Commentary on: Nightingale Health

Papyan V, Donoho DA, Donoho DL.

For: Proceedings of NBER Conference

**Introduction**

We congratulate Obermeyer and Mullanaithan for spotlighting a new era in medical research. Nightingale Health creates a platform where donations of health data can meetup with donations of ML researcher time and expertise, enabling research that's currently impeded because of legal and other limitations on the sharing of medical research data. Over time, this initiative can lead to entirely new customs in medical research and potentially to the unleashing of a great deal of research energy as traditional barriers to research are shattered.

If data donors and research practitioners will take advantage of this new platform en masse, we can envision many new research teams will form, leading to many advances in medical science and also, eventually, in human health care.

Our comments mirror the oral remarks we made in person in Toronto in September, and are divided into three sections, based on our three areas of expertise.

**Comments, Part I:** Common Task Framework (DLD)

The platform concept – join researchers with data – has been proven to work in field after field, across decades.

Today's favorite biometric technologies – fingerprint recognition, face recognition, retinal scanning, voice recognition, speech-to-text, were all developed using the so-called Common Task Framework (CTF) in DARPA-sponsored research from the mid 1980's to the mid-2000's. Under CTF, there is a publicly available shared dataset, a defined task and defined task performance metric, and researchers compete with each other to improve the performance metric. At set intervals, a leaderboard is updated and researchers get to see how others are doing, and thereby to understand who is currently "winning" and how far "off the pace" their own efforts are. Periodically, contest ceremonies are conducted, at which winners are proclaimed, and some sort of reward is bestowed on the winner.

As operationalized by DARPA in the 1990's, DARPA contracted with NIST under the leadership of Jonathon Phillips [3,4] to create a series of datasets that could be used in a series of annual biometrics challenges. Year after year, researcher performance improved on the various biometrics challenges, typically progressing so that after about 5-10 years of such annual

challenges, researchers had succeeded in constructing models that approached or exceeded human-level performance.

The same pattern of improvement until reaching human levels of performance was made in challenge after challenge, this held up regardless of the underlying modeling technology. In particular, much of this success predates deep learning models. Specific examples can be found in the excellent work of Isabelle Guyon and collaborators, on a variety of challenge problems, eg [5].

We maintain that the recent successes during 2012-2022 in machine vision and natural language processing based on deep learning and its elaborations, are merely a continuation of established patterns of successful CTF deployment, with new classes of datasets and a new class of models. From this viewpoint, the big event of the deep learning era was the bright idea of Fei Fei Li and collaborators to create the IMAGENET dataset and the ILSVRC competition, after that, progress in image recognition proceeded according to customary patterns.

In short we are saying that the ‘secret sauce’ of machine learning is the CTF rather than the specific technology. From this viewpoint, Yann Le Cun made a bigger impact by developing the MNIST dataset and publishing it, than by the specifics of any actual ML models he constructed for use with MNIST. Those early neural nets models have been superseded, but MNIST is still powering research papers today.

From our perspective, Nightingale Health seeks to bring this CTF `secret sauce’ into medical research, and we believe with the right CTF set-up, should be just as successful in medical research as it has been in other fields.

Let us emphasize a key element of the CTF that has always been present in instances where CTF has succeeded. That element is not yet in evidence with Nightingale Health, and including this element should be considered post haste.

The so-far missing ingredient is *reward;* for CTF to really work, there needs to be some benefit to the participants who reach the pinnacle of the leaderboard. For example, DARPA gave generous research awards to winners of the annual contest installation; while in the NETFLIX challenge the winning team split an award of US$ 1,000,000. The rewards don’t necessarily need to be purely monetary, but they need to be attention-getting and convincing. One can argue that the victory by the University of Toronto team in the ISLVRC 2012 led indirectly to outside financial rewards to the team members, who now enjoy high salaries at major research institutions.

Hence, Nightingale Health could consider ways to offer rewards. These could include not merely rewards for winning a challenge, but also for donating data and for developing a challenge.

To an audience of economists, this point must seem obvious; but among academics this type of essential, yet `crass’ observation, might be omitted, because seemingly inappropriate in a ‘high-minded discussion. Yet reward has been present (if sometimes implicitly so) in all the success stories of CTF we are aware of.

**Comments, Part II: The next decade's paradigm will not be the last decade's paradigm**

Over the last decade, many machine learning systems achieved unprecedented performance, sometimes superhuman, on an assortment of learning tasks arising in a variety of disciplines. In most cases, landmark results were fueled by exponential growth in proprietary image, text, and speech data available to "big tech" hegemons; in the arrival of the 'cloud', powered by hegemon constructions of massive data centers distributed globally; and, despite the failure of Moore's law at the CPU level, massive performance boosts in individual compute power following widespread adoption GPU technology. On top of this, there were numerous algorithmic and architectural advances introduced by the deep learning community [1].

Nightingale's platform aims to bring some of the last decade's innovations into the medical research context. This is exciting, and very promising. However, just as AI researchers who were not employed at hegemons faced roadblocks in exploiting data and compute during the last decade, there may well be, during the coming decade, obstacles for medical researchers and their ML collaborators to follow last decade's successful roadmap using Nightingale.

For the new platform to fully follow the roadmap of the last decade, Nightingale must provide to its users access to massive compute and massive data. In terms of compute, currently on Nightingale, a GPU hour costs $1. Ideally, Nightingale would allow researchers to use national clusters (e.g. Compute Canada) or university clusters (e.g. Stanford's Sherlock cluster) since these are significantly more affordable to academics. Ideally, Nightingale would also allow non-researchers to use cloud computing, which often provides GPUs at a lower price point (e.g., preemptable GPUs) and would also allow integration with other cloud computing services. Thus, what has worked previously in empirical ML seems to require that Nightingale broaden its compute strategy.

In terms of data, currently all datasets in Nightingale are labeled; Nightingale seems to be all-in on the paradigm of supervised learning. The supervised learning approach faces two important cosiderations. First, getting expert labels on data can be expensive; and second, there is a radically larger amount of unlabelled data, than there is of labelled data.

Since the beginning of this decade, thought leadership in AI has begun to challenge the `labelled-data' paradigm. In particular, recently-reported advances in *self-supervised learnin*g seem to show that training deep learning models no longer requires large labeled datasets. To the point for those interested in medicine, recent work [2] has shown the effectiveness of self-supervised learning as a pretraining strategy for medical image classification.

Thus, current thought leadership in empirical ML seems to require that Nightingale broaden its data strategy, to diversify its datasets and allow for productive exploitation of unlabeled data.

**Comments, Part III:** Medical Datasets and Artificial Intelligence

Nightingale provides an excellent playground for machine learning scientists to test new methods and algorithms on a well-curated corpus of medical datasets, and allows for the adoption of new datasets after their development. In that way, it may serve to highlight certain particular clinical use cases that are impactful (global ophthalmologic health; chest radiographs for tuberculosis, and so on) and focus the attention of the computational field on optimization for those use cases. However, the two largest challenges to the development of impactful artificial intelligence systems are data exfiltration [i.e. getting data out of the medical context they are curretly locked inside, such as a surgeon's video microscope, or ahospital EMR ]and operationalization of machine learning models. Platforms functioning in the intermediate step between data exfiltration and productization, such as Nightingale, have a unique opportunity to shape the course of the entire ecosystem.

First, by providing incentives for the development of large, high-quality datasets, and structures to manage terms of use, IP and compensation, Nightingale enable the use of machine learning to identify and address important clinical and public health challenges. Nightingale could provide incentives not only for ML researchers to vie for the top of the leaderboard (as in Part I), but also for medical stakeholders to develop and contribute deidentified, large unlabelled, sparsely labeled, or coarsely labeled datasets (as in Part II) in selected high impact areas. These incentives would reward the careful curation and dataset exfiltration from the HIPAA domain to the non-HIPAA domain, two principal challenges to the development and deployment of code against data. Incentives could be aligned with the mission of other larger organizations and given in the form of cash or model credits. For example, one could imagine a "Nightingale x PEPFAR Challenge" to generate datasets that support machine learning applications in HIV/AIDS research.

During dataset development, the sensitive nature of medical data and strict privacy regulations under HIPAA and GDPR generate significant challenges and costs. Incentives can help encourage and focus efforts to overcome these barriers across institutions. Furthermore, a transparent, standardized and open framework for formulating terms of use, intellectual property agreements and compensation would accelerate institutional participation. Nightingale's incentives could be aligned with the mission of other larger organizations, such as government agencies or healthcare providers, and could be given in the form of cash or model credits.

Second, Nightingale could provide structures and frameworks for model outputs and deployment inputs to promote impact within the public health and healthcare environments. One of the key challenges in the development and deployment of machine learning algorithms in healthcare is the need to ensure that the models are able to produce reliable, interpretable, and actionable outputs that can be used by healthcare providers and other stakeholders. This

involves not only developing and training the models, but also ensuring that the outputs are suitable for use in real-world settings.

To address this challenge, Nightingale should provide structures and frameworks to shape the outputs of machine learning models and receive return inputs from operational deployments. This could involve the development of tools and frameworks to support the deployment of machine learning models in healthcare settings, as well as mechanisms for monitoring and evaluating the performance and impact of these models.

For example, Nightingale could provide metrics for assessing the accuracy, precision, and interpretability of model outputs. Nightingale could also provide mechanisms for incorporating feedback from healthcare providers and other stakeholders into the model development process, such as mechanisms for soliciting and incorporating user feedback on model performance and outputs.

By providing these structures and frameworks, Nightingale can provide a machine learning playground that not only allows for academic development and skill building, but also promotes the development of machine learning algorithms that are actionable and impactful in real-world healthcare settings.

**Conclusion:**

The Nightingale platform has the potential to drive significant advances in medical research by bringing together researchers and data donors. By incorporating elements of the successful Common Task Framework, such as rewards and leaderboards, Nightingale can encourage participation and drive progress in the field. The ability to perform compute on local clusters would significantly accelerate adoption. In contrast to the supervised datasets currently available, methods which use unsupervised or semi-supervised data are gaining in popularity. Nightingale could provide incentives for the development of large, sparsely labeled datasets in selected high impact areas, and provide structures to shape the outputs of machine learning models and receive feedback from operational deployments. Positioned at the center of a medical artificial intelligence research workflow, Nightingale has the potential to shape the future of artificial intelligence in healthcare and accelerate innovation in global public health.

**References**

[1] Sevilla J, Heim L, Ho A, Besiroglu T, Hobbhahn M, Villalobos P. Compute trends across three eras of machine learning. arXiv preprint arXiv:2202.05924. 2022 Feb 11.

[2] Azizi, Shekoofeh, et al. "Big self-supervised models advance medical image classification." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[3] PJ Phillips, H Wechsler, J Huang, PJ Rauss The FERET database and evaluation procedure for face-recognition algorithms Image and vision computing, 1998

[4] HI Christensen, PJ Phillips Empirical evaluation methods in computer vision

World Scientific Publishers 2002

[5] I Guyon, S Gunn, A Ben-Hur, G Dror - Result analysis of the nips 2003 feature selection challenge Advances in neural information processing systems, 2004