Yezheng Li

2678093088, yezhengli9@outlook.com,
4119 Powelton Avenue, Philadelphia, PA, 19104

Education

University of Pennsylvania, PA, USA
PhD candidate, Applied Math and Computational Science Advisor: Hongzhe Li *Sept.2015–Nov. 2020*
University of Pennsylvania, PA, USA
Master of engineering, Computer and Information Science *Sept.2015–May. 2020*
Social activity: Wharton quantitative and trading group
Nanjing University, CN
Undergraduate, major in Information and Computational Mathematics *Sept.2011–June 2015*

- **Ranking** Sept. 2011-Mar. 2014: Major 1/137 Overall 1/137 **GPA** Major 96.46/100 Overall 95.39/100

University of California Davis, CA, USA
Exchange student *March.2014–July 2014*

Internships

Philips Lighting Research → Signify, Cambridge, MA, USA
Data science research intern, supervised by Dr. Jasleen Kaur *May 2018 – Aug.2018*

There are multiple time series signals collected for each street light (every 15 mins, 30mins, 1 hour, etc.)
(1) energy meter reading time series , (2) switching points time series, (3) faults reported from devices.
Philips Lighting has already created a Smart Alarm System reporting abnormal behaviors (of street lights) based on (1, 2).
My task is to explain this abnormal behaviors. 50% abnormal behaviors can be explained by following three approaches.

- explained by (3): time series matching of device faults and SmartAlarm system.
- Explained by manufacture information (of street lights): lifetime prediction of street lights based on weather data, switching cycles, dimming level, etc. We can identify that street lights produced by certain manufacturer has relatively shorter lifetime (more abnormal behavior). The function of lifetime prediction is given and since there are fewer maintenance information, optimization/ parameter estimation is not difficult.
- explained by street-level abnormal behaviors: By time series clustering (together with geographic/spatial clustering), street-level abnormal behaviors (for several days, for a weeek, etc.) can be identified .

Cities are not only in US, but also Southeast Asia, South America, etc. We use postgresql on Amazon Redshift, data visualization via Tableau and plotly@python, etc. Database of each city is 10-200GB with 20,000-40,000 street lights.

Programming skills (github: yezhengli-Mr9)

- Python, C++/C, java, R, tensorflow, Apache Spark, SQL (MIcrosoft SQL server, mysql, postgresql),kdb+ (c++ import 3598 US stocks one month within 5 mins-single thread), MongoDB, Neo4j, MATLAB, MAPLE, Mathematica, TEXm multiprocess/ mulithread programming, data visualization via Tableau, Plotly, etc. NodeJS. ETL pipeline (parsing/ reformating JSON object from logs, standardize timestamps.)

  Machine learning NLP, computer vision tasks, including deep learning (trained by multiple GPUs). AWS (EMR, S3, EC2), GCP, etc. MapReduce with Hadoop. Entry-level: SSIS, Go, scala, spark.
- Projects related to artificial intelligence:
  - Computational linguistics **(2017)**: github.com/yezhengli-Mr9/CIS530-computational-linguistics
    - Email spam filter via bag of words (naive Bayes), hidden Markov model (including Viterbi algorithm).
    - Text embeddings: term-document matrix, term-context matrix, word2vec, GIoVe, ELMo, BERT embeddings etc. Sentence embeddings, paragraph embeddings.
    - Text classification: Document classification classifier of encrypted documents with AWS deployment (python flask + nginx, uWSGI). **(2018)** github.com/yezhengli-Mr9/doc-class (For the startup HeavyWater@Philadelphia).
    - Name entity recognition (on Spanish. Transfer learning of parameters from English task to the Spanish task. bidirectional LSTM and use BERT pretrained embeddings to improve results), question answering (attention flow network, dynamic coattention network), CBOW model, LSTM-CRF model, (implement conference papers' github code from English into a different language) etc.
  - Machine perception course projects: **(2018)**

    Optical flow, image stitching; neural networks: GAN, variational autoencoders, etc.

    Object detection in neuron-surgery videos (independent study ROBO 599 with Prof. Jianbo Shi) **(2019)**

    Use top-down Mask-rcnn and bottom-up ExtremeNet (together with hourglass network) to instruments (as well as detect neuron, spin cords – which is not my focus) ok f in neuron-surgery videos. Project is implemented by pytorch with multiple GPUs (at most 4 at the same time).

- Projects related to system design:
  - SQL: Database system for 2012 sports meeting Nanjing University, Database system for Chinese stock market with connection to specific stock-market analysis software (**2013**), Pyspark SQL for graph BFS and PageRank (**2018**).
  - Parallel sorting (via multiprocess, multithread, shared memory, etc.), SMTP-POP3 servers (email storages in Mbox formats and connection to other email clients, for example, Thunderbird), Multicast for distributed chat system (different multicast ordering – FIFO/ CAUSAL /TOTAL ordering) **(C++, 2017)**
  - PennCloud (a simplified Google Drive system; joint work with Yuding Ai, Thomas Greening, Yeru Liu, Li Huang): github.com/yezhengli-Mr9/CIS505-networked-system **(c++, 2017)**

    Build a distributed system analogous to Google Drive with SMTP email server connect to external mail system (Hotmail, QQ, etc.); with Cloud Drive upload of various files. Fault tolerance and consistency of the model are also emphasized. Also run AWS with partial features realized.

    Backend and frontend are connected with grpc and google protocol buffer.
  - Search engine (crawler&indexer&PageRank, I am in charge of indexer). With 200 GB data on AWS S3 bucket, 20-35 GB on mysql RDS (inverted index table with 1.7 1.9 × 10^8 (word,url) pairs). There is a version of indexer run on EMR. Deployed on AWS. This is from CIS555 course project. **(java, 2020)**

Research experience (Google scholar: Yezheng Li)

Generative adversarial net for cross-tissue missing imputation of GTEx

Yezheng Li, Hongzhe Li *May 2020 – Oct. 2020*

- Generative adversarial net for missing imputation of GTEx with cross-tissue information and SNPs.

Multi-sample estimation of Bacterial centroid log ratio matrix in metagenomics data

Yezheng Li, Yuanpei Cao, Hongzhe Li *March 2019 – Now*

- Multi-sample estimation of Bacterial centroid log ratio matrix via nuclear norm minimization.

Object detection in neuron-surgery videos

Guided by Dr. Jianbo Shi *Jan 2019 – May 2019*

- Use top-down Mask-rcnn and bottom-up ExtremeNet to detect neuron, spin cords, instruments in neuron-surgery videos. Project is implemented by pytorch with multiple GPUs. (ROBO 599: independent study with Prof. Jianbo Shi).

Two-sample test statistic of community structures via adjacency matrices

Yezheng Li, Hongzhe Li *Aug. 2017- Dec. 2017*

- Design test of network community structure (theory):

  We propose a test statistic for two-sample hypothesis test of independent views based on eigenspace geometry of adjacency matrices. Customized normalization enables us to handle membership blocks of different sizes.

Stochastic processes and control theory(undergraduate thesis) Nanjing University, CN

Supervised by Prof. Wanyang Dai *Sept. 2015-June 2015*

- Stochastic systems and optimal resource-performance controls in future wireless communication networks

ARock new: Asynchronous Parallel Coordinate Updates

Zhimin Peng, Yezheng Li, Yerong Li, Wotao Yin *Aug. 2014- Mar. 2015*

- Finding a fixed point to a nonexpansive operator, i.e., $x^* = Tx^*$, abstracts many problems in numerical linear algebra, optimization, and other areas of scientific computing. To solve fixed-point problems, we propose ARock, an algorithmic framework in which multiple agents (machines, processors, or cores) update x in an asynchronous parallel fashion. (My contribution: several operators and unittests forARock-new – a private github system)

Convex optimization Nanjing University, CN

Suprevised by Prof. Bingsheng He *April 2013-April 2014*

- Topics on Unified Framework of Alternating Direction Method, convex optimization. Witness an outstanding work ”Chen, Caihua, et al. ”The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent.”

Courses

- AB testing (Udacity), CIS550 Database and Information Systems FNCE 717 Financial Derivatives (audit), CIS 555 Internet and Web Systems (upcoming)
- ROBO 599 independent study, CIS 680 Vision and Learning (audit)
- CIS530 Computational linguistics, CIS800 Peeking into black box of DL models, CIS 502 Analysis of algorithm, CIS 505 Software system, CIS 580 Computer Vision and Computational Photography, CIS 521 Introduction to AI
- STAT 991 Statistical Network Data Analysis, STAT 541 Statistical methodology, STAT 530 Probability, STAT 531 Stochastic processes, BSTA 789 Big data, STAT 971 Linear model , STAT 972 Advanced topic in statistics
- AMCS 608/ 609 Analysis, ESE 605 Modern convex optimization, STAT 770, AMCS 603 Numerical linear algebra, ECON 681 Microeconomics theory