IS THE ALGORITHM PLOTTING AGAINST US?

A Layperson's Guide to the Concepts, Math, and Pitfalls of AI

KENNETH WENGER

Supplemental visuals to the audio book

Keep the conversation going!

Scan the QR code to keep in touch with everything that's happening at Working Fires Foundation.

CHAPTER 1: POLARIZATION AND ITS CONSEQUENCES

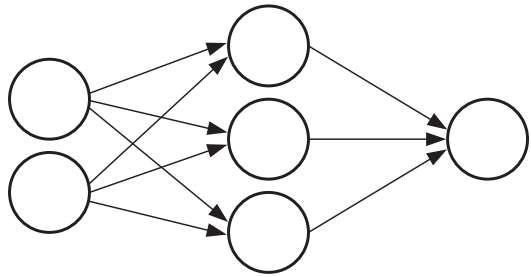

Figure 1.1 A simple artificial neural network. Information flows from left to right. The *two circles* in the first layer are the input nodes. The *three circles* in the middle are the processing neurons, and the *one circle* on the right represents the output value. The output in a neural network signifies a "prediction" on the input.

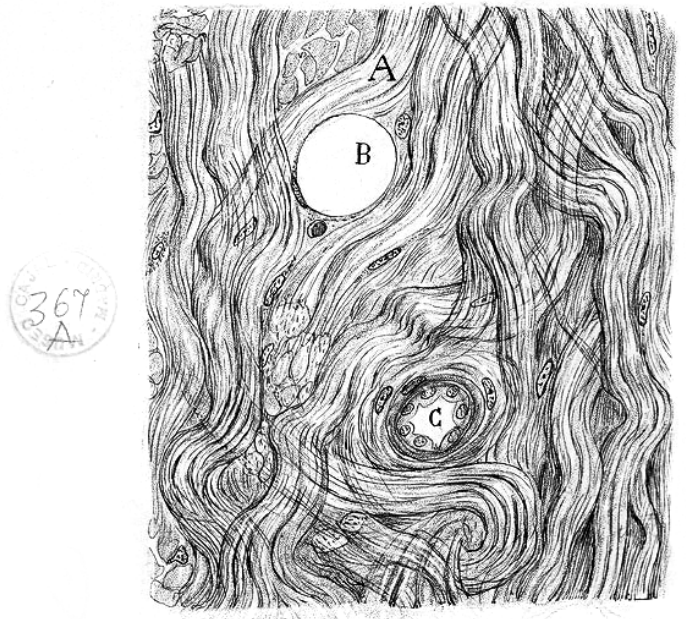

Figure 1.2 Tumor cells of the covering membranes of the brain, 1890. *Cajal Institute (CSIS), Madrid.*

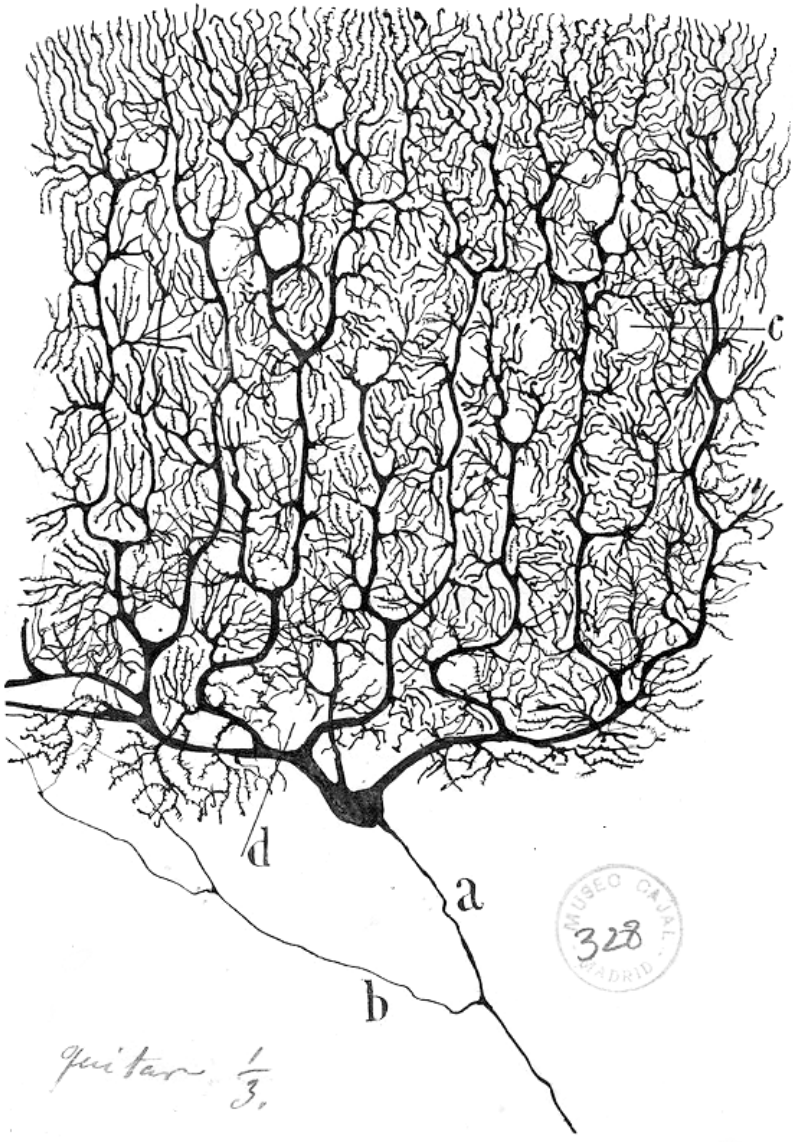

Figure 1.3 A purkinje neuron from the human cerebellum, ca. 1900. *Cajal Institute (CSIS), Madrid.*

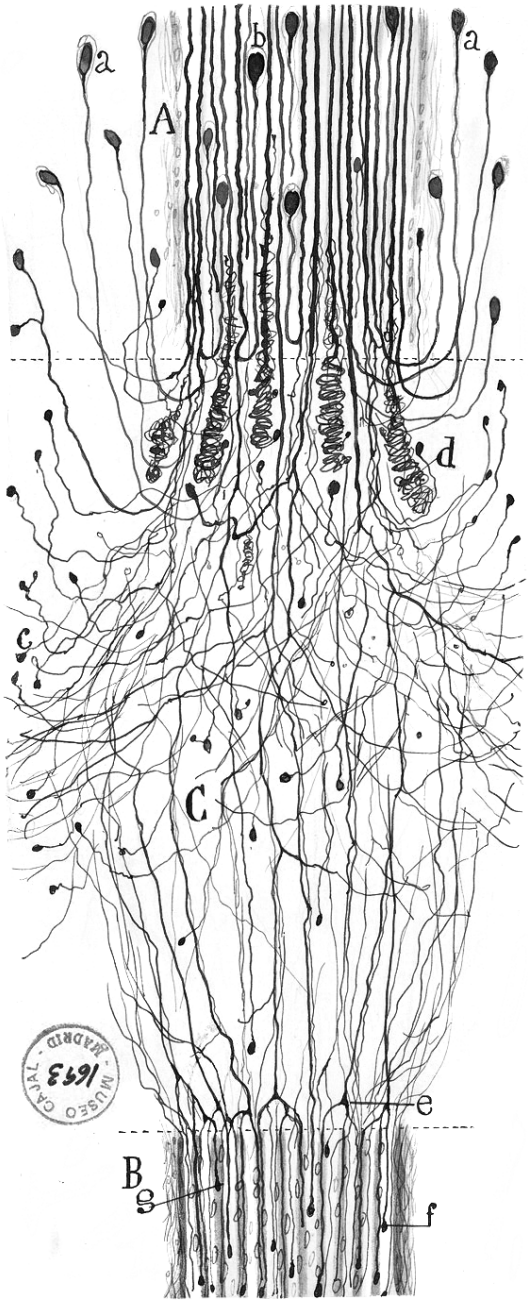

Figure 1.4 A cut nerve outside the spinal cord, 1913. *Cajal Institute (CSIS), Madrid.*

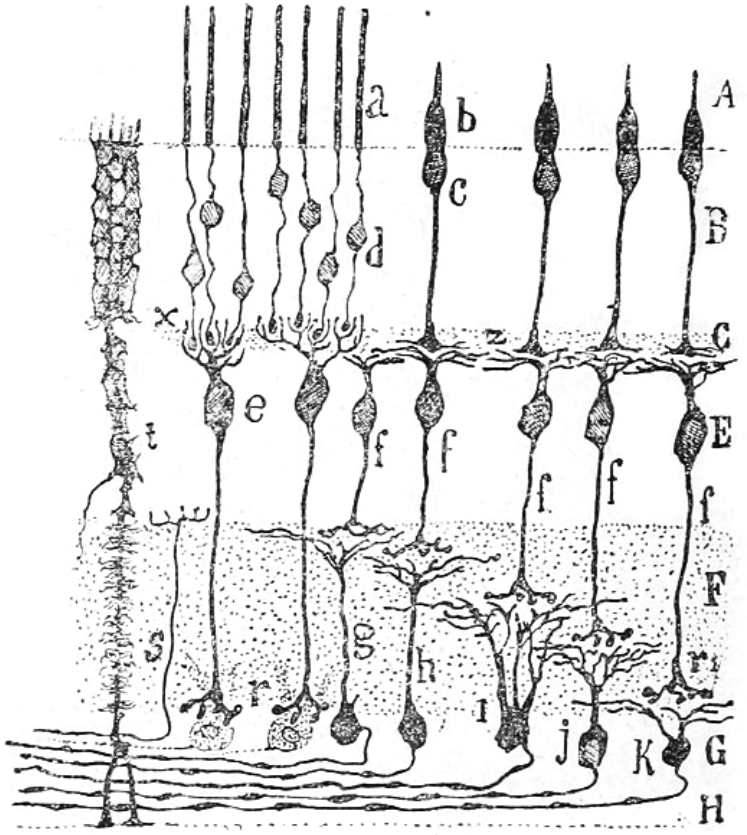

FIG. 27.

Coupe transversale de la rétine d'un mammifère.

Figure 1.5 "Coupe transversal de la rétine d'un mammifère" (Cross section of the retina of a mammal). *Illustration from* Les nouvelles idées sur la structure du système nerveux: chez l'homme et chez les vertébrés *(Paris: C. Reinwald, 1894), 112.*

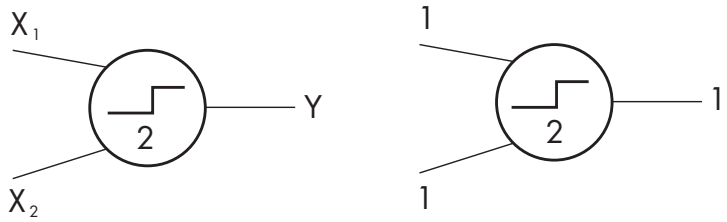

Figure 1.6 A McCulloch-Pitts neuron with two inputs, a firing threshold of 2, and one output. In the example on the right, both input signals have a value of 1.

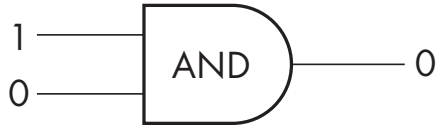

Figure 1.7 An AND logic gate with two inputs and one output. If either of the inputs is 0, the output is 0. If both inputs are 1, the output is 1.

Table 1.1 AND Gate and Possible Input/Output Combinations

Input 1	Input 2	Output
0	0	0
0	1	0
1	0	0
1	1	1

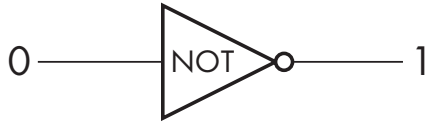

Figure 1.8 A NOT logic gate, with one input and one output. If the input is 0, the output is 1. If the input is 1, the output is 0.

Table 1.2 NOT Gate and Possible Input/Output Combinations	
Input 1	Output
0	1
1	0

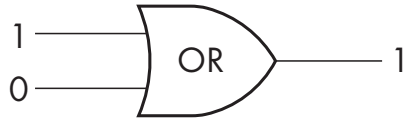

Figure 1.9 An OR logic gate with two inputs and one output. If either input is 1, the output is 1. If both inputs are 0, the output is 0.

Table 1.3 OR Gate and Possible Input/Output Combinations		
Input 1	Input 2	Output
0	0	0
0	1	1
1	0	1
1	1	1

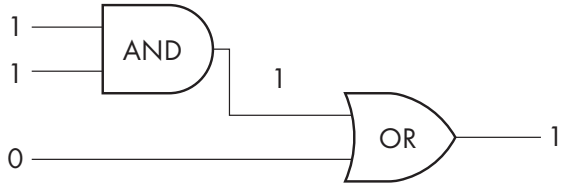

Figure 1.10 An AND gate and an OR gate automatically controlling access into the EU.

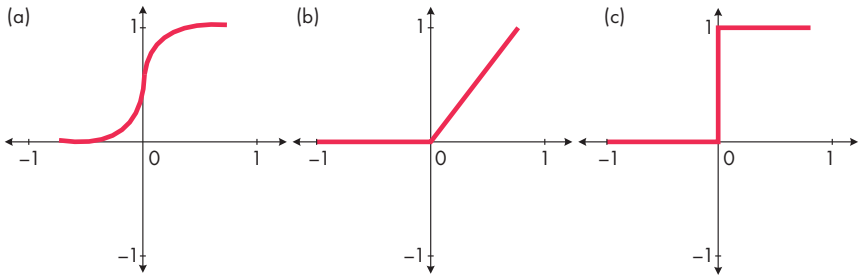

Figure 1.11 Three activation functions we refer to often in this chapter. The sigmoid activation function (*a*) outputs a value between 0 and 1 for all its inputs, with an input 0 producing 0.5 as output. The rectified linear unit activation function (*b*) outputs 0 for all negative inputs, and outputs the unmodified input value for all positive inputs. The threshold function (*c*) outputs a 0 for all input values below the threshold value; for input values equal to or greater than the threshold, it outputs a 1.

Equation 1

$$V = \left(\sum_{j=1}^{n} w_j x_j \right) + b$$

Equation 2

$$O = f(w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 + w_5 x_5) + b$$

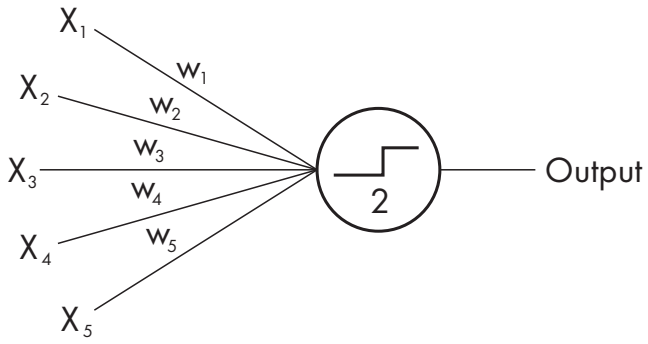

Figure 1.12 A Rosenblatt perceptron with five input connections. Each connection has a weight (*w*) associated with it. The value of the weight can be any number between 0 and 1. The output of the perceptron depends on the weighted sum operation passing the threshold test.

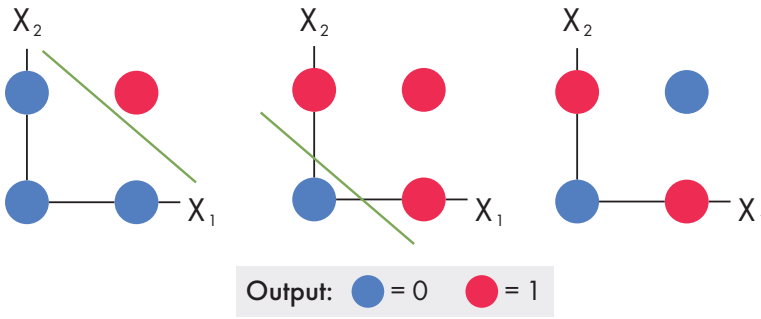

Figure 1.13 Possible outputs of the AND, OR, and XOR logic gates. The AND gate outputs a 1 only when both X_1 and X_2 are 1. This is shown as the *red circle*; otherwise, it outputs a 0 (*blue circle*). The outputs of an AND gate are linearly separable: we can draw a single line that separates the two classes of outputs, 0 and 1. The same is true for the or gate. But for the XOR gate, we cannot draw a single straight line to separate the classes of outputs: the outputs of the XOR gate are not linearly separable.

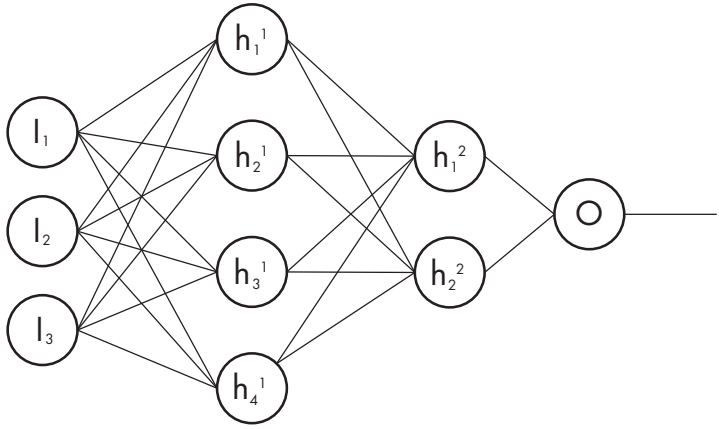

Figure 1.14 A multilayer perceptron (MLP) with two hidden layers. Three input nodes are connected to the first hidden layer (h^1), and h^1 is connected to the second hidden layer (h^2), which is connected to the output neuron.

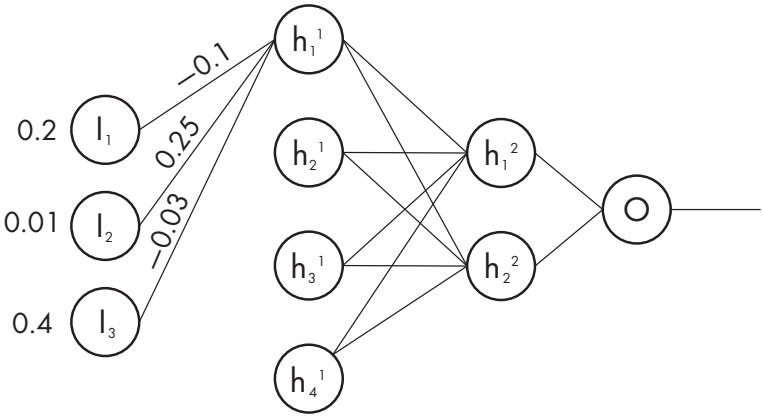

Figure 1.15 The weight values of the connections to h_1^1. Note that we have removed the connections to the other neurons in h^1 only for visibility reasons.

Equation Series A

$h_1^1 = \text{ReLu}(w_{1,1}{}^1 I_1 + w_{1,2}{}^1 I_2 + w_{1,3}{}^1 I_3)$

$= \text{ReLu}(-0.1*0.2 + 0.25*0.01 + (-0.03*0.4))$

Recall that ReLu functions output 0 for negative values.

$= \text{ReLu}(-0.0295)$

$= 0$

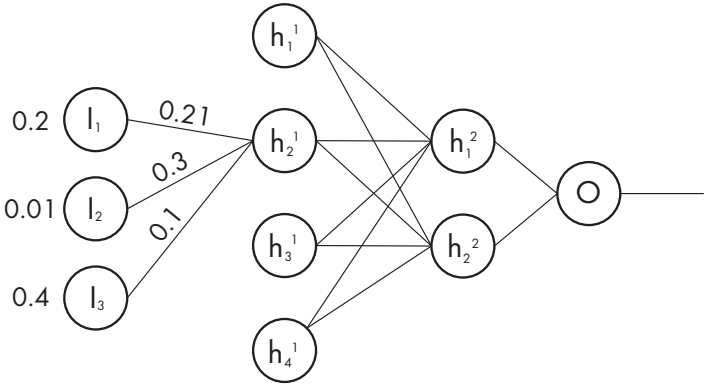

Figure 1.16 The weight values of the connections to h_2^1 (as in fig. 1.15, simplified for visibility).

Equation Series B

$h_2^1 = \text{ReLu}(w_{2,1}{}^1 I_1 + w_{2,2}{}^1 I_2 + w_{2,3}{}^1 I_3)$

$= \text{ReLu}(0.21*0.2 + 0.3*0.01 + 0.1*0.4)$

$= \text{ReLu}(0.085)$

$= 0.085$

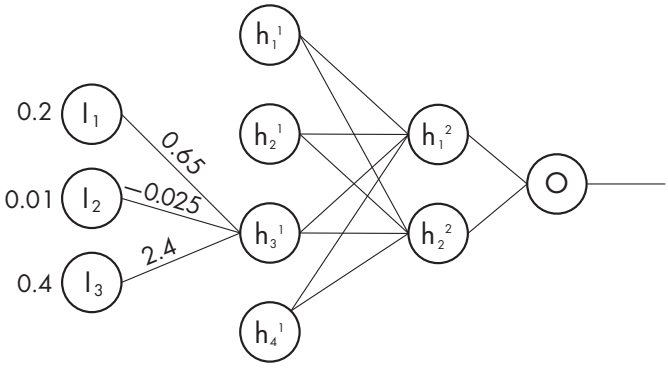

Figure 1.17 The weight values of the connections to h_3^1.

Equation Series C

$h_3^1 = \text{ReLu}(w_{3,1}{}^1 I_1 + w_{3,2}{}^1 I_2 + w_{3,3}{}^1 I_3)$

$= \text{ReLu}(0.65*0.2 + (-0.025)*0.01 + 2.4*0.4)$

$= \text{ReLu}(1.1)$

$= 1.1$

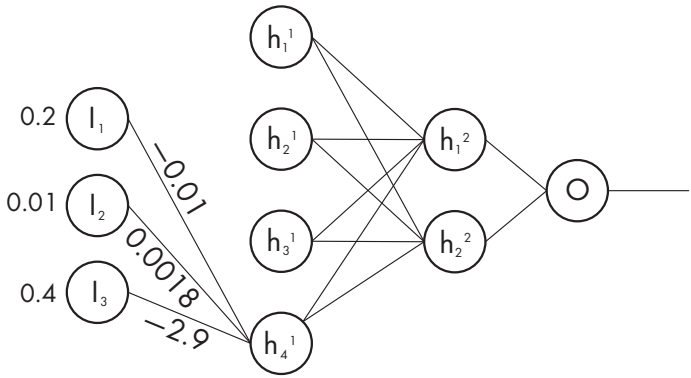

Figure 1.18 The weight values of the connections to h_4^1.

Equation Series D

$h_4^1 = \text{ReLu}(w_{4,1}{}^1 I_1 + w_{4,2}{}^1 I_2 + w_{4,3}{}^1 I_3)$

$= \text{ReLu}(-0.01*0.2 + 0.0018*0.01 + (-2.9)*0.4)$

$= \text{ReLu}(-1.16)$

$= 0$

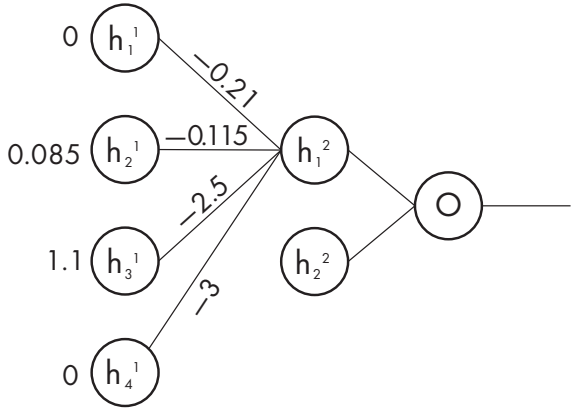

Figure 1.19 The inputs and weight values of the connections to h_1^2. Note that each input into layer h^2 is the output of h^1 [0, 0.085, 1.1, 0], which was calculated in the previous steps.

Equation Series E

$h_1^2 = \text{ReLu}(w_{1,1}^2 h_1^1 + w_{1,2}^2 h_2^1 + w_{1,3}^2 h_3^1 + w_{1,4}^2 h_4^1)$

$= \text{ReLu}(-0.21*0 + (-0.115)*0.085 + (-2.5)*1.1 + (-3)*0)$

$= \text{ReLu}(-2.75)$

$= 0$

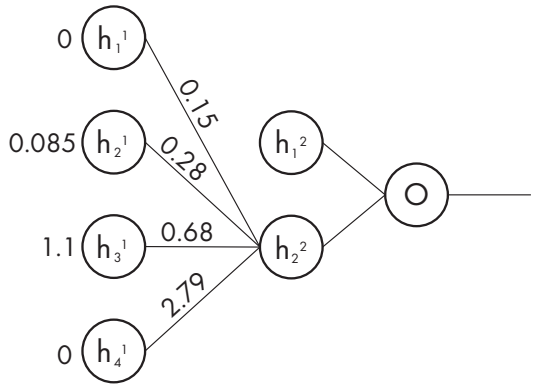

Figure 1.20 The inputs and weight values of the connections to h_2^2.

Equation Series F

$h_2^2 = \text{ReLu}(w_{2,1}^2 h_1^1 + w_{2,2}^2 h_2^1 + w_{2,3}^2 h_3^1 + w_{2,4}^2 h_4^1)$

$= \text{ReLu}(0.15*0 + 0.28*0.085 + 0.68*1.1 + 2.79*0)$

$= \text{ReLu}(0.77)$

$= 0.77$

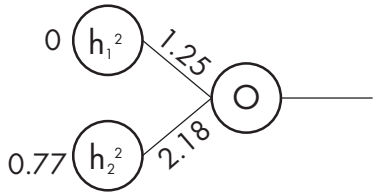

Figure 1.21 The output of h^2 as the input into the output layer, along with the connection weights to the output layer.

Equation Series G

$$O = \text{sigmoid}(w_{1,1}^{3} h_1^{2} + w_{1,2}^{3} h_2^{2})$$
$$= \text{sigmoid}(1.25*0 + 2.18*0.77)$$
$$= \text{sigmoid}(1.68)$$
$$= \frac{1}{1 + e^{-(1.68)}}$$
$$= 0.84$$

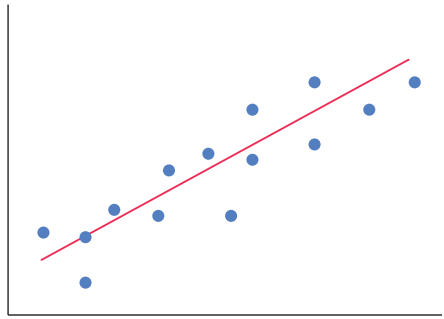

Figure 1.22 A best-fit line (*red line*) running through a set of data points (*blue dots*). We can use the best-fit line to predict the value for missing data samples.

Table 1.4 Feature Description of the Boston Housing Price Data Set

1	Per capita crime rate by town
2	Proportion of residential land zoned for lots over 25,000 sq. ft.
3	Proportion of nonretail business acres per town
4	Charles River dummy variable (1 if tract bounds river; 0 otherwise)
5	Nitric oxides concentration (parts per 10 million)
6	Average number of rooms per dwelling
7	Proportion of owner-occupied units built prior to 1940
8	Weighted distances to five Boston employment centers
9	Index of accessibility to radial highways
10	Full-value property-tax rate per $10,000
11	Pupil-teacher ratio by town
12	B – 1000(Bk – 0.63)^2 where Bk is the proportion of Black residents by town*
13	LSTAT: % lower status of the population
14	Median value of owner-occupied homes (in thousands of dollars)

**Note:* *Row 12 bears addressing, but to not break the flow of the current explanation, it will be addressed immediately following the end of this chapter.*

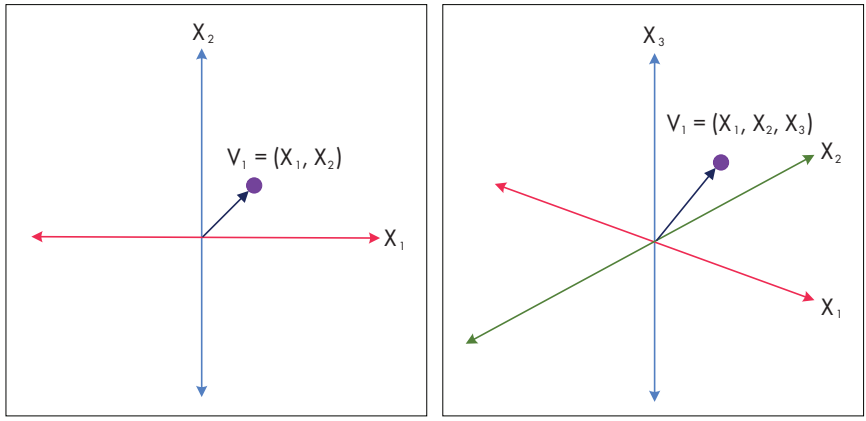

Figure 1.23 A vector in 2D space (*left*) and a vector in 3D space (*right*). A vector is just a point in some space with an arrow running from the origin to the point.

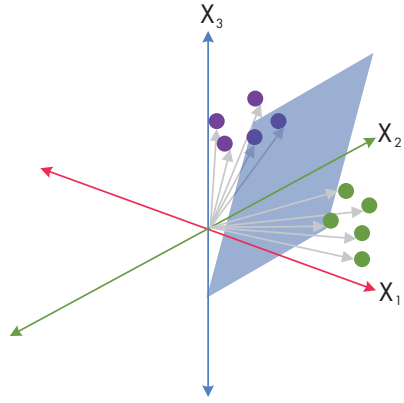

Figure 1.24 Two classes of vectors in a 3D space separated by a hyperplane. The vectors to the left of the plane (pointing to the *purple points*) belong to one class, and the vectors to the right of the plane (pointing to the *green points*) belong to a different class. Why is this useful? If we get a new sample and we do not know what class it belongs to, all we need to do is interpret it as a vector and see on which side of the plane it falls; then we can predict its class.

CHAPTER 2: HELLO, PANDA!

Figure 2.1 *Left,* an input image, Leonardo da Vinci's *Lady with an Ermine; center,* the result of an edge detection operation; *right,* the result of a Gaussian filter operation. Both operations were performed following the process described in figure 2.2.

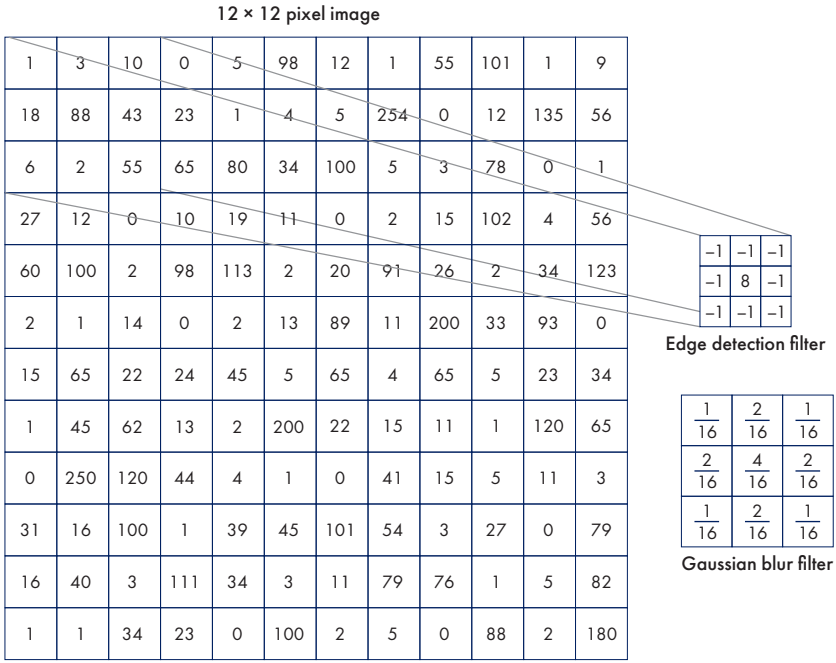

Figure 2.2 *Left,* a 12 × 12 pixel image with the associated value for each pixel; *above right,* an edge detection filter matrix; *below right,* a Gaussian blur filter matrix.

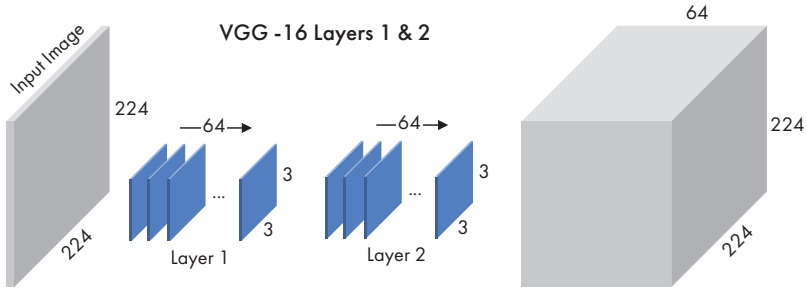

Figure 2.3 The VGG-16 neural network architecture consists of several convolutional layers, where each layer is a collection of 3 × 3 filter matrices. Layers 1 and 2 both contain collections of sixty-four 3 × 3 filter matrices. The output of these two layers is a volume of sixty-four images, called feature maps, the same size as the input image. *Author's rendering based on Karen Simonyan and Andrew Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition" (poster presented at the Third International Conference on Learning Representations, San Diego, CA, 2015), https://doi.org/10.48550/arXiv.1409.1556.*

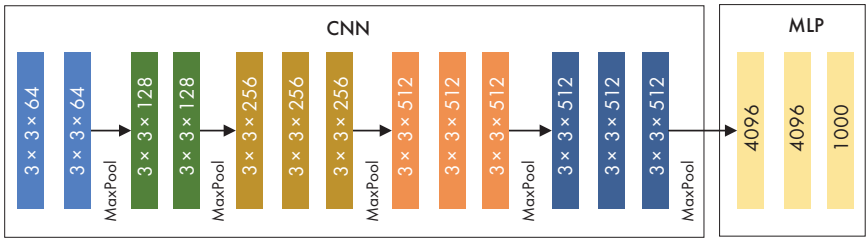

Figure 2.4 The VGG-16 neural network architecture is divided into a convolutional feature-extraction portion and an MLP classifier portion. The CNN section itself is divided into blocks of layers of equal numbers of filter matrices, with a MaxPool operation between blocks. The final output of this neural network as shown in the MLP section comprises 1,000 values. This neural network can classify objects into 1,000 different categories. *Author's rendering based on Karen Simonyan and Andrew Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition" (poster presented at the Third International Conference on Learning Representations, San Diego, CA, 2015), https://doi.org/10.48550/arXiv.1409.1556.*

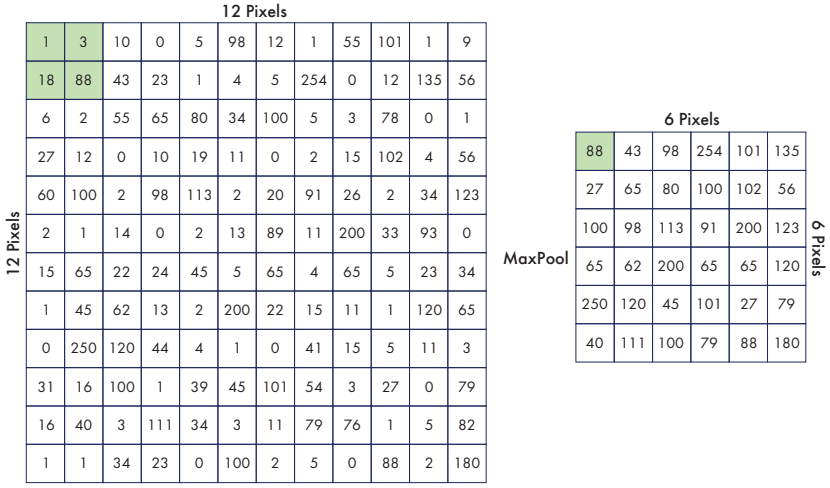

Figure 2.5 The MaxPool operation serves to reduce the size of the data that must be processed in each block. It is performed by selecting a 2 × 2 pixel window on the top left of the input image and then selecting the pixel with the highest value in that window. This pixel is kept and becomes the first pixel in the output image; the other pixels are discarded. We then slide the 2 × 2 pixel window to the right and continue processing the input image. In this example, our first 2 × 2 pixel window contains the following pixel values: 1, 3, 18, 88. The highest value is 88, so we keep this pixel and discard the rest (*right*). The next window contains: 10, 0, 43, 23. The highest value is 43, so we keep this pixel and discard the rest. Thus, we reduce the input image from 12 × 12 pixels to 6 × 6 pixels.

Table 2.1 Four Classes of Batteries Our Neural Network Must Identify

Class label	Class
0	9V
1	AAA
2	AA
3	Button

Table 2.2 Probability Outputs for Each Predicted Class

Predicted class	Prediction
0	0.98
1	0.01
2	0.007
3	0.003

Table 2.3 Probability Outputs Conveying a Neural Network's Lower Confidence Level on the Correct Classification

Predicted class	Prediction
0	0.025
1	0.65
2	0.32
3	0.005

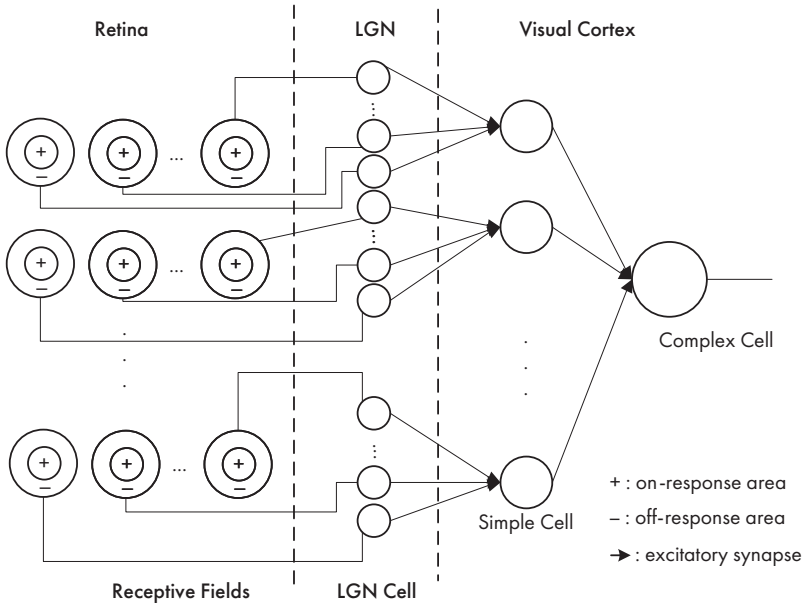

Figure 2.6 Hubel and Wiesel's hierarchy model. Information flows from the retina to the brain's visual cortex in a cat. It's hard not to see a strong resemblance to artificial neural networks. *Author's rendering based on Hubel and Wiesel 1962.*

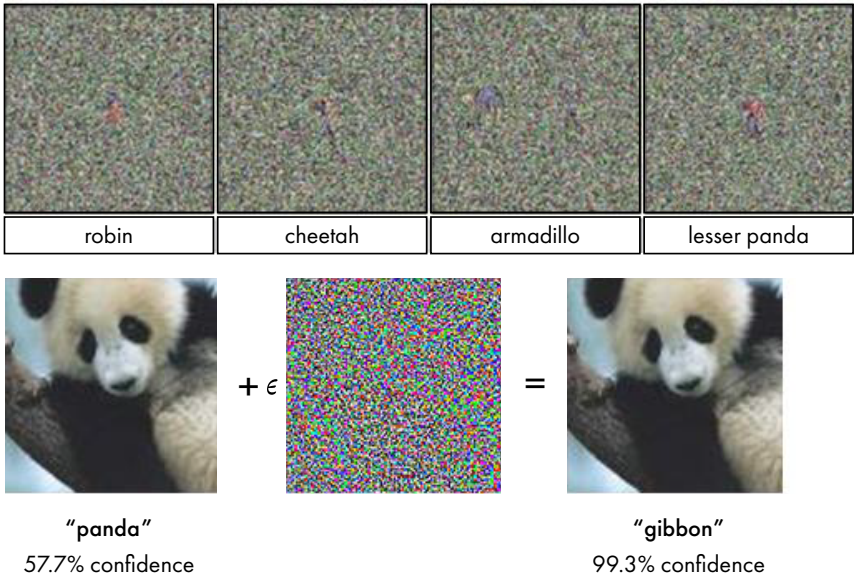

Figure 2.7 Two different kinds of artificial optical illusions created by different researchers that have been able to fool neural networks. *Top*, four examples of noisy-looking images that caused a well-trained neural network to predict "robin," "cheetah," "armadillo," and "lesser panda" for each sample. *Bottom*, an image of a panda: the trained neural network predicts that this is an image of a panda with 57.7 percent confidence; researchers then modify the image by adding a small amount of noise, resulting in an image that looks no different to us but causes the neural network to reclassify the image as a "gibbon" with 99.3 percent confidence. *Image of "panda"/"gibbon" from Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and Harnessing Adversarial Examples" (poster presented at the Third International Conference on Learning Representations, San Diego, CA, 2015), https://doi.org/10.48550/arXiv.1412.6572. Image of noisy "robin," "cheetah," "armadillo," and "lesser panda" from Anh Nguyen, Jason Yosinski, and Jeff Clune, "Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images," in* 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), *427–36 (Washington, DC: IEEE Computer Society, 2015).*

CHAPTER 3: ANSWERING AN AGE-OLD QUESTION

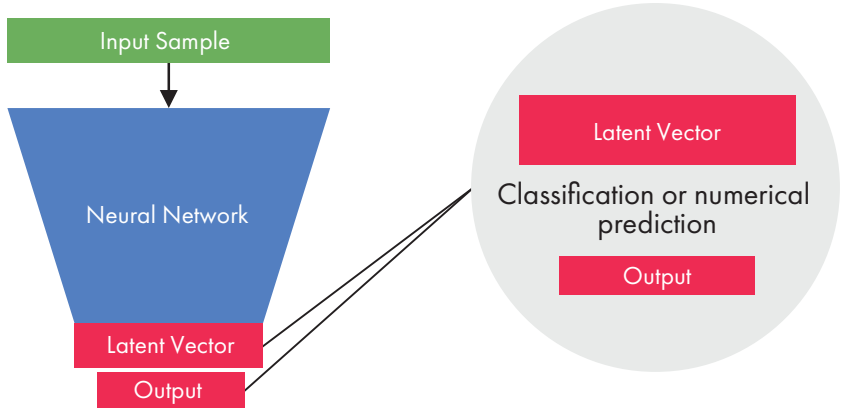

Figure 3.1 A neural network as a funnel of information. Information is input at the top and compressed through the network down to a latent vector representation. The output of the network is then typically calculated by performing linear regression (if predicting a continuous value) or logistic regression (if performing some classification).

Table 3.1 Distribution of Boys according to Height in a Survey

Height (m)	Number of boys
1.64	30
1.66	80
1.68	200
1.7	400
1.72	220
1.74	50
1.76	20

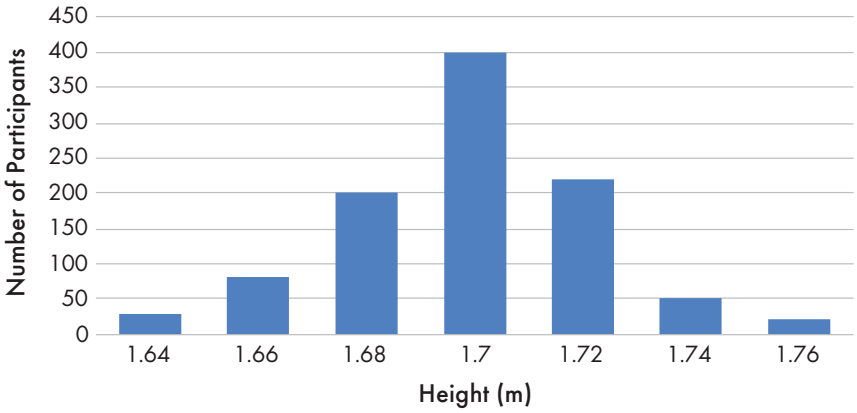

Figure 3.2 Histogram of one thousand participants in a survey, arranged according to height.

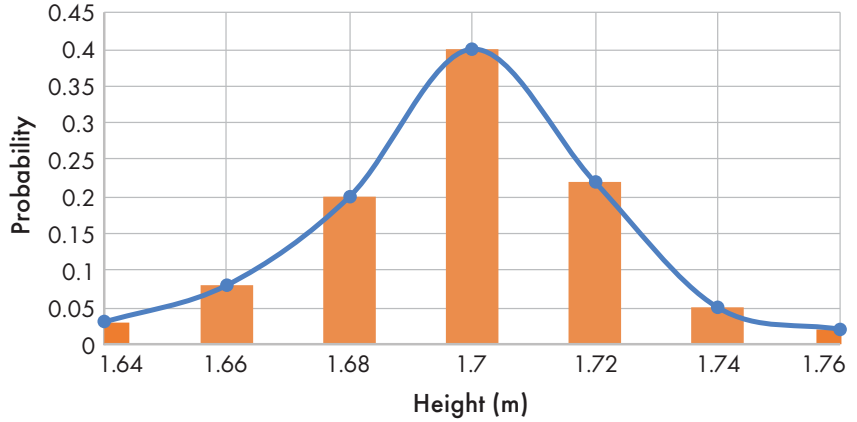

Figure 3.3 Probability distribution for each possible measurement (*orange bars*). The *blue line* shows the continuous probability for the height of the surveyed participants in the 1.64–1.76 m range.

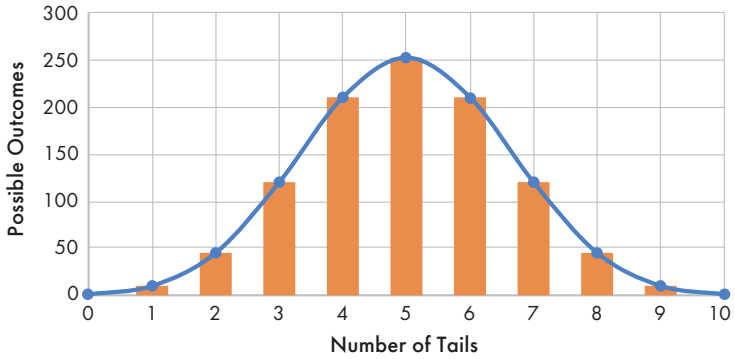

Figure 3.4 The probability distribution function for the coin-flipping experiment. The x-axis shows the number of tails we can get in an experiment comprising ten coin flips. The y-axis shows the probability of flipping any number of tails in ten trials.

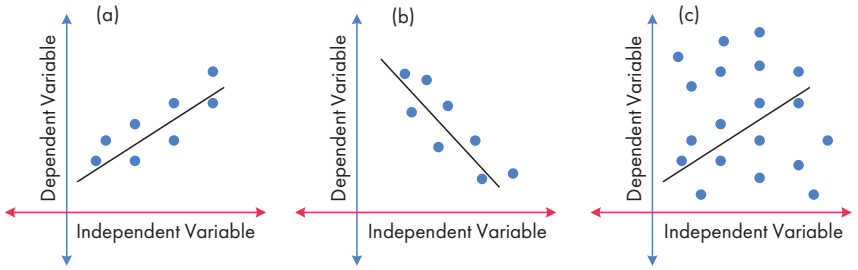

Figure 3.5 Three scatter plots with best-fit lines (*black lines*) running through the data points (*blue dots*). Subplot (*a*) shows a positive linear relationship between the dependent and independent variable. Subplot (*b*) shows a negative linear relationship between the dependent and independent variable. Subplot (*c*) shows data that does not exhibit a linear relationship between dependent and independent variables.

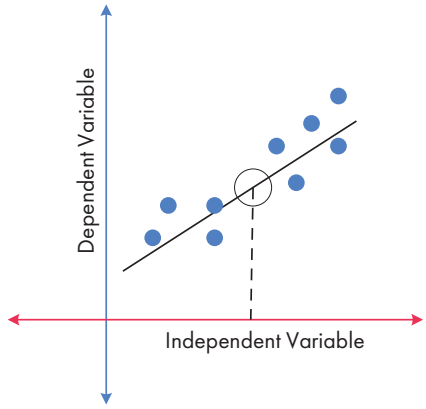

Figure 3.6 How a linear regression model (i.e., a best-fit line) can predict information for missing data in the data set. The *segmented line* indicates a value for the independent variable that does not correspond to any known value in our data set (visualized as a gap in the *blue dots*). The best-fit line can be used to approximate the missing information.

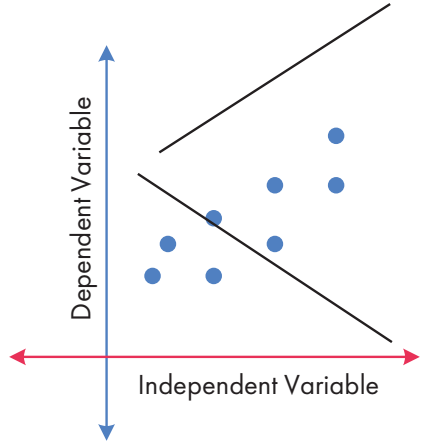

Figure 3.7 The two *black lines* are simply two possibilities among infinite ways to randomly place a line over a data set.

Equation 3

$$\hat{y} = \theta_1 x + \theta_0$$

Equation 4

$$L = \frac{1}{2}(h_\theta(x) - y)^2$$

Equation 5

$$\theta = \theta - \alpha \frac{\partial L}{\partial \theta}$$

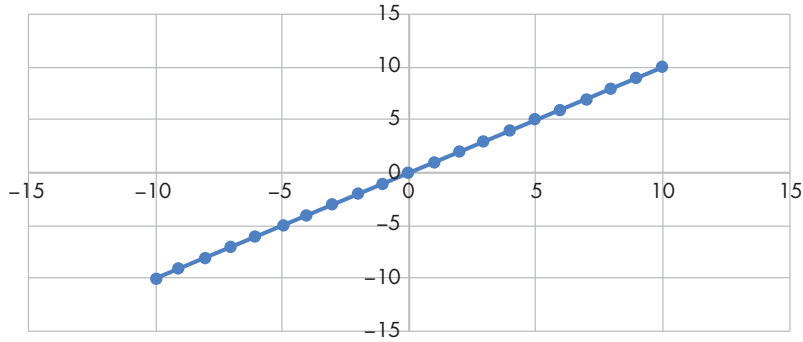

Figure 3.8 A plot of the line $y = x$.

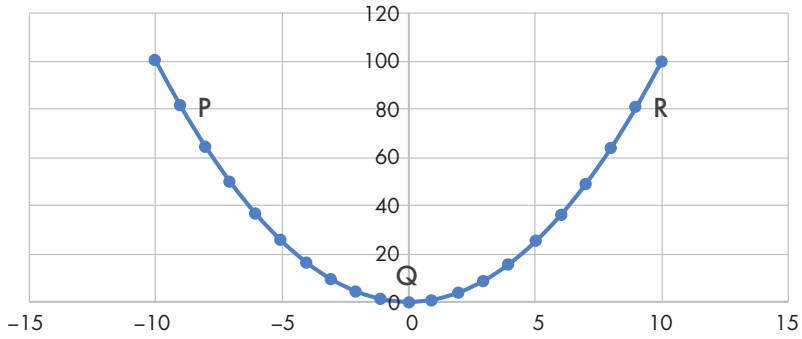

Figure 3.9 A plot of the line $y = x^2$.

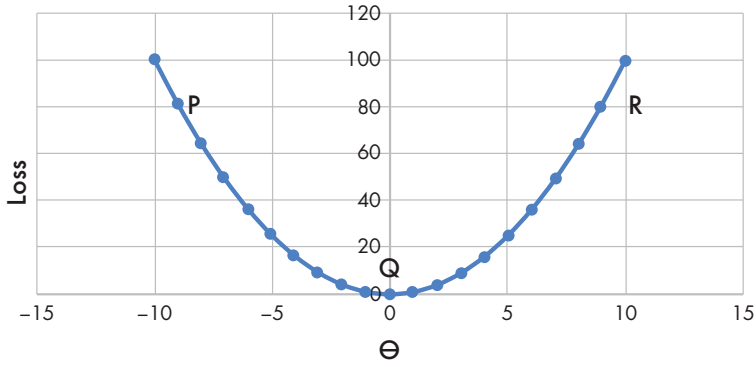

Figure 3.10 What a loss surface for $L = \frac{1}{2}(h_\theta(x) - y)^2$ might look like if the input data sample contains a single feature (i.e., x_1).

Equation 6

$$y = \theta_1 x + \theta_0$$

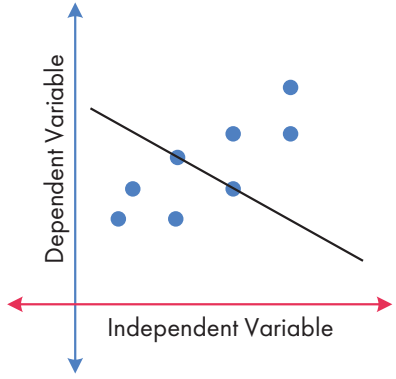

Figure 3.11 A line running through two random points in the data set. The *black line* is clearly not a best-fit line, yet it runs through two points.

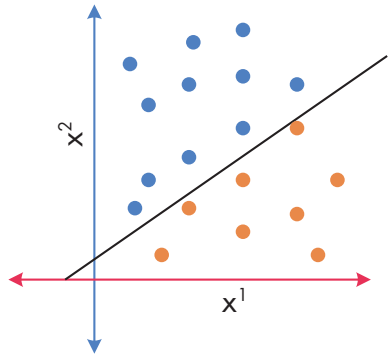

Figure 3.12 A logistic regression model separating two classes of samples (*blue dots* and *orange dots*) into different categories. In this figure, x_1 and x_2 are simply the two features describing our 2D data.

Equation 7

$$y = \frac{1}{1 + e^{-(\theta x)}}$$

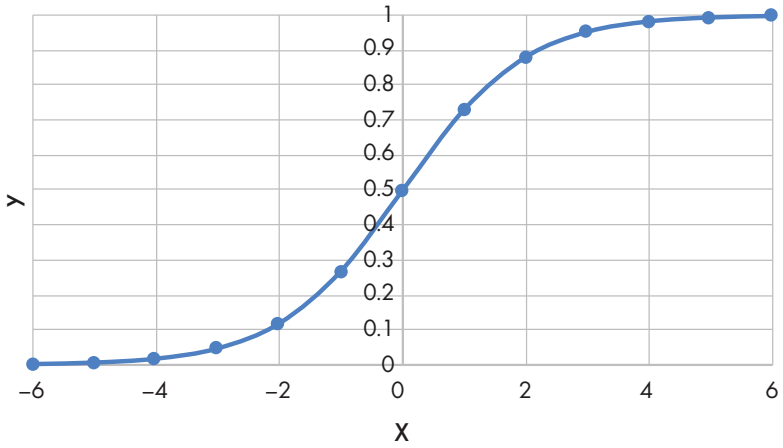

Figure 3.13 Possible values of y ranging between 0 and 1 for any input x for the logistic function, $\frac{1}{1 + e^{-(x)}}$.

Equation 8

$$L = -y \log\big(h_\theta(x)\big) - (1 - y) \log\big(h_\theta(x)\big)$$

Equation 9

$$L = -1 \log(1) - (1 - 1) \log(1) = 0$$

Equation 10

$$L = -1 \log(0.0001) - (1 - 1) \log(0.0001) = 4$$

Equation 11

$$\mathrm{P}(x) = p^x (1 - p)^{1 - x}$$

Equation 12

$$\mathrm{P}(y|x;\, \theta) = h_\theta(x)^y (1 - h_\theta(x))^{1 - y}$$

CHAPTER 4: INTELLIGENT DISCOURSE

Table 4.1 Predicting and Observing Recidivism in One Hundred Cases (First Hypothetical)

	Observed behavior	
Predicted behavior	**Suspect reoffended**	**Suspect did not reoffend**
Suspect reoffends	35	*5*
Suspect does not reoffend	20	40

Table 4.2 Predicting and Observing Recidivism in One Hundred Cases (Second Hypothetical)

	Observed behavior	
Predicted behavior	**Suspect reoffended**	**Suspect did not reoffend**
Suspect reoffends	35	20
Suspect does not reoffend	*5*	40

SOURCES

Works whose influence can be found in specific chapters are listed below.

Introduction: Living with Lions

Asher Hamilton, Isobel. "Elon Musk's Neuralink Wants to Embed Microchips in People's Skulls and Get Robots to Perform Brain Surgery." Impact Lab, Aug. 4, 2021. https://www.impactlab.com/2021/08/04/elon-musks-neuralink-wants-to-embed-microchips-in-peoples-skulls-and-get-robots-to-perform-brain-surgery/.

Reuters. "Maasai Now Track Lions instead of Killing Them." NBC News, Oct. 15, 2009. https://www.nbcnews.com/id/wbna33240556.

Yasukawa, Olivia, and Thomas Page. "Lion-Killer Maasai Turn Wildlife Warriors to Save Old Enemy." CNN, Feb. 8, 2017. https://www.cnn.com/2017/02/07/africa/maasai-tanzania-wildlife-warriors/index.html.

Polarization and Its Consequences

Amari, S., and M. A. Arbib, eds. *Competition and Cooperation in Neural Nets: Proceedings of the U.S.-Japan Joint Seminar Held at Kyoto, Japan February 15–19, 1982*. Lecture Notes in Biomathematics Series. Berlin: Springer, 1982. https://doi.org/10.1007/978-3-642-46466-9.

Burke, Robert E. "Sir Charles Sherrington's *The Integrative Action of the Nervous System*: A Centenary Appreciation." *Brain* 130, no. 4 (Apr. 2007): 887–94. https://doi.org/10.1093/brain/awm022.

Computer History Museum. "The Engines." The Babbage Engine. Accessed May 2, 2022. https://www.computerhistory.org/babbage/engines/.

Deng, J., W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei. "ImageNet: A Large-Scale Hierarchical Image Database." In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–55. Washington, DC: IEEE Computer Society, 2009. https://doi.org/10.1109/CVPR.2009.5206848.

Gardner, Martin. "Mathematical Games: The Fantastic Combinations of John Conway's New Solitaire Game 'Life.'" *Scientific American* 223 (October 1970): 120–23. https://www.scientificamerican.com/magazine/sa/1970/10-01/.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." *Communications of the ACM* 60, no. 6 (June 2017): 84–90. https://doi.org/10.1145/3065386. Originally published in the proceedings of the 2012 Conference and Workshop on Neural Information Processing Systems, *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, edited by F. Pereira, C. J. Burges, L. Bottou, and K. Q. Weinberger.

Large Movie Review Dataset. Accessed May 2, 2022. https://ai.stanford.edu/~amaas/data/sentiment/.

LeCun, Y., B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. "Backpropagation Applied to Handwritten Zip Code Recognition." Paper for AT&T Bell Laboratories, Holmdel, NJ, Sept. 1989. http://yann.lecun.com/exdb/publis/pdf/lecun-89e.pdf.

McCulloch, Warren, and Walter Pitts. "A Logical Calculus of Ideas Immanent in Nervous Activity." *Bulletin of Mathematical Biophysics* 5 (1943): 127–47. https://doi.org/10.1007/BF02478259.

Minsky, Marvin, and Seymour A. Papert. *Perceptrons: An Introduction to Computational Geometry*. Exp. ed. Cambridge, MA: MIT Press, 1987.

Ramón y Cajal, Santiago. *Texture of the Nervous System of Man and the Vertebrates*. Vol. 1. Translated and edited by Pedro Pasik and Tauba Pasik. Vienna: Springer, 1999.

Roberts, Eric. "History: The 1940's to the 1970's." Neural Networks (website). Accessed May 2, 2022. https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/History/history1.html.

———. "History: The 1980's to the Present." Neural Networks (website). Accessed May 2, 2022. https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/index.html.

Roberts, Siobhan. "The Lasting Lessons of John Conway's Game of Life." *New York Times*, Dec. 28, 2020. https://www.nytimes.com/2020/12/28/science/math-conway-game-of-life.html.

Sherrington, Charles Scott. *The Integrative Action of the Nervous System*. New Haven, CT: Yale University Press, 1906.

Swaine, M. R., and Paul A. Freiberger. "Analytical Engine." *Encyclopedia Britannica*, May 26, 2020. https://www.britannica.com/technology/Analytical-Engine.

Hello, Panda!

Elsayed, Gamaleldin, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alex Kurakin, Ian Goodfellow, and Jascha Sohl-Dickstein. "Adversarial Examples that Fool both Computer Vision and Time-Limited Humans." arXiv.org, May 22, 2018. https://arxiv.org/abs/1802.08195.

Great Learning Team and Avinash Thite. "Introduction to VGG16: What Is VGG16?" *Great Learning*, Oct. 1, 2021. https://www.mygreatlearning.com/blog/introduction-to-vgg16/.

Harvard University. "A Nobel Partnership: Hubel & Wiesel." Harvard Brain Tour. Accessed May 2, 2022. https://braintour.harvard.edu/archives/portfolio-items/hubel-and-wiesel.

Hubel, David H., and Torsten N. Wiesel. "Early Exploration of the Visual Cortex." *Neuron* 20 (March 1998): 401–12. https://doi.org/10.1016/S0896-6273(00)80984-8.

Katz, Bernard. "Stephen William Kuffler, 24 August 1913–11 October 1980." *Biographical Memoirs of Fellows of the Royal Society* 28 (Nov. 1982): 225–59. https://doi.org/10.1098/rsbm.1982.0011.

Lian, Yanbo, Ali Almasi, David B. Grayden, Tatiana Kameneva, Anthony N. Burkitt, and Hamish Meffin. "Learning Receptive Field Properties of Complex Cells in V1." *PLOS Computational Biology*, Mar. 2, 2021. https://doi.org/10.1371/journal.pcbi.1007957.

Nutan. "Deep Convolutional Networks VGG16 for Image Recognition in Keras." Medium, Aug. 28, 2020. https://medium.com/@nutanbhogendrasharma/deep-convolutional-networks-vgg16-for-image-recognition-in-keras-a4beb59f80a7.

Simonyan, Karen, and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." arXiv.org, Apr. 10, 2015. https://arxiv.org/abs/1409.1556.

University of Bonn. "'Math Neurons' Identified in the Brain: When Performing Calculations, Some Neurons Are Active when Adding, Others when Subtracting." ScienceDaily, Feb. 14, 2022. https://www.sciencedaily.com/releases/2022/02/220214121241.htm.

University of Oxford. Visual Geometry Group (homepage). Accessed May 2, 2022. https://www.robots.ox.ac.uk/~vgg/.

Answering an Age-Old Question

Asimov, Isaac. "Runaround." In *I, Robot*, 25–45. New York: Bantam, 1991.

Bryson, Arthur E., Jr., and Yu-Chi Ho. *Applied Optimal Control: Optimization, Estimation, and Control*. New York: Taylor & Francis, 1975.

Hintze, Arend. "What an Artificial Intelligence Researcher Fears about AI." *Scientific American*, July 14, 2017. https://www.scientificamerican.com/article/what-an-artificial-intelligence-researcher-fears-about-ai/.

LeCun, Y. "A Theoretical Framework for Back-Propagation." In *Proceedings of the 1988 Connectionist Models Summer School*, edited by D. Touretzky, G. Hinton, and T. Sejnowski, 21–28. Pittsburgh: Carnegie Mellon University, 1988. http://yann.lecun.com/exdb/publis/pdf/lecun-88.pdf.

Piper, Kelsey. "Why Elon Musk Fears Artificial Intelligence." Vox, Nov. 2, 2018. https://www.vox.com/future-perfect/2018/11/2/18053418/elon-musk-artificial-intelligence-google-deepmind-openai.

Russell, Stuart, and Peter Norvig, eds. *Artificial Intelligence: A Modern Approach*. 4th ed. Hoboken, NJ: Pearson, 2021.

Shead, Sam. "Elon Musk Says DeepMind Is His 'Top Concern' when It Comes to A.I." CNBC, July 29, 2020. https://www.cnbc.com/2020/07/29/elon-musk-deepmind-ai.html.

Intelligent Discourse

Agence-France Presse in Shanghai. "World's Best Go Player Flummoxed by Google's 'Godlike' AlphaGo AI." *The Guardian*, May 23, 2017. https://www.theguardian.com/technology/2017/may/23/alphago-google-ai-beats-ke-jie-china-go.

"AlphaGo—the Movie: Full Award-Winning Documentary." YouTube video posted by DeepMind on Mar. 13, 2020, 1:30:27. https://www.youtube.com/watch?v=WXuK6gekU1Y.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. "Courtroom Equations Wrongly Flagging Blacks as Future Criminals." *Tampa Bay Times*, May 23, 2016. https://www.tampabay.com/news/publicsafety/crime/courtroom-equations-wrongly-flagging-blacks-as-future-criminals/2278656/.

BBC. "Google AI Defeats Human Go Champion." BBC News, May 25, 2017. https://www.bbc.com/news/technology-40042581.

Bostrom, Nick. "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents." *Minds and Machines* 22, no. (May 2012): 71–85. https://doi.org/10.1007/s11023-012-9281-3.

Cummings, M. L. "Automation Bias in Intelligent Time Critical Decision Support Systems." AIAA Meeting Paper, AIAA 1st Intelligent Systems Technical Conference, Sept. 20–22, 2004, Chicago. https://doi.org/10.2514/6.2004-6313.

Dressel, Julia, and Hany Farid. "The Accuracy, Fairness, and Limits of Predicting Recidivism." *Science Advances* 4, no. 1 (Jan. 2018). https://doi.org/10.1126/sciadv.aao5580.

Gorner, Jeremy. "Chicago Police Use 'Heat List' as Strategy to Prevent Violence." *Chicago Tribune*, Aug. 21, 2013. https://www.chicagotribune.com/news/ct-xpm-2013-08-21-ct-met-heat-list-20130821-story.html.

Lin, Zhiyuan "Jerry," Jongbin Jung, Sharad Goel, and Jennifer Skeem. "The Limits of Human Predictions of Recidivism." *Science Advances* 6, no. 7 (Feb. 2020). https://doi.org/10.1126/sciadv.aaz0652.

Russell, Stuart. *Human Compatible: Artificial Intelligence and the Problem of Control*. Repr. ed. New York: Penguin Books, 2020.

Silver, David, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. "Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm." arXiv.org, Dec. 5, 2017. https://arxiv.org/abs/1712.01815.

Silver, David, Thomas Hubert, Julian Schrittwieser, and Demis Hassabis. "AlphaZero: Shedding New Light on Chess, Shogi, and Go." DeepMind, Dec. 6, 2018. https://www.deepmind.com/blog/alphazero-shedding-new-light-on-chess-shogi-and-go.

Simonite, Tom. "Algorithms Were Supposed to Fix the Bail System. They Haven't." *Wired*, Feb. 19, 2020. https://www.wired.com/story/algorithms-supposed-fix-bail-system-they-havent/.

Stroud, Matt. "Heat Listed." The Verge, May 24, 2021. https://www.theverge.com/c/22444020/chicago-pd-predictive-policing-heat-list.

U.S. Department of Health and Human Services. "Heart Disease and African Americans." Office of Minority Health, last modified Jan. 1, 2022. https://minorityhealth.hhs.gov/omh/browse.aspx?lvl=4&lvlid=19#:~:text=In%202018%2C%20African%20Americans%20were,their%20blood%20pressure%20under%20control.

Yudkowsky, Eliezer. "MIRI Announces New 'Death with Dignity' Strategy." *LessWrong*, Apr. 1, 2022. https://www.lesswrong.com/posts/j9Q8bRmwCgXRYAgcJ/miri-announces-new-death-with-dignity-strategy.

Yudkowsky, Eliezer, Anna Salamon, Carl Shulman, Steven Kaas, Tom McCabe, and Rolf Nelson. *Reducing Long-Term Catastrophic Risks from Artificial Intelligence*. San Francisco: Singularity Institute, 2010. https://intelligence.org/files/ReducingRisks.pdf.

Završnik, Aleš. "Criminal Justice, Artificial Intelligence Systems, and Human Rights." *ERA Forum* 20 (2020): 567–83. https://doi.org/10.1007/s12027-020-00602-0.

From Chat Rooms to Chatbots

Bonifacic, I. "Alphabet's Isomorphic Labs Is a New Company Focused on AI-Driven Drug Discovery." Engadget, Nov. 4, 2021. https://www.engadget.com/alphabet-isomorphic-labs-announcement-193546886.html.

Coldewey, Devin. "Isomorphic Labs Is Alphabet's Play in AI Drug Discovery." Tech Crunch, Nov. 4, 2021. https://techcrunch.com/2021/11/04/isomorphic-labs-is-alphabets-play-in-ai-drug-discovery.