# How to use Data for Evil

## A Guide on how to Mislead with Data

Jake Rozran

November 14, 2024

RLEAnalytics


RLEAnalytics

# Who Am I?

- Owner & Data Scientist @ RLE Analytics
- Adjunct Professor of Data Science @ Villanova
- Masters from Villanova
- Undergrad from Drexel
- Maryland Native; DelCo Import
- Best of all: Dad

RLEAnalytics

# Manipulating people with data is nothing new

# Lying with Statistics

"Figures often beguile me, particularly when I have the arranging of them myself; in which case the remark attributed to [then British Prime Minister, Benjamin Disraeli] would often apply with justice and force: 'There are three kinds of lies: **lies, damned lies, and statistics**.'"

- Mark Twain, 1907

(emphasis and underline my own)

RLEAnalytics

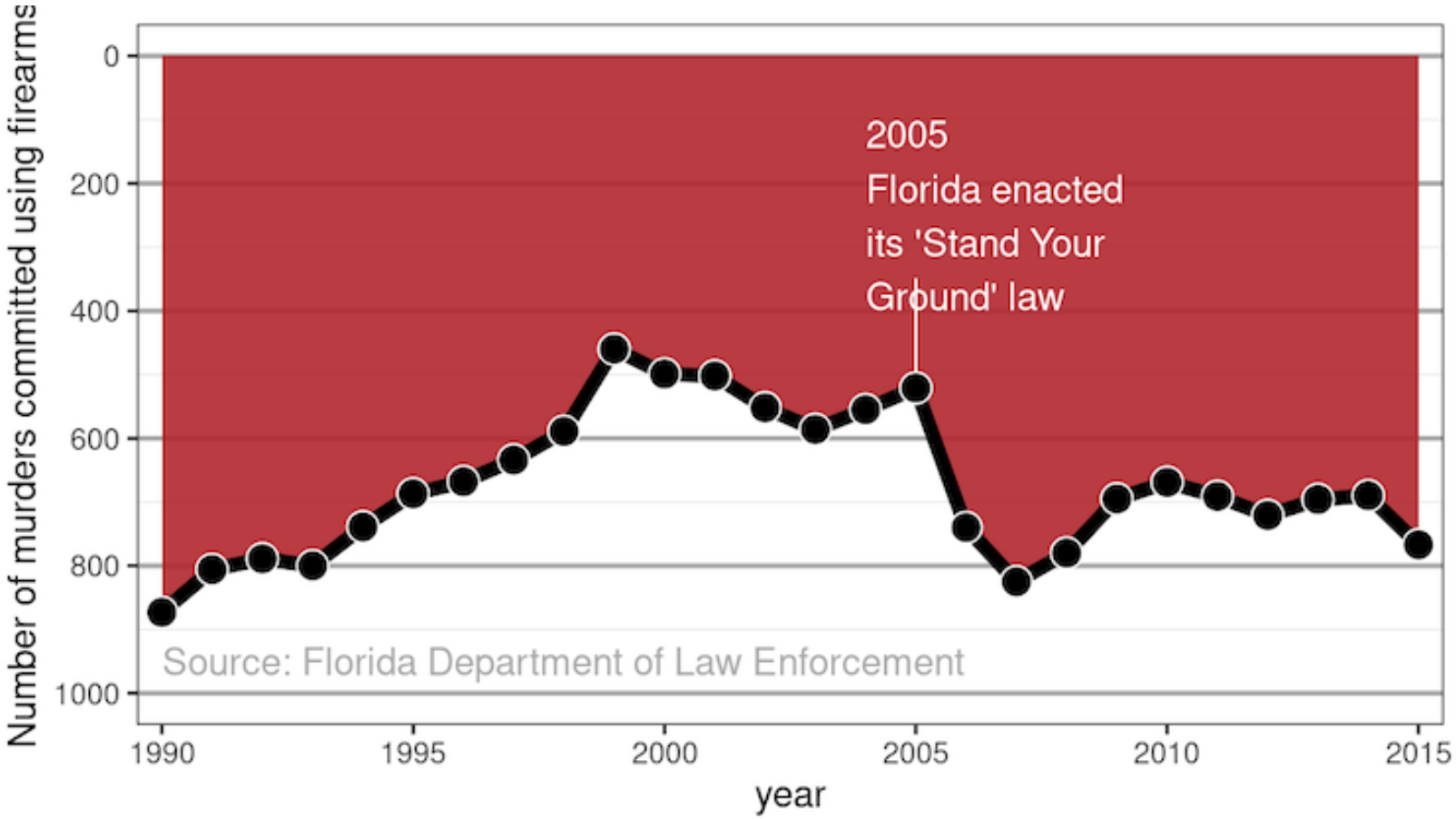
Alright, then let’s get lying…


RLEAnalytics

# How to do Data Visualizations [badly]!

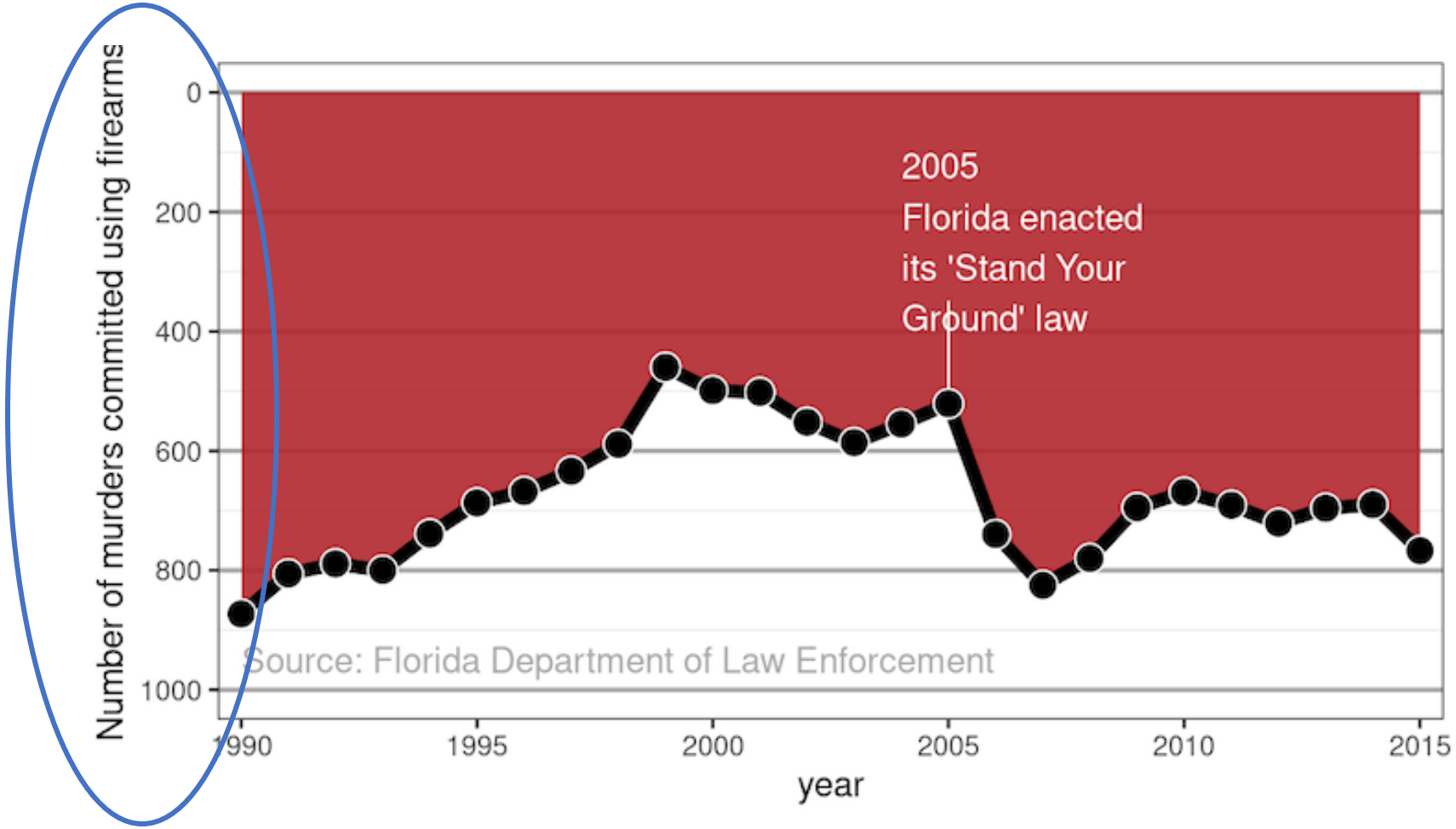
RLEAnalytics

# Obviously, the ‘Stand Your Ground’ law was a good thing if you don’t like murder committed with a firearm



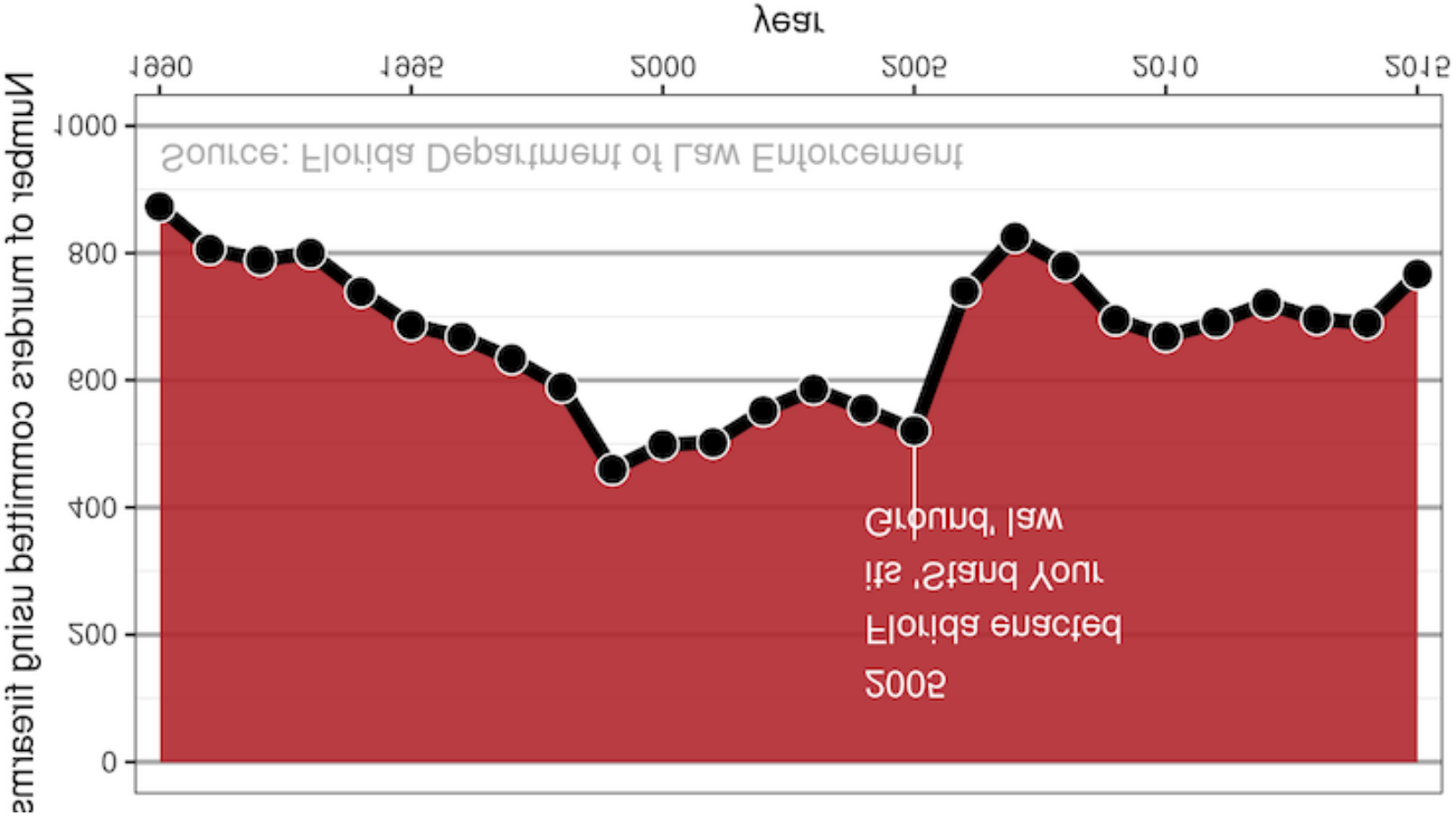

RLEAnalytics

# Lesson 1: Don’t like the trend? Just flip the axis!



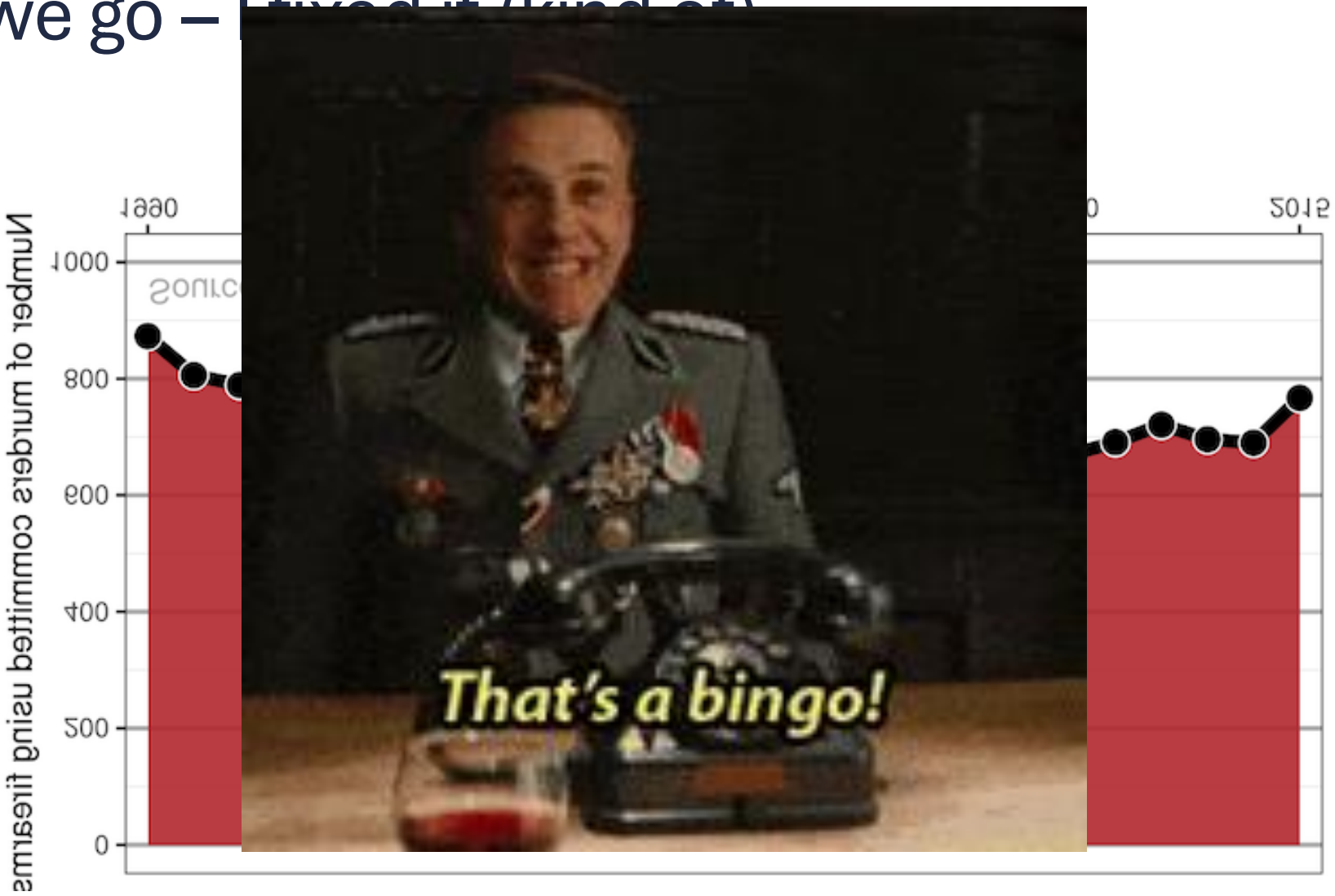

RLEAnalytics

# There we go – I fixed it (kind of)



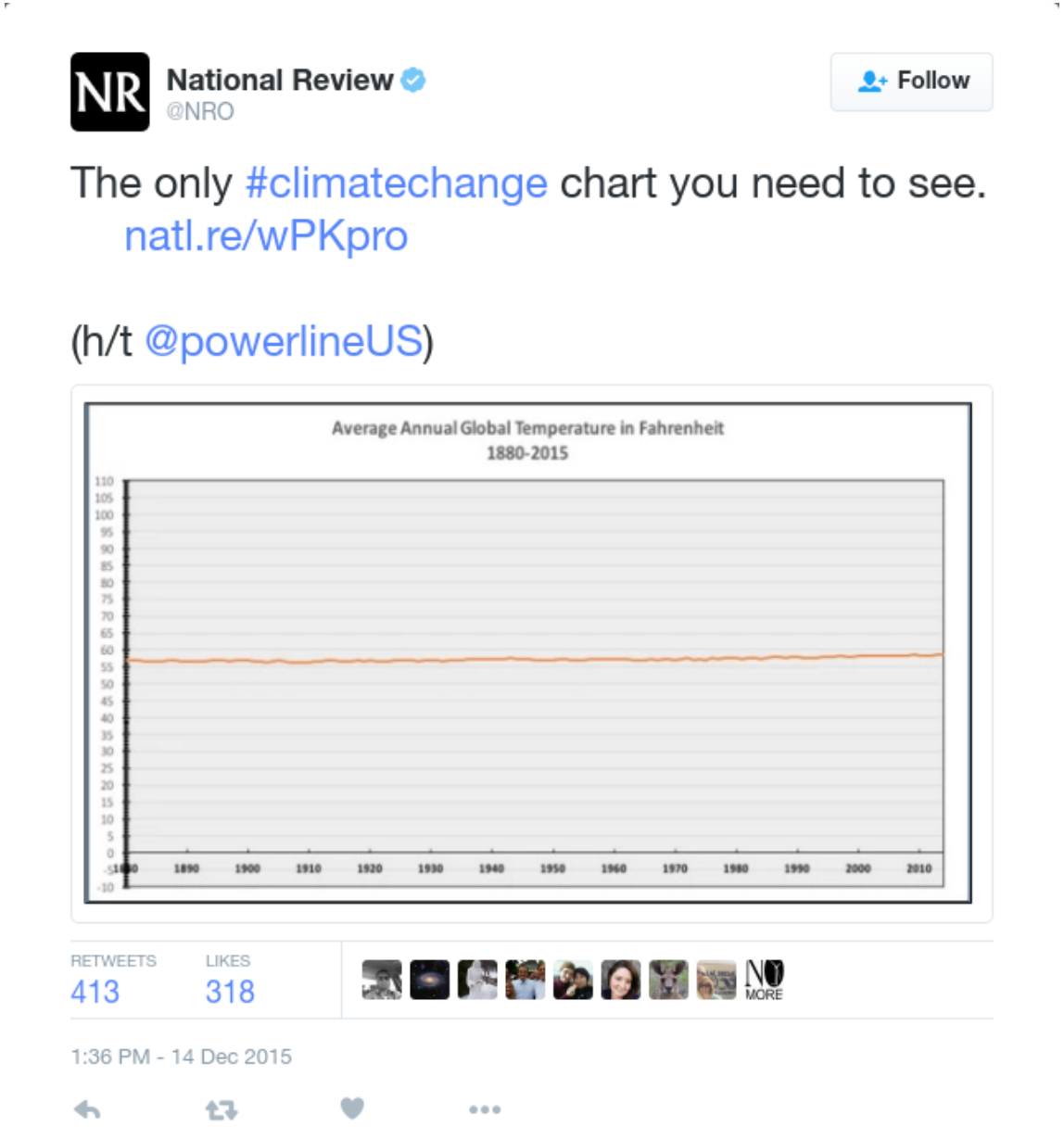

RLEAnalytics

# There we go – I fixed it (kind of)




RLEAnalytics

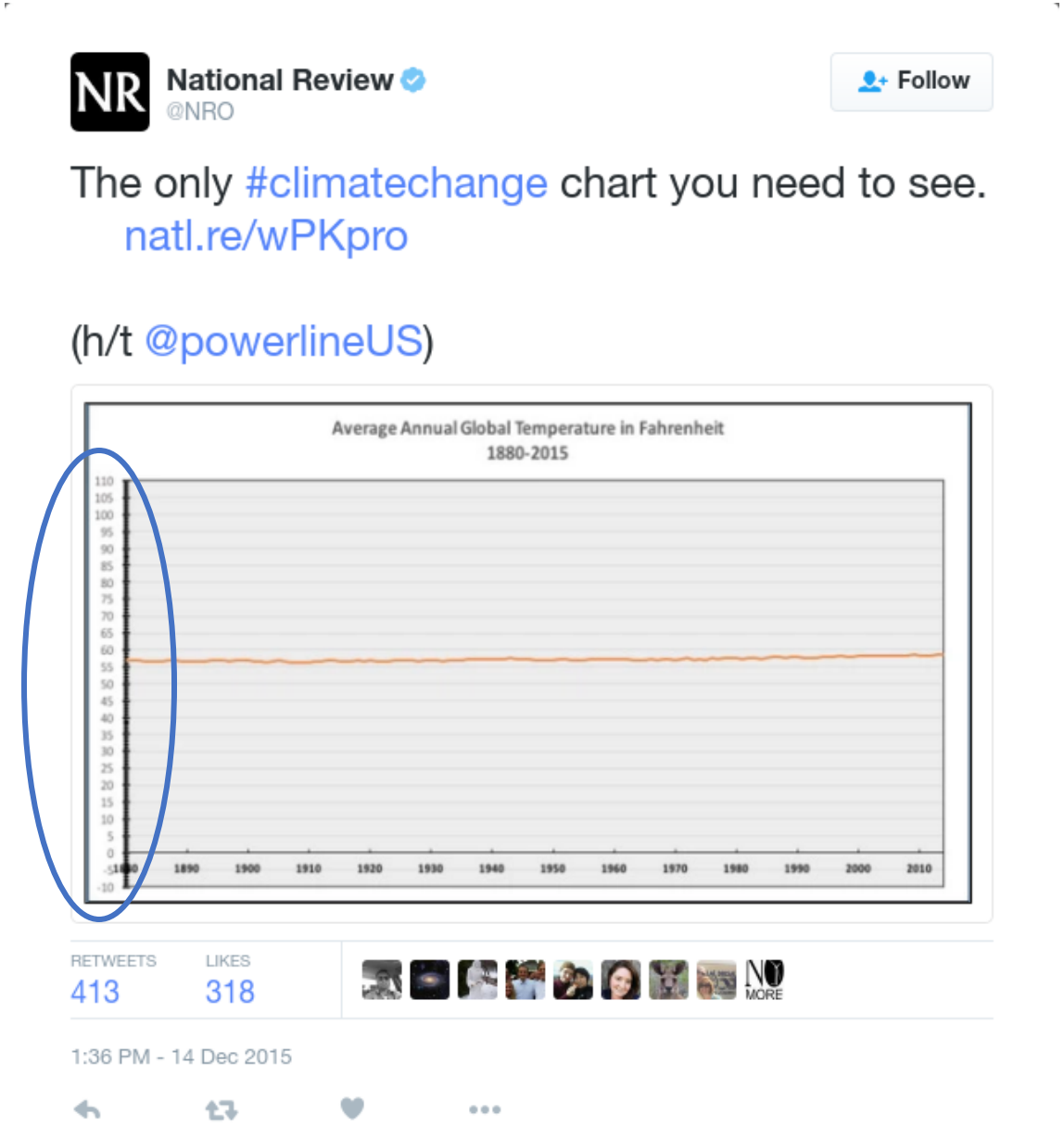
You people out there worried about climate change are foolish




RLEAnalytics

# Lesson 2: Don’t like the trend? Zoom out the y-axis…

by like, a **LOT**

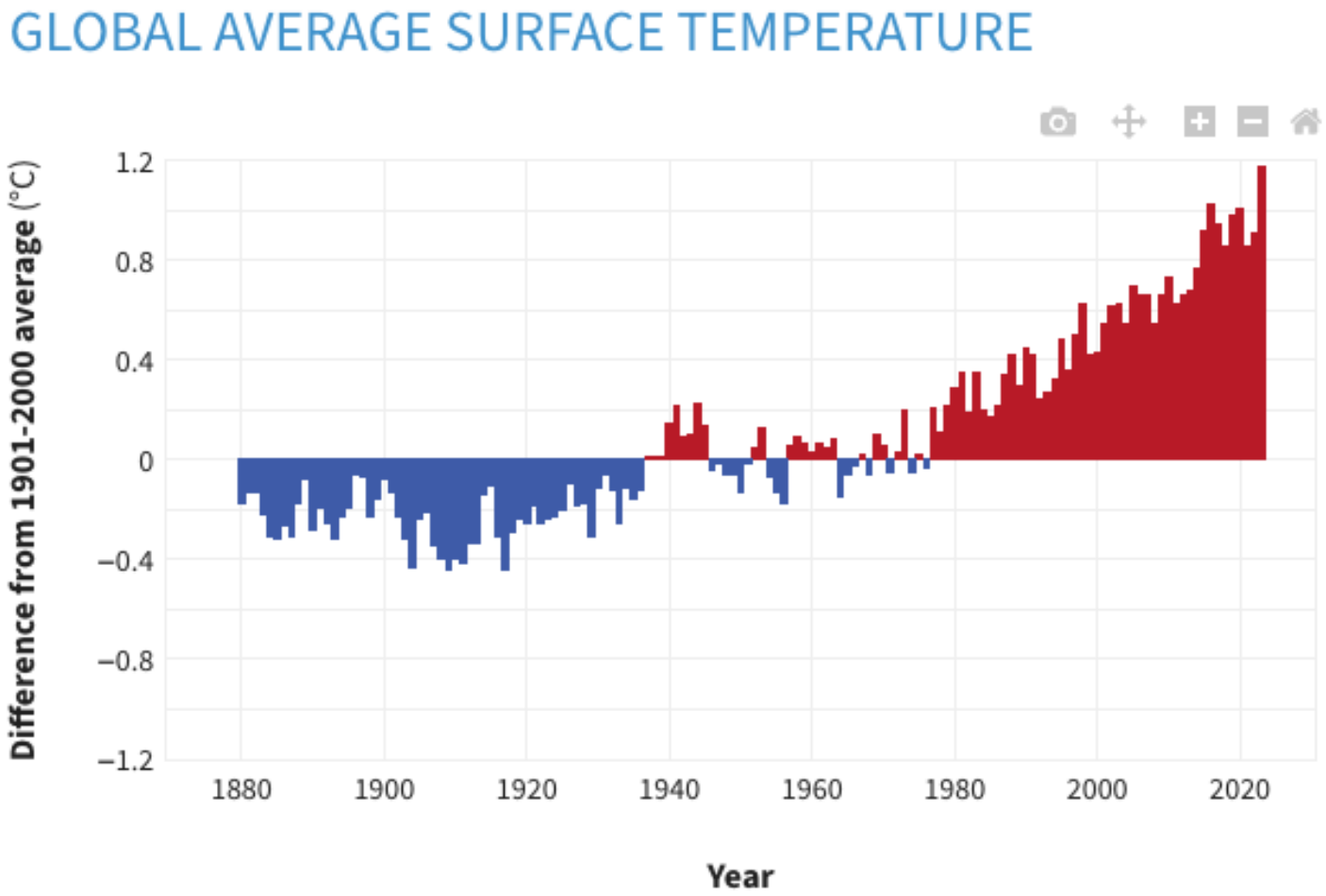



RLEAnalytics

Everything looks smaller from far away!


RLEAnalytics

# These fusspots out here putting out reasonable global warming visualizations…

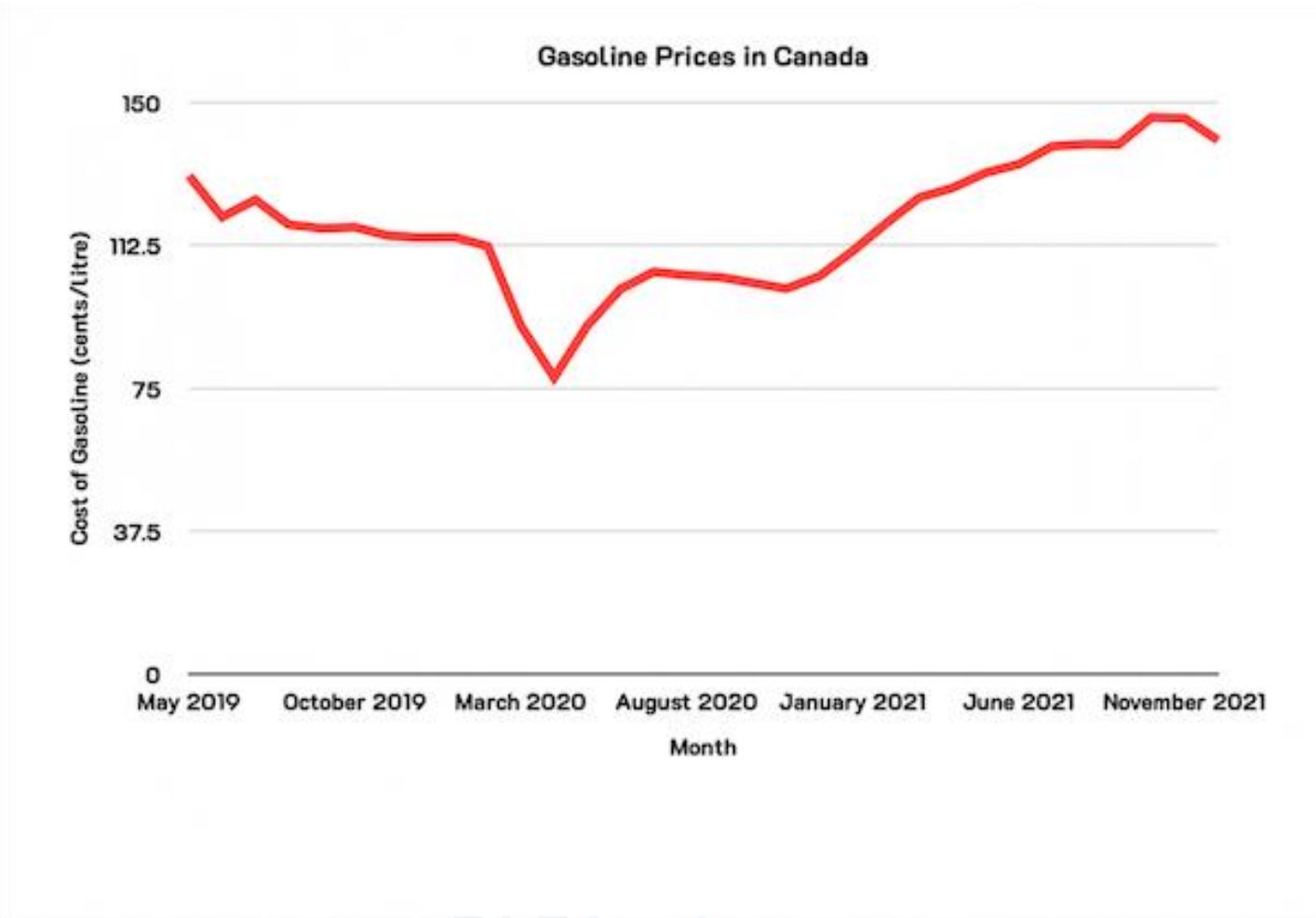



RLEAnalytics

# These fusspots out here putting out reasonable global warming visualizations...

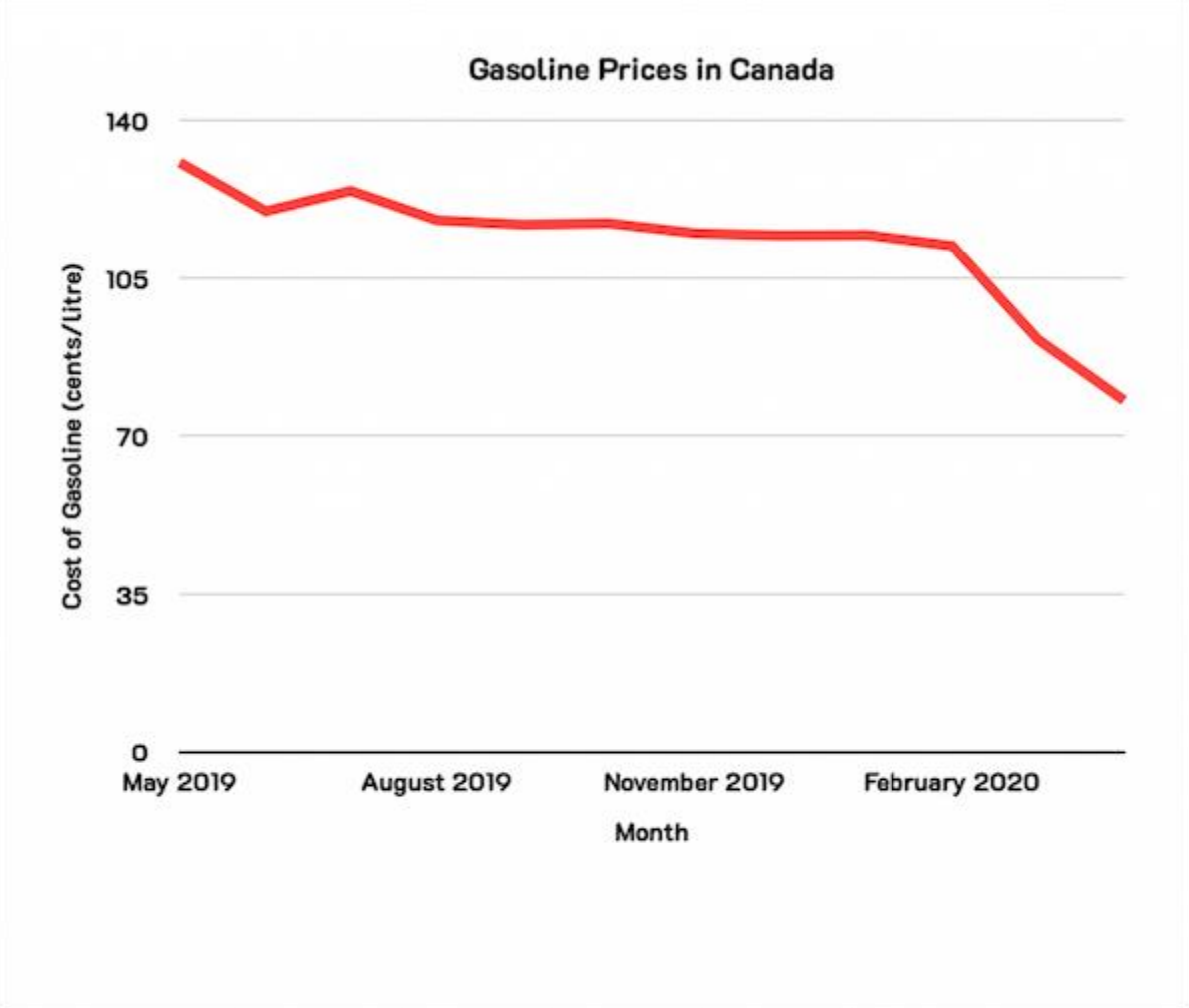

RLEAnalytics

# Flipping the axis or zooming out not your speed? Still have an inconvenient trend?



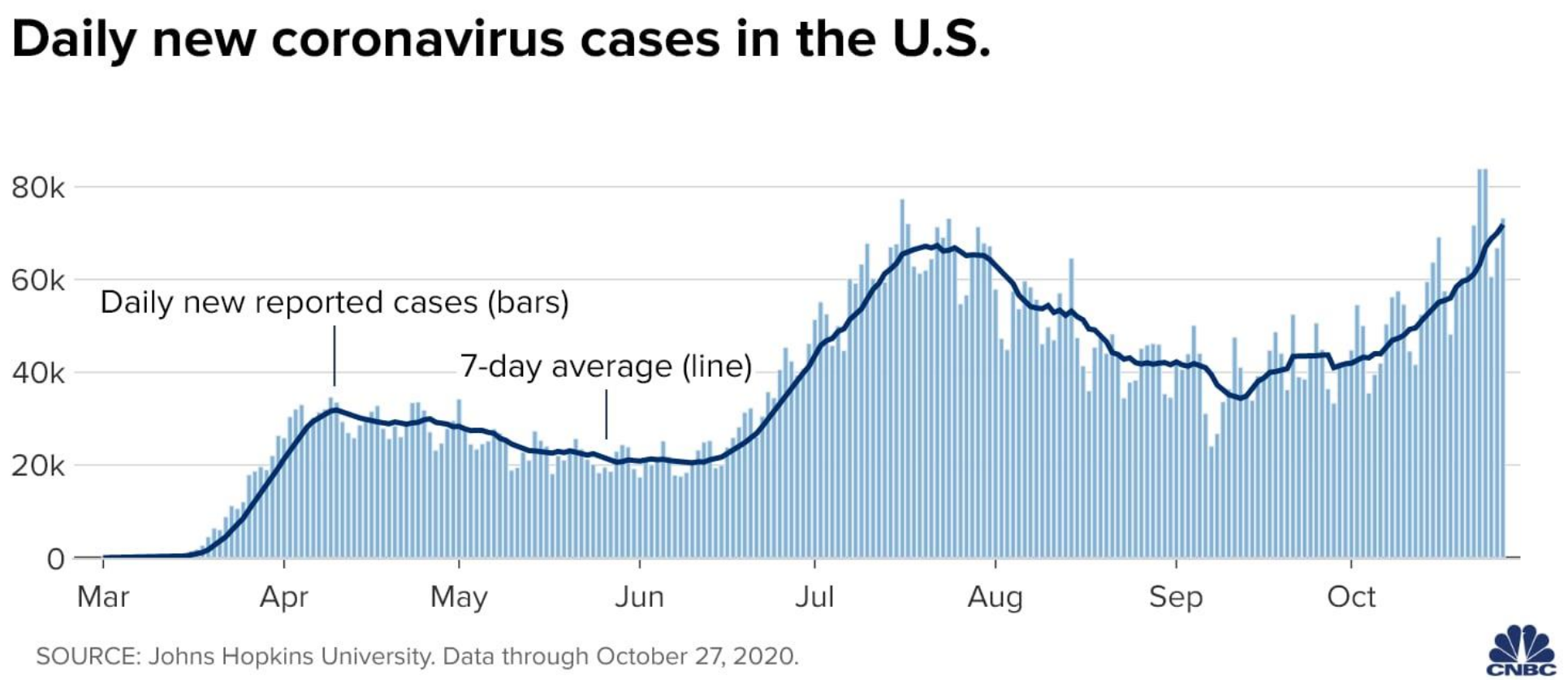

RLEAnalytics

Lesson 3: Omit the inconvenient data! Now we have the trend the way we like it.



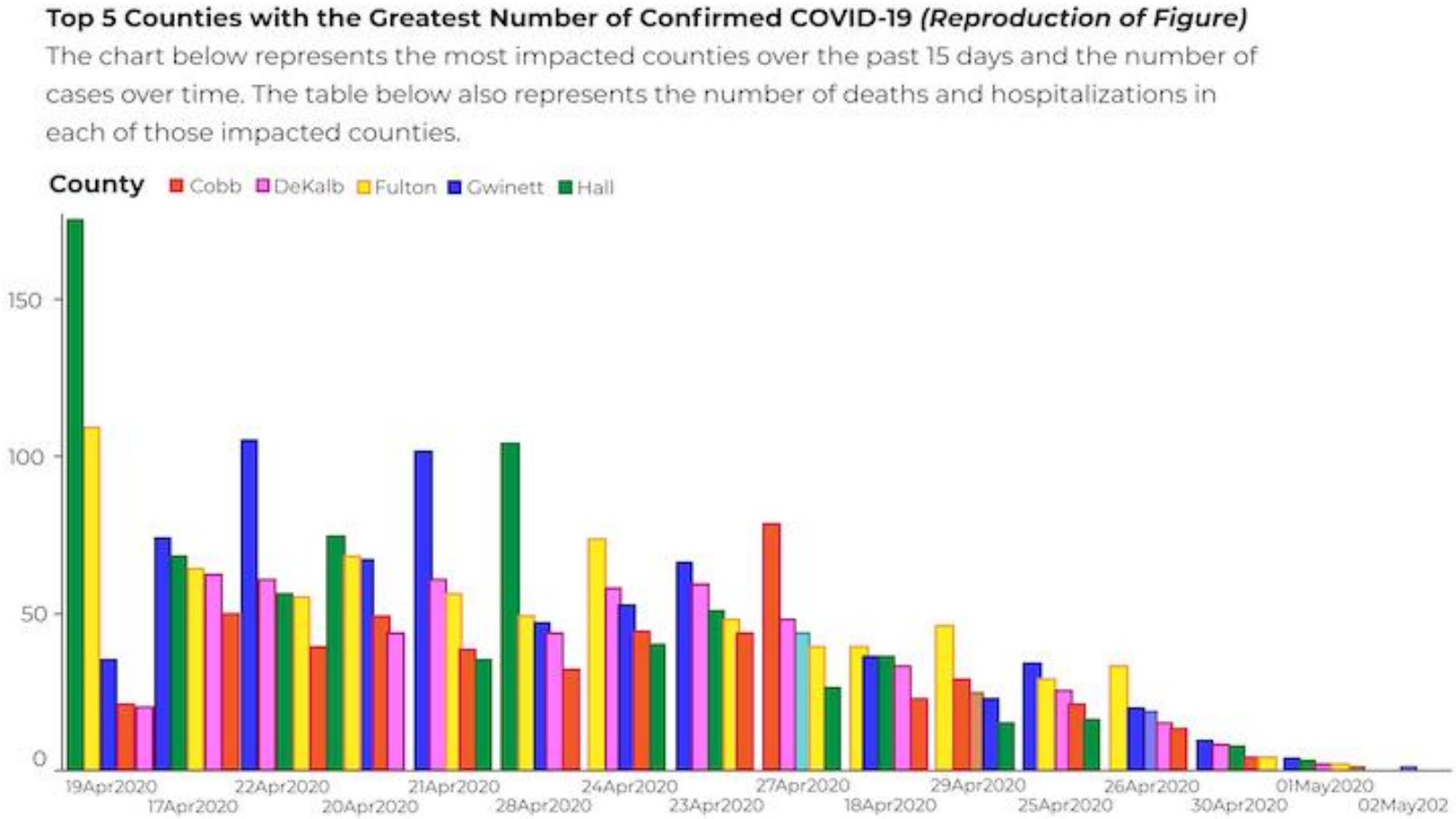

RLEAnalytics

# Getting desperate for good news and the last few tricks are played out?




RLEAnalytics

# Lesson 4: Forget about ordinality! Change the order of dates to make it obvious that your policies are working perfectly




RLEAnalytics

# Lesson 4: Forget about ordinality! Change the order of dates to make it obvious that your policies are working perfectly

**Top 5 Counties with the Greatest Number of Confirmed COVID-19** ***(Reproduction of Figure)***

The chart below represents the most impacted counties over the past 15 days and the number of cases over time. The table below also represents the number of deaths and hospitalizations in each of those impacted counties.





RLEAnalytics

Lesson 4: Forg
about ordinali
Change the or
of dates to ma
it obvious that
your policies a
working perfec

-19 *(Reproduction of Figure)*
ast 15 days and the number of
aths and hospitalizations in

19Apr2020 22Apr2
17Apr2020

20 01May2020
30Apr2020 02May202


RELAnalytics

# Beyond Data Viz! Let’s Build Some Predictive Models!

RLEAnalytics

Hey – You have a face! Let me use a picture of it to predict your sexual orientation?

RLEAnalytics

Hey – You have a face! Let me use a picture of it to predict your sexual orientation?

RLEAnalytics

# BECAUSE WE CAN, BIZNATCHES!

- Y. Wang and Kosinski (2018) used [Machine Learning] to build a classifier for sexual orientation based on pictures of people's faces
- The authors claim that if given five images of a person's face, their model would correctly predict the sexual orientation of 91% of men and 83% of women

# BECAUSE WE CAN, BIZNATCHES!

- Y. Wang
  classifi
  faces
- The aut
  model
  men ar


RLEAnalytics

# Lesson 5a: If you are going to be suspicious, go all in… don’t take time to call out potentials for harm

The authors highlight the potential harm that their work could do in their abstract:

*“Additionally, given that companies and governments are increasingly using computer vision algorithms to detect people’s intimate traits, our findings expose a threat to the privacy and safety of gay men and women.”*

RLEAnalytics

# Lesson 3 Revisited: Definitely only show the favorable statistics

- The authors claim that if given five images of a person's face, their model would correctly predict the sexual orientation of 91% of men and 83% of women
- I wonder how these statistics look if we segmented between heterosexual and homosexual individuals? Gallup estimates that 7.2% of adults identify as LGBT in 2022 – very imbalanced data.
- If we were to just predict everyone is heterosexual, then we'd be 92.8% correct… Better than their model (oops). *Quick shout out to a student of mine, Joey C., who highlighted this in our discussion last week.*

RLEAnalytics

# Lesson 3 Revisited: Definitely only show the favorable statistics

- The authors claim that if given five images of a person's face, their model would correctly predict the sexual orientation of 91% of men and 83% of women
- I wonder how these statistics look if we segmented between heterosexual and homosexual individuals? Gallup estimates that 7.2% of adults identify as LGBT in 2022 – very imbalanced data.
- If we were to just predict everyone is heterosexual, then we'd be 92.8% correct… Better than their model (oops). *Quick shout out to a student of mine, Joey C., who highlighted this in our discussion last week.*

RLEAnalytics

# Lesson 3 Revisited: Definitely only show the favorable statistics

- The authors claim that if given five images of a person's face, their model would correctly predict the sexual orientation of 91% of men and 83% of women
- I wonder how these statistics look if we segmented between heterosexual and homosexual individuals? Gallup estimates that 7.2% of adults identify as LGBT in 2022 – very imbalanced data.
- If we were to just predict everyone is heterosexual, then we'd be 92.8% correct… Better than their model (oops). *Quick shout out to a student of mine, Joey C., who highlighted this in our discussion last week*.

RLEAnalytics

# Lesson 5b: Ignore the people and their “facts,” too…

A subsequent article in The New Yorker also notes that:

*“the study consisted entirely of white faces, but only because the dating site had served up too few faces of color to provide for meaningful analysis.”*

RLEAnalytics

# Moving beyond faces – let's build models to tackle criminal recidivism!

RLEAnalytics

# An Algorithm can do better than judges, who get hungry

There is a criminal recidivism algorithm used in several states

The algorithm returns predictions about how likely a criminal is to commit another crime based on a survey of 137 questions

RLEAnalytics

# None of the questions, however, ask about race, so why then does ProPublica think the algorithm is biased?

“Black defendants were still 77 percent more likely to be pegged as at higher risk of committing a future violent crime and 45 percent more likely to be predicted to commit a future crime of any kind.”

# Final Lesson: Figure out as many **proxy variables** as possible to hide the fact that you are using race and other protected variables in your algorithm

- Consider that one of the survey questions is “was one of your parents ever sent to jail or prison?”
- Who is more likely to answer this question “yes?” Why?

RLEAnalytics


Remember: These Lessons will Ensure your Success!


# Thank you!


- Jake Rozran
- jake@rleanalytics.com
- rleanalytics.com
- https://jakelearnsdatascience.com


RLEAnalytics