Data Science and Analytics Module descriptions (2019-2020)

CS620C Structured programming

Programming fundamentals: variables, types, expressions and assignment; simple I/O; Conditional and iterative control structures (if statements and while loops); Strings and string processing; Use of class APIs for creating objects and calling methods; Understanding data abstraction and encapsulation; Problem solving: understanding and developing algorithms; Implementing algorithms as simple programs.
Introduction to algorithms and data structures. Review of elementary programming concepts suitable for the implementation of abstract data types (operators, types and expressions; control of flow; methods; recursion; input & output); Algorithms for searching: linear, bounded linear and binary searches; Algorithms for sorting: selection, insertion, bubble and quick sorts; Fundamental linear data structures: stacks, queues, linked lists; Object-oriented programming: encapsulation and information hiding, classes, interfaces, class hierarchies, inheritance, polymorphism, basic exception handling; Analysis of basic algorithms.

CS621C Spatial databases

Conceptual spatial data modeling; relational database models for spatial data; spatial data structures; Structured Query Language (SQL); spatial data integrity; spatial data manipulation; Geographic Information Systems (GIS); spatial analysis techniques including statistical approaches for spatially-informed decision making; spatial data visualisation; basic transaction processing; database security; spatial data formats; alternative database models for geospatial data.

ST661 R for data analytics

The module objective is to equip students to carry out data science projects in R. Basics of R: Data structures, control structures, functions. Exploratory data analysis. Reports with Rmarkdown. Data wrangling with dplyr. Visualisation: ggplot, shiny. Big data techniques, Sparklyr.

ST463/ ST683 Linear models 1

Scatterplots and regression. Simple linear regression: estimation, parameter inference, prediction, analysis of variance, R2, F-test, correlation, residual plots, assessing normality. The multiple regression model and its applications. Matrix notation, mean vectors and covariance matrices. Least-squares estimation. The multivariate normal distribution. Hypothesis testing and confidence intervals, prediction. Analysis of variance, R2, sequential sums of squares, general F-tests. Added variable plots. Model checking via testing for lack of fit and residual plots.
Regression diagnostics: residuals, leverage, outliers, influence and Cook's Distance.

ST465/ ST685 Linear models 2

Polynomials and factors in regression. Weighted least squares. Transformation of response and predictors, Box-Cox method, variance stabilizing transformation. Collinearity and variance inflation factors. Model selection: forwards, backwards and stepwise. All possible regressions, Mallow's Cp and PRESS. Cross-validation. Auto-correlation, Durbin-Watson test. AR(1) models. Linear mixed model theory, single random effects, multiple random effects, repeated measures, random coefficient models.

ST663 Statistical methods for data science

This course covers probability and statistical techniques for data analytics and data science. Topics include: Exploratory data analysis and visualisation. Probability basics, independence, Bayes theorem. Probability models for data, including Binomial, Poisson, Exponential and Normal. Parameter estimation; method of moments and maximum likelihood. Confidence intervals and hypothesis testing: one and two samples, paired samples, proportions. Simple linear and multiple regression. Analysis of Variance. Case studies in R.

NCG608 Introduction to Geographical Information Science

The module will introduce the main complementary methodologies with geocomputation: geographical information systems; spatial statistics; exploratory spatial data analysis; and the science of spatial data handling. It covers fundamental concepts, techniques and ideas that shape Spatial eHumanities and associated GIS Software. The module will be taught as a mix of lectures and practicals. The lecture series will introduce key concepts and analytical approaches used within GI Science including; the foundations of GIS, spatial data models, data input and output, core spatial modelling and specialist analytical approaches and techniques. The practicals will be based around FOSS software including QGIS and R

NCG612 Case studies in data science and analytics

In this course, students will explore real-world case studies in data analytics. The range of topics will include areas such as crime pattern analysis, house price prediction, the modelling of epidemics and the analysis of textual data. The topics will be introduced by a mixture of lecturers from the NCG and external practitioners.

Emphasis will be based on reflection on the case studies, in terms of appropriateness of analytical approaches, reliability and robustness of findings, and potential future work to address unanswered questions.

NCG613 Data analytics project
Students will complete a significant data analytics project. This will require the identification of analytical questions to be addressed; identification of appropriate sources of data; consideration of appropriate analytical techniques to apply; choice of software tools to be used; and reporting of and reflection upon the results of the analysis. Students will work in teams and assessment will be based on a project diary, a final report and a team presentation.

NIR605 Critical data studies
There is a long history of governments, businesses, science and citizens producing and utilising data in order to monitor, regulate, profit from, and make sense of the world. In general, data are taken at face value. This module, however, will critically interrogate the nature of data, how they are being produced, organised, analysed and employed, and how best to make sense of them and the work they do. In other words, it will employ a more philosophical approach to data. The course will provide:
(1) a detailed overview of big data, open data and data infrastructures;
(2) an introduction to thinking conceptually about data, data infrastructures, data analytics and data markets;
(3) a critical discussion of the technical shortcomings and the social, political and ethical consequences of the data revolution;
(4) an analysis of the implications of the data revolution to academic, business and government practices.
The core book accompanying the course will be: Kitchin, R. (2014) The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences. Sage, London.

ST464/ST684 Statistical machine learning
Introduction to statistical learning. Classification: logistic regression, LDA and QDA. Unsupervised learning: PCA, clustering methods. Resampling methods: cross-validation and the bootstrap. Linear Model selection and regularisation. Modelling with regression splines and smoothing splines. Tree-based methods. Support vector machines. Implementation in R.

ST466/ST686 Advanced statistical modelling
Categorical data methods. Generalized linear models, exponential families. Logistic regression for binary responses and Binomial counts. Loglinear models for Poisson counts, two and three-way tables. Introduction to Bayesian inference. Bayesian analysis for simple models: conjugate prior distributions, inference for mean of normal data, linear regression model. Simulation based Bayesian analysis: using Bugs/jags from R. Predictive distribution and model checking.

ST662 Topics in data analytics
Data science overview. Data analytics life cycle. A selection of topics such as the following will be covered: Data manipulation and cleaning. Statistical programming skills in software such as SAS or R. Interactive data visualisation. Missing data techniques. Classification with naive Bayes. Time series. Text analysis. Graphs and social networks. Guest Lectures to present examples of data science projects in industry.

CS615C Internet solutions engineering for data scientists
The course starts with an introduction to information processing followed by look at components and architecture of WWW and Internet. The course then deals with client-side and server-side technologies, frameworks, programming languages and protocols. Specifically HTML5, CSS, JavaScript, JQuery, AJAX, Server-side technology (ASP.NET, JSP, PHP), SOAP Web services and REST web services are explained using practical and real-world examples. All the concepts described with focus on progressive enhancements (separation of structure, presentation and behaviour) and best practices in web application development. The course also deals with mobile web development and using standard web technologies (HTML5, CSS and JavaScript) to create hybrid mobile apps. Internet solutions engineering using best-practice design patterns are examined and used to deploy a large-scale multi-tiered assignment.

ST606/GT606/CS648 Masters Project and Dissertation