Joseph E. Gonzalez, Postdoc UC Berkeley Curriculum Vitae

Email: jegonzal@eecs.berkeley.edu Last updated: November 9th, 2014

Vision: *By exploiting properties of machine learning algorithms to develop scalable systems, and leveraging advances in system design to develop scalable machine learning algorithms, we can solve important real-world problems, address the greater machine learning life cycle, and define the abstractions that will drive future research in these complementary fields.*

I am the co-founder of Dato Inc. (formerly GraphLab Inc.) and in January 2016, I will be joining the EECS department at UC Berkeley as an assistant professor.

Research Summary

My research addresses the challenges of designing and building large-scale machine learning algorithms and systems. In particular, my thesis work introduced algorithms, abstractions, and systems for scalable inference in graphical models and played a key role in defining the space of parallel inference algorithms and graph processing systems. The resulting GraphLab and PowerGraph systems have defined state-of-the-art performance in tasks ranging from predicting ad preferences in social networks to solving complex protein modeling tasks and led to the founding of GraphLab Inc. (now Dato.com) which has successfully commercialized my research. As a postdoc in the UC Berkeley AMPLab, I worked on unifying graph-processing systems with general purpose dataflow frameworks and introducing transaction processing primitives into the design of machine learning algorithms.

For more information visit `http://eecs.berkeley.edu/~jegonzal`.

Education

Ph.D., Machine Learning: December 2012. Machine Learning Department, School of Computer Science at Carnegie Mellon University.

Title: *"Parallel and Distributed Algorithms and Systems for Probabilistic Reasoning."*
Thesis Advisor: Carlos Guestrin

M.S., Machine Learning: December 2009. Machine Learning Department, School of Computer Science at Carnegie Mellon University.

B.S. with Honors, Computer Science: June 2006. California Institute of Technology.

Awards

- **Nominated for ACM Dissertation Award [2013]:** My thesis was nominated by CMU for the ACM Dissertation Award.
- **AT&T Labs Fellowship (2007):** Graduate research stipend for academic achievement as an underrepresented minority.
- **NSF Graduate Research Fellowship (2007):** Graduate research stipend for 3 years.
- **NASA Space Act Award (2005):** Awarded for a sizeable contribution to space exploration.

- **NASA Inventions and Contributions Board Award (2005):** Awarded for the development of an innovative new technology that has made a contribution to space exploration.
- **Upper Class Merit Award (Twice) (2004,2005):** full tuition for academic excellence, research, and faculty recognition.
- **Presidential Award (2002-2006):** I was awarded tuition for research and academic achievements.

Publications

[1] Veronika Strnadova-Neeley, Aydin Buluc, Jarrod Chapman, John Gilbert, Joseph Gonzalez, and Leonid Oliker. Efficient data reduction for large-scale genetic mapping. In *ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (BCB '15)*, 2015.

[2] Neeraja J. Yadwadkar, Bharath Hariharan, Joseph Gonzalez, and Randy Katz. Faster jobs in distributed data processing using multi-task learning. In *SIAM International Conference on Data Mining (SDM '15)*, 2015.

[3] Daniel Crankshaw, Peter Bailis, Joseph E. Gonzalez, Haoyuan Li, Zhao Zhang, Michael J. Franklin, Ali Ghodsi, and Michael I. Jordan. The missing piece in complex analytics: Low latency, scalable model management and serving with velox. In *Conference on Innovative Data Systems Research (CIDR '15)*, 2015.

[4] Xinghao Pan, Stefanie Jegelka, Joseph E. Gonzalez, Joseph K. Bradley, and Michael I. Jordan. Parallel double greedy submodular maximization. In *Neural Information Processing Systems (NIPS '14)*, 2014.

[5] Joseph E. Gonzalez, Reynold S. Xin, Ankur Dave, Daniel Crankshaw, Michael J. Franklin, and Ion Stoica. Graphx: Graph processing in a distributed dataflow framework. In *11th USENIX Symposium on Operating Systems Design and Implementation (OSDI 14)*, pages 599–613, 2014.

[6] Veronika Strnadova, Aydin Buluc, Leonid Oliker, Joseph Gonzalez, Stefanie Jegelka, Jarrod Chapman, and John Gilbert. Fast clustering methods for genetic mapping in plants. In *16th SIAM Conference on Parallel Processing for Scientific Computing*, 2014.

[7] David Bader, Aydın Buluç, John Gilbert, Joseph Gonzalez, Jeremy Kepner, and Timothy Mattson. The graph blas effort and its implications for exascale. In *SIAM Workshop on Exascale Applied Mathematics Challenges and Opportunities (EX14)*, 2014.

[8] Xinghao Pan, Joseph E. Gonzalez, Stefanie Jegelka, Tamara Broderick, and Michael I. Jordan. Optimistic concurrency control for distributed unsupervised learning. In *NIPS '13*, 2013.

[9] Evan Sparks, Ameet Talwalkar, Virginia Smith, Xinghao Pan, Joseph Gonzalez, Tim Kraska, Michael I Jordan, and Michael J Franklin. MLI: An api for distributed machine learning. In *International Conference on Data Mining (ICDM)*. IEEE, December 2013.

[10] Reynold Xin, Joseph E. Gonzalez, Michael Franklin, and Ion Stoica. Graphx: A resilient distributed graph system on spark. In *SIGMOD Grades Workshop*, 2013.

[11] Joseph E. Gonzalez, Yucheng Low, Haijie Gu, Danny Bickson, and Carlos Guestrin. Powergraph: Distributed graph-parallel computation on natural graphs. In *OSDI '12*, 2012.

[12] Yucheng Low, Joseph Gonzalez, Aapo Kyrola, Danny Bickson, Carlos Guestrin, and Joseph M. Hellerstein. Distributed graphlab: A framework for machine learning and data mining in the cloud. In *Proceedings of Very Large Data Bases (PVLDB)*, August 2012.

[13] Amr Ahmed, Mohamed Aly, Joseph Gonzalez, Shravan Narayanamurthy, and Alex Smola. Scalable inference in latent variable models. In *Conference on Web Search and Data Mining (WSDM)*, 2012.

[14] Joseph Gonzalez, Yucheng Low, Arthur Gretton, and Carlos Guestrin. Parallel gibbs sampling: From colored fields to thin junction trees. In *Artificial Intelligence and Statistics (AISTATS)*, May 2011.

[15] Joseph Gonzalez, Yucheng Low, and Carlos Guestrin. *Scaling Up Machine Learning*, chapter Parallel Inference on Large Factor Graphs. Cambridge U. Press, 2010.

[16] Y. Low, J. Gonzalez, A. Kyrola, D. Bickson, C. Guestrin, and J. M. Hellerstein. Graphlab: A new parallel framework for machine learning. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2010.

[17] J. Gonzalez, Y. Low, C. Guestrin, and D. O'Hallaron. Distributed parallel inference on large factor graphs. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, July 2009.

[18] J. Gonzalez, Y. Low, and C. Guestrin. Residual splash for optimally parallelizing belief propagation. In *Artificial Intelligence and Statistics (AISTATS)*, April 2009.

[19] R. Chamberlain, J. Gonzalez, G. Gutt, and E. Taylor. New line of sight algorithm renders superlative TINs superfluous. Technical Report D-32587, JPL, 2005.

Publications Under Review

- Peter Bailis, Joseph E. Gonzalez, Ali Ghodsi, Michael J. Franklin, Joseph M. Hellerstein, Michael I. Jordan, and Ion Stoica. Asynchronous Complex Analytics in a Distributed Dataflow Architecture. *SIGMOD'16*. 2016

Selected Invited Talks

- **[2014] OSDI Conference** *GraphX: Graph Processing in a Distributed Dataflow Framework*
- **[2014] Annual meeting of the International Society for Bayesian Analysis (ISBA)** *Concurrency Control For Scalable Bayesian Inference.*
- **[2014] Tutorial at the International Conference for Machine Learning (ICML)** *Emerging Systems for Large-Scale Machine Learning.*
- **[2014] Session on Graph Algorithms Building Blocks at the International Parallel and Distributed Processing Systems (IPDPS)** *GraphX: Unifying Table and Graph Analytics.*
- **[2014] NetApp ATG University Day** *Large Scale Graph Analytics: Applications and Systems*
- **[2014] Keynote Speaker: Workshop on Big Graph Mining at the International World Wide Web Conference (WWW)** *From Graphs to Tables: The Design of Scalable Systems for Graph Analytics.*
- **[2013] SIAM CSE'13 Minisymposium Frontiers in Large-Scale Graph Analysis** Large-Scale Graph-Structured Machine Learning: GraphLab in the Cloud and GraphChi in your PC
- **[2012] OSDI Conference** *PowerGraph: Distributed Graph-Parallel Computation on Natural Graphs*
- **[2011] IDGA Data Center Consolidation Summit** *GraphLab: A New Parallel Framework for Machine Learning*

- **[2010] DARPA Future Ideas Symposium** Invited speaker at the DARPA future ideas symposium.
- **[2009] UAI Conference.** *Distributed Parallel Inference on Large Factor Graphs*

Teaching

Advising Graduate Students As a post-doc at UC Berkeley I was given the opportunity help advise a group of exceptional graduate students as they entered graduate school:

- **Dan Crankshaw [2013-2014]** Studying the design of systems for managing the life cycle of machine learning models.
- **Ankur Dave [2013-2014]** Studying the design of graph processing systems.
- **Orianna DeMasi [2013-2014]** Studying active learning algorithms for the tuning of BLAS libraries and computer architectures.
- **Josh Rosen [2013]** Worked on a range of projects in the context of dataflow programming and optimizations.
- **Xinghao Pan [2013 - 2014]** Studying the application of transactional models to a range of tasks in machine learning.

Guest Lectures Throughout my postdoc and even during graduate school I had the opportunity to guest lecture in numerous classes:

- **[2014] Berkeley Graduate Database Systems (CS 286)** Graph Processing Systems
- **[2014] Berkeley Graduate Statistical Learning Theory (CS 281A)** Linear Regression and the Bias Variance Tradeoff.
- **[2014] Berkeley Introduction to Data Science (CS 194-16)** Introduction to Graph Analytic
- **[2012] Berkeley class on Analyzing Big Data with Twitter.** Big Learning with Graphs.
- **[2012] CMU Machine Learning with Large Datasets class.** Large-scale Machine Learning on Graph Structured Data.

Teach Assistant: Both as a graduate student and as an undergraduate I was given the opportunity to help teach classes:

- **[2009] Teaching Assitant** for the CMU Masters Machine Learning class. I designed problem sets and exams, gave recitation lectures, and mentored student projects. This was an ambitious class. *(Geoff Gordon: guestrin@cs.cmu.edu)*
- **[2007] Teaching Assitant** for the CMU Graduate Machine Learning class. I designed problem sets and exams, gave recitation lectures, and mentored student projects. *(Carlos Guestrin: guestrin@cs.cmu.edu)*
- **[2004,2005] Head Teaching Assistant** for the Caltech Introductory Computer Science Course (CS2): Redesigned introductory computer science. *(Al Barr: barr@cs.caltech.edu)*

Workshop Organizer

- **[2014] DIMACS Workshop Organizer** I organized the DIMACS workshop on the Systems and Analytics of Big Data (`http://dimacs.rutgers.edu/Workshops/Analytics/`)
- **[2013] NIPS Workshop Organizer** I helped organize the third annual NIPS "Big Learning: Algorithms, Systems, and Tools" workshop. (`http://biglearn.org`)
- **[2012] NIPS Workshop Organizer** I helped organize the second annual NIPS "Big Learning: Algorithms, Systems, and Tools" workshop. (`http://biglearn.org`)
- **[2011] NIPS Workshop Organizer** I organized and led the workshop entitled "Big Learning: Algorithms, Systems, and Tools for Learning at Scale" For more information visit the workshop website `http://biglearn.org`
- **[2009] NIPS Workshop Organizer** I organized and led the first NIPS BigLearn workshop entitled "Large-Scale Machine Learning: Parallelism and Massive Datasets." For more information visit the workshop website `http://www.select.cs.cmu.edu/meetings/biglearn09`

Grant Writing and Funding Experience

- [**2015**] NSF Panelist
- [**2013**] working with Carlos Guestrin, I raised 6.75M in series A funding for the GraphLab Inc. startup.
- [**2010**] Applied for and was awarded a grant to have early access to the Intel Single-chip Cloud Computer (SCC) as part of the Many-core Applications Research Community.
- **[2008 - 2009] Helped Lead a DARPA Interdisciplinary Sciences and Technology Study (ISAT) Group** to investigate the future of parallel machine learning from an interdisciplinary perspective I also participated in the final Woodshole annual ISAT meeting to prepare a proposal for the DARPA director.
- [**2008**] Applied for and was awarded funding for (BAA 08-34) "Machine Learning and AI in the context of Multicore and Cluster Computing."

Reviewing

- **[2014] OSDI Extended Review Committee**
- **[2014] HotCloud Review Committee**
- **[2014] ICML**
- **[2013] Transactions on Pattern Analysis and Machine Intelligence**
- **[2013] NIPS**
- **[2013] Super Computing**
- **[2012] Parallel Computing**
- **[2010] ICML**
- **[2009] JMLR**
- **[2007] JMLR**
- **[2007] IPSN**

Industry Experience

- **GraphLab Inc. (2013-Present):** I co-founded the GraphLab startup to commercialize my research and help lead the initial technology roadmap. I now consult part-time on technology direction.
- **Yahoo! Research (2011):** Developed the next generation of the GraphLab abstraction to enable large-scale machine learning on natural graphs derived from social media and web-content. *(Alex Smola: smola@yahoo-inc.com)*
- **AT&T Labs Research (2007):** Developed models for statistically assessing DSL quality from limited noisy data. *(Steven Phillips: phillips@research.att.com)*
- **Intern at ADAPT (2006)** Worked on an automated AdWords auction agent. I developed and implemented models for assessing word value. *(Alex Bcker: alex@caltech.edu)*
- **Microsoft Developer Internship (2005):** Worked with MSN Search team developing techniques to use behavioral information to identity search spam. *(Greg Hullender: greghull@windows.microsoft.com)*
- **Caltech Research Fellowship (2004):** Developed a new query-less search technology that uses prior reading interests to identify novel documents. *(Alex Bcker: alex@caltech.edu)*
- **NASA Jet Propulsion Labs Fellowship (2003):** Developed a new algorithm for efficiently evaluating line-of-sight on digital elevation maps at JPL. *(Robert Chamberlain: rgc@jpl.nasa.gov)*

Publicly Released Software

- **GraphX (Scala)** GraphX is the graph computation framework built into the widely adopted Apache Spark open-source project. `https://spark.apache.org/graphx/`
- **GraphLab/PowerGraph (C++)** A sophisticated API for building parallel and distributed machine learning algorithms on top of multicore and cloud architectures. GraphLab generalizes the MapReduce abstraction to support iterative asynchronous computation on graph structured dependent data. `http://graphlab.org`
- **Distributed SplashBP (C++)** This library implements the SplashBP algorithm for factor graph inference in the distributed setting using MPI. `http://www.select.cs.cmu.edu/code/mpi_splash.tar.gz`
- **Shared Memory SplashBP (C++)** This library implements the SplashBP algorithm for Markov random fields inference. `http://www.select.cs.cmu.edu/code/parallelmrf_src.tar.gz`

References

Carlos Guestrin (Thesis Advisor)
Amazon Professor of Machine Learning
University of Washington
CEO, GraphLab Inc.
`guestrin@cs.washington.edu`

Michael J. Franklin (Postdoc PI)
Thomas M. Siebel Professor of Computer Science
Chair of the Computer Science Division of EECS, UC Berkeley
`franklin@cs.berkeley.edu`

Ion Stoica
Professor of Computer Science, UC Berkeley
CEO, Databricks
`istoica@cs.berkeley.edu`

Michael I. Jordan
Pehong Chen Distinguished Professor
Department of EECS and Statistics, UC Berkeley
`jordan@cs.berkeley.edu`

Alexander J. Smola
Researcher, Google
Professor, Carnegie Mellon University
`alex@smola.org`