Russi Chatterjee

- Github: [@ixaxaar](https://github.com/ixaxaar) Stackoverflow: [@ixaxaar](https://stackoverflow.com/users/2134957/ixaxaar) Twitter: [@markovinchains](https://twitter.com/markovinchains)
- Email root@ixaxaar.in / russichatterjee@gmail.com

In short

- Data Scientist and Engineer with ~ 8 years exp.
- Experience in Ad-tech, healthcare, linguistics, social networks, and IOT
- Blend of both data science and engineering experience
- Held leading roles for > 4 years

Stack

Languages: Scala, Python, bash ML frameworks: scikit-learn, scikit-image, networkx, pytorch, CoreNLP, KenLM, openIE, ML-LIB, pymc3, pandas, numpy, scipy Big data: Spark, presto, hive, storm, AWS EMR, Athena, Glue catalog Databases: cassandra, elasticsearch, postgre / postGIS Other: Lucene, GraphX, java topology suite

Experience

1. Senior Data Scientist, Lifesight.io (05-2018 - present)

- Geo-location based ad analytics
- Lead the data science team
- Worked on petabyte-scale data
- Userprofiles
  - A daily-updated dataset of every known activity of 1.7 billion users
- Cross devices
  - Measure affinity between users, identify household-level user clusters
- Behavioral Segmentation - segment users based on places and visitation patterns
- Affluence estimation
- Optimal radius of attribution - model to estimate the extent of places
- Visibility of billboards, traffic density estimation

2. Data Scientist, Reverie language technologies (07-2016 - 08-2017)

- Creating an NLP stack for Indian languages

- Individual contributor role amongst a team of researchers

- Worked on several cutting-edge problems for oft-ignored languages

- Transliterate a search query from Indian languages to English

  - e.g. “ ” “feynman lectures hardbound”

    - Employs a seq2seq-attn neural net and Language models

- Extract terms out out of a search query and label the terms
    - { “feynman lectures”: “book”, “hardbound”: “type” }
    - Employs Knowledge bases and mapped taxonomies
- Cross-linguistic entity recognition
- Kneser-Ney ARPA Language models
- Language model supported spell correction
- Extractive sentence summarization
- Generative Neural language model as autocomplete

3. Data Scientist & Tech Lead, Practo (05-2015 - 07-2016)

- Lead the search team
- Conceptualized and executed “Cerebro”, the search engine that powers practo.com

```
  - Medical query expansion using SNOMED-CT
  - Robust autocorrect using combination of string distances
  - Search ranking algorithm and evaluation
  - Mungle and derive insights about user behavior (click streams)
  - Conceptualization, architecture and engineering processes
  - Stack: Scala, Akka, Lucene, CoreNLP, Spray, Spark, Python, scikit-learn
```

4. Software Engineer, on contract, Showt (07-2014 - 01-2015)

- Data engineering role
- Involved designing distributed systems at a scale of 20k QPS

```
Trends: trends detection in social streams
  - Define a model for "trending" and algorithm to detect trending, implement microservice
  - Evaluate and validate algorithm with twitter data as reference
  - Stack: Python, node.js, ZeroMQ, redis, cassandra

Rylai: Milestone / Event detection in social streams
  - Infrastructure for sharded counting
  - Data pipeline for event delivery (feeds and notifications)
  - Stack: Scala, storm, cassandra
```

5. Software Engineer, MagnetWorks (08-2013 - 06-2014)

- IOT analytics product for heavy industries
- Extremely hands-on, full-stack developer role

Frontend dashboards:

- Dynamic web dashboards built on backbone
- Stack: D3, NVD3, Backbone, Bootstrap, JQuery, HTML-CSS-JS

Analytics backend for time series data, ETL engine, ML algorithms job runner

- Compute engine of various time series metrics
- Stack: node.js, R, RServe, CouchDB, PostgreSQL

Time series data capture and API backend

- Magnetlang - python-based custom DSL for ETL
- Stack: python, Django, pandas, numpy

6. Software Engineer, Toshiba (02-2013 - 08-2013)

- Wrote interrupt subsystem & ABI of TOPPERS microkernel
- Stack: Assembly languages (ARM and Tensilica), C, gcc

7. Software Engineer, Wipro (02-2011 - 02-2013)

- Linux device driver for Maxim 9240
- Bootloader and remote firmware updater for Renesas M16C
- Stack: C, Linux kernel, Assembly language

Open source

- **Deep Learning**

1. pytorch-dnc : Differential Neural Computers, Sparse Access Memory and Sparse Differentiable Neural Computers. (replicating code unreleased by google)
2. pytorch-dni : Implementation of Deepmind's "Decoupled Neural Interfaces Using Synthetic Gradients".
3. subLSTM : Implementation of subLSTMs from the paper "Cortical microcircuits as gated-recurrent neural networks".
4. OpenNMT-memnets : A fork of OpenNMT supporting memory networks for seq2seq.
5. awd-dnc-lm : A Neural language model with augmented memory.

- **Computational Neuroscience**

6. Spectrum : Higher-order spectrum estimation toolkit, for analysing EEG recordings
7. Cortical Microcircuit : Implementation of a model of a cortical microcircuit (a 1mm^2 of the human brain), in pyNEST, from the paper "The cell-type specific cortical microcircuit: relating structure and activity in a full-scale spiking network model."
8. S-transform : The Stockwell transorm.

- **Mathematics**

monoid.space, a book on learning how to “implement” a few math theories (logic, group theory, category theory, algebraic topology and HoTT) using type theory in Agda (in progress).

Education

- B-Tech., Computer Science, NIT, Rourkela - 2006-2010
- Internship on robustness of image watermarking, Indian Statistical Institute, Kolkata
- Thesis on Computer Vision - Texture Segmentation Using Gabor Filters and Wavelets
- Coursera certificates in Statistical Inference, Computational Neuroscience, The Brain and Space, Synapses, Neurons & Brains, Bioinformatics