Experimental Results

Two important lessons

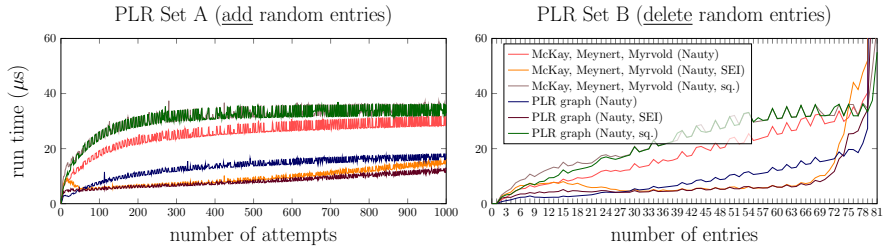
The Experimental Results section **often determines if a paper is accepted or rejected**.

- **We cannot write this section well, unless we design our experiments well.**
- **Do better than previous work.**
  - More data points.
  - Larger data sets.
  - More experiments.
  - Better presentation. High-quality figures; worthy of publication.

Image source `http://cdn.shopify.com/s/files/1/0329/0333/products/quality_assurance_GOOD_ENOUGH_IS_NOT_GOOD_ENOUGH_grande.jpg?v=1397950645`

Pgfplots example

Here's a plot from one of my papers using pgfplots:



- Many data points!
- We can use LaTeX commands in the figures.
- Compiles from data file.

Experimental Results

In the Experimental Results section, we answer:

- **Decisions**: What experiments do we perform? What are we testing? Why these experiments?
- **Setup**: What hardware, datasets, etc., do we use? Why these?
- **Baseline**: What do we compare to? Why this baseline?
- **Observations**: What do we observe from the experiments?
- **Deductions**: What do we deduce from the observations? (*Important*: separate "observations" from "deductions".)

(And maybe more, e.g., how observations relate to past work, etc.)

Important: **We improve on prior work** including in experiments.

- Use more data points; a larger data set; additional experiments; etc.
- Improve the presentation too.

Decisions

What experiments do we perform?

- Usually this is summarized at the beginning of the section, and the details of each experiment has its own subsection.

Why these experiments?

- We need to answer: *Why do we choose to perform these experiments?* And not other experiments. What are we testing? (Don't say "performance"; it's too vague.)
- Sometimes we follow the style in previous work; we point this out.

Why this way?

- We need to answer: *Why do we perform these experiments in this way?* (And not some other way.)

Setup

What hardware do we use?

- Usually CPU speed (and number of cores), memory, GPUs, etc.
- Often helpful to use a small table.
- Don't need to write ™ or ® (this is the company's responsibility).

What dataset(s) do we use?

- Describe the dataset and its statistics.

Other design decisions? E.g. maybe we set the "block size to 32".

Why did we choose these? Sometimes needed; sometimes obvious.

Baseline

We usually compare a proposed method with a state-of-the-art method.

What is the baseline?

- This should have been introduced in detail in the Related Work section.

Why do we (as impartial scientists) choose this baseline?

- We can easily manipulate the results by choosing an "easy" baseline. **Very common reason for a paper to be rejected.**

How has it been implemented?

- We can easily manipulate the results by poorly implementing the baseline. **Another common reason for a paper to be rejected.**
- What I've written in papers:
  - "We use the version of [baseline] downloaded from [URL]."
  - "All coding was performed by [the first author] who spent a comparable amount of time on optimization."
  - "Where possible, identical code was re-used in both implementations."

We recognize our human biases.
We describe our efforts to overcome them.

Observations and Deductions

Step 1: **What are we looking at?** We direct attention to a table or figure. (Where should the reader look? What is it?)

> *In Figure 1, we plot the speedup of [the proposed search engine] vs. [the baseline search engine] as [some variable] varies.*

Step 2: **Observation.** We identify the important points of the figure or table. (Which points should the reader look at?)

> *We observe that, as [some variable] increases, the speedup of [the proposed search engine] increases. In fact, when [something], we see a speedup of around [something].*

Step 3: **Deductions.** What do we conclude from these observations?

> *These observations indicate that [the proposed search engine] is most beneficial when [something], which is likely to be true for [large-scale search engines].*

Suits usual paragraph structure: topic sentence, deductions, conclusion. (Steps 1 and 2 are usually much longer than this.)

There are three steps: (1) What are we looking at? (2) Observation. (3) Deductions.

Steps 1 and 2 are factual; they're describing measurements. Step 3 is more opinionated. The reader may not agree with our *deductions*, but they should be able to understand how we came to our conclusions.

Disaster: I've seen students go from Step 1 to Step 3 within a single sentence:

> *As shown in Figure 1, [the proposed search engine] outperforms [large-scale search engines].*

Here "As shown in Figure 1, ..." simultaneously means both "displayed" and "demonstrated". (Then these students fill in the rest of the paragraph with "salesmanship", or something not needed in the Experimental Results.)

Long papers may have separate sections "Experimental Results" (Steps 1 and 2) and "Discussion" (Step 3).

Example

(1) What are we looking at? (2) Observation. (3) Deductions.

> The variance in effectiveness for TREC topics 826 – 850 across 100 topically partitioned distributed IR system instantiations ~~are~~ is illustrated in Figure 1. While effectiveness for some topics are consistent, others are clearly not. A shift in mean effectiveness due to outliers is also observed for several topics. Our proposed approach for significance testing can be used to eliminate such ambiguity and help researchers derive more accurate conclusions.
>
> — Jayasinghe et al., SIGIR, 2014.

What to write varies wildly.

Critique...

(1) What are we looking at? (2) Observation. (3) Deductions.

> Figure 1 shows the results of the microbenchmark experiment, looking at two different matrix sizes. We see that even on a single GPU, both SemCache++ and CUBLASXT are faster than the baseline because they overlap communication with computation. SemCache++ is faster than CUBLASXT because it is able to minimize communication. The A matrix is cached on both GPUs, as are the results of the DGEMMs. Hence, the DAXPY can be performed with no additional communication. In contrast, CUBLASXT, which does not leave the DGEMM results on the GPUs, must communicate the results of the DGEMMs back to the GPUs to perform the DAXPY.
>
> — Al-Saber and Kulkarni, PPoPP, 2015.

- Step 1 is clear: we know where to look, and what we will find there.
- Step 2 is succinct, but the observations end mid-sentence.
- Step 3 is missing (?). We make deductions from the observations.
- The part in black should be in the Proposed Solutions section.

Critique...

> Table 2 demonstrates the CPU and GPU runtime results for five key steps of the FMM algorithm, as well as the total runtime of a complete FMM iteration. Since the coefficient matrices associated with the P reconditioning and Direct passes have similar coefficient matrix dimensions that do not require coefficient matrix decompositions, clustering technique is only applied to the Downward and Evaluation passes. From Table 2, we observe 20X to 45X speedups for P reconditioning passes, 13X to 45X speedups for Direct passes, 1.2X to 4X speedups for Upward passes (non-critical kernel), 15X to 33X speedups for Downward passes, 20X to 30X speedups for Evaluation passes, and 18X to 30X speedups for the overall FMM iteration.
>
> — Zhao and Feng, DAC, 2011.

Completely out of order!!!

- We should begin with something like Table 2 tabulates [something]. When then give the extra details.
- Then observations. Then deductions.
- *Illogical*: Tables cannot demonstrate things!

Ceci n'est pas une pipe (This is not a pipe)

— The Treachery of Images
https://en.wikipedia.org/wiki/The_Treachery_of_Images

The famous pipe. How people reproached me for it! **And yet, could you stuff my pipe?** *No, it's just a representation, is it not? So if I had written on my picture 'This is a pipe', I'd have been lying!*

—René Magritte

Be careful with: "increases exponentially"

People like to say data points "increase exponentially"—it sounds slick. However, this term has a mathematical meaning. (It's asymptotically $\Theta(2^n)$.)

There are functions like 2^n and e^n which increase exponentially. If our data follows such a trend, it's okay to say it has **exponential growth** and it **increases exponentially**. However....

There are functions which do not increase exponentially.

- If our data is bounded above by say n^{100}, then it increases too slowly to be exponential. This is **sub-exponential growth**.
- If our data has approximately the same growth as $n!$, then it increases too rapidly to be exponential. This is **super-exponential growth**.

Also, if our function does not make sense for $n \geq$ [some constant], then we should not use asymptotic terminology to describe it.

Averaging

Sometimes, we present results that are **averaged over repeated experiments**.

The reason we repeat experiments is to **estimate the variance in measurements**. I.e., how much the measurement changes from experiment to experiment.

If we write the average of some measurement, we need to also write the variance.

"Various" vs. "different"

Table 1. The average execution time of PDE with different number of processors and the corresponding speedup compared with the sequential execution.

Number of Cores n	Time (second)	Speedup
1	1002.01	1

— Liu et al., GECCO, 2016.

"Different" and "various" are often not synonyms:

- Problem 1:
  - *Correct*: "Dogs are **different** to cats."
  - *Incorrect*: "Dogs are **various** to cats."
- Problem 2: Different meanings:
  - *Example 1*: "We live in **different** houses."
  - *Example 2*: "We live in **various** houses."

Example 1 means: We don't live together.
Example 2 means: We live in multiple houses.

"Various" vs. "different" (cont.)

This is technically correct:

> Table 1. The average execution time of PDE **with different numbers of processors** and the corresponding speedup compared with ~~the~~ sequential execution.

but it's hard to understand. (It's natural to think "different numbers of processors to what?")

This is better written:

> Table 1. The average execution time of PDE **as the number of processors varies** and the corresponding speedup compared with ~~the~~ sequential execution.

Variable; various; varying; and varies. They have the same root.

Mental link: "the number of processors" is the "variable;" (because it varies).

Theoretical analysis

Theoretical analyses

Instead of experiments, some papers give a theoretical analysis. This is typical in cryptography, coding, and theoretical computer science. And even these papers have experiments.

I can't say much without giving a course on mathematical writing.

So I'll mention a few things I see wrong repeatedly...

Big-O notation

Big-O notation $O(\cdots)$ is **asymptotic notation**, this means some variable must be going to infinity.

- *Reasonable assumption*: e.g., if the variable is the number of docIDs in an inverted list, and an inverted list might contain 100000+ docIDs.
- *Poor assumption*: e.g., if the variable is the number of blocks in a code stripe, which is probably around 10 or 20. **Do not use big-O notation**: the variable is not going to infinity.

Big-O notation is generally used in two ways:

1. We write $f(n) = O(n^2)$ when the function $f = f(n)$ as bounded above asymptotically by the function $n \mapsto n^2$.
2. We write $f(n) = n^2 + O(n)$ when we can write $f(n) = n^2 + g(n)$ for some function g satisfying $g(n) = O(n)$.

Big-O notation (cont.)

The purpose of using big-O notation in computer science is the massive simplification it offers:

- $O(10 \log_2 n + 10^6)$ is the same as $O(\log n)$.
- $O(6.124n^2 + 0.224n)$ is the same as $O(n^2)$.
- $O(nm)$ is ?!?! In multivariable big-O notation, we need to identify which variables are going to infinity.

Using big-O notation, we can asymptotically compare two algorithms, filtering out minor optimizations and implementation details.

If simplifying like this does not make sense, then don't use big-O notation.

Other problems

- The percentage symbol % (used in code) should not by used to mean a mod b.
- We should not write $1/ab$, as it could be interpreted as either $1/(ab)$ or $(1/a)b = \frac{1}{a}b$. (Instead, we write $1/(ab) = \frac{1}{ab}$ or $\frac{1}{a}b = \frac{b}{a} = b/a$.)
- The symbols "$<$" and "$>$" mean "less than" and "greater than". Brackets are typeset `$\langle 1,2,...,n \rangle$`, which displays as $\langle 1, 2, \ldots, n \rangle$.

That's all the new material!!