CSE 599K
Empirical Research Methods

Winter 2025

SE meets Science

Today

- Recap: analysis validity
- SE principles for rigorous Science
- Two example studies

Analysis validity

Analysis validity: open questions

External validity
- Does the experiment generalize (to larger population, other subjects, etc.)?
- How representative is the sample?

Internal validity
- Does the experiment isolate the variable(s) of interest?
- Does the experiment control for confounders and unwanted effects?

Construct validity
- Does the experiment measure what it claims to measure?
- Do the proxy measures and tools adequately measure the concept of interest?

(Statistical) conclusion validity
- Are the conclusions valid based on the chosen statistical test and sample size?
- Are the conclusions valid based on the observed significance (p value)?

SE principles for rigorous science

Science to practice is not a one-way street!

Let's improve **scientific rigor** with
SE principles and **best practices!**

Design reviews

- Design reviews are common in practice.
- Embrace and value **pre-registrations**.

- RFCs and public discussions (e.g., GH) provide valuable context.
- **Public (open) reviews** should be a no-brainer!

Hark no more: On the preregistration of chi experiments, Cockburn et al., CHI 2018
https://openreview.net/

Quality assurance

- Modern code review is incremental (not holistic).
- Move to **pre-acceptance artifact evaluations.**

- Software testing is the most common QA approach in practice.
- Require **evidence for artifact testing**.

Expectations, outcomes, and challenges of modern code review, Bacchelli and Bird, ICSE 2013
Modern code review: a case study at Google, Sadowski et al., ICSE 2018

Process

Merge conflicts (branches) are resolved by branch authors.

Expect **resolution** (knowledge) of **conflicting results**.

Don't expect others to resolve your merge conflict!

Process

Merge conflicts (branches) are resolved by branch authors.

Expect **resolution** (knowledge) of **conflicting results**.

No premature optimizations.

Focus on **design validity** before scrutinizing artifacts.

Science is a collaborative effort!

Software Engineering is a **collaborative effort**.
We should view science the same way!

Science as Amateur Software Development

1. How can software engineering principles improve the rigor of data analyses?
2. Are these principles equally applicable to computational notebooks?
3. Describe three specific quality control mechanisms.
4. McElreath attributes a significant number of incorrect (scientific) studies to "sloth". What are the specific issues he is calling out, and what solutions does he propose?
5. Provide an argument for why or why not general-purpose programming languages such as Python are an adequate choice for data analysis.

Two example studies

An example study: design

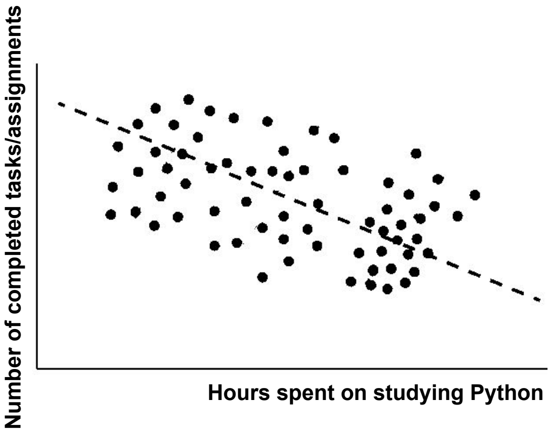
Goal:

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

Methodology:

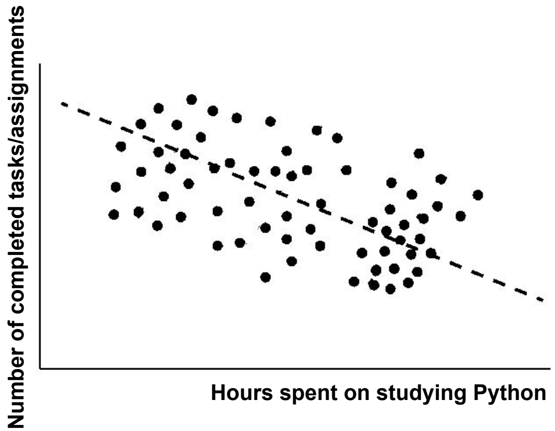
- ~100 participants are randomly selected in front of CSE.
- Each participant is given a high-level overview of the study.
- Each participant decides on how long to study before attempting to solve any coding assignment.
- Each participant solves as many coding assignments as possible in one hour (after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

An example study: conclusions



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

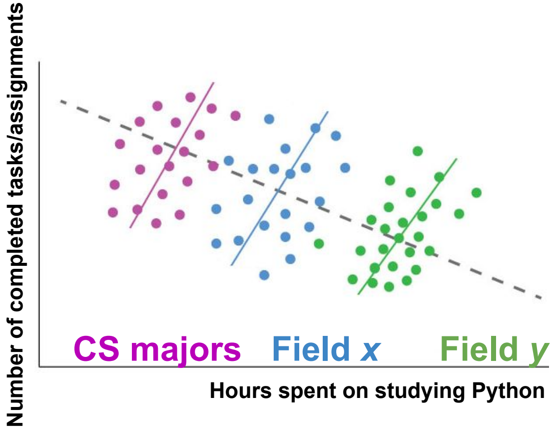
An example study: conclusions



What may cause this result?

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

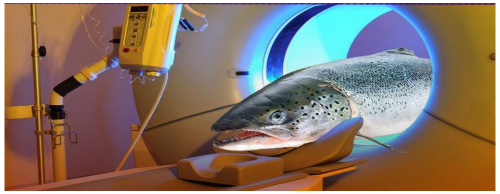
An example study: Simpson's paradox



Where did this study fail?

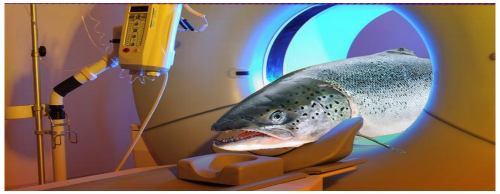
This phenomenon is called: **Simpson's paradox.**

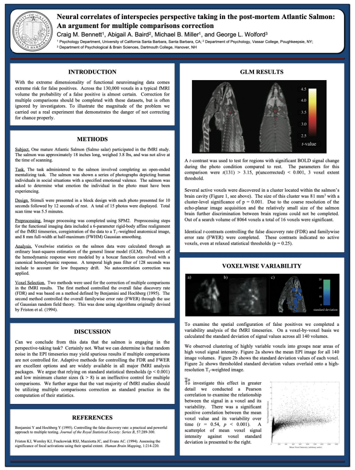
Another example study



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

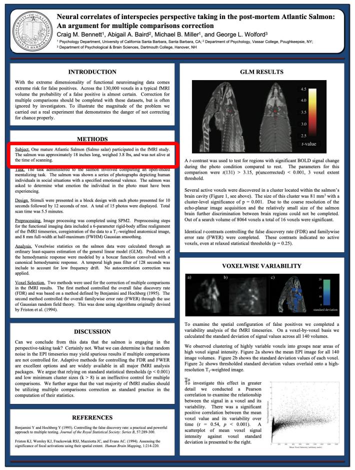
Another example study: design





Another example study: design

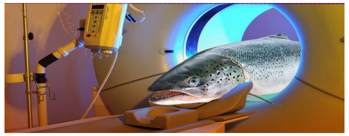
Subject: One mature **Atlantic Salmon** (Salmo salar) participated in the **fMRI study**. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was **not alive at the time of scanning**.

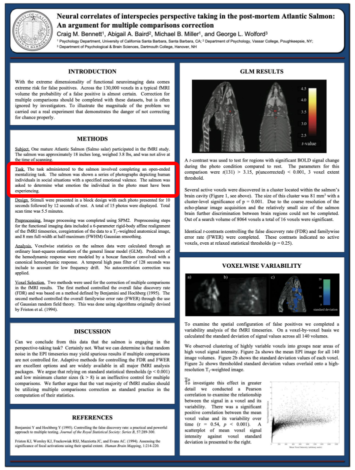


Another example study: design

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**.



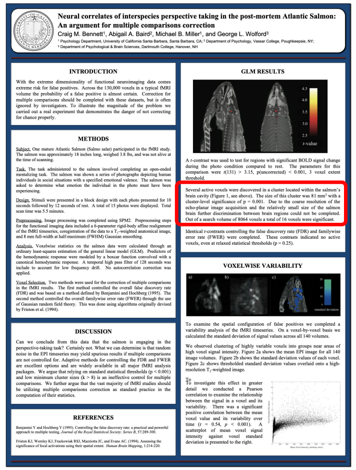


Another example study: conclusions

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.**

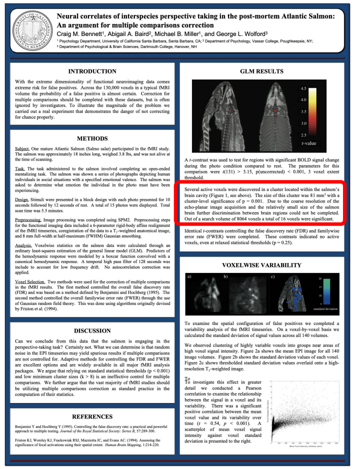


Another example study: conclusions

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**?



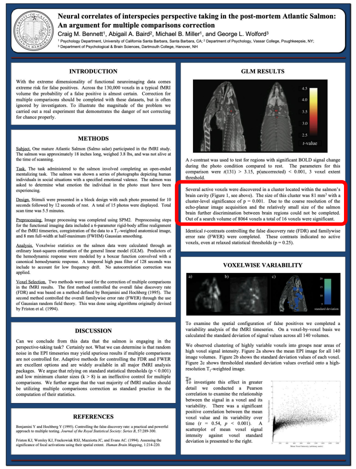
Where did this study fail (on purpose)?

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**?



Valid data analysis goes well beyond implementation correctness.