TEACHING DATA AND COMPUTATIONAL JOURNALISM

Charles Berret & Cheryl Phillips

Columbia Journalism School

Knight Foundation

Columbia Journalism School
Pulitzer Hall
2950 Broadway
New York, NY 10027

Teaching Data and Computational Journalism
by Charles Berret and Cheryl Phillips

ISBN: 978-0-692-63745-6

Printed in the USA by Rosemont Press in Deer Park, NY

This book was set in Adobe Jenson Pro and Arno Pro,
both by Robert Slimbach.

The cover, layout, and graphics were designed by Jessica Griscti.
Find her at jesslovestype.com

CONTENTS

PREFACE

The digital revolution ushered in fundamental changes in how information is structured. It also brought changes in how governments and corporations use information to exercise power. Governments now influence communities through the management of large data sets, such as in the allocation of services through predictive policing. They hold exclusive access to data that would help us to understand which policies are working, or how vulnerable populations are affected by the exercise of public policy. Corporations write opaque algorithms to determine who gets insurance at what price. These developments challenge journalism to move well beyond adaptation to social media or the adoption of new technologies for visualization. They implicate journalism's public purpose. Encouragingly, a new facet of journalistic practice is emerging, adapting technology to reporting in the public interest.

This is an important reason why we must teach journalists to work with data: There are vital questions to be asked that require numeracy, and there are big stories to find and tell in new ways. The intellectual history of journalism reveals a continuous interrogation of emerging technologies for their relevance to the profession's public purpose and concerns. We need journalists to be positioned to assess techniques like natural language processing and facial recognition for their relevance and promise as tools of reporting, as well as for their ethical dangers.

This is where journalism education may play a leadership role. Integrating computation, data science and other emerging technologies into public-spirited reporting is an ideal mission for journalism. These schools can access the full resources of a university. The mission also relieves journalism educators of the risk of teaching perishable digital skills, and perishable platforms. Data journalism curricula respond to objective change in the sheer amount of information that is stored digitally today – information that requires computation to access and interrogate. Teaching journalists to be literate about these changes and some to be specialists requires committing ourselves to using data, computation, and emerging technologies as essential tools of our profession.

Steve Coll
Dean & Henry R. Luce Professor
Columbia Journalism School

EXECUTIVE SUMMARY

Over the past century, journalism schools have developed solid foundations for teaching shoe-leather reporting techniques. Hundreds of universities teach how to interview, how to develop sources, how to cover a beat, and how to write a breaking news story, a feature, a sports dispatch, or an investigative piece.

But the practice of data journalism has been largely left out of the mainstream of journalism education, even as the field's relatively small core of devotees has honed it into a powerful and dynamic area of practice. For decades, data journalists have competed for the profession's highest prizes and secured positions of distinction within the most competitive news organizations, yet our research has found that relatively few journalism schools offer courses in this area, let alone a concentration, even as these schools have expanded instruction in presentation-focused digital skills.

The authors of this report believe that all journalism schools must broaden their curricula to emphasize data and computational practices as foundational skills. To place data journalism in the core of journalism education will mark a crucial advance in what schools can offer their students. Journalists who understand data and computation can more effectively do their job in a world ever more reliant on complicated streams of information.

Beyond teaching, too few journalism schools support faculty research into tools and techniques of data-driven reporting, despite rich opportunities for developing theories and applications that may change journalistic practice. Journalism schools that embrace research in their missions can transform themselves into innovation hubs, introducing new tools and techniques to the profession and across their universities, instead of merely preparing students to enter the field.

This report offers a snapshot of the state of data journalism education in the United States and outlines models for both integrating the use of data journalism into existing academic programs and establishing new degrees that specialize in data-driven and computational reporting practices. While we focus on the state of education in one country, we hope that the results may also be useful internationally.

But first, a definition. When we say "data journalism," we mean using data for the journalistic purpose of finding and telling stories in the public interest. This may take many forms: to analyze data and convey that analysis in written form, to verify data found in reports, to visualize data, or to build news apps that help readers to explore data themselves. This field also encompasses the use of computation—algorithms, machine learning, and emerging technologies—to more effectively mine both structured and unstructured information to find and tell stories. The ability to use, understand, and critique data amounts to a crucial literacy that may be applied in nearly every area of journalistic practice.

We interviewed more than 50 journalists, educators, and students, and we evaluated more than 100 journalism programs across the nation. This report features a chapter detailing quantitative findings, such as the number of U.S. journalism programs offering classes in data, computation, and related tech skills. We also include a chapter of qualitative findings in which our interviews and classroom observations offer some color and texture to this picture of the present state of data journalism education and its potential.

AMONG OUR FINDINGS:

- Many journalism programs offer few courses in data journalism, and nearly half offer no classes at all.
- The classes offered are largely introductory, and the need is still largely for the basics, such as knowing how to use a spreadsheet, understand descriptive statistics, negotiate for data, and clean a messy data set and then "interview" it to find a story.
- The field offers a few foundational textbooks, but beyond that lacks a broad and strong core of literature to help teach both the history and practice of data journalism.
- Many journalism programs do not have a faculty member skilled in data journalism. Hiring professional journalists as adjuncts may pose many challenges, one of which is that job openings outnumber qualified applicants.
- Graduates with data journalism skills are better equipped to succeed, our interviews show. Faced with a decision to hire an entry-level reporter with no data skills or one who knows how to use a spreadsheet or query a database, the data skills provide a key edge.

10 AMONG OUR RECOMMENDATIONS:

- Journalism schools can collaborate across the university to meet the burgeoning need for instruction in data and computation but should be wary of trying to outsource too much—while understanding how to do math, statistics, or computer programming is an important component, data journalism is much more than that.
- Journalism programs can integrate alternative teaching methods to help fill the gaps in their own faculty. Examples include cooperative teaching among different university departments, online courses, and independent tutorial packs.
- Journalism programs can choose among several models of instruction, all of which begin with a key component: at least one required class in analyzing data for stories—what historically has been termed computer-assisted reporting (CAR).
- Journalism schools that embrace both teaching and research into data journalism methods will be poised to fundamentally improve the way future journalists will inquire into matters of public interest and communicate with their audiences.

Following our findings, this report outlines several model curricula and general recommendations. We offer a model for a core, required course in data journalism. Then we suggest ways of introducing data and computation into existing journalism classes such as Ethics and Global Reporting. Next comes a set of full model curricula for degrees and concentrations in data and computational journalism. Finally, we address a range of institutional concerns on matters ranging from finding teachers to providing technological infrastructure.

Our objective is not to replace or diminish shoe-leather reporting in journalism instruction, but to augment it with data-driven and computational techniques. This report is meant to describe the state of data journalism education, to underline the urgency of incorporating these skills to equip the next generation of reporters, and to offer guidelines for moving journalism schools in this direction.

INTRODUCTION

Journalism schools have a long history of avoiding the call for instruction in quantitative skills. A little over a century ago, when the idea of teaching journalism at the college level was practically unthinkable, Joseph Pulitzer wrote an essay arguing for the potential and civic importance of journalism education—all as a response to several schools refusing the money he had hoped to donate in order to establish such a school. In this 1904 essay in the North American Review, Pulitzer outlined the skills he thought journalists would need in order to serve this lofty civic role. It was an ambitious list, highlighting law and ethics, history and literature, truth and accuracy, as well as a range of mathematical and scientific disciplines. Among these, Pulitzer specifically insisted on educating journalists in statistics.

> Everybody says that statistics should be taught. But how? Statistics are not simply figures. It is said that nothing lies like figures—except facts. You want statistics to tell you the truth. You can find truth there if you know how to get at it, and romance, human interest, humor and fascinating revelations, as well. The journalist must know how to find all these things—truth, of course, first.[1]

This proposal for a statistics curriculum was largely left behind in the wave of journalism programs that were established in the twentieth century, including at Columbia, the school that bears Pulitzer's name. Core reporting classes have taught students to gather, analyze, and present information, mostly through shoe-leather reporting and writing skills.

Data journalism and other quantitative reporting methods, on the other hand, have been developed largely in the field. Working journalists were the ones who first saw the potential of analyzing and presenting data, and of adopting tools such as spreadsheets and databases for stories, visualizations, and apps. Much of the instruction has come through professional workshops, not in classrooms.

---

1 Pulitzer, "Planning a School of Journalism," p. 53.

To be sure, journalism programs have offered classes in research methodology, descriptive statistics, and basic numeracy. The Accrediting Council on Education in Journalism and Mass Communications (ACEJMC), which accredits roughly a fourth of the journalism programs in the nation, lists the ability to "apply basic numerical and statistical concepts" among the core competencies it expects its accredited schools to teach. This accreditation process is designed to be a voluntary process that helps schools maintain quality by meeting a set of national standards. Benchmarks such as statistical concepts are useful, but they don't get to the heart of what it means to teach data journalism.

Many of the statistics and numeracy courses required for journalism majors are theoretical in nature, rather than journalistic. Despite the inclusion of basic statistics in schools of journalism and communications, the use of data and computation as applied to journalism has remained a set of niche practices, often omitted from journalism programs, our analysis shows.

Some of the early quantitative reporting methods that made the jump from practice to the classroom came with the move of key data journalists to academia. Over time, video and web-based multimedia, such as slideshows or timelines, were integrated into journalism instruction. Web and multimedia skills instruction now seems to outnumber data journalism instruction.

To get a better sense of what is being taught, we collected information on the course offerings of 113 programs located within the United States, including Puerto Rico. Four of the programs held a provisional accreditation and the remainder were fully accredited with ACEJMC. Chapter 2 of this report includes a longer discussion of our findings, while full tables of our data can be found in the appendix.

Of the 113 programs, 93 offer multimedia instruction—how to design a website, launch a blog, or shoot video for the Web. The average number of multimedia classes was 3. Far fewer journalism programs offer data analysis or visualization. A little more than half of these universities, 59 of the 113 schools we reviewed, regularly offer one or more data journalism classes. Among the 59 that teach data journalism, the average number of data journalism classes offered was 2.8, with a median of 2. The average across all 113 schools in our study was 1.4 data journalism classes each.

We defined a data journalism class as being focused on the intersection of data and journalism, and using spreadsheets, statistical software, relational databases, or programming toward that end. We excluded courses in numeracy, research methodologies, and statistics unless they included an explicit focus on data journalism.

The 59 programs we identified as offering data journalism included a wide range of courses. At a minimum, programs offered courses that taught students to use spreadsheets to analyze data for journalistic purpose. At the other end of the spectrum, some schools provided that basic data journalism instruction and far more, teaching multiple classes in programming skills, such as scraping the Web, building news apps, or creating advanced data visualizations.

But those more advanced programs were rare. Of the 59 programs we identified that teach at least one data journalism class, 27 of the schools offered just one course, usually foundational. Fourteen journalism programs offered two classes. Just 18 of the 59 schools offered three or more classes.

For those students who learn data journalism, a robust job market awaits. But when it comes to teaching data journalism, it's difficult to find journalists to do it full time. Many work as adjuncts, but the pay is low. Additionally, even as some universities add classes in web development and coding, they have not kept pace with offering courses in computer-assisted reporting skills like learning how to analyze and understand data.

For advanced positions in data journalism—jobs that deal with statistics and mapping, novel forms of data visualization, rich online databases, and machine learning—little is available in the way of data journalism education preparation. Students who study both computer science and data journalism are well positioned to move into some of these more challenging jobs, but there is a dearth of such job candidates, say data journalists.

This paper will delve into the state of data journalism education today and present the lessons learned from those who have taught, studied, and practiced data journalism—what doesn't work and should be abandoned, and what works and how it can be more widely adopted.

In the hope of providing practical guidance from leaders in the data journalism world, we will offer model curricula designed to reach a broad swath of educational institutions, from public land-grant universities to private institutions, for both undergraduate programs and graduate-level study, as well as possible concentrations and specialized efforts in graduate programs.

The goal throughout is to help journalism education move toward a more cohesive and thoughtful vision, one that will help to educate journalists who understand and use data as a matter of course—and as a result, produce journalism that may have more authority, yield stories that may not have been told before, and develop new forms of journalistic storytelling.

This vision of bringing data journalism into the mainstream of journalism education has yet one more, broader mission: improving the future of journalism education programs from a research perspective. The practice of data journalism—analyzing, sifting, and telling stories from information—will increase the contribution of journalism schools to the range of data-centered fields emerging across university campuses.

Data journalism "also can be a bridge to other parts of the university," said James T. Hamilton, an economist and director of the journalism program at Stanford University, which in 2015 launched the Stanford Computational Journalism Lab. He pointed to possible collaborations with social scientists as just one example.

The authors and a committee of professors and professional data journalists agree that if journalists and journalism educators want to innovate, then equipping our students with practical data skills and, more importantly, a data frame of mind, is a vital part of the path forward for the students, faculty, and administrators.

CHAPTER 1: DEFINING THE FIELD OF STUDY

WHAT'S IN A NAME

In our view, data journalism as a field encompasses a suite of practices for collecting, analyzing, visualizing, and publishing data for journalistic purposes. This definition may well be debated. The history of data journalism is full of arguments about what it should be called and what it includes.

In fact, data journalism has been evolving ever since CBS used a computer to successfully predict the outcome of the presidential election in 1952. As technology has advanced, so has the ability of journalists to tap that technology and use it for important storytelling.

One key definition of data journalism can be found in a 2014 report by Alexander Howard for the Tow Center for Digital Journalism and Knight Foundation. Data journalism is "gathering, cleaning, organizing, analyzing, visualizing, and publishing data to support the creation of acts of journalism," Howard wrote. "A more succinct definition might be simply the application of data science to journalism, where data science is defined as the study of the extraction of knowledge from data."[2]

But news games, drone journalism, and virtual reality—approaches that some may not consider mainstream data journalism today—may represent a much more dominant presence tomorrow. Or data journalism may evolve in yet another direction, perhaps into common applications for machine learning and algorithms. Data journalists are already working more with unstructured information (text, video, audio) as opposed to the historical elements of data journalism (spreadsheets and databases full of rows and columns of numbers).

2 Howard, "The Art and Science of Data-Driven Journalism," p. 4.

"I think the one good thing about the name discussion is that people are realizing there are different kinds of approaches to data for journalism," said Brant Houston, the Knight Chair in Investigative and Enterprise Reporting at the University of Illinois at Urbana-Champaign and a former executive director of Investigative Reporters and Editors (IRE).

The ever-evolving practice of data journalism has at heart represented what journalists do best—push against the boundaries of what is expected. Editors used to argue that readers wouldn't understand a scatterplot published in the newspaper. Today, the *New York Times*'s Upshot, Fivethirtyeight.com, and others regularly provide informative data graphics and visualizations.

At the same time, each generation of data journalists has informed the next and balanced the desire to try new methods with foundational ethics and transparency.

Our study aims to provide a broad evaluation of many areas of journalistic practice involving data and to identify best practices for teaching these skills and the "data frame of mind" that goes with them. In doing so we looked at multiple forms of data journalism and defined them as best we could to ensure clear communication.

FOUR KEY AREAS OF DATA JOURNALISM

For this report, we will divide data journalism into four categories, acknowledging that overlap is inevitable in practice. Examples of journalism that fall under each of these headings can be found in the appendix.

DATA REPORTING

DEFINITION: Obtaining, cleaning, and analyzing data for use in telling journalistic stories.

INCLUDES:

» Deploying computer-assisted reporting or analysis for writing journalistic stories

» Practicing precision journalism, as introduced by Philip Meyer, including the use of social science research methods in the interest of journalism

» Visualizing data—mapping and charting—for use in exploration and analysis

» Programming to obtain and analyze data for writing journalistic stories

TECHNIQUES AND TECHNOLOGIES:

» Invoking public records law to negotiate for data

» Using web scraping tools and techniques (ranges from tools to knowledge of Python programming language)

» Using relational database software (can range from Microsoft Access to MySQL)

» Understanding statistical concepts and software or programming languages with statistical packages (SPSS or R among others)

» Using mapping and visualization tools and software (Tableau, Esri mapping software, QGIS, Google Fusion)

DATA VISUALIZATION AND INTERACTIVES

DEFINITION: Using code for digital publishing (HTML/CSS/JavaScript/jQuery) as well as programming and database management to build interactive journalistic work. This overlaps with design work, which falls outside of traditional definitions of data journalism. But visualizations and apps also can be integral to the storytelling process.

INCLUDES:

» Visualizations developed and designed as interactive charts and graphics for presentation, including the use of code

» Interactive applications, including searchable databases and games that help readers explore and understand a news story; these applications can be a key part of the utility of a data journalism project

TECHNIQUES AND TECHNOLOGIES:

- The use of code, which is defined as HTML and CSS and also could include JavaScript
- The use of visualization software or programs, ranging from Tableau visualizations to the D3 JavaScript Library
- Database management and programming, including Python, web frameworks such Django, Flask and Ruby on Rails, and more
- Mapping applications, including QGIS, CartoDB, Esri, TileMill, GeoDjango, and more
- Server knowledge and the use of GitHub, versioning, and Agile software development techniques

EMERGING JOURNALISTIC TECHNOLOGIES

DEFINITION: New developments using data and technology.

- Drone Journalism
- Sensor Journalism
- Virtual and Augmented Reality Journalism

DRONE TECHNOLOGIES:

"Drone journalism is generally defined as the use of unmanned aerial systems to gather photos, video and data for news. What separates Drone Journalism from drone photography is the application of journalistic ethics and consideration of the public interest when using [drones]." — Matt Waite, a professor of practice at the University of Nebraska and founder of the Drone Journalism Lab

Drone technologies can include an airframe, defined by configuration (such as fixed wing or multirotor); an autopilot of varying capabilities (full automation, minor stability assistance, return-to-home fail-safe functionality); a control system (manual control through radio signals, automated flight through software and Bluetooth wireless connection); and a sensor (camera, video camera, multispectral camera, other physical sensor).

SENSOR TECHNOLOGIES:

Sensor technologies include a wide range of software and hardware to measure physical conditions like air quality, motion, or noise levels. These can be used to gather data with a small, portable computer or microcontroller. The Raspberry Pi is a low-cost, credit card–sized computer that has a variety of input/output pins for mounting devices like sensors. Similarly, Arduino is an open-source microcontroller platform that is widely used for prototyping with electronic components like sensors. Some universities have already begun teaching sensor journalism with specific project-based classes, such as to test environmental conditions like air and water quality.

VIRTUAL AND AUGMENTED REALITY TECHNOLOGIES:

Virtual reality (VR), long heralded as an emerging digital technology, finally appears poised to enter the broad consumer market. Samsung, Oculus, and

Google have developed consumer VR headsets along with controllers to facilitate interactivity using your hands and feet. From a production standpoint, panoramic images and videos may be stitched together from an array of cameras, while the company Jaunt is developing a standalone camera to capture 3D video in 360-degree, immersive format. Yet questions of narrative, audience interaction, and journalistic values have yet to be settled with these technologies, even as the *New York Times*, Los Angeles Times, and PBS "Frontline" have launched exploratory ventures to use VR. Journalism schools need to not only provide exposure and instruction in this emerging technology, but also to inquire into values and best practices.

COMPUTATIONAL JOURNALISM

DEFINITION: The use of algorithms, machine learning, and other new methods to accomplish journalistic goals. This area overlaps with data reporting and emerging technologies.

INCLUDES:

» Algorithms that help journalists mine unstructured data in new ways

» New digital platforms to better manage documents and data

TECHNOLOGIES:

» Programming languages like Python, Ruby, and R

» Frameworks and applications like Jupyter that enable journalists to mix code and prose as they perform analysis and show the steps in their work

» Platforms like Overview that facilitate the use of complicated computational processes like natural language processing and topic modeling

20 A BRIEF HISTORY OF COMPUTERS AND JOURNALISTS

In 1967, Philip Meyer had just returned to Knight Ridder's Washington Bureau from a Nieman Fellowship at Harvard University, where he had delved into a different area of computational methods: social science. Social science methodologies, including statistical tests and surveys, had recently been used by academics to detail the reasons behind the 1965 Watts riots in Los Angeles. Meyer believed similar methodologies could have great impact in journalism. He wasn't back at work for long when he was able to put that belief into practice.

In July 1967, an early morning raid of an unlicensed bar in Detroit resulted in rioting. Crowds of people ran through the streets, burning, looting, and shooting. Theories abounded as to why the rioting had occurred. Some experts thought it was done by those "on the bottom rung of society" with no money or education. A second theory was that it was caused by transplanted and unassimilated Southerners.

Meyer, on loan to Knight Ridder's *Detroit Free Press*, reached out to friends who were social scientists to devise a survey, cobble together funding, and train interviewers. In the survey, respondents, who were guaranteed anonymity, were asked to assess their own level of participation in the riots. They were also asked to indicate whether they considered rioting a crime, whether they supported fines or jail for the looters, and whether they considered African Americans in Detroit to be better off than those elsewhere.

The survey results contradicted the earlier theories and pointed to a different explanation—that the relative good fortune of many African Americans highlighted more deeply the gap felt by those who were left behind.

The Free Press's coverage of the rioting, including Meyer's "swift and accurate investigation into the underlying causes," won the Pulitzer Prize for Local General Reporting in 1968 and launched a new era in the use of computational methods in the service of journalism. Meyer's seminal book, *Precision Journalism: A Reporter's Introduction to Social Science Methods* was published in 1973 and argued that journalists trained in social science methods would be better equipped for journalistic work and provided guidelines for journalists to understand those methods.[3] "The tools of sampling, computer analysis, and statistical inference increased the traditional power of the reporter without changing the nature of his or her mission," Meyer wrote, "to find the facts, to understand them, and to explain them without wasting time."[4]

That pioneering work by Meyer is commonly thought to be the beginning of what has been termed either precision journalism or computer-assisted reporting. His approach inspired other journalists. Their work in turn inspired a movement and the creation of a training ground. Two academic institutions

3 In later editions, the name changed to *The New Precision Journalism* (2013).

4 Meyer, *Precision Journalism*, p. 3.

in particular, Indiana University and the University of Missouri, supported the development of that training ground.

But in the wider academic world, computational methods applied to reporting largely did not have an impact on other university programs or how journalism was taught. Instead, professional journalists taught other professional journalists the new techniques, and only as those data journalists began to enter academia did data journalism education begin to take a wider hold in that setting.

By the 1980s, as desktop personal computers took the place of typewriters, and editing terminals were used with digital publishing systems, reporters began to use software on PCs to great effect. In 1986, Elliot Jaspin, a reporter at the *Providence Journal-Bulletin*, used databases to match felons and bad driving records to school bus drivers.

In 1988, Bill Dedman, a reporter for the *Atlanta Journal Constitution*, using data from a 9-track tape and with analysis by Dwight Morris and input from the Hubert H. Humphrey School of Public Affairs at the University of Minnesota, showed that banks were redlining African Americans on loans throughout Atlanta, and eventually the country, while providing services in even the poorest white neighborhoods. That series, "The Color of Money," won a Pulitzer Prize in Investigative Reporting.

By 1989, Jaspin launched the Missouri Institute for Computer-Assisted Reporting (MICAR) at the University of Missouri. Soon, he was teaching computer-assisted reporting to students at the university and holding boot camps for professional journalists. Four years later, in 1994, a Freedom Forum grant would help the institute boost its presence and become a part of IRE as NICAR—the National Institute for Computer-Assisted Reporting.

In 1990, at Indiana University, former journalist turned professor James Brown worked with IRE to organize the first computer-assisted reporting conference, sponsored by IRE. He created a fledgling group called the National Institute for Advanced Reporting (NIAR).

"Andy Schneider, a two-time Pulitzer winner, had just joined our faculty as the first Riley Chair professor. One day we were talking about how so few journalists used computers in their reporting," Brown recalled in an email. "In 1990, I don't know of any schools that had such skills integrated into the curriculum. At that time, any undergraduate in even the smallest school of business knew how to use a spreadsheet. We decided to do something about it and that was how NIAR started."

NIAR would host six conferences before deciding to fold to avoid duplicating efforts by IRE and MICAR, Brown said. Still, the Indiana conferences trained more than 1,000 journalists and were a precursor to a new era. In 1993, IRE and MICAR (which later would be renamed to NICAR), held a computer-assisted reporting conference in Raleigh, North Carolina, that drew several hundred attendees. That marked the beginning of an annual event that continues today, where new generations of reporters and editors learn to use spreadsheets or query data and to use maps and statistics to arrive at news-

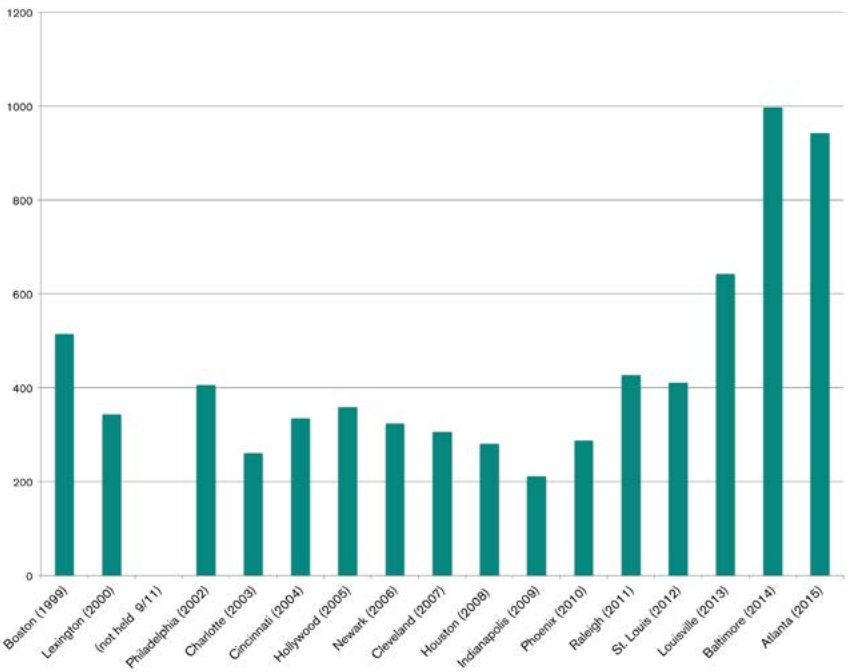
worthy findings.

In 1993, the same year as the Raleigh computer-assisted reporting conference, the *Miami Herald* received the Pulitzer Prize for Public Service after reporter Steve Doig used data analysis and mapping to show that weakened building requirements were the reason Hurricane Andrew had so devastated certain parts of Miami.

Much of this new computer-assisted reporting came about because as the Internet emerged and became more accessible, so too did the concept of using a computer in reporting. But NICAR and the University of Missouri in particular had a broad and deep impact. A good number of the most prominent practitioners of data journalism learned their skills from NICAR and from other journalists trying to solve similar data challenges.

This pattern is perhaps most visible through tracking the careers of the NICAR trainers themselves. Sarah Cohen was part of a *Washington Post* team that received the 2002 Pulitzer Prize in investigative reporting for detailing the District of Columbia's role in the neglect and death of 229 children in protective care, and Jennifer LaFleur has won multiple national awards for the coverage of disability, legal, and open government issues. Both were NICAR trainers.

Another NICAR trainer was Tom McGinty, now a reporter at the *Wall*



NICAR conferences over time. The conference was not held in 2001 because of 9/11.
Source: IRE

Street Journal and the data journalist for "Medicare Unmasked," which received the 2015 Pulitzer Prize in Investigative Reporting. Jo Craven McGinty was also a NICAR trainer and later worked as a database specialist at the *Washington Post* and at the *New York Times*; she now writes a data-centric column for the *Wall Street Journal*. Her analysis about the use of lethal force by Washington

police was part of a Post series that received the Pulitzer Prize for Public Service and the Selden Ring Award for Investigative Reporting in 1999.

Journalist David Donald moved on from his NICAR training role to head data efforts at the Center for Public Integrity and is now data editor at American University's Investigative Reporting Workshop.

Aron Pilhofer was an IRE/NICAR trainer and led IRE's campaign finance information center. He went on to work at the Center for Public Integrity and the *New York Times*, where he founded the paper's first interactives team. Today, Pilhofer is digital executive editor at the *Guardian*.

Justin Mayo, a data journalist at the *Seattle Times*, graduated from the University of Missouri and worked in the NICAR database library and as a NICAR trainer. He has paired with reporters on work that has opened sealed court cases and changed state laws governing logging permits. Mayo was involved in data analysis and reporting on an investigative project on problems with prescription methadone policies in the state of Washington, which received a Pulitzer Prize for Investigative Reporting in 2012 and in covering a mudslide that received a Pulitzer Prize for Breaking News Reporting in 2015.

Clearly, working at NICAR has meant building powerful skills. So, too, has attending conferences and boot camps. The students who attended early NICAR boot camps were "missionaries" who returned to their newsrooms to teach computational journalism skills to their colleagues, Houston recalled. For years, the conferences and boot camps were "the only place where people have had an extensive amount of time to try out new techniques."[5]

By the late 1990s, as the increasing prominence of the Internet led more news organizations to post stories online, journalism education offered even more digitally focused instruction: multimedia, online video skills, and HTML coding, among others.

Two strands, data and digital, represent distinct uses of computers within journalism. Early calls for journalism schools to adapt to changing technological conditions were answered mainly with the addition of digital classes—learning how to build a web page, create multimedia, and curate content.

Many of the early digitally focused journalism instructors faced a battle in trying to introduce new concepts into print journalism traditions. Data journalism instructors—focusing more on data analysis for use in stories—have faced similar challenges.

Meanwhile, by the 1990s, a few universities had begun teaching data analysis for storytelling. Meyer, who in 1981 became Knight Chair at the University of North Carolina, was teaching statistical analysis as a reporting method. Indiana University, with Brown, the professor who launched the first

5 For a more complete look at the long and storied history of computer-assisted reporting, the spring/summer 2015 edition of the IRE Journal provides a detailed and engaging recounting by Jennifer LaFleur, NICAR's first training director in 1994 and now the senior data editor at the Center for Investigative Reporting/Reveal. Brant Houston details that history in "Fifty Years of Journalism and Data: A Brief History," Global Investigative Journalism Network, November 12, 2015.

CAR conference, began incorporating the methods into classes. And Missouri offered computer-assisted reporting instruction, thanks to Jaspin; Brant Houston, an early NICAR director who later became IRE's executive director; and others. Other universities began to introduce basic classes or incorporate spreadsheets into existing classes.

Houston's *Computer-Assisted Reporting: A Practical Guide* became one of the few foundational texts available on the subject. His book, now in its fourth edition, lays out the basics of computer-assisted reporting: working with spreadsheets and database managers as well as finding data that can be used for journalism, such as local budgets and bridge inspection information. What Houston detailed in that first edition became essentially a core curriculum for data journalism from 1995 through the present day. Houston's work codified the principles and practices of computer-assisted reporting from the perspective of its burgeoning community.

But throughout those two decades, journalists still learned these skills primarily through the NICAR conferences or from other journalists. For many years, for example, Meyer and Cohen taught a NICAR stats and maps boot camp at the University of North Carolina geared toward teaching professional journalists.

Since then, boot camps have become a popular model, used by universities and other journalism training organizations, often in coordination with IRE/NICAR. A key tenet of the boot camp is practical, hands-on training, using data sets that journalists routinely report on, such as school test scores. To sum up this model, Houston said it's all about "learning by doing."

Many boot camp graduates have gone on to robust data journalism careers and have also moved into teaching in journalism programs, both as adjuncts and full-time faculty, where they have integrated those teaching techniques into their classes. These journalists essentially took the curriculum from NICAR and introduced it into the wider academic world.

In 1996, Arizona State University lured Doig from the *Miami Herald* to the academic life where he has been teaching data journalism ever since, serving as the Knight Chair in Journalism and specializing in data journalism. The stats and maps boot camp eventually migrated to ASU as well.

As journalism programs began to offer these classes, they focused on the basics covered in Houston's book: negotiating for data, cleaning it, and using spreadsheets and relational databases, mapping, and statistics to find stories.

In 2005, ASU benefited from a push by the Carnegie Corporation of New York and the John S. and James L. Knight Foundation to revamp journalism education. The school expanded its focus on all things data and multimedia with the founding of News21. That program has focused heavily on using data to tell important and far-reaching stories while teaching hundreds of students journalism at the same time.

At Columbia, the first course on computer-assisted reporting was offered in 2003, when Tom Torok, then data editor at the *New York Times*, taught a one-credit elective. With the founding of the Stabile Center for Investigative

Journalism in 2006, some data-driven reporting methods were integrated into the coursework for the small group of students selected for the program. The number of offerings in data and computation at Columbia has risen steadily since the founding of the Tow Center for Digital Journalism in 2010 and the Brown Institute for Media Innovation in 2012. In addition to research and technology development projects, these centers brought full-time faculty and fellows to teach data and computation, as well as supplied grants to support the creation of new journalistic platforms and modes of storytelling.

Columbia has also launched several new programs in recent years that situate data and computational skills within journalistic practice. One is a dual-degree program in which students simultaneously pursue M.S. degrees in both Journalism and Computer Science — and those students must be admitted to both programs independently. In 2014, the Columbia Journalism School established a second data program, The Lede, in part to aid students in developing the broad skillset they would need to be a competitive applicant to both Journalism and CS. The Lede is a non-degree program that provides an intensive introduction to data and computation over the course of one or two semesters. Most students arrive with little or no experience with programming or data analysis, but after three to six months they emerge with a working knowledge of how databases, algorithms, and visualization can be put to narrative use. Post Lede, many students are competitive applicants for the dual degree, but others go directly into the field as reporters.

The emergence of these initiatives in journalism schools reflects the extent to which data-driven reporting practices have broadened in the last decade. In the 2000s, journalists began to move well beyond CAR, trying out advanced statistical analysis techniques, crowdsourcing in ways that ensured data accuracy and verification, web scraping, programming, and app development.

In 2009, IRE began working to attract programmers and journalists specializing in data visualization, said executive director Mark Horvit. It always offered hands-on sessions in analyzing data, mapping, and statistical methods. Added to that now are sessions on web scraping, multiple programming languages, web frameworks, and data visualization, among other topics. The sessions have even included drone demonstrations. The challenge has become balancing the panels so that there is enough of each type of data journalism. As a result, the annual conferences have grown tremendously, from around 400 at the CAR conference each year in the early 2000s to between 900 and 1,000 attendees today.

Other groups began addressing data journalism as well as pushing for new methods of digital journalism. The Society of Professional Journalists wanted to teach its members about data and joined with IRE to do so, sponsoring regional two- or three-day Better Watchdog Workshops. Minority journalism associations began to provide data journalism training, often in collaboration with IRE or its members or under the Better Watchdog theme.

The Online News Association's annual conference focuses on the larger world of digital journalism. Many of its panels feature coding for presentation,

cutting-edge developments in digital web-based products, audience development, and mobile. It also offers panels on data journalism and programming.

Still, a gap has persisted. At times, new organizations formed to fill some of the needs. In 2009, Pilhofer, then at the *New York Times*, Rich Gordon from Northwestern University, and Associated Press correspondent Burt Herman, who was just finishing a Knight Fellowship at Stanford, created a loosely knit organization that brings together journalists and technologists, hence the name Hacks/Hackers. Its mission is to create a network of people who "rethink the future of news and information." Even as some groups have tried to fill gaps in data journalism instruction, what exactly counts as data journalism remains a rough boundary, with few distinctions between data journalism and digital/web skills. In this paper, we continue to sharpen the focus on what will improve the level of data journalism education, not overall digital instruction.

In 2013, a group of journalists used Kickstarter to raise $34,000 and create ForJournalism.com, a teaching platform to provide tutorials on spreadsheets, scraping, building apps, and visualizations. Founder Dave Stanton said the group wanted to focus on teaching programmatic journalism concepts and skills and offer subjects that weren't being taught. "You didn't really even have these online code school things," he said. "There were a few. The problem was there was no context for journalism."

A PATH TO DIVERSITY IN DATA JOURNALISM

Journalism training organizations and the journalists in NICAR have grappled with the challenges found in tech elsewhere of low minority and female participation in the industry. Most of the practitioners of data journalism were, and are, white men. IRE has worked with minority journalism organizations to provide training in data journalism. The conference planners also worked to include diversity on its panels and in its training sessions.

In the mid-1990s, women at the NICAR conferences began a tradition of heading out to dinner together. In 2000, at the Lexington conference, fewer than a dozen women attended. Over time, the number of women grew so large that the dinner couldn't be held because finding a restaurant was difficult. Then in 2011, in St. Louis, the dinner was reprised by three of the original members. In 2015, the dinner drew 100 women. It likely would have drawn more but was capped for space limitations. The event goes a long way toward emphasizing that women are welcomed and celebrated at this intersection of technology and journalism.

With data coursework lacking in so many schools, the strongest presence of data journalism in most of academia has been the study of changing newsrooms by sociologists and communication scholars. Their work aims to document and explain data practices within ongoing scholarly conversations about media, technology, information, and society.

Elsewhere in academia, narrative uses of data and computation have emerged independently. Besides the work of quantitative social scientists, like those who inspired the work of Meyer, significant movements in the arts and humanities treat data either as a novel inroad to their traditional objectives or as a means to reinterpret those objectives. Probably the broadest of these movements falls under the heading of the "digital humanities." One of its leading figures, Franco Moretti of Stanford University's English department, has developed methods of "distant reading" by which one asks questions of a set of books larger than any one person could read in a lifetime. Dennis Tenen, a professor in Columbia University's English and Comparative Literature department who has also taught at the Journalism School, identifies himself as a practitioner of computational cultural studies and argues that most disciplines have by now developed computational methods that have either complemented or supplanted their earlier practices.[6]

Several universities have founded centers and institutes devoted to work at the nexus of data, computation, and humanistic endeavors. The University of Illinois, Urbana-Champaign, for instance, hosts the Institute for Computing in Humanities, Arts, and Social Sciences, or I-CHASS, a partnership between the university and the National Center for Supercomputing Applications. The institute helps develop partnerships among social scientists and computing experts, engineers, data scientists, and computer scientists. Their collaborations have included work on large-scale video analysis, research into climate change, and even digitizing and analyzing the papers of Abraham Lincoln.

The uses of data and computation in architecture, geography, and economics also reflect the manner in which these disciplines adopted new tools and methods in recent decades. In journalism, our history is not so different. Like data journalism, computational work in the humanities and social sciences is growing, and this is reflected in the relatively healthy academic job market for digital humanists compared with the job market for traditional scholars.

Overall, we see data science and computational methods being introduced into disciplines across universities that, like journalism, have not been particularly quantitative in the past. Practices involving the use of data and computational methods may be bundled into entirely new departments,

6 Franco Moretti, Graphs, Maps, Trees: Abstract Models for Literary History (London: Verso, 2007) and Dennis Tenen, "Blunt Instrumentalism," in Debates in the Digital Humanities, forthcoming in 2016, University of Minnesota Press.

centers, research institutes, and degree programs (such as data science and computational media). It is not the purpose of a program in data journalism to compete with these other disciplines, but to develop a curriculum that is intrinsically journalistic—one that reflects a mission to find and tell stories in the public interest—as well as develop partnerships and collaborations with other disciplines.

One example of unexpected interdepartmental collaboration at Columbia has been with the Earth Institute, which has curated a massive database of climate data and offers courses in Python programming in which several Journalism students have enrolled. This course focuses on large time-series data sets, which enables data journalists to put the climate into context in their stories.

In 2013, Jean Folkerts, John Maxwell Hamilton, and Nicholas Lemann—all journalism school deans and two of the three of them longtime professional journalists—published "Educating Journalists: A New Plea for the University Tradition." The paper focused on "universities' role in journalism as a profession" but it also discussed how this transformation in journalism could be a boon for the schools that educate journalists. The authors wrote:

> That journalism is going through profound changes does not vitiate—in fact, it enhances—the importance of journalism schools' becoming more fully participant in the university project. Done properly, that will produce many benefits for the profession at a critical time. Journalism schools should be oriented toward the future of the profession as well as the present, and they should not be content merely to train their students in prevailing entry-level newsroom practices.[7]

Key among their recommendations was this: "We see all three of these early strains in journalism education—practice-oriented, subject matter-oriented, and research-oriented—as essential. And all of them can and should be applied, with potentially rich results, to the digital revolution. Journalism schools should embrace all three, not choose one and reject the others."[8]

Journalism programs, with their ability to communicate to a general audience and their potential to analyze and visualize data for story, are a perfect partner for other departments. For example, at Stanford's new Computational Journalism Lab (co-founded by one of this report's authors), faculty are working on several projects with professors from other academic disciplines whose research mission touches on the same data. One goal is that data sets can be collected, analyzed, and used in academic research as well as for journalistic storytelling. In some instances, new methods of analysis can be developed in concert with important public accountability journalism projects.

7 Folkerts, Hamilton, and Lemann, "Educating Journalists," p. 4.

8 Ibid., p. 12.

Talk to deans of journalism schools today and you will hear the same refrain and the belief that data journalism, while not a savior, is an increasingly important component of how journalism education can evolve.

Steve Coll, the dean of the Columbia Graduate School of Journalism, describes the emergence of instruction in data-driven reporting practices as a recognition that data journalism is about more than just publishing stories through digital media, but about developing reporting methods appropriate to the complexity of the world today.

"Data journalism and tools like sensors look powerful because, in comparison to the way journalism schools have responded to previous iterations of technological change, this one runs deep, and to the heart of professional practice. It's not about shifting distribution channels, or shifting structures of audience," Coll said. "It was very tempting, in many ways necessary, for journalism schools to rush over to the teaching of tools, the teaching of platforms, the teaching of changing audience structure. But that transformation often had little to do with the core, enduring purpose of journalism, which is to discover, illuminate, hold power to account, explain, illustrate."

Journalism schools, by necessity, adapted many new tools to respond to the massive and rapid shift to digital media. But delving into data journalism brings journalism back to its journalistic mission and moves it ahead in its research mission at the same time, Coll said.

"What we're really seeing now is that this is a durable change in the structure of information, and therefore a need to durably change a journalist's knowledge in order to carry out their core democratic function. Not to build a business model, not to reach more people, not to have more followers, but to actually discover the truth—you need to learn this."

The rise of data analysis may also foster cross-campus collaboration. Journalism schools, as they embrace data analysis within their already powerful ability to tell stories, are uniquely suited to be robust participants and even leaders in developing means of storytelling with data.

Our research, which is focused on journalism schools, may not account for programs where data analysis is centered in another school or department that teaches this subject to students throughout the university. For undergraduates, in particular, there is little reason to offer in-house classes in subjects that students have free rein to study in another department. Yet it would require a great deal of latitude and initiative for students to construct hybrid degrees this way. Journalism students can sometimes be better served by cross-departmental initiatives that pair instructors for team teaching and connect journalism students with other disciplines that focus on data and computation. Northwestern, Stanford, Boston University, Columbia, Georgia Tech, Syracuse, and others have worked to build these interdisciplinary initiatives.
By establishing these interdepartmental bridges, schools can create pathways of collaboration between journalism, its partner disciplines of communication and media studies, and the other areas of research that share an interest in the future of technology and society.

Even as cross-departmental work increases, another challenge for journalism education will be to identify which data courses need to be framed journalistically and which others can be learned through classes framed within the methodology of other departments. In order to learn statistics, for example, students may be encouraged to register in classes offered by the math, statistics, or even political science department. The principles and objectives of these classes could apply within journalistic work, but that may not always be the case. These classes are often taught from a research or theoretical perspective. A statistics class that emphasizes survey methodology, for example, could be less useful for a journalism student.

Journalists do not often work with samples, but they do work with entire data sets. For data journalism education in particular, a more useful statistics class might be the type of instruction Meyer provided both in college courses and in IRE/NICAR boot camps, using social science to address journalistic challenges. Accommodating both techniques in a research or statistics class could foster collaboration instead of silos. In other instances, outsourcing a course may make sense. Mapping skills necessary for journalists, for example, are the same types of skills necessary for other disciplines in academia.

Yet the task of developing and adopting a data journalism curriculum comes with its own challenges. The high rate of change in digital tools, platforms, and programming languages means that there is more to teach and that classes themselves must be updated frequently. It is difficult to decipher which new techniques are just passing fads and which have the potential to remain relevant for even ten years. For this reason, it is important for classes to be designed so that they teach data and computation as fundamental styles of inquiry. Students can learn enough about the concepts behind a technique to be able to more easily learn new tools that address the technique—as opposed to focusing on the discrete tools used from time to time.

There are exceptions—the Unix command line, for example, has been as fundamental and immutable as any computing tool. This is a text-based application, still favored by developers for many tasks on Mac and Linux systems, for controlling the computer using typed commands instead of a graphical interface. And many of its core utilities remain essentially unchanged since the 1970s. Yet it is far more common to cite such examples as the ActionScript language for Adobe Flash, which was taught at several journalism schools less than a decade ago and is all but abandoned by developers today. The silver lining is that ActionScript shares many features with programming languages such as JavaScript and Python, so it may have offered a path for a student to develop other proficiencies. But it also highlights the importance of selecting techniques for journalism classes with long-term considerations in mind.

CHAPTER 2: STATE OF THE FIELD: OUR QUANTITATIVE DATA

THE SCOPE OF OUR STUDY

For this report, we collected and analyzed information on 113 journalism schools, roughly one-quarter of the nation's journalism programs, and gathered 63 syllabi for courses on topics spanning data-driven journalism, computational journalism, data visualization, and other methods. We combined that with a series of in-depth interviews with more than 50 professors and professional journalists (many of whom are adjuncts), and we spoke with ten students or recent graduates. We also attended nine classes and participated in three massive open online courses (MOOCs).

For years, anecdotal evidence has indicated that U.S. journalism schools have fallen behind in data instruction, or rather, started from behind and have not caught up with the field as it has been practiced in newsrooms. A key tenet of this field is that using data to report and tell stories can result in a more powerful story. As LaFleur described it in her IRE article: "understand the data, interview the data, report the data." That is the process we tried to follow for this report.

We first collected the course offerings of 113 programs accredited (fully or provisionally) by the Accrediting Council on Education in Journalism and Mass Communications. Accreditation is a voluntary process for journalism schools. We used the ACEJMC programs simply because they represented a significant portion of journalism schools and their curriculum requirements include two that fit in with the concept of providing data journalism instruction: "apply basic numerical and statistical concepts" and "apply current tools

and technologies appropriate for the communications professions in which they work, and to understand the digital world."

We scraped what we could from the journalism program websites and hand-entered the remainder. To verify the data, we then emailed or called programs that had listed either no classes in data journalism or very few classes. This yielded changes in our numbers for several programs where the online course descriptions were not accurate. In soliciting this feedback, we also heard from 11 schools where the department is revamping its curriculum and considering adding data journalism. Sixteen schools did not respond to multiple emails or phone calls. We then revisited every program website for all 113 programs and double-checked the data.[1]

We also collected information on multimedia offerings of each program so that we could compare multimedia course offerings with data journalism course offerings.

OUR FINDINGS

A little more than half of the universities we reviewed—59 of the 113 schools—offer one or more data journalism courses. We defined a data journalism class as being focused on the intersection of data and journalism, and using spreadsheets, statistical software, relational databases, or programming toward that end. We included in the data journalism category only those programming classes that went beyond basic HTML and CSS. For the purposes of this report, we considered classes on HTML, CSS, and JavaScript to be focused on digital/design journalism, not data journalism. We also excluded courses in numeracy and communications research methodologies and statistics unless the course offerings explicitly included a journalism focus. The appendix includes tables detailing the full results of our analysis.

For Aaron Williams, who is four years out of college, it was not surprising to hear that our analysis showed 54 of the 113 programs don't offer a standalone class on data journalism. Williams has worked in data journalism at the Los Angeles Times, the Center for Investigative Reporting, and now as interactive editor at the San Francisco Chronicle. Almost everything he knows he learned from colleagues at NICAR, he said. "I didn't even really know about data journalism as a discipline, nor did my instructors . . . until basically I was a senior," Williams recalled.

Of the 59 programs we identified that teach at least one data journalism class, 27 of the schools offer just one course, usually foundational. Fourteen offer two classes. Just 18 of the 59 schools teaching data journalism offer three or more classes in this subject.

At a minimum, these programs offer courses that teach students to use

1 It should be noted that information on a degree program's website does not necessarily reflect the present state of their curriculum. We reached out to professors and administrative staff in order to confirm our data, but this was not always possible.

spreadsheets to analyze data for journalistic purposes. At the other end of the spectrum, some schools provide far more, teaching multiple classes in programming skills, such as scraping the Web, building news apps, or creating advanced data visualizations. But programs with multiple classes are rare.

A significant number of programs offer some instruction in data journalism, even if they don't provide a standalone class. Of the 113 ACEJMC-accredited programs, 69 integrate some data journalism into other reporting and writing courses, our analysis showed. In most cases, this entails introducing the concepts of using spreadsheets or basic analysis as part of reporting and writing classes or certain topic classes, such as business journalism.

Again, tables summarizing these findings can be found in the appendix, while the remainder of this chapter will dig deeper into our analysis of syllabi and course offerings in data journalism.

TEACHING DATA FUNDAMENTALS: ROWS AND COLUMNS

Data journalism professors say that the foundational data class is the most important because it lays down key mindsets and skills that are a prerequisite for more advanced learning. Steve Doig of ASU believes the core data syllabus should consist of negotiating for data, thinking critically about data, and using spreadsheets to analyze data.

It is difficult to overstate the value of spreadsheets for managing information. When we asked former CUNY professor Amanda Hickman, now an Open Lab senior fellow at BuzzFeed, how she defines data, she replied, "anything tabular."

For the foundational computer-assisted reporting classes, the syllabus analysis and interviews indicate that the coursework is comprehensive, providing a strong base in critical thinking and basic concepts surrounding the use of data to find and tell stories. Students are taught similar concepts: critical thinking and developing a "data frame of mind"—in other words, being able to question data in a disciplined way, make sense of discrepancies, and find the underlying patterns and outliers that are important to the analysis.

Most of the classes include some type of hands-on learning. Many of them focus first on spreadsheets, then SQL, followed by mapping and statistical concepts. Others include basic data visualization, using Tableau or Google Fusion as a way into the subject. Multiple professors said the hands-on approach reinforces the critical thinking concepts, including helping students to understand what structured data look like and how information of any kind can be structured for better understanding.

Another key feature of the 63 syllabi we reviewed was an exercise in requesting and negotiating for data from a governmental body. Dan Keating, who works at the *Washington Post* and teaches a long-standing class in computer-assisted reporting at the University of Maryland, said that finding what "no one has ever known before" is a defining part of his class.

Many CAR courses break down this way:

Hard Skills

- Searching for and finding documents and data that enable the journalist to make statements of fact, including public requests, deep research, and scraping skills
- Understanding data structures and how to clean and standardize data into a form that is useful
- Analyzing data using spreadsheets, databases, mapping, and visualization
- Learning advanced statistical methods that illuminate data

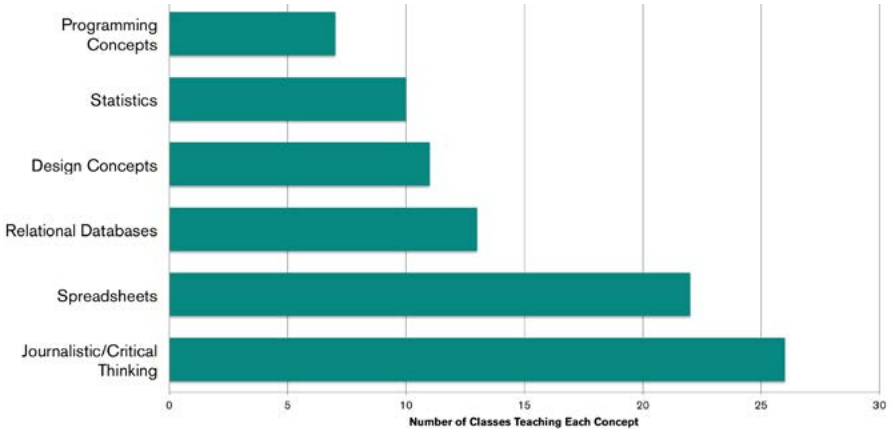
Guiding concepts

» Finding what "no one has known before"

» Developing data-driven storytelling techniques, including how to use numbers effectively in prose and how to tell a story visually

» Thinking of data as an asset in the reporting process

Whether following the guiding concepts or applying the hard skills, journalism students today must be well grounded in both the importance of data and the tools to use data in storytelling. "If you don't deal with data as a journalist, you're shutting yourself down," said McGinty of the *Wall Street Journal*.



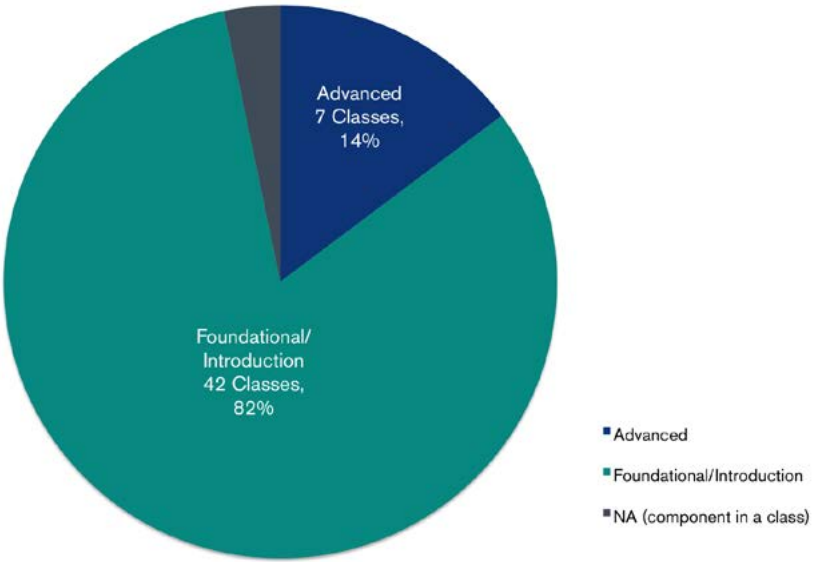
Analysis from our collection of syllabi.

Key concepts overlap, so multiple concepts can be taught in one class.

36 TEACHING ADVANCED DATA SKILLS: VISUALIZATION AND PROGRAMMING

Advanced instruction in data journalism today is limited. Only 14 of the 113 AEJMC-accredited programs surveyed for this study teach programming beyond HTML/CSS to journalism students. And only 11 of the 113 offer coursework in emerging areas of data journalism, such as drones, virtual reality, and computational methods.

In fact, based on the analysis of syllabi and journalism programs, even some classes described as advanced primarily teach basic tenets of spreadsheet use. Part of the reason is that this is still where the need is greatest, said professors and trainers. "It is unbelievable how much time I spend teaching the basics," said Jaimi Dowdell, the senior training director for IRE.



Analysis from our collection of syllabi. Advanced classes included data visualization, programming languages, and other emerging methods such as machine learning.

Foundational classes included spreadsheets, basic relational database understanding, and descriptive statistics.

However, teaching the basic CAR curriculum is not enough, argued Kevin Quealy, a graphics editor at the *New York Times* and adjunct professor of journalism at New York University. "To do data work at a high level, one or two semesters of courses is very inadequate," he said.

Many journalism programs offer design classes, but often those classes focus on basic design tenets, overall web design, or static infographics. Teaching students the concepts and skills needed to visualize data in an interactive way or to build a web application is more rare.

Not all data journalism educators are convinced that data visualization for news presentation should even be considered part of a data journalism curriculum. However, most agree that it is vital to teach visualization for the purpose of analysis. Alberto Cairo, who is leading an effort to fill a data visualization gap in his role as Knight Chair in Visual Journalism at Miami University, believes that even basic visualization instruction goes a long way toward literacy.

First, data journalists need to know how to do basic exploratory visual analysis, Cairo said. And second, even journalists who practice data visualization need to start with the exploratory analysis. They need to know—just like the CAR specialists—how to "interview" the data, he said.

One challenge for traditional journalism schools, which may lack a strong journalism design component and may already have difficulty teaching a CAR or data analysis class, is whether they should tap professionals or recruit or train faculty to incorporate data visualization. To that, Cairo and other academics and professionals we interviewed suggest that such schools collaborate with other parts of a university to fill the gap.

For our analysis, we differentiated between web and digital technologies aimed at presentation and the data skills needed to tell a story. This can be a difficult boundary line. News applications, for example, are focused on design, but, based on our interviews, there is a key difference in building a new website or a multimedia presentation and building something like ProPublica's "Dollars for Docs," which enabled readers to drill into the story of pharmaceutical industry payments to doctors and also made it possible for other journalists to find and tell other stories. Meanwhile, "Snow Fall," the New York Times's much-touted (and Pulitzer Prize-winning) interactive story of skiers caught in a Washington state avalanche, wasn't about data and it wasn't about furthering the use of the data; it used design skills to make the story an immersive multimedia experience for the reader.

Just 14 journalism schools in our data set teach programming beyond HTML and CSS, based on their course descriptions. At present, the programming languages most often used in classes on data-driven reporting are SQL, Python, and R. Instructors focusing on data analysis often incorporate SQL, and some will introduce R. Some instructors also teach web frameworks, such as Django and Ruby on Rails, and some visualization professors teach JavaScript and other skills, though fewer go into the D3 library developed by Mike Bostock, a former New York Times graphic editor.

Deen Freelon, a communication studies professor at American University, takes a different approach, teaching "code for the purposes of analysis" in a course open to both communications and journalism students. "I just got back from my last class where I was teaching students how to analyze Twitter data," he said.

While advanced classes are rare, there is a clear demand for this knowledge. In the tech world, short programs designed to train web developers have emerged as financially viable businesses. These code schools have shown that some of these skills can be taught in considerably less time than a four-year degree. The Lede Program at Columbia, which offers a summer boot camp as well as an intensive two-semester certification program in computational skills for journalists, has drawn students interested in gaining key data skills in a short period of time.

Maggie Mulvihill, a clinical professor of journalism at Boston University, is raising revenue for computational journalism efforts there through holding

week-long camps on storytelling with data for non-journalism professionals.

Integrating data journalism exposes students to the field, highlighting this as an area that they might choose to practice, but it is also an important step for students developing a foundation of journalistic skills. As noted, 69 of the 113 AEJMC-accredited programs already integrate some data journalism into reporting and writing courses, and on this front there is some good news: several schools expressed interest in adding data journalism in a systematic way to their programs. When, in order to verify or data, we contacted each of the programs that had listed either no data journalism class or just one, 11 responded that they are actively working to add data journalism to their curricula. At the University of Alabama in Tuscaloosa, for example, the school does not offer a standalone CAR class, but it now includes components of data analysis instruction in three separate journalism classes.

ALTERNATIVE DATA JOURNALISM INSTRUCTION: THE STATE OF ONLINE COURSES

One response to the widespread lack of instruction in data journalism, and instructors capable of teaching it, has been to enlist respected teachers for massive open online courses, or MOOCs. Doig is one of those teachers, and he suggests MOOCs offer great benefit for certain classes, providing expert instruction and hands-on training.

He was an instructor in two MOOCs focused on data journalism, one organized by the European Journalism Centre, which drew 25,000 people to enroll, and the other by Rosental Alves of the Knight Center for Journalism in The Americas at the University of Texas School of Journalism, which drew more than 4,000.

"One strength is that there are a wide variety of MOOCs out there created by top faculty at major institutions like Harvard and MIT and Stanford," Doig wrote in a memo on the subject. "Their existence begs the question of why should your institution go to the trouble of creating and staffing a class that covers the same ground. (Of course, one reason would be to collect the tuition from your students!)"

He suggested that a partial journalism curriculum could be crafted out of MOOC offerings combined with video content from journalism-related sources such as IRE and the Poynter Institute's News University. However, being unable to provide individual feedback, MOOCs would come up short for classes in newswriting or basic reporting, he said.

Our research assistant participated in three MOOCs to help us develop a sense of how well the virtual courses teach data journalism. He found that MOOCs are best at offering introductory exposure, but one should not expect to reach in-depth knowledge. MOOCs may be useful for developing an initial foundation in a subject, or for reinforcing a fading proficiency, but may be

lacking in terms of teaching reporting techniques, critical thinking, or creative skills. The three MOOCs he participated in were effective at teaching tools, and our RA reported that he was often excited to learn a feature or technique within an application. However, finding ways to apply these tools outside of exercises may require person-to-person interaction in a classroom setting.

In order for MOOCs to be viable resources, they must be maintained. Many classes referenced lost and outdated information. Broken links, missing materials, and redesigned websites often made it difficult to navigate through the lessons.

Our participant's experience pointed to the issues raised by Doig, but the ASU professor does think MOOCs could still be an optional resource for a data journalism course. "Students eager to go beyond what is offered in the classroom (alas, almost certainly a minority) can be pointed to online sources that will give them that content," Doig wrote. "To that end, it might be a good idea to develop a list of MOOCs and such that journalism instructors could sample and offer to their students."

TEXTBOOKS: LITTLE CONSENSUS

Our analysis also found one more gap in curricula—a strong core of textbooks. The concepts and skills of this field were described in fairly consistent ways throughout our interviews and the text of the syllabi, but data journalism instructors share only a few core books in common. In fact, most didn't use a textbook at all but provided a list of selected readings.

Of the syllabi, more than 70 different textbooks were required, but there was no consensus on which books were preferred. The most popular book—Brant Houston's *Computer-Assisted Reporting: A Practical Guide*—was required in just 14 percent of the classes.

Five courses required membership in IRE, and 23 of the courses required students to buy a book published through IRE. They included various editions of Houston's and IRE's *The Investigative Reporter's Handbook: A Guide to Documents, Databases, and Techniques* and Sarah Cohen's *Numbers in the Newsroom: Using Math and Statistics in News.* Various editions of Philip Meyer's book, *New Precision Journalism* or *Precision Journalism*, were required in nine of the courses.

Eight of the courses required *The Data Journalism Handbook*, which was produced as a combined effort by data journalists around the globe. The online book is an initiative of the European Journalism Centre and the Open Knowledge Foundation and is available free on the Web in English, Russian, Spanish, French, and Georgian.

In 17 classes, no text was required. The lesson here may be that online reading works best for these classes. But it also could mean that despite its long history, data journalism is still a nascent subject within journalism schools and there may be a dearth of effective textbooks beyond the few that are commonly assigned.

CHAPTER 3: QUALITATIVE FINDINGS: INTERVIEWS AND OBSERVATIONS

IDENTIFYING WHAT TO TEACH

> "Data journalism isn't easy to define or to teach. It is constantly changing and best practices are evolving. One needs to learn a lot by doing, too."
> – Jonah Newman of the *Chicago Reporter*

Our interviews echoed many of the findings in our quantitative data, so rather than repeat those findings, this chapter focuses on how the professors put the concepts into practice in the classroom. It is intended to provide a roadmap of existing pedagogical work in data journalism and offer insights into common challenges.

We interviewed nearly 50 teachers and practitioners, and while there is a diversity of thought, there also is a consensus when it comes to the foundations of data journalism curricula: critical thinking, mastery of key data skills, and teaching programming concepts so that students will be able to learn new tools as needed.

David Boardman, dean of Temple University's School of Media and Communication, suggested that data journalism is about learning higher and more complex levels of analysis. This includes learning more sophisticated tools and software and almost certainly some level of programming.

In a data journalism class, having that critical thinking skill means that the students learn to treat data in an ethical way, so that rather than bending the data to represent a particular view, the goal is toward truth and accuracy.

"I would always err on the [teaching of] critical thinking skills," said LaFleur of the Center for Investigative Reporting. "That is the harder skill to ingrain in people. You can learn how to click things and write a line of code."

In general, those who teach data journalism focus on hands-on methods. In the beginning, the professors will provide data to students to analyze. LaFleur, for example, uses hands-on training with one data set and then will introduce a similar data set and assign the students to ask the same types of questions, but on their own.

By the middle of a course, students often have to obtain their own data, submitting public records requests. The students then move on to data that require more complex analysis. By doing this, the professors are doing two key things: teaching the critical thinking that goes with negotiating for information and understanding the bounds of that information. At the same time, the students are using basic tools to accomplish their goals, be they spreadsheets or a relational database. In some classes, the focus is on writing a memo by the end of the course on a possible story. In other courses, the professors expect the students to report and write a story. This last step—either a memo or a story—once again helps the student use critical thinking, this time pairing that with the skills of storytelling.

The key is learning how to obtain mastery, said Ira Chinoy, an associate professor at the University of Maryland who previously led the data journalism efforts at the *Washington Post*.

Chinoy relates this to the 2009 "Miracle on the Hudson" and how the pilot used reflexive mastery to land the US Airways plane on the river after bird strikes caused both engines to fail. In class, when students get discouraged about bad interactions or conversations in pursuit of their databases, Chinoy brings up "Sully" Sullenberger's actions and says, "Do you think he could have done that on his first day of pilot school?"

Chinoy emphasized that the information should not always be presented to the students up front. He likes to give them a chance to come up against obstacles. They also need to develop a sense of when data could be problematic, what are signs of that, and what is each student's best practice for examining the data.

Whether data journalists need to program remains an active debate. But when we delved into this issue, we found that we first need to define what we mean in terms of data journalism. To some, "code" means web development and design—back to the concept of HTML and CSS. "Programming" means writing programs that enable advanced mining of data or algorithms that could identify patterns.

The bottom line is that to do more advanced data journalism, its practitioners need, at a minimum, to understand how programming works. This could be considered the start of computational thinking. Just scraping information from the Web can involve simple programming using Python, and understanding what is possible with programmatic solutions is critical for journalists looking at websites and other troves of information, much of which is not just in rows and columns.

As students develop the ability to recognize computational solutions to some of these problems, some of them may then learn how to program. But even those who don't take the coding path should still be able to understand how solutions like these can be a part of their journalistic practice. The ability to work with data and think in terms of computation is a skill broader and more necessary than any specific tool or programming language. It is vital that we don't confuse the two.

Mark Hansen, a professor of journalism at Columbia and director of its Brown Institute for Media Innovation, also focuses on teaching both programming skills and the mindset of computational journalism. The idea, according to Hansen, is that by using programming, journalists can think beyond rows and columns as they search for answers in data of all forms, whether structured or unstructured.

Nicholas Diakopoulos, a computer scientist at the University of Maryland, has been teaching a number of classes in data journalism beyond the introductory level. And he also provides a course on coding in the sophomore year. His aim, he said, is to move the undergraduate students from the track of learning CSS/HTML basic web skills to understanding web development and news apps. Beyond that, he's offering a class on computational journalism with a focus on Python, text analysis and aggregation, recommender systems, and writing stories with code behind it.

Diakopoulos suggested that students could also take computer science classes if they want to learn how to be hackers—meaning, in the original sense of the word, anyone fluent enough with computers to use them creatively.

It's all about working with data in a principled way, Diakopoulos said. He ties this to CAR and Philip Meyer's crusade to bring the scientific way of thinking into journalism 50 years earlier; in other words, thinking methodically, thinking about how to frame an experiment, gather data, and use rigorous methods to build evidence of some finding of journalistic importance.

In the end, data journalism is about teaching how to find the story, using an increasing array of data techniques, said David Donald, data journalist in residence at the School of Communication at American University and data director of AU's Investigative Reporting Workshop. "You're still talking about story and how data needs to be vetted and be expressed in a way that gets into the public's brain easily," Donald said. "From the investigative side, you are looking for evidence in the data."

Developing that computational ability will become even more important to handle the vast amounts of data in today's world. More tools will come and go, but data journalism, at its core, will enable journalists to do their job in a more expansive way, said Coll of Columbia.

"I think that [data journalism] will be around for a while" Coll said. "It will be around for a whole set of iterations of platforms and distribution systems, and even media. So we get virtual reality, or we get 3D, or we don't. That's a whole different set of questions. This is going to be about how you report on government, how you report on corporations, how you tell wheat from chaff."

Data journalists are starting to address this type of coverage, but it takes a deeper level of data journalism capability—the computational journalism slice of data journalism. Some of it involves presenting data in new, journalistic ways. The "Surgeon Scorecard," published by ProPublica, is one example. ProPublica used extensive Medicare data and collaborated with leaders in the field to evaluate the performance of surgeons.

In other iterations, this level of computational journalism means examining information in new, more complex ways. Examples of that type of data journalism are the *Wall Street Journal*'s coverage of the Medicare system, which received the 2015 Pulitzer Prize in Investigative Reporting, and Reuters' 2014 project examining influence in the Supreme Court.

INSTITUTIONAL CHALLENGES: RESOURCES

Depending on the university, some students need more support with technology. Some students still do not own personal laptops and rely on school computer labs for their assignments, for example. Other students may be using personal computing devices that are not equipped with what they need to do data journalism. Students who use a tablet (such as an iPad) as their primary tool will face barriers.

Meredith Broussard, who taught data journalism at Temple University until 2015, said that ensuring that her students had the equipment they needed for her class was a major priority. Many of her students relied on a tablet, which meant equipping computer labs with the necessary equipment and platforms—or even lending laptops to students for the term.

Brant Houston of the University of Illinois Urbana-Champaign also pointed to the availability of resources as an important issue—especially for universities that draw students from economically disadvantaged populations.

Journalism schools can help these students by investing in up-to-date lab equipment and by working to create an environment that makes it easy for students to access needed software and to install it on their own devices. Journalism school administrators should consider more frequent audits and surveys of professors to identify which software will be most useful for their students.

And for students who are working on their own personal laptops, some professors hold provisioning sessions to help students install the needed software at the beginning of the term.

INSTITUTIONAL CHALLENGES: FACULTY EXPERTISE

There is no secret that a divide exists between the professional journalism world and the academic world. This chasm continues even with faculty when it comes to who teaches data journalism and the impact it will have on the department.

Of course, each brand of data journalism instructor may have his or her own biases. Those who started as professional journalists, or who still work in a newsroom and teach as an adjunct, believe that they can convey the critical thinking skills needed to succeed in a newsroom environment more effectively than a professor whose experience is in research.

On the other hand, Diakopoulos of the University of Maryland believes that faculty should hold PhDs, and that while it would be good to be able to hire someone with 25 years of experience in data journalism, it's an unrealistic expectation at this stage of the field's development. His goal, he said, is to teach thinking through research. Still, he admitted that this is a struggle.

Some data journalists and journalism professors take issue with Diakopoulos, suggesting that such a model of data journalism professors with PhDs is unrealistic in a world where data journalism emerged from the professional practice, not academia.

Wherever journalism schools find the necessary faculty, just hiring a new professor to specialize in data journalism will not solve the problem, said Doig from ASU. "One difficulty with having somebody like me is everybody else can say, 'Ah, we don't have to worry about data journalism now.' In reality, I teach maybe two sections of 20 students each semester. That's a fraction of our total student load," Doig said. "So believing that it is somehow being taken care of by one specialist like that, that isn't the case."

To help solve at least some of the issues, Doig has provided short video tutorials to other professors for basic government reporting classes.

While Doig believes it would be good to have a required data journalism course, he also questions whether that is possible. "How are you going to find the faculty to teach that?" he asked. "There's not enough people in town who could teach that."

Professional track and academic track faculty members agree that for now, pulling in professional journalists to serve as adjuncts will continue to be necessary and that relying on professional journalists alone will not solve the problem.

For Dustin Harp at the University of Texas, Arlington, this conundrum was solved through her own initiative. She had never taught data journalism but decided the students needed the class, so she did some research and created one. Some colleagues asked her why. She has tenure and no one asked her to take on the extra work. But the students needed the class, Harp said. She used lynda.com for tutorials and learned the same information before teaching her students.

"The thing is I'm a qualitative researcher, I'm not a numbers person, I'm not a numbers cruncher, so it was very crazy and daunting. After I said I was going to do it, it was on the schedule, I was like what have I gotten myself into?" Harp said. "But I follow the field. . . . I'm aware that data journalism is, it's a tool our students need to be more competitive to get jobs."

INSTITUTIONAL CHALLENGES: STUDENT ENGAGEMENT

Journalism programs need to do a better job of persuading or even requiring students to take a data journalism class. Students may shy away because they believe they aren't any good at math. "A lot of students are scared of 'that math thing,'" said one journalism student at Northwestern University.

Resistance to math is an issue far broader than the field of journalism, but it will need to be addressed if teaching data journalism is to be taken seriously. This applies to both teachers and students, some of whom may have chosen to pursue journalism in part because they thought that it would require little or no math.

The problems go deeper than just convincing people they can handle math. Even in universities with entire programs focused on teaching programming, data journalism, and even data visualization, some students have reported that it wasn't easy to find out about these opportunities. Some of the reasons have to do with silos within schools and departments for specific programs as well as specific tracks with emphasis on specific types of journalistic practice.

Rich Gordon, co-founder of the Knight Lab and director of digital innovation at the Medill School of Journalism at Northwestern University, agrees that a gap exists between the basic CAR course and the much more advanced program through Knight Lab, which brings in technologists and works with the technologists to develop new applications for journalism. Journalism

students going through a normal degree plan may have the opportunity to take a basic CAR class, but most won't ever be exposed to the work at the lab, he said.

In general, data journalism courses are electives and draw only a few students out of the total enrolled in each journalism program. Some of that has to do with capacity, but another issue is the lack of visibility. Often, other professors don't treat the classes as vital to a journalism career.

"There's some student interest in CAR," said one University of Missouri journalism graduate. "But there would be more if it were expressed as an option for students early on."

Universities can address this issue, said Mike Reilley, professor of practice at Arizona State University, who regards universities as "'too siloed." Reilley advocates team-taught courses and cooperation between departments.

Some students work their way through to find what they need. For instance, one student who took Temple University's undergraduate class in data journalism had taken classes in programming with Python through the university's business school. She told our researchers she planned on learning more data journalism as she wanted to continue doing this type of journalism when she graduated. But institutional changes could make data journalism much more accessible.

Several students suggested that schools should offer a track that could include a journalistically focused statistics class, a class that focuses on databases, a class with a focus on reporting with data, and others that delve into more in-depth data reporting and data visualization.

One of the authors of this report co-taught a spring 2015 watchdog reporting class with an engineering professor. Five computer science students embedded into project teams of journalism students. The journalism students learned new data skills, and the computer science students learned techniques in reporting and writing. The class offered challenges, too. Next time around, there may be a more defined way for the journalism students to take on data challenges of their own and continued emphasis on having the computer science students learn skills such as interviewing.

CHAPTER 4: MODEL CURRICULA IN DATA AND COMPUTATION

INTRODUCTION AND SUMMARY OF CURRICULAR RECOMMENDATIONS

The preceding chapters offer a picture of the state of education in data and computational journalism in the United States, as well as an argument for the necessity and even urgency of journalism schools committing to teach these subjects. What follows in this chapter are model curricula and guidelines that we hope will facilitate this transition. We intend these models to be flexible; we hope that this information can be applied across schools despite variation in term length, academic units, and time to degree.

This chapter is divided into five sections. The first is a model for an introductory class in data and computation that we recommend as a requirement for all journalism students. The second section offers ways of integrating data and computational instruction into core classes and certain electives. The last three sections present full model curricula for a range of degrees. The first of those three is a track or concentration in data and computational journalism. It is flexible in order to be generally applicable at the undergraduate or graduate level. The other two are models for advanced graduate work. The first presumes a student with some reporting experience or a journalism degree in hand who wishes to develop expertise-driven reporting skills—that is, to write about complicated subjects from a position of deep understanding. The final model is for a research-driven, lab-based graduate degree in emerging media and technological innovation.

Above all, we recommend that all programs have a required foundational course in data journalism, teaching basic principles of data analysis for the purpose of finding stories while cultivating a sense of the general techniques

and possibilities of data-driven reporting. The premise of this course is that all reporters must be prepared to use data in their work and to recognize when this approach is needed.

Considering that many journalism programs are designed to cover a dizzying range of material in a short period of time, some of our readers might be wondering where to find room for a required class in data journalism. Schools that have an existing class on basic numeracy for journalists could rework the class to include a greater emphasis on data-driven reporting methods. Another opening might lie in multimedia classes, many of which were introduced only in the past decade. Compared with data journalism, which frames a mindset for gathering and presenting information, multimedia instruction often centers on teaching tools with uncertain shelf lives. It might be time to consider retiring the audio slideshow from required coursework to make room for data skills.

MODEL 1. INTEGRATING DATA AS A CORE CLASS: *Foundations of Data Journalism*

This is a model for a required introductory course at the graduate or undergraduate level. What follows is a narrative account of how such a class may proceed in developing data literacy among beginning journalists. We realize that this course may need to fit the unique contours of different journalism programs, some of which contain boot camps and other introductory programs with idiosyncratic durations and varying levels of intensity and focus on different skills. The point is for this course to be given equal footing with other skills or subject matters that are currently treated as essential in a journalism education.

COURSE DESCRIPTION: This course is an introduction to the collection, analysis, presentation, and critique of structured information by journalists. As students are introduced to the basics of reporting and the range of journalistic methods that they may pursue in later coursework, an introduction to data and computation is an essential component of their journalism education.

Over the course of a term, students should begin to develop a frame of mind in which they approach every story looking for data possibilities. They should understand how to use basic methods using spreadsheets and relational databases. They should get a primer on using and understanding statistical concepts. They should learn how to take their data findings and locate the people who illustrate those findings for their stories. They will learn how to convert their data analysis into a pitch for a journalistic story.

Students will learn how to find data online, how to maintain personal records as they report stories, and how to use simple visualization methods to find new information: how keeping a timeline can help reveal discrepancies and how cross-checking sources of information may lead to new avenues of

inquiry. The use of data in these contexts will benefit students no matter which area of journalism they choose to practice. Just like interviewing, which is a ubiquitous journalistic skill, the art of gathering and understanding data should extend widely across the field of journalism.

The trouble with data is that it so often appears clinical or detached from the richness of people's lives. To reduce things to abstractions may seem limiting to some students. Early exercises may help to counter this presupposition. If you ask the class to gather information about each other such as their birth dates, blood types, eye color, and birthplace, they may see within a 20-minute exercise how interesting data can be when we learn something from data that we care to know.

From this point, the class may move to more journalistic exercises with spreadsheets. As students become more comfortable with spreadsheets, the class may turn to methods of data analysis such as pivot tables and other plotting methods.

SKILLS: This class should prepare students to use spreadsheets and databases to find and tell stories.

Central concerns include: spreadsheet training, how to find data, clean it, look for patterns and outliers, and question the biases and omissions in how it was gathered. Instructors may choose to use an introductory data set with a good story for beginners to find (examples are listed in the appendix).

Students should also learn how to critically assess claims surrounding data. Reporting on data may go astray when it presumes this information is complete and accurate. Reporters should be trained to look for problems in data. It is necessary to question every source of information.

Another foundational aspect of data-driven reporting is to recognize patterns and anomalies in data. Two skills that should emerge from this class are to look for trends and to identify outliers. Every data point is a possible source or anecdote.

This class will introduce data visualization, but mainly as a means to explore a data set. Using an approachable program such as Excel, Fusion Tables, or Tableau, students will learn to display data in graphic form as an inroad to asking journalistic questions. The goal is not to design a graphic for publication, but to graph for the sake of understanding the data. Students may think of this as a research method or a sketchpad for further reporting. Instruction should include discussion of the ways that different visualization methods can be misleading.

Along the way, this class can cover basic numeracy and descriptive statistics—skills that every journalist needs to know. This may include reminders about how to calculate percentage and percent change, working with units and measures, and even identifying large numbers like billions and trillions. Once those are covered, the material could move to statistics principles and methods such as standard deviation and regression analysis.

TOPICS: Data sources, importing data, negotiating for data, checking the veracity of data, data cleaning, using formulas in spreadsheets, querying data-

bases, finding social significance in the data, writing a data story, visualizing a data story.

COURSE STRUCTURE: Mix of hands-on practice and lectures, primarily using spreadsheet tools and perhaps relational database software; some limited exposure to data visualization for story exploration.

EXAMPLE ASSIGNMENTS

- **HOMEWORK:** Bring a piece of data journalism to critique in class.
- **HOMEWORK:** Find a data set and explain why it's interesting and what it might reveal.
- **CLASSWORK:** Discuss basic data analysis and cleaning on prepared example data.
- **SPREADSHEET ASSIGNMENT:** Analyze a government's payroll, including overtime, or examine a city or county budget. This could be a bridge inspection data set, a city budget, or city payroll.
- **FINAL:** Produce a data story in three assignments: pitch, draft, final submission.

CLASSROOM SUPPORT

Since many students will enter college with little or no prior training in data and computation—and worse still, a bevy of uncertainties about their abilities—we recommend a range of support resources including open lab sessions, teaching assistants, and online resources to review tools and methods.

Open lab sessions give students extra assistance. Matt Waite runs "maker hours" that are well attended by his students at Nebraska. At Northwestern's Knight Lab, students have weekly open hours in which to build and discuss new digital tools.

In our interviews, observations, and personal experience, a teaching assistant (TA) is a considerable asset to classes in data and computation. TAs can offer in-class help when students encounter minor bumps. Here's a common scenario: a student forgets to type a single character while learning a programming language and cannot understand the error message or parse what's missing from that line. Confronted with that situation, many students may not want to interrupt the instructor and as a result could be left behind. The TA can quickly assist with a problem like that. TAs also can be on hand during open lab sessions so that multiple students can get help at once instead of waiting for the instructor.

MODEL 2. INTEGRATING DATA AND COMPUTATION INTO EXISTING COURSES AND CONCENTRATIONS:

General Guidelines for the Undergraduate and Graduate Levels

The basic principles of data journalism should be as familiar to students as writing a lede, shooting b-roll, or tweeting updates to a developing story. To integrate data skills into journalism instruction means introducing these concerns across the curriculum.

Our central recommendation is for journalism schools to treat data and computation as core skills for all students. Data journalism must be taught as a foundational method in introductory classes, a distinct theme in media law and ethics, a reporting method suitable to any specialized reporting course, and a subject in which interested students can pursue advanced coursework or a concentration.

Moreover, because data and algorithms are increasingly important topics to understand in order to report on issues in business, politics, technology, and health, among others, subject area reporting classes should include material that prepares students to approach these information sources with proper skepticism and to explain them clearly in writing. In the models that follow, we point to a few ways that data journalism can be integrated into courses that are commonly offered in journalism schools.

One notable difference between graduate and undergraduate programs is that a master's program often begins with a boot camp in which students are quickly brought up to speed on a wide range of skills. For the majority of students, who enter without a declared concentration, a boot camp may point toward areas of unexpected interest. To integrate data and computational journalism into graduate programs, it must be given equal footing alongside other areas where students may choose to specialize. An introductory module on data journalism will benefit students as much as learning the basics of photojournalism. Moreover, thematic elective coursework such as environmental and political reporting should integrate data instruction to the same degree that it would emphasize such distinct approaches as photojournalism, broadcast, and long-form journalism.

Introductory journalism classes are necessarily broad. Some classes are thematic, covering material from the basic history and general practices of journalism to the range of technologies and reporting techniques that constitute the modern media. Others focus entirely on the practice of journalism. Either way, data and computation must have a place foundational courses.

At the undergraduate level, this should apply to students pursuing either a major or a minor in journalism. Coursework toward the minor also should integrate some measure of data and computational instruction.

Schools may also consider working more coursework in data and computation for other programs and concentrations. Students focused on investigative reporting, for instance, would benefit from additional coursework on finding stories in data, perhaps even as an additional requirement.

INTRODUCTORY AND REQUIRED JOURNALISM CLASSES

Integrating Data and Computation

BASIC GRAPHICS, VIDEO, AND MULTIMEDIA

HOW AND WHY TO INTEGRATE DATA: Different schools may teach a variety of visual tools under the heading of graphic, video, multimedia, or digital media. There are productive ways for data and computation to be integrated into these lessons, however the classes are structured. Data visualization would dovetail with instruction in other graphical storytelling methods such as design and video, for instance, while a general familiarity with news apps could be developed in multimedia classes.

SKILLS TO INTEGRATE: Simple tools for building charts, maps, and timelines. Include building maps and basic data charts, visualizations and timelines, plus an overview on news apps.

POSSIBLE ASSIGNMENTS:

- Use simple tools (Google Fusion, CartoDB, or Esri's Story Maps) to locate the availability of a public service across a geographic area.
- Use simple online charting tools to illustrate changes in the annual budgets of several government offices.
- Include data visualization within a video to provide context and enhance the story.

MEDIA LAW AND ETHICS

HOW AND WHY TO INTEGRATE DATA: Legal considerations form one of the core concerns of data journalists: making public records requests can be one of the most fruitful avenues for reporting, but also one of the most frustrating.

Journalism students should learn the relevant public records laws at the state and federal levels.

They should also address the common misconception that data is sterile, objective, or in some sense detached from human experience. On the contrary, all data exists because someone has chosen to gather it, adn the use of data has social and ethical consequences.

Courses should include material on the verification of photos (through metadata or crowdsourcing) and ethical considerations surrounding leaked or sensitive data, as well as source protection and digital security in conditions of pervasive surveillance.

SKILLS TO INTEGRATE: Becoming familiar with a range of ethical questions surrounding the use of data. Scrutinizing data for bias, errors, and incompleteness.

POSSIBLE ASSIGNMENTS:

» Prepare a critical response paper on legal and ethical concerns surrounding leaked data. This could take the form of an essay or even a mock editorial responding to a sensitive story.

» File a Freedom of Information Act (FOIA) or other public records request, then follow up with needed negotiations. This may be framed as preparation for a project in a subsequent term, if and when the records come through.

HISTORY OF JOURNALISM

HOW AND WHY TO INTEGRATE DATA: Understanding history is especially valuable during times of apparent change. To observe the field of journalism evolving over the centuries can make journalism students more conscious participants in the process of inventing its future. It may also help to temper the widespread view that journalism is witnessing unprecedented upheaval due to technology. Looking back, we see that institutions come and go, new technologies are often disruptive before settling into routine, and the mission and practice of the profession are perennially under revision. Data and computation are in many ways emblematic of our time, but not exclusive to it. These topics have a long history in journalism. This class needs to tell that story.

Two distinct strands of historical concern should be covered. One is to recount the historical uses of data in the news. For example, a striking and memorable early case of data-driven journalism dates to the antebellum period in the United States, when Harriet Beecher Stowe compiled the accounts of several escaped slaves, aggregated advertisements from Southern newspapers offering rewards for their return, and published several tables of data as a rebuttal to claims that her novel *Uncle Tom's Cabin* had exaggerated the reality of slavery. Likewise, one might point to Philip Meyer's use of data to undermine racial stereotypes in the coverage of the 1967 Detroit riots. These two cases highlight the enduring value of data for asserting truths that might

otherwise be denied. More broadly, where these stories place data journalism in historical context, it will not only form a canon to orient students in this area of practice, but it will also reveal that data journalism, for all its glamorous novelty, is rooted in a tradition of quality work.

SKILLS TO INTEGRATE: Acquiring a sense of how the journalistic profession has developed over time, especially in terms of how journalists have chosen to depict the world to their audiences. Appreciating how data and computational journalism fit into historical context.

POSSIBLE ASSIGNMENTS:

» **HOMEWORK:** Find and analyze a chart, graph, map, or other data visualization published in a newspaper at least 50 years ago.

» **TERM PAPER:** Consider a contemporary concern surrounding emerging technology, such as algorithmic transparency or the Snowden leaks, in the context of other historical cases.

ADVANCED CLASSES AND ELECTIVES: *Integrating Data and Computation*

INVESTIGATIVE REPORTING

HOW AND WHY TO INTEGRATE DATA: Many of the tools and methods of computational and data-driven journalism were developed through investigative reporting. Fluency with spreadsheets, databases, and other mainstays of computer-assisted reporting will enable students to conduct deep investigations with the full range of resources at their disposal.

SKILLS TO INTEGRATE: Compiling the backgrounds of people and organizations with the use of data. Turning documents into data. Making public records requests and negotiating for data.

POSSIBLE ASSIGNMENTS:

» Tracing shell company ownership through public records.

» Examining medical device reports for problems in devices sold by specific companies.

NARRATIVE REPORTING AND FEATURE WRITING

HOW AND WHY TO INTEGRATE DATA: Great feature writing is built on facts and compelling narratives. This course should incorporate some data-driven and computer-assisted reporting methods, teaching students to frame, explain, and give context to data that will help to tell their story. This class should highlight that words and numbers are both sources of data. The instructor may consider inviting a guest lecture from a professor in the digital humanities to highlight novel approaches developed in this field for understanding literature and the arts through a computational lens.

SKILLS TO INTEGRATE: General grasp of using numbers to support a narrative. Using spreadsheets to organize chronologies of the main characters in the course of reporting. Using large-scale textual analysis tools to organize, index, and annotate documents.

POSSIBLE ASSIGNMENTS:

» Use Overview or Document Cloud to explore a large cache of documents, such as the Congressional Record, Wikipedia, or a recent leak.

» Organize reporting for a long-form narrative piece by placing sources, quotes, and chronologies in a spreadsheet.

» Analyze tax return (IRS Form 990) data on arts nonprofits to evaluate their finances.

SOCIAL MEDIA SKILLS

HOW AND WHY TO INTEGRATE DATA: The use of social media by contemporary news organizations goes hand in hand with the use of analytics to drive traffic. If students are taught to run social media feeds, they also should be taught to understand the analytics for these platforms. Moreover, the ability to mine the social web to interpret social trends and public opinion will be an asset in reporting.

SKILLS TO INTEGRATE: Gathering and interpreting web analytics. Scraping or otherwise aggregating social media content for analysis use in a story.

POSSIBLE ASSIGNMENTS:

» Use Twitter analytics to determine the rate of growth in followers, retweeting activity, or the most popular stories, sections, writers, and days of the week.

» Use Google analytics to aggregate several streams of traffic data and generate more complicated (second-order) insights.

» Use Google Trends to do a story on patterns in search data.

» Analyze social media data to produce chart of attention around a recent news event.

» (Advanced) Use scraped Twitter data to tell a story (perhaps through sentiment analysis).

BUSINESS AND ECONOMIC REPORTING

HOW AND WHY TO INTEGRATE DATA: The ability to gather, analyze, and critique financial data is an essential component of business reporting. Many classes already include some instruction on reading and interpreting data. As more of this data has become generally available, while some of it has become more complicated and difficult to interpret, business reporting classes will need to adapt and offer more advanced instruction.

SKILLS TO INTEGRATE: The ability to gather and analyze data from a variety of sources, including Bloomberg terminals and APIs (application programming interfaces) for financial information. Advanced spreadsheet analysis and financial/budget analysis training.

POSSIBLE ASSIGNMENTS:

- Spreadsheet assignment: find a story in a company's public financial statements.
- Build a personal dashboard of APIs to track financial information for a story.
- Analyze whether you can predict earnings or stock price through a factor like CEO salary.

DIGITAL DESIGN AND VISUAL COMMUNICATION

HOW AND WHY TO INTEGRATE DATA: Digital design courses in journalism schools serve to introduce students to layout design, editorial graphics, and the principles of visual critique. In order to integrate data and computation, such a course should include material on data visualization and at least an introduction to the idea of news apps and web development.

SKILLS TO INTEGRATE: Basic charts, graphs, and maps. A visual critique to know which styles of visualization are good for which kinds of data and to pinpoint cases in which visual forms can conceal or distort the data.

POSSIBLE ASSIGNMENTS:

- Find a data visualization you like, then dissect, explain, analyze, and critique it.
- Find some data, identify what's interesting about it, and visualize your findings.
- Mock up (design, don't program) a news app.

HOW AND WHY TO INTEGRATE DATA: When journalists cover other countries, numbers will often help both them and their audience to picture these unfamiliar and often complicated matters with greater clarity. A global reporting class should teach students to find, assess, and accurately convey facts and figures about foreign countries and subjects with an international scope. On a deeper level, such a class should teach students to find stories by gathering and scrutinizing data from global sources.

SKILLS TO INTEGRATE: How to gather, evaluate, and use data from multiple international sources. How to evaluate what data can communicate about international development patterns. How to use data to complete an investigative project focused on an international issue.

POSSIBLE ASSIGNMENTS:

» Use a data set from a large international organization such as the UN to find a story, then learn how the organization gathered its data and discuss the limitations and biases that may result.

» Find data that deepens your understanding of an international story in the news, complicates the prevailing narrative, or reveals another side of it.

SCIENCE AND ENVIRONMENTAL REPORTING

HOW AND WHY TO INTEGRATE DATA: Data is a crucial component of scientific topics in the news. The ability to interpret research papers and scrutinize experimental methods will make students far better reporters on these subjects. Students should emerge from this class with an understanding of the scientific method, randomized controlled experiments, statistical significance, and other factors that they will encounter while reporting on topics in science and the environment. If possible, they should also have the opportunity to use their own data sources, such as sensors for air or water quality.

SKILLS TO INTEGRATE: How to gather, evaluate, and use data on specialized scientific topics, and to critically assess published research.

POSSIBLE ASSIGNMENTS:

» Reading and analyzing data stories covering these topics and reverse engineering how the reporters told this story.

» Drafting data analysis of key data sets and story pitch memos.

» Reading a research paper, evaluating the evidence (including the statistical arguments used), and summarizing in plain language for a non-technical reader.

» Setting up a sensor network to test air quality across campus (class project).

MODEL 3. CONCENTRATION IN DATA & COMPUTATION

A data journalism concentration should begin with several core, required classes before moving into a track of electives offering data journalism analysis, visualization, and online research/backgrounding.

The curriculum detailed below should provide a framework for a school to begin offering specialized coursework to students who wish to concentrate in data-driven reporting or computational journalism.

This section describes some of the courses that may form such a degree. Depending on the availability of instructors and other resources, classes like these may form either the mandatory core of a concentration in data and computation, or else a range of electives.

Please note that we would not expect any journalism school to offer all of these classes, nor only these, in its data and computational curriculum. This is just one picture of the skills and thematic exposure that could constitute a journalism degree specializing in data and computation.

CORE CLASSES

Required for Concentration in Data & Computation

FOUNDATIONS OF DATA JOURNALISM

This is the course outlined in the opening of this chapter (see full description on page 50) as a requirement for all journalism students. If students enter journalism school without declared concentrations, this introduction will be suitable for future data concentrators to learn the basics before proceeding to other required courses and electives. Schools may also choose to require applicants to be specifically accepted into the data concentration, in which case it may be advisable to offer a summer boot camp (see "Note on Incoming Skills, Technical Literacies, and Specialized Boot Camps," page 74) to get students up to speed on the tools and methods they will need. In this case, data concentrators may be placed in a more advanced fall foundations course with their peers.

INTRODUCTION TO JOURNALISTIC PROGRAMMING

COURSE DESCRIPTION: The purpose of this course is to introduce students to several foundational computer-programming skills that they will use to find and tell stories. This should be a requirement of those who concentrate in data and computation, but also open to students from other tracks.

COURSE STRUCTURE: Meets twice weekly, first for lecture and then for an intensive workshop.
SKILLS: The Unix command line; basic Python programming for scraping, parsing, connecting to APIs; introduction to JavaScript for web work.
TOOLS: Bash utilities, Jupyter/IPython Notebook, Pandas, Matplotlib, JavaScript.

EXAMPLE ASSIGNMENTS:

» Test proficiency with the command line with a quiz, or even a screencast demonstrating completion of a series of tasks using Bash alone.

» Story assignment reported and submitted in Jupyter/IPython notebook.

STATISTICS FOR JOURNALISM

COURSE DESCRIPTION: The methods and principles of statistics have proven to be powerful tools in the hands of journalists. This course should be a rigorous introduction to stats work taught from within a framework of journalistic concerns. That means the course is story-based, in the sense of precision journalism and the CAR tradition.
COURSE STRUCTURE: Weekly lectures with in-class exercises, regular homework, and a final exam.
SKILLS: Developing and testing hypotheses; understanding and applying the central limit theorem, normal distribution, and confidence intervals; Frequentist versus Bayesian statistics; linear regression; analysis of variance.
TOOLS: R Studio, Excel, MySQL, Microsoft Access, SAS, SPSS (proprietary) or PSPP (F/OSS).

EXAMPLE ASSIGNMENTS:

» Analyze crime statistics, look for a trend, and try to explain its cause.

» Look at the distribution of cancer cases and try to decide if there is evidence of an increase in more polluted areas.

» Analyze statistical evidence for U.S. and international cases to predict whether reducing the number of guns would have an effect on gun violence.

» Analyze the stats in a research paper and report them in plain language.

DISTRIBUTION OF ELECTIVES

For the concentration, the school may offer elective courses to fulfill requirements in two or three areas of data and computational work. We have divided these into three categories: presentation/visualization, analysis for story, and journalistic programming. As a matter of designing degree requirements, a program might choose to require at least one class from each category in addition to fulfilling overall credit requirements.

PRESENTATION & VISUALIZATION

- Data Visualization
- Visual Journalism with Data and Computation
- Advanced Data Visualization
- Advanced Journalistic Mapping

ANALYSIS FOR STORY

- Writing About Data
- Statistical Analysis for Journalism
- Advanced Computational Reporting Methods (Using CAR)

JOURNALISTIC PROGRAMMING

- Introduction to Journalistic Programming
- Methods of Collecting Data and Automating Reporting
- News App Development
- Advanced Computational Journalism

ELECTIVE COURSEWORK

Graduate Degree with Concentration in Data & Computation

METHODS OF COLLECTING DATA & AUTOMATING REPORTING

COURSE DESCRIPTION: This course focuses on developing expertise in gathering data, cleaning it, storing it in a database, and retrieving it with ease. It also emphasizes building automated tools to serve as data sources in reporting.
COURSE STRUCTURE: Weekly workshop or lab-based instruction.
SKILLS: Web scraping, APIs, cron jobs, bash scripting, digitizing paper documents, regular expressions, parsing text and data, fuzzy string matching, record linkage, content analysis.
TOOLS: Python, Beautiful Soup, Mechanize, Scrapy, Tabula, SQL, MongoDB, data formats (CSV, JSON), Tesseract (OCR), Twitter bots.

EXAMPLE ASSIGNMENTS:

- **CLASSWORK:** Design Google Alerts to monitor subjects of interest.
- **HOMEWORK:** Write a program to scrape the Congressional Record for everything a particular representative has said on the floor of the House.
- **HOMEWORK:** Write a web scraper in Python and automate it with a cron job.
- **HOMEWORK:** Build a web app or Twitter bot to post useful information from an API.

» **GROUP PROJECT:** Build a sensor network to automatically post temperature or air quality measurements online.

» **FINAL PROJECT:** Gather a useful body of data, previously unavailable, and share it publicly.

VISUAL JOURNALISM WITH DATA AND COMPUTATION

COURSE DESCRIPTION: This course covers a range of methods, media, and formats for the graphic presentation of information. Readings should introduce principles of visual design and integrate these into regular assignments. Beginning with a fairly basic program like Tableau, the class should highlight the effective and accurate presentation of information in graphic form. By the middle of the term, students should branch out into using a programming library such as D3 to design their own graphics outside the constraints of existing software.

COURSE STRUCTURE: Weekly seminar to discuss readings, followed by hands-on workshop.

SKILLS: Data visualization, news apps, GIS/mapping for presentation.

TOOLS: Tableau, JavaScript, D3, QGIS, CartoDB.

EXAMPLE ASSIGNMENTS:

» **HOMEWORK:** Use Tableau to find a story in a previously unexplored data set.

» **FINAL PROJECT:** Create an original visualization or interactive piece programmed by hand (presumably in D3 or even pure JavaScript if it was taught in an earlier class).

ADVANCED DATA ANALYSIS & JOURNALISTIC ALGORITHMS

COURSE DESCRIPTION: This course should build upon the core, required classes to bring together data and computation for finding stories and making predictions using algorithmic and computational analysis.

COURSE STRUCTURE: Weekly lecture and workshop with regular homework and a final project.

SKILLS: Python for machine learning, clustering, classifying documents, standardizing and matching algorithms.

TOOLS: R, Python (Pandas, MatPlotLib, SciPy, scikit-learn), clustering algorithms (k-means, k-nearest neighbor clustering), topic modeling algorithms (LDA or NMF).

EXAMPLE ASSIGNMENTS:

» **CLASSWORK:** Record linkage for data cleaning, for example, analyze Federal Election Commission data to find top donors, which requires regularization of names, best done with machine learning.

» **HOMEWORK:** Analyze State of the Union speeches since 1790 to make a visualization of how key topics have changed over time.

» **HOMEWORK:** Implement clustering to detect outliers in a data set.

» **FINAL PROJECT OPTION:** Build an election or market prediction model.

» **FINAL PROJECT OPTION:** Reverse engineer a pricing, lending, or credit score algorithm.

CYBERSECURITY SKILLS

We recommend that students concentrating in data and computation take a module on digital security because it will be a necessary consideration in their work.

Every news organization should have an information technology staff that is capable of securing its digital infrastructure and advising staff about security risks and countermeasures. In practice, this is not enough. More editorial staff should be trained in digital security in order to assess and address risks to the organization, its sources, and its readers. This training would dovetail with the technical skills that students in a computational journalism course are already learning.

Cybersecurity has become an increasingly salient ethical concern, especially in the wake of the Snowden leaks, but digital security skills are rarely taught in journalism schools today. Just as no journalism student should graduate without some sense of libel law, no student should leave without knowing at least the dangers of insecure communication channels and practices, and ideally also some solutions.

For introductions to encryption and digital security for journalists, see Micah Lee's "Encryption Works" handbook, published by the Freedom of the Press Foundation, and "Security for Journalists" by Jonathan Stray.

ADVANCED DATA VISUALIZATION

COURSE DESCRIPTION: This course would pick up from the data visualization skills developed in the core course in visual journalism. The technical aspects should be conducted entirely through programming. The most likely tools are JavaScript and D3, but others will certainly emerge. The key point is that data visualization at this level should be programmatic so that the software itself does not limit design possibilities.

COURSE STRUCTURE: It may be designed to alternate between seminars (high-level reading, discussion, and analysis of visual communication and information design principles, focusing on how it is most effective and where it can be misleading) and lab classes (advanced practical instruction in application and coding frameworks for info design).

SKILLS: Designing for clarity, precision, impact.

TOOLS: D3, JavaScript, or another suitable programming framework.

EXAMPLE ASSIGNMENTS:

» **HOMEWORK:** Regular data assignments in different media: static web, video, interactive.

» **FINAL PROJECT:** An original analysis of unexplored data, presented in an original visualization programmed more or less from scratch, with cross-platform consistency.

ADVANCED JOURNALISTIC MAPPING

COURSE DESCRIPTION: This course should build on previous coursework in mapping to cover more advanced manipulations of data, to develop a higher degree of design sophistication, and to develop a high level of news judgment in the selection of timely, compelling, and original topics. This involves using GIS technologies, joining that spatial data with other information, using density and other spatial analysis to inform stories, not just building presentations.

COURSE STRUCTURE: Hands-on workshop and lab.

TOOLS: Both GIS analysis software and presentation software, including Esri, QGIS, CartoDB, Leaflet, sensors like DustDuino.

SKILLS: Clustering, binning, heat maps, joining different geographic data sets.

EXAMPLE ASSIGNMENTS:

» **HOMEWORK:** Weekly pitches and journalistic mapping assignments.

» **FINAL PROJECT:** An interactive map or set of maps telling a story about a timely or unexplored subject, and/or a narrative story using findings from the mapping analysis.

» **CLASS PROJECT:** Build a sensor network or otherwise amass an unexplored data set, then work in small groups to build a package of maps to explore the data.

ADVANCED JOURNALISTIC TEXT MINING

COURSE DESCRIPTION: Text is data. The purpose of this class is to teach journalists to gather, analyze, and present stories using large amounts of textual data. This may build on material from the course "Methods of Collecting Data and Automating Reporting."

COURSE STRUCTURE: Weekly lecture and lab, working toward a final project.

SKILLS: Web scraping, analyzing large bodies of text, sentiment analysis, topic modeling.

TOOLS: Overview, DocumentCloud, Natural Language Toolkit (NLTK) or Stanford NLP.

EXAMPLE ASSIGNMENTS:

» **HOMEWORK:** Use sentiment analysis to reproduce the before and after tone change of a story.

» **FINAL PROJECT:** Build a scraper to crawl a significant chunk of the Web, for example, collecting the blogosphere of a country that's in the news and learning what people are talking about.

» **FINAL PROJECT:** Gather and analyze a large body of documents, such as looking for a story in a leaked cache of documents.

ADVANCED COMPUTATIONAL JOURNALISM

COURSE DESCRIPTION: This course should reflect the state of computational tools in journalistic practice while looking toward novel applications of emerging and unexplored tools. By this time, students should have already developed a strong foundation of programming and data analysis skills. This class should build on that foundation and encourage in-depth, independent projects centered on reporting stories or developing a piece of software.
COURSE STRUCTURE: Meets twice weekly, once for lecture and once for lab.
TOOLS: Python, Ruby, or a similarly powerful and versatile scripting language. Additionally, physical computing tools like Arduino, which is often programmed in Java.

EXAMPLE ASSIGNMENTS:

» **HOMEWORK:** Use regular expressions to mine the Congressional Record for a senator's stated positions on a political issue over the course of his or her career.

» **MOCK CODING INTERVIEW:** In the common style of interviewing for tech jobs, solve a given programming problem on a whiteboard and narrate your line of thinking.

» **FINAL PROJECT:** An in-depth story reported using an advanced tool, including but not limited to a journalistic algorithm, machine learning, analysis of personally gathered data, or development of a piece of software.

CAPSTONE OR THESIS PROJECT

A thesis in data journalism will arise, ideally, from work with instructors and an adviser. It could take the form of a reported story, a technical report, a piece of software, or a substantial design piece such as a map or data visualization.

A capstone project for concentrators in data and computation may take a class of students and coordinate a project using and honing the skills they have developed in their earlier coursework. Each student's work should then be supplemented with an individual contribution such as a reported piece or data visualization.

MODEL 4. ADVANCED GRADUATE DEGREE:

Expertise-Driven Reporting on Data & Computation

While most undergraduate and master's programs are designed to offer journalistic newcomers a set of skills that can be applied to a wide range of subjects, there is also demand for mid-career journalists to return for coursework in which they can develop deep expertise to report on complicated subjects. As highly technical topics have come to permeate matters from international politics to everyday life, journalism schools may wish to offer classes that prepare students and equip mid-career journalists to report on such issues as cyberwar, data breaches, and cryptocurrency that require specialized skill when writing for a general audience. The uses of data, machine learning, and computational models may also aid these reporters in finding and telling these stories.

Data and computational journalism are ideal subjects for a mid-career degree because of the time and mentorship that could be devoted to developing this set of skills. Since it is directed at students who already know how journalism works, this degree could provide a level of depth and focus that may be difficult to reach during a standard journalism program or while working a full-time job.

COURSES

FOUNDATIONS OF DATA-DRIVEN JOURNALISM

(as detailed on page 50, but adapted to the level of advanced students)

REPORTING ABOUT DATA

COURSE DESCRIPTION: : The goal of this course is to prepare students to understand and critically assess reports, studies, scholarly work, and other information sources that are based in data and technical work. This will be an essential skill as each student develops a focus as an expert reporter on a topic centered on data, computation, technology, or the experimental sciences.

COURSE STRUCTURE: Small seminar focused on the discussion and analysis of readings and case studies, culminating in a longform piece.

EXAMPLE ASSIGNMENTS:

- **RESPONSE PAPERS:** Weekly analysis and reflection on class readings.
- **TERM PAPER:** A substantial long-form reporting project on a topic of the student's choosing, possibly developed as an outgrowth of a weekly response paper.

THESIS

COURSE DESCRIPTION AND STRUCTURE: Seminar in which students develop independent reporting projects, sharing progress during class and meeting regularly with an adviser to build toward a master's thesis.

ELECTIVES

One of the goals of a mid-career, expertise-driven degree in journalism is for students to develop a deep understanding of the field they are reporting. To this end, this degree should offer several elective slots for taking classes in other departments that contribute directly to the subject of the thesis.

Students may also consider auditing courses with skill requirements above their level (for example, if assignments must be submitted in the C programming language, which is still the case in some traditional computer science classes).

EXAMPLE ELECTIVES AND JUSTIFICATION:

- An earth science or geology course focused on climate data.
- A digital humanities course that uses computational techniques to explore historical archives, literary works, or leaked caches of documents, to name just a few examples.
- Any number of computer science courses in which students could learn the technical basis and academic concerns surrounding issues of interest in their reporting, such as computer vision or cryptography.
- A course in digital security could help a journalist not only to protect sensitive sources, but also to report on such matters as public key encryption or onion routing, and to assess new developments in these fields.
- A graduate course in statistical modeling, whether taken in the statistics department or in a quantitative social science such as sociology.

The point of elective courses should be to permit students to craft a coursework plan that is suitable to their own unique interests as they develop the capacity for expertise-driven reporting in some area related to data, computation, and emerging technologies.

MODEL 5. ADVANCED GRADUATE DEGREE: *Emerging Journalistic Techniques and Technologies*

Investigative reporting is in many ways the research and development wing of journalism. According to Brant Houston, "It's the only place where people have had an extensive amount of time to try out new techniques."

CAR, data journalism, and computational journalism are some of the clearest examples of this phenomenon at work. These practices have developed where reporters have had the time or inclination to work with new tools and platforms. Universities are ideally suited to cultivate this stance toward journalistic practice—not merely teaching the wisdom of the field as it exists, but developing entirely new approaches based on encounters with other disciplines and unexplored tools.

If journalism schools were to take up the mantle of encouraging work that seems to happen only under these permissive conditions—not just through grants and innovation labs, but perhaps through coursework as well—then universities could also act as R&D labs in a way that investigative reporting has in the past.

This curriculum is in many ways the least structured and most speculative one we offer. It is an open question whether these degrees should be offered at the master's or doctoral level. One might also ask whether the degree should require any coursework or simply provide an open platform for research.

OLD AND NEW TECHNOLOGIES

The history of technology often appears to move in regular cycles of emergence and obsolescence, but in fact old technologies are rarely eclipsed entirely. We must be cautious with the concept of "emerging" technologies because we risk missing the continued utility of old ones.

For example, microcontrollers like the Arduino have minimal computing power by contemporary standards, but they are powerful enough to process a set of programmed instructions for projects like gathering sensor data. These devices have proved especially useful *because* of their simplicity, not in spite of it. Similarly, as we promote spaces for journalism schools to explore technologies so new that their uses are not yet apparent, it will be worthwhile to maintain a perspective broad enough to consider the utility of seemingly obsolete technologies.

We should also bear in mind the long histories of platforms like virtual reality and holograms as part of our cultural imagination, if not yet as successful mass products. In past conceptions of the future, we may rediscover promising avenues for innovation.

One objective for this program would be to help teach algorithms for journalism, machine learning, and artificial intelligence.

Drones and virtual reality are two platforms that are being actively explored for their journalistic potential. Matt Waite established the Drone Journalism Lab at the University of Nebraska precisely to explore the journalistic applications of these devices. Likewise, the Brown Institute at Columbia has sponsored several Magic Grants to support teams of journalists exploring the narrative potential of immersive virtual reality.

Many other emerging technologies have been recognized for their journalistic potential. At the time of our writing, immersive virtual reality headsets seem poised to enter the market to enthusiastic reception. Augmented reality presents similar possibilities: broadcast journalists may soon arrive in our living rooms as holograms.

The point is not to speculate on the arrival of these devices, nor to promote innovation for its own sake, but to consider the role of journalism schools in developing and shaping the use of new devices.

SETTING AND GEAR FOR EMERGING TECHNOLOGY LABS

Although many coding and design projects may require nothing more than a laptop, a variety of hardware should be on hand for students interested in experimenting with sensors and other hardware that can be used for journalistic projects.

Ideally, cheap devices and components can be provided on an honor system, and more expensive gear checked out through an equipment room. A thriving example of this model is the Interactive Telecommunications Program at New York University. The program also designates a few shelves for donating useful scrap materials, such as old electronics to be dismantled for components or sheer curiosity.

A WORD ON SAFETY

Most innovation labs will feature at least one tool or device that requires safety training. Most journalism students will arrive without having had experience handling soldering irons or electrical wiring.

Any lab that includes these devices must provide some safety infrastructure. Soldering irons should be used with some means of ventilation. Fire hazards require a nearby extinguisher. And many circumstances may require safety gloves or goggles.

When Amanda Hickman arrived at the BuzzFeed lab in San Francisco, one of her first tasks was to evaluate safety. The BuzzFeed team has purchased safety goggles and fire extinguishers, for example, because it is working with saws and soldering irons.

Tinkering equipment can be quite cheap, but any technology lab must cover some basics. These components are the bread and butter of hacker and maker circles, so they are easy to find. Because they offer such useful inroads to experimenting with technology, they are valuable for journalism schools to cultivate spaces of innovation.

Most electrical prototyping starts with a solderless breadboard, a flat plastic case with an underlying grid of connections or building circuits. A simple electronic device like an air quality meter can be built from scratch by placing components like wires, resistors, knobs, buttons, and sensors across the grid. And connecting a breadboard device to a simple computer like an Arduino or a Raspberry Pi enables users to issue commands and gather data from the equipment. A starter set for such a project would generally run under $100, far less than cameras and other equipment that journalism students are often required to purchase.

Beyond the small computers used for prototyping, more substantial computers should be on hand for projects that call for it. If possible, an emerging technologies lab should have machines that allow students to gain firsthand experience working with news-bound technology such as immersive 3D cameras, VR headsets, and drones instead of relying on secondhand accounts. These skills and literacies fit into a larger constellation of technical concerns that may give rise to media innovation along unforeseen paths.

CHAPTER 5: INSTITUTIONAL RECOMMENDATIONS

Steps Toward Bringing Data and Computation into Your Journalism School

FACULTY DEVELOPMENT AND RECRUITMENT

For many journalism schools, integrating specialized coursework in data and computation will present something of a chicken-and-egg conundrum. There is professional demand for data journalists in part because they are relatively scarce, so while schools may wish to prepare their graduates for this emerging field, the field itself may not yet have enough teachers in its ranks.

But specialized coursework will meet only some of the need highlighted in this study: data and computation must also continue to be integrated into many classes where it is now neglected, from boot camps to capstones and theses. It is worth recalling that editors once resisted the use of photography as a journalistic tool. Journalism schools must prepare students to bring data and computation to any story that needs it.

This may require many journalism faculty to be trained to work with data and computational tools. At the very least, journalism instructors should be conscious of when a student's work may benefit from data, even if the student must go elsewhere for targeted instruction. Scheduling guest lectures may also serve as a transitional solution.

TRAININGS OR MODULES

In 2014, NICAR introduced data journalism exercises for academics interested in teaching data journalism but in need of a little help. These promise to be useful to journalism instructors who would like to teach data.

The benefit of NICAR and other journalism training organizations could go well beyond the modules, though. New tools are developing quickly, and it is critical for faculty to continue to grow, learn, and change as the field itself develops.

In fact, NICAR filled a vacuum that existed because many academic institutions didn't address new tools or skills. Meanwhile, personal computing tools became more powerful, digital information sources became more commonplace, and news organizations increasingly relied on digital methods of gathering and distributing news. Just as print newspapers were slow to recognize the power of data-driven reporting and the Internet, so too have journalism schools been reluctant to change. Teaching institutions must adapt or risk being unable to fulfill their goals and mission, both to their students and to the profession.

INCOMING SKILLS, TECHNICAL LITERACIES, AND BOOT CAMPS

Many graduate programs in journalism readily enroll students with little to no prior experience as reporters. There is an implicit assumption that their undergraduate work will provide a foundation to begin learning to think like a reporter and produce stories in a variety of platforms.

With data and computational journalism, though, there may be more substantial gaps to bridge in terms of math skills and technical literacies. Oftentimes, students must learn to use a variety of unfamiliar software in order to even begin working with data, statistics, and programming languages. As it stands, it should be fairly straightforward to teach the average journalism student to think about data, to find stories in a spreadsheet, and even to think critically about the numbers. Reporters have always needed to see inside complicated issues and to ask tough questions in order to get the story right.

Math and tech skills may require extra time. This skill gap could be ameliorated with a summer boot camp that focuses largely on building skills, tools, and technical literacies, while deferring instruction in reporting until the regular term begins. This way, when students enter their regular master's coursework, they will be equipped with some fluency in the data and computational tools that they will use as concentrators.

In the case that data concentrators go through an extended boot camp, it may be appropriate for their fall introduction to data journalism class to be separate and more advanced than the data journalism course that is required of all students. For an example of how this coursework could be structured, we have listed the offerings of Columbia's Lede Program in the appendix.

Many colleges and universities provide computer labs and studios for classes. The primary advantage is the certainty that each student will have a workstation with the necessary technical specifications and software installed. The primary disadvantage is that students may graduate without the tools they need to practice the skills they have learned.

Although their newsroom workstation could potentially be outfitted with the tools they need, if any of those students became freelancers, they may be out of luck if they left class without bringing along them the tools they learned in school.

Providing server space for students is a great way to begin teaching them the Unix command line and to provide resources for data-intensive projects. But several institutional concerns arise. Schools are required by law to maintain the confidentiality of student data, and so the security of student servers may become a concern. Students might instead begin by working on virtual machines using a program such as the free, cross-platform VirtualBox in order to become acquainted with running a machine from the command line.

BENEFITS OF DISTANCE OR ONLINE LEARNING

Using MOOCs in complementary fashion with data journalism courses could help professors integrate new skills into what they offer, said Doig from ASU.

In addition, distance learning and virtual classrooms may provide structure and support that MOOCs lack. Journalism schools may consider coordinating partnerships in which students cross-enroll in specialized coursework and take the class over a video stream. The student would participate in class, submit work, and receive credit like any other student. This approach could fill coursework gaps in cases where it is otherwise difficult to find an instructor.

Stanton, the founder of ForJournalism.com, offers a cautionary word: maintaining online courses is a problem for any program that produces tutorials or screencasts. Without updating, the value of the offerings diminishes quickly, Stanton said. The ForJournalism.com tutorial on building a web framework with Django is based on an older version of the open source software, for example.

Stanton suggests that universities create a consortium of universities where each participating university would take ownership of specific topics in which it had expert faculty. It would create labs to provide the technical instruction in those areas and offer screencast tutorials on the basics. Then each school could build on that foundational learning in projects specific to their programs.

FOSTERING COLLABORATION

Journalism schools should build collaborative partnerships with other disciplines. Many professional schools, journalism included, have tended to operate as silos within universities because they draw their culture and concerns from a field of practice rather than a tradition of academic discourse. That stance must shift because journalism itself is shifting. As a result, we should recognize that journalism is not a narrow set of traditional newsroom skills, but instead encompasses whatever tools and methods have, in one way or another, been made journalistic. Practitioners of data-driven and computational journalism have thrived by embracing interdisciplinarity in their work. Several journalism schools have begun to build bridges with computer science departments by opening research centers, co-teaching and cross-listing classes, and even developing joint degree programs. This is a promising start. Not only will journalism schools benefit from acting as leaders in interdisciplinary collaboration, but they also should be naturally suited to this role as a field situated at the intersection of many other disciplines.

NOTE ON SPECIALIST FACULTY IN DATA AND COMPUTATION

An integrated data journalism curriculum presents a unique challenge. In the state of the field as it has developed and exists today, data journalism is usually a lone course, or element of a course, taught by one specialist instructor. Often, the instructor is a professional journalist working as an adjunct more for the love of spreading the word than for the money. To achieve a fully integrated curriculum, the overall faculty at journalism programs would need to commit to change, and administrations would need to foster training for faculty. The change needs to be broad. There should not be a single faculty member juggling all the classes in data and computational skills, nor should guest lectures from that faculty member suffice in broadening the class to account for data. Journalism schools must commit to the idea that they cannot train information professionals to work in an increasingly complicated world of information without developing these crucial literacies. It must be integrated across the board.

APPENDIX

TABLES FROM OUR ANALYSIS

Classes Offered by Subject at ACEJMC-Accredited Journalism Programs

DATA JOURNALISM

Number of Classes	Number of Programs	Percent of Total
No class	54	48%
One class	27	24%
Two classes	14	12%
Three or more classes	18	16%

CLASSES WITH DATA JOURNALISM AS COMPONENT

Number of Classes	Number of Programs	Percent of Total
No Classes	44	38%
One Class	31	27%
Two Classes	22	19%
Three Classes	9	8%
Four or More Classes	7	6%

MULTIMEDIA

Number of Classes	Number of Programs	Percent of total
No classes	20	18%
One class	31	27%
Two classes	12	11%
Three classes	16	14%
Four or more classes	34	30%

PROGRAMMING BEYOND HTML/CSS

Number of Classes	Number of Programs	Percent of Total
No Classes	99	88%
One Class	6	5%
Two Classes	5	4%
Three or More Classes	3	3%

Note: This analysis of programming classes is focused on those courses taught within a journalism program. It should be noted that a fair number of schools pointed to collaborations with other departments where journalism students were able to take advanced programming or computer science classes.

NOTABLE STORIES

Below we list several examples, for reference, of stories that are emblematic of the categories we define in Chapter 1.

DATA REPORTING

- "Drugging Our Kids," San Jose Mercury News, 2014
- "Methadone and the Politics of Pain," The Seattle Times, 2012

DATA VISUALIZATION AND INTERACTIVES

- ProPublica's "Dollars for Docs," 2010
- The *Washington Post*'s visualization of the missing Malaysian jet, 2014

EMERGING JOURNALISTIC TECHNOLOGIES

DRONE EXAMPLES:

- "Tanzania: Initiative to Stop the Poaching of Elephants," CCTV Africa, 2014
- Because of regulatory issues with the Federal Aviation Administration, the use of drones for journalism is not widespread in spite of significant interest on the part of industry and academia. Uses foreseen when regulations become more permissible include news photography and videography, scanning news locations for use in 3D models and 360-degree video applications, remotely sensed data gathering through visible images or multispectral images, mapping of areas of interest at higher temporal resolutions than currently available and as sensor distributors or sensor-based data gatherers.

SENSOR EXAMPLES:

- WNYC's Cicada Tracker project in 2013 recruited interested listeners to use sensors to identify where cicadas would emerge.
- USA Today's "Ghost Factories" investigation in 2012 used X-ray gun sensors to scan the soil.
- The Houston Chronicle's 2005 investigative story "In Harm's Way" used sensors to examine air quality near oil refineries and factories.

VIRTUAL AND AUGMENTED REALITY EXAMPLES:

- The New York Times sent out more than a million Google Cardboard kits to subscribers in 2015 as it launched its first VR story, "The Displaced," a piece detailing children displaced by war.
- Stanford University's Department of Communication, home to the Stanford Virtual Human Interaction Lab, has scheduled a VR class for the winter 2016 quarter as part of its curriculum for its master's in journalism program.

STORY EXAMPLES:

» The 2014 Wall Street Journal investigation into Medicare

» "The Echo Chamber," a 2014 Reuters investigation into influence at the Supreme Court

PLATFORM EXAMPLE:

» PDF repository DocumentCloud or Overview, developed by Jonathan Stray

TOOLS, RESOURCES AND METHODS DISCUSSED IN THE REPORT

The ethics of software may also shape decisions about the tools and techniques you teach. "Free" software is licensed in an effort to promote freedom of computing, in a manner analogous to freedom of speech. Free software may be copied, altered, used, and shared freely. A related form of software licensing, titled "open source," is very similar to free software, but instead emphasizes the public availability of code.

Proprietary software may also have certain advantages. Often the interface design is more polished, support services are provided, and in some cases they simply run better on demanding tasks.

But the gap between free and proprietary software has become narrower in recent years, and many professionals in fact prefer to use free and open source software on more than ideological grounds. F/OSS software is often more secure because it can be openly vetted by security researchers. For the same reason, particularly popular applications may have many talented and dedicated developers, as well as a support community of fellow users rather than call center or online service.

Given the expense of proprietary software and its inevitable obsolescence, there are few advantages to using these applications in data and computation classes instead of free and open-source ones.

GUIDE TO COMMON TOOLS FOR DATA AND COMPUTATIONAL JOURNALISM

The following list of common tools for data and computational journalism is quoted from the Lede Program at Columbia.

PROGRAMMING LANGUAGES

C is a heavy-lifting programming language that is the language of choice for the Computer Science Department. It's far faster than Python or JavaScript and introduces you to the nitty-gritty of computer science.

Git is something called a version control system—it's not a programming language, but programmers use it often. Version control is a way of keeping track of the history of your code, along with providing a structure that encourages collaboration. GitHub is a popular cloud-based service that makes use of git, and we make heavy use of it during the Lede Program.

HTML isn't technically a programming language, it's a markup language. A HyperText Markup Language, to be exact. HTML is used to explain what different parts of web pages are to your browser, and you use it extensively when learning to scrape web pages.

JavaScript is a programming language that's in charge of interactivity on the Web. When images wiggle or pop-ups annoy you, that's all JavaScript. The popular interactive data visualization framework D3 is built using JavaScript.

Python is a multipurpose programming language that is at home crunching, parsing text, or building Twitter bots. We use Python extensively in the Lede.

R is a programming language that is used widely for mathematical and statistical processing.

TOOLS FOR DATA AND ANALYSIS

Beautiful Soup and ***lxml*** are tools used for taking data from the Web and making it accessible to your computer.

D3 is a JavaScript library for building custom data visualizations.

IPython Notebooks are an interactive programming environment that encourage documentation, transparency, and reproducibility of work. When you're done with your analysis, you'll be able to put your work up for everyone to see—and check!

NLTK (Natural Language Toolkit) is a Python library built to process large amounts of text. Whether you're analyzing congressional bills, Twitter outrages, or Shakespearean plays, NLTK has you covered.

OpenRefine (previously Google Refine) is downloadable software that helps you sort and sift dirty data, cleaning it to the point where you can start your actual analysis.

Pandas is a high-performance data analysis tool for Python.

QGIS (geographic information system) is an open-source tool used to work with geographic data, from reprojecting and combining data sets to running analyses and making visualizations.

Scitkit-learn is a Python package for machine learning and data analysis. It's the Swiss Army knife of data science: it covers classification, regression, clustering, dimensionality reduction, and so much more.

Web scraping is the process of taking information off of websites and making use of it on your computer. A lot of times documents aren't easily available in accessible formats, and you need to scrape them in order to process and analyze them.

DATA FORMATS

An ***API (application programming interface)*** is a way for computers to communicate to one another. For us, this generally means sharing data. We'll be coding up Python scripts to talk to and request data from machines around the world, from Twitter to the U.S. government.

CSVs (comma-separated values) are the most common format for data. It's a quick export away from Excel or Google Spreadsheets, and you'll find yourself working from CSVs more often than any other format. Although "comma-separated" is in the name, a CSV can arguably also use tabs, pipes, or any other character as a field delimiter (although the tab-separated one can also be called a TSV).

GeoJSON and ***Topojson*** are specifically formatted JSON files that contain geographic data.

JSON stands for JavaScript Object Notation, and it's a slightly more complicated format than a CSV. It can contain lists, numbers, strings, sub-items, and all sort of complexities that are great for expressing the nuance of real-world data. Data from an API is often formatted as JSON.

SQL (Structured Query Language) is a language to talk to databases. You'll sometimes find data sets in SQL format, ready to be imported into your database system of choice.

TECH TEAM REPORT

Another useful resource for understanding the tools of data journalism was prepared at Stanford by an interdisciplinary team of computer science and data journalism students in a Spring 2015 course on watchdog reporting. The report is available here: http://cjlab.stanford.edu/tech-team-report/

ONLINE COURSES AND MOOCS

- Doing Journalism with Data: First Steps, Skills and Tools (http://datajournalismcourse.net/)
- School of Data (http://schoolofdata.org/)
- The Knight Center for Journalism in the Americas offers a number of MOOCs as distance learning for journalists (https://knightcenter.utexas.edu/distancelearning)

USEFUL DATA SETS FOR CLASSWORK AND ASSIGNMENTS

- Baby name census data—clean data, always varies from year to year, papers always cover it (Top 1000 baby names by year can be found at https://www.ssa.gov/oact/babynames/limits.html
- Greenhouse gas data (NOAA has a number of searchable datasets at http://www.esrl.noaa.gov/gmd/dv/data/)
- Student grade distributions for your college
- This is a small data set used in a lot of the School of Data Examples: The GRAIN database of land grabs (http://datahub.io/dataset/grain-landgrab-data/resource/af57b7b2-f4e7-4942-88d3-83912865d116)
- World Bank Open Data (http://data.worldbank.org/)
- The Guardian Databases (http://www.theguardian.com/news/datablog/interactive/2013/jan/14/all-our-datasets-index)
- The Eurostat Databases (http://ec.europa.eu/eurostat/help/new-eurostat-website)
- UK Government Databases (https://data.gov.uk/data/search?res_format=RSS)
- National and International Statistical Services by region and country (https://en.wikipedia.org/wiki/List_of_national_and_international_statistical_services)
- Global Health Observatory Data Repository (http://apps.who.int/gho/data/node.home
- Business Registry Databases (https://www.investigativedashboard.org/business_registries/)
- Google's list of Public Data (http://www.google.com/publicdata/directory#)
- Open Spending (https://openspending.org/)
- Datahub (http://datahub.io/)

- Open Access Directory (http://oad.simmons.edu/oadwiki/Main_Page)
- Data Portals (http://dataportals.org/)
- NASA's Data Portal (https://data.nasa.gov/)

PHILIP MEYER'S RECOMMENDED TEXTS

- John Tukey, *Exploratory Data Analysis* (Upper Saddle River, NJ: Pearson Education. 1977)
- James A. Davis, *The Logic of Causal Order* (Thousand Oaks, CA: Sage, 1985)
- Robert P. Abelson, *Statistics as Principled Argument* (Hillsdale, NJ: Lawrence Erlbaum Associates, 1995)

DATA JOURNALISM ARTICLES, PROJECTS, AND READING LISTS USED IN INSTRUCTION

MOOC EXAMPLES:

- Cairo, Alberto. "Recommended Resources for My Infographics and Visualization Courses." Personal. The Functional Art: An Introduction to Information Graphics and Visualization, October 11, 2012. http://www.thefunctionalart.com/2012/10/recommended-readings-for-infographics.html.
- "Cameroon—Cameroon Budget Inquirer." Accessed September 23, 2015. http://cameroon.openspending.org/en/.
- Downs, Kat, Dan Hill, Ted Mellnik, Andrew Metcalf, Cory O'Brien, Cheryl Thompson, and Serdar Tumgoren. "Homicides in the District of Columbia—The *Washington Post*." News. The *Washington Post*, October 14, 2012. http://apps.washingtonpost.com/investigative/homicides/.
- "Find My School .Ke." Accessed September 23, 2015. http://findmyschool.co.ke/.
- Keefe, John, Steven Melendez, and Louise Ma. "Flooding and Flood Zones | WNYC." News. WNYC. Accessed September 23, 2015. http://project.wnyc.org/flooding-sandy-new/index.html.
- Kirk, Chris, and Dan Kois. "How Many People Have Been Killed by Guns Since Newtown?" Slate, September 16, 2013. http://www.slate.com/articles/news_and_politics/crime/2012/12/gun_death_tally_every_american_gun_death_since_newtown_sandy_hook_shooting.html.

» Lewis, Jason. "Revealed: The £1 Billion High Cost Lending Industry | The Bureau of Investigative Journalism." Journalism. The Bureau of Investigative Journalism, June 13, 2013. https://www.thebureauinvestigates.com/2013/06/13/revealed-the-1billion-high-cost-lending-industry/.

» Nguyen, Dan. "Who in Congress Supports SOPA and PIPA/PROTECT-IP? | SOPA Opera." News. ProPublica, January 20, 2012. http://projects.propublica.org/sopa/.

» Rogers, Simon. "Government Spending by Department, 2011-12: Get the Data." The Guardian, December 4, 2012, sec. UK news. http://www.theguardian.com/news/datablog/2012/dec/04/government-spending-department-2011-12.

» ———. "John Snow's Data Journalism: The Cholera Map That Changed the World." The Guardian, March 15, 2013, sec. News. http://www.theguardian.com/news/datablog/2013/mar/15/john-snow-cholera-map.

» ———. "Wikileaks Data Journalism: How We Handled the Data." The Guardian, January 31, 2011, sec. News. http://www.theguardian.com/news/datablog/2011/jan/31/wikileaks-data-journalism.

» ———. "Wikileaks Iraq War Logs: Every Death Mapped." The Guardian, October 22, 2010. http://www.theguardian.com/world/datablog/interactive/2010/oct/23/wikileaks-iraq-deaths-map.

» Rogers, Simon, and John Burn-Murdoch. "Superstorm Sandy: Every Verified Event Mapped and Detailed." The Guardian, October 30, 2012. http://www.theguardian.com/news/datablog/interactive/2012/oct/30/superstorm-sandy-incidents-mapped.

» Serra, Laura, Maia Jastreblansky, Ivan Ruiz, Ricardo Brom, and Mariana Trigo Viera. "Argentina's Senate Expenses 2004-2013." News. La Nacion, April 3, 2013. http://blogs.lanacion.com.ar/ddj/data-driven-investigative-journalism/argentina-senate-expenses/.

» Shaw, Al, Jeremy B. Merrill, and Zamora, Amanda. "Free the Files: Help ProPublica Unlock Political Ad Spending." ProPublica, September 4, 2015. https://projects.propublica.org/free-the-files/.

» "Where Does My Money Go?" Accessed September 23, 2015. http://wheredoesmymoneygo.org/.

LEDE PROGRAM CURRICULUM

The Lede Program at Columbia Journalism School is a post-baccalaureate in which students from a variety of backgrounds learn data and computation skills over the course of one or two semesters. The program was designed to help students rapidly elevate their skills in these areas, especially if they were considering applying for Columbia's highly demanding dual-degree program in journalism and computer science.

In the context of this report, the one-semester version of the Lede represents a promising "extended boot camp" in which students who have been accepted into a data journalism master's program may attend for a full summer before their peers in order to develop the skills that will help them get the most out of their education.

The following course descriptions were pulled on November 5, 2015, from: http://www.journalism.columbia.edu/page/1060-the-lede-program-courses/908

FOUNDATIONS OF COMPUTING

During this introduction to the ins and outs of the Python programming language, students build a foundation upon which their later, more coding-intensive classes will depend. Dirty, real-world data sets will be cleaned, parsed and processed while recreating modern journalistic projects. The course will also touch upon basic visualization and mapping, and how to use public resources such as Google and Stack Overflow to build self-reliance.
FOCUS: Familiarize yourself with the data-driven landscape
TOPICS & TOOLS INCLUDE: Python, basic statistical analysis, OpenRefine, CartoDB, pandas, HTML, CSVs, algorithmic story generation, narrative workflow, csvkit, git/GitHub, Stack Overflow, data cleaning, command line tools, and more

DATA AND DATABASES

Students will become familiar with a variety of data formats and methods for storing, accessing and processing information. Topics covered include comma-separated documents, interaction with website APIs and JSON, raw-text document dumps, regular expressions, text mining, SQL databases, and more. Students will also tackle less accessible data by building web scrapers and converting difficult-to-use PDFs into useable information.
FOCUS: Finding and working with data
TOPICS & TOOLS INCLUDE: SQL, APIs, CSVs, regular expressions, text mining, PDF processing, pandas, Python, HTML, Beautiful Soup, IPython Notebooks, and more

ALGORITHMS

Machine learning and data science are integral to processing and understanding large data sets. Whether you're clustering schools or crime data, analyzing relationships between people or businesses, or searching for a single fact in a large data set, algorithms can help. Through supervised and unsupervised learning, students will generate leads, create insights, and figure out how

to best focus their efforts with large data sets. A critical eye toward applications of algorithms will also be developed, uncovering the pitfalls and biases to look for in your own and others' work.

FOCUS: Analyzing your data

TOPICS & TOOLS INCLUDE: linear regression, clustering, text mining, natural language processing, decision trees, machine learning, scikit-learn, Python, and more

DATA ANALYSIS STUDIO

In this project-driven course, students refine their creative workflow on personal work, from obtaining and cleaning data to final presentation. Data is explored not only as the basis for visualization, but also as a lead-generating foundation, requiring further investigative or research-oriented work. Regular critiques from instructors and visiting professionals are a critical piece of the course.

FOCUS: Applying your skillset

TOPICS & TOOLS INCLUDE: Tableau, web scraping, mapping, CartoDB, GIS/QGIS, data cleaning, documentation, and more

WORKS CITED

Cohen, Sarah, Chengkai Li, Jun Yang, and Cong Yu. 2011. "Computational Journalism: A Call to Arms to Database Researchers." In *Proceedings of the 5th Biennial Conference on Innovative Data Systems Research*, Sailorman. Asilomar, CA: ACM.

De Maeyer, Juliette, Manon Libert, David Domingo, François Heinderyckx, and Florence Le Cam. Forthcoming. "Waiting for Data Journalism: A Qualitative Assessment of the Anecdotal Take-up of Data Journalism in French-speaking Belgium." *Digital Journalism*.

European Journalism Center. 2010. "Data-Driven Journalism: What Is There to Learn?"

Folkerts, Jean, John Maxwell Hamilton, and Nicholas Lemann. 2013. "Educating Journalists: A New Plea for the University Tradition." New York: Columbia Journalism School.

Hamilton, James T., and Fred Turner. 2009. "Accountability Through Algorithm: Developing the Field of Computational Journalism." Center for Advanced Study in the Behavioral Sciences Summer Workshop.

Howard, Alexander B. 2014. "The Art and Science of Data-Driven Journalism." Tow Center for Digital Journalism. A Knight Report. http://towcenter.org/wp-content/uploads/2014/05/Tow-Center-Data-Driven-Journalism.pdf

Karlsen, Joakim, and Eirik Stavelin. 2014. "Computational Journalism in Norwegian Newsrooms." *Journalism Practice* 8(1): 34–48. doi:10.1080/17512786.2013.813190.

Moretti, Franco. 2007. *Graphs, Maps, Trees: Abstract Models for Literary History.* London: Verso.

Pulitzer, Joseph. 1904. "Planning a School of Journalism – The Basic Concept in 1904." *The North American Review* (178:5).

Royal, Cindy. 2010. "The Journalist as Programmer: A Case Study of The New York Times Interactive News Technology Department." Presented at the International Symposium for Online Journalism, Austin, TX, April 23.

Schudson, Michael. 1978. *Discovering the News: A Social History of American Newspapers*. New York: Basic Books.

Schudson, Michael. 1996. *The Power of News*. Cambridge, MA: Harvard University Press.

Tenen, Dennis. Forthcoming. "Blunt Instrumentalism." In *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press.

UNESCO. 2007. *Model Curricula for Journalism Education*. http://unesdoc.unesco.org/images/0015/001512/151209E.pdf

Zelizer, Barbie. 1995. "Words Against Images: Positioning Newswork in the Age of Photography." In *Newsworkers: Toward a History of the Rank and File*, edited by Bonnie Brennen and Hanno Hardt. Minneapolis: University of Minnesota Press: 135–59.

ACKNOWLEDGMENTS

We could not have done this without the assistance of Maxwell Foxman and Joscelyn Jurich, two Ph.D. students at Columbia. Max and Joscelyn traveled far and wide, scoured the Web, and checked innumerable facts in order for this report to even come close to depicting the state of data journalism education. Their thoughtful memos, perceptive comments—and yes, meticulous spreadsheets—were vital contributions to this research.

Many of the insights in this report should be credited to our advisory committee: Sarah Cohen, Meredith Broussard, Steve Doig, Michelle Minkoff, Shazna Nessa, Jeremy Singer-Vine, Jonathan Stray, Matt Waite, and Derek Willis. Our committee met twice, first to launch the project and frame its mission, then seven months later to review our findings and help us to refine our conclusions.

A number of journalists and journalism teachers also offered in-depth interviews on their experience, the state of the field, and their vision for the future, and for that we thank them for sharing their time and their insights: Jonathon Berlin, Rahul Bhargava, David Boardman, R. B. Brenner, Matt Carroll, Ira Chinoy, Brian Creech, Catherine D'Ignazio, Nick Diakopoulos, David Donald, Jaimi Dowdell, Deen Freelon, David Herzog, Mark Horvit, Brant Houston, Mike Jenner, Dan Keating, Jennifer LaFleur, Darnell Little, Kathy Matheson, Tom McGinty, Philip Meyer, George Miller, T. Christian Miller, Maggie Mulvihill, Jonah Newman, Ben Poston, Kevin Quealy, Mike Reilley, Simon Rogers, Judd Slivka, Margot Susca, Ben Welsh, Aaron Williams, and Zach Wise.

Special thanks go out to those who invited us into their classrooms to watch how they teach. Catherine D'Ignazio of Emerson College offered our very first observation and also introduced us to her colleagues at the MIT Media Lab who are tackling similar questions about teaching data. Zach Wise and Larry Birnbaum invited us to observe their team-taught class at Northwestern. Amanda Hickman hosted us twice for her class at the CUNY Graduate School of Journalism. Meredith Broussard and George Miller each allowed us to observe their classes at Temple University. Likewise, Dan Keating graciously hosted an in-class observation at the University of Maryland.

We would also like to thank the students who spoke to us about their experience learning to use data in their reporting: Matt Bernardini, Simeng Dai, George Dumontier, Alex Duner, Jasmine Han, John Hilliard, Austin Huguelet, Ashley Jones, Anne Li, Mary Ryan, and Nicole Zhu. We wish them fruitful careers as they shape this field of practice.

Through the course of this bi-coastal research project, the Columbia and Stanford communities have provided an unimaginable level of support and inspiration. Mark Hansen, the East Coast director of the Brown Institute for Media Innovation, has articulated a remarkable vision for the union of journalism and computation, as well as a model for bicoastal collaboration.

Susan McGregor shared her deep understanding of journalism, pedagogy, and digital security at crucial stages of this project, in addition to hosting us in her classroom twice. Giannina Segnini, James Madison Visiting Professor on First Amendment Issue at the Columbia School of Journalism, offered her considerable expertise in the use of data for global investigative reporting. Those from Stanford who provided critical input and support include Hearst Professional in Residence Dan Nguyen, Peninsula Press Managing Editor Vignesh Ramachandran and Geri Migielicz, the Lorry I. Lokey Visiting Professor in Professional Journalism, who is co-teaching a virtual reality class in the winter of 2015.

Steve Coll expressed an elegant and convincing case for data journalism's place in contemporary practice—from which we borrowed unabashedly and at length. This report was substantially improved by Jay Hamilton, Sarah Cohen, Jonathan Stray, Jonathan Soma, and Chris Anderson, who read our first draft and offered remarkably useful feedback. We thank Marcia Kramer for the thought and care with which she edited this report.

We are indebted to the John S. and James L. Knight Foundation for funding this research. Knight's commitment to both the future of journalism education and to innovation in journalistic practice dovetailed in this project, and we hope that it is a worthy contribution to Knight's greater mission.

Finally, we thank the person who conceived of this project, recruited our team, and served as a sagely advisor at every step: Sheila Coronel, director of Columbia's Stabile Center for Investigative Journalism. Sheila recognized not only the urgency of developing a data curriculum based on an empirical study of the field, but also the value of sharing the results with other journalism educators. What could easily have been just a research study, or else just a curriculum development project, was envisioned by Sheila as something that could be greater than the sum of its parts.