Prescription

Big Data refers to the large and often complex datasets generated in the modern world: data sources such as commercial customer records, internet transactions, environmental monitoring. This course provides an introduction to the theory and practice of working with Big Data. Students enrolling in this course should be familiar with the basics of machine learning, data mining, statistical modelling and with programming.

Course learning objectives

Students who pass this course should be able to:

1. Identify properties and challenges of very large data sets in order to determine appropriate analysis techniques to apply a specific Big Data task.
2. Identify challenges in high-dimensional data and choose appropriate dimensionality reduction methods, from a software library such as Weka, to solve high-dimensional problems.
3. Analyse regression and clustering data to choose appropriate analysis methods with good parameter settings from a software library such as R to address regression and clustering problems and to generate data visualisations.
4. Based on an understanding of Hadoop MapReduce and Apache Spark, implement relevant algorithmic analysis of Big Data problems using appropriate machine learning libraries.

Course content

Section 1 Introduction to Big Data

- What is Big Data ?
- Where does Big Data come from?
- What we can do and what we should do with Big Data ?
- Typical examples of Big Data analysis in real word

Section 2 Machine learning for high-dimensional data

- Data Preprocessing and Introduction to Feature Manipulation
- Machine learning for high-dimensional data, dimensionality reduction and feature selection (and possibly missing data analysis) Wrapper, filter and embeded dimensionality reduction method
- The techniques covered will include sequential forward selection, sequential backword selection, and other machine learning methods such as decision trees, random forest, support vector machines, genetic programming (and possibly particle swarm optimisation).

Section 3 Regression, Clustering and other Techniques in Big Data

- Regression: ridge regression, local regression, lasso; curse of dimensionality
- Generalized additive models; case study on intelligible models in healthcare applications.

- Clustering and resampling methods.

Section 4 Big Data Tools/Project

- Hadoop MapReduce
- Apache Spark
- Spark Machine Learning Libraries

Withdrawal from Course

Withdrawal dates and process:
https://www.wgtn.ac.nz/students/study/course-additions-withdrawals

Lecturers

Bing Xue (Coordinator)

bing.xue@vuw.ac.nz 04 4635542

352 Cotton, Kelburn

Mengjie Zhang

Mengjie.Zhang@vuw.ac.nz 04 4635654

355 Cotton, Kelburn

Teaching Format

There are three slots for this course (one hour duration per slot). Two hours will be lectures and the other hour will be tutorials.

Student feedback

Student feedback on University courses may be found at:
www.cad.vuw.ac.nz/feedback/feedback_display.php

Dates (trimester, teaching & break dates)

- Teaching: 02 March 2020 - 07 June 2020
- Break: 13 April 2020 - 27 April 2020
- Study period: 08 June 2020 - 11 June 2020
- Exam period: 12 June 2020 - 27 June 2020

Class Times and Room Numbers

02 March 2020 - 22 March 2020

- **Monday** 11:00 - 11:50 – LT102, Murphy, Kelburn
- **Tuesday** 11:00 - 11:50 – LT102, Murphy, Kelburn
- **Wednesday** 11:00 - 11:50 – LT102, Murphy, Kelburn

27 April 2020 - 07 June 2020

- **Monday** 11:00 - 11:50 – LT102, Murphy, Kelburn
- **Tuesday** 11:00 - 11:50 – LT102, Murphy, Kelburn
- **Wednesday** 11:00 - 11:50 – LT102, Murphy, Kelburn

Other Classes

no

Set Texts and Recommended Readings

Required

There are no required texts for this offering.

Mandatory Course Requirements

In addition to achieving an overall pass mark of at least 50%, students must:

- submit reasonable attempts for at least two out of the three assignments. (Justification: The practical skills that are obtained in the assignments are a critical part of the CLO's, and engagement with a minimum of two of the assignments is considered essential.)

If you believe that exceptional circumstances may prevent you from meeting the mandatory course requirements, contact the Course Coordinator for advice as soon as possible.

Assessment

Assessment Item	Due Date or Test Date	CLO(s)	Percentage
Assignment 1 (3 weeks) (Analysis and report)	Week 4/5	CLO: 1,2	20%
Assignment 2 (3 weeks) (Analysis and report)	Week 7	CLO: 3	20%
Assignment 3 (3 weeks) (Analysis and report)	Week 11	CLO: 4	20%
Final exam (2 hours)	Exam Period	CLO: 1,2,3,4	40%

Penalties

The penalty for assignments that are handed in late without prior arrangement is one grade reduction per day. Assignments that are more than one week late will not be marked.

Extensions

Individual extensions will only be granted in **exceptional personal circumstances**, and should be negotiated with the course coordinator before the deadline whenever possible. Documentation (eg,

medical certificate) may be required.

Submission & Return

All work should be submitted through the ECS submission system, accessible through the course web pages. Marks and comments will be returned through the ECS marking system

Workload

In order to maintain satisfactory progress in COMP 424, you should plan to spend an average of at least 10 hours per week on this paper. A plausible and approximate breakdown for these hours would include:

- Lectures and tutorials: 3
- Readings: 2-4
- Assignments: 3-5

However, since this is multidisciplinary course, students with different background may need different amounts of time to work on different sections/assignments of the course, i.e. could be more or could be less.

Teaching Plan

See: https://ecs.wgtn.ac.nz/Courses/COMP424_2020T1/LectureSchedule

Communication of Additional Information

All online material for this course can be accessed at https://ecs.wgtn.ac.nz/Courses/COMP424_2020T1/

Links to General Course Information

- Academic Integrity and Plagiarism: https://www.wgtn.ac.nz/students/study/exams/integrity-plagiarism
- Academic Progress: https://www.wgtn.ac.nz/students/study/progress/academic-progess (including restrictions and non-engagement)
- Dates and deadlines: https://www.wgtn.ac.nz/students/study/dates
- Grades: https://www.wgtn.ac.nz/students/study/progress/grades
- Special passes: Refer to the Assessment Handbook, at https://www.wgtn.ac.nz/documents/policy/staff-policy/assessment-handbook.pdf
- Statutes and policies, e.g. Student Conduct Statute: https://www.wgtn.ac.nz/about/governance/strategy
- Student support: https://www.wgtn.ac.nz/students/support
- Students with disabilities: https://www.wgtn.ac.nz/st_services/disability/
- Student Charter: https://www.wgtn.ac.nz/learning-teaching/learning-partnerships/student-charter
- Terms and Conditions: https://www.wgtn.ac.nz/study/apply-enrol/terms-conditions/student-contract
- Turnitin: http://www.cad.vuw.ac.nz/wiki/index.php/Turnitin
- University structure: https://www.wgtn.ac.nz/about/governance/structure
- VUWSA: http://www.vuwsa.org.nz

Offering CRN: 31156

Points: 15
Prerequisites: One of (COMP 307, 309, STAT 393, 394); STAT 193 or ENGR 123 or approved background in Statistics;

Restrictions: COMP 473 (2016-2018)
Duration: 02 March 2020 - 28 June 2020
Starts: Trimester 1
Campus: Kelburn