Chapter 4
Hypotheses, Questions, and Evidence

The intensity of the conviction that a hypothesis is true has no bearing on whether it is true or not.

P.B. Medawar
Advice to a Young Scientist

The great tragedy of Science, the slaying of a beautiful hypothesis by an ugly fact.

T.H. Huxley
Biogenesis and Abiogenesis

An argument is a connected series of statements intended to establish a proposition … Argument is an intellectual process. Contradiction is just the automatic gainsaying of anything the other person says.

Monty Python
The Argument Sketch

The first stages of a research program involve choice of interesting topics or problems, and then identification of particular issues to investigate. The research is given direction by development of specific questions that the program aims to answer. These questions are based on an understanding—an informal model, perhaps—of how something works, or interacts, or behaves. They establish a framework for making observations about the object being studied. This framework can be characterised as a statement of belief about how the object behaves—in other words, a *hypothesis*.

Many hypotheses concern some aspect of the physical world: whether something is occurring, whether it is possible to alter something in a predictable way, or whether a model is able to accurately predict new events. Astronomers use nuclear physics to predict the brightness of stars from their mass and chemical composition, for example, while a geneticist may seek to know whether substituting one gene for another can improve the health of a cell.

In computer science, some hypotheses are of this kind. We examine the limits of speech recognition, ask whether Web search can be used effectively by children, or predict how well a service will respond to increasing load. Other hypotheses are constructive. For example, we propose new technologies and explore their limitations and feasibility, or propose theorems that imply that there may be new solutions to long-standing algorithmic problems. Regardless of field, if you wish to achieve

J. Zobel, *Writing for Computer Science*, DOI 10.1007/978-1-4471-6639-9_4

robust research outcomes it is essential to have a hypothesis. This chapter concerns hypotheses and research questions, and how we use evidence to confirm or disprove them.

Hypotheses

In outline, an example research program might proceed as follows.

- A researcher investigating algorithms might speculate as to whether it is possible to make better use of the cache on a CPU to reduce computational costs.
- Preliminary investigation might lead to the *hypothesis* that a tree-based structure with poor memory locality will be slower in practice than an array-based structure with high locality, despite the additional computational cost.
- The hypothesis suggests the *research question* of whether a particular sorting algorithm can be improved by replacing the tree structure with the array structure.
- The *phenomenon* that should be observed if the hypothesis is correct is a trend: for example, as the number of items to be sorted is increased, the tree-based method should increasingly show a high rate of cache misses compared to the array-based method.
- The *evidence* is the number of cache misses for several sets of items to be sorted. Alternatively, external evidence might be used, such as changes in execution time as the volume of data changes.

As this example illustrates, the structure of the research program flows from having a definite research question and hypothesis.

A hypothesis or research question should be specific and precise, and should be unambiguous; the more loosely a concept is defined, the more easily it will satisfy many needs simultaneously, even when these needs are contradictory. And it is important to state what is *not* being proposed—what the limits on the conclusions will be. Consider an example. Suppose P-lists are a well-known data structure used for a range of applications, in particular as an in-memory search structure that is fast and compact. A scientist has developed a new data structure called the Q-list. Formal analysis has shown the two structures to have the same asymptotic complexity in both space and time, but the scientist intuitively believes the Q-list to be superior in practice and has decided to demonstrate this by experiment.

This motivation by belief, or instinct, is a crucial element of the process of science: since ideas cannot be known to be correct when they are first conceived, it is intuition or plausibility that suggests them as worthy of consideration. That is, the investigation may well have been undertaken for subjective reasons; but the final report on the research—that is, the published paper—must be objective.

Continuing the example above, the hypothesis might be encapsulated as

✗ Q-lists are superior to P-lists.

But this statement is not sufficient as the basis of an experiment: success would have to apply in all applications, in all conditions, for all time. Formal analysis might be

able to justify such a result, but no experiment will be so far-reaching. In any case, it is rare for a data structure to be completely superseded—consider the durability of arrays and linked lists—so in all probability this hypothesis is incorrect. A testable hypothesis might be

✓ As an in-memory search structure for large data sets, Q-lists are faster and more compact than P-lists.

Further qualification may well be necessary.

✓ We assume there is a skew access pattern, that is, that the majority of accesses will be to a small proportion of the data.

The qualifying statement imposes a scope on the claims made on behalf of Q-lists. A reader of the hypothesis has enough information to reasonably conclude that Q-lists do not suit a certain application; this limitation does not invalidate the result, but instead strengthens it, by making it more precise. Another scientist would be free to explore the behaviour of Q-lists under another set of conditions, in which they might be inferior to P-lists, but again the original hypothesis remains valid.

As the example illustrates, a hypothesis must be testable. One aspect of testability is that the scope be limited to a domain that can feasibly be explored. Another, crucial aspect is that the hypothesis should be capable of falsification. Vague claims are unlikely to meet this criterion.

✗ Q-list performance is comparable to P-list performance.

✗ Our proposed query language is relatively easy to learn.

The exercise of refining and clarifying a hypothesis may expose that it is not worth pursuing. For example, if complex restrictions must be imposed to make the hypothesis work, or if it is necessary to assume that problems that are currently insoluble must be addressed before the work can be used, how interesting is the research?

A form of research where poor hypotheses seem particularly common is "black box" work, where the black box is an algorithm whose properties are poorly understood. For example, some research consists of applying a black-box learning algorithm to new data, with the outcome that the results are an improvement on a baseline method. (Often, the claim is to the effect that "our black box is significantly better than random".) The apparent ability of these black boxes to solve problems without creative input from a scientist attracts research of low value. A weakness of such research is that it provides no insights into the data or the black box, and has no implications for other investigations. In particular, such results rarely tell us whether the same behaviour would occur if the same approach were applied to a different situation, or even to a new but similar data set.

That is, the results are not *predictive*. There may be cases in which it is interesting to observe the behaviour of an algorithm on some data, but in general the point of experimentation is to confirm models or theories, which can then be used to predict future behaviour. That is, we use experiments to learn about more general properties, a characteristic that is missing from black-box research.

A related problem is the re-naming fallacy, often observed in the work of scientists who are attempting to reposition their research within a fashionable area. Calling a network cache a "local storage agent" doesn't change its behaviour, and if the term "agent" can legitimately be applied to any executable process then the term's explanatory power is slim—a particular piece of research is not made innovative merely by changing the terminology. Likewise, a paper on natural language processing for "Web documents" should presumably concern some issues specific to the Web, not just any text; a debatable applicability to the Web does not add to the contribution. And it seems unlikely that a text indexing algorithm is made "intelligent" by improvements to the parsing. Renaming existing research to place it in another field is bad science.

It may be necessary to refine a hypothesis after initial testing; indeed, much of scientific progress can be viewed as refinement and development of hypotheses to fit new observations. Occasionally there is no room for refinement, a classic example being Einstein's prediction of the deflection of light by massive bodies—a hypothesis much exposed to disproof, since it was believed that significant deviation from the predicted value would invalidate the theory of general relativity. But more typically a hypothesis evolves in tandem with refinements in the experiments.

However, the hypothesis should not follow the experiments. A hypothesis will often be based on observations, but can only be regarded as confirmed if it is able to make successful predictions. There is a vast difference between an observation such as "the algorithm worked on our data" and a tested hypothesis such as "the algorithm was predicted to work on any data of this class, and this prediction has been confirmed on our data". Another perspective on this issue is that, as far as possible, tests should be blind. If an experiment and hypothesis have been fine-tuned on the data, it cannot be said that the experiment provides confirmation. At best the experiment has provided observations on which the hypothesis is based. In other words: first hypothesize, then test.

Where two hypotheses fit the observations equally well and one is clearly simpler than the other, the simpler should be chosen. This principle, known as Occam's razor, is purely a convenience; but it is well-established and there is no reason to choose a complex explanation when another is available.

Defending Hypotheses

One component of a strong paper is a precise, interesting hypothesis. Another component is the testing of the hypothesis and the presentation of the supporting evidence. As part of the research process you need to test your hypothesis and if it is correct—or, at least, not falsified—assemble supporting evidence. In presenting the hypothesis, you need to construct an argument relating your hypothesis to the evidence.

For example, the hypothesis "the new range searching method is faster than previous methods" might be supported by the evidence "range search amongst n elements requires $2 \log_2 \log_2 n + c$ comparisons". This may or may not be good evidence, but it is not convincing because there is no argument connecting the evidence to the

hypothesis. What is missing is information such as "results for previous methods indicated an asymptotic cost of $\Theta(\log n)$". It is the role of the connecting argument to show that the evidence does indeed support the hypothesis, and to show that conclusions have been drawn correctly.

In constructing an argument, it can be helpful to imagine yourself defending your hypothesis to a colleague, so that you play the role of inquisitor. That is, raising objections and defending yourself against them is a way of gathering the material that is needed to convince the reader that your argument is correct. Starting from the hypothesis that "the new string hashing algorithm is fast because it doesn't use multiplication or division" you might debate as follows:

- I don't see why multiplication and division are a problem.

  On most machines they use several cycles, or may not be implemented in hardware at all. The new algorithm instead uses two exclusive-or operations per character and a modulo in the final step. I agree that for pipelined machines with floating-point accelerators the difference may not be great.
- Modulo isn't always in hardware either.

  True, but it is only required once.
- So there is also an array lookup? That can be slow.

  Not if the array is cache-resident.
- What happens if the hash table size is not 2^8?

  Good point. This function is most effective for tables of size 2^8, 2^{16}, and so on.

In an argument you need to rebut likely objections while conceding points that can't be rebutted, while also admitting when you are uncertain. If, in the process of developing your hypothesis, you raised an objection but reasoned it away, it can be valuable to include the reasoning in the paper. Doing so allows the reader to follow your train of thought, and greatly helps the reader who independently raises the same objection. That is, you need to anticipate concerns the reader may have about your hypothesis. Likewise, you should actively search for counter-examples.

If you think of an objection that you cannot refute, don't just put it aside. At the very least you should raise it yourself in the paper, but it may well mean that you must reconsider your results.

A hypothesis can be tested in a preliminary way by considering its effect, that is, by examining whether there is a simple argument for keeping or discarding it. For example, are there any improbable consequences if the hypothesis is true? If so, there is a good chance that the hypothesis is wrong. For a hypothesis that displaces or contradicts some currently held belief, is the contradiction such that the belief can only have been held out of stupidity? Again, the hypothesis is probably wrong. Does the hypothesis cover all of the observations explained by the current belief? If not, the hypothesis is probably uninteresting.

Always consider the possibility that your hypothesis is wrong. It is often the case that a correct hypothesis at times seems dubious—perhaps in the early stages, before it is fully developed, or when it appears to be contradicted by initial experimental

evidence—but the hypothesis survives and may even be strengthened by test and refinement in the face of doubt. But equally often a hypothesis is false, in which case clinging to it is a waste of time. Persist for long enough to establish whether or not it is likely to be true, but to persist longer is foolish.

A corollary is that the stronger your intuitive liking for a hypothesis, the more rigorously you should test it—that is, attempt to confirm it or disprove it—rather than twist results, and yourself, defending it.

Be persuasive. Using research into the properties of an algorithm as an example, issues such as the following need to be addressed.

- Will the reader believe that the algorithm is new?

  Only if the researcher does a careful literature review, and fully explores and explains previous relevant work. Doing so includes giving credit to significant advances, and not overrating work where the contribution is small.

- Will the reader believe that the algorithm is sensible?

  It had better be explained carefully. Potential problems should be identified, and either conceded—with an explanation, for example, of why the algorithm is not universally applicable—or dismissed through some cogent argument.

- Are the experiments convincing?

  If the code isn't good enough to be made publicly available, is it because there is something wrong with it? Has the right data been used? Has enough data been used?

Every research program suggests its own skeptical questions. Such questioning is also appropriate later in a research program, where it gives the author an opportunity to make a critical assessment of the work.

Forms of Evidence

A paper can be viewed as an assembly of evidence and supporting explanations; that is, as an attempt to persuade others to share your conclusions. Good science uses objective evidence to achieve aims such as to persuade readers to make more informed decisions and to deepen their understanding of problems and solutions. In a write-up you pose a question or hypothesis, then present evidence to support your case. The evidence needs to be convincing because the processes of science rely on readers being critical and skeptical; there is no reason for a reader to be interested in work that is inconclusive.

There are, broadly speaking, four kinds of evidence that can be used to support a hypothesis: proof, modelling, simulation, and experiment.

Proof. An proof is a formal argument that a hypothesis is correct (or wrong). It is a mistake to suppose that the correctness of a proof is absolute—confidence in a proof may be high, but that does not guarantee that it is free from error; it is common

for a researcher to feel certain that a theorem is correct but have doubts about the mechanics of the proof.[1]

Some hypotheses are not amenable to formal analysis, particularly hypotheses that involve the real world in some way. For example, human behaviour is intrinsic to questions about interface design, and system properties can be intractably complex. Consider an exploration to determine whether a new method is better than a previous one at lossless compression of images—is it likely that material that is as diverse as images can be modelled well enough to predict the performance of a compression algorithm? It is also a mistake to suppose that an asymptotic analysis is always sufficient. Nonetheless, the possibility of formal proof should never be overlooked.

Model. A model is a mathematical description of the hypothesis (or some component of the hypothesis, such as an algorithm whose properties are being considered) and there will usually be a demonstration that the hypothesis and model do indeed correspond.

In choosing to use a model, consider how realistic it will be, or conversely how many simplifying assumptions need to be made for analysis to be feasible. Take the example of modelling the cost of a Boolean query on a text collection, in which the task is to find the documents that contain each of a set of words. We need to estimate the frequency of each word (because words that are frequent in queries may be rare in documents); the likelihood of query terms occurring in the same document (in practice, query terms are thematically related, and do not model well as random co-occurrences); the fact that longer documents contain more words, but are more expensive to fetch; and, in a practical system, the probability that the same query had been issued recently and the answers are cached in memory. It is possible to define a model based on these factors, but, with so many estimates to make and parameters to tune, it is unlikely that the model would be realistic.

Simulation. A simulation is usually an implementation or partial implementation of a simplified form of the hypothesis, in which the difficulties of a full implementation are sidestepped by omission or approximation. At one extreme a simulation might be little more than an outline; for example, a parallel algorithm could be tested on a sequential machine by use of an interpreter that counts machine cycles and communication costs between simulated processors; at the other extreme a simulation could be an implementation of the hypothesis, but tested on artificial data. A simulation is a "white coats" test: artificial, isolated, and conducted in a tightly controlled environment.

A great advantage of a simulation is that it provides parameters that can be smoothly adjusted, allowing the researcher to observe behaviour across a wide spectrum of inputs or characteristics. For example, if you are comparing algorithms for removal of errors in genetic data, use of simulated data might allow you to control the error rate, and observe when the different algorithms begin to fail. Real data may have unknown numbers of errors, or only a couple of different error rates, so in some sense can be less informative. However, with a simulation there is always the risk

[1] Which can, of course, lead to the discovery that the theorem is wrong after all.

that it is unrealistic or simplistic, with properties that mean that the observed results would not occur in practice. Thus simulations are powerful tools, but, ultimately, need to be verified against reality.

Experiment. An experiment is a full test of the hypothesis, based on an implementation of the proposal and on real—or highly realistic—data. In an experiment there is a sense of *really doing it*, while in a simulation there is a sense of *only pretending*. For example, artificial data provides a mechanism for exploring behaviour, but corresponding behaviour needs to be observed on real data if the outcomes are to be persuasive.

In some cases, though, the distinction between simulation and experiment can be blurry, and, in principle, an experiment only demonstrates that the hypothesis holds for the particular data that was used; modelling and simulation can generalize the conclusion (however imperfectly) to other contexts.

Ideally an experiment should be conducted in the light of predictions made by a model, so that it confirms some expected behaviour. An experiment should be severe; seek out tests that seem likely to fail if the hypothesis is false, and explore extremes. The traditional sciences, and physics in particular, proceed in this way. Theoreticians develop models of phenomena that fit known observations; experimentalists seek confirmation through fresh experiments.

Use of Evidence

Different forms of evidence can be used to confirm one another, with say a simulation used to provide further evidence that a proof is correct. But the different forms should not be confused with one another. For example, suppose that for some algorithm there is a mathematical model of expected performance. Encoding this model in a program and computing predicted performance for certain values of the model parameters is not an experimental test of the algorithm and should never be called an experiment; it does not even confirm that the model is a description of the algorithm. At best it confirms claimed properties of the model.

When choosing whether to use a proof, model, simulation, or experiment as evidence, consider how convincing each is likely to be to the reader. If your evidence is questionable—say a simplistic and assumption-laden model, an involved algebraic analysis and application of advanced statistics, or an experiment on limited data—the reader may well be skeptical of the result. Select a form of evidence, not so as to keep your own effort to a minimum, but to be as persuasive as possible.

Having identified the elements a research plan should cover, end-to-start reasoning suggests how these elements should be prioritized. The write-up is the most important thing; so perhaps it should be started first. Completing the report is certainly more important than hastily running some last-minute experiments, or quickly browsing the literature to make it appear as if past work has been fully evaluated.

Some novice researchers feel that the standards expected of evidence are too high, but readers—including referees and examiners—tend to trust work that is already published in preference to a new, unrefereed paper, and have no reason to trust work

where the evidence is thin. Moreover, experienced researchers are well aware that skepticism is justified. It has been said, with considerable truth, that most published research findings are false; and unpublished findings are worse.

This means that a paper must be persuasive. Your written work is the one chance to persuade readers to accept the ideas, and they will only do so if the evidence and arguments are complete and convincing.

Approaches to Measurement

A perspective on the history of science is that it is also a history of development of tools of measurement. Our understanding of the laws of physics followed from development of telescopes, voltmeters, thermometers, and so on. Each improvement in the measurement technology has refined our understanding of the underlying properties of the universe.

From this perspective, the purpose of experimentation is to take measurements that can be used as evidence. A good choice of measure is essential to to practical system improvement and to persuasive and insightful writing. The measurements are intended to be a consequence of some underlying phenomenon that is described by a theory or hypothesis. In this approach to research, phenomena—the eternal truths studied by science—cannot change, but the measurements can, because they depend on the context of the specific experiment. Measurements can be quantitative, such as number or duration or volume—the speed of a system, say, or an algorithm's efficiency relative to a baseline. They can also be qualitative, such as an occurrence or difference—whether an outcome was achieved, or whether particular features were observed.

As you develop your research questions, then, you should ask *what is to be measured?* and *what measures will be used?* For example, when examining an algorithm, will it be measured by execution time? And if so, what mechanism will be used to measure it? This question can be tricky to answer for a single-threaded process running on a single machine. For a distributed process using diverse resources across a network, there probably is no perfect answer, only a range of choices with a variety of flaws and shortcomings, each of which needs to be understood by you and by your readers.

There is then a critical, but more subtle, question: you need to be satisfied that the properties being measured are logically connected to the aims of the research. Typically, research aims are *qualitative*. We seek to improve an interface, accelerate an algorithm, extract information from an image, generate better timetables for lectures, and so on. Measurement is *quantitative*; we find a property that can be represented as a quantity or value. For example, the effectiveness of machine translation systems is sometimes assessed by counting the textual overlap (words or substrings) of a computer translation with that made by a human. However, such a measure is obviously imperfect: not only are there many possible human translations, but a highly overlapping text can still be incoherent, that is, not a good translation.

As another example, we might say that the evidence for the claim that a network is qualitively improved is that average times to transmit a packet are reduced—a quantity that can be measured. But if the aim of network improvement is simplified to the goal of reducing wait times, then other aspects of the qualitative aim (smoothness of transmission of video, say, or effectiveness of service for remote locations) may be neglected.

In other words, once a qualitative aim is replaced by a single quantitative measure, the goal of research in the field can shift away from achievement of a practical outcome, and instead consist entirely of optimization to the measure, regardless of how representative the measure is of the broader problem. A strong research program will rest, in part, on recognition of the distinction between qualitative goals and different quantitative approximations to that goal.

The problem of optimization-to-a-measure is particularly acute for fields that make use of shared reference data sets, where this data is used for evaluation of new methods. It is all too easy for researchers to begin to regard the standard data as being representative of the problem as a whole, and to tune their methods to perform well on just these data sets. Any field in which the measures and the data are static is at risk of becoming stagnant.

Good and Bad Science

Questions about the quality of evidence can be used to evaluate other people's research, and provide an opportunity to reflect on whether the outcomes of your work are worthwhile. There isn't a simple division of research into "good" and "bad", but it is not difficult to distinguish valuable research from work that is weak or pointless.

The merits of formal studies are easy to appreciate. They provide the kind of mathematical link between the possible and the practical that physics provides between the universe and engineering.

The merits of well-designed experimental work are also clear. Work that experimentally confirms or contradicts the correctness of formal studies has historically been undervalued in computer science: perhaps because standards for experimentation have not been high; perhaps because the great diversity of computer systems, languages, and data has made truly general experiments difficult to devise; or perhaps because theoretical work with advanced mathematics is more intellectually imposing than work that some people regard as mere code-cutting. However, many questions cannot be readily answered through analysis, and a theory without practical confirmation is of no more interest in computing than in the rest of science.

Research that consists of proposals and speculation, entirely without a serious attempt at evaluation, can be more difficult to respect. Why should a reader regard such work as valid? If the author cannot offer anything to measure, arguably it isn't science. And research isn't "theoretical" just because it isn't experimental. Theoretical work describes testable theories.

The quality of work can be unclear if the terminology used to describe it is over-inflated. Sometimes such terminology is used to avoid having to define terms properly. A *hypernet*, for example, sounds much more powerful than a network; but who knows if there is really a difference. Researchers use such terminology to make cloudy, big-picture claims that are rarely justified by their actual outcomes.

Terms that in common usage describe aspects of cognition or consciousness, such as "intelligent" or "belief", or even "aware", are particularly slippery. They sound like ordinary concepts we are all familiar with. But in their common usage they are not well defined; and when terms are borrowed from common usage their meaning changes. These terms anthropomorphize the computational behaviour to create a sense of specialness or drama, when in fact what is being described may well be highly mechanical and deterministic, and possibly isn't very interesting. This is a form of the renaming fallacy noted earlier. Thus, while we might have an impression of what the author means when they claim that a system is intelligent, that impression is vague. Successful science is not built on vagueness.

A particular example is the widely misused term "semantic", which, strictly speaking, concerns the meaning of a concept, as distinct from its syntax or representation. But computers are machines for processing representations: enriching the representation by, say, addition of further descriptors does not "bridge the semantic gap". At best, it shifts the problem, from one of computing in the absence of descriptors to one of creating and then making use of descriptors. For example, a text indexing technique that "maps terms to concepts, allowing semantic retrieval" might be no more than a trivial function in which an ontology is used to map words, correctly or otherwise, to labels; retrieval then proceeds as usual, but with labels instead of terms as queries. An additional resource (a dictionary) has been introduced into the process, but the method isn't semantic, and certainly isn't particularly intelligent.

Some science is not simply weak, but can be described as pseudoscience. Inevitably, some claimed achievements are delusional or bogus. Pseudoscience is a broad label covering a range of scientific sins, from self-deception and confusion to outright fraud. A definition is that pseudoscience is work that uses the language and respectability of science to gain credibility for statements that are not based on evidence that meets scientific standards. Much pseudoscience shares a range of characteristics: the results and ideas don't seem to develop over time, systems are never quite ready for demonstration, the work proceeds in a vacuum and is unaffected by other advances, protagonists argue rather than seek evidence, and the results are inconsistent with accepted facts. Often such work is strenuously promoted by one individual or a small number of devotees while the rest of the scientific community ignores it.[2]

[2] An example of pseudoscience in computing are schemes for high-performance video compression that promised delivery of TV-quality data over low-bandwidth modems. In the 1990s, the commercial implications of such systems were enormous, and this incentive created ample opportunities for fraud. In one case, for example, millions of dollars were scammed from investors with tricks such as hiding a video player inside a PC tower and hiding a network cable inside a power cable. Yet, skeptically considered, such schemes are implausible. For example, with current technology, even a corner of a single TV-resolution image—let alone 25 frames—cannot be compressed into

An example is what might be described as "universal" indexing methods. In such methods, the object to be indexed—whether an image, movie, audio file, or text document—is manipulated in some way, for example by a particular kind of hash function. After this manipulation, objects of different type can be compared: thus, somehow, documents about swimming pools and images of swimming pools would have the same representation. Such matching is clearly an extremely difficult problem, if not entirely insoluble; for instance, how does the method know to focus on the swimming pool rather than some other element of the image, such as children, sunshine, or its role as a metaphor for middle-class aspirations?[3]

In some work, the evidence or methods are internally inconsistent. For example, in a paper on how to find documents on a particular topic, the authors reported that the method correctly identified 20,000 matches in a large document collection. But this is a deeply improbable outcome. The figure of 20,000 hints at imprecision—it is too round a number. More significantly, verifying that all 20,000 were matches would require many months of effort. No mention was made of the documents that weren't matches, implying that the method was 100 % accurate; but even the best document-matching methods have high error rates. A later paper by the same authors gave entirely different results for the same method, while claiming similar good results for a new method, thus throwing doubt on the whole research program. And it is a failure of logic to suppose that the fact that two documents match according to some arbitrary algorithm implies that the match is useful to a user.

The logic underlying some papers is outright mystifying. To an author, it may seem a major step to identify and solve a new problem, but such steps can go too far. A paper on retrieval for a specific form of graph used a new query language and matching technique, a new way of evaluating similarity, and data based on a new technique for deriving the graphs from text and semantically (that word again!) labelling the edges. Every element of this paper was a separate contribution whose merit could be disputed. Presented in a brief paper, the work seemed worthless. Inventing a problem, a solution to the problem, and a measure of the solution—all without external justification—is a widespread form of bad science.[4]

(Footnote 2 continued)
the 7 kilobytes that such a modem could transmit per second. Uncompressed, the bandwidth of a modem was only sufficient for one byte per row per image, or, per image, about the space needed to transmit a desktop icon. A further skeptical consideration in this case was that an audio signal was also transmitted. Had the system been legitimate, the inventor must have developed new solutions to the independent problems of image compression, motion encoding, and audio compression.

[3] In another variant of this theme, objects of the same type were clustered together using some kind of similarity metric. Then the patterns of clustering were analyzed, and objects that clustered in similar ways were supposed to have similar subject matter. Although it is disguised by the use of clustering, to be successful such an approach assumes an underlying universal matching method.

[4] An interesting question is how to regard "Zipf's law". This observation—"law" seems a poor choice of terminology in this context—is if nothing else a curious case study. Zipf's books may be widely cited but they are not, I suspect, widely read. In *Human Behaviour and the Principle of Least Effort* (Addison-Wesley, 1949), Zipf used languages and word frequencies as one of several examples to illustrate his observation, but his motivation for the work is not quite what might be expected. He states, for example, that his research "define[s] objectively what we mean by the term

We need to be wary of claimed results, not only because we might disagree for technical reasons but because the behaviour of other researchers may not be objective or reasonable. Another lesson is that acceptance of (or silence about) poor science erodes the perceived need for responsible research, and that it is always reasonable to ask skeptical questions. Yet another lesson is that we need to take care to ensure that our own research is well founded.

Reflections on Research

Philosophers and historians of science have reflected at length on the meaning, elements, and methods of research, from both practical and abstract points of view. While philosophy can seem remote from the practical challenges of research, these reflections can be of great benefit to working scientists, who can learn from an overall perspective on their work. Being able to describe what we do helps us to understand whether we are doing it well.

Such philosophies and definitions of science help to establish guidelines for the practical work that scientists do, and set boundaries on what we can know. However, there are limits to how precise (or interesting) such definitions can be. For example, the question "is computer science a science?" has a low information content.[5] Questions of this kind are sometimes in terms of definitions of science such as "a process for discovering laws that model observed natural phenomena". Such definitions not only exclude disciplines such as computing, but also exclude much of the research now undertaken in disciplines such as biology and medicine. In considering definitions of science, a certain degree of skepticism is valuable; these definitions are made by scientists working within particular disciplines and within the viewpoints that those disciplines impose.[6]

(Footnote 4 continued)
personality" (p. 18), explains the "drives of the Freudian death wish" (p. 17), and "will provide an objective language in terms of which persons can discuss social problems impersonally" (p. 543). It "will help to protect mankind from the virtual criminal action of persons in strategic political, commercial, social, intellectual and academic positions" (p. 544) and "as the authority of revealed religion and its attendant ethics declines, something must take its place … I feel that this type of research may yield results that will fulfill those needs" (p. 544). Perhaps these extraordinary claims are quirks, and in any case opinions do not invalidate scientific results. But it has been argued that the behaviour captured by Zipf's conjecture is a simple consequence of randomness, and, for the example for which the conjecture is often cited (distribution of words in text), the fit between hypothesis and observation is not always strong.

[5] Two philosophers are arguing in a bar. The barman goes over to them and asks, "What are you arguing about?"

"We're debating whether computer science is a science", answers one of them.
"And what do you conclude?" asks the barman.
"We're not sure yet," says the other. "We can't agree on what 'is' means".

[6] But, in fairness, the views here have the same limitations, as they are those of a computer scientist who believes that the discipline stands alongside the traditional sciences.

It is true that, considered as a science, computing is difficult to categorize. The underlying theories—in particular, information theory and computability—appear to describe properties as eternal as those of physics. Yet much research in computer science is many steps removed from foundational theory and more closely resembles engineering or psychology.

A widely agreed description of science is that it is a method for accumulating reliable knowledge. In this viewpoint, scientists adopt the belief that rationality and skepticism are how we learn about the universe and shape new principles, while recognizing that this belief limits the application of science to those ideas that can be examined in a logical way. If the arguments and experiments are sound, if the theory can withstand skeptical scrutiny, if the work was undertaken within a framework of past research and provides a basis for further discovery, then it is science. Much computer science has this form.

Many writers and philosophers have debated the nature of science, and aspects of science such as the validity of different approaches to reasoning. The direct impact of this debate on the day-to-day activity of scientists is small, but it has helped to shape how scientists approach their work. It also provides elements of the ethical framework within which scientists work.

One of the core concepts is *falsification*: experimental evidence, no matter how substantial or voluminous, cannot prove a theory true, while a single counter-example can prove a theory false. A practical consequence of the principle of falsification is that a reasonable scientific method is to search for counter-examples to hypotheses. In this line of reasoning, to search for supporting evidence is pointless, as such evidence cannot tell us that the theory is true. A drawback of this line of reasoning is that, using falsification alone, we cannot learn any new theories; we can only learn that some theories are wrong. Another issue is that, in practice, experiments are often unsuccessful, but the explanation is not that the hypothesis is wrong, but rather that some other assumption was wrong—the response of a scientist to a failed experiment may well be to redesign it. For example, in the decades-long search for gravity waves, there have been many unsuccessful experiments, but a general interpretation of these experiments has been that they show that the equipment is insufficiently sensitive.

Thus falsification can be a valuable guide to the conduct of research, but other guides are also required if the research is to be productive. One such guide is the concept of *confirmation*. In science, confirmation has a weaker meaning than in general usage; when a theory is confirmed, the intended meaning is not that the theory is proved, but that the weight of belief in the theory has been strengthened. Seeking of experiments that confirm theories is an alternative reasonable view of scientific method.

A consequence is that a hypothesis should allow some possibility of being disproved—there should be some experiment whose outcomes could show that they hypothesis is wrong. If not, the hypothesis is simply uninteresting. Consider, for example, the hypothesis “a search engine can find interesting Web pages in response to queries”. It is difficult to see how this supposition might be contradicted.

In the light of these descriptions, science can be characterized as an iterative process in which theory and hypothesis dictate a search for evidence—or “facts”—

while we learn from facts and use them to develop theories. But we need initial theories to help us search for facts.

Thus confirmation, falsification, and other descriptions of method help to shape research questions as well as research processes, and contribute to the practice of science. We need to be willing to abandon theories in the face of contradictions, but flexible in response to failure; contradictions may be due to an incorrect hypothesis, faulty experimental apparatus, or poor measurement of the experimental outcomes. We need to be ready to seek plausible alternative explanations of facts or observations, and to find experiments that yield observations that provide insight into theories. That is, theories and evidence are deeply intertwined. A scientific method that gives one primacy over the other is unlikely to be productive, and, to have high impact, our research programs should be designed so that theory and evidence reinforce each other.

A "Hypotheses, Questions, and Evidence" Checklist

Regarding *hypotheses and questions*,

- What phenomena or properties are being investigated? Why are they of interest?
- Has the aim of the research been articulated? What are the specific hypotheses and research questions? Are these elements convincingly connected to each other?
- To what extent is the work innovative? Is this reflected in the claims?
- What would disprove the hypothesis? Does it have any improbable consequences?
- What are the underlying assumptions? Are they sensible?
- Has the work been critically questioned? Have you satisfied yourself that it is sound science?

Regarding *evidence and measurement*,

- What forms of evidence are to be used? If it is a model or a simulation, what demonstrates that the results have practical validity?
- How is the evidence to be measured? Are the chosen methods of measurement objective, appropriate, and reasonable?
- What are the qualitative aims, and what makes the quantitative measures you have chosen appropriate to those aims?
- What compromises or simplifications are inherent in your choice of measure?
- Will the outcomes be predictive?
- What is the argument that will link the evidence to the hypothesis?
- To what extent will positive results persuasively confirm the hypothesis? Will negative results disprove it?
- What are the likely weaknesses of or limitations to your approach?