Translational challenges of biomedical machine learning solutions in clinical and laboratory settings

Carlos Vega, Miroslav Kratochvíl, Venkata Satagopam, and Reinhard Schneider carlos.vega@uni.lu

Luxembourg Centre for Systems Biomedicine, Université du Luxembourg, Esch-sur-Alzette, Luxembourg

Summary and goals

- This poster summarizes the challenges of porting ML-based solutions to clinical and lab settings.
- 2 case studies highlight a mismatch between expectations and result interpretation between the researchers.
- We point out some sources of common problems, and outline how interpretable inference reporting can help to detect the issues and overcome the challenges.

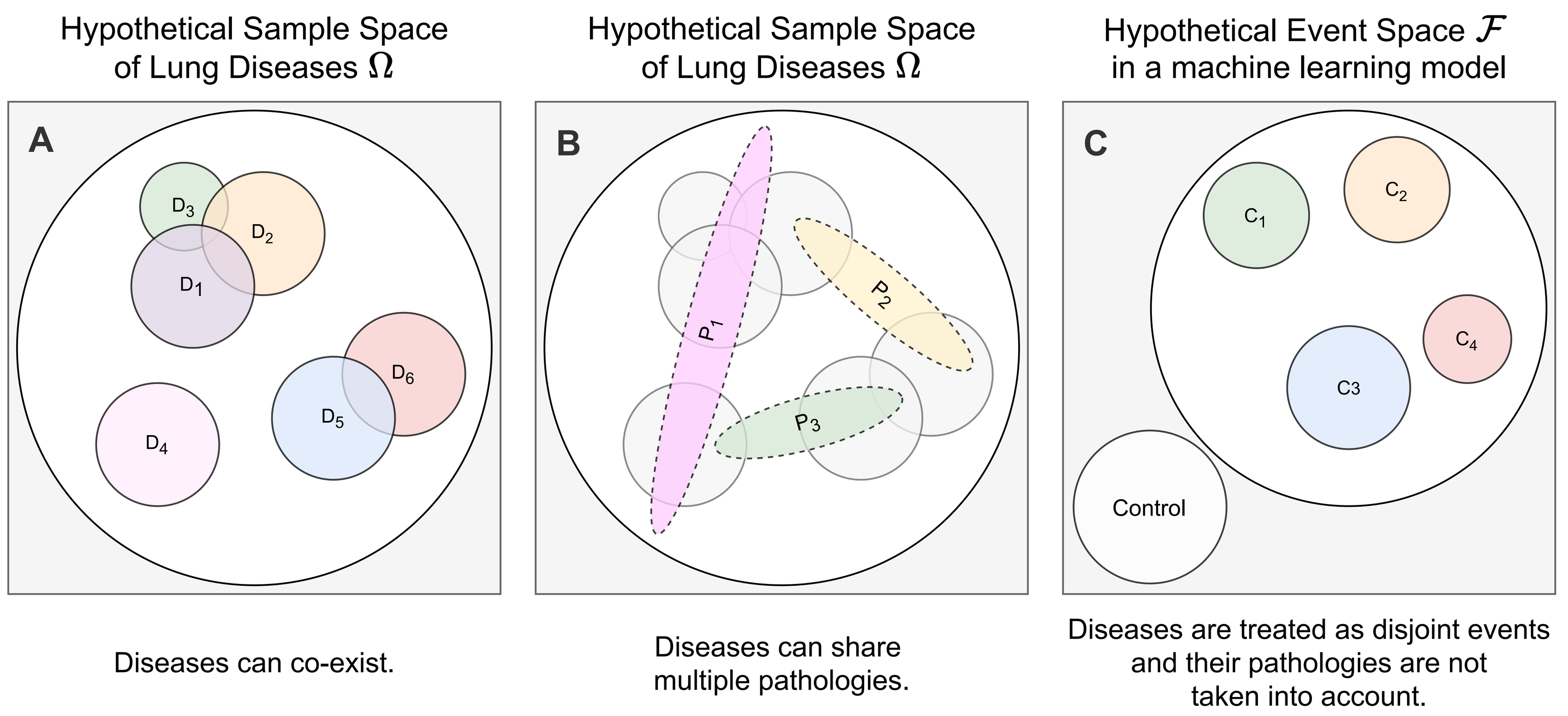
"For translational medicine, all training sets and predictions are merely blurry **shadows of reality.**"

UNIVERSITÉ DU LUXEMBOURG

LCSB

Extended abstract!

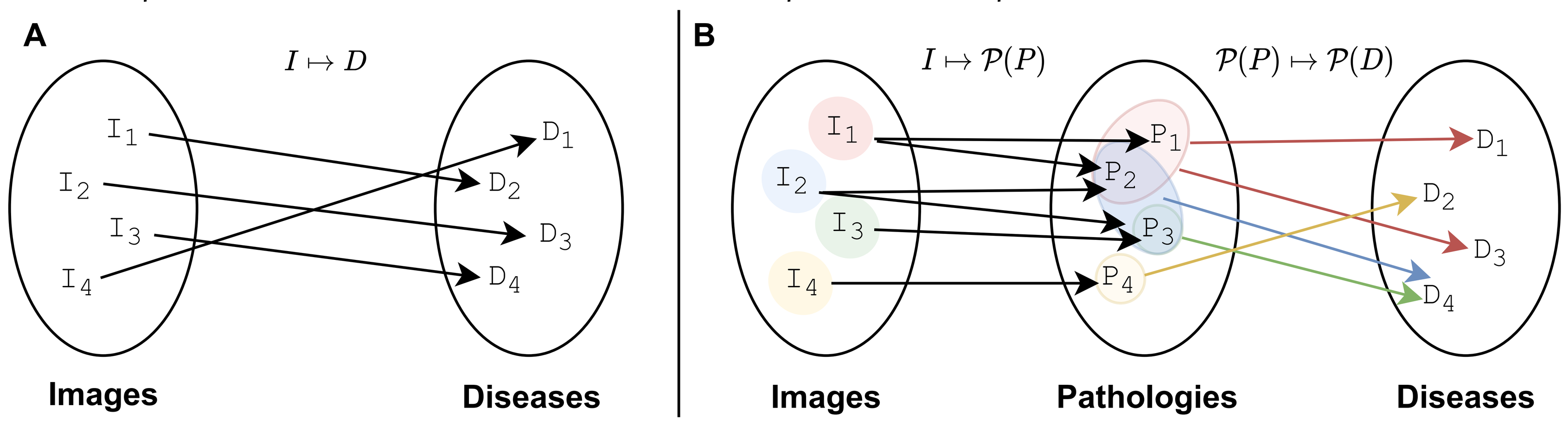
1. Chest X-ray image diagnosis of COVID-19



At the pandemic outset, researchers rushed to develop solutions from crowd-sourced repositories to predict the COVID-19 severity and outcome from chest X-ray (CXR) images. Questionable methods and poor annotation of the datasets spawned a multitude of problems. Commonly, the solutions were based on binary or multi-class classification methods that considered a small subset of diseases.

Moreover, these solutions assume mutual exclusion of the classes while, in fact, many lung diseases may co-exist. For example, COVID-19 and Tuberculosis share abnormalities such as fibrosis and opacities, and produce a spectrum of pathologies that evolves over time, requiring a combination of tests (e.g. blood, sputum) for their diagnosis.

Attempts to diagnose lung diseases with just CXR images are thus unnecessarily partial and defy the multimodal nature of diagnosis. Hence, regardless of the reported evaluation metric, binary and multinomial classification solutions are rarely suited for real clinical settings, mainly due to unrealistic assumptions about the nature of the predicted phenomena.



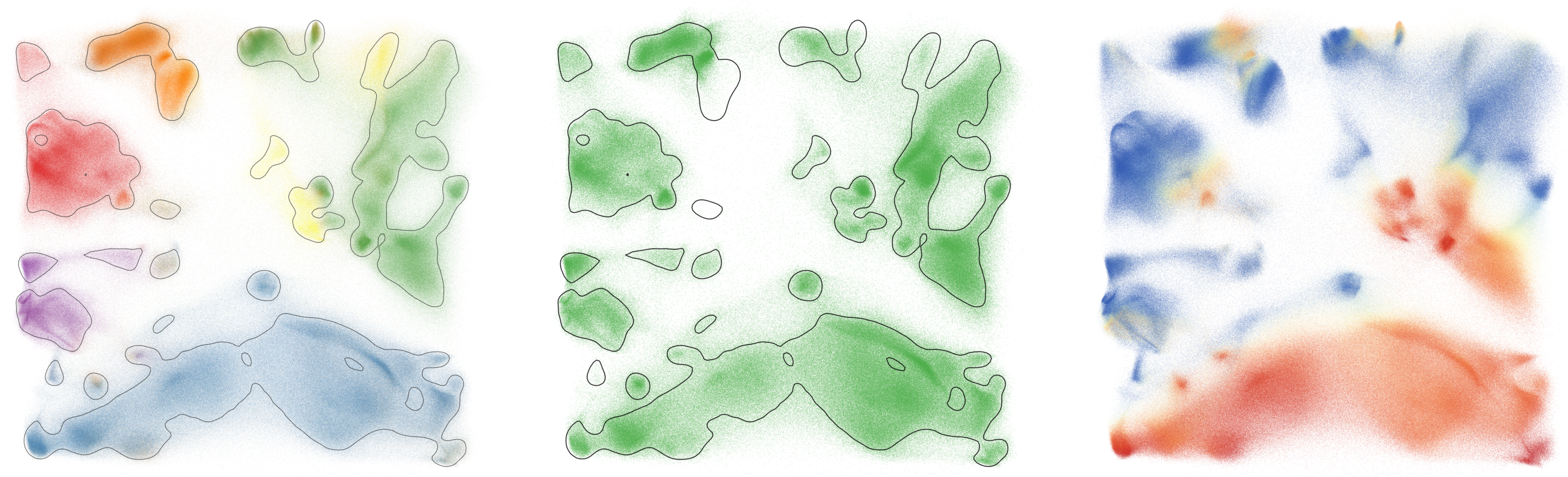
A Conventional models make use of a single valued mapping, $I \mapsto D$. **B** An example of an alternative method employs multivariate and multivalued functions for a more representative mapping. $\mathcal{P}(P)$ and $\mathcal{P}(D)$ denote the respective powersets of the set of pathologies P and set of diseases D.

2. Disease status detection in cytometry

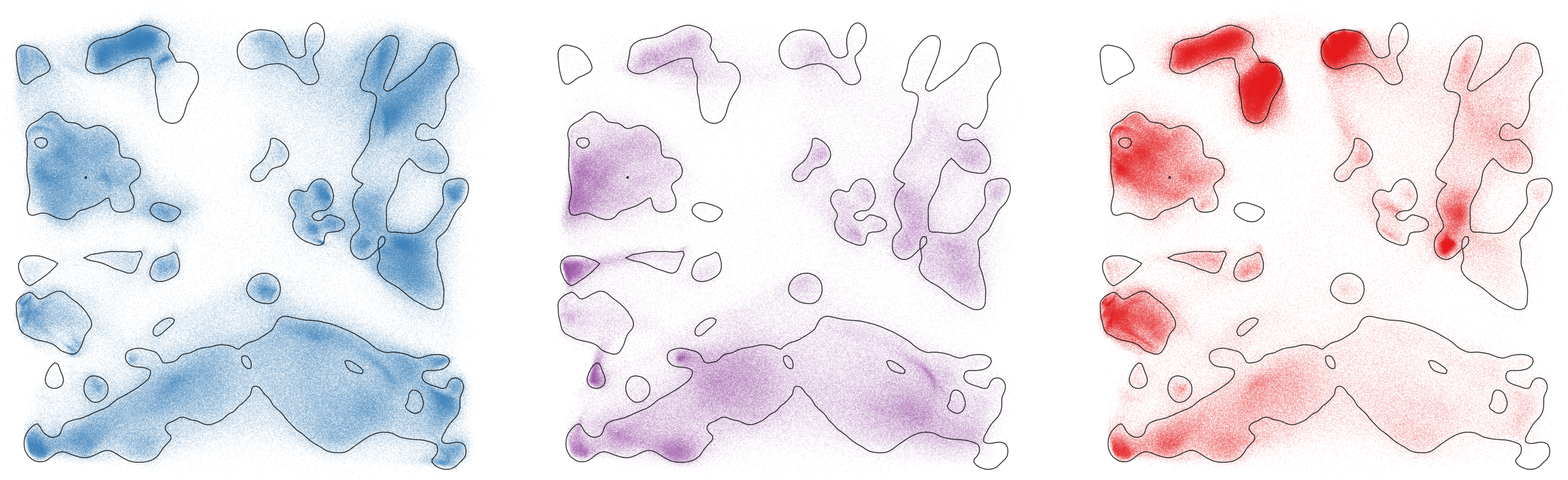
Studies in cytometry are usually limited to a few dozen samples, extremely prone to batch effects, and plagued by result discretization that was often viewed as a remedy to the data complexity. In turn, the field met similar challenges as in case study 1.

The new ML-powered dimensionality reduction methods, such as t-SNE, UMAP, TriMap, EmbedSOM, ivis, scvis:

- × Allow (wide) misuse of the 2D data for base analysis.
- × Introduce danger of losing details (can be relieved by user interaction).
- ✓ Avoid discretization error, communicate 'unclear' results comprehensibly.
- ✓ Provide QC for free — all dataset problems are immediately visible.



Above From left: an imprecise but useful color assigment of 'phenotype' to cells; cell subset present in a 'healthy' donor; cells colored by presence of a specific surface protein (CD28).
Below Cell subsets present in donors with celiac disease (left), Crohn disease (mid), and type-II refractory celiac disease (right). Visualization allows quick diagnosis of changes in cell composition, while avoiding the need for classifiers that are problematic to understand outside ML community.



Poster challenge How fast can you spot the major differences in cell composition between healthy (green) and the diseased samples (lower row)?

Dataset in case study 2: van Unen, Vincent, et al. "Visual analysis of mass cytometry data by hierarchical stochastic neighbour embedding reveals rare cell types." *Nat comm* 8.1 (2017).