Pan **Tong**

Data Scientist

3026 Brookview Dr, Pearland, TX, 77584

832-589-6203 | nickytong@gmail.com | nickytong.github.io | nickytong

pan-tong-98155127 | Citizen: China, Permanent resident: US

"Be the change that you want to see in the world."

Summary

I am a data scientist eager to reveal unforeseen patterns, evaluate intriguing hypothesis and make predictions that help informed decision-making. I am keen to practice full-stack data science for improved productivity.

Education

University of Texas Health Science Center at Houston — *Houston, USA*

Ph.D. in Biomathematics and Biostatistics — *2008 - 2013*

- Thesis: integrative biomarker identification and classification using high throughput assays

Huazhong University of Science and Technology — *Wuhan, China*

B.E. in Bioinformatics — *2004 - 2008*

- Thesis: microRNA prediction using sequence conservation for fruit flies

Work Experience

Department of Bioinformatics and Computational Biology, UT MD Anderson Cancer Center — *Houston, Texas*

Research/Senior Statistical Analyst — *2013 - PRESENT*

- Develop novel methodologies for data analysis and publish research paper
- Provide collaborators with visualizations, reports, analytic insights, and recommendations enabling effective experiment planning, preclinical drug prioritization, hypothesis generation and validation

Data Science Projects

StackBooks: best programming books recommended by Stack Overflow experts

Techniques: Python | SQL | pandas |web crawling | Scrapy | text mining | sentiment analysis | D3 | web development | AWS Cloud

- Analyzed the entire Stack Overflow questions and answers (50G), indexed and managed the data with a SQL database to extract all books discussed in different tags.
- Scraped Amazon website for Book information including cover page, title, authors and book summary
- Performed text mining on the posts and questions including sentiment analysis and word frequency and visualized the data with D3/NVD3 library
- Built a website with Flask and deployed it on Amazon EC2 using Beanstalk, now live at: `http://stackbooks.us-west-2.elasticbeanstalk.com/`

Forecasting how many visitors a restaurant will receive (Kaggle)

Techniques: Machine learning | time series forecasting | feature engineering | cross-validation | model stacking | scikit-learn | pandas

- Performed exploratory data analysis for the Recruit Holdings' restaurant data including reservation data, visitor data, and restaurant characteristic data
- Built an AROMA model for each restaurant genre and predicted number of visitors each restaurant would receive during the 1-month test period
- Carried out feature engineering to extract visitor and reservation summary by week of day across stores, holiday lagging effect, encoding of restaurant genre and location.
- Implemented a fixed length cross-validation strategy for time series data and built a stacking model consisting of level 1 models light GBM, GBM, K-nearest neighbor and level 2 linear model.

Course projects spanning big data, text mining, machine learning and social network analysis

Techniques: Apache Spark | AWS Cloud | text mining | TF-IDF | topic modeling | supervised machine learning | social network analysis | node importance | link prediction

- *Introduction to Apache Spark and AWS*: Analyze the ebooks from the Gutenberg Project (the largest collection of free ebooks) by building a Spark cluster on EC2. Ebook in RDF format was manipulated using Python RDFlib and queried with SPARQL. From the collection of unstructured data, I extracted the number of times each ebook had mentioned males and females and finally stored the structured data in csv format on AWS S3.
- *Applied Text Mining in Python*: (I) Spam email classification. For feature engineering, I obtained TF-IDF for single word, 2-gram 5-gram as the primary features and added features such as email length, number of digits mentioned, and number of non-word characters. A L2 penalized logistic regression model was used to make the final prediction. (II): Evaluate the quality of paraphrases generated by human using document similarity calculated based on word similarity. (III) Topic modeling of news using Gensim's Latent Dirichlet Allocation (LDA) model.
- *Applied Social Network Analysis in Python*: (I) Given a company's email communication network, predict whether an employee will receive a manager-level salary. I extracted different node metrics including degree, closeness, betweenness and page rank, which department the person is working at and optimized a gradient boosted decision trees classifier using cross-validation that achieved an AUC of 0.95. (II) Given an employee connection network, predict future connections that employees will build later.

Building RESTful web service using Python Django and R

Techniques: Experimental design | Django | REST framework | Python-R interface

- Implemented various statistical randomization schemes using R including: complete randomization, stratified complete randomization, block randomization, stratified block randomization and minimization randomization
- Enabled seamless bi-directional communication between R and Python using Rserve
- Build a RESTful web service using Django which served as core computation engine empowering clinical trial design functions.

Predicting chemotherapy response using high throughput gene expression data

Techniques: mixture model | maximum likelihood estimation| outlier detection | feature selection | classification | cross-validation

- Developed the bimodality index approach using mixture models to identify features with sustained abnormality pattern. An R package SIBER was published to facilitate bimodality index applications.
- Built prediction models to classify chemotherapy response of cancer patients based on genomic expression data using linear discriminant analysis (LDA), k-nearest neighbors, support vector machine (SVM), random forest, artificial neural network and gradient boosting.

Generating client-side code from screenshot of graphical user interface (GUI) using deep neural network

Techniques: object tagging | image processing | convolution neural network | object classification | transfer learning | tensorflow-keras

- Tagged training and testing data from GUI images using imglab to generate bounding boxes and object labels for different GUI widgets.
- Using Faster RCNN (consisting of a region proposal network and an object classification network) as the object detection model, trained a Keras implementation to detect GUI elements on a cloud GPU machine.
- Assembled the detected widgets with bounding boxes into a hierarchical structure which was later compiled into client-side code.

SKILLS AND TOOLS

- High level languages: R (9 years) | Python (4 years) | SQL (2 years) | SAS/IML (1 year)
- Low level language: C++ (2 years) | Java (1 year)
- Machine learning tools: scikit-learn (3-years) | pandas (3 years) | Tensorflow-Keras (1 year) | Pig (1 year) | Hive (1 year) | Spark (2 years)
- Web technology: Django (2 years) | Javascript (2 years) | HTML (2 years) | MySQL (2 years) | MongoDB (1 year)

Software Products

- StackBooks: `http://stackbooks.us-west-2.elasticbeanstalk.com/`
- REST API for experimental design: `https://github.com/nickytong/graphmydata`
- drexplorer for dose-response analysis: `https://github.com/nickytong/drexplorer`
- SIBER for detecting outlier measurements: `https://github.com/nickytong/SIBER`

SELECTED PUBLICATIONS

Google Scholar: `https://scholar.google.com/citations?user=F_jaJakAAAAJ&hl=en`

- Tong, Pan, and Hua Li. "Mining Massive Genomic Data for Therapeutic Biomarker Discovery in Cancer: Resources, Tools, and Algorithms." Big Data Analytics in Genomics. Springer International Publishing, (2016): 337-355.
- Tong, Pan, et al. "SIBER: systematic identification of bimodally expressed genes using RNAseq data." Bioinformatics (2013): 605-613.
- Tong, Pan, and Kevin R. Coombes. "integIRTy: a method to identify genes altered in cancer by accounting for multiple mechanisms of regulation using item response theory." Bioinformatics (2012): 2861-2869.
- Tong, Pan, et al. "drexplorer: A tool to explore dose-response relationships and drug-drug interactions." Bioinformatics (2015): btv028.
- Akbani, Rehan, et al. "A pan-cancer proteomic perspective on The Cancer Genome Atlas." Nature communications (2014): 3887.