Common Knowledge in Research

Version 2.0 by Sherry in June 2021

Notes

1. The slides CANNOT cover all the common knowledge you need to know for research.
2. Please READ the policies and guidelines of the conference websites before your paper submission **BY YOURSELF**. You can learn a lot. **DO NOT BE LAZY~**
3. Please attend PDEV6770 and LANG5001. Keep all the things you have learned in mind.

What You Need to Have Before Conducting PhD in Database Area?

- **Technical Skills**
  - Programming languages such as C, C++, Java, Python, R…
  - Basic Linux skills
- **Theoretical Background**
  - Knowledge about NP optimization problems
    - e.g., A compendium of NP optimization problems
    - e.g., Computers and intractability: a guide to the theory of NP-completeness
    - e.g., COMP 5711 and COMP 5712 in HKUST
  - Knowledge about data structures and advanced algorithms
    - e.g., COMP3711 in HKUST
    - e.g., Graph Theory
- **Maths and English are very important!**

A PhD in HKUST

Suggestions about Taking PhD in CSE in HKUST…

- **Research is our major duty in PhD, so remember to read papers and conduct research every week!**
- **1-2 courses will be good enough in one semester (otherwise, you may have no time for research 🥵)**
  - According to most people, UG courses are usually more difficult than PG courses.
  - I think the following plan may be suitable for most people (especially in the semester when you have **TA duties**):
    - **1 PG course**
    - **1 UG course + 1 PG course**
    - **2 PG courses**

I took 3 PG courses in my first semester. I almost died in the final week. 😅

Our Weekly Meeting Routine

- **One small group meeting**
  - **More discussion is better! Asking questions is not stupid.**
  - Everyone shares their research progress

Only about research. Do not need to talk about our homework/courses/TA duties…

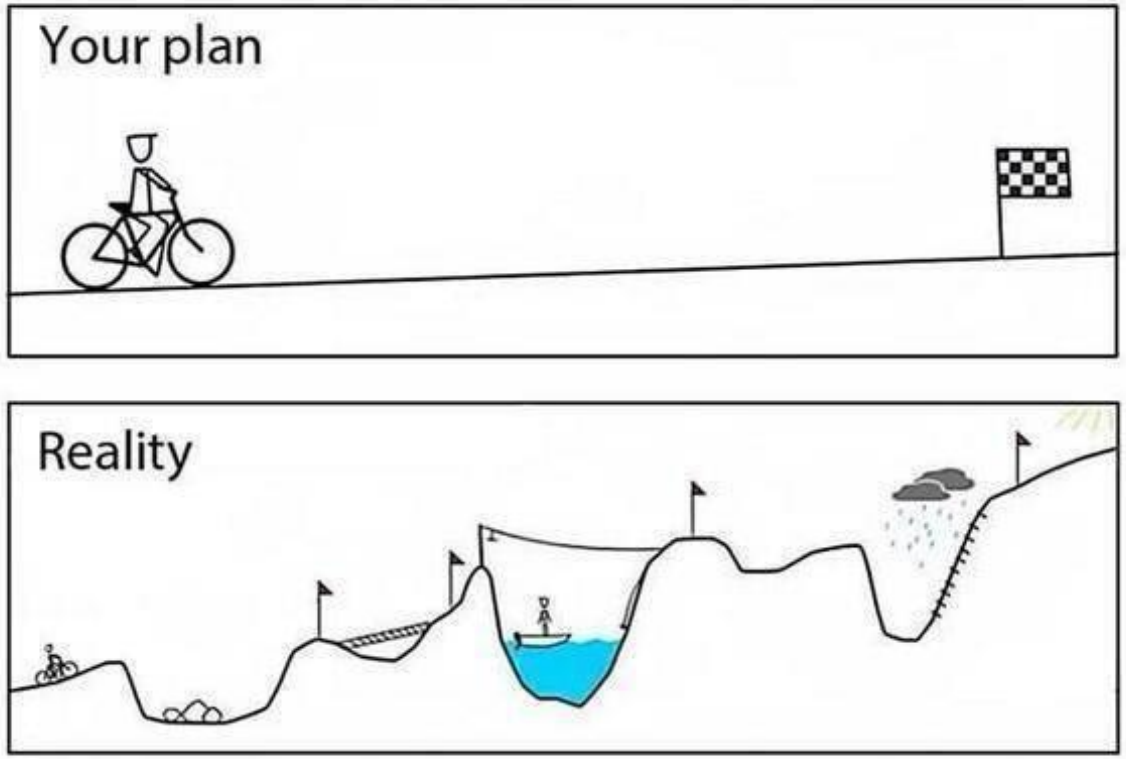
- **One large group meeting**
  - 1-2 people from each group presents the details of his/her current work

About Our Group!

- **Blockchain:** https://www.cse.ust.hk/blockchain/index.html
- **DB for AI**
- **Graph:** http://www.cse.ust.hk/biggraph/index.html
- **Knowledge Graph (KG):** https://www.cse.ust.hk/knowledgegraph/home.html
- **Machine Learning (ML):** https://www.cse.ust.hk/dbml
- **Spatial Crowdsourcing (STC):** https://www.cse.ust.hk/stc

May not be suitable for every PhD in the world, just some experience from my friends and me…

General Research Suggestions for PhD

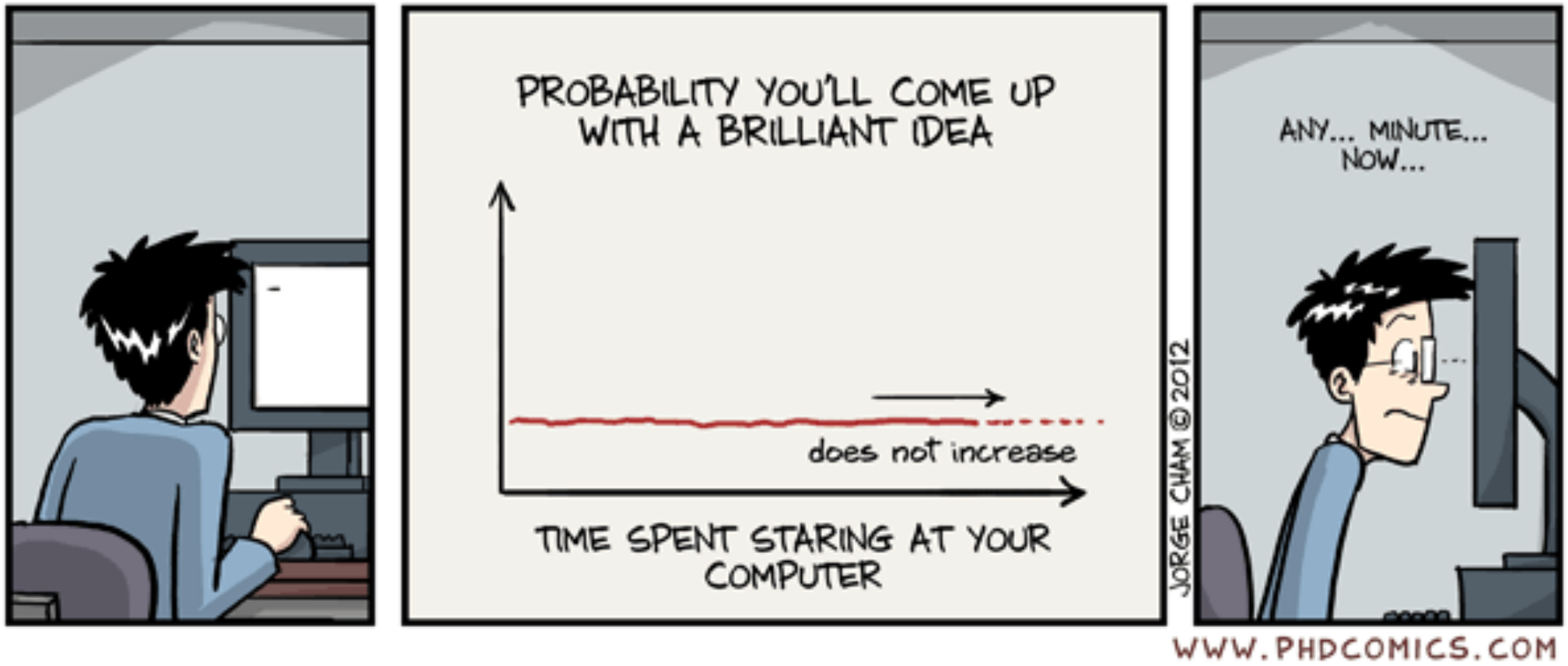


How to Find Interesting Topic?

1. Search for the **surveys** in certain topics in recent 3 years
2. Look for the proceedings in **DB/DM top conferences** in recent 3 years
   - SIGMOD/VLDB/ICDE/KDD/ICDM/TKDE/VLDBJ….
3. Find the hot topics and tutorials from **GitHub** (the awesome xxx/must-read-papers)
   - GNNPapers: https://github.com/thunlp/GNNPapers
   - Awesome-KGQA: https://github.com/BshoterJ/awesome-kgqa
   - KG group https://github.com/heathersherry/Knowledge-Graph-Tutorials-and-Papers
4. Follow the works published in **famous researchers** in this area
   - e.g., Jure Leskovec https://cs.stanford.edu/people/jure/
5. Read one paper in your interested topic (from top conference), then
   - Read its **related works** to know the state-of-the-arts techniques
   - Read the papers that **cite this paper** via Google Scholar

   In this way, you can find the **small "paper community"** that study this topic

How to Start Your Work?



How to Start Your Work?

Meanwhile, write down your thoughts and details in slides or latex (in case you forget)

1. Choose your **favorite topic**
2. Find the **datasets** used in most papers in this topic
3. Conduct some **experiments** to reproduce the results of state-of-the-arts techniques
4. Find the **limitations** in these experimental results, and conduct an **analysis**:
   - Why there are these bad cases? Are these bad cases common in most scenario?
   - Why is this step so slow?
5. Motivated by your analysis, start to think about whether there is a **research problem**
   - Can this be solved by any other existing method? 😅
   - Can this be solved by an engineering combination of other existing methods? 😅
   - If not, start to think whether you can formulate this as a research problem 🤩
6. **Implement** your naïve solution and start to improve it…

How to Present Ideas to the Others?

This is the content you need to write in your **Introduction**.

- **Describe the problem**
  - Is this problem important? Or just some rare cases in real-world scenarios? Is there any statistics or existing publication that proves its importance? Do you have motivating examples?
- **Present a summary (instead of details) of the solutions from the previous works**
  - What are the drawbacks of these previous works?
- **Present the general idea of your solution**
  - Why it is better than the previous works?

Use a clear and simple motivating example (better with figures) to illustrate your idea!

- **Present the technical challenges (2~3 challenges) of your solution**
  - Is it just an engineering modification or combination of the previous works?
  - If parts of your work are based on any previous work, do you **clarify the border** between previous works and **your contributions**?

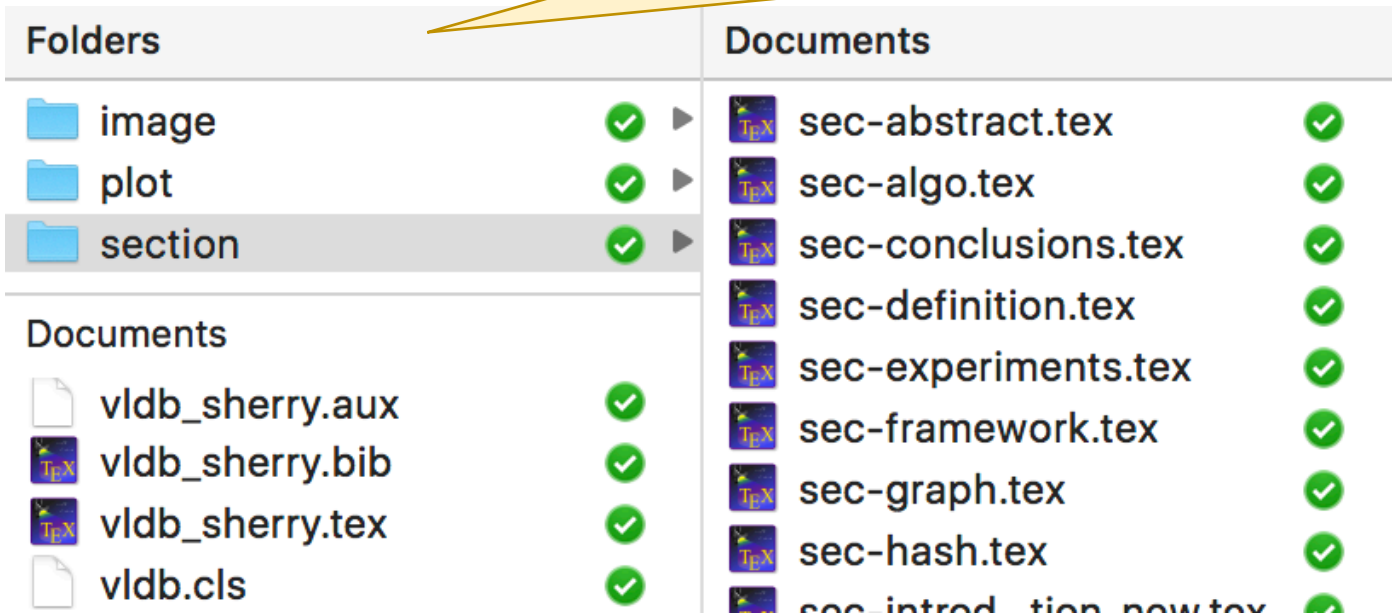
Notes: We present the idea to our supervisors and group members to **seek for comments and advices.** They can give their opinions whether your idea is **feasible and convincing**. **DO NOT ask them to directly give you the complete solution.** Otherwise, they can do another PhD based on your topic.

Which Tools to Use to Write Papers?

- **Paper Writing: LATEX**
  - **Tutorial:** http://www.docs.is.ed.ac.uk/skills/documents/3722/3722-2014.pdf
  - **Useful applications: TextStudio** https://www.texstudio.org/

- **Collaboration with other people**
  - **GitHub**
  - **Dropbox**
  - **Overleaf**

TIPS: Organize the files in folders, so that you and your collaborators can work on different parts of the paper without conflict.



How to Write Good Papers?

Good tool for typo and grammar checking:
Google online doc

- **Paper writing is as important as actual works:**
  - Treat your paper as **a textbook that will be published** instead of an experimental report or a course report.
  - If you are not a native English speaker and this is the first time you submit a research paper, it is better to revise each section **AT LEAST THREE TIMES.**

Ask your friends for help. Let them read your paper and tell you which parts they cannot understand. Revise that parts by yourself until **everyone understands everything in the introduction and abstract**. The key idea of a good paper should be understood by any undergraduate student.

  - Check the spelling and grammar **BY YOURSELF** before passing the paper to your coauthors and supervisors. DO NOT use them as Grammarly.

Reviewers may accept your paper if they understand your idea and solution CLEARLY.
They will definitely NOT accept your paper if it looks complicate and hard to follow.

How to Write Good Papers?

Suggestions:

1. **Use good notations**. You need to decide which notations to use at the very beginning of your work. DO NOT wait until you finish all the writing. DO NOT use **too many** notations in one paper, it confuses the readers.
2. **Use examples (with figures)** to demonstrate your algorithms or workflows.
3. **Be concreate (especially in experiment discussion)**. Concreate numbers (e.g., "87.3%") are ALWAYS BETTER than "many/a large number of/a few/largely improve/slightly improve".
4. **Explain every technical phrase** (e.g., "global similarity") when it appears in your text for the first time.
5. Use **vector graphics** (pdf, svg, eps…) instead of raster graphics (jpeg, gif, …).
6. Avoid redundant patterns/words/sentences in the text. Try to use synonyms and different sentence patterns to express the same meaning.
7. Make the "a/an, the, (plural)" in your paper correct!!! Otherwise it always confuses your readers!

- "We work on a KB." → We work on one KB. No specific.
- "We work on KBs." → We work on all KBs. No specific.
- "We work on the KB." → We only work on the specific KB that we mentioned above.

P.S. You can send your draft to the grammar tutor of our department (**dalton AT cse.ust.hk**). She will check your grammar and send some feedbacks after one week. However, she is not an expert in understanding your technical details. You need to check most of the technical parts **BY YOURSELF**.

What is Important in Paper Format?

- **Capitalize** the titles of your sections and subsections
- Use bullet points (\begin{itemize}), bold fonts, and italic fonts to **highlight the important parts**
- Leave a space between words and citation, e.g., "Kruskal's algorithm [1]", not "Kruskal's algorithm[1]". Use "~\cite{}" in LaTex.
- In the References section, make sure each citation **has the same format**

- M. C. Phan, A. Sun, Y. Tay, J. Han, and C. Li. Pair-linking for collective entity disambiguation: Two could be better than all. Transactions on Knowledge and Data Engineering, 31(7):1383–1396. IEEE, 2018.
- Z. Fang, Y. Cao, Q. Li, D. Zhang, Z. Zhang, and Y. Liu. Joint entity linking with deep reinforcement learning. The World Wide Web Conference, pages 438–447. ACM, 2019.

Choose either style. Use one style for all the papers.

- M. C. Phan, A. Sun, Y. Tay, J. Han, and C. Li. Pair-linking for collective entity disambiguation: Two could be better than all. TKDE, 31(7):1383–1396. IEEE, 2018.
- Z. Fang, Y. Cao, Q. Li, D. Zhang, Z. Zhang, and Y. Liu. Joint entity linking with deep reinforcement learning. WWW, pages 438–447. ACM, 2019.

- DO NOT just copy and paste the BibTex from Google Scholar. Most BibTex from Google Scholar is messy. **Check and Fix it by yourself. DO NOT BE LAZY~**
- **One suggestion is to directly use the BibTex from DBLP, which is more standard.**

What Conferences/Journals Do We Target At?

- **Conferences (you need to be aware of the deadlines every year):**
  - SIGMOD (2 ddls, usually one in June/July, the other in September/October), 12 pages
  - PODS (2 ddls, usually one in June, the other in December)
  - VLDB (ddl on 1st of each month), 12 pages
    - A rejected paper is only allowed to be resubmit after 12 months
  - ICDE (2 ddls, usually one in June/July, the other in October/November), 12 pages with references
  - KDD (February), 9+2 pages
  - Other top conferences in ML, such as ICML, NeurIPS, ICLR…
- **Journals:**
  - TKDE, VLDBJ, TODS…

Suggestions:

1. **Schedule a recycle plan**. You need to know the notification date of each conference every year. If a paper is rejected from the first submission, estimate the time to revise the paper and prepare to submit it to one of the next deadlines.
2. **Check the conference websites** for new publications and keynotes, to follow the hot topics. e.g., https://sigmod2020.org/sigmod_research_list.shtml

What is Important During Paper Submission?

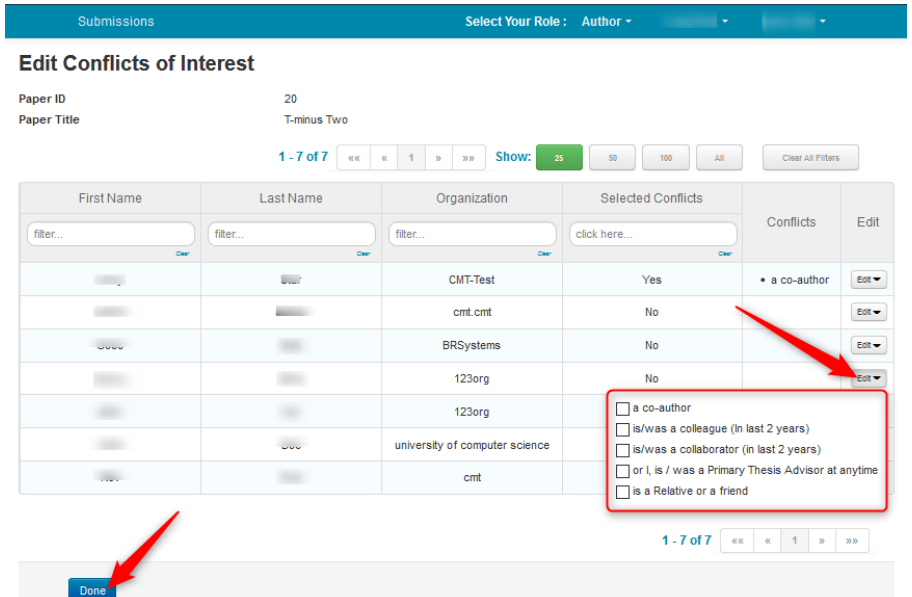
- **Fill in "Domain Conflict":**
  - **ust.hk; connect.ust.hk; cse.ust.hk**
  - Do not enter public webmail providers as institution domains
  - And the domains of your coauthors

DO NOT miss anyone!! This is important in academic honesty.

What is Important During Paper Submission?

- **Fill in the Conflict of Interest List (COI list):**
  - Add all the recent colleagues/recent collaborators/thesis advisor/friend/relatives of all the coauthors. You may need to read their Google Scholar or DBLP.
  - For fairness, these people will not have chance to bid and review your paper because you are too familiar with each other.
  - Ask your supervisor, your senior and your coauthors if you do not know who to fill in.

DO NOT miss anyone!!! This is important in academic honesty.



Some Common Knowledge of Paper Reviews

- **Mode of Reviews**
  - **Single-blind:** The reviewers can view the names of author, but the author does not know who the reviewers are. (e.g., VLDB, ICDE, applied science track of KDD)

    DO NOT submit low-quality papers to single-blind conferences. This damages the reputation of you and your coauthors.
  - **Double-blind:** The reviewers cannot view the names of author, and vice versa. (e.g., SIGMOD, research track of KDD)
  - **Triple-blind:** The reviewers and handling editors (PC chairs) cannot view the names of author, and vice versa. (e.g., ICDM)
  - **Open review:** The paper review is public. Anyone can review and comment the paper. (e.g. ICLR) https://iclr.cc/archive/www/doku.php%3Fid=iclr2017:faq.html)

Notes:

1. Check the conference websites for the modes of reviews every time before your submission.
2. For conferences in double-blind and triple-blind, you should not leak any information about your identities in the paper. **DO NOT mention** your name, your affiliations, your github address, the arXiv version of your paper (if there is any).
3. Different conferences have different rules in different years. Make sure you **read EVERY WORD in the submission guideline**. **DO NOT BE LAZY**!!

Some Common Knowledge of Paper Reviews

- **How to submit your code in a double-blind review conference?**
  - Nowadays almost all the conferences accept supplement materials in the submission.
  - You can create an anonymous Dropbox/GitHub/GoogleDrive account, and submit the corresponding link of your code in the “supplement materials”.

    Please provide a detailed README file, so that the reviewers can check the code and test.

  - As an example, please check the “Availability” part in SIGMOD2021 for more details:
    - https://2021.sigmod.org/calls_papers_sigmod_research.shtml

Some Common Knowledge of Paper Reviews

- **What will happen after you submit the paper?**

1) **Reviewers bid for papers they want to review:** During bidding, reviewers can access abstracts of all non-conflicting papers. They can bid as per their expertise and interest.

> **TIPS:** Before submission, you may read the list of PCs (e.g., http://sigmod2020.org/org_sigmod_pc.shtml , or in the COI candidate list) in the conference website and guess who will bid for your paper. Make sure you discuss or cite their works properly in your paper.

2) **Reviewers read the papers and write comments.** The common aspects include:
   - Originality/Novelty of the paper
   - Technical Quality of the paper
   - Impact/Outreach of the paper
   - Presentation
   - Reproducibility
   - Three positive aspects of the paper
   - Three negative aspects of the paper

> **TIPS:** Ask yourself the **SAME QUESTIONS** before you submit. What scores will you give to your paper regarding each of these issues if you are a reviewer? **Do you really want to submit?**

Some Common Knowledge of Paper Reviews

- **What will happen after you submit the paper?**

3) A **discussion** is usually initiated by the Chair or Meta-Reviewer after reviews have been submitted for papers with divergent recommendations. Reviewers' identities can be shared or kept anonymous during discussion.

Make sure your idea convinces everyone!

P.S. Usually (for most conferences and journals) there will be three reviewers for one paper. Sometimes if it is hard to decide whether a paper should accepted or rejected, the meta-reviewer or chair may invite one or two more reviewers for the paper.

- If you want to know more details about the review process, you can read some guidelines in CMT: https://cmt3.research.microsoft.com/docs/help/index.html (In the "How To" section, there is a introduction about what chairs/meta-viewers/reviews will do during the review).
- ACM SIGMOD 2021 Reviewer Instructions & Policies: https://docs.google.com/document/d/1NXgppEWth1OFXx1vEois5feAVicc6HzkjllrTHvBQ5s/edit?usp=sharing
- Anyway, the most important thing is to **submit a good work**!!

How to Make Good Slides for Presentation?

This is applicable to both **idea presentation** and **paper presentation** in the conference.

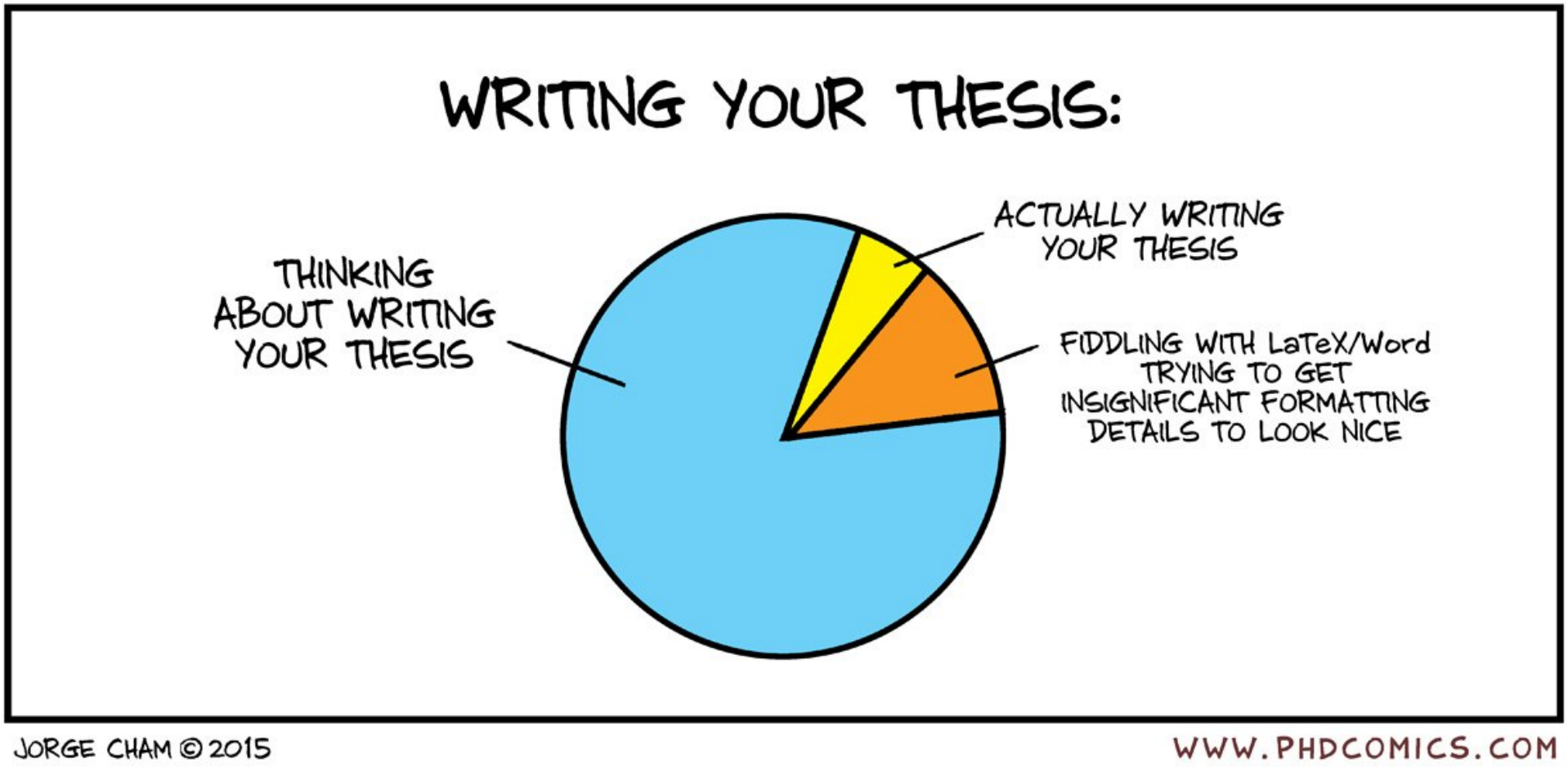
- **Make an Outline**
  - Background and Motivation (to make the audiences understand your problem)
  - Existing Solutions and Their Drawbacks
  - Your Solution/Framework/Algorithms
  - Experimental Results
  - Summary
- **DO NOT put too many words in one page**
  - Use bullets points instead of large paragraphs of text
- **Use figures and animations to demonstrate a simple running example of your algorithm**
- **Insert page number in each page,** so that it is convenient for the audiences to ask questions
- **Prepare some backup slides** that demonstrate technical details of your solution, or any problems that the audiences may ask

Some Good Materials for PhD

- How to Read a Paper, *by Prof. Srinivasan Keshav*.
- A Research to Engineering Workflow, *by Dr. Dustin Tran*.
- On Coding, Ego and Attention, *by Jose Browne*.
- Pre-requirement in CS PhD (Chinese), by *BJTU Synergy Lab*
- Crafting-Your-Research-Future, by *Prof. Charles X. Ling and Prof. Qiang Yang*
- How to Have Your Abstract Rejected, *by Mary-Claire van Leunen and Richard Lipton*

How to Be a Qualified PhD?

- **Be self-motivated. Time flies!**



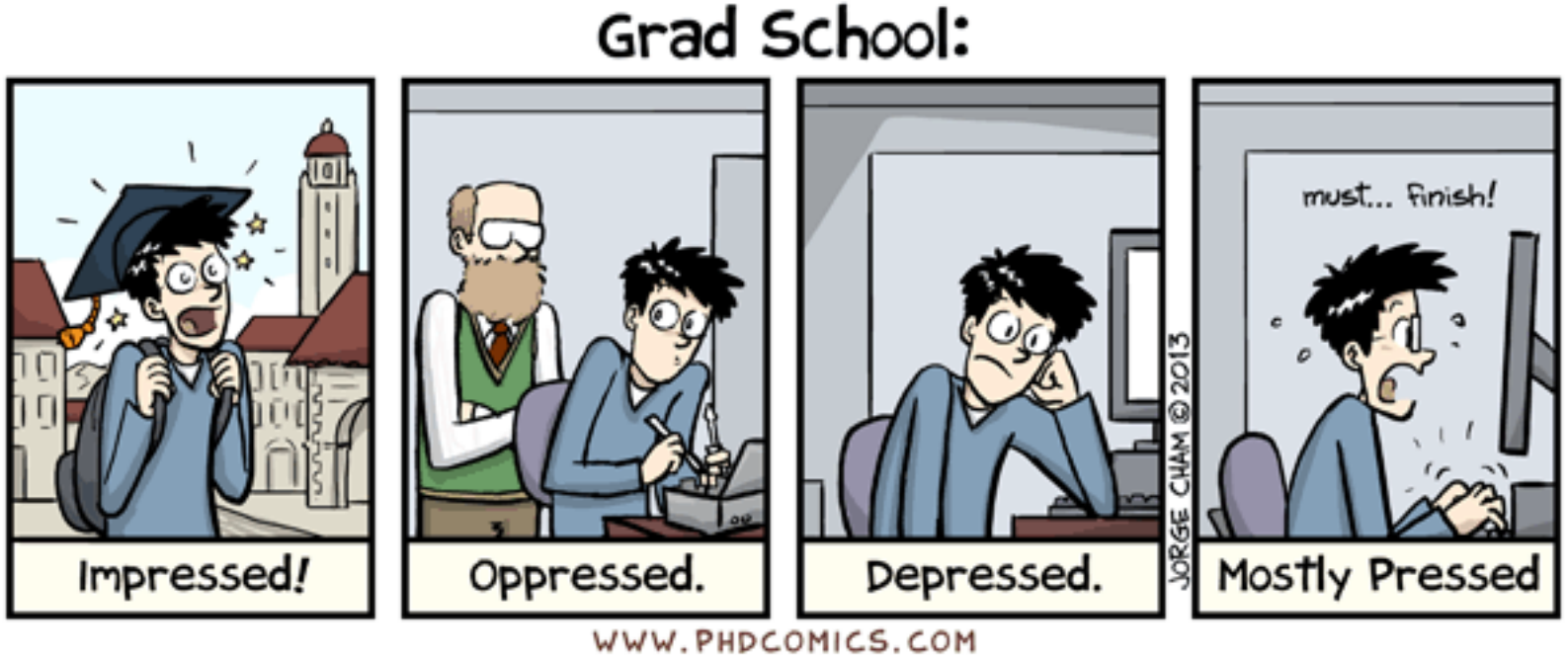
How to Be a Qualified PhD?

- **Conduct research independently.**
- **Seek for knowledge and solutions independently.**
  - **DO NOT wait** for your supervisors and seniors to give good ideas/pass good papers/pass good tutorials to you.
  - It is good to talk with your friends/group mates/lab mates/roommates/classmates (I believe at least one of them is doing PhD as well). **Learn something from their research habits and experience.**

This is the most important ability of PhD.

How to Be a Qualified PhD?

- Most people will come across **bad moments/failures/depression** during the PhD journey



- This is common and normal! Because doing research (and handling a lot of non-research issues such as courses/TA duties/projects at the same time) is **not easy**.
- Your **health** and **happiness** are always the most important thing.

Plagiarism

This is a very important issue, which is strongly related to the **reputation** of your **supervisor** and **every group member**.

Different societies have different policies. But most of them are similar. We take the policies from ACM as an example here.

Plagiarism (Defined from ACM Society)

- **Level I - Incidental violation:**
  - Verbatim copying of small portions (less than 2 sentences) of another author's paper with citing, but not clearly differentiating what text has been copied (e.g., not applying quotation marks correctly),
  - Or not citing the source correctly,
  - Or self-plagiarism/redundant publication, if there are citations but they are poorly placed or misleading
- **Penalty:**
  - Write **a letter of apology** to parties (e.g., editors, authors of prior works, co-authors) identified during the investigation, including an admission of wrong-doing.
  - Published items in the ACM Digital Library will be **updated with a corrigendum noting** the reasons for change and addressing the issue. Unpublished items will be **rejected** without further review.

Plagiarism (Defined from ACM Society)

- **Level II - Low-level violation:**
  - Verbatim copying of small portions (more than 3 sentences) of another author's paper with citing, but not clearly differentiating what text has been copied (e.g., not applying quotation marks correctly),
  - Or not citing the source correctly,
  - Or self-plagiarism/redundant publication, without citations

- **Penalty :**
  - Write **a letter of apology** to parties (e.g., editors, authors of prior works, co-authors) identified during the investigation, including an admission of wrong-doing.
  - Published items in the ACM Digital Library will be **withdrawn**. Unpublished items will be **rejected** without further review.

Plagiarism (Defined from ACM Society)

- **Level III - Moderate violation:**
  - Verbatim copying,
  - Or near-verbatim copying,
  - Or purposely paraphrasing **several elements** of another author‘s work,
  - Or copying elements of another author's work without citing the sources or without clearly delineating the source material
- **Penalty:**
  - Write **a letter of apology** to parties (e.g., editors, authors of prior works, co-authors) identified during the investigation, including an admission of wrong-doing.
  - Published items in the ACM Digital Library will be **withdrawn**. Unpublished items will be **rejected** without further review.
  - Banned from contributing to **any related ACM Venue** for **a full calendar year**.
  - **A letter explaining the violation, findings, and penalties** will be sent to the Dean, chair, or supervisor of each party found in violation.

Plagiarism (Defined from ACM Society)

- **Level IV - Significant violation:**
  - Verbatim copying,
  - Or near-verbatim copying,
  - Or purposely paraphrasing **a significant portion** of another author's work without citing the sources and without clearly delineating the source material
- **Penalty:**
  - Write **a letter of apology** to parties (e.g., editors, authors of prior works, co-authors) identified during the investigation, including an admission of wrong-doing.
  - Published items in the ACM Digital Library will be **withdrawn**. Unpublished items will be **rejected** without further review.
  - Banned from submission to **any ACM Venue** for the **next two years**.
  - **A letter explaining the violation, findings, and penalties** will be sent to the Dean, chair, or supervisor of each party found in violation.

Plagiarism (Defined from ACM Society)

- **Level V - Severe violation:**
  - Repeated violations (contains two or more violations mentioned above)
- **Penalty:**
  - Write **a letter of apology** to parties (e.g., editors, authors of prior works, co-authors) identified during the investigation, including an admission of wrong-doing.
  - Published items in the ACM Digital Library will be **withdrawn**. Unpublished items will be **rejected** without further review.
  - Banned from submission to **any ACM Venue** for the **next five years**.
  - The case and evidence will be forward to the ACM Committee on Professional Ethics for consideration.
  - **A letter explaining the violation, findings, and penalties** will be sent to the Dean, chair, or supervisor of each party found in violation.

Plagiarism

- **Policies:**
  - ACM (applicable to SIGMOD, KDD…) https://www.acm.org/publications/policies/plagiarism-overview
  - IEEE (applicable to ICDE, ICDM…) https://www.acm.org/publications/policies/plagiarism-overview Section 8.2.4d, Page 105
  - VLDB https://www.vldb.org/pvldb/policies.html

Please attend PDEV6770 seriously. There are sessions such as *Publication Ethics, Academic Integrity & Intellectual Property,* and *Research Ethics Case Studies.* Most of these sessions are **compulsory for all HKUST RPG students**. The contents include (but not limited to):

- Judging What is Plagiarism & How to Avoid It
- Differences between Plagiarism and Copyright Violation
- Hong Kong Copyright Law and Legitimate Use and Fair Dealing
- HKUST Intellectual Property Policies

Please keep everything you have learned in mind.

Plagiarism – How to Avoid?

Conclusions:

1. **Cite EVERY paper and material** you refer to during your writing/coding **properly**.
2. **DO NOT USE DIRECT QUOTES From Published Material:** In 99.99% of the cases, the information you want from a research article is an objective result or interpretation. How the author stated this information, i.e., *their prose*, is of little importance compared to the results or interpretations themselves. Take the information and ***put it into your own words***; avoid paraphrasing since this can potentially lead to plagiarism.
3. Draw **EVERY** component in the figures **BY YOURSELF**. DO NOT CROP others' figures.
4. If a theorem is already proved in another work, cite that work and use the theorem. **DO NOT write the proof again**.
5. **DO NOT copy** anything from your **previous papers** or **any paper of your coauthors**. This is self-plagiarism. All the penalties apply to self-plagiarism.
6. DO NOT ask questions such as "I think this paper has also copied something from that paper, why it is not punished?". *Others are idiots* does not mean that *you also have to be an idiot*.
7. A paper will be related with the reputation of you and your coauthors **for a whole life time**. DO NOT take any risk.

Wish You All The Best in Your Research Path

I sincerely acknowledge all the collaborators of my previous works and my group members for sharing their good research habits and experience with me.

I also acknowledge my friends, Richard, Rongxin, Yepang, Ying, and Guangneng, for sharing their common knowledge of research with me. ☺

Feel free to share your comments and knowledge in the slides!