Introduction

The Stanford Data Science Initiative is all about high-impact data science. This is data science that has a large-scale impact on society. High-impact data science requires cutting-edge research, awareness of important opportunities, and the means to have impact at scale.

**Cutting-edge research.** SDSI researchers are developing the algorithms and analytic techniques that unlock subtle relationships, improve computational efficiency (sometimes by several orders of magnitude!), and make sophisticated techniques much easier to use.

**Awareness of important opportunities.** Our deep engagement with the corporate members of SDSI provides extraordinary visibility on the most important problems, opportunities, and real-world constraints. Our corporate members inform, influence, and inspire our research agenda!

**The means to have impact at scale.** Large-scale impact requires real-world datasets, powerful computers, and the resources to apply inventions to change the world. By partnering with industry, government, and nonprofits, Stanford’s innovations are leading to a better society, more successful companies, and improved scholarship.

The following exemplary applications serve to make these points. The first describes the use of DeepDive to fight human sex trafficking on the web. The second illustrates how computer vision and machine learning are providing precise mapping of worldwide poverty, thus enabling more effective humanitarian actions. The third example is about automated approaches to cancer diagnosis that outperform human pathologists. The example on text-based crisis counseling describes the analysis of 660,000 text messages from 15,000 counseling sessions to improve effectiveness, while the example on human resources in corporations uses natural language processing on emails to predict employee retention. The example on hospital-bred infections describes a data-driven study using 500,000 Medicare patient records. The gender bias example illustrates the presence of bias in machine learning…and what to do about it. In a preview from an upcoming article, new work on data science theory sheds light on unobserved mutations in genomics. The article from Science describes the use of networks to describe various complex systems.
TABLE OF CONTENTS

Memex Human Trafficking ................................................................. 1
Stanford Researchers Use Dark of Night and Machine Learning to Shed Light on Global Poverty.. 3
Computers Trounce Pathologists in Predicting Lung Cancer Type, Severity ......................... 6
Stanford Research Could Improve Counseling on Crisis Help Lines ..................................... 9
Which Will Get You Further: Fitting In or Standing Out? ...................................................... 11
How Hospitals Avoid Penalties for Making Patients Sick ..................................................... 14
Removing Gender Bias from Algorithms ................................................................................. 17
Unseen Distributions in Genomics: Recovering the Structure of the Unseen Portion of a Distribution ............................................................................................................................ 20
Higher-Order Organization of Complex Networks ..................................................................... 21
Human trafficking is a serious humanitarian problem. Traffickers use physical or financial force to coerce victims into providing labor. Such labor is often advertised online.

These online advertisements contain crucial information about workers: services offered, location, pay rates, and other attributes. However, the advertisements are often textual, like Craigslist ads. They contain useful data but cannot be queried using SQL databases or statistical tools such as R.

DeepDive is a tool for structured data extraction that allows us to transform the raw set of advertisements into a single clean structured database table. Analysts can now use standard tools to exploit this information that was previously inaccessible, and thereby identify and help human trafficking victims.

Problem Description

Human trafficking is the use of physical, financial, cultural, or other methods to coerce labor from individuals. Human trafficking is not only a serious humanitarian problem; it is also a serious challenge to law enforcement and military institutions who must grapple with the organizations who engage in human trafficking. If we could use data-centric tools to better identify human trafficking victims and understand the activities of traffickers, we could help people in need as well as help governmental institutions trying to address an ugly and difficult problem.

Dataset Availability

Many trafficking victims are channeled into forced factory labor or sex work. The trafficker benefits by taking some part of the victim’s pay for engaging in this work. Since human trafficking is substantially an economic phenomenon, it is believed that traffickers advertise their victims’ services, disguised as a standard commercial offering. Of course, like advertising in general, many of these trafficking-driven advertisements are likely to be on the Web.

If we could analyze online advertisements for factory or sex workers, we might be able to help human trafficking victims in several important ways:

- Identifying which advertisements are likely to describe trafficked individuals, as opposed to non-trafficked factory or sex workers
- Using advertisement-embedded data to shed light on the habits and internal structure of traffickers or trafficking organizations
- Helping law enforcement prioritize their scarce resources to focus on helping individuals who are in the greatest danger, or on traffickers who are the most dangerous

Advertisement data is surprisingly informative. Imagine a large set of text advertisements for sex work. These advertisements --- a mixture of structured HTML fields and unstructured text --- are similar in form to advertisements for more conventional goods. An advertisement may contain the name of the worker, contact information if the customer is interested, physical characteristics, and price. Like advertisements on, say, Craigslist, any one data value might be missing. But there are some values that are commonly observed, and which are crucial for a sex worker (whether trafficked or not) interested in making a sale.
The Analytics Opportunity

The set of advertisements for this work is surprisingly large. Our collaborators have obtained more than 30 million advertisements for sex work from online sources. These advertisements contain extremely rich data. In principle, we could do the following with the advertisement data:

- Traffickers may move their victims from place to place in order to keep them more isolated and easier to control. If we detect individuals in the advertisement data who post multiple advertisements from different physical locations, it might be a signal that the individual is being trafficked.
- Non-trafficked sex workers show economic rationality: they charge as much as possible for services, and avoid engaging in risky services. Charging non-market rates or engaging in risky services may be a signal the individual is being trafficked.
- Traffickers may have multiple victims simultaneously. If the contact information for multiple workers across multiple advertisements shows contains consecutive phone numbers, it might suggest one individual purchased several phones at one time. The existence of such a person may be a signal that the individual is being trafficked.

These analytics questions embody information about human trafficking, and to someone new in the area they might sound surprising. But they are entirely straightforward from a data management point of view. Given the right dataset, an analyst or data scientist could easily answer these questions using SQL plus Python, or R, or Stata, or any number of other products.

Unfortunately, in their textual form these advertisements cannot be queried using a standard database or analyzed using standard statistical or business intelligence tools. The only common tool for managing unstructured data is a Google-style search engine, which retrieves a document-at-a-time and cannot enable the business-intelligence-style analytics that we require. This is a shame: the advertisement data is a huge opportunity that conventional data management tools cannot exploit.

That is where DeepDive comes in.

Structured Data Extraction with DeepDive

We need to transform the collection of raw advertisement texts into a structured table suitable for analyst processing. Instead of millions of documents, we want a single table with millions of rows. Crucially, the analyst wants that table to contain the following five columns:

- URL where the advertisement was found
- Phone number of the person in the advertisement
- Name of the person in the advertisement
- Location where the person offers services
- Rates for services offered

If we had such a table in a traditional SQL database, the analyst could write the query that would enable the analytics questions listed above.

DeepDive solves this problem. It is a tool for structured data extraction. It transforms the raw text corpus into a high-quality database that downstream tools can process. After DeepDive has populated the table, the analyst can ignore the fact the data was ever derived from documents in the first place. From the analyst’s perspective, there is simply a new and remarkable dataset to analyze.
Stanford researchers use dark of night and machine learning to shed light on global poverty

An interdisciplinary team of Stanford scientists is identifying global poverty zones by comparing daytime and nighttime satellite images in a novel way.

Stanford researchers use machine learning to compare the nighttime lights in Africa indicative of electricity and economic activity with daytime satellite images of roads, urban areas, bodies of water and farmland. (Image credit: Marshall Burke)

One of the biggest challenges in fighting poverty is the lack of reliable information. In order to aid the poor, agencies need to map the dimensions of distressed areas and identify the absence or presence of infrastructure and services. But in many of the poorest areas of the world such information is rare.

“There are very few data sets telling us what we need to know,” said Marshall Burke, an assistant professor in Stanford’s Department of Earth System Science and an FSE Senior Fellow at the Freeman Spogli Institute. “We have surveys of a limited number of households in some countries, but that’s about it. And conducting new surveys in hard-to-reach corners of the world, such as parts of sub-Saharan Africa, can be extremely time-consuming and expensive.”

A new poverty mapping-technique developed by interdisciplinary researchers at Stanford offers cause for hope. The technique is based on millions of high-resolution satellite images of likely poverty zones. To analyze these images, the researchers used machine learning, a discipline within the broader field of artificial intelligence. In machine learning, scientists provide a computational model with raw data and an objective – but do not directly program the system to solve the problem. Instead, the idea is to design an algorithm that learns how to solve the puzzle by combing through the data without direct human intervention.
The researchers began their poverty-mapping project knowing that nighttime lights provide an excellent proxy for economic activity by revealing the presence of electricity and the creature comforts it represents. That was half of the raw data that their system needed.

“Basically, we provided the machine-learning system with daytime and nighttime satellite imagery and asked it to make predictions on poverty,” said Stefano Ermon, assistant professor of computer science. “The system essentially learned how to solve the problem by comparing those two sets of images.”

Learning to spot poverty

Burke, Ermon and fellow team members David Lobell, an associate professor of Earth system science, undergraduate computer science researcher Michael Xie and electrical engineering PhD student Neal Jean detailed their approach in a paper for the proceedings of the 30th AAAI Conference on Artificial Intelligence.

Their basic technique – directing a model to compare images to predict a specific value – is a variant of machine learning known as transfer learning. Ermon likens this to how the skills for driving a car are transferable to riding a motorcycle. In the case of poverty mapping, the model used daytime imagery to predict the distribution and intensity of nighttime lights – and hence relative prosperity.

It then “transferred” what it learned to the task of predicting poverty. It did this by constructing “filters” associated with different types of infrastructure that are useful in estimating poverty. The system did this time and again, making day-to-night comparisons and predictions and constantly reconciling its machine-devised analytical constructs with details it gleaned from the data.

“As the model learns, it picks up whatever it associates with increasing light in the nighttime images, compares that to daytime images of the same area, correlates its observations with data obtained from known field-surveyed areas and makes a judgment,” Lobell said.

Those judgments were exceptionally accurate. “When we compared our model with predictions made using expensive field-collected data, we found the performance levels were very close,” Ermon said.

Highly effective machine-learning models can be very complex: The model the team developed has more than 50 million tunable, data-learned parameters. So although the researchers know what their mapping model is doing, they don’t know exactly how it is doing it. “To a very real degree we only have an intuitive sense of what it is doing,” said Lobell. “We can’t say with certainty what associations it is making, or precisely why or how it is making them.”

Next generation of surveying

Ultimately, the researchers believe, this model could supplant the expensive and time-consuming ground surveys currently used for poverty mapping.

“This offers an unbelievable opportunity for cheap, scalable and surprisingly accurate measurement of poverty,” Burke said. “And the beauty with developing and working with these huge data sets is that the models should do a better and better job as they accumulate more and more information.”

The availability of information is something of a limiting factor. Right now satellite coverage of impoverished areas is spotty. More imagery, acquired on a more consistent basis, would be needed to
give their system the raw material to take the next step and predict whether locales are inching toward prosperity or getting further bogged down in misery.

But such data restraints could soon be lifted – or at least mitigated.

“There’s a huge number of new high-resolution satellite images that are being taken right now that should be available in the next 18 months,” Burke said. “That should help us predict in time as well as space. Also, there are several micro-sat companies that plan to provide images of the planet almost daily, and we’re rapidly getting enough satellites up to do that. I don’t think it will be too long before we’re able to do cheap, scalable, highly accurate mapping in time as well as space.”

Even as they consider what they might be able to do with more abundant satellite imagery, the Stanford researchers are contemplating what they could do with different raw data – say, mobile phone activity. Mobile phone networks have exploded across the developing world, said Burke, and he can envision ways to apply machine-learning systems to identify a wide variety of prosperity indicators.

“We won’t know until we try,” Lobell said. “The beauty of machine learning in general is that it’s very useful at finding that one thing in a million that works. Machines are quite good at that.”
Computers trounce pathologists in predicting lung cancer type, severity

Automating the analysis of slides of lung cancer tissue samples increases the accuracy of tumor classification and patient prognoses, according to a new study.

Stanford Medicine News Center, August 16, 2016

Assessing a biopsied slice of tissue to determine the grade and severity of a tumor can be highly subjective, but Stanford researchers found that computers could be trained to make accurate assessments of lung cancer tissue. Science Photo/Shutterstock

Computers can be trained to be more accurate than pathologists in assessing slides of lung cancer tissues, according to a new study by researchers at the Stanford University School of Medicine.

The researchers found that a machine-learning approach to identifying critical disease-related features accurately differentiated between two types of lung cancers and predicted patient survival times better than the standard approach of pathologists classifying tumors by grade and stage.

“Pathology as it is practiced now is very subjective,” said Michael Snyder, PhD, professor and chair of genetics. “Two highly skilled pathologists assessing the same slide will agree only about 60 percent of the time. This approach replaces this subjectivity with sophisticated, quantitative measurements that we feel are likely to improve patient outcomes.”

The research was published Aug. 16 in Nature Communications. Snyder, who directs the Stanford Center for Genomics and Personalized Medicine, shares senior authorship of the study with Daniel Rubin, MD, associate professor of radiology and of medicine. Graduate student Kun-Hsing Yu, MD, is the lead author of the study.
Although the current study focused on lung cancer, the researchers believe that a similar approach could be used for many other types of cancer.

“Ultimately this technique will give us insight into the molecular mechanisms of cancer by connecting important pathological features with outcome data,” said Snyder.

Assessing severity of cancer

For decades, pathologists have assessed the severity, or “grade,” of cancer by using a light microscope to examine thin cross-sections of tumor tissue mounted on glass slides. The more abnormal the tumor tissue appeared — in terms of cell size and shape, among other indicators — the higher the grade. A stage is also assigned based on whether and where the cancer has spread throughout the body.

Often a cancer’s grade and stage can be used to predict how the patient will fare. They also can help clinicians decide how, and how aggressively, to treat the disease. This classification system doesn’t always work well for lung cancer, however. In particular, the lung cancer subtypes of adenocarcinoma and squamous cell carcinoma can be difficult to tell apart when examining tissue culture slides. Furthermore, the stage and grade of a patient’s cancer doesn’t always correlate with their prognosis, which can vary widely. Fifty percent of stage-1 adenocarcinoma patients, for example, die within five years of their diagnosis, while about 15 percent survive more than 10 years.

The researchers used 2,186 images from a national database called The Cancer Genome Atlas obtained from patients with either adenocarcinoma or squamous cell carcinoma. The database also contained information about the grade and stage assigned to each cancer and how long each patient lived after diagnosis.

The researchers then used the images to “train” a computer software program to identify many more cancer-specific characteristics than can be detected by the human eye — nearly 10,000 individual traits, versus the several hundred usually assessed by pathologists. These characteristics included not just cell size and shape, but also the shape and texture of the cells’ nuclei and the spatial relations among neighboring tumor cells.

“We began the study without any preconceived ideas, and we let the software determine which characteristics are important,” said Snyder, who is the Stanford W. Ascherman, MD, FACS, Professor in Genetics. “In hindsight, everything makes sense. And the computers can assess even tiny differences across thousands of samples many times more accurately and rapidly than a human.”
Bringing pathology into the 21st century

The researchers homed in on a subset of cellular characteristics identified by the software that could best be used to differentiate tumor cells from the surrounding noncancerous tissue, identify the cancer subtype, and predict how long each patient would survive after diagnosis. They then validated the ability of the software to accurately distinguish short-term survivors from those who lived significantly longer on another dataset of 294 lung cancer patients from the Stanford Tissue Microarray Database.

This brings cancer pathology into the 21st century and has the potential to be an awesome thing for patients and their clinicians.

Identifying previously unknown physical characteristics that can predict cancer severity and survival times is also likely to lead to greater understanding of the molecular processes of cancer initiation and progression. In particular, Snyder anticipates that the machine-learning system described in this study will be able to complement the emerging fields of cancer genomics, transcriptomics and proteomics. Cancer researchers in these fields study the DNA mutations and the gene and protein expression patterns that lead to disease.

“We launched this study because we wanted to begin marrying imaging to our ‘omics’ studies to better understand cancer processes at a molecular level,” Snyder said. “This brings cancer pathology into the 21st century and has the potential to be an awesome thing for patients and their clinicians.”

The work is an example of Stanford Medicine’s focus on precision health, the goal of which is to anticipate and prevent disease in the healthy and precisely diagnose and treat disease in the ill.

Stanford co-authors of the study are former postdoctoral scholar Ce Zhang, PhD; professor of pathology Gerald Berry, MD; professor of bioengineering, of genetics and of medicine Russ Altman, MD, PhD; and assistant professor of computer science Christopher Re, PhD.

The study was supported by the National Cancer Institute and the National Institutes of Health (grants U01CA142555 and 5U24CA160036-05).

Stanford’s Department of Genetics also supported the work.
Stanford News, August 10, 2016

**Stanford research could improve counseling on crisis help lines**

Many people now text rather than call for help, allowing computer scientists to study anonymous data files and learn which counseling tactics work best.

By Marina Krakovsky

Crisis hotlines have been around for years, but until recently there’s been very little data on which counseling strategies seemed most effective at helping people cope. The recent emergence of text-based crisis help lines is changing that.

Designed for people who prefer texting to talking, these services generate large datasets of anonymous counseling sessions, raw material that computer scientists can study to identify the words and techniques that seem to improve outcomes.

As more people text rather than call crisis hotlines, Stanford computer scientists are developing new techniques to discover the counseling strategies that seem to work. (Image credit: TommL / Getty Images)

“Until now, most research on counseling has remained small-scale, looking at voice transcripts of only a few dozen sessions,” said Jure Leskovec, an associate professor of computer science at Stanford.

Working with graduate students Tim Althoff and Kevin Clark, Leskovec analyzed 660,000 text messages from 15,000 crisis counseling sessions. In a recent scholarly paper the researchers identified several techniques associated with successful sessions, such as personalizing exchanges, quickly getting to the root of the problem, and using words and phrases to steer conversations onto a positive track.

Leskovec said he believes such findings could be used to train counselors how to respond most effectively when a person in the midst of a crisis reaches out for help.

“We can look at orders of magnitude more data than previous studies allowed, to gain new insights and precisely quantify which counseling strategies worked,” he said.

**5 phases of conversation**

For this study, the researchers developed new methods of natural language analysis to determine how the words and phrases that counselors used influenced whether distressed texters reported feeling better at the end of the conversation.

In particular, they contrasted the language used by counselors who are very successful at getting texters to report feeling better with the language of those counselors who were generally less successful.
Researchers discovered that all counseling conversations followed five stages: introduction, problem setting, problem exploration, problem solving, and wrap-up.

Each stage can be characterized by the words counselors as well as texters use. For example, the introduction stage was marked by greetings on both sides and the wrap-up stage showed texters expressing appreciation and counselors using words like “any time,” “good night” and “appreciate.”

**Key to success**

These stages were independent of the topic, which could be anything from relationship troubles to thoughts of suicide. But by analyzing and comparing how the most successful and least successful counselors progressed through the stages, the researchers discovered one key difference.

“Successful counselors quickly got to the heart of the issue and spent more of the conversation dealing with the problem,” Althoff said. “The less successful counselors took a lot more time to get to know the problem.”

This finding is related to another interesting pattern: successful counselors tend to respond more effectively to ambiguous messages. Presented with exactly the same situation — a breakup with a boyfriend or girlfriend, for example — a successful counselor typically asks more clarifying questions. They paraphrase responses to make sure they understand, and they thank the texter for reaching out.

In short, successful counselors do more to draw out the terse texter and get to the crux of the person’s problem. As Althoff explained, this means that successful counselors tend to talk more. They personalize their messages to the specific texter and situation so their comments sound natural. The study showed that texters tended to talk more about certain topics once counselors broached those topics. So counselors can put texters in a better frame of mind by making subtle changes to their own language.

“If you talk about the future, I will be more likely to talk about the future,” as Althoff said. “If I talk positively, you’ll be more likely to talk positively.”

This type of analysis can be applied to training crisis counselors in the short run and, as the language analysis techniques improve, perhaps even lead to the development of automated conversational agents that support counselors during training and actual conversations.

“These sorts of applications become possible when we bring the power of natural language analysis and artificial intelligence to bear on extremely large datasets of real crisis conversations,” Leskovec said.
Which Will Get You Further: Fitting In or Standing Out?

The most successful employees do a bit of both, striking a balance between integration and nonconformity.

Stanford Graduate School of Business: September 2, 2016|by Eilene Zimmerman

Is it better to fit in or stand out? That question has vexed all of us at one time or another, from teenagers to aspiring executives to sociologists. The answer, says Stanford Graduate School of Business professor Amir Goldberg: It depends. If you’re the kind of person who stands out culturally — you wear sweater vests and bowties to the office — then to succeed you will need to fit into your organization structurally, by being part of a tight-knit group of colleagues. And if you stand out structurally — you aren’t a member of any one clique at work, but have friends across departments — then you better fit in culturally (so ditch the bowties). Goldberg and his coauthor Sameer Srivastava at the University of California, Berkeley, explain the reasoning behind that answer in their new paper, “Fitting In or Standing Out? The Tradeoffs of Structural and Cultural Embeddedness,” soon to be published in the American Sociological Review. The paper, written in collaboration with Christopher Potts, an associate professor of linguistics at Stanford, and graduate researchers Govind Manian and Will Monroe, explores the relationship between fitting in, standing out, and success within an organization.

The modern corporate world generally celebrates those who stand out from the pack, rewarding them with promotions and salary increases. “No one wants to be perceived as average or replaceable, especially in tech companies that value innovation, diversity, and creativity,” says Goldberg. Yet fitting into a company is also important. It creates a larger, motivating sense of identity for employees and enables them to be productive members of the organization, which by definition depends on cooperation. The result is a conflicting pressure on workers to fit into an organization and, at the same time, stand out. Goldberg and his colleagues wanted to learn more about that tension and find ways to resolve it.

No one wants to be perceived as average or replaceable, especially in tech companies that value innovation, diversity, and creativity.
Amir Goldberg

They took a novel approach, examining the language used in corporate emails. The researchers were given access to a mid-sized technology company’s complete archive
of email messages exchanged among 601 full-time employees between 2009 and 2014. As a way to protect privacy and confidentiality, only emails exchanged among employees were analyzed. Those sent to or received from people outside the company were excluded, as were messages exchanged among the seven members of the executive team and with the firm’s attorneys. The raw data was stored on secure research servers installed at the company. Goldberg, Srivastava, Monroe, and Potts created an algorithm that could analyze the natural language in emails; the most difficult part of the analysis was figuring out what categories to use to organize language and filter out what wasn’t meaningful.

“We wanted to look at language that reflected cultural style and the level of cultural alignment, not language that referred only to a job’s function or the industry,” says Goldberg. Using an established language-analysis program, the researchers looked for specific words such as “I” or “we,” as well as “should” or “would.” They also looked for expletives. In some companies, says Goldberg, swearing is a way to assert authority. “In that culture, imagine if a new person comes in and uses very polite language,” he says. “That would reflect a low level of cultural assimilation in a company where cursing is acceptable.”

They closely examined the structure and pattern of email interactions to learn more about small, tight-knit cliques within the organization. Previous research has shown that bridging the gaps between these cliques benefits employees because of the information that flows to them from different groups. “They are able to connect networks that are usually disconnected, so they serve an important role,” says Goldberg. “They have their tentacles stretched into different worlds, and stand out because of that.”

To learn how this relates to an employee’s success, Goldberg, Srivastava, and their collaborators relied on human resource data supplied by the company. In addition to employee age, gender, and tenure, it identified all employees who had left the company and whether their departure was voluntary or involuntary. That data enabled the researchers to correlate professional success with fitting in and standing out. “Involuntary exits are a sharp sign of negative attainment, because employees are at greatest risk of experiencing that when their performance is weak,” Goldberg says.

The researchers concluded that employees in the firm can be characterized by their levels of cultural and structural “embeddedness,” after measuring their general cultural assimilation as well as the strength of their attachment to various network cliques. This led the researchers to identify four organizational archetypes: “doubly embedded actors,” “diseMBEDded actors,” “assimilated brokers” and “integrated nonconformists.”
Those most likely to get ahead are what they call “assimilated brokers,” meaning people who are high on cultural fit and low on network cliquishness, and their mirror images, the “integrated nonconformists,” meaning people who are part of a tight-knit group but still stand out culturally.

“The assimilated broker is the ultimate networker, the person with friends in marketing, customer service, engineering,” Goldberg says. “She is well-connected across the firm but not really a part of any one group.” Yet she blends in culturally, speaking and dressing the same as everyone else. The integrated nonconformist, on the other hand, has the security and mutual commitment that comes from being part of a clique but has not fully assimilated into the corporate culture. “He doesn’t repeat the same old Star Wars references, but does inject the occasional hip-hop reference into conversation. He wears an ironic, vintage button-down shirt rather than a T-shirt with the company logo.”

What the researchers call a “doubly embedded” employee, meaning someone who is both culturally compliant and part of a dense network, never stands out and isn’t likely to move ahead. Think of the geeky software engineer who is part of his tight little team, but doesn’t interact outside that group. He isn’t aware of what’s happening in the company as a whole and knows no more than his immediate peers. Because of that, a doubly embedded actor brings nothing new to the table. “He is a carbon copy of every other geeky member of his team,” says Goldberg. “He isn’t perceived as unique or indispensable in any way.” The archetype most likely to lose their job is the “disembedded actor,” who interacts with people in different parts of the organization but isn’t part of any one group and doesn’t fit in culturally either. “Think of a cocktail party where this guy tries to network with you but the conversation is really painful,” says Goldberg. “That’s the disembedded actor, and being a bad networker is far worse than not trying to network at all.”

Comparing their email analysis to the company’s personnel records, the researchers found that workers identified as doubly embedded actors were over three times more likely to be involuntarily terminated (i.e., fired) than those identified as integrated nonconformists.

Clearly, both fitting in and standing out are important for career success, but the lesson, says Goldberg, is that if you blend in structurally and culturally — especially in tech firms, which put a premium on creativity and innovation — you will be seen as bland and unremarkable. At the same time, if you dress and speak differently than your peers but you also aren’t part of any one group, you will wind up being perceived as odd and threatening.

The goal is to find a balance between the two. “Either maintain your place as part of a tight-knit group but stand out by behaving a little weirdly, or be the smooth networker who knows what’s going on across the organization but also knows how to blend in culturally,” says Goldberg. “You want to distinguish yourself from the pack without making anyone in the pack uncomfortable.”
How Hospitals Avoid Penalties for Making Patients Sick

Lax reporting requirements make it easier to change records.

March 7, 2016 | by Edmund L. Andrews

Reuters/Lucas Jackson

Listen to This Story
For years now, health care reformers have been pushing hospitals to do what should be obvious: protect patients from hospital-bred infections that make them sicker than they already are.

The U.S. Centers for Disease Control estimates that about 4% of hospital patients — more than 1 million people a year — acquire serious and often life-threatening infections during the course of being treated. One study estimated that hospital-acquired conditions contributed to more than 98,000 deaths in 2002 alone. The CDC estimates that the direct economic cost of those infections is between $28 billion and $34 billion a year.

Starting in 2008, Medicare stopped reimbursing hospitals for hospital-acquired infections. The idea was to hit careless hospitals where it hurt and give them a financial incentive to clean up their act.

Unfortunately, the evidence so far suggests that the penalties have had little impact.

Now, a study led by Mohsen Bayati at Stanford Graduate School of Business offers an explanation: Instead of spurring tighter standards of care, the penalties are prompting hospitals to simply “upcode” their records and claim that patients were already infected at the time they were admitted.

Upcoding — or assigning an inaccurate billing code, usually to increase reimbursement — is difficult to track, except through auditing, because hospital claims for reimbursement don’t include laboratory and nursing reports from a patient’s stay. But the researchers were able to estimate how often hospitals changed their records by comparing data from states that require very strict reporting of hospital-acquired infections with data from states that have weak reporting requirements.

They found that hospitals in weakly regulated states reported substantially lower rates of hospital-acquired infections as well as higher rates of infections that were “present on admission.” Those differences couldn’t be explained by differences in patient demographics or other factors. In other words, hospitals that didn’t have to be as transparent about in-house mishaps were in fact reporting fewer of them.
Using conservative assumptions, the researchers estimate that they found 11,000 cases of upcoded infections per year — about $200 million in unwarranted payments — in a sampling of about 492,000 Medicare patients who had hospital stays between 2009 and 2010.

In itself, $200 million is a drop in the bucket. The real cost, says Bayati, is that the non-payment policy isn’t providing the incentive for hospitals to reduce the huge losses that stem from unnecessary illness and death. Indeed, he argues, the penalties may be creating the false impression that the policy is working and that hospital standards of care are higher than they really are.

That insight runs counter to the conventional wisdom among health care reform advocates. Some advocacy groups have pushed for even tougher financial penalties, and Congress has moved in that direction. A section of President Obama’s 2010 Affordable Care Act authorizes the federal government to reduce Medicare reimbursement rates by 1% for hospitals that rank in the top quartile for hospital-acquired conditions, or HACs. Those regulations became effective in 2014.

But those stronger penalties are based on an assumption that the reported rates are accurate. In fact, however, one implication of the new study is that hospitals with higher reported rates may actually be the better ones — those that are more truthful in their reporting and possibly have higher standards of care.

“Medicare’s current plan to increase penalties through the HAC Reduction Program does not address these concerns, and may in fact exacerbate the problem since hospitals with high HAC rates will face even greater financial pressure to engage in upcoding,” the researchers write.

Bayati and his colleagues — Hamsa Bastani, a PhD candidate at Stanford’s Electrical Engineering Department; and Joel Goh, a former Stanford graduate student who is now a Harvard Business School professor — argue that a more effective solution is to increase the mandatory reporting and expand the use of targeted auditing. Hospitals showed much less evidence of upcoding when states required detailed reporting that included the patient’s identity, the impact of the infection, and its root cause.

The researchers caution that it’s not enough to simply have a system for reporting adverse events, because some reporting systems are so lax that they have almost no impact at all. The key is to require enough information, and auditing, that upcoding becomes impractical.

“The more data a state has regarding the circumstances of an adverse event, the harder it is for a provider to be untruthful,” the researchers argue in their paper. As Supreme Court Justice Louis Brandeis wrote a century ago, “Sunlight is said to be the best of disinfectants.”
Removing gender bias from algorithms

September 25, 2016 9.31pm EDT, The Conversation

Can machine learning help us find – and reduce – gender bias

Machine learning is ubiquitous in our daily lives. Every time we talk to our smartphones, search for images or ask for restaurant recommendations, we are interacting with machine learning algorithms. They take as input large amounts of raw data, like the entire text of an encyclopedia, or the entire archives of a newspaper, and analyze the information to extract patterns that might not be visible to human analysts. But when these large data sets include social bias, the machines learn that too.

A machine learning algorithm is like a newborn baby that has been given millions of books to read without being taught the alphabet or knowing any words or grammar. The power of this type of information processing is impressive, but there is a problem. When it takes in the text data, a computer observes relationships between words based on various factors, including how often they are used together.

We can test how well the word relationships are identified by using analogy puzzles. Suppose I ask the system to complete the analogy “He is to King as She is to X.” If the system comes back with “Queen,” then we would say it is successful, because it returns the same answer a human would.

Our research group trained the system on Google News articles, and then asked it to complete a different analogy: “Man is to Computer Programmer as Woman is to X.” The answer came back: “Homemaker.”
Examples of bias detected in machine learning word analysis. James Zou, CC BY-ND

**Investigating bias**

We used a common type of machine learning algorithm to generate what are called “word embeddings.” Each English word is embedded, or assigned, to a point in space. Words that are semantically related are assigned to points that are close together in space. This type of embedding makes it easy for computer programs to quickly and efficiently identify word relationships.

After finding our computer programmer/homemaker result, we asked the system to automatically generate large numbers of “He is to X as She is to Y” analogies, completing both portions itself. It returned many common-sense analogies, like “He is to Brother as She is to Sister.” In analogy notation, which you may remember from your school days, we can write this as “he:brother::she:sister.” But it also came back with answers that reflect clear gender stereotypes, such as “he:doctor::she:nurse” and “he:architect::she:interior designer.”

The fact that the machine learning system started as the equivalent of a newborn baby is not just the strength that allows it to learn interesting patterns, but also the weakness that falls prey to these blatant gender stereotypes. The algorithm makes its decisions based on which words appear near each other frequently. If the source documents reflect gender bias – if they more often have the word “doctor” near the word “he” than near “she,” and the word “nurse” more commonly near “she” than “he” – then the algorithm learns those biases too.

**Making matters worse**

Not only can the algorithm reflect society’s biases – demonstrating how much those biases are contained in the input data – but the system can potentially amplify gender stereotypes. Suppose I search for “computer programmer” and the search program uses a gender-biased database that associates that term more closely with a man than a woman.

The search results could come back flawed by the bias. Because “John” as a male name is more closely related to “computer programmer” than the female name “Mary” in the biased data set, the search program could evaluate John’s website as more relevant to the search than Mary’s – even if the two websites are identical except for the names and gender pronouns.
It’s true that the biased data set could actually reflect factual reality – perhaps there are more “Johns” who are programmers than there are “Marys” – and the algorithms simply capture these biases. This does not absolve the responsibility of machine learning in combating potentially harmful stereotypes. The biased results would not just repeat but could even boost the statistical bias that most programmers are male, by moving the few female programmers lower in the search results. It’s useful and important to have an alternative that’s not biased.

**Removing the stereotypes**

If these biased algorithms are widely adopted, it could perpetuate, or even worsen, these damaging stereotypes. Fortunately, we have found a way to use the machine learning algorithm itself to reduce its own bias.

Our debiasing system uses real people to identify examples of the types of connections that are appropriate (brother/sister, king/queen) and those that should be removed. Then, using these human-generated distinctions, we quantified the degree to which gender was a factor in those word choices – as opposed to, say, family relationships or words relating to royalty.

Next we told our machine-learning algorithm to remove the gender factor from the connections in the embedding. This removes the biased stereotypes without reducing the overall usefulness of the embedding.

When that is done, we found that the machine learning algorithm no longer exhibits blatant gender stereotypes. We are investigating applying related ideas to remove other types of biases in the embedding, such as racial or cultural stereotypes.
Unseen Distributions in Genomics: Recovering the Structure of the Unseen Portion of a Distribution

Gregory Valiant and James Zou

What can one say about the unseen portion of a distribution---the portion of a distribution that is supported on domain elements that have never been observed in a dataset? Perhaps surprisingly, one can say a lot. For example, the classical "Good-Turing" missing mass estimator suggests that the ratio of the total number of domain elements observed a single time, to the sample size, provides an accurate estimate of the total amount of probability mass accounted for by these unobserved domain elements. (For example, if you want to estimate the probability that the next word in this paragraph is a new word that has not appeared above, the Good-Turing estimate is the ratio of the number of words appearing once, to the total word count thus far.)

A recent series of work of Gregory Valiant and his collaborators (including publications in STOC, NIPS, FOCS, and an upcoming paper in Nature Communications) goes beyond simply estimating the probability mass accounted for by unseen elements, and gives an algorithm for recovering the structure of this portion of the distribution. While it is impossible to say which unobserved domain elements have nonzero probability, their algorithm gives a provably accurate representation of the frequency spectrum---describing the approximate number of unseen elements that arise with various probability values. This ability to understand the structure of the unseen portion of a distribution has several implications, including enabling accurate inferences about larger datasets. For example, given a dataset of size n, their approach yields accurate estimates of the expected number of new domain elements that would be observed in larger datasets, of size up to log n drawn from the same distribution (which, as they show, is information theoretically optimal).

This work has a number of practical implications, particularly to the problems of anomaly analysis (what is the structure of the distribution over anomalies that have not been seen yet?) and evaluating the benefits of collecting larger datasets (how much more data is necessary before one is likely to see most phenomena?). In a collaboration with James Zou (a new faculty joining Stanford's Biomedical Data Sciences Department) which will appear in Nature Communications, they explore applications of this approach to genomics---a setting for which unobserved mutations abound. Their techniques provide insights into basic questions about the distributions of various types of mutations within and across different demographics. Additionally, they provide answers to questions of the form "How many new, medically relevant SNP mutations are we likely to discover if, instead of 60k genomes, we sequence 100k, or 1M?" Their techniques, providing concrete and theoretically principled answers to these basic questions, provide an essential tool that enables an informed and strategic approach to designing the next phase of large genomic studies.
NETWORK SCIENCE

Higher-order organization of complex networks

Austin R. Benson,1 David F. Gleich,2 Jure Leskovec3

Networks are a fundamental tool for understanding and modeling complex systems in physics, biology, neuroscience, engineering, and social science. Many networks are known to exhibit rich, lower-order connectivity patterns that can be captured at the level of individual nodes and edges. However, higher-order organization of complex networks—at the level of small network subgraphs—remains largely unknown. Here, we develop a generalized framework for clustering networks on the basis of higher-order connectivity patterns. This framework provides mathematical guarantees on the optimality of obtained clusters and scales to networks with billions of edges. The framework reveals higher-order organization in a number of networks, including instances of M 101. In particular, the units in neuronal networks and hub structure in transportation networks. Results show that networks exhibit rich higher-order organizational structures that are exposed by clustering based on higher-order connectivity patterns.

Networks are a standard representation of data throughout the sciences, and higher-order connectivity patterns are essential to understanding the fundamental structures that control and mediate the behavior of many complex systems (1–7). The most common higher-order structures are small network subgraphs, which we refer to as network motifs (Fig. 1A). Network motifs are considered building blocks for complex networks (1, 8). For example, feedforward loops (Fig. 1A, M 1) have proven fundamental to understanding transcriptional regulation networks (8); triangular motifs (Fig. 1A, M 2–4) are crucial for social networks (9); open bidirectional wedges (Fig. 1A, M 5–6) are key to structural hubs in the brain (10); and two-hop paths (Fig. 1A, M 7–10) are essential to understanding air traffic patterns (11). Although network motifs have been recognized as fundamental units of networks, the higher-order organization of networks at the level of network motifs largely remains an open question.

Here, we use higher-order network structures to gain new insights into the organization of complex systems. We develop a framework that identifies clusters of network motifs. For each network motif M, a different higher-order clustering may be revealed (Fig. 1B), which means that different organizational patterns are exposed, depending on the chosen motif.

Conceptually, given a network motif M, our framework searches for a cluster of nodes S with two goals. First, the nodes in S should participate in M. Second, the set S should avoid cutting instances of M, which occurs when only a subset of the nodes from a motif are in the set S (Fig. 1B). More precisely, given a motif M, the higher-order clustering framework aims to find a cluster (defined by a set of nodes S) that minimizes the following ratio:

\[ \sigma_M(S) = \frac{\text{cut}_M(S, \mathcal{S})}{\min \{\text{vol}_M(S), \text{vol}_M(\mathcal{S})\}} \]

where \( \mathcal{S} \) denotes the remainder of the nodes (the complement of S), \( \text{cut}_M(S, \mathcal{S}) \) is the number of instances of motif M with at least one node in \( S \) and one in \( \mathcal{S} \), and \( \text{vol}_M(S) \) is the number of nodes in instances of M that reside in \( S \). Equation 1 is a generalization of the conductance metric in spectral graph theory, one of the most useful graph partitioning scores (11). We refer to \( \sigma_M(S) \) as the motif conductance of S with respect to M.

Finding the exact set of nodes \( S \) that minimizes the motif conductance is computationally infeasible (12). To approximately minimize Eq. 1 and, hence, to identify higher-order clusters, we developed an optimization framework that provably finds near-optimal clusters (supplementary materials (13)). We extend the spectral graph clustering methodology, which is based on the eigenvalues and eigenvectors of matrices associated with the graph (11), to account for higher-order structures in networks.

The resulting method maintains the properties of traditional spectral graph clustering: computational efficiency, ease of implementation, and mathematical guarantees on the near-optimality of obtained clusters. Specifically, the clusters identified by our higher-order clustering framework satisfy the motif Cheeger inequality (14), which means that our optimization framework finds clusters that are at most a quadratic factor away from optimal.

The algorithm (illustrated in Fig. 1C) efficiently identifies a cluster of nodes S as follows:

1. Given a network and a motif \( M \) of interest, form the motif adjacency matrix \( W_M \) whose entries \((i,j)\) are the co-occurrence counts of nodes i and j in the motif \( M \).
2. \( (W_M)_{ij} \) = number of instances of \( M \) that contain nodes i and j.

Fig. 1. Higher-order network structures and the higher-order network clustering framework.

(A) Higher-order structures are captured by network motifs. For example, all 13 connected three-node directed motifs are shown here. (B) Clustering of a network based on motif \( M_1 \). (C) The higher-order network clustering framework. Given a graph and a motif of interest (in this case, \( M_1 \)), the framework forms a motif adjacency matrix \( W_M \) by counting the number of times two nodes co-occur in an instance of the motif. An eigenvector of a Laplacian transformation of the motif adjacency matrix is then computed. The ordering of the nodes provided by the components of the eigenvector \( \psi \) produces nested sets \( S = \{S_1, \ldots, S_n\} \) of increasing size. We prove that the set \( S \) with the smallest motif-based conductance, \( \sigma_M(S) \), is a near-optimal higher-order cluster (15).
Fig. 2. Higher-order cluster in the C. elegans neuronal network. [See (29).] (A) The four-node bi-fan motif, which is overexpressed in neuronal networks (1), intuitively, this motif describes a cooperative propagation of information from the nodes on the left to the nodes on the right. (B) The best higher-order cluster in the C. elegans frontal neuronal network based on the motif in (A). The cluster contains three ring motor neurons (RMEL, -L, and -R; cyan) with many outgoing connections, which serve as the source of information; six inner labial sensory neurons (IL2DL, -VR, -R, -DR, -VL, and -L; orange) with many incoming connections, serving as the destination of information; and four URA motor neurons (purple) acting as intermediaries. These RME neurons have been proposed as pioneers for the nerve ring (21), whereas the IL2 neurons are known regulators of nociception (22), and the higher-order cluster exposes their organization. The cluster also reveals that RHI serves as a critical intermediary of information processing. This neuron has incoming links from three RME neurons, outgoing connections to five of the six IL2 neurons, and the largest total number of connections of any neuron in the cluster. (C) Illustration of the higher-order cluster in the context of the entire network. Node locations are the true two-dimensional spatial embedding of the neurons. Most information flows from left to right, and we see that RMEV, -R, and -L and RHI serve as sources of information to the neurons on the right.

- Step 2: Compute the spectral ordering \( \sigma \) of the nodes from the normalized motif Laplacian matrix constructed via \( W \) (15).

- Step 3: Find the prefix set of \( \sigma \) with the smallest motif conductance (the argument of the minimum), formally, \( S = \arg \min \sigma(S) \), where \( \sigma(S) = \sigma(S) \).

For triangular motifs, the algorithm scales to networks with billions of edges and, typically, only takes several hours to process graphs of such size. On smaller networks with hundreds of thousands of edges, the algorithm can process motifs up to size 9 (15). Although the worst-case computational complexity of the algorithm for triangular motifs is \( \Theta(m^{15}) \), where \( m \) is the number of edges in the network, in practice, the algorithm is much faster. By analyzing 16 real-world networks where the number of edges \( m \) ranges from 159,000 to 2 billion, we found the computational complexity to scale as \( \Theta(m^{13}) \). Moreover, the algorithm can easily be parallelized, and sampling techniques can be used to further improve performance (26).

The framework can be applied to directed, undirected, and weighted networks, as well as motifs (15). Moreover, it can also be applied to networks with positive and negative signs on the edges, which are common in social networks (friend versus foe or trust versus distrust edges) and metabolic networks (edges signifying activation versus inhibition) (15). The framework can be used to identify higher-order structure in networks where domain knowledge suggests the motif of interest. In the supplementary materials, we also show that when a domain-specific higher-order pattern is not known in advance, the framework can also serve to identify which motifs are important for the modular organization of a given network (15). Such a general framework allows complex higher-order organizational structures in a number of different networks by using individual motifs and sets of motifs. The framework and mathematical theory immediately extend to other spectral methods, such as localized algorithms that find clusters around a seed node (17) and
Fig. 3. Higher-order spectral analysis of a network of airports in Canada and the United States. (A) The three higher-order structures used in our analysis. Each motif is “anchored” by the blue nodes i and j, which means our framework only seeks to cluster together the blue nodes. Specifically, the motif adjacency matrix adds weight to the (i, j) edge on the basis of the number of third intermediate nodes (green squares). The first two motifs correspond to highly connected cities, and the motif on the right connects nonhubs to nonhubs. (B) The top 50 most populous cities in the United States, which correspond to nodes in the network. The edge thickness is proportional to the weight in the motif adjacency matrix $W_M$. The thick, dark lines indicate that large weights correspond to popular mainline routes. (C) Embedding of nodes provided by their corresponding components of the first two nontrivial eigenvectors of the normalized Laplacian for $W_M$. The marked cities are eight large U.S. hubs (green), three West Coast nonhubs (red), and three East Coast nonhubs (purple). The primary spectral coordinate (left to right) reveals how much of a hub the city is, and the second spectral coordinate (top to bottom) captures west-east geography. (D) Embedding of nodes provided by their corresponding components in the first two nontrivial eigenvectors of the standard, edge-based (non–higher-order) normalized Laplacian. This method does not capture the hub and geography found by the higher-order method. For example, Atlanta, the largest hub, is in the center of the embedding, next to Salina, a nonhub.

algorithms for finding overlapping clusters (18). To find several clusters, one can use embeddings from multiple eigenvectors and k-means clustering (13, 19) or can apply recursive bipartitioning (13, 20).

The framework can serve to identify a higher-order modular organization of networks. We apply the higher-order clustering framework to the C. elegans neural network, where the four-node “bi-fan” motif (Fig. 2A) is overexpressed (7). The higher-order clustering framework then reveals the organization of the motif within the C. elegans neural network. We find a cluster of 20 neurons in the frontal section with low bi-fan motif conductance (Fig. 2B). The cluster shows a way that nicotinic is controlled. Within the cluster, ring motor neurons (RME1, V or –R), proposed pioneers of the nerve ring (27), propagate information to inner labial sensory neurons, regulators of nicotinic (23), through the neuron RHI (Fig. 2C). Our framework contextualizes the importance of the bi-fan motif in this control mechanism.

The framework also provides new insights into network organization beyond the clustering of nodes based only on edges. Results on a transportability reachability network (28) demonstrate how it finds the essential hub interconnection airports (Fig. 3). These appear as a peak on the primary spectral direction (Fig. 3C) when two-hop motifs (Fig. 3A) are used to capture highly connected nodes and nonhubs. The first spectral coordinate of the normalized motif Laplacian embedding was positively correlated with the airport city’s metropolitan population with Pearson correlation 99% confidence interval (0.33, 0.53). The second spectral direction identified the west-east geography in the North American flight network (it was negatively correlated with the airport city’s longitude with Pearson correlation 99% confidence interval (−0.68, −0.50)). On the other hand, edge-based methods confine geography and hub structure. For example, Atlanta, a large hub, is embedded next to Salina, a nonhub, with an edge-based method (Fig. 3D).

Our higher-order network clustering framework unifies motif analysis and network partitioning—two fundamental tools in network science—and reveals new organizational patterns and modules in complex systems. Prior efforts along these lines do not provide worst-case performance guarantees on the obtained clustering (29) and do not reveal which motifs organize the network (28) but rely on expanding the size of the network (26, 27). Theoretical results in the supplementary materials (13) also explain why classes of hypergraph partitioning methods are more general than previously assumed and how motif-based clustering provides a rigorous framework for the special case of partitioning directed graphs. Finally, the higher-order network clustering framework is generally applicable to a wide range of network types, including directed, undirected, weighted, and signed networks.

REFERENCES AND NOTES
RESEARCH | REPORTS

12. Minimizing $w(s)$ is non-deterministic polynomial-time hard (NP-hard), which follows from the NP-hardness of the traditional definition of conducitance (19).
13. See the supplementary materials on Science Online.
14. Formally, when the motif has three nodes, the selected cluster $S$ satisfies $w(s) \leq 4 \sqrt{w(v)}$, where $w(v)$ is the smallest motif conducitance of any possible node set $S$. This inequality is proved in the supplementary materials.
15. The normalized motif Laplacian matrix is $\Delta = D^{-1/2} W D^{-1/2}$, where $D$ is a diagonal matrix with the row-sums of $W$ on the diagonal. $I$ is the identity matrix. The inverse square root of the diagonal $D^{-1/2}$ is the same matrix with the same entries as before. The spectral ordering is the by-value ordering of $\Delta^{-1/2}$, where $v$ is the eigenvector corresponding to the second smallest eigenvalue of $\Delta W$, i.e., $\alpha_i$ is the index of $\Delta^{-1/2} W$ with the $i$th smallest value.

ACKNOWLEDGMENTS
The authors thank R. Sosl for insightful comments. A.R.B. was supported by a Stanford Graduate Fellowship. D.F.G. was supported by NSF (CPS 1219796 and IIS 1420518). J.L. was supported by NSF (IIS-1148437 and CNS-1010021). Trans-AIR initiative Big Data to Knowledge (BD2K), Defense Advanced Research Projects Agency (DARPA and Simplifying Complexity in Scientific Discovery (SIMPLEX) Boeing, LightSpeed and Volkswagen. Software implementations and the data sets used to obtain the results in this manuscript are available at http://navier.stanford.edu/NavierOrder/

SUPPLEMENTARY MATERIALS
www.sciencemag.org/content/353/6295/8595/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S13
Tables S1 to S12
References (S0–S4)
18 November 2015; accepted 15 May 2016
10.1126/science.aaf6529
The Stanford Data Science Initiative is Stanford’s university-wide program for big data. SDSI conducts research on massive data, new algorithms, and advanced analytics with strong ties to application areas across Stanford, with industry, and worldwide.

sdsi.stanford.edu