REAL BIG DATA

Data scientists can help us make smarter decisions about some of society's most vexing problems

If we have to be hospitalized, most of us hope to leave healthier than when we entered – or at least well on the road to recovery. But not everyone does.

Some patients head home and find themselves getting sicker instead of better, requiring even more treatment. Some even end up back in the hospital. What's the difference between patients who make a full recovery and patients who don't? One major factor is what doctors and nurses call "frailty" — a constellation of factors that include age, nutrition, psychological health, social supports and more.

When UNCG Assistant Professor of Nursing Deborah Lekan did her dissertation on frailty several years ago, she did a painstaking analysis of information drawn from electronic health records, which had just begun to change how nurses and other providers cared for patients.

"I basically had PDF copies of nursing documentation and physician notes," she says.

Getting the information was time consuming and limited by how many records she could analyze herself.

But now, she and collaborators at UNCG and in Greensboro's Cone Health System are harnessing the power of computers, sophisticated statistical techniques, and machine learning to dive much deeper.

The goal: Identify patients at risk of not fully recovering, in real time, and improve the care nurses and others provide for them.

"The tool we're working on would allow clinicians to strategically target interventions based on the risk factors present for an individual," Lekan says.

Her collaborators include Cone Health's Director of Nursing Research Marjorie Jenkins, as well as statistician Thomas McCoy and computer scientists Somya Mohanty and Prashanti Manda at UNCG.

Using cutting-edge data science techniques, the team is analyzing reams of information from patients — everything from physician notes to lab results and medications.

The idea is to develop machine learning techniques that can uncover patterns and insights that human practitioners might miss on their own.

BIG DATA

It's part of a burgeoning field of research in "big data" that allows scientists to scoop up the vast amounts of data increasingly available in our digital world, discover new insights, and even train computers to make predictions – such as which patients aren't ready to leave the hospital.

The work often requires cross-disciplinary collaborations among computer scientists, statisticians, and specialists in academic fields as diverse as public health, government finance, and social justice. The techniques are being used to better diagnose disease, direct disaster response efforts, understand how social media drives news coverage and public opinion, and much more.

"I can train an algorithm to tell me certain inferences about data that are not possible when we look at it through human eyes," says Assistant Professor Somya Mohanty. "You need machine learning to help you out with that."

Dr. Mohanty has applied machine learning methods to a diverse set of problems. In cybersecurity research funded by the U.S. Department of Energy, he worked to use machine learning to predict attacks on computer networks. He's also part of an effort – funded first by the National Oceanic and Atmospheric Administration and now a UNCG Giant Steps Research Development Grant – to track tweets during and after hurricanes, to help identify areas of focus for disaster recovery efforts.

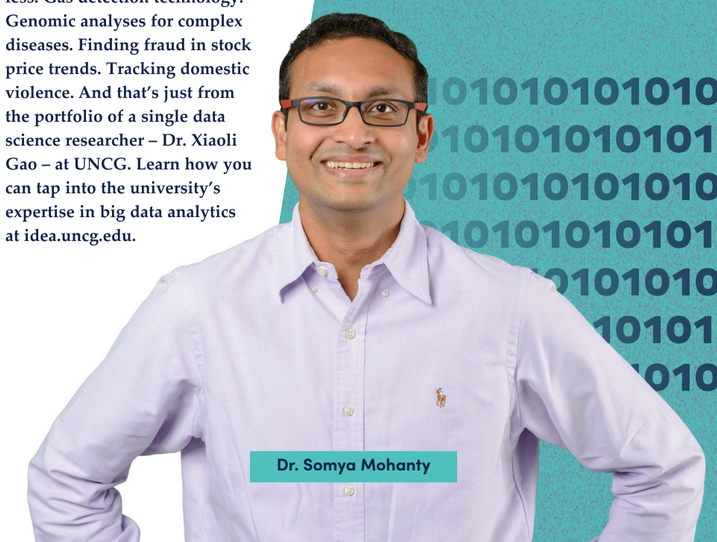
You can't do this work on a desktop. At UNCG, researchers harness powerful computers on campus, access cloud computing capabilities, or use UNC System supercomputers. The files can be so huge that only part of a given set of data can be processed at any time – there simply isn't enough memory.

Much of that work is facilitated through UNCG's Institute for Data, Evaluation and Analytics, or IDEA. The institute connects data scientists and researchers from other fields – economics, education, sociology, nursing, kinesiology, geography, and more – to collaborate on data-focused projects. It also helps connect scientists with individuals and organizations outside the university that need help with data-focused research and data-informed decision making.

Big data applications are limitless: Gas detection technology. Genomic analyses for complex diseases. Finding fraud in stock price trends. Tracking domestic violence. And that's just from the portfolio of a single data science researcher – Dr. Xiaoli Gao – at UNCG. Learn how you can tap into the university's expertise in big data analytics at idea.uncg.edu.

WANT TO LEARN DATA SCIENCE?
UNCG is launching a master's degree with a concentration in data analytics and informatics. The interdisciplinary program gives students grounding in the math, statistics, and computer science needed to tackle a wide range of data-focused challenges, in areas ranging from business and social media to health care and genetics.

UNCG students, under the supervision of Mohanty, are working with local government agencies to help them develop software to analyze spending patterns, identify anomalous expenditures, and predict future spending. Another project focuses on bettering access to often messy courthouse data.



Dr. Somya Mohanty

Dr. Prashanti Manda

Manda and Mohanty are mining more than a century's worth of academic journal papers across several fields to better understand what factors make some papers so influential, while so many others are published and then virtually disappear from collective academic memory. The work is supported by a Microsoft grant and includes access to the software giant's Azure cloud computing platform.

TEACHING COMPUTERS TO READ

UNCG researchers are also using big data to turn an inquisitive lens on academia itself.

Assistant Professor Prashanti Manda, in collaboration with Mohanty, is teaching computers how to read complex academic journal articles. The goal is to make research more accessible, cut down on duplicative research, and help scientists expand their knowledge faster.

In the last few decades, academics have published increasing numbers of papers on all sorts of topics. That means that while there's more knowledge out there, it's also harder to sort through the volume of information available.

A scientist focused on honeybee genetics, for example, might be interested in a very narrow topic, such as the role of a specific gene or a certain protein. But finding the relevant research can be very time consuming.

Papers are sometimes marked up in ways that make searching them electronically easier. In those cases, a person can go through and tag parts of the paper with specific terms, say "protein" or "disease." While that makes the papers more searchable, it is tedious and time consuming.

"With how many papers get published each day, each year, nobody can keep up," Manda says. "I want to develop text mining and natural language processing methods that can automatically 'read papers' and do this annotation themselves."

While the results of that process won't be perfect, it will produce annotated papers that people can double-check for accuracy much faster than they can do markups from scratch.

With support from a UNCG Giant Steps Research Development Grant, Manda is working with Dr. Olav Rueppell, a UNCG biologist and honeybee expert, to make the large volume of bee research more accessible to scientists.

TACKLING COMMUNITY CHALLENGES

Researchers at UNCG are using data science to deepen their impact on some of our region's most vexing issues, in areas such as housing, opioid abuse, and even government budgets.

Many of these projects are part of the MetroLab initiative, a partnership between the University and Guilford County. (Photo: UNCG researchers meet with Guilford County Budget Director Michael Halford.) Through the initiative, students and professors are developing tools to help local government understand the data they have better, and in the process create better solutions to community problems.

In one of the first MetroLab projects, researchers at UNCG partnered with local law enforcement, emergency medical services, and other first responders in Guilford County to tackle opioid overdoses.

The ambitious GCSTOP program, funded by an N.C. General Assembly allocation, collects and shares data across agencies to allow for rapid, targeted interventions. "In year one, we want to reduce opioid-related deaths in the county by 20 percent," says program leader Chase Holleman.

Researchers are also tracking evictions in the county and what happens to those evicted. The data help guide the decisions of the pilot Eviction Diversion Program, a collaboration among UNCG, the Greensboro Housing Coalitions, the 18th Judicial District, and Guilford County.

"We're uncovering inequalities and opportunities for investment," says Center for Housing and Community Studies Director Stephen Sills. "It helps us understand what we can do differently."

As a MetroLab participant, UNCG belongs to a network of 44 cities, five counties, and 59 research universities. The research-driven, town-gown partnerships serve as testbeds for urban innovation.

TRENDING QUESTIONS

Across campus, Dr. Aaron Beveridge in the Department of English is using home-brewed software to better understand how social media influences the media and popular opinion.

The project sparked a few years ago, while Beveridge was in graduate school, when he watched "Tonight Show" host Jimmy Fallon make a claim without any discernible evidence.

"He said something like, 'We just caused this to trend worldwide,'" Beveridge recalls. The assistant professor, who focuses on digital rhetoric, was skeptical. "Is that true? Can I question that? Because to say that something trends worldwide would be such a cultural phenomenon and so powerful – to make that claim without giving data is an unfair thing to do."

The notion that a particular idea is "trending" – in recent years, based on lists of trending items that show up in social media networks and online platforms – is used for more than just late-night laughs. Journalists cite the idea that something is trending in news stories to explain why it might be important. Politicians tout trends as social proof to buttress their positions.

Beveridge wanted to know what it really means when we say something is trending and what that might tell us – or not tell us – about how widespread an idea is.

But there were some roadblocks.

The sources of online trend data, social networks like Twitter and search engines like Google, produce far more information than a single professor like Beveridge – or even a whole building of professors – could realistically expect to digest just by clicking refresh in a web browser. Social media companies sell access to their aggregated data, but it's expensive.

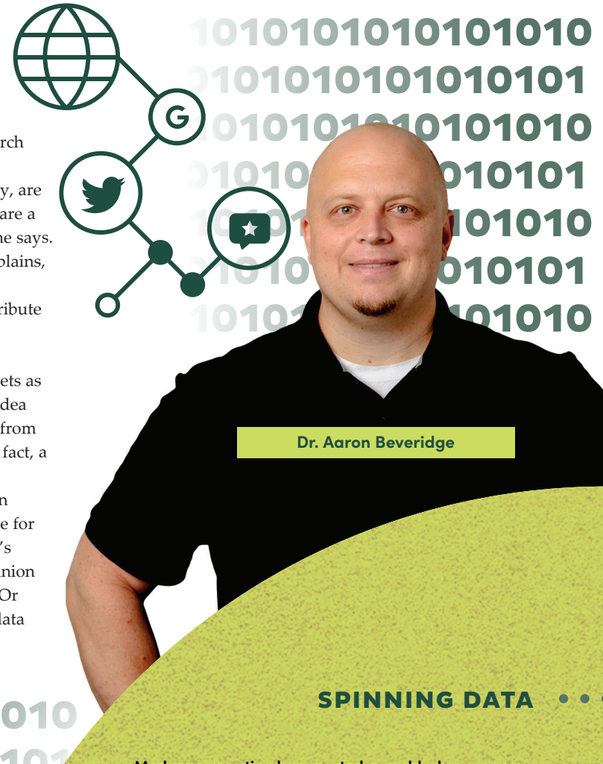
Beveridge, though, turned this problem into an opportunity. Working with a technically savvy collaborator, he helped produce free software that can be used to vacuum up data from Twitter, Google, Tumblr, and other sites. The software, MassMine, was developed with funding from the National Endowment for the Humanities.

It's been invaluable to his research since graduate school.

"What I'm interested in, broadly, are the ways in which social networks are a macroscopic form of persuasion," he says. Seeing something as a trend, he explains, creates authority behind it. Calling something a trend could even contribute to it trending.

Beveridge is now conducting a study on news articles that use tweets as evidence that a particular topic or idea is "trending." He's compiling data from Twitter to see if those topics are, in fact, a major part of online conversations.

"We're going to show how often tweets become the implied evidence for a trend or a public opinion when it's actually not a significant public opinion from a data standpoint," he says. "Or maybe it is. We're going to let the data speak."



Dr. Aaron Beveridge

Dr. Shan Suthaharan

SPINNING DATA • • • • • INTO GOLD

Modern computing has created a world where reams of data are being collected by millions of devices on a seemingly endless array of topics – consumer purchasing patterns, electronic patient records, electricity consumption, and more.

Machine learning techniques allow computers to plow through large datasets and discover hidden connections, yielding insights that allow aspects of business, society, or human behavior to be improved. "It's the kind of knowledge that can cause a noticeable impact on something," says Shan Suthaharan, a computer science professor and author of Springer's 2016 "Machine Learning Models and Big Data Classification."

But right now, those techniques are typically designed to tackle specific problems – whether it's mining health information or power records. And that limits how fast society can find the digital gold buried in all the data that's being collected.

Suthaharan is focused on developing intelligent machine learning models that function like master keys – a concept he calls "transformative knowledge discovery." "You want to make the machine smarter," Suthaharan says. "And you want the model to work across multiple disciplines and domains."

In his quest to develop models that can be applied to many different datasets and address wildly different questions, Suthaharan is working on a wide array of projects – from detecting retinal diseases using images of eyes to classifying fruits and vegetables according to nutrition content. The common thread is his interest in figuring out how to train computers to intelligently deal with very different kinds of data and still find meaningful patterns.

Suthaharan is also working on a second obstacle to widespread use of big data applications: privacy. As big data techniques uncover hidden connections in complex datasets, they can sometimes uncover hidden identities.

There are two kinds of information that can be private, he says. One is categorical information, such as someone's age or whether they have a certain disease. But the other type is numerical pattern information – patterns of data that can unintentionally reveal things someone does not want known.

Imagine, for example, data that a power company collects on household electricity use. Patterns in power usage could reveal when someone isn't home, making their homes more vulnerable to burglary.

"Privacy may be compromised," Suthaharan says.

How do you protect privacy when the whole point of big data is to ferret out actionable insights and make predictions based on hidden patterns?

Suthaharan is working on smarter machine learning techniques that take privacy into consideration. You have to teach the computer to weigh security concerns, he says, while searching for useful insights. "It's about optimization."

There's a trade-off with this approach – the more accurate a model is, the less able it is to protect privacy. His solution to that tricky question? Flexible mechanisms that allow individual data owners to decide how much they want to share.

SPEAKING FOR PATIENTS

In Lekan and Jenkins' patient frailty project, the goal is to let data speak for patients in a way that patients themselves and their providers might not always be able to.

The project pulls together clinical measurements, notes from nurses and doctors, demographic information, and other data – and sets machine learning algorithms loose on them to see if they can predict which patients will need additional care to properly recover.

"One of the benefits of our models is they actually tell us how important different features or variables are in making this decision," Mohanty says. "We will be able to tell them 'OK, this is a particular key variable which makes this a person who is at high risk of readmission.'"

The project started about two years ago and is still in the early stages. Just navigating the legal and technical issues involved in using large amounts of real patient data has taken some time. This summer, analysis was done on a subset of the entire data, and now the researchers are adding more records.

Patient readmission, especially among older patients, is a critical issue. Because of changes to how Medicare pays hospitals, they can face penalties if patients leave and then must be readmitted within a month.

But the frailty project will break ground in other ways, too. One of the first research outcomes will be a kind of roadmap, illustrating how UNCG and Cone Health collaborated to tap into the massive amounts of data in Cone's electronic health record system. It will provide a guide to the legal and technical issues that researchers and hospitals face in using such data.

Electronic health record systems have been widely adopted in the last 10 years, and the software Cone uses is one of the most commonly used systems, so the research could have applications at thousands of hospitals.

Lekan's vision is an app on a hospital's electronic health records that could improve care. "It could red-alert the nurse," she says. "It could prioritize treatments and flag the care specialists we want to loop in. Timely interventions are critical."

By Mark Tosczak • Photography by Jiyoung Park, contributing photography by Martin W. Kane • Learn more at idea.uncg.edu | compsci.uncg.edu | english.uncg.edu | nursing.uncg.edu

Dr. Marjorie Jenkins & Dr. Deborah Lekan