Information for New Grad Students

(See Acknowledgements and License)

Goals of the Document

This document aims to set the expectations for how to produce sound quantitative research. You should use the document as a guide to catch up on different topics and terminology that you need to understand. Not only is good experimental methodology essential to produce good science, but a poor methodology is likely to see your paper rejected.

Mistakes Can Happen

History has shown that one should not believe every result in a paper without critical thought. Shortcomings can happen. In computing, shortcomings happen, maybe from ignorance, inappropriate extrapolation, oversight, or technical limitations. The following is a selection of examples:

- Binary decision diagrams (BDDs) were proposed as a good abstraction for context sensitive analysis in programs. The result influenced a number of works exploring and using BDDs for that purpose. Follow-up work [LH2006] showed that the original experimental result had collapsed cycles in the call graph and thus had less precision than other work.
- PIN made a fundamental and lasting contribution to the area of dynamic binary instrumentation. Nevertheless, the original paper [LCM+2005] contains a flaw in the summarization of the experimentation results.
- Early works evaluating the prospective performance gains of using GPGPUs showed large speedup factors. A debunking paper showed that these works ignored important types of overhead and that the overall speedup for the whole system is much lower.
  http://dl.acm.org/citation.cfm?id=1816021
  http://sbel.wisc.edu/Courses/ME964/Literature/LeeDebunkGPU2010.pdf
- Sometimes students also reflect on their work and respond to criticism:
  http://www.eecs.berkeley.edu/~sangjin/2013/02/12/CPU-GPU-comparison.html
- Panny [Panny2010] tells the history of experiments and theorems related to binary search trees that were contradicted by subsequent work.

[Panny2010] Panny, Wolfgang. "*Deletions in random binary search trees: A story of errors.*" Journal of Statistical Planning and Inference 140.8 (2010): 2335-2345. http://dx.doi.org/10.1016/j.jspi.2010.01.028

[ZC2004] J. Zhu and S. Calman. *Symbolic pointer analysis revisited*. In Proceedings of PLDI 2004, pages 145–157. ACM Press, 2004.

[LH2006] Lhoták O., Hendren L., *Context-sensitive points-to analysis: is it worth it?*, CC 2006, Vienna, March 2006.

[LCM+2005] Chi-Keung Luk, Robert Cohn, Robert Muth, Harish Patil, Artur Klauser, Geoff Lowney, Steven Wallace, Vijay Janapa Reddi, Kim Hazelwood. "*Pin: Building Customized Program Analysis Tools with Dynamic Instrumentation*," in Proceedings of the ACM SIGPLAN 2005 Conference on Programming Language Design and Implementation (PLDI). Chicago, Illinois, USA. June 2005, pages 191-200.

The Value of Data

Data is the key to validate practical innovation [DFJ+2012, PP2007, Tic1998]. Data can also provide insight towards new research directions.

When needing data for research, three situations may arise: (1) no data is available, (2) some sort of data is available, and (3) the perfect data is available. Based on experience, Category 1 is the most frequently occurring situation followed by Category 2. Category 3 will mostly only occur for follow-up paper.

Therefore, love the data you already have. If data is available, use it to the extent that it is applicable to the experiment, and no further. Make sure that any claims you make do not overreach the data used.

[DFJ+2012] Desprez F, Fox G, Jeannot E, Keahey K, Kozuch M, Margery D, Neyron P, Nussbaum L, Perez C, Richard O, et al. Supporting Experimental Computer Science. Technical Report, Argonne National Laboratory Technical Memo, 2012.

[PP2007] Peterson L, Pai VS. Experience-Driven Experimental Systems Research. ACM Communications 2007; 50(11):38–44, doi:10.1145/1297797.1297820. URL http://doi.acm.org/10.1145/1297797.1297820.

[Tic1998] Tichy WF. Should Computer Scientists Experiment More? IEEE Computer 1998; 31(5):32–40.

Terminology For Experimentation

At the moment, a comprehensive glossary for terminology relating to experimentation is unavailable. The following glossary and pointers provide information:

- http://www.itl.nist.gov/div898/handbook/glossary.htm
- Ronald F. Boisvert. Reproducibility in Computing. Talk at Dagstuhl Seminar 16111 on Rethinking Experimental Methods in Computing. March 14, 2016.

Experimental Design

An experiment is something that should not be taken lightheartedly. There exist structured methods for planning and executing experiments. Experiment design is an established statistical discipline which deals with how to run experiments efficiently given a set of controlled and random variables.

- Chapter 1-11 of Raj Jain. The Art of Computer Systems Performance Analysis. Wiley. 1991. http://www.amazon.com/The-Computer-Systems-Performance-Analysis/dp/047150336
- Chapters 1-7 of David Lilja: Measuring Computer Performance: A Practitioner's Guide. http://www.amazon.com/Measuring-Computer-Performance-Practitioners-Guide/dp/0521646707
- S. E. Maxwell and H. D. Delaney. Designing Experiments and Analyzing Data: a Model Comparison Perspective. Routledge, 2004.
- Larry B. Christensen, R. Burke Johnson, Lisa A. Turner. Research Methods, Design, and Analysis, 11th Edition. http://www.amazon.com/Research-Methods-Design-Analysis-11th/dp/0205701655
- Andrew J. Ko, Thomas D. LaToza, and Margaret M. Burnett. A Practical Guide to Controlled Experiments of Software Engineering Tools with Human Participants.
- Statistical Analysis: An Interdisciplinary Introduction to Univariate & Multivariate Methods by Sam Kash Kachigan. Radius Pr. 1986.
- Les Kirkup: Experimental Methods: An Introduction to the Analysis and Presentation of Data http://www.amazon.com/Experimental-Methods-Introduction-Analysis-Presentation/dp/0471335797
- NIST/SEMATECH e-Handbook of Statistical Methods, http://www.nist.gov/itl/sed/gsg/handbook_project.cfm

Interesting links from this submission site: http://www.cs.amherst.edu/ccm/wea08/submission.html

- A Theoretician's Guide to the Experimental Analysis of Algorithms by David S. Johnson http://www.research.att.com/~dsj/papers.html
- Algorithm Engineering by Camil Demetrescu, Irene Finocchi, and Giuseppe F. Italiano. http://citeseer.ist.psu.edu/712588.html
- How to Present a Paper on Experimental Work with Algorithms by Catherine C. McGeoch and Bernard M.E. Moret. http://citeseer.ist.psu.edu/328002.html
- Presenting Data from Experiments in Algorithmics by Peter Sanders http://citeseer.ist.psu.edu/504969.html

Benchmarks

Often data is unavailable for the specific practical innovation under consideration. Consequently, you will have to design a workload or use a benchmark that is appropriate for your purposes. “The choice of benchmarks and benchmarking methodology can therefore have a significant impact on a research field, potentially accelerating, retarding, or misdirecting energy and innovation” [Blackburn et al, OOPSLA, 2006].

Researchers have investigated properties and criteria for benchmark selection and creation.

- The DaCapo paper describes how the process and lessons learnt in creating a good benchmark. (See DaCapo citation in the section on Example Artifacts)
- Books on experimentation in computing contain material on the use and misuse of benchmarks. (See the book by Raj Jain and the book by David Lilja in the section on Experimental Design)

Data Analysis

The essential step after conducting the experiment is to analyze the data. For decades, statisticians have researched theory, methods, and tools to draw valid conclusions from data. Understand when techniques are appropriate to use. Report precisely any assumptions you have made, and which techniques you have used.

- Catherine C. McGeoch. A Guide to Experimental Algorithmics. Cambridge University Press. 2012

Look at and understand your data; do not apply statistical methods blindly. For example, it is often useful to visualize it first (the online NIST handbook cited in the part on experimental design also presents a number of useful data visualization methods).

Parametric statistics

- See the book by Raj Jain and the book by David Lilja in the section on Experimental Design.
- A Brief Introduction to Inferential Statistics Dario Basso in Experimental Methods for the Analysis of Optimization Algorithms. Appendix A. http://dx.doi.org/10.1007/978-3-642-02538-9

- Paul Cohen. Empirical Methods for Artificial Intelligence. Chapter 4.

Non-parametric statistics

- Paul Cohen. Empirical Methods for Artificial Intelligence. Chapter 5.
- Larry Wassermann, All of statistics (has both parametric and non-parametric method, including modern methods, but presented in very concise way)

If you encounter a specific term used in statistics, check the “Encyclopedia of Measures and Statistics” by Neil J. Salkind.

Artifacts

Several conferences are now recognizing high-quality artifacts in order to encourage and reward creation of software tools, proofs, databases, etc. that can be reused by others

An artifact associated with a publication should be (taken from http://evaluate.inf.usi.ch):

- **Consistent with the paper.** Does the artifact substantiate and help to reproduce the claims in the paper?
- **Complete.** What is the fraction of the results that can be reproduced?
- **Well documented.** Does the artifact describe and demonstrate how to apply the presented method to a new input?
- **Easy to reuse.** How easy is it for others to reuse it?

These resources provide information on artifacts:

- Artifact evaluation website (http://www.artifact-eval.org/)

Example Artifacts

Artifacts can be considered successful by various metrics. Judith Bishop advocates “PEAR: Perception, engagement, adoption, recruitment.” An artifact can be valued on how it is perceived by the community, whether it engaged the community, how others adopted the artifact and the ideas associated with the artifact, and finally whether the artifact gained influence in recruitment decisions of industry.

- DaCapo Benchmark Suite
  Stephen M. Blackburn, Robin Garner, Chris Hoffmann, Asjad M. Khang, Kathryn S. McKinley, Rotem Bentzur, Amer Diwan, Daniel Feinberg, Daniel Frampton, Samuel Z. Guyer, Martin Hirzel, Antony Hosking, Maria Jump, Han Lee, J. Eliot B. Moss, Aashish Phansalkar, Darko Stefanović, Thomas VanDrunen, Daniel von Dincklage, and Ben Wiedermann. 2006. The DaCapo benchmarks: Java benchmarking development and analysis. In Proceedings of the 21st annual ACM SIGPLAN conference on Object-oriented programming systems, languages, and applications (OOPSLA '06). ACM, New York, NY, USA, 169-190. doi: 10.1145/1167473.1167488
- MiBench benchmark
  M. R. Guthaus, J. S. Ringenberg, D. Ernst, T. M. Austin, T. Mudge and R. B. Brown, "MiBench: A free, commercially representative embedded benchmark suite," Workload Characterization, 2001. WWC-4. 2001 IEEE International Workshop on, 2001, pp. 3-14. doi: 10.1109/WWC.2001.990739
- JikesRVM (formerly known as Jalapeno virtual machine)
  B. Alpern, C. R. Attanasio, J. J. Barton, M. G. Burke, P. Cheng, J.-D. Choi, A. Cocchi, S. J. Fink, D. Grove, M. Hind, S. F. Hummel, D. Lieber, V. Litvinov, M. F. Mergen, T. Ngo, J. R. Russell, V. Sarkar, M. J. Serrano, J. C. Shepherd, S. E. Smith, V. C. Sreedhar, H. Srinivasan, and J. Whaley. 2000. The Jalapeño virtual machine. IBM Syst. J. 39, 1 (January 2000), 211-238. doi: 10.1147/sj.391.0211
- SimpleScalar: http://www.simplescalar.com/
- DIMACS Challenge: http://dimacs11.cs.princeton.edu/

- SteinLIB (http://steinlib.zib.de/steinlib.php), MIPLib (http://miplib.zib.de/), TSPLib (http://comopt.ifi.uni-heidelberg.de/software/TSPLIB95/), andORLibrary (http://people.brunel.ac.uk/~mastjjb/jeb/info.html)
- UMass Trace Repository: http://traces.cs.umass.edu/index.php
- Qualitas Corpus, a curated collection of software systems intended to be used for empirical studies of code artefacts: http://qualitascorpus.com/
  Ewan Tempero, Craig Anslow, Jens Dietrich, Ted Han, Jing Li, Markus Lumpe, Hayden Melton and James Noble 'Qualitas Corpus: A Curated Collection of Java Code for Empirical Studies' 2010 Asia Pacific Software Engineering Conference (APSEC2010), pp336–345, December 2010.
- Dror Feitelson's Parallel Workloads Archive: http://www.cs.huji.ac.il/labs/parallel/workload

See also [Dror Feitelson: Experimental Computer Science: The Need for a Cultural Change, USENIX Workshop on Experimental Computer Science, http://www.cs.huji.ac.il/~feit/papers/exp05.pdf] for more artifact repository links (and good high level discussion).

High-level Guides and Stories on Experimentation

The following work provides anecdotal evidence and experience on experimentation:

- Ian P. Gent, Stuart A. Grant, Ewen MacIntyre, Patrick Prosser, Paul Shaw, Barbara M. Smith, Toby Walsh. How Not To Do It. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.16.559
- Stephen M. Blackburn, Amer Diwan, Matthias Hauswirth, Peter F. Sweeney, José Nelson Amaral, Tim Brecht, Lubomr Bulej, Cliff Click, Lieven Eeckhout, Sebastian Fischmeister, Daniel Frampton, Laurie J. Hendren, Michael Hind, Antony L. Hosking, Richard E. Jones, Tomas Kalibera, Nathan Keynes, Nathaniel Nystrom, and Andreas Zeller. The Truth, the Whole Truth, and Nothing but the Truth: A Pragmatic Guide to Assessing Empirical Evaluations. TOPLAS, 2016
- Jan Vitek and Tomas Kalibera. Repeatability, Reproducibility, and Rigor in Systems Research. In Proceedings of the Ninth ACM International Conference on Embedded Software (EMSOFT). ACM, New York, NY, USA, 33-38. 2011. http://dx.doi.org/10.1145/2038642.2038650

The following work shows the pitfalls and proposes concrete approaches:

- T. Kalibera and R. Jones, "Rigorous Benchmarking in Reasonable Time". Proceedings of the 2013 ACM/SIGPLAN International Symposium on Memory Management (ISMM), 2013. http://dx.doi.org/10.1145/2555670.2464160
- Oliveira, A., S. Fischmeister, A. Diwan, M. Hauswirth, and P. Sweeney, "Why You Should Care About Quantile Regression", Proceedings of the Eighteenth International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), Houston, USA, March, 2013. http://dx.doi.org/10.1145/2451116.2451140
- Philip J. Fleming and John J. Wallace. 1986. How Not to Lie with Statistics: the Correct Way to Summarize Benchmark Results. Commun. ACM 29, 3 (March 1986), 218-221. http://dx.doi.org/10.1145/5666.5673
- Tim Harris: Do Not Believe Everything You Read in the Papers. https://timharris.uk/misc/2016-nicta.pdf

Replication and Reproduction Studies

The ability to confirm technical innovation reported by other researchers is an essential part of the scientific method.

- Michael A. Heroux. Editorial: ACM TOMS Replicated Computational Results Initiative http://dx.doi.org/10.1145/2743015
- James M. Willenbring . Replicated Computational Results (RCR) Report for “BLIS: A Framework for Rapidly Instantiating BLAS Functionality” http://dx.doi.org/10.1145/2738033

Places to store artifacts

- Dagstuhl DARTS an open repository for storing artifacts: https://www.dagstuhl.de/publikationen/darts/
- An open repository for storing large artifacts: https://zenodo.org
- Commercial repository for code:
    - Github.com
    - Bitbucket.com

Acknowledgements

This document has been created as a collaborative effort at the Dagstuhl Seminar 16111 “Rethinking Experimental Methods in Computing”. The current version of the document has been drafted by Sebastian Fischmeister (University of Waterloo), Richard Jones (University of Kent), and Luis Paquete (University of Coimbra) with input from the whole Dagstuhl Seminar Group: Umut Acar (Carnegie Mellon University), José Nelson Amaral (University of Alberta), David Bader (Georgia Institute of Technology), Judith Bishop (Microsoft Corporation), Ronald Boisvert (NIST), Marco Chiarandini (University of Southern Denmark), Markus Chimani (Universität Osnabrück), Emilio Coppa (Sapienza University of Rome), Daniel Delling (Apple Inc.), Camil Demetrescu (Sapienza University of Rome), Amer Diwan (Google), Dmitry Duplyakin (University of Colorado), Eric Eide (University of Utah), Erik Ernst (Google), Norbert Fuhr (Universität Duisburg-Essen), Paolo Giarrusso (Universität Tübingen), Andrew Goldberg (Amazon.com), Matthias Hagen (Bauhaus-Universität Weimar), Matthias Hauswirth (University of Lugano), Benjamin Hiller (Konrad-Zuse-Zentrum), Tomas Kalibera (Northeastern University), Marco Lübbecke (RWTH Aachen), Catherine McGeoch (Amherst College), Kurt Mehlhorn (MPI für Informatik), Eliot Moss (University of Massachusetts), Ian Munro (University of Waterloo), Petra Mutzel (TU Dortmund), Mauricio Resende (Amazon), Celso Carneiro Ribeiro (Fluminense Federal University), Peter Sanders (KIT --- Karlsruher Institut für Technologie), Nodari Sitchinava (University of Hawaii at Manoa), Peter Sweeney (IBM TJ Watson Research Center), Walter Tichy (KIT --- Karlsruher Institut für Technologie), Petr Tuma (Charles University), Dorothea Wagner (KIT --- Karlsruher Institut für Technologie), and Roger Wattenhofer (ETH Zurich).

Licensing