Steve Bicko Cygu

P.O Box 570 - 00200, Nairobi

+254 713 246 604 cygubicko@gmail.com www.stevecygu.com

Profile summary

- I am a self-motivated and highly organized **Data Scientist** with 5 years of experience in the fields of data science, machine learning, mathematical and statistical modeling, and computational skills. Throughout my career, I have demonstrated proficiency in working with large and complex datasets, employing various analytical and statistical approaches to tackle a diverse range of challenging problems.
- With a strong foundation in quantitative analysis and data presentation, I have effectively contributed to informed decision-making processes for product improvement and business development. My goal-oriented nature drives me to set clear objectives and closely monitor product performance to identify areas for enhancement and growth.
- As a collaborative team player, I thrive in diverse work environments, leveraging my interpersonal skills to foster effective teamwork and achieve collective goals. My passion lies at the intersection of computational approaches and public health, where I have actively contributed to the development of machine learning and artificial intelligence solutions.
- Being an active member of the R open-source community, I am deeply invested in the advancement of R packages, furthering the accessibility and efficiency of data science tools. Additionally, my involvement as a Co-Investigator in ongoing projects has given me valuable experience in grant writing and scientific publication.
- If you seek a dedicated and innovative Data Scientist with a proven track record of leveraging data for impactful insights, I am confident that my diverse skill set and passion for data-driven solutions make me an ideal candidate for your team. Let's connect and explore how I can contribute to the success and growth of your organization.

Professional experience

Nov., 2022 – Date

Data scientist

African Population and Health Research Center (APHRC), Nairobi – Kenya

- Exploration, transformation and loading (ETL) development
  - Harmonization of various data sets from APHRC's microdata portal and health and demographic surveillance sites (HDSS) within East Africa into common data models (CDMs). These CDMs are currently being implemented in Observational Health Data Sciences and Informatics (OHDSI) ATLAS for cross-site comparison with other studies.
  - Developed code base in R and SQL to automate the process of generating observational tables and CDMs for similar data sets. The codes have been adopted by other team-members working on similar projects.
- Data mining, visualization and modelling
  - Developed R and python codes to extract data from various data portals hosted within APHRC and from the HDSS.
  - Worked on standardization and merging of the data sets from various platforms to a useable and shareable format.
  - Developed R shiny dashboards to provide real-time visualization of surveys, data sets in APHRC's microdata portal and to fit and summarize simple models.
  - Developed pipelines for automated feature engineering, data preprocessing, hyper-parameter tuning and model selection for over 100 data sets which had never been utilized.
  - Identified over 10 data use-cases and worked with other team-members to build various predictive machine learning algorithms to identify health risk factors.
  - Developed R code (which will be developed into R package) to automate the process of data mining, visualization and modelling for similar data sets.
- Evidence products for learning
  - Led the cross-site analysis and writing of 2 draft manuscript for the HDSS data sets and fitted models to understand causes of mortality and morbidity among the residence of HDSS.
  - Provided technical supervision to the implementation of machine learning algorithms to predict loss-to-follow in Nairobi urban settlements for a two-stage population-based epilepsy prevalence study.

- Capacity building
  - Conducted bi-weekly training on introduction to data science to the researchers and HDSS data managers.
  - Trained data managers within INSPIRE network on data harmonization and standardization using CMDs and OHDSI.
- Contribution to common good and supervision
  - Introduced, developed and conducted on-screen interviews for junior data scientists joining the Center.
  - Led and mentored a team of 6 junior data scientists through weekly working sessions to train them on various aspects of data science.
- Project management and coordination
  - Managed and coordinated project for Global Partnership for Sustainable Development Data (GPSDD) on the training in data science for climate and health outcomes.
  - Conceptualized and developed proposal for the supermarket data project which aimed to investigate the relationship between non-communicable diseases and the purchasing patterns of residence of Nairobi city.

Nov., 2022 – Date **Research fellow**
International Digital Health and AI Research Collaborative (I-DAIR), Geneva – Switzerland

- Developed a no-code platform for end-to-end ML/AI prediction
  - Implemented the back-end python code for data exploration, feature engineering and selection, data partitioning, hyper-parameter tuning, model training, and prediction.
  - Implemented Django APIs to connect back-end with the web front-end interface.
  - Worked with the DevOps team to improve the web interface and customization to specific needs of the public health researchers.
- Installation of Research Infrastructure (RI), the so-called "RI Nodes"
  - Supervised installation of hardware and software for high performance computing and big data analytics, ML and AI with a data protection fabric.
- Capacity building
  - Worked with public health researchers in identifying various use-cases and compared the performance of the platform with the existing cloud-based platforms and R implementation.
  - Conducted training on the use of no-code platform and RI Nodes to the Researchers at APHRC.

Sep., 2018 - Sep., 2022 **Data science and machine learning research assistant**
MacTheobio Lab McMaster University, Hamilton – Canada

- Research
  - Worked on the application and development of computational approaches to understand public challenges.
- Developed software packages
  - R package (**pcoxtime**) for penalized survival analysis (available on CRAN)
  - R package (**varpred**) for isolated confidence intervals and bias-corrected predictor effects (on github)
  - R package (**satpred**) pipeline for survival analysis training and prediction (on github)
  - R package (**glmnetsurv**) a user friendly wrapper for fitting and validation of survival models using **glmnet** (on github)

May, 2021 - Aug., 2022 **Data scientist (Remote)**
African Population and Health Research Center (APHRC), Nairobi – Kenya

- Developed food system database
  - Identified various sources of data on food systems in Africa and developed R codes to periodically mine, clean and store the data into a standard database.
  - Conducted a gap analysis for the existing data sets on food system and came up with mitigation measures on how to go about missing data.

  - Helped in refining a dashboard to visualize country-specific indicators for food systems.
- Data management, exploration, visualization and analysis
  - Conducted daily data quality checks and cleaning and shared generated reports with the data managers for progress reports and policy briefs.
  - Created visual presentations for bi-weekly reports to help in formulation of data-driven research questions.
  - Implemented models to investigate the impact of the Innovative Partnership for Universal Sustainable Healthcare (i-PUSH) program on reproductive, maternal, newborn and child health services uptake and outcomes.
- Capacity building and training
  - Developed a training module on and conducted a weekly hands-on training *introduction to R* for public health researchers.

May - Aug., 2018 **Data analyst**
African Population and Health Research Center (APHRC), Nairobi – Kenya

- Data collection
  - Helped in scripting questionnaires using ODK to enable mobile-based data collection.
  - Developed R shiny dashboards for real-time survey data collection status update and monitoring.
  - Helped in sample size calculation.
- Data management
  - Performed daily data quality checks, cleaning and callbacks in cases of missing information or discrepancies.
  - Provided daily summary of the achieved targets to the field managers.
  - Developed pipelines and R codes for automatic data quality checks and report generation.
- Data exploration, visualization and analysis
  - Generated summary statistics to provide an overview of the data and inform further analysis.
  - Developed dashboards to visualize summary statistics and model results.
  - Conducted analysis of the collected data using various statistical methods (both basic and advance) in order answer various research questions.

Mar., 2017 - May, 2018 **Data processing manager**
Dalberg Research, Nairobi – Kenya

- Supported business development
  - Helped the business development team to develop proposals for new projects and clients.
  - Reviewed client reports and provided technical assistance to clients.
  - Prepared client-ready final reports and presentations.
  - Developed data management SOPs.
- Data collection
  - Helped in sample size calculation and questionnaire design.
  - Helped in scripting questionnaires using ODK and SurveyCTO to enable mobile-based data collection.
  - Developed R scripts for real-time survey data quality checks, collection status update and monitoring.
- Data management
  - Ensured daily data quality checks, cleaning and callbacks in cases of missing information or discrepancies were performed.
  - Developed R package (**safisha**) for automatic data quality checks and report generation.
- Data exploration, visualization and analysis
  - Developed dashboards to visualize summary statistics and model results.
  - Conducted analysis of the collected data using various statistical methods (both basic and advance) in order answer various research questions.

- Team management and capacity building
  - Managed and supervised a team of 10 junior data processing clerks and 5 data executives.
  - Started a weekly data enthusiasts meet-ups to provide a training opportunity for fresh graduates who were interested in data and related fields.

Jan., 2016 - Mar., 2017 **Research student**
South African Centre of Excellence in Epidemiology Modelling and Analysis, Stellenbosch – South Africa

- Developed statistical and mathematical models for the analysis of epidemiological data.
- Developed R software package for the estimation of reference ranges.

Jan., 2014 - Sep., 2014 **Data processing manager**
Infotrak Research and Consulting, Nairobi – Kenya

- Provided technical assistance to business development team through proposal development and sample size calculations.
- Helped in survey tools design and development.
- Supervised data collection and entry of paper questionnaire data files.
- Performed database management and provided data access to clients once surveys were completed.
- Provide daily real-time updates on survey status and reports on the achieved daily targets.
- Conducted data cleaning, processing and analysis.
- Supervised and trained a team of 7 junior data processing clerks.

Teaching Experience

Sep., 2018 - Sep., 2022 **Teaching assistant**
McMaster University, Hamilton – Canada

- Introduction to data science theory
- An introduction to bioinformatics
- Introduction to mathematical and scientific programming
- Population ecology

May, 2021 - Oct., 2021 **Lead trainer**
African Population and Health Research Center (APHRC), Nairobi – Kenya

- Data management
- Data analysis
- Programming in R

Academic administrative experience

2020 – 2022 Vice President, McMaster University Chapter of SIAM.
2012 – 2013 Student Representative (Academic Director), Maasai Mara University.
2011 Treasurer Information Club, Maasai Mara University.
2010 Chairman Mathematics club, Maasai Mara University.

Education

2018 – 2022	**PhD in Computational Science and Engineering**, McMaster University, Canada.
2016 – 2017	**MSc in Mathematics**, Stellenbosch University, South Africa.
2014 – 2015	**MSc in Mathematical Sciences**, University of Cape Town, South Africa.
2009 – 2013	**BSc in Applied Statistics with Computing**, Maasai Mara University, Kenya.
2004 – 2007	**Kenya Certificate of Secondary Education**, Waondo Secondary School, Kenya.

Honors and awards

2016	South African Centre of Excellence Epidemiological Modeling and Analysis Scholarship.
2014	African Institute for Mathematical Sciences (Next Einstein Initiative) Scholarship.

Conferences attended

2019	Compute Ontario HPC Summer School, Hamilton, Canada
2016	Bayesian Analysis of Longitudinal Studies, Stellenbosch, South Africa
	Quantitative Bias Analysis with Epidemiological Data, Stellenbosch, South Africa
	Mathematical Modeling for Infectious Diseases, Cape Town, South Africa
	Clinic on Meaningful Modeling of Epidemiological Data (MMED)

Digital skills

- Data science: R, C++, Python, SQL, OMOP.
- Machine learning: R, Python.
- Data analysis: R, SPSS, Stata.
- Data visualization: R shiny and dashboards, Microsoft Excel, PowerPoint.
- Questionnaire scripting: ODK, Kobo and Google form.

Publications

1. **Steve Cygu**, Hsien Seow, Jonathan Dushoff, and Benjamin M. Bolker. *Comparing machine learning approaches to incorporate time-varying covariates in predicting cancer survival time. Sci Rep 13, 1370 (2023). https://doi.org/10.1038/s41598-023-28393-7.*
2. **Steve Cygu**, Benjamin M. Bolker, and Jonathan Dushoff. *Outcome plots: uncertainty estimation and bias correction for predictions and effects in simple and generalized linear (mixed) models; Preprint.*
3. **Cygu, Steve**, and Benjamin M. Bolker. *"pcoxtime: Penalized Cox Proportional Hazard Model for Time-dependent Covariates." arXiv preprint arXiv:2102.02297 (2021)*; submitted to *Journal of Statistical Software*
4. **Steve Cygu**, Helen Payne, Denise Lawrie, Debbie Glencross, Martin Nieuwoudt. *"Determining paediatric Immune Biomarker Reference Ranges using a model-based, age-continuous estimation method." Preprint.*

Open-source software developed

1. R package (**pcoxtime**) for penalized survival analysis (available on CRAN)
2. R package (**vareffects**) for isolated confidence intervals and bias-corrected predictor effects (on github)
3. R package (**satpred**) pipeline for survival analysis training and prediction (on github)
4. R package (**glmnetsurv**) a user friendly wrapper for fitting and validation of survival models using **glmnet** (on github)
5. R package for the clinicians to estimate age-related in health South African children (on github)
6. R package (**safisha**) for automated data quality and report generation
7. R scripts for scrapping web pages and Twitter to extract data and store them in usable formats (web dashboards)