Big Data Hero, Richard Freeman , PhD Lead Data Engineer, Solution Architect and Data Scientist at JustGiving

Congratulations! Richard Freeman has been selected as this month's techUK 'Big Data Hero'.

The purpose of techUK's 'Big Data Hero' campaign is to highlight the benefits of pursuing a career involving Big Data and Data Analytics in order to try and reduce the Big Data skills gap we are currently facing in the UK.

If the UK is to fully capitalise on the Big Data revolution then we need to attract more people into the industry. We think highlighting the careers and experiences of industry leaders will inspire people to consider a career in Big Data and Data Analytics, which will help secure the UK's Big Data future.

What is your current role and what are your responsibilities?

I currently work for JustGiving, a tech-for-good company and the number one platform for online giving in the world that's helped 28.5 million users in 196 countries raise $4.2 billion for over 27,000 good causes.

I'm leading the architecture and development of our in-house big data platform RAVEN and big data production systems hosted in Amazon Web Services (AWS). In addition I work as a data scientist specialising in scalable streaming analytics, machine learning (ML) and natural language processing (NLP) algorithms. I also enjoy sharing my technical experience and knowledge internally and externally relating to AWS, Azure, stream processing, serverless stacks, ML and NLP. I am a peer-reviewer for several large conferences and journals, present at industry conferences, open source my code and write technical blog posts.

What is your academic/career background and how did you get to where you are today?

I have a solid background in computer science with a Master of Engineering (MEng) in computer systems engineering and Doctorate (Ph.D.) in machine learning, artificial intelligence and natural language processing both from the University of Manchester.

Rather than join a specialised vendor in my Ph.D. area of expertise, I decided to broaden my skills and gain more client exposure by joining Capgemini, a large global consulting, technology and outsourcing services company. I was in the Business Intelligence & Information Management and later in the Financial Services groups, working my way from being a developer to a solutions architect. I helped deliver large scale projects for Fortune Global 500 companies in insurance, retail banking, financial services, financial regulation and central government.

I then joined Michael Page where I worked as an architect on a global transformation programme across 34 countries. I led and designed the technical solutions for search, multi-channel communication, business intelligence, text analytics, job board integration, and advertising solutions.

What made you choose a career in Big Data and Data Analytics?

For data science, I've always been interested in artificial intelligence, machine learning and natural language processing, and particularly at ways to make scalable systems and robots more intelligent and responsive. I do a variety of tasks such as data preparation, running experiments, deriving insight, creating visuals, presenting data stories, and educating others. For the data engineering, my interest comes from my background as a solution architect where I enjoy building cloud-based systems, that store and process data to derive new insight and knowledge. I'm also interested in big data pipelines and automating the whole machine learning process, as this helps the data scientists and analysts save time preparing the data for running their algorithms, metrics and key performance indicators at scale.

What does a typical day involve?

JustGiving is still a start-up at heart, so there is no typical day and I get involved in various tasks, such as data and report requirements capture, engineering new data pipeline, investigating operational issues, running data experiments, analysing unstructured data looking for useful patterns, exploring new ways to use the data to answer questions, presenting a data story, and sharing my knowledge and experience. This means that I work closely with marketing, product managers and product analysts to understand their data needs and what metrics are important for them. Speaking to others outside your specialist area helps to broaden your views, gives you a new perspective and new ideas. On the technical side, I work with engineers, data analysts, developers, business intelligence analysts, operations, and data scientists to support their data and platform requirements.

What do you most enjoy about your work?

I am passionate about working with huge data sets, as you face different kinds of performance and operational issues that require you to think differently in order to scale your data warehouse, ETL processes, and algorithms and how you present your results. Building and running large-scale data infrastructure and algorithms, are active areas in academia and industry. They are evolving at a fast pace, with new tooling being introduced every few months. I like to use cloud solutions in an innovative way to improve our in-house data science platform, enhance our business processes, and make data insights available to internal and external users.

I am also a big fan of Python and have successfully encouraged its uptake in the company and wider community for the data pipelines, data science and serverless computing. I also enjoy working with different organisations, charities, universities and giving back to the wider technical community.

What do you think the future holds for Big Data?

I see more people using machine learning, real-time analytics, graph analytics and natural language processing in their products and applications. For real-time analytics, there is a growing demand from consumers that are much more data aware and impatient. For example they want to know what is happening right now, see the results of their action, and use more intelligent applications and websites that adapt as they are interacting with them. On the infrastructure side, I see serverless computing and Platform as a Service (PaaS) infrastructure in the public cloud such as AWS and Azure becoming more prominent. Functions in serverless computing are particularly interesting for me, as they can auto-scale within a second, are highly available and are low cost. They are low cost as you only pay for the time your code is executed, rather than for an always-on machine or container like in more traditional cloud infrastructure. The open source frameworks and programming languages will also continue to grow compared to closed vendor specific products and languages, e.g. Apache Hadoop framework, Python, R, SQL. The same goes for the data storage and access: cloud storage, data warehouses and data lakes will store data in more open rather than proprietary formats, and this will be more accessible over standard application program interfaces (APIs) or open protocols. There will also be growing requirements to analyse unstructured and multimedia data sources.

We will also see more companies making the transition from using strategies decided by a few at the top to becoming more data-driven at each level. For example the testing of new products or features, identifying new opportunities and strategic decisions will come more and more from the data analysis, insight and predictions. This will require more staff to get involved in data capture, data preparation, running experiments using algorithms, data visualisation and presenting results. As such, new data orientated jobs based on creating and training data models, and automating existing processes will emerge, disrupting some of the existing specialist fields such as accountancy and law.

The artificial intelligence (AI), automation, internet of things and robotics will also replace some existing blue and white collar jobs so we will need to think about training and upskilling people to the changing landscape, and possibly introduce some kind of universal basic income. Some draw parallels with the shift seen during the industrial revolution from the agrarian or pre-industrial times, and believe that for AI to take off, we need two things to happen: the human cost becomes higher than AI alternative and for AI to be deployed in a scalable way.

In the much the longer term, quantum computing will also disrupt the field again in terms of how we process, analyse and store data, and will transform areas like cyber security and existing AI.

What needs to be done to inspire people to pursue careers in Big Data?

I think it's a lot easier to get people interested in big data and data science than it used to be, thanks to the likes of Google and Facebook that make it fashionable to be smart and work within technology. In addition, the growing number of young and flexible startup companies with infrastructures in the public cloud, are successfully competing and winning market shares from large established companies. Employers need to be willing to educate and upskill existing staff or graduates rather than solely recruit people with existing data engineering or data science skills. For inspiring existing staff, we need to show the benefits, use cases and data sources most relevant to them, which makes them more productive and their job easier. With more data exploration tools available, staff in other departments outside IT or finance, such as customer support, marketing and product managers will be self-serving on the data and insights.

For people who have not worked in industry, I think we need to start early in schools then universities. Teachers and lecturers in non-computer science subjects could make data more visual and interactive in their respective fields. I think that almost any subject can benefit, for example even in English literature you can draw a relationship graph of the characters and their connections linked to main themes, events and locations, or in history you can have and interactive visual graph and time evolving map representations.

What advice would you give to someone considering a career in Big Data and Data Analytics?

Whether you are a graduate, already working in an organisation or not from a technical background you can benefit from analysing and understanding data. For example data journalist are typically not from a technical or scientific background, yet are able to do simple analysis and create an interesting data story for the general public.

It's about self-motivation, when things move at such a fast pace, you can look broadly across the sector to gain a general understanding but also need to focus your energy on one specific course or project and complete it. The industry also tends to repackage old technologies with some improvements as new trending ones like cyber security, cognitive computing, chatbots, virtual reality and artificial neural networks at the moment (October 2016), so I would follow your heart for the areas you are truly interested in and want to focus on, rather than the latest trend.

In terms of gaining the knowledge it is a lot easier than it used to be, for example in the past you had to pay for specific vendor training and there was a cost to the product itself. You can now access the learning materials, data sources, and tools all for free, so there is no excuse not to get started today! For the learning materials, a lot of the content is available for free in massive open online courses, forms, blogs, and source code repositories. Equally there are numerous free data sources like ML datasets, open data, news feeds and social media you can use. There are many tools out there, some are graphical but in my view you should learn to program in SQL, and Python or R. All three have the ability to do data science at scale thanks to frameworks like Apache Spark. I particularly like Python as it benefits from being an efficient development language with a solid test framework and numerous data science packages. As a data engineer or data scientist, expect to spend a lot of time on data preparation. This is an important process to master, which involves the cleaning, parsing, enriching and shaping the data so that it can be used in the machine learning algorithms and experiments. Overall remember that the processes, tools and data sources are always evolving, so there is no one off training course you can do. You will need be self-motivated and open to constantly learn and adapt to the data ecosystem.

In my opinion, with the lack of clarity and impact from the Brexit and probable cuts in EU funding for research, I would recommend that you learn another European language to remain mobile within Europe. In addition this will also open your mind and give you an insight into other cultures and values, and how they use their data. Cloud computing also means that you no longer need a physical presence in a country to operate in it, so you need to be open to build systems across regions and analyse data from many countries.

Some jobs and professions will be replaced, and some human expertise will be lost, but we will still rely on the data and algorithms. For example once driverless transportation is widely adopted and considered safer, cheaper and more convenient than human drivers, future generations may not wish to drive a car or even have a driving licence. However humans will still be involved in the systems that automate the driving, the creative analysis of the telemetry and IoT data, the supervision and monitoring of the ecosystem, and the wider participation in the transport industry and sharing economy.

In summary if you want to have a career in data science or data engineering, think about the metrics you want to calculate, hypothesis you want to validate with an experiment, metrics that will benefit your business, what actions will your audience take with your results, growth opportunities for a business then work back to see what data, algorithms and infrastructure you need for the task. I think that being curious, inquisitive and having an experimental mind are important qualities.