Projects for UCL Computer Science M.Sc. Programmes and 4th Year Undergraduate Students (2023/24)

Vasileios Lampos
Department of Computer Science
University College London
v.lampos@ucl.ac.uk

About our research

We conduct research in the emerging area of digital or computational epidemiology. We develop methods for using non-traditional, but evidently informative, data sources – such as online search activity – to provide early warnings for epidemics, understand health-related concepts better, but to also support the earlier diagnosis of serious health conditions. Our core methodological focus is in machine learning and natural language processing. Our disease prevalence estimates for influenza[1] and COVID-19[2] are fed directly to the UK Health Security Agency (UKHSA; formerly known as Public Heath England), and are included in their weekly disease surveillance reports.[3] Novel and effective solutions to the tasks described below could be incorporated to the information we send to UKHSA on a weekly basis, i.e. your research project might also have a tangible impact.

General information

Below we briefly describe a few projects or broader research tasks that we are interested in. Please read all of them as information provided in one project may be implied in the next one. We are open to discuss different research directions that fit within our general research scope, can be supported by data sets that are already at our disposal, and will not require a further ethics review.

Project 1 – Influenza prevalence forecasting using online search activity

Influenza is an infectious disease that causes hundreds of thousands of deaths annually. Early indications for an emerging seasonal epidemic can help health authorities to prevent excessive spread and prepare more effectively, e.g. by re-allocating resources and by administering antivirals. Traditional epidemiological surveillance systems are predominantly based on doctor (GP) visitation rates and laboratory confirmed cases. However, both approaches seldom provide timely disease rate estimates. Specifically for influenza, GP visitation rates only capture a small and demographically biased part of the population (Wagner et al., 2018).

Alternative sources of information such as web search activity can be modelled to provide quite accurate estimates of the current flu rates, a task commonly referred to as *now*-casting (Lampos et al., 2015, 2017). Online searches have a very strong *recency* attribute resulting to more timely disease prevalence estimates compared to traditional approaches. Accurate forecasting models, on the other hand, can significantly increase the utility of digital epidemiology, allowing more time for planning and deploying public health interventions. At the same time, their development poses new challenges. In this project, you will use web search activity data to (choose one of the following):

- Develop neural network architectures for influenza rate forecasting with uncertainty estimates. See some recent work of ours in this topic in Morris et al. (2023).
- Propose frameworks for hyperparameter validation in influenza rate forecasting models with neural networks. Hyperparameter validation in disease rate forecasting models cannot follow common practices in machine learning. Infectious diseases are not static (e.g. viruses mutate and their transmissibility changes), and the same holds for web search behaviour as well as other related factors such as the way people interact with national health systems. Hyperparameter validation should consider these properties. When it does not, forecasting accuracy is

[1] Flu detector, fludetector.cs.ucl.ac.uk
[2] COVID-19 models, covid.cs.ucl.ac.uk
[3] UKHSA's flu and COVID-19 weekly surveillance reports, gov.uk/government/statistics/national-flu-and-covid-19-surveillance-reports-2022-to-2023-season

inferior. So, the task here is to find ways to optimise hyperparameter validation automatically, as a component of the entire training process that is tailored to the special characteristics of our task, and eventually learn more accurate forecasters.

- Develop solutions for epidemiology-aware optimisation (or hybrid approaches) of neural network forecasters. Epidemiology uses variations of compartmental models to capture the dynamics of an infectious disease. Compartmental models are often defined as sets of differential equations that provide a description of how parts (compartments) of a population are affected by the disease at certain points in time. These models are based on expert knowledge and other factors that are derived from epidemiological analysis. In this way, they are limited to various design assumptions that may not hold especially for novel diseases or variants, and similarly to traditional disease surveillance systems may be carrying over biases present in the collected epidemiological information. On the other side of the spectrum, disease forecasting models that are based on web search activity, are sometimes producing forecasts that violate basic disease properties, something that is hard to diagnose during a flu season (i.e. when performing forecasts in real time), but becomes obvious in retrospective modelling. Hence, an interesting task could be to bring in traditional epidemiological models inside a neural network forecaster, and use them as a regulariser / soft-constraint during optimisation. Notably, recent work has shown how ordinary differential equation (ODE) systems can be solved by neural networks (Xu et al., 2021). Please note that this project will have a strong research element (increased level of difficulty).

Project 2 – COVID-19 incidence and health burden modelling at finer geographies using the Google symptoms data set

In this task, we will use a novel data set of symptom-related web search activity that has been released by Google.[4] In addition to covering more than 400 symptom categories, this data set offers very fine geographical granularity. The focus of the modelling will be on COVID-19. Supervised learning models for COVID-19 incidence indicators (confirmed cases, deaths, hospitalisations etc.) will be developed nationally as well as sub-nationally. Can we improve model accuracy by learning more complex models (for *now*-casting or forecasting) using an expanded set of observations (training samples) from the many different sub-national locations? How are we going to tackle the limitation of not being able to access enough historical data (i.e. many different COVID-19 waves/seasons)?

[4] Google symptoms search trends, pair-code.github.io/covid19_symptom_dataset/

Project 3 – Transferring a disease model from one location to another

In previous work, we showed how to transfer disease prevalence models, trained in a source country using past web search activity and disease rates, to a target country, where we assumed to have access only to web search activity (Zou et al., 2019; Lampos et al., 2021). Hence, from a transfer learning perspective, no ground truth was available for the target domain to assist us in calibrating the transfer. We investigated the transfer of regularised linear models, and the mapping of features (web search queries) from the source to the target location based on semantic and frequency time series similarity. The aim of this project would be to investigate the transfer of more complex models (e.g. neural network architectures) as well as more abstract mappings between the domains' feature spaces.

Data and computational resources

Data for the aforementioned projects is either publicly available or will be provided by us. Common data providers are Google, UKHSA, the Royal College of General Practitioners (RCGP), and the Centers for Disease Control and Prevention (CDC). Additional data sources may be sought based on the choice or formulation of a project. You will be able to use our group's computational resources (including some basic GPU support), if necessary.

Ethics

These projects will use aggregate information. Data for individual users is not available to us. Given this, the projects described here have obtained an ethics exemption by UCL Computer Science.

References

Vasileios Lampos, Maimuna S. Majumder, Elad Yom-Tov, Michael Edelstein, Simon Moura, Yohhei

Hamada, Molebogeng X. Rangaka, Rachel A. McKendry, and Ingemar J. Cox. 2021. Tracking COVID-19 using online search. *Nature (npj) Digital Medicine*, 4(17).

Vasileios Lampos, Andrew C. Miller, Steve Crossan, and Christian Stefansen. 2015. Advances in nowcasting influenza-like illness rates using search query logs. *Scientific Reports*, 5(12760).

Vasileios Lampos, Bin Zou, and Ingemar J. Cox. 2017. Enhancing feature selection using word embeddings: The case of flu surveillance. In *Proc. of the 26th International World Wide Web Conference*, pages 695–704.

Michael Morris, Peter Hayes, Ingemar J. Cox, and Vasileios Lampos. 2023. Neural network models for influenza forecasting with associated uncertainty using Web search activity trends. *PLOS Computational Biology*, 19(8).

Moritz Wagner, Vasileios Lampos, Ingemar J. Cox, and Richard Pebody. 2018. The added value of online user-generated content in traditional methods for influenza surveillance. *Scientific Reports*, 8(13963).

Winnie Xu, Ricky T. Q. Chen, Xuechen Li, and David Duvenaud. 2021. Infinitely Deep Bayesian Neural Networks with Stochastic Differential Equations. *arXiv preprint (2007.04504)*.

Bin Zou, Vasileios Lampos, and Ingemar J. Cox. 2019. Transfer Learning for Unsupervised Influenza-like Illness Models from Online Search Data. In *Proc. of the 28th International Web Conference*, pages 2505–2516.