DRIVENDATA

IMPACT REPORT

10 years of data and AI solutions
for problems that matter

Letter from the founders

When we started DrivenData a decade ago, “data science for social good” was an idea, but not yet a movement. Data science was booming in the private sector, but in nonprofits, NGOs, and public institutions, its potential remained untapped.

Working out of the Harvard Innovation Lab, we set out to close that gap. In addition to offering our own data science consulting services, we developed a crowdsourcing platform that connects social sector organizations with top data science talent. Through our challenges, teams compete for prizes by developing the best machine learning models to solve social-impact problems. Among hundreds of models submitted per competition, our benchmarking process identifies proven solutions that organizations can further develop.

From the beginning, and across all the work we do, our mission has been about more than technical excellence. We help organizations operate at the intersection of what’s possible with technology and what’s worth doing from a societal and ethical perspective. Over our first decade, we explored and adopted new technologies to design practical, ethical, and innovative data solutions. In doing so, we’ve helped prove that data science and AI can be used not just to optimize profits, but also to improve lives and contribute to a growing global movement dedicated to using data for good.

Along the way, we’ve had the privilege of:

Partnering with **128 organizations**, from the World Bank and the Gates Foundation to Candid, Microsoft, IDEO.org, and NASA.

Designing **80+ data science competitions**, tapping a worldwide network of problem solvers, and awarding **$4.3 million** in prizes.

Building a thriving competition platform and community of **125,000+ impact-minded developers** eager to contribute their skills to social good.

Delivering **97 consulting projects** where we’ve built everything from custom datasets and predictive models to full-scale AI applications that advance education, health, conservation, human rights, and more.

Pioneering the use of AI and ML on **video, audio, imagery, text, and tabular data**, widening the possibilities for meaningful analysis and innovation.

Creating **open resources for the field,** including standardized templates and an ethics checklist for data scientists.

Ten years later, the world looks very different. Technology has evolved at a breathtaking pace. New tools are transforming what's possible, but they've also raised new questions about privacy, equity, and trust.

For many social sector organizations, the challenge isn't whether to use these technologies; it's how to use them responsibly, effectively, and sustainably. The risk isn't a lack of innovation, but uncertainty about the effectiveness of implementation. In response, we continue to help organizations realize the benefits of data science and AI by providing services that extend from R&D to full implementation support.

This report shares the story of our first decade, organized around three themes that capture how we think about lasting impact: **Projects, Portfolio,** and **People**. In using data science for social impact, the most powerful work starts with problem-solving (Projects). By mining ideas from our suite of projects (Portfolio), we transfer knowledge from one domain to solve related complex social impact challenges. Over time, the accumulated and shared knowledge increases practitioner expertise and organizations' capacity (People). When all three come together, real change happens.

As we look back on the first ten years of DrivenData, we're filled with gratitude for our partners, our community, and the shared belief that data and AI can make the world better. We hope the reflections and examples in this report spark ideas, raise new questions, and inspire you to join us in shaping what comes next.

With appreciation,

Greg Lipstein, Peter Bull, and Isaac Slavitt
Co-founders of DrivenData

DrivenData by the numbers

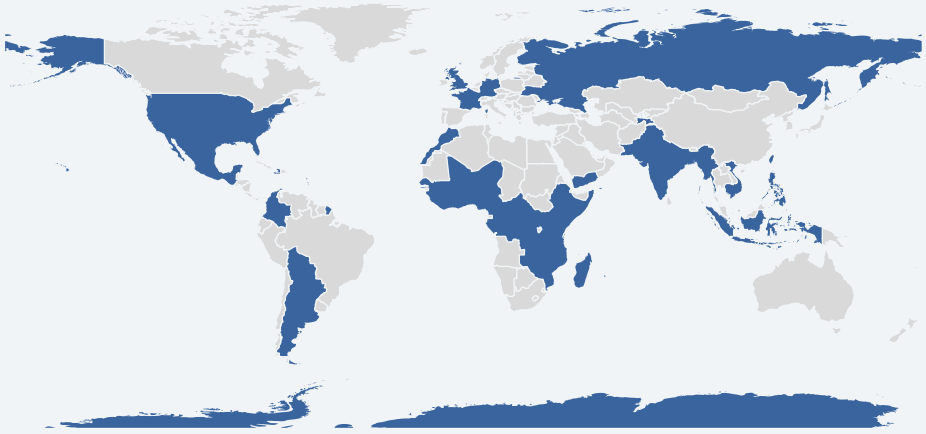
DrivenData runs data science competitions and works directly with mission-driven organizations to tackle real-world challenges in areas like health, education, conservation, disaster response, and more. We help social impact organizations harness their data to work smarter and offer more impactful services using data science, machine learning, and AI.

+Mars

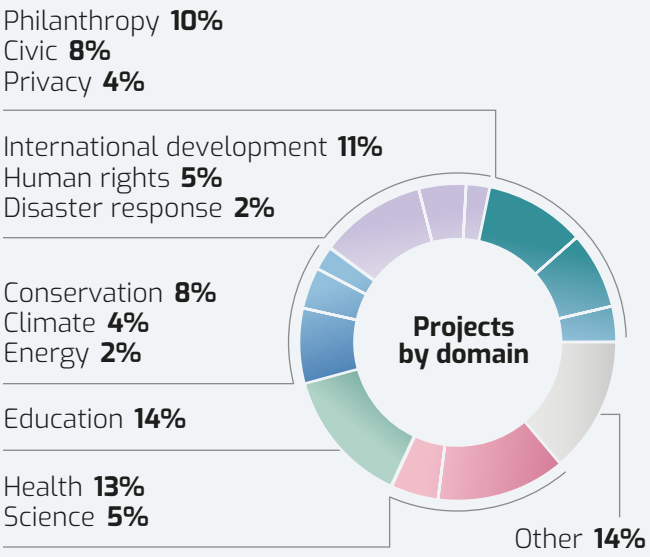
166 projects

128 partners

57 countries

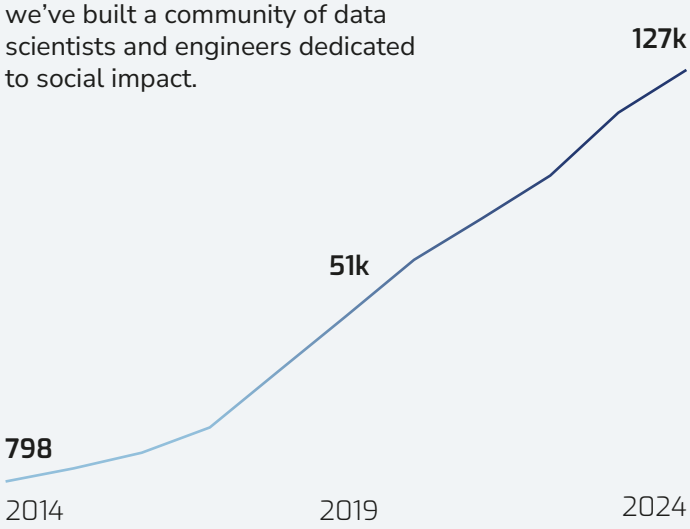


A breadth of social impact domains



125k+ DrivenData community members

Through our competition platform, we've built a community of data scientists and engineers dedicated to social impact.



10k+ GitHub stars

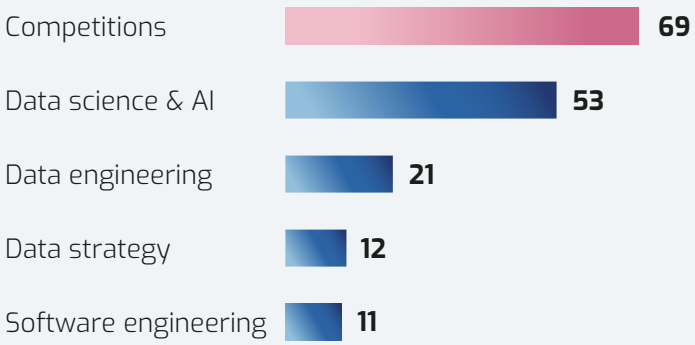
We maintain a number of popular open source projects for the data science, machine learning, and software engineering communities.

Cookiecutter Data Science

://cloudpathlib

deon

97 consulting projects, 69 competition projects



All numbers in the report reflect work between 2014 and 2024.
Some competition projects included multiple competitions.

Projects

01

Projects

Understanding project impact

The primary avenue for impact in our work is at the project level. This is where we apply data science to specific problems to effect change.

The starting point for any project, and what ultimately defines its impact, is establishing the purpose and scope. With a decade of practical experience on more than 150 successful projects, and a staff with deep technical expertise, we've gained enormous insight into what works and what is worth doing.

"DrivenData has some of the smartest, most caring, and humble folks I've had the privilege of working with. If you want to deploy cutting-edge AI for social impact, this is the team to work with."

— Andrew Means, Senior Director at Salesforce.org

WHAT OUR PARTNERS SAY

Together with funders and partner organizations, we assess each project through the lens of feasibility and value. Our feasibility assessments cover technology, methods, and cost. Our value assessments cover the utility to the organization, end users, and the wider community of social impact practitioners. Our aim is to shape and deliver work that is ethical, cost-effective, pragmatic, innovative, and sustainable.

We also benefit from a cross-domain perspective that lets us apply learnings and methods from an individual project to similar technical challenges in seemingly unrelated contexts. While the subject matter of our work may vary widely, all of our projects are deeply rooted in the use of data science and machine learning to help social sector organizations do their work more effectively.

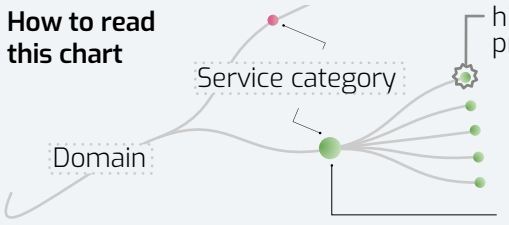
Transformative projects from our first decade

The following projects demonstrate ways that data science and artificial intelligence (AI) have been applied in the social impact space to change the world for the better. These projects showcase the range of pressing societal issues that data science can help address and reflect the diversity of methods, data, tools, and solutions we use in our work. Taken together, they not only tell a story of past successes but also point to the depth and breadth of positive impact that is possible in the years ahead.

A closer look at our projects

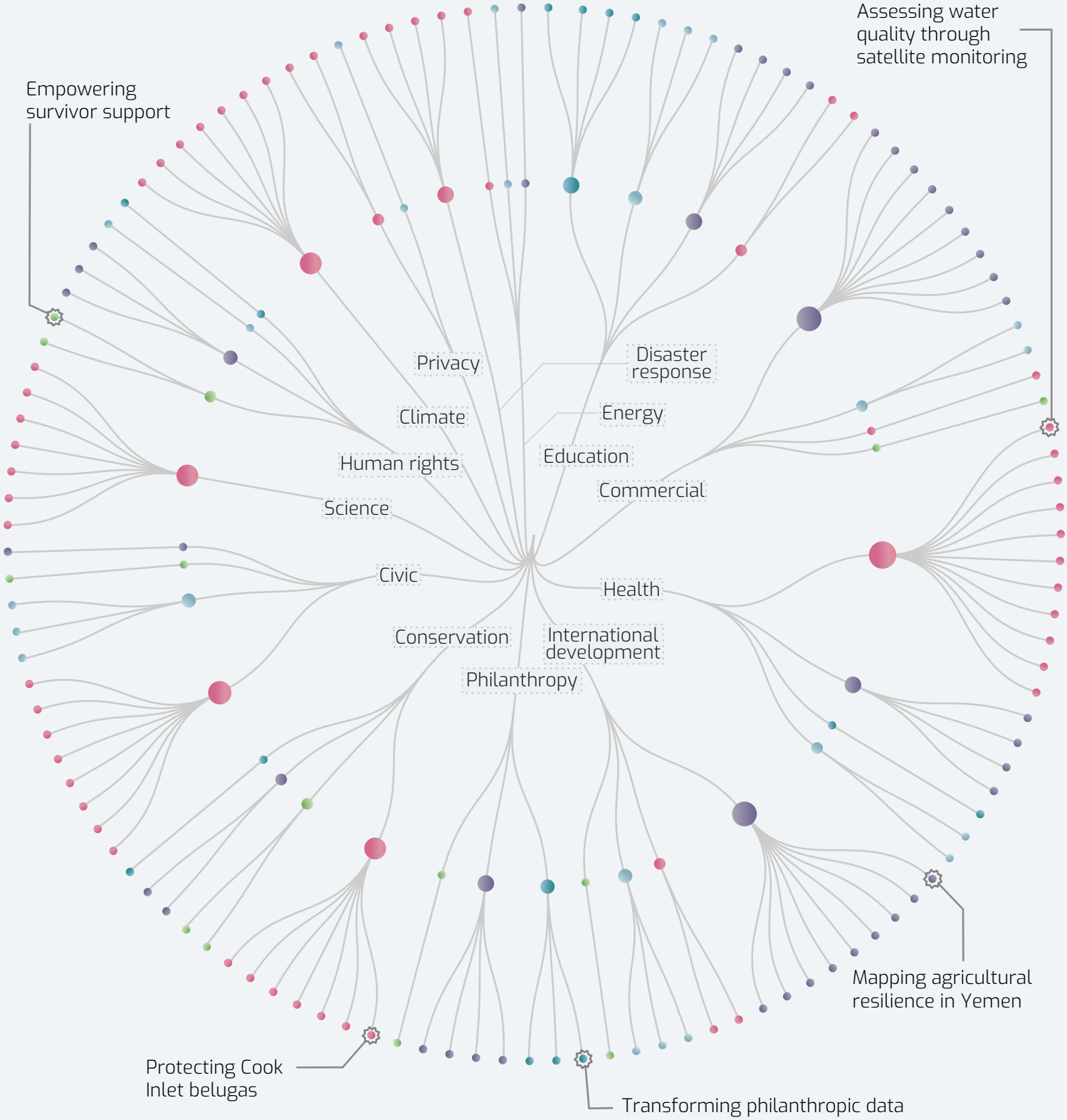
A comprehensive project list broken by domain and service category

How to read this chart



Service category

- Data engineering
- Data science & AI
- Software engineering
- Data strategy
- Competition



Protecting Cook Inlet belugas

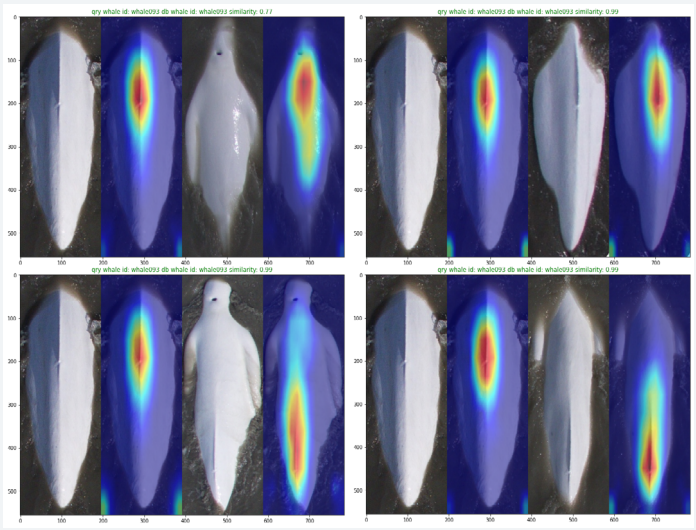
CONSERVATION COMPETITION

Partners: Bureau of Ocean Energy Management, National Oceanic and Atmospheric Administration, Wild Me, and the NASA Tournament Lab

Beluga whales are extremely sociable mammals that live in seasonally ice-covered Arctic and Sub-Arctic waters. Years of hunting and vessel traffic have dramatically reduced the Cook Inlet beluga whale population to the point of extinction, with just over 300 estimated to be alive in 2022. NOAA Fisheries monitors Cook Inlet belugas through annual aerial photo surveys that use uncrewed systems to photograph whales at river mouths. Historically, researchers manually identified individual belugas based on their distinct color, scars/marks, and dorsal ridge features, a very time-consuming task.

We helped BOEM, NOAA, and our other partners improve photo monitoring processes by designing and hosting the Where's Whale-do? challenge. Over 400 solvers developed over 1,000 automated solutions for matching photos of individual whales against an image database, a task called re-identification. To ensure the winning models would work well for NOAA, models were tested for generalizability under different conditions, such as querying across years or querying top-view images against side-view images. To ensure end users could examine and understand model behavior, finalists competed for bonus prizes that rewarded explainable methods. Explainability reports showed the dorsal ridge and surrounding areas, along with scars and other marks, were the most important features used by winning models. The winning reports and models, as well as the competition dataset, are published in open-access repositories.

Aerial photograph of endangered Cook Inlet beluga whales. DrivenData competition solvers developed models to re-identify individual whales, enabling non-invasive monitoring. Image sources: NOAA Fisheries



Explainability heatmaps show distinctive dorsal ridge features (highlighted in yellow/green) used for re-identification. Image sources: Competition winner Raphael Kiminya

The models and explainability techniques identified in "Where's Whale-do" are now actively used for non-invasive wildlife monitoring, and not just for Cook Inlet belugas. After the competition ended, our partners at Wild Me confirmed that the winning models outperformed the then-state-of-the-art models in this space. They developed the MiewID algorithm based on the winning models' approaches and found

that it worked well not only with belugas but also with other whale species, dolphins, and even lions and leopards. The new algorithm is now integrated into Wild Me's software platform, which NOAA Fisheries biologists use. The competition's winning explainability technique has also been integrated into the platform to power a match visualization feature.

This work leveraged our community and crowdsourcing approaches to tackle a difficult technical conservation problem. With new methods and models for re-identification, we enabled our partners to monitor and protect critical wildlife populations without invasive physical tagging.

Empowering survivor support

HUMAN RIGHTS SOFTWARE ENGINEERING

Partner: EverFree

Human trafficking and exploitation constitute a severe humanitarian crisis. As of 2021, an estimated 50 million people worldwide were living in modern slavery, forced to work or marry against their will.

EverFree provides comprehensive care for survivors of exploitation, and unites partners in measuring impact, improving care, and stopping exploitation. Their "lifemap" assessment methodology provides survivors a research-backed, participant-centered way to determine their strengths and vulnerabilities around core dimensions of well-being and freedom. Working with survivors, case workers use the assessment to identify priorities and develop individualized care plans.

We partnered with EverFree to design and build a secure, data-driven, production web application called Freedom Lifemap. The application's user-friendly interface allows survivors to complete the assessment, engage with visual summaries of their responses as they set priorities, celebrate achievements, and track their progress over time. In addition, data from the platform enhances the ability of anti-trafficking communities and organizations to identify effective interventions, allocate resources, and drive policy change.

In 2024, the new Freedom Lifemap platform was deployed to 20 partner organizations worldwide, which conducted over 2,000 assessments with over 1,700 participants.

Freedom Lifemap is an example of the power of thoughtfully designed data applications to drive social good. Based on user-centered design and best practices in data and software engineering, Freedom Lifemap enables survivors to receive the individualized support they need to thrive.

"DrivenData did an outstanding job building our Freedom LifeMap platform, providing us with innovative solutions to visualize complex data and track survivor pathways."

— Katie Rootlieb, COO of EverFree

WHAT OUR PARTNERS SAY

Mapping agricultural resilience in Yemen

INTERNATIONAL DEVELOPMENT | DATA SCIENCE & AI

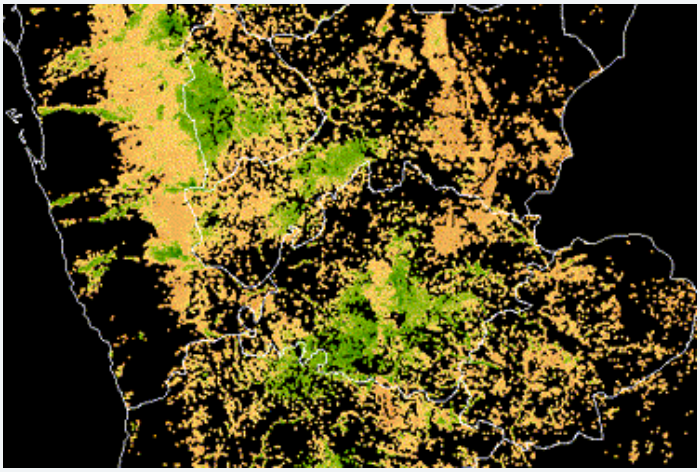
Partners: World Bank

After years of conflict, Yemen faces critical challenges in food security and agricultural development. The World Bank sought a comprehensive measurement of agricultural crop patterns and climate risks across the country to inform central statistics and development programs, but traditional ground-level data collection was impractical and potentially dangerous.

We helped the World Bank estimate crop health and its correlates using satellite imagery. We identified relevant datasets and worked with experts to remotely label over 210 square kilometers of high-resolution satellite images of Yemen. We then used the data and labels to train models that produced a comprehensive 10-meter-resolution map covering all of Yemen, processing millions of satellite image tiles to generate predictions of crop types and health over a six-year period. Significant effort went into scaling inference to handle the cumulative total of 27 billion predictions.

Our work revealed a significant expansion in cropland starting in 2019, which correlated strongly with changes in local rainfall patterns. The analysis also identified increasing climate vulnerabilities, particularly due to greater seasonal rainfall variability. Our models and the data we sourced continue to serve as decision-making resources at the World Bank, as insights from our analysis have been used to shape strategy and target interventions.

This work exemplifies how data science unlocks the power of large-scale, passively collected data. Using satellite imagery and machine learning, we were able to provide comprehensive views of conditions on the ground and changes over time, enabling more informed decision-making.



A map of NDVI (Normalized Difference Vegetation Index), a satellite-derived index that measures vegetation greenness, over cropland in Yemen. DrivenData's analysis supported World Bank food security programs in the conflict-affected region. Image source: DrivenData.

Transforming philanthropic data

PHILANTHROPY | DATA ENGINEERING

Partners: Candid

Transparency and accountability are foundational to effective philanthropy. However, access to reliable information about non-profit organizations and their funding relationships remains a challenge.

Candid maintains a leading database of millions of non-profits, built from the reams of public records that it collects annually. These records document the grants and donations flowing between organizations. Matching records of these flows between existing or new organizations is a critical and difficult part of the process. Public records inevitably contain variations in names, addresses, and other details, which can lead to duplication, mis-matching, and missing connections. Achieving highly accurate

matched records typically requires labor-intensive, human-in-the-loop review.

Over several years of phased work, DrivenData helped Candid develop tools and models to improve matching, identify duplicate records, and focus manual reviewers' time on the trickiest data anomalies. We also implemented tools that improve model visibility and data management workflows.

"DrivenData helped Candid significantly improve our entity matching algorithm, which has had a direct impact on the quality of data we provide to the social sector. Beyond the technical deliverables, they worked alongside our team to build lasting capacity in data science and AI, not just deliver a solution."

— Shane Ward, VP of Data at Candid

WHAT OUR PARTNERS SAY

The result is an improved data processing system and a more sustainable database that helps funders and organizations spend and raise money in more efficient, accountable ways, so each dollar improves the world.

Our work with Candid exemplifies our customer-centric approach to data engineering and applied machine learning, a necessity for projects in the social impact movement. Through this work, we met the challenge of both improving the data and creating an accountable, transparent data-handling process. The system was custom-tailored to Candid's specifications and made operational on their internal infrastructure.

Assessing water quality through satellite monitoring

HEALTH COMPETITION DATA SCIENCE & AI

Partners: NASA, NOAA, EPA, USGS, DOD's Defense Innovation Unit, Berkley AI Research, Microsoft AI for Earth, and the NASA Tournament Lab

Inland water bodies, such as lakes and reservoirs, face threats from harmful algal blooms (HABs), which are often caused by cyanobacteria. Cyanobacteria blooms produce toxins that are harmful to humans, pets, and aquatic life. To keep the public safe and healthy, water quality managers collect manual samples to measure cyanobacteria levels and then issue drinking water advisories or close recreational areas when needed.

While field samples are accurate, manual sampling is time- and labor-intensive. Field teams must travel to the water bodies, collect samples at designated points, and then send them to labs for toxin analysis. The geographic coverage and measurement frequency needed for comprehensive, timely action are simply more than is possible to do manually.

To meet the challenge of bringing satellite monitoring to inland waterways, DrivenData organized the Tick Tick Bloom machine learning competition, in which participants developed novel approaches to detect cyanobacteria from Sentinel-2 satellite imagery. Participants trained models on a newly aggregated dataset of thousands of manually collected in situ cyanobacteria measurements, sourced from 14 data providers across the U.S. The winning models and novel dataset from the competition are published in open-access repositories.

Building on the winning solutions, DrivenData developed an open source package called CyFi, short for Cyanobacteria Finder. CyFi makes it easy to generate cyanobacteria estimates for a given date and location by running the trained machine learning model under the hood, enabling satellite-based detection of HAB outbreaks in lakes, reservoirs, and rivers. In a benchmark comparison, we found that CyFi performs at least as well as other satellite-based tools for HAB identification. Because of its high resolution, CyFi can detect HABs in much smaller bodies of water than other systems, providing coverage for 10 times as many lakes across the U.S.

To make CyFi as useful as possible, we conducted user interviews with water quality managers, used human-centered design methods to gain insight into current workflows, and gathered their input on how CyFi could support monitoring and decision-making.

With CyFi in the workflow, water quality managers can better allocate resources for in situ sampling, make more informed decisions about when to issue public health warnings, keep humans safe, and ensure the health of aquatic life that rely on small inland water bodies.

This project exemplifies the full lifecycle from dataset development and research to modeling and benchmarking to application development. Advancing the science and proving the approach were the first critical steps, and the competition approach enabled DrivenData to rapidly iterate on potential algorithmic solutions through crowdsourcing. Building a useful and usable tool was the next step to creating a practical impact. DrivenData remains engaged in promoting and supporting CyFi and its use with the community of water authorities and regulators.

A cyanobacteria bloom in Lake Erie as captured by satellite imagery. Through a competition and follow-on work, DrivenData developed an approach to estimate harmful algae cell counts from Sentinel-2 imagery and made the trained machine learning model available in an open source package called CyFi. Image source: NASA Earth Observatory

Portfolio

02

Portfolio

With a decade of experience and more than 160 completed projects, DrivenData operates with a portfolio perspective that allows us to see and make connections between projects in different domains, addressing distinct challenges. The benefits of mining our rich portfolio of diverse projects include:

- **Cross-cutting learning**: Working across many domains enables us to effectively identify where innovation in one area might add value in another. Having applied approaches in one context on one type of data can make it easier and cheaper to bring them to bear on another, and we get to build on the experience and lessons learned along the way.

- **Playing the long game**: Impact through innovation often takes time to materialize and depends on external factors beyond our control. Sometimes, the right data is not available. Other times, technology is not ready. Efforts in one project may highlight these deficiencies, which are then addressed in subsequent projects. Approaching our projects as steps in an ongoing progression of innovation and improvement allows us to play the long game for impact.

- **Staying sustainable**: The diversification we get from the portfolio approach offsets risks and enables us to pursue more innovative work. Maintaining a balanced portfolio enables us to sustainably serve the social sector with competitive pricing and cutting-edge expertise.

In this chapter, we share examples of how our portfolio approach compounds impact across our own work and that of others.

> "From the beginning at DrivenData, we knew what we were ultimately trying to maximize was the combined impact of our entire portfolio of projects."
>
> — Isaac Slavitt, Co-founder

OUR VISION

Compounding learning

Our portfolio perspective inspires us to borrow methods and insights and adapt them to other domains of problem-solving. The technical skills and knowledge that we've developed in one area can be applied effectively to other domains and contexts to solve very different problems.

For example, seemingly unrelated problems like supporting early detection of pests that are threatening crops in Africa and extracting skills from resumes to support job matching are both rooted in a natural language processing technique called named entity recognition (NER).

We've employed computer vision models across many subject areas, including mapping kelp forests for conservation efforts and detecting lesions in cervical biopsies for earlier, more accurate cancer diagnoses. And we've seen decision-tree models top the leaderboards for a range of satellite imagery-based tasks, including estimating the amount of water in snowpack, toxic algae cell counts, and the amount of particulate matter (PM2.5) in the air.

We also seek opportunities to build on the expertise and context we've already developed and to make a deeper impact in a given domain. For example, over the past 6 years, we've worked extensively with camera-trap data to support wildlife conservation efforts. We've run machine learning competitions focused on species classification and depth estimation to develop effective methodologies. By transforming research code into robust, customizable pipelines, we developed Zamba, an open-source Python package that leverages machine learning to streamline camera-trap data analysis. Zamba powers Zamba Cloud, a no-code web application that makes advanced custom model training capabilities accessible to non-programmers. By enabling conservationists to efficiently analyze large datasets and train models tailored to their specific ecological contexts, our work advances wildlife monitoring, research, and evidence-based conservation decision-making.

Enabling downstream work

One of the most valuable things for data science work is good data. This is both an input to our work and, at times, an output from it. Open, machine-learning-ready data for social impact applications supports research and development in areas where it is needed most.

Powering cutting-edge insights into human conversation

To advance the science of coaching, BetterUp Labs set out to explore the deep dynamics of human conversation. This required designing and executing the collection and analysis of large-scale multimodal (video, audio, text) data from the ground up. We engineered a robust system to handle the full lifecycle of newly collected multimodal conversational data. The pipeline supported raw data ingestion, validation, cross-modal synchronization, and complex feature extraction.

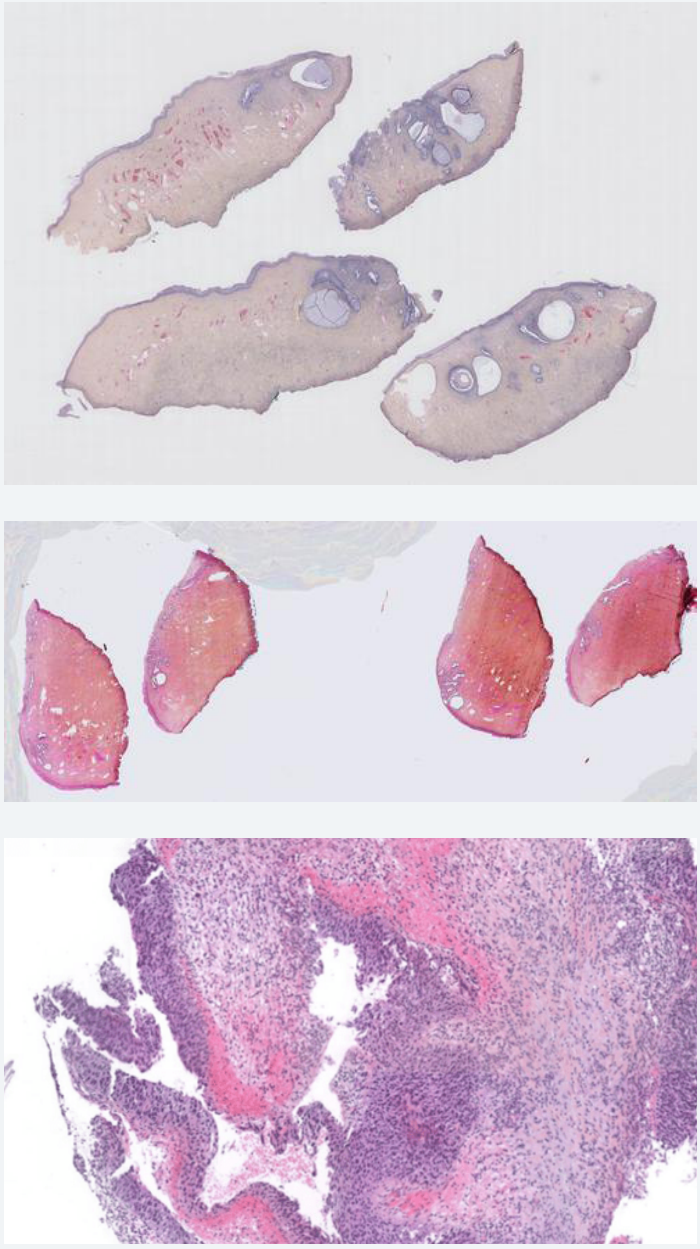
A screenshot from Zamba Cloud, showing the species prediction and bounding box detection for an image. This application, developed by DrivenData, enables conservationists to train and run machine learning models used to process camera trap data. Image source: DrivenData

The result was the CANDOR Corpus, a novel, large-scale research asset containing a 1,000+ hour dataset of real American conversations from 2020, enriched with hundreds of behavioral measures. This dataset was published in Science Advances and continues to drive ongoing research into the science of conversation and behavior.

Advancing computational pathology

Machine learning models have the potential to help pathologists work more efficiently and accurately, and focus on cases that truly benefit from manual review and expertise. Through projects with the French Society of Pathology and Health Data Hub, we've demonstrated this potential in practice. The TissueNet challenge showed that machine learning models could detect lesions in cervical biopsies with near-clinical accuracy, and the VisioMel challenge produced models for melanoma relapse prediction that significantly outperformed those based on tabular clinical features.

The datasets for these challenges were time- and labor-intensive to create, as annotating whole slide images requires specialized training and careful examination of microscopic tissue. These competitions created a lasting impact beyond their initial scope, as both the TissueNet and VisioMel datasets have been published for open sharing and continue serving as research resources for computational pathology.



Example portions of cervical tissue from microscopic slides in the TissueNet dataset. After the TissueNet competition concluded, the hand-labeled dataset was made available to support computational pathology research. Image source: DrivenData

Building momentum

Our portfolio lens pushes us to apply a range of techniques within a domain and develop deep expertise applying techniques across domains. But it also pushes us to take on risky work and explore what is possible through new pairings of technique and domain. Projects that explore and demonstrate what can be done do not necessarily show immediate benefit, but may flourish over time and with contributions by others. Engagement in cutting-edge research can create a virtuous cycle of technological advancements that stimulate additional investment, experimentation, and innovation.

Electronic monitoring for fisheries

In the N+1 Fish, N+2 Fish competition, we sought proof-of-concept applications to automate the monitoring practices that enable sustainable fishing, including detection, classification, counting, and measurement of fish. The competition results demonstrated impressive accuracy and revealed untapped potential in these tools, inspiring additional investment. Our partners further developed this work and created openEM, an open-source package that enables fisheries to affordably verify compliance with sustainability standards.

Supporting Mars planetary science

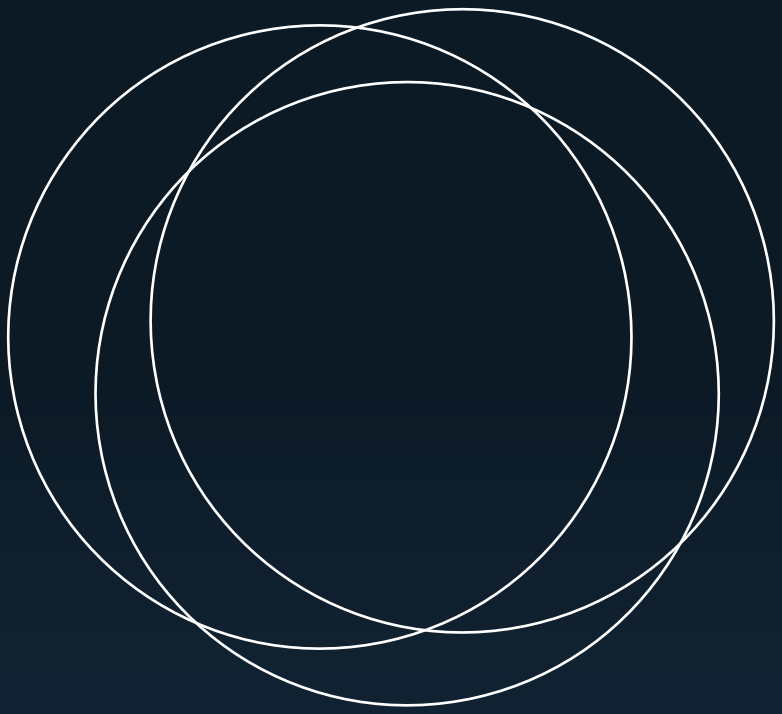
We ran two research competitions focused on building models to automatically analyze mass spectrometry data for Mars exploration. The first challenge used data collected by evolved gas analysis, and the second used gas chromatography data from soil and rock samples. These two challenges were proof-of-concept projects to assess the feasibility of combining data collected from different instruments in a single machine learning application. The results and learnings from the challenges have been published in a scientific journal and support the longer-term goal of deploying methods to autonomously guide space operations.

"The results of our challenge could not only support NASA scientists to more quickly analyze data, but also provide a proof-of-concept for the application of machine learning techniques on complex datasets for future space missions."

— Victoria Da Poian, NASA Goddard Space Flight Center Data Scientist & Engineer

WHAT OUR PARTNERS SAY

People



03

People

Building human, organizational, and social capital

Many of the most profound and durable advancements in data science for social good emerge from the vibrant ecosystem of practitioners, mission-driven organizations, funders, and the ever-expanding pool of shared knowledge and open source tools. What we are developing together through our work is human, organizational, and social capital.

At DrivenData, we've dedicated resources and time to building such capital in order to give back to the community and strengthen the movement of data science for social good. Through our work, we provide professional development opportunities for data scientists and AI/ML practitioners, both internal and external to our organization; we strengthen funding institutions and empower social impact organization teams; and we expand horizons by providing tools, knowledge, guidance, and resources to better the practice of data science for social good.

"As we look toward the future, we believe that investment in the social impact 'commons' of shared knowledge, resources, and tools, combined with the vibrant community of practitioners, is one of the most powerful engines for progress in our field."

— Greg Lipstein, Co-founder

OUR VISION

Supporting practitioners

At the heart of DrivenData's long-term impact is nurturing a vibrant community of data scientists, AI/ML modelers, and developers from around the world. These practitioners are the engine that continuously transforms technical expertise into social change.

One of the main ways DrivenData supports the practitioner community is through our competition platform. Seen through the lens of human capital, the competition process creates a space where data scientists can develop and demonstrate their skills, form teams and collaborate with one another, compete and learn from others, engage with new datasets, and broaden their understanding of critical problems. Over the past decade, the competition process hosted by DrivenData alone has grown into a powerful force for innovation: we've logged almost 300,000 submissions and awarded more than $4 million in prizes across a diverse range of challenges.

80k
Total competition participants

297k
Total submissions

$4.3M
Total prize money awarded

The process is designed to maximize both learning and impact. Competitions typically run for 2-3 months, giving participants time to deeply engage with each problem. As part of competitions, we provide careful framing of real-world applications, ML-ready datasets, user-friendly data documentation, and modeling tutorials and benchmarks to get folks started. Many participants engage with competitions to experiment with new data types or methodologies. Participants learn from one another by forming teams and sharing code or resources on the user forum or community code board. Over the past decade, more than 5,000 teams have been formed on our competition platform, connecting data scientists through a shared interest in social good.

Perhaps most importantly, we've seen how this community model builds lasting capacity in data science for social good. Data scientists in our solver community often showcase work from a data science competition on their resumes or professional profiles, leveraging their experience to advance their careers. In one very direct example of competition work advancing careers, Aivin Solatorio's prize-winning solution in our 2018 Pover-T Tests Competition impressed our World Bank partners so much that they hired him as a consultant and eventually as a full-time employee, where he has continued to contribute to best-practice data science and the application of machine learning to poverty and development data.

Building organizational capacity

Many social sector organizations know they could benefit from advanced technology, but struggle to see concrete paths forward. The rapid pace of AI/ML technology change and development exacerbates the problem.

Social impact organizations tend to fall into two groups. Many organizations lack the internal expertise to discriminate among available options and the confidence to commit to solutions that best balance reliability, impact, and cost-effectiveness.

Other organizations face a different type of constraint. They have data engineering teams that manage internal data systems, models, and research projects, but they simply do not have time for the additional research and experimentation necessary to meet all the organization's needs.

At DrivenData, we've learned that one of our most significant contributions is helping social impact organizations envision what's possible with data science and AI, and to conduct rapid prototyping and proof-of-concept work to advance the vision. Our work addresses both why and how organizations engage with data science.

We help organizations (with and without internal data science and AI/ML capacity) make confident investments by demonstrating the potential of machine learning models and instilling best practices for managing data workflows that support them. We help organizations build or enhance

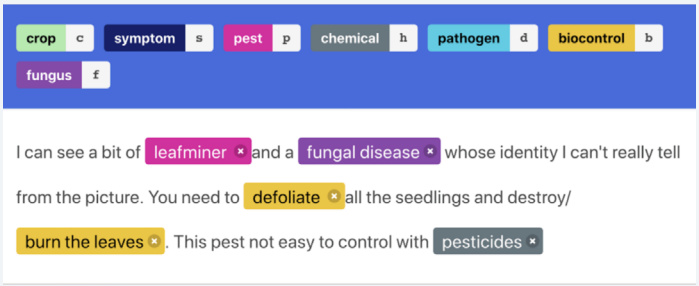
internal capacity through experimentation, rapid prototyping, documentation, training, collaboration, coaching, hiring, and the development of data and AI/ML strategies.

Empowering agricultural support with entity recognition

We worked with CABI's Plantwise program, which aims to help farmers lose less of what they grow to plant health problems by providing timely, appropriate, and actionable advice through a network of plant doctor clinics.

The project focused on automating the recognition of agricultural entities (such as crops, pests, diseases, and chemicals) in WhatsApp and Telegram messages among plant doctors, to surface emerging trends and threats. These messages contain valuable real-time information on plant health and crop issues, yet without systematic, automated analysis, it is very difficult to identify where farmers are receiving bad advice or to surface important patterns, such as emerging pests.

Integrated into our work were trainings on reproducible data science, technical guides for git and GitHub, and best practices for project management of technical work. We worked side by side with the CABI team, annotating the data, building models, and reviewing each other's code. The result of the project was not just trained entity extraction models but also an up-skilled team.



Example hand-labeled plant doctor message showing various entity types. DrivenData and CABI used natural language processing techniques to detect patterns in crop health and agricultural guidance. Image source: DrivenData

Sharing knowledge

The strength and durability of our collective social impact movement also depend on the community resources we create. By sharing knowledge, together we help each other move towards reproducible, responsible, effective, and ethical practices.

Over the past decade, we've applied lessons from our project work and invested in developing open-source tools to share our knowledge. Below are a few examples.

Cookiecutter Data Science

Cookiecutter Data Science (CCDS) is a logical, reasonably standardized, but flexible project structure for data science. With over 8,500 GitHub stars, CCDS has helped countless teams standardize their data science workflows, adopt best practices that support reproducibility, and keep data scientists organized and on track. As the original version was published over 8 years ago, we recently released an updated version that embraces changes in the landscape of data science tooling and MLOps over the past 5 years and looks to the future.

Deon

Deon provides an ethics checklist for data science projects. We created deon to help data scientists across the sector be more intentional in their choices and more aware of the ethical implications of their work. Checklists make sure big questions do not slip through the cracks, and tough conversations happen even (especially) in fast-moving environments. Deon provides concrete, actionable reminders to developers who influence how data science is done.

"DrivenData's ongoing investment in open source tools reflects our belief that good tools make good work possible. Our knowledge sharing is part of our contribution to building a better future together."

— Peter Bull, Co-founder

OUR VISION

Developer tools

DrivenData has published several open source packages that address specific developer pain points: cloudpathlib provides a consistent and easy interface in Python to access files in cloud storage like S3 and Azure; erdantic is a simple tool for drawing entity relationship diagrams that show how data model classes are connected; and nbautoexport automatically exports Jupyter notebooks to various file formats (.py, .html, and more) upon save while using Jupyter to facilitate code reviews.

Benchmarked models from DrivenData competitions

Our hosted competitions provide us with even more power to tap into a worldwide community and create shared knowledge. Part of our mission is to enable data scientists and mission-driven organizations to learn from the work done in these competitions. To this end, the code submitted by the winners is released under an open source license for others to learn from, use, and adapt.

These tools also serve as the basis for many conference talks and workshops we've given. Our capacity-building presentations extend our impact, helping practitioners leverage these tools to implement more effective solutions within their organizations.

We have oh-so-many opinions about doing data science well, and we love to share them! Some of our greatest hits are:

- Actionable insights come from actions, not from insights
- Data science is software
- Actionable ethics for data scientists
- Data science and human-centered design

We are also committed to sharing our non-technical learnings to support the broader data science for social good community. By sharing what has worked and what has not, we aim to help others in the sector work more effectively and amplify their impact.

Looking ahead

As we reflect on the past decade of innovation in data science and AI for social impact and look toward the next decade, we believe that this is a pivotal moment. The field has evolved dramatically—from early investments in dataset development and initial experiments in applying data science, to today's massive data collections and sophisticated AI solutions. So too has our capacity to use these resources to solve social problems.

The path forward demands that we tackle several pressing and fundamental challenges:

- How do practitioners keep up with rapid technological advances?
- How do organizations know when and how to invest in AI/ML for maximal impact?
- How do funders identify the most promising areas of investment and the organizations ready to lead in those areas?
- How do we transform basic research and innovation into sustained and impactful implementations?
- How do we accelerate the pace of innovation and problem-solving to keep up with emerging and ever more complex challenges?
- How do we ensure equitable dissemination and use of these powerful tools across the social sector?
- How do we mitigate potential harm and deploy AI responsibly, transparently, ethically, and reliably?

This report is in part an approach to tackling these challenging questions.

In essence, we keep moving forward project by project. We gain greater insight and capabilities from the portfolio of our collective work. We continue to advance data science for social good through shared knowledge and investments in human, organizational, and social capital.

And we don't travel solo. Every breakthrough we've witnessed has come from bringing together passionate problem-solvers with subject-matter experts from mission-driven organizations and research institutions.

We're proud to be part of an amazing, growing community dedicated to harnessing technology for social good. We are committed to expanding this community, sharing what we learn, and working together to create meaningful change. The challenges ahead are significant, but so are the opportunities to make a difference.

We look forward to working with you to shape the future.

A final word(play)

We believe data science should be rigorous *and* fun. Our punny competition titles help set that tone. Here are some we're especially proud of:

Hakuna Ma-data

Identify over 50 different types of animals in camera trap images from the Serengeti.

Where's Whale-do?

Protect endangered Cook Inlet beluga whales by identifying individuals from drone imagery.

Kelp Wanted

Help researchers estimate the extent of Giant Kelp Forests by segmenting Landsat imagery.

Tick Tick Bloom

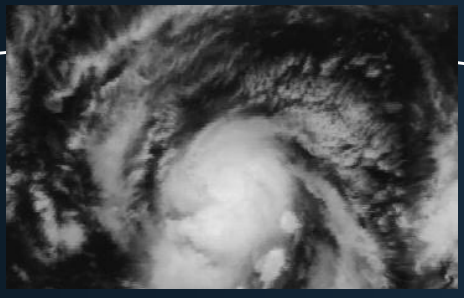
Use satellite imagery to estimate toxic algae concentrations that help protect public health.

Pri-matrix Factorization

Build machine learning models for identifying wildlife from camera trap video footage.

N+1 fish, N+2 fish

Support sustainable fisheries by extracting counts, size, and species of discarded catch from videos.



Wind-dependent Variables

Inform humanitarian response efforts by predicting wind speeds of tropical storms.

Mars Spectrometry

Analyze mass spectrometry data from rock and soil samples for Mars mission science.

DengAI

Predict the number of dengue fever cases in two locations from climate variables.

DRIVENDATA