Ching-Chuan (Jamal) Chen

Data Scientist / Data Engineer

CONTACT

- chingchuan-chen.github.io/
- +886-966-676-326
- zw12356@gmail.com
- www.linkedin.com/in/celestial0230
- github.com/ChingChuan-Chen

SKILLS

R / MatLab	Master
Statistics	Advanced
Statistical Learning	Advanced
SQL / Python	High-Intermediate
LaTeX / Bash	Intermediate
C++ / C#	Basic

LANGUAGES

Chinese	Native speaker
English	Advanced
Japanese	Intermediate (JLPT N3)

REFERENCES

Jeng-Min Chiou
Research Fellow
Institute of Statistical Science
Academia Sinica
+886-2-2783-5611 ext. 312
jmchiou@stat.sinica.edu.tw

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 currently working as a senior data scientist in Trend Micro which is a global well-known cybersecurity company. I am capable of developing efficient, well-designed applications of machine learning, statistical methodologies, and data visualization.

I'm also ...

- ☑ a statistician who developed several systems to detect the significant changes in the wafer manufacturing processes.
- ☑ a skilled distributed computing engineer who has experience in the distributed database who built a system to process 6 billion data per day.
- ☑ an experienced programmer skilled at R, Python, Shell, MATLAB, Scala, and SQL.
- ☑ a machine learning expert who won third place in an internal competition over 100 teams.

WORK EXPERIENCES

Trend Micro Inc., Taipei, Taiwan

Senior Data Scientist, Consumer, 01-2019 - Present

Taiwan Semiconductor Manufacturing Company, Taichung, Taiwan

Senior Data Scientist / Engineer, CIM Department, 09-2018 - 01-2019
Data Scientist / Engineer, CIM Department, 07-2016 - 08-2018

Objective
Develop several systems of quality control on wafer manufacturing based on the data with high volume and complexity.

Projects

1. WAT chart change detection / data engineer / data scientist
   - ★WAT is wafer acceptance test which is tested while finishing the process of a wafer. There are no detailed orders in data. It only contains date information.
   - ★Proposed an algorithm to detect the daily changes based on statistics. It is useful to detect the changes between upper control limit and lower control limit.
   - ★Parallelly processed 3 billions of records of data and output results of detections in 2 hours. It is done by the R language and the MPI.
2. Find out the key factors making wafer low yield / data scientist
   - ★Engage the statistical hypotheses to compare the yields of wafer produced by the different equipments.
3. Performance improvement of control chart change detection / data engineer
   - ★Proposed a new architecture powered by MPI to improve the speed of detection algorithm.
   - ★Reduced the implementation time from 8 hours to 40 minutes in the new architecture.
   - ★Proposed an algorithm to dispatch the detection jobs with different running time which depends on data size.
4. Reduce the rate of out-of-control quality measurements in the process / data scientist
   - ★Through the correlations, dig out the key factors to raise the rate of out-of-control.
   - ★Use a robust correlation coefficient estimation with robust variance estimations to avoid the influential points.
5. Build up a development environment for the data scientists / data engineer
   - ★Construct a consistent and centralized controlled development environment for the data scientists.
   - ★Write a customized RStudio server Dockerfile to ensure everyone gets the same environment.
6. Propose a methodology to recognize the relationships of quality measurements for the advanced process control / data scientist
   - ★Propose a methodology to estimate the effects of variables via a regression model.
7. Data pipeline for processing history data and quality measurements data / organizer / data engineer
   - ★Propose a fast and reliable data pipeline powered by Spark in Scala and Python. (UDAF is written in Scala.)
   - ★Any new ETL can be set flexible and easy for users. This UI is done via R shiny.
   - ★Good monitoring for job implementation and data quality.
   - ★Stored 6 billions of records of data into Hive with full automation and good data quality.
8. Propose, validate and construct a big data solution / organizer / data engineer
   - ★For the messy data query requirement for data analysts, I tested several solutions with SQL-like query language to test.
   - ★Tested several big data solutions like Cassandra, Drill, and Hive.
   - ★Made number of machine learning jobs can be done in 20 times shorter computing time than Oracle database.

Academia Sinica, Taipei, Taiwan

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

Objective
Develop the methodologies based on functional PCA applied to transportation.

Projects

1. Imputation of functional data / data scientist
   - ★Used 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
   - ★Built a data pipeline to transform crawled data from XML to JSON and store them into a MongoDB.
   - ★Developed 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 the performance of critical functions.

Ministry of National Defense, Taichung, Taiwan
Alternative military service, National Immigration Agency, 10-2014 - 09-2015

AWARDS

12-2017

TSMC Kaggle Competition for the Defect Recognition
Third Place

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

08-2014

Competition for Data Analysis with R in Taiwan
Honourable Mention

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

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

EDUCATION

09-2012 09-2014

National Cheng Kung University, Tainan, TW
Master of Statistics

GPA: 4.0 / 4.0

Thesis:
A Classification Approach Based on Density Ratio Estimation with Subspace Projection

Abstract:
To overcome the curse of dimensionality on density ratio estimation, we propose a methodology to perform dimension reduction. With several real datasets, our results show that the proposed a method is better than the ones without dimension reduction.

09-2008 06-2012

National Cheng Kung University, Tainan, TW
Bachelor of Economics and Statistics (Double major)

GPA: 3.5 / 4.0

Ching-Chuan (Jamal) Chen, chingchuan-chen.github.io/
+886-966-676-326, zw12356@gmail.com