Serena Peruzzo

serena.peruzzo@gmail.com | www.serenaperuzzo.com | LinkedIn | GitHub

EDUCATION

M.Sc. Statistics | La Sapienza University of Rome Sep 2009 – Apr 2011
Final grade 110/110 cum laude
Final dissertation on 'Cluster Leaderships and International Development Strategies'

B.Sc. Statistics and IT for Business | La Sapienza University of Rome Sep 2005 – May 2009
Final grade 110/110 cum laude
Final dissertation on 'Analytical models for Purchase Parity Power'

PROFESSIONAL EXPERIENCE

Lead Data Scientist | Eggbun Education, London/Remote Jun 2017 – Present
Distributed startup building an Intelligent Tutoring System for learning Asian languages

- Research and define an AI roadmap for Eggbun
- Research, build and implement ML algorithms to add intelligence to the virtual tutor and personalise the learning to the users
- Educate the rest of the team on AB testing and support them with experiment design and analysis
- Support the marketing team building metrics and tools for analysis of customer's behaviour

Lead Data Scientist | BulbThings, London/Remote Oct 2016 – May 2017
A London startup providing an enterprise app for physical asset management leveraging IoT technologies.

- Work closely with the CPO and CTO to define the data analytics features and roadmap for BulbThings
- Research and identify opportunities for improving quality and intelligence of BulbThings analytics
- Support the marketing team building metrics and tools for analysis of customer's behaviour

Researcher | Technical University of Eindhoven, Computer Science Dept. Dec 2015 – Sep 2016

- Research on data-driven predictive maintenance with focus on neural networks and unsupervised machine learning for extracting features from multivariate time series.

Data Scientist | Portent.io, London/Remote June 2015 – Nov 2015
A London startup providing predictive analytics for the entertainment industry

- Improve predictive model using a large movie database and complementary datasets
- Prototype features and models
- Identify, research, and analyse, new data sources to improve model accuracy
- Automated a system for anomaly detection in social-media data streams

Participant | Recurse Center Feb 2015 – May 2015
A self-directed educational retreat for programmers

Analysis of Shakespearean Plays: Two stage unsupervised approach to document analysis and classification combining topic modelling and clustering analysis. Presented the comparison of Natural Language Processing results vs. traditional classification results for comedies and tragedies.

RC's social interactions: An analysis of social interactions at Recurse Center mining the history of the internal chat tool. Direct interactions were summarised in a graph form and represented in D3.js.

Statistical Analyst | Symbolix Pty Ltd, Melbourne Mar 2012 – Dec 2014
A Melbourne based consultancy that helps businesses access, analyse and act on data

- Designed and implemented methodologies for data handling and management
- Applied statistical methods to data and interpret the output
- Built, maintained, and documented scripts (e.g. in R/SQL) for repeatable analysis
- Communicated analysis outputs and made recommendations through formal and informal reporting (data packs and visualizations in html/R shiny/Tableau).

Relevant projects:

SmartMeter classification, major electricity wholesaler: Designed and constructed unsupervised clustering algorithm based on a custom distance metric that uses outward smart meter data to classify and cluster customers into 3 phases plus outliers.

Queensland Regional Travel Surveys (with Ipsos/IView), Dept. Transport and Main Roads QLD: Provided a sampling methodology designed to control bias and match the new ASGS (Australian Statistical Geographical Standard). Post data collection provided statistical audits, imputation of missing data (using stochastic imputation, hot-decking and custom-designed methodology), and post stratification weights.

Major component analysis, Woolnorth Wind Farm: Use data on equipment commissioning dates, and relevant failure dates, to provide failure rate analysis of electrical components.

Automated bat call noise filtering and classification, Hydro Tasmania: Explore large datasets of audio fingerprints using unsupervised Bayesian classification (AutoClass) to obtain a signal vs noise classification of the recordings and allow automation of the screening process.

SKILLS AND TOOLS

Programming: R, Python, SQL, awk, VBA
Software: R, MongoDB, mySQL Workbench , Office Suite, Tableau, Prezi
Languages: Bilingual Italian-English, Fluent French

PUBLICATIONS

- S. Peruzzo, M. Holenderski & J. J. Lukkien, *Pattern-based feature extraction for fault detection in quality relevant process control,* ETFA 2017
- *Data Cleaning a Necessary Chore,* The Recompiler, Issue Six - March 2017
- *Data Driven Literary Analysis,* Code Words, Issue Seven - October 2016
- Hull C., Stark E., Peruzzo S. & Sims C. 2013. *Avian collisions at two wind farms in Tasmania, Australia: Taxonomic and ecological characteristics of colliders versus non-colliders*, New Zealand Journal of Zoology, 40:1, 47-62.

PROFESSIONAL AFFILIATIONS, SHORT COURSES AND OTHER

Institute of Analytics Professionals Australia (IAPA) – Member
Data mining in R – *Learning with case studies* – Course member
RC Start, *mentorship program for new programmers run by Recurse Centre* – Mentor
PyData, Amsterdam – Speaker
K2 Data Science, *online data science bootcamp for working professionals* – Mentor