# The Sequence of Things to Come: An Interview with Wojciech Szpankowski about the Future of Information Science

by Luke Redington

Computers baffle me. For questions large and small—from how to rid my screen of yet another pop-up ad to whether the advent of Heartbleed means I might as well tweet all my financial data—I must ask others. When it comes to people who should be asked about computers, Wojciech Szpankowski is very near the top of the list. Szpankowki is Saul Rosen Professor of Computer Science at Purdue University and director of the Center for Science of Information (CSoI), an NSF organization which brings together leading researchers from around the world. In April 2014, I had the opportunity to interview him. We made something of an odd couple: A paper-and-pen loving graduate student in English and a renowned mathematical theorist whose work with a wide variety of algorithms has made him a leader in information science. But it worked. I focused my questions on the cultural impacts of developments in computer science. Szpankowski's answers illuminated the ways which the work at CSoI is futuristic, far-reaching, and yet already woven into the fabric of every day life.

Unsure at first how the odd couple arrangement would work, I bring as a conversation starter a *New York Times* op ed piece about public perceptions of big data by Gary Marcus and Ernest Davis. Szpankowski had read it already, and liked it, but he was bursting with enthusiasm to discuss a different op ed piece which had recently appeared in *The Financial Times* called "Big Data: Are We Making a Big Mistake?" by Tim Harford. I'm not about to try to curb Szpankowski's enthusiasm, even if it meant jettisoning the start of my interview plan. So, I ask, "What do you like about the editorial?"

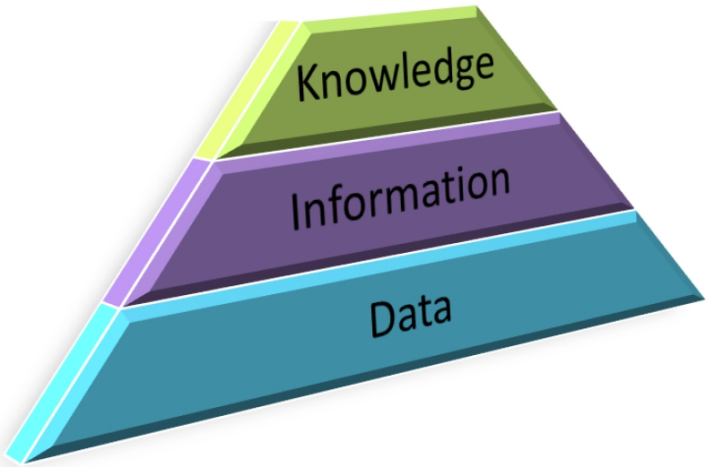
*Szpankowski in his office at Purdue University*

*Photo by Vincent Walter*

"It has a great quote: 'Big data has arrived, but big insight has not.'" For Szpankowski, Harford's proverb encapsulates a central challenge facing information scientists. He explains this challenge in terms of a habit which has for many people become as

vital and subconscious as breathing: Asking Google a question. “When we do a query on Google,” Szpankowski continues, “We want answers. But what is Google actually doing? It is searching through data. Data contains information, surely, and data and information are relatively easy to define. But what we want when we seek answers is knowledge, and this is harder to define.”

As a student of language, I think most people—myself included—would have roughly equal levels of difficulty defining data, information, and knowledge. However, Szpankowski is referencing a key development in the history of information science—the moment in 1948 when the terms data and information were given clear definitions by Claude Shannon, the founder of the field. Shannon was then working at Bell Laboratories, using his background as an electrical engineer to improve the technologies behind a device in tremendous popular demand—the telephone. The main improvement that interested Shannon in 1948 was to figure out how to keep phone conversations understandable as they were occurring over ever longer stretches of cable. In crafting a theory that quantified this loss of comprehension, Shannon coined several key terms and framed relationships between them, a feat which still inspires (and beguiles) information scientists. In his 2011 book *The Philosophy of Information,* Luciano Floridi explains the losses and gains resulting from Shannon's decision to put data, information, and knowledge in a mathematical, hierarchical arrangement. Information, Floridi contends, has a long history in Western thought as a rich but messy term. Various influential thinkers have used information to mean things as lofty as enlightenment to as pedestrian as the accurate conveyance of a verbal communique by a messenger-on-foot from its author to its audience (80-1). Floridi notes that Shannon clearly favored the latter view of information, a disposition still prevalent in information theory today, summed up in the notion that “information consists of data” (82). What is gained through this view is that information can be made quantifiable, as Shannon did in his landmark 1948 paper “A Mathematical Theory of Communication.” Once quantifiable, information can be monitored as it crosses distances and changes formats. Its integrity, after undergoing these changes, can be measured. What is lost through this view of information is a comprehensive means of clarifying the upward relation information has to knowledge.




*Graphic by Mike Atwell and Evan Smalley*

Szpankowski's initial example about Google encapsulates both the difficulty and the importance of placing conceptual parameters around “knowledge.” It's what we're after when we pose a question to a search engine, or a stock broker, or a magic eight ball. And, it's what Harford says has not arrived.

Not *yet* arrived, Szpankowski counters. As he thinks about what lies ahead for CSoI, he's optimistic. “For the next five years, CSoI is focusing on how to analyze the movement from data

to information to knowledge." Szpankowski's own work addresses several challenges in analyzing this movement, specifically, how to respond to the astronomical size of the databases that must be queried in any hopes to find information, much less knowledge. "Imagine you have a big bag of data," he says, gesturing widely at an empty space on his office floor. I imagine the data as marbles of various bright colors, but I do not interrupt him to vet the validity of my poetic license. "You want to know, with a high degree of certainty, aspects of the entire bag of data. But, you have only a handful of data" he says while extending a cupped, upturned hand, which I fill with more marbles.

Believing I see where this is going, I jump in with one of the few things about statistics I know to ask: "Is the sample representative?"

He grins and peers at me over his glasses. "That is the question."

"Can information science help answer this question?"

"Yes. We are working on describing how 'sketchy' the information can be and still provide an answer that is right with a high degree of probability."

Another challenge stemming from the vastness of today's databases is data compression, also a major focus of Szpankowski's work. "Nowadays, data is huge, multi-modal, and structured." The huge part, I get. Szpankowski helps me with the "multi-modal" and "structured" parts by talking about social networks. "Novels are made up of just one kind of data—words on a page." (I want to interject, "Yeah, but it's the *best* kind of data." Somehow, I refrain.) That's not how it is with data in social networks. It's from all kinds of media. Plus, a network is structured information." He shows me a graph. It looks like a square, white paper target that has been shot with a thousand black ink pellets by someone with better than average aim. The black dots, he explains, represent users of a particular social network, and their placement on the page expresses the nature of their relationship with other users. Because the spatial relationship reflects the social nature of a relationship, that spatial-structural component of the information becomes the really useful part. It becomes the knowledge sought by marketers, law enforcement agencies, actuaries, or public health officials.

So, in these many cases where the structure is what matters, Szpankowski is tackling the challenge of how to compress the data while leaving its structure in tact. As an illustration of this challenge, consider this map. According the staff of The Mariner's Museum of Newport News, Virginia, it was published in a Dutch atlas in 1694 (The Mariner's Museum). At first glance, the map struck me as surprisingly accurate given its date. Then, I noticed the teensy-weensy error that has made this map an historical curiosity: California is an island.


*Source: http://ageofex.marinersmuseum.org*

As if that weren't sufficiently curious, consider that the California-as-island myth was well over 150 years old by the time this map was published, and that accurate information about California's attachment to the mainland was already available in Europe. Evidently the pro-island faction had circulated their view more widely (The Mariner's Museum). One need not be a cartographer to understand what went wrong here: Too much hubris, not enough data. Spanish conquistador Hernán Cortéz, founder of the pro-island movement, trusted his explorer instincts so fully that he did not feel the need to traverse the entire length of the Baja Peninsula before deciding that neither it nor any of the land stretching north of it ever touched the mainland (The Mariner's Museum). Many others followed his lead, including more than a few cartographers. Honest cartographers, when they knew they were drawing beyond the data, wrote "Here be Monsters" or "*Terra Incognita."* We cannot, in fairness, judge the honesty of the cartographer behind this 1694 map; we only know that the result seamlessly blends good map making with bad data.

Today, the challenge is exactly the opposite: Too much data. And because most data information scientists work with is collected through automated processes, very little of it is likely to be "bad" data in the mathematical sense. Computers misbehave, but they don't miscount. And, more to the point, the data information scientists work with is intricately structured because it is arises from the complexity of real relational networks, real purchasing habits, and real satellite images of the earth's surface. So today, information scientists are like cartographers tasked with putting a full-sized map on a postage stamp. They constantly wrestle with questions such as which data is most relevant for which end-users and how to achieve data-compression while minimizing data-distortion. Considering this, Szpankowski sighs and leans back in his chair, saying, "Shannon calculated how much data could fit through a channel of a given size without losing too much data. He never thought about structures."

Szpankowski thinks about structures a lot these days. His recent work suggests that the new frontier of dealing with big structures is to think small. Structured data is built of patterns, and Szpankowski is currently working on new ways to identify patterns, even if they are small and scattered across huge data sets. For an example of why pattern detection matters, think about the way marketers make use of patterns. If you shill (insert egregious dollar amount here) for a latte each Thursday morning, this habit shows up in databases as a pattern. If you pay for your (insert pretentious, unpronounceable name of your beverage here) with a credit card, the dams burst and you cause a "data flood," to use a phrase made popular by Jame's Gleick's best-selling book *The Information: A History, A Theory, A Flood.* Instantly, your credit card company and all those "third parties" named in the privacy agreement open their reservoirs and take in the data. These "third parties" are basically people who want to sell you stuff, or at least advertisers working for companies who want to sell you stuff. They make highly educated guesses about what they should try to get you to buy based on their ability to navigate floods of data and identify patterns. Have you been skipping the whipped cream the last few weeks? A diet supplement ad is on its way. Have you recently purchased a pair of running shoes? A message goes out to all sporting goods stores within a thirty mile radius to send you coupons for stay-dry socks. If there is a recent pattern of rainy mornings in your area, expect additional ads for light-weight, water-proof jackets. So far, these are small patterns—a drop in the information ocean. But since detecting them involves looking at a multitude of data points over an extended period

of time, the process still requires immense number crunching capability. Advertisers, though, want to operate on an entirely different scale of magnitude. They want to track as many patterns as possible for as many people as possible and to crunch the numbers as quickly as possible.

Szpankowski's recent work can help pattern-tracking endeavors by capitalizing on a paradox: If computers are given a larger, more sophisticated definition of "pattern," they can more effectively identify small features of patterns in huge databases. He explains: "What if you are looking for a pattern of three elements—a, b, and c. Well, I've begun asking, 'What if you don't care whether those elements occur in the main part of the pattern or in a sub-string of the pattern?'" He can tell from the look on my face that any drama or irony tied up in these "what if's" is utterly lost on me, so he bails me out, continuing, "Looking for patterns in sub-strings is extremely important in detecting cyber-intrusions."

Now, I nod vigorously. I've been reading recently about Heartbleed and scrambling to change my passwords.

I've gleaned that Szpankowski's work has potential impacts in economics and cyber security. But before I can flesh out these huge topics, he tosses another into the mix. "Now, I'm working with protein networks, looking at the ways enzymes interact to form proteins." Here, too, the ability to compress patterns is like possessing the Rosetta Stone. "We don't have a machine which can read an entire DNA sample at once. You have to split it into smaller copies, but then you get many overlapping copies. Let's say, represented as data, the DNA sample needing to be sequenced is 10 to the ninth power.[1] So, we cut the data in to "reads" of 100 to 300 data bits. But then, you have millions of these reads. What to do? You have to find a better way to locate patterns in all this data."

We are now late in the interview and have covered enough ground that I feel ready to test my comprehension. "So, let me see if this describes the challenge: If you identify a sub-string that's too short, you don't compress the data very much. If you mis-identify a sub-string, you lose data in your compression."

"Yeah. Yeah."

"Does this have applications in biology? In oncology, maybe?"

"I don't know. Too much of a stretch for me." His broad grin and the impish twinkle in his eye tell me that by "too much of a stretch," he simply means he's never looked into it. "Several other CSoI members are pursuing those sorts of applications, like Ananth Grama here at Purdue and David Tse at Stanford."

I can't believe he wants to skip over the applications. That's the juicy part, the payoff. It's the thing that keeps the science world spinning. Then, I pause and think perhaps my cultural assumptions about science might be surfacing. Having anticipated this might happen, I journaled about these assumptions as part my interview preparation. That journaling process produced the following question, which despite its lengthy wind-up, seemed worth voicing in full as a means of capping off the interview: "We Americans tend to be pragmatically minded people. I can see this as I think back to the science classes I took. My education has always been heavily tilted toward the humanities, so these were just basic, required science classes. Anyway, as I think back on them, it seems Ben Franklin was given as much attention and held in at least as high regard as Isaac Newton. I think it's because the things he invented could be put to immediate, widespread use—the lightning rod, better indoor heating, bifocal glasses. The more I read about the history

1 More commonly known as one billion.

of science, though, the more I begin to think that all discovery is useful. It's just that some discoveries are useful immediately, and others are useful after more time has passed. So, when non-scientists think about the value of the work being done at CSoI, would you encourage them to think mostly about the inherent value of discovery rather than the pursuit of projects with immediate applications? If so, what examples come to—"

"Einstein."

"Pardon?"

"Einstein. All Einstein's experiments were thought experiments. His elevator, etc. He did not have any applications in mind. I'm from Poland, but I've been here thirty years. One thing America gave me is a competitive nature. That's good—if you don't push too hard. And, it gave me a push to do something new. I believe that science works best when you bring together the right people at the right time to ask the right questions."

And this is CSoI's operative philosophy. It is a place where people like Szpankowski can, in his own words, "Keep asking the big questions." And yet it is also a place that fosters special collaborations between researchers primarily pursuing theory and researchers who keep an eye toward application, as is the case with fellow CSoI member Ananth Grama, the colleague to whom Szpankowski alluded. It is a place, in fact, that has succeeded in transcending place. As a consortium, it overcomes spatial boundaries. CSoI members frequently collaborate, even when they hail from different universities all over the world. As a center—rather than an academic department—CSoI reaches into disciplines and across them in innovative ways. I could not help but feel it is a place where the future is in particularly close reach.

*Luke Redington is a staff writer for CSoI.*



Works Cited

Floridi, Luciano. *The Philosophy of Information.* Oxford: Oxford UP, 2011. Print.

Gleick, James. *The Information: A History, a Theory, a Flood.* New York: Pantheon, 2010. Print.

Harford, Tim. "Big Data: Are we making a big mistake?" *The Financial Times. (FT.com)*

28 March 2014. Web.

Marcus, Gary and Ernest Davis. "Eight (No, Nine!) Problems With Big Data." *New York Times.*

The Opinion Pages. 6 April 2014. Web.

Mariner's Museum, The. “Exploration Through the Ages: The Southern Continent, 1694.” 29 July 2014. Web.

Shannon, Claude. “A Mathematical Theory of Communication.” *The Bell System Technical Journal,* Vol. 27 (July/Oct 1948). 379–423, 623–656. Print.