# Data-Centric AI w/ Deep Learning

David Ruby

5/5/2023

## I. Introduction

What is Data-Centric Artificial Intelligence?

"Data-centric AI encompasses methods and tools to systematically characterize, evaluate, and monitor the underlying data used to train and evaluate models. At the ML pipeline level, this means that the considerations at each stage should be informed in a data-driven manner. We term this a data-centric lens. Since data is the fuel for any ML system, we should keep a sharp focus on the data, yet rather than ignoring the model, we should leverage the data-driven insights as feedback to systematically improve the model." [1]

This view of Artificial Intelligence where models are driven more by data than algorithms is only possible in the age of deep learning. Before diving more into this new focus on data, we will need to take a closer look at changes to AI due to Deep Learning.

## II. Deep Learning

Until the late 2000's deep neural nets were not able to shine against other machine learning methods. This changed in the late 2009-2010 with several simple but important algorithmic improvements: activation functions, weight initialization schemes, optimization schemes. With these improvements nets of 10 or more layers, allowing deep learning to start to shine. In 2014, 2015, 2016, further improvements introduced with batch normalization, residual networks, and depthwise separable convolutions. Now arbitrarily complex networks are possible. [2]

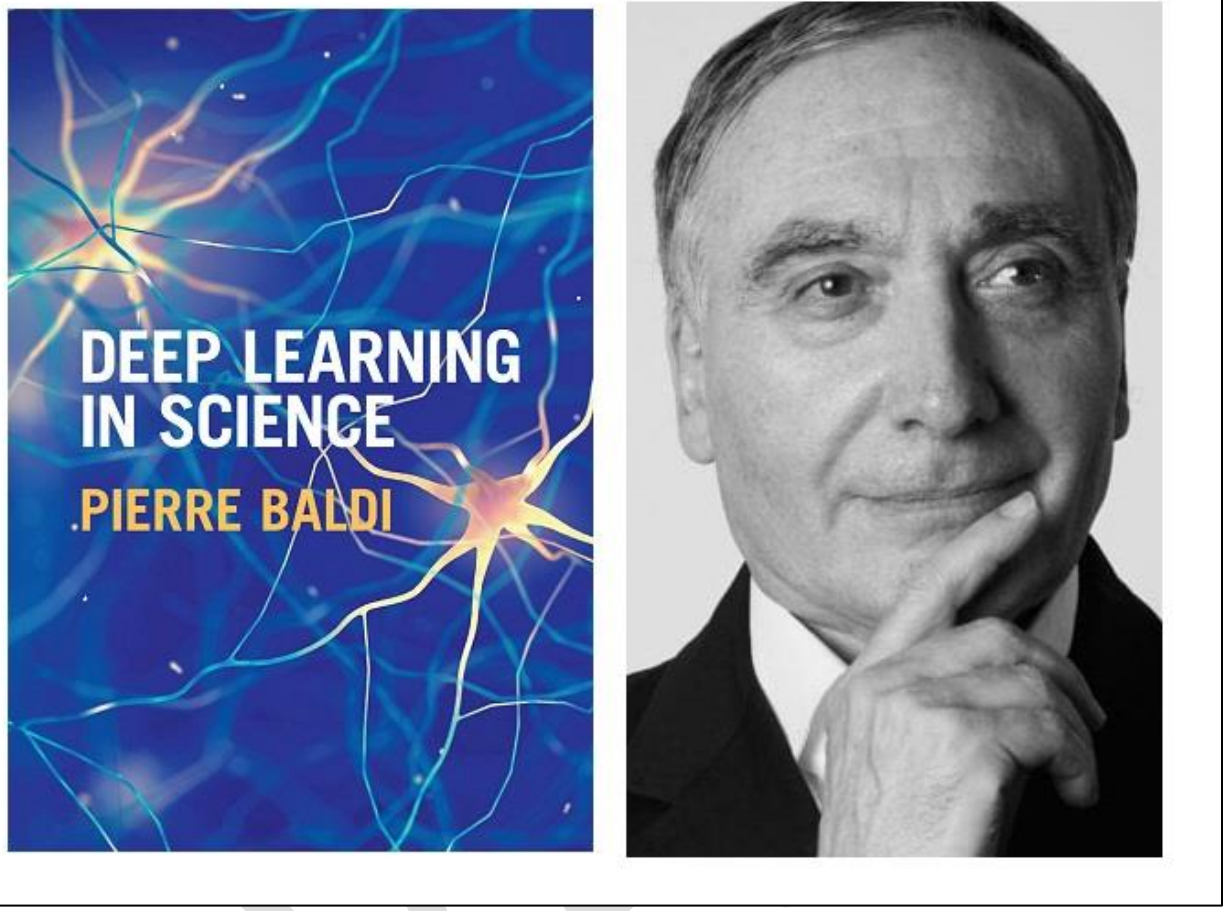


*Figure 1 Deep Learning & Neuroscience [3]*

Although many researchers approach Deep Learning purely from an algorithmic perspective, some researchers do emphasize the connection to biological neurological systems. Dr. Pierre Baldi with his Deep Learning text encourages researchers to understand these ties. [3]

## A. Neurons and Deep Learning

Deep learning has its root in traditional machine learning, which in turn has multiple roots. One root definitely includes the work from neurobiology in understanding brain functioning, and the basic unit the neuron. The perceptron, and its perceptron learning algorithm has always been an important part of machine learning history. [4] Later Yann LeCun took inspiration from animal visual cortex architecture such as the early work of Hubel & Wiesel with cats to with his work with hand-written character recognition. [5] [6]

# III. References


[1] N. Seedat, F. Imrie and M. van der Schaar, *DC-Check: A Data-Centric AI checklist to guide the development of reliable machine learning systems,* arXiv, 2022.

[2] F. Chollet, Deep Learning with Python, Second Edition, Manning, 2021.

[3] P. Baldi, Deep Learning in Science, Cambridge University Press, 2021.

[4] S. J. Russell, S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, Pearson, 2020.

[5] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-Based Learning Applied to Document Recognition," in *Proceedings of the IEEE*, 1998.

[6] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of Physiology,* vol. 160, pp. 106-154, 1962.

[7] Y. Zhang, W. Han, J. Qin, Y. Wang, A. Bapna, Z. Chen, N. Chen, B. Li, V. Axelrod, G. Wang, Z. Meng, K. Hu, A. Rosenberg, R. Prabhavalkar, D. S. Park, P. Haghani, J. Riesa, G. Perng, H. Soltau, T. Strohman, B. Ramabhadran, T. Sainath, P. Moreno, C.-C. Chiu, J. Schalkwyk, F. Beaufays and Y. Wu, *Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages,* arXiv, 2023.

[8] D. Zha, Z. P. Bhat, K.-H. Lai, F. Yang and X. Hu, *Data-centric AI: Perspectives and Challenges,* arXiv, 2023.

[9] J. D. Ullman and J. Widom, A First Course in Database Systems, Pearson/Prentice Hall, 2008.

[10] K. P. Murphy, Probabilistic Machine Learning: An Introduction, MIT Press, 2022.

[11] M. Mohri, A. Rostamizadeh and A. Talwalkar, Foundations of Machine Learning, second edition, MIT Press, 2018.

[12] C. Meisel and K. A. Bailey, "Identifying signal-dependent information about the preictal state: A comparison across ECoG, EEG and EKG using deep learning," *EBioMedicine,* vol. 45, p. 422–431, 2019.

[13] D. Koller and N. Friedman, Probabilistic Graphical Models: Principles and Techniques, MIT Press, 2009.

[14] I. Goodfellow, Y. Bengio and A. Courville, Deep Learning, MIT Press, 2016.

[15] E. Ghirardini, G. Sagona, A. Marquez-Galera, F. Calugi, C. Navarron, F. Cacciante, S. Chen, F. Di Vetta, L. Dadà, R. Mazziotti, L. Lupori, E. Putignano, P. Baldi, J. Lopez-Atalaya, T. Pizzorusso and L. Baroncelli, "Cell-specific vulnerability to metabolic failure: the crucial role of parvalbumin expressing neurons in creatine transporter deficiency," *Acta Neuropathologica Communications,* vol. 11, March 2023.

[16] A. Geron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, 2nd ed., O'Reilly Media, Inc., 2019.

[17] D. P. Friedman, A. Mendhekar, Q. Su, G. L. Steele and P. Norvig, The Little Learner: A Straight Line to Deep Learning, MIT Press, 2023.

[18] E. Charniak, Introduction to Deep Learning, MIT Press, 2019.

[19] E. Alpaydin, Introduction to Machine Learning, MIT Press, 2014.