

# CS430-15 Foundations of Data Analytics

**26/27**

**Department**

Computer Science

**Level**

Undergraduate Level 4

**Module leader**

Ayse Sunar

**Credit value**

15

**Module duration**

10 weeks

**Assessment**

Multiple

**Study location**

University of Warwick main campus, Coventry

---

## Description

### Introductory description

Students will study techniques for how to go from raw data to a deeper understanding of the patterns and structures within the data, to support making predictions and decision making.

### Module aims

To understand the foundational skills in data analytics, including preparing and working with data; abstracting and modeling an analytic question; and using tools from statistics, learning and mining to address the question

### 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.

Data Analytics involves being about to go from raw data to a deeper understanding of the patterns and structures within the data, to support making predictions and decision making. The course will cover a number of topic, including:

Introduction to analytics, case studies - How analytics is used in practice. Examples of successful analytics work from companies such as Google, Facebook, Kaggle, and Netflix. Suggestions for the course project.

Basic tools: command line tools, plotting tools, programming tools - The wide variety of tools available to work with data, including unix/linux command line tools for data manipulation (sorting, counting, reformatting, aggregating, joining); tools such as gnuplot for displaying and visualising data; advanced programming tools such as Perl and Python for powerful data manipulation.

Statistics: Probability recap, distributions, significance tests, R - The tools from statistics for understanding distributions and probability (means, variance, tail bounds). Hypothesis testing for determining the significance of an observation.

Database: Data quality, data cleaning, Relational data - Problems found in realistic data: errors, missing values, lack of consistency, and techniques for addressing them.

Regression: linear regression, least squares, logistic regression - Predicting new data values via regression models. Simple linear regression over low dimensional data, regression for higher dimensional data via least squares optimisation, logistic regression for categorical data.

Matrices: Linear Algebra, SVD, PCA - Matrices to represent relations between data, and necessary linear algebraic operations on matrices. Approximately representing matrices by decompositions (Singular Value Decomposition and Principal Components Analysis). Application to the netflix prize.

Clustering: hierarchical, k-means, k-center - Finding clusters in data via different approaches. Choosing distance metrics. Different clustering approaches: hierarchical agglomerative clustering, k-means (Lloyd's algorithm), k-center approximations. Relative merits of each method.

Classification: Trees, NB, Support Vector Machines - Building models to classify new data instances. Decision tree approaches and Naive Bayes classifiers. The Support Vector Machines model and use of Kernels to produce separable data and non-linear classification boundaries.

Data Sharing: Privacy, Anonymization, Risks - The ethics and risks of sharing data on individuals. Technologies for anonymising data: k-anonymity, and differential privacy.

Graphs: Social Network Analysis, metrics, relational learning - Graph representations of data, with applications to social network data. Measurements of centrality and importance. Recommendations in social networks, and inference via relational learning.

## Learning outcomes

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

- Understand the principles and purposes of data analytics, and articulate the different dimensions of the area.
- Work with and manipulate a data set to extract statistics and features, coping with missing and dirty data.
- Apply basic data mining machine learning techniques to build a classifier or regression model, and predict values for new examples.

- Identify issues with scaling analytics to large data sets, and use appropriate techniques to scale up the computation.
- Appreciate the need for privacy, identify privacy risks in releasing information, and design techniques to mediate these risks.

## Indicative reading list

[Reading lists can be found in Talis](#)

## Subject specific skills

Working with data;  
abstracting and modelling;

## Transferable skills

Communication skills;  
Problem solving.

---

## Study

### Study time

Type	Required	Optional
Lectures	30 sessions of 1 hour (20%)	3 sessions of 1 hour
Practical classes	5 sessions of 1 hour (3%)	
Private study	115 hours (77%)	
Total	150 hours	

### Private study description

Private study and revision

### Costs

No further costs have been identified for this module.

---

## Assessment

You do not need to pass all assessment components to pass the module.

Students can register for this module without taking any assessment.

## Assessment group D

	<b>Weighting</b>	<b>Study time</b>	<b>Eligible for self-certification</b>
CS430 Project	40%		No
CS430 Project: report and code			
Centrally-timetabled examination (On-campus)	60%		No
CS430 examination			

---

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

## Assessment group R3

	<b>Weighting</b>	<b>Study time</b>	<b>Eligible for self-certification</b>
In-person Examination - Resit	100%		No
CS430 resit exam			

---

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

## Feedback on assessment

Written feedback and mark breakdown for assignments based on the created rubric.

[Past exam papers for CS430](#)

---

## Availability

### Courses

This module is Optional for:

- Year 5 of UCSA-G504 MEng Computer Science (with intercalated year)
- UCSA-G503 Undergraduate Computer Science MEng
  - Year 4 of G503 Computer Science MEng

- Year 4 of G503 Computer Science MEng
- UCSA-G4G3 Undergraduate Discrete Mathematics
  - Year 4 of G4G1 Discrete Mathematics
  - Year 4 of G4G3 Discrete Mathematics
- Year 5 of UCSA-G4G4 Undergraduate Discrete Mathematics (with Intercalated Year)