Big Data, information and political campaigns: an application to the 2016 US Presidential Election

Presentation largely based on

- *Politics and Big Data: Nowcasting and Forecasting Elections with Social Media*, 2017

Preamble

Big Data are those labeled, for strange reasons, with the capitalized "**Big**". Nevertheless, they are still "**Data**" (with also a Capital letter!)

Therefore…**good statistical techniques** are required in order to extract meaningful results from such sources

What Big Data are not

Big Data are not **just** a data collection with a very large-N

That is, a very *large survey of citizen participation* cross-nationally is **not**, strictly speaking, Big Data

What Big Data are : 3 main attributes (at the same time!)

"*volume*" - data exceed the capacity of traditional computing methods to store and process them

"*frequency*" - data come from streams or complex event processing, i.e., size per unit of time matters

"*unpredictability*" - data come in the many different forms, they are raw, messy, **unstructured**, not ready for processing, and so on

Their origin: *administrative data*; *transaction data*; ***social media data***

Let's focus on the latter ones

Why studying Big Data?

To make a long answer short…

…Always more (and more) data are available out there in **time** and **space!**

Can we really ignore this?

Big Data Analytics

Dan Ariely

6 gennaio 2013 ·

Segui

Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it...

The main approaches (with a specific emphasis on electoral campaigns)

Main approach	Sub-approaches
Computational	Volume data
	Endorsement data
Sentiment Analysis (SA)	Automated Sentiment Analysis (ontological dictionaries)
	Machine Learning
Supervised Aggregated Sentiment Analysis (SASA)	ReadMe (Hopkins & King 2010)
	iSA (Ceron, Curini & Iacus 2016)

Computational approaches

Purely Quantitative

Endorsement data: counting #followers, #likes

Volume data: counting the # of mentions related to a party or candidate or the occurrence of particular hashtags (such as the party name) etc.

More followers, likes, mentions, more votes!

Limits: Endorsement/volume data measure the **degree of public attention** or awareness around each candidate/party. Anything more?

So…add some **sentiment** to that!

Sentiment analysis

Analyzing the ***stance of the comments***: positive, negative or neutral to infer

Usually, more positive, more votes!

But also other approaches:

How to map tweets into votes for the US Presidential Race 2012 (Ceron, Curini & Iacus 2014):

- a) the tweet includes an explicit statement related to the intention to vote for a candidate/party
- b) the tweet includes a statement in favor of a candidate/party together with an hastag connected to the electoral campaign of that candidate/party
- c) the tweet includes a negative statement opposing a candidate/party together with an hastag connected to the electoral campaign of another candidate/party
- Also retweets that satisfy any of the previous conditions

Sentiment analysis: ML vs. SASA

Notice that in social science as well as in electoral studies, what matters in forecasting attempts is the **aggregated distribution of opinion** or share of votes rather than the individual opinion or vote behaviour

Estimating a "good" aggregate measure with the lowest possible error is what is relevant here!

Therefore, SASA approaches better….?

SM advantages with respect to political campaigning & elections

You listen, you do not ask!

Less affected by the **social desirability bias** that often plagues survey on "hot" topics (i.e., racism, **sympathy toward terrorism**…more on this, this afternoon!)?

- ✓ But also Brexit, the "*Shy-Tory*" ("*Shy-Trump*"?) effect

SM advantages with respect to political campaigning & elections

A **geo-localized** analysis is possible…

…as well as a **real-time analysis** of the electoral campaigning (i.e., which is the impact of a TV debate on the popularity of candidates?)

Through that it becomes possible to capture (and often anticipate) sudden change in public opinion (so called "momentum"): **nowcasting** the present!

Let's see an example based on the US 2016 Presidential Campaign

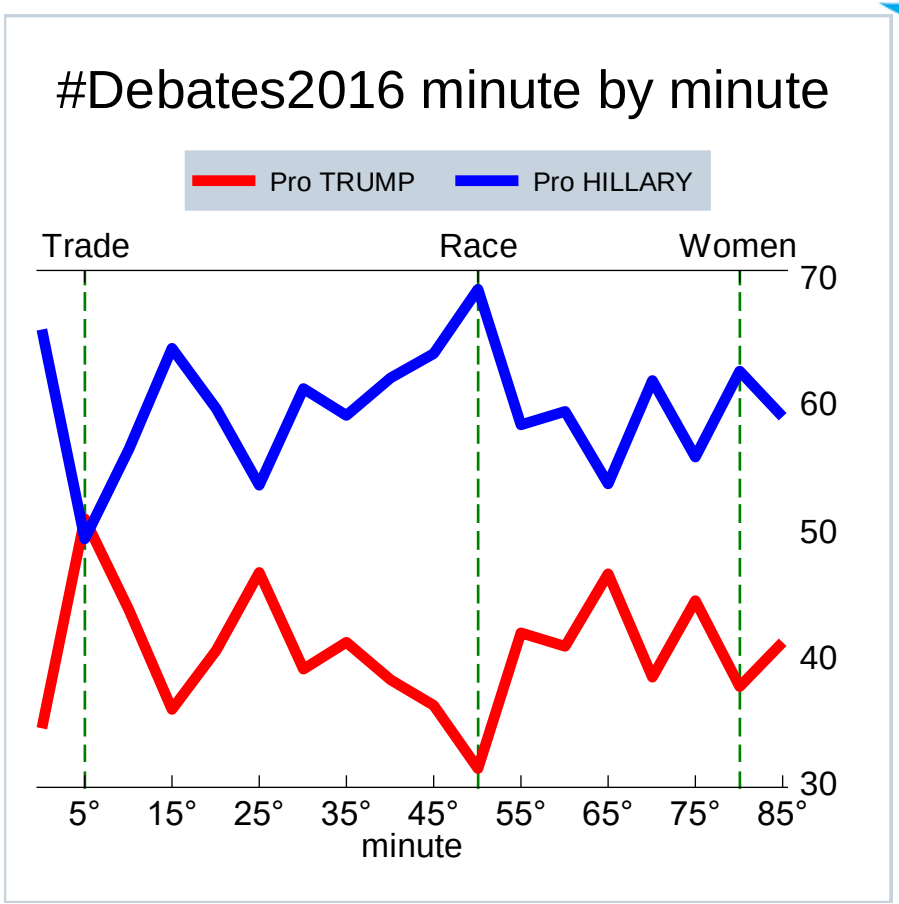
US Presidential Race 2016: the Debates

US Presidential Debates: First One

????

62%

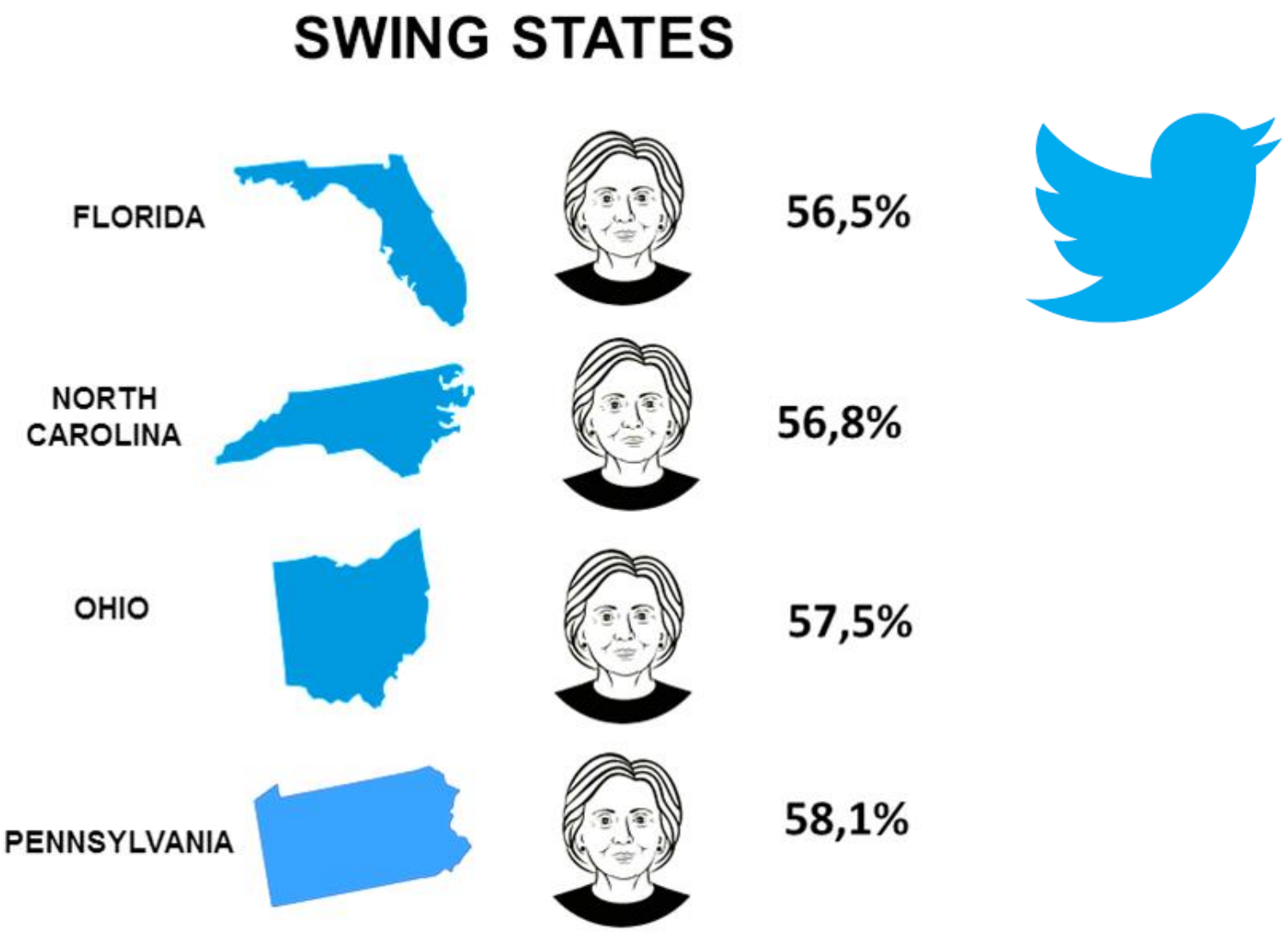
27%



59.5%

40.5%

US Presidential Race 2016: First Debate



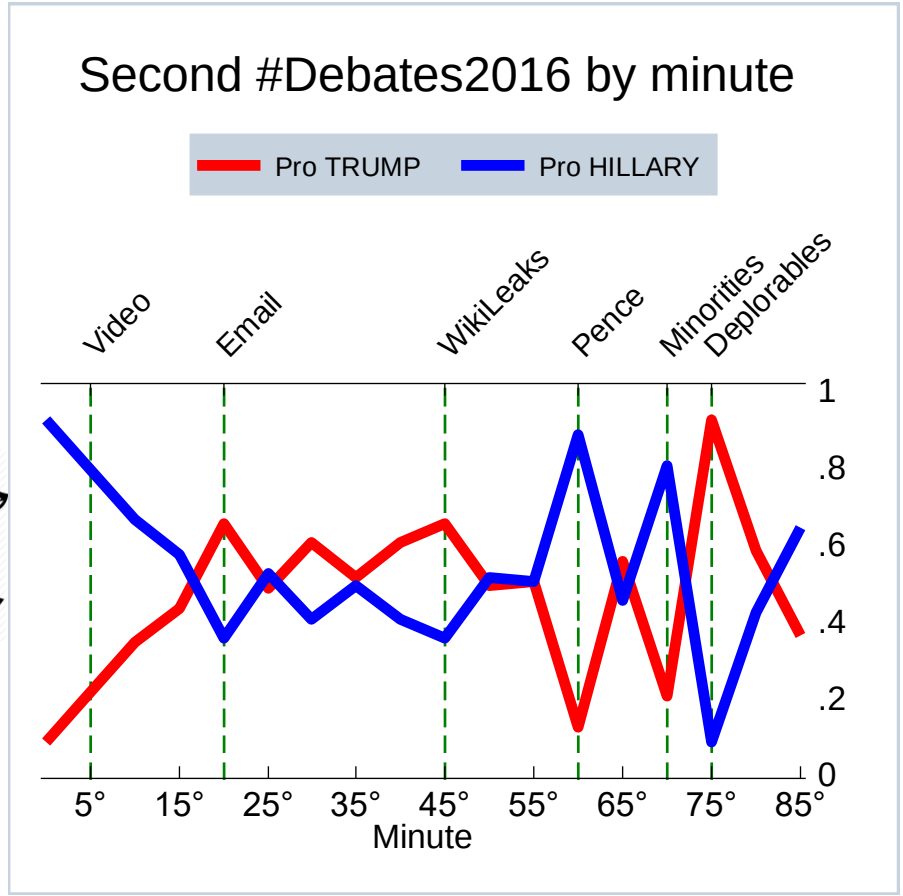
US Presidential Race 2016: the Debates

US Presidential Debates: Second One

????

57%

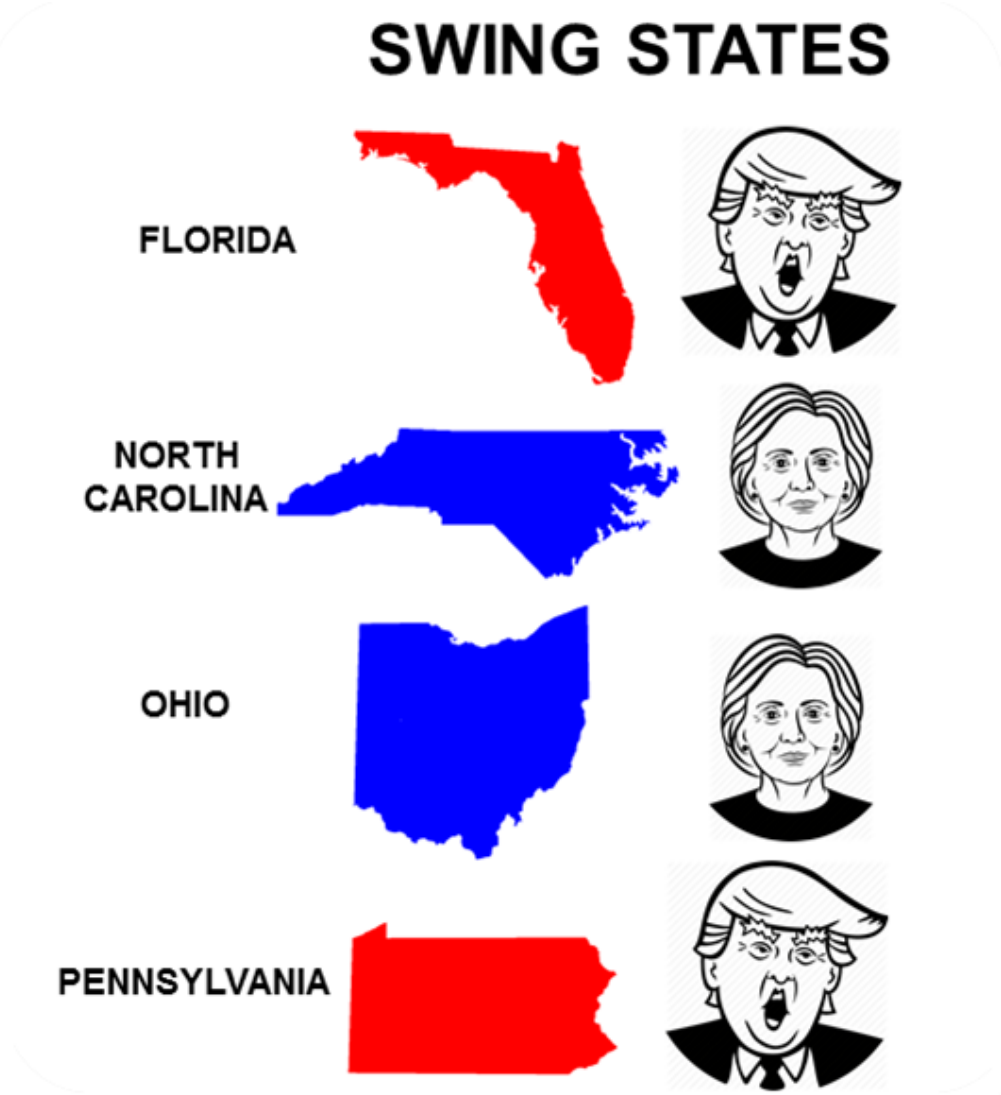
34%



51.5%

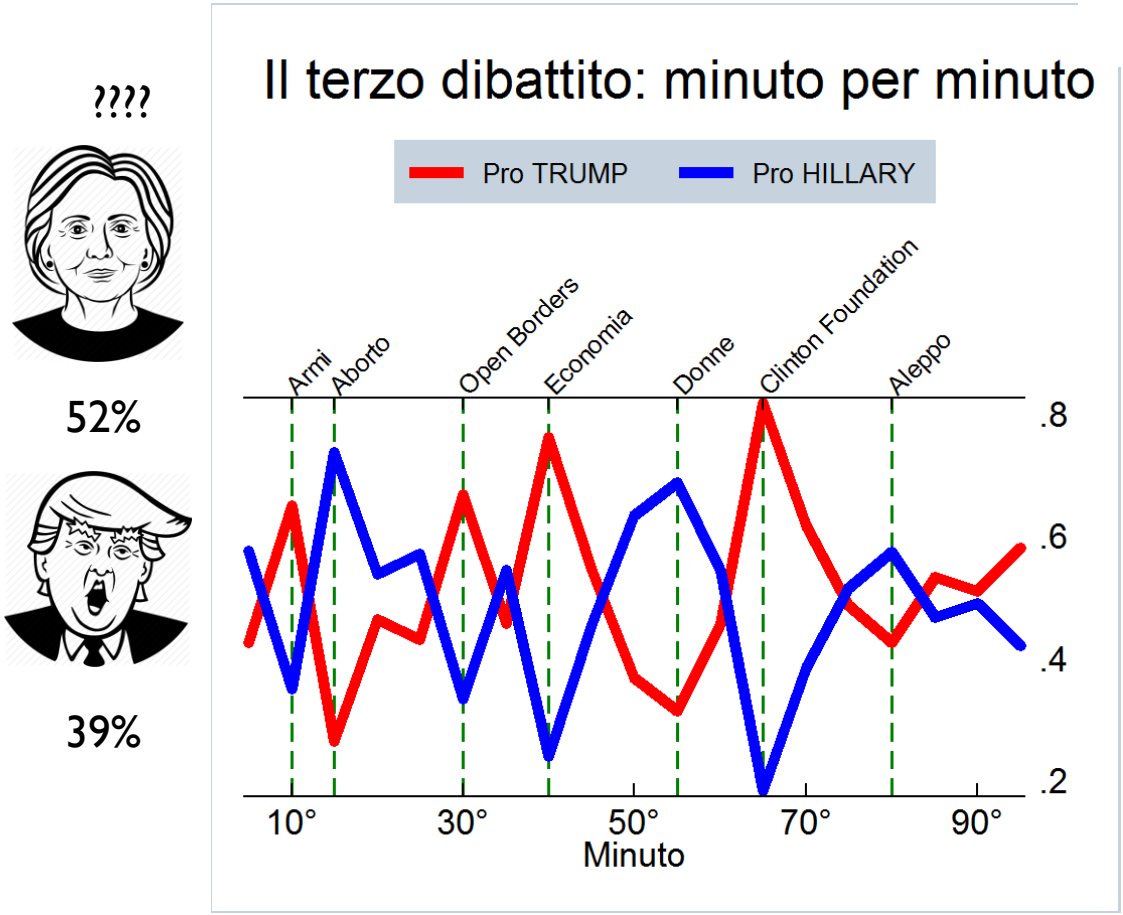
48.5%

US Presidential Race 2016: Second Debate



US Presidential Race 2016: the Debates

US Presidential Debates: Third One



????

52%

39%

50.5%

49.5%

Limits of Social Media data

- The **real profiles** behind social media accounts are **not known** in most cases
- The population on Social Media is (can be?) a **biased sample** from the demographics population (and in the surveys?)
- The population of Social Media under observation, **changes** according to the topic
- Social media **are not the same** everywhere (no FB but VK in RUSSIA, no Twitter but Sina Weibo in China, etc)

(possible solutions to some of these issues exist)

Beyond nowcasting…

Can we also **forecast** the electoral final result?

This is actually quite fascinating because…

…to validate the predictive accuracy of a model we need to have an **independent measure** of the observed outcome that the model is trying to predict

In this respect **forecasting an election** is one of the few exercises on collective social events where an **independent measure of the outcome** that you want to try to predict is clearly available, i.e., the vote-share of candidates (and/or parties) at the ballots

A meta-analysis

239 electoral forecasts related to 94 different elections, held between 2007 and 2015 in 22 countries, covering all the five continents (Japan included!)

Our DV: the **MAE of each social-media based forecast** (we focused just on vote-share, not seat-share!)

Within our sample, the **average value** of MAE is 7.39

MAE of electoral surveys for same elections: 2.22

Note, however, that the **variance of MAE** within the social-media based forecasts: s.d 6.65, i.e., some social-media forecasts **were as good as (if not better than)** surveys

A meta-analysis

Our research question

- ✓ **when social media analysis** is able to provide accurate forecast and **when not** (method, context, other elements prompting the coherence between online opinions and offline behavior, etc.)

How to map posts/tweets into votes?

Computational approach: more volume (more discussion), more votes!

Sentiment approach: more positive posts, more votes!

SASA approach: more positive posts (at the aggregate level), more votes!

Which factors matter?

The method through which you extract information from social media is crucial!

SASA method decreases the MAE by 3.4 points if compared to forecasts based on a mere **computational approach** and by **2.2 points** if compared to **other SA techniques**, which **are not more effective** than computational methods in improving the accuracy of the prediction (and **iSA beats ReadMe**!)

Which factors matter?

Institutions do matter!

When elections are held under PR, the MAE decreases by a remarkable 2.13 points if compared to plurality (on-line *sincere* vs. off-line *strategic* vote effect?)

Volume also matters! Having more information on citizens' preferences decreases the error, though only when the **turnout rate is sufficiently high**, i.e. when we can expect to observe an actual behavior that is somehow consistent with the declared one

Which factors matter?

Other Prediction's Attributes?

No Academic vs. Non Academic impact

No time effect (whether prediction was made ex-ante/ex-post) impact

No per-user comment impact

Which factors matter?

Other Prediction's Attributes?

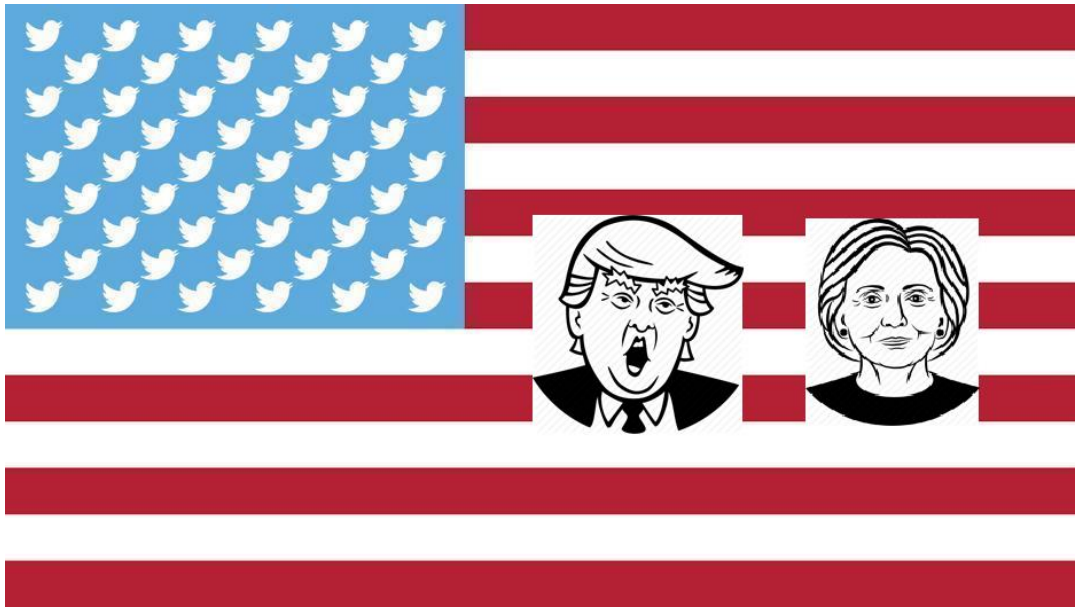
Weighting the predicted share of votes according to the socio-demographic features of the users has a negligible impact. How is possible?

Weighting procedure limits?

No socio-demographic representativeness; but political? And in terms of topic coverage?

Which factors matter?

- Finally, taking into account polls when doing social media electoral forecasts **reduces the MAE considerably**
- …let's go back to the US 2016 Presidential Campaign



U.S. 2016 Presidential Elections

The Method we employed

Three steps process

First: we focused on all posts coming from U.S., written in English (no Spanish or other languages), explicitly mentioning on of the two main candidates via Twitter API. Between 3/5M tweets on a daily-basis

The Method we employed

Three steps process

- To forecast the **nationwide popular** vote via iSA, we counted only tweets coming from the United States
- To estimate the **state-by-state results**, we analyzed tweets geolocated in each state, using the geolocation information metadata attached to each tweet

That means that while we were able to **effectively** monitor the Twitter discussion about the campaign nationally, **we could not be** as precise for individual states, because only a fraction of tweets (2 to 5 percent) give location data in the US case

The Method we employed

Three steps process

Second: *econometric calibration*: from 19th of September till the 2nd of October, we run an econometric model to explain the survey data **at the national level** in terms of our sentiment analysis estimate

- In particular, we considered only the n**egative sentiment against Donald Trump** and the online **voting intentions expressed in support of Hillary Clinton**
- Together these two variables are able to explain .97% of variance in the polls avarage taken from Real Clear Politics

The Method we employed

Three steps process

Third: since the 3rd of October we relied on social-media data only (i.e., the two previous variables weighted according to the analysis just illustrated), to produce our final estimates of the electoral forecast

Final results?

Quite good at the national scale %: +1.2% Hillary (actual restul: +2.1%)

The Washington Post
Democracy Dies in Darkness

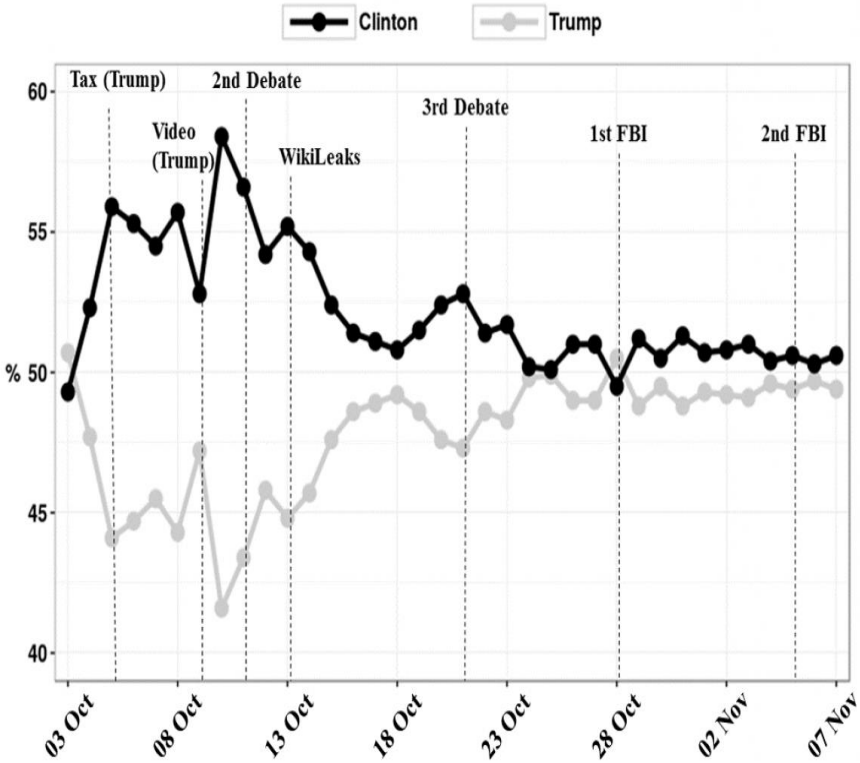
Monkey Cage

How pollsters could use social media data to improve election forecasts

By Andrea Ceron, Luigi Curini and Stefano Iacus December 21, 2016

Twitter logo. (Leon Neal/AFP/Getty Images)

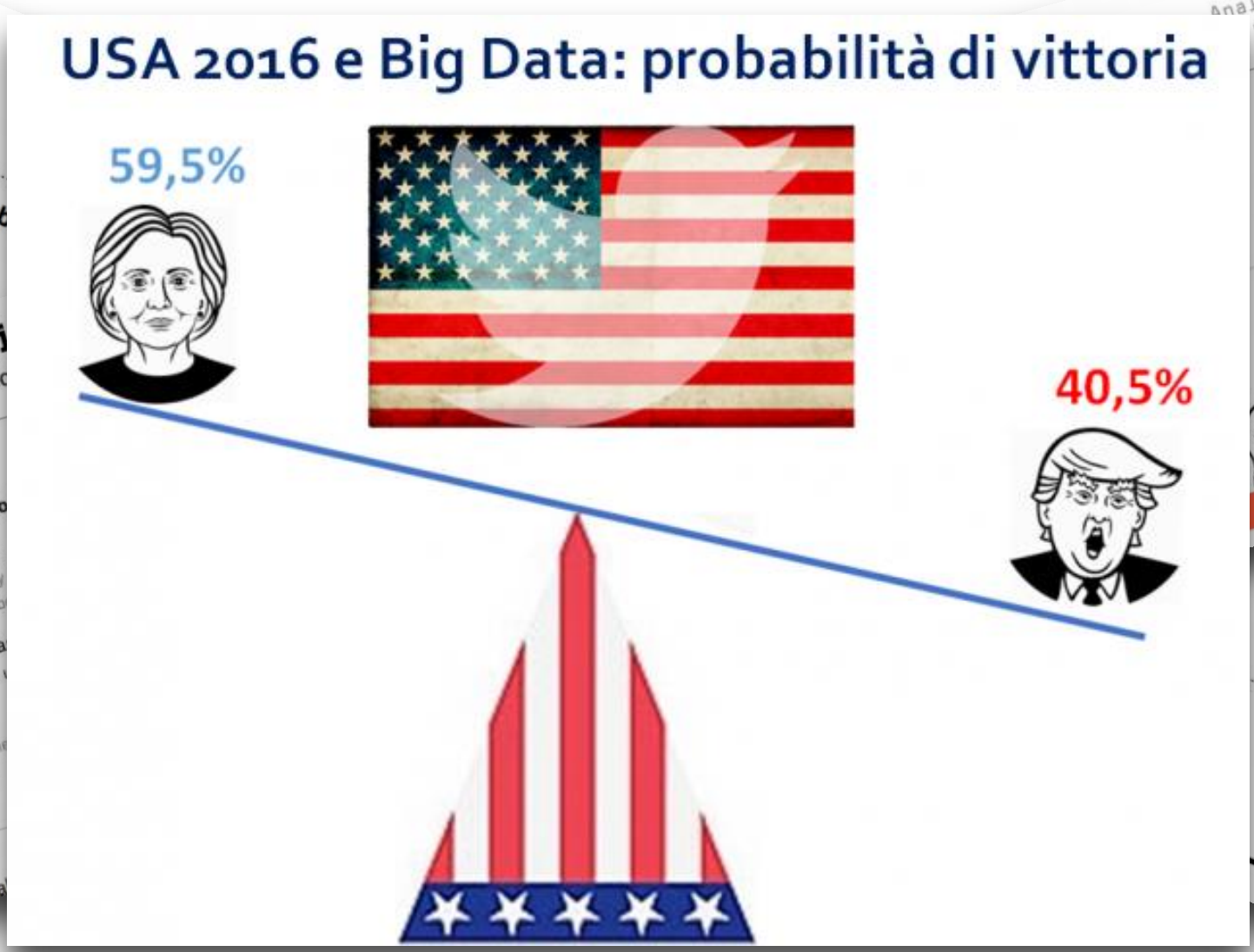
Donald Trump's Nov. 8 victory surprised almost everyone. But if pollsters had looked at Twitter, they might have recognized that the race was close — or so we learned in our recent research.



The Method we employed

But what about the outcome of the election?

Big Data vs. the rest



The Method we employed

Ohio, Florida, Nevada and **Colorado** were basically not in competition (contrary to State-surveys). Even in mid October, the first two ranked consistently for Trump, and the latter two for Clinton

Pennsylvania race was much closer than expected, with predictions moving back and forth, favoring one candidate and then the other (Trump up between the 3rd and 6th of November. At the end Clinton at 51.5%). Same for **Michigan** (Clinton: 50.5%)

The Trump rise in the other **Midwest States** (Wisconsin or Iowa) couldn't have been predicted via Twitter

The Method we employed

The inaccurate results are probably affected by the **limits of the geolocated data** and the fact that, instead of **calibrating the social media results with state specific surveys**, we relied on the national data.

This was not a deliberate choice; we did so because no state polls were available every day, while national ones were

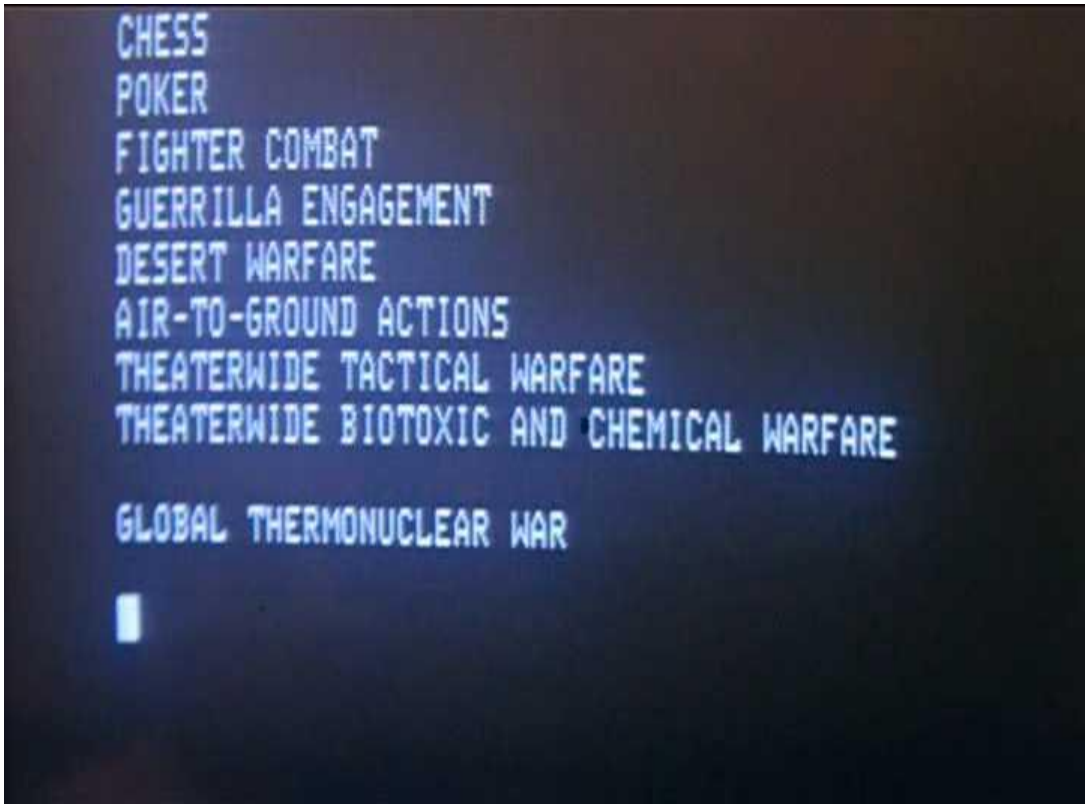
Conclusion

Public opinion has profoundly changed. And the way to measure it must change as well…

…there is no longer the option to avoid listening to social media information or, even worst, considering it as merely noise

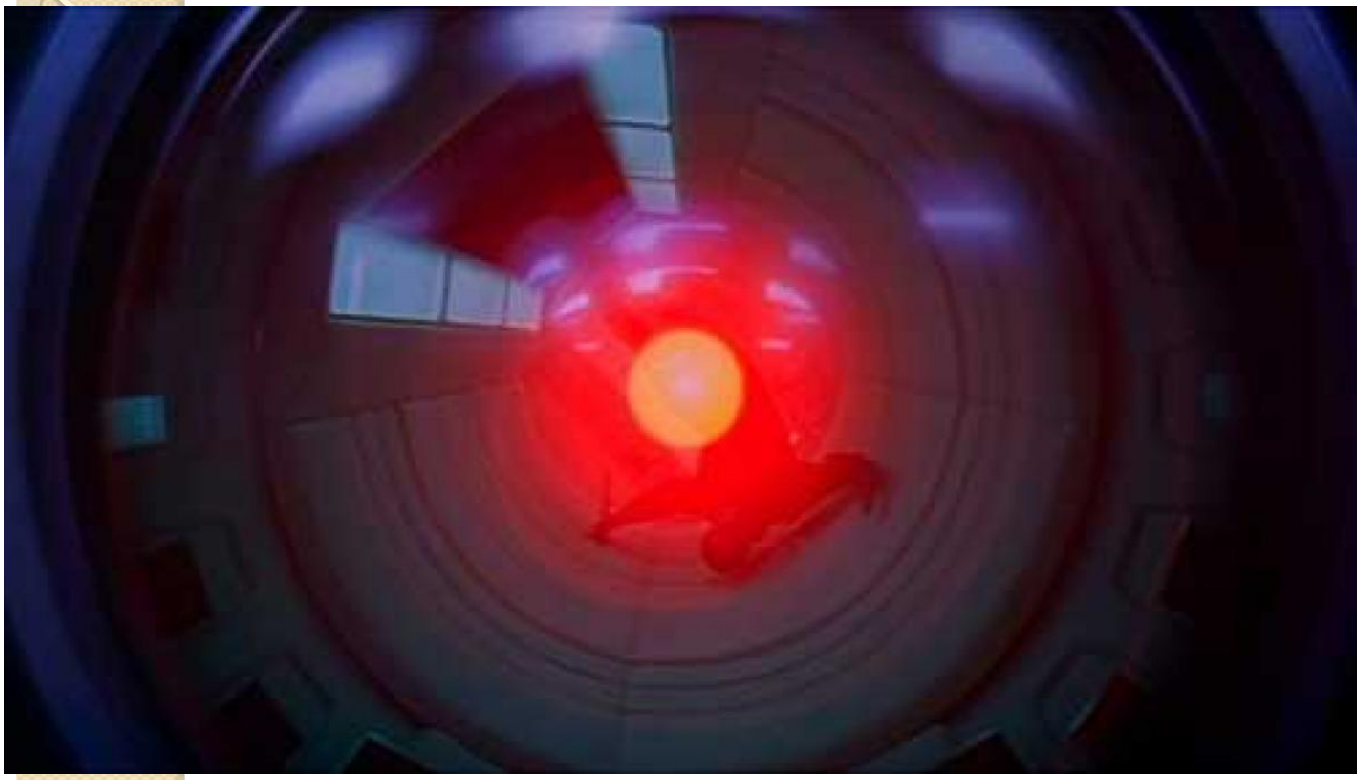
But beware of the challenges!

Conclusion



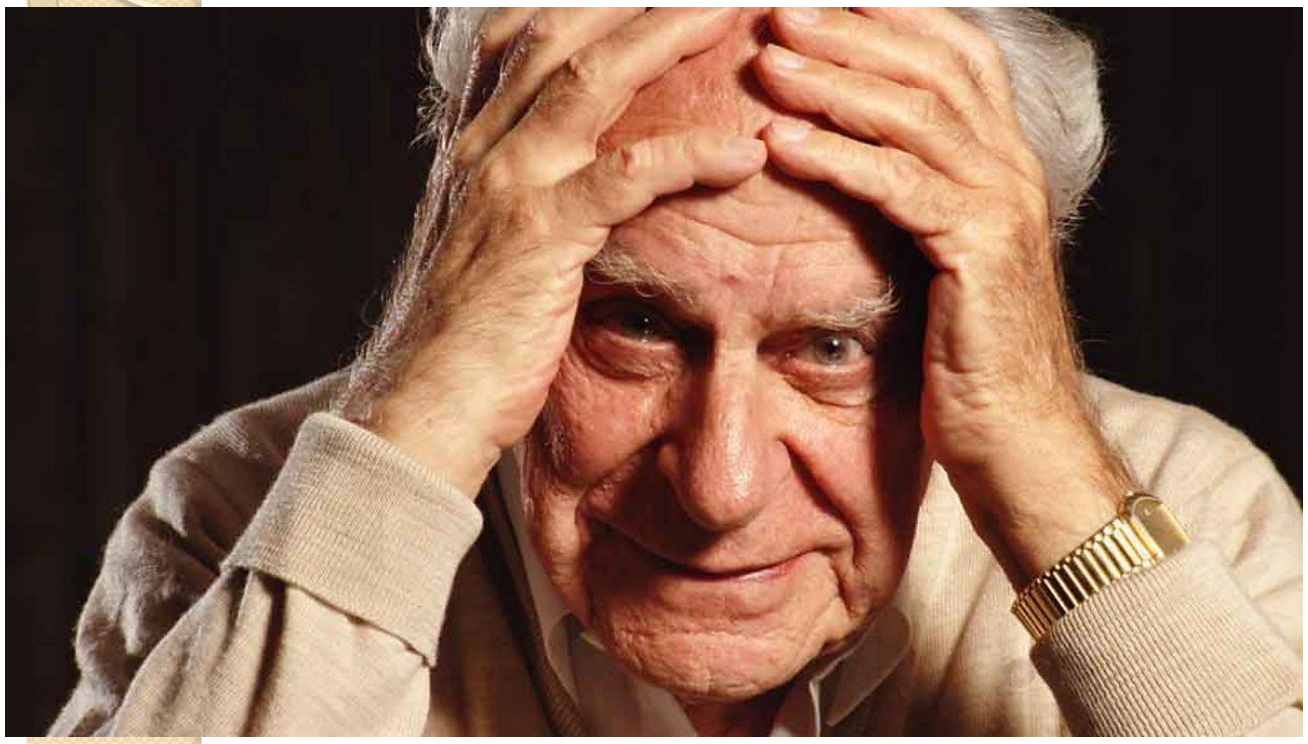
"To err is human, to really mess things up requires a computer"

Conclusion



For example: behind a ML algorithm there is no any explicative model!

Conclusion

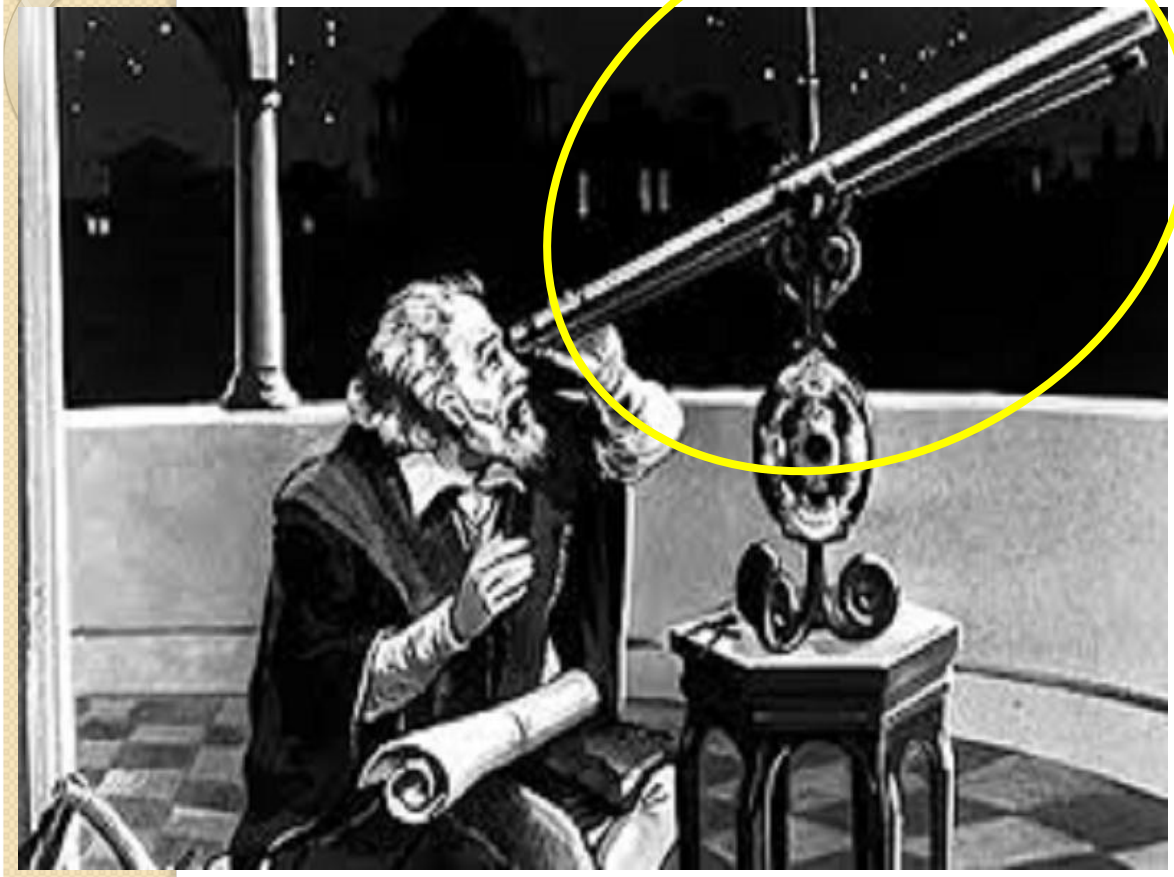


"**Induction** is not so much wrong as impossible"

Without a **theoretical understanding** of the world, how would we even know what to describe?

This remains true also in a BIG DATA world!

Conclusion



Big data has the power to transform and expand the universe of answerable social science questions…but we need **new questions** now!

Two final sentences

Big Data are simply "today's data" to (better) understand our world

ONLY More GOOD Data is better than less