**Jin (Ryan) Kang**

[kangjin2014@gmail.com](mailto:kangjin2014@gmail.com) |11 Charlotte, Toronto M5V 2H5 | 647-655-0007

Data Science Blog: <https://dare2ai.wordpress.com> |Github: https://github.com/kangjin2014

**Data Engineer Consultant***, iFuture.ai, Toronto Aug 2017 – Present*

• Big Data Pipeline

- Architect data warehouse in Hadoop eco-system on AWS or Google Cloud

- Developed real-time streaming data pipelines using Kafka to ingest data into NoSQL databases and HDFS

- Experimented with Docker containers deployment in clusters using Kubernetes

- Orchestrate workflow automation using Apache Airflow and Linux Crontab

- Manage the projects using Github and Dockerhub

• Click-through rate prediction modeling for mobile advertising client:

- Used Pyspark to explore, process, and engineer features from billions of purchase/click-view records

- Use Spark ML and Python Sklearn to build cross-validated regression/classification models such as random forest,

decision tree, SVM etc

- The model achieved 20% improvement with an AUC statistics of 0.81 and was selected by client for pre-production

deployment

**Data Science Trainee,** *WeCloudData Inc., Toronto Apr 2017 – Jul 2017*

• Collaborative-filtering recommender project:

- Developed a web application using Kafka, Spark on EMR, Redshift, DynamoDB and S3 for real-time

product recommendation based on collaborative filtering algorithms

- Developed web application backend using Flask and scaled up the app using Docker/K8s

• Large-scale ETL and machine learning classification

- Used Scrapy on AWS to scrape and collect online data (TBs)

- Leveraged image recognition tools to label samples for machine learning (10x faster rate than previous)

- Developed machine learning classification algorithm using Python Sklearn and NLP (TF-IDF) and

achieved **95%** **classification prediction accuracy**

• Digital marketing ETL and dashboard

- Built a data application to collect and process data collected from Google Analytics and social media APIs (Google

Analytics, AdWords, Twitter, Facebook, Eventbrite, Meetup, internal database, etc) to help the client track the

company’s digital marketing and SEO effort

- Visualize the reports in open source dashboard for insightful data-driven decision making

**Graduate Research & Teaching Assistant,** *Ryerson University, Toronto Sep.2014 – May.2017*

- Biostatistics modeling by Python and Matlab

- Automation of data collection, process and analysis from machine hardwares (IoT)

- Statistics, Math Modeling (100+ hours lecturing, 250 students)

**EDUCATION**

Data Science Immersive Bootcamp, WeCloudData 2017

Master of Engineering, Chemical Engineering, Ryerson University 2014-2017

Bachelor of Science, Xiamen University