 Summary

Data scientist familiar with gathering, cleaning, and organizing data for use by technical and non-technical personnel. Advanced understanding of statistical, algebraic, and other analytical techniques. Highly organized, motivated and diligent with significant background in predictive analytics.

**LUCHAO QI**

**SOFTWARE PORTFOLIO**

R Packages

* MRIPCA: Principal component analysis (PCA) on MRI data <https://github.com/LuchaoQi/MRIPCA>
* MRIcloudT1volumetrics: Volumetric analysis of MRI data <https://github.com/bcaffo/MRIcloudT1volumetrics>

R Shiny Web Applications

* Clustering analysis using K-means, t-SNE, and UMAP <https://github.com/LuchaoQi/Shiny_clustering>
* BMI Calculator <https://luchao-qi.shinyapps.io/BMI_Calculator/>

Hosted R Shiny website comparing machine learning algorithms (PCA, kmeans, UMAP, and t-SNE) & visualized clustering results using ggplot2 and plotly

Demo: https://github.com/LuchaoQi/Shiny\_clustering

**PROJECT EXPERIENCE**

Reinforcement Learning: Super Mario Bros (NES) 02/2020 – 05/2020

AI that Learns to Play Super Mario Bros Using Deep Q-Network (DQN) in TensorFlow Demo: <https://github.com/LuchaoQi/Reinforcement_Learning>

Recognized for expertise in utilizing OpenAI Gym; emulated Nintendo Entertainment System for designing and building a reinforcement learning environment through Nes-Py in Python

Pioneered the process of implementing 5 hidden layers as an agent in TensorFlow for structuring a convolutional neural network (CNN) model

Drastically reduced training downtime by 20% using Adam optimizer while training the agent on deep Q- learning

Achieved 2X faster than average of human players with trained agent to complete tasks successfully

Natural Language Processing: Amazon Rating Prediction 09/2019 – 12/2019

Use of Machine Learning to Detect Fake or Abusive Amazon Product Reviews Demo: <https://www.kaggle.com/luchaoqi/amazon-review-rating-prediction>

Exceeded expectations in using pandas, NumPy and dfply for processing Amazon Food Review data

Initiated the use of NLTK for feature construction while tokenizing the unstructured text of user reviews

Proven success in using the bag-of-words model (unigram/bigram) with scikit-learn for converting text to vector

Forecasted clients’ ratings using logistic regression with 0.94 AUC and also steered a 3% increase in negative review detection to identify abusive entities (sellers & reviewers) via random forest

Investigating Yelp User Funnels, Key Performance Indicators (KPIs) 01/2018 - 03/2018

Yelp User & Restaurant Performance Analysis Through SQL. Demo: <https://github.com/LuchaoQi/Yelp_Data_Set_SQL>

Excelled in utilizing Xpaths & BeautifulSoup for programming web crawler to scrape/parse unstructured data from Yelp

Migrated a file of 10GB into the database developed using MySQL Workbench

Determined and visualized the geographical distribution of restaurants with average ratings using Tableau

Measured customer engagement using well-designed metrics (bracket retention, DAU/MAU) while also conducting A/B testing for exploring other methods for improving upon KPIs

Paul C. Lauterbur Lab, Shenzhen, China 11/2016 – 05/2017

Research Associate

Utilized spectral analysis for the project on EMG signal pattern recognition for hand gestures

Supported and participated in R&D electrophysiology, cardiovascular telemetry, general drug safety, and electroretinography studies

Maintained, reviewed, and reconciled safety database and clinical database and evaluated safety trends and risk assessment

Pioneered the process of using Fast Fourier Transform for denoising time-domain signals of 200 gestures

Generated an 82% accuracy rate in using support vector machines (SVM) for classifying hand movements

Delivered extensive and innovative insights for medical rehabilitation systems while increasing the neural network training accuracy by 3%

**Data Analyst Intern**

• Experience writing unit tests for R Packages  
• Experience interfacing R with external APIs and other programming languages.

<https://www.ziprecruiter.com/jobs/tekreqs-inc-473c3612/senior-r-programmer-data-science-platform-8fb2a15d?job_id=c8d5a646b0ede5bae503c9b440ce7cdd>

**Next Generation Sequencing (NGS): RNA-Seq** Nov 2018 - Jan 2019

Differential gene expression (DGE) analysis & Gene set enrichment analysis (GSEA) of RNA-Seq data

Demo: <https://github.com/LuchaoQi/NGS>

Created tools (Bash script, R, Python) to perform pipeline analysis from downloading the raw Sequence Read Archive (SRA) gene data to investigating the differentially expressed gene matrix

Performed gene set enrichment analysis (GSEA) of profiles obtained from Gene Expression Omnibus (GEO)

Identified significant (p-value < 0.05) co-occurring or mutually exclusive mutated driver genes across different cancer types using Fisher’s exact test, Chi-Square test and Permutation test

Identified 50 over-represented genes that may have associations with disease phenotypes

**Investigation of Yelp user funnels, Key Performance Indicators (KPIs)** Nov 2018 - Jan 2019

Performance analysis of Yelp users & restaurant using SQL

Demo: <https://github.com/LuchaoQi/Yelp_Data_Set_SQL>

Wrote web crawler to scrape and parse unstructured data from Yelp using Xpaths, BeautifulSoup in Python

Created a database using MySQL workbench and imported ~10 GB data file into the database

Visualized geographic distribution of restaurants with average ratings using Tableau

Performed metrics analysis (bracket retention, DAU/MAU) using SQL to measure customer engagement and making suggestions for ways to improve upon KPIs via A/B testing

## **PUBLICATIONS**

1. **Qi L**, Zhang Q, Tan Y, et al. Non-contact High-frequency Ultrasound Microbeam Stimulation: A Novel Finding and Potential Causes of Cell Responses. *IEEE Trans Biomed Eng* 2019.
2. **Qi L**, Zhang Q, Lam KH, et al. Calcium fluorescence response of human breast cancer cells by 50-MHz ultrasound microbeam stimulation. Presented at 2017 IEEE International Ultrasonics Symposium (IUS), 6-9 Sept. 2017 2017.

你如果听了我之前关于data简历上的项目讲解的话，可以发现我一直在强调pipeline的问题，那么在做一个data相关的项目的时候，1. 先要告诉你的读者(可能是hr/recruiter/interviewer)，你为什么要做这个项目，也就是motivation，use case或者要达到的目的和项目功能。2. 说明data的获取来源，比如爬虫获得还是怎样，当然了，从raw data开始是工业界的人最希望看到的。如果实在data相对不是很complex，那么也可以先写一下如何进行的预处理(pre-processing)，包括怎么做的extraction，cleaning等等。3. 预处理之后可以想想怎样存取数据或者进行realtime的data互动，比如存到数据库中？那么这个过程包含数据库的设计。4. 开始使用数学模型来完成你的目的，比如用一些ml的常用算法，然后基于常用模型的优化，调参，比对各个模型的优势等等；5. 测试，是如何测试的最后结果？测试的结果怎么样？为什么你做的东西有价值？这个环节能量化就量化。能量化的结果最有说服力。6. 迭代整个pipeline，再不停优化。要说明优化办法。

数据科学简历修改指导：https://youtu.be/XuYUVlvzZPo