

How Are Data Scientists Contributing to Social Good Right Now?

Originally published at <https://www.iflexion.com/blog/data-science-for-social-good>

Published: June 10, 2020 | By Martin Anderson



Though the study and interpretation of statistical data dates back at least to the eighth century, the modern meaning of 'data science' is inextricably linked with revolutionary innovations in AI software development over the past ten years. With the power of neural networks has come the responsibility of defining ethical frameworks, laws, and boundaries for how such systems will be used, and for how the data is obtained, treated and interpreted.

It's a work in progress, fueling a debate that has, predictably, been defined by the most negative implications of machine learning and big data — from deepfakes through to AI-powered military killing machines, intractable scoring algorithms in credit, health and insurance evaluations, biased or intrusive use of public data for law enforcement, and increased unemployment through AI automation, among many other social points of contention.

Consequently, data science has entered the public's consciousness in the context of a disruptive threat rather than a social and societal good.

However, municipal government, NGOs and many charitable or crowdsourced concerns are exploiting data science at least as eagerly as big business and high-level governments around the world, and for much the same reason: the digital revolution of the last twenty years has made high-volume data available to the masses, while the open source revolution in machine learning and the advent of GPU acceleration has democratized the power to exploit that data for unarguable public benefit.

Data can now be exploited for unarguable public benefit owing to the open source revolution in machine learning and the advent of GPU acceleration.

AI Projects for Disaster Response and Relief

In disaster scenarios, information can be both scarce and overwhelming. With the central communications infrastructure compromised, ad-hoc communication technologies such as local Bluetooth networks can combine with social media to form an unconventional but useful lifeline for rescue services and volunteer organizations.

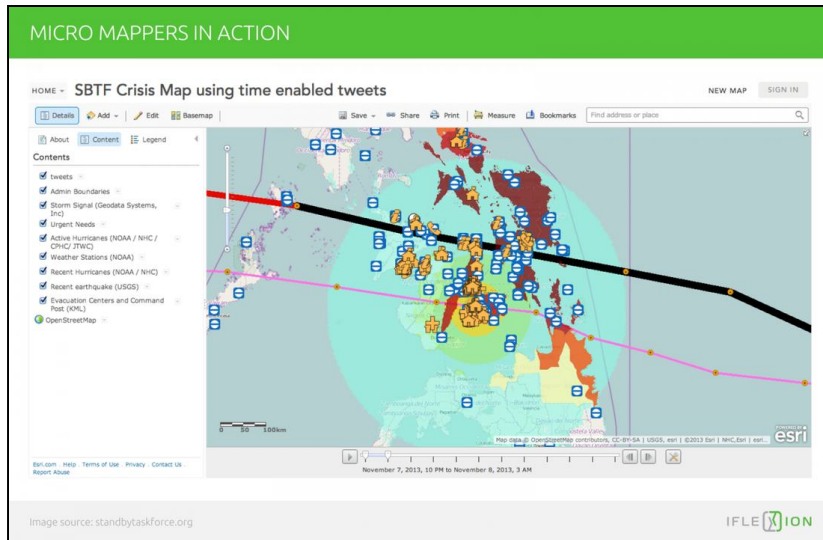
In certain cases, secondary surveillance and monitoring systems, from closed loop CCTV networks to satellite feeds, can provide additional data where normal channels have been disrupted.

The Artificial Intelligence for Disaster Response (AIDR) repository is a collaboration among data scientists and engineers across Asia and the west, with the aim of providing open-source monitoring and analysis tools to develop action maps from social network messages in the wake of a disaster.

The initiative was born in the wake of a 7.7 magnitude earthquake in Pakistan in 2017, when the United Nations asked the Qatar Computer Research Institute (QCRI) to help interpret the torrent of social media posts from the affected zone in order to help target relief efforts.

Within a few hours of the earthquake, over 35,000 geolocated tweets had been collected via QCRI's Micro Mappers NLP (Natural Language Processing) worker threads, which ran filters over the rapidly accruing dataset to identify disaster hot-spots to be passed on to the triage processes overseen by the UN and volunteer aid.

The Micro Mappers provided 14,000 tweets to the Digital Humanitarian Network's Standby Volunteer Task Force (SBTF), and within 30 hours 341 images from the tweets had been assimilated into AIDR in order to generate a human-tagged machine learning classifier algorithm capable of identifying pertinent messages with a near to 80% accuracy.



AIDR, which constitutes the interpretive framework for the data, leverages convolutional neural networks (CNNs) that are additionally capable of creating segmentation maps to visualize crisis spots, aided by UAV systems.



The system is collaborative: AIDR's tagger module initially relies on human-tagged posts as the ground truth for a classification algorithm which will then spontaneously classify tweets within its estimated catchment range of pertinence. AIDR has subsequently been used, together with diverse intelligence-gathering tools, in a number of disaster scenarios around the world.

In the interval, various other similar machine learning-driven disaster relief systems have been developed:

- The American Red Cross's Missing Maps project combines crowdsourced mapping with a Facebook machine learning layer to identify informal communities that are not registered in popular mapping systems.
- New research out of the University of Texas uses transfer learning via the VGG-16 CNN to collate and map social media posts in disaster scenarios.
- The US Department of Defense is in the third year of its XView challenge to AI engineers, with the most recent iteration concentrating on [computer vision algorithms](#) to more effectively analyze satellite and aerial image streams. The aim is to more effectively ingest and process data for post-disaster damage assessment.

Using Data Science for Medical Diagnostics

It can take the length of a career for a medical professional to gain enough experience to correctly diagnose medical conditions on the basis of data such as X-rays, CT scans, and other investigative technologies, placing an effective diagnostician among the most highly-prized of medical consultants.

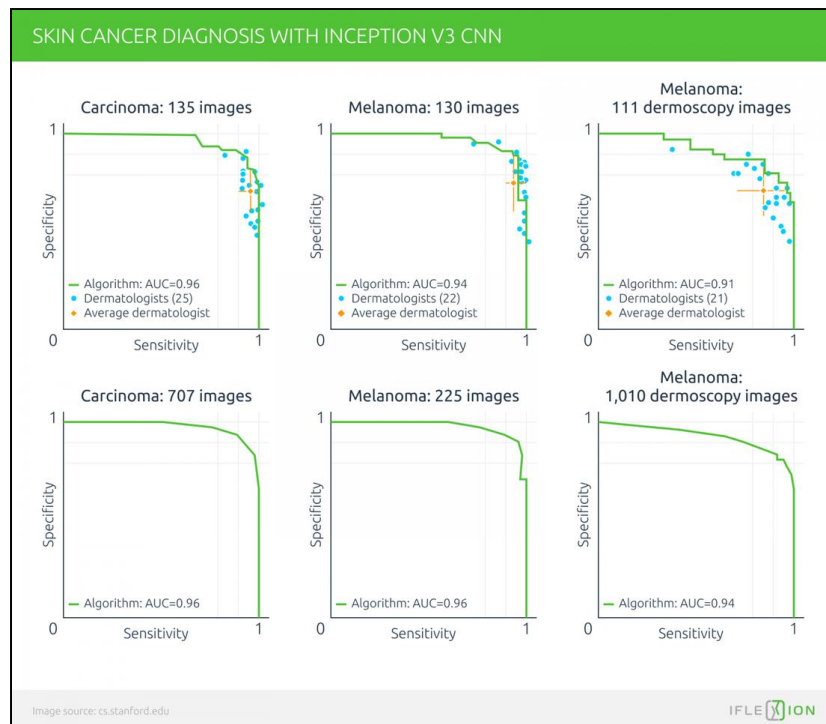
Since neural networks are almost exclusively designed to recognize and evaluate patterns in large data-sets, AI's contribution to better patient outcomes is growing every year. Combined with new streams of data from wearable devices and IoT-based streams, the potential for data science to improve and lengthen life offers unprecedented opportunities for [healthcare app developers](#) in this sector.

Open-source medical data-sets such as the Breast Cancer Wisconsin (Diagnostic) Data Set are available for medical researchers to leverage with deep learning frameworks, with [more than 40 other comparable data-sets](#) available in the US alone.

The American Association For Cancer Research also operates the cBio Cancer Genomics Portal, which provides access to data from 5,000 tumor samples across 20 cancer studies.

Beyond the USA, the International Collaboration on Cancer Reporting (ICCR) publishes a range of publicly available data-sets for evidence-based research.

In terms of [AI applications](#), a number of real-world projects are enjoying success. Research out of the Stanford Artificial Intelligence Laboratory offers a skin cancer diagnosis algorithm derived from 130,000 skin disease images processed and classified by a neural network. The AI achieved a success rate equal to human counterparts almost immediately after it began operating.



Researchers in Beijing are successfully using machine learning and big data to diagnose malignant thyroid nodules, with a 90% accuracy. Also, recent findings from the Sidney Kimmel Cancer Center offer thyroid nodule analysis through a combination of non-invasive ultrasound and machine learning provisioning from Google — potentially cutting down the need for invasive biopsies.

In this case, the AI was trained on ultrasound images from 121 patients who had received biopsies for the affected regions visualized. The machine learning algorithm was subsequently able to determine patterns and establish parameters for future diagnostic procedures.

Furthermore, major AI contributor Google AI has developed a cancer detection algorithm that's 16% more accurate than a trained pathologist in interpreting tumors from medical imaging.

However, Google notes that such algorithms, despite their efficacy, are not intended to replace human diagnosticians, since they are focused on particular pathologies, and will not pick up on any adjacent problems that a scan might reveal.

Improving Air Quality in Cities

Meteorological data presents one of the most confounding challenges for statisticians and data scientists seeking to create worthwhile prediction models. In the case of air quality, unnatural factors such as carbon monoxide, nitrous oxide, and methane gas add yet more variables to weather patterns and air flows, which are already difficult to model accurately.

Current approaches to model-building favor traditional long-term data evaluation, with little global consensus regarding investment in real-time air monitoring systems. While sensor networks are relatively affordable, machine learning equipment and expertise is usually more difficult to obtain, and connectivity for remote sensors presents an additional hurdle, as we will see.

A number of global initiatives and science competitions are leading the way in the development of potential affordable new workflows and methodologies for meteorological data science. Many are international initiatives, offering the hope for consensus regarding which machine learning and data science techniques will be most effective in tackling air pollution in real-time monitoring systems.

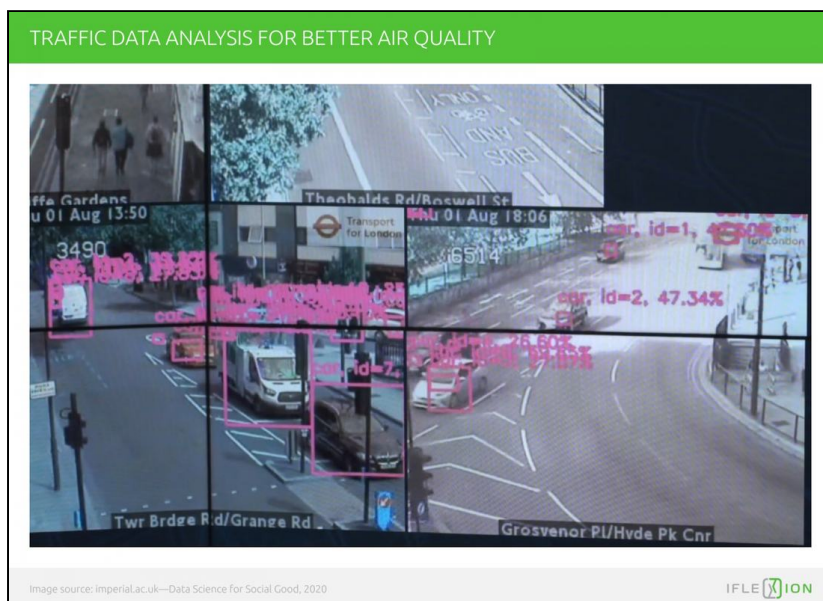
MODEL DEVELOPMENT FOR ML-BASED AIR QUALITY PREDICTION SYSTEMS							
Model/ algorithm	Artificial neural network	Radial basis function	Deep belief network	Decision tree	Improved decision tree	Support vector machine	Improved support vector machine
Big-data based	Yes	Yes	Yes	No	Yes	No	Yes
Air quality factors	Physical, Bio, Chemical	Physical Chemical	Biological Chemical	Physical, Bio, Chemical	Biological Chemical	Chemical	Chemical
Structured data-sets	Yes	Yes	Yes	No	Yes	Yes	Yes
Training data	60%	82%	76%	55%	70%	60%	78%
Real-time prediction	Yes	No	Yes	No	Yes	Yes	Yes
Simplicity	Yes	Yes	No	Yes	No	Yes	Yes
Accuracy	Err: 3.7x10 ⁻⁴	R ² =0.6020	Err: -1.6(O ₂)	70%	85%	RMSE: 40.7%	Err:0.089
Upstream flow	Yes	No	Yes	No	Yes	Yes	Yes
Downstream flow	Yes	No	Yes	Yes	No	Yes	Yes
Regression based	Yes	Yes	Yes	Yes	No	No	No
Sensors used	No	Yes	No	Yes	No	Yes	Yes
Robustness	Yes	Yes	Yes	No	No	Yes	Yes
Flexibility	Yes	No	Yes	No	Yes	No	No
Data clustering	Possible	Impossible	Possible	Possible	Impossible	Possible	Possible

Data source: researchgate.net—Air Quality Prediction: Big data and Machine Learning Approaches

IFLE 

Unsurprisingly, investment is easiest to find where meteorological issues have a direct and uncontested effect on society. Therefore, the topic of air pollution in cities is currently the 'loss leader' for the sector.

Imperial College London's Data Science for Social Good Fellowship Programme (DSSG) includes an initiative to improve air traffic data via [machine learning](#), with a view to developing new regulatory systems for improving air quality in cities.



Conventional traffic data is often low-resolution and regional, with the output data being too aggregated and infrequent to permit ad-hoc interventions or experiments with traffic regulation. So to achieve improved coverage and greater data resolution through frequency, the DSSG project uses a GPU to interpret the image streams of 600 Transport For London CCTV cameras in real time.

The project is aimed at informing better policy decisions, serving as background data for new green route developments in cities, and at providing open-source datasets for better air quality predictions for meteorological offices.

Machine learning is also powering the AirQo Ugandan Air Quality Forecast Challenge, which has spent three years building new sensor-driven monitoring services for urban areas in sub-Saharan Africa. The vast amounts of data generated require a great deal of computation, a logistical problem that has been alleviated by a 2019 grant from Google to provide a cloud-based platform for the project's data science needs.

Research out of China, one of the world's most affected areas in terms of air pollution, has proposed an air quality forecasting framework called DAQFF, comprising two deep neural networks. The architecture leverages one-dimensional convolutional neural networks (CNNs) and a bi-directional long short-term memory (LSTM) recurrent neural network (RNN).

Dozens of other research efforts around the world are making headway toward standardized approaches to air quality data analysis through AI.

Waiting for 5G

In common with many proposed civic and environmental monitoring frameworks, live air quality analysis systems require effective sensor networks in outlying locations that may be poorly served in terms of bandwidth availability. Arguably the biggest bottleneck in a machine learning workflow is not the computing power, but the lack of even a meagre bandwidth from a sensor's location.

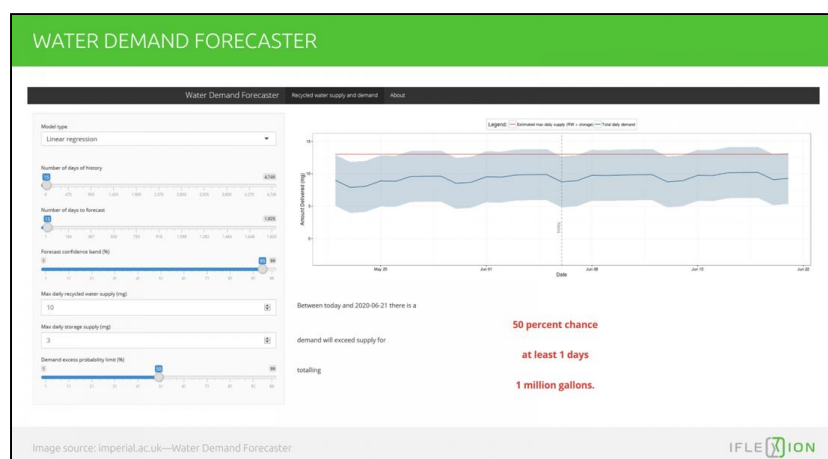
Thus many of the air quality monitoring systems envisioned are essentially awaiting the roll-out of 5G in the near future. The 5G spectrum includes 'low-band' or 'sub-6' 5G frequencies that will be suitable for lightweight data packets over considerable distances from the transmitter exchange.

The coming of 5G is about to give a boost to AI and IoT driven air quality monitoring systems.

Data Science for Social Good: Worldwide Initiatives

Reporting the full spectrum of active sectors, projects and resources in AI-driven data science is beyond the scope of this article, but initiatives around the world are lively and varied:

- Medic Mobile is a US-based organization working with Datakind.org to develop machine learning algorithms capable of anticipating the medical needs of outlying communities.
- The Urban Sanitation Challenge is using Time Series Forecasting to analyze waste patterns in under-served regions of Kenya, and eventually create more extended networks of health and sanitation provisioning.
- The volunteer-driven DataCorps project in California is using data science in its Water Demand Forecaster (see image below), so that drought-stricken regions in the state can benefit from the same kind of analytical prediction system for water as has long been implemented for electricity consumption. The aim is to avoid the necessity of importing expensive potable water to the afflicted communities.



- A UK-based data-sharing and machine analytics project around homelessness has created prediction systems for a homeless shelter, in order to better understand how statistics can indicate future demand, and what the key indicators might be for a client returning to the hostel in the future. Volunteers have also used open-source visualization program Gephi to map patterns in hostel residents' use of the Citizen's Advice Bureau (see image below).



- A project in Chicago worked with one of the city's leading providers of mental health outreach services, creating an analytical dashboard powered by data from several siloed providers. As well as incorporating the outreach company's internal client database, the data hub was

able to connect to sources from the Cook County Jail and the Illinois Department of Healthcare and Family Services. The resulting dashboard offers new and valuable insights into trends in resource allocation and client behavior.

- In Australia, an AI platform is analyzing satellite image feeds to identify areas of deforestation, with the aid of crowd-sourced volunteer efforts. The satellite images are classified by the volunteers before being passed to a neural network that evaluates and weights the ecological significance of the data, ultimately deploying new algorithms capable of providing accurate and actionable information regarding deforestation.

Conclusion

Negative news about [data science in politics](#) risks to entirely frame the debate, as well as to define the regulatory landscape and poor public profile likely to ensue for the sector. Therefore, it's important to consider that nearly all potentially harmful new technologies have transpired to have beneficial uses, and that the good works of AI-driven data science are in evidence and growing all over the world.