Newsjournal of the Society for Industrial and Applied Mathematics
sinews.siam.org

siam news

Volume 52/ Issue 6
July/August 2019

Life Sciences Special Issue

Read about the application of mathematics and computational science to a wide variety of life sciences topics in this ***special issue.***

A collaboration between researchers at North Carolina A&T State University, North Carolina State University, and the Food Bank of Central & Eastern North Carolina (FBCENC) has helped optimize food distribution at the county level. Pictured from left: Steven Jiang (NC A&T State University), Lauren Davis (NC A&T State University), Charlie Hale (FBCENC), Irem Sengul Orgut (Lenovo), Reha Uzsoy (NC State University), Julie Ivy (NC State University), and Isaac Amoako Nuamah (NC A&T State University). Image credit: Robert Lasson.

In an article titled "Food Distribution: How to Give Fairly with Less Waste" on page 4, Jenny Morber explores a mathematical model that helps food banks balance equity and effectiveness under uncertain capacity.

Mathematical Modeling Gains Days for Brain Cancer Patients

By Matthew R. Francis

Glioblastoma, or glioblastoma multiforme, is a particularly aggressive and almost invariably fatal type of brain cancer. It is infamous for causing the deaths of U.S. Senators John McCain and Ted Kennedy, as well as former U.S. Vice President Joe Biden's son Beau. Though glioblastoma is the second-most common type of brain tumor—affecting roughly three out of every 100,000 people—medicine has struggled to find effective remedies; the U.S. Food and Drug Administration has approved only four drugs and one device to counter the condition in 30 years of research. The median survival rate is less than two years, and only about five percent of all patients survive five years beyond the initial diagnosis.

Given these terrible odds, medical researchers strive for anything that can extend the effectiveness of treatment. The nature of glioblastoma itself is responsible for many obstacles; brain tumors are difficult to monitor noninvasively, making it challenging for physicians to determine the adequacy of a particular course of therapy.

Kristin Rae Swanson and her colleagues at the Mayo Clinic believe that mathematical models can help improve patient outcomes. Using magnetic resonance imaging (MRI) data for calibration (see Figure 1, on page 2), they constructed the proliferation-invasion (PI) model — a simple deterministic equation to estimate how cancer cells divide and spread throughout the brain. Rather than pinpoint every cell's location, the model aims to categorize the general behavior of each patient's cancer to guide individualized treatment.

During her presentation at the American Association for the Advancement of Science 2019 Annual Meeting, which took place in Washington, D.C., earlier this year, Swanson noted that every glioblastoma patient reacts differently to the same treatment. She hopes that use of the PI model might help predict patient response to a given regimen. "The model is able to provide a sort of virtual control," Andrea Hawkins-Daarud, Swanson's collaborator at the Mayo Clinic, said. "With a virtual control, you can consider how the size of the tumor changes over time. Then you can begin thinking through a lot of different possible response metrics."

See **Brain Cancer** *on page 2*

Engineered for Function: The Power of Biologically-Constrained Neural Networks for Neurosensory Integration

By Charles B. Delahunt, Charles Fieseler, and J. Nathan Kutz

New opportunities to build bio-inspired models for neurosensory integration arise from data-driven modeling methods. The emergence of rich multimodal data recordings of neurosensory processing systems also enhances biologically-motivated models. We are entering a golden age of access to biological data and structures, made possible by a diverse set of genetically-tailored organisms and a range of recording and stimulation methods, including functional magnetic resonance imaging, electrode arrays, calcium imaging, and optogenetics. Given that many biological architectures resulted from millions of years of competitive pressure to robustly accomplish certain tasks, understanding the network architecture promises payoffs in terms of novel, valuable functionalities that one can apply to machine learning (ML) and artificial intelligence (AI) methods.

Recognizing how the form and structure of neural pathways transform input stimuli into motor-neuron-driven behavioral responses is of particular interest. Data-driven models can integrate neurosensory information into a modern learning and control theoretic framework, enhancing our comprehension of the role of network structure and function. We highlight two model organisms that exploit different network architectures for functionality: the *Manduca sexta* moth and *Caenorhabditis elegans* (roundworm). The former uses a large, randomly-connected network for processing sensory (olfactory) information and learning, while the latter functions via a small, stereotyped connectivity graph. These two alternative approaches offer insight into the range of neurosensory strategies that produce robust and stable behaviors for organisms in environments with noisy stimuli.

Neuroscience and ML techniques have existed in partnership for many decades. For instance, neural networks (NNs) were inspired by the Nobel Prize-winning work of David Hubel and Torsten Wiesel, who demonstrated that NNs in the primary visual cortex of cats are organized in hierarchical cell layers to process visual stimuli. The study of NNs is currently textbook material for both neuroscientists and the deep learning community. Anatomical studies and neuronal recordings provide increasing detail of biological structure that helps us move beyond abstract models. Biologically-constrained architectures are critical when explaining the exceptional performance of neurosensory integration with limited and noisy input stimuli, and thus have the power to further revolutionize NN design.

Manduca sexta and Fast Learning

The insect olfactory network—including the antennal lobe (AL) and mushroom bodies (MB)—has evolved for robust, rapid learning. The AL-MB's key anatomical features include competitive inhibition, random and sparse connectivity, neuromodulator stimulation, Hebbian weight updates (*fire together*, *wire together*), and large dimension shifts between layers [3, 10, 14]. Though these elements are endemic in biological NNs, researchers do not typically apply them in the context of ML [1, 13]. A model loosely based on the honeybee MB with Hebbian updates was a step toward application of biological

See **Neurosensory Integration** *on page 3*

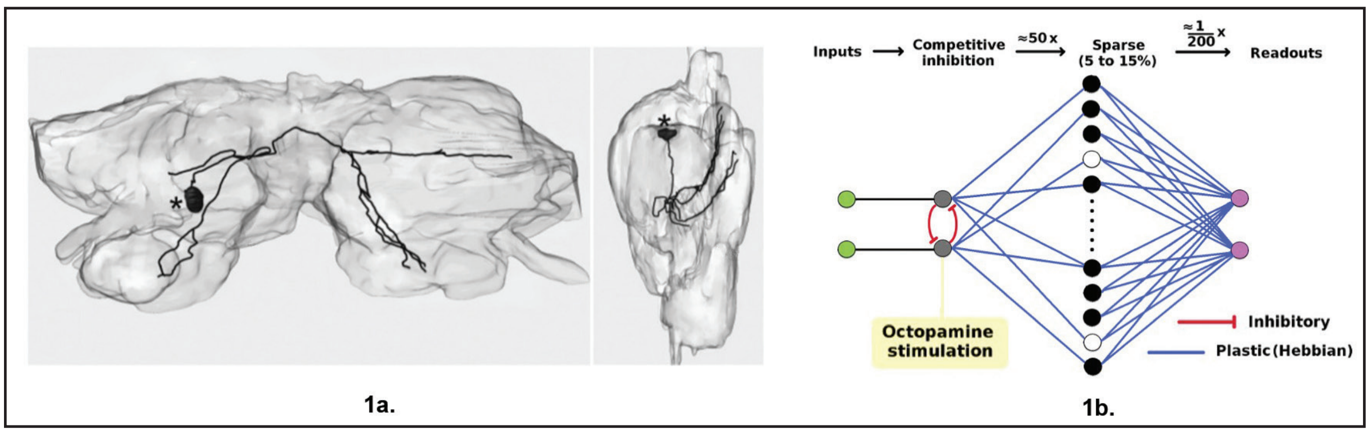

Figure 1. *Olfaction processing in the antennal lobe.* **1a.** *Three-dimensional reconstruction of a dopaminergic neuron in Manduca sexta.* **1b.** *A neural network (NN) based on the Manduca sexta antennal lobe and mushroom bodies. This NN outperforms standard machine learning methods at rapid learning of digits from the Modified National Institute of Standards and Technology database. Figure 1a courtesy of [13], 1b courtesy of [4].*

Nonprofit Org
U.S. Postage
PAID
Permit No 360
Bellmawr, NJ

siam
SOCIETY for INDUSTRIAL and APPLIED MATHEMATICS
3600 Market Street, 6th Floor
Philadelphia, PA 19104-2688 USA

siam news
Volume 52/ Issue 6/ July/August 2019

4 Food Distribution: How to Give Fairly with Less Waste
Food banks are often tasked with effectively distributing a limited food supply to vulnerable populations. They aim to fairly provide food to those with the greatest need while simultaneously minimizing waste. Jenny Morber overviews a mathematical model that helps food banks balance equity and effectiveness under conditions of uncertain capacity. Specifically, she describes a robust optimization model for capacity uncertainty that distributes goods by absorbing random capacity changes across multiple counties.

5 Trust Me. QED.
Michael Heroux emphasizes the need to recognize the importance of computational reproducibility. Fully providing the resources necessary for researchers to repeat an experiment has traditionally presented many challenges. However, Heroux insists that given both computation's increasingly critical role in science and the availability of new tools and processes, the time has come to improve replicability.

6 Technical and Organizational Challenges for Data Scientists
Catherine Micek of 3M recounts her trajectory from an initial academic appointment to an industrial position. The emergence of the field of data science coincided nicely with Micek's transition to industry, allowing her to explore a range of roles including data scientist, software developer, and predictive modeler across the insurance, energy, and finance sectors.

8 The Virtual Ecologist
The small island of Isle Royale in Lake Superior is a natural laboratory harboring an idealized predator-prey system of wolves and moose. This presents the perfect locale for researchers to study the effect of climate on predator-prey dynamics. Easton White and Alan Hastings describe how mathematical models enable such analysis by evaluating the effects of different climatic conditions over several decades of experiment on multiple replicates of the island.

Brain Cancer

Continued from page 1

The team discovered that absolute tumor size was a less important metric than tumor position on the growth curve. Swanson and her colleagues use the term "days gained" to describe the result: does the treatment turn back the clock on cancer proliferation and buy the patient more time? Estimating days gained requires an understanding of the time-dependent growth kinetics pertaining to the individual's cancer, which is precisely what the PI model attempts to do.

A Model for Tumor Growth

As for many other tumors, neurosurgeons commonly begin glioblastoma treatment by surgically removing as much of the cancer as possible before following up with chemotherapy and radiation. However, glioblastoma is more diffuse than most cancers; because the tumor extends into healthy tissue, it is nearly impossible for surgeons to remove all cancer cells without damaging the brain.

To make matters worse, the degree of diffusivity varies widely among patients, and MRI scans alone are not particularly good at distinguishing the nuances of these cases. "Doctors don't really have a clean way of knowing the difference between one patient's tumor being really diffuse and another patient's tumor being really nodular, or which tumor is growing faster than another," Hawkins-Daarud said. "MRI detects what the cancer cells have done to the environment, but it can't specifically say 'this is a tumor.' It can't identify the boundary [of the glioblastoma]."

The uncertainty in measuring that boundary means that clinicians struggle to determine which treatments are working and which require adjustment. However, glioblastoma's diffusivity also makes it amenable to a reaction-diffusion model — a common type of equation in mathematical biology. The PI model approximates the tumor's growth in space and time by treating it as a continuous fluid [3]:

$$\frac{\partial c}{\partial t} = \nabla \cdot \left(D \nabla c \right) + \rho c \left(1 - c \right),$$

where c is the tumor cell density. The free parameters D and ρ respectively quantify the cancer cells' diffusion and rate of proliferation. Assuming a spherical tumor, the solution to the PI equation far from the tumor center takes the form of a traveling wave with velocity $2\sqrt{D\rho}$ and steepness D/ρ.

These parameters are not directly measurable. To infer their values, Swanson's team used MRI measurements for 160 glioblastoma patients [1]. They obtained an estimate of tumor growth and proliferation by comparing two MRI scans for each patient, then applied a Bayesian framework [2] to quantify uncertainties in both the data and model. These efforts yielded a means of classifying patient responses to treatment in terms of days gained (see Figure 2).

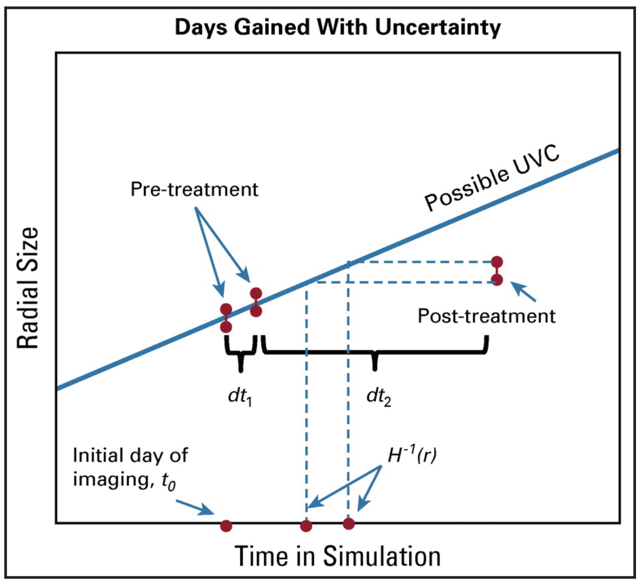

Figure 2. *Estimated radial size of a tumor before and after treatment, where the treated tumor size corresponds to an earlier stage of growth according to the model. This allows researchers to estimate the days gained with a particular treatment. Figure courtesy of [2].*

"I don't think [the PI model] is good at giving precise boundaries of tumor cell density throughout the brain," Hawkins-Daarud said. "However, it is good with helping us conceptually 'bin' patients into categories." The days-gained metric identified via the PI model proved to be a much better predictor than tumor size alone, thanks to incorporation of cancer kinetics.

"The difference in overall survival for patients with a larger days-gained value was statistically significant over those who had the smaller days-gained value," Hawkins-Daarud continued. "Our hope is that [the model] will be able to identify when a therapy is truly failing and you should change it, or when a therapy is being useful and you should stay on it — even though it looks like it may not be as good as you might expect."

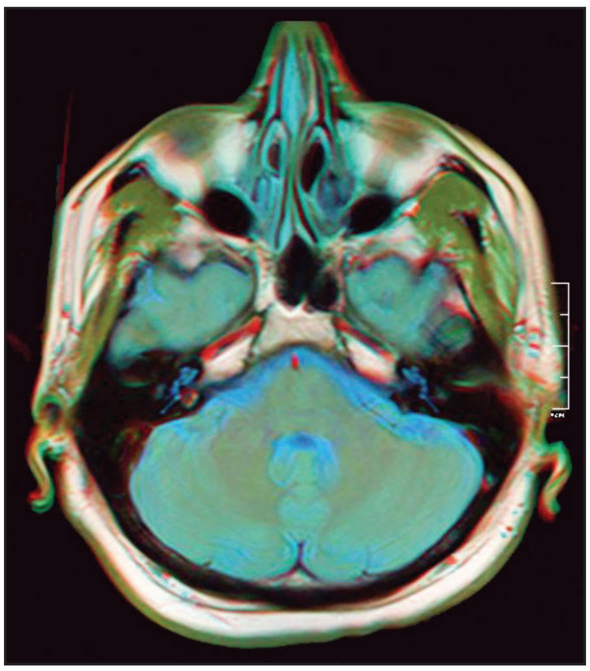

Figure 1. *Magnetic resonance imaging scan of the brain. Public domain image.*

Hope is a Thing with Equations

The PI model is deterministic and treats tumors as continuous fluids, whereas real glioblastoma consists of discrete cells that spread more haphazardly. For this reason, Swanson, Hawkins-Daarud, and many other researchers are combining forces to create better models that incorporate cancer kinetics, machine learning, and cellular automata, along with a wider range of medical data. The preliminary results of these efforts are not yet published, but Hawkins-Daarud believes that they hold a great deal of promise.

Even so, the problem can still seem insurmountable. Cancer is not a single disease, but rather a large set of conditions with many causes and a number of common features. The PI model enables better understanding of glioblastoma's specific traits; however, this does not work for most cancers, which metastasize and are non-diffuse. Yet hope is a relative thing in cancer research — for mathematical oncologists as much as for doctors and patients.

"The math isn't going to cure the cancer," Hawkins-Daarud said. "But I think that math can certainly help optimize the process of finding a cure. We are actually in the midst of talking to various drug companies to try and incorporate our response metrics into the clinical trials to see if we can speed up the proceedings."

Even a few months of extra time acquired from improved treatments is significant to glioblastoma patients and their loved ones. While math alone will not provide this time, the PI model shows that it can help gain some valuable days.

References

[1] Baldock, A.L., Ahn, S., Rockne, R., Johnston, S., Neal, M., Corwin, D., …Swanson, K.R. (2014). Patient-specific Metrics of Invasiveness Reveal Significant Prognostic Benefit of Resection in a Predictable Subset of Gliomas. *PLoS One*, *9*(10), e99057.

[2] Hawkins-Daarud, A., Johnston, S.K., & Swanson, K.R. (2019). Quantifying Uncertainty and Robustness in a Biomathematical Model-Based Patient-Specific Response Metric for Glioblastoma. *JCO Clin. Cancer Inform.*, *3*, 1-8.

[3] Rockne, R., Alvord Jr., E.C., Rockhill, J.K., & Swanson, K.R. (2009). A mathematical model for brain tumor response to radiation therapy. *J. Math. Biol.*, *58*(4-5), 561-578.

Matthew R. Francis is a physicist, science writer, public speaker, educator, and frequent wearer of jaunty hats. His website is BowlerHatScience.org.

siam news

ISSN 1557–9573. Copyright 2019, all rights reserved, by the Society for Industrial and Applied Mathematics, SIAM, 3600 Market Street, 6th Floor, Philadelphia, PA 19104-2688; (215) 382-9800; siam@siam.org. To be published 10 times in 2019: January/February, March, April, May, June, July/August, September, October, November, and December. The material published herein is not endorsed by SIAM, nor is it intended to reflect SIAM's opinion. The editors reserve the right to select and edit all material submitted for publication.

Advertisers: For display advertising rates and information, contact Kristin O'Neill at marketing@siam.org.

One-year subscription (nonmembers): Electronic-only subscription is free. $73.00 subscription rate worldwide for print copies. SIAM members and subscribers should allow eight weeks for an address change to be effected. Change of address notice should include old and new addresses with zip codes. Please request address change only if it will last six months or more.

Printed in the USA.
siam *is a registered trademark.*

Neurosensory Integration

Continued from page 1

constraints [7]. MothNet applied tighter constraints and calibration to *in vivo* electrode recordings [5] and modeled the full AL-MB — subject to neuromodulators during learning. Unlike standard NN models, the olfaction processing model is dynamic in nature and contains neurons that obey a firing rate model of the form

$$\frac{d\boldsymbol{x}}{dt} = f\left(\boldsymbol{x}, \boldsymbol{W}\right),$$

where $\boldsymbol{x}$ depicts the neuronal dynamics and $\boldsymbol{W}$ represents the connection weights trained for odor classification at the readout neurons (see Figure 1, on page 1).

New rewarded odors are subjected to neuromodulators, which provide weight updates between neurons i and j as $\triangle \boldsymbol{W}_{ij} = \gamma f_i(t) f_j(t)$ — where $f_j(t)$ is the firing rate of the $j\mathrm{th}$ neuron. A crucial payoff of this approach was reproduction of the actual insect's rapid learning; MothNet attained close to 80 percent accuracy on the Modified National Institute of Standards and Technology (MNIST) database, given just one to 10 samples per class. It outperformed standard ML techniques—including specialized one-shot methods—in this rapid learning regime [4]. These results indicate that one can usefully port the AL-MB structure to a ML context.

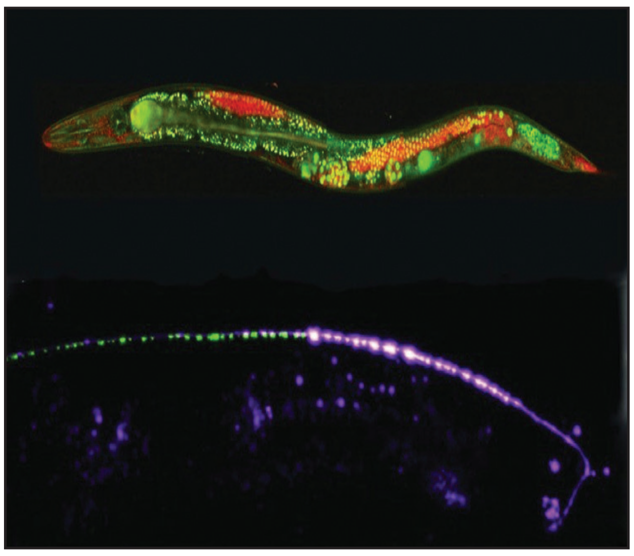

Figure 2. *Whole-brain imaging has produced novel datasets of the activity of a network with stereotyped connectivity. Figure courtesy of [8].*

In short, tighter biological constraints can yield novel, useful results. Such constraints also inspire new perspectives; the idea of massive training data—which NNs take for granted—is alien to the insect AL-MB. A bug requiring 60,000 training samples (routine for MNIST studies) would be dead. Striving for 99.9 percent accuracy is also unfamiliar to the insect AL-MB, since it has no need for such precision. Insect architectures instead excel at rapid, low-fidelity learning and outperformed ML methods in this regime. Given the rich variety of biological NNs, many other learning architectures and mechanisms—including dynamic data processing and energy/power constraints associated with neuronal wiring—remain unexplored in the context of ML.

C. elegans and the 300

C. elegans has 302 neurons for performing all of its varied life-sustaining tasks, including chemotaxis, predator avoidance, and mating. These actions typically involve a sequence of primary behaviors — such as forward crawling, backward crawling, omega turns, and head sweeps. The worms tend to live in *noisy* stimulus environments and use a small number of sparsely-connected neurons to robustly navigate. The full connectome is characterized, as is almost every aspect of *C. elegans* anatomy.[1] Despite a wealth of detailed knowledge, current state-of-the-art ML has difficulty capturing how the worms survive and compute neurosensory information. *C. elegans* may have broader principles of network design that one can potentially learn in order to engineer robust functional behaviors with limited resources.

Much existing modeling work on *C. elegans* does not include the dynamical structure of the worm [2, 6, 9], whose connectome encodes a number of critical behaviors. Indeed, the connectomic structure itself seems ideally designed for the integration of proprioceptic feedback for efficient locomotion [9]. It is thus highly likely that *C. elegans*' specific wiring diagram is engineered for its functional repertoire of behaviors. Emergence of the worm's whole brain imaging will further revolutionize our ability to posit models constrained by known structure and dynamics (see Figures 2 and 3) [8, 12]. Such data can allow us to mathematically move toward plausible data-driven control models of the form

$$\frac{d\boldsymbol{x}}{dt} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{u},$$

where the data alone initiates the discovery of matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ and control signal $\boldsymbol{u}$ [11]. We can perform this regression process in a supervised or unsupervised fashion, therefore framing *C. elegans* in a classic control context and providing a firm theoretical foundation for the characterization of neuronal control laws. It is very likely that *C. elegans* will be the first model organism that researchers understand completely from a neurosensory integration perspective.

Outlook: New Data, New Models

Our work has demonstrated the importance of incorporating organism-specific knowledge into modeling efforts to truly comprehend their design principles. We can only realize the amazing promise of robust signal processing in noisy environments using small networks via the newly possible combination of *in vivo* datasets and physiological constraints. The emergence of both mathematical methods and innovative recordings allows for significant improvements in our understanding of neurosensory integration, network functionality, and robustness. We therefore might also expect the imbuing of NN architectures with biological constraints to yield significant improvements in ML and AI structures. This could lead to increased robustness, significant reduction in training data, and/or more energy- and memory-efficient network designs. Ultimately, the rich interplay between neuroscience and ML is set to accelerate high dividends in science and technology.

References

[1] Bartunov, S., Santoro, A., Richards, B., Marris, L., Hinton, G.E., & Lillicrap, T. (2018). Assessing the scalability of biologically-motivated deep learning algorithms and architectures. *Adv. Neur. Info. Process. Syst.*, 9390-9400.

[2] Boyle, J.H., Stefano, B., & Cohen, N. (2012). Gait modulation in C. elegans: an integrated neuromechanical model. *Front. Comp. Neurosci., 6*(10).

[3] Dacks, A.M., Riffell, J.A., Martin, J.P., Gage, S.L., & Nighorn, A.J. (2012). Olfactory modulation by dopamine in the context of aversive learning. *J. Neurophysiol., 108*(2), 539-550.

[4] Delahunt, C.B., & Kutz, J.N. (2019). Putting a bug in ML: The moth olfactory network learns to read MNIST. *Neur. Net.* To be published.

[5] Delahunt, C.B., Riffell, J.A., & Kutz, J.N. (2018). Biological Mechanisms for Learning: A Computational Model of Olfactory Learning in the Manduca sexta Moth, with Applications to Neural Nets. *Front. Comp. Neurosci., 12*(102).

[6] Fieseler, C., James, K.-G., & Kutz, J.N. (2018). The control structure of the nematode Caenorhabditis elegans: Neuro-sensory integration and proprioceptive feedback. *J. Biomech., 74*, 1-8.

[7] Huerta, R., & Nowotny, T. (2009). Fast and Robust Learning by Reinforcement Signals: Explorations in the Insect Brain. *Neur. Comp., 21*(8), 2123-2151.

[8] Kato, S., Kaplan, H.S., Schrödel, T., Skora, S., Lindsay, T.H., Yemini, E., … Zimmer, M. (2015). Global brain dynamics embed the motor command sequence of Caenorhabditis elegans. *Cell, 163*(3), 656-669.

[9] Kunert, J.M., Proctor, J.L., Brunton, S.L., & Kutz, J.N. (2017). Spatiotemporal feedback and network structure drive and encode Caenorhabditis elegans locomotion. *PLoS Comp. Biol., 13*(1), e1005303.

[10] Martin, J.P., Beyerlein, A., Dacks, A.M., Reisenman, C.E., Riffell, J.A., Lei, H., & Hildebrand, J.G. (2011). The neurobiology of insect olfaction: Sensory processing in a comparative context. *Prog. Neurobiol., 95*, 427-447.

[11] Proctor, J.L., Brunton, S.L., & Kutz, J.N. (2016). Dynamic mode decomposition with control. *SIAM J. Appl. Dyn. Sys., 15*(1), 142-161.

[12] Shipley, F.B., Clark, C.M., Alkema, M.J., & Leifer, A.M. (2014). Simultaneous optogenetic manipulation and calcium imaging in freely moving C. elegans. *Front. Neuro. Circ., 8*(28).

[13] Srinivasan, S., Greenspan, R.J., Stevens, C.F., & Grover, D. (2018). Deep(er) Learning. *J. Neurosci., 38*(34), 7365-7374.

[14] Wilson, R.I., Turner, G.C., & Laurent, G. (2004). Transformation of olfactory representations in the Drosophila antennal lobe. *Science, 303*(5656), 366-370.

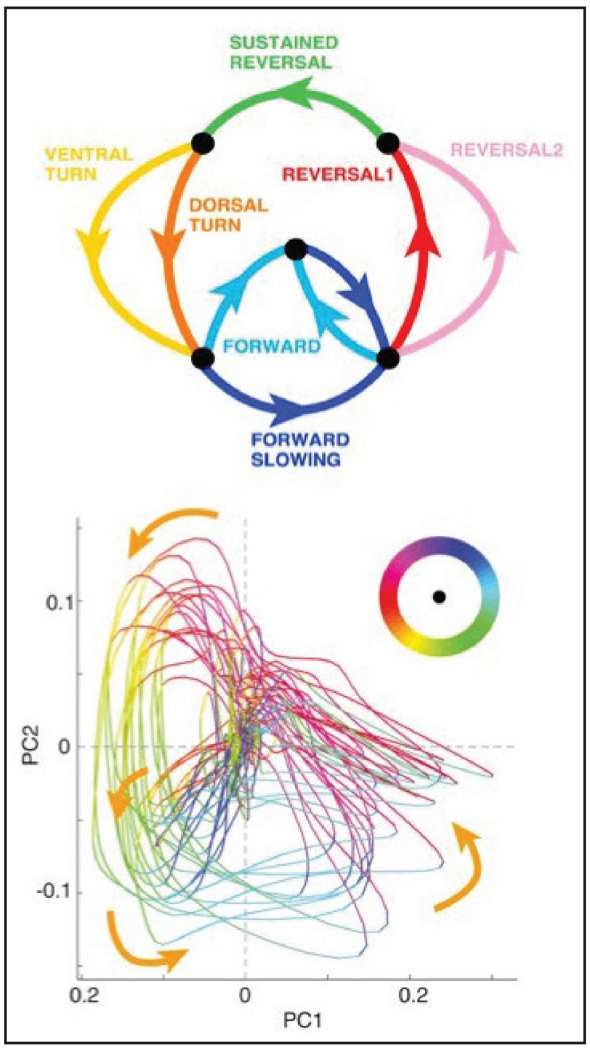

Figure 3. *The neural data from Figure 2 live on a low-dimensional manifold with discrete states. Figure courtesy of [12].*

Charles B. Delahunt is a postdoctoral fellow in the Department of Applied Mathematics at the University of Washington, where he develops bio-inspired neural network architectures and machine learning (ML) algorithms. He also applies ML methods to global health problems at Global Good in Bellevue, Wash. Charles Fieseler is a graduate student in the Department of Physics at the University of Washington, where he develops data-driven biophysical models of the interaction of neuronal networks and sensory processing with biomechanics. J. Nathan Kutz is a professor of applied mathematics at the University of Washington, where he works at the intersection of data analysis and dynamical systems.

SIAM at the Coalition for National Security Research

On May 22, 2019, the Coalition for National Security Research—of which SIAM is a member—jointly hosted the second annual Science, Technology, & Innovation Exchange briefing with the Department of Defense (DoD) Basic Research Office. The event took place on Capitol Hill in Washington, D.C., and showcased remarks by members of Congress, DoD officials, and scientists. A corresponding poster exhibition highlighted the many ways in which the U.S. research enterprise confronts scientific challenges currently facing the defense community. SIAM Science Policy Fellowship recipient Robert Edman (Adventium Labs) presented a poster that emphasized applied mathematics' relevance to the DoD mission.

SIAM Science Policy Fellowship recipient Robert Edman (Adventium Labs) presents a poster illustrating the importance of applied mathematics to the Department of Defense mission at the second annual Science, Technology, & Innovation Exchange briefing, which took place in May on Capitol Hill in Washington, D.C. Here he converses with Bindu Nair (Office of the Secretary of Defense) and Representative Bill Foster (D-IL). Photo courtesy of Eliana Perlmutter.

---

[1] https://www.wormatlas.org/

Food Distribution: How to Give Fairly with Less Waste

By Jenny Morber

When a fledgling company writes its business plan, the first worry is often demand — namely whether enough people will buy the product. Nonprofit supply chains face different motivations and uncertainties, and those that address food scarcity may instead ask how to best distribute a limited supply. But what does "best" mean? Food banks aim to fairly provide the most food to those with the greatest need while also minimizing waste. Sometimes the former and latter goals are at odds.

In a recent paper, Irem Sengul Orgut (Lenovo), Julie Ivy and Reha Uzsoy (North Carolina State University), and Charlie Hale (Food Bank of Central & Eastern North Carolina (FBCENC)) describe a model to help food banks balance equity and effectiveness when food supply is uncertain [1]. The researchers draw on an eight-year partnership with the FBCENC to optimize food distribution at the county level.

The FBCENC, a Feeding America affiliate based in Raleigh, N.C., distributes food to 34 counties through six branch locations. Unfortunately, high need and low supply make it impossible to satisfy demand. Hence, the FBCENC instead seeks to dispense food across counties in proportion to the population experiencing poverty, so that each food-insecure person ideally receives the same amount of donated food. In short, it strives for equity. However, data suggests that distribution is inequitable, with some counties remaining underserved and others reporting food waste.

More effective food distribution strategies would minimize food waste by directing the majority of donations to high-capacity food banks. But this solution means that people in some areas would be consistently underserved. "If you only care about equity, one very trivial solution is just to not ship anything," Sengul Orgut said. "It's perfectly equitable but it's not effective. On the other hand, if you only want to ship out maximum food, you would look at each distribution location and stack its capacity to the limit. But then you are not looking at the population size that the locations serve, and their distribution ends up becoming inequitable. So we have these two contradicting objectives."

Robust Food Distribution Algorithm

```
if $\Delta\min_{i\in J}\left\{\frac{\hat{C}_i}{D_i}\right\} < S$ then
    $B = \left\{j \in J \mid \frac{\hat{C}_j}{D_j} = \min_{i\in J}\left\{\frac{\hat{C}_i}{D_i}\right\}\right\}$
    $\varphi_j^{min} = \min\left\{1, \frac{\hat{C}_j D_B - \hat{C}_B D_j}{D_B \varepsilon_j}\right\}, \forall j \in J$
    if $\sum_{l=1}^{n} \varphi_l{}^{min} \geq \rho$ then
        $\varphi_j^* \leftarrow \varphi_j^{min}$
        quit
    else
        $J_E = \left\{j \in J \mid \frac{\hat{C}_j - \varepsilon_j}{D_j} > \frac{\hat{C}_B}{D_B}\right\}$
    end if
else
    $\varphi_j = \min\left\{\frac{\hat{C}_j\Delta - SD_j}{\varepsilon_j\Delta}, 1\right\}$
    if $\sum_{l=1}^{n} \varphi_l{}^{min} \geq \rho$ then
        $\varphi_j^* \leftarrow \varphi_j^{min}$
        quit
    else
        $J_E = \left\{j \in J \mid \frac{\hat{C}_j - \varepsilon_j}{D_j} > \frac{S}{\Delta}\right\}$
    end if
end if

$\varphi_j \leftarrow 1, \forall j \in J_E$
$J_I = (J_E)^C$
Sort $J_I$ in a decreasing order of $\frac{\hat{C}_j - \varepsilon_j}{D_j}$
while $j \leq |J_I|$ do
    $\varphi_j = \frac{\rho - |J_E| - j + 1 - \sum_{l=j+1}^{|J_I|}\frac{C_{J_I(l)}}{\varepsilon_{J_I(l)}} + \frac{C_{J_I(j)}}{D_{J_I(j)}}\sum_{l=j+1}^{|J_I|}\frac{D_{J_I(l)}}{\varepsilon_{J_I(l)}}}{1 + \frac{\varepsilon_{J_I(j)}}{D_{J_I(j)}}\sum_{l=j+1}^{|J_I|}\frac{D_{J_I(l)}}{\varepsilon_{J_I(l)}}}$
    if $\varphi_j \leq 1$ then
        for $1 \leq i \leq j$ do
            $\varphi_{J_I(i)}^* \leftarrow 1$
        end for
        $\varphi_{J_I(j)}^* \leftarrow \varphi_j$
        for $i + 1 \leq i \leq |J_I|$ do
            $\varphi_{J_I(i)}^* \leftarrow \frac{\hat{C}_{J_I(i)}D_{J_I(j)} - \hat{C}_{J_I(j)}D_{J_I(i)} + \varepsilon_{J_I(j)}D_{J_I(i)}\varphi_j}{\varepsilon_{J_I(i)}D_{J_I(i)}}$
        end for
        quit
    end if
end while
$X_j^* \leftarrow D_i\, min_{j\in J}\left\{\frac{\hat{C}_j - \varepsilon_j\varphi_j^*}{D_j}\right\}, \forall j \in J$
```

Figure 1. *The robust optimization algorithm for capacity uncertainty provides a practical roadmap for food distributors who seek to optimize the sometimes-competing goals of effectiveness and equity via the robust optimization model for capacity uncertainty (C-RM). Figure courtesy of [1].*

Adding to this complexity is the fact that the capacity—amount of food each agency can receive in a given week or month—fluctuates in the FBCENC distribution area. Demand remains mostly stable, but budget, transportation, workforce, and food storage capability change over time. Some food banks must temporarily close if they fail to meet required equity targets, and volunteer numbers wax and wane with the changing seasons.

Fairness in the Face of Uncertainty

The robust optimization model for capacity uncertainty (C-RM) helps food banks distribute goods in a way that can absorb random capacity changes across the region's multiple counties. The team reasoned that while capacity may be uncertain, it is likely uncertain within a range. The C-RM allows users to specify a particular capacity robustness within this range while attempting to maximize the total amount of distributed food. One can then balance equity and effectiveness according to each decision-maker's specific needs.

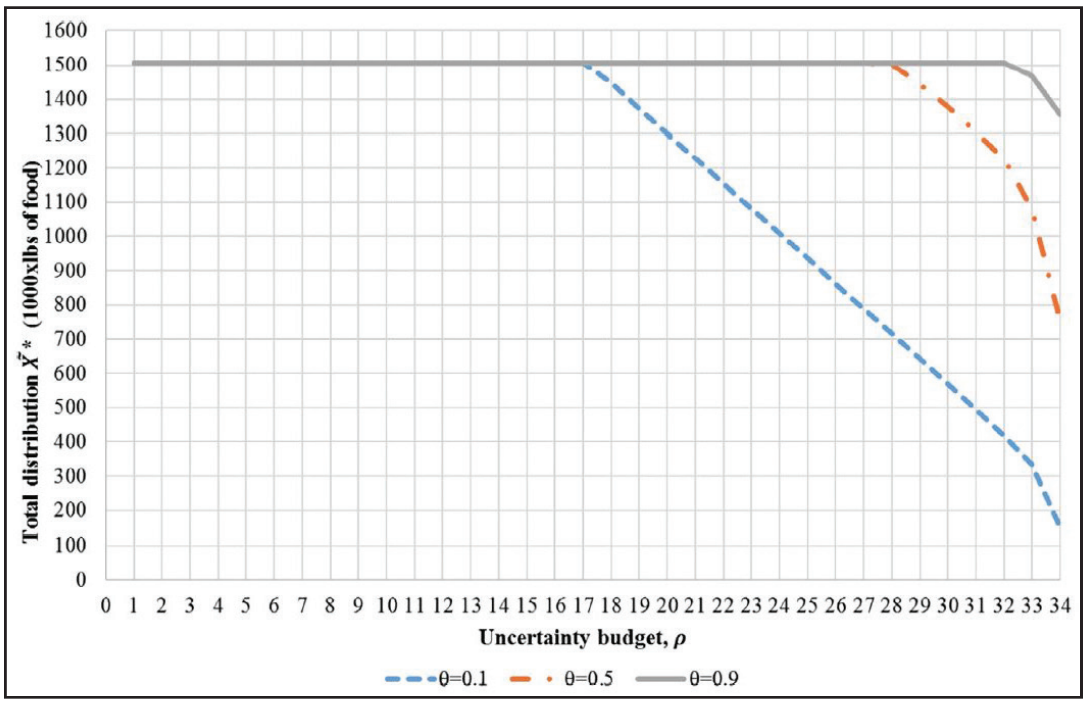

Figure 2. *The plot of total optimal food distribution versus uncertainty budget ρ for changing capacity range ϑ indicates that uncertainty and greater capacity fluctuations necessitate decreases in total food distribution to prevent food waste. Figure courtesy of [1].*

The researchers formulated the C-RM as follows, where n is the number of counties in the service region:

$$\max\sum_{l=1}^{n} X_l, \qquad (1)$$

subject to the following constraints:

$$\frac{X_j}{\Sigma_{l=1}^{n} X_l} - \frac{D_j}{\Sigma_{l=1}^{n} D_l} = 0 \forall j \in J, \qquad (2)$$

where j is the county index, X_j is the amount of food shipped to county j, J is the set of counties in the food bank service area, D_j is the demand of county j, and $\widehat{C}_j$ is the nominal capacity value for county j.

$$X_j \leq \widehat{C}_j - \varepsilon_j\varphi_j \forall j \in J, \qquad (3)$$

where ε_j is the maximum negative deviation of county j's capacity from its nominal value $\widehat{C}_j$, and φ_j is the fraction of deviation in county j's capacity from its nominal value.

$$\sum_{l=1}^{n} \varphi_l \geq \rho, \qquad (4)$$

where ρ is the robustness control parameter (the uncertainty budget).

$$\sum_{l=1}^{n} X_l \leq S, \qquad (5)$$

where S is total supply.

$$\varphi_j \leq 1 \forall j \in J. \qquad (6)$$

The objective function (1) seeks to maximize the total amount of food shipped from the food bank while maintaining equity, as expressed in (2). With constraint (3), the model prevents capacity overestimation by stipulating that the quantity of food sent to a county must be less than or equal to its nominal capacity minus the allowed deviation. Since county agencies are not as well equipped for long-term food storage as their suppliers, overestimation of agency capacity leads to greater food waste.

To avoid an overly-conservative solution, ρ limits the number of parameters that can take their worst-case values. In this model, the minimum capacity-to-demand ratio R is given by

$$R = \min_{j\in J}\{\frac{C_j}{D_j}\},$$

where C_j is a random variable that fluctuates over the interval $\left[\widehat{C}_j - \varepsilon_j, \widehat{C}_j\right]$. R reaches its minimum and maximum when capacities of all counties respectively achieve their lowest and highest values. Unlike in previous models, it must be treated as a random variable. Because county capacities are uncertain, the county with the minimum CD ratio may fluctuate. The same is true of the value of that ratio.

The C-RM's optimal solution finds deviations from an agency's nominal capacity that satisfy constraint (4) with minimal effect on total food distribution. Sengul Orgut et al. again discern that a bottleneck county's capacity-to-demand ratio sets the overall food distribution when capacity is the constraining factor. In this case, one can express the excess capacity fraction as

$$\varphi_j^{\min} = \min\{1, \frac{\widehat{C}_j D_B - \widehat{C}_B D_j}{D_B\varepsilon_j}\}.$$

No bottleneck exists if the nominal instance is supply-constrained instead. Here, one defines the proportion of excess capacity fraction by

$$\varphi_j^{\min} = \min\{1, \frac{\widehat{C}_j\Delta - SD_j}{\varepsilon_j\Delta},$$

where $\triangle$ is the capacity deviation coefficient and $S = \Delta\min_{j\in J}\frac{\widehat{C}_j - \varepsilon_j\varphi_j}{D_j}$. If the calculated excess capacity fraction is greater than or equal to the uncertainty budget ρ, total distribution is equal to the nominal problem, such that $\Sigma_{l=1}^{n} X_l^* = \min\{S, R\Delta\}$. But if the excess capacity fraction is less than ρ, a set of counties can achieve maximum deviation from their nominal values without affecting the optimal solution. That set is expressed as

$$J_E = \{j \in J \mid \frac{\widehat{C}_j - \varepsilon_j}{D_j} > \min\{\frac{S}{\Delta}, \frac{\widehat{C}_B}{D_B}\}\}.$$

One can determine the optimal solution by listing the counties in decreasing order of their $\frac{\widehat{C}_j - \varepsilon_j}{D_j}$ ratios. The optimal distribution is $X_i^* = D_i\min_{j\in J}\{\frac{\widehat{C}_j - \varepsilon_j\varphi_j^*}{D_j}\}$.

To help food banks implement this model, the team translated it into an algorithm that first detects whether the problem is supply- or capacity-constrained, and then finds optimized distributions for each county according to the model. This is the robust optimization algorithm for capacity uncertainty (see Figure 1).

But does it work in real life? To illustrate the model's applicability to real-life scenarios, the researchers input historical data from the FBCENC and ran an experiment based on the assumption that the food bank has received 2,600,255 pounds of food — the average monthly dry goods donation in 2014. They set $\widehat{C}_j$ equal to the 90th percentile of the empirical distribution of the amount of food shipped to each county from July 2012 to June 2015, and performed experiments for capacity deviation coefficient $\theta = \{0.1, 0.5, 0.9\}$ to examine capacity uncertainty's effect on total food distribution (see Figure 2).

As the robustness control parameter increases, one must reduce food shipments to protect against larger deviations from nominal capacity. To optimize distribution at a real food bank, employees can set capacity ranges using historical data and their own experiences/goals. A less risk-averse food bank—or one with well-known capacity ranges—would be expected to set θ closer to one, while a more risk-averse bank—or one with increased capacity fluctuation—may set θ at a lower value.

See **Food Distribution** *on page 7*

Trust Me. QED.

By Michael A. Heroux

Consider a standard SIAM journal article containing theoretical results. Each theorem has a proof that typically builds on previous developments. Since every theorem stems from a firm foundation, the research community can trust a result without further evidence. One could thus argue that a theorem does not require a proof because surely an author would not publish it if no proof existed to back it up. However, respectable reviewers and editors expect proofs without exception, and papers containing proof-less theorems will likely go unpublished.

Next, consider a paper with computational results and look for details on the generation of these results. Suppose your task is to qualitatively obtain the same outcome. Does the article contain enough information to support you in this task? If your experience is like mine, you will find that while some authors provide details that give you a chance to succeed, many others do not. Contrary to the rigor of justifying theoretical conclusions, the meticulousness (or lack thereof) applied to many published computational results is the equivalent of "Trust me. QED."

In defense of the insufficiency currently associated with computational results, we must acknowledge that fully capturing the input conditions and specifying the necessary execution environment to repeat a computational experiment has traditionally been very challenging. Providing this information for another scientist's use is even more so. Furthermore, scientists trained in formal academic environments were seldom exposed to the tools, practices, and processes used to confirm result reproducibility. But with computation's increasingly critical role in science and engineering—and the availability of new tools, practices, and processes to make the job easier—we can and must improve reproducibility.

SOFTWARE AND PROGRAMMING

The past decade has seen the emergence of new platforms that support rigorous software management. Environments like GitHub and GitLab provide users with the ability to develop, test, and integrate software changes using efficient collaborative workflows. The wide usage and accessibility of these platforms ensure that community members can easily document and publish a description of the software environment used to compute a result. Additionally, container technologies such as Docker support encapsulation of the full software environment that enables portable execution on many computer systems with very little overhead. With these improved tools, scientific software developers can adopt new workflows and practices that make reproducible computational results more feasible.

Reproducibility as the Key Focus

Both the growing importance of computation in science and engineering, as well as the new tools, practices, and processes presently available, yield more opportunities to elevate the quality of computational science. Raising expectations for reproducible results provides the incentive for realizing these opportunities.

Reproducibility as a fundamental goal in computational science is very powerful. First, the computation should be repeatable with realization of the same qualitative result. Second, the result must be usable, trustworthy, and extensible as a step toward scientific progress without concern of moving in the wrong direction. While the details can be more complicated and require careful articulation, the basic concepts should be easy to grasp.

Demanding Reproducibility Improves Productivity and Sustainability

Reproducibility expectations dictate that the software, input data, and execution environment used to produce published results must be available in the future. Validation of computational results using an independent software, data, and execution environment would provide the most rigorous evidence of reproducibility (commonly termed "replicability"). However, verification of an author's findings using his or her own environment is still valuable and arguably the first step in a prudent approach to obtaining trustworthy results.

The increased incentive to improve developer productivity and software sustainability is a serendipitous outcome of pursuing reproducible results. Authors have compelling reasons to invest in source code management and annotation, data provenance, improved documentation, automated build tools, and comprehensive tests. The higher workload will initially delay results, raise costs, and perhaps require a research team to invest in skills or people that are not directly focused on scientific questions. Yet in the long run, researchers will see overall improvement in scientific output, especially when accounting for the increased trustworthiness of computational results. The ensuing increase in quality is fundamentally valuable.

Metadata and Meta-computation

As we have observed, trustworthy computational results require something beyond published outcomes. In most cases, program source code, documentation, input data, and details about the computing environment are essential. A complete software container that provides the entire computing environment in a single file is even better. However, metadata is not sufficient in all situations. In particular, some computational results are obtained from computing environments with limited access, such as supercomputing centers. In these cases, one can perform additional computations — like testing conservation properties or operator symmetries.

For example, if a computation involves the application of a symmetric linear transformation A to a vector—as would happen in a Krylov iterative solver—one can compute the expressions $x^T(Ay)$ and $y^T(Ax)$ for two random vectors x and y and obtain the same scalar result, up to round-off error. Putting this simple test into a preamble computation prior to executing the main code is inexpensive and useful. Meta-computations can provide suitable substitutes or supplements to metadata for boutique computing environments, wherein reviewers or authors may be unable to access the same computing environment in the future. In addition to supercomputers, experimental computer systems, testbeds, and configurable hardware systems are all transient computational environments that can be difficult to re-instantiate at a later date. If their performance is inexpensive, meta-computations can also help debug new functionality and assure the sanity of novel software environments.

See **Trust Me** *on page 7*

Technical and Organizational Challenges for Data Scientists

By Catherine Micek

As a freshman in college, I decided that I wanted to be a math professor. I'd always valued learning and had been interested in teaching from a young age. My desire to teach mathematics was twofold. I enjoyed the challenge of communicating a dense subject by breaking down complex ideas and explaining them step by step, and I loved learning about the applications of mathematics across diverse fields like physics, computer science, and music. As a professor, I knew I could continue learning through research while sharing my love of math with students. With this goal in mind, I pursued a Ph.D. in applied mathematics.

CAREERS IN MATHEMATICAL SCIENCES

Yet by the time I obtained my degree, I was no longer certain that I wanted to continue in academia. Furthermore, the then-recent 2008 financial crisis meant that the academic job market was poor. It was also very important for me to stay near my family in Minnesota, and academic career paths would have likely not allowed that. So I decided to accept a job as a visiting assistant professor at Augsburg College in Minneapolis, Minn., and start exploring industry career options. The buzz around data science began shortly thereafter.

Management magazines like the *Harvard Business Review* were touting data science as a vehicle that could enable companies to transform their operations, decision-making, and product development. Because the underlying business problems were complex, data scientists would need a strong quantitative and computational skill set to solve them. Although my Ph.D. was not in statistics or machine learning, I had developed extensive computational skills during my thesis work and was excited about learning new areas of mathematics. My transition to data science thus began early in my career with my first industry job as a predictive modeler at the Travelers Companies, Inc. in 2012.

Since then, my data science positions have encompassed a variety of technical roles—such as data scientist, software developer, and predictive modeler—across diverse industries including insurance, energy, and finance. My job responsibilities have included the technical work of solving data science problems and automating the solutions, and the organizational work of educating and training my business colleagues about the field itself. Each data science job that I have held falls within the broader field of decision sciences. To me, the role of data scientist involves using sophisticated mathematical and computational techniques to develop algorithms or analyses that extract meaningful information from data. As a data scientist, I have built statistical models to price insurance policies, created custom machine learning algorithms to detect anomalies on the electrical grid, and developed forecasting models to predict sales.

The data scientist's algorithm or analysis is often embedded in a *data science product*, which automates data processing through the algorithm via a *software delivery mechanism*. The level and technical sophistication of automation depends on the team's collective skill set. For example, when my team included software engineers, the implementation was a custom-built software package fully automated from end to end. But with no developers to assist with automation, the data scientists had to devise simple, semi-automated implementations — such as an R script run on an ad-hoc basis to generate Excel workbook output.

An *end user* consumes the output of the data science product. In my experience, he/she is typically a business analyst; I have thus often had to translate mathematical output into actionable business information. For instance, when explaining a prediction interval for a sales forecast, my description would be something like "a prediction interval quantifies the uncertainty in our forecast by providing a best- and worst-case estimate of sales." The goal is to present the analyst with a high-level description of the math that explains how one can use the information for decision-making. This explanation is especially important if the business analyst is not relying solely on a model for the forecast, but rather combining the model output with other pieces of business intelligence.

When I begin tackling a data science business problem, I find it helpful to break the problem into five primary components: the business question, data, algorithm or analysis, delivery mechanism, and communication of results. I think of each component as either *technical* or *organizational*. The data, algorithm and analysis, and delivery mechanism constitute the technical components, while the business question and communication of results make up the organizational elements. In this framework, it becomes clear that the required skill set for an effective data scientist is a blend of hard and soft skills: a range of technical abilities in mathematics, programming, and software engineering; an understanding of the business domain; and strong communication techniques.

The primary point of debate that I have observed seems to be not *if* a blend of hard and soft skills is required for data scientists, but *what* skills should comprise that blend. I have heard domain experts argue that understanding the domain is more important than the sophistication of the associated data science tools. Conversely, I have witnessed technical experts argue

See **Data Scientists** *on page 8*

Catherine Micek (third from left) and her team, K-Means Business, hard at work during 3M's Data Intelligence Global Hackathon, which took place in April 2019. The team placed third. Image courtesy of Ghulam Jafferi.

Trust Me

Continued from page 5

Reproducible and Replicable

Two distinct approaches tend to emerge as research communities assume expectations of reproducible scientific results. The first is the activity of a reviewer using an author's experimental environment to duplicate that author's result. This activity can be thought of as verification, answering the question, "Did the author do things right?" A second endeavor involves using a different experimental environment to obtain a consistent result; this can be considered validation, answering the question, "Did the author do the right thing?" Independent validation surely results in the most powerful evidence of correctness, but both activities are important to improving the trustworthiness of scientific results.

As communities develop distinct activities in pursuit of trustworthiness, taxonomies arise that allow common synonyms to take on specific meanings. In computational science communities, "reproducible" is typically associated with the verification activity and "replicable" with the validation activity. However, not all communities use these terms consistently. For example, the Association for Computing Machinery (ACM) defines "reproducible" and "replicable" in essentially the opposite way [2-3]. Even so, the concepts remain the same and both actions are valuable.

Expecting Reproducibility: How to Get There

Introducing new research approaches will increase the time and effort required to produce computational results. We will need novel tools, methodologies, and workflows, and may have to recruit or collaborate with new people to gain additional expertise. Investing in productivity and sustainability can reduce future time and effort in pursuit of reproducibility and improve the quality of our work, but not immediately. Any improvement strategy should be incremental, both within a given research team and across collective teams in a community (see Figure 1).

Identifying teams within a community that have already made progress in reproducible computational results is a good starting point. Colleagues can often more readily adapt these teams' approaches than seek methods from dissimilar software communities. These early adopters can provide inspiration and practical advice to others. Many options exist to help teams work toward reproducible results, meaning that the *how* of improved reproducibility—beyond the need to move ahead incrementally—is straightforward. Adapting incentive systems is more challenging.

The real first step toward improving reproducibility of computational results is to expose its value, and SIAM can play a central role toward that end by rewarding authors for assuring reproducible results (see Figure 2). The ACM awards badges for papers whose results have been reviewed [1], and many conferences do the same. SIAM can also provide recognition for society members who help lead the community toward reproducible computational science. Beyond SIAM, funding agencies and employers—especially academic institutions—can also play important roles by rewarding people whose computational work is consistently reproducible.

Computational scientists are inherent problem solvers. Given the challenge and incentive to make our computational results reproducible, we will develop effective and efficient ways to meet that challenge. We will also serendipitously improve our productivity and the sustainability of our software environments, ultimately moving from "Trust me. QED." to trustworthy.

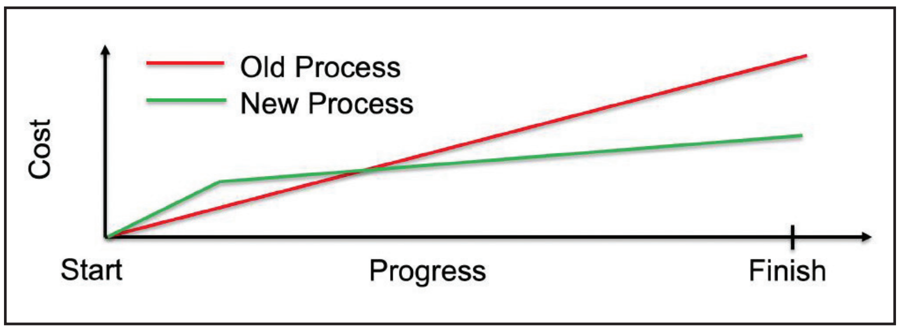

Figure 1. *Introducing increased reproducibility expectations will heighten time and cost toward obtaining computational results. We must raise expectations incrementally, improving how we work while pursuing our results. Finding and learning from early adopters while gradually increasing expectations over the span of a few years has been effective in other communities—such as the Association for Computing Machinery—where reproducibility efforts are rewarded. Figure courtesy of Michael A. Heroux.*

References

[1] Association for Computing Machinery (2018). Artifact Review and Badging. Retrieved from https://www.acm.org/publications/policies/artifact-review-badging.

[2] Heroux, M.A., Barba, L.A., Parashar, M., Stodden, V., & Taufer, M. (2018). Toward a Compatible Reproducibility Taxonomy for Computational and Computing Sciences. Technical report. *Office of Scientific and Technical Information, U.S. Department of Energy*. Retrieved from https://www.osti.gov/biblio/1481626-toward-compatible-reproducibility-taxonomy-computational-computing-sciences.

[3] National Academies of Sciences, Engineering, and Medicine (2019). *Reproducibility and Replicability in Science*. Washington, D.C: The National Academies Press.

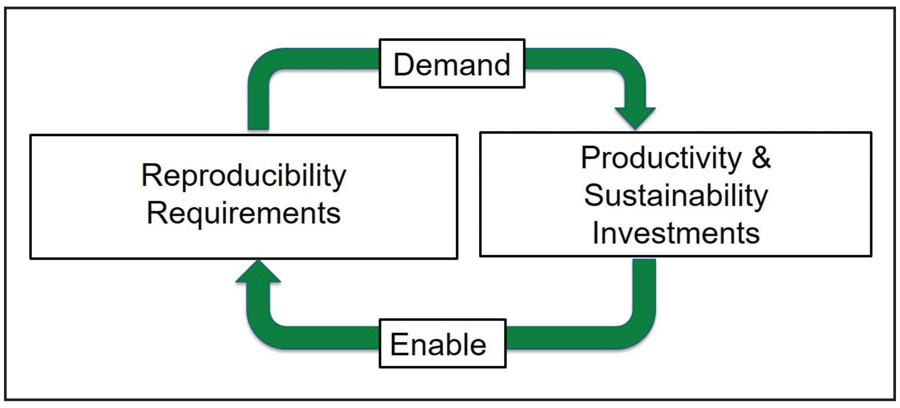

Figure 2. *Adapting our incentive systems to expect improved reproducibility will increase demand for heightened productivity and sustainability, which will in turn enable our desired reproducibility improvements. SIAM can play a central role by introducing reproducibility incentives in its publications and giving recognition to community members who are leaders in reproducible computational science. Figure courtesy of Michael A. Heroux.*

Michael A. Heroux is a senior scientist at Sandia National Laboratories, director of S&W Technologies for the U.S. Department of Energy's Exascale Computing Project, and scientist-in-residence at St. John's University in Minnesota. His research interests include all aspects of scalable scientific and engineering software for new and emerging parallel computing architectures.

Best Paper Research Award at the SIAM International Conference on Data Mining

Hemank Lamba (second from left) and Leman Akoglu (third from left) of Carnegie Mellon University receive the Best Research Paper Award from Conference Program Co-chairs Jennifer Neville (Purdue University) and Xifeng Yan (University of California, Santa Barbara) at the 2019 SIAM International Conference on Data Mining, which took place in Calgary, Alberta, Canada in early May. Their paper and presentation were titled "Learning On-the-Job to Re-Rank Anomalies from Top-1 Feedback." Photo courtesy of Tanya Berger-Wolf.

Food Distribution

Continued from page 4

Bypassing the Bottlenecks

The results indicate that perfect equity means some food banks consistently receive less food than their capacity allows. What if distributors were more equity-flexible? In their current and previous work, Sengul Orgut et al. show that small deviations from perfect equity yield large increases in the total quantity of distributed food [2]. Using the robust optimization algorithm for equity deviation (E-RA), the team explores how distribution is affected when fewer counties are allowed to deviate from perfect equity, but can do so by a larger margin. The E-RA assumes that capacity is deterministic and known, but deviation from equity is a random variable. This nontraditional use of the robust optimization method permits large deviations at the more granular county level but maintains overall equity within the larger system.

The researchers introduce parameter a_j to define the proportion of total food shipped to any single county, equal to $\frac{X_j}{\Sigma_{l=1}^{n} X_l}$. The value of a_j can vary over the range of equity requirements, so that

$$a_j \in [\frac{D_j}{\triangle} - K, \frac{D_j}{\triangle} + K]\; j \in J.$$

Counties with low CD ratios are at high risk of becoming bottlenecks. The model allows underserving of such counties, allocating the remaining food to those with greater capacity. This solution works in actuality because the FBCENC enables its distributors to ship excess food to whomever can take it, sacrificing equity in lieu of food waste. The team's model again provides a practical algorithm for real-world use (see Figure 3).

Sengul Orgut et al. use historical data to show that the model allows considerable increase in total distribution in other counties, and that distribution deviates from perfect equity only by about one percent in bottleneck counties. Unlike many academic models, this work provides food banks with a practical tool for real decision-making. "In some of the other work we do, you must estimate probability distributions and things like that to solve a model," Sengul Orgut said. "But in this case, it's very simple to explain. You're just looking at ranges."

This work and other related papers by Sengul Orgut and her colleagues [1-2] highlight the often-overlooked fact that capacity is just as important as supply when demand is high and decision-makers seek equity. In the future, they plan to include stochastic supply to address the problem of supply scarcity.

The aforementioned algorithms are useful for any system with high demand and uncertain capacity. In addition to implementation in food banks across the U.S., potential applications include natural disaster response, wartime supplies, and allocation of funds for growth and supply storage infrastructure within similar organizations.

References

[1] Sengul Orgut, I., Ivy, J.S., Uzsoy, R., & Hale, C. (2018). Robust optimization approaches for the equitable and effective distribution of donated food. *Euro. J. Operation. Res., 269*(2), 516-531.

[2] Sengul Orgut, I., Ivy, J.S., Uzsoy, R., & Wilson, J.R. (2015). Modeling for the equitable and effective distribution of donated food under capacity constraints. *IIE Trans., 48*(3), 252-266.

Jenny Morber holds a B.S. and Ph.D. in materials science and engineering from the Georgia Institute of Technology. She is a professional freelance science writer and journalist based out of the Pacific Northwest.

Robust Optimization Algorithm for Equity Deviation

```
Sort $J$ in an increasing order of $\frac{C_j}{D_j}$ ratios
$\omega_j \leftarrow 0, \forall j \in J$
while $i \le |J|$ do
    for $1 \le l \le i$ do
        $\omega_l \leftarrow \frac{C_{J(i+1)} D_{J(l)} - C_{J(l)} D_{J(i+1)}}{K\Delta C_{J(i+1)}}$
    end for
    if $\sum_{l \in J} \omega_l > \frac{\Gamma_K}{2}$
        $R \leftarrow \frac{\sum_{l=1}^{i} C_{J(l)}}{\sum_{l=1}^{i} D_{J(l)} - \frac{\Gamma_K}{2} K\Delta}$
        if $R > \min_{l \in J} \left\{\frac{C_l}{D_l - K\Delta}\right\}$
            $R \leftarrow \min_{l \in J} \left\{\frac{C_l}{D_l - K\Delta}\right\}$
        end if
        $\sum_{l=1}^{n} X_l^{*} \leftarrow \min\{S, R\Delta\}$
        Quit
    end if
end while
```

Figure 3. *The robust optimization algorithm for equity deviation (E-RA) allows food banks to relax equity requirements for some bottleneck counties in order to achieve greater food distribution for the overall service area. Real-world scenarios that utilize this algorithm show that even very small deviations from perfect equity generate large increases in total food distribution for these counties. Figure courtesy of [1].*

The Virtual Ecologist

By Easton White and Alan Hastings

Isle Royale is perhaps one of nature's best natural experiments. The small island is located in Lake Superior, about 25 kilometers from the Minnesotan mainland. Over 100 years ago, moose began to inhabit the island; wolves followed in the late 1940s.[1] Wolves are the only predators of the moose and hunt them almost exclusively. This dynamic established an idealized predator-prey system that researchers have studied for the last six decades.

Suppose we want to better understand the role of climate in this predator-prey system. To do so, a biologist would first create hundreds of exact replica Isle Royales, each with different climatic conditions over several decades of experiment (due to the long generation times of the relevant species). This would allow us to examine past climatic effects as well as future impact on the predator-prey relationship. Of course, such an experiment would be ridiculous and impossible. This is what makes ecology difficult. Data is sparse, experiments can be challenging (if not unfeasible), and our actions affect the systems under study. We can, however, perform these experiments in silico; in other words, on our computers.

Mathematical models have a long history in the field of ecology. Italian mathematician Leonardo of Pisa (commonly known as Fibonacci) developed perhaps the earliest model of population growth around the year 1200. He envisioned a scenario where a pair of newborn rabbits were allowed to grow and reproduce. Assuming rabbits can reproduce at one month of age, one rabbit pair would exist at the end of the first month, two rabbit pairs at the end of the second month, three rabbit pairs after the third month, and so forth. This leads to the Fibonacci sequence.

Ecology has come a long way since Fibonacci's time. In 1838, Pierre-François Verhulst applied the logistic function to model populations whose growth depended on population size. In the early 1900s, Alfred Lotka and Vito Volterra independently developed differential equations to study both competition and predator-prey dynamics. The advancements continue to this day. Researchers have gained a better understanding of simple population models and built more detailed and specific models. These improvements occur because of new questions and novel tools, including increased computational power. Several authors have examined the role of models in ecology. Richard Levins emphasized the trichotomy between generality, realism, and precision in models [6]. Evelyn Pielou categorized models into different types—including models for forecasting and models for understanding—based on their applications [10]. We build on these papers with a discussion and examples of the main uses of models in practice today.

Seeking Understanding

Perhaps the most common purpose of a model is to better understand how the natural world works. One can do so by reducing nature to extremely simple models—like the Lotka-Volterra differential equations to represent competition—or building more detailed models that are comparable with real systems.

Brett Melbourne and Alan Hastings utilized a series of stochastic, discrete-time models to better understand the growth and spread of *Tribolium spp.*—an invasive pest better known as the flour beetle—in a laboratory setting [9]. These models varied in complexity, ranging from those with only demographic stochasticity to those with environmental stochasticity and biases in the sex ratio. Melbourne and Hastings compared their models to actual experimental data and determined that the most complicated, detailed models were required to explain spatial spread.

Elisa Benincà and her collaborators also built models to study nature, examining a rocky intertidal community that cycled between bare rock, barnacles and algae, and mussels [1]. The team sought to determine the drivers of this cyclic behavior, but only had a single replicate. Therefore, they built a model of differential equations and included temperature as a seasonal forcing term. Benincà et al. discovered that sustaining the cyclic oscillation required seasonal changes in temperature. This discovery would have been impossible without a detailed model where one could vary seasonal changes — an impractical field experiment.

Moose began inhabiting Lake Superior's Isle Royale as early as the 1900s; now they are an iconic part of the ecosystem. Image courtesy of the National Park Service.

Making Predictions

One of the most desired benefits of models is undoubtedly their ability to make predictions about the natural world. For instance, one might wish to forecast the number of fish that will populate a fishery in the coming year. In the age of big data, researchers can combine streams of data to predict future fish populations. Yet this process only works with a lot of available data and—most importantly—a similarity between the future and the past. As another example, scientists have used models to calculate future species distributions under the influence of climate change. Such models rely on past occurrence data—correlated with environmental conditions—to extrapolate into the future. They do not often include details on dispersal limitations or species interactions. In a changing world, process-based (or mechanistic) models are necessary for prediction. These models can be difficult to build because they require an understanding of the system's underlying processes. However, they integrate uncertainty and provide transparent assumptions, making them appropriate for projection [3].

Wolves have been the primary predator of moose on Isle Royale, a small island in Lake Superior, since 1948. Image courtesy of the National Park Service.

Generating Hypotheses

One of the most fruitful uses of mathematical models has been in the development of new hypotheses about the natural world. Early work by Lotka and Volterra gave rise to the predator-prey equations that bear their names. They found that predators and prey could coexist with cyclic behavior. In 1934, Georgii Gause tested these hypotheses in protozoan predator-prey experiments [4]. He found that typically, predators quickly drive prey to extinction — which was not in line with Lotka and Volterra's predictions. Gause instead discovered conditions (e.g., prey refuges) that enabled coexistence. Decades later, Robert May suggested that complex dynamics (e.g., limit cycles, chaos) could arise from simple discrete-time equations [7-8]. This finding resulted in new experiments and field studies to better understand the incidence of these dynamics in nature [5].

Designing Experiments

Mathematical models can also guide experimental design. In a recent paper, Easton White asked how many years lent confidence to estimation of long-term population trends [11]. To address this question in the context of the moose population on Isle Royale, one would need to construct hundreds of new Isle Royales, place a moose population on each, and sample for different lengths of time. This, of course, is not practical. Creating a mathematical model to build virtual moose populations and estimate the minimum time required for confidence is a better alternative [12]. Researchers have utilized this approach to address a variety of other questions and either aid or substitute for experiments. On a related note, because a model formalizes one's thinking about a system, it can also reveal gaps in his/her understanding of that system, which then provide guidance on the types of data one should collect.

Conclusions and Future Direction

Mathematical models have a long history in ecology and will continue to play a large role in the field. This is especially true of more sophisticted models that have resulted from recent advances in computing. As collected field data becomes sufficient for classification as "big data," models will also become increasingly important in building theory. Additionally, ecological models must also account for rare, black swan-types of events, and come to grips with uncertainty with regard to either parameter estimation or the model structure itself. Approaches like partially observable Markov decision processes have gained traction in dealing with uncertainty [2]. How complicated models should become remains an open question.

References

[1] Benincà, E., Ballantine, B., Ellner, S.P., & Huisman, J. (2015). Species fluctuations sustained by a cyclic succession at the edge of chaos. *Proc. Nat. Acad. Sci., 112*(20), 6389-6394.

[2] Chadès, I., Nicol, S., Rout, T.M., Péron, M., Dujardin, Y., Pichancourt, J.-B., … Hauser, C.E. (2017). Optimization methods to solve adaptive management problems. *Theor. Ecol., 10*(1), 1-10.

[3] Cuddington, K., Fortin, M.-J., Gerber, L.R., Hastings, A., Liebhold, A., O'Connor, M., & Ray, C. (2013). Process-based models are required to manage ecological systems in a changing world. *Ecosphere, 4*(2), 1-12.

[4] Gause, G.F. (1934). *The Struggle for Existence*. Baltimore, MD: Williams and Wilkins Company.

[5] Hastings, A., Hom, C.L., Ellner, S., Turchin, P., & Godfray, H.C.J. (1993). Chaos in Ecology: Is Mother Nature a Strange Attractor? *Ann. Rev. Ecol. Syst., 24*, 1-33.

[6] Levins, R. (1966). The strategy of model building in population biology. *Amer. Sci., 54*(4), 421-431.

[7] May, R.M. (1974). Biological populations with nonoverlapping generations: Stable points, stable cycles, and chaos. *Science, 186*, 645-647.

[8] May, R.M. (1976). Simple mathematical models with very complicated dynamics. *Nature, 261,* 459-467.

[9] Melbourne, B.A., & Hastings, A. (2009). Highly variable spread rates in replicated biological invasions: fundamental limits to predictability. *Science, 325*, 1536-1539.

[10] Pielou, E.C. (1981). The usefulness of ecological models: a stock-taking. *Quart. Rev. Bio., 56*(1), 17-31.

[11] White, E.R. (2019). Minimum time required to detect population trends: the need for long-term monitoring programs. *BioSci., 69*(1), 40-46.

[12] Zurell, D., Berger, U., Cabral, J.S., Jeltsch, F., Meynard, C.N., Münkemüller, T.,…Grimm, V. (2010). The virtual ecologist approach: Simulating data and observers. *Oikos, 119*(4), 622-635.

Easton White is a research associate in the Department of Biology at the University of Vermont. He is a quantitative ecologist who works on questions related to spatial ecology, fisheries, conservation biology, invasive species, and science education. Alan Hastings is a distinguished professor in the Department of Environmental Science and Policy at the University of California, Davis. He is the editor-in-chief of Theoretical Ecology and works on spatial ecology, invasive species, conservation biology, and structured population models.

---

[1] https://isleroyalewolf.org

Data Scientists

Continued from page 6

that technical expertise is more important because the domain is learnable. Since it is unrealistic for data scientists to be experts in all five areas, business teams typically rank the importance of these skills when hiring.

My duties as a data scientist have differed from team to team, so data scientists should expect to accommodate the expertise of their business teams. On teams where I was the sole data science expert, I was simultaneously responsible for handling all of a problem's technical components (including software engineering) and learning the domain well enough to effectively communicate with the domain experts. In contrast, as a data scientist on a team of technical experts (data scientists, back-end developers, front-end developers, etc.), I have focused more on a problem's technical components and less on the organizational factors. I have learned a lot from each role, but the emphasis on what I was required to learn (e.g., domain expertise, software engineering, or mathematics) has differed widely across positions. I am currently a data scientist for the finance organization at 3M. My job is to develop prediction and classification algorithms for the finance department, assist in the operationalization of these algorithms, and educate finance colleagues about data science techniques.

In conclusion, I would advise aspiring data scientists to think about which components of a data science business problem interest them the most. Are you drawn solely to the mathematics of the algorithm, exclusively to the technical elements, or to a mix of technical and organizational factors? I personally enjoy employing a broad array of mathematical and computational skills to solve problems, identifying the use of mathematics in specific domain applications, and teaching mathematics as I communicate results. This comprehensive blend of hard and soft skills is a good fit for me. What is a good fit for you?

Catherine (Katy) Micek is a data scientist at 3M. She holds a Ph.D. in applied mathematics from the University of Minnesota.