Viz Bang!

Data consulting @ nexus-data.info

A response/rebuttal to Ed Yong's article in The Atlantic about my talk at ASHG 2015

CC: @theatlantic @edyong

I wanted to clarify and correct some of the claims made about my work in an article in the Atlantic by Ed Yong. You can find it here: http://www.theatlantic.com/science/archive/2015/10/no-scientists-have-not-found-the-gay-gene/410059/. I have reached out to the Atlantic via Twitter about this but have heard nothing back as of this posting.

I am not interested in a public argument but Ed made fundamentally inaccurate claims that could damage my professional reputation. I would have appreciated the chance to explain the analytical procedure in much more detail than was possible during my 10-minute talk but he didn't give me the option.

Essentially, Ed's claim that inappropriate statistics were used is not credible because he clearly misunderstood the analytical procedure. This is not a matter of opinion. He mischaracterized several crucial things as I will explain in detail below. He is free to disagree or critique my work but not if he is getting the basic facts wrong. I like constructive criticism. I have had several good debates on Twitter and Tumblr with people who don't agree with my findings but understood the approach. This is not an attempt to shut down valid concerns and questions.

I want to address the most troubling issue first: Ed's misinterpretation of the approach I took in building my models. All models (from the very first to the final one) were built using JUST the training data. Only after we had created the model did we test their performance on the test data (the algorithm didn't 'see' these during model creation). If performance was unsatisfactory, we remade the model by selecting a different set of predictors/features/data based on information from the TRAINING set and then reevaluating on the test set. **This approach is used widely in statistical/predictive modeling field. It is not an insidious issue or data manipulation, despite how Ed chose to present it.** If he would look into the literature, he would see many papers emphasizing the importance of model tuning and feature selection. If this approach is wrong, someone needs to tell Amazon, Netflix, Google, and just about everyone doing statistical modeling and machine learning.

The second issue I want to discuss is his claim that we needed multiple testing correction. Again he is misunderstanding the approach and rationale. We did not need to correct for multiple testing because we did one hypothesis test. We are not testing whether each of the 6000 marks/loci are significantly associated with sexual orientation. If we had done that, multiple testing correction would have certainly been warranted. But we didn't. **The single test we did was to ask whether the final model we had built was performing better than random guessing. It seemed to be because its p-value was below the nearly universal statistical threshold of 0.05.**

Third, I have always been clear that it is too early to determine the nature of the relationship of the genes to sexual orientation. **I believe he was at my talk so I am not sure how he thought I said anything about causation.** I said they might be promising candidates because their functions seemed to make sense in the context of sexual orientation. Those were the extent of my statements.

Finally, he brings up the possibility that none of these findings will be borne out. That is true. **We need to validate them in a much more highly powered study.** We need a validation set of data.

I never claimed to have found the ‘gay gene(s)’. All I presented on were results that I believed were interesting, credible, and worthy of followup. He disagrees with me about that, which is valid and not something I have an issue with at all. But in order to conclude that my approach was incorrect, he needs to understand what I did, which is something that he failed to do.

Tuck Ngun

#ashg15

1 note Oct 10th, 2015

MORE YOU MIGHT LIKE

A brief digression from pretty pictures

I was fortunate enough to be able to share my with at #ashg15. I just wanted to respond to comments from @epgntxeinstein about the study. To @epgntxeinstein, thank you for taking the time to write about it. I think they made many valid points but I wanted to clarify and give some context.

First off, I didn’t take anything they said personally except for this: “Some poor young lad gets up on stage at #ASHG15 having worked hard to generate this story and is now being eviscerated by people like me.”

It would be a big mistake to reduce me to that. I take issue with the implication that I’m going to wilt under scrutiny. Whatever you dish out, I can take it. Trust me, I’ve had to deal with a lot worse as someone who grew up gay and an outsider. Dealing with critiques about my work are nothing compared to dealing with people telling me I’m going to hell.

More specific responses:

1. We used a sequencing approach (RRBS) not arrays. We did find microbial DNA sequences which we filtered out. Some could have gotten through I suppose.

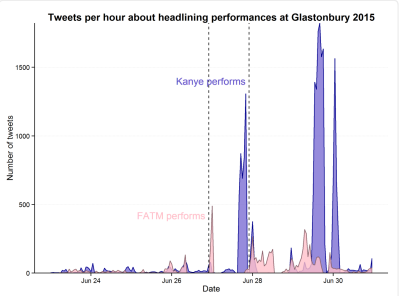
2. “What is not described in the abstract is that the cells used were from saliva samples”

pjie2 asked:

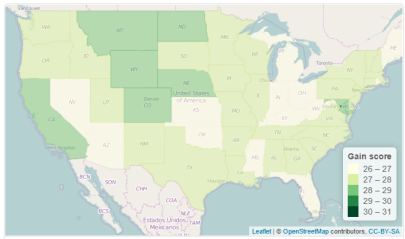
Your response seems to be self-contradictory. First, you say "If performance was unsatisfactory, we remade the model...", but then "The single test we did was to ask whether the final model we had built was performing better than random guessing." How often did you 'remake' the model, i.e. how many versions of the model did you try? This is the bit that needs a multiple testing correction. Getting p=0.05 on the best out of 20 different versions of the model is not a significant result.

Hi pjie2,

My collaborators and I are going to issue a joint statement to clear this all up. Until then I have been advised to not talk about this issue on social media. Look out for the statement!



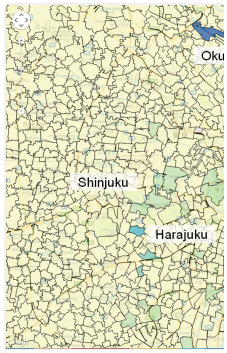
A refinement on my previous post. It’s often instructive to compare two things. So I’m going to compare



A few years ago the Obama administration implemented a program called CARS, a.k.a. Cash for Clunkers. I decided to explore how successful the program was in each state. To measure success, I devised a metric I call the Gain score. It’s a composite measure of the amount of liquid capital that flowed into the economy and the increase in fuel efficiency of the vehicles purchased through the program. Washington DC came out tops. Click here to get to an interactive version of this map. It’s fun!

#Choropleth #obama
#cash for clunkers
#way back wednesday #prius #cars

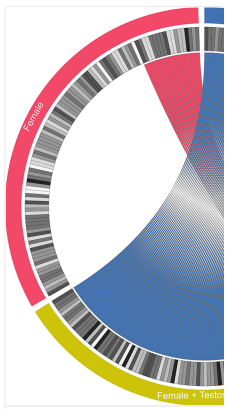
I was at the Hammer in Westwood today and saw Mark Bradford’s exhibit there. By far my favorite piece was a massive painting he did in the front lobby that blurs the line between art and data viz. It’s a choropleth depicting the number of AIDS cases in the US in 2008. His intent was to bring up the question of the accuracy of visualizations because they are supposed to tell us something about



Museum density by n
central Tokyo. I’m pla
in the fall and one of t
doing when I travel is
this played a role in d
my hotel was. Oku (at
division of it) had the
concentration of muse
Harajuku and Ueno. D
from Open Street Map
great but I wish I coul
subdivisions of the wa

Source: dl.dropboxuser
#travel #nippon #jap
#osm #open street ma

1 note



A recent visualization
really happy with. I di
friend/colleague. It sh
expression patterns o
that are expressed at
male vs female cells (
the gray bars). When

True. We only had so much space in the abstract. But to address the point more directly, there is almost no way of getting to the best tissue (the brain) for the population we were interested in. I would have been much happier with the ‘right’ tissue type.

3. “Also not described is the marginal, uncorrected significance for this underpowered study”

A) The only source for it being uncorrected is a tweet from Leonid Kruglyak. He is not involved at all with the study and got it wrong. The p-value was for a single test: is the model’s performance different than just random guessing? As for marginal, that’s up to you to decide. It’s why I show the actual p-value.

B) Yes, we were underpowered. The reality is that we had basically no funding. As with tissue type, the sample size was not what we wanted. But do I hold out for some impossible ideal or do I work with what I have? I chose the latter.

3. “Why use a new algorithm to identify these predictive markers, did current approaches not yield any results?”

In short, yes. The algorithm (FuzzyForest) is available for anyone to use (look on Github) and extensively documented. It was mainly used for feature selection. The model was ultimately constructed using random forest, a widely used and well tested algorithm.

4. “…like everyone else in the history of epigenetics studies they could not resist trying to interpret the findings mechanistically. So they go there: they talk about the genes implicated.”

Let’s be real here: no one is going to pay attention unless you talk about implicated genes. It’s all about interpretability. We ultimately want to understand what’s going on in terms of the biology so of course we’re going to talk about any genes that seem related and are interesting.

5. “We should only present biomarker studies when they are shown to perform robustly as biomarkers…If we have an intriguing preliminary observation, we present it as such and

tweets about Florence + The Machine (who was one of the other headliners at this festival) to Kanye.

#data visualization #ggplot2 #sentiment analysis #time series #kanye west #florence and the machine #fatm #twitter

the present but are inherently ‘flawed’ because they are based on data from the past.

#hammer museum #mark bradford #art #choropleth #history #hiv #aids #data visualization #empiricism

1 note

in testosterone, the va
these genes become
(indicated by the blue
minority are relatively
remain female-typical
red lines).

#data visualization #g
#gene expression #ci

1 note

do not claim that we have generated '…strong support to the hypothesis that epigenetics is involved in sexual orientation.'"

They are absolutely right that this is a preliminary study and that is how I always talk about it (see Science and Buzzfeed stories about this study where they actually talk to me and not just quote another source). Preliminary though this study may be, I stand by that statement and my work. However, this view is subjective and people are free to disagree with me as @epgntxeinstein has. I like debate and we should have as much of it as we can.

6. "@NatureNews [and] the American Society for Human Genetics…should know better…not blindly accept that the numbers generated from DNA methylation studies are inherently meaningful."

'Meaningfulness' is a fair standard but I think both organizations thought that our study was interesting, which is an acceptable standard too in my opinion.

I hope it is obvious I am not taking any of this personally with the single exception noted at the start. I welcome discussion, debate and criticism. That's how the world works best.

Tuck Ngun