Sungwook Yoon

Contact

Palo Alto, CA	sungwook.yoon@gmail.com	http://sungwookyoon.com	480 277 3458

Positions

present	Principal Data Scientist	MapR
2014 - 2015	Data Scientist	MapR
2013 - 2014	Sr. Data Scientist	Vectra Networks
2012 - 2013	Architect	Seven Networks
2012 - 2012	Data Scientist	Identified
2008 - 2012	Research Scientist	PARC (Palo Alto Research Center)
2006 - 2008	Assistant Research Professor	ASU (Arizona State University

Education

Ph.D.	Computer Engineering	Purdue University	AI, Machine Learning, Planning
M.E.	Electrical Engineering	Seoul National University	Video Compression, ATM
B.E.	Electrical Engineering	Seoul National University	Control and Instrumentation

Awards

2011	Best Paper Award Runner Up	Journal of Artificial Intelligence
2011	Best Paper Award Runner Up	International Conference on Automated Planning and Scheduling
2009	Best Machine Learning	International Learning and Planning Competition
2006	(Unofficial) Winner	International Probabilistic Planning Competition
2004	Winner	International Probabilistic Planning Competition

Projects Summary (details available)

Data Science Engagements	MapR	Various Customer Engagements / Teachings
Malware Expression Detection	Vectra Networks	Malware/Fraud Detection from PCAP MetaData
Mobile App Data Analysis	Seven Networks	App Data / Power Usage Analysis on Mobile Phone
Career Analysis	Identified	People's Job Transition Analysis
Workforce Planning	PARC	Call Center Staffing Solution
GILA (Machine Learning)	ASU	Machine Learning from Expertise

Skills

Expert Level	R, SQL, C, Lisp, Hadoop FS, MapReduce, Spark, Drill, Scheme, OpenTSDB
Mid Expert Level	Scala, H2O, Python, Django, Hive, Elasticsearch, Kafka, Storm, HBase
Intermediate	Bash, Linux Sysadmin, C++,Java, Ruby, GCE, AWS, Impala,Pig
Minimal Experience	Javascript, JQuery

Publications: 20+ Publications at only top journals, **JAIR**, **JMLR** and top conferences like AAAI, **NIPS**, **ICML**, **UAI**, ICAPS, IJCAI.

BigData Implementation Experiences (Successfully Delivered Solutions)

Hadoop Systems	MapReduce, Hive, Impala, Spark, Apache Drill, HBase, Storm, Kafka, OpenTSDB
Spark Systems	Core Spark, MLLib, SparkSQL, GraphX
MapR Systems	Cluster Installation, Cluster Validation, MapR DB
ML Systems	H2O, Spark MLLib, R Hadoop
Elasticsearch Systems	Elasticsearch on MapR, Kibana, Logstash, ES-Hadoop

Machine Learning and Data Science Experiences

Public Speech	Strata San Jose 2015, LA Big Data Camp 2014
MeetUps	Hadoop/Spark Talk @{LA, Portland, Seattle, Utah, Dallas} Speech all rated higher than 4 out 5 stars
Private ML classes	Delivered Machine Learning Workshops for Several customers
ML Development	Developed Spark/Scala/HiveHQL/Pig codes for various customers
ML Competition	Won 2009 International ML for planning competition Our work was the only ML technique that worked across domains

Selected Commercial/Government Development Experiences

Real-Time Network Anomaly Detection System @ Multiple Customers	We worked with a few fortune 500 companies for their IT security data analysis projects for several months. We used Spark to enrich the streaming data and used ES to Spark to ingest into Elasticsearch Visualization We developed machine learning system using Spark MLLib for the baseline analysis and traffic pattern We also used Spark GraphX to develop consistent network topology of the customer network. PageRank algorithm and Connected Component analysis in GraphX help the customer easily find significant lateral data movement. We used Scala for Spark development. Upon customers request, we perform Pyspark demo.

Data Ingestion Into MapR @ Multiple Customers	We performed several data ingestion services for multiple customers. Mostly from existing databases or streaming log text data, We ingested into either MapR FS, MapR DB or OpenTSDB. The tools used are, Sqoop, Spark Streaming, Logstash or Bash codes
Real-Time App Data Processing @ Multiple Customers	We performed for several customers on their raw application data log processing, ingestion and visualization. Depends on the type of the data, we used the most fitting methods. Be it, Logstash, Bash processing, Spark, Drill or Sqoop For visualization, we used Elasticsearch + Kibana solution to show real time data ingestion. For BI use cases, we used Drill ODBC to connect Bigdata to BI applications like Spotfire or Tableau.
Data Science Engagements @ MapR	Use Case Discovery with several customers Machine Learning code developed in Scala, Spark, H2O Lead successful workshop with customer on Machine Learning on Hadoop Delivered successful engagements in Use Case Discovery and Code development Developed Machine Learning on Hadoop course and lab material
Malware Expression Detection @Vectra	Vectra Networks specializes in detecting malware expression on packet flow. The product sits on the clients network, sniff packets then find malware expression in the packet flow. It is impossible to detect the infection, but the malware expressions are pretty limited. Among many expressions of malware, I specialized in DDoS detection (detecting clients asset joining DDoS attack). I analyzed, developed and produced the algorithm to the production level
Mobile App Data Analysis @Seven	Seven Networks optimizes out unnecessary traffics from mobile data use. I developed metrics for internal performance measure. I developed several visualization techniques for our products field performance. My visualization created unique views on our products behavior as well as mobile apps behavior in relation to phones screen activity as well as radio activity. This led to development of novel optimization ideas and implementations. I used Hive to access Big Data and I used R for visualization. For preprocessing the data for R, I used python and C-Sharp

Career Analysis @Identified	From massive data on peoples career (over 40 million people), we constructed peoples career path graphs. Used Java/SQL/Python/R. Modeled as HMM with LDA. I used NGramDistance to model emission probabilities from Job function to regularized Job Title. Developed mechanism for learning transition probabilities.
Workforce Planning @PARC	Web Component Development, for Xerox project, 2010. This is a Xerox project. Xerox/ACS maintains call centers and we try to optimize number of working call center agents. We modeled the call arrivals as bucketed Poisson arrival. Our call arrival model was highly accurate with more than .96 R-square measures. We then used AI planning technology to plan for agents hiring and firing strategies. We used Erlang to identify the number of agents needed to satisfy the service level. The final product was SaaS. We used JQuery/Tomcat/Servlet/Hibernate framework
Generalized Integrated Learning Architecture @ASU	DARPA GILA project, 2007. In military campaign through the air, the air-space is scare resource that every unit needs to share. We have access to the data that was recorded from the air-space managers operation. We learned from the data on how he/she selected particular missions and when she/he asked for modification. I made the knowledge representation framework. I designed the machine learning algorithm for the highly skewed data distribution. I coded and delivered in Java