Algorithms week 1:
Algorithms in Journalism

Jonathan Stray
Columbia Lede Program
July 18, 2018

Words <-> numbers

Are there quantitative statements here?

In a country where a sitting congressman told a crowd that evolution and the Big Bang are "lies straight from the pit of hell," where the chairman of a Senate environmental panel brought a snowball into the chamber as evidence that climate change is a hoax, where almost one in three citizens can't name the vice president, it is beyond dispute that critical thinking has been abandoned as a cultural value

Anti-Intellectualism Is Killing America, Psychology Today, 23 June 2015

Regression

The Boston Globe

July 20, 2003
(updated July 25, 2003)

A Boston Globe analysis of traffic tickets and warnings, from every police department in Massachusetts, shows differences in race, sex and age in who gets a fine, and who gets a break, for the same offenses.

	Statistical Significance	Odds Ratio	95.0% C.I. for Odds Ratio Lower	Upper
RACE (reference category = White)				
Black	.000	1.403	1.264	1.557
Hispanic	.000	1.424	1.218	1.664
Asian	.000	1.135	1.063	1.212
AGE GROUP (ref categ = < 25 years)				
26 - 40 years	.000	.350	.305	.401
40 - 55 years	.000	.592	.555	.631
55 - 70 years	.000	.666	.638	.695
Over 70	.000	.776	.746	.806
MPG OVER LIMIT (ref categ = <10 MPH)	.000			
10 to 15 MPH OVER	.000	18.269	16.357	20.404
16 to 20 MPH OVER	.000	50.319	44.911	56.377
MORE THAN 20 MPH OVER	.000	157.017	138.595	177.887
GENDER (ref categ = FEMALE)	.000	.762	.738	.786

Text analysis

Investigations

Whistleblowers say USAID's IG removed critical details from public reports

The Post obtained draft versions of 12 audits by the inspector general's office, covering projects from the Caribbean to Pakistan to the Republic of Georgia between 2011 and 2013. The drafts are confidential and rarely become public. The Post compared the drafts with the final reports published by the inspector general's office and interviewed former and current employees. E-mails and other internal records also were reviewed.

The Post tracked changes in the language that auditors used to describe USAID and its mission offices. The analysis found that more than 400 negative references were removed from the audits between the draft and final versions.

Sentiment analysis used by *Washington Post*, 2014

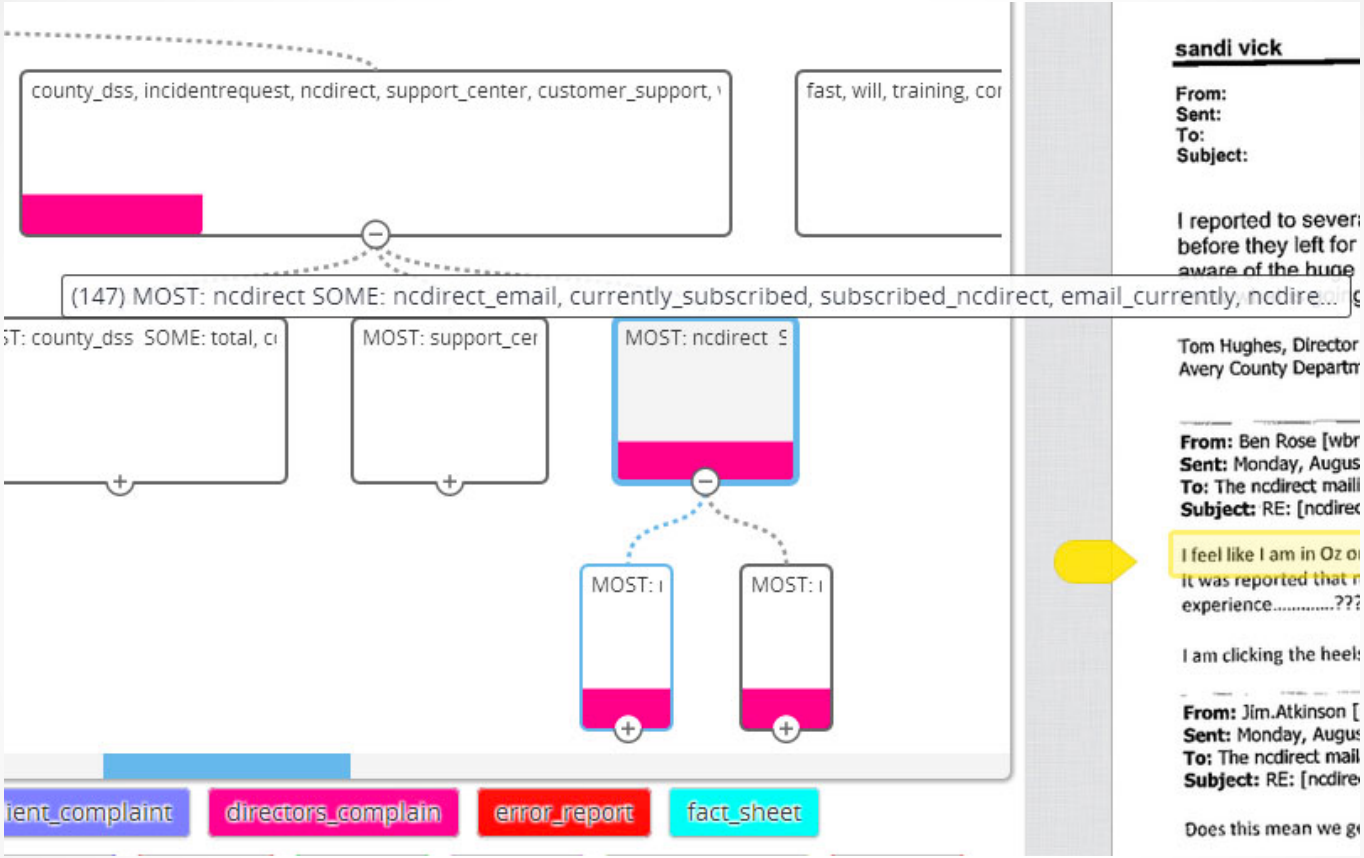
Records: DHHS downplayed food stamp issues

By Tyler Dukes

Posted: 1:00 p.m. today

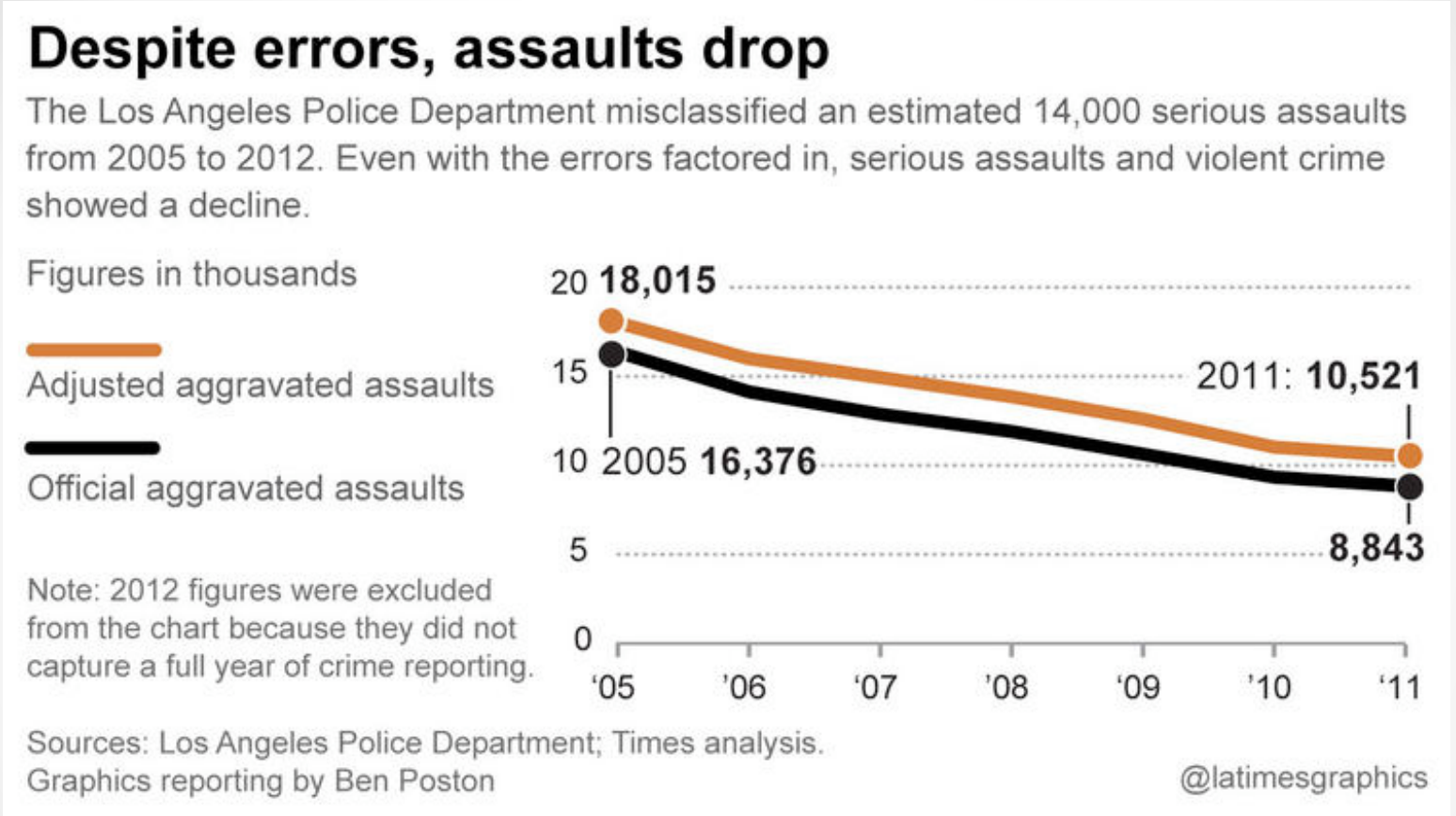
Technical troubles with a new system meant that almost 70,000 North Carolina residents received their food stamps late this summer. That's 8.5 percent of the number of clients the state currently serves every month. The problem was eventually traced to web browser compatibility issues. WRAL reporter Tyler Dukes obtained 4,500 pages of emails — on paper — from various government departments and used DocumentCloud and Overview to piece together this story.

https://blog.overviewdocs.com/completed-stories/

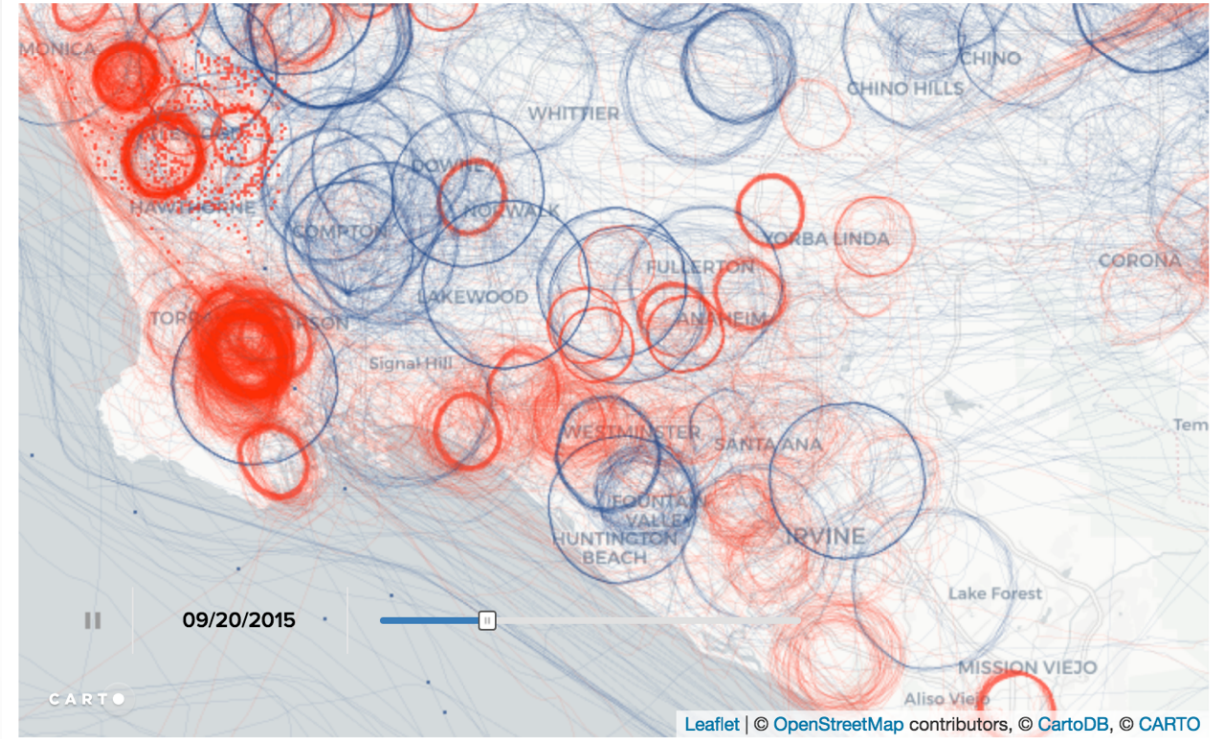


Overview's "topic tree" (unsupervised TF-IDF clustering) used to find a group of key emails from a mailing list (Tyler Dukes / WRAL)

Machine learning



LAPD Underreported Serious Assaults, Skewing Crime Stats for 8 Years
Los Angeles Times, 2015



U.S. government surveillance planes identified by machine learning on flight path data (Peter Aldhous/Buzzfeed)

Data cleanup

Example

Ingredient Phrase	1	tablespoon	fresh	lemon	juice
Ingredient Labels	QUANTITY	UNIT	COMMENT	NAME	NAME

Let $\{x^1, x^2, \ldots, x^N\}$ be the set of ingredient phrases, e.g. {"½ cups whole wheat flour", "pinch of salt", ...} where each x^i is an ordered list of words. Associated with each x^i is a list of tags, y^i.

For example, if $x^i = [x^i_1, x^i_2, x^i_3]$ = ["pinch", "of", "salt"] then $y^i = [y^i_1, y^i_2, y^i_3]$ = [UNIT, UNIT, NAME] . A tag is either a NAME, UNIT, QUANTITY, COMMENT or OTHER (i.e., none of the above).

The goal is to use data to learn a model that can predict the tag sequence for any ingredient phrase we throw at it, even if the model has never seen that ingredient phrase before. We approach this task by modeling the conditional probability of a sequence of tags given the input, denoted p(tag sequence | ingredient phrase) or using the above notation, $p(y \mid x)$.

Extracting Structured Data From Recipes Using Conditional Random Fields
Erica Greene, New York Times, 2015

For years now, organizations like the Center for Responsive Politics, the National Institute on Money in State Politics, the Sunlight Foundation have been standardizing donor names using a combination of automated analysis and human review. It's been an amazing service, and one I've used many times, but it's also made me wonder: Could machine learning accurately make the same judgments? Could we model the intuition of these expert standardizers and generalize it to any campaign finance dataset -- federal, state or local?

The short answer is yes. And this writeup will show you how it's done.

The intuition of the method I use follows four distinct steps:

- First, preprocess the data. Split donor names into first, middle and last; make all strings either all lowercase or uppercase; and add a few fields we'll need for testing later.
- Second, break our universe of 100,000 campaign contributions into smaller chunks that can be processed more quickly and efficiently.
- Third, use machine learning to run pairwise comparisons of individual donations to determine whether they are from the same person. If they are, link them together into miniature graphs.
- And finally, find all of the graphs of multiple connected donations and assign them a unique donor ID.

FEC Standardizer, Chase Davis https://github.com/cjdd3b/fec-standardizer/wiki

Simulations

Hillary Clinton has an **85% chance** to win.

CHANCE OF WINNING

85%
Hillary Clinton

15%
Donald J. Trump

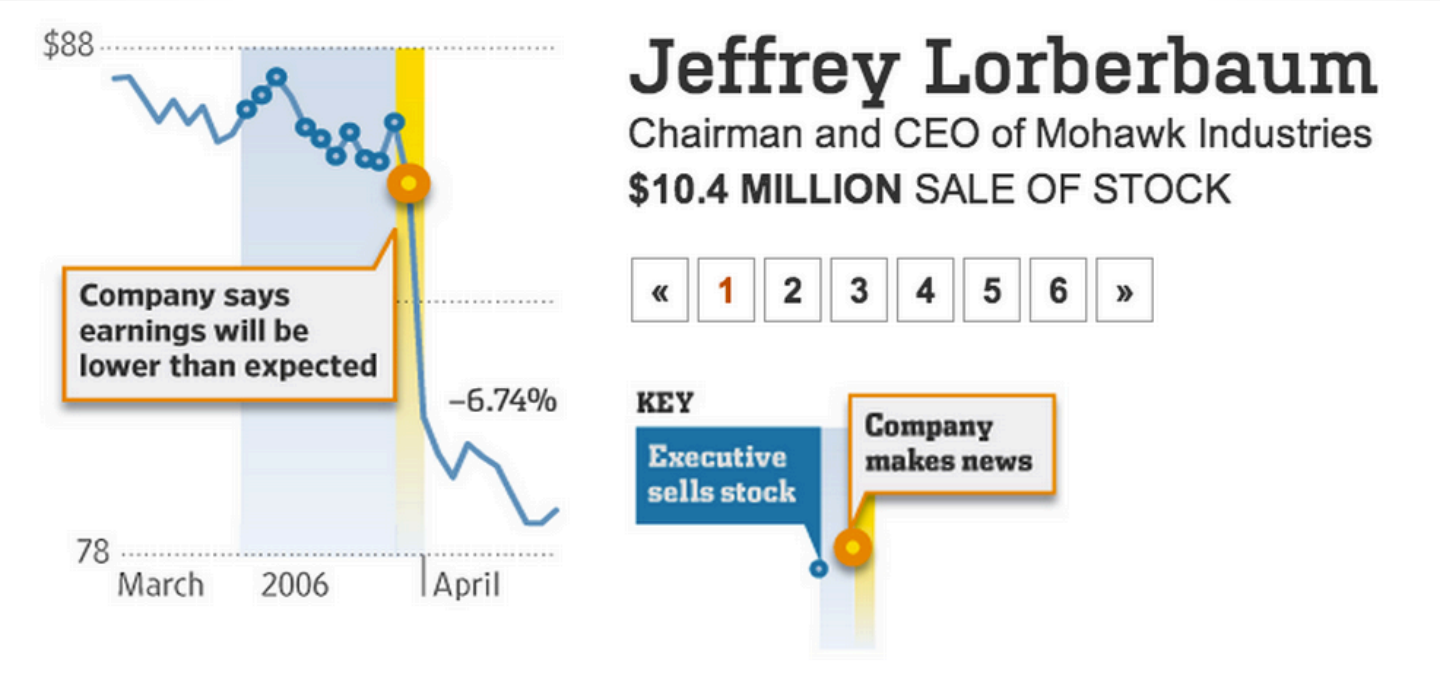
Last updated Tuesday, November 8 at 10:20 PM ET

Forecast history ↘ Recent changes ↘ State by state ↘ Other forecasts ↘ Likely scenarios ↘ Explore paths ↘

The Upshot's elections model suggests that Hillary Clinton is favored to win the presidency, based on the latest state and national polls. A victory by Mr. Trump remains possible: Mrs. Clinton's chance of losing is about the same as the probability that an N.F.L. kicker misses a 37-yard field goal.

New York Times 2016 Election Predictions

Randomization to detect insider trading



Randomization to detect secret payments?

The simulation confirmed that it is **extremely unlikely** that, by random chance alone, a set of payments near a specific date would almost equal $130,000.

For each of the 10,000 sets, we generated a "closeness" value—the difference between their "best match" and $130,000. For instance, if the "best match" was $130,014.29, the "closeness" value would be $14.29.

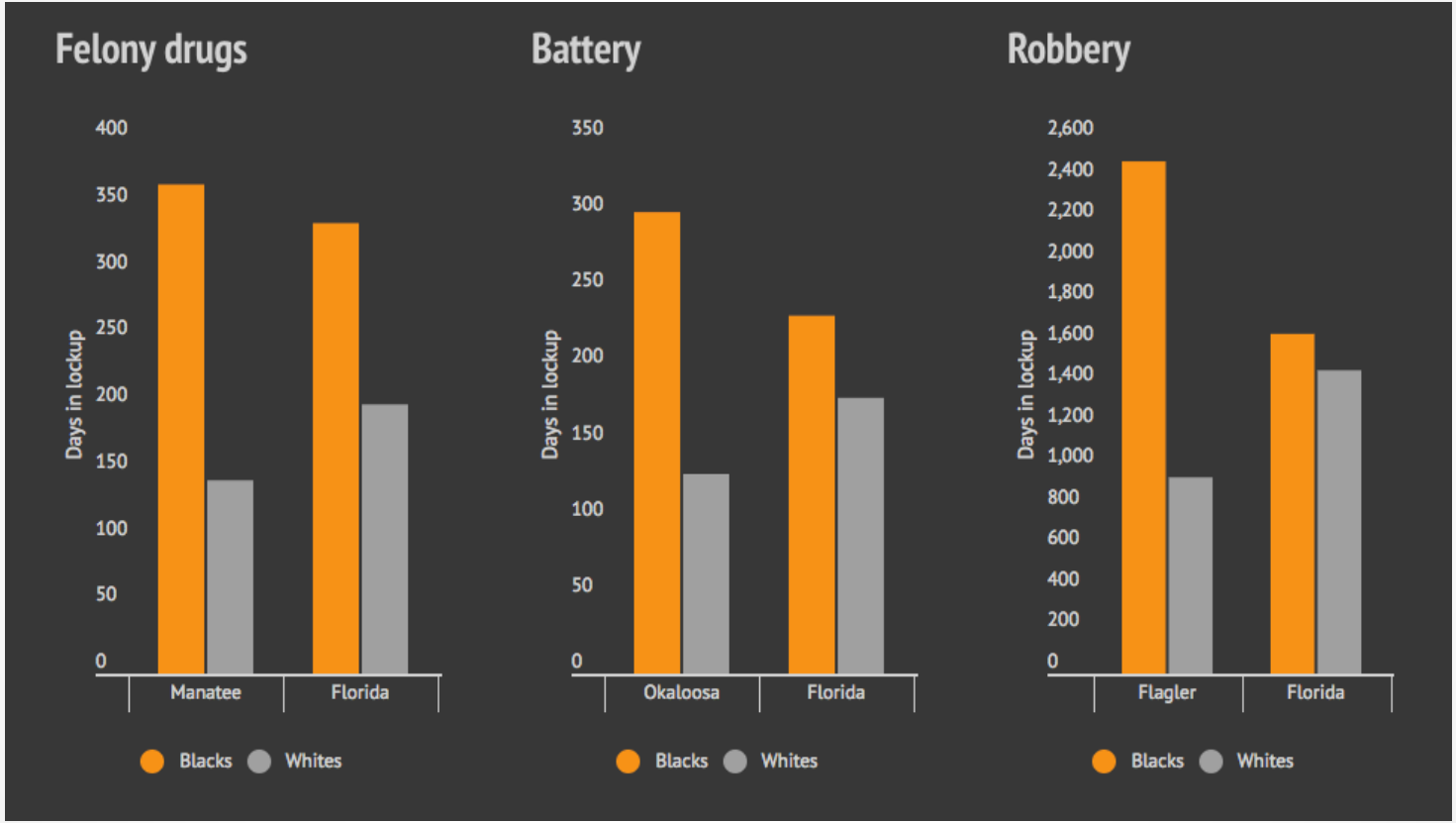
Percentile	Closeness to $130,000
0.001	$0.24
0.005	$1.44
0.01	$2.75
0.02	$5.42
0.03	$8.32
0.04	$11.32
0.05	$14.77
0.1	$33.48
0.2	$93.25
0.3	$283.59
0.4	$7,717.36
0.5	$20,340.55

Statistical Model Strongly Suggests the Stormy Daniels Payoff Came from the Trump Campaign,

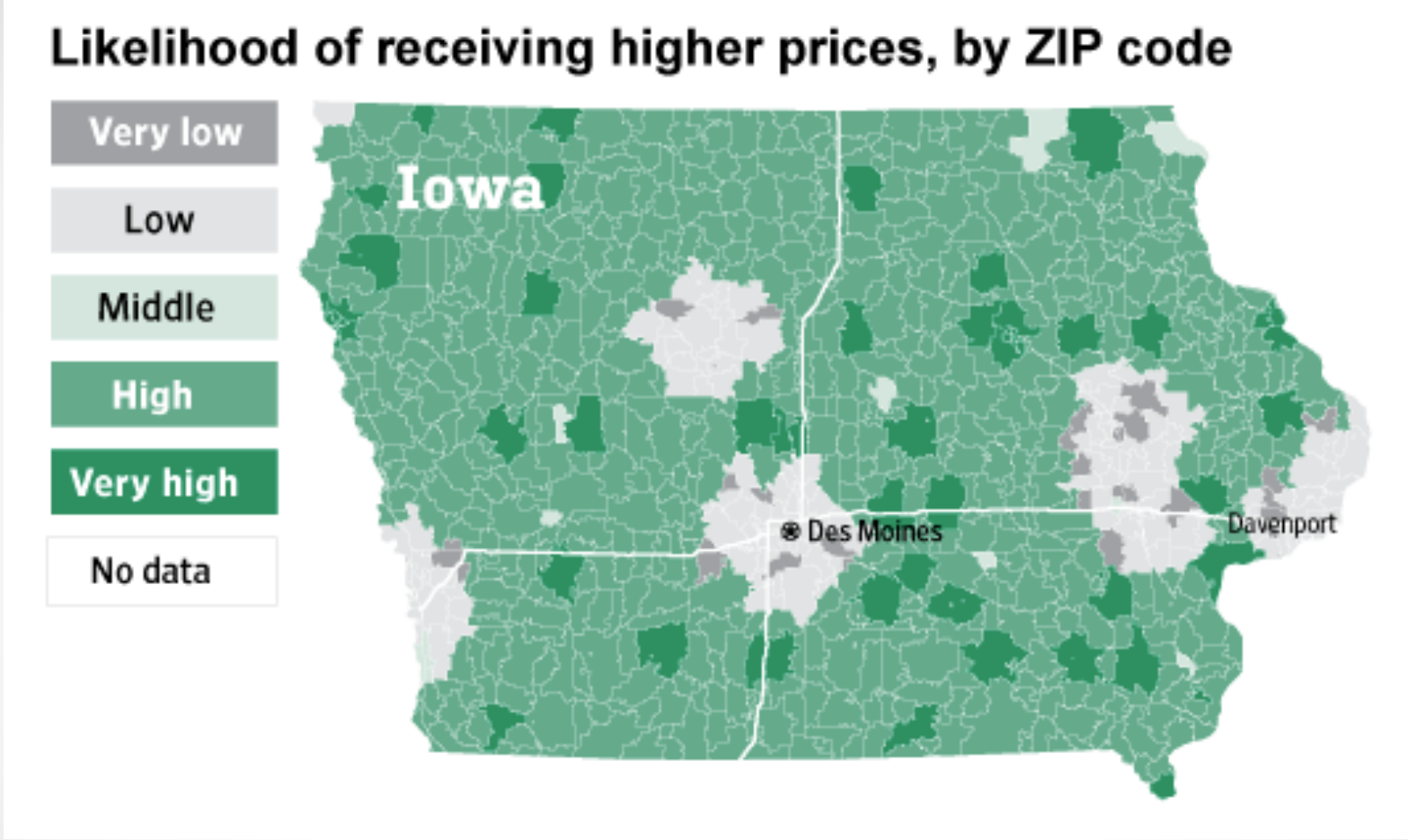
Will Stancil

Journalism on Algorithms

Florida sentencing analysis adjusted for "points"



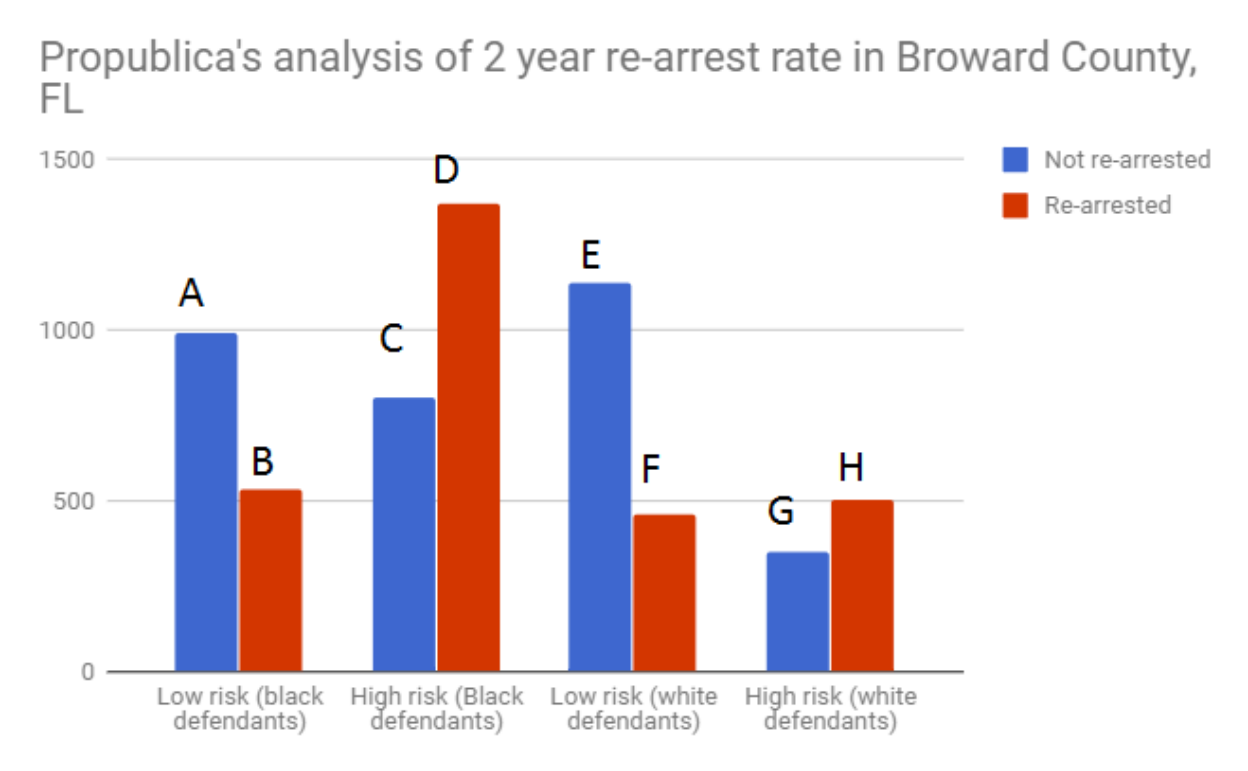
Bias on the Bench, Michael Braga, Herald Tribune



Websites Vary Prices, Deals Based on Users' Information

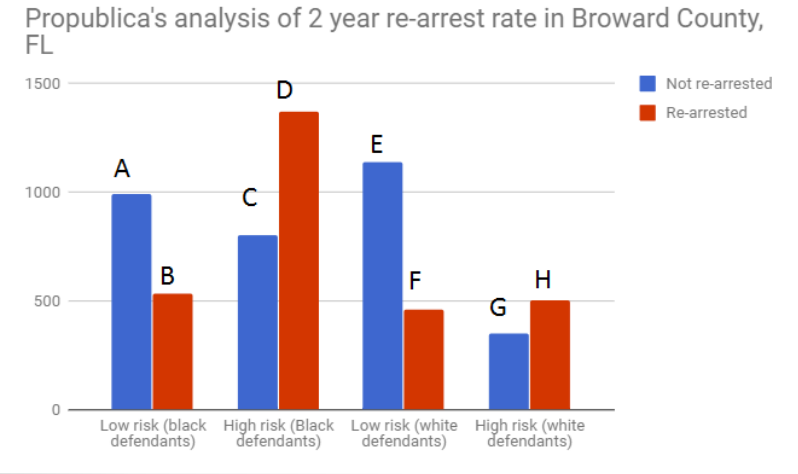
Valentino-Devries, Singer-Vine and Soltani, WSJ, 2012

	Low risk (black d	High risk (Black (	Low risk (white d	High risk (white defendants)
Not re-arrested	990	805	1139	349
Re-arrested	532	1369	461	505



Stephanie Wykstra, personal communication

	Low risk (black d	High risk (Black (	Low risk (white d	High risk (white defendants)
Not re-arrested	990	805	1139	349
Re-arrested	532	1369	461	505



ProPublica argument

False positive rate

P(high risk | black, no arrest) = C/(C+A) = 0.45

P(high risk | white, no arrest) = G/(G+E) = 0.23

False negative rate

P(low risk | black, arrested) = B/(B+D) = 0.28

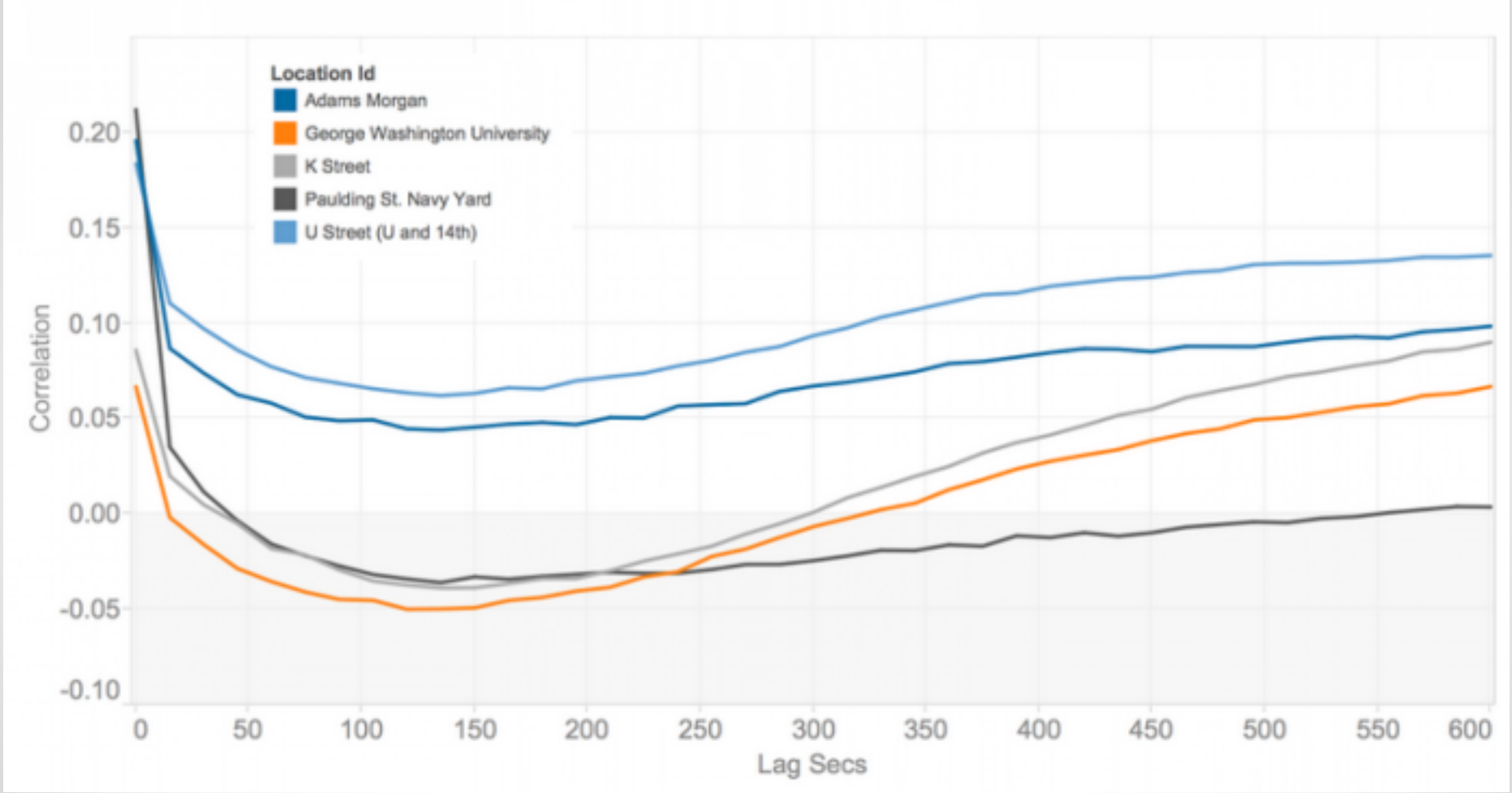
P(low risk | white, arrested) = F/(F+H) = 0.48

Northpointe response

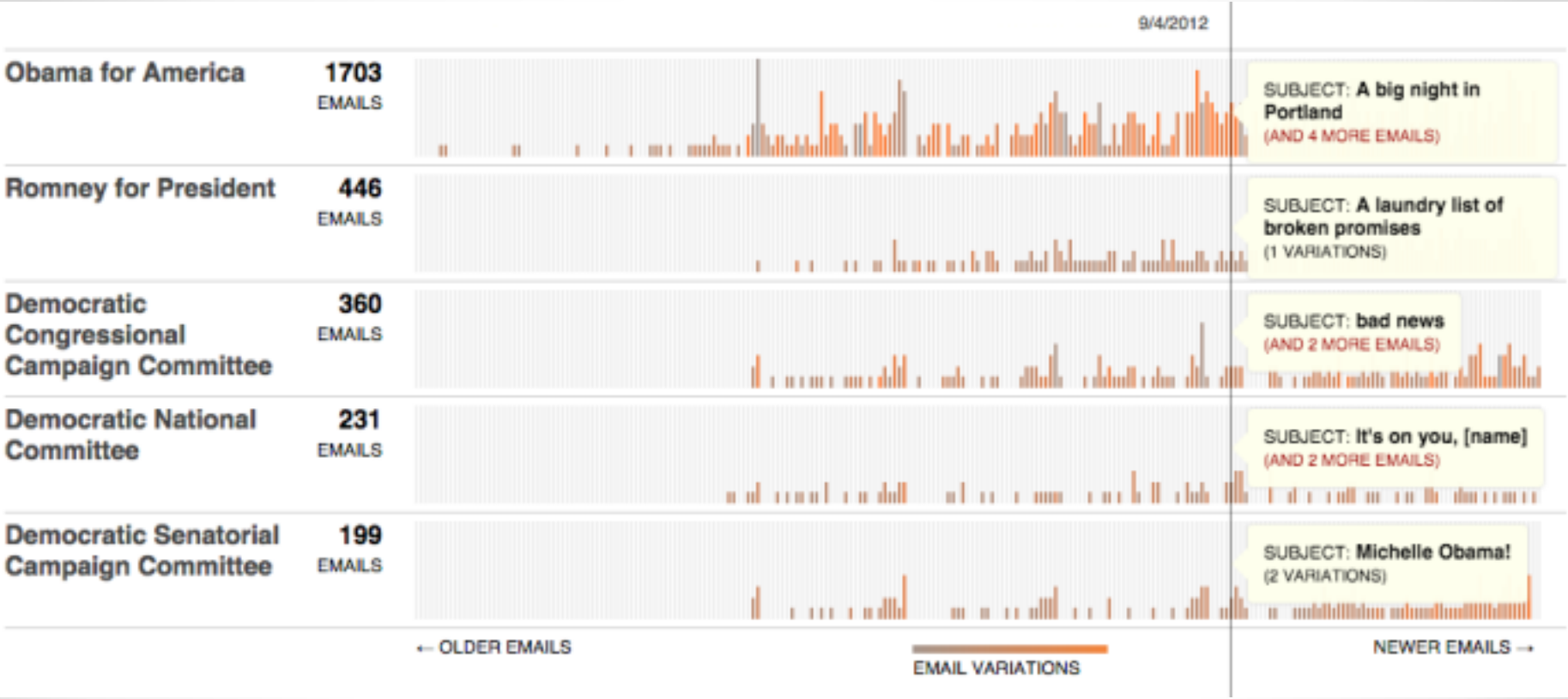
Positive predictive value

P(arrest | black, high risk) = D/(C+D) = 0.63

P(arrest | white, high risk) = H/(G+H) = 0.59



- *How Uber surge pricing really works,* Nick Diakopolous



Message Machine

Jeff Larson, Al Shaw, ProPublica, 2012

Algorithm Tips

Find tips for stories on algorithms

What is this?

This is a growing list of potentially newsworthy algorithms used by the U.S. government. As more decisions are influenced by algorithms, more algorithmic accountability may be needed. But where to start? Here, you can find algorithms warranting a closer look. Read about our criteria and sources.

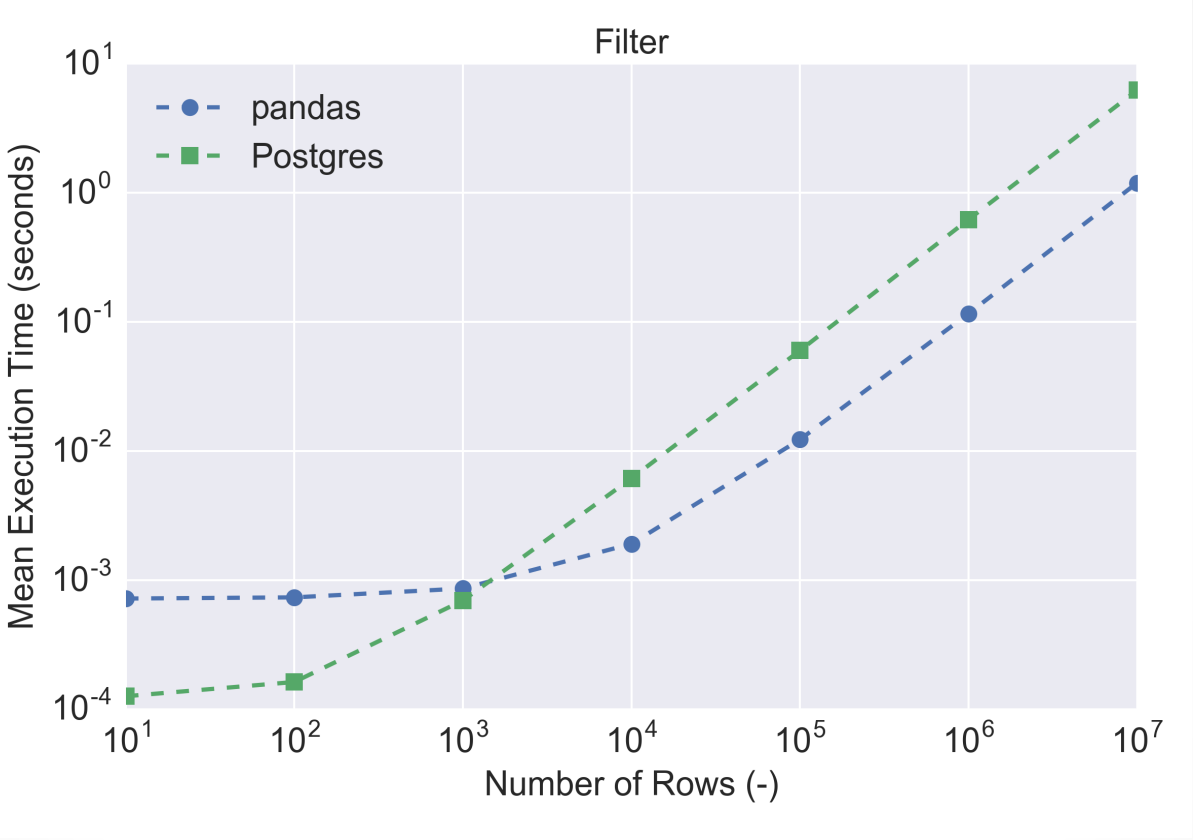
How to get started

Search for interesting algorithms using keywords relating to facets such as agency (e.g., Dept. of Justice) or topic (e.g., health, police, etc.). Then, on our resources page, learn how to submit FOIA requests about algorithms, or find news articles and research papers about the uses and risks of algorithms.

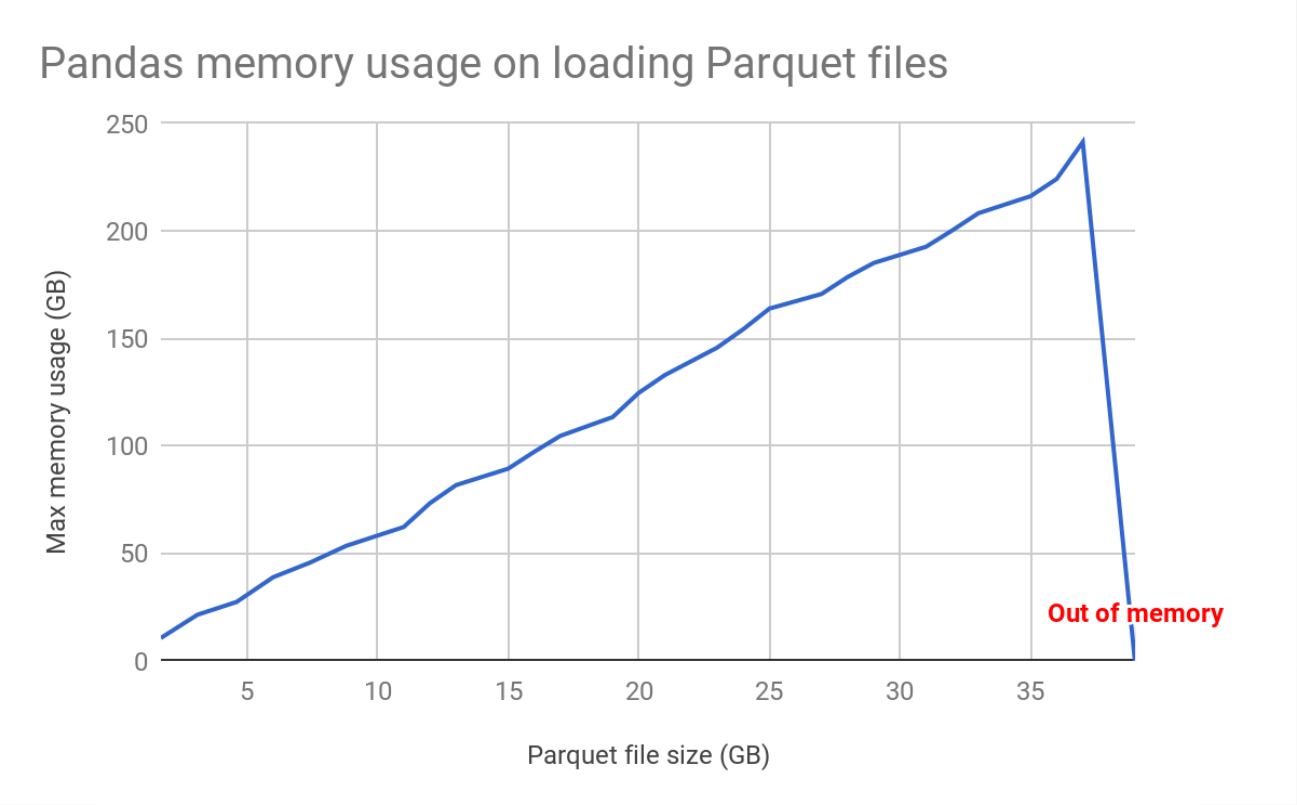
Want to help?

There are several ways to get involved. You can submit your own tip about a government algorithm. Or, volunteer to screen the tips we receive from our users. International collaborators welcome: we want to make this a global resource. Contact us to learn how you can collaborate to help us expand the tip list.

Cost of Algorithms: Time and Space



Time cost: How long will your code take to run?
Absolute time and trends vs size.
How many "steps"? What about loops?



Space cost: memory requirements or file usage

How big is your data? How many bytes to store a dataframe?

How many bytes to store e.g. an image?

An algorithm to find the maximum value

```
def max(x):
    m = x[0]
    for i in range (1, len(x)):
        if x[i] > m:
            m = x[i]

    return m
```

Match donors – all pairs comparison algorithm

```
def sum_donations_by_donor(d):
    for i in range(0, len(d)):
        name = d['name'][i]
        total = d['amount'][i]

        for j in range(i+1, len(d)):
            if d['name'][j] == name:
                total += d['amount'][j]

        print('Donor ' + name + ' gave ' + str(total))
```

Why mathematical formalism?

How do you know you're right?

percent change = (new – old)/old

So what's the old value, given new and percent change?