The VADA Program
Visual and Automated Disease Analytics Graduate Training Program

Monday Program

Time	Activity/Topic	Speaker/Led by
7:45am – 8:30am	Breakfast (A206 Chown)	
8:30am – 8:45am	Welcome to summer school and overview of week	Dr. Lisa Lix and Dr. Pourang Irani
8:45am – 9:45am	Q&A with Dr. Meghan Azad and Dr. Kozeta Miliku on Big Data Challenge	Dr. Lisa Lix
9:45am – 12:00pm	Work time for Big Data Challenge	
12:00pm – 1:00pm	Lunch – John Buhler Research Centre Atrium	
1:00pm – 4:30pm	Work time for Big Data Challenge	

Tuesday Program

Time	Activity/Topic		
7:45am – 8:30am	Breakfast served (A206 Chown)		
8:30am – 10:00am	Exploring Missing Data Techniques – A207 Chown – Kristine Kroeker	Developing Predictive Models – B207 Chown – Dr. Rob Balshaw	Microbiota, Microbiomes and Metagenomes - A315 Chown – Dr. Gary Van Domselaar
10:00am – 10:15am	Coffee break (A206 Chown)		
10:15am – 10:45am	Bear pit session with faculty advisors		
10:45am – 12:15pm	Work time for Big Data Challenge		
12:15pm - 1:00pm	Lunch – John Buhler Research Centre Atrium		
1:00pm – 1:45pm	Check in with Dr. Meghan Azad and Dr. Kozeta Miliku on Big Data Challenge		
1:45pm – 4:00pm	Work time for Big Data Challenge		
4:00pm – 5:30pm	Big Data Challenge presentations (A&B 207 Chown) Judges: Dr. Andriy Koval, Dr. Kozeta Miliku, Dr. Wendy Young and Dr. Chris Greene		
6:00pm – 9:00pm	Faculty and judge Dinner – Cibo Waterfront Cafe	Student social event – Coronation Bowling Centre	

The VADA Program
Visual and Automated Disease Analytics Graduate Training Program

Wednesday Program

Time	Activity/Topic	Speaker/Led by
7:45am – 8:15am	Breakfast - A206 Chown	
8:15am – 8:30am	Overview of day and introductions	
8:30am – 12:00pm	Statistical methods for Neuroimaging Data Analysis – A&B207 Chown	**Dr. Linglong Kong** – University of Alberta
12:00pm – 1:00pm	Lunch – John Buhler Research Centre Atrium **Separate Round Table Discussions:** • Asset based VADA student development: resources, skills and experience available within Island Health **- Dr. Wendy Young** – 403 Brodie • Title TBD **- Dr. Pourang Irani** – 404 Brodie	
1:00pm – 3:00pm	Techniques for Reproducible Visualization in R: Graph as a Sequence of Three Functions	**Dr. Andriy Koval** – University of Central Florida
3:30pm – 5:00pm	Networking Reception – Canadian Museum for Human Rights Sifton Terrace	

Thursday Program

Time	Activity/Topic	Speaker/Led by
7:45am – 8:15am	Breakfast - A206 Chown	
8:15am – 8:30pm	Overview of day	
8:30am – 10:30am	Using Github For Data Science: Version Control, Project Management, Promotion – A&B207 Chown	**Dr. Andriy Koval** – University of Central Florida
10:30am – 11:15am	Integrating GitHub and R for reproducible epidemiological and genomics research for precision medicine approaches to psychiatry - A&B207 Chown	**Dr. Kaarina Kowalec** – University of Manitoba
11:15am – 12:00pm	Automated Pipelines for the Analysis of Genomic Data and Their Clinical Applications – A&B 207 Chown	**Dr. Xiao-Qing Liu** – University of Manitoba
12:00pm – 1:00pm	Lunch – John Buhler Research Centre Atrium **Separate Round Table Discussions:** • Using data analytics to monitor and transform health care – **Dr. Michael Routeledge** – 405 Brodie • What kind of training do you need to become a successful health data scientist – **Dr. Joon Lee** – 404 Brodie • Digital analytics optimization for value based health care – **Xue Yao** – Buhler Atrium	
1:00pm – 2:30pm	Data Science, AI, and Health: Where are we Headed? - A&B207 Chown	**Dr. Joon Lee** – University of Calgary
2:30pm – 2:45pm	Refreshment Break – 206 Chown	

2:45pm – 4:30pm	Data visualization strategies and tools for microbial genomic epidemiology - A&B207 Chown	**Anamaria Crisan** – University of British Columbia

Friday Program

Time	Activity/Topic	Speaker/Led by
7:45am – 8:15am	Breakfast - A206 Chown	
8:15am - 8:30am	Overview of Day	
8:30am – 10:30am	Jupyter Notebook: A Modern Tool for Open, Reproducible, and Distributable Data Science – A&B207 Chown	**Adrian Zetner** – Public Health Agency of Canada
10:30am – 10:45am	Break	
10:45am – 12:15pm	Career Panel - A&B207 Chown **Dr. Andriy Koval** – University of Central Florida **Dr. Laura Cowen –** University of Victoria **Anamaria Crisan** – University of British Columbia	
12:15pm – 1:15pm	Lunch - John Buhler Research Centre Atrium **Separate Round Table Discussions:** • Validating Models in AI **– Dr. George Tzanetakis** - 405 Brodie Centre • Impact of Electronic Health Data Quality on Research Findings **- Dr. Lisa Lix** – 404 Brodie Centre	
1:15pm – 1:45pm	Question and Answer Session on Big Data Programs at University of Manitoba – A&B 207 Chown	Speaker TBD
1:15pm – 2:15pm	Debrief and Evaluations	

The VADA Program
Visual and Automated Disease Analytics Graduate Training Program

Session Descriptions and Speaker Biographies

Tuesday June 11th

Exploring Missing Data Techniques

Tuesday, June 11th 8:30am – 10:00am

Session Description: The advantages and disadvantages of different techniques for handling missing data will be discussed. A few commonly used missing data methods will be applied and compared using R.

Bio: Kristine is a Data Analyst in the Data Science Platform mainly working with Manitoba's administrative data and clinical databases. She leads workshops in R Studio. She completed her B.Sc (Hons) in Statistics and M.Sc in Community Health Sciences both from the University of Manitoba.

Developing Predictive Models

Tuesday, June 11th 8:30am – 10:00am

Session Description: Predictive modeling has a simple goal: to make the best possible predictions in new data. But there are many ways to achieve this – some good, some not so good, and some bad. Using the multiple logistic regression model and tools from the R tidyverse, a "near best practice" approach to building, refining, and testing a predictive model using R will be reviewed.

Bio: Rob Balshaw is a Senior Biostatistician with the Data Science Platform at the George and Fay Yee Centre for Healthcare Innovation. Before this, Rob was a Senior Scientist for 5 years at the BC Centre for Disease Control, and for 14 years was head of biostatistics at Syreon Corporation (Vancouver), a contract research organization conducting clinical trials for the pharmaceutical industry. He has been building, refining, and testing predictive models for more than 25 years and continues to learn about how it should be done.

Microbiota, Microbiomes and Metagenomes

Tuesday, June 11th 8:30am – 10:00am

Session Description: This session will serve as an overview to omics data and will cover how this data is generated, sources of variability and confounders as well as how to analyze this data and the role of machine learning within analysis.

Bio: Dr. Gary Van Domselaar is the Chief of the Bioinformatics Laboratory at the National Microbiology Laboratory in Winnipeg Canada, and Adjunct Professor in the Department of Medical Microbiology at the University of Manitoba. Dr. Van Domselaar's lab combines novel analytical systems and advanced visualization systems to research and control disease. His work incorporates metagenomics, infectious disease genomic epidemiology, genome annotation, bacterial population structure analysis, and genome wide association studies to understand and respond to infectious disease threats.

The VADA Program
Visual and Automated Disease Analytics Graduate Training Program

Wednesday June 12th

Statistical Methods for Neuroimaging Data Analysis

Wednesday June 12th 8:30am – 12:00pm

Session description: As modern imaging techniques have developed, massive imaging data can be observed over both time and space (ex: magnetic resonance imaging (MRI), functional magnetic resonance imaging (fMRI), and diffusion tensor imaging (DTI)). Advanced statistical methods on imaging data have been proposed, studied and applied in various fields. This short course aims to provide a practical introduction to and overview of recent advanced statistical challenges and developments for analyzing and modelling medical image data quantitatively. The course material is applicable to a wide variety of medical and biological imaging problems. The topics include tract-based analysis, multi-scaled statistical methods, fMRI processing methods, diffusion imaging methods, brain image and genetics. While presenting the statistical fundamentals, we emphasize the concepts, methods and their real-world implementation.

Bio: Dr. Linglong Kong is an associate professor at the department of Mathematical and Statistical Sciences at the University of Alberta. Currently, Linglong is serving as associate editor of Journal of the American Statistical Association, Applications & Case Studies, International Journal of Imaging Systems and Technology, Canadian Journal of Statistics, and as the ASA Statistical Imaging Session program chair. His research interests include high-dimensional data analysis, neuroimaging data analysis, robust statistics and quantile regression, and statistical machine learning.

Techniques for Reproducible Visualization in R: Graph as a Sequence of Three Functions

Wednesday June 12th 1:00pm – 3:00pm

Session Description: Rarely do applied data science projects produce any given graph only once. The need to generate plots of the same form using different inputs, different options, and in multiple contexts require the analyst to structure the operations involved in graph production as customizable functions. This session will demonstrate a technique for organizing workflows that generate reproducible data visualizations. The technique divides the production of data visualizations into three sections, each governed by a dedicated function. The first function prepares the data for graphing, the second produces the graphic, and the third prints the image to disk. This applied session will walk the learner through the stages of developing such a chain of functions and demonstrate the advantages of such an approach in reproducible projects. Data and starter scripts will be provided. Software in focus: RStudio, R, specifically ggplot2 package.

Bio: Dr. Koval is an assistant professor with the Department of Health Management and Informatics at University of Central Florida. His background is in quantitative methods and he has a strong interest in data-driven models of human aging. Dr. Koval's program of work proposes to develop a system for population health surveillance that would focus on chronic diseases, with particular focus on mental health and substance use conditions, which tend to have high comorbidity rates, polysubstance use patterns, and slowly progressing pace of development.

The VADA Program
Visual and Automated Disease Analytics Graduate Training Program

Thursday June 13th

Using GitHub for Data Science: Version Control, Project Management, Promotion

Thursday June 13th 8:30am – 10:30am

Session Information: This session offers a hands-on walk through a minimalistic project designed to introduce the learner to basic operational utility of git and GitHub in data science projects. While git(hub) is best known for providing version control and streamlining collaboration in software development, its applications in data analytics projects can enhance the robustness of the produced solutions and help make results more visible and accessible to the community. The topics to be covered include: 1) project kickoff from a standard stencil, 2) using GitHub client for version control 3) using GitHub Issues for structuring tasks in project development 3) team control and communication and 4) designing an effective README page that faces the public. Data and starter scripts will be provided. Software in focus: R, RStudio, GitHub client.

Bio: Dr. Koval is an assistant professor with the Department of Health Management and Informatics at University of Central Florida. His background is in quantitative methods and he has a strong interest in data-driven models of human aging. Dr. Koval's program of work proposes to develop a system for population health surveillance that would focus on chronic diseases, with particular focus on mental health and substance use conditions, which tend to have high comorbidity rates, polysubstance use patterns, and slowly progressing pace of development.

Integrating GitHub and R for reproducible epidemiological and genomics research for precision medicine approaches to psychiatry

Thursday June 13th 10:30am – 11:15am

Session description: Schizophrenia is one of the top 15 leading causes of disability worldwide, with an average of ~30 years lost. Those with schizophrenia are at 15-25 times higher risk for suicide, compared to the general population. Features associated with poor outcomes in schizophrenia include childhood adversities (e.g. neglect), family history of psychiatric disorders, and schizophrenia genomic burden. There are currently no means of predicting, at first presentation, who will experience a poor outcome in schizophrenia and previous studies were limited by small sample size, and did not consider a range of outcomes. In this session, Dr. Kowalec will describe an investigation of poor clinical outcomes in schizophrenia using large population-based data and comprehensive genomic data from Sweden, Norway, and Denmark. Dr. Kowalec will describe the utility of GitHub and R for reproducible, trans-Nordic research.

Bio: Dr. Kaarina Kowalec joined the College of Pharmacy at the University of Manitoba in January 2019. Dr. Kowalec's research focus is on precision medicine approaches to neurology and psychiatry. Dr. Kowalec uses large administrative health datasets and genomic data to identify individuals with neurological or psychiatric disorders who are at high risk for experiencing poor outcomes (e.g. treatment-resistance, adverse drug reactions, mortality).

The VADA Program
Visual and Automated Disease Analytics
Graduate Training Program

Automated Pipelines for the Analysis of Genomic Data and Their Clinical Applications

Thursday June 13th 11:15am – 12:00pm

Session description: Next generation sequencing (NGS) technology has made it possible to have genomic profiles in patients with suspected genetic disorders. It has been proposed that whole genome sequencing should be used as the first-tier genetic test for such patients as the new standard of care. Nowadays, one of the greatest challenges of applying the NGS information in a clinical setting is the time, given that an average of >2 weeks is needed to analyze and interpret the data. In this session I will review some of the automated pipelines which have been developed to shorten the diagnosis time using NGS data. The advantages and disadvantages of these pipelines will be discussed, and examples of their clinical applications will be presented.

Bio: Dr. Xiao-Qing Liu is a genetic epidemiologist and assistant professor at the Department of Obstetrics and Gynaecology of the University of Manitoba. She is interested in using epidemiological and biostatistical methods to gene mapping of monogenic and complex disorders. She obtained her master's degree in epidemiology from the Public Health School of Johns Hopkin University and in medical genetics from Peking Union Medical College. She graduated with a medical degree from the Jiamusi University Medical College in China.

Data Science, AI, and Health: Where are we Headed?

Thursday June 13th 1:00pm – 2:30pm

Session Information: Currently, there is a lot of, and rapidly growing, interest and enthusiasm for how data technologies, artificial intelligence in particular, are poised to transform the entire field of health. This talk will discuss how we got to this point, the current state of cutting-edge health data science research, and where we are headed next. The truly multidisciplinary nature of health data science will be dissected including ethical, equity, and training issues. An overview of research happening at the Data Intelligence for Health Lab in the Cumming School of Medicine, University of Calgary will also be provided.

Bio: Dr. Joon Lee is the Director of the Data Intelligence for Health Lab and an Associate Professor of Health Data Science in the Cumming School of Medicine, University of Calgary. Prior to joining the University of Calgary, he held a faculty appointment in the School of Public Health and Health Systems at the University of Waterloo for 6 years. He holds a PhD in Biomedical Engineering from the University of Toronto and completed postdoctoral training in Medical Data Science at the Harvard-MIT Division of Health Sciences and Technology. His research applies data science, machine learning, artificial intelligence, natural language processing, mobile technology, and biostatistics to a wide range of health domains including intensive care medicine, aging, and population health surveillance. He received the Early Researcher Award from the Ontario Ministry of Research, Innovation and Science in 2016.

The VADA Program
Visual and Automated Disease Analytics Graduate Training Program

Data Visualization Strategies and Tools for Microbial Genomic Epidemiology

Thursday June 13th 2:45pm – 4:30pm

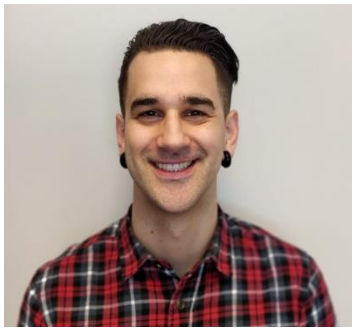
Session Information: New technologies are enabling public health agencies to collect more data of many different types, which can be used to inform public health policy and practice. Yet, this new "big data" is challenging to analyze and to communicate to stakeholders that need to make decisions. In this session a brief introduction to data visualization research will be provided – what it is, how it can be used, and what tools exist to help visualize data. Visualization of public health microbial genomic epidemiology will then be discussed. Different visualization techniques that are used in public health will be discussed as will how to use this knowledge to build data visualization tools and a visualization recommender system. As public health continues to evolve, it is becoming critical to build robust and actionable data visualization tools that support growing challenges of public health data-driven decision making.

Bio: Anamaria Crisan is a Vanier Canada Scholar and UBC Public Scholar in her final year of study in Computer Science at the University of British Columbia. Under the joint supervision of Dr. Tamara Munzner and Dr. Jennifer Gardy, she is researching how to visualize the heterogenous collections of genomic and public health data that support investigations of disease outbreaks. Prior to her PhD, Ana worked as a researcher with the British Columbia Centre for Disease Control and, separately, at a Vancouver-based start-up where she was responsible for the research and development of a commercially deployed prostate cancer genomic classifier. She holds a MSc in Bioinformatics from the University of British Columbia and a BSc in Computer Science from Queen's University. You can learn more about her research on her website.

Friday June 14th

Jupyter Notebook: A Modern Tool for Open, Reproducible, and Distributable Data Science

Friday June 14th 8:30am – 10:30am



Session Information: In the world of modern data science there exists a plethora of development and distribution tools available to the budding data scientist. Foremost amongst these options is the Jupyter Notebook: an open-source, web-based application that simplifies the process of sharing code, results, and associated documentation across multiple languages. This workshop will teach the basics of Jupyter from installation up to distribution using real world examples in R and Python. At the end of this 2-hour session you will have learned the tools and developed an understanding of using modern computational notebooks for reproducible data science.

Bio: Adrian Zetner is a computational biologist and member of the Bioinformatics section at the National Microbiology Laboratory. He collaborates with fellow scientists on genomics-related public health projects as well as heading the department's online and in-person training programs. Despite his initial reluctance, he has now embraced Jupyter notebooks as the ideal vehicle for computer-aided scientific analysis and communication. He is a graduate of the University of Manitoba Microbiology program, and a self-taught R champion.