# Data Science Storytelling and Narrative

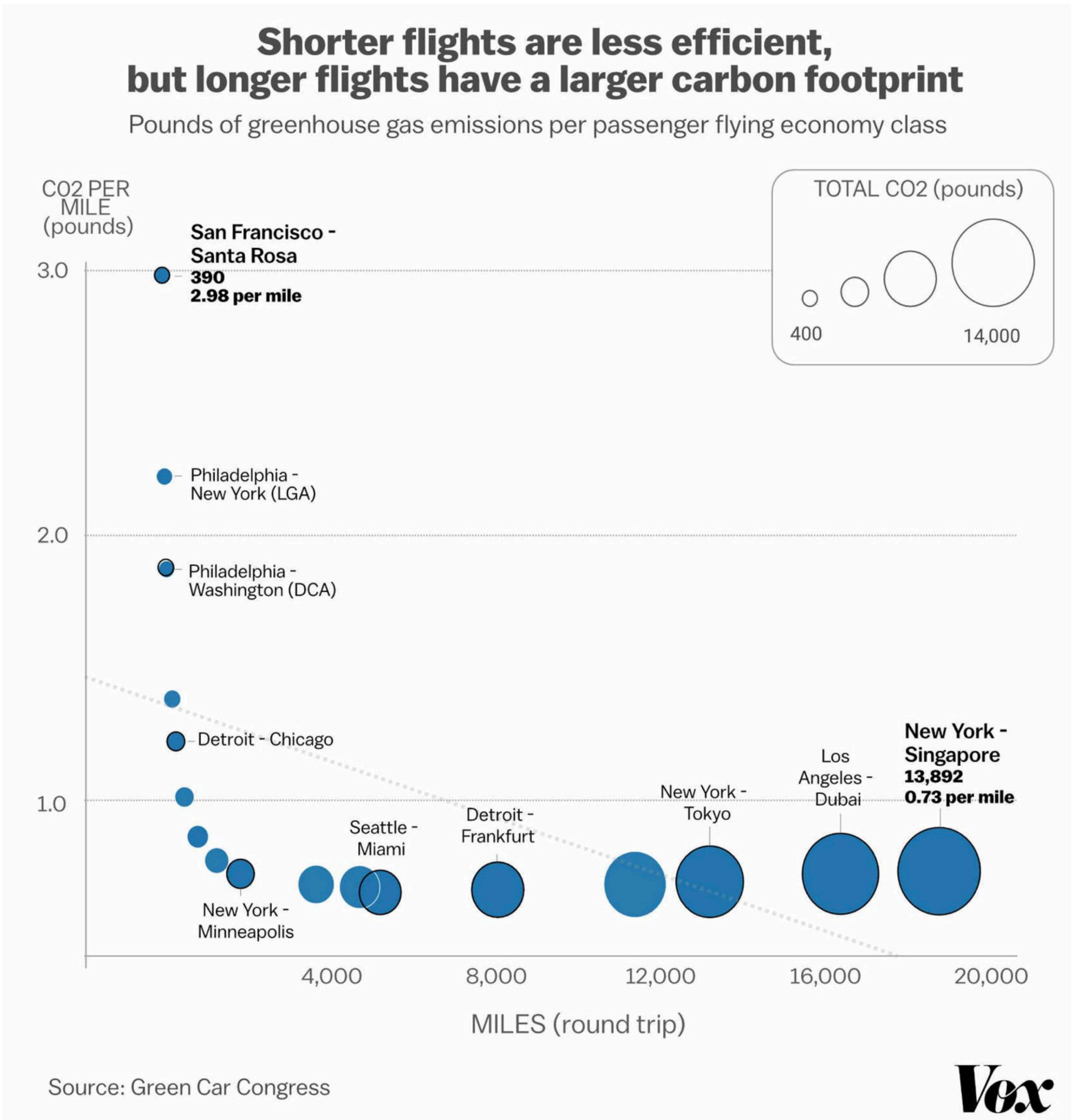
Roger D. Peng
Stephanie C. Hicks

Advanced Data Science
Term 1
2019

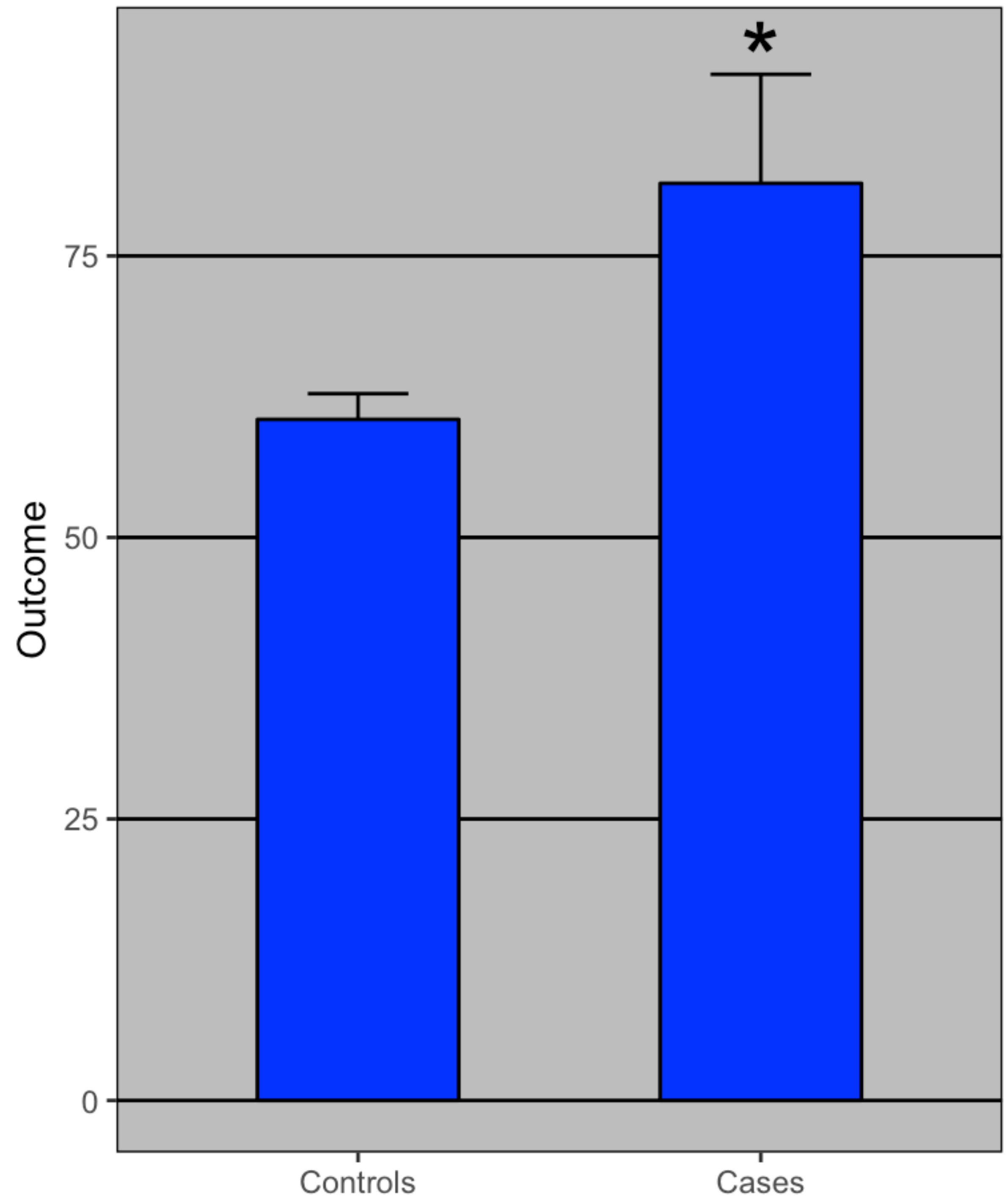
# Follow Up: Evaluating Plots

- What comparison am I being asked to make?
- Is the plot helping me to make that comparison?
- How well can I evaluate the strength of the evidence?

# What Comparison?



# Dynamite Plots Must Die




https://simplystatistics.org/2019/02/21/dynamite-plots-must-die/

# Just the Facts?

# Data Storytelling

- The story / narrative communicates a central dramatic argument based on evidence and data
- Dimension reduction for analytic results
- Often we disagree on the story but agree on the evidence
- You are negotiating with your audience to get them to accept your central argument

# Data Storytelling

- Central dramatic argument / theme
- Thematic structure
- Story causality
- Format
- Presentation
- Trust

# Central Dramatic Argument

American Journal of Respiratory and Critical Care Medicine

Home > All AJRCCM Issues > **Vol. 197, No. 6 | Mar 15, 2018**

## Long-Term Coarse Particulate Matter Exposure Is Associated with Asthma among Children in Medicaid

**Corinne A. Keet** [1], **Joshua P. Keller** [2], and **Roger D. Peng** [2]

+ Author Affiliations

https://doi.org/10.1164/rccm.201706-1267OC    **PubMed:** 29243937

Received: June 28, 2017    Accepted: November 21, 2017

Comments

| Abstract | Full Text | References | Supplements | Cited by | PDF | Related |
|---|---|---|---|---|---|---|

https://www.ncbi.nlm.nih.gov/pubmed/29243937

# Central Dramatic Argument

FiveThirtyEight

**Politics** Sports Science & Health Economics Culture

JUN. 21, 2018, AT 2:38 PM

## Spying Doesn’t Pay — Unless You’re Really Good At It

By Jeff Asher

Filed under Espionage

https://fivethirtyeight.com/features/spying-doesnt-pay-unless-youre-really-good-at-it/

# Central Dramatic Argument?

## Statistical Models for Earthquake Occurrences and Residual Analysis for Point Processes

YOSIHIKO OGATA*

This article discusses several classes of stochastic models for the origin times and magnitudes of earthquakes. The models are compared for a Japanese data set for the years 1885–1980 using likelihood methods. For the best model, a change of time scale is made to investigate the deviation of the data from the model. Conventional graphical methods associated with stationary Poisson processes can be used with the transformed time scale. For point processes, effective use of such *residual analysis* makes it possible to find features of the data set that are not captured in the model. Based on such analyses, the utility of seismic quiescence for the prediction of a major earthquake is investigated.

KEY WORDS: Akaike information criterion; Epidemic-type models; Conditional intensity; Likelihood; Marked point process; Seismic quiescence; Trigger models.

**"For point processes, effective use of such residual analysis makes it possible to find features of the data set that are not captured in the model."**

http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1988.10478560

# Central Dramatic Argument

**"The purpose of the story is to take a character from ignorance of the truth of the theme to embodiment of theme through action."**

**-Craig Mazin, *Scriptnotes Podcast*, Ep. 403**

# Thematic Structure

**Everything is fine (a state of ignorance, stasis)**

**We have a problem and we *must* do something about it**

**But...maybe we can ignore the problem and continue on....**

**A moment of doubt (have I been living a lie??)**

**Because we act according to the theme, we prevail**

**Stasis achieved, everything is fine, but different**

# Story Causality

# Story Causality

**https://youtu.be/vGUNqq3jVLg?t=47**

# Story Causality

**NOAA dataset analysis 1**

**NOAA dataset analysis 2**

# Most Poisoned Baby Name

## Not So Standard Deviations

A statistics (etc.) blog by Hilary Parker


Posted on **January 30, 2013**

**← Previous** **Next →**

### Hilary: the most poisoned baby name in US history

I've always had a special fondness for my name, which — according to Ryan Gosling in "Lars and the Real Girl" — is a scientific fact for most people (Ryan Gosling constitutes scientific proof in my book). Plus, the root word for Hilary is the Latin word "hilarius" meaning cheerful and merry, which is the same root word for "hilarious" and "exhilarating." It's a great name.

Several years ago I came across this blog post, which provides a cursory analysis for why "Hillary" is the most poisoned name of all time. The author is careful not to comment on the details of why "Hillary" may have been poisoned right around 1992, but I'll go ahead and make the bold causal conclusion that it's because that was the year that Bill Clinton was elected, and thus the year Hillary Clinton entered the public sphere and was generally reviled for not wanting to bake cookies or something like that. Note that this all happened when I was 7 years old, so I spent the formative years of 7-15 being called "Hillary Clinton" whenever I introduced myself. Luckily, I was a feisty feminist from a young age and rejoiced in the comparison (and life is not about being popular).

https://hilaryparker.com/2013/01/30/hilary-the-most-poisoned-baby-name-in-us-history/

# Central Dramatic Argument

**Defining "poisoning" as the relative loss of popularity in a single year and controlling for fad names, "Hilary" is absolutely the most poisoned woman's name in recorded history in the US.**

**https://hilaryparker.com/2013/01/30/hilary-the-most-poisoned-baby-name-in-us-history/**

# Everything is Fine

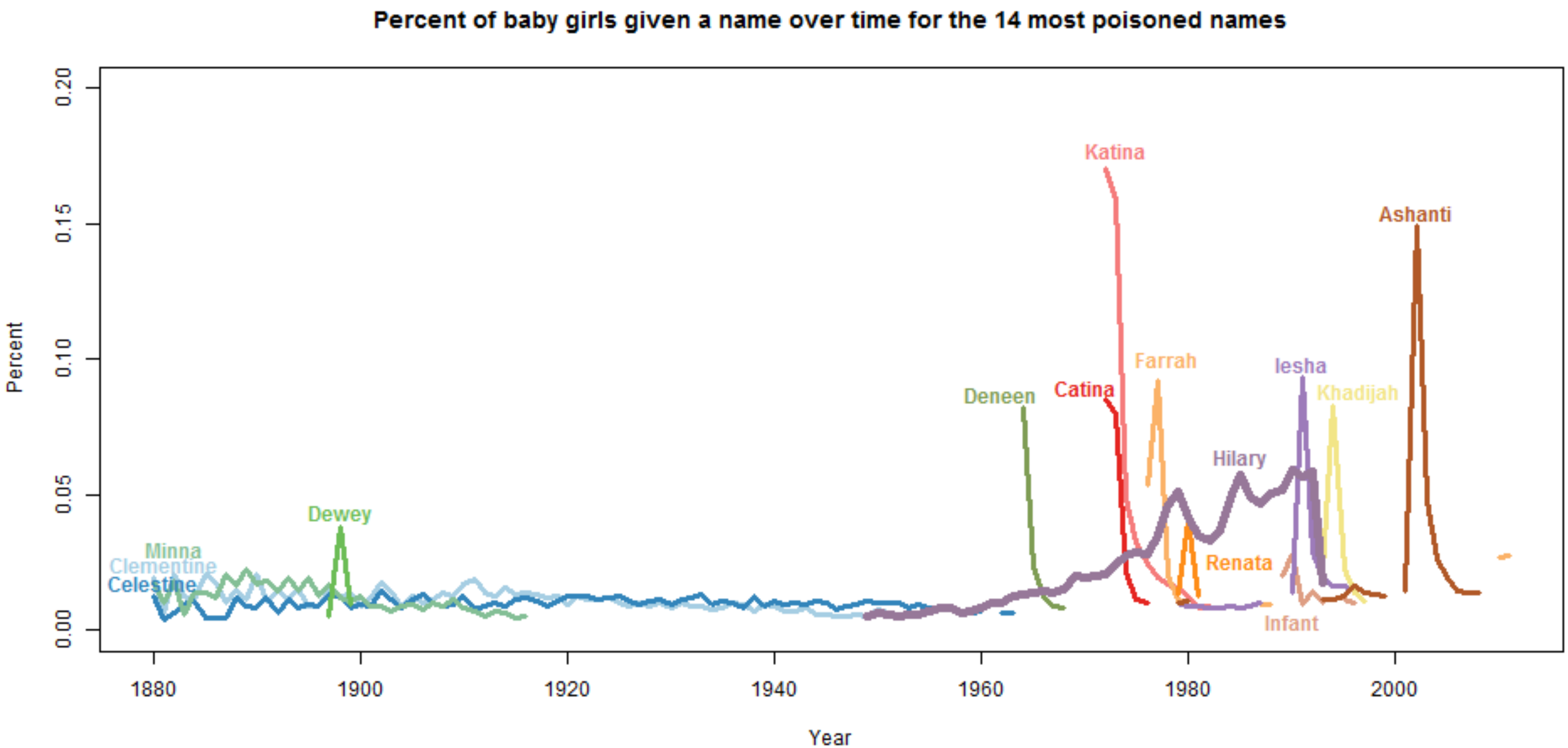
**Is Hilary/Hillary really the most rapidly poisoned name in recorded American history? An analysis.**

I will follow up this post with more details on how to perform web-scraping with R (for this I am infinitely indebted to my friend Mark — check out his storyboard project and be amazed!). For now, suffice it to say that I was able to collect from the social security website the data for every year between 1880 and 2011 for the 1000 most popular baby names. For each of the 1000 names in a given year, I collected the raw number of babies given that name, as well as the percentage of babies given that name, and the rank of that name. For girls, this resulted in 4110 total names.

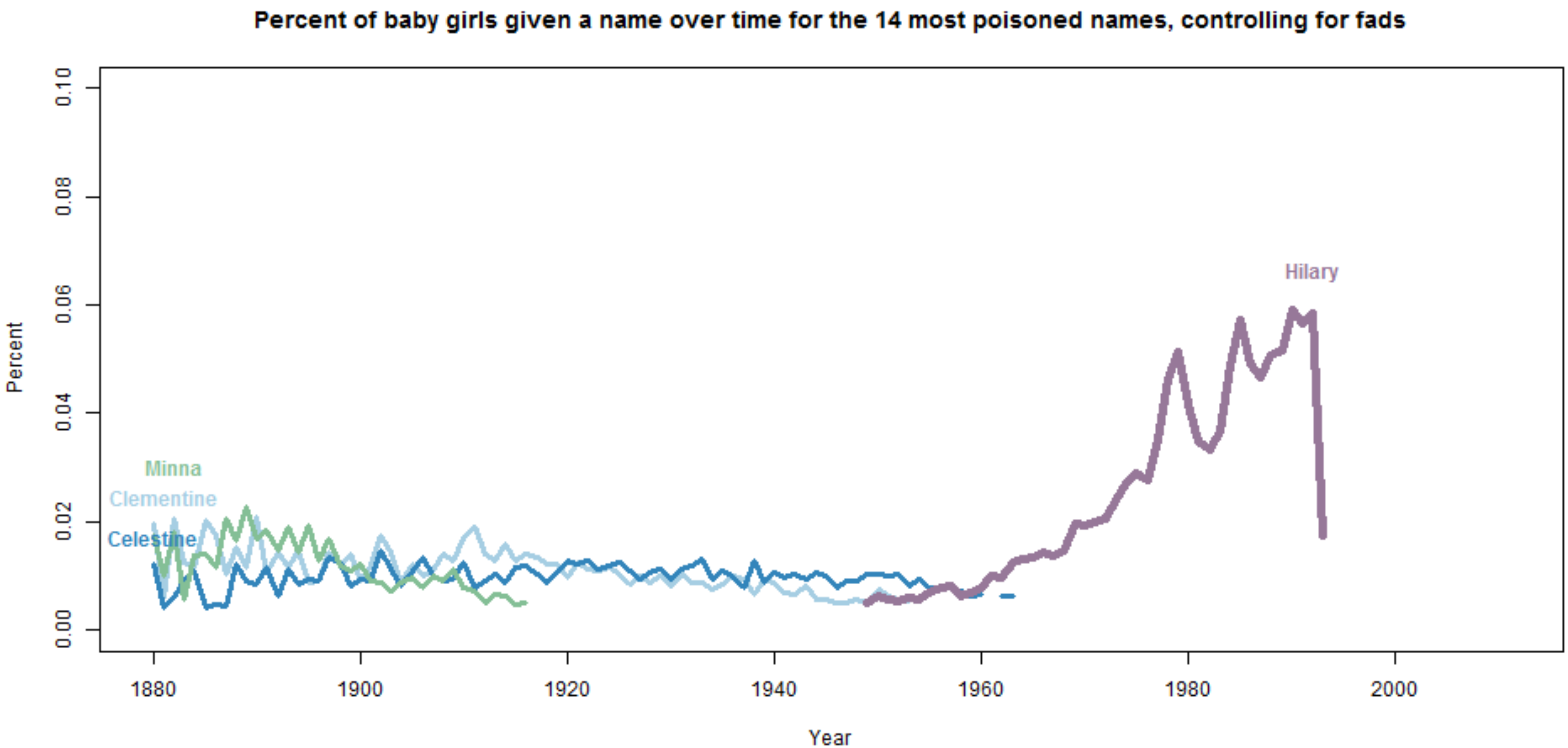
# See??

| Name | Loss (%) | Year |
|---|---|---|
| Farrah | 78 | 1978 |
| Dewey | 74 | 1899 |
| Catina | 74 | 1974 |
| Deneen | 72 | 1965 |
| Khadijah | 72 | 1995 |
| Hilary | 70 | 1993 |
| Clementine | 69 | 1881 |
| Katina | 69 | 1974 |
| Renata | 69 | 1981 |
| Iesha | 69 | 1992 |
| Minna | 68 | 1883 |
| Ashanti | 68 | 2003 |
| Celestine | 67 | 1881 |
| Infant | 67 | 1991 |

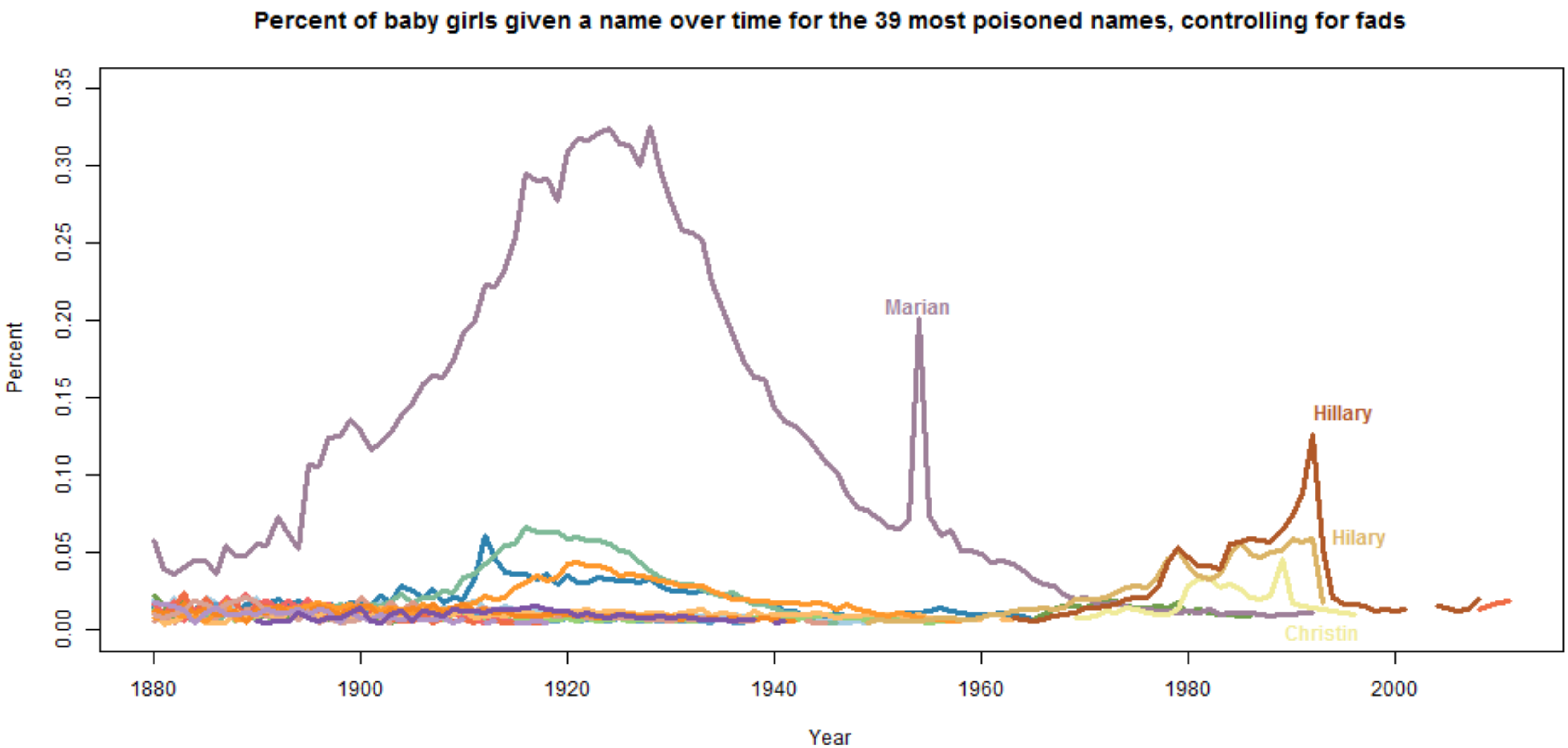
# A Moment of Doubt



# Reframing the Question



# Compared to What...?




**I also did a parallel analysis for boys, and aside from fluctuations in the late 1890s/early 1900s, the only name that comes close to this rate of poisoning is Nakia, which became popular because of a short-lived TV show in the 1970s.**

# Format

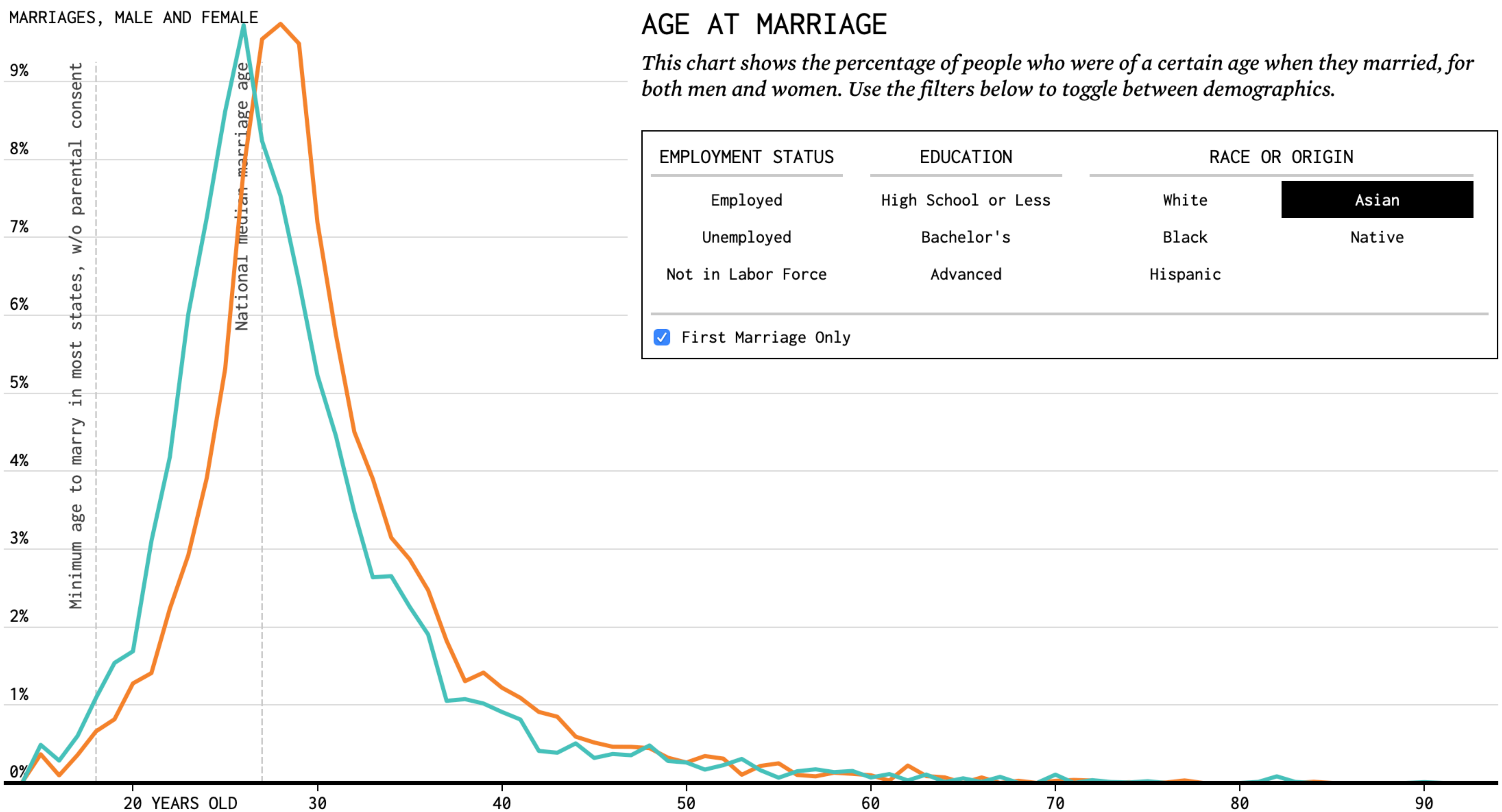
- The format of the presentation should match nature of the story being told
- Blog post
- Report
- Paper
- **Email** - Intention / obstacle, expectation / deviation
- **Interactive presentation**

# Email

- Short format, concise, ideally 3 sentences
- I have this intention and/or have this expectation
- I am facing the following obstacle or deviation from my expectation
- [Ask a yes/no question] or indicate negative control [I will do something unless you say otherwise]

**https://simplystatistics.org/2011/09/23/getting-email-responses-from-busy-people/**

# Is It Your Time?

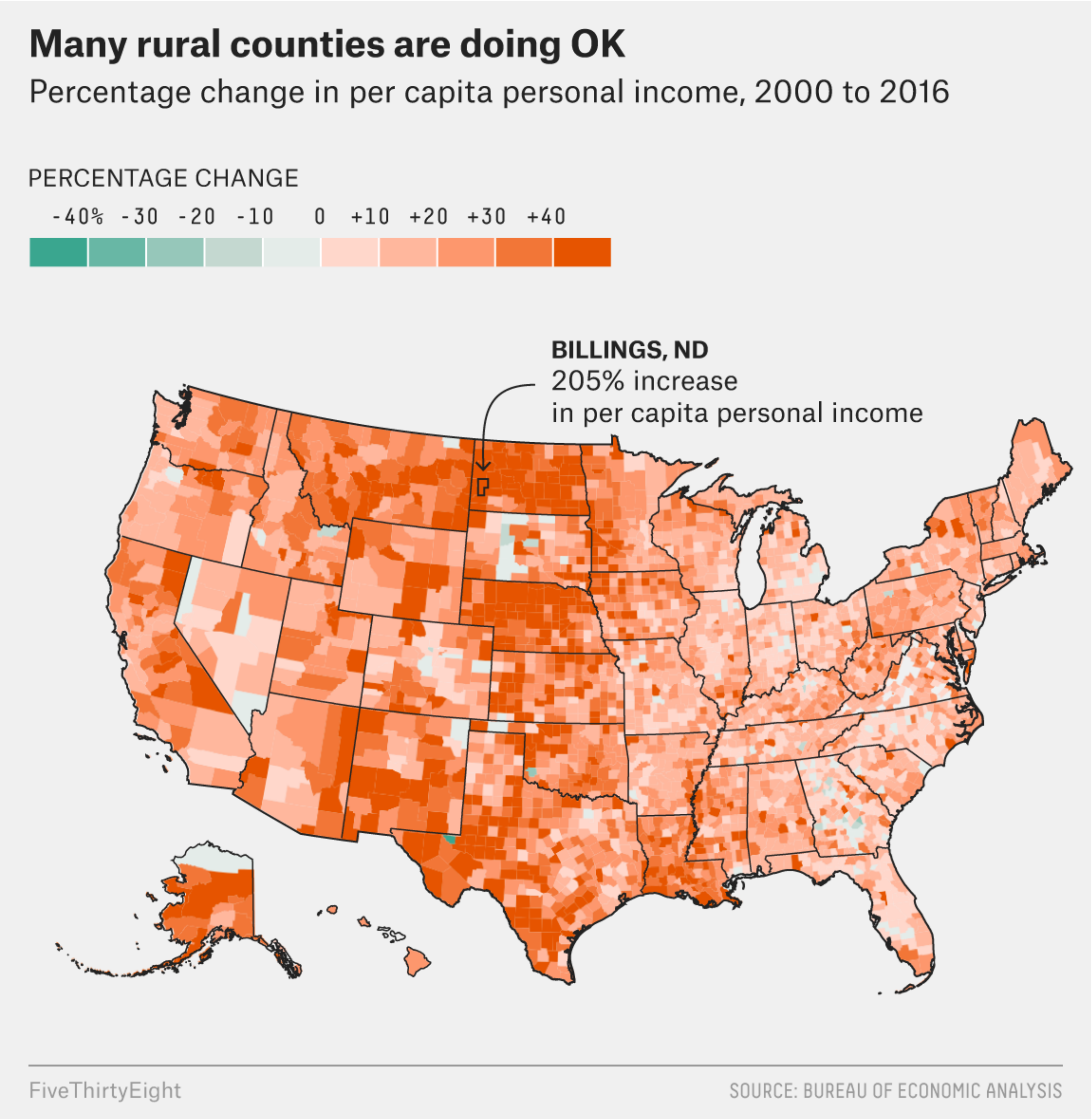



https://flowingdata.com/2016/03/03/marrying-age/

# Presentation

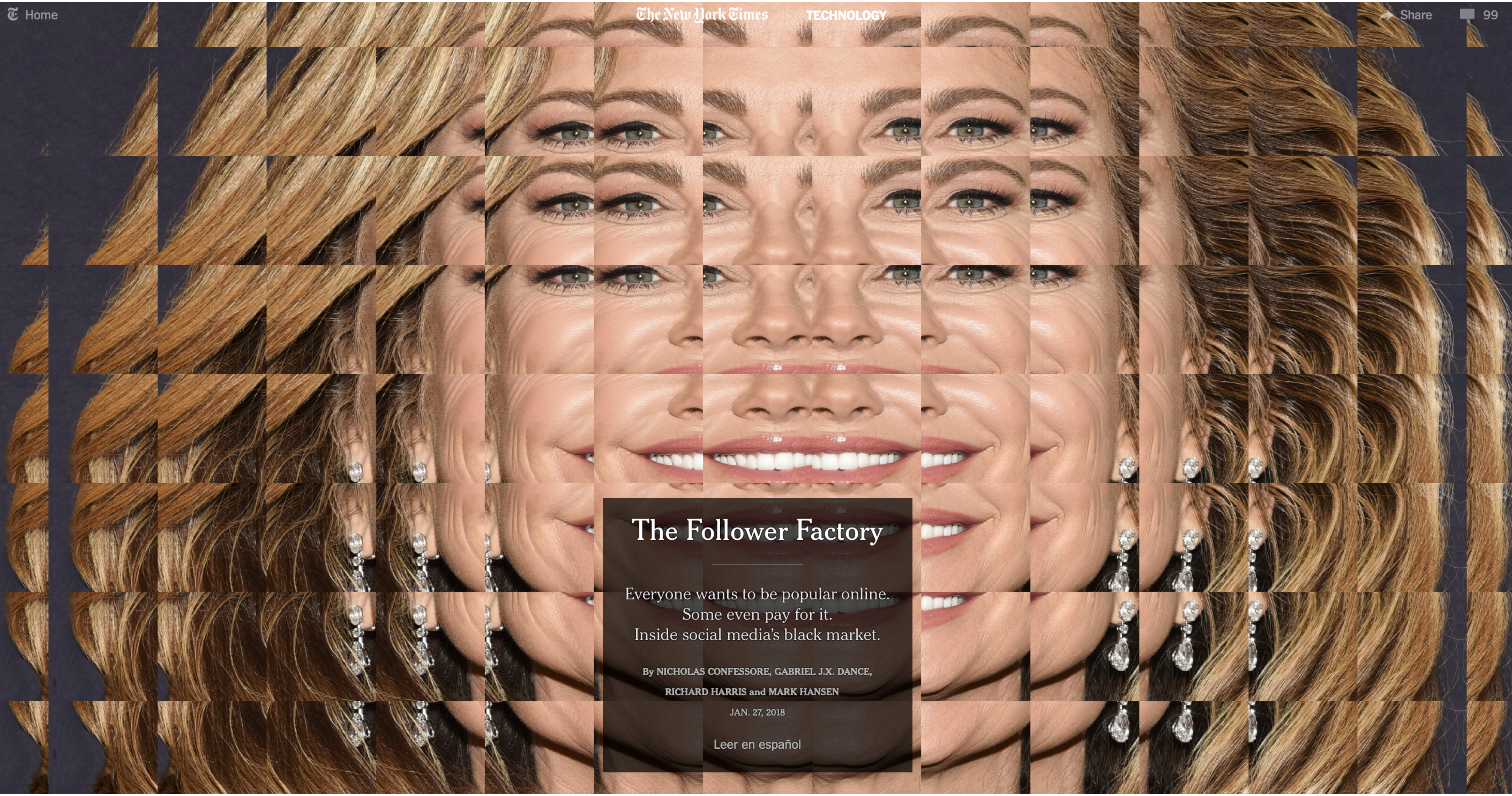
- Pacing - how much time can you afford with this audience?
- Context - what is the audience's background? What do they already know? What don't they know?
- Elements
    - What aspects of the analysis should be included that support the story?
    - Trust-building elements - do I need additional details to get this audience to trust me?
- Abstraction / Detail spectrum

# Where Blue-Collar America Is Strongest

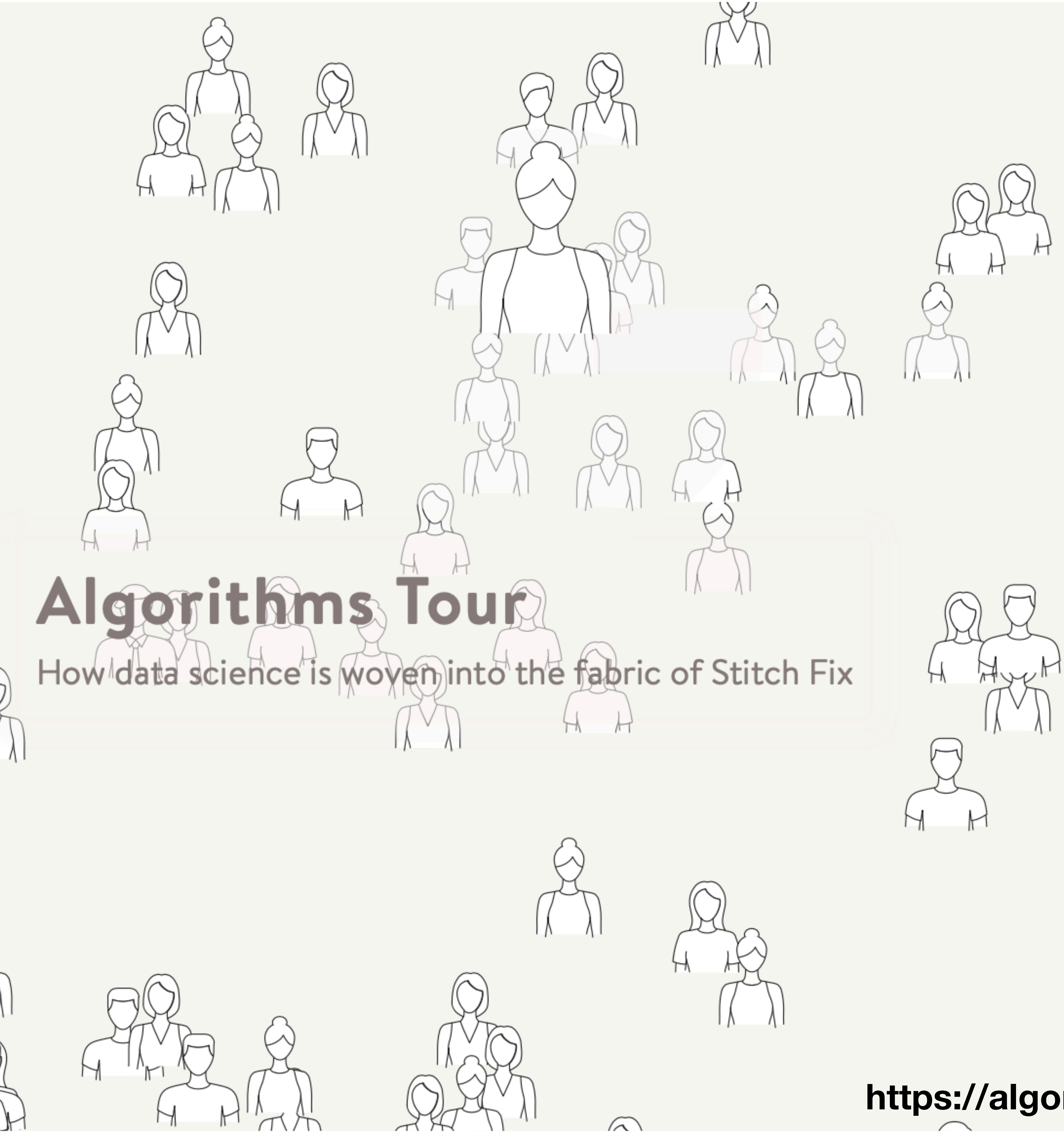



https://fivethirtyeight.com/features/has-trump-made-the-working-class-great-again-not-if-youre-black-or-female/

# The Follower Factory

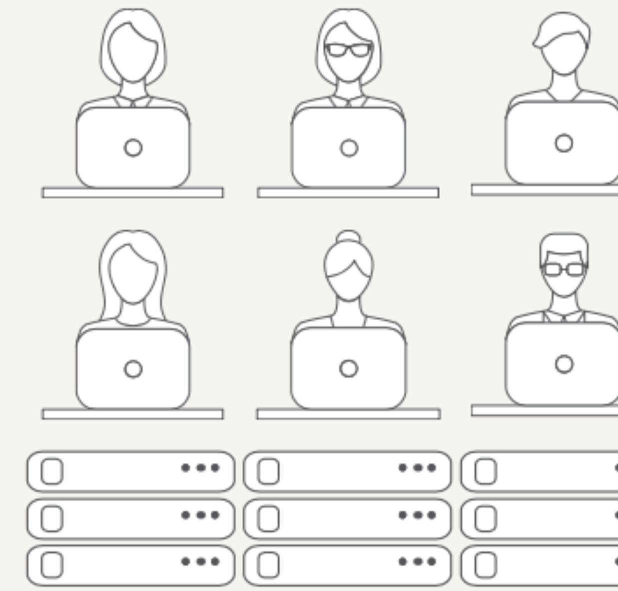



https://www.nytimes.com/interactive/2018/01/27/technology/social-media-bots.html



# Algorithms Tour

How data science is woven into the fabric of Stitch Fix



$$\log \frac{p}{1-p} = a + X\beta + Zb$$

...

$$\min_{a} \sum_{i} \sum_{j} a_{ij} q_{ij}$$

$$s.t. \quad a_{ij} \in \{0,1\}, \ \forall i,j$$

$$\sum_{j} a_{ij} = 1 \ \forall i$$

$$\sum_{i} a_{ij} < k_j \ \forall j$$

...

$$\frac{\partial x}{\partial t} = f(x_t, u_t, w_t)$$

...

$$p(i \rightarrow j)$$

$$= logit(\beta_0 + \beta_1 x \dots)$$

https://algorithms-tour.stitchfix.com

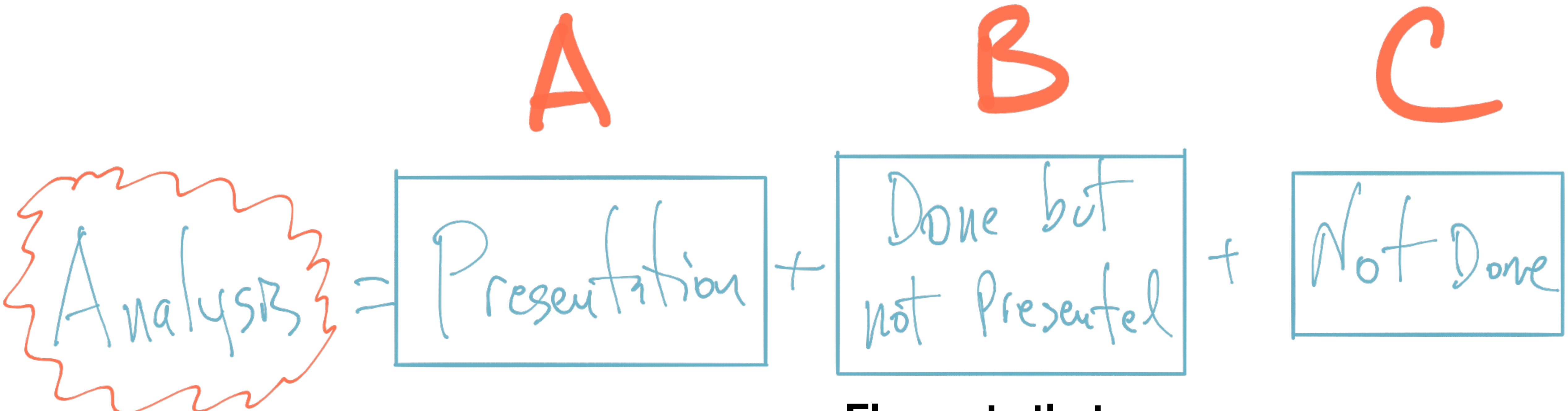
# Building Trust

- An analysis/presentation is will be heavily discounted if the audience does not trust you

- They story you tell is part of a larger negotiation with the audience to accept the analysis

- Your story may need to be altered if the audience does not know you

# Trustworthy Analysis




**Elements that support the story**

**Elements that are relevant but do not change the story**

**Orthogonal elements**

https://simplystatistics.org/2018/06/04/trustworthy-data-analysis/

# Trusting vs. Believing

- Trust
    - I accept the analysis, the data were analyzed properly and thoroughly
    - Trust is particular to the **analysis** and the **person** doing the analysis
- Believing
    - I believe the conclusion / central argument, is true
    - Depends on context, previous work, factors outside the analysis

# Other Factors

- Meme filter:
  - Can my data be misinterpreted?
  - What if your data were broken down into bite-sized chunks?
- News filter:
  - What’s the pulse of what’s going on in the news?
- How might this impact the people involved in your work?

**Vivian Peng, "Ethics of Data Storytelling" - https://youtu.be/CgYDsDBQAwU**

# Summary

- Story is the means by which you deliver your central dramatic argument

- How the story is told depends on your relationship with the audience and the audience's background

- Data stories can be told in many different formats but the basic technique is the same