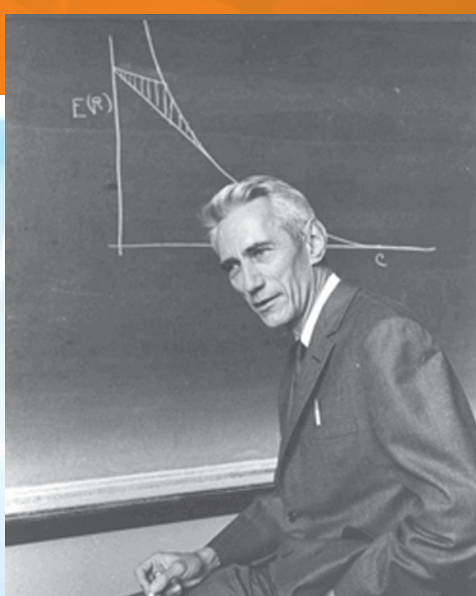
TOWARD A NEW SCIENCE OF INFORMATION

It's often said we live in the information age. With our mobile phones, iPods, iPads, DVDs, wireless networks, and the Internet, we are awash in information. What people may not realize is that the foundations for transmission of information in today's technologies were laid back in the 1940s by Claude Shannon in his seminal paper on "A Mathematical Theory of Information."



Claude Shannon

> **We need to focus more on how information is managed in living systems and how this brings about higher-level biological phenomena.... We need to describe the molecular interactions and biochemical transformations that take place in living organisms, and then translate these descriptions into the logic circuits that reveal how information is managed. This analysis should not be confined to the flow of information from gene to protein, but should also be applied to all functions operating in cells and organisms, including chemical interactions and transformations as well as physical phenomena, such as electrical signaling and mechanical processes.**
>
> — PAUL NURSE, "Life, logic and information," *Nature 2008*

Published in 1948, Shannon's work established fundamental limits on the compression, transmission, and storage of information in simple communication systems—and we've been riding the Shannon "wave" ever since. These are the principles that govern, for example, how many songs you can fit in a 16 Gigabyte iPod and how well your streaming video works over your WiFi connection.

At the time, Shannon deliberately made many simplifying assumptions, acknowledging that in the future, other work would be needed to account for aspects not encompassed within the scope of the problem he defined. He ignored issues of time-critical transmission of data, multi-point communication, spatially structured information, and message content and context. Nevertheless, it was enough to help spur a trillion-dollar communications industry that has had profound personal, societal, and economic effects. Enough to make one ask: What wonders might be in store for us next?

"Shannon basically gave us fundamental modes for storage and communication point to point," says Wojciech Szpankowski, director of the Center for Science of Information (CSoI). "In today's applications, there isn't just point to point communication; everything is dynamic, from biological networks, to the Internet, to wireless. And this changes the situation completely."

Researchers at CSoI are working to advance the science of information to transcend the limits of Shannon's theory. Headquartered at Purdue University, the center brings together researchers at nine partner institutions from computer science, information theory, and applications areas such as natural sciences, economics, and social sciences to advance the field.

"We are part of the Shannon legacy," says Szpankowski. "The goal is to extend Shannon theory beyond its original objective to incorporate temporal, spatial, structural, semantic, and contextual features in order to better understand the flow of information in today's applications: in biology, modern communications, economics, and others."

In the mathematical definition of information that Shannon introduced, the main objective was the statistical reduction of uncertainty. In a broader sense, however, information is something that has to be related to the objectives of the recipient.

An example of time-critical information is an airline passenger waiting for a flight, explains CSoI associate director Ananth Grama of Purdue. "You need to find out what time your flight is leaving, and if you get a text message after the flight has already left, saying that you're flight is going to be on time, that information is worthless to you," he notes. Center researchers are working on ways to incorporate the "value" of information with respect to time, "overcoming the simplifying assumptions that Shannon imposed–time doesn't matter, delay doesn't matter--but we know very well that if the message arrives very late it might be completely useless," emphasizes Szpankowski.

Another of Shannon's simplifying assumptions was to ignore the content of messages. "If you study the traffic flow of trucks on the highway, you may not care what they carry. Whether they have bananas or jet fuel or whatever–in other words, the semantics doesn't matter. But in other applications, it does matter very much," says Szpankowski.

Other problems involve data structures that are not easily tractable in classical information theory. "We've gotten to a point where information is a lot more sophisticated that just bags of zeros and ones," says Grama. "Data may have structure—geometry, as in chemical structures, DNA, proteins. In biology, we're talking about very large networks of cellular interactions. In social sciences, we're talking about social networks. In economics we're talking about transactions. Data are higher order–it's not just simply bits anymore."

The mission of the Center for Science of Information is to advance science and technology through a new quantitative understanding of the representation, communication, and processing of information in biolog¬ical, physical, social, and engineered systems. In the process, the center aims to prepare a new generation of scientists equipped to develop and apply emerging theories.

This area of work is expected to have many applications for industry and government organizations. The center currently is interacting with a range of companies, large and small, in the retail, software, aerospace, and biotech sectors, in addition to federal agencies, says Grama, who serves as director of knowledge transfer for CSoI. "Each has a large amount of data, and they are trying to find effective and rigorous ways of extracting actionable information from their data sets." □

A CONVERSATION WITH THE DIRECTOR

Wojciech Szpankowski

"I see myself as an enabler," says Wojciech Szpankowski, director of the Center for Science of Information and the Saul Rosen Professor of Computer Science at Purdue University.

"The team we have is incredible–the best researchers in information sciences in the world. This is a one-hundred member center with about fifty faculty, fifty students, and some collaborators. So it is huge. The people are top-notch. Of course you can't tell them to do anything," he laughs. "The only thing you can do is to motivate them, and when they become enthusiastic about something, they will do wonders.

"So that is why we are using a process of extended, one-day workshops, held at many regional sites. We sit for six to seven hours and we ask questions, and people pick up problems and begin to work on them. We need to establish natural collaborations–this is the hardest part because the center is big. And they are indeed being established."

EDUCATIONAL PROGRAMS AT CSOI

A new undergraduate course, titled "Introduction to the Science of Information," has been developed by CSoI to provide a new synthesis of this emerging discipline and to provide undergraduate students a foundation to pursue higher studies at Center institutions.

The course was launched during fall 2011 at Purdue University by CSoI associate director Mark Daniel Ward through the University Honors Program. A related course is scheduled to be given at Bryn Mawr College in spring or fall of 2012.

On another front, the Center's inaugural summer school brought students and faculty from ten universities together during summer 2011. Students gained a broad understanding of the emerging field of science of information while establishing connections with one another and the Center's faculty. The four-day intensive included ten surveys and tutorials spanning the breadth of the Center's research thrusts. Students learned new tools and applications in daily laboratory offerings.

"I learned what information really is, as it was a vague concept I took for granted before the summer school. Specifically, I learned about how information science should be studied from a multi-disciplinary approach," says one participant. "I also learned about graduate schools and different programs/departments that might be of interest to me in the future from talking to professors and students." Lectures with video/audio/slides and laboratory exercises are available at http://soihub.org/summer-school.php

In Their Own Words...

EDUCATION AND OUTREACH: TAKING THE SCIENCE OF INFORMATION FROM THE CLASSROOM TO THE COMMUNITY

Deepak Kumar

CSoI Associate Director for Education & Diversity
Professor of Computer Science, Bryn Mawr College

"Computing, for me, is very intimately related with the study of information. Even though the idea of information transcends computing and computation, I think computing is what makes it more interesting and makes it possible to apply the science of information.

Models of information or underlying principles of information ultimately come to life in computers, and so my goal, what's exciting to me, is I can take these ideas and introduce them to students in many different disciplines.

Here's a question we can actually ask and answer today. Were all the works we associate with Shakespeare today actually written by him? We can apply computational techniques and mathematical analyses to the texts, to the entire body of work that's claimed to be Shakespeare's, and we can analyze it within a matter of, if not minutes, then hours, and come up with metrics of evaluation that tell you yes, they're all in the same style.

Another example: we now have available the texts of all of the U.S. president's states of the union speeches over the years. Let's analyze it by various metrics. When did the word woman actually appear in those speeches? When did issues of race come up? How many times? We can plot the frequencies, and when you do that you actually see when cultural and social awareness of some of these issues came about.

There's this whole new field that's coming up now called digital humanities. Information theorists would be interested in the efficiency of language–how many words, how long a text needs to be to express the same set of ideas in different languages. We can look at translations of the same text and come up with some metrics about how efficient with respect to the size of the text a particular language might be. And actually Shannon sort of dropped some of these ideas way back when, because according to Shannon's theory, redundancy in language is actually a good thing, because there is loss of data over transmission, and so that's the key to error correction.

English, it turns out, is only about 50 percent efficient. You can see that when you text now–you don't send text messages with complete words and phrases. There are a whole bunch of these issues we didn't even know how to get a handle on, that we can now.

The science of information is one of these areas that brings many disciplines together. And so the idea of creating a new undergraduate course on science of information is founded on those ideas: how can we engage students who are in a college where everything is by department, right, and you have to take x courses in this discipline and y in that, and so forth–How can we break those boundaries?

Science of information is a natural place–it combines computing, it combines information, and it brings all these other disciplines together. And so the new course we're creating has those goals. How can we engage students so that no matter what they're studying, they'll bring a lot more back into their own studies than they would have otherwise. So when an English major comes and takes a course like that, goes back to the English department, what is she bringing back to that? Some of the examples I've described are in that realm: suddenly the paradigm of computing, the way of thinking about information is relevant to their own disciplines in a new way. That's exciting to me.

As somebody responsible for education and diversity, I'm actually looking well beyond information in terms of outreach. For instance, in fall 2010, there was a movie that came out called *Top Secret Rosies,* and it's about women mathematicians at the University of Pennsylvania during World War II. Instead of Rosie the Riveter, during the war, these were their mathematical cousins. Nobody knew about their work because it was all classified. The filmmaker making the documentary interviewed these women, who were still alive.

It's about computing history, it's about women in computing history, it's about an important piece of that, so as part of the center, we actually rented the local theater in town, made it free to the public, invited the filmmaker, screened the movie, and had her do a question and answer session. We had 300 members of the public, and when the question and answers started, there were actually a few people who stood up and said they were one of these top secret Rosies, or their grandmother was one. One of the things we're trying to do is make sure that we organize and host events like these. We did this in March at Bryn Mawr Theater in Pennsylvania, and partners in Purdue also held a similar event."