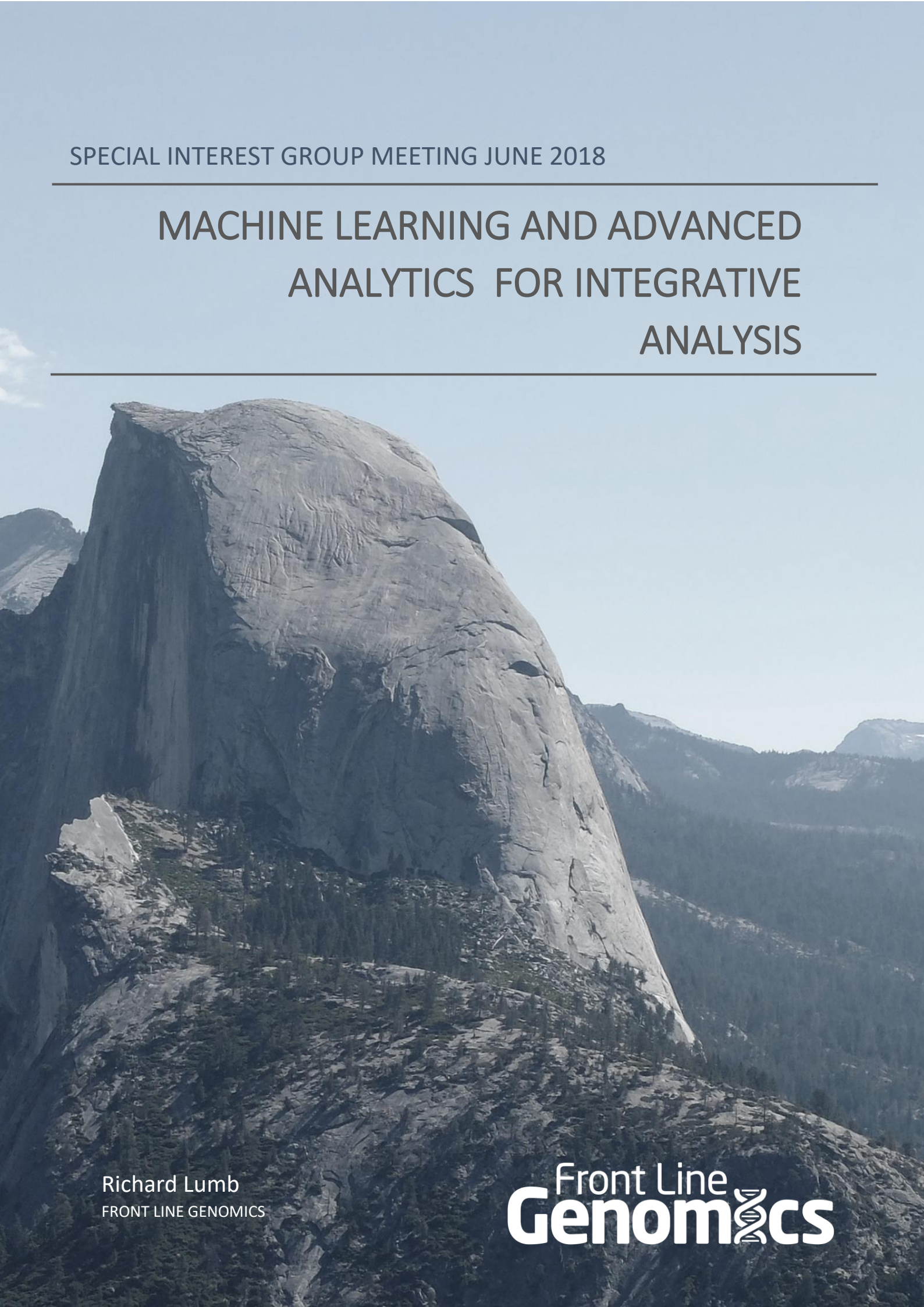
SPECIAL INTEREST GROUP MEETING JUNE 2018

MACHINE LEARNING AND ADVANCED ANALYTICS FOR INTEGRATIVE ANALYSIS



Richard Lumb
FRONT LINE GENOMICS

Front Line Genomics

Foreword

In the big data era of medicine, data is growing explosively in both volume and variety. The field of precision medicine research lacks the computational tools to reliably integrate multiply types of bio-data. There is a strong need of integrative machine learning models in order to better make use of heterogeneous information in decision making and knowledge discovery.

To explore and examine how these models could be developed and implemented, experts in the field were brought together for a meeting under Chatham House Rules, in Boston on June 20th. In order to maintain a balanced perspective, the participants of the meeting were selected from a range of backgrounds, with many different specialties and focuses.

The purpose of this meeting was to outline the challenges and plot a route for machine learning experts and data scientists to wisely optimize the use of these data. A key focus here was on the standardization of ontologies, data collection, capture, and refinement. Without these standards set in place, ML is limited in its use.

This report presents an accurate reflection of the discussions that took place at this meeting and does not constitute an official statement from any of the individual participants, their organizations, or Front Line Genomics.

Attendees:

Pete Stetson, Chief Health Informatics Officer, Deputy Physician-In-Chief, Office of the Physician-In-Chief, **Memorial Sloan Kettering Cancer Center**

Heming Xing, Senior Principal Scientist, Precision Immunology Cluster, **Sanofi**

Riccardo Sabatini, Chief Data Scientist, **Orionis Biosciences**

Marghoob Mohiyuddin, Research Leader, Bioinformatics, **Roche**

Lee Lichtenstein, Associate Director, Somatic Computational Methods, **The Broad Institute**

Renato Umeton, Head of Data Science, **Dana-Farber Cancer Institute**

Linda Zhou, Director, Research and Life Sciences Solutions, **Western Digital**

Brad Chapman, Senior Research Scientist, **Harvard University**

Vibhor Gupta, Director, **Pangaea Group**

Pablo Cingolani, Principal Scientist, **AstraZeneca**

Nathanael Fillmore, Associate Director for Machine Learning and Predictive Analytics, **MAVERIC, VA Boston Healthcare System**

Bin Li, Director, Computational Biology, **Takeda**

Jack Pollard, Head of Cancer Bioinformatics, **Sanofi**

Joanna Fueyo, Visiting Scholar, Bioinformatics, **Boston University**

Mark Kon, Professor of Mathematics and Statistics, **Boston University**

John Quackenbush, Professor of Computational Biology and Bioinformatics, **Harvard T.H. Chan School of Public Health**

Leonid Peshkin, Lecturer on Systems Biology, **Harvard Medical School**

Will Chen, VP Product Management and Business Development, Precision Medicine, **Elsevier**

Amy Lu, Visiting Scholar, Bioinformatics and Genomics, **Boston Children's Hospital**

Jason LaBonte, Head of Product, **DataVant**

Frances Shaw, Producer, **Front Line Genomics**

Richard Lumb, CEO and Founder, **Front Line Genomics**

Report Author: Frances Addison, Staff Writer, Front Line Genomics

Points of Discussion:

Standardization

- **There needs to be standard ontologies to harmonize data effectively.** It is possible to rework definitions for individual data points on small datasets, but it isn't scalable; when working with big genomic datasets, it is impossible. Instead, researchers need to be building standardized ontologies that can be used for multiple datasets, instead of being used once and then abandoned, and then working to that. These ontologies need to be built collaboratively so that they can be used freely and easily by different groups.

- **Current ontology strategies may only be pseudocompatible.** Building ontologies that can be standardized across clinical trials is very difficult, because the parameters are changing for each case. This means that different trials' data cannot be examined effectively together because they are not comparable.

- **Disease classifications are not standardized**, which limits data sharing and collaboration between healthcare organizations. Particularly with highly varied conditions like cancer, there are many different options for measuring disease progression, treatment success, disease stage, and so on. Without this standardization, possible research collaborations to develop improved therapies are being held back.

Gathering Data

- **Data reporting is frequently inaccurate**, and this is delaying researchers' ability to use it effectively. Most projects are formulated and planned on the basis of the data that they believe they will have access to from the descriptions of the datasets. However, this reporting is often inaccurate and the expected data is not available, meaning that researchers have to go back and retrofit their use case for the data that is actually available.

- **Use cases do not work in the way they have done traditionally.** In the past, use cases were used to define the data. Now, the data is being obtained first and then used to define the use case.

- **Data needs to be cleanly formatted before being submitted to a consortium.** If each research group does the necessary data processing at the front end, then collective databases can be better structured and more usable. This would also remove the need for central databases to spend large amounts of time and money processing

their datasets at a later time.

- Data capture, refinement, and analysis need to be considered in coherence. All three make up the overall process of data handling; if they are thought about individually, then other aspects of the process will suffer and the output will be weaker as a result. Developers need to be working on the whole picture.

- Generating high quality data is time consuming. It is much quicker to generate large quantities of data of a substandard quality, and so this is commonly the only option for groups that have limited manpower. To encourage everyone to start generating data of a higher quality, which can then enable improved ML and AI techniques, the community needs to provide better solutions than those currently available.

Machine Learning Algorithms

- Identifying variant relevance can be difficult for machine learning algorithms. Each patient will have millions of genetic variants that can be identified by data analysis algorithms, but the same algorithms are unlikely to be able to ascribe biological significance. To prevent large numbers of irrelevant variants being identified, the learning process of the algorithm needs to be heavily constrained to bias identification in favor of relevant variants.

- Understanding learning constraints is imperative when developing new algorithms. If too much data is used, then the algorithm will be less able to pick out the variants that are most important to the research being performed. Large training sets will also extend the learning process and may be difficult to obtain when examining rare variants. At the same time, if too little data is data is used, the algorithm will not be trained effectively.
  The model being used will constrain the data and, simultaneously, the data will constrain the model.

- Unsupervised learning models might resolve the limited scalability of human data processing. Current supervised learning models for algorithms require researchers to input the desired data labels against which the data can be sorted. Because of the complexity and size of genetic data, this approach can severely limit the scalability of learning processes. Unsupervised models, which do not require labels to be provided, might enable researchers to avoid this problem in the future.

Culture

- Pharma companies are moving in the right direction with data sharing, but they're still not the best examples to follow. Pre-competitive collaborations are enabling

companies to work together to further research and disease understanding, but they are highly unlikely to be willing to share any clinical data, because it has the potential to harm their drug development programs.

- Deep learning has a confidence problem. Generally, researchers are not confident in the results being supplied to them by deep learning processes because of inaccuracies in the techniques in the past. This is made worse by the fact that the accuracy of certain deep learning techniques can be hard to clearly quantify. Without confidence in these results, the techniques will not be used as often as they might be otherwise.

- There is a legal aspect to consider when using AI and ML clinically. Currently, a doctor is still making any final decisions; any machine learning tools are only being used in an advisory capacity. This means that the doctor is still directly responsible for patient care in a legal sense. If ML and AI tools start to be developed that are capable of making treatment decisions alone then there are significant legal concerns regarding responsibility and accountability.

- Everyone is facing the same problems, particularly around gathering clean data. This is encouraging because it increases the possibility of collaborations that can start to resolve these issues, but it will require communication and universal agreement on how data should be handled.

- There is limited communication between ML developers and end users. As a result, the people using these tools are not the ones with input into how they could be improved to increase usability or functionality. For such tools to improve in a way that can aid research and clinical applications, there needs to be much stronger communication between the two groups.