Lecture 1. Data Science?

Josh Clinton & Will Doyle

Today's Agenda

1. Meet the Instructor(s):

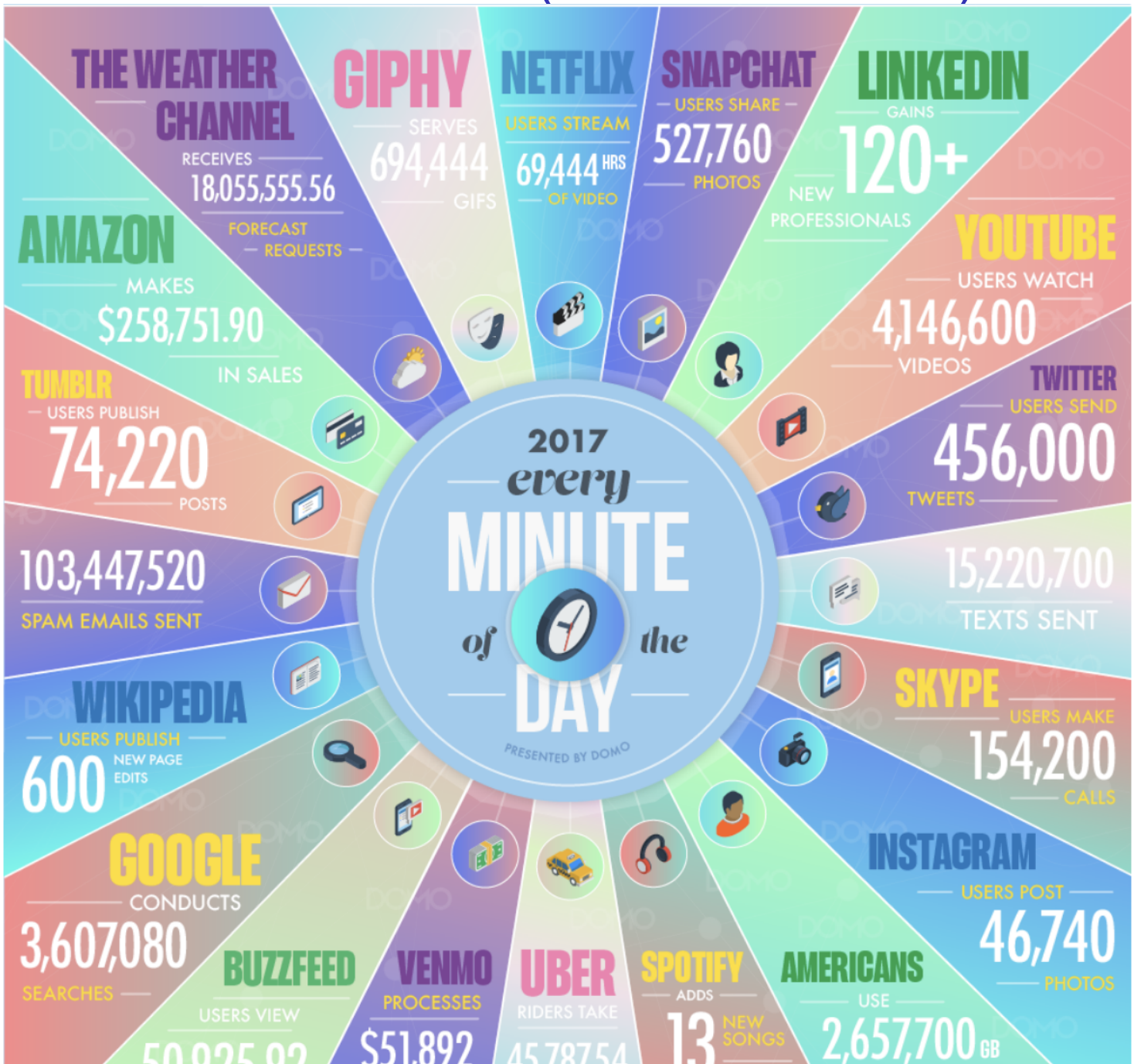
- Prof. Clinton: *josh.clinton@vanderbilt.edu*
- Prof. Doyle: *w.doyle@vanderbilt.edu*
- Jennifer Barnes (TA): *jennifer.n.barnes@vanderbilt.edu*

Course Email: *vu.dsci1000@gmail.com*

2. Course Motivation & Objectives

- Content: Critical Thinking, Analysis, Presentation
- Skills: Computing and Analysis using R

Too Much Information? (even in 2017...)



Course Motivation via “Tweets"

- “It is a capital mistake to theorize before one has data.” Sherlock Holmes
- “Torture the data, and it will confess to anything.” Ronald Coase, Nobel Prize Laureate in Economics
- “Here’s an open secret of the big data world: all data is dirty. All of it.” Meredith Broussard, *Artificial Unintelligence: How Computers Misunderstand the World*
- “Big Data processes codify the past. They do not invent the future.” Cathy O’Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*

Course Motivation

Data are more prevalent than ever!

But does quantification equal precision? Or knowledge? Or understanding?

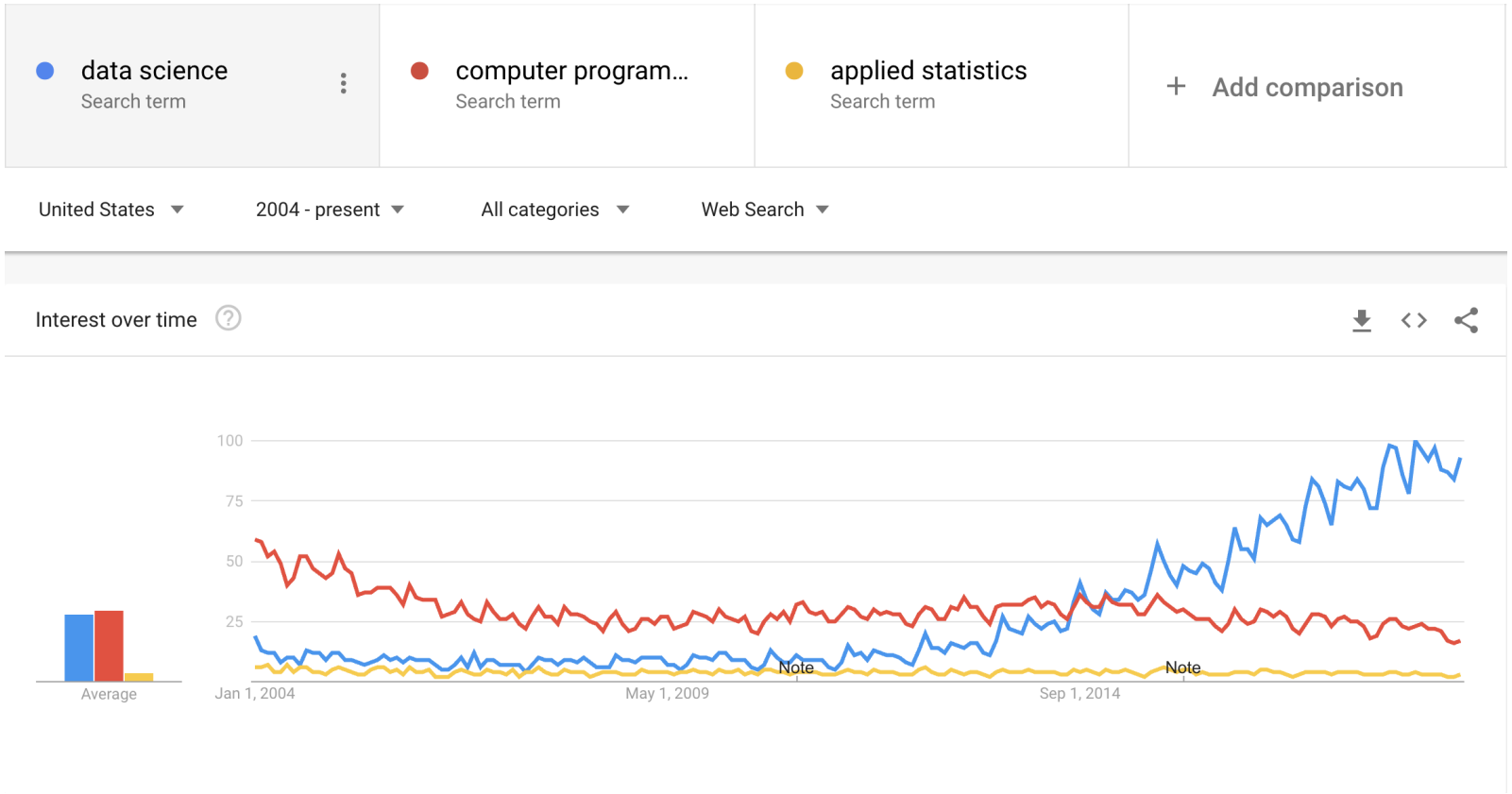
- Numbers without context are meaningless.
- *How* is data collected? What is actually being *measured*?
- What might be wrong/misleading/misinterpreted?
- Theory-driven? Data-Driven?

Data-driven decisions and conclusions require multiple skills. . . . . . and such skills are in high demand.

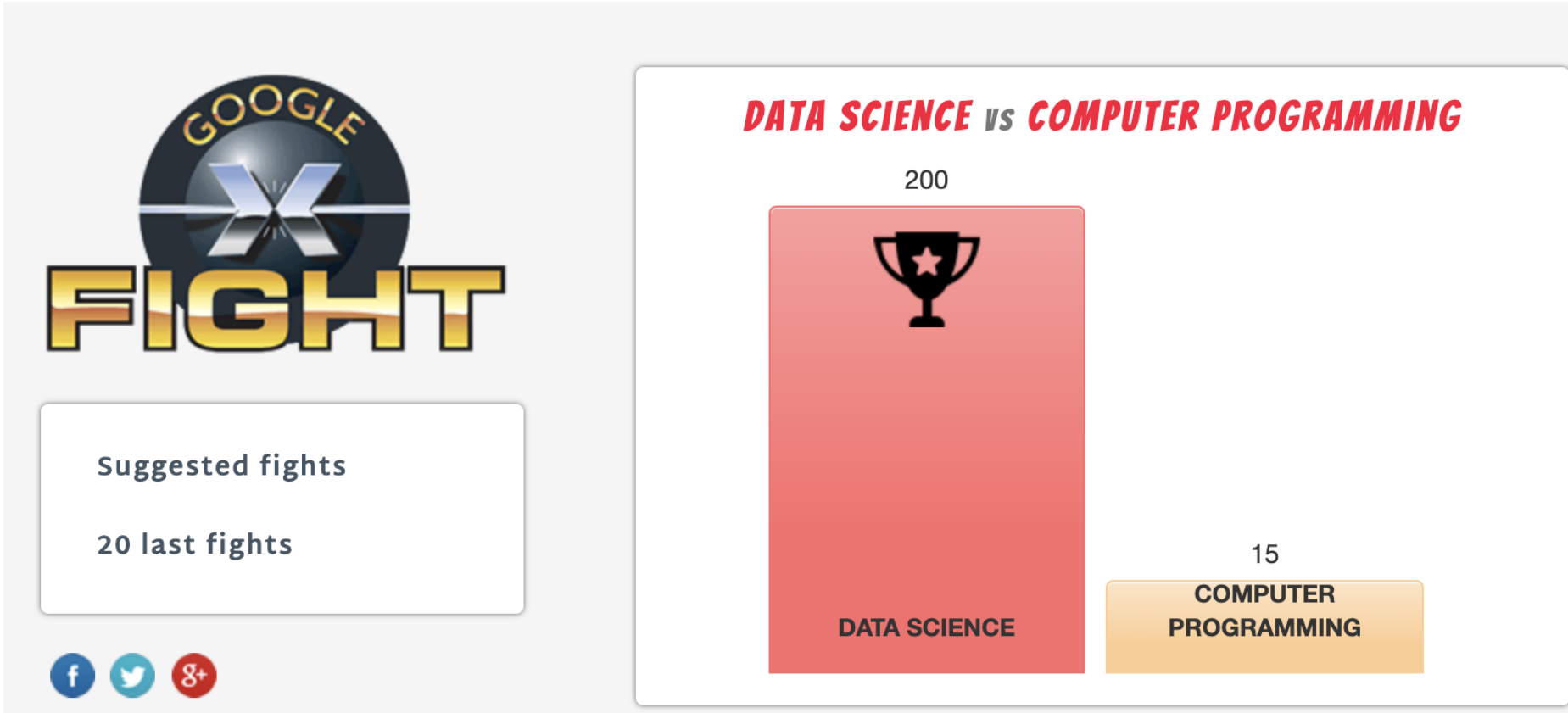
What is “Data Science”?

- No agreed upon definition.
- Set of skills: statistics, programming, communication, critical-thinking.

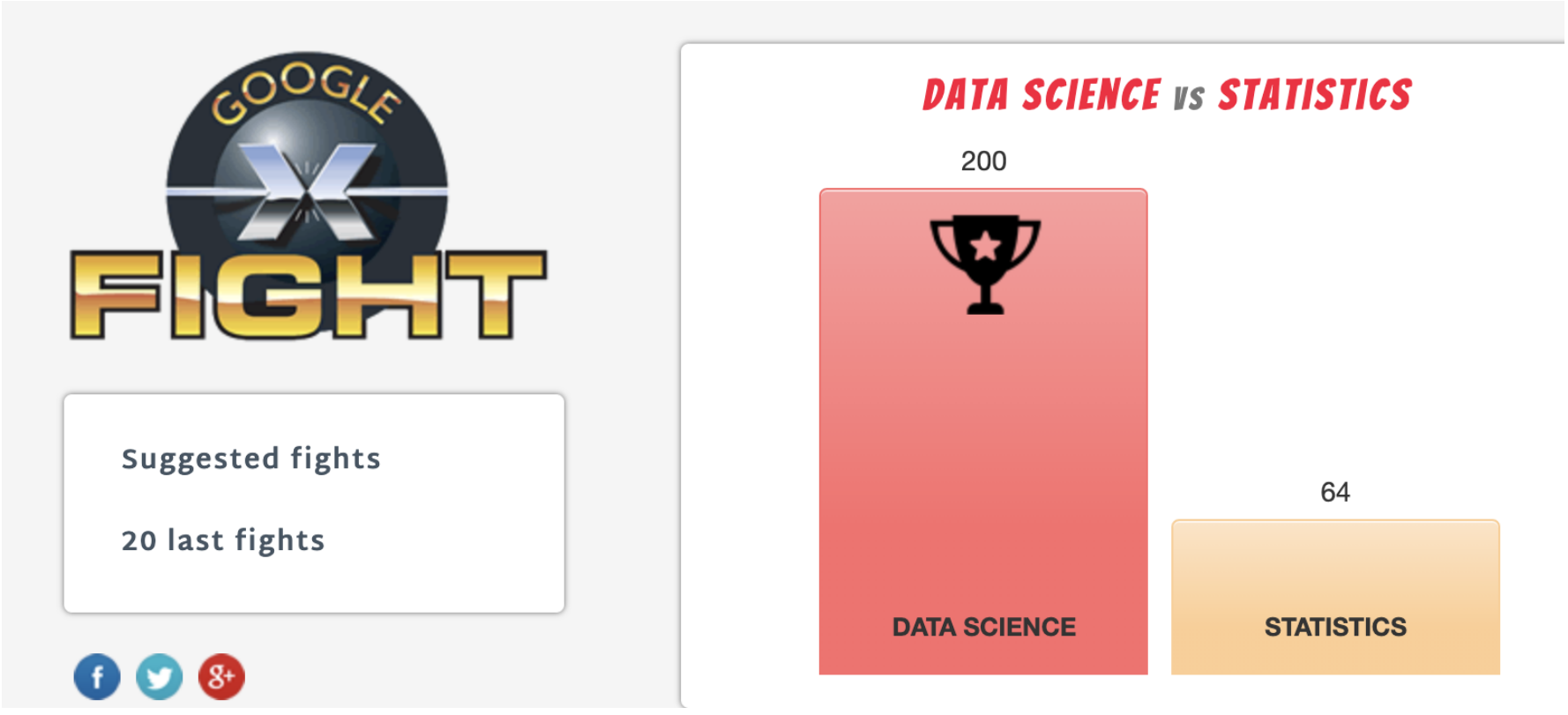
Everyone is doing it?



Everyone is doing it?



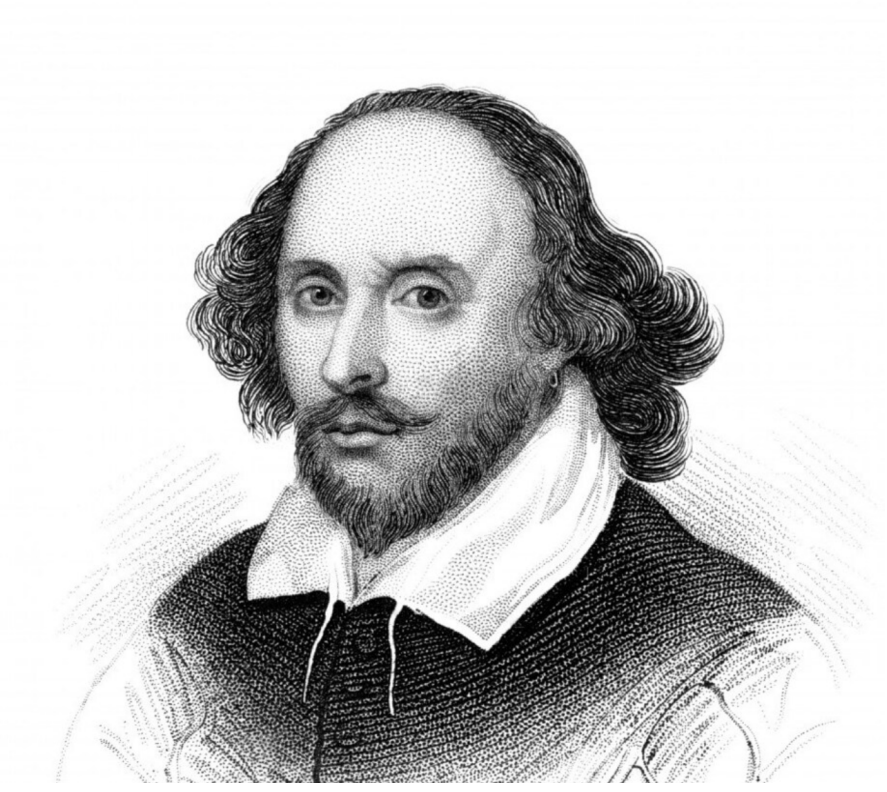
Everyone is doing it?



For nearly everything!

Is Shakespeare by Shakespeare?

Big debate about Shakespeare finally settled by big data: Marlowe gets his due



Predict Elections. . .

Finding Fame With a Prescient Call for Obama

By Stephanie Clifford

Nov. 9, 2008

At 9:46 p.m., blogging on his site FiveThirtyEight.com, Nate Silver called the presidential election for Barack Obama. The television networks followed suit about an hour and 15 minutes later after most polls in Western states closed.

Of course, Mr. Silver had a head start: he had forecast that Senator Obama would beat Senator John McCain back in March.

In an election season of unlikely outcomes, Mr. Silver, 30, is perhaps the most unlikely media star to emerge. A baseball statistician who began analyzing political polls only last year, he introduced his site, FiveThirtyEight.com, in March, where he used his own formula to predict federal and state results and run Election Day possibilities based on a host of factors.

. . . by using Twitter "sentiment"?

C. Aggregration

The winner was decided as the person having the higher Positive versus Total count ratio (PvT Ratio), calculated as

$$Ratio = |P|/|T \quad (1)$$

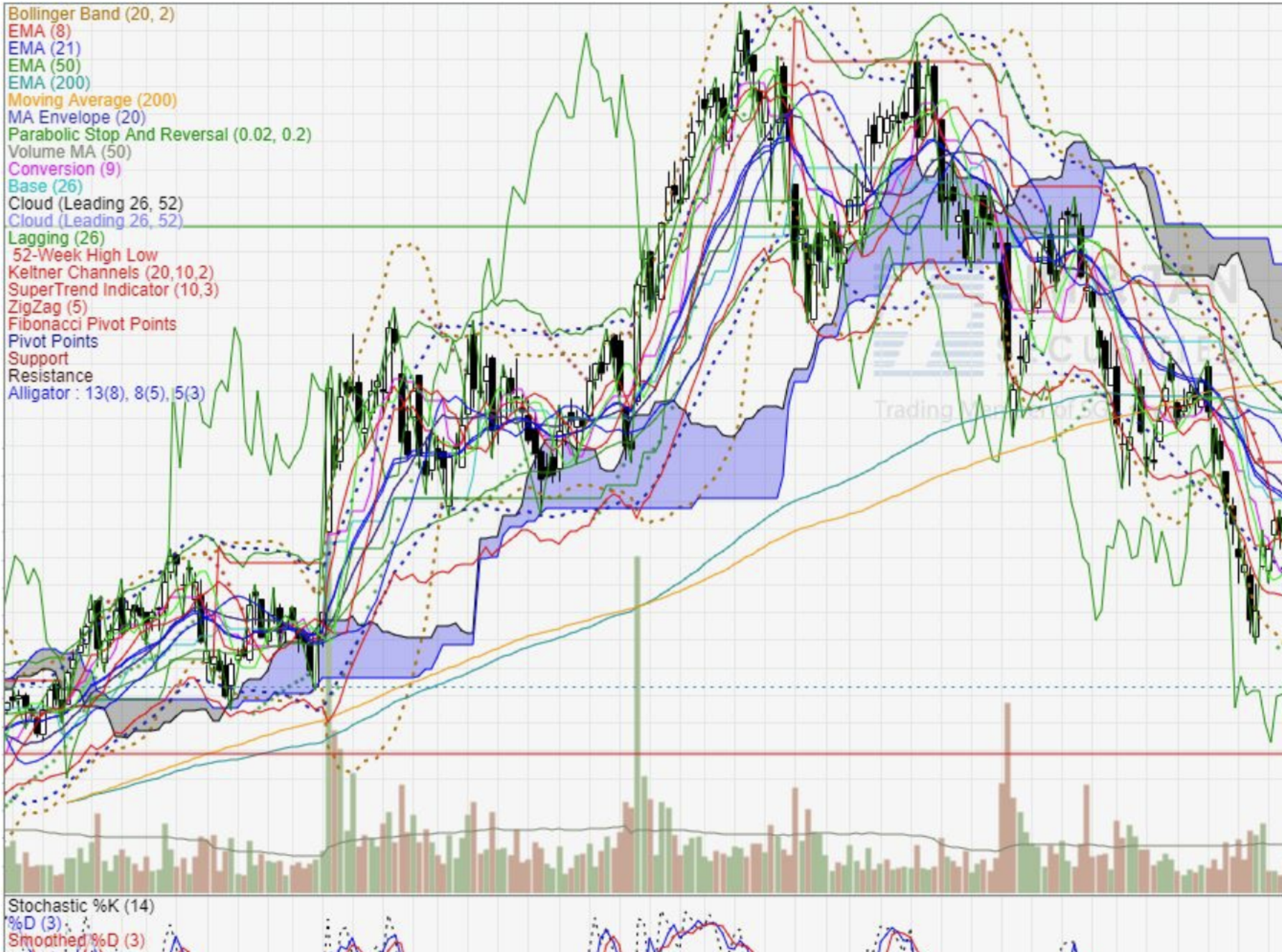
View Source

Here, P constitutes the tweets classified to be positive for the candidate (by the candidate's sentiment analyzer), T constitutes all the tweets classified as related to the candidate (by the entity classifier).

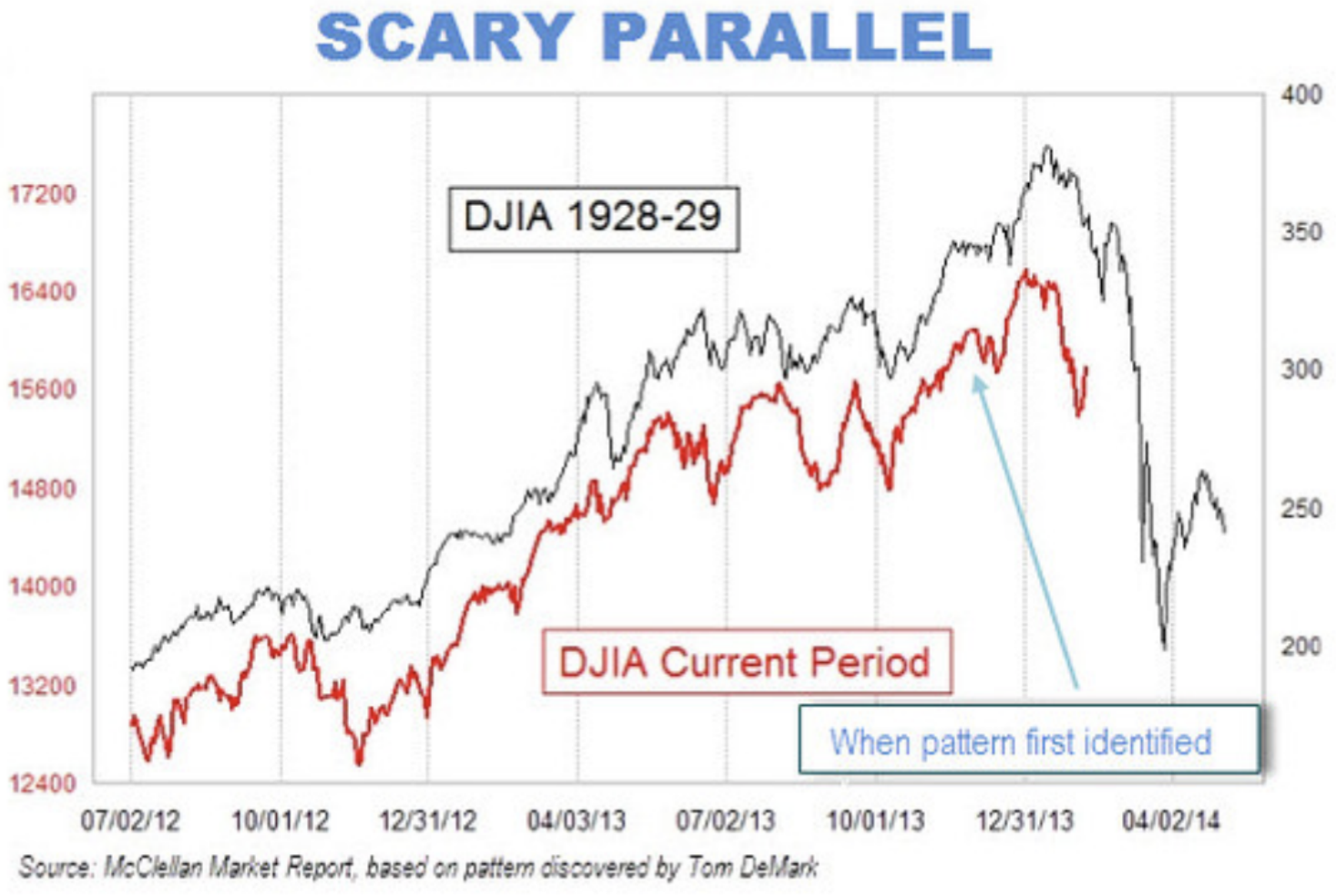
Table VI Pvt RATIO FOR canadidates

Candidate	*Positive*	*Negative*	*Total*	*PvT Ratio*
Donald Trump	2681	2170	4851	0.553
Hillary Clinton	1378	2410	3788	0.364

Perform "Technical" Analysis of stock markets. . .



. . . by using past trends "predict" future market conditions?



Identify markers for Breast Cancer Risk?

NIH NATIONAL CANCER INSTITUTE

NEWS RELEASES

NCI Press Release

BRCA Exchange aggregates data on thousands of BRCA variants to inform understanding of cancer risk

Posted: January 9, 2019

Contact: NCI Press Office
240-760-6600

A global resource that includes data on thousands of inherited variants in the *BRCA1* and *BRCA2* genes is available to the public. The BRCA Exchange was created through the BRCA Challenge, a long-term demonstration project initiated by the Global Alliance for Genomics and Health (GA4GH) to enhance sharing of *BRCA1* and *BRCA2* data. The resource, available through a website and a new smartphone app , allows clinicians to review expert classifications of variants in these major cancer predisposition genes as part of their individual assessment of complex questions related to cancer prevention, screening, and intervention for high-risk patients.

Credit: BRCA Challenge

The five-year BRCA Challenge project was funded in part by the National Cancer Institute (NCI), part of the National Institutes of Health, and through the Cancer Moonshot℠. A paper detailing the development of the BRCA Exchange was published January 8, 2019, in *PLOS Genetics*.

"This project has yielded a meta-analysis of *BRCA1* and *BRCA2* variants collected from multiple sources to understand how experts annotate specific mutations in the two genes," said Stephen J. Chanock, M.D., director of NCI's Division of Cancer Epidemiology and Genetics and lead author of the paper. "There's an urgent need for sharing data in cancer predisposition research. The BRCA Exchange is proof of principle that large-scale collaboration and data sharing can be achieved and can provide the latest and best quality information to

And the causes(s) of autism?

Early report

Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children

A J Wakefield, S H Murch, A Anthony, J Linnell, D M Casson, M Malik, M Berelowitz, A P Dhillon, M A Thomson, P Harvey, A Valentine, S E Davies, J A Walker-Smith

Summary

Background We investigated a consecutive series of children with chronic enterocolitis and regressive developmental disorder.

Methods 12 children (mean age 6 years [range 3–10], 11 boys) were referred to a paediatric gastroenterology unit with a history of normal development followed by loss of acquired skills, including language, together with diarrhoea and abdominal pain. Children underwent gastroenterological, neurological, and developmental assessment and review of developmental records. Ileocolonoscopy and biopsy sampling, magnetic-resonance imaging (MRI), electroencephalography (EEG), and lumbar puncture were done under sedation. Barium follow-through radiography was done where possible. Biochemical, haematological, and immunological profiles were examined.

Findings Onset of behavioural symptoms was associa

Introduction

We saw several children who, after a pe[r]iod of apparent normality, lost acquired skills, includ[i]ng com[mu]nication. They all had gastrointestinal [s]ymptoms, [i]ncluding abdominal pain, diarrhoea, and [b]loating and, i[n] some cases, food intolerance. We d[e]scribe [t]he clinical f[i]ndings, and gastrointestinal feature[s] of these chi[l]dren.

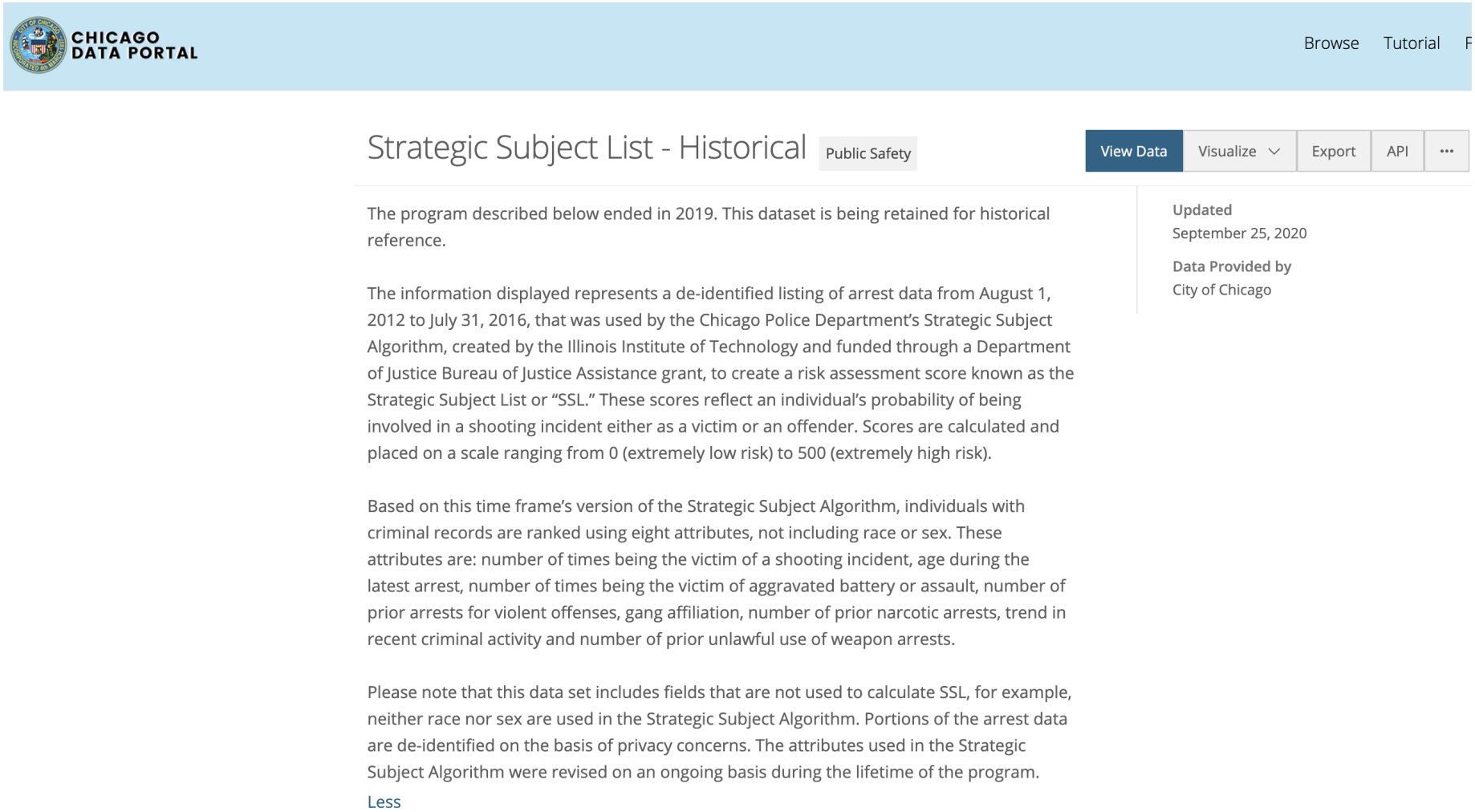
Patients and meth[o]ds

12 children, cons[e]cutively [a]dmitted to t[h]e department of paediatric gastr[o]enterology [w]ith a his[t]ory of a pervasive developmental [d]isorder with loss [o]f acquired skills and intestinal symptoms ([d]iarrh[o]ea, abdominal [p]ain, bloating and food intolerance), were inv[e]stigated. All children were admitted to the ward fo[r] 1 week, accomp[a]nied by their parents.

Clinical investigations

[W]e took histori[e]s including details of immunisations and e[x]posure to infect[i]us diseases, and assessed the children. In 11 cas[e]s the history [w]as obtained by the senior clinician (JW-S). Neuro[l]ogical and psychiatric assessments were done by

Name the 1400 people most likely to cause or experience gun violence in Chicago?



CHICAGO DATA PORTAL

Browse Tutorial F

Strategic Subject List - Historical

Public Safety

View Data | Visualize | Export | API | ...

The program described below ended in 2019. This dataset is being retained for historical reference.

The information displayed represents a de-identified listing of arrest data from August 1, 2012 to July 31, 2016, that was used by the Chicago Police Department's Strategic Subject Algorithm, created by the Illinois Institute of Technology and funded through a Department of Justice Bureau of Justice Assistance grant, to create a risk assessment score known as the Strategic Subject List or "SSL." These scores reflect an individual's probability of being involved in a shooting incident either as a victim or an offender. Scores are calculated and placed on a scale ranging from 0 (extremely low risk) to 500 (extremely high risk).

Based on this time frame's version of the Strategic Subject Algorithm, individuals with criminal records are ranked using eight attributes, not including race or sex. These attributes are: number of times being the victim of a shooting incident, age during the latest arrest, number of times being the victim of aggravated battery or assault, number of prior arrests for violent offenses, gang affiliation, number of prior narcotic arrests, trend in recent criminal activity and number of prior unlawful use of weapon arrests.

Please note that this data set includes fields that are not used to calculate SSL, for example, neither race nor sex are used in the Strategic Subject Algorithm. Portions of the arrest data are de-identified on the basis of privacy concerns. The attributes used in the Strategic Subject Algorithm were revised on an ongoing basis during the lifetime of the program.

Less

Updated
September 25, 2020

Data Provided by
City of Chicago

Spread misinformation?

Donald J. Trump Retweeted

Kevin McCullough @KMCRadio · 17h

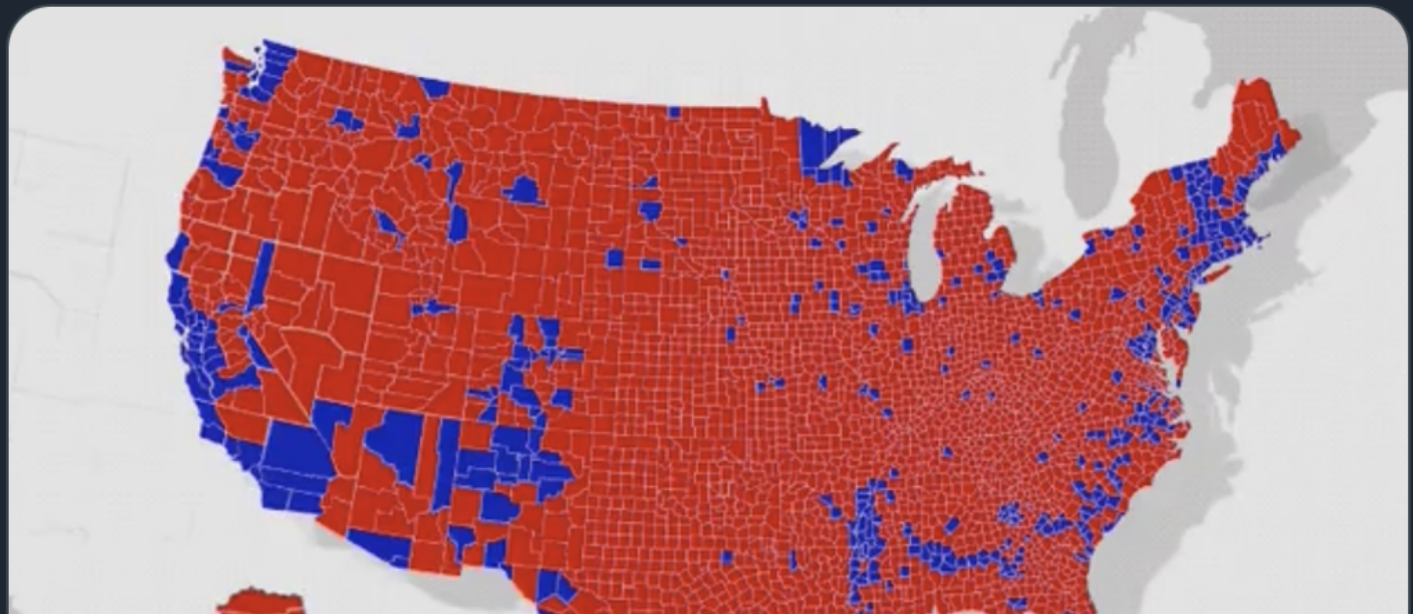
DIIVIDED NATION: 🚨🚨

America is 50 states.
Minus the states in question Trump won 25, Biden won 16.

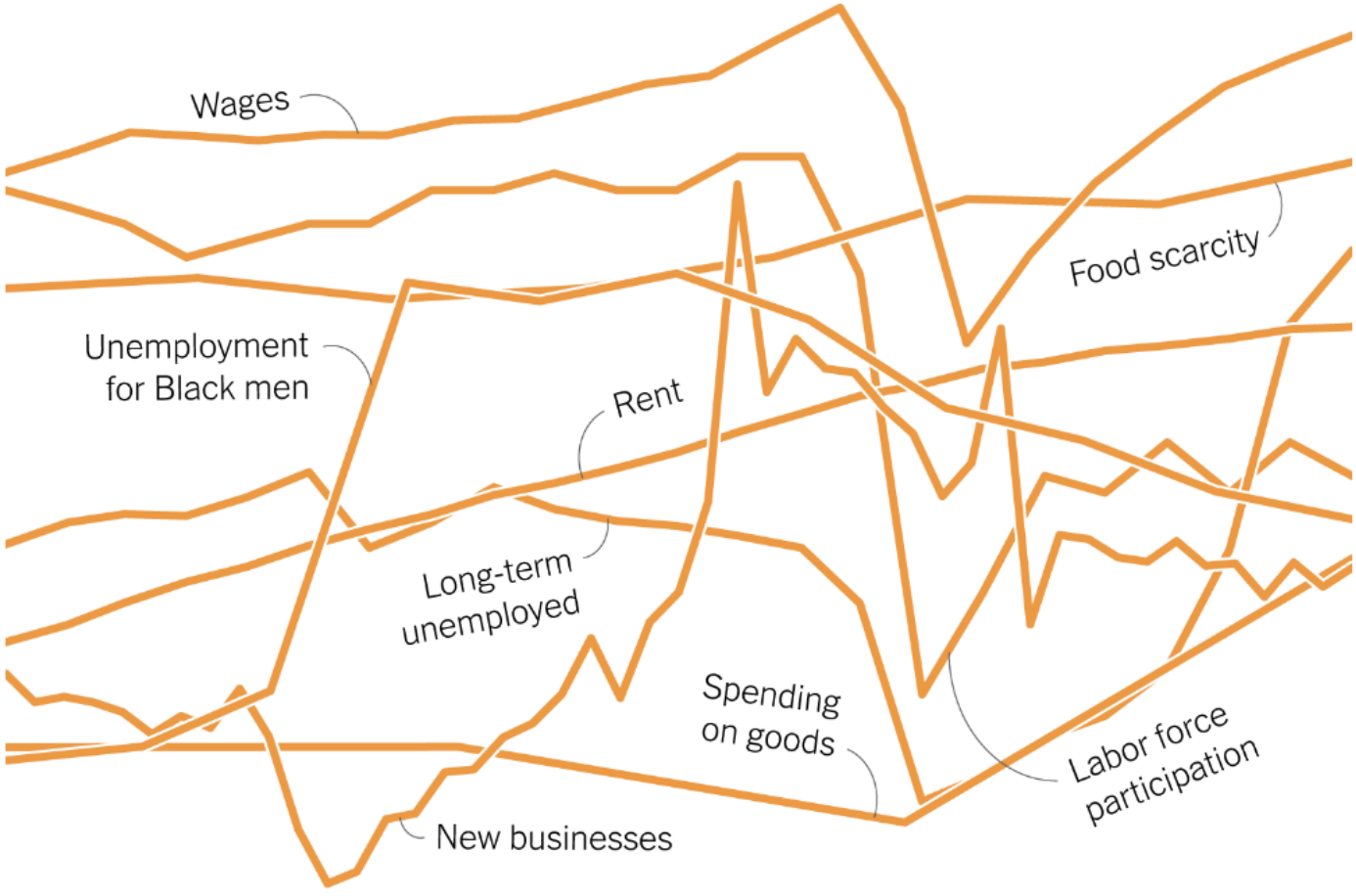
Those states house 2974 counties.
Even with the "votes in question" Trump won 2496, Biden on 477.

Trump won 84% of America, Biden "won" 16%.

#PickSixCheatBigJoeDid



Confuse the Public?



How the U.S. Economy Is Actually Doing, in 9 Charts

The unemployment rate doesn't tell the whole story, so we talked to a panel of economists to find out what other measures can shed light.

(Unintentionally?) Create Inequality?

Journal of Marketing Research
2019, Vol. 56(6) 960-980

DOI: 10.1177/0022243719852959
journals.sagepub.com/home/mrj
SAGE

When Words Sweat: Identifying Signals for Loan Default in the Text of Loan Applications

Oded Netzer, Alain Lemaire, and Michal Herzenstein

Abstract

The authors present empirical evidence that borrowers, consciously or not, leave traces of their intentions, circumstances, and personality traits in the text they write when applying for a loan. This textual information has a substantial and significant ability to predict whether borrowers will pay back the loan above and beyond the financial and demographic variables commonly used in models predicting default. The authors use text-mining and machine learning tools to automatically process and analyze the raw text in over 120,000 loan requests from Prosper, an online crowdfunding platform. Including in the predictive model the textual information in the loan significantly helps predict loan default and can have substantial financial implications. The authors find that loan requests written by defaulting borrowers are more likely to include words related to their family, mentions of God, the borrower's financial and general hardship, pleading lenders for help, and short-term-focused words. The authors further observe that defaulting loan requests are written in a manner consistent with the writing styles of extroverts and liars.

What we will do. . .

- How to think about data. (What are we doing and what can we do with data?)
- Work with data in a modern, widely-used, statistical programming language.
- How to analyze? How to present? [Reproducible!]
- Empower you to collect, and analyze new data to describe and understand the world

What we will NOT do...

- Talk (too much) about any one topic.
- Provide a comprehensive understanding of any subject.
- Cover the latest and greatest Data Science tools and methods (but `tidyverse`,`ggplot`,`tidymodels`)

What does “Introduction” mean?

- This is not “Foundations”...
- Give you experience in running code (copy, paste, & tweak)
- Not going through every function in detail...
- Or the math that justifies what we are doing and why...
- Focus on intuition and motivation.

Some of what we will talk about:

- Predicting elections in the United States using polls
- Looking at why some movies may make more money than others
- Predicting college admission and enrollment decisions
- Identifying "clusters" of voters (unsupervised learning)
- Analyzing twitter data (sentiment analysis)
- Predicting who wrote contested documents (Textual Forensics)

How to succeed:

Before Class

- Download the lecture notes (and data) prior to class.
- Put in a separate folder and see if you can load the data and knit the code. (Libraries?)
- Review the lecture notes to get a sense of where we are going. Figure out where you get lost in the notes.

How to succeed:

Before Class

- Download the lecture notes (and data) prior to class.
- Put in a separate folder and see if you can load the data and knit the code. (Libraries?)
- Review the lecture notes to get a sense of where we are going. Figure out where you get lost in the notes.

During Class:

- During class (if not before), try to predict what each code chunk will do before you run it.
- During class (and after), ask questions about points of confusion. If you have a question so do others!

How to succeed:

Before Class

- Download the lecture notes (and data) prior to class.
- Put in a separate folder and see if you can load the data and knit the code. (Libraries?)
- Review the lecture notes to get a sense of where we are going. Figure out where you get lost in the notes.

During Class:

- During class (if not before), try to predict what each code chunk will do before you run it.
- During class (and after), ask questions about points of confusion. If you have a question so do others!

After Class:

- Tweak the code to do something similar, but different. Can you do figure out how to use the code to do something different?
- Understand that you will not "think in code" – your goal is to give yourself tools to get the right answer.

You Can Do This