When and How Data Journalists Disclose their Methods

#nicarmethods

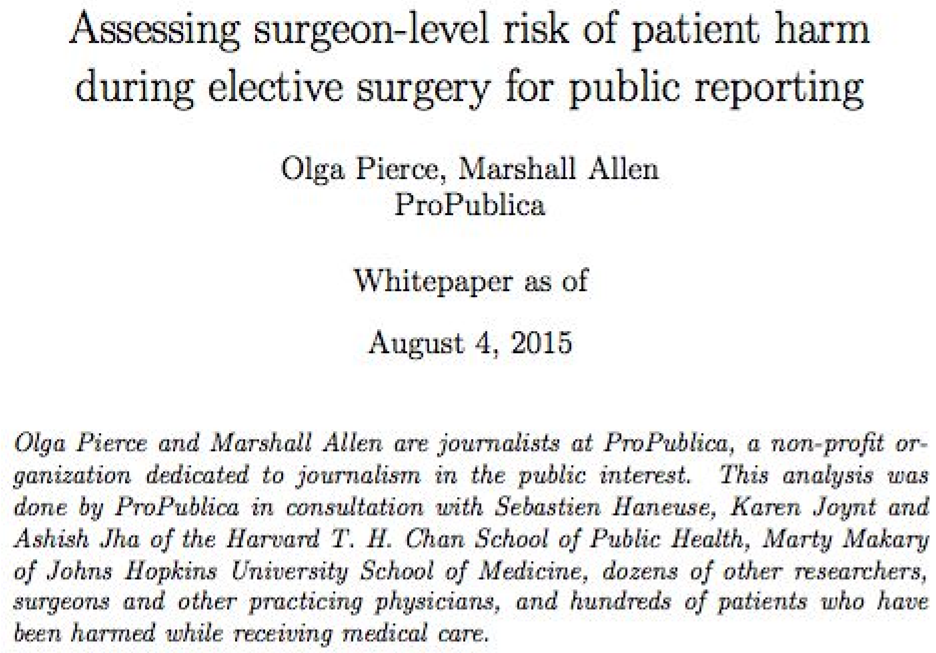
Chris Groskopf, Quartz
@onyxfish

Ryann Jones, ProPublica
@ryanngro

Stuart Thompson, WSJ
@stuartathompson

Aaron Williams, Washington Post
@aboutaaron

Examples of when we do



Assessing surgeon-level risk of patient harm during elective surgery for public reporting

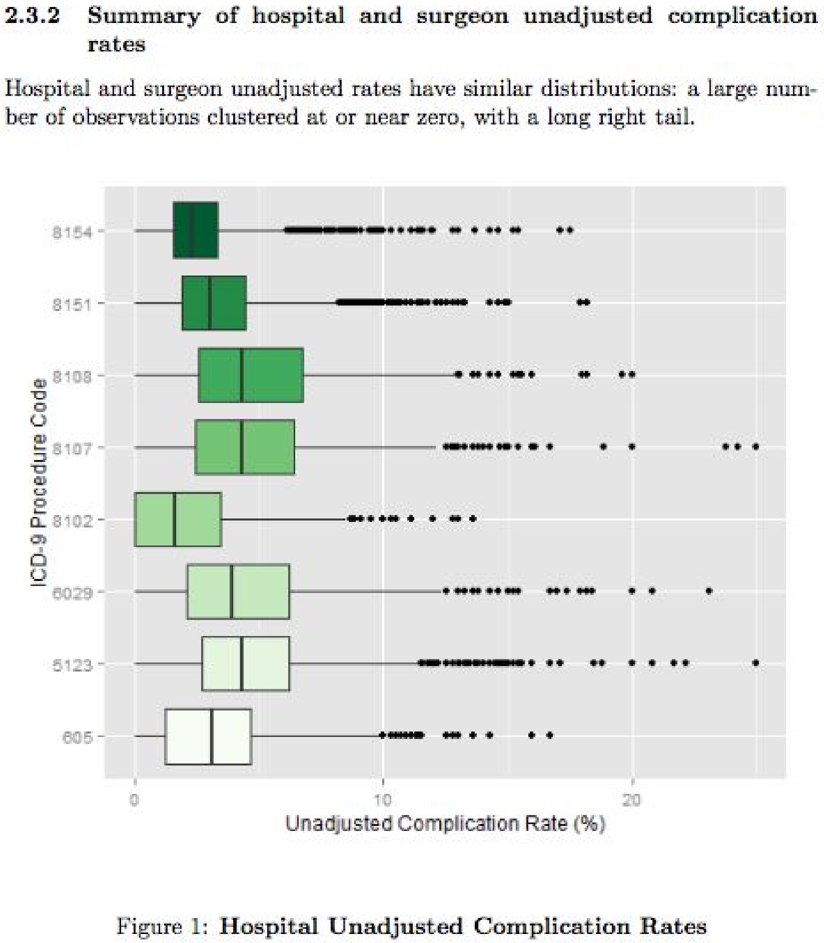
Olga Pierce, Marshall Allen
ProPublica

Whitepaper as of

August 4, 2015

Olga Pierce and Marshall Allen are journalists at ProPublica, a non-profit organization dedicated to journalism in the public interest. This analysis was done by ProPublica in consultation with Sebastien Haneuse, Karen Joynt and Ashish Jha of the Harvard T. H. Chan School of Public Health, Marty Makary of Johns Hopkins University School of Medicine, dozens of other researchers, surgeons and other practicing physicians, and hundreds of patients who have been harmed while receiving medical care.

https://static.propublica.org/projects/patient-safety/methodology/surgeon-level-risk-methodology.pdf



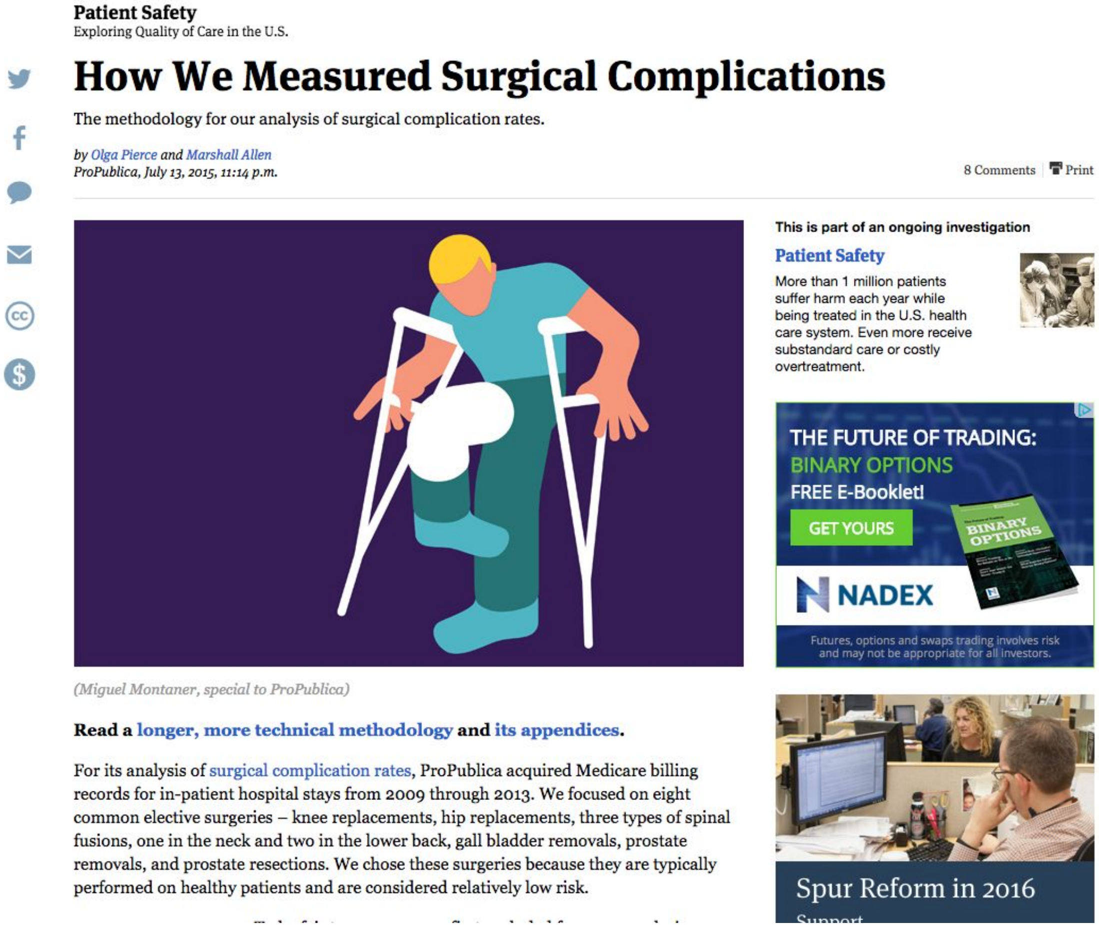
2.3.2 Summary of hospital and surgeon unadjusted complication rates

Hospital and surgeon unadjusted rates have similar distributions: a large number of observations clustered at or near zero, with a long right tail.

ICD-9 Procedure Code
8154
8151
8108
8107
8102
6029
5123
605
0
10
20
Unadjusted Complication Rate (%)

Figure 1: **Hospital Unadjusted Complication Rates**

#nicarmethods



Patient Safety
Exploring Quality of Care in the U.S.

How We Measured Surgical Complications

The methodology for our analysis of surgical complication rates.

by Olga Pierce and Marshall Allen
ProPublica, July 13, 2015, 11:14 p.m.

8 Comments | Print

This is part of an ongoing investigation

Patient Safety

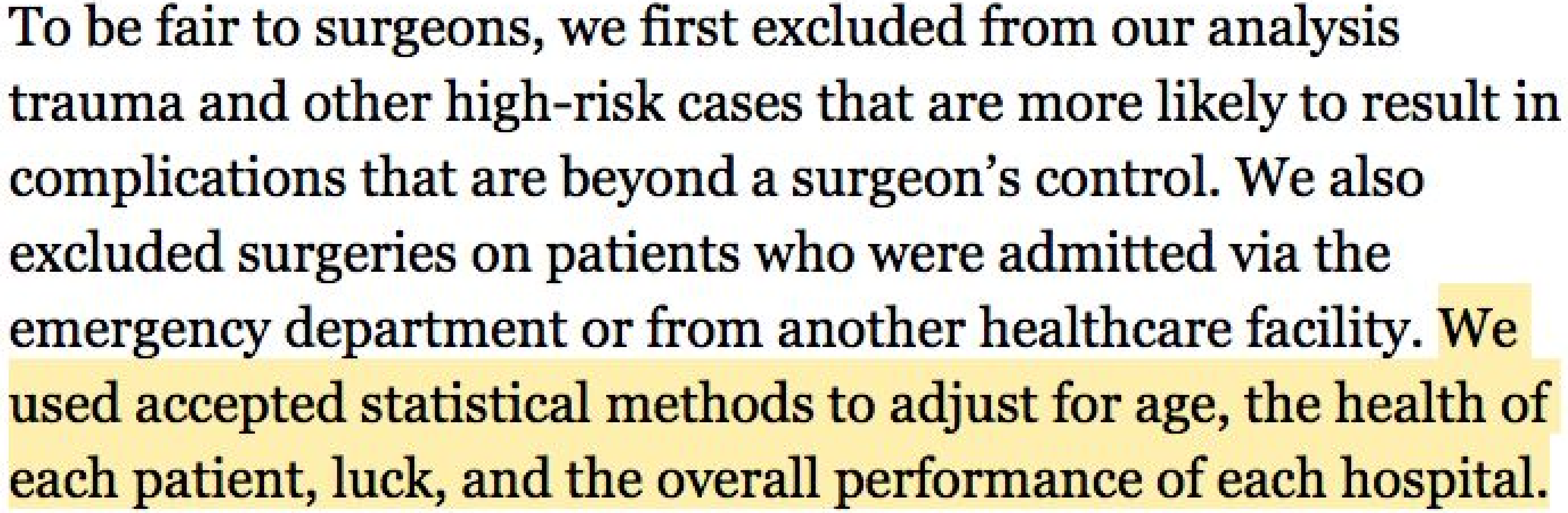
More than 1 million patients suffer harm each year while being treated in the U.S. health care system. Even more receive substandard care or costly overtreatment.

(Miguel Montaner, special to ProPublica)

Read a longer, more technical methodology and its appendices.

For its analysis of surgical complication rates, ProPublica acquired Medicare billing records for in-patient hospital stays from 2009 through 2013. We focused on eight common elective surgeries – knee replacements, hip replacements, three types of spinal fusions, one in the neck and two in the lower back, gall bladder removals, prostate removals, and prostate resections. We chose these surgeries because they are typically performed on healthy patients and are considered relatively low risk.

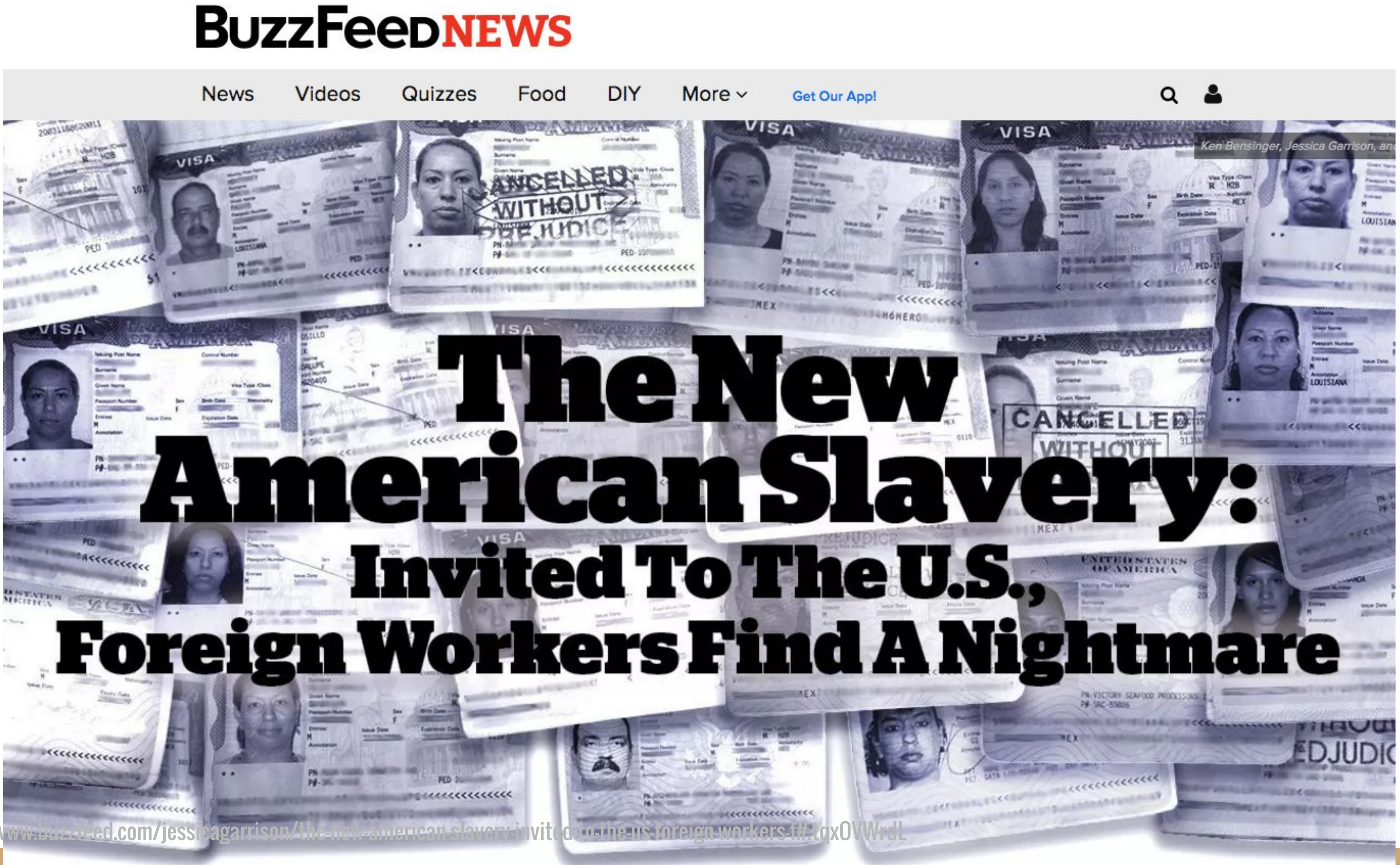
https://www.propublica.org/article/surgeon-level-risk-short-methodology



To be fair to surgeons, we first excluded from our analysis trauma and other high-risk cases that are more likely to result in complications that are beyond a surgeon’s control. We also excluded surgeries on patients who were admitted via the emergency department or from another healthcare facility. We used accepted statistical methods to adjust for age, the health of each patient, luck, and the overall performance of each hospital.

Should we all do this with complex analysis?

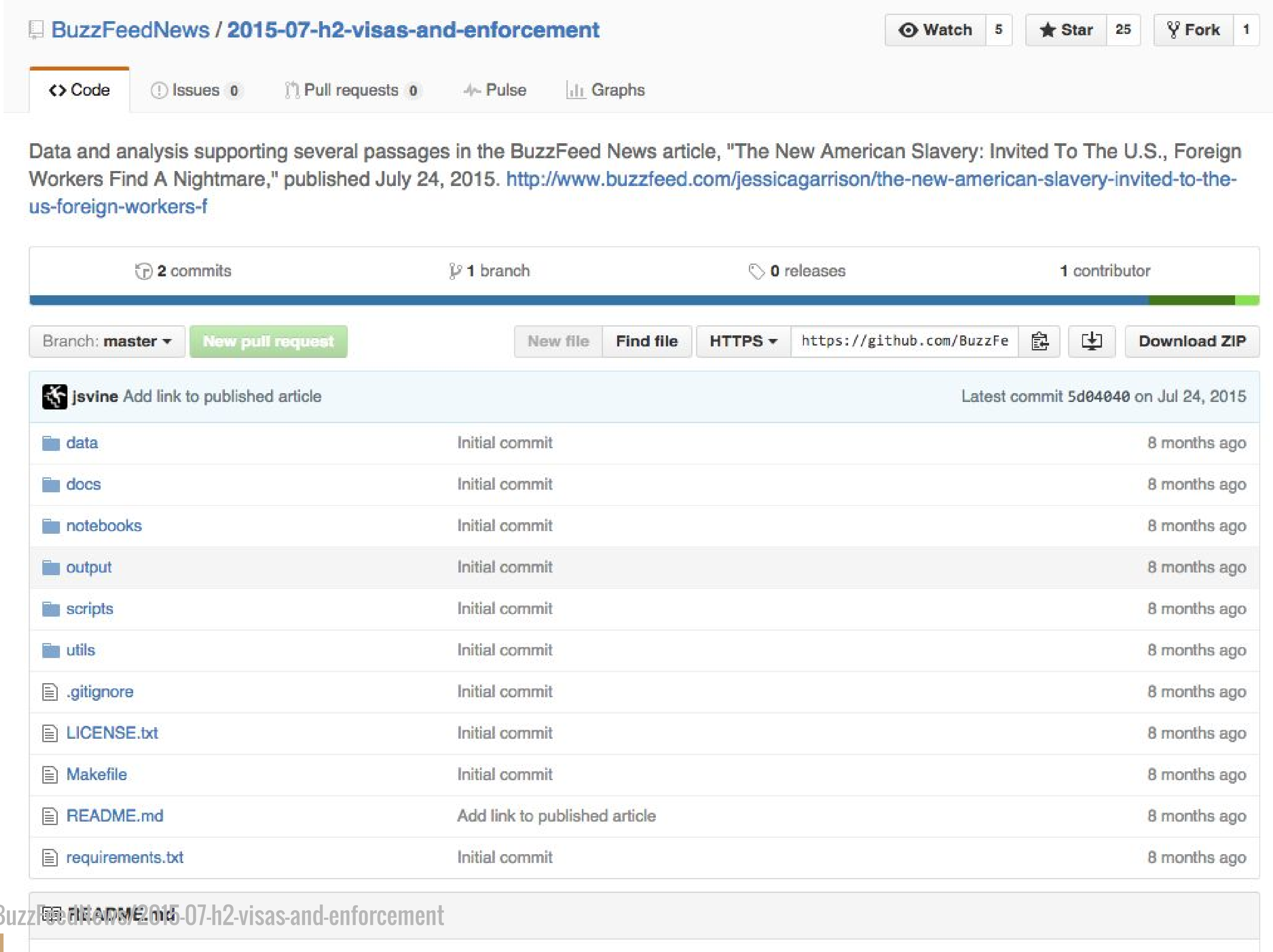
Do we have time?



The New American Slavery: Invited To The U.S., Foreign Workers Find A Nightmare

Ken Bensinger, Jessica Garrison, and

http://www.buzzfeed.com/jessicagarrison/the-new-american-slavery-invited-to-the-us-foreign-workers-f#.tqxOVWrdL

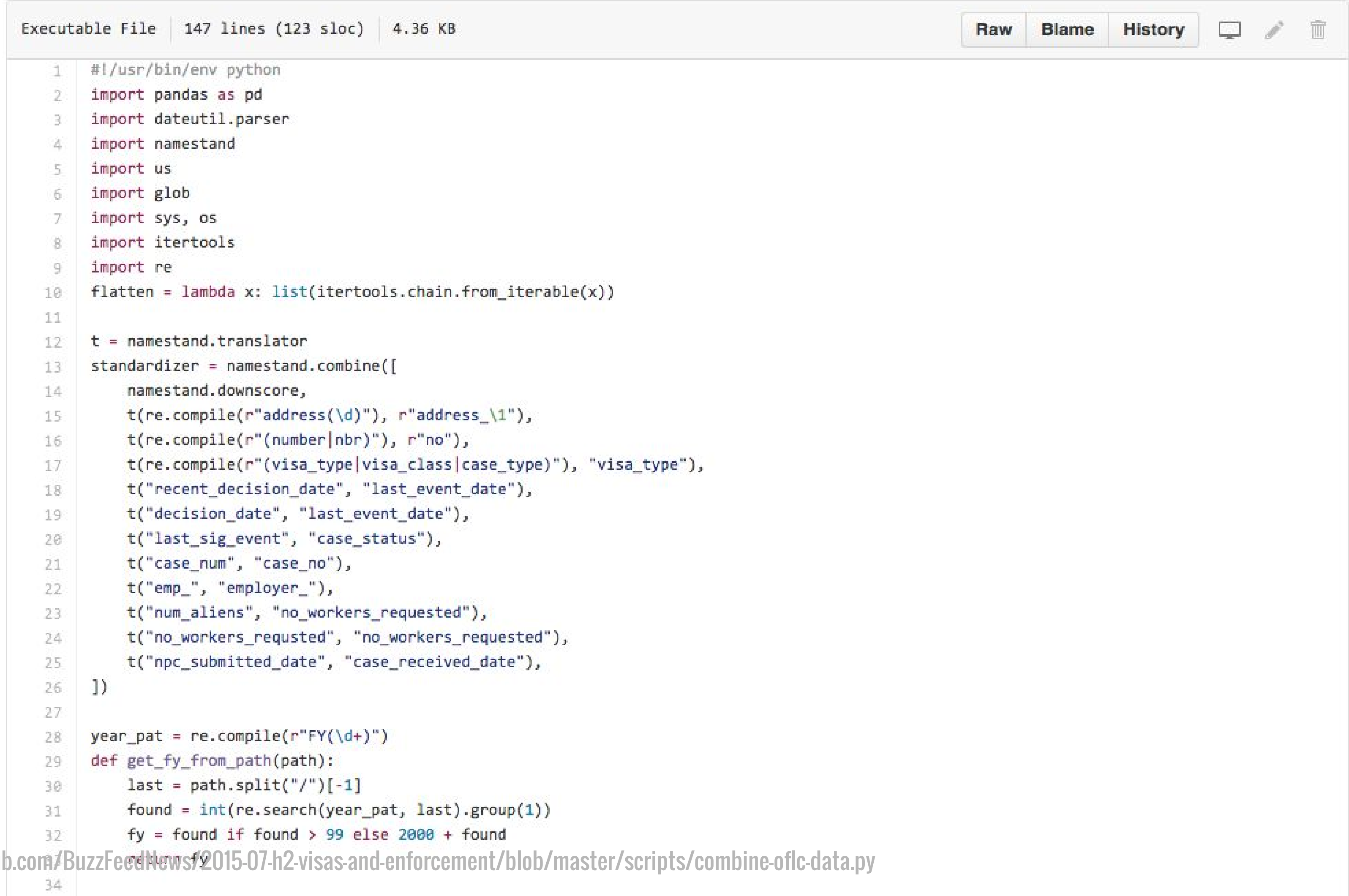


Data and analysis supporting several passages in the BuzzFeed News article, "The New American Slavery: Invited To The U.S., Foreign Workers Find A Nightmare," published July 24, 2015. http://www.buzzfeed.com/jessicagarrison/the-new-american-slavery-invited-to-the-us-foreign-workers-f

jsvine Add link to published article — Latest commit 5d04040 on Jul 24, 2015

data	Initial commit	8 months ago
docs	Initial commit	8 months ago
notebooks	Initial commit	8 months ago
output	Initial commit	8 months ago
scripts	Initial commit	8 months ago
utils	Initial commit	8 months ago
.gitignore	Initial commit	8 months ago
LICENSE.txt	Initial commit	8 months ago
Makefile	Initial commit	8 months ago
README.md	Add link to published article	8 months ago
requirements.txt	Initial commit	8 months ago

https://github.com/BuzzFeedNews/2015-07-h2-visas-and-enforcement



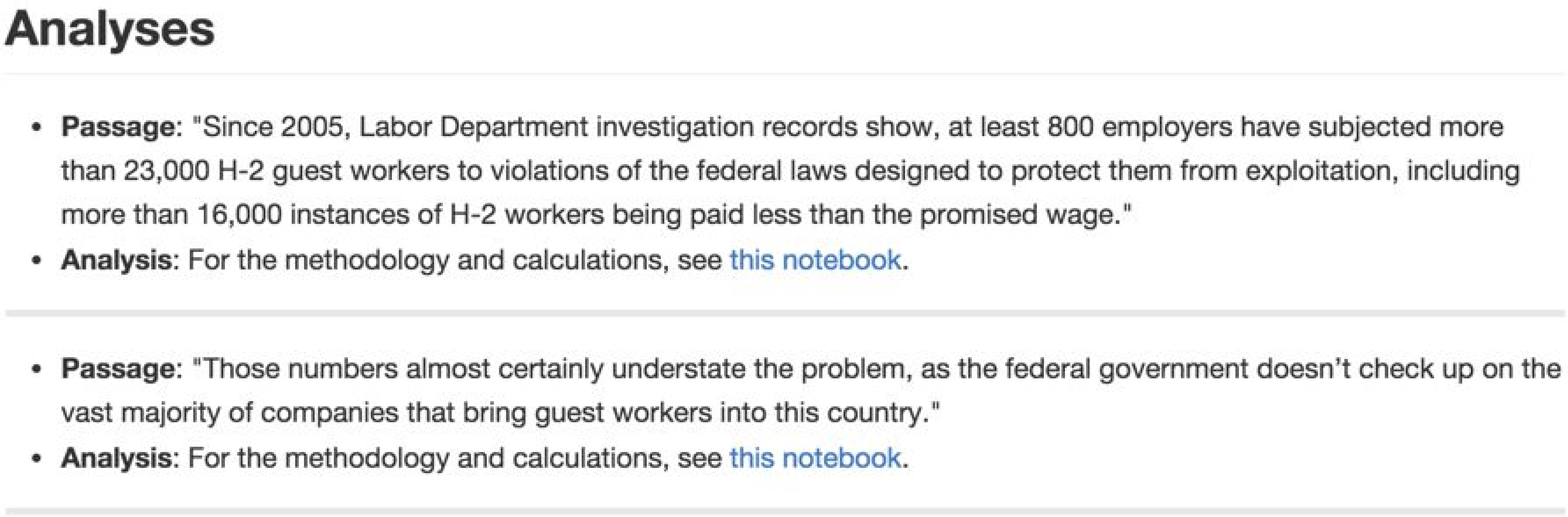
```python
#!/usr/bin/env python
import pandas as pd
import dateutil.parser
import namestand
import us
import glob
import sys, os
import itertools
import re
flatten = lambda x: list(itertools.chain.from_iterable(x))

t = namestand.translator
standardizer = namestand.combine([
    namestand.downscore,
    t(re.compile(r"address(\d)"), r"address_\1"),
    t(re.compile(r"(number|nbr)"), r"no"),
    t(re.compile(r"(visa_type|visa_class|case_type)"), "visa_type"),
    t("recent_decision_date", "last_event_date"),
    t("decision_date", "last_event_date"),
    t("last_sig_event", "case_status"),
    t("case_num", "case_no"),
    t("emp_", "employer_"),
    t("num_aliens", "no_workers_requested"),
    t("no_workers_requsted", "no_workers_requested"),
    t("npc_submitted_date", "case_received_date"),
])

year_pat = re.compile(r"FY(\d+)")
def get_fy_from_path(path):
    last = path.split("/")[-1]
    found = int(re.search(year_pat, last).group(1))
    fy = found if found > 99 else 2000 + found
    return fy

```

https://github.com/BuzzFeedNews/2015-07-h2-visas-and-enforcement/blob/master/scripts/combine-oflc-data.py

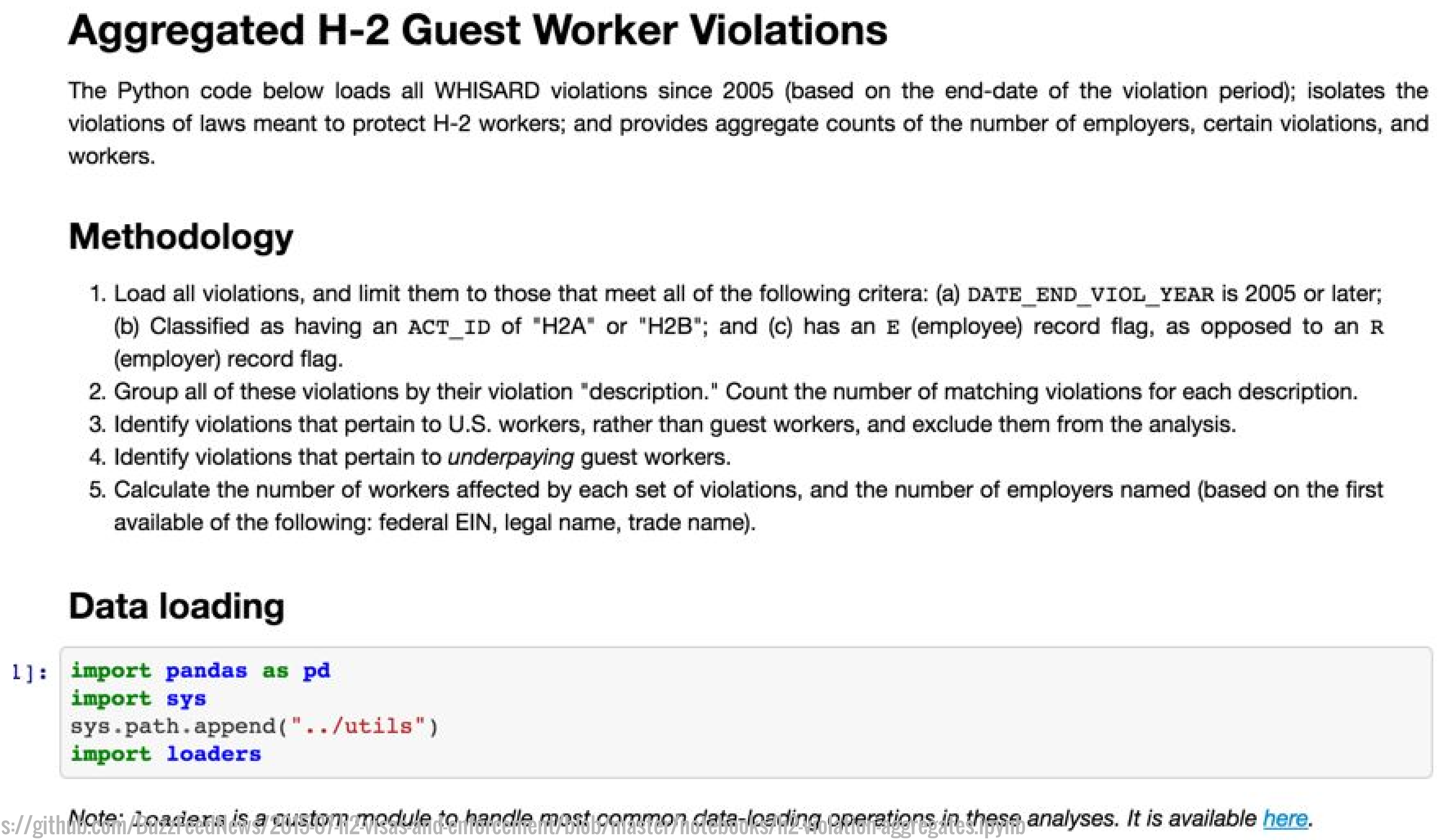


Analyses

- **Passage**: "Since 2005, Labor Department investigation records show, at least 800 employers have subjected more than 23,000 H-2 guest workers to violations of the federal laws designed to protect them from exploitation, including more than 16,000 instances of H-2 workers being paid less than the promised wage."
- **Analysis**: For the methodology and calculations, see this notebook.

- **Passage**: "Those numbers almost certainly understate the problem, as the federal government doesn't check up on the vast majority of companies that bring guest workers into this country."
- **Analysis**: For the methodology and calculations, see this notebook.

https://github.com/BuzzFeedNews/2015-07-h2-visas-and-enforcement/blob/master/scripts/combine-oflc-data.py



Aggregated H-2 Guest Worker Violations

The Python code below loads all WHISARD violations since 2005 (based on the end-date of the violation period); isolates the violations of laws meant to protect H-2 workers; and provides aggregate counts of the number of employers, certain violations, and workers.

Methodology

1. Load all violations, and limit them to those that meet all of the following critera: (a) `DATE_END_VIOL_YEAR` is 2005 or later; (b) Classified as having an `ACT_ID` of "H2A" or "H2B"; and (c) has an `E` (employee) record flag, as opposed to an `R` (employer) record flag.
2. Group all of these violations by their violation "description." Count the number of matching violations for each description.
3. Identify violations that pertain to U.S. workers, rather than guest workers, and exclude them from the analysis.
4. Identify violations that pertain to *underpaying* guest workers.
5. Calculate the number of workers affected by each set of violations, and the number of employers named (based on the first available of the following: federal EIN, legal name, trade name).

Data loading

```
1]: import pandas as pd
    import sys
    sys.path.append("../utils")
    import loaders
```

Note: `loaders` is a custom module to handle most common data-loading operations in these analyses. It is available here.

https://github.com/BuzzFeedNews/2015-07-h2-visas-and-enforcement/blob/master/notebooks/h2-violation-aggregates.ipynb

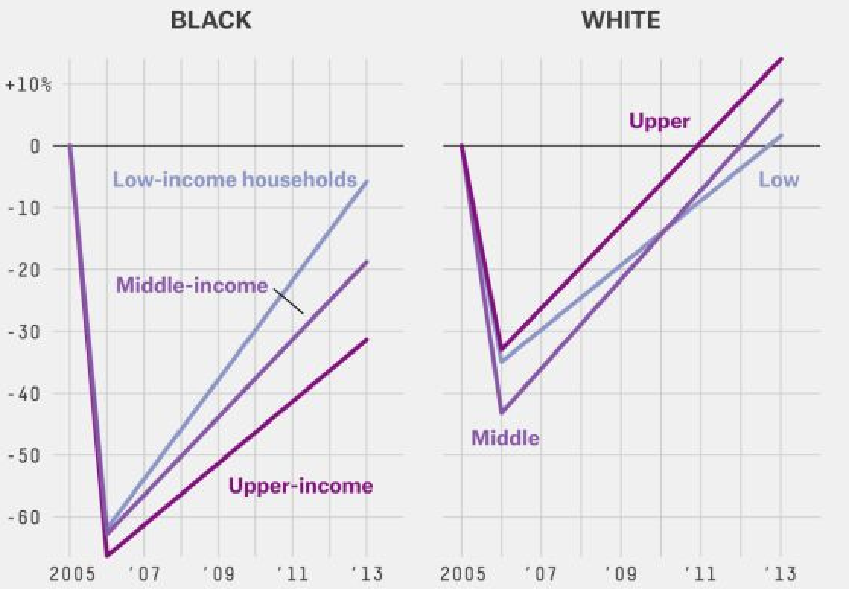
Does this help most readers understand what we did to the data?

What is the goal for doing this?

Examples of when we do... *sort of*

The Black Middle Class Has Struggled To Rebound

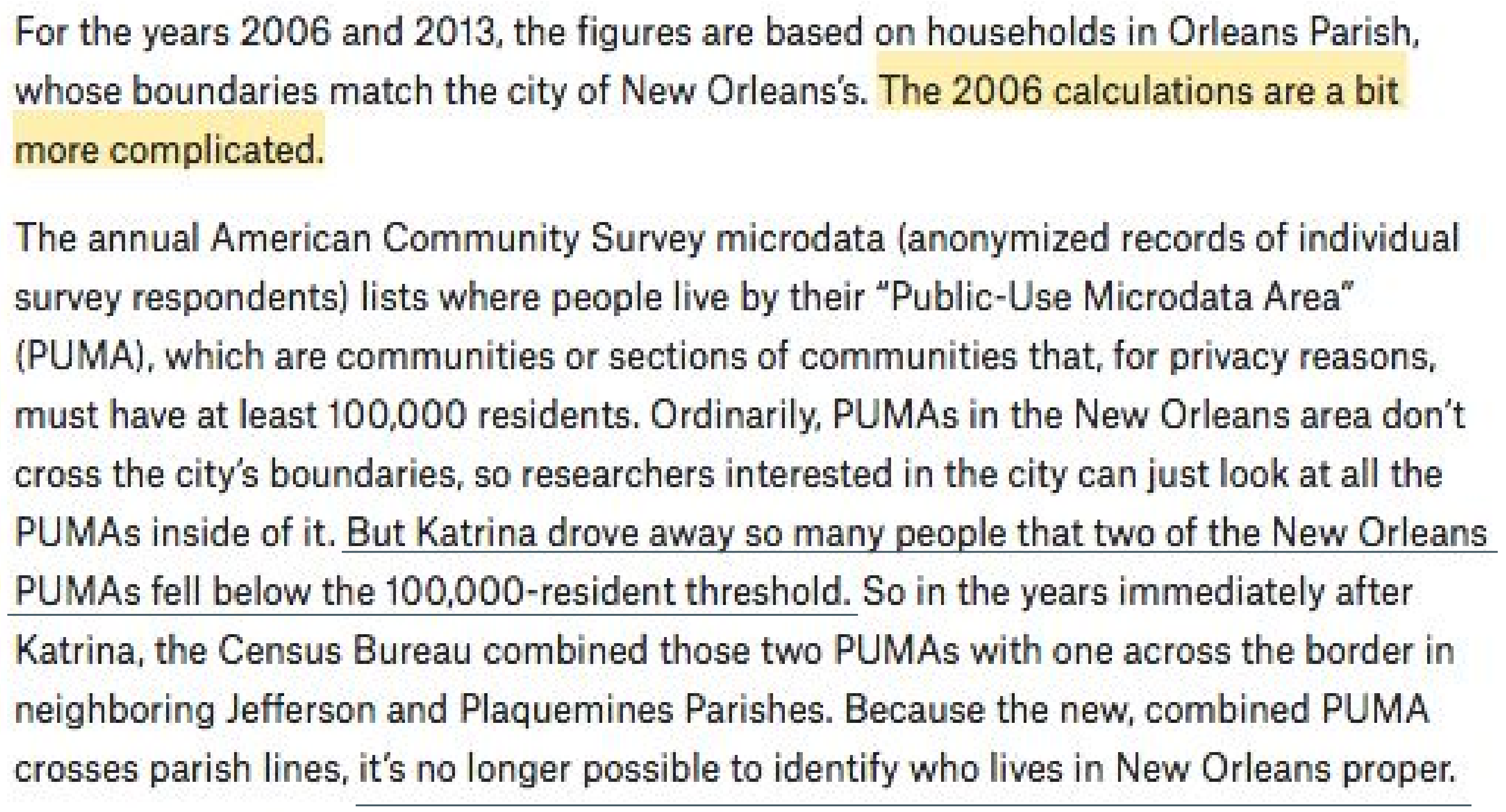
Percent change since before Katrina in the number of black and white households by income level, adjusted for inflation



FIVETHIRTYEIGHT

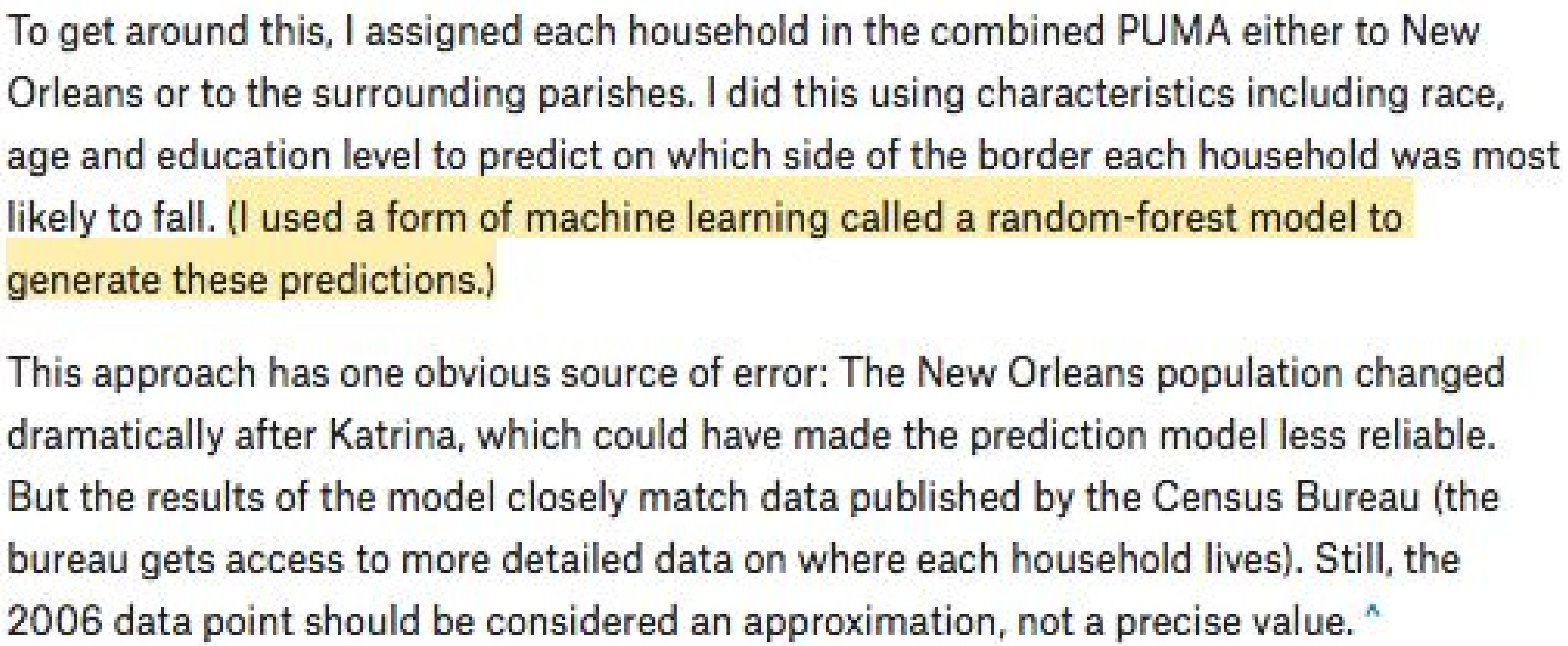
BASED ON AMERICAN COMMUNITY SURVEY DATA

http://fivethirtyeight.com/features/katrina-washed-away-new-orleanss-black-middle-class/



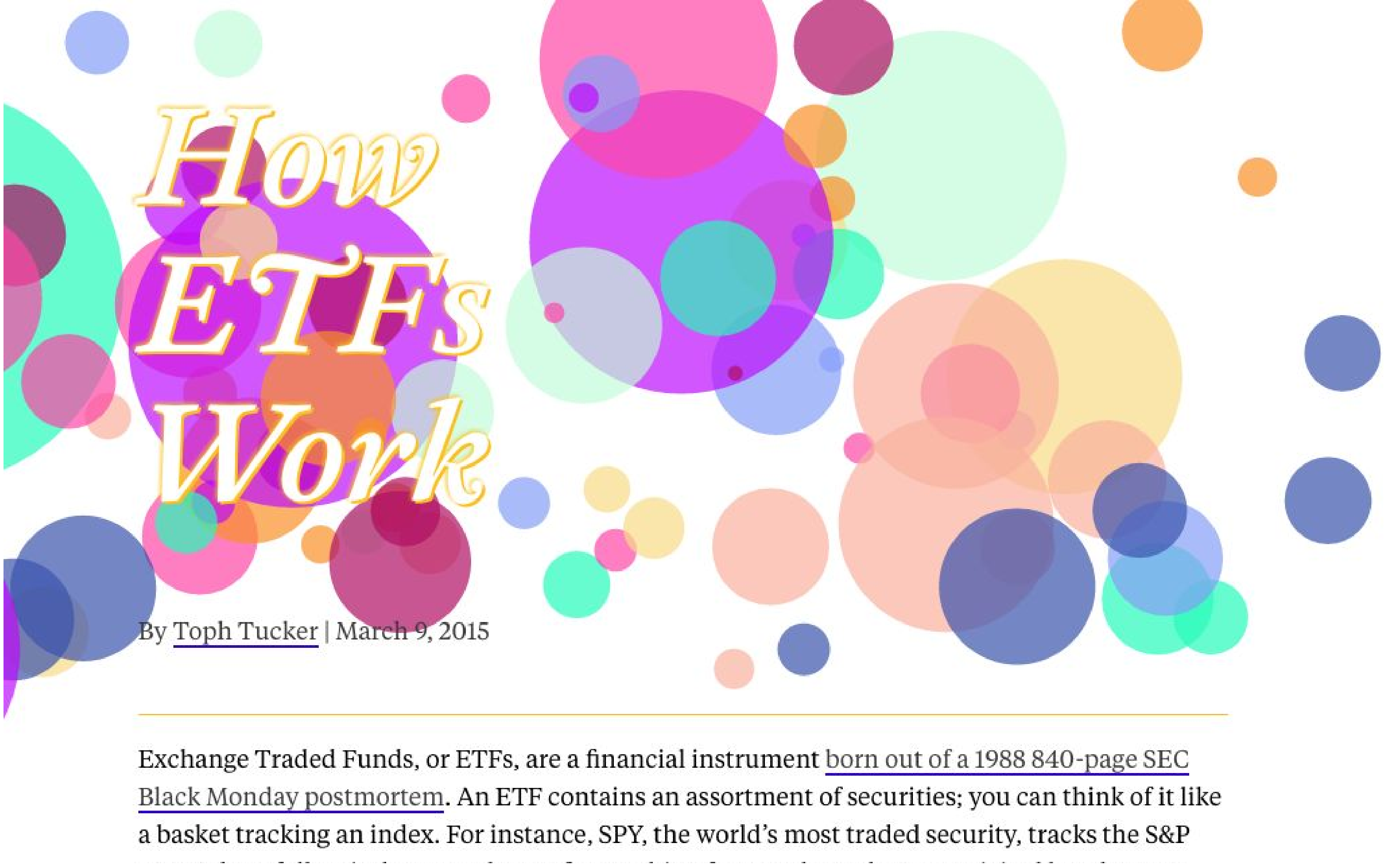
For the years 2006 and 2013, the figures are based on households in Orleans Parish, whose boundaries match the city of New Orleans's. The 2006 calculations are a bit more complicated.

The annual American Community Survey microdata (anonymized records of individual survey respondents) lists where people live by their "Public-Use Microdata Area" (PUMA), which are communities or sections of communities that, for privacy reasons, must have at least 100,000 residents. Ordinarily, PUMAs in the New Orleans area don't cross the city's boundaries, so researchers interested in the city can just look at all the PUMAs inside of it. But Katrina drove away so many people that two of the New Orleans PUMAs fell below the 100,000-resident threshold. So in the years immediately after Katrina, the Census Bureau combined those two PUMAs with one across the border in neighboring Jefferson and Plaquemines Parishes. Because the new, combined PUMA crosses parish lines, it's no longer possible to identify who lives in New Orleans proper.



To get around this, I assigned each household in the combined PUMA either to New Orleans or to the surrounding parishes. I did this using characteristics including race, age and education level to predict on which side of the border each household was most likely to fall. (I used a form of machine learning called a random-forest model to generate these predictions.)

This approach has one obvious source of error: The New Orleans population changed dramatically after Katrina, which could have made the prediction model less reliable. But the results of the model closely match data published by the Census Bureau (the bureau gets access to more detailed data on where each household lives). Still, the 2006 data point should be considered an approximation, not a precise value. ^

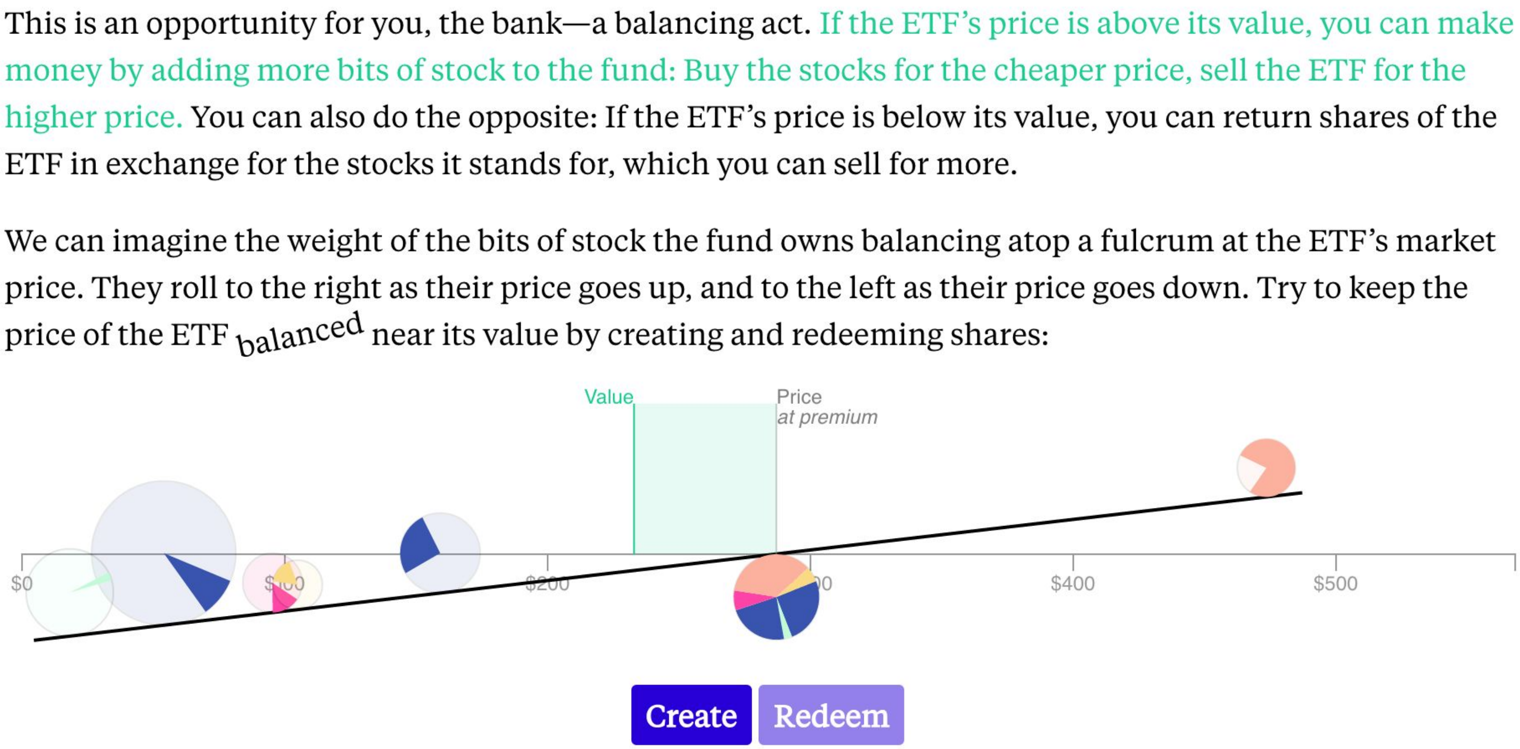


How ETFs Work

By Toph Tucker | March 9, 2015

Exchange Traded Funds, or ETFs, are a financial instrument born out of a 1988 840-page SEC Black Monday postmortem. An ETF contains an assortment of securities; you can think of it like a basket tracking an index. For instance, SPY, the world's most traded security, tracks the S&P

http://www.bloomberg.com/features/2016-etf-files/toy/



This is an opportunity for you, the bank—a balancing act. If the ETF’s price is above its value, you can make money by adding more bits of stock to the fund: Buy the stocks for the cheaper price, sell the ETF for the higher price. You can also do the opposite: If the ETF’s price is below its value, you can return shares of the ETF in exchange for the stocks it stands for, which you can sell for more.

We can imagine the weight of the bits of stock the fund owns balancing atop a fulcrum at the ETF’s market price. They roll to the right as their price goes up, and to the left as their price goes down. Try to keep the price of the ETF balanced near its value by creating and redeeming shares:

http://www.bloomberg.com/features/2016-etf-files/toy/

Editors: Joel Weber, Thomas Houston; thanks also to Eric Balchunas

Methodology: The core metaphor relies on an analogy between the center of mass, `( ∑ mass * position ) / total mass`, and net asset value, `( ∑ shares * price ) / total shares`. The rate of change of stocks follows a log-normal random walk, which is smooth only for the sake of usability; for a more realistic (rougher and nowhere-differentiable) model, consider fractional Brownian motion. The premium at which the ETF trades also follows a random walk, weakly attracted to a fair 1:1 ratio. Market impact is simplistic and linear; there is no bid-ask spread; the ETF rebalances continuously; and the expense ratio is assumed to be negligible. Tweet at @tophtucker with corrections or pedantry.

Methodology: The core metaphor relies on an analogy between the center of mass, (**∑ mass * position) / total mass**, and net asset value, (∑ shares * price) / total shares. The rate of change of stocks follows a **log-normal random walk,** which is smooth only for the sake of usability; **for a more realistic (rougher and nowhere-differentiable) model**, **consider fractional Brownian motion**. The premium at which the ETF trades also follows a random walk, weakly attracted to a fair 1:1 ratio. Market impact is simplistic and linear; there is no bid-ask spread; the ETF rebalances continuously; and the expense ratio is assumed to be negligible. Tweet at @tophtucker with **corrections or pedantry**.

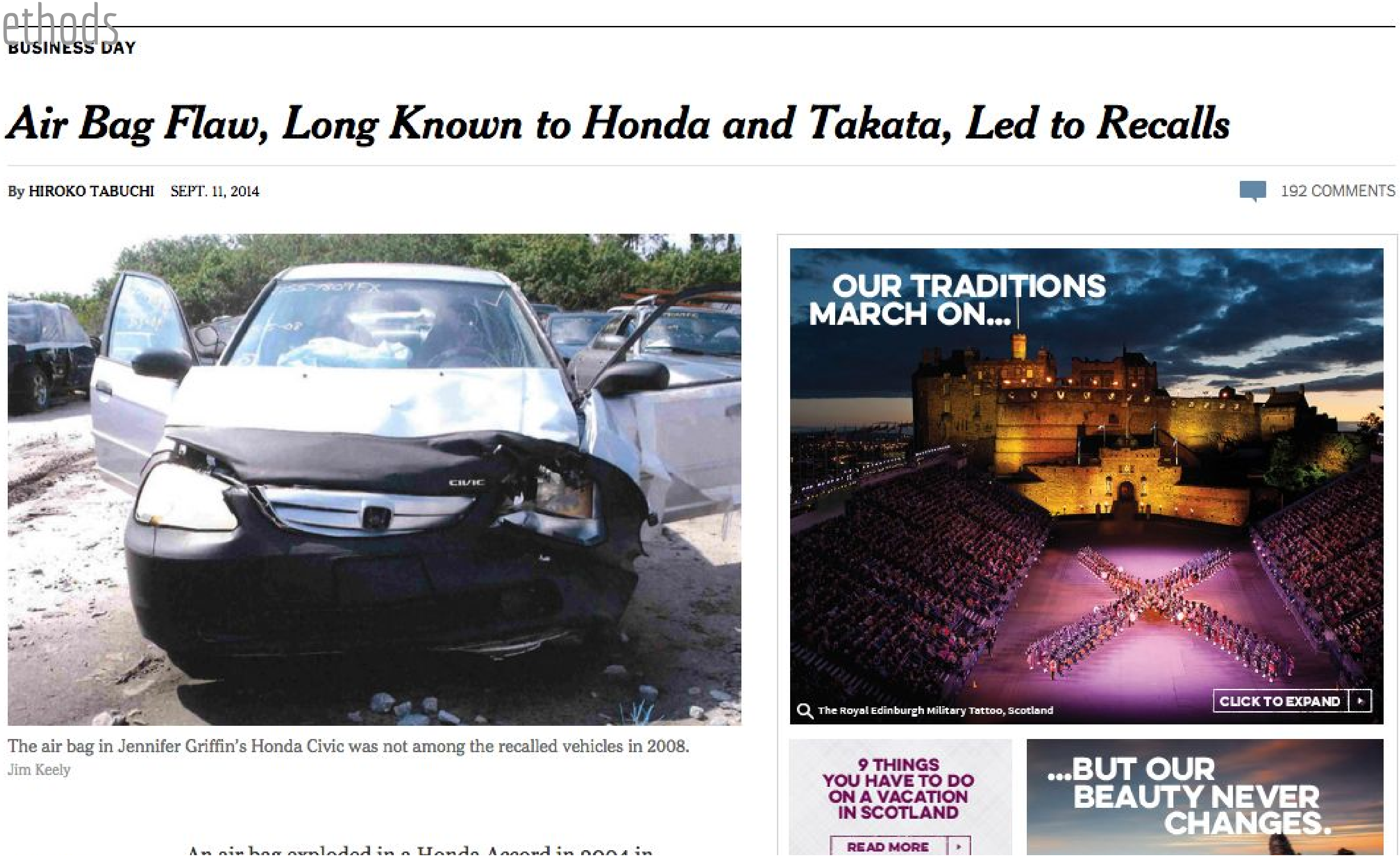
What parts of complex analysis should we explain?

Are readers equipped to understand it?

And do they care?

Examples of when we *don't*

Most of the time?



BUSINESS DAY

Air Bag Flaw, Long Known to Honda and Takata, Led to Recalls

By HIROKO TABUCHI SEPT. 11, 2014

192 COMMENTS

The air bag in Jennifer Griffin's Honda Civic was not among the recalled vehicles in 2008.

Jim Keely

An air bag exploded in a Honda Accord in 2004 in

http://www.nytimes.com/2014/09/12/business/air-bag-flaw-long-known-led-to-recalls.html

Daeil Kim

The Machine Learning Approach
A sample comment. We will preprocess this data for the algorithm
TOKENIZE
FILTER
DATA IS READY FOR TRAINING!

http://www.slideshare.net/mortardata/daeil-kim-at-the-nyc-data-science-meetup

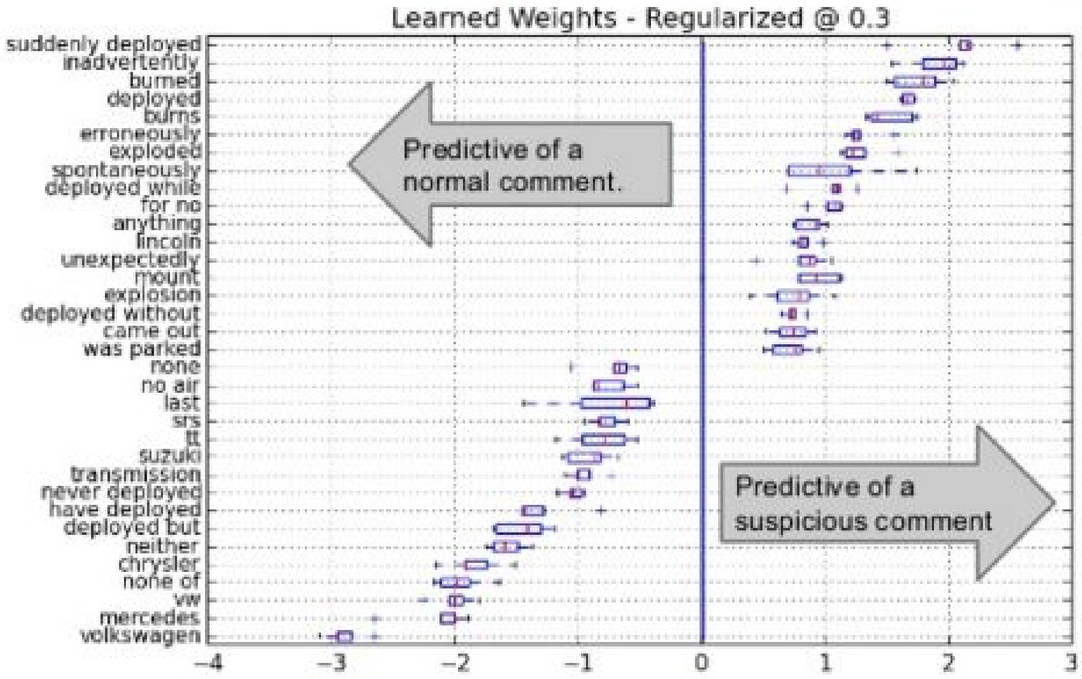
Daeil Kim

Data Scientist at NYT

Develop a prediction algorithm based on whether a comment is suspicious.

Tokenize → Filter → Train → Cross-validation

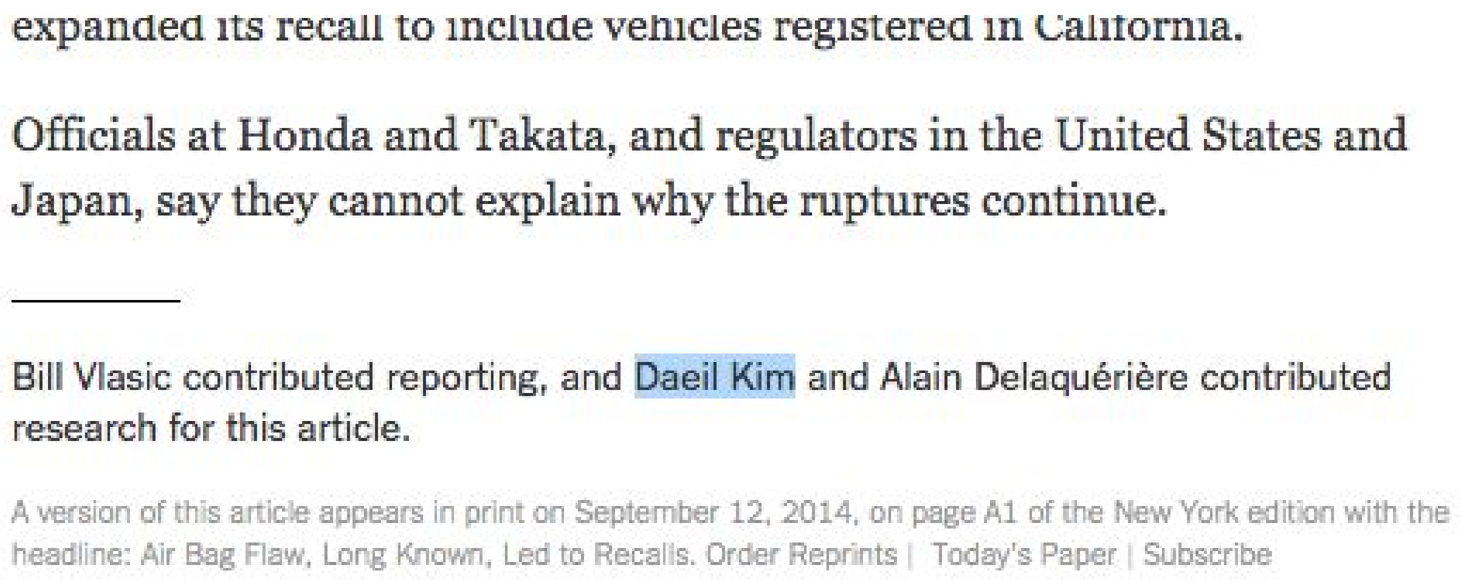
The most predictive words / features



After training the model, we then applied this on the full dataset.

We looked for comments that Hiroko didn't label as being suspicious, but the algorithm did to follow up on (374 / 33K total).

***Result:* 7 new cases where a passenger was injured were discovered from those comments she missed.**



expanded its recall to include vehicles registered in California.

Officials at Honda and Takata, and regulators in the United States and Japan, say they cannot explain why the ruptures continue.

Bill Vlasic contributed reporting, and Daeil Kim and Alain Delaquérière contributed research for this article.

A version of this article appears in print on September 12, 2014, on page A1 of the New York edition with the headline: Air Bag Flaw, Long Known, Led to Recalls. Order Reprints | Today's Paper | Subscribe

Is there any harm for readers in ***not*** explaining your methods?

Do readers have different expectations for transparency depending on news org?

#nicarmethods

Chris Groskopf	@onyxfish
Ryann Jones	@ryanngro
Aaron Williams	@aboutaaron
Stuart Thompson	@stuartathompson