Riding the Big Data Wave
Trinity College embraces data-focused programs in the arts and sciences

Every two days, more data are created than were recorded from earliest civilization until 2003. Every minute, YouTube adds seventy-two hours of new video, Google maps another neighborhood, and billions more tweets and texts merge into the ever-growing data stream. We are awash in pixels and bytes, a plentiful and renewable resource awaiting skilled and thoughtful minds to find meaning in this swell of digital matter.

“The unprecedented volume of information flooding our society will be a wave that washes over all universities, and if you’re not thinking proactively about it, you will drown,” predicts Robert Calderbank, the Phillip Griffiths Professor of computer science, mathematics, and electrical and computer engineering, and director of Information Initiative at Duke (iiD), a new interdisciplinary program designed to increase “big data” research on campus. “It will change every university and the work they do.”

Trinity College of Arts & Sciences is already addressing the challenge. Humanities labs at the Smith Warehouses hum with interdisciplinary projects involving data visualization, allowing students to experience art and history in a new way. Social scientists analyze massive sets of employment and health data through the secure Triangle Research Census Data Center to find revealing patterns of human behavior. And at Gross Hall, a $20-million renovation encourages scholars in various disciplines to collaborate with the big-data brain trust behind the iiD—a group that includes mathematicians, computer scientists, statisticians, electrical engineers, and more.

“Everything we do in our contemporary society is digital,” says Cathy Davidson, the John Hope Franklin Humanities Institute Professor of interdisciplinary studies and Ruth F. Devarney Professor of English. Indeed, a whopping 98 percent of our stored data is now digitized, up from 25 percent at the dawn of the millennium. “Humanists must engage profoundly and completely with the virtual, the digital, and the material to understand the world we live in today. It’s a tremendously exciting moment.”
Boundary Issues
BorderWork(s) Labs map the world’s complexities.

Borders are never merely lines drawn in dust. They’re meant to contain or exclude, threaten or pacify, unite or fracture. In other words, they’re about conflict. The humanities lab initiative BorderWork(s) looks at how these powerful lines of demarcation change lives and shape history. In the process, students learn how to turn complex, hard data into visual statements with emotional impact.

Offered through the Franklin Humanities Institute, these interdisciplinary labs bring together faculty members and students from the social sciences and humanities to work on themed projects and independent studies. The goal of the research is to create visual documentation—a map or an exhibit, an app, or even a film series—that can be used for educational purposes or as source material for future global policy initiatives.

“We want students to create projects that have wider impact than a paper they hand back to me,” says Philip Stern, one of BorderWork(s)’ three codirectors and an associate professor of history. “These projects will have another life, a continuing life.”

Stern’s two-year ongoing project, “Maps, Art and Empire,” resulted in a website and cartography exhibit curated by the students themselves that ran for three months at the Nasher Museum.

Under Stern’s supervision, students researched specific zones of European empires and selected maps of Africa, India, and South America from the Rubenstein Rare Book & Manuscript Library to illustrate their findings. More than simply old maps hanging on white walls, the exhibit depicted how and why colonialism took root and spread. “We try to think about the relationship between text and data and how we quantify and visualize all these complicated concepts involving borders,” says Stern.

Erika Weinthal, another BorderWork(s) codirector and an associate pro-

A Perfect Fit
An aspiring filmmaker and environmentalist finds her niche at FHI.

“I have lived in West Bank, Palestine, and have backpacked in Jordan and Israel, so I’ve personally witnessed the environmental impact of conflicts in the region,” says freshman Leo Lou, one of the students cataloguing UNEP film for Erika Weinthal’s BorderWork(s) project. When Lou, who plans to major in public policy with a minor in environmental sciences and policy and a certificate in ethics, found out about the lab at the Franklin Humanities Institute open house, she knew she’d found her ideal class.

“It was a no-brainer,” she says. “I had experience and extensive interest in documentary filmmaking and storytelling, and on top of that, I’m extremely interested in border-conflict resolutions.” And while her interests and personal experiences made her well-suited to the project, she says it has exceeded her expectations, in part, because of Weinthal’s extensive professional connections in the environmental peacekeeping world. “The project is so unique in that it offers us an authentic and comprehensive look into some of the core operations of UNEP in post-conflict regions—something barely anyone gets to do since the UN takes great caution in deciding what the public does and does not see,” she notes.

Weinthal also brought to the lab guests from the UN and the Woodrow Wilson
fessor of environmental policy, has an entirely different data problem to solve in her new lab. In a scene she describes as “something out of James Bond,” Weinthal was given a huge cache of films shot in conflict and disaster zones from the United Nations Environmental Programme (UNEP). Collected over a decade, this unedited footage depicts scenes of environmental destruction in Afghanistan, Democratic Republic of Congo, Haiti, Iraq, Liberia, Nigeria, and Sierra Leone, among other countries.

“Only a small number of people had ever looked at this material, and even then only minimally,” she recalls. “I wasn’t even sure where to begin.” But now, students enrolled in her UNEP film lab are helping to catalogue all the footage—a huge job—so that the 200 hours of raw film can be analyzed and edited into a series of five-minute videos to be used by the UNEP or other organizations for educational purposes.

For her students, the project (co-funded by Bass Connections) has been a powerful reminder of what visual data can do that statistical data sets cannot. “When we look at things only quantitatively, we’re often missing a big part of the story,” says Weinthal. “The inequalities built into the variables, into the analysis itself, how the data [are] collected [and] analyzed, can mask injustices and important narratives.”

BorderWork(s) labs are designed, she says, to break down these barriers and get to a deeper truth.

Center, some of whom were involved in the operations depicted in the films. “I’ve been able to sit down with most of the visitors for on-camera interviews and gain insights from their decades of work and expertise,” says Lou. “I absolutely plan to continue with the project for as long as it takes.”

Charting a New Course

Data visualization reveals a complex network of student and faculty involvement in humanities projects.

Public perception of humanities scholarship seems so shrouded in the myth of the reclusive scholar,” notes Christina Chia Ph.D. ’04, associate director for programs and communication at the Franklin Humanities Institute. “We wanted to show how humanities research and teaching actually work at Duke.”

A network seemed the obvious way to document the phenomenon, so Chia and her colleagues created a series of data visualizations. The finished graphics show how students come to the labs in the first place, as well as “the glue” that connects them to multiple projects.

At any given time, there are at least two undergraduate labs and one graduate lab in full swing at FHI. Housed in the loft-like Smith Warehouse complex, each lab has three to five faculty members, support staff, and graduate students, and most of them are working on satellite projects of their own. Students funneling into this creative vortex often take up research roles in one or more of these projects, too.

“We started to see in the rosters of students the overlaps, the names that kept popping up,” she says. “I kept telling people stories about these students doing all this amazing work, and finally someone said, ‘Where’s your data?’”

The graphics make it all very clear. For instance, the BorderWork(s) Focus cluster, “Humanitarian Challenges,” has been an effective portal to the labs. From there students have moved onto more advanced research projects, such as the joint BorderWork(s) and Bass Connections course on energy and peace-building, as well as a graduate student-led project on South Sudan. Chia hopes that lab directors can use the visuals to help assess their work and new initiatives.

“We’re teaching through these very active projects,” says Chia. “Students are contributing to research led by faculty and becoming researchers themselves. It’s incredibly effective.”
Follow the Numbers
Information, Society & Culture projects bring big data to the humanities.

It’s easy to forget that Galileo was a philosopher as well as a physicist and that Isaac Newton was as much a theologian as mathematician. Brilliant discoveries and innovation have always sprung from interdisciplinary learning. But too often quantitative disciplines are kept at arm’s length from qualitative ones, says Cathy Davidson, the John Hope Franklin Humanities Institute professor of interdisciplinary studies. “The industrial age brought each field to its own silo, its own expertise,” she says. “We need to come up with a new model. We need to rethink for this era.”

In a digital age preoccupied with technology and its chief byproduct, big data, that new model might look something like the “Information, Society & Culture” humanities project offered through Bass Connections, the $50 million donor-funded initiative that debuted in the winter of 2013.

The project consists of seven interdisciplinary classes designed to explore how information flows through society, how it can change perceptions of culture, and even rewrite history. “If you want to make a societal impact, you need courses with some cultural currency,” says Hans J. Van Miegroet, professor and chair of the Department of Art, Art History & Visual Studies, who co-developed the theme with Robert Calderbank, professor of electrical and computer engineering and director of the Information Initiative at Duke (iiD).

Like all of the multiple Bass Connections initiatives, those under the Information, Society & Culture theme combine lectures, research, and group projects; and are co-taught by professors in fields as disparate as dance and engineering, archaeology and computer science. That shared expertise, Van Miegroet says, gives students a richer, deeper experience in a subject. “Hybrid students are the future,” he says. “But the knowledge base cannot be shallow.”

Davidson’s course, “Making Data Matter,” is co-taught with a computer science professor from UNC-Chapel Hill, whose research on urban renewal in Asheville became the focus of the class project. A graduate student and three undergraduates analyzed real-estate records, oral histories, and data sets to create a crowdsourcing tool to help the citizens of Asheville’s Southside neighborhood chart the changes within their community.

“It was like working for a start-up,” laughs Davidson. The students brought their individual interests to the effort: A sociology student pored over old real-estate petitions to find evidence of the type of citizen resistance found in other social movements. A statistics major crunched population, employment, and income statistics to design an economic map of the changing community. Readings on urban renewal provided historical context, and lectures on data security and privacy broadened students’ understanding of IT issues.

Students could see for themselves “the cultural patterns emerge through data,” Davidson says.

Nowhere was this more evident than in the impressive crowd-sourcing website developed by Lalita Maraj, a junior in computer science, and Rachel Anderson, an undeclared sophomore.

Using Big Board, a chat room managed by Google Maps that digitizes property documents, Maraj and Anderson found a way to bring all the relevant data from Asheville under one virtual roof. The website, located on GitHub, is designed to be easy to navigate and includes a message board and a discussion tool to manage conversation. The open-source code allows anyone access—there are no proprietary claims, but the original is protected—and Davidson and her students hope the work they began will continue in Asheville and other cities, too. And when Big Board turned out to be too technical for non-scientists to use, Maraj and Anderson built an alternative site specifically designed for “community-sourcing.”

Like so many other humanities projects, this one will continue to grow with the power of collaborative resources. Certainly the original hybrid scholars would approve: “We build too many roads and not enough bridges,” noted Isaac Newton, who, though speaking of the past, might as well have been talking about our future. ■
Invariably, students imagine archaeology—even in the twenty-first century—as a low-tech and decidedly dusty pursuit popularized in movies and novels. Which is why Maurizio Forte, founder of the DIG@Lab, begins his classes by dispelling such romantic notions. Forte, also a professor of classical studies, art history, and visual studies, then dazzles students with the technology and modern methodology that are revolutionizing and transforming the field.

Forte, a pioneer and leader in the field of digital archaeology, describes himself with a chuckle as “borderline Archimedes,” referring to the ancient Greek inventor. In one sense, he is a traditionalist, who has coordinated archaeological projects in Italy, Ethiopia, Egypt, Oman, India, China, Honduras, Mexico, Spain, and Turkey. But he is also an innovator, who started using computer applications in archaeology as a student in the early 1980s. “My first book was on virtual archaeology, and it was criticized by many in the community,” he recalls. “And now we have Ph.D.s in the field.”

Forte excavates real and virtual landscapes with the help of cutting-edge technology, such as 3D documentation, geographic information systems (GIS), remote sensing, spatial technologies, aerial photography, open software WEB-VRGIS, and virtual reality. As a result, digs can be more environmentally friendly, and the complex data gleaned from a site’s artifacts and structures can be digitally shared around the world.

“All these methodologies return information we didn’t have before,” he says. “We have the capacity to analyze models through different soils. We can get information about color, texture, composition, physical evidence of some features, and transform all this into a virtual data set to increase our perception.”

Samantha Bernstein, a junior majoring in mechanical engineering and minoring in classics, took Forte’s “Principles of Archaeology” class and was amazed by the amount of technology and technical knowledge archaeologists must have at their disposal. “As an engineer, it was so exciting to see all of the new inventions that he is putting to use to discover more about ancient civilizations,” she says. “In his lab he has a drone that can fly around a site and take pictures at predetermined locations, so cool!”

In Forte’s Bass Connections project, “Envisioning the Digital Landscape,” part of this year’s Information, Culture & Society theme, students explored Vulci, Italy; Catalhoyuk, Turkey; and Venice. They recreated aspects of ancient Venice with remarkable accuracy and excavated a Roman landscape through remote sensing software, satellite imagery, and geophysical data. Using the same technology, one group of students even discovered a previously unknown site of Roman villas and other buildings and made a digital map of it.

“Only four months in a lab, and they made a discovery,” says Forte, still amazed by how quickly students mastered the software and algorithms. “It was 90 percent the students’ work. And some of them didn’t have any background in the field.”

Indeed, most students don’t have a background in archaeology. “Many of my students come from different backgrounds—computer science, biology, engineering—and they immediately find a link between their primary academic interest and archaeology.” One recent student wrote a paper on robotics in archaeology; another did a DNA project. “Very soon they have removed all their stereotypes of archaeology, and they start to think something different is possible.”
Match.Duke
Big data analytics enable students to find their ideal adviser.

The right academic adviser can ensure that a student makes the most of all that Duke has to offer. Yet too often the relationship between student and adviser is more of an arranged marriage than a perfect pairing. When ElizaBeth Fox arrived at Duke three years ago, as associate dean of Trinity College and director of the Academic Advising Center, she quickly understood why: The system for matching first-year students with advisers was outdated.

Fox organized background information from 1,400 incoming students and 110 academic advisers on two Excel spreadsheets and then toggled between to make manual matches. The system allowed advisers to take twelve students or zero students, but nothing in between. Advisers’ schedules fluctuated term-to-term, creating a logistical nightmare. Fox realized she needed to make the system more flexible to draw more faculty participation, so she offered the option of three, six, nine, or twelve students per adviser. The plan worked, and soon she had more than doubled the number of advisers willing to take on students. But if the old paper-pushing system was impractical before, it was doubly so now.

“I would simply run out of time,” recalls Fox. “Some matches were good, but most were relatively random, basically luck of the draw.” The system began to collapse under its own weight. One semester the spreadsheet was corrupted, and Fox lost all her work. The memory still makes her shudder. “I’ll admit, I was pretty frustrated,” she says, ruefully.

Fox shared her concerns with Ed Gomes, senior associate dean for the Trinity College Office of Technology Services, who oversees IT planning for arts and sciences departments and programs, and he and Fox worked with his team to find a solution.

“There are many ways that companies take data and build algorithms that match people with the right people or service. They do it in the medical system to create customized health care,” says Gomes.

And if the ads are to be believed, they do it with some success in the dating world, too. Gomes laughs at the comparison. “I would hope that the types of questions being asked to help you find an ideal mate would not be the same types of questions asked to help you find an ideal adviser for your academic career,” he says.

Gomes and Fox decided it was important to keep the human element in the process so that Fox and her team could use what they knew from talking to students and faculty members to inform final placement.

The new system was up and running for the 2013 freshman class. Gomes provided Fox with an IT troubleshooter to help smooth out glitches in real time during the rollout. Since then, students and advisers have expressed satisfaction with their matches, and the experience has been a professional-development boon for Fox’s staff. Fox’s advising coordinator, Christie Phipps, a Duke employee for twenty-five years who had never run a database before, is now an expert in the new software.

“I quantified the number of human hours saved and the added value to our employees’ portfolios, and the cost savings to the university is substantial,” Fox says. “We’re doing twice as much with the same amount of human resources. That’s exactly what good data systems, big or small, are supposed to do.”

Mining for Meaning
Professor Seth Sanders finds unique data sets within mountains of U.S. population information.

Sometimes you can find data which no one has been using or has been forgotten,” says Seth Sanders, professor of economics and public policy. “And sometimes you find a data set that no one has noticed.” Sanders specializes in sifting through massive, amorphous data sets to find the hidden research gems. His work explores the economic consequences of teen childbearing, the performance of gay/lesbian families in the economy, public health and the African-American Great Migration, and gender and racial demographic trends among the highly educated.

Applying the tools of econometrics, Sanders chips away until patterns between people and behaviors begin to emerge. “Most of my research comes from data collected for reasons other than my intended purpose,” he says. “It’s a creative process, not just a statistical exercise. The challenge is how to take it apart and then how to put it all together.”

Take, for instance, his research on the Great Migration. He was curious to find out whether railroad proximity played a significant role in determining which African-American families moved north and which remained in their small Southern communities. “It turns out you need massive amounts of data for this,” he laughs. By combining survey information on railroad tracks laid in the late 1800s with the population data of 80 million people, Sanders was able to show that being born on a railroad line had an enormous effect on increasing the rate of migration. “No one had put a number to this,” he says. “But big data lets you become very precise about patterns that might have been discussed only generally in history.”

Sanders’ students also have benefited from his big-data expertise. Deborah Rho, one of Sanders’ Ph.D. students, re-
Global Education Draws Interest Across Disciplines

Yet data show there’s room for more communication between programs.

Whether through Focus clusters for first-year students, January’s Winter Forum, DukeEngage, or other service-learning opportunities, Duke’s commitment to nurturing global citizens has never been stronger. But the expansion of courses, labs, externships, and interdisciplinary initiatives has occurred more often than not in a decentralized manner, says Keith Whitfield, vice provost for academic affairs. “We’re so busy implementing and doing that we haven’t looked at how our many global programs are connected relative to our curriculum,” he says. It was a question ripe for big-data analysis.

To find out how students make their way through the various offerings, Whitfield turned to his friend and colleague, sociology professor James Moody, for help collecting and analyzing Trinity student data. “I approached it as a network problem,” says Moody, who used his expertise in social-network theory to create a visualization of student participation. “I wanted to see how individuals are linked through activities and behaviors.”

By crunching data from 18,000 student records from 2008 to 2012, Moody found that most students take advantage of multiple global-education opportunities during their time at Duke, with interest in one often leading to interest in another. “This suggests a community of students who interact together in multiple settings, likely enriching their experience in any one of the programs,” says Moody. “That about half of the students who graduate with distinction have not been involved in one of the global-education programs suggests room to grow. And it raises new questions: What other activities are engaging those students and how we might help facilitate linkages within that community?”

Whitfield plans to track this dynamic year to year with the goal of improving communication and cohesiveness between programs. “At some point, I’d like to see we have a capstone to bring together all these experiences and see what it means to the students,” he says. “Global issues flow through so many majors, we must make sure all these pieces meet the desires and interests of our students.” Not to mention the needs of an ever-changing world.

A recently completed study of income and employment information from workers in twenty-seven states to pinpoint specific patterns of income disparity between immigrant workers and native workers. These data are housed in the Triangle Research, Census Data Center, where scholars can access large administrative and survey data with special permission from the U.S. Census Bureau. “The big advantage of the data for this study is that there are large enough samples of immigrants,” says Rho. “The vast majority of panel data sets that labor economists use do not have enough immigrants because of the nature of the sampling or the sample size.” Rho’s findings overwhelmingly support the argument that low-skilled immigrant workers are among the most proactive and determined in the American economy.

The data reveal that when immigrants come to the U.S. as low-skilled workers they earn much less than their American counterparts. However, their income rises more rapidly than low-skilled Americans, and within ten years the income gap is nearly closed. The rate at which an immigrant worker is out of the workforce is much lower than a native worker, and the immigrant worker is much more likely to be working two jobs. “This pattern is extremely consistent across time—it’s been known for thirty years by economists—but we haven’t been able to determine why immigrant earnings rise so rapidly with time in the U.S. until now,” says Sanders. “It has taken huge data sets that track almost every American worker on every job that they hold to see the different work patterns of immigrants.”
Welcome to the Big Data Hub
A newly sleek and modern Gross Hall is ground zero for Duke’s big data initiative.

At a recent Friday meeting in Gross Hall, Tom Nechyba, professor of economics and public policy studies, and colleagues from the Bass Connections education theme were discussing how to measure the success of a music therapy intervention for children with autism. The goal of the therapy is to help the children read emotions in others and to express themselves. In the conversation, someone recalled hearing biomedical engineering professor Guillermo Sapiro talk about his specialized cameras that can track mouth movements to detect degrees of smiling. Suddenly, a possible solution was at hand, and with it, an opportunity to share resources across vastly different disciplines.

“That wouldn’t have happened if we weren’t all sharing this space,” says Nechyba, referring to the newly redesigned Gross Hall, home to the Information Initiative at Duke (iiD), an interdisciplinary program that will vastly expand the university’s big-data capabilities and the Social Science Research Institute, directed by Nechyba. The $29 million renovation features communal spaces, labs, team rooms, and classrooms designed to encourage innovative research. “The collaboration is already happening, and we’ve only been here three months,” he says.

Collaboration is the guiding principle of the iiD. The initiative brings together leaders in the fields of computational science, Bayesian statistics, image processing, genomics, remote sensing, wireless devices, social science, health care, signal processing, finance, machine learning, and graduate students from various departments who’ve taken up residence in the bright, new space. “Robert’s group has this extraordinary tool kit to bring to social sciences,” says Nechyba. “They can find manifolds in data and ask us if we think they’re interesting from a social science perspective, and we can provide the theoretical and conceptual frameworks.”

Nechyba and his SSRI team—particularly Carol Ripple and Rachel Franke—are focused on data security and privacy issues. A corporation or government entity lending its data to Duke wants assurances it will be protected from hackers, theft, and misuse, says Nechyba, but the university wants to ensure the data can be used effectively for scholarly purposes, too.

“So, how to marry the protected data infrastructure with big-data needs?” he asks.

Much of the data is highly sensitive. (Case in point: The iiD group is currently working on research projects for the U.S. Department of Homeland Security.) Typically social science data are kept in locked rooms and on computers disconnected from the Internet. But that system means caches of data are studied in isolation. So SSRI is building a virtual bunker to protect valuable information while allowing approved researchers to compare large data sets to one another.

“Our solution is easily scalable, makes it convenient for collaboration, and you can be in different parts of the world and access it,” says Nechyba. Access will be restricted with the latest authentication measures, of course, and privacy will be protected. “You know immediately if there is a security breach,” he says. “There can be trip wires so you know something has happened. The technical aspects are evolving.”

“It’s a brave experiment,” Calderbank says proudly, and one that our era requires. “The U.S. is a service economy now, and we must become better at doing information on a large scale.”