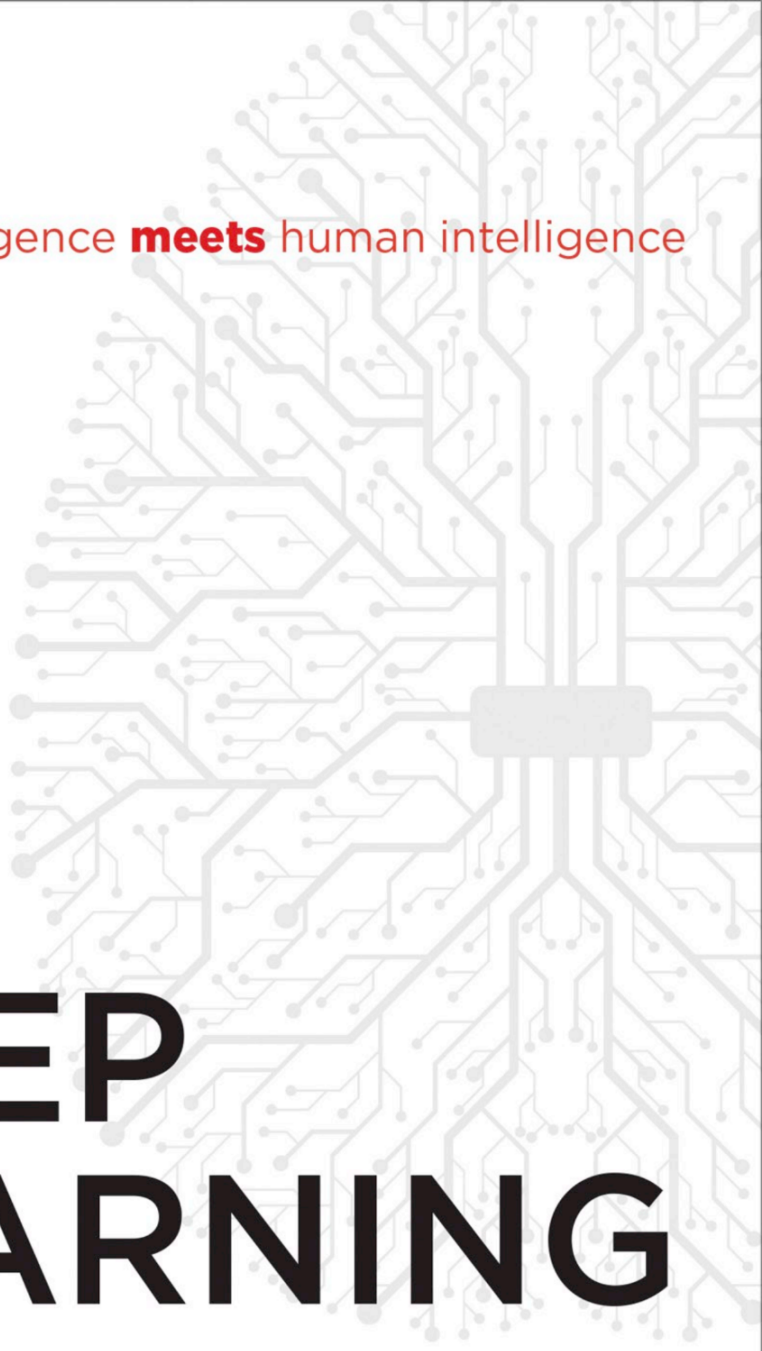


artificial intelligence **meets** human intelligence



THE  
**DEEP  
LEARNING**  
REVOLUTION

TERRENCE J. SEJNOWSKI

# The Deep Learning Revolution

Terrence J. Sejnowski

The MIT Press  
Cambridge, Massachusetts  
London, England

© 2018 Massachusetts Institute of Technology

All rights reserved. No part of this book may be reproduced in any form by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from the publisher.

This book was set in ITC Stone Sans Std and ITC Stone Serif Std by Toppan Best-set Premedia Limited. Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Names: Sejnowski, Terrence J. (Terrence Joseph), author.

Title: The deep learning revolution / Terrence J. Sejnowski.

Description: Cambridge, MA : The MIT Press, 2018. | Includes bibliographical references and index.

Identifiers: LCCN 2017044863 | ISBN 9780262038034 (hardcover : alk. paper)

Subjects: LCSH: Machine learning. | Big data. | Artificial intelligence--Social aspects.

Classification: LCC Q325.5 .S45 2018 | DDC 006.3/1--dc23 LC record available at <https://lcn.loc.gov/2017044863>

10 9 8 7 6 5 4 3

# Contents

Preface ix

## Part I: Intelligence Reimagined 1

- 1 The Rise of Machine Learning 3
- 2 The Rebirth of Artificial Intelligence 27
- 3 The Dawn of Neural Networks 37
- 4 Brain-style Computing 49
- 5 Insights from the Visual System 63

## Part II: Many Ways to Learn 79

- 6 The Cocktail Party Problem 81
- 7 The Hopfield Net and Boltzmann Machine 91
- 8 Backpropagating Errors 109
- 9 Convolutional Learning 127
- 10 Reward Learning 143
- 11 Neural Information Processing Systems 161

## Part III: Technological and Scientific Impact 169

- 12 The Future of Machine Learning 171
- 13 The Age of Algorithms 195
- 14 Hello, Mr. Chips 205
- 15 Inside Information 219
- 16 Consciousness 233
- 17 Nature Is Cleverer Than We Are 245
- 18 Deep Intelligence 261

Acknowledgments	269
Recommended Reading	275
Glossary	281
Notes	285
Index	321

## Preface

The recent progress in artificial intelligence (AI) was made by reverse engineering brains. Learning algorithms for layered neural network models are inspired by the way that neurons communicate with one another and are modified by experience. Inside the network, the complexity of the world is transformed into a kaleidoscope of internal patterns of activity that are the ingredients of intelligence. The network models that I worked on in the 1980s were tiny compared with today's models, which now have millions of artificial neurons and which are dozens of layers deep. What made it possible for deep learning to make big breakthroughs on some of the most difficult problems in artificial intelligence was persistence, big data, and a lot more computer power.

We're not good at imagining the impact of a new technology on the future. Who could have predicted in 1990, when the Internet went commercial, what impact it would have on the music business? On the taxi business? On political campaigns? On almost all aspects of our daily lives? There was a similar failure to imagine how computers would change our lives. Thomas J. Watson, the president of IBM, is widely quoted as saying in 1943: "I think there is a world market for maybe five computers."<sup>1</sup> What's hard to imagine are the uses to which a new invention will be put, and inventors are no better than anyone else at predicting what those uses will be. There is a lot of room between the utopian and doomsday scenarios that are being predicted for deep learning and AI, but even the most imaginative science fiction writers are unlikely to guess what their ultimate impact will be.

The first draft of *The Deep Learning Revolution* was written in a few focused weeks after hiking in the Pacific Northwest and meditating on the remarkable recent shift in the world of artificial intelligence, which had its origin many decades earlier. It is a story about a small group of researchers challenging an AI establishment that was much better funded and at the

time the “only game in town.” They vastly underestimated the difficulty of the problems and relied on intuitions about intelligence that proved to be misleading.

Life on earth is filled with many mysteries, but perhaps the most challenging of these is the nature of intelligence. Nature abounds with intelligence in many forms, from humble bacterial to complex human intelligence, each adapted to its niche in nature. Artificial intelligence will also come in many forms that will take their particular places on this spectrum. As machine intelligence based on deep neural networks matures, it could provide a new conceptual framework for biological intelligence.

*The Deep Learning Revolution* is a guide to the past, present, and future of deep learning. Not meant to be a comprehensive history of the field, it is rather a personal view of key conceptual advances and the community of researchers who made them. Human memory is fallible and shifts with every retelling of a story, a process called “reconsolidation.” The stories in this book stretch over forty years, and even though some are as vivid to me as if they occurred yesterday, I am well aware that the details have been edited by my memory’s retellings over time.

Part I provides the motivation for deep learning and the background needed to understand its origins; part II explains learning algorithms in several different types of neural network architectures; and part III explores the impact that deep learning is having on our lives and what impact it may have in years to come. But, as the New York Yankees’ philosopher Yogi Berra once said: “It’s tough to make predictions, especially about the future.” Text boxes in eight of the chapters to follow provide technical background to the story; timelines at the beginning of the three parts keep track of events that bear on that story and extend over sixty years.

The 2019 Turing Award recognized Geoffrey Hinton, Yann LeCun, and Yoshua Bengio for their pioneering work on deep learning. The Turing Award is the highest honor in computer science and is named after Alan Turing, who pioneered the fundamental theory underlying modern digital computers. Digital computers were invented after Nobel Prizes were established, so the Turing Award is widely considered to be like a Nobel Prize for computing. This book tells the story of the 30-year journey that led these pioneers to a revolution in artificial intelligence and the influence that their achievements are having on the world.

# I Intelligence Reimagined

## Timeline

**1956—The Dartmouth Artificial Intelligence Summer Research Project** gave birth to the field of AI and motivated a generation of scientists to explore the potential for information technology to match the capabilities of humans.

**1959—David Hubel and Torsten Wiesel** published “Receptive Fields, Binocular Interaction and Functional Architecture in the Cat’s Visual Cortex,” which reported for the first time the response properties of single neurons recorded with a microelectrode. Deep learning networks have an architecture similar to the hierarchy of areas in the visual cortex.

**1962—Frank Rosenblatt** published *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*, which introduced a learning algorithm for neural network models with a single layer of variable weights—the precursor of today’s learning algorithms for deep neural network models.

**1969—Marvin Minsky and Seymour Papert** published *Perceptrons*, which pointed out the computational limitations of a single artificial neuron and marked the beginning of a neural network winter.

**1979—Geoffrey Hinton and James Anderson** organized the Parallel Models of Associative Memory workshop in La Jolla, California, which brought together a new generation of neural network pioneers and led to publication of Hinton and Anderson’s collected volume by the same title in 1981.

**1987—The First Neural Information Processing Systems (NIPS) Conference and Workshop** was held at the Denver Tech Center, bringing together researchers from many fields.





Copyrighted image

### Figure 1.2

Beer Bottle Pass. This challenging terrain was near the end of the 2005 DARPA Grand Challenge for a vehicle to drive unassisted by a human through a 132-mile off-road desert course. A truck in the distance is just beginning the climb. Courtesy of DARPA.

products that its operating systems control, hoping to repeat its successful foray into the cell phone market. Seeing a business that had not changed for 100 years transformed before their eyes, automobile manufacturers are following in their tracks. General Motors paid \$1 billion for Cruise Automation, a Silicon Valley start-up that is developing driverless technology, and invested an additional \$600 million in 2017 in research and development.<sup>2</sup> In 2017, Intel purchased Mobileye, a company that specializes in sensors and computer vision for self-driving cars, for \$15.3 billion dollars. The stakes are high in the multitrillion-dollar transportation sector of the economy.

Self-driving cars will soon disrupt the livelihoods of millions of truck and taxi drivers. Eventually, there will be no need to own a car in a city when a self-driving car can show up in a minute and take you safely to your destination, without your having to park it. The average car today is only used 4 percent of the time, which means it needs to be parked somewhere 96 percent of the time. But because self-driving cars can be serviced and parked outside cities, vast stretches of city land now covered with parking lots can be repurposed for more productive uses. Urban planners are already

thinking ahead to the day when parking lots become parkland.<sup>3</sup> Parking lanes along streets can become real bike lanes. Many other car-related businesses will be affected, including auto insurance agencies and body shops. No more speeding or parking tickets. There will be fewer deaths from drunk drivers and from drivers falling asleep at the wheel. Time wasted commuting to work will be freed for other purposes. According to the U.S. Census Bureau, in 2014, 139 million Americans spent an average of 52 minutes commuting to and from work each workday. That amounts to 29.6 billion hours per year, or an astounding 3.4 million years of human lives that could have been put to better use.<sup>4</sup> Highway capacity will be increased by a factor of four by caravanning.<sup>5</sup> And, once developed and widely used, self-driving cars that can drive themselves home without a steering wheel will put an end to grand theft auto. Although there are many regulatory and legal obstacles in the way, when self-driving cars finally become ubiquitous, we will indeed be living in a brave new world. Trucks will be the first to become autonomous, probably in 10 years; taxis in 15 years and passenger cars in 15 to 25 years from start to finish.

The iconic position that cars have in our society will change in ways that we cannot imagine and a new car ecology will emerge. Just as the introduction of the automobile more than 100 years ago created many new industries and jobs, there is already a fast-growing ecosystem being created around self-driving cars. Waymo, the self-driving spin-off from Google, has invested \$1 billion over 8 years and has constructed a secretive testing facility in California's central valley with a 91-acre fake town, including fake bicycle riders and fake auto breakdowns.<sup>6</sup> The goal is to broaden the training data to include special and unusual circumstances, called edge cases. Rare driving events that occur on highways often lead to accidents. The difference with self-driving cars is that when one car experiences a rare event, the learning experience will propagate to all other self-driving cars, a form of collective intelligence. Many similar test facilities are being constructed by other self-driving car companies. These create new jobs that did not exist before, and new supply chains for the sensors and lasers that are needed to guide the cars.<sup>7</sup>

Self-driving cars are just the most visible manifestation of a major shift in an economy being driven by information technology (IT). Information flows through the Internet like water through city pipes. Information accumulates in massive data centers run by Google, Amazon, Microsoft, and other IT companies that require so much electrical power that they need to be located near hydroelectric plants, and streaming information generates so much heat that it needs rivers to supply the coolant. In 2013, data

centers in the United States consumed 10 million megawatts, equivalent to the power generated by thirty-four large power plants.<sup>8</sup> But what is now making an even bigger impact on the economy is how this information is used. Extracted from raw data, the information is being turned into knowledge about people and things: what we do, what we want, and who we are. And, more and more, computer-driven devices are using this knowledge to communicate with us through the spoken word. Unlike the passive knowledge in books that is externalized outside brains, knowledge in the cloud is an external intelligence that is becoming an active part of everyone's lives.<sup>9</sup>

### Learning How to Translate

Deep learning is used at Google today in more than 100 services, from Street View to Inbox Smart Reply and voice search. Several years ago, engineers at Google realized that they had to scale up these compute-intensive applications to cloud levels. Setting out to design a special-purpose chip for deep learning, they cleverly designed the board to fit into a hard disk drive slot in their data center racks. Google's tensor processing unit (TPU) is now deployed on servers around the world, delivering an order-of-magnitude improvement in performance for deep learning applications.

An example of how quickly deep learning can change the landscape is the impact it has had on language translation—a holy grail for artificial intelligence since it depends on the ability to understand a sentence. The recently unveiled new version of Google Translate based on deep learning represents a quantum leap improvement in the quality of translation between natural languages. Almost overnight, language translation went from a fragmented hit-and-miss jumble of phrases to seamless sentences (figure 1.3). Previous computer methods searched for combinations of words that could be translated together, but deep learning looks for dependencies across whole sentences.

Alerted about the sudden improvement of Google Translate, on November 18, 2016, Jun Rekimoto at the University of Tokyo tested the new system by having it translate the opening of Ernest Hemingway's "The Snows of Kilimanjaro" into Japanese and then back into English—with the following result (guess which one is the original Hemingway):

1: Kilimanjaro is a snow-covered mountain 19,710 feet high, and is said to be the highest mountain in Africa. Its western summit is called the Masai "Ngaje Ngai," the House of God. Close to the western summit there is the dried and frozen carcass of a leopard. No one has explained what the leopard was seeking at that altitude.

Copyrighted image

### Figure 1.3

Japanese signs and menus instantly translated into English by Google Translate, which is now an app on your smart phone. This is especially useful if you need to find the right train in Japan.

2: Kilimanjaro is a mountain of 19,710 feet covered with snow and is said to be the highest mountain in Africa. The summit of the west is called “Ngaje Ngai” in Masai, the house of God. Near the top of the west there is a dry and frozen dead body of leopard. No one has ever explained what leopard wanted at that altitude.<sup>10</sup>

(Hemingway is #1.)

The next step will be to train larger deep learning networks on paragraphs to improve continuity across sentences. Words have long cultural histories. Vladimir Nabokov, the Russian writer and English-language novelist who wrote *Lolita*, came to the conclusion that it was impossible to translate poetry between languages. His literal translation of Aleksandr Pushkin’s *Eugene Onegin* into English, annotated with explanatory footnotes on the cultural background of the verses, made his point.<sup>11</sup> Perhaps Google Translate will be able to translate Shakespeare someday by integrating across all of his poetry.<sup>12</sup>

### Learning How to Listen

Another holy grail of artificial intelligence is speech recognition. Until recently, speaker-independent speech recognition by computers was

limited to narrow domains, such as airline reservations. Today, it is unlimited. A summer research project at Microsoft Research by an intern from the University of Toronto in 2012 dramatically improved the performance of Microsoft's speech recognition system (figure 1.4).<sup>13</sup> In 2016, a team at Microsoft announced that its deep learning network with 120 layers had achieved human-level performance on a benchmark test for multi-speaker speech recognition.<sup>14</sup>

The consequences of this breakthrough will ripple through society over the next few years, as computer keyboards are replaced by natural language interfaces. This is already happening with digital assistants as Amazon's Alexa, Apple's Siri, and Microsoft's Cortana leapfrog one another into homes everywhere. Just as typewriters became obsolete with the widespread use of

Copyrighted image

#### Figure 1.4

Microsoft Chief Research Officer Rick Rashid in a live demonstration of automated speech recognition using deep learning on October 25, 2012, at an event in Tianjin, China. Before an audience of 2,000 Chinese, Rashid's words, spoken in English, were recognized by the automated system, which first showed them in subtitles below Rashid's screen image and then translated them into spoken Chinese. This high-wire act made newsfeeds worldwide. Courtesy of Microsoft Research.

to record your electroencephalogram (EEG) and muscle activity while you sleep. In the course of each night, you will enter into slow-wave sleep and, periodically, into rapid-eye-movement (REM) sleep, during which you will dream, but insomnia, sleep apnea, restless leg syndrome, and many other sleep disorders can disrupt this pattern. If you had trouble sleeping at home, sleeping in a strange bed connected by wires to ominous medical equipment can be a real challenge. A sleep expert will look over your EEG recordings and mark the sleep stages in blocks of 30 seconds, which takes several hours to score each eight hours of sleep. You will eventually get back a report on abnormalities in your sleep pattern and a bill for \$2,000.

The sleep expert will have been trained to look for telltale features that characterize the different sleep stages, based on a system devised in 1968 by Anthony Rechtschaffen and Alan Kales.<sup>18</sup> But, because the features are often ambiguous and inconsistent, experts agree only 75 percent of the time on how to interpret them. In contrast, Philip Low, a former graduate student in my lab, used unsupervised machine learning to automatically detect sleep stages with a time resolution of 3 seconds and a concordance with human experts of 87 percent, in less than a minute of computer time. Moreover, this required recording from only a single location on the head rather than many contacts and a bundle of wires that take a long time to put on and take off. In 2007, we launched a start-up company, Neurovigil, to bring this technology to sleep clinics, but they showed little interest in disrupting their cash flow from human scoring. Indeed, with an insurance code to bill patients, they had no incentive to adopt a cheaper procedure. Neurovigil found another market in large drug companies that run clinical trials and need to test the effects of their drugs on sleep patterns, and it is now entering the market for long-term care facilities, where elderly often have progressive sleep problems.

The sleep clinic model is flawed because health problems can't be reliably diagnosed based on such restricted circumstances: Everyone has a different baseline, and departures from that baseline are the most informative. Neurovigil already has a compact device, the iBrain, which can record your EEG at home, transmit the data to the Internet and analyze the data longitudinally for trends and anomalies. This will allow doctors to detect health problems early when it is easier to treat them and to stop the development of chronic illnesses. There are other diseases whose treatment would benefit from continuous monitoring, such as type 1 diabetes, for which the level of sugar in the blood could be monitored and regulated by delivery of insulin.

Access to cheap sensors that can record data continuously is having a major impact on diagnosis and treatment of other chronic diseases.

There are several lessons to be learned from the Neurovigil experience. Although having better and cheaper technology does not translate easily into a marketable new product or service, even a far superior one, when an incumbent is entrenched in the market, there are secondary markets where the new technology can have a more immediate impact and buy time to improve and better compete. This is how the technologies of solar energy and of many other new industries entered the market. In the long run, sleep monitoring and new technologies with demonstrated advantages will reach patients at home and eventually be integrated into medical practice.

### Learning How to Make Money

More than 75 percent of trading on the New York Stock Exchange is automated (figure 1.6), fueled by high-frequency trades that move into and out of positions in fractions of a second. (When you don't have to pay for each transaction, even small advantages can be parlayed into big profits.) Algorithmic trading on a longer time scale takes into account longer-term trends based on big data. Deep learning is getting better and better at making both more money and higher profits.<sup>19</sup> The problem with predicting the financial markets is that the data are noisy and conditions are not stationary—psychology can change overnight after an election or international conflict. This means that an algorithm that predicts stock values today may not work tomorrow. In practice, hundreds of algorithms are used and the best ones are continually combined to optimize returns.

Back in the 1980s, when I was consulting for Morgan Stanley on neural network models of stock trading, I met David Shaw, a computer scientist who specialized in designing parallel computers. On leave of absence from Columbia University, Shaw was working as a quantitative analyst, or “quant,” in the early days of automated trading. He would go on to start his own investment management firm on Wall Street, the D. E. Shaw Group, and he is now a multibillionaire. The D. E. Shaw Group has been highly successful, but not as successful as another hedge fund, Renaissance Technologies, which was founded by James Simons, a distinguished mathematician and former chair of the Mathematics Department at Stony Brook University. Simons made \$1.6 billion in 2016 alone, and this wasn't even his best year.<sup>20</sup> Called “the best physics and mathematics department in the world,”<sup>21</sup> Renaissance “avoids hiring anyone with even the slightest whiff of Wall Street bona fides.”<sup>22</sup>



## Latency versus position timeline

Latency

Copyrighted image

### How long position held

**Figure 1.6**

Machine learning is driving algorithmic trading, which is faster than traditional long-term investment strategies and more deliberate than high-frequency trading (HFT) in stock markets. Many different kinds of machine learning algorithms are combined to achieve best returns.

No longer involved in the daily operation of D. E. Shaw, David Shaw is now engrossed in D. E. Shaw Research, which has built a special-purpose parallel computer, called “Anton,” that performs protein folding much faster than any other computer on the planet.<sup>23</sup> Simons has retired from overseeing Renaissance and has started a foundation that funds research on autism and other programs in the physical and biological sciences. Through the Simons Institute for the Theory of Computing at UC Berkeley, the Simons Center for the Social Brain at MIT, and the Flatiron Institute in New York, Shaw’s philanthropy has had a major impact on advancing computational methods for data analysis, modeling, and simulation.<sup>24</sup>

Financial services more broadly are undergoing a transformation under the banner of financial technology, or “fintech,” as it has come to be called. Information technology such as block chain, which is a secure Internet ledger that replaces financial middlemen in transactions, is being tested on a small scale but could soon disrupt multitrillion-dollar financial markets. Machine learning is being used to improve credit evaluation on loans, to accurately deliver business and financial information, to pick up signals on

social media that predict market trends, and to provide biometric security for financial transactions. Whoever has the most data wins, and the world is awash with financial data.

### Learning the Law

Deep learning is just beginning to affect the legal profession. Much of the routine work of associates in law firms who charge hundreds of dollars an hour will be automated, especially in large, high-value commercial offices. In particular, technology-assisted review, or discovery, will be taken over by artificial intelligence, which can sort through thousands of documents for legal evidence without getting tired. Automated deep learning systems will also help law firms comply with the increasing complexity of governmental regulations. They will make legal advice available for the average person who cannot now afford a lawyer. Not only will legal work be cheaper; it will be much faster, a factor that is often more important than its expense. The world of law is well on its way to becoming “Legally Deep.”<sup>25</sup>

### Learning How to Play Poker

Heads-up no-limit Texas hold 'em is one of the most popular versions of poker, commonly played in casinos, and the no-limit betting form is played at the main event of the World Series of Poker (figure 1.7). Poker is challenging because, unlike chess, where both players have access to the same information, poker players have imperfect information, and, at the highest levels of play, skills in bluffing and deception are as important as the cards that are dealt.

The mathematician John von Neumann, who founded mathematical game theory and pioneered digital computers, was particularly fascinated with poker. As he put it: “Real life consists of bluffing, of little tactics of deception, of asking yourself what is the other man going to think I mean to do. And that is what games are about in my theory.”<sup>26</sup> Poker is a game that reflects parts of human intelligence that were refined by evolution. A deep learning network called “DeepStack” played 44,852 games against thirty-three professional poker players. To the shock of poker experts, it beat the best of the poker players by a sizable margin, one standard deviation, but it beat the thirty-three players overall by four standard deviations—an immense margin.<sup>27</sup> If this achievement is replicated in other areas where human judgment based on imperfect information is paramount, such as politics and international relations, the consequences could be far reaching.<sup>28</sup>

Copyrighted image

### Figure 1.7

Heads-up no-limit Texas hold 'em. Aces in the hole. Bluffing in high stakes poker has been mastered by DeepStack, which has beaten professional poker players at their own game by a wide margin.

### Learning How to Play Go

In March 2016, Lee Sedol, the Korean Go 18-time world champion, played and lost a five-game match against DeepMind's AlphaGo (figure 1.8), a Go-playing program that used deep learning networks to evaluate board positions and possible moves.<sup>29</sup> Go is to Chess in difficulty as chess is to checkers. If chess is a battle, Go is a war. A 19×19 Go board is much larger than an 8×8 chessboard, which makes it possible to have several battles raging in different parts of the board. There are long-range interactions between battles that are difficult to judge, even by experts. The total number of legal board positions for Go is  $10^{170}$ , far more than the number of atoms in the universe.

In addition to several deep learning networks to evaluate the board and choose the best move, AlphaGo had a completely different learning system, one used to solve the temporal credit assignment problem: which of the many moves were responsible for a win, and which were responsible for a loss? The basal ganglia of the brain, which receive projections from the entire cerebral cortex and project back to it, solve this problem with a temporal difference algorithm and reinforcement learning. AlphaGo used the same learning algorithm that the basal ganglia evolved to evaluate sequences of

A large rectangular area in the center of the page is marked with the text "Copyrighted image". This area is intended to contain a photograph of Demis Hassabis and Ke Jie, but the image itself is not visible.**Figure 1.10**

Demis Hassabis (left) and Ke Jie meet after the historic Go match in China in 2017, holding a board with Ke Jie's signature. Courtesy of Demis Hassabis.

in good form. Their performances follow an inverted U-shaped curve, with their best ones in an optimal state between low and high levels of arousal. Athletes call this being “in the zone.”

AlphaGo also defeated a team of five top players on May 26, 2017. These players have analyzed the moves made by AlphaGo and are already changing their strategies. In a new version of “ping-pong diplomacy,” the match was hosted by the Chinese government. China is making a large investment in machine learning, and a major goal of their brain initiative is to mine the brain for new algorithms.<sup>34</sup>

The next chapter in this Go saga is even more remarkable, if that is possible. AlphaGo was jump-started by supervised learning from 160,000 human Go games before playing itself. Some thought this was cheating—an autonomous AI program should be able to learn how to play Go without human knowledge. In October, 2017, a new version, called AlphaGo Zero, was revealed that learned to play Go starting with only the rules of the game, and trounced AlphaGo Master, the version that beat Ke Jie, winning 100 games to none.<sup>35</sup> Moreover, AlphaGo Zero learned 100 times faster and with 10 times less compute power than AlphaGo Master. By completely ignoring human knowledge, AlphaGo Zero became super-superhuman.

There is no known limit to how much better AlphaGo might become as machine learning algorithms continue to improve.

AlphaGo Zero had dispensed with human play, but there was still a lot of Go knowledge handcrafted into the features that the program used to represent the board. Maybe AlphaGo Zero could improve still further without any Go knowledge. Just as Coca-Cola Zero stripped all the calories from Coca-Cola, all domain knowledge of Go was stripped from AlphaZero. As a result, AlphaZero was able to learn even faster and decisively beat AlphaGo Zero.<sup>36</sup> To make the point that less is more even more dramatically, AlphaZero, without changing a single learning parameter, learned how to play chess at superhuman levels, making alien moves that no human had ever made before. AlphaZero did not lose a game to Stockfish, the top chess program already playing at superhuman levels. In one game, AlphaZero made a bold bishop sacrifice, sometimes used to gain positional advantage, followed by a queen sacrifice, which seemed like a colossal blunder until it led to a checkmate many moves later that neither Stockfish nor humans saw coming. The aliens have landed and the earth will never be the same again.

AlphaGo's developer, DeepMind, was cofounded in 2010 by neuroscientist Demis Hassabis (figure 1.10, left), who had been a postdoctoral fellow at University College London's Gatsby Computational Neuroscience Unit (directed by Peter Dayan, a former postdoctoral fellow in my lab and winner of the prestigious Brain Prize in 2017 along with Raymond Dolan and Wolfram Schultz for their research on reward learning). DeepMind was acquired by Google for \$600 million in 2014. The company employs more than 400 engineers and neuroscientists in a culture that is a blend between academia and start-ups. The synergies between neuroscience and AI run deep and are quickening.

### **Learning How to Become More Intelligent**

Is AlphaGo intelligent? There has been more written about intelligence than any other topic in psychology except consciousness, both of which are difficult to define. Psychologists since the 1930s distinguish between fluid intelligence, which uses reasoning and pattern recognition in new situations to solve new problems, without depending on previous knowledge, and crystallized intelligence, which depends on previous knowledge and is what the standard IQ tests measure. Fluid intelligence follows a developmental trajectory, reaching a peak in early adulthood and decreasing with age, whereas crystallized intelligence increases slowly and asymptotically as you age until fairly late in life. AlphaGo displays both crystallized and

fluid intelligence in a rather narrow domain, but within this domain, it has demonstrated surprising creativity. Professional expertise is also based on learning in narrow domains. We are all professionals in the domain of language and practice it every day.

The reinforcement learning algorithm used by AlphaGo can be applied to many problems. This form of learning depends only on the reward given to the winner at the end of a sequence of moves, which paradoxically can improve decisions made much earlier. When coupled with many powerful deep learning networks, this leads to many domain-dependent bits of intelligence. And, indeed, cases have been made for different domain-dependent kinds of intelligence: social, emotional, mechanical, and constructive, for example.<sup>37</sup> The “*g* factor” that intelligence tests claim to measure is correlated with these different kinds. There are reasons to be cautious about interpreting IQ tests. The average IQ has been going up all over the world by three points per decade since it was first studied in the 1930s, a trend called the “Flynn effect.” There are many possible explanations for the Flynn effect, such as better nutrition, better health care, and other environmental factors.<sup>38</sup> This is quite plausible because the environment affects gene regulation, which in turn affects brain connectivity, leading to changes in behavior.<sup>39</sup> As humans increasingly are living in artificially created environments, brains are being molded in ways that nature never intended. Could it be that humans have been getting smarter over a much longer period of time? For how long will the increase in IQ continue? The incidence of people playing computers in chess, backgammon, and now Go has been steadily increasing since the advent of computer programs that play at championship levels, and so has the machine augmented intelligence of the human players.<sup>40</sup> Deep learning will boost the intelligence not just of scientific investigators but of workers in all professions.

Scientific instruments are generating data at prodigious rate. Elementary particle collisions at the Large Hadron Collider (LHC) in Geneva generate 25 petabytes of data each year. The Large Synoptic Sky Telescope (LSST) will generate 6 petabytes of data each year. Machine learning is being used to analyze the huge physics and astronomy datasets that are too big for humans to search by traditional methods.<sup>41</sup> For example, DeepLensing is a neural network that recognizes images of distant galaxies that have been distorted by light bending by “gravitational lenses” around another galaxy along the line of sight. This allows many new distant galaxies to be automatically discovered. There are many other “needle-in-a-haystack” problems in physics and astronomy for which deep learning vastly amplifies traditional approaches to data analysis.

## The Shifting Job Market

Introduced by banks in the late 1960s to dispense cash to account holders 24/7, a much-welcomed convenience for those in need of cash before or after normal banking hours, automated teller machines (ATMs) have since acquired the ability to read handwritten checks. And though they reduced routine work for bank tellers, there are more bank tellers than before providing customers with personalized services such as mortgage and investment advice, and new ATM repair jobs<sup>42</sup>—just as the steam engine displaced manual laborers, on the one hand, but gave rise to new jobs for skilled workers who could build and maintain steam engines and drive steam locomotives, on the other. So, too, Amazon’s online marketing has displaced many workers from local brick-and-mortar retail stores but has also created 380,000 new jobs for workers in the distribution and delivery of the goods sold by it and by the many businesses under its umbrella.<sup>43</sup> And as jobs that now require human cognitive skills are taken over by automated AI systems, there will be new jobs for those who can create and maintain these systems.

Job turnover is nothing new. Farmworkers in the nineteenth century were displaced by machines, and new jobs were created at city factories made possible by machines, all of which required an educational system to train workers in new skills. The difference is that, today, the new jobs being opened up by artificial intelligence will require new, different, and ever-changing skills in addition to traditional cognitive skills.<sup>44</sup> So we will need to learn throughout our lifetimes. For this to happen, we will need a new educational system that is based at the home rather than the school.

Fortunately, just as the need for finding new jobs has become acute, the Internet has made available free massive open online courses (MOOCs) to acquire new knowledge and skills. Though still in their infancy, MOOCs are evolving rapidly in the education ecosystem and hold great promise for delivering quality instruction to a wider range of people than ever before. When coupled with the next generation of digital assistants, MOOCs could be transformational. Barbara Oakley and I developed a popular MOOC called “Learning How to Learn” that teaches you how to become a better learner (figure 1.11) and a follow-up MOOC called “Mindshift” that teaches you how to reinvent yourself and change your lifestyle (both MOOCs will be described in chapter 12).

As you interact with the Internet, you are generating big data about yourself that is machine readable. You are being targeted by ads generated

Copyrighted image

# Learning How to Learn

**Figure 1.11**

“Learning How to Learn,” a massive open online course (MOOC) that teaches you how to become a better learner is the most popular MOOC on the Internet, with over 3 million learners. Courtesy of Terrence Sejnowski and Barbara Oakley.

from the digital bread crumbs you have left behind on the Internet. The information you reveal on Facebook and other social media sites can be used to create a digital assistant that knows you better than almost anyone else in the world and will not forget anything, becoming, in effect, your virtual doppelganger. By pressing both Internet tracking and deep learning into service, the educational opportunities for the children of today’s children will be better than the best available today to wealthy families. These grandchildren will have their own digital tutors, who will accompany them throughout the trajectory of their education. Not only will education become more individualized; it will become more precise. There are already a wide range of educational experiments under way throughout the world at programs like the Kahn Academy and funded by the Gates, Chan-Zuckerberg, and other philanthropic foundations that are testing software to make it possible for all children to progress at their own pace throughout their formal education and to adapt to the specific needs of each child.<sup>45</sup> The widespread availability of digital tutors will free teachers from the repetitive parts of teaching, like grading, and allow them to do what humans do best—emotional support for struggling students and intellectual inspiration for gifted students. Educational technology—edtech—is moving rapidly ahead, and the transition to precision education could be quite fast compared to self-driving cars because the obstacles it must overcome are much less daunting, the demand is much greater, and education in the U.S. is a trillion-dollar market.<sup>46</sup> One major concern will be who has access to the internal files of the digital assistants and digital tutors.

## Is Artificial Intelligence an Existential Threat?

When AlphaGo convincingly beat Lee Sedol at Go in 2016, it fueled a reaction that had been building over the last several years concerning the



generally known. *The Deep Learning Revolution* tells that story and explores the origins and consequences of deep learning from my perspective both as a pioneer in developing learning algorithms for neural networks in the 1980s and as the president of the Neural Information Processing Systems (NIPS) Foundation, which has overseen discoveries in machine learning and deep learning over the last thirty years. My colleagues and I in the neural network community were for many years the underdogs, but our persistence and patience eventually prevailed.

## 2 The Rebirth of Artificial Intelligence

Marvin Minsky was a brilliant mathematician and a founder of the MIT Artificial Intelligence Laboratory (MIT AI Lab).<sup>1</sup> Founders set the direction and the culture of a field, and, thanks in no small part to Minsky, artificial intelligence at MIT in the 1960s was a bastion of cleverness. Bubbling over with more ideas per minute than anyone else I knew, he could convince you that his take on a problem was right, even when common sense told you otherwise. I admired his boldness and his cleverness—but not the direction that he took AI.

### Child's Play?

Blocks World is a good example of a project that came out of the MIT AI Lab in the 1960s. To simplify the problem of vision, Blocks World consisted of rectangular building blocks that could be stacked to create structures (figure 2.1). The goal was to write a program that could interpret a command, such as “Find a large yellow block and put it on top of the red block,” and plan the steps needed for a robot arm to carry out the command. This seems like child's play, but a large, complex program had to be written, one that became so cumbersome that it could not be readily debugged and was effectively abandoned when the student who wrote the program, Terry Winograd, left MIT. This seemingly simple problem was much harder than anyone thought it would be, and, even if it had succeeded, there was no direct path from Blocks World to the real world, where objects come in many shapes, sizes, and weights, and not all angles are right angles. Compared to a controlled laboratory setting where the direction and level of lighting can be fixed, in the real world, lighting can vary dramatically from place to place and time to time, which greatly complicates the task of object recognition for computers.

Copyrighted image

### Figure 2.1

Marvin Minsky watching a robot stacking blocks around 1968. Blocks World was a simplified version of how we interact with the world, but it was far more complex than anyone imagined, and was not solved until 2016 by deep learning.

In the 1960s, the MIT AI Lab received a large grant from a military research agency to build a robot that could play Ping-Pong. I once heard a story that the principal investigator forgot to ask for money in the grant proposal to build a vision system for the robot, so he assigned the problem to a graduate student as a summer project. I once asked Marvin Minsky whether the story was true. He snapped back that I had it wrong: “We assigned the problem to undergraduate students.” A document from the archives at MIT confirms his version of the story.<sup>2</sup> What looked like it would be an easy problem to solve proved to be quicksand that swallowed a generation of researchers in computer vision.

### Why Vision Is a Hard Problem

We rarely have difficulty identifying what an object is despite differences in the location, size, orientation, and lighting of the object. One of the earliest ideas in computer vision was to match a template of the object with the pixels in the image, but that approach failed because the pixels of the two images of the same object in different orientations don't match. For example, consider the two birds in figure 2.2. If you shift the image of one bird over the other, you can get a part to match, but the rest is out of register; but you can get a fairly good match to an image of another bird species in the same pose.

Copyrighted image.

### Figure 2.2

Zebra finches consulting with each other. We have no difficulty seeing that they are the same species. But because they have different orientations to the viewer it is difficult to compare them with templates even though they have almost identical features.

Progress in computer vision was made by focusing not on pixels but on features. For example, birders have to become experts in distinguishing between different species that may differ in only a few subtle markings. A practical and popular book on identifying birds has only one photograph of a bird (figure 2.3), but many schematic drawings pointing out the subtle differences between them.<sup>3</sup> A good feature is one that is unique to one bird species, but because the same features are found on many species, what makes it possible to identify a bird is the unique combination of several field marks such as wing bars, eye stripes, and wing patches. And when these field marks are shared by closely related species, there are calls and songs that distinguish one from another. Drawings or paintings of birds are much better at directing our attention to the relevant distinguishing features than are photographs, which are filled with hundreds of less relevant features (figure 2.3).

The problem with this features-based approach is not just that it is very labor intensive to develop feature detectors for the hundreds of thousands of different objects in the world, but that, even with the best feature detectors, ambiguities arise from images of objects that are partially occluded, which makes recognizing objects in cluttered scenes a daunting task for computers.

Copyrighted image

### Figure 2.3

Distinctive feature that can be used to discriminate between similar birds. The arrows point toward the location of where to find wing bars that are especially important for telling apart families of warblers: Some are conspicuous, some obscure, some double, some long, some short. From Peterson, Mountfort, and Hollom, *Field Guide to the Birds of Britain and Europe*, 5th ed., p.16.

Little did anyone suspect in the 1960s that it would take fifty years and a millionfold increase in computer power before computer vision would reach human levels of performance. The misleading intuition that it would be easy to write a computer vision program is based on activities that we find easy to do, such as seeing, hearing, and moving around—but that took evolution millions of years to get right. Much to their chagrin, early AI pioneers found the computer vision problem to be extremely hard to solve. In contrast, they found it much easier to program computers to prove mathematical theorems—a process thought to require the highest levels of intelligence—because computers turn out to be much better at logic than we are. Being able to think logically is a late development in evolution and, even in humans, requires training to follow a long line of logical propositions to a rigorous conclusion, whereas, for most problems we need to solve to survive, generalizations from previous experiences work well for us most of the time.

### Expert Systems

Popular in the 1970s and 1980s, AI expert systems were developed to solve problems like medical diagnosis using a set of rules. Thus an early expert system, MYCIN, was developed to identify the bacteria responsible for infectious diseases such as meningitis.<sup>4</sup> Following the expert system approach, MYCIN's developers had first to collect facts and rules from infectious disease experts, as well as symptoms and medical histories from the patients, then to enter these into the system's computer, and finally to program the

Copyrighted image

#### Figure 2.4

Terry Sejnowski talking about scaling laws for the cortex shortly after he moved to the Salk Institute in 1989. Courtesy of Ciencia Explicada.

winging it. “The fly can see, it can fly, it can navigate, and it can find food. But what is truly remarkable is that it can reproduce itself. MIT owns a supercomputer that costs \$100 million: it consumes a megawatt of power and is cooled by a huge air-conditioner. But the biggest cost of the supercomputer is human sacrifice in the form of programmers to feed its voracious appetite for programs. That supercomputer can’t see, it can’t fly, and although it communicates with other computers, it can’t mate or reproduce itself. What is wrong with this picture?”

After a long pause, a senior faculty member spoke, “Because we haven’t written the vision program yet.” (The Department of Defense had recently poured \$600 million into its Strategic Computing Initiative, a program that ran from 1983 to 1993 but came up short on building a vision system to guide a self-driving tank.)<sup>9</sup> “Good luck with that,” was my reply.

Gerald Sussman, who made several important applications of AI to real-world problems, including a system for high-precision integration for orbital mechanics, defended the honor of MIT’s approach to AI with an appeal to the classic work of Alan Turing, who had proven that the Turing machine, a thought experiment, could compute any computable function. “And how long would that take?” I asked. “You had better compute quickly or you will be eaten,” I added, then walked across the room to pour myself a cup of coffee. And that was the end of the dialogue with the faculty.

“What is wrong with this picture?” is a question that every student in my lab can answer. But the first two rows of my lunchtime audience were stumped. Finally, a student in the third row offered this reply: “The digital computer is a general-purpose device, which can be programmed to compute anything, though inefficiently, but the fly is a special-purpose computer that can see and fly but can’t balance my checkbook.” This was the right answer. The vision networks in the fly eye evolved over hundreds of millions of years, and its vision algorithms are embedded in the networks themselves. This is why you can reverse engineer vision by working out the wiring diagram and information flow through the neural circuits of the fly eye, and why you can’t do that for a digital computer, where the hardware by itself needs software to specify what problem is being solved.

I recognized Rodney Brooks smiling in the back of the crowd, someone I had once invited to a workshop on computational neuroscience in Woods Hole on Cape Cod, Massachusetts. Brooks is from Australia, and, in the 1980s, he was a junior faculty member in the MIT AI Lab, where he built walking robotic insects using an architecture that did not depend on digital logic. He would eventually become the lab’s director and go on to found iRobot, the company that makes Roombas.

The room where I gave my lecture that afternoon was huge and filled with a large contingent of undergraduate students, the next generation looking to the future rather than the past. I talked about a neural network that learned how to play backgammon, a project I collaborated on with Gerald Tesauro, a physicist at the Center for Complex Systems Research at the University of Illinois in Urbana-Champaign. Backgammon is a race to the finish between two players, with pieces that move forward based on each roll of the dice, passing over one another on the way. Unlike chess, which is deterministic, backgammon is governed by chance: the uncertainty with every roll of the dice makes it more difficult to predict the outcome of a particular move. It is a highly popular game in the Middle East, where some make a living playing high-stakes backgammon.

Rather than write a program based on logic and heuristics to handle all possible board positions, an impossible task given that there are  $10^{20}$  possible backgammon board positions, we had the network learn to play through pattern recognition by watching a teacher play.<sup>10</sup> Gerry went on to create the first backgammon program that played at world-championship levels by having the backgammon network play itself (a story that will be told in chapter 10).

After my lecture, I learned that there was a front page article in the *New York Times* that morning about how government agencies were slashing

funding for artificial intelligence. Although this was the beginning of an AI winter for mainstream researchers, it didn't affect me or the rest of my group, for whom the neural network spring had just begun.

But our new approach to AI would take twenty-five years to deliver real-world applications in vision, speech, and language. Even in 1989, I should have known it would take this long. In 1978, when I was a graduate student at Princeton, I extrapolated Moore's law for the exponential increase in computing power, doubling every 18 months, to see how long it would take to reach brain levels of computing power and concluded it would happen in 2015. Fortunately, that did not deter me from charging ahead. My belief in neural networks was based on my intuition that if nature had solved these problems, we should be able to learn from nature how to solve them, too. The twenty-five years I had to wait was not even a blink of the eye compared to the hundreds of millions of years it took nature.

Inside the visual cortex, neurons are arranged in a hierarchy of layers. As sensory information is transformed cortical layer by cortical layer, the representation of the world becomes more and more abstract. Over the decades, as the number of layers in neural network models increased, their performance continued to improve until finally a critical threshold was reached that allowed us to solve problems we could only dream about solving in the 1980s. Deep learning automates the process of finding good features that distinguish different objects in an image, and that is why computer vision is so much better today than it was five years ago.

By 2016, computers had become a million times faster and computer memory had increased by a million times from megabytes to terabytes. It became possible to simulate neural networks with millions of units and billions of connections, compared with networks in the 1980s that had only hundreds of units and thousands of connections. Though still tiny by the standards of a human brain, which has a hundred billion neurons and a million billion synaptic connections, today's networks are now large enough to demonstrate proof of principle in narrow domains.

Deep learning in deep neural networks has arrived. But before there were deep networks, we had to learn how to train shallow networks.





### 3 The Dawn of Neural Networks

The only existence proof that any of the hard problems in artificial intelligence can be solved is the fact that, through evolution, nature has already solved them. But there were clues in the 1950s for how computers might actually achieve intelligent behavior, if AI researchers would take an approach that was fundamentally different from symbol processing.

The first clue was that our brains are powerful pattern recognizers. Our visual systems can recognize an object in a cluttered scene in one-tenth of a second, even though we may have never seen that particular object before and even when the object is in any location, of any size, and in any orientation to us. In short, our visual system behaves like a computer that has “recognize object” as a single instruction.

The second clue was that our brains can learn how to perform many difficult tasks through practice, from playing the piano to mastering physics. Nature uses general-purpose learning to solve specialized problems, and humans are champion learners. This is our special power. The organization of our cerebral cortex is similar throughout, and deep learning networks are found in all our sensory and motor systems.<sup>1</sup>

The third clue was that our brains aren’t filled with logic or rules. Yes, we can learn how to think logically or follow rules, but only after a lot of training, and most of us aren’t very good at it. This is illustrated by typical performances on a logical puzzle called the “Wason selection task” (figure 3.1).

The correct selections are the card with “8” and the brown card. In the original study, only 10 percent of subjects got the right answer.<sup>2</sup> But most subjects had no trouble getting the right answer when the logic test was grounded in a familiar context (figure 3.2).

Reasoning seems to be domain specific, and the more familiar we are with a domain, the easier it is for us to solve problems in that domain. Experience makes it easier to reason within a domain because we can use examples we have encountered to intuit solutions. In physics, for example,

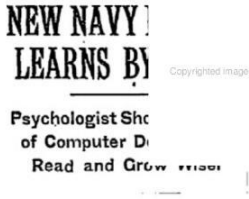
Copyrighted image.

### Figure 3.3

Pandemonium. Oliver Selfridge imagined that there were demons in the brain that were responsible for extracting successively more complex features and abstractions from sensory inputs, resulting in decisions. Each demon at each level is excited if it is a match to input from an earlier level. The decision demon weighs the degree of excitement and importance of its informants. This form of evidence evaluation is a metaphor for current deep learning networks, which have many more levels. From Peter H. Lindsay and Donald A. Norman, *Human Information Processing: An Introduction to Psychology*, 2nd ed. (New York: Academic Press, 1977), figure 3-1. Wikipedia Commons: <https://commons.wikimedia.org/wiki/File:Pande.jpg>.

The traditional way that an engineer solves this problem is to handcraft the weights based on analysis or an ad hoc procedure. This is labor intensive and often depends on intuition as much as on engineering. An alternative is to use an automatic procedure that learns from examples, the same way that we learn about objects in the world. Many examples are needed including those not in the category, especially if they are similar, such as dogs if the goal is to recognize cats. The examples are passed to the perceptron one at a time and corrections are automatically made to the weights if there is a classification error.

The beauty of the perceptron learning algorithm is that it is guaranteed to find a set of weights automatically if such a set of weights exists and if enough examples are available. The learning takes place incrementally after each of the examples in the training set is presented and the output compared with the correct answer. If the answer is correct, no changes are made to the weights, but if it isn't correct (1 when it should be 0, or 0 when



Copyrighted image

Copyrighted image

Copyrighted image

**Figure 3.4**

Frank Rosenblatt at Cornell deep in thought. He invented the perceptron, an early precursor of deep learning networks, which had a simple learning algorithm for classifying images into categories. Article in the *New York Times*, July 8, 1958, from a UPI wire report. The perceptron machine was expected to cost \$100,000 on completion in 1959, or around \$1 million in today's dollars; the IBM 704 computer that cost \$2 million in 1958, or \$20 million in today's dollars, could perform 4,000 multiplies per second, which was blazingly fast at the time. But the much less expensive Samsung Galaxy S6 phone, which can perform 34 billion operations per second, is more than a million times faster. Photo courtesy of George Nagy.

### Box 3.1 The Perceptron



A perceptron is a neural network with one artificial neuron that has an input layer and a set of connections linking the input units to the output unit. The goal of a perceptron is to classify patterns presented to input units. The basic operation performed by the output unit is to sum up the values of each input ( $x_i$ ) multiplied by its connection strength, or weight ( $w_i$ ), to the output unit. In the diagram above, a weighted sum of the inputs ( $\sum_{i=1, \dots, n} w_i x_i$ ) is compared to the threshold  $\theta$  and passed through a step function that gives an output of “1” if the sum is greater than the threshold and an output of “0” otherwise. For example, the input could be the intensities of pixels in an image, or more generally, features that are extracted from the raw image, such as the outline of objects in the image. Images are presented one at a time, and the perceptron decides whether or not the image is a member of a category, such as the category of cats. The output can only be in one of two states, “on” if the image is in the category or “off” if it isn’t. “On” and “off” correspond to the binary values 1 and 0, respectively. The perceptron learning algorithm is

$$\Delta w_i = \alpha \delta x_i$$

$$\delta = \text{output} - \text{teacher},$$

where both the output and teacher are binary, so that  $\delta = 0$  if the output is correct, and  $\delta = +1$  or  $-1$  if the output is not correct, depending on the difference.

it should be 1), then the weights are changed slightly so that the next time the same input is given, it is closer to getting the correct answer (box 3.1). It is important that the changes occur gradually so that the weights can feel the tugs from all the training examples, and not just from the last one.

If this explanation of perceptron learning isn't clear, there is a much neater geometric way to understand how a perceptron learns to classify inputs. For the special case of two inputs, it is possible to plot the inputs on a two-dimensional graph. Each input is a point in the graph and the two weights in the network determine a straight line. The goal of learning is to move the line around so that it cleanly separates the positive and negative examples (figure 3.5). For three inputs, the space of inputs is three-dimensional, and the perceptron specifies a plane that separates the positive and negative training examples. The same principle holds even in the general case, when the dimensionality of the space of inputs can be quite high and impossible to visualize.

Eventually, if a solution is possible, the weights will stop changing, which means the perceptron has correctly classified all of the examples in the training set. But, in what is called "overfitting," it is also possible that there are not enough examples in the set, and the network has simply memorized the specific examples without being able to generalize to new ones. To avoid overfitting, it is important to have another set of examples,



**Figure 3.5**

Geometric explanation for how two object categories are discriminated by a perceptron. The objects have two features, such as size and brightness, which have values  $(x,y)$  and are plotted on each graph. The two types of objects (pluses and squares) in the panel on the left can be separated by a straight line that passes between them; this discrimination can be learned by a perceptron. The two types of objects in the other two panels cannot be separated by a straight line, but those in the center panel can be separated by a curved line. The objects in the panel on the right would have to be gerrymandered to separate the two types. The discriminations in all three panels could be learned by a deep learning network if enough training data were available.

called a “test set,” that wasn’t used to train the network. At the end of training, the classification performance on the test set is a true measure of how well the perceptron can generalize to new examples whose respective categories are unknown. Generalization is the key concept here. In real life, we never see the same object the same way or encounter the same situation, but if we can generalize from previous experience to new views or situations, we can handle a broad range of real-world problems.

## SEXNET

As an example of how a perceptron can be used to solve a real-world problem, consider how you would tell a male from a female face, taking away hair, jewelry, and secondary sexual characteristics such as Adam’s apples, which tend to be larger in males. Beatrice Golomb, a postdoctoral fellow in my lab in 1990, used faces of college students from a database she obtained as inputs to a perceptron that was trained to classify the sex of a face with an 81 percent accuracy (figure 3.6).<sup>8</sup> The faces that the perceptron had difficulty classifying were also difficult for humans to classify, and members of my lab achieved an average performance of 88 percent on the same set of faces. Beatrice also trained a multilayer perceptron (which will be introduced in chapter 8) that achieved a 92 percent accuracy,<sup>9</sup> better than people from my lab. At a talk she gave at the 1991 Neural Information Processing Systems (NIPS) Conference, she concluded: “Since experience improves performance, this should suggest that people in the lab need to spend more time engaged in discriminating sex.” She called her multilayer perceptron the “SEXNET.” In the question-and-answer period, someone asked whether SEXNET could be used to detect transvestite faces. “Yes,” said Beatrice, to which Ed Posner, the founder of the NIPS conferences, retorted, “That would be the DRAGNET.”<sup>10</sup>

**Figure 3.6**

What is the sex of this face—male or female? A perceptron was trained to discriminate male from female faces. The pixels from the image of a face (top) are multiplied by the corresponding weights (bottom), and the sum is compared to a threshold. The size of each weight is depicted as the area of the pixel. Positive weights (white) are evidence for maleness and negative weights (black) favor femaleness. The nose width, the size of the region between the nose and mouth, and image intensity around the eye region are important for discriminating males, whereas image intensity around the mouth and cheekbone is important for discriminating females. From M. S. Gray, D. T. Lawrence, B. A. Golomb, and T. J. Sejnowski, “A Perceptron Reveals the Face of Sex,” *Neural Computation* 7 (1995): 1160–1164, figure 1.

Embarrassingly simple distributions of points in two dimensions cannot be separated by a perceptron (figure 3.5, nonlinear). It turned out that the tank perceptron was not a tank classifier, but a time of day classifier. It is much more difficult to classify tanks in images; indeed, it cannot be done with a perceptron. This also shows that, even when a perceptron has learned something, it may not be what you think it has learned. The final blow to the perceptron was a 1969 tour de force mathematical treatise, *Perceptrons* by Marvin Minsky and Seymour Papert.<sup>14</sup> Their definitive geometric analysis showed that the capabilities of perceptrons are limited: they can only separate categories that are linearly separable (figure 3.5). The cover of their book illustrates a geometric problem that Minsky and Papert proved the perceptron could not solve (figure 3.7). Although, at the end of their book, Minsky and Papert considered the prospect of generalizing single- to multiple-layer perceptrons, one layer feeding into the next, they doubted there would ever be a way to train even these more powerful perceptrons. Unfortunately, many took this doubt to be definitive, and the field was abandoned until a new generation of neural network researchers took a fresh look at the problem in the 1980s.

In a perceptron, each input contributes independent evidence to the output unit. But what if several inputs need to be combined in ways that make decisions dependent on the combination and not on each input separately? This is why a perceptron cannot distinguish whether a spiral is connected or not: a single pixel carries no information on whether it is on the inside or the outside. Although in multilayer feedforward networks, combinations of several inputs can be formed in intermediate layers between the input and output units, no one in the 1960s knew how to train a network with even a single layer of such “hidden units” between the input and output layers.

Frank Rosenblatt and Marvin Minsky had been classmates at the Bronx High School of Science in New York City. They debated their radically different approaches to artificial intelligence at scientific meetings, where participants tilted toward Minsky’s approach. But despite their differences, each man made important contributions to our understanding of perceptrons, which is the starting point for deep learning.

When Rosenblatt died in a boating accident in 1971 at age 43, the backlash against perceptrons was in full swing, and there were rumors that he might have committed suicide, or was it an outing gone tragically wrong?<sup>15</sup> What became clear was that a heroic period of discovering a new way of computing with neural networks was closing; a generation would pass before the promise of Rosenblatt’s pioneering efforts was realized.