

1 Submitted to *Journal of Vision*.

2 EMERGING TREND IN VISION SCIENCE

3 **Deep learning: Using machine learning**  
4 **to study biological vision**

5

6 Najib J. Majaj<sup>1</sup> and Denis G. Pelli<sup>1,2</sup>

7 <sup>1</sup>Center for Neural Science and <sup>2</sup>Department of Psychology, New York University

8

9 **ABSTRACT**

10 Today many vision-science presentations employ machine learning, especially the version  
11 called “deep learning”. Many neuroscientists use machine learning to decode neural responses.  
12 Many perception scientists try to understand how living organisms recognize objects. To them,  
13 deep neural networks offer benchmark accuracies for recognition of learned stimuli. Originally  
14 machine learning was inspired by the brain. Today, machine learning is used as a statistical tool  
15 to decode brain activity. Tomorrow, deep neural networks might become our best model of brain  
16 function. This brief overview of the use of machine learning in biological vision touches on its  
17 strengths, weaknesses, milestones, controversies, and current directions. Here, we hope to help  
18 vision scientists assess what role machine learning should play in their research.

## 19 INTRODUCTION

20 What does machine learning offer to biological-vision  
21 scientists? Machine learning was developed as a tool for  
22 automated classification, optimized for accuracy.  
23 Machine learning is used in a broad range of applications  
24 (Brynjolfsson, 2018), e.g. regression in stock market  
25 forecasting and reinforcement learning to play chess, but  
26 here we focus on classification. Physiologists use it to  
27 identify stimuli based on neural activity. To study  
28 perception, physiologists measure neural activity and  
29 psychophysicists measure overt responses, like pressing  
30 a button. Physiologists and psychophysicists are starting  
31 to consider deep learning as a model for object  
32 recognition by human and nonhuman primates (Cadieu  
33 et al., 2014; Ziskind et al., 2014; Yamins et al., 2014;  
34 Khaligh-Razavi & Kriegeskorte, 2014; Testolin, Stoianov,  
35 & Zorzi, 2017). We suppose that most of our readers  
36 have heard of machine learning but are wondering  
37 whether it would be useful in their own research. We  
38 begin by describing some of its pluses and minuses.

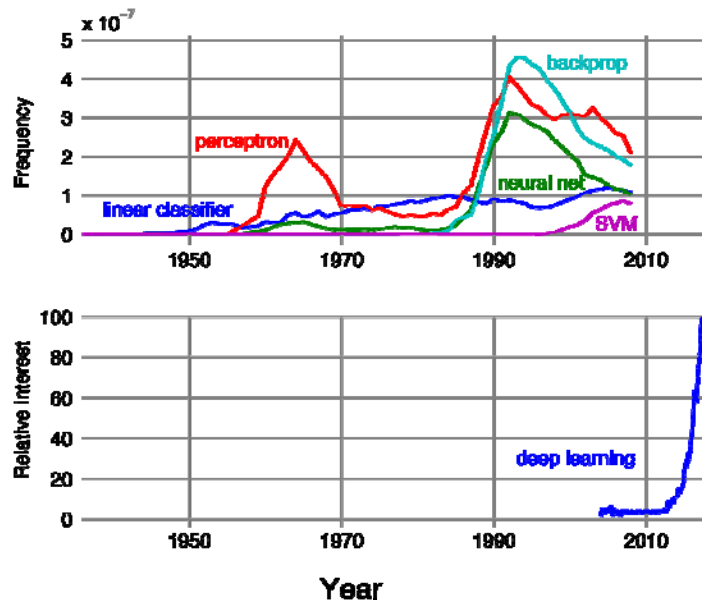
## 39 PLUSES: WHAT IT'S GOOD FOR

40 At the very least, machine learning is a powerful tool for  
41 interpreting biological data. A particular form of machine learning, *deep learning*, is very popular

## GLOSSARY

***Machine learning*** is any computer algorithm that learns how to perform a task directly from examples, without a human providing explicit instructions or rules for how to do so. In one type of machine learning, called “supervised learning,” correctly labeled examples are provided to the learning algorithm, which is then “trained” (i.e. its parameters are adjusted) to be able to perform the task correctly on its own and generalize to unseen examples.

***Deep learning*** is a newly successful and popular version of machine learning that uses backprop (defined below) neural networks with multiple hidden layers. The 2012 success of AlexNet, then the best machine learning network for object recognition, was the tipping point. Deep learning is now ubiquitous in the internet. The idea is to have each layer of processing perform successively more complex computations on the data to give the full “multi-layer” network more expressive power. The drawback is that it is much harder to train multi-layer networks (Goodfellow et al. 2016). Deep learning ranges from discovering the weights of a multilayer network to parameter learning in hierarchical belief networks.



**Figure 1. Top:** The frequency of appearance of each of five terms — “linear classifier”, “perceptron”, “support vector machine”, “neural net” and “backprop” — in books indexed by Google in each year of publication. Google counts instances of words and phrases of  $n$  words, and calls each an “ngram”. Frequency is reported as a fraction of all instances of ngrams of that length, normalized by the number of books published that year (ngram / year / books published). The figure was created using Google’s ngram viewer (<https://books.google.com/ngrams>), which contains a yearly count of ngrams found in sources printed between 1500 and 2008. **Bottom:** Numbers represent worldwide search interest relative to the highest point on the chart for the given year for the term “deep learning” (as reported by <https://trends.google.com/trends/>).

43 (Fig. 1). Is it just a fad? For computer vision, the old  
44 paradigm was: feature detection, followed by  
45 segmentation, and then grouping (Marr, 1982). With  
46 machine learning tools, the new paradigm is to just  
47 define the task and provide a set of labeled examples,  
48 and the algorithm builds the classifier. (This is  
49 “supervised” learning; we discuss unsupervised learning  
50 below.)

51 Unlike the handcrafted pattern recognition (including  
52 segmentation and grouping) popular in the 70’s and 80’s,  
53 deep learning algorithms are generic, with little domain-  
54 specificity.<sup>1</sup> They replace hand-engineered feature  
55 detectors with filters that can be learned from the data.  
56 Advances in the mid 90’s in machine learning made it  
57 useful for practical classification, e.g. handwriting  
58 recognition (LeCun et al., 1989; Vapnik, 1999).

59 Machine learning allows a neurophysiologist to decode  
60 neural activity without knowing the receptive fields  
61 (Seung & Sompolinsky, 1993; Hung et al., 2005).

62 Machine learning is a big step in the shifting emphasis in  
63 neuroscience from *how* the cells encode to *what* they

64 encode, i.e. what that code tells us about the stimulus (Barlow, 1953; Geisler, 1989). Mapping a  
65 receptive field is the foundation of neuroscience (beginning with Weber’s 1834/1996 mapping of

**Neural nets** are computing systems inspired by biological neural networks that consist of individual neurons learning their connections with other neurons in order to solve tasks by considering examples.

**Supervised learning** refers to any algorithm that accepts a set of labeled stimuli — a training set — and returns a classifier that can label stimuli similar to those in the training set.

**Unsupervised learning** discovers structure and redundancy in data without labels. It is less widely used than supervised learning, but of great interest because labeled data are scarce while unlabeled data are plentiful.

**Cost function.** A function that assigns a real number representing cost to a candidate solution by measuring the difference between the solution and the desired output. Solving by optimization means minimizing cost.

**Gradient descent:** An algorithm that minimizes cost by incrementally changing the parameters in the direction of steepest descent of the cost function.

---

<sup>1</sup> Admittedly, these networks still demand tweaking of a few parameters, including number of layers and number of units per layer.

66 tactile “sensory circles”). This once required single-cell recording, looking for minutes or hours at  
67 how one cell responds to each of perhaps a hundred different stimuli. Today it is clear that  
68 characterization of a single neuron’s receptive  
69 field, which was invaluable in the retina and V1, fails to  
70 characterize how higher visual areas encode the  
71 stimulus. Machine learning techniques reveal “how  
72 neuronal responses can best be used (combined) to  
73 inform perceptual decision-making” (Graf, Kohn,  
74 Jazayeri, & Movshon, 2010). The simplicity of the  
75 machine decoding can be a virtue as it allows us to  
76 discover what can be easily read-out (e.g. by a single  
77 downstream neuron) (Hung et al. 2005). Achieving  
78 psychophysical levels of performance in decoding a  
79 stimulus object’s identity and location from the neural  
80 response shows that the measured neural performance  
81 has all the information needed for the subject to do the  
82 task (Majaj et al. 2015; Hong et al. 2016).

83 For psychophysics, Signal Detection Theory (SDT)  
84 proved that the optimal classifier for a known signal in  
85 white noise is a template matcher (Peterson, Birdsall, &  
86 Fox, 1954; Tanner & Birdsall, 1958). Of course, SDT  
87 solves only a simple version of the general problem of  
88 object recognition. The simple version is for known  
89 signals, whereas the general problem includes variation

**Convexity:** A real-valued function is called “convex” if the line segment between any two points on the graph of the function lies on or above the graph (Boyd & Vandenberghe, 2004). A problem is convex if its cost function is convex. Convexity guarantees that gradient descent will always find the global minimum.

**Generalization** is how well a classifier performs on new, unseen examples that it did not see during training.

**Cross validation** assesses the ability of the network to generalize, from the data that it trained on, to new data.

**Backprop**, short for “backward propagation of errors”, is widely used to apply gradient-descent learning to multi-layer networks. It uses the chain rule from calculus to iteratively compute the gradient of the cost function for each layer.

**Hebbian learning** and spike-timing-dependent plasticity (**STDP**). According to Hebb’s rule, the efficiency of a synapse increases after correlated pre- and post-synaptic activity. In other words, neurons that fire together, wire together (Löwel & Singer, 1992).

90 in viewing conditions and diverse objects within a category (e.g. a chair can be any object that  
91 affords sitting). SDT introduces the very useful idea of a mathematically defined ideal observer,  
92 providing a reference for human performance (e.g. Geisler, 1989; Pelli et al., 2006). However,  
93 one drawback is that it

94 doesn't incorporate learning. Deep learning, on the other  
95 hand, provides a pretty good observer that learns, which  
96 may inform studies of human learning.<sup>2</sup>

***Support Vector Machine (SVM)***

is a type of machine learning algorithm for classification. SVMs generalize well. An SVM uses the “kernel trick” to quickly learn to perform a nonlinear classification by finding a boundary in multidimensional space that separates different classes and maximizes the distance of class exemplars to the boundary (Cortes & Vapnik, 1999).

***Convolutional neural networks (ConvNets)***

have their roots in the Neocognitron (Fukushima 1980) and are inspired by the simple and complex cells described by Hubel and Wiesel (1962). ConvNets apply backprop learning to multilayer neural networks based on convolution and pooling (LeCun et al., 1989; LeCun et al., 1990; LeCun et al., 1998).

97 These networks might reveal the constraints imposed by  
98 the training set on learning. Further, unlike SDT, deep  
99 neural networks cope with the complexity of real tasks. It  
100 can be hard to tell whether behavioral performance is  
101 limited by the set of stimuli, their neural representation,  
102 or the observer's decision process (Majaj et al. 2015).  
103 Implications for classification performance are not readily  
104 apparent from direct inspection of families of stimuli and  
105 their neural responses. SDT specifies optimal  
106 performance for classification of known signals but does  
107 not tell us how to generalize beyond a training set.  
108 Machine learning does.

109 **MINUSES: COMMON COMPLAINTS**

---

<sup>2</sup> In the same spirit, “sequential ideal observer” and “accuracy maximization” modeling generalized ideal observer calculations to include a shallow form of supervised learning (Geisler, 1989; Burge & Jains, 2017).

110 Some biologists point out that neural nets do not match what we know about neurons (e.g.,  
111 Crick, 1989; Rubinov, 2015). Biological brains learn on the job, while neural networks need to  
112 converge before they can be used. Furthermore, once trained, deep networks generally  
113 compute in a feed-forward manner while there are major recurrent circuits in the cortex. But this  
114 may simply reflect the different ways that we use artificial and real neurons. The artificial  
115 networks are trained for a fixed task, whereas our visual brain must cope with a changing  
116 environment and task demands, so it never outgrows the need for the capacity to learn.

117 It is not clear, given what we know about neurons and neural plasticity, whether a backprop  
118 network can be implemented using biologically plausible circuits (but see Mazzone et al., 1991,  
119 and Bengio et al., 2015). However, there are several promising efforts to implement more  
120 biological plausible learning rules, e.g. spike-timing-dependent plasticity (Mazzone et al., 1991;  
121 Bengio et al., 2015; Sacramento, Costa, Bengio, & Senn, 2017).

122 Engineers and computer scientists, while inspired by biology, focus on developing machine  
123 learning tools that solve practical problems. Thus, models based on these tools often do not  
124 incorporate known constraints imposed by biological measurements. To this, one might counter  
125 that every biological model is an abstraction and can be useful even while failing to capture all  
126 the details of the living organism.

127 Some biological modelers complain that neural nets have alarmingly many parameters. Deep  
128 neural networks continue to be opaque. Before neural-network modeling, a model was simpler  
129 than the data it explained. Deep neural nets are typically as complex as the data, and the  
130 solutions are hard to visualize (but see Zeiler & Fergus, 2013). However, while the training sets  
131 and learned weights are long lists, the generative rules for the network (the computer programs)  
132 are short. Traditionally, having very many parameters has often led to overfitting, i.e. good  
133 performance on the training set and poor performance beyond it, but the breakthrough is that

134 deep-learning networks with a huge number of parameters nevertheless generalize well.

135 Furthermore, Bayesian nonparametric models offer a disciplined approach to modeling with an  
136 unlimited number of parameters (Gershman & Blei, 2011).

137 Some statisticians worry that rigorous statistical tools are being displaced by deep learning,  
138 which lacks rigor (Friedman, 1998; Matloff, 2014, but see Breiman, 2001; Efron & Hastie, 2016).  
139 Assumptions are rarely stated. There are no confidence intervals on the solution. However,  
140 performance is typically cross-validated, showing generalization. Deep learning is not convex,  
141 but it has been proven that convex networks can compute posterior probability (e.g. Rojas,  
142 1996). Furthermore, machine learning, and statistics seem to be converging to provide a more  
143 general perspective on probabilistic inference that combines complexity and rigor.

144 Some physiologists note that decoding neural activity to recover the stimulus is interesting and  
145 useful but falls short of explaining what the neurons do. Some visual psychophysicists note  
146 some salient differences between performance of human observers and deep networks on  
147 tasks like object recognition and image distortion (Ullman et al. 2016; Berardino et al. 2017).

148 Some cognitive psychologists dismiss deep neural networks as unable to “master some of the  
149 basic things that children do, like learning the past tense of a regular verb” (Marcus et al., 1992).

150 Deep learning is slow. To recognize objects in natural images with the recognition accuracy of  
151 an adult, a state-of-the-art deep neural network needs five thousand labelled examples per  
152 category (Goodfellow et al., 2016). But children and adults need only a hundred labelled letters  
153 of an unfamiliar alphabet to reach the same accuracy as fluent native readers (Pelli et al. 2006).

154 Overcoming these challenges may require more than deep learning.

155 These current limitations drive practitioners to enhance the scope and rigor of deep learning.

156 But bear in mind that some of the best classifiers in computer science were inspired by

157 biological principles (Rosenblatt, 1957; 1958; Rumelhart et al., 1986; LeCun, 1985; LeCun et al.



158 1989; LeCun et al. 1990; Riesenhuber & Poggio, 1999; and  
159 see LeCun, Bengio, Hinton 2015). Some of those  
160 classifiers are now so good that they occasionally exceed  
161 human performance and might serve as rough models for  
162 how biological systems classify (e.g. Yamins, et al. 2014;  
163 Khaligh-Razavi & Kriegeskorte, 2014; Ziskind, Hénaff,  
164 LeCun, & Pelli, 2014; Testolin, Stoianov, & Zorzi, 2017).

## 165 **MILESTONES IN CLASSIFICATION**

166 **Mathematics vs. engineering.** The history of machine

167 learning has two threads: mathematics and engineering. In

168 the *mathematical* thread, two statisticians, Fisher and later Vapnik, developed mathematical

169 transformations to uncover categories in data, and proved that they give unique answers. They

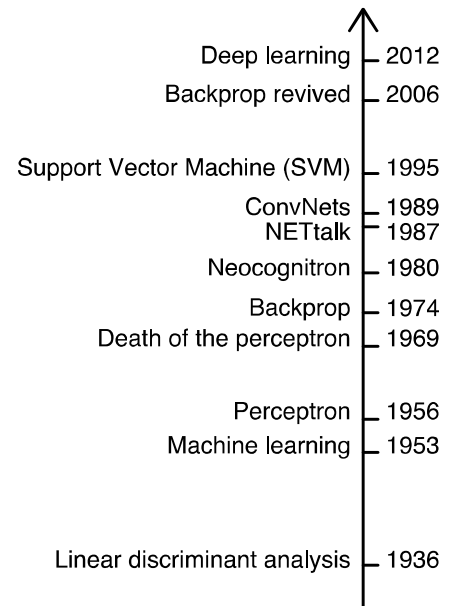
170 assumed distributions and proved convergence.

171 In the *engineering* thread, a loose coalition of psychologists, neuroscientists, and computer

172 scientists (e.g. Turing, Rosenblatt, Minsky, Fukushima, Hinton, Sejnowski, LeCun, Poggio,

173 Bengio) sought to reverse-engineer the brain to build a machine that learns. Their algorithms

174 are typically applied to stimuli with unknown distributions and lack proofs of convergence.



175 **1936: Linear discriminant analysis.** Fisher (1936)

**Figure 2.** Milestones in classification.

176 introduced linear discriminant analysis to classify two

177 species of iris flower based on four measurements per flower. When the distribution of the

178 measurements is normal and the covariance matrix between the measurements is known, linear

179 discriminant analysis answers the question: Supposing we use a single-valued function to

180 classify, what linear function  $y = w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4$ , of four measurements  $x_1, x_2, x_3, x_4$

181 made on flowers, with free weights  $w_1, w_2, w_3, w_4$ , will maximize discrimination of species?<sup>3</sup>

182 Linear classifiers are great for simple problems for which the category boundary is a hyperplane

183 in a small number of dimensions. However, complex problems like object recognition typically

184 require more complex category boundaries in a large number of dimensions. Furthermore, the

185 distributions of the features are typically unknown and may not be normal.

186 Cortes & Vapnik (1995) note that the first algorithm for pattern recognition was Fisher's optimal

187 decision function for classifying vectors from two known distributions. Fisher solved for the

188 optimal classifier in the presence of gaussian noise and known covariance between elements of

189 the vector. When the covariances are equal, this reduces to a linear classifier. The ideal

190 template matcher of signal detection theory is an example of such a linear classifier (Peterson et

191 al., 1954). This fully specified simple problem can be solved analytically. Of course, many

192 important problems are not fully specified. In everyday perceptual tasks, we typically know only

193 a "training" set of samples and labels.

194 **1953: Machine learning.** The first developments in machine learning were to play chess and

195 checkers. "Could one make a machine to play chess, and to improve its play, game by game,

196 profiting from its experience?" (Turing, 1953). Arthur Samuel (1959) defined *machine learning*

---

<sup>3</sup> Linear discriminant analysis is an outgrowth of regression which has a much longer history. Regression is the optimal least-squares linear combination of given functions to fit given data and was applied by Legendre (1805) and Gauss (1809) to astronomical data to determine the orbits of the comets and planets around the sun. The estimates come with confidence intervals and the fraction of variance accounted for, which rates the goodness of the explanation.

197 as the “Field of study that gives computers the ability to learn without being explicitly  
198 programmed.”

199 **1958: Perceptron.** Inspired by physiologically measured receptive fields, Rosenblatt (1958)  
200 showed that a very simple neural network, the perceptron, could learn to classify from training  
201 samples. Perceptrons combined several linear classifiers to implement piecewise-linear  
202 separating surfaces. The perceptron learns the weights to use in a linear combination of feature-  
203 detector outputs. The perceptron transforms the stimulus into a binary feature vector and then  
204 applies a linear classifier to the feature vector. The perceptron is piecewise linear and has the  
205 ability to learn from training examples without knowing the full distribution of the stimuli. Only the  
206 final layer in the perceptron learns.

207 **1969: Death of the perceptron.** However, it quickly became apparent that the perceptron and  
208 other single-layer neural networks cannot learn tasks that are not linearly separable, i.e. cannot  
209 solve problems like connectivity (Are all elements connected?) and parity (Is the number of  
210 elements odd or even?); people solve these readily (Minsky & Papert, 1969). On this basis,  
211 Minsky and Papert announced the death of artificial neural networks.

212 **1974: Backprop.** The death of the perceptron showed that learning in a one-layer network was  
213 too limited. This impasse was broken by the introduction of the backprop algorithm, which  
214 allowed learning to propagate through multiple-layer neural networks. The history of backprop is  
215 complicated (see Schmidhuber, 2015). The idea of minimization of error through a differentiable  
216 multi-stage network was discussed as early as the 1960s (e.g. Bryson, Denham, & Dreyfus,  
217 1963). It was applied to artificial neural networks in the 1970s (e.g. Werbos, 1974). In the 1980s,  
218 efficient backprop first gained recognition, and led to a renaissance in the field of artificial neural  
219 network research (LeCun, 1985; Rumelhart, Hinton, & Williams, 1986). During the 2000s  
220 backprop neural networks fell out of favor, due to four limitations (Vapnik, 1999): **1. No proof of**

221 **convergence.** Backprop uses gradient descent. Gradient descent with a nonconvex cost  
222 function with multiple minima is only guaranteed to find a local, not the global minimum of the  
223 cost function. This has long been considered a major limitation, but LeCun et al. (2015) claim  
224 that it hardly matters in practice in current implementations of deep learning. **2. Slow.**  
225 Convergence to a local minimum can be slow due to the high dimensionality of the weight  
226 space. **3. Poorly specified.** Backprop neural networks had a reputation for being ill-specified,  
227 an unconstrained number of units and training examples, and a step size that varied by  
228 problem. “Neural networks came to be painted as slow and fussy to train [,] beset by voodoo  
229 parameters and simply inferior to other approaches.” (Cox & Dean, 2014). **4. Not biological.**  
230 Lastly, backprop learning may not to be physiological: While there is ample evidence for  
231 Hebbian learning (increase of a synapse’s gain in response to correlated activity of the two cells  
232 that it connects), such changes are never propagated backwards, beyond the one synapse, to a  
233 previous layer. **5. Inadequate resources.** With hindsight it is clear that backprop in the 80’s was  
234 crippled by limited computing power and lack of large labeled datasets.

235 **1980: Neocognitron,** the first convolutional neural network. Fukushima (1980) proposed and  
236 implemented the Neocognitron, a hierarchical, multilayer artificial neural network. It recognized  
237 stimulus patterns (deformed numbers) despite small changes in position and shape.

238 **1987: NETtalk,** the first impressive backprop neural network. Sejnowski et al. (1987) reported  
239 the exciting success of NETtalk, a neural network that learned to convert English text to speech:  
240 “*The performance of NETtalk has some similarities with observed human performance. (i) The*  
241 *learning follows a power law. (ii) The more words the network learns, the better it is at*  
242 *generalizing and correctly pronouncing new words. (iii) The performance of the networks*  
243 *degrades very slowly as connections in the network are damaged: no single link or processing*  
244 *unit is essential. (iv) Relearning after damage is much faster than learning during the original*  
245 *training...*”

246 **1989: ConvNets.** Yann LeCun and his colleagues combined convolutional neural networks with  
247 backprop to recognize handwritten characters (LeCun et al., 1989; LeCun et al., 1990). This  
248 network was commercially deployed by AT&T, and today reads millions of checks a day  
249 (LeCun, 1998). Later, adding half-wave rectification and max pooling greatly improved its  
250 accuracy in recognizing objects (Jarrett et al., 2009).

251 **1995: Support Vector Machine (SVM).** Cortes & Vapnik (1995) proposed the support vector  
252 network, a learning machine for binary classification problems. SVMs generalize well and are  
253 free of mysterious training parameters. Many versions of the SVM are convex (e.g. Lin, 2001).

254 **2006: Backprop revived.** Hinton & Salakhutdinov (2006) sped up backprop learning by  
255 unsupervised pre-training. This helped to revive interest in backprop. In the same year, a  
256 supervised backprop-trained convolutional neural network set a new record on the famous  
257 MNIST handwritten-digit recognition benchmark (Ranzato et al., 2006).

258 **2012: Deep learning.** Geoff Hinton says, “It took 17 years to get deep learning right; one year  
259 thinking and 16 years of progress in computing, praise be to Intel.” (Cox & Dean, 2014; LeCun,  
260 Bengio, & Hinton, 2015). It is not clear who coined the term “deep learning”.<sup>4</sup> In their book, *Deep*  
261 *Learning Methods and Applications*, Deng & Yu (2014) cite Hinton et al. (2006) and Bengio  
262 (2009) as the first to use the term. However, the big debut for deep learning was an influential  
263 paper by Krizhevsky et al. (2012) describing AlexNet, a deep convolutional neural network that  
264 classified 1.2 million high-resolution images into 1000 different classes, greatly outperforming  
265 previous state-of-the-art machine learning and classification algorithms.

---

<sup>4</sup> The idea of “deep learning” is not exclusive to machine learning and neural networks (e.g. Dechter, 1986)

## 266 **CONTROVERSIES**

267 The field is growing quickly, yet certain topics remain hot. For proponents of deep learning, the  
268 ideal network is composed of simple elements and learns everything from the training data. At  
269 the other extreme, computer vision scientists argue that we know a lot about how the brain  
270 recognizes objects, which we can engineer into the networks before learning (e.g. gain control  
271 and normalization). Some engineers look to the brain only to copy strengths of the biological  
272 solution, others think there are useful clues in its limitations as well (e.g. crowding).

273 **Is deep learning the best solution for all visual tasks?** Deep learning is not the only thing in  
274 the vision scientist's toolbox. The complexity of deep learning may be unwarranted for simple  
275 problems that are well handled by, e.g. SVM. Try shallow networks first, and, if they fail, go  
276 deep.

277 **Why object recognition?** The visual task of object recognition as has been very useful in  
278 vision research because it is an objective task that is easily scored as right or wrong, is  
279 essential in daily life, and captures some of the magic of seeing. It is a classic problem with a  
280 rich literature. Deep neural nets solve it, albeit with a million parameters. Recognizing objects is  
281 a basic life skill, including recognition of words, people, things, and emotions. The concern that  
282 the research focus on object recognition might be merely an obsession of the scientists rather  
283 than a central task of biological vision is countered by hints that visual perception is biased to  
284 interpret the world as consisting of discrete objects even when it isn't, e.g. when we see animals  
285 in the clouds.

286 Of course, there are many other important visual tasks, including interpolation (e.g. filling in) and  
287 extrapolation (e.g. estimating heading). The inverse of categorization is synthesis. Human  
288 estimation of one feature, e.g. of brightness or speed, is imprecise and adequately represented  
289 by roughly 7 categories (Miller, 1956). For detection of image distortion, a simple model with

290 gain-control normalization is better than current deep networks (Berardino et al. 2017).

291 Scientists, like the brain, use whatever tool works best.

292 **Deep learning is not convex.** A problem is convex if the cost function is convex, i.e. if the line  
293 between any two points on the function lies on or above the function. This guarantees that  
294 gradient descent will find the global minimum. For some combinations of stimuli, categories, and  
295 classifiers, convexity can be proved. In machine learning, kernel methods, including learning by  
296 SVMs, have the advantage of easy-to-prove convexity, at the cost of limited generalization. In  
297 the 1990s, SVMs were popular because they guaranteed fast convergence even with a large  
298 number of training samples (Cortes & Vapnik, 1995). Thus, when the problem is convex, the  
299 quality of solution is assured, and one can rate implementations by their demands for size of  
300 network and training sample. However, cost functions for deep neural networks are not convex.  
301 Unlike convex functions, nonconvex functions can have multiple minima and saddle points. The  
302 challenge in high dimensional cost functions is the saddle points, which greatly outnumber the  
303 local minima, but there are tricks for not getting stuck at saddle points (Dauphin et al. 2014).  
304 Although deep neural networks are not convex, they do fit the training data, and generalize well  
305 (LeCun, Bengio, & Hinton, 2015).

306 **Shallow vs. deep networks.** The field's imagination has focused alternately on shallow and  
307 deep networks, beginning with the Perceptron in which only one layer learned, followed by  
308 backprop, which allowed multiple layers and cleared the hurdles that doomed the Perceptron.  
309 Then SVM, with its single layer, sidelined the multilayer backprop. Today multilayer deep  
310 learning reigns; Krizhevsky et al. (2012) attributed the success of AlexNet to its 8-layer depth; it  
311 performed worse with fewer layers. Some people claim that deep learning is essential to  
312 recognize objects in real world scenes. For example, the "Inception" 22-layer deep learning  
313 network won the Image Net Real World Challenge in 2014 (Szegedy et al. 2015).

314 The need for depth is hard to prove, but, in considering the depth vs. width of a feed-forward  
315 neural network, Eldan and Shamir (2016) show that a radial function can be approximated by a  
316 3-layer network with far fewer neurons than the best 2-layer network (also see Telgasky, 2015).  
317 Object recognition implies a classification function that assigns one of several discrete values to  
318 each image. Mhaskr et al. (2017) suggest that for real-world recognition the classification  
319 function is typically compositional, i.e. a hierarchy of functions, one per node, in feed-forward  
320 layers, in which the receptive fields of higher layers are ever larger. They argue that scalability  
321 and shift invariance in natural images require compositional algorithms. They prove that deep  
322 hierarchical networks can approximate compositional functions with the same accuracy as  
323 shallow networks but with exponentially fewer training parameters.

324 **Supervised vs. unsupervised.** Learning algorithms for a classifier can be supervised or not,  
325 i.e. need labels for training, or don't. Today most machine learning is *supervised* (LeCun,  
326 Bengio, & Hinton, 2015). The images are labeled (e.g. "car" or "face"), or the network receives  
327 feedback on each trial from a cost function that assesses how well its answer matches the  
328 image's category. In *unsupervised* learning, no labels are given. The algorithm processes  
329 images, typically to minimize error in reconstruction, with no extra information about what is in  
330 the (unlabeled) image. A cost function can also reward decorrelation and sparseness (e.g.  
331 Olshausen and Field, 1996). This allows learning of image statistics and has been used to train  
332 early layers in deep neural networks. Human learning of categorization is sometimes done with  
333 explicitly named objects — "Look at the tree!" — but more commonly the feedback is implicit.  
334 Consider reaching your hand to raise a glass of water. Contact informs vision. On specific  
335 benchmarks, where the task is well-defined and labeled examples are available, supervised  
336 learning can excel (e.g. AlexNet), but unsupervised learning may be more useful when few  
337 labels are available.



## 338 **CURRENT DIRECTIONS**

339 **What does deep learning add to the vision-science toolbox?** Deep learning is more than  
340 just a souped-up regression (Marblestone et al., 2016). Like Signal Detection Theory (SDT), it  
341 allows us to see more in our behavioral and neural data. In the 1940's, Norbert Wiener and  
342 others developed algorithms to automate and optimize signal detection and classification. A lot  
343 of it was engineering. The whole picture changed with the SDT theorems, mainly the proof that  
344 the maximum-likelihood receiver is optimal for a wide range of simple tasks (Peterson et al.,  
345 1954). In white noise a traditional receptive field computes the likelihood of the presence of a  
346 signal matching the receptive field weights. It was exciting to realize that the brain contains  $10^{11}$   
347 likelihood computers. Later work added prior probability, for a Bayesian approach. Tanner &  
348 Birdsall (1958) noted that, when figuring out how a biological system does a task, it is very  
349 helpful to know the optimal algorithm and to rate observed performance by its *efficiency* relative  
350 to the optimum. SDT solved detection and classification mathematically, as maximum likelihood.  
351 It was the classification math of the sixties. Machine learning is the classification math of today.  
352 Both enable deeper insight into how biological systems classify. Of course, as noted above,  
353 SDT is restricted to the case of known signals in additive noise, whereas deep learning can  
354 solve real world object recognition like detecting a dog in a photo after training on labeled  
355 examples. In the old days we used to compare human and ideal classification performance  
356 (Pelli et al. 2006). Today, we also compare human and machine learning. Deep learning is the  
357 best model we have today for how complex systems of simple units can recognize objects as  
358 well as the brain does. Deep learning, i.e. learning by multi-layered neural networks using  
359 backprop, is not just AlexNet but also includes ConvNets and other architectures of trained  
360 artificial neural networks. Several labs are currently comparing patterns of activity of particular  
361 layers to neural responses in various cortical areas of the mammalian visual brain (Yamins et al.  
362 2014; Khaligh-Razavi & Kriegeskorte, 2014).

363 **What computer scientists can learn from psychophysics.** Computer scientists build  
364 classifiers to recognize objects. Vision scientists, including psychologists and neuroscientists,  
365 study how people and animals classify in order to understand how the brain works. So, what do  
366 computer and vision scientists have to say to each other? Machine learning accepts a set of  
367 labelled stimuli to produce a classifier. Much progress has been made in physiology and  
368 psychophysics by characterizing how well biological systems can classify stimuli. The  
369 psychophysical tools (e.g. threshold and signal detection theory) developed to characterize  
370 behavioral classification performance are immediately applicable to characterize classifiers  
371 produced by machine learning (e.g. Ziskind, Hénaff, LeCun, & Pelli, 2014; Testolin, Stoianov, &  
372 Zorzi, 2017).

373 **Psychophysics.** “Adversarial” examples have been presented as a major flaw in deep neural  
374 networks (Mims, 2018; Hutson, 2018). These slightly doctored images of objects are  
375 misclassified by a trained network, even though the doctoring has little effect on human  
376 observers. The same doctored images are similarly misclassified by several different networks  
377 trained with the same stimuli (Szegedy, et al., 2013). Humans too have adversarial examples.  
378 Illusions are robust classification errors. The blindspot-filling-in illusion is a dramatic adversarial  
379 example in human vision. While viewing with one eye, two finger tips touching in the blindspot  
380 are perceived as one long finger. If the image is shifted a bit so that the fingertips emerge from  
381 the blindspot the viewer sees two fingers. Neural networks lacking the anatomical blindspot of  
382 human vision are hardly affected by the shift (but see Azulay & Weiss, 2018). The existence of  
383 adversarial examples is intrinsic to classifiers trained with finite data, whether biological or not.  
384 In the absence of information, neural networks interpolate and so do biological brains.  
385 Psychophysics, the scientific study of perception, has achieved its greatest advances by  
386 studying classification errors (Fechner, 1860). Such errors can reveal “blindspots”. Stimuli that  
387 are physically different yet indistinguishable are called *metamers*. The systematic understanding

388 of color metamers revealed the three dimensions of human color vision (Palmer, 1777; Young,  
389 1802; Helmholtz, 1860). In recent work, many classifiers have been trained solely with the  
390 objects they are meant to classify, and thus will classify everything as one of those categories,  
391 even doctored noise that is very different from all of the images. It is important to train with  
392 sample images that represent the entire test set.

## 393 **CONCLUSION**

394 Machine learning is here to stay. Deep learning is better than the “neural” networks of the  
395 eighties. Machine learning is useful both as a model for perceptual processing, and as a  
396 decoder of neural processing, to see what information the neurons are carrying. The large size  
397 of the human cortex is a distinctive feature of our species and crucial for learning. It is  
398 anatomically homogenous yet solves diverse sensory, motor, and cognitive problems. Key  
399 biological details of cortical learning remain obscure, but, even if they ultimately preclude  
400 backprop, the performance of current machine learning algorithms is a useful benchmark.

## 401 **RESOURCES**

402 We recommend textbooks on deep learning by Goodfellow, Bengio, & Courville (2016) and Ng  
403 (2017). There are many packages for optimization and machine learning in MATLAB and  
404 Python.

405

## 406 **ACKNOWLEDGEMENTS**

407 Thanks to Yann LeCun for helpful conversations. Thanks to Aenne Brielmann, Augustin  
408 Burchell, Kaitie Holman, Katerina Malakhova, Cristina Savin, Laura Suciu, Pascal Wallisch, and  
409 Avi Ziskind for helpful comments on the manuscript. We thank our editor, David Brainard, and  
410 both reviewers, Nikolaus Kriegeskorte and anonymous, for many constructive suggestions.  
411 DGP was supported by NIH grant R01 EY027964.

412 **REFERENCES**

413

414 Azulay, A., & Weiss, Y. (2018). Why do deep convolutional networks generalize so poorly to  
415 small image transformations? *arXiv preprint arXiv:1805.12177*.

416 Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends in Machine*  
417 *Learning*, 2(1), 1–127.

418 Bengio, Y., Le, D. H., Bornschein, J., Mesnard, T., & Lin, Z. (2015). Towards biologically  
419 plausible deep learning. *arXiv preprint arXiv:1502.04156*.

420 Berardino, A., Laparra, V., Ballé, J., & Simoncelli, E. (2017). Eigen-Distortions of Hierarchical  
421 Representations. In *Advances in Neural Information Processing Systems* (pp. 3533-3542).

422

423 Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

424 Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by  
425 the author). *Statistical science*, 16(3), 199-231.

426 Brynjolfsson, E. (2018). Machine learning will be the engine of global growth. *Financial Times*.  
427 July 26, 2018.

428 Bryson, A. E., Denham, W. F., & Dreyfus, S. E. (1963). Optimal programming problems with  
429 inequality constraints. *AIAA journal*, 1(11), 2544-2550.

430 Burge, J., & Jains, P. (2017). Accuracy maximization analysis for sensory-perceptual tasks:  
431 Computational improvements, filter robustness, and coding advantages for scaled additive  
432 noise. *PLoS computational biology*, 13(2), e1005281.

433 Cadieu, C. F., Hong, H., Yamins, D. L., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J. &  
434 DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate IT cortex for core  
435 visual object recognition. *PLoS computational biology*, 10(12), e1003963.

436 Caporale, N., & Dan, Y. (2008). Spike timing-dependent plasticity: a Hebbian learning rule.  
437 *Annu. Rev. Neurosci.*, 31, 25-46.

438 Cox, D. D., & Dean, T. (2014). Neural networks and neuroscience-inspired computer vision.  
439 *Current Biology*, 24(18), R921-R929.

440 Crick, F. (1989). The recent excitement about neural networks. *Nature*, 337(6203), 129.

441

442 Dauphin, Y. N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., & Bengio, Y. (2014). Identifying  
443 and attacking the saddle point problem in high-dimensional non-convex optimization.  
444 In *Advances in neural information processing systems* (pp. 2933-2941).

445 Dechter, R. (1986). Learning while searching in constraint-satisfaction-problems. In *Proceedings*  
446 *of the Fifth AAAI National Conference on Artificial Intelligence* (pp. 178-183). AAAI Press.

- 447 Deng, L., & Yu, D. (2014). Deep learning: methods and applications. *Foundations and Trends®*  
448 *in Signal Processing*, 7(3–4), 197-387.
- 449 Efron, B., & Hastie, T. (2016). *Computer Age Statistical Inference: Algorithms, Evidence, and*  
450 *Data Science* (Vol. 5). Cambridge University Press.
- 451 Eldan, R., & Shamir, O. (2016). The power of depth for feedforward neural networks.  
452 In *Conference on Learning Theory* (pp. 907-940).
- 453 Fechner, GT (1860). Elements of psychophysics. Breitkopf & Härtel, Leipzig (reprinted in 1964  
454 by Bonset, Amsterdam); German translation by HE Adler (1966): Elements of psychophysics.
- 455 Fisher, R. A. (1922). The goodness of fit of regression formulae, and the distribution of  
456 regression coefficients. *Journal of the Royal Statistical Society*, 85(4), 597-612.  
457 doi:10.2307/2341124.
- 458 Friedman, J. H. (1998). Data mining and statistics: What's the connection? *Computing Science*  
459 *and Statistics*, 29(1), 3-9
- 460 Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism  
461 of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4), 193-202.
- 462 Galton, F. (1877). Typical laws of heredity. *Nature*, 15(389), 512-514.
- 463 Gauss, C.F. (1809). *Theoria motus corporum coelestium in sectionibus conicis solem*  
464 *ambientium*. Hamburg: Friedrich Perthes und I. H. Besser.
- 465 Geisler, W. S. (1989). Sequential ideal-observer analysis of visual discriminations.  
466 *Psychological review*, 96(2), 267.
- 467 Gershman, S. J., & Blei, D. M., (2011) A tutorial on Bayesian nonparametric models.  
468 [arXiv:1106.2697](https://arxiv.org/abs/1106.2697) [stat.ML]
- 469 Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge: MIT Press.  
470 <http://www.deeplearningbook.org>
- 471 Helmholtz, H. von (1860/1925). *Handbuch der physiologischen Optik*, volume II. Leopold Voss,  
472 Leipzig, third edition. Translated as *Treatise on Physiological Optics*, volume II. The Optical  
473 Society of America, 1925. Edited by James P. C. Southall.
- 474 Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets.  
475 *Neural computation*, 18(7), 1527-1554.
- 476 Hong, H., Yamins, D. L., Majaj, N. J., & DiCarlo, J. J. (2016). Explicit information for category-  
477 orthogonal object properties increases along the ventral stream. *Nature neuroscience*, 19(4),  
478 613.
- 479 Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional  
480 architecture in the cat's visual cortex. *The Journal of physiology*, 160(1), 106-154.

- 481 Hung, C. P., Kreiman, G., Poggio, T., & DiCarlo, J. J. (2005). Fast readout of object identity from  
482 macaque inferior temporal cortex. *Science*, 310(5749), 863-866.
- 483 Hutson, M. (2018). Hackers easily fool artificial intelligences. *Science*, 361, 6399, 215.
- 484 Ivakhnenko, A. G. (1971). Polynomial theory of complex systems. *IEEE Transactions on*  
485 *Systems, Man and Cybernetics*, 4, 364-378.
- 486 Ivakhnenko, A. G. & Lapa, V. G. (1965). *Cybernetic Predicting Devices*. CCM Information  
487 Corporation.
- 488 Jarrett, K., Kavukcuoglu, K., & LeCun, Y. (2009). What is the best multi-stage architecture for  
489 object recognition? In *IEEE 12th International Conference on Computer Vision*. pp. 2146-2153.
- 490 Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised,  
491 models may explain IT cortical representation. *PLoS computational biology*, 10(11), e1003915.
- 492 Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep  
493 convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097-  
494 1105.
- 495 LeCun, Y. (1985). Une procedure d'apprentissage pour reseau a seuil asymmetrique (A  
496 learning scheme for asymmetric threshold networks). In *Proceedings of Cognitiva 85*, Paris,  
497 France. pp. 599-604.
- 498 LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D.  
499 (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4),  
500 541-551.
- 501 LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., & Jackel,  
502 L. D. (1990). Handwritten digit recognition with a back-propagation network. In *Advances in*  
503 *neural information processing systems* (pp. 396-404).
- 504 LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to  
505 document recognition. *Proceedings of the IEEE*, 86(11), 2278-2324.
- 506 LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.
- 507 Legendre, A. M. (1805). *Nouvelles méthodes pour la détermination des orbites des comètes*  
508 (No. 1). F. Didot.
- 509 Lin, C. J. (2001). On the convergence of the decomposition method for support vector  
510 machines. *IEEE Transactions on Neural Networks*, 12(6), 1288-1298.
- 511 Lowel, S., & Singer, W. (1992). Selection of intrinsic horizontal connections in the visual cortex  
512 by correlated neuronal activity. *Science*, 255(5041), 209.

- 513 Majaj, N. J., Hong, H., Solomon, E. A., & DiCarlo, J. J. (2015). Simple learned weighted sums of  
514 inferior temporal neuronal firing rates accurately predict human core object recognition  
515 performance. *Journal of Neuroscience*, 35(39), 13402-13418.
- 516 Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., Xu, F., & Clahsen, H. (1992).  
517 Overregularization in language acquisition. *Monographs of the society for research in child*  
518 *development*, i-178.
- 519 Marr, D. (1982). *Vision: A computational investigation into the human representation and*  
520 *processing of visual information*. San Francisco, CA: Freeman and Company.
- 521 Matloff, N. (2014). Statistics: Losing Ground to CS, Losing Image Among Students. *Revolutions*.  
522 August 26, 2014. [http://blog.revolutionanalytics.com/2014/08/statistics-losing-ground-to-cs-losing-image-among-](http://blog.revolutionanalytics.com/2014/08/statistics-losing-ground-to-cs-losing-image-among-students.html)  
523 [students.html](http://blog.revolutionanalytics.com/2014/08/statistics-losing-ground-to-cs-losing-image-among-students.html)
- 524 Marblestone, A. H., Wayne, G., & Kording, K. P. (2016). Toward an integration of deep learning  
525 and neuroscience. *Frontiers in computational neuroscience*, 10.
- 526 Mazzone, P., Andersen, R. A., & Jordan, M. I. (1991). A more biologically plausible learning rule  
527 for neural networks. *Proceedings of the National Academy of Sciences*, 88(10), 4433-4437.
- 528 Mhaskar, H., Liao, Q., & Poggio, T. A. (2017). When and why are deep networks better than  
529 shallow ones? In *AAAI* (pp. 2343-2349).
- 530 Miller, G. A. (1956). The magic number seven plus or minus two: Some limits on our capacity for  
531 processing information. *Psychological review*, 63, 91-97.
- 532 Mims, C. (2018). Should artificial intelligence copy the human brain? *The Wall Street Journal*,  
533 August 4, 2018. [https://www.wsj.com/articles/should-artificial-intelligence-copy-the-human-brain-](https://www.wsj.com/articles/should-artificial-intelligence-copy-the-human-brain-1533355265)  
534 [1533355265](https://www.wsj.com/articles/should-artificial-intelligence-copy-the-human-brain-1533355265)
- 535 Minsky, M., & Papert, S. (1988). *Perceptrons: An Introduction to Computational Geometry*.  
536 Cambridge, MA: MIT press.
- 537 Ng, A. (2017). *Machine Learning Yearning*. <http://www.mlyearning.org/>
- 538 Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by  
539 learning a sparse code for natural images. *Nature*, 381(6583), 607.
- 540 Palmer, G. (1777). *Theory of Colour and Vision*, London: Leacroft.
- 541 Pearson, K., Yule, G. U., Blanchard, N., & Lee, A. (1903). The law of ancestral heredity.  
542 *Biometrika*, 2(2), 211-236.
- 543 Pelli, D. G., Burns, C. W., Farell, B., & Moore-Page, D. C. (2006). Feature detection and letter  
544 identification. *Vision research*, 46(28), 4646-4674.
- 545 Peterson, W. W. T. G., Birdsall, T., & Fox, W. (1954). The theory of signal detectability.  
546 *Transactions of the IRE professional group on information theory*, 4(4), 171-212.

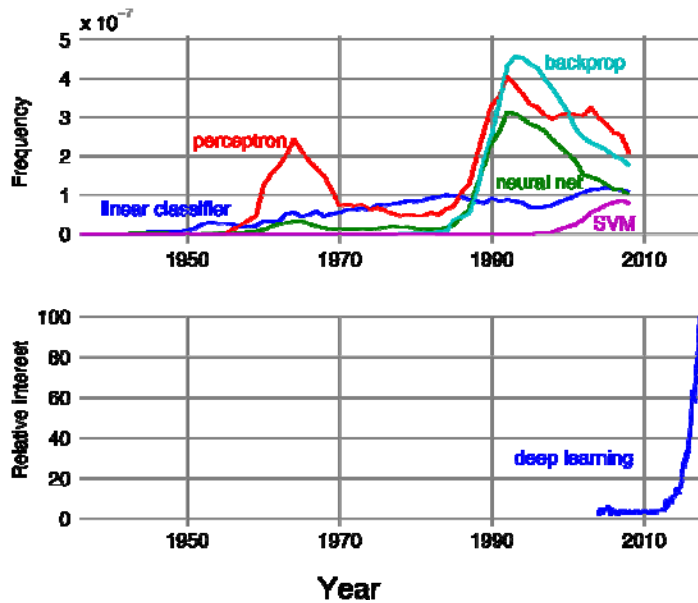
- 547 Ranzato, M. A., Huang, F. J., Boureau, Y. L., & LeCun, Y. (2007). Unsupervised learning of  
548 invariant feature hierarchies with applications to object recognition. In *2007 IEEE conference on*  
549 *computer vision and pattern recognition*, pp. 1-8.
- 550 Ranzato, M. A., Poultney, C., Chopra, S., & LeCun, Y. (2007). Efficient learning of sparse  
551 representations with an energy-based model. In *Proceedings of NIPS*.
- 552 Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in  
553 cortex. *Nature neuroscience*, 2(11), 1019-1025.
- 554 Rojas, R. (1996). A short proof of the posterior probability property of classifier neural networks.  
555 *Neural Computation*, 8(1), 41-43.
- 556 Rosenblatt, F. (1958), The Perceptron: A Probabilistic Model for Information Storage and  
557 Organization in the Brain, *Psychological Review*, 65, 6, pp. 386–408.
- 558 Rubinov, M. (2015). Neural networks in the future of neuroscience research. *Nature Reviews*  
559 *Neuroscience*.
- 560 Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-  
561 propagating errors. *Nature*, 323, 533-536. doi:10.1038/323533a0.
- 562 Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM*  
563 *Journal of research and development*, 3(3), 210-229.
- 564 Samuel, A. L. (1967). Some studies in machine learning using the game of checkers. II—recent  
565 progress. *IBM Journal of research and development*, 11(6), 601-617.
- 566 Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, 61,  
567 85-117.
- 568 Sejnowski, T. J., & Rosenberg, C. R. (1987). Parallel networks that learn to pronounce English  
569 text. *Complex systems*, 1(1), 145-168.
- 570 Seung, H. S., & Sompolinsky, H. (1993). Simple models for reading neuronal population codes.  
571 *Proceedings of the National Academy of Sciences*, 90(22), 10749-10753.
- 572 Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R.  
573 (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.
- 574 Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. &  
575 Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference*  
576 *on computer vision and pattern recognition* (pp. 1-9).
- 577 Tanner Jr, W. P., & Birdsall, T. G. (1958). Definitions of  $d'$  and  $\eta$  as psychophysical measures.  
578 *The Journal of the Acoustical society of America*, 30(10), 922-928.
- 579 Telgarsky, M. (2015). Representation benefits of deep feedforward networks. *arXiv preprint*  
580 *arXiv:1509.08101*.



- 581 Testolin, A., Stoianov, I., & Zorzi, M. (2017). Letter perception emerges from unsupervised deep  
582 learning and recycling of natural image features. *Nature Human Behaviour*, 1(9), 657.
- 583 Turing, A.M. (1953). 'Digital computers applied to games'. in 'Faster than thought', ed. B.V.  
584 Bowden, London 1953. Published by Pitman Publishing.
- 585 Ullman, S., Assif, L., Fetaya, E., & Harari, D. (2016). Atoms of recognition in human and  
586 computer vision. *Proceedings of the National Academy of Sciences*, 113(10), 2744-2749.
- 587 Vapnik, V. (2013). *The nature of statistical learning theory*. Springer Science & Business Media.
- 588 Weber, E. H. (1834/1996). *EH Weber on the tactile senses*. Psychology Press. Translated by  
589 Helen E. Ross from E.H Weber (1834) *De Tactu*.
- 590 Werbos, P. (1974). *Beyond regression: New tools for prediction and analysis in the behavioral*  
591 *sciences*. PhD thesis, Harvard University.
- 592 Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014).  
593 Performance-optimized hierarchical models predict neural responses in higher visual  
594 cortex. *Proceedings of the National Academy of Sciences*, 111(23), 8619-8624.
- 595 Young, T. (1802). The Bakerian Lecture. On the theory of light and colours, *Philosophical*  
596 *Transactions of the Royal Society of London* 92, 12-48. doi: 10.1098/rstl.1802.0004
- 597 Yule, G. U. (1897). On the theory of correlation. *Journal of the Royal Statistical Society*, 60(4),  
598 812-854.
- 599 Zeiler, M. D., & Fergus, R. (2013). Visualizing and understanding convolutional networks. arXiv  
600 preprint arXiv:1311.2901.
- 601 Ziskind, A.J., Hénaff, O., LeCun, Y., & Pelli, D.G. (2014) The bottleneck in human letter  
602 recognition: A computational model. Vision Sciences Society, St. Pete Beach, Florida, May 16-  
603 21, 2014, 56.583. <http://f1000.com/posters/browse/summary/1095738>  
604

605 FIGURES AND CAPTIONS

606



607

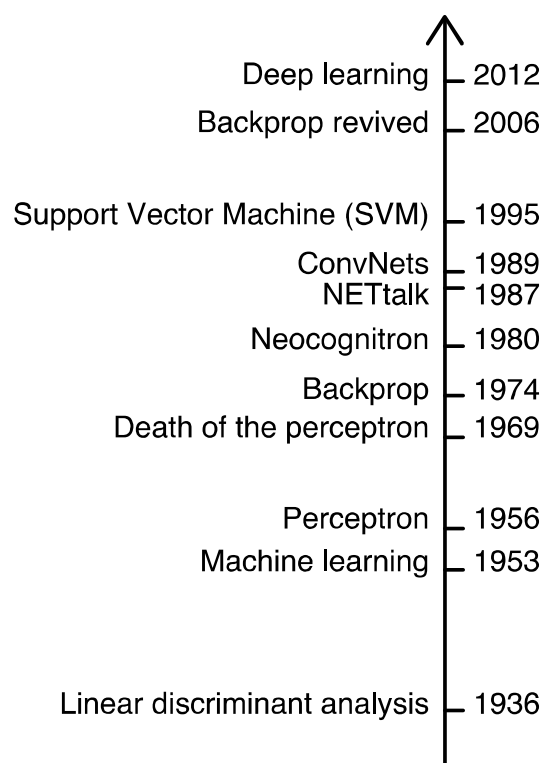
608

609 **Figure 1. Top:** The frequency of appearance of each of five terms — “linear classifier”, “perceptron”, “support vector  
610 machine”, “neural net” and “backprop” — in books indexed by Google in each year of publication. Google counts  
611 instances of words and phrases of  $n$  words, and calls each an “ngram”. Frequency is reported as a fraction of all  
612 instances of ngrams of that length, normalized by the number of books published that year  
613 (ngram / year / books published). The figure was created using Google’s ngram viewer  
614 (<https://books.google.com/ngrams>), which contains a yearly count of ngrams found in sources printed between 1500  
615 and 2008. **Bottom:** Numbers represent worldwide search interest relative to the highest point on the chart for the  
616 given year for the term “deep learning” (as reported by <https://trends.google.com/trends/>).

617

618

619



620  
621  
622

**Figure 2.** Milestones in classification.

