COMP229: Introduction to Data Science
Lecture 1: module overview and expectations

Vitaliy Kurlin, vitaliy.kurlin@liverpool.ac.uk
Autumn 2018, Computer Science department
University of Liverpool, United Kingdom

The scanner for registrations will be from week 2.

Background and contact details

PhD 2003 Maths, Moscow State University.

Taught large classes in the UK since 2007.

Now: Senior Lecturer in Computer Science.

Offices: George Holt building 201D in CS,
office 36 in the Materials Innovation Factory.

Personal webpage: http://kurlin.org

Best contact: email vitaliy.kurlin@liverpool.ac.uk.

For all admin, e-mail csstudy@liverpool.ac.uk.

What is important for COMP229?

The keyword is *Science*, which is "a systematic enterprise that builds and organises knowledge in the form of *testable explanations*" [Wikipedia].

Hence COMP229 will involve many *rigorous definitions* and *logical proofs* very similarly to COMP109 and COMP116 (not pre-requisites).

Bad news: COMP229 can be your hardest module. **Good news**: it is optional, so you can choose another module within the first 2 weeks.

What is Data Science?

Definition: Data Science is an interdisciplinary field that uses *scientific methods* and algorithms to extract knowledge from data [Wikipedia].

Other less strict names for Data Science:
pattern recognition (old), data mining (new).

My view: Data Science is within an intersection of Machine Learning (or AI) and Statistics.

Data Science in COMP229 is viewed as a part of unsupervised learning: predict without any help.

Yann LeCun's classification

Types of machine learning

Yann Lecun's Black Forest cake

- **"Pure" Reinforcement Learning** (cherry)
  - The machine predicts a scalar reward given once in a while.
  - **A few bits for some samples**
- **Supervised Learning** (icing)
  - The machine predicts a category or a few numbers for each input
  - Predicting human-supplied data
  - **10→10,000 bits per sample**
- **Unsupervised/Predictive Learning** (cake)
  - The machine predicts any part of its input for any observed part.
  - Predicts future frames in videos
  - **Millions of bits per sample**

- **(Yes, I know, this picture is slightly offensive to RL folks. But I'll make it up)**

Brief syllabus of COMP229

Metric spaces: clouds, distances, invariants

Statistical learning: regression, hypotheses

Dimensionality reduction: linear algebra

Topological graph analysis: classifications

Graph visualisation: planarity of graphs

Clustering methods: single-edge, k-means

Computational geometry: shapes of clouds

Learning outcomes

At the end of the module you should know how to

- describe modern problems and tools in data clustering and dimensionality reduction,
- formulate a real data problem in a rigorous form and suggest potential solutions,
- choose the most suitable approach or algorithmic method for given real-life data,
- visualise high-dimensional data and extract hidden non-linear patterns from the data.

Lecture times and rooms

Timetable: 10 weeks, 3 lectures per week.

Lectures: 30 hours, your extra work: 120 hours.

Wednesday at 9.00-9.50 in 502-LT2.

Wednesday at 12.00-12.50 in ELEC-ELT.

Friday at 10.00-10.50 in 502-LT3.

The lecture on Friday 16th November is shifted to Friday 7th December (the same time and room).

To test your skills, compute without a calculator:
$51 \times 24 - 49 \times 93 + 27 \times 51 + 44 \times 49 = ?$

Resources on VITAL

Slides will be regularly updated. If you notice typos, please e-mail me. The discussion board is for asking constructive questions (anonymously).

The COMP229 wiki page is for sharing your contributions, e.g. extra examples on the content of the slides or your links to helpful resources.

Exemplar behaviour and good non-anonymous contributions can be rewarded by references.

What is exemplar behaviour?

- Coming to lectures in time.
- Asking constructive questions:
  in lectures, by e-mail or on VITAL.
- Contributing to the COMP229 wiki.
- Focusing on understanding, not marks.

A typical unacceptable question about the exam: is slide 8 of lecture 4 needed for the exam?

Answer: all content of the slides can be in the exam, with modified numbers and similar logic.

Textbooks on Data Science

Textbooks aren't needed, listed below as extras (finding them online is a good exercise for you, I couldn't add them to the reading list, sorry).

- David Lane, Introduction to Statistics
- Sinan Ozdemir, Principles of Data Science
- Glenberg, Andrzejewski, Learning from Data: An Introduction to Statistical Reasoning

If you find anything better, please e-mail me or add to the COMP229 wiki page on VITAL.

Exam is the only assessment

COMP229 is a **new** non-programming module.

Hence there is no point to ask about past exams.

The exam will be based on the slides. All typical exam questions will be covered in the lectures.

Your mark for the module = 100% exam consisting of 50% multiple choice, 50% written questions.

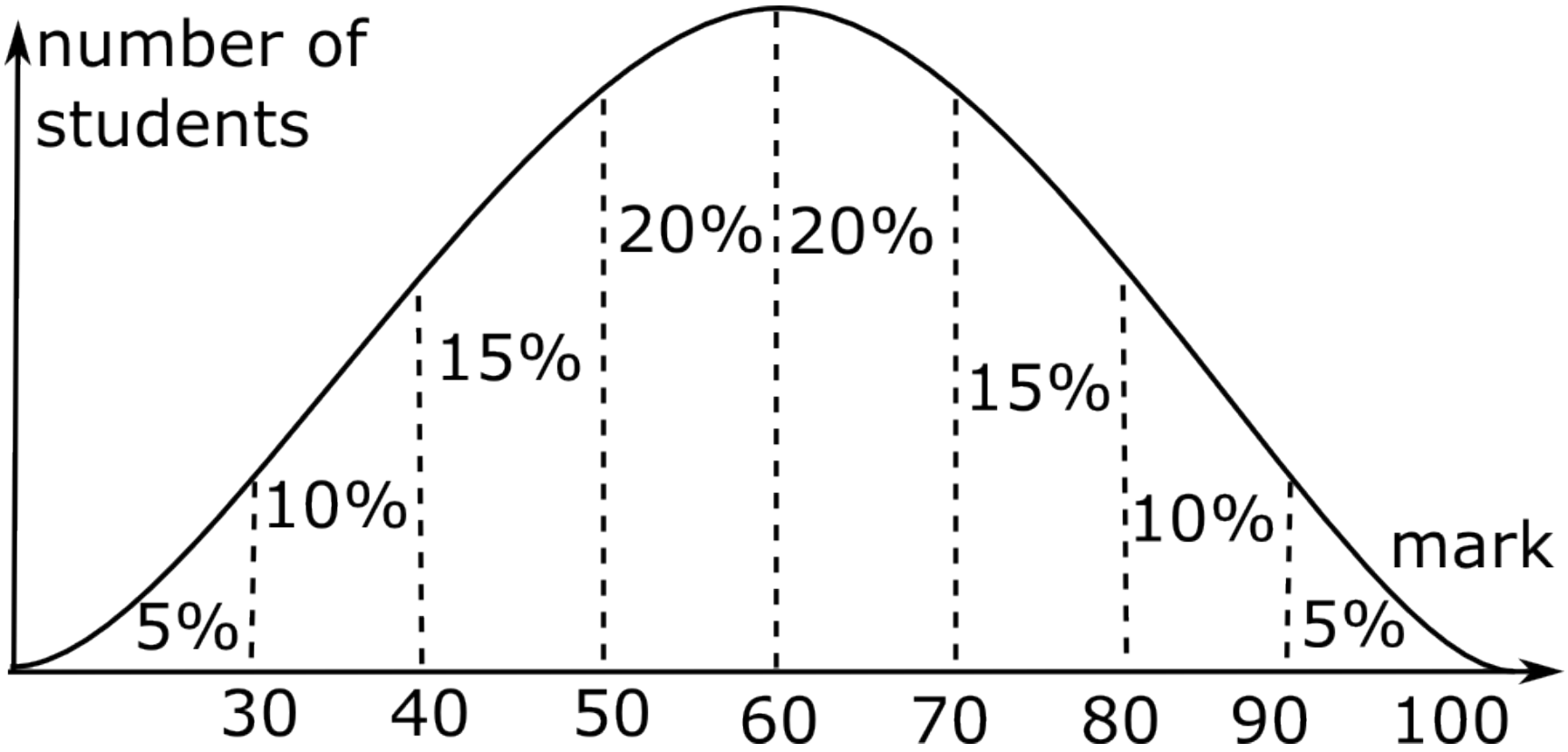
Time: 2.5 hours, all questions will be marked.

The exam will contain problems that need proofs.

Calculators aren't allowed to avoid waste of time.

A distribution of exam marks

The pass mark is 40, first class marks are 70+.



The expected average mark is within [55,65], the above distribution is only one potential example.

Instructions for the exam

INSTRUCTIONS TO CANDIDATES

HANDWRITING: If the examiner cannot read your handwriting you may be awarded a mark of zero.

You must not commence writing until instructed to do so by an invigilator.

1. Fill in the above lines before beginning to answer the questions.
2. Where it is practicable to do so, write on BOTH sides of the page. Where rough work is necessary, use the left-hand page for this purpose and cross out before handing in the book. *No part of this book may be torn off.*
3. Begin each answer on a new page and write the number of the question in the margin. If an answer requires more than one page, show the number of the question on each page.
4. If two or more books are used, fasten them together. Fasteners may be obtained from the Invigilators.
5. Fold over and seal the gummed edges to conceal your name.
6. If during the exam you have any questions or feel ill please inform the invigilator.

No books or papers may be taken to an examination desk. Your mobile phone and any other electronic data storage devices (including smartwatches) must be switched off and placed in a clear bag under your seat for the duration of the exam. Examination writing books, whether used or unused, must not be taken out of the examination room.

CANDIDATES should enter below in the left hand column the numbers of the questions answered in the order in which they have been attempted.

Question Number	For Examiners use only	

Question Number	For Examiners use only	

TOTAL MARK

What's important for the exam?

Apart from the clear handwriting, correct spelling and good English grammar, the main advice is to *seriously* consider your studies and exams.

Coming late to lectures isn't serious. In this case even if you get 100%, don't expect a reference.

The exam is designed for 2.5 hours. At first glance a marker will see the time when you've left the exam: if you leave in 1.5 hours instead of 2.5, what's your expected max mark instead of 100?

Expectations and feedback

During the lecture my students are quiet. You can ask questions, also before or after any lecture.

For anonymous questions and feedback:

- e-mail vitaliy.kurlin@liverpool.ac.uk or
- put in COMP229 folder after any lecture
- post at the discussion board on VITAL

The bottom line required by the university:
reply to student e-mails within 3 working days.

In the first 10 weeks I'll try to reply faster within few hours, the rest depends on your feedback.

How to succeed in the module

Non-native English speakers can spend at least 1 hour per day on improving English skills: online, on paper, talking to people in university clubs, ...

Subscribe to e-mail alerts about new posts.

Post anonymously at the discussion board on VITAL by ticking a box below your message.

Edit the COMP229 wiki page, e.g. by giving links to external resources. Your excellent contributions can be highlighted in references. Your career will depend on your connections, not on exam marks.

Another professor's quote

"A university-level module is not a series of bite-sized chunks of easily digestible pap presented with the sole aim of being prepared for an exam. Disquieting it may seem, study of a specialist topic to degree level at an historic university is complicated."

The extra advice (that I would really appreciate 20 years ago): your future career will largely depend on your *emotional intelligence*, e.g. the ability to communicate well with others, not on your marks.

Your questions and the quiz

Revisions will be at the end of every lecture.

To benefit from the lecture, now you could

- ask or submit your anonymous questions to the COMP229 folder after the lecture;
- write down your summary in 2-3 phrases, e.g. list key concepts you have learned.

Question. Within what period can you switch to another module and have you computed this?

$51 \times 24 - 49 \times 93 + 27 \times 51 + 44 \times 49 = ?$

Answer to the quiz and summary

Answer: you can switch within the first 2 weeks.

- All the resources of COMP229 are on VITAL.
- How to learn better: teach your classmates (explain concepts, test your understanding).
- COMP229 exam requires logical proofs.
- Practice your handwriting and English.

$$51 \times 24 - 49 \times 93 + 27 \times 51 + 44 \times 49 =$$
$$51 \times (24 + 27) - 49 \times (93 - 44) = 51^2 - 49^2 =$$
$$(51 - 49) \times (51 + 49) = 2 \times 100 = 200.$$