

## **Data Mining and Knowledge Discovery – an introductory course with focus on veterinary epidemiology applications**

The increasing quantity of electronic data available (including clinical, laboratorial or research data) is allowing epidemiological intelligence to move from hypothesis testing to knowledge discovery. Data mining technology allows researchers to analyse large observational data sets, previously collected for different purposes, to summarize the data in novel ways or to find unsuspected relationships. Rather than asking constrained questions, researchers can “let the data speak for themselves”, using computer algorithms to unveil hidden patterns. This data-driven process adopts fundamentally different approaches to traditional question-driven methods. Data mining has been used in epidemiology for early detection of outbreaks (syndromic surveillance), studies of antimicrobial resistance, among others. This course provides an introduction to machine learning techniques, focusing on unsupervised techniques – those in which training datasets are not required. Participants are expected to have a basic knowledge of biostatistics.

### **Workshop specifications**

The instructors are prepared to offer this 2-day course as a pre- or post-conference workshop. The minimum number of participants is 10, and the maximum number of participants would be 25. Participants are required to bring their own laptop computer. The course exercises will use *RapidMiner*, the world-leading open-source system for data mining available freely from Rapid-I (<http://rapid-i.com/content/view/26/84/>). Participants will receive download and installation instructions upon registration. Participants can bring a dataset of their own to explore these knowledge discovery techniques; datasets will also be provided as part of the course.

### **Workshop Content**

**Day 1** – Introduction to Data Mining and Knowledge Discovery. The theory will cover concepts of supervised and unsupervised machine learning. Participants will work with hands-on exercises in Data Preparation, Variable Reduction/Transformation and Data Visualization of large datasets.

**Day 2** – Unsupervised Machine Learning techniques. Theory will include techniques to discover patterns, hidden structures and associations in large datasets. This will cover association rules and clustering methods, including Self-Organizing/Kohonen Maps, K-Means, and hierarchical approaches. Participants will work on exercises using the datasets provided. Participants wishing to bring their own dataset can explore these with the instructors’ assistance, according to time availability during the course.

### **Course Fees**

Course fees are proposed as 350€ for professionals and 250€ for students.

## Instructors' Biographies

### **Crawford Revie ([crevie@upei.ca](mailto:crevie@upei.ca))**

Crawford is a computing scientist who came to veterinary epidemiology by way of a doctorate in mathematical modelling. He currently holds the Canada Research Chair in *Epi-informatics*, roughly translated as the application of a wide range of informatics tools and approaches to epidemiology. Prior to moving to the Atlantic Veterinary College he was based in Glasgow where he was a member of emerging groups in the areas of Veterinary Informatics and Quantitative Epidemiology at the two Scottish vet schools. A major focus of his research over the past decade has been the application of data-driven models to disease control in a range of veterinary contexts. Recently he has also led research projects in the area of syndromic surveillance, working with a range of animal species in both Canada and sub-Saharan Africa.



### **Fernanda Dórea ([fdorea@upei.ca](mailto:fdorea@upei.ca))**

Fernanda is a veterinarian from Brazil. She contributed to various animal health programs in Brazil, her experience ranging from field work in the Amazon, to the 2 years as an Epidemiologist in the headquarters of the Brazilian Ministry of Agriculture, Livestock and Food Supply. Driven by her growing interest in quantitative epidemiology, she pursued a Masters degree in Infectious Disease Modeling at the University of Georgia, USA. She is now venturing into computer sciences, exploring ways to automatically extract surveillance information from electronically available data in animal health. Her doctorate work, under the supervision of Drs. Javier Sanchez and Crawford Revie at the University of Prince Edward Island, Canada, is scheduled for completion in the Spring of 2013.

