Sam Ritchie sam@mentat.org
https://samritchie.io Mobile: +1.703.863.8561

Education

Princeton University Princeton, NJ
BSE in Mechanical and Aerospace Engineering *Sep 2005 - Jun 2009*

Recent Work Experience

Mentat Collective Boulder, CO
Principal Researcher *Jan 2021 - Present*

- **Emmy Computer Algebra System**: Emmy is a Clojure(Script) computer algebra system designed for math and physics investigation. The library ships with a collection of viewers that make it possible to author and publish interactive, executable articles, posts and books with embedded simulations.

Alphabet X Mountain View, CA
Research Engineer, Blueshift *Oct 2019 - Jan 2021*

- **Caliban**: Designed and implemented Caliban, a tool that makes it trivial for researchers to scale their experiments from a local script to running code on thousands of GPU-enabled Cloud machines. Caliban has reduced the time it takes a researcher to become productive from 3 weeks to 1 hour, a 2-order-of-magnitude speedup.
- **Blueshift Research Archive**: Designed and implemented UV, a metric reporting framework that forms the core of a suite of tools designed to solve the problem of collecting vast amounts of data during thousands of machine learning experiments and synthesizing them down into world-class research papers. Led the design of the rest of this suite, called the Blueshift Research Archive.

Stripe, Inc San Francisco, CA
Senior Software Engineer, Machine Learning Infrastructure *Jan 2016 - June 2018*

- **Semblance**: Extended Summingbird into a system that accepts a user's declarative description of a machine learning feature and computes it in both batch and realtime with sub-100 millisecond latency for use in training and scoring machine learning models.
- **Machine Learning Workflow Manager**: Designed Stripe's core machine learning feature generation, model training, deployment and evaluation workflow and implemented many of the components. This system currently powers all machine learning applications at Stripe (fraud detection via Radar, customer risk evaluation, etc).
- **Machine Learning Explanations**: Contributed to the development of a state-of-the-art technique to generate explanations of the predictions of Stripe's "black box" fraud models with millisecond latency. Presented this work at Strange Loop and on the TWiML&AI Podcast.

PaddleGuru Boulder, CO
Founder, Lead Developer *Jan 2014 - Oct 2016*

- **PaddleGuru**: PaddleGuru.com is an athletic timing, registration and payment processing platform used by thousands of athletic competitions. The system is full-stack Clojure, with a React-based Clojurescript front-end that communicates via websocket with a Clojure backend running on Dokku. PaddleGuru may be the techiest product in this space, for better or for worse.
- **Open Source**: Released a number of successful open-source front end libraries, including Om-Bootstrap and wrappers around Stripe, Mandrill and a number of other web development services.

Twitter San Francisco, CA
Senior Software Engineer, Platform Engineering *Dec 2011 - Nov 2013*

- **Summingbird**: Developed and released Summingbird, a MapReduce platform written in Scala that powers the majority of Twitter's realtime analytics workflows. I was profiled in Wired magazine for this work. Summingbird is open-source and is one of the most popular streaming MapReduce platforms (2k+ GitHub stars, dozens of corporate users).
- **Streaming Infrastructure**: Deployed and maintained production code on dozens of Hadoop and Storm based applications at the core of Twitter's analytics products, internal and external.
- **Revenue Team**: Worked on and deployed Twitter's first generation advertiser analytics dashboard. A later Summingbird implementation reduced the number of machines required to run this system from 80 to 5, and the number of humans from 8 to 1.

REDD Metrics Philadelphia, PA

Lead Developer *June 2010 - Sep 2011*

- **Global Forest Watch**: Designed and built a globe-scale deforestation monitoring and prediction system (FORMA) with the support of the World Bank and the World Resources Institute. FORMA is the foundation for WRI's Global Forest Watch product. It trains on, and applies classification algorithms to, terabytes of NASA MODIS satellite data using a Clojure workflow built on Apache Hadoop. FORMA's scale and stability was unprecedented in the deforestation monitoring community when we launched in 2011.
- **Open Source**: During this work I became a committer or maintainer of many of the leading Hadoop data processing libraries, including Cascalog, Cascading, Scalding and Pallet. I released a cloud-agnostic provisioning tool called Pallet-Hadoop that deploys and manages the hundreds of machines required to process satellite data at scale.

ThinkFun, Inc Alexandria, VA

Digital Projects Director, Lead Developer *Apr 2009 - Sep 2011*

- **Digital Games**: Developed an interactive suite of successful ($150K+ revenue/yr) iPhone and Android applications based on ThinkFun's mechanical brainteaser puzzle line; Rush Hour alone has amassed millions of downloads. Worked with schools around the country to develop a creative problem-solving curriculum called "Brainlab" based on these brainteaser puzzles.

Notable Projects

- **Summingbird**: Summingbird was the first Lambda architecture implementation. The library lets you write MapReduce programs that look like native Scala or Java collection transformations; Summingbird generates code from these programs that can run on Hadoop, Storm, or Spark alone or in a hybrid batch/realtime mode that offers your application very attractive fault-tolerance properties. This work resulted in a paper accepted at VLDB 2014.
- **Algebird**: Algebird is Scala's abstract algebra library. Algebird was designed with streaming aggregations in mind, and implements a number of data structures and combinators that are useful in a streaming environment. The Monoid, for example, is a core concept of Summingbird. Algebird opened the door for large-scale deployments at Twitter of approximate data structures from the research world, notably the HyperLogLog and CountMinSketch.
- **Bijection**: Bijection is an implementation of invertible functions in Scala; the library allows for type-safe conversion between isomorphic data types spread across various Scala libraries. The interfaces in Bijection are foundational to Twitter and Stripe's Scala codebase, and to the other Scala projects referenced here.
- **Chill**: Chill provides a number of enhancements to the Kryo JVM serialization library; notably, serializers for all Scala primitives and collection types, and plugins that make it easy to use Kryo in Hadoop and Storm jobs. Scalding, Cascalog, Spark, Summingbird and many other projects use Chill to manage serialization across their various distributed system implementations.
- **Cascalog**: Cascalog is a Datalog implementation in Clojure that compiles logic queries into Hadoop jobs. Cascalog powers dozens of companies, and powered Twitter's analytics dashboard until Summingbird deposed it. I've maintained Cascalog since late 2011 and authored many core features and modules, including midje-cascalog and cascalog-contrib.
- **RV-10 Airplane Build**: I'm 95% through a build of a Vans Aircraft RV-10 4 seater airplane. The RV-10 is a single engine, fuel-injected aircraft with a cruising speed of 200mph, a service ceiling of 20,000" and a range of 1000 miles. This project is my connection to my old life as an aerospace engineer.
- **More Projects**: You can find long-form descriptions of more of my work at https://samritchie.io/projects.

Technical Skills

- **Languages**: Fluent in Scheme, Clojure, ClojureScript, Python, Scala and Java, less so with JS, Haskell and R.
- **Open Source**: I'm a committer on Cascalog, Cascading, Scalding, Summingbird, and have developed lots of Scala and Clojure.

Talks and Publications

1. **The REPL Podcast**: *Executable Textbooks with Sam Ritchie,* 13 Jan 2023
2. **re:Clojure 2020**: *Functional Physics & the Preservation of Society,* 4 Dec 2020
3. **CoRecursive Podcast**: *Portal Abstractions with Sam Ritchie,* 4 Apr 2020
4. **This Week in ML & AI Podcast**: *Exploring Black Box Predictions,* 25 Nov 2017
5. **Strange Loop 2017**: *Just So Stories for AI: Explaining Black Box Predictions* (slides), 28 Sep 2017

6. **VLDB 2014**: *Summingbird: A Framework for Integrating Batch and Online MapReduce Computations,* Sep 2014
7. **Data Day Texas**: *The Road to Summingbird: Stream Processing at Every Scale,* Jan 2014
8. **Wired Magazine**: *How the Nephew of Computer Science Royalty Remade Twitter,* Nov 2013
9. **Clojure/Conj 2013**: *Cascalog 2.0: Datalog in Realtime* (slides), Nov 2013
10. **PNW Scala 2013**: *Taking Hadoop Realtime with Summingbird,* Sep 2013
11. **Boston Storm Users**: *Summingbird, Scala & Storm* (slides), Sep 2013
12. **CUFP 2013**: *Realtime MapReduce at Twitter,* Sep 2013
13. **Bay Area Storm Users**: *Introduction to Summingbird* (w/ Oscar Boykin), Sep 2013
14. **GigaOm**: *Twitter Open Sources Summingbird,* Sep 2013
15. **Twitter Eng Blog**: *Streaming MapReduce with Summingbird,* Sep 2013
16. **AK Data Summit**: *Summingbird: Streaming MapReduce at Twitter* (slides), Jun 2013