Anastasia Valeeva

<https://www.facebook.com/anastasiajourno>

twitter @anastasiajourno

How to build a data hypothesis

14:00 - 15:15 guided discussion and brainstorming

15:30 - 16:45 data expedition: 1 idea -- many data stories

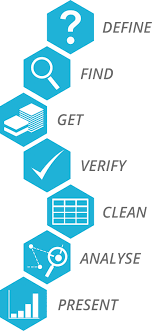
Are you already using data hypotheses (or something like this) and came here to learn more / get new ideas?

Are you new to data storytelling and want to explore what it is?

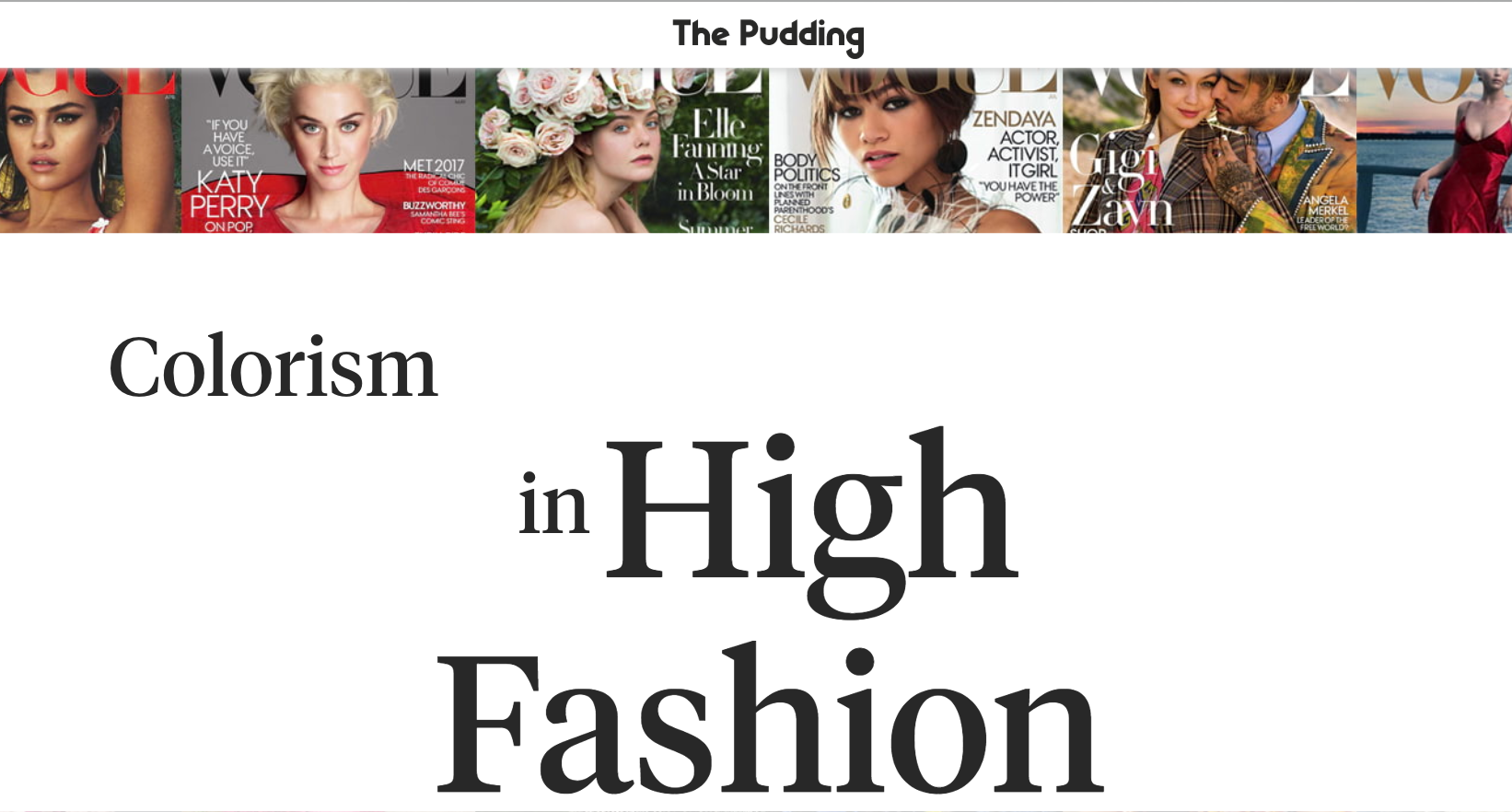
First, we start off by defining together what is a data hypothesis and what are its defining features.

They often say, a data story can start either from a question that you ask to the data, or with the data that you explore.

School of Data Pipeline:



Let’s read these stories and formulate a hypothesis in 1 sentence with no question mark at the end.



<https://pudding.cool/2019/04/vogue/>



<https://www.hanna-und-ismail.de/english/>

So what is defining a good / working hypothesis?

Based on the previous research / exploration

Can be verified with data

Specific about the data and indicators to use (operationalize)

It’s something newsworthy

A good hypothesis:

• Can either be proven or disproven with data

• The topic is important to the public

• Is specific about what is being measured

• The data is available

* 1 extra (later… hold on)

Benefits of a hypothesis:

• Helps in management and time planning

• Holds your story together

• Helps you to get rid of unnecessary information

• Is a tool that you can use to generate story ideas

Hypothesis is a guiding star for your story



Your turn:

*“Online space is unequal”*

* Take an idea and formulate a hypothesis in one sentence
* Don’t go looking for the data (yet) but describe it precisely
* Don’t forget about the ‘so what?’ question

**Women MPs receive more abuse than male MPs on social media (define a country; period; topic)**

**(it can be tested; is specific; the data is available)**

**find yourself on a spectrum between too specific and too broad**

**Not everyone in the world has (good) access to internet**

**- Distance to travel**

**- Bandwidth**

**- Costs**

**>>So what: Is it already a human right?**

**right wing parties have more fake accounts on social media and spread more fake news online than left wing parties**

**Female youtubers receive more gender based abused than male**

**Access to the Internet is more limited for citizens in poorer countries than for citizens in poorer countries (In other words, access to the Internet correlates with GDP)**

**Digital inequality is widening during pandemic in terms of education**

**Recent research data indicate that online space is unequal due to limitations one meets regarding their gender, race or social status despite the fact big tech companies tell us the opposite.**

**Access to the Internet is more limited for citizens in poorer countries than for citizens in richer countries (In other words, access to the Internet correlates with GDP)**

**Women are more harassed than men online.**

**Men post more on Twitter than women (trying to guess from the names, with Gender API, Genderize.io, et Gender Guesser.-)**

Weak hypothesis:

* One half or both halves of the hypothesis cannot be proved with data.
* The hypothesis is unspecific
* There is no data
* The topic is too broad. The topic is better for a book than a single story
* The hypothesis is too narrow: it only measures how one factor influences a trend and discounts other data sources that might also contribute to the trend.
* The hypothesis has already been proven true and is common knowledge.

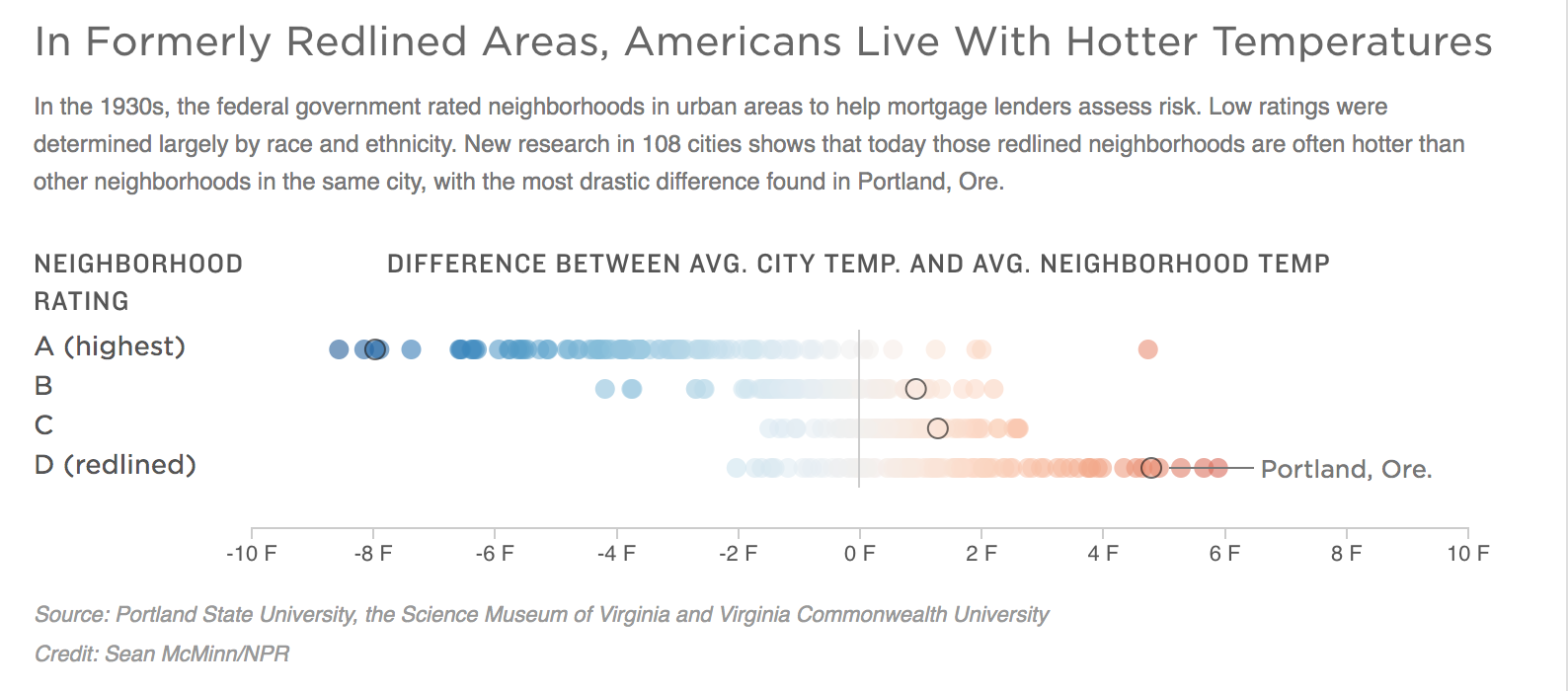
Good news -- you can make a weak hypothesis stronger by working on making it more specific and verifiable, looking for the alternative data if there is none, exploring the topic better to account for contributing factors but also focus better on your own story.

Suggest some ways to make this weak hypothesis better:

There are not enough kindergartens in our city.

Now let’s dive a little deeper.

Let’s listen to this 3 minute data story:



<https://www.npr.org/2020/01/14/795961381/racist-housing-practices-from-the-1930s-linked-to-hotter-neighborhoods-today>

Can you list the arguments?

Describing the problem:

* People of colour live in poorer neighbourhoods
* Poorer neighbourhoods are hotter than wealthy ones
* Wealthy neighbourhoods have more trees
* Poor neighbourhoods have more pavement

Factor/cause:

* In 1930 city planners “redlined” coloured neighbourhoods
* Still in 2020, formerly redlines neighbourhoods are hotter

Consequences/ Impact/ Victims/Who is benefiting:

* Heat has consequences for financial and human health
* The cooler neighbourhoods are now unaffordable for inhabitants of poor neighbourhoods who are poor and of color

Solution

Warning from Ben: Don't take a correlation or some occurrences as a cause for something else that 'seems' connected. Please dont just connect it because you think it does. It's of grave concern and many people keep doing it wrong.

Also a good hypothesis:

• Measures the problem, causes, impact and solutions

Problem

Impact

Cause

Solution

**Tools:**

1. Data questions for generating data hypotheses:

o Problem questions: how big is the problem? How expensive is it? Is the problem getting better or worse?

o Impact questions: Who is affected by the problem? How? Are some groups more affected than others?

o Cause questions: What are the causes of the problem? What factors have made it worse?

o Solution questions: What is the solution to the problem? How can we measure effectiveness? Has anybody tried to solve the problem and did she succeed?

Try this: The impact of pandemic on the domestic violence

2) Enriching stories with data:

One of the most important skills for a data journalist is to recognize opportunities to transform an ordinary story into a data story. Read more stories. 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 is unclear what the journalist is trying to prove. In others, there is an anecdote which can lead to a perfect data story.

<https://www.theguardian.com/world/2020/sep/16/holocaust-us-adults-study>

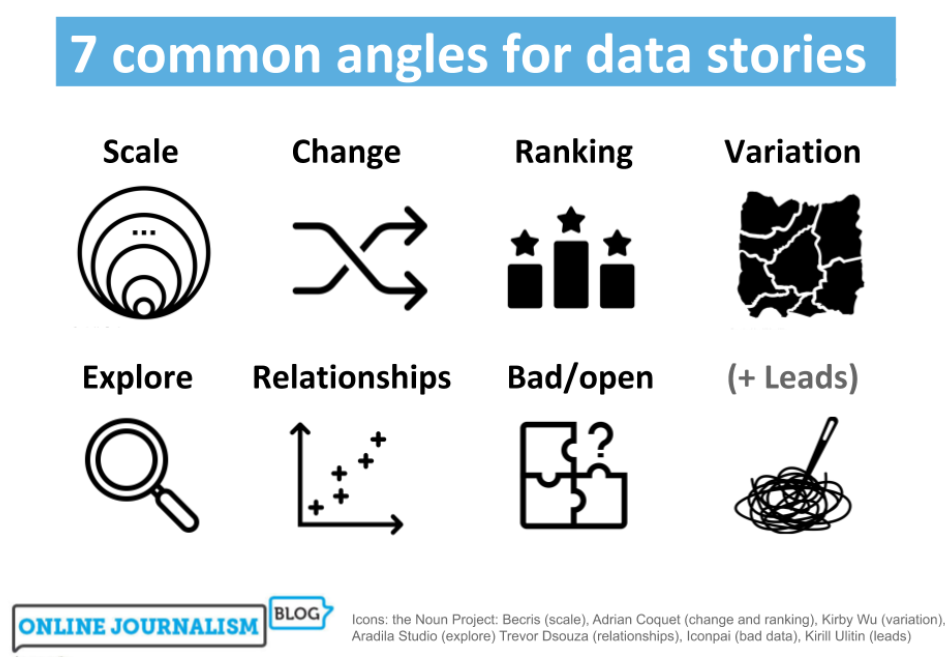
3) Turning on a data mindset

The shop on my street got closed due to the coronavirus --

<https://www.washingtonpost.com/opinions/2020/04/01/one-block-more-than-120-jobs-lost/?arc404=true>

Your turn:

I started ordering take out food more often --



<https://gijn.org/2020/08/18/from-relationships-to-ranking-angles-for-your-next-data-story/>

4) Generating ideas from a dataset:

<https://docs.google.com/spreadsheets/d/1Yj3TASk4ULTvG7Q07LiIpOyn2tfZJKBClbmx4Fe8vkI/edit#gid=0>

Data Matrix:

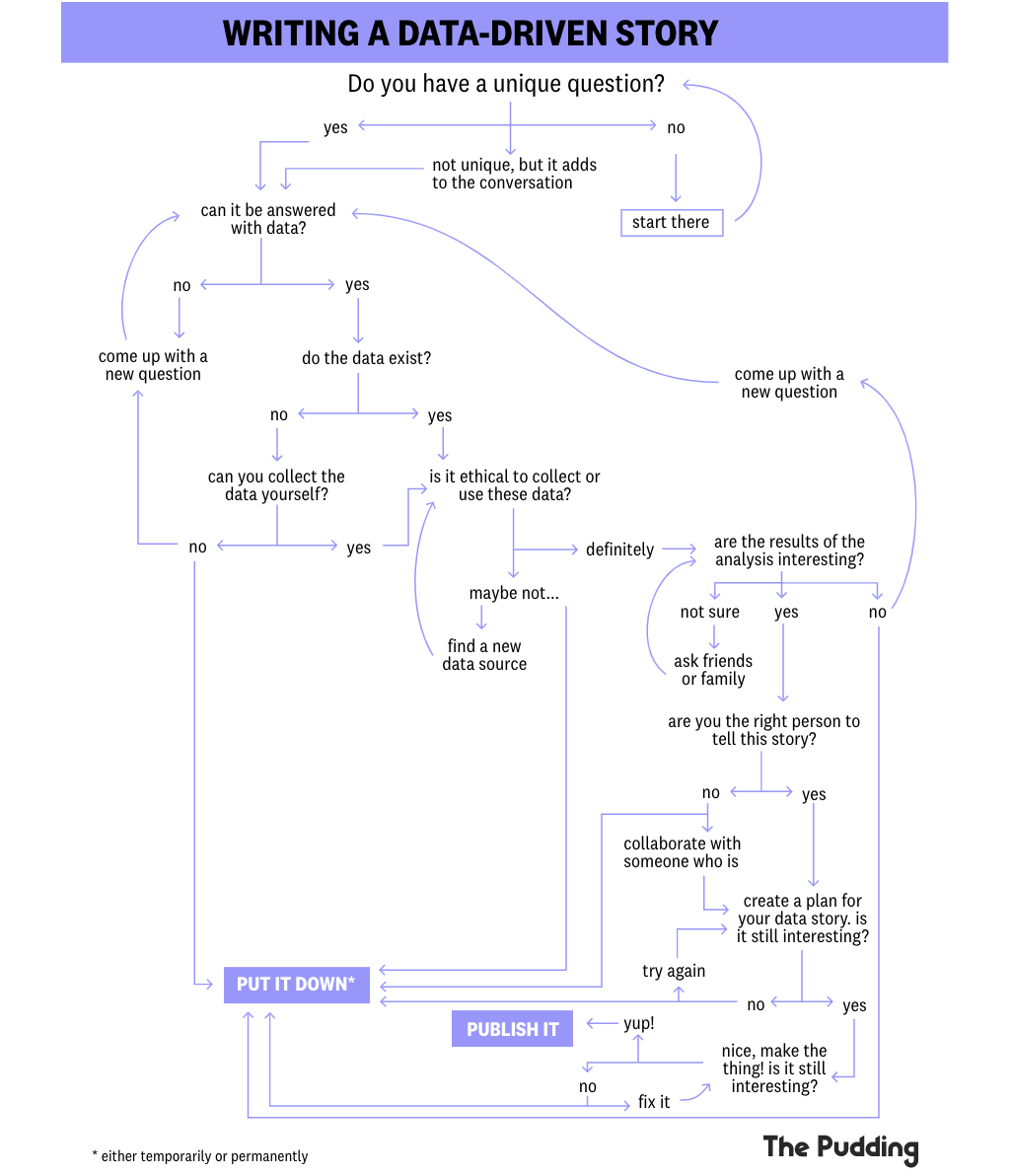
|  |  |  |  |
| --- | --- | --- | --- |
|  | reason | sex | year |
| reason |  |  |  |
| sex |  |  |  |
| year | + |  |  |

Generate ideas from a dataset with the following headlines:

vehicle\_id, car brand, car model, accident date, accident time, payment date, payment time, crossroad, accident type, fine amount, fine status (paid or not paid), paid amount.

5) Data Expedition

<https://schoolofdata.org/data-expeditions/guide-for-guides/?__cf_chl_jschl_tk__=2f973a71d506dcf93e4e6597e24d8318691c0a7f-1600262337-0-AVfMEPNUVFZq6tNSPjIuMpoe4ZeeVvEe_qCrDUJslKzadY3dQ3LHogV0wyJx7TGnTtCXHGCVlI9lOTyx-H4GojfrVmsmNssC7a3SnnHSCJK5DXX7fWOL3HM8t271XTE77uJ__cjmBC2DO-QUtrQCTB10AbXUXrRKcU7wyaeTFvI6JAZFz0kaXLsoCI61VI8epJ93Y8yE20i0ONkkl-QUa14RKkmg77bApltfhL9ubIAR_yZ0bJQZwsl36aW-H2rUwaB_heJBqHCuw2gmq0IM6v2xRbFgY1QWa77FLDS0n9LIn6n3Ts92TQQ5MDa6Rkpy6A>



<https://pudding.cool/process/pivot-continue-down/>

How to get from idea to execution:

<https://www.youtube.com/watch?v=aTOaONyON3o&feature=youtu.be>

**Resources:**

Eva Constantaras’ Manual

<https://en.arij.net/materials/arij-hosts-data-journalism-training-manual-in-arabic/>

Recent MOOC on data journalism:

<https://live-journalismcourses.pantheonsite.io/course/data-journalism-and-visualization-with-free-tools/>

Collection of data stories:

<https://datajournalismawards.org/projects/>