Data Science for Good: What Problems Fit?

Julia Koschinsky, Ph.D.
GeoDa Center for Geospatial Analysis and Computation
Arizona State University
P.O. Box 875302, Tempe, AZ 85287
(480) 965-7533
julia.koschinsky@asu.edu

1. ABSTRACT

Making sense of emerging sources of big, open, and administrative data has become paramount. This analysis assesses key characteristics of projects that are widely assumed to generate new and actionable insights and have social impacts. I review 72 use cases by prominent organizations in the "data science for good" community to determine the types of problems where data science techniques add value. The four main categories I identify are 1) improving data infrastructure by combining data with higher temporal and spatial resolution and automating data analysis to enable more rapid and locally specific responses, 2) predicting risk to help target prevention services, 3) matching supply and demand more efficiently through near-real time predictions for optimized resource allocation, and 4) using administrative data to assess causes, effectiveness and impact. In almost all cases, the insights that are generated are based on an automated process, are localized, in near real-time and disaggregated.

Keywords

Data science, problem, actionable insight, impact

2. THE CHALLENGE

Making sense of big data, open data, administrative data, social media data and combinations of these data has become paramount [1]. We are not only looking for insights but for actionable insights that can augment existing government and nonprofit practices. Government and nonprofits are not alone in figuring out this puzzle: According to a 2011 survey of 3,000 companies in 30 industries and 100 countries, for almost four of 10 respondents, "the leading obstacle to widespread analytics adoption is lack of understanding of how to use analytics to improve the business" [2]. Even if one can figure out how to gain not only new but actionable insights from data, another challenge is the translation of these insights into impacts. Insights need to be "closely linked to business strategy, easy for end-users to understand and embedded into organizational processes so that action can be taken at the right time." [2].

A key reason why it is non-trivial to translate data analytics results into insights and impacts for social good is that this effort requires collaboration across traditionally siloed disciplines, skillsets and departments with their own jargons, cultures and ways of thinking. In order to inform a social or public problem with data analytics, this problem needs to be defined from the perspective of the people making decisions that influence the problem resolution, so that technological and statistical solutions can ultimately be embedded within these decision workflows. Not all techniques in the toolboxes of computer and data scientists meaningfully inform public and nonprofit problems. Therefore the question arises which types of public and nonprofit questions and problems can be informed by data science.

Like technology, data science does not improve social outcomes by itself. At its best, it augments existing implementation processes [3]. Figuring out which organizational processes are particularly prone to such an augmentation and which agencies are ready to adopt or expand data-driven cultures is important to translating insights into impacts. And even when data analytics proves to make existing operational processes more efficient, the question still remains whether the outcomes are socially and politically desirable or not [4]. This skepticism is reflected in the growing criticism of *Minority Report*-style surveillance (informed by predictive modeling of crime such as in Chicago) and of the "governance of algorithms," which are often black-boxed and outside of the realm of accountability to residents [5, 6].

Another prominent critique of technological and data-driven data-for-good projects is that they promote a perspective that assumes that "there's an app" for every problem [7]. There is a tendency to offer band-aid solutions that, for instance, might help manage the process of serving a few homeless persons a little better but ignore long-term structural problems such as inequality, racial discrimination or shifts to lower paying service sector jobs that cannot be easily fixed with a civic tech tool or predictive model developed during a hackathon weekend.

As more cities are displaying the results of quantitative indicators on dashboards, critics point out that the choices about what data are collected, how indicators are measured, what goals they represent, and how they are displayed are innately political rather than merely technical [8, 9]. Arguments to "just let the data speak" as if they were objectively representing an independent truth are misleading.

3. QUESTION AND METHODOLOGY

This analysis assesses key characteristics of projects that are widely assumed to generate new and actionable insights and have social impacts. To do so, I conduct a preliminary review of 72 use cases by prominent organizations in the "data science for good" community to determine the types of problems where data science techniques add different kinds of value. I chose organizations that focus on data science methods such as machine learning or predictive modeling and impact measurement. Efforts that primarily specialize in the visualization of raw data or basic statistical analysis are not included. Related projects in civic tech (such as Code for America's) are also excluded since most of these projects focus more on technological advances to government problems than on data analytics.

Four organizations were chosen that represent well-known efforts in the data-for-good community: DataKind, Bayes Impact, the Data Science for Social Good (DSSG) Fellowship (University of Chicago), and New York's Mayor's Office of Data Analytics

Bloomberg Data for Good Exchange Conference.
28-Sep-2015, New York City, NY, USA.

(MODA). This is a convenience sample that is not designed to be representative. It will be extended over time as more projects are documented online. However, it is noteworthy that there is already substantial overlap in the questions and problems addressed by the four organizations, suggesting that the sample does effectively capture some common trends.

To identify the universe of use cases for this paper, I chose the projects listed on the organization's websites in mid-July 2015 (specifically the seven winning projects of the 2014 Bayes Impact 24-hour Hackathon[1], 23 projects on DataKind's project page[2], 38 fellowship projects of DSSG[3], and four projects from MODA's 2013 Annual Report[4]). Table 1 contains the categorization, name, agency, sponsor, problem, short description, data, and method used in each of these projects.

4. FINDINGS

4.1 Types of Problems

The four main categories that I identify to classify the 72 use cases in terms of the problems that data science helps to address are 1) improving data infrastructure by combining data with higher temporal and spatial resolution and automating data analysis to enable more rapid responses, 2) targeting limited resources to highest risks for prevention efforts, 3) matching supply and demand more efficiently through near-real time predictions for optimized resource allocation, and 4) using existing data to assess performance and impact. In almost all of the cases, the insights that are generated are based on an automated process, are localized, in near real-time and disaggregated. This section discusses these findings in more detail.

The range of problem areas of the use cases is so broad that the choice of problem area does not seem to be a constraining factor. How the problem is defined and what data are available within a given problem area seems to be more relevant. Common problem areas in the sample include health problems, non-completion of school or service programs, government corruption, human rights violations, neighborhood blight (abandoned properties), access to funding for nonprofits, poverty and homelessness, as well as government operations (such as fire, building codes, and policing).

One of the key differences between traditional quantitative analysis in the social sciences (e.g. using multivariate regression models) and the use cases analyzed here is that their reliance on machine learning methods comes with a shift in focus from descriptive to predictive and prescriptive insights [10]. Descriptive insights address the question what has happened and why; predictive insights focus on what could happen while prescriptive insights inform choices about how to respond to what has happened or could happen. The vast majority of use cases in this sample produce prescriptive or predictive insights, i.e. insights that are (or at least appear to be) actionable without requiring additional analysis. Projects often include the full pathway from identifying patterns in past and current data (e.g. why students dropped out of school in the past) to predicting future behavior (e.g. risk scores of who might soon drop out) to informing intervention strategies (e.g. to help prevent drop-outs).

The use cases general fit Santos' [3] criteria for actionable results: They a) inform a better-than-usual selection of response that b) can be implemented in a feasible and efficient way and that c) are related to an improved outcome. One of the reasons why the results are actionable is the reliance on disaggregated units of analysis rather than aggregates, which is driven by the increasing availability of electronically generated data at this scale. This disaggregation makes insights actionable at the individual level since it focuses the analysis on a unit that matches that of decisionmakers. An example is the disaggregation of smart meter readings for the total household to estimate the energy that could potentially be saved by individual appliances.

These are the four categories I identify to characterize the types of problems and value added by data science in the sample (letters also used in Table 1):

A. **Improving data infrastructure to enable faster and local responses**
by combining data with higher temporal and spatial resolution and automating data analysis to enable more rapid and locally specific responses

B. **Predicting risk to help target prevention services**
assisting a service provider with targeting of limited prevention resources based on prediction of elevated risk

C. **Detecting space-time clusters to help match supply and demand**
predicting optimized allocation of resources to better match supply and demand across a complex system

D. **Using administrative data to assess causes, effectiveness and impact**
improving services or systems through an assessment of effectiveness based on analysis of existing administrative data or combining past data on process and outcome

These categories are not mutually exclusive: The same use case can be part of multiple categories at different stages. For instance, a case could start with building a data infrastructure, then proceed with estimating elevated risks in sub-samples and conclude with assessing the effectiveness of service delivery. To reflect this, some use cases in Table 1 are classified in more than one category. This following sections illustrate each of these categories with examples (see Table 1 for more details).

Improving Data Infrastructure to Enable Faster and Local Responses

Several use cases pertain to automating and centralizing data access, usually for data that are more location-specific and timely than traditional censuses (although often not as complete). Examples include the creation of a central atlas of businesses in New York after Hurricane Sandy as part of post-disaster aid and the merging of data for different medications to identify negative side effects of interactions between them. There are also examples of improving automated measurement, e.g. testing the measurement of poverty with proxies from satellite images such

[1] http://bayeshack.devpost.com/submissions

[2] http://www.datakind.org/projects/

[3] http://dssg.io/projects/

[4]http://www.nyc.gov/html/analytics/downloads/pdf/ annual_report_2013.pdf

as roof type or light patterns; or generating inflation estimates in near-real time with greater local accuracy through web scraping. This data infrastructure and analysis serves as the foundation for work in other categories, such as targeting resources or assessing program effectiveness. For instance, the automated process of identifying roof types was used to help a nonprofit allocate resources to households with less wealth (measured by living in homes with thatched as opposed to metal roofs).

Predicting Risk to Help Target Prevention Services

A very common use case that all four organizations worked on is related to assisting stakeholders with prioritizing where limited prevention resources should be targeted based on predicted elevated risk of an undesirable outcome. The typical scenario here is that a nonprofit or government intervenes to address a problem such as a fire, structurally unsafe home, school non-completion, or human rights violation. Interventions involved sending fire and housing inspectors to homes to detect unsafe conditions in advance, having school counselors help at-risk students stay on track for graduation, or using publicity campaigns to put public pressure on representatives to stop human rights violations as they were automatically tracked in near-real time. The problem in these cases is that the intervention is limited to a sample of the total population due to resource constraints. Data science is used to first model the determinants of elevated risk based on past incidents and then use the results from these models to predict risks for new cases. As part of an early warning system, an estimated prediction of highest risk helps identify which sub-groups to prioritize.

Use case examples include the estimation of risk scores for individual students (for dropping out of school or college undermatching), patients (for adverse health outcomes such as obesity or maternal mortality), offenders (for re-committing domestic violence) or officers (for police brutality). Other examples quantify the risk that a particular home will be abandoned, catches fire, violates housing codes, or contains lead; that an accident occurs in a mine; that a company fails to report hazardous waste; or that an area is subject to concentrations of 311 complaints or disaster impacts. A good example of an early warning system is a use case where phone-based sensors to track the temperature of vaccine coolers in Africa are used to predict power outages (DataKind, NexLeaf and MedicMobile). This information allows stakeholders to intervene in advance and prevent vaccine spoilage. A related group of use cases is the search for targets with elevated risk of engaging in illegal activities, providing enough information to identify existing or emerging targets. In this case data patterns are used to identify potential child prostitution rings by location and phone number, probable fraud in contract bidding by contractor, or likely money laundering by particular businesses.

Detecting Space-Time Clusters to Help Match Supply and Demand

In a set of related use cases data analytics helps to match supply and demand more efficiently by optimizing the allocation of resources across an entire complex system in near-real time. In these cases there is typically a supplier of a good or service and a population of people or entities using this service. In contrast to the previous example, the goal here is not to identify an at-risk sample within this population but to more optimally distribute goods or services across the entire population. The problem here is that the use of services varies across space and time, leading to mismatches between supplied services and demand for them. Data analytics identifies when these concentrations in demand are likely to occur to help optimize the matching supply of services. For instance, such matching can occur between buses and riders to avoid overcrowding; between shared bikes and stations to aid rebalancing; between employers and job seekers aided by training programs; between surplus food and nearby food banks; or between hospices and terminally ill children. In one example, traditional federal labor census data were supplemented with online data streams from employment sites to provide workforce agencies with more localized and timely information on how to better target their services (DSSG, Department of Labor, CareerBuilder and Skills for Chicagoland's Future).

Using Administrative Data to Assess Causes, Effectiveness and Impact

Another common use case pertains to program or systems evaluation and impact measurement using existing data collected for other purposes. The goal here is to improve a sponsor’s service delivery through an assessment of the effectiveness of its services based on administrative data they or others have been collecting as part of service delivery. This is more affordable than to collect new data solely for evaluation purposes, as is often the case. In many of these examples services were delivered with the help of technology (through phone message texting, mobile apps, or websites). As a result, databases of users and a log of the service delivery itself were collected automatically. The question to be addressed through data science is how services can be delivered more effectively, e.g. to prevent dropping out of services, avoid disconnects between mentors and students, or improve a provider’s chance of obtaining crowd-sourced funds online. Examples include the mining of text messages or other electronic administrative data to identify successful practices (e.g. nonprofit fundraising) or interactions (e.g. between tutors and students, crisis mentors and teens) among some stakeholders that can be used to improve sub-optimal practices or relations among others.

In related examples of impact measurement, DataKind helped the health nonprofit Nurse-Family Partnership determine that its programs increased children’s vaccine rates. It also assisted the New York City Parks Department in an evaluation of the impact of tree pruning, which turns out to reduce the risk of hazardous tree conditions during storms in New York City. Furthermore, new small businesses that utilized the New York City Business Atlas were able to open two and a half months earlier than those not utilizing the atlas (MODA).

Particularly interesting examples of program or systems evaluation combined data sources in a way that connected an outcome with data on the process that helped generate this outcome. This connection provides insights into the full life cycle of a problem to aid program or system reform: For instance, one use case designed more effective homeless prevention services by linking data from two service providers that were assisting people before and after they became homeless. Another case merged data on financial contributions with that on voting behavior to document the influence of contributions on voting decisions and track this relationship for each political representative. When influence on voting is purchased, this documents a conflict with how the political process is supposed to function, warranting reform. A third example combined judgments on human rights violations by the European Court of Human Rights with data on whether the judgment was enforced. Again, here the effectiveness of judgments is assessed in terms of their implementation. One of the reasons why these connections had not been made before is because of the technical challenge of combining datasets without common IDs, data structures and formats.

4.2 Sponsors and Data

All but three projects had a sponsoring agency, which typically defines the problem or question, often shares internal data, and uses the results. Some projects are designed to automate the process of generating analytic results, so the process can be integrated with day-to-day operations (e.g. DataKind's project with Benetech or MODA's project with New York’s fire department).

Almost all projects are based on individual-level units of analysis that can be identified by their characteristics and/or location (such as persons, businesses, parcels, bikes, contracts, or organizations). A few projects analyze aggregated data to identify overall trends in a country or region, e.g. in risk factors of maternal mortality (sponsored by Mexico's Office of the President and implemented by DSSG), mismatches between terminally ill children and hospice services in the U.K. (DataKind), or child poverty across a city or country (DataKind's work for DC Action for Children and North East Child Poverty Commission). Some combine aggregated with disaggregated data to first target a geographic area and then identify particular individuals at risk within these areas (e.g., DSSG's analysis to help the Illinois Department of Human Services target its services for improved birth outcomes).

Since DSSG is located in Chicago, many of DSSG's projects are conducted in collaboration with sponsors in Chicago. Similarly, all of MODA's projects are in New York City although other cities have started to adopt some of their best practices (e.g. the City of New Orleans improved its distribution of smoke detectors through estimates of fire risk that were informed by MODA's efforts and aided by Enigma.io). DataKind has the largest share of international projects, including several projects in Africa, Southeast Asia, and the UK (reflecting the work of its international chapters in the UK, Bangalore, Singapore, and Dublin). DSSG worked with international sponsors from Mexico, Qatar, Australia, and Costa Rica. Especially the projects sponsored by human rights nonprofits had a global focus (e.g. by Amnesty International, Ushaidi, or Benetech).

Half of the sponsors were nonprofits. Many of these nonprofits shared administrative data that they had collected for operational purposes and were now looking to re-analyze to answer other questions. For instance, Bayes Impact used logs from DonorsChoose, a crowdfunding platform for teachers, to predict the funding success of proposed projects as they are being proposed to allow teachers to adjust their project descriptions in real time (DataKind also worked with this sponsor but not as part of their listed projects). Another example mentioned before is NexLeaf, a nonprofit that uses cell phones to monitor temperatures in order to prevent vaccine spoilage due to power outages in Africa. DataKind helped develop an early warning system using NexLeaf's text message data.

Of the other half of sponsors, 36% were government agencies. The majority were local governments, including many police departments and school districts that shared internal data for re-analysis. Health departments, transit agencies, and planning departments (concerned with addressing housing abandonment) were also frequently represented, often using a combination of internal and open city data (such as 311 or parcel data). New York's Mayor's Office of Data Analytics specialized in helping city agencies improve operations through analytics with internal and open data from the fire department, buildings, and emergency management. One of the challenges for managing an open data infrastructure like New York’s is related to data quality and decentralized cleaning of data. Data-for-good projects are one of many efforts involved in cleaning open data but there is often no centralized way to access cleaned data, annotated metadata or code/workflows to clean updated versions of the same data. As a result, there is a lot of duplication of efforts that could be streamlined more.

The federal Department of Labor also sponsored a project to estimate risks of mining accidents to prioritize targeting of inspector visits (Bayes Impact). In another project they sponsored DSSG started developing an open source real-time labor market information system to help match local labor demand and education/job training more efficiently. The only private agency in the sample was the World Bank (4 projects), with two projects on detecting fraud in contract bidding (DSSG), one on automating the measurement of poverty through detection of light patterns and one on measuring inflation faster and more locally through scraped web data (both DataKind). Three projects were sponsored by public-private/nonprofit partnerships, for instance involving collaborations between the City of Memphis and community-based organizations to address blight and housing abandonment (DSSG). In Chicago, Skills for Chicagoland's Future received data from the online employment site CareerBuilder to assess which additional skills job seekers need to secure a job (DSSG).

The three projects without a sponsor were general public interest projects. In two cases they were based on open data, to predict side effects of drug interactions and identify potential money laundering (both Bayes Impact). In one case users had to submit their own building energy use data to obtain recommendations for where energy could be saved (DSSG).

4.3 Methods and Tools

Machine learning or predictive modeling was the most frequently used method in the sample, especially for predicting/preventing risk and for better matching supply and demand (categories B and C). Common examples of specific methods include decision tree learning, optimization and cluster algorithms, natural language processing, and neural networks. One of the most common use case was the generation of scores to estimate risks. Exploratory data analysis and data or map visualizations were also used often, sometimes in preliminary analyses or to present and communicate the results. Several projects conducted remote sensing analysis to classify images for improving the measurement of poverty. Propensity score matching was most often used for those projects seeking to measure impacts.

Many project descriptions (e.g. DataKind's) do not contain details about the tools that were used for the analysis. Hence the information summarized here is based on the small subset of projects that does mention what tools were applied. One of the noteworthy trends that emerged from the assessment of this subset is the common use of open-source tools and open-source code, which enables the sharing of code and replication of the analysis in other places and contexts. Many projects make their code available on GitHub. Since all groups (except for Bayes Impact’s hackathon) are engaged in ongoing data science projects, there are multiple efforts to scale existing projects and replicate them in other domains, coding events, or cities. These efforts are aided by platforms to help organizations replicate data-for-good projects (such as the datalook.io website).

Projects relied on 1) Python tools (e.g. pandas for data processing and analysis, scikit-learn for machine learning, NLTK for natural language processing, scraperwiki for scraping web pages, matplotlib for statistical graphics, and the flask web framework); 2) Data-Driven Documents (d3.js javascript library) for web-based data visualization; 3) R packages for statistical analysis (e.g. gbm (Generalized Boosted Models), ggmaps for statistical plots, and Matching for propensity score matching); and 4) cloud mapping solutions such as Mapbox and OpenStreet Map for web mapping, Foursquare’s Quattroshapes for geocoding, and Elasticsearch gazetteer for geographic indexing.

4.4 Limitations

This analysis was designed as a preliminary overview of some important state-of-the-art projects in the data science for good community in the U.S. and some of its international partners. It is characterized by several limitations: In many cases, the project descriptions were very short (e.g. less than a page), which could result in misclassifications (e.g. regarding project scope or organizations involved in the project) and has led to missing data (e.g. on methods and tools that were applied). Since some of the 2015 DSSG projects are still ongoing, a more detailed project description will only become available later in 2015. Further, since the categorization of projects is part of a first preliminary review it is somewhat ad hoc and subjective. It will be tested and adjusted with additional use cases in the near future.

5. CONCLUSION

Of all of the use cases in the sample, the ones that are most controversial are those closest to surveillance and most at risk of violating privacy rights. Many of the crime-related examples fall into this category. For instance, Chicago's police department has been criticized for sending officers to the homes of potential offenders based on modeling results, which suggested that these residents could be likely to commit an offense. The possibility that the intervention itself might further facilitate the undesired outcome is often ignored here. But even the estimation of which students are at risk of dropping out of school or services could be controversial depending on how the intervention in response to the risk estimation is structured (e.g. whether students are informed about the availability of services with or without letting them know that they are estimated to be at risk of dropping out).

The extent to which data-science-for-good insights are translated to actual impacts is related to how much a project's solution is embedded within the decision processes of institutions that have impacts. Ironically, this integration and related project scoping and networking tends to take much more time than the technical implementation of a data-driven solution. The four agencies that implemented the 72 projects differ in their approach: On the one hand, Bayes Impact conducted a 24-hour hackathon. Within this framework, there is no time for an in-depth understanding of the agency processes that a data-driven solution could augment.

On the other hand, Mike Flowers who directed MODA under the Bloomberg administration spent a significant amount of time understanding workflows, networking with stakeholders and building trust to prepare for a meaningful integration of data-driven strategies and outcomes within agencies. As expected, the hackathon results run the risk of having a short-lived impact unless someone continues the development in another forum (e.g. many of the Bayes Impact project links are already broken). On the other hand, MODA's impact, for instance, on the New York fire department and New York's open data infrastructure (DataBridge) have been sustained because they were institutionalized.

DataKind and the Data Science for Social Good fellowship fall in-between these two examples. While DataKind also conducts hackathons, they employ full-time staff to spend several months scoping projects with a focus on sustained impacts after the volunteer contributions. In addition, the organization has been moving towards longer-term engagements with their own data scientist staff (Data Corps). The DSSG fellowship engages data scientists for three months. It also provides staff support for project scoping and management to ensure that project results are relevant to sponsoring agencies and that they are integrated within a sponsor's decision processes. Several projects are extended over multiple summers.

The "killer app" critique [6] is related to the extent of this integration effort: Arguably, the more non-technical work is invested in the integration of the sponsor's decision processes (understanding the process, networking, trust building, etc.), the more data-driven solutions have a chance of improving the delivery of infrastructure services (like 911 response times, fighting fires or issuing licenses to new businesses more efficiently) or nonprofit services (like preventing vaccine spoilage or disseminating micro grants).

The political dimensions of the choices embedded in data-driven solutions deserve more room in public discussions. For instance, when we rely on social media to aid disaster response, do we inadvertently end up prioritizing areas where people tweet more, potentially shortchanging areas on the wrong side of the digital divide like those with more seniors or low-income residents? This was one of the lessons from civic tech where apps to report pot holes to the city resulted in lots of reports from whiter wealthier areas, which actually had fewer potholes than other areas but a higher propensity to report them. At the same time, new civic tech efforts to improve large-scale sensor data collection related to infrastructure promise to not only advance the efficiency but also the equitable distribution of infrastructure improvements [11].

One of the goals of this preliminary meta-level analysis of data-science-for-good projects is to aid the discussion of how our technical and statistical solutions are related to these larger questions in order to ensure that the impacts remain sustainable, equitable and address privacy concerns.

6. REFERENCES

1. Chen, H., Chiang, R. H. L., Storey, V. 2012. Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly* 36, 4 (December), 1165-1188.
2. LaValle, S., Lesser, E. Shockley, R., Hopkins, M.S., and Kruschwitz, N. 2011. Big Data, Analytics and the Path From Insights to Value. *MIT Sloan Management Review* 52, 2 (Winter), 21-31.
3. Boba Santos, R. 2014. The Effectiveness of Crime Analysis for Crime Reduction: Cure or Diagnosis? *Journal of Contemporary Criminal Justice* 30, 147-168.
4. Toyama, K. 2015. *Geek Heresy: Rescuing Social Change from the Cult of Technology.* PublicAffairs, New York, NY.
5. Townsend, A. 2013. *Smart Cities: Big Data, Civic Hackers, and the Quest for a New Utopia*. W. W. Norton & Company, New York, NY.
6. Boyd, D., Levy, K. and Marwick, A. 2014. The Networked Nature of Algorithmic Discrimination. *Data & Discrimination: Collected Essays* (Eds. Seeta Peña Gangadharan and Virginia Eubanks), 43-57.
7. Morozov, E. 2014. *To Save Everything, Click Here. The Folly of Technological Solutionism*. PublicAffairs.
8. Mattern, S. 2015. Mission Control: A History of the Urban Dashboard. *Places Journal*, March.
9. Kitchin, R., Lauriaulta, T. P., McArdle, G. 2015. Knowing and governing cities through urban indicators, city benchmarking and real-time dashboards. *Regional Studies, Regional Science* 2, 1, 6–28.
10. IBM Corporation. 2013. Descriptive, predictive, prescriptive: Transforming asset and facilities management with analytics. IBM Software Thought Leadership White Paper (October).
11. Adibhatla, V., Henke, G. and Atwater, P. 2015. *Street Quality Identification Device.* ARGO Labs Working Paper.

Table 1. Summary of 72 Data-Science-for-Good Projects that were Reviewed

Cat	Project Name	Agency	Sponsor	Problem	Short Description	Data	Methods
A	Scraping Websites to Collect Consumption and Price Data	DataKind	World Bank	Measuring inflation locally + timely	Automate access to online data about inflation in Africa.	Data scraped from web	data visualization
A	Want to Save 268 Days of Data Acquisition?	DataKind	Microfinance Information Exchange	Access to info for micro loans	Automate access to online data relevant for microfinancing.	open online microfinance data	scraping data from web
AD	Driving Small Business Growth with Analytics	MODA	New Business Acceleration Team (NBAT)	Fragmented data sources	Centralize data without common ID to create a census of businesse in New York. Measure impact of time saved by system.	Dept. of Consumer Affairs, Dept of Health + Mental Hygiene, Dept of Environmental Protection, tax parcels	data and map visualization, impact measurement
AB	Using the Simple to Be Radical	DataKind	GiveDirectly	Measuring poverty	Automate the identification of poor villages who are eligible for mobile phone-based cash transfers.	satellite images	satellite image processing and machine learning to differentiate roof types
A	Shining a Light on Poverty	DataKind	World Bank	Measuring poverty	Automate the measurement of poverty with light data.	satellite images	satellite image analysis, correlations
A	Mapping Poverty to Beat It	DataKind	DC Action for Children	Measuring poverty	Convert PDF documents of child well-being indicators to interactive online maps.	KIDS COUNT	data and map visualization
A	Delving into Child Poverty Data	DataKind	North East Child Poverty Commission	Measuring poverty	Give NECPC a more real time understanding of child poverty and communicate the data in a more actionable way to encourage immediate responses.	Citizens Advice Bureau	data and map visualization
A	Drug Safety in Your Pocket	Bayes Impact	No Sponsor	Unknown drug interactions	Predict novel interactions for pairs of drugs that do not have a historical interaction record.	AERS (Federal Drug Adverse Event Reporting System) dataset	data classified by the RxNorm hierarchy; neural networks
A	Case Foundation: A Hairball to Help Non-Profits Untangle Strategy	DSSG	Case Foundation	Nonprofit funding	Identify networks of similar nonprofits and funders that support them.	Data scraped from web (tweets, news)	natural language processing (TF-IDF) and social network analysis
B	Predicting College Persistence among High School Students	DSSG	KIPP Chicago - Public Schools	Edu: College non-completion	Predict a student's risk of struggling in college.	Data from KIPP Chicago - College Prep Public Schools	predictive modelling
B	Mesa Public Schools: Undermining Undermatching	DSSG	Mesa public schools	Edu: College undermatching	Predict gifted students likely to undermatch in college enrollment.	Data from Mesa public schools and college test scores	predictive modelling
B	Early Warning Systems for Struggling Students	DSSG	Montgomery County Public Schools	Edu: High school non-completion	Improve existing early warning system to identify high school students at risk of not graduating.	Data from Montgomery County Public Schools	risk scores based on predictive modelling (Poisson regression, random forest, ordinal regression trees, Cox regression)
B	Identifying High School Students Who May Not Graduate on Time	DSSG	Wake County and Arlington Public School Districts	Edu: High school non-completion	Scale existing early warning system to identify high school students at risk of not graduating.	school district data	risk scores based on predictive modelling
B	Keep In Touch: Robust Retention Strategies for Health Leads	DSSG	Health Leads	Health: Disconnect from services	Identify patients' risk of dropping out of health-related service provision.	Data from Health Leads	predictive modelling
B	IDHS Project: Better Birth Outcomes	DSSG	Illinois Department of Human Services	Health: Adverse births	Identify women at greatest risk of adverse birth for targeting support services.	Data from Illinois Department of Human Services	machine learning, predictive modeling, mapping visualization
BD	Defining the Undefinable, Measuring the Unmeasurable	DSSG	Nurse-Family Partnership	Health: Adverse births	Determine risk factors associated with dropping out early or not achieving program goals.	Data from Nurse-Family Partnership	impact evaluation
B	NorthShore: mining medical data to tackle the obesity crisis	DSSG	NorthShore University Health System	Health: Child obesity	Predict obesity risk for a child based on hospital's medical records to guide health interventions.	NorthShore University Health System's Electronic Medical Records	linear regression
B	Predictive Modeling for Public Health: Preventing Childhood Lead Poisoning	DSSG	Chicago Department of Public Health	Health: Lead contamination	Identify homes at high risk of lead contamination.	blood test results, building and inspection records, census data	machine learning; claddification algorithms (logistic regression, support vector machines, random forests)
B	Maternal Mortality in Mexico: Distilling Data into Policy Strategies	DSSG	Mexico's Office of the President	Health: Maternal mortality	Determine the risk factors associated with maternal mortality.	birth and death, patient discharge records, hospital data, Census data - back to 1990	exploratory analysis and predictive modeling (logistic regression, decision tree, random forest)
B	Warm Calls and Persuadability: Enroll America	DSSG	Enroll America	Health: Uninsured	Predicting an individual's probability of enrolling in subsidized health insurance.	Data from Enroll's GetCoveredAmerica campaign	correlation matrix, Lasso regression
B	Keeping it Cool: Using Mobile Technology to Preserve Vaccines	DataKind	NexLeaf	Health: Spoiled vaccines	Decrease spoiled vaccines due to power outages through phone-based temperature monitoring.	Internal data from NexLeaf	data visualization, correlation analysis

B	Improving Long-Term Financial Soundness by Identifying Causes of Home Abandonment in Mexico	DSSG	Infonavit	Hsg: Abandoned properties	Determine the risk factors associated with housing abandonment.	Data from Infonavit, census and home surveys, loans	exploratory data analysis
BD	Easing the Distress of Neighborhoods with Data	DSSG	City of Memphis and CDCs	Hsg: Abandoned properties	Determine risk that a home becomes abandoned.	unemployment, poverty, income; and real estate data	clustering algorithm, random forest classifier, hedonic regression, propensity score matching
BD	Proactive Blight Reduction and Neighborhood Revitalization	DSSG	City of Cincinnati	Hsg: Abandoned properties	Early warning system for when and where properties are likely to become blighted.	City of Cincinnati data	predictive modelling and impact analysis
BA	Cook County Land Bank Part 2: A Real Estate Finder for Vacant Properties	DSSG	Cook County Land Bank Authority	Hsg: Abandoned properties	Suggest which abandoned properties should be prioritized for purchase to avoid further neighborhood decline.	parcel and neighborhood data (crime, demographics and socio-economics)	predictive modelling
B	An automated filter for the Department of Buildings (DOB) B+ program	MODA	Department of Buildings	Building code violations	Estimate risk of building violations to prioritize building inspections.	Department of Building and 311 data	predictive modelling
B	Predictive Analytics for Smarter City Services	DSSG	City of Chicago	City complaints	Identify areas and times with higher risk of complaints about graffiti, potholes, etc.	City of Chicago 311 data (open)	predictive modelling
B	Predictive analytics of crime	DSSG	Chicago Police Department	Crime	Detect emerging crime problems at daily level to allocate officers more effectively.	Data from Chicago Police Department and University of Chicago's Crime Lab	predictive modelling
B	Out for Justice: A decision support system for police departments	Bayes Impact	San Francisco Police Department	Crime	Help San Francisco police department (SFPD) optimize patrol car placement.	SFPD 911 data and OSM street data	machine learning (boosted Poisson regression trees); optimization algorithms; predictive model
B	Preventing Domestic Violence with Data-driven Action	Bayes Impact	High Point, NC Police Dept.	Crime: Domestic violence	Prevent recidivism among domestic violence offenders.	High Point Police Department's data	predictive modelling
B	Detecting and Visualizing Prostitution Rings	Bayes Impact	Thorn	Crime: Child prostitution	Help discover and geolocate previously unnoticed prostitution rings in U.S.	Data scraped from web (adult posts)	soft text matching
B	How Network Analysis Can Help Identify Money Laundering Schemes	Bayes Impact	No Sponsor	Crime: Money laundering	Identify individuals and businesses that could be laundering their money.	open data (UK business registries, incl. offshore companies)	network analysis
B	Identifying Fraud & Collusion in International Development Projects	DSSG	World Bank Group	Crime: Fraud in contract bidding	Identify likely fraud in contract bidding.	International contract bidding data from World Bank Group	predictive modelling
B	Clean Development: Data Mining for Corruption Risks	DSSG	World Bank Group	Crime: Fraud in contract bidding	Identify likely fraud in contract bidding.	International contract bidding data from World Bank Group	predictive modelling
B	Early Warning Indicators for Adverse Police Interactions	DSSG	Charlotte-Mecklenburg PD	Crime: Police brutality	Develop early warning system to flag officers at risk of engaging in adverse interactions.	Charlotte-Mecklenburg Police Department data	predictive modelling + early warning systems
B	Using Data for a More Transparent Government	DSSG	Harris School of Public Policy	Government corruption	Automatically detect which congressional allocations were earmarked.	Congressional documents	machine learning (Support Vector Machine, Name Identity Recognizer)
B	Predictive Enforcement of Pollution and Hazardous Waste Violations	DSSG	Energy Policy Inst., University of Chicago	Hazardous waste violations	Predict a company's risk of severe environmental violations.	EPA data: reporting, monitoring, inspection, enforcement (RMP + RCRA)	predictive modelling
B	QCRI: Tapping Twitter for Faster Disaster Relief	DSSG	Qatar Computation Research Institute	Disaster response	Use tweets to aid disaster response in near-real time.	Tweets	mapping, classification, extraction (CRF), clustering/merging (SDLA algorithm)
BA	Disaster Response and Recovery	MODA	Office of Emergency Management	Disaster response	Analyze 911 and 311 data in near-real time to aid faster disaster response.	Various city data sources, city's 311 and 911 data.	geocoding, pattern analysis, data integration
BA	FDNY's Risk Based Inspection System (RBIS)	MODA	FDNY	Fire	Forecast risk of fire to prioritize fire inspections.	Data from FDNY and DataBridge (housing)	predictive modelling
BD	Out On a Limb - For Data	DataKind	NYC Parks Department	Hazardous tree conditions	Determine if tree pruning reduces risk of hazardous tree conditions during storm.	Internal data from NYC Parks Department	data visualization, impact analysis and predictive modeling
B	Mine Risk Evaluator	Bayes Impact	Department of Labor	Mine accidents	Help mine inspection managers to prioritize their next inspection.	Department of Labor data	predictive modelling
B	Energywise	DSSG	No Sponsor	Energy conservation	Assess which actions result in high energy use that could be reduced.	User submits building energy data	predictive modelling
B	Making Smart Meters Smarter	DSSG	Pecan St., Elevate Energy, Oak Park	Energy conservation	Address challenge of load disaggregation to estimate energy use of individual appliances.	Pecan Street and Oak Park (ComEd, Smart Cities, Green Button, Direct, ISEIF) data	neural networks, hidden Markov models, sparse coding
B	Ushahidi: Machine Learning for Human Rights	DSSG	Ushaidi	Human rights violations	Automatically identify high-risk situations in real-time.	Text message data from Ushaidi	machine learning and natural language processing

B	Predicting and preventing human rights abuses	DataKind	Amnesty International	Human rights violations	Automatically identify high-risk situations in future based on patterns of past urgent messages.	Amnesty International internal data	text analysis and predictive modelling
BA	Strengthening Global Human Rights Through Mapping	DataKind	Benetech	Human rights violations	Automatically flag and map concentrations in human rights violations in near-real time.	Benetech data (collected through Martus)	text analysis and automated reports/maps
BA	Predicting and Preventing Nonprofit Financial Default	DataKind	GuideStar	Nonprofit default	Predict risk of nonprofit financial default.	Supplemented GuideStar data (operations) with GreatNonprofits (nonprofit reviews)	predictive modelling
C	Divvy Part 2: A Crystal Ball for Rebalancing	DSSG	Divvy	Unbalanced shared bikes	Rebalancing bike-share bikes to avoid supply-demand mismatches at a station.	Internal data from Divvy	predictive modelling (Poisson regression)
C	CTA: Why Bus Crowding Happens and How Data Can Help	DSSG	Chicago Transit Authority	Unbalanced bus loads	Optimize allocation of buses to avoid overcrowding.	Data from Chicago Transit Authority	bus service simulations to forecast impact of adding or removing service
C	Identifying New Opportunities for Food Bank Donation from Food Service Retail	DSSG	Feeding America	Food waste vs. hunger	Identify total pounds of surplus food within a service area of a food bank.	Data about local foodservice channels from Feeding America	estimation of total amount of food available from different sources
CA	Improving Local Labor Market Matching with High Frequency Resume and Jobs Data	DSSG	Department of Labor	Skills gap	Develop real-time labor market information system to close skills gap.	federal, local, private, and public business and labor market datasets	labor demand models and skills gap analysis
C	Identifying Skills Gaps to Reduce Unemployment	DSSG	Skills for Chicagol.'s Future; Careerbuilder	Skills gap	Identify which additional skills job seekers need to find a job.	Data from CareerBuilder	text analysis
CA	Finding 30,000 Missing Children	DataKind	Shooting Star Chase	Unbalanced match terminally ill children-hospices	Match terminally ill children with nearby hospices with extra capacity.	Integrated data from Shooting Star Chase with public data on hospice and healthcare sector and demographic data	multi-layer map visualization
C	Anticipating Back to School Numbers, Before Summer Vacation	DSSG	Chicago Public Schools	Unknown student-school match	Estimate in spring how many students will enroll in a school in fall to save costs.	City of Chicago and Chicago Public Schools	data and map visualization, machine learning
DC	Insight for DonorsChoose	Bayes Impact	DonorsChoose	Funding: Nonprofits	Help teachers understand the best way to get their projects funded on DonorsChoose.	Data from DonorsChoose	predictive modelling
DC	Helping Great Causes Get Funded	DataKind	Global Giving	Funding: Nonprofits	Improve online crowd-funding for nonprofits through analysis of successful organizations.	Internal data from Global Giving	decision trees and regression analysis
D	Clustering Arts Organizations to Help Them Thrive	DataKind	Cultural Data Project	Funding: Nonprofits	Improve training based on better understanding what helps arts organizations succeed.	Financial and programmatic data from 11,000+ arts and cultural institutions	cluster analysis to identify peer communities
DC	Australian Conservation Foundation Project: Engage and Protect	DSSG	Australian Conservation Foundation	Funding: Nonprofits	Improve effectiveness of fundraising through data analysis and experiments.	Internal data from Australian Conservation Foundation	data exploration, clustering techniques, predictive modeling, experiments
D	Going Mobile	DataKind	Grameen Foundation	Funding: Access to micro loans	Improve effectiveness of text message help for subsistence farmers.	text message data from Grameen Foundation	text analysis
D	Understanding Text data to Help Disadvantaged Families	DataKind	Buttle UK	Funding: Access to micro loans	Improve provision of micro grants based on analysis of past patterns.	Buttle UK reports	natural language processing
D	Learning from Text Messages to Help Teens in Crisis	DataKind	Crisis Text Line	Teen crises	Improve text-based teen crisis intervention through analysis of existing patterns.	Crisis Text Line internal data	text analysis
D	A Different Kind of House Call	DataKind	Mobilizing Health	Health: Access in poor rural areas	Improve doctor-patient e-relationship through analysis of existing patterns.	Mobilizing Health internal data	text analysis
D	The Match Game: Measuring the National Impact of Nurse-Family Partnership	DSSG	Nurse-Family Partnership	Health: Adverse births	Determine if program is effective at improving health outcomes for mothers and babies.	Data from Nurse-Family Partnership, National Immunization Survey	propensity score matching
D	Improving Access to Education by Supporting Tutors	DataKind	The Access Project	Edu: Low-income student access: best univ.s	Improve tutor-student relationship through analysis of existing patterns.	The Access Project internal data	text analysis and data visualization
D	Uncovering the ABCs of Successful Online Mentoring	DataKind	iCouldBe	Edu: High school non-completion	Improve e-mentoring drop-out prevention program based on analysis of past patterns.	Internal data from iCouldBe	predictive modelling
D	Tracking the Paths of Homelessness	DSSG	Chicago Alliance to End Homelessness	Homelessness	Identify which housing types are most effective at providing housing stability.	Anonymized data from Chicago Alliance to End Homelessness (HMIS)	data visualization (Sankey diagram)
DA	Sharing data to learn about homelessness	DataKind	St Mungo’s Broadway	Homelessness	Improve provision of services to homeless residents by understanding what services they need before they become homeless.	Linked data from St Mungo’s Broadway and Citizens Advice	data and map visualization, network analysis
DA	Tracing Policy Ideas From Lobbyists Through State Legislatures	DSSG	Sunlight Foundation	Government corruption	Determine to what extent lobbyists are writing legislative bills that are considered for adoption.	Data scraped from web and Sunglight Foundation data	text analysis and impact analysis

DA	Let the Sun Shine on Politics	DataKind	Sunlight Foundation	Government corruption	Determine if financial contributions influence political representatives' votes.	OpenSecrets lobbying database + Sunlight Foundation data on fundraising activities of politicians	affinity score metric to track relation between donations and voting over time
DA	Shining Light on International Human Rights Case Law	DataKind	HURIDOCS	Human rights violations	Allow HURIDOCS to track whether human rights case judgments were enforced.	Integrate ECHR HUDOC database (Caselaw Analyzer) with Council of Ministers data on case execution.	scraping data from web and integrating it
D	Heatmaps for Habitats: Enriching Conservation Sensor Data	DSSG	TEAM Network	Species conservation	Determine how temperature changes affect species movement.	TEAM data	radial basis interpolation, heatmaps