How to Lie with Statistics

April 4, 2019

Data Science CSCI 1951A

Brown University
Instructor: Ellie Pavlick
HTAs: Wennie Zhang, Maulik Dang, Gurnaaz Kaur

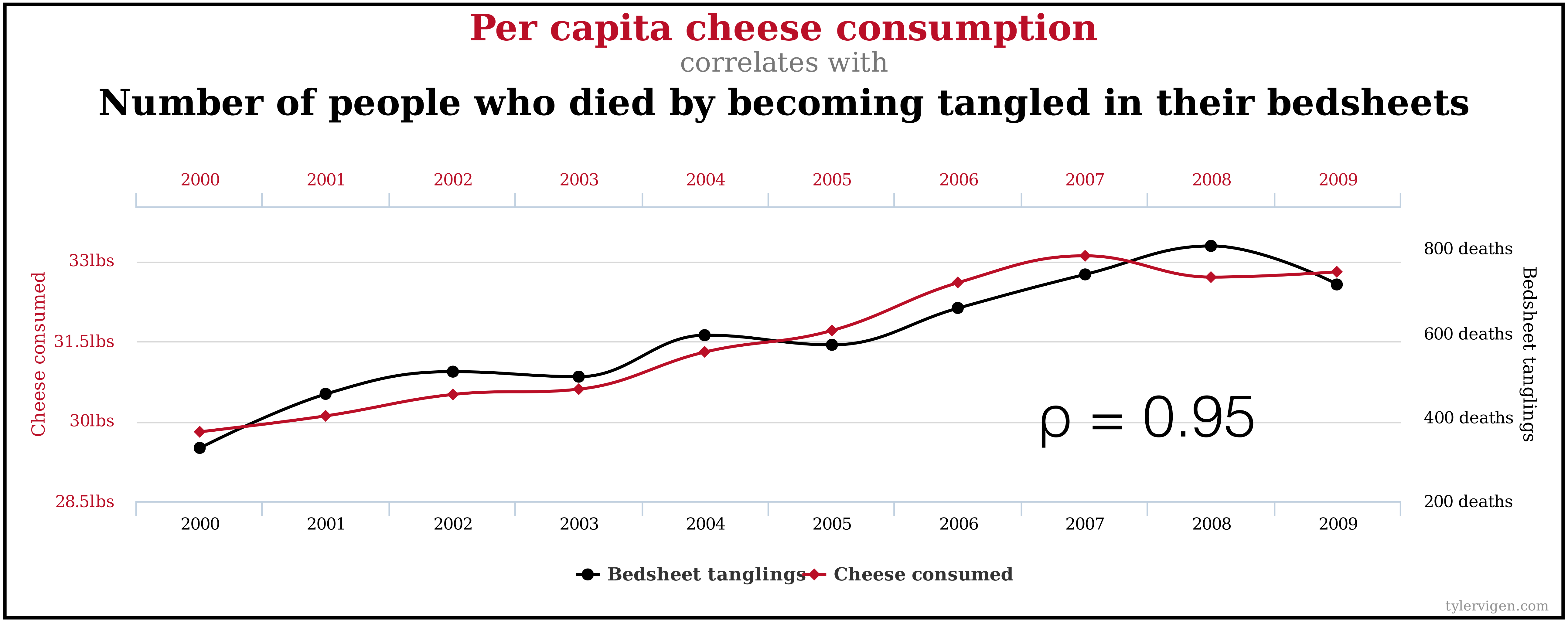
Announcements

- Sundry final project updates: expectations, posters, grading
- Cloud/grid space

Today

- Finish D3 demo
- P-Hacking, Researcher Degrees of Freedom
- (Hopefully) In-Class activity

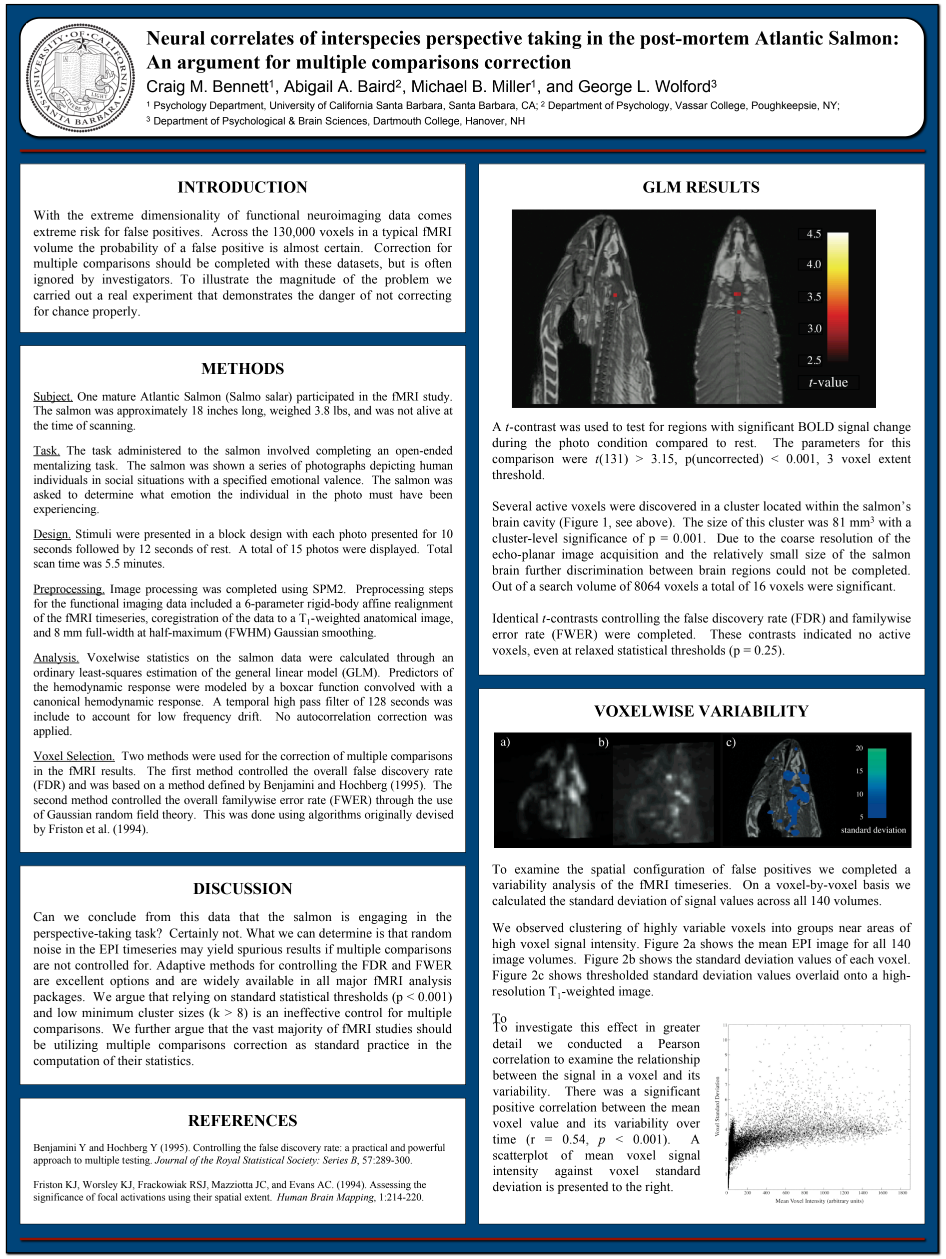
You can find almost anything if you look hard enough.



https://en.wikipedia.org/wiki/Data_dredging

http://www.tylervigen.com/spurious-correlations

You can find almost anything if you look hard enough.



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

Craig M. Bennett[1], Abigail A. Baird[2], Michael B. Miller[1], and George L. Wolford[3]

[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 asked to determine what emotion the individual in the photo must have been experiencing.

Design. Stimuli were presented in a block design with each photo presented for 10 seconds followed by 12 seconds of rest. A total of 15 photos were displayed. Total scan time was 5.5 minutes.

Preprocessing. Image processing was completed using SPM2. Preprocessing steps for the functional imaging data included a 6-parameter rigid-body affine realignment of the fMRI timeseries, coregistration of the data to a T_1-weighted anatomical image, and 8 mm full-width at half-maximum (FWHM) Gaussian smoothing.

Analysis. Voxelwise statistics on the salmon data were calculated through an ordinary least-squares estimation of the general linear model (GLM). Predictors of the hemodynamic response were modeled by a boxcar function convolved with a canonical hemodynamic response. A temporal high pass filter of 128 seconds was include to account for low frequency drift. No autocorrelation correction was applied.

Voxel Selection. Two methods were used for the correction of multiple comparisons in the fMRI results. The first method controlled the overall false discovery rate (FDR) and was based on a method defined by Benjamini and Hochberg (1995). The second method controlled the overall familywise error rate (FWER) through the use of Gaussian random field theory. This was done using algorithms originally devised by Friston et al. (1994).

DISCUSSION

Can we conclude from this data that the salmon is engaging in the perspective-taking task? Certainly not. What we can determine is that random noise in the EPI timeseries may yield spurious results if multiple comparisons are not controlled for. Adaptive methods for controlling the FDR and FWER are excellent options and are widely available in all major fMRI analysis packages. We argue that relying on standard statistical thresholds ($p < 0.001$) and low minimum cluster sizes ($k > 8$) is an ineffective control for multiple comparisons. We further argue that the vast majority of fMRI studies should be utilizing multiple comparisons correction as standard practice in the computation of their statistics.

REFERENCES

Benjamini Y and Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57:289-300.

Friston KJ, Worsley KJ, Frackowiak RSJ, Mazziotta JC, and Evans AC. (1994). Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping*, 1:214-220.

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.

Several active voxels were discovered in a cluster located within the salmon's brain cavity (Figure 1, see above). The size of this cluster was 81 mm^3 with a cluster-level significance of $p = 0.001$. Due to the coarse resolution of the echo-planar image acquisition and the relatively small size of the salmon brain further discrimination between brain regions could not be completed. Out of a search volume of 8064 voxels a total of 16 voxels were significant.

Identical t-contrasts controlling the false discovery rate (FDR) and familywise error rate (FWER) were completed. These contrasts indicated no active voxels, even at relaxed statistical thresholds ($p = 0.25$).

VOXELWISE VARIABILITY

To examine the spatial configuration of false positives we completed a variability analysis of the fMRI timeseries. On a voxel-by-voxel basis we calculated the standard deviation of signal values across all 140 volumes.

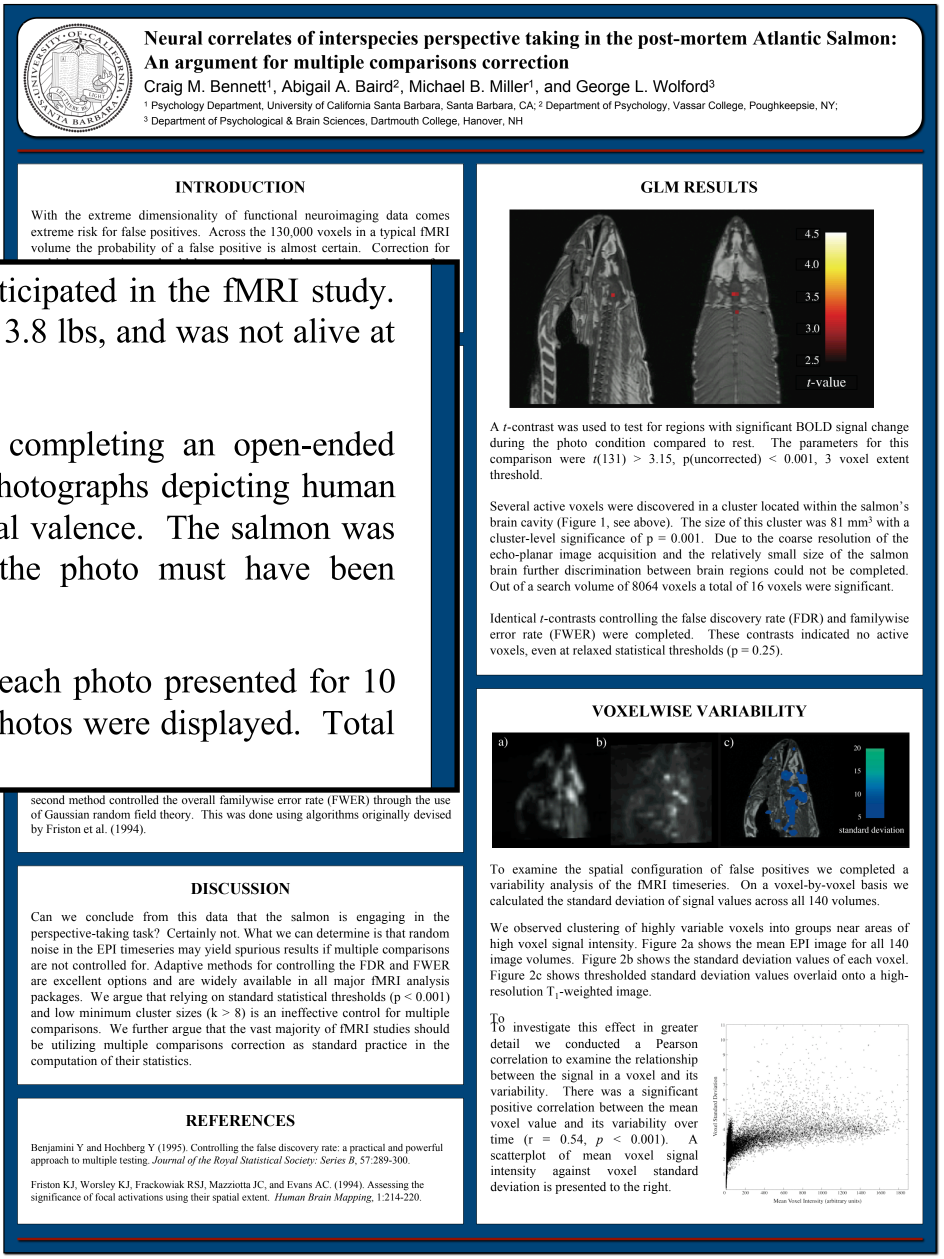
We observed clustering of highly variable voxels into groups near areas of high voxel signal intensity. Figure 2a shows the mean EPI image for all 140 image volumes. Figure 2b shows the standard deviation values of each voxel. Figure 2c shows thresholded standard deviation values overlaid onto a high-resolution T_1-weighted image.

To

To investigate this effect in greater detail we conducted a Pearson correlation to examine the relationship between the signal in a voxel and its variability. There was a significant positive correlation between the mean voxel value and its variability over time ($r = 0.54$, $p < 0.001$). A scatterplot of mean voxel signal intensity against voxel standard deviation is presented to the right.

Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon

You can find almost anything if you look hard enough.


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

Craig M. Bennett[1], Abigail A. Baird[2], Michael B. Miller[1], and George L. Wolford[3]

[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

second method controlled the overall familywise error rate (FWER) through the use of Gaussian random field theory. This was done using algorithms originally devised by Friston et al. (1994).

DISCUSSION

Can we conclude from this data that the salmon is engaging in the perspective-taking task? Certainly not. What we can determine is that random noise in the EPI timeseries may yield spurious results if multiple comparisons are not controlled for. Adaptive methods for controlling the FDR and FWER are excellent options and are widely available in all major fMRI analysis packages. We argue that relying on standard statistical thresholds ($p < 0.001$) and low minimum cluster sizes ($k > 8$) is an ineffective control for multiple comparisons. We further argue that the vast majority of fMRI studies should be utilizing multiple comparisons correction as standard practice in the computation of their statistics.

REFERENCES

Benjamini Y and Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57:289-300.

Friston KJ, Worsley KJ, Frackowiak RSJ, Mazziotta JC, and Evans AC. (1994). Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping*, 1:214-220.

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.

Several active voxels were discovered in a cluster located within the salmon's brain cavity (Figure 1, see above). The size of this cluster was 81 mm^3 with a cluster-level significance of $p = 0.001$. Due to the coarse resolution of the echo-planar image acquisition and the relatively small size of the salmon brain further discrimination between brain regions could not be completed. Out of a search volume of 8064 voxels a total of 16 voxels were significant.

Identical t-contrasts controlling the false discovery rate (FDR) and familywise error rate (FWER) were completed. These contrasts indicated no active voxels, even at relaxed statistical thresholds ($p = 0.25$).

VOXELWISE VARIABILITY

To examine the spatial configuration of false positives we completed a variability analysis of the fMRI timeseries. On a voxel-by-voxel basis we calculated the standard deviation of signal values across all 140 volumes.

We observed clustering of highly variable voxels into groups near areas of high voxel signal intensity. Figure 2a shows the mean EPI image for all 140 image volumes. Figure 2b shows the standard deviation values of each voxel. Figure 2c shows thresholded standard deviation values overlaid onto a high-resolution T_1-weighted image.

To investigate this effect in greater detail we conducted a Pearson correlation to examine the relationship between the signal in a voxel and its variability. There was a significant positive correlation between the mean voxel value and its variability over time ($r = 0.54$, $p < 0.001$). A scatterplot of mean voxel signal intensity against voxel standard deviation is presented to the right.

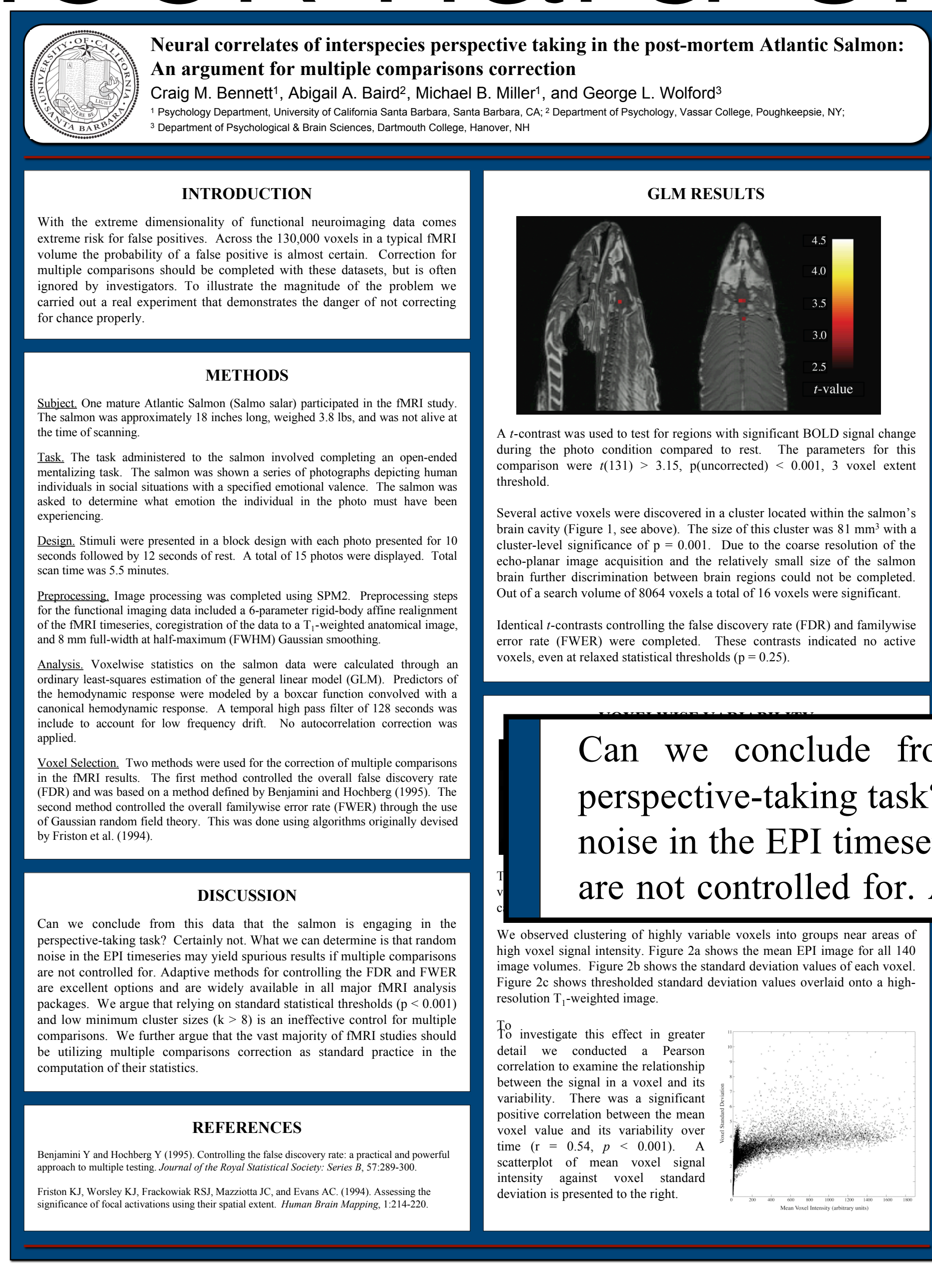
Subject. One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning.

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.

Design. Stimuli were presented in a block design with each photo presented for 10 seconds followed by 12 seconds of rest. A total of 15 photos were displayed. Total scan time was 5.5 minutes.

Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon

You can find almost anything if you look hard enough.



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

Craig M. Bennett[1], Abigail A. Baird[2], Michael B. Miller[1], and George L. Wolford[3]

[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 asked to determine what emotion the individual in the photo must have been experiencing.

Design. Stimuli were presented in a block design with each photo presented for 10 seconds followed by 12 seconds of rest. A total of 15 photos were displayed. Total scan time was 5.5 minutes.

Preprocessing. Image processing was completed using SPM2. Preprocessing steps for the functional imaging data included a 6-parameter rigid-body affine realignment of the fMRI timeseries, coregistration of the data to a T_1-weighted anatomical image, and 8 mm full-width at half-maximum (FWHM) Gaussian smoothing.

Analysis. Voxelwise statistics on the salmon data were calculated through an ordinary least-squares estimation of the general linear model (GLM). Predictors of the hemodynamic response were modeled by a boxcar function convolved with a canonical hemodynamic response. A temporal high pass filter of 128 seconds was include to account for low frequency drift. No autocorrelation correction was applied.

Voxel Selection. Two methods were used for the correction of multiple comparisons in the fMRI results. The first method controlled the overall false discovery rate (FDR) and was based on a method defined by Benjamini and Hochberg (1995). The second method controlled the overall familywise error rate (FWER) through the use of Gaussian random field theory. This was done using algorithms originally devised by Friston et al. (1994).

DISCUSSION

Can we conclude from this data that the salmon is engaging in the perspective-taking task? Certainly not. What we can determine is that random noise in the EPI timeseries may yield spurious results if multiple comparisons are not controlled for. Adaptive methods for controlling the FDR and FWER are excellent options and are widely available in all major fMRI analysis packages. We argue that relying on standard statistical thresholds ($p < 0.001$) and low minimum cluster sizes ($k > 8$) is an ineffective control for multiple comparisons. We further argue that the vast majority of fMRI studies should be utilizing multiple comparisons correction as standard practice in the computation of their statistics.

REFERENCES

Benjamini Y and Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57:289-300.

Friston KJ, Worsley KJ, Frackowiak RSJ, Mazziotta JC, and Evans AC. (1994). Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping*, 1:214-220.

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(uncorrected) < 0.001, 3 voxel extent threshold.

Several active voxels were discovered in a cluster located within the salmon's brain cavity (Figure 1, see above). The size of this cluster was 81 mm^3 with a cluster-level significance of $p = 0.001$. Due to the coarse resolution of the echo-planar image acquisition and the relatively small size of the salmon brain further discrimination between brain regions could not be completed. Out of a search volume of 8064 voxels a total of 16 voxels were significant.

Identical t-contrasts controlling the false discovery rate (FDR) and familywise error rate (FWER) were completed. These contrasts indicated no active voxels, even at relaxed statistical thresholds ($p = 0.25$).

We observed clustering of highly variable voxels into groups near areas of high voxel signal intensity. Figure 2a shows the mean EPI image for all 140 image volumes. Figure 2b shows the standard deviation values of each voxel. Figure 2c shows thresholded standard deviation values overlaid onto a high-resolution T_1-weighted image.

To investigate this effect in greater detail we conducted a Pearson correlation to examine the relationship between the signal in a voxel and its variability. There was a significant positive correlation between the mean voxel value and its variability over time ($r = 0.54$, $p < 0.001$). A scatterplot of mean voxel signal intensity against voxel standard deviation is presented to the right.

Can we conclude from this data that the salmon is engaging in the perspective-taking task? Certainly not. What we can determine is that random noise in the EPI timeseries may yield spurious results if multiple comparisons are not controlled for. Adaptive methods for controlling the FDR and FWER

Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon

Hypothesis Testing (again!)

Hypothesis Testing (again!)

- p-value = Probability of obtaining an effect equal to or more extreme than the one observed, presuming the null hypothesis is true

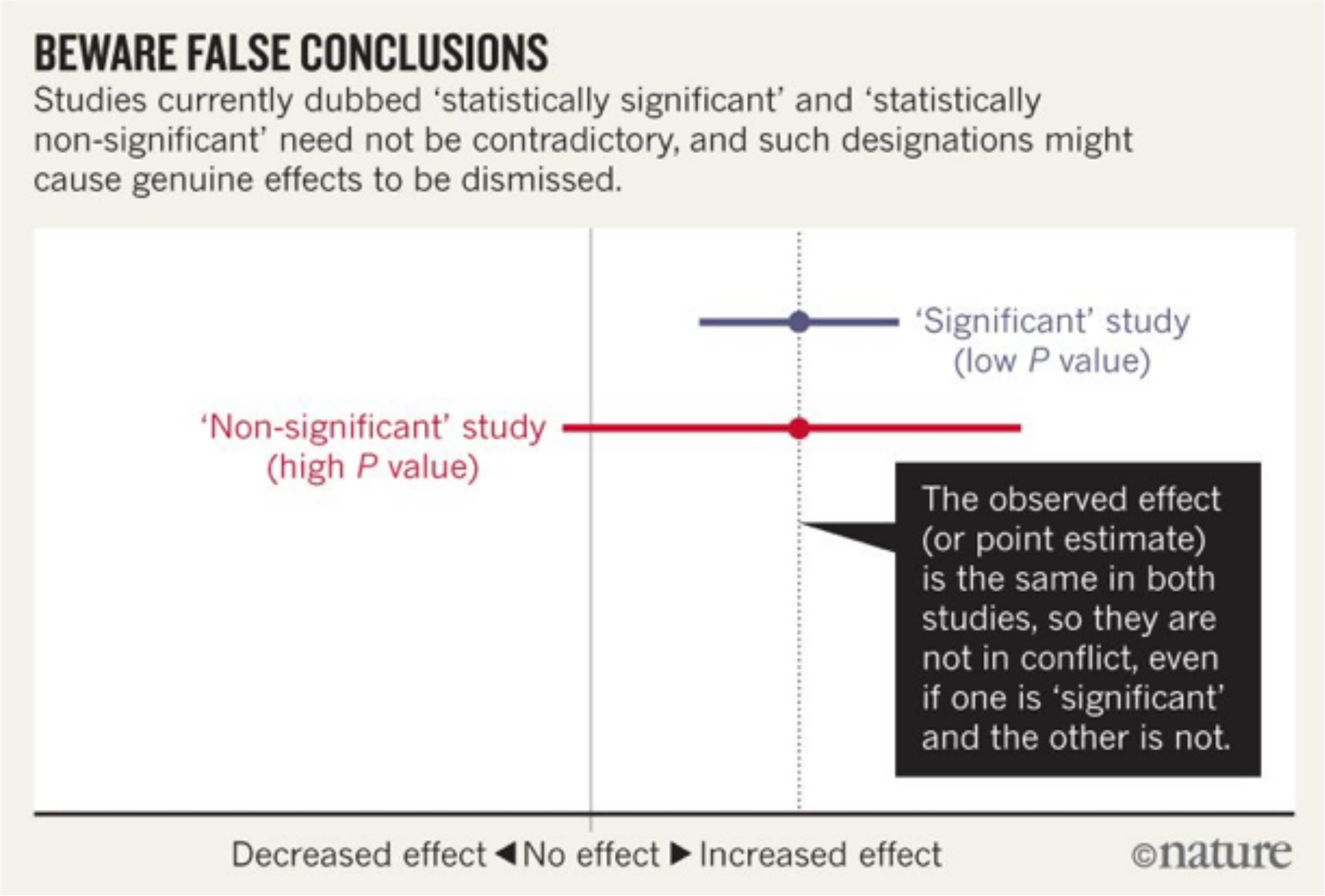
Credit: CS1951A Spring 2017

Hypothesis Testing (again!)

- ☑ p-value = Probability of obtaining an effect equal to or more extreme than the one observed, presuming the null hypothesis is true
- ☐ ***NOT*** the probability that the null or the alternative hypothesis are correct or incorrect

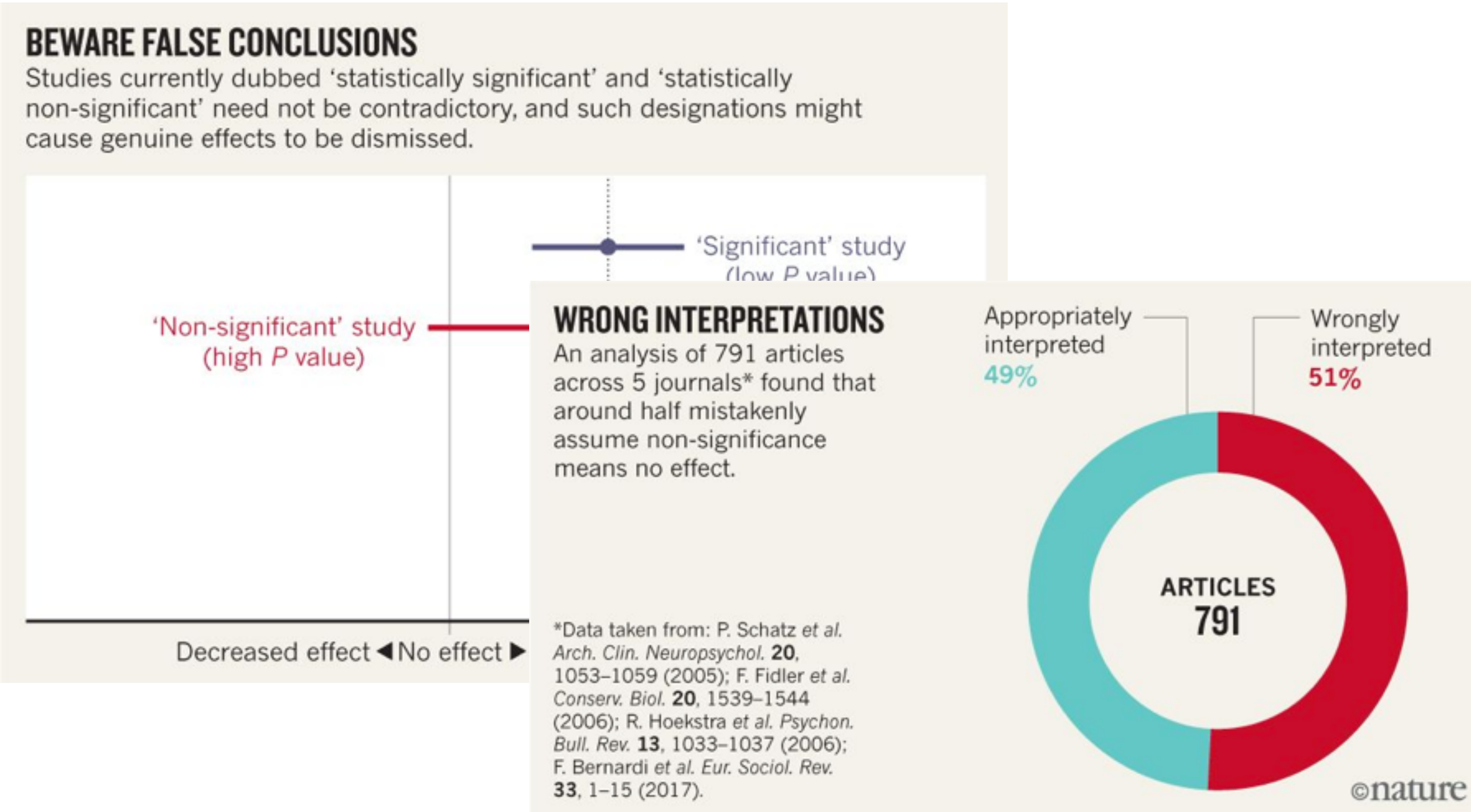
Credit: CS1951A Spring 2017

War on P-Values!



https://www.nature.com/articles/d41586-019-00857-9

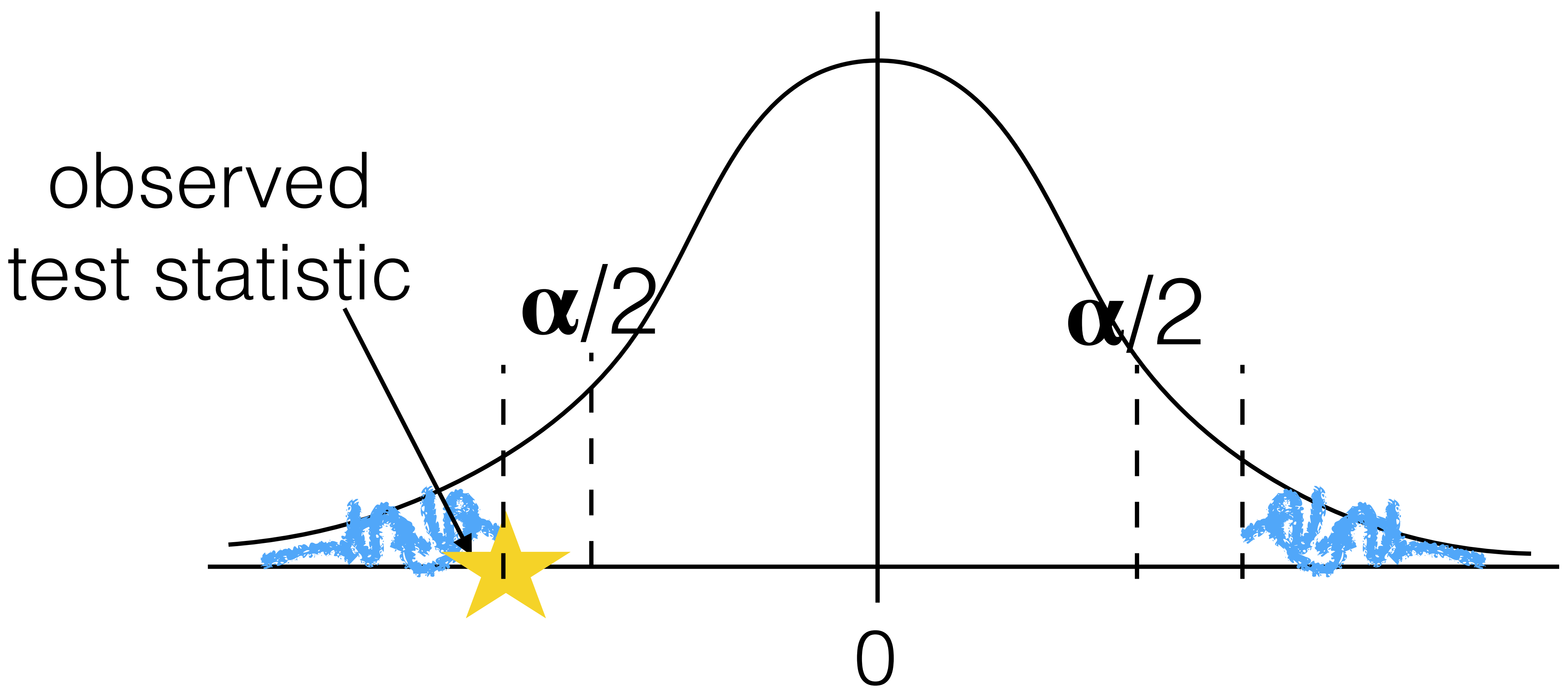
War on P-Values!



https://www.nature.com/articles/d41586-019-00857-9

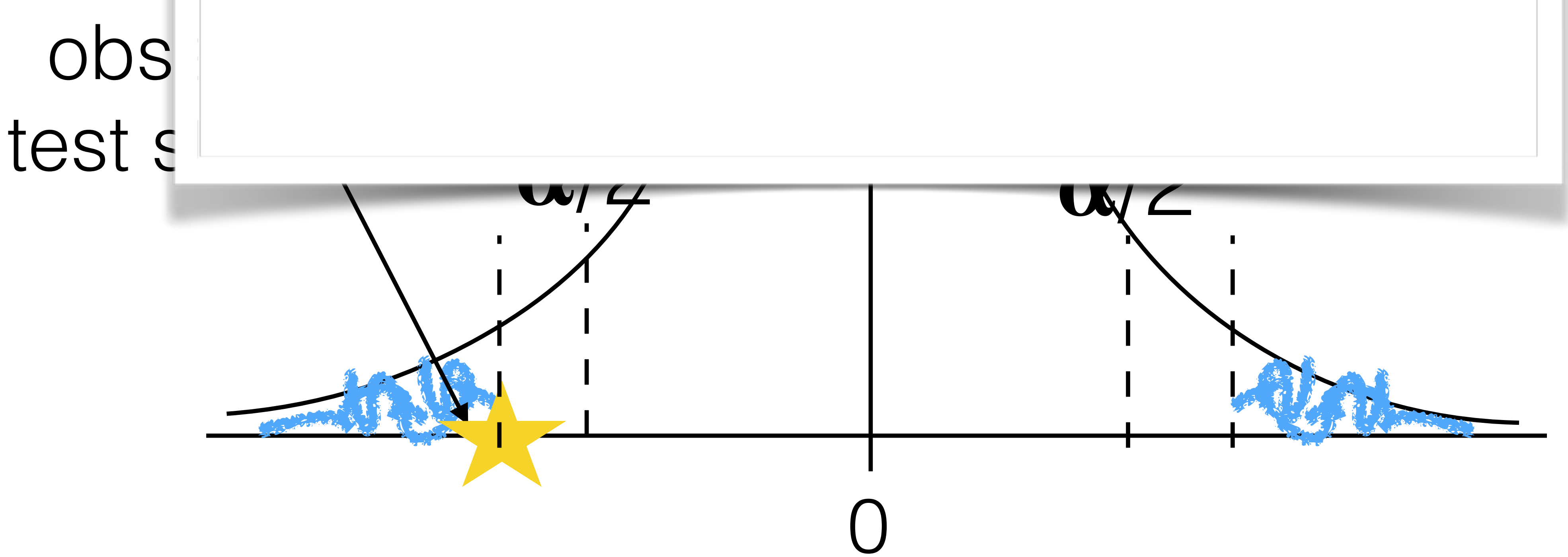
Hypothesis Testing (again!)

p-value = cumulative density of values more extreme than observed statistic



Hypothesis Testing (again!)

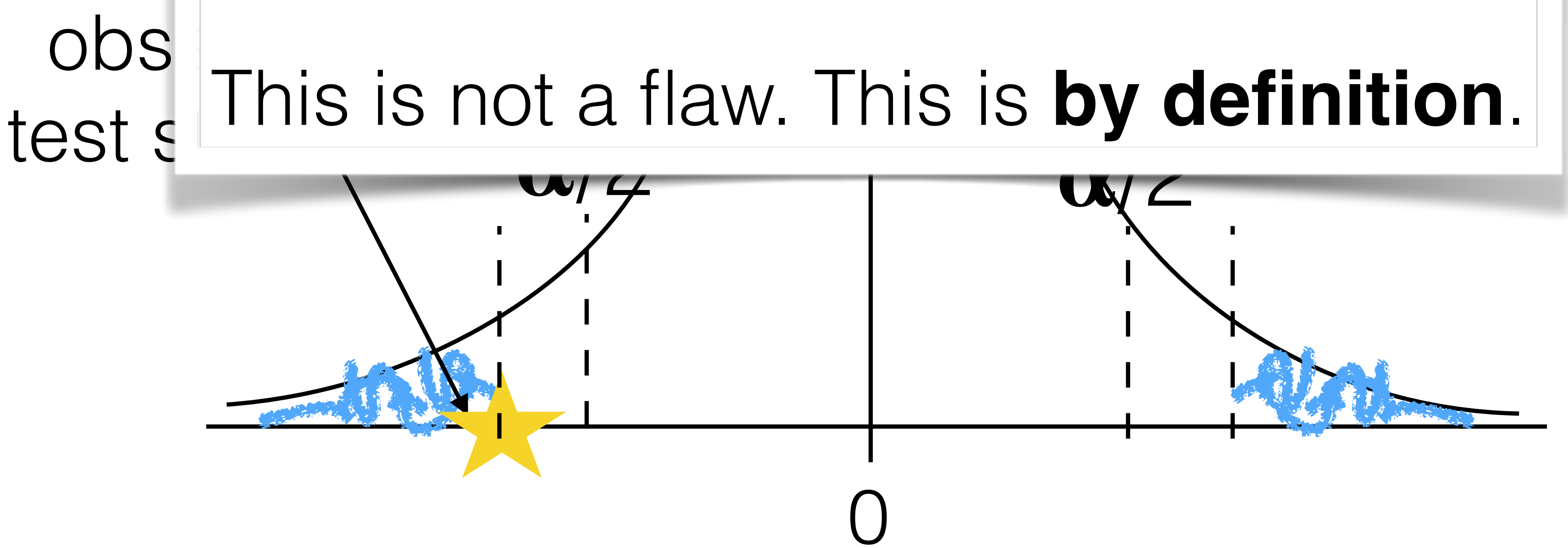
If we run the same test on 100 random samples, we **should expect get a significant effect 100*α times**.



Hypothesis Testing (again!)

If we run the same test on 100 random samples, we **should expect get a significant effect 100*α times**.

This is not a flaw. This is **by definition**.



Multiple Comparisons

Multiple Comparisons

Multiple Comparisons

Historians are all fluff and emotion, scientists are all logic and rigor. I will prove this with my data and my science.

Multiple Comparisons

24,393 discussion posts from "Science and Math" forums

20,575 discussion posts from "History" forums

5,569 "strongly subjective" words, subdivided into categories

For each word, test whether there is a significant difference in its usage between History forums and Science forums

Multiple Comparisons

Crim , You are failing to see the difference between small-scale , verifiable negatives , like the empty box example , and large-scale unverifiable negatives , like the non-existence of god , or extraterrestrial life somewhere in the universe . David Hume is the philosopher who first articulated the idea that you ca n't prove a large-scale unverifiable negative . Given our knowledge of the universe and our lack of the ability to gather information about life-forms in other systems , this is precisely the sort of logical fallacy Hume described . Hume saw a problem with making generalizations based on a limited number of observations . This is called Hume 's problem , and is the basis for the claim that you can not prove or disprove an unverifiable negative .

Screaming just means you 're emotional about your opinion . And the sovereign authority of the state -- i.e. its People , which is the supreme sovereign authority of that state -- may construe that , or any other law , as it pleases regarding its domestic policy . The SC can explicitly state that the world is flat ; but that does n't make it so , since it has no such power over heaven and earth ; and it likewise has no power to grant or deny the international sovereignty of states . It may rule on cases that come before it , and pass them into subordinate case-law ; however this can not affect the actual sovereignty of the states in question , any more than it can make the Earth flat , or make England and France into the 51st and 52nd states..

Multiple Comparisons

Crim , You are failing to see the difference between small-scale , verifiable negatives , like the empty box example , and large-scale unverifiable negatives , like the non-existence of god , or extraterrestrial life somewhere in the universe . David Hume is the philosopher who first articulated the idea that you ca n't prove a large-scale unverifiable negative . Given our knowledge of the universe and our lack of the ability to gather information about life-forms in other systems , this is precisely the sort of logical fallacy Hume described . Hume saw a problem with making generalizations based on a limited number of observations . This is called Hume 's problem , and is the basis for the claim that you can not prove or disprove an unverifiable negative .

Screaming just means you 're emotional about your opinion . And the sovereign authority of the state -- i.e. its People , which is the supreme sovereign authority of that state -- may construe that , or any other law , as it pleases regarding its domestic policy . The SC can explicitly state that the world is flat ; but that does n't make it so , since it has no such power over heaven and earth ; and it likewise has no power to grant or deny the international sovereignty of states . It may rule on cases that come before it , and pass them into subordinate case-law ; however this can not affect the actual sovereignty of the states in question , any more than it can make the Earth flat , or make England and France into the 51st and 52nd states..

Multiple Comparisons

absolute actual actually ambiguous arbitrary attraction beautiful belief believe chaos chaotic coherence confusing contemplate correctly debate difficulty disprove doomsday eternity ethical exact exactly extremely faith false friction fundamental hmm ignorance imagination imagine improbable incapable incredible incredibly insight insulting intelligent interesting irrelevant know knowing knowledge liar love mean moreover must mysterious mystery need okay overcome perfect perfectly pleasure pretty problematic quite rather rational realistic really reject shark sorry star stars suffering super sure surely think tremendous true truth understand virus weird will

aggression alliance alliances ambivalent anger angry atrocities bad beast best blame brutal brutality burden childish contempt courage crusade demonize denial deny desire despotism devastated disagree disastrous dispute domination dramatic evil evils extermination facts fascism fascist fear felt forget genius genocide great greatest greatly greatness greed grievances guilt happiness hero honorable horrible horrific horror hypocrisy hysteria idiocy idiot inevitable inferior insane justification kid knew liberty lie lies mad majesty massacre mentality mess moderate moral morality motivation myth nationalism notorious opinions opposition oppression oppressive partisan patriot patriotic peculiar persecution perverted precious prejudice pride propaganda prosecute protest provoke racist racists radical radicals rebellious revenge ridiculous sacrifice scarcely sentiment sentiments slaves struggle superiority support supporter suppose supremacy sympathy terror traitor traitorous treason tribute tyrannical tyranny tyrant unacceptable unpopular views vital willing worse worst

Multiple Comparisons

absolute actual actually ambiguous arbitrary attraction beautiful belief believe chaos
chaotic coherence confusing contemplate correctly debate difficulty disprove doomsday
eternity ethical exact exactly extremely faith false friction fundamental hmm ignorance
imagination imagine improbable incapable incredible incredibly insight insulting
intelligent interesting irrelevant know knowing knowledge liar love mean moreover must
mysterious mystery need okay overcome perfect perfectly pleasure pretty problematic
quite rather rational realistic really reject shark sorry star stars suffering super sure surely
think tremendous true truth understand virus weird will

81

aggression alliance alliances ambivalent anger angry atrocities bad beast best blame brutal
brutality burden childish contempt courage crusade demonize denial deny desire despotism
devastated disagree disastrous dispute domination dramatic evil evils extermination facts
fascism fascist fear felt forget genius genocide great greatest greatly greatness greed
grievances guilt happiness hero honorable horrible horrific horror hypocrisy hysteria idiocy
idiot inevitable inferior insane justification kid knew liberty lie lies mad majesty massacre
mentality mess moderate moral morality motivation myth nationalism notorious opinions
opposition oppression oppressive partisan patriot patriotic peculiar persecution perverted
precious prejudice pride propaganda prosecute protest provoke racist racists radical radicals
rebellious revenge ridiculous sacrifice scarcely sentiment sentiments slaves struggle
superiority support supporter suppose supremacy sympathy terror traitor traitorous treason
tribute tyrannical tyranny tyrant unacceptable unpopular views vital willing worse worst

129

Multiple Comparisons

24,393 discussion posts from "Science and Math" forums

20,575 discussion posts from "History" forums

5,569 "strongly subjective" words, subdivided into categories

For each word, test whether there is a significant difference in its usage between History forums and Science forums

Multiple Comparisons

24,393 discussion posts from "Science and Math" forums

20,575 discussion posts from "History" forums

5,569 "strongly subjective" words, subdivided into categories

For each word, test whether there is a significant difference in its usage between History forums and Science forums

Multiple Comparisons

24,393 discussion posts from
"Science and Math" forums

20,575 discussion posts from
"History" forums

5,569 "strongly subjective" words,
subdivided into categories

For each word, test whether there
is a significant difference in its
usage between History forums and
Science forums

Multiple Comparisons

$\boldsymbol{\alpha} = 0.05$

(set in advance like good scientists)

Multiple Comparisons

$\alpha = 0.05$

5,569 “strongly subjective” words

We expect 278 of those to show a difference by random chance alone.

210 words showed significant differences in usage between Science and History

Multiple Comparisons

Bonferroni Correction

$p = 0.05 / 5{,}567 = 0.0000089$

Multiple Comparisons

Bonferroni Correction

p = 0.05 / 5,567 = 0.0000089

Stricter p-value to maintain a 5% “false positive” rate

Multiple Comparisons

absolute actual actually ambiguous arbitrary attraction beautiful belief believe chaos chaotic coherence confusing contemplate correctly debate difficulty disprove doomsday eternity ethical exact exactly extremely faith false friction fundamental hmm ignorance imagination imagine improbable incapable incredible incredibly insight insulting intelligent interesting irrelevant know knowing knowledge liar love mean moreover must mysterious mystery need okay overcome perfect perfectly pleasure pretty problematic quite rather rational realistic really reject shark sorry star stars suffering super sure surely think tremendous true truth understand virus weird will

aggression alliance alliances ambivalent anger angry atrocities bad beast best blame brutal brutality burden childish contempt courage crusade demonize denial deny desire despotism devastated disagree disastrous dispute domination dramatic evil evils extermination facts fascism fascist fear felt forget genius genocide great greatest greatly greatness greed grievances guilt happiness hero honorable horrible horrific horror hypocrisy hysteria idiocy idiot inevitable inferior insane justification kid knew liberty lie lies mad majesty massacre mentality mess moderate moral morality motivation myth nationalism notorious opinions opposition oppression oppressive partisan patriot patriotic peculiar persecution perverted precious prejudice pride propaganda prosecute protest provoke racist racists radical radicals rebellious revenge ridiculous sacrifice scarcely sentiment sentiments slaves struggle superiority support supporter suppose supremacy sympathy terror traitor traitorous treason tribute tyrannical tyranny tyrant unacceptable unpopular views vital willing worse worst

Multiple Comparisons

absolute actual actually ambiguous arbitrary attraction beautiful belief believe chaos chaotic coherence confusing contemplate correctly debate difficulty disprove doomsday eternity ethical exact exactly extremely faith false friction fundamental hmm ignorance imagination imagine improbable incapable incredible incredibly insight insulting intelligent interesting irrelevant know knowing knowledge liar love mean moreover must mysterious mystery need okay overcome perfect perfectly pleasure pretty problematic quite rather rational realistic really reject shark sorry star stars suffering super sure surely think tremendous true truth understand virus weird will

aggression alliance alliances ambivalent anger angry atrocities bad beast best blame brutal brutality burden childish contempt courage crusade demonize denial deny desire despotism devastated disagree disastrous dispute domination dramatic evil evils extermination facts fascism fascist fear felt forget genius genocide great greatest greatly greatness greed grievances guilt happiness hero honorable horrible horrific horror hypocrisy hysteria idiocy idiot inevitable inferior insane justification kid knew liberty lie lies mad majesty massacre mentality mess moderate moral morality motivation myth nationalism notorious opinions opposition oppression oppressive partisan patriot patriotic peculiar persecution perverted precious prejudice pride propaganda prosecute protest provoke racist racists radical radicals rebellious revenge ridiculous sacrifice scarcely sentiment sentiments slaves struggle superiority support supporter suppose supremacy sympathy terror traitor traitorous treason tribute tyrannical tyranny tyrant unacceptable unpopular views vital willing worse worst

When am I at risk of “multiple comparisons” errors?

When am I at risk of “multiple comparisons” errors?

- You are literally running the same test multiple times (“tuning the random seed”)

When am I at risk of “multiple comparisons” errors?

- You are literally running the same test multiple times (“tuning the random seed”)
- You are running a large number of experiments and then looking for the ones that are significant after-the-fact

How could I have done this better?

How could I have done this better?

- Perform one test — count total number of subjective words in each population and do a single test for population proportion

How could I have done this better?

- Perform one test — count total number of subjective words in each population and do a single test for population proportion
- What problems could still exist?

Researcher Degrees of Freedom

Researcher Degrees of Freedom

"Researcher degrees of freedom can lead to a multiple comparisons problem, even in settings where researchers perform only a single analysis on their data. The problem is there can be a large number of potential comparisons when the details of data analysis are highly contingent on data, without the researcher having to perform any conscious procedure of fishing or examining multiple p-values."

— Andrew Gelman and Eric Loken

The garden of forking paths: Why multiple comparisons can be a problem… Gelman and Loken (2013).

Researcher Degrees of Freedom

"Researcher degrees of freedom can lead to a multiple comparisons problem, even in settings where researchers perform only a single analysis on their data. The problem is there can be a large number of potential comparisons when the details of data analysis are highly contingent on data, without the researcher having to perform any conscious procedure of fishing or examining multiple p-values."

— Andrew Gelman and Eric Loken

The garden of forking paths: Why multiple comparisons can be a problem… Gelman and Loken (2013).

Researcher Degrees of Freedom

"Researcher degrees of freedom can lead to a multiple comparisons problem, even in settings where researchers perform only a single analysis on their data. The problem is there can be a large number of potential comparisons when the details of data analysis are highly contingent on data, without the researcher having to perform any conscious procedure of fishing or examining multiple p-values."

— Andrew Gelman and Eric Loken

The garden of forking paths: Why multiple comparisons can be a problem… Gelman and Loken (2013).

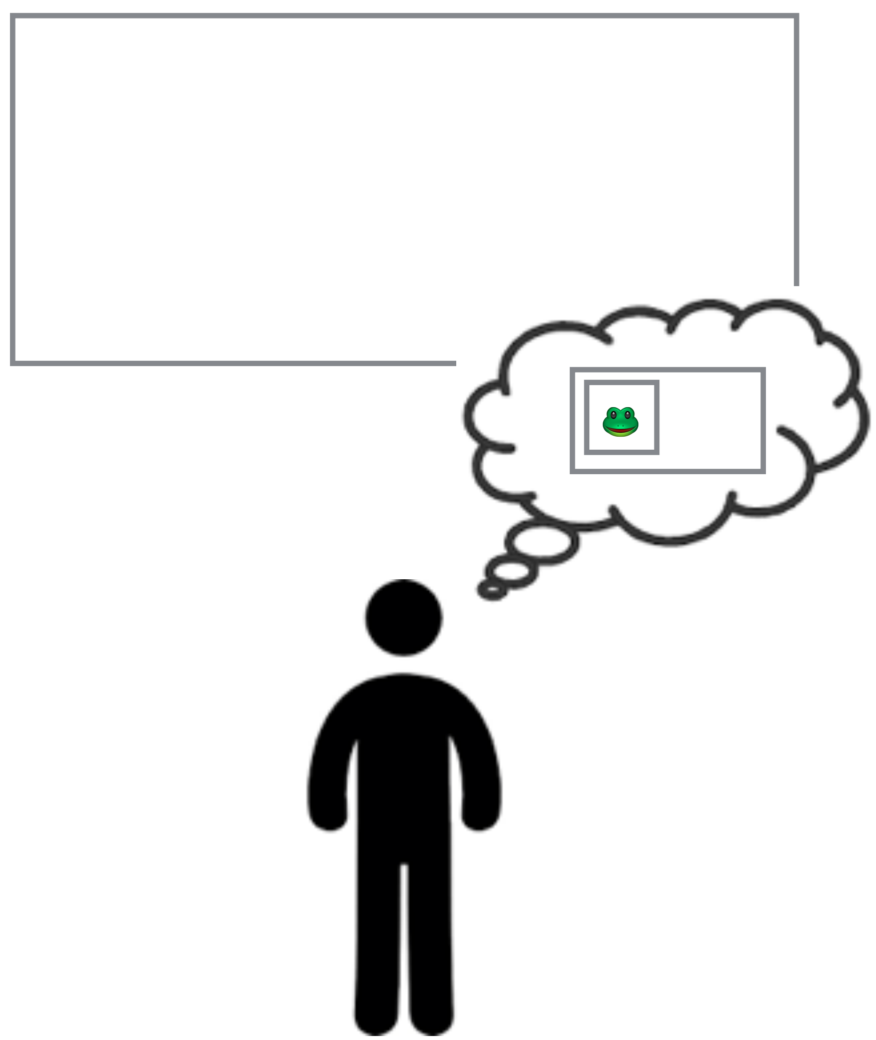
Researcher Degrees of Freedom

Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. Bem (2011).

The garden of forking paths: Why multiple comparisons can be a problem… Gelman and Loken (2013).

Researcher Degrees of Freedom

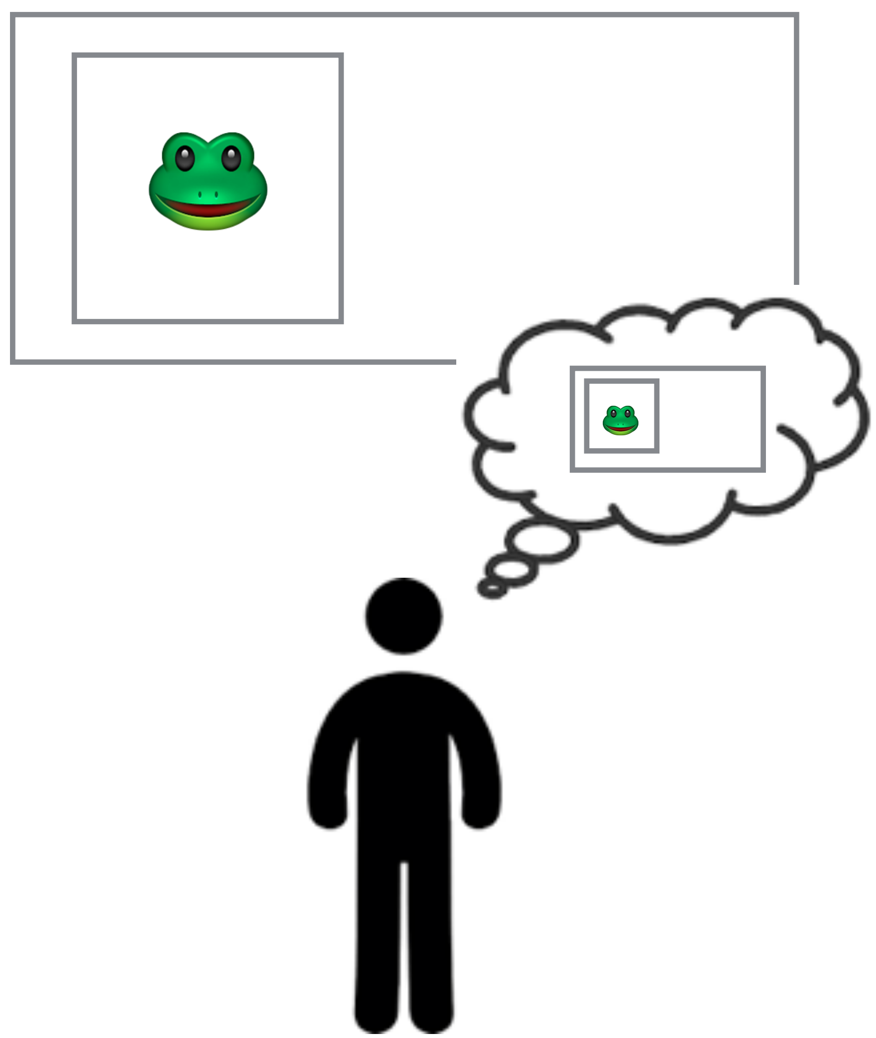
Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. Bem (2011).



The garden of forking paths: Why multiple comparisons can be a problem… Gelman and Loken (2013).

Researcher Degrees of Freedom

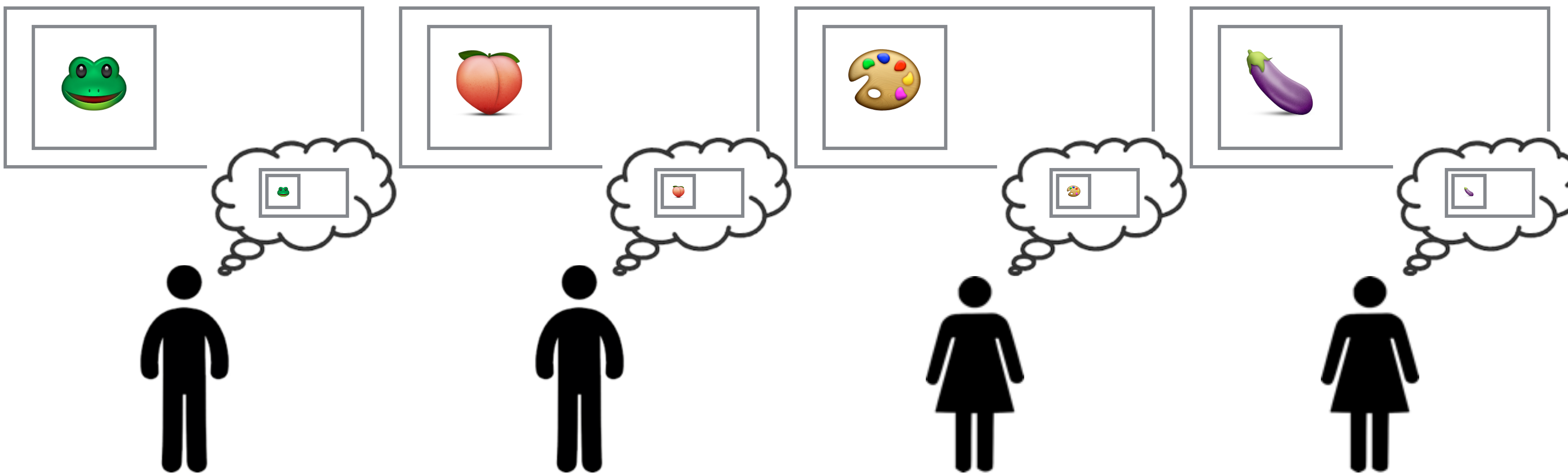
Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. Bem (2011).



The garden of forking paths: Why multiple comparisons can be a problem… Gelman and Loken (2013).

Researcher Degrees of Freedom

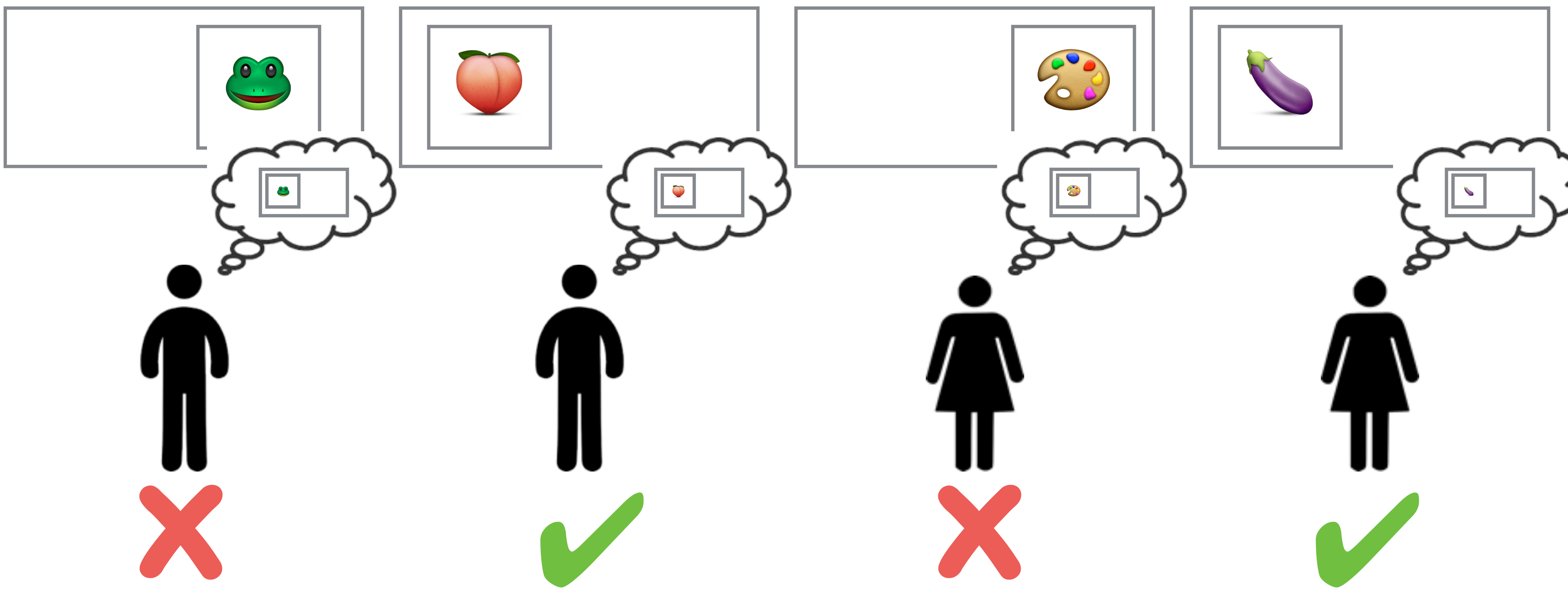
Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. Bem (2011).



The garden of forking paths: Why multiple comparisons can be a problem… Gelman and Loken (2013).

Researcher Degrees of Freedom

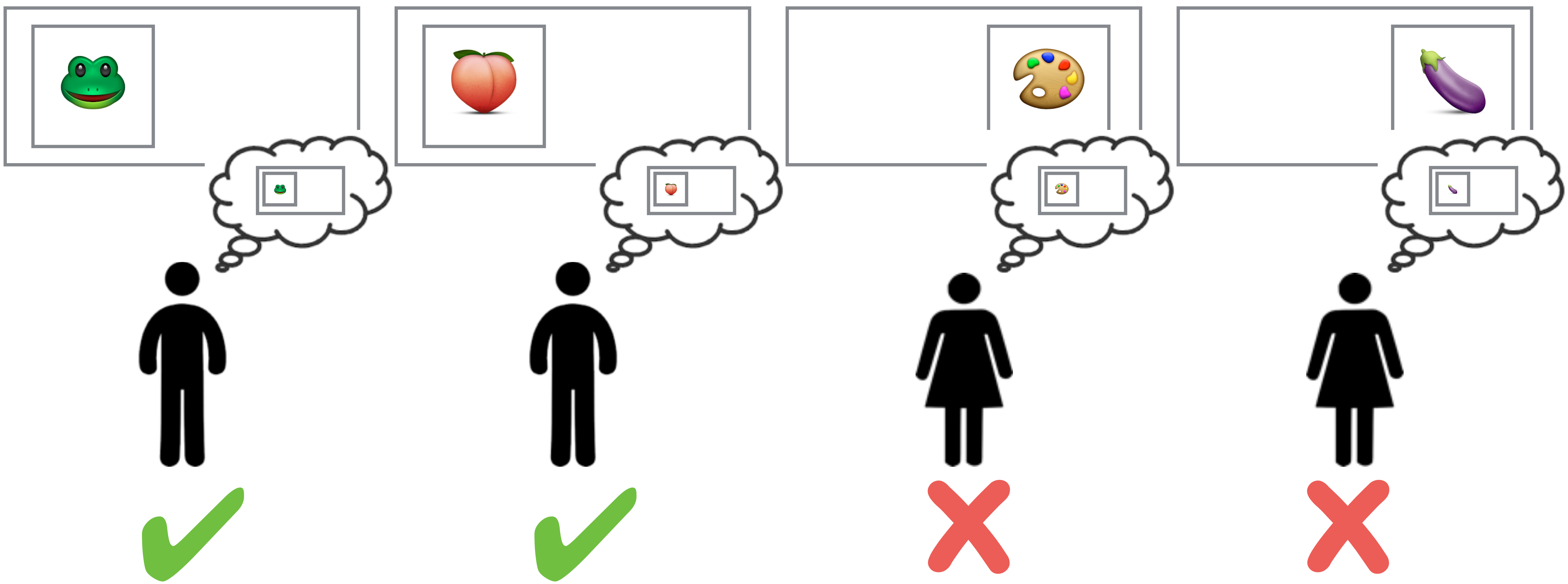
“We show precognitive effects exist for erotic images”



The garden of forking paths: Why multiple comparisons can be a problem… Gelman and Loken (2013).

Researcher Degrees of Freedom

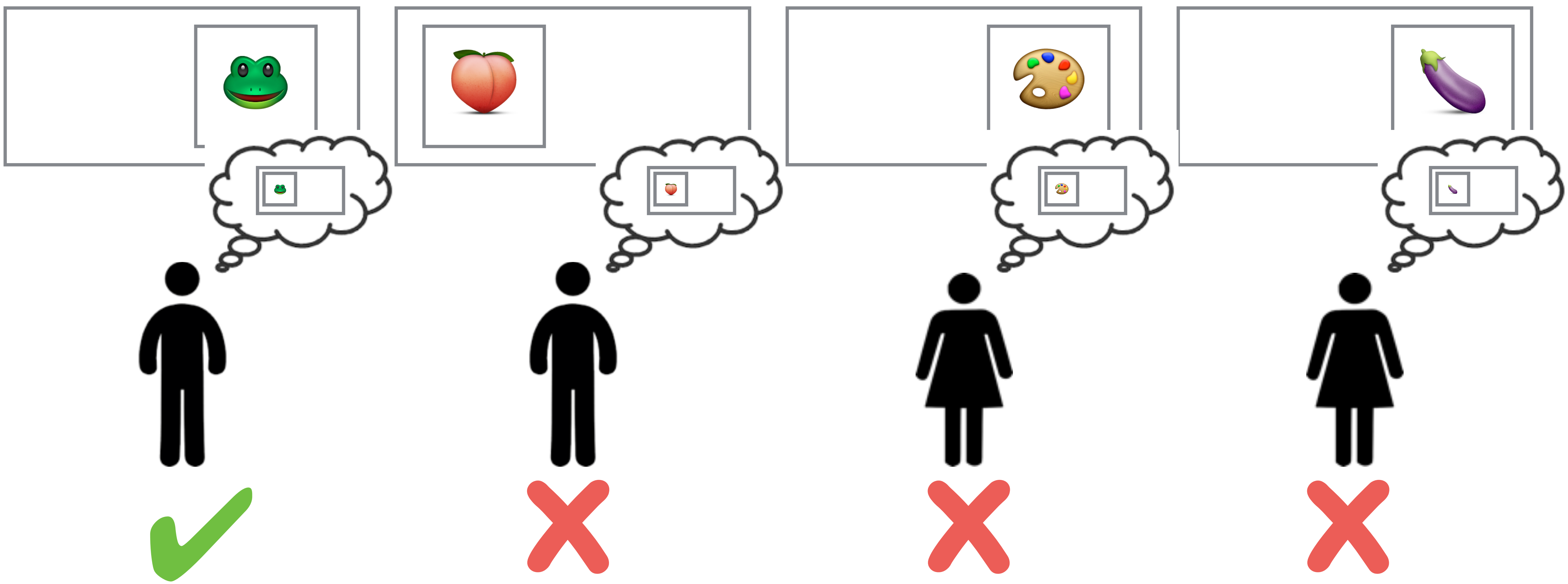
"We show precognitive effects exist in men"



The garden of forking paths: Why multiple comparisons can be a problem… Gelman and Loken (2013).

Researcher Degrees of Freedom

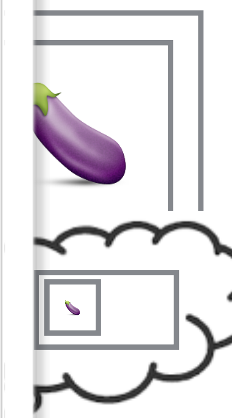
"We show precognitive effects exist in men for frog-related images."



The garden of forking paths: Why multiple comparisons can be a problem… Gelman and Loken (2013).

Researcher Degrees of Freedom

"We are not saying the scientific claims in these papers are necessarily wrong…What we are saying is that the evidence in these research papers is not as strong as stated….To put it another way, we view these papers—despite their statistically significant p-values—as exploratory, and when we look at exploratory results we must be aware of their uncertainty and fragility…."



The garden of forking paths: Why multiple comparisons can be a problem… Gelman and Loken (2013).

Intermediate Task	Avg	CoLA	SST	MRPC	QQP	STS	MNLI	QNLI	RTE	WNLI
ELMo with Intermediate Task Training										
Random^E	70.5	38.5	87.7	79.9/86.5	86.7/83.4	80.8/82.1	75.6	79.6	**61.7**	33.8*
Single-Task^E	71.2	39.4	**90.6**	77.5/84.4	86.4/82.4	79.9/80.6	75.6	78.0	55.6	11.3*
CoLA^E	71.1	39.4	87.3	77.5/85.2	86.5/83.0	78.8/80.2	74.2	78.2	59.2	33.8*
SST^E	71.2	38.8	**90.6**	80.4/86.8	87.0/83.5	79.4/81.0	74.3	77.8	53.8	43.7*
MRPC^E	71.3	40.0	88.4	77.5/84.4	86.4/82.7	79.5/80.6	74.9	78.4	58.1	**54.9***
QQP^E	70.8	34.3	88.6	79.4/85.7	86.4/82.4	81.1/82.1	74.3	78.1	56.7	38.0*
STS^E	71.6	39.9	88.4	79.9/86.4	86.7/83.3	79.9/80.6	74.3	78.6	58.5	26.8*
MNLI^E	72.1	38.9	89.0	80.9/86.9	86.1/82.7	81.3/82.5	75.6	79.7	58.8	16.9*
QNLI^E	71.2	37.2	88.3	81.1/86.9	85.5/81.7	78.9/80.1	74.7	78.0	58.8	22.5*
RTE^E	71.2	38.5	87.7	81.1/87.3	86.6/83.2	80.1/81.1	74.6	78.0	55.6	32.4*
WNLI^E	70.9	38.4	88.6	78.4/85.9	86.3/82.8	79.1/80.0	73.9	77.9	57.0	11.3*
DisSent WP^E	71.9	39.9	87.6	**81.9/87.2**	85.8/82.3	79.0/80.7	74.6	79.1	61.4	23.9*
MT En-De^E	72.1	40.1	87.8	79.9/86.6	86.4/83.2	81.8/82.4	75.9	79.4	58.8	31.0*
MT En-Ru^E	70.4	**41.0**	86.8	76.5/85.0	82.5/76.3	81.4/81.5	70.1	77.3	60.3	45.1*
Reddit^E	71.0	38.5	87.7	77.2/85.0	85.4/82.1	80.9/81.7	74.2	79.3	56.7	21.1*
SkipThought^E	71.7	40.6	87.7	79.7/86.5	85.2/82.1	81.0/81.7	75.0	79.1	58.1	52.1*
MTL GLUE^E	72.1	33.8	90.5	81.1/87.4	86.6/83.0	82.1/83.3	**76.2**	79.2	61.4	42.3*
MTL Non-GLUE^E	**72.4**	39.4	88.8	80.6/86.8	**87.1/84.1**	**83.2/83.9**	75.9	**80.9**	57.8	22.5*
MTL All^E	72.2	37.9	89.6	79.2/86.4	86.0/82.8	81.6/82.5	76.1	80.2	60.3	31.0*
BERT with Intermediate Task Training										
Single-Task^B	78.8	56.6	90.9	88.5/91.8	89.9/86.4	86.1/86.0	83.5	**87.9**	69.7	**56.3**
CoLA^B	78.3	**61.3**	91.1	87.7/91.4	89.7/86.3	85.0/85.0	83.3	85.9	64.3	43.7*
SST^B	78.4	57.4	**92.2**	86.3/90.0	89.6/86.1	85.3/85.1	83.2	87.4	67.5	43.7*
MRPC^B	78.3	60.3	90.8	87.0/91.1	89.7/86.3	86.6/86.4	**83.8**	83.9	66.4	**56.3**
QQP^B	79.1	56.8	91.3	88.5/91.7	**90.5/87.3**	88.1/87.8	83.4	87.2	69.7	**56.3**
STS^B	79.4	61.1	92.3	88.0/91.5	89.3/85.5	86.2/86.0	82.9	87.0	71.5	50.7*
MNLI^B	**79.6**	56.0	91.3	88.0/91.3	90.0/86.7	87.8/87.7	82.9	87.0	**76.9**	**56.3**
QNLI^B	78.4	55.4	91.2	**88.7/92.1**	89.9/86.4	86.5/86.3	82.9	86.8	68.2	**56.3**
RTE^B	77.7	59.3	91.2	86.0/90.4	89.2/85.9	85.9/85.7	82.0	83.3	65.3	**56.3**
WNLI^B	76.2	53.2	92.1	85.5/90.0	89.1/85.5	85.6/85.4	82.4	82.5	58.5	**56.3**
DisSent WP^B	78.1	58.1	91.9	87.7/91.2	89.2/85.9	84.2/84.1	82.5	85.5	67.5	43.7*
MT En-De^B	73.9	47.0	90.5	75.0/83.4	89.6/86.1	84.1/83.9	81.8	83.8	54.9	**56.3**
MT En-Ru^B	74.3	52.4	89.9	71.8/81.3	89.4/85.6	82.8/82.8	81.5	83.1	58.5	43.7*
Reddit^B	75.6	49.5	91.7	84.6/89.2	89.4/85.8	83.8/83.6	81.8	84.4	58.1	**56.3**
SkipThought^B	75.2	53.9	90.8	78.7/85.2	89.7/86.3	81.2/81.5	82.2	84.6	57.4	43.7*
MTL GLUE^B	**79.6**	56.8	91.3	88.0/91.4	90.3/86.9	**89.2/89.0**	83.0	86.8	74.7	43.7*
MTL Non-GLUE^B	76.7	54.8	91.1	83.6/88.7	89.2/85.6	83.2/83.2	82.4	84.4	64.3	43.7*
MTL All^B	79.3	53.1	91.7	88.0/91.3	90.4/87.0	88.1/87.9	83.5	87.6	75.1	45.1*

Intermediate Task	Avg	CoLA	SST	MRPC	QQP	STS	MNLI	QNLI	RTE	WNLI
ELMo with Intermediate Task Training										
Random^E	70.5	38.5	87.7	79.9/86.5	86.7/83.4	80.8/82.1	75.6	79.6	**61.7**	33.8*
Single-Task^E	71.2	39.4	**90.6**	77.5/84.4	86.4/82.4	79.9/80.6	75.6	78.0	55.6	11.3*
CoLA^E	71.1	39.4	87.3	77.5/85.2	86.5/83.0	78.8/80.2	74.2	78.2	59.2	33.8*
SST^E	71.2	38.8	**90.6**	80.4/86.8	87.0/83.5	79.4/81.0	74.3	77.8	53.8	43.7*
MRPC^E	71.3	40.0	88.4	77.5/84.4	86.4/82.7	79.5/80.6	74.9	78.4	58.1	**54.9***
QQP^E	70.8	34.3	88.6	79.4/85.7	86.4/82.4	81.1/82.1	74.3	78.1	56.7	38.0*
STS^E	71.6	39.9	88.4	79.9/86.4	86.7/83.3	79.9/80.6	74.3	78.6	58.5	26.8*
MNLI^E	72.1	38.9	89.0	80.9/86.9	86.1/82.7	81.3/82.5	75.6	79.7	58.8	16.9*
QNLI^E	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	58.8	22.5*
RTE^E	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	55.6	32.4*
WNLI^E	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	57.0	11.3*
DisSent WP^E	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	61.4	23.9*
MT En-De^E	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	58.8	31.0*
MT En-Ru^E	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	60.3	45.1*
Reddit^E	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	56.7	21.1*
SkipThought^E	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	58.1	52.1*
MTL GLUE^E	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	61.4	42.3*
MTL Non-GLUE^E	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	57.8	22.5*
MTL All^E	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	60.3	31.0*
Single-Task^B	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	69.7	**56.3**
CoLA^B	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	64.3	43.7*
SST^B	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	67.5	43.7*
MRPC^B	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	66.4	**56.3**
QQP^B	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	69.7	**56.3**
STS^B	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	71.5	50.7*
MNLI^B	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	**76.9**	**56.3**
QNLI^B	78.4	55.4	91.2	**88.7/92.1**	89.9/86.4	86.5/86.3	82.9	86.8	68.2	**56.3**
RTE^B	77.7	59.3	91.2	86.0/90.4	89.2/85.9	85.9/85.7	82.0	83.3	65.3	**56.3**
WNLI^B	76.2	53.2	92.1	85.5/90.0	89.1/85.5	85.6/85.4	82.4	82.5	58.5	**56.3**
DisSent WP^B	78.1	58.1	91.9	87.7/91.2	89.2/85.9	84.2/84.1	82.5	85.5	67.5	43.7*
MT En-De^B	73.9	47.0	90.5	75.0/83.4	89.6/86.1	84.1/83.9	81.8	83.8	54.9	**56.3**
MT En-Ru^B	74.3	52.4	89.9	71.8/81.3	89.4/85.6	82.8/82.8	81.5	83.1	58.5	43.7*
Reddit^B	75.6	49.5	91.7	84.6/89.2	89.4/85.8	83.8/83.6	81.8	84.4	58.1	**56.3**
SkipThought^B	75.2	53.9	90.8	78.7/85.2	89.7/86.3	81.2/81.5	82.2	84.6	57.4	43.7*
MTL GLUE^B	**79.6**	56.8	91.3	88.0/91.4	90.3/86.9	**89.2/89.0**	83.0	86.8	74.7	43.7*
MTL Non-GLUE^B	76.7	54.8	91.1	83.6/88.7	89.2/85.6	83.2/83.2	82.4	84.4	64.3	43.7*
MTL All^B	79.3	53.1	91.7	88.0/91.3	90.4/87.0	88.1/87.9	83.5	87.6	75.1	45.1*

Science doesn't happen linearly. Exploratory analysis is fine (essential, actually!) just know that it is exploratory.

When am I at risk of “researcher DoF” errors?

When am I at risk of “researcher DoF” errors?

- Always. You always are. That is why scientific results require consensus from many similar studies. No one study “proves” anything.

When am I at risk of “researcher DoF” errors?

- Always. You always are. That is why scientific results require consensus from many similar studies. No one study “proves” anything.

- But in particular—if you are refining your experimental design during the experiment, esp. in response to observed results

When am I at risk of “researcher DoF” errors?

- Always. You always are. That is why scientific results require consensus from many similar studies. No one study “proves” anything.

- But in particular—if you are refining your experimental design during the experiment, esp. in response to observed results (this is often unavoidable, but just acknowledge it)

Rules to live by…

Rules to live by…

- Define your hypothesis ahead of time, based on independent data

Rules to live by…

- Define your hypothesis ahead of time, based on independent data
- When possible, pre-register your methods

Rules to live by…

- Define your hypothesis ahead of time, based on independent data
- When possible, pre-register your methods
- The point of significance testing is to indicate levels of uncertainty, not to certify of “truth”

Rules to live by…

- Define your hypothesis ahead of time, based on independent data
- When possible, pre-register your methods
- The point of significance testing is to indicate levels of uncertainty, not to certify of "truth"
- Stay Curious! "Recognize the actual open-ended aspect of your projects…and analyze your data with this generality in mind" (Gelman and Loken)

So, on that note….let’s p hack!!!