Jamal Ching-Chuan Chen

Data Scientist / Data Engineer

github.com/ChingChuan-Chen

CONTACT

6	+886-966-676-326
\geq	zw12356@gmail.com
in	www.linkedin.com/in/celestial0230

SKILLS

R / MatLab	Master
Statistics	Advanced
Statistical Learning	Advanced
SQL / Python	High-Intermediate
LaTeX / Bash	Intermediate
C++ / C#	Basic
M LANGUAGES	
• Chinese	Native speaker
● English	Advanced

Japanese REFERENCES

Jeng-Min Chiou Institute of Statistical Science Academia Sinica +886-2-2783-5611 ext. 312

diate (JI PT N3)

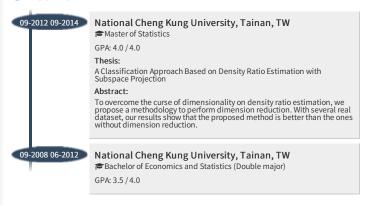
Sheng-Mao Chang Associate Professor Department of Statistics National Cheng Kung University +886-6-275-7575 ext. 53632 smchang@mail.ncku.edu.tw

SUMMARY

I am a data scientist and sometimes work as data engineer. About personality, I would say I am enthusiastic. I enjoy helping people to solve the difficulties they encountered.

- A person enjoying solving problems and sharing knowledge with people.
- A statistician worked deeply with data visualization, statistical methodologies and statistical learnings
- A engineer with creativity, critical observation and leadership.
- A experienced programmer skilled with R, Python, Shell, MATLAB, Scala and SQL.
- A skilled data engineer in data streaming / ETL, distributed computing and distributed

EDUCATION



WORK EXPERIENCES

Trend Micro Inc., Taipei, Taiwan Senior Data Scientist, Consumer, 01-2019 - Present

Along with a group of data engineers and domain experts, develop an IPS on network flows via statistical learning.

Projects

- Network behavior analysis / data scientist
 - *According to device, summarize the network flows to profile device behaviors.
 - ★Define a good score and threshold for device profiling with statistical sense.

Taiwan Semiconductor Manufacturing Company, Taichung, Taiwan

Senior Data Scientist / Engineer, CIM Department, 09-2018 - 01 Data Scientist / Engineer, CIM Department, 07-2016 - 08-2018

Develop automation systems on quality control during wafer processing from a big volume of data (3 billions per day).

- 1. WAT chart change detection / data engineer / data scientist
 - ★WAT is wafer acceptance test which is examined when finishing the process of a
 - $\bigstar \text{Common changing point analysis cannot be used in these data because there is no detailed orders in data. It only contains date information.$
 - ★I propose a algorithm to detect the daily changes based on statistics.
 - ★it is effective to detect the changes between upper control limit and lower control limit.
- 2. control chart change detection performance improvement / data engineer
 - \bigstar it originally use Hadoop MapReduce to split data into csv stored in chart level, then implement R on each chart.
 - \bigstar I propose to use MPI to accelerate the implementation time. I reduce the implementation time from 8 hours to 40 minutes.
 - $\bigstar The first key to reduce time is an algorithm to distribute the jobs with different running time which depends on data size.$
 - ★The second key is to use MPI without memory limit instead YARN container with only 1GB memory. (Note that because we cannot change the memory size of YARN container.)
- 3. Build up a development environment for data scientist / data engineer
 - $\bigstar Rely$ on Docker technology, we can provide a consistent and centralized controled development environment for data scientist.
 - ★I write a customized RStudio server Dockerfile to ensure everyone get the same
- 4. Data pipeline for processing history data and measurements data / organizer / data
 - ★Done by Spark written in Scala and Python. (UDAF is written in Scala.)
 - ★Good design on job and provide multiple configurations for users
 - ★Well monitoring for job implementation and data quality.
- 5. Propose, validate and construct a big data solution / organizer / data engineer

- $\bigstar \mbox{For the messy data query requirement for data analysts, I tested several SQL-on-hadoop solutions to test.$
- ★Cassandra is unable to get data by different primary key.
- $\bigstar \text{It spend too much time on query for Drill.}$
- ★Hive on Tez is chosen as our final solution.

Academia Sinica, Taipei, Taiwan

Research Assistant, Institute of Statistical Science, 09-2015 - 06-2016

Objective

Complete at least one research in the field of functional data analysis.

Projects

- 1. Imputation of functional data / data scientist
 - ★Use functional clustering to impute missing values in traffic data with lower RMSE than other methods.
- 2. Create a data streaming for researches (data from Taiwan freeway bureau) / organizer / data engineer
 - $\bigstar \text{Build}$ a data pipeline to transform crawled data from XML to JSON and store them into a MongoDB.
 - ★Develop a platform to view the data with d3.js via R shiny.
- 3. Travel Time Estimation / data scientist
 - ★Study journals about travel time estimation.
 - ★Realize the algorithms in journals and summarize the pros and cons.
- 4. Organize and refactor the source codes of previous researches / organizer
 - ★Study the previous researches and learn how FPCA works.
 - ★Remove several redundant blocks and improve performance of key functions.

JOURNALS

milr: Multiple-Instance Logistic Regression with Lasso Penalty

Ping-Yang Chen, Ching-Chuan Chen, Chun-Hao Yang, Sheng-Mao Chang and Kuo-Jung Lee *The R Journal* (2017) 9:1, pages 446-457.
♦ https://journal.r-project.org/archive/2017/RJ-2017-013/index.html

AWARDS



TSMC Kaggle Competition for the Defect Recognition

A internal competition in TSMC. There are over 100 teams to assist wafer factory decrease labor cost on the categorization of defects. There are only 3000 defect/reference images provided. The goal is do our best to get high accuracy rate on testing set (1200 images.). I used a 6-layer convolution nerval network with two Xception modules and win third place in 91.2% accuracy rate.



Competition for Data Analysis with R in Taiwan

◆ Honorable Mention

A national competition in Taiwan. There are over 30 teams to do a brainstorming on the data from a system to register the actual selling price of real estate. Each team have one day to come out a topic, and apply R language to complete and demostrate the results. Our team chose to predict the price of house from the messy data via LASSO approach.