A data journalism primer

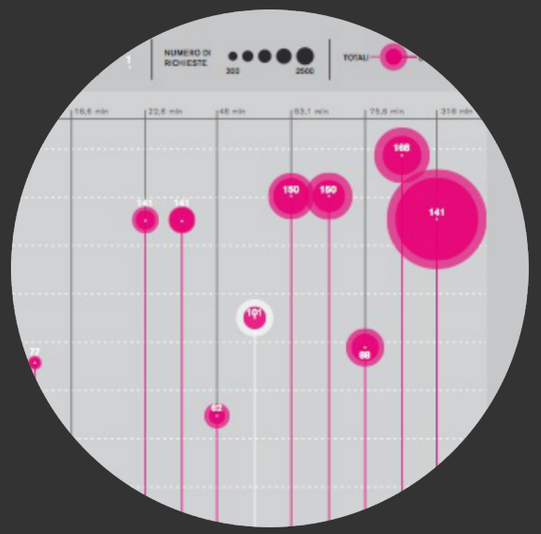
Alice Corona

alice-corona.nl
@alice_corona

JADS · Den Bosch, 18 May 2017

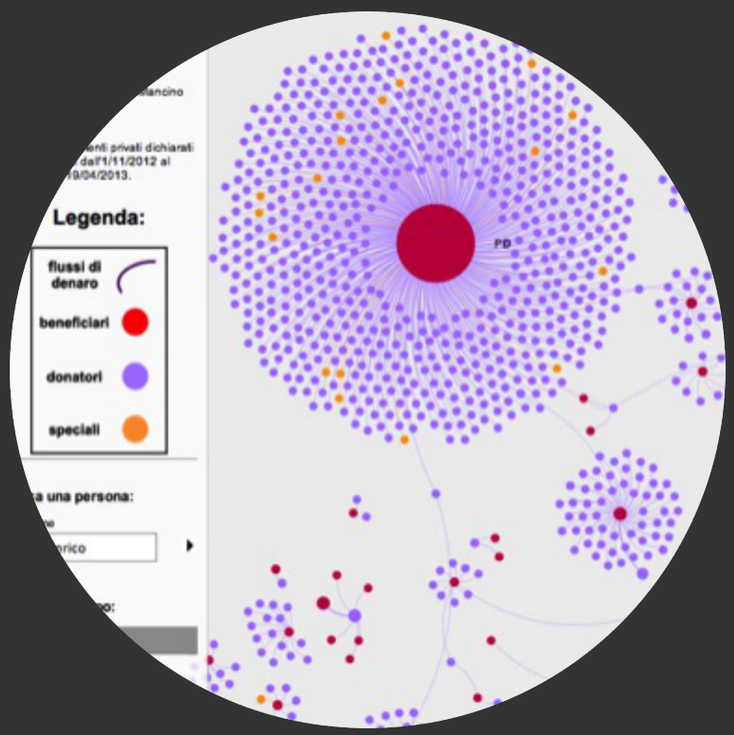
Alice Corona

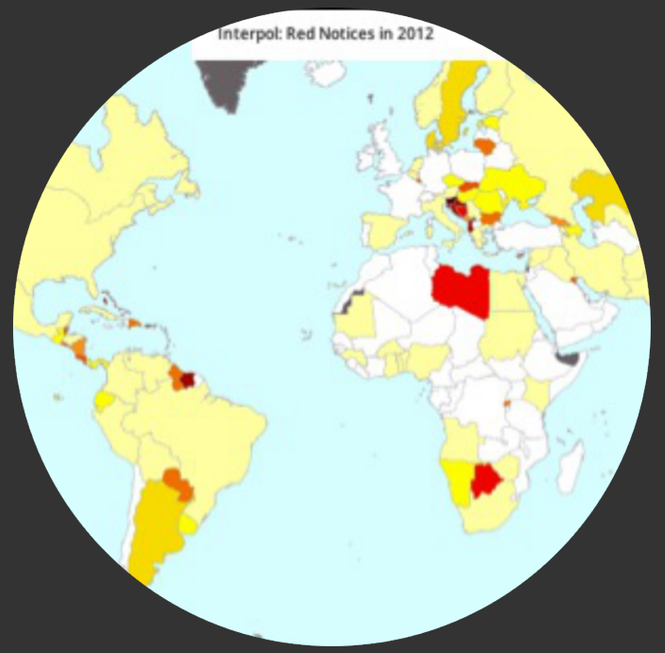
Data-Driven Storyteller · Consultant · Trainer · Researcher



I am a data journalist and a researcher specialized in data-driven methodologies for storytelling.

I've worked as a data journalist and I've counselled NGOs, non-profits and businesses on researching and producing data-driven stories. I've given talks, trainings and workshops on data-driven storytelling techniques and best practices. I was lead data journalist and trainer at Silk.co, a semantic web platform for publishing and visualizing data. Before, I've crafted data journalism stories and data visualizations for digital and print Italian news outlets.





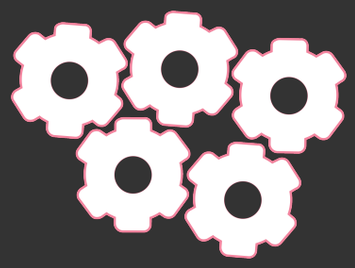
Alice Corona

alice-corona.nl
@alice_corona

JADS · Den Bosch, 18 May 2017

Outline

Introduction to data journalism



Elements of a data journalist's workflow

Tips and tricks

Alice Corona

alice-corona.nl
@alice_corona

Introduction to data journalism

Alice Corona

alice-corona.nl
@alice_corona

JADS · Den Bosch, 18 May 2017

What is data journalism?

Computer-assisted reporting (CAR)

Data-driven journalism (ddj)

Database journalism

Infographics

Precision journalism

Predictive journalism

Hacker journalist

Open data journalism

News apps

Alice Corona

alice-corona.nl
@alice_corona

What is data journalism?

Journalism

Alice Corona

alice-corona.nl
@alice_corona

Data journalism can be seen as a **method** and a series of **techniques** that journalists can apply in their reporting to uncover or reinforce stories using data as their source.

In its essence, its processes are hidden. It is in the “back end”. There are no “front end” cues, like data visualisations, that automatically let you tell a data journalism article apart from a regular journalism article

Alice Corona

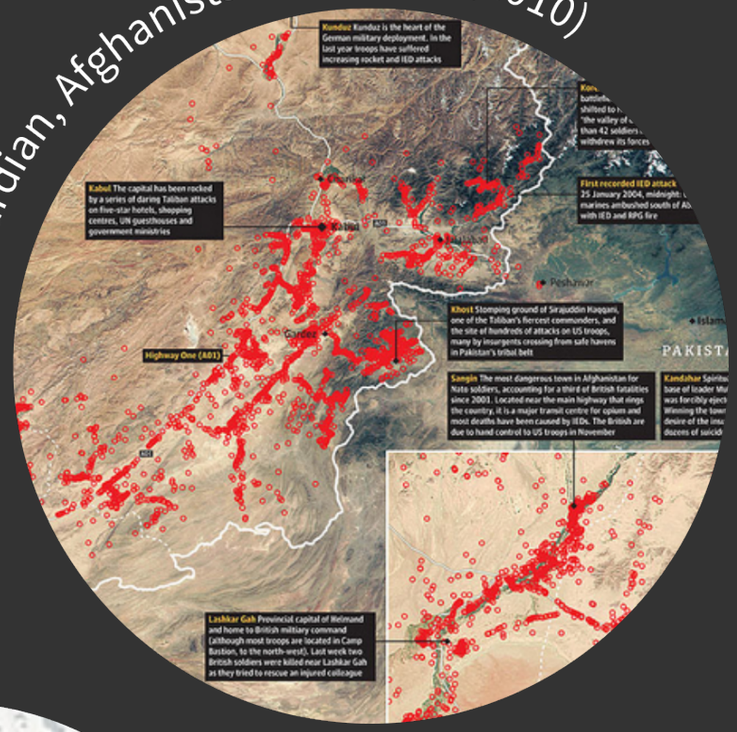
alice-corona.nl
@alice_corona

what is data journalism

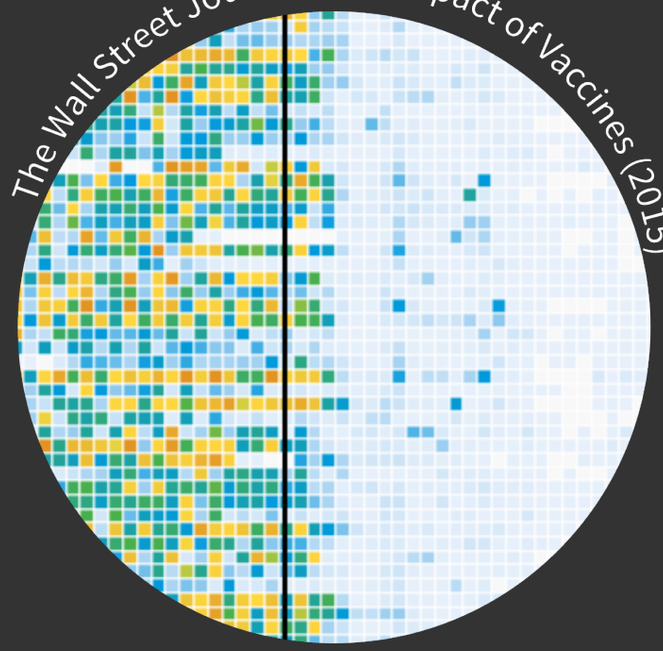
What data?

ICIJ, Panama Papers (2016)

John Snow, Map of Cholera Outbreak in London (1854)



The Guardian, Afghanistan War Logs (2010)



The Wall Street Journal, The Impact of Vaccines (2015)

Journalism++ & Others, The Migrant Files (2013)

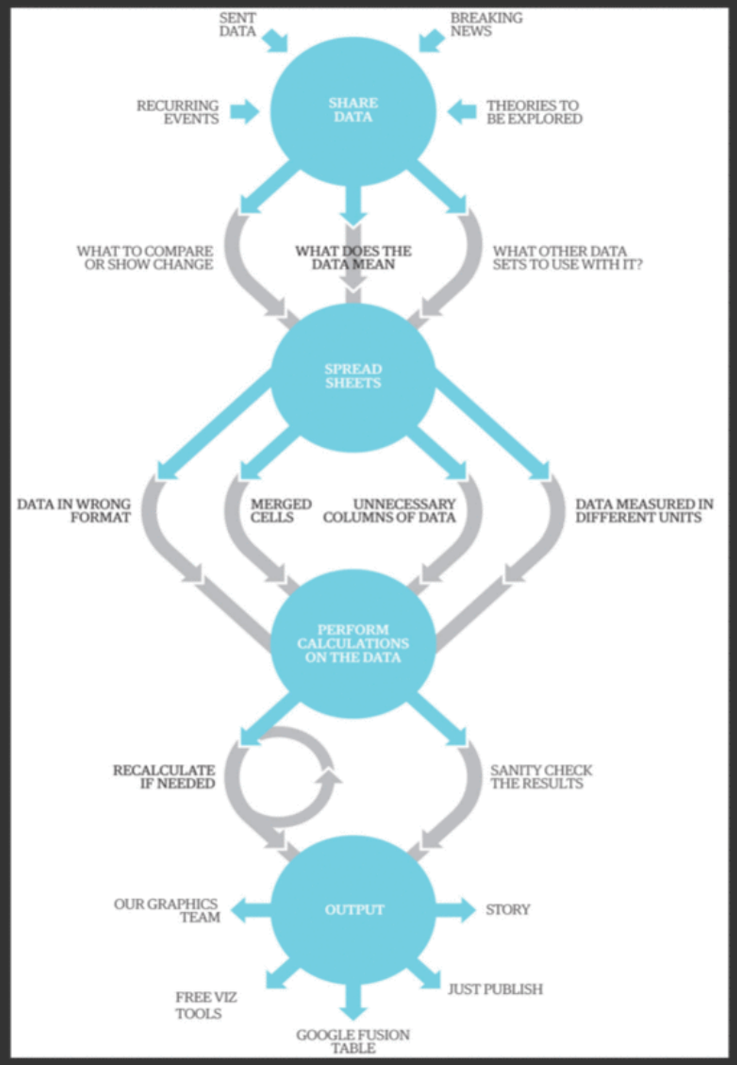
WNYC, Cicada Tracker (2013)

Alice Corona

alice-corona.nl
@alice_corona

How?

So many workflows, so little time…



Alice Corona

alice-corona.nl
@alice_corona

Elements of a data journalist's workflow

Alice Corona

alice-corona.nl
@alice_corona

JADS · Den Bosch, 18-6 August 2016

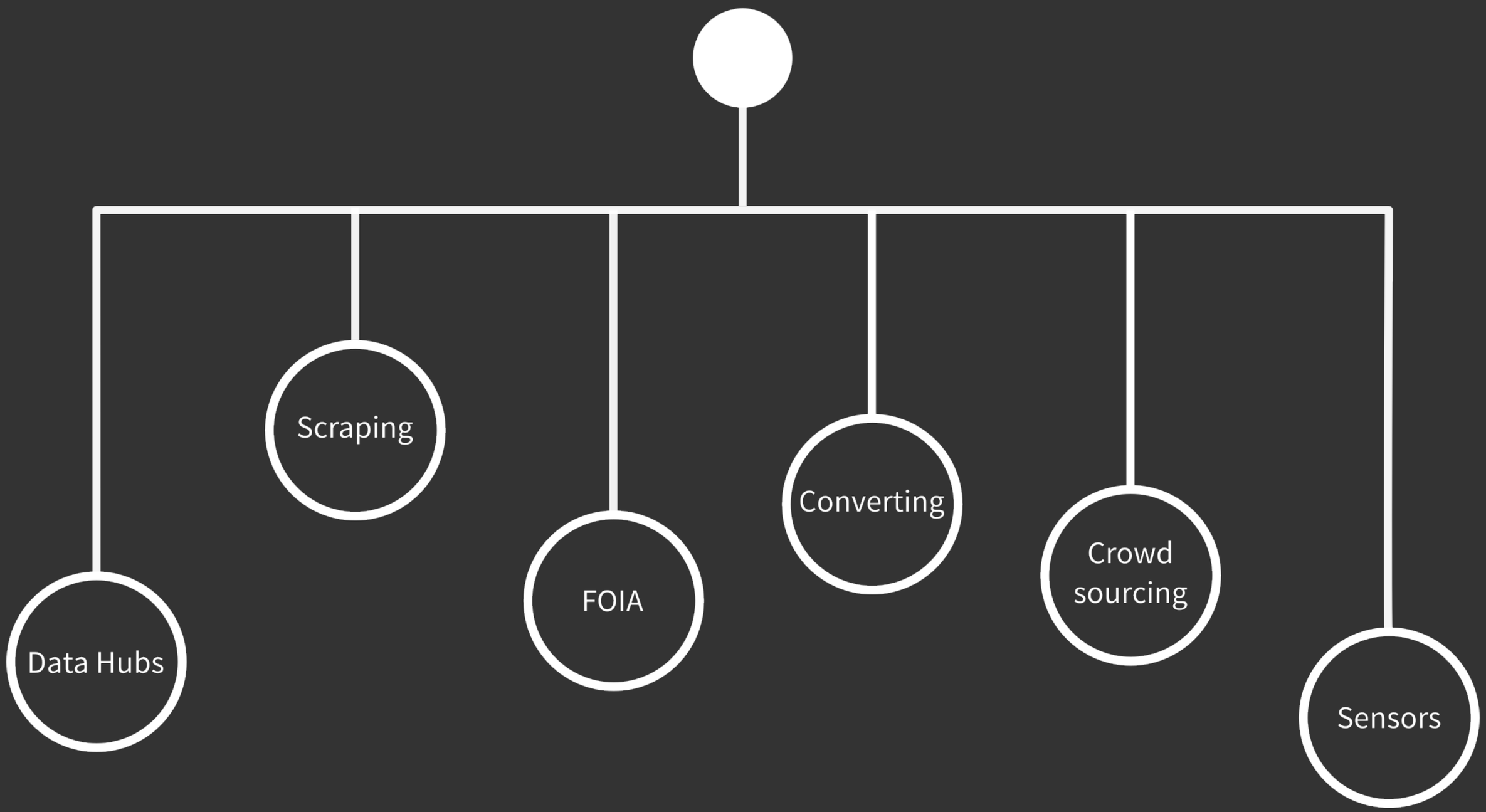
a data journalist's workflow

Data Sourcing

Alice Corona

alice-corona.nl
@alice_corona

a data journalist's workflow

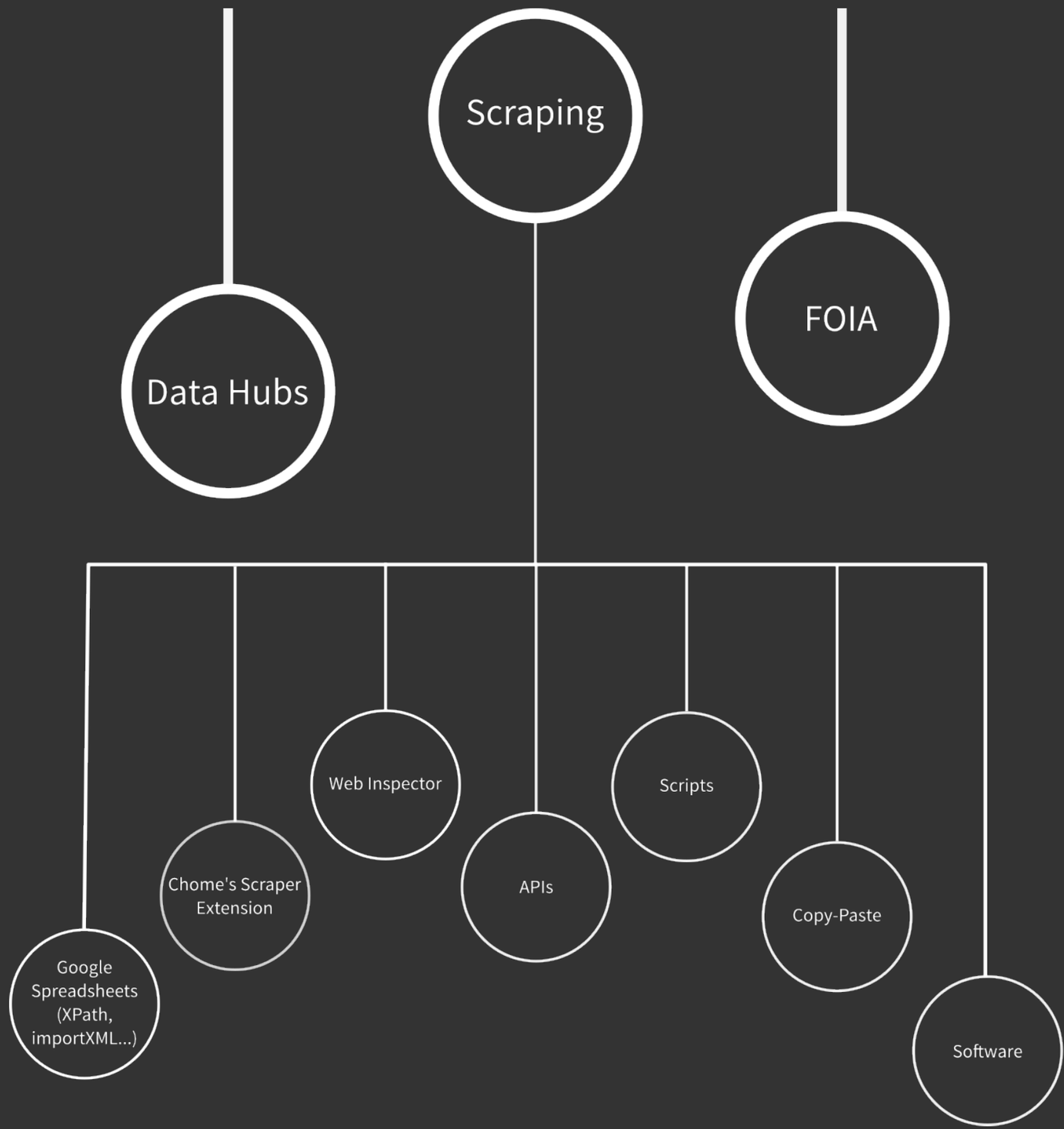


Alice Corona

alice-corona.nl
@alice_corona

JADS · Den Bosch, 18-6 August 2016

a data journalist's workflow



Alice Corona

alice-corona.nl
@alice_corona

When collecting data

* Investigate the **metadata** of your dataset: who collected the data, why did they collect it, why are they publishing it, how they collected it (always read the methodology), and make sure you understand what is computed with each variable. Write down potential biases, conflicts of interests, inaccuracies, etc.
- Consider the **ethical / legal aspects of scraping**, especially when using your own scripts to harvest data.

Alice Corona

alice-corona.nl
@alice_corona

a data journalist's workflow

Data Cleaning

Alice Corona

alice-corona.nl
@alice_corona

JADS · Den Bosch, 18-6 August 2016

In theory...

*"Process of **detecting and correcting (or removing)** corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data" (Wikipedia)*

Alice Corona

alice-corona.nl
@alice_corona

…and in practice

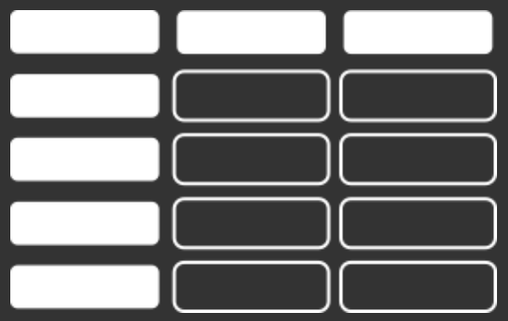
- Check data **constraints / consistency / format** within same column (example: 18/05/2017 vs. May 18 2017 ….)
- Check **measuring conventions** to make sure you have what you expect to use and, if not, transform accordingly (example: 18/05/2017 vs. 05/18/2017; 10.33 vs. 10,33; etc.)
- Check for consistent **handling of NULL values** and transform into desired format (example: NULL, N/A, *empty cell*, 0; - ; etc.)
- Check for consistent handling of **multiple values within same cell** and eventually transform into desired format
- Check **spelling, wording and capitalisation** to identify possible duplicates (example: Mr. John Smith; Smith, John; JOhn smith; etc.)
- **Disambiguate**, after spelling/wording/capitalisation checks (example: are there two John Smith that are actually different people? Might want to name them something like John Smith 1 and John Smith 2)
- **Remove duplicates** (after disambiguation and spelling/wording/capitalisation and all other checks)
- Data **enhancement**: combining datasets on the basis of a common column (example: column mobile number prefix to derive country column)
- Strive for high level of **data granularity** per column (example:address column with "Street,City,Country" could be duplicated and split to add city and country columns. Of course, do this reasonably with your goals.)
- Make sure the dataset follows guidelines for **proper structuring** and, if not, act accordingly to achieve this (1 type of variable, and in one consistent format, per column; one observation per row)

Alice Corona

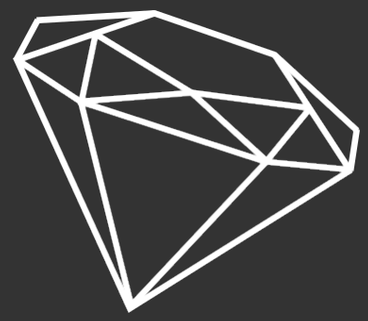
alice-corona.nl
@alice_corona

a data journalist's workflow

How?



Spreadsheets



Open Refine

Code it

Alice Corona

alice-corona.nl
@alice_corona

When cleaning data

- Biggest danger is loss of information. To mitigate, always **keep the originals** (of the whole dataset, of each column before transforming it…) and maybe also a copy of the dataset before each important transformation.
- **Document** steps. This is a must if you want to be transparent about your work, but it will also help you to replicate the process if needed in the future.

Alice Corona

alice-corona.nl
@alice_corona

a data journalist's workflow

Data Analysis

Alice Corona

alice-corona.nl
@alice_corona

JADS · Den Bosch, 18-6 August 2016

a data journalist's workflow

Where's the smoking gun?

Alice Corona

alice-corona.nl
@alice_corona

JADS · Den Bosch, 18-6 August 2016

When analysing data

- **Beware of findings and trends** that could be motivated by a change in the data collection or publishing methodology!
- Don’t hesitate to look for **more data to corroborate** the meaning of your findings, for example for earlier data points. (Although it’s better to plan first what you need and then try to stick to the plan, or you get lost in stories)
- **Think like a journalist** (what stories are interesting, new, provocative…).
- Make sure you **compare apples with apples and oranges with oranges.**

Alice Corona

alice-corona.nl
@alice_corona

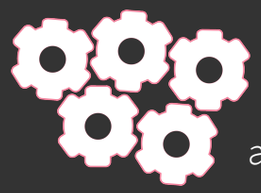
a data journalist's workflow

Data Storytelling

Alice Corona

alice-corona.nl
@alice_corona

JADS · Den Bosch, 18-6 August 2016



a data journalist's workflow

Journalistic storytelling

It's a very specific form of storytelling that involves the production and communication of newsworthy and **verified** information. As a journalist, you should abide to a very strict code of conduct (**deontology**) throughout all your work, including during the storytelling process.

Alice Corona

alice-corona.nl
@alice_corona

a data journalist's workflow

Define your audience and act accordingly

Medium

Components

Storytelling strategies

Alice Corona

alice-corona.nl
@alice_corona

a data journalist's workflow

Same data, different stories (1/2)

source: Martin Rosenbaum in "The Data Journalism Handbook"

Measurement

'Local councils across the country spent a total of £x billion on paper clips last year'

Proportion

'Last year local councils spent two-thirds of their stationery budget on paper clips'

Internal comparison

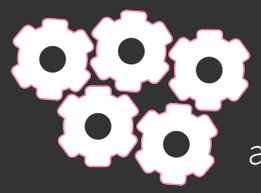
'Local councils spend more on paper clips than on providing meals-on-wheels for the elderly'

External comparison

'Council spending on paper clips last year was twice the nation's overseas aid budget'

Alice Corona

alice-corona.nl
@alice_corona



a data journalist's workflow

Same data, different stories (2/2)

Change over time

'Council spending on paper clips has trebled in the past four years'

"League Tables"

Borsetshire Council spends more on paper clips for each member of staff than any other local authority, at a rate four times the national average'

Analysis by categories

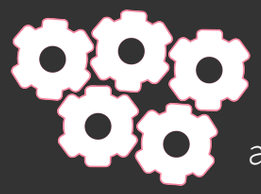
Councils run by the Purple Party spend 50% more on paper clips than those controlled by the Yellow Party'

Association

'Councils run by politicians who have received donations from stationery companies spend more on paper clips, with spending increasing on average by £100 for each pound donated'

Alice Corona

alice-corona.nl
@alice_corona



a data journalist's workflow

Find the humans behind the data

- Contextualise numbers
- Interview, make phone calls, quote
- Document more in depth specific stories found in the dataset with words, emotions, photos, videos, etc.
- Whatever else works for story!

Alice Corona

alice-corona.nl
@alice_corona

When narrating data

- Data visualisations take ambiguity away from the data. Use your storytelling to **add the nuances** behind averages and other statistics.
- **Process of finding the data can be the story itself.** A missing dataset, a dataset published in a specific point in time…all these could make for interesting stories.
- **Be transparent.** Publish datasets, methodology and all relevant files along with your story. Unless you have ethical/privacy concerns.
- **Become familiar with journalism's deontology.** You'll be a better journalist. And avoid legal troubles.

Alice Corona

alice-corona.nl
@alice_corona

a data journalist's workflow

Data Visualisation

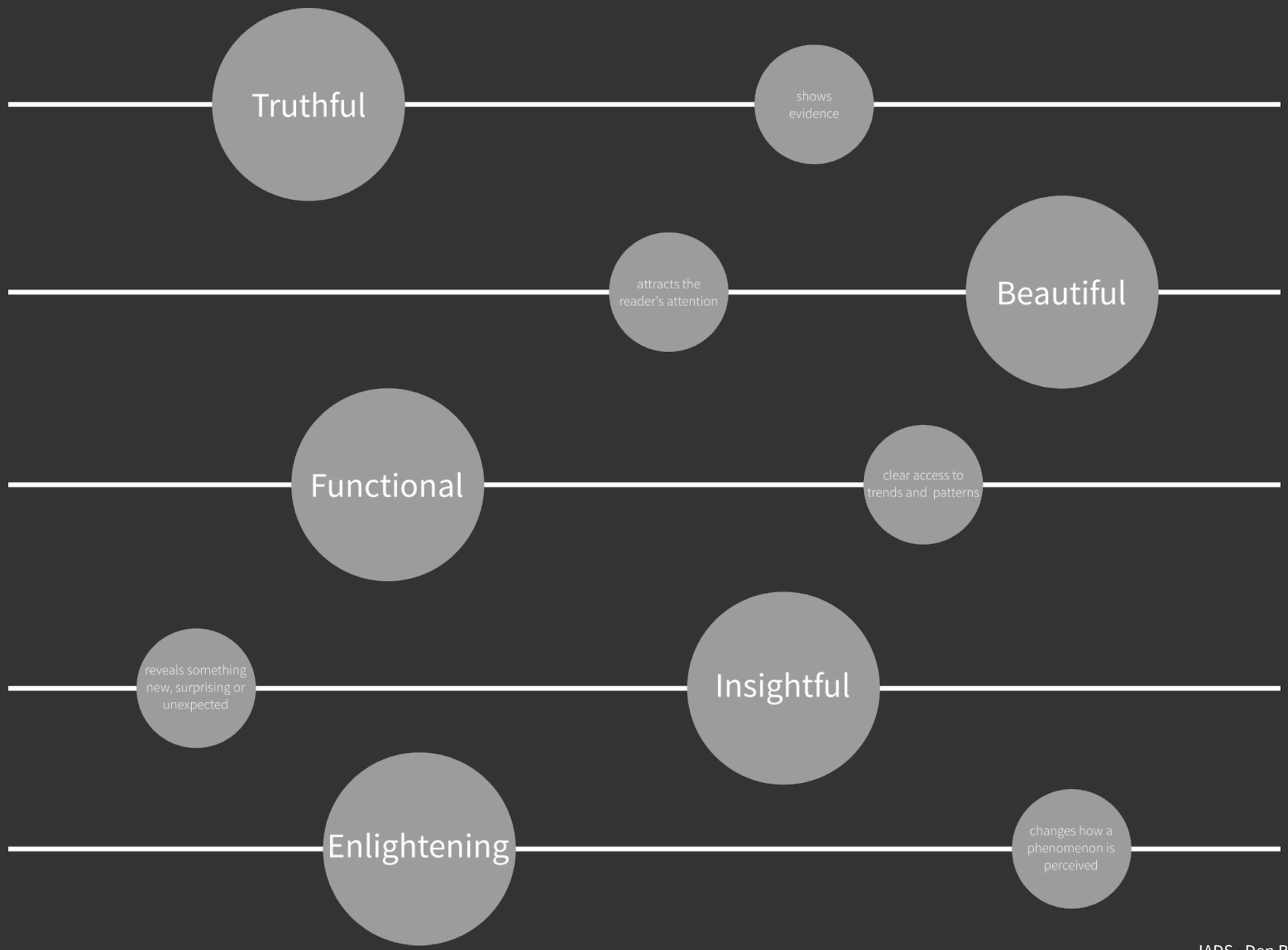
Alice Corona

alice-corona.nl
@alice_corona

JADS · Den Bosch, 18-6 August 2016

a data journalist's workflow

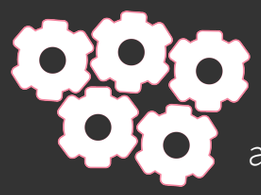
Alberto Cairo's guiding values for data visualisation



Alice Corona

alice-corona.nl
@alice_corona

JADS · Den Bosch, 18-6 August 2016



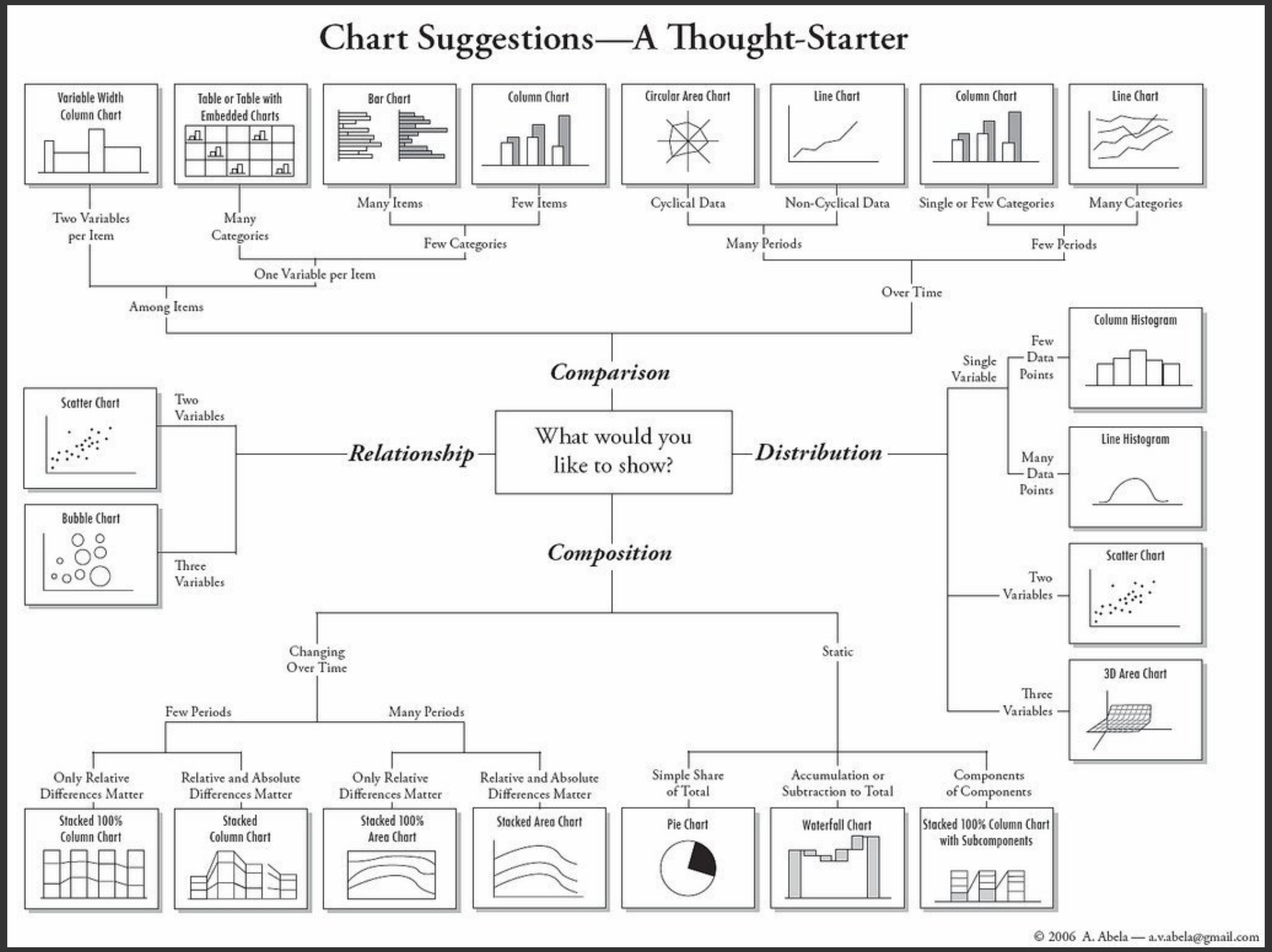
a data journalist's workflow

When choosing a specific graphic form, you should always consider your visualisation's **function** and choose accordingly.

Alice Corona

alice-corona.nl
@alice_corona

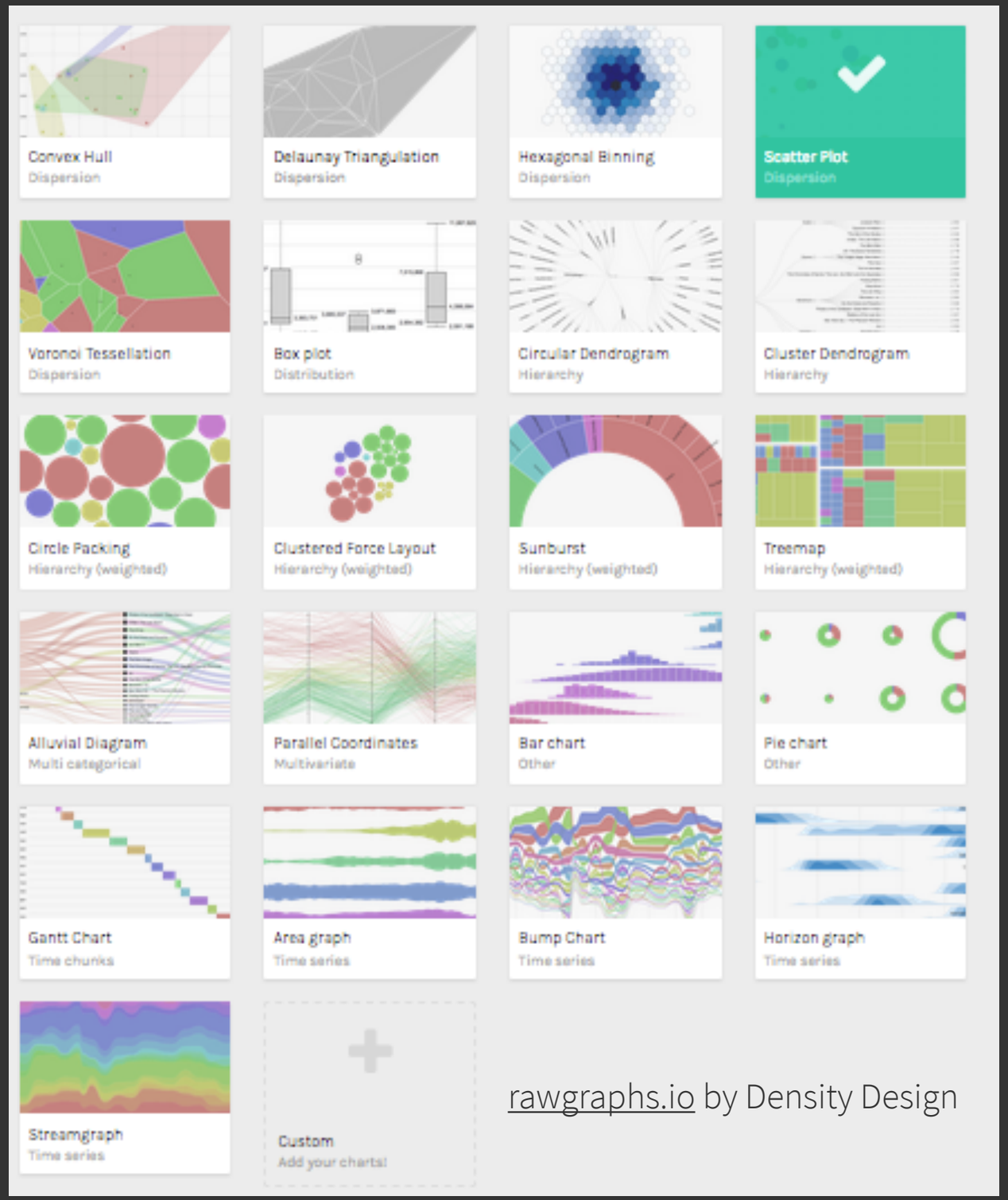
a data journalist's workflow



Alice Corona

alice-corona.nl
@alice_corona

a data journalist's workflow

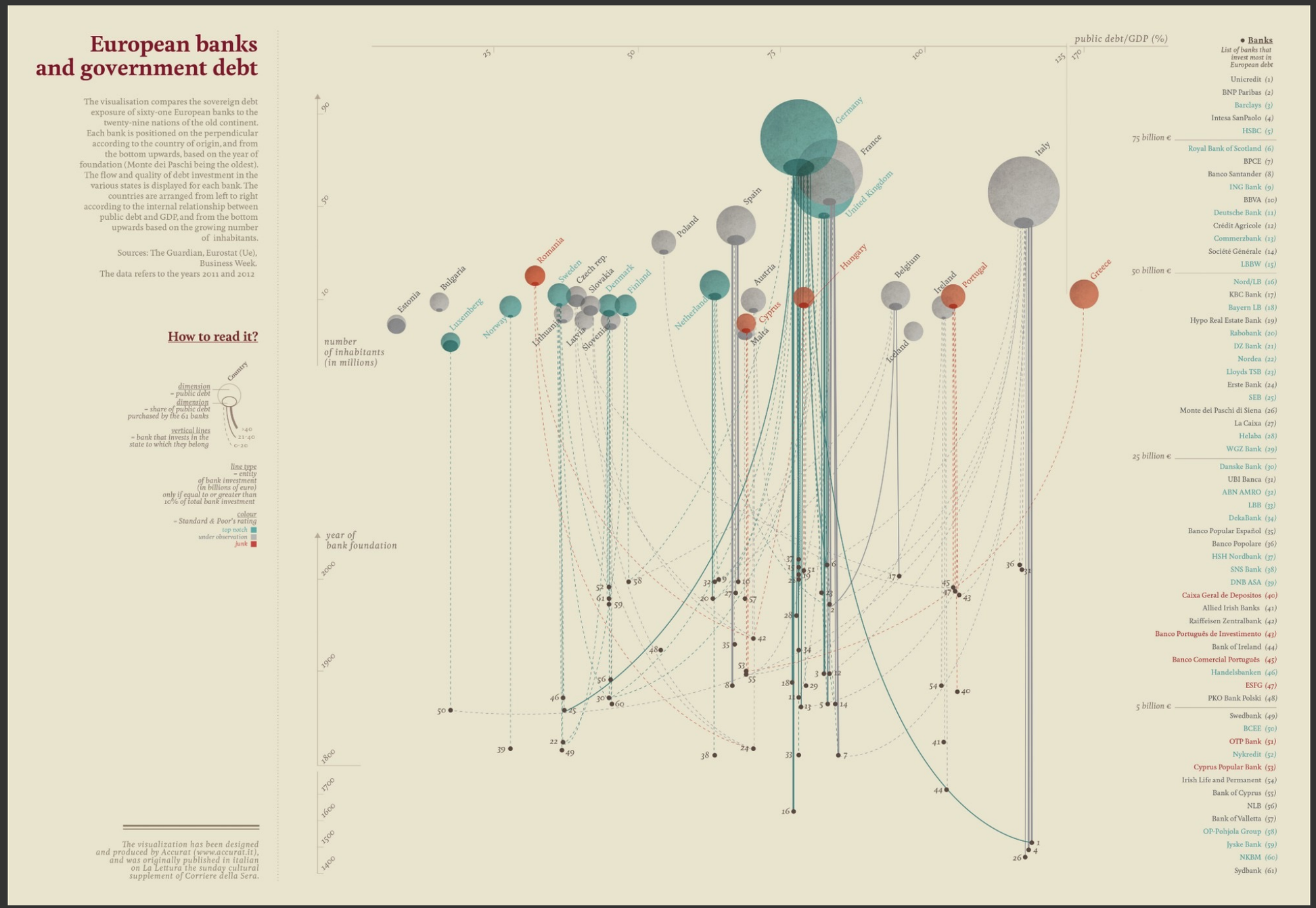


rawgraphs.io by Density Design

Alice Corona

alice-corona.nl
@alice_corona

a data journalist's workflow



European Banks and Government Debt, Accurat for La Lettura

Alice Corona

alice-corona.nl
@alice_corona

JADS · Den Bosch, 18-6 August 2016

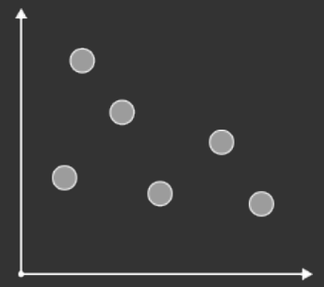
a data journalist's workflow

Visual Encodings

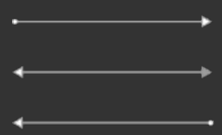
lenght

area

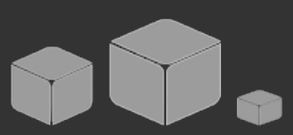
hue



position



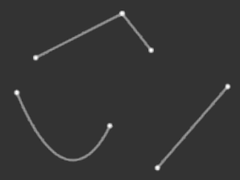
direction



volume

saturation

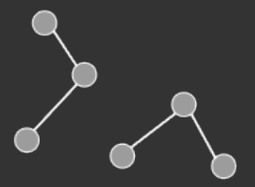
proximity



angle/curve
& slope

shape

texture



connections

Alice Corona

alice-corona.nl
@alice_corona

When visualising data

- **Do you need a data visualizaions?** Not every data journalism piece needs visuals. Ask yourself whether it will add value or clutter.
- **Show only what it makes sense to show.** You don't have to show everything just because you can. Think about what is not necessary and remove it. (data-ink ratio)
- **Graphics and interactivity doesn't exempt you from storytelling.** You can dump all the data in a visualisation and add filters for the user to explore the dataset, but try to make the main story easy to parse. Have a headline, introduction and everything needed for context.
- **Annotate, highlight and use hierarchies.** Add labels and make sure to give the right emphasis to what needs it.
- Your main focus is not to make art. Your visualisation is part and parcel of your journalistic work and thus it must be up to high **journalistic standards, values and deontology.**

Alice Corona

alice-corona.nl
@alice_corona

Collaborate!

Alice Corona

alice-corona.nl
@alice_corona

a data journalist's workflow

Alice Corona

alice-corona.nl
@alice_corona

The Fallen of World War II by Neil Halloran

JADS · Den Bosch, 18-6 August 2016

Alice Corona

Data-Driven Storyteller · Consultant · Trainer · Researcher

GET IN TOUCH

alice-corona.nl

alice-corona@riseup.net

FOLLOW

medium.com/@alice_corona

twitter.com/alice_corona

github.com/ali-ce

Credits for the icons: The Noun Project (zidney Evangeline White, Juan Pablo Bravo, Jems Mayor, Creative Stall, Bonegolem, Chris Homan,Gregor Cresnar, xicons.co, Dinosoft Labs, Artem Kovyazin, Ecem Afacan, Barracuda, Viktor Vorobyev, Wilson Joseph, Aenne Brielmann,Becris, Kirby Wu, Chameleon Design, Meaghan Hendricks, Christopher Holm-Hansen, Miguel C Balandrano, Jae Aquino, kokiri studio)

Alice Corona

alice-corona.nl
@alice_corona