Crowdsourcing Data Science for Social Good

Submitted by Kirk Borne, Steven Mills, and the Data Scientists at Booz Allen Hamilton
(If accepted, this paper will be presented by Dr. Kirk Borne)

Abstract

This paper describes the origin, motivations, target problems, results, and future of the Data Science Bowl international competitions. We focus on the roles of crowdsourcing, data partners, participants, sponsors, and social engagement in furthering the cause of data science for social good in addressing significant global problems. Details are provided on the first three Data Science Bowl competitions, including impact statements from members of the communities that are the direct beneficiaries of the winning machine learning solutions. Finally, an invitation is extended to the data science community to join us in the upcoming and future competitions, to experience the benefits and rewards of this competitive yet supportive community, and to change the world.

Introduction to the Data Science Bowl

On Thursday Feb.18, 2016, two cardio (heart) scientists from the U.S. NIH (National Institutes of Health) National Heart, Lung, and Blood Institute (NHLBI) went online for a Reddit AMA ("Ask Me Anything") session to discuss the importance, value, and benefits of cardio imaging analytics for improving heart disease diagnosis and (ultimately) for saving lives! [REF1] The context for this AMA discussion was the 2nd Annual Data Science Bowl. This Science Bowl was the second of now three completed data science competitions on the Kaggle.com platform that has been devoted entirely to addressing major social good challenges through data analytics and machine learning.

The Data Science Bowl (DSB) was started in 2014-2015, organized by and sponsored (with prize funds) from Booz Allen Hamilton Corporation (hereafter, Booz Allen), and hosted by Kaggle. The DSB has been run each year since then, and the plan is to continue the DSB for the foreseeable future.

The general structure of the competition has not changed since the first one: the DSB runs for three months, is open to all data scientists worldwide (except those from the sponsoring institutions, who may participate but are not permitted to win prize money), is judged with a strictly objective performance metric (that is described in the competition rules), and offers a range of cash prizes to the teams who produce the top-performing machine learning algorithms (as measured against an unseen validation test set). Within this consistent framework, the targeted social challenge problem is different each year.

The first DSB (completed in spring 2015, with a total cash purse of $175,000 provided by Booz Allen) was focused on ocean health through classification of 100+ varieties of plankton species as seen in a large trove of subsurface ocean images. The second DSB (completed in spring 2016, with a total cash purse of $200,000 provided by Booz Allen) was devoted to improving heart

health diagnosis through cardio imaging analytics of MRI heart scans. The third DSB (completed in spring 2017, with a total cash purse of $1 million provided by a private foundation: see below) was aimed at improving early detection of lung cancer in radiological CT scans.

Crowdsourcing Data Science

A common feature of the DSB challenge problems has been the type of data being investigated: large images stored in a non-traditional domain-specific data format. Some of the machine learning techniques that have been applied to these imaging analytics challenges include machine vision, deep learning, and convolutional neural networks. Another common characteristic of many online machine learning competitions (including the DSBs) is the necessity for feature engineering – designing, extracting, and testing different features in the images for input into the image classification models. The variety of different feature sets and evaluation criteria, in addition to the application of different machine learning techniques, will differ in interesting and distinct ways between the different teams in the competition. Consequently, one of the most significant defining characteristics of Data Science For Social Good activities, including online competitions like the DSB, is their diversity!

Many researchers of citizen science projects (in which a nontrivial science challenge problem is posed to the public) have argued that it is precisely the diversity of perspectives, approaches, backgrounds, insights, skillsets, and internal biases that lead to the best (most accurate, most insightful, unbiased) crowdsourced solutions. [REF2] In a sense, it is similar to the old cartoon of the blind men feeling the elephant: individually they may have a self-centric (biased) description of what they are sensing, but together they contribute to a complete unbiased community-based (and more accurate) understanding of the whole thing.

Philanthropy and volunteerism, of course, have deeper roots than citizen science and crowdsourcing, though the latter goes back at least 300 years to the Longitude Prize. This prize was established by the British government in 1714 as motivation for innovative individuals to find a solution to the longstanding challenge for seafarers to measure their ship's longitudinal position when far from fixed landmarks (such as coastlines). The quest for the ultimate solution and final prize award continued for over 100 years before the British Parliament closed the Longitude Office in 1828. [REF3] It is a fascinating story with enduring scientific lessons for all about perseverance, iterative solutions, and establishing an experimental mindset. [REF4]

In the modern era, citizen science projects (e.g., Galaxy Zoo) and frameworks (Zooniverse.org) have greatly accelerated public participation and crowdsourcing of solutions to time-intensive science problems. [REF5] In data science, the community of data scientists have volunteered and amplified their contributions to solving societal challenges through hackathons, online competitions, "data for good" philanthropic organizations (e.g., DataKind, Bayes Impact, DataLook, and Data Analysts for Social Good), and organizations with dedicated "Data Science for Social Good" programs (e.g., Booz Allen, UN Global Pulse).

Philanthropy has been a significant feature of the DSB. Significant in-kind resources (human, scientific, and technological) have been contributed by Booz Allen for each DSB. In the first two years of the competitions, the prize purse was also contributed by Booz Allen. In the third year, the significantly increased prize purse was contributed by the Laura and John Arnold Foundation. Additional computing resources and conference passes were provided in the different years by NVIDIA Corporation and Amazon Web Services. Numerous other partners and sponsors have also contributed their time, talents, resources, and social capital to the success of the DSB (these are listed on the datasciencebowl.com website).

Of course, the most critical contributor each year to the DSB competition has been the data provider, who has also communicated, explained, and developed the challenge problem statement, as well as providing assistance in the identification, production and delivery of the validation test sets (i.e., the training sets for the machine learning classification algorithms). We will say more about those below.

In year 3 of the DSB, additional social engagement prizes (in smaller monetary amounts than the competition awards) were awarded to individuals through a random selection (lottery) from among the many hundreds of online "socialites" (i.e., the Twitter and Instagram user community) who mentioned the DSB or some aspect of the competition in their online social posts (e.g., photos of team members, personal stories associated with cancer, etc.).

The Data Science Bowl Story

The DSB was first conceived by Booz Allen, a leading provider of management and technology consulting services. A recognized pioneer in the data science field, Booz Allen's client and pro bono work proved the impact of data science for social good. Our work with the US Holocaust Memorial Museum, for example, focused on preventing episodes of state-sponsored mass killing. The nature of this and other work accomplished through dozens of smaller hackathons and special events inspired our team of data scientists to take action on a much grander scale.

We wanted to harness the power of the data science community to bring together the best and brightest in the field to solve a seemingly impossible societal challenge, something bigger than what could be achieved solo. It gives these analysts *"the opportunity to contribute to something bigger than themselves,"* says Booz Allen Senior Vice President Josh Sullivan.

Thousands of participants devote significant time and skill to discover solutions that push the boundaries of what's possible. In the first two years alone we've facilitated the donation of an estimated 180,000 hours of support against some of society's most complex, challenging problems. Extensive outreach, partnerships, engagement, and media coverage bring a global audience to these challenges, giving a voice to those who cannot speak for themselves.

The idea of tapping into the growing data science community also appealed to the people at Kaggle, the world's largest online global data science community with over 700,000 members. Together, Booz Allen and Kaggle brought the first Data Science Bowl to life in 2014.

From there, the concept grew and we are now planning for the DSB's fourth annual competition.

The Approach

The DSB unites data scientists, organizations, industry experts, and citizens to design revolutionary solutions to global problems. This global online competition brings together all backgrounds and skill levels in a simultaneously supportive and competitive environment.

Tutorials help everyone start out on equal footing, and we work closely with Kaggle to support the participants and build awareness for the social problem being tackled. Engagement planning and sponsorships are key to boosting visibility for the DSB. The additional awareness and interest helps to ensure that the competition outcomes far outlast the time window of the DSB competition and help improve society both directly and indirectly.

Participants work hard on these challenges. For example, it can take hours to run and test their algorithms on the data sets within their environments. We are always seeking sponsors to provide technology and tools to help accelerate their findings and to increase participant engagement and enjoyment. Consequently, we have a purposeful approach to each DSB competition, along these three streams of activities:

- Booz Allen data scientists work with partners, clients, and individuals to identify potential challenge areas in which the data science community can make connections between innovative analytic techniques (such as deep learning and AI) and global problems. We look at these challenges differently from previous attempts and address thorny issues like inaccessible, inaccurate, or "noisy" data.
- Booz Allen then works with a targeted and closed group of core partners to identify and prepare data sets that have never been available for public use and to determine if we can make them public—important to maintaining a fair and intriguing competition. We then work with the partners to design the scope of both the competition (making sure that the solutions will have practical application) and the engagement activities that will accelerate the adoption and the scalability of the winning solutions within the appropriate industry or application domain.
- Booz Allen works with platform partner Kaggle to support participants throughout the competition via tutorials and machine learning kernels, as well as via engagement and sponsorship activities that increase community awareness and that ensure outcomes will have the greatest societal impact as possible.

The Competitions – DSB 1: Predicting Ocean Health, One Plankton at a Time

In 2014, Booz Allen data scientists worked with Oregon State University (OSU) marine scientists to unleash data science on the world of marine biology for the first time, by launching the first-ever National Data Science Bowl (NDSB). NDSB participants examined more than 100,000 underwater images to assess ocean health at a massive speed and scale. The winning solution

produced an algorithm that reduced time of data analysis by 50 percent and boasted a 10 percent increase in accuracy.

Plankton are key to Earth's massively intricate ecosystems. These organisms capture 25% of the CO2 released from burning fossil fuels every year. They also form the foundation for marine and terrestrial food chains. Because they are susceptible to small changes in temperature or water chemistry, plankton populations serve as an indicator for broader ocean health. A drop in plankton populations can be a predictor of devastating effects on our world. Creating algorithms that enable rapid assessment of plankton populations gives us the ability to measure ocean health at a speed and scale never before possible.

More than 1,000 teams participated in the inaugural 2014-2015 DSB, contributing over 15,000 machine learning solutions to the challenge problem. They examined over 100,000 underwater images provided by the OSU Hatfield Marine Science Center (HMSC) and their research partners. The top-placing team consisted of a group of graduate students and postdocs at the University of Ghent (Belgium). (As a consequence of the winning team coming from outside the USA, the NDSB thereafter has been named the DSB!) The #1 winning algorithm was a deep learning classification algorithm, using convolutional neural networks. [REF6] The prize-winning (top 3) teams' algorithms will be key to monitoring ocean health faster and with greater accuracy than ever before. The results surpassed the researchers' greatest expectations.

The DSB algorithms exceeded what marine science researchers thought possible – in some cases achieving human-level performance. The real-time insights they provide are an important leap forward, and had been impossible through manual image classification and analysis. As significant as the algorithms, the DSB event itself has energized work in this area. According to the HMSC, the DSB brought a fundamentally different kind of image classification technique to the field of plankton imaging. The Director of the HMSC, Dr. Bob Cowen, said this at the completion of the DSB 1: "We're excited to receive the winning algorithms from the Data Science Bowl and to test and validate these proofs of concepts in our own labs." Previous methods, even semi-automated ones, still required manual verification of images which delays the scientific process by months to years. The Data Science Bowl specifically allowed HMSC to:

- Quantitatively and comprehensively evaluate a suite of classification schemes to find the very best ones – which is critical to facilitating more rapid analysis of the marine environment to understand and manage impacts of ocean and climate change.
- Connect their scientific team with the top machine learning and computer vision researchers in the world. This has set off a cascade of activity at OSU, and also in France (University of Pierre and Marie Curie) in quickly utilizing the solutions from the competition in a variety of plankton imaging systems.
- Continue to work with DSB winning teams, such as the authors of the 3rd place solution, Dr. Benjamin Graham of the University of Warwick, on implementation of his technique ("Sparse Convolutional Neural Nets"). Other Data Science Bowl data scientists continued to engage with HMSC researchers long after the competition, allowing them to analyze vast volumes of data that would have otherwise been left idle. [REF7]

The Competitions – DSB 2: Transforming How We Diagnose Heart Disease

Data Science Bowl 2 tackled cardiovascular diseases, which cause 30 percent of all deaths in the U.S.—more than all forms of cancer combined. The challenge was to develop an algorithm of key heart disease indicators by evaluating thousands of MRI scans. The National Institutes for Health (NIH) are sharing the winning approach with members of the medical community in hopes of changing the way doctors analyze these images, saving 20 minutes per image scan for cardiologists. [REF8]

Heart disease is the major cause of death worldwide. Working with NIH, we focused the second DSB on how to combat this deadly disease and help doctors diagnose heart conditions earlier.

Declining cardiac function is a key indicator of heart disease. The gold standard test to accurately determine cardiac function uses Magnetic Resonance Imaging (MRI). But reading MRIs is a manual and slow process. Making this measurement process more efficient carries broad implications for advancing the science of heart disease treatment.

The 2015-2016 competition (DSB 2) focused on creating an open source algorithm to accurately measure cardiac function. The top-placing team consisted of a pair of Wall Street quants with no formal medical training. The winning (top 3) teams' algorithms revealed ground-breaking gains. The previous competition already proved that the data science community can successfully attack a long-standing problem and deliver solutions with significant value.

Qui Liu and Tencia Lee, the winners of the 2015-2016 Data Science Bowl, are an ideal example. Liu, a PhD in Physics with a hedge fund background, and Lee, a Math graduate and hedge fund trader, first met on the DSB discussion forum. Neither have a medical background. They were ranked eighth and ninth on the leaderboard respectively and decided they were stronger together than apart. This type of collaboration is common. Lee said, *"What drew me to the challenge was its inherent complexity. The data set had a lot of quirks that required us to think through the unique scenarios and redirect our algorithms multiple times. In the end, we came to a solution that proved viable for multiple data sets."* Their solution exceeded researchers' expectations by achieving near human-level accuracy. After the competition, Lee worked with researchers at NIH to implement the winning algorithm for further validation and testing.

Our aim from the onset was to provide cardiologists with an objective, reference diagnostic model, eliminating measurement bias. This reality is now taking shape. Since the competition closed in April 2016, NIH continues to test and disseminate the findings from the DSB competition.

Cardiologists from the NIH Heart, Lung & Blood Institute affirm the value of convening communities across data scientists and the medical industry for critical problem-solving. Dr. Michael Hansen of the NIH said this at the completion of the DSB 2: *"Analyzing MRI scans is traditionally an expensive and incredibly time-consuming process for doctors to do manually for*

real-time imaging and interventional procedures. If we are successful in automating the process, we open up new possibilities." [REF9]

The Competitions – DSB 3: Improving Lung Cancer Screening and Detection

This year (2017), DSB competitors used machine intelligence to fight lung cancer, which strikes 225,000 people in the U.S. every year and accounts for $12 billion in healthcare costs. Early detection is critical to give patients the best chance at recovery and survival. The large false positive rate from low-dose CT Scans have minimized the potential effectiveness of early detection diagnostic imaging tests of this type. With nearly 97% of the positive test results being False Positives, primary care physicians and medical specialists have been reluctant to pursue complex, expensive, or even invasive additional testing on a patient population in which 29 out of 30 of the patients with a "positive" diagnosis actually do not need anything of that sort. In those cases, the most likely causes for the false positive readings from the images are bruises or inflammation in the lungs, which are neither cancerous nor seriously consequential to the patient's health.

Using a data set of thousands of high-resolution lung scans provided by the National Cancer Institute, the DSB 3 participants developed algorithms that accurately determine when lesions in the lungs are cancerous. (Approximately 2000 images were in the training set, and about 500 images were in the validation test set.)

The top-placing team consisted of two researchers from China's Tsinghua University who have no formal medical background. The second-place team consisted of two software and machine learning engineers based in the Netherlands, one of whom came in third in the DSB 2. [REF10] The third-place team consisted of members who work for a Netherlands-based company Aidence that applies deep learning to medical image interpretation. The winning teams' algorithms will dramatically reduce the false-positive rate that plagues the current detection technology, get patients earlier access to life-saving interventions, and give radiologists more time to spend with their patients.

Keyvan Farahani, Program Director at the NIH National Cancer Institute said this about the DSB 3 competition: *"Reducing the false-positive rate of low-dose CT scans is a critical step towards making these scans available to more patients."*

Anthony Goldbloom, CEO of Kaggle, said this about DSB 3: *"This is one of the most important competitions Kaggle has ever hosted. Recent breakthroughs in deep neural networks may make it feasible to diagnose lung cancer from CT scans with higher accuracy than previously possible. The interest in this year's Data Science Bowl has been unprecedented for a competition of this size. The results are incredibly promising."*

Preliminary findings from DSB 3 suggest that the top 10 solutions outperform best-in-class models by 10% with a moderate (but clinically meaningful) decrease in false positives. Medical scientists at NIH (and their colleagues) are currently working to integrate ensemble models of

the winning solutions into clinical prototypes. A first step in getting the results into clinical use is integrating the algorithms into practical, open-source, clinic-ready software and systems. A new competition is now addressing exactly that challenge.

As a follow-on competition to DSB 3, on August 8 2017, the Bonnie J. Addario Lung Cancer Foundation launched the Lung Cancer Early Detection "Concept to Clinic" Challenge along with DrivenData.org. [REF11] The total prize purse is $100,000, to be distributed in stages to the top teams in amounts corresponding to their algorithm's ranking at the end of various stages of the competition. These prizes will be distributed to different categories of participants [REF12]:

- Data scientists to build out the machine learning algorithms;
- Software engineers to develop backend functions and data pipelines to run the tool;
- Engineers and designers to build out the user interface; and
- Community contributors to enrich the documentation, discussions and outreach.

The ultimate goal of the Addario Foundation's Lung Cancer Early Detection competition is to build an open source software application that puts the advances from the machine learning algorithms of DSB 3 into the hands of practicing clinicians – to take the results from concept to clinic: from machine learning theory to clinical practice.

Data Science Bowl Impact Statement

The DSB "Data Science for Social Good" competition is the largest and most prestigious of its kind, garnering involvement of tens of thousands of participants and advocates as well as corporate, academic and non-profit partners and sponsors. We provide here a quick summary of past competitions' impact.

In its first three years, the DSB has produced amazing results, expanding every year in different dimensions. In 2017, we achieved a 53% increase in the number of participants while the prize purse jumped four-fold and online engagement skyrocketed. In 2017, approximately 10,000 DSB participants joined together to form over 2000 teams, who contributed over 18,000 algorithm solutions, which employed an estimated 160,000 hours of volunteer effort.

The DSB taps into a growing, global, yet tightly focused data science community that is armed and dedicated with the tools, talents, and techniques to make an impact on our world.

With its global participation and promotion, the 3rd Data Science Bowl reached more than 28.12 million via social media in 2017, driving 165.83M online impressions and over 58,000 posts, 26,000+ site sessions on DataScienceBowl.com, and extensive media coverage.

The Future

The Data Science Bowl is a competitive but supportive virtual community where data science experts and novices alike can learn, expand their skill sets, and change the world as we know it.

At Booz Allen, we want to advance the art of data science by showing that the most pressing problems can be met by smart solutions when the right tools are implemented. In the coming years, sectors such as energy, defense, or finance may see their largest struggles deconstructed as part of a Data Science Bowl competition. We'll continue leading the charge.

We are always looking for new challenge problems requiring advanced learning algorithms to be applied to significantly complex datasets, in order to address (and hopefully someday solve) a major societal challenge. Please go to datasciencebowl.com, or contact a Booz Allen data scientist, to contribute your ideas for the next challenge. Specifically, we solicit your inputs in at least these three areas:

- What is the next big challenge we should tackle?
- How do we amplify the impact even more?
- How can we help the DSB participants (competitors) through technology, mentorship, improved collaboration, and/or other means?

Summary

Since the early days of the Big Data and Data Science movements, many data scientists have been devoted to applications of those data collections and advanced algorithms to solving societal problems. Subsequently, informal groups (e.g., meetups) and entire organizations have coalesced around particular approaches (e.g., data hackathons; and online competitions), societal needs, or application domains (e.g., persons; cities; and the environment). The range and depth of those applications have a tremendous reach: from human health to ocean health, from pet shelters to sheltering persons in need, from healthy lifestyle change to climate change, from smart cities to smart farms, from precision law enforcement to precision medicine, from improving commuter traffic to stopping human trafficking.

The world is data now. Data are the new asset to fuel insightful discoveries, smarter decisions, essential innovations, and better outcomes. And the winning approach to drive those insights, innovations, and outcomes is a combination of smart humans and smart algorithms: crowdsourced data science for social good.

The Data Science Bowl has reached into the previously untapped potential around the world among those who strive to deliver something greater than themselves. The DSB competitions have forged meaningful connections and productive partnerships within global communities that do not normally intersect – data scientists, technology enthusiasts, and domain-specific organizations – who believe passionately in open data crowdsourcing – to generate solutions for society's most challenging problems. It is a competitive yet supportive experience through which individuals can harness their passion, unleash their curiosity, and amplify their impact to effect change on a global scale.

"Solutions uncovered in the Data Science Bowl competition will have a lasting impact for our environment and food supply chain," said Booz Allen data scientist Roman Salasznyk after the

first DSB in 2015. Similar statements of purpose and value have been (and will be) said about the DSB in 2016, 2017, and beyond.

REFERENCES

[REF1] "Data Science For Good - For All," https://www.linkedin.com/pulse/data-science-good-all-kirk-borne

[REF2] "Ideas for Citizen Science in Astronomy," https://arxiv.org/abs/1409.4291

[REF3] "Longitude Rewards," https://en.wikipedia.org/wiki/Longitude_rewards

[REF4] Sobel, Dava, "Longitude: The True Story of a Lone Genius Who Solved the Greatest Scientific Problem of His Time" (Walker Books: 1995)

[REF5] "Galaxy Zoo: Morphological Classification and Citizen Science," https://arxiv.org/abs/1104.5513

[REF6] "Assessing Ocean Health at a Massive Speed and Scale," http://www.datasciencebowl.com/competitions/assessing-ocean-health-at-massive-speed-and-scale/

[REF7] "The Data Science Bowl's Impact Runs Deep," http://www.datasciencebowl.com/dsb_impact_runs_deep/

[REF8] "Transforming How We Diagnose Heart Disease," http://www.datasciencebowl.com/competitions/transforming-how-we-diagnose-heart-disease/

[REF9] "A Medical Perspective on the Quality of the Left Ventricular Volume and Ejection Fraction Measurements in the 2016 Data Science Bowl Competition," https://www.kaggle.com/c/second-annual-data-science-bowl/forums/t/19839/a-medical-perspective-on-the-quality-of-the-left-ventricular-volume-and

[REF10] "2017 Data Science Bowl, Predicting Lung Cancer: 2nd Place Solution Write-up," http://blog.kaggle.com/2017/06/29/2017-data-science-bowl-predicting-lung-cancer-2nd-place-solution-write-up-daniel-hammack-and-julian-de-wit/

[REF11] "Early Detection in Lung Cancer," https://concepttoclinic.drivendata.org/lung-cancer

[REF12] "Lung Cancer Early Detection Challenge: Concept to Clinic," https://biometry.nci.nih.gov/cdas/approved-projects/1559/