mdif
media
development
investment
fund

KUANG KENG
KUEK SER

Best Practices for Data Journalism

Table of Contents

BEST PRACTICES FOR DATA JOURNALISM
13 COMMANDMENTS

1 When setting up your data team, it is absolutely ok to start small. The diversity of skills matters more than team size. However, avoid having a single data journalist and expecting that person to do everything.

2 All data team members, regardless of their discipline, are part of the news desk. They should be treated as journalists and sit within the desk to encourage communication and collaboration.

3 The data team should not be seen as a "service desk" called in to help solve a technical issue or add a feature midway through a project. They should be involved in the project as early as possible.

4 When providing data journalism training for your newsroom, include managers and editors at least in the introduction or overview part. You want them to empower the journalists and be able to edit data stories.

5 Always question the quality of your data. Apply the same journalism ethics and standards - such as accuracy and accountability - to data-driven stories.

6 Before doing anything to the dataset, keep a copy of the original in case you make mistakes and lose access to the data source.

7 When dealing with a data sample, scrutinize its size, sampling method, margin of error and confidence level to avoid misleading your audience.

8 Be extra careful in making cause and effect conclusions among data points. Correlation between two variables does not mean causation.

9 The goal of data visualization is to communicate information effectively. A common mistake among designers is aiming for powerful graphics rather than powerful journalism.

10 Perform user-testing on your design, and redesign it based on user feedback. User-testing is not necessarily expensive or resource-intensive.

11 Avoid using pie charts to present your data as they are often less effective and widely misused.

12 Journalism is about people. Don't put numbers or visualization above people. Humanize your data by telling the stories of people who have been affected by the data.

13 Don't hide your data. Open your data to the public, including your competitors. Exposing data to experts and other users allows more findings and insights to be extracted.

HOW TO USE THIS GUIDE?

If you are a newsroom manager

If you plan to set up a data journalism team, or want to make your existing data journalism desk more efficient, **Sections 1-4** are most relevant to you. Section 7, on monetization and data journalism business models, and parts of Section 8, *Building Data Journalism in Newsroom*, should also be of interest.

If you are a journalist

If you're an aspiring data journalist, or you're already telling stories with data but looking to do this better, focus on **Sections 1-4**. Also, Sections 5-6 explain the techniques and tools for data journalism. For further reading, practice and external support, check the resources in Section 8.

If you are an editor

Even if you don't do data analysis or create visualization, your role in guiding and bulletproofing data-driven stories is vital. **Sections 2 and 4** will help you guide your data journalists. Parts of Section 5 - *Data Integrity and Bulletproof Your Data* - will help you ask the right questions when editing stories. For editing graphics, please refer to the good data visualization elements under *Visualization* in Section 6. There are also several resources specifically selected for editors and an online course *Bulletproof Data Journalism* listed in Section 8.

Data journalism stories by Pajhwok Afghan News

Pajhwok Afghan News is Afghanistan's largest independent news agency and the country's sole media outlet to use data to inform the public. It uses simple data visualization with strong narrative to appeal to an audience with low data and digital literacy, and is often cited by advocates as an example of how data journalism can be done even under extremely difficult conditions.

Coal mine closure investigation by Oxpeckers

In an award-winning series, the Oxpeckers Center for Investigative Environmental Journalism, Africa's first journalistic investigation unit focusing on environmental issues, used data heavily in its reporting. The unique data it collected revealed, for example, that since at least 2011 no large coal mines operating in South Africa have been rehabilitated, and are simply abandoned. The center has two full-time staff and works with more than a dozen associate journalists.

High and Dry by Nepali Times

The use of stunning images and an engaging video make this story stand out, and, while data is not the main focus, its inclusion and the use of visualization help to highlight the gravity of the erratic rainfall. Nepali Times, an English weekly newspaper based in Kathmandu, has no dedicated data team. Its small editorial team, including a self-taught data/online producer, works together to produce interactive and data stories.

WHY DATA **JOURNALISM?**

Briefly, data journalism is a set of approaches, skills and tools for using data for journalistic purposes.

> "Data can be the source of data journalism, or it can be the tool with which the story is told - or it can be both," says data journalism guru Paul Bradshaw in Data Journalism Handbook.

Data journalism came to prominence in the 2000s due to several factors:

- The amount of digital data generated and made public rose exponentially.
- The increased use of the Internet by governments and the public.
- The global open data movement that contributed to the increased publication of government data.
- New computational tools lowered the barrier for journalists to use data in their work.

To be clear, data journalism does not replace traditional journalism, but rather complements and enhances what journalists have been doing for centuries.

Media organizations and journalism schools have invested in data journalism because it has been proven to:

- Find stories that would not have been found through traditional reporting.
- Find insightful or important stories hidden in data.
- Verify or clarify claims more authoritatively with evidence
- Tackle bigger stories that involved a huge amount of information or data.
- Communicate information quickly, effectively and memorably.
- Set your reporting apart from your competitors.
- Engage the audience in more innovative and personalized storytelling approaches.
- Increase audience engagement through social media-friendly visualization.

GETTING **STARTED**

It's fine to start small.
The Los Angeles Times news app team was just two people for a long time, and The Guardian data projects team in the UK has only three journalists.

However, it's not advisable to have just a single data journalist and expect him/her to do everything. Data journalism requires multiple skills from different disciplines.

The diversity of skills is more important than the size of the team. The key skills are Code, Design, and Journalism.

For example, a 2-member team may consist of a reporter who can code and a designer; a programmer-cum-designer and a reporter; or other combinations. There are also data teams with limited coding capacity that work with an organization's IT department.

If you plan to hire externally for the team - whether for a reporter, programmer or designer - stress the need for an ability in critical thinking. Essentially, the person is a journalist. Look for generalists with skills in at least two of the Code, Design and Journalism areas, and be willing to teach them the third. The easiest to teach is coding.

Potential candidates should be "aggressively collaborative" as the team will need to work with members across your organization. The knowledge or skills needed in data journalism are evolving and may vary from region to region. The following are some popular skills used in data journalism to help managers hire the right candidates.

DATA CRUNCHING	Spreadsheet (Excel), SQL, statistics, OpenRefine
CODING FOR WEBSITE/APPLICATION	HTML, CSS, Javascript, PHP
CODING FOR DATA	R, Python, D3, Google Map/Earth API
DESIGN	Adobe Illustrator, mapping softwares like QGIS, data visualization tools like Highcharts, Tableau and Carto, infographics tools like Infogr.am.

Alternatively, you can build the team by training existing staff. Look for reporters who use spreadsheets in their work and present numbers effectively in their stories. A background or interest in mathematics is helpful. The team's programmer or designer should have a passion to deploy their skills for journalism and be prepared to be involved in reporting.

How the Team Works

All team members are part of the news desk and should sit within it to encourage communication and collaboration. They should be treated as journalists and attend editorial meetings, interview sources and be given bylines. The coder and the reporter should be encouraged to share and learn skills from each other - so the coder thinks like a reporter when analyzing data or designing a news app. All team members are responsible for all the work they produce. This is a big shift from traditional media operations where IT departments mostly work separately from the news desk. Data team coders should not be 'lent' to the IT department to work on platform work like the CMS or website redesign. Keep them focused on the data team's journalistic projects.

The team should both produce its own data-driven stories and collaborate with other journalists on their reporting. Producing the former generally takes more time and resources, but has a stronger impact - such as an investigation into deaths by the police or climate change data. The team can work with specialist journalists or with outside experts.

Support and collaboration from news desk videographers and photographers is important for story presentation. These are stories that can set your reporting apart from your competitors, and prove the value of your team to colleagues and management.

Stories in the second category require a shorter turnaround time and are smaller in scale - such as a map for a flood story or a series of charts for a report on poverty. The stories are led by the journalists, and the role of the data team is to use data to present the story more effectively. These are good opportunities for the data team to familiarize the newsroom with their work, and helps raise data literacy overall.

The data team should not be seen as a "service desk" that is called in to help solve a technical issue during a project or asked to create graphics at the end of the reporting process. It should be involved in the project as early as possible.

If the team has no editor, the news desk should assign one with a background in mathematics and design to guide the team. The editor is responsible for finding and picking the team's projects, fact-checking and editing, evaluating the team's performance and progress, and securing resources. As the team grows, it should have its own data editor or senior data journalist as team leader.

Team members should be given self-learning time in their work schedule so they can expand or strengthen their skill-sets. This is especially important for teams formed by existing staff or those in countries where the talents mentioned previously are less than abundant.

What Project/Stories to Start With?

The team can start on two fronts: graphics and data.

For graphics, cut out those that are ineffective (dull), unnecessary (simply break up the text) or that add little or no value (don't help tell the story more effectively). Focus resources on improving the quality of essential visual elements, and optimize those graphics for platforms where your audience is most active. The impact of good graphics can be monitored quickly and easily through simple metrics such as the number of shares and likes on social media.

In newsrooms with no data culture, it can be a 'tough sell' to convince journalists to use data in their reporting. Building a strong and well-received data-driven story shows the newsroom the value that data can bring, spurring interest in data journalism.

A quick way to kick-start a data project is to pick an issue where data is readily available, especially in countries where government data is not public. Another consideration is to piggy-back on similar projects by media in other countries so your team can reproduce or innovate based on them. Issues like healthcare, government budget/spending and voting are some low-hanging fruit. Choose one with broad public appeal and a relatively long shelf life to maximize the impact. (more on this in Section Data Journalism...).

TOWARDS SUSTAINABLE DATA JOURNALISM **IN NEWSROOMS**

Grow Your Team

Beside producing content, a data team can be a lead innovator and data evangelist in the newsroom, so it's crucial to ensure it is growing in both intelligent and technical capacity.

If you operate in regions where there's a vibrant data journalism community - such as the United States, Europe or Latin America - there are plentiful external resources. Conferences such as those organized by the National Institute for Computer-Assisted Reporting (NICAR), workshops by outside contractors or local journalism organizations like the Investigative Reporters and Editors (IRE) in the US, mentorship by other data journalists and partnership with specialist journalism groups like the Center for Public Integrity are great ways to upgrade your team's firepower. See Resources section for more such resources, including fellowships and networks.

In countries where there is little data journalism, peer learning, mentorship and external training might be less easy to come by. This is where partnership with outside experts and hosting events can be helpful.

University academics and think-tank researchers are good partners for data projects as they have both data and expertise, and many are keen to work with the media to increase public awareness on a subject and boost publicity for their work.

Civic hackers - programmers, designers or data scientists who deploy their professional skills to solve civic problems - are also potential partners (and a good source for future hiring). You can find them in non-profit technology advocacy groups like Code for Africa or media-tech groups like Hacks/Hackers, a global movement with over 100 local chapters in most major cities that aims to marry technology with journalism. Sending your team members to participate in journalism-themed hackathons can be an effective way to expand their knowledge and grow their network. The Global Editors Network's (GEN) Editors Lab program is a worldwide series of hackathons held for newsrooms.

When designing internal training on data journalism, consider the practices below:

1. Pitch data skills as less work.
2. Understand participants' workflow and the challenges they face in learning new skills. Select data skills that fit their workflow and make it more efficient.
3. Customize training for your newsroom as every newsroom is different.
4. Prepare several sessions for different groups of journalists. You are unlikely to be able to train everyone at the same time.
5. Customize training for different groups as their roles and specialized areas are different.
6. Use participants' stories as examples or in exercises.
7. Use real data relevant to topics covered by the participants.
8. Start with tools/skills that have a low learning curve and will be used frequently.
9. Prepare step-by-step tutorials with visuals that speak the participants' language for them to refer to when they want to deploy their newly learned skills.
10. Include editors at least in the introduction or overview part. You want them to empower the journalists, who will be pitching data stories.

In markets where media organizations and journalists are reluctant to share knowledge and exchange information, data journalists can tap into an international network of digital journalists. The US-based IRE manages several mailing lists, including the most popular NICAR-L listserv, which is dedicated to subjects related to computer-assisted reporting. It's open for all journalists, and many subscribers are based outside the US. Requests for technical solutions and advice get a swift response from other subscribers. Another international online forum is the GEN's Data Journalism Awards Slack, a group launched in 2016 with over 350 members worldwide.

Don't forget the developer members in your team.

Without journalism experience, they might lack the ability to put the data into a broader political, social and economic context and see trends that matter to readers. They also need storytelling skills. Ideally, you want them to understand best journalistic practices like accuracy, fact-checking and accountability, be able to generate story ideas from data, study the topics and write stories.

Encouraging and Nurturing Data Culture in the Newsroom

Forming a data team is the first step towards establishing data as an integral part of newsroom culture and practice.

Data journalism practitioners call this building a "data state of mind" among newsroom members. This is about the newsroom being aware of what data is available to help strengthen its reporting, and how to access it. The more members of the newsroom with such a mindset, the more data becomes a key part of the process, and the higher the reporting quality. This can also free up time and resources to focus on more sophisticated and impactful projects.

There is no one-size-fits-all antidote when it comes to spreading data literacy across the newsroom. Once the data team has adapted to its workflow and built visibility within the newsroom, it can start promoting data journalism to newsroom colleagues.

At The Guardian, the Data Projects team holds workshops for colleagues who are interested in learning Excel. It's important to target only those most open to learning new skills and most likely to actually use them.

La Nacion, a data journalism pioneer and leader in Latin America, suggested that everyone in the newsroom, including editors, should be equipped with basic spreadsheet knowledge as this is the foundation that serves as a "gateway" to other forms of data journalism.

Some newsrooms found that providing a user-friendly tool for journalists to create their own charts makes them more independent in working with their own data, and reduces the burden on colleagues with specialized graphic skills. The open-source Chartbuilder created by Quartz, a US-based online media, is probably the charting tool that created the most buzz. With Chartbuilder, Quartz said its reporters "published thousands of charts that make their stories more compelling and the data they're writing about more understandable".

Many publications, including The Wall Street Journal, CNBC and the Australian Broadcasting Corporation (ABC), have adopted the tool and customized it to suit their own use. Some were inspired by Chartbuilder and built their own custom charting tool. The set-up of Chartbuilder requires someone with moderate coding skills, but once it's up and running you can produce a chart in 30 seconds with no coding knowledge. You can also use the cloud version hosted on GitHub without having to install it.

Many successful data teams have produced stories that have been a "game-changer" in shifting their newsrooms' attitude to data journalism. Such stories often look into topics that people care about and feature personalization that connects the data with individual users.

For example, The Sun Sentinel, a local daily newspaper in Florida, published a database on the results of nail salon inspections around Miami that allows users to search the reports of their nail station. It attracted hundreds of thousands of hits to the newspaper's website. Three months later, the same model was used on restaurant inspections and drew millions of hits. As a result, the newspaper hired a new developer to work exclusively with investigative reporters to produce more data-driven stories. (report - integrating data journalism)

German media outlet Zeit Online had its game-changer story during the Fukushima nuclear disaster in 2011. It published an interactive map showing the concentrations of Germans living at varying distances from nuclear plants. The map went viral on social media and brought in huge traffic.

Data teams should recognize the characteristics of such stories - topics of broad appeal, less time-sensitive issues, impactful and personalized - and use them to guide their own production.

Collaborative Newsroom

There are certain underlying factors behind successful data journalism.

A research report from the Storybench project at Northeastern University's School of Journalism published in May 2017 interviewed 72 digital journalists, including data journalists, in leading digital newsrooms around the world, and found three themes of best practice behind their successes: team-based collaboration, an open source ethos within and between newsrooms, and mobile-driven story presentation.

The first theme is fundamental for the sustainability of data journalism in a newsroom, while the second could contribute to the development of digital journalism in a region.

You can employ talented people with diverse skillsets but more important, according to the report, is to "construct collaborative environments in which these players can work together as a team". Traditional newsroom operations tend to segregate news production into departments like design, photo and research, and organize the workflow in which ideas and information always move in one direction. By contrast, the successful digital newsrooms interviewed allow "nimble, multifaceted teams to self-organize, coming together organically to produce an editorial product".

A more collaborative newsroom allows ideas from different disciplines to collide and spark. Below are some practices widely found by many leading digital newsrooms to be effective in forging collaboration.

1. Sit programmers, designers and journalists together. Proximity works.
2. Use an instant communication tool like Slack for the whole newsroom.
3. Have 15-minute daily stand-up meetings, where those in the newsroom share briefly what they did yesterday and what they plan to do today. This keeps everyone on the same page and finds areas for collaboration.
4. Break down perceptions: i.e. developers and designers should not "serve" journalists. They are an integral part in finding and presenting stories and designing how users interact with them.
5. Have regular skill sharing or show-and-tell sessions to foster appreciation for each other's work. Attendance should be voluntary.
6. Encourage developers and designers to write, and journalists to code.
7. Involve everyone early in the project to build a shared sense of ownership.
8. Start with projects on less time-sensitive issues to experience the collaborative process without rushing to meet a deadline.
9. Use Google documents or a project management tool like Trello to track progress.

DATA JOURNALISM STRATEGY FOR **NEWSROOMS WITH LIMITED RESOURCES**

Journalists in countries where there is little or no data journalism often think that data-driven stories require vast resources. Reading behind-the-scenes stories from data-driven projects by major news outlets can strengthen these perceptions as those projects often involve months of work and a large team.

Successful data journalism by small or resource-challenged newsrooms usually doesn't get as much publicity. But you can find plenty of them among South American media, local US newspapers and nominees for international data journalism awards.

Topic Selection

Knowing how to pick your battles is crucial.
You should invest your resources in projects more likely to bring the highest return - based on journalistic impact and audience engagement. Projects should have an effect on individuals and society. They can cover policy changes or a rise in public awareness, and they should engage a broad audience that can be measured by various analytics.

The key to producing such projects is story selection. Below are some characteristics that data teams should look for when conceiving projects.

Evergreen or recurring topics

Evergreen topics are not time-sensitive and attract continuous public attention, such as election campaigns, terror attacks, refugee crises, the immigration debate, affordable housing, health issues, police violence and corruption. Recurring topics include government annual budgets, economic indicators like unemployment, elections, festivals and food hygiene.

Projects on evergreen topics have a longer shelf life than event-based stories, and can generate

Financial plan
INVOICE
Business Company
BUSINESS
Economy of the European Union

more traffic over the long-run than major breaking news. They can be remarketed when the topic is back in the news or serve as backgrounders for related stories. Projects on recurring topics can be ‘recycled’ with updated data. In short, they are more cost-effective.

Topics with broad and strong public interest

Avoid becoming obsessed with your data or story idea. If the subject matter is of little interest to most people, the project won’t go far, even if it involves in-depth investigation or great presentation. Topics that people care about include (urban) traffic, money, crime, jobs, and health issues like air pollution or garbage collection. For specialist, or niche, media, it means topics that resonate with most of your audience.

Data that can be personalized

One advantage of using data in reporting is the ability to connect the subject matter with individual users. Data personalization is highly engaging because users can access information directly related to them.

Examples are news apps that allow users to view crime records in their neighborhood or websites that compare users’ income with their parents’ generation.

Don’t forget to look for data stories by other media outlets on the same topic. Know the datasets they used, the data analysis and processes they did and what results they achieved. That can help in your project planning.

Build Reusable and Easy Tools

Most data-driven stories are better told through interactive or visual components, or a different website layout that is unavailable in traditional Content Management System (CMS). Sometimes a news app is the best way to deliver data to users (more on this in chapter Telling Stories with Data).

When such a need arises, the data team will ask developers to build what’s required. Over time, the team will find some components or presentation formats are used repeatedly. This is when tools and templates should be built for journalists to create the components or plug in different types of content into the layout. The user interface should be easy enough for journalists to edit the content directly without touching any code.

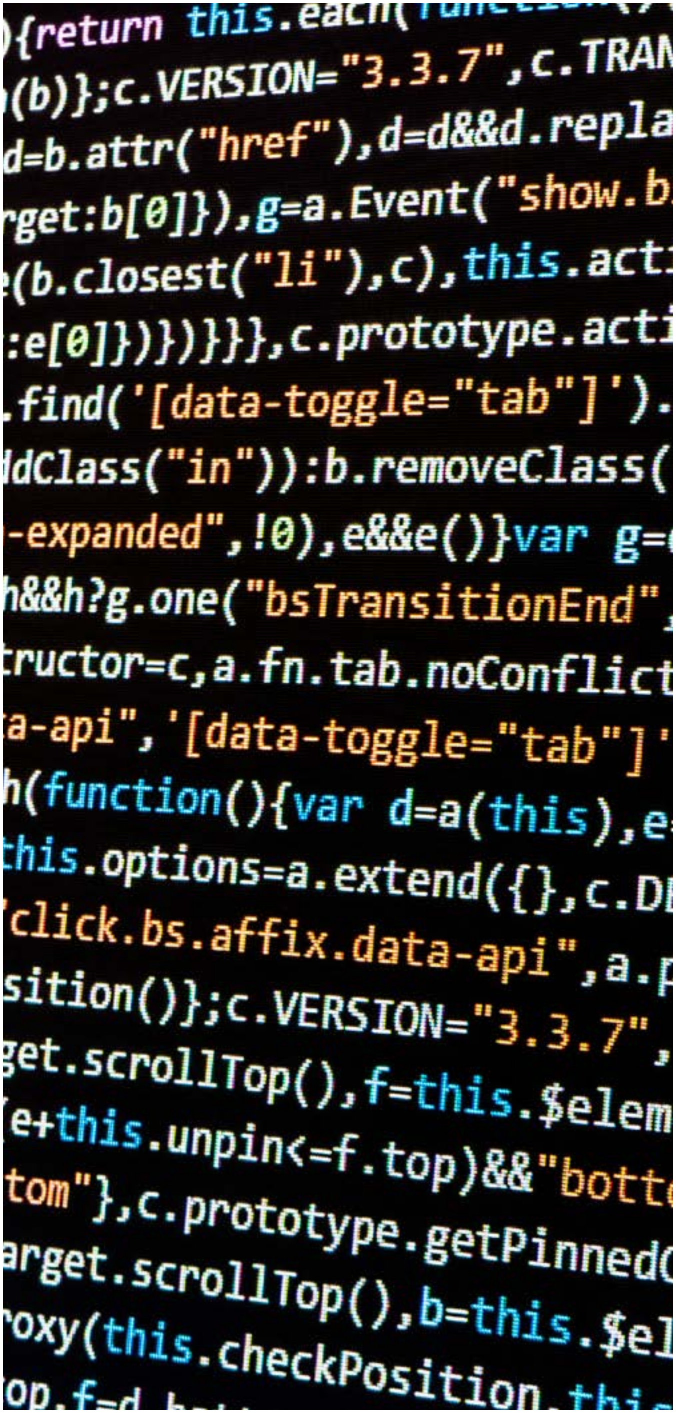
Such tools and templates can shorten the production time and free up developers to focus on more important things. Avoid investing resources in developing something that is rarely used.

Data visualization components often used in data-driven stories include timelines, interactive charts and maps, sortable and searchable data tables, and grid image galleries. Data teams should have access to tools that can create these components easily and quickly.

There are open-source publishing tools that make building templates simpler and faster. Check out Tarbell by the Chicago Tribune New Applications Team which allows editors to update a page through a Google spreadsheet instead of copy/paste code. Its latest version contains documentation and tutorials. The Al-Jazeera America newsroom used the Tarbell template 125 times in 15 months to make special projects.

If your newsroom has no developer or IT team, you might want to hire a journalist-cum-developer as one of your data team members. Refer to Section Getting Started.

MORE MEDIA OUTLETS, INCLUDING NPR, PROPUBLICA AND WNYC, HAVE MADE THEIR TOOLS OPEN SOURCE.
NPR
PROPUBLICA
WNYC



```
){return this.each(
(b)};c.VERSION="3.3.7",c.TRAN
d=b.attr("href"),d=d&&d.repla
rget:b[0]}),g=a.Event("show.b
(b.closest("li"),c),this.act
:e[0]})})}}},c.prototype.acti
.find('[data-toggle="tab"]').
ddClass("in")):b.removeClass(
-expanded",!0),e&&e()}var g=
h&&h?g.one("bsTransitionEnd",
ructor=c,a.fn.tab.noConflict
a-api",'[data-toggle="tab"]'
h(function(){var d=a(this),e
this.options=a.extend({},c.D
'click.bs.affix.data-api",a.p
sition()};c.VERSION="3.3.7",
get.scrollTop(),f=this.$elem
e+this.unpin<=f.top)&&"bott
tom"},c.prototype.getPinned(
arget.scrollTop(),b=this.$el
oxy(this.checkPosition,thi
op,f=d.h
```

DATA JOURNALISM **WORKFLOW**

In this document, we use a fairly broad definition for a data-driven story. It can be a quick turnaround short story such as a report using economic data to rebut an official's claim that a nation's economy is strong. It can also cover investigative reporting that relies heavily on data, like the Panama Papers, and news apps that allow users to better access data. They may look very different and require varying resource levels, but they have a common thread - without the data, there's no story.

Why Do You Need Data?

Before searching for data, ask yourself what it's for. Following are some common factors.

You have a question that needs data to answer, or a hypothesis that needs data to prove or disprove.

You want to know whether criminal activities in a neighborhood have risen or fallen after a special crime-fighting program was implemented 5 years ago.

You want to know whether more people were killed by police this year compared to previous years.

You want to fact-check a government official's claim that a big street protest this year led to a drop in the number of foreign tourists.

You have a dataset that needs questioning.

Your local city hall makes public the data of restaurants' food hygiene ratings.
The government just released the latest census data. You want to find stories from that data.
You received a large number of leaked documents related to government contracts.

You have a dataset that is useful for the public.

The government published all its tender results online. You want to compile them into a user-friendly database for the public to access.
One of Texas Tribute's most popular data-driven

stories is the Government Salaries Explorer, which allows users to review the pay of more than 500,000 government employees in Texas.

Something is better explained with data.

You want to use election data to explain the issue of gerrymandering.

You want to use global data to explain the issue of climate change.

If you can frame your question, hypothesis or reason in more specific terms, this helps make your reporting more efficient. Instead of "Has crime risen or fallen in Gotham City?" use "Has the intentional homicide rate risen or fallen in Gotham City despite the city tripling its crime-fighting spending in the last 3 years?"

Where to Find Data?

Government data

Most data used by journalists is government data, and most of that is published by governments' statistical departments. Wikipedia has a list of national and international statistical services. Governments that have an open data policy usually have an open data portal that serves as a one-stop center for all government data. Here is a list of over 2,600 open data portals.

Journalists can file freedom of information (FOI) requests to get government data in countries with FOI laws. In countries where government information is closely guarded, accessing data can be challenging, but not impossible. There are many success stories in South America: LA Nacion in Argentina, OjoPúblico and Convoca in Peru, and Poderopedia in Chile (now expanded to Venezuela and Columbia) are some media organizations that produce great data journalism without an open data environment.

Sometimes, the data you're looking for is buried deep in government websites. Other times, it may be scattered in government reports published by different agencies. An investigation by La Nacion on bus subsidies in Argentina is a good case of searching data on different government websites.

When government data is not accessible or is unreliable, journalists should look into other data sources.

International/regional data portals

Many international bodies like the World Bank and the United Nations have their own open data portals. Some investigative journalism organizations also build international databases for fellow journalists, like the Investigative Dashboard and OpenCorporates.

Buzzfeed data editor Jeremy Singer-Vine maintains a weekly newsletter of useful datasets for journalists. These datasets are archived in a Google spreadsheet.

Journals and research reports

Papers published in journals or research papers published by think-tanks or academics on the topic you are reporting can lead you to the data you need. Even if the reports don't include the data, they often reference the data source.

Academics and universities

They collect government data and sometimes their own data related to their subject matter. They are partners who can provide expertise on your data analysis and reporting, and are often incentivized to work with the media as it helps publicize their work and institution.

Professional bodies and trade associations

Many professional bodies - like boards of engineers, bar councils and medical practitioner associations - are entrusted with certain powers to safeguard the ethics and quality of their professions. They collect and maintain data on their members, including location, qualifications and disciplinary records. Trade associations share the same practice. The more established ones also collect data to conduct research regarding their industry.

NGOs and advocacy groups

NGOs might collect more and better data than governments on their subject matter.

Crowdsourced data

Asking your audience to provide you with data is usually a last resort on an issue your audience cares about but where no-one has collected any data. It can, though, be a challenge to verify crowdsourced data. However, the benefits of filling the data vacuum or complementing existing incomplete or flawed data can outweigh the shortcomings.

For example, US news website ProPublica launched a project in August 2017 to collect hate-crime data from victims because there is no reliable national hate-crime data. NGO France Libertés built a dataset of tap water prices across France by collecting over 5,000 scanned utility bills from consumers.

Crowdsourced data journalism is not just about asking users for data points. Some successful crowdsourced data journalism projects asked users to help on tasks that can't be computer-automated. The Guardian, for example, asked readers to sieve out important information from some 700,000 documents in the MP's Expenses project. La Nacion's investigation into the death of a prosecutor asked selected users to help listen to over 40,000 audio recordings and categorize them.

API (Application programming interface)

Many online services offer part of their data or content to the public for free. You need a developer to request the data through an interface called API that enables interaction between software or applications. Popular online apps like Twitter, Uber and Waze have good documentation on how to use their API to access their data. Using Uber API, a report in The Washington Post found that the ride-hailing app in Washington D.C. provides a better service in areas with more white people.

Tools to Access Data

Below are some tools that increase the efficiency in finding, saving and converting data into machine-readable format (structured data such as CSV, JSON or spreadsheet).

Google Advanced Search

Makes your data searching quicker and easier, unearthing information buried in websites.

» Google Advanced Search tutorial

Data scraping tools

Use a technique to extract data from websites and save it in your designated destination in a structured format (usually spreadsheet).

DocumentCloud

If you are reporting on or publishing primary source documents, this is a good tool to search, analyze, annotate and highlight data in your documents. It's free for journalism organizations, but you have to apply for an account.

DocumentCloud

Converting PDF to spreadsheet

Content in a PDF document is not structured data, and it can be a challenge to convert the content into structured data like a spreadsheet so it can be analyzed and processed.

Tabula - free and open source tool designed for journalists.

Cometdocs - free until a limit
However, these only work for PDF documents with digital material e.g., text created with software. For non-digital material like scanned documents, photos or hand-written documents, you need OCR (Optical Character Recognition) software to convert into digital material.

Data Integrity

FOR NON-CODERS:

DataMiner

Import.io

Google Spreadsheet (tutorial)

OutWit

FOR CODERS:

Python with the Beautiful Soup library

RESOURCE:

Scraping for Journalists by Paul Bradshaw

» More tips on PDF conversion

Once you have your data, question its quality - just as you research the sources who provide information for your story. You want to make sure the data meets the usual journalistic standards such as accuracy and clarity.

Your data might not be able to answer all these questions, but that doesn't mean you can't use it. The decision to use or not is an editorial one, just like whether you want to use the information provided by your sources. The circumstances vary, but what every data journalist can and should do is to be open and honest to their users. Let them know the weaknesses and limitations of your data. Explain your decision to use that data if necessary.

FURTHER READING ON THIS TOPIC:

Evaluating data quality, Columbia Journalism School

Data Biography

A good practice is to build a data biography - a document to record all the answers to the questions above. Data science and information design expert Heather Krause has shared a data biography template online.

FURTHER READING ON THIS TOPIC:

Data Biographies: How To Get To Know Your Data

The following questions could help you evaluate the data.

1. Is the data accurate? Does it reflect the real-world objects?
2. Do you have enough data? Is all necessary data present?
3. Did you get the data for the relevant time period? What is the most recent data? If newer data is not available, why not?
4. Are all data values clearly defined by the data provider? Does it come with a data dictionary/codebook?
5. Is there another source that collects similar or parallel data? Which is more reliable? Why are they different?
6. Who collects, processes, publishes and maintains the data? They could be different parties.
7. Why, when and how was the data collected? The motive might lead to biased data.
8. Did the collection method change over time? Are data points at different times comparable?
9. Does the data come with disclaimers, limitations or assumptions?
10. Who has used the data before? What were the analyses made using the data?
11. If you want to compare between data points, are they apples-to-apples comparisons? For example, the definition of underage child varies in different countries.

Data Cleaning

When you decide to use the dataset and have it in machine-readable format, you need to make sure the data is clean. Keep a copy of the original dataset in case you make any irreversible mistakes and lose access to the data source. Rename the files with a set of naming standards and store them together with their data biography and codebook so they can be easily retrieved.

When you're about to start cleaning or processing your dataset, it is a recommended practice to log every step. This not only allows your colleagues or editors to 'bulletproof' your data and reporting, it also gives you a clear path to trace and reverse the processes. It can be a useful reference when you deal with similar datasets in the future.

Data cleaning often means removing errors as below:

- Misspelling or discrepancies - e.g. "Mohamad" and "Mohd.", "Burma" and "Myanmar"
- Bad data - e.g. negative values in age data
- Blank fields and missing data
- Duplicates
- Totals differ from aggregates
- Names are in different orders - e.g. Chinese put surname (family name) before first name

FURTHER READING ON THIS TOPIC:

The Quartz guide to bad data

Data cleaning is crucial when you are dealing with raw data input by a human. If your data has been compiled or processed by a secondary source or a data warehouse like the World Bank or statistical departments, it's usually clean.

There are many available data cleaning tools. The most common non-coding tools among data journalists are spreadsheet programs (Microsoft Excel, Libre Office Calc, Google Sheets) and OpenRefine, an open source tool designed for cleaning and transforming messy data. OpenRefine can also analyze and visualize data. Most of its features require no or minimal coding knowledge.

Data cleaning also involves transforming data:

- Split data into more than one data point - e.g. first and last name, street name and postal code.
- Transform data points into the same unit - e.g. U.S. customary unit and SI unit, GDP (nominal) and GDP (PPP)
- Transform money values so they have a common time basis.
- Transform date and time into consistent formats.
- Transform the data structure in a spreadsheet to a proper layout e.g., only one row for header, one data point in each row.

TUTORIAL

Cleaning Data in Excel

OpenRefine tutorial by Dan Nguyen

FOR PROGRAMMERS, PYTHON AND R ARE THE COMMON PROGRAMMING LANGUAGE FOR DATA WRANGLING.

Python tutorial (online book)

Python tutorial (websites)

Finding Stories from Data

We have already covered *(Why Do You Need Data)* several reasons that prompt you to use data in your reporting. This part discusses the technical details involved to 'interview' your data to find the answers you want.

Note: There are many data analysis tools, including more advanced software like SPSS and programming languages like Python, R and SQL. This guide focuses on spreadsheet programs like Excel and Google Sheets as these are common tools for the beginner.

The basics

In most cases you just need to perform basic statistical and mathematical procedures to find insight from the data.

You should understand the concepts below and know how to use spreadsheet program functions to find:

- Total
- Maximum/minimum values
- Averages: mean, median, mode
- Percentage
- Percent change
- Rate: per capita, per cases
- Distribution of data

The basic spreadsheet functions you should know are:

- Sort
- Filter
- Perform simple calculations
- Use built-in mathematical formulas (SUM, MEAN, MEDIAN, MAX, MIN etc.)
- Join/split variables
- Join data points from 2 or more datasets based on common variables
- Summarize/aggregate data points with Pivot Table
- Make histogram or density plot to see data distribution pattern

TUTORIAL

Beginner, intermediate and advanced spreadsheets tutorial

Dealing with data sample

With a data sample that aims to represent the bigger entity, such as an election poll, learn about the concepts of sample size, sampling methods, margin of error and confidence level. These help you select more reliable datasets and avoid misleading your audience.

To represent the bigger entity, select samples so they don't over- or under-represent the subgroups in the bigger entity. For example, if the bigger entity to be represented is the population of a country and 60% of the population is aged 15-30, the samples should have the same composition. Hence online polls such as those run by news outlets on Facebook or Twitter cannot represent the bigger entity as the samples (respondents) are not controlled.

Since we use a smaller data sample to represent a much bigger entity, there is always uncertainty. In statistics, that uncertainty is calculated as 'margin of error' (MOE) and 'confidence level' (CL).

MOE measures the range or error of the results. For example, in a poll that has a MOE of 3% and a CL of 95%, 50% of the respondents say they will vote for candidate A. This means that the support for candidate A ranges from 47% to 53%. The 95% CL means if the same population is sampled repeatedly, the results would fall within this range (47%-53%) 95 times out of 100.

There is a checklist of 8 items in this article that helps you to decide whether a poll should be taken seriously.

Another common mistake among journalists is using 'mean' instead of 'median' in their reporting. Journalists often need to use the average value to explain the characteristics of a group of people or a phenomenon e.g., average household income to indicate economic status. 'Mean', often referred to as 'average', is calculated by dividing the sum of values in a collection by the number of values. 'Mean' is a good measurement when all the values are relatively close.

However, when a dataset is skewed - a small number of extremely high or low values in a dataset can distort the 'mean' to a higher or lower value - it's better to use the 'median'.
'Median' is the midpoint in an ordered list of values - the point at which half the values are higher and half lower. For example, in a society where wealth distribution is uneven, a small number of super-rich people can inflate the 'mean' income, so 'median' is a better measurement of average income.

As a rule of thumb, you should calculate both the mean and median. If they are close, stick with the 'mean', if they are far apart, because of extreme outliers, use the 'median'.

FURTHER READING ON THIS TOPIC:

Sampling and margins of error

Five key things to know about the margin of error in election polls

What to look for

When you are familiar with the concepts and functions above, you can start to 'interview' the data with different analyses depending on the questions you have.

Look out for:	
TREND	how the variables have changed over time or across groups.
CONTRAST	the differences among comparable variables.
OUTLIERS	data points that are far from the average (but double check to make sure they are not errors).

FURTHER READING ON THIS TOPIC:

[Finding stories in spreadsheets]

Combine and compare

Analyzing a dataset can prompt more questions that require other datasets to answer.
For example, when you compare the number of murder cases in two cities, you need their population numbers to calculate the number of cases per capita for a fair comparison. To investigate further the reasons behind a rise or fall in the murder rate, you might want to look at police numbers or drug abuse cases in those cities.

Another typical situation that requires more than one dataset is when you need to prove or disprove a hypothesis. For example, if the education department claims that having more exercise books in schools produces better academic results, you need at least two datasets - the number of exercise books used in schools and their academic results - to examine it.

Correlation and causation

Examining different datasets helps you to investigate the relationship between phenomena, but be extra careful in drawing cause and effect conclusions. Correlation between two variables does not mean causation. For example, the homeless population and crime rate in a city are correlated, but this does not necessarily mean homeless people commit crime. There could be other factors like unemployment or drug abuse. Make it clear to your audience when you describe the relationship between variables. There is a hilarious book on false causation called Spurious Correlations.

Visualization as exploration tool

However, many insights in the data are hidden from these 'interview' techniques. You need to convert the data into visual forms to see them. Relationships between variables, trends and insights within geospatial data are best recognized using visualization.

The goal of visualization at this stage is for exploratory data analysis rather than presentation to end-users. Start by visualizing one variable in your dataset in different chart types (such as maps if your data is geospatial). Repeat this with other variables, then use multiple variables in each visualization. Looking at multiple visualizations at the same time can help reveal insights. There is no hard rule but this process gets faster with experience.

The techniques and tools of data visualization will be discussed in the next chapter.

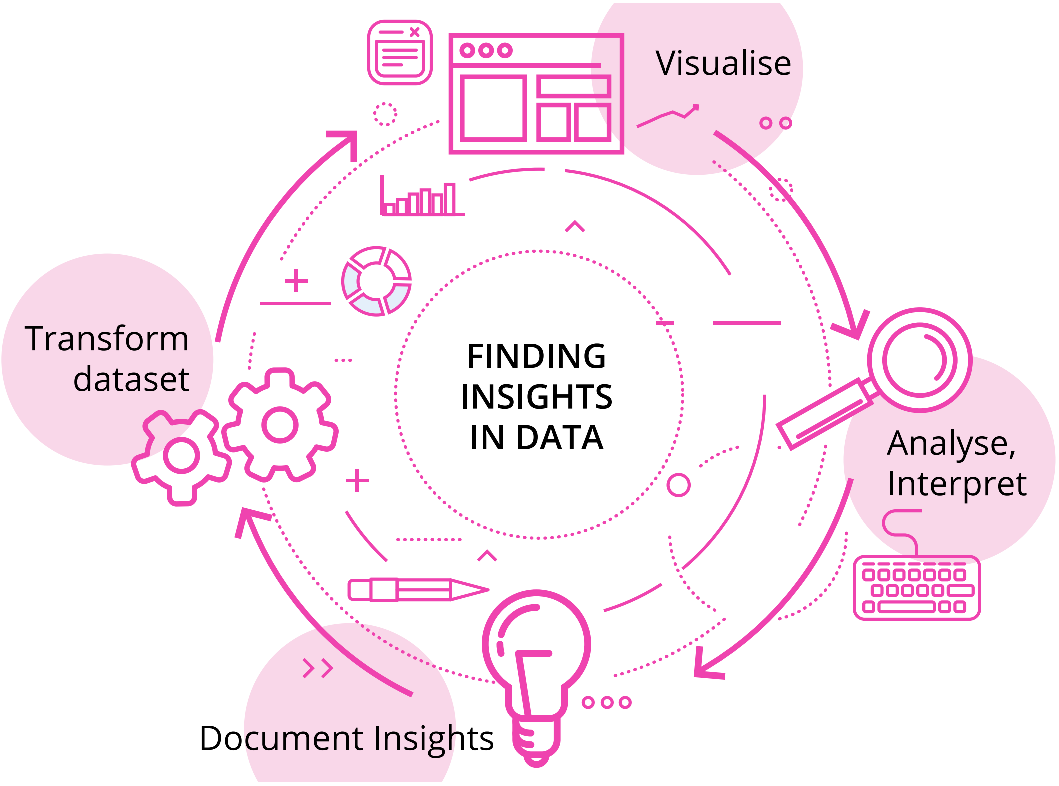
Bulletproof your data

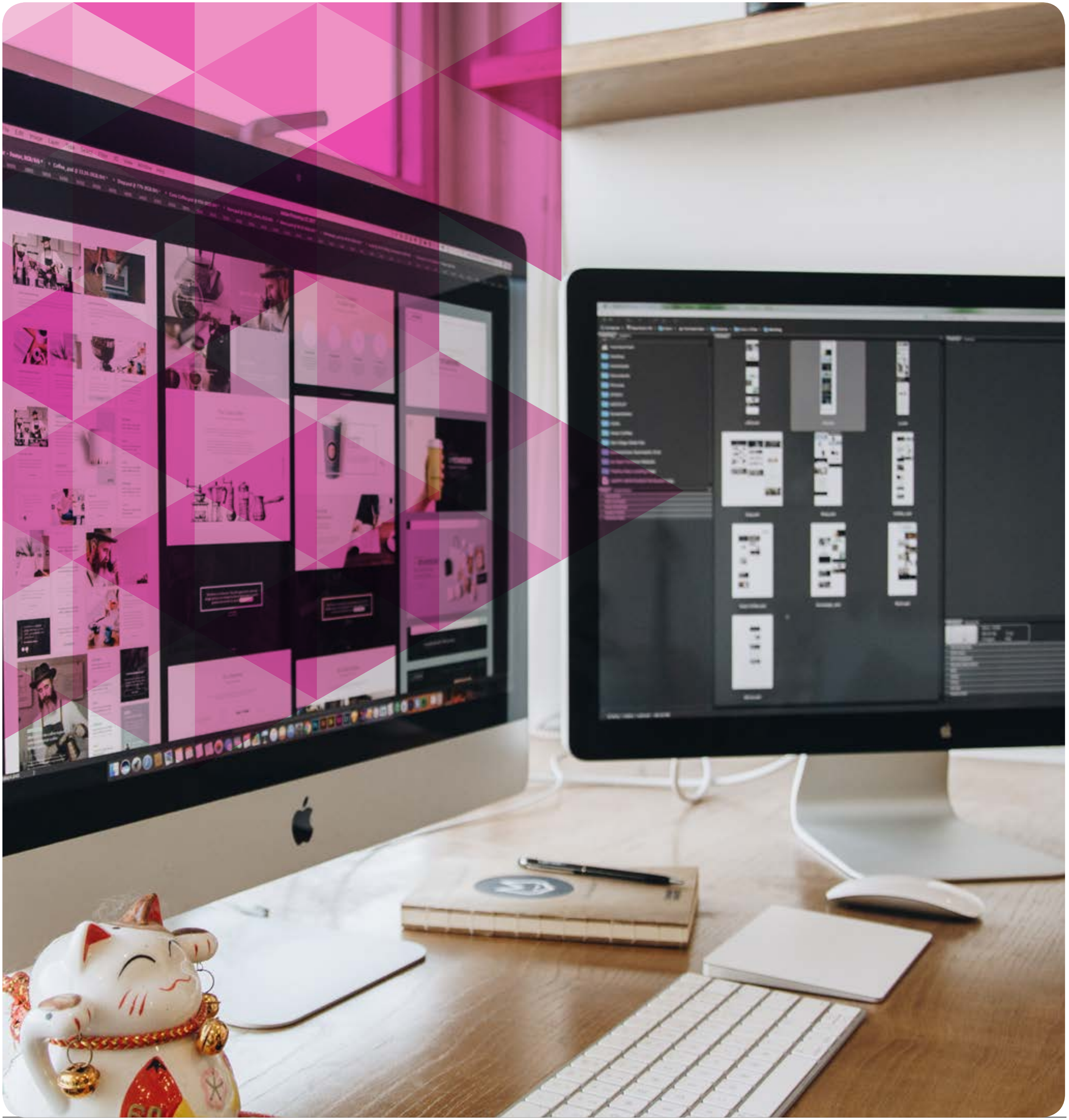
As in other forms of journalism, data journalism is vulnerable to mistake and bias. All the journalism ethics and standards apply to data-driven stories, but some steps are different from traditional reporting. ProPublica has published a guide on bulletproofing data.

There is also an online tutorial on this topic.

FURTHER READING ON THIS TOPIC:

The challenges and possible pitfalls of data journalism, and how you can you avoid them





TELLING STORIES **WITH DATA**

An important part of data journalism is the effective communication of data to your audience. Done properly, data presentation can be more engaging, comprehensible and impactful than traditional storytelling.

There are different ways to communicate data. Below are 6 categories outlined by data journalism guru Paul Bradshaw.

VISUALIZATION
NARRATION
SOCIAL COMMUNICATION
HUMANIZATION
PERSONALIZATION
UTILIZATION

We will look deeper into visualization as it is the most common way in data journalism to communicate data.

Visualization

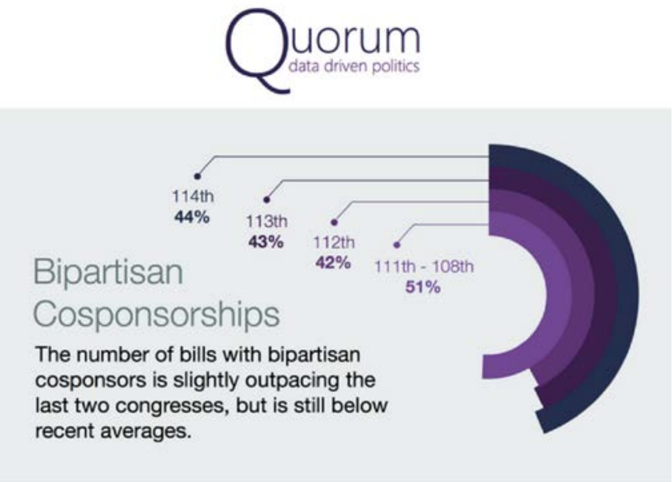
Good and bad visualization

THERE ARE REASONS WHY VISUALIZATION IS BETTER THAN VERBAL OR TEXT COMMUNICATION:

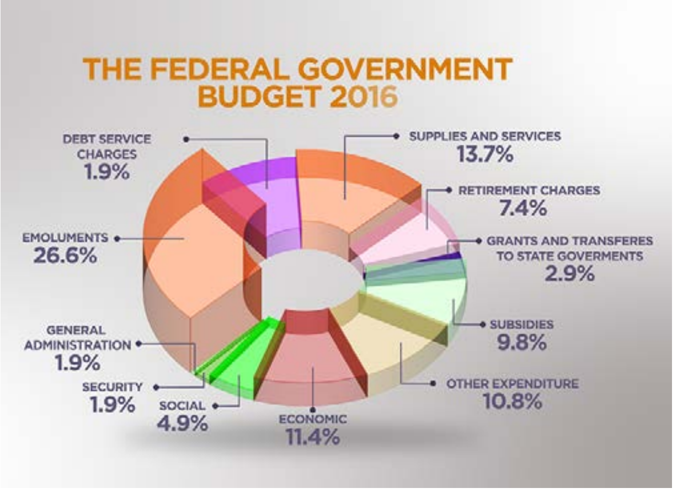
The human brain processes images much faster than text.

Most audiences prefer to consume and share visual content.

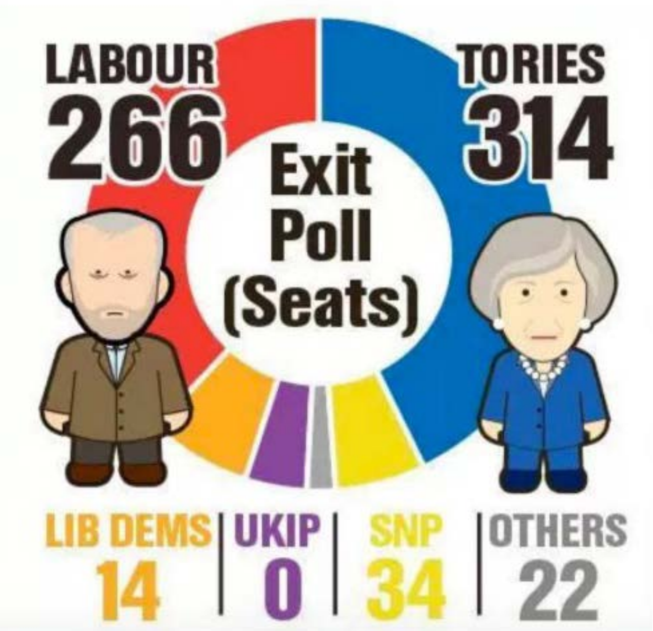
It cannot be emphasized enough that the goal of data visualization is to communicate information effectively. A common mistake among designers is to go for a powerful graphic rather than powerful journalism. This results in fancy and decorative visualizations that are unclear, confusing or misleading. (Google “bad data visualizations” for some examples of this)



Source



Source



Source

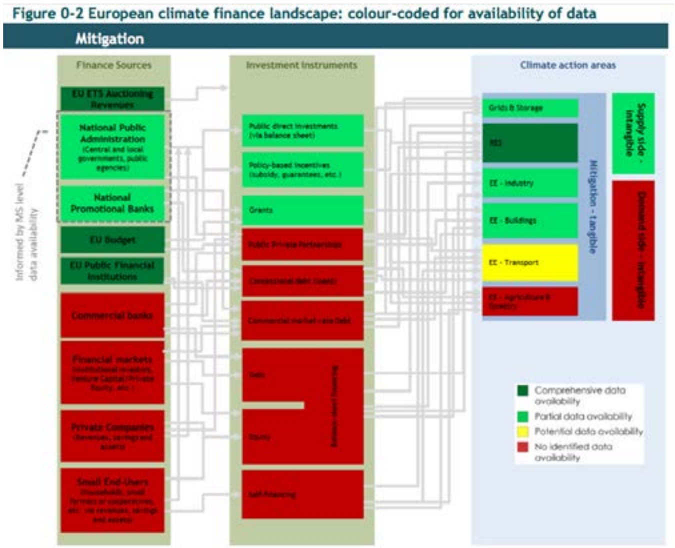
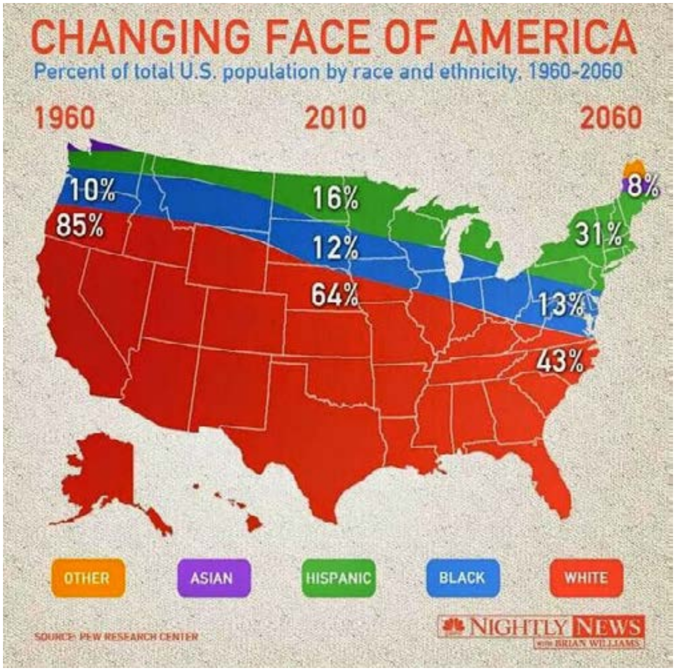


Figure 0-2 European climate finance landscape: colour-coded for availability of data
Mitigation
Finance Sources
EU ETS Auctioning Revenues
National Public Administration (Central and local governments, public agencies)
National Promotional Banks
EU Budget
EU Public Financial Institutions
Commercial banks
Financial markets
Private Companies
Small End-Users
Informed by MS level data availability
Investment Instruments
Public direct investments (via balance sheet)
Policy-based incentives (subsidy, guarantees, etc.)
Grants
Public Private Partnerships
Concessional debt (loans)
Commercial market-rate debt
Debt
Equity
Self-financing
Climate action areas
Grids & Storage
RES
EE - Industry
EE - Buildings
EE - Transport
Mitigation - tangible
Supply side - intangible
Demand side - intangible
Comprehensive data availability
Partial data availability
Potential data availability
No identified data availability

Source



Source

FURTHER READING ON THIS TOPIC:

Good and bad design

Data visualization expert Alberto Cairo has four principles for good visualization:

- Based on good data
- Attract attention
- Don't frustrate readers
- Show the right amount of data

Effective data visualizations - where information has been thoughtfully selected and presented in a clear, simple and straightforward manner - help viewers save time. For more complicated visualizations, the design complexity should reflect the data complexity. Take a look at the "Data looks better naked" series by Darkhorse Analytics to see how 'less is more' in data visualization.

Design with users in mind

But how do you know your visualization will not frustrate your readers? This is where design thinking can help. Put simply, design thinking - also known as user-centered design - is a mindset that makes decisions based on what the user, not the creator, wants. It's a relatively new concept in journalism where journalists are used to deciding what content to push to their audience without asking them.

Start by defining your goal. What is the single most important message you want your audience to take away from your visualization? Avoid trying to squeeze in too many messages. When you are clear with the message, select only enough data points to be visualized. For example, if you want to show that the US ranks second in the medal count in the 2014 Sochi Winter Olympics, you don't have to show the entire medal count, just the top 10 countries will suffice.

Next, identify your targeted or potential users, and understand them. What is their data literacy level? What kind of charts do they understand better? These questions will help you decide which type of charts to use. Always use charts that users are most familiar with.

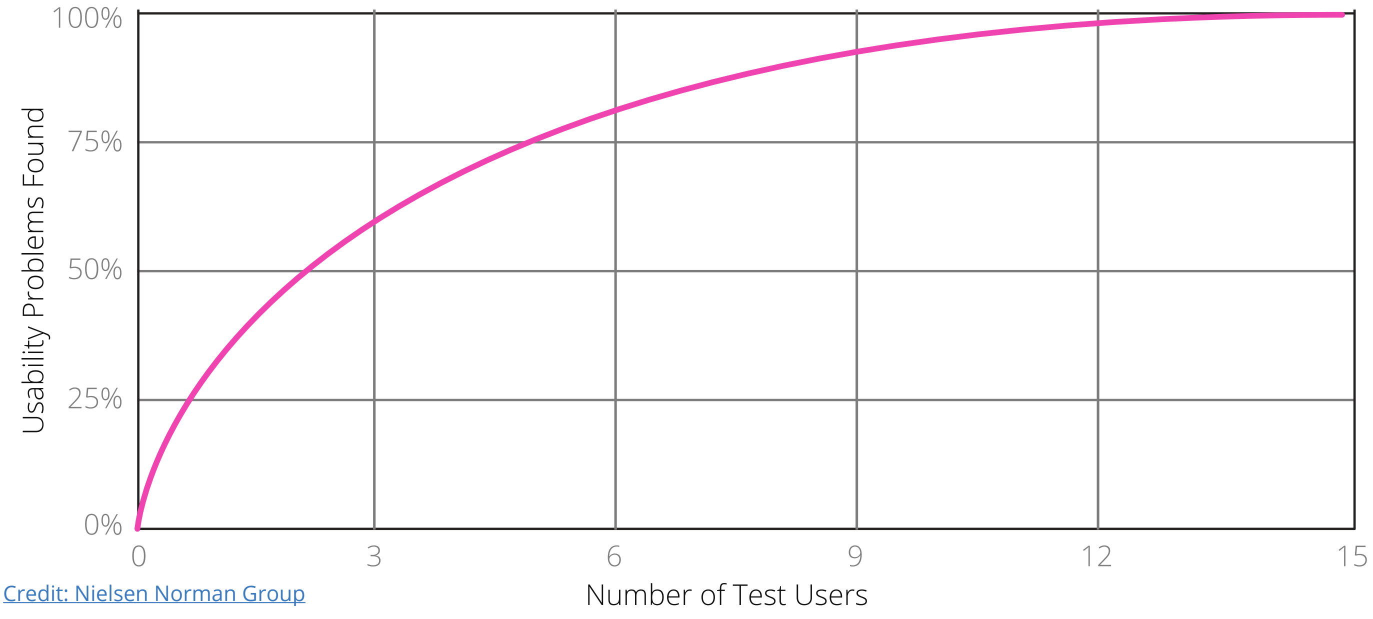
Knowing how they will view or interact with your visualization (social media, mobile/desktop, print or TV broadcast) or whether they understand the terminologies you plan for your visualization (median or random sampling) will affect the language, size and layout of your visualization.

The key to your design thinking should be knowing what device users use to access your product. For example, about 80% of Malaysiakini users access its content through mobile devices, so it used a mobile-first design approach in the production of a data-driven news game called Undi Power (jointly developed by Malaysiakini and the author of this guide). The layout, animated graphics, animated charts and user interface are all designed for mobile screens before adjusting for a desktop screen. The social sharing function of the news game is also optimized for Facebook, the most popular social media platform in Malaysia.

Bear in mind a saying among developers at news organizations with a majority of users on mobile - *if it doesn't work on mobile, it doesn't work.*

The next important step is to perform user-testing on your design, and be prepared to redesign based on user feedback. Try to simulate the real-world situation where users consume your visualization. For example, if your visualization is a static image for online publication, show it to users for 10 seconds (or longer if it's a more complex visualization) without any verbal explanation, and ask them what message they got from it.

User-testing need not be resource intensive. Studies found that just 5 users can find 85% of the usability problems in your design. However, don't ask your newsroom colleagues to be your test users unless journalists are your target users. Instead, find users in other non-journalistic departments or among your friends and family as long as they fit your targeted user profile. If you can, repeat the test-and-redesign process for another two rounds - which means 15 users in total.



Credit: Nielsen Norman Group

Which chart to use?

The main factors to consider when selecting a chart are: the type of data, the message you want to get across, and the charts your audience are most familiar with. There are at least 60 types of charts, but those listed below are usually sufficient to effectively communicate data.

- **Bar chart (or column chart)** - to compare discrete, numerical data across categories. Good at showing accurate differences even when these are small.
- **Line graph** - to compare numerical data over a continuous time span. Effective in showing trends and relationships between two variables.
- **Scatter plot** - to compare relationship or interaction of two variables. Might not be correctly interpreted by common users without further explanation.
- **Pictogram** - similar to bar chart, but the use of icons makes it more entertaining. Can also be used to show composition.
- **Map** - to compare geospatial data. Can be further grouped into bubble map, dot distribution map and choropleth map.
- **Network graph** - to show how entities are interconnected. Often used to visualize the relationship of a group of people or organizations.

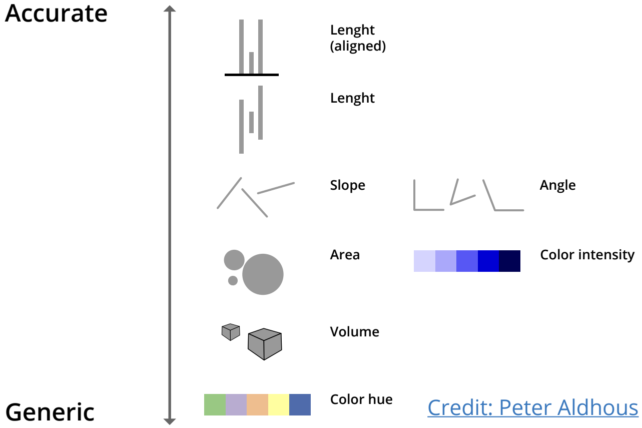
You may notice this list doesn't include the widely-used pie chart. Many data journalists don't like pie charts as they can be less effective and widely misused. In most cases, a pie chart can be replaced by a bar chart.

The [Data Visualization Catalogue] is an online library of 60 chart types.

Use a pie chart only when:

- you want to show parts of a whole (often a bar chart can do a better job)
- slices always add up to 100%
- you have fewer than five slices
- the size of the slices should be significantly different
- you don't need to make an accurate comparison (the brain is bad at comparing angles)

The diagram below is another guideline for chart selection, ranking visual elements by how accurately the brain can estimate the numbers presented by them. When you want users to accurately compare data points, try to use the elements near the top of the hierarchy.



Credit: Peter Aldhous

FURTHER READING ON THIS TOPIC:

Which chart should I use, and why?
Information design for the human brain

When to use a table?

In some cases a table works better than visualization:

- when there are only a few data points
- when it's unnecessary to show comparison, trend or relationship
- when it's important to show precise value

After creating your chart, there are other components you need to decide, including the headline, label, legend, typography and color.

The following part covers some general guidelines. A good book on this topic is Designing Data Visualizations by Noah Iliinsky and Julie Steele.

The following part covers some general guidelines. A good book on this topic is [Designing Data Visualizations by Noah Iliinsky and Julie Steele](#).

Six Golden Rules

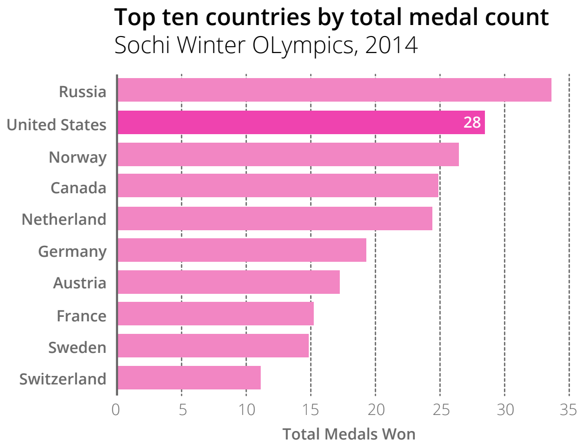
In her data journalism manual, published by the United Nations Development Programme, data journalism trainer Eva Constantaras lists five golden rules for data visualization design. We have added one more below.

1 **No 3D effects:** These distort the data as the part closer to the viewer appears larger. Also avoid color gradients and drop shadows. These are unnecessary visual noise.

2 Sort data from largest to smallest: This makes visual comparison much easier.

3 Choose maximum 2 colors or shades of the same color for your graphic: Stay away from rainbow colors. This makes visualizations look more professional and credible.

4 Your headline should tell the story: Tell your audience the key message. For example, for a bar chart showing Olympic medal counts by country, instead of "Top 10 countries by total medal count, Sochi Winter Olympics, 2014", use a more specific headline like "US ranks 2nd in medal count in the Sochi Winter Olympics 2014".



5 Use direct labeling whenever possible: Don't let your audience jump back and forth between a graph and a legend.

6 Make your message stand out from the rest: Use color tone, labeling or other visual effects to draw attention to the element that carries your message. Look at the medal count chart above, the bar representing US has a stronger yellow and label because the story is about the US team's performance at the Games.

Data visualization tools

The tools listed below are those commonly used in journalism projects - and tend to have extensive manuals and tutorial and a strong user community where you can get advice and view examples.

Generally, open source tools allow you to make highly customized charts as you have access to the source code. Paid tools often provide a free version that allows you to access the basic features or limit the publication of your visualization.

Visualising Data, a website created by data visualization specialist Andy Kirk, has compiled a catalogue of tools to make charts.

Chart making tools

Product	User interface	Paid/Free	Interactive?	Customization	Learning curve
Highcharts Cloud	GUI	Paid (free for non-commercial)	Yes	Medium	Low
Highcharts API	Javascript	Paid (free for non-commercial)	Yes	High	Medium
Datawrapper	GUI	Paid	Yes	Low	Low
Plotly Chart Studio	GUI	Paid	Yes	Low	Low
Plotly JS	Javascript	Open source	Yes	High	Medium
Tableau Public*	GUI	Free	Yes	Medium	Medium
Chartbuilder	GUI	Open source	No	Low	Low
D3.js	Coding	Open source	Yes	High	High

*Tableau is the only listed tool that allows a non-coder to build multiple charts on a dashboard that can interact with each other. This is good when you don't know coding but want to build a news app for users to explore complicated datasets.

Mapping tools

Simple and free:
Google My Maps
Google Earth Pro
Google Fusion Tables

Powerful GUI programs:
QGIS (free)
Carto (paid)
ArcGIS (paid)
MapBox (paid)

Require coding:
Google map API
D3.js

Infographics tools

Infographics tools are good for building pictograms as they offer a series of ready-made icons and images. Most infographics tools also have their own built-in chart-making features.

Chart-making features:
Piktochart (paid)
Infogram (paid)
Venngage (paid)

Other tools

Look out for:	
Timeline JS (free)	a simple tool to present multimedia information in timeline format.
StoryMapJS (free)	a simple tool to present map-based multimedia information.
Esri Story Map (paid)	a non-coding tool that lets you combine maps with other story-telling components to form an immersive experience.
Gephi, NodeXL and Google Fusion Tables	tools to make network graphs.
Color Brewer 2.0 and Color Picker	tools that help to choose color schemes for choropleth maps.
Color Hunt	a library of color palettes for designers.
Color Oracle	a color blindness simulator.

Animated visualization

Done well, animation can supercharge the effectiveness of your data visualization and make it more engaging. There are minutes-long data journalism video and short animation (less than 10 seconds). Some are part of an interactive story package, some are standalone animation published in GIF (Graphics Interchange Format) that loops indefinitely. GIF is especially popular in data visualization due to its small size and the loop effect.

It is good at:

- showing changes over time
- comparing differences
- presenting a data-based narrative
- explaining a series of processes

ProPublica news app developer Lena Groeger has been advocating the use of GIF in information design. This presentation by her showed many GIF animation examples. She also published a tutorial to make GIF animation.

Animated visualization can also be effective when you want to present your information in stages rather than throwing everything at the user. Here is an example from Hindustan Times that compares the fastest running athletes in the world.

FURTHER READING ON THIS TOPIC:

Nine tips for making a data journalism animation

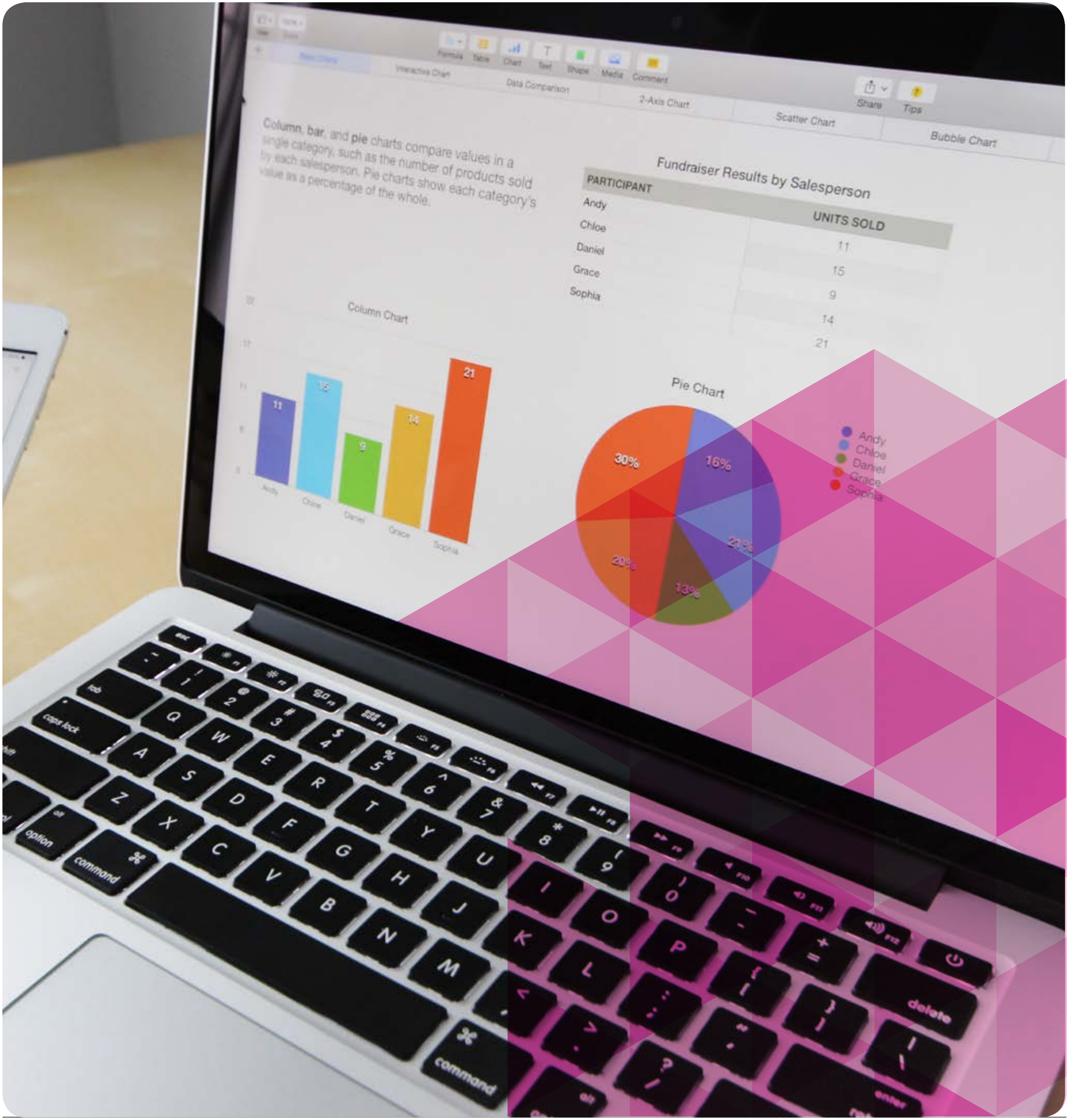
ON REPEAT:

How to Use Loops to Explain Anything

Although visualization plays a major role in most data-driven stories, it has to be complemented by other traditional storytelling elements such as text, image and video. In some cases data visualization is not the primary medium to communicate the data. Next, we look at other ways of communicating data.

Narration

Almost all data-driven stories come with text narrative. This helps explain content that can't be visualized. There are tips on how to simplify data insights (page 14-19) and tips for writing numbers for web readers.



Column, bar, and pie charts compare values in a single category, such as the number of products sold by each salesperson. Pie charts show each category's value as a percentage of the whole.

Fundraiser Results by Salesperson

PARTICIPANT	UNITS SOLD
Andy	11
Chloe	15
Daniel	9
Grace	14
Sophia	21

Social communication

Data can be highly sociable if you pick issues that resonate with users, produce the right visualizations or give users access to the data.

The Financial Times has built a huge Instagram following since 2015 partly by posting data visualizations optimized for the social media platform. Many publications found GIF animation to be a good format to present data content on social media.

Manila-based Rappler launched a social campaign on road safety on the back of its data-driven series on the increasing road traffic accident rates in the Philippines. Data crowdsourcing that involves user participation can also build an engaged community. Opening up your dataset, whether via an API or news app, enables users to utilize and share parts of the dataset relevant to them. All these approaches help your content reach a broader audience and build social relationships.

Humanization

A major pitfall of data-driven stories is putting numbers above journalism. Ultimately, journalism should be about people, but data can sometimes turn human stories into hard cold numbers.

It is imperative to humanize data by going to the ground and telling the stories of people whose lives are affected by the data. This helps you connect with your audience and create empathy.

A good example is the Pulitzer Prize-winning series Failure Factories by the Tampa Bay Times. The prologue features 24 animated charts that illustrate the problem of 5 public schools. It was followed by personal stories of affected students and their parents.

Another example is Hindustan Times' Class of 2018. This special series features interactive data stories, photos and detailed personal stories of teachers and students.

Personalization

Personalization brings users closer to the data, making it more engaging and relevant.

In data-driven stories, personalization is often through interactive features or news apps that invite users to input their details before showing data based on those inputs.

For example, when the Malaysian election commission proposed redrawing voting boundaries, news website Malaysiakini developed a 'Redelineation Checker' that invites users to input their electoral districts to check if they will be affected by the proposal.
A New York Times data interactive went further by using readers' IP addresses to estimate where they were and use that information to show data related to that location.

South African investigative reporting website Oxpeckers, with Code for Africa, created an app called #MineAlert that sends information on the status of mining operations in a user's area.

Utilization

Another way to communicate your data is to build tools that allow other users to use your data based on their own needs. Data personalization is part of this approach, though some tools don't involve personalization. For example, in 2010 ProPublica built an open database of payments made by pharmaceutical companies to doctors and a news app to access that data. Over 125 local outlets have used the database for their own reporting and investigation. Another example is the news app by Public Radio International that let users explore and visualize terrorist attack data in 2016.

FURTHER READING ON THIS TOPIC:

How to Build a News App

How to Make a News App in Two Days, as Told by Six People Who Tried It for the First Time

TUTORIAL:

First News App

How to build a news app that never goes down and costs you practically nothing

How to build a News app – NPR

OTHER THINGS THAT NEED **ATTENTION**

Open Your Data and Tools

Sharing your data and tools can go against the culture of most news publication markets where outlets compete for exclusive stories, audience attention and faster news delivery. However, more journalists have found that sharing data, tools, technical know-how and newsroom workflow across news organizations has created a synergy that elevates the industry's capabilities as a whole. The wheel doesn't have to be reinvented by everyone every time.

For example, one media outlet produced a story using government census data. It converted the data, originally in PDF format, into a spreadsheet, cleaned it and shared it online. Other outlets can skip those basic steps and develop more complicated tools to analyze the cleaned data for different stories. When the second outlet shared its tools, other outlets can modify them in turn for their own use. This creates an ecosystem that helps everyone inside it to utilize their resources more efficiently and grow their capabilities.

In an interview with the Tow Center for Digital Journalism, La Nacion multimedia and interactive development manager Momi Peralta made a good case for sharing data especially in countries where access to information is difficult.

"This is not only a data revolution. It is an open innovation revolution around knowledge. Media must help open data, especially in countries with difficult access to information.

"The way to go for us now is to use data for journalism but then open that data. We are building blocks of knowledge and, at the same time, putting this data closer to the people, the experts and the ones who can do better work than ourselves to extract another story or detect spots of corruption.

"It makes lots of sense for us to make the effort of typing, building datasets, cleaning, converting and sharing data in open formats, even organizing our own 'datafest' to expose data to experts. Open data will help in the fight against corruption. That is a real need, as here corruption is killing people."

A research report by Storybench at Northeastern University's School of Journalism concluded that the ethos of open-source sharing is one of the three themes of best practice behind successful digital newsrooms.

In ecosystems where such a culture is flourishing, like the US, leading digital journalists come together and share best practices, step-by-step tutorials and production tools at journalism conferences, specialized workshops and on websites. This document is just another guide that builds on that increasingly growing open source knowledge.

Google Sheets and GitHub are the most common online platforms to share data and other information. Most journalists are familiar with Google Sheets that allow them to share and work on the same spreadsheet. However GitHub, an online service for users to store, share and collaborate on digital projects, is less well known among journalists. Most open source materials have a GitHub page.

TUTORIAL:

How to use GitHub and the terminal: a guide

Monetization and Business Model

The disruption posed by the Internet has wreaked havoc on the business models of many traditional newsrooms, forcing them to downscale their operations. While searching for new revenue sources, newsrooms are cautious in investing their resources into new ventures - holding back growth in data journalism in countries where this genre is still relatively new.

Most leading newsrooms in digital journalism see data journalism as a new way to strengthen the quality and broaden the reach of their journalism instead of a new asset to be monetized. They found data journalism is able to:

1. Attract more audience: data-driven stories on evergreen/recurring issues are able to generate a long-tail traffic. Personalized data make stories more engaging and shareable. For example, the data library of Texas Tribune, a nonprofit media organization that covers the state of Texas, contributed a majority of its traffic.
2. Strengthen brand and product line: Data journalism sets your reporting apart from your competitors. It is not easily replicable. Many award-winning investigative reports are data-based.
3. Serve as a catalyst for a higher level of technology adoption and innovation in the newsroom.

There are also a handful of outlets specializing in data journalism that are building a business model around data. In 2014, ProPublica, a non-profit news organization known for its investigative and data journalism, launched Data Store to sell its datasets. The new initiative pulled in $30,000 revenue in the first 5 months and hit $200,000 in 2016. To balance between openness and revenue goals, the pricing of data varies depending on who is buying. For example, at launch, a state's data on doctors receiving payments from pharmaceutical companies costs $200 for journalists and $2,000 for academics. Those data are also sold to other companies. Yelp, a popular mobile app in the US that collects and publishes reviews on local businesses, has inked a deal with ProPublica to include ProPublica's ER wait-time data and other related data on its hospital listings. ProPublica has expanded the model to manage sales of other journalism organizations' datasets, helping turn those resources into revenue.

Another example of selling data is UK-based social enterprise OpenCorporates, which collects and publishes company data. Clients who want to use the data for private commercial purposes have to pay.

Apart from data, the skills in processing and presenting data can be monetized, too.
During the London 2012 Summer Olympics, New York Times interactive team turned the live raw data from the International Olympic Committee into embeddable widgets and other packages to be sold to other news outlets.

Katadata ('talk data' in English) is an Indonesian online media company that provides in-depth economic and business information which are supported by profound data and analyses. Besides selling its datasets, Katadata also provides data analysis and visualization services to private companies and government bodies. In August 2017, it worked with a government agency to produce a 17-meter long data visualization on the country's ecnomic history which was recognized as the longest infographic in Indonesia. The exhibition in which the infographic was shown was attended by Indonesian President Joko Widodo.

RESOURCES

BUILDING DATA JOURNALISM IN THE NEWSROOM

Want to start a small data journalism team in your newsroom? Here are 8 steps

Big data in small organizations: Constantaras. Strategies from setting up data teams in developing countries

How Argentina’s La Nación became a data journalism powerhouse in Latin America

Integrating Data Journalism into Newsrooms

Collaborative, Open, Mobile: A Thematic Exploration of Best Practices at the Forefront of Digital Journalism

Diving into Data Journalism: Strategies for getting started or going deeper

ESSENTIAL DATA JOURNALISM SKILLS

Finding Stories in Spreadsheets

The Curious Journalist’s Guide to Data

Numbers in The Newsroom: Using Math and Statistics in News

Data Journalism Manual by ODECA

MaryJo Webster’s training materials

Tools for journalists

DATA JOURNALISM NETWORKS
NICAR listserv
Data Journalism Award Slack group

DATA JOURNALISM FELLOWSHIPS
ICFJ
School of Data
List of grants and fellowships

DATA JOURNALISM MOOC (MASSIVE OPEN ONLINE COURSES)
Managing Data Journalism Projects
Doing Journalism with Data: First Steps, Skills and Tools
Mistakes We Made So You Don’t Have To: Data Visualisation, Journalism and the Web
Bulletproof Data Journalism
Cleaning Data in Excel

WHERE TO FIND GOOD DATA JOURNALISM?
List of nominees of the Data Journalism Awards by Global Editors Network
List of Online Journalism Awards winners
Monthly digest of the best of data visualization by Visualising Data

NEW WEBSITES SPECIALIZING IN DATA JOURNALISM
FiveThirtyEight
Katadata
IndiaSpend
ProPublica

DATA BLOG/VERTICAL OF MAJOR NEWS OUTLET
New York Times - Upshot
Washington Post - Data blog
South China Morning Post - Infographics
Hindustan Times - Interactives
Tempo - Investigations
The Guardian - Datablog
Caixin - DataNews

RESOURCES FOR EDITORS
Editors: Bulletproofing your data stories
Data Editing Road Map

OTHERS
Collection of code being written by news organizations and individuals
Collection of DDJ books and tips
Collection of Data Journalism books
Collection of programming resources for journalists

Guide Author

KUANG KENG KUEK SER

Kuang Keng Kuek Ser is an award-winning digital journalist. He produces and consults on data-driven reporting and interactive journalism projects. Keng is also the founder of DataN, a training program that lowers the barrier for newsrooms and journalists with limited resources to integrate data journalism into daily reporting. He has more than 10 years of experience in digital journalism. He was awarded a Fulbright scholarship in 2013 to further his studies at New York University's Studio 20. In 2015, Keng was selected as a Google Journalism Fellow and a Tow-Knight Fellow.

You can reach him at kuangkeng@gmail.com or follow @kuangkeng on Twitter.

READ MORE GUIDES AT WWW.KBRIDGE.ORG/GUIDES mdif media development investment fund

MDIF New York
37 W 20th St., #801
New York, NY 10011, USA
1 (212) 807-1304

MDIF Prague
Salvatorska 10
110 00 Prague 1, Czech Republic
(420) 224-312-832

www.mdif.org

Sida Funded by The Swedish International
Development Cooperation Agency (Sida)