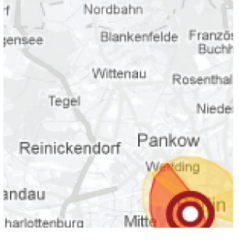
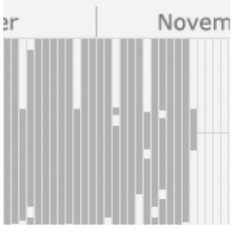
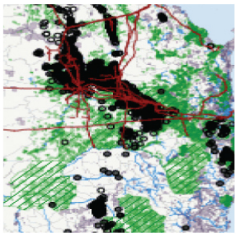
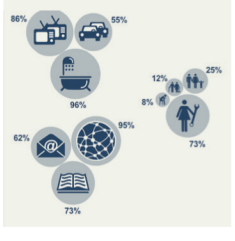
/Theory/In/Practice

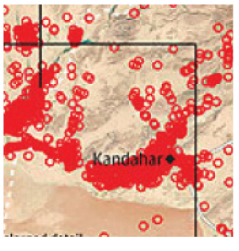
The Data Journalism Handbook

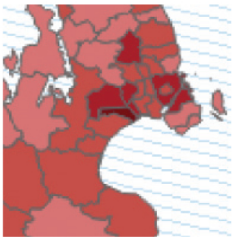
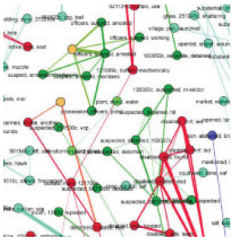
How Journalists Can Use Data to Improve the News

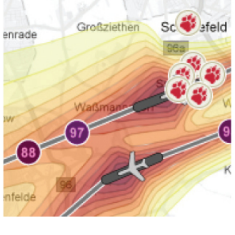



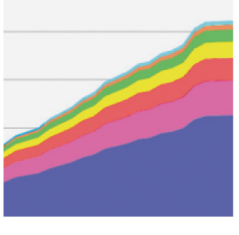
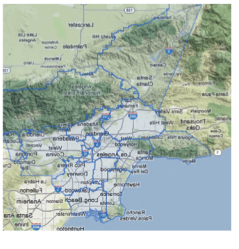
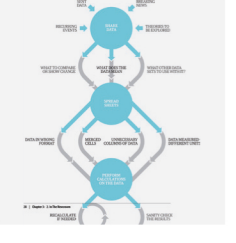
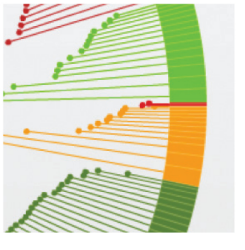







Edited by Jonathan Gray, Liliana Bounegru, and Lucy Chambers

O'REILLY®

Data/Journalism

The Data Journalism Handbook

When you combine the sheer scale and range of digital information now available with a journalist's "nose for news" and her ability to tell a compelling story, a new world of possibility opens up. With *The Data Journalism Handbook*, you'll explore the potential, limits, and applied uses of this new and fascinating field.

This valuable handbook has attracted scores of contributors since the European Journalism Centre and the Open Knowledge Foundation launched the project at MozFest 2011. Through a collection of tips and techniques from leading journalists, professors, software developers, and data analysts, you'll learn how data can be either the source of data journalism or a tool with which the story is told—or both.

- Examine the use of data journalism at the BBC, the *Chicago Tribune*, the *Guardian*, and other news organizations
- Explore in-depth case studies on elections, riots, school performance, and corruption
- Learn how to find data from the Web, through freedom of information laws, and by "crowd sourcing"
- Extract information from raw data with tips for working with numbers and statistics and using data visualization
- Deliver data through infographics, news apps, open data platforms, and download links

A project of the European Journalism Centre and the Open Knowledge Foundation

Purchase the ebook edition of this O'Reilly title at oreilly.com and get free updates for the life of the edition. Our ebooks are optimized for several electronic formats, including PDF, EPUB, Mobi, APK, and DAISY—all DRM-free.

Strata
Making Data Work

Strata is the emerging ecosystem of people, tools, and technologies that turn big data into smart decisions. Find information and resources at oreilly.com/data.

US $24.99 CAN $25.99
ISBN: 978-1-449-33006-4

Twitter: @oreillymedia
facebook.com/oreilly

O'REILLY®
oreilly.com

The Data Journalism Handbook
Edited by Jonathan Gray, Liliana Bounegru, and Lucy Chambers

A project of the European Journalism Centre and the Open Knowledge Foundation.

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

Editor: Shawn Wallace
Production Editor: Kristen Borg
Proofreader: O'Reilly Production Services

Cover Designer: Karen Montgomery
Interior Designer: David Futato
Illustrator: Kate Hudson

July 2012: First Edition.

Revision History for the First Edition:
2012-07-11 First release
See *http://oreilly.com/catalog/errata.csp?isbn=9781449330064* for release details.

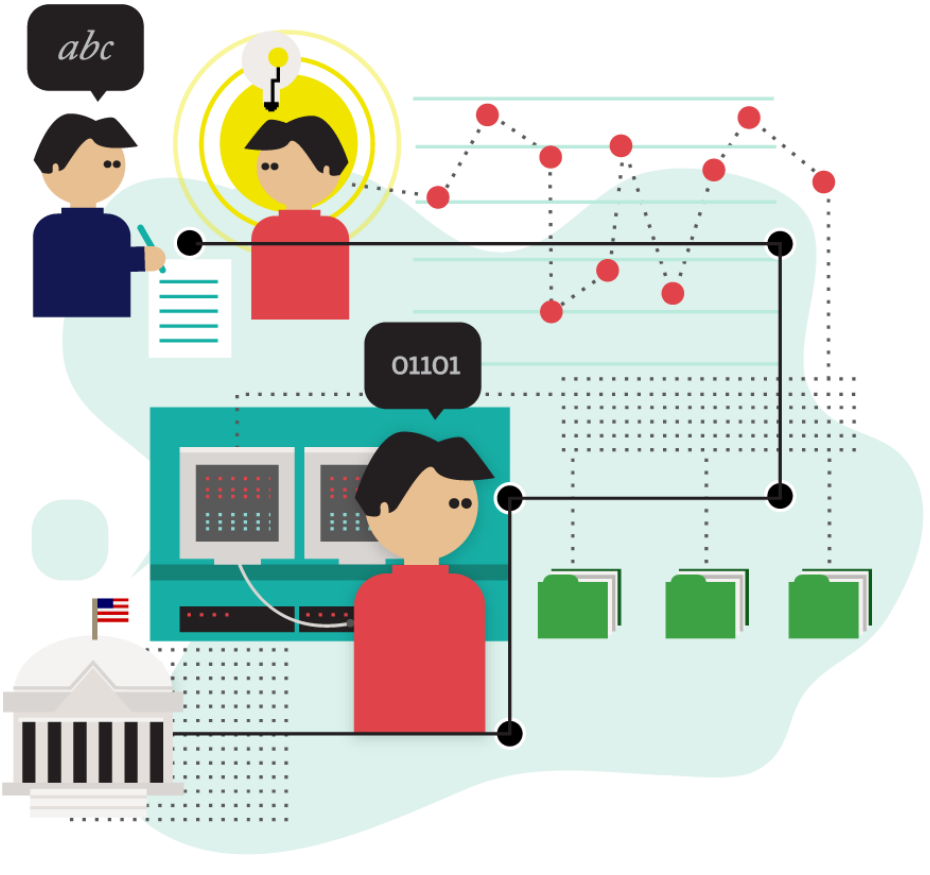
ISBN: 978-1-449-33006-4

[LSI]

1342026449

CHAPTER 5

Understanding Data



Once you've got your data, what do you do with it? What should you look for? What tools should you use? This section opens with some ideas on improving your data literacy, tips for working with numbers and statistics, and things to bear in mind while working with messy, imperfect and often undocumented datasets. We go on to learn about how to get stories from data, data journalists' tools of choice, and how to use data visualization to give you insights into the topic you're looking at.

Become Data Literate in Three Simple Steps

Just as literacy refers to "the ability to read for knowledge, write coherently, and think critically about printed material," data-literacy is the ability to consume for knowledge, produce coherently, and think critically about data. Data literacy includes statistical literacy, but also understanding how to work with large datasets, how they were produced, how to connect various datasets, and how to interpret them.

Figure 5-1. Digging into data (photo by JDHancock, http://www.flickr.com/photos/jdhancock/3386035827/)

Poynter's News University offers math classes for journalists, in which reporters get help with concepts such as percentage changes and averages. Interestingly enough, these concepts are being taught simultaneously near Poynter's offices, in Floridian schools to fifth grade pupils (age 10-11), as the curriculum attests.

That journalists need help in math topics normally covered before high school shows how far newsrooms are from being data-literate. This is a problem. How can a data journalist make use of a bunch of numbers on climate change if she doesn't know what a confidence interval means? How can a data reporter write a story on income distribution if he cannot tell the mean from the median?

A reporter certainly does not need a degree in statistics to become more efficient when dealing with data. When faced with numbers, a few simple tricks can help her get a much better story. As Max Planck Institute professor Gerd Gigerenzer says, better tools will not lead to better journalism if they are not used with insight.

Even if you lack any knowledge of math or stats, you can easily become a seasoned data-journalist by asking 3 very simple questions.

1. How was the data collected?

Amazing GDP growth

The easiest way to show off with spectacular data is to fabricate it. It sounds obvious, but data as commonly commented upon as GDP figures can very well be phony. Former British ambassador Craig Murray reports in his book, *Murder in Samarkand*, that growth rates in Uzbekistan are subject to intense negotiations between the local government and international bodies. In other words, it has nothing to do with the local economy.

GDP is used as the number one indicator because governments need it to watch over their main source of income—VAT. When a government is not funded by VAT, or when it does not make its budget public, it has no reason to collect GDP data and will be better off fabricating them.

Crime is always on the rise

"Crime in Spain grew by 3%," writes El Pais. Brussels is prey to increased crime from illegal aliens and drug addicts, says RTL. This type of reporting based on police-collected statistics is common, but it doesn't tell us much about violence.

We can trust that within the European Union, the data isn't tampered with. But police personnel respond to incentives. When performance is linked to clearance rate, for instance, policemen have an incentive to report as much as possible on incidents that don't require an investigation. One such crime is smoking pot. This explains why drug-related crimes in France increased fourfold in the last 15 years while consumption remained constant.

What you can do

When in doubt about a number's credibility, always double-check, just as you'd have if it had been a quote from a politician. In the Uzbek case, a phone call to someone who's lived there for a while suffices ("Does it feel like the country is 3 times as rich as it was in 1995, as official figures show?").

For police data, sociologists often carry out victimization studies, in which they ask people if they are subject to crime. These studies are much less volatile than police data. Maybe that's the reason why they don't make headlines.

Other tests let you assess precisely the credibility of the data, such as Benford's law, but none will replace your own critical thinking.

2. What's in there to learn?

Risk of Multiple Sclerosis doubles when working at night

Surely any German in her right mind would stop working night shifts after reading this headline. But the article doesn't tell us what the risk really is in the end.

Take 1,000 Germans. A single one will develop MS over his lifetime. Now, if every one of these 1,000 Germans worked night shifts, the number of MS sufferers would jump to 2. The additional risk of developing MS when working in shifts is 1 in 1,000, not 100%. Surely this information is more useful when pondering whether to take the job.

On average, 1 in every 15 Europeans totally illiterate

The above headline looks frightening. It is also absolutely true. Among the 500 million Europeans, 36 million probably don't know how to read. As an aside, 36 million are also under 7 (data from Eurostat; *http://bit.ly/eurostat-numeracy*).

When writing about an average, always think "an average of what?" Is the reference population homogeneous? Uneven distribution patterns explain why most people drive better than average, for instance. Many people have zero or just one accident over their lifetime. A few reckless drivers have a great many, pushing the average number of accidents way higher than what most people experience. The same is true of the income distribution: most people earn less than average.

What you can do

Always take the distribution and base rate into account. Checking for the mean and median, as well as mode (the most frequent value in the distribution) helps you gain insights in the data. Knowing the order of magnitude makes contextualization easier, as in the MS example. Finally, reporting in natural frequencies (1 in 100) is way easier for readers to understand than using percentage (1%).

3. How reliable is the information?

The sample size problem

"80% dissatisfied with the judicial system", says a survey reported in Zaragoza-based Diaro de Navarra. How can one extrapolate from 800 respondents to 46 million Spaniards? Surely this is full of hot air.

When researching a large population (over a few thousand), you rarely need more than a thousand respondents to achieve a margin of error under 3%. It means that if you were to retake the survey with a totally different sample, 19 times out of 20, the answers you'll get will be within a 3 percentage points interval of the value you would have found, had you asked every single person.

Drinking tea lowers the risk of stroke

Articles about the benefits of tea-drinking are commonplace. This short item in Die Welt saying that tea lowers the risk of myocardial infarction is no exception. Although the effects of tea are seriously studied by some, many pieces of research fail to take into account lifestyle factors, such as diet, occupation, or sports.

In most countries, tea is a beverage for the health-conscious upper classes. If researchers don't control for lifestyle factors in tea studies, they tell us nothing more than "rich people are healthier—and they probably drink tea."

What you can do

The math behind correlations and error margins in the tea studies are certainly correct, at least most of the time. But if researchers don't look for co-correlations (e.g., drinking tea correlates with playing sports), their results are of little value.

As a journalist, it makes little sense to challenge the numerical results of a study, such as the sample size, unless there are serious doubts about it. However, it is easy to see if researchers failed to take relevant pieces of information into account.

— *Nicolas Kayser-Bril, Journalism++*

Tips for Working with Numbers in the News

- The best tip for handling data is to enjoy yourself. Data can appear forbidding. But allow it to intimidate you and you'll get nowhere. Treat it as something to play with and explore and it will often yield secrets and stories with surprising ease. So handle it simply as you'd handle other evidence, without fear or favor. In particular, think of this as an exercise in imagination. Be creative by thinking of the alternative stories that might be consistent with the data and explain it better, then test them against more evidence. "What other story could explain this?" is a handy prompt