

---

## EDUCATION

**Columbia University** **New York, NY** **Expected Dec 2017**

- Master of Science, Data Science, GPA: 4.13/4
- Relevant Coursework – Probability and Statistics, Machine Learning for Data Science, Applied Machine Learning, Big Data Analytics, Algorithms in Data Science, Storytelling with Data
- Course Assistant for Machine Learning (edX Columbia University MicroMasters Program) – Summer 2017

**IIIT Hyderabad** **Hyderabad, India** **Jul 2007 – Jul 2013**

- Master of Science (by Research), Computer Science, July 2013, CGPA: 4/4
- B.Tech. (Hons.) in Computer Science, May 2011. CGPA: 3.67/4
- Dean's list in 3 out of 8 semesters for academic excellence
- Teaching Assistant - IT Workshop, C Programming, Computer Vision and Introduction to Databases (Head TA)

---

## SKILLS

Python (IPython, Numpy, PySpark, Scipy, Pandas, Matplotlib, NetworkX, Scikit-Learn, Keras); SQL; R; Spark;  
JavaScript (jQuery, D3.js), Web Development (Python - Flask), Mobile App Development (iOS - Swift)

---

## DATA SCIENCE PROFESSIONAL EXPERIENCE

**Twitter** **Data Science Intern** **Jun 2017 – Sept 2017**

- **Media Data Acquisition Module for PyCX Library** (*Tools Used: Python, PySpark, Scala, Scalding*)
  - Improved PyCX (the PySpark framework for Data Analysis) with increased Twitter data source access. Also, used PyCX to generate relevant tweet data to train GANs for automatic tweet generation (internal research project).
  - Implemented PyCX Media Sampler to acquire data using Scalding Jobs(Scala) and PySpark notebooks for ML/DL tasks
  - Twitter Toxicity Intern Challenge – Supervised Classification challenge for the detection of toxic tweets. Achieved 87% accuracy and stood 2<sup>nd</sup> amongst the Science Org Interns.

**TCS Innovation Labs** **Researcher – Applied Data** **Aug 2013 – Jul 2016**

- **Analyzing GitHub Experts** (*Tools Used: Numpy, Pandas, Scikit-Learn, NetworkX, D3.js, Flask, AWS*)
  - Analyzed the code data of several JAVA experts to discern their skillset, tools and techniques used most often and ways to tag and classify these into subfields for visualization.
  - Publication – Saxena R., Katiyal A. and Pedanekar N.; “I Know What You Coded Last Summer: Mining Candidate Expertise from GitHub Repositories”, CIKM Workshop on Data-Driven Talent Acquisition, Oct 2016

---

## ACADEMIC DATA SCIENCE PROJECTS

- **Machine Learning Model Implementations** (*Tools Used: Numpy, Pandas, Matplotlib, Seaborn*)
  - Implemented Ridge Regression, Naïve Bayes, KNN and Logistic Regression for problems like vehicle mileage prediction and spam/ham email classification.
  - Implemented Ada-Boost to boost a weak least squares classifier to predict room occupancy given weather data. Also, implemented K-Means to cluster the data generated from a weighted mixture of Gaussians.
  - Implemented Probabilistic Matrix Factorization to predict movie ratings and find similar movies based on MovieLens Dataset and Non-Negative Matrix Factorization for topic detection on New York Times dataset.
- **Applied Machine Learning Projects** (*Tools Used: Numpy, Pandas, Scikit-Learn, Keras, Matplotlib, Seaborn*)
  - Text Classification, Clustering and Topic Modeling of Complaints about Boston Traffic conditions.
  - Stood 2<sup>nd</sup> amongst 100 teams participating in In-Class Kaggle Classification Analysis of Bank's Marketing Campaign to analyze Subscription Status. Achieved best ROC-AUC score of 0.798 using ensemble methods.
  - Regression Analysis ( $R^2$ : 0.62) to predict Market Rate for Apartments based on NYC Housing Survey (NYCHVS)
- **Data Visualization Projects** (*Tools Used: Numpy, Pandas, Scikit-Learn, NetworkX, D3.js, Flask, AWS*)
  - Quantified Self: Analyzing Personal Garmin Data <http://bit.ly/2vdiox7>
  - Marvel vs DC Superheroes compared using D3.js. <http://bit.ly/2wlvZAe>

---

## RESEARCH EXPERIENCE: THESIS AND PUBLICATIONS

- Successfully defended dissertation titled “Spatio-Spectral and Statistical Approaches to Pixel Unmixing and Improving Classification Accuracies.” Published Research papers at Major Remote Sensing Conferences:
  - Katiyal A., Rajan KS; “An Unmixing Framework to improve class accuracies using detected High Importance Local Regions”, IEEE IGARSS, July 2013.
  - Katiyal A., Rajan KS; “Spatio-Spectral method for Estimating classified regions with high confidence using Modis Data”, ISRSE35, April 2013. Awarded **Best Young Scientist Presentation Award** at the conference.

---

## HOBBIES

- Quantified Self Enthusiast, Avid Runner