Saptarshi Guha — June 2014

Contact Information

3201 23rd St, Apt 203
San Francisco, CA
USA, 94110

+1 765 543 8189
saptarshi.guha@gmail.com
http://people.mozilla.org/~sguha/

Work Experience

Mozilla, Mountain View, CA, USA — **March 2010 — current**

Senior Data Scientist

Job responsibilities include framing statistical hypotheses for experiments conducted by the marketing/engagement and engineering teams.

- Analyzed results of engineering experiments e.g. reduced memory usage(the MemShrink project) and reduced garbage collection times via the Firefox Telemetry subsystem. The results of the analyses statistically validated the improvements in algorithms.
- Assisted in redesigning automated testing (the TALOS automated testing system). The objective was to reduce false positives consequently making the testing alerts more informative.
- Comparing Firefox Nightly performance across builds using box plots, shift plots, QQ plots and approximate distributions. The information is taken from hundreds of Telemetry measurements to build dashboards for engineers to monitor.
- Used time series model to detect errors in log file collection. Proposed to use this to send alert messages when errors in log file collection occur.
- Designed the requirements of the Firefox Health Report data collection. This data set collects longitudinal data on Firefox profiles.
  - Model changes in performance and usage to understand long term effects of new features on profiles.
  - Build retention and activity models to answer questions such as 'what makes a profile keep using Firefox' and 'how often is the browser used'.
  - Growth models that forecast adoption rates for daily builds of Firefox. With this information engineers know how long they need to wait to receive Telemetry measurements.
  - Use activity models to inform sample selection in Telemetry Experiments.
- Building product indices that measure performance, profile involvement and customization. These indices are used as a 'state of Firefox' measure and disseminated across the organization.

Revolution Analytics, Palo Alto, CA, USA — **September 2010 — March 2010**

Solutions Architect

- Designed a statistical tool to detect out of band (both shocks and systemic changes) values in network metrics across a data center.
- Designed an R integrator for Hbase.

GE Capital International Services, Bangalore, India — **2002 — 2004**

Business Analyst

- Developed tracking systems to monitor effectiveness of marketing campaigns
- Constructed and implemented experimental designs for advertising campaigns
- Built statistical models to market JC Penney credit cards and promotional offers

Education

Purdue University, West Lafayette, Indiana, USA — **August 2004 — August 2010**

Doctor of Philosophy

- Dissertation Title: ***Statistical Programming Environments for Large Data Sets.***
- Advisor: Prof. William Cleveland

Indian Statistical Institute, New Delhi, India — **August 2000 — June 2002**

Masters in Statistics

- Specialization in Mathematical Statistics and Probability.

Presidency College, Kolkata, India — **August 1997 — July 2000**

Bachelors in Statistics

Presentations

Terra + R = RTerra: Using Terra for Fast R Extensions, Invited talk at Bay Area R Users Group, 2013 (see http://bit.ly/1bauuac and http://people.mozilla.org/~sguha/)

A Streaming Statistical Algorithm for Detection of SSH Keystroke Packets in TCP Connections, Saptarshi Guha, Paul Kidwell, Ashrith Barthur, William Cleveland, John Gerth and Carter Bullard, INFORMS Computing Society Conference, Monterey, 2011.

Distributed Data Analysis, Invited talk at ISBS, Portoroz, 2010.

RHIPE: Subsetting and Analyzing Massive Data With R, Invited talk at High Performance Computing Section, R In Finance Conference, April, 2010.

RHIPE: Examples with Massive Data and R, Invited talk at Bay Area R Users' Group,March, 2010.

Visualization Databases: Tools Involved, ASA Invited Presentation at the American Statistical Association, JSM 2009.

Resultant-Vector Banking Of Graphical Displays: Geometry And Statistical Properties, Joint Statistical Meeting, Denver, 2008.

Posters

Visualization Databases for Analysis of Large and Complex Data Hafen, R. P., Guha, S. and Cleveland, W. S., Second Annual U.S. Department of Homeland Security Annual University Network Summit, Washington, DC, 2008.

Papers

A Rules-Based Statistical Algorithm for Keystroke Detection, Guha, S., Kidwell. P, Barthur., Cleveland, W.S., Bullard, C. and Gerth J., ICS 2011, Monterrey.

Visualization Databases for the Analysis of Large Complex Datasets, Guha, S., Hafen, R. P., and Cleveland, W. S., Proc. of the 12th International Conference on Artificial Intelligence and Statistics, 2009.

Skills

Experience with statistical methodologies such as Design of Experiments, Non Parametric methods, General Linear Mixed Models, Non Linear Modeling, Mixed Effects Modeling, Survival Modeling etc for both descriptive and predictive models.

Experience using 'data mining' tools such as random forests, decision trees, boosting and bagging, multivariate data analysis.

Experience with a range of ***Big Data*** technologies e.g. RHIPE, Hadoop MapReduce, NoSQL technologies(HBase) and Pig.

Comfortable in R, C , Java,Lua and Python.

Experience using SAS,SQL and LaTeX for data analysis and reporting.

Designed and implemented an open source R package (in R, C and Java) for a seamless integration of R and Hadoop. (http://datadr.org/)

Honours and Awards

Amazon Web Services' Research Grant for $3000, September 2009

Purdue Research Fellowship, 2007-2008

Ross Fellowship, 2004-2005

Languages

Fluent in Bengali and English.

Organizations

American Statistical Association

Activities

Cycling, swimming and scuba diving.

Referees

Prof. William S. Cleveland
Professor
Purdue University
West Lafayette, IN, USA
phone: ***available on request***
e-mail: ***available on request***

Richard Kittler
VP Professional Services
Revolution Analytics
Palo Alto, CA, USA
phone: ***available on request***
e-mail: ***available on request***

Gilbert C. FitzGerald
Director of Analytics
Skype
California,USA
phone: ***available on request***
e-mail: ***available on request***