The Graduate School VLAG

WAGENINGEN UNIVERSITY & RESEARCH

Postgraduate Course

Big Data Analysis in the Life Sciences

Wageningen University, 21–23 October 2019

BACKGROUND

Technological and scientific advances have pushed the Life sciences into the BIG data era. The increased volume and heterogeneity of available data challenges scientists to master techniques for comprehensive data analysis. Extracting meaningful (biological) information from large datasets is increasingly becoming a challenge in all fields of life sciences. Thus, the ability to select and deploy analysis tools and algorithms has become an indispensable skill for all researchers. In this course, we aim to introduce participants to techniques for comprehensive data analysis of large, heterogeneous datasets and extract relevant information for elucidating biological design principles. The course is modular and focuses on data generation, mining, analysis, data integration, and visualization

TARGET GROUP

This course is aimed at anyone working with big datasets in the life sciences who has an interest in learning more about tools and possibilities for big data analysis. The course is introductory.

Required knowledge: Basic knowledge of statistics.

Prerequisite: Experience of computer programming would be useful but is not mandatory. Computer practicals will require using R. An R tutorial will be sent in advance to ensure that all participants have the required level.

COURSE LECTURERS

- **Dr Maria Suarez Diez**
Systems & Synthetic Biology, WUR
- **Dr Edoardo Saccenti**
Systems & Synthetic Biology, WUR
- **Jasper Koehorst MSc**
Systems & Synthetic Biology, WUR
- **Dr Guido Hooiveld**
Division of Human Nutrition and Health, WUR

PROGRAMME

Day 1:
- Linked data retrieval: querying biological resources using RDF and SPARQL.
- Brief introduction to Big data theoretical concepts: volume, variability, veracity, velocity
- Introduction to the Resource Description Framework (RDF) data format: triples, data types and objects
- Practical: Federated queries: advanced mining of (multiple) biological and biochemical database (such as UniProt, Reactome or EBI) using SPARQL
- Data reduction using multivariate statistics: PCA, theory and practical aspects
- Variance-Covariance matrix and correlations
- Data decomposition and PCA solution
- Methods for dimensionality assessment (statistical tests and computational approaches (permutation and cross-validation)
- Interpretation of PCA model: loading and biplots and limitations

Day 2:
- Sparse approaches to PCA: sparse PCA, groupwise PCA
- Introduction to PLS regression
- Collinearity
- Prediction and classification with PLS: Discriminant analysis Overfitting, Cross-validation and model optimization
- Sparse approaches to PLS
- Multilevel data integration using PLS and sparse PLS for different data types

Day 3:
- Network reconstruction: Associations networks based on information theoretical methods and correlation analysis
- Brief introduction to Bayesian networks
- R tools and practical on network reconstruction exploration
- Data visualization
- Introduction to Cytoscape and R to visualize omics data such as expression data

COURSE FEE [1]

PhD candidates affiliated with VLAG/EPS/PE&RC/WIAS/WASS/SENSE	€ 225
All other PhD candidates	€ 450
Postdocs, academic staff / Non-profit	€ 625
Participants from the private sector / For-profit	€ 1200

[1]includes course materials, lunches/tea/coffee and one course dinner

REGISTRATION AND INFORMATION

https://www.vlaggraduateschool.nl/en/courses/course/BDA19.htm

For information contact:
Cornelia van Bree-Evers
E-mail: cornelia.vanbree-evers@wur.nl