How-to-PhD

A Dummies Guide towards Research

Boris Glavic[1]

DBGroup

Illinois Institute of Technology

2023-03-28

[1] bglavic@iit.edu

Outline

1. What is (CS) Research?
2. How to Find and Cite Stuff?
3. How to Formalize and Prove Stuff?
4. How to Build Stuff?
5. How to Measure Stuff?
6. How to Communicate Stuff?
7. Conclusions

Who the f*** am I?

- **Primary Key (almost)**: Boris Glavic
- **Location**: Chicago, USA
- **Job title**: Associate Professor
- **What I like**: good research
- **Waht I don't like**: bad research

What is this talk about?

- As a new Ph.D. student you are immediately confronted with the enigma of scientific research
- **You are faced with many challenging questions:**
  - What is (CS) research?
  - What is the reality of life in academia?
  - How to do literature search?
  - How to find a (good) thesis topic?
  - How to learn about your research community?
  - How to answer theoretical research questions and formalize a problem?
  - How to build systems?
  - How to conduct scientific experiments?
  - How to communicate your research findings?
  - How to manage your adviser?

What is this talk about?

- As a new Ph.D. student you are immediately confronted with the enigma of scientific research
- **You are faced with many challenging questions:**
  - **What is (CS) research?**
  - What is the reality of life in academia?
  - **How to do literature search?**
  - How to find a (good) thesis topic?
  - How to learn about your research community?
  - **How to answer theoretical research questions and formalize a problem?**
  - **How to build systems?**
  - **How to conduct scientific experiments?**
  - **How to communicate your research findings?**
  - How to manage your adviser?

What is this talk about?

Deadly sins

- we will discuss the many deadly sins, traps to avoid as a researchers
- Don't sin and go to research hell

What is this talk about?

How to ascend?

- we will discuss how to ascend to research heaven

What is (CS) research?

What is research?

- **Basic science**: study stuff that exists in the world
- **Engineering discipline**: design stuff and evaluate it

What is (CS) research?

What is CS research

- CS is both a basic science and an engineering discipline
- We study fundamental properties of the world (e.g., complexity theory)
- We design new things and evaluate them (e.g., database systems & algorithms)

What is (CS) research?

Developing hypotheses (models) about the world

- Hypothesis have to be falsifiable!
- **Example**: *Is attending this talk a waste of time?*

What is (CS) research?

Formalizing models and making predictions

- We can formalize models that encode hypothesis and then make predictions
- **Example** *If attending the talk is a waste of time, then people attending the talk would not have learned anything new compared to people not attending the talk*

What is (CS) research?

Designing and conducting experiments to test hypothesis

- Designing experiments
  - **Example**: *let's split the workshop attendees into a control group that has to leave the room and a study group that attends the talk and compare their insights into research after the talk*
- Collect evidence for or against hypothesis based on careful interpretation of experimental results
  - **Example**: *some of the students leaving the room may have talked to a good mentor in the meanwhile*

Outline

Fear not!

- Everybody is a sinner to some degree!
- . . . but as in popular religions we can redeem ourselves by repenting and improving our behavior!

- Discuss deadly sins related to the questions posed before
- . . . and discuss how to ascend to (research) heaven

Outline

1. What is (CS) Research?
2. How to Find and Cite Stuff?
3. How to Formalize and Prove Stuff?
4. How to Build Stuff?
5. How to Measure Stuff?
6. How to Communicate Stuff?
7. Conclusions

Deadly sins

- **1.** Only being negative (**wrath**, **envy**)
- **2.** Ignore related work (**pride**)
- **3.** Excessive & lazy citation (**gluttony**, **sloth**, **pride**)
- **4.** Only citing upwards (**envy**, **sloth**)

Negativity

Typical Examples

- *System **INSERT COMPETITOR** is crap, because it does not support **INSERT SLIGHT VARIANT OF THE PROBLEM***
- *Clearly the authors of **INSERT COMPETITOR** are idiots, because their approach does not perform well on **INSERT RANDOM UNREALISTIC CORNER CASE***
- ***INSERT COMPETITOR** is inferior to our system, because we did implement **INSERT SMALL AND OBVIOUS EXTENSION***

Negativity

Why this is bad

- By being one-sided we loose objectivity
- We are not giving credit where credit is due
- Create a toxic community

Negativity

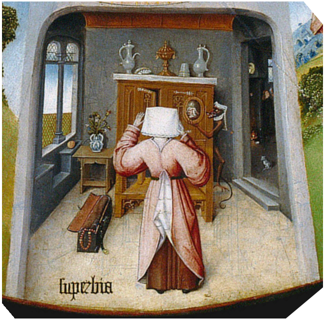
Why are people sinning?

- Misguided assumption that to elevate ones research it is necessary to disqualify / denigrate other research
- Strong emphasis on novelty in the community creates need to distinguish your work from others

Strategic or ignorant non-citation

Typical Examples

- *Ignore competitors because they are too similar*
- *Do not put in the effort to identify relevant related work*



Strategic or ignorant non-citation

Why this is bad

- Generates large amounts of overly similar papers
- The wheel is reinvented over and over again

Strategic or ignorant non-citation

Why are people sinning?

- Misguided attempts to claim novelty
- Time constraints
- Arrogance

Excessive & lazy citation

Typical Examples

- Cite many papers from the same project that overlap a lot in content
- Cite irrelevant / less relevant work
- Bias towards citing your own work

Excessive & lazy citation

Why this is bad

- Confusing the reader instead of highlighting the most relevant work

Excessive & lazy citation

Why are people sinning?

- Increase one citation count
- Not investing the time to identify the most relevant related work
- Lack of understanding of the field

Only citing upwards

Typical Examples

- *Cite big shots in the field only*
- *Cite only papers from top-10 universities*
- *Cite only papers from SIGMOD / VLDB / PODS*

Only citing upwards

Why this is bad

- Ignores good work published outside of top conferences and not from top universities
- Only quality / relevance of the work should count!

Only citing upwards

Why are people sinning?

- Time-consuming to search in other venues / for different authors
- Disrespect for venues / authors

How to ascend?

- **1.** Spend the effort to identify the objectively most important work
- **2.** Make citation decisions only based on quality / relevance of the work
- **3.** Be careful about citing your own work
- **4.** Choose *"standing on the shoulders of giants"* over *"defecating on the heads of gnomes"*

Outline

1. What is (CS) Research?
2. How to Find and Cite Stuff?
3. How to Formalize and Prove Stuff?
4. How to Build Stuff?
5. How to Measure Stuff?
6. How to Communicate Stuff?
7. Conclusions

Deadly sins

- 1. Avoiding formalization / theory (**sloth**, **pride**)
- 2. Omit proving *"trivial"* results (**sloth**, **pride**)
- 3. Overindulging in formalisms (**lust**, **gluttony**)

Avoiding formalization / theory

Typical Examples

- *"I do systems work, formalizations are useless non-sense"*
- *"What's the point of all this heavy notation?"*

Avoiding formalization / theory

Why this is bad

- Lack of formal problem definitions and notation leads to ambiguity / verbosity
- Proofs and notation help developing a field

Avoiding formalization / theory

Why are people sinning?

- Lack of background in theory
- Lack of appreciation for the benefits

Omitting proofs

Typical Examples

- *Proofs are omitted because of lack of space*
- *Proofs are omitted as they seem trivial*

Omitting proofs

Why this is bad

- A result that seems obvious may still be wrong
- External validation is important, but not possible without access to proofs
- Both the author and the reader can learn something for almost every proof

Omitting proofs

Why are people sinning?

- Lack of time
- Underestimation of complexities
- Overestimation of capabilities
- Not knowing that there are anonymous ways of providing supplementary materials

Overindulgence

Typical Examples

- *Introducing formal notation that is not utilized*
- *Using unnecessarily complex formal notation*

Overindulgence

Why this is bad

- Off-putting to readers: lot of investment for little reward
- Correctness is hard to verify
- Notation distracts from content

Overindulgence

ILLINOIS INSTITUTE OF TECHNOLOGY

Why are people sinning?

- Assumption that formal notation equals depth
- Lack of appreciation for KISS

How to ascend?

- **1.** A good formalization eliminates ambiguities of your ideas and exposes problems
- **2.** A good formalization helps others to understand your work
- **3.** By proving properties of the concepts you introduce, you learn more about your ideas
- **4.** Keep it lean and mean
- **5.** Don't be afraid of iterating over notation until it is appropriate

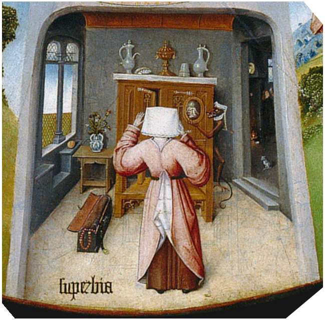
Outline

1. What is (CS) Research?
2. How to Find and Cite Stuff?
3. How to Formalize and Prove Stuff?
4. How to Build Stuff?
5. How to Measure Stuff?
6. How to Communicate Stuff?
7. Conclusions

Deadly sins

- **1.** Not implementing your algorithms (**pride**, **sloth**)
- **2.** Hard-coding your experiments (**sloth**)
- **3.** Not sharing code (**envy**, **sloth**)

No implementation



Typical Examples

- *No implementation of the algorithms*

No implementation

Why this is bad

- Missed opportunity to learn more about an idea / algorithm
- Problems are often just identified once they arise during implementation

No implementation

Why are people sinning?

- Lack of skills
- Lack of understanding what can be learned by implementing an algorithm

Hard-coded experiments

Typical Examples

- *Implementing specific experiments instead of a general algorithm*
- *"Simulating" the algorithm based on poor assumptions*

Hard-coded experiments

Why this is bad

- Results may not be representative of how an actual implementation may behave
- Problems may not materialize for the specific workload used in the experiment

Hard-coded experiments

Why are people sinning?

- Time crunch
- Overestimation of what can be learned from the behavior of the hard-coded examples
- Lack of implementation skills

Not sharing code

Typical Examples

- *Building a system and not open-sourcing it*
- *Not participating in reproducibility efforts*

Not sharing code

Why this is bad

- Lack of reproducibility and transparency
- The community can make progress if research can build on existing results

Not sharing code

Why are people sinning?

- Shame (my code is not good enough)
- Not willing to put in the time
- Under-appreciation of the benefits

How to ascend?

- **1.** Go the extra mile and fully implement your algorithm
- **2.** Building a full system is a lot of work but pays dividends in the long run
- **3.** Share your code! People may actually start to use your system!

Outline

1. What is (CS) Research?
2. How to Find and Cite Stuff?
3. How to Formalize and Prove Stuff?
4. How to Build Stuff?
5. **How to Measure Stuff?**
6. How to Communicate Stuff?
7. Conclusions

Deadly sins

- **1.** Bad hypothesis or lack of hypothesis (**sloth**)
- **2.** Apples & beef jerky comparisons (**sloth**)
- **3.** Only showing positive results (**envy**, **pride**)
- **4.** Lack of interpretation (**sloth**)

Bad hypothesis or lack of hypothesis

Typical Examples

- *We ran our system on workloads X, Y, Z*
- *We evaluated whether our system is better*

Bad hypothesis or lack of hypothesis

Why this is bad

- Confirmation bias
- Experiments that do not lead to insights

Bad hypothesis or lack of hypothesis

Why are people sinning?

- Coming up with good hypotheses is hard
- It is easier to describe what you have done then why you have done it
- Feeling the pressure to demonstrate how great your work is

Apples & beef jerky comparisons

Typical Examples

- *Comparing a standalone implementation against DBMS for performance*
- *Evaluating a system on use cases it was not designed for*

Apples & beef jerky comparisons

Why this is bad

- Unfair comparisons lead to unsound conclusions
- The field needs an even playing ground to make progress

Apples & beef jerky comparisons

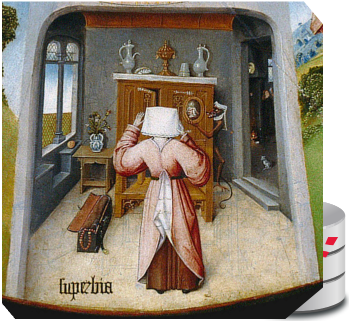
Why are people sinning?

- Lack of understanding of how such comparisons affect outcomes
- Lack of code availability
- Cherry-picking

Only showing positive results

Typical Examples

- *Our system outperformed competitors on **INSERT CHERRY-PICKED WORKLOADS***



Only showing positive results

Why this is bad

- Incomplete picture of the behavior of an approach
- Other research cannot build on your results
- Hurting other research that is not cherry-picking

Only showing positive results

Why are people sinning?

- Misguided impression that research that acknowledges limitations is less likely to be published
- Anxiety about your research being valued

Lack of interpretation

Typical Examples

- *System **X** did run 10 times faster than system **Y***
- *On workload **X**, system **Y** showed surprising results*

Lack of interpretation

Why this is bad

- More important than **how** approaches perform is **why** do they perform like this
- The even playing ground thing

Lack of interpretation

Why are people sinning?

- Interpretation is hard and requires more work

How to ascend?

- **1.** Formulate hypothesis upfront **before** you design your experiments
- **2.** Reflect on experimental results
- **3.** Show the full picture
- **4.** Identify meaningful comparisons

Outline

1. What is (CS) Research?
2. How to Find and Cite Stuff?
3. How to Formalize and Prove Stuff?
4. How to Build Stuff?
5. How to Measure Stuff?
6. How to Communicate Stuff?
7. Conclusions

Deadly sins

- **1.** Not motivating the problem (**sloth**)
- **2.** Not exploring & explaining choices (**sloth**)
- **3.** Lack of good examples (**sloth**)
- **4.** Too much / little technical details (**greed**, **sloth**)
- **5.** Lack of guidance for the audience (**pride**)

Motivation

Typical Examples

- *We improve the performance of* ***X***
- *We present a new algorithm for* ***X***
- *We study* ***INSERT UNMOTIVATED SMALL VARIATION ON EXISTING PROBLEM***

Motivation

Why this is bad

- Not giving the audience a reason to care
- Not telling the community how this work advances the state-of-the-art

Motivation

Why are people sinning?

- Lack of reflection on the *"why"*
- Lack of appreciation that a good motivation goes a long way

Exploring & explaining choices

Typical Examples

- *We use* ***INSERT RANDOM HEURISTIC***
- *To improve performance we* ***INSERT CORNER WE DID CUT***

Exploring & explaining choices

Why this is bad

- If the *"why"* is not clear, the *"how"* does not matter much
- Audience cannot judge soundness of your choices

Exploring & explaining choices

Why are people sinning?

- Reflection from the inside is hard
- Choices that are clear to you may not be clear to *"outsiders"*

Lack of good examples

Typical Examples

- *Introduce a technical concept without providing an example*
- *Argue a point without giving an example*

Lack of good examples

Why this is bad

- Good examples help the audience to follow what you are saying and confirm their understanding

Lack of good examples

Why are people sinning?

- Coming up with good, simple examples for complex concepts is hard
- Once you studied a problem long enough, things start to look trivial

Too much / little technical details

Typical Examples

- *Providing details that are irrelevant for the contribution*
- *Omitting details that are critical for understanding your approach*

Too much / little technical details

Why this is bad

- Details that distract from the main points
- Not giving the audience the chance to understand what you are doing

Too much / little technical details

Why are people sinning?

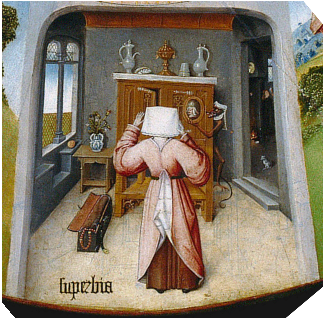
- Finding a good balance is hard
- Lack of reflection on *"Is this detail needed to understand the approach?"*

Lack of guidance for the audience

ILLINOIS INSTITUTE OF TECHNOLOGY

Typical Examples

- *Diving into technical details too early*
- *Omitting summaries of what has been discussed so far*
- *Omitting outlines of what is to come*
- *Not providing the motivation for what things will be used for*
- *Not exploiting the structure of a paper / talk*



Lack of guidance for the audience

Why this is bad

- Loosing the audience

Lack of guidance for the audience

Why are people sinning?

- Lack of space / time
- Things that are obvious to you are most likely not obvious to the audience!

How to ascend?

- **1.** Identify realistic use cases early on
- **2.** Clearly specify your contributions
- **3.** Spend the time to come up with good examples
- **4.** State the reasons for your choices
- **5.** Provide appropriate guidance to the audience

Outline

1. What is (CS) Research?
2. How to Find and Cite Stuff?
3. How to Formalize and Prove Stuff?
4. How to Build Stuff?
5. How to Measure Stuff?
6. How to Communicate Stuff?
7. Conclusions

Take-aways

- **We are all sinners**
    - don’t despair over mistakes
    - reflect on your behavior and improve
    - **research is a life-long learning experience**
- **Sound morals are essential**
    - science needs objective, rational, and honest scientists!
- **Withstand temptations**
    - Many "sins" lead to short term gains
    - . . . but will eventually ruin your reputation / negatively affect the quality of your research

Your PhD is just the beginning

- Finding good mentors is critically important
- Learn from positive / negative examples
- Don't despair! You are doing good work!
- Don't get overly confident / too comfortable either
- Have fun!

Things we did not cover

- How to find a good thesis topic / develop *"research taste"*
- How to become involved in the community?
- How to manage your adviser?
- How to establish collaborations?
- How to become involved in the community?
- How to manage your time?
- How to balance professional / personal life?