DB TSAI

E-mail: dbtsai@dbtsai.com *Web: www.dbtsai.com* *LinkedIn: www.linkedin.com/in/dbtsai* *GitHub: www.github.com/dbtsai*

Summary

- I specialize in big data machine learning algorithms and infrastructure with strong background in theoretical statistics and mathematics.
- I lead and architect the Netflix offline recommendation training pipeline using Apache Spark, and build the common shared machine learning library used in several teams at Netflix. The work accelerates the innovation through a series of A/B tests to provide best experience for our members.
- I am devoted to sharing my knowledge through talks to help and learn from others. It's all about our human nature of wanting to connect and collaborate with others. This helps my company to build technical brand for recruiting.
- I help to build the machine learning infrastructure team at Netflix through giving talks at conference, and the connections in the open source communities.
- I have been actively involved with the open source Apache Spark development as a committer. I contribute code and designs of various distributed machine learning algorithms to the project, and also mentor contributors by leading directions, code review, and merging their contributions.

Experience

- **Apache Spark** — A fast and general engine for large-scale data processing
  Project Management Committee (PMC) Member *June 2017 to current*
  Committer *May 2015 to current*
  - My contributions, https://github.com/apache/spark/commits/master?author=dbtsai
  - Implement new features and designs such as L-BFGS, and Multinomial / Binomial Logistic Regression, etc.
  - Conduct code review for other contributors, and guide them until the code is merged.
  - Fix various bugs, write documentation and perform performance optimization.
- **Netflix, Los Gatos, CA** — A Leading Provider of Internet Streaming Media Available Worldwide
  Research Engineer *Apr. 2015 to current*
  - Lead and architect the personalized recommendation pipelines and machine learning infrastructure using Apache Spark.
  - Architect and implement **Distributed Time Travel Machine for Feature Generation using Apache Spark**, which enables our researchers to quickly try ideas with new features on historical data such that running offline experiments and transitioning to online A/B tests is seamless. This framework reduces the time to bring an offline experiments to online A/B tests from months to weeks, and significantly removes the offline/online discrepancy because of sharing the feature generation logics between offline/online. U.S. Patent filed February 2016. Patent Pending.
  - Implement categorical feature learner in Netflix's in-house GBDT (Gradient Boosting Decision Tree) implementation as part of the global algorithm effort to incorporate the country and language categorical signals for global launch.
  - Implement Weighted Logistic Regression in open source Apache Spark ML which is used in Netflix's personalized page algorithms for constructing the rows in the homepage.
  - Work closely with Apache Spark community to merge our changes, and implement new features for our needs.
- **SF Big Analytics Meetup, CA** — Covering the Topics of Data Infrastructure, Data Pipelines and Analytics
  Founder and Co-Organizer *Jan. 2015 to current*
  - https://www.meetup.com/SF-Big-Analytics/
  - Organized 19 talks in the first year.
  - The membership has increased from 0 to 3098 members in 10 months.
- **SF Machine Learning Meetup, CA** — People with Shared Interests of Machine Learning and Big Data
  Co-Organizer *Jun. 2013 to July 2015*
  - http://www.meetup.com/sfmachinelearning/

  - Hold the meetup monthly, and invited famous speakers in industry and academic to give talks.

- **Alpine Data Labs, San Francisco, CA** — The Leader in Data Science for Big Data
  Machine Learning Engineer *Apr. 2013 to Aug. 2014*
  - Developed scalable Multinomial Logistic Regression and Linear Regression with elastic-net regularization which linearly combines the L1 and L2 penalties in Apache Spark. Implemented OWLQN for L1/L2 regularized optimization.
  - Developed scalable algorithms such as Decision Tree, Variable Selection based on Information Gain, exact one-pass Linear Regression with L2 penalty, and PCA in Hadoop MapReduce.
  - Migrated build infrastructure from ANT to SBT for better third party library dependency management using the Maven central repository, better integration with Jenkins for continuous integration, better developement/debuging experience for developers, and easier release build.
- **KeeKa, StartX 2012 Summer, Stanford, CA** — A Social Network Connecting People through Fashion
  Co-founder and Software Engineer *Jan. 2012 to Mar. 2013*
  - Planned the strategies and invented a disruptive product.
  - Designed the architecture of the website, including deployment, front-end, and back-end systems.
  - Coordinated the designer, front-end team, and back-end team and performed the code review to ensure reliability, effectiveness, progress, and productivity.

Selected Publications (https://www.dbtsai.com/publications/)

- **MLlib: Machine Learning in Apache Spark**,
  Xiangrui Meng, Joseph Bradley, Burak Yavuz, Evan Sparks, Shivaram Venkataraman, Davies Liu, Jeremy Freeman, **DB Tsai**, Manish Amde, Sean Owen, Doris Xin, Reynold Xin, Michael J. Franklin, Reza Zadeh, Matei Zaharia, and Ameet Talwalkar
  Journal of Machine Learning Research, Vol 17 (34) pp. 1-7 *Apr. 2016*
- **Distributed Time Travel for Feature Generation**,
  Hossein Taghavi, Prasanna Padmanabhan, **DB Tsai**, Faisal Zakaria Siddiqi, and Justin Basilico
  US Patent Pending *Feb. 2016*
- **Quantum Zeno and anti-Zeno effect of nanomechanical resonator measured by a point contact**,
  Po-Wen Chen, **DB Tsai**, and Philip Bennett
  Physical Review B 81, 115307 *Mar. 2010*
- **Optimal control of the silicon-based donor-electron-spin quantum computing**,
  DB Tsai, Po-Wen Chen, and Hsi-Sheng Goan
  Physical Review A 79, 060306(R) *June 2009*
- **A Guide to Having Fun with the Next Generation Linux, Ubuntu by S.-W. Lee and D.-B. Tsai**, Taipei
  ISBN: 9867199979, GrandTech Press. Among the top 5 best-selling computer-science books from Nov. 2006 to Jan. 2007 in the Chinese book market in Taiwan. *Sept. 2006*

Selected Talks (https://www.dbtsai.com/talks/)

- **Discussing Scaling Machine Learning at ScaledML as One of the Panelists**, Stanford, CA
  Scaled Machine Learning Conference *Mar. 25th, 2017*
- **Netflix's Recommendation ML Pipeline using Apache Spark**, Boston, MA
  Spark Summit East *Feb. 8th, 2017*
- **Distributed Time Travel for Feature Generation**, New York, NY
  Spark Summit East *Feb. 17th, 2016*
- **Large-Scale Elastic-Net Regularized Generalized Linear Models**, San Francisco, CA
  Spark Summit *June 15th, 2015*
- **Lambda Architecture with Apache Spark**, Galvanize, San Francisco, CA
  Next.ML Conference *Jan. 17th, 2015*
- **Large-Scale Machine Learning with Apache Spark**, Moscone Center, San Francisco, CA
  Internet of Things Conference *Oct. 20th, 2014*
- **Multinomial Logistic Regression with Apache Spark**, Alpine Data Labs, San Francisco, CA
  SF Machine Learning Meetup *May 1st, 2014*

Education

- **Stanford University**, California, U.S.A.
  ABD in Applied Physics Ph.D. program *Sept. 2010 to June 2012*
  M.S. in Electrical Engineering *Sept. 2010 to June 2012*
- **National Taiwan University**, Taipei, Taiwan
  M.S. in Physics *Sept. 2006 to July 2008*
- **National Cheng Kung University**, Tainan, Taiwan
  B.S. in Physics *Sept. 2002 to June 2006*