Lizhen Shi

http://ww2.cs.fsu.edu/~lshi/

Office: MCH 102F
Tel: (205)542-8124
E-mail: Lizhen9.shi@gmail.com / lshi@cs.fsu.edu

Computer Science
Florida State University
600 W College Ave, Tallahassee, FL 32306

EDUCATION

Florida State University   Tallahassee FL
Ph.D. Candidate in Computer Science (GPA: 4.0/4.0)   Expected May 2020
- Dissertation: Computational Strategies for Efficient and Scalable Genomics Analysis
- Committee: Michael Mascagni (Advisor), Xin Yuan, Peixiang Zhao, and Peter Beerli

Auburn University   Auburn AL
M.S. in Computer Science (GPA: 4.0/4.0)   Aug. 2015
- Thesis: Feature Enhancement and Performance Evaluation of BioPig Analytics
- Committee: Weikuan Yu, Sanjeev Baskiyar, and Saad Biaz

North China Electric Power University   Baoding China
B.S. in Computer Science   Aug. 2007

RESEARCH INTERESTS

Data Science, Machine Learning, Big Data, Distributed Computing, Computational Biology, Image Processing, Natural Language Processing

PROFESSIONAL EXPERIENCE

Kaggle Master   08/2017 - present
- Participated in 16 predictive modeling competitions in a wide variety of domains hosted by companies and research institutes.
- Competed with statisticians and data miners from all over the world, aiming for the best models.
- Got 1 gold, 3 silver and 1 bronze medals.

Florida State University   08/2015 – present
- Built a generic tool for representing DNA sequence as a vector using word embedding in NLP and Locality Sensitive Hashing.
- Designed a hybrid clustering algorithm based on LPA for metagenome read clustering
- Developed a scalable sequence clustering tool named SpaRC based on Apache Spark.

- Designed a cross-layer scheduler for improving DAG-structured query processing in MapReduce.

Lawrence Berkeley National Lab   05/2015 – 08/2015
- Upgraded BioPig toolkit from Hadoop 1 to Hadoop 2 (2.7.0).
- Tuned Hadoop parameters for BioPig performance optimization.
- Implemented K-mer Similarity to extended BioPig toolkit functionality

Auburn University   08/2014 – 05/2015
- Implemented an MPI program to sort millions of integers using multiple sorting algorithms.
- Developed an android project that supports downloading files from top cloud storage providers.

PwC, CleNET, and Kingbase *(Beijing, China)*   06/2007 – 12/2012
- Developed/Maintained projects in various domains using C#, Java languages and SQL server database.

PUBLICATIONS

Lizhen Shi, Zhong Wang. *Computational Strategies for Scalable Genomics Analysis.* Genes 10.12 (2019): 1017.

Lizhen Shi, Bo Chen. *A Vector Representation of DNA Sequences Using Locality Sensitive Hashing*. (bioRxiv)

Lizhen Shi, Volkan Sevim, Michael Mascagni, Zhong Wang. *Leveraging long-read sequencing for cost-effective metagenome clustering.* (To be submitted)

Kexue Li, Lili Wang**, Lizhen Shi**, Li Deng, Zhong Wang. *Deconvolute individual genomes from metagenome sequences through short read clustering*. (Accepted by PeerJ with Revision)

Lizhen Shi, Xiandong Meng, Elizabeth Tseng, Michael Mascagni, Zhong Wang. *SpaRC: scalable sequence clustering using Apache Spark*. Bioinformatics 35.5 (2018): 760-768.

Ji Huang, Stefania Vendramin, **Lizhen Shi**, Karen M McGinnis. *Construction and optimization of a large gene Co-expression network in maize using RNA-Seq data*. Plant physiology 175.1 (2017): 568-583.

Lizhen Shi, Zhong Wang, Weikuan Yu, Xiandong Meng. *A Case Study of Tuning MapReduce for Efficient Bioinformatics in the Cloud*. Parallel Computing 61 (2017): 83-95.

Lizhen Shi, Zhong Wang, Weikuan Yu, Xiandong Meng. *Performance Evaluation and Tuning of BioPig for Genomic Analysis*. Proceedings of the 2015 International Workshop on Data-Intensive Scalable Computing Systems. ACM, 2015.

INVITED TALKS AND PRESENTATIONS

Nov 2019. "A vector representation of DNA sequences using Locality Sensitive Hashing." FSU CS expo, Tallahassee, FL.

Nov 2018. "SpaRC: scalable sequence clustering using Apache Spark." FSU CS expo, Tallahassee, FL.

Nov 2015. "Performance Evaluation and Tuning of BioPig for Genomic Analysis." International Workshop on Data-Intensive Scalable Computing Systems (DISCS) in conjunction with the ACM/IEEE Supercomputing Conference (SC'15), Austin, TX.

DATA SCIENCE EXPERIENCE

Participated in 16 predictive modeling competitions on Kaggle and got rich hands-on experience on data-intensive analytics, machine learning, image processing, etc. The projects and my achieved ranks are listed below:

Project	Rank
Web Traffic Time Series Forecasting	7/1095 (Top 1%)
Santander Product Recommendation	20/1786 (Top 2%)
Zillow Prize: Zillow's Home Value Prediction (Zestimate)	41/3775 (Top 2%)
The Nature Conservancy Fisheries Monitoring	101/2293 (Top 5%)
TensorFlow Speech Recognition Challenge	124/1314 (Top 10%)
Melbourne University AES/MathWorks/NIH Seizure Prediction	101/478 (Top 22%)
Outbrain Click Prediction	152/979 (Top 16%)
Recruit Restaurant Visitor Forecasting	276/2157 (Top 13%)
Predicting Molecular Properties	309/2749 (Top 12%)
Google Cloud & YouTube-8M Video Understanding Challenge	375/655 (Top 58%)
Two Sigma Financial Modeling Challenge	502/2066 (Top 25%)
Bosch Production Line Performance	597/1373 (Top 44%)
Corporación Favorita Grocery Sales Forecasting	929/1674 (Top 56%)
Quora Question Pairs	1394/3304 (Top 43%)
Mercedes-Benz Greener Manufacturing	1444/3831 (Top 38%)
Personalized Medicine: Redefining Cancer Treatment	101/1386 (Top 8%)

TEACHING EXPERIENCE

Instructor
- Microcomputer Applications for Business and Economics (CGS 2100): Spring 2020 (583 students)
- Computer Fluency (CGS 2060): Spring 2020 (680 students)
- Java for Non-Majors (CGS 3416): Fall 2019 (26 students)
- Computer Fluency (CGS 2060): Summer 2019 (358 students)
- Java for Non-Majors (CGS 3416): Spring 2019 (32 students)

Teaching Assistant
- Concurrent, Parallel, and Distributed Programming (COP 5570): Spring 2017- Fall 2018

Lab Assistant
- Computer Fluency (CGS2060, CGS2100): Fall 2015 – Fall 2016

HONORS / AWARDS

Second Place in CS expo presentation, FSU 2019
Second Place in CS expo presentation, FSU 2018
Best Employee Award, Kingbase 2009
Second Place in campus singing competition, NCEPU 2006

SERVICE (Ad Hoc Reviewer)

Computers in Biology and Medicine
Journal of RNA and Genomics
Recent patents on computer science

REFERENCES

Michael Mascagni (Research Advisor)
Professor
Department of Computer Science
Florida State University
253 Love Building
1401 Academic Way
Tallahassee FL 32306-4530
(850)644-3290
mascagni@fsu.edu

Xin Yuan (Committee)
Professor and Chair
Department of Computer Science
Florida State University
259 Love Building
1401 Academic Way
Tallahassee FL 32306-4530
(850)644-9133
xyuan@cs.fsu.edu

Zhong Wang (Internship Supervisor)
Genome Analysis Group Lead
DOE Joint Genome Institute
2800 Mitchell Drive
Walnut Creek, CA 94598
(925) 296-5795
ZhongWang@lbl.gov

Bob Myers (Teaching Supervisor)
Associate Teaching Faculty
Department of Computer Science
Florida State University
105C Love Building
1401 Academic Way
Tallahassee FL 32306-4530
(850) 644-0972
myers@cs.fsu.edu