# ST419-15 Advanced Topics in Data Science

## 23/24

**Department**
Statistics
**Level**
Undergraduate Level 4
**Module leader**
Dario Spano
**Credit value**
15
**Module duration**
10 weeks
**Assessment**
Multiple
**Study location**
University of Warwick main campus, Coventry

---

## Description

### Introductory description

This module runs in Term 2 and will be comprised of three selected topics in the area of computational challenges associated with data analysis. The topics may change year to year.

Some examples of topics from previous academic years:
Deep Learning for Natural Language Processing, Decision Trees and Random Forests, Model Comparison and Selection, Artificial Neural Networks, Introduction to Reinforcement Learning and Modelling the Written Word: Compression and Human-Computer-Interfaces.

Students will be given selected advanced research material for independent study and examination.

This module is available for students on a course where it is a listed option and as an Unusual Option to students who have completed the prerequisite modules.

Pre-requisites: ST219 Mathematical Statistics B OR ST220 Introduction to Mathematical Statistics OR OR CS260 Algorithms.

Module web page

## Module aims

Data Science is an important frontier in the mathematical sciences and employers across a number of sectors are looking for graduates with strong computational and statistical skills. The aim of this module is to provide students with a working knowledge of three selected topics that emphasize the interplay between data analysis and computation.

## Outline syllabus

This is an indicative module outline only to give an indication of the sort of topics that may be covered. Actual sessions held may differ.

Topics will vary from year to year. The module's advanced topic will include material for independent study and assessment This material will reflect current trends, innovations or expertise in the area.

Example syllabi include:

1. Numerical optimization: algorithms to find the minimizer of convex function e.g. gradient descent, Newton and quasi-Newton methods. Rates of convergence and computational costs associated with these algorithms in general and / or specific settings. Discussion of relative merits of different methodology.
2. The MapReduce programming model: modern approaches to scaling-up computation via distribution and parallelization, such as the map-reduce programming model, and systems such as Hadoop and Spark.
3. Streaming algorithms: algorithms to process massive streams of data. Hashing, sketching and randomization. Probabilistic counting, counting distinct elements, count-min sketch.
4. High-dimensional regression and variable selection: methods for regression with large datasets, and methods for determining which covariates are important. Statistical and computational issues relating to large numbers of covariates and / or measurements. Ridge regression, the LASSO, and variable selection. Cross-validation. Screening.
5. Coding theory: compression and error detection. Lossless coding, entropy, Shannon's theorem. Symbol and dictionary-based approaches. Error-correcting codes, parity checks, Hamming (7,4)-code.
6. Deep Learning for Natural Language Processing, introduction to the theory and practice of deep NNs with focus on the applications in natural language processing (NLP). Neural network architectures such as convolutional NNs, recurrent NNs, attention mechanisms, transformers, sequence-to-sequence learning and (time-permitting) generative adversarial networks and variational autoencoder. key concepts of artificial NNs, such as activation functions, layers, weights and gradient descent for fitting a NN;
7. Decision Trees and Random Forests: CHAID, C4.5, C5.0, ID3 and CART algorithms, bagging, boosting, random forests, the pros and cons of such approaches. Software examples (e.g. R package CHAID and rpart).
8. Model comparison and selection, Scientific validity in the context of the data analytics workflow, Basic model-agnostic assessment of supervised/predictive models, Performance quantification of models. Predictive model validationin R/mlr and python/sklearn Julia/MLJ. Statistical formulation of the supervised learning setting,

Bias-variance trade-off, cross-validation and re-sampling estimators, Estimators of the generalization loss and the loss‘s variance, Hypothesis testing for pairwise and portmanteau model comparison, Meta-strategies for automated model improvement, Interaction of model tuning and model validation workflows

9. Artificial Neural Networks. Artificial neural networks (NNs) are a class of learning algorithms for regression, classification, and unsupervised learning that mimic real neural networks. They are very flexible and have become hugely popular in recent years. This topic will provide an introduction to the theory and practice of artificial NNs for supervised learning, building up from simple single layer feed-forward networks to complex multi-layer 'deep' architectures. We will cover some theory such as universal approximation theorems, as well as practicalities like training and regularization. Convolutional Nns. Advanced component: recurrent NNs and unsupervised NNs.
10. Reinforcement Learning: Reinforcement Learning (RL) is one of the main subfields of machine learning, alongside supervised and unsupervised learning, that focuses on decision making under implicit feedback. As such, it is heavily employed and developed in areas such as robotics and Al engines in games like Go and Chess.
    This topic will introduce the field of RL and standard agent-environment framework, covering Bellman's equations, dynamic programming, Monte Carlo and Temporal-Difference learning. Advanced Component: eligibility traces, function approximation.
11. Modelling the Written Word: Modelling of written words, viewed as streams of symbols from a finite alphabet, is a rich field with an extensive literature. This topic will provide an introduction to some probabilistic approaches to this problem and will show how these models can be used to efficiently store written text and also to provide efficient mechanisms for entry of text into computer systems which can be used without mastering the keyboard. Advanced Component: Grammar-based language models.

## Learning outcomes

By the end of the module, students should be able to:

- Demonstrate understanding of the three selected topics.
- Appreciate the computational challenges associated with data analysis and use some techniques developed to meet these challenges.
- Be able to critically appraise the use of these topics.
- Be able to understand current research and developments in the three selected topics.
- Be critical of current research developments in the three selected topics.

## Indicative reading list

General texts in the correct area:
Hastie, T. and Tibshirani R. (2009)"The Elements of Statistical Learning ", Corr. 9th printing 2017 edition; Springer
Bishop, C.M. (2008) "Pattern Recognition and Machine Learning" ; Springer-Verlag New York
Other texts will be specified depending on the topics covered.

View reading list on Talis Aspire

### Subject specific skills

This will depend on the topic but will be the ability to understand, evaluate and apply various complex statistical, computational and machine learning tools to a variety of datasets. Students will develop skills in the use of appropriate software.

### Transferable skills

The general understanding of data from a variety of contexts. The ability to identify and find new data analysis techniques and to then learn them from suitable documentation. Be able to learn coding type software. To be able to appraise and discuss recent research in a technical area.

---

## Study

## Teaching split

| Provider | Weighting |
|---|---|
| Computer Science | 66% |
| Statistics | 34% |

## Study time

| Type | Required | Optional |
|---|---|---|
| Lectures | 30 sessions of 1 hour (20%) | 2 sessions of 1 hour |
| Private study | 120 hours (80%) | |
| Total | 150 hours | |

### Private study description

Study of advanced topic, weekly revision of lecture notes and materials, wider reading, practice exercises and preparing for examination.

## Costs

No further costs have been identified for this module.

---

## Assessment

You must pass all assessment components to pass the module.

Students can register for this module without taking any assessment.

### Assessment group B3

| | Weighting | Study time |
|---|---|---|
| In-person Examination | 100% | |

The examination will contain one compulsory question on the advanced topic and four additional questions of which the best marks of TWO questions will be used to calculate your grade.

---

- Answerbook Pink (12 page)
- Students may use a calculator

### Assessment group R2

| | Weighting | Study time |
|---|---|---|
| In-person Examination - Resit | 100% | |

The examination will contain one compulsory question on the advanced topic and four additional questions of which the best marks of TWO questions will be used to calculate your grade.

---

- Answerbook Pink (12 page)
- Students may use a calculator

### Feedback on assessment

Solutions and cohort level feedback will be provided for the examination.

Past exam papers for ST419

---

## Availability

### Anti-requisite modules

If you take this module, you cannot also take:

- ST343-15 Topics in Data Science

## Courses

This module is Optional for:

- Year 1 of TMAA-G1PE Master of Advanced Study in Mathematical Sciences

- Year 1 of TMAA-G1PD Postgraduate Taught Interdisciplinary Mathematics (Diploma plus MSc)
- Year 1 of TMAA-G1P0 Postgraduate Taught Mathematics
- Year 1 of TMAA-G1PC Postgraduate Taught Mathematics (Diploma plus MSc)
- TMAA-G1PF Postgraduate Taught Mathematics of Systems
    - Year 1 of G1PF Mathematics of Systems
    - Year 1 of G1PF Mathematics of Systems
- Year 1 of TESA-H1B1 Postgraduate Taught Predictive Modelling and Scientific Computing
- Year 1 of TSTA-G4P1 Postgraduate Taught Statistics
- USTA-G300 Undergraduate Master of Mathematics,Operational Research,Statistics and Economics
    - Year 3 of G300 Mathematics, Operational Research, Statistics and Economics
    - Year 4 of G300 Mathematics, Operational Research, Statistics and Economics

This module is Option list A for:

- USTA-G1G3 Undergraduate Mathematics and Statistics (BSc MMathStat)
    - Year 3 of G1G3 Mathematics and Statistics (BSc MMathStat)
    - Year 4 of G1G3 Mathematics and Statistics (BSc MMathStat)
- USTA-G1G4 Undergraduate Mathematics and Statistics (BSc MMathStat) (with Intercalated Year)
    - Year 4 of G1G4 Mathematics and Statistics (BSc MMathStat) (with Intercalated Year)
    - Year 5 of G1G4 Mathematics and Statistics (BSc MMathStat) (with Intercalated Year)

This module is Option list B for:

- Year 4 of USTA-G304 Undergraduate Data Science (MSci)
- Year 4 of UCSA-G4G3 Undergraduate Discrete Mathematics
- Year 5 of UCSA-G4G4 Undergraduate Discrete Mathematics (with Intercalated Year)

This module is Option list D for:

- USTA-G300 Undergraduate Master of Mathematics,Operational Research,Statistics and Economics
    - Year 4 of G30C Master of Maths, Op.Res, Stats & Economics (Operational Research and Statistics Stream)
    - Year 4 of G30C Master of Maths, Op.Res, Stats & Economics (Operational Research and Statistics Stream)
- Year 5 of USTA-G301 Undergraduate Master of Mathematics,Operational Research,Statistics and Economics (with Intercalated

This module is Option list E for:

- Year 4 of USTA-G300 Undergraduate Master of Mathematics,Operational Research,Statistics and Economics
- Year 5 of USTA-G301 Undergraduate Master of Mathematics,Operational Research,Statistics and Economics (with Intercalated

This module is Option list F for:

- Year 3 of USTA-G300 Undergraduate Master of Mathematics,Operational Research,Statistics and Economics
- USTA-G301 Undergraduate Master of Mathematics,Operational Research,Statistics and Economics (with Intercalated
    - Year 3 of G30H Master of Maths, Op.Res, Stats & Economics (Statistics with Mathematics Stream)
    - Year 4 of G30H Master of Maths, Op.Res, Stats & Economics (Statistics with Mathematics Stream)