CSE P 590 Building Data Analysis Pipelines

Fall 2024



Course introduction



A loosely related story

One week ago ... in Vienna, Austria

Benchmarks and Replicability in Software Engineering Research: *Challenges and Opportunities*



René Just

ISSTA 24

University of Washington

PLSE

A loosely related story

Two weeks ago ... directions and weather for Vienna



Temperature	7 C	Precipitation: 64 Humidity: 89% Wind: 32 km/h tion Wind	%			Satu	Weath rday 5:00 Clo
9	9	9	9	9	9	11	11
6AM	9AM	12 PM	3PM	6PM	9 PM	12AM	ЗAM
Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat
	·	-	-	- 之	\rightarrow	- 之	\rightarrow
10° 9°	11° 9°	12° 11°	18° 12°	22° 12°	22° 12°	19° 11°	21° 10
Innere Stad	Warnir It 9 - GeoSphe						
Fresh snow I	between 120	and 200 cm i	s possible.				

A loosely related story

Two weeks ago ...



🛞 Snow Warning

Innere Stadt Posted 19 hours ago

Recommended actions

TAKE ACTION to protect yourself. Widespread deep snow and/or significant ice coverage with significant disruption to road, rail and air transport. High risk of drivers becoming stranded. Avoid making non-essential journeys.

Source: GeoSphere Austria

Info & updates

Fresh snow between 120 and 200 cm is possible.

Weather Saturday 5:00 AM Cloudy

3AM

Sat

 \rightarrow

21° 10°

12AM

Fri

19° 11°

6PM

Wed

22° 12°

Thu

22° 12°

What happened?



VS.



- Incorrect data: Wind speed entered as mm/h (as opposed to km/h).
- Incorrect assumption: Data (mm/h) interpreted as snow fall.
- **No contextualization:** No consideration of the likelihood of such a snow storm, in the context of warm temperatures and historical data.

Valid data analysis: a simplified checklist



VS.



- Analysis grounded in a **conceptual model?**
- Clear operationalization (implementation)?
- Implementation consistent with the model?
- **Proper** use of **statistical methods**?
- Data interpreted in **context** of **prior knowledge?**
- Explored and validated **alternative hypotheses**?

Today

- Logistics and course overview
- Your background and expectations
- Data analysis: a birds-eye view
- A first data analysis task

Logistics and course overview

The CSEP 590 team

Instructor

- René Just (CSE2 338)
- Office hours: After class and by appointment
- <u>rjust@cs.washington.edu</u>

Teaching assistant

- Hannah Potter
- Office hours: by appointment
- <u>hkpotter@cs.washington.edu</u>

Logistics

- CSE2 G10, Mon, 6:30pm 9:20pm.
- Lectures, discussions, and in-class exercises.
- Course material, schedule, etc. on website: <u>https://homes.cs.washington.edu/~rjust/courses/CSEP590</u>
- Submission of assignments and Ed Discussion via Canvas: <u>https://canvas.uw.edu/1746473</u>

Course overview: the big picture

- **09/30:** Course introduction
- **10/07:** Analysis design and validity
- **10/14:** Data wrangling
- **10/21:** Statistical modeling
- **10/28:** Statistical significance and power
- **11/04:** Advanced statistical modeling
- **11/11:** No class
- **11/18:** Data visualization and reporting
- **11/25:** Big data
- **12/02:** Big data









Course overview: the big picture

•	09/30: Course introduction	
•	10/07: Analysis design and validity	In-class exercise
•	10/14: Data wrangling	In-class exercise
٠	10/21: Statistical modeling	In-class exercise
•	10/28: Statistical significance and power	In-class exercise
•	11/04: Advanced statistical modeling	HW 1
•	11/11: No class	
•	11/18: Data visualization and reporting	HW 2
•	11/25: Big data	
•	12/02: Big data	In-class exercise
	Class sessions have 2 parts: lecture and	in-class activity

Course overview: in-class exercises

In-class exercises (graded activities) have two parts

- 1. In-class part: Small-group work on a problem set
- 2. Take-home part: Reflection and submission of answers

What if I can't attend a class meeting?

- Work individually on the in-class exercise or work remotely with a partner.
- In-class exercise submissions are due at the end of the week.

Course overview: the big picture

09/30: Course introduction **10/07:** Analysis design and validity **10/14:** Data wrangling **10/21:** Statistical modeling **10/28:** Statistical significance and power 0 11/04: Advanced statistical modeling **11/11:** No class **11/18:** Data visualization and reporting **11/25:** Big data

Ouestions?

• 12/02: Big data

In-class exercise In-class exercise In-class exercise In-class exercise HW 1

HW 2

In-class exercise

Course overview: grading

- **30%** Homeworks
- 60% In-class exercises
- **10%** Participation



Course overview: the even bigger picture

This course

- is feedback-driven and evolves -- your input matters!
- covers a wide range of data analysis topics
- provides a hands-on experience for data analysis

This course <u>is not</u>

- a comprehensive course on statistical methods
- a tutorial on existing BI systems









Course overview: the even bigger picture

Other (UW) resources

- INFO 270: Calling Bullshit: Data reasoning in a digital world <u>https://callingbullshit.org</u>
- Practical Statistics for HCI

https://depts.washington.edu/madlab/proj/ps4hci/

• Statistical Analysis and Reporting in R http://depts.washington.edu/madlab/proj/Rstats/

Course overview: expectations

- Engage in discussions
- Reason about analysis design and validity
- Read a few research papers
- Work with the R programming language
- Have fun!

Your background and expectations

Your background and expectations



Introduction and a very brief survey

- **Role:** What is your current role?
- **Experience:** What is your experience with data analysis?
- **Top-2 expectations:** What do you expect from this course?

Data analysis: a birds-eye view

Data analysis vs. data analytics vs. data science

Many conflicting definitions and nuanced distinctions

This course uses *data analysis* as an **umbrella term**, covering all aspects from **design**, over **implementation** and **data collection**, to **statistical analysis** and **contextualization** of results.

An example study: design

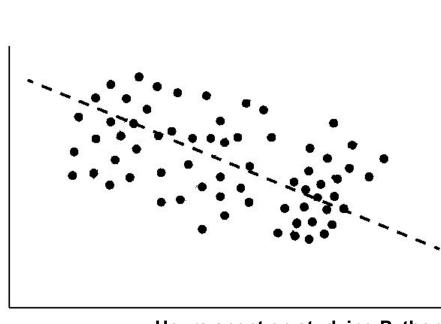
Goal:

Studying the **relationship** between **time spent** on **studying** Python and **success rate** in completing coding assignments.

Methodology:

- ~100 participants are randomly selected in front of CSE.
- Each participant is given a high-level overview of the study.
- Each participant decides on how long to study before attempting to solve any coding assignment.
- Each participant solves as many coding assignments as possible in one hour (after studying).

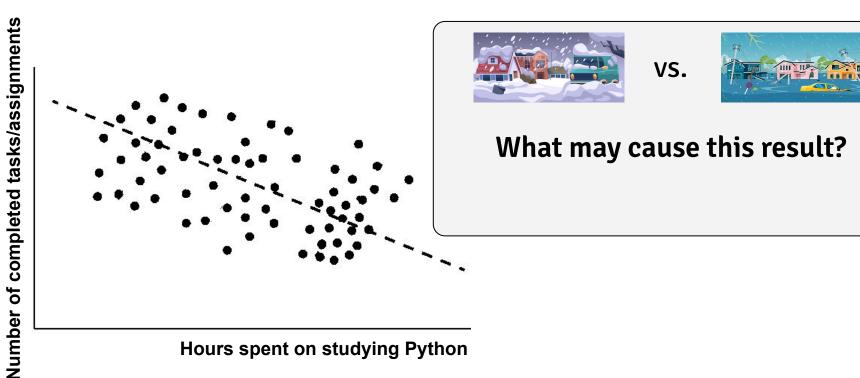
An example study: conclusions



Hours spent on studying Python

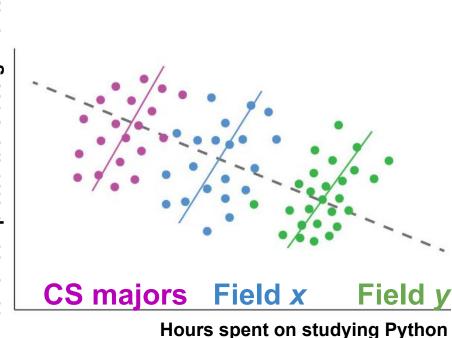
Conclusion: Spending more time on learning Python makes you a worse Python programmer.

An example study: conclusions



Conclusion: Spending more time on learning Python makes you a worse Python programmer.

An example study: Simpson's paradox



This phenomenon is called: **Simpson's paradox.**

An example study: Simpson's paradox



- Clear operationalization (implementation)?
- Implementation consistent with the model?
- Proper use of statistical methods?
- Data interpreted in context of prior knowledge?
- Explored and validated **alternative hypotheses**?



Hours spent on studying Python

Where did this study fail?



Another example study



http://www.prefrontal.org/files/posters/Bennett-Salmon-2009.pdf

Another example study: design



An argument for multiple comparison Craig M. Bennett ¹ , Abigail A. Baird ² , Michael	B. Miller ¹ , and George L. Wolford ³ ta Barbara, CA ² Department of Psychology, Vasaar College, Poughiseepsie, NY;	
INTRODUCTION With the extreme dimensionality of functional neuroimaging data comes	GLM RESULTS	
extreme risk for faise positives. Across the 130,000 words in a typical MRI volume the probability of a faile 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 property.		
METHODS	2.5 (-value	
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.	A t-contrast was used to test for regions with significant BOLD signal change	
Task. The task administered to the salmon involved completing an open-ended metalizing task. The salmon was shown a series of photographic 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.	during the photo condition compared to rest. The parameters for this comparison were $r(131) > 3.15$, p(uncorrected) < 0.001, 3 vote letter threshold. Several entry event diverse of the letter bend within the almost's brain a night (Figure 1, see above). This site of this cluster was 1 mm ³ within shorts event significances of $p = 0.001$. Thus the of this cluster was 1 mm ³ within the elebrothymetry and the state of the shorts was 1 mm ³ within the elebrothymetry and the regions could not be completed. Out of a search volume of B604 voxels a total of 16 voxels were significant. Out of a search volume of B604 voxels a total of 16 voxels were significant. Out of a search volume of B604 voxels a total of 16 voxels were significant. Out of a search volume of B604 voxels a total of 16 voxels were significant. Out of a search volume of B604 voxels a total of 16 voxels were significant.	
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.		
Pergrozessing. Image processing was completed using SPML. Preprocessing steps for the functional imaging data included a 6-parameter rigid-body affine realignment of the MRI timesentes, coregistration of the data to a T ₁ -weighted mattonical image, and 8 mm full-width at half-maximum (PWHM) Gaussian smoothing.		
<u>Analysis</u> . Vorcelwise statistics on the summon 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 corvolved with a canonical hemodynamic response. A temporal high pass filter of 128 seconds was include to account for low frequency drift. No subcorrelation correction was	VOXELWISE VARIABILITY	
applied. <u>Youd Sciencian</u> . Two methods were used for the correction of multiple comparisons in the (MRI routh). The first method controlled the overall failer discovery rate (PRN) and was howed on a method defined by Beginmini and Hochberg (1995). The second method controlled the overall familywise error rate (PWFR) through the use of Claussian methods that the two; This was done using algorithms originally devised by Friston et al. (1994).	a) b) c)	
DISCUSSION	To examine the spatial configuration of false positives we completed a variability analysis of the IMRI timeseries. On a voxel-by-voxel basis we calculated the standard deviation of signal values across all 140 volumes.	
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 trunseries may yield serious results if multiple comparisons are not controlled for. Adaptive methods for controlling the FDR and FWER are excellent options and are widely witable in all marking TMRI analysis packages. We argue that relying on standard statistical thresholds ($p < 0.001$) and low minimum clutter sizes ($k > 3$) is an infective control for multiple	We observed chattering of highly variable voxels into groups not areas of high voxel signal intensity. Figure 2a shows the mean EP1 image for all 140 Figure 2a shows thereholded standard deviation values overlaid onto a high- resolution T ₁ -weighted image.	
comparisons. We further argue that the vast majority of MRI studies should be utilizing multiple comparisons correction as standard practice in the computation of their statistics.	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	
REFERENCES	positive correlation between the mean voxel value and its variability over	

REFERENCES

Benjamini Y and Hochberg Y (1995). Controlling the fulse discovery rate: a practical and p approach to multiple testing. *Journal of the Royal Statistical Society: Series 8*, 57:289-300.

on KJ, Wonsley KJ, Frackowiak RSJ, Mazzietta JC, and Evans AC. (1994). Assessi ficance of focal activations using their spatial extent. Hawaw Brain Mapping, 1:214-220

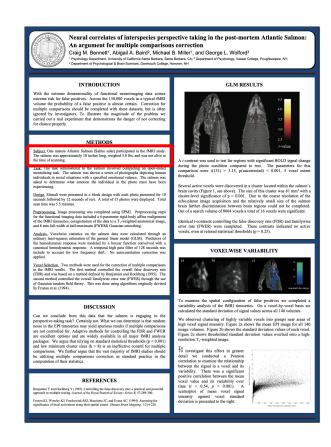
deviation is presented to the right

time (r = 0.54, p < 0.001). A

scatterplot of mean voxel signal ntensity against voxel standard

Another example study: design

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.

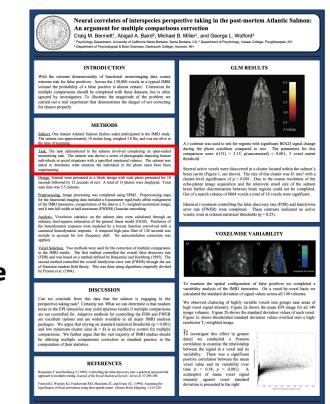


Another example study: design

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: [...] **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**.



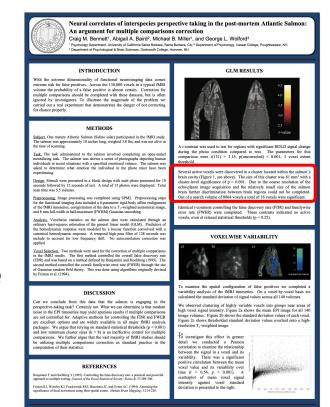


Another example study: conclusions

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: [...] 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.

Results: Several active voxels were discovered [...] Out of a search volume of 8064 voxels a total of **16 voxels** were significant.



Another example study: conclusions

Interpretation of pure noise

- Noisy data source
- Multiple hypotheses tested on the same data
- An argument for multiple comparisons correction
- Analysis grounded in a conceptual model?
- Clear operationalization (implementation)?
- Implementation consistent with the model?
- **Proper** use of **statistical methods**?
- Data interpreted in context of prior knowledge?
- Explored and validated **alternative hypotheses**?

Where did this study fail (on purpose)?

Neural correlates of interspecies perspective taking in the post-mortem Atlantic Sal An argument for multiple comparisons correction Craig M. Bennett', Abgal A. Baird', Michael B. Miller', and George L. Wolford ¹ Hyperhop Desemter Multiple of Carling Man Beach. Salt hatens, Adv. "Operated of Psychology. Your Colege, Paughaeyan, NY: Hyperhop Desemter Multiple of Carling Man Beach. Salt hatens, Adv. "Operated of Psychology. Your Colege, Paughaeyan, NY: Hyperhop Desemter Multiple Carling Man Beach. Salt hatens, NY:				
INTRODUCTION	GLM RESULTS			
With the extreme idmensionality of functional neuroimaging data comes termer nick for faith positives. Across the 10,000 vocatio in a poychal (MB1 volume the probability of a fulse positive is almost certain. Cerrection for imported by investigators. To Illustrate the magnitude of the problem we certife out a real experiment that demonstrates the danger of not correcting for chance properly.				
METHODS Sabjaci, One mature Atlantic Salmon (Salmo salar) participated in the MRI study. The salmon was approximately 18 metes kong, weighed 3.8 ms, and was not alive at the time of samming. Task: The task administered to the salmon was shown a series of photographic depicting human	25 J-value A r-contrast was used to test for regions with significant BOLD sign during the photo condition compared to rest. The parameter comparison wer (131) > 3.15, p(monoreted) < 0.001, 3 vo			
individuals in social situations with a specified emotional valence. The salmon was adode 10 determine what emotion the individual in the photo must have been experiencing. <u>Dasing</u> , Simmli were presented in a block design with each photo presented for 10 sconds followed by 21 scondor of rest. A total of 15 photos were displayed. Total scan time was 5.5 minutes.	threshold. Several active voxels were discovered in a cluster located within th brain activ() (Figure 1, see above). The size of this cluster was 81 r cluster-level significance of $p = 0.001$. Due to the coarse resolu- clo-poltant image acquisition and the relatively small size of t brain further discrimination between brain regions could not be 0.01 of a search volume of 864 to volume 1 atol 161 for souch were sign			
Perprocessing, Image processing was completed using SPM2. Proprocessing steps for the functional imaging data included a 6-parameter right-body affine realignment of the MRI timeseries, coregistration of the data to a T,-weighted anatomical image, and 8 mm full-width at half-maximum (FWHM) Gaussian smoothing. Analysis, Vorchwise statistics on the salmon data were calculated through an	Identical <i>i</i> -contrasts controlling the false discovery rate (FDR) and i error rate (FWER) were completed. These contrasts indicated voxels, even at relaxed statistical thresholds (p = 0.25).			
Pathyzia, Vecewie a sistence on the general inter model (GLM). Predictors of onlinary least-squares estimation of the general inter 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 asentiated.	VOXELWISE VARIABILITY			
Stead Education. Two methods were used for the corrections of multiple comparisons in the PMI reasons. The first method controlled the overall fields discovery rate (PDR) and was based on a method defined by Benjamini and Hechberg (1995). The second method controlled the overall finallysisies error rate (PVER) through the use of Ginessian random field theory. This was done using algorithms originally devised by Fraston et al. (1954).	a) b) c)			
DISCUSSION	To examine the spatial configuration of false positives we co variability analysis of the fMRI timeseries. On a voxel-by-voxe calculated the standard deviation of signal values across all 140 volu			
Can we conclude from this data that the nations is engaging in the prepercive-taking that C-trainity out Waits we can determine its that random moise in the EPI timeseries may pited parious results if multiple comparisons are concruted left. And apprive methods for controlling the TDM and FWER are exceeding objects and are widely available in all major MRI analysis packages. We argue that relying an standard statistical thresholds (σ 40001) and low minimum cluster sizes ($\kappa > 8$) is an ineffective control for multiple comparisons. We further argue that the variantity of MRI analysis due to the statistical statistical thresholds (σ 40001) and low minimum cluster sizes ($\kappa > 8$) is an ineffective control for multiple to this statistical statistical statistical thresholds be utilizing multiple comparisons correction as standard practice in the comparisons. We for analysis.	caccuated the sindual deviation of signal values across all 140 volus We observed classifier of highly values loves in min groups are high voted signal intensity. Figure 2a shows the near PFI maps i mage volumes. Figure 2a shows the instantial deviation values overhald or resolution T, weighted mage.			
REFERENCES Breijanzini Y and Heckberg Y (1995). Centrolling the failer discovery rate: a practical and powerful	content in significant positive correlation between the mean voce value and its variability over time $(r = 0.54, p < 0.001)$. A satterplot of mean voxel signal intensity against voxel signal			

Another example study: conclusions

Interpretation of pure noise

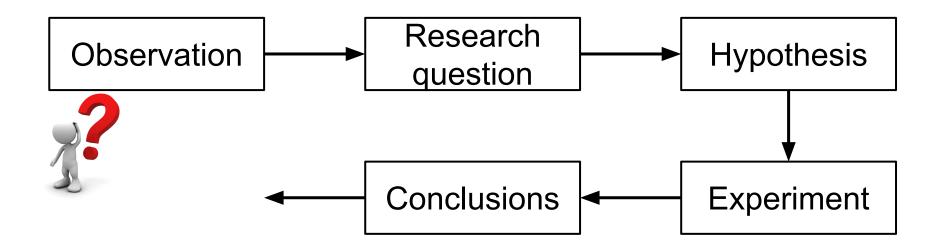
- Noisy data source
- Multiple hypotheses tested on the same data
- An argument for multiple comparisons correction
- Analysis grounded in a conceptual model?
- Clear operationalization (implementation)?
- Implementation consistent with the model?
- **Proper** use of **statistical methods**?
- Data interpreted in context of prior knowledge?
- Explored and validated **alternative hypotheses**?

Sound data analysis goes well beyond implementation correctness.

Craig M. Bennett ¹ , Abigal A. Baird ⁹ , Michael B. Miller ¹ , and George L. Wolford ¹ Proceedings of the stress of the stress of the stress that the start is that stress of the stre				
INTRODUCTION	GLM RESULTS			
With the extreme dimensionality of functional neuroimaging data comes termer nisk for false positives. Across the 10,000 vockit in a posical MRI volume the probability of a false positive is almost certain. Correction for ingrood 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.	43 40 35 30			
HETHODS Stability. One metters Atlantic solutions (Solato solar) participated in the MMI andy. The analous was approximately 15 includes long, weights 3.3.8 Bes, and was not after at the time of committing. Task. The task sharing that solar solarised completing an open-ended metalizing task. The asiance was advected metalizing an open-ended metalizing task. The asiance was advected metalizing an open-ended individual in noted intuines with a specific metanismum and the solar was individual in noted intuines with a specific metanismum and the solar was advected as a specific metalized as a specific m	23 Fulle A recentrat was used to test for regions with significant HOLD signa during the photo condition compared to rest. The parameters comparison were (131) > 3.15, plancerreted) < 0.001, 3 vox threshold.			
saked to determine what emotion the individual in the photo must have been experiencing. Design, Slimitul were presented in a block design with each photo presented for 10 seconds followed by 12 accords of rest. A total of 15 photos were displayed. Total seam time was 5.5 minutes. Preprocessing, image processing was completed using SPM2. Preprocessing steps	Several active voxels were discovered in a cluster located within the brain cavity (Figure 1, see above). The size of this cluster was 81 mm cluster-level significance of p = 0.001. Due to the coarse resolution echophanar image acquisition and the relatively small size of the brain further discrimination between brain regions could not be co Out of a search volume of 8864 voxels a total of 16 voxels were signing the search volume of 8864 voxels a total of 16 voxels were signing the search volume of 8864 voxels a total of 16 voxels were signing the search volume of 8864 voxels a total of 16 voxels were signing the search volume of 8864 voxels a total of 16 voxels were signing the search volume of 8864 voxels a total of 16 voxels were signing the search volume of 8864 voxels a total of 16 voxels were signing the search volume of 8864 voxels a total of 16 voxels were signing the search volume of 8864 voxels a total of 16 voxels were signing the search volume of 8864 voxels a total of 16 voxels were signing the search volume of 8864 voxels a total of 16 voxels were signing the search volume of 8864 voxels a total of 16 voxels were signing the search volume of 8864 voxels a total of 16 voxels were signing the search volume of 8864 voxels a total of 16 voxels were signing the search volume of 8864 voxels a total of 16 voxels were signing the search volume of 8864 voxels a total of 16 voxels were signing the search volume of 8864 voxels a total of 16 voxels were signing the search volume of 8864 voxels a total of 16 voxels were signing the search volume of 8864 voxels a total of 16 voxels were signing the search volume of 8864 voxels a total of 16 voxels were signing the search volume of 8864 voxels a total of 16 voxels were signing the search volume of 8864 voxels a total of 16 voxels were signing the search volume of 8864 voxels a total of 16 voxels were signing the search voxels a total search volume the search voxels a total search voxels a total search voxels total search voxels a to			
for the functional imaging data included a 6-parameter rigid-body affine realignment of the MRI timeseries, coregistration of the data to a 1, weighted matomical image, and 8 mm full-width a thalf-maximum (FWHM) Gaussian smoothing. <u>Analyzis</u> , Voxelwise statistics on the salmon data were ealculated through an	Identical <i>t</i> -contrasts controlling the false discovery rate (FDR) and fa error rate (FWER) were completed. These contrasts indicated <i>t</i> voxels, even at relaxed statistical thresholds (p = 0.25).			
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.	VOXELWISE VARIABILITY			
Approximation of the second se	a) b) c) c) a c c) c			
DISCUSSION	To examine the spatial configuration of false positives we com variability analysis of the fMRI timeseries. On a voxel-by-voxel calculated the standard deviation of signal values across all 140 volum			
Can we conclude from this data that the saftone in engaging in the prepretive-taking skills? Cratishi you Shva the can determine its him random noise in the EPI timeseries may yield sparsors results if multiple comparisons are not controlled of a Anghviry methods to constraining an EPA and FWLM packages. We argue that relying on tunnider datatistical thresholds (ϕ - 6.007) and low minimum cluster size (k-8.5 %) in an interfective control for multiple comparisons. We further argue that the vant majority of MRI stadies should be utilizing multiple comparisons correction as standard practice in the comparison of their statistics.	We observed chararing of highly variable vende is an group, near high vend sign in the home TP Hunger is image volumes. Figure 2b shows the initiated deviation values of en- fryme 2b shows the holded at marked deviation values overhid one resolution T_weighted image.			
REFERENCES	positive correlation between the mean voxel value and its variability over			
Benjamini Y and Hechberg Y (1995). Centrolling the fulse discovery rate: a practical and powerful approach to multiple testing. <i>Journal of the Royal Statistical Society: Series 8</i> , 57:289-300.	time (r = 0.54, p < 0.001). A scatterplot of mean voxel signal intensity against voxel standard			
Fristen KJ, Wonsley KJ, Frackowiak RSJ, Mazzietta JC, and Evans AC. (1994). Assessing the significance of focal activations using their spatial extent. <i>Hawase Brain Mapping</i> , 1:214–220.	deviation is presented to the right.			

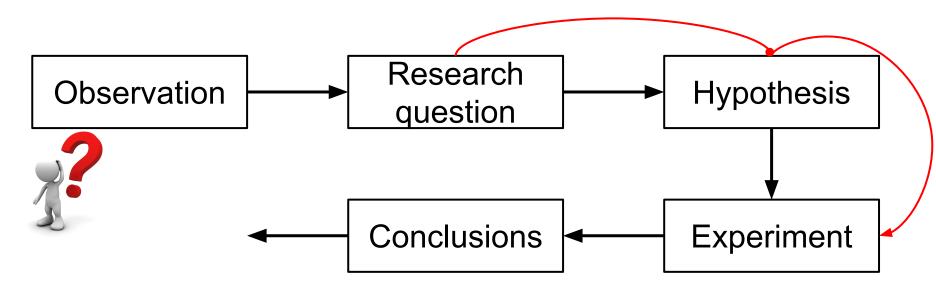


The scientific method



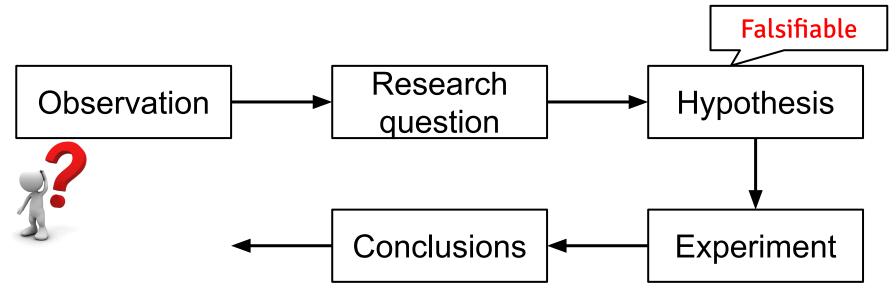
Seems pretty simple ... what's important?

The scientific method

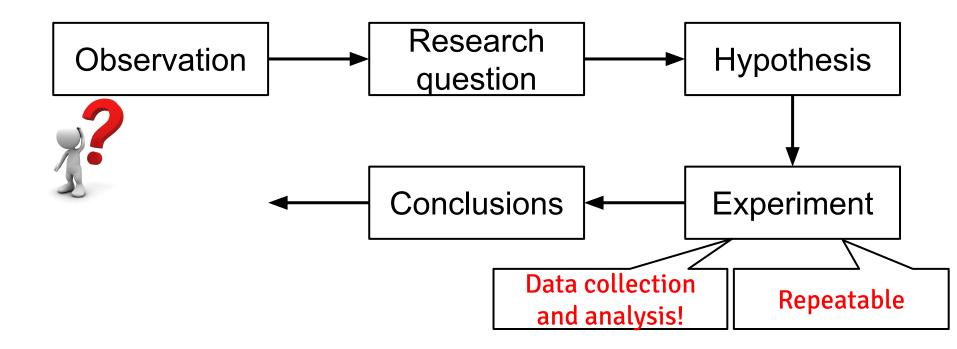


Operationalization/hypothesis formalization

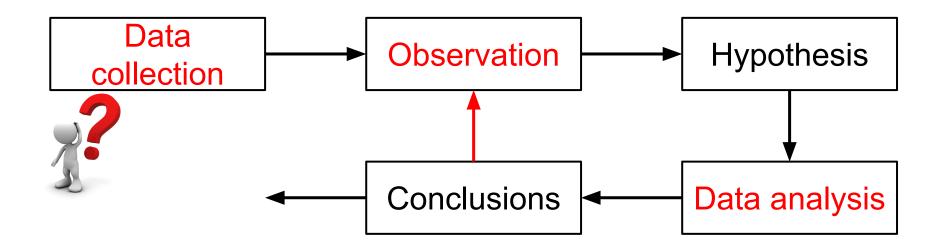
The scientific method



The scientific method

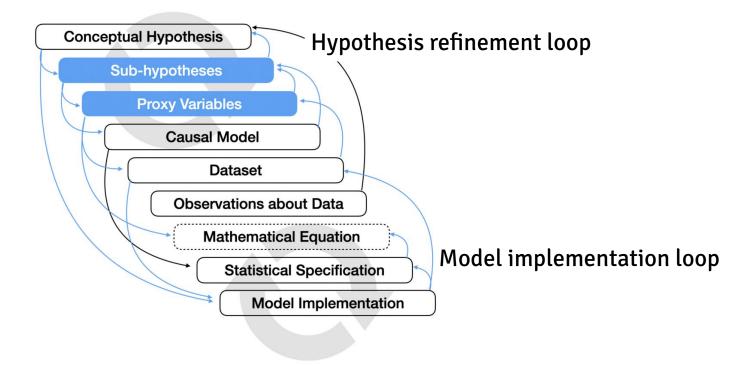


The scientific method: common mistake



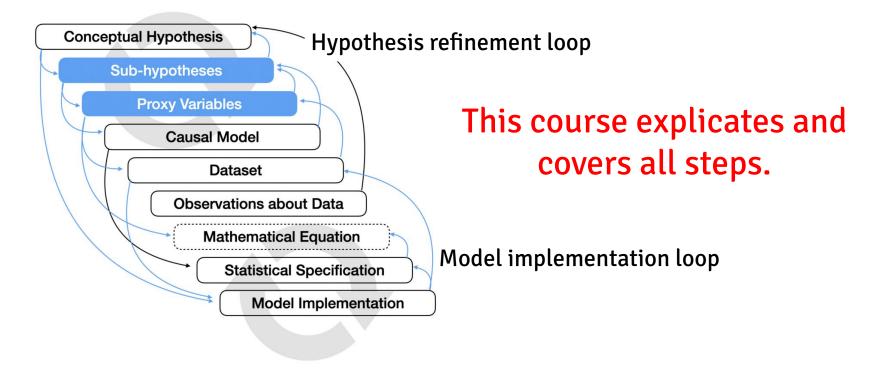
"If you torture the data long enough, it will confess." [Ronald Harry Coase]

A more nuanced view on hypothesis formalization



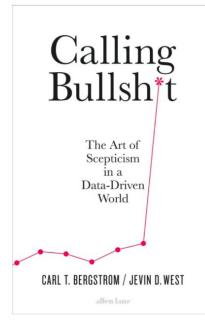
Hypothesis formalization: Empirical findings, software limitations, and design implications, Jun et al., TOCHI 2022

A more nuanced view on hypothesis formalization



Hypothesis formalization: Empirical findings, software limitations, and design implications, Jun et al., TOCHI 2022

Why should you care?



Make informed decisions based on valid data analyses.

Collaborators, students, reviewers:

- These results are bad and cannot be true.
- If you don't trust my intuition, run your own experiments.
- These results are entirely expected.
- I have computed all the data; which statistical test should I use to show that my results are significant?
- Most papers are wrong or later obsolete, so who cares?
- I don't understand these intervals, can you give a p value?

Collaborators, students, reviewers:

- These results are bad and cannot be true.
- If you don't trust my intuition, run your own experiments.
- These results are entirely expected.
- I have computed all the data; which statistical test should I use to show that my results are significant?
- Most papers are wrong or later obsolete, so who cares?
- I don't understand these intervals, can you give a p value?

Avoid confirmation bias; always scrutinize your results.

Collaborators, students, reviewers:

- These results are bad and cannot be true.
- If you don't trust my intuition, run your own experiments.
- These results are entirely expected.
- I have computed all the data; which statistical test should I use to show that my results are significant?
- Most papers are wrong or later obsolete, so who cares?
- I don't understand these intervals, can you give a p value?

Transform intuition and expectations into testable hypotheses!

Collaborators, students, reviewers:

- These results are bad and cannot be true.
- If you don't trust my intuition, run your own experiments.
- These results are entirely expected.
- I have computed all the data; which statistical test should I use to show that my results are significant?
- Most papers are wrong or later obsolete, so who cares?
- I don't understand these intervals, can you give a p value?

"Statistical significance is the least interesting thing about the results" [Sullivan and Fein: Using effect size -- or why the p value is not enough]

A first data analysis task

A first data analysis task



Context

- Your team semi-automatically patches SW bugs with *AutoCoder*.
- A new tool *AutoPatcher* is available: promising (benchmark) results.

Guiding questions

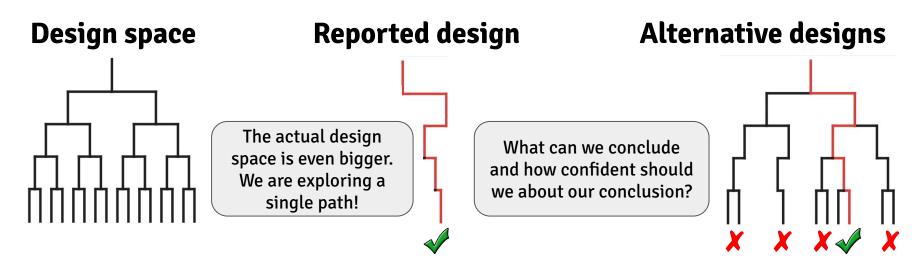
- Is AutoPatcher better than AutoCoder?
- Should your team adopt *AutoPatcher*?

Set up

- Small groups (~6 students)
- Discuss and document an analysis design: https://tinyurl.com/48uz6wau
- Report design (decisions) to the class

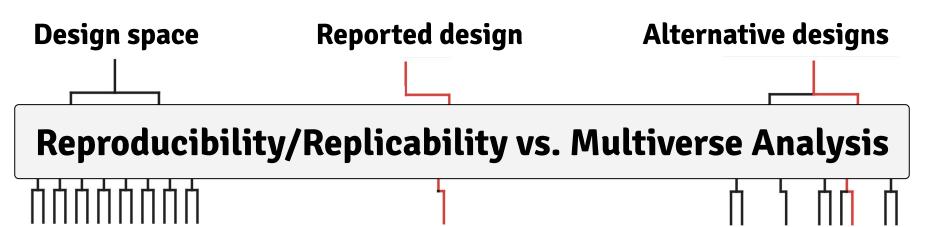
Should your team adopt *AutoPatcher*?

- 1. Define proxy for patch success (plausible vs. correct)
- 2. Choose evaluation benchmark (external vs. internal)
- 3. Aggregation (mean vs. median)
- 4. Choose statistical test (T vs. U)



Should your team adopt *AutoPatcher*?

- 1. Define proxy for patch success (plausible vs. correct)
- 2. Choose evaluation benchmark (external vs. internal)
- 3. Aggregation (mean vs. median)
- 4. Choose statistical test (T vs. U)



Artifact badges (ACM publications)





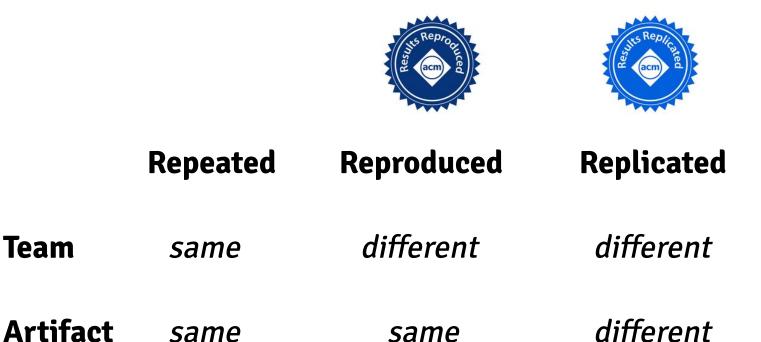


Pre-publication (Publishing team)

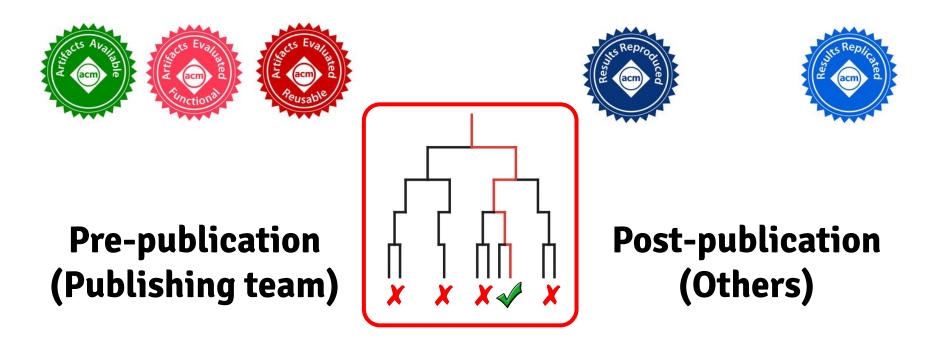
Post-publication (Others)

Reproduce vs. Replicate (It's confusing, I know)

Team



Robust analysis results != robust conclusions



Replication can improve confidence in conclusions.

Open discussion

- 1. Define proxy for patch success (plausible vs. correct)
- 2. Choose evaluation benchmark (external vs. internal)
- 3. Aggregation (mean vs. median)
- 4. Choose statistical test (T vs. U)

