



How to talk to your bioinformatician?

January Weiner 

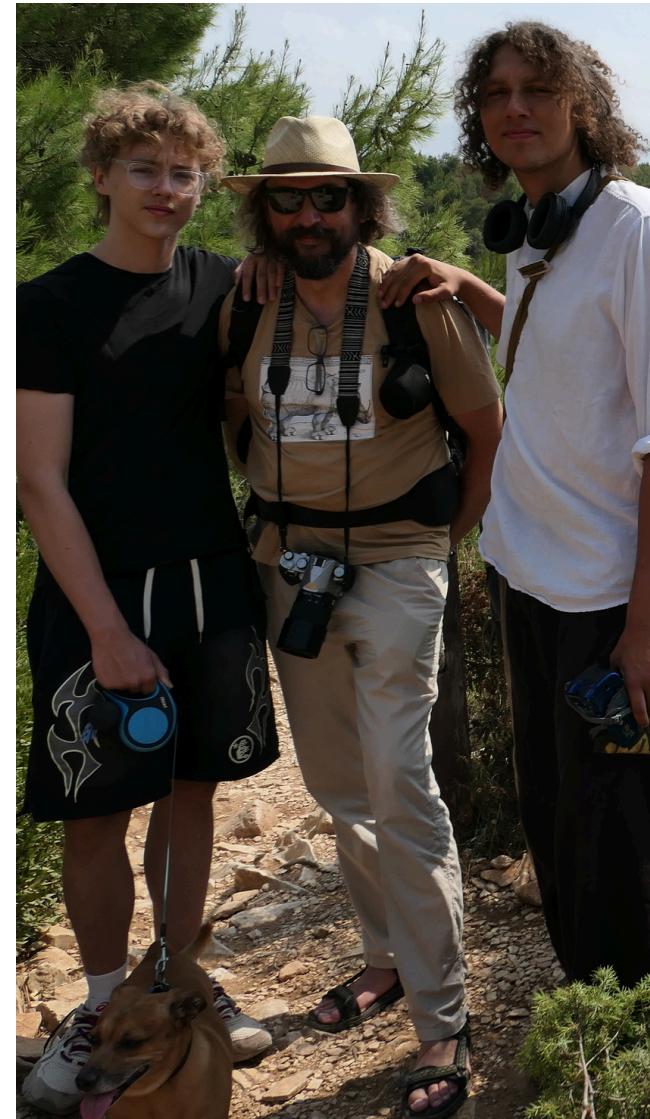
Core Unit for Bioinformatics, BIH@Charité

Then and now...

How it was going

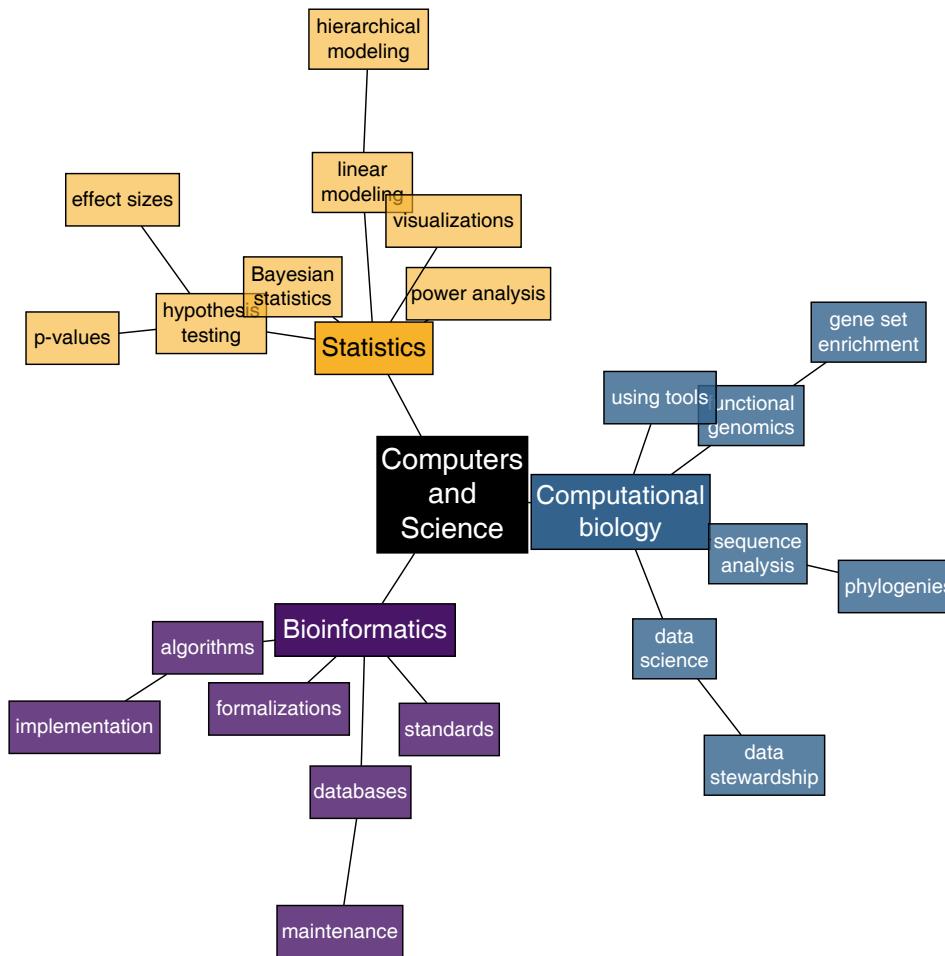


How it is going



Core Unit for Bioinformatics, BIH@Charite

What is this thing we call “bioinformatics”?



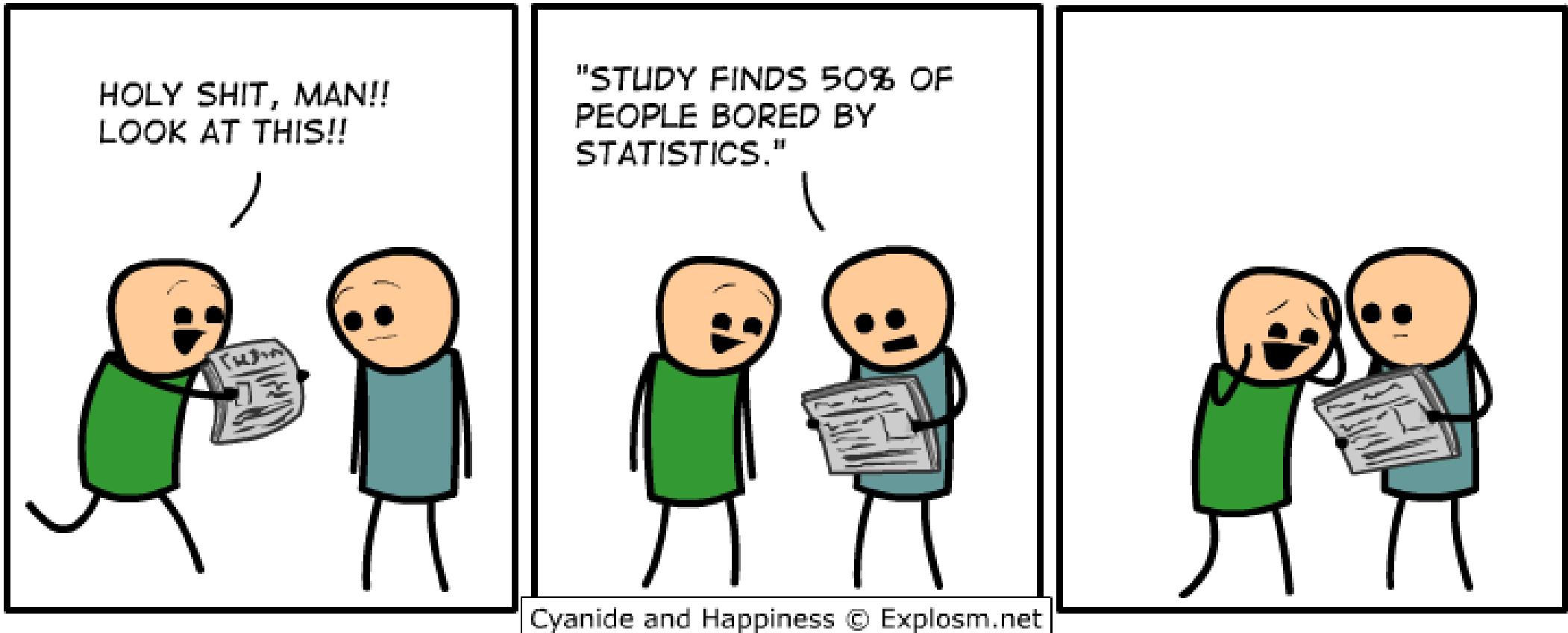


My key advice to you

Communication is key

- Keep explaining your project – teach us
- Work iteratively
- Meet frequently

Statistics



What is a p-value?

H_0 : The null hypothesis, no effect

H_1 : The alternative hypothesis, there is an effect

We run a test, we get a p-value, say 0.03. It is a probability.

Probability of *what*, exactly?

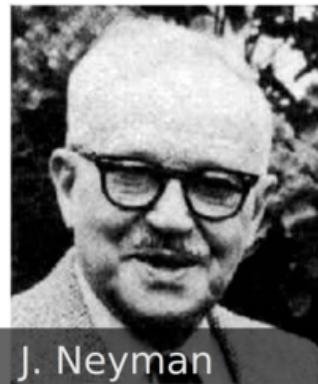
1. Probability that H_0 is true (probability that there is no difference), given the data
 2. Probability that H_1 is true (probability that there is a difference), given the data
 3. Probability that the data is random
 4. Probability that the observations are due to random chance
 5. Probability of getting the same data by random chance
- Probability of observing an effect at least as extreme given that H_0 is true

Our intuition is bayesian, not frequentist

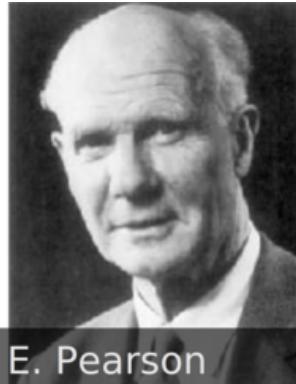
Frequentist Statistics



R.A. Fisher

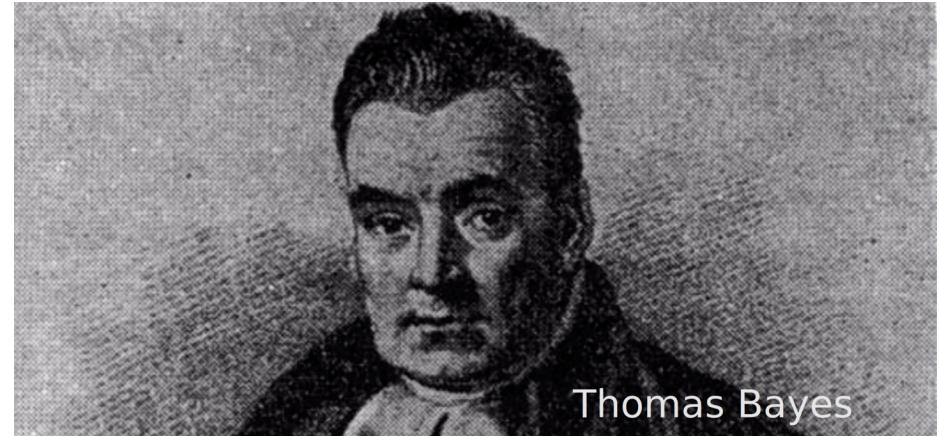


J. Neyman



E. Pearson

Bayesian Statistics



Thomas Bayes

1. Probability is defined as the long-run frequency of events
2. Parameters (like the “true value”) are fixed but unknown quantities.
3. Asking about the probability of a hypothesis does not make sense

1. Probability represents a degree of belief or certainty about an event
2. Parameters are treated as random variables with their own probability distributions.
3. Asking about the probability of a hypothesis is the main goal

Why is that important?

P-values are part of scientific language

- Always use effect sizes
- Never rely on p-values alone

Know their limits:

- they control only type I errors (false positives)
- they **do not** control type II errors (false negatives)

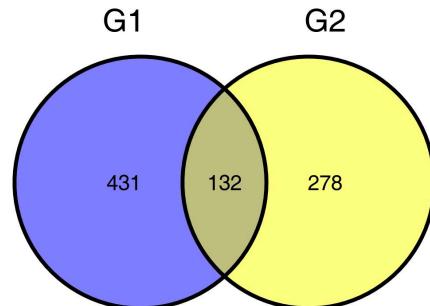


How Venn diagrams can fool scientists

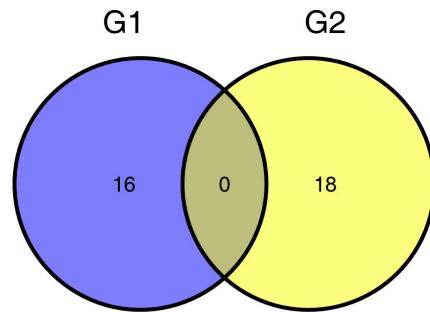
COVID-19 study, both COVID-19 patients and non-COVID-19 patients are compared in two groups of people, *G1* and *G2*.

We wanted to know whether the influence of COVID-19 is different in these two groups.

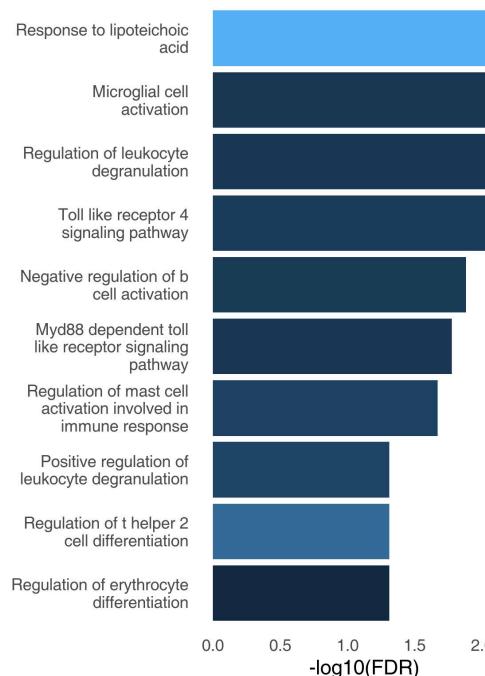
A Differentially expressed genes



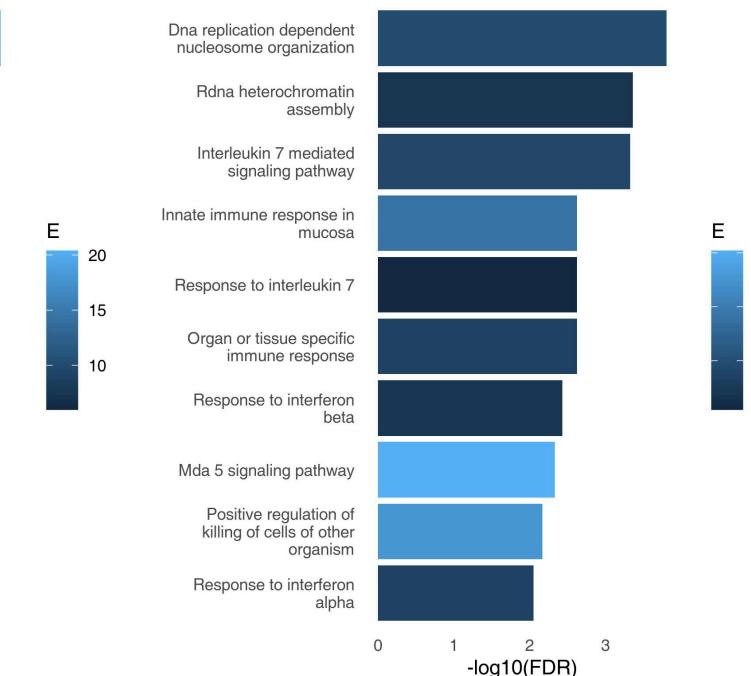
B Enriched GO terms



C G1



D G2

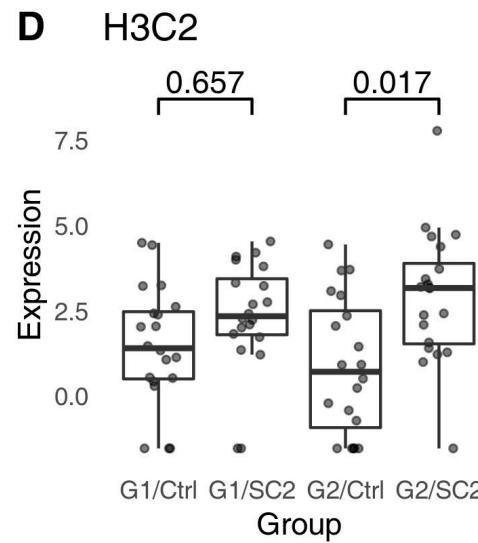
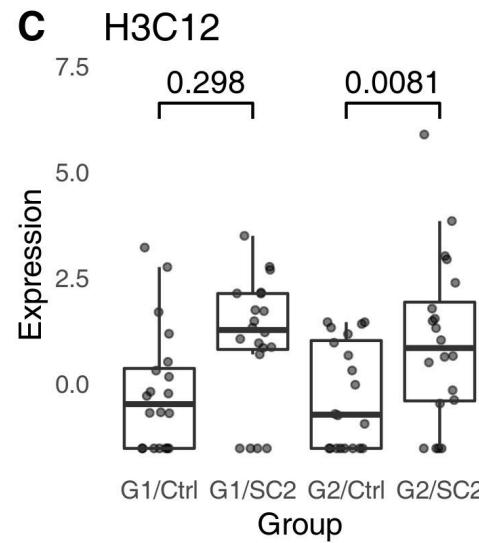
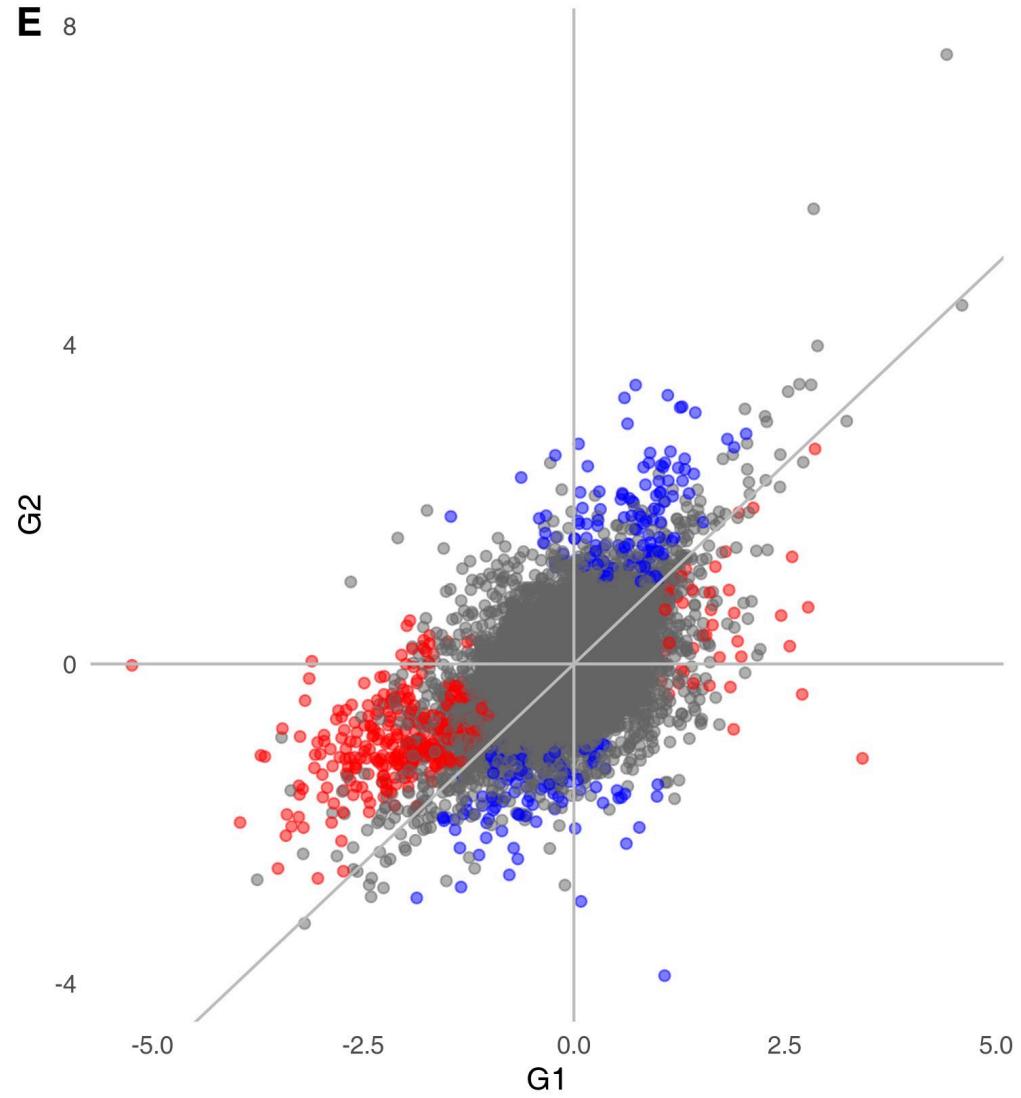
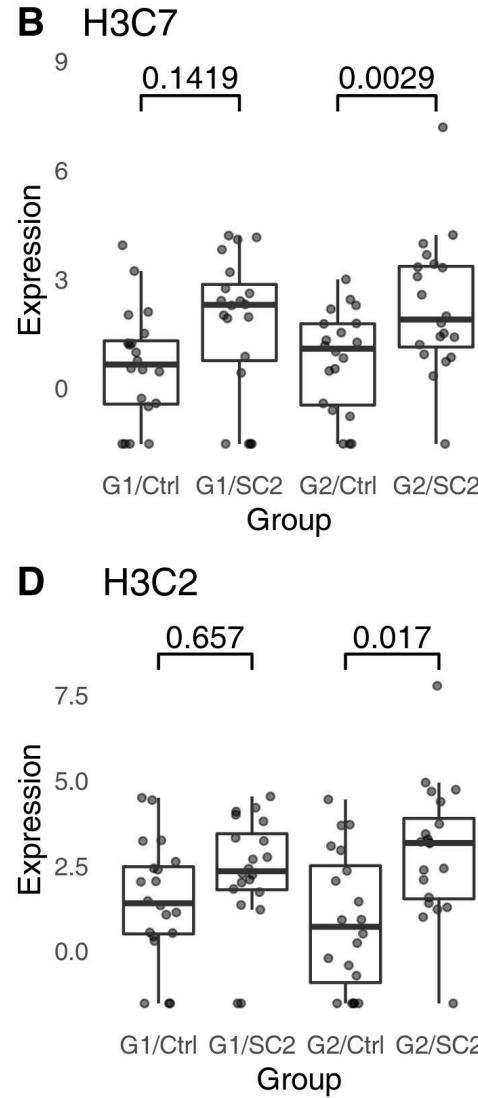
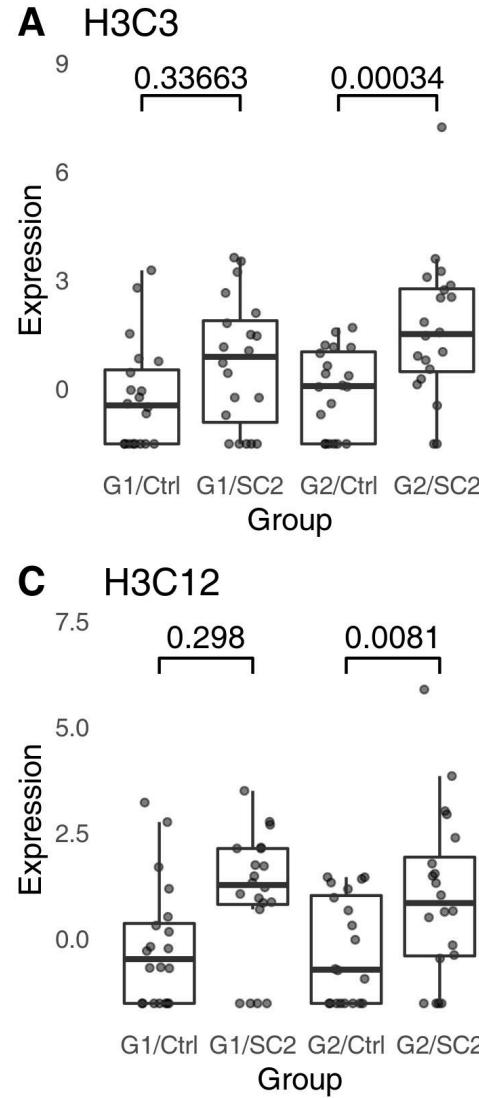


Venn diagrams may indicate erroneous statistical reasoning in transcriptomics. Weiner, Obermayer and Beule,

Core Unit for Bioinformatics, BIH@Charité

The results are artifacts!

Groups G1 and G2 were randomly drawn from the same population. They were not different at all.



What happens is, we are comparing significance with non-significance

The Difference Between “Significant” and “Not Significant” is not Itself Statistically Significant

(Andrew Gelman and Howard Stern)

If a gene is significant in one comparison, and not significant in another, that does not mean that there is a difference between the two groups.

It simply means that we *failed* to detect the difference in one of the comparisons, but that is actually quite likely to happen!

 Therefore:

Don’t say “there is no difference”. Say “we did not detect a difference”.

Reproducibility

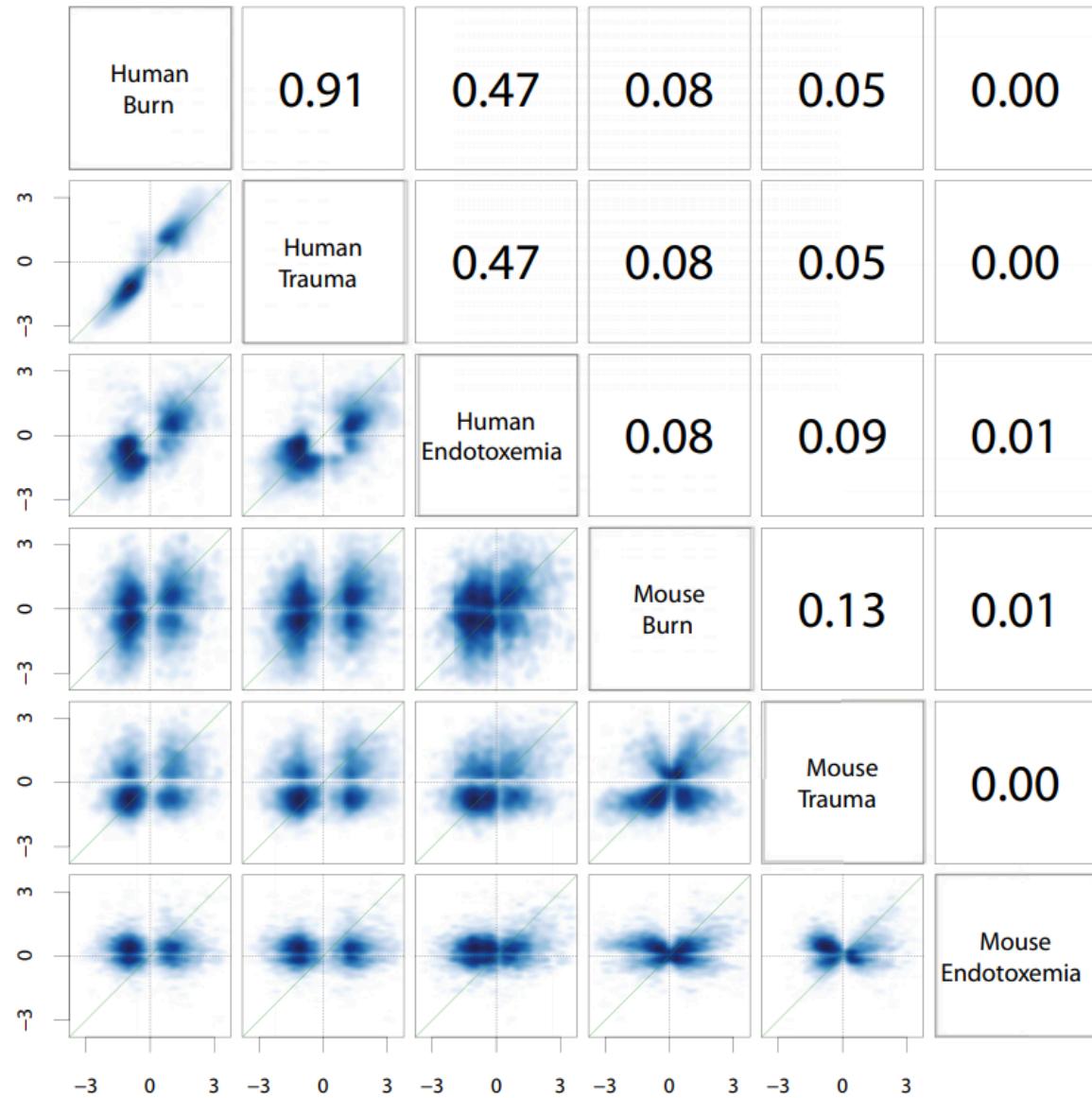
Tale of two papers



Genomic responses in mouse models poorly mimic human inflammatory diseases

Junhee Seok^{a,1}, H. Shaw Warren^{b,1}, Alex G. Cuenca^{c,1}, Michael N. Mindrinos^a, Henry V. Baker^c, Weihong Xu^a, Daniel R. Richards^d, Grace P. McDonald-Smith^e, Hong Gao^a, Laura Hennessy^f, Celeste C. Finnerty^g, Cecilia M. López^h, Shari Honari^f, Ernest E. Moore^h, Joseph P. Mineiⁱ, Joseph Cuschieri^j, Paul E. Bankey^k, Jeffrey L. Johnson^h, Jason Speer^h

Tale of two papers



Tale of two papers



Genomic responses in mouse models poorly mimic human inflammatory diseases

Junhee Seok^{a,1}, H. Shaw Warren^{b,1}, Alex G. Cuenca^{c,1}, Michael N. Mindrinos^a, Henry V. Baker^c, Weihong Xu^a, Daniel R. Richards^d, Grace P. McDonald-Smith^e, Hong Gao^a, Laura Hennessy^f, Celeste C. Finnerty^g, Cecilia M. López^h, Shari Honari^f, Ernest E. Moore^h, Joseph P. Mineiⁱ, Joseph Cuschieri^j, Paul E. Bankey^k, Jeffrey L. Johnson^h, Jason Speer^l



Genomic responses in mouse models greatly mimic human inflammatory diseases

Keizo Takao^{a,b} and Tsuyoshi Miyakawa^{a,b,c,1}

^aSection of Behavior Patterns, Center for Genetic Analysis of Behavior, National Institute for Physiological Sciences, Okazaki, Aichi 444-8585, Japan;

^bCore Research for Evolutional Science and Technology, Japan Science and Technology Agency, Kawaguchi, Saitama 332-0012, Japan; and ^cDivision of Systems Medical Science, Institute for Comprehensive Medical Science, Fujita Health University, Toyoake, Aichi 470-1192, Japan

Tale of two papers



Genomic responses in mouse models poorly mimic human inflammatory diseases

Junhee Seok^{a,1}, H. Shaw Warren^{b,1}, Alex G. Cuenca^{c,1}, Michael N. Mindrinos^a, Henry V. Baker^c, Weihong Xu^a, Daniel R. Richards^d, Grace P. McDonald-Smith^e, Hong Gao^a, Laura Hennessy^f, Celeste C. Finnerty^g, Cecilia M. López^h, Shari Honari^f, Ernest E. Moore^h, Joseph P. Mineiⁱ, Joseph Cuschieri^j, Paul E. Bankey^k, Jeffrey L. Johnson^h, Jason Speer^l



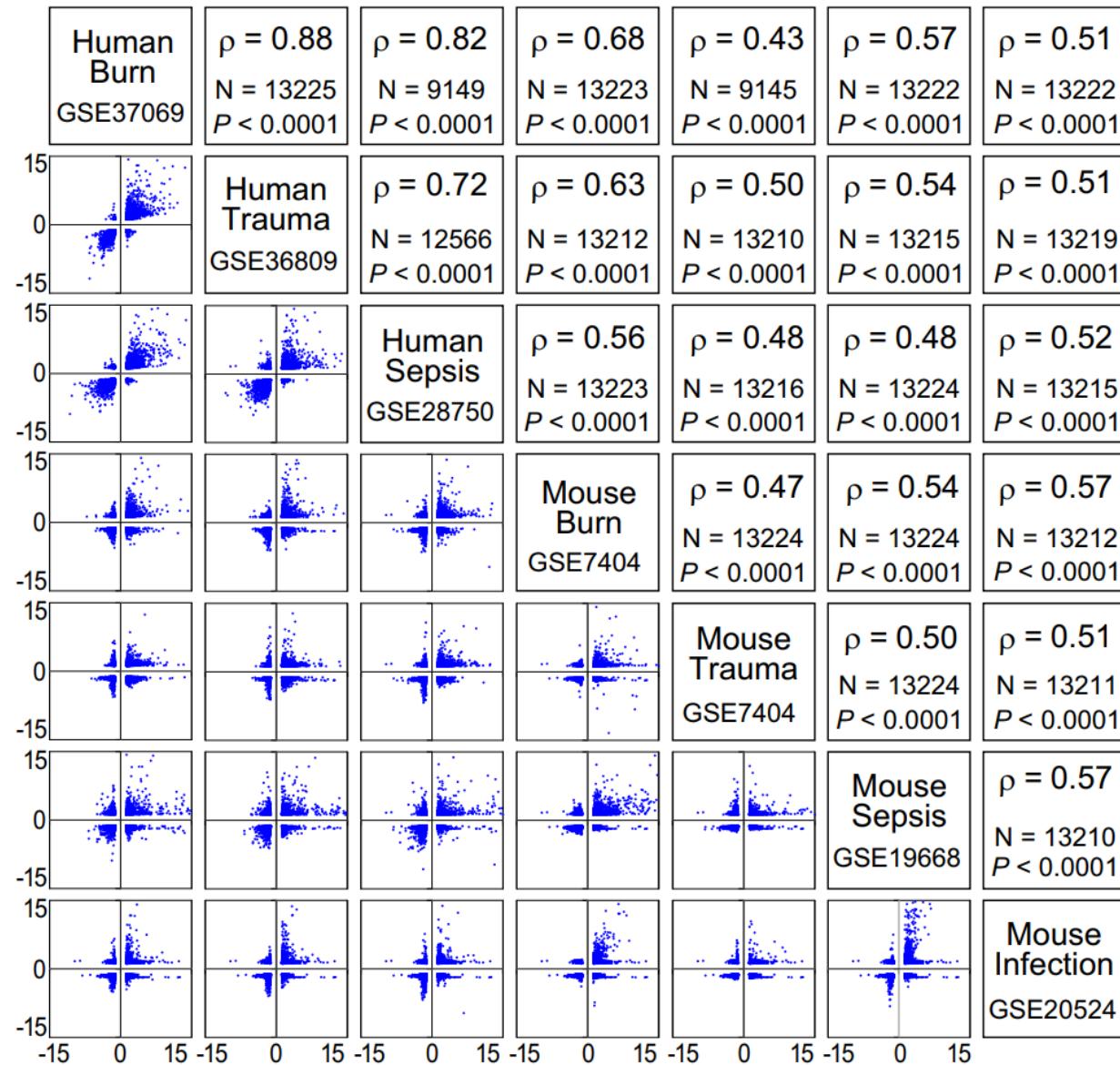
Genomic responses in mouse models **greatly** mimic human inflammatory diseases

Keizo Takao^{a,b} and Tsuyoshi Miyakawa^{a,b,c,1}

^aSection of Behavior Patterns, Center for Genetic Analysis of Behavior, National Institute for Physiological Sciences, Okazaki, Aichi 444-8585, Japan;

^bCore Research for Evolutional Science and Technology, Japan Science and Technology Agency, Kawaguchi, Saitama 332-0012, Japan; and ^cDivision of Systems Medical Science, Institute for Comprehensive Medical Science, Fujita Health University, Toyoake, Aichi 470-1192, Japan

Tale of two papers



Lessons learned

- *A lot* depends on how you analyze your data
- This in turn depends on the questions you ask
- The average “Methods” section is not sufficient for reproducible science!

Attempt to replicate 53 high-impact cancer biology papers:

” Second, none of the 193 experiments were described in sufficient detail in the original paper to enable us to design protocols to repeat the experiments, so we had to seek clarifications from the original authors.” (Errington et al., 2021)

Errington TM, Mathur M, Soderberg CK, Denis A, Perfito N, Iorns E, Nosek BA. Investigating the replicability of preclinical cancer biology. *Elife*. 2021 Dec 10;10:e71601.

Excel and gene names

The screenshot shows a research article from PLOS Computational Biology. The title of the article is "Scientists rename human genes to stop Microsoft Excel from misreading them as dates". The article is categorized as a "RESEARCH ARTICLE" and is marked as "OPEN ACCESS" and "PEER-REVIEWED". The lead authors are Mandhri Abeysooriya, Megan Soria, Mary Sravya Kasu, and Mark Ziemann. The article was published on July 30, 2021. A red oval highlights the title "Gene name errors: Lessons not learned". Another red oval highlights the publication date "July 30, 2021". The illustration at the bottom left is by Alex Castro / The Verge.

Home > Genome Biology > Article

Scientists rename human genes to stop Microsoft Excel from misreading them as dates

PLOS COMPUTATIONAL BIOLOGY

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Gene name errors: Lessons not learned

Mandhri Abeysooriya, Megan Soria, Mary Sravya Kasu, Mark Ziemann

Version 2 Published: July 30, 2021 • <https://doi.org/10.1371/journal.pcbi.1008984>

Illustration by Alex Castro / The Verge

If you buy something from a Verge link, Vox Media may earn a commission. [See our ethics statement.](#)

Thank you

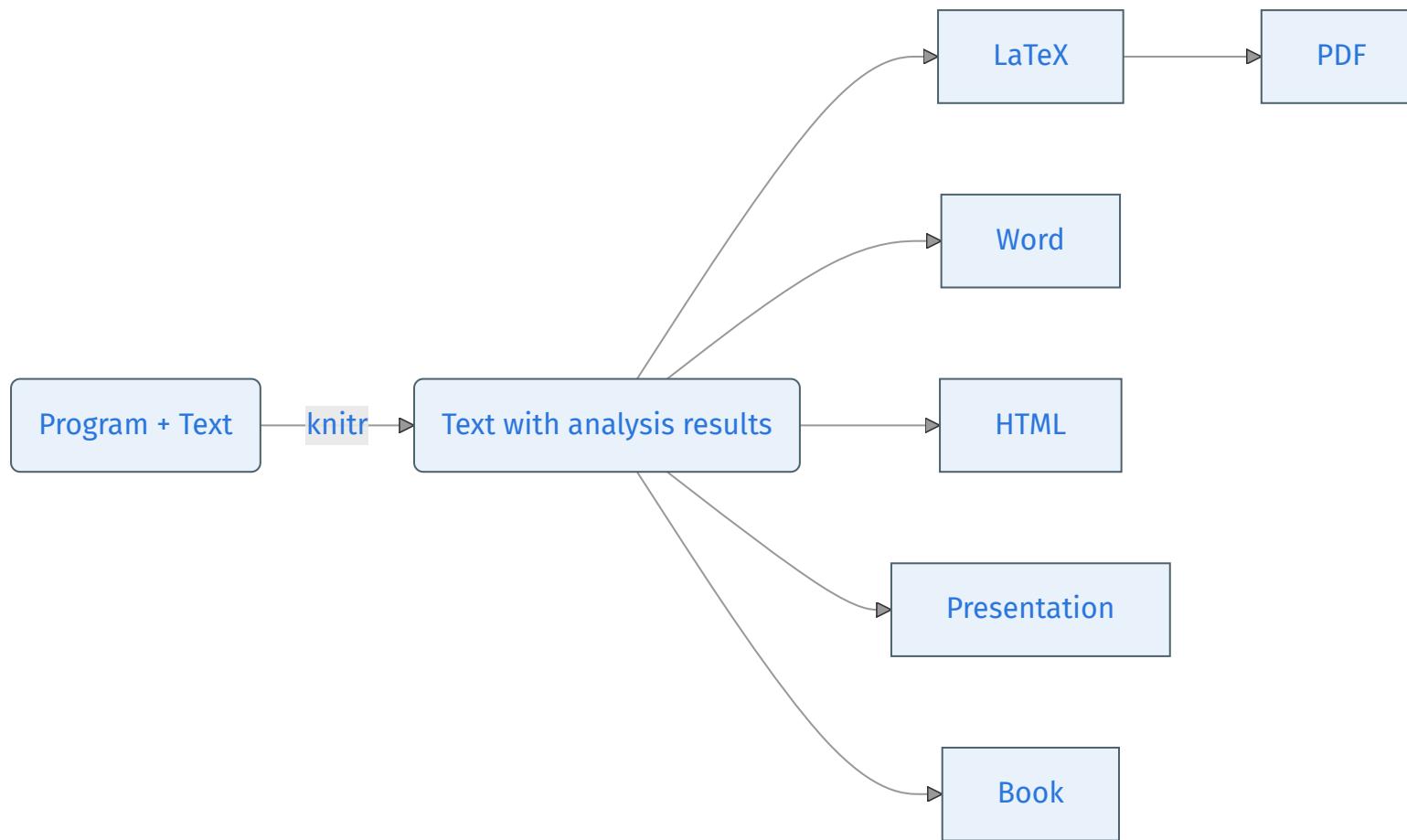
You can find a longer presentation along its source code at
<https://github.com/bihealth/howtotalk>

Parts of this have been expanded to a longer text which can be found at
<https://bihealth.github.io/howtotalk-book/>

A 5 day R crash course book is available at
<https://bihealth.github.io/RCrashcourse-book/>



Reproducible workflows with Rmarkdown



This can be Rmarkdown, Quarto, Jupyter... the goal is that your code and your text are in one place, and the results of your calculations are entered automatically into the text.

Reproducible workflows with Rmarkdown

In systems such as R markdown, you can put directly your analysis results in your text. For example, when I write that the *p*-value is equal to 0.05, I am writing this:

- 1 In systems such as R markdown, you can put directly your
- 2 analysis results in your text. For example, when I write that the
- 3 \$p\$-value is equal to `r p`, I am writing this:

The *p*-value above is not entered manually (as 0.05), but is the result of a statistical computation. If the data changes, if your analysis changes, the *p*-value above will automatically change as well.

Identifiers

- Never use “pure numerical” identifiers (1, 2, 3, ...)
- Never remove columns with “unneeded” identifiers or you can get “involuntary anonymization”
- If your identifier is, say, S1, then always refer to that sample as S1, not Smp. 1 or 1 or Sample 1
- Better use a unique prefix for a study / cohort / experiment, like RCDB2024_S1, RCDB2024_S2, RCDB2025_S1, RCDB2025_S2
- Composite identifiers are fine, as long as you use them consistently: WT_treatment_1, KO_control_2 and not WT_treatment_1, KO-2-control, WT_ctrl12

How (not to) work with Excel

Avoid manually modifying Excel files

- Manual changes cannot be tracked automatically
- You have to record every change you make
- Otherwise, this is not reproducible science!

How (not to) work with Excel

Never use formatting for data

Never encode information as formatting, always use explicit columns

Color / font size / font style cannot be read automatically

A	B
1	Sample ID (red=treated)
2	Sample 1
3	Sample 2
4	Sample 3
5	Sample 4
6	Sample 5
7	Sample 6
8	Sample 7
9	
10	

How (not to) work with Excel

Don't combine values and comments

Make a separate column for comments

Otherwise the values might be lost¹

No:

	A	B
1	Sample	Measurement
2	Sample 1	10
3	Sample 2	20 (morning)
4	Sample 3	30
5	Sample 4	40
6	Sample 5	99 (out of range)
7		

Yes:

	A	B	C
1	Sample	Measurement	Comment
2	Sample 1		10
3	Sample 2		20 morning
4	Sample 3		30
5	Sample 4		40
6	Sample 5		99 out of range
7			

How (not to) work with Excel

Don't put meta-information into column names

Make a separate excel sheet for column meta information

	A	B	C
1	Sample - standard identifier from the Redcap db	Measurement (ug/ml; using the standard plate reader)	
2	Sample 1	11	
3	Sample 2	13	
4	Sample 3	28	
5	Sample 4	1.5	
6	Sample 5	32	
7			
8			

The statistical testing roulette



Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction

Craig M. Bennett¹, Abigail A. Baird², Michael B. Miller¹, and George L. Wolford³

¹ Psychology Department, University of California Santa Barbara, Santa Barbara, CA; ² Department of Psychology, Vassar College, Poughkeepsie, NY;

³ Department of Psychological & Brain Sciences, Dartmouth College, Hanover, NH

INTRODUCTION

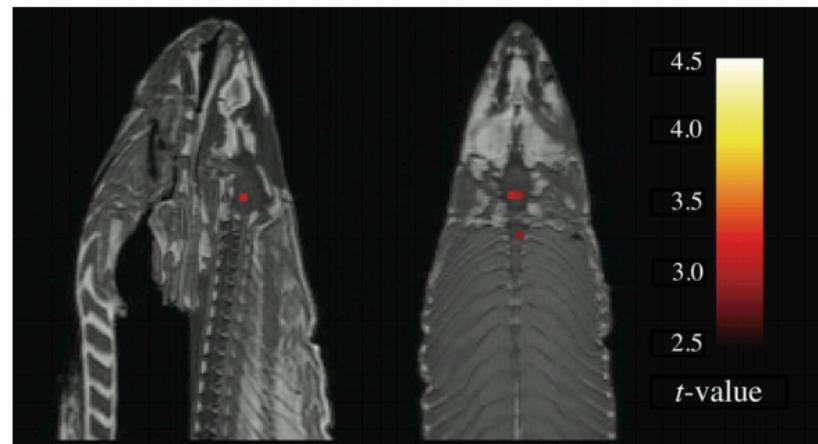
With the extreme dimensionality of functional neuroimaging data comes extreme risk for false positives. Across the 130,000 voxels in a typical fMRI volume the probability of a false positive is almost certain. Correction for multiple comparisons should be completed with these datasets, but is often ignored by investigators. To illustrate the magnitude of the problem we carried out a real experiment that demonstrates the danger of not correcting for chance properly.

METHODS

Subject. One mature Atlantic Salmon (*Salmo salar*) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning.

Task. The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence. The salmon was

GLM RESULTS

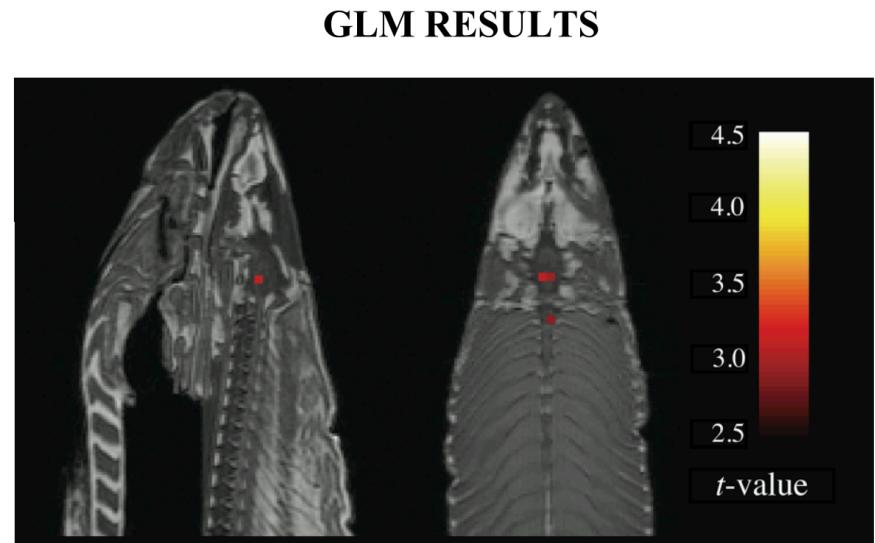


A *t*-contrast was used to test for regions with significant BOLD signal change during the photo condition compared to rest. The parameters for this comparison were $t(131) > 3.15$, $p(\text{uncorrected}) < 0.001$, 3 voxel extent threshold.

The statistical testing roulette

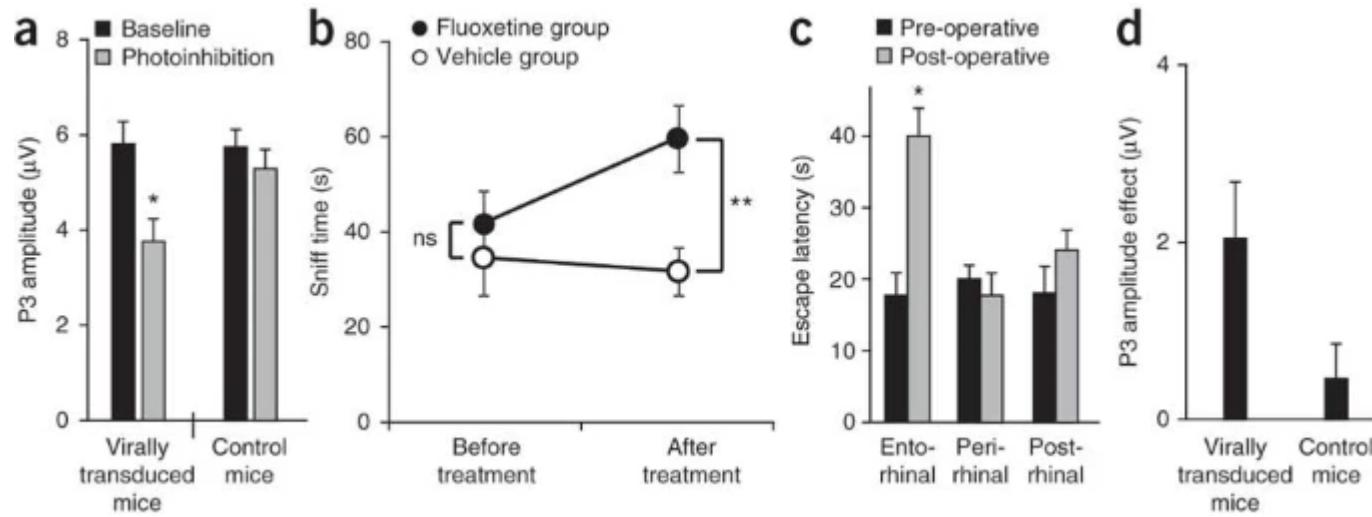
Subject. One mature Atlantic Salmon (*Salmo salar*) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning.

Task. The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been experiencing.

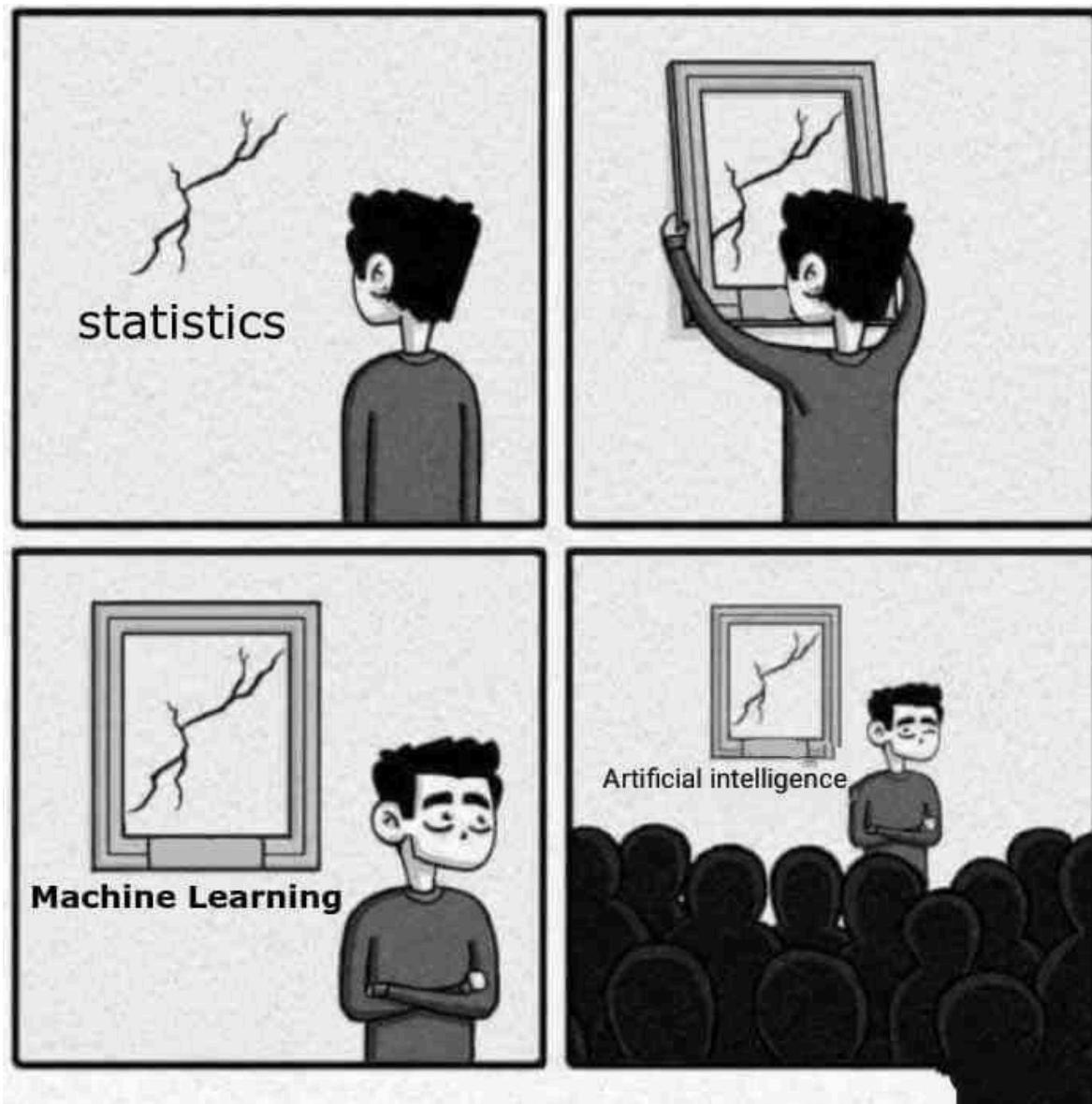


The error is widespread

Nieuwenhuis et al. found that half of the scientists who could have committed this error, did in fact commit this error.



Will “AI” change the field?



- New deep learning methods are useful, but hard to use
- Some of them are truly revolutionizing the field
- There is still place for simpler ML algorithms
- Ready to use LLMs (ChatGPT & Co.) have their use, but also limitations

