The CODATA-RDA Research Data Science Applied workshops on Extreme sources of data, Bioinformatics and IoT/Big-Data Analytics | (smr 3178)

Monday, 24 July 2017

09:00 - 18:00 Extreme Sources of Data

Location: Adriatico Guest House

09:00 **Welcome** *15'*

09:15 **Introduction to Python programming language I** *2h45'*

12:00 **Lunch break** *2h0'*

14:00 **Introduction to Python programming language II** *1h45'*

15:45 **Coffee break** *30'*

16:15 **Python hands-on session** *1h45'*

09:00 - 18:00 IoT/Big Data Analytics

Location: Adriatico Guest House

09:00 **Introduction to Big Data & IoT Analytics Problem Scope. Analysis of Large Scale Real-Time and Streaming Data** *3h0'*

12:00 **Lunch break** *2h0'*

14:00 **Lab (Day 1) Set up simulator, services and IDE environment (Hortonworks Sandbox)** *1h45'*

15:45 **Coffee break** *30'*

16:15 **Group Discussion on Problem** *1h45'*

09:30 - 18:00 Bioinformatics

Abstract: In this module, we'll explore various technique used in Medical Genetics to find genomic variations responsible for complex and Mendelian traits. We'll start with association analysis in complex traits using genotyping data in thousands of individuals and will conclude with the analysis of Exome sequences in families with a Mendelian Pathology. For each technique, we'll illustrate the correlated biological and statistical problems, the rationale and potential pitfalls. In the afternoon, all the students will apply the techniques using hand-on tutorials.

Location: Adriatico Guest House

09:30 **Workshop Introduction** *15'*
Speaker: Luca Bortolussi (University of Trieste)

09:45 **Finding a needle in a haystack finding pathological mutations in the human genome** *1h15'*
Speaker: Pio d'Adamo (University of Trieste/Children's Hospital Burlo Garofolo)

11:00 **Coffee break** *30'*

11:30 **Finding a needle in a haystack finding pathological mutations in the human genome** *1h30'*
Speaker: Pio d'Adamo (University of Trieste/Children's Hospital Burlo Garofolo)

13:00 **Lunch break** *1h0'*

14:00 **Lab Session: Finding a needle in a haystack finding pathological mutations in the human genome** *1h30'*
Speaker: Pio d'Adamo (University of Trieste/Children's Hospital Burlo Garofolo)

15:30 **Coffee break** *30'*

16:00 **Lab Session: Finding a needle in a haystack finding pathological mutations in the human genome** *2h0'*
Speaker: Pio d'Adamo (University of Trieste/Children's Hospital Burlo Garofolo)

Tuesday, 25 July 2017

09:00 - 17:00 Extreme Sources of Data

Location: Adriatico Guest House

09:00 **Introduction to Particle Physics** *45'*

10:00 **The ATLAS experiment at the LHC** *45'*

11:00 **Discovering particles with ATLAS** *45'*

12:00 **Lunch break** *2h0'*

14:00 **An introduction to the Grid** *45'*

14:45 **Monte Carlo samples and tools** *45'*

15:30 **Coffee break** *30'*

16:00 **TBD** *1h0'*

09:00 - 18:00 Bioinformatics

Abstract: Next Generation Sequencing (NGS) technologies have led to discoveries of new diagnostic, prognostic and therapeutic targets. Despite these discoveries, treatment of cancer patients, detection of cancer biomarkers and prediction of therapy response remain largely unsolved problems. These difficulties are hindering the realisation of effective approaches to personalized medicine; and data needs to be better exploited to systematically elucidate the mechanisms and causes underlying cancer origination and development.

Cancers accumulate genetic mutations that allow their cells to proliferate out of control. Mutations occur randomly, are inherited through cell divisions, and orchestrate cancer initiation and development with accumulation patterns differing between individuals. NGS technologies are routinely used to detect mutations in tumoral biopsies, and free-access large collections of cancer datasets are now available. Cancer mutation profiles are incredibly heterogenous, and we observe few common mutations across patients even if their cancers have similar histological classification. Tumor Heterogeneity (TH) is intimately related to Cancer Evolution, and is considered to lead to the emergence of drug-resistance mechanisms, relapse and failure of treatments. Quantification of TH across cancer types and patients is of the utmost importance in modern cancer research.

I will present a causal framework to infer, from DNA sequencing data, Graphical Models that recapitulates the progression of the tumors (i.e., evolutionary models). This inference problem has several formulations, according to the type of NGS data that we have access to. I will discuss an approach that combines Statistics, Machine Learning and Formal Methods to infer models from single-sample data; and then I will move on to the problem of studying Cancer Evolution from multi-samples of the same individual. These two problems are orthogonal, and I will discuss attempts at defining a unique framework to study Cancer Evolution. Example applications with real data will be presented and discussed.

Location: Adriatico Guest House

09:00 **Data Science approaches to infer Cancer Progression Models** *1h30'*
Speaker: Giulio Caravagna (The University of Edinburgh)

10:30 **Coffee break** *30'*

11:00 **Data Science approaches to infer Cancer Progression Models** *1h30'*
Speaker: Giulio Caravagna (The University of Edinburgh)

12:30 **Lunch break** *1h30'*

14:00 **Lab Session: Data Science approaches to infer Cancer Progression Models** *1h30'*
Speaker: Giulio Caravagna (The University of Edinburgh)

15:30 **Coffee break** *30'*

16:00 **Lab Session: Data Science approaches to infer Cancer Progression Models** *2h0'*
Speaker: Giulio Caravagna (The University of Edinburgh)

09:00 - 18:00 IoT/Big Data Analytics
Location: Adriatico Guest House

09:00 **Data Collection, Preparation, Data Quality and Data Integration** *3h0'*

12:00 **Lunch break** *2h0'*

14:00 **Lab (Day 2) Ingesting and capturing real time events and streams (Apache NiFi & Kafka)** *1h45'*

15:45 **Coffee break** *30'*

16:15 **Lab Review** *1h45'*

Wednesday, 26 July 2017

09:00 - 17:00 Extreme Sources of Data
Location: Adriatico Guest House

09:00 **Introduction to ATLAS Open Data platform/tools** *3h0'*

12:00 **Lunch break** *2h0'*

14:00 **Introduction to hands-on session** *1h45'*

15:45 **Coffee break** *30'*

16:15 **ATLAS Open Data: Q&A** *45'*

09:00 - 18:00 Bioinformatics

Abstract: DNA aligners (such as BLAST, Bowtie or BWA) are very fast tools that allow searching occurrences of (short) DNA sequences in one or more (big) genomes. The idea behind these tools is to pre-process the genome file and build an index; such an index permits to search a DNA sequence in time proportional to its length, rather than to the length of the genome. Indexing accelerates DNA alignment by millions of times, but it introduces a problem: the index could be several times bigger than the text, exceeding the computer's RAM size. This is particularly concerning in view of recent developments in DNA sequencing technologies: projects such as the 1000 Genomes Project are producing thousands of sequenced genomes, which should be indexed in order to quickly align DNA sequences on them. Not all hope is lost, however. Two genomes from the same species are 99.99% identical, so compression techniques can be exploited to greatly reduce the index size. In this lecture I will introduce a famous compression and indexing technique that is having a huge impact in bioinformatics: the Burrows-Wheeler transform (BWT). We will see - both in theory and practice - how BWT-based aligners can achieve extremely high search speeds while taking (up to) thousands of times less space than the input collection of genomes.

09:00 **Aligning DNA sequences on compressed collections of genomes** *1h30'*
Speaker: Nicola Prezza (Technical University of Denmark)

10:30 **Coffee break** *30'*

11:00 **Aligning DNA sequences on compressed collections of genomes** *1h30'*
Speaker: Nicola Prezza (Technical University of Denmark)

12:30 **Lunch break** *1h30'*

14:00 **Aligning DNA sequences on compressed collections of genomes** *1h30'*
Speaker: Nicola Prezza (Technical University of Denmark)

15:30 **Coffee break** *30'*

16:00 **Lab Session: Aligning DNA sequences on compressed collections of genomes** *2h0'*
Speaker: Nicola Prezza (Technical University of Denmark)

09:00 - 18:00 IoT/Big Data Analytics
Location: Adriatico Guest House

09:00 **Data Analysis and Visualization** *3h0'*

12:00 **Lunch break** *2h0'*

14:00 **Lab (Day 3) Real time event processing, low latency query and visualization (HBase, Hive, Storm)** *1h45'*

15:45 **Coffee break** *30'*

16:15 **Lab Review & Industry Use Case** *1h45'*

Thursday, 27 July 2017

09:00 - 15:00 Bioinformatics

Abstract: High-throughput data sets from next-generation sequencing provide a rich but highly complex picture of the biological processes assayed. Statistical challenges abound, arising from high dimensionality, strong heterogeneity and general low replication of the data. In this talk, I will describe how techniques from machine learning and computational statistics can be effectively used to answer some of these questions. I will focus on the issues of statistical testing for epigenomic data such as ChIP- and BS-Seq, and determining isoform proportions/ splicing ratios from low coverage RNA-Seq data.

09:00 **Using machine learning to address challenges in high-throughput biology** *1h30'*
Speaker: Guido Sanguinetti (The University of Edinburgh)

10:30 **Coffee break** *30'*

11:00 **Using machine learning to address challenges in high-throughput biology** *1h30'*
Speaker: Guido Sanguinetti (The University of Edinburgh)

12:30 **Lunch break** *1h0'*

13:30 **Warp up and hands out certificates** *1h30'*
Speakers: Alberto Policriti (University of Udine), Guido Sanguinetti (The University of Edinburgh)

09:00 - 18:00 Extreme Sources of Data
Location: Adriatico Guest House

09:00 **Laboratory session: ATLAS Open Data hands-on tutorial** *3h0'*

12:00 **Lunch break** *2h0'*

14:00 **Laboratory session: ATLAS Open Data hands-on tutorial** *1h45'*

15:45 **Coffee break** *30'*

16:15 **Laboratory session review** *1h45'*

09:00 - 18:00 IoT/Big Data Analytics
Location: Adriatico Guest House

09:00 **Advanced Analytics** *3h0'*

12:00 **Lunch break** *2h0'*

14:00 **Lab (Day 4) Analyzing social media and customer sentiment (Apache NiFi & Solr)** *1h45'*

15:45 **Coffee break** *30'*

16:15 **Lab Review** *1h45'*

Friday, 28 July 2017

09:00 - 18:00 Extreme Sources of Data

Location: Adriatico Guest House

09:00 **Presentations of previous day's work/results by participants I** *3h0'*

12:00 **Lunch break** *2h0'*

14:00 **Presentations of previous day's work/results by participants II** *1h45'*

15:45 **Coffee break** *30'*

16:15 **Live virtual-visit to ATLAS Control Room** *1h0'*

17:15 **Summary session and hand-out of certificates** *45'*

09:00 - 18:00 IoT/Big Data Analytics

Location: Adriatico Guest House

09:00 **Wrap up: Relevant Big Data/IoT Use Cases & Challenges** *3h0'*

12:00 **Lunch break** *2h0'*

14:00 **Lab (Day 5) Group project presentations** *1h45'*

15:45 **Coffee break** *30'*

16:15 **TBD** *1h45'*