BIG DATA Revolution

INNOCENTIVE®

Big Data and the Crowd

Workshop Report

Big Data and the Crowd Workshop Report

Contents

1- Introduction

Recent years have witnessed dramatic improvements in data collection, storage and sharing; significant innovations in the field of cloud computing; the development of sophisticated, new tools for data analysis; and the exponential rise of the 'internet of things'. These interconnected sub-trends can be viewed as multiple aspects of the on-going big data revolution, which is driving a paradigmatic shift in the way businesses, organisations and governments operate.

The crowdsourcing revolution is another technology-enabled mega-trend that is transforming our world. This movement has been facilitated by communications technology advances that have dramatically expanded the opportunities for deploying the crowd in areas such as content-production, research and innovation. The exponential rise of the online innovation competition provides a good example of this phenomenon in action.

InnoCentive convened a workshop to explore the potential interconnections between these two technological mega-trends. The event took place at Millbank Tower in Westminster, on 12th September 2012. Individuals from a diverse range of organisations were invited to attend and commercial companies, academic institutions, public bodies and NGOs were all represented. Ryan Shuttleworth, a Technical Evangelist from Amazon Web Services, presented on big data, while InnoCentive staff presented on the rise of the crowd and facilitated the workshop's discussion and ideation sessions.

This report provides a brief summary of the Big Data Revolution Workshop.

2- Attendees List

The following is a complete list of the individuals who attended the workshop, along with details of their respective organisations:

Name	Company
Helen Andrews	Olswang
Sanjit Atwal	Squawka
John Callahan	ONR
Graham Cameron	E.ON
Marco Cardinale	British Olympic Association
David Chan	City University London
James Dennis	Royal Holloway
Siobhan Gibney Gomis	InnoCentive
Alan Grogan	RBS
Julio Guijarro	HP
Mirsoslav Hamouz	Alertme.com
Cecilia Liao	Deloitte
Hamish McArthur	Mondo Visione
Moritz Neutard	InnoCentive
Stephanie Okimoto	DHS
Nick Pope	BioSpring
Tristram Riley Smith	CSAP
Dannielle Roberts	Sport England
Simon Schneider	InnoCentive
Ryan Shuttleworth	Amazon
Herbie Skeete	Mondo Visione
Jonathon Slater	InnoCentive
Iain Sterland	Boots
Jay Stow	InnoCentive
Dave Strain	Pitney Bowes

3- The Big Data Revolution

After the initial welcome and introductions, Ryan Shuttleworth presented on the big data revolution.

With the increasing proliferation of incredibly large datasets, organisations all over the world are looking to innovate in the ways they manage, analyse and utilise the data they have. It is not only its' growing volume that makes the data difficult to use, but also its' increasing complexity and variety. Significant challenges in data analysis can actually start at relatively low volumes of data, depending on the specific context and the resources or technologies available for deployment. The progress of cloud computing has played a central role in the big data revolution, enabling dramatically improved data collection and processing. Cloud technology is crucial in allowing large-scale collaboration in the compilation, organisation and utilisation of big datasets.

The rapid progress made in areas relating to human genomic science during recent years, provides a good example of the radical nature of the big data revolution. One case examined regarded efforts to develop cancer drugs: in order to enable the design of these pharmaceuticals, computational chemistry algorithms needed to be developed that could identify protein targets against a massive bustle of background noise. This comprised a serious big data issue, as 21 million compounds needed to be tested and a highly accurate level of analysis needed to be achieved. Such an operation required over 12 years-worth of computing hours and a few years ago would have demanded over $20 million in infrastructure investments. Using cutting edge cloud systems however, this procedure was accomplished in just three hours, for a cost of under $1500. Progress of this magnitude clearly represents a significant leap forward, especially as the technology has applications in virtually every area of scientific, commercial and organisational endeavour.

4- Hopes and Fears Discussion

The workshop was asked to discuss their hopes and fears regarding the big data revolution.

The group discussed privacy issues and the subject of data-ownership. Significant problems highlighted concerned openness, transparency and consent in data-management and developing a good technique for open data-collection was emphasised as a desirable, potential innovation solution. New models deploying innovative strategies to incentivise data-sharing amongst individuals could be a powerful tool for building massive datasets free from the suspicion of 'big brother' exploitation. Another useful, potential technology innovation mentioned involved the opportunity to systematically build privacy controls into big data architectures on a foundational level. The problems involved in integrating datasets were also discussed and there is certainly scope for the development of useful technologies and techniques to assist in this area.

5- The Rising Crowd

Crowdsourcing can be defined as the practice of outsourcing defined tasks or goals to an undefined public. The model has a long and colourful history, with innovation competition crowd-solving facilitating many of the significant advances of the industrial revolution. Modern communications technology has revolutionised crowd-deployment in recent years and enabled a major paradigm shift to occur in the area. Crowds now collaborate in the design of open source software, assist in the development of commercial products and undertake a wide range of collective decision-making roles.

The radical nature of these developments has been dramatically demonstrated by the crowds of the Arab Spring Revolutions, as social networking enabled peer-to-peer, decentralised coordination to play a key role in shifting the balance of power between people and their rulers. Furthermore, movement towards increased decentralisation and de-commercialisation could be reversing the major economic trends of the 20th Century. The potential for crowdsourced employment to significantly alter everyday lifestyles might also be highly consequential.

Overall, the 'new crowd' is changing the way we work as a system, on both a micro and a macro level. Crowd systems balance order attained through centralised hierarchy with order attained through decentralised coordination and this may represent progress towards a more mature and sophisticated system, perhaps better able to cope in an increasingly complex world. But the idea raises an important question: how are we to work this new system, both globally and on a local organisational level?

6- Big Data and the Crowd

These two great technological mega-trends are deeply intertwined and several broad interconnections can be readily identified. There is a clear association between open data and open organisation, as the working coordination of crowd systems clearly requires transparent information to be made publicly available. The link between open data and open innovation is also natural, as crowds need access to relevant information in order to maximise their innovative potential. Furthermore, it is logical to see how a large crowd could assist in dealing with an especially difficult dataset, potentially taking on data that is both too voluminous for traditional organisations to tackle and too complex for current computers to analyse.

A variety of systems have been deployed that seek to exploit the apparent synergies between crowdsourcing techniques and big data technologies. Amazon's Mechanical Turk system breaks major projects down into bite-size chunks and then crowdsources the work, compensating participants with financial rewards paid per task they accomplish. Many of these projects are strongly data-oriented. The Citizen Science Alliance's Zooniverse project also uses the crowd to tackle large, complex datasets: mobilising a network of 600,000 people to analyse photographs of the universe and help classify and categorise galaxies and celestial bodies. PatientsLikeMe harnesses the crowd for data-collection purposes, collating large amounts of health data from volunteers and then making it available for medical research.

There have also been several notable data-themed innovation competitions. The Netflix Prize (2006) offered $1 million to innovators able to come up with a predictive algorithm that improved Netflix's movie recommendation system's success rate by 10% or more. The reward was collected by a collaboration of teams who combined their software in order to collectively attain the 10% improvement demanded. The GoldCorp Challenge offered $500,000 to innovators who could analyse the company's mining data and successfully predict the best places to prospect in the future. The competition was a resounding success: not only was $3 billion worth of gold discovered in the new seams, GoldCorp also attained access to a diverse range of valuable new prospecting technologies. The ALS Prize4Life Challenge, run by InnoCentive, asks the crowd to predict the functional deterioration rate of people diagnosed with ALS (Amyotrophic Lateral Sclerosis). Innovators use the training datasets supplied to develop their algorithms, before testing their solutions online against hidden test datasets and receiving instant feedback on their success through a public leader-board updated in real-time.

7- Thoughts and Reactions Discussion

The group discussed big data and the crowd and the issues raised so far in the workshop presentations.

The problem of organising datasets so that they can be used by the crowd was emphasised as significant. Suitable techniques for data preparation are clearly required and it was mentioned that good models for data visualisation could be very useful. Organising data in an attractive way would likely help to attract more solvers, but may risk forcing innovators down a specific innovation path effectively prescribed by the way the data was initially visualised. An innovation challenge based around developing new tools for data-visualisation could be useful across a wide variety of areas.

The potential for artificial intelligence analytics systems to be trained by human crowds was also discussed. It is conceivably possible to use the crowd to correct or confirm results provided by data-analysis software and for the computer to systematically use crowd-feedback to improve its own processes. Designing software that could undertake this kind of learning could provide valuable general purpose technology, useful across a whole spectrum of fields and disciplines.

The difference between crowd 'intelligence' and crowd 'wisdom' was also emphasised. Crowd intelligence is demonstrated by clever innovations that win inducement challenges whilst crowd wisdom depends on the amalgamation of numerous individual decisions building up to a wise collective decision. The potential to combine these two distinct crowdsourcing strategies was highlighted.

8- Big Data Challenges

Before embarking on an ideation session to brainstorm ideas for big data innovation challenges, the group was given a brief presentation detailing some of the basic fundamentals of technical challenge design. A simple formula for challenge concept development was outlined, breaking the key questions down into the following categories:

- Context
- Aims and Goals
- Victory Criteria
- Success Metrics
- Structure/Events
- Reward
- Datasets

Certain significant competition design strategies and techniques were also highlighted. Challenge abstraction was emphasised as significant: where a specific challenge is reduced to its core fundamentals and reconstituted in abstracted terms. This encourages innovators from outside the traditional sector to participate in a competition and helps to avoid over-prescriptive solution definition. Gamification was drawn out as a useful model to be applied: accentuating the 'game' element of a challenge can encourage more people to get involved and increase the effort participants are willing to contribute. Multi-level crowdsourcing was also underlined as a potentially fruitful avenue for exploration when designing data challenges. Synergistically combining competitive crowd-solving with crowd-data-collection or collaborative open source innovation could significantly improve the effectiveness of a challenge.

Big data innovation challenges can be applied in a number of different areas: facilitating the development of useful models or simulations; advanced data analysis solutions; or improved fundamental data tools (e.g. data visualisation techniques). Data challenges have some distinctive advantages relative to other types of innovation competition: there is a strong tendency towards solid, quantifiable success metrics; instant-testing with real-time success feedback is enabled; the possibility for individuals and small organisations to participate is maximised; and there is a good opportunity to quantify returns on innovation investment post-challenge.

9- Challenge Concept Ideation

The workshop now turned to brainstorming ideas for potential innovation challenges, with the group asked to consider relevant problems or opportunities within the field. The following possible competition ideas were generated:

1. Obtaining Quality Data Challenge – There is often a difficulty in collecting high-quality data, as self-selection bias and dishonest responses can greatly distort results. An ideation challenge could focus on improving data collection techniques.
2. Define Robust Data Challenge – There appear to be tipping points in data science concerning quantity and quality of data. Beyond these theoretical thresholds datasets seem to significantly improve in utility and robustness. A challenge could seek theories to explain this phenomenon, rewarding well-supported identification of important threshold values in a range of specific contexts.
3. Research Data-Mining Challenge – There is huge scope for the advancement of data-mining technology that can efficiently extract and present useful information from big online datasets. Such a challenge could usefully be applied to glean material from a public research portal.
4. London Transport Prediction Challenge – A challenge could focus on the extraction of data relating to London's transport network, potentially facilitating improved traveling experience and efficiency. Twitter datasets could be analysed for relevant information that could be used to help predict specific individual journey times.
5. Marine Pandemic Prediction Challenge – There is little understanding of how diseases spread through populations of fish, although such knowledge would be desirable as it could improve the efficiency of the fishing industry. A challenge could reward the development of a model able to accurately simulate marine pandemics and successfully predict their consequences.
6. Online Identity Verification Challenge – Online identity verification represents a significant big data challenge, as systems need to be developed that can successfully distinguish isolated illegitimate access attempts amidst a noisy background of legitimate requests. Technology that could do this effectively, and in real-time, would be valuable indeed, making this another interesting area to potentially deploy an innovation challenge.

10- Challenge Concept Development

The London Transport prediction challenge was selected as an idea that seemed worthy of further concept development. The purpose of this exercise was to generate debate around certain challenge architectural options and provide insights into the intricacies of the overall design process. The group discussed the basic elements of challenge design one by one, following the framework outlined earlier in the afternoon (see 'Big Data Challenges' section).

It was decided that the overall aim of the challenge would be to develop software that could accurately predict transport times in London and use this information to improve network efficiency and traveller experience.

Attempting to develop a 'the winner is…' statement encouraged discussion around attaining the appropriate balance of audacity and achievability within the victory conditions. Rewarding technology that could analyse multiple data-feeds to accurately predict journey times was one option discussed, although it was argued that this may be an overly ambitious target for a single-stage competition. Perhaps the first stage of the challenge should be to develop software that can identify relevant events and incidents through real-time Twitter analysis, before launching a separate sub-challenge to utilise this information in a way that enables travellers to better plan their own journeys.

Possible success metrics could measure the accuracy of journey-time predictions by testing innovator's algorithms against a hidden validation dataset. Other useful metrics might include measuring how quickly the information could be extracted and quantifying the volume of data that the system can analyse effectively.

Alternative competition structures were discussed and it was noted that a challenge of this complexity would probably work best if broken down into a progressive series of sub-challenges. The first stage might concern the identification of relevant events from specific data-feeds, with the second stage focusing on deploying this information to make useful predictions and the third stage rewarding innovators who could elegantly turn this know-how into a user-friendly smartphone app. Live events should constitute an important aspect of almost any data grand challenge.

Reward amounts were not really speculated upon, although it was noted that intellectual property considerations would need to be addressed within challenge design because the data-mining technology involved could be extremely valuable.

There are many datasets that could be analysed to assist in accomplishing this challenge. Social media feeds such as Twitter would be useful, as would historical TfL datasets or meteorological information. The question of whether all participants would be required to use the same data was raised, leading to discussion regarding whether innovators should be encouraged to use their own initiative and draw on whatever data they see fit.

11- Conclusion

In conclusion, it seems there is much scope for the crowd to assist in the development of advanced big data technology. Many ideas for potential innovation challenges were conceived at the workshop, not only to facilitate the advancement of predictive analytics tools but also to design improved systems for collecting, organising and presenting data. The concept development around the London Transport Prediction Challenge demonstrated how these ideas could begin to be worked up, but also helpfully highlighted some of the complexities involved in prize design. The fact that multiple stage grand challenges would probably be needed to tackle the biggest innovation hurdles was also implied during the formulation of this competition concept.

There is plenty of potential to deploy the challenge model in the field of big data and many areas exist where data technology innovations could have a dramatic impact across multiple sectors. Pioneering organisations that seize the opportunities presented by big data and the crowd will get a head start on their rivals and will be able to play a leading role in building the exciting future that these grand technology revolutions promise.