INSIDE THE NUMBERS

From Broadway to biomedicine and the DC Beltway, Big Data is helping Northwestern experts unlock the secrets of creativity and connection. The findings could change everything

With America's presidential election season in full swing, voters nationwide are making choices among exuberant candidates. Their decisions are informed by debates that can make or break a voter's trust. Determined to pinpoint the qualities that make a candidate's rhetoric effective and instill confidence in voters — think John F. Kennedy's telegenic charisma or Abraham Lincoln's reputation for honesty — data scientists and linguists at Northwestern analyzed presidential speeches from 1976 to 2012 to find patterns that translated into spikes at the polls. They discovered that a little-known linguistic technique could strengthen a candidate's credibility time after time.

Called linguistic style matching, the technique involves mimicking an opponent's speech patterns — specifically words with no intrinsic meaning of their own, such as first- or third-person pronouns and prepositions. For example, candidates who begin a rebuttal with the word "we" instead of "I," after their opponent has just used "we" to make a point, will seem more credible than those who choose a different construction.

The study's lead investigator, **Brian Uzzi,** leadership and organizational change, had examined this linguistic tactic in employee-employer negotiations and found it proved beneficial. To parse the effectiveness of the technique in 36 years of presidential speech data, he needed both data scientists and social scientists on his team.

"Computational scientists are especially adept at finding interesting relationships in data, but they also care about making predictions that have external validity — that go beyond the data set in front of them," Uzzi says. "That's where social scientists come in."

A Formula for Creativity

An emerging cross-disciplinary field, computational social science aims to gain insights about human behavior using data science. It encourages fascinating research collaborations, including some that use advanced software to allow scientists to see real-time correlations between tiny movements in facial muscles and human emotions. The applications can be life-altering: Scientists can now sift through millions of electronic health records to characterize various types of a disease and customize treatments to increase the likelihood of a patient's recovery. Collaborators range from the Department of Physics and Astronomy (see page 24) to management and medicine.

For example, Uzzi worked with **Luis Amaral,** a computational scientist in chemical and biological engineering, to determine what factors spur creativity in individuals and teams. To understand the effectiveness of complex teams, they analyzed two types of projects — scientific research and Broadway musicals — that require robust teams with members of varying expertise.

The researchers analyzed the size of the teams, the ratio of senior members to newcomers, and the tendency of senior members to repeat collaborations, gathering data from more than 30 scientific journals and 107 years of Broadway shows. In both disciplines, Uzzi and Amaral found that teams with a higher number of senior members, with seniority closely linked to expertise, produced more successful outcomes. But if those senior members continued to work exclusively with each other, they were less successful than teams that built new collaborations.

Identifying the factors that spur successful collaborations was key to Uzzi and Amaral, codirectors of the Northwestern Institute on Complex Systems (NICO), the University's hub for cross-disciplinary research in systems spanning science, technology, and human behavior. NICO collaborations have contributed to the understanding of various complex networks, including neurons, the US power grid, and social structure.

Understanding that novelty was both a hallmark of successful collaborations and an essential feature of creativity, Uzzi set out to define it using analytical tools. Dismantling misconceptions that novelty is spontaneous and independent of convention, Uzzi and his team found that novelty is actually rooted in established ideas.

Luis Amaral, chemical and biological engineering

Photo by C. Jason Brown

"We used to think of novelty and conventionality as existing on the same spectrum — as you moved away from one, you moved toward the other," Uzzi says. "What we found is that innovation is 90 percent conventional knowledge and 10 percent novelty."

Using familiar ideas as a foundation for new expressions tends to have a greater impact on audiences. Uzzi uses Darwin's famous *On the Origin of Species,* which introduced readers to his theory of evolution, as an example.

"Darwin spends the first 700 pages citing conventional knowledge about animal breeding, an idea that was very familiar to his readership. It's only in the last 70 pages that he introduces his novel idea," Uzzi says.

To develop their formula, the research team looked for novelty among a set of 28 million scientific papers, using paper citations as clues to the amount of conventional knowledge that informed a new idea.

Taking a similar approach, Amaral and another team examined both scientific papers and US movies to determine what made certain examples of each significant. The concept of significance, or how people recognize the creative work of others, was measured by citations in scientific papers and by references to visual choices on screen.

"The way the director sets up a scene or focuses an image can be an homage to films that came before," says Amaral, an expert in complex systems.

Amaral's team mined the Internet Movie Database to gather data on reviews, awards, public opinion, and box office returns for 15,000 films. They ranked the movies with the most citations and cross-referenced titles with the National Film Registry of the Library of Congress, the team's benchmark for cinematic significance. *The Wizard of Oz* and *The Godfather* ranked among the most-referenced movies in cinematic history.

As with Uzzi's Broadway musical study, this research suggested the importance of blending novelty with a proven foundation to stimulate creativity that resonates with audiences.

Data for a Deeper Diagnosis

Like the movie website Amaral's team used to sift through films, databases containing patient health records continue

Photo by Callie Lipkin

Brian Uzzi, management and organizations

to grow. Every day, healthcare providers choose treatment courses that ideally are customized to a person's ailment and likely reaction, and the data on the patient's diagnosis and response are logged in electronic health records (EHR). This offers new ways to study millions of diseases and their treatments, contributing to the rise of precision medicine, an approach that uses computational science and genomics.

"Looking at gene sequences alone won't reveal much if you don't know the impact of those sequences. That's where EHRs come in," says **Justin Starren,** chief of the Feinberg Division of Health and Biomedical Informatics in the Department of Preventive Medicine. "EHRs give you the ability to ask clinical questions among hundreds of millions of cases."

Starren is a co-investigator of the Electronic Medical Records and Genomics (eMERGE) Network, a consortium of sites with EHRs and genomics biobanks. Northwestern is a founding member of eMERGE, which pools records for analysis at scale. Running such large studies has already provided researchers with insights on human diseases. For example, scientists now know that many diseases, such as diabetes and autism, are typically not caused by a few common gene variants but by thousands of different rare variants. Only by observing how each variant behaves can scientists understand and treat the diseases.

"The idea behind precision medicine is that once we gather enough of this data, we can start conducting molecular-based treatment," says Starren, who also directs the Center for Data Science and Informatics at the Feinberg School of Medicine. "For many diseases, like cancer, picking the right treatment requires knowing what is happening at the molecular level within the cells."

Posttreatment data combined with genetics can also help keep beneficial drugs on the market, he says. The clinical trials conducted before a drug is approved involve only a small number of people, as compared to the number who will later use the approved drug. Because of this, side effects are sometimes discovered after approval. EHR data monitoring may detect such effects early, and with enough data, scientists can learn whether a side effect is related to a rare gene variant. Otherwise, drugs that can benefit many people may be pulled off the market to prevent severe side effects in a few.

Despite such potential, expanded use of EHRs faces the ongoing challenge of current privacy regulations that often automatically inhibit the use of patient records. Projects such as the National Institutes of Health's Precision Medicine Initiative are trying to make this easier, Starren says, but there is a long way to go.

"There is no technology that can absolutely guarantee that patient information will stay private, but we at least need to allow patients to make the risk-benefit trade-off for themselves rather than make it for them," he says.

Information Overload

Data production is increasing at a startling speed. According to Google CEO Eric Schmidt, humanity produces in just two days an amount of data greater than the total sum created prior to 2003. As this data tsunami becomes more difficult to process, research collaborations between computational scientists and social scientists at Northwestern are proving that a connected world allows for knowledge and creativity breakthroughs that were impossible a decade ago.

This summer, the Feinberg School will host a Biomedical Big Data Day, and the Kellogg School will host the Second Annual International Conference on Computational Social Science, bringing together leaders in the field.

Uzzi says the Kellogg conference will highlight the emergence of a new methodology — a computational one — that is powerfully augmenting the traditional ways in which science is done. — *Monika Wnuk*

Landmark Discovery Makes Waves

Vicky Kalogera, physics and astronomy, has played a prominent role in proving Albert Einstein right.

Vicky Kalogera

The director of Northwestern's Center for Interdisciplinary Exploration and Research in Astrophysics (CIERA) is a longtime member of a global team that, earlier this year, announced the first direct evidence of gravitational waves, a key prediction in Einstein's theory of relativity. With these data, the team also made the first direct observation of two black holes colliding. They did so using twin, 2.5-mile-long L-shaped antennas — the Laser Interferometer Gravitational-Wave Observatory, or LIGO — based in Louisiana and Washington State. The tools, and hundreds of researchers, formed the LIGO Scientific Collaboration, an effort that harnessed diverse expertise to interpret the unique data discovered.

NRM spoke with Kalogera about the team's search for signals in the data and the role of computational science in astrophysics.

How have big data and computational science impacted astronomy?

Astronomy is a very data-driven science. My early interaction with data was small scale and focused on theoretical simulation to explain observations in astronomy. I joined LIGO in 2000, working on pure astrophysics, and it wasn't until about eight years ago that I became involved with major challenges in data science. In LIGO, we were searching for very faint signals in many terabytes of noisy data. As astrophysicists, we didn't have all the tools we needed to detect those signals, so we collaborated with experts in applied math and statistics. Doing so, we were able to efficiently and robustly tease out the signals from deep within the noise.

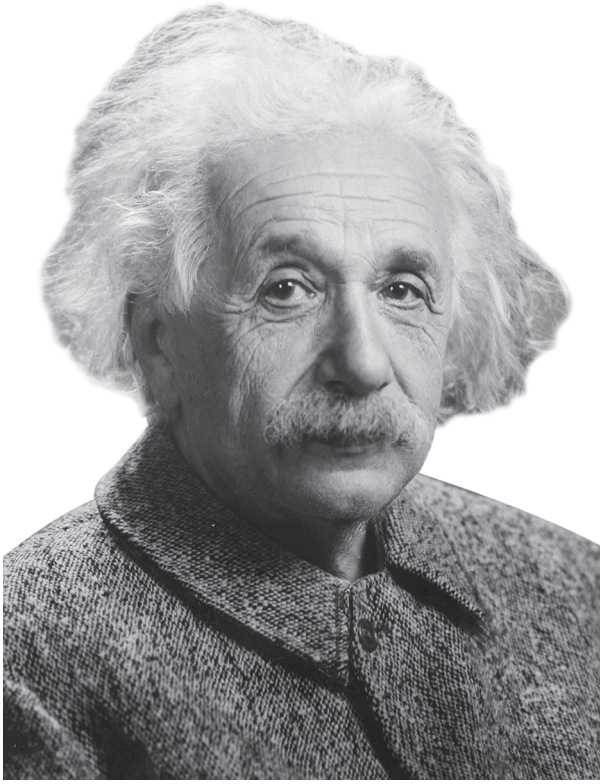
We can't see black holes through telescopes. Was there another way you examined the data received from the LIGO detectors?

In astronomy, we are used to analyzing faint features in gorgeous images, but what we got from the LIGO detectors were long data time sequences that were not especially pleasing to the eyes. What's fascinating, however, is that this type of detector mimics how our ears work. Our sight is limited to the direction in which we point our eyes, but our ears pick up sound from all around us. Similarly, you don't point your gravitational wave detector in one direction; instead, it picks up signals from all directions and converts them into sound waves with frequencies that humans can hear and differentiate.

How might computational science advance astronomical discovery in the future?

We're working with the National Science Foundation to build a telescope capable of surveying the whole sky every three days for 10 years. We'll be taking the longest movie ever of the universe. If you wanted to play all of the data collected in sequence as a movie, that "film" would run for more than a year. The data collected will be 100 times bigger than anything we're used to in astronomy. We will rely heavily on knowledge and methods from computational science to develop algorithms that return the interesting discoveries we anticipate.

—Monika Wnuk



THE GRAVITY OF IT ALL

For the first time, scientists have observed ripples in the fabric of spacetime called gravitational waves, arriving at the Earth from a cataclysmic event in the distant universe. The discovery confirms a major prediction of Albert Einstein's 1915 general theory of relativity and opens an unprecedented new window onto the cosmos. Northwestern's **Vicky Kalogera** and **Shane Larson,** both physics and astronomy, and **Selim Shahriar,** electrical engineering and computer science, were part of the team making this historic discovery.

1 Billion
It was more than 1 billion years ago, that the two black holes collided with one another

30x
One black hole was estimated to be 30 times as massive as the sun and the other 29 times larger

.004
The energy caused by the gravitational wave vibrated the LIGO instruments just four one-thousandths of a diameter of a proton.

1,000+
The number of scientists and engineers at 80 institutions in 15 countries that comprised the LIGO Scientific Collaboration

1915
Albert Einstein lays out his general theory of relativity, revolutionizing the way we understand gravity.

1916
Einstein predicts the existence of gravitational waves, while adding that he doubts anyone will ever be able to detect them.

1922
"Gravitational waves move at the speed of thought," says Sir Arthur Stanley Eddington, a renowned English astronomer and physicist, expressing skepticism for the supposed ripples in spacetime.

1957
Felix Pirani, a young postdoc at the University of North Carolina, explains for the first time how scientists might detect gravitational waves.

1960s
Joseph Weber builds the first "resonant bar detectors."

1971
Robert Forward builds the first interferometer for gravitational wave detection. It's a tabletop experiment.

1972
Rainer Weiss publishes the first serious analysis of the experimental challenges of gravitational wave detection with interferometers.

1999
Construction of LIGO's original gravitational wave detectors is completed.

2010
LIGO is redesigned to see 10 times farther out into the Universe.

September 14, 2015
A century after Einstein's revolutionary general theory of relativity, the LIGO collaboration makes the first detection of a gravitational wave.