# CSE 503

Software Engineering

Winter 2021

**Empirical Research**

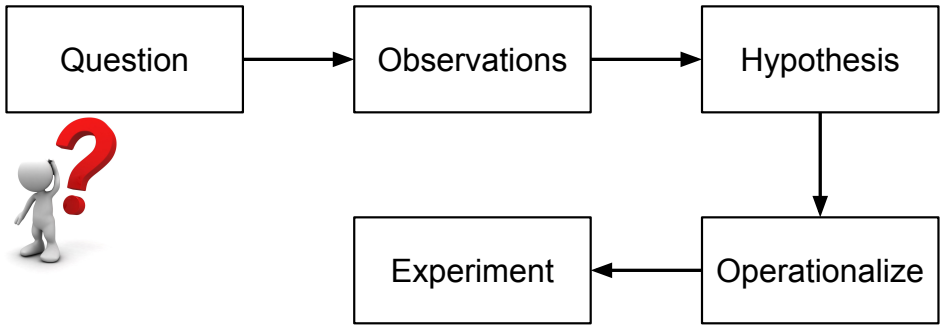
February 24, 2021

## Today

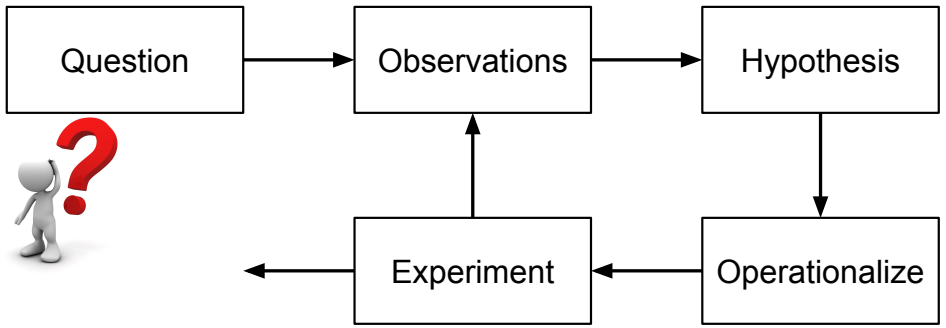
- **The scientific method**
- Study design and validity
- Paper discussion

## The scientific method


## The scientific method

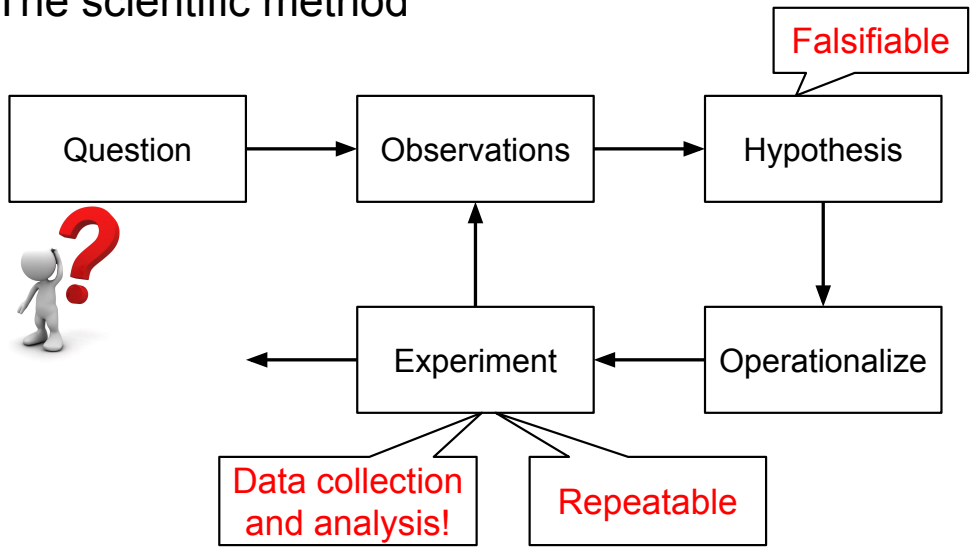


## The scientific method




Seems quite simple. What's important?

## The scientific method



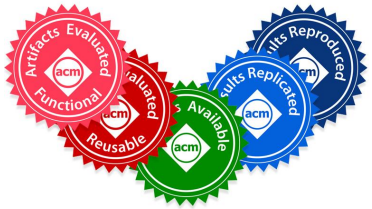

## Repeatability, replicability, and reproducibility

- **Repeatability**
  - Same research questions
  - Same experimental setup and artifacts
  - Same team
- **Reproducibility**
  - Same research questions
  - Same experimental setup and artifacts
  - Different team
- **Replicability**
  - Same research questions
  - Different experimental setup and artifacts
  - Different team


Note: the ACM defined replicability and reproducibility in the opposite way of most other scientific fields .. recently fixed!

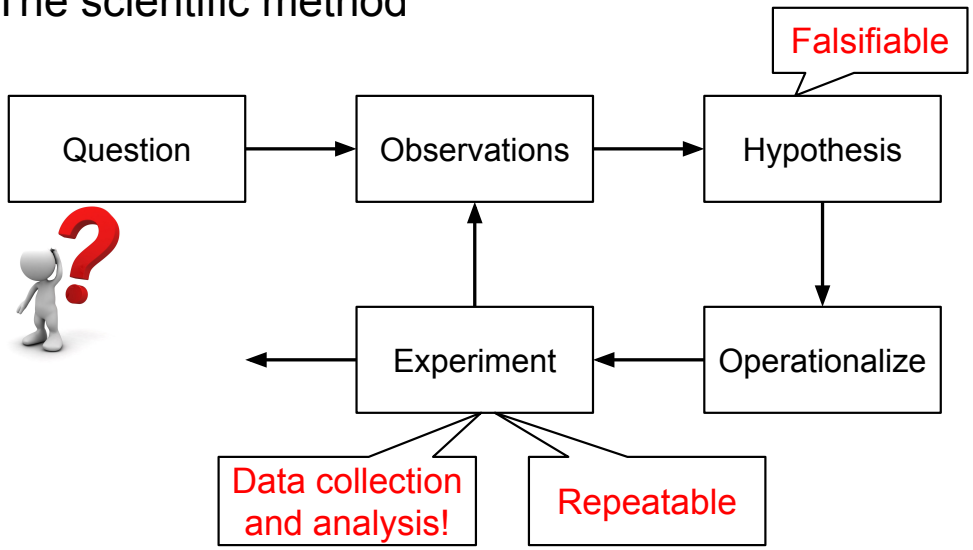
## Repeatability, replicability, and reproducibility

- **Repeatability**
  - Same research questions
  - Same experimental setup and artifacts
  - Same team
- **Reproducibility**
  - Same research questions
  - Same experimental setup and artifacts
  - Different team
- **Replicability**
  - Same research questions
  - Different experimental setup and artifacts
  - Different team

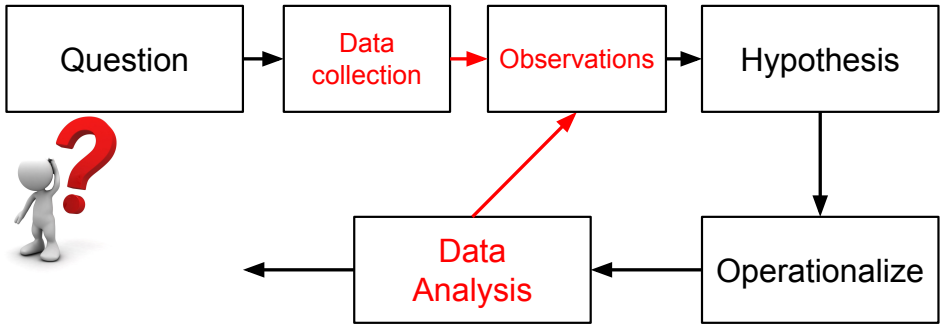



Does this even matter?

## The scientific method




## The scientific method: common mistake




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

## My favorite “scientific” 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?
- I don’t understand these intervals, can you give a p value?
- Most papers are wrong or later obsolete, so who cares?
- Don’t be naive, science is about papers not impact.

## My favorite “scientific” 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?
- I don’t understand these intervals, can you give a p value?
- Most papers are wrong or later obsolete, so who cares?
- Don’t be naive, science is about papers not impact.

Avoid confirmation bias; always assume you screwed up :)

## My favorite “scientific” 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?
- I don’t understand these intervals, can you give a p value?
- Most papers are wrong or later obsolete, so who cares?
- Don’t be naive, science is about papers not impact.

Transform intuition and expectations into testable hypotheses!

## My favorite “scientific” 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?
- I don’t understand these intervals, can you give a p value?
- Most papers are wrong or later obsolete, so who cares?
- Don’t be naive, science is about papers not impact.

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

## My favorite “scientific” 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?
- I don’t understand these intervals, can you give a p value?
- Most papers are wrong or later obsolete, so who cares?
- Don’t be naive, science is about papers not impact.

No comment!

## Today

- The scientific method
- **Study design and validity**
- Paper discussion

# Kinds of conceptual variables

### Dependent variable

- Outcome variable -- the measured response.

### Independent variable

- Experimental variable -- systematically manipulated/controlled.

### Covariate

- Experimental variable -- measurable but not controllable.

# Study designs

### Between subjects design

- Independent variable(s) take on exactly one value for each subject.

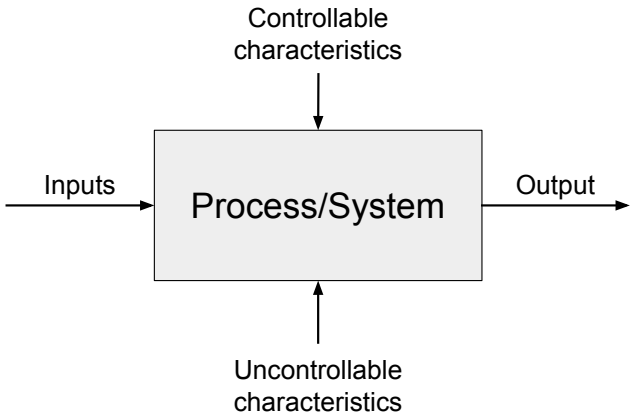
### Within subjects design

- Independent variable(s) take on multiple/all possible values for each subject.
- Repeated measures design.

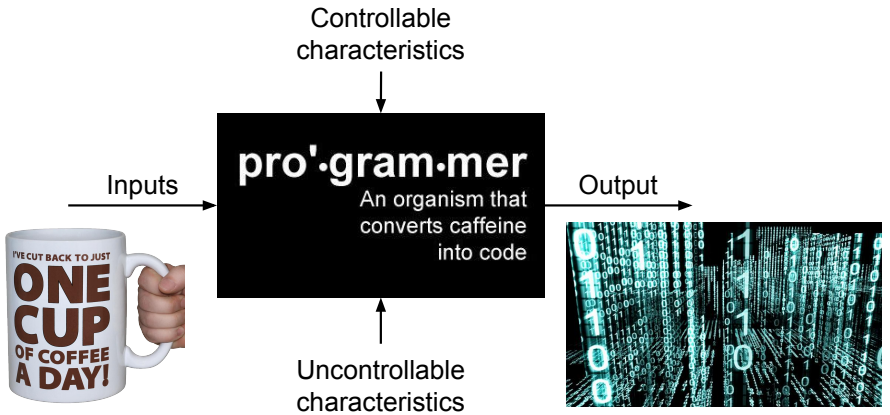
### Mixed design

- A mixed design of between-subjects variables and within-subjects variables.

# Example




# Example



# Example experiment

pro'·gram·mer
An organism that converts caffeine into code

## High-level research question:

Does coffee consumption improve programmer productivity and code quality?

**Operationalization 1:**

- 20 participants code for 20 weeks: on project 1 on Mondays with coffee; on project 2 on Fridays without coffee.
- Code quality: number of defects encountered in each project.
- Productivity: number of lines of code written.
- Coffee consumption: dollars spent on coffee (Monday receipts).

**Operationalization 2:**

- 20 participants, randomly assigned to two groups of 10: each participant gets the same coding assignment.
- Code quality: number of defects encountered in the assignment.
- Productivity: number of lines of code written.
- Coffee consumption: Participants in group 1 get a free 64oz coffee.

# Types of variables

- **Categorical** (nominal)
  - Unordered set of values
  - Example: [HCI, PLSE, Robotics, UbiComp]
- **Dichotomous** (dichotomized or "natural" dichotomy)
  - Categorical with exactly two possible values
  - Example: [Day, Night]
- **Ordinal**
  - Ordered set of values (no assumption about equidistant values)
  - Example: [low, medium, high]
- **Continuous/Interval**
  - Ordered values (equidistant values)
  - Example: [0..100]

# Kinds of studies

## Experiment

- Independent **variable(s)** are **directly manipulated**/controlled.
- Repeatable with a testable hypothesis.
- Randomization (e.g., counterbalancing for within-subjects designs).

## Observational study

- **Variables** are **not manipulated**/controlled.
- Useful if an experiment is impractical/unethical.
- Greater risk of spurious correlations.

For example: assessing the harm of smoking.

# Kinds of studies

## Experiment

- Independent **variable(s)** are **directly manipulated**/controlled.
- Repeatable with a testable hypothesis.
- Randomization (e.g., counterbalancing for within-subjects designs).

## Observational study

- **Variables** are **not manipulated**/controlled.
- Useful if an experiment is impractical/unethical.
- Greater risk of spurious correlations.

**Parachute use to prevent death and major trauma when jumping from aircraft: randomized controlled trial**

*BMJ* 2018 ; 363 doi: https://doi.org/10.1136/bmj.k5094 (Published 13 December 2018)
Cite this as: *BMJ* 2018;363:k5094

# Kinds of studies

## Experiment

- Independent **variable(s)** are **directly manipulate**
- Repeatable with a testable hypothesis.
- Randomization (e.g., counterbalancing for within-

## Observational study

- **Variables** are **not manipulated**/controlled.
- Useful if an experiment is impractical/unethical.
- Greater risk of spurious correlations.

**Parachute use to prevent death and major trauma when jumping from aircraft: randomized controlled trial**

*BMJ* 2018 ; 363 doi: https://doi.org/10.1136/bmj.k5094 (Published 13 December 2018)
Cite this as: *BMJ* 2018;363:k5094

# Kinds of studies

## Experiment

- Independent **variable(s)** are **directly manipulated**/controlled.
- Repeatable with a testable hypothesis.
- Randomization (e.g., counterbalancing for within-subjects designs).

## Observational study

- **Variables** are **not manipulated**/controlled.
- Useful if an experiment is impractical/unethical.
- Greater risk of spurious correlations.

## Case study

- Focus on one particular subject (“deep dive”).
- Useful for qualitative analyses and interpretation of results.

# Today

- The scientific method
- Study design and validity
- **Paper discussion**

# Is computer science science?

Paper discussion:

- CS = science, engineering, and mathematics.
- *“CS is a grab bag of tenuously related areas thrown together”*
- *“CS is not a science, and its ultimate significance has little to do with computers”*
- *“Computing is not a science because it studies man-made objects”*
- *“Most scientific fields have saturated”*
- *“Science will never again yield revelations as monumental as the theory of evolution, general relativity, quantum mechanics, …”*
- *“Has computer science already made all the big discoveries it’s going to? Is incremental progress all that remains?”*
- CS constantly forms new relationships with other fields => new fields.
- Overclaiming (empty promises) hurts the credibility of CS.
- Is the scientific method applicable to CS?

# Is computer science science?

## Paper discussion:

- CS = science, engineering, and mathematics.
- „
- *Latour defines science-in-the making as the processes by which scientific facts are proposed, argued, and accepted. A new proposition is argued and studied in publications, conferences, letters, email correspondence, discussions, debates, practice, and repeated experiments. It becomes a "fact" only after it wins many allies among scientists and others using it. To win allies, a proposition must be independently verified by multiple observations and there must be no counterexamples.*

  ***Latour sees science-in-the making as a messy, political, human process, fraught with emotion and occasional polemics.***

# Should computer scientists experiment more?

## Paper discussion:

1. Is computer science an experimental science?
2. What can we learn from the Knight-and-Leveson experiment?
3. Traditional scientific method isn't applicable.
4. The current level of experimentation is good enough (1998 vs. 2020).
5. Experiments cost too much.
6. Demonstrations will suffice (proof of concept is good enough).
7. There is too much noise in the way (the easy way out).
8. Progress will slow.
9. Technology changes too fast.
10. You'll never get it published.
11. Feature comparison is good enough (comparison on paper or verbally).
12. Trust your intuition.
13. Trust the experts.
14. Flawed experiments (unrealistic assumptions etc.).
15. Competing theories (RISC vs. CISC, OO vs. functional programming).
16. Soft Science and Misuse.