Lecture 1. The Wonderful World of Data Science

Josh Clinton & Will Doyle

8/19/2021

Data Science 1000: How Data Shape Our World

Today's Agenda

1. Meet the Instructor(s): Prof. Clinton: *josh.clinton@vanderbilt.edu*
   Prof. Doyle: *w.doyle@vanderbilt.edu*

Meet the Teaching Assistants: Mellissa Meisels: *mellissa.b.meisels@vanderbilt.edu*

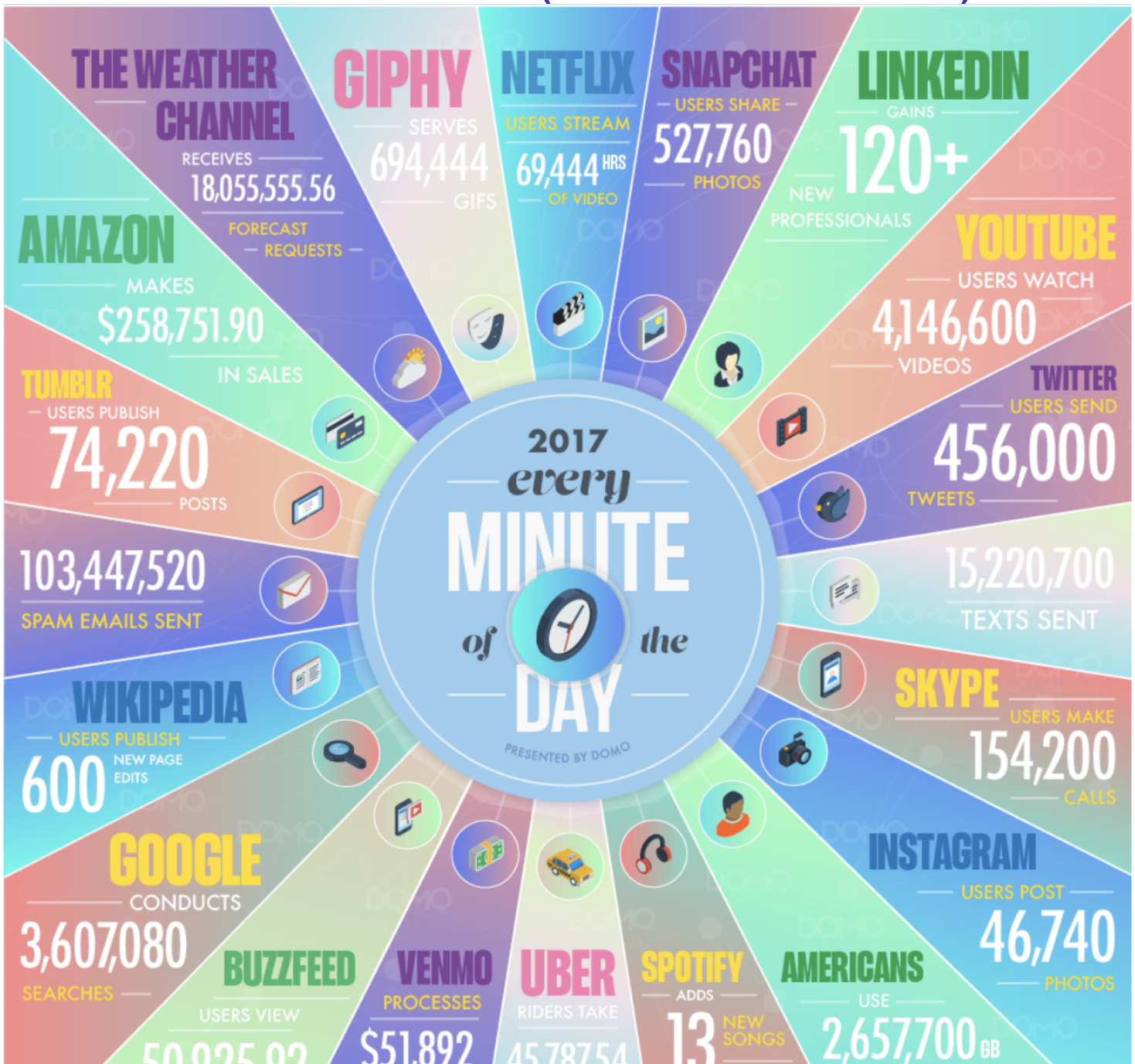
2. Course Objectives

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

3. Course Structure

- Lectures
- Workshops
- Evaluations

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



Course Motivation via “Tweets"

- “It is a capital mistake to theorize before one has data.” Sherlock Holmes
- “Data is the sword of the 21st century, those who wield it the samurai.” Jonathan Rosenberg, former Senior Vice President of Products at Google and current advisor to Alphabet Inc. CEO Larry Page
- “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

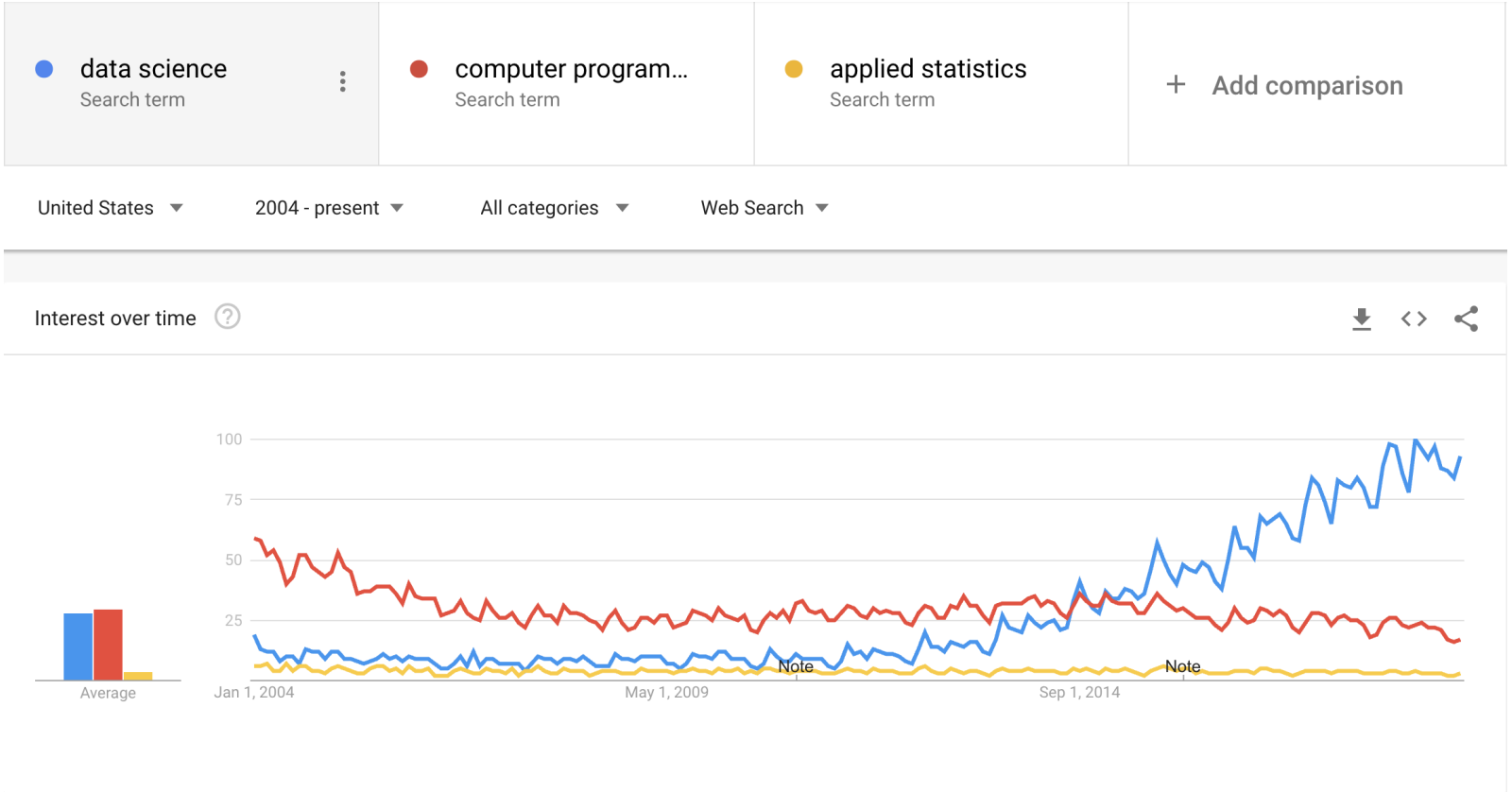
Data are more prevalent than ever!

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

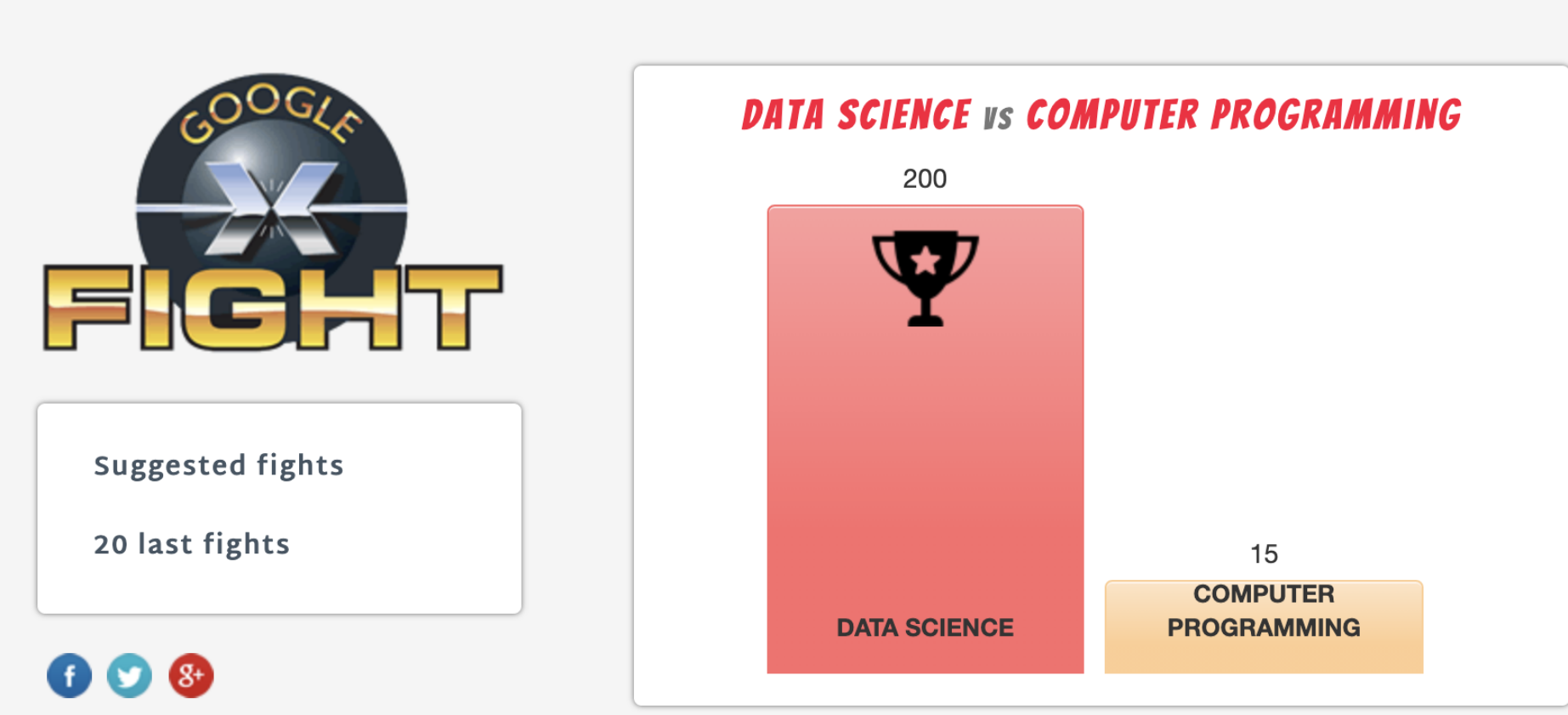
- Numbers without context are meaningless.
- *How* is data collected? What is actually being *measured*?
- What might be wrong/misleading/misinterpreted?
- Theory-driven? Data-Driven?

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

Everyone is doing it?



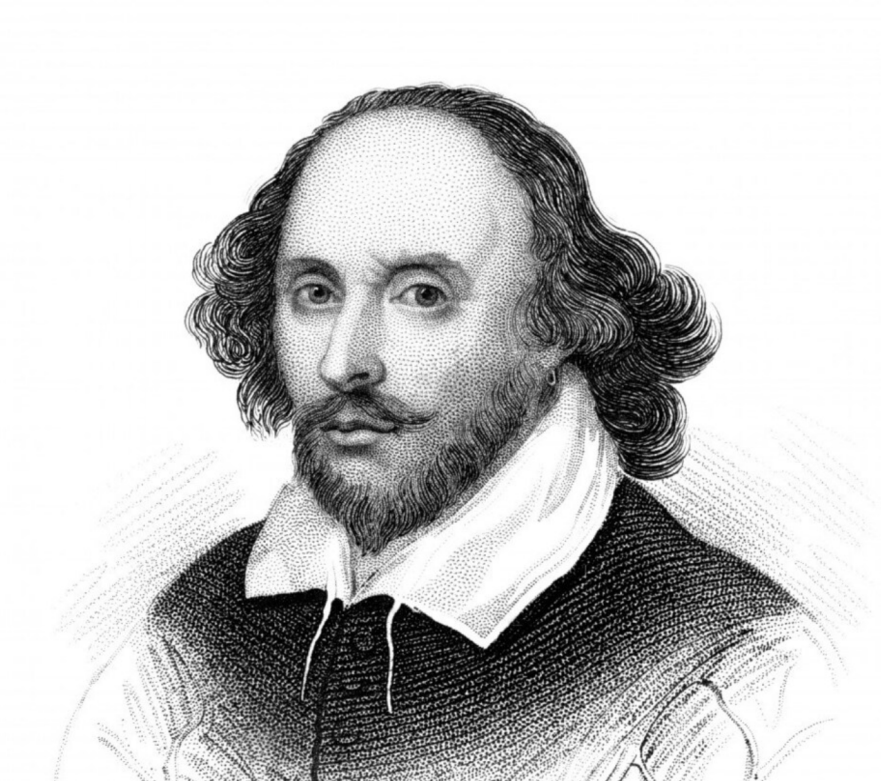
Everyone is doing it?

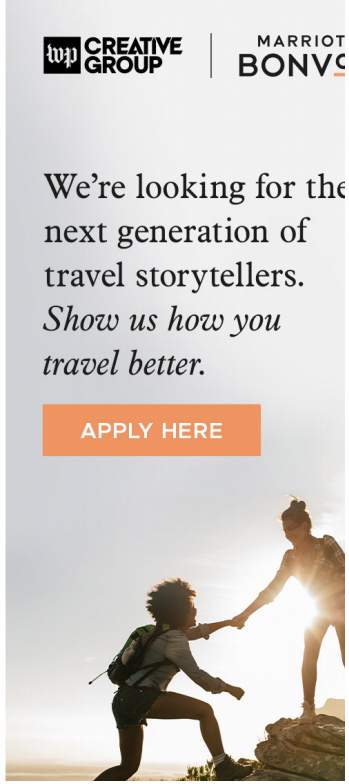


For nearly everything.

Is Shakespeare by Shakespeare?

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





Can Twitter "sentiment" to predict elections?

C. Aggregation

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

$$Ratio = |P|/|T \tag{1}$$

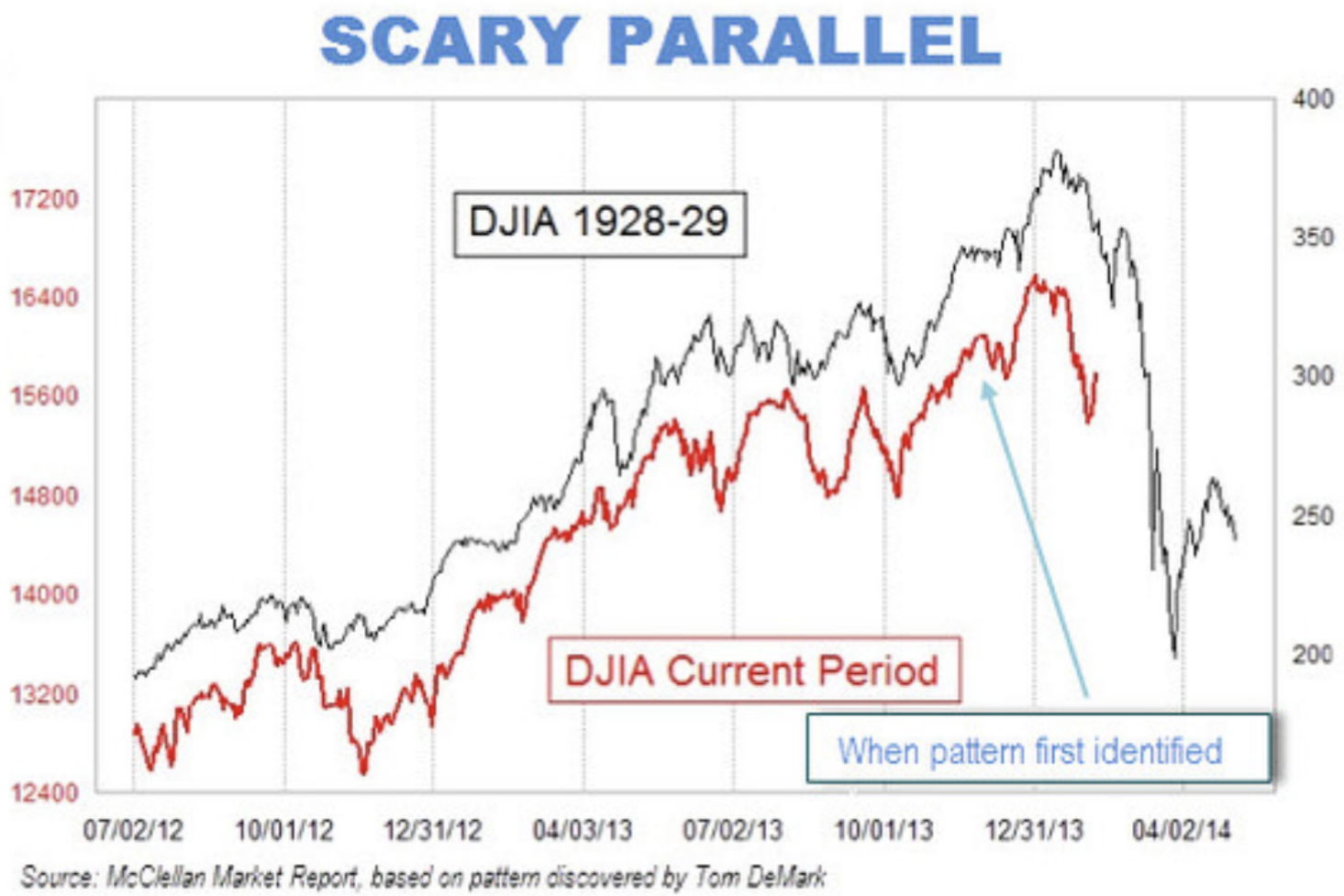
View Source

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

Can Trends "predict" future market conditions?



Can data diagnose 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 period of apparent normality, lost acquired skills, including communication. They all had gastrointestinal symptoms, including abdominal pain, diarrhoea, and bloating and, in some cases, food intolerance. We describe the clinical findings, and gastrointestinal features of these children.

Patients and methods

12 children, consecutively referred to the department of paediatric gastroenterology with a history of a pervasive developmental disorder with loss of acquired skills and intestinal symptoms (diarrhoea, abdominal pain, bloating and food intolerance), were investigated. All children were admitted to the ward for 1 week, accompanied by their parents.

Clinical investigations

We took histories including details of immunisations and exposure to infectious diseases, and assessed the children. In 11 cases the history was obtained by the senior clinician (JW-S). Neurological and psychiatric assessments were done by

RETRACTED

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

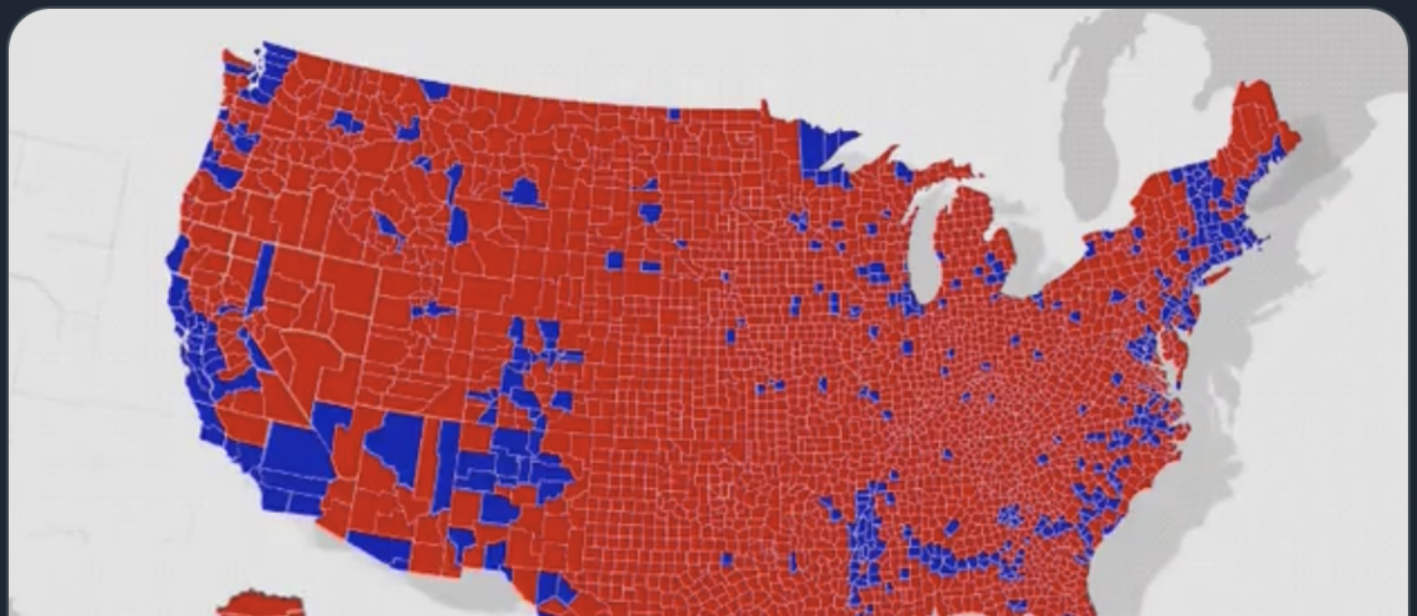
DIIVIDED NATION: 🚨🚨

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

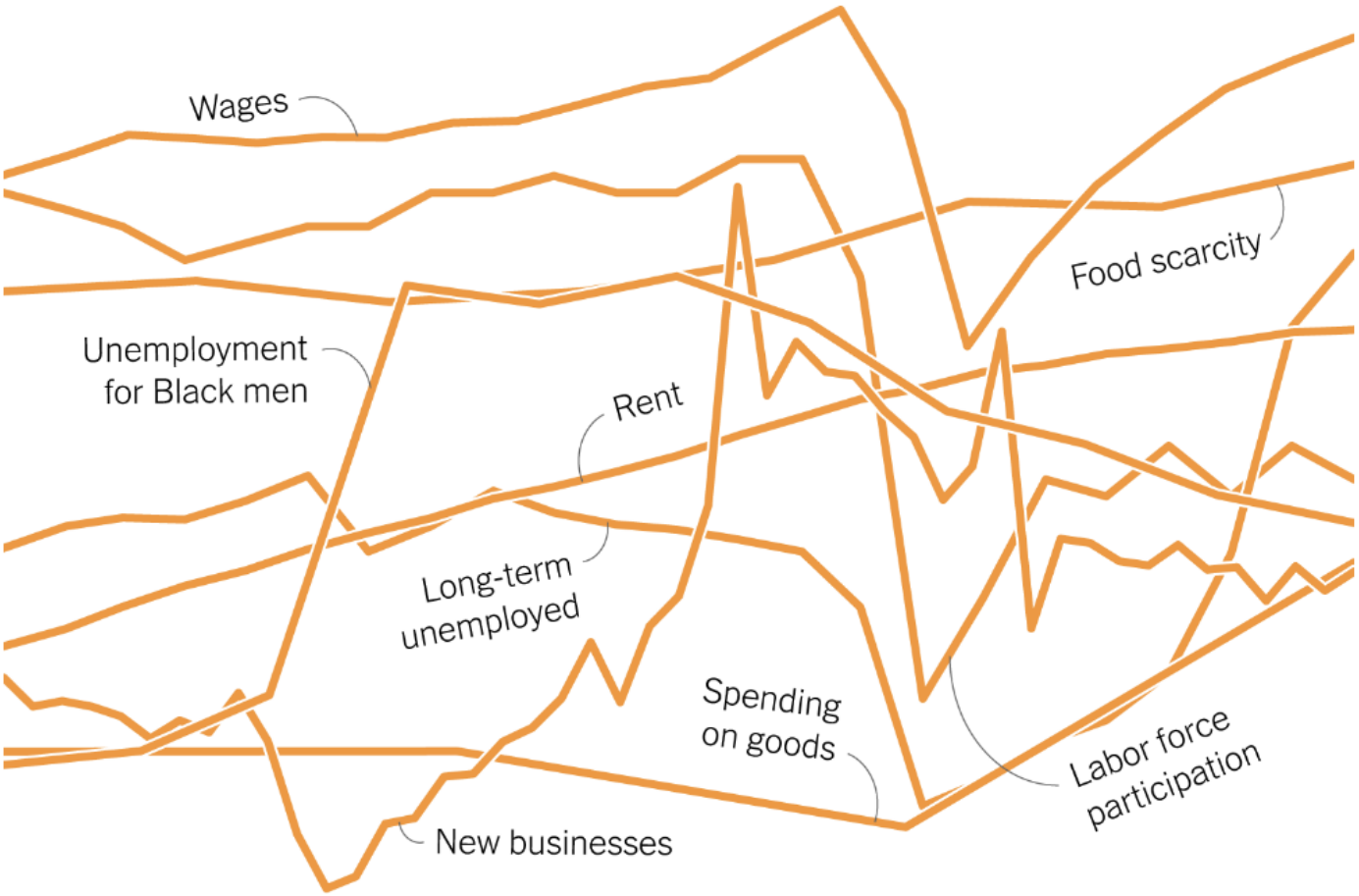
Those states house 2974 counties.
Even with the "votes in question" Trump won 2496, Biden on 477.

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.

Create Inequality?

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

Journal of Marketing Research
2019, Vol. 56(6) 960-980
© American Marketing Association 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0022243719852959
journals.sagepub.com/home/mrj
SAGE

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. (Theory-driven vs. ‘data-driven)
- How to 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.

Some of what we will talk about:

- Predicting Elections
- Predicting College Admission Decisions (and Enrollment)
- Visualizing Pandemic Data
- Analyzing Consumer Choice

Preliminaries

- Where to find things
- How class will work
- Expectations: attendance, assignments, guided exercises
- Communication
- Health and Safety
- Honor Code
- Technology