Zongheng Yang

zongheng.y@gmail.com http://zongheng.me
Mobile: (510) 502-7057 github.com/concretevitamin

Education

University of California, Berkeley Aug 2012–Dec 2015 (expected)
B.A., Computer Science (Honors) & Mathematics (double major) GPA: 3.848 out of 4.0

Research Experience

Undergraduate Researcher, AMPLab, UC Berkeley Sep 2013–present

- **ZipG**. Designed and implemented *ZipG*, a distributed graph store that achieves three properties existing graph stores fall short of achieving simultaneously: (1) support for sophisticated queries (e.g., Facebook's TAO API); (2) scalability with increasing graph data sizes; (3) query interactivity. ZipG achieves all of the above using a fundamentally new approach: executing queries directly on a compressed representation of the input graph.

  I am the lead author on the ZipG paper, which is under submission. Advised by Rachit Agarwal and Ion Stoica.

- **Planck.** Contributed to *Planck*, a performance prediction system for analytics jobs, by *(i)* investigating prediction models based on information collected from Spark job stages/tasks, and *(ii)* implementing and conducting large-scale evaluations on EC2. Paper to appear in NSDI 2016.

System building:

- **SparkR**. With Shivaram Venkataraman, I co-authored the initial prototype of *SparkR*, the R frontend for Spark. Grew its traction by regularly reviewing pull requests and engaging users for ∼1 year. The project was merged into official Apache Spark in early 2015 and is now widely used.

- **KeystoneML.** *KeystoneML* is a machine learning pipeline library on top of Spark. I contributed many of its natural language processing (NLP) components: *(i)* the distributed and highly scalable Stupid Backoff language model, based on a Google paper (Brants 07), and *(ii)* several optimized NLP building blocks, e.g. a space-efficient n-gram featurizer.

- Other AMPLab contributions: ml-matrix, SparkR's AMPCamp training exercises.

Publications

Zongheng Yang, Rachit Agarwal, Evan Ye, Anurag Khandelwal, Ion Stoica. ZipG: Serving Queries on Compressed Graphs. *In submission to SIGMOD 2016 (research track).*

Shivaram Venkataraman, **Zongheng Yang**, Michael J. Franklin, Benjamin Recht, Ion Stoica. Planck: Efficient Performance Prediction for Large-Scale Advanced Analytics. *Accepted to USENIX Symposium on Network System Design and Implementation (NSDI), 2016.*

Shivaram Venkataraman, **Zongheng Yang**, Davies Liu, Eric Liang, Xiangrui Meng, Reynold Xin, Ali Ghodsi, Michael J. Franklin, Ion Stoica, Matei Zaharia. SparkR: Scaling R Programs with Spark. *In submission to SIGMOD 2016 (industrial track).*

(Poster) Evan R. Sparks, Shivaram Venkataraman, Tomer Kaftan, **Zongheng Yang**, Vaishaal Shankar, Michael J. Franklin, Benjamin Recht. KeystoneML: Simplifying end-to-end Machine Learning at Scale. *Poster at AMPLab retreat, May 2015.*

Talks

AMPCamp@China Intel Shanghai Campus, China. May 2015.
A Sneak Peek at Recent Updates from AMPLab.

AMPCamp 5 Berkeley, CA, U.S. November 2014.
(Demo part; with Shivaram Venkataraman) *SparkR: Enabling Interactive Data Science at Scale*.

Spark Summit 2014 San Francisco, CA, U.S. July 2014.
SparkR: Interactive R programs at Scale.

Graduate-level Coursework

CS262A: Systems seminar (Prof. John Kubiatowicz)
Project: Probing Distributed Linear Algebra Operators in the Cloud. Companion poster.

CS263: Programming Language Theory (Prof. George Necula)
Survey: Making Dependent Types Practical.

CS288: Natural Language Processing (Prof. Dan Klein)
Projects: *(i)* LM, *(ii)* Speech Recognition, *(iii)* Parsing, *(iv)* Reranking, *(v)* Word Alignment.

CS294: Big Data seminar (in progress; Prof. Ion Stoica)
Presented and led discussion on TAO, Facebook's social graph store.

Industry Experience

Software Engineering Intern, Databricks May–Aug 2014

- Optimized Spark SQL's performance by implementing new physical plan algorithms (e.g. broadcast join) and low-level profiling (YourKit, dstat, byte code inspection). Co-authored a Databricks blog post on performance gains: bit.ly/1rDfk6g.
- Contributed 10+ patches to Spark: implemented new AST nodes in Spark SQL's parser and query planner; prototyped a cost-based optimization framework, by using the Hadoop API and Hive statistics to fetch table sizes.

Software Engineering Intern, Twitter June–Aug 2013

- Integrated approximate data structures (HyperLogLog, CountMinSketch) into TSAR, Twitter's event processing framework, by implementing corresponding execution plans and proper serialization. This work enabled cardinality approximation and heavy hitter counting in the Twitter Analytics service on petabyte-scale data.

Honors and Awards

EECS Honors Degree Program (1.8% of all declared EECS/CS majors), UC Berkeley

Dean's Honors (Fall 2012, Spring 2013, Fall 2013), College of Letters and Science, UC Berkeley

Skills

Languages	Scala, Java, C/C++, Python, R, Haskell, OCaml, Scheme, numpy, bash, Javascript.
Technologies	Spark, Hadoop (MapReduce, HDFS, Hive), AWS (EC2, S3), JVM, Thrift, Git, Linux.

Publicly Released Software

- **SparkR** (R, Scala, Spark): github.com/amplab-extras/SparkR-pkg hosts the prototype we developed; it is now part of official Spark.
- **KeystoneML** (Scala, Spark, Distributed NLP): github.com/amplab/keystone/pull/71 contains the bulk of my NLP contributions.

The highlighted texts point to corresponding files, which can be found at zongheng.me.