Module 2: From Data to Stories

STUDENT WORKBOOK

This module instills basic knowledge of data formats, the skills to find data online and the concepts to transform data into stories. Starting with a review of data formats, the unit moves on to Google search techniques to find different data types, introduces the formal process for designing a data project, evaluates the hypothesis of stories that use data and provide a chance for students to transform basic data fact sheets into stories. After completing this module, you will be able to:

- Identify basic data formats
- Search for data in different formats
- Automate data searches
- Transform fact sheets into simple stories
- Develop a hypothesis and questions for a story
- Practice evaluating the hypotheses of other stories

Contents

Getting Started........................................................................................................................................3
Lesson 1: Common Data Formats ........................................................................................................4
  Data formats: Machine-readable, Computer Generated, Structured ..............................................5
  Portable Document Formats (PDF) ......................................................................................................6
  From Document Formats to Machine Readable Data Formats .........................................................7
Lesson 2: Finding Data Online .............................................................................................................9
  Using Advanced Search .......................................................................................................................10
  Setting Up Alerts ..............................................................................................................................13
  Advanced Google Searches: Scavenger Hunt! ..................................................................................14
  Using Data Portals ............................................................................................................................15
  National Databases ............................................................................................................................17
  Navigating National Databases ......................................................................................................... Error! Bookmark not defined.
Getting Started

Public databases like the World Bank’s DataBank have a wealth of data about more subjects than we can count or imagine: [http://databank.worldbank.org/](http://databank.worldbank.org/)

But as you can see on this site, a lot of the jargon around data can be confusing. Looking at the first few categories, what do GDP growth (annual %); GDP (current US$); GDP per capita (current US$); GNI per capita, Atlas method (current US$); Exports of goods and services (% of GDP) mean? We don’t expect everyone who uses economic data in their reporting to be an economist.

Rather, this unit will help you start navigating your way around complex data by not only introducing some basic data formats and some questions that we should always ask of datasets, but also the resources we need to find out what is beyond our knowledge to help us understand what the data is measuring and how.
Lesson 1: Common Data Formats

Modern data analysis relies on software to do the heavy lifting involved in data analysis for us. We cannot work with data until we convert data into a format that the computer understands so that it can organize data into rows, columns, and cells.

Much of what stops citizens from using data, either intentionally or unintentionally, is that data is provided in formats that can’t be immediately used or read by a computer. This lesson explains such data formats and the processes to transform them.

Data analysis, storytelling and visualization all depend on a computer program being able to read our data. Unfortunately, often data comes in formats that computers do not understand.
Data formats: Machine-readable, Computer Generated, Structured

In these data formats, computer software recognizes an explicit structure to the data - most commonly in a table - with columns and rows that organize and describe discrete data points. Excel and CSV are common examples.

- **Excel file (XLS):** data is saved as a table readable by Microsoft Excel
- **Comma separated values (CSV):** Plain text file with each data entry separated by a comma

These formats are typically the best suited for analysis, and you can easily work with them in a spreadsheet program - like Excel. When searching for data, if you can find Excel or CSV formats, this is a good sign that you won’t have to spend a lot of time cleaning and formatting.

Note that CSV (comma-separated values) and TSV (tab-separated values) formats are formats for “encoding” tabular data. In simple words, CSV and TSV files are plain text files in which:

- Each line represents a row and
- Within each line, a comma (for CSV) or a tab character (for TSV) separates columns

Excel files also uses on a similar structure, but relies on Microsoft software.

**Tools:**

Google spreadsheets, Microsoft Excel are commonly available tools that help you work these formats.

From Evidence to Stories: Thinking Like a Data Journalist
PDF files come in a few different varieties.

- The first question to ask is if they are computer-generated or not? That is, if a file was saved in a PDF format or if it was actually printed and scanned back in as an image not generated by software.
- The next question is if the data within the PDF is structured, as in, it’s available in columns and rows published in a table.
- Finally if it is searchable - which has to do with whether it was generated by a computer. Basically searchable means that you can highlight the text and the computer recognizes the letters and numbers as characters.
From Document Formats to Machine Readable Data Formats

Typically, the best suited data formats for analysis are structured and machine readable – like CSV or Excel. When you find data in other formats, say a PDF, it’s useful to convert it into a structured and machine readable format.

**Data in PDFs**

PDFs often contain structured, computer generated tables but a PDF is not a data format. The table has to be converted into a format that can be opened by a spreadsheet program. So these data tables require extraction into a data format through special software. You will practice extracting data in the Scraping lab.

*Tools:* Tabula, CometDoc, PDFtoExcel, Zamzar

**Data in Scanned Images**

These are primarily image files that are read as one giant block instead of discrete parts. These require Optical Character Recognition Software to recognize the text in the file. Usually, these used to be computer generated, but then someone printed the document and scanned it back into the computer, turning it into giant image file.

*Examples:* Some PDF and all bitmap images (GIF, JPEG, PNG, BMP)

*Tools:* Google Docs OCR, Document Cloud
Data in Unstructured Formats
Some data has been generated by a computer but does not have a structure recognized by machines. Examples of this include data that has been entered into a text document in paragraph format and some data on websites. Basically, in this case, a developer has to teach the computer what the pattern is in the data and then extract it into a data format.

Tools: Python or Ruby programming languages to scrape data using https://morph.io/

Less Common Data formats
Some data, especially large databases, are saved in packages designed to be coded into websites or read by statistical software like Stata or R. These require conversion to CSV or Excel for use with spreadsheets software.

Examples: JSON (JavaScript Object Notation) or XML (extensible Markup Language) for programming and .SAV or .R. Try using https://konklone.io/json/ to convert JSON files to CSV.
Lesson 2: Finding Data Online

In the digital age, more data is more available than ever before. In fact, sometimes it feels like we are drowning in data and it is difficult to find the data we are actually looking for. In this lesson, we will explore ways to find data online both through portals and by searching for it. We will also look at options for when the data we want isn’t available and we need to collect data ourselves through ‘crowdsourcing’ or sensors.

The flowchart created by Paul Bradshaw shows common ways journalists try to access data and what they do when they face roadblocks along the way. This should be a reference chart for you when you start your own data search, hit a wall, and don’t know what the next step is: http://onlinejournalismblog.com/2011/09/06/gathering-data-a-flow-chart-for-data-journalists-2/

From Evidence to Stories: Thinking Like a Data Journalist
Using Advanced Search

There are many sources on data on the internet. A useful technique of finding data online is to use Google’s advanced search.

**GOOGLE ADVANCED SEARCH**

Open [http://www.google.com/advanced_search](http://www.google.com/advanced_search)

A screen with several search fields appears. The following table explains various search options within Google advanced search. It also provides alternative shortcuts to perform the same search using the regular Google search that you may be familiar with.

<table>
<thead>
<tr>
<th>Google advanced search feature</th>
<th>Alternative option on regular Google search</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All these words</strong> is like a regular Google search</td>
<td>Type in all the words you want to find in the regular search bar</td>
</tr>
<tr>
<td><strong>Exact word or phrase</strong> helps find results in</td>
<td>Use quotes to search – for example “Ministry of”</td>
</tr>
</tbody>
</table>

From Evidence to Stories: Thinking Like a Data Journalist
<table>
<thead>
<tr>
<th>Which the words appear in the exact order you mention</th>
<th><em>Labor, Public Service and Human Resources Development</em></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Any of these words</strong> helps find results where any of the mentioned words appear</td>
<td>Use OR between words in a search – for example, <em>agriculture OR farming OR crops</em></td>
</tr>
<tr>
<td><strong>None of these words</strong> will filter out search results with words that you specify</td>
<td>Type the minus sign before the word you want to emit in a search – for example, <em>Montenegro -tourism</em></td>
</tr>
<tr>
<td><strong>Language:</strong> specify the language of the results</td>
<td>-</td>
</tr>
<tr>
<td><strong>Region:</strong> limit results to only websites from a geographical region</td>
<td>-</td>
</tr>
<tr>
<td><strong>Last Update:</strong> limit results to recent content</td>
<td>-</td>
</tr>
</tbody>
</table>
| **Site or domain:** Narrow search to specific website | Use this format to search - *site:url*  
For example: *site:http://www.who.int/*  
Note that the website address has to be EXACT.  
- **CORRECT**  
  *site: https://www.unodc.org/*  
- **WRONG**  
  *site: WorldHealthOrganization.org* |
| **Filetype:** Search only for files with a specific extension (for instance: xls, pdf, csv, doc) | Use this search format - *Filetype:[extension]*  
For example, here is a search term to look for XLS files:  
- **CORRECT**  
  *filetype:xls*  
- **WRONG**  
  *filetype:Excel* |
Now let’s try using Google advanced search:

- Use ‘any of these words’ to find content about malnutrition, hunger or starvation in your country.
- Use ‘none of these words’ to find information about malnutrition not about children.
- Find content about Zika only in French.
- Find content about Zika only from Brazilian websites
- Find content about Zika published in the last week.
- Search the Ministry of Health website for Excel files
- Search for PDFs about maternal health in your country.
Setting Up Alerts

If you interested in a particular topic, you can also use the following techniques to receive alerts or updates when something new appears online.

Google Alerts to follow topics

- **Step 1:** Sign into your Gmail
- **Step 2:** Go to https://www.google.com/alerts
  - Alternatively, you can use http://www.talkwalker.com/alerts
- **Step 3:** Create alert. Be specific. Put in the topic and region or person of interest.
- **Step 4:** Select how often, source, language, region and how many.
- **Step 5:** Turn alerts on and off as you follow stories.

Change Detection to track new content uploaded on websites

- **Step 1:** Open www.changedetection.com/
- **Step 2:** Open a website that regularly (but not too frequently) uploads new data or reports
- **Step 3:** Copy the URL of that website into the search window of the change detection software
- **Step 4:** Receive alerts when new content is uploaded to the site
Advanced Google Searches: Scavenger Hunt!

Use Google Search to find:

- A PDF report on education in your country
- A PDF report on the UNICEF website about immunization in the region
- The 2016 national budget for your country
- Your country’s annual exports from www.tradingeconomics.com/
- An Excel file with data about migrants from your country
- An estimated population projection from the national statistics website
- The inflation rate in your country for the last 20 years
- News about HIV in your country from the last month
Using Data Portals

With a global push for open data many governments, international organizations are creating their own open data portals. These portals are a source of rich data and it’s important to understand how to use a variety of interfaces to access and download desired data.

International, government, civil society and university databases are all fantastic sources of data. However, they all have their own interface that is a little bit different and require some exploration to understand how to navigate.

This is a general guide for how to navigate databases:

**Select a database**

In many cases, a website will have many databases and the first step is to select which database you want to search. For example, on the World Bank Data portal, you can select to search only for health data, only for education data or development indicators, among many more options.

**Select a geographical region**

There are many ways to compare how your geographical area compares

From Evidence to Stories: Thinking Like a Data Journalist
to others. You can compare neighboring cities, states, or countries, regions with a similar level of economic development or population.

**Select indicators**

Often databases will allow you to check boxes to identify which indicators you want to compare. It is best to select a wide range, look for interesting trends, and narrow down your focus later.

**Select a time period**

There is a higher probability of finding enough data points to identify trends over a large span of time. In many cases, data will be collected in different countries in different years so it is best to start with a wide search and then narrow down the time period once you know what years have data points.

**Select a format**

Often databases will allow you to see a table, map or visualization of the data. These can be useful overview tools. What we are most interested in is downloading the data either in CSV or Excel format. Visualizations can be useful to identify patterns but generally we want to work with the raw dataset ourselves.
National Databases

There are several places to access national data portals:

- [https://www.opendatasoft.com/a-comprehensive-list-of-all-open-data-portals-around-the-world/](https://www.opendatasoft.com/a-comprehensive-list-of-all-open-data-portals-around-the-world/)
- [https://investigativedashboard.org/](https://investigativedashboard.org/)
International Databases

In addition to national databases, there are many international data sources:

World Health Organization
United Nations
Population Reference Bureau
UNICEF Data
The Guardian's world government data portal
Google's public data directory
The data hub
DBPedia Datasets
Factual
Free GIS data
List of open data resources
Energy data repositories

From Evidence to Stories: Thinking Like a Data Journalist
World Research Institute
Data wrangling
Quora thread: "Where can I find large datasets open to the public?"
Directory of APIs
Infochimps
Datamarket
Offshore Leaks
Investigative Dashboard
Open Corporates
Natural Earth data
UNEP Data
Transparency International Corruption Index
Land Ownership Database
Gapminder World
Global Data Lab
Navigating International Databases

Try this example to download data about Sudan and its neighboring countries from an international database:


3. Under ‘Country’, select:
   - Albania
   - Armenia
   - Azerbaijan
   - Belarus
   - Bosnia and Herzegovina
   - Croatia
   - Cyprus
   - Georgia
   - Kazakhstan
   - Kosovo
   - Kyrgyzstan
   - Moldova
   - Montenegro
   - Romania
   - Russian Federation
   - Serbia
   - Tajikistan
   - The former Yugoslav Republic of Macedonia

From Evidence to Stories: Thinking Like a Data Journalist
• Turkey
• Turkmenistan
• Ukraine
• Uzbekistan

4. Scroll down, and click **Series**. The available indicators are listed.

5. Click on the **Filter** icon and select **Medical Resources and Usage**.

6. Check the boxes for **Nurses and midwives (per 1,000 people)** and **Physicians (per 1,000 people)**.

7. Now, click **Years**. The available years are listed.

8. Select the years of your interest, say the last 15 years. Click on **Apply Changes**.

9. Then click on **Table** on the top right corner when your selection is ready. You can always click on the menu on the right to change selections

10. Click the **Download Options** button, and download your data as an Excel file.

11. Open your data file in Excel.

**Making Requests for Government Data**

Article 19 of the Universal Declaration of Human Rights states that everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers.

Many countries that have access to information laws lack rigorous regulations and procedures to respond to data and information requests. With the law still very new in many countries, it’s essential that journalists actively submit requests to ensure that these procedures and regulations are developed and pave the way for data sharing systems between government and citizens.

To find out more information about freedom of information in general visit [http://foiadvocates.net/](http://foiadvocates.net/)

To find out the status of access to information laws in your country check: [http://www.freedominfo.org/regions/](http://www.freedominfo.org/regions/)

These are many of the excuses that you will get for denying access to information requests. Remember, they are just excuses! You have to be specific in your requests and persistent in order to get the data you need.

From Evidence to Stories: Thinking Like a Data Journalist
• “We don’t have that data on a computer.”
• High fees
• Delay tactics
• “Your request was unclear.”
• Sending the wrong data
• “Our database is too complicated to give you access.”
• “Our database software is proprietary.”
• “That information is protected by privacy law.”
Lesson 3: Alternative Data Sources

Often, when official data sources are not available, organizations may use “crowd-sourcing” to solicit data from citizens or from a trained network of volunteers. They may also use sensors, citizen reports, media reports or leaked data as alternative data sources.

For example, take a look at the Afghan Election Violence Map: http://www.tfp.nai.org.af/map/main

In this map, election monitors in Afghanistan sent SMS alerts documenting violence and irregularities on election day. The monitors were trained in how to use the system. The data was used by journalists to report on the status of polling stations across the country.
Crowd-sourced data

One of the most well-known platforms for collecting ‘crowd-sourced’ data is Ushahidi. This platform has been deployed to map natural disasters, political crises, and other events where live data collection can inform a response. This system enables an open or closed network to submit reports of incidents (such as violence, ballot box tampering, police harassment), which is sent to a centralized system for verification, addition to the database, and mapping.

For example, a website called Uchaguzi\(^2\) was set up using Ushahidi for Kenyan Election Monitoring. In this example, media houses tapped into a centralized network of election monitors who were tracking and categorizing election incidents.

\(^2\)https://www.facebook.com/ushahidi/photos/a.193585313994844.42244.116038145082895/543807175639321/?type=1&theater

From Evidence to Stories: Thinking Like a Data Journalist
Citizen Media Contributions

In many cases, media houses engage their audience to send in their own raw material or eye witness reports via SMS, video, or photos – referred to as citizen media contributions.

For instance, Al Jazeera\(^3\) is one of many media who have resorted to social media and citizen media accounts to report on the situation within Syria. Most media houses have verification policies in place but they are not immune to publishing false content.

\(^3\) [http://www.stream.aljazeera.com/story/201105112039-003652](http://www.stream.alazeera.com/story/201105112039-003652)

From Evidence to Stories: Thinking Like a Data Journalist
Sensors

Data collected by sensor data has been used for reporting on environmental issues. For instance, a media house may distribute small, inexpensive sensors to a trained community of volunteers to gather data such as air quality, water temperature, or earthquake activity.

For example, Ekuatorial⁴ is an attempt to collect data about the Indonesia rainforest, an area that is remote and difficult to monitor. Sensors have been left with members of communities of these remote areas to collect ground level data on environmental conditions which are combined with national and satellite data to track environmental degradation.

---

⁴ http://ekuatorial.com/en
Drones

Increasingly, media are using unmanned aircraft to estimate the size of protests, measure garbage dumps, calculate the rate of deforestation, and gather other useful information from an aerial view.

For example the Wall Street Journal shared videos taken by drones to demonstrate the scale of protests in Hong Kong: http://www.wsj.com/video/aerial-drone-captures-scale-of-hong-kong-protests/76AA792E-7AB9-4D2B-88BB-E9B5F9D707EC.html

Though the technology is available, privacy issues and creating the effect that protesters are being spied on by drones that could be from government can cause suspicion.
Mining Newsroom Data

For subjects with poor official documentation, crowd-sourcing media reports on the topic can often yield a rich store of data. This strategy has been used to collect data on violence against women, people killed in police shootings, and Chinese aid to Africa.

For example ‘The Migrant Files’ sources reports of migrant deaths in the Mediterranean from global media reports. When combined, these reports paint a much more complete picture of migrants’ deaths in their journey to Europe.

---

5 http://www.themigrantsfiles.com/

From Evidence to Stories: Thinking Like a Data Journalist
Risks

There is an inherent problem with using crowd-sourcing to measure public service delivery for several reasons: those with poor access to public services usually also have poor access to the telecommunications needed to report on the issue.

- **Selection bias**: only people with time, resources and motivation are likely to contribute
- **Verification process**: people could flood the system with erroneous reports and it is difficult to find out which are real
- **Context**: from crowd sourced data, we only know what the crowd tells us so much of the contextual information that we would usually include to explain data, is lost.
- **Privacy**: sometimes personal data about contributors can be accessed by agencies that might want to target critics

For example, in this power outage reporting system in Nairobi\(^6\), slum areas - with poor access to electricity in the first place - are under-reporting power outages.

\(^6\) [http://poweralerts.kenyapower.co.ke/tweetmap](http://poweralerts.kenyapower.co.ke/tweetmap)

*From Evidence to Stories: Thinking Like a Data Journalist*
Lesson 4: Planning a Data Story

The prevalence of data in the modern world has changed the way humans receive information. Now, more insight and solutions can be developed by enriching traditional information channels with data. This lesson will review how journalists organize data projects to ensure a successful story.

When the Sun Sets in Turkana

Take a look at this news story: https://www.youtube.com/watch?v=Ga8CEYVALo4

http://www.internewskenya.org/summaries/internews52e7747b74fff.pdf

In this example, the journalist investigates a news story: the drought in Turkana and widespread starvation, with a data lens:

- She looks at climate data to determine that droughts are increasing in severity and frequency.
- She uses health data to determine the health impact of malnutrition on children.
- She uses international aid data to determine if there is a long-term solution to the problem: investing in food security instead of humanitarian aid.

The key to success for any data journalist is organization. Unlike in many other kinds of journalism, how you decide to organize your information and narrate your story can make or break your story. The process we will follow for organizing a data journalism story consists of six steps:

From Evidence to Stories: Thinking Like a Data Journalist
As we proceed through this course, we will cover the tasks and skills required for each step. In this lesson, we will cover the two planning stages: background and hypothesis and questions.

**Background**
When first identifying a topic for a data story, the first step is to search for other data stories produced by other journalists on the same topic. This serves several purposes. It familiarizes you with how other journalists have approached the issue, where and what kind of data he or she used and what storytelling strategy was effective.

**Cases of violence against women: Is mediation the best option?**

In this story, a team of journalists decided to cover the issues of domestic violence services in Afghanistan.

**Case studies**

Using advanced Google search techniques, they identified three similar data stories in the media:

- *India is a Nation of Violent, Stressed Men*, IndiaSpend, India
- *Till Death Do Us Part*, Post and Courier, USA
- *Most Dangerous Transport System for Women*, Global Post, Global

From these examples the journalists noted:

- The government’s ability to provide services is key in determining whether or not victims of domestic violence survive their experience.
- Rates of domestic violence and reporting vary widely across geographical regions
- Visualizations can be effective in showing the scope of the problem

*Reports/Data*
In the next stage, journalists identify reports and data related to the subject of the investigation. These reports can be found through searches, through data requests to the government and CSOs or through the creation of data for the investigation.

For the domestic violence story, journalists identified the following key reports:

<table>
<thead>
<tr>
<th>Report</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Justice through the Eyes of Afghan Women: Cases of Violence against</td>
<td>United Nations Assistance Mission in Afghanistan</td>
</tr>
<tr>
<td>Women Addressed through Mediation and Court Adjudication UNAM?OCHA</td>
<td></td>
</tr>
<tr>
<td>An Update on Implementation of the Law on Elimination of Violence</td>
<td></td>
</tr>
<tr>
<td>against Women in Afghanistan</td>
<td></td>
</tr>
<tr>
<td>AFGHANISTAN Ending Child Marriage and Domestic Violence</td>
<td></td>
</tr>
<tr>
<td>USIP Women’s Access to Justice in Afghanistan</td>
<td></td>
</tr>
<tr>
<td>World Bank Gender Data Portal</td>
<td></td>
</tr>
</tbody>
</table>

**Methodology**

To evaluate the source of the data, journalists answer the following questions (here with sample answers from the first report)

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who gathered the data?</td>
<td>United Nations Assistance Mission in Afghanistan</td>
</tr>
<tr>
<td>When was the data gathered?</td>
<td>Detailed information from 18 of Afghanistan’s 34 provinces for the one-year period October 2012 to September 2013 with technical review by the UN Office of the High Commissioner for Human Rights</td>
</tr>
<tr>
<td>What time period does the data cover?</td>
<td>Two years (2012-2013)</td>
</tr>
<tr>
<td>How was the data gathered?</td>
<td>Field monitoring and analysis of police and court records</td>
</tr>
</tbody>
</table>

**Important findings**

Finally, journalists read the executive summary of the report and write down 3-5 interesting findings.

1. The number of reports of domestic violence is rising
2. The economic and social vulnerability of women remains constant
3. Most cases of domestic violence are registered through the police
4. Most cases of domestic violence are resolved through mediation

From Evidence to Stories: Thinking Like a Data Journalist
Hypothesis and Questions

The most important stage of the data story process is the formulation of the hypothesis and questions. In scientific research, a researcher develops a hypothesis with a suspected set of ideas and then builds an experiment to support the hypothesis. The same process is true for data journalism. The journalist, using his or her news nose, develops a hypothesis that can be proved or disproved with data. News articles have a habit of bringing up more questions than they answer, and following up with a hypothesis and investigation can uncovered details that lead to a data journalism story.

Building a Hypothesis for Data and Investigative Stories

1. A hypothesis gives you something to verify, instead of trying to uncover a secret. People do not give up their secrets without a very good reason. They are much more likely to offer confirmation of information that is already in your possession, simply because most people hate to lie. A hypothesis enables you to ask them to confirm something, rather than to advance information. It also puts you in the position of someone who is open to discovering that there is more to the story than he or she thought at first, because you are willing to accept that there are facts beyond what you suspected at the start.

2. A hypothesis increases your chances of discovering secrets. A lot of what we call “secrets” are simply facts that no one ever asked about. A hypothesis has the psychological effect of making you more sensitive to the material, so you can ask those questions. As the French investigator Edwy Plenel said, “If you want to find something, you have to be looking for it.” We would add that if you’re really looking for something, you’ll find more than you were looking for.

3. Hypotheses makes it easier to manage your project. Having defined what you’re looking for, and where to start looking for it, you can estimate how much time the initial steps of the investigation will require. This is the first step to treating an investigation as a project that you can manage. We’ll return to this point at the end of this chapter.

4. Hypotheses are a tool that you can use again and again. When you can work in a methodical way, your career will change. More important, you will change. You will no longer need someone to tell you what to do. You will see what needs to be done to combat some of the chaos and suffering in this world, and you will be able to do it. Isn’t that why you became a journalist in the first place?

5. A hypothesis virtually guarantees that you will deliver a story, not just a mass of data. Editors want to know that at the end of a specific period of time – a specific investment of resources – there will be a story to publish. A hypothesis hugely increases the likelihood of that outcome. It enables you to predict a minimum and maximum positive result for your work, as well as a worst case.

---

7 http://unesdoc.unesco.org/images/0019/001930/193078e.pdf

From Evidence to Stories: Thinking Like a Data Journalist
- The worst case is that verification of the hypothesis will quickly show there is no story, and the project can be ended without wasting significant resources.

- The minimum positive outcome is that the initial hypothesis is true, and can be quickly verified.

- The maximum is that if this hypothesis is true, others must logically follow, and either a series of related stories or one big story will result.

Tips for a strong hypothesis:

- Posits a theory that can either be proven or disproven with data

- Is specific about what is being measured

- The data is available

- The topic is important to the public

Below is an example of how to transform a weak hypothesis into a strong hypothesis:

1. Children in this country are dying of malnutrition.

2. Most children in this country who die below the age of five die of malnutrition. *(Add an indicator that can be measured “under five cause of death.”)*

3. Most of the children in this country who die below the age of five die of malnutrition and live in the poorest provinces. *(Add a geographical measure and an economic measure)*

4. Most of the children in this country who die below the age of five die of malnutrition and live in the poorest provinces despite a donor funded feeding program that was supposed to cut malnutrition rates in half over five years. *(Add a measure of whether the solution to the problem is working or not)*
Hypothesis

After completing the background section for their investigation the Afghan journalists developed the following hypothesis:

**Government programs cannot keep up with the increased demand for domestic violence services.**

As you can see this hypothesis posits two theories that can be proven or disproven with data:

- Domestic violence reporting rates are increasing
- Government services to respond to these reports are inadequate.

Please evaluate the following hypotheses, writing **S** (Strong) or **W** (Weak). If weak, please rewrite in the space provided.

1. The Ministry of Health should spend more on public health ____
   ____________________________________________________________
   ____________________________________________________________

2. Average primary school test scores are rising but fewer rural and low income families are able to send their children to school.____
   ____________________________________________________________
   ____________________________________________________________
   ____________________________________________________________

From Evidence to Stories: Thinking Like a Data Journalist
3. The decline in government spending for public health over the last five years has contributed to a lack of progress in achieving Millennium Development Goals to improve health. ___
______________________________________________________________________________
______________________________________________________________________________

4. Lack of hospitals are responsible for high disease rates. ___
______________________________________________________________________________
______________________________________________________________________________
______________________________________________________________________________

5. The rate of under-five child deaths is going down. ___
______________________________________________________________________________
______________________________________________________________________________
______________________________________________________________________________

6. For the Turkana story above, what do you think the journalist’s hypothesis was?
______________________________________________________________________________
______________________________________________________________________________

Questions
Once you have a strong hypothesis, you should develop at least five questions that can be answered with data to prove or disprove your hypothesis. All the questions should be able to be answered with a number. Other types of questions, such as interview questions, will be developed after analysis.

The questions you write should:

• Measure the trend.
• Compare different groups.
• Measure the causes.
• Measure the impact.

Questions
Remember the hypothesis for the domestic violence story:

From Evidence to Stories: Thinking Like a Data Journalist
**Government programs cannot keep up with the increased demand for domestic violence services.**

These are the questions the journalists wrote:

1. Are the number of cases being reported each year going up or down?
2. What is the age breakdown of women registering domestic violence cases under the new law? Are younger women registering cases?
3. What percentage of women have experienced domestic violence in Afghanistan?
4. How bad is the situation for women in AF compared to the rest of the world?
5. What percentage of cases that are registered go to court? What happens to the rest of the cases?
6. What kinds of violence are being prosecuted under the new law?
7. What percentage of cases that make it to court result in a conviction?

**Lesson 5: Enriching Stories With Data**

Many stories that include data and statistics are not truly data journalism stories because they do not use data to explain the underlying issue. One of the most important skills as a data journalism is to recognize opportunities to transform an ordinary story into a data story. For each of the stories below, read the existing story and transform it into an idea for the data story. In some cases, the journalist has a hypothesis but fails to prove it with data. In other cases, there is data and statistics but it unclear what the journalist is trying to prove.

**Enriching Stories with Data: Health Reporting**

![Decline of Fistula Infection among Women](image)

From Evidence to Stories: Thinking Like a Data Journalist
Here is a news article that claims that cases of urine fistula among women in Sudan have declined due to the intensification of awareness campaigns:

Let’s critically evaluate the claims in this news story from a data perspective – and further suggest what data would be needed to support the claim.

**Hypothesis:**

_____________________________________________________________________________________
_____________________________________________________________________________________
_____________________________________________________________________________________
_____________________________________________________________________________________

**Consider:**

- What data would you need to answer whether fistula infection is going down?
- What data would you need to understand factors?
- What other maternal complications could you compare this disease to?
- Do you trust the data?
- What other data would you want to assure you that there is a trend?

**Questions:**

_____________________________________________________________________________________
_____________________________________________________________________________________
_____________________________________________________________________________________
_____________________________________________________________________________________
_____________________________________________________________________________________
_____________________________________________________________________________________

From Evidence to Stories: Thinking Like a Data Journalist
Two million measles vaccines delivered to Sudan for mass immunisation

The arrival of the vaccines comes following the signing of the mid-term review report report of the 2013-2016 joint country programme of cooperation agreement between the Sudanese government and UNICEF.

The agreement aims to reach an estimated 3.8 million children across the country with health, education, water, sanitation, nutrition, hygiene promotion and protection interventions.

While UNICEF said Sudan had made some progress towards improving the health and quality of life for children across the country, more needed to be done.

“Despite the gains made, more still needs to be done to address the myriad and dire needs of children in Sudan,” said UNICEF representative Geert Cappelaere. This includes more investment in boosting school attendance levels, the management of acute malnutrition, sanitation coverage and reducing the outbreak of diseases.

Here is a news article that attempts to link resources for health care available in Sudan to the impact on the health of children in the country: http://www.sudantribune.com/spip.php?article54473

Here is a statement from the article: “While UNICEF said Sudan had made some progress towards improving the health and quality of life for children across the country, more needed to be done.”

Hypothesis:

Consider

To find out whether investments are improving healthcare in Sudan – here’s the kind of data to find:

- How much of the national budget is allocated to healthcare?
- What is the per capita government spending on healthcare?
- How does per capita healthcare spending compare to other countries?

Questions:

From Evidence to Stories: Thinking Like a Data Journalist
Enriching Stories with Data: Conflict Reporting

Sudanese presidency calls for monitoring gasoline distribution to prevent a looming fuel crisis

June 9, 2015 (KHARTOUM) – The Sudanese presidency has instructed the ministry of oil to activate the work of the committee tasked with monitoring the distribution of gasoline in order to bring what it called the “petroleum products manipulation” to an end.

Sudan lost 75% of its oil reserves after the southern part of the country became an independent nation in July 2011, denying the north billions of dollars in revenues. Oil revenue constituted more than half of Sudan’s revenue and 90% of its exports.

In September 2013, protests erupted in Sudan’s major towns following an announcement by the government that it was reducing subsidies on fuel and other basic commodities, leading to calls for regime change.

At least 200 protesters died, 15 of them children and more than 800 others have been detained.

This is a very bland article about gas shortages in Sudan, not touching on protests related to the gas shortages: http://www.sudantribune.com/spip.php?article55283

Hypothesis:

Consider

To enrich this story, can we find data related to:

- How many people have died in oil related conflicts in the country?
- How do deaths related to oil protests compare to other kinds of civil unrest?
- Which exact distribution irregularities is the article referring to? Can these be quantified?
- What manipulations is the author referring to and who is responsible for oversight?

Questions:
Enriching News Stories with Data: Aid reporting

This is an article about aid from Germany to Sudan to address the humanitarian crisis:

Hypothesis:

Consider

The trouble is that while it gives a total amount of aid, it doesn’t put this number in to context, comparing it to other donors or explaining where the money is being spent. Can you find:

- How much international aid does Sudan receive each year?
- Where is it spent? By sector? Region? Per capita?
- How can data be used to ensure accountability in aid spending?

Questions:
Enriching News Stories with Data: Economic reporting

Sudan’s inflation rate declines to \(19.8\%\) in May

June 6, 2015 (KHARTOUM) - Sudan’s Central Bureau of Statistics (CBoS) reported on Saturday that the monthly inflation rate has dropped to \(19.8\%\) in May from \(24.3\%\) in April.

The CBoS said in its monthly bulletin released Saturday the average index has increased by \(5.9\) points in May, pointing the price index in urban areas reached \(465.9\) points compared to \(459.3\) points in April.

According to the bulletin, the price index of the food and beverages has reached \(481.3\) points compared to \(469.7\) points in April while the annual rate of price change in urban areas has decreased to \(22\%\) compared to \(22.3\%\) last year.

This is a typical dry economic story reporting on changing inflation rates:
http://www.sudantribune.com/spip.php?article55242

Hypothesis:

Consider

It does not do a good job of linking inflation rates to how changing costs impact the wallet of Sudanese citizens.

- What data is available to put inflation in a larger context?

From Evidence to Stories: Thinking Like a Data Journalist
• How much has price inflation of household goods affected a Sudanese family’s household budget?

• How does an average Sudanese wage compare to the average monthly household expenses?

• On average, how many people does a Sudanese citizen with a job support financially?

• How can the price of consumer goods be made into a subject that is interesting to the public?

Questions:

_____________________________________________________________________________________
_____________________________________________________________________________________
Enriching News Stories with Data: Economic reporting

In Sudan, Aid Groups Struggle With Massive Measles Outbreak

Sudan currently is trying to persuade the international community to relieve it of its outstanding foreign debt, which the International Monetary Fund estimates at $46 billion.

UNICEF representative Cappelaere tells VOA that creditors should not just look at the debt from a purely financial perspective. He says he would like to see that debt relief, if granted to Sudan, made conditional on the government spending some of that money for the health and well-being of its people.

Read: [http://www.voanews.com/content/aid-agencies-struggling-to-contain-massive-sudan-outbreak/2750234.html](http://www.voanews.com/content/aid-agencies-struggling-to-contain-massive-sudan-outbreak/2750234.html)

Here is a news article which mentions that Sudan is trying to persuade the international community to cancel its foreign debt in the context of the massive measles outbreak in the country. You need to evaluate the claim that cancelling foreign debt can improve health outcomes, and further suggest the data that you would need to assess this claim.

Hypothesis:
Consider

- What data would we need to evaluate whether canceling Sudan’s external debt would lead to better healthcare outcomes?
- Can we compare other countries who have had their external debt canceled? Did it improve healthcare outcomes? What other factors were at play?
- Who could we interview about how canceling external debt would impact healthcare?

Questions:
Practice: Find Data Angles for Public Interest Topics

• For each of the following events, list a data angle that could explain the issue:

  • A teachers’ strike in which teachers are demanding a pay raise and reduced class size

  • A food shortage in two states in Sudan

  • An outbreak of measles among children in rural areas

  • A program to install solar panels in villages without electricity

In this exercise, we will brainstorm public interest topics that would benefit from statistical analysis to put the issue into context for the reader.

Scenario
You have been asked to map out different scenarios in which a trigger event is an opportunity for a deeper data driven analysis of a phenomenon. For each of the following events, list a hypothesis that could explain the issue:

1. A teachers’ strike in which teachers are demanding a pay raise and reduced class size

   Hypothesis:
   __________________________________________________________
   __________________________________________________________
   __________________________________________________________
   __________________________________________________________
   __________________________________________________________

2. A food shortage in two states in Sudan
3. An outbreak of measles among children in rural areas

Hypothesis:
________________________

4. A program to install solar panels in villages without electricity

Hypothesis:
________________________
Exercise: Analyzing Fact Sheets

Overview
In this exercise we will look at data that has been analyzed and visualized to provide an overview of health in your country. The goal is to interpret the data, begin to explore where the data came from, identify what is most interesting about the data, write a hypothesis and put together a story based on the findings.

To begin, open one of the followin:

Global burden of diseases, injuries, and risk factors profile:
http://www.healthdata.org/results/country-profiles

WHO country profile:
• http://www.who.int/gho/countries/en/ (Choose a country, select Country Profiles and select General health statistical profile)
• UNICEF country profile
  http://www.unicef.org/infobycountry/ (Select a country and select statistics)

Background
For each of the data sources, answer the following questions:

1. Who gathered the data?
   ________________________________________________________________
   ________________________________________________________________
   ________________________________________________________________

2. When was the data gathered?
   ________________________________________________________________
   ________________________________________________________________
   ________________________________________________________________

3. What time period does the data cover?
   ________________________________________________________________
   ________________________________________________________________
   ________________________________________________________________

From Evidence to Stories: Thinking Like a Data Journalist
4. How was the data gathered?

_____________________________________________________________________________________
_____________________________________________________________________________________
_____________________________________________________________________________________

**Understanding the indicators**

- What do the indicators mean?
- Can I look up definitions of indicators I don’t understand?
- What are the differences between the categories?
- What indicators are not included in this data that would provide more context?

*Underline 3-5 data points can answer these questions*

- What is interesting or surprising about this data?
- What in this data could help citizens make better decisions about their health?
- What in this data could help policymakers make better decisions about health spending?
- What in this data could explain big picture health trends in the country?

**Hypothesis**

*Write down a hypothesis that you can prove with this data:*

_____________________________________________________________________________________
_____________________________________________________________________________________

**Prepare the story**

Put your data points in order

- Start with the data that is most important to answer your hypothesis
- Add data that goes into further detail or provides context about the trend

_____________________________________________________________________________________

From Evidence to Stories: Thinking Like a Data Journalist
Module 2: From Data to Stories

Write the story

Evaluate the stories

1. Did the story answer the hypothesis?

2. Did each data point in the story support the hypothesis?

3. Did the data points come in a logical order in the story?