A snapshot of my data projects

Matt Malishev
@darwinanddavis

Data are fickle beasts. They're a myriad of shapes and sizes, chaotic in their natural raw form, and often stubborn and disobedient. Despite their unfriendly nature, we need them, as they are keys to nature's mysteries. As scientists, we're very good at capturing them, cracking them open, and rummaging around to get what we need. However, merely operating is not enough. We need the right sets of tools and insight to extract pieces that tell the best story. This makes good data analysis equal parts art and science.

The art comes from finding creative ways to make complex data compelling. The science is applying technical know-how to convert raw numbers into output. My selected projects present a range of data and problem sets that blend the creative and technical elements of my data science toolbox. The tabs in the page header contain a snapshot of each project, including collaborators, the aims, outcomes, and example outputs of my data analysis, and links to downloadable files and the project's online home. As well as showcasing my data analysis in both independent and collaborative spaces, the projects highlight three crucial elements driving my interests as a data scientist.

A data science lens

First, I'm most comfortable using modelling insight and a computational lens to crack a problem. Tapping into my background as a modeller and computational scientist brings insight from diverse fields, such as math and computer science. The Meta-analysis project illustrates this well. In this 13-person project across 9 institutions, we conducted a meta-analysis on how disease burden compromises metabolic growth and reproduction of roaming animals, both wild and livestock, and its consequences on biomass recycling at a landscape scale. The project involved many hands on deck, but integral to the overall aim was including a model to bolster the larger idea with computational methods. The result was a disease transmission model using host energetics to track how infected animals can shift supply and turnover rates of ecosystem nutrients. The delight for me in this project came from being able to fill some of the technical gaps in the group, of which many members were more field and lab trained. This underscores the value of applying data scientist tools to tackle unique problems and provide necessary analytical expertise in a collaborative project.

In the Meta-analysis project, I also used my experience in data mining and wrangling to design a keyword query bot and a data scraper bot to bear some of the weight of data extraction and streamline work efficiency in the group. This is a nice example of exploiting data analytics to improve workflow and enhance output to meet project deadlines. The project page has a detailed rundown with user instructions and downloadable open source links to the bots.

Data diversity

Second, a major reason I enjoy data analytics is the diversity of data types and systems I work on. This cultivates new ways of problem solving and plays to my strength of switching between projects. Here are two project examples where I use a wide spectrum of data types and analyses.

In the Spatial project, we built an individual-based simulation model integrating weather and microclimate data with Light Detection and Ranging (LIDAR) habitat data to forecast dispersal potential of animals in space and time. The model then used GPS telemetry data of animal locations to predict species occurrence under habitat change. The project involved mining and diagnosing messy data, spatial and population modelling, integrating theory into model frameworks, and data presentation. I was able to borrow from familiar methods, such as mining geocode data, and apply novel ones, such as spatial modelling, to an ecology problem combining a wide range of technical data. The project also taught me how to overcome

the many technical obstacles of data analysis. For example, creating reproducible workflows for analysing a variety of data was the glue combining the intricate layers of the project.

In the Time-series project, we took field temperature data for human risk sites of water-borne parasites then used wavelet analysis to reveal how human disturbance transforms these areas into exposure hotspots. We meshed detailed environment, host-parasite ecology, and spatial data with unique analysis that bridged ecology and epidemiology in an applied way. The valuable lesson here was that data analysis and communication can and should reach non-specialist audiences. Turning numbers into a story and making your data sing are part of the art of good data science. The value of creative and compelling data viz makes the abstract nature of data more familiar and improves science communication. When done right, it can also soften the underlying complexity of big data and the methods by identifying where valuable insights lie.

Adapting to the right tools

Finally, collaborations need a balance among technical, communication, and bigger picture. Too much computational authority on a project can be counter-productive. Therefore, good science means adapting to the right data tools and being accountable for your expertise. A good example is the Disease project. The research problem is well understood: Improve our understanding of how host-parasite ecology translates to human health risk. Our solution was building novel simulation models using detailed field habitat data and lab data on host-parasite ecology. Our outcomes show high-risk human exposure periods occur much earlier in the transmission season than previously thought. This project helped me navigate the strengths and limitations of using computational tools in applied scenarios. Some may argue complex models beget complex results. I believe complex models give different answers to simpler ones. For example, using models to fill uncertainty and data deficiency gaps in classic lab and field heavy research can help flesh out patterns. However, identifying when their scope may be limited is also part of the process. I believe some failed management practices, such as disease prevention, bio-control, and land disturbance, stem from failing to analyse and interpret data correctly. When you realise this amounts to time, resource, and monetary costs, it stresses why data-driven decisions need the right tools and analyses.

The final Emory Coding Club project rounds out my technical expertise and collaborative work. The concept is simple: create and maintain a regular and informative teaching environment for coding and data science. The coding club is a springboard for gaining new technical know-how in research methods and analysis techniques, as well as backend programming and public liaison. This helped me adapt to group and user feedback to improve my communication and project design.

Applying a data science lens to crack problems, working with diverse data, and an ability to adapt to the right tools are three elements that stitch together my projects and drive my interest as a data scientist. My combined lead and collaborative roles among the projects illustrate my data analysis insight, academic and non-specialist communication, and analysis toolbox for adapting to a range of diverse projects.