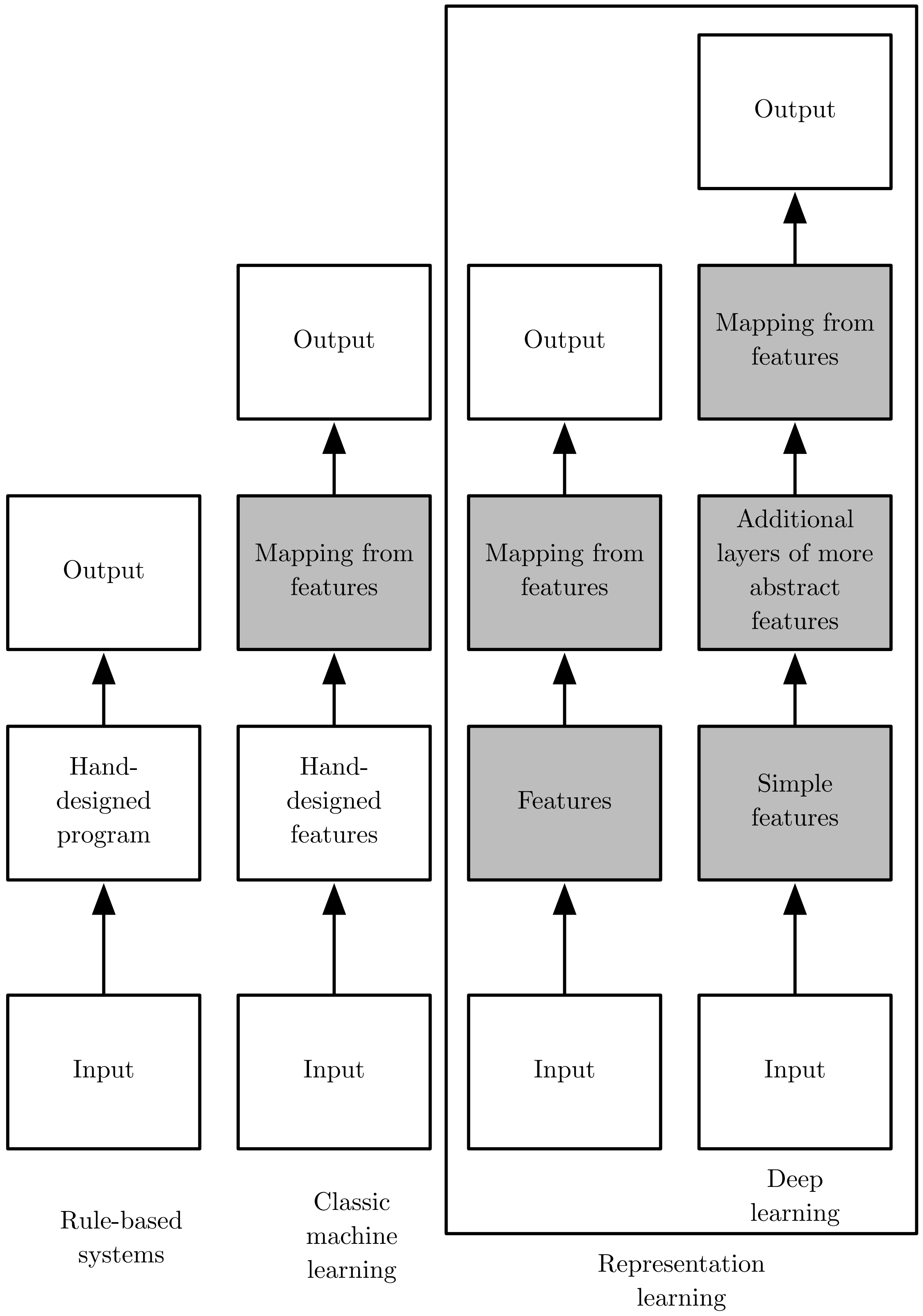
Credit: Zeiler and Fergus 2014

Deep Learning

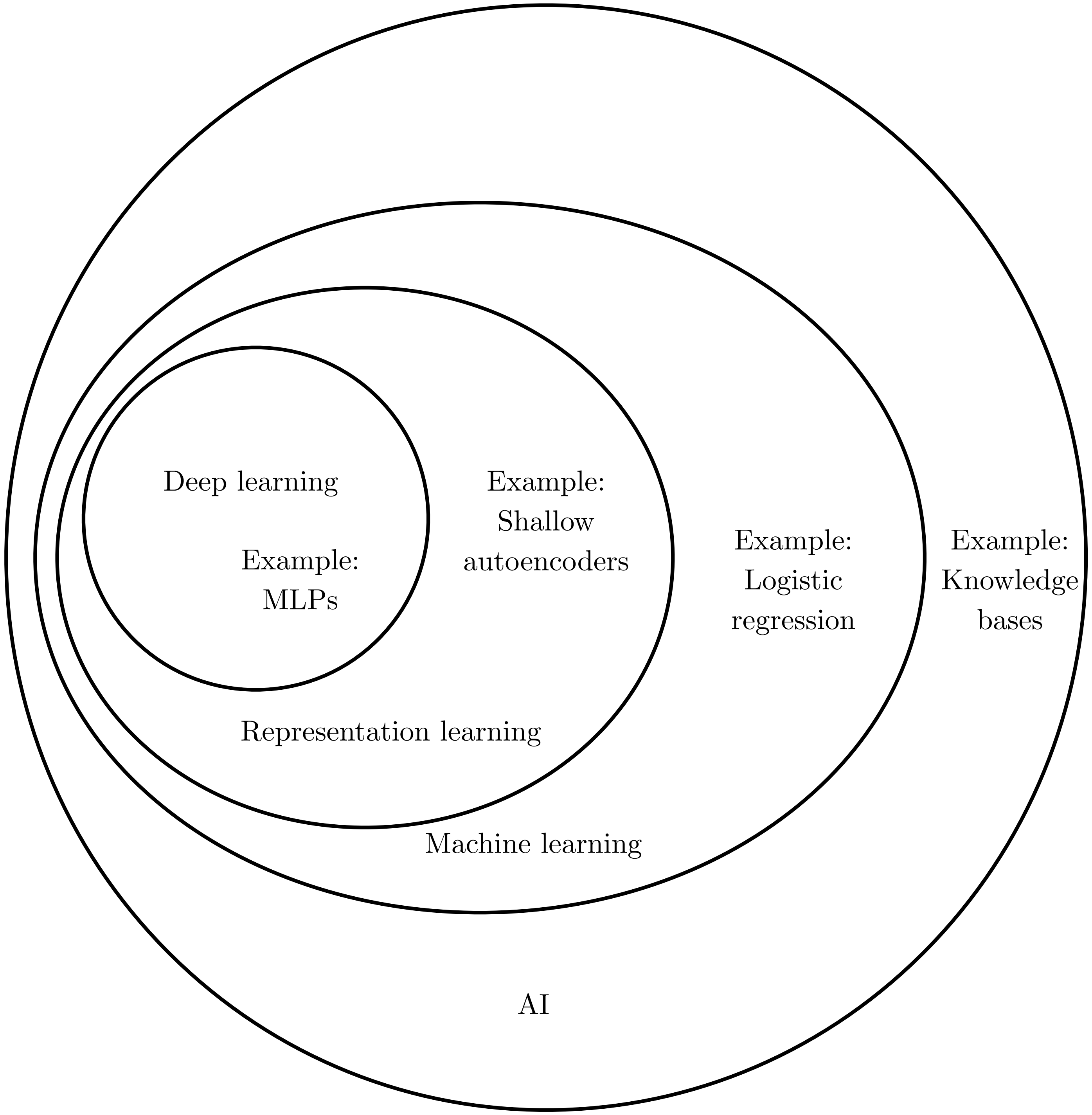
- Until now, we have considered one of the possible interpretation of deep learning, i.e., that it allows to learn the right representation for the data.

- Another possible perspective on deep learning is that depth enables a computer to learn a multi-step computer program.

- Each layer of the network can be thought as the state of the computer's memory after executing a set of instructions in parallel.



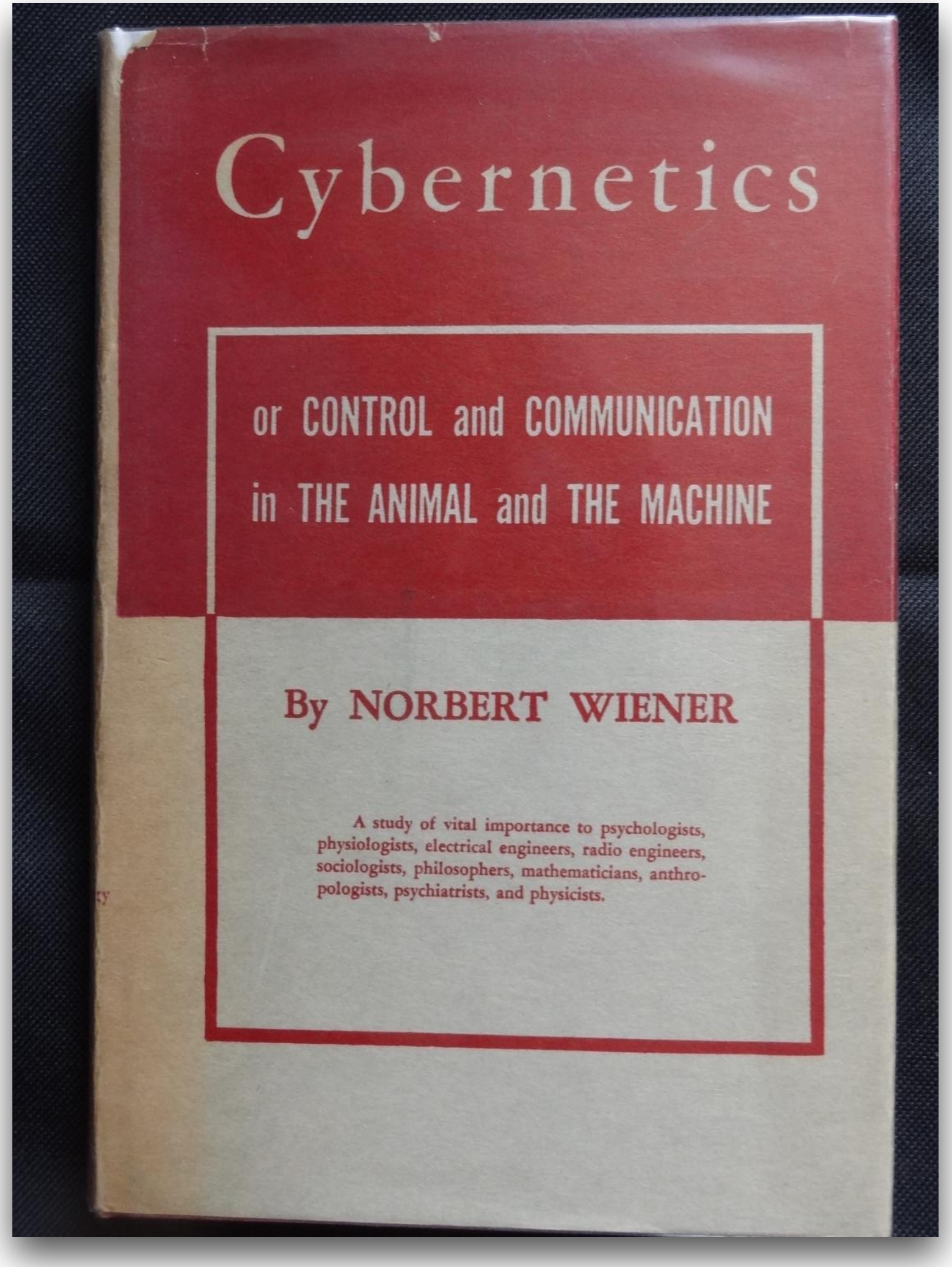
Credit: Goodfellow et al. 2016



Credit: Goodfellow et al. 2016

From Theories of Biological Learning to Deep Learning

- Three waves:
  - Cybernetics (1940s-1960s)
  - Connectionism (1980s-1990s)
  - Deep learning (2006-today)
- Some of the earliest learning algorithms were intended to be computational models of the brain. As a result, one of the names used for deep learning is *artificial neural networks (ANNs)*.



Cybernetics

or CONTROL and COMMUNICATION
in THE ANIMAL and THE MACHINE

By NORBERT WIENER

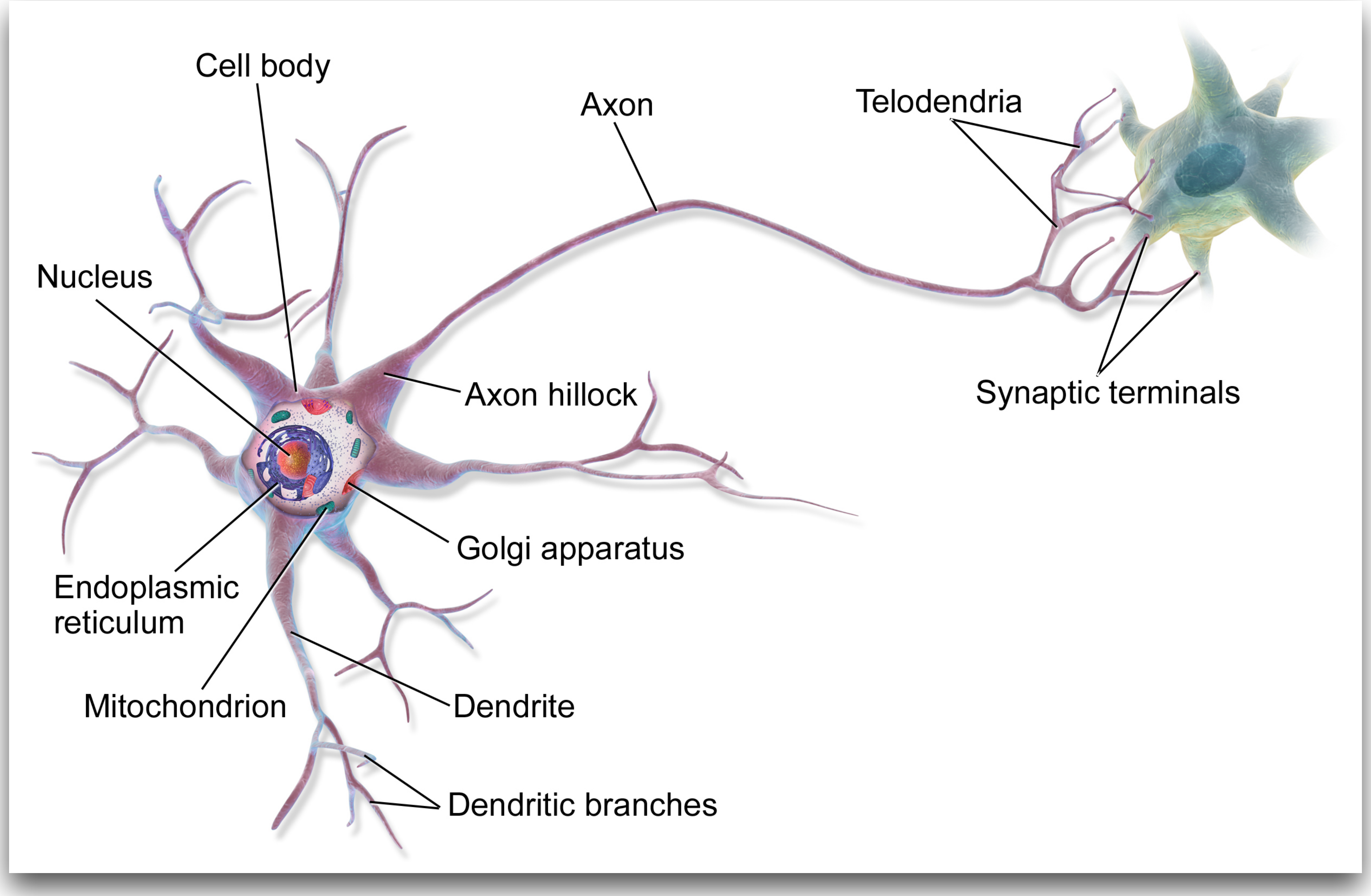
A study of vital importance to psychologists, physiologists, electrical engineers, radio engineers, sociologists, philosophers, mathematicians, anthropologists, psychiatrists, and physicists.

Artificial Neural Networks and Neuroscience

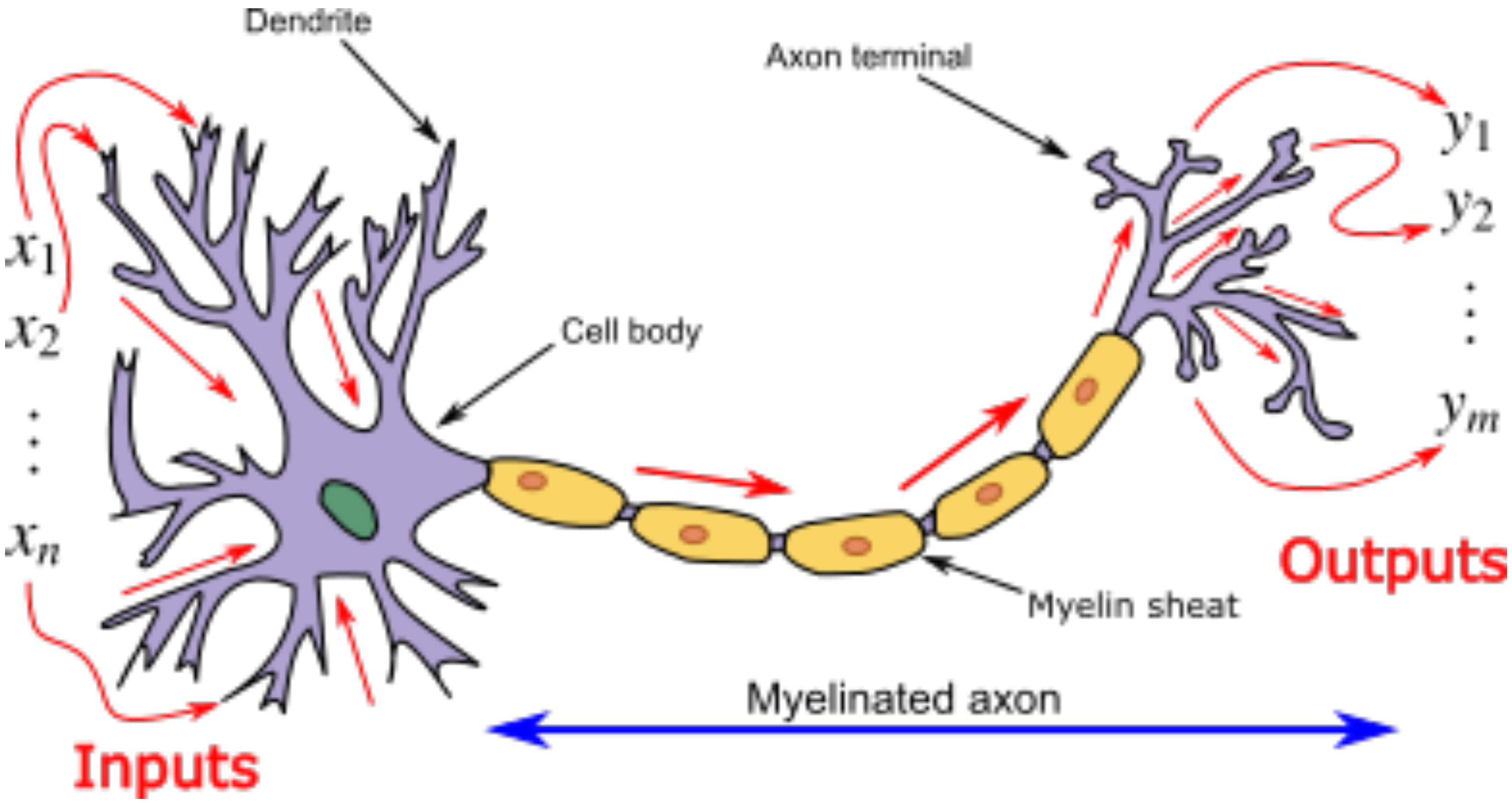
- The earliest predecessors of modern deep learning were simple linear models motivated from a neuroscience perspective.

- These models were designed to take a series of n input values $x_1, x_2, \ldots, x_n$ and associate them to an output y.

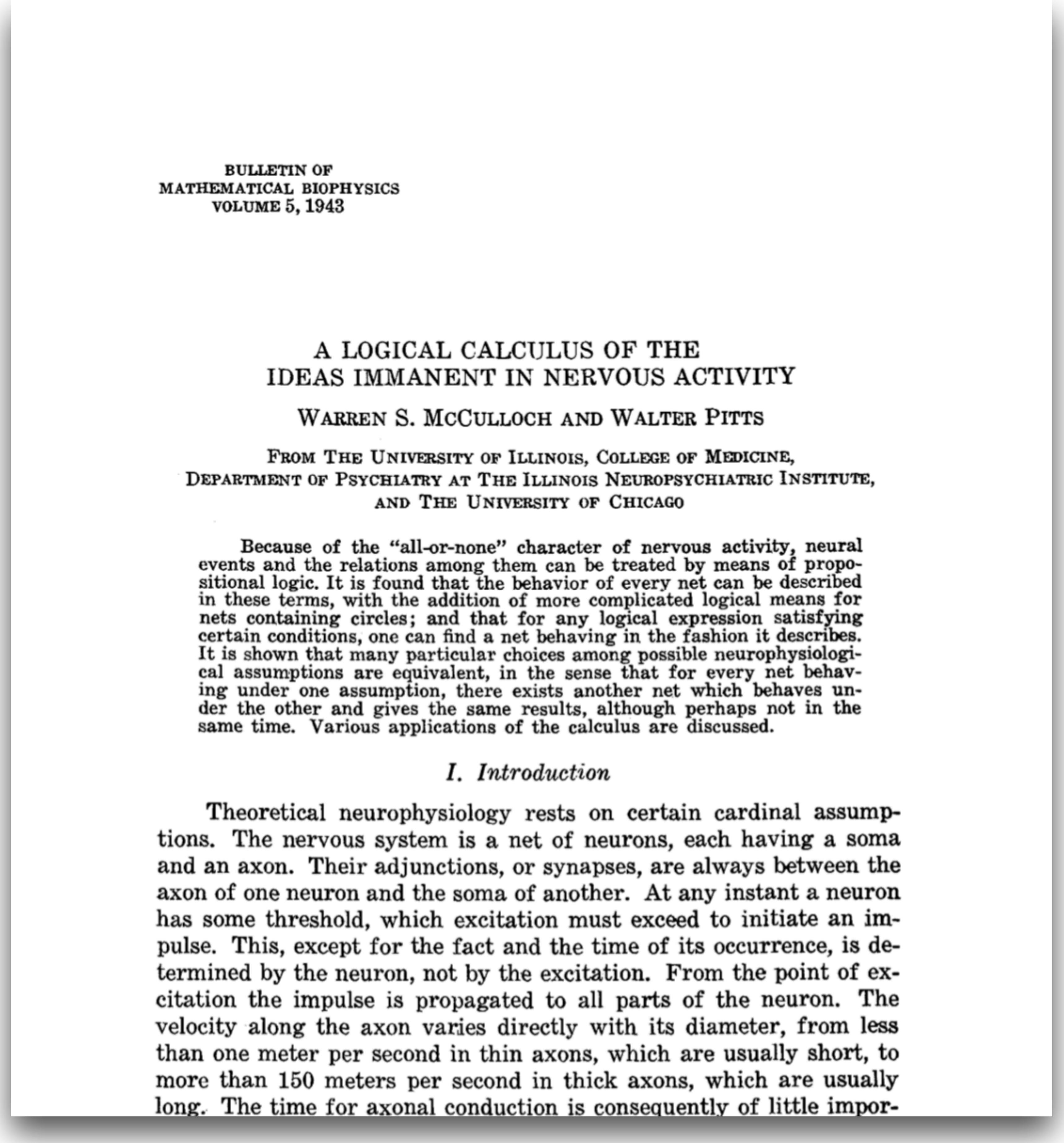
- These models would be based or learn a set of weights: $y = f(\mathbf{x}, \mathbf{w}) = w_1 x_1 + \ldots + w_n x_n$



Credit: Wikimedia



Credit: Wikimedia



BULLETIN OF
MATHEMATICAL BIOPHYSICS
VOLUME 5, 1943

A LOGICAL CALCULUS OF THE IDEAS IMMANENT IN NERVOUS ACTIVITY

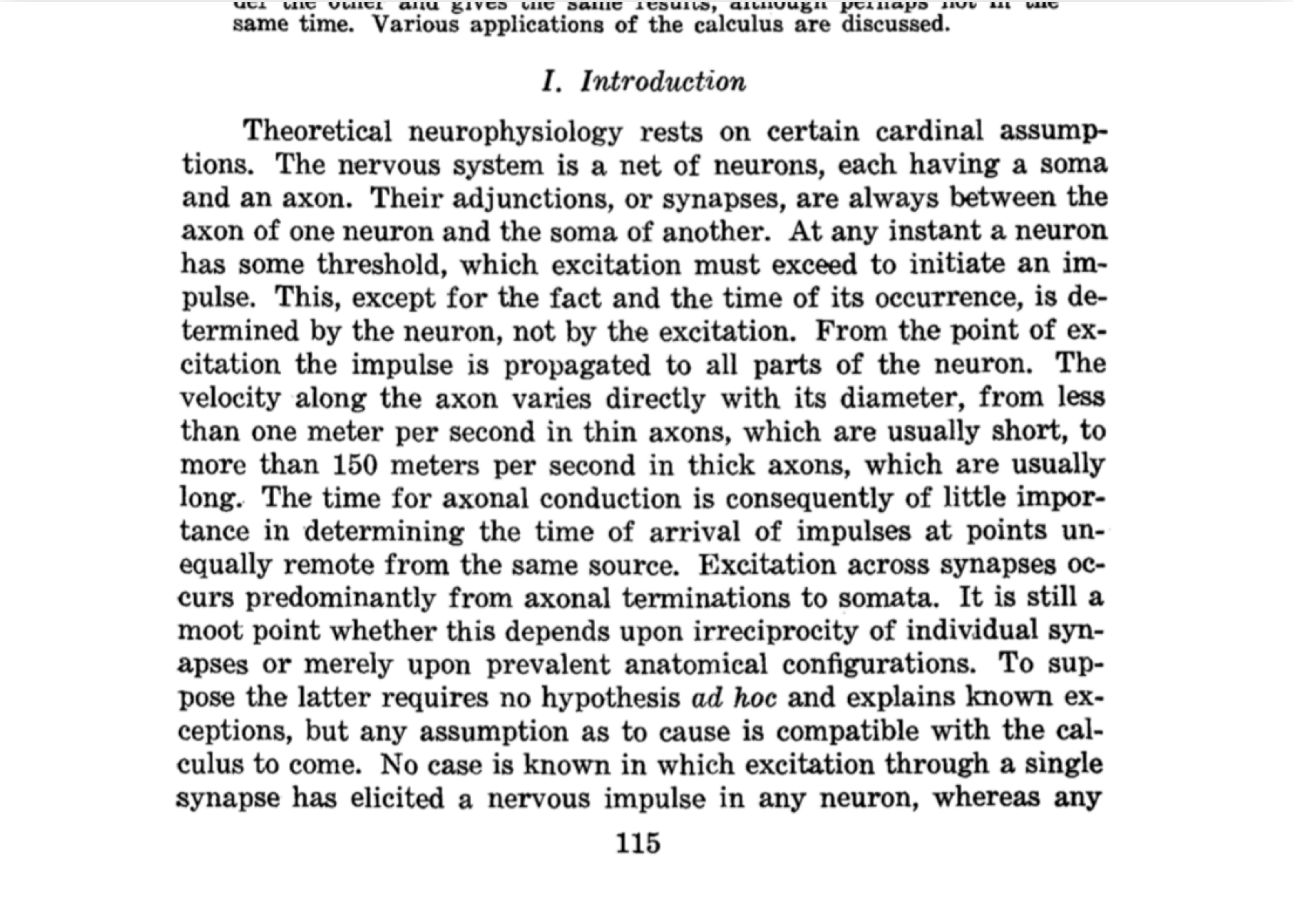
WARREN S. MCCULLOCH AND WALTER PITTS

FROM THE UNIVERSITY OF ILLINOIS, COLLEGE OF MEDICINE,
DEPARTMENT OF PSYCHIATRY AT THE ILLINOIS NEUROPSYCHIATRIC INSTITUTE,
AND THE UNIVERSITY OF CHICAGO

Because of the "all-or-none" character of nervous activity, neural events and the relations among them can be treated by means of propositional logic. It is found that the behavior of every net can be described in these terms, with the addition of more complicated logical means for nets containing circles; and that for any logical expression satisfying certain conditions, one can find a net behaving in the fashion it describes. It is shown that many particular choices among possible neurophysiological assumptions are equivalent, in the sense that for every net behaving under one assumption, there exists another net which behaves under the other and gives the same results, although perhaps not in the same time. Various applications of the calculus are discussed.

I. *Introduction*

Theoretical neurophysiology rests on certain cardinal assumptions. The nervous system is a net of neurons, each having a soma and an axon. Their adjunctions, or synapses, are always between the axon of one neuron and the soma of another. At any instant a neuron has some threshold, which excitation must exceed to initiate an impulse. This, except for the fact and the time of its occurrence, is determined by the neuron, not by the excitation. From the point of excitation the impulse is propagated to all parts of the neuron. The velocity along the axon varies directly with its diameter, from less than one meter per second in thin axons, which are usually short, to more than 150 meters per second in thick axons, which are usually long. The time for axonal conduction is consequently of little impor-



der the other and gives the same results, although perhaps not in the same time. Various applications of the calculus are discussed.

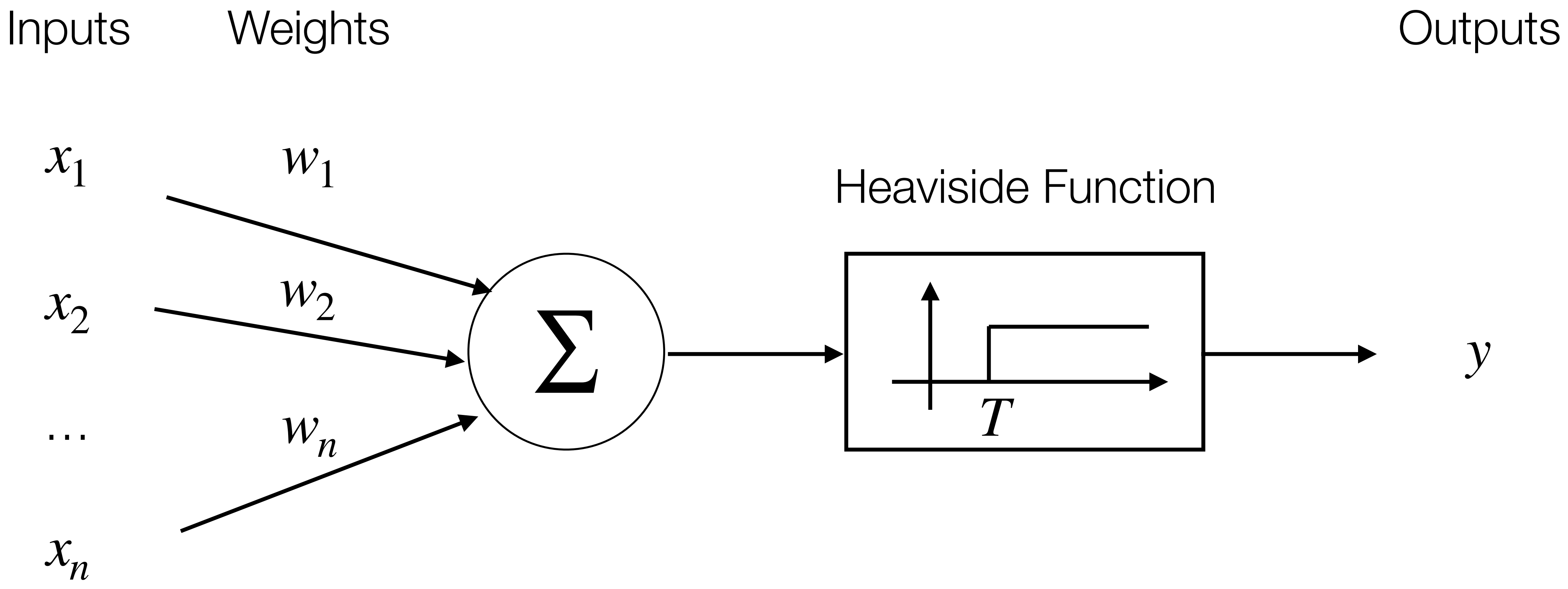
I. Introduction

Theoretical neurophysiology rests on certain cardinal assumptions. The nervous system is a net of neurons, each having a soma and an axon. Their adjunctions, or synapses, are always between the axon of one neuron and the soma of another. At any instant a neuron has some threshold, which excitation must exceed to initiate an impulse. This, except for the fact and the time of its occurrence, is determined by the neuron, not by the excitation. From the point of excitation the impulse is propagated to all parts of the neuron. The velocity along the axon varies directly with its diameter, from less than one meter per second in thin axons, which are usually short, to more than 150 meters per second in thick axons, which are usually long. The time for axonal conduction is consequently of little importance in determining the time of arrival of impulses at points unequally remote from the same source. Excitation across synapses occurs predominantly from axonal terminations to somata. It is still a moot point whether this depends upon irreciprocity of individual synapses or merely upon prevalent anatomical configurations. To suppose the latter requires no hypothesis *ad hoc* and explains known exceptions, but any assumption as to cause is compatible with the calculus to come. No case is known in which excitation through a single synapse has elicited a nervous impulse in any neuron, whereas any

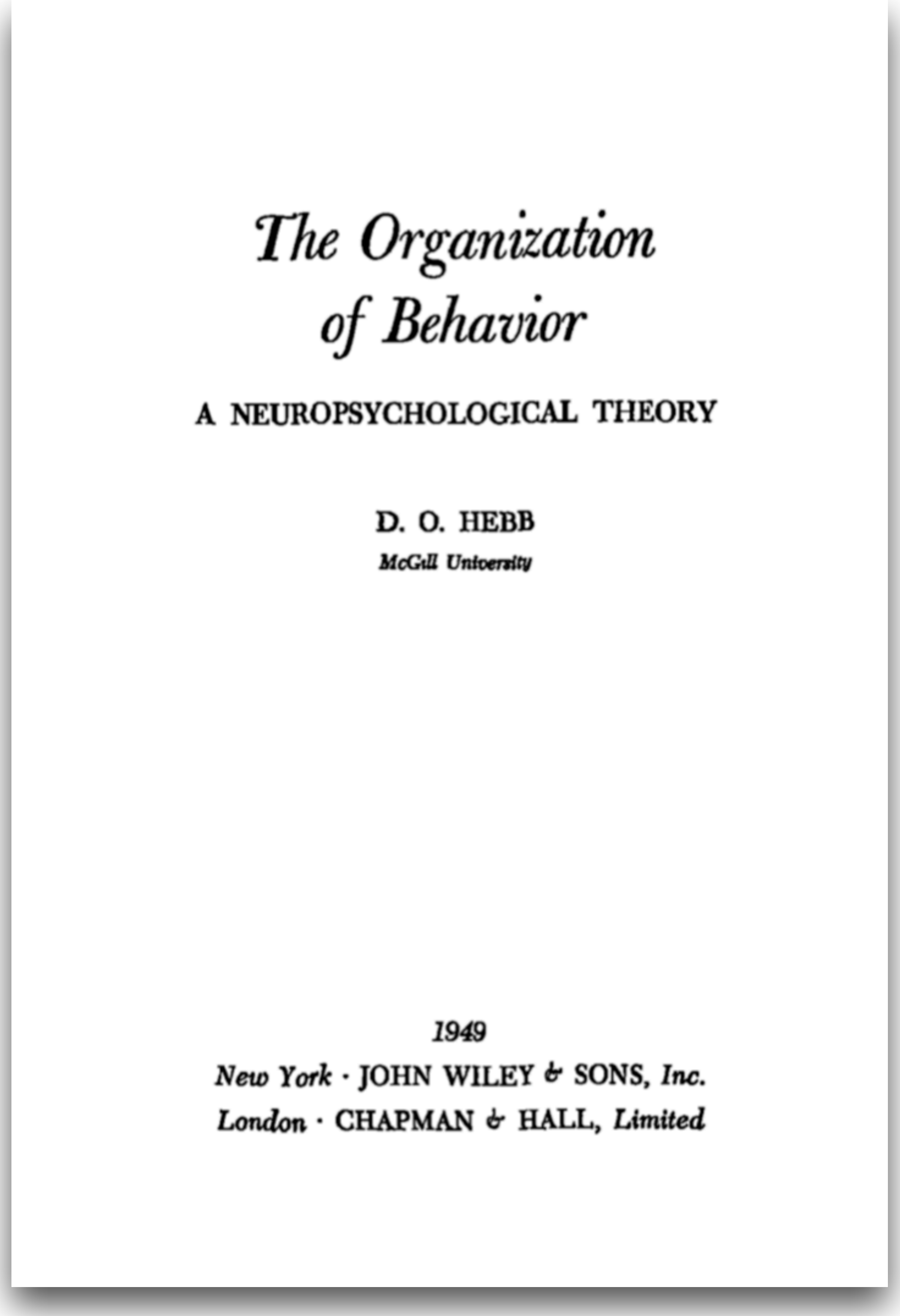
The McCulloch-Pitts Neural Model

- In “A Logical Calculus of the Ideas Imminent in Nervous Activity” (1943), McCulloch and Pitts suggested a model about how thought executes.
- This is the original inspiration of current deep learning models.
- The set of operations is defined over two values:
  - True (1)
  - False (0)
- The calculus contained NOT, AND, OR. By changing the (fixed) values of the weights, you can obtain different functions.

McCulloch-Pitts Model of Neuron



In the McCulloch-Pitts model, the values of the weights are fixed.



The Organization of Behavior

A NEUROPSYCHOLOGICAL THEORY

D. O. HEBB

McGill University

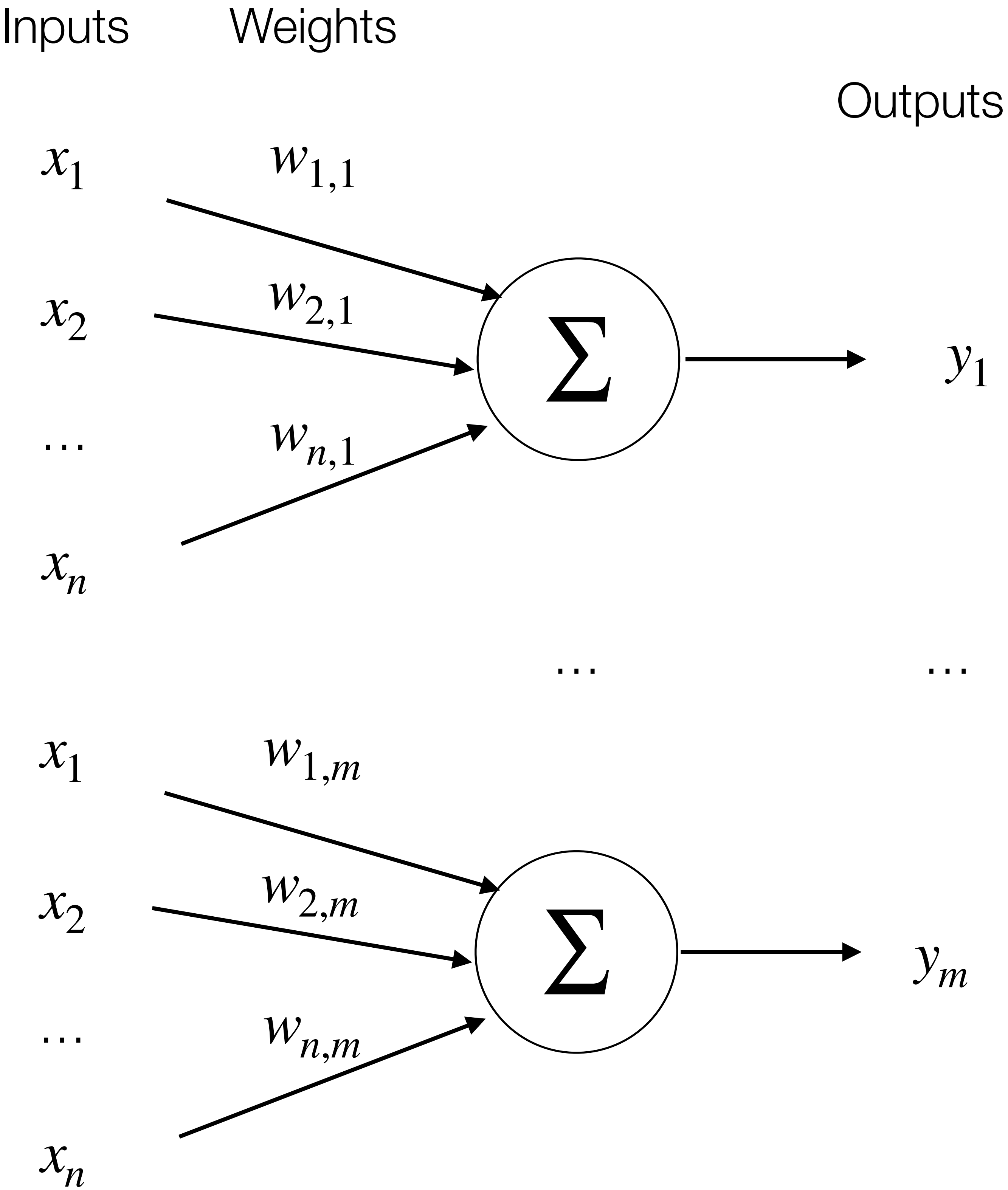
1949

New York · JOHN WILEY & SONS, *Inc.*
London · CHAPMAN & HALL, *Limited*

Hebb’s Law

- From Hebb’s “The Organization of Behavior” (1949): “When an axon cell A is near enough to excite cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such as that A’s efficiency, as one of the cells firing B is increased”.
- This is usually referred to as “Hebb’s Law”.
- First simulations of artificial neural networks in 1950s based on Hebb’s model.
- Weights of the models are also called synaptic connectivity.

Hebb's Model of Neuron



The Hebbian network model has n-node input layer:
$\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$
and an m-node output layer
$\mathbf{y} = [y_1, y_2, \ldots, y_n]^T$

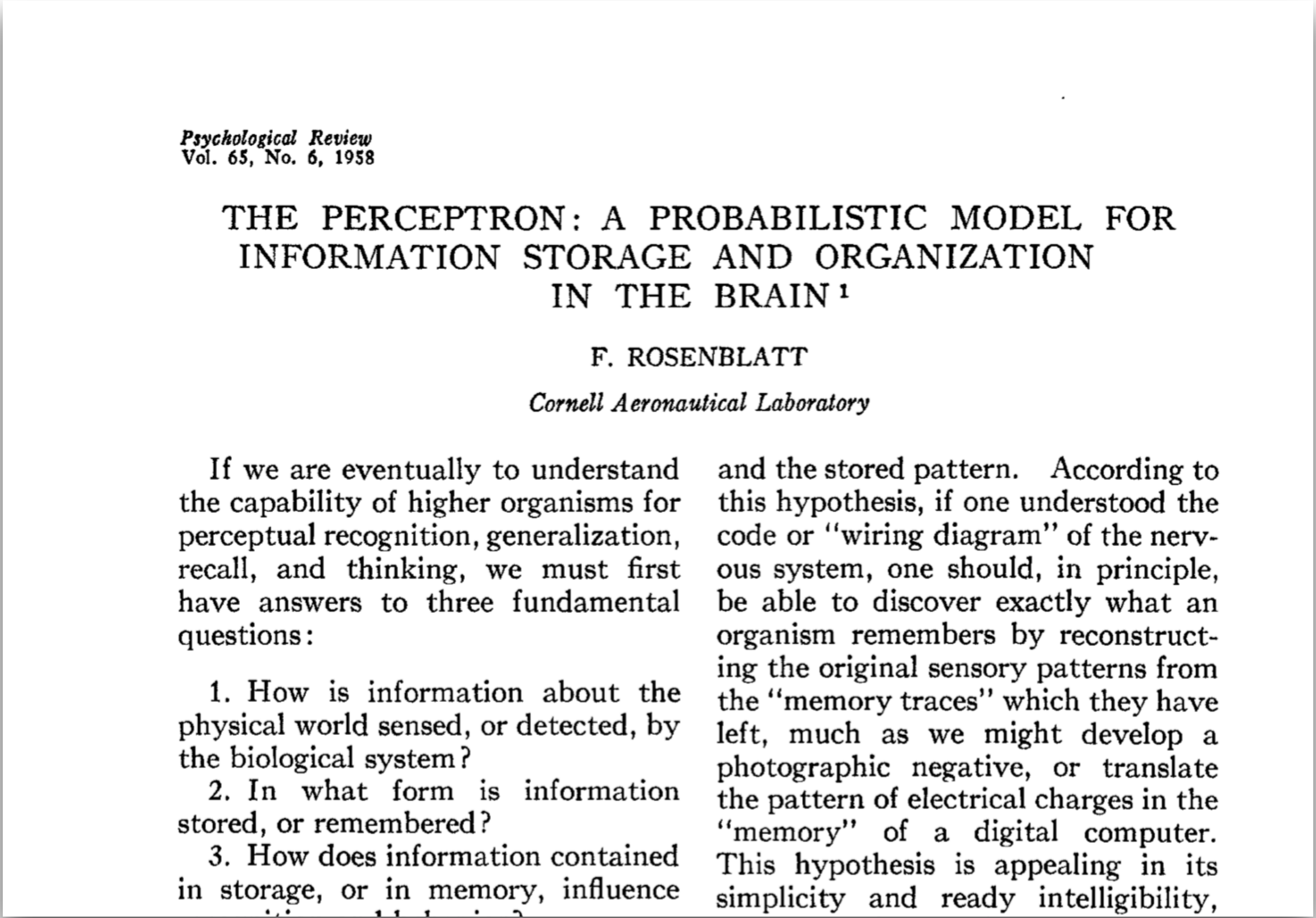
Each output is connected to all input as follow:

$$y_i = \sum_{j=1}^{n} w_{i,j} x_j$$

The learning rule is the following:

$$w_{i,j}^{new} \leftarrow w_{i,j}^{old} + \eta x_j y_i$$

η is the learning rate.



Psychological Review
Vol. 65, No. 6, 1958

THE PERCEPTRON: A PROBABILISTIC MODEL FOR INFORMATION STORAGE AND ORGANIZATION IN THE BRAIN[1]

F. ROSENBLATT

Cornell Aeronautical Laboratory

If we are eventually to understand the capability of higher organisms for perceptual recognition, generalization, recall, and thinking, we must first have answers to three fundamental questions:

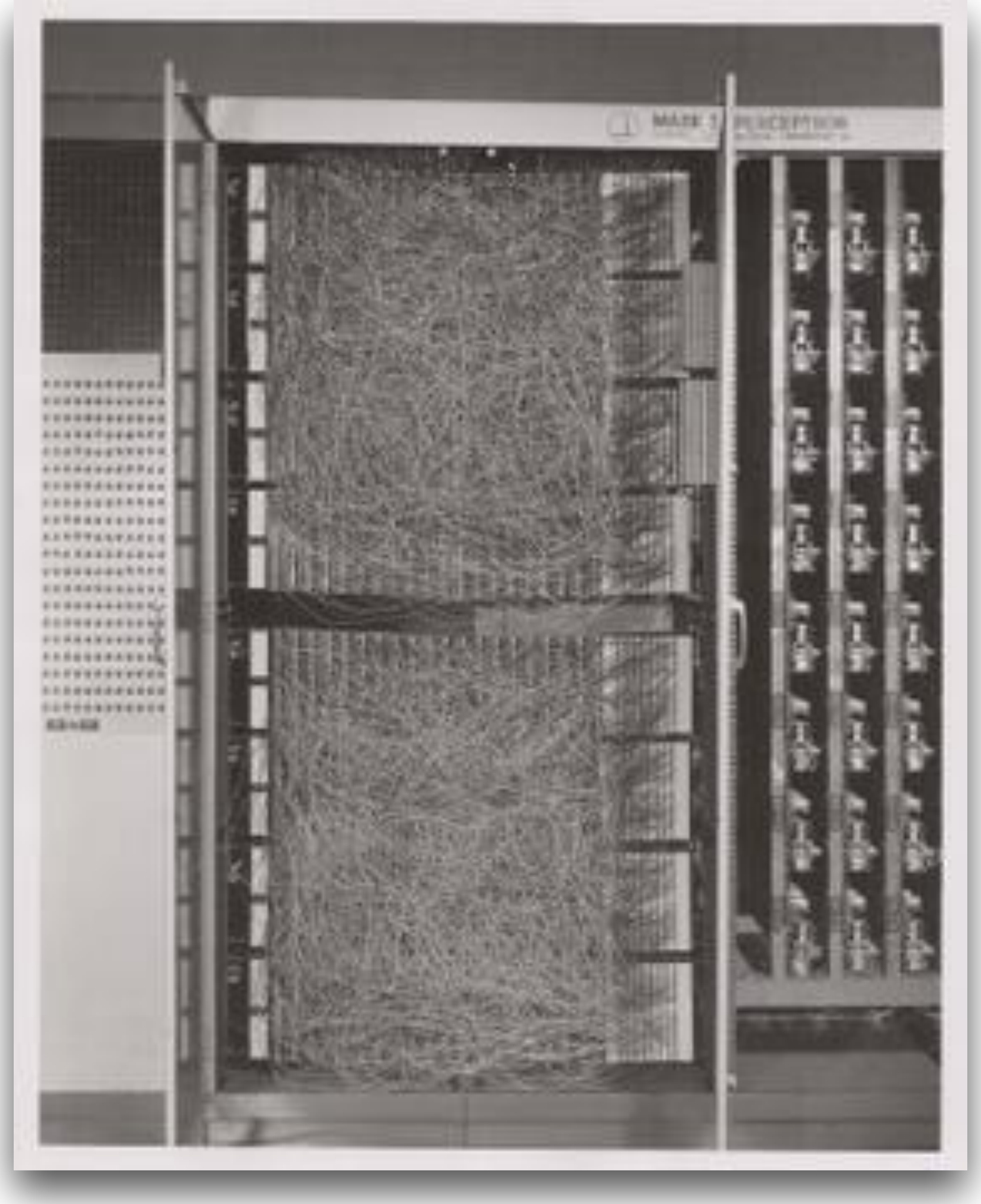
1. How is information about the physical world sensed, or detected, by the biological system?
2. In what form is information stored, or remembered?
3. How does information contained in storage, or in memory, influence

and the stored pattern. According to this hypothesis, if one understood the code or "wiring diagram" of the nervous system, one should, in principle, be able to discover exactly what an organism remembers by reconstructing the original sensory patterns from the "memory traces" which they have left, much as we might develop a photographic negative, or translate the pattern of electrical charges in the "memory" of a digital computer. This hypothesis is appealing in its simplicity and ready intelligibility,

Rosenblatt’s Perceptron Model

- Frank Rosenblatt’s perceptron models the first one with variables weights that were learned from examples:
  - Learning the weights of categories given examples of those categories.
- The perceptron was intended to be a machine rather than a program.
  - First implementation was actually for IBM 704.

IBM 704



Mark I Perceptron

Image for image recognition

400 photocells randomly connected to the neurons

Weights encoded in potentiometers

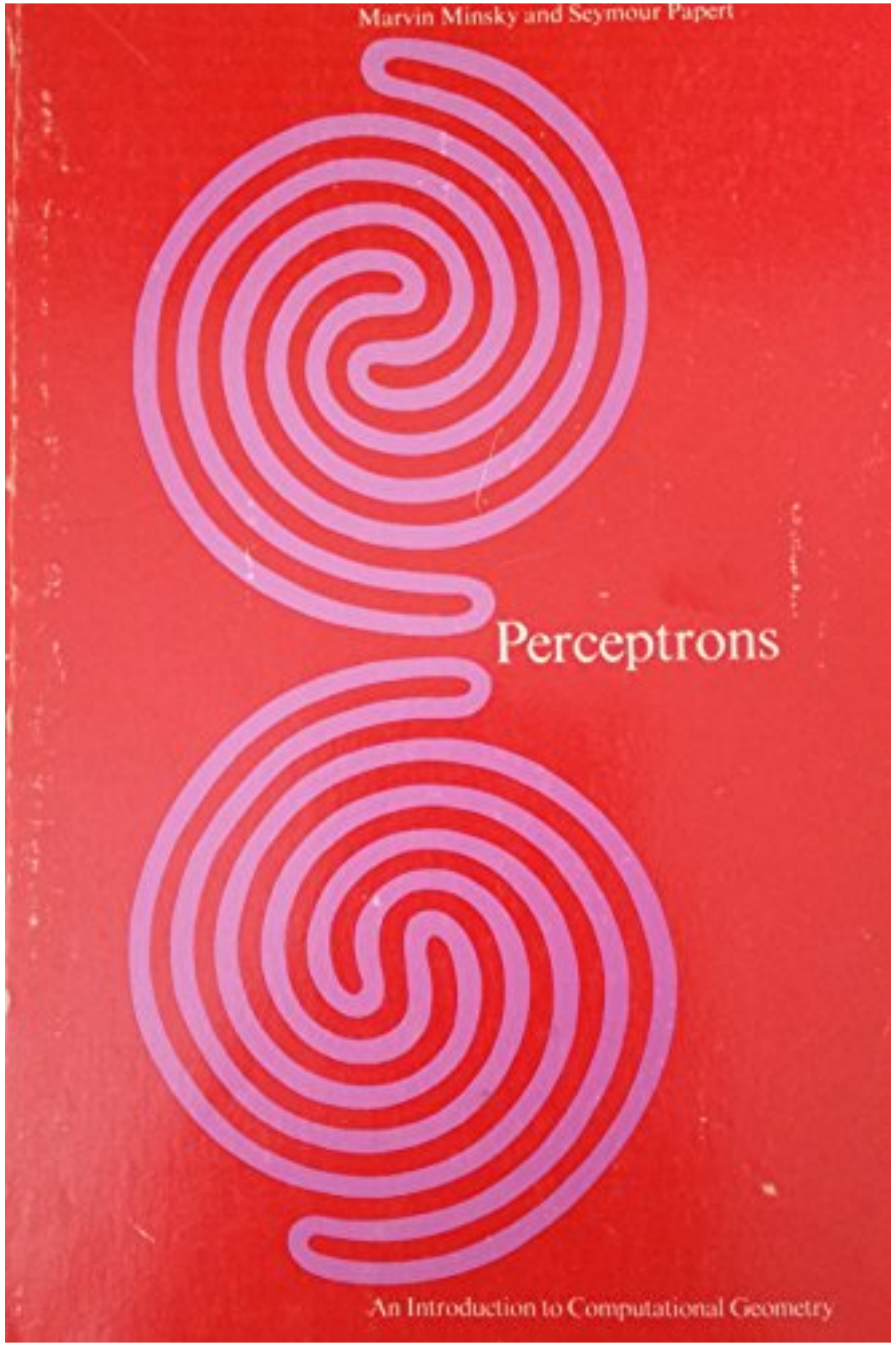
Limitations of Perceptrons

- Linear models have many limitations.

- Most famously, they cannot learn the XOR function where $f([0,1], \mathbf{w}) = 1$ and $f([1,0], \mathbf{w}) = 1$ but $f([1,1], \mathbf{w}) = 0$ and $f([0,0], \mathbf{w}) = 0$.

- This was observed by Minsky and Papert in 1969 in Perceptrons.

- This was the first major dip in the popularity of neural networks.


Marvin Minsky and Seymour Papert
Perceptrons
An Introduction to Computational Geometry

References

- Chapter 1 of Ian Goodfellow, Yoshua Bengio and Aaron Courville. Deep Learning. MIT Press. 2016.