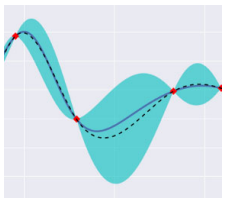
UNIVERSITY OF CAMBRIDGE
Computer Laboratory

# R244: Large-Scale Data Processing and Optimisation

## Course Guide



*Eiko Yoneki*

*Systems Research Group*
*University of Cambridge Computer Laboratory*

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

## R244 Course Objectives

- Understand key concepts of dataflow programming for scalable data processing
- Understand how to build distributed systems in data driven approach
- Understand a large and complex parameter space in computer system's optimisation and applicability of Machine Learning approach

- Research skills
  - Establish basic research domain knowledge in large data processing and Optimisation with ML
  - Obtain your view of research area for thinking forward
  - NOT to learn ML tools for ML applications

2

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

## Course Structure

https://www.cl.cam.ac.uk/~ey204/teaching/ACS/R244_2023_2024

Session 1: Introduction

Session 2: Data Flow Programming: Map/Reduce to TensorFlow to ML

Session 3: Large-scale Graph Data Processing

Session 4: Hands-on Tutorial: Distributed Data Flow Programing

Session 5: Probabilistic Programming + BO : Guest lecture (Brooks Paige)

Session 6: Optimisation in ML Compiler

Session 7: Optimisation of Computer Systems (HW design, DB…)

Session 8: Project Study Presentation (2022.11.29 @10:00)

3

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

## Reading Club

- Reading Club (not Lecture Class!)
  - 4~5 Paper review presentations and discussion per session (10-20 minutes presentation + discussion)
  - Each of you will present ~2 reviews during the course
    - Presented slides need to be emailed to me on the following day
  - *Review_Log*: minimum 1 log per session
    - Email me by noon Tuesday every week
    - Template of review log on the webpage
    - Prepare questions
  - Active participation to review discussion!

4

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

# *Review_Log*

**Paper Review Log: Session x**

**Name and (crsid):**

**Paper Title and Authors**

**1. Paper Summary (<100 words)**
**Describe a brief summary (extract essentials)**

**2. Punch-line of the Paper (<200 words):**
**What is the significant contribution?**
**What is the difference from the existing work?**

**3. Any major criticism to the authors (<150 words)**
**Any criticism and suggestions to the authors?**

5

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

# *Course Work: Reports 1&2*

- Review report on full length of paper (<1800 words)
  - Pick a paper from listed papers in R244
  - Describe the contribution of paper in depth with criticism
  - Crystallise the significant novelty in contrast to the other related work
  - Suggestion for future work

  Submission Deadline: 2023.11.10 12:00

- Survey report on sub-topic in data centric networking (<2000 words)
  - Pick up ~5 papers as core papers in your survey scope
  - Read them and expand your reading through related works
  - Comprehend your view and finish as your survey paper

  Submission Deadline: 2023.12.08 12:00

6

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

## *Study of Open Source Project*

- Pick an open source project from R244 scope
- Open Source project normally comes with new proposal of system/networking architecture
- Understand the prototype of proposed architecture, algorithms, and systems through running an actual prototype
- Any additional work
  - Writing applications
  - Extending prototype to another platform
  - Benchmarking using online large dataset
- Some projects are rather large and may require extensive environment and time; make sure you are able to complete this assignment



UNIVERSITY OF CAMBRIDGE
Computer Laboratory

## *Course Work: Reports 3*

- Report on project study and exploration of a prototype (<2500 words)
  - Project selection by November 10, 2023 (16:00)
    - Title and brief description (>150 words) by email
  - Project presentation on November 29, 2023
  - Final report on the project study by January 16, 2024

Try to finish by the end of 2023!

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

# Plan Reading Papers for Reports

- Through October and November, pick/read papers from all sessions with your interests.
- Scope in the survey topic towards the end of November for writing up a survey report by December 8.
- Think through potential open-source projects in November, present initial study/plan on November 29. Actual implementation/experiment work possibly after November 29.


9

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

# Candidates of Open Source Project

http://www.cl.cam.ac.uk/~ey204/teaching/ACS/R244_2023_2024/opensource_projects.html

- List is not exhausted and discuss with me if you find more interesting one for you
- Expectation of workload on open source project study is about intensive 3-7 full days work except writing up report
- One approach: pick one in the session topic, which you are interested in along your survey report

10

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

## *Important Dates*

- November 10 (Friday) 16:00
  - Mini Project selection
- November 10 (Friday) 12:00
  - Review report
- December 8 (Friday) 12:00
  - Survey report
- January 16, 2024 (Tuesday)
  Try to finish by the end of 2023!
  - Open-source project study report

11

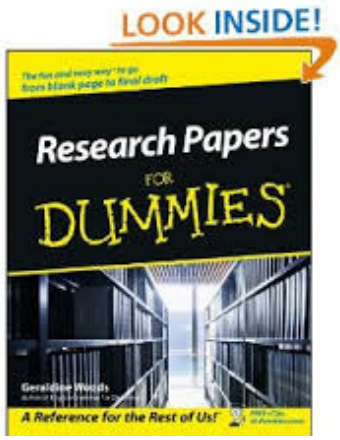
UNIVERSITY OF CAMBRIDGE
Computer Laboratory

## *Assessment*

- The final grade for the course will be provided as a letter grade or percentage and the assessment will consist of two parts:
- 25%: for a reading club (presentation, participation, tutorial session exercise and *review_log – no mark)*:
  - 10%: Presentation
  - 15%: Participation
- 75%: for the three reports: with marks
  - 15%: Intensive review report
  - 25%: Survey report
  - 35%: Project study

12

## How to Read a Paper?






UNIVERSITY OF CAMBRIDGE
Computer Laboratory

## How to Read a Paper?

- Scope of R244 is wide
- ...includes distributed systems, OS, networking, programming language, database, ML…
- Type of papers
  - Building a real system
  - Proposing algorithm/logic on architecture design
  - Optimising computer systems
  - New idea

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

## *Critical Thinking*

- Reading research paper is not like reading a textbook
- Most importantly the paper may not show the *truth*
  - there is no right and wrong, just good and bad
  - There are inherently subjective qualities…but you can't get away with just your opinion: must argue
- Critical thinking is the skill of marrying subjective and objective judgment of a piece of work



UNIVERSITY OF CAMBRIDGE
Computer Laboratory

## *First Let's Argue for…*

- What is the problem?
- What is important?
- Why isn't it solved in previous work?
  e.g. Why graph specific parallel processing? MapReduce is not good enough?
- What is the approach?
  e.g. Graph specific MapReduce
- Why is this novel/innovative?
  e.g. Iterative operation for graph parallel

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

## *And Now against…*

- Problem is overstated (or oversold)
- Problem does not exist
- Approach is broken
  - It does not work for all the algorithms…
  - It does not scale…
- Solution is insufficient
  - Only works when data is in memory…
- Evaluation is unfair/biased
  - Use HPC for experiment



UNIVERSITY OF CAMBRIDGE
Computer Laboratory

## *So Which is RIGHT Answer?*

- There isn't one!
  - Most of arguments are mostly correct…
- Your judge on what is valuable on topic
- In this course, we'll be reviewing a selection of 20+ papers (4-5 per week)
  - All of these papers were peer-reviewed and published
  - However you can pick your opinion on papers!

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

## *Reviewing Tips & Tricks*

- Identify a core/major idea of the topic
- Read related work and/or background section and read key other papers on the topic
- Capture the author's claim of contribution in *introduction* section and judge if it is delivered
- Understand the methodology that demonstrates paper's approach
- Capture what authors evaluate and judge if that is a good way to evaluate the proposed idea
- For theory/algorithm paper, capture what it produces as a result (rather than how)



UNIVERSITY OF CAMBRIDGE
Computer Laboratory

## *Key in Review Comments*

- What do YOU think?
  - Where you finally get to explain your opinion!
  - You should aim to give *a judgement* on the work
  - Your judgement should be backed by your argument
- Questions for the authors

S. Hand'10

## *How to Review a Paper Aid…*

- S. Keshav: How to Read a Paper, ACM SIGCOMM Computer Communication Review 83 Volume 37, Number 3, July 2007.
- T. Roscoe: Writing Reviews for Systems Conferences, 2007.
- Simon Peyton-Jones: How to write a great paper and give a great talk about it, Microsoft Research Cambridge.
- David A. Patterson: How to Have a Bad Career in Research/Academia, 2001.

See course web page for the paper links.





## *Structure of Presentation*

- Cover 3 things in your presentation

1. Background/context
   - What motivated the authors?
   - What else was going on in the research community?
   - How have things changed since?
2. What is problem to be tackled?
   - What is the problem they tried to solve?
   - What are the key ideas?
   - What did the authors actually do?
   - What were the results?
3. Your opinion of the paper
   - What you agree and what you disagree?
   - What is the strength and weakness of their approach?
   - What are the key takeaway?
   - What was the impact (possible impact)?

S. Hand'10

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

## *Preparing…*

- Not too much basics: remember, others would have read the paper
  - Brief overview
  - Do not make exact repeat of the paper

- Aim: generate discussion – spit your straight opinion about the paper to stir the discussion
  - Explore the arguments they make and the conclusions they draw. What is your opinion on it?
  - When you argue, state clearly the point of argument



UNIVERSITY OF CAMBRIDGE
Computer Laboratory

## *Presenting…*

- Practice beforehand to ensure length of your presentation

- Getting nervous is normal!
  - We are in the same boat and we help each other to understand the paper
  - Presentation is a tool to provide a discussion forum
- Try not to get defensive or angry at questions
  - It is not your paper !

## *Listening Presentation…*

- You need to get involved

- Ask questions from your review – bring your *review_log* copy
- Always be respectful of the speaker





## *How to Write Reviews (Report 1)*

- Paper Summary
  - Provide a brief summary of the paper
  - At this stage you should try to be objective
- Problem
  - What is the problem? Why is it important? Why is previous work insufficient?
- Solution or Approach
  - What is their approach?
  - How does it solve the problem?
  - How is the solution unique and/or innovative?
  - What are the details?
- Evaluation is unfair/biased
  - How do they evaluate their solution?
  - What questions do they answer?
  - What are the strength/weakness of the system and evaluation itself?

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

## *How to write Survey paper (Report 2)*

- Demonstrate a summary of recent research results in a novel way that integrates and adds understanding to work in the research area
- Must expose relevant details associated, but it is important to keep a consistent level of details and to avoid simply listing the different works
- For example:
  - Define the scope of your survey
  - Classify and organize the trend
  - Critical evaluation of approaches (pros/cons)
  - Add your analysis or explanation (e.g. table, figure)
  - Add reference and pointer to further in-depth information


27

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

## *Summary*

- R244 course web page:

  http://www.cl.cam.ac.uk/~ey204/teaching/ACS/R244_2023_2024

  Email: eiko.yoneki@cl.cam.ac.uk

- Slides of presentation, forms, other information will be on the web

- Please email me your presentation slides after the session

28