Darya **Vanichkina**

Data Scientist & Consultant

Australian Citizen - Marsfield, New South Wales, Australia

📱 +61 420889939 | ✉ d.vanichkina@gmail.com | ⌂ daryavanichkina.com | dvanic | in daryavanichkina

Profile

PhD-qualified data scientist with 10 years hands-on experience in big data, machine learning and statistics and team, client and end-to-end project management. I am passionate about using analytics solutions to gain actionable insights & leading for performance and individual growth while supporting business goals. I am looking to join a data-driven organisation that values excellence and performance, so that I may be continually challenged as I move from an individual contributor towards a management role.

Skills

Programming For Data Analysis: R (tidyverse, base, data.table, shiny), Python (numpy, pandas, matplotlib, seaborn, plotly, web scraping, regex, Jupyter notebooks, networkX, sqlite3), SQL, MongoDB, git & GitHub (code review), Bash, awk, HTML, CSS & LaTeX; geospatial data analysis with sf, stars & QGIS; Hugo & RMarkdown for dynamic reporting.
Machine Learning & Statistics: Descriptive & inferential statistics, supervised & unsupervised learning approaches. Hypothesis testing & experimental design. GLM. Time series. Survival Analysis. PCA, tSNE, UMAP. Clustering. Python (Scikit-learn, tensorflow, spacy, nltk) & R (tidymodels, caret, underlying packages, bioconductor).
High-Performance & Cloud Computing for Big Data: Deploying analytics on HPC clusters (PBS-Pro, SLURM) & AWS.
Software Development: Version control with git, code review, automated testing, continuous integration, Agile.
Communication: Collaborating with data owners to turn loosely-defined questions into clearly prioritised projects. Visualisation & dashboarding in PowerBI, R, Python. Developing reports and presentations. Public speaking.
Project management: Scoping, designing, running & coordinating projects using Waterfall & Agile methods (JIRA, Confluence, GitHub). Managing stakeholders, clients & analysts. Delivering outcomes within time frames & budgets. Preparing grants & project proposals.
Leadership: Managing a team with distinct personalities & backgrounds to achieve outcomes in the presence of multiple competing interests. Leading change.
Other: Using Eventbrite, REDCap, Qualtrics & other tools to organise events & administer surveys; OpenRefine for data cleaning; Advanced MS Office suite & Adobe Photoshop, Illustrator & In-Design.

Experience

University of Sydney, Sydney Informatics Hub *Sydney, Australia*
Data Scientist & Analytics Trainer *Feb 2020 - Present*

Key responsibilities:

- Provide diverse researchers with analytics support by working on 2-12 week data science projects utilising machine learning, statistics & high performance and cloud computing.
- Manage projects and stakeholder relationships where other analysts and software engineers are carrying out the technical work, to ensure scoped outcomes and time frames are met.
- Develop, deliver and manage a program of advanced data science training at the University of Sydney.
- Coach staff in best training practices and improve internal processes and procedures.
- Manage SIH partnerships with national organisations.

Key achievements:

Australian obesity corpus

- Resolved encoding issues in a corpus of 27 000+ articles about obesity from 12 Australian newspapers.
- Implemented near-duplicate detection, visualisation and removal using minhash and locality-sensitive hashing ("reimplemented Turnitin in Python") to ensure subsequent statistical analyses were not invalid due to duplicates.
- Used topic modelling to create sub-corpora for further analysis.

- Translated analytics spreadsheets used by researchers to a standalone Python script, automating and scaling one-vs-all and one-vs-one sub-corpus analysis.
- Provided comprehensive client-interpretable documentation.
- *Key tools: Python: pandas, chardet, spacy, seaborn, gensim, datasketch; Git + GitHub; Jira; fdupes*.

Online sports wagering

- Integrated 12+ million online wagering transactions from 50 000 clients of 6 major online wagering operators, co-writing the first academic publication & a report for Responsible Wagering Australia on customer protection tool use among Australian online wagerers. Engineered over 200 features for analysis.
- This revealed that mandatory vs opt-in deposit limits are likely to reduce the risk of harms, so was covered by media worldwide.
- *Key tools: R: data.table, tidyverse, Rmarkdown, lubridate, gtsummary, tidyquant, lmertest, broom; Git + GitHub*.

Geospatial API data

- Scraped 30 years of vegetation data for an Ecology client, comparing discrepancies between scrapes & identifying correct data by overlapping with BOM records.
- Used a high-performance computing cluster to download, crop and process 30 years worth of satellite imagery as input for machine learning.
- Client has subsequently engaged with SIH for a comissioned project.
- *Key tools: R: Rmarkdown, tidyverse, sf, raster, tmap, jsonlite, data.table; Git + GitHub*.

Test comparison

- Evaluated whether a specific test was comparable with an internationally used standard.
- Provided report for clients with confidential answer, as well as next actionable steps and recommendations on future statistical work.
- *Key tools: R: Rmarkdown, tidyverse; Git + GitHub*.

Project management

- Managed 5 projects executed by other staff, including coaching in best analysis methods, communicating with clients and negotiating temporary reduction in staff time due to competing analyst project loads. This included work:
- Using logistic regression and random forest models to predict whether genetic markers and demographics could predict the need for oxygen therapy among flu patients. [They couldn't - and my role was to communicate and resolve this negative result with the client].
- Developing a social network analysis of 20 000 students to model the relationship between student belonging/social network expansion and their academic performance. Pilot results showed we could use the data provided to generate social network, and the client has engaged SIH to continue this work.
- Integrating 20 years of records from the manually curated OMIA database into the international European Variant Archive (EVA), including resolving issues resulting from manual data entry. I was the only person with both software and biology experience on the team, so this project involved substantial communication mediation. Records have been standardised and will be part of the next EVA release.
- Developing software for storage and querying outcomes of mass spectrometry experiments. This software is currently undergoing user acceptance testing.
- Determining the association between cancer severity and type and sleep disorders using beta-regression on actigraphy data. Client provided funding to SIH for this analysis, and has used it as part of a publication.
- *Key tools: git & GitHub; GitHub projects; Jira; Trello.*

Dashboarding & reporting

- Developed a PowerBI dashboard of training performance for University executive & stakeholders.

Training

- Improved Python and R-based machine learning training developed for internal use for fee-for-service delivery for clients outside the university.
- Established and developed a program of short-format online training, increasing training attendance by over 200% while reducing staff-hour costs per attendee
- *Key tools: R: caret & tidymodels + Rmarkdown, xaringan, GitHub pages; Python: scikit-learn, pandas + mkdocs.*

Partnerships

- Delivered training in Python for Geoscience to members of the Petroleum Exploration Society of Australia (including staff of Santos, BHP, Beach Energy, Origin etc). *Key tools: Python: pandas, shapefile, lasio, obspy, requests, cartopy, scikit-learn, tensorflow, dask. AWS EC2 & S3 for learner support. GitHub pages.*

- Scoped out a new partnership with the `eHealth@Sydney` conference to run a Digital Health bootcamp in conjunction with the conference (delivered Feb 2021). *Key tools: Python: scipy, pyross, MedSpaCy.*

- Coordinated & delivered the Business School Data Science Summer School to PhD candidates and researchers at all career stages, including the Deputy Director of the Institute of Transport & Logistics Studies (30 attendees). *Key tools: Python & SQL.*

Recruitment

- Served on selection committee in the successful recruitment of 1 software engineer and 1 analyst.

National activities

- Provided national and international guidance on the pivot to teaching online through a series of webinars, focus groups and presentations, reaching over 200 attendees.
- Delivered invited presentations at eResearch Australasia (2020, 2021) & served on the steering committee of the ARDC Skills Summit in 2020 & 2021 (over 150 attendees each year).

University of Sydney, Sydney Informatics Hub *Sydney, Australia*
Data Science Group Lead (Secondment) *Aug 2019 - Feb 2020*

Key responsibilities:

- Lead team effectively and positively.
- Maintain working relationships and strong reputations with clients.
- Scope, manage and deliver outstanding projects and consultations that meet client needs within time and budget constraints.
- Contribute to team leadership activities such as strategy, reporting, outreach and recruitment.
- Work reflectively, sharing critical insights into how team processes can be improved.

Key achievements:

Personnel management

- Managed team of 4 data scientists & software engineers.
- Carried out end of year performance planning and development review and goal setting, coaching staff in aligning personal career objectives with unit goals.
- Supported staff member through successfully applying for a role overseas.
- Led performance review process liaising with University human resources.
- Recruited 4 data scientists and was part of unsuccessful recruitment of a software engineer.

Process improvement

- Improved hiring processes, incorporating a technical task and phone screening steps.
- Improved performance planning and development processes by drafting template KPIs that map to staff position descriptions and overall SIH aims.
- Prepared a template repository for all analytics projects, enabling documentation and client communication to be integrated with actual project code, ensuring information is retained after staff leave SIH.

Project management

- Prepared 4 project scopes which were subsequently approved.
- My staff delivered key projects around geospatial software for mapping of experimental farm data, integrating protein expression experiments, querying the NYT API for comments, and language processing for media.

University of Sydney, Sydney Informatics Hub *Sydney, Australia*
DATA SCIENTIST & ANALYTICS TRAINER *Oct 2018 - Jul 2019*
As Data Analytics Trainer, I was responsible for establishing a program of advanced data science training at the University of Sydney.

Key responsibilities:

- Identify priorities for the development of data science training at the University of Sydney.
- Develop and deliver courses. Support staff in making contributions to these courses.
- Adapt business bootcamp from the US to the Australian context, and mentor academics in developing and delivering bootcamp-style training.
- Establish processes and procedures for training management at SIH.
- Identify and build relationships with stakeholders interested in data science training across the university.

Key achievements:

- Established SIH's coordinated training program, standardising the activities, reporting and management of training among staff across SIH's four diverse teams. Prior to this, the four teams operated independently & ad-hoc, which was inefficient & made quantitating performance & reporting up impossible.
- Mentored other instructors and provided upskilling in best data science and training practices.
- Developed, organised and delivered courses in machine learning and geospatial data analysis using R and Python.
- Developed, delivered, managed and ran a 5-day Data Science Summer School for the Business school, garnering positive feedback from the Deputy Dean and a request to hold it again.

Centenary Institute of Cancer Medicine and Cell Biology *Sydney, Australia*
BIOINFORMATICS POSTDOCTORAL RESEARCH OFFICER *Oct 2015 - Oct 2018*
Key achievements:

- Developed tools and methods for intron retention analysis using long read sequencing.
- Prepared first report on data science challenges of analysing intron retention using sequencing data (published as an academic paper).
- Deployed HiC analytics pipeline on AWS
- Established analysis pipeline using Artemis HPC. Awarded Artemis Grand Challenge allocation.
- Investigated Glioblastoma Multiforme patient samples to assess whether intron retention could be used to predict cancer grade or treatment outcomes.

Institute of Molecular Bioscience, University of Queensland *Brisbane, Australia*
PHD CANDIDATE & TUTOR *Dec 2010 - Mar 2016*
Key achievements:

- Developed a novel method of identifying miRNA targets with a ~10x improvement over existing approaches.
- President of the IMB Students association, SIMBA, leading a team of 6. Supported one of my team in establishing the IMBar which continues to this day.
- Science outreach as an IMB science Ambassador and ATSE Wonder of Science Ambassador, engaging with school and community groups, visiting academic, industry and government dignitaries. Visited primary schools in rural Queensland to encourage students to pursue STEM.
- Characterised the neuronal depolarisation transcriptome using microarrays, identifying a ncRNA with a critical role in normal activity altered in schizophrenia.
- First described the oligodendrocyte precursor transcriptome using a combination of bulk and single-cell sequencing (results published as 2 manuscripts).
- Tutor/lead tutor at the School of Chemistry and Molecular Bioscience

Karolinska Institutet *Stockholm, Sweden*
VISITING RESEARCHER *Sep 2014 - Dec 2014*
Key achievements:

- Awarded a competitive Boehrinher Ingelheim Travel Grant to work at the Department of Medical Biochemistry and Biophysics at Karolinska.
- Used SciLifeLab high performance computing cluster to analyse sequencing and microarray data to gain insights into white matter development in mice.

Education

University of Sydney *Sydney, Australia*
Senior Fellowship, UK Higher Education Academy *Oct 2021*

Graduate Certificate in Educational Studies (Higher Education) *Feb 2019 - Mar 2020*

Fellowship, UK Higher Education Academy *Jan 2020*

University of Queensland *Brisbane, Australia*
PhD, Genomics & Bioinformatics *Nov 2010 - Mar 2016*

Lomonosov Moscow State University, Department of Molecular Biology *Moscow, Russia*
Specialist, Biochemistry with a major in Molecular Biology *Sep 2005 - Jun 2010*

Training development & delivery

This is a list of training I have developed and/or delivered. All training delivered in-person unless otherwise specified.

2021	Petroleum Exploration Society of Australia Python for Geosciences (online, attendees from Santos, BHP, Origin, Beach Energy etc)
2021	eHealth@Sydney Python for analysing disease spread & electronic health records (online)
2020	Sydney Informatics Hub Machine learning in R (online and in-person); Machine learning in python (online); Publication-quality tables in R (online); Profiling R code (online); Teaching at the Informatics Hub
2020	University of Sydney Business School Business School Data Science Summer School
2020	Carpentries Carpentries Instructor Training (online)
2019	Monash University Carpentries Instructor Training
2019	University of Sydney Business School Business School Data Science Summer School
2019	Sydney Informatics Hub Machine learning in R (in-person x 4); Machine learning in python (in-person x 4)
2019	UNSW Geospatial data analysis
2018	Women Who Code Sydney R for Data Analysis (https://github.com/dvanic/wwc2018)
2018	University of Sydney Brain and Mind Centre R for Biomedical Researchers
2018	Sydney Informatics Hub Geospatial data analysis
2018	CSIRO Introduction to SQL for the CSIRO Data School
2018	Macquarie University Introduction to R for reproducible scientific analysis
2017	University of Technology, Sydney Introduction to R, Unix and version control
2016	University of Sydney Introduction to python, Unix and version control
2016	pyconAU Introduction to data management, python for data analysis and version control
2015	University of Queensland Lead tutor for courses in microbiology, cell biology and bioinformatics
2014	University of Queensland Awarded competitive grant to develop flipped learning materials for bioinformatics course. Subsequent feedback indicated that ”these materials were a lifesaver for the course when the COVID pandemic and associated pivot to online learning occurred”.
2014	University of Queensland Lead tutor for courses in bioinformatics and cell biology
2013	University of Queensland Lead tutor for courses in bioinformatics, cell biology and microbiology
2012	University of Queensland Tutor/lead tutor for courses in microbiology and cell biology
2011	University of Queensland Tutor/lead tutor for courses in microbiology and genetics

Professional development

2021	Research Portfolio Mentorship Program University of Sydney program to support emerging leaders
2020	Dare To Lead University of Sydney based on the work of Brene Brown
2020	Deep Learning for Natural Language Processing Monash University
2020	Introduction to Deep Learning and TensorFlow Monash University
2020	Introduction to Power BI Datacamp
2019	Quality Coaching Conversations: Skills to Thrive University of Sydney
2019	Women in Leadership Course: Foundation Skills Centre for Continuing Education,University of Sydney
2018	Interpersonal Skills and Effective Communication Centre for Continuing Education,University of Sydney
2018	Introduction to Machine Learning Datacamp

Awards and certifications

2020	RStudio Certified Trainer RStudio Inc
2019	HLTAID011 First Aid Australian Red Cross
2018	Carpentries Trainer The Carpentries
2014	Graduate Teaching Assistant (GTA) University of Queensland
2014	Carpentries Instructor The Carpentries
2014	Boehringer Ingelheim Travel Grant Boehringer Ingelheim Fond
2014	Student Poster Prize 35th Lorne Genome Conference
2014	Blue Card Commission for Children and Young People
2013	1st place, 3 minute thesis competition Institute for Molecular Bioscience,University of Queensland
2011	ANZ Trustees Scholarship for Medical Research ANZ Bank
2011	Radiation Use Licence Queensland Government
2009	R.B. Khesin Award for Outstanding Junior Thesis Lomonosov Moscow State University