EDWARD O. CANNON

+44797 · 409 · 3957 ⋄ eoc210@googlemail.com

PROFILE

Principal data scientist with over 10 years experience. Leading teams of up to 10 people, to develop methodologies for data analysis to provide insight in a variety of fields ranging from retail analytics, failure prediction, fluid optimization pipelines, to pharmaceutical industry and healthcare.

TECHNICAL SKILLS

Big Data Technologies	Hadoop, Hive, HBase, Pig, AWS, Azure, Mongo DB, Apache Spark
Computer Languages	Python, R, Java, C++, LaTeX, XML, XPATH, SQL
Operating Systems	Windows, Mac OS X, UNIX, Linux
Software Packages	MOE, Pipeline Pilot, Spotfire, Sybyl, Tableau, Brandwatch, Clarabridge
Machine Learning	Weka, R, Python
Scientific Knowledge	Chemistry, Chemoinformatics, Bioinformatics, Statistics
Tools	MS Office Suite, Git, SVN, CMake, Maven, Visual Studio, Eclipse Atlassian Stash & FishEye, TeamCity, CircleCI, Qt, Swig, Pycharm, Jupyter Notebooks

PROFESSIONAL EXPERIENCE

QuantumBlack / McKinsey May 2018 - June 2019
Principal (Jr) Data Scientist *London, UK*

· Established and lead best practices. Developed risk register for data science engagements. Active contributor to internal assets. Established and lead the chemogenomics group.

· **Patient Enrolment Site Selection** - *Lead data scientist, managed team of 3 data scientists.*
· Objective: Identification of optimum trial sites world wide to minimise cost, whilst maximising patient enrolment in the shortest possible time.
· Developed a RandomForest approach to predict trial complexity and enrolment rate at site level.
· Used mixed Integer Programming (Google OR-Tools) to optimise site selection.
· Constructed cross-validation pipelines to analyse model performance, across multiple metrics.
· Tools: Python, Git, JIRA, Confluence, AWS hosted environment.

· **Oil Optimisation** - *Lead data scientist, managed team of 5 data scientists.*
· Objective: Maximisation of raw oil flow rate to provide 1% lift in performance ($100M profit).
· Analysed GAP software simulation data, developed a custom API, developed optimisers using genetic algorithm (DEAP) and a Bayesian (PYGPGO) framework.
· Tools: Python, Git, JIRA, Confluence, Azure hosted environment.

Capgemini Oct 2015 - April 2018
Data Scientist - Senior Applications Consultant *London, UK*

· Established and implemented gold standard rules for project development and defined best coding practices. Actively recruited at all stages for Capgemini. Managed, reviewed and set objectives/targets for less senior data scientists. Created and lead a fortnightly python club consisting of 14 data scientists.

- **Predictive Maintenance** - *Lead data scientist, managed team of 6 data scientists.*
- Objective: Building of a Long-Short-Term Memory recurrent neural networks for *failure prediction* of combined heat & power engines using *deep learning* (keras, tensorflow).
- Created custom *loss and activation* functions based on the Weibull distribution time to failure prediction.
- Engaged, presented, and interacted with national and international senior stakeholders, growing the account.
- Tools: Python & MS Azure. An agile scrum based methodology was adopted.

- **Consumer Marketing Goods** - *Lead Data Scientist, managed team of 10 data scientists.*
- Objective: Development of analytic methods for brand analysis.
- Identified brands' return on investment by media channel using Vector AutoRegression Models, generating client savings of ∼£30k a year.
- Developed a custom service to validate influencers across social media channels for brand managers, using Dataiku, saving client over 1000 man-hours.
- Generated client savings of ∼£50K a year by creating a custom weather index service, and identifying the impact of weather on sales volume.
- Built ontologies and integrated multiple data sources to identify brand issues, and generate recommendations.
- Conducted market research by analysing brand trends in social media channels, based on demographic, weather, and EPOS data. Identified key drivers in social media spikes. Lead and managed the migration of code base to Atlassian Stash.
- Tools: Python, R, Dataiku, Clarabridge, Brandwatch, Hadoop, Pyspark.

- **Retail Analytics** - *Lead data scientist, managed 1 junior data scientist.*
- Objective: Identification of key attributes driving store performance.
- Developed regression models to predict store total sales per sqft, created a store performance ranking.
- Analysed store performance using catchement, geographical and store type data. Used demographic data to identify which mosaic segments correlate with sales at client stores. Identified customer behaviour including: in store, online and click & collect, the amount customers spend and the frequency of their purchases.
- Tools: Python, Java, R and postgres SQL. An agile scrum based methodology was adopted.

Starcount May 2014 - Sep 2015
Data Scientist Consultant - Head of Data Development *London, UK*

- Supervised and mentored a team of 5-10 data scientists and engineers and managed external contract developers on projects worth > $50K/year.
- Lead data extraction and pre-processing from social media APIs.
- Developed cronned pipelines and used Hive & Pig to query data and obtain high level statistics.
- Developed Python applications for data segmentation and clustering.
- Managed, developed, and optimised the data warehouse, in Hadoop, AWS cluster and MongoDB databases.

OpenEye Scientific Software Inc. Oct 2010 - Mar 2014
Life Science Consultant - Scientific Software Developer *Boston, MA, USA*

- **Lexichem Toolkit** - *Lead developer & data scientist.*
- Developed a computational framework for automatic conversion of chemical names into structures and vice versa in more than 10 languages (e.g. English, French, Chinese, Japanese, Spanish, etc). Annual sales > $1million.
- Developed, designed, maintained, bug fixed, tested and released software across multiple platforms (Windows, Mac OS X, Linux) in a timely manner (4 releases per year).
- Developed a green field application for Lexichem using Qt in an agile development framework. The application offered the user the possibility to automatically extract and convert chemical names into structures and vice-versa from locally uploaded documents and online data.
- Published original research papers and presented at national and international conferences.

Cambridge University Jul 2008 - Jun 2010
Postdoctoral Research Associate - Polymer Informatics Data Scientist *Cambridge, UK*

- Developed a green field Ajax application to calculate physico-chemical features of polymers. The app was developed using the Google Web Toolkit, Restlet and an Apache Derby database.
- Developed polymer ontologies and semantic web services for the Dutch Polymer Institute.
- Published original research papers and presented at international conferences and meetings.

Novartis Pharmaceuticals Jun 2005 - Aug 2005
Research Placement *Horsham, UK*

- Three month placement using Pipeline Pilot, R and Spotfire.

Lubrizol Ltd. Jul 2003 - Jul 2004
Research Placement *Manchester, UK*

- One year research placement synthesising block copolymers by atom transfer radical polymerisation for use as pigment dispersants in automotive coats.

EDUCATION

University of Cambridge Jun 2008
PhD - Chemoinformatics *Cambridge, UK*
Thesis: Chemical Informatics of Banned Substances. Used machine learning and data mining techniques to classify and virtual screen chemical substances taken from the World Anti-Doping Agency's Prohibited List. Supervisor Dr John Mitchell.

University of Sheffield Oct 2005
MSc - Chemoinformatics *Sheffield, UK*
Thesis: A novel circular substructure fingerprint for virtual screening using multiple bioactive reference structures. Supervisors Dr Peter Willett and Dr Val Gillet.

University of York Jul 2004
MChem Chemistry (1st) *York, UK*

INTERESTS

Skiing, dancing, and going down the gym. Active participant to London Data Science meetups.

SELECTED PUBLICATIONS

- Cannon, E. O. (2012). New Benchmark for Chemical Nomenclature Software. *Journal of Chemical Information and Modeling* 52, 1124-1131.
- Colas, Cannon et al. (2009). A Subtype Discovery Methodology as a R Package with Use Case on Chemoinformatics Data. *IEEE-EMBC '08*.
- Cannon et al. (2007). Support Vector Inductive Logic Programming Outperforms the Naive Bayes Classifier and Inductive Logic Programming for the Classification of Bioactive Chemical Compounds. *Journal of Computer-Aided Molecular Design* 21, 269-280.
- Cannon et al. (2006). Chemoinformatics-based Classification of Prohibited Substances Employed for Doping in Sport. *Journal of Chemical Information and Modeling* 46, 2359-2380.