HHS Public Access

Author manuscript

*Spine J*. Author manuscript; available in PMC 2022 October 01.

Published in final edited form as:

*Spine J*. 2021 October ; 21(10): 1606–1609. doi:10.1016/j.spinee.2020.08.012.

# Sharpening the Resolution on Data Matters: A Brief Roadmap for Understanding Deep Learning for Medical Data

**Allen Schmaltz, PhD**[1], **Andrew L. Beam, PhD**[1,2]

[1]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA

[2]Department of Biomedical Informatics, Harvard Medical School, Boston, MA

**Keywords**

machine learning; artificial intelligence; big data; natural language processing

## Introduction

Statistics has historically been the workhorse for analyzing clinical trials, building prediction models, and analyzing most types of medical data. However, recently large amounts of available healthcare data and increasing computing power have enabled techniques from the field of machine learning to play an ever larger role in clinical applications. The fields of machine learning and statistics in fact have many things in common, but they can appear at first glance to be quite different given the way each community describes its goals and techniques (Beam and Kohane 2018). This false sense of difference is further heightened because much of the recent work in machine learning has been described under the umbrella term of "artificial intelligence" (AI). It should be noted that AI refers to a *goal* (i.e. computers that behave "intelligently") and does not in itself describe *a method to achieve that goal*. Much of the recent and rapid progress *towards* the goal of medical AI has been enabled by advancements in the subfield of machine learning known as *deep learning* (LeCun, Bengio, and Hinton 2015; Hinton 2018; Schmidhuber 2015). However, beyond the AI hyperbole, the successes afforded by the use of deep learning to date have been more modest and narrow in high-risk settings, such as medicine, but are nonetheless likely to expand in coming years (Topol 2019; Ghassemi et al. 2018), though many challenges remain (Beam, Manrai, and Ghassemi 2020). The relevant literature stretches multiple decades and is rapidly growing, and we provide here a high-level overview of deep learning, with an emphasis on medical applications and natural language data. We offer a key insight for applications of deep learning in medicine, a dichotomy that presents both challenges and opportunities for real-world applications: Deep learning models can amplify biases and other



The sponsors did not influence the work represented in this manuscript in any fashion

issues in the underlying data, but those same models can be leveraged to uncover data issues and patterns that might not be readily discoverable via more parsimonious models.

## Historical Redux: The Old New

Deep learning is broadly defined as the use of *artificial neural networks*, a class of machine learning models loosely inspired by how biological brains are believed to process information. The core ideas underpinning modern deep learning are not new. The early "connectionist" work by researchers in the 1940's and 1950's led to the McCulloch-Pitts neurons (McCulloch and Pitts 1943) and the perceptron (Rosenblatt 1958), among other developments, which underpin current notions of neural networks. Interestingly, this line of work, with which we can also group the work of mathematicians such as Norbert Wiener in the 1940's and earlier on feedback systems and related areas, was subsequently largely overshadowed through at least the 1990's by symbolic/logic-based approaches, championed by John McCarthy and others. This latter line of work emerged from the 1956 Dartmouth Workshop, out of which came the term "artificial intelligence" (McCarthy 1988). Work on neural networks did, however, continue in the latter half of the 20th century by a small, yet dedicated group of researchers, leading to core architectures and learning approaches used today (Schmidhuber 2015; LeCun, Bengio, and Hinton 2015). However, it is not until the more recent decade, with the emergence of powerful graphics processing units (GPUs), which are particularly amenable to training modern neural network models with large amounts of data, that deep learning has become the dominant paradigm for large data settings.

## What distinguishes deep learning from statistics?

Due to the convergence of varying academic fields and subdisciplines, there is a large amount of overlapping, inconsistent, oblique, and otherwise confusing terminology in the field of deep learning. The distinctions between "deep learning", "machine learning", and "statistics", as well as the more amorphous, catch-all area of "artificial intelligence", are themselves blurry, but there are still meaningful distinctions that can be made. Most successful deep learning projects leverage vast amounts of data and enormous amounts of computing power to build very complicated (i.e., high parameter) models. In contrast, techniques from statistics have traditionally used less complex models to analyze studies with more modest amounts of available data, and the goal of models is often quite different (Breiman 2001). Most statistical models are designed to yield interpretable quantities that provide information about the effect or an association between the variable and the outcome of interest, whereas most deep learning models are primarily focused on prediction and in general lack interpretable quantities such as these.

For example, a typical study in Orthopaedics might involve the study of patients (on the order of hundreds or thousands), with a dozen or so covariates and a single outcome. An analysis of such data would likely use some variant of linear regression or logistic regression depending on the characteristics of the data, available information, etc.---i.e., the standard concerns of applied statistics. Importantly, in this scenario the use of deep learning may not be particularly advantageous, or even advisable to consider, since the additional complexity

may not be necessary to learn the patterns in small datasets (Christodoulou et al. 2019), with the additional disadvantage that a deep learning model might "overfit" the data and result in a model that fails to generalize to new data. Such "small N" scenarios with only a few variables are drastically different from the context in which deep learning models have had their most convincing successes. For example, deep learning models built to understand natural language data are routinely built using *millions* or *billions* of data points (e.g., a large sample of sentences crawled from the internet), with each covariate, a word in English, for example, itself existing in a very high-dimensional space, with long-distance, time-varying dependencies. It is this type of scenario (and analogously for images and video input) that deep neural networks have achieved considerable advances in recent years. It is, thus, not the case that deep learning is a good fit for all problems in medicine and healthcare.

## When is deep learning the preferred approach in medical settings?

Deep learning does not obviate the need for the basic principles of statistics, but there are some rules-of-thumb for using deep learning in healthcare applications. Concretely, when the data provided to the model consists of text, images, and/or videos, deep neural networks will typically be more effective than traditional regression models, assuming there is a sufficiently large amount of such data available. In these scenarios, deep learning models are able to automatically encode complex dependencies that can be challenging for human annotators to articulate and label. Conversely, the relative advantage of deep learning on "tabular data" (e.g., data that easily fits into a spreadsheet's rows and columns) has been less convincing, and thus traditional statistical models might be preferred in these settings. Relatedly, as noted further below, deep learning can also be used in analysis settings where the primary goal is to discover unknown patterns, or to verify patterns, in high-dimensional data. Finally, it should also be noted that deep learning approaches are largely not yet ready for day-to-day clinical deployment with patients (Beede et al. 2020), but can be useful for medical researchers analyzing data in less-critical scenarios.

## Natural Language Processing: Signal from Text

A considerable amount of data in medicine is available in unstructured text from electronic health records and related sources. Natural Language Processing (NLP), the study of language by computational means and the resulting applications (Jurafsky and Martin 2008), for which deep learning has become the primary approach within the last five years, can be used for analysis and prediction in the context of such data. The types and variations of models that are commonly used abound and are evolving, but roughly coalesce around a handful of base approaches, each of which has a rich history (Schmidhuber 2015; LeCun, Bengio, and Hinton 2015) and which can be combined together. With such models, NLP has moved from explicitly extracting covariates or features from the data by hand, to instead imposing generic constraints on the model which can automatically discover useful features in the data, a common evolution observed across subfields of modern machine learning (Sutton 2019). Models currently used in NLP research for medicine include Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber 1997), which can model temporal dependencies, a lens with which we can view the left-to-right dependencies in English, for example. LSTM networks improve the stability of earlier recurrent neural

network (RNN) models (e.g., Elman 1990), enabling them to train on more complicated data. Bi-directional LSTMs (Graves and Schmidhuber 2005) enable learning left and right dependencies, providing additional context useful for encoding language. In addition to these sequential models, a particular type of feed-forward network, convolutional neural networks (CNNs) (Fukushima 1980, 2013; LeCun et al. 1990 inter alia), which leverage a mathematical operation particularly well suited to image data known as a "convolution", have also been widely used.

A third commonly used architecture is the Transformer (Vaswani et al. 2017), which makes use of self-attention (Parikh et al. 2016). Transformers have found particular use in training very high parameter models, which can be trained in an "unsupervised" manner on natural language text by predicting the next word in a sentence given the previous words (Radford et al. 2019; Brown et al. 2020), or by randomly masking a word and having the model "fill in the blank" (Devlin et al. 2018). This unsupervised "pre-training" serves as an initial step to learn useful patterns (i.e., the structure and distribution of the observed language), after which domain-specific tasks can be trained with (typically much) smaller labeled datasets. This pattern of "pre-training" deep learning models on generic tasks before "fine-tuning" them towards a specific one has yielded similar advancements in computer vision and signal processing.

## Bias and Safety Concerns

Neural networks can amplify societal biases already contained in data sets, and models are dependent upon the data on which they are trained. Additionally, since we do not always have labels for the particular quantities of interest, it can be necessary to use proxies. Independent of the type of model, labels, such as those relating to cost, used as proxies for health needs or outcomes, have the potential to introduce biases (Obermeyer et al. 2019). In NLP applications, often we may have a large amount of text, but only a limited number of human annotated labels for our quantities of interest. For example, with EHR data, we may only have ICD-9 or ICD-10 diagnosis codes, which are typically used for billing, but we might want to use these codes for other types of tasks. However, these codes are often a poor surrogate for the true clinical state of a patient (Agniel, Kohane, and Weber 2018), and training a deep learning model to predict them may not yield a clinically relevant model. In practice, creating a dataset with enough high-quality labels for a deep learning model may be an insurmountable bar for many real applications.

## Data and Model Checks: Introspecting Inference

Perhaps counter-intuitively, deep neural models can themselves be leveraged to identify biases and other data issues. Deep learning models can be used to "introspect" a given dataset to identify possible errors or biases in a way that would be difficult with other approaches in high-dimensions. Using imputation-style losses, such as predicting censored parts of the input (e.g., randomly masking the identity of certain words, and then training a model to predict such words), we can pre-train networks on large amounts of unlabeled data. We can then use exemplar auditing (Schmaltz and Beam 2020) to introspect our predictions. This can be used to identify patterns in the data at resolutions more fine-grained than our

available labels, and it can also be used defensively to identify model errors or issues with the underlying training data, including with regard to protected attributes (Chen, Johansson, and Sontag 2018). This line of work, which is at the intersection of many fields in machine learning, is an active line of research that aims to improve the safety and reliability of deep learning by strategically auditing the manner in which a model is using the data it has been given.

## The need for humility with deep learning applications in medicine

The aforementioned models can be useful to medical researchers, but the distance to deployment in most clinical settings remains significant. Even with the various available tools now at our disposal, the extant state of deep learning is such that careful field studies are needed when deploying any algorithm. Translating existing research to clinical applications involves technical issues in machine learning, such as accounting for differences between training and test distributions, but it also involves logistical factors and assessing human preferences for computer interfaces. The deployment of EHRs has been problematic for caregivers (Gawande 2018), and a similar trajectory is possible with a naive application of deep learning algorithms. Nonetheless, the potential benefits of learning from large-scale, high-dimensional datasets is such that the effort is a worthwhile one. The solution is for clinicians, machine learning researchers, and statisticians to work together to advance progress toward real-world applications of deep learning.

## Acknowledgments

Funding Disclosures:

Dr. Beam was supported by awards from the NIH NHLBI (award #: 7K01HL141771) and the NIH NINDS (award #: 1R61NS113341)

Dr. Schmaltz was supported by an award from the NIH NINDS (award #: 1R61NS113341)

## References

Agniel Denis, Kohane Isaac S., and Weber Griffin M.. 2018. “Biases in Electronic Health Record Data due to Processes within the Healthcare System: Retrospective Observational Study.” BMJ, k1479. [PubMed: 29712648]

Beam Andrew L., and Kohane Isaac S.. 2018. “Big Data and Machine Learning in Health Care.” JAMA: The Journal of the American Medical Association 319 (13): 1317–18. [PubMed: 29532063]

Beam Andrew L., Manrai Arjun K., and Ghassemi Marzyeh. 2020. “Challenges to the Reproducibility of Machine Learning Models in Health Care.” JAMA: The Journal of the American Medical Association, 1. 10.1001/jama.2019.20866.

Beede Emma, Baylor Elizabeth, Hersch Fred, Anna Iurchenko Lauren Wilcox, Ruamviboonsuk Paisan, and Vardoulakis Laura M.. 2020. “A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy.” In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1–12. CHI ’20. New York, NY, USA: Association for Computing Machinery.

Breiman Leo. 2001. “Statistical Modeling: The Two Cultures.” Statistical Science: A Review Journal of the Institute of Mathematical Statistics 16 (3): 199–231.

Brown Tom B., Mann Benjamin, Ryder Nick, Subbiah Melanie, Kaplan Jared, Dhariwal Prafulla, Neelakantan Arvind, et al. 2020. “Language Models Are Few-Shot Learners.” arXiv[cs.CL], arXiv. http://arxiv.org/abs/2005.14165.

Chen Irene, Johansson Fredrik D., and Sontag David. 2018. “Why Is My Classifier Discriminatory?” In Advances in Neural Information Processing Systems 31, edited by Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, and Garnett R, 3539–50. Curran Associates, Inc.

Christodoulou Evangelia, Ma Jie, Collins Gary S., Steyerberg Ewout W., Verbakel Jan Y., and Van Calster Ben. 2019. “A Systematic Review Shows No Performance Benefit of Machine Learning over Logistic Regression for Clinical Prediction Models.” Journal of Clinical Epidemiology 110 (6): 12–22. [PubMed: 30763612]

Devlin Jacob, Chang Ming-Wei, Lee Kenton, and Toutanova Kristina. 2018. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.” arXiv[cs.CL]. arXiv. http://arxiv.org/abs/1810.04805.

Elman Jeffrey L.1990. “Finding Structure in Time.” Cognitive Science 14 (2): 179–211.

Fukushima Kunihiko. 1980. “Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position.” Biological Cybernetics. 10.1007/bf00344251.

———.2013. “Artificial Vision by Multi-Layered Neural Networks: Neocognitron and Its Advances.” Neural Networks: The Official Journal of the International Neural Network Society 37 (1): 103–19. [PubMed: 23098752]

Gawande Atul. 2018. “Why Doctors Hate Their Computers.” New Yorker 12. http://www.tramuntalegria.com/wp-content/uploads/2018/11/Why-Doctors-Hate-Their-Computers-The-New-Yorker.pdf.

Ghassemi M, Naumann T, Schulam P, and Beam AL. 2018. “Opportunities in Machine Learning for Healthcare.” arXiv Preprint arXiv. https://arxiv.org/abs/1806.00388.

Graves Alex, and Schmidhuber Jürgen. 2005. “Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures.” Neural Networks: The Official Journal of the International Neural Network Society 18 (5-6): 602–10. [PubMed: 16112549]

Hinton Geoffrey. 2018. “Deep Learning—A Technology With the Potential to Transform Health Care.” JAMA. 10.1001/jama.2018.11100.

Hochreiter S, and Schmidhuber J. 1997. “Long Short-Term Memory.” Neural Computation 9 (8): 1735–80. [PubMed: 9377276]

Jurafsky Daniel, and Martin James H.. 2008. “Speech and Language Processing: An Introduction to Speech Recognition, Computational Linguistics and Natural Language Processing.” Upper Saddle River, NJ: Prentice Hall.

LeCun Yann, Bengio Yoshua, and Hinton Geoffrey. 2015. “Deep Learning.” Nature 521 (7553): 436–44. [PubMed: 26017442]

LeCun Yann, Boser Bernhard E., Denker John S., Henderson Donnie, Howard RE, Hubbard Wayne E., and Jackel Lawrence D.. 1990. “Handwritten Digit Recognition with a Back-Propagation Network.” In Advances in Neural Information Processing Systems 2, edited by Touretzky DS, 396–404. Morgan-Kaufmann.

McCarthy John. 1988. “The Logic and Philosophy of Artificial Intelligence.” Kyoto Prize Lecture. https://www.kyotoprize.org/wp-content/uploads/2019/07/1988_A.pdf.

McCulloch Warren S., and Pitts Walter. 1943. “A Logical Calculus of the Ideas Immanent in Nervous Activity.” The Bulletin of Mathematical Biophysics 5 (4): 115–33.

Obermeyer Ziad, Powers Brian, Vogeli Christine, and Mullainathan Sendhil. 2019. “Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations.” Science 366 (6464): 447–53. [PubMed: 31649194]

Parikh Ankur P., Täckström Oscar, Das Dipanjan, and Uszkoreit Jakob. 2016. “A Decomposable Attention Model for Natural Language Inference.” arXiv [cs.CL]. arXiv. http://arxiv.org/abs/1606.01933.

Radford Alec, Wu Jeffrey, Child Rewon, Luan David, Amodei Dario, and Sutskever Ilya. 2019. “Language Models Are Unsupervised Multitask Learners.” OpenAI Blog 1 (8): 9.

Rosenblatt F1958. “The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain.” Psychological Review 65 (6): 386–408. [PubMed: 13602029]

Schmaltz Allen, and Beam Andrew. 2020. “Exemplar Auditing for Multi-Label Biomedical Text Classification.” arXiv [cs.CL]. arXiv. http://arxiv.org/abs/2004.03093.

Schmidhuber Jürgen. 2015. “Deep Learning in Neural Networks: An Overview.” Neural Networks: The Official Journal of the International Neural Network Society 61 (January): 85–117. [PubMed: 25462637]

Sutton Rich. 2019. “The Bitter Lesson.” 3 13, 2019. http://www.incompleteideas.net/IncIdeas/BitterLesson.html.

Topol Eric J.2019. “High-Performance Medicine: The Convergence of Human and Artificial Intelligence.” Nature Medicine 25 (1): 44–56.

Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, Gomez Aidan N., Kaiser Ł. Ukasz, and Polosukhin Illia. 2017. “Attention Is All You Need.” In Advances in Neural Information Processing Systems 30, edited by Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, and Garnett R, 5998–6008. Curran Associates, Inc.