

Perspective

Breaking away from labels: The promise of self-supervised machine learning in intelligent health

Dimitris Spathis,^{1,*} Ignacio Perez-Pozuelo,² Laia Marques-Fernandez,³ and Cecilia Mascolo¹¹Department of Computer Science and Technology, University of Cambridge, CB3 0FD Cambridge, UK²MRC Epidemiology Unit, School of Clinical Medicine, University of Cambridge, CB2 0SL Cambridge, UK³Addenbrooke's Hospital, Cambridge University Hospitals NHS Foundation Trust, CB2 0QQ Cambridge, UK*Correspondence: ds806@cam.ac.uk<https://doi.org/10.1016/j.patter.2021.100410>

THE BIGGER PICTURE Machine learning (ML) touches every area of science, and medicine especially is well poised to benefit the most. Hospital and nonhospital settings generate unprecedented amounts of data that if used correctly can unlock advances in new diagnostics and contribute to preventive medicine. The established paradigm of ML (supervised) requires the collection of input data (such as vitals or imaging) coupled with annotations from experts (such as indications of arrhythmia). New self-supervised models promise to do without annotations by using clever transformations of the input data only and achieve remarkable performance in an array of clinical tasks. This perspective gives a brief overview of the fundamental methodologies that enable these advances and discusses further challenges and opportunities.



Development/Pre-production: Data science output has been rolled out/validated across multiple domains/problems

SUMMARY

Medicine is undergoing an unprecedented digital transformation, as massive amounts of health data are being produced, gathered, and curated, ranging from in-hospital (e.g., intensive care unit [ICU]) to person-generated data (wearables). Annotating all these data for training purposes in order to feed to deep learning models for pattern recognition is impractical. Here, we discuss some exciting recent results of self-supervised learning (SSL) applications to high-resolution health signals. These examples leverage unlabeled data to learn meaningful representations that can generalize to situations where the ground truth is inadequate or simply infeasible to collect due to the high burden or associated costs. The most prominent bottleneck of deep learning today is access to labeled, carefully curated datasets, and self-supervision on health signals opens up new possibilities to eliminate data silos through general-purpose models that can transfer to low-resource environments and tasks.

INTRODUCTION

Underutilized medical data and the label gap

Medical data have the power to transform lives. Advances in the ways in which we collect, process, interpret, and use these data can be used to save lives and transform our society. Although the overwhelming majority of current medical research now focuses on clinical data (labs, imaging, vitals, etc.), the average person visits a doctor only around five times a year. Further, recent advances in wearable sensing and mobile computing, alongside their widespread and growing adoption, have created new pathways for the collection of health and well-being data outside of the laboratory and hospital settings, in a longitudinal fashion. These devices can be used to “fill the gaps” that are often found in traditional clinical data, opening up new research and com-

mercial directions for large-scale lifestyle monitoring and providing sources of truth in nondisease scenarios. For example, millions of people worldwide use such devices to track their physical activity and sleep,¹ with increasingly more sophisticated predictive capabilities and a wider range of sensors used to monitor these human behaviors and activities.

Concurrent to this self-monitoring revolution, seemingly disparate forces such as mature open-source scientific software libraries, easier data crowdsourcing and labeling, and the repurposing of specialized hardware (graphics cards) have enabled dramatic improvements in predictive modeling. Many machine learning (ML) tasks have achieved impressive performance, ranging from object recognition in images to outperforming experts in breast cancer screening. The common denominator in all these cases has been the curation of high-quality large datasets



that allow models to exploit latent patterns and subsequently generalize in real-world scenarios. However, especially in medicine, where erroneous predictions can have grave consequences, the roll-out and adoption of such systems have been met with resistance, mostly citing algorithm interpretability reasons.

Similar to how social networks learn our online behaviors, wearable and mobile devices monitor our activities in the real world. By tracking our sleep, steps, and eating and working habits, they create a holistic understanding of the most important components of our everyday health, until now only possible through subjective surveys. Although we recognize the value of such datasets, advances in ML for health and mobile sensing have not kept up with other areas. For example, over the last decade devices such as Fitbit or the iPhone have been collecting multi-modal sensor data at an unprecedented temporal resolution. However, effectively leveraging these datasets has presented many challenges, leading to these data being frequently overlooked for scientific and medical research.² Central to this problem is obtaining quality annotations and ground truth, which can be costly, burdensome, and at times, even impossible, given the granularity of these data. In this article, we discuss the potential of self-supervised methods toward bridging the label gap in biomedical data.

BREAKING AWAY FROM LABELS

Supervised learning: Reaching its limits

Deep supervised learning requires a decent amount of labels and samples in order to achieve good performance. (Arguing about the optimal dataset/label size is definitely a very empirical problem at this moment and that is why we do not quote exact numbers here, since it depends on the complexity of the given problem and the model. There is evidence that accuracy plateaus faster with supervised models compared to self-supervised ones, and the gains are mostly in the low-data regime.³) These manual labels—in the best scenario—are easier to obtain through crowdsourcing (cf. Imagenet), but in some cases, it is virtually impossible. For example, annotating wearable sensor timeseries for human activity recognition tasks *a posteriori* is not feasible without a video recording. On the other hand, given the fact that the amount of unlabeled data (e.g., all the images on the internet or the entire Fitbit user base) is considerably more heterogeneous and representative than some limited datasets, ongoing research and interest in this area have grown significantly. However, unsupervised learning is hard and, until recently, was less efficient than supervised learning. The first promising unsupervised studies in the area of health signals employed the successful paradigm of word2vec and reported results on a par with supervised models.⁴ However, the static vectors produced with these methods have limitations against context-specific representations.⁵

Self-supervised learning: The quest for the best data representation

A simple yet exciting emerging idea is to obtain labels “for free” from the input data (\mathbf{x}) through various transformations and, then, use conventional supervised objectives to predict them (\mathbf{v}_{SSL}). The representations obtained this way would be meaningful for downstream tasks with limited labeled data and linear classifiers

(see Figure 1). This has been coined self-supervised (or predictive) learning (SSL) due to learning the supervision directly from the data. (The terminology surrounding unsupervised and SSL is a bit blurry. Unsupervised learning is used for a wide range of models, ranging from autoencoders⁶ and Boltzmann machines⁷ to principal component and cluster analysis. SSL can be seen as a subset of unsupervised learning, where supervisory signals are learned directly from the data.⁸ However, the two terms are sometimes used interchangeably.) Even before this term was coined, researchers used to handcraft pretext tasks, which exploited unlabeled data. The most common tasks involved predicting distorted versions of the spatial characteristics of image data by means of rescaling, rotating, patching, shuffling, colorizing, and inpainting missing parts.

However, one could argue that devising these increasingly complex pre-training tasks resembles traditional feature engineering that neural networks promised to automate. Therefore, more generic recent methods have switched their focus from inventing single data transformations to comparing such views in the latent space and, therefore, offering elegant methods of implicit clustering between pseudo-positive and negative samples. Notably, SimCLR⁹ achieved—for the first time—performance on a par with supervised models, by proposing a two-network training method for visual representations, which maximizes agreement between differently transformed views of the same sample via a contrastive cosine similarity loss in the latent space. More recently, BYOL claimed better results even without the negative pairs in its training objective through a similar two-network approach.¹⁰ This sounds surprisingly similar to another family of models: generative adversarial networks (GANs), where the objective draws from game-theoretic principles, and two networks contest with each other in a game to generate more realistic data (see Figure 1). A useful taxonomy is introduced in,¹¹ where SSL models are grouped into three categories: generative (e.g., autoencoders), contrastive (e.g., SimCLR), and generative-contrastive (e.g., GANs or adversarial autoencoders). The main difference across these categories is the objective, ranging from reconstruction and contrastive losses to distributional divergences. We expect to see more overlap between generative, adversarial, and contrastive training in the future.¹²

These methods have shown promise that can indeed generalize to other data beyond images; however, data augmentations or objectives might need to be adapted when moving to a different type of input data.^{13,14} Furthermore, approaches such as the arrow of time¹⁵ exploit the temporal—rather than spatial—information of the data, by either artificially reverting the input sequences so as to distinguish between the correct and the reverse order (see Figure 1), or just predicting the future. We believe that models that anticipate and forecast the future are more robust and generalizable (for example, across hospitals).¹⁶ This seems to be particularly effective in language tasks as well, where models such as GPT-3⁵ outperform every other method by just slicing the data in such a way so as to predict the next word.

Transfer learning: The art of fine-tuning

Transfer learning is the natural application of SSL. The term transfer describes a set of methods toward preserving and reusing previously acquired information, applied possibly to a slightly

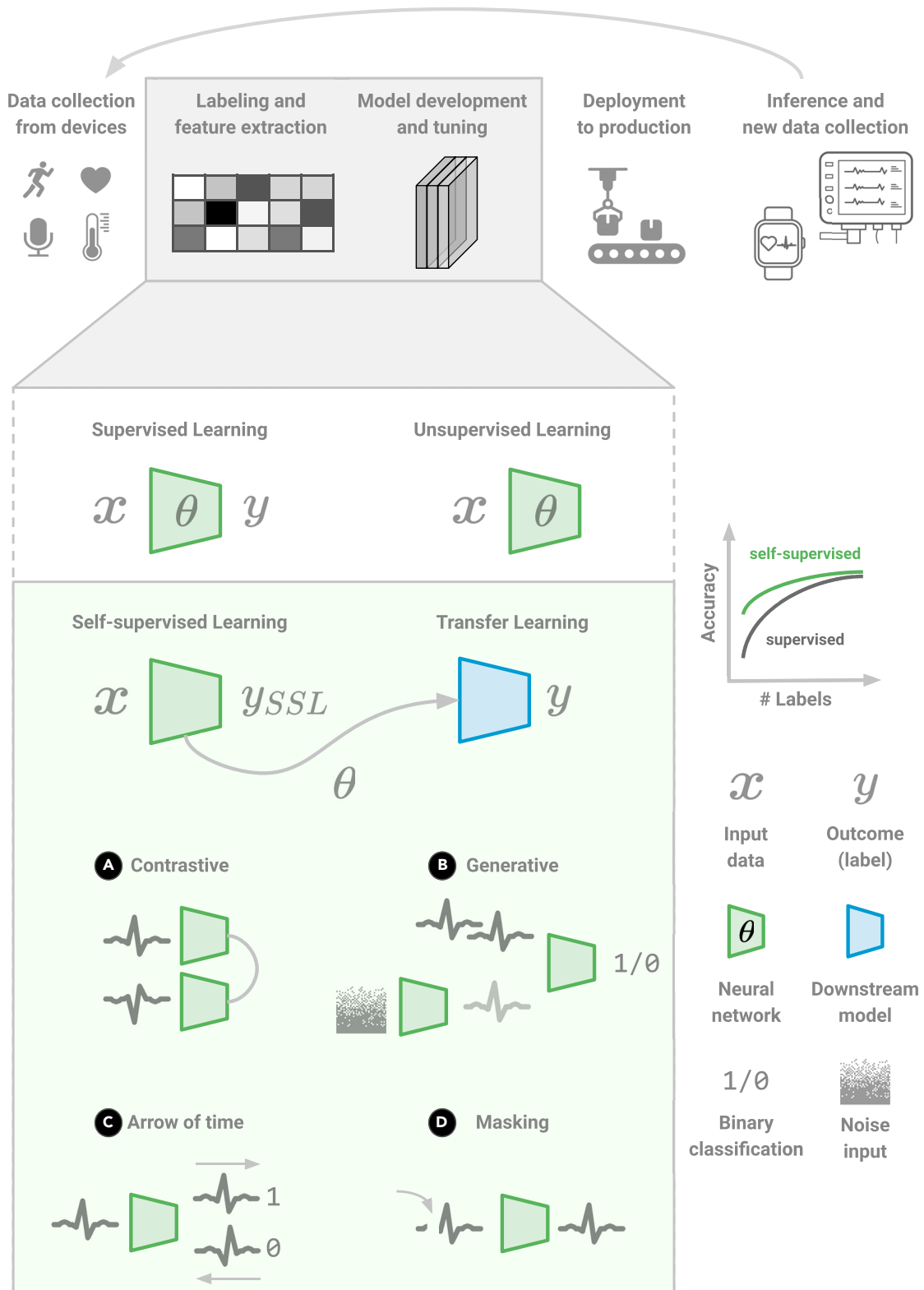


Figure 1. Self-supervised learning for health signals

Here, we illustrate the case of ECG signals and the prominent methods, which leverage unlabeled data with self-supervised learning. (A) Contrastive training maximizes the agreement between the original and the distorted view (flipped, rotated, or other augmentations).

(legend continued on next page)

different domain. This stored information can further accelerate the training of a downstream task with usually limited training data. Modern transfer learning uses pre-trained networks as fixed feature extractors in a) linear downstream models, where a logistic regression classifier is trained to classify a new dataset based on the self-supervised embeddings, or b) further fine-tuning the model in a downstream task. This has shown remarkable results in vision and language domains, where the learned embeddings can be directly applied to smaller datasets.^{5,9}

SELF-SUPERVISION FOR HEALTH SIGNALS

We will now present some recent results of self-supervised models applied to biomedical signal data (for a comprehensive view of this topic, the reader may be interested in Chowdhury and colleagues' review on SSL on medicine).¹⁷

Learning generalized physiological representations

While everyone can download off-the-shelf pre-trained models to further customize vision or language tasks, this is not the case for health signals. To this end, we recently developed Step2Heart,¹⁸ a self-supervised model that exploits the multi-modal data of modern wearables to learn meaningful representations, which generalize to several outcomes with transfer learning. The model maps activity data to future heart rate (HR) responses (implicitly applying the arrow of time principle) and can be used as a feature extractor for wearable data. For pre-training, we used a joint quantile loss function that accounts for the long tails of HR data, while downstream, we aggregated the window-level features to user-level ones and showcased the value captured by the learned embeddings through strong performance at inferring physiologically meaningful variables, ranging from anthropometrics to fitness, outperforming unimodal autoencoders and common biomarkers. For instance, the embeddings achieved an area under curve (AUC) of 0.68 for cardio-fitness prediction and an AUC of 0.80 for physical activity energy expenditure. Obtaining these outcomes in large populations can be valuable for downstream health-related inferences that would normally be costly and burdensome (for example, a fitness test requires expensive laboratory treadmill equipment and respiration instruments). Additionally, low fitness is a stronger predictor of mortality than diabetes, hypertension, and smoking.¹⁹ Overall, this is one of the first multi-modal self-supervised models for behavioral and physiological data, with implications for large-scale health and lifestyle monitoring (open source code and pre-trained models are available on <https://github.com/sdimi/Step2heart>).

Improving human activity recognition

A staple task in mobile health is that of human activity recognition (distinguishing between walking, sitting, running, etc., through activity sensors). This task is fundamental for the development of higher-level health monitoring applications. With SelfHAR,²⁰ we recently showed that large unlabeled mobile accelerometer

datasets can be leveraged to complement small, labeled ones. Our approach combines teacher-student self-training, which distills the knowledge of unlabeled and labeled datasets, while allowing for data augmentation, and multi-task self-supervision, which learns robust signal-level representations by predicting distorted versions of the input. SelfHAR achieved up to a 12% increase in performance (F1 score) using the same number of model parameters, by using up to 10 times fewer labeled data compared to supervised approaches. This work showed how to effectively distill, filter, and use unlabeled data orders of magnitude larger than supervised datasets. In a subsequent study, we studied the impact of the combinations of timeseries transformations in SimCLR models for activity recognition.¹³

Subject-aware biosignal representations

Commonly, datasets with a small number of subjects, such as electroencephalograms (EEG), manifest high intersubject variability. Therefore, Cheng et al. proposed a self-supervised model with an adversarial subject identifier to minimize subject-specific content.¹² They developed domain-inspired augmentations such as frequency-based perturbations to augment the signal, because the power in certain EEG frequency bands has been shown to be highly correlated with different brain activities. Through that, they found that temporal specific transformations (cutout and delay) are the most effective ones. Last, they investigated two scenarios: (1) using subject-specific distributions to compute the contrastive loss and (2) promoting subject invariance through adversarial training, finding that promoting subject invariance increases classification performance when training with a small number of subjects.

Data-efficient cardiac arrhythmia classification

Electrocardiogram (ECG) data are ubiquitous in healthcare settings and are increasingly common in personal devices such as the Apple Watch. Kiyasseh et al.¹⁴ proposed CLOCS, which leverages temporal and spatial invariances of ECG leads based on the two key observations: adjacent ECG segments of shorter duration will continue to share context, and recordings from different leads (at the same time) will reflect the same cardiac function and, thus, share context. A new idea was to define a positive pair as a representation of transformed instances that belong to the same patient. By doing so, the model implicitly personalizes the learned representations to each patient. Driven by this, they designed a new contrastive objective that outperformed supervised and generic self-supervised methods (in terms of AUC) such as BYOL, most notably, with only 25% of labelled training data.

Improving patient monitoring

Yèche et al. took the idea of inducing priors on contrastive losses a step forward.²¹ They design an objective that preserves the time dependency of the representations of the timeseries segments and outperforms unsupervised and supervised methods in predicting intensive care unit (ICU) decompensation, length

(B) Generative models such as GANs involve two networks that contest with each other in a game to generate more realistic data.

(C) Time-aware models try to guess whether the signal follows the arrow of time.

(D) Masked models hide part of the signal and challenge the model to predict the original one. The representations learned from these methods can be reused on downstream transfer learning tasks with linear models (blue box). Self-supervision is more label efficient in low-data regimes (top right graph).³

of stay, and sepsis onset (on the MIMIC dataset). The versatility of this approach is twofold: when fully unsupervised, it is competitive to supervised models, and when used in a supervised manner, it outperforms contrastive methods. Tonekaboni et al. independently arrived to a similar formulation by ensuring that in the encoding space, the distribution of signals from within a neighborhood is distinguishable from the distribution of non-neighboring signals.²² Their models surpassed competitors such as the triplet loss and contrastive predictive coding in predicting diverse outcomes, ranging from atrial fibrillation to human activity recognition, assessed with clustering metrics such as the silhouette score and classification metrics such as the area under the precision-recall curve. They also showed better clusterability over other contrastive losses. Both studies highlighted the generality of such models, which can be reused in multiple downstream tasks.

FURTHER IMPLICATIONS

Impact on physician workload

Apart from improving accuracy, self-supervision has the potential to lower physicians' workload. Most healthcare systems are overstretched, staff suffer from burnout, and predictive clinical tools could serve to relieve some pressure, especially in increasingly aged populations. These patients require more investigations—such as imaging or ECGs—hence adding to the overload of clinicians. Thus, employing clinical domain experts to label large datasets whilst working in understaffed departments is an expensive and unrealistic expectation that contributes to the bottleneck for the adoption of ML. Radiologists were the first to acknowledge this vicious cycle, which is likely to propagate to other medical specialties. Therefore, it is of paramount importance to develop efficient models that require minimal—or zero—annotations and physician burden. Self-supervised models and the automation of time-consuming tasks are a feasible solution toward more efficient medicine, potentially enabling more scalable patient screening, preventing delays in time-sensitive diagnoses, clinical management decisions, and hence, improving prognoses.

As a result, investing resources only on supervised models will slow down the adoption of ML in clinical settings and will disproportionately benefit high-income countries. Instead, label-efficient models pre-trained on large populations could generalize to global settings, which is of special relevance now that hospitals resume their elective operations whilst still fighting the COVID-19 pandemic.

Limitations

Self-supervised models tend to achieve remarkable transfer results with a fraction of labeled data. However, training these models can be computationally prohibitive, since it has been empirically shown that they require more data and more training iterations.⁵ Additionally, SSL is a multi-step process, involving a first step of generating augmentations, then training the pretext model, and last, fine-tuning to the target dataset, which makes the entire pipeline more complicated than simple end-to-end supervised models. Additionally, tracking progress in the area of health signals is not trivial, since there are many different tasks, modalities, baselines, and evaluation metrics. For instance, in

some papers, the sign when measuring the gap between SSL and supervised models is positive and, in some, negative, which seems to point to benchmark dependencies.

Conclusion and future outlook

We demonstrated the potential of models that learn meaningful representations directly from unstructured data and presented some recent results in the area of biomedical signals. The underlying challenge here is to find the best representation for biomedical signals, which range from generic ones, such as in BYOL,¹⁰ to ones based on medical intuition and physiology.^{18,22} Given that we have strong statistical priors about the nature of these signals, we are particularly excited by the prospect presented by the latter. Zooming out, Andrew Ng has recently expressed some alarm that there is a considerable gap between proof-of-concept models and actual real-life deployments, due to differences in sensors, protocols, or data collection methods: “In contrast, any human [doctor] can walk down the street to the other hospital and do just fine” (<https://www.fastcompany.com/90630654/stanford-ai-experts-healthcare>). Sequential transfer learning, as seen, for example, in Step2Heart¹⁸ or in numerous recent works,¹⁰ is probably the first step to validating that learned representations can generalize across different tasks while being label-efficient. The next steps should focus on demonstrating how these models can perform equally well in changing environments (e.g., different hospitals, populations, or devices). Some exciting new approaches toward this direction include disentangled autoencoders and meta-learning for domain generalization. In short, access to high-quality labeled datasets is the main bottleneck of transferring ML advances to critical fields such as healthcare, and SSL appears to be a feasible solution.

ACKNOWLEDGMENTS

D.S. was supported by the Embiricos Trust Scholarship of Jesus College Cambridge and EPSRC through Grant DTP (EP/N509620/1). I.P. was supported by GlaxoSmithKline and EPSRC through an iCase fellowship (17100053).

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

- Althoff, T., Hicks, J.L., King, A.C., Delp, S.L., and Leskovec, J. (2017). Large-scale physical activity data reveal worldwide activity inequality. *Nature* 547, 336–339.
- Hicks, J.L., Althoff, T., Kuhar, P., Bostjancic, B., King, A.C., Leskovec, J., and Delp, S.L. (2019). Best practices for analyzing large-scale health data from wearables and smartphone apps. *NPJ Digital Med.* 2, 1–12.
- Newell, A., and Deng, J. (2020). How Useful Is Self-Supervised Pretraining for Visual Tasks? (CVPR).
- Aggarwal, K., Joty, S., Fernandez-Luque, L., and Srivastava, J. (2019). Adversarial Unsupervised Representation Learning for Activity Time-Series (AAAI), pp. 834–841.
- Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., and Dhariwal, P. (2020). Language Models Are Few-Shot Learners (NeurIPS).
- Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A., and Bottou, L. (2010). Stacked denoising autoencoders: learning useful

- representations in a deep network with a local denoising criterion. *J. machine Learn. Res.* 11, 3371.
7. Salakhutdinov, R., and Hinton, G. (2009). Deep Boltzmann Machines (AISTATS).
 8. Jing, L., and Tian, Y. (2020). Self-supervised visual feature learning with deep neural networks: a survey. *IEEE Trans. Pattern Anal. Machine Intelligence* 43, 4037–4058.
 9. Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A Simple Framework for Contrastive Learning of Visual Representations (ICML).
 10. Grill, J.B., Strub, F., Alché, F., Tallec, C., Richemond, P.H., and Buchatskaya, E. (2020). Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning (NeurIPS).
 11. Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., and Tang, J. (2021). Self-supervised learning: generative or contrastive. *IEEE Trans. Knowledge Data Eng.* <https://doi.org/10.1109/TKDE.2021.3090866>.
 12. Cheng, J.Y., Goh, H., Dogrusoz, K., Tuzel, O., and Azemi, E. (2020). Subject-Aware Contrastive Learning for Biosignals. *arXiv*, 2007.04871.
 13. Tang, C.I., Perez-Pozuelo, I., Spathis, D., and Mascolo, C. (2020). Exploring Contrastive Learning in Human Activity Recognition for Healthcare (NeurIPS Machine Learning for Mobile Health workshop).
 14. Kiyasseh, D., Zhu, T., and Clifton, D.A. (2021). CloCS: Contrastive Learning of Cardiac Signals across Space, Time, and Patients (ICML).
 15. Wei, D., Lim, J.J., Zisserman, A., and Freeman, W.T. (2018). Learning and Using the Arrow of Time (CVPR), pp. 8052–8060.
 16. Chen, H., Lundberg, S., Erion, G., Kim, J.H., and Lee, S.I. (2020). Forecasting Adverse Surgical Events Using Self-Supervised Transfer Learning for Physiological Signals. *arXiv*, 2002.04770.
 17. Chowdhury, A., Rosenthal, J., Waring, J., and Umeton, R. (2021). Applying self-supervised learning to medicine: review of the state of the art and medical implementations. *Informatics* 8, 59.
 18. Spathis, D., Perez-Pozuelo, I., Brage, S., Wareham, N.J., and Mascolo, C. (2021). Self-Supervised Transfer Learning of Physiological Representations from Free-Living Wearable Data (ACM CHIL).
 19. Mandsager, K., Harb, S., Cremer, P., Phelan, D., Nissen, S.E., and Jaber, W. (2018). Association of cardiorespiratory fitness with long-term mortality among adults undergoing exercise treadmill testing. *JAMA Netw. Open* 1, 6.
 20. Tang, C.I., Perez-Pozuelo, I., Spathis, D., Brage, S., Wareham, N., and Mascolo, C. (2021). SelfHAR: Improving Human Activity Recognition through Self-Training with Unlabeled Data (ACM IMWUT).
 21. Yèche, H., Dresdner, G., Locatello, F., Huser, M., and Ratsch, G. (2021). Neighborhood Contrastive Learning Applied to Online Patient Monitoring (ICML).
 22. Tonekaboni, S., Eytan, D., and Goldenberg, A. (2021). Unsupervised Representation Learning for Time Series with Temporal Neighborhood Coding (ICLR).

About the authors

Dimitris Spathis recently completed his PhD in computer science at the University of Cambridge and is now interning in Microsoft Research. He has degrees in AI and computer science and has previously worked in Telefonica Research, Qustodio, and Ocado. His research enables deep neural networks to learn richer and label-efficient representations of high-dimensional real-world data (mobile sensors, time-series, audio, or other modalities), motivated by challenges in health. Lastly, he serves on the program committees of top AI conferences such as AAAI, IJCAI, and KDD, and his research projects have been featured in international media (BBC, Guardian, Forbes, Times, NPR, Venturebeat).

Ignacio Perez Pozuelo is with the University of Cambridge and the Alan Turing Institute, focusing on human-activity recognition using multimodal wearable sensors. He uses these behavioural phenotypes to further understand the impact of physical activity and sleep on health. Ignacio has worked on deriving sleep inferences from multi-modal data using deep learning approaches. Further, he has worked on time-series forecasting of digital biomarkers using physical activity as well as on activity classification using semi-supervised and self-supervised learning approaches for large, unlabelled datasets.

Laia Marques-Fernandez is a doctor at Cambridge University Hospitals NHS Foundation Trust (Addenbrooke's Hospital). She has a medical degree from the University of Barcelona with a year abroad at the University of Bologna. She is currently working as a junior doctor both in the hospital (in obstetrics and gynaecology, paediatrics, and emergency medicine) and in the community (in sexual and reproductive health at iCaSH Cambridge). Her recent research studies women's health from a population level down to the biological characteristics of disease. She envisions a future where doctors and AI work together to improve patient care and physician workload through predictive medicine.

Cecilia Mascolo is a Full Professor of mobile systems in the Department of Computer Science and Technology, University of Cambridge, UK. She is co-director of the Centre for Mobile, Wearable System and Augmented Intelligence. She is also a fellow of Jesus College Cambridge and the recipient of an ERC Advanced Research Grant. Her research interests are in mobile systems and data for health, human mobility modelling, sensor systems and networking, and mobile data analysis. More details can be found at www.cl.cam.ac.uk/users/cm542.