CHARLIE DICKENS

dickens.c07@gmail.com ◇ c-dickens.github.io

RESEARCH INTERESTS

- The design & implementation of small data summaries which allow approximate analysis of large data.
- Scalable machine learning algorithms through the application of data sketches or other methods.
- Data analysis and machine learning over large data streams, and to what extent this may preserve privacy.

EMPLOYMENT

Yahoo! *June 2021 - present*
Research Scientist *London*

- Co-developed a compression scheme to reduce the size of data summaries used for reporting. **Outcome:** Compressed summaries are $15-25\%$ smaller and are only $\approx 5\%$ larger than theoretically optimal compression. Published internally and productionised.
- Discovered and rectified a deficiency in our most-popular reporting tool. **Outcome:** most expensive step in generating distinct count reports made 20% faster. Code alteration pushed to production.
- Co-designed diffferentially private data summaries for cardinality estimation. **Outcome:** These algorithms are faster by a factor of k than prior state-of-the-art (for a size k-bit data structure) and use a factor of at least 8 less noise to achieve the same privacy guarantee. Published at NeurIPS.
- Built a prototype server for private distributed aggregation. **Outcome:** delivered to Research VP and awaiting approval for engineering resource allocation to build into a product.
- Showed that brute-force engagement analytics incur a $\mathbf{10^6}\times$ **larger** space-usage than using data sketches; advocated against migrating these analyses to brute-force cloud providers to reduce data analysis costs.
- Managed cross-continental working relationships as the entirety of my team is based in California. Achieved **consistently high performance reviews** despite being the only team member based in the UK.

Verizon Media *June 2020 - November 2020*
Research Intern *Remote (UK)*

- Studied matrix computation using small space summaries for high-dimensional ridge regression.
- Proposed method was faster, requiring only one expensive summarisation step (versus logarithmically many from prior work), and converged at a rate **3× faster** than competing methods.

Amazon *June 2019 - December 2019*
Applied Science Intern *Cambridge*

- Designed an interpretable, user-friendly anomaly detector using random forests. **Outcome:** Method is competitive with prior state-of-the-art but enjoys the benefit of operating over multidimensional streaming data. Delivered a report outlining how to integrate the algorithm seamlessly with the existing anomaly detector.
- **Initiated and led** a weekly reading group/tutorials on "Data sketching in machine learning" for lab scientists to advocate for the adoption of scalable machine learning methods.

The Alan Turing Institute *May 2017*
Group Facilitator for Defence Science and Technology Laboratory (DSTL)

- **Managed** a team of diverse backgrounds to develop tools for detecting anomalous network behaviour.
- Delivered a report on the project's findings to **DSTL stakeholders** outlining findings and future directions.

EXTRACURRICULAR

Apache Software Foundation DataSketches Project Management Committee *Oct. 2021-present*

- Project committer since October 2021. **Promoted** to PMC in July 2023.

- Developed **project management skills** such as identifying and improving upon deficiencies in current product and collaborating on future roadmap.
- Focus on PMC has been external advocacy (outcome: BigDataLDN 2023 presenter) and website/documentation redevelopment (outcome: revamped website deployment scheduled Autmumn/Winter 2023).

West Midlands Velodrome Campaign *2018-present*

- Led campaign to publish evidence explaining why a velodrome was not build for the Birmingham 2022 Commonwealth Games. **Started a petition that amassed ≈ 10k signatures** and delivered to Birmingham City Council Leader;
- **Outcome:** Birmingham City Council and West Midlands Combined Authority allocated funding for a business case analysis (expected completion in winter 2023).

Mentoring

- Currently volunteer as an "industrial supervisor" for a UCL MSc Machine Learning student. Aim is to develop a mergeable and scalable language model with small space overhead.
- **Successfully mentored** two students to highly competitive MSc places in Machine Learning at the University of Edinburgh and University of Illinois at Urbana-Champaign.
- **Volunteered** as a "maths mentor" for GCSE students in group and 1:1 settings during my MSci and PhD.

EDUCATION

University of Warwick *October 2016 - 2021*
PhD Computer Science *Supervisor: Prof. Graham Cormode*

- Thesis: *"On the Efficiency of Finding & Using Tabular Data Summaries: Scalability, Accuracy, and Hardness."*
- Awarded the *Faculty of Science, Engineering, and Medicine Thesis Prize in Computer Science*
- Visiting Graduate Student for the Foundations of Data Science program at the Simons Institute for the Theory of Computing, Berkeley, California, August-October 2018.
- Enrichment Year Student at Alan Turing Institute for Data Science & AI, London, 09/2017-08/2018.

University of Birmingham *2012 - 2016*
MSci Mathematics Class I

- Thesis title: *"Probabilistic and Algorithmic Aspects of Spectral Graph Theory"*

AWARDS

- **Faculty of Science, Engineering, & Medicine Thesis Prize in Computer Science (June 2022).** Awarded for best thesis in Computer Science
- **Turing Enrichment Award (October 2017 - October 2018).** Scholarship (one of seventeen awarded nationally) to study at the UK's National Institute for Data Science and Artificial Intelligence
- **British Colloquium for Theoretical Computer Science Speaker Bursary (March 2018)** One of ten bursaries awarded for PhD speakers
- **Warwick Postgraduate Colloquium in Computer Science Best Presentation (June 2017)** Awarded for best presentation in *Foundations* track
- **London Mathematical Society Research Bursary (Summer 2016)** One of twenty bursaries awarded nationally for undergraduate researchers
- **Catenian Association Prize for Public A Level Performance (Summer 2012)** Upper School Prize in Mathematics and Electronics

TECHNICAL SKILLS & PROGRAMMING

- Proficient in Python (including NumPy, Matplotlib, Pandas, Scikit-learn etc.) MATLAB.
- Working knowledge of C++; capable of building C++ bindings to Python code.

RESEARCH PUBLICATIONS

Order-Invariant Cardinality Estimators Are Differentially Private. C. Dickens, J. Thaler, D Ting, *NeurIPS 2022*

Subspace Exploration: Bounds on Projected Frequency Estimation. G. Cormode, C. Dickens, D.P Woodruff, *ACM Principles of Database Systems (PODS) 2021 (One of five PODS papers invited to SIGMOD)*

Iterative Hessian Sketch in Input Sparsity Time. G. Cormode, C. Dickens, *NeurIPS 2019 Workshop on Beyond First Order Methods in Machine Learning*

Leveraging Well-Conditioned Bases: Streaming and Distributed Summaries in Minkowski p-Norms. G. Cormode, C. Dickens, D.P. Woodruff, *International Conference on Machine Learning 2018 (Invited for long talk)*

Frequent Directions as a Tool for Learning with Small Space. C. Dickens, *Technical Report (Verizon Media internship work)*

Interpretable Anomaly Detection with Mondrian Pólya Forests on Data Streams. C. Dickens, E. Meissner, P. G. Moreno, T. Diethe, *Technical Report (Amazon internship work)*