Writing for

Computer Science & Engineering

Ki-Il Kim

Department of Computer Science and Engineering

CNU

Experimentation

Introduction

- The use of experiments to verify hypothesis is one of the central elements of science
  - Implementation tried against test data
  - Confirming hypotheses about algorithms and systems
  - An experiment can verify that a system can complete a specified task and can do so with reasonable use of resources
  - A tested hypothesis becomes part of scientific knowledge if it is sufficiently well described and constructed, and if it is convincingly demonstrated
- Some people disagree with the view that rigorous experiments are essential in computer science or may hold a low opinion of papers that have no new theory and are "merely" experimental
  - Such view are in stark contract to the role of experiments in other disciplines
  - Experiments are an essential part of sound science

Introduction

- Experiments in computing
  - Take diverse forms, from tests of algorithms performance to human factors analysis
  - Same principles underlying good experimentation : test should be fair rather than constructed to support the hypothesis
  - If the design of tests seems biased towards the intended contribution, readers will not be persuaded by the results
- The topic of this chapter is the design, execution, and description of experiments in computing

Baselines

- ❖ A first step in the design of experiments
  - Identify the benchmarks against which your contribution will be measured
  - Identify an appropriate baseline
    - ➢ No sensible research would advocate that their new sorting algorithm was a breakthrough on the basis that it is faster than bubblesort. Instead, the algorithm should be compared to the best previous method
- ❖ Baseline
  - Compelling if it is implemented to a high standard, and it may be that comparison to a baseline is difficult because an implementation of a competing method must be obtained.
  - However, without such a comparison it may be impossible for the reader to know whether the new method offers an improvement.
  - This is barrier to entry : there is a barrier to entry does not excuse poor science

Baselines

- A danger in an ongoing research program is to fail to update the choice of baseline
- Example
  - Text indexing in work in the 1980s on signature file performance was compared to that of inverted files as reported in papers form the 1970s
  - Papers on signature files even in the 2000s continued to quote these baselines, despite dramatic improvements in inverted file
  - New work in signature files was compared to previous work in the same area, but not to relevant work on other pertinent technology
- A similar problem can arise when a well-known, widely available implementation becomes commonly used as a reference point.
  - Benefit : use of the common resource means that readers can have confidence that the baseline is accurate
  - Suffer : the advances that are being described may not be cumulative

Baselines

- It is critical that baseline should be identified early in the research program
  - What is the point of developing new method if existing method provide a satisfactory solution

Persuasive Data

❖ For work that involves experiments, it is critical that you have access to appropriate data, and that you understand it well. In general, you need to consider

- What data my be available, and whether it is created by you or sourced from elsewhere
- What specific mechanisms will be used to gather and standardized the data
- Whether the data will be sufficient in volume or quality to give a robust answer to the question
- What domain knowledge may be required to properly interpret the data
- What the limits, biases, flaws, and properties of the data are likely to be, and how these problems will be addressed or managed
- What the results will be like if the data supports the hypothesis; or, alternatively, what they will be like if the hypothesis is false

Persuasive Data

- Some experiments depend on human annotation of data, to provide a gold standard or ground truth
  - For example, in document classification, human annotation many indicates the topic of document: politics, entertainment, sport, and so on
  - Gathering of annotations can be the dominant cost of an experiment
- In the process of developing new algorithms, researchers typically use as a testbed a data set with which they are familiar
  - If the algorithm is parameterized in some way; this testbed can be used for tuning to identify the parameter values that give the best performance
  - If parameters have been derived by tuning, the only way to establish their validity is to see if they give good behavior on other data
- Choosing parameters to suit data, or choosing data to suit parameters invalidates the research

Persuasive Data

- The research in some fields is underpinned by the availability and use of reference data sets
- An underlying point is that persuasive research requires appropriate data, and thus you need to be confident that you can obtain good data before committing to a particular research question
- Ask whether a single data set is sufficient, or whether multiple data sets are required. Multiple data set are sufficiently independent

Persuasive Data

- Sometimes appropriate data can be artificial, or simulated. Such data can allow a thorough exploration of the properties of an algorithm
- Another question is estimation of the volume of data required
  - To what volume of data should your claims apply?
  - If you are making claims about terabytes, but testing on megabytes, you are asking the reader to believe that your results can be extrapolated a million-fold.
  - For example, if you are comparing two parsers according to their ability to accurately extract phrases from English text, it may be that statistical principles will tell you that a collection of 100 examples is unlikely to be sufficient for the anticipated improvement to be detected.

Interpretation

- When checking experimental design or outcomes, consider whether there are other possible interpretations of the results
- Consider for instance the problem of finding whether a file stored on disk contains a given string
  - Directly scans the files
  - Scans a compressed form of the file
  - Further tests would be needed to identify whether the speed gain was because the second algorithm used fewer machine cycles or because the compressed file was fetched more quickly from disk
- Care is particularly needed when checking the outcome of negative or failed experiments
  - The failure of an experiment typically leads to it being redesigned
- It is always worth considering whether the results obtained are sensible – boundary condition

Interpretation

- Conclusions should be sufficiently supported by the results. Success in a special case does not prove success in general, so be aware of factors in the test that may make it special
- Don’ t draw undue conclusions or interferences.
  - If one method is faster than another on a large data set, and they are of the same speed on a medium data set, that does not imply that the second is faster on a small data set -> different costs dominate at different scale
- Don’ t overstate your conclusions
  - If a new algorithm is somewhat worse than an existing one, it is wrong to describe them as equivalent
  - A reader might infer that they are equivalent if the different is small, but it is not honest for you to make that claim

Interpretation

- Another aspect of interpretation is that numerical measures allow numerical manipulation, but such manipulation does not always make sense if applied to the qualitative goal we wish to achieve
  - One system may achieve a 20% higher score than another under some measure of user satisfaction, but it makes little sense to say that the user is 20% more satisfied

Interpretation

❖ Predictivity

- The main reason that we experiment and measure is to provide evidence about the behavior of a system in general
- We use measurements on the data we have to hand to make predictive claims about what will happen in the future, when the same system is applied to new data; the conclusions in our papers are usually about properties of systems, not their behaviors on the data we have already seen

Robustness

- Experiments should be as far as possible be independent of the accuracy of measurements or quality of the implementation
  - Ideally an experiment should be designed to yield a result that is unambiguously either true or false
  - This is not possible, another form of confirmation is to demonstrate a trend or pattern of behavior
- Results may includes some anomalies or peculiarities. These should be explained or at least discussed. Don’ t discard anomalies unless you are certain they are irrelevant; they may represent problems you haven’ t consider
  - As the graph shows, the algorithm was much slower on two of the data set. We are still investigate this behavior

Robustness

- A common failing in experimental work is that complex processes are tested as a whole, but not as components.
  - Many proposed methods are pipelines or composites of one kind or another, in which independent elements are combined to give a result
  - Search engine : crawler, parser, query engine
  - If a research proposed a new engine comprised of entirely new components, but only tested it as a whole, the reader would not learn to what extent each component was valuable
- Similarly, an experimental regime should include separate investigation of each relevant variable – the reader needs to know what factors are influencing the outcomes
  - If increasing a value has an effect, what happens with a further increase?
  - You should explore factors broadly enough to make trends and patters clear

Performance of Algorithms

- Potential approaches to assessment of the performance of algorithms
  - Proof, mathematical modelling, simulation and experimentation
- How these are used flows from questions about what the assessment of the algorithm is intended to achieve, and what the likely limitations of and constraints on the experiments
  - Basics of evaluation : The basis of evaluation should be made explicit. For comparison, specify not only the environment but also the criteria used for comparison. A comparison should have a realistic basis. Simplifying assumptions can be used to make mathematical analysis traceable, but can give unrealistic models.
  - Processing time : Time over some given input is one of the principal resources used by algorithms; others are memory, disk space, and disk and network traffic. Measurements of CPU time can be unreliable.
  - Memory and disk requirements : It is often possible to trade memory requirements against time, not only choice of algorithm but also by changing the way disk is used and memory is accessed
  - Disk and network traffic : Disk costs have two components, the time to fetch the first bit of requested data and the time required to transmit the requested data. Because of the sophistication of current disk and the complexity of their interaction with CPU and operating system, exact mathematical descriptions of algorithm behavior are unattainable

Coding for Experimentation

- In computer science research, in principle at least, the sole reason for coding is to build tools and probes for generating, observing, or measuring phenomena. The choice of what to measure guides the process of coding and implementation
- The basic rule is to keep things simple
  - If efficiency is not being measured, for example, don’ t waste time squeezing cycles from code
  - If a database join algorithm is being measured, it may not necessary to implement indexes, and it is almost certainly unnecessary to write an SQL interpreter

Coding for Experimentation

❖ Coding for an experiment, there are several other such rules or guideline that might seem obvious

- One task, one tool: decompose the problem into separate pieces of code. In most cases, trying to create a single piece of code that does everything is just not productive
- Be aware that you many need to trade ease of implementation against realism of the result
- Cut the right corners. Coding for a day to save an hour’ s manual work is a waste of time, even if coding is the more principled approach
- Use the right tool, not the most conventional tool
- Don’ re-code unnecessarily. Use libraries
- Find an independent way of verifying that the output is correct
- For long-running processes, consider developing mechanism for periodically saving state, so that weeks of work isn’ t lost.

Coding for Experimentation

- In environments such as the Unix family of operating system
  - A program is often tested by being run from the command-line with output directed to the screen
  - Parameters may be passed in as arguments, but to simplify coding they may be defined as constants within programs
- A more reliable, repeatable approach is to run all experiments from scripts.
  - Parameter settings are captured within the script
  - The settings used last time can be commented out
  - Output from the script can be directed to a logfile and kept indefinitely
  - If the output is well designed, it should include information such as input file names, code versions, parameter values, and date and time
- Using simple Unix tools, it is straightforward to take data directly form a log file and produces a graph or other summary of the results

Coding for Experimentation

- A practical consideration is whether the experiments are feasible at all
  - Experimentation can require
    - storage of large volumes of data
    - implementation of production-quality code
    - execution over months
    - with repetitions after failure
    - access to particular machines or configurations
    - use of humans for evaluation of results
    - access to restricted data set
    - use of particular pieces of software, or most of these things at the same time
- Before proceeding too far with a research question, you need to be confident that you will have the resource required to undertake the experiments that are needed for a persuasive outcome

Describing Experiments

- Your interpretation and understanding of the results can be as important as the results themselves
  - When describing the outcomes of an experiment, don’t just compile dry lists of figures or a sequence of graphs
  - Analyze the results and explain their significance, select typical results and explain why they are typical, theorize about anomalies, show why the results confirm or disprove the hypotheses, and make the results interesting
- Experiments are only valuable if they are carefully described. The description should reflect the care taken of the hypothesis
- Results are valueless if they are some kind of singleton event: repetition of the experiment should yield the same outcomes. And result are equally valueless if they cannot be repeated by other researchers

Describing Experiments

- Researchers must decide which results to report
  - Reported results should be a fair reflection of the experiment’s outcomes
- If a test fails on some data sets and succeeds on others, it is unethical to conceal the failure, and the existence of failure should be stated as prominently as that of success. Likewise, reporting just one success might lead the reader to wonder whether it was no more than a fluke
- The experimental outcomes reported in a paper may represent only a fraction of the work that was undertaken in a research program
- Another strategy that keeps researchers honest, and helps to describe and publicize their work, is to make code and data available online

Describing Experiments

❖ Part of reporting of experiments is description of the data that was used

- Typically, readers need to know how the data was gathered and created
- How you version of the data might be obtained or recreated
- What the shortcomings of the data are, that is, in what ways it might be uncertain, incomplete, or unreliable
- What aspects of the research question are not tested by the data

Writing for

Computer Science & Engineering

Ki-Il Kim

Department of Computer Science and Engineering

CNU

Statistical Principles

Introduction

- We use experiments and take observations to study the behavior of a system, to test hypotheses, to investigate the effect of manipulations and optimizations, and, overall, to produce evidence for our argument
  - The elementary material of evidence is measurement: the reduction of complex phenomena to numerical scores than can be recorded, compared, and analyzed
- Raw numbers, however, are dangerously deceptive in their apparent certainty
  - If we find in an experiment that our system has higher score than that of a competitor, we are easily convinced that our system is superior
    - The measurement itself may be inaccurate or misleading. It many be only an approximation to the real-world quality that we are attempting to measure
    - Even if the measurement is appropriate, the value it provides in a single test may be subject to variability and randomness – in the choice of experimental inputs, in the conditions in which the experiment is run or in human assessment of the outcomes

Introduction

- To gain trust in experiments, we need to repeat them, giving sets of results to which we can apply statistical methods.
  - Having multiple experiment results not only provide aggregated, stable measurements, but let us use the tools of statistical inference to determine how confident we should be in our conclusions
- Repeated experiments also provide insights into system behavior through allowing us to observe variability, success, and failure in a systematic way
  - All researchers should be aware of relevant statistical principles, and be able to judge when use of statistic is necessary for their works

Variables

- The ideal experiment examines the effect of one variable on the behavior of an object being studied
  - How does increasing the volume of data affect execution time?
  - Can the vision system track rapidly moving objects?
  - How much compression can be achieved without visibly degrading the image
  - If no other variables are present, it is easy to be confident that the variable does indeed affect the behavior in the way observed
  - The test environment should be designed to minimize the effect of extraneous factors

Samples and Populations

- Taking a measure of an event gives us a numerical score quantifying that event
  - The numerical score has a sense of exactness to it, but exactness is often misleading
  - We run the experiment again, a different score might be produced, due to factors such as change in the take, change in the input to the same task, variability in the experimental setup that is beyond our control
- The variability may arise naturally in the system or environments. In either case, the existing of variability demonstrates that a single test is not enough, and that the variability should be reported and analyzed along with the overall outcomes
- In computer science research, many people view statistics are no more than reporting of average and deviations.

Samples and Populations

- “Algorithm NEW is typically faster than algorithm OLD”
  - Claim : NEW is faster on average
  - But, an average of WHAT?
  - If the interned meaning is only that NEW is faster than OLD on average for the runs undertaken in the experiments, what is it about these runs that make them representative or predictive?
- Population : the set of all possible runs
  - If NEW is indeed faster than OLD on average across the whole population, the claim is a reasonable one, but in all likelihood the population is infinite, as it must contain all combinations of input data
  - It is necessary to resort to taking a sample; to do so, it is necessary to understand what the population consists of
  - In medicine, the population could be all people, or perhaps all sick people, or possibly all people for whom other medications have failed.

Samples and Populations

- A straightforward case of varying inputs is whether this is due to randomization
  - Classifier that achieve a certain effectiveness score on a collection of 100,000 items, with 10,000 of items randomly selected for training and 90,000 for the test -> effectiveness will differ for different random partitionings of the items into training and test sets
  - The correct approach is to perform multiple randomizations, to yield a distribution of outcomes that can be analyzed statistically.

Aggregation and Variability

- A simple way to aggregate the multiple scores achieved by running an experiment multiple times is to report an average of the outcomes
  - Average is reported as the arithmetic mean
  - The average score is more stable and representative of the experiment than any single score
  - It is better predictor of what score would be achieved if the experiment were run again
- The importance of randomness in the selection or generation of inputs to tests is something that is often missed by novice researchers

Aggregation and Variability

- An average is a reasonable estimate of typical behavior depends on the kind of event being measured.
  - Typical network delay for a round-trip of a packet, average may well be meaningless
  - Some delays are effectively infinite. The distribution of such delays often consists of a large number of fast response and a small number of extremely slow response; the average is therefore somewhat slower than the fast times, but in a range where no value were observed at all
- A consequence of this reasoning is that there are cases where the maximum or the minimum may be the best value to report
  - The time taken for a distributed system to process a problem may vary depending on a range of variables, all of which have the effect of interfering with the system
  - It may be appropriate to report the fastest time observed, while noting the variance

Reporting of Variability

- Averaging provides valuable insight into typical behavior, but it is often also appropriate to report variability
  - Cases the average included wild outliner
- A commonly reported descriptive statistic is standard deviation
  - It qualitied variability in a single value, which are in the same units as the mean, and that is a key input to statistical inference
- Average scores should only be reported with a precision that corresponds to the accuracy of the average
  - If only a few instance of a highly variable phenomenon are observed, then reporting many decimal places gives a false impression of exactness
  - "The average running time of our algorithm is 1.161s" makes the result seem much more precise than it really is

Randomness and Error

- In some circumstances, it is possible for an experiment to succeed, or at least appear to succeed, by lock; there might be an atypical pattern to the data, or variations in system response might favour one run over another
  - Particularly true for timing, which can be affected by other users, system overhead, inability of most operating system to accurately allocate clock cycles to process, and cache
- Comparing the speed of two algorithms (NEW and OLD) for the same task
  - Take the same input and produce the same output – NEW is faster than OLD by several percent
- Possible to conclude
  - NEW is better implemented than was OLD. Perhaps the same care was not taken with OLD
  - OLD uses more buffer space than NEW, leading to poor behavior on this particular computer
  - OLD uses floating-point operations that are not supported in hardware
  - At compile-time, OLD was accidentally buit with debug option enable, slowing it down
  - Inaccuracies in the timing mechanism randomly flavored NEW.
  - OLD was run first, and was delayed while the input was copied to memory; NEW accessed the input directly form cache
  - The particular input chosen happened to favor NEW

Visualization of Results

- We use computers to produce results, and can also use computers to help to digest them
  - Apply statistics and use visualization
- Visualization of data is substantial field in its own right, with a wide range of established techniques and principles
- Elementary approaches to re-interpretation of data via graphs can yield valuable insights
- Graphs can also be used to interpret data from a variety of perspectives

Visualization of Results

✗

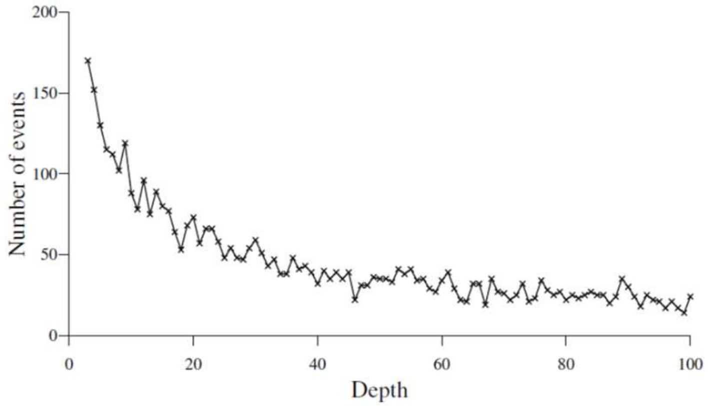


FIGURE 7. *The number of events observed at each depth; depths 1 and 2 have been omitted for reasons of scale.*

✓

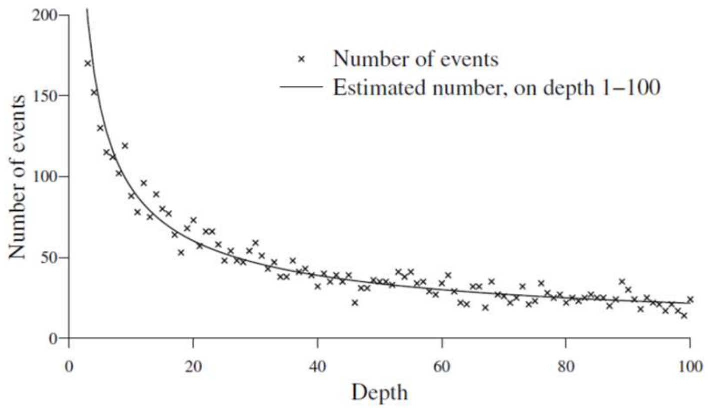
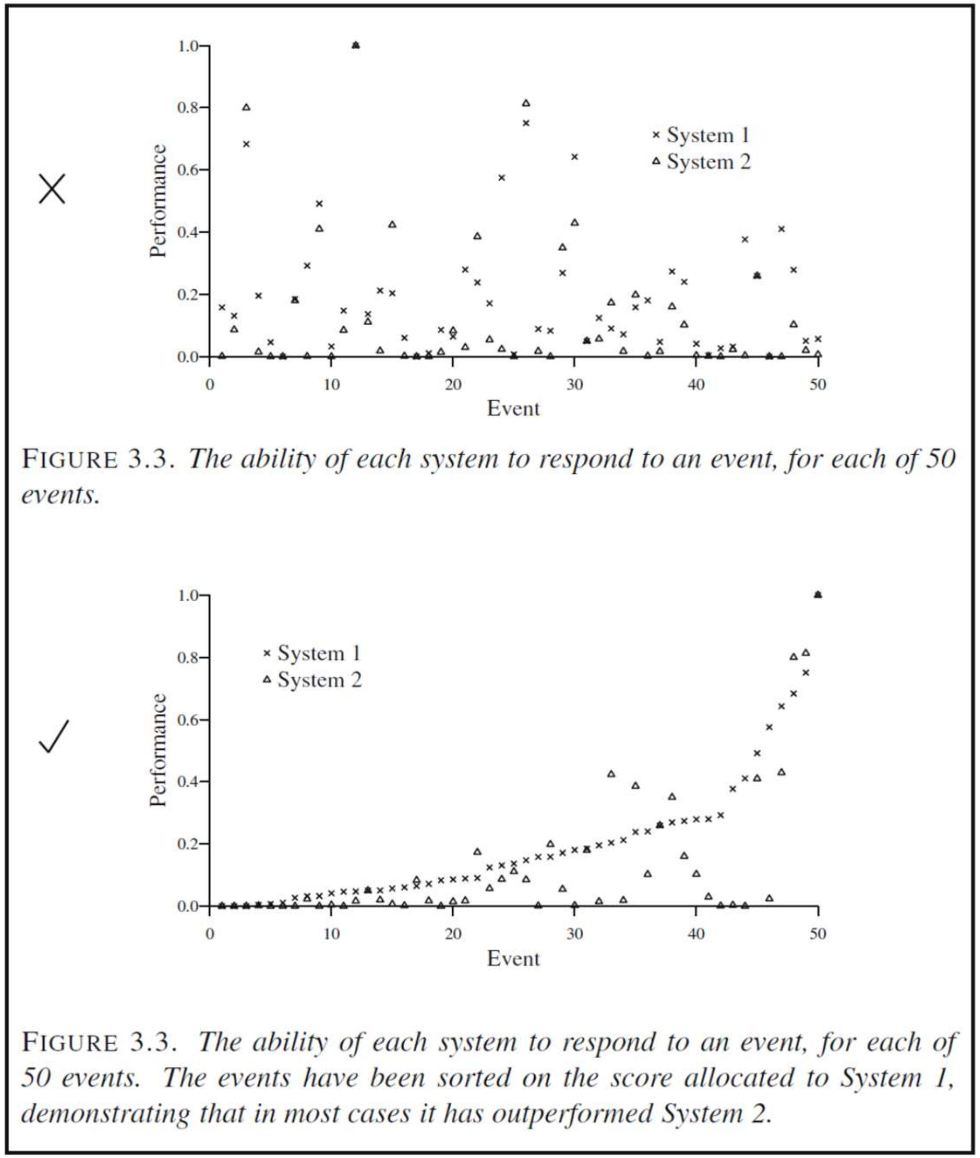


FIGURE 7. *The number of events observed at each depth; depths 1 and 2 have been omitted for reasons of scale. The solid line shows a best-fit to the points.*

Visualization of Results



✗

FIGURE 3.3. *The ability of each system to respond to an event, for each of 50 events.*

✓

FIGURE 3.3. *The ability of each system to respond to an event, for each of 50 events. The events have been sorted on the score allocated to System 1, demonstrating that in most cases it has outperformed System 2.*

Visualization of Results

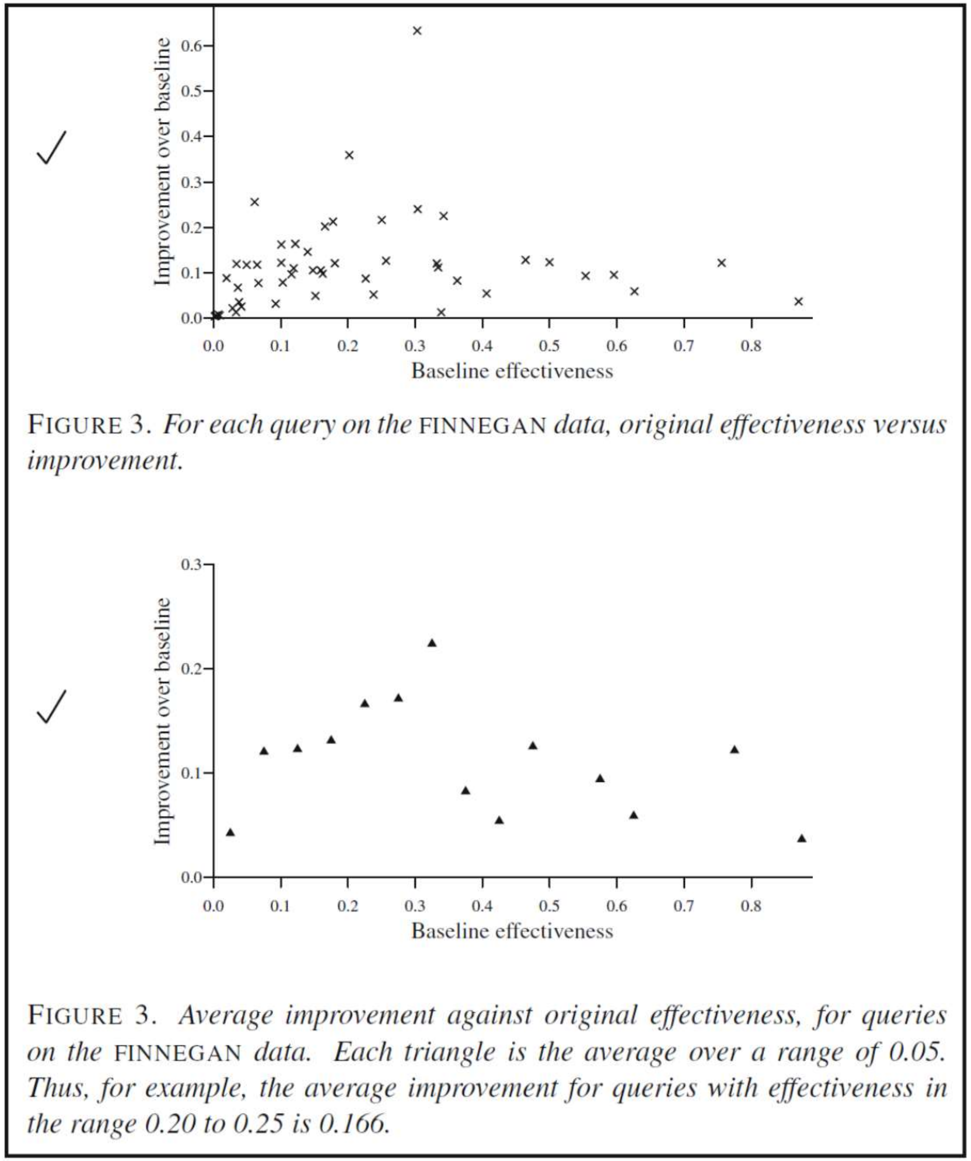


FIGURE 3. *For each query on the* FINNEGAN *data, original effectiveness versus improvement.*

FIGURE 3. *Average improvement against original effectiveness, for queries on the* FINNEGAN *data. Each triangle is the average over a range of 0.05. Thus, for example, the average improvement for queries with effectiveness in the range 0.20 to 0.25 is 0.166.*

Q
&
A