



## 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. Explain the challenges in high-dimensional data and choose appropriate dimensionality reduction methods, from a software library such as KNIME and Orange, 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. Use their understanding of tools such as Hadoop MapReduce and Apache Spark to implement relevant algorithmic analysis of Big Data problems using appropriate machine learning libraries.

## Course content

The course is primarily offered in-person, but there will also be a remote option and there will be online alternatives for all the components of the course for students who cannot attend in-person.

Students taking this course remotely must have access to a computer with camera and microphone and a reliable high speed internet connection that will support real-time video plus audio connections and screen sharing. Students must be able to use Zoom; other communication applications may also be used. A mobile phone connection only is not considered sufficient. The computer must be adequate to support the programming required by the course: almost any modern windows, macintosh, or unix laptop or desktop computer will be sufficient, but an Android or IOS tablet will not.

If the assessment of the course includes tests, the tests will generally be run in-person on the Kelburn campus. There will be a remote option for students who cannot attend in-person, but the remote option imposes extra costs on the School and will be limited to students with a strong justification (for example, being enrolled from overseas). The remote test option will use the ProctorU system for online supervision of the tests. ProctorU requires installation of monitoring software on your computer which also uses your camera and microphone, and monitors your test-taking in real-time. Students who will need to use the remote test option must contact the course coordinator in the first two weeks to get permission and make arrangements.

=====

### 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 world

### 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 embedded dimensionality reduction method
- The techniques covered will include sequential forward selection, sequential backward 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

---

### Qi Chen (Coordinator)

qi.chen@vuw.ac.nz

---

### Bach Nguyen

bach.nguyen@vuw.ac.nz

---

### Bing Xue

bing.xue@vuw.ac.nz 04 4635542

352 Cotton, Kelburn

## Teaching Format

This course will be offered in-person and online. For students in Wellington, there will be a combination of in-person components and web/internet based resources. It will also be possible to take the course entirely online for those who cannot attend on campus, with all the components provided in-person also made available online.

Two lectures per week, with associated assignments. Additional content may be provided through video resources.

## Student feedback

Student feedback on University courses may be found at:  
[www.cad.vuw.ac.nz/feedback/feedback\\_display.php](http://www.cad.vuw.ac.nz/feedback/feedback_display.php)

## Dates (trimester, teaching & break dates)

- Teaching: 22 February 2021 - 28 May 2021
- Break: 05 April 2021 - 18 April 2021
- Study period: 31 May 2021 - 03 June 2021
- Exam period: 04 June 2021 - 19 June 2021

## Class Times and Room Numbers

### 22 February 2021 - 04 April 2021

- **Monday** 13:10 - 14:00 – LT001, Hugh Mackenzie, Kelburn
- **Thursday** 13:10 - 14:00 – LT101, Murphy, Kelburn

### 19 April 2021 - 30 May 2021

- **Monday** 13:10 - 14:00 – LT001, Hugh Mackenzie, Kelburn
- **Thursday** 13:10 - 14:00 – LT101, Murphy, Kelburn

## Set Texts and Recommended Readings

### Required

There are no required texts for this offering.

## Mandatory Course Requirements

There are no mandatory course requirements for this course.

*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 (25 hours)	Monday Week 5	CLO: 1,2,3	20%
Assignment 2 (25 hours)	Friday Week 8	CLO: 1,2,3	25%
Test (50 Minutes)	Monday Week 10	CLO: 1,2,3	25%
Assignment 3 (25 hours)	Tuesday Second Week of Assessment Period	CLO: 4	30%

## 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 AIML 427, 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: 2
- 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/AIML427\\_2021T1/LectureSchedule](https://ecs.wgtn.ac.nz/Courses/AIML427_2021T1/LectureSchedule)

## Communication of Additional Information

All online material for this course can be accessed at [https://ecs.wgtn.ac.nz/Courses/AIML427\\_2021T1/](https://ecs.wgtn.ac.nz/Courses/AIML427_2021T1/)

## 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-progress> (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/](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:** [33069](#)

**Points:** 15

**Prerequisites:** one of (AIML 420, 421, COMP 307, 309, STAT 393, 394); one of (ENGR 123, STAT 193, MATH 177, QUAN 102) or comparable background in Statistics

**Restrictions:** COMP 424, COMP 473 (2016-2018);

**Duration:** 22 February 2021 - 20 June 2021

**Starts:** Trimester 1

**Campus:** Kelburn