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

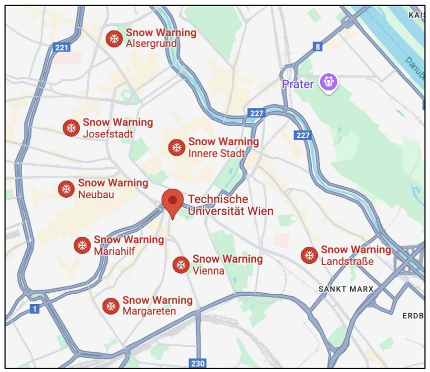
ISSTA 24

René Just
University of Washington

PLSE

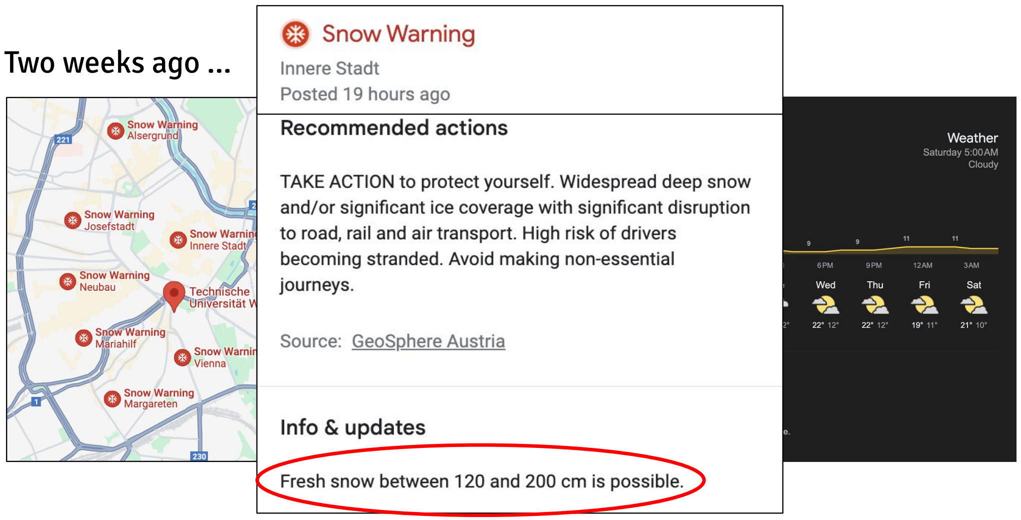
A loosely related story

Two weeks ago … directions and weather for Vienna



A loosely related story

Two weeks ago …



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
- rjust@cs.washington.edu

Teaching assistant
- Hannah Potter
- Office hours: by appointment
- hkpotter@cs.washington.edu

Logistics

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

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**
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	**In-class exercise**
10/14: Data wrangling	**In-class exercise**
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**

Questions?

Course overview: grading

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

Questions?

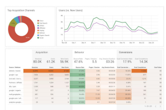
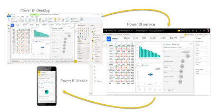
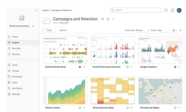
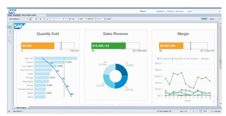
Course overview: the even bigger picture

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
  https://callingbullshit.org
- 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.

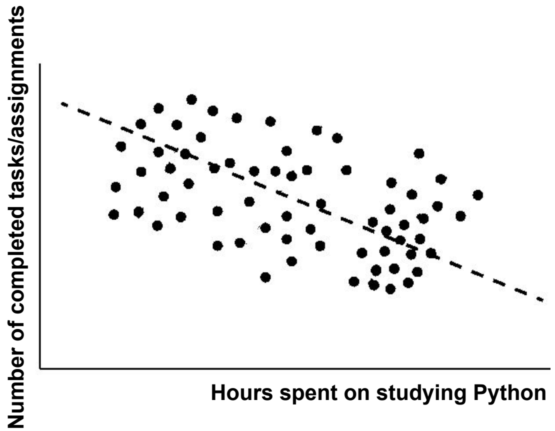
An example study: design

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

Methodology:

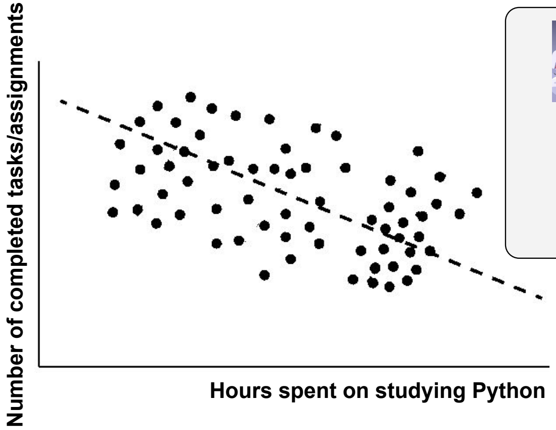
- ~100 participants are randomly selected in front of CSE.
- Each participant is given a high-level overview of the study.
- Each participant decides on how long to study before attempting to solve any coding assignment.
- Each participant solves as many coding assignments as possible in one hour (after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

An example study: conclusions



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

An example study: conclusions

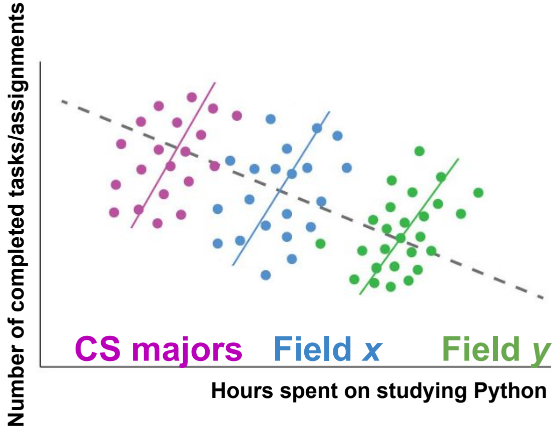


vs.

What may cause this result?

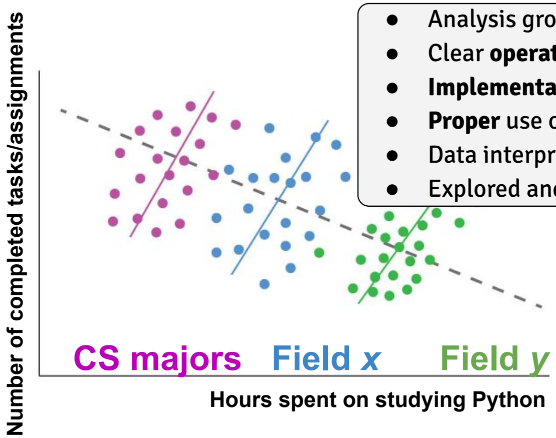
Conclusion: Spending more time on learning Python makes you a worse Python programmer.

An example study: Simpson's paradox



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

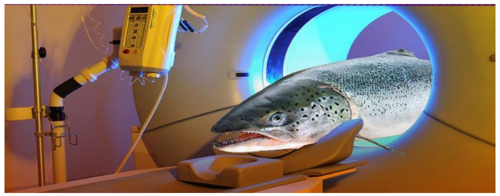
An example study: Simpson's paradox



- 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**?

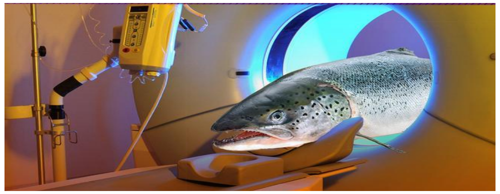
Where did this study fail?

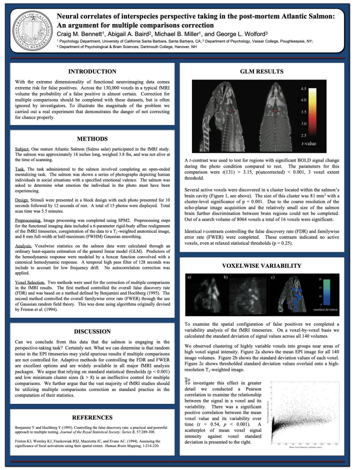
Another example study



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

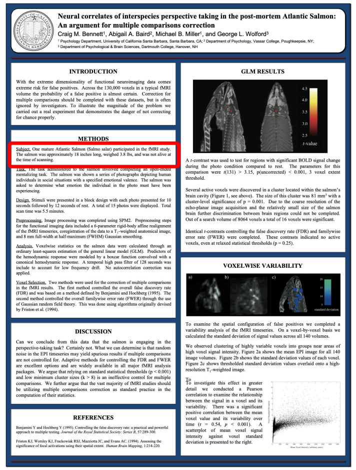
Another example study: design





Another example study: design

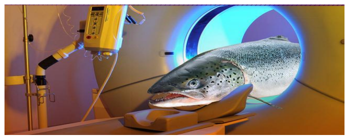
Subject: One mature **Atlantic Salmon** (Salmo salar) participated in the **fMRI study**. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was **not alive at the time of scanning**.

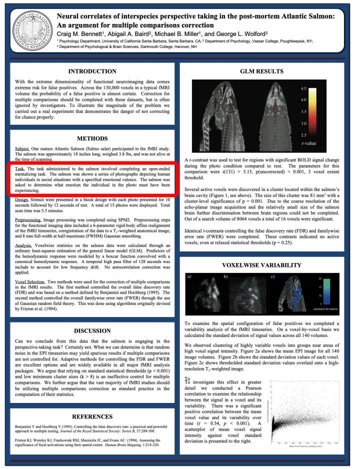


Another example study: design

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**.



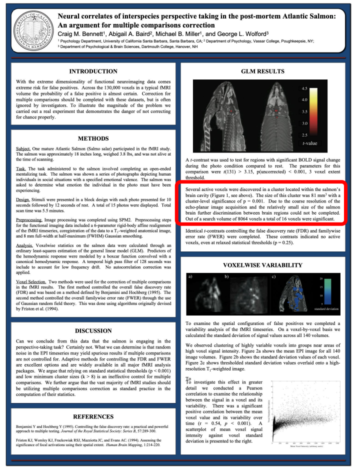


Another example study: conclusions

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.**

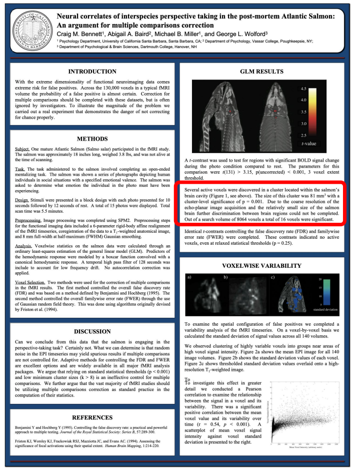


Another example study: conclusions

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**?



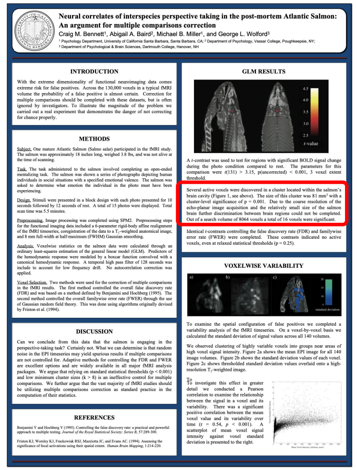
Where did this study fail (on purpose)?

Another example study: conclusions

Interpretation of pure noise

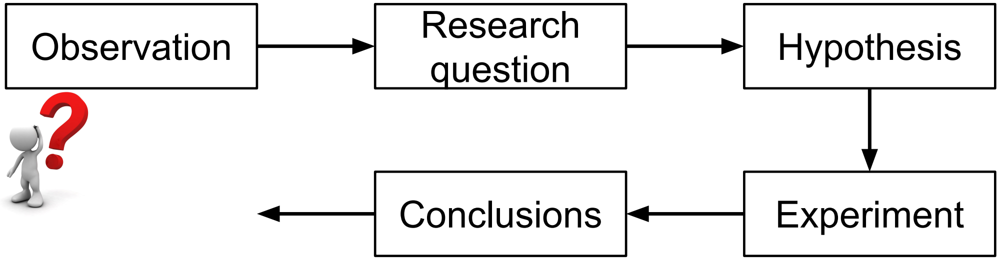
- Noisy data source
- Multiple hypotheses tested on the same data
- An argument for multiple comparisons correction

- 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**?



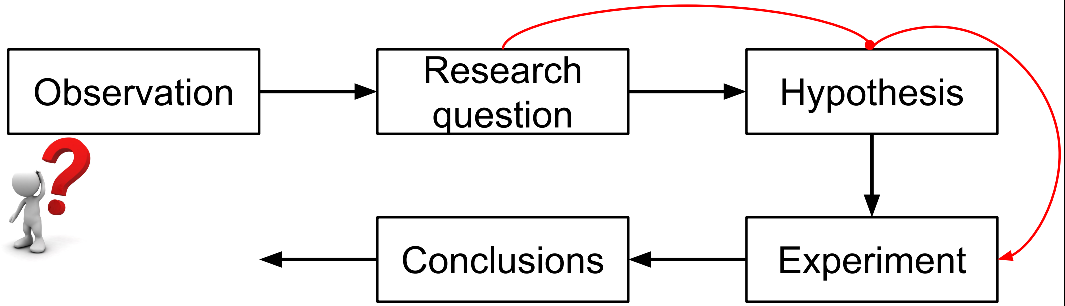
Sound data analysis goes well beyond implementation correctness.

The scientific method



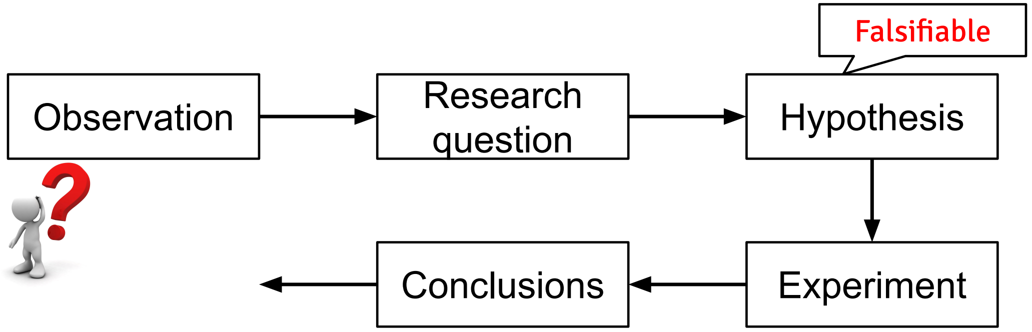
Seems pretty simple ... what’s important?

The scientific method

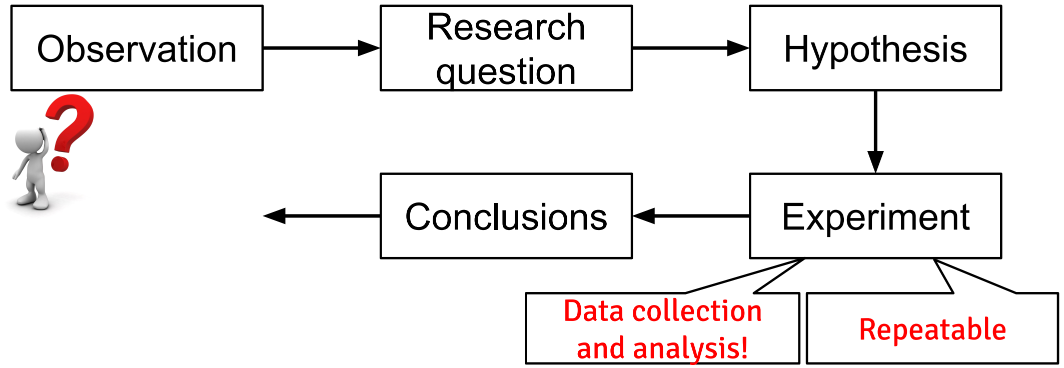


Operationalization/hypothesis formalization

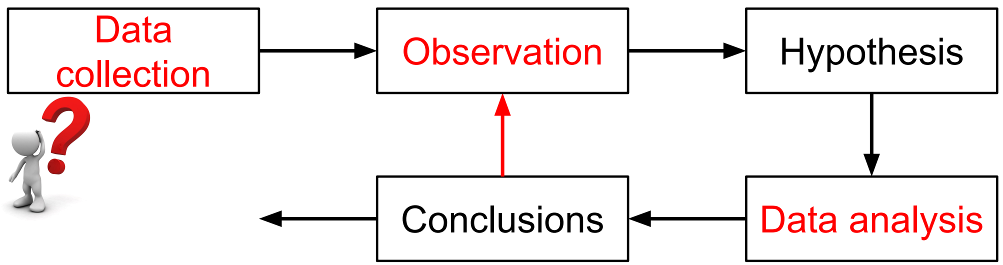
The scientific method



The scientific method

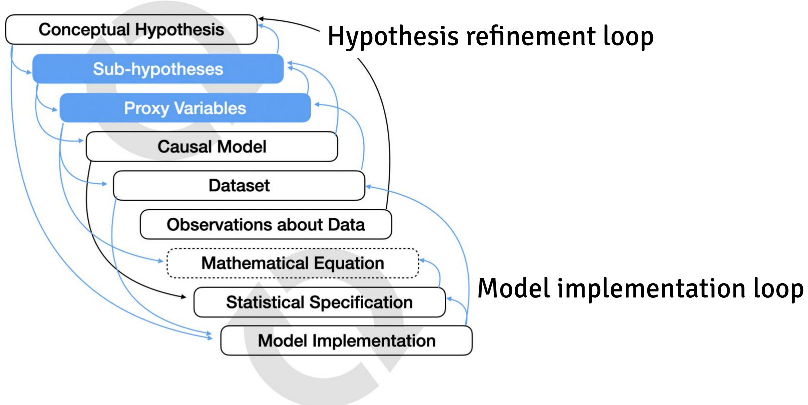


The scientific method: common mistake



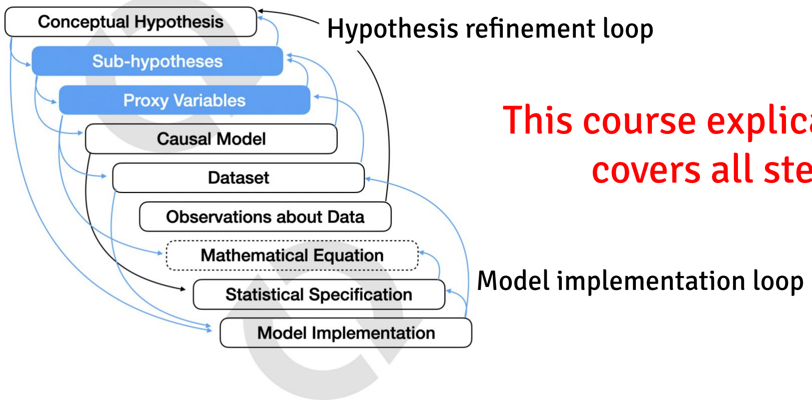
"If you torture the data long enough, it will confess."
[Ronald Harry Coase]

A more nuanced view on hypothesis formalization



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

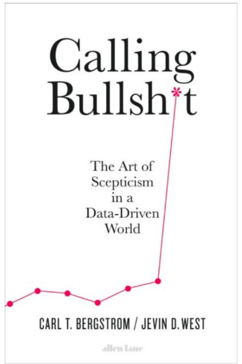
A more nuanced view on hypothesis formalization



This course explicates and covers all steps.

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.

Why I care: my favorite quotes

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?

Why I care: my favorite quotes

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.

Why I care: my favorite quotes

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!

Why I care: my favorite quotes

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

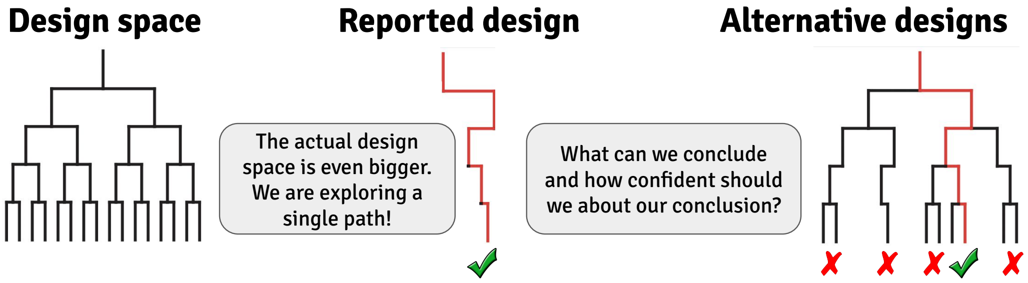
- Is *AutoPatcher* better than *AutoCoder*?
- Should your team adopt *AutoPatcher*?

Set up

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

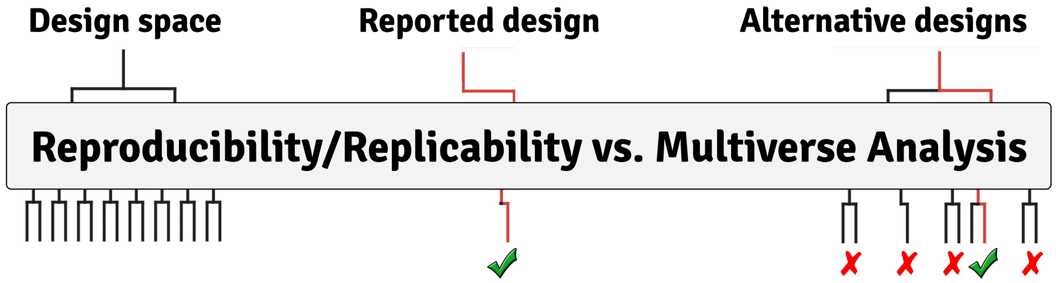
Should your team adopt *AutoPatcher*?

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)

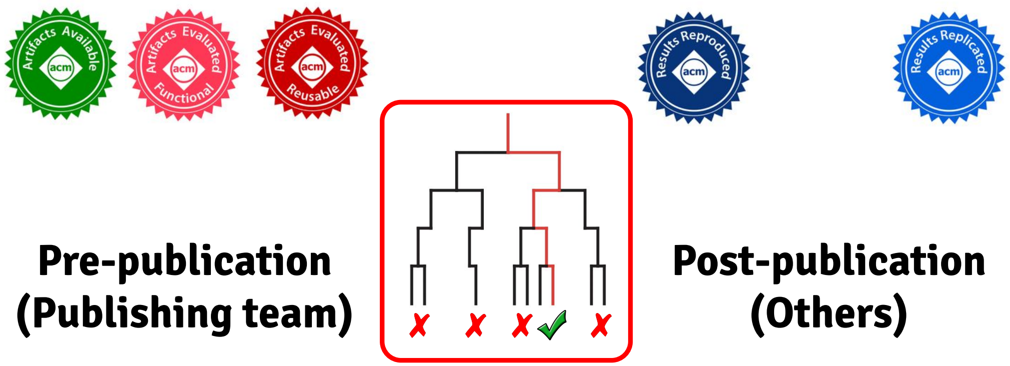
Pre-publication (Publishing team)

Post-publication (Others)

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

	Repeated	Reproduced	Replicated
Team	*same*	*different*	*different*
Artifact	*same*	*same*	*different*

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)